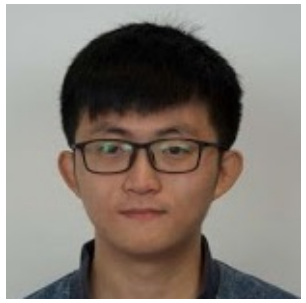
A Distant Supervision Corpus for Extracting Biomedical Relationships Between Chemicals, Diseases and Genes

Dongxu Zhang*, UMass,

Sunil Mohan*, CZI Science,

Michaela Torkar, CZI Science,

Andrew McCallum UMass



Chan Zuckerberg Initiative

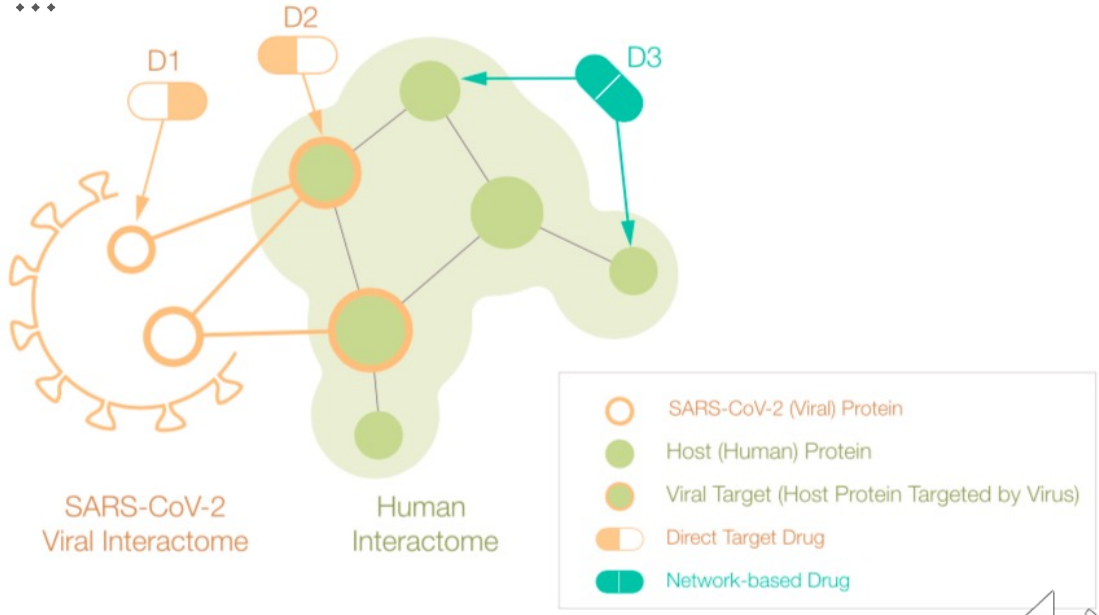
UMassAmherst | College of Information & Computer Sciences

Biomedical Relation Extraction: Motivation

- Assist curators of Biomedical KBs:
  - CTD, CMAP, ProteinDB, KEGG, ...

Over 1M new papers p.a. !

- Build Human Interactome Knowledge Graphs:
  Drugs, Genes/Proteins, Diseases, Symptoms, ...
  - Understanding disease mechanisms
    (*Lee et al, 2019*)
  - Predict poly-pharmacy side-effects
    (*Zitnick et al, 2018*)
  - Drug re-purposing
    (*Morselli Gysi et al, 2020*)



Previous Biomedical RE Corpora

- Several labeled corpora, esp. since ~2006
- Limitations
  - Fewer documents
  - Fewer entity types, one or small number of relation types
  - Entity linking sometimes not included

- Curating Direct labeling of Entities and Relationships in text
  - Very useful for ML/NLP, but …
  - Very time consuming
  - Curated corpora tend to be small

