A computational method for the extraction of pharmacogenomic relationships from text

Adrien Coulet[1,2], Nigam Shah[2], Yael Garten[2], Mark Musen[2], Russ Altman[2]

1 LORIA, INRIA Nancy – Grand-Est

2 Stanford University

The NCBO and PharmGKB

- A joint project

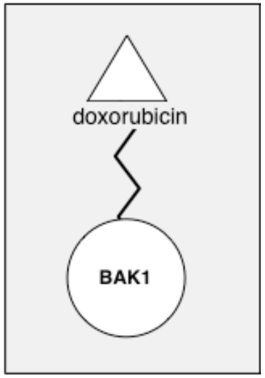
NATIONAL CENTER FOR BIOMEDICAL ONTOLOGY & PharmGKB

- Content of PharmGKB

  - <u>Current</u>:

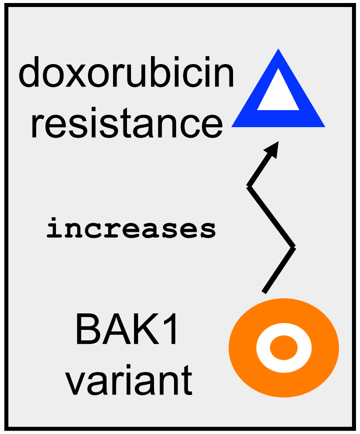
  pharmacogenomics (PGx) relationships

  Gene – Drug ; Gene – Disease ; Drug – Disease



  - <u>Goal</u>:

  to provide more precise relationships



Population of PharmGKB

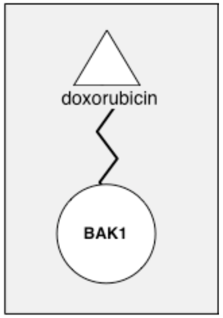
Sentence 1: BAK1 gene polymorphism affects doxorubicin resistance.

Sentence 2: Resistance to Doxorubicin is influenced by BAK1 variants.

Sentence 3: Doxorubicin induces BAK1 activity.

Scientific literature

PharmGKB curators



Population of PharmGKB

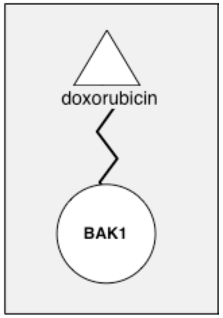
Sentence 1: BAK1 gene polymorphism affects doxorubicin resistance.

Sentence 2: Resistance to Doxorubicin is influenced by BAK1 variants.

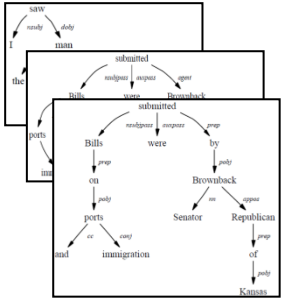
Sentence 3: Doxorubicin induces BAK1 activity.

Scientific literature

PharmGKB curators

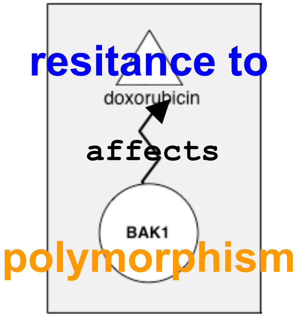


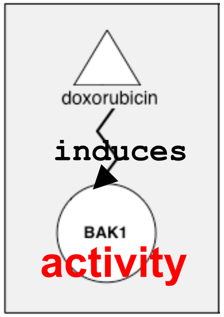
Dependency Graph parsing



Dependency Graphs of sentences

Relation extraction





Population of PharmGKB

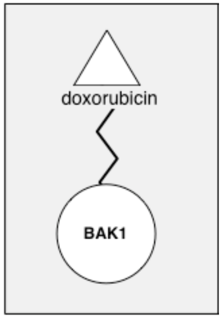
Sentence 1: BAK1 gene polymorphism affects doxorubicin resistance.

Sentence 2: Resistance to Doxorubicin is influenced by BAK1 variants.

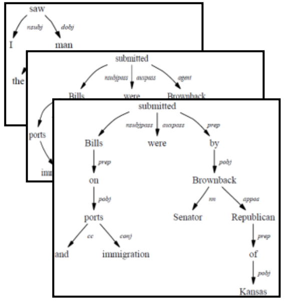
Sentence 3: Doxorubicin induces BAK1 activity.

Scientific literature

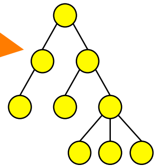
PharmGKB curators



Dependency Graph parsing

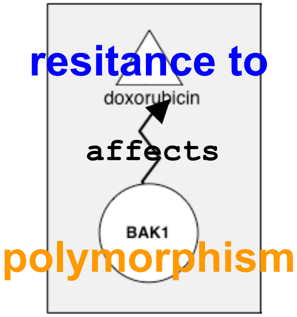


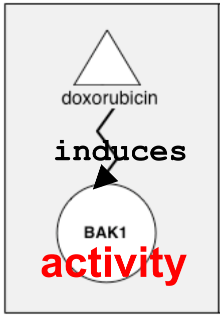
Dependency Graphs of sentences



ontology

Relation extraction





Outline

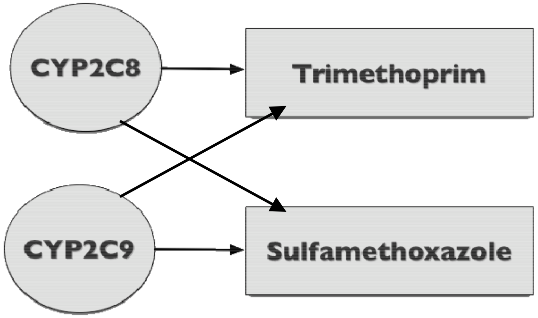
1. Limitations of co-occurrences

2. Construction of a knowledge base
    1. Algorithm to extract raw relationships
    2. Semi-automated ontology building
    3. Knowledge base content from 1 & 2

Limitations of co-occurrence (that we wanted to solve)

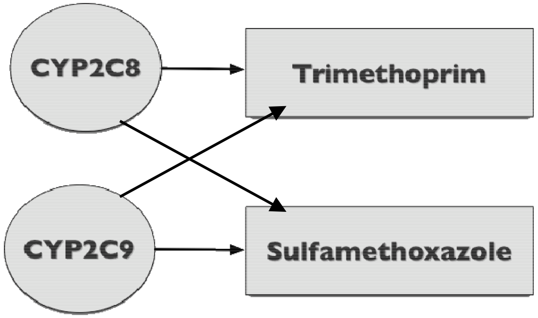
1. Avoid false positive connections

"Trimethoprim inhibits activity of CYP2C8 while sulfamethoxazole inhibits CYP2C9 activity."



Limitations of co-occurrence (that we wanted to solve)

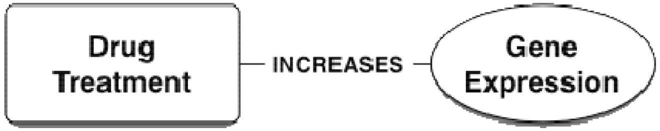
1. Avoid false positive connections



"Trimethoprim inhibits activity of CYP2C8 while sulfamethoxazole inhibits CYP2C9 activity."

2. Characterize fine-grain semantics of relationships

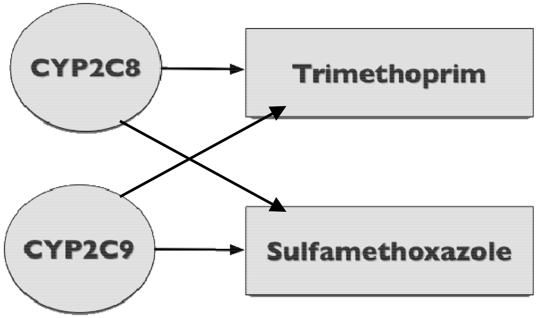
"CYP3A4 mRNA expression was increased significantly by rifampicin exposure in human hepatocytes."



Limitations of co-occurrence (that we wanted to solve)

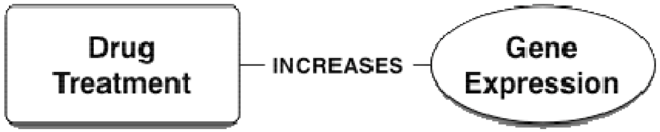
1. Avoid false positive connections

“Trimethoprim inhibits activity of CYP2C8 while sulfamethoxazole inhibits CYP2C9 activity.”



2. Characterize fine-grain semantics of relationships

“CYP3A4 mRNA expression was increased significantly by rifampicin exposure in human hepatocytes.”



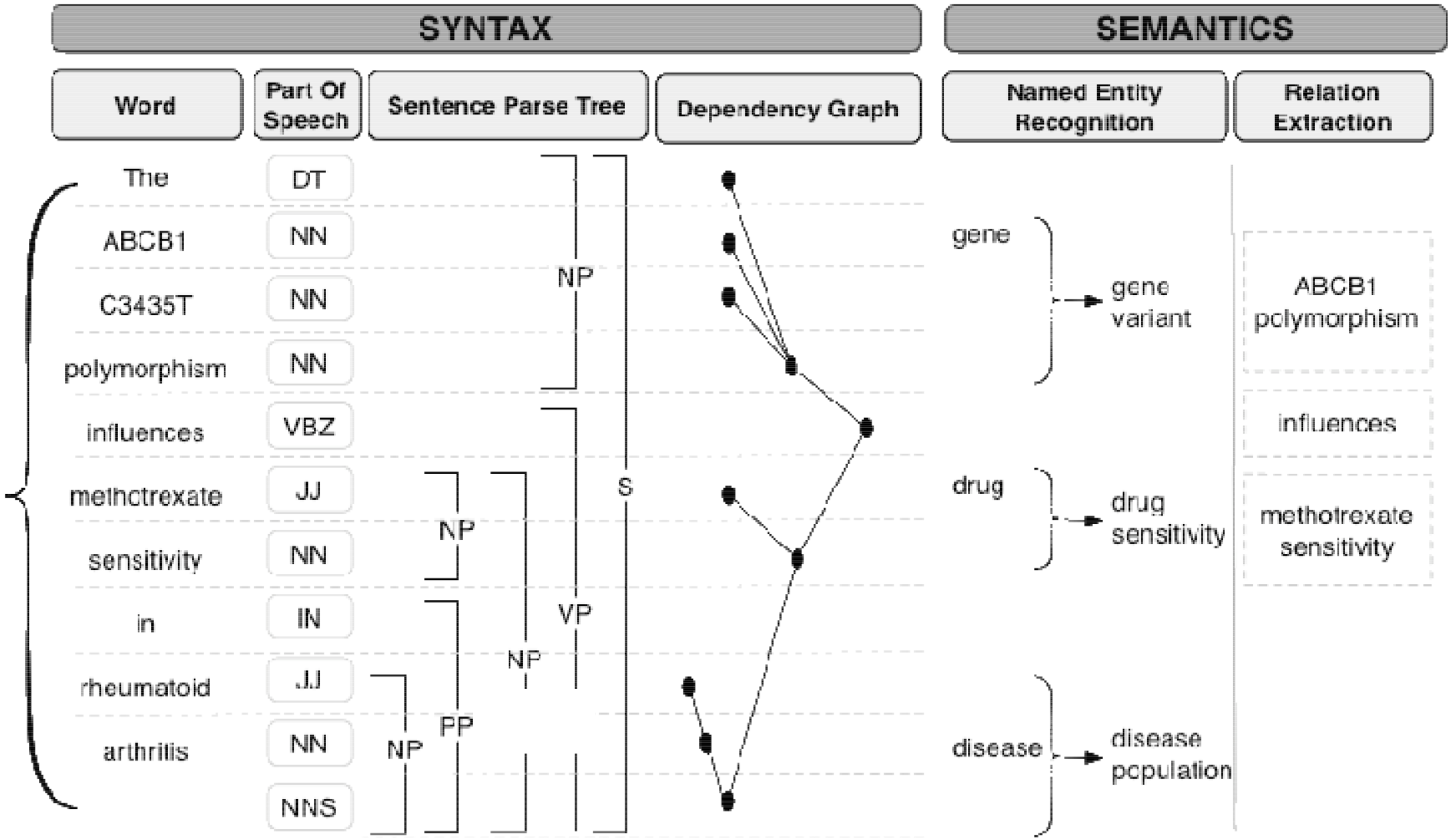
3. To consolidate synonyms (normalize):
   - Between complex entity names:

synthesis of PGE2
PGE2 formation → dinoprostone_synthesis
Prostaglandin E2 production

