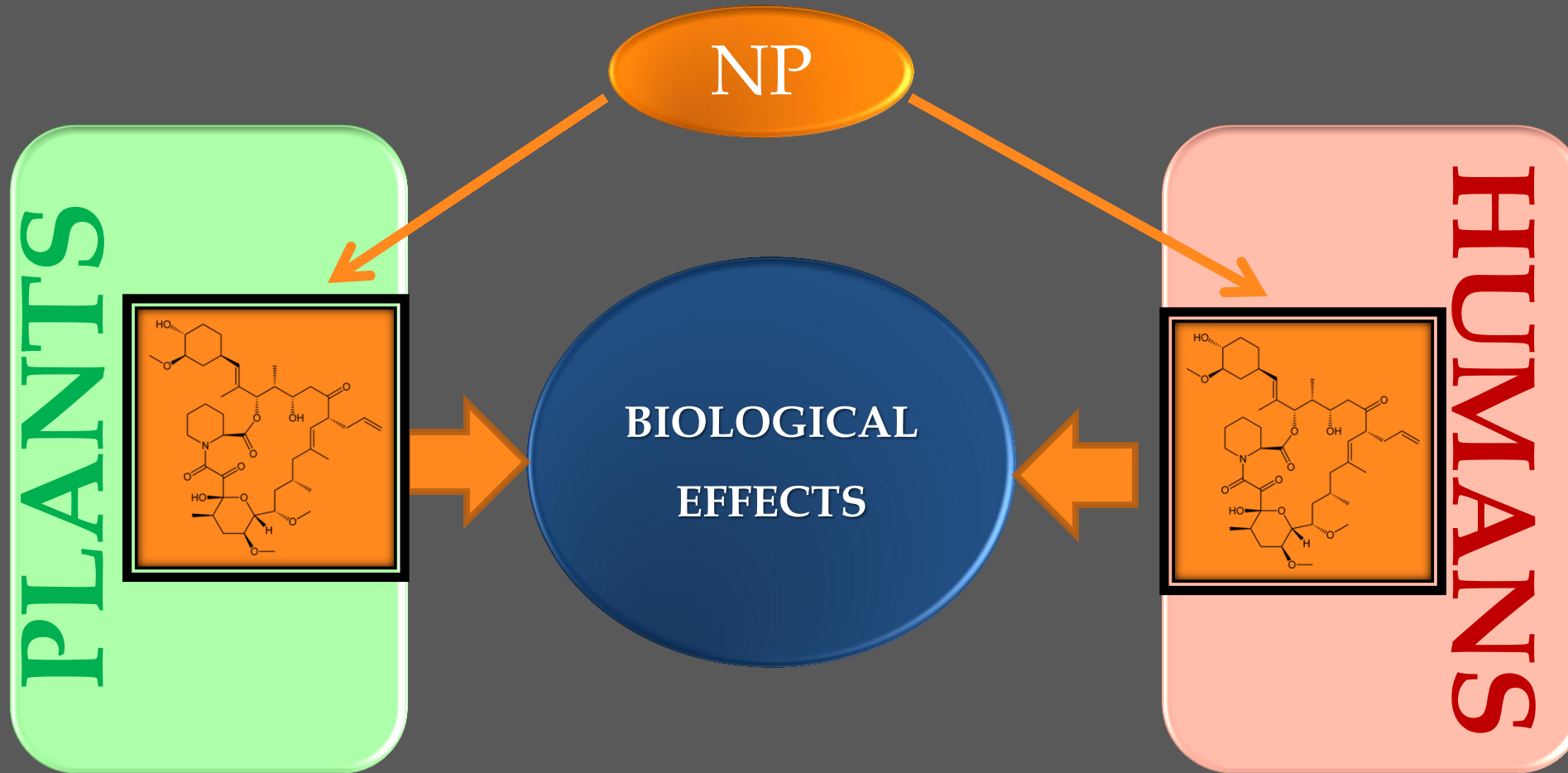




# Targeting Natural Products for Drug Discovery by Mining Biomedical Information Resources

*E. Muratov, N. Baker, N. Rice, D. Fourches, and A. Tropsha*  
*University of North Carolina at Chapel Hill, USA*

# Let's hypothesize ...(or dream?)



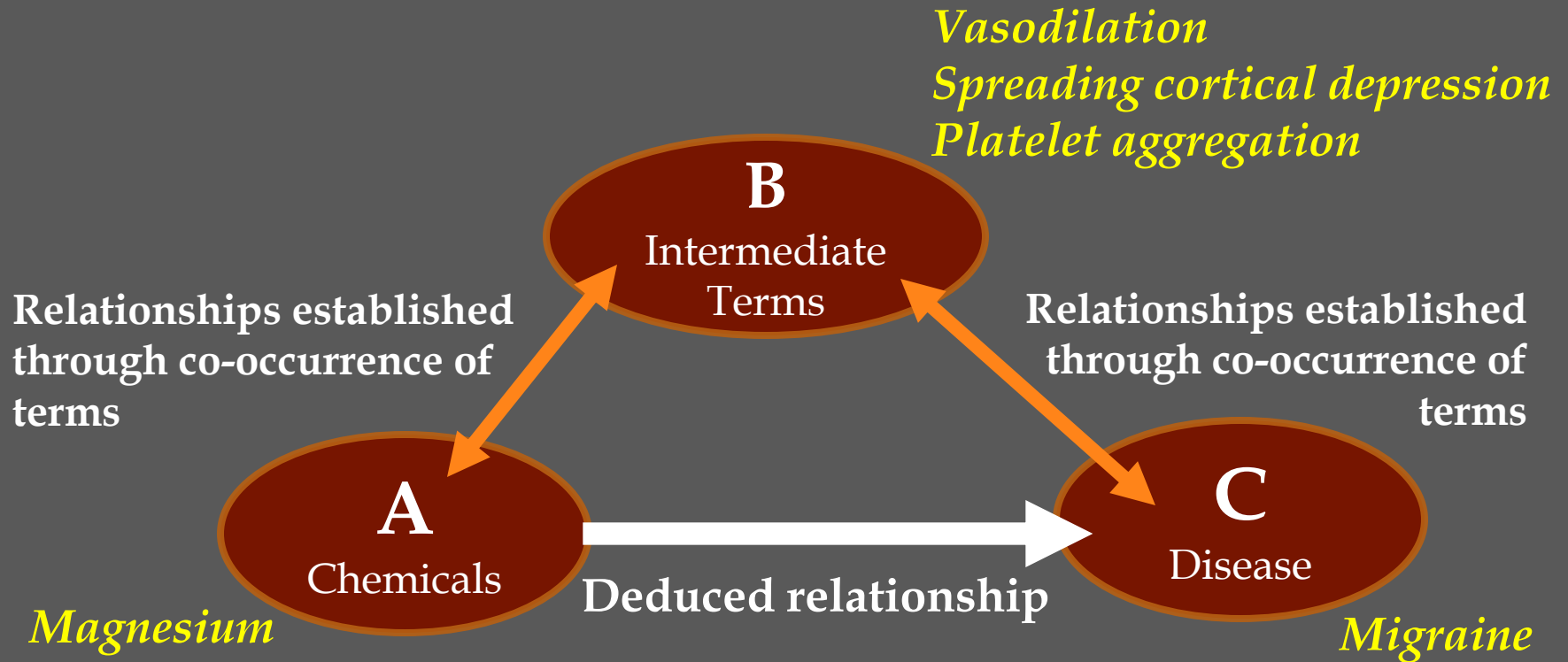
Many NPs will interact within similar biochemical pathways: thus we shall study similarities between plants and humans at the biochemical pathway level to discern novel drug-target-disease associations.

# Research question(s)

---

- Parallel screening of natural products (NP) is a typical approach to identifying target-specific hits, which are then modified by medicinal chemistry approaches;
- Targets are discovered mostly serendipitously;
- Results of biological testing are reported in scientific literature and digital databases;
- Can we employ **literature/data mining** approaches to rationalize search for targets of NP-derived compounds in the context of the systems chemical biology paradigm?

