



Mid-Term meeting of MSC ITN EID project BIGCHEM
grant agreement no. 676434

Coordinator's report

Modena, 23 October 2017

Coordinator: **HelmholtzZentrum münchen**
German Research Center for Environmental Health

Content

1. Scientific overview
2. Training
3. Networking
4. Management



1. Scientific overview

1. Reasons for this project

2. Research objectives of the network

3. Scientific highlights so far

2. Training

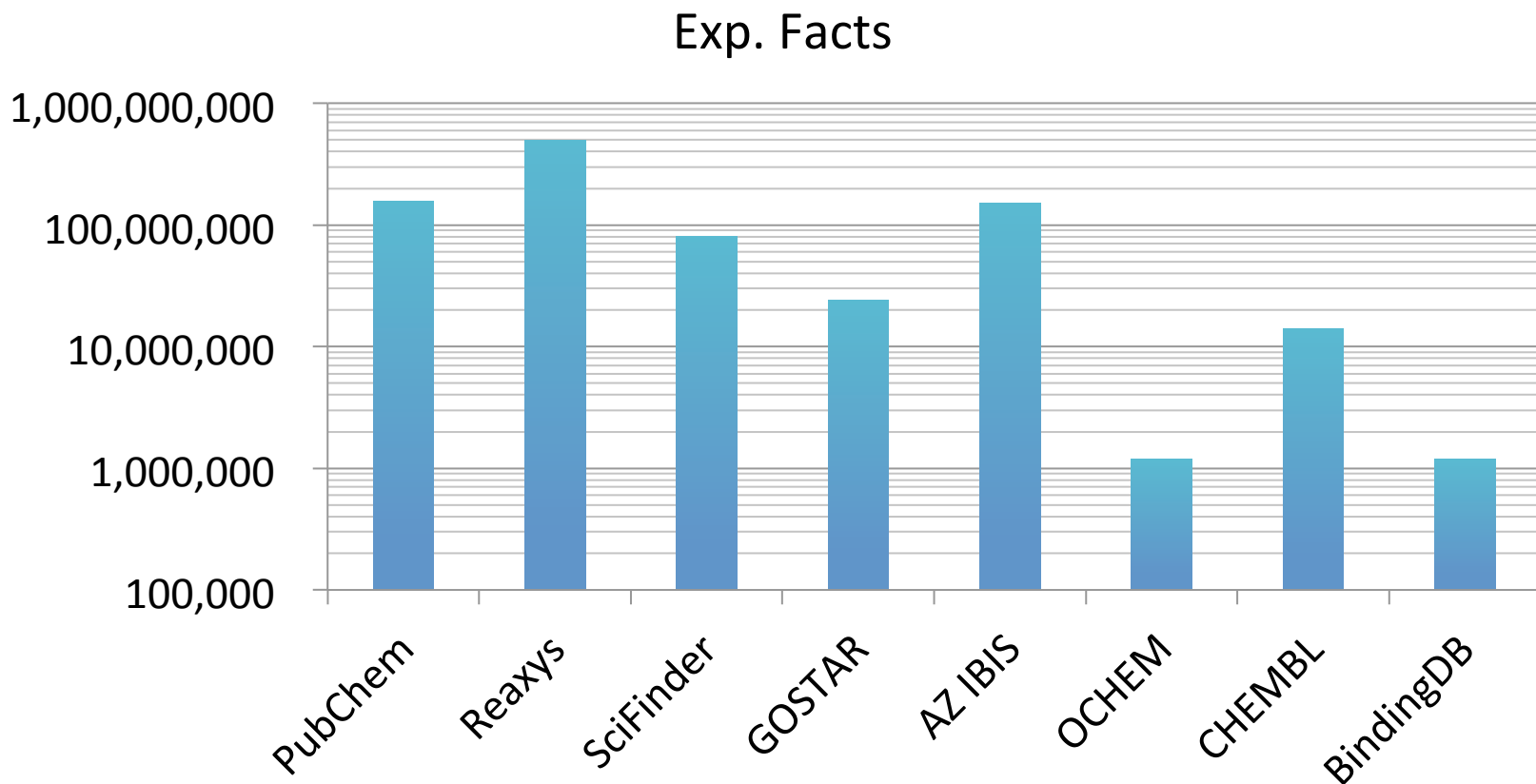
3. Networking

4. Management



Reasons for this research

The increasing **volume of biomedical data in chemistry and life sciences** requires development of **new methods and approaches for their analysis**



Big Data Sources - Large Chemical Database

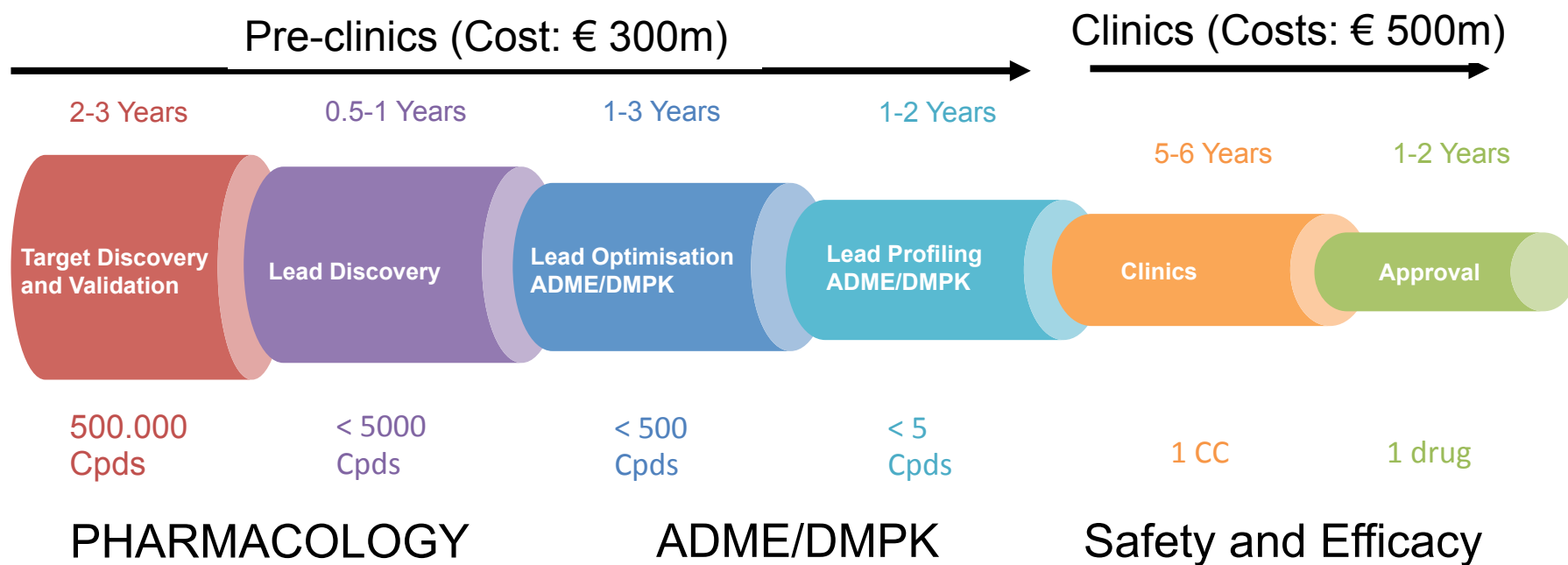


Reasons for this research

- Increasing widespread of new screening technologies
→ challenges of Big Data in intersectorial areas
- Increasing size of the market for Big Data
- Leadership in this area is absolutely critical for EU
- Contributes to further growth and prosperity of our society
- Demand for chemoinformatics specialists, but relevant training programmes are limited and fragmented



Big Data Sources - Process of Drug Discovery



Profiling and screening in the virtual space can help to identify most promising candidates

Slide courtesy of Dr. C. Höfer, Vitilis



BIGCHEM objectives

- Integration of research and teaching activities
- Bring together the premier European academic and industrial expertise in chemoinformatics
- Provide the most comprehensive curriculum in chemoinformatics/Big Data analysis available to date.



Scientific highlights: working packages

- Visualizing and data mining

Uni Bonn, Uni
Strasbourg,
BI (1-3)

- Promiscuity analysis

HMGU, Uni
Bonn, LDC,
AZ (4-5)

- Accessing new chemical space based on predictive models

ETHZ, Uni
Modena, Uni
Bern, AZ, Uni
Bonn, BI (6-9)

- Secure sharing of information

HMGU, AZ
(CWI) (10)

Bigchem fellows



Fellows individual projects



WP1: Visualizing and data mining (1-3)



ESR1: Machine learning methodologies for utilizing Big Compound Data

Raquel Rodríguez Pérez

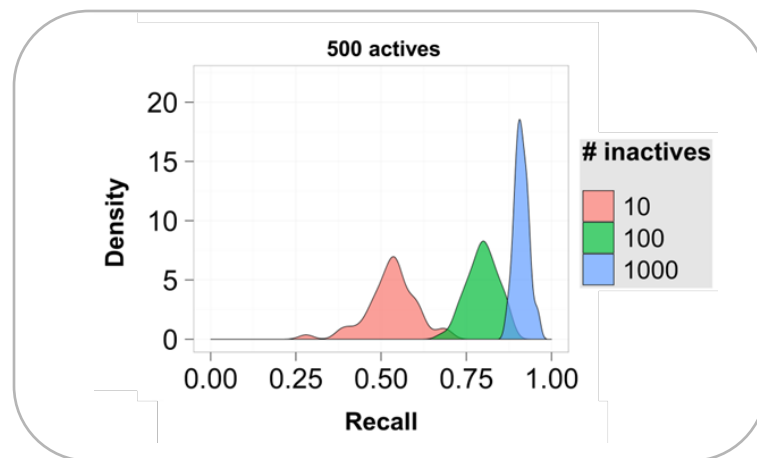


ESR1 Major achievements

- Application of **support vector machines** classification (SVM) and regression (SVR)

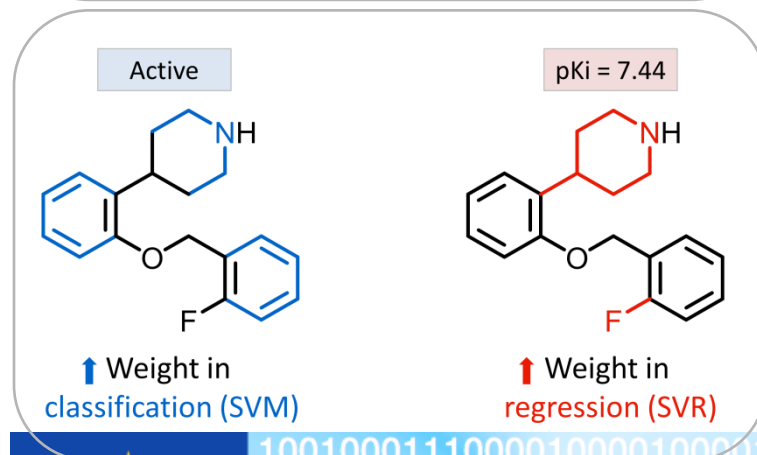
- Project 1:** Influence of **training set composition and size** on SVM activity predictions

Rodríguez-Pérez *et al.* *J. Chem. Inf. Model.* **2017**, *57*, 710-716.