   - Between relationships:

inhibit
repress → INHIBIT
antagonize

Several steps of text processing enable extracting relationship semantics



SYNTAX

SEMANTICS

Word	Part Of Speech
The	DT
ABCB1	NN
C3435T	NN
polymorphism	NN
influences	VBZ
methotrexate	JJ
sensitivity	NN
in	IN
rheumatoid	JJ
arthritis	NN
	NNS

Named Entity Recognition

gene → gene variant

drug → drug sensitivity

disease → disease population

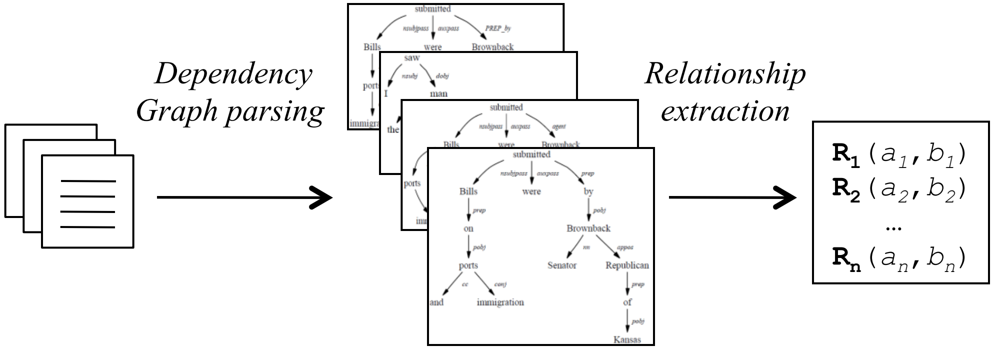
Relation Extraction

ABCB1 polymorphism

influences

methotrexate sensitivity

Issue: we extracted heterogeneous relationships



Dependency Graph parsing

Relationship extraction

$R_1(a_1, b_1)$
$R_2(a_2, b_2)$
…
$R_n(a_n, b_n)$

MEDLINE abstracts

Dependency Graphs of sentences

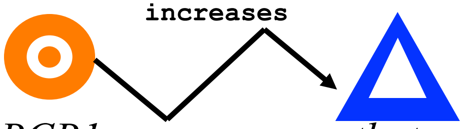
Raw relationships

~17,000,000 abstracts

~87,000,000 dependency graphs

~41,000 raw relationships

Example:



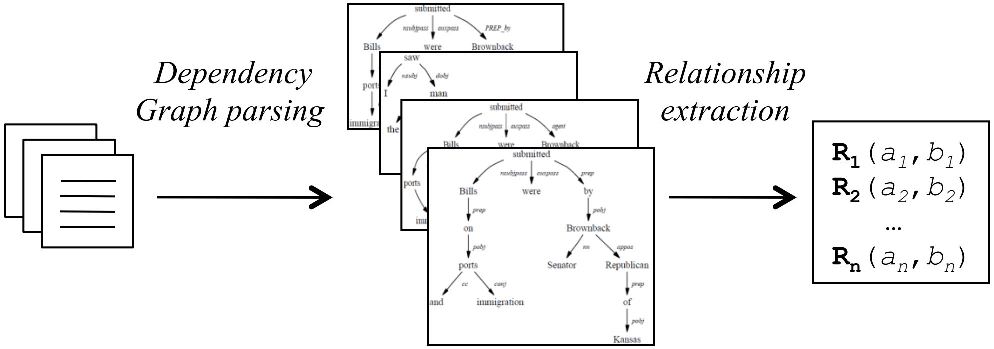
ABCB1 variant

methotrexate sensitivity

increase(P-gp_variants, methopterin_intolerance)

augment(ABCB1_SNPs, methotrexate_sensitivity)

Issue: we extracted heterogeneous relationships



Dependency Graph parsing

Relationship extraction

$\mathbf{R_1}(a_1, b_1)$
$\mathbf{R_2}(a_2, b_2)$
…
$\mathbf{R_n}(a_n, b_n)$

MEDLINE abstracts

Dependency Graphs of sentences

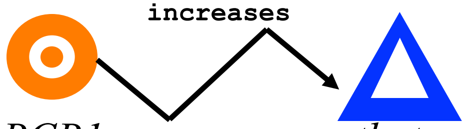
Raw relationships

~17,000,000 abstracts

~87,000,000 dependency graphs

~41,000 raw relationships

Example:



ABCB1 variant

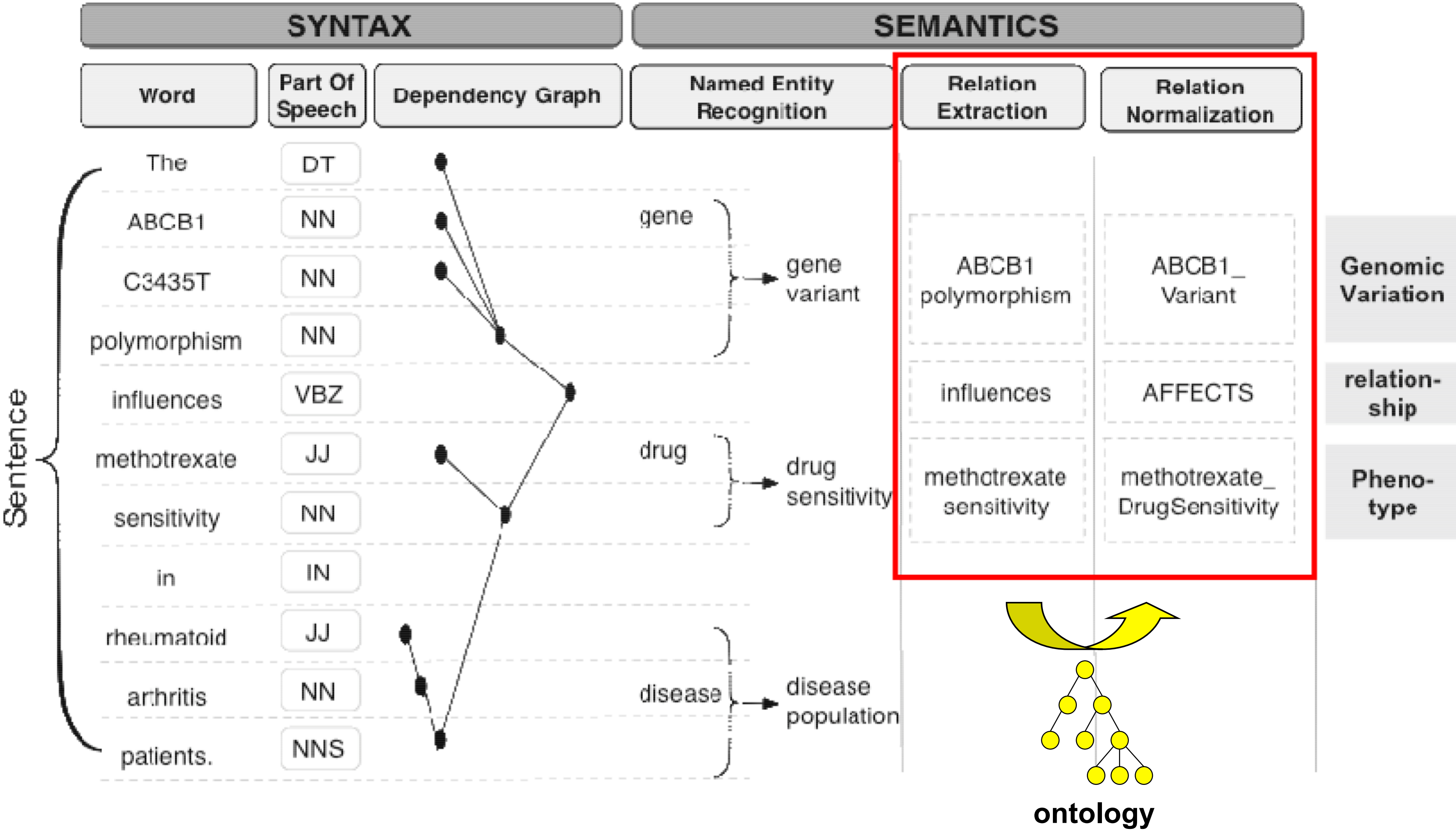
methotrexate sensitivity

increase(P-gp_variants, methopterin_intolerance)

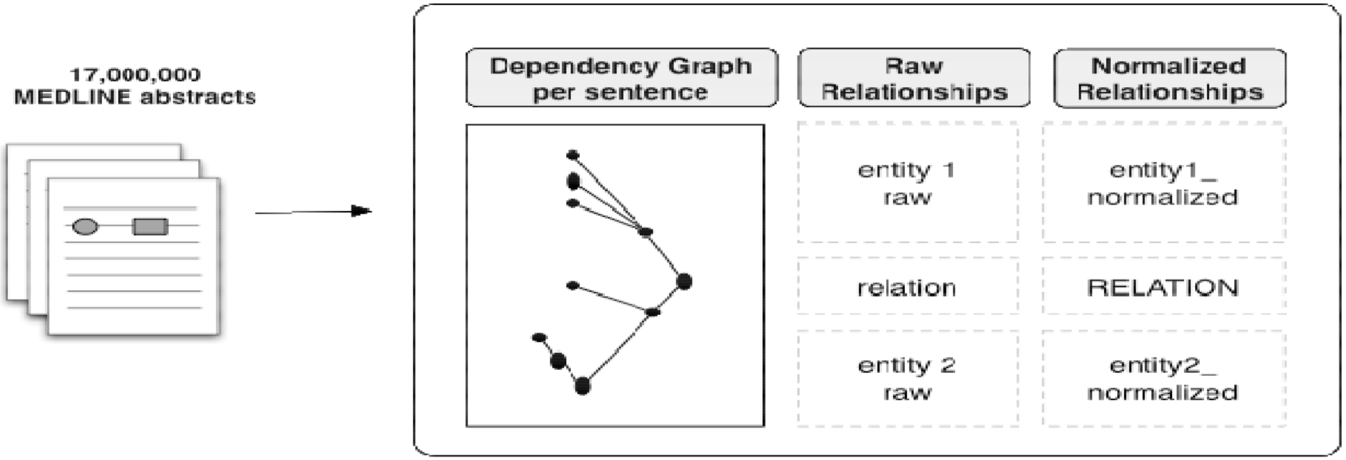
augment(ABCB1_SNPs, methotrexate_sensitivity)

- There is no relation ontology for most of specialized domains
- We created one from extracted relationships

We built and use an ontology to normalize relationships



We manually created a PGx ontology “bottom-up”



