Generative models in biomedicine

Antal Péter

Computational Biomedicine workgroup
Artificial Intelligence research group
Department of Measurement and Information Systems,
Budapest University of Technology and Economics

Overview

- Introduction
- Origins of deep belief networks (DBNs)
- Large-scale challenges for DBNs
  - Genetics of multimorbidities
  - Laboratory diagnostic tests in sequential decisions
  - Drug-multitarget interaction prediction
- Artificial creativity
  - De novo molecule generation using complex priors

Artificial intelligence and machine learning in computational biomedicine

- Knowledge engineering
- Study design
- Genetic measurements
- Data engineering
- Data analysis
- Interpretation
- Decision support

Computational Biomedicine Laboratory (ComBine.Lab): http://bioinfo.mit.bme.hu/

ComBineLab.hu: tools

- **BayesEye: Bayesian, systems-based data analysis**
  - Bayesian model averaging over causal structures.
- **BayesCube: Probabilistic decision support**
  - Semantically enriched Bayesian and decision network models.
- **BysCyc/QSF** (Bayesian Encyclopedia):
  - Large-scale quantitative, semantic data and knowledge fusion
- **QDF**: Kernel-based fusion methods for drug repositioning
  - Multi-aspect rankings and multi-aspect metrics in drug discovery
- **Variant Meta Caller**: precision NGS
  - Next-generation sequencing pipelines
- **VB-MK-LMF**: drug-target interaction prediction
  - Variational Bayesian Multiple Kernel Logistic Matrix Factorization

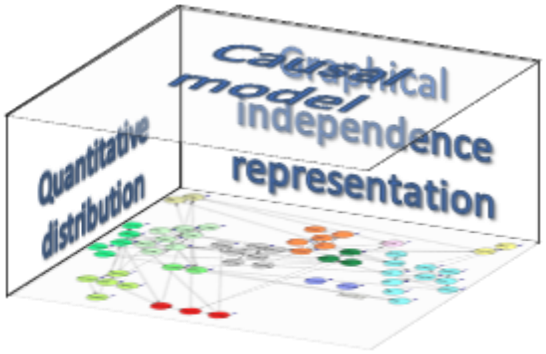
- ... see Tools @ http://bioinfo.mit.bme.hu/

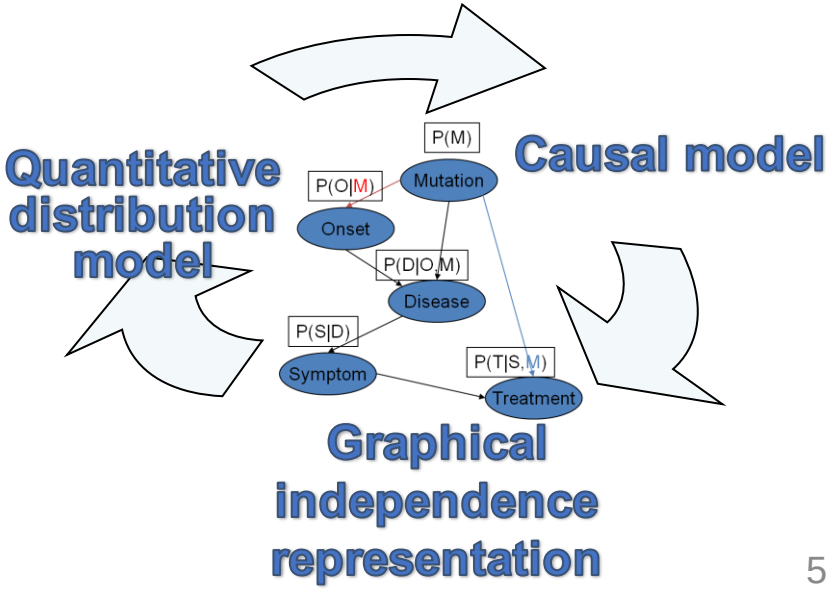
Probabilistic graphical models: Bayesian Networks

- A directed acyclic graph (DAG)
- Nodes are random variables
- Edges represent direct dependence (causal relationship)
- Local models: $P(X_i|Pa(X_i))$
- Offers three interpretations

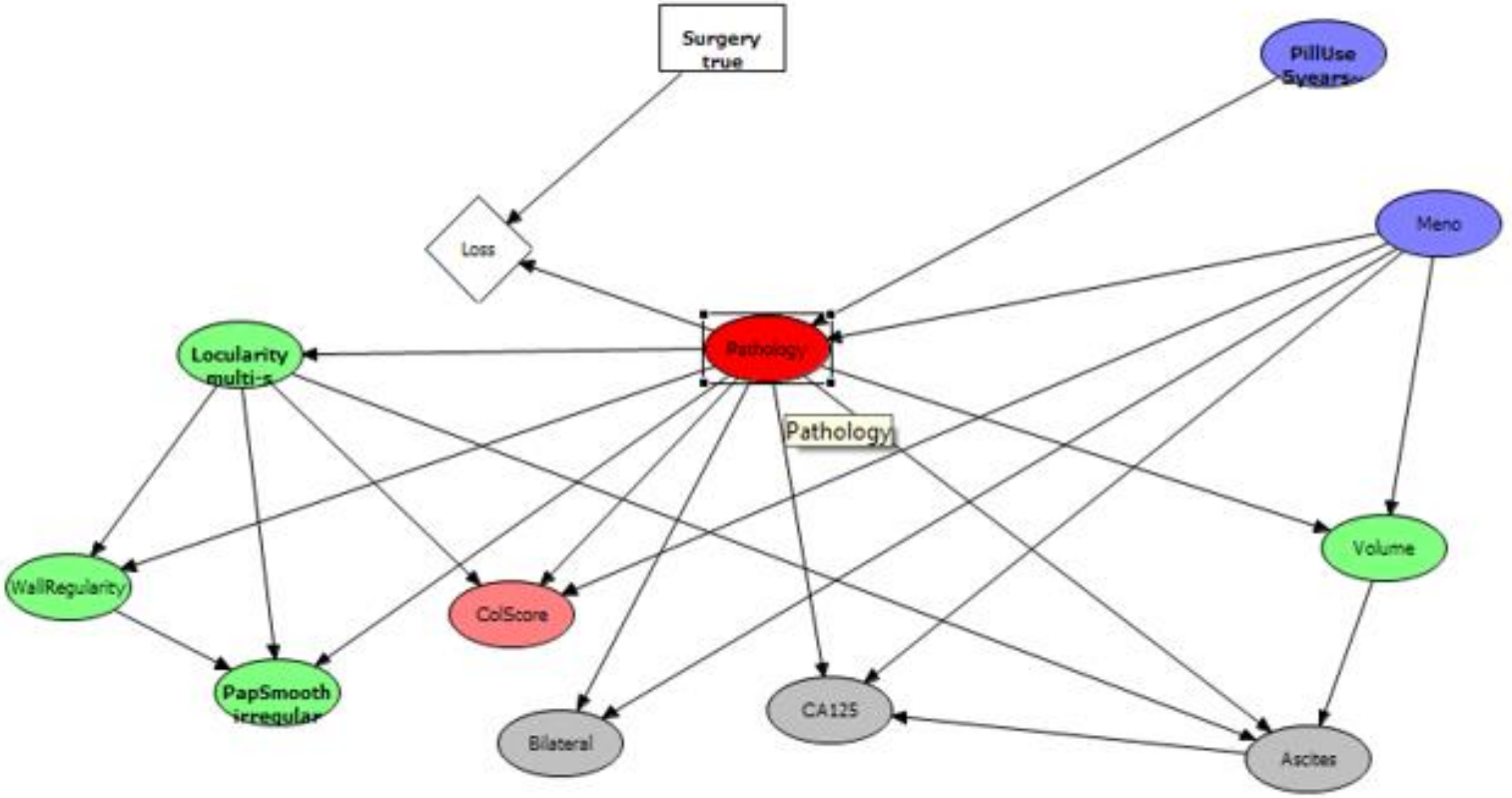
Thomas Bayes
(c. 1702 – 1761)

$$P(Model \mid Data) \propto P(Data \mid Model)P(Model)$$



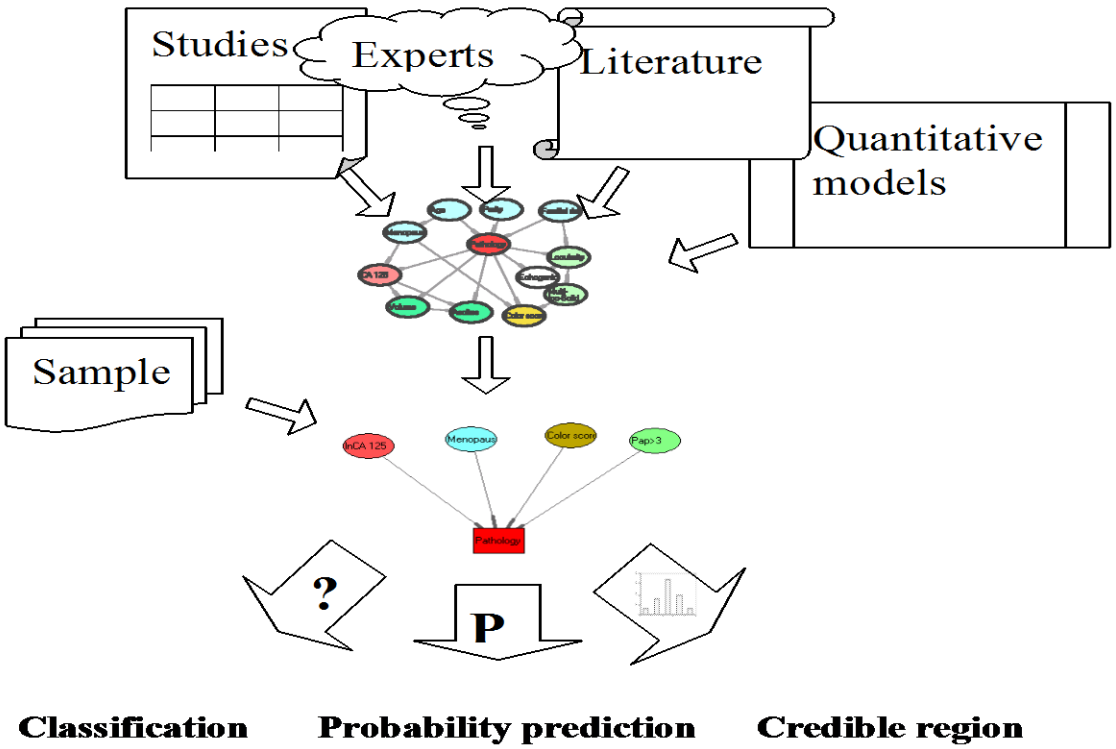


Generative models



Antal, P., Fannes, G., Timmerman, D., Moreau, Y. and De Moor, B., Using literature and data to learn Bayesian networks as clinical models of ovarian tumors. *Artificial Intelligence in medicine*, *30*(3), pp.257-281, 2004

„Informed" conditional models



Informed selection of:

- structure,
- hyperparameters,
- parameters,
- output combination,
- etc.

P. Antal, G. Fannes, D. Timmerman, Y. Moreau, B. De Moor: Bayesian Applications of Belief Networks and Multilayer Perceptrons for Ovarian Tumor Classification with Rejection, *Artificial Intelligence in Medicine*, vol. 29, pp 39-60, 2003

From deep belief networks to deep learning

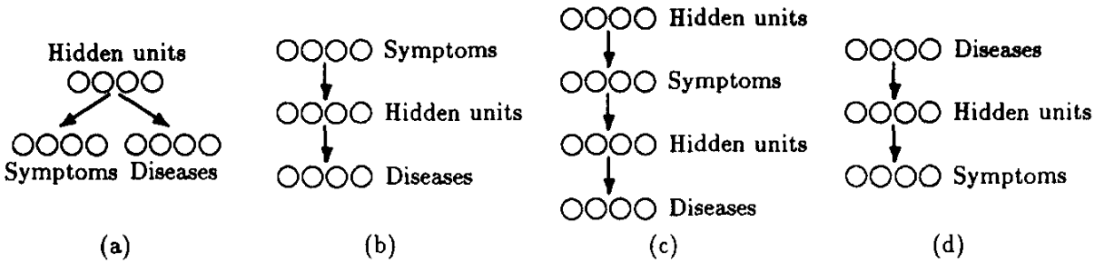


Fig. 4. Four network architectures for a medical diagnosis problem.

Neal, R.M., 1992. Connectionist learning of belief networks. *Artificial intelligence*, *56*(1), pp.71-113.

Neal, R.M. and Hinton, G.E., 1998. A view of the EM algorithm that justifies incremental, sparse, and other variants. In *Learning in graphical models* (pp. 355-368). Springer, Dordrecht.

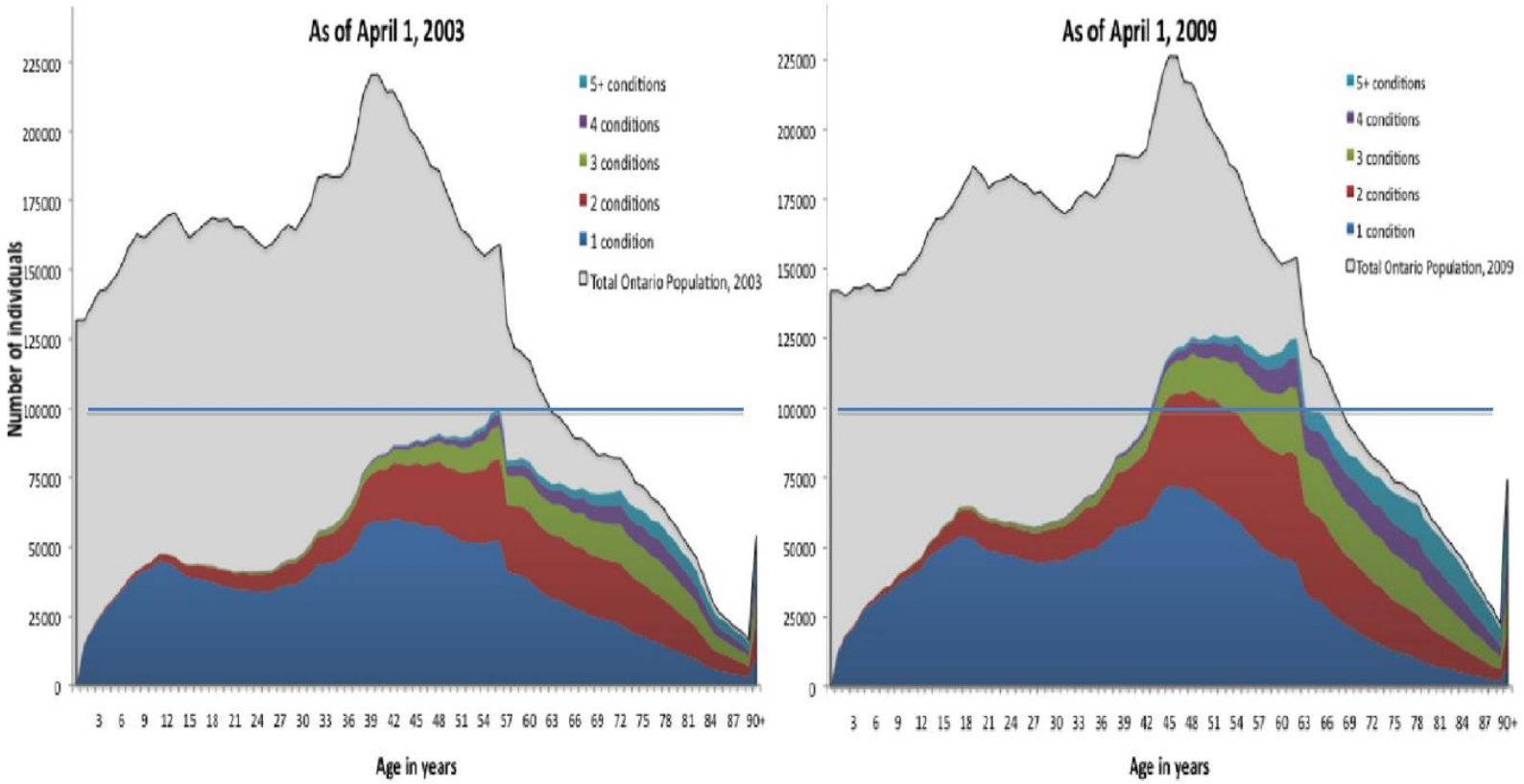
Hinton, G.E., Osindero, S. and Teh, Y.W., 2006. A fast learning algorithm for deep belief nets. *Neural computation*, *18*(7), pp.1527-1554.

LeCun, Y., Bengio, Y. and Hinton, G., 2015. Deep learning. *Nature*, *521*(7553), pp.436-444.

