

Helmholtz Zentrum München

Dr. Igor V. Tetko

Institute of Structural Biology, HelmholtzZentrum München & BIGCHEM GmbH

Institute of Structural Biology Wednesday, January 26, 2021

HelmholtzZentrum münchen German Research Center for Environmental Health



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

>



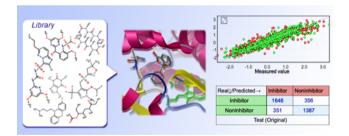
2 300	Employees (60% female, 70 nations)
21	Young Investigator Groups
49	Institutes / Research Units
31	Appointments with universities
1 448	Publications
26	ERC Grants (total)
3	Translational centers
> 40	Clinical research projects with partners
274	Million € total budget
45,7	Million € third-party funding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

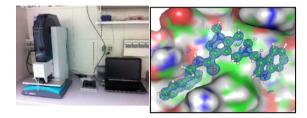
Chemoinformatics Igor Tetko



Institute of Medicinal Chemistry Oliver Plettenburg

X-ray Crystallography

Robert Janowski/Niessing



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)

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, ...

Assay Development & Screening Platform

Kamyar Hadian

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



Accelerated Early staGe drug dlScovery



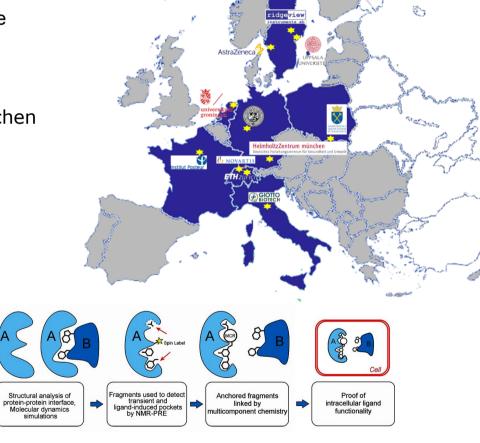
EC HORIZON 2020 Marie Skłodowska-Curie Innovative Training Network (ITN)

Coordination: Michael Sattler, Helmholtz Zentrum München

UNOVARTIS









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				Cited by			
BigChem	Horizon2020 Marie Skłodowska-Curie Innovative Training Network Europe Doctorate Verified email at bigchem.eu			Since 2016				
	big data chemoinformatics cheminformatics			h-index i10-index	16 29	16 29		
TITLE		CITED BY	YEAR			1100		
	Warning in drug discovery Wang, M Olivecrona, T Blaschke 23 (6), 1241-1250	520	2018			825 550		
	design through deep reinforcement learning nke, O Engkvist, H Chen atics 9 (1), 48	325	2017		2017 2018 2019 202	275		
Automating drug di G Schneider Nature Reviews Drug D		233	2018		2017 2018 2019 202	20 2021		
	erative Autoencoder in <i>De Novo</i> Molecular Design ona, O Engkvist, J Bajorath, H Chen 37 (1-2), 1700123	175	2018	Co-authors		VIEW ALL		
	nges and opportunities for big data analysis in chemistry I Koch, JL Reymond, H Chen 35 (11-12), 615-621	71	2016	Hongm	eneca R&D Gothenb ning Chen eneca R&D Mölndal	vrg O >		
	of in silico drug design methods for drug repurposing N Sturm, A Tinivella, O Engkvist, H Chen, G Rastelli ogy 8, 298	69	2017	Profession	Bajorath sor of Life Science In s Blaschke			
QSAR without bord EN Muratov, J Bajorath Chemical Society Revie	n, RP Sheridan, IV Tetko, D Filimonov, V Poroikov,	51	2020	Phd str	udent, AstraZeneca/l	>		
	- 13 chemical space using deep generative models ke, S Ulander, JL Reymond, H Chen, O Engkvist atics 11 (1), 1-14	51	2019	Univer:	ouis Reymond sity of Bern	>		
	ES strings improve the quality of molecular generative models nsson, O Prykhodko, EJ Bjerrum, C Tyrchan, atics 11 (1), 1-13	45	2019	Univers	Arús-Pous sity of Bern Jannik Bjerrum	>		
network	ar generation method using latent vector based generative adversarial	34	2019	Raque	al Scientist - Machine I Rodríguez-Pérez Scientist, Novartis In	e lear		
Journal of Cheminform	atics 11 (1), 74				dre Varnek sor of Chemistry, Uni	versit >		
M Awale, R Visini, D Pr	ig data challenge for molecular diversity robst, J Arus-Pous, JL Reymond ournal for Chemistry 71 (10), 661-666	27	2017	Michae Apheria	e l Withnall s Al	>		

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



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.

