



5th International Symposium on Semantic Mining in Biomedicine (SMBM)

Institute of Computational Linguistics

University of Zurich, Switzerland

September 4, 2012

From biomedical information integration to knowledge discovery



Olivier Bodenreider

Lister Hill National Center
for Biomedical Communications
Bethesda, Maryland - USA

Semantic mining

- ◆ Extract information from structured and unstructured sources
 - From text: text mining
 - From ontologies and knowledge bases
- ◆ Integrate information
 - From structured and unstructured sources
- ◆ Aggregate information
 - Subsumption reasoning
- ◆ Use the extracted information for a meaningful purpose
 - Hypothesis generation / knowledge discovery
 - Better information retrieval
 - Question answering

Outline

- ◆ Knowledge, integration and aggregation
- ◆ Knowledge sources
 - Structured sources
 - Relations extracted from text
- ◆ Integrating relations from text mining and ontologies
- ◆ Biomedical Knowledge Repository



KNOWLEDGE, INTEGRATION AND AGGREGATION

Definitional knowledge

◆ Definitional knowledge

- Universally true
- Examples
 - Lung cancer *has_location* Lung
 - Myocardial infarction *isa* Cardiovascular disease
 - Liver *part_of* Abdomen (canonical anatomy, in a given species)
- Typically found in ontologies
- Useful as background knowledge

Assertional knowledge

◆ Assertional knowledge

- True in a given context
- Examples
 - Aspirin *treats* headache
 - IL-13 *inhibits* COX2
 - Chest pain *manifestation_of* Myocardial infarction
 - Ciprofloxacin *causes* Tendon rupture
- Typically found in knowledge bases (and in text)
- Useful for knowledge discovery, question answering, biocuration support, etc.



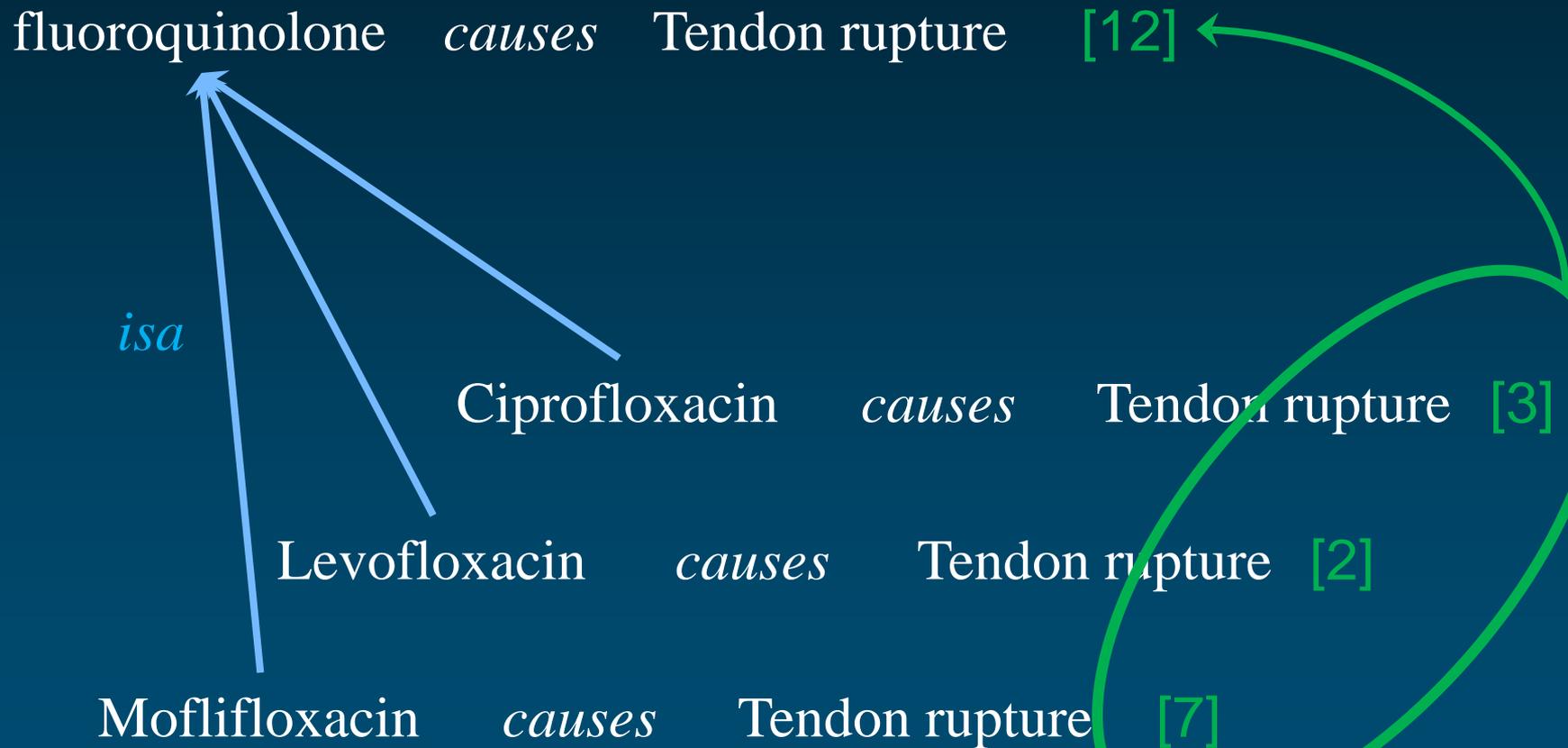
Definitional vs. assertional knowledge

- ◆ Definitional knowledge
 - Universally true
 - Typically found in ontologies
 - Useful as background knowledge
- ◆ Assertional knowledge
 - True in a given context
 - Typically found in knowledge bases (and in text)
 - Useful for knowledge discovery, question answering, biocuration support, etc.

Why integrate assertional and definitional knowledge?

- ◆ To bridge the granularity mismatch
 - Differences in granularity between
 - What is expressed in in text (or structured sources)
 - What is needed in “semantic mining” applications
- ◆ To increase statistical power
 - Low frequency for individual, fine-grained assertions
 - Higher frequency when frequencies are aggregated at a coarser level

Aggregating frequencies



Bridging the granularity mismatch

- ◆ A researcher is interested in glycosylation and its implications for one disorder: congenital muscular dystrophy.

Link between glycosyltransferase activity and congenital muscular dystrophy?



All Databases PubMed Nucleotide Protein Genome Structure PMC

Search Gene for 9215[uid] [Go](#) [Clear](#) [Save Search](#)

Limits Preview/Index History Clipboard Details

Display Full Report Show 20 Send to

All: 1 Current Only: 1 Genes Genomes: 1 SNP GeneView: 1

1: LARGE like-glycosyltransferase [*Homo sapiens*]
 GeneID: 9215 updated 02-Jul-2007

LARGE
(GeneID: 9215)

Phenotypes

has_associated_disease

Muscular dystrophy, congenital, type 1D
[MIM: 608840](#)

Congenital muscular dystrophy, type 1D Provided by GOA

GeneOntology

Function	Evidence
acetylglucosaminyltransferase activity	TAS PubMed

Process	Evidence
N-acetylglucosamine metabolic process	TAS PubMed
carbohydrate biosynthetic process	IEA
glycosphingolipid biosynthetic process	TAS PubMed
muscle maintenance	ISS
protein amino acid glycosylation	TAS PubMed

Component	Evidence
integral to Golgi membrane	TAS PubMed
integral to membrane	IEA
membrane	IEA



All Databases PubMed Nucleotide Protein Genome Structure PMC

Search Gene for 9215[uid] [Save Search](#)

Limits Preview/Index History Clipboard Details

Display Full Report Show 20 Send to

All: 1 Current Only: 1 Genes Genomes: 1 SNP GeneView: 1

LARGE
(GeneID: 9215)

1: LARGE like-glycosyltransferase [*Homo sapiens*]

GeneID: 9215

updated 02-Jul-2007

Phenotypes

Muscular dystrophy, congenital, type 1D
[MIM: 608840](#)

GeneOntology

has_molecular_function

Provided by [GOA](#)