# Swanson's ABC approach to drug discovery via text mining\*

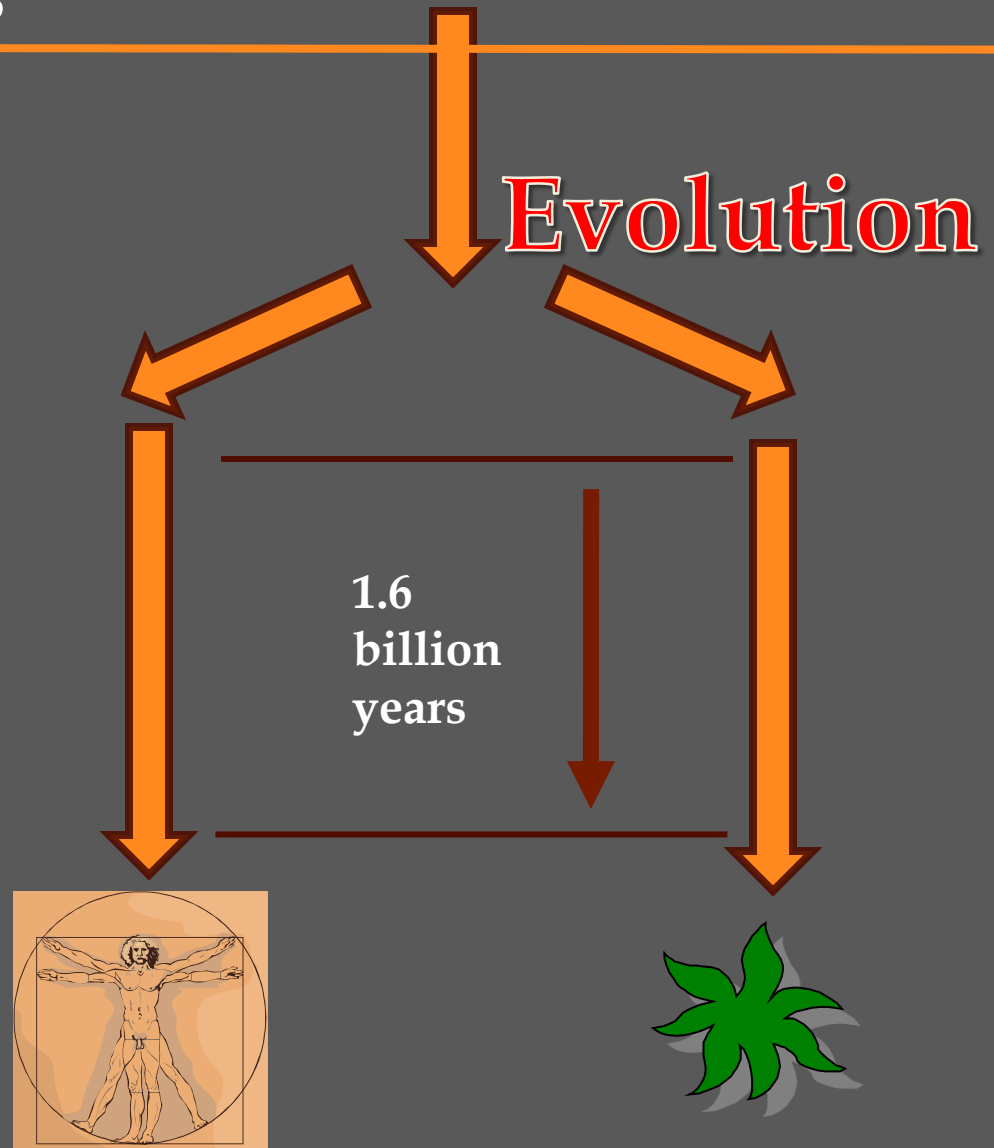


\*Swanson DR. Medical literature as a potential source of new knowledge.  
*Bull Med Libr Assoc* 1990;78(1):29-37

# Commonality of biological pathways between arabidopsis and humans

Arabidopsis and humans

- 1.6 billion years to evolve separately, but still many orthologous genes
- Among cancer genes, 70% have orthologs in *Arabidopsis*
- Why? These proteins serve similar basic cellular functions.



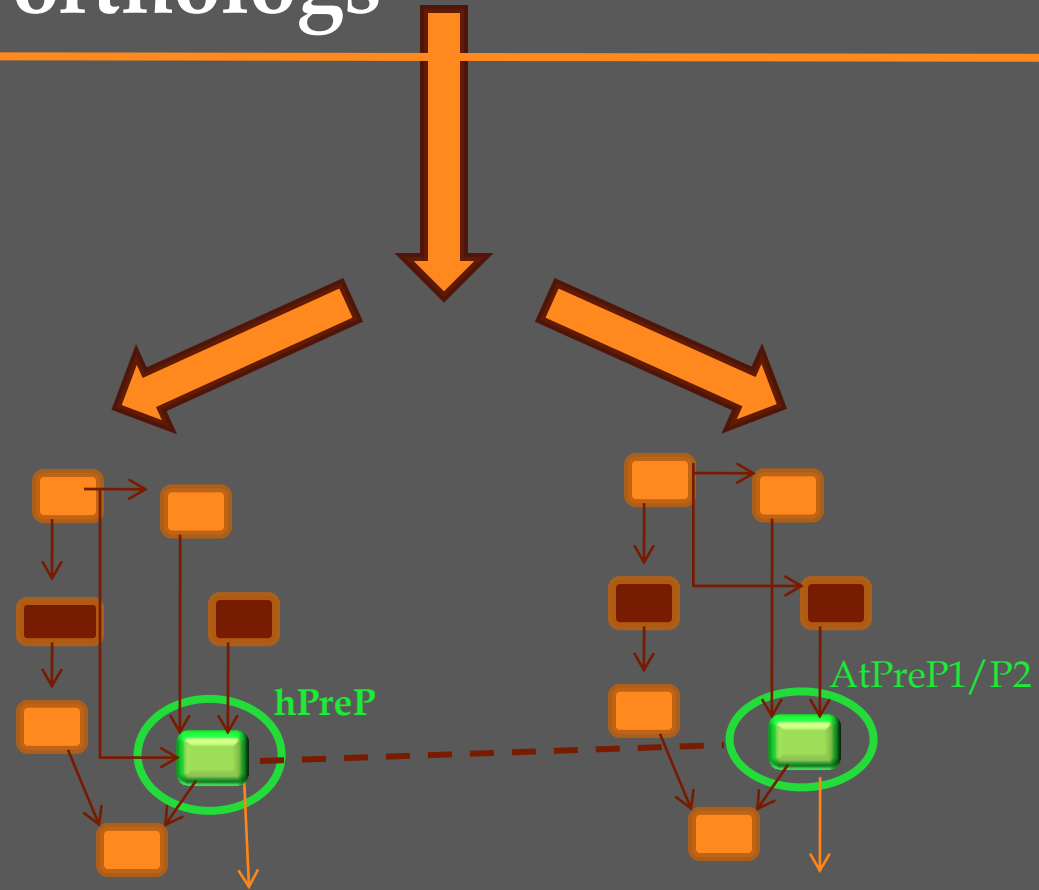
# Example of proteins - orthologs

## Arabidopsis

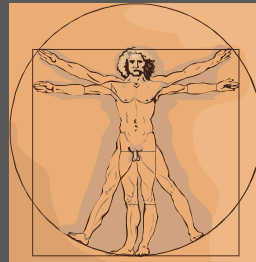
AtPreP1 and AtPreP2 – proteases; digest small unstructured peptides

## Human ortholog

PITRM1 – degrades human peptide plaques  $A\beta$



Alikhani, N., M. Ankarcona and E. Glaser. "Mitochondria and Alzheimer's Disease: Amyloid-Beta Peptide Uptake and Degradation by the Presequence Protease, hPreP." *Journal of Bioenergetics and Biomembranes* 41.5 (2009): 447-451.



Degrades peptide plaques ... Connection to Alzheimer's?