- Project 2:** Prioritized **structural features** for compound activity and potency predictions

Rodríguez-Pérez *et al.* *ACS Omega.* **2017**, *2*, 6371-6379.



ESR2: Computational compound screening and profiling by large-scale mining of pharmaceutical data

Until September 30, 2017:

Benedict Mutimba

New hiring under discussion



Boehringer
Ingelheim



universität**bonn**

Rheinische
Friedrich-Wilhelms-
Universität Bonn

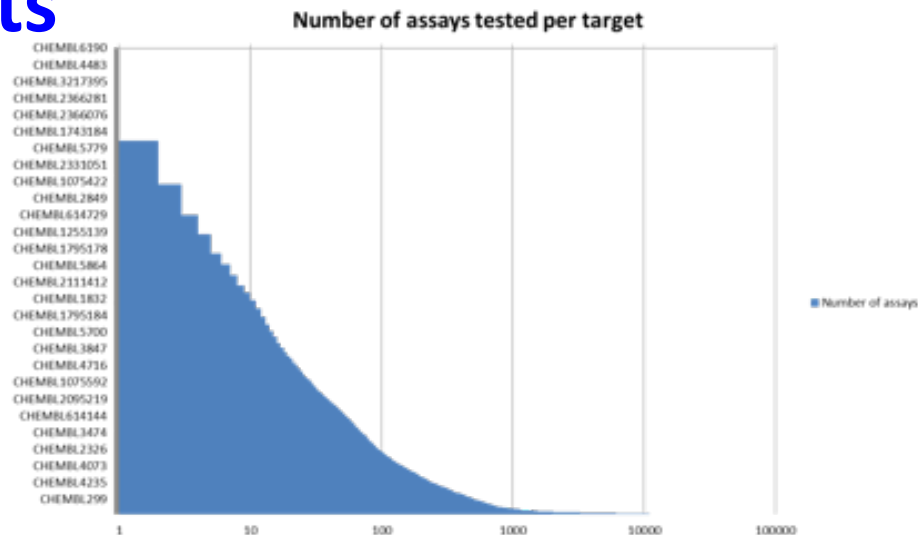


10010001110000100001000011
CC(=O)CC1=CC=CC=C1C(O)=O
11 10 100010 0100 00
CC(=O)N1=CC=CC=C1C(O)=O
BigChem

ESR2 Major Achievements

Scientific project:

Explore strategies to create models from different & heterogenous data sources



Approach:

- Compilation of data sets from ChEMBL with multiple assays per target, each assay with at least 10 compounds. 5 targets selected
- Model building procedure: RF; RDKit descriptors.
- Model validation procedure defined, including „leave-one-assay-out“ scenarios and addition of data from completely unrelated targets

Preliminary Results:

- Suitable models can be obtained from training sets based on heterogenous assays
- Leave-one-assay-out models' performances comparable to those of models using all assays.

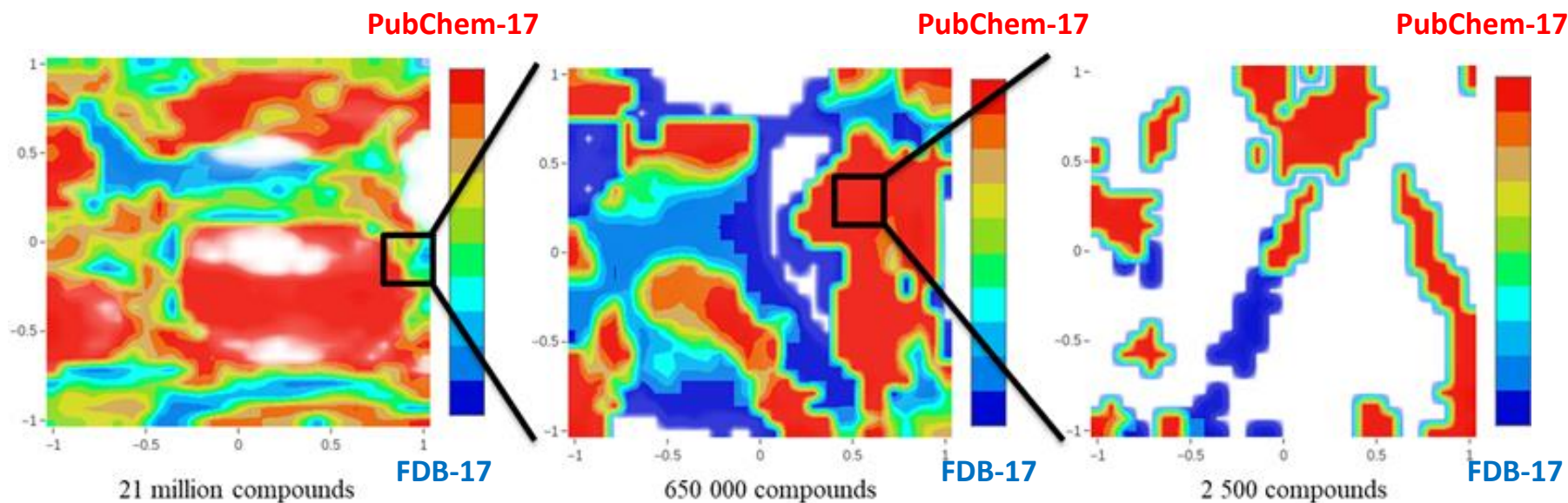


ESR3: Big data visualisation and modelling using Generative Topographic Mapping (GTM) approach

Arkadii Lin



ESR3: Big Data analysis using Generative Topographic Method



Hierarchical GTM “zooming” of the chemical space occupied by FDB-17 (blue) vs PubChem-17 (red, on the fuzzy classification maps). For each handpicked zone on a map, a local GTM model with identical parameters is refitted to the local residents only

WP2: Promiscuity analysis (4-5)



ESR4: Development of frequent hitters filters for HTS screening

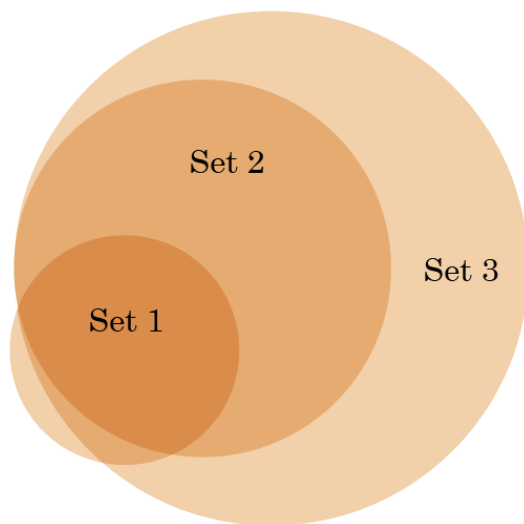
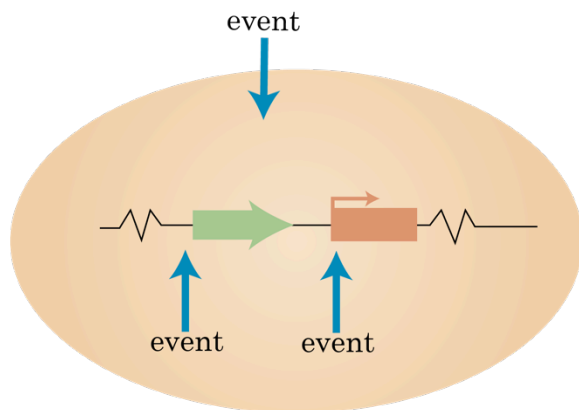
Dipan Ghosh



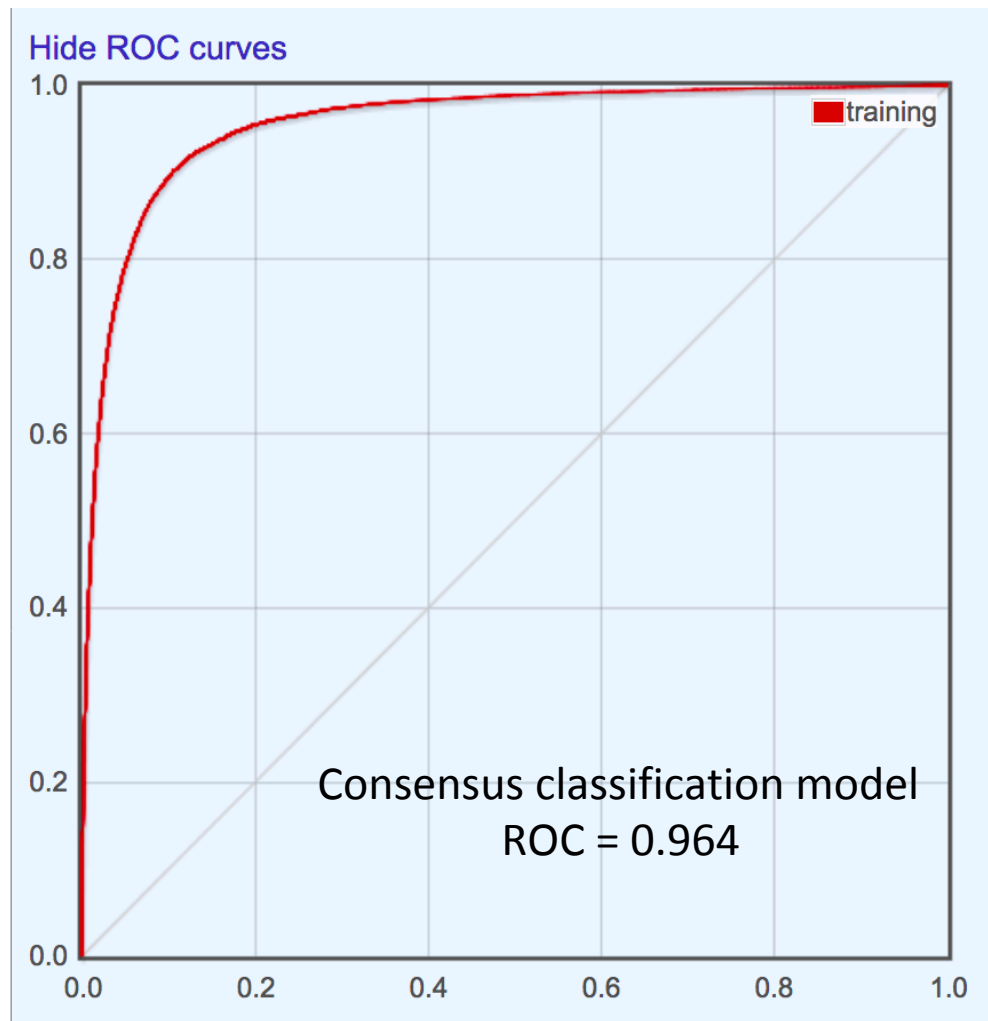
HelmholtzZentrum münchen
German Research Center for Environmental Health