	Online chemical database		
н	tome · Database · Models ·		Iog in crea A+ a- Privacy
	Welcome to OCHEM! Your possible actions	Check out the properties available on OCHEM	Latest active users
	Explore OCHEM data Search chemical and biological data experimentally measured, publican deropende trajulal access by our user, You can also spont your data. Create Q&AR models Middl CSAR models	OCHEM contains 3377/228 records to 709 properties (with at least 50 records) collected from 15094 sources Melting Point (ogPow logBs Lessitivity LogDim Vortege Deptile LogDim Vortege Deptile Control Case-02 Paper disolation End Control Case-02 Paper disolation End Control Case-02 Paper disolation End Control Case-02 Paper disolation Control Case-03 Paper disolation Case-03 Case-03 Paper disolation<	 jzung: Ms. Jessica Zung seconds ago aldo guzmar: Prof. ALDO JAVIER GUZMA DUXTAN about 1 hours ago about 4 hours ago
	properties. The models can be based on the experimental data justicle in our database. Run predictions Apply one of the available models to predict property you are interested in provised of compounds. Screen compounds with ToxAlerts	ECSD aquatic NEC aquatic LOEC aquatic LOEC aquatic Constant Consta	Winblay: Dr. Vincent Blay about 4 hours ago Lind: Mr. Ruxian Gazzov bout 7 hours ago Iggradow Mr. Tegar Yuniarta about 9 hours ago
erties	Screen you compound litaries against structural alerts for such responds an admagneticity, allo resolutation, aquesus succity, etc. Optimise different properties for your moleculas (e.g., reduce ther huckpy or improve their ANAE properties) umg her state-of-the MoDoriman utility based on matched onlocalar pars Tuttorials Check our video lutrihis lo know more about the OCHEM statures. Our acknowledgements	IC50 inhibition Density pKa (smiles as ob. cond.) DMSO Solubility ^{bayte} ^{bay}	Latest published models
ts	Feedback and help User's manual Check an online user's manual	Newfoundy Mic NamoToxiday monthly NamoToxiday ECS0 Genotoxic carcinogenicity, mutagenicity Flash point 54.0X(1) Ready biodegradability Binding constant H.cs, Henry's law constant, volumeric HIV_ECS0 HIV_IC50 B.0.0%, spi-inhibitor Toluene solubility logPer Lipoxygenase inhibitor HIV Active Compounds mutagen logPersone logPeyotw	Critical micelle concentration model publish by echmstry more than a year ago Drug-Induced Rhabdomyolysis model public by cingshuang0501
oute		IC50 cell proliferation (C50 tabler IC50 tabler as logERBA (qualitative) PAINS_10K (C50 FPRs) log RP AR Severe Skin Disorder tubulin inhibitors class logPradw topPalw tubulin inhibitors. AbrakemenFile herg act read phospholipidesis status TTR-binding qualitative form three Retension Factor Chromatographic Hydrophobicity Index ket/s DILL Retension Provider Abraham descriptor A Abraham descriptor B Abraham descriptor S Abraham descriptor E Abraham descriptor L	more than a year ago y guinea pig_oral_LD50 model published by pirotex: more than a year ago y Log(C50 model published by amitju more than a year ago
nges (NIH, EPA ToxCast)	CERAPP Agonist CERAPP Anagonist CERAPP Binding CERAPP Agonist_Class CERAPP_Anagonist_Class CERAPP_Binding_Class tw/st NR-AhR NR-PPAR-gamma NR-Aromatase NR-AR-LBD NR-ER-LBD NR-ER NR-AR SR-MMP SR-P53 SR-HSE SR-ATAD5 SR-ARE NR-Aromatase Tox21 NR-AR Tox21 NR-AR-LBD Tox21 NR-ER Tox21 SR-HSE SR-ATAD5 SR-ARE NR-Aromatase Tox21 NR-AR Tox21 SR-ARE Tox21 SR-ATAD5 Tox21 SR-HSE Tox21 SR-MMP Tox21 SR-AP63 Tox21 NR-BR Tox21 AnadomicsTree NanoToxicity LCS0 equilitative NanoToxicity ECS0 (qualitative NanoTo	CMAPARA Binding model published by ited more than a year ago by dipamit20 more than a year ago Si RARE model published by amaziz more than a year ago Si RARAE model published by amaziz more than a year ago

