An Updated Comparison of Selected Public and Commercial Bioactive Chemistry Databases

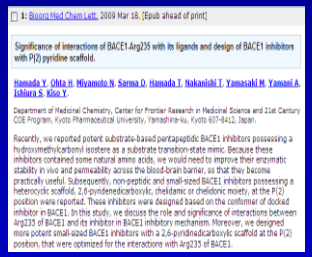
Christopher Southan

The International Conference for Science & Business Information
Sitges, Spain, October 2009

ChrisDS Consulting

http://www.cdsouthan.info/Consult/CDS_cons.htm

Entity Relationships:
in vitro activity-to-compound-to-protein mapping

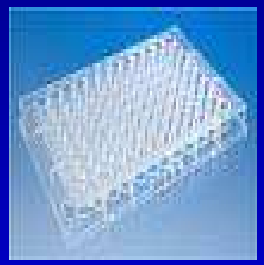

1: Bioorg Med Chem Lett. 2009 Mar 18. [Epub ahead of print]

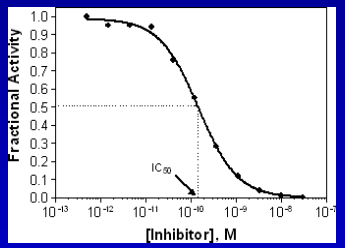
Significance of interactions of BACE1 Arg235 with its ligands and design of BACE1 inhibitors with P(2) pyridine scaffold.

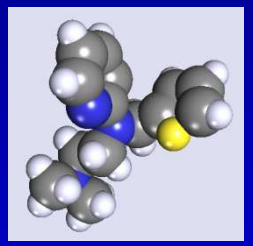
Hanada Y, Ohta H, Miyamoto N, Sarma D, Hamada T, Nakanishi T, Yamasaki M, Yamani A, Ishiura S, Kiso Y.

Department of Medicinal Chemistry, Center for Frontier Research in Medicinal Science and 21st Century COE Program, Kyoto Pharmaceutical University, Yamashina-ku, Kyoto 607-8412, Japan.

Recently, we reported potent substrate-based pentapeptidic BACE1 inhibitors possessing a hydroxymethylcarbonyl isostere as a substrate transition-state mimic. Because these inhibitors contained some natural amino acids, we would need to improve their enzymatic stability in vivo and permeability across the blood-brain barrier, so that they become practically useful. Subsequently, non-peptidic and small-sized BACE1 inhibitors possessing a heterocyclic scaffold, 2,6-pyridinedicarboxylic, chelidamic or chelidonic moiety, at the P(2) position were reported. These inhibitors were designed based on the conformer of docked inhibitor in BACE1. In this study, we discuss the role and significance of interactions between Arg235 of BACE1 and its inhibitor in BACE1 inhibitory mechanism. Moreover, we designed more potent small-sized BACE1 inhibitors with a 2,6-pyridinedicarboxylic scaffold at the P(2) position, that were optimized for the interactions with Arg235 of BACE1.







MAQALPWLLLWMGAGVLPAHGTQHGIRLPLRSGLGG
APLGLRLPRETDEEPEEPGRRGSFVEMVDNLRGKSGQ
GYYVEMTVGSPPQTLNILVDTGSSNFAVGAAPHPFLHR
YYQRQLSSTYRDLRKGVYVPYTQGKWEGELGTDLVSI
PHGPNVTVRANIAAITESDKFFINGSNWEGILGLAYAEI
ARPDDSLEPFFDSLVKQTHVPNLFSLQLCGAGFPLNQSE
VLASVGGSMIIGGIDHSLYTGSLWYTPIRREWYYEVIIV
RVEINGQDLKMDCKEYNYDKSIVDSGTTNLRLPKKVFE
AAVKSIKAASSTEKFPDGFWLGEQLVCWQAGTTPWNI
FPVISLYLMGEVTNQSFRITILPQQYLRPVEDVATSQDD
CYKFAISQSSTGTVMGAVIMEGFYVVFDRARKRIGFAV
SACHVHDEFRTAAVEGPFVTLDMEDCGYNIPQTDESTL
MTIAYVMAAICALFMLPLCLMVCQWRCLRCLRQQHD
DFADDISLLK

Document → Assay → Result → Compound → Sequence

Unstructured data → Structured data

Expert extraction and curation

[2]