Digests peptides

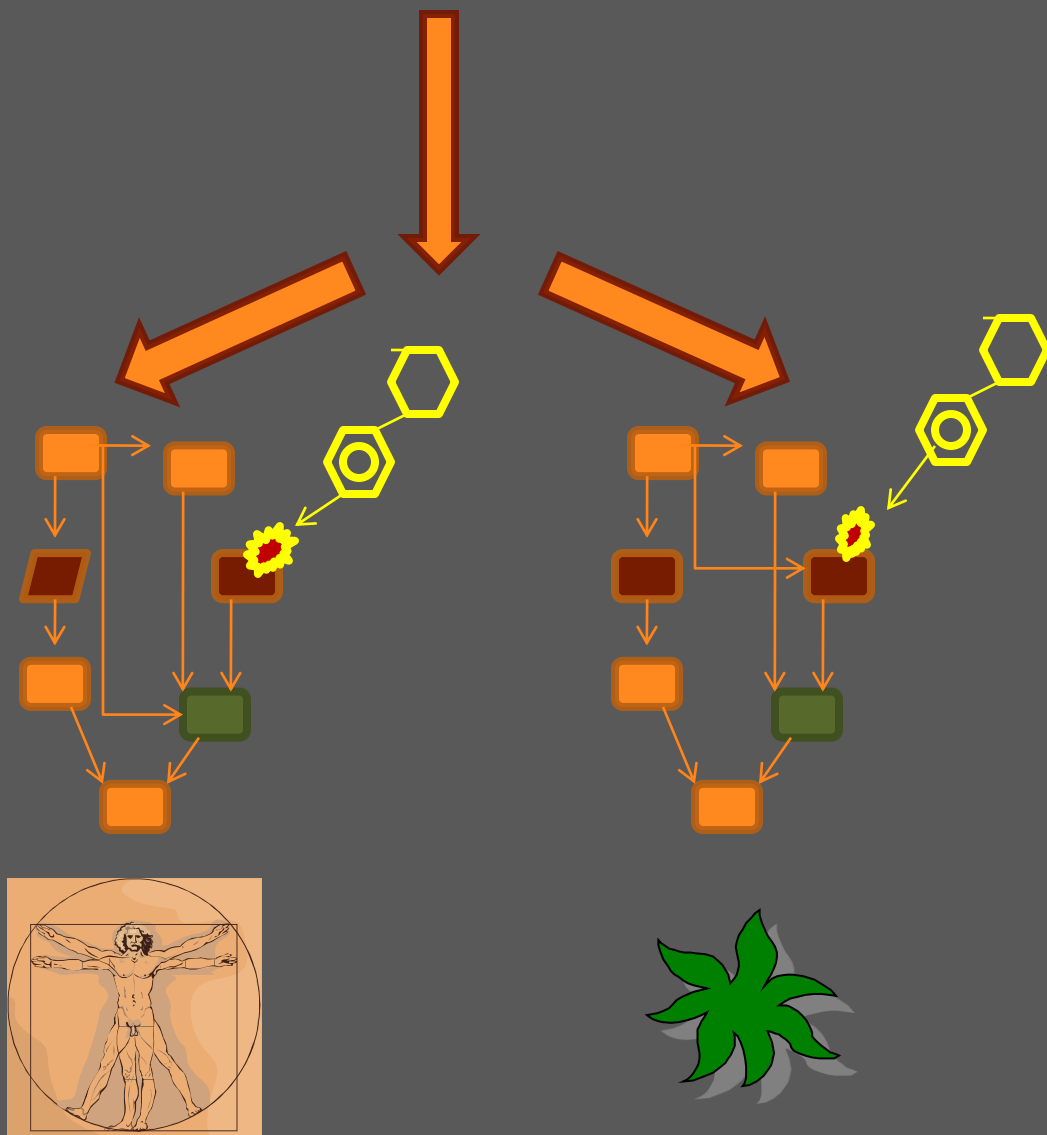
Xu, X. M. and S. G. Moller. "The Value of Arabidopsis Research in Understanding Human Disease States." *Current opinion in biotechnology* (2010).

# Opportunity for new drugs

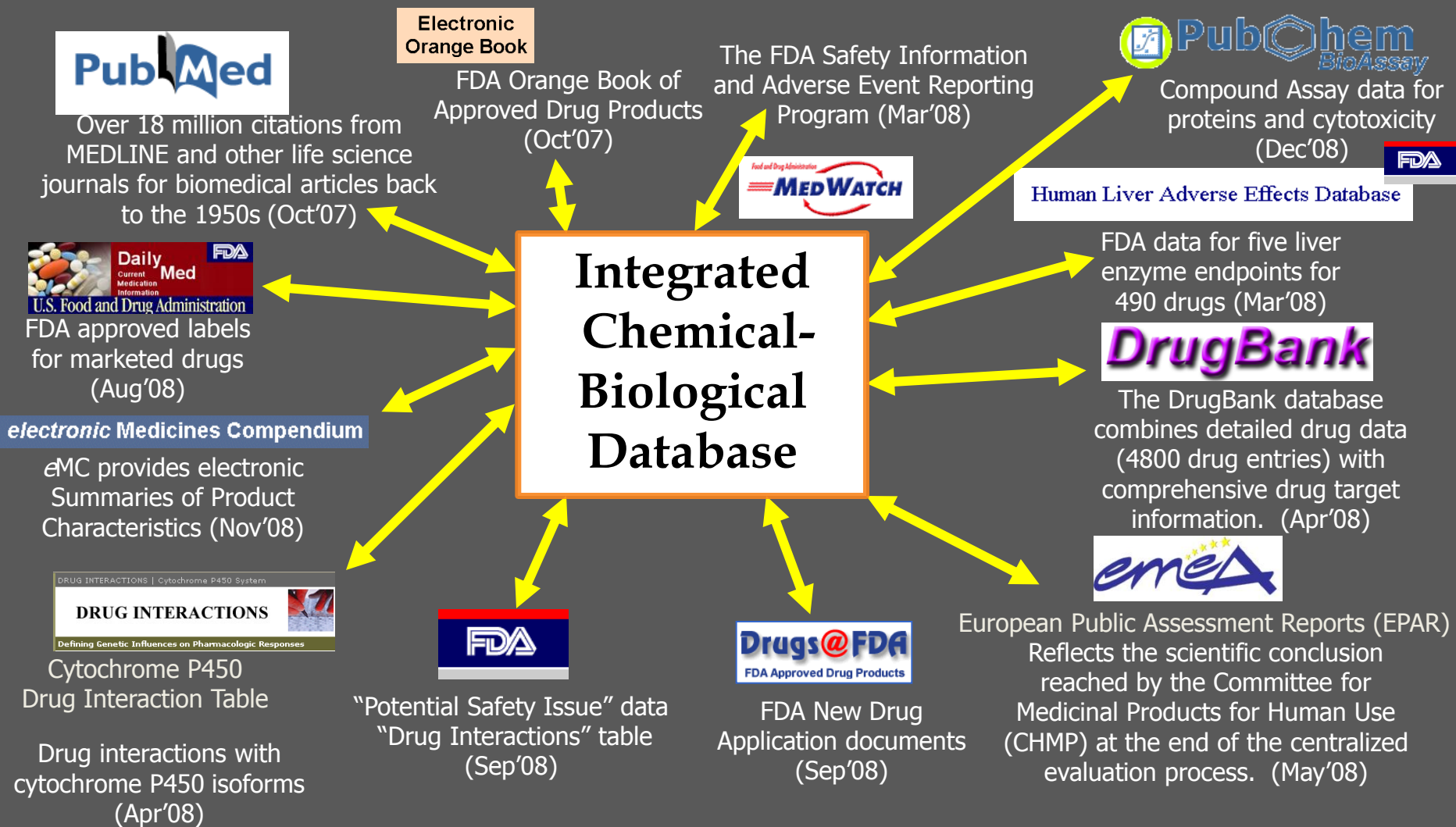
Chemicals that modulate the protein in the plant could be used to modulate similar protein in humans



