Institute of Structural Biology

Big Data in Chemistry

Dr. Igor V. Tetko

Institute of Structural Biology (STB), Helmholtz Zentrum München (HMGU)

Neuherberg, 18 October 2016

HelmholtzZentrum münchen
German Research Center for Environmental Health

Outline

- Sources
- Example of Big Data
- Data quality and complexity
- Annotation of large virtual sets
- Deep learning
- Secure data sharing
- Outlook

Big Data Sources

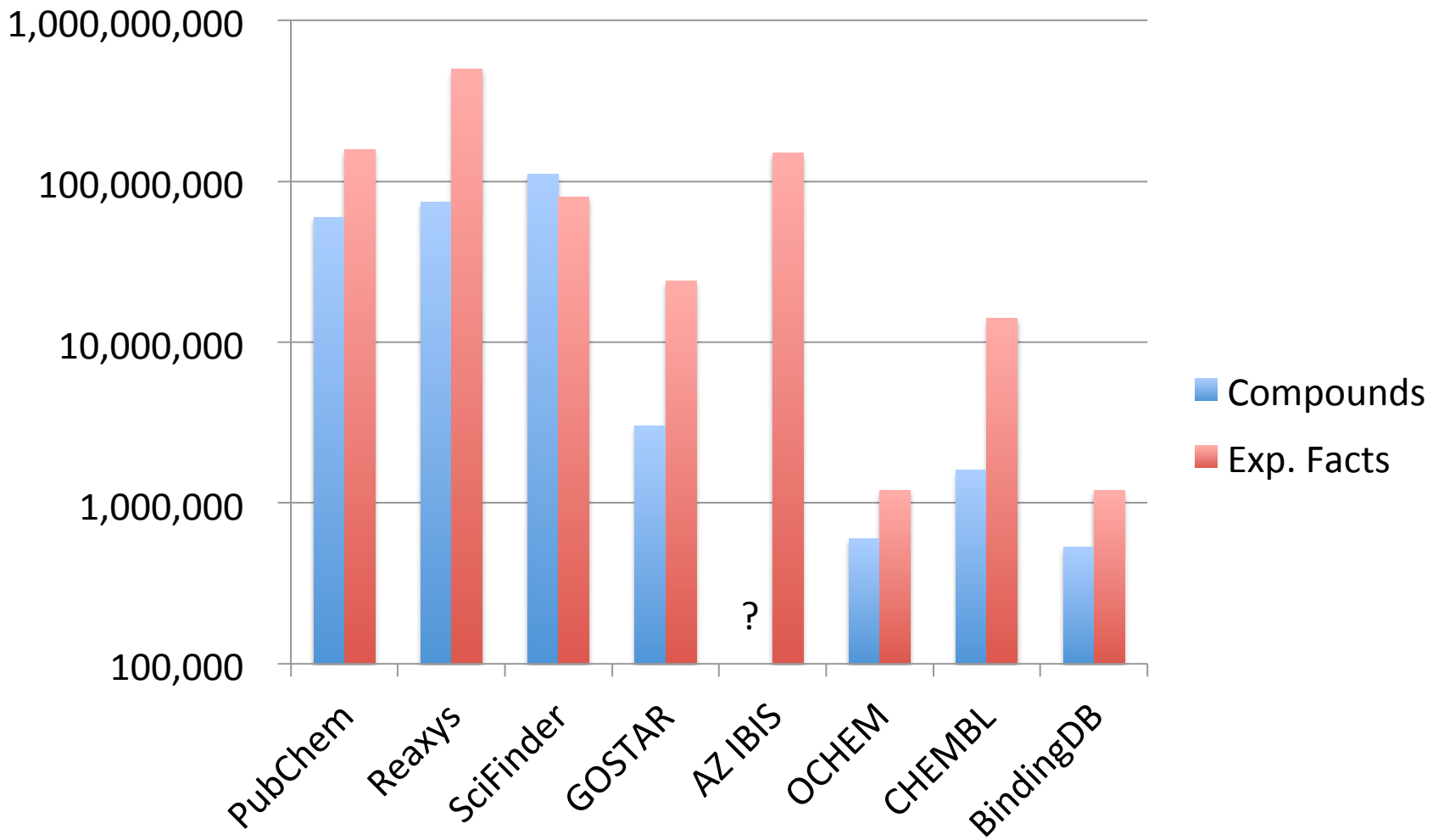
Do we really have Big Data in chemistry?

What kind of large data do we have?

Big Data definition

Big data is a term for data sets that **are so large** or **complex** that traditional data processing applications are inadequate (Wikipedia)

Large Chemical Database



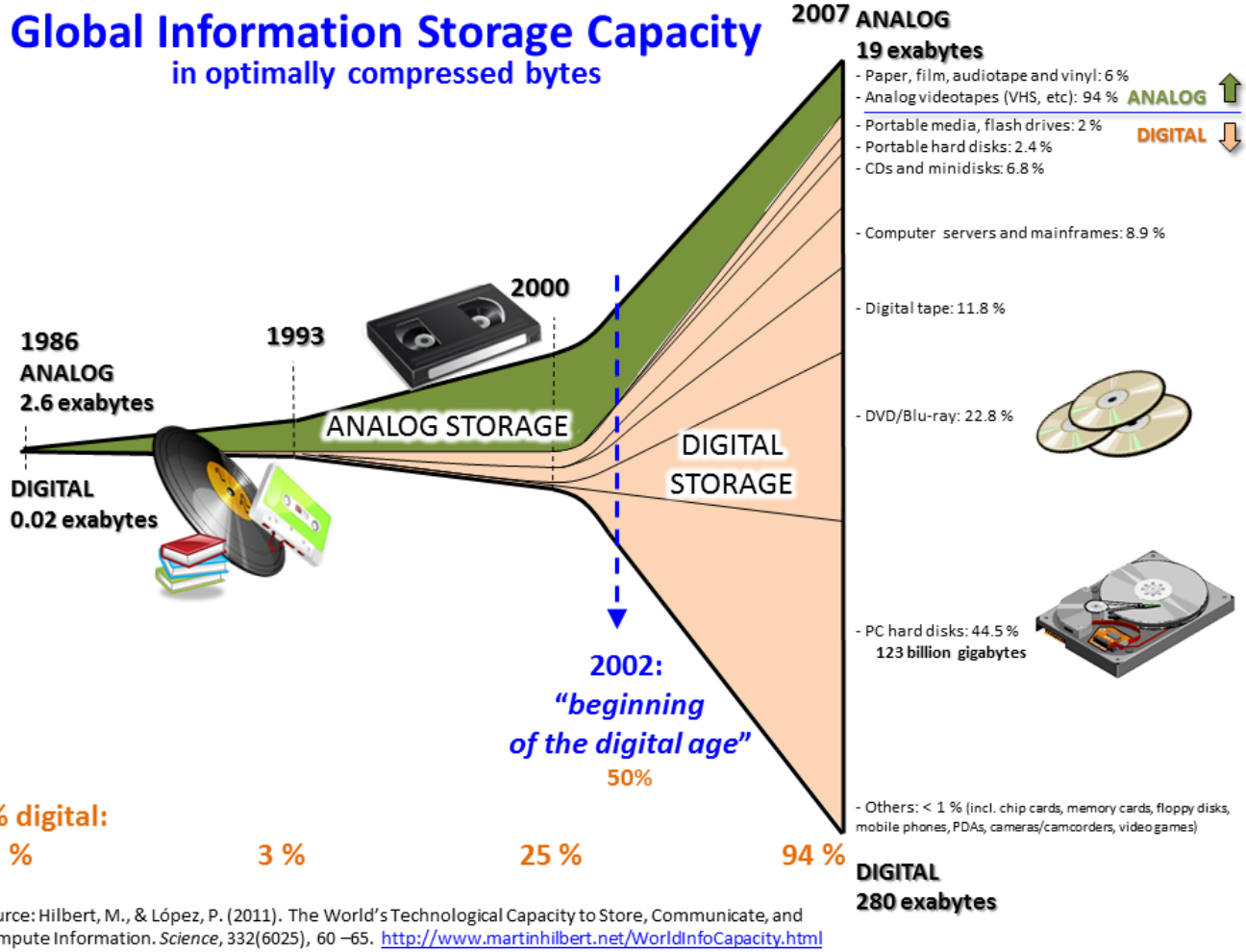
Data Types

Database	Main data types
ChEMBL v. 21[1]	Data mined from literature and PubChem HTS assays
BindingDB[2]	Experimental protein-small molecule interaction data
PubChem[3]	Bioactivity data from HTS assays
Reaxys[4]	Literature mined property, activity and reaction data
SciFinder (CAS)[5]	Experimental properties, ^{13}C and ^{1}H NMR spectra, reaction data
GOSTAR[6]	Target-linked data from patents and articles
AZ IBIS[7]	AZ *in-house* SAR data points
OCHEM[8]	Mainly ADMET data collected from literature

1) Papadatos G, et al. J Comput Aided Mol Des 2015;29(9)885-96.
2) Gilson MK, et al. Nucleic Acids Res 2016;44(D1):D1045-53.
3) Kim S, et al. Nucleic Acids Res 2016;44(D1):D1202-13.
4) http://www.elsevier.com/solutions/reaxys
5) http://www.cas.org/products/scifinder
6) http://www.gostardb.com
7) Muresan S et al. Drug Discov Today 2011;16(23-24):1019-30.
8) Sushko I, et al.. J Comput Aided Mol Des 2011;25(6):533-54.

Big Data sizes

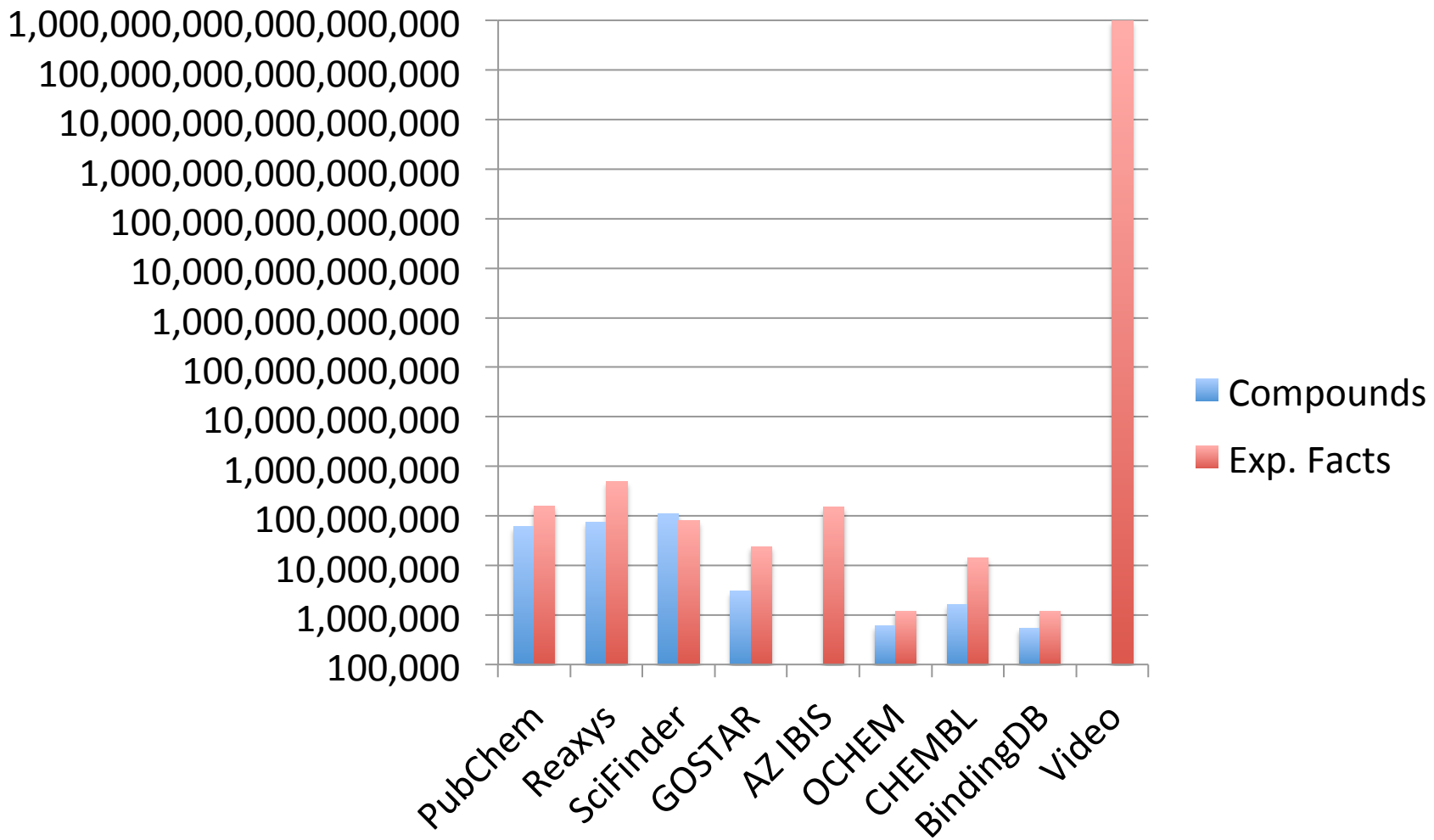
Big data is a term for data sets that are so large or complex that traditional data processing applications are inadequate (Wikipedia)



Source: Hilbert, M., & López, P. (2011). The World's Technological Capacity to Store, Communicate, and Compute Information. *Science*, 332(6025), 60 –65. http://www.martinhilbert.net/WorldInfoCapacity.html

1 exabyte: 10^{18}

Large Chemical Database



Big Data are relative to a field

- Methods to analyze such data do not exist
- We may not sufficient technical resources (speed, memory) to use the existing methods
- We may not have knowledge to use the existing methods

Thus the Big Data can appear due to:

Physical challenges (hardware)

Knowledge challenges (informatics, software)

Example of Big Data

Which data are really big ones?

