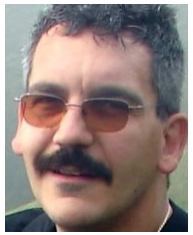
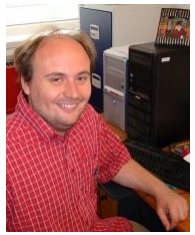
Laboratory of Chemoinformatics
UMR 7140 CNRS/UniStra

Laboratory of Chemoinformatics in October 2022


A. Varnek (Professor)


D. Horvath (DR2 CNRS)


G. Marcou (MCF)


F. Bonachera (IR CNRS)


O. Klimchuk (IE UniStra)

+ 1 Postdoc + 10 PhD students

« Mirror » Laboratory


R. Staub (Ass. Professor)


P. Gantzer (postdoc)

Why Chemoinformatics?

- **Chemoinformatics is a major discipline in theoretical chemistry**

  Molecular informatics, *30*(1), 20-32

- **"Chemoinformatics" and "Cheminformatics" emerged after 1997**

  Thomson's Web of Science database

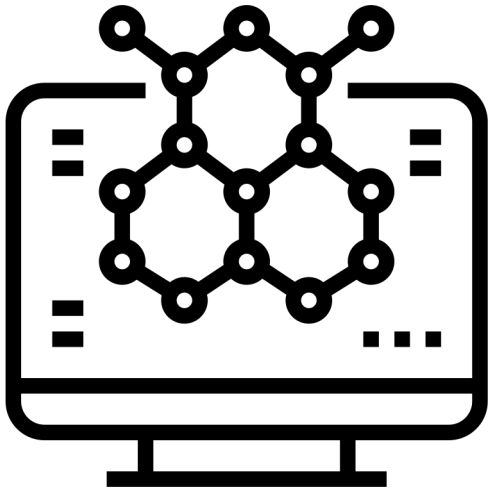
- **Social and innovation challenges**
  - ✔ Exponential growth of chemical information

    Chemical Abstract Service. 10th million chemical substance in 1990, 196 million substances in 2023
  - ✔ Research and development concern

    Novartis first chemoinformatics tools in 1995
  - ✔ REACH directive

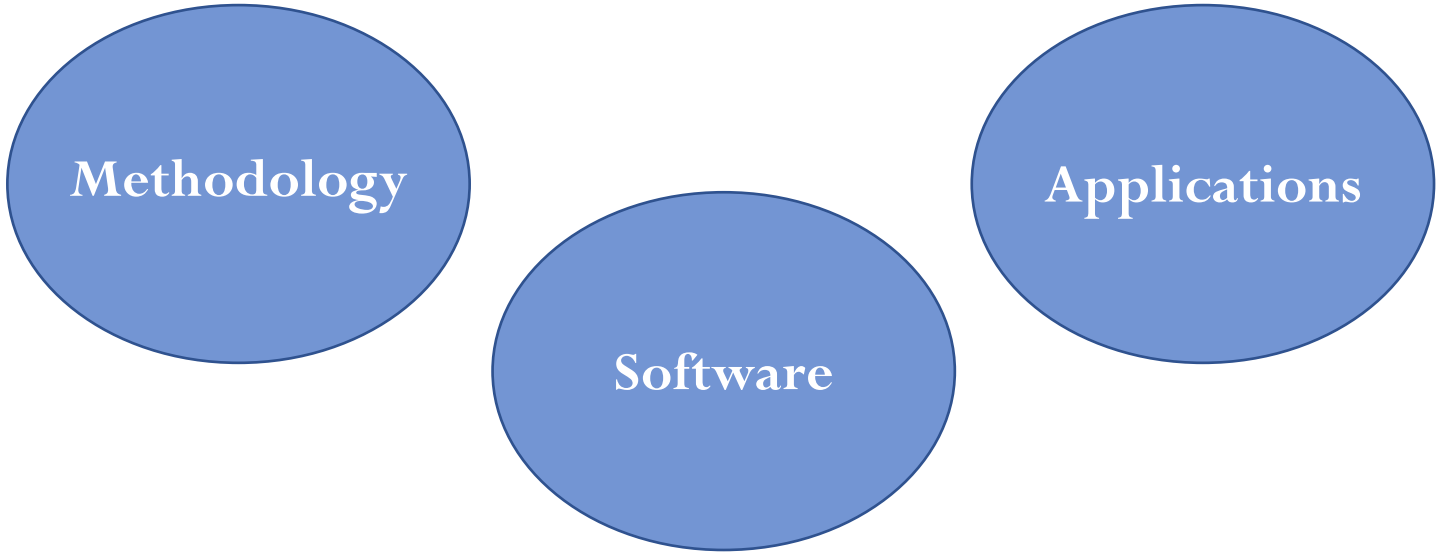
    Since 01/06/2008, 22395 substances have been registered
  - ✔ US Environmental Protection Agency under the (new) Toxic Substances Control Act

    45031 cases since 1971
  - ✔ chemoinformatics approaches have been proposed by OECD since



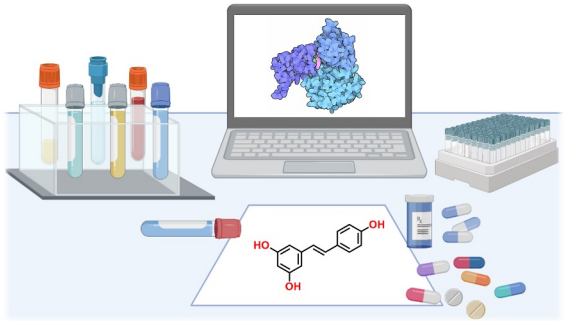
ChemOinformatique
STRASBOURG

Approaches and tools for computer-aided molecular design



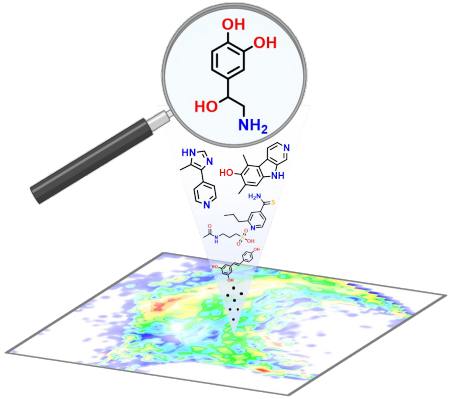
4 key research areas

1 AI-driven design of new compounds and materials



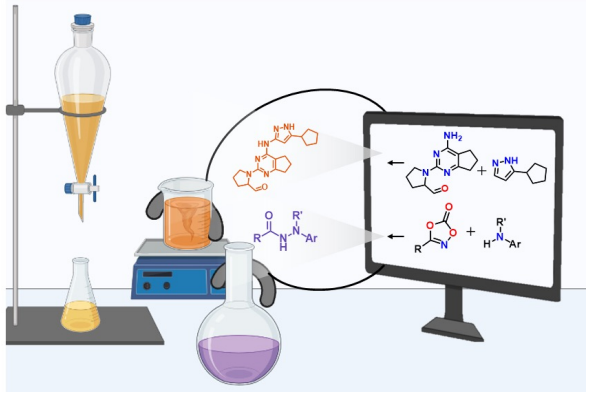
Chem. Soc. Rev., **2020**, *49*, 3525
J. Chem. Inf. Model., **2019**, *59*, 4569

2 Analysis of ultra-large chemical spaces



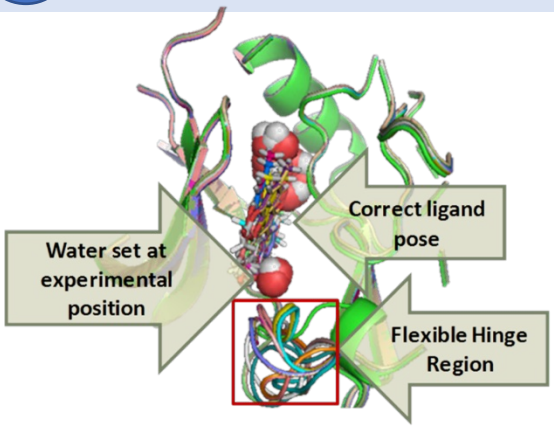
J. Chem. Inf. Model. **2021**, *61*, 179
Drug Disc. Today: Technology, **2019**, *32*, 99

3 Chemical reactions mining



Scientific Reports, **2021**, *11*, 3178
J. Chem. Inf. Model., **2021**, *61* , 554

4 *De novo* design



J Med Chem. **2018**, *61*, 5719
J. Chem. Inf. Model., **2019**, *59*, 1472

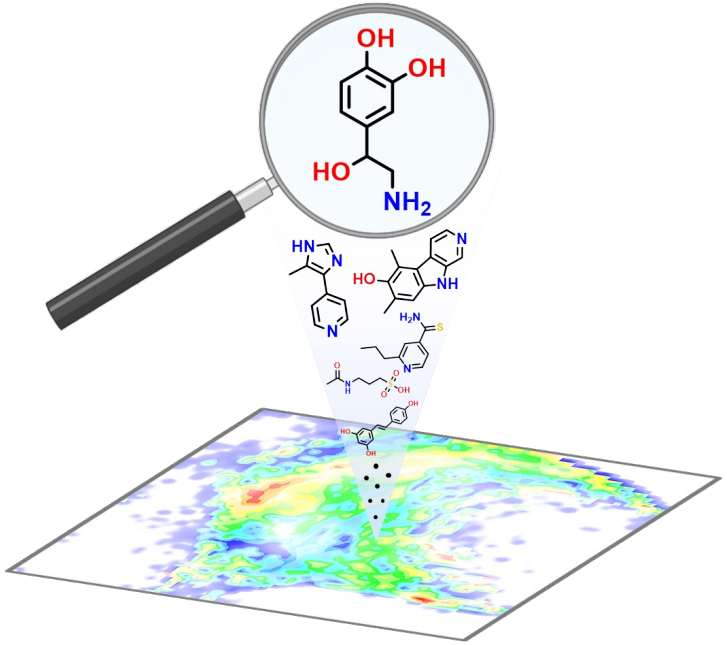
Competitive advantage:

