HelmholtzZentrum münchen
German Research Center for Environmental Health

Introduction to Chemoinformatics

Dr. Igor V. Tetko

Helmholtz Zentrum München - German Research Center for Environmental Health (GmbH)
Institute of Bioinformatics & Systems Biology (HMGU)

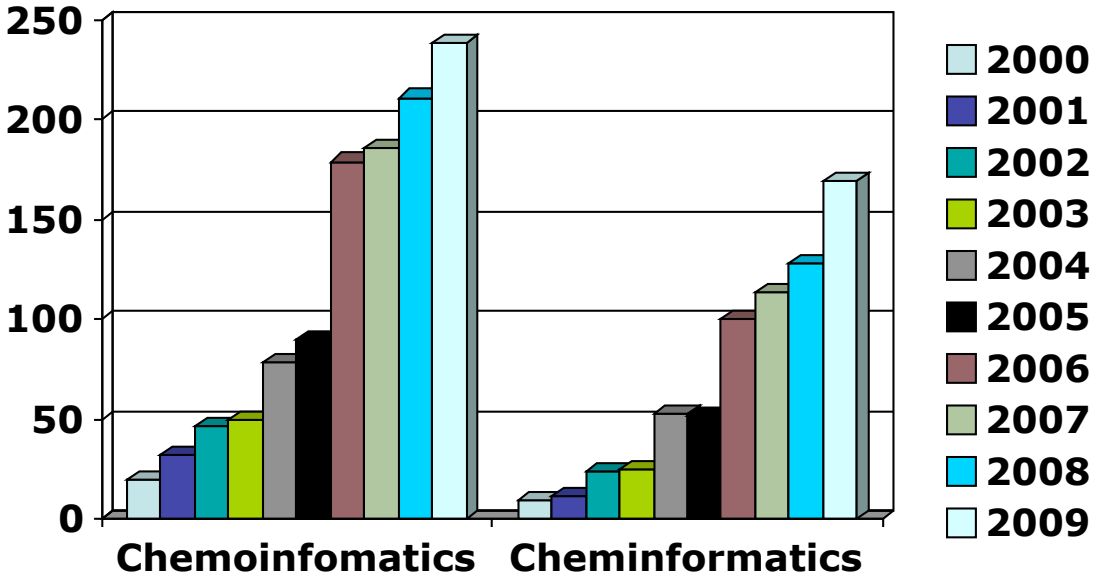
Kyiv, 10 August 2009, Summer School

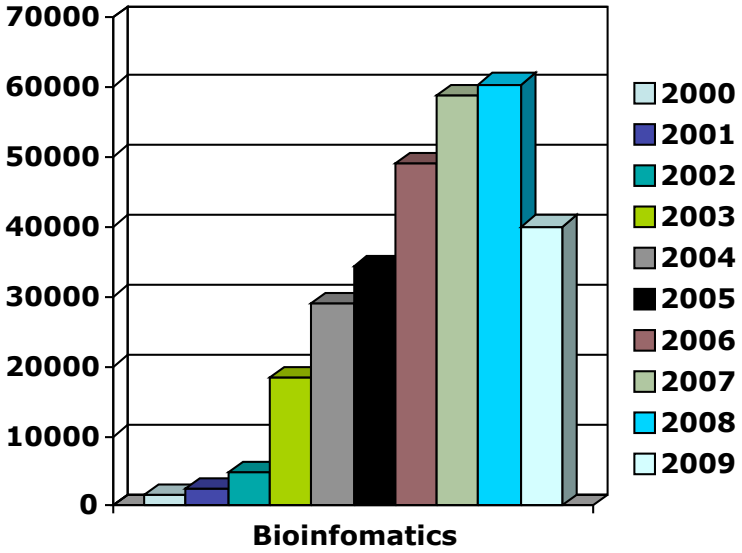
HELMHOLTZ ASSOCIATION

Layout

- Chemoiformatics??? What does it mean?
- Role of chemoinformatics
  - in drug discovery
  - Chemical Biology, NIH Roadmaps
  - REACH, chemical safety
- Definitions of chemoinformatics
- OCHEM
- Overview of the course

Cheminformatics & Bioinformatics keywords in SCOPUS



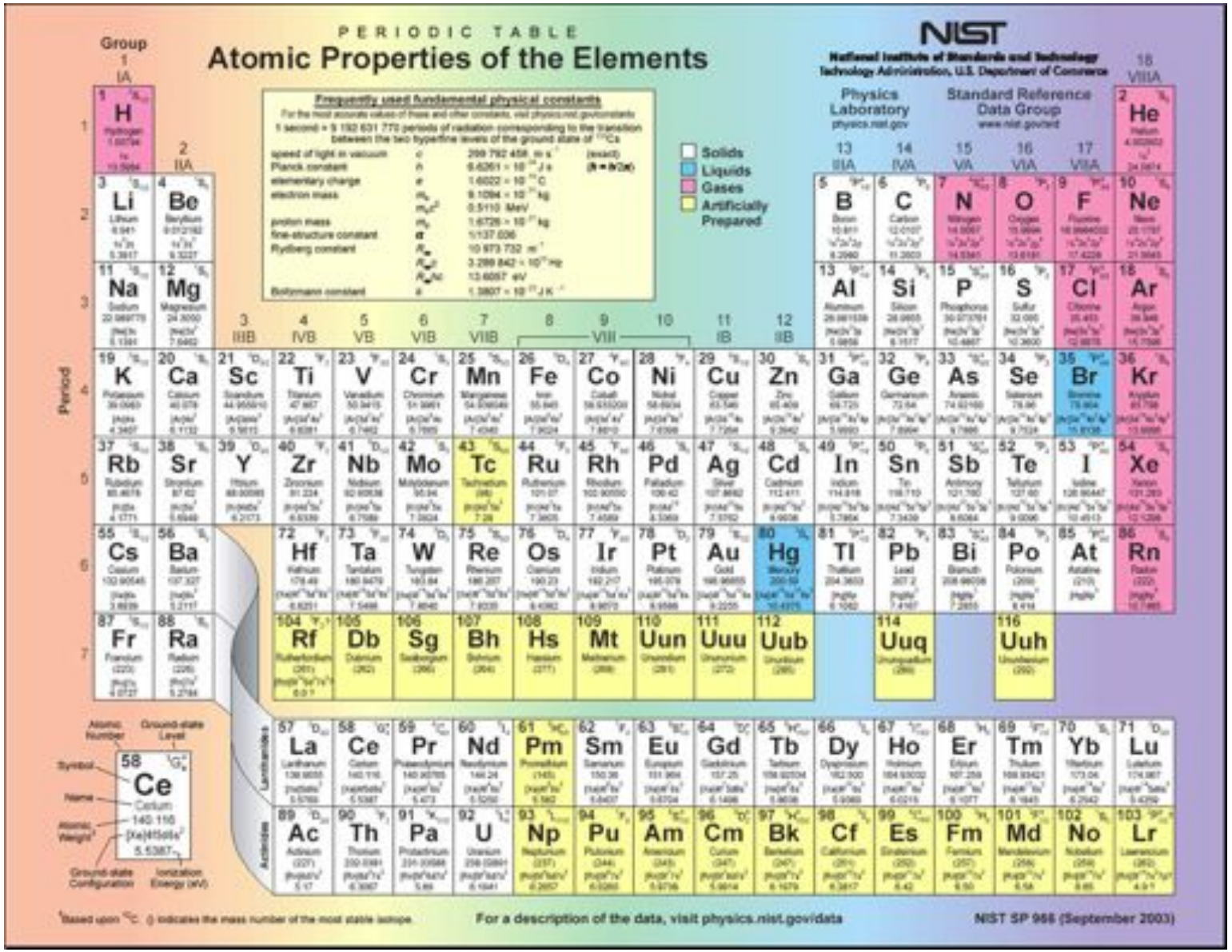


There was no dedicated journal, many authors does not explicitly use world chemoinformatics

