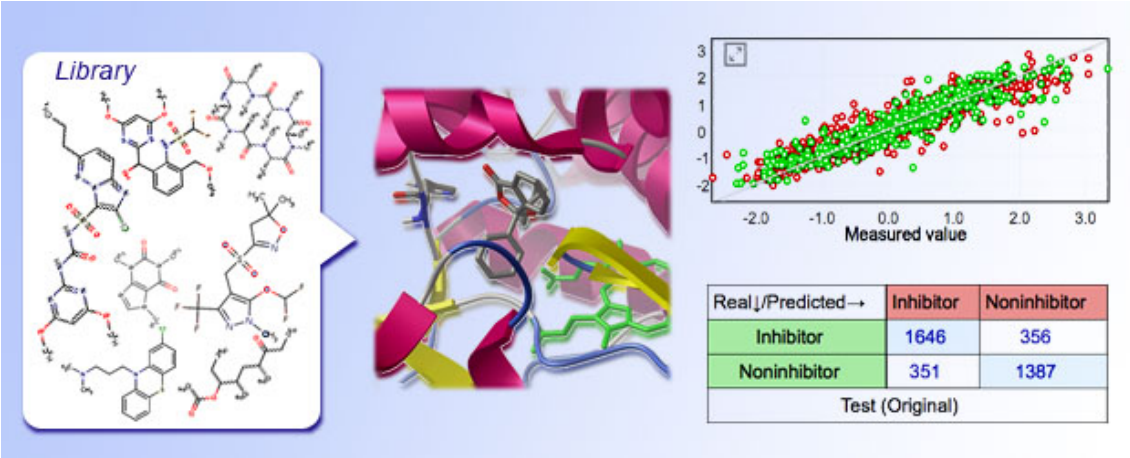


Helmholtz Zentrum München

Dr. Igor V. Tetko

Institute of Structural Biology, HelmholtzZentrum München & BIGCHEM GmbH

Wednesday, January 26, 2021

Institute of Structural Biology

HelmholtzZentrum münchen
German Research Center for Environmental Health

big chem

Helmholtz Association: Key Figures

Germany´s largest research organization

- 18 research centres
- Budget: 3.8 billion €, more than 36,000 staff

Mission

- We contribute to solving grand challenges
- We research systems of great complexity with large scale facilities
- We contribute to shaping our future

Research fields

- **Health**, Energy, Earth and Environment, Structure of Matter, Key Technologies, Transport and Space

Health Research Centers

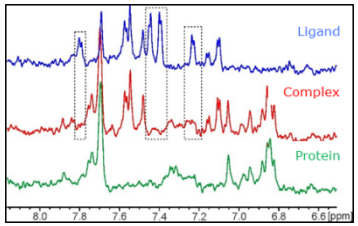
- Budget: 550 million €, about 5.500 staff
- Common Diseases: cancer, cardiovascular diseases, metabolic diseases, lung diseases and allergies, neurodegenerative diseases, infectious diseases
- Contribution to diagnosis, treatment, prevention
- Education and training of the next generation of scientists
- Move into the field of Precision Medicine

Helmholtz Zentrum München (HMGU): Key Figures

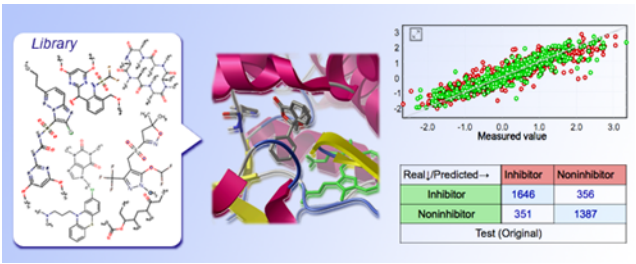
> 2 300	**Employees** (60% female, 70 nations)
21	**Young Investigator Groups**
49	**Institutes / Research Units**
31	**Appointments with universities**
1 448	**Publications**
26	**ERC Grants (total)**
3	**Translational centers**
> 40	**Clinical research projects** with partners
274	**Million € total budget**
45,7	**Million € third-party funding**
20	**Spin-off companies (since 1997)**
5	**Products**

Helmholtz Zentrum München
Structure-based drug discovery & technology platforms

Fragment Based Screening NMR hit validation
Grzegorz Popowicz, Ana Messias
Michael Sattler

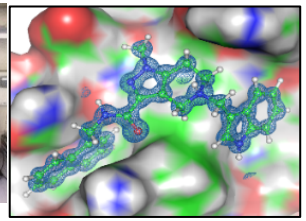


Chemoinformatics
Igor Tetko



Institute of Medicinal Chemistry
Oliver Plettenburg

X-ray Crystallography
Robert Janowski/Niessing



Protein Expression & Purification Facility
Arie Geerlof

- Protein production & support for structural biology & drug discovery
- Know-how, training, resources, i.e. expression vectors, general use proteins: TEV, precision, Cas9, …