- Original methodology for predictive structure-property modeling and *de novo* design

Potential applications:

- Development of new synthetically feasible molecules and materials possessing desirable properties

Analysis of ultra-large chemical spaces using molecular cartography



Generative Topographic Mapping

- visualization of both individual data (chemical structures) and their probability distribution on a 2D map;
- molecules populating a given area on the map are expected to have similar properties;
- prediction of properties of new molecules projected on the map which, therefore, can be used as a virtual screening tool;
- suitable for Big Data analysis (billions of molecules)

Case study:

- Comparison of commercial and biologically relevant chemical spaces

ChemSpace Atlas

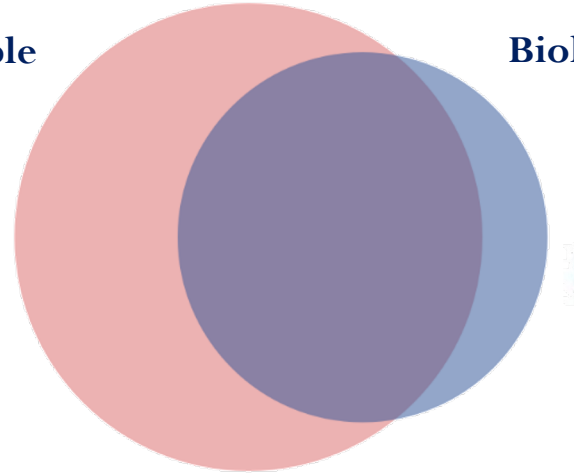
Analysis of ultra-large chemical spaces

Collaboration with Inst. Org. Chem., Kiev, Ukraine

Commercially available chemotypes

ZINC DATABASE

>1.3 billion cmpds



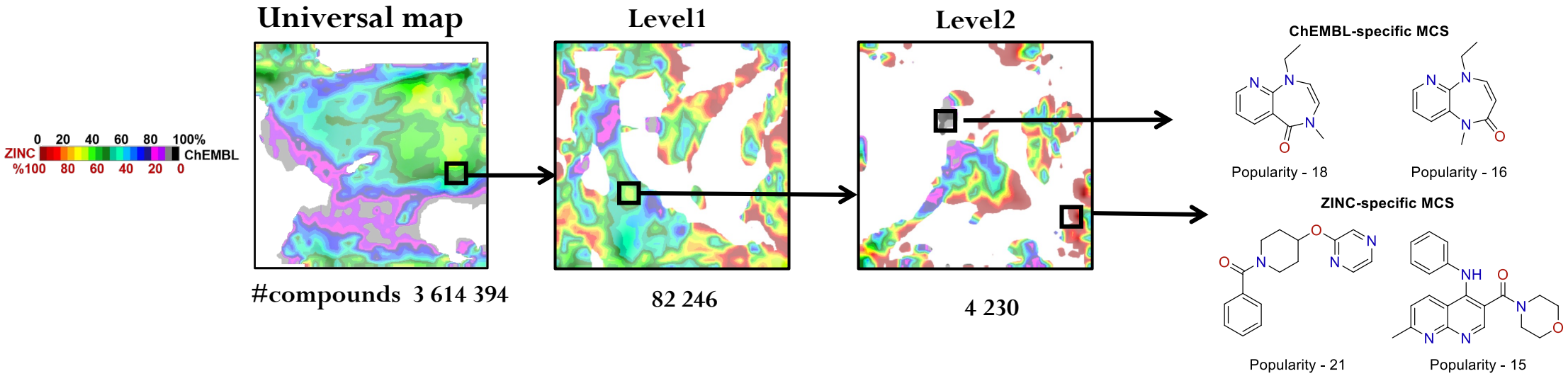
Biologically relevant chemotypes

ChEMBL

>1.8 M cmpds

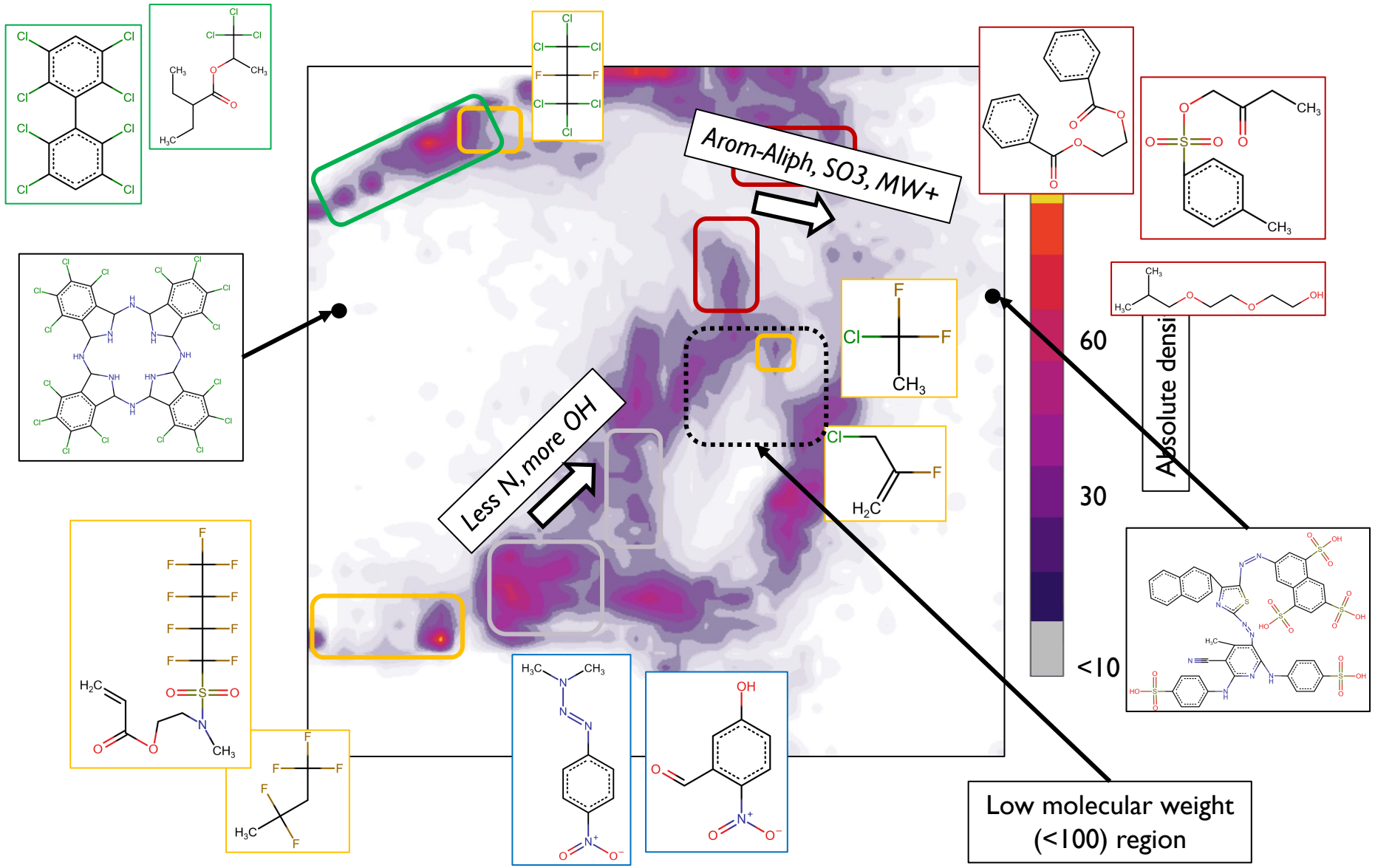
> **100K** chemotypes which have never been biologically tested

