HelmholtzZentrum münchen
German Research Center for Environmental Health

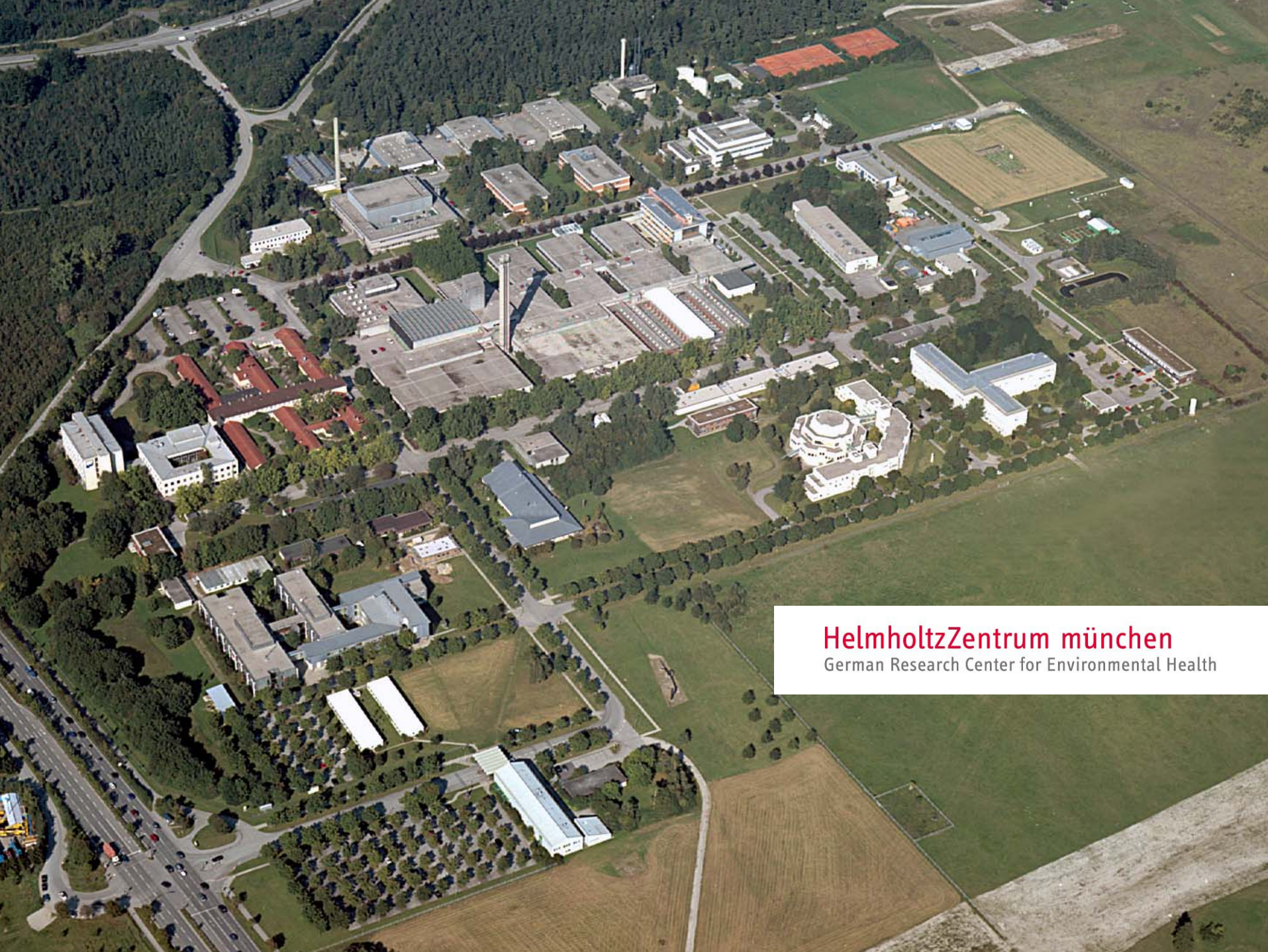
Modeling of physicochemical properties for perfluorinated compounds using a data integration approach