Function	Evidence
acetylglucosaminyltransferase activity	TAS PubMed

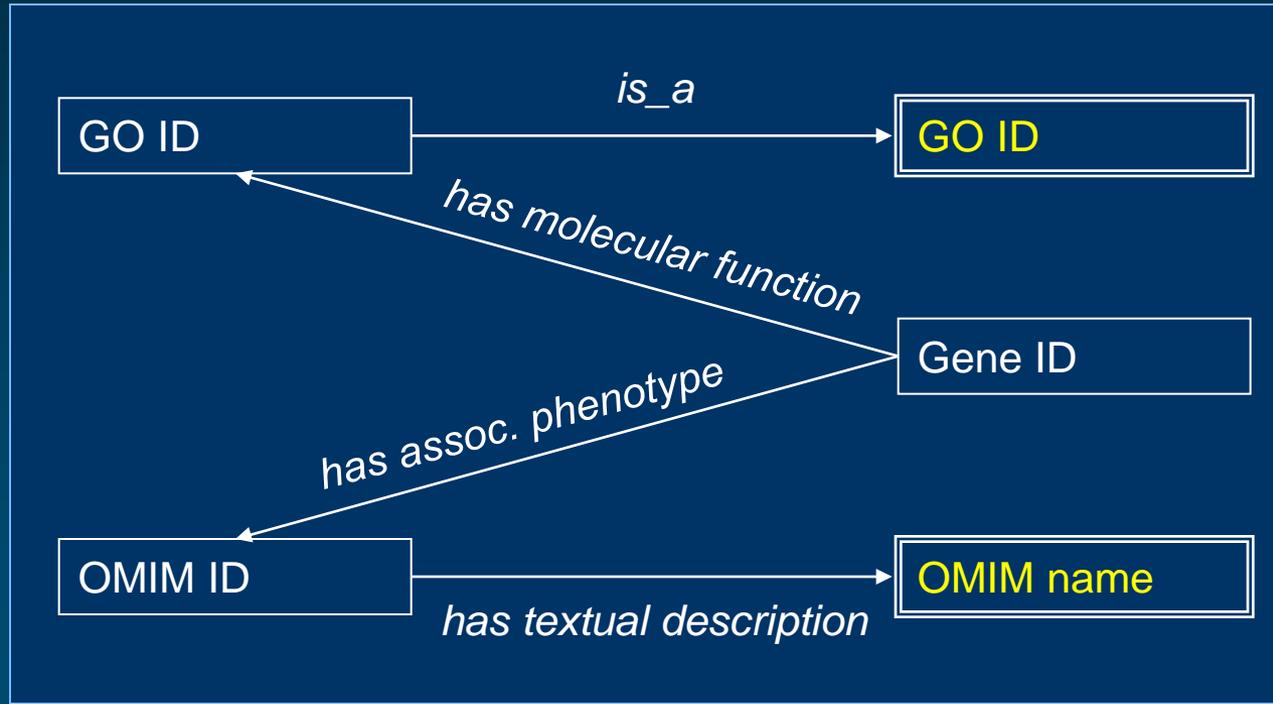
acetylglucosaminyltransferase activity

Process	Evidence
N-acetylglucosamine metabolic process	TAS PubMed
carbohydrate biosynthetic process	IEA
glycosphingolipid biosynthetic process	TAS PubMed
muscle maintenance	ISS
protein amino acid glycosylation	TAS PubMed

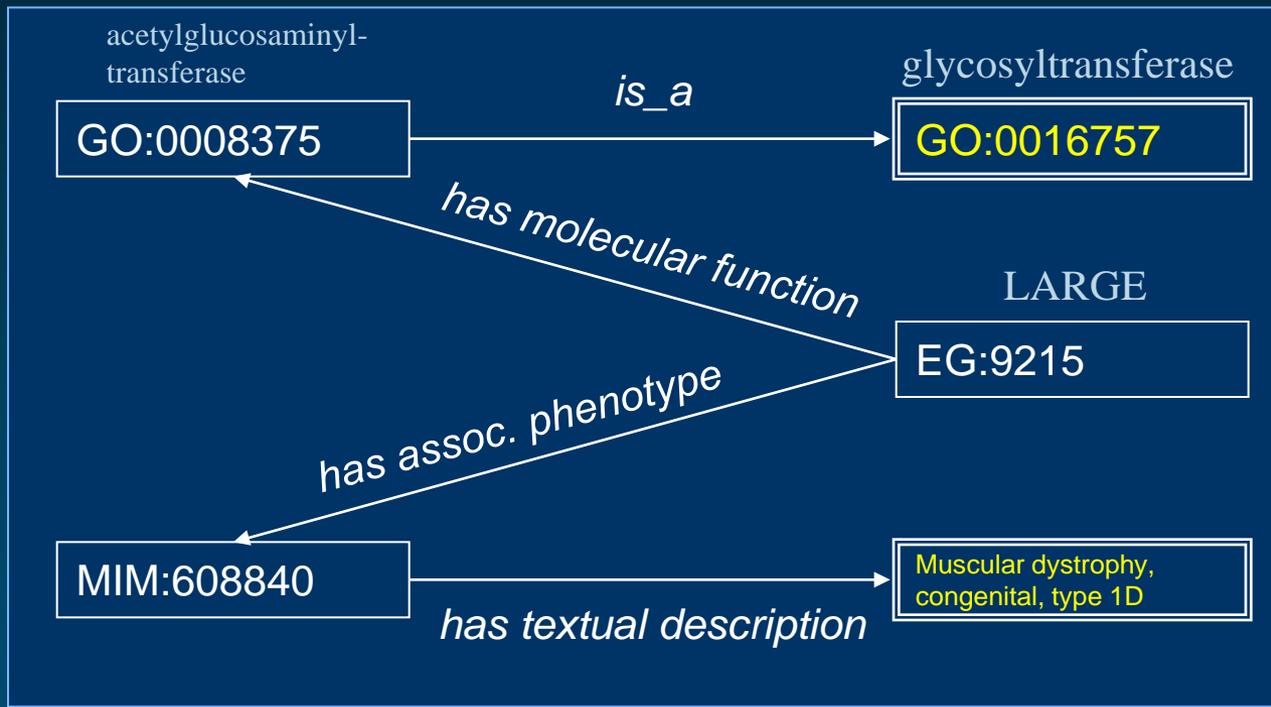
Component	Evidence
integral to Golgi membrane	TAS PubMed
integral to membrane	IEA
membrane	IEA

Using SPARQL to test a hypothesis

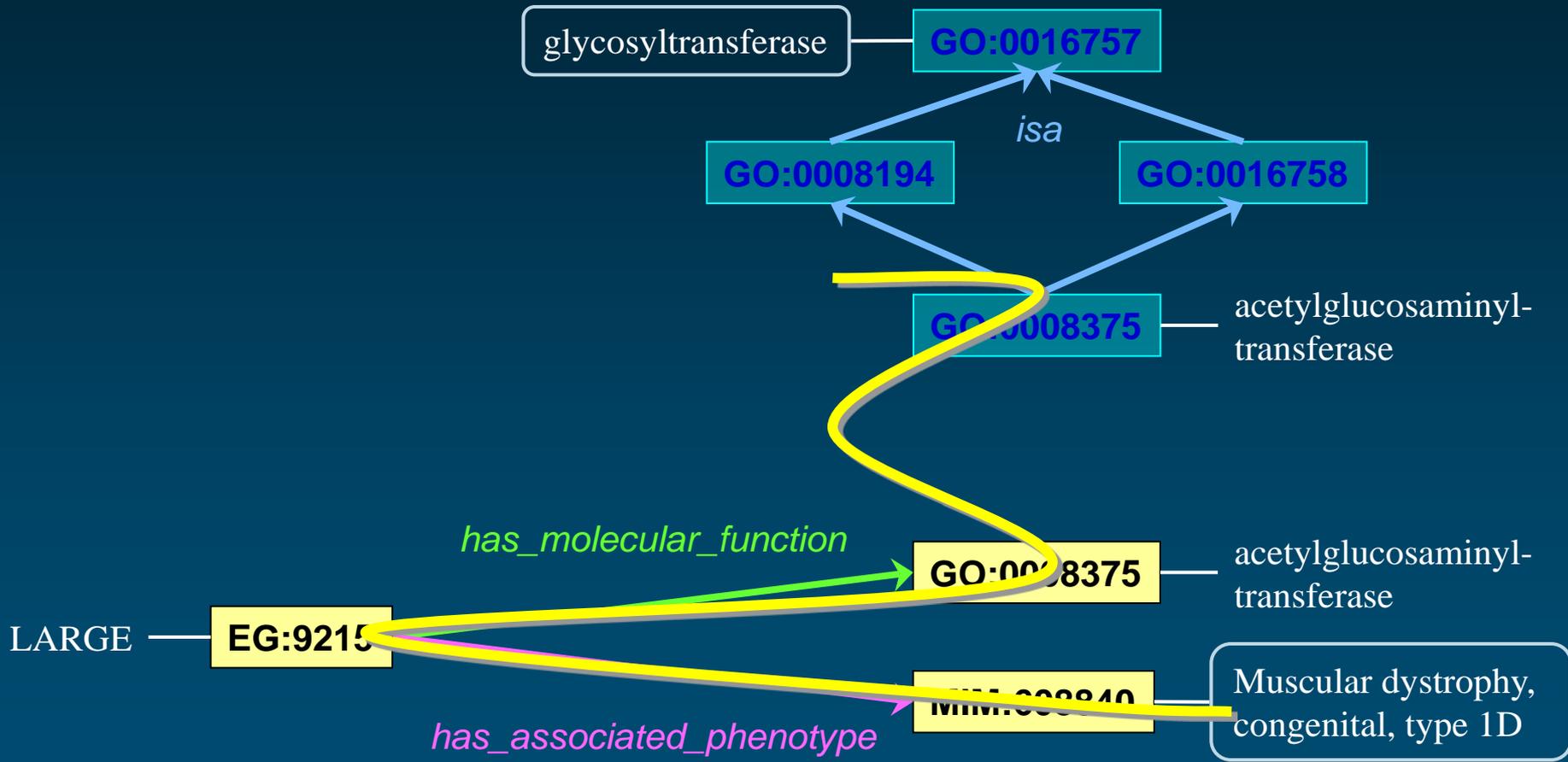
Find all the genes annotated with the GO molecular function glycosyltransferase or any of its descendants and associated with any form of congenital muscular dystrophy



Results Instantiated graph



From *glycosyltransferase* to *congenital muscular dystrophy*



KNOWLEDGE SOURCES

Knowledge sources

- ◆ Ontologies – **definitional knowledge (mostly)**
 - Terminology integration systems
 - Unified Medical Language System (NLM)
 - BioPortal (NCBO)
- ◆ Relations extracted from text – **assertional knowledge (mostly)**
 - Text corpus
 - MEDLINE
 - Relation extraction system
 - SemRep (NLM), MedLEE (Columbia)
 - Commercial systems, specialized systems



Unified Medical Language System



◆ SPECIALIST Lexicon

- 460,000 lexical items
- Part of speech and variant information

◆ Metathesaurus

- 8M names from over 160 terminologies
- 2.7M concepts
- 16M relations

◆ Semantic Network

- 133 high-level categories
- 7000 relations among them

Lexical
resources

Terminological
resources

Ontological
resources



Metathesaurus Basic organization

◆ Concepts

- Synonymous terms are clustered into a concept
- Properties are attached to concepts, e.g.,
 - Unique identifier
 - Definition