17,000,000 MEDLINE abstracts

Dependency Graph per sentence

Raw Relationships

Normalized Relationships

entity 1 raw

entity1_ normalized

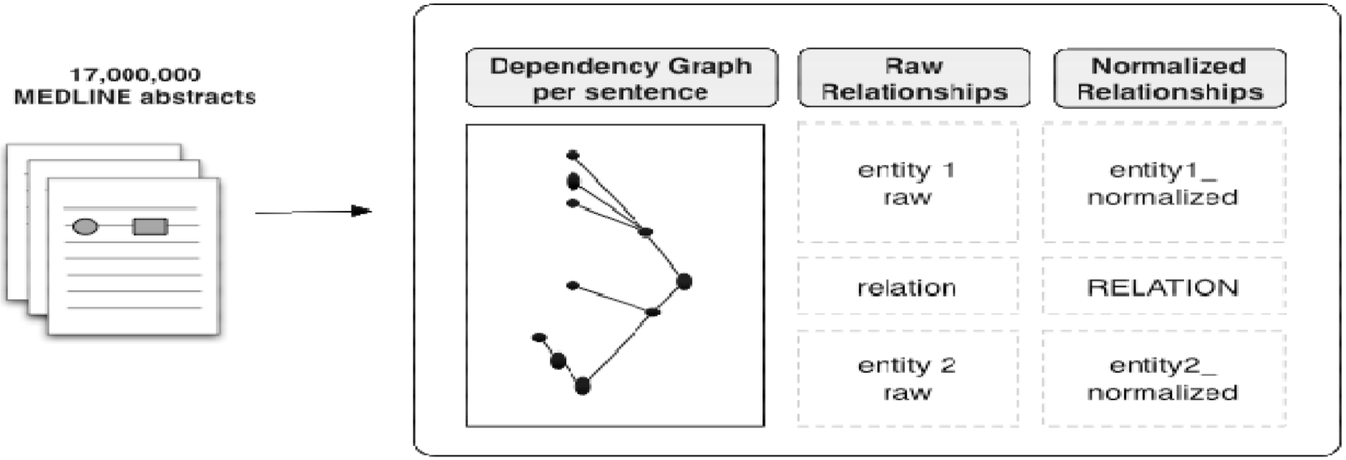
relation

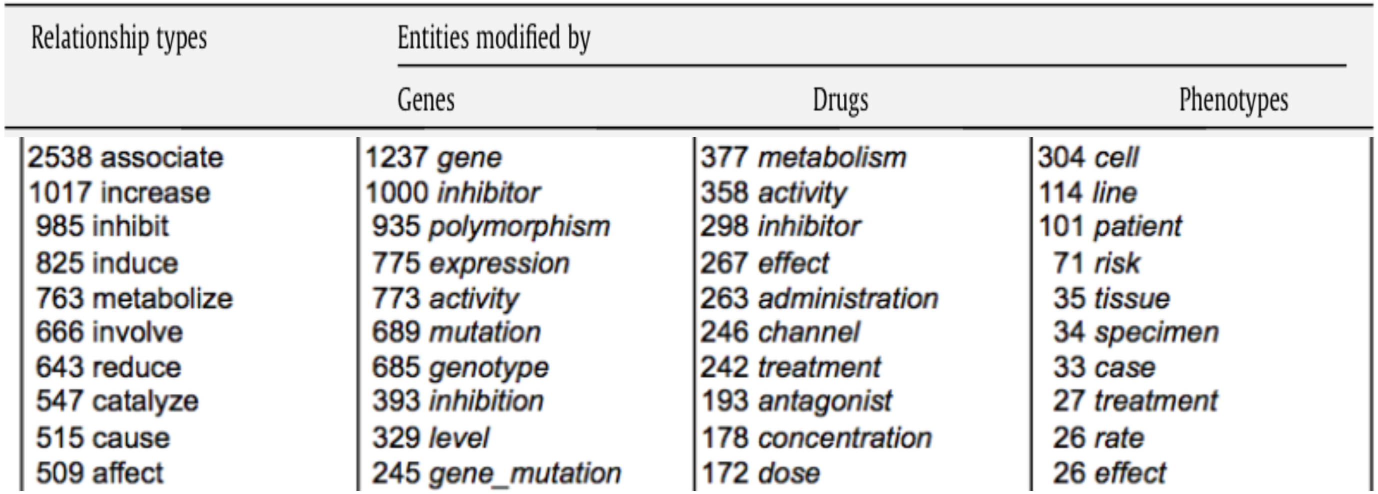
RELATION

entity 2 raw

entity2_ normalized

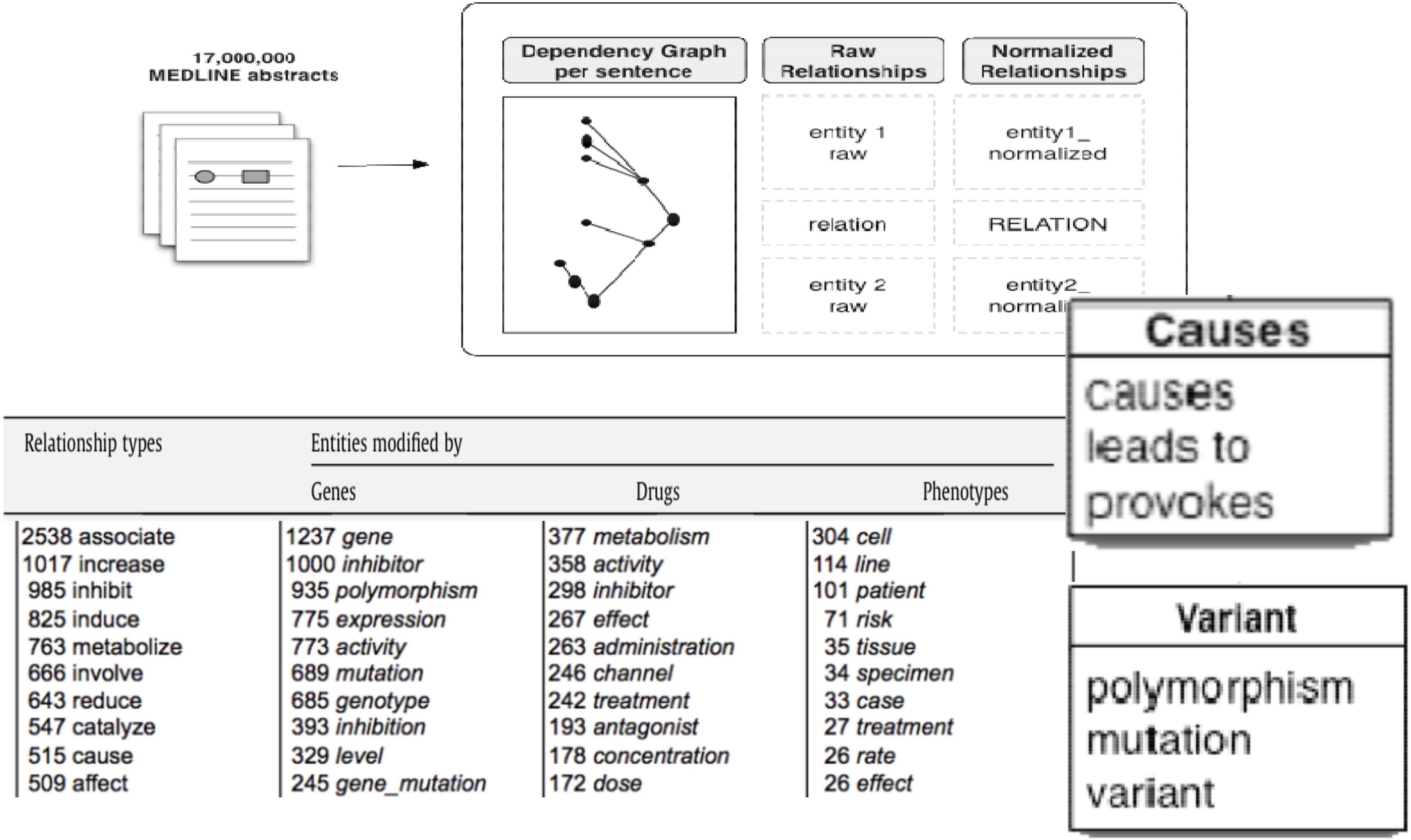
We manually created a PGx ontology "bottom-up"





Relationship types	Entities modified by		
	Genes	Drugs	Phenotypes
2538 associate	1237 *gene*	377 *metabolism*	304 *cell*
1017 increase	1000 *inhibitor*	358 *activity*	114 *line*
985 inhibit	935 *polymorphism*	298 *inhibitor*	101 *patient*
825 induce	775 *expression*	267 *effect*	71 *risk*
763 metabolize	773 *activity*	263 *administration*	35 *tissue*
666 involve	689 *mutation*	246 *channel*	34 *specimen*
643 reduce	685 *genotype*	242 *treatment*	33 *case*
547 catalyze	393 *inhibition*	193 *antagonist*	27 *treatment*
515 cause	329 *level*	178 *concentration*	26 *rate*
509 affect	245 *gene_mutation*	172 *dose*	26 *effect*

We manually created a PGx ontology “bottom-up”



Relationship types	Entities modified by		
	Genes	Drugs	Phenotypes
2538 associate	1237 *gene*	377 *metabolism*	304 *cell*
1017 increase	1000 *inhibitor*	358 *activity*	114 *line*
985 inhibit	935 *polymorphism*	298 *inhibitor*	101 *patient*
825 induce	775 *expression*	267 *effect*	71 *risk*
763 metabolize	773 *activity*	263 *administration*	35 *tissue*
666 involve	689 *mutation*	246 *channel*	34 *specimen*
643 reduce	685 *genotype*	242 *treatment*	33 *case*
547 catalyze	393 *inhibition*	193 *antagonist*	27 *treatment*
515 cause	329 *level*	178 *concentration*	26 *rate*
509 affect	245 *gene_mutation*	172 *dose*	26 *effect*