Wolfram Teetz

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

Idstein, Fresenius, 19. Juni 2010

HELMHOLTZ ASSOCIATION



HelmholtzZentrum münchen

German Research Center for Environmental Health

HelmholtzZentrum münchen
German Research Center for Environmental Health

REACH
The new EU chemicals legislation

- Efficiently use scarce and scattered information available

- Reduce animal testing - use in silico and in vitro information on related compounds.

HELMHOLTZ
ASSOCIATION

HelmholtzZentrum münchen
German Research Center for Environmental Health

CADASTER

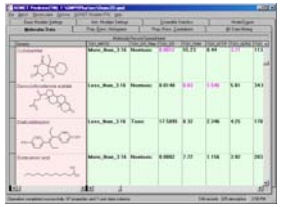
- Practical guidance to integrated risk assessment
- Full risk assessment for four compound classes (incl. PFC)
- Explicit uncertainty in data and in models
- Applicability domain
- Exemplify integration of information for risk assessments for large numbers of substances.
- How to use non-testing information for regulatory decision

SEVENTH FRAMEWORK PROGRAMME

FP7-funded 2009-01-01 end: 2012-12-31

HELMHOLTZ ASSOCIATION

ITS

Method	Costs	Time	Quality
Blind *in-vitro* testing of all compounds			
Blind *in-silico* testing of all compounds 			
in-silico testing plus intelligent testing of selected compounds +			

What compound to measure?

What property to measure?

What property to predict?

What property can be predicted from another one? -> Knowledge

The accuracy + AD has to be estimated in the end !

Properties of molecules

- Data are lost after publication of an article
- The original sources of data are difficult to track
- The conditions of experiments are frequently not provided
- The conversion between different units is error prone
- Current databases do not allow community correction of errors
- The tracking of changes (by users) is required

Models

- Most published models are never used
- Implementation can be as difficult as new model development
- Different implementations can produce different results

- “Wiki” approach to data handling
  users can add, modify and delete data

- Mandatory reference to an origin of information
  each record in a database should contain a reference to a source (article, book, proceeding or even personal communications), where the data were published

- Storing rich information
  we store measurement conditions to increase data quality

- Several tools to support decision making
  integration with other web-services (validation of molecule names against PubChem database, automatic fetching of article information from PubMed), duplicate records management

- Aimed at model building
  convenient to build training sets from data - filter by property, article and export data either to internal modelling tools or download as Excel file

Overview | Applicability domain

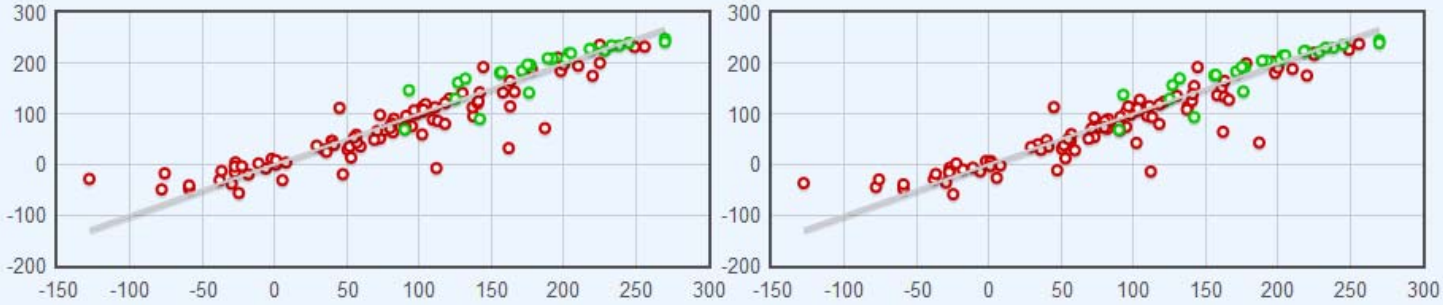
Model name: BarunMPBP.BP-PERFORCE-28apr10v1 [rename]
Public ID is 2627848
Predicted property: Boiling Point
Training method: ANN
measured in °C