**Potential drug candidates**



# Information resources for bioactive chemicals are abundant and growing



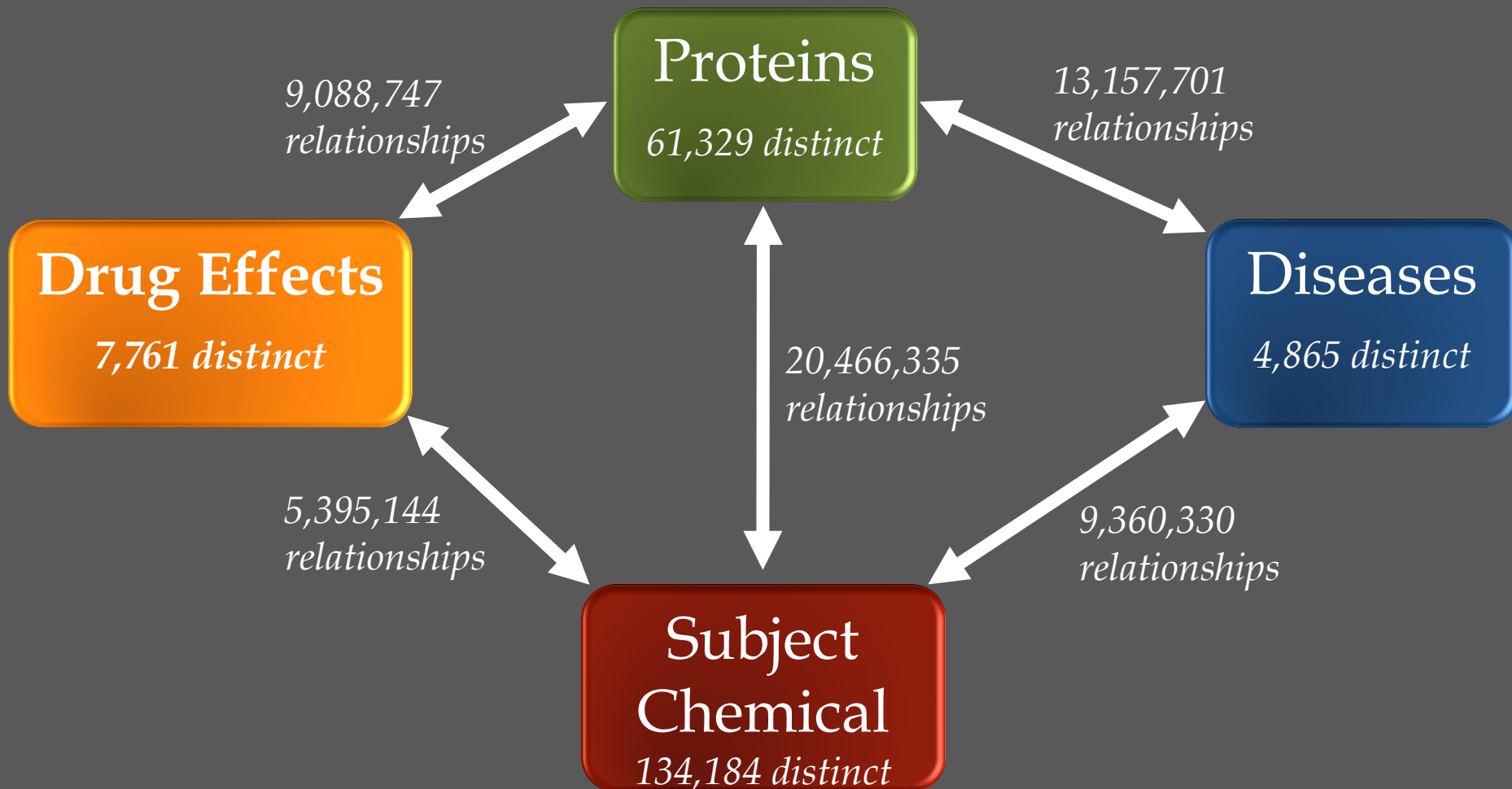
# ChemoText

---

- A chemical-centric database composed of MeSH (Medical Subject Heading) annotations extracted from PubMed's Medline
- Processed and organized to be useful in computational drug research methods
- Captures essential relationships in drug research: drug-protein-disease

Baker NC, Hemminger BM. Mining connections between chemicals, proteins, and diseases extracted from Medline annotations. *J Biomed Inform.* 2010, 43(4):510-9.

# Exploring PubMed as one of the largest chemical biology databases with ChemoText



- 2008 Medline baseline: 16,880,015 records
- 6,635,344 records had subject chemicals

# Example: Disease-protein co-annotation

---

PMID- 15610510

...

DP - 2004 Dec

TI - Alterations of glucosylceramide-beta-glucosidase levels in the skin of patients with psoriasis vulgaris.

MH - Epidermis/\*enzymology/pathology

MH - **Glucosylceramidase**/\*genetics/\*metabolism

MH - Humans

MH - **Psoriasis**/\*metabolism/pathology/\*physiopathology

MH - RNA, Messenger/metabolism

MH - Reverse Transcriptase Polymerase Chain Reaction

MH - Water/metabolism

MH - **beta-Glucosidase**

**Disease in humans**

**Protein in humans**

# What is stored in ChemoText from these Medline records?

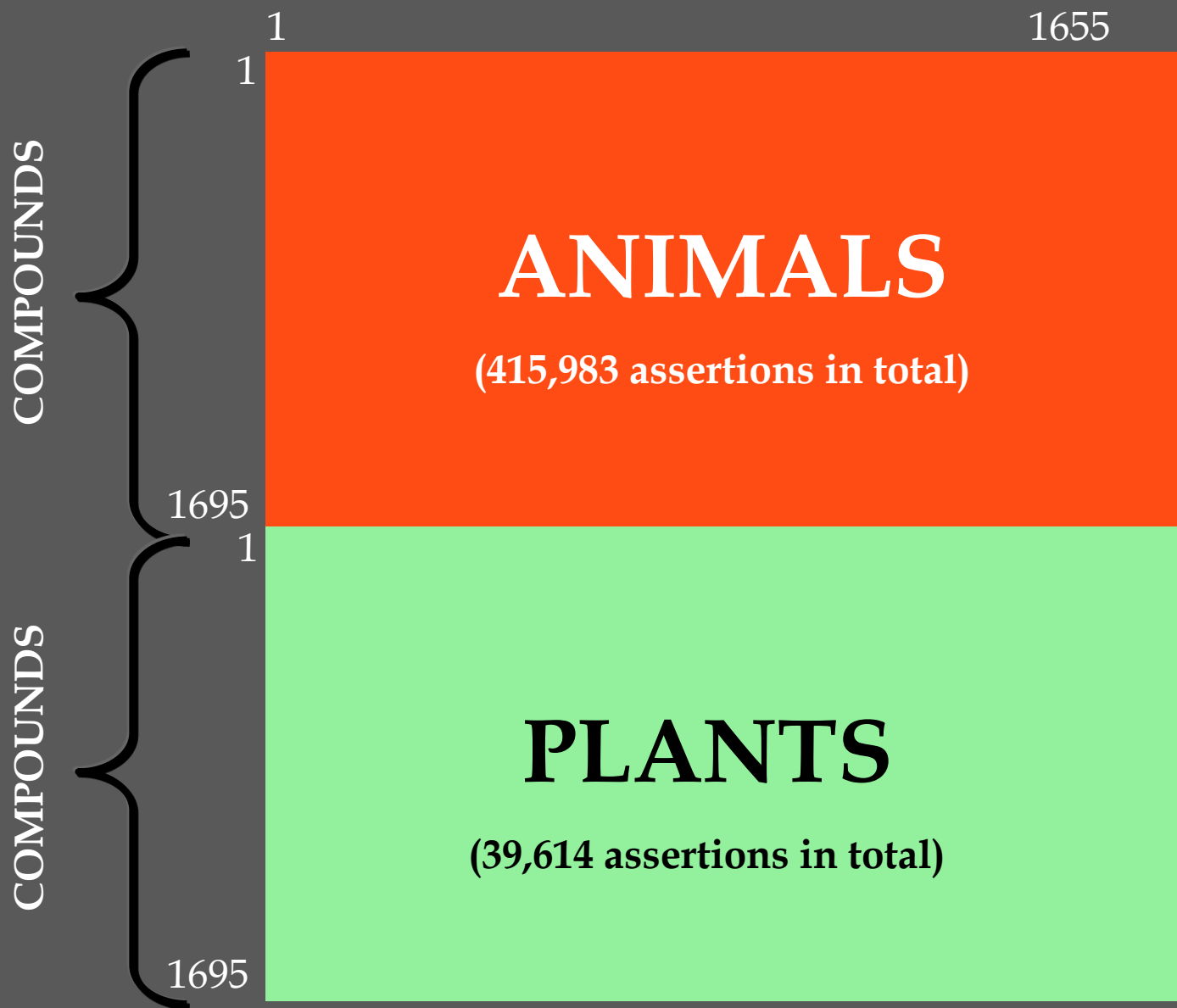
## Protein - disease table (over 13 million records)