DBN biomed challenge (1): genetics of multimorbidities

Multimorbidity: prevalence

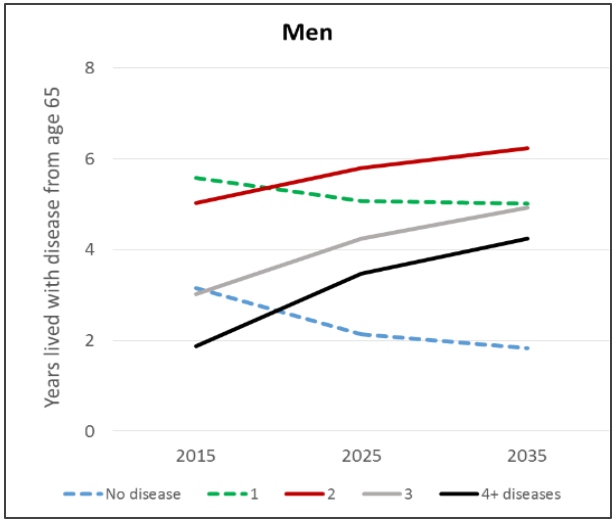
Simultaneous occurrence of multiple chronic conditions.

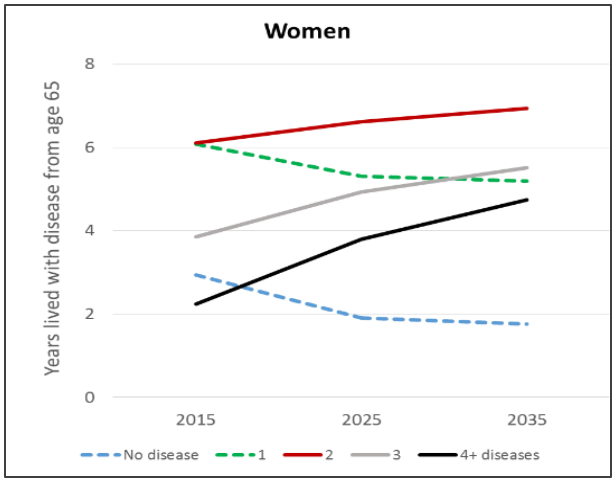


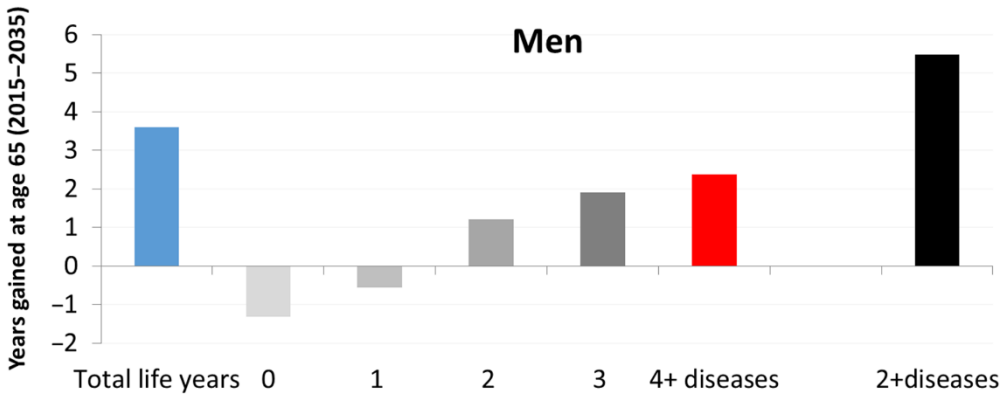
Ontario, Canada: 2003 (n = 12,242,273), 2009 (n = 13,068,845).

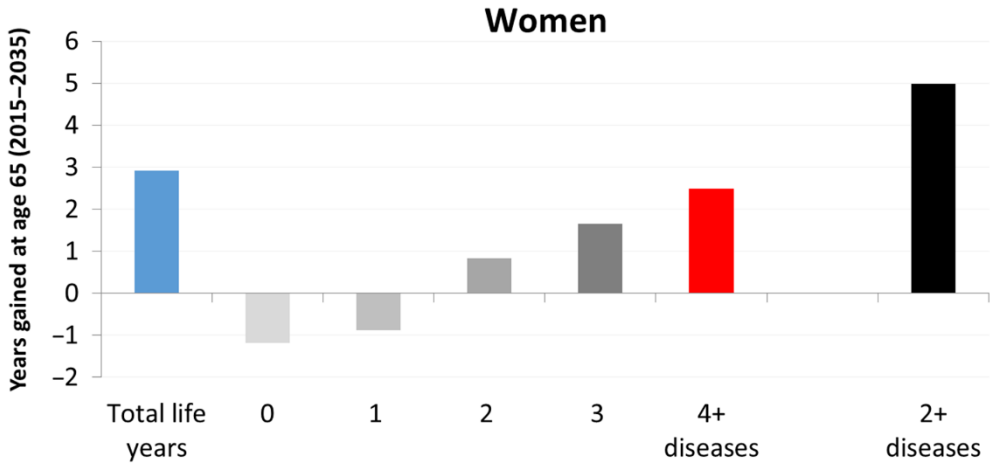
Pefoyo, Anna J. Koné, et al. "The increasing burden and complexity of multimorbidity." *BMC public health* 15.1 (2015): 415.

Multimorbidity: 2015-2035





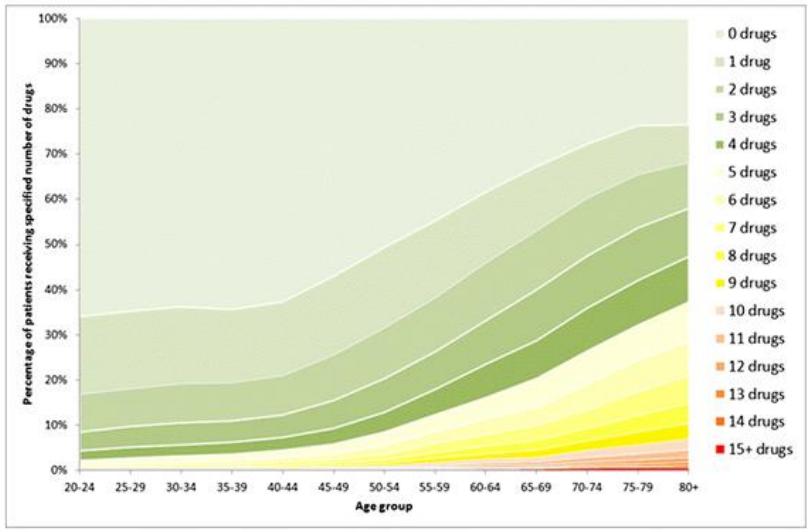




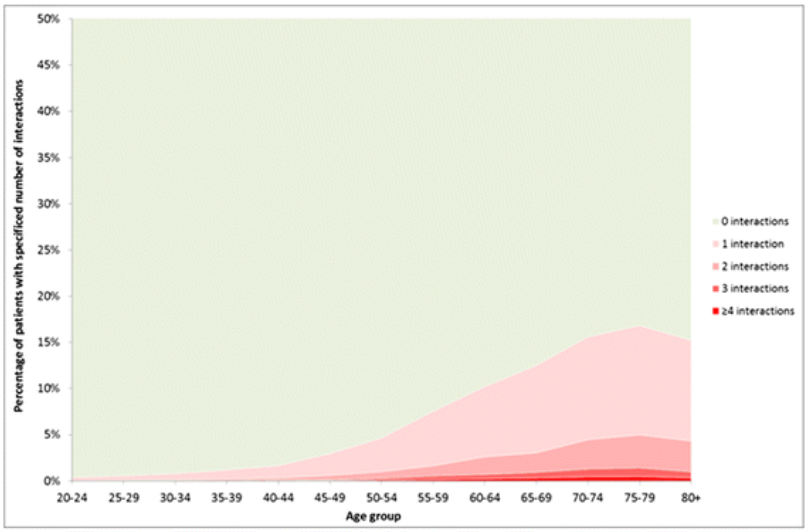
Kingston, Andrew, et al. "Projections of multi-morbidity in the older population in England to 2035: estimates from the Population Ageing and Care Simulation (PACSim) model." *Age and ageing* 47.3 (2018): 374-380.

Polipharmacy

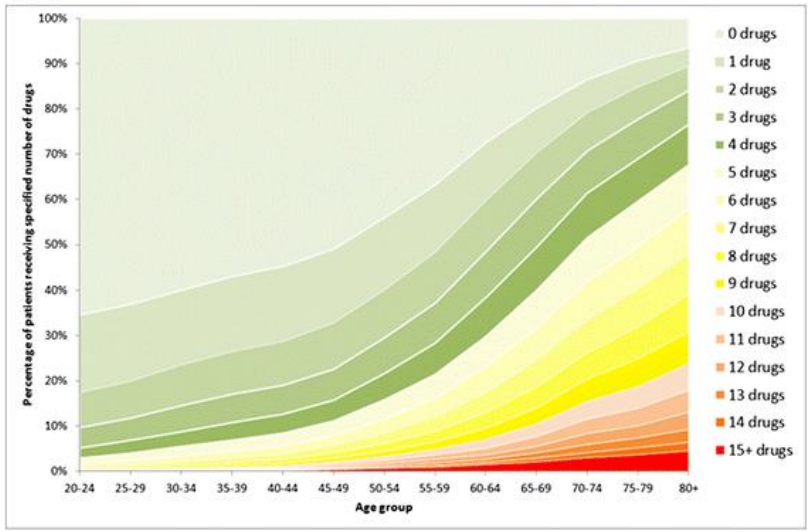
1995



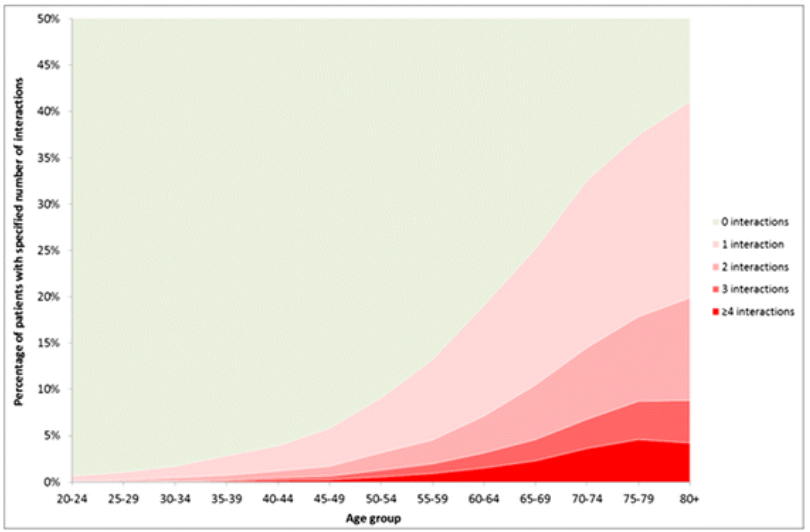
1995



2010



2010



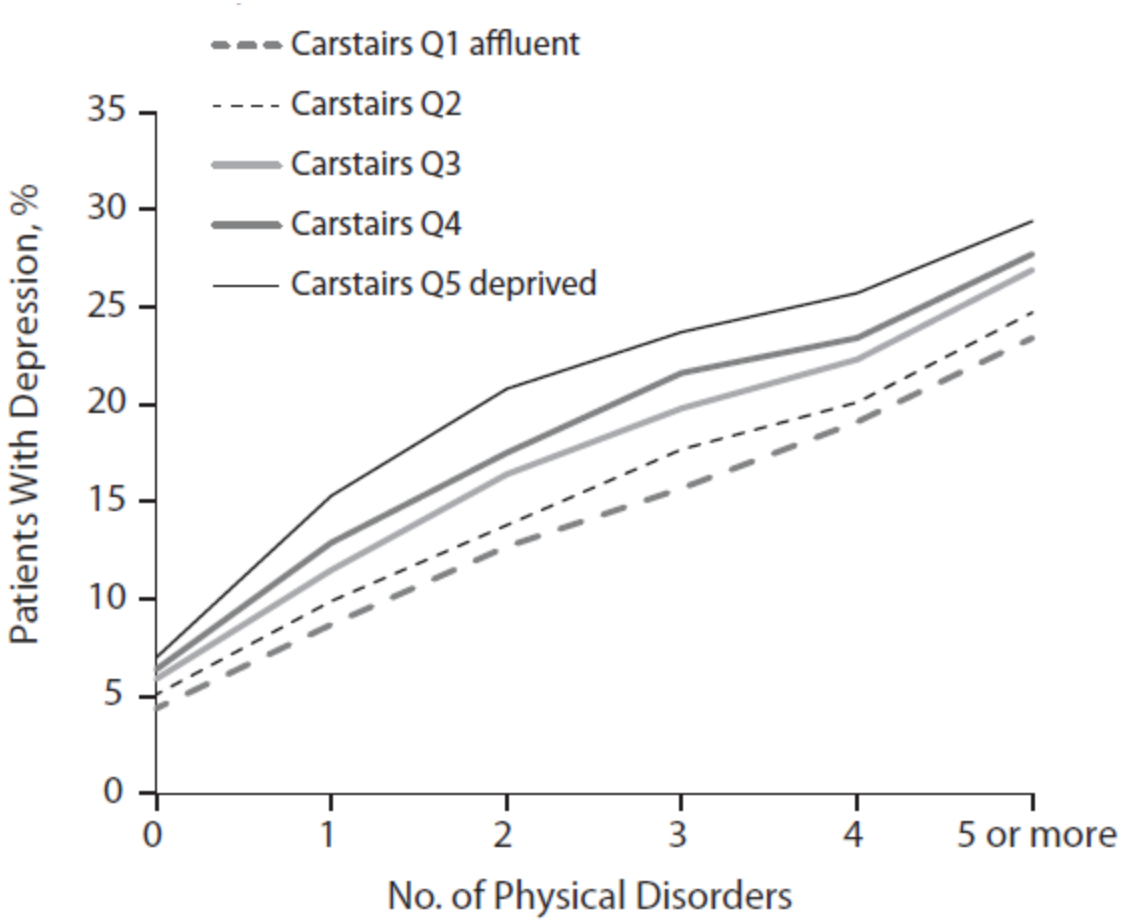
Guthrie, Bruce, et al. "The rising tide of polypharmacy and drug-drug interactions: population database analysis 1995–2010." *BMC medicine* 13.1 (2015): 74.

„Multiple chronic conditions: an emerging healthcare challenge”

- Between 2015 and 2035, the number of older people with more than two illnesses (‘multi-morbidity’) will almost double, from 5.2 million in 2015 to 9.8 million in 2035.
- Increases of more than 50% are projected in the number of older people affected by most individual diseases and impairments –the largest increases being for numbers having cancer (179.4%, or 2.2 million) and diabetes (118.1%, or 1.7 million).
- The number of older people in the population with more than four diseases (‘complex multi-morbidity’) will increase from 9.8% (952,400) in 2015 to 17.0% (2,453,200) in 2035.
- Two-thirds of those with more than four diseases will have mental ill-health (dementia, other cognitive impairment, depression) by 2035 – a total of 1.75 million people, an increase of 600,000 from 2015.

https://esrc.ukri.org/news-events-and-publications/evidence-briefings/multiple-chronic-conditions-an-emerging-healthcare-challenge/

Multimorbidity: depression



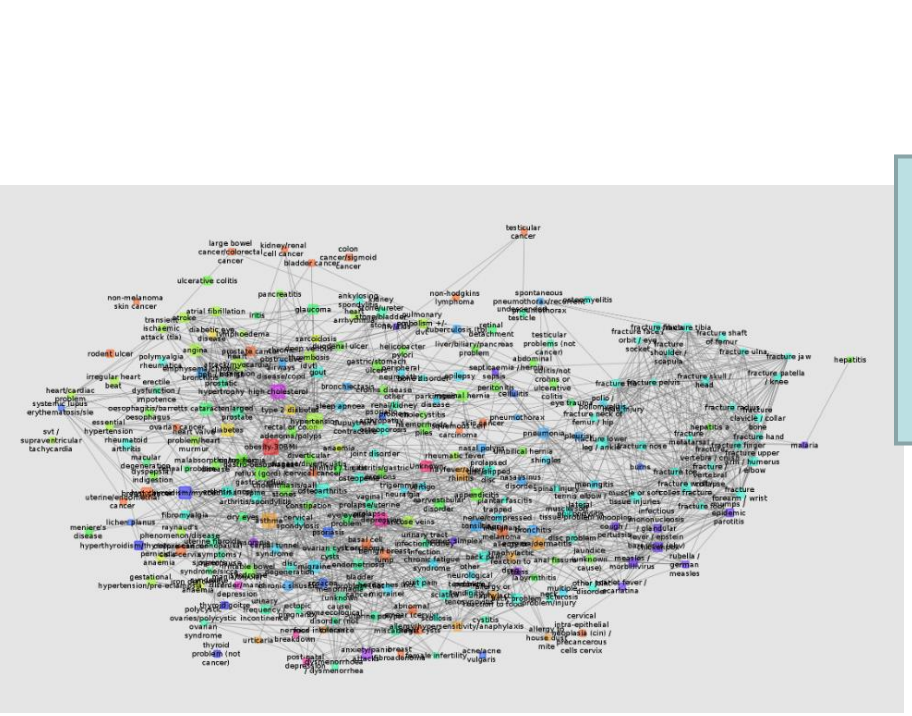
Smith, Daniel J., et al. "Depression and multimorbidity: a cross-sectional study of 1,751,841 patients in primary care." *The Journal of clinical psychiatry* 75.11 (2014): 1202-8.