What data sizes are "big" ones?

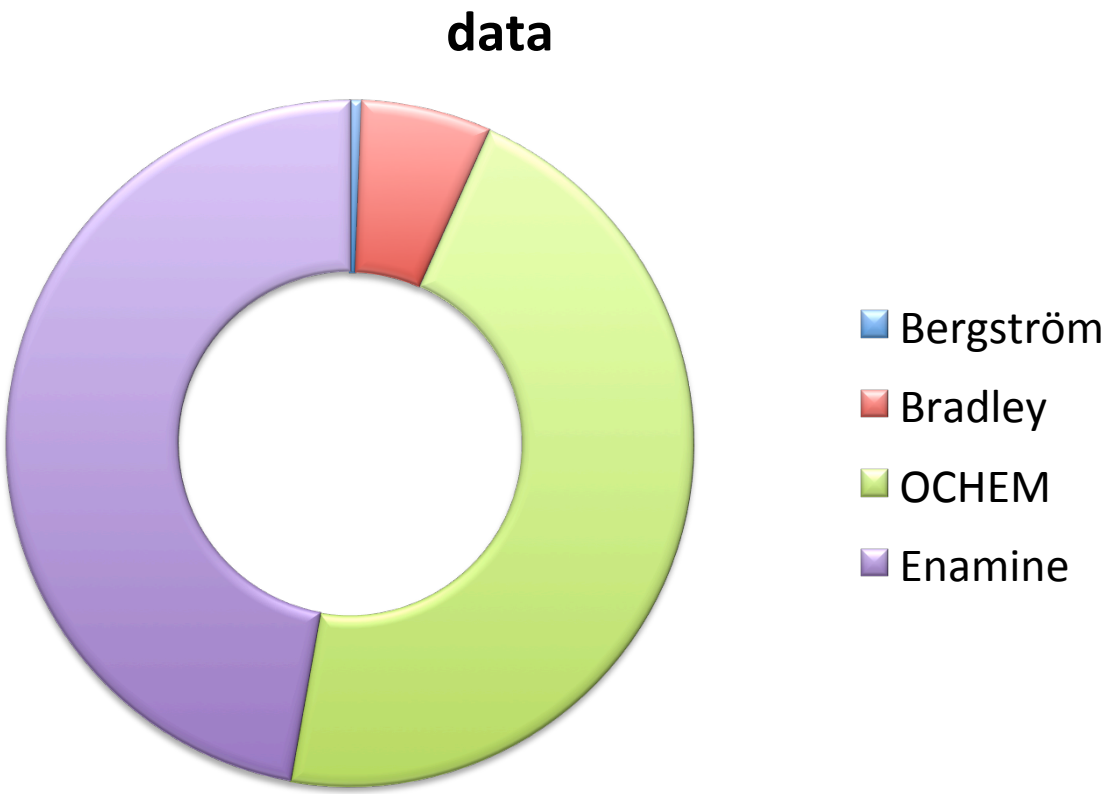
"General melting point prediction based on a diverse compound data set and artificial neural networks" Karthikeyan et al. J. Chem. Inf. Model. 2005, 45(3), 681-90. N = 4173

→ Large data set ~50k

→ Big data set ~250k

Melting Point Datasets

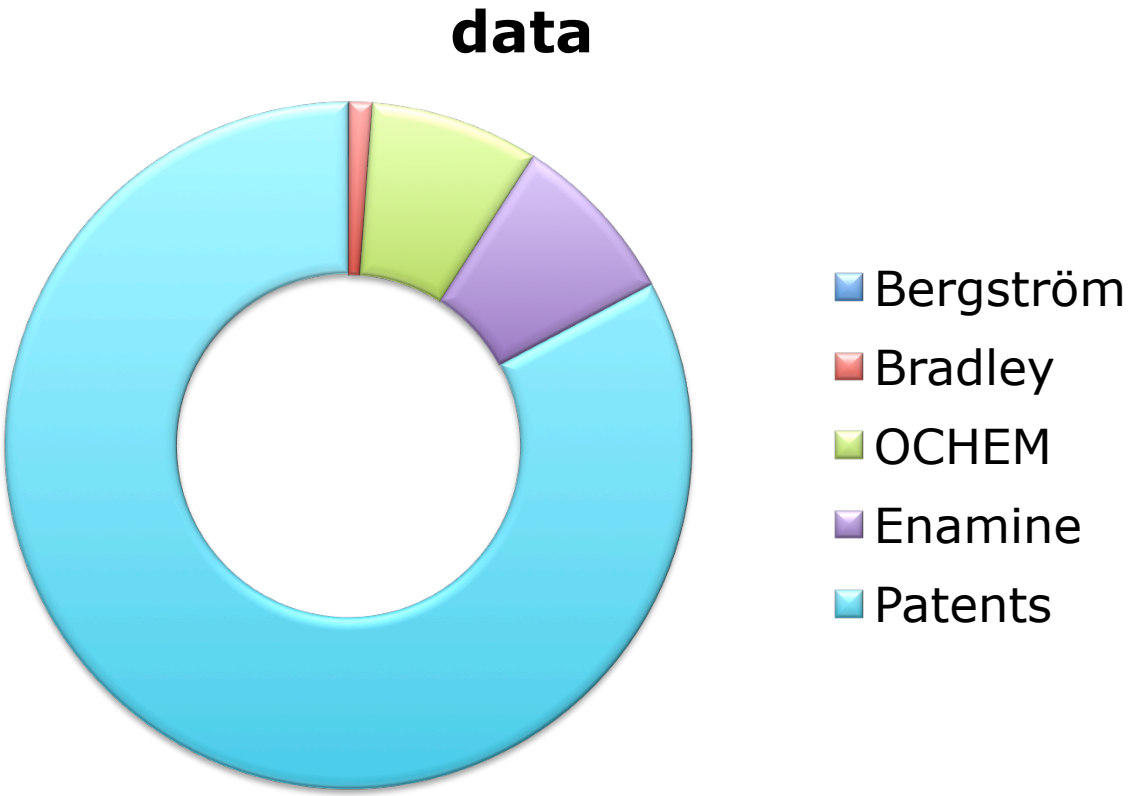
- Bergström 277
- Bradley 2886
- OCHEM 22404
- Enamine 21883



Tetko et al *J. Chem. Inf. Model*. 2014, 22;54(12):3320-9.

275k Melting Point Datasets

- Bergström 277
- Bradley 2886
- OCHEM 22404
- Enamine 21883
- PATENTS 228079



COMBINED: OCHEM + Enamine + Bradley + Bergström

Tetko et al *J. Chemoinformatics, 2016, 8, 2.*

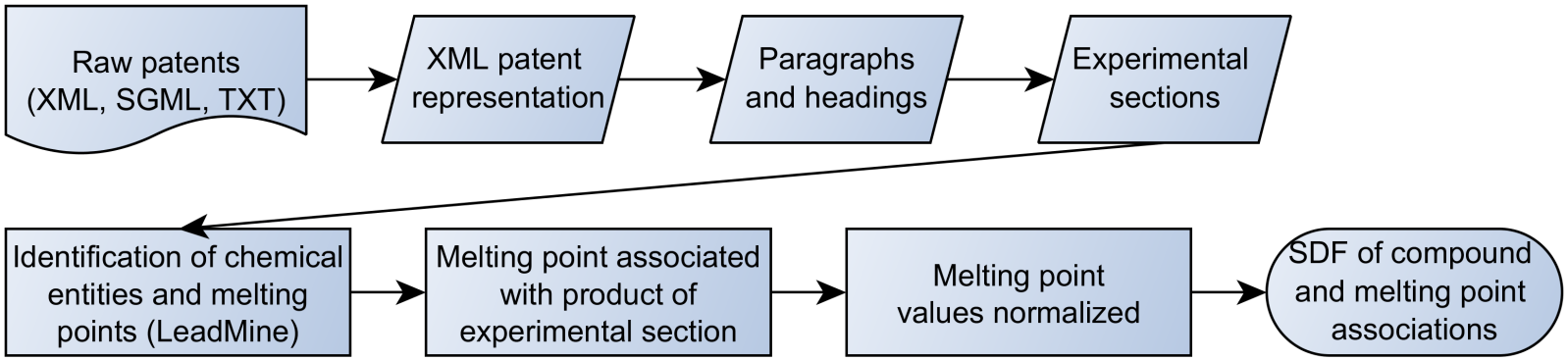
Extraction of MP information from patents

- [0835] To a solution of 2-amino-4,6-dimethoxybenzamide (0.195 g, 0.99 mmol) and 5-(2-(tert-butyldimethylsilyloxy)ethoxy)-6-phenylpicolinaldehyde (0.355 g, 0.99 mmol) in N,N-dimethyl acetamide (10 ml), was added NaHSO3 (0.264 g, 1.49 mmol) and p-toluenesulfonic acid monohydrate (0.038 g, 0.198 mmol). The reaction mixture was heated at 120° C. for 16 h. After that time the reaction was cooled to rt and the solvent was removed under reduced pressure. The reaction mixture was then diluted with water (150 mL) and neutralized with NaHCO3. The precipitated solids were collected by filtration, washed with water and dried to give 2-(5-(2-(tert-butyldimethylsilyloxy)ethoxy)-6-phenylpyridin-2-yl)-5,7-dimethoxyquinazolin-4(3H)-one (0.500 g, 94%) as an off-white solid: 1H NMR (400 MHz, DMSO-d6) δ 11.08 (s, 1H), 8.35 (d, J=8.98 Hz, 1H), 8.21 (d, J=2.34 Hz, 2H), 7.82 (d, J=8.59 Hz, 1H), 7.44-7.52 (m, 3H), 6.81 (d, J=2.34 Hz, 1H), 6.58 (d, J=2.34 Hz, 1H), 4.24-4.32 (m, 2H), 3.94-4.00 (m, 2H), 3.92 (s, 3H), 3.86 (s, 3H), 0.85 (s, 9H), 0.08 (s, 6H); ESI MS m/z 534 [M+H]+.

- http://www.google.com/patents/US20140140956

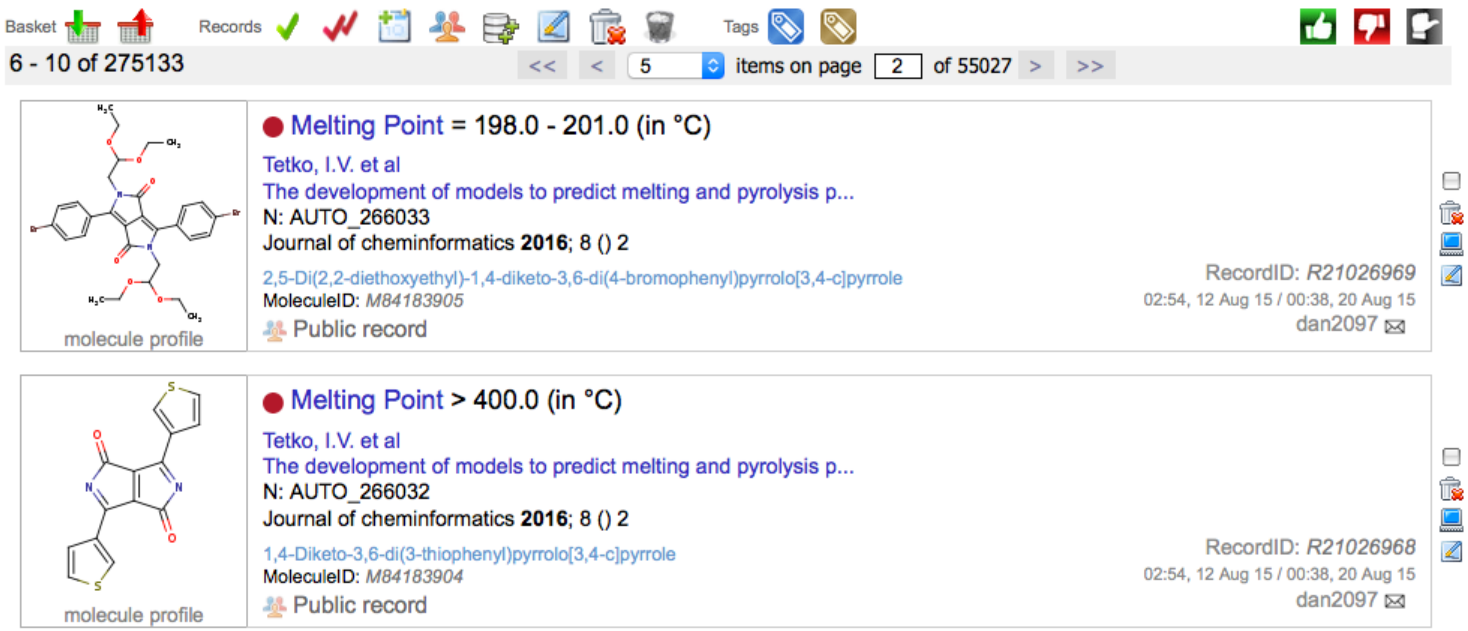
Extracting of melting points from patents