PubMed ID	Protein	Disease	Species
15610510	Beta-Glucosidase	Psoriasis	Humans
15610510	Glucosylceramidase	Psoriasis	Humans

## Protein - chemical table (over 20 million records)

PubMed ID	Protein	Subject Chemical	Species
15610510	Beta-Glucosidase	Podophyllotoxin	Podophyllum peltatum

# Overview of extracted information stored in data matrices



# Curation of extracted information : MeSH Tree

---

Anatomy [A]

Organisms [B]

Diseases [C]

Chemicals and Drugs [D]

Analytical, Diagnostic and Therapeutic Techniques and Equipment [E]

Psychiatry and Psychology [F]

**Phenomena and Processes [G]** ← *Relevant for this study*

Disciplines and Occupations [H]

Anthropology, Education, Sociology and Social Phenomena [I]

Technology, Industry, Agriculture [J]

Humanities [K]

Information Science [L]

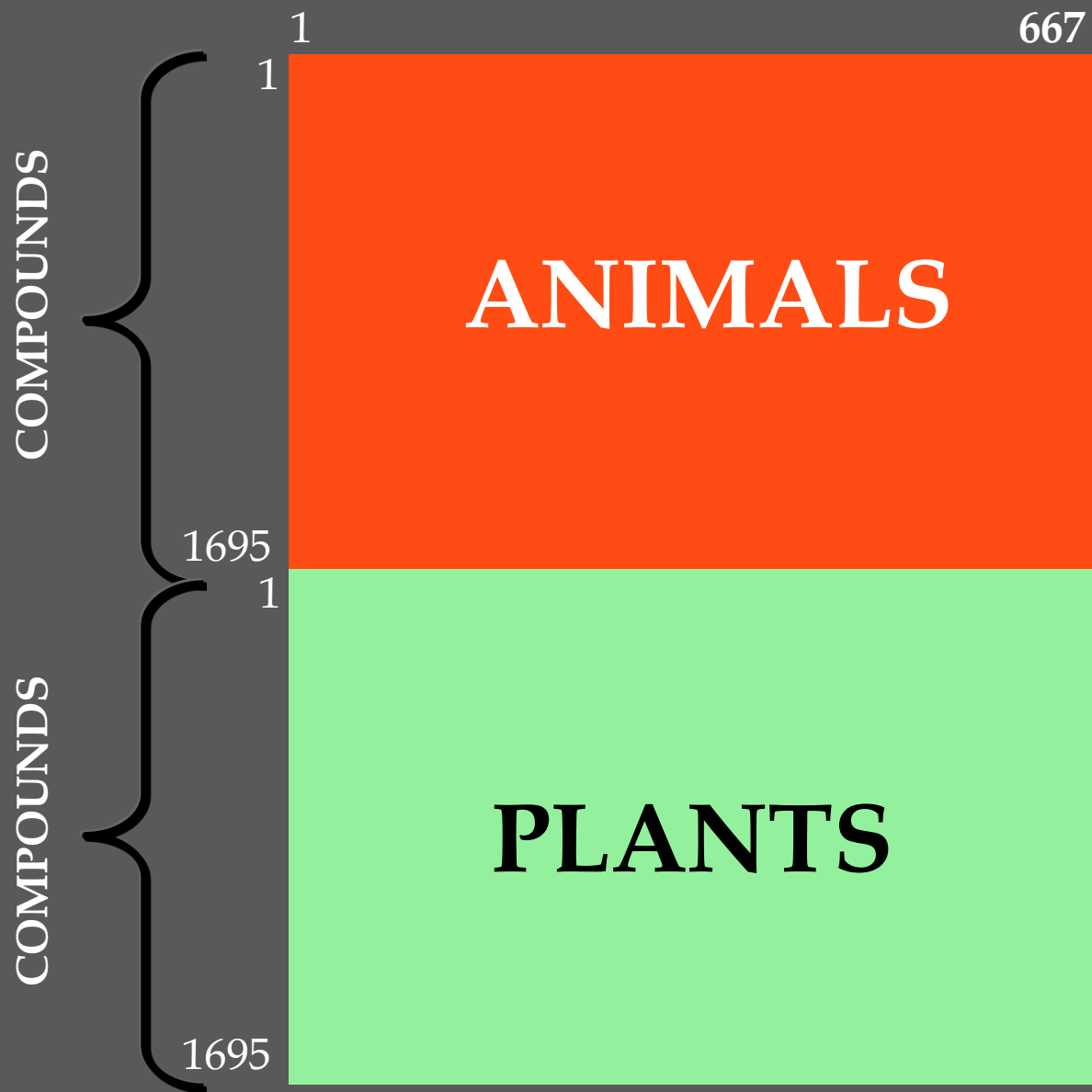
Named Groups [M]

Health Care [N]

Publication Characteristics [V]

Geographical [Z]

# Overview of reduced information stored in data matrices



# Data Explorer

NATURAL PRODUCT project

MATRIX GENERATOR | **EXPLORER** | RESULTS

PREVIOUS: 2,2'-azobis(2-amidinopropane)      NEXT

FIRST: **20 / 1695**      GO TO:      LAST

Select Matrix: C:\WORK\NAT\_PRODUCTS\UNITED\_MATRIX\_T5.txt

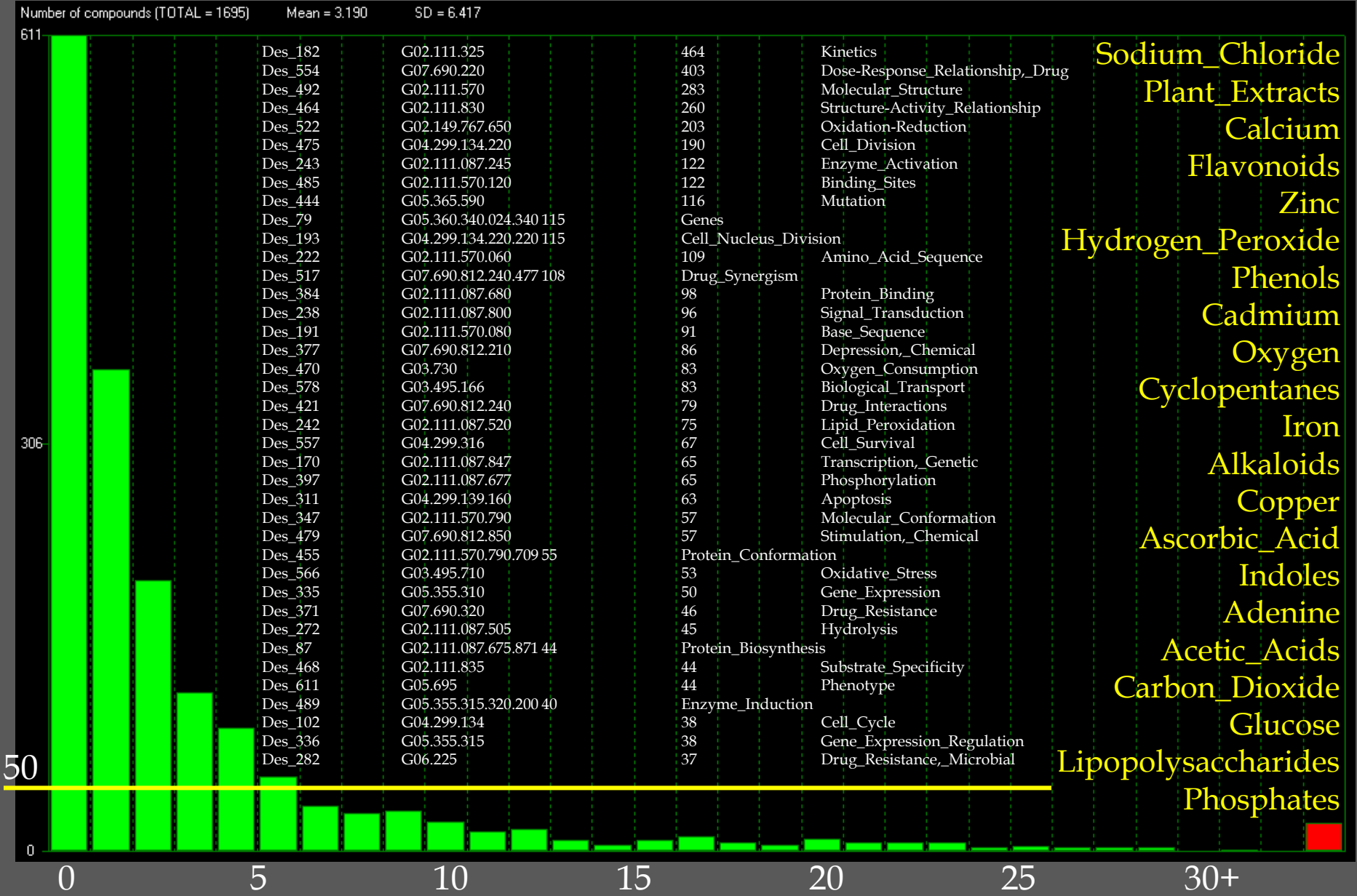
Convert Names: # common patterns = 4      **BATCH Search**

