Хемоинформатика – основные понятия и области применения

Alexandre Varnek
Laboratory of Chemoinformatics, University of Strasbourg

ASFM 2014, Moscow, 9-10 December 2014

UNIVERSITÉ DE STRASBOURG

ChémoInformatique STRASBOURG

Laboratory of Chemoinformatics

Master in Chemoinformatics (2001)

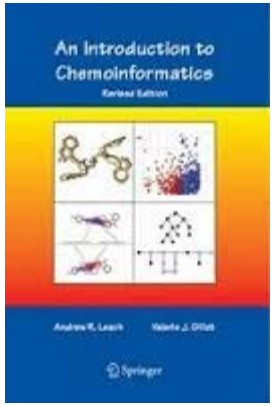
MASTER Chem Info

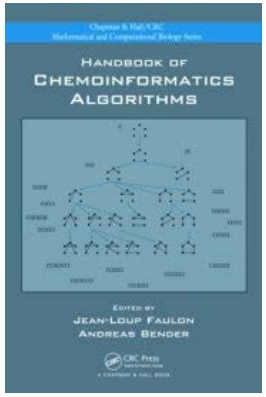
Master "In Silico Drug Design" (2010)

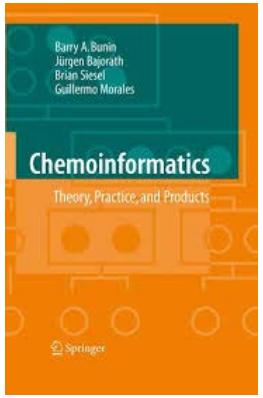
Strasbourg - Paris-Diderot - Milan

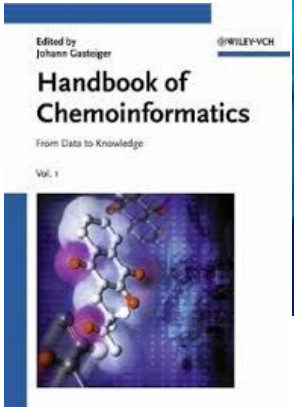
Double diploma Strasbourg / Kazan (2013)

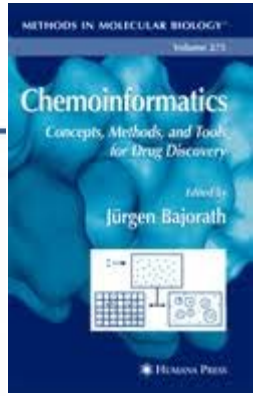
Selected books in chemoinformatics

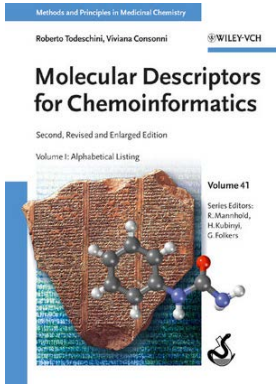


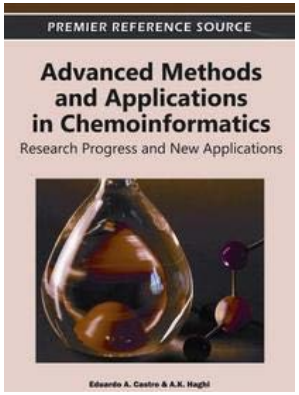


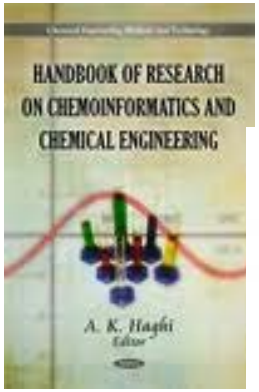


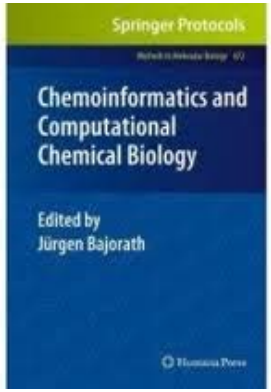












РОССИЙСКИЙ НАУЧНЫЙ ФОНД

ПОДДЕРЖКА И РАЗВИТИЕ

НОВОСТИ О ФОНДЕ ДОКУМЕНТЫ КОНКУРСЫ КОНТАКТЫ

Классификатор Фонда

КОД НАИМЕНОВАНИЕ

03-700 Многомасштабное компьютерное моделирование структуры и свойств сложных химических систем и материалов

03-705 Хемоинформатика

грант РНФ 2014

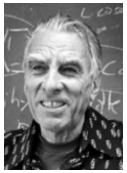
Chemoinformatics:

new disciline combining several „old“ fields

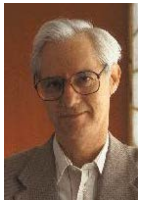
- Chemical databases
- Structure-Activity modeling (QSAR)
- Structure-based drug design
- Computer-aided synthesis design


Michael Lynch


Peter Willett


Corwin Hansch


Johann Gasteiger


Irwin D. Kuntz


Hans-Joachim Böhm


Elias Corey


Ivar Ugi

OUTLOOK

- **Fundamentals of chemoinformatics**
- **Structure-Property modeling: case studies**
  - **Anti-thrombotics**
  - **Tautomeric equilibria**
  - **Ionic Liquids**
  - **Chemical reactions**
- **Materials design**

Chemoinformatics: fundamentals

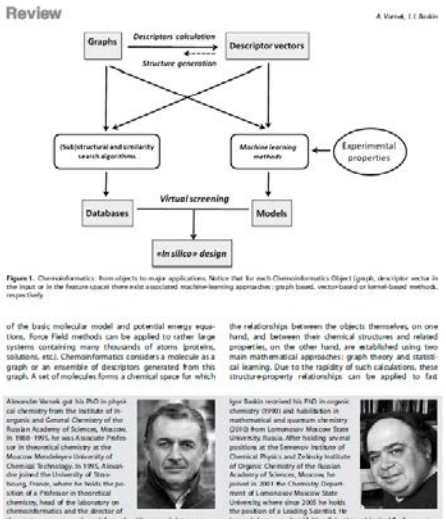
Review

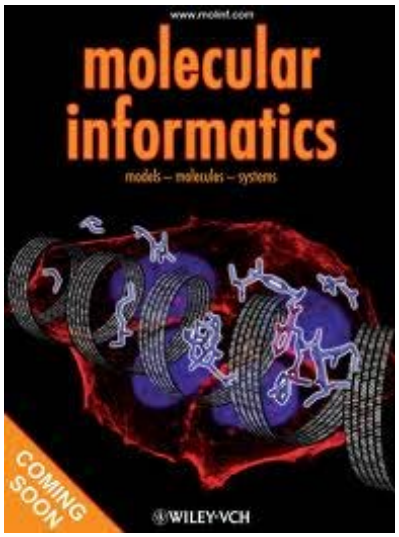
DOI: 10.1002/minf.201000100

Chemoinformatics as a Theoretical Chemistry Discipline

Alexandre Varnek*[a] and Igor I. Baskin[b]

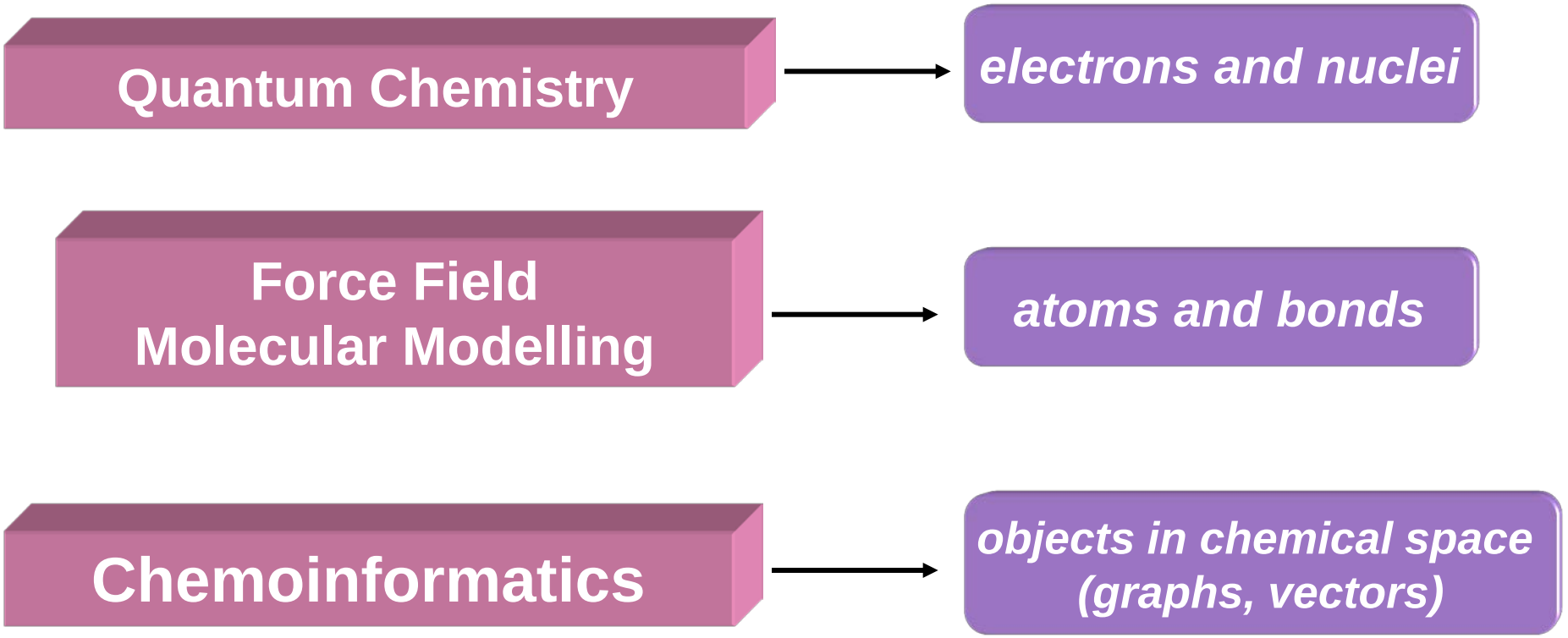
Mol. Inf. **2011**, *30*, 20 – 32



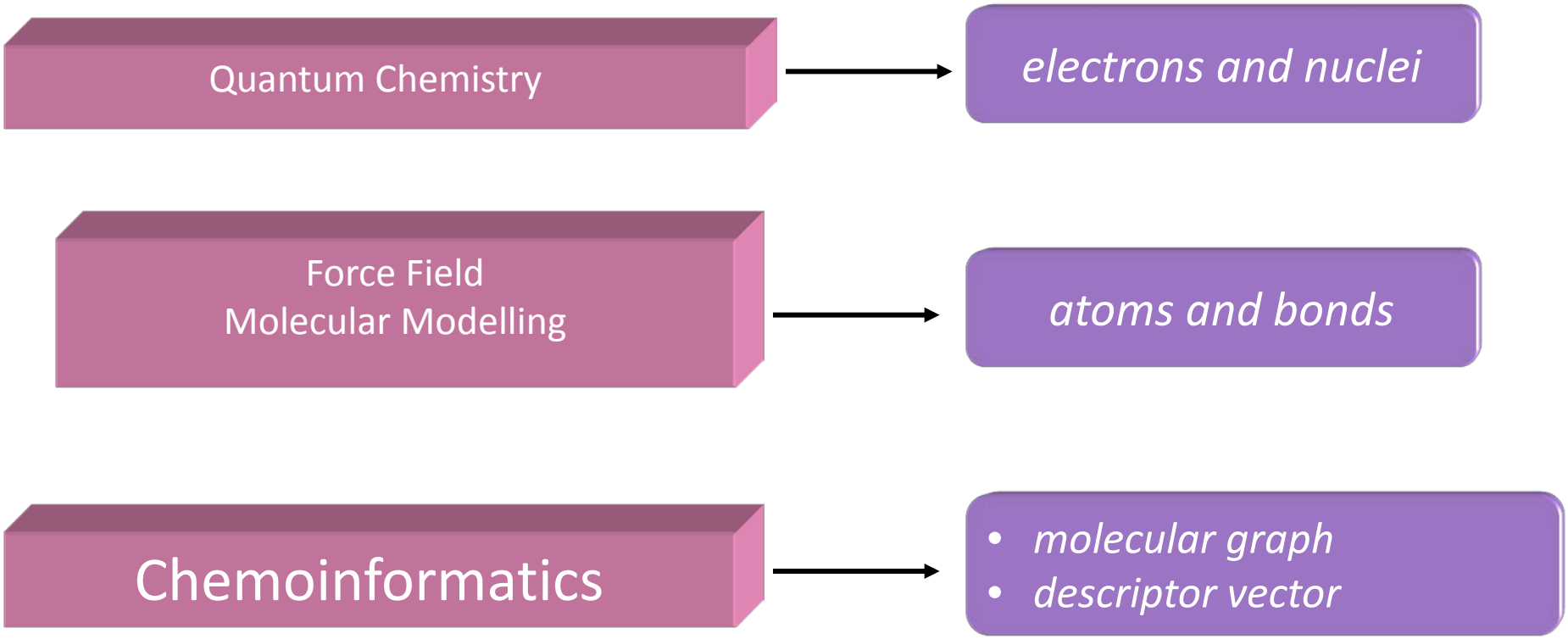


Chemoinformatics is defined as individual discipline characterized by its own molecular model, basic concepts, major applications and learning approach

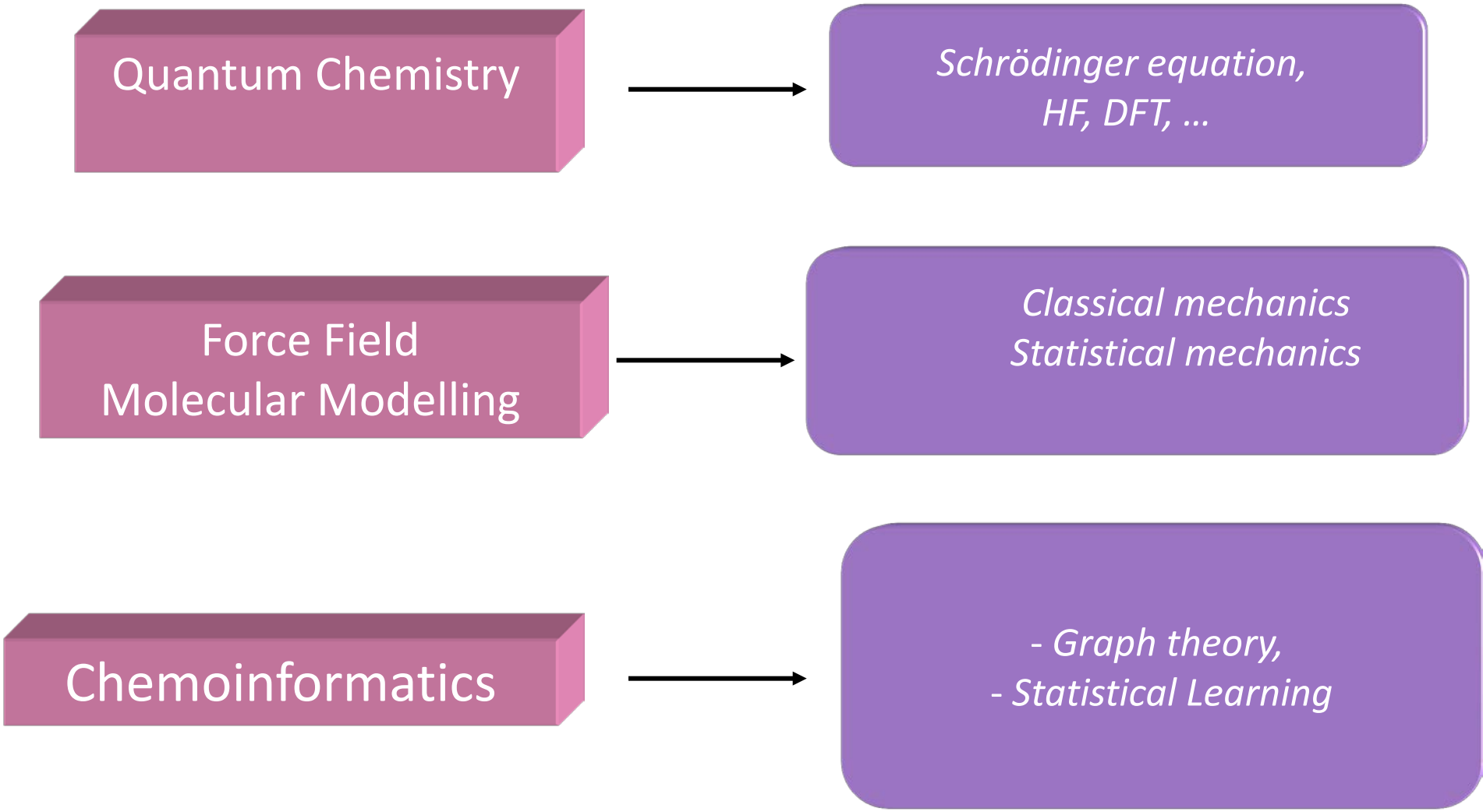
Molecular Model



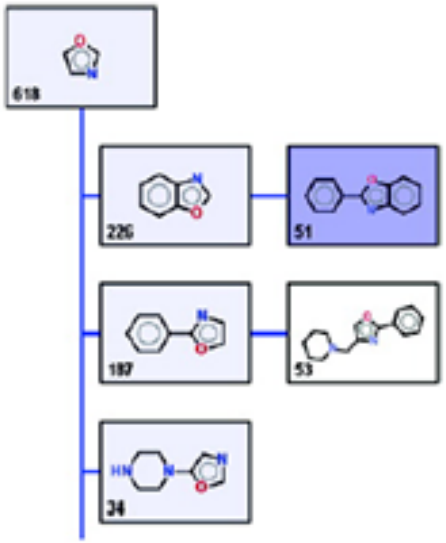
Molecular Model



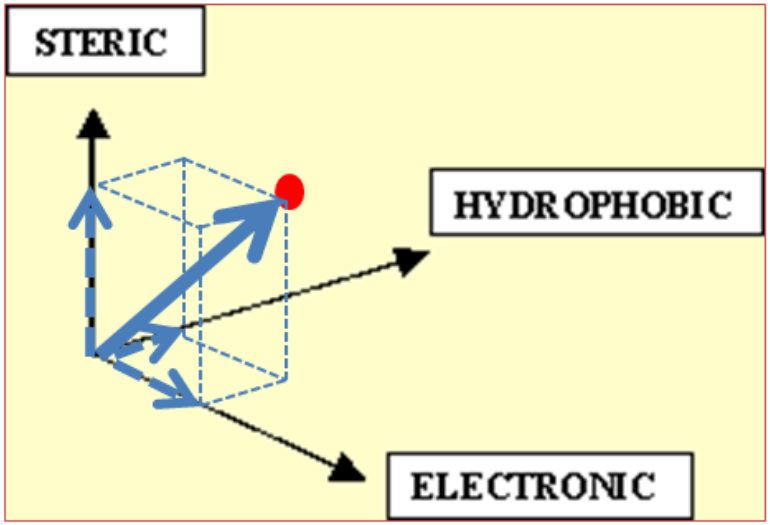
Basic mathematical approaches



Chemical Space paradigm



graphs-based



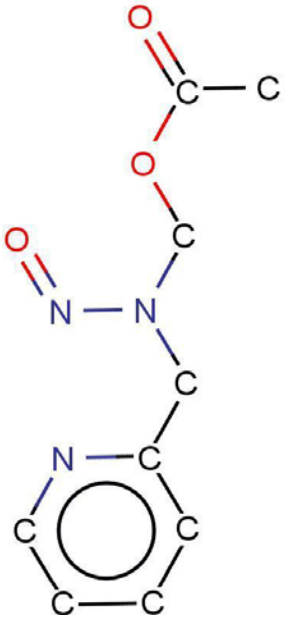
descriptors -based

SPACE = objects + metric

Molecular Descriptors :

ensemble of topological, electronic, geometry parameters calculated directly from molecular structure