Funding

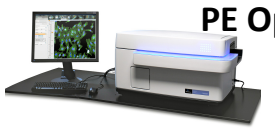
EU Horizon 2020 ITN "AEGIS"
EU Horizon 2020 ITN "*BIGCHEM*"
EU Horizon 2020 ITN "*AIDD*"
EU Horizon 2020 ITN "*RNAct*"
Novel methods in SBDD: BMWi-ZIM, TV, VIP+

Publications

Dawidowski *Science* (2017)
Jagtap *J Med Chem* (2016)
Riebold *Nature Medicine* (2015)
Gilsbach *J Med Chem* (2015)
Piccoli *Mol Cell Biol (*2014),
Zierer *Angew Chem (*2014)

Assay Development & Screening Platform
Kamyar Hadian

- In-house compound library
- Biochemical screening
  - AlphaScreen, FP, …
- Cell-based assays and screening



PE Operetta®

http://www.helmholtz-muenchen.de/stb

Accelerated Early staGe drug dIScovery

AEGIS

Accelerated Early staGe drug dIScovery

EC HORIZON 2020 Marie Skłodowska-Curie
Innovative Training Network (ITN)

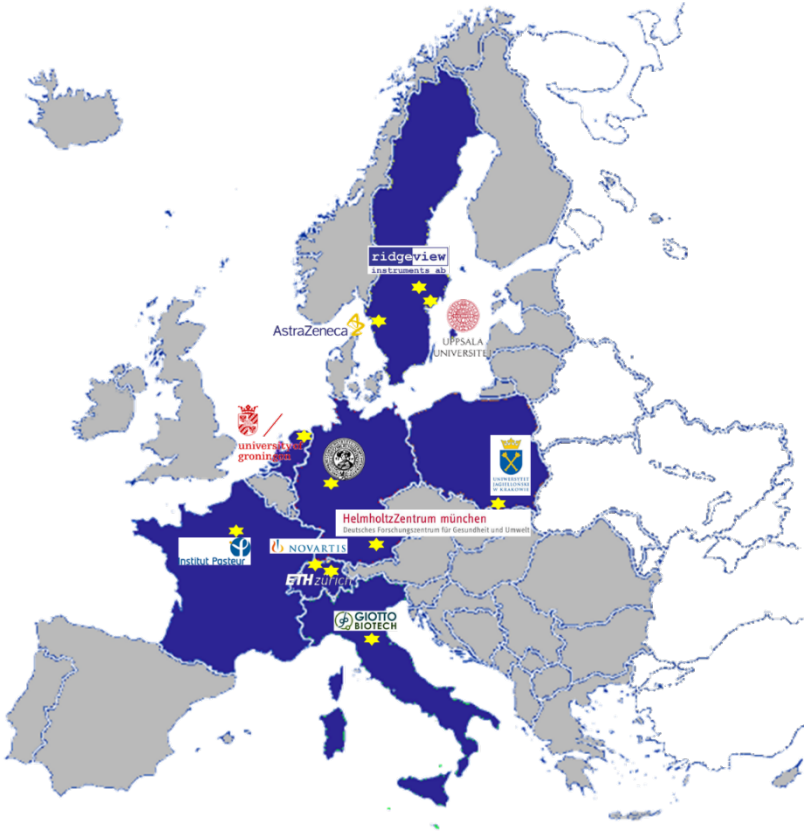
Coordination:
Michael Sattler, Helmholtz Zentrum München

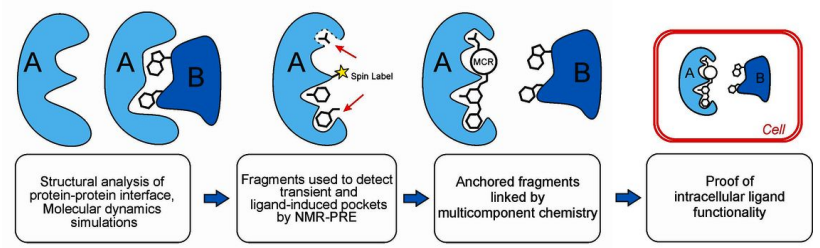
NOVARTIS

AstraZeneca

ridgeview instruments ab

GIOTTO BIOTECH





BigChem http://bigchem.eu

big data in chemistry + informatics = chemoinformatics

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

The BIGCHEM project will provide **innovative education in large chemical data analysis**. The innovative research program will be implemented with the target users, **large pharma companies and SMEs**, which generate and analyze large chemical data as well as will promote technology transfer from academy to industrial applications.