[EState]
Correl. limit: 0.95
Supersab, 1000 iterations, 3 neurons
5-fold cross-validation

66 filtered descriptors
Supersab, 1000 iterations, 3 neurons

Calculated in 68 seconds

Data Set	Original			Recalculated		
	R^2	RMSE	MAE	R^2	RMSE	MAE
Training set: BarunMPBP.BP-28apr10v1-93 (93 records)	0.84	32.40	20.88	0.85	31.90	20.17
Test set: BarunMPBP.BP-PERFORCE-28apr10v1-25 (25 records)	0.77	24.54	20.23	0.80	22.07	18.07



Overview | Applicability domain

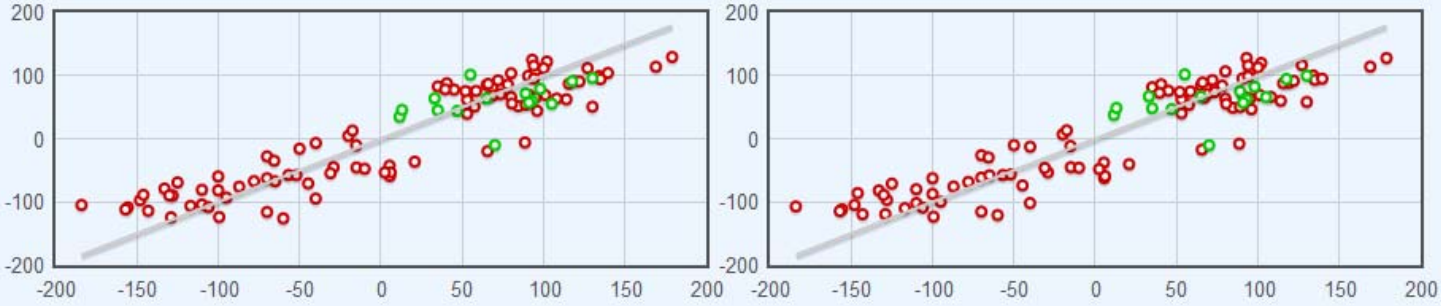
Model name: BarunMPBP.MP-PERFORCE-28apr10v1 [rename]
Public ID is 5666671
Predicted property: Melting Point
Training method: ANN
measured in °C

[EState]
Correl. limit: 0.95
Supersab, 1000 iterations, 3 neurons
5-fold cross-validation

87 filtered descriptors
Supersab, 1000 iterations, 3 neurons

Calculated in 35 seconds

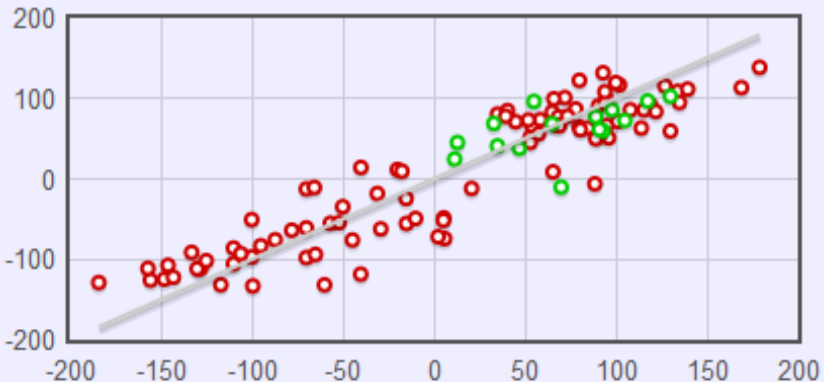
Data Set	Original			Recalculated		
	R_2	RMSE	MAE	R_2	RMSE	MAE
Training set: BarunMPBP.MP-28apr10v1-93 (93 records)	0.84	36.96	30.59	0.85	36.69	30.38
Test set: BarunMPBP.MP-PERFORCE-28apr10v1-15 (15 records)	0.21	35.09	29.43	0.22	34.12	28.34