> **20 K** chemotypes missing from the commercial chemical space

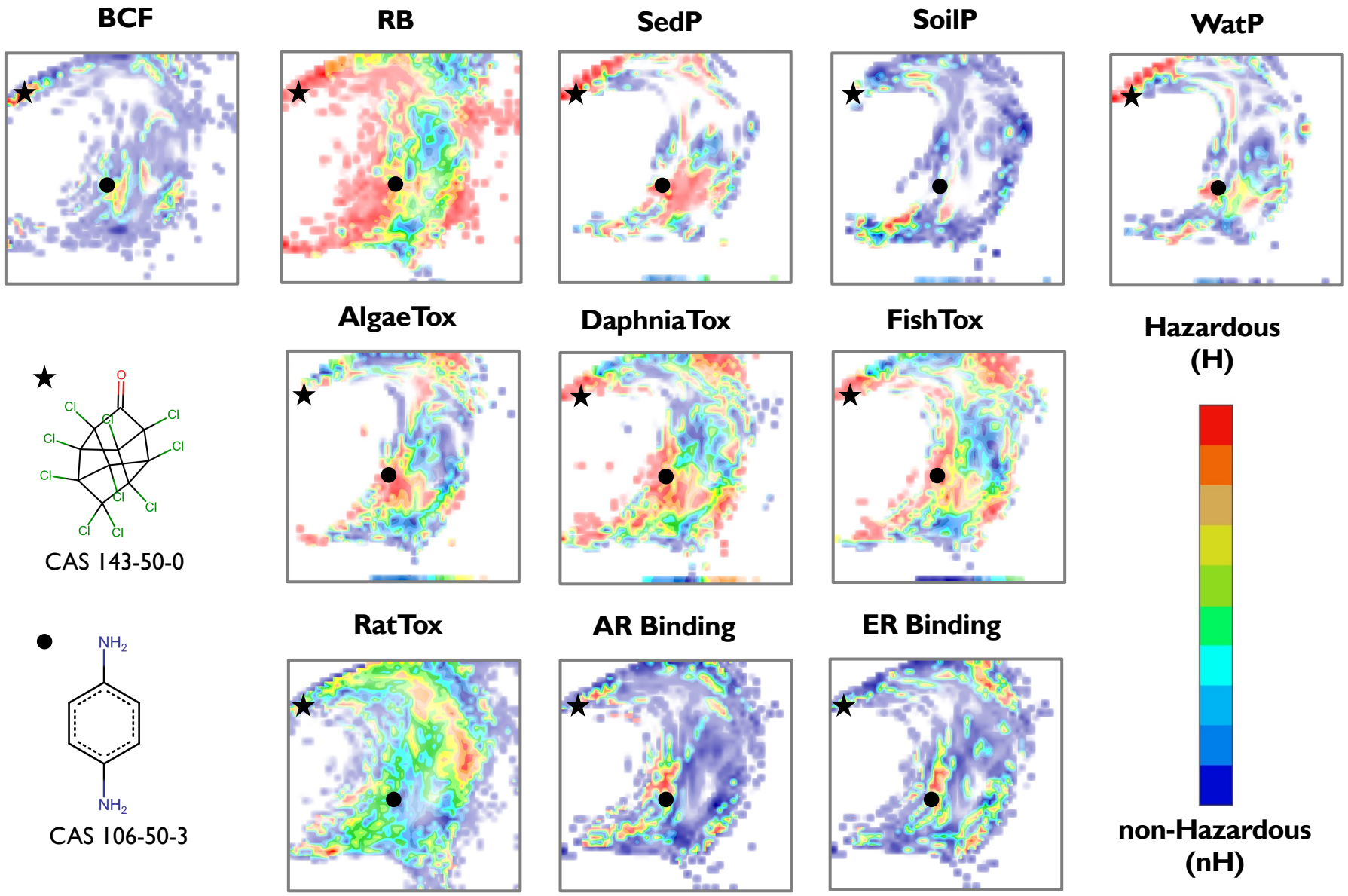


Yu. Zabolotna *et al.*, *« Searching for hidden treasures »*, *J. Chem. Inf. Model.* 2021, 61, 1, 179–188

REACH-chemical space

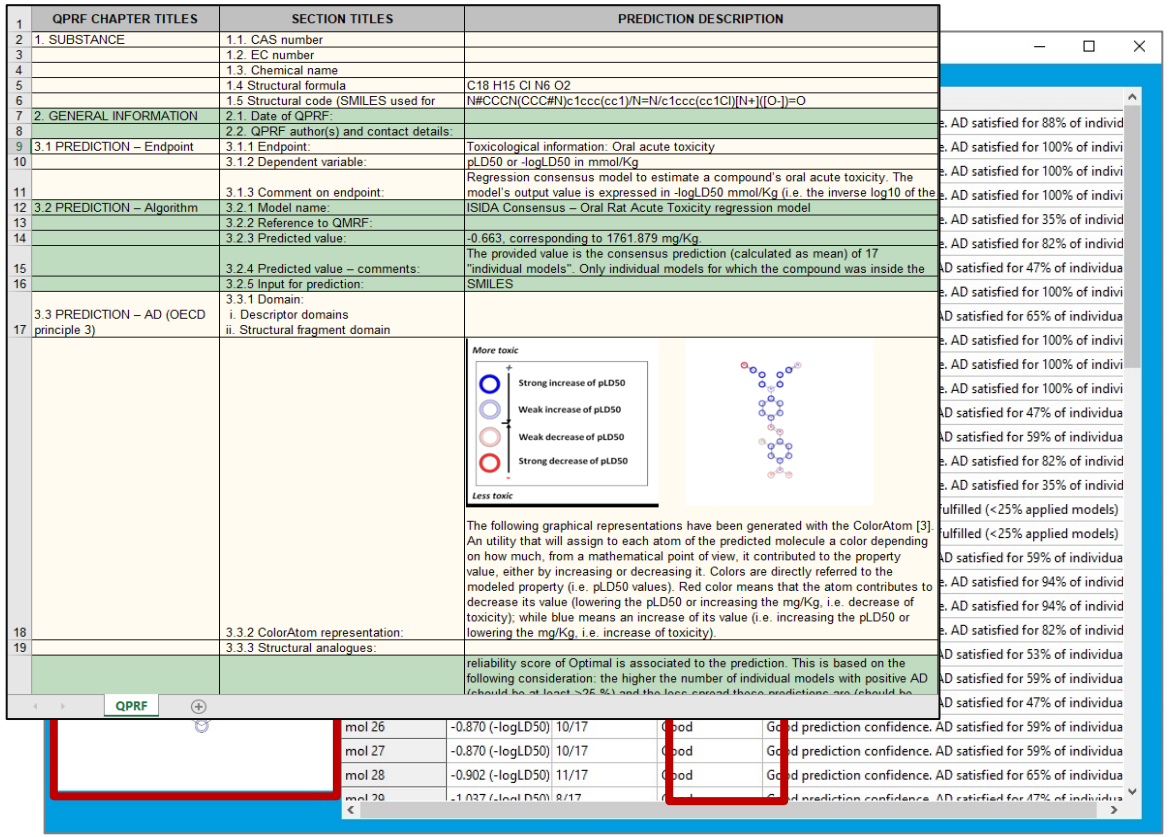


New compounds profiling



ISIDA/Predictor

- Clear **reliability evaluation**
- **Mechanistic interpretation**: ColorAtom
- **Automatic QPRF** (QSAR Prediction Reporting Format) generation



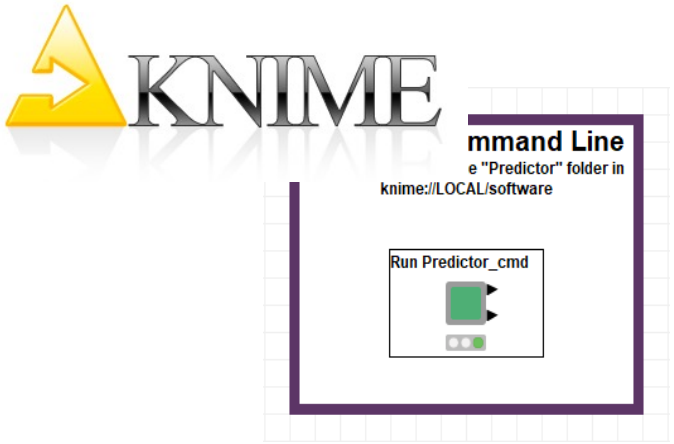
	QPRF CHAPTER TITLES	SECTION TITLES	PREDICTION DESCRIPTION
1			
2	1. SUBSTANCE	1.1. CAS number	
3		1.2. EC number	
4		1.3. Chemical name	
5		1.4 Structural formula	C18 H15 Cl N6 O2
6		1.5 Structural code (SMILES used for	N#CCCN(CCC#N)c1ccc(cc1)/N=N/c1ccc(cc1Cl)[N+]([O-])=O
7	2. GENERAL INFORMATION	2.1. Date of QPRF:	
8		2.2. QPRF author(s) and contact details:	
9	3.1 PREDICTION – Endpoint	3.1.1 Endpoint:	Toxicological information: Oral acute toxicity
10		3.1.2 Dependent variable:	pLD50 or -logLD50 in mmol/Kg
11		3.1.3 Comment on endpoint:	Regression consensus model to estimate a compound's oral acute toxicity. The model's output value is expressed in -logLD50 mmol/Kg (i.e. the inverse log10 of the
12	3.2 PREDICTION – Algorithm	3.2.1 Model name:	ISIDA Consensus – Oral Rat Acute Toxicity regression model
13		3.2.2 Reference to QMRF:	
14		3.2.3 Predicted value:	-0.663, corresponding to 1761.879 mg/Kg.
15		3.2.4 Predicted value – comments:	The provided value is the consensus prediction (calculated as mean) of 17 "individual models". Only individual models for which the compound was inside the
16		3.2.5 Input for prediction:	SMILES
17	3.3 PREDICTION – AD (OECD principle 3)	3.3.1 Domain: i. Descriptor domains ii. Structural fragment domain	
18		3.3.2 ColorAtom representation:	More toxic Strong increase of pLD50 Weak increase of pLD50 Weak decrease of pLD50 Strong decrease of pLD50 Less toxic The following graphical representations have been generated with the ColorAtom [3]. An utility that will assign to each atom of the predicted molecule a color depending on how much, from a mathematical point of view, it contributed to the property value, either by increasing or decreasing it. Colors are directly referred to the modeled property (i.e. pLD50 values). Red color means that the atom contributes to decrease its value (lowering the pLD50 or increasing the mg/Kg, i.e. decrease of toxicity); while blue means an increase of its value (i.e. increasing the pLD50 or lowering the mg/Kg, i.e. increase of toxicity).
19		3.3.3 Structural analogues:	
			reliability score of Optimal is associated to the prediction. This is based on the following consideration: the higher the number of individual models with positive AD