about 10x lower if journals and references are exclued

Dmitrii Ivanovich Mendeleev, 1834-1907

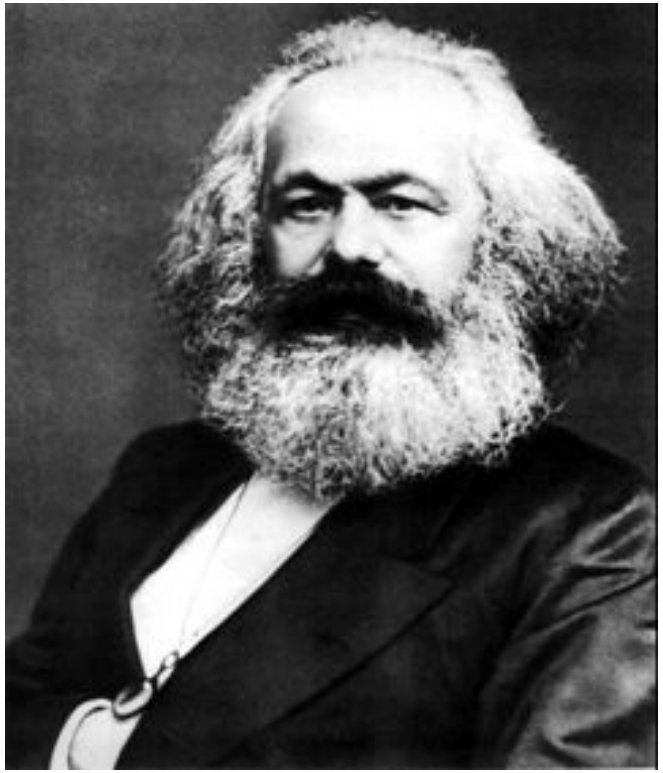
Discoverer of the Periodic Table — An Early "Chemoinformatician"


PERIODIC TABLE
Atomic Properties of the Elements
NIST
National Institute of Standards and Technology
Technology Administration, U.S. Department of Commerce
Physics Laboratory
physics.nist.gov
Standard Reference Data Group
www.nist.gov/srd
Frequently used fundamental physical constants
Solids
Liquids
Gases
Artificially Prepared
Group
Period
Lanthanides
Actinides
For a description of the data, visit physics.nist.gov/data
NIST SP 966 (September 2003)

Why Mendeleev?

Faced with a large amount of data, with many gaps, Mendeleev:

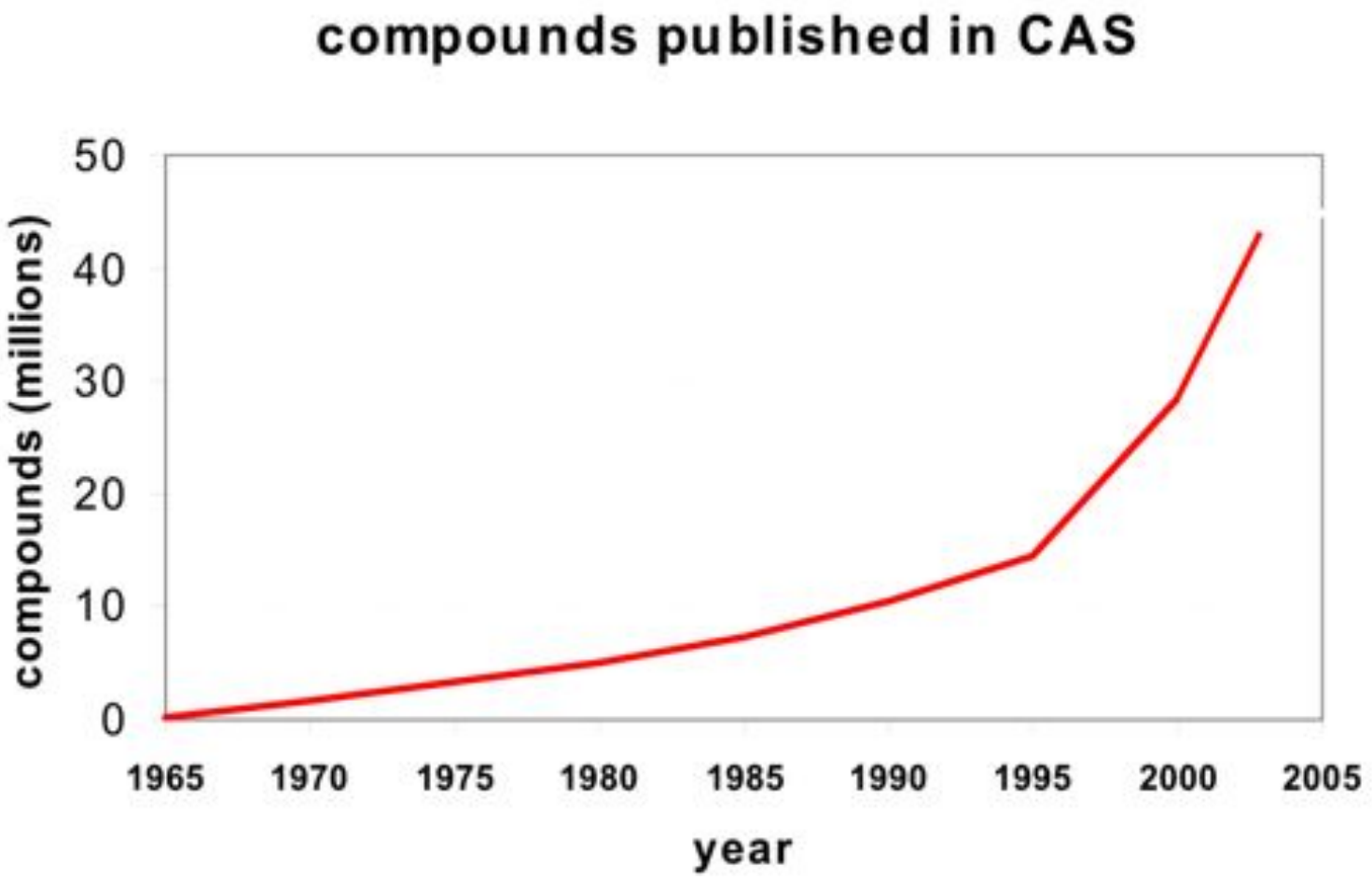
- Sought patterns where none were obvious,
- Made predictions about properties of unknown chemical substances, based on observed properties of known substances,
- Created a great visualization tool!