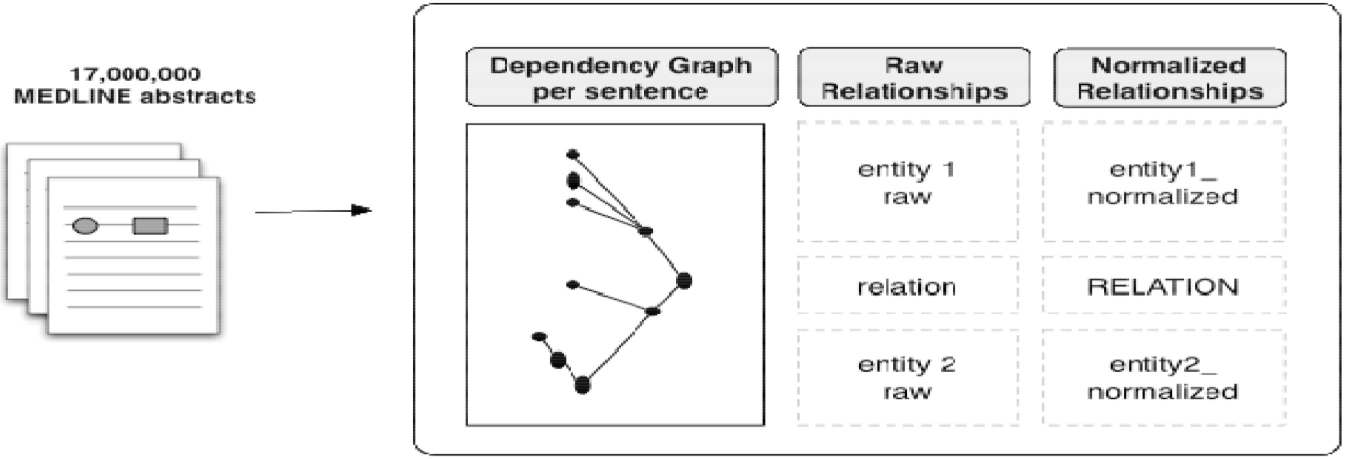
Causes

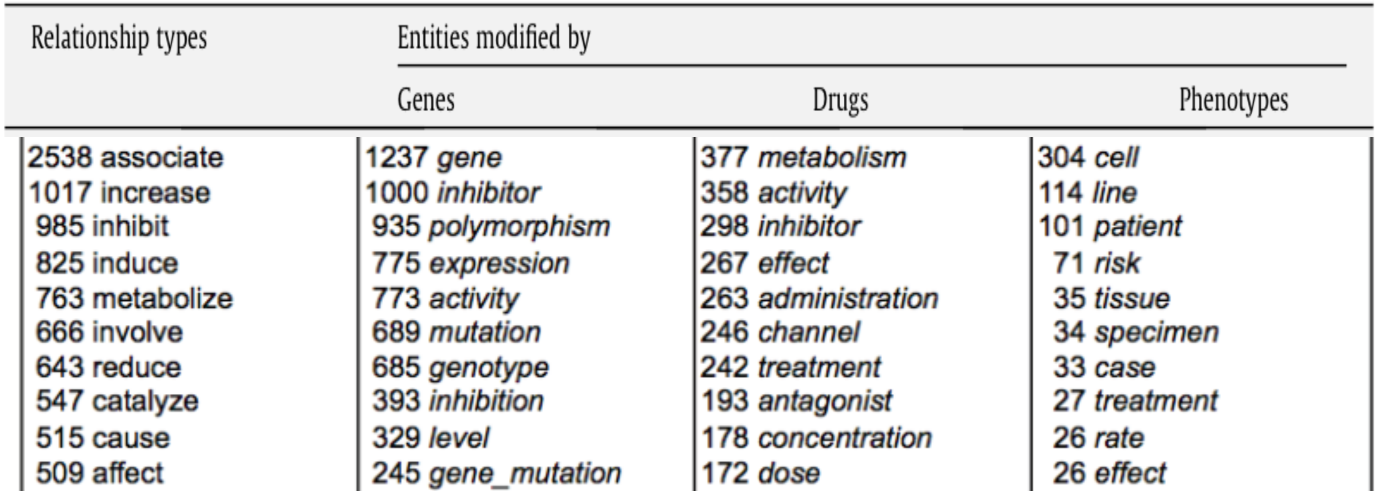
causes
leads to
provokes

Variant

polymorphism
mutation
variant

We manually created a PGx ontology "bottom-up"

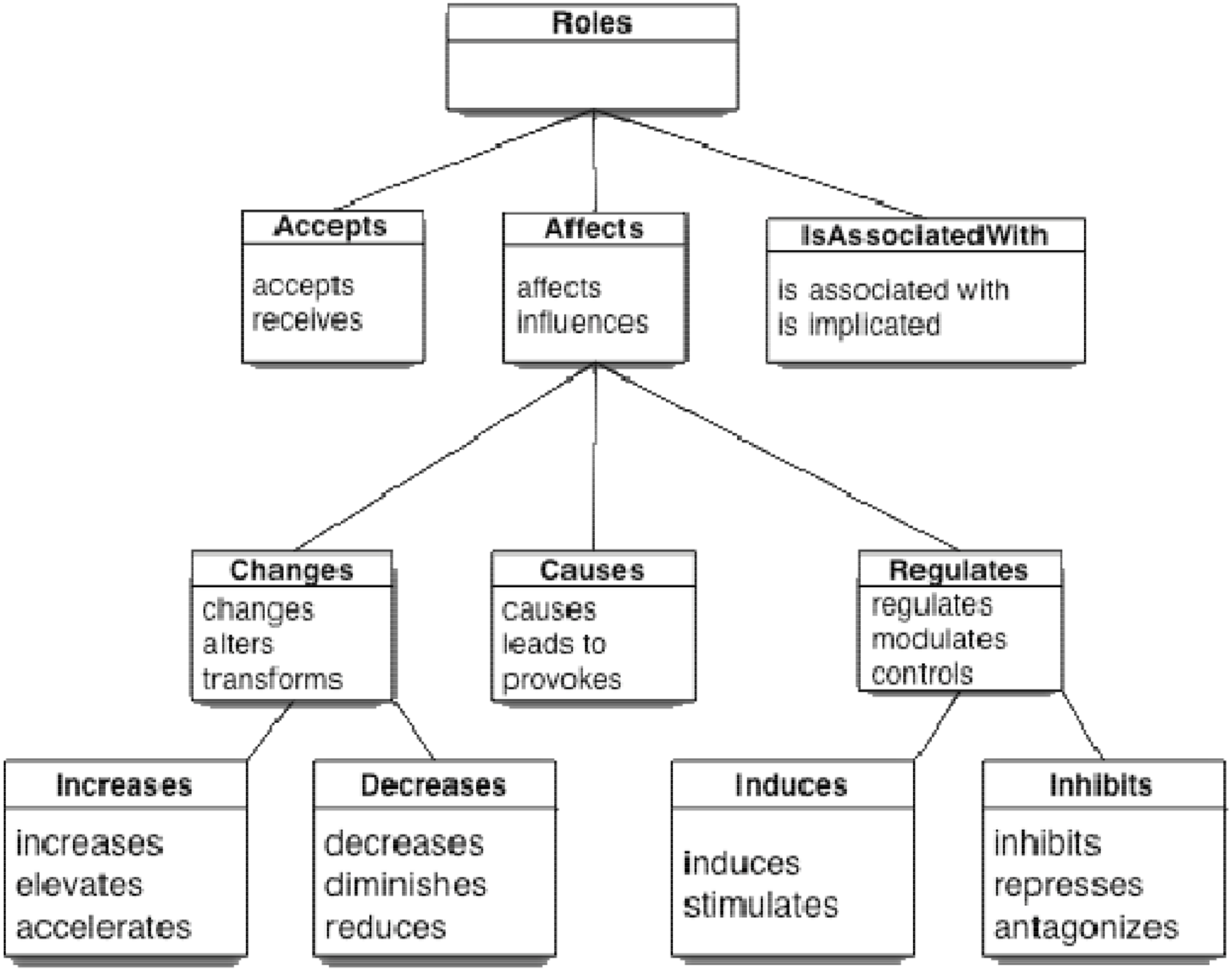




Relationship types	Entities modified by		
	Genes	Drugs	Phenotypes
2538 associate	1237 *gene*	377 *metabolism*	304 *cell*
1017 increase	1000 *inhibitor*	358 *activity*	114 *line*
985 inhibit	935 *polymorphism*	298 *inhibitor*	101 *patient*
825 induce	775 *expression*	267 *effect*	71 *risk*
763 metabolize	773 *activity*	263 *administration*	35 *tissue*
666 involve	689 *mutation*	246 *channel*	34 *specimen*
643 reduce	685 *genotype*	242 *treatment*	33 *case*
547 catalyze	393 *inhibition*	193 *antagonist*	27 *treatment*
515 cause	329 *level*	178 *concentration*	26 *rate*
509 affect	245 *gene_mutation*	172 *dose*	26 *effect*

237 concepts 76 roles

Snapshot of the role hierarchy



Roles

Accepts
accepts
receives

Affects
affects
influences

IsAssociatedWith
is associated with
is implicated

Changes
changes
alters
transforms

Causes
causes
leads to
provokes

Regulates
regulates
modulates
controls

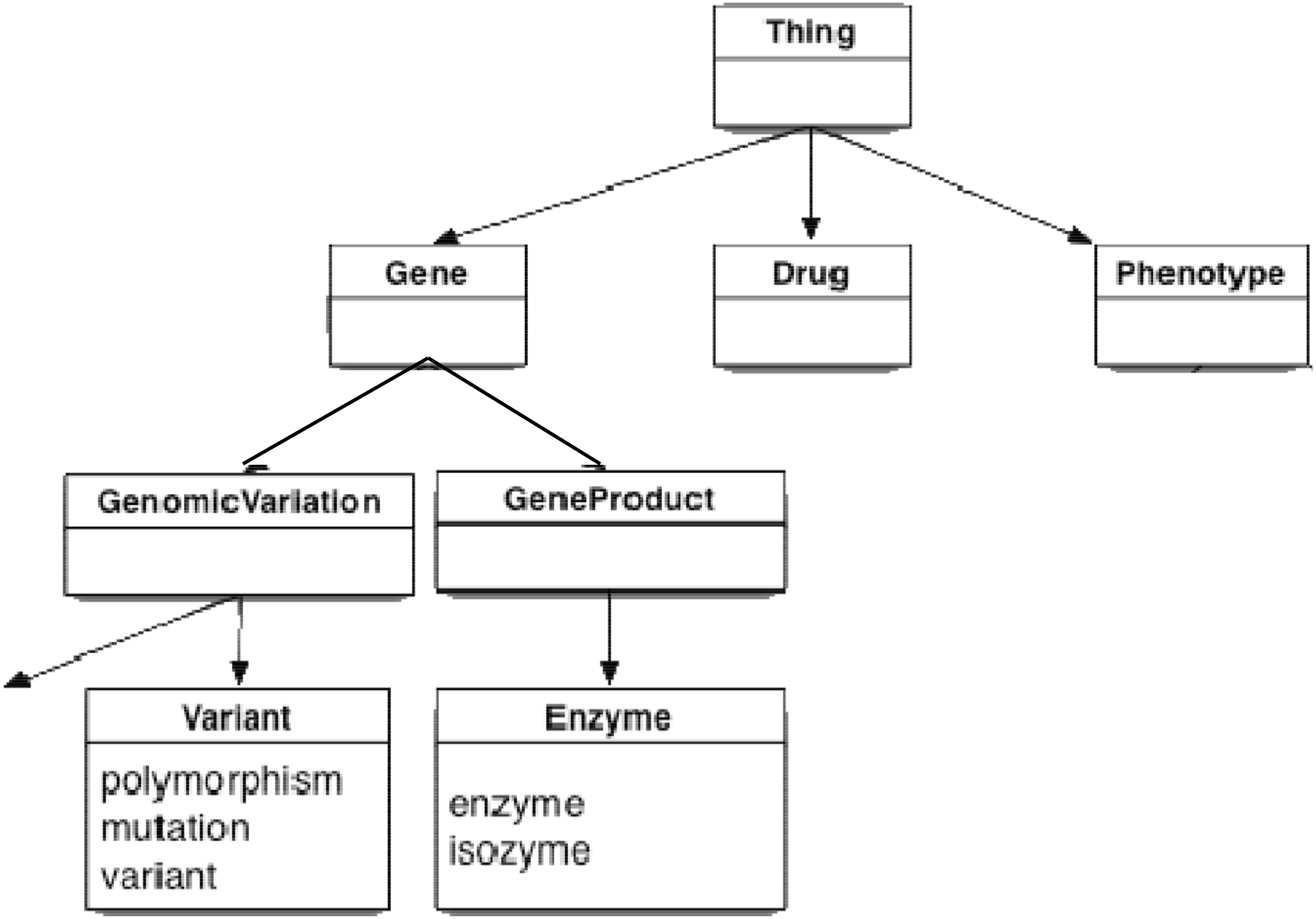
Increases
increases
elevates
accelerates