◆ Relations

- Concepts are related to other concepts
- Properties are attached to relations, e.g.,
 - Type of relationship
 - Source



Organize terms

- ◆ Synonymous terms clustered into a concept
- ◆ Preferred term
- ◆ Unique identifier (CUI)

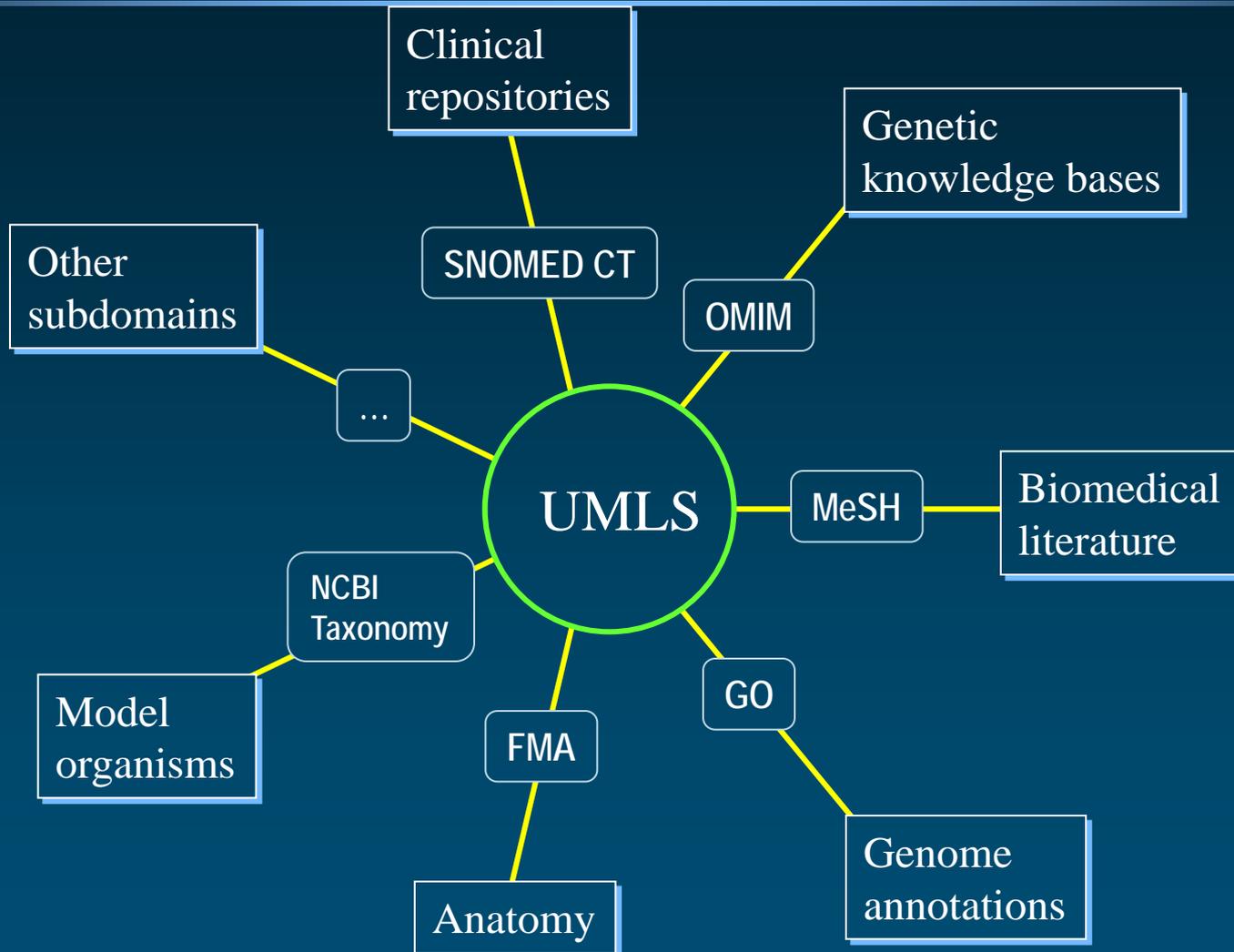
Addison Disease	MeSH	D000224
Primary hypoadrenalism	MedDRA	10036696
Primary adrenocortical insufficiency	ICD-10	E27.1
Addison's disease (disorder)	SNOMED CT	363732003