Model name: MP BP Model - Perforce
Training method: ANN

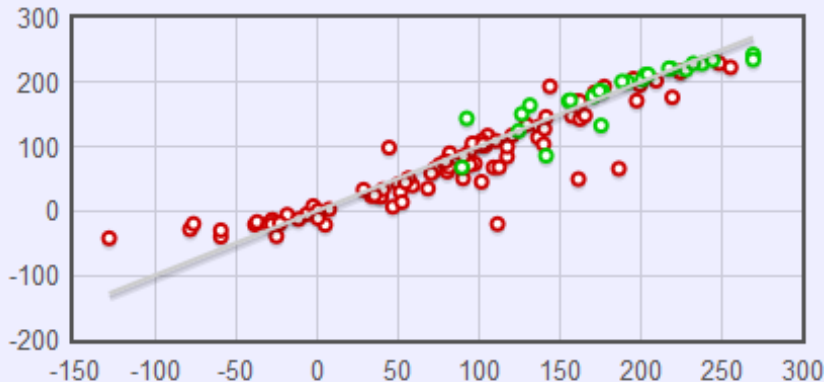
Property: Melting Point measured in °C (Details..)

Data Set	R_2	RMSE	MAE
Barun MPBP-MPBP Train(93)	0.85	35.65	29.20
PERFORCE MPBP(15)	0.32	31.91	25.89



Property: Boiling Point measured in °C (Details..)

Data Set	R_2	RMSE	MAE
Barun MPBP-MPBP Train(93)	0.85	31.76	20.52
PERFORCE MPBP(25)	0.78	22.78	17.02



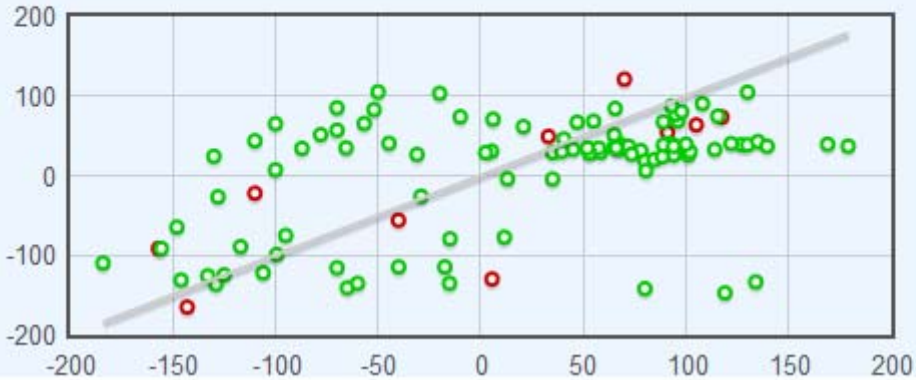
Save the model

Please enter your model's name: Melting Point 10->108

Overview | Applicability domain

Model name: Melting Point, set 3731, 72 [rename]
Public ID is 2344045
Predicted property: Melting Point
Training method: ANN
measured in °C

Data Set	R^2	RMSE	MAE
Training set: NewMP10 (10 records)	0.63	62.42	51.63
Test set: NewMP108 (98 records)	0.21	84.70	65.21