Decreases
decreases
diminishes
reduces

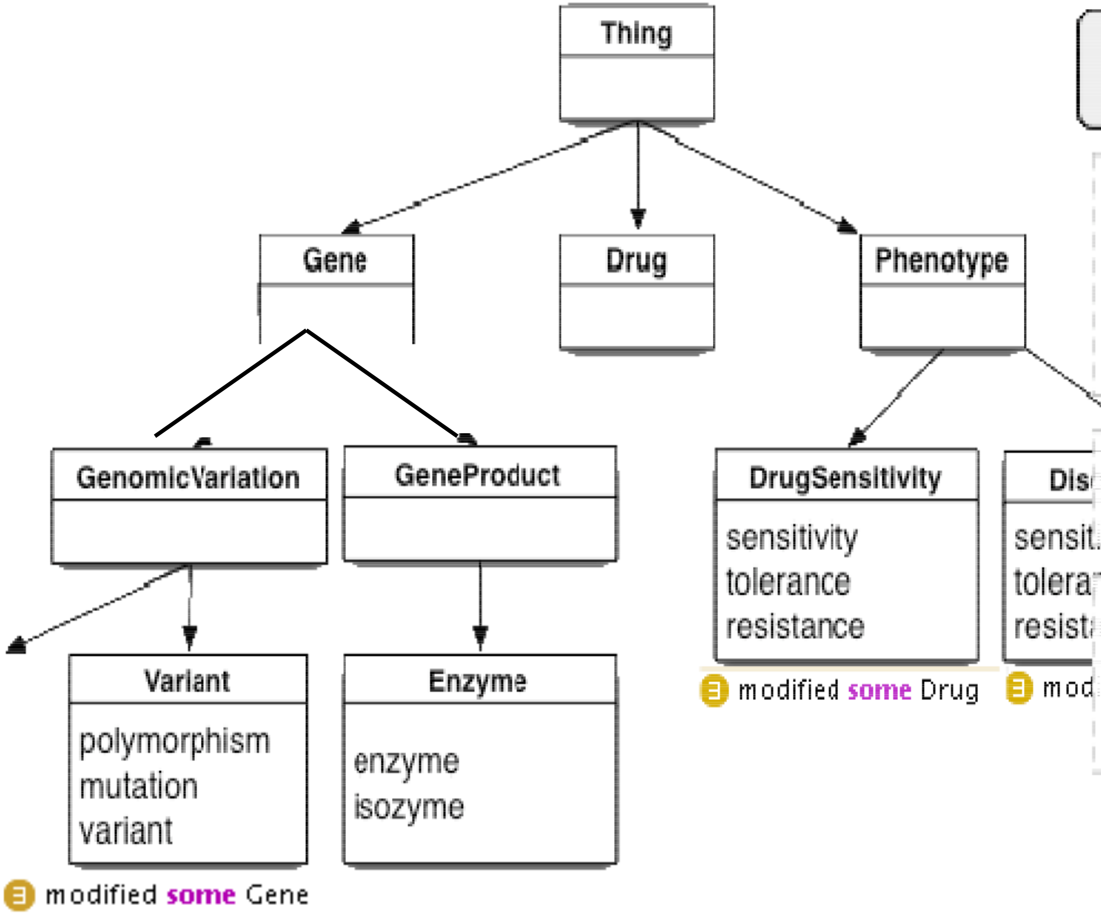
Induces
induces
stimulates

Inhibits
inhibits
represses
antagonizes

Snapshot of the concept hierarchy

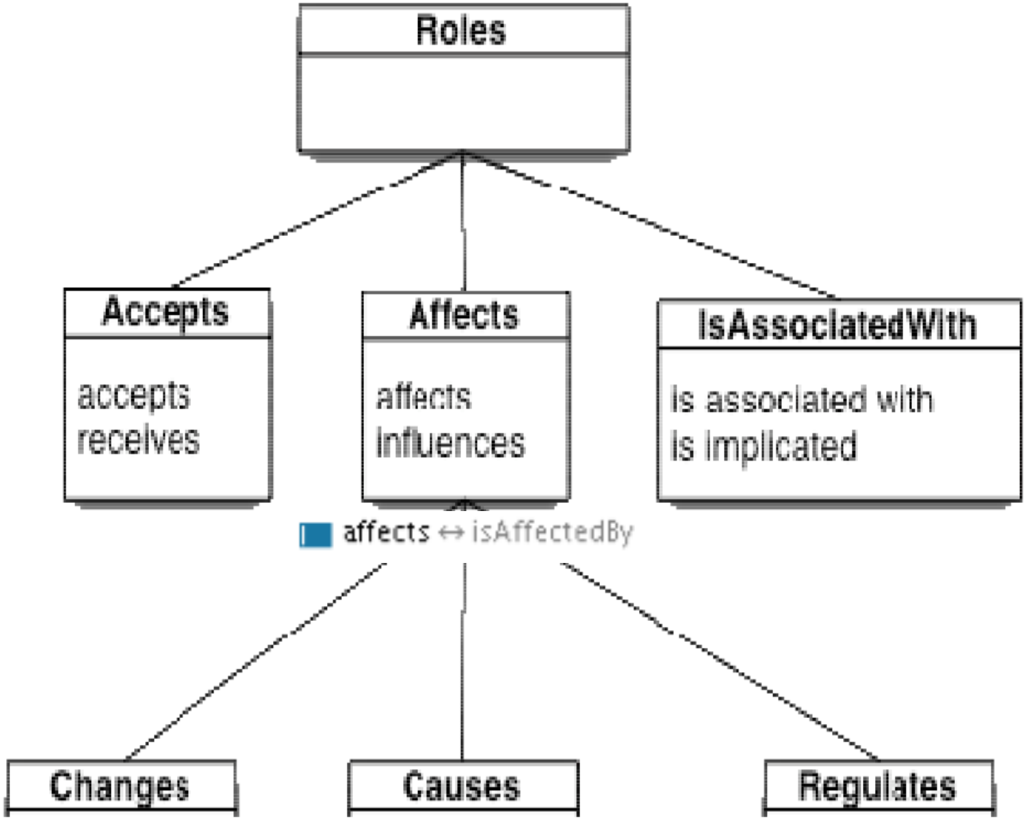


We use the ontology to normalize the raw relationship (subject, relation and object)



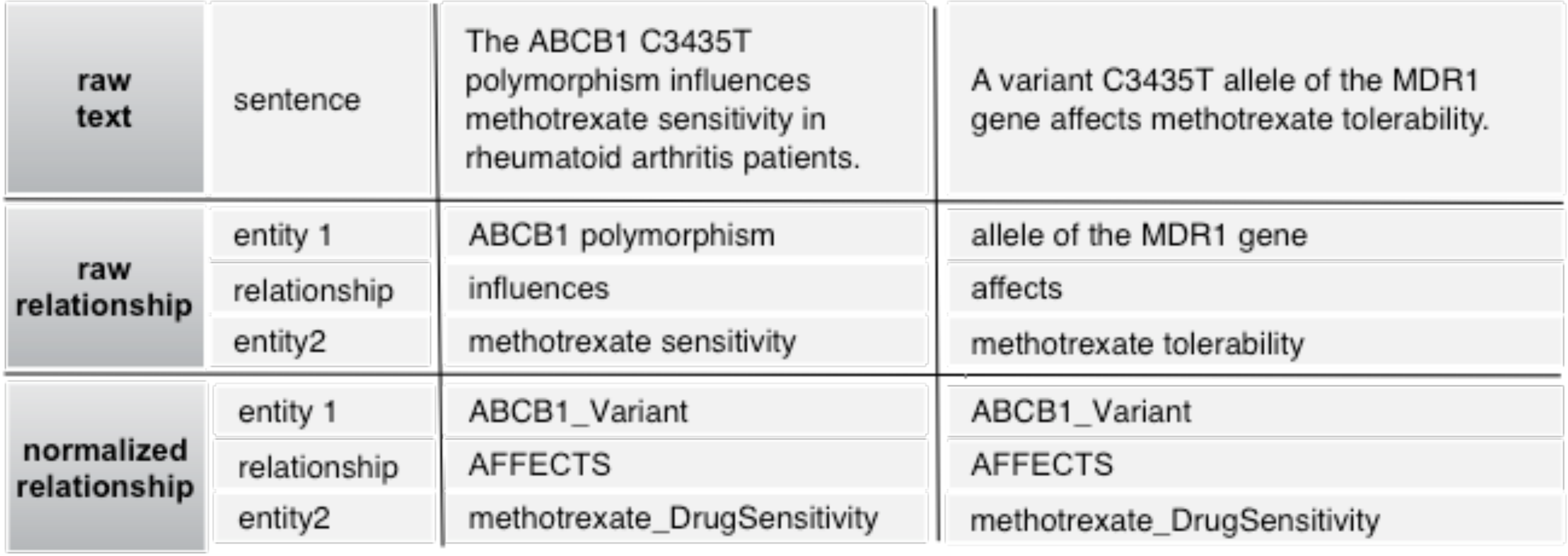
Raw Relation Extraction	Relation Normalization
ABCB1 polymorphism	ABCB1_ Variant
influences	AFFECTS
methotrexate sensitivity	methotrexate_ DrugSensitivity

We use the ontology to normalize the raw relationship (subject, relation and object)



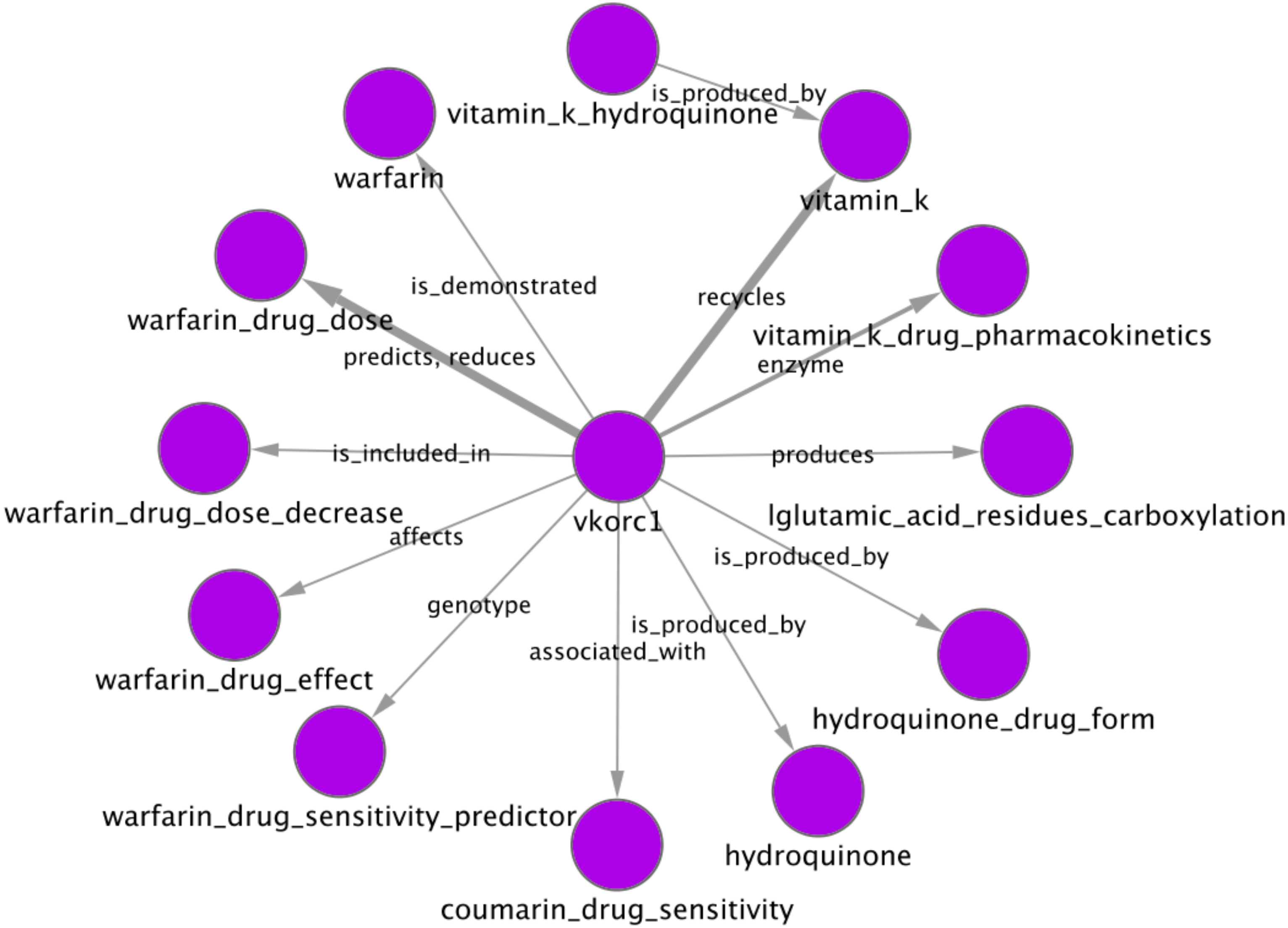
Raw Relation Extraction	Relation Normalization
ABCB1 polymorphism	ABCB1_ Variant
influences	AFFECTS
methotrexate sensitivity	methotrexate_ DrugSensitivity