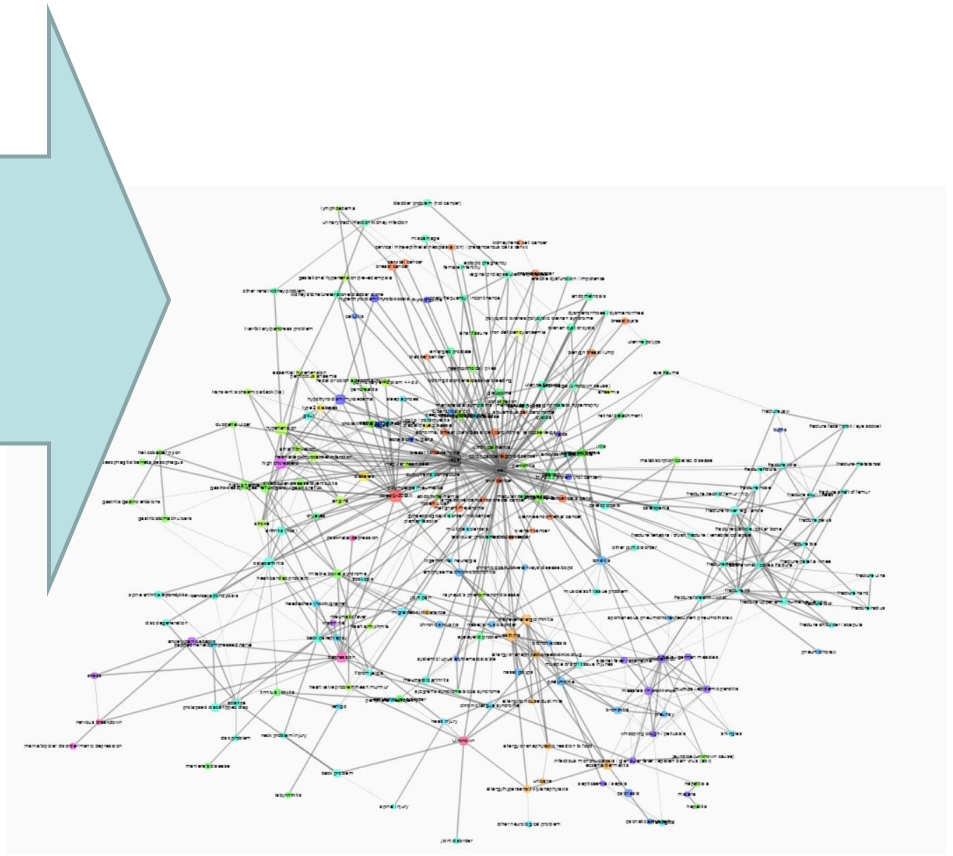
UK Biobank 2006-2010

UK Biobank is a national and international health resource with unparalleled research opportunities, open to all bona fide health researchers. …. **It is following the health and well-being of 500,000 volunteer participants and provides health information....**

Elliott, P., & Peakman, T. C. (2008). The UK Biobank sample handling and storage protocol for the collection, processing and archiving of human blood and urine. *International Journal of Epidemiology*, *37*(2), 234-244.
Collins, R. (2012). What makes UK Biobank special?. *The Lancet*, *379*(9822)

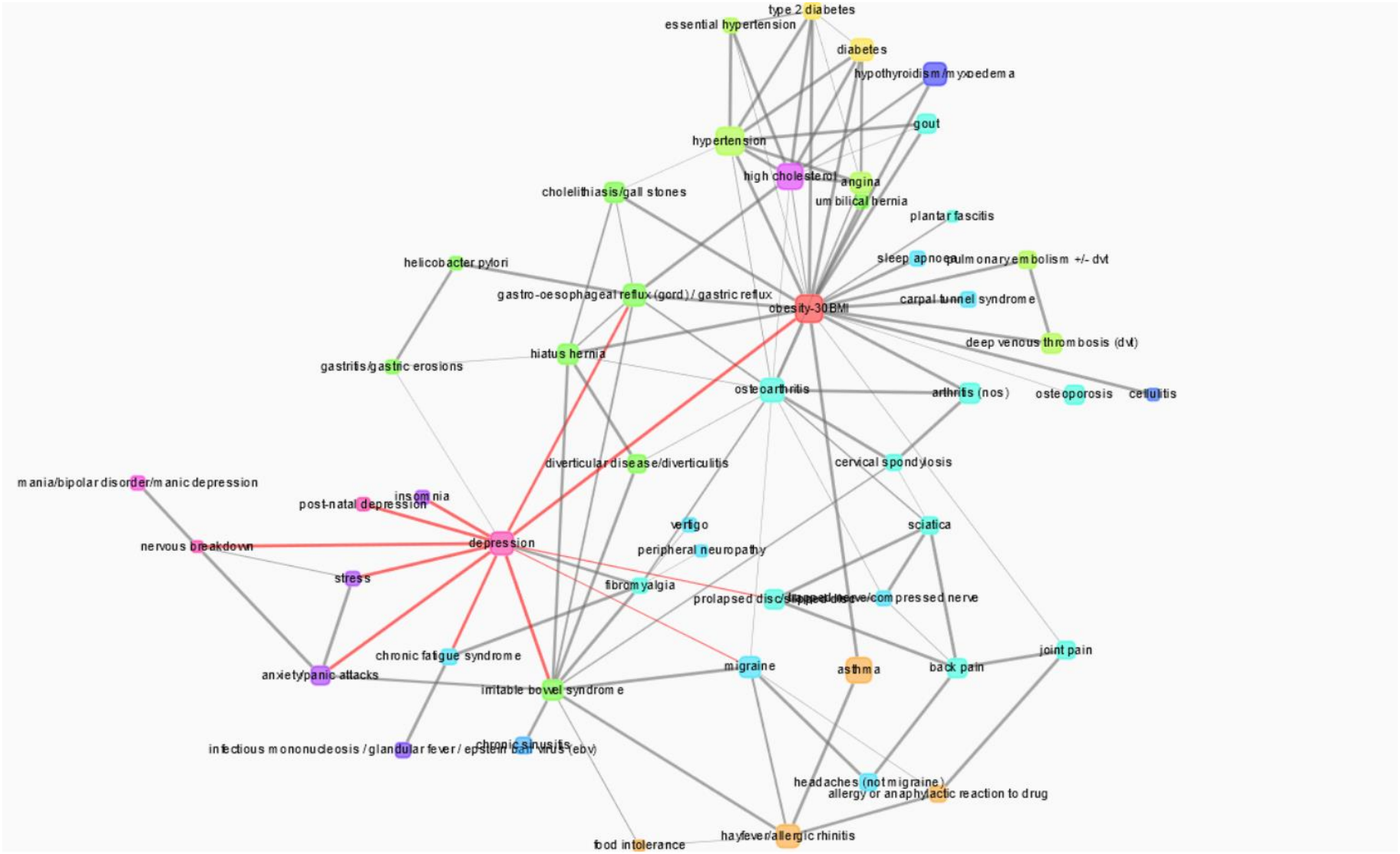
From associations to direct dependencies II. (off:80%)





Marx, P., Antal, P., Bolgar, B., Bagdy, G., Deakin, B. and Juhasz, G., 2017. Comorbidities in the diseasome are more apparent than real. *PLoS computational biology*, *13*(6), p.e1005487.

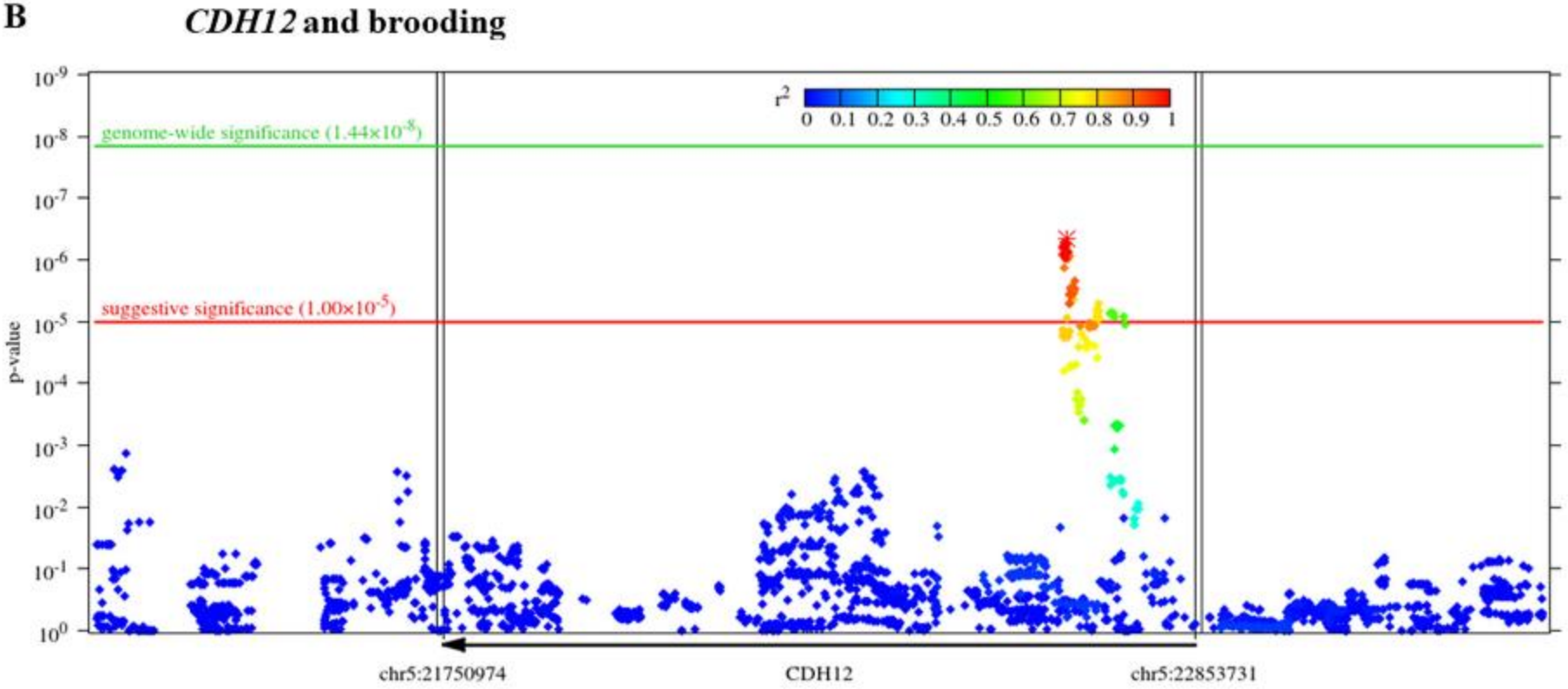
Depression multimorbidity cluster



http://bioinformatics.mit.bme.hu/UKBNetworks/full/index.html#/

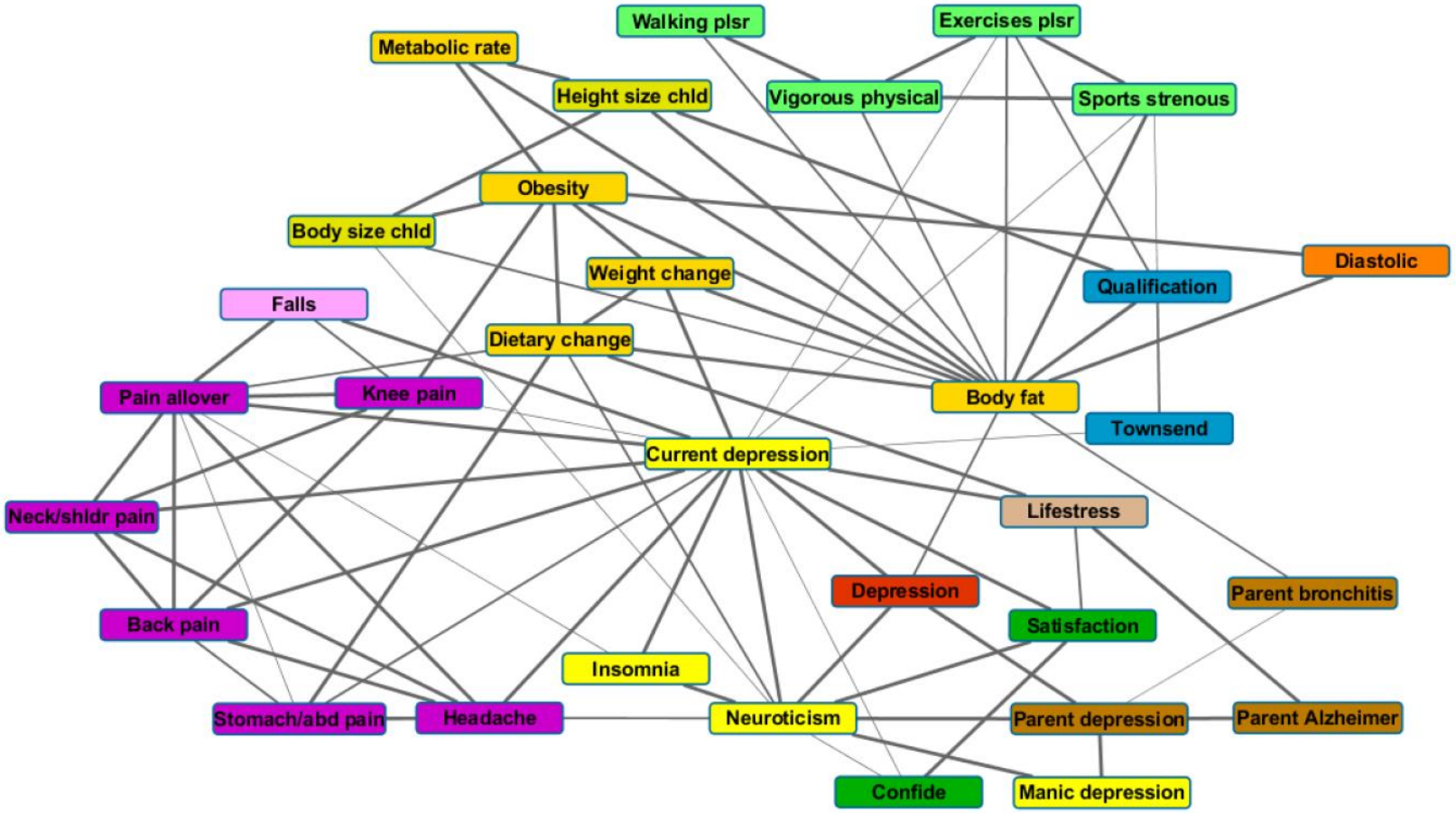
Genetics of depression multimorbidities

Brooding



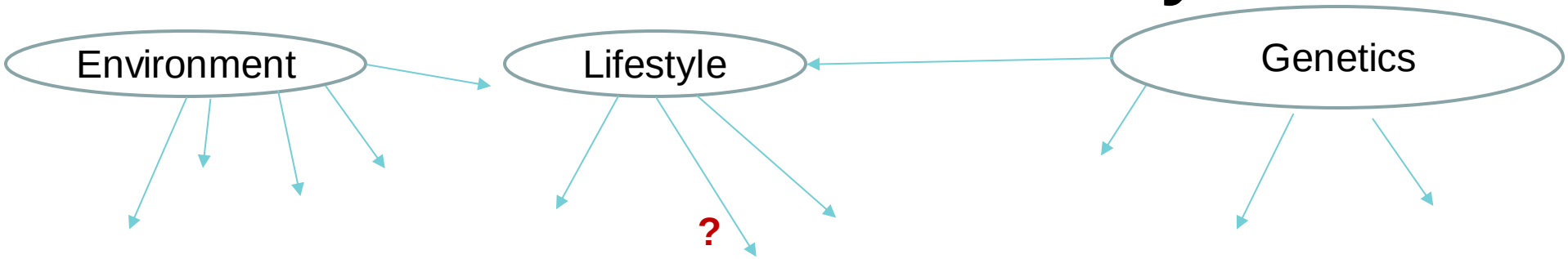
Nora Eszlari Andras Millinghoffer, Peter Petschner, Xenia Gonda, Daniel Baksa, Attila J. Pulay, Janos Rethelyi, John Francis William Deakin, Peter Antal, Gyorgy Bagdy, Gabriella Juhasz, Genome-wide association analysis reveals KCTD12 and miR-383-binding genes in the background of rumination, Translational Psychiatry (9: 119), 2019