Introducing *ChemDisGene*

- New labeled dataset for Biomedical Relation Extraction
- Distant (document level) labels for
  - 18 relation types
  - Between 3 entity types (Chemicals, Diseases, Genes / Protein Gene-products)
- Two corpora
  - Large, automatically derived: ~80k abstracts
  - Manually curated: 523 abstracts

Outline

- Task definition
- ChemDisGene: Automated Derivation
- ChemDisGene: Manual Curation Process
- Dataset Statistics
- Benchmark Models

Task Definition

- Input
  - PubMed® Abstracts
  - Entity mentions linked: Chemicals, Diseases and Genes / Proteins.
- Output

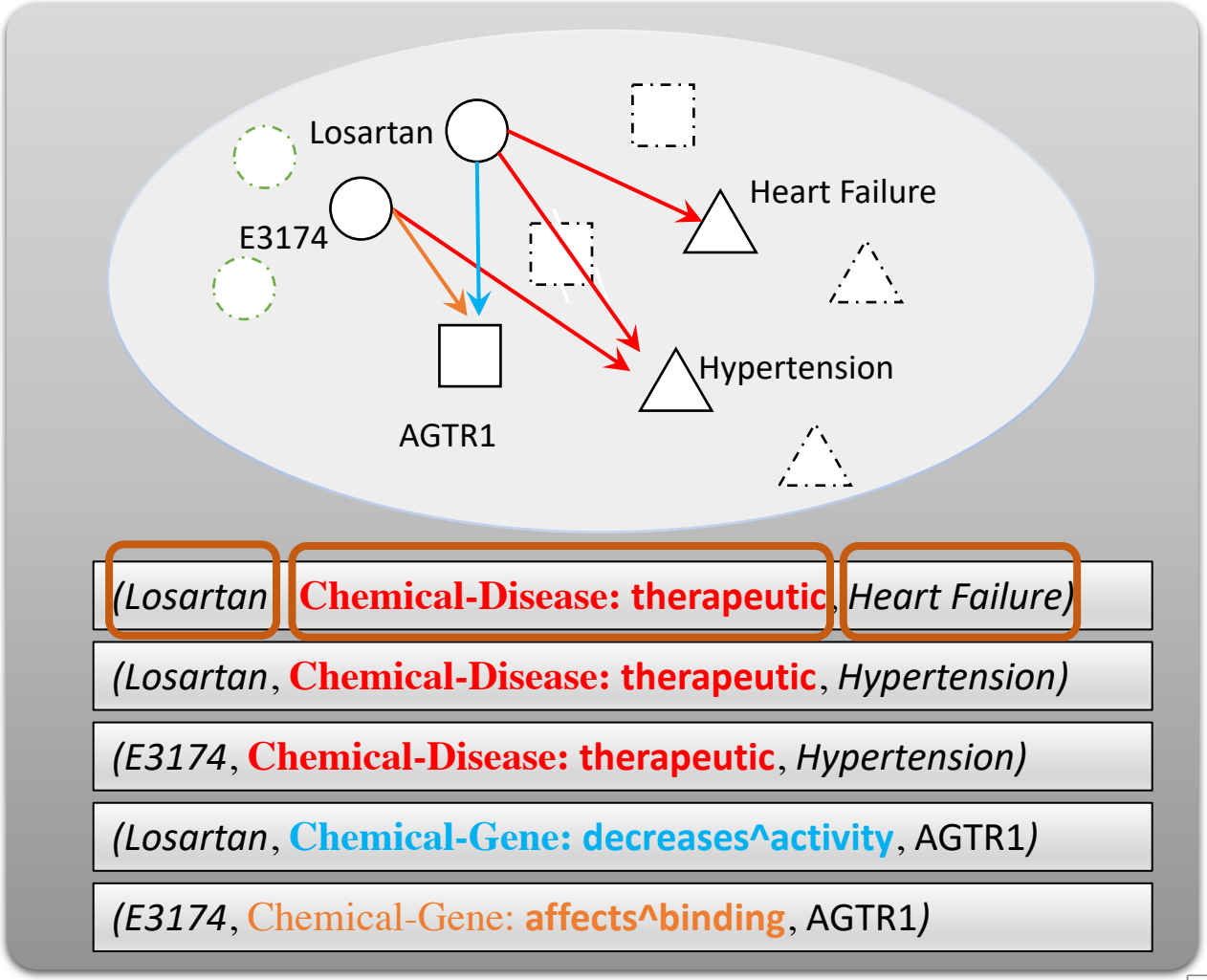
Losartan**: a selective **angiotensin II type 1 (AT1) receptor** antagonist for the treatment of **heart failure