Example: two sentences but one fact

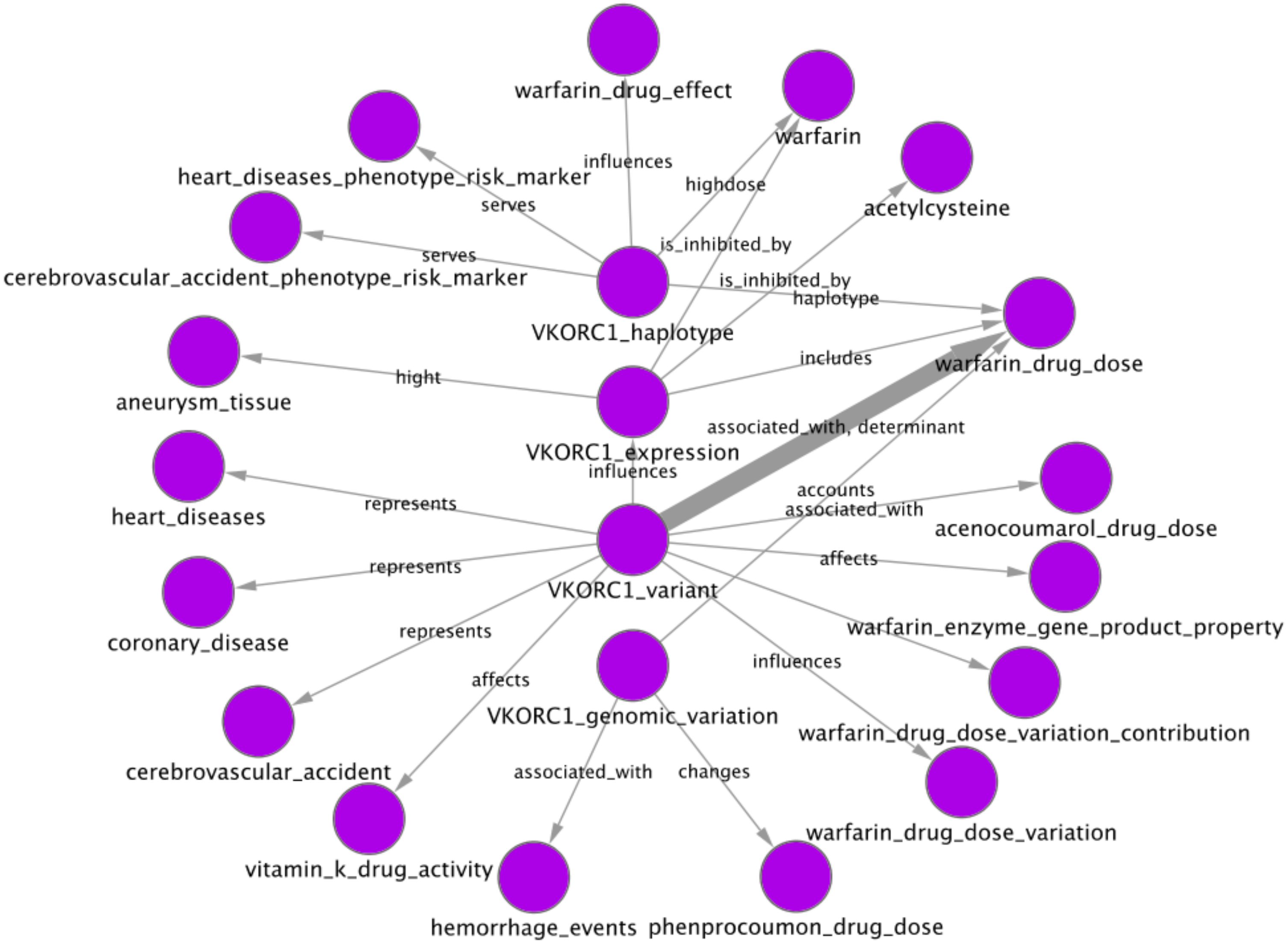


raw text	sentence	The ABCB1 C3435T polymorphism influences methotrexate sensitivity in rheumatoid arthritis patients.	A variant C3435T allele of the MDR1 gene affects methotrexate tolerability.
raw relationship	entity 1	ABCB1 polymorphism	allele of the MDR1 gene
	relationship	influences	affects
	entity2	methotrexate sensitivity	methotrexate tolerability
normalized relationship	entity 1	ABCB1_Variant	ABCB1_Variant
	relationship	AFFECTS	AFFECTS
	entity2	methotrexate_DrugSensitivity	methotrexate_DrugSensitivity