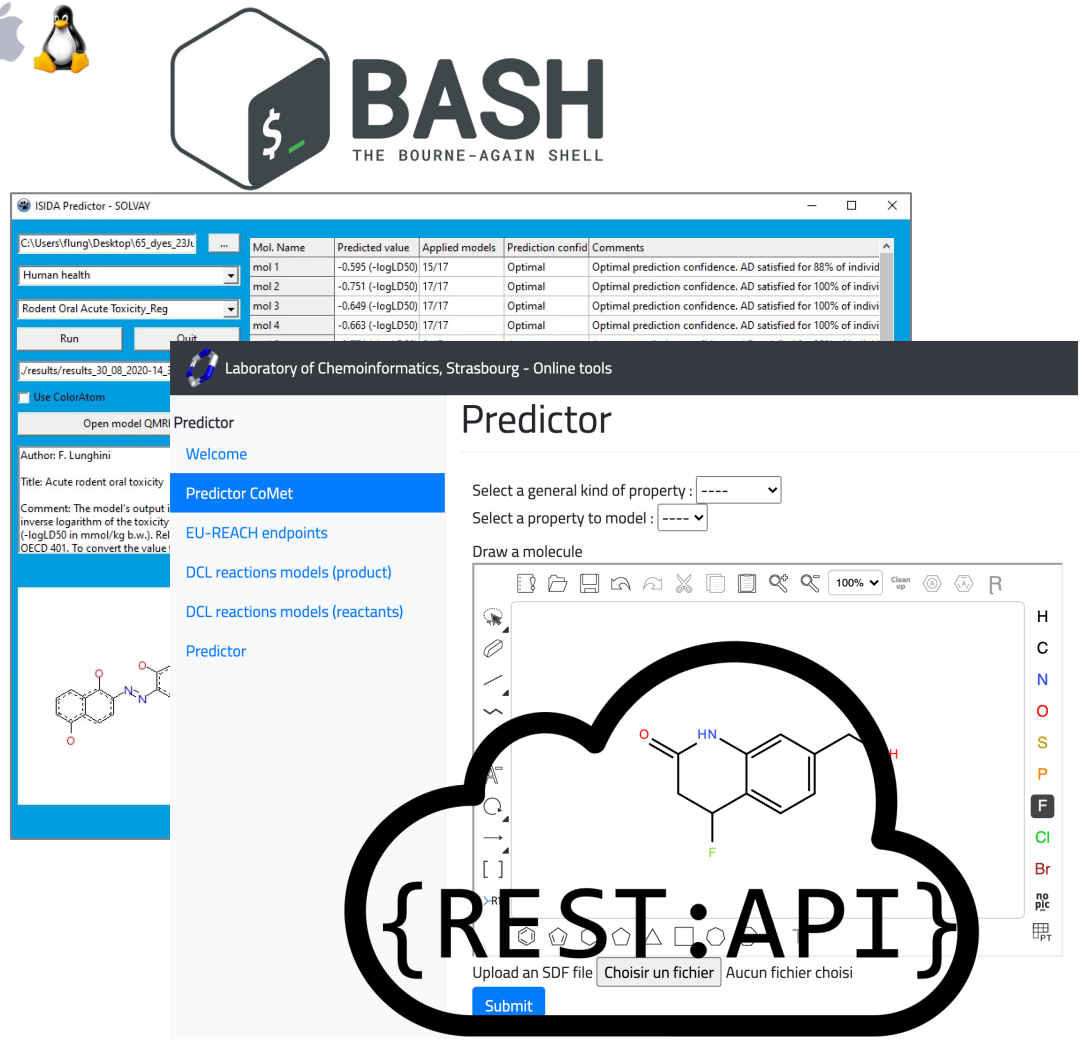
mol 26	-0.870 (-logLD50)	10/17	Good	Good prediction confidence.
mol 27	-0.870 (-logLD50)	10/17	Good	Good prediction confidence.
mol 28	-0.902 (-logLD50)	11/17	Good	Good prediction confidence.

*http://infochim.u-strasbg.fr/cgi-bin/predictor_reach.cgi

Specifications

- Supported Systems
- Command line
- Graphical User Interface
- Web Interface
- REST
- KNIME Integration

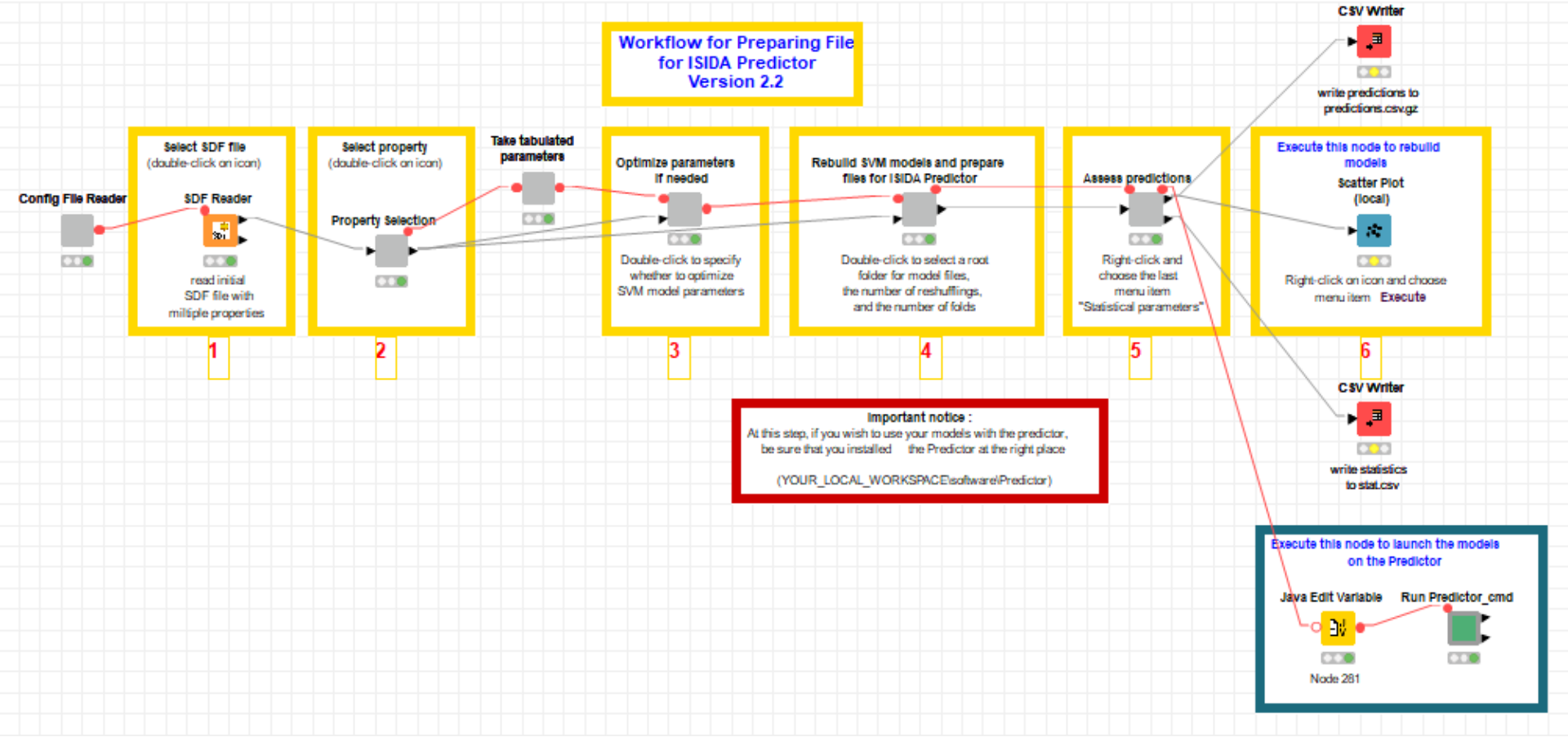




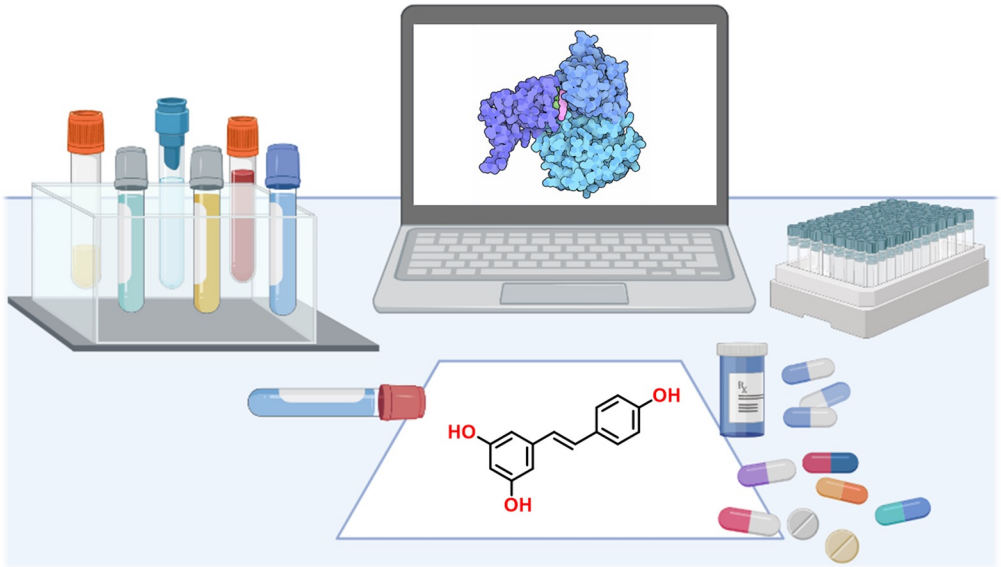
Models

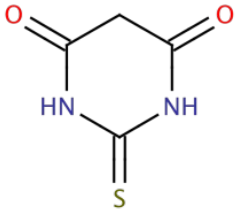
- **Biological properties**
  - ✔ ChEMBL activities (>600)
  - ✔ Antiviral (flavivirus, coronavirus, HIV)
  - ✔ Antiparasitic (Malaria)
  - ✔ Antibacterial (S. Aureus, E. Coli)
- **Medical properties**
  - ✔ Drug-drug interactions (carbamazepine and psychotropes)
  - ✔ Diagnostic (multiple sclerosis)
- **Thermodynamics**
  - ✔ Oxydoreduction potential
  - ✔ Viscosity
  - ✔ Ionic conductivity
  - ✔ Metal-ligand affinity
  - ✔ Vapor-liquid equilibria
  - ✔ Transition state temperatures (boiling point, melting point, glass transition)
  - ✔ Solubility
  - ✔ Auto-ignition temperature
- **Regulatory**
  - ✔ Ready biodegradability
  - ✔ Half-life
  - ✔ Acute toxicity
  - ✔ Aquatic toxicity
  - ✔ Endocrine disruption (Estrogen receptor, androgen receptor)

KNIME Integration



AI-driven design of new compounds



$$PROPERTY_{pred} = \mathbf{f}($$  $$)$$

- tuneable ISIDA descriptors
- special deep neural network architectures
- coupling of AI tools with chemical cartography
- hundreds of ML models for different PhysChem properties and bioactivities integrated in a user-friendly interface

Case studies:

- antivirals (SARS COV2)
- enantioselective catalysts
- efficient CO_2 absorbents
- tubulin activity modulators