Marie Skłodowska-Curie Innovative Training Network
European Industrial Doctorate

BIGCHEM project publications http://bigchem.eu

BIGCHEM publications

FOLLOW

Horizon2020 Marie Skłodowska-Curie Innovative Training Network European Industrial Doctorate

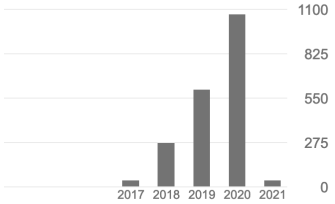
Verified email at bigchem.eu

big data chemoinformatics cheminformatics

TITLE	CITED BY	YEAR
The rise of deep learning in drug discovery H Chen, O Engkvist, Y Wang, M Olivecrona, T Blaschke Drug discovery today 23 (6), 1241-1250	520	2018
Molecular de-novo design through deep reinforcement learning M Olivecrona, T Blaschke, O Engkvist, H Chen Journal of cheminformatics 9 (1), 48	325	2017
Automating drug discovery G Schneider Nature Reviews Drug Discovery 17 (2), 97	233	2018
Application of Generative Autoencoder in *De Novo* Molecular Design T Blaschke, M Olivecrona, O Engkvist, J Bajorath, H Chen Molecular informatics 37 (1-2), 1700123	175	2018
BIGCHEM: challenges and opportunities for big data analysis in chemistry IV Tetko, O Engkvist, U Koch, JL Reymond, H Chen Molecular informatics 35 (11-12), 615-621	71	2016
On the integration of in silico drug design methods for drug repurposing E March-Vila, L Pinzi, N Sturm, A Tinivella, O Engkvist, H Chen, G Rastelli Frontiers in pharmacology 8, 298	69	2017
QSAR without borders EN Muratov, J Bajorath, RP Sheridan, IV Tetko, D Filimonov, V Poroikov, ... Chemical Society Reviews	51	2020
Exploring the GDB-13 chemical space using deep generative models J Arús-Pous, T Blaschke, S Ulander, JL Reymond, H Chen, O Engkvist Journal of cheminformatics 11 (1), 1-14	51	2019
Randomized SMILES strings improve the quality of molecular generative models J Arús-Pous, SV Johansson, O Prykhodko, EJ Bjerrum, C Tyrchan, ... Journal of cheminformatics 11 (1), 1-13	45	2019
A de novo molecular generation method using latent vector based generative adversarial network O Prykhodko, SV Johansson, PC Kotsias, J Arús-Pous, EJ Bjerrum, ... Journal of Cheminformatics 11 (1), 74	34	2019
Chemical space: big data challenge for molecular diversity M Awale, R Visini, D Probst, J Arus-Pous, JL Reymond CHIMIA International Journal for Chemistry 71 (10), 661-666	27	2017

Cited by

	All	Since 2016
Citations	2035	2032
h-index	16	16
i10-index	29	29



Co-authors VIEW ALL

- Ola Engkvist, AstraZeneca R&D Gothenburg O…
- Hongming Chen, Astrazeneca R&D Mölndal
- Jürgen Bajorath, Professor of Life Science Inform…
- Thomas Blaschke, Phd student, AstraZeneca/Unive…
- Igor V. Tetko, Group Leader at Helmholtz Zentr…
- Jean-Louis Reymond, University of Bern
- Josep Arús-Pous, University of Bern
- Esben Jannik Bjerrum, Principal Scientist - Machine lear…
- Raquel Rodríguez-Pérez, Senior Scientist, Novartis Institut…
- Alexandre Varnek, Professor of Chemistry, Universit…
- Michael Withnall, Apheris AI

Up to now ~ 70 articles, including five highly cited articles according to the Web of Science

ICANN19

Organising committee

General Chair

Igor Tetko, ENNS, Helmholtz Zentrum München (GmbH), Germany

Fabian Theis, Helmholtz Zentrum München (GmbH), Germany

Honorary Chair

Vera Kurkova, Czech Academy of Sciences (ENNS President)

Steering Committee

Vera Kurkova, Czech Adacemy of Sciences (President of ENNS)

Erkki Oja, Aalto University, Finland (ex-President of ENNS)

Wlodzislaw Duch, Nicolaus Copernicus University, Poland (ex-President of ENNS)

Alessandro Villa, University of Lausanne, Switzerland (ex-President of ENNS)

Cesare Alippi, Politecnico di Milano, Italy and Università della Svizzera Italiana, Switzerland