C0001403

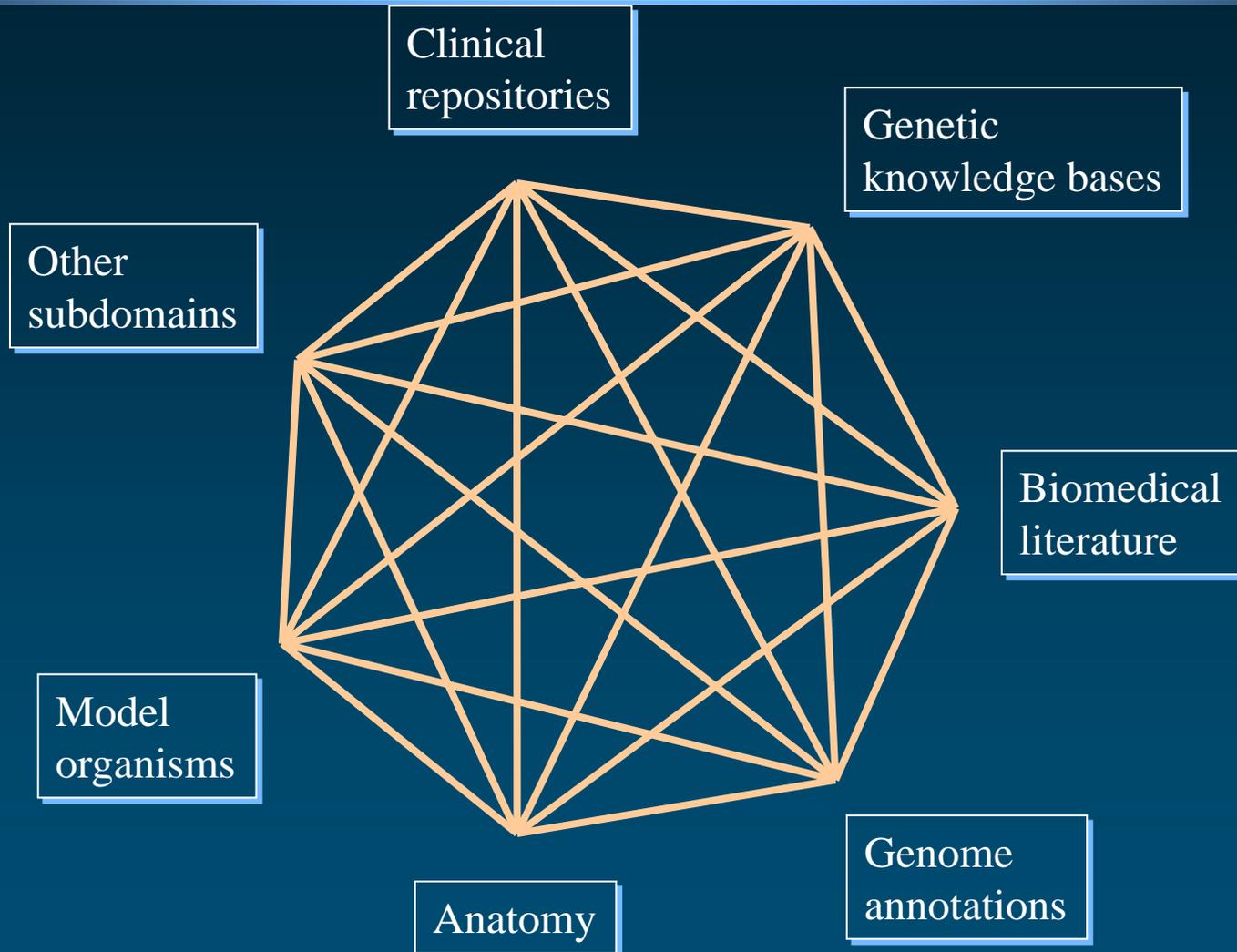
Addison's disease



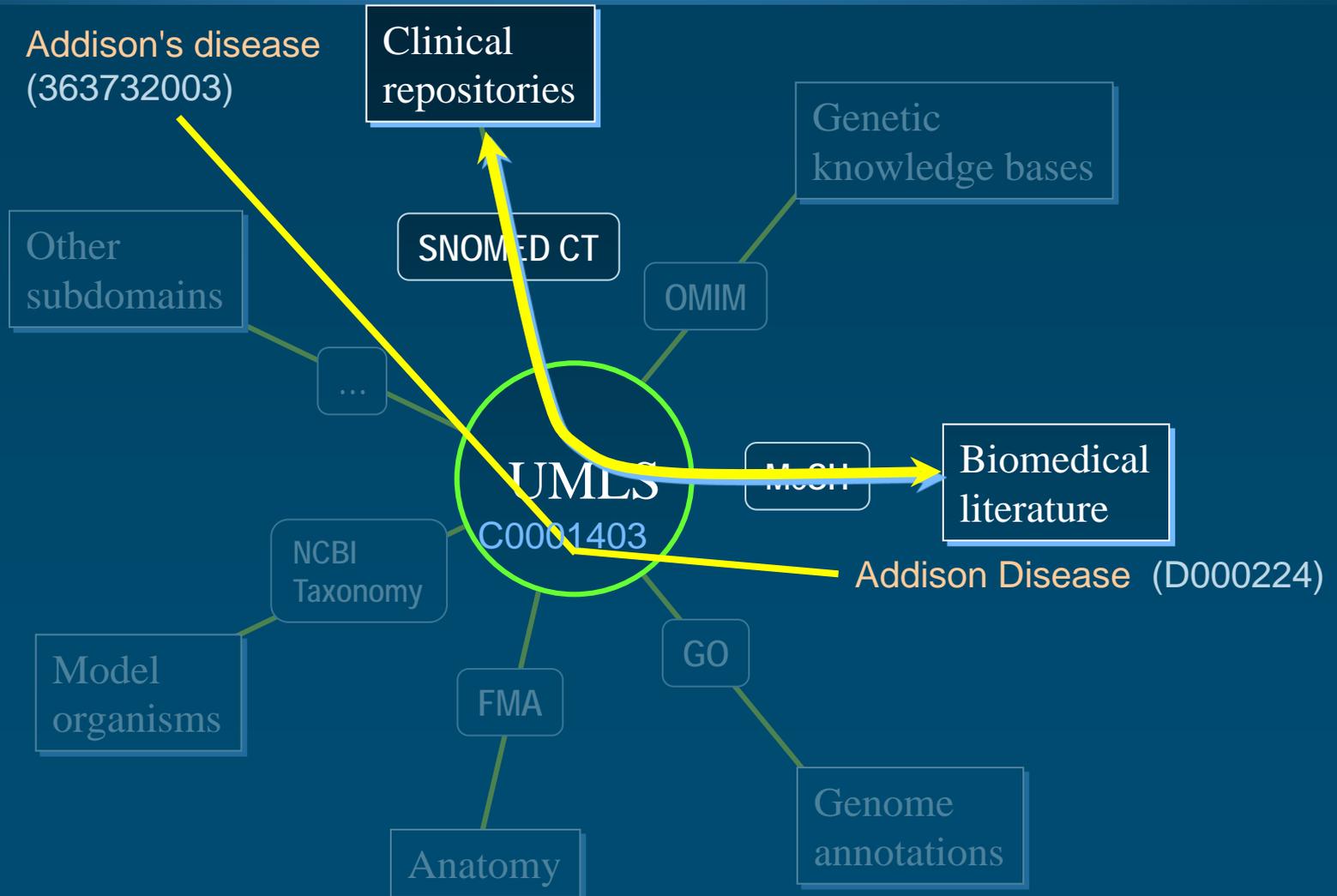
Integrating subdomains



Integrating subdomains

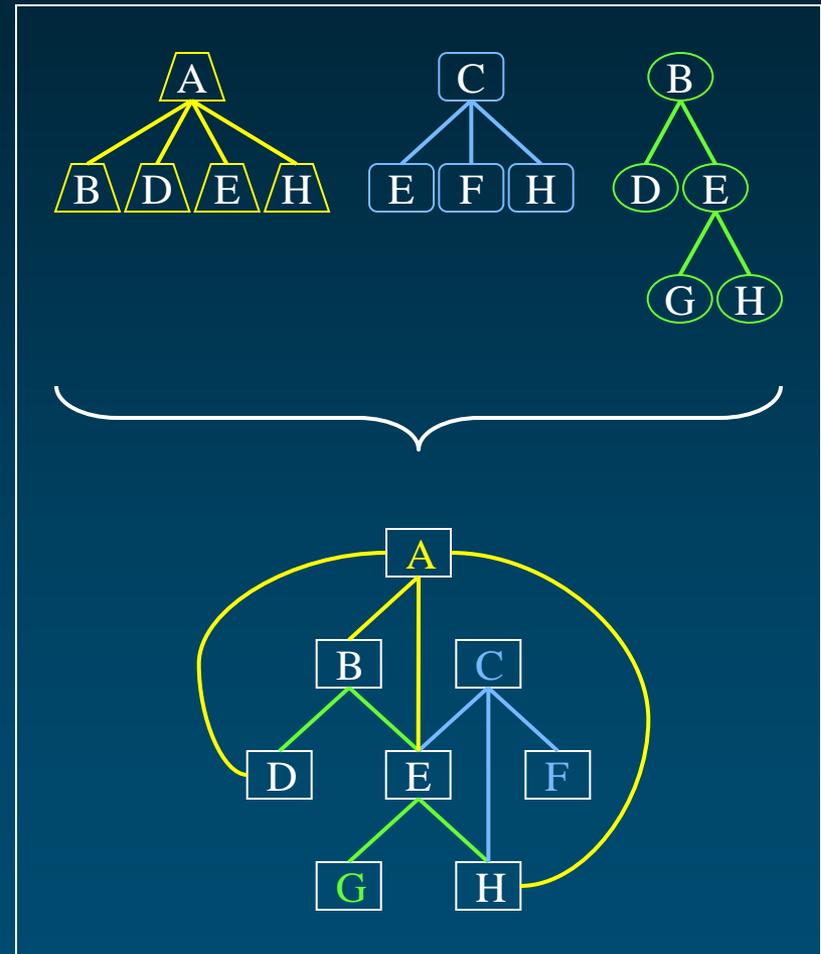


Trans-namespace integration



Organize concepts

- ◆ Inter-concept relationships: hierarchies from the source vocabularies
- ◆ Redundancy: multiple paths
- ◆ One graph instead of multiple trees (multiple inheritance)



SemRep

- ◆ Part of the Semantic Knowledge Representation project at NLM
 - Tom Rindflesch & Marcelo Fiszman
- ◆ Knowledge extraction system for the automatic summarization system SemanticMEDLINE
 - <http://skr3.nlm.nih.gov/SemMedDemo/>



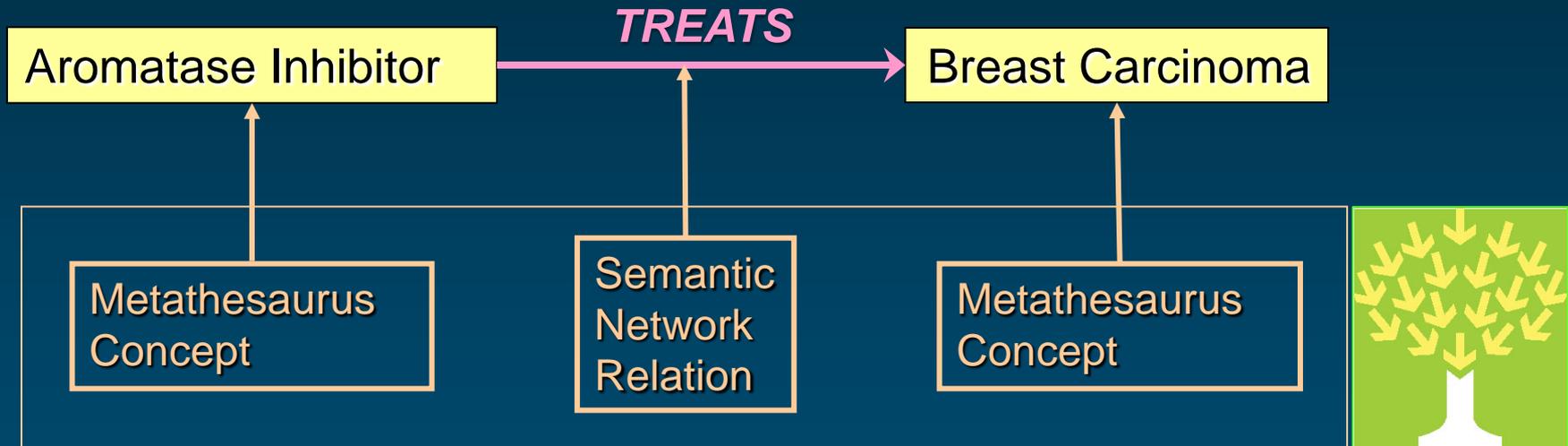
SemRep

- ◆ Extract semantic predications from biomedical research literature (MEDLINE citations)
- ◆ Based on
 - Generalizations about the structure of English
 - Structured domain knowledge: UMLS
- ◆ Balances linguistic insight with practical implementation
 - Underspecified syntax
 - Core predications only
 - Limited by domain



SemRep: Extract Predication

... Exemestane after non-steroidal aromatase inhibitor **for** post-menopausal women with advanced **breast cancer**