ESR4: Analysis of inhibitors of luciferase



>360,000 compounds



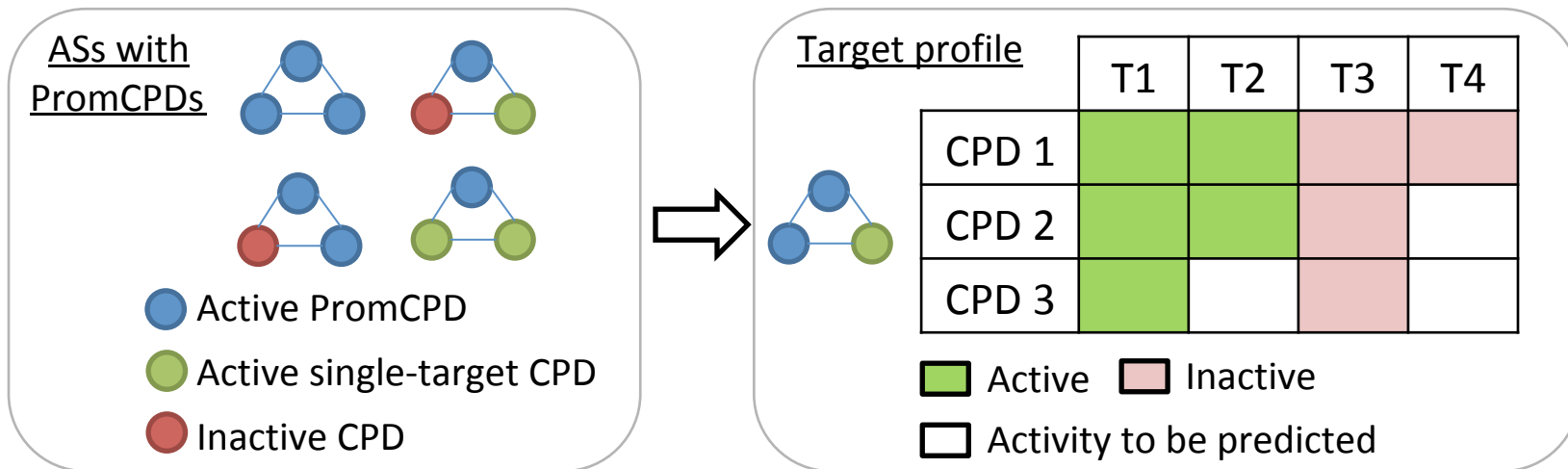
ESR5: Analysis of compound promiscuity and selectivity patterns

Laurinne David



ESR5: Work progress

- **Project 1:** Systematic evaluation of analog series (ASs) containing promiscuous compounds (PromCPDs)
- **Objective:** Assess the potential of analog series for target prediction and derive new target hypotheses for analogs



WP3: Accessing new chemical space (6-9)



ESR6: Integrating public data into the exploration of uncharted chemical space

Josep Arús-Pous



u^b

AstraZeneca 

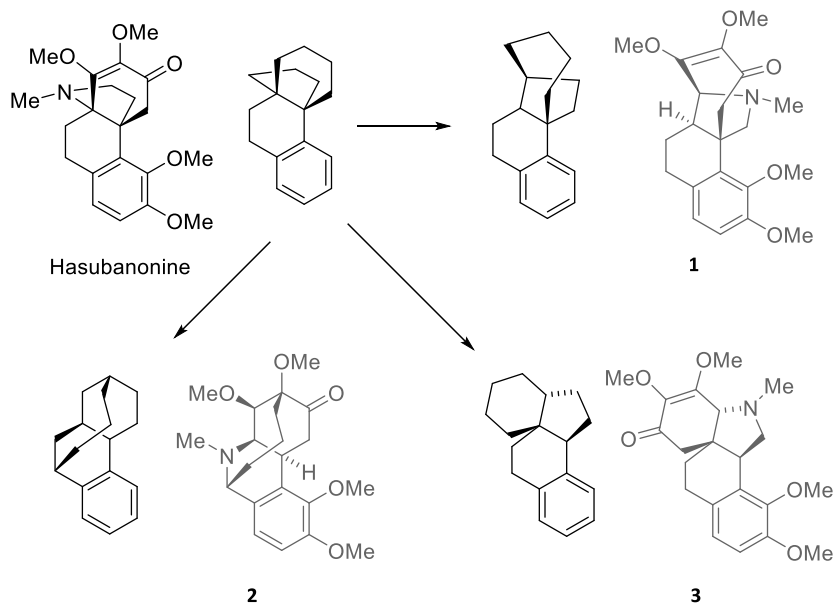
^b
UNIVERSITÄT
BERN



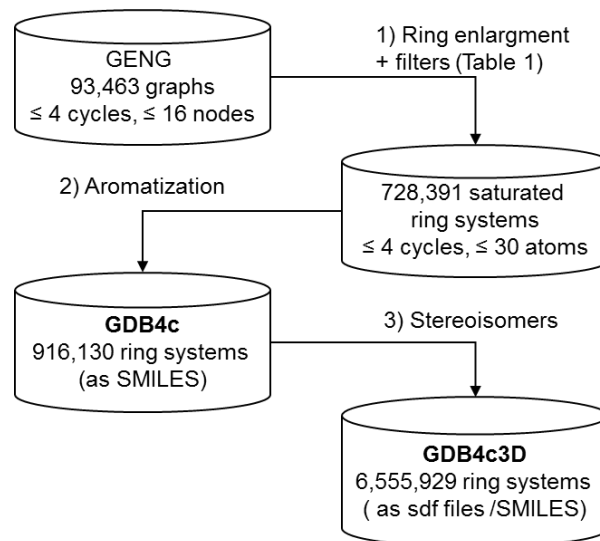
GDB4c

- **Ring systems (RS)** are cyclic cores without heteroatoms of molecules.
- We have enumerated 916,130 2D RS with up to 4 rings and the corresponding 6,555,929 3D stereoisomers.
- The majority of RS (98.6 %) are novel (not found in known molecules) and represent natural product like, chiral, 3D-shaped macrocycles

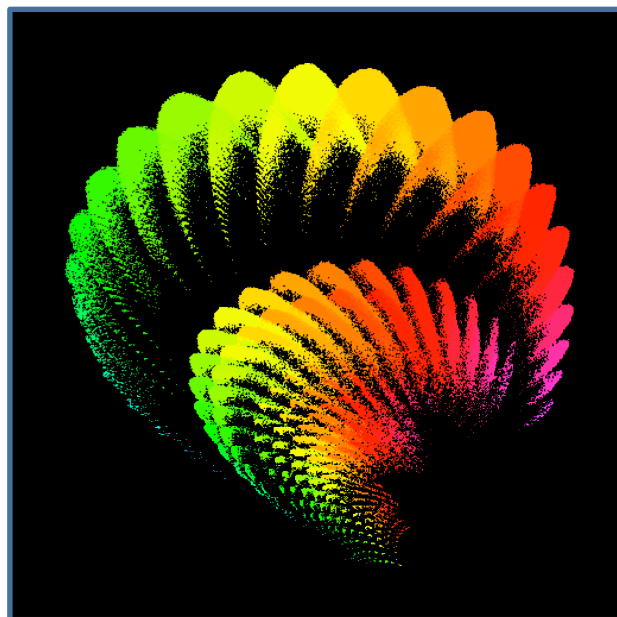
2. Natural product redesign with GDB4c



1. Database enumeration



3. Similarity map of GDB4c colored by HAC



Visini, R.; Arus-Pous, J.; Awale, M.; Reymond, J.-L. *J. Chem. Inf. Model.* **2017**, acs.jcim.7b00457.

ESR7: Exploration of uncharted regions of the chemical space by reaction-driven *de novo* design

Xuejin Zhan



ETH zürich



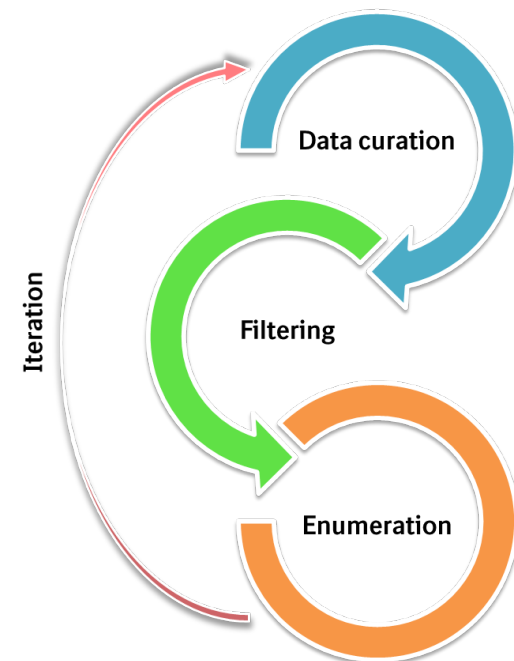
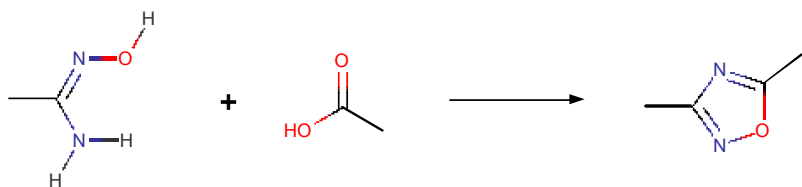
Boehringer
Ingelheim