Workflow



Extraction of MP information from patents

[0835] To a solution of 2-amino-4,6-dimethoxybenzamide (0.266 g, 1.36 mmol) and 3-(5-(methylsulfinyl)thiophen-2-yl)benzaldehyde (0.34 g, 1.36 mmol) in N,N-dimethylacetamide (17 mL) was added $NaHSO_3$ (0.36 g, 2.03 mmol) and p-toluenesulfonic acid monohydrate (0.052 g, 0.271 mmol) at rt. The reaction mixture was heated at 120° C. for 12.5 h. After that time the reaction was cooled to rt, concentrated under reduced pressure and diluted with water (20 mL). The precipitated solids were collected by filtration, washed with water and dried. The product was purified by flash column chromatography (silica gel, 95:5 chloroform/methanol) to give 5,7-dimethoxy-2-(3-(5-(methylsulfinyl)thiophen-2-yl)phenyl)quinazolin-4(3H)-one (0.060 g, 10%) as a light yellow solid: mp 289-290° C.; ^{1}H NMR (400 MHz, DMSO-d_6) δ 12.19 (br s, 1H), 8.48 (s, 1H), 8.18 (d, J=7.81 Hz, 1H), 7.90 (d, J=8.20 Hz, 1H), 7.72 (d, J=3.90 Hz, 1H), 7.55-7.64 (m, 2H), 6.77 (d, J=2.34 Hz, 1H), 6.54 (d, J=1.95 Hz, 1H), 3.88 (s, 3H), 3.84 (s, 3H), 2.96 (s, 3H); ESI MS m/z 427 $[M+H]^+$.



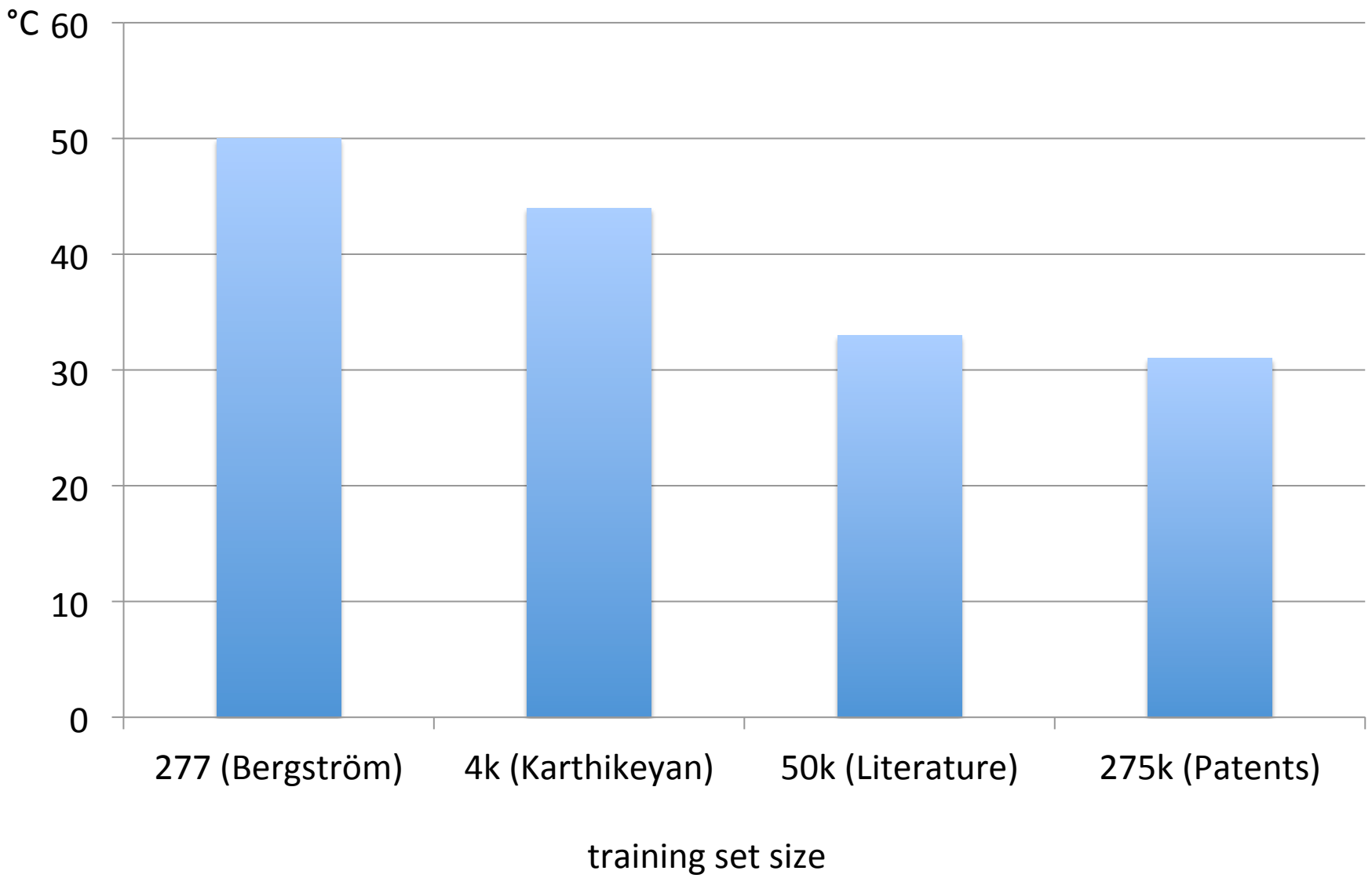
Modeling of MP data

Package name	Type of descriptors	Number of descriptors	Matrix size, billions	Non zero values, millions	Sparseness
Functional Groups	integer	595	0.18	3.1	33
QNPR	integer	1502	0.45	6.3	49
MolPrint	binary	688634	205	8.1	7200
Estate count	float	631	0.19	10	14
Inductive	float	54	0.02	11	1
ECFP4	binary	1024	0.31	12	25
Isida	integer	5886	1.75	18	37
ChemAxon	float	498	0.15	23	1.5
GSFrag	integer	1138	0.34	24	5.7
CDK	float	239	0.07	27	2
Adriana	float	200	0.06	32	1.3
Mera, Mersy	float	571	0.17	61	1.1
Dragon	float	1647	0.49	183	1.5

Large → Big

- Neural Networks was too slow (ensemble training!)
  → SVM was used
  - Support of parallel calculations (48 core)
  - Support of grid analysis (>1000 CPUs)
- Storage of full data matrix -> sparse data matrix

Prediction errors for Bergström drug like compounds using models developed with different training sets

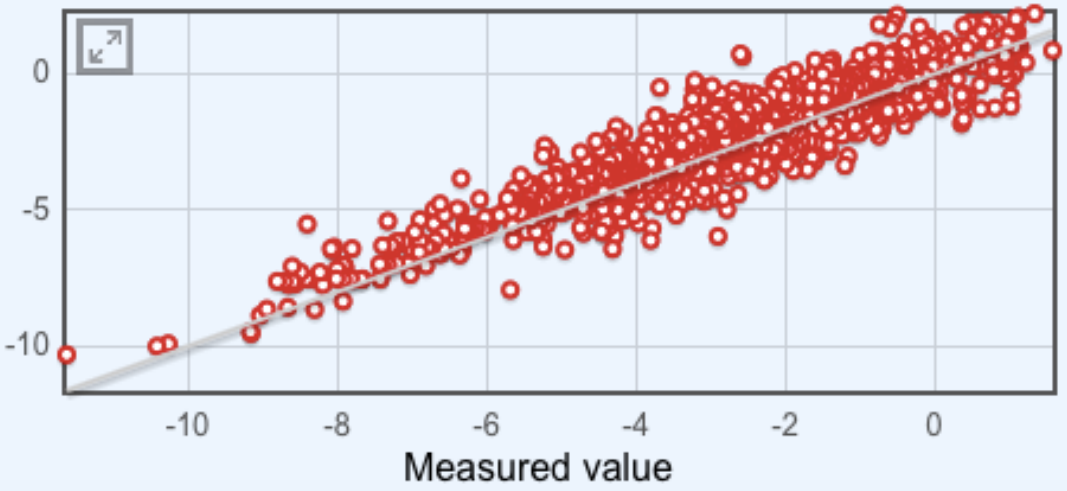


Prediction of Huuskonen set using ALOGPS logP and MP based on 50k measurements

logS = 0.5 – 0.01(MP-25) – log Kow

Predicted property: **Aqueous Solubility** modeled in log(mol/L)
Training method: MLRA

Data Set	#	R2	q2	RMSE	MAE
Training set: logS Huuskonen	1311 records	0.838 ± 0.009	0.81 ± 0.01	0.9 ± 0.02	0.71 ± 0.01

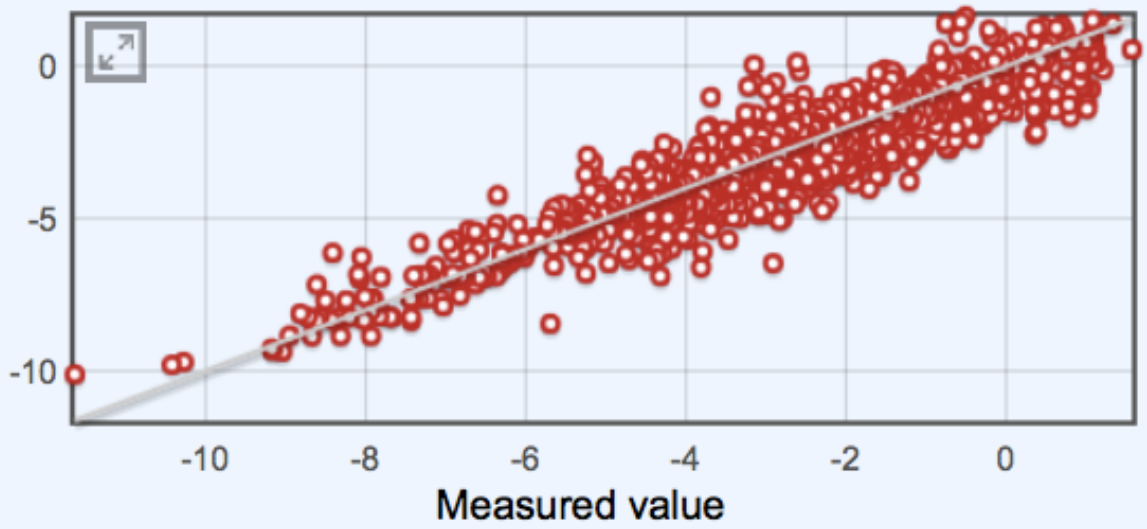


Prediction of Huuskonen set using ALOGPS logP and MP based on 230k measurements

$$logS = 0.5 - 0.01(MP-25) - log\ Kow$$

Predicted property: **Aqueous Solubility** modeled in log(mol/L)
Training method: MLRA