Databases of Bioactive Compounds

Public

DrugBank

PubChem

Commercial

GVK BIO

Dictionary of Natural Products

MDL Drug Data Report

WOMBAT

WOrld of Molecular BioAcTivity

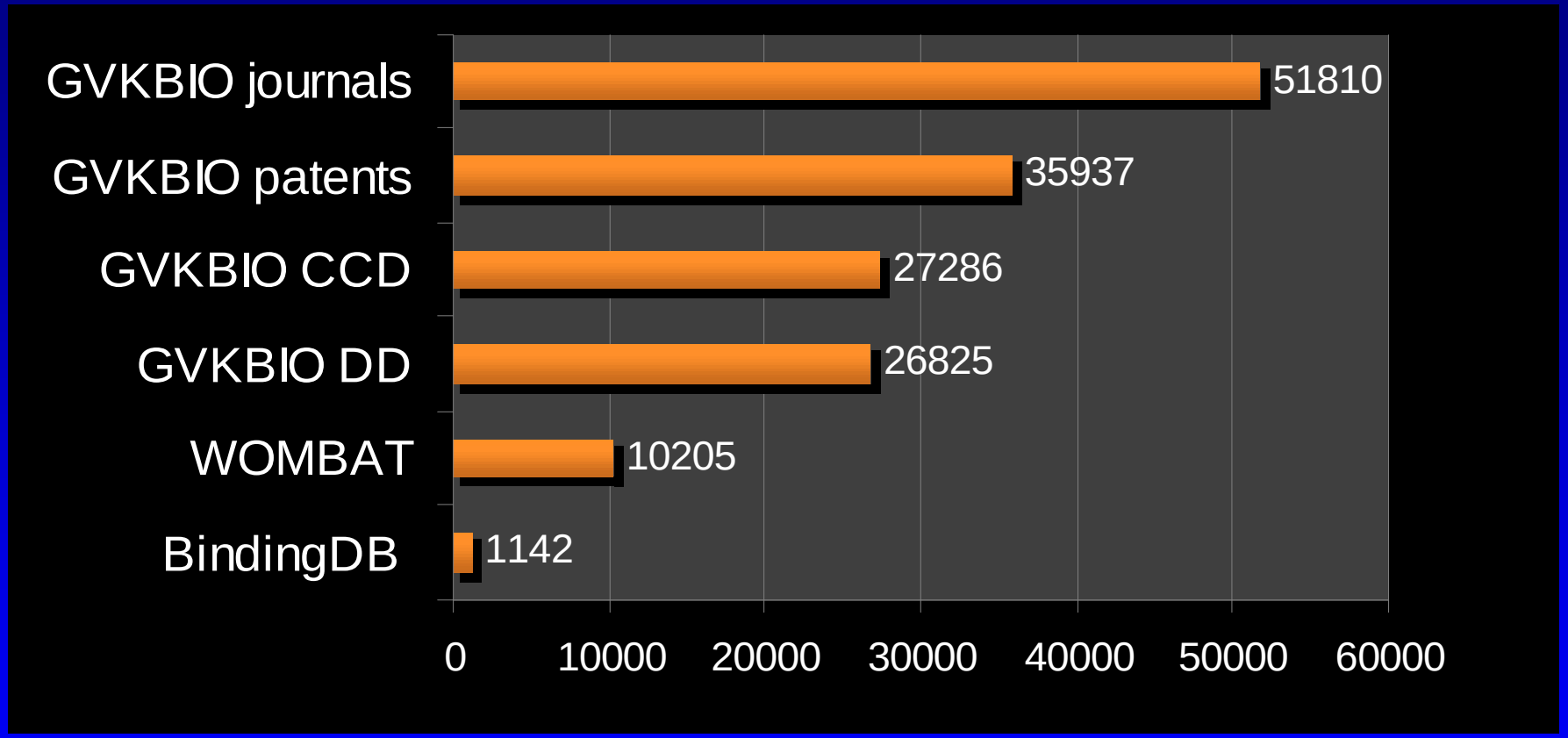
Comparing Compound Sets

- Filter sources and subsets to normalise compound content
- Compare protein mappings, document counts, and compound ratios
- Produce an all-vs.-all compound overlap matrix
- Review overlap and content differences for 2008
- Compare between 2006 and 2008
- Make selected Venn-type comparisons
- Compare selected larger merges

Filtration of Sources and Subsets

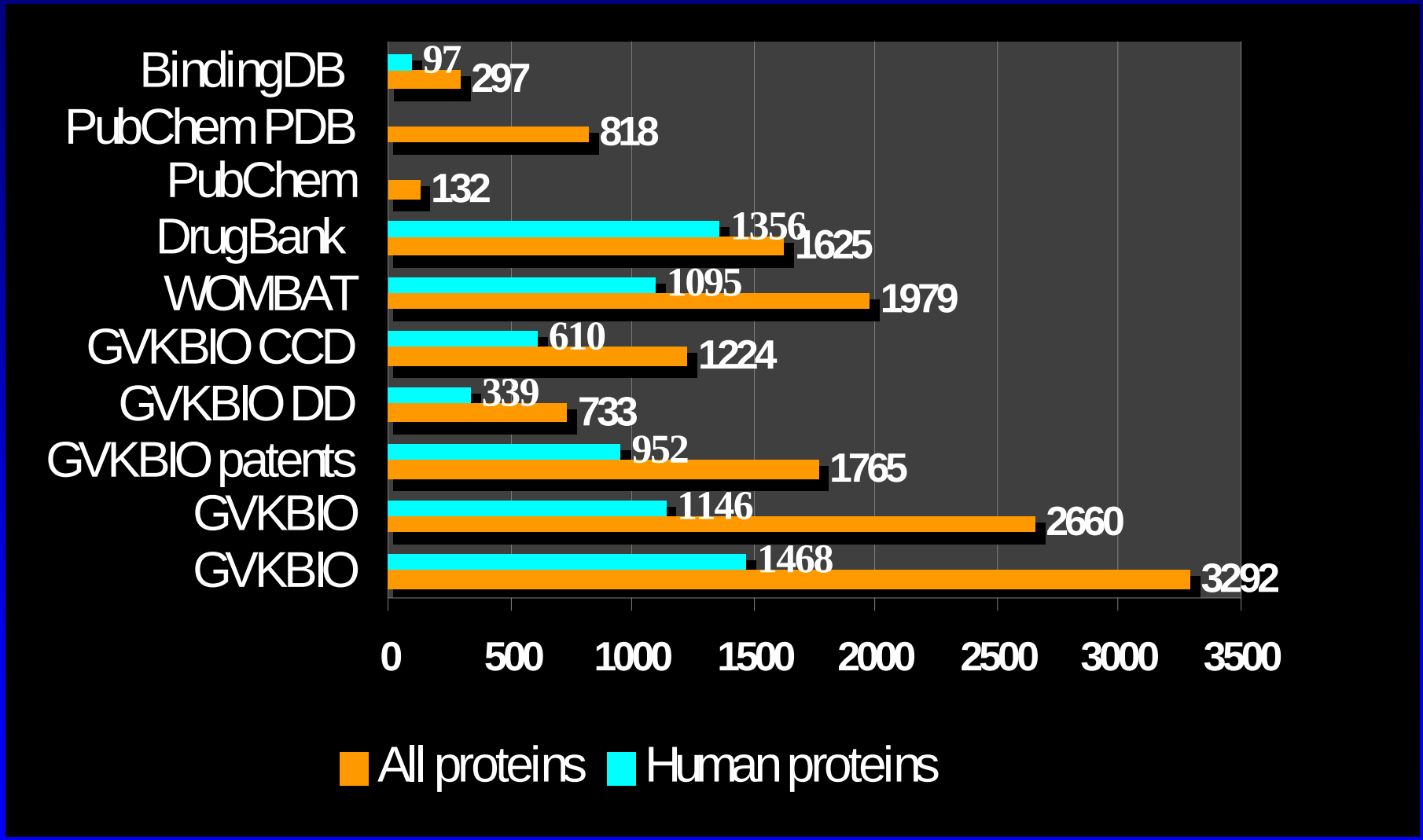
Dataset	Filtered cpds	Filtration reduction
GVKBio	2,054,151	-8%
GVKBio Journals	658,198	-8%
GVKBio patents	1,484,218	-7%
GVKBIO DD	3,675	-4%
GVKBIO CCD	8,864	-1%
GVKBIO BACE1	5,228	-11%
GVKBIO BACE1 journals	389	-6%
GVKBIO BACE1 patents	4,901	-11%
WOMBAT	180,856	-18%
PubChem	14,965,539	-23%
PubChem Prous	4,652	-2%
PubChem PDB	5,706	-8%
PubChem actives	7,472	-3%
PubChem pharmacol	5,311	-63%
PubChem MLSMR	233,284	-1%
PunChem BindingDB	24,203	-4%
PubChem ChEBI	7,428	-31%
DrugBank all	4,545	-7%
DrugBank approved	1,341	-3%
DrugBank experimental	2,999	-6%
DNP	144,383	-26%
MDDR	176,600	-4%
MDDR launched	1,435	-5%