ANIMAL related information		PLANT related information	
G02.111.087.160.750	Glycolysis	G02.111.325	Kinetics
G02.111.325	Kinetics	G02.111.087.520	Lipid_Peroxidation
G08.686.815	Sex_Characteristics	G04.299.134.220	Cell_Division
G02.111.087.520	Lipid_Peroxidation	G07.690.220	e-Response_Relationship_
G02.111.087.245	Enzyme_Activation	G03.495.710	Oxidative_Stress
G05.355.180.230	DNA_Fragmentation		
G02.111.087.505	Hydrolysis		
G04.299.139.160	Apoptosis		
G02.111.087.677	Phosphorylation		
G07.690.812.240	Drug_Interactions		
G02.111.570.790.700	Protein_Conformation		
G03.730	Oxygen_Consumption		
G04.299.134.220	Cell_Division		

Automatic search of common profiles



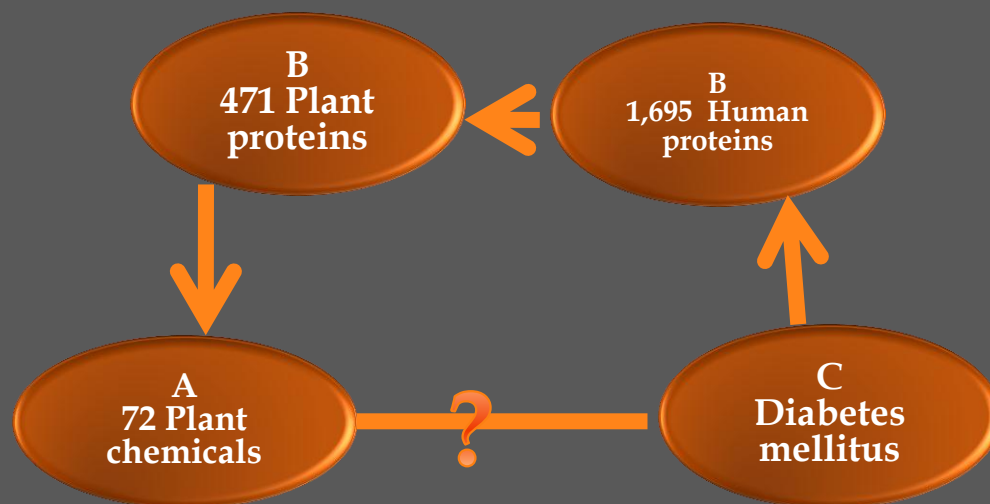
# General observations for compounds



# npABC: Step by step

## *Diabetes mellitus example*

- Look for all human protein annotations in ChemoText that co-occur with diabetes annotation
- Using the same protein names, look for articles about plants that are annotated with the protein.



- Look for plant small molecule chemicals annotated with any of the plant proteins

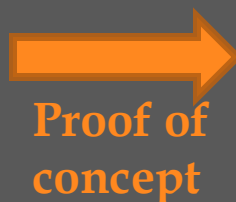
Was this compound linked to diabetes?  
-yes: proof of concept  
-no: new hypothesis

# Hypothesis set: chemicals with link to diabetes

26 compounds already linked to diabetes through the literature (ChemoText)

**First Yr:** first year a link appears between compound and diabetes

**PMID Ct:** The number of PubMed articles which directly link the compound to diabetes (ChemoText contains data through 2008 only)



Compound	First Yr	PMID Ct
Vitamin E	1949	363
Ascorbic Acid	1949	305
Folic Acid	1962	58
Quercetin	1977	25
Choline	1950	22
resveratrol	2006	18
alginic acid	1983	17
Citric Acid	1984	10
Genistein	1998	8
Salicylic Acid	1959	8
Lutein	1997	4
caffeic acid	2000	3
Erythritol	1996	3
alpha-carotene	1999	3
proanthocyanidin	2004	3
Vitamin K 1	1993	2
Phloroglucinol	2004	1
naringenin	2006	1
Methoxsalen	1969	1
steviol	2008	1
Thapsigargin	2003	1
zeaxanthin	2008	1
kaempferol	2000	1
Abscisic Acid	2007	1
gibberellic acid	1989	1
Chlorophyll	1963	1