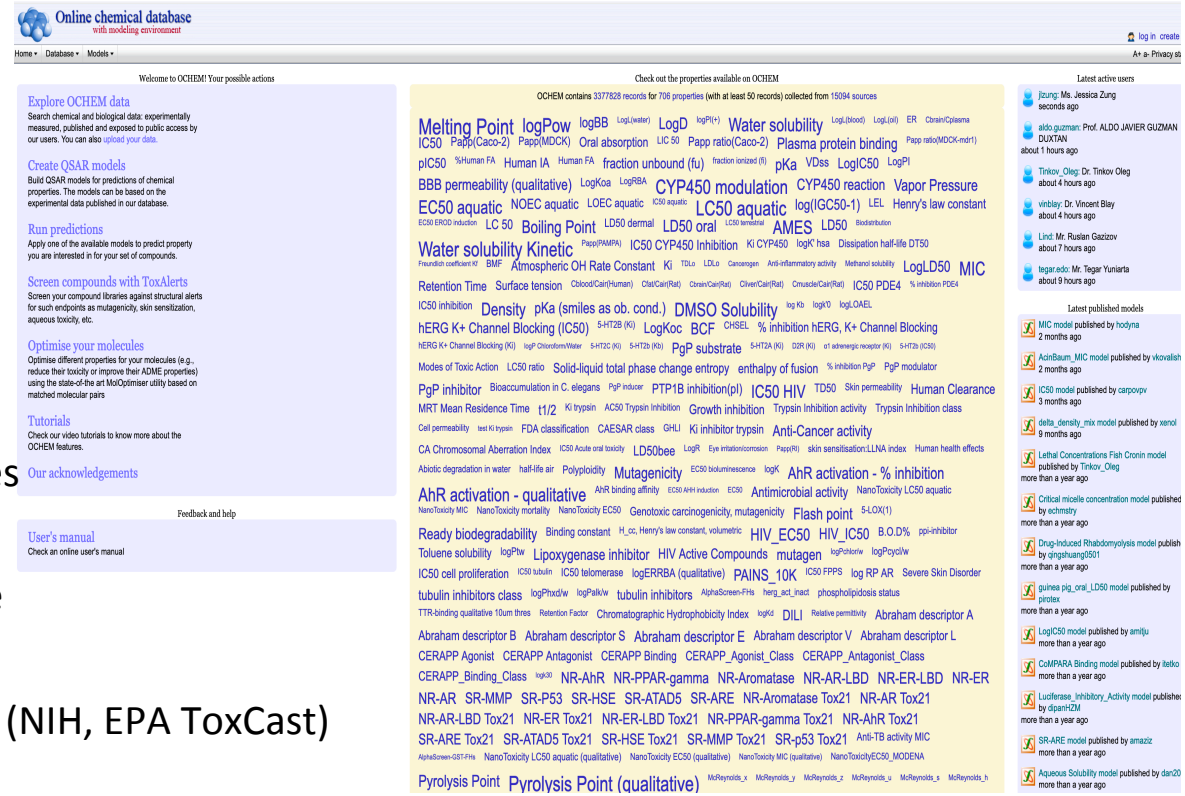
Jérémie Cabessa, Université Paris 2 Panthéon-Assas, France

Maxim Fedorov, Skoltech, Russia

OCHEM http://ochem.eu overview

- **Physico-chemical properties:** logP, water solubility, melting point, pyrolysis, vapor pressure, ODT, etc.
- **Biological activity:** estrogen receptors; endocrine disruptors; AMES mutagenicity; *in vivo* toxicity
- **Environmental endpoints:** ready biodegradability; fish toxicity; environmental toxicity, daphnia, etc.

- In total 162 published models*
- 7375 registered users
  - 600 commercial
  - 450 governmental
- ca 37M tasks were executed
- ca 3.4M data points for 692 properties
- >25M uploaded private data points
- Academic groups regularly contribute
- OCHEM is used for teaching
- Top-performing models in challenges (NIH, EPA ToxCast)



* As from 18.01.2021

OCHEM modeling

- Comprehensive modeling
- Multitask learning (up to 100 properties)
- >20 descriptor blocks
- Feature net (“model in model”)
- Consensus models
- GPU + CPU modern methods (~20)
- Supports models
  - >1,000,000 compounds
  - >200,000,000,000 descriptors*
  - >1,000 servers
  - up to 1GB in size (Java limit)
- Model private/publishing
- Export, import, web/REST services
- Conditions, external descriptors
- ToxAlerts