Molecular graph



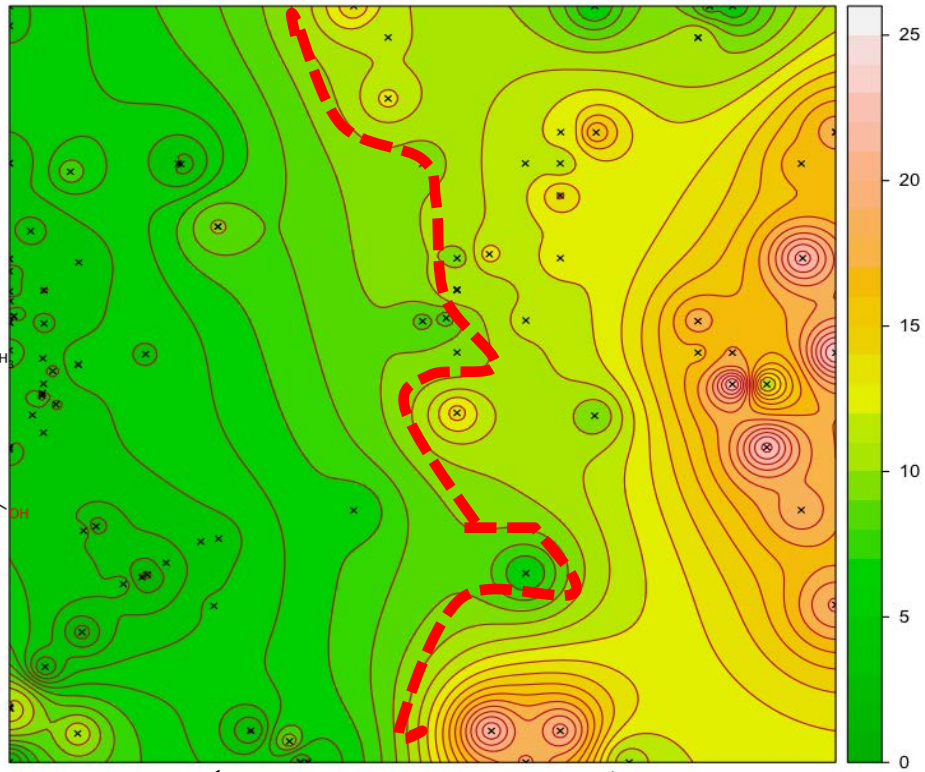
- **Topological indices,**
- **Atomic charges,**
- **Inductive descriptors,**
- **Substructural fragments,**
- **Molecular volume and surface, …**

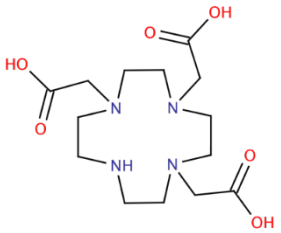
Descriptor vector

Descriptors
D_1
D_2
…
D_i
…

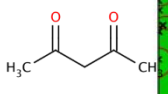
> 5000 types of descriptors are reported

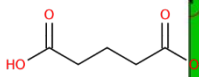
Chemical space visualization

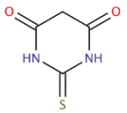




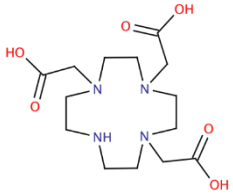
+ Lu^{3+}

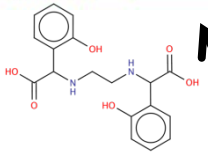






Weak binders





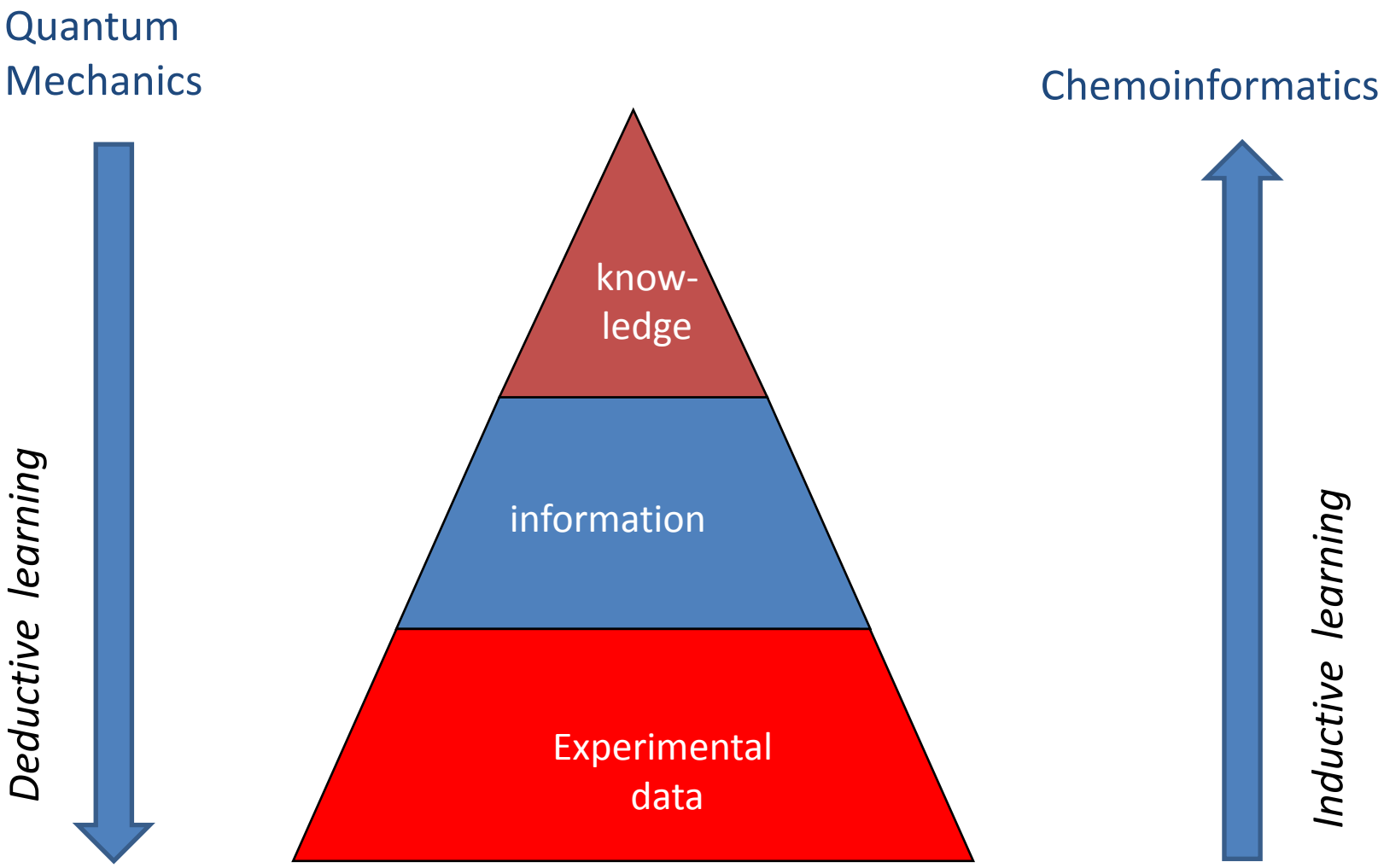
Strong binders

Generative Topographic Mappping of the set of Lu^{3+} binders

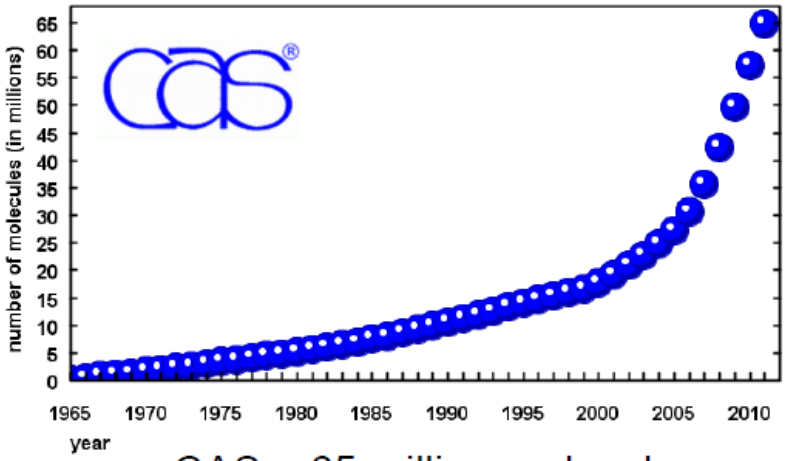
Contours correspond to different logK values

H. Gaspar, I. Baskin, G. Marcou, A. Varnek *unpublished results*

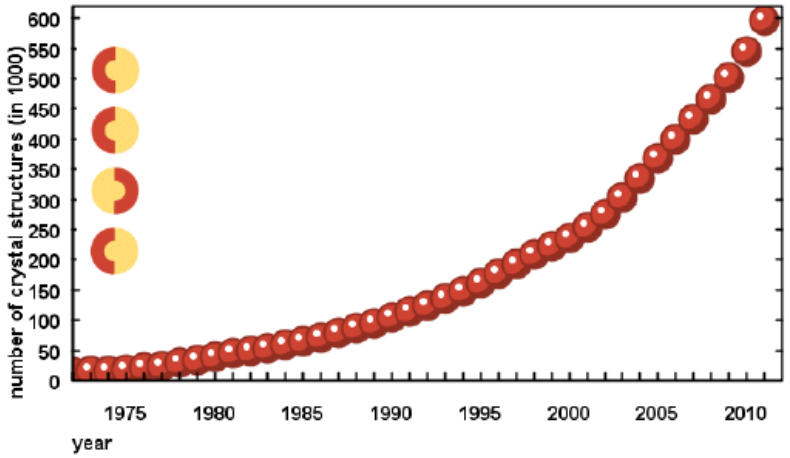
Chemoinformatics: Learning from Data



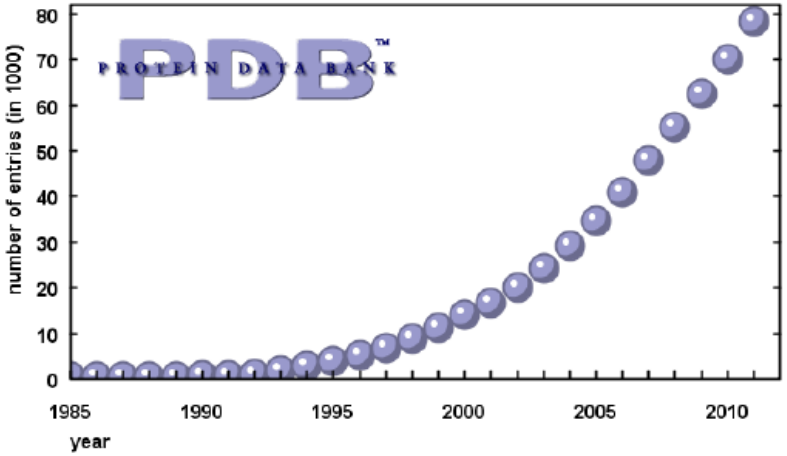
Data Explosion in Chemistry



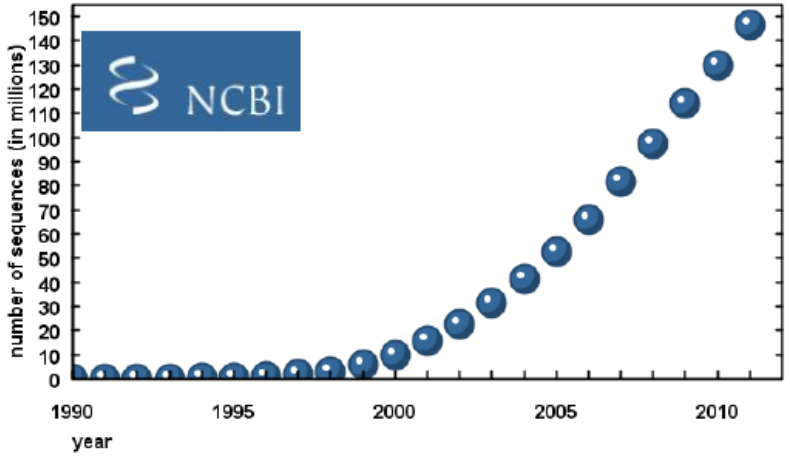
CAS – 65 million molecules



CCDC – 600'000 structures



PDB – 78'000 proteins



GenBank – 145 million sequences

THALESNano

"Good reactions"™

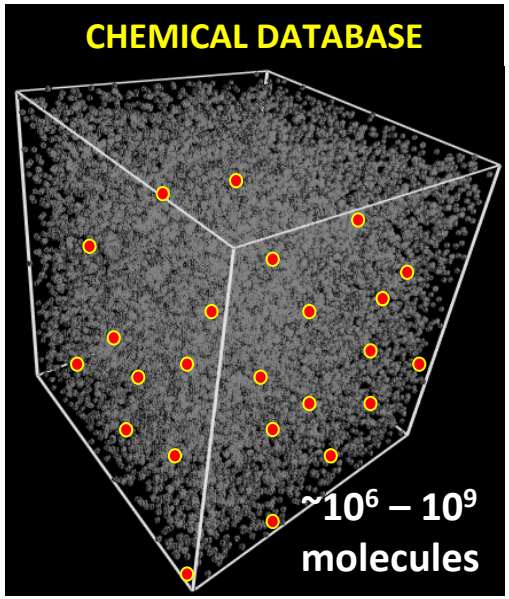
H-Cube® Continuous-flow Hydrogenation Reactor

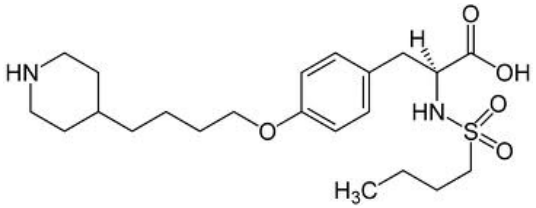
A revolutionary bench-top standalone hydrogenation reactor combined with automated liquid handler and laboratory automation software

Up to 100 reactions a day

≈ 30.000 reactions a year

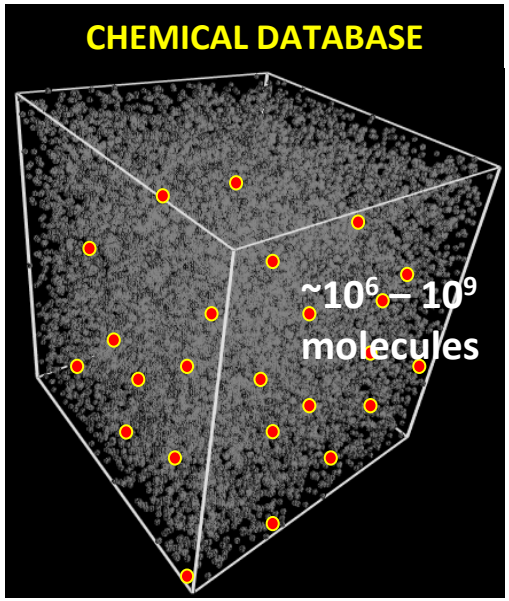
Virtual screening : finding the needle in the haystack