Why he? Why at that time?

"The **Law** of transformation of **Quantity** into **Quality"**
by Karl Marx

Quantity aspects of chemoinformatics



Major challenges of Chemoinformatics

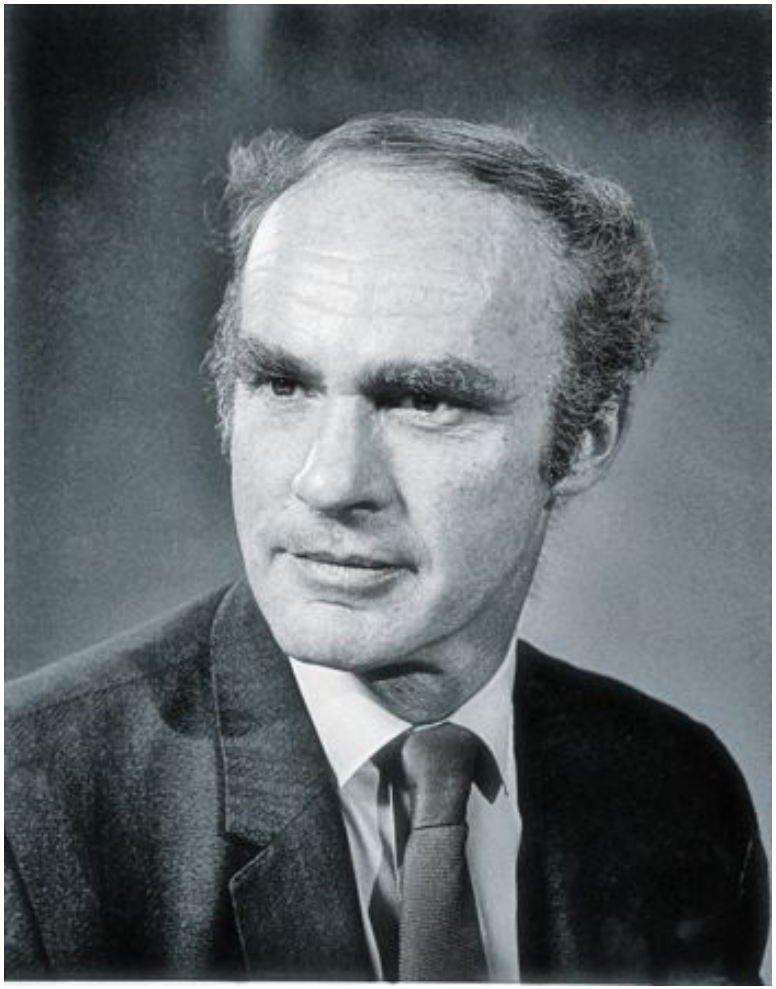
Millions of structures
Thousand of publications

Storage, organization and search of information
QSAR / QSPR studies: prediction of properties and activities

Chemical biology & drug discovery
In silico environmental toxicology, REACH

QSAR/QSPR - Quantitative Structure-Activity (Property) Relationship Studies

Goal of chemistry: Synthesis of Properties



The most fundamental and lasting objective
of synthesis is
not production of new compounds
but
production of properties

George S. Hammond
Norris Award Lecture, 1968

Where Chemoinformatics is required?

Pharma companies

- data collection and handling
- model development QSAR/QSPR, *in silico* design
- *In vitro, in vivo* data analysis and interpretation

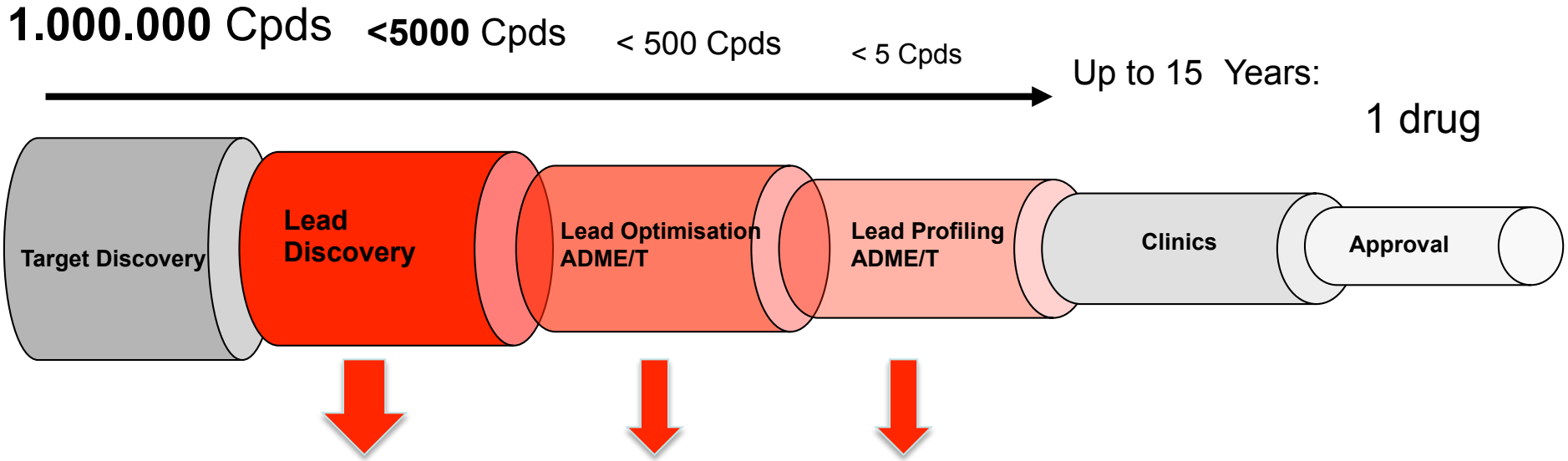
REACH

- > 140,000 compounds
- Development of *in silico* methods to predict toxicity of chemical compounds

Chemical industry

- Design of new properties; chemical synthesis
- Prediction of toxicity of chemicals BEFORE they enter market