Unified Medical Language System

Several Evaluations

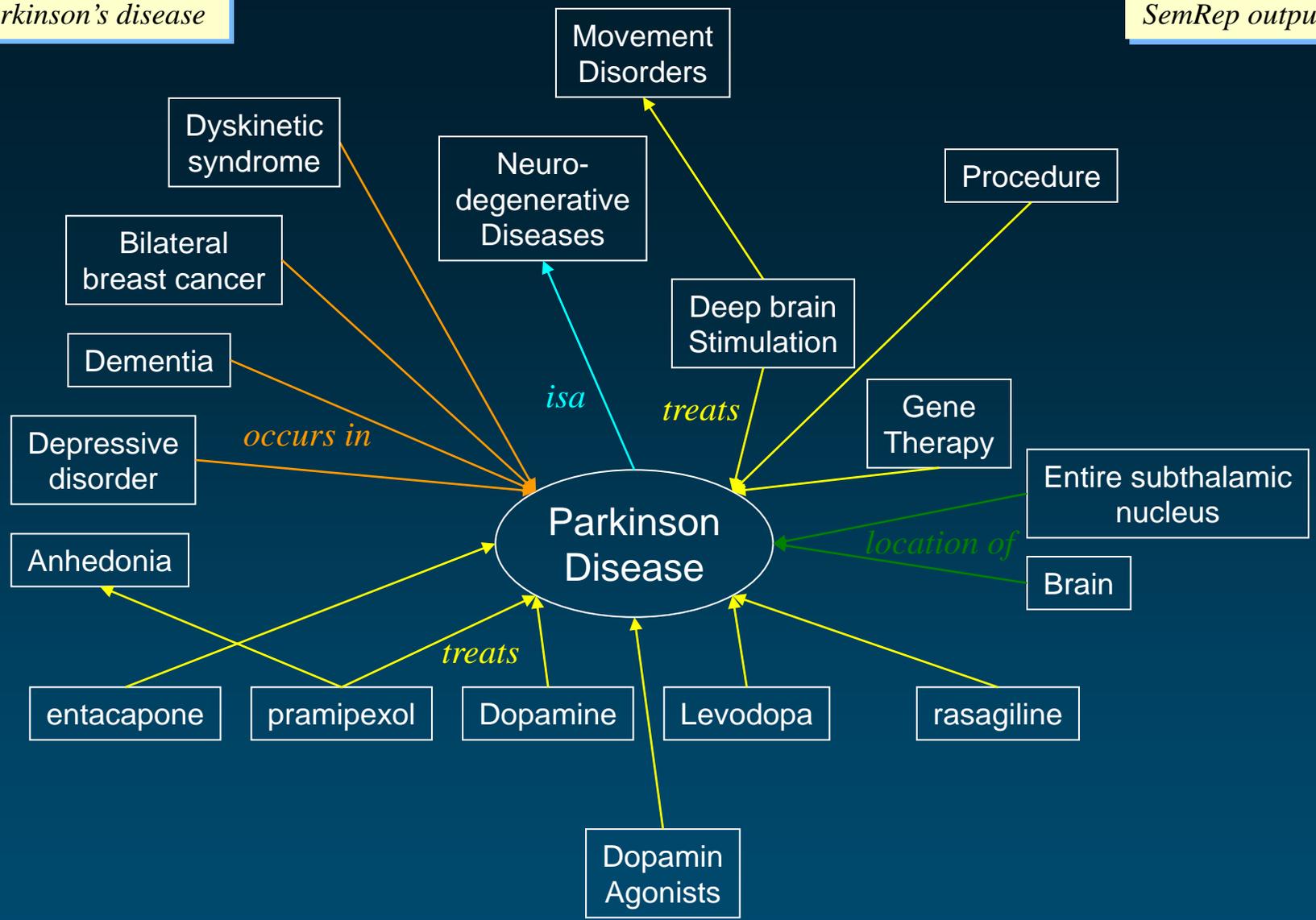
- ◆ Focused on biomedical subdomains, e.g.
 - Clinical treatment, genetic etiology of disease, pharmacogenomics
- ◆ Focused on structure, e.g.
 - Hypernymic predications, comparatives, nominalizations
- ◆ Overall
 - Precision is around 75% (lower for molecular biology)
 - Recall is around 60%

Predication Database: SemMedDB

- ◆ Processed all of MEDLINE
 - More than 21 million citations
 - Titles and abstracts
- ◆ SemRep predications extracted
 - 57 million predications (through 06/30/2012)
- ◆ Made available to the research community
 - MySQL database
 - RDF triples