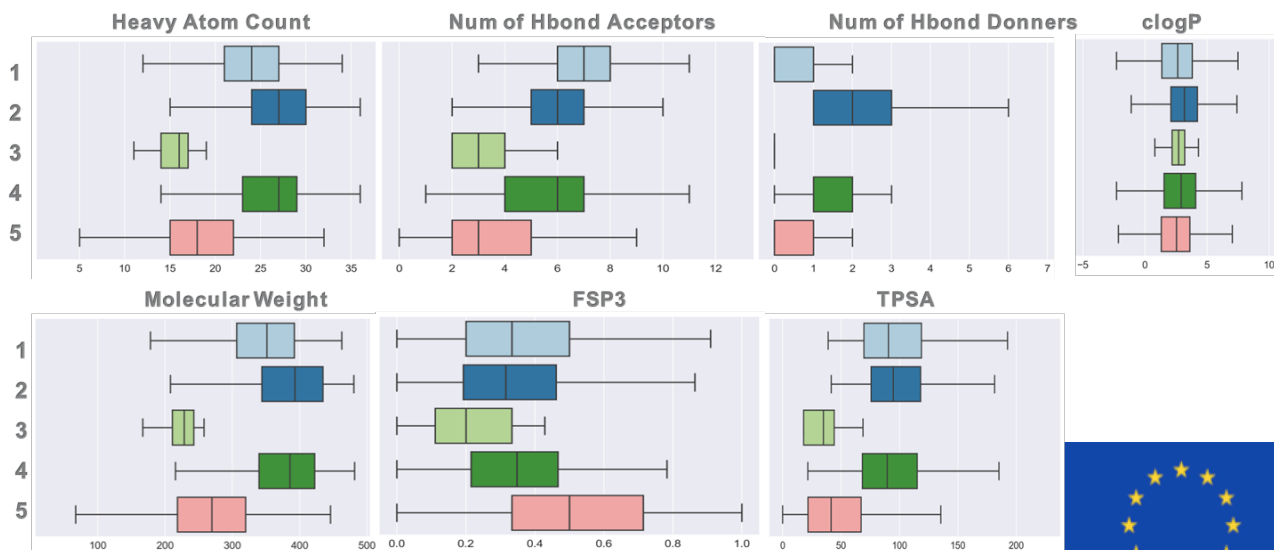
Major Achievements: Build an automatically User-friendly Chemical Enumeration Framework

Reaction-driven enumeration software established:

- Automatically data curation and enumeration
- Automatically calculation of molecular properties
- Automatically generated graphic reports after enumeration



Property distribution of five selected reactions:



ESR8: Accessing new chemical space for lead optimization based on QSAR models

Thomas Blaschke



AstraZeneca 

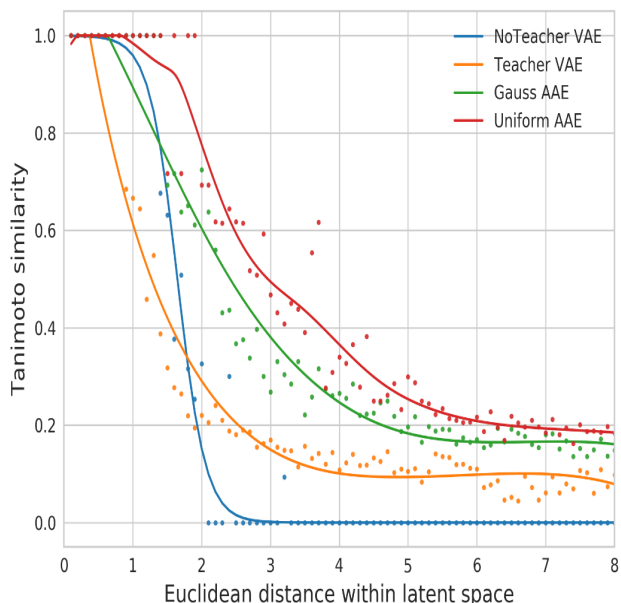
 universität**bonn**

Rheinische
Friedrich-Wilhelms-
Universität Bonn

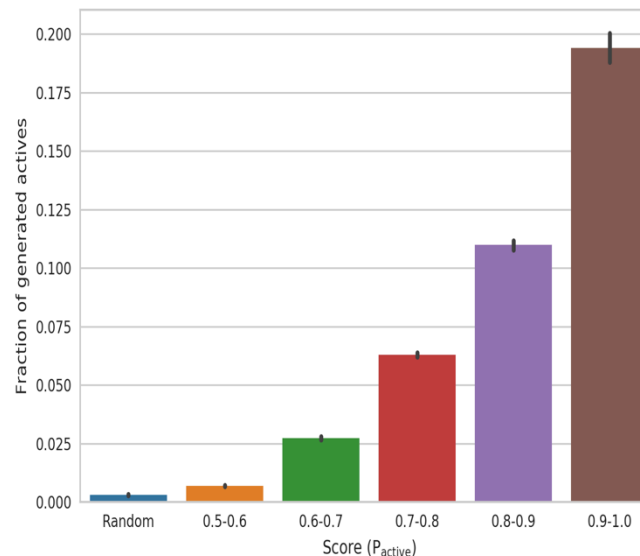


Generative autoencoder in *de novo* molecular design

- Introduced adversarial autoencoder (AAE) to embed molecules into a numeric representation (latent space) and hence generate novel structures.
- The chemical similarity principle is preserved in the latent space generated by AAE.
- Combining with Bayesian optimization method, AAE generative model has been applied on QSAR model (DRD2 model) guided structure generation (inverse QSAR).



Similarity vs distance in latent space



Probability of finding DRD2 active compounds increases at searched solutions

(submitted) publication: "Application of generative autoencoder in *de novo* molecular design", T. Blaschke, et al *Mol. Inf.*



ESR9: Integrated ligand- and structure based approaches for drug design and discovery using big data

Until September 1, 2017:

Eric March Vila

New hiring under discussion



UNIVERSITÀ DEGLI STUDI
DI MODENA E REGGIO EMILIA

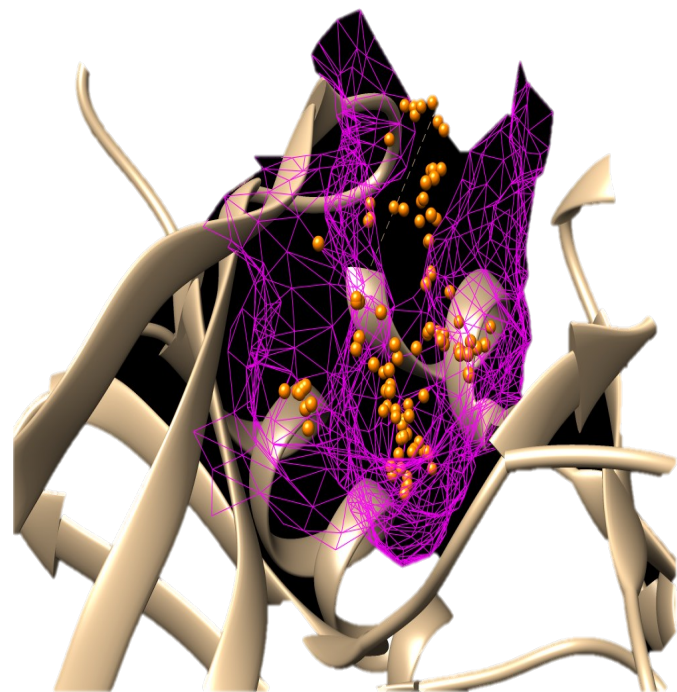
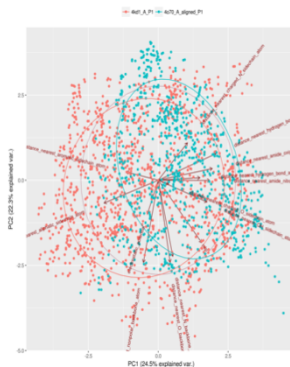
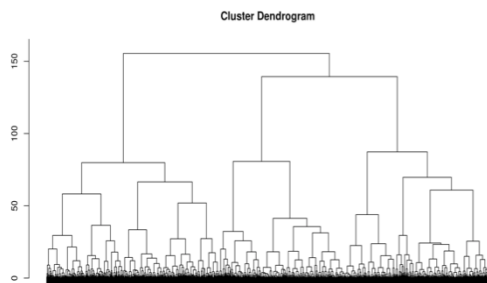
AstraZeneca 



Computational approaches for predicting compound polypharmacology

Goal: develop a computational tool to detect, describe and compare protein target binding sites. The tool will ultimately be tested using chemogenomics data, and then applied to predict compound polypharmacology.

- Binding pocket prediction using *fpocket*
- Calculation of protein pocket descriptors (geometrical, electrostatic, hydrophobic)
- Principal component analysis (PCA) of the resulting surface descriptors
- Cluster analysis and definition of common surface patches



Cdk2 structure (4lyn.pdb) showing alpha spheres within the pocket surface mesh

E. March-Vila, L. Pinzi, N. Sturm, A. Tinivella, O. Engkvist, H. Chen, G. Rastelli. *Front. Pharmacol.* **2017**, Volume 8, 298.

WP5: Secure sharing of information (10)



ESR10: Secure sharing of information

Michael Withnall



HelmholtzZentrum münchen

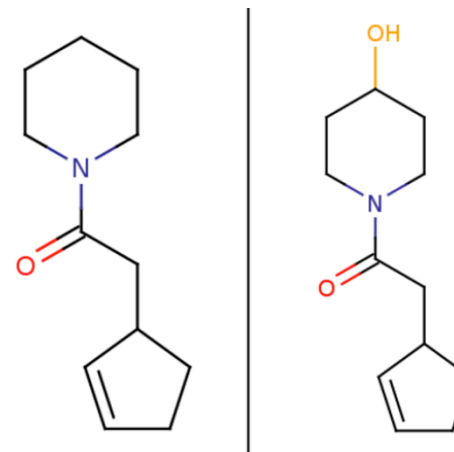
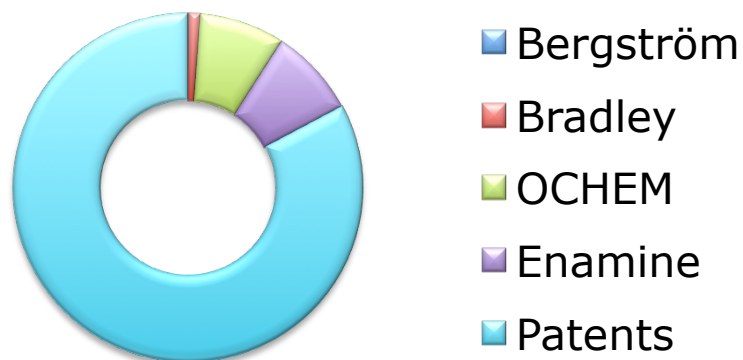
German Research Center for Environmental Health