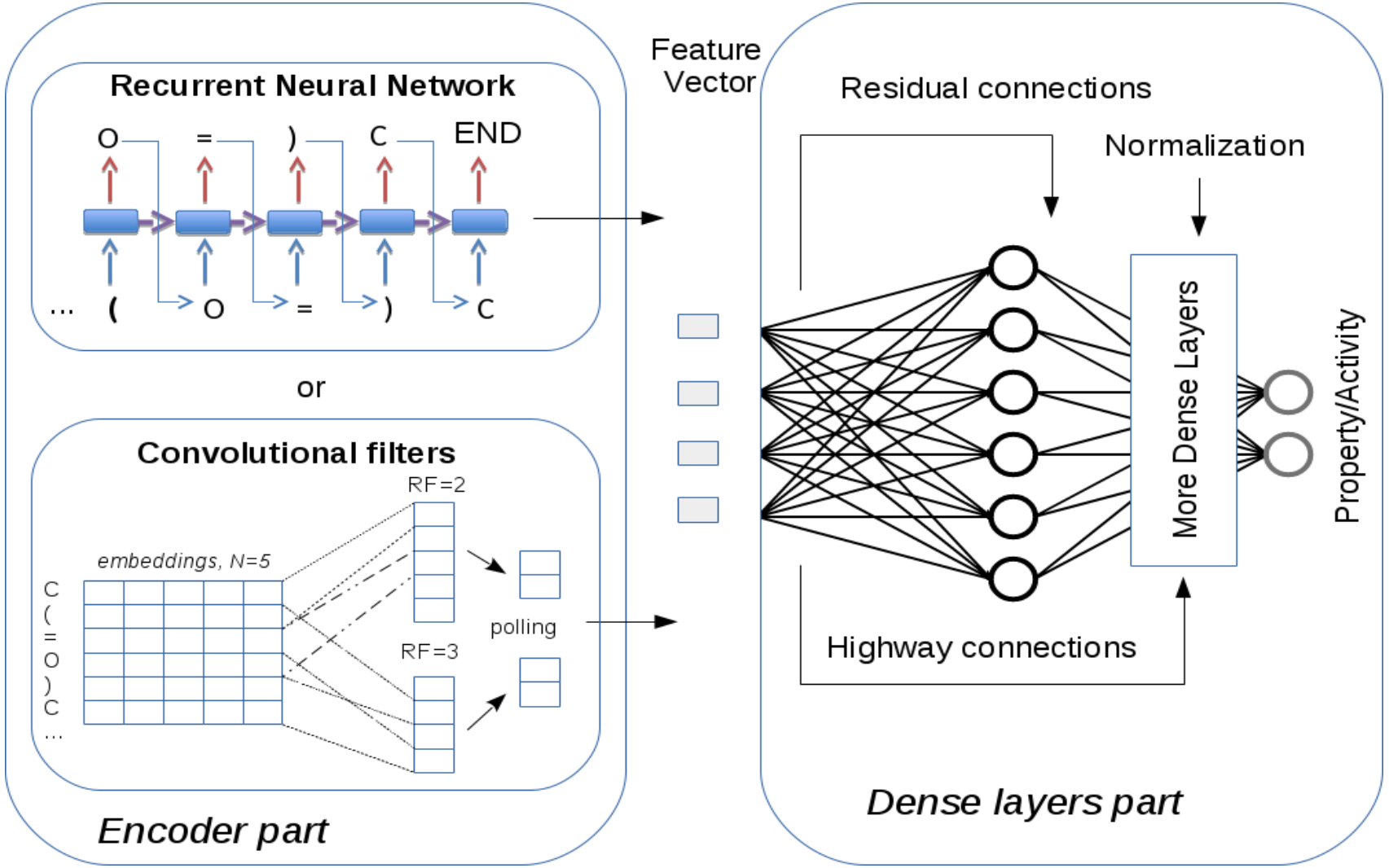
Predicted property: LLNA skin sensitization
Training set: TRAINING-SARpy-SKIN-SENS-giugno20 OK.xlsx

Metrics AUC for Training set Validation: Cross-Validation (84 models)

	LSSVMG	ASNN	PLS	KNN
ALogPS, OEstate	0.74	0.68	0.61	0.64
CDDD	0.8	0.74	0.75	0.71
CDK2 (cons,topol,geom,elec,hybrid) 3D:corina	0.75	0.71	0.56	0.71
ChemaxonDescriptors (pH 0 - 14:1) 3D:corina	0.76	0.7	0.59	0.68
Dragon6 (2D blocks: 1 28)	0.64	0.66	0.59	0.65
Dragon6 (3D blocks: 1-29) 3D:corina	0.76	0.72	0.57	0.65
Fragmentor (length:2 - 4)	0.72	0.7	0.59	0.63
GSFrag (F + L)	0.69	0.69	0.61	0.61
InductiveDescriptors 3D:corina	0.69	0.71	0.57	0.67
JPlogP	0.73	0.74	0.59	0.67
MAP4	0.71	0.65	0.59	0.67
MORDRED (All) 3D:corina	0.77	0.73	0.57	0.68
Mera, Mersy 3D:corina	0.73	0.69	0.55	0.67
OEstate	0.74	0.67	0.63	0.68
PyDescriptor 3D:corina	0.71	0.71	0.7	0.67
QNPR (length:1 - 3)	0.68	0.62	0.58	0.58
RDKIT (3D blocks: 1-11 15-16) 3D:corina	0.77	0.72	0.56	0.65
SIRMS (labels:charge+logp+hb+refractivity)	0.76	0.73	0.59	0.67
Spectrophores (accuracy=20) 3D:corina	0.68	0.6	0.52	0.6
StructuralAlerts	0.67	0.64	0.58	0.51
alvaDesc (3D blocks: (only) 1-30) 3D:corina	0.75	0.71	0.57	0.68