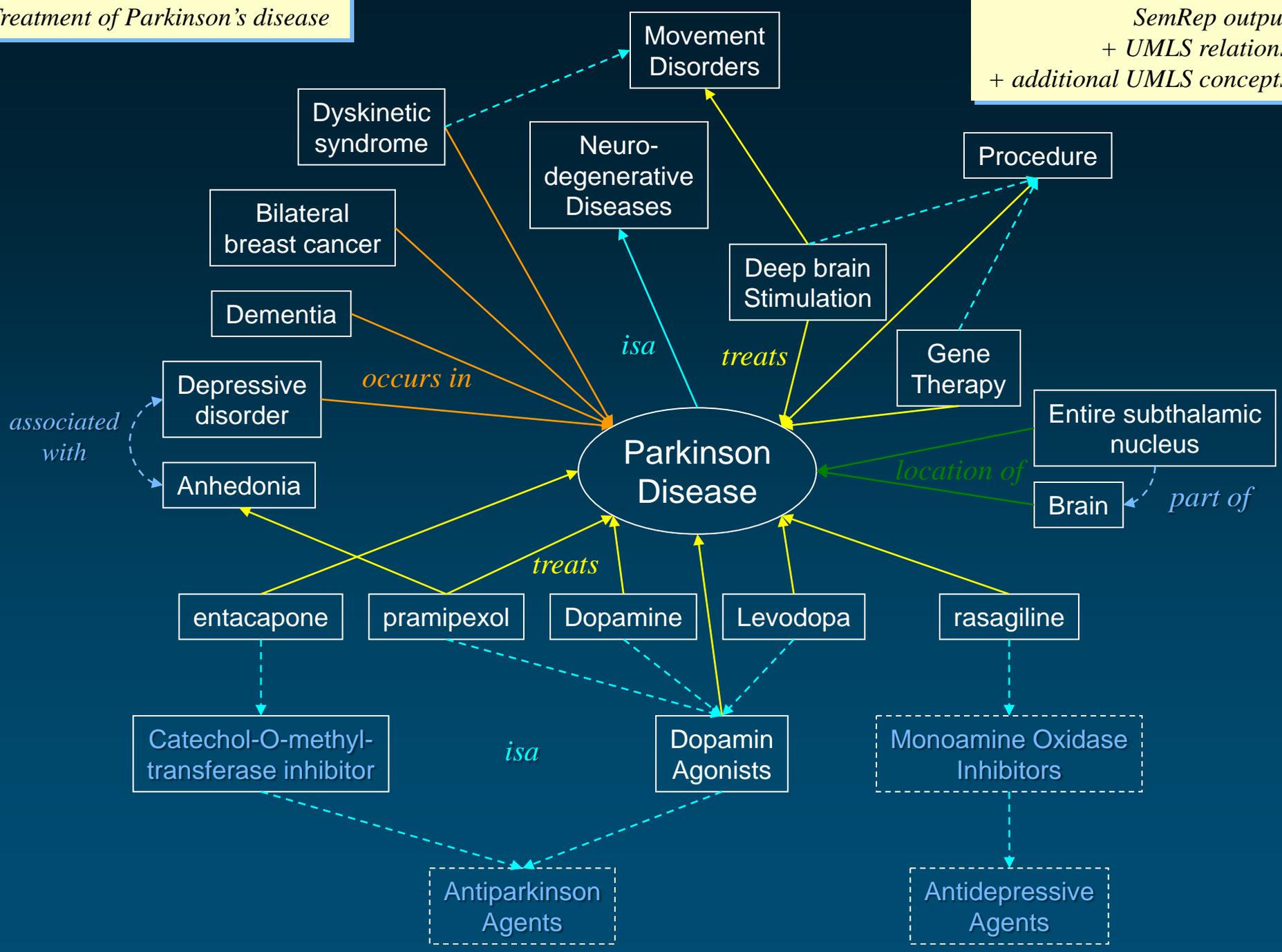


INTEGRATING RELATIONS FROM TEXT MINING AND ONTOLOGIES

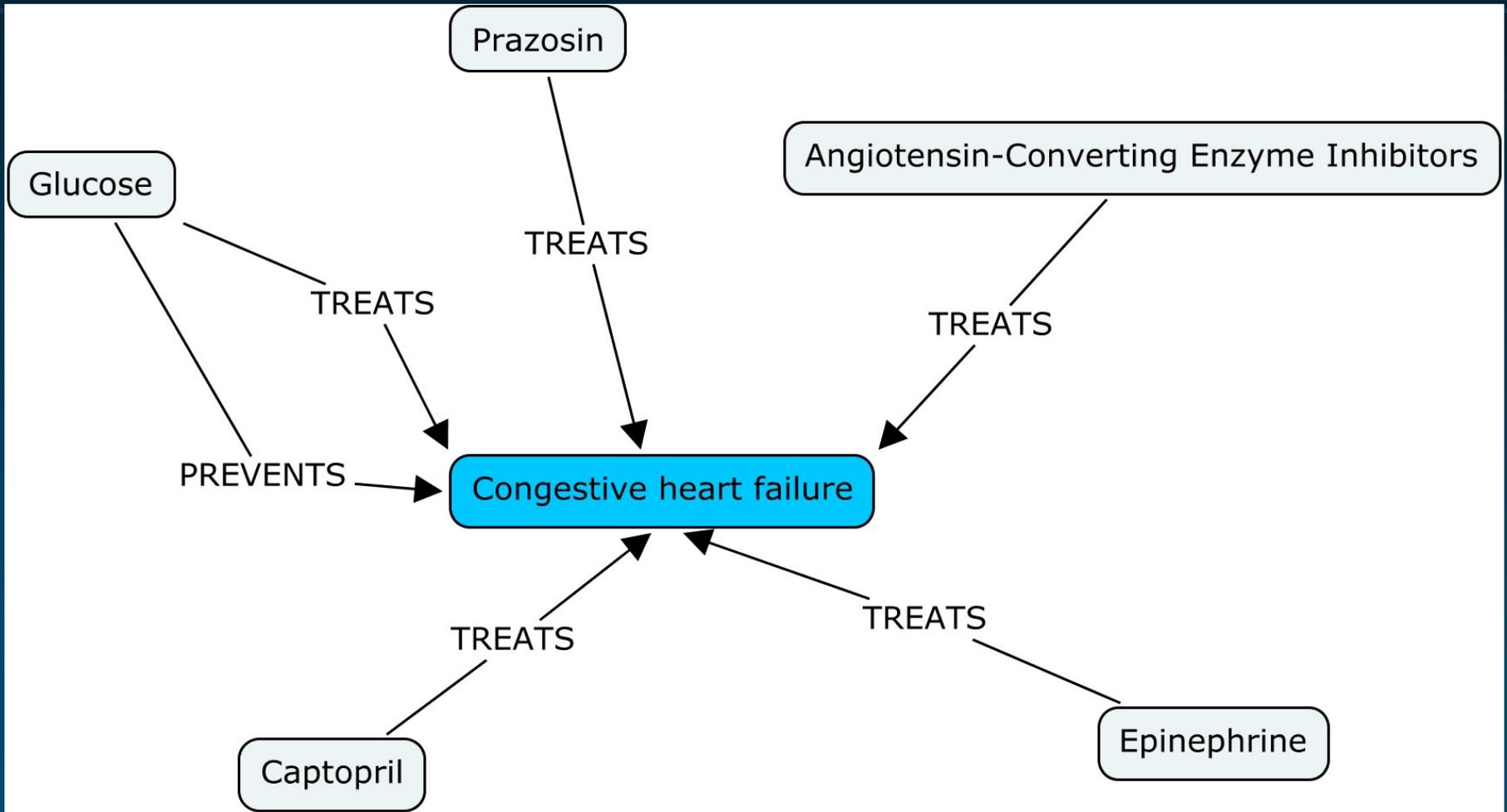


Treatment of Parkinson's disease

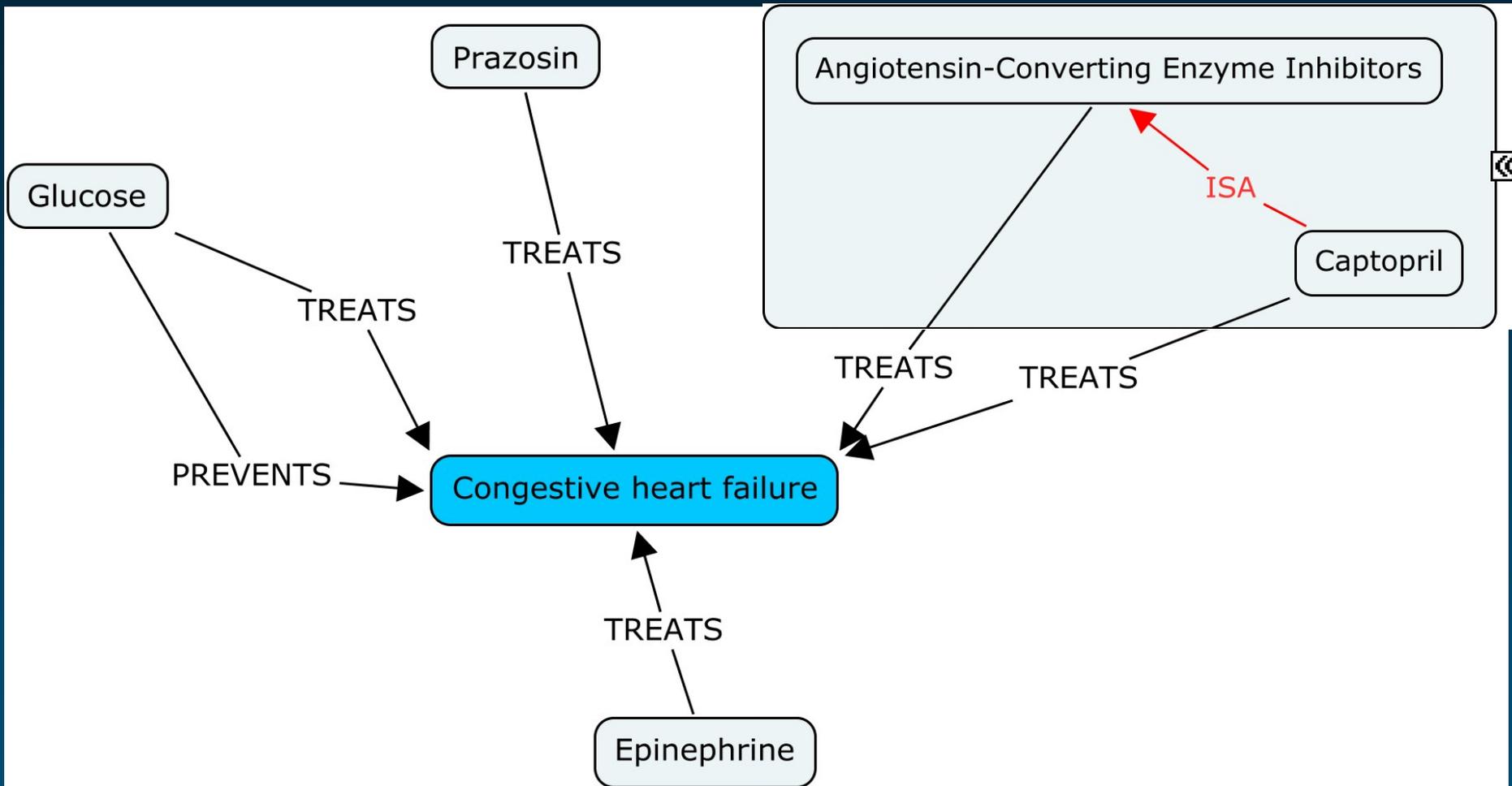
SemRep output
+ UMLS relations
+ additional UMLS concepts