ISIDA package

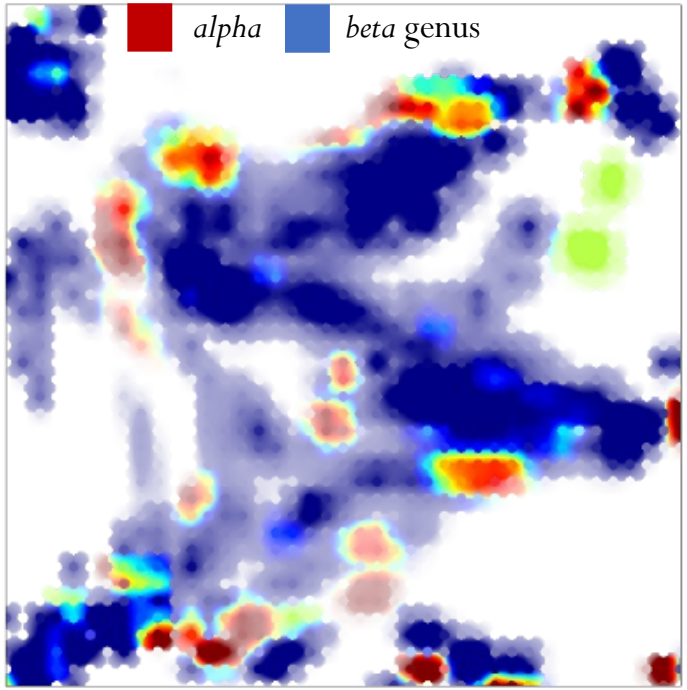
Discovery of new SARS-CoV-2 M^{pro} inhibitors

collaboration with the Chumakov Center, Russia

- The data for SARS-COV2 were not available at the very beginning of the COVID19 pandemic.
- *SARS-COV Relevant Antiviral Space map* was built for CoVs ligands studied before 2020

ZINC DATABASE

1.3 billion cmpds

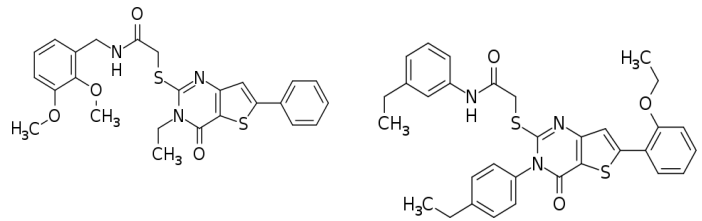


574 hits

Docking to SARS-CoV2 M^{pro}

10 hits

Experiment



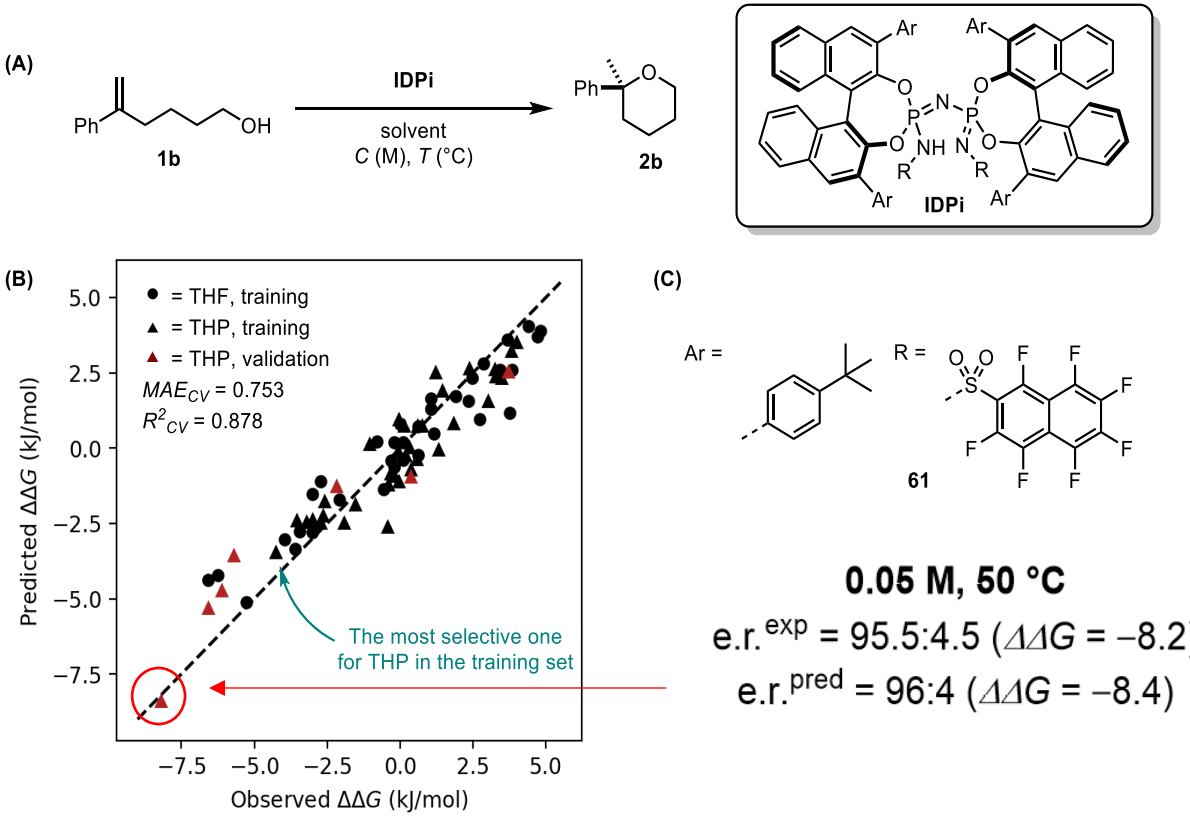
Confirmed M^{pro} inhibitors

D. Horvath et al. *Molecular Informatics*, 2020, **39**(12), 2000080

M.Yu. Zakharova et al *Frontiers in Pharmacology*, 2021, **12**, 773198

Computer-aided design of enantioselective catalysts

Collaboration with ICReDD, Hokkaido Univ.



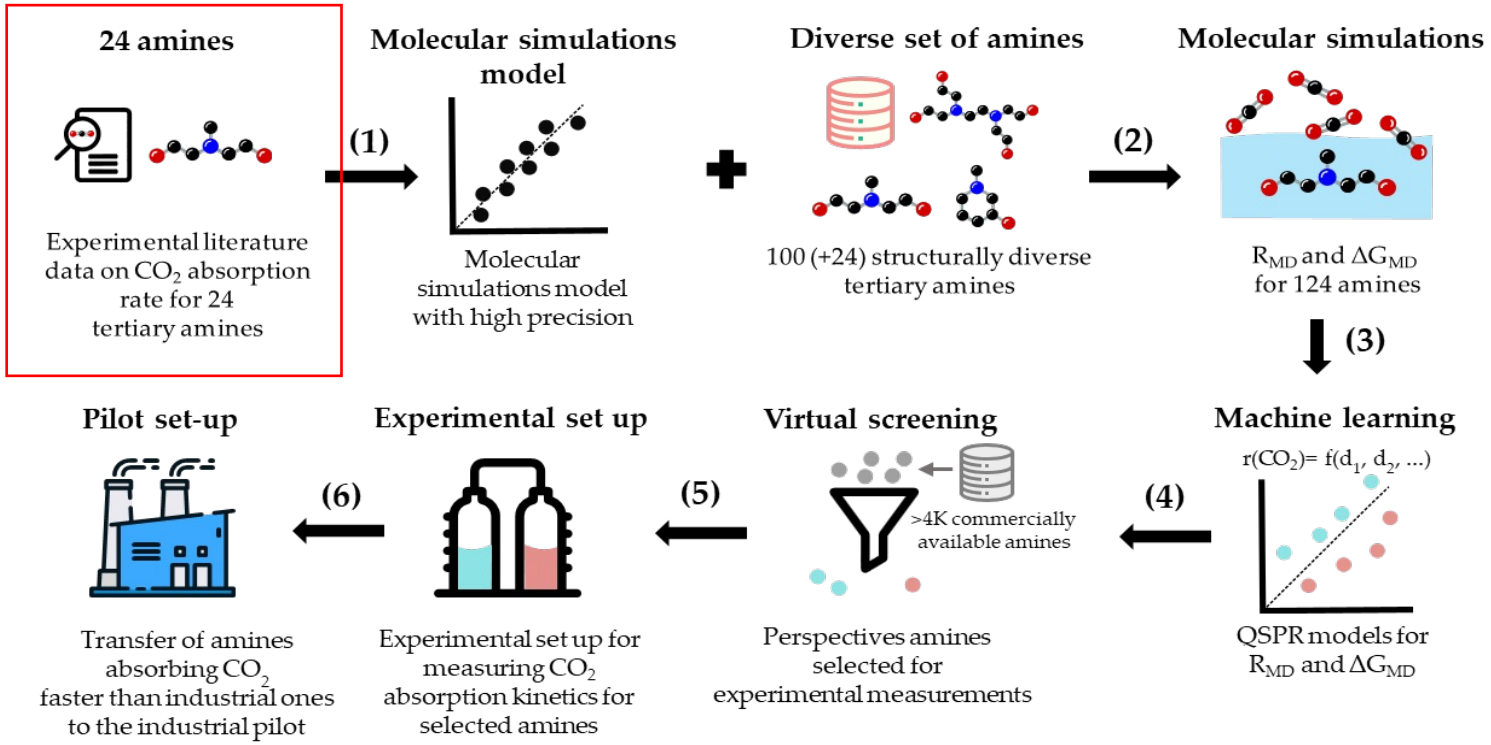
Benjamin LIST

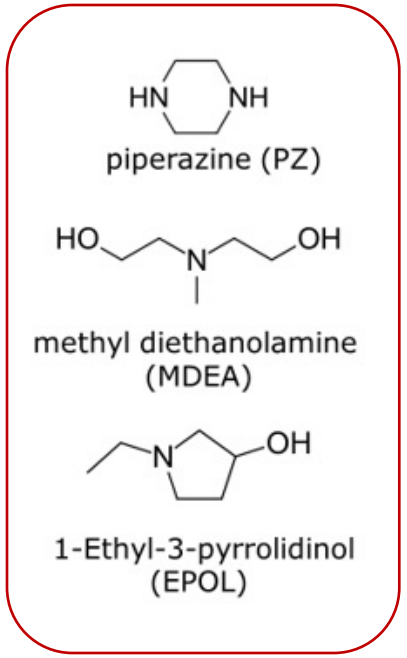
Max-Planck-Institut für Kohlenforschung, Germany
ICReDD, Hokkaido University, Japan

Machine-learning SVM models based on the ISIDA descriptors were used to design efficient catalysts of stereoselective hydroalkoxylation reaction. The designed molecule is much more selective (**95.5%**) than the best catalyst from the training set (**82%**).