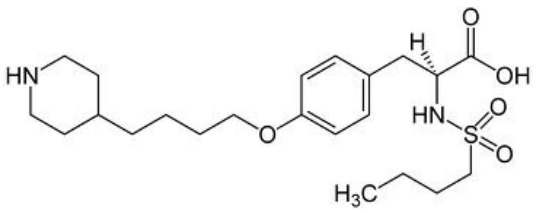


Chemoinformatics:

pattern recognition in chemistry



model

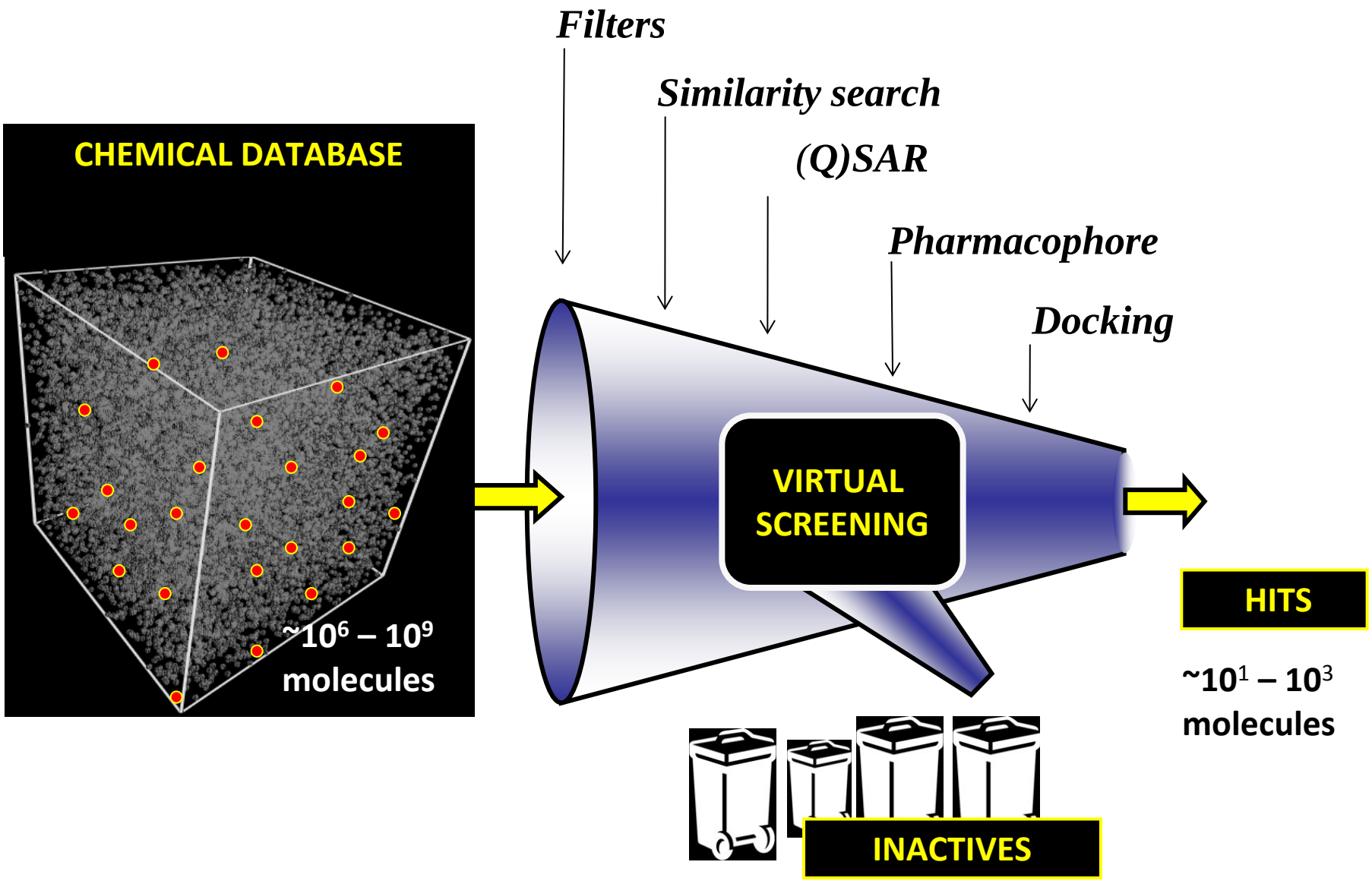


- **Specific structural motifs,**
- **Selected molecular properties (shape, fields, …),**
- **Interaction patterns,**

- Mathematical equations

Activity = F (structure)

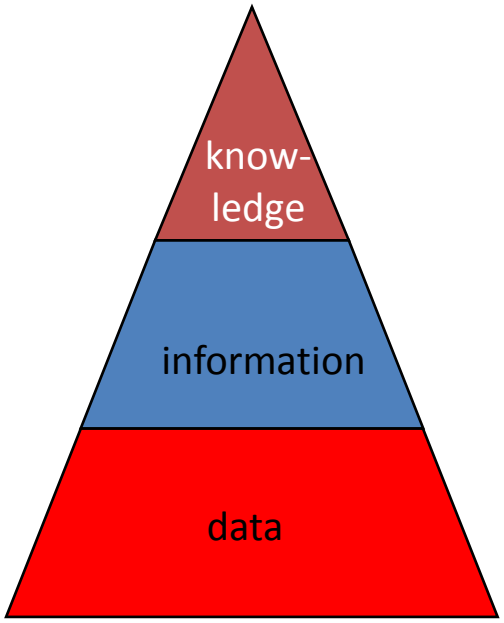
Chemoinformatics: Virtual screening "funnel"



(Quantitative) Structure-Property Relationships (Q)SPR

Property = F (structure)

= F (descriptors)



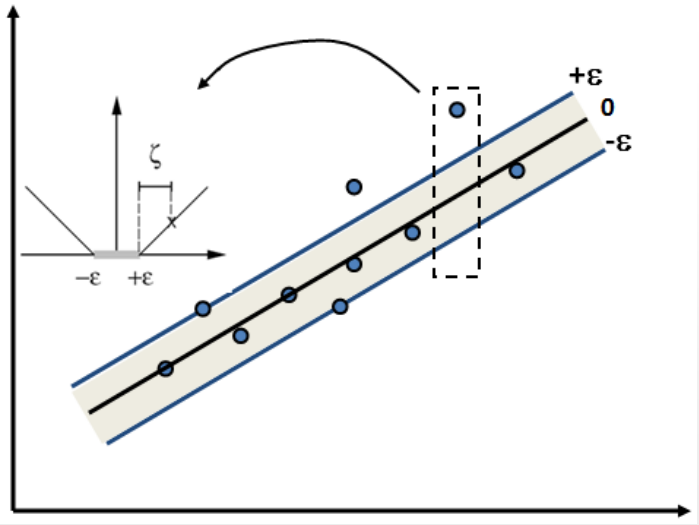
- Two main types of models:
  - **classification** (prediction of classes)
  - **regression** (prediction of numbers)

Machine learning: *Regression models*

Multiple Linear Regression (MLR)

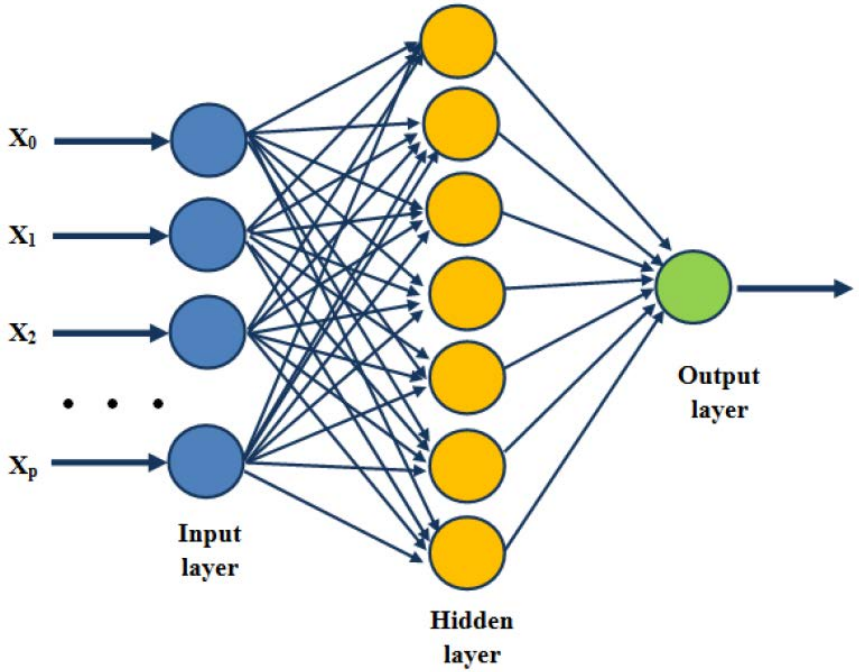
$$\text{Property} = a_0 + \sum_{i=1}^{k} a_i \cdot X_i$$

Support Vector Regression (SVR)



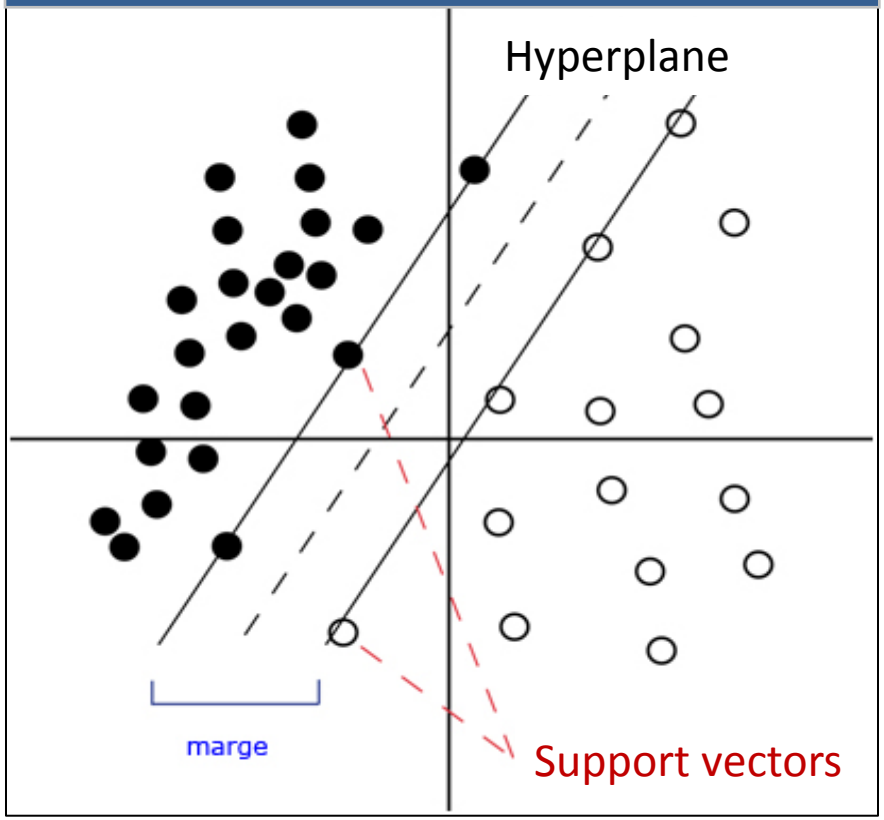
Associative Neural Networks (ASNN)

Ensemble of Neural Networks



Machine learning: *Clasification Models*

Support Vector Classification (SVC)

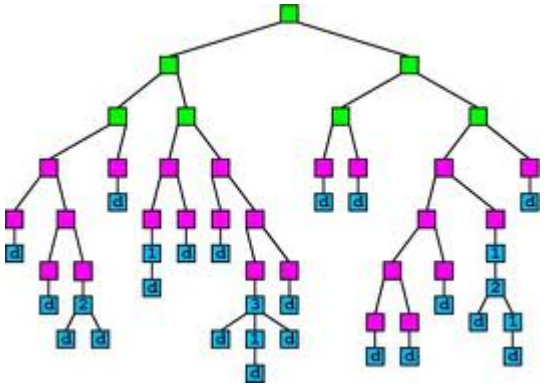


Naive Bayes (NB)

Probabilistic Method

$$P(Y|a,b) = \frac{P(a|Y) \times P(b|Y) \times P(Y)}{P(a) \times P(b)}$$

Decision Trees



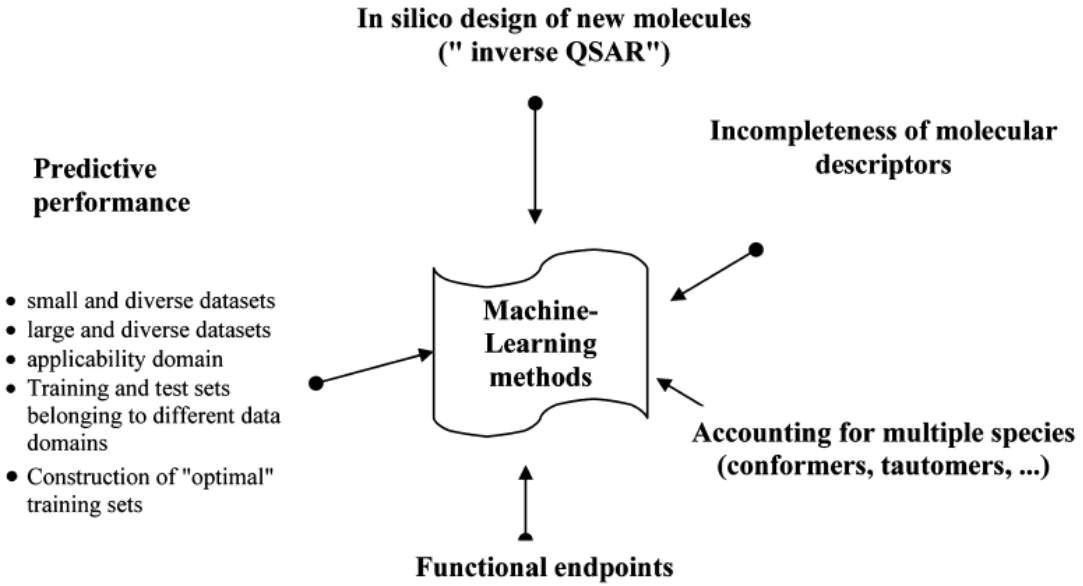
JOURNAL OF CHEMICAL INFORMATION AND MODELING

Perspective

pubs.acs.org/jcim

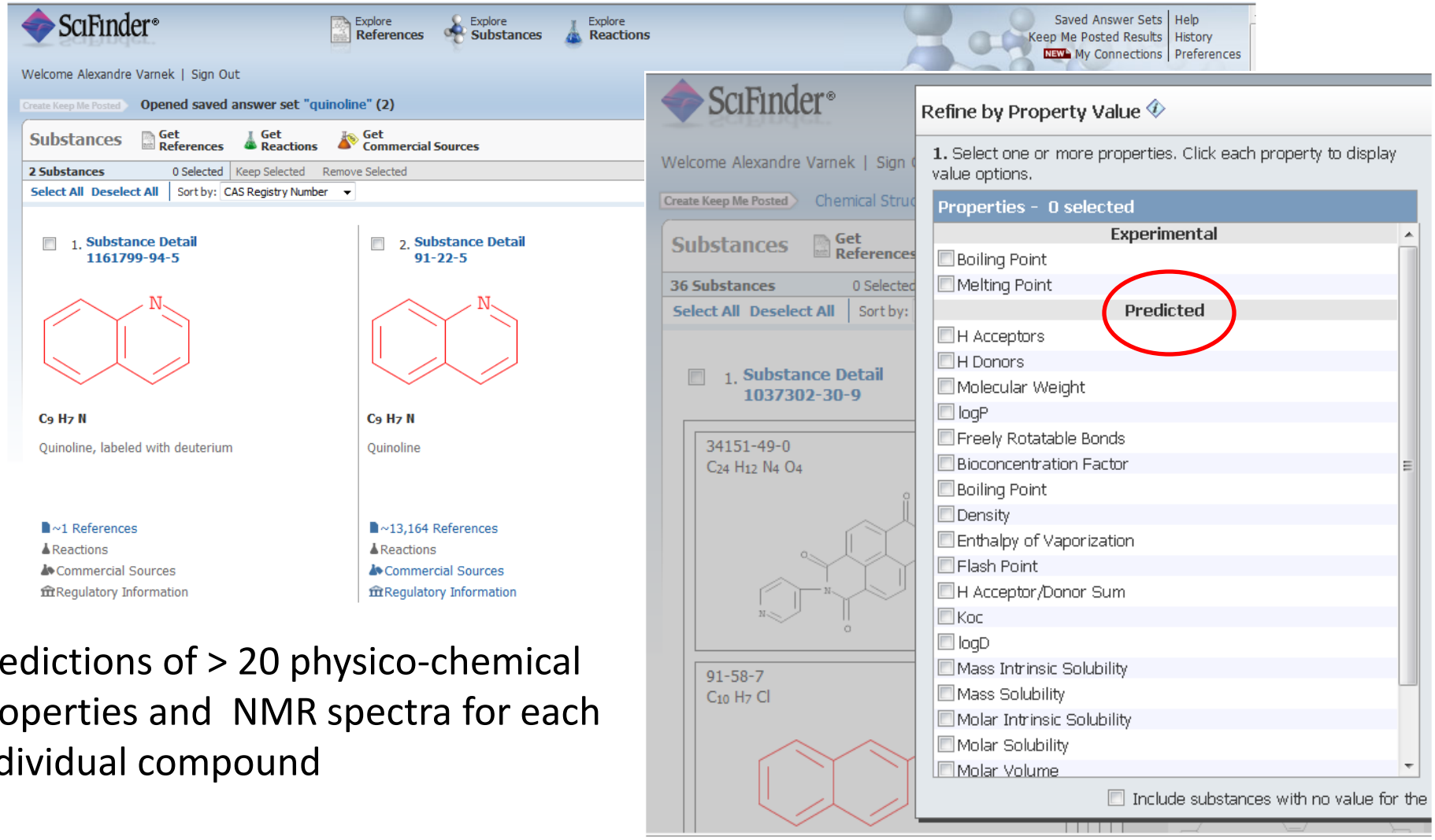
Machine Learning Methods for Property Prediction in Chemoinformatics: *Quo Vadis*?