Document Counts

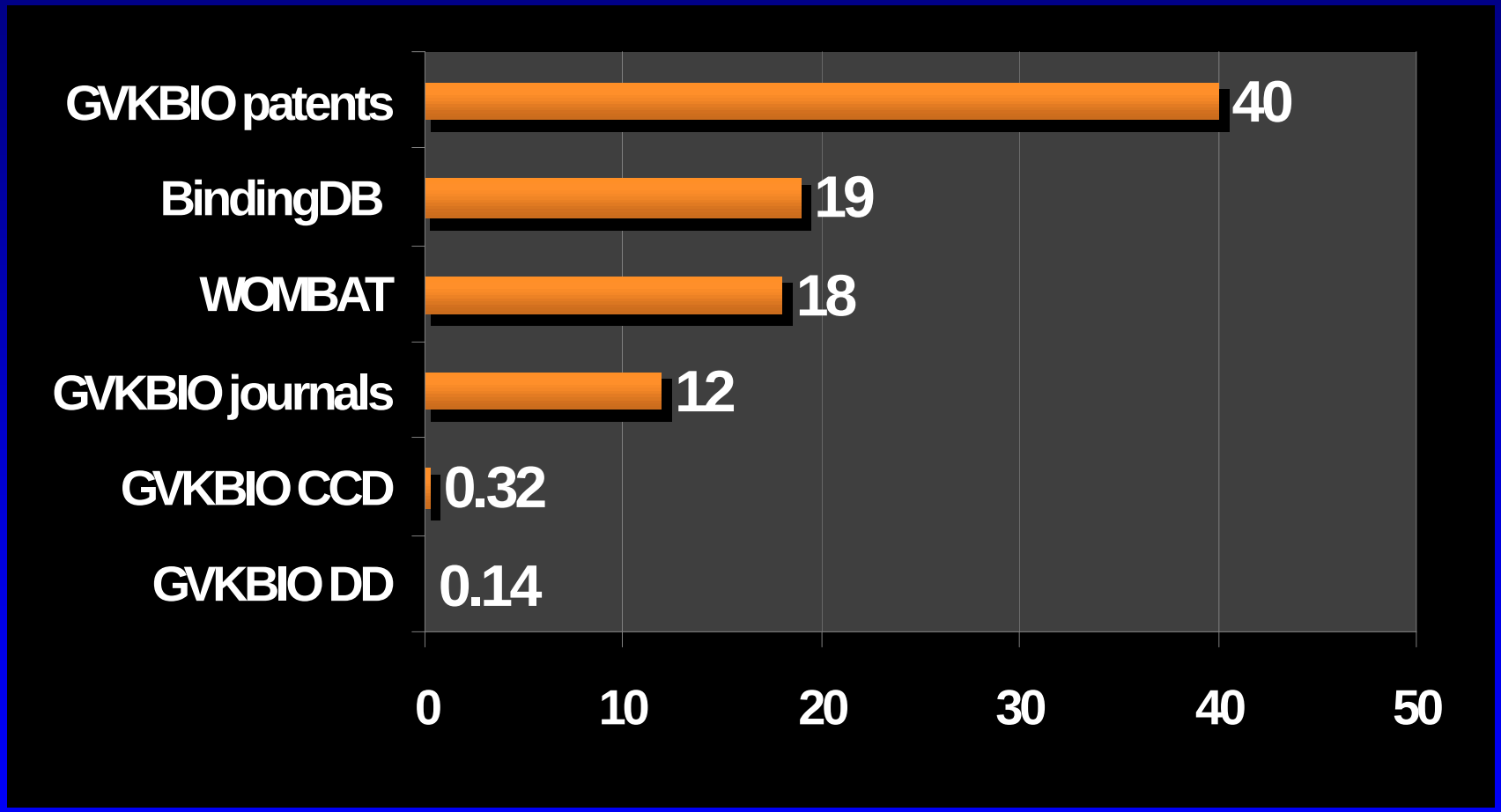


[6]