# Hypothesis set: chemicals with no direct link to diabetes

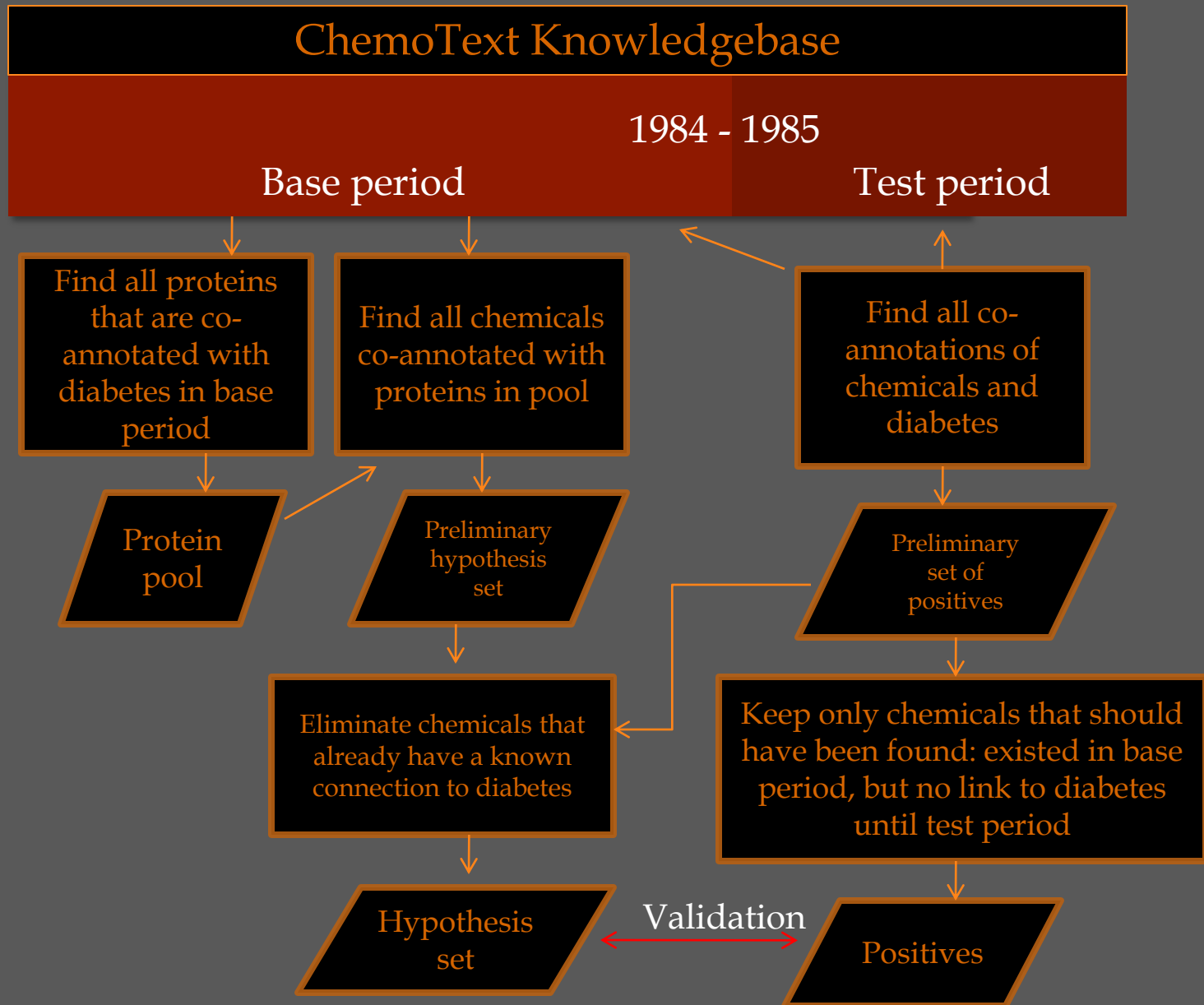
46 plant chemicals  
*predicted* to have a link  
to diabetes through  
intermediary protein  
annotations, but no  
direct link (yet) in the  
literature  
(ChemoText)



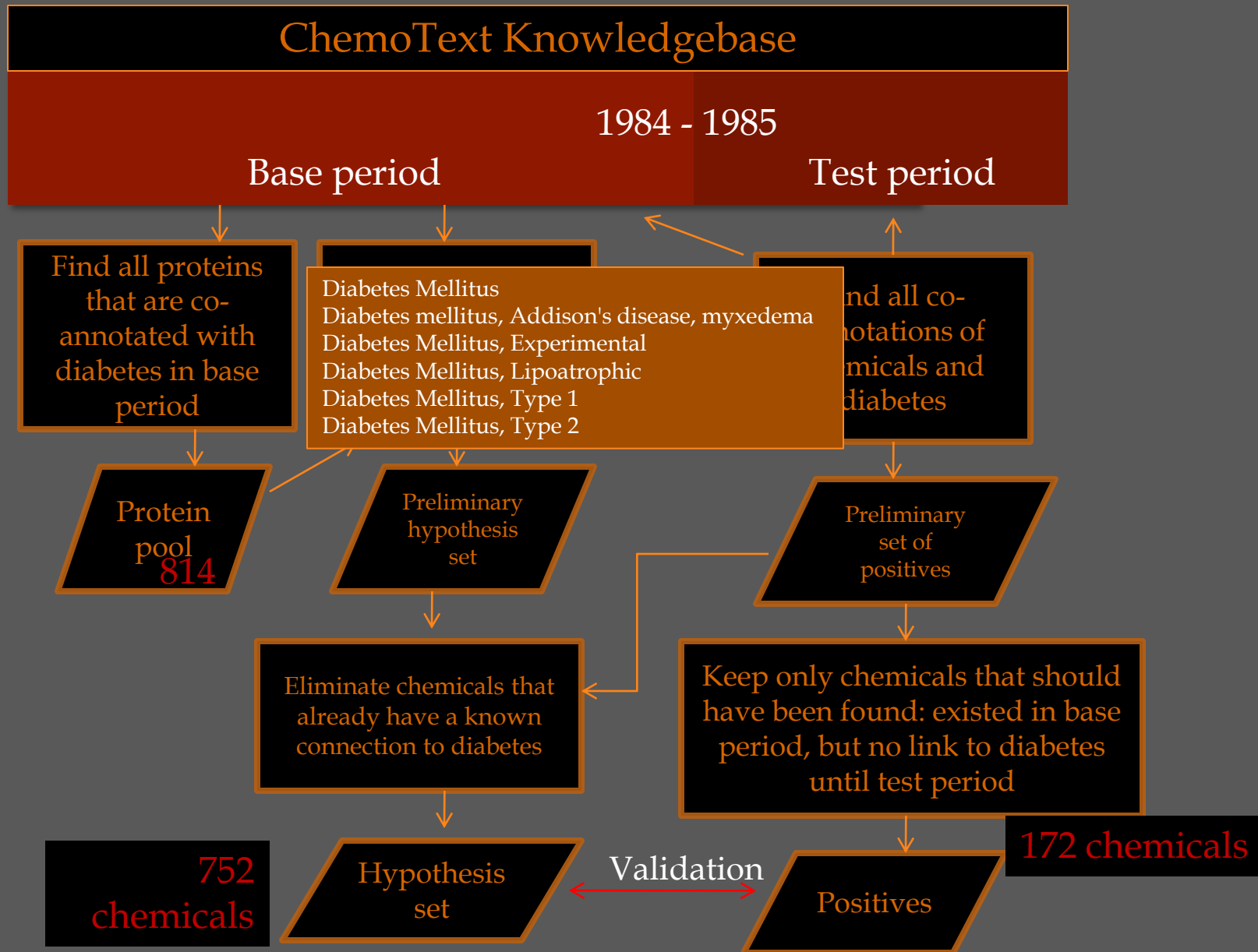
**Predictions**  
**(very preliminary!)**

dhurrin  
echinenone  
ethylene  
flavan-3-ol  
dihydrocamelexic acid  
deoxynivalenol  
crepenynic acid  
chlorophyll b  
camalexin  
jasmonic acid  
caffeoylquinic acid  
brassinolide  
aristolochic acid I  
Adenosine Diphosphate Glucose  
8',8',8'-trifluoroabscisic acid  
5,11-methenyltetrahydrohomofolate  
Camptothecin  
phytoene  
Zeatin  
xanthoxin  
violaxanthin  
Uridine Diphosphate Xylose  
thalianol  
sinapinic acid  
sinapine  
scopolin  
sclareolide  
sclareol  
saponarin  
Quinic Acid  
Protochlorophyllide  
indoleacetic acid  
methyl jasmonate  
Gossypol  
GR24 compound  
indoleacetamide  
4-coumaric acid  
Plastoquinone  
gallic acid  
plastoquinol  
morlin  
N(6)-(delta(2)-isopentenyl)adenine  
Oxalic Acid  
pheophorbide a  
gibberellin A12  
Isopentenyladenosine