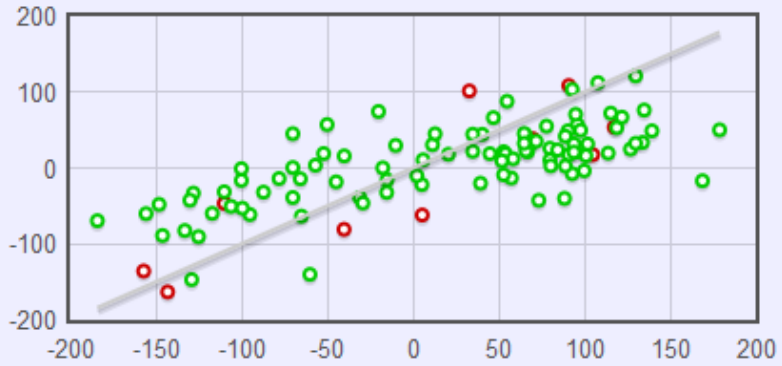
Save the model

Please enter your model's name: Melting Point from BP

Model name: Melting Point, set 3730, 67
Training method: ANN

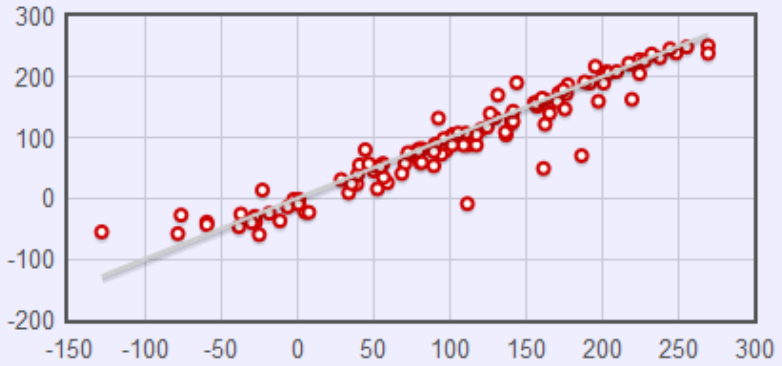
Property: Melting Point measured in °C (Details..)

Data Set	R_2	RMSE	MAE
NewBP118(10)	0.73	53.63	48.16
NewMP108(98)	0.49	65.53	55.27



Property: Boiling Point measured in °C (Details..)

Data Set	R_2	RMSE	MAE
NewBP118(118)	0.91	27.04	16.55
NewMP108(0)	NaN	NaN	NaN



SITE STRUCTURE

- ECO
  - Description
    - General information
    - Synergism
    - Methodology
    - Training
      - Content and quality
      - Schools
      - Structure
    - Dissemination
  - REACH
  - Contact
- Open Positions ✔
  - Experienced Researcher
  - Long Term
    - QSAR/QSPR modeling
    - Toxicity of nano-materials
  - Short Term

Open Positions

Marie Curie Initial Training Networks (ITN) are aimed at improving the career perspectives of European researchers who are in the first five years of their career by offering structured training in well defined scientific and technological areas as well as providing complementary skills and exposing the researchers to other sectors including private companies.

A number of open positions is available within two conceptual work packages listed below