Losartan (COZAAR) is the prototype of a new class of potent and selective angiotensin II (AII) type 1 (AT(1)) receptor antagonists with the largest published preclinical and clinical data base. .... Losartan (parent compound), has moderate affinity for the AT(1) receptor (competitive inhibition). Losartan is well-absorbed orally as an active drug and is rapidly converted via oxidation in the human liver to a more potent metabolite (designated E3174) with an affinity 20- to 30-times greater for the AT(1) receptor (non-competitive inhibition). E3174 has a half-life of 6 - 9 h; elimination is via renal and hepatic routes. Antihyper-tensive and, in heart failure patients, haemodynamic activity is observed over a 24 h period with once daily dosing. Over 6 million patients have been treated for hypertension with continued excellent tolerability. Clinical experience in heart failure is growing, and ...

Task Definition (contd.)

- Input
  - PubMed® Abstracts
  - Entity mentions detected and linked.

- Output: Relationships
  - (Entity, Relation-type, Entity) triplets
  - Distant labels: Annotated at the Document Level



Outline

- Task definition
- ChemDisGene: Automated Derivation
- ChemDisGene: Manual Curation Process
- Dataset Statistics
- Benchmark Models

Comparative Toxicogenomics Database (CTD)

http://ctdbase.org

- Expert curated KB of Relationships, between Chemicals, Diseases, Genes/Proteins, …
- Primary contributions, extracted from research papers (*full text*)
- Distant Labels: Relationships associated with PubMed ID
- Relationship types:
  - Binary: "lead acetate *Chemical* *results in decreased expression* of 14-3-3ZETA mRNA *Gene*"
  - Complex: Quercetin$_{\text{Disease}}$ *inhibits the reaction*
    [[24-hydroxycholesterol$_{\text{Chemical}}$ *co-treated with*
    27-hydroxycholesterol$_{\text{Chemical}}$ *co-treated with*
    cholest-5-en-3 beta,7 alpha-diol$_{\text{Chemical}}$]
    results in increased expression of ITGB1$_{\text{Gene}}$ mRNA]