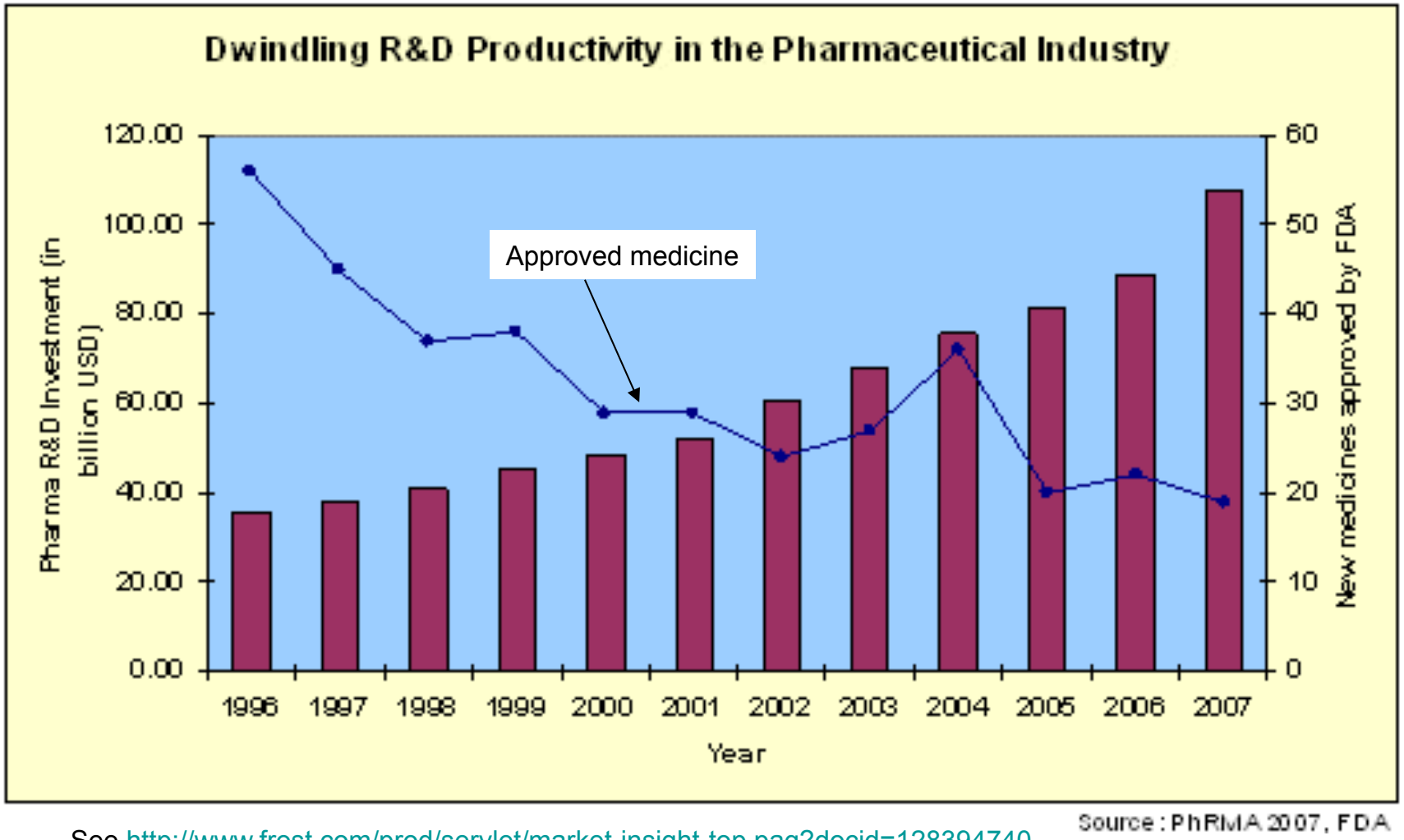
Pharma R&D: Cost and Productivity issues
Compound numbers



In vitro and in vivo property determination:
Millions of screens for solubility, stability, absorption, metabolism, transport, reactive products, drug interactions, etc etc

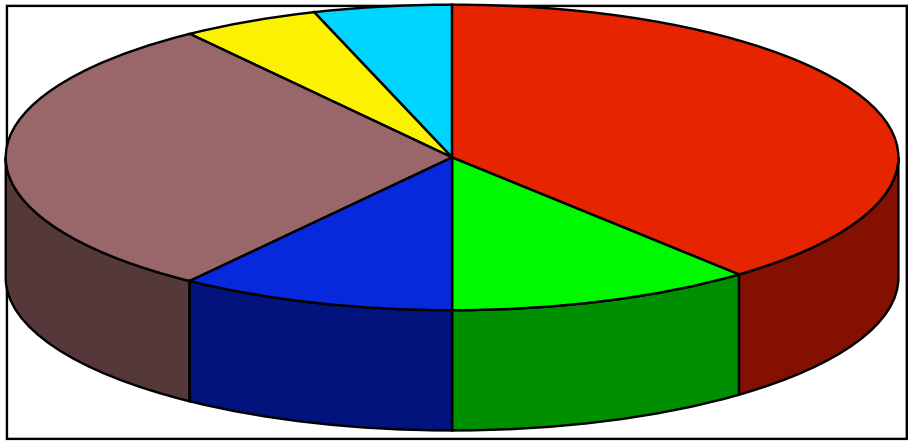
Preclinics Costs: > $300m PER COMPOUND to reach approval

Declining R&D productivity in the pharmaceutical industry



See http://www.frost.com/prod/servlet/market-insight-top.pag?docid=128394740

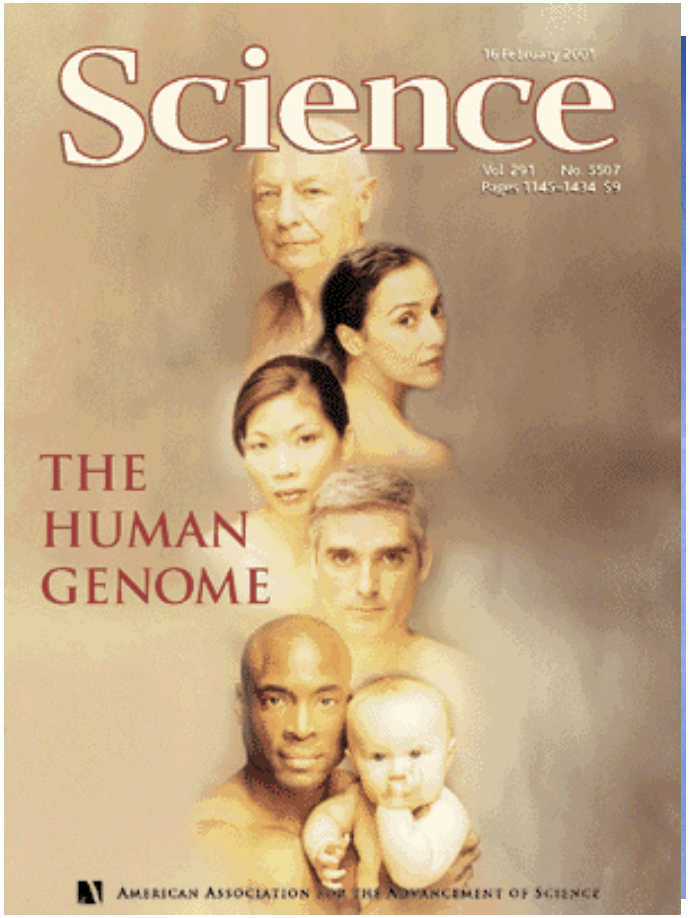
Pharma R&D Cost and productivity: Reasons for compound failure



TOP four reasons are connected to compound
Absorption, **D**istribution, **M**etabolism and **E**xcretion, all of which may contribute to lack of efficacy and **T**oxicity: **ADME/T** issues

Chemoinformatics!