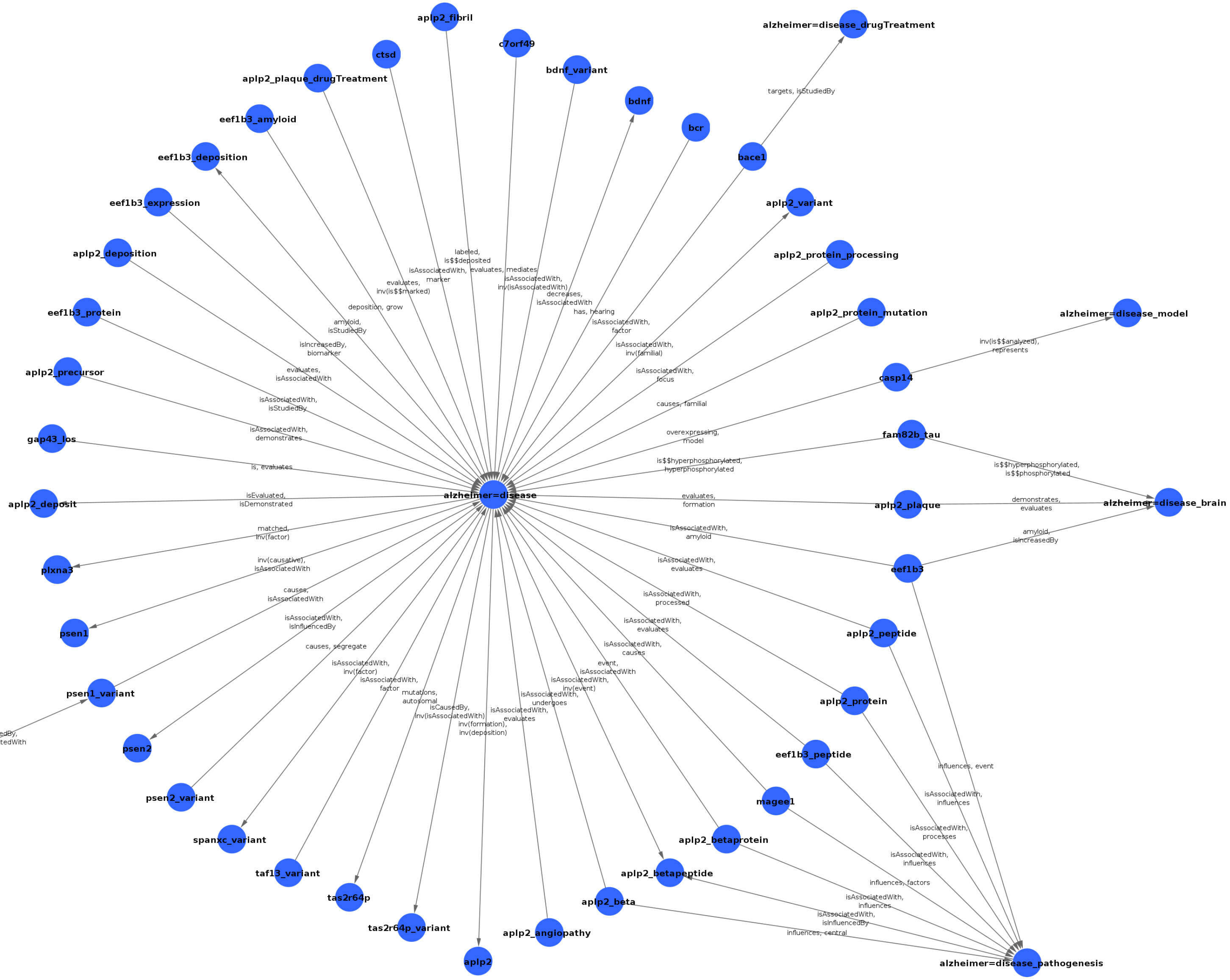
Example of network (1/3): VKORC1



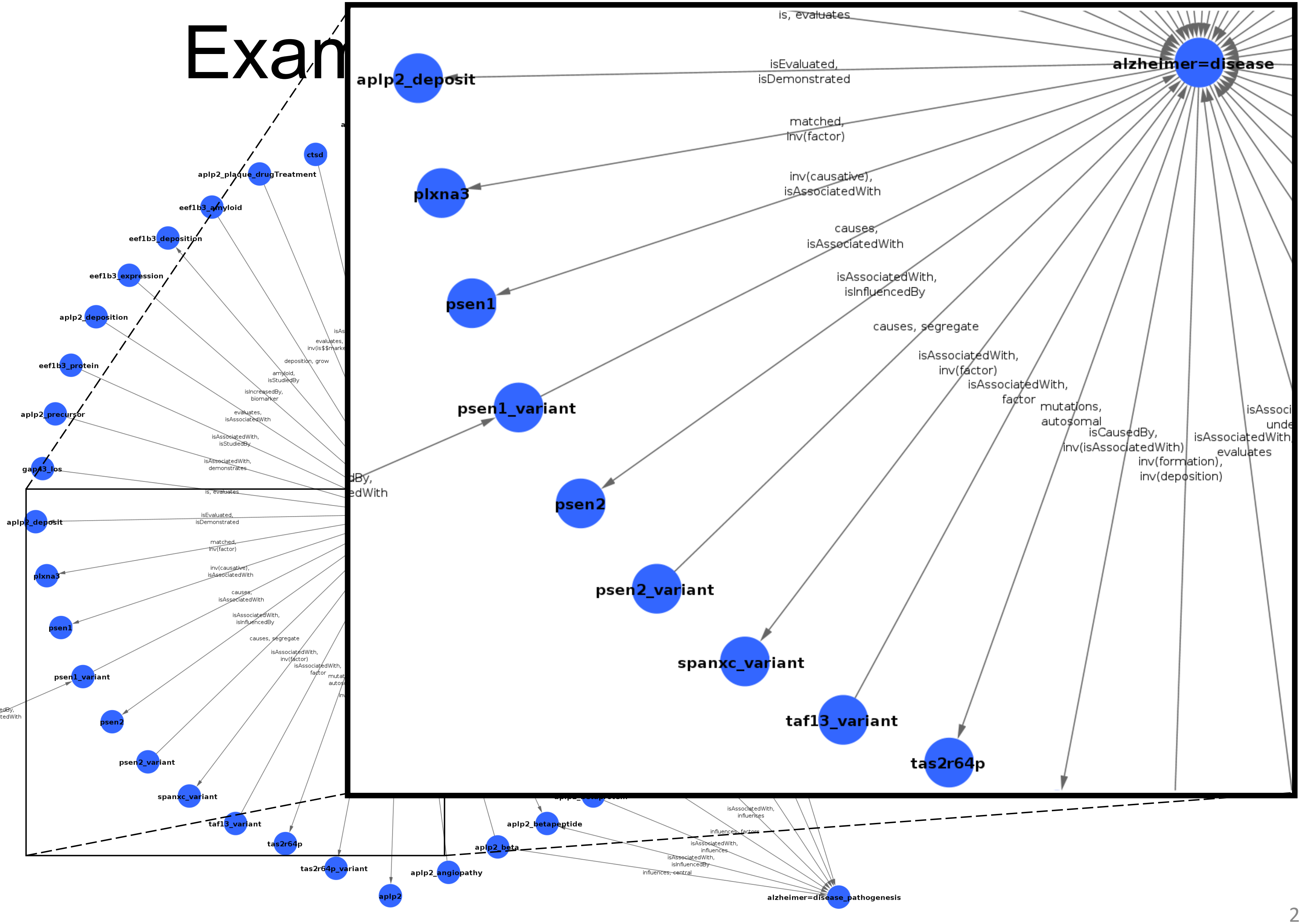
Example of network (2/3): modified by VKORC1



Example of network (3/3): AD



Exam



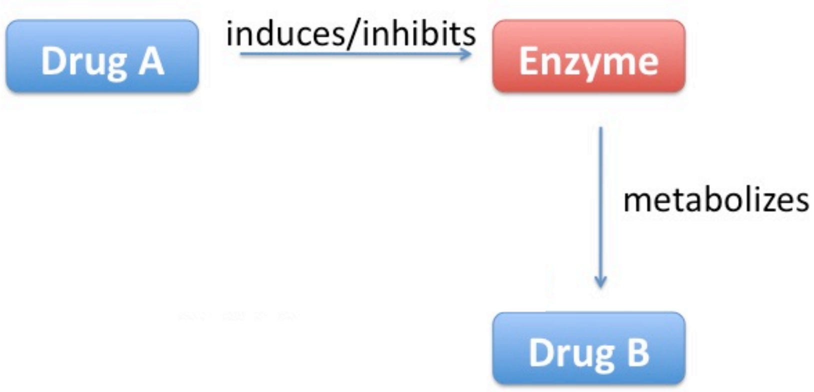
alzheimer=disease
is, evaluates
aplp2_deposit
isEvaluated, isDemonstrated
plxna3
matched, inv(factor)
psen1
inv(causative), isAssociatedWith
causes, isAssociatedWith
psen1_variant
isAssociatedWith, isInfluencedBy
psen2
causes, segregate
isAssociatedWith, inv(factor)
psen2_variant
isAssociatedWith, factor
spanxc_variant
mutations, autosomal
taf13_variant
isCausedBy, inv(isAssociatedWith)
tas2r64p
inv(formation), inv(deposition)
isAssociatedWith, evaluates
ctsd
aplp2_plaque_drugTreatment
eef1b3_amyloid
eef1b3_deposition
eef1b3_expression
aplp2_deposition
eef1b3_protein
aplp2_precursor
gap43_los
deposition, grow
amyloid, isStudiedBy
isIncreasedBy, biomarker
evaluates, isAssociatedWith
isAssociatedWith, isStudiedBy
isAssociatedWith, demonstrates
tas2r64p_variant
aplp2
aplp2_angiopathy
aplp2_betapeptide
aplp2_beta
isAssociatedWith, influences
influences, factors
isAssociatedWith, isInfluencedBy
influences, central
alzheimer=disease_pathogenesis

Resulting Knowledge Base

- Useful
  - For curation and knowledge summarization @PharmGKB
  - For knowledge discovery

e.g.Predicting Drug-Drug interaction

=>*Yael Garten's PhD thesis*



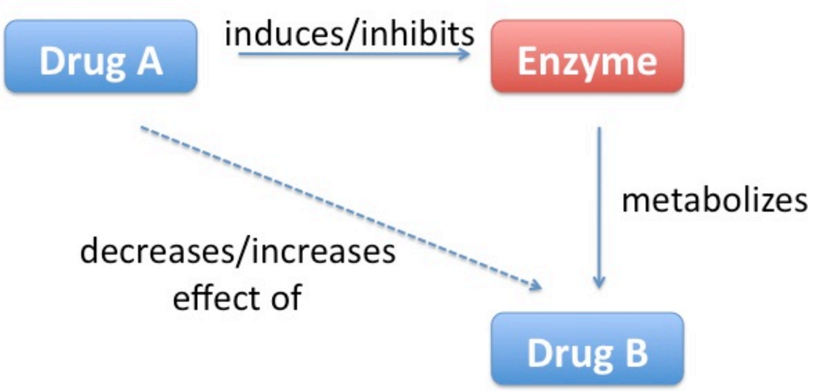
Resulting Knowledge Base

- Useful
  - For curation and knowledge summarization

    @PharmGKB
  - For knowledge discovery

    e.g.Predicting Drug-Drug interaction

    =>*Yael Garten's PhD thesis*



Resulting Knowledge Base

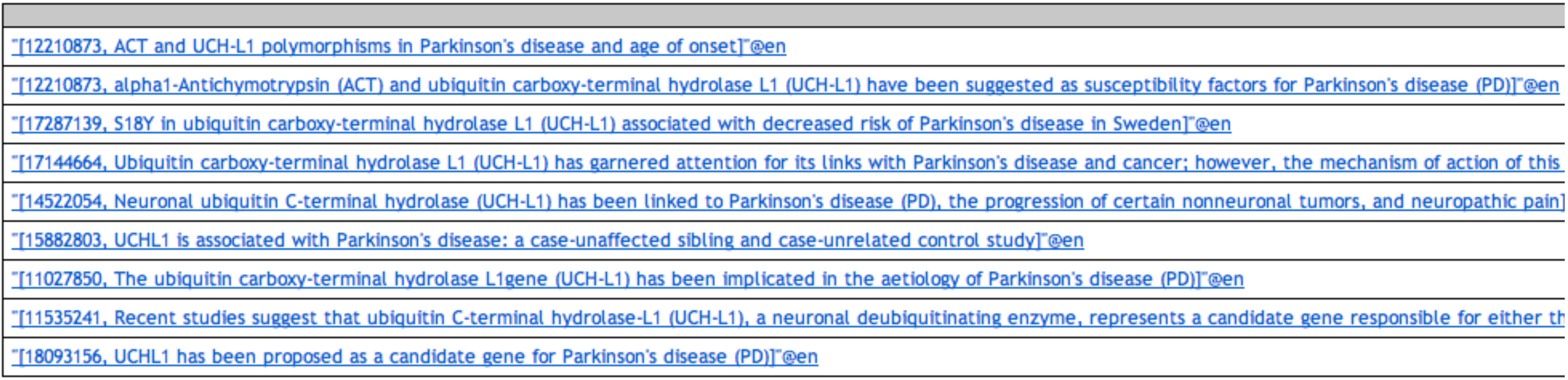
- Online
  - SPARQL endpoint

    http://sparql.bioontology.org/webui/

  - Example of queries

    http://www.loria.fr/~coulet/material/sparql_queries

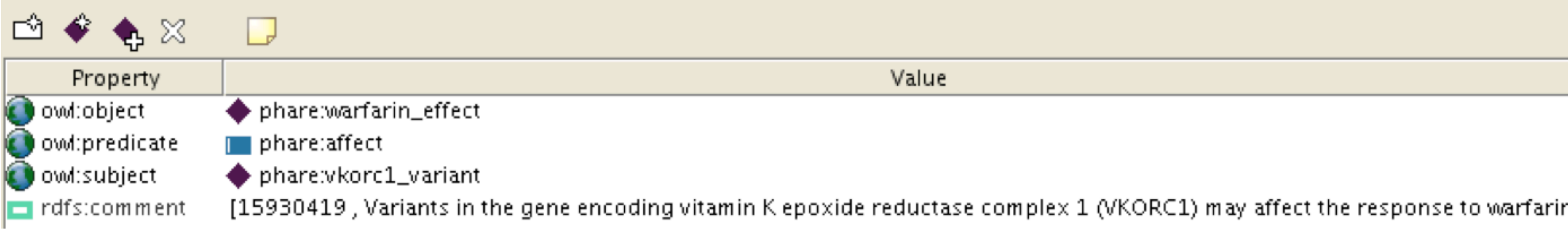
```
###all sentences where the gene UCHL1 is involved in a relation
select $y
from <rmi:phare.owl#pd>
where $rel <http://www.w3.org/2000/01/rdf-schema#comment> $y
and $rel <owl:annotatedSource> <http://www.stanford.edu/~coulet/
phare.owl#uchl1>;
```



"[12210873, ACT and UCH-L1 polymorphisms in Parkinson's disease and age of onset]"@en
"[12210873, alpha1-Antichymotrypsin (ACT) and ubiquitin carboxy-terminal hydrolase L1 (UCH-L1) have been suggested as susceptibility factors for Parkinson's disease (PD)]"@en
"[17287139, S18Y in ubiquitin carboxy-terminal hydrolase L1 (UCH-L1) associated with decreased risk of Parkinson's disease in Sweden]"@en
"[17144664, Ubiquitin carboxy-terminal hydrolase L1 (UCH-L1) has garnered attention for its links with Parkinson's disease and cancer; however, the mechanism of action of this
"[14522054, Neuronal ubiquitin C-terminal hydrolase (UCH-L1) has been linked to Parkinson's disease (PD), the progression of certain nonneuronal tumors, and neuropathic pain]
"[15882803, UCHL1 is associated with Parkinson's disease: a case-unaffected sibling and case-unrelated control study]"@en
"[11027850, The ubiquitin carboxy-terminal hydrolase L1gene (UCH-L1) has been implicated in the aetiology of Parkinson's disease (PD)]"@en
"[11535241, Recent studies suggest that ubiquitin C-terminal hydrolase-L1 (UCH-L1), a neuronal deubiquitinating enzyme, represents a candidate gene responsible for either th
"[18093156, UCHL1 has been proposed as a candidate gene for Parkinson's disease (PD)]"@en

Resulting Knowledge Base

To improve! (1/2)

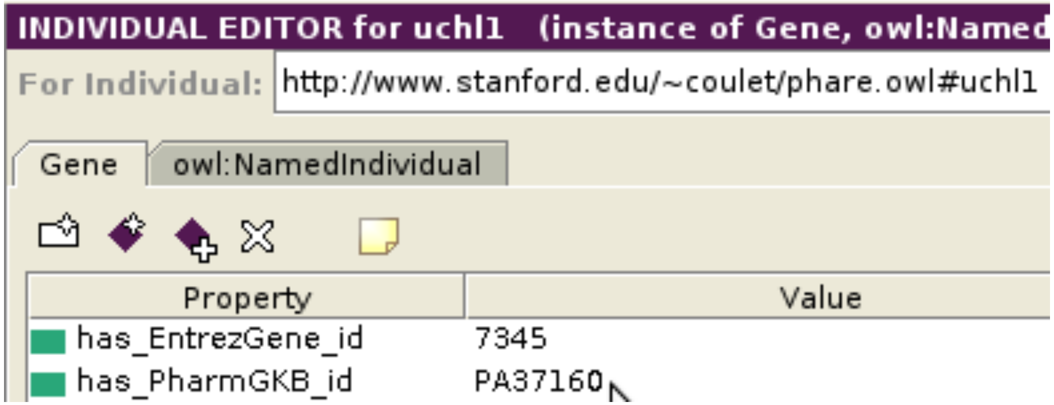


Property	Value
owl:object	phare:warfarin_effect
owl:predicate	phare:affect
owl:subject	phare:vkorc1_variant
rdfs:comment	[15930419 , Variants in the gene encoding vitamin K epoxide reductase complex 1 (VKORC1) may affect the response to warfarin

- Representation of provenance
  - One relation is one triplet
  - Provenance is encoded as an `rdfs:comment`

Resulting Knowledge Base

To improve! (2/2)



INDIVIDUAL EDITOR for uchl1   (instance of Gene, owl:Named

For Individual: http://www.stanford.edu/~coulet/phare.owl#uchl1

Property	Value
🟩 has_EntrezGene_id	7345
🟩 has_PharmGKB_id	PA37160

- Connections with the Linked Data Cloud?
  - IDs from Entrez Gene, DrugBank, MeSH

Questions?