AstraZeneca 

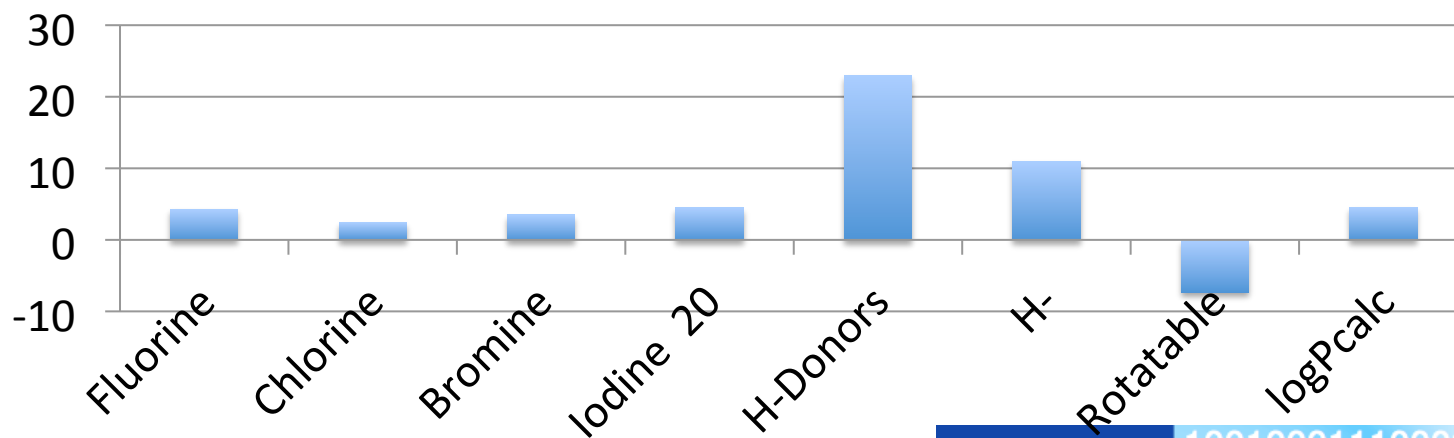


Matched Molecular Pair Analysis on Large Melting Point (MP) Dataset

MP Data (275k)



➤ 917,831 unique pairs



1. Scientific overview

2. Training

1. The training programme and the career development achievements

2. Secondments

3. Complementary skills

4. Training events open to external participants.

3. Networking

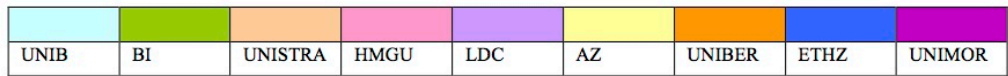
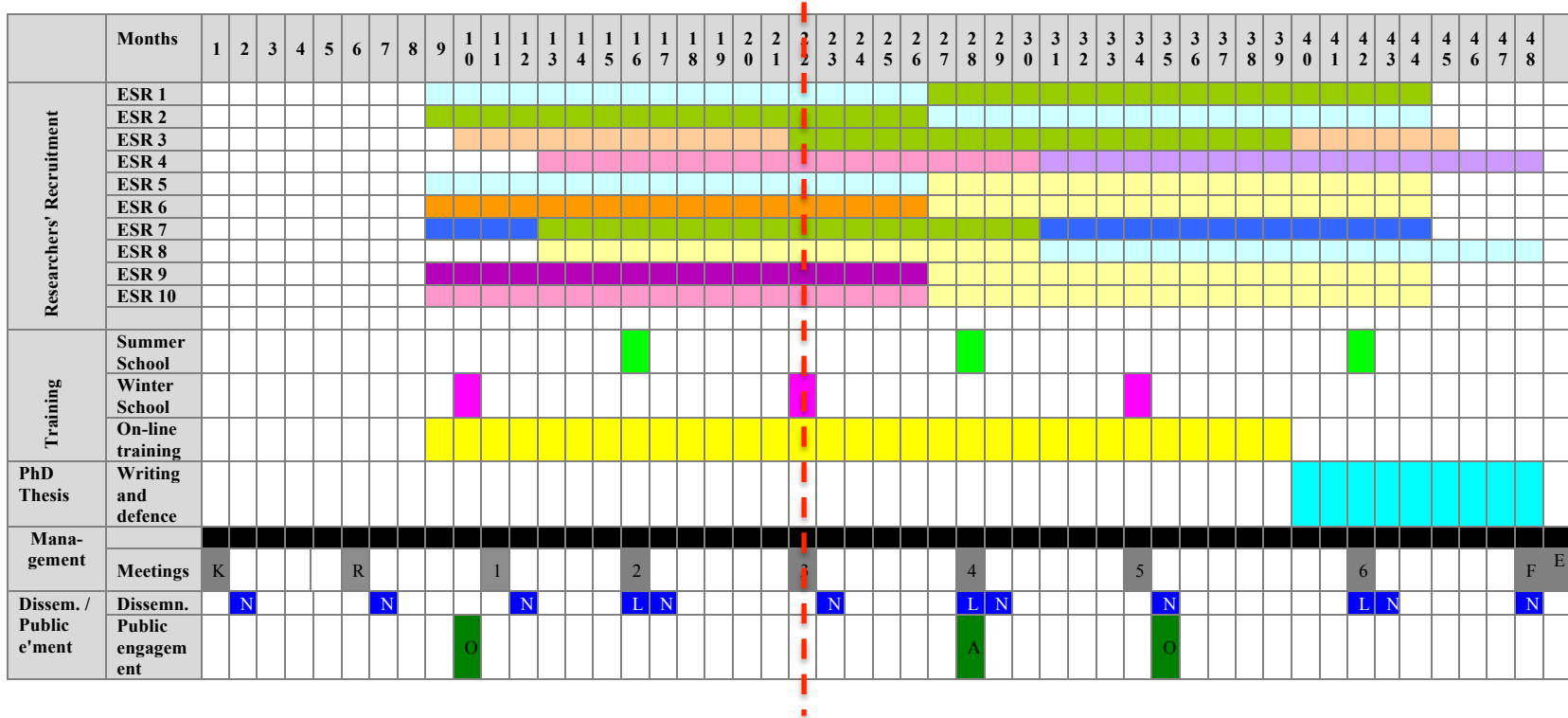
4. Management