NIH Roadmaps



Chemical Biology

Dr. Elias Zerhouni
Former NIH director (2002-2008)
$29.5 billions in 2008 for
27 Institutes

Ukraine GDP is ca $180 billions
Belarus GDP is ca $60 billions

NIH Roadmaps

New pathways to discovery

- understand **complex biological systems**
- understand their **connections**
- build **better "toolbox"** for researchers

Research teams of the future

Re-engineering the clinical research team

Molecular Libraries Screening Center Network (MLSCN)

- *10 centers were created*
- *250 assays to screen >200,000 molecules*

Chemoinfomatics!

- ***PubChem** database to handle data*

REACH

Registration, **E**valuation, **A**uthorisation and Restriction of **Ch**emical substances

European **Ch**emicals **A**gency (ECHA) in Helsinki

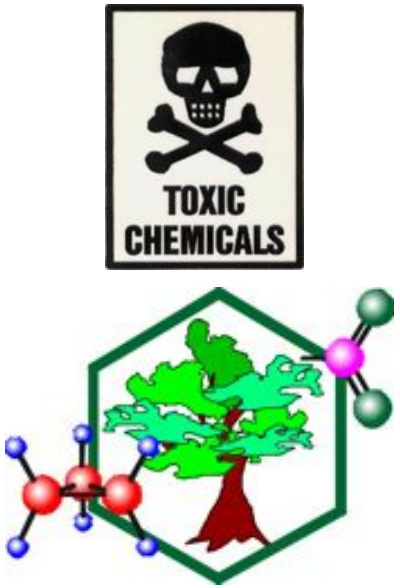
REACH and Environmental Chemoinformatics

> 140,000 chemicals to be registered … is a lot!

It is expensive to measure all of them ($200,000 per compound), a lot of animal testing => **€24 billions over next 10 years**

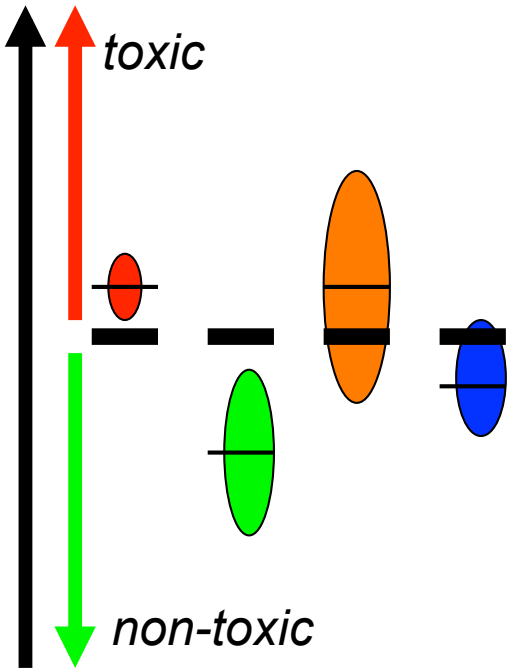
QSAR models can be used to prioritize compounds

- Compound is predicted to be toxic
  - Biological testing will be done to prove/ disprove the models
- Compound is predicted to be not toxic
  - tests can be avoided, saving money, animals
  - but ... only if we are confident in the predictions



Requirements of biological testing following QSAR model prediction

model prediction	prediction confidence	
	high	low
toxic, EC50 > LIMIT	**strong need**	**moderate need (depends on other properties)**
non-toxic, EC50 < LIMIT	**no need**	**low need (depends on other properties)**



Acceptance of decisions is more accurate if confidence intervals (prediction errors) are known and are taken into analysis: concept of applicability domain.

CAse studies on the development and application of in-silico techniques for environmental hazard and risk assessment

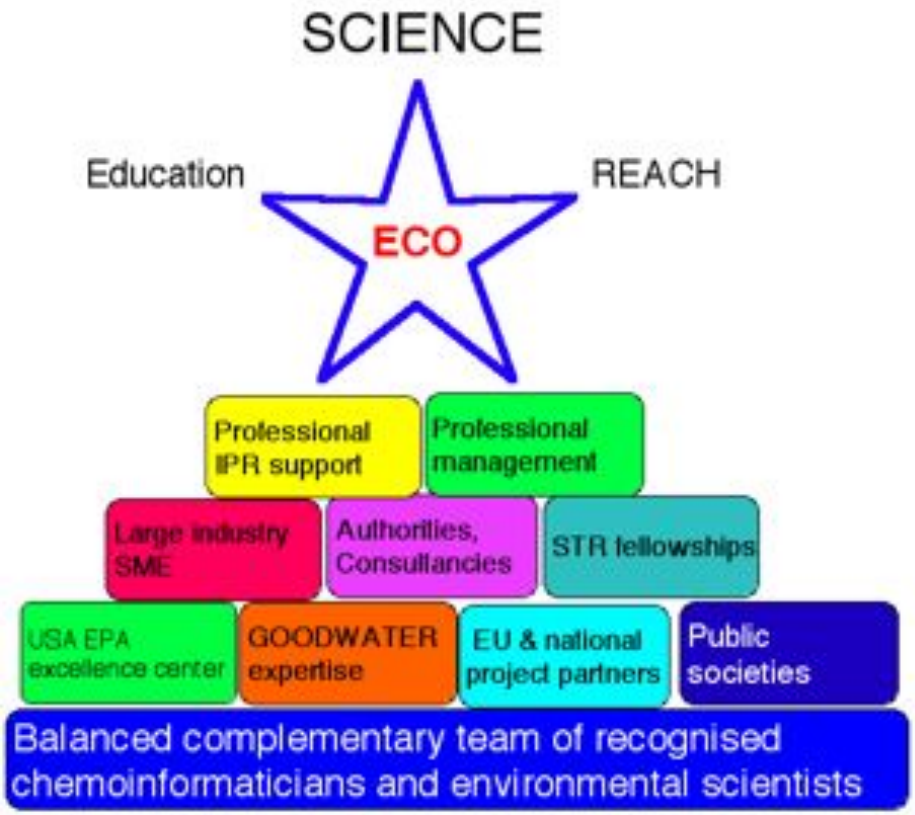
Environmental ChemOinformatics (ECO)
Marie Curie Initial Training Center

Main goal: training of new personnel with respect to REACH

Team: 7 partners from 5 EU countries

Offered training:

- 11 PhD students
- 1 postdoc
- 37 short-term fellowships (3-12 months)
- **in total:** ca 50 years of positions