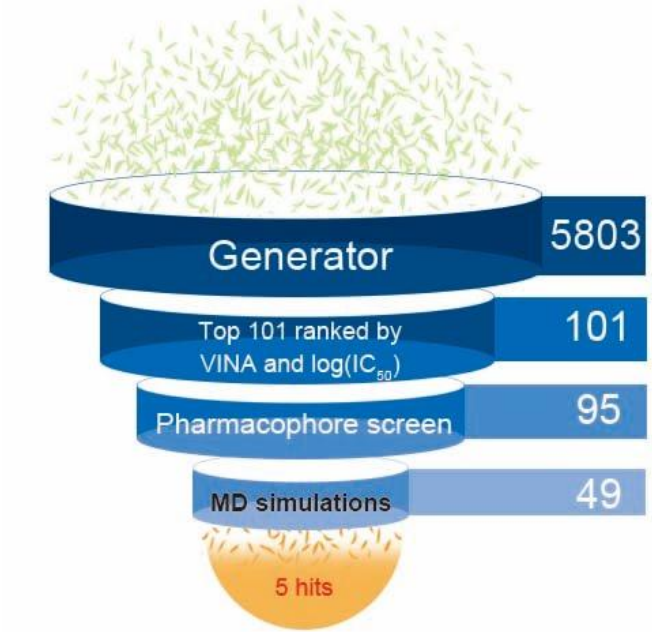
* Sparse format, DOI:10.1186/s13321-016-0113-y

Machine Learning directly from chemical structures



Karpov, P.; Godin, G.; Tetko, I.V. Transformer-CNN: Swiss knife for QSAR modeling and interpretation. *J. Cheminform.* **2020**, *12*, 17, doi:10.1186/s13321-020-00423-w.

Computer aided drug design of Mdmx inhibitors

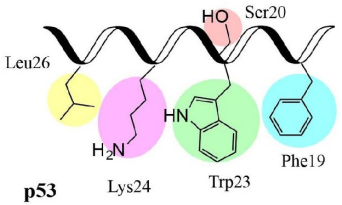


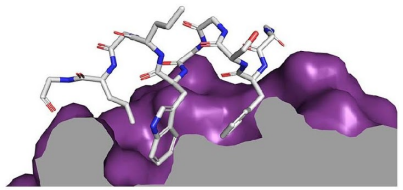
	RMSD				-*log*(IC_{50})	ΔG (kcal/mol)
Compound	**avg[a]**	**std[b]**	**min[c]**	**max[d]**		
WK298	2.183	0.662	0.543	4.227	-4.7	-4.1
3021	4.675	0.379	0.627	5.947	-5.2	-13.0
92	1.605	0.454	0.506	3.711	-7.7	-10.8
100	1.738	0.680	0.480	5.831	-7.9	-6.9
34	2.789	0.696	0.668	5.176	-7.9	-6.7
39	4.407	1.184	0.778	7.960	-7.6	-6.7

a. avg = the average; b. std = standard deviation; c. min = the minimum; d. max = the maximum

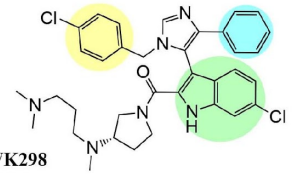
Xia, Z.; Karpov, P.; Popowicz, G.; Tetko, I.V. Focused Library Generator: case of Mdmx inhibitors. *J. Comput. Aided. Mol. Des.* **2020**, *34*, 769-782.

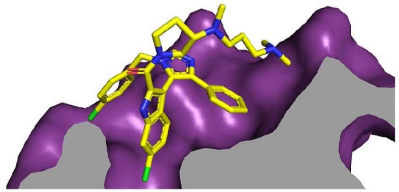
a)



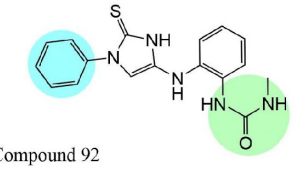


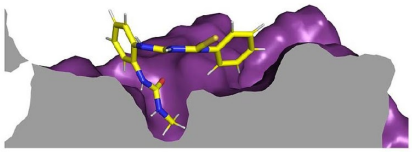
b)

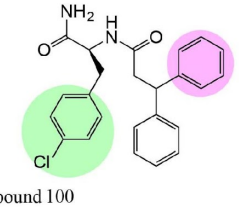


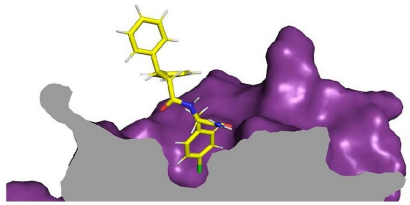


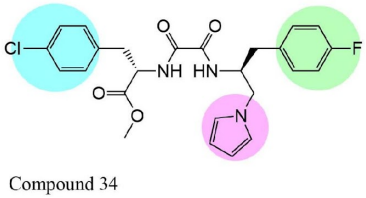
c)