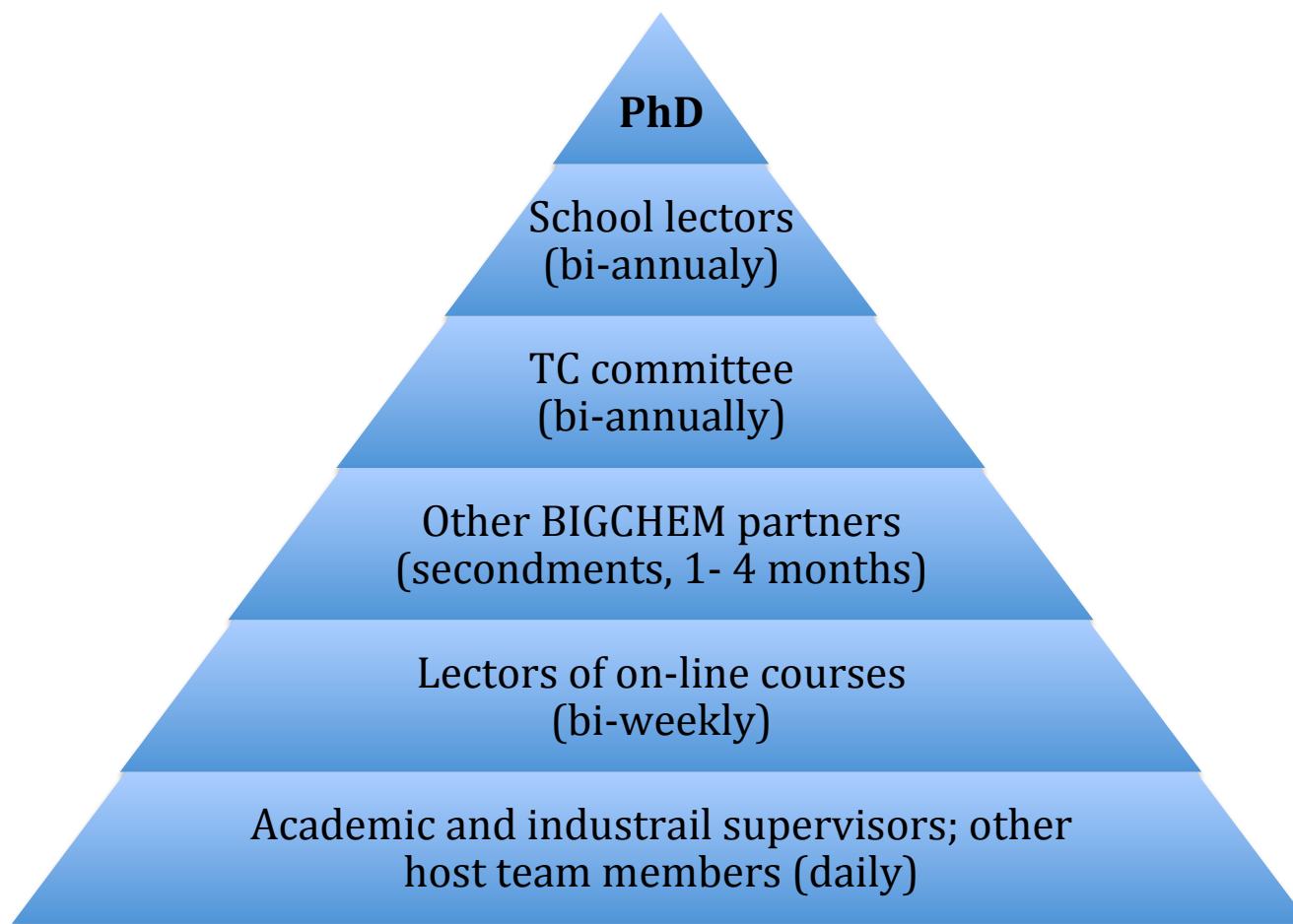
Training activities



Training Schedule



Training organisation



Training in BIGCHEM Schools

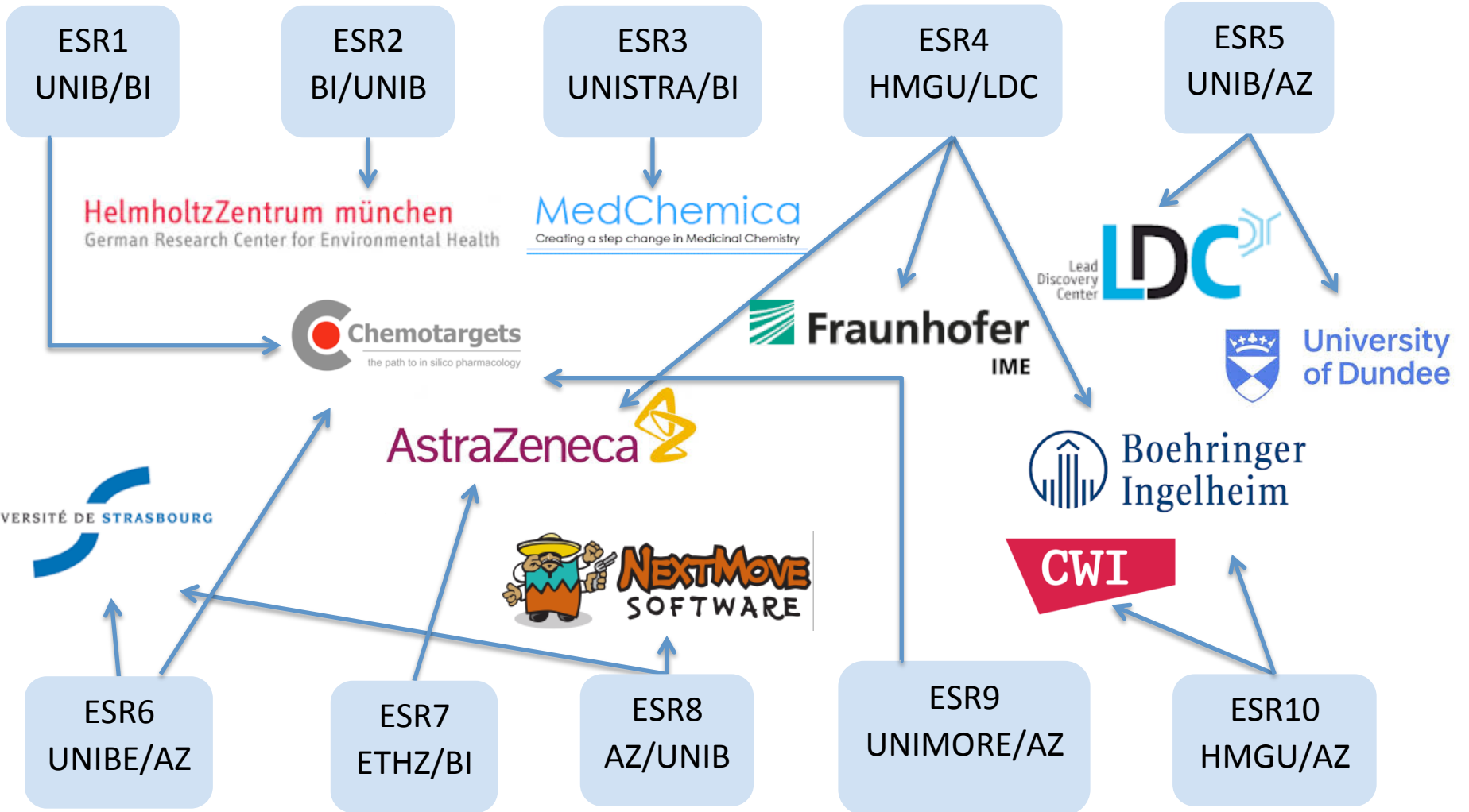
	School	Lead Institution	Location	Date
WS1	Introduction to chemoinformatics	HMGU/BI	Munich	October 2016
SS1	Chemical Data Resources	UNIBE/ETHZ	Barcelona	April 2017
WS2	Computer-Aided Drug Discovery	UNIMORE	Modena	October 2017
SS2	Chemical Space and ADMETox profiling – with Strasbourg Summer School on Chemoinformatics	UNISTRA	Strasbourg	June 2018
WS3	Virtual and HTS screening	LDC/BI	N.N.	N.N.
SS3	Final Closing School	HMGU/AZ	N.N.	N.N.



Training by on-line courses

	Title	Speaker	Date
1	Small molecule drug discovery	Jan Kriegl	29.9.16
2	On-line chemoinformatics tools	Igor Tetko	13.10.16
3	Chemoinformatics as a theoretical chemistry discipline	Alexandre Varnek	26.10.16
4	FreeWilson analysis and R-group QSAR model	Hongming Chen	2.11.16
5	Compiling, curating and enriching biochemical data collections	Bernd Beck	16.11.16
6	Similarity search	Uwe Koch	7.12.2016
7	Chemical Space Networks and SAR visualization	Martin Vogt	11.1.17
8	Phenotypic screening	Ola Engkvist	1.2.2017
9	De novo design	Gisbert Schneider	22.2.17
10	Compound collection informatics	Thierry Kogej	8.3.17
11	Introduction to the Reference Interaction Site Model (RISM)	Ekaterina Ratkova	2.3.17
12	Structure-Activity Modelling: QSAR	Uwe Koch	3.5.17
13	Machine learning for chemoinformatics	Francesca Grisoni	17.5.17
14	Foundations of virtual screening: SAR landscapes	Jürgen Bajorath	7.6.17
15	Structure-based virtual screening	Giulio Rastelli	21.6.17
16	Computational Approaches for Target Prediction	Mahendra Awale	5.7.17
17	Molecular de novo design through deep reinforcement learning	Marcus Olivecrona	6.9.17
18	In silico ADME	Susanne Winiwarter	21.9.17
19	Success stories of structure-based drug discovery	Ana Messias	11.10.17
20	ComFA and MFTA methods in drug discovery	Eugene Radchenko	08.11.17

Secondments



Professional/Complementary skills

- Training in BIGCHEM Schools
- Training at host organizations:
 - Scientific writing
 - Language courses in the respective language of the host country



Training events open to external participants

- SS1: morning lectures open to the public, > 30 external participants
- WS2: morning lectures open to the public, external participants (38 registered + 26 Uni Modena students)
- > half of on-line courses and Schools lectures are freely available at the BIGCHEM website



1. Scientific overview
2. Training
- 3. Networking**
 1. How the Network functions and how the beneficiaries cooperate in practice
 2. Interaction with private sector
 3. Dissemination and outreach activities
4. Management



Cooperation within the network

- Join supervision of fellows
- Join articles
- Contribution to the fellows training:
 - On-line courses
 - Training during Schools
- Secondments
- Rotational composition of Management Team



Interaction with private sector

- Training and supervision of fellows (in daily work, in Thesis Committees)
- Join development of new computational methods
- Translation of academic ideas to practical exploitation in industry



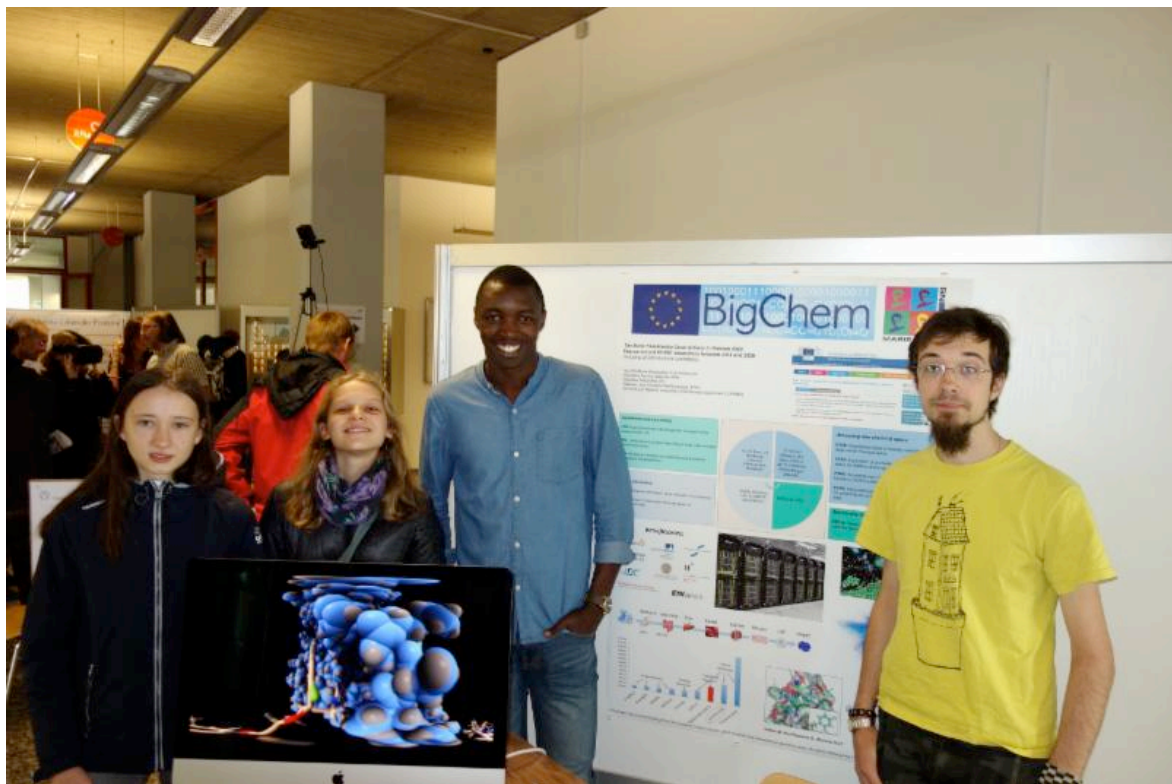
Dissemination of project results

- 7 published articles + 1 accepted for publication
- Participation in meetings and conferences: 7 posters, 16 oral presentations
- BIGCHEM website: publications, lectures from on-line courses and schools
- Professional networks: ResearchGate and LinkedIn




Outreach activities

- Newsletters: 4 delivered, 5 will be done; 329 subscribers
- Universities Open Days:
 - TUM: 2016 & 2017
 - ETHZ: 2017
 - Uni Bern: 2017
- Videos in YouTube







 Technische Universität München

Big data in Chemistry:
 Projekt BIGCHEM

Helmholtz-Zentrum,
 Institut für Strukturelle Biologie

Tag der offenen Tür

Helmholtz-Zentrum München
 German Research Centre for Environmental Health

BIGCHEM: Big Data In Chemistry, M

Overview

Beneficiaries:

AstraZeneca, Bayer, Boehringer Ingelheim, IDIC, Merck, Novartis, Roche, TUM, ETH Zurich

In one year, BIGCHEM has trained over 100 researchers from 15 countries and 165 institutions.

Research

... for training very ...
 ... by large-scale ...
 ... using ...
 ... for HTS ...
 ... on ...

Partners:

BI, Uni ...
 Strada ...
 Chem ...
 Mech ...

HMGU, Bonn, U ...
 AZ, Hel ...
 Mech ...

Research training

A1. Theoretical research ...
 ... of PhD projects with ...
 ... and career opportunities for ...

A2. Experimental approach ...
 ... in computational ...
 ... biological ...
 ... methods in order to ...
 ... of models, as well as ...

A3. Transferable skills ...
 ... research ...
 ... in primary ...
 ... through multidisciplinary ...