Envirome - life style - depression

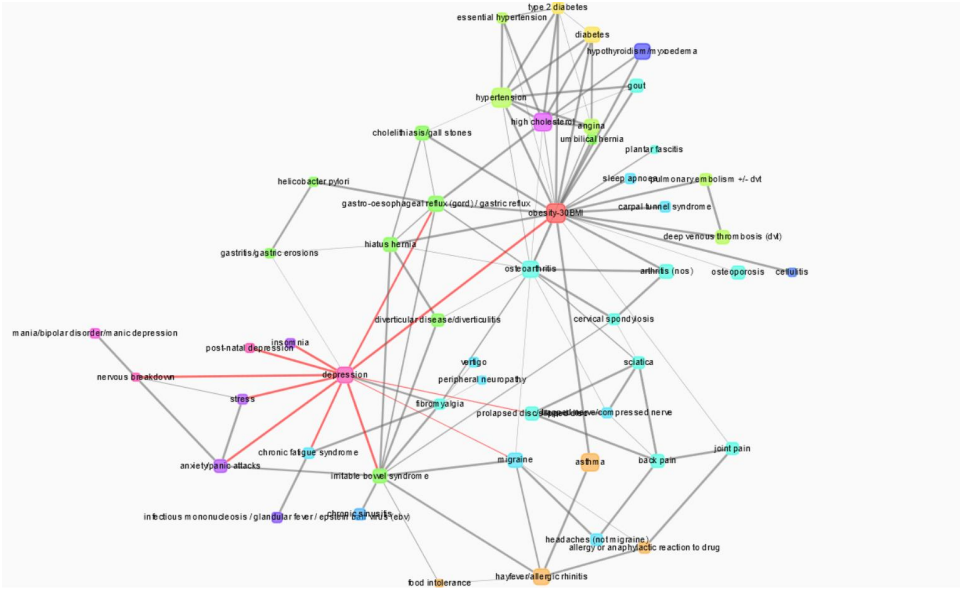


Hullam, G., Antal, P., Petschner, P., Gonda, X., Bagdy, G., Deakin, B. and Juhasz, G., 2019. The UKB envirome of depression: from interactions to synergistic effects. *Scientific reports*, *9*(1), pp.1-19.

GxExLS ⇨ multimorbidity cluster



[Deep] conditional generative model



Bruncsics, B. and Antal, P., 2019, July. A multi-trait evaluation of network propagation for GWAS results. In *2019 IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB)* (pp. 1-6). IEEE.

DBN biomed challenge (2): clinical laboratory parameters of multimorbidities

Laboratory testing

Large-scale laboratory test data sets are untapped resources, but they are complex:

- incomplete,
- continuous, but with establised reference thresholds,
- longitudinal,
- heterogeneous medical scenarios:
  - **Prevention**: recognition of a pre-disease state.
  - **Screening**: early diagnosis of a given disease.
  - **Exploratory diagnostics**:, inference of cause(s).
  - **Differential diagnostics**: select most plausible explanation.
  - **Monitoring:** track effect of intervention.

Hypothesis: Laboratory tests have a complex, rich dependency structure.

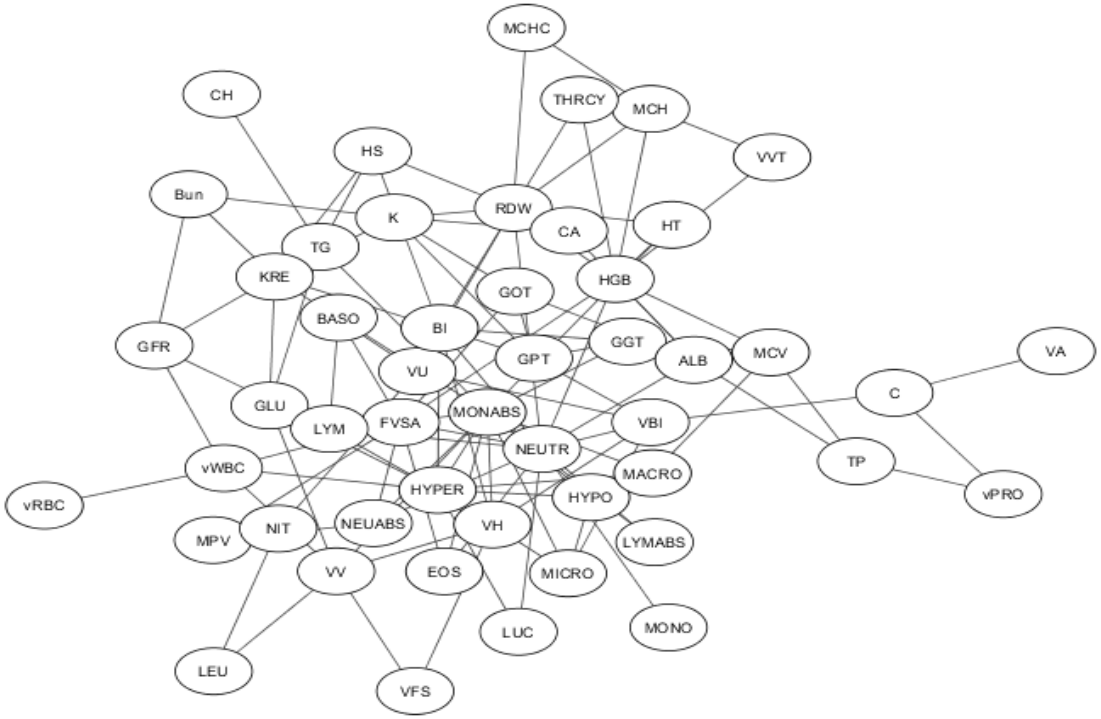
Goals & Data

In cooperation with the Central Laboratory of Semmelweis University:

- **Prune requested tests**: Predict that certain requested tests are confidently predictable based on earlier measurements from the patient's history and from current measurements.
- **Extend requested tests**: Predict that the value of certain not requested tests are abnormal with high confidence.
- **Data set:**
  - Patients: 202,976
  - Laboratory tests: 2078
  - Visits (~orders): 1,376,758
  - Valid tests: 37,354,817 (now: 70 million)

Bayesian map of laboratory tests

We estimated the a posteriori probabilities of edges using a DAG-based Markov Chain Monte Carlo simulation.



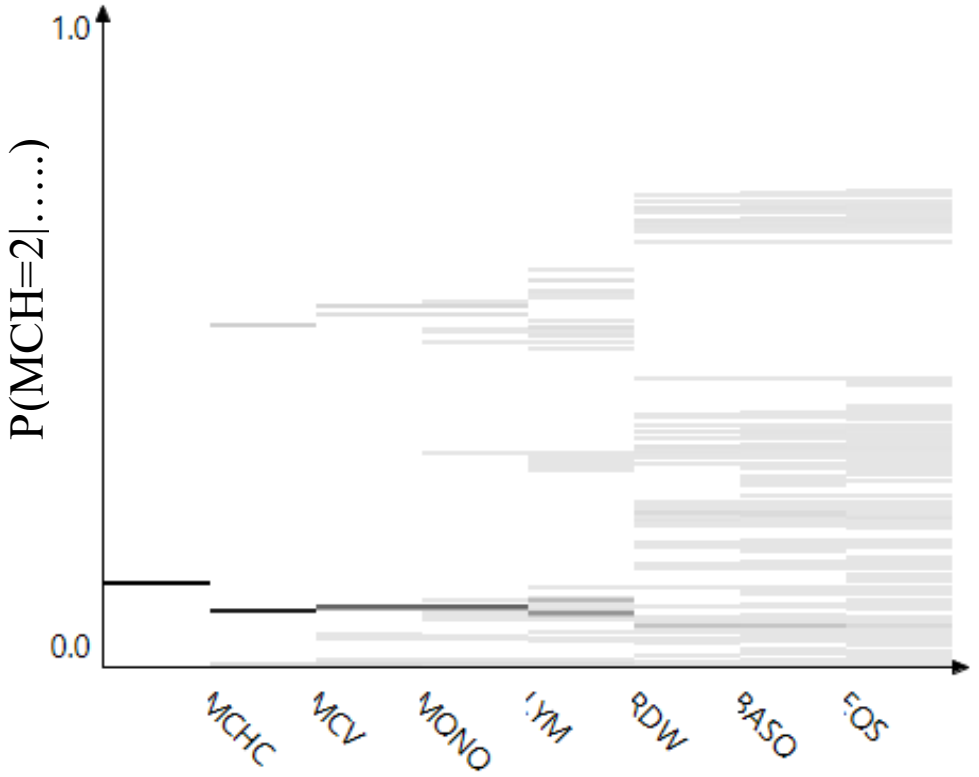
The map of edges with posteriors between [0.75-1.0].

Guenfoud, Z. and Antal, P., 2018, October. Bayesian exploration of dependencies of laboratory tests and evaluation of test redundancy. In *2018 3rd International Conference on Pattern Analysis and Intelligent Systems (PAIS)* (pp. 1-6). IEEE.

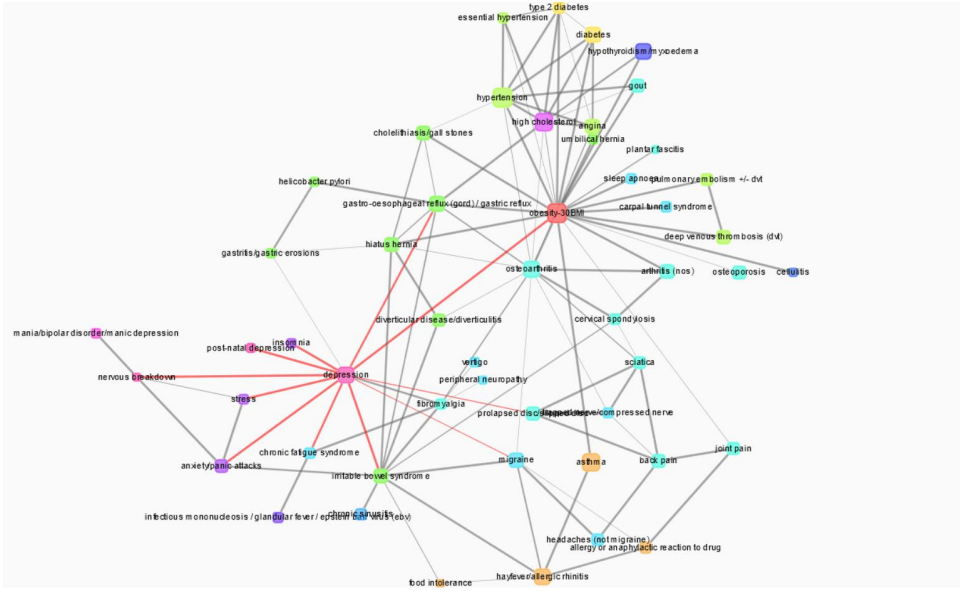
Prediction of laboratory tests

- In silico/virtual laboratory test is based on the calculation of the conditional probability distribution of a laboratory test given the outcome(s) of other test(s).

The conditional probability of an abnormal Mean Corpuscular Hemoglobin



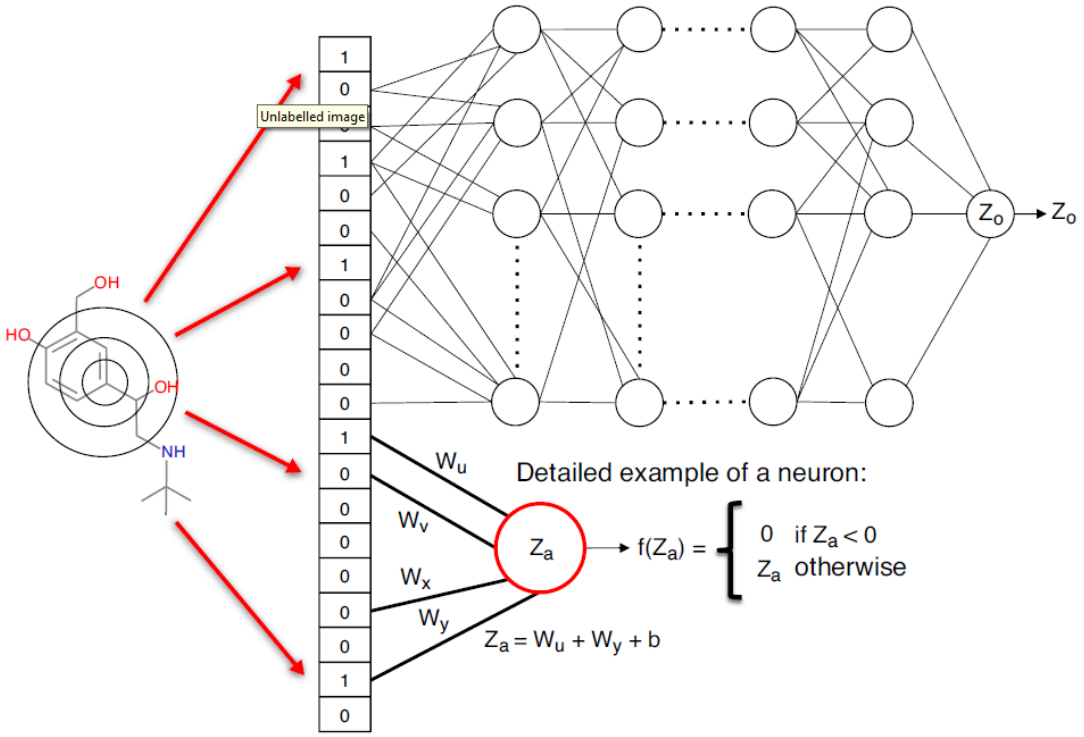
Multimorbidities ⇨ laboratory tests



[Deep] conditional generative model

DBN biomed challenge (3): drug-target interaction prediction

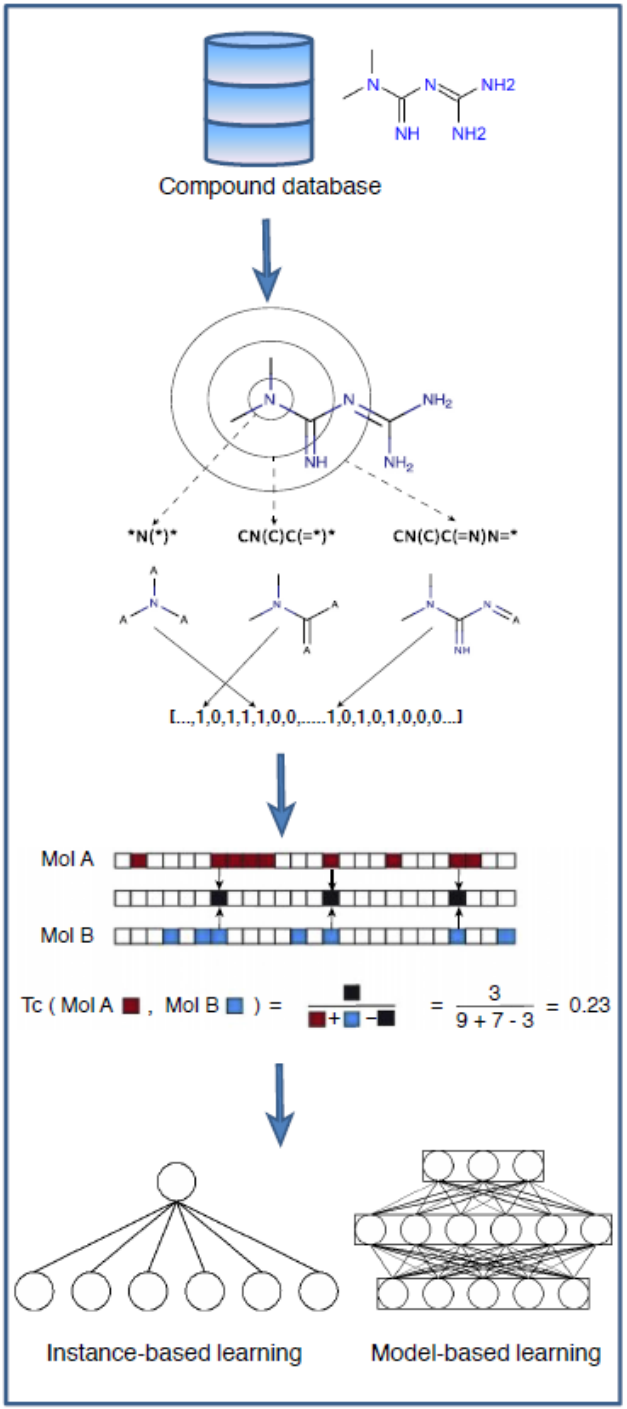
Drug-target interaction prediction



Kövesdi, I., Dominguez-Rodriguez, M.F., Ôrfi, L., Náray-Szabó, G., Varró, A., Papp, J.G. and Mátyus, P., 1999. Application of neural networks in structure–activity relationships. *Medicinal research reviews*, *19*(3), pp.249-269.