Alexandre Varnek*,† and Igor Baskin†,‡



Main challenges of machine learning methods in chemoinformatics

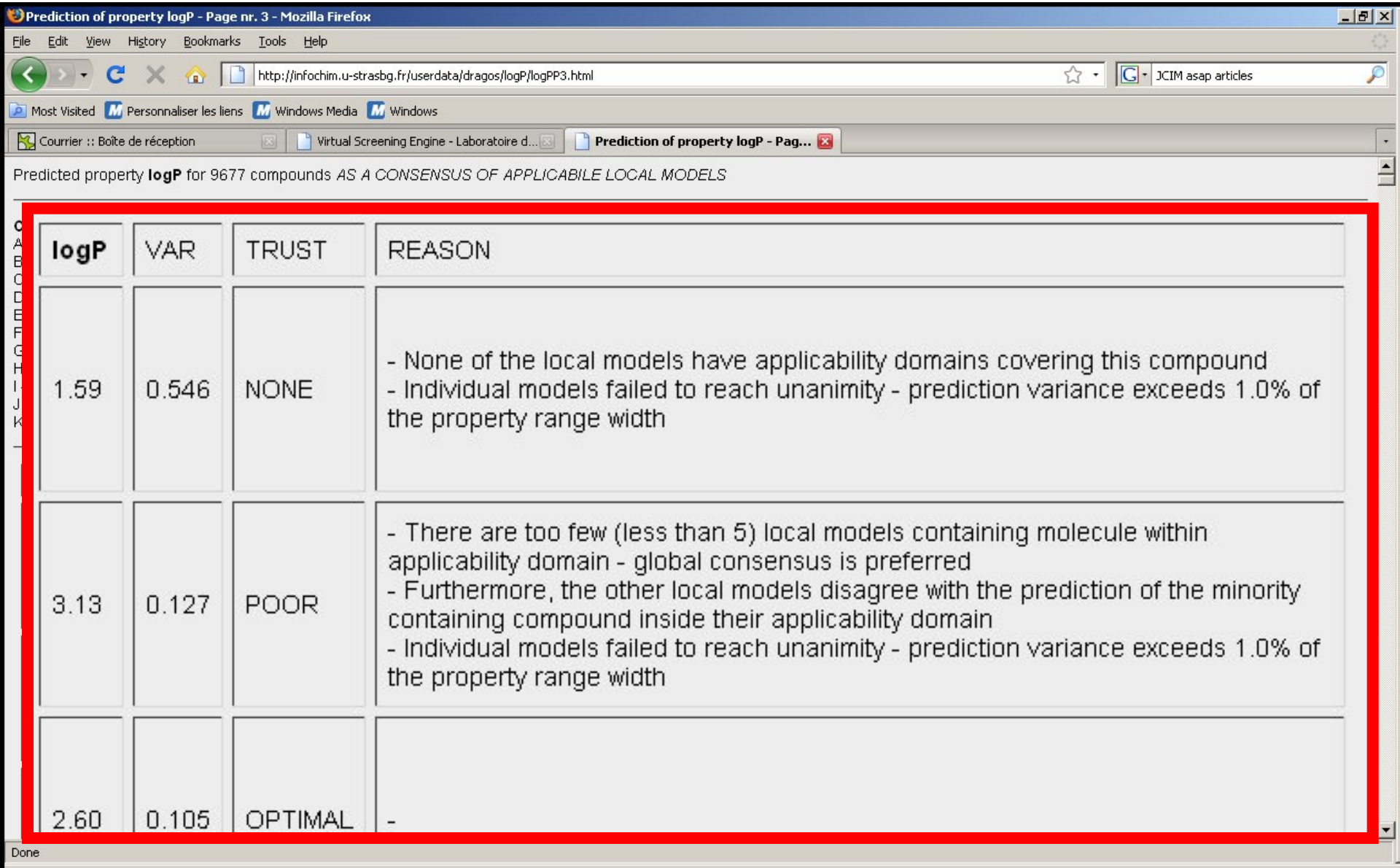
Chemoinformatics tools in SciFinder:



predictions of > 20 physico-chemical properties and NMR spectra for each individual compound

ISIDA property prediction WEB server

infochim.u-strasbg.fr/webserv/VSEngine.html



logP	VAR	TRUST	REASON
1.59	0.546	NONE	- None of the local models have applicability domains covering this compound - Individual models failed to reach unanimity - prediction variance exceeds 1.0% of the property range width
3.13	0.127	POOR	- There are too few (less than 5) local models containing molecule within applicability domain - global consensus is preferred - Furthermore, the other local models disagree with the prediction of the minority containing compound inside their applicability domain - Individual models failed to reach unanimity - prediction variance exceeds 1.0% of the property range width
2.60	0.105	OPTIMAL	-

CoMet project

models for stability constants of metal-ligand complexes in water

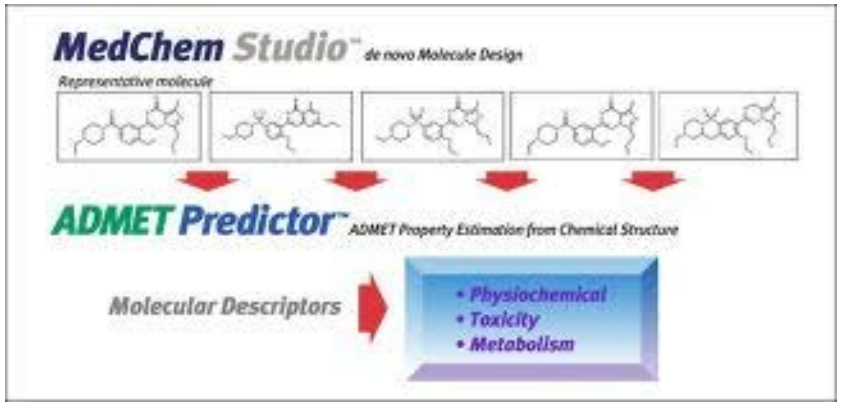
in red: *the models are available*

H																	He
Li	Be											B	C	N	O	F	Ne
Na	Mg											Al	Si	P	S	Cl	Ar
K	Ca	Sc	Ti	V	Cr	Mn	Fe	Co	Ni	Cu	Zn	Ga	Ge	As	Se	Br	Kr
Rb	Sr	Y	Zr	Nb	Mo	Tc	Ru	Rh	Pd	Ag	Cd	In	Sn	Sb	Te	I	Xe
Cs	Ba	La	Hf	Ta	W	Re	Os	Ir	Pt	Au	Hg	Tl	Pb	Bi	Po	At	Rn
Fr	Ra	Ac															

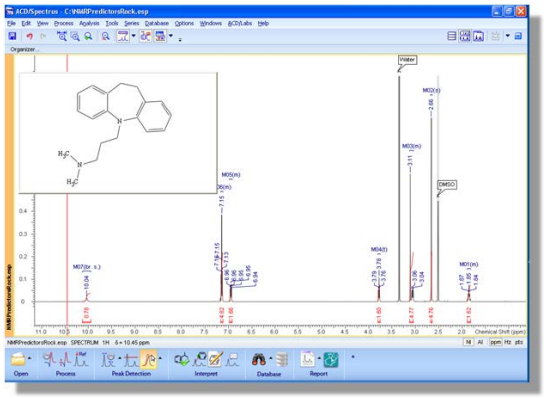
lanthanides	Ce	Pr	Nd	Pm	Sm	Eu	Gd	Tb	Dy	Ho	Er	Tm	Yb	Lu
actinides	Th	Pa	U	Np	Pu	Am	Cm	Bk	Cf	Es	Fm	Md	No	Lr

Commericial and Public Software

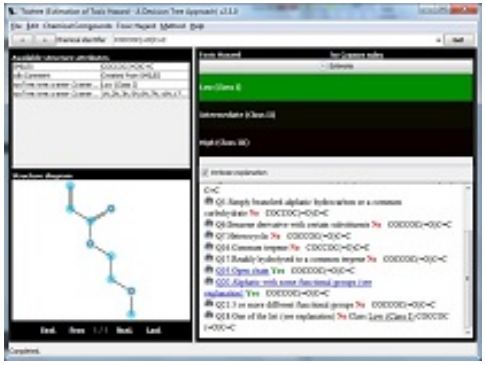
simulations plus, inc.
integrating science and software



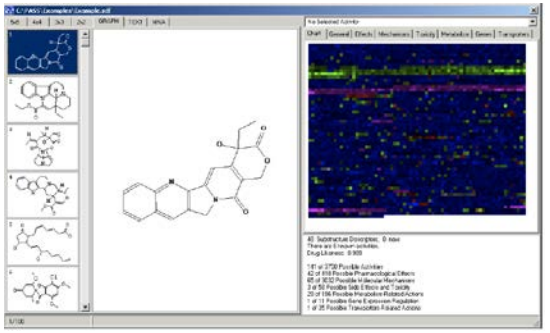
ACD/Labs



Toxtree



PASS



REACh regulation

- The European Union adopted Regulation on the **R**egistration, **E**valuation, **A**uthorisation, and **R**estriction of **Ch**emicals (the "REACH Regulation"), which entered into force on June 1, 2007.

- REACH imposes requirements of information of physico-chemical, toxicology and eco-toxicology parameters for the chemicals, production of which exceeds 1 ton.

- More than 30.000 compounds must be tested. Total cost estimated (EU Commission) over a 11 -15 year period is €2.8 - €5.2 bn

No Data, No Market!

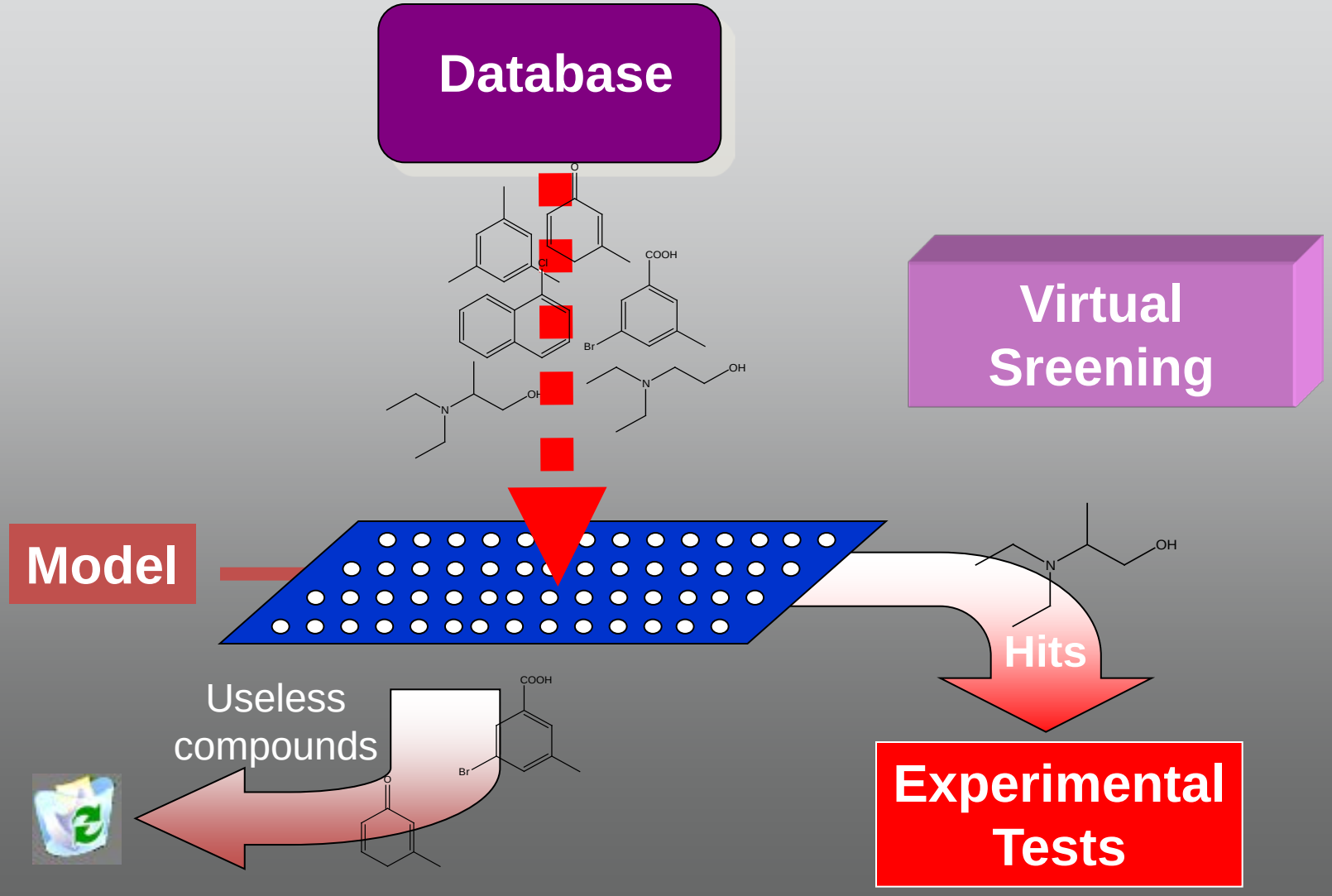
Chemoinformatics:

areas of application

- Drug design (pharmacodynamics and pharmacokinetics),
- Prediction of physico-chemical properties,
- Materials design,
- Synthesis design,
- Molecular spectra simulations