Definitions of **Chemoinformatics**

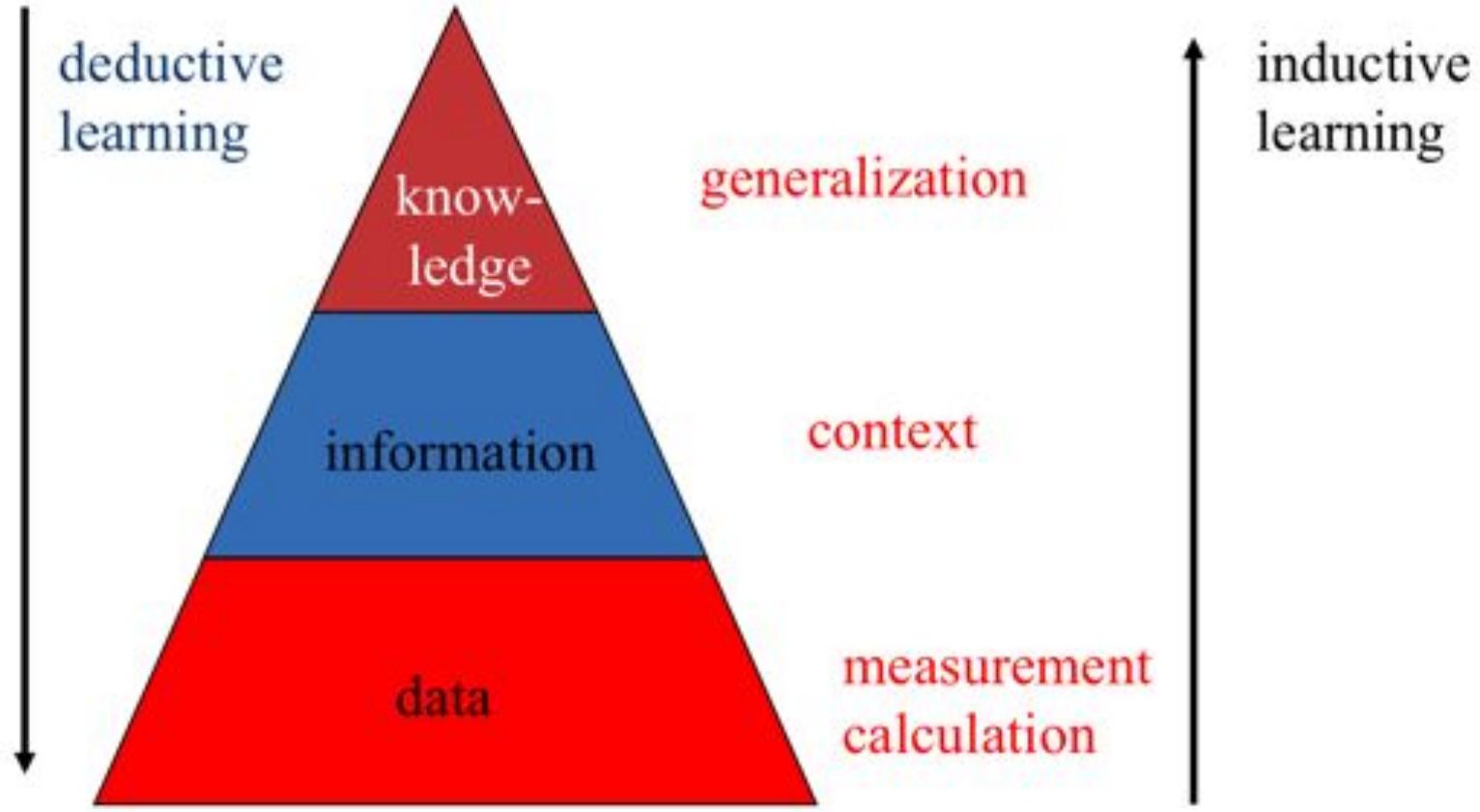
Chemoinformatics is a generic term that encompasses the design, creation, organization, management, retrieval, analysis, dissemination, visualization, and use of chemical information. *G. Paris, 1988*

Chemoinformatics is the mixing of those information resources to transform data in to information and information in to knowledge for the intended purpose of making better decisions faster in the area of drug lead identification and optimization. *F.K. Brown, 1998*

Chemoinformatics is the application of informatics methods to solve chemical problems. *J. Gasteiger, 2004*

Chemoinformatics is a field dealing with molecular objects (graphs, vectors) in multidimensional chemical space. *A. Varnek & A. Tropsha, 2007*.