Colwell, L.J., 2018. Statistical and machine learning approaches to predicting protein–ligand interactions. *Current opinion in structural biology*, *49*, pp.123-128.

Machine learning in chemoinformatics



Chemical feature extraction
Compounds retrieved from database were characterized by the chemical substructure fragments or other methods

Chemical fingerprint creation
The presence and absence of particular substructure fragments were used to create a chemical fingerprint for similarity comparison

QSAR/QSPR modeling
Given known compound properties, the chemical features can be used to train machine learning models (instance-based or model-based) for compound property predictions

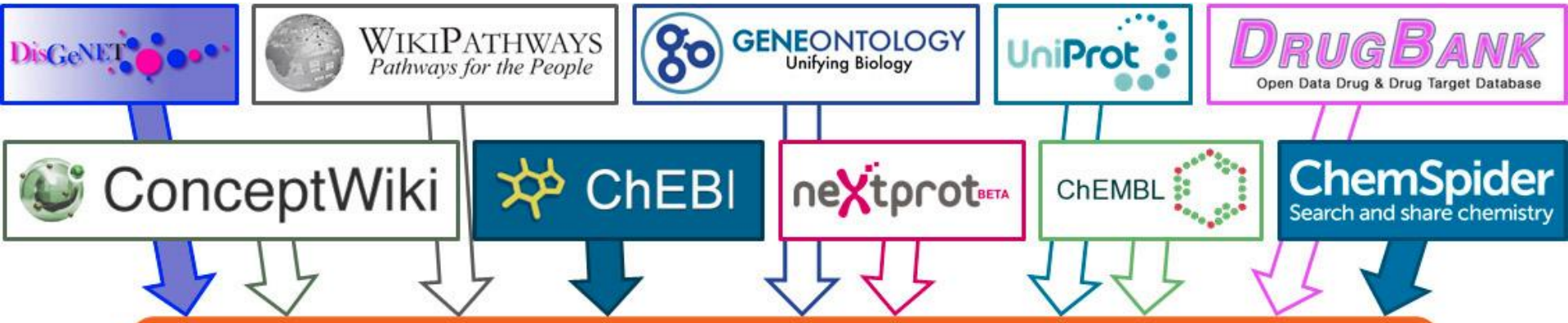
Lo, Y.C., Rensi, S.E., Torng, W. and Altman, R.B., 2018. Machine learning in chemoinformatics and drug discovery. *Drug discovery today*.

Open Pharmacological Space

Precursor: Gene Ontology: tool for the unification of biology, Nature, 2000

- Discovery Platform for **cross-domain** fusion.
- Public, curated, linked data.
  - The data sources you **already use**, **integrated** and **linked** together: *compounds*, *targets*, *pathways*, *diseases* and *tissues*.
- Everything in **triples**: **Subject-predicate-object**

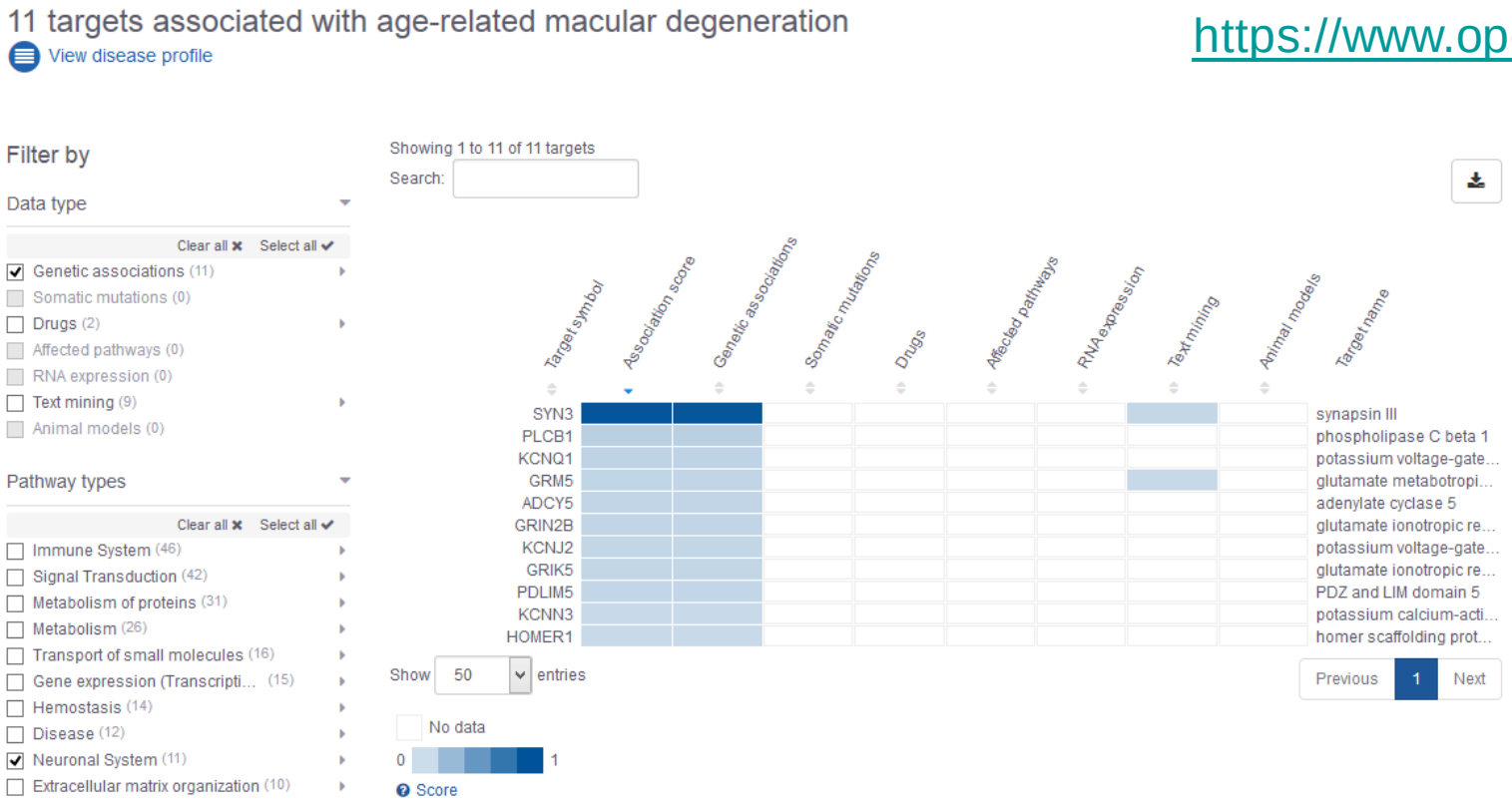
Open PHACTS



Open PHACTS

Physicochemical data	Identifiers	Pharmacological data
Molecular weight & formula	Synonyms	Activity type, value, concentration
H-Bond acceptors / donors	SMILES	Assay description
Polar surface area, AlogP	InChI / InChIkey	Target organism
Melting point	ChemSpider ID	Target name
...and more	...and more	...and more

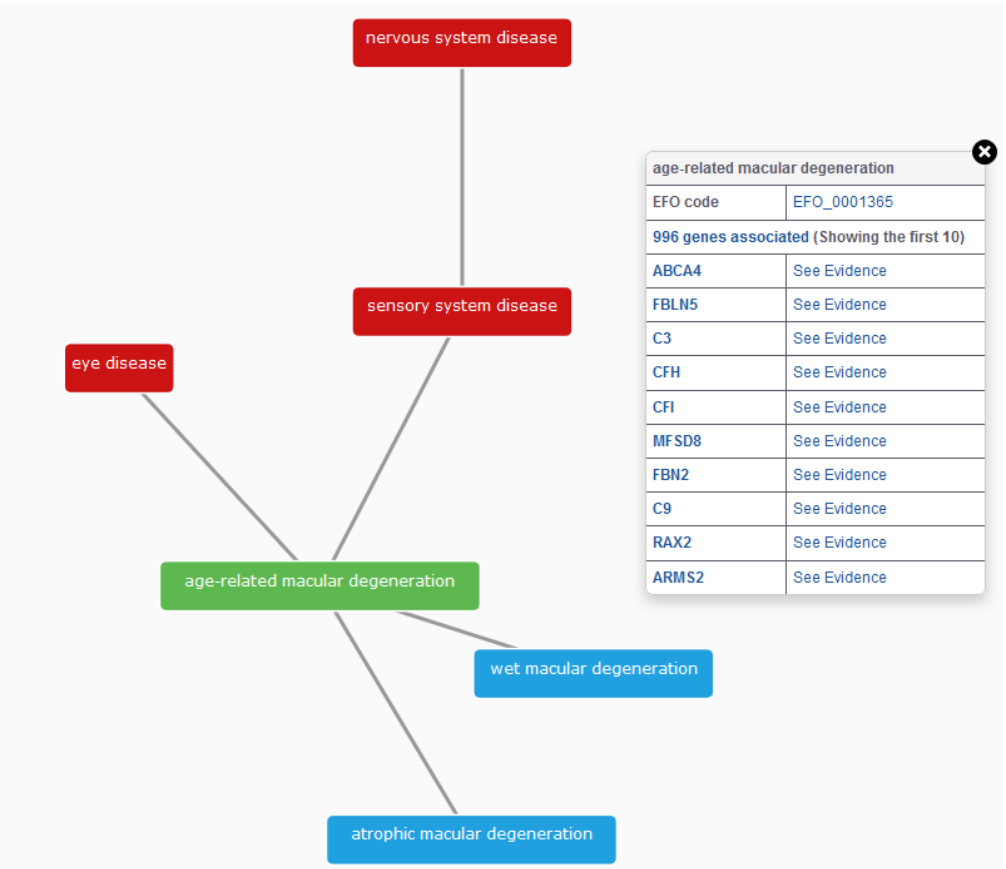
Open Targets I.



Khaladkar, M., Koscielny, G., Hasan, S., Agarwal, P., Dunham, I., Rajpal, D. and Sanseau, P., 2017. Uncovering novel repositioning opportunities using the Open Targets platform. *Drug discovery today*.

Koscielny, G., An, P., Carvalho-Silva, D., Cham, J.A., Fumis, L., Gasparyan, R., Hasan, S., Karamanis, N., Maguire, M., Papa, E. and Pierleoni, A., 2016. Open Targets: a platform for therapeutic target identification and validation. *Nucleic acids research*, *45*(D1), pp.D985-D994.

Open Targets II.



age-related macular degeneration	
EFO code	EFO_0001365
996 genes associated (Showing the first 10)	
ABCA4	See Evidence
FBLN5	See Evidence
C3	See Evidence
CFH	See Evidence
CFI	See Evidence
MFSD8	See Evidence
FBN2	See Evidence
C9	See Evidence
RAX2	See Evidence
ARMS2	See Evidence

Evidence for SYN3 in age-related macular degeneration

SYN3
synapsin III

May be involved in the regulation of neurotransmitter release and synaptogenesis.

Genetic associations

Somatic mutations

Drugs

Affected pathways

RNA expression

Text mining

Source: Europe PMC

Shown are the 3 articles where target and disease are found in the same sentence.

Bioactivity databases I.

ChEMBL is a database of bioactive drug-like small molecules, it contains 2-D structures, calculated properties (e.g. logP, Molecular Weight, Lipinski Parameters, etc.) and abstracted bioactivities (e.g. binding constants, pharmacology and ADMET data).

https://www.ebi.ac.uk/chembl

- Targets: 10,774
- Compound records: 1,715,667
- Distinct compounds: 1,463,270
- Activities: 13,520,737
- Publications: 59,610

PubChem

Bioactivity databases II.

Compounds:	97,127,348
Substances:	252,300,917
BioAssays:	1,067,565
Tested Compounds:	3,417,415
Tested Substances:	5,591,261
RNAi BioAssays:	173
BioActivities:	239,680,570
Protein Targets:	12,159
Gene Targets:	58,186

Bioactivity databases III:ExCAPE-DB

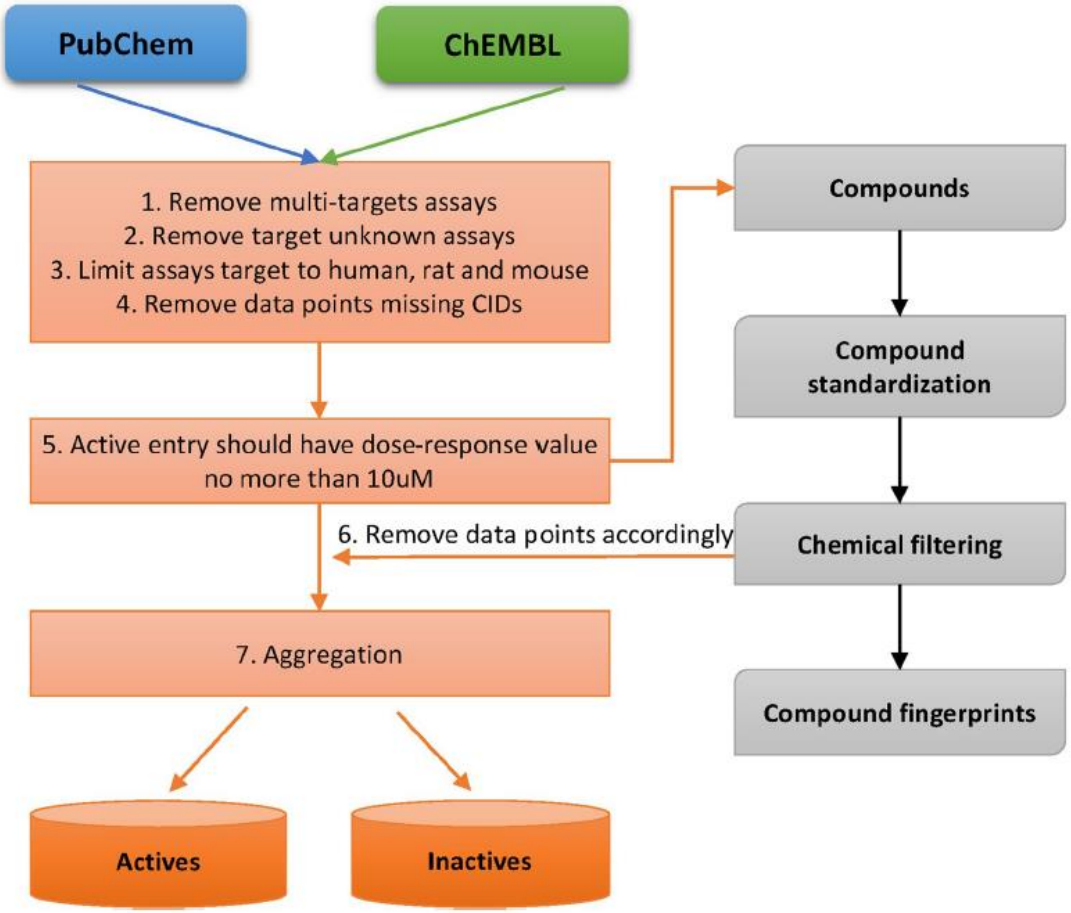


Table 1 Public chemogenomics dataset

	ChEMBL	PubChem	ExCAPE-DB
Actives			
# SAR data points	1,259,338	439,288	1,332,426
# Compounds	566,143	263,119	593,156
Inactives			
# SAR data points	1,530,908	68,948,609	69,517,737
# Compounds	416,655	654,562	719,192
Total			
# SAR data points	2,790,246	69,387,897	70,850,163
# Compounds	710,324	828,317	998,131
# Targets	1644	1588	1667

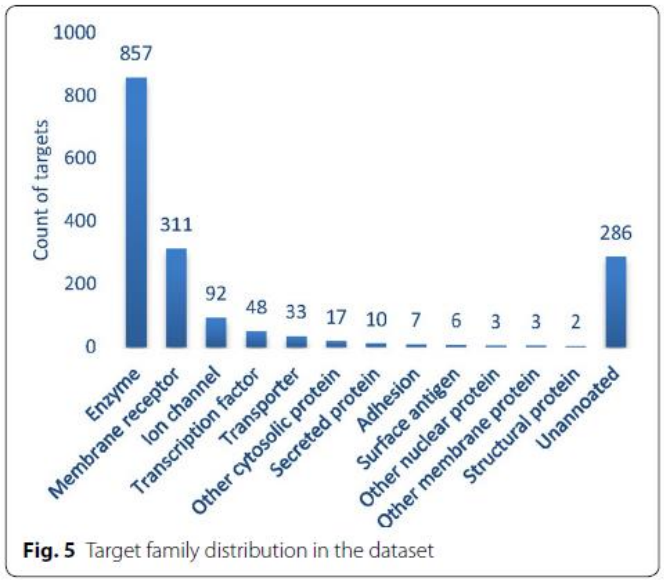


Fig. 5 Target family distribution in the dataset

Sun, J., Jeliazkova, N., Chupakhin, V., Golib-Dzib, J.F., Engkvist, O., Carlsson, L., Wegner, J., Ceulemans, H., Georgiev, I., Jeliazkov, V. and Kochev, N., 2017. ExCAPE-DB: an integrated large scale dataset facilitating Big Data analysis in chemogenomics. *Journal of cheminformatics*, *9*(1), p.17.