Design of effective solvents for CO_2 capture

Collaboration with TOTAL Energies and MINES ParisTech

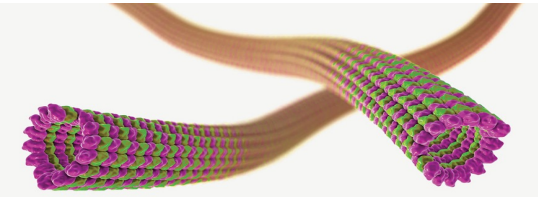




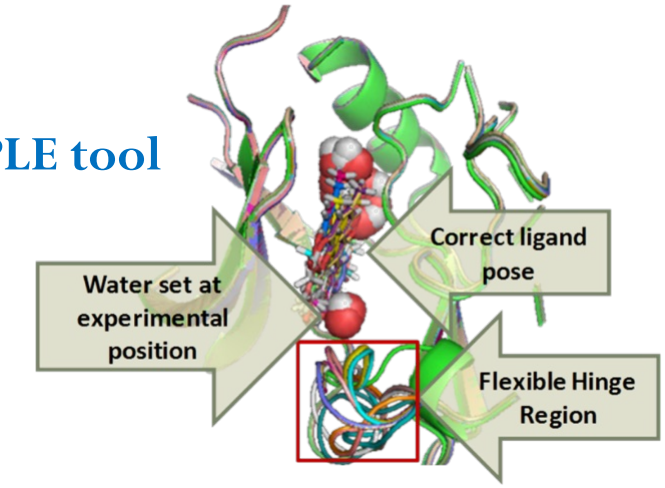
Using kinetic experiments, molecular simulations and machine learning, a class of tertiary amines that absorb CO2 faster than a typical commercial solvent has been identified

A. Orlov *et al. Communications Chemistry.*, 5, 37, 2022

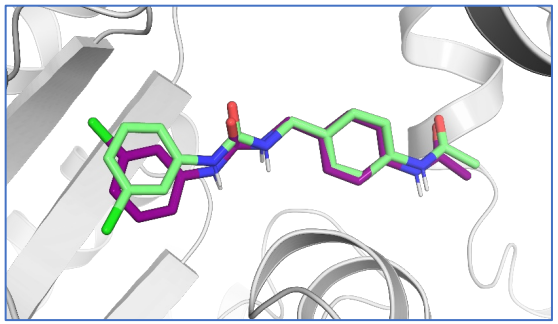
In silico design of Tubulin activity modulators

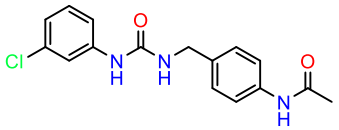


S4MPLE tool



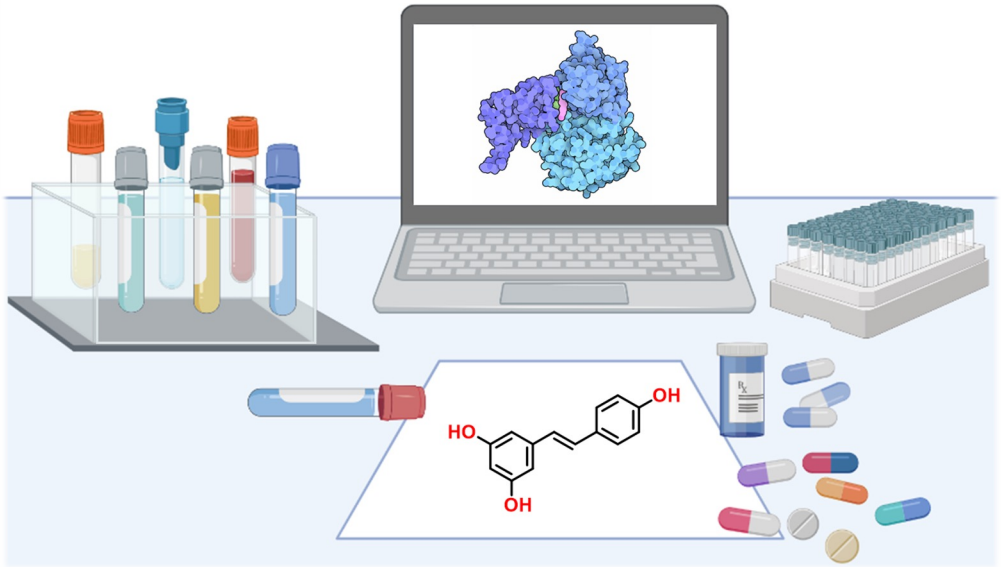
flexible docking



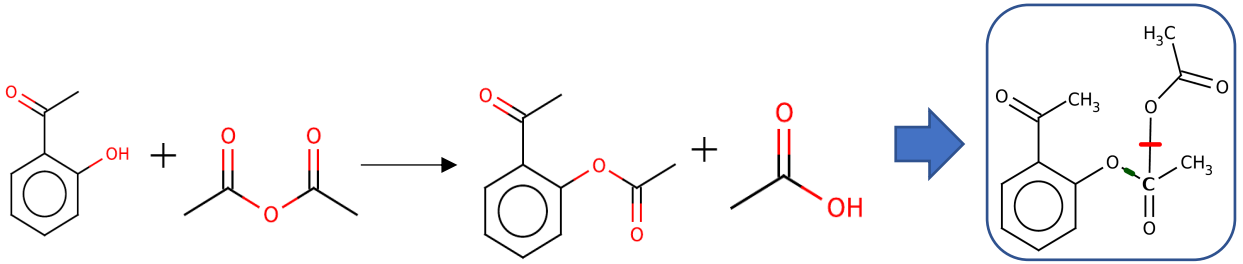


- For ~ 60% of computationally selected compounds, crystal structures validate binding modes as predicted by in-house docking tool S4MPLE
- Some of them display a clear microtubule depolymerizing effect

Chemical reaction mining



Condensed Graph of Reaction



- representation of a chemical transformation as a single molecular graph (pseudomolecule) leading to a significant simplification of machine-learning modelling of chemical reactions
- ML models for reaction kinetics and thermodynamics, optimal reaction conditions
- retrosynthesis

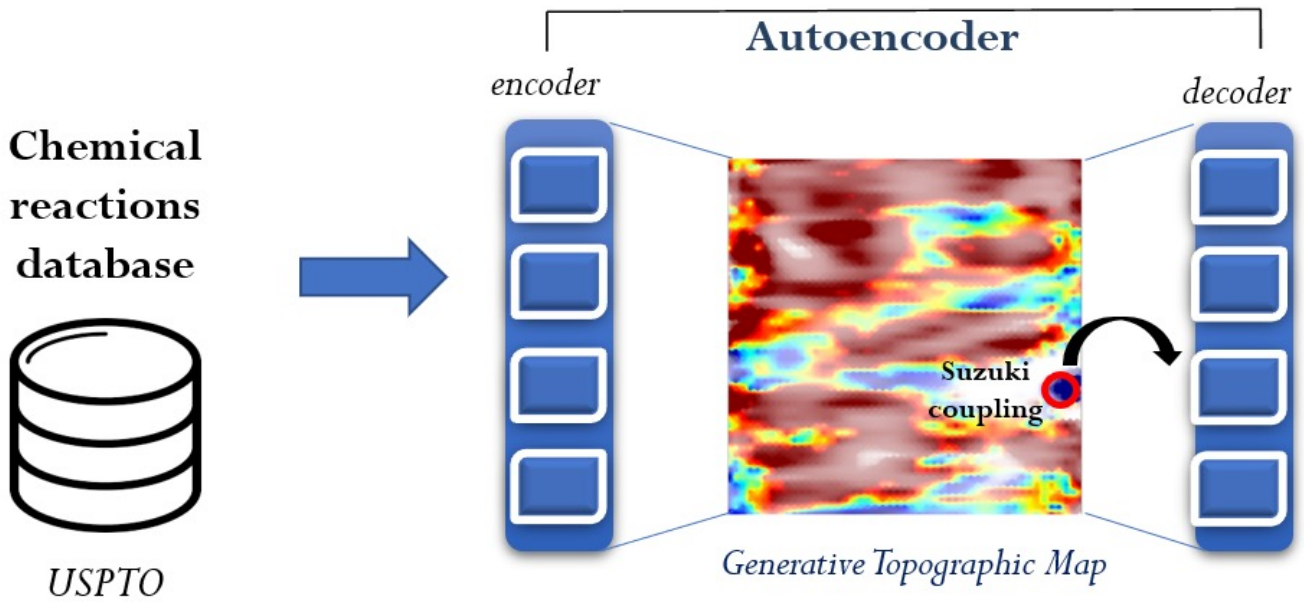
Case studies:

- AI-driven design of new chemical transformations

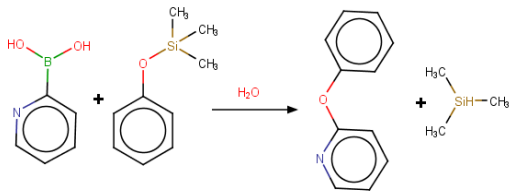
CGRTools package

AI-driven design of new Suzuki-like reactions

Collaboration with ICReDD, Hokkaido Univ.