Structure-Property modeling: case studies

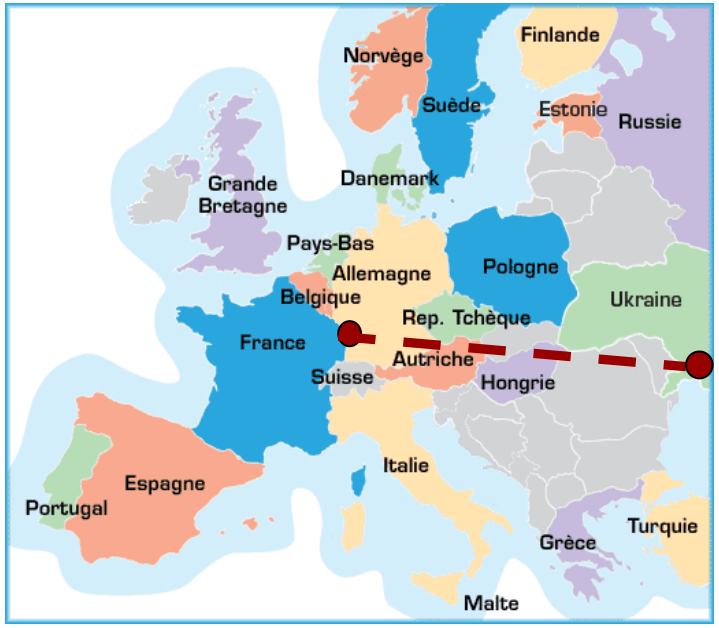
Screening and hits selection



In silico design of new antithrombotics

University of Strasbourg - Bogatsky Institute in Odessa

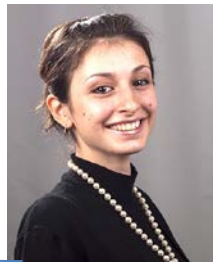
A. Varnek



S. Andronati

V. Kuzmin

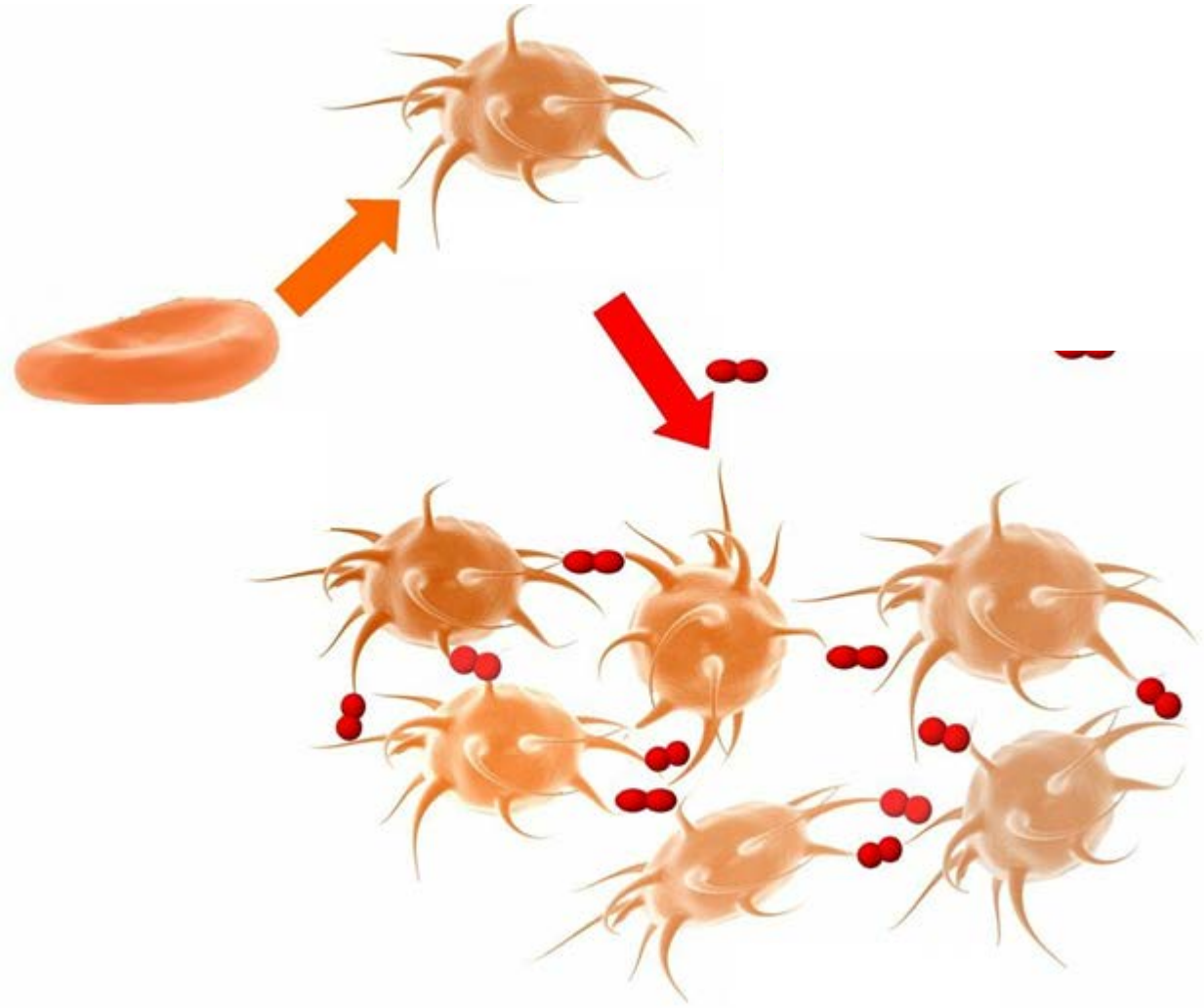
P. Polishchuk



T. Khristova

Co-supervised PhD project

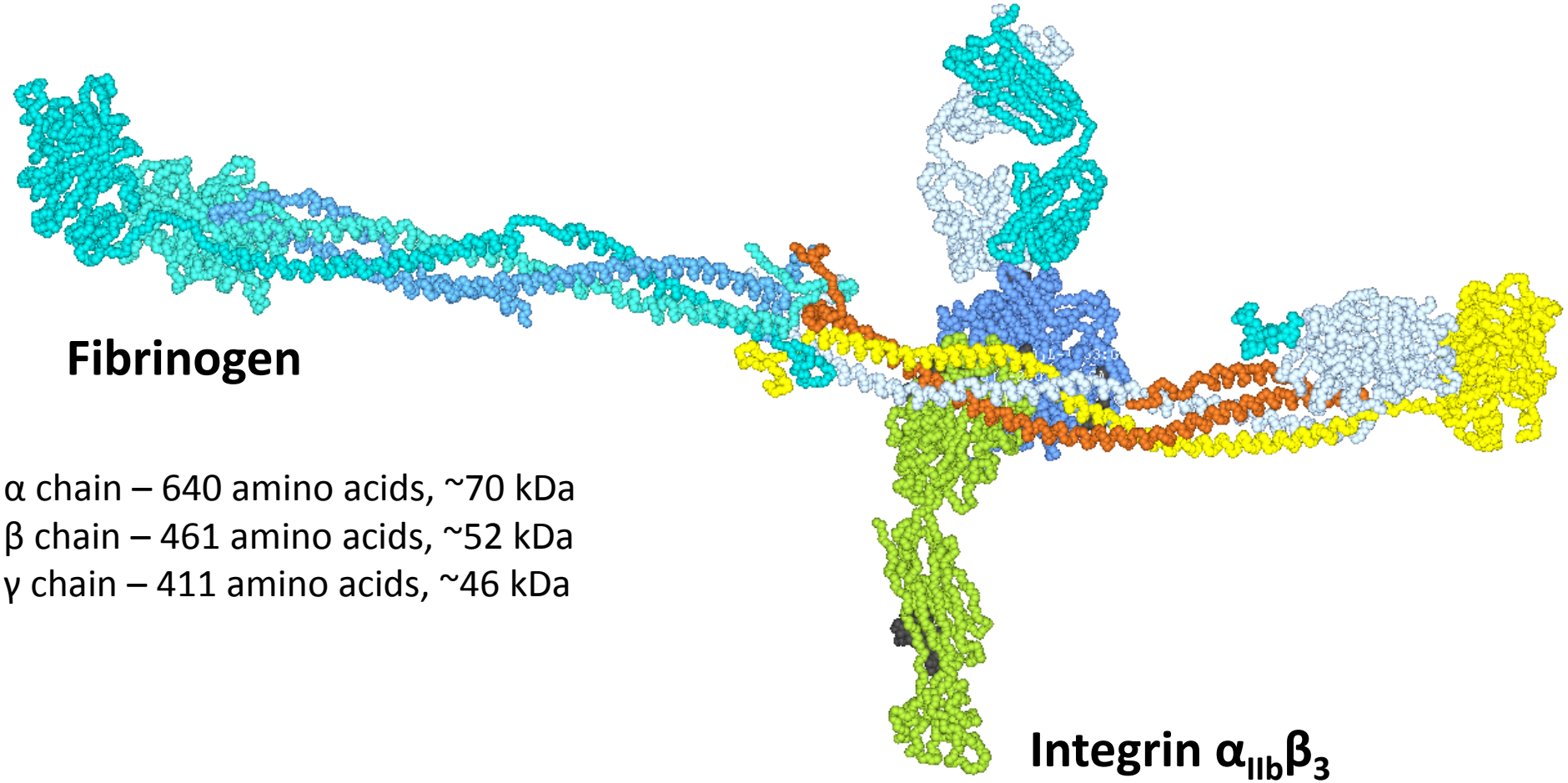
Platelets aggregation



S. Andronati and A. Krysko. Bulletin of Pharmacology and Pharmacy (Ukraine). 2007, № 10. 17 - 27

Protein-Protein interactions:

Complex of *Integrin* $\alpha_{IIb}\beta_3$ and *Fibrinogen*



Fibrinogen

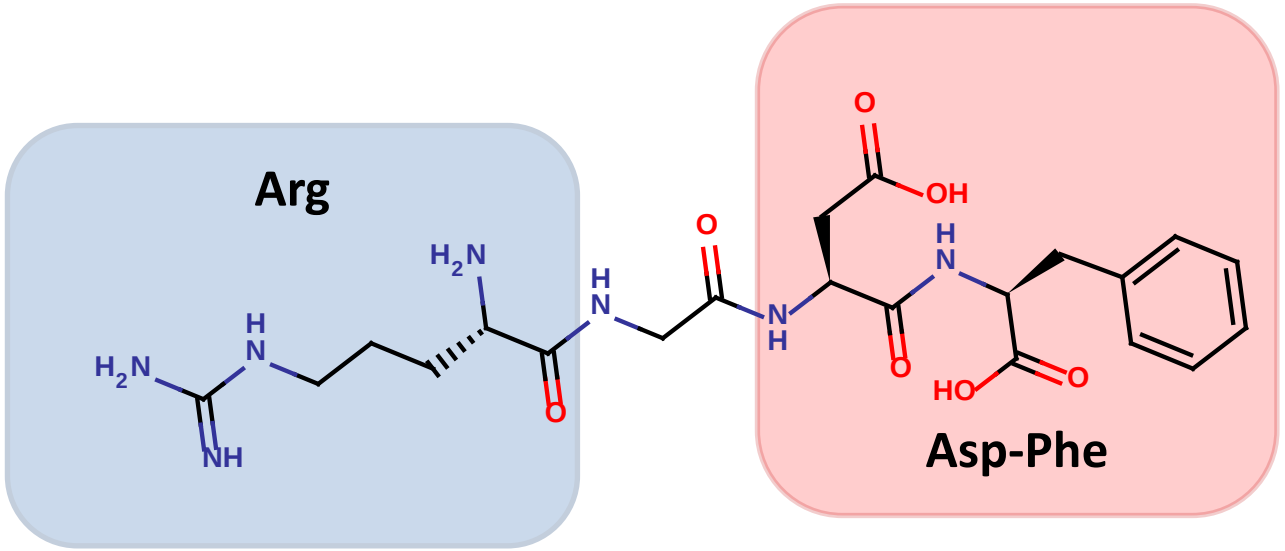
α chain – 640 amino acids, ~70 kDa
β chain – 461 amino acids, ~52 kDa
γ chain – 411 amino acids, ~46 kDa

Integrin $\alpha_{IIb}\beta_3$

α_{IIb} – 870 amino acids, ~94 kDa
β_3 – 762 amino acids, ~84 kDa

Protein-protein docking with Hex 6.3

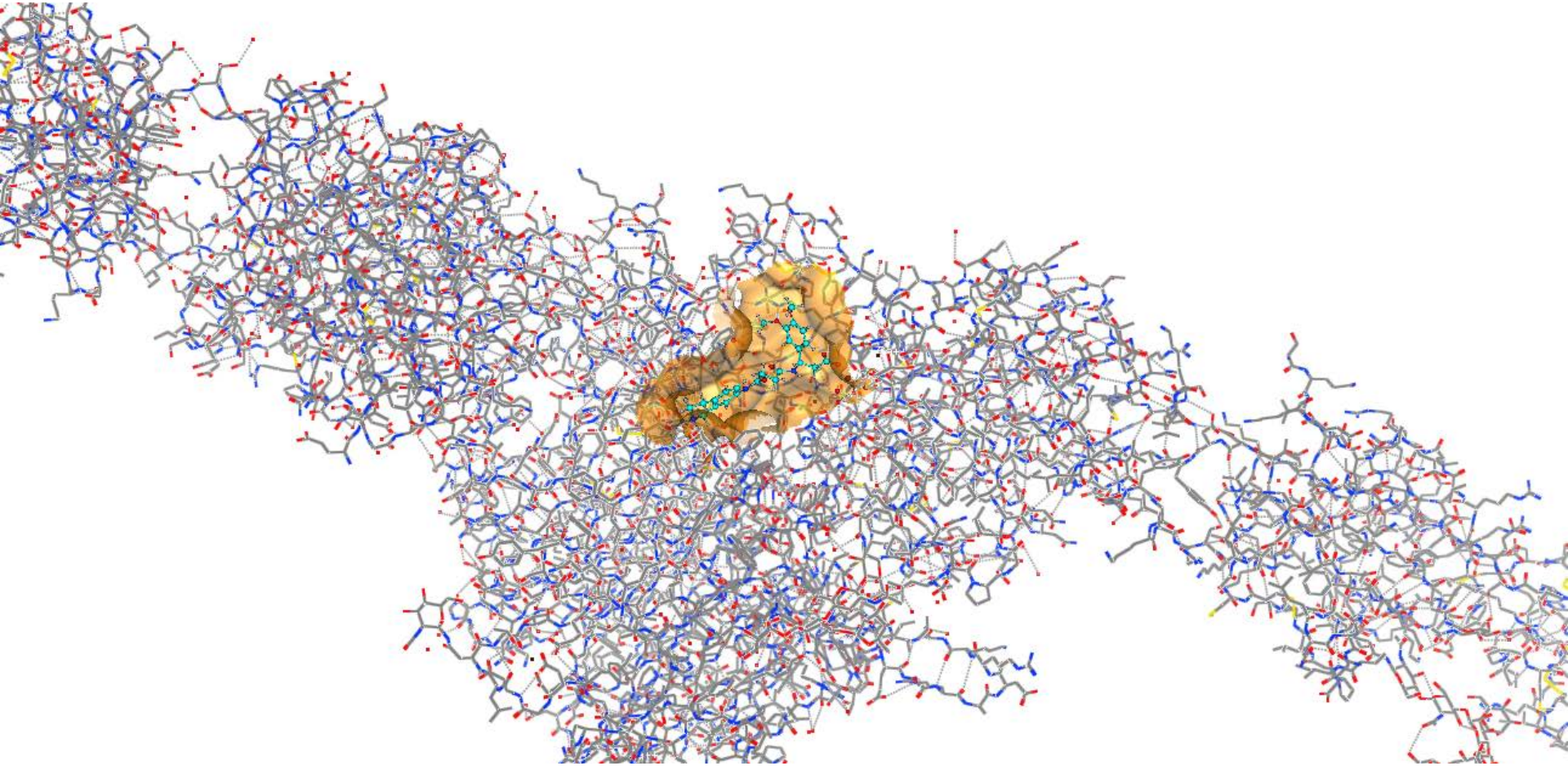
Residues of fibrinogene interacting with binding site of integrine $\alpha_{IIb}\beta_3$



Arg-Gly-Asp-Phe
(RGD)

Protein-Ligand complex:

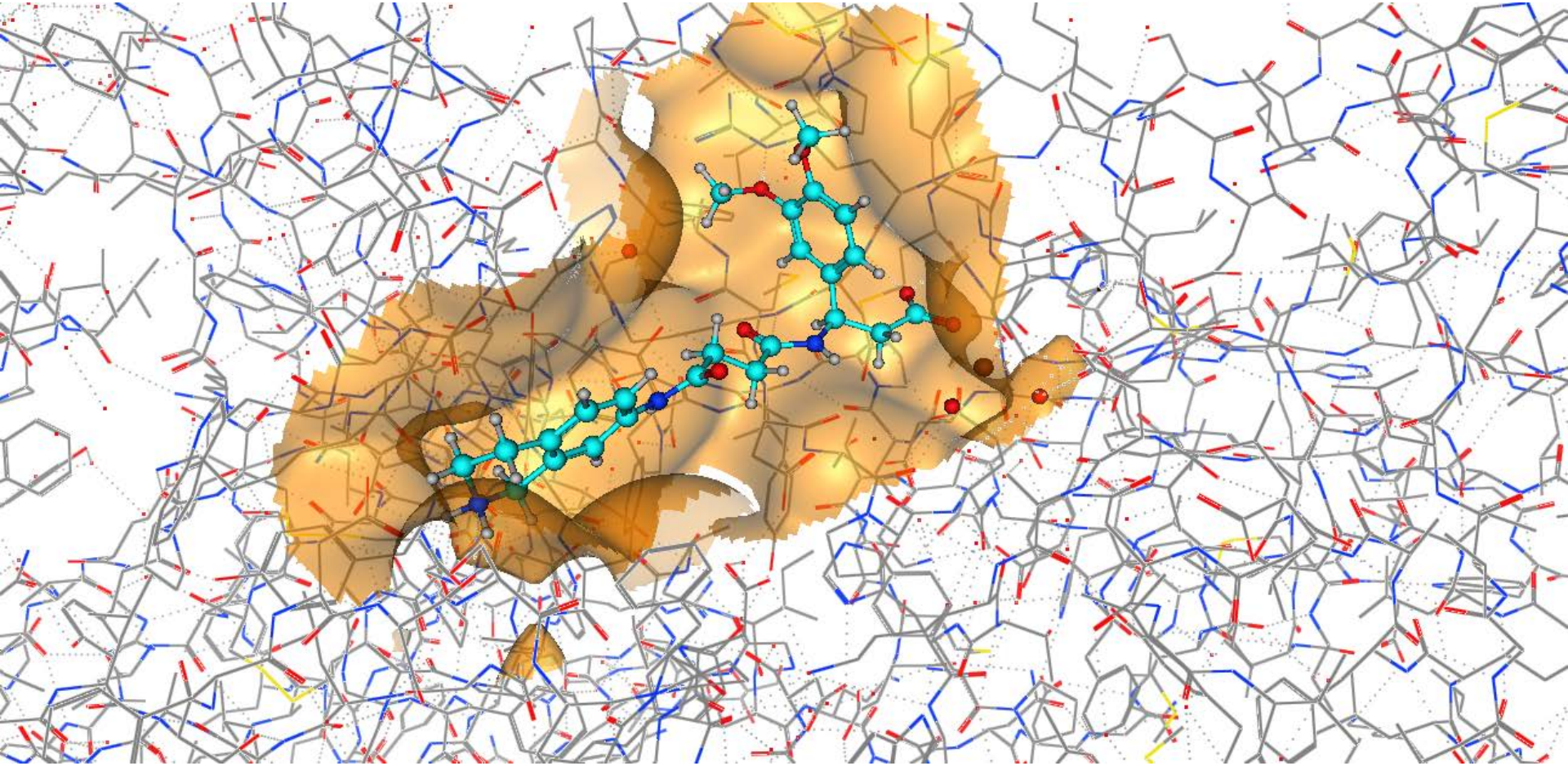
Binding of RGD-mimetic to integrin $\alpha_{IIb}\beta$



Ligand to protein docking with *MOE*

Protein-Ligand complex:

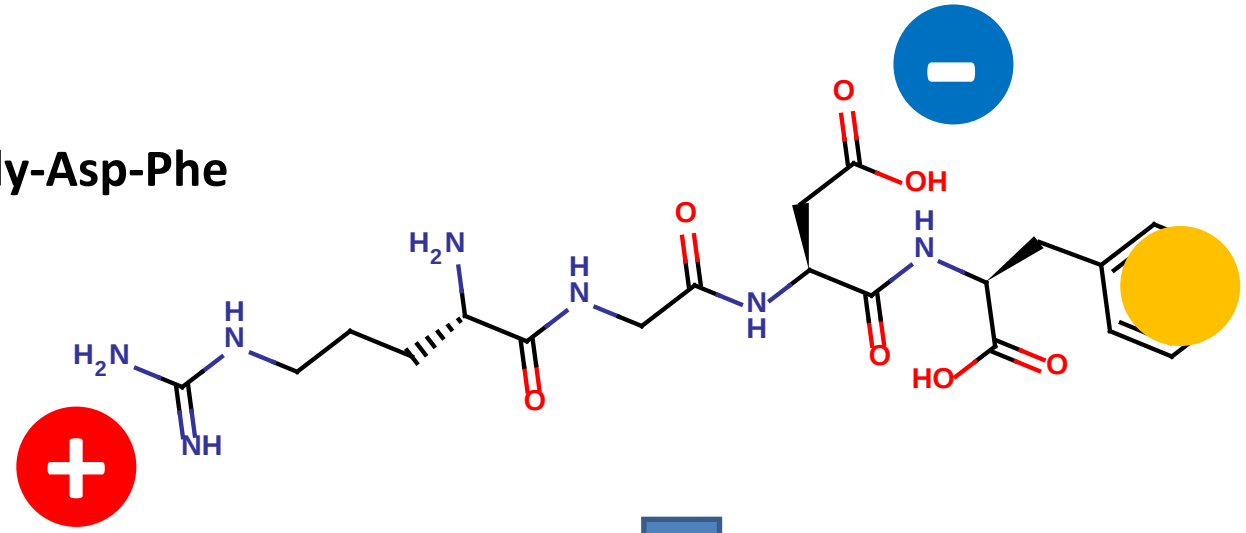
Binding of RGD-mimetic to integrin $\alpha_{IIb}\beta$



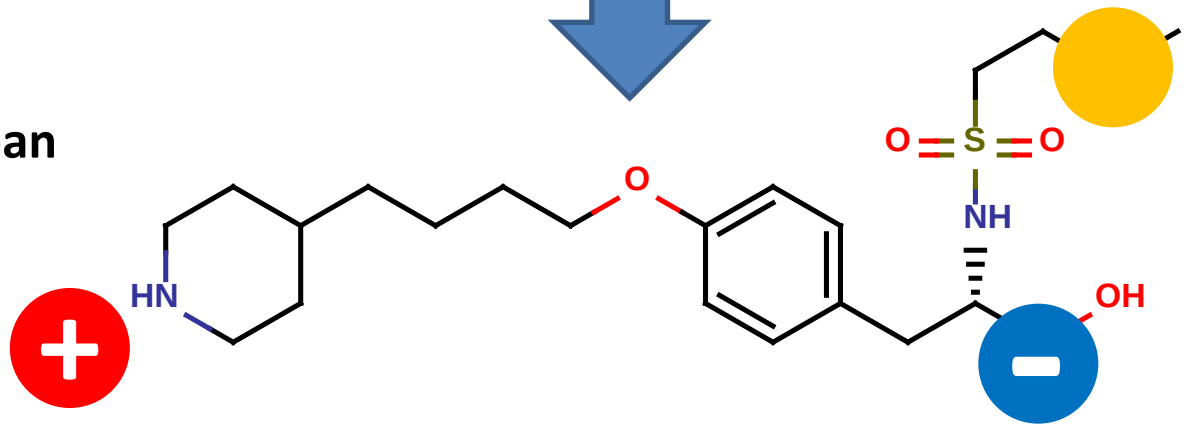
Ligand to protein docking with *MOE*

What is in common between these two molecules ?