Drug-target interaction prediction I.

- Drug/compound information
  - Fingerprints, pharmacophore properties, etc.
  - Similarities
- Target information
  - Protein vs. binding site/pocket
  - Sequence/../complete structure
  - Similarities
- Interaction data
  - Indirect/direct
  - Binary/rank/scalar
  - IC50, Ki,..
  - Complete/incomplete

Drug-target interaction prediction II.

- Goal
  - New drugs for a given target
  - New targets for a given compound
  - Multitask learning
    - Targets for a novel drug
    - Drugs for a novel target
    - Interaction between novel drugs and targets.
- (Sequentiality)

A benchmark DTI task

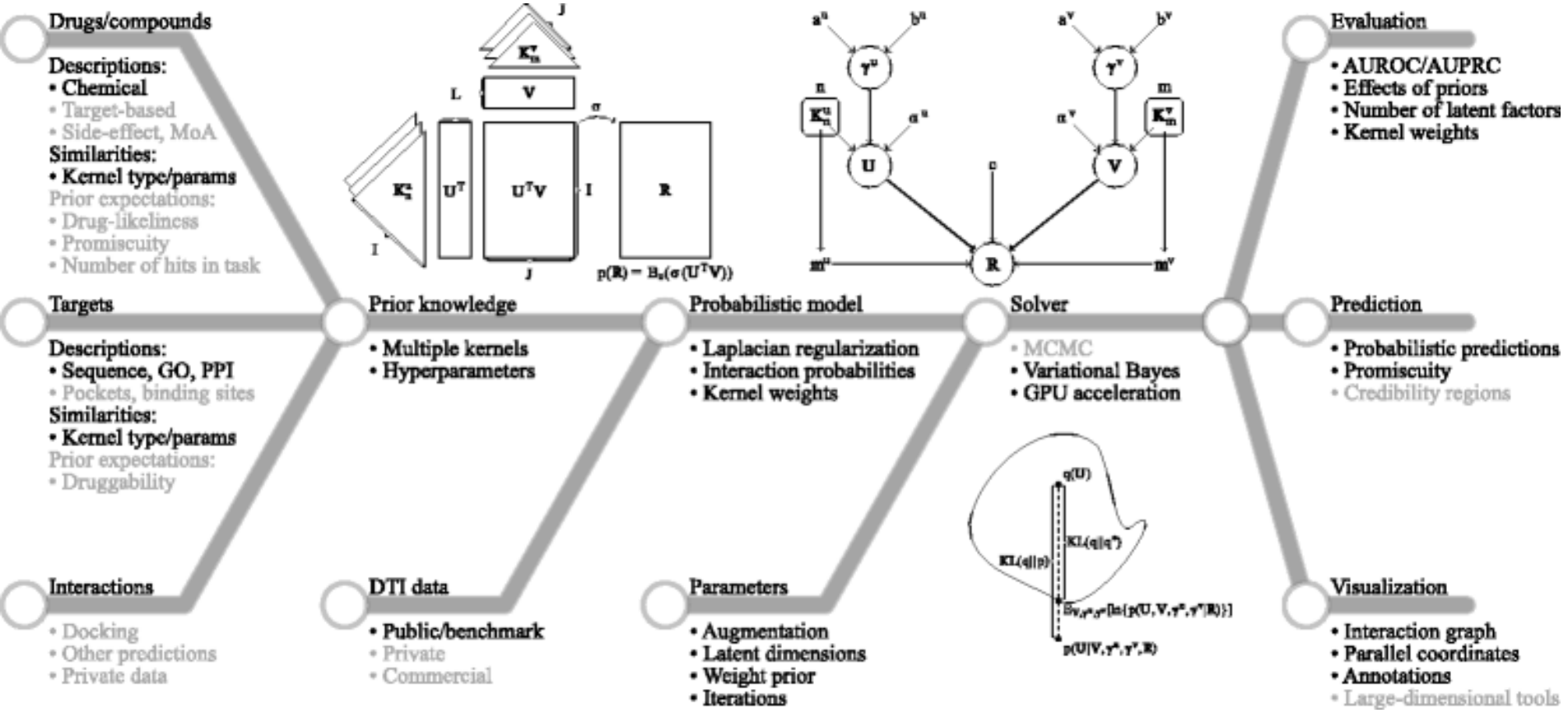
Statistics	Enzyme	Ion channel	GPCR	Nuclear receptor
No. of drugs	445	210	223	54
No. of target proteins	664	204	95	26
(Total in human genome)	(2741)	(292)	(757)	(49)
No. of drug–target interactions	2926	1476	635	90
Average degree of drugs	6.57	7.02	2.84	1.66
Average degree of targets	4.40	7.23	6.68	3.46

Yamanishi Y, Araki M, Gutteridge A, Honda W, Kanehisa M. Prediction of drug-target interaction networks from the integration of chemical and genomic spaces. Bioinformatics. 2008; 24(13):232–40. doi:10.1093/bioinformatics/btn162.

Multitask DTI prediction

- Approaches
  - Network methods
  - ….
  - **Pairwise conditional approaches or pairwise kernel methods**
  - **Matrix factorization methods**

Fusion of drugs, targets and interactions

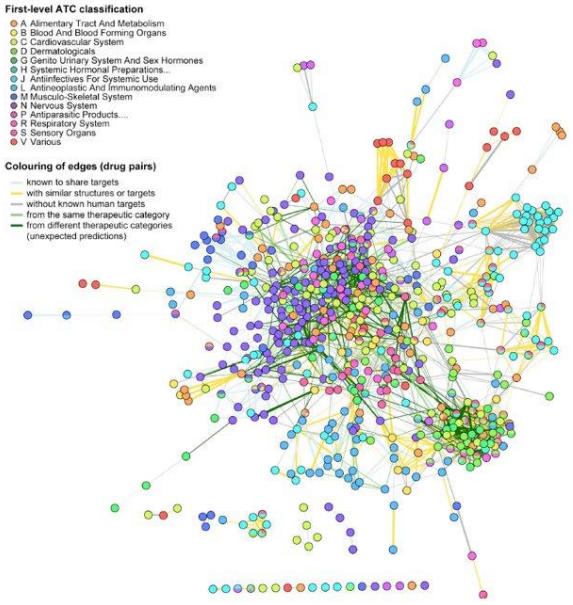


Bolgár, Bence, and Péter Antal. "VB-MK-LMF: fusion of drugs, targets and interactions using variational Bayesian multiple kernel logistic matrix factorization." *BMC Bioinformatics* 18.1 (2017): 440.

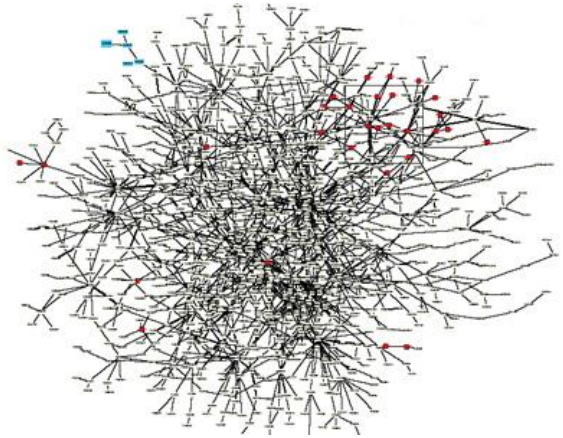
DBNs in DTI

- Wang, Y. and Zeng, J., 2013. Predicting drug-target interactions using restricted Boltzmann machines. *Bioinformatics*, *29*(13), pp.i126-i134.
- Liang, M., Li, Z., Chen, T. and Zeng, J., 2014. Integrative data analysis of multi-platform cancer data with a multimodal deep learning approach. *IEEE/ACM transactions on computational biology and bioinformatics*, *12*(4), pp.928-937.
- Sridhar, D., Fakhraei, S. and Getoor, L., 2016. A probabilistic approach for collective similarity-based drug–drug interaction prediction. *Bioinformatics*, *32*(20), pp.3175-3182.
- Hamanaka, M., Taneishi, K., Iwata, H., Ye, J., Pei, J., Hou, J. and Okuno, Y., 2017. CGBVS-DNN: Prediction of Compound-protein Interactions Based on Deep Learning. *Molecular informatics*, *36*(1-2), p.1600045.
- Ghasemi, F., Mehridehnavi, A., Fassihi, A. and Pérez-Sánchez, H., 2018. Deep neural network in QSAR studies using deep belief network. *Applied Soft Computing*, *62*, pp.251-258.
- Lee, I., Keum, J. and Nam, H., 2019. DeepConv-DTI: Prediction of drug-target interactions via deep learning with convolution on protein sequences. *PLoS computational biology*, *15*(6), p.e1007129.

Fingerprints ⇨ target activities



[Deep] conditional generative model



DBN biomed challenge (+): de novo molecule generation

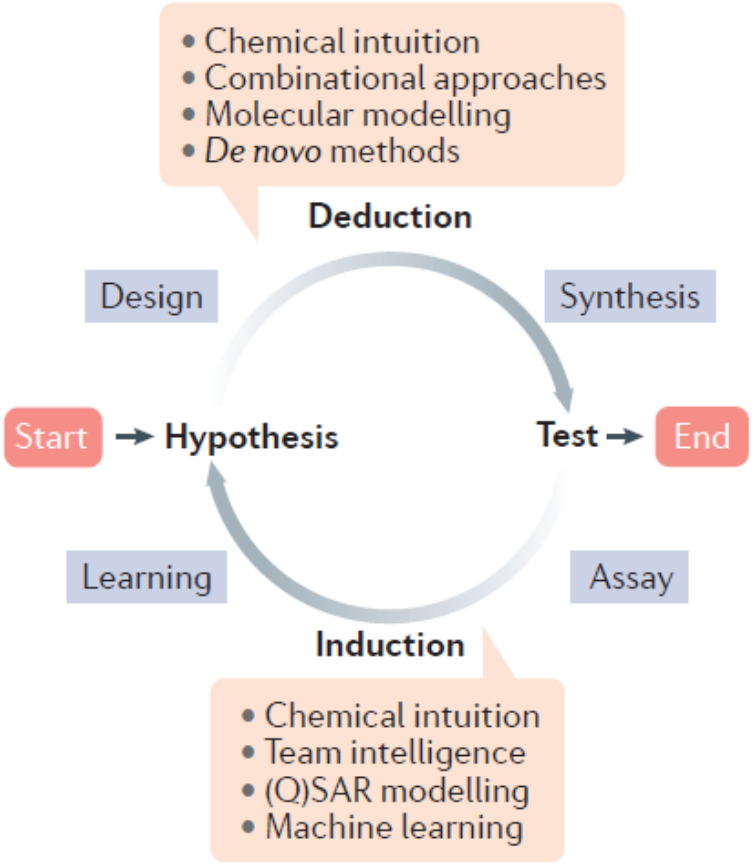
Automated discovery systems

- Langley, P. (**1978**). Bacon: A general discovery system. Proceedings of the Second Biennial Conference of the Canadian Society for Computational Studies of Intelligence (pp. 173-180). Toronto, Ontario.
- D. R. Swanson et al.: An interactive system for finding complementary literatures: a stimulus to scientific discovery, Artificial Intelligence, 1997
- Chrisman, L., Langley, P., & Bay, S. (**2003**). Incorporating biological knowledge into evaluation of causal regulatory hypotheses. Proceedings of the Pacific Symposium on Biocomputing (pp. 128-139). Lihue, Hawaii.
- R.D.King et al.: The Automation of Science, Science, 2009
- Rzhetsky, A. et al.: 2015. Choosing experiments to accelerate collective discovery. *PNAS*, *112*(47), pp.14569-14574.

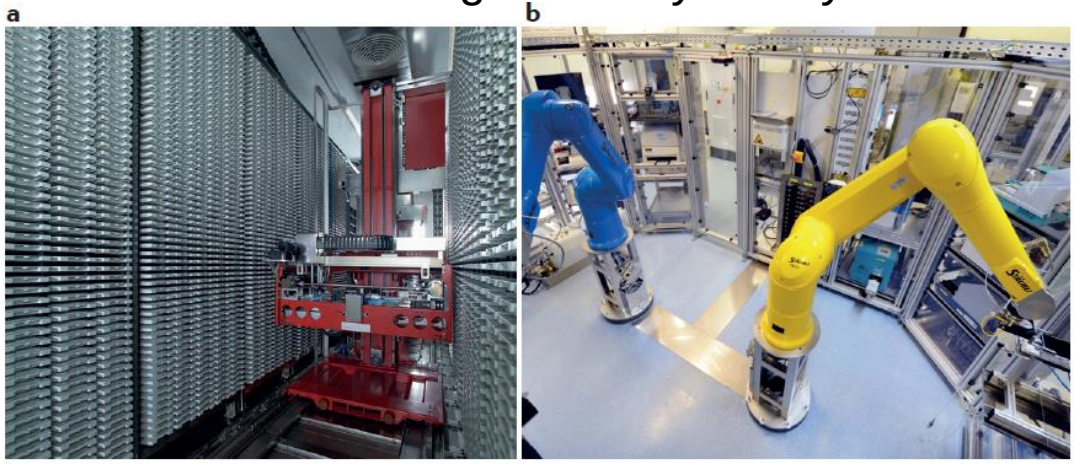
Artificial creativity???

Boden, M.A., 2009. Computer models of creativity. *AI Magazine*, *30*(3), pp.23-23.

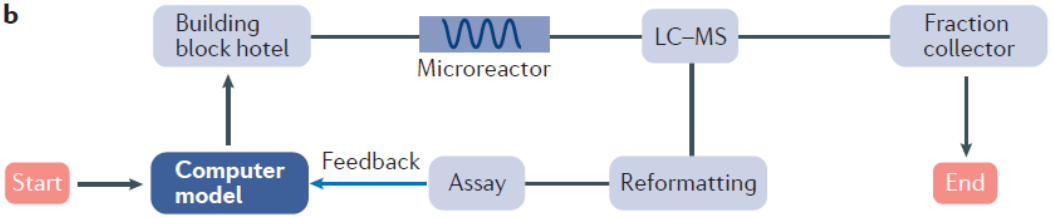
Automating drug discovery



Automated drug discovery facility



Active learning with microfluidics



Schneider, Gisbert. "Automating drug discovery." *Nature Reviews Drug Discovery* 17.2 (2018): 97.

De novo molecular design I.