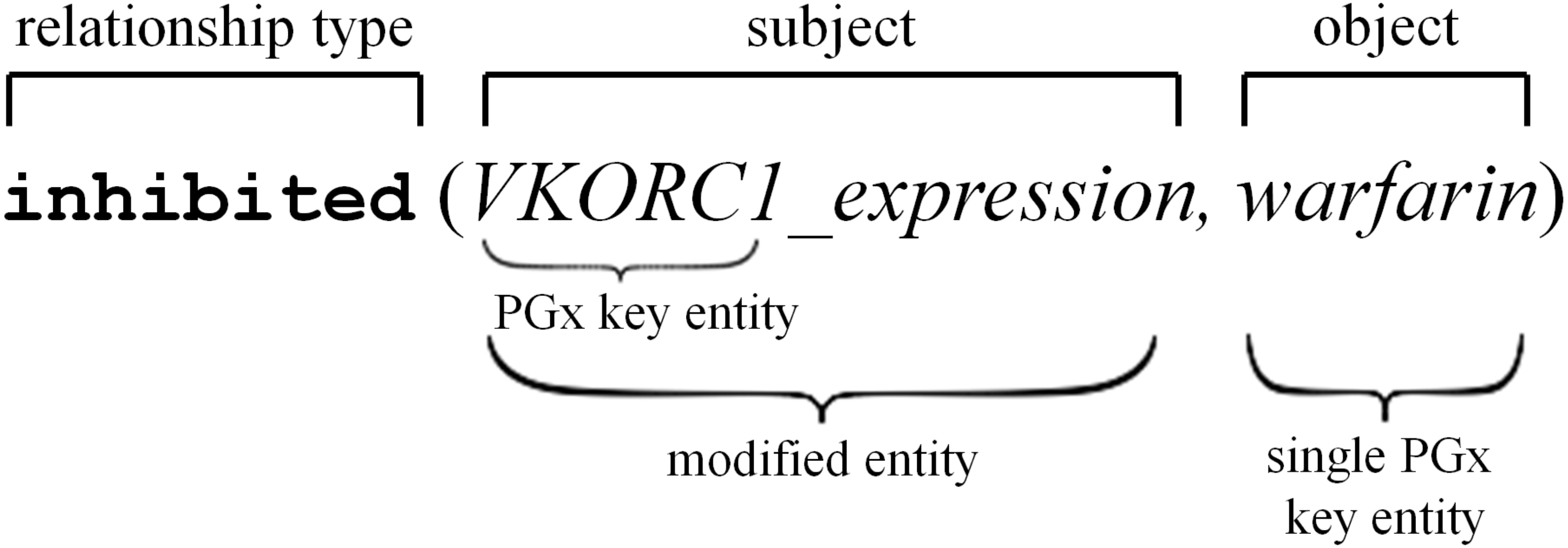
Coulet *et al. Journal of Biomedical Informatics* 43(6), December 2010

or

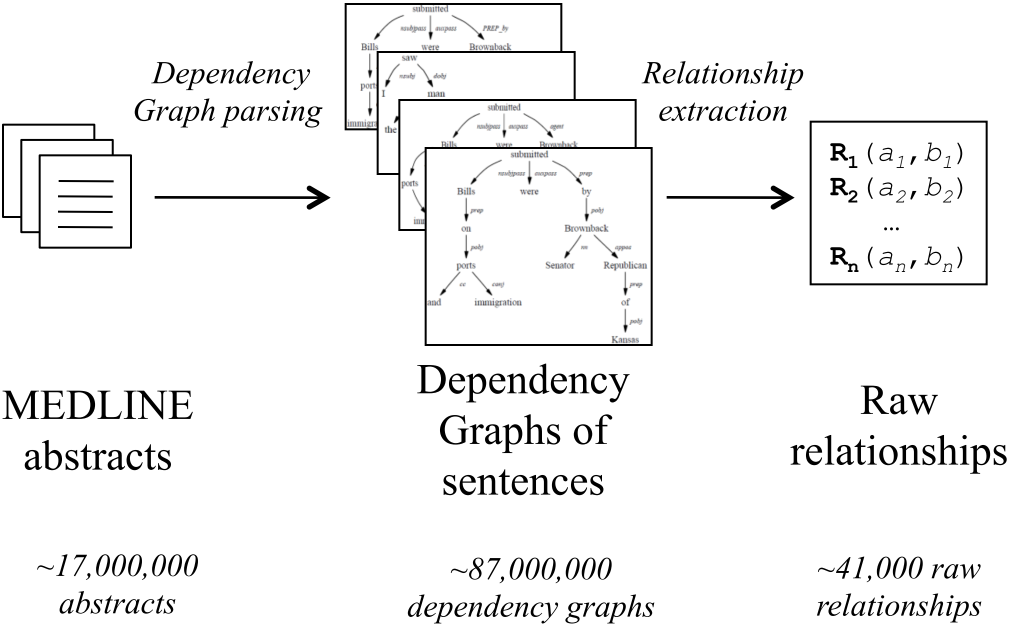
adrien.coulet@loria.fr

Thanks

And thanks to Yael Garten for many slides


relationship type
subject
object
inhibited(VKORC1_expression, warfarin)
PGx key entity
modified entity
single PGx key entity

The method extracts high quality typed relationships



Dependency Graph parsing

Relationship extraction

$R_1(a_1, b_1)$
$R_2(a_2, b_2)$
...
$R_n(a_n, b_n)$

MEDLINE abstracts

Dependency Graphs of sentences

Raw relationships

~17,000,000 abstracts

~87,000,000 dependency graphs

~41,000 raw relationships

Evaluation:
Randomly selected 220 raw relationships: classified into 3

"polymorphisms in VKORC1 are associated with warfarin dose."

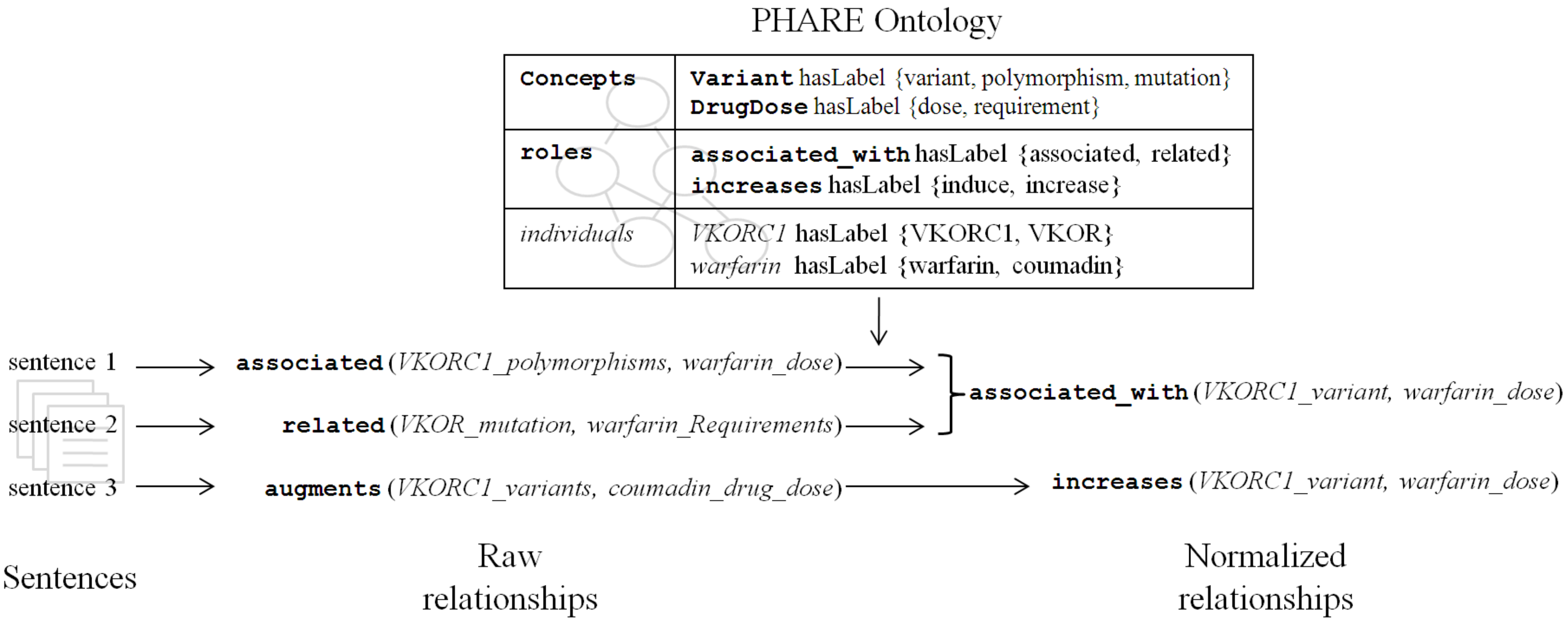
•associated(VKORC1_polymorphisms,warfarin_dose)
= true and complete

• associated (VKORC1_polymorphisms, warfarin)
= true and incomplete

•polymorphisms (VKORC1, warfarin_dose)
= false

Results:

- 87.7% were complete or incomplete true positives
  - 70% true and complete
  - 17.7% true and incomplete
- 12.3% were false positives



PHARE Ontology

Concepts	**Variant** hasLabel {variant, polymorphism, mutation}
	DrugDose hasLabel {dose, requirement}

roles	**associated_with** hasLabel {associated, related}
	increases hasLabel {induce, increase}

individuals	*VKORC1* hasLabel {VKORC1, VKOR}
	warfarin hasLabel {warfarin, coumadin}

sentence 1 → **associated** (*VKORC1_polymorphisms, warfarin_dose*) →

sentence 2 → **related** (*VKOR_mutation, warfarin_Requirements*) →

associated_with (*VKORC1_variant, warfarin_dose*)

sentence 3 → **augments** (*VKORC1_variants, coumadin_drug_dose*) → **increases** (*VKORC1_variant, warfarin_dose*)

Sentences

Raw relationships

Normalized relationships