The positions are available within two workpackages

WP 1 - Urgent problems of QSAR/QSPR modeling for REACH

WP 2 - Computational and in vitro screening methods for nano-materials

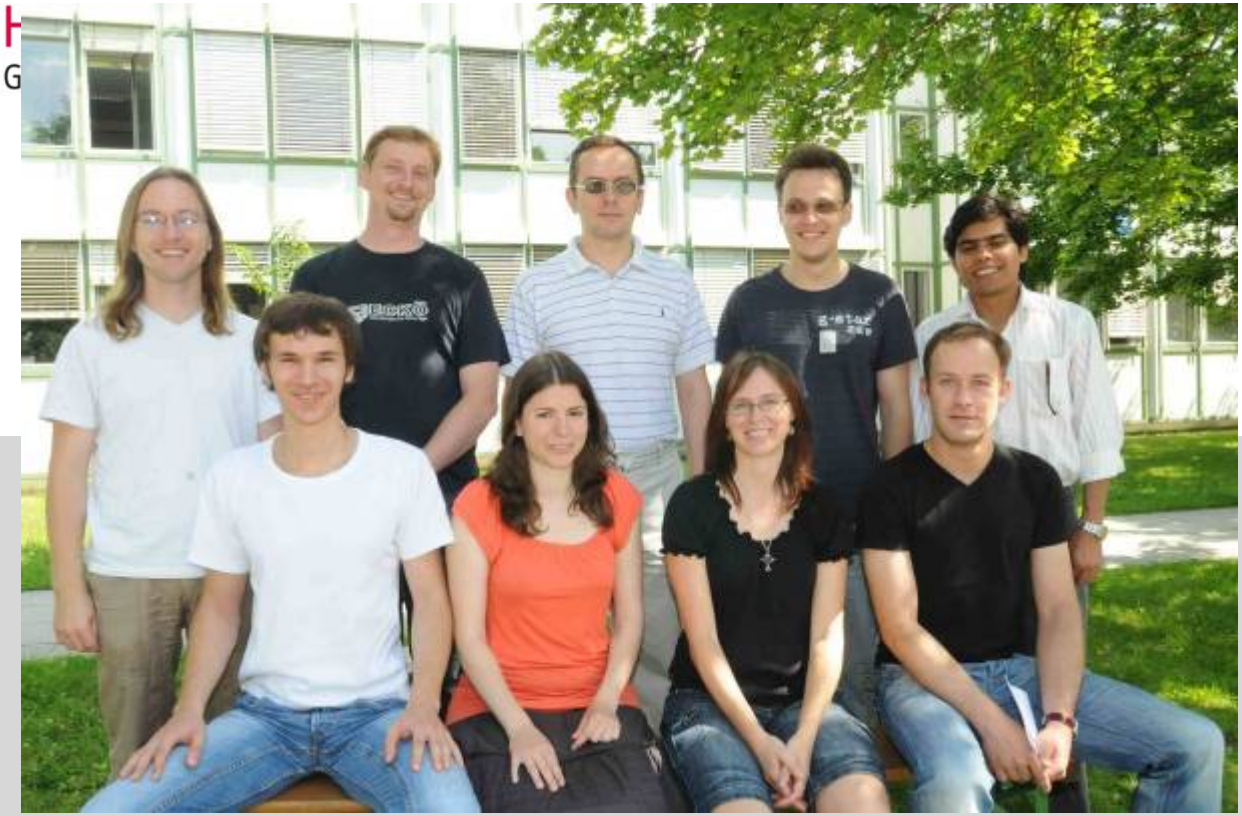
There are

one 1-year duration ER (Experienced Researchers, postdoc position) position,

still **2** of **11** Long Term Fellowships (LTF) (**ESR** positions, Early Stage Researcher, PhD positions), each one of 36 months duration, and

37 Short Term Fellowships (STF) (**ESR** positions), each one of 3-12 months duration.

Acknowledgements



Funding

GO-Bio BMBF http://qspr.eu

Germany-Ukraine grant UKR 08/006

FP7 Marie Curie ITN ECO http://eco-itn.eu

FP7 CADASTER http://www.cadaster.eu

Our group

Iurii Sushko

Sergii Novotarskyi

Anil K. Pandey

Robert Körner

Stefan Brandmaier

Matthias Rupp

Wolfram Teetz

Eva Schlosser

Collaborators:

Dr. G. Poda (Pfizer)

Dr. C. Ostermann (Nycomed)

Dr. C. Höfer (DMPKore)

Prof. A.S. Dimoglo (Turkey)

Prof. A. Tropsha (NC, USA)

Prof. T. Oprea (New Mexico, USA)

Prof. A. Varnek (Strasbourg, France)

Prof. R. Mannhold (Düsseldorf, Germany)

Prof. R. Todeschini (Milano, Italy)

+ many other colleagues