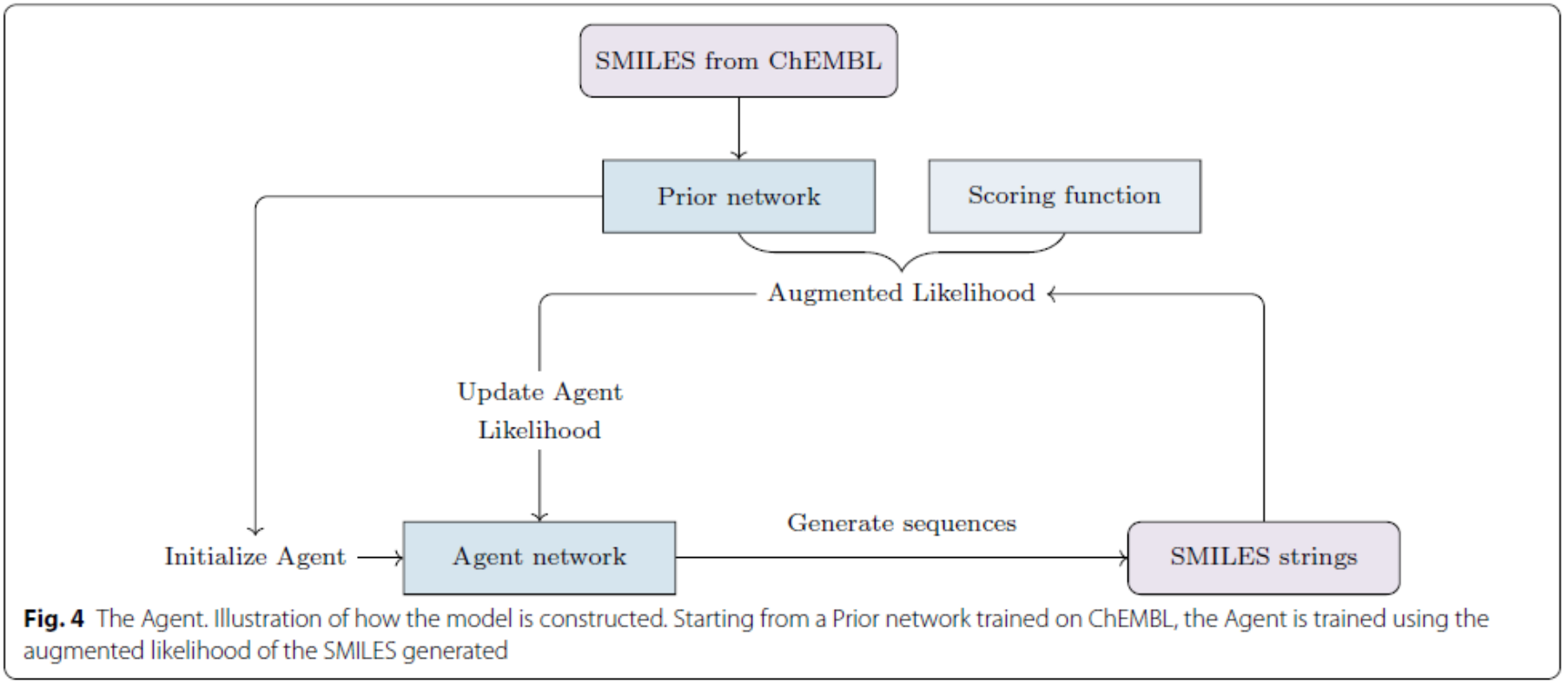
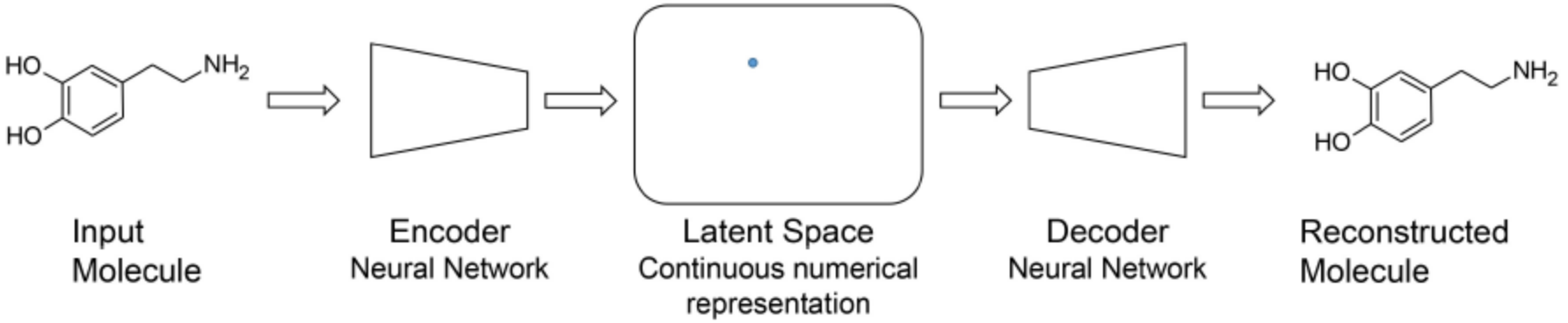


Fig. 4 The Agent. Illustration of how the model is constructed. Starting from a Prior network trained on ChEMBL, the Agent is trained using the augmented likelihood of the SMILES generated

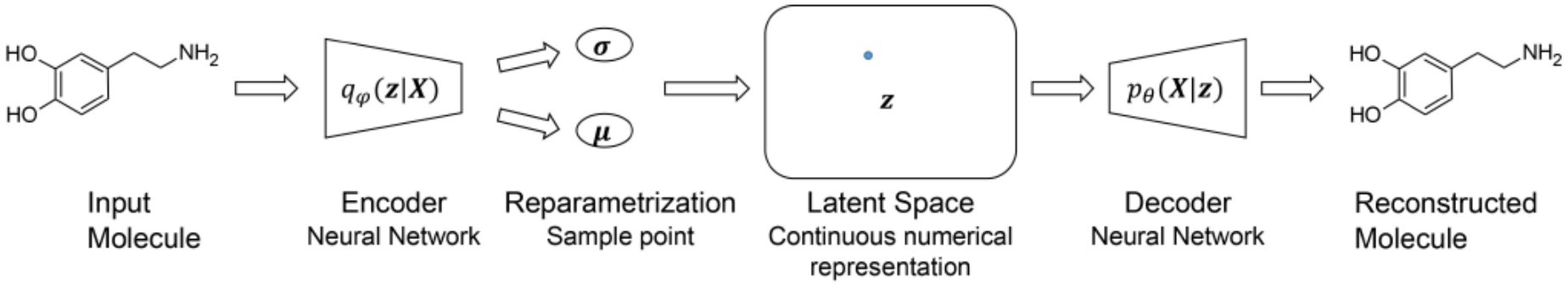
Olivecrona, M., Blaschke, T., Engkvist, O. and Chen, H., 2017. Molecular de-novo design through deep reinforcement learning. *Journal of cheminformatics*, *9*(1), p.48.

De novo molecular design II.

Autoencoder



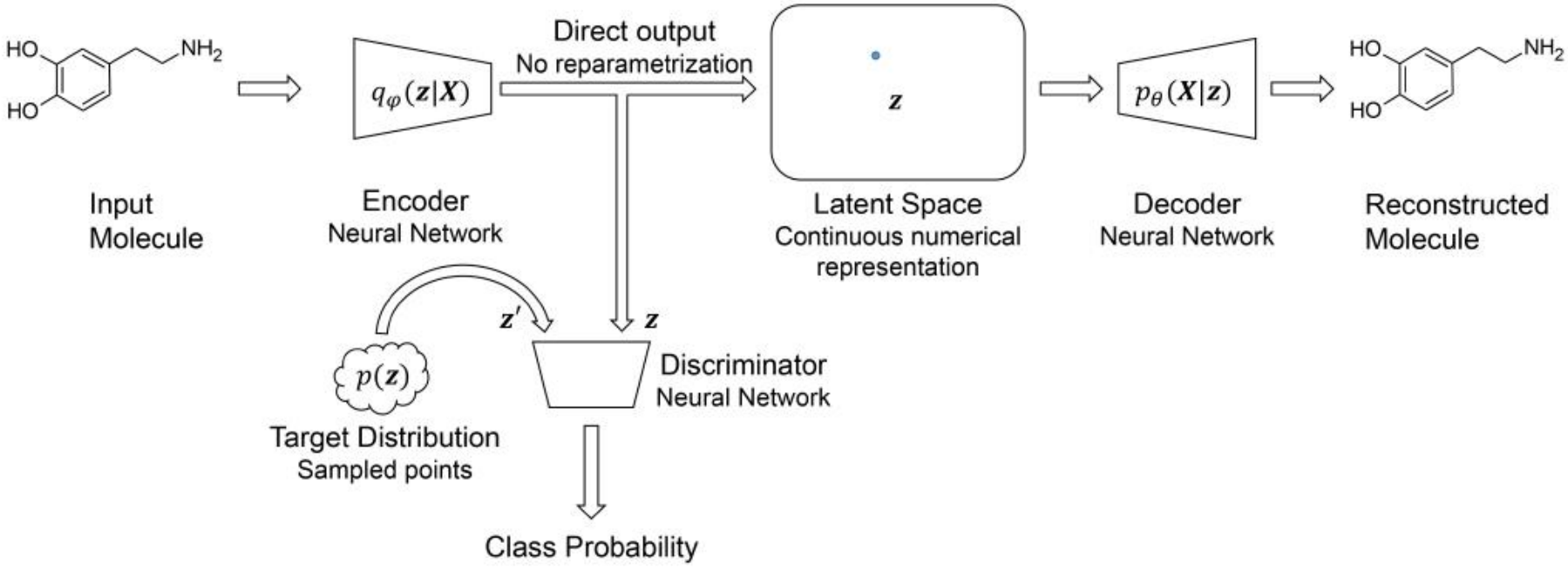
Variational autoencoder



Blaschke, T., Olivecrona, M., Engkvist, O., Bajorath, J. and Chen, H., 2018. Application of generative autoencoder in de novo molecular design. *Molecular informatics*, *37*(1-2), p.1700123.

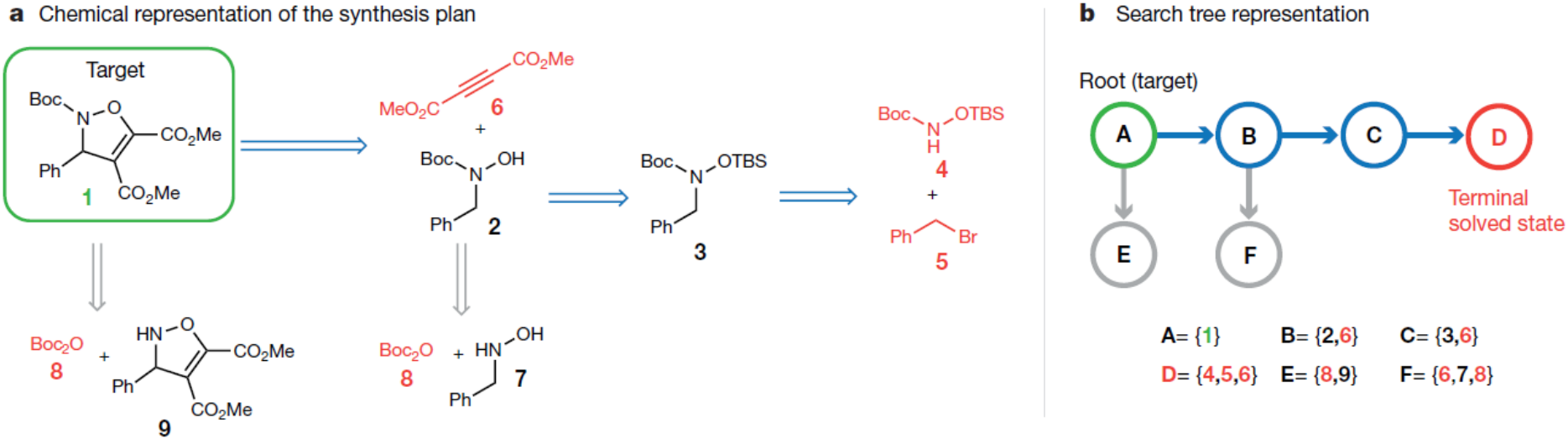
De novo molecular design III.

Generative adversarial autoencoder neural network



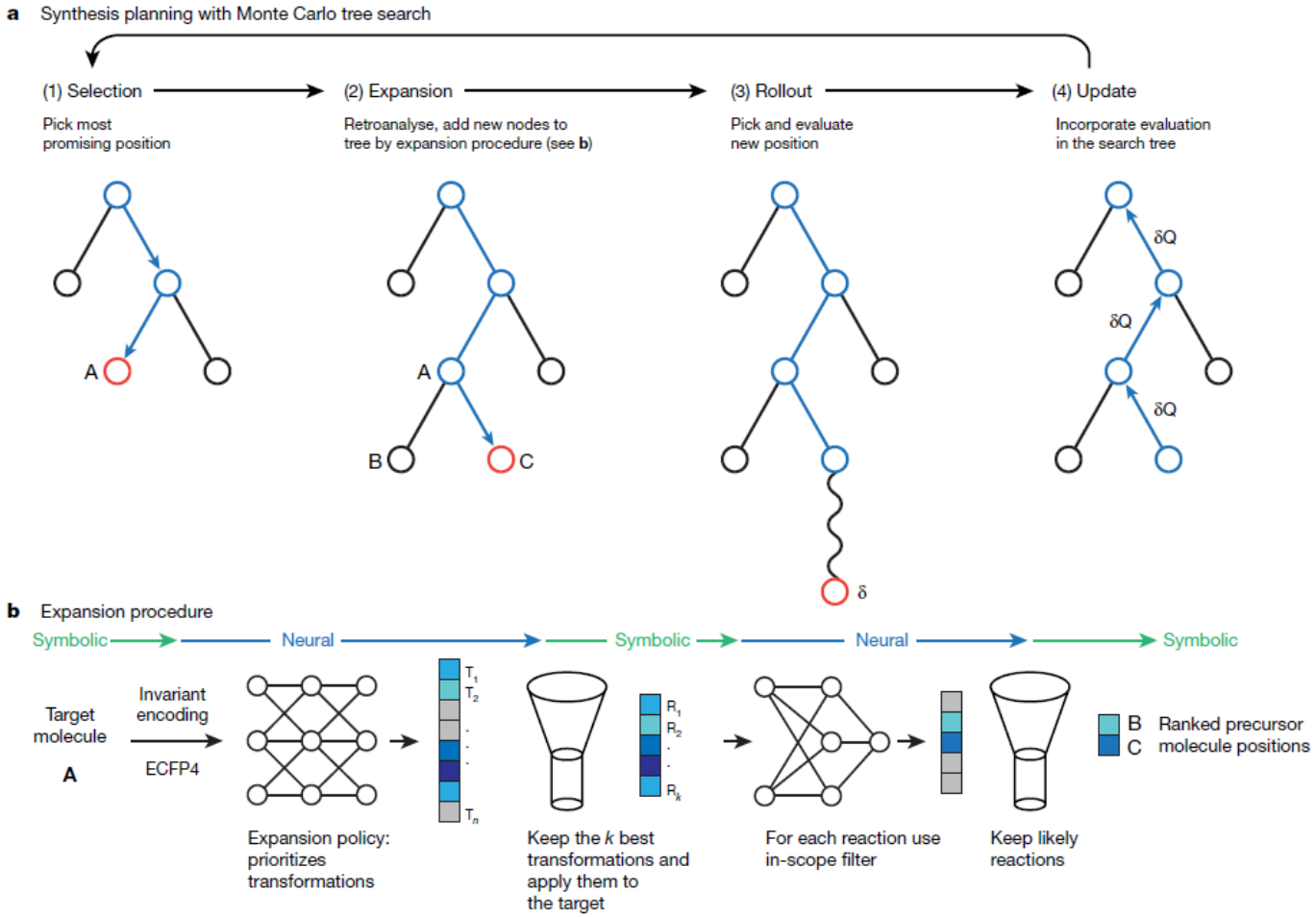
Blaschke, T., Olivecrona, M., Engkvist, O., Bajorath, J. and Chen, H., 2018. Application of generative autoencoder in de novo molecular design. *Molecular informatics*, *37*(1-2), p.1700123.

Chemical syntheses by deep artificial intelligence I.



Segler, M.H., Preuss, M. and Waller, M.P., 2018. Planning chemical syntheses with deep neural networks and symbolic AI. *Nature*, *555*(7698), p.604.

Chemical syntheses by deep artificial intelligence II.

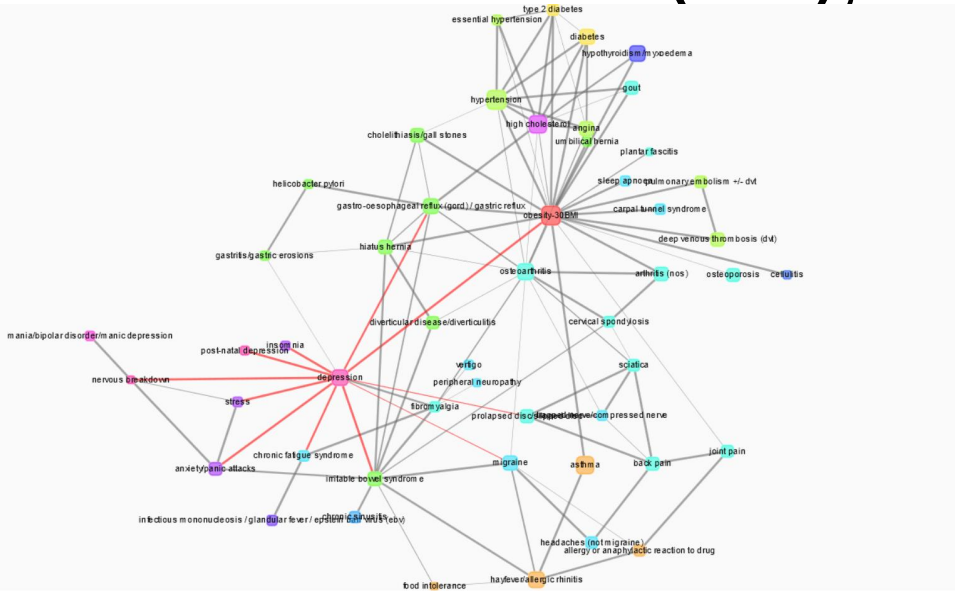


Segler, M.H., Preuss, M. and Waller, M.P., 2018. Planning chemical syntheses with deep neural networks and symbolic AI. *Nature*, *555*(7698), p.604.