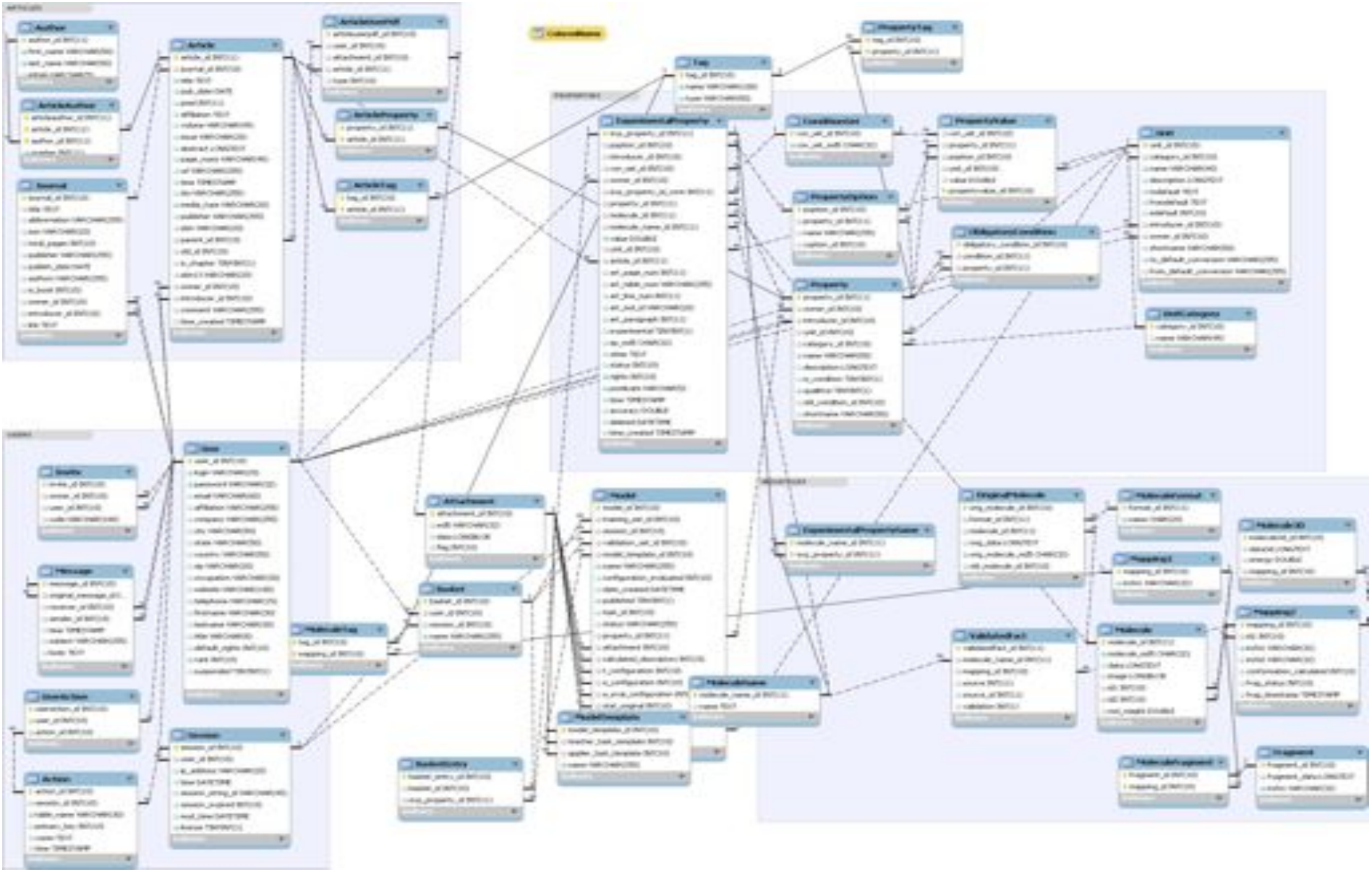
Chemoinfomatics: data to knowledge

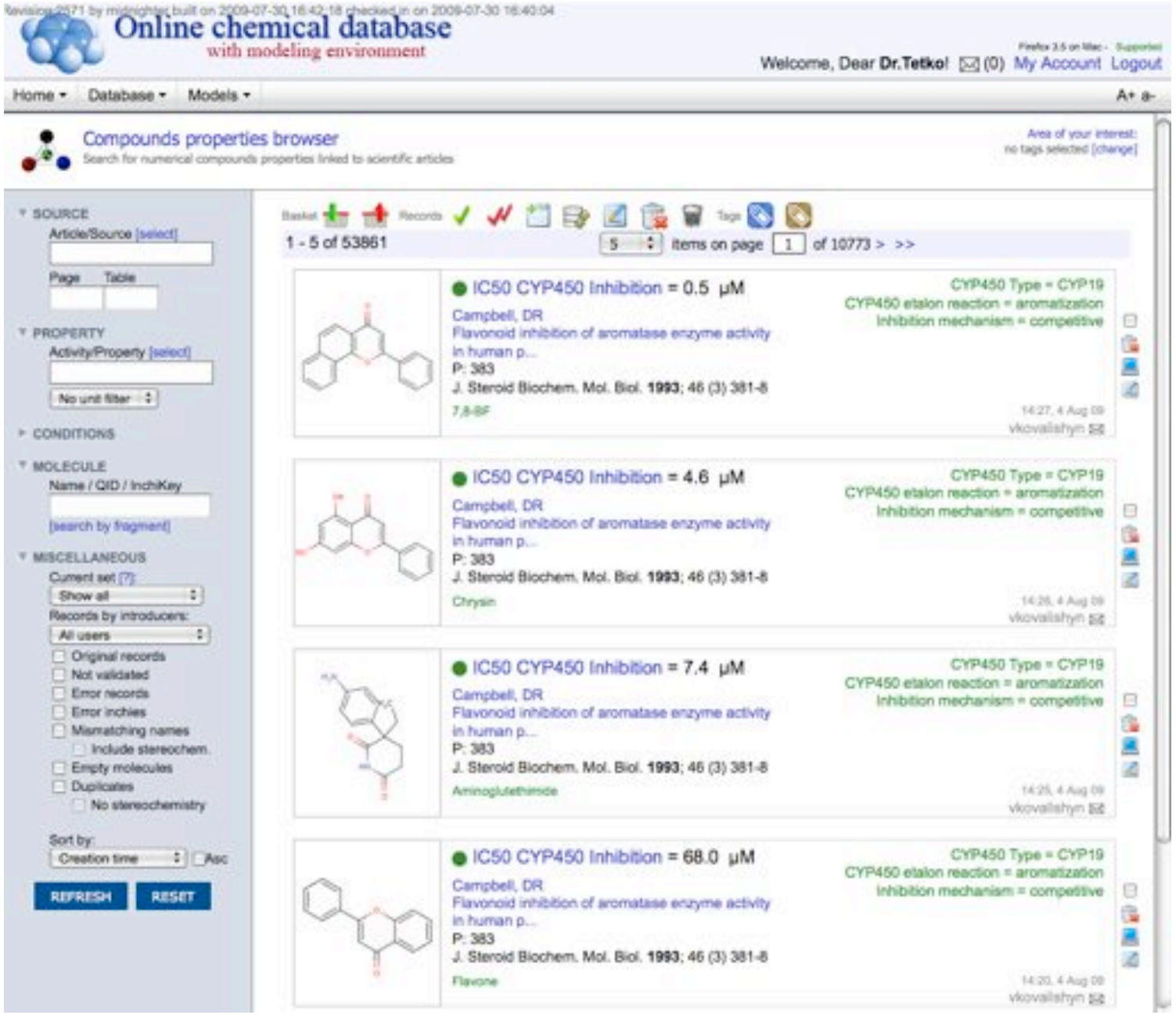


What is required to create a good model?

- Data -- the most essential part!
- Appropriate representation of molecules (choice of descriptors is crucial!)
- Good statistical methods

OCHEM - On-line CHEmical database and Modeling environment

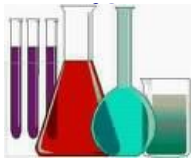



Online chemical database
with modeling environment
Welcome, Dear Dr.Tetko! (0) My Account Logout
Home Database Models
Compounds properties browser
Search for numerical compounds properties linked to scientific articles
SOURCE
Article/Source [select]
Page Table
PROPERTY
Activity/Property [select]
No unit filter
CONDITIONS
MOLECULE
Name / QID / InchiKey
[search by fragment]
MISCELLANEOUS
Current set [?]
Show all
Records by introducers:
All users
Original records
Not validated
Error records
Error inchies
Mismatching names
Include stereochem.
Empty molecules
Duplicates
No stereochemistry
Sort by:
Creation time
Asc
REFRESH
RESET
1 - 5 of 53861
items on page
of 10773 > >>
IC50 CYP450 Inhibition = 0.5 µM
Campbell, DR
Flavonoid inhibition of aromatase enzyme activity in human p...
P: 383
J. Steroid Biochem. Mol. Biol. 1993; 46 (3) 381-8
7,8-BF
CYP450 Type = CYP19
CYP450 etalon reaction = aromatization
Inhibition mechanism = competitive
IC50 CYP450 Inhibition = 4.6 µM
Chrysin
IC50 CYP450 Inhibition = 7.4 µM
Aminoglutethimide
IC50 CYP450 Inhibition = 68.0 µM
Flavone

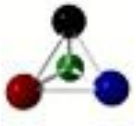
Database schema
Simplified overview

Property

logP = 0.5
Melting Point = 100



Filtering
Toxicology, Biology,
Partition coefficient.

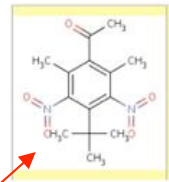


Condition

Temperature,
pH, species,
tissue, method

Tags
Toxicology, Biology,
Partition coefficient.