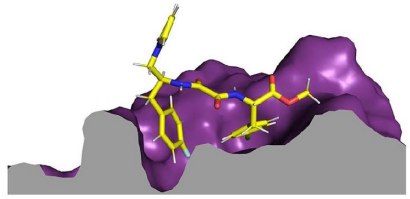


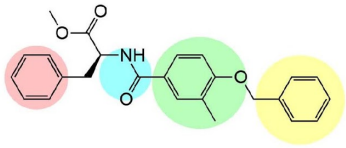




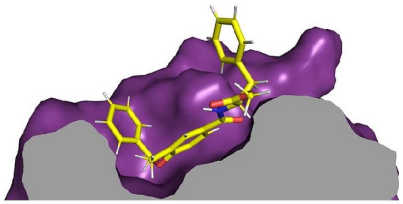






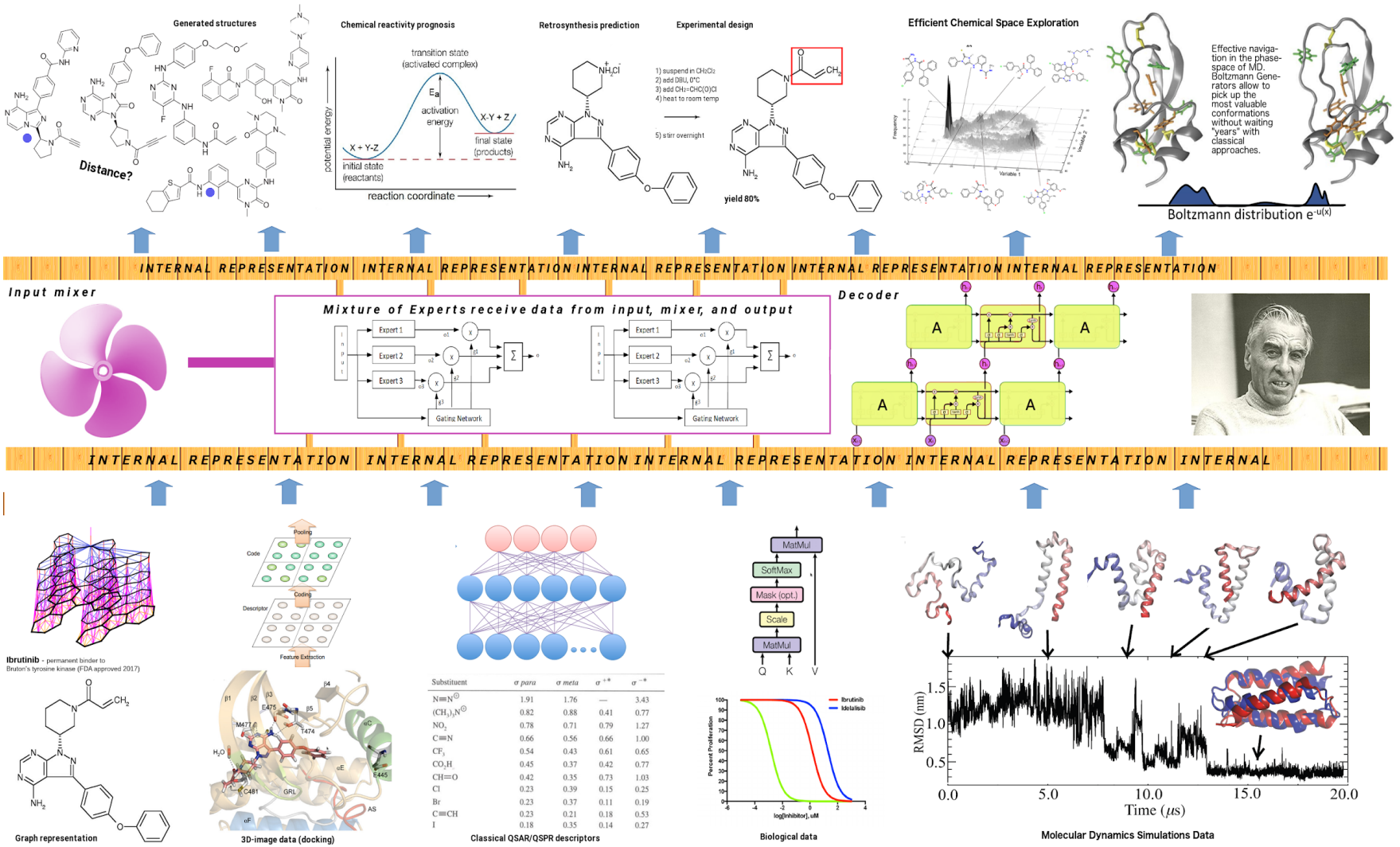


f)