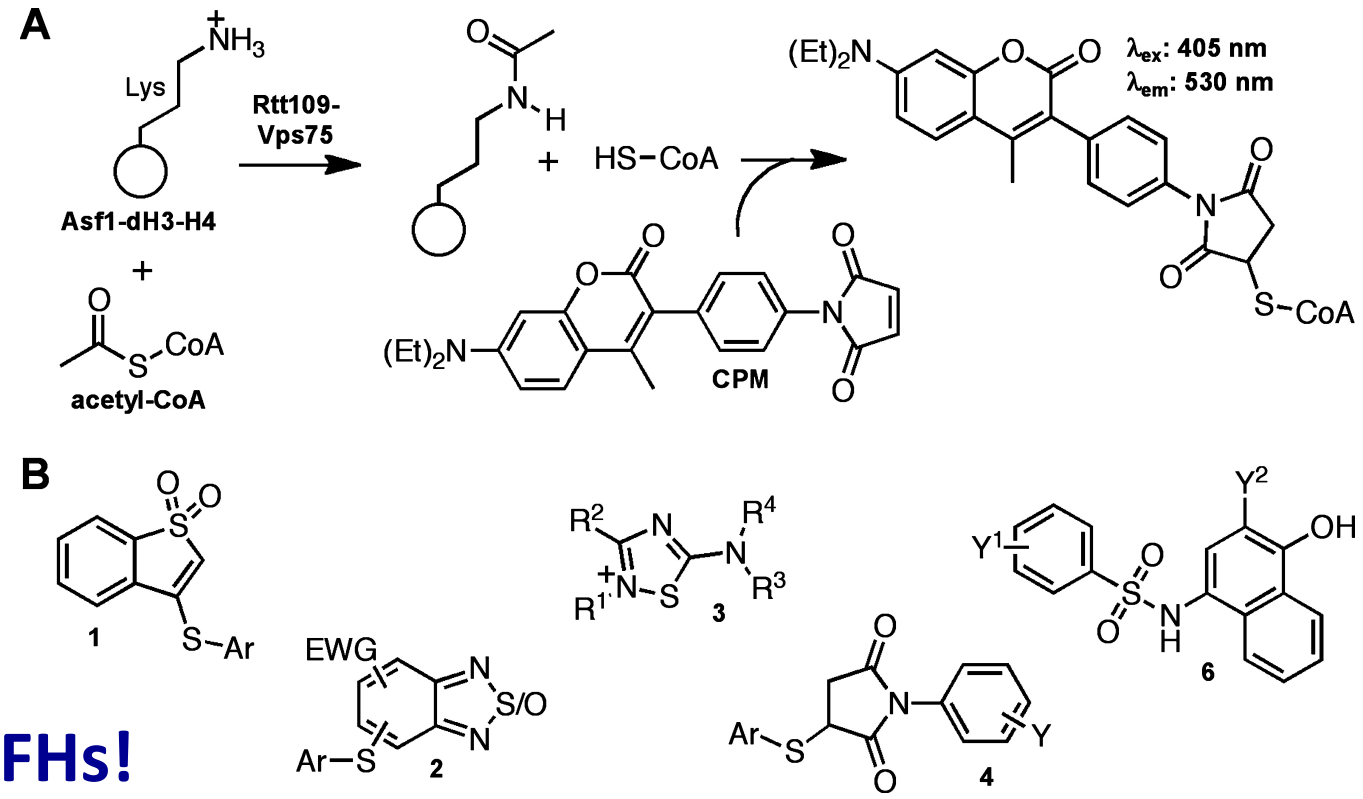
Data Set	#	R_2	q_2	RMSE	MAE
Training set: logS set	1311 records	0.842 ± 0.009	0.83 ± 0.01	0.84 ± 0.02	0.64 ± 0.02



Big Data Quality and Complexity

Why is it very important?

How domain specific analysis could help?



99.8% FHs!

Susceptibility of CPM-based HTS to screening compound-based interference. (A) Assay schematic for the CPM-based HTS used in this study. The assay measures the HAT activity of the Rtt109–Vps75 complex, which catalyzes the transfer of an acetyl moiety from acetyl-CoA to specific lysine residues on the Asf1–dH3–H4 substrate complex to produce acetylated histone residues and coenzyme A (CoA). Addition of the thiol-scavenging probe CPM leads to a highly fluorescent adduct by reacting with the CoA byproduct, which is used to quantify HAT activity via fluorescence intensity measurement. (B) Representative assay interference chemotypes identified during post-HTS triage.

Dahlin et al *J. Med. Chem.* 2015, 58, 2091-2113.

Promiscuous compounds filters

Journal of Medicinal Chemistry

Article

pubs.acs.org/jmc

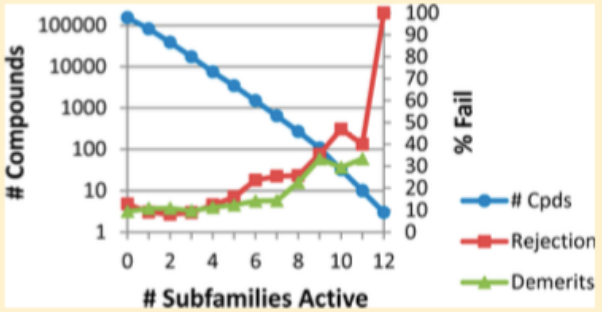
Rules for Identifying Potentially Reactive or Promiscuous Compounds

Robert F. Bruns* and Ian A. Watson

Lilly Research Laboratories, Eli Lilly and Company, Lilly Corporate Center, Indianapolis, Indiana 46285, United States

Supporting Information

ABSTRACT: This article describes a set of 275 rules, developed over an 18-year period, used to identify compounds that may interfere with biological assays, allowing their removal from screening sets. Reasons for rejection include reactivity (e.g., acyl halides), interference with assay measurements (fluorescence, absorbance, quenching), activities that damage proteins (oxidizers, detergents), instability (e.g., latent aldehydes), and lack of druggability (e.g., compounds lacking both oxygen and nitrogen). The structural queries were profiled for frequency of occurrence in druglike and nondruglike compound sets and were extensively reviewed by a panel of experienced medicinal chemists. As a means of profiling the rules and as a filter in its own right, an index of biological promiscuity was developed. The 584 gene targets with screening data at Lilly were assigned to 17 subfamilies, and the number of subfamilies at which a compound was active was used as a promiscuity index. For certain compounds, promiscuous activity disappeared after sample repurification, indicating interference from occult contaminants. Because this type of interference is not amenable to substructure search, a "nuisance list" was developed to flag interfering compounds that passed the substructure rules.



Promiscuous compounds filters

Journal of Medicinal Chemistry Article

J. Med. Chem. **2010,** *53,* 2719–2740 **2719**
DOI: 10.1021/jm901137j

New Substructure Filters for Removal of Pan Assay Interference Compounds (PAINS) from Screening Libraries and for Their Exclusion in Bioassays

Jonathan B. Baell*[,†,‡] and Georgina A. Holloway[†,‡]

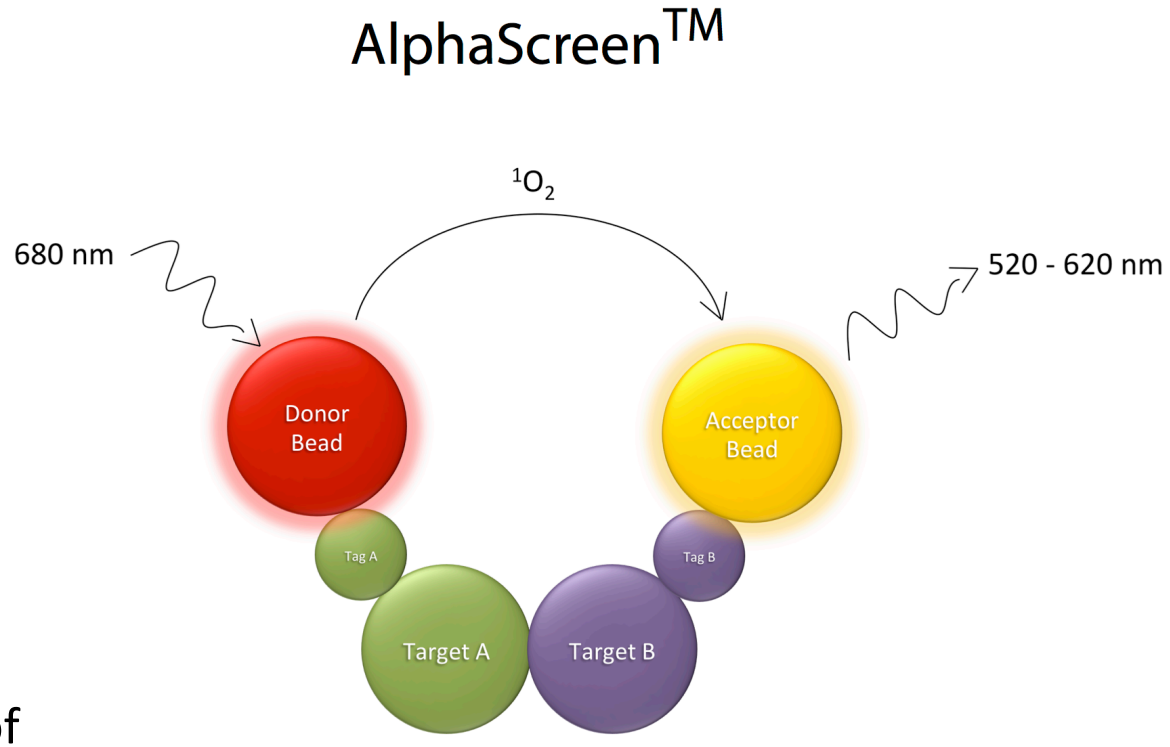
[†]*The Walter and Eliza Hall Institute of Medical Research, 1G Royal Parade, Parkville, Victoria 3052, Australia and* [‡]*Cancer Therapeutics-CRC P/L, 4 Research Avenue, La Trobe R&D Park, Bundoora, Victoria 3086, Australia*

Received July 31, 2009

This report describes a number of substructural features which can help to identify compounds that appear as frequent hitters (promiscuous compounds) in many biochemical high throughput screens. The compounds identified by such substructural features are not recognized by filters commonly used to identify reactive compounds. Even though these substructural features were identified using only one assay detection technology, such compounds have been reported to be active from many different assays. In fact, these compounds are increasingly prevalent in the literature as potential starting points for further exploration, whereas they may not be.

Pan Assay INterference compoundS (PAINS) Filters

- color quenching
- singlet oxygen quenching
- auto-fluorescence
- covalent binding
- inherently "sticking" compounds
- disrupt the interaction between the tag of the protein and binding site of the detection system



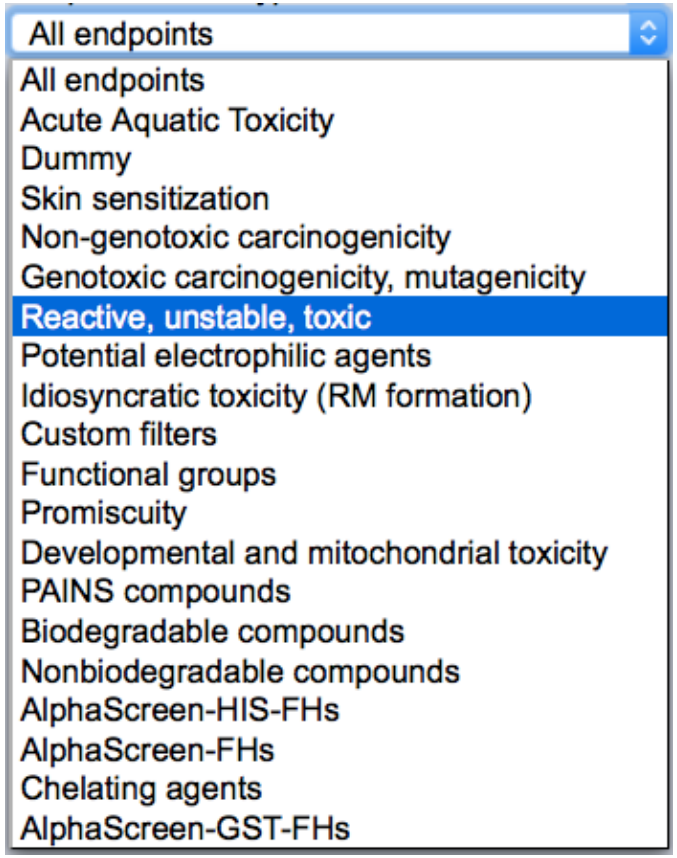
~ 500 filters based on N = 93212 compounds

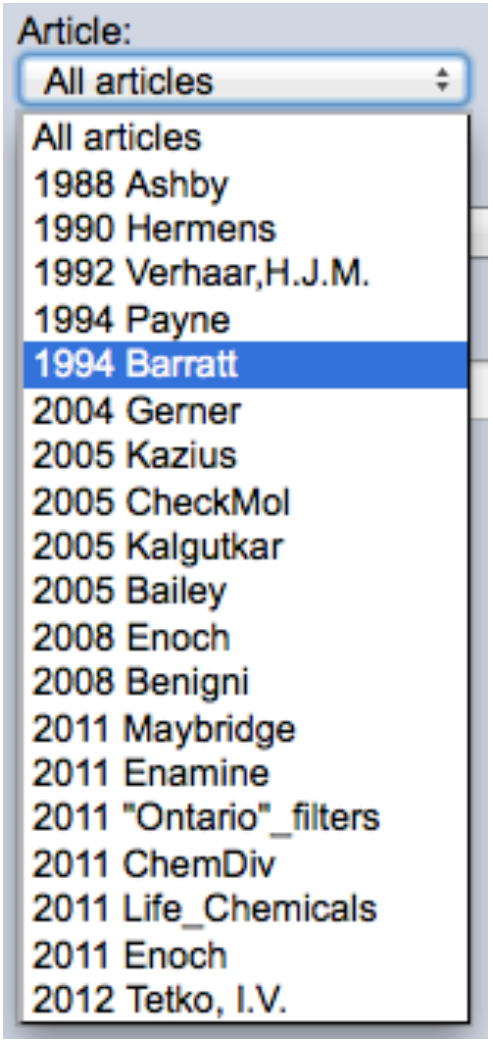
Baell and Holloway, J. Med. Chem., 2010, 53:2719-40.