New reactions



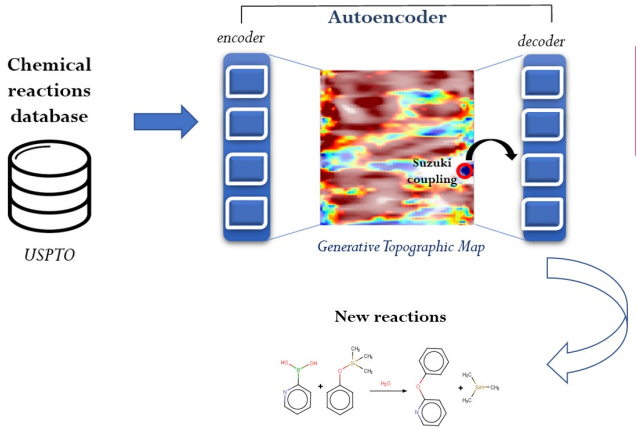
Reactions feasibility assessment with DFT

- Each reaction was encoded by a single *Condensed Reaction Graph (CGR)*
- Special *SMILES/CGR* strings were used as an input

- **13 new (with respect to the training data) Suzuki-like reactions have been detected**
- **5 of them have been found in the literature**

W. Bort *et al.*, *Scientific Reports*, 2021, **11**, 3178

AI-driven design of new Suzuki-like reactions



W. Bort *et al.*, *Scientific Reports*, 2021, **11**, 3178

ChemOinformatique

STRASBOURG

Research

Master in Chemoinformatics

- one of the first programs in the field (since 2001)
- 5 double diploma agreements
- Erasmus-Mundus program *Chemoinformatics PLUS*

International events

- Summer Schools in Chemoinformatics (since 2008)
- French –Japanese workshops (since 2008)
- French –Israeli workshops (since 2018)

9th Strasbourg Summer School on Chemoinfomatics

Strasbourg, 24 - 28 June 2024

9th Strasbourg Summer School on Chemoinfomatics

Strasbourg, 24-28 June 2024

- Plenary lectures
- Short oral presentations
- Poster session
- Tutorials

- Cultural program
- Beer & Bretzel party
- Vine & Cheese party

- Conference diner

- Pre-conference
  - “Crash course in Chemoinformatics”
    - for beginners
  - Hackathon
  - Advanced skills

Strasbourg, 2022

Education in Chemoinformatics

- **Master in Chemoinformatics** (2001) unique in Europe

- **Master *In Silico Drug Design*** (2011) – with Paris-Diderot and Milan

- **Double diploma programs**

  - 5 agreements signed with universities of Kiev, Milan, Ljubljana, Tel-Aviv, Lisbon

- **Erasmus Mundus program "Chemoinformatics PLUS"**

Mundus Chemoinfo+

European Master in Chemoinformatics

7 partners for 6 tracks

(Particular situation for Kyiv and the track no.6)

TRACK	YEAR 1		Mandatory Summer School	YEAR 2		Optional Summer School
	S1	S2		S1	S2	
In Silico Drug Design	Strasbourg	Milan		Paris	Internship	
Chemoinformatics and Physical Chemistry	Milan	Milan		Strasbourg	Internship	
Chemoinformatics for Biophysical and Computational Chemistry	Ljubljana	Ljubljana		Strasbourg	Internship	
Chemoinformatics for Organic Chemistry	Lisbon	Lisbon		Strasbourg	Internship	
Chemoinformatics and Materials Informatics	Ramat-Gan	Ramat-Gan		Strasbourg	Internship	

First promotion: 2022-2024

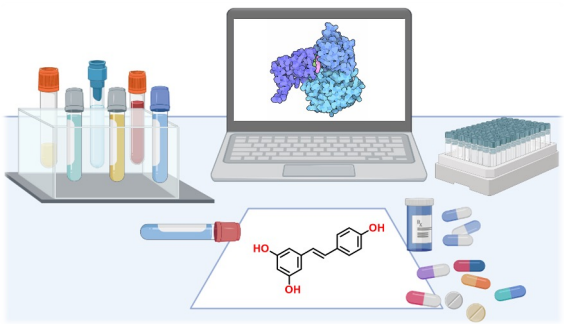
https://masterchemoinfoplus.chimie.unistra.fr/

EMJMD Primary Topics

- **Chemoinformatics**
  - Coding of chemical structures
  - Chemical space
  - Chemical similarity and diversity
  - Chemical databases and data sources
  - Molecular descriptors
  - Data science
  - IA and QSAR
  - Generative models
- **Drug Design**
  - Chemical libraries of biological interests
  - Pharmacodynamics, pharmacokinetics
  - ADME/Toxicity
  - Environmental fate
  - Protein ligand-docking
  - Virtual screening
  - Profiling of chemical libraries
  - Structural determination and modeling of macromolecules
  - Biological environment
- **Quantum Chemistry**
  - Conventional quantum chemical methods
    - Semi-empirical, Hartree-Fock, DFT
  - Physical motivations of quantum chemistry calculation methods
  - Applicability domain of quantum chemical models
  - Main software packages for quantum calculations

- **Molecular Modeling**
  - Molecular mechanics and molecular dynamics
  - Force fields and empirical potential energy functions
  - Molecular modeling as a tool in chemical research
  - Intra- and supra-molecular interactions
  - Emerging properties at macroscopic scales
  - Thermodynamics ensembles
  - Solvation hydrophilic and hydrophobic
  - Conformational analysis and empirical representations
  - Rational choice of methods
  - Evaluation of the reliability of results
- **Data mining and artificial intelligence**
  - Methods
    - Classification, Regression, Clustering
  - Validation
  - Tuning method parameters
  - Ensemble modeling
  - Active learning
  - Multi-task learning
  - Semi-supervised learning and transductive inference
  - Recommender systems
  - Generative models and adversarial learning
  - Autoencoders

1 AI-driven design of new compounds and materials

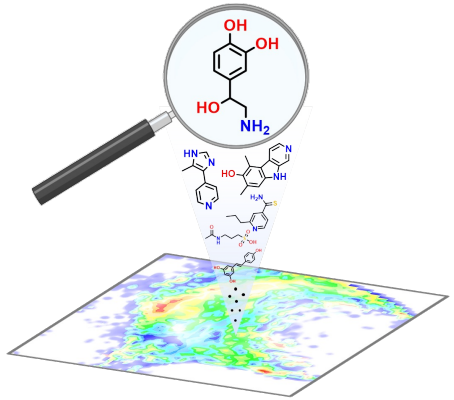


Machine-Learning / Artificial intelligence modeling

- tuneable ISIDA descriptors
- special deep neural networks architectures
- coupling of AI tools with chemical cartography
- Hundreds of ML models for different physchem properties and bioactivities integrated in a user-friendly interface

ISIDA package

2 Analysis of ultra-large chemical spaces

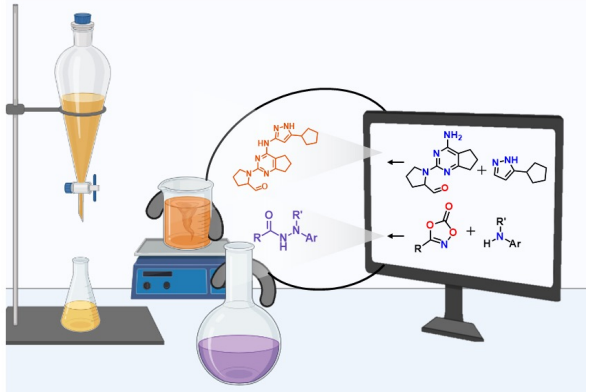


Chemical cartography with Generative Topographic Mapping

- visualization of both individual data (chemical structures) and their probability distribution on a 2D plane;
- prediction of properties of new molecules projected on the map and, therefore, can be used as a virtual screening tool;
- suitable for Big Data analysis (billions of molecules)

ChemSpace Atlas

3 Chemical reactions mining



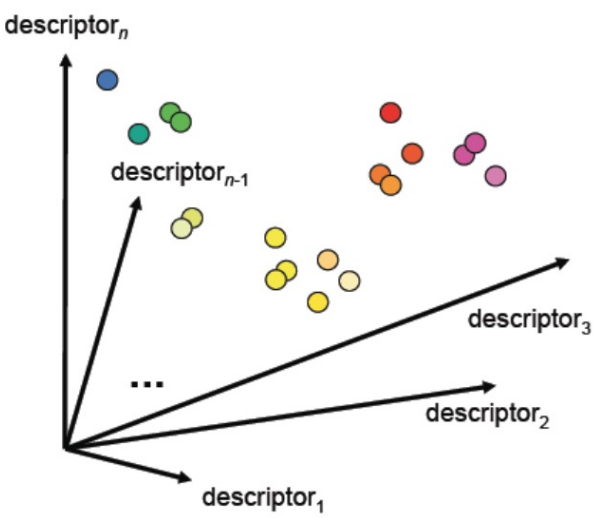
Condensed Graph of Reaction

- representation of a chemical transformation as single molecular graph (*pseudomolecule*) leading to significant simplification of machine-learning modelling of chemical reactions
- ML models for reaction kinetics and thermodynamics, optimal reaction conditions
- retrosynthesis

CGRTools package

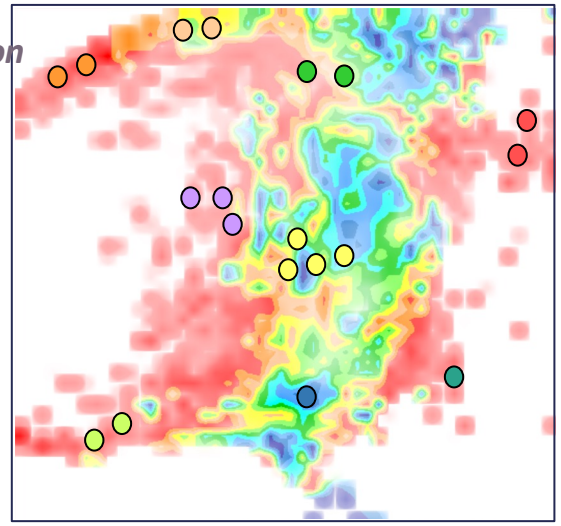
Thank you!