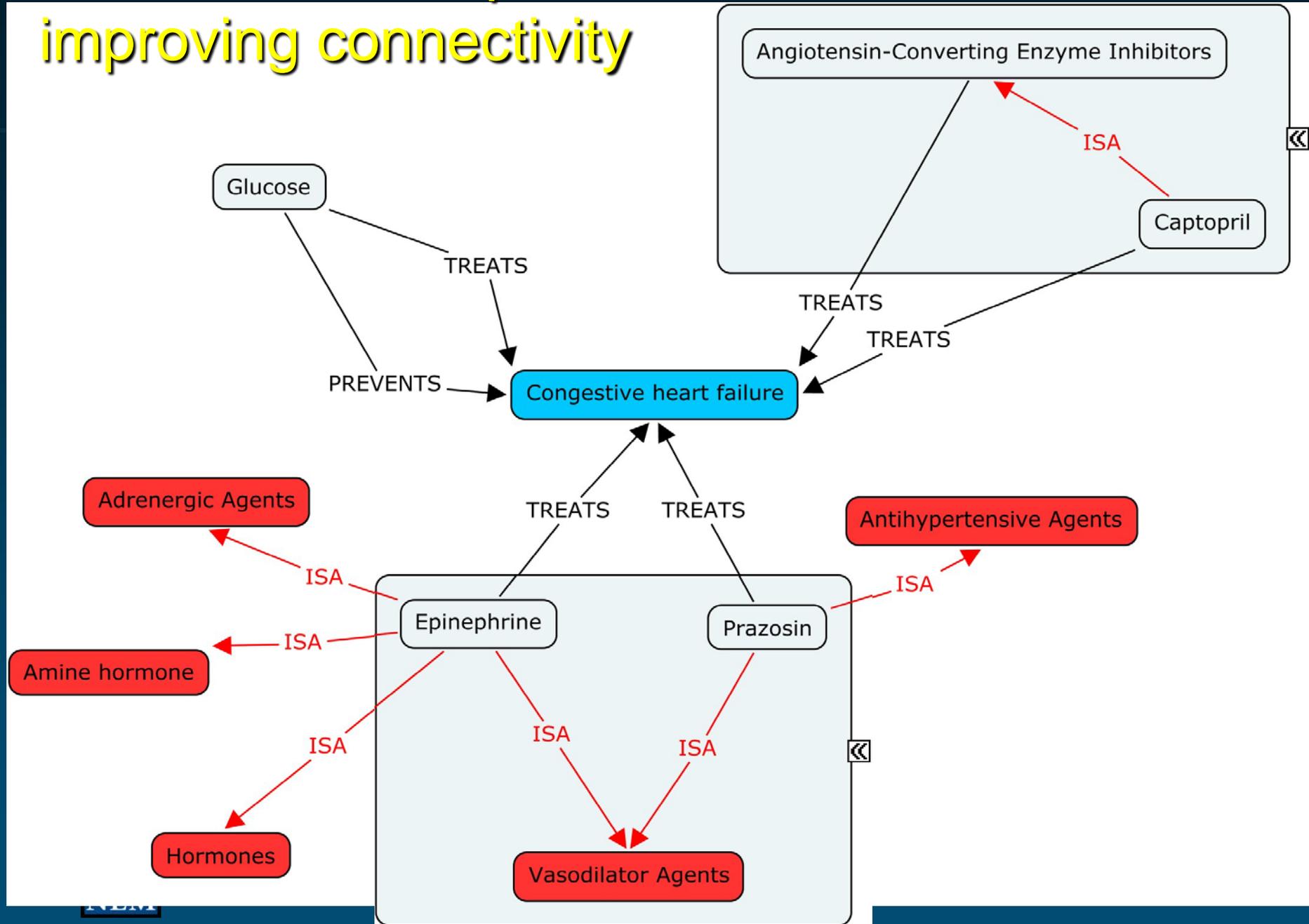
Original graph



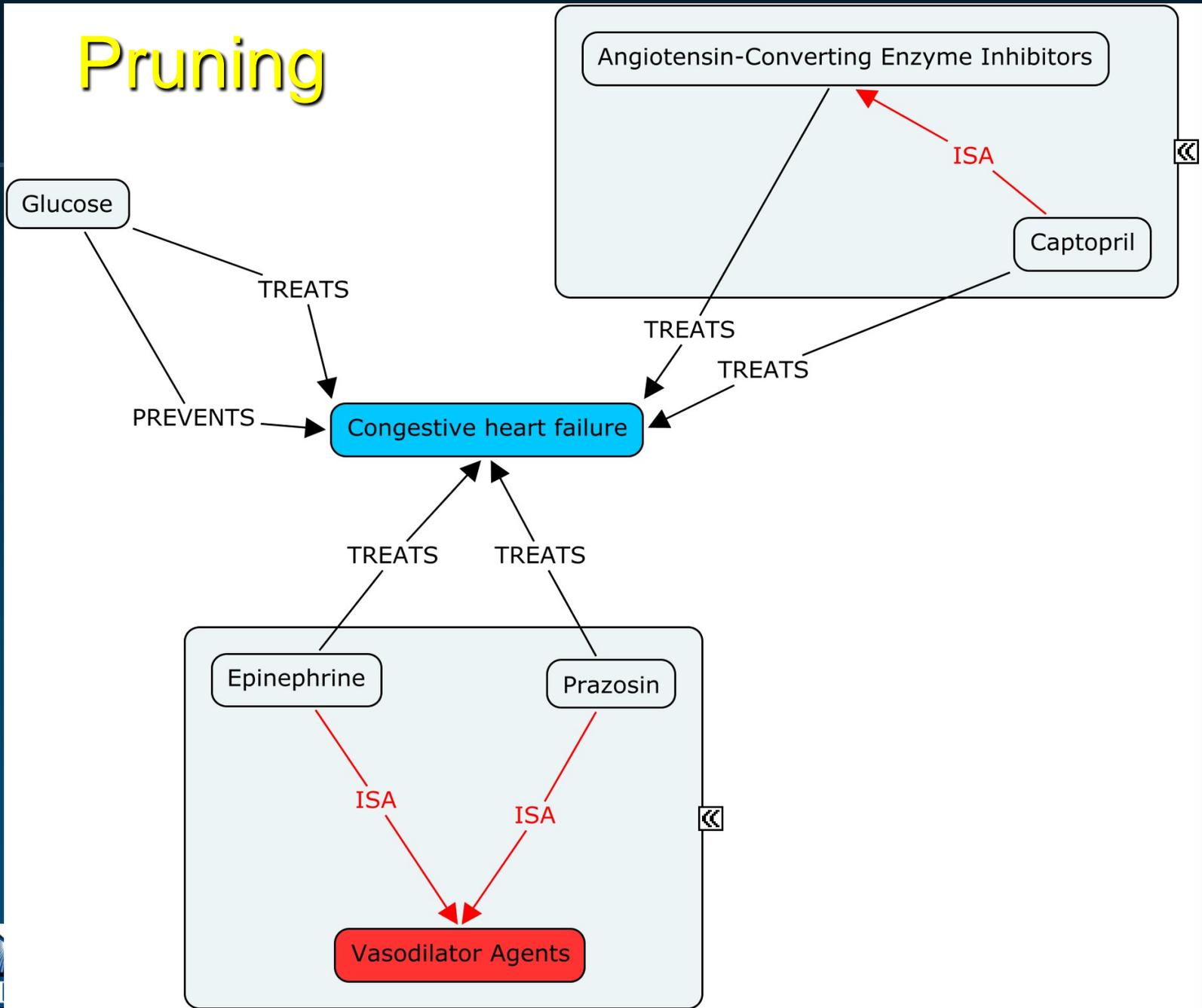
Adding hierarchy between any two concepts in the graph