Structural & Toxic Alerts at http://ochem.eu

- Screening of compounds against published groups, frequent hitters
- Filter alerts by endpoints or publications
- Create or upload custom SMARTS rules

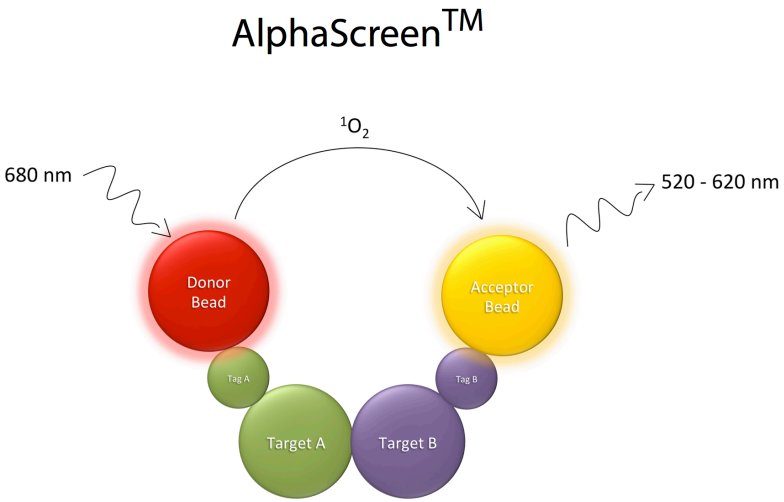
>500 functional groups
>2.3k alerts in total



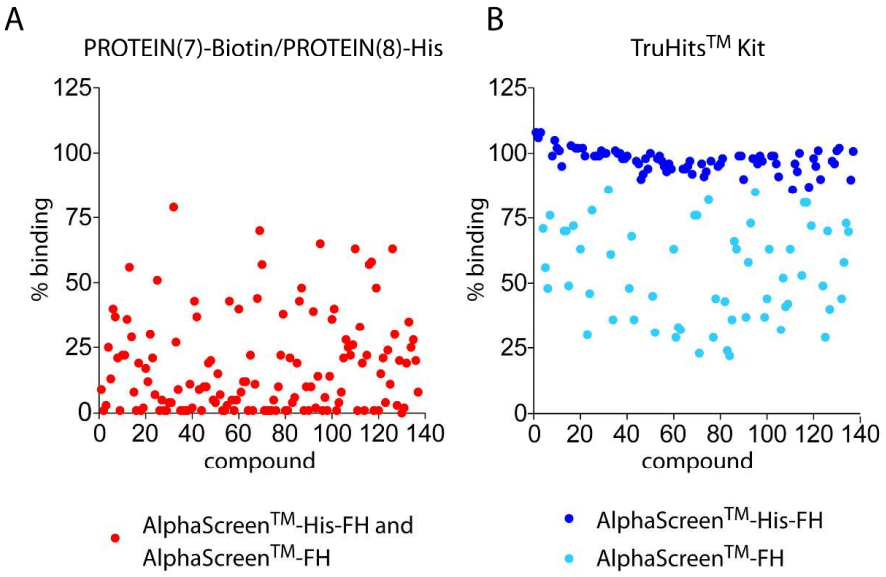


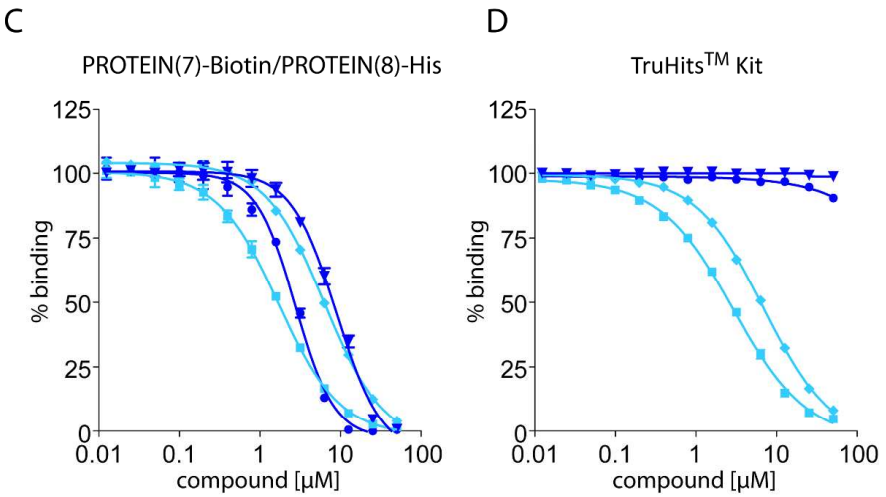
Sushko et al, J. Chem. Inf. Model, 2012, 52(8):2310-6.

Identification of AlphaScreen-HIS Frequent Hitters



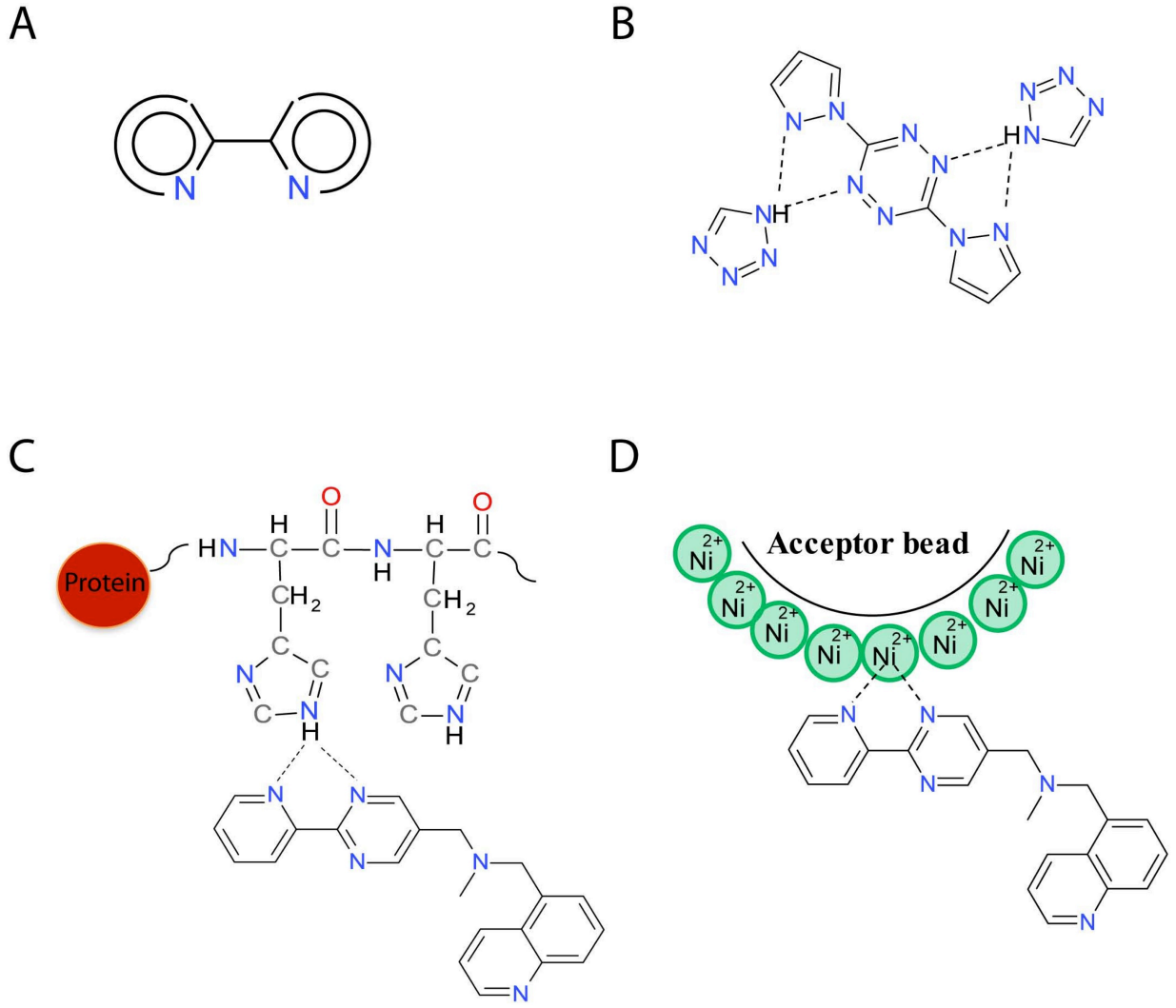
TrueHits™ :
Streptavidin Donor bead
Biotinylated Acceptor beads





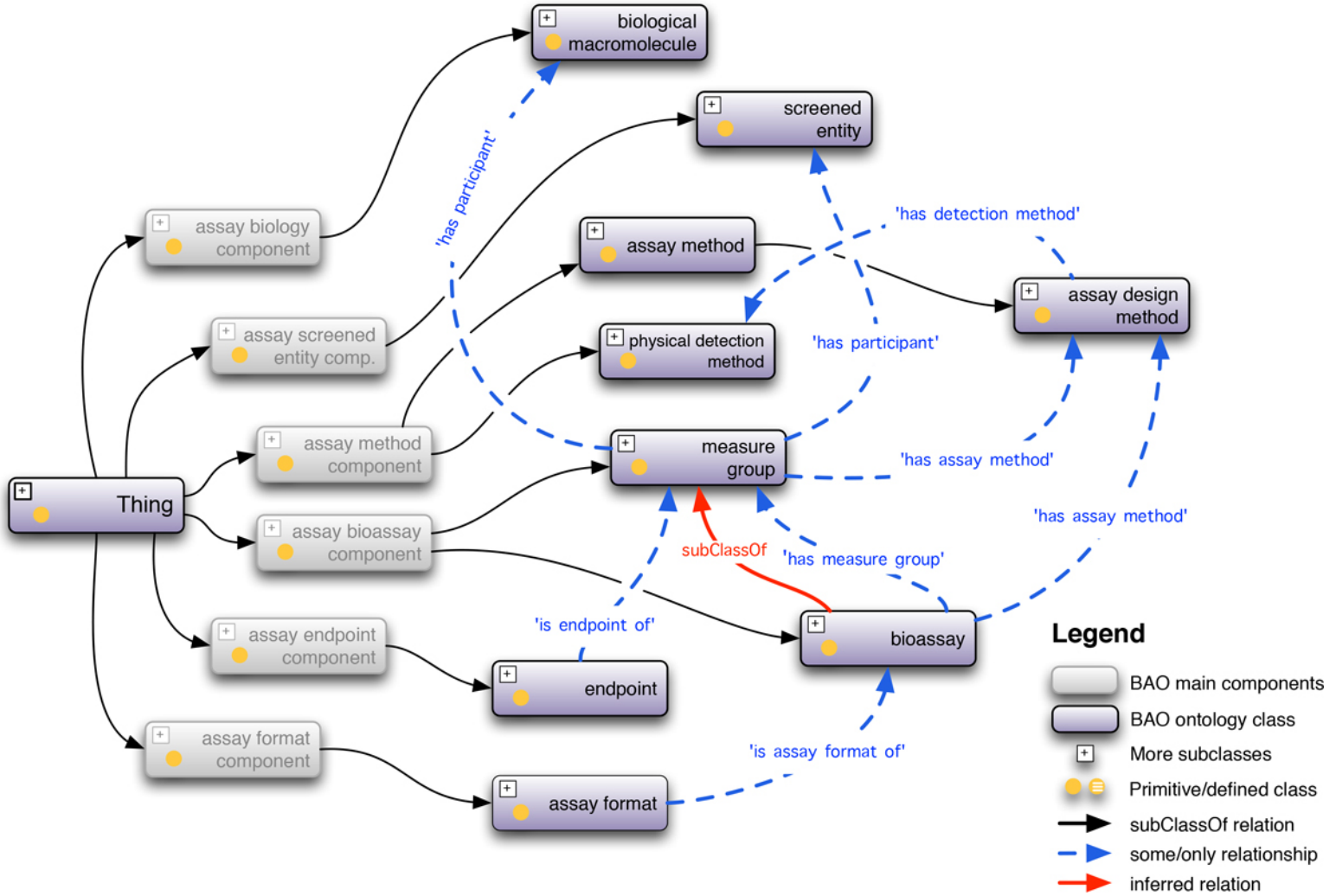
Schorpp et al J. Biomol. Screen. 2014, 9, 715-726.

Mode Of Action of AlphaScreen-HIS Frequent Hitters



Schorpp et al J. Biomol. Screen. 2014, 9, 715-726.