# Validation of Chemotext-based approach

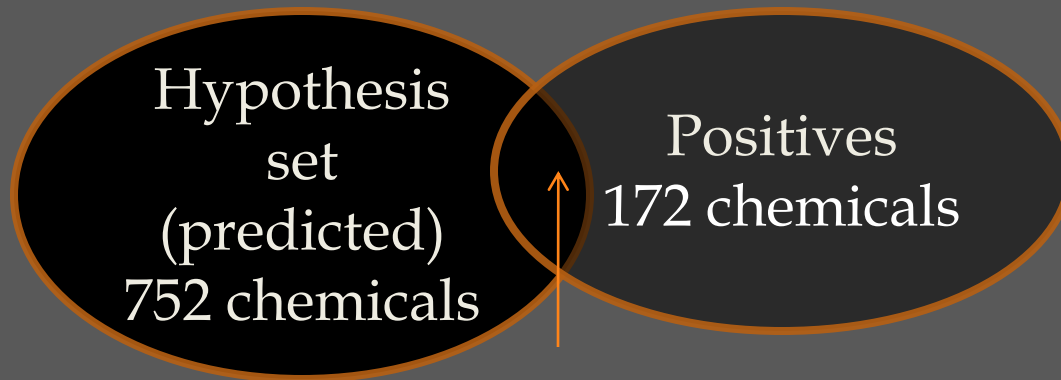


# Validation of Chemotext-based approach



# Results summary

---



True positives: 105  
chemicals

Recall:  $TP / (TP + FN) = 105 / 172 = 61\%$

Precision:  $TP / (TP + FP) = 105 / 752 = 16\%$

# Results – top 20

Before 1985		1985 and after			
Avg Rank	Chemical Name	FirstYr	ArtCt	DisQual	ChemQual
1	Gallic Acid	0	0		
2	Vinblastine	0	0		
3	Vincristine	2003	1	chemically induced	adverse effects
4	Gossypol	0	0		
5	Capsaicin	1989	24	physiopathology	pharmacology
6	Tetrahydrocannabinol	1985	5	drug therapy	pharmacology
7	Lignin	2009	2	complications	therapeutic use
8	Guaiacol	0	0		
9	Strychnine	0	0		
10	acid citrate dextrose	0	0		
11	alpha-Tocopherol	1989	42	blood	blood
12	Latex	1996	2	drug therapy	adverse effects
13	ochratoxin A	0	0		
14	Rhamnose	0	0		
15	Eugenol	2000	2	blood	pharmacology
16	Hymecromone	0	0		
17	Methoxsalen	0	0		
18	Tea	1994	28	epidemiology	chemistry
19	Emetine	0	0		
20	Rotenone	2009	1	chemically induced	pharmacology

Level of stringency: subject drug co-annotated with diabetes. These drugs may be connected to diabetes complications.

# Gallic acid and diabetes – recent paper

[Chem Biol Interact.](#) 2011 Jan 15;189(1-2):112-8. Epub 2010 Nov 13.

Insulin-secretagogue, antihyperlipidemic and other protective effects of gallic acid isolated from Terminalia bellerica Roxb. in streptozotocin-induced diabetic rats.

[Latha RC](#), [Daisy P.](#)

Department of Biotechnology, Holy Cross College, Teppakulam PO, Main Gauard Gate, Trichy, Tamil Nadu, India. latha2611@gmail.com

## Abstract

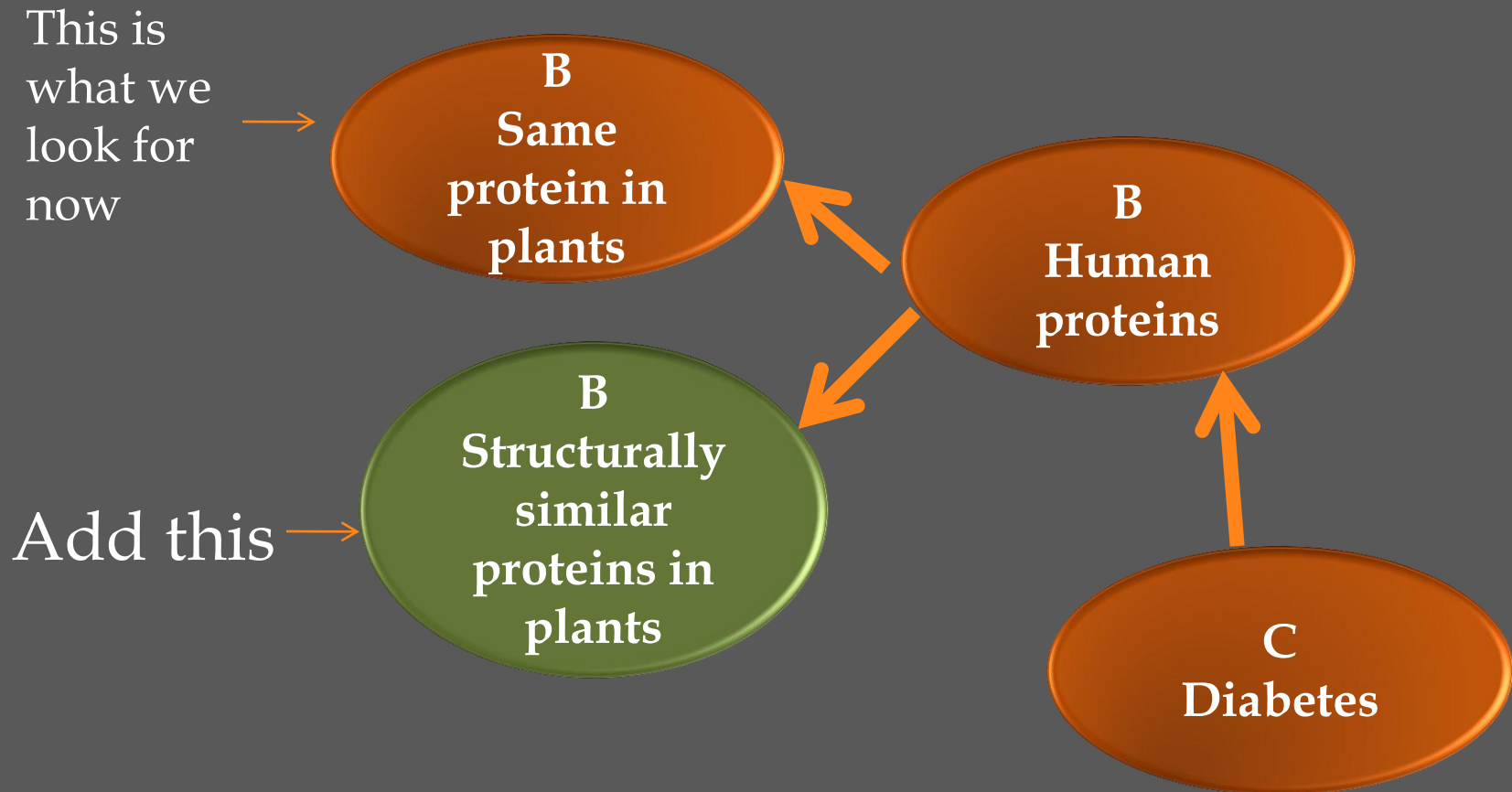
Diabetes mellitus causes derangement of carbohydrate, protein and lipid metabolism which eventually leads to a number of secondary complications. Terminalia bellerica is widely used in Indian medicine to treat various diseases including diabetes. The present study was carried out to isolate and identify the putative antidiabetic compound from the fruit rind of T. bellerica and assess its chemico-biological interaction in experimental diabetic rat models. Bioassay guided fractionation was followed to isolate the active compound, structure was elucidated using  $(1)H$  and  $(13)C$  NMR, IR, UV and mass spectrometry and the compound was identified as gallic acid (GA). GA isolated from T. bellerica and synthetic GA was administered to streptozotocin (STZ)-induced diabetic male Wistar rats at different doses for 28 days. Plasma glucose level was significantly ( $p < 0.05$ ) reduced in a dose-dependent manner when compared to the control. Histopathological examination of the pancreatic sections showed regeneration of  $\beta$ -cells of islets of GA-treated rats when compared to untreated diabetic rats. In addition, oral administration of GA (20mg/kg bw) significantly decreased serum total cholesterol, triglyceride, LDL-cholesterol, urea, uric acid, creatinine and at the same time markedly increased plasma insulin, C-peptide and glucose tolerance level. Also GA restored the total protein, albumin and body weight of diabetic rats to near normal. Thus our findings indicate that gallic acid present in fruit rind of T. bellerica is the active principle responsible for the regeneration of  $\beta$ -cells and normalizing all the biochemical parameters related to the patho-biochemistry of diabetes mellitus and **hence it could be used as a potent antidiabetic agent.**

# How to improve

---

- Not to rely on predictions only; pathways and chemical systems biology mechanisms must be elucidated
- To extend Intermediate B proteins list by addition of phenologs and different proteins with similar function to plant and animal proteins with same names and same function
- Careful curation and analysis of literature data and obtained results
- Optimization and rigorous external validation of hypothesis and selected hits, both computationally and experimentally

# Next steps in npABC: expand definitions of “similar” proteins



Thank you!