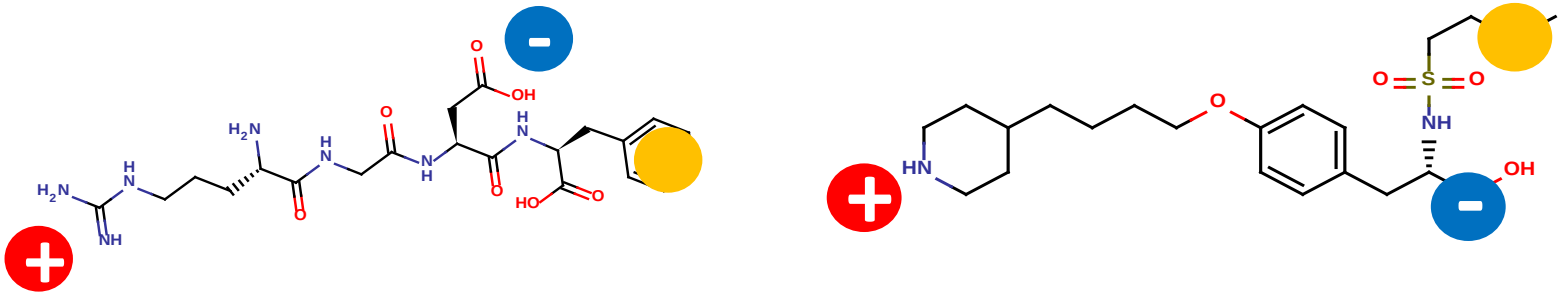
Arg-Gly-Asp-Phe

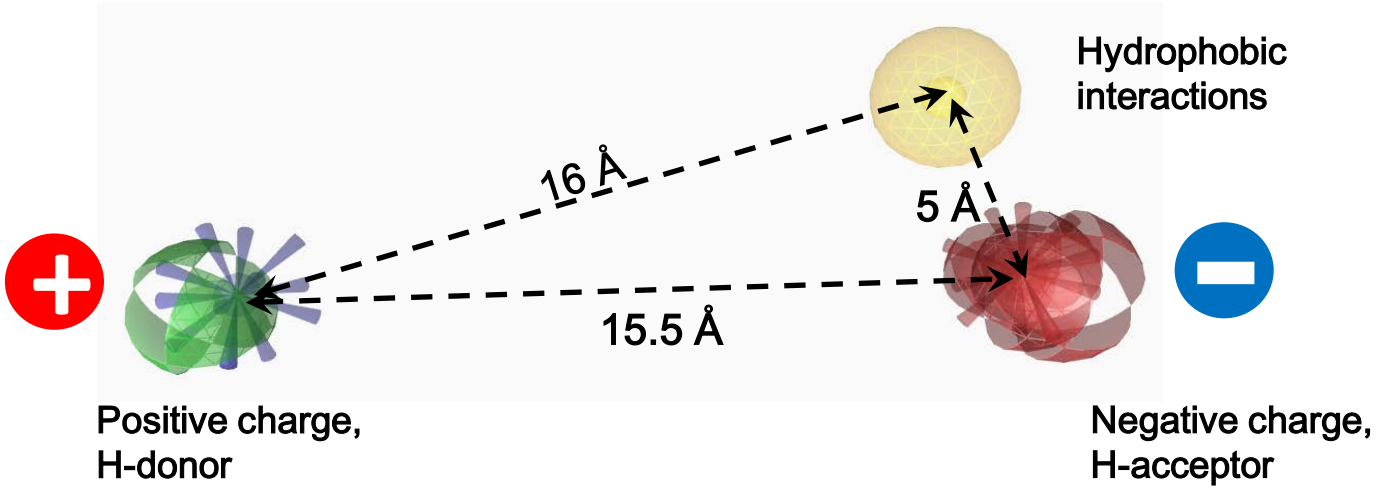


Tirofiban

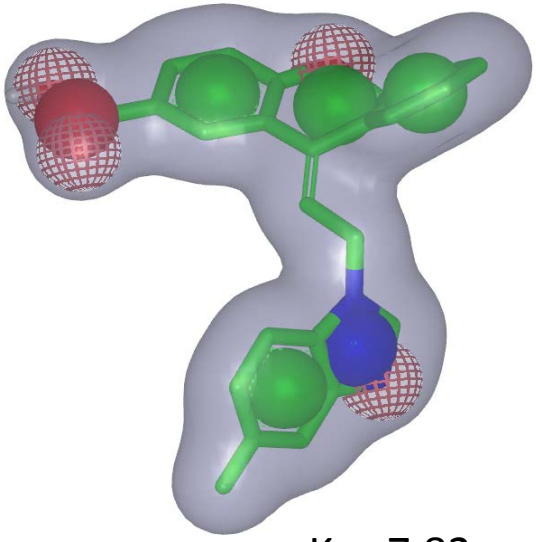


Pharmacophore model of ligand complementary to integrine $\alpha_{IIb}\beta_3$

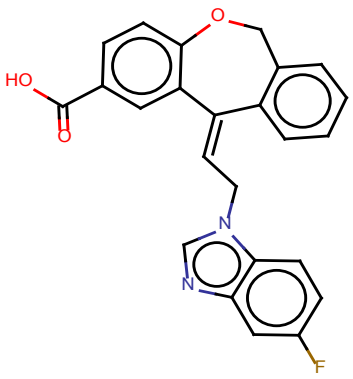


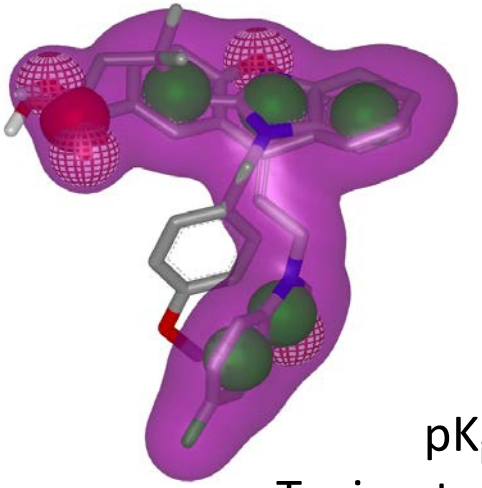


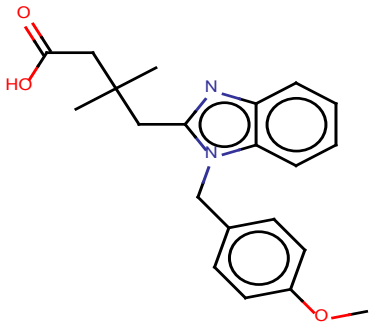
Molecular Shape similarity analysis



$pK_i = 7.82$

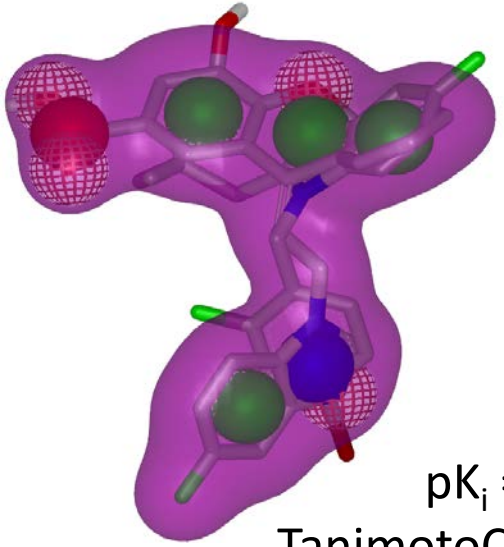


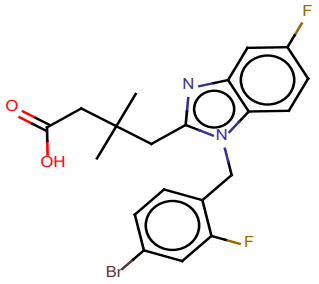




$pK_i = 7.51$

TanimotoCombo = 0.74



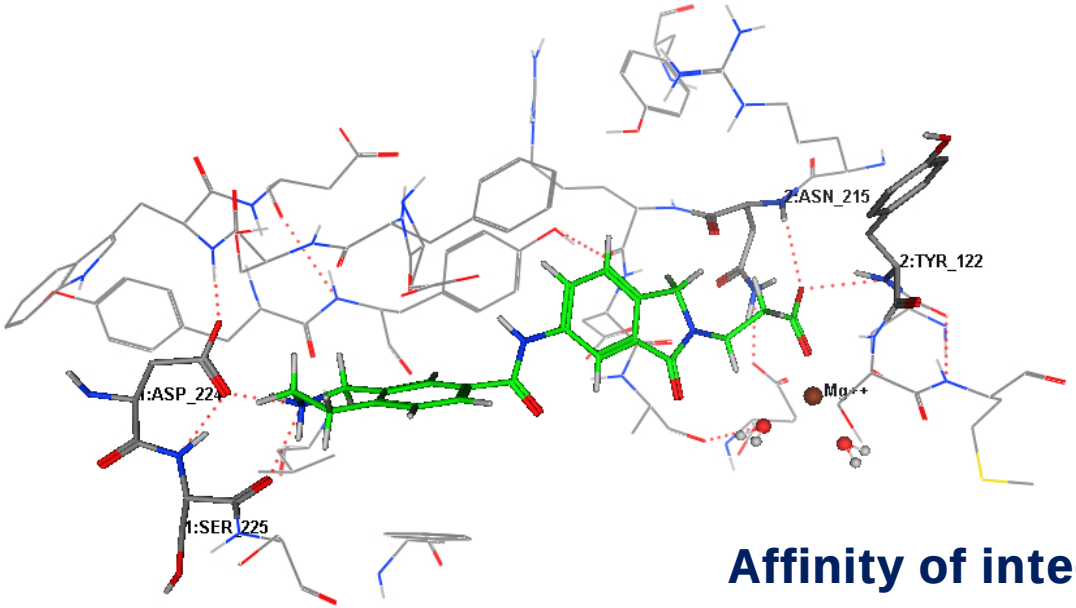


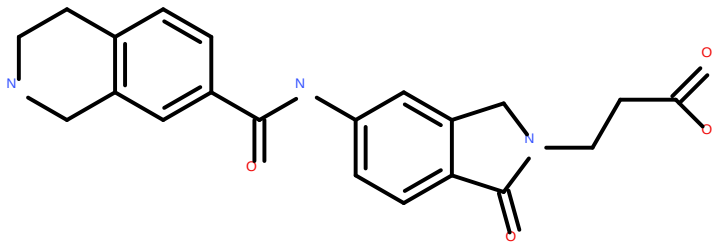
$pK_i = 7.82$

TanimotoCombo = 0.67

Experimental validation:

inhibition of platelet aggregation in human platelet-rich plasma





Affinity of integrin $\alpha_{IIb}\beta_3$ inhibition (IC_{50})

exp = 6.5 nM;

pred = 9.3 nM

Krysko A.A. et al. *Bioorganic & Medicinal Chemistry Letters*, (2011), **21**, 5971–5974

Virtual screening of designed RGD-peptidomimetics

Virtual library generation

6930
(24066 stereisomers)

Pharmacophore model

6783

QSAR

Predicted affinity for $\alpha_{IIb}\beta_3$ pIC_{50} >= 8
Predicted antiaggregation activity pIC_{50} >= 7

143

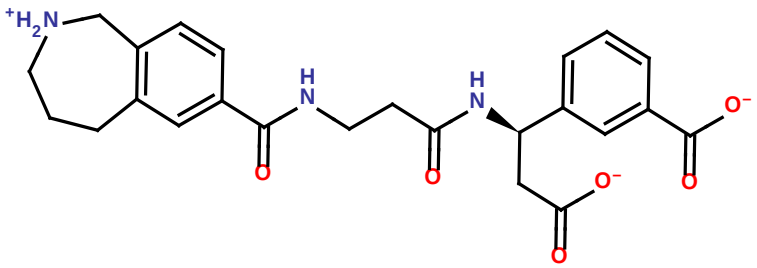
Docking

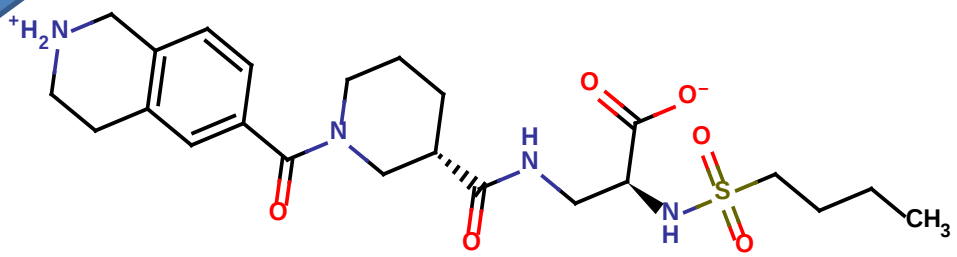
42

Synthetic feasibility

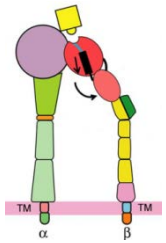
ADME/Tox

19





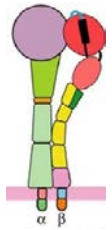
Eight potent αIIbβ3 antagonists with two types of binding modes were developed:



Open headpiece

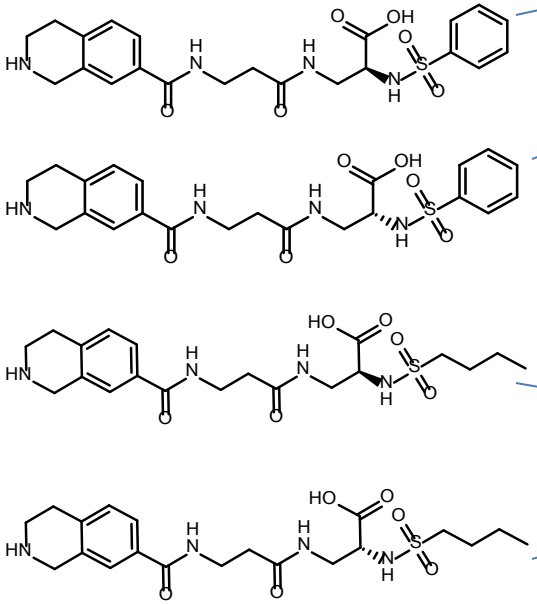
4 “classical”

4 “non classical”

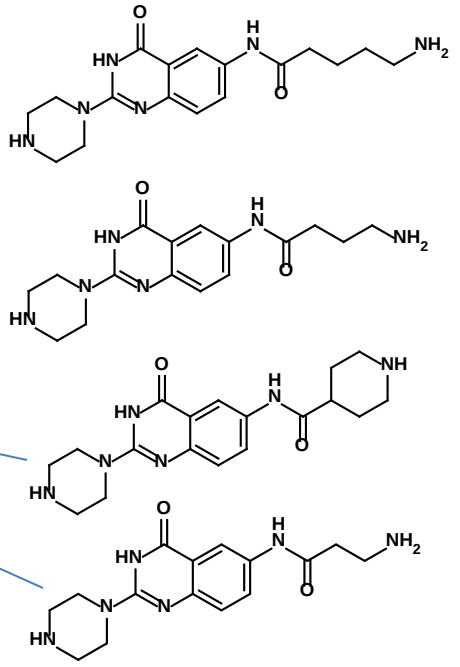


Closed headpiece

Affinity for αIIbβ3



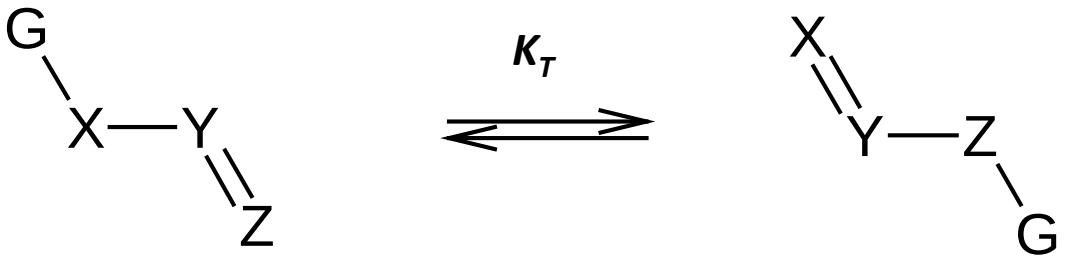
Tirofiban



Prediction of Tautomer Equilibrium Constants

IUPAC definition

Isomerism of the general form:



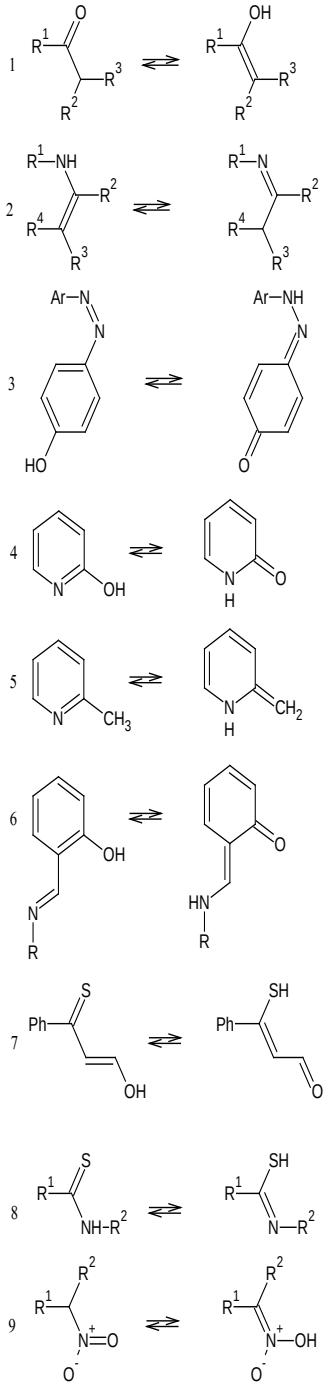
where the isomers (called tautomers) are readily interconvertible

Goal: to build a model predicting logK_T

T.I. Madzhidov, A.S. Petrovsky, R.I. Nugmanov, T.R. Gimadiev, I.I. Baskin, I.S. Antipin, A. Varnek *in preparation*

CONSIDERED TAUTOMERIC EQUILIBRIA

	N
• Keto-Enol	269
• Amino-Imino	105
• Azo-Hydrazine	91
• Pyridol-Pyridone	36
• Pyridinoid-Pyridonoid	6
• Phenol-Imine – Keto-Amine	37
• Thione-Enol – Keto-Thiol	17
• Amine-Thione – Imine-Thiol	82
• Nitro-Acid	13
• Classical Form -Zwitterion	19
• Ring-Chain	50