Outreach activities II

- MC Ambassadors:
 - ESR1 and ESR7: gave talks to Master students
 - ESR6: Research Night at Uni Bern



1. Scientific overview
2. Training
3. Networking
4. Management
 1. Recruitment report
 2. Deliverables
 3. Milestones
 4. Ethical issues, if applicable
 5. Management meetings
 6. Financial aspects
 7. Critical implementation risks and mitigation actions
 8. Any proposed re-orientations of the networks' activities
 9. Document management and Open Research Data, if applicable

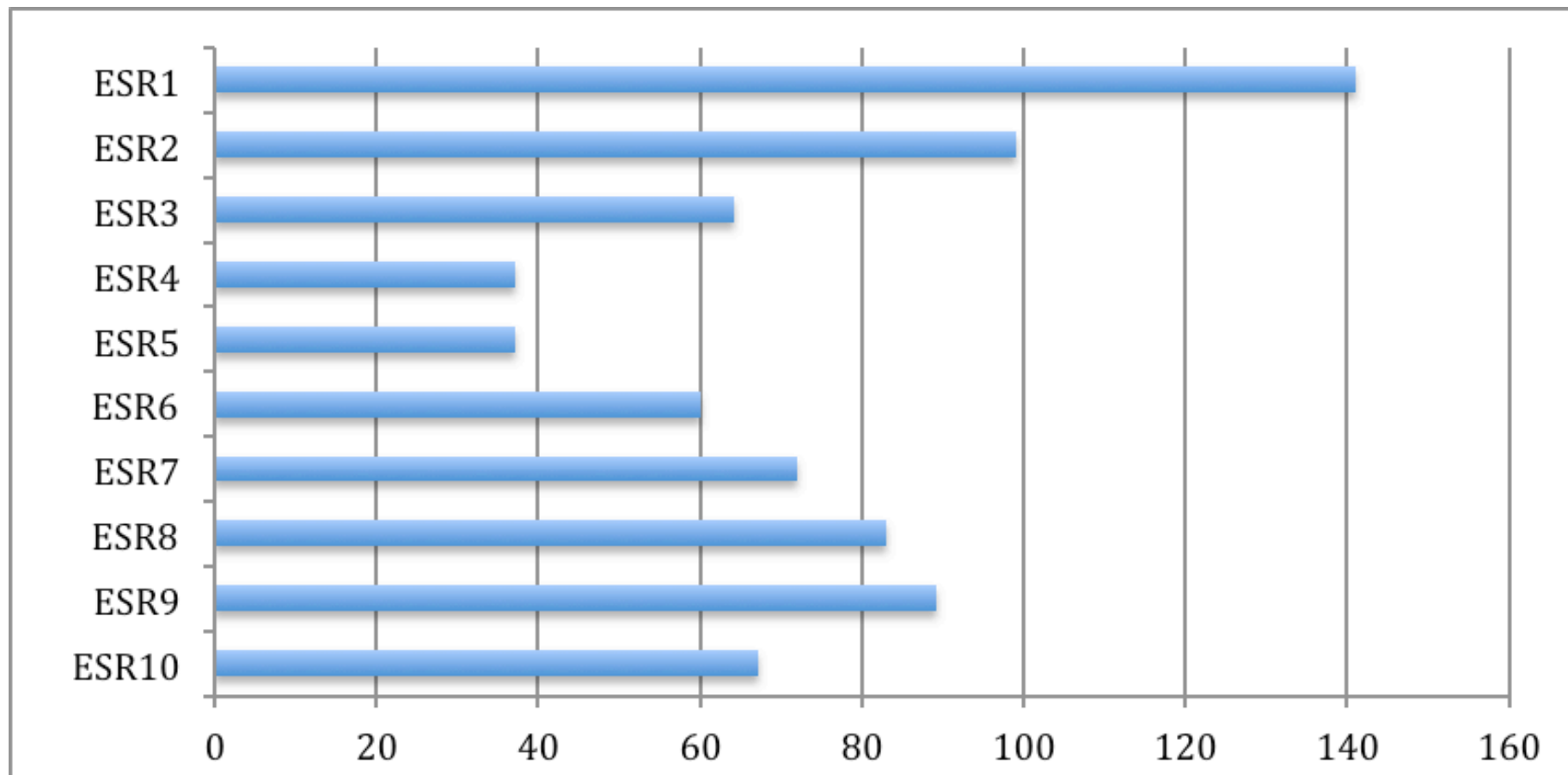


Recruitment

- 10 ESR positions announced internationally (EURAXESS, CCL, LinkedIn, KAGGLE, Naturejobs, DADD, PHDJobs...)
- Recruitment website: information and on-line application
- Central recruitment
- 277 applicants, 749 applications
- Applicants from 54 countries
- 24% female and 76% male candidates



Number of applications per ESR position



Recruitment: pre-selection of candidates

- Applications screened by both academic and industrial partners
- Eligibility was checked by PM
- Reviewed considering scientific background, computational skills, mobility experience, etc.
- Best matches of profiles and academic records
- Selected for the recruitment meeting:
 - 1 leading + 1-2 strong back-up candidates



Recruitment meeting

- April 4th, 2016, at HMGU
- 22 invited candidates (4 per Skype)



Positions ESR2 & ESR9

- Positions will be re-open due to resignation of fellows
- New hiring is currently discussed by the respective partners



Deliverables

Nr.	Title	Delivery date
D6.1	Minutes of the kick-off meeting	29.02.2016
D6.2	Web site and application system for fellows	29.02.2016
D6.9	Consortium agreement	04.10.2017
D6.13	Supervisory Board of the network	29.02.2016
D5.9	Updated training curriculum	30.06.2016
D6.3	Minutes of the recruitment meeting	09.06.2016
D7.1	POPD - Requirement No. 6	28.07.2016
D7.2	POPD - Requirement No. 5	28.07.2016
D7.3	POPD - Requirement No. 4	28.07.2016
D7.4	NEC - Requirement No. 3	28.07.2016
D7.5	NEC - Requirement No. 2	28.07.2016
D7.6	OEI - Requirement No. 1	28.07.2016
D5.1	Preparation of CDPs	30.01.2017
D5.2	1st Winter school report	21.12.2016



Deliverables 2

Nr.	Title	Delivery date
D6.4	Publication of newsletter	21.12.2016
D6.10	Progress Report	31.01.2017
D5.3	1st Summer school report	29.06.2017
D6.5	Organization of Open Days	29.06.2017
D1.1	Overview of public and <i>in house</i> data collected	21.08.2017
D2.1	Overview of HTS data	21.08.2017
D3.1	Review of selected targets	21.08.2017
D4.1	Overview of strategies for data sharing	Postponed*
D5.4	First overview of ESRs progress towards their PhDs	24.08.2017

*approved by the Project Officer



Milestones

Nr.	Title	Reported date
1	Launch of the web site of the project	29.02.2016
2	Selection of ESR fellows	28.07.2016
3	Launch of the on-line courses	29.09.2016
4	Schools started	17.10.2016
5	Targets identified	21.12.2017
6	Identify secure sharing information protocols	22.12.2017
7	Enrolment of ESR in PhD programs	15.01.2017
8	Data for model development collected	24.04.2017
9	HTS and activity data collected	21.03.2017
10	1st Evaluation of ESRs	01.08.2017

Ethical issues

- Does not apply to the project

Management meetings

- 18.01.2016 Kick-off meeting
- 04.04.2016 Recruitment meeting
- 06.05.2016 General Assembly and Supervisory Board meetings (per Skype)
- 20.10.2016 General Assembly and Supervisory Board meetings
- 19.4.2017 General Assembly
- 20.4.2017 Supervisory Board meeting
- Once a month: Management team meeting