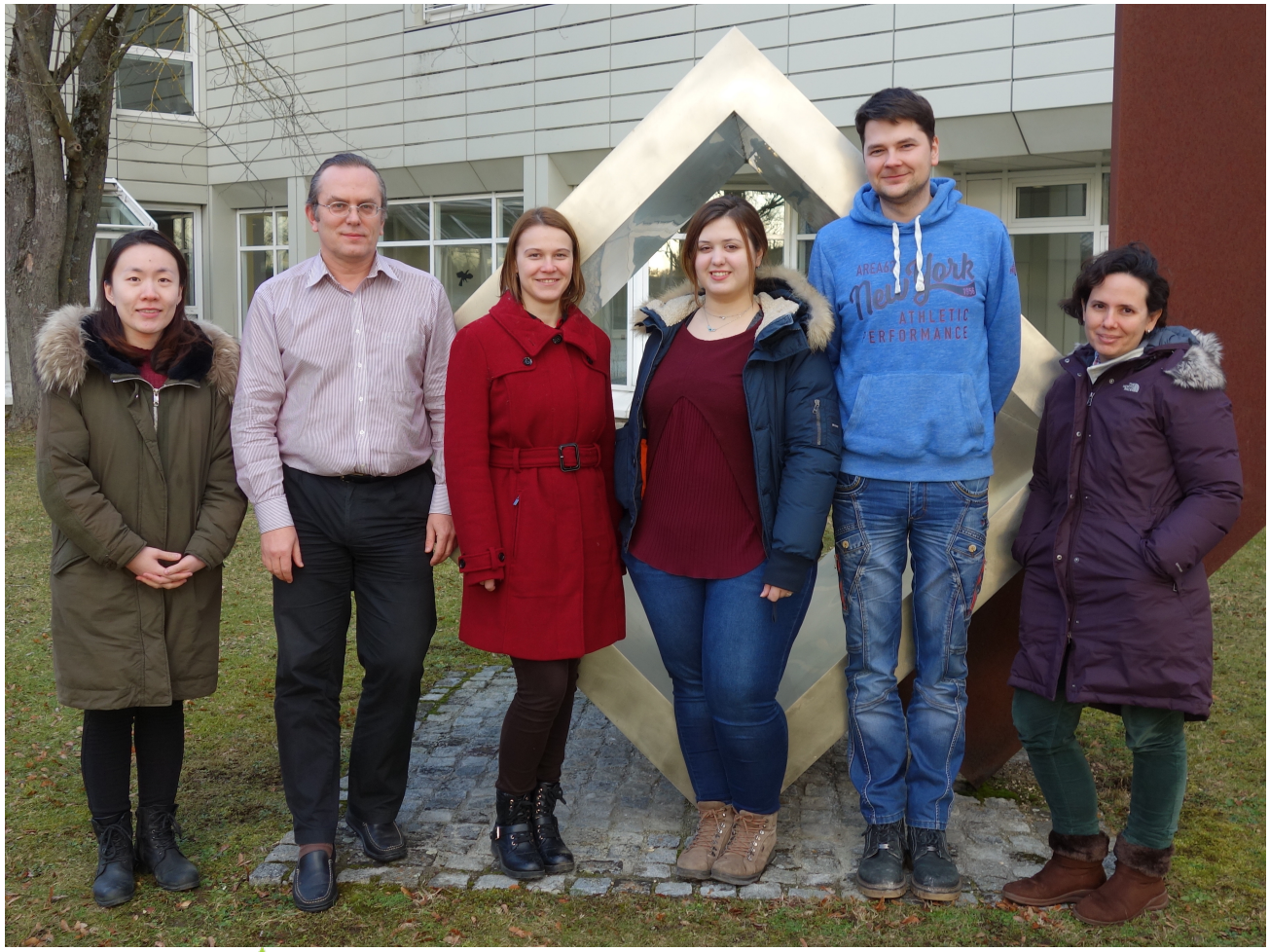


Tetko, I.V.; Karpov, P.; Van Deursen, R.; Godin, G. State-of-the-art augmented NLP transformer models for direct and single-step retrosynthesis. *Nat. Comm.* **2020**, *11*, 1-11, doi:10.1038/s41467-020-19266-y.

AI "One Chemistry" model for drug discovery



Acknowledgement



Pavel Karpov
Zhonghua Xia
Mark Embrechts
Joseph Yap
Dipan Ghosh
Michael Withnall
Monica Campillos
Genny Cau
Elena Golosovkaia

Alexander von Humboldt
Stiftung/Foundation

ERA CVD

MARIE CURIE ACTIONS

eeco
environmental chemoinformatics
marie curie initial training network

BigChem