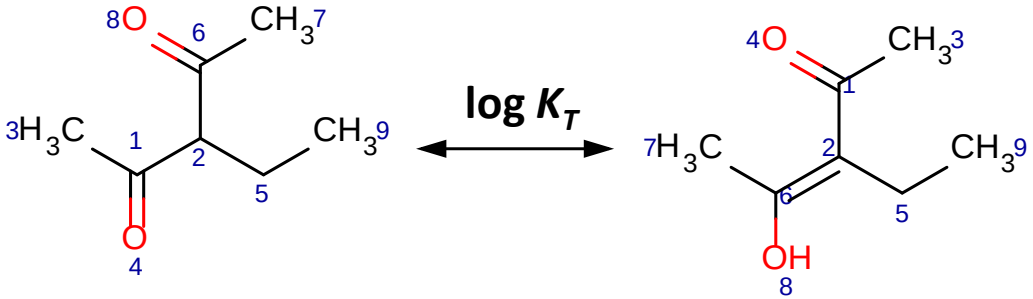
Data workflow

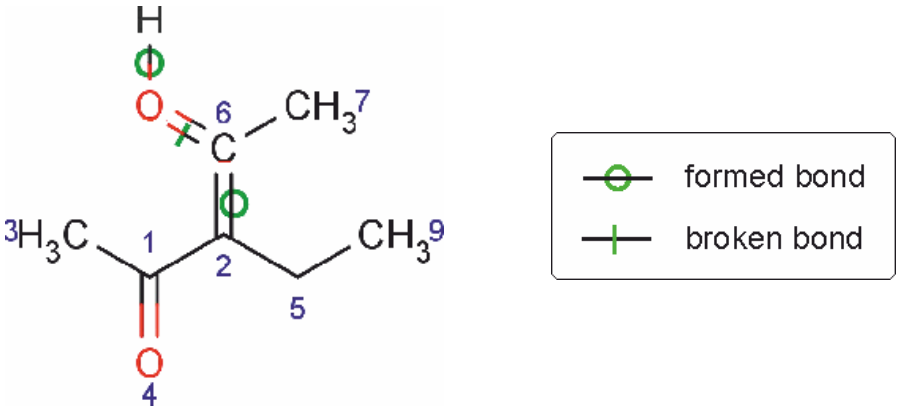
Database

Reactions

Condensed Graph of Reaction (CGR)

723 tautomeric equilibrium ($\log K_T$)
11 types of tautomerism
12 pure solvents
7 water-organic solvent mixtures





Tautomers: Models Performance

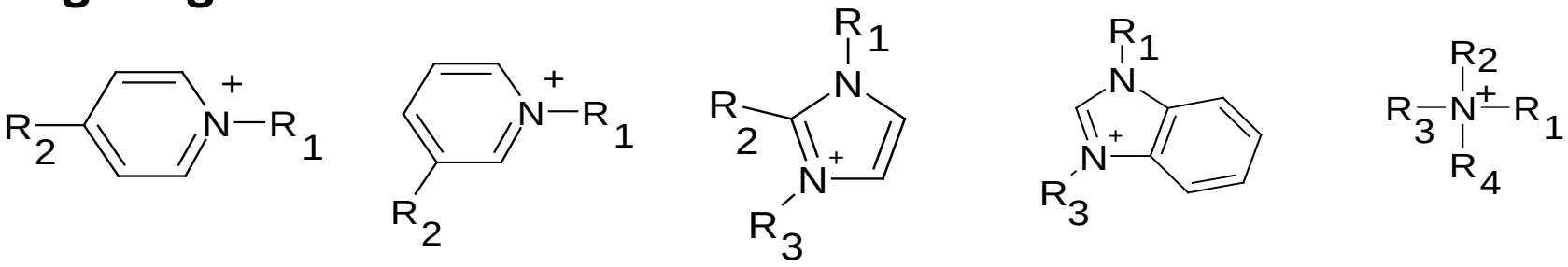
	Our model (cross-val)	Our model (ext set)	QM (semi-empirical)[1]	QM (DFT) [2]	ChemAxon
# molecules	723	31	643	31	127
Correct most stable tautomer (%)	**85**	**81**	**55**	**45**	**56**
RMSE (log K_T)	0.86	1.34	5.99	8.99	5.38

[1] IEF-PCM/PM6
[2] IEF-PCM/B3LYP/6-311++G(d,p)

Ionic Liquids

Ionic Liquids are composed of

large organic cations:

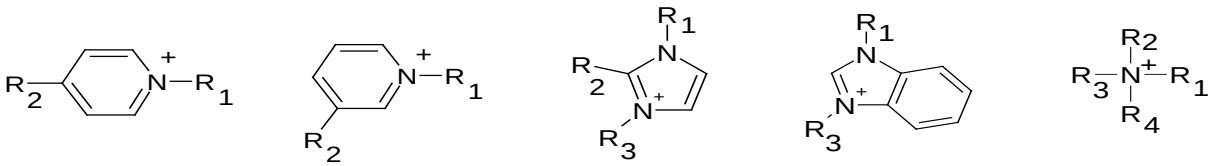


and anions:

PF_6^-, Cl^-, BF_4^-, $CF_3SO_3^-$, $[CF_3SO_2)_2N]^-$

Ionic Liquids

Large organic cations:

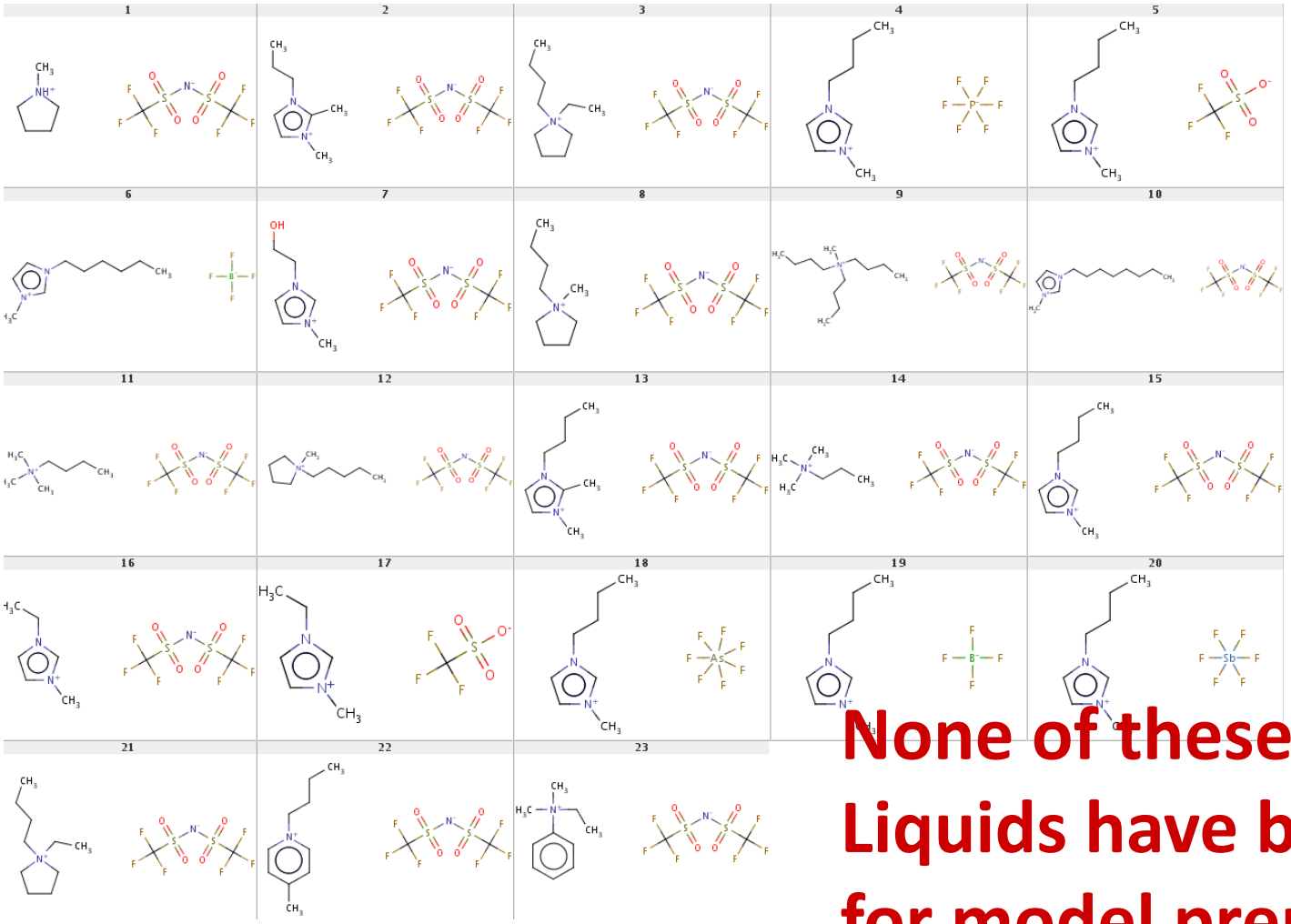


anions:

PF_6^-, Cl^-, BF_4^-, $CF_3SO_3^-$, $[CF_3SO_2)_2N]^-$

There exist 10^{18} combinations of ions that could lead to useful ionic liquids

Viscosity predictions on 23 new ILs

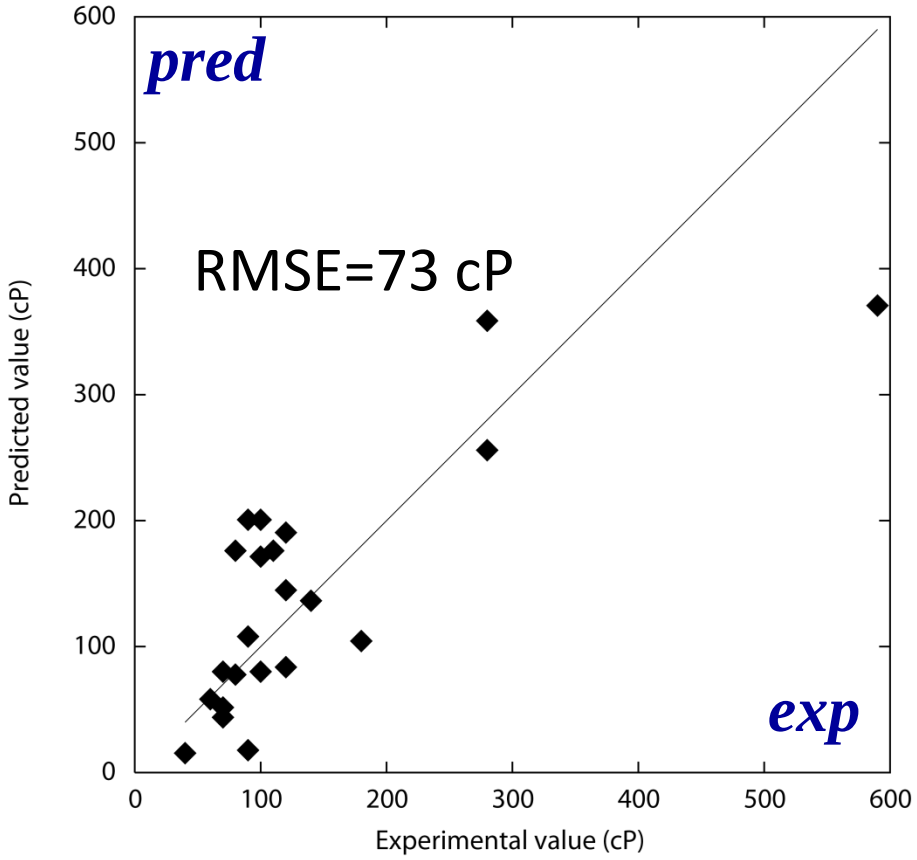


Solvionics company

None of these Ionic Liquids have been used for model preparation

Ionic Liquids viscosity:

Experimental validation of the Neural Networks models

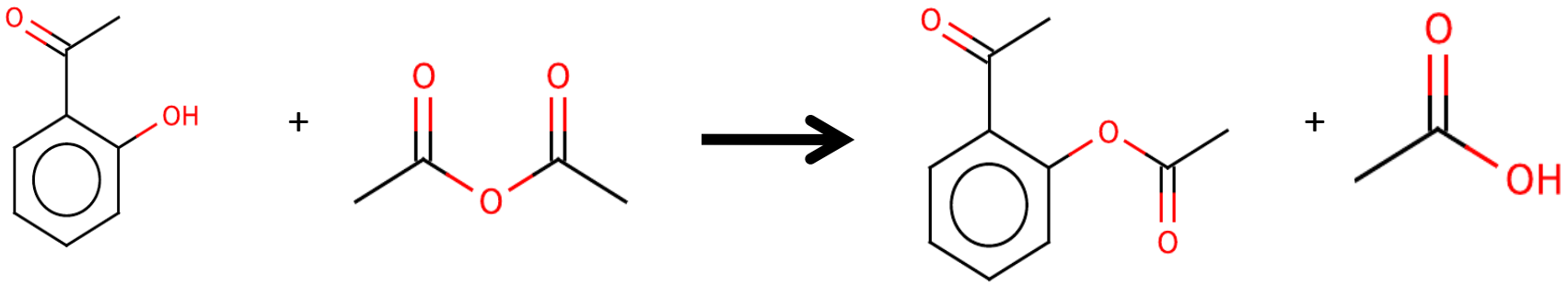


- ***prediction error (~70 cP) is similar to the "noise" in the experimental data used for the training of the model***

G. Marcou, I. Billard , A. Ouadi and A. Varnek,
J. Phys. Chem. B, 2011, **115** (1), 93–98

Prediction of optimal reaction conditions

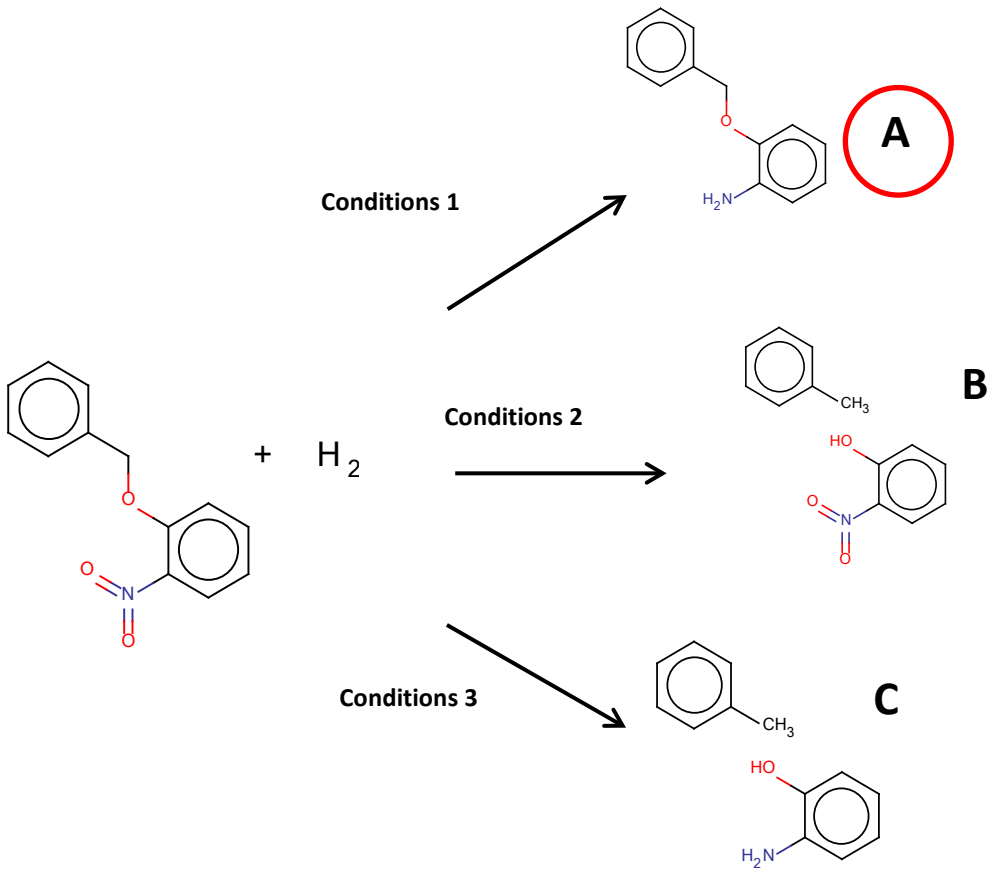
What is a reasonable way to encode a chemical reaction ?



Chemical reactions are difficult objects :

- many species;
- two types of species: reactants and products;
- multi-step reactions,
- dependent on experimental conditions

Selective hydrogenation



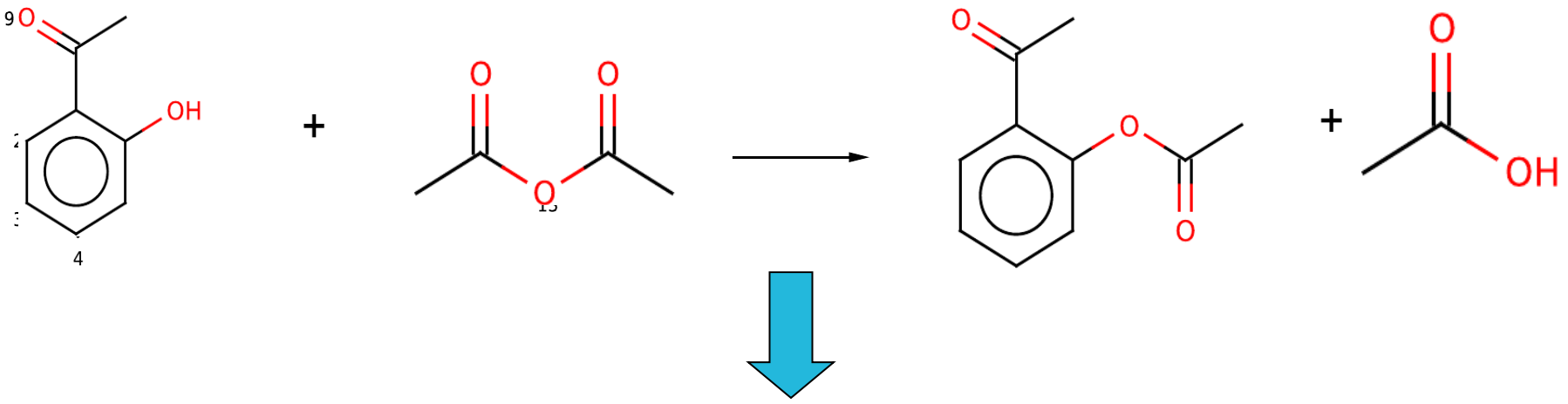
Condition =

- catalyst,
- solvent,
- additives

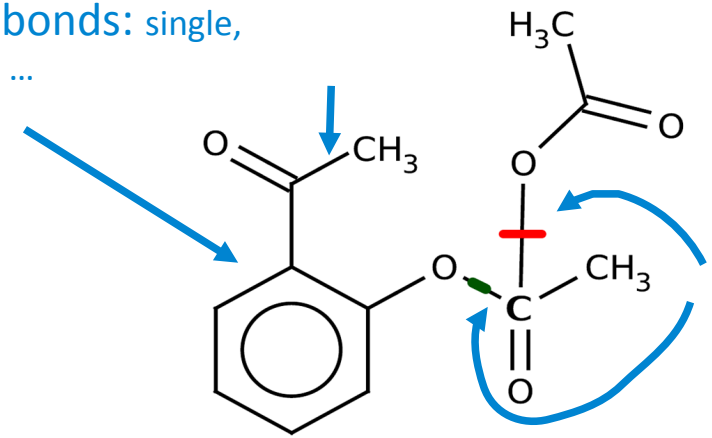
Example reaction leading to products A, B and C under specific conditions.