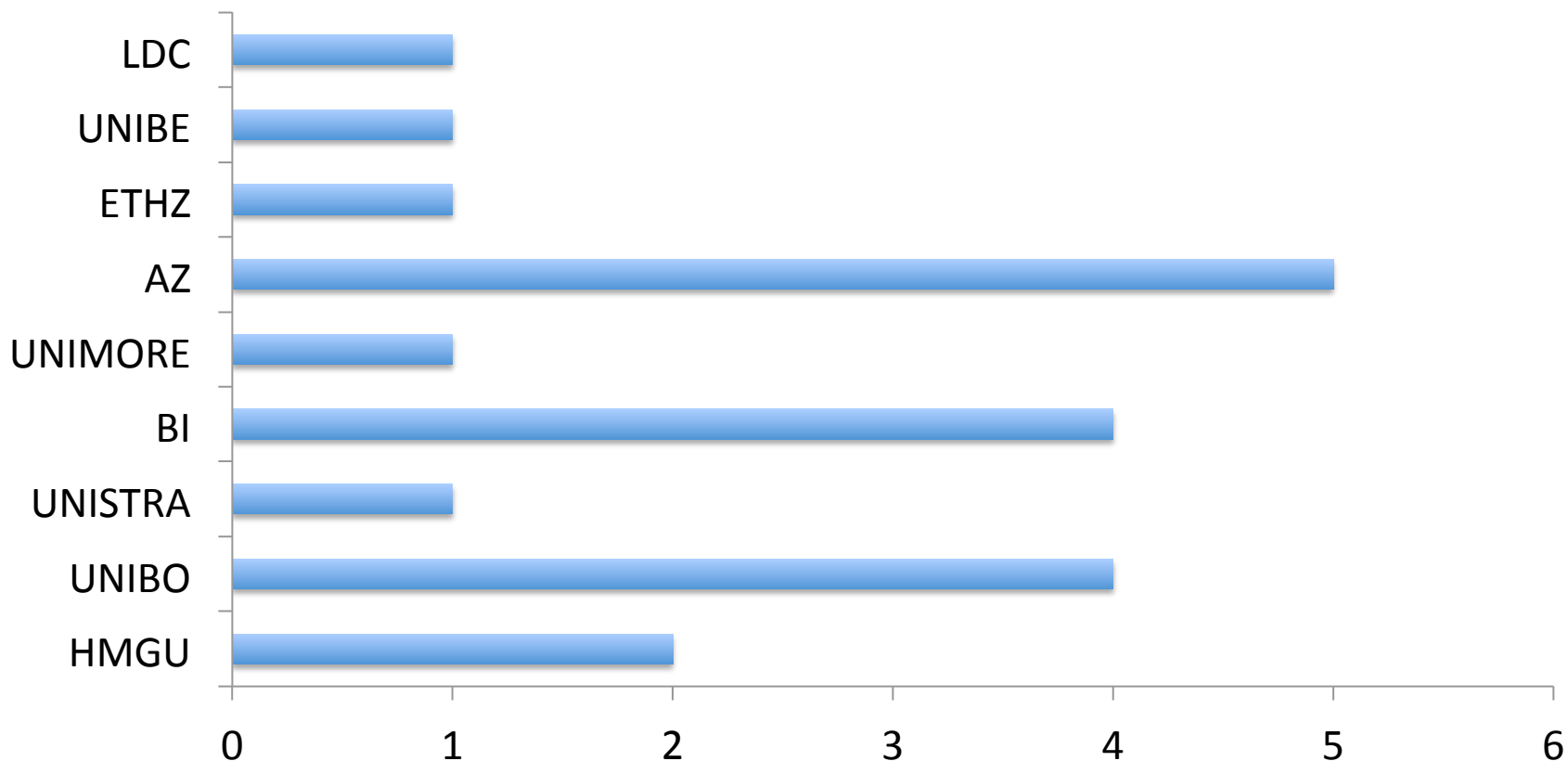


Management structure



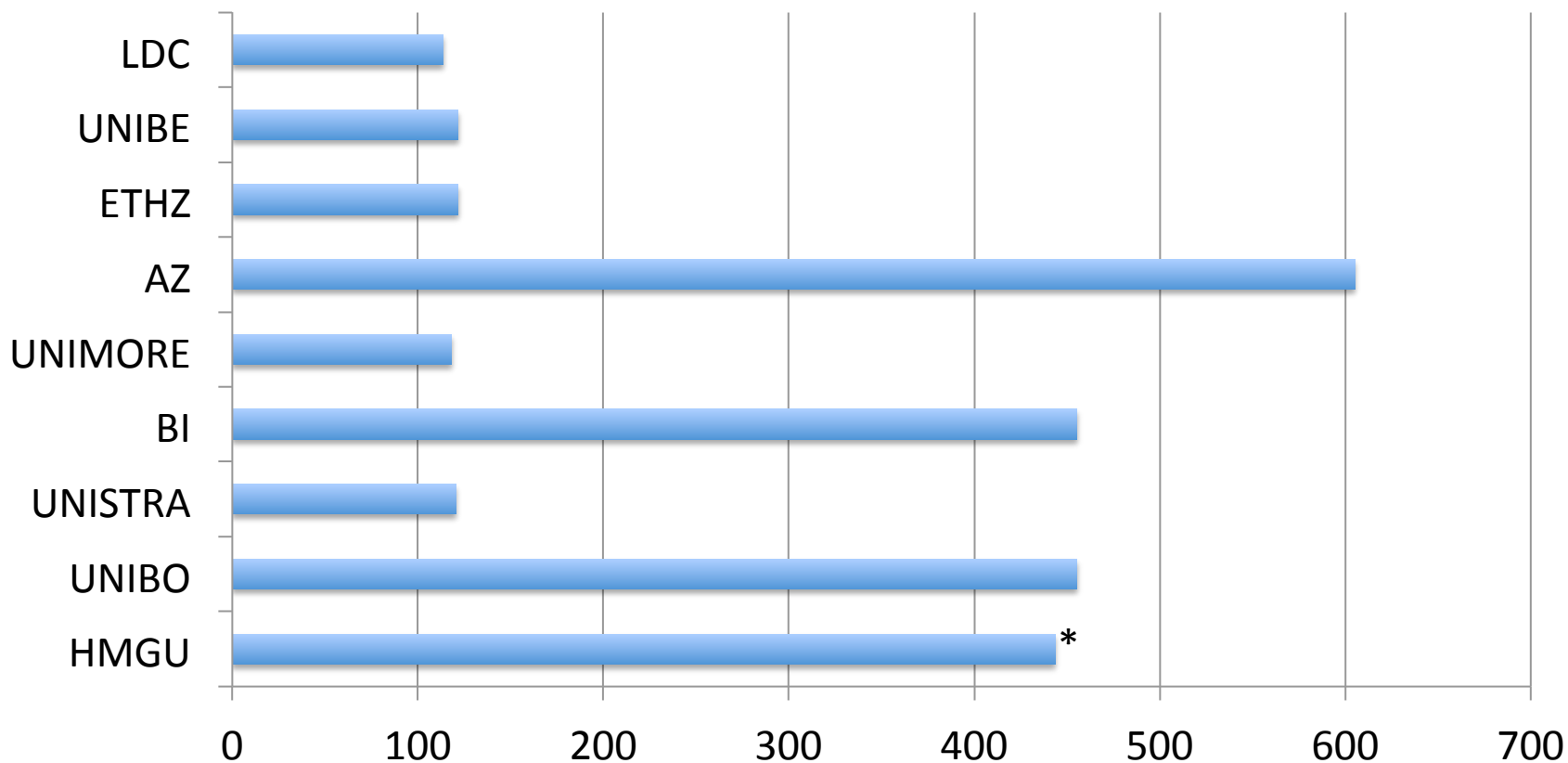
Financial aspects

ESR employed



Financial aspects

Total budget (€)

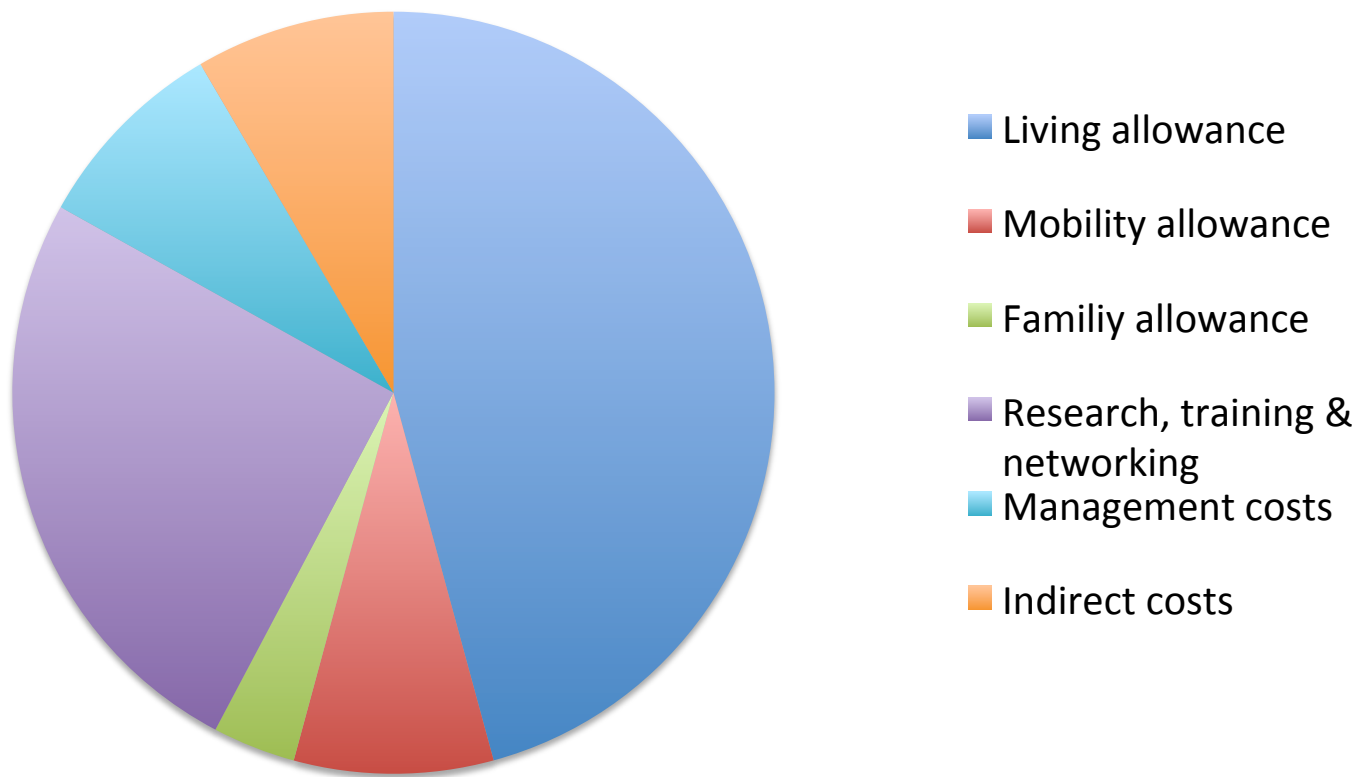


*Including 216k€ to support the management costs



Financial aspects

Budget distribution



Common pot for the network wide events



Grant Agreement amendment

- Signed by HMGU on October 13, 2017
- Modification of the employment order of ESR3 & 7
- Relocation of ESR8 from HMGU to the Uni Bonn
- Annex I was modified to reflect these changes as well as:
 - Updated composition of Thesis Committees
 - Updated training curriculum
 - New leader of curriculum: Hongming Chen
 - New associated partner: University of Dundee



Critical implementation risks

- Two fellows have left the project
- The employment of new fellows
 - Difficulties with enrollment in PhD programs
 - Difficulties with financing of 3rd year



Document management and open access

The screenshot shows the Zenodo website interface. At the top, the Zenodo logo is on the left, followed by a search bar containing 'BIGCHEM'. To the right of the search bar are links for 'Upload' and 'Communities', and buttons for 'Log in' and 'Sign up'. Below the search bar, the page indicates 'Found 7 results.' and shows a pagination control with '1' selected. On the left side, there are three filter panels: 'Access Right' with 'Open (7)' selected, 'File Type' with 'Pdf (7)' selected, and 'Keywords' with a list of terms like 'Design (2)', 'Learning (2)', 'Molecular (2)', 'Based (1)', 'Bidrug (1)', 'Chemogenomics (1)', 'De (1)', 'Discovery (1)', 'Drug (1)', and 'Equation (1)'. Below these is a 'Type' panel with 'Publication (7) +' selected. The main content area displays three search results, each with a date, version, document type, and 'Open Access' status, followed by a 'View' button. The first result is 'Matched Molecular Pair Analysis on Large Melting Point Datasets: A Big Data Perspective' by Michael Withnall, Hongming Chen, and Igor Tetko, uploaded on September 13, 2017. The second result is 'BIGCHEM: Challenges and Opportunities for Big Data Analysis in Chemistry' by Tetko, Igor, Engkvist, Ola; Koch, Uwe; Reymond, Jean-Louis; Chen, Hongming, uploaded on December 13, 2016. The third result is 'Does 'Big Data' exist in medicinal chemistry, and if so, how can it be harnessed?' by Tetko, Igor; Engkvist, Ola; Chen, Hongming, uploaded on December 13, 2016. The fourth result is 'On the Integration of In Silico Drug Design Methods for Drug Repurposing' by March-Vila, Eric; Pinzi, Luca; Sturm, Noé; Tinivella, Annachiara; Engkvist, Ola; Chen, Hongming; Rastelli, Giulio, uploaded on July 4, 2017.

<http://bigchem.eu> is the main hub of project documents



Thank you for your attention!