Protein Counts



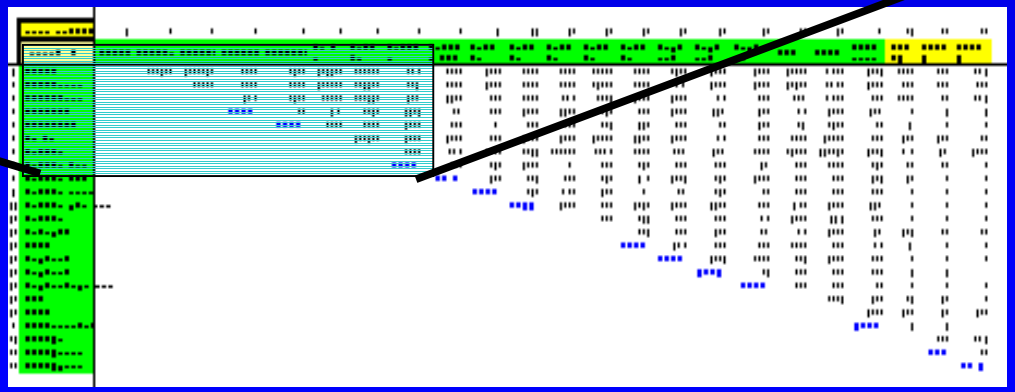
Compounds-per-protein



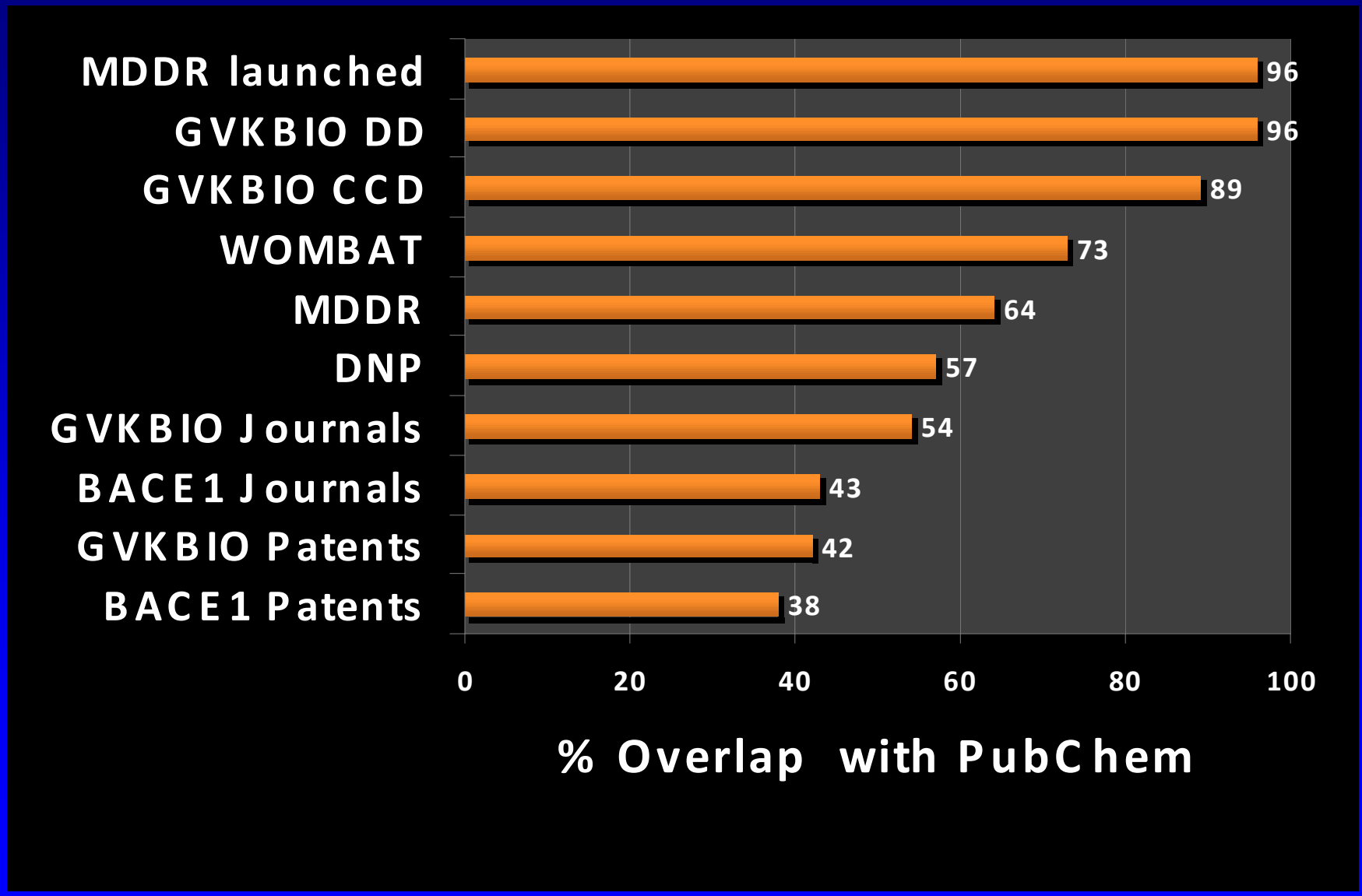
[8]

Pair-wise Comparison Matrix: 23 X 23

	GVKBIO	GVKBIO Journals	GVKBIO Patents	GVKBIO DD	GVKBIO CCD	WOMBAT	PubChem
GVKBIO	2,054,151	658,198	1,484,218	2,847	6,178	171,178	925,845
GVKBIO Journals		658,198	88,265	2,779	5,492	169,734	361,192
GVKBIO Patents			1,484,218	1,404	3,149	45,564	633,115
GVKBIO DD				3,675	33	1,060	3,513
GVKBIO CCD					8,864	2,652	7,925
WOMBAT						180,856	133,124
PubChem							14,965,539