Add UMLS concepts for improving connectivity



Pruning



Congestive heart failure

PREVENTS TREATS TREATS TREATS TREATS TREATS

Glucose

Epinephrine

Prazosin

Captopril

Caloric Agent

Vasodilator Agents

Angiotensin-Converting Enzyme Inhibitors

replacement preparation

hypotensive agent

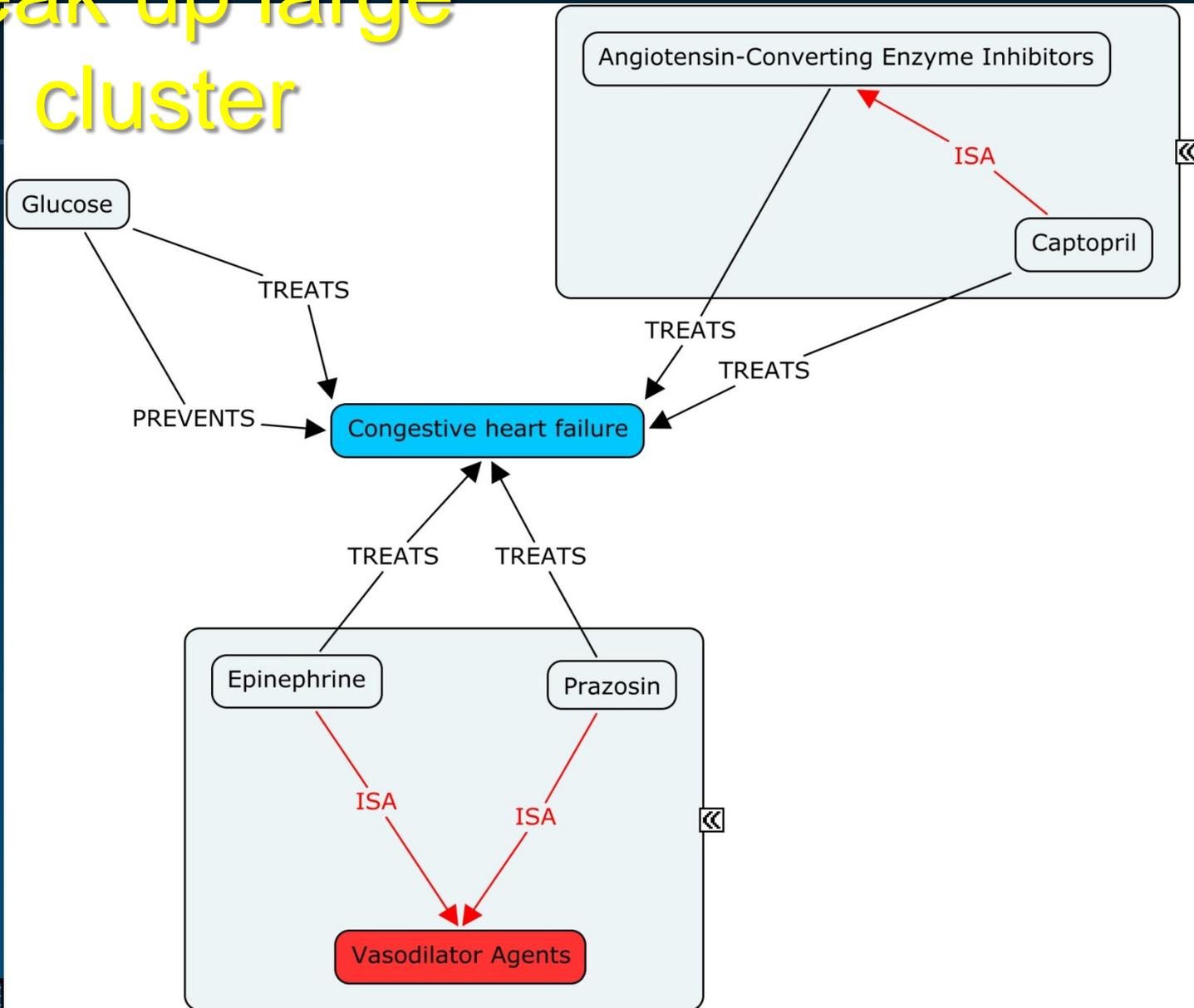
cardiovascular drug

Pharmacologic Substance

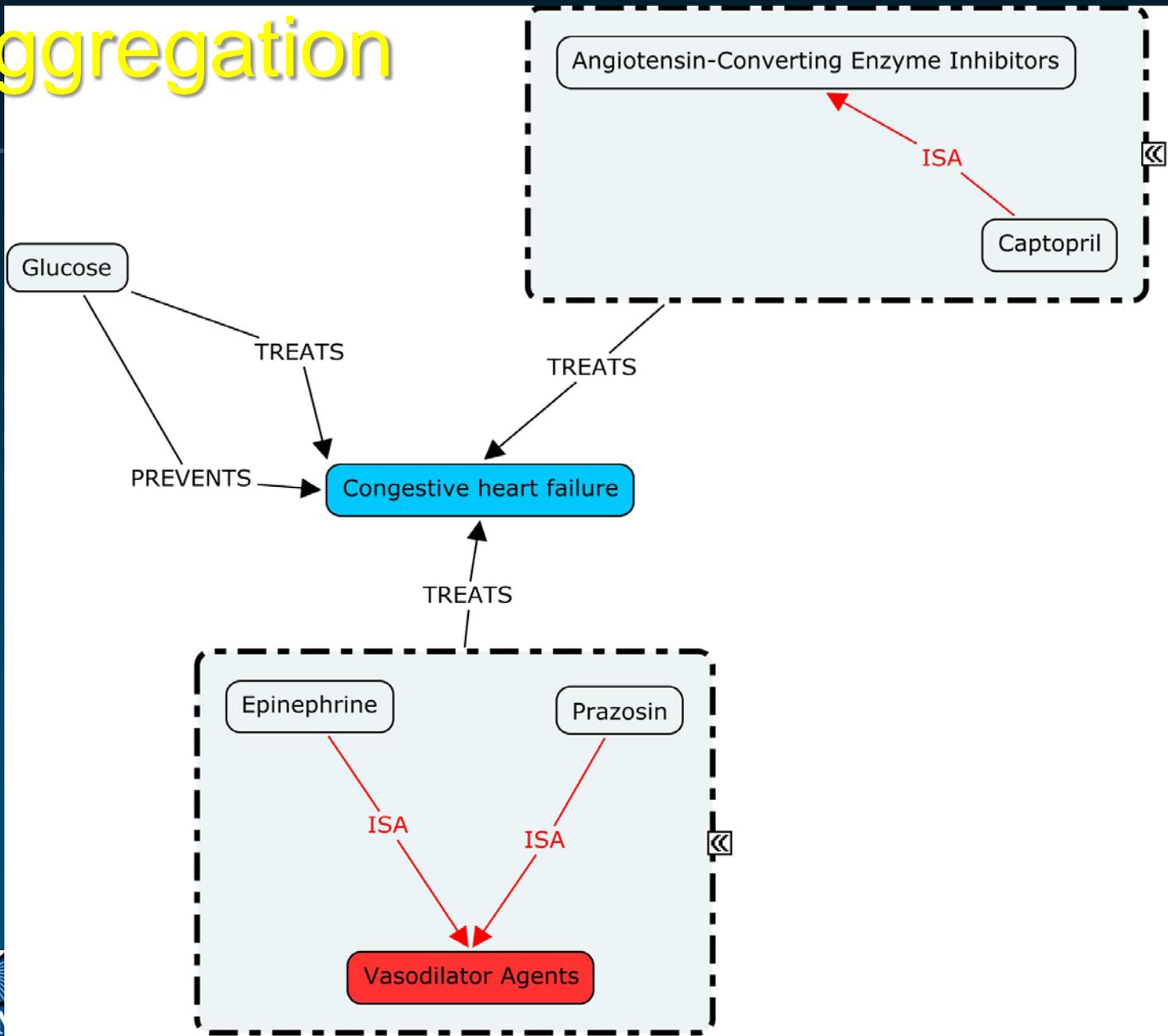
ISA



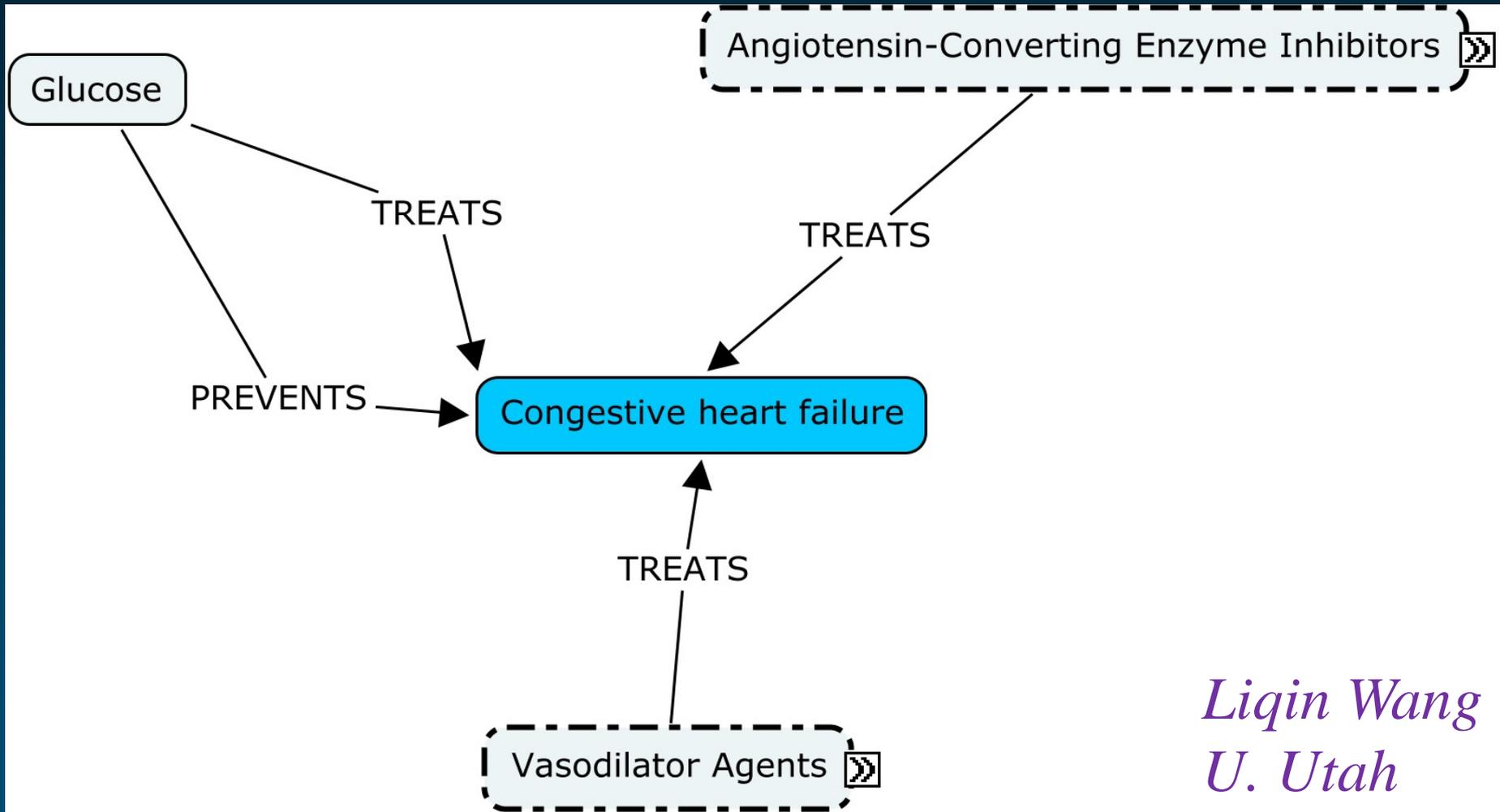
Break up large cluster



Aggregation



After aggregation



Liqin Wang
U. Utah

BIOMEDICAL KNOWLEDGE REPOSITORY

Biomedical Knowledge Repository

- ◆ Integrated set of relations
 - From the UMLS Metathesaurus
 - Extracted from MEDLINE by SemRep
- ◆ Together with metadata
 - Source of the relations (provenance)
- ◆ Semantic Web technologies
 - RDF store (Virtuoso)



Representation

Non-contextualized
relation (semantic
type level)

Pharm. substance *treats* Disease or Syndrome

↑
ACE Inhibitors

↑
Cardiovascular disease

Non-contextualized
relation (class level)

↑
Captopril *treats*

↑
Congestive heart failure

Contextualized
relation (instance level)

↑
Captopril *treats*
PMID:12345

↑
Congestive heart failure

Metadata

↓
PMID:12345

publication_date

9/4/2012



Status

- ◆ Experimental
- ◆ Fully populated
 - UMLS 2012AA
 - 50M relations extracted from MEDLINE
- ◆ SemMedDB available for download
- ◆ UMLS in RDF not yet available for download
- ◆ Not available as a SPARQL endpoint
 - Licensing issues
 - Lack of access control in RDF stores



Potential applications

- ◆ Multi-document summarization
 - Semantic MEDLINE “plus”
- ◆ Information retrieval of relations
 - Beyond keywords or concepts
- ◆ Simple question answering
 - Which drugs treat congestive heart failure?
- ◆ Knowledge discovery
 - Swanson’s paradigm (e.g., finding “B”s)
 - Patterns of relations

A knowledge discovery platform

Sleep. 2012 Feb 1;35(2):279-85.

A closed literature-based discovery technique finds a mechanistic link between hypogonadism and diminished sleep quality in aging men.

Miller CM, Rindfleisch TC, Fizman M, Hristovski D, Shin D, Rosemblat G, Zhang H, Strohl KP.

National Institutes of Health, National Library of Medicine, Cognitive Science Branch, Bethesda, MD 20894, USA. millercm@mail.nih.gov

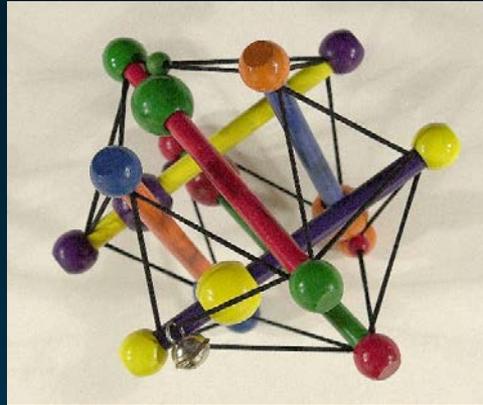
Abstract

STUDY OBJECTIVES: Sleep quality commonly diminishes with age, and, further, aging men often exhibit a wider range of sleep pathologies than women. We used a freely available, web-based discovery technique (Semantic MEDLINE) supported by semantic relationships to automatically extract information from MEDLINE titles and abstracts.

DESIGN: We assumed that testosterone is associated with sleep (the A-C relationship in the paradigm) and looked for a mechanism to explain this association (B explanatory link) as a potential or partial mechanism underpinning the etiology of eroded sleep quality in aging men.

MEASUREMENTS AND RESULTS: Review of full-text papers in critical nodes discovered in this manner resulted in the proposal that testosterone enhances sleep by inhibiting cortisol. Using this discovery method, we posit, and could confirm as a novel hypothesis, cortisol as part of a mechanistic link elucidating the observed correlation between decreased testosterone in aging men and diminished sleep quality.

CONCLUSIONS: This approach is publically available and useful not only in this manner but also to generate from the literature alternative explanatory models for observed experimental results.



Medical Ontology Research

Contact: olivier@nlm.nih.gov

Web: mor.nlm.nih.gov



Olivier Bodenreider

Lister Hill National Center
for Biomedical Communications
Bethesda, Maryland - USA

IEEE International Conference on Bioinformatics and Biomedicine (BIBM 2012)

Philadelphia, USA October 4-7, 2012

BIBM 2012



Home

Important Dates

Paper Submission

Workshop Organizers

Invited Speaker

Accepted Papers

The First International Workshop on the role of Semantic Web in Literature-Based Discovery

<http://knoesis.org/swlbd2012/>

(SWLBD2012)

in conjunction with

The IEEE International Conference on Bioinformatics and Biomedicine (BIBM 2012)

<http://www.ischool.drexel.edu/ieeebibt/bibt12/>

October 4-7, 2012, Philadelphia PA, USA