Bio Assays Ontology relationships

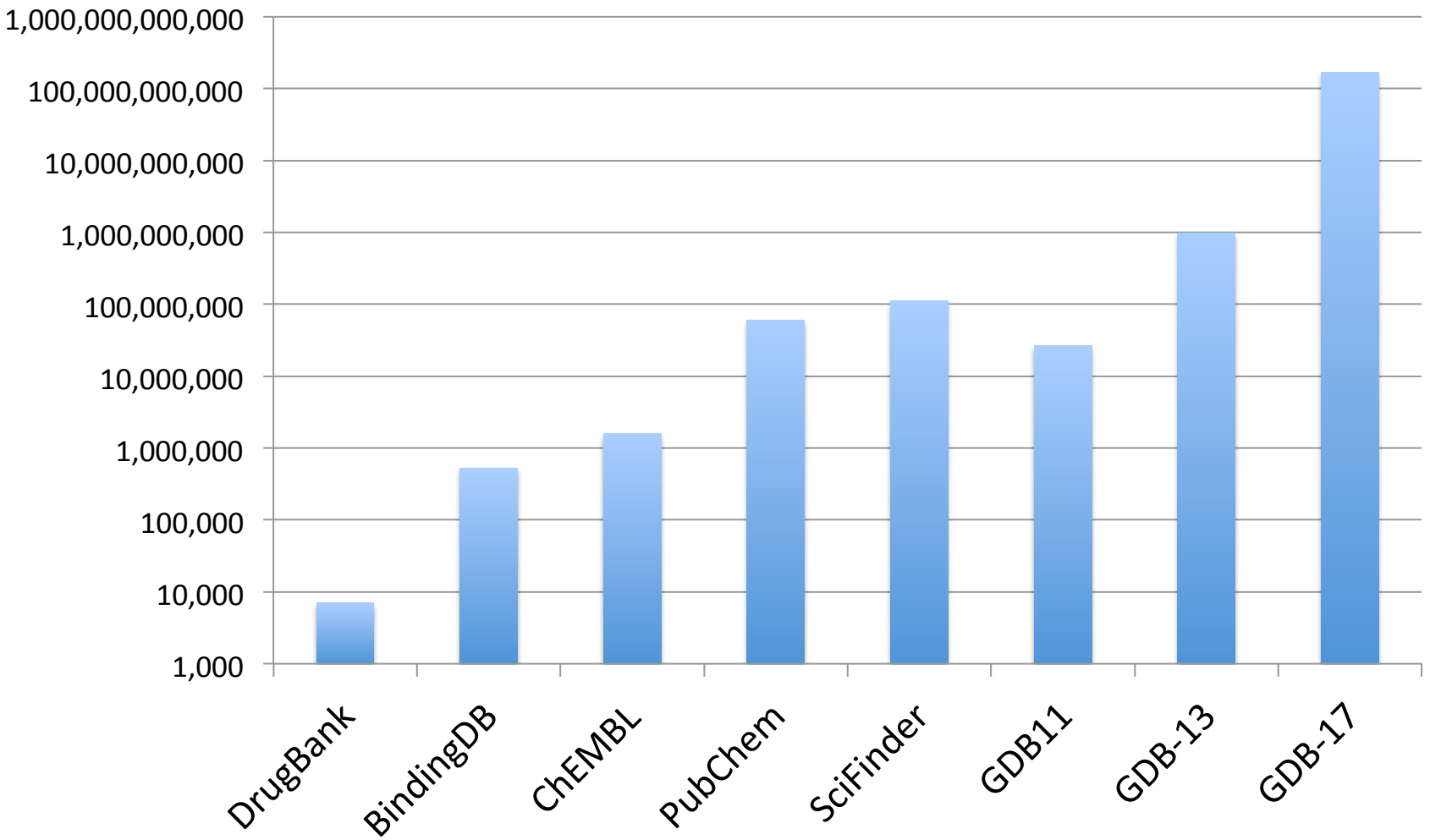


Abeyruwan, U. et al "Evolving BioAssay Ontology (BAO): Modularization, Integration and Applications," Journal of Biomedical Semantics, vol. 5, no. 1:S5, 2014.

Annotation of large chemical spaces

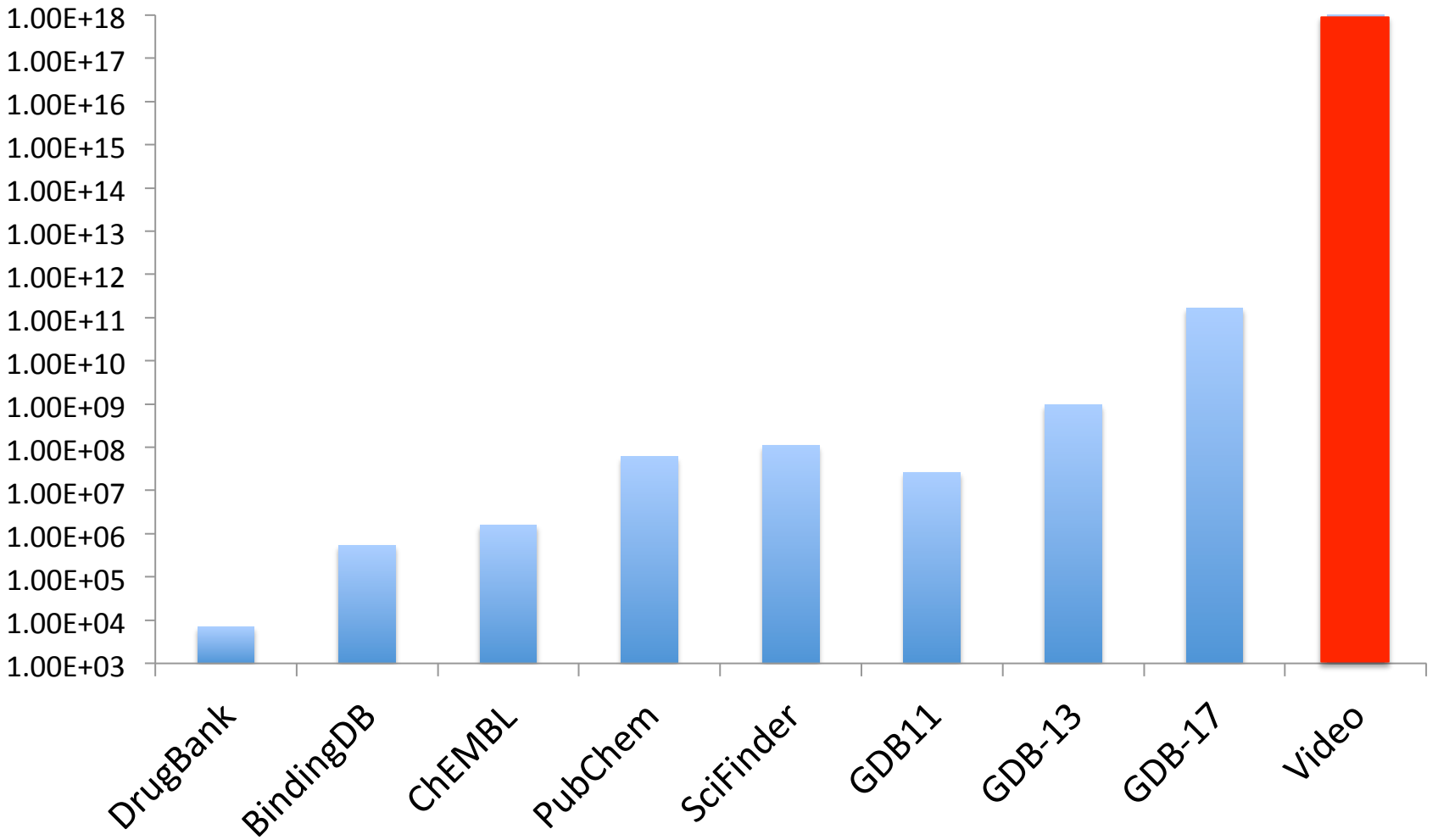
Big Data, which have been always in chemistry.

Virtual chemical spaces



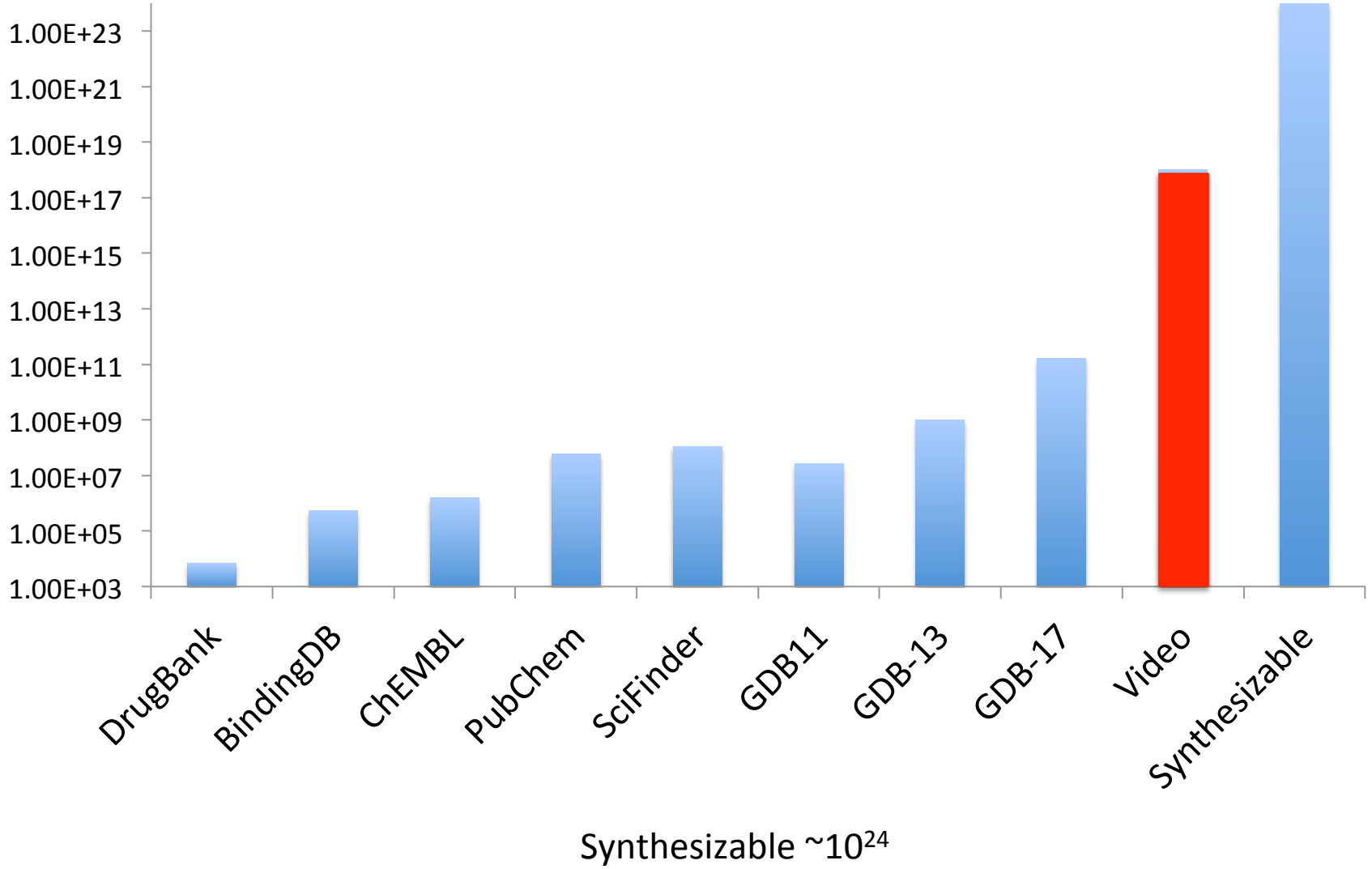
Synthesizable ~10^{24} and total space is ~ 10^{60}