The goal is to predict reaction conditions leading selectively to product A.

Reactions representation: Condensed Graphs of Reactions



Conventional bonds: single, double, aromatic, …

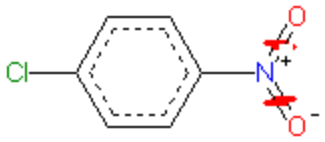
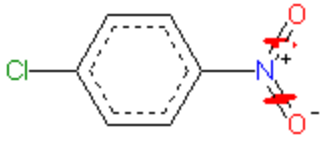


Dynamical bonds:
created single, broken single, …

Notice: atom-to-atom mapping is required

chier Edition Affichage Favoris Outils ?

Précédente Rechercher Favoris

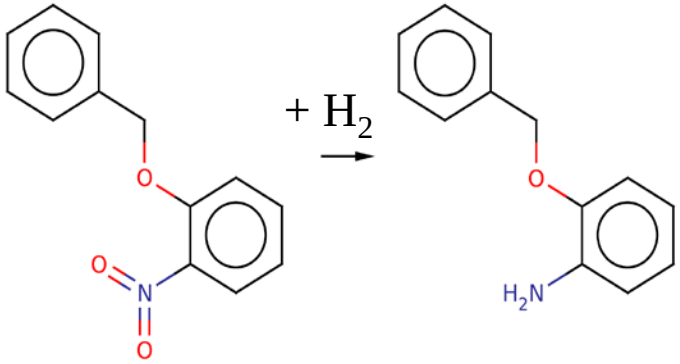
3745	Ir/CaCO3 (5%)	DMF	None	25	1	4-Chloro-phenylamine	75	101650_1259		0.513
	Ir/CaCO3 (5%)	DMF	None					101650_582	c:c-C-O>>c:c-C-O	99
	Ir/CaCO3 (5%)	DMF	None					101650_873	c:c-C-O>>c:c-C-O	100
	Ir/CaCO3 (5%)	DMF	None					101650_873	C-O-c:c>>C-O-c:c	100
	Ir/CaCO3 (5%)	DMF	None					101650_873	c:c:c-O>>c:c:c-O	100
	Ir/CaCO3 (5%)	DMF	None					101650_1298	c:c:c-O>>c:c:c-O	100
3697	Pt/C 10%	THF	None	25	1	4-Chloro-phenylamine	78	101650_1243		0.513
	Pt/C 10%	THF	None					101650_566	c:c-C-O>>c:c-C-O	100
	Pt/C 10%	THF	None					101650_707	c:c-C-O>>c:c-C-O	58

Search deta

- DATABASE
  - Name : C:\c
  - CpdNb : 387
  - DescNb : 3
- QUERY
  - Name : C:\c
  - CpdNb : 1
  - DescNb : 17
- METRIC
  - Type : Tanir
  - CutOff : 49
- FILTER
  - Property : Yi
  - Value : 70
- FRAGMENTATI
  - Type : I(AB,
- WEIGTH
  - Type : reduc

erminé

Internet



Sub A

Conditions suggested by the program

	catalyst	*solvent*	*additif*
1	Pt/C (10%)	THF	None
2	Pt/C (10%)	DMF	None
3	Ir/CaCO3 (5%)	EtOH	NEt3 (5 %)
4	Ir/CaCO3 (5%)	Hexane	None
5	Ir/CaCO3 (5%)	DMF	None

A. Varnek, in "*Chemoinformatics and Computational Chemical Biology*", J. Bajorath, Ed., Springer, 2010

Materials design

Why material informatics is difficult

- **Materials data is more inconsistent** - different measurements and different conditions are used by different labs.
- **Materials data is more unpredictable** - many properties can not be calculated by higher level methods (DFT), especially for amorphous materials. MD is not appropriate for electronic properties.
- **Materials data is more process-dependent** - material properties depend not only on their chemical composition but also on how they are fabricated

Descriptors can be used to quantify the controllable parts of the factors involved, but not the effects of defect and impurities

Courtesy of Prof. C. M Breneman

Material Quantitative Structure-Property Relationship (MQSPR)

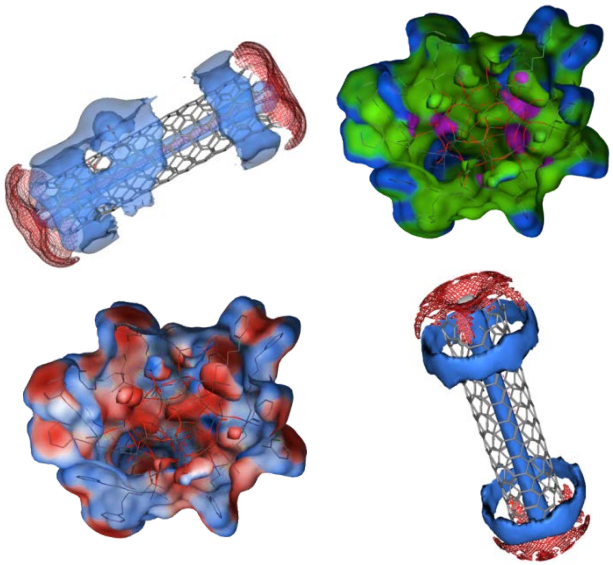
$$Material\ Property = \mathbf{f}\,(descriptors)$$

Molecular Descriptors

Physicochemical Descriptors

Macroscopic properties

Morphology information



Descriptors ⟷ Model ⟷ Property

Representing Potential-Energy Surfaces by Artificial Neural Networks

Eur. Phys. J. B (2014) 87: 152
DOI: 10.1140/epjb/e2014-50070-0

THE EUROPEAN PHYSICAL JOURNAL B

Colloquium

Next generation interatomic potentials for condensed systems

Christopher Michael Handley and Jörg Behler[a]

THE JOURNAL OF CHEMICAL PHYSICS **139**, 054112 (2013)

Permutation invariant polynomial neural network approach to fitting potential energy surfaces

Bin Jiang and Hua Guo[a)]

Representing Potential Energy Surfaces with Neural Networks and High Dimensional Model Representations

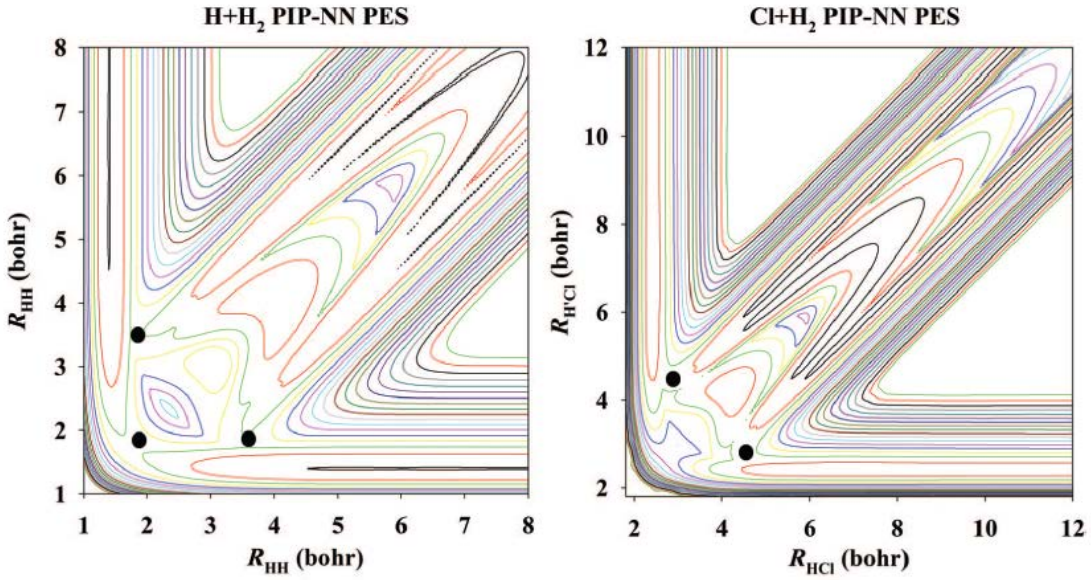
Sergei Manzhos and Tucker Carrington

THE JOURNAL OF CHEMICAL PHYSICS **122**, 084104 (2005)

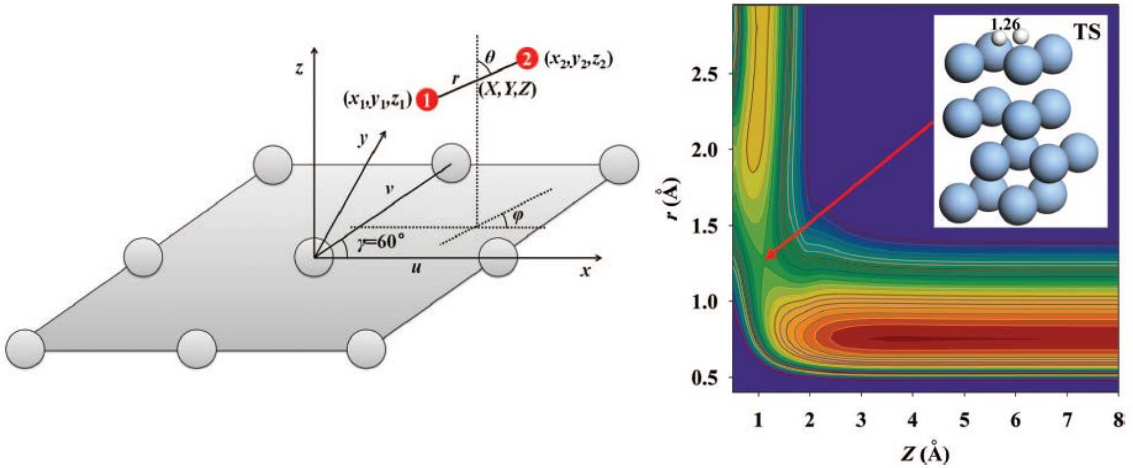
Ab initio potential-energy surfaces for complex, multichannel systems using modified novelty sampling and feedforward neural networks

L. M. Raff

Guo's Permutation-Invariant Neural Network for Approximating Potential-Energy Surfaces



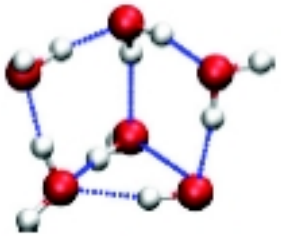
Three- and four-atom reaction systems (*J. Chem. Phys.*, **2013**, *139*, 054112; *J. Chem. Phys.*, **2013**, *139*, 204103; *J. Chem. Phys.*, **2014**, *140*, 044327)



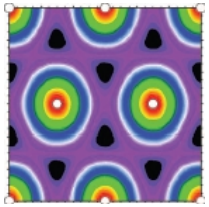
Molecule-surface interactions H_2 + Ag(111) and H_2 + Cu(111) (*J. Chem. Phys.*, **2014**, *141*, 034109)

Behler's Permutation-Invariant Symmetry-Adapted Neural Network for Approximating Potential-Energy Surfaces.

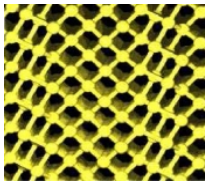
Applications



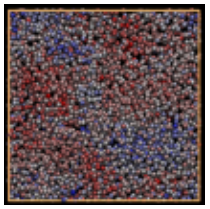
Water clusters (*J. Chem. Phys.*, **2012**, *136*, 064103; *J. Phys. Chem. A*, **2013**, *117*, 7356; *Z. Phys. Chem.*, **2013**, *227*, 1559)



Molecule-surface interactions (*J. Chem. Phys.*, **2007**, *127*, 014705)



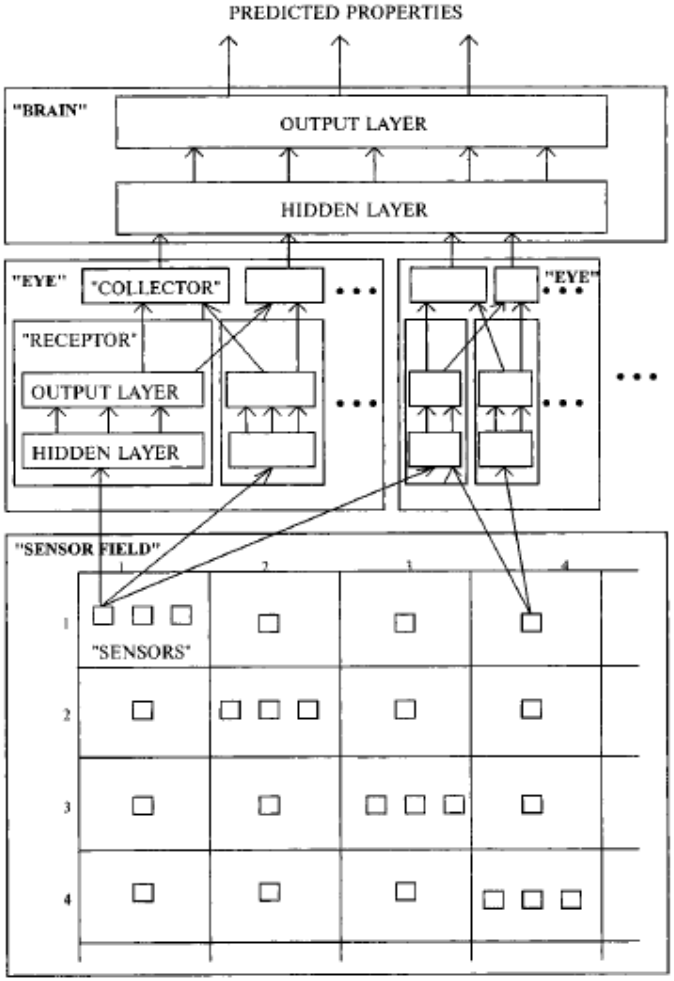
Metadynamics simulations of the high-pressure phases of silicon (*Phys. Rev. Lett.*, **2008**, *100*, 185501)



Large-scale MD simulations of phase change materials GeTe (*J. Phys. Chem. Lett.*, **2013**, *4*, 4241)

The First Permutation-Invariant Neural Network

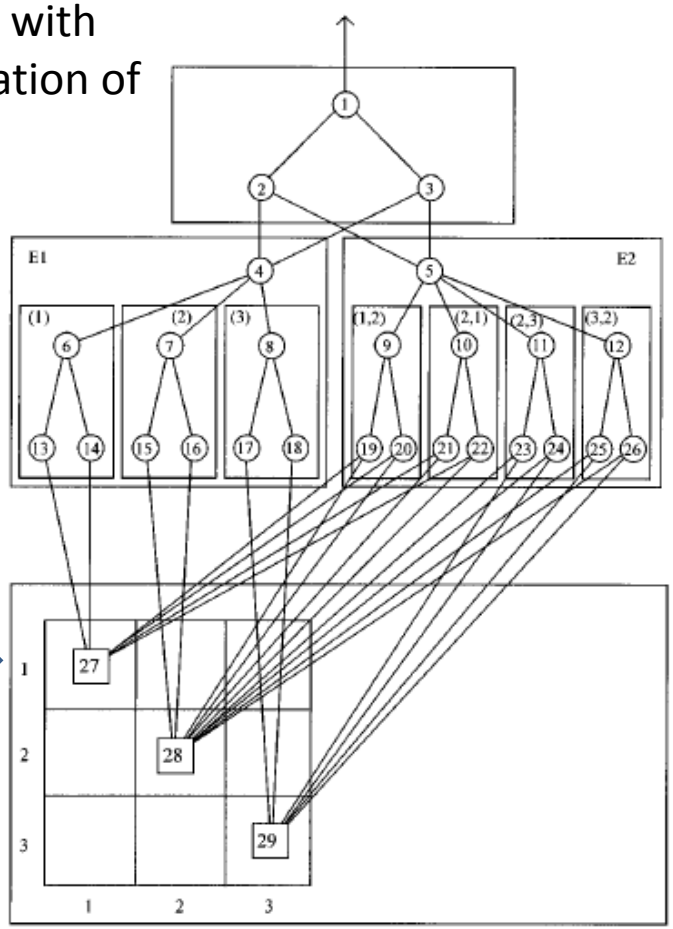
Initially developed for chemoinformatics purposes (QSPR modeling)



Output is invariant with respect to permutation of identical atoms

Properties of atoms (on diagonal)

Distances or bond orders (off-diagonal)



General architecture

Application to 3-atom system

I.I.Baskin et al., *J. Chem. Inf. Comput. Sci.*, **1997**, *37*, 715

Neural Network Potentials:

- now applicable to condensed systems (solids, large clusters, liquids)
- NN reproduces total energies very accurately (also in value!)
- provide analytic derivatives (forces, stress)
- fast and linear scaling with system size
- can be constructed using any electronic structure method

Limitations of NN Potentials:

- need many reference calculations
- limited transferability, no extrapolation
- limited number of chemical elements

Outlook:

- useful tool to speed up extended simulations

Jörg Behler, Energy Landscapes, Chemnitz 2010

Chém Informatique
STRASBOURG

Connecting People