Data Point

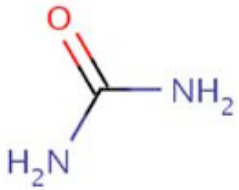


LD50 oral = 10.0 g/kg
Species = rat
Schmeiser, HH
Evaluation of health risks caused by musk ketone...
P: 295 T: 1 L: 1
Int J Hyg Environ Health **2001**; 203 (4) 293-299
Fragrances CADASTER molecules
81-14-1 ; FRA-043 ; Musk ketone
09:02, 23 Mar 09
maraluini

Introducer
Bill G., Sergey B.
Date of modification
Informationsystem

Structure

Benzene. Urea. ...

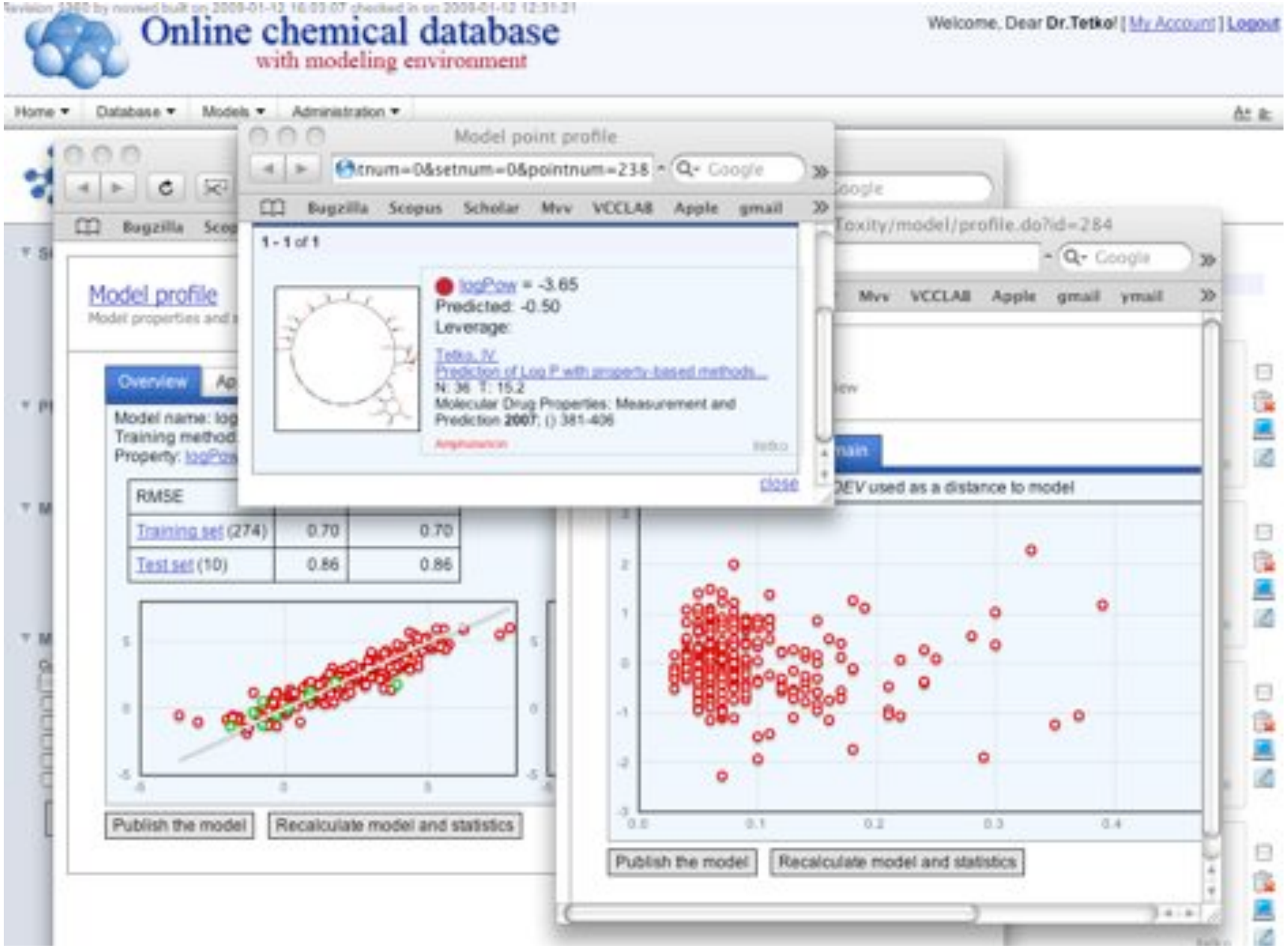


Manipulation
Editing
Organization
Working sets<

Article

Garberg, P
"In vitro models for ..."

Descriptors


Online chemical database
with modeling environment
Welcome, Dear Dr.Tetko! | My Account | Logout
Home
Database
Models
Administration
Model point profile
tnum=0&setnum=0&pointnum=238
Bugzilla
Scopus
Scholar
Mvv
VCCLAB
Apple
gmail
1 - 1 of 1
logPow = -3.65
Predicted: -0.50
Leverage:
Tetko, IV
Prediction of Log P with property-based methods...
N: 36 T: 15.2
Molecular Drug Properties: Measurement and
Prediction 2007; () 381-406
close
Model profile
Overview
RMSE
Training set (274)
0.70
0.70
Test set (10)
0.86
0.86
Publish the model
Recalculate model and statistics
Toxity/model/profile.do?id=284
ymail
Publish the model
Recalculate model and statistics

Offered course

Monday (10.08, 10:00-13:00)

- Introduction to Chemoinformatics
- From 2D to 4D descriptors (Prof. V.E. Kuzmin, OSU, Ukraine)
- Developing chemoinformatics models: One can't embrace unembraceable (Dr I.V. Tetko, HMGU, Germany)

Tuesday (11.08, 10:00-13:00)

- Role of 3D information in chemoinformatics: problem of chirality (Prof. V.E. Kuzmin)
- Introduction to predictive toxicology (Prof. W. Peijnenberg, 2 lectur

Wednesday (12.08, 10:00-13:00)

- Introduction to OCHEM: database (Mr S. Novotarskyi, HMGU)
- Practical session (Mr S. Novotarskyi & I. Sushko, HMGU)

Thursday (13.08, 10:00-13:00)

- Introduction to OCHEM: models (Mr I. Sushko, HMGU, Germany)
- Overview of machine learning methods available at OCHEM (Dr. I. Tetko)
- Practical session (Mr S. Novotarskyi & I. Sushko, HMGU, Germany)

HMGU Team

Acknowledgement

Prof. W. Peijnenburg
Prof. V. Kuz'min
Mr I. Sushko
Mr S. Novotarskyi

Team members
Mr A.K. Pandey
Mr R. Körner
Mr S. Brandmayer
Dr. M. Rupp

BMBF GO-Bio grant 0313883, CADASTER FP7 EU project,
Germany-Ukraine bilateral grant

Thank you for your attention!