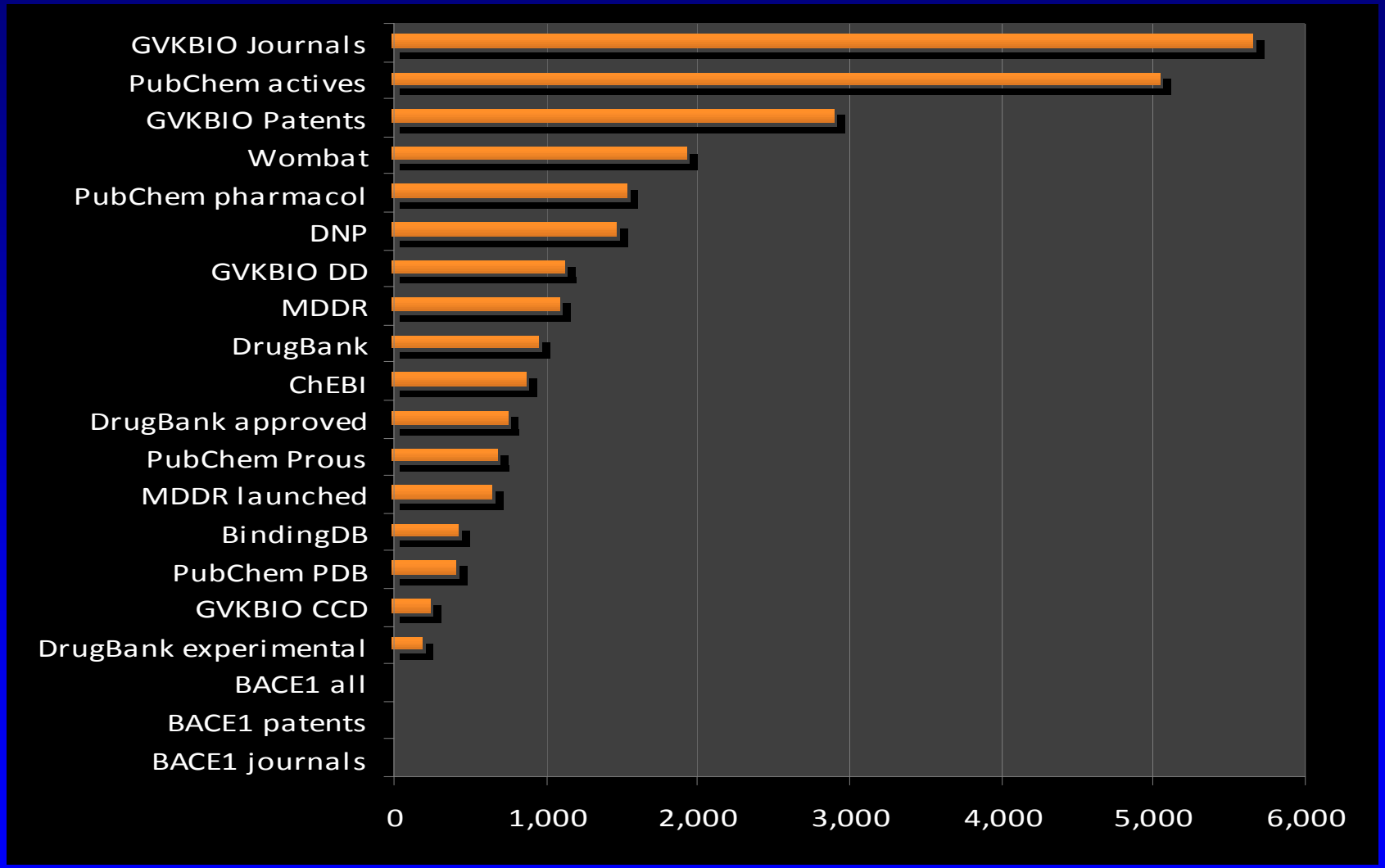


Coverage of Commercial Databases by PubChem



[10]

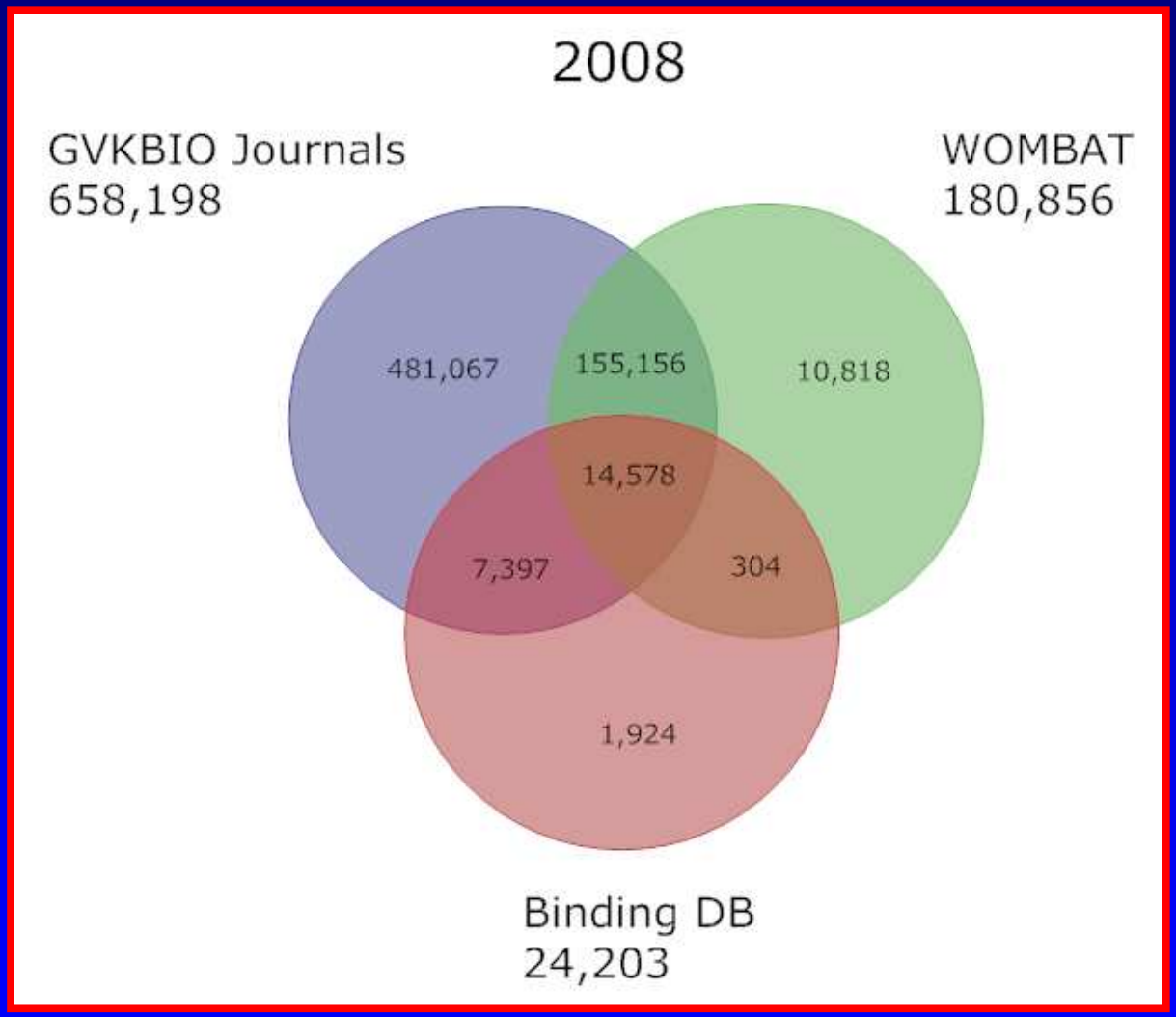
Molecular Libraries-Small Molecule Repository