DBNs in our research

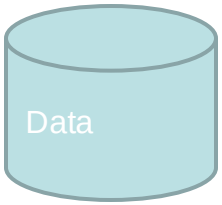
Genetics of multimorbidities

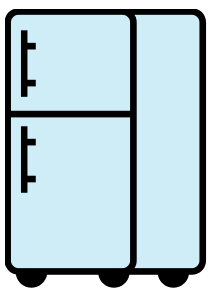
- OTKA 119866: Bayesian, systems-based methods for analyzing large health data sets, 2016-2020
- UK Biobank – research project No.1602, 2013-2017, 2017-2020
- Participants: Gabriella Juhász (SE), Péter Antal (BME)

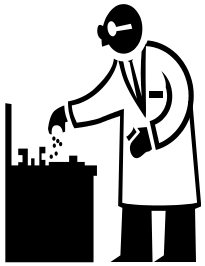


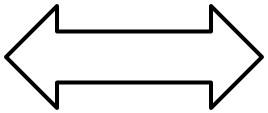
Complex models of laboratory tests in sequential decision support

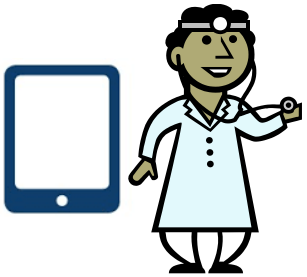
- In cooperation with Department of Laboratory Medicine, Semmelweis University











Drug-target interaction prediction

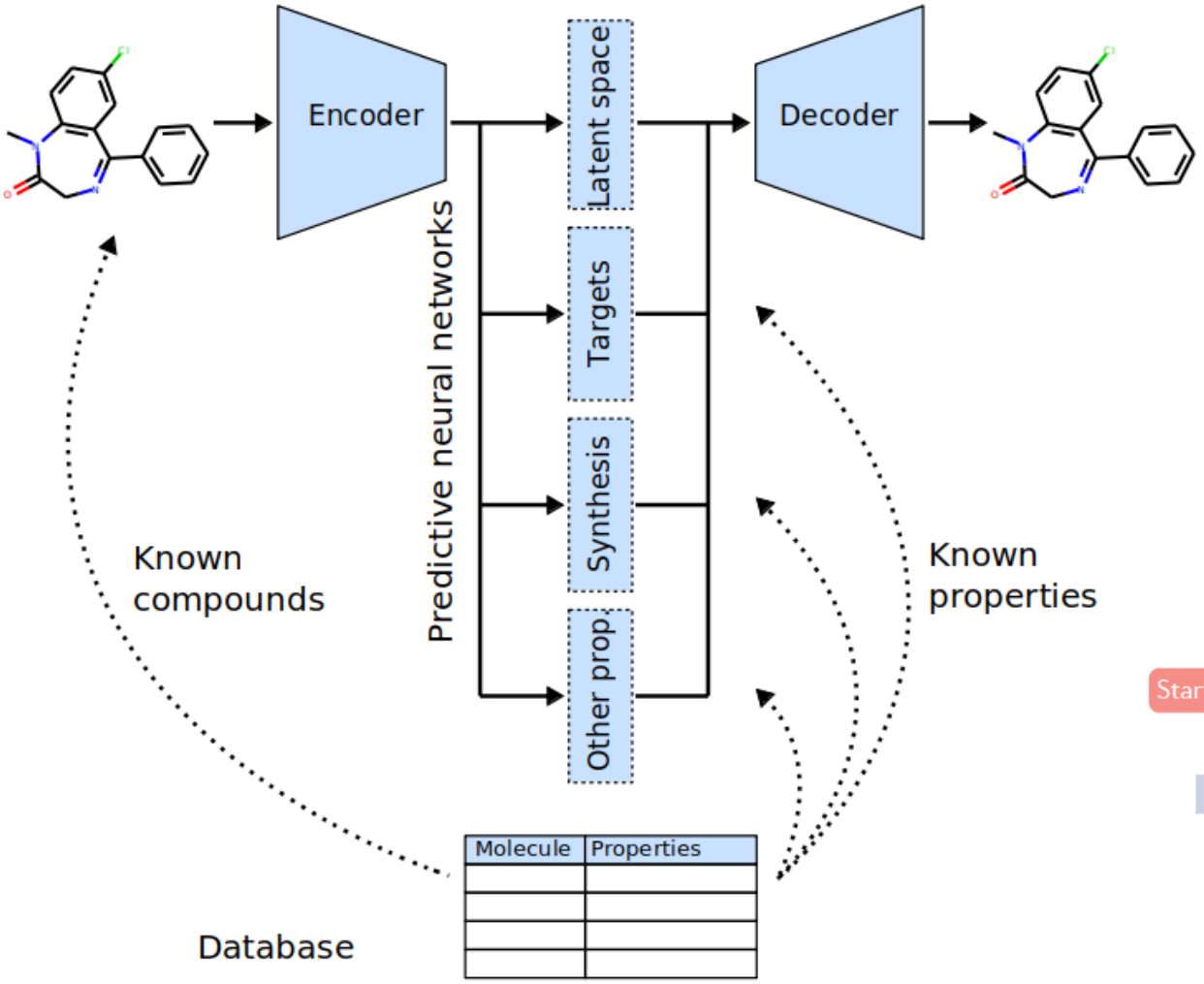
MELLODDY: privacy-preserving federated learning in drug discovery
https://www.imi.europa.eu/projects-results/project-factsheets/melloddy

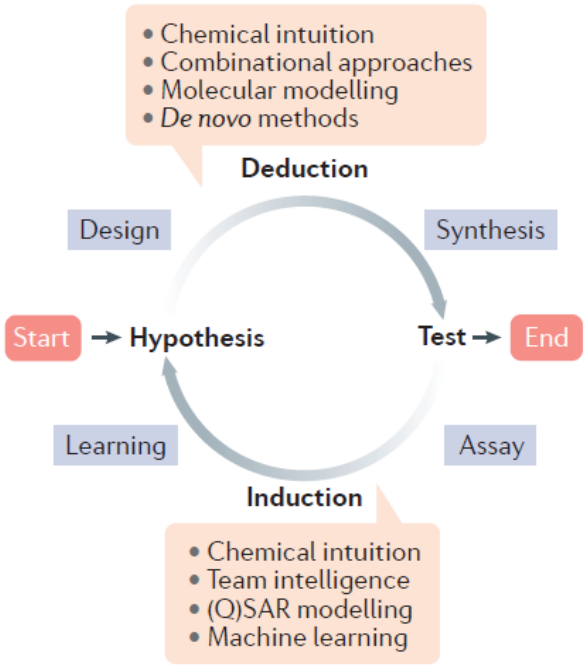
- IMI2 project:
- Participants
  - 10 big pharmas
  - 2 universities
  - 4 companies
  - 1 global IT company
- 2019-2021

Automated (early) drug discovery

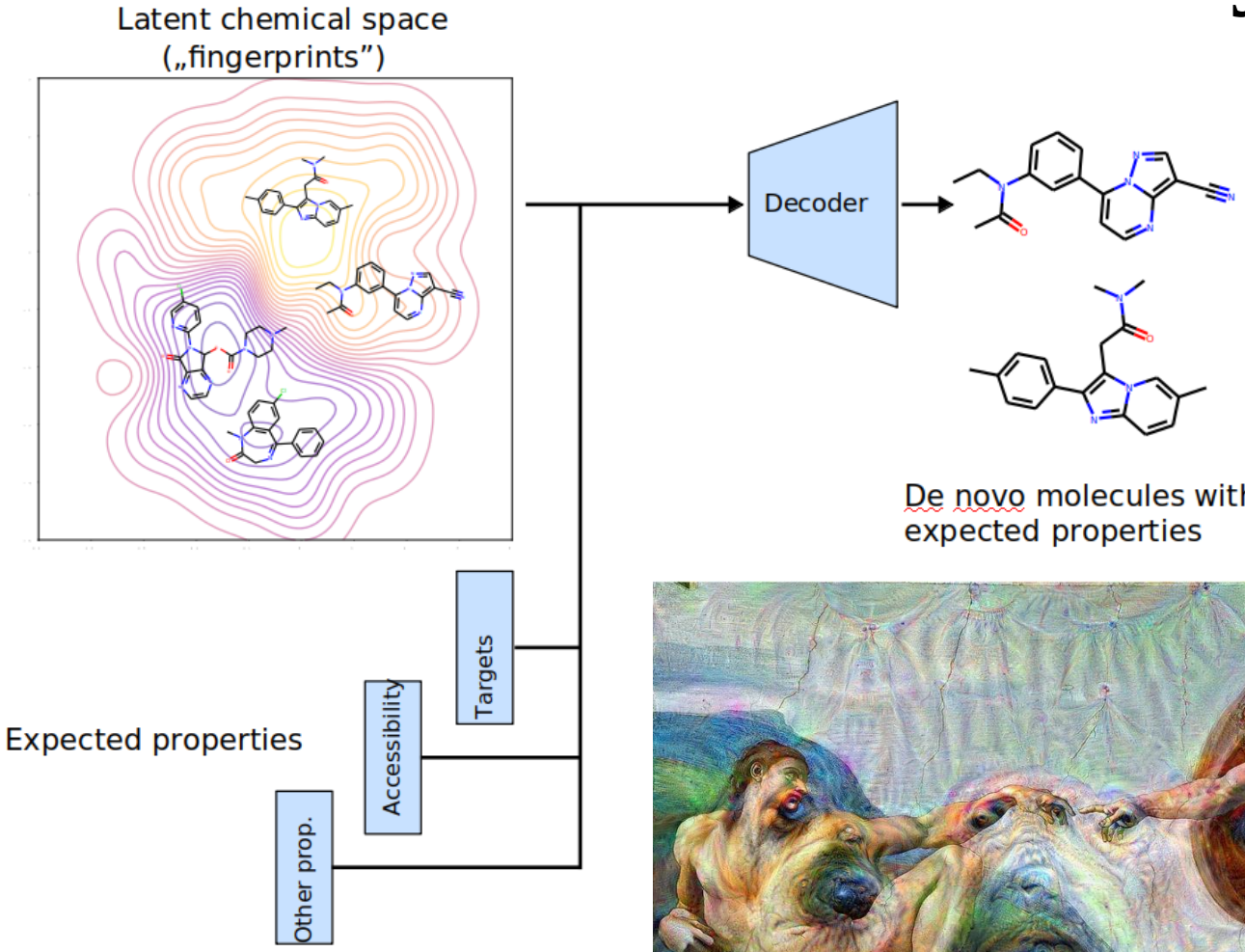
- Schneider, Gisbert, et al. "Virtual screening for bioactive molecules by evolutionary de novo design." Angewandte Chemie International Edition 39.22 (2000): 4130-4133.
- Schneider, Gisbert, and Uli Fechner. "Computer-based de novo design of drug-like molecules." Nature Reviews Drug Discovery 4.8 (2005): 649.
- Schneider, Gisbert, ed. De novo molecular design. John WileySons, 2013.
- Schneider, Gisbert. "Generative Models for Artificially-intelligent Molecular Design." Molecular informatics 37.1-2 (2018)
- Sanchez-Lengeling, Benjamin, and Alán Aspuru-Guzik. "Inverse molecular design using machine learning: Generative models for matter engineering." Science 361.6400 (2018): 360-365.
- Merk, Daniel, et al. "De novo design of bioactive small molecules by artificial intelligence." Molecular informatics 37.1-2 (2018): 1700153.
- Schneider, Gisbert, and David E. Clark. "Automated De Novo Drug Design–“Are we nearly there yet?”." Angewandte Chemie (2019).

De novo molecule generation: learning





De novo molecule generation: artificial creativity



Google: Deep dream

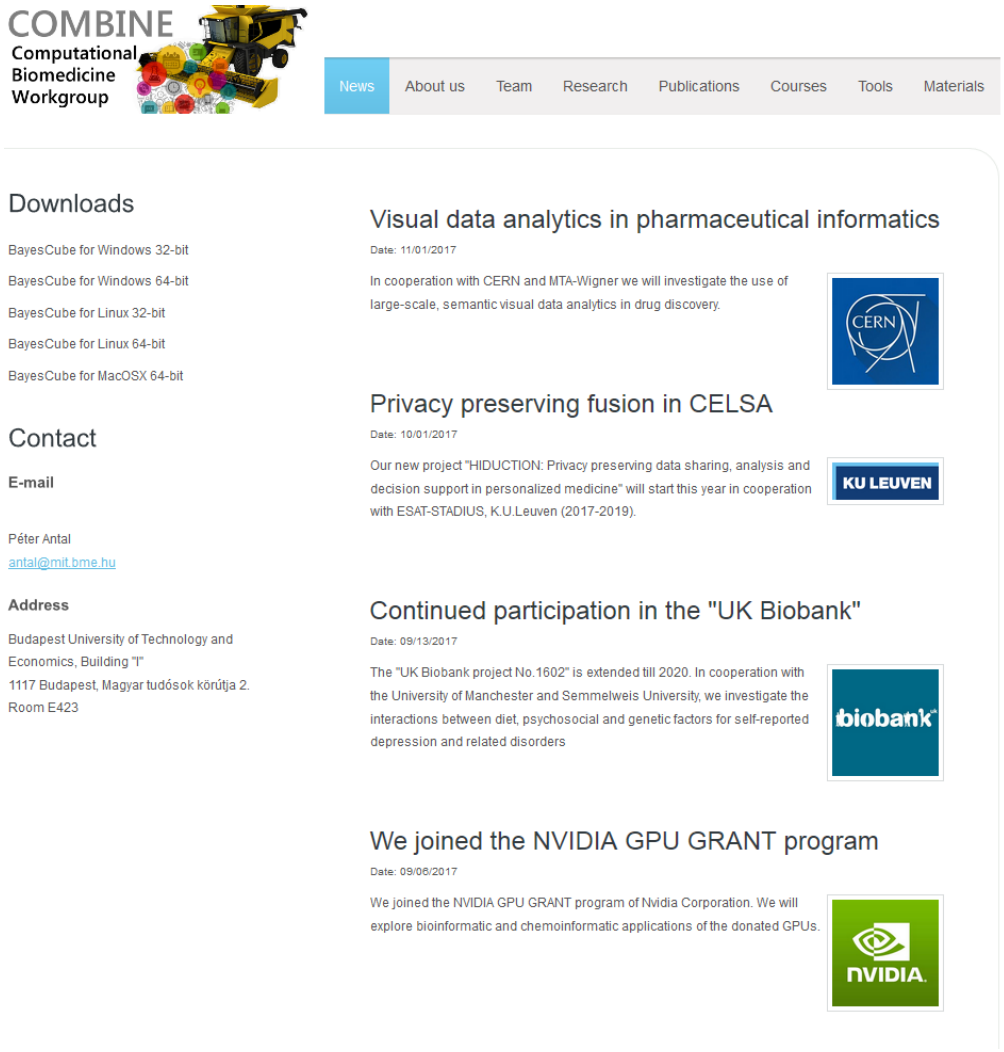
Intelligent de novo generation: polypharmacology and multitargets

- L Bolognesi, M. "Polypharmacology in a single drug: multitarget drugs." Current medicinal chemistry 20.13 (2013): 1639-1645.
- Medina-Franco, José L., et al. "Shifting from the single to the multitarget paradigm in drug discovery." Drug discovery today 18.9-10 (2013): 495-501.
- Zhang, Weilin, Jianfeng Pei, and Luhua Lai. "Computational multitarget drug design." Journal of chemical information and modeling 57.3 (2017): 403-412.
- Proschak, Ewgenij, Holger Stark, and Daniel Merk. "Polypharmacology by Design: A Medicinal Chemist's Perspective on Multitargeting Compounds." Journal of medicinal chemistry 62.2 (2018): 420-444.

Summary

- Complex generative models in biomedicine
  - Standard models for general decision support
  - Flexible models for highly incomplete data
  - Artificial creativity

Computational Biomedicine (ComBine) lab



Team
Bence Bolgár
Bence Bruncsics
András Gézsi
Gábor Hullám
András Millinghoffer
Péter Sárközy
Péter Antal

http://bioinfo.mit.bme.hu/

Thank you for you attention!