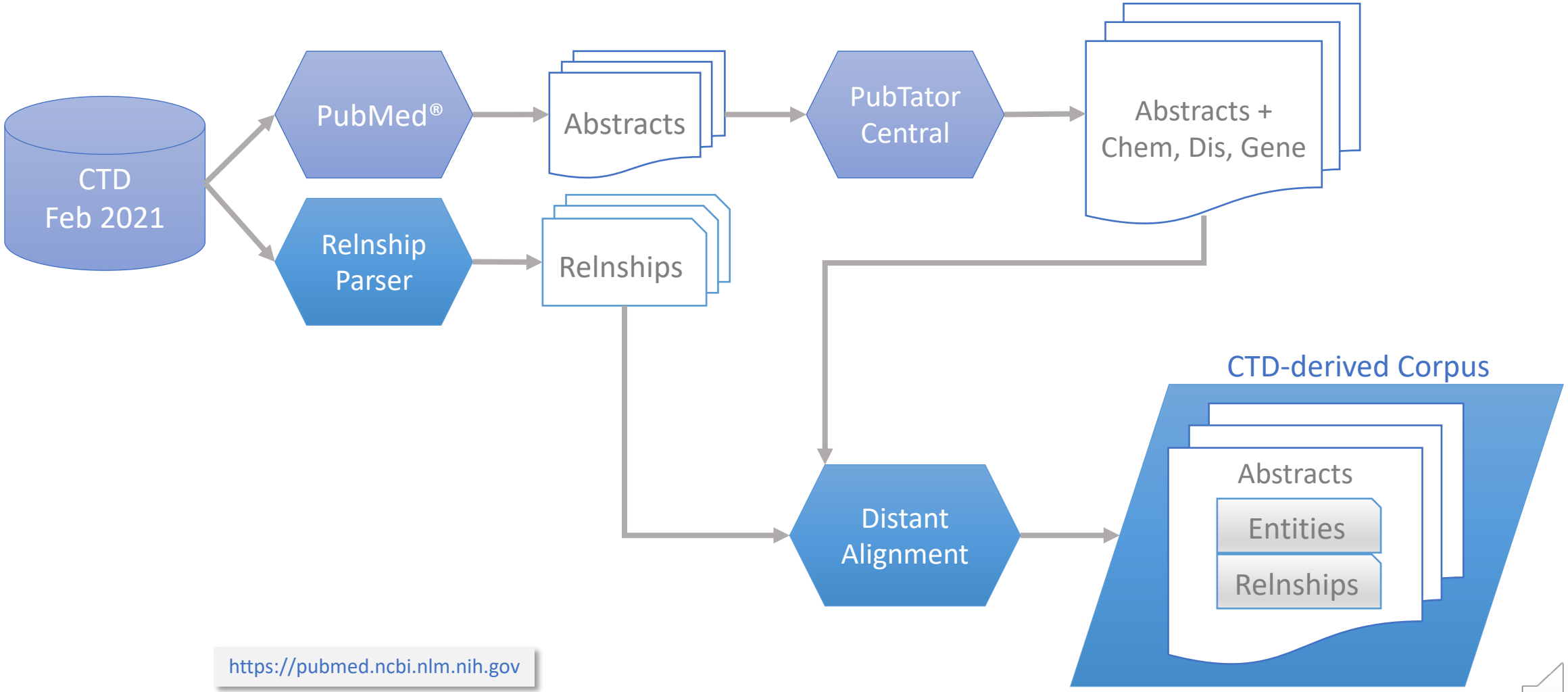
CTD Derivation Target: ChemDisGene

- Only Binary Relationships,
  associated with Abstracts (not full-text)

- Reduced Relation Types:
  - Chemical – Disease: 2 types
  - Chemical – Gene: 6 types * (up to 3 degrees: *affects*, {*increases*, *decreases*})
  - Gene – Disease: 2 types

- Entities in Relationships:
  Detect and link all mentions

Derivation from CTD



https://pubmed.ncbi.nlm.nih.gov

Relationship Parser

- Abstract fine-grained CTD relationships to reduced types
- Extract Binary relationships from complex and nested interactions in CTD

Complex nested interaction in CTD:

Quercetin$_{\text{Disease}}$ *inhibits the reaction*
[[24-hydroxycholesterol$_{\text{Chemical}}$ *co-treated with*
27-hydroxycholesterol$_{\text{Chemical}}$ *co-treated with*
cholest-5-en-3 beta,7 alpha-diol$_{\text{Chemical}}$]
results in increased expression of ITGB1$_{\text{Gene}}$ mRNA]

Extracted binary relationships:

expression-increases(24-hydroxycholesterol, ITGB1)
expression-increases(27-hydroxycholesterol, ITGB1)
expression-increases(cholest-5-en-3 beta,7 alpha-diol, ITGB1)