MLSMR 233,284, PubChem actives 7,472

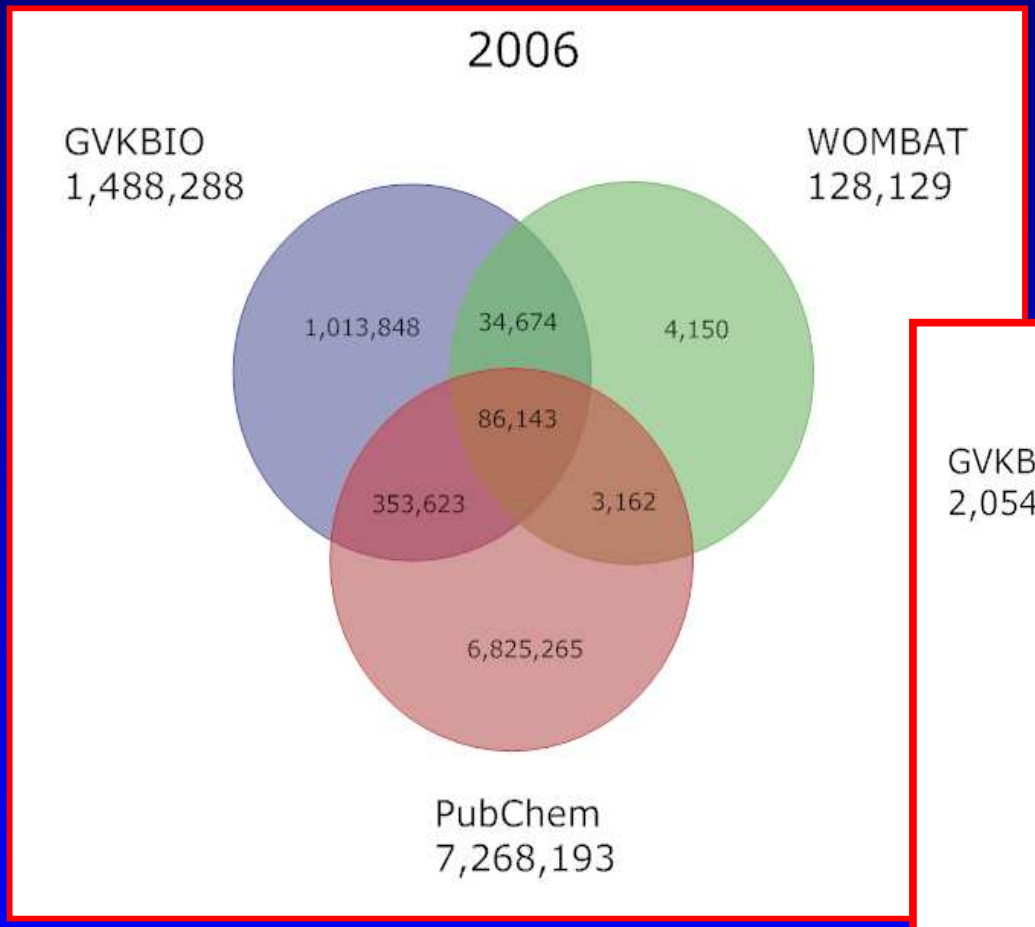
[11]