GTM property profiling

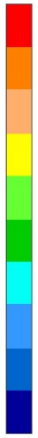


N-dimensional space

dimensionality reduction



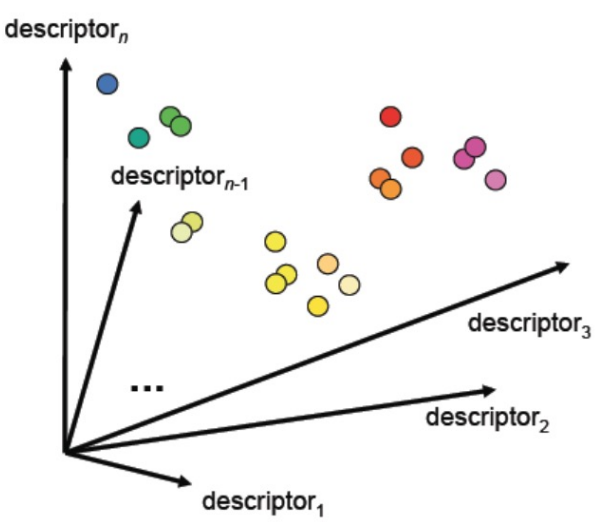
2D space



Toxic

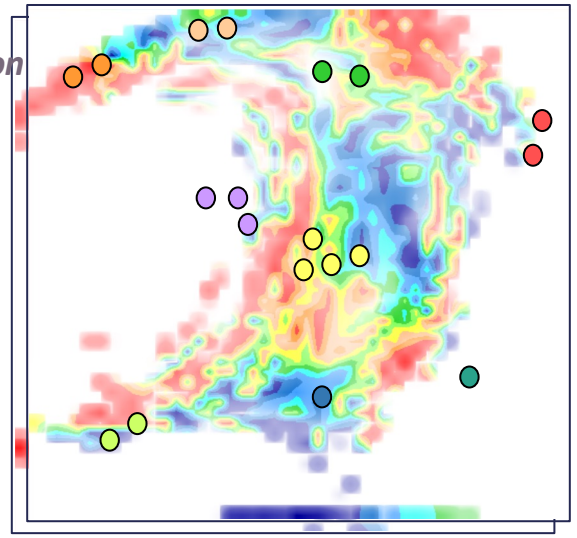
non-Toxic

GTM property profiling

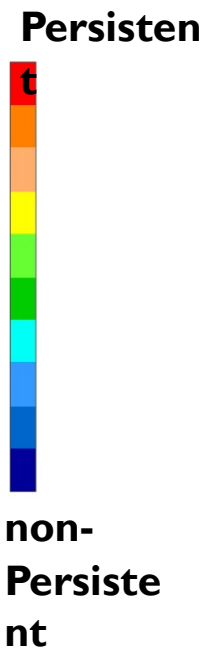


N-dimensional space

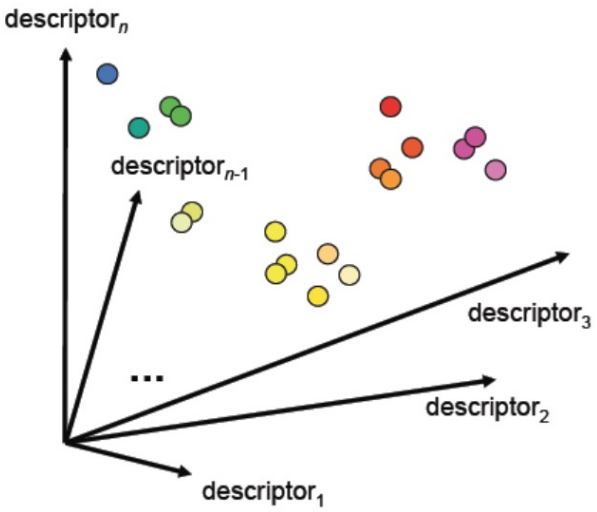
dimensionality reduction



2D space

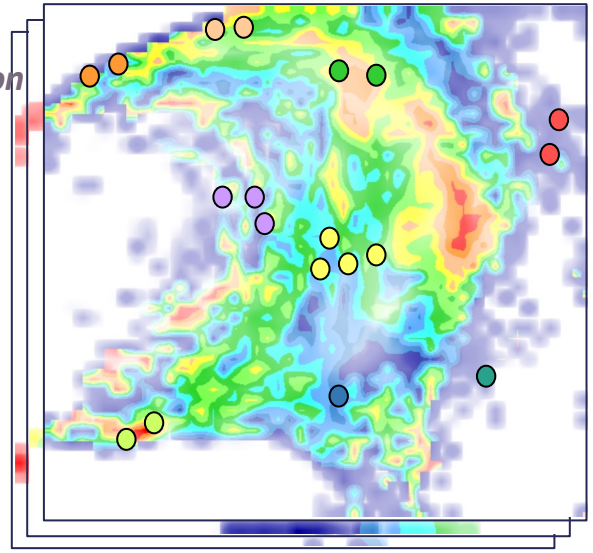


GTM property profiling

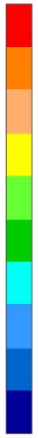


N-dimensional space

dimensionality reduction



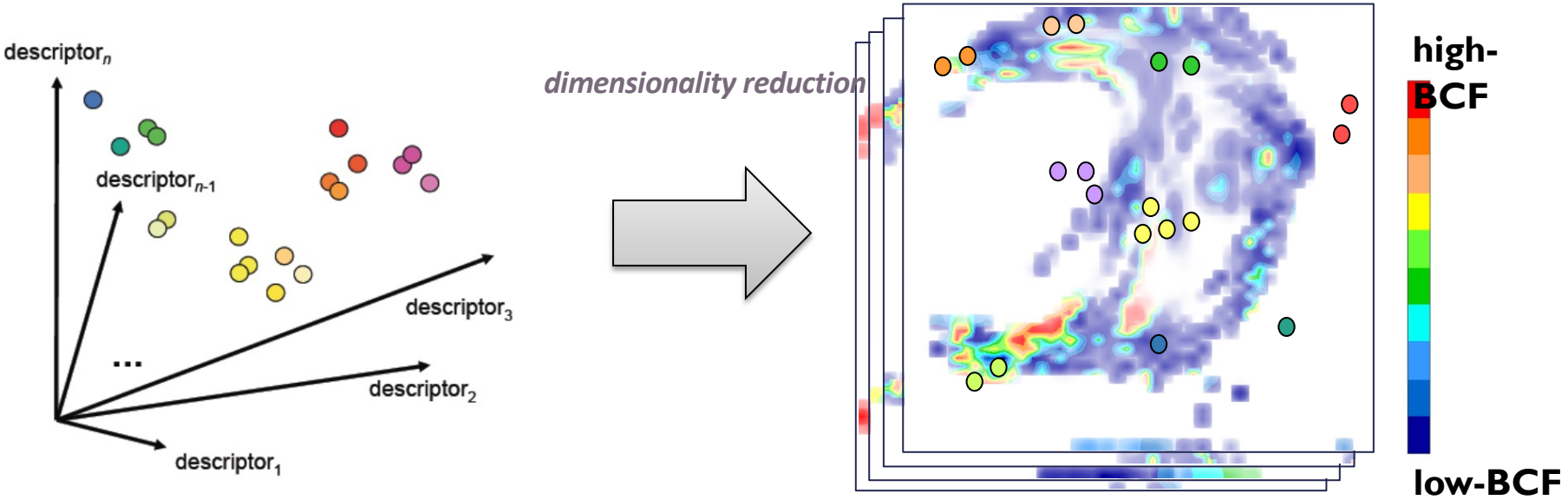
2D space



Binder

non-Binder

GTM property profiling



N-dimensional space

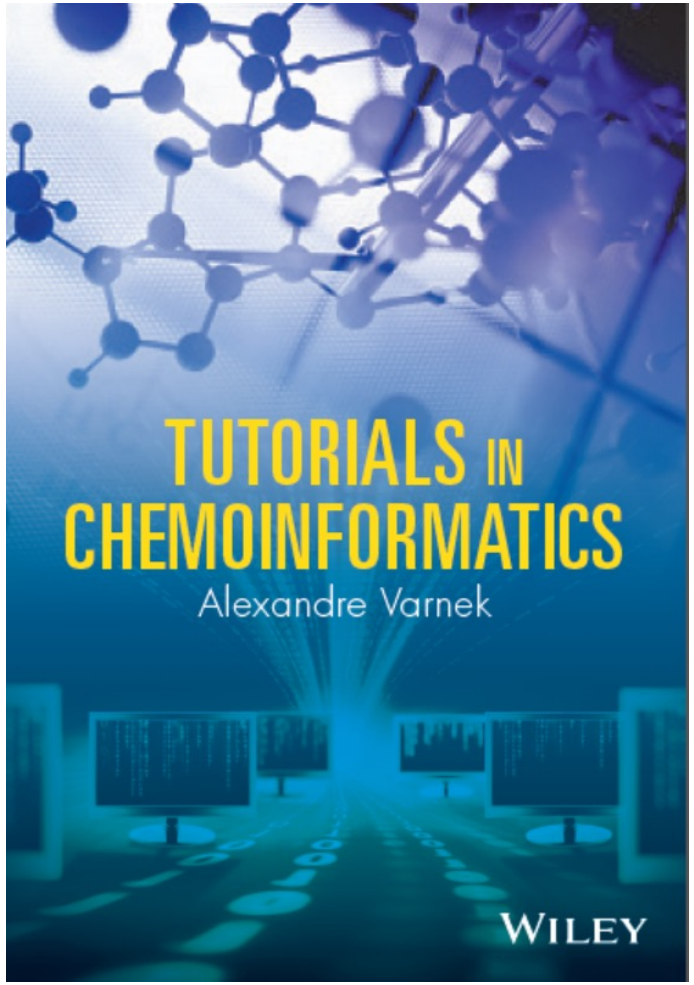
2D space

Text-books in chemoinformatics

« Introduction to Chemoinformatics »
by Igor Baskin, Timur Madzhidov and Alexandre Varnek

in 6 volumes, 2013-2019

« Fundamentals of Chemoinformatics »,
WILEY, 2024



2017

Laboratory of Chemoinformatics 2017- 2022

Publications	88
Chapitres de livre	7
Thèses soutenus	10
Thèses actuellement en cours	10
Contrats industriels	10
Projets ANR	2
Marie Curie ITN H2020	2