Entity Linking and Distant Alignment

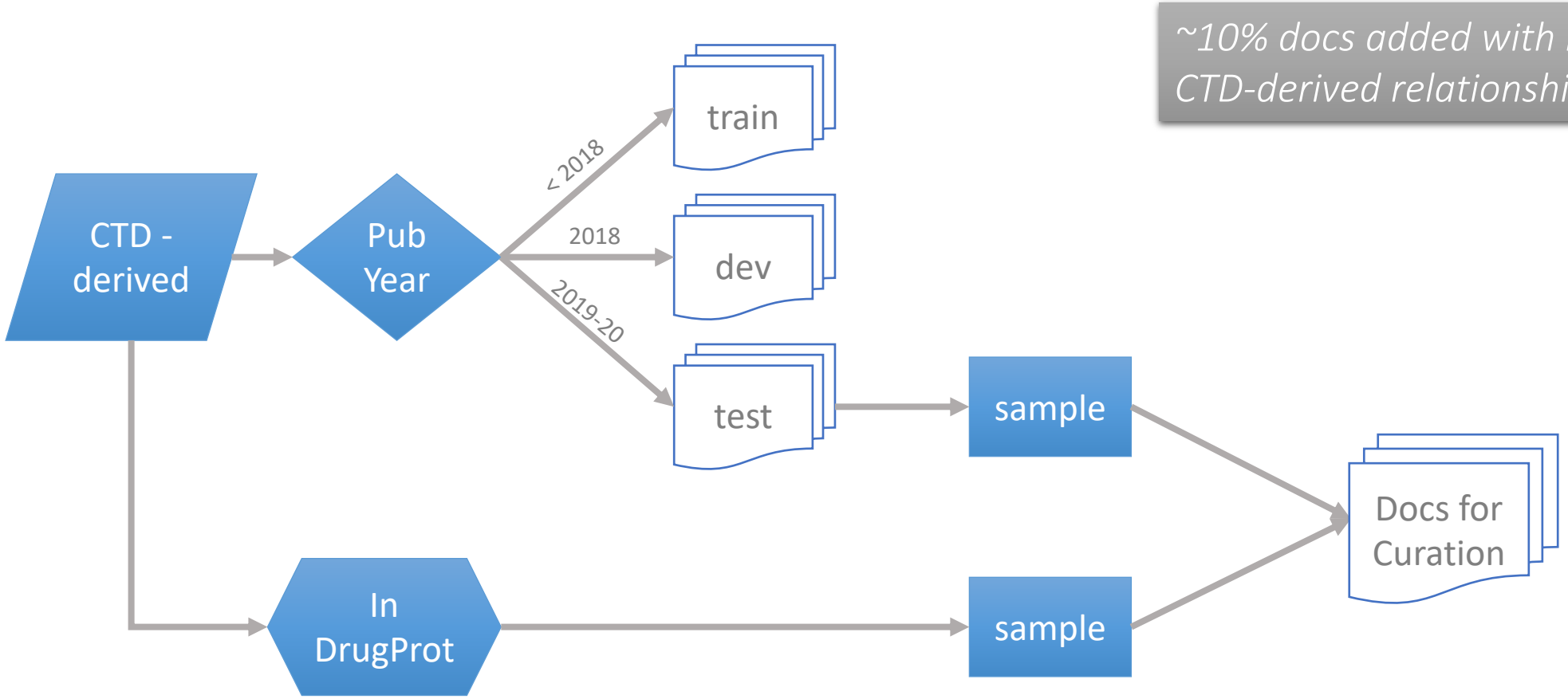
- PubTator Central used to link mentions of
  - Chemicals (MeSH),
  - Diseases (MeSH and OMIM), and
  - Genes/Proteins (NCBI Gene)

https://www.ncbi.nlm.nih.gov/research/pubtator/

- Alignment to extracted Relationships:
  - Reject relationships whose entities not detected in abstract

- *Noisy association: the abstract text may not actually express the relationship.*
  - 78% confidence based on curated sample