- 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 Our acknowledge
- >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

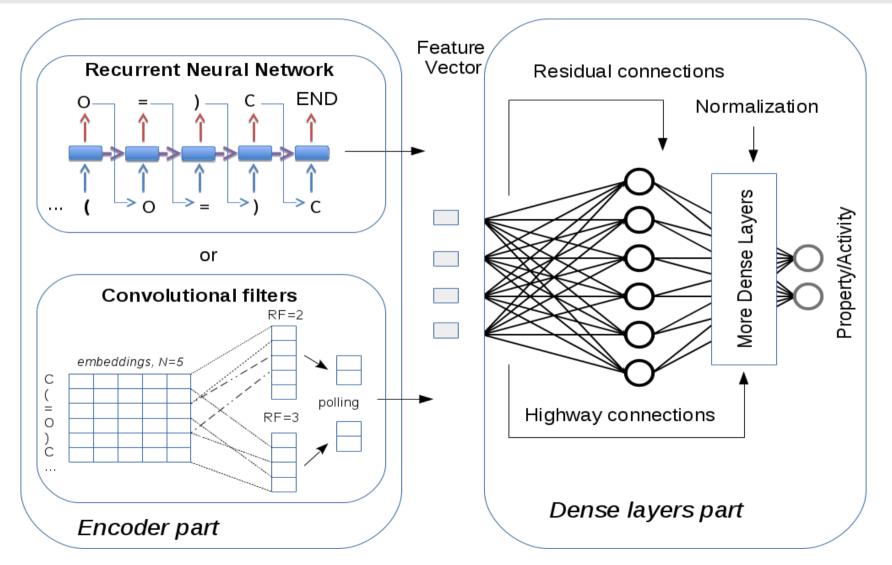
- Comprehensive modeling
- Multitask learning (up to 100 properties)
- >20 descriptor blocks
- Feature net ("model in model")
- Consensus models
- GPU + CPU modern methods (~20)
- Supports models
 - >1,000,000 compounds
 - >200,000,000,000 descriptors*
 - >1,000 servers
 - up to 1GB in size (Java limit)
- 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 of Training set Validation: Cross-Validation (84 models)					
		LSSVMG	ASNN	PLS	KNN
AL	ogPS, OEstate	0.74	0.68	0.61	0.64
	CDDD	0.8	0.74	0.75	0.71
CDK2 (cons,topo	l,geom,elec,hybrid) 3D:corina	0.75	0.71	0.56	0.71
ChemaxonDescr	iptors (pH 0 - 14:1) 3D:corina	0.76	0.7	0.59	0.68
Dragor	16 (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
Fragm	entor (length:2 - 4)	0.72	0.7	0.59	0.63
G	SFrag (F + L)	0.69	0.69	0.61	0.61
Inductive	Descriptors 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
MORDF	RED (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
PyDes	scriptor 3D:corina	0.71	0.71	0.7	0.67
QNI	PR (length:1 - 3)	0.68	0.62	0.58	0.58
RDKIT (3D blo	cks: 1-11 15-16) 3D:corina	0.77	0.72	0.56	0.65
SIRMS (labels:c	harge+logp+hb+refractivity)	0.76	0.73	0.59	0.67
Spectrophore	s (accuracy=20) 3D:corina	0.68	0.6	0.52	0.6
St	ructuralAlerts	0.67	0.64	0.58	0.51
alvaDesc (3D bl	ocks: (only) 1-30) 3D:corina	0.75	0.71	0.57	0.68

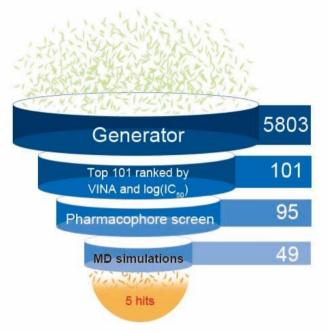
* Sparse format, DOI:10.1186/s13321-016-0113-y

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.