Virtual chemical spaces



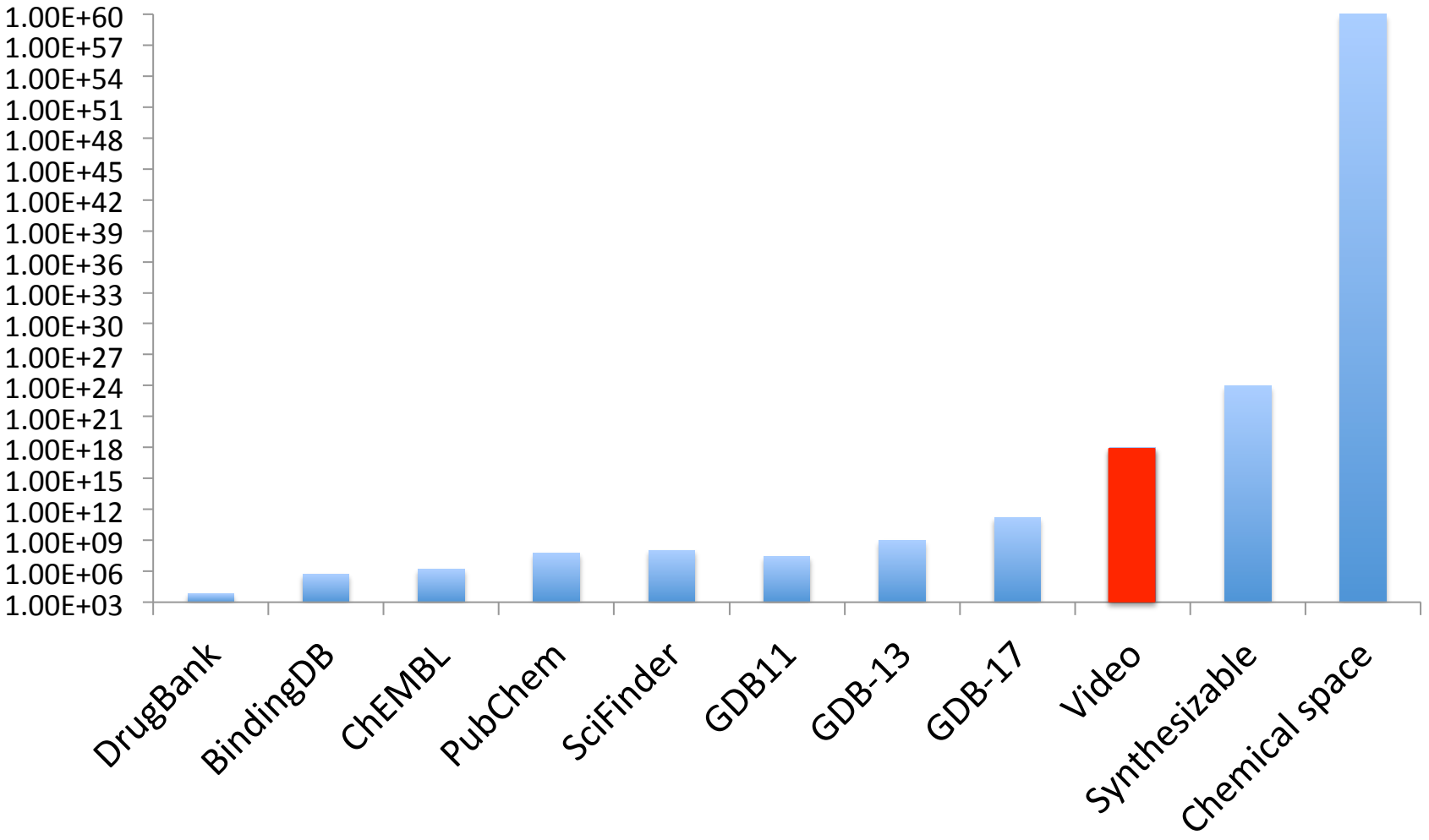
GDB*__N__ – all possible chemicals with ≤ **N atoms**

Virtual chemical spaces



Synthesizable ~10^{24}

Virtual chemical spaces



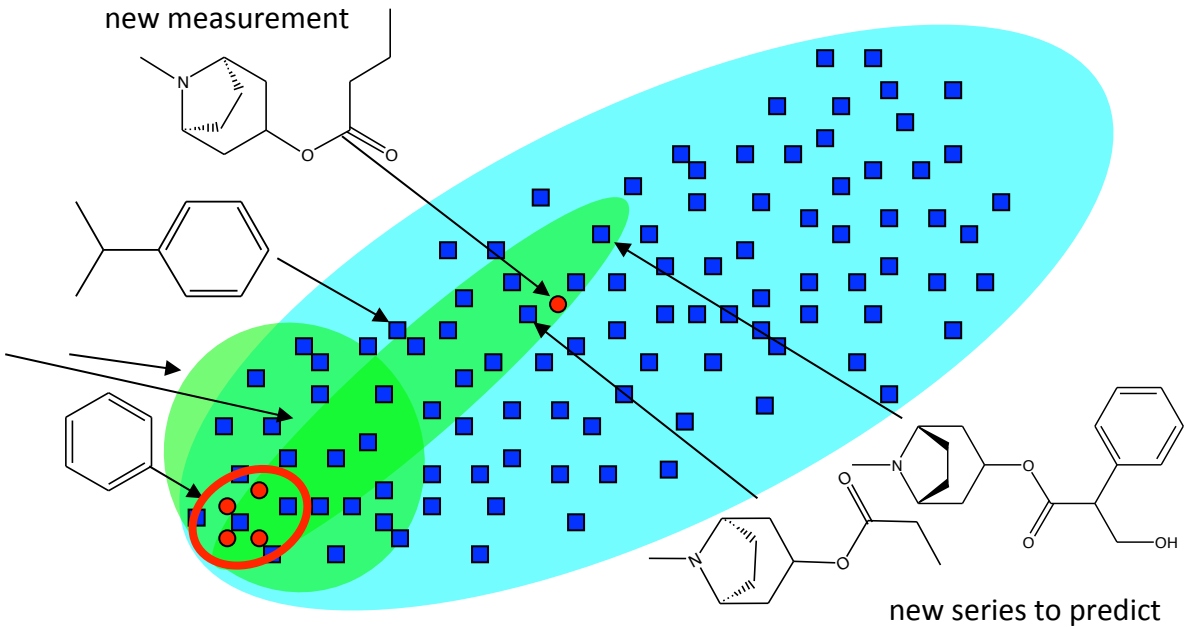
Synthesizable ~10^{24} and total "drug – like" space is ~ 10^{60}

Annotation of compounds

- ALOGPS 2.1* (prediction of logP and water solubility of chemical compounds)
- ~ 100,000 molecules per minute
- Annotation of GDB-17 will take ~3 years of calculations using one core
- ~10 minutes on Leibniz Supercomputing Centre with 241,000 cores

*Tetko, I.V. *J. Chem. Inf. Comput. Sci.* **2001**, 41, 1407-1421.

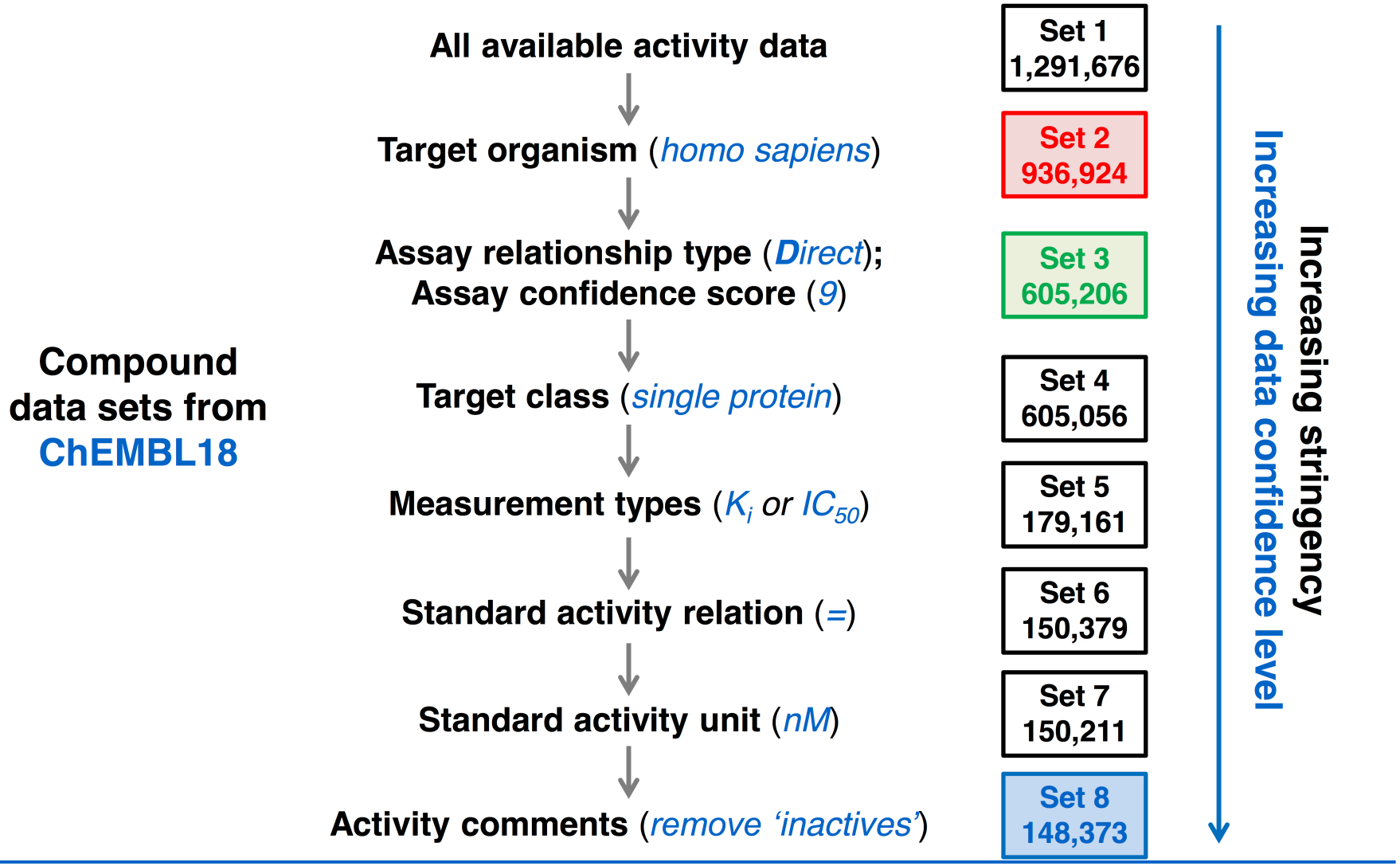
We can't predict unpredictable!



New machine learning approaches

Which methods can help us with Big Data?

Data Sets with Varying Confidence Levels



Courtesy of Prof. J. Bajorath

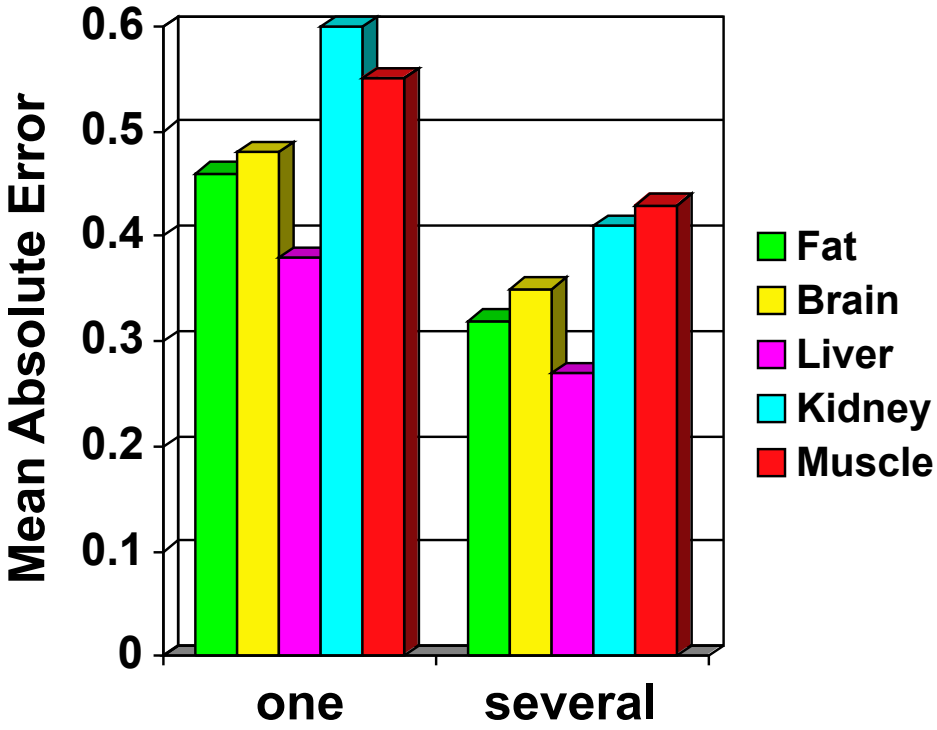
Multi-task learning

Problem:

• prediction of tissue-air partition coefficients
• small datasets 30-100 molecules (human & rat data)

Results:

simultaneous prediction of several properties increased the accuracy of models



Varnek, A. et al J. Chem. Inf. Model. 2009, 49, 133-44.

Renaissance of neural networks

Deep learning

- Massive neural networks with thousands of neurons and layers
- New learning methods (dropout technique)

Examples of the use of deep learning technology:

- Recognition of Chinese characters with human accuracy
- Victory in Go-tournament
- Diagnostics of breast cancer

Baskin, I.I.; Winkler, D.; Tetko, I.V. A renaissance of neural networks in drug discovery. *Expert opinion on drug discovery* **2016**, 11(8):785-95.