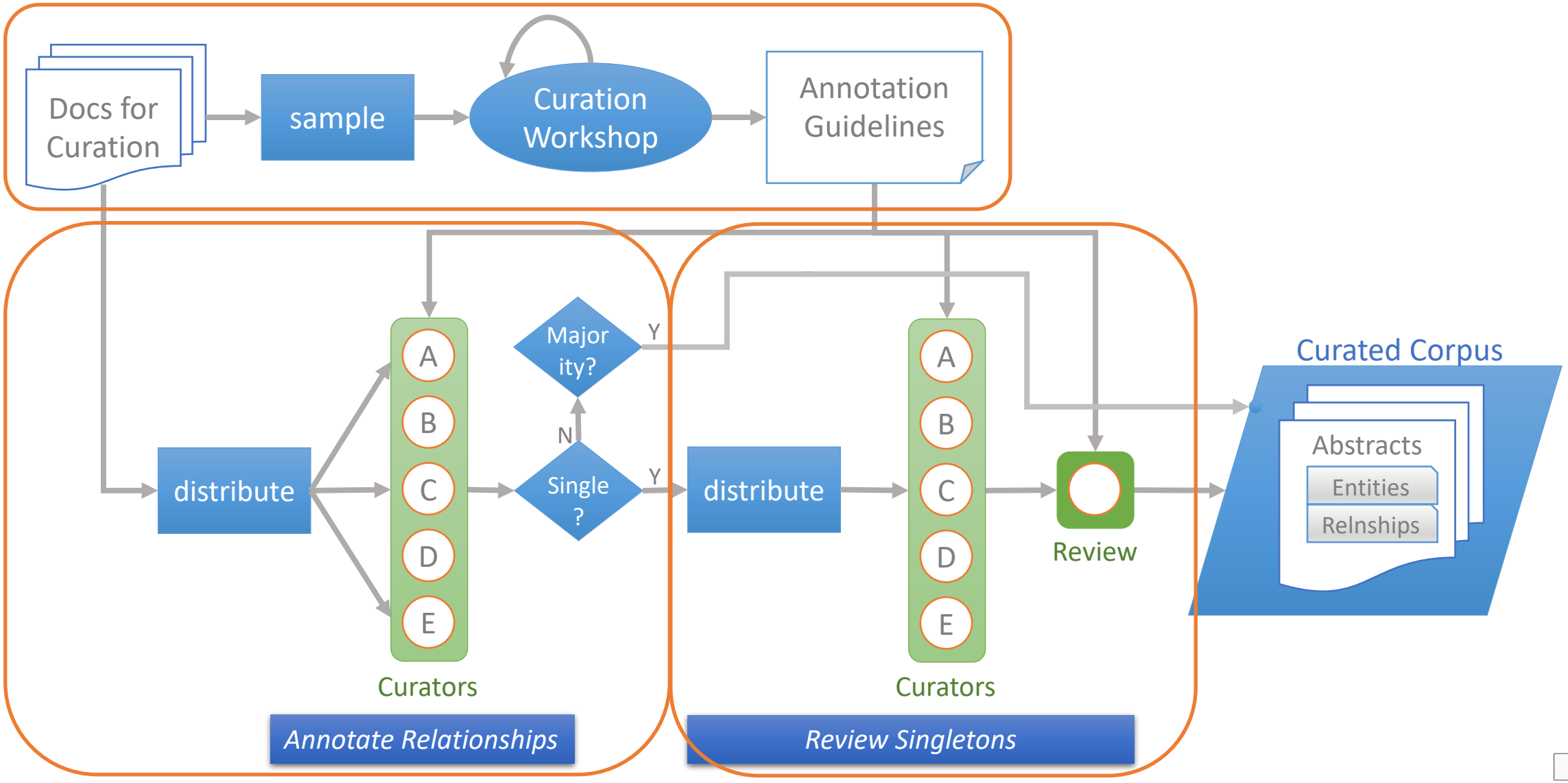
Document Selection



Outline

- Task definition
- ChemDisGene: Automated Derivation
- ChemDisGene: Manual Curation Process
- Dataset Statistics
- Benchmark Models

Curation Process



Annotation Guideline examples

A: *Don't record investigated or motivating relationships that remain unknown and hypothetical.*

"Gene A is a therapeutic target for treatment of Disease X; it may therefore have a potential role in treatment of Disease Z."