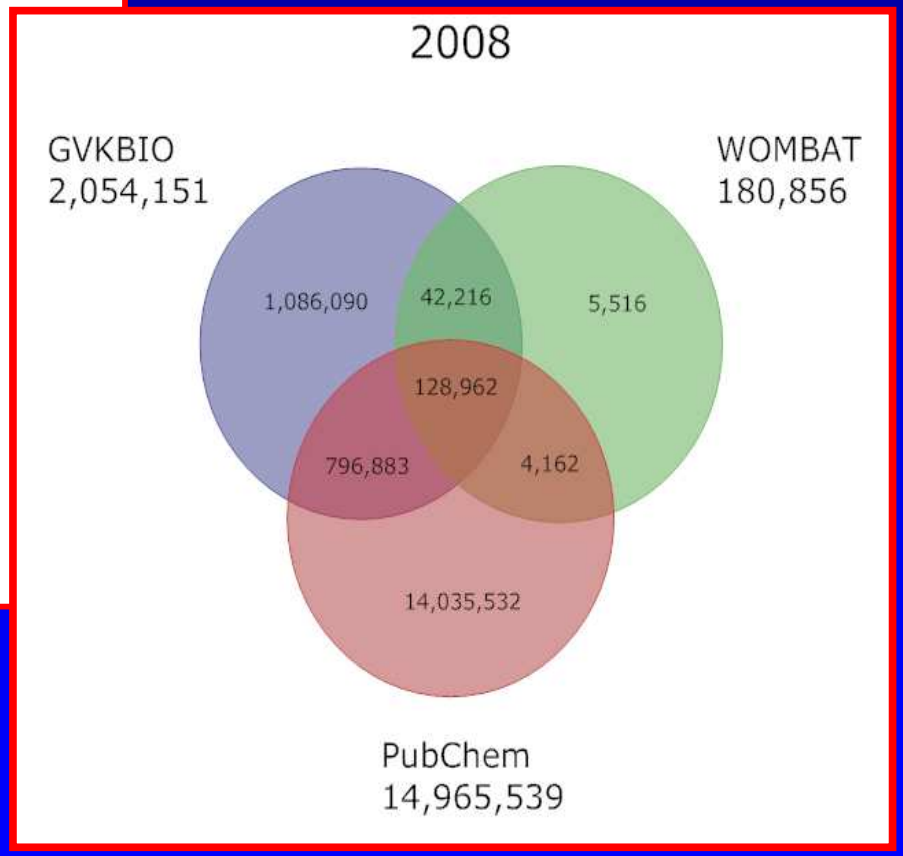
Comparison of Journal Extractions



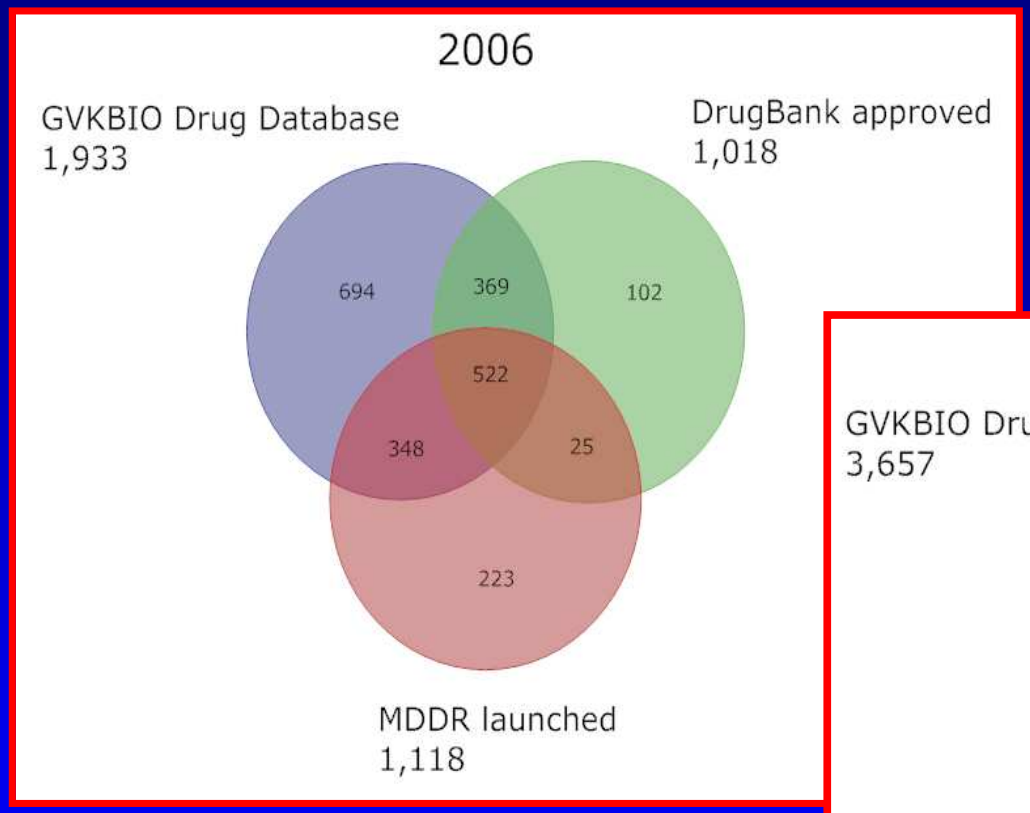
Document ratios GVK:WOM:BDb 50:9:1

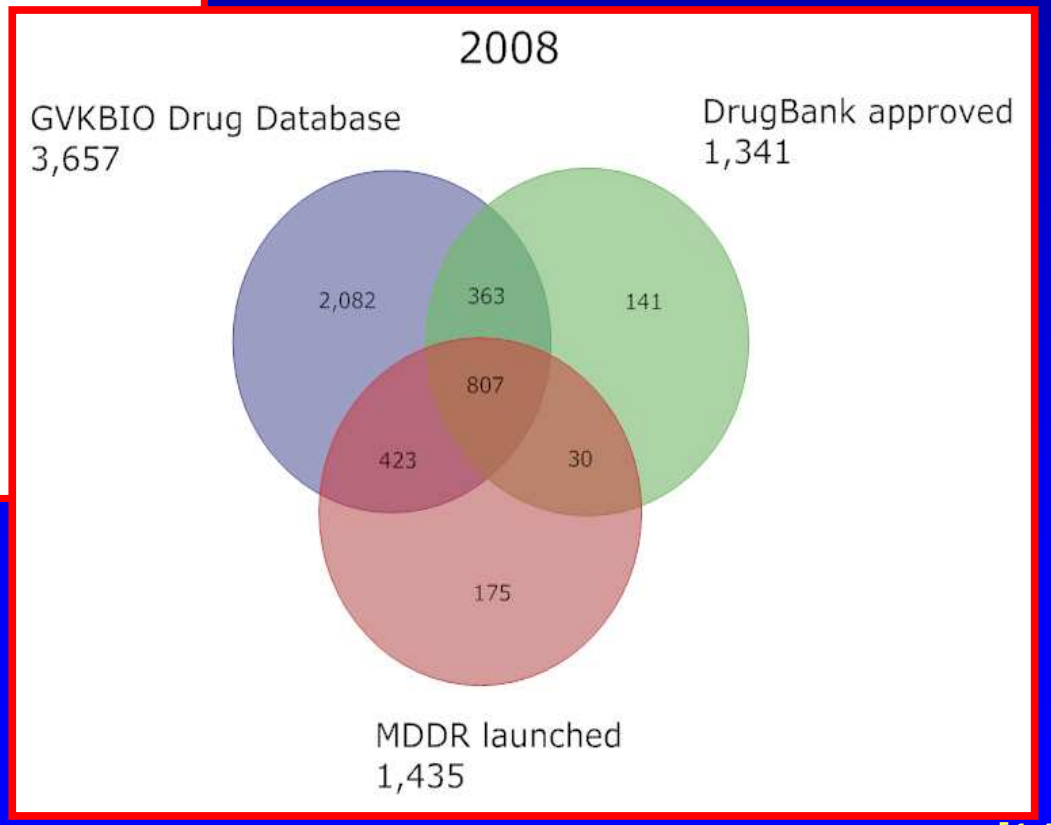
GVKBIO vs WOMBAT vs PubChem



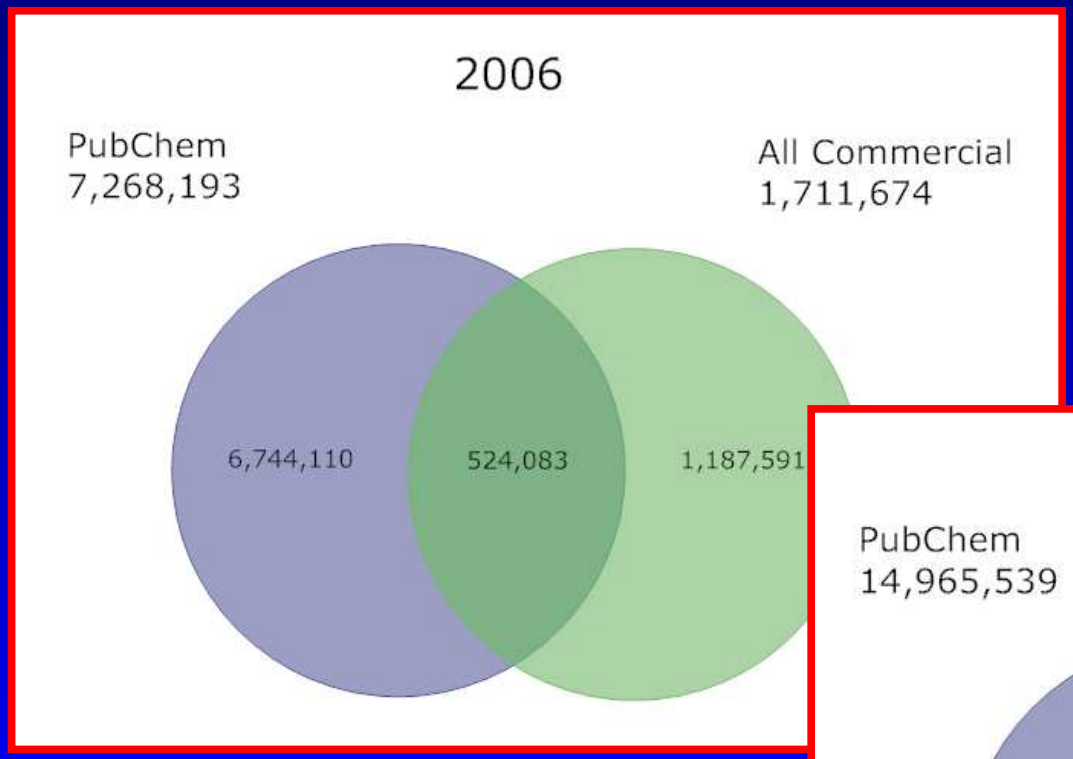


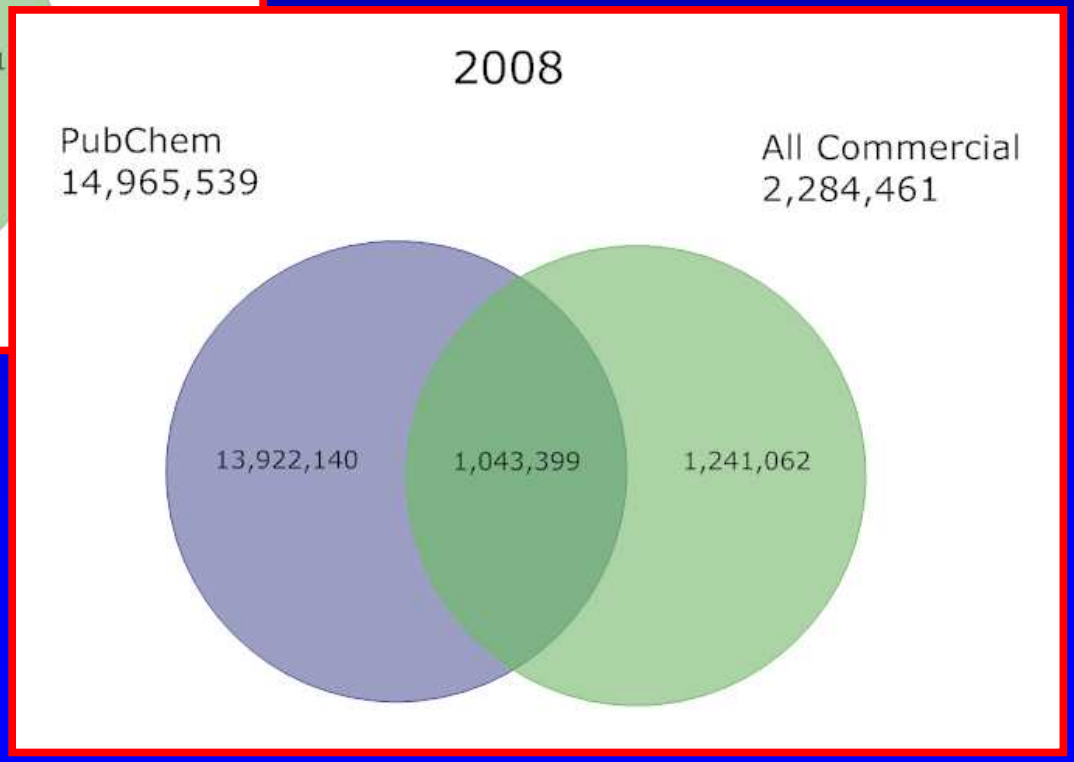
Comparison of Approved Drug Collections





Public vs Commercial Total Merges





Conclusions

- Database utility assessments are inadequate without direct comparisons of compound content, document counts and activity-to-protein mapping counts
- Both shared and unique content provide value
- Based on compound content *per se* the pendulum is swinging in the public direction
- But journal and patent compound-assay-protein mapping is still covered at a larger scale by commercial databases
- Public sources have essential complementarity to commercial ones for the exploration of bioactive chemical space
- Users can get the best of both worlds

References and Acknowledgments

1502 *Current Topics in Medicinal Chemistry*, 2007, 7, 1502-1508

Complementarity Between Public and Commercial Databases: New Opportunities in Medicinal Chemistry Informatics

Christopher Southan*, Péter Várkonyi and Sorel Muresan*

PMID: 17897036

Research article Open Access Highly accessed

Quantitative assessment of the expanding complementarity between public and commercial databases of bioactive compounds

Christopher Southan, Péter Várkonyi, Sorel Muresan

Journal of Cheminformatics 2009, **1**:10 (6 July 2009)

www.jcheminf.com/content/1/1/10

Thanks to: *Tudor Oprea, Steve Byant, Paul Thiessen, Yanli Wang and Jens Sadowski*