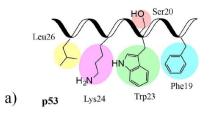
Computer aided drug design of Mdmx inhibitors

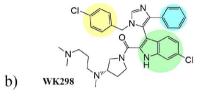


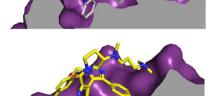
	RMSD				- <i>log</i> (IC ₅₀)	ΔG (kcal/mol)	
Compound	avg ^a	std ^b	min ^c	max ^d			
WK298	2.183	0.662	0.543	4.227	-4.7	-4.1	
3021	4.675	0.379	0.627	5.947	-5.2	-13.0	
92	1.605	0.454	0.506	3.711	-7.7	-10.8	
100	1.738	0.680	0.480	5.831	-7.9	-6.9	
34	2.789	0.696	0.668	5.176	-7.9	-6.7	
39	4.407	1.184	0.778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

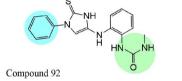
a. avg = the average; b. std = standard deviation; c. min = the minimum; d. max = the maximum

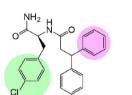
Xia, Z.; Karpov, P.; Popowicz, G.; Tetko, I.V. Focused Library Generator: case of Mdmx inhibitors. *J. Comput. Aided. Mol. Des.* **2020**, *34*, 769-782.

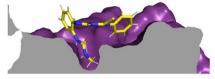


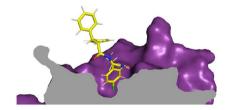


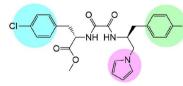








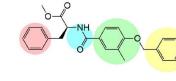


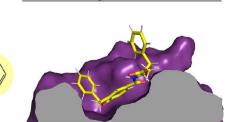




Compound 100

c)





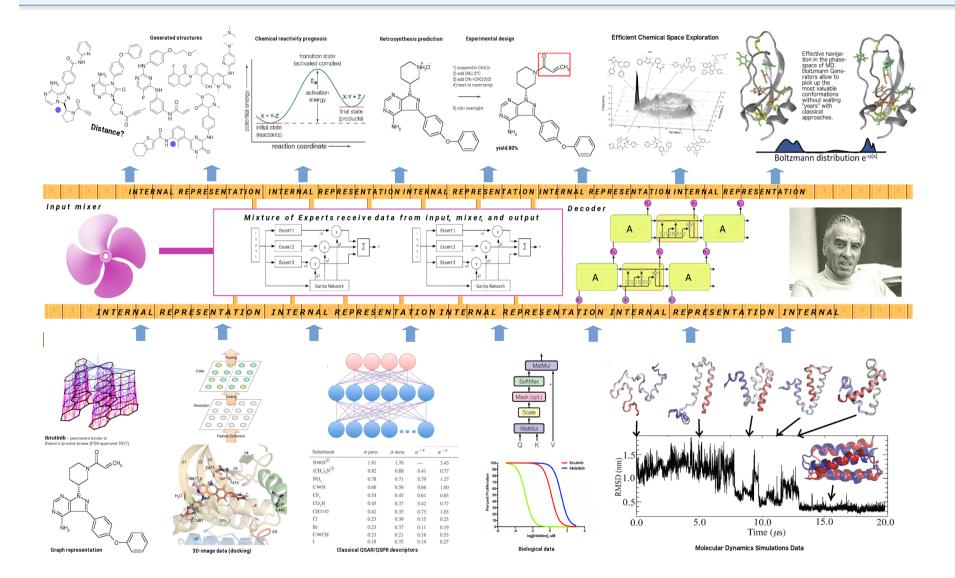
Compound 39

f)

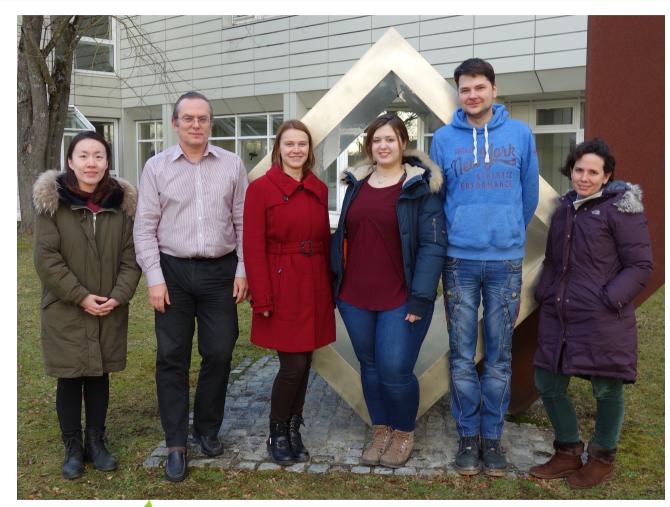


Tetko, I.V.; Karpov, P.; Van Deursen, R.; Godin, G. State-of-the-art augmented NLP transformer models for direct and single-step retrosynthesis. *Nat. Comm.* **2020**, *11*, 1-11, doi:10.1038/s41467-020-19266-y.

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