Massively Multitask Networks for Drug Discovery

Bharath Ramsundar[*,†,°] RBHARATH@STANFORD.EDU
Steven Kearnes[*,†] KEARNES@STANFORD.EDU
Patrick Riley[°] PFR@GOOGLE.COM
Dale Webster[°] DRW@GOOGLE.COM
David Konerding[°] DEK@GOOGLE.COM
Vijay Pande[†] PANDE@STANFORD.EDU
([*]Equal contribution, [†]Stanford University, [°]Google Inc.)

259 datasets
- 128 PubChem
- 17 MUV
- 102 DUD-E
- 12 Tox21

Descriptors:
ECFP4
RDKit

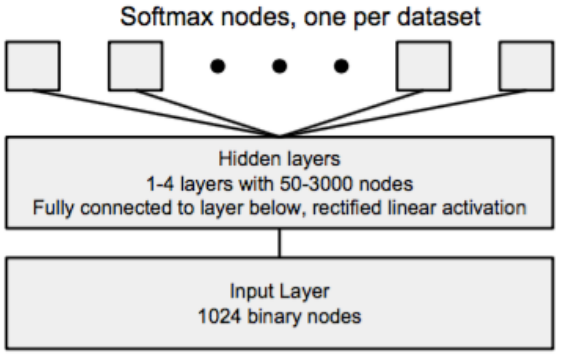


Figure 1. Multitask neural network.

Total ~ 40M datapoints for 1.6M compounds

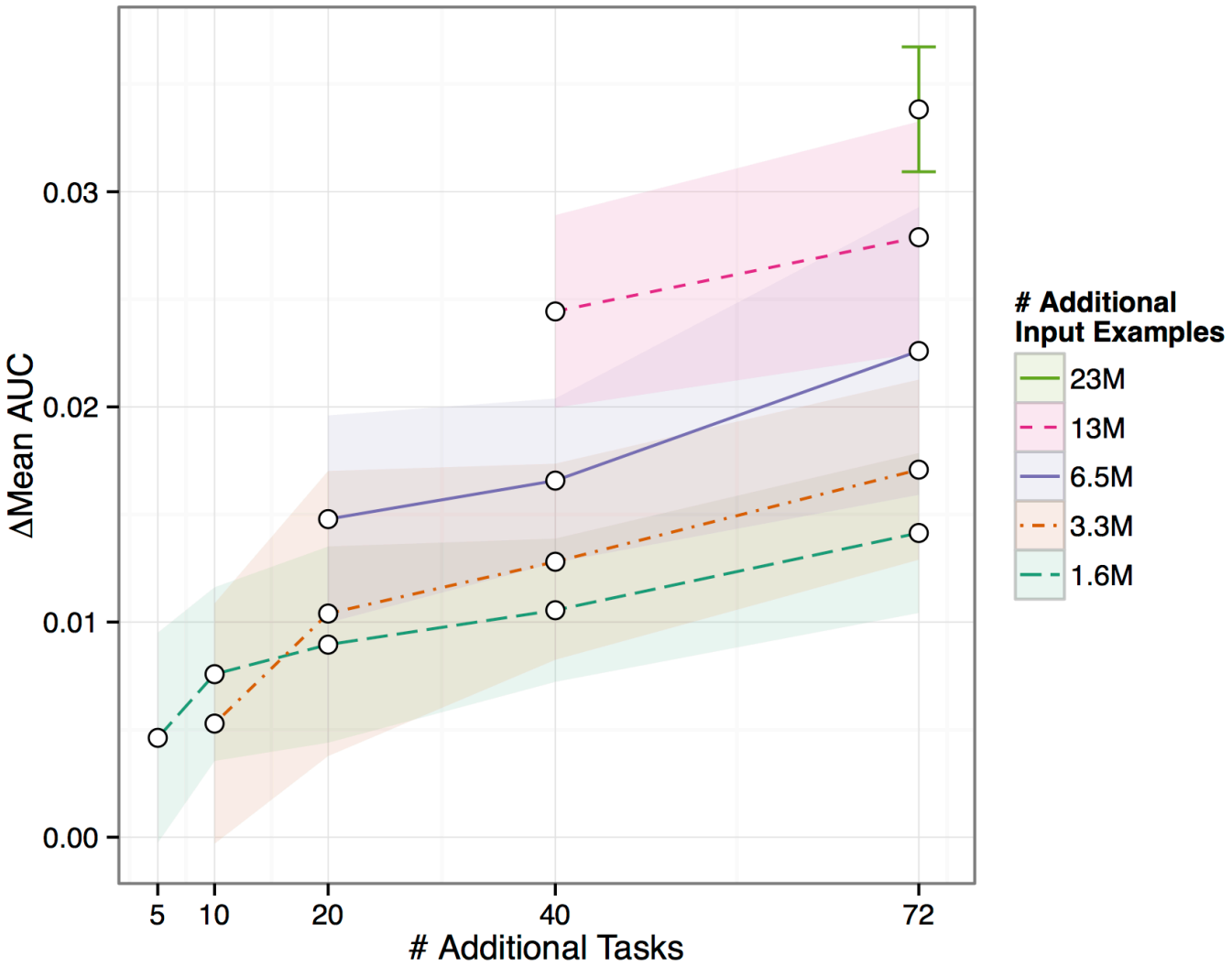
http://adsabs.harvard.edu/abs/2015arXiv150202072R

Multitask Networks Learning Results

- Massively multitask networks obtain predictive accuracies significantly better than single-task methods.
- The predictive power of multitask networks improves as additional tasks and data are added.
- The total amount of data and the total number of tasks both contribute significantly to multitask improvement.
- Multitask networks afford limited transferability to tasks not in the training set.

http://adsabs.harvard.edu/abs/2015arXiv150202072R

Multitask benefit from increasing tasks and data independently.



http://adsabs.harvard.edu/abs/2015arXiv150202072R

Secure Information Sharing

How can we share information but not data?

How can we enable cooperation between industries?

Secure Sharing of information

- CINF/COMP workshop was organized during ACS in 2005 by Prof. Oprea
- Various structure representation (descriptors) were proposed
- Several methods for secure sharing were introduced

- But in the theoretical limit*
  - SMILES representation of molecules: CCC, CNCCC, c1ccccc1
  - Zipping of structures requires < 1 bit per atom
  - Structure with 32 atoms requires < 32 bits
  - Any descriptor or their combination with > 32 bits could be used to decode a molecule (in theory)

*Tetko, I.V.; Abagyan, R.; Oprea, T.I. J. Comput. Aided. Mol. Des. 2005, 19, 749-764.

Currently used technologies

"Honest broker"

- Receives descriptors (or structures)
- Develop models and do not reveal the underlying data

Sharing relationships between structures

- Matched Molecular Pairs (changes in property due to change of groups)

Sharing developed models

- Structural alerts
- Computational prediction models

Sharing reliable predictions (surrogate data)*

*Tetko, I.V.; Abagyan, R.; Oprea, T.I. J. Comput. Aided. Mol. Des. 2005, 19, 749-764.

Multi-party secure computation

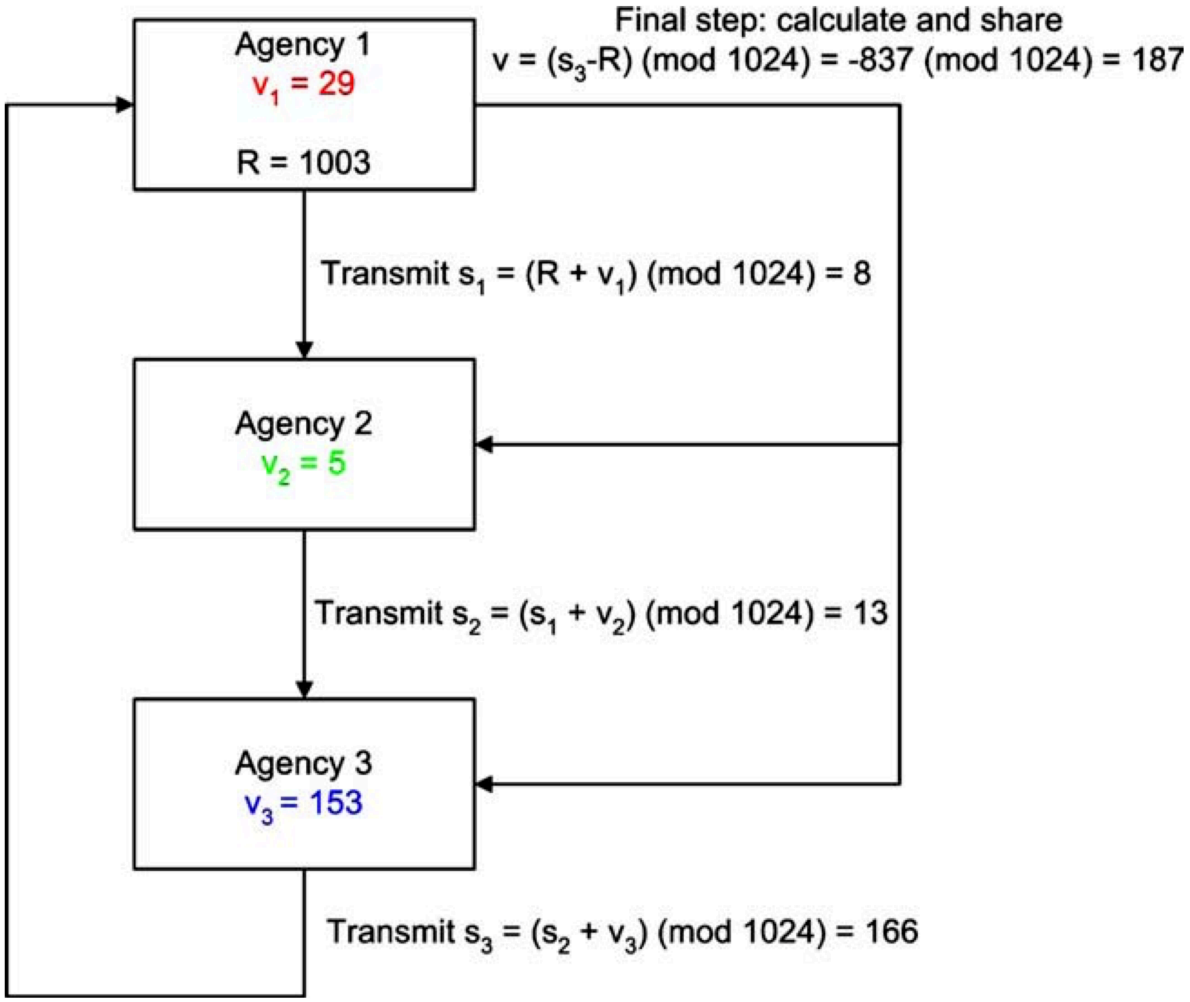
Journal of Computer-Aided Molecular Design (2005) 19: 739–747
DOI 10.1007/s10822-005-9011-5

Secure analysis of distributed chemical databases without data integration

Alan F. Karr[a,*], Jun Feng[a], Xiaodong Lin[a], Ashish P. Sanil[a], S. Stanley Young[a] & Jerome P. Reiter[b]

[a]*National Institute of Statistical Sciences Research, Triangle Park, NC 27709-4006, USA;* [b]*Duke University, Durham, NC 27708, USA;* [c]*University of Cincinnati, Cincinnati, OH USA;* [d]*Bristol-Myers Squibb, Princeton, NJ USA*

Secure summation



Conclusions

Expectations

- ✓ Improved prediction of properties, and activities
- ✓ Improved poly-pharmacology
- ✓ Search of new chemistry (top down exploration and *de novo* design)
- ✓ Prediction of *in vivo* enpoints

Challenges

- ✓ New machine learning approaches (deep learning)
- ✓ Integration of diverse data and *a priory* knowledge (ontology, pathways, *in vitro, in vivo,* simulation results, different errors, etc.)
- ✓ Applicability domain
- ✓ Secure data sharing
- ✓ Data visualization
- ✓ De novo design

Further reading

- Tetko, I. V.; Engkvist, O.; Koch, U.; Reymond, J. L.; Chen, H., BIGCHEM: Challenges and Opportunities for Big Data Analysis in Chemistry. *Mol Inform* **2016**, 35(11-12):615-621 (Open Access).
- Tetko, I.V.; Engkvist, O.; Chen, H. Does 'Big Data' exist in medicinal chemistry, and if so, how can it be harnessed? *Future Med Chem*. **2016** 8(15):1801-1806 (Open Access).

Acknowledgements

Dr. O. Engkvist (AZ)
Dr. H. Chen (AZ)
Dr. U. Koch (LDC)
Prof. J.-L. Reymond (Uni Bern)
Dr. I. Baskin (MSU)
Dr. D. Winkler (CSIRO)
And BIGCHEM partners

Dr. Y. Sushko
Dr. S. Novotarskyi
Mr. R. Körner
Mrs. E. Salmina
Dr. K. Hadian (TOX, HMGU)
Dr. A. Williams (USA)
Dr. D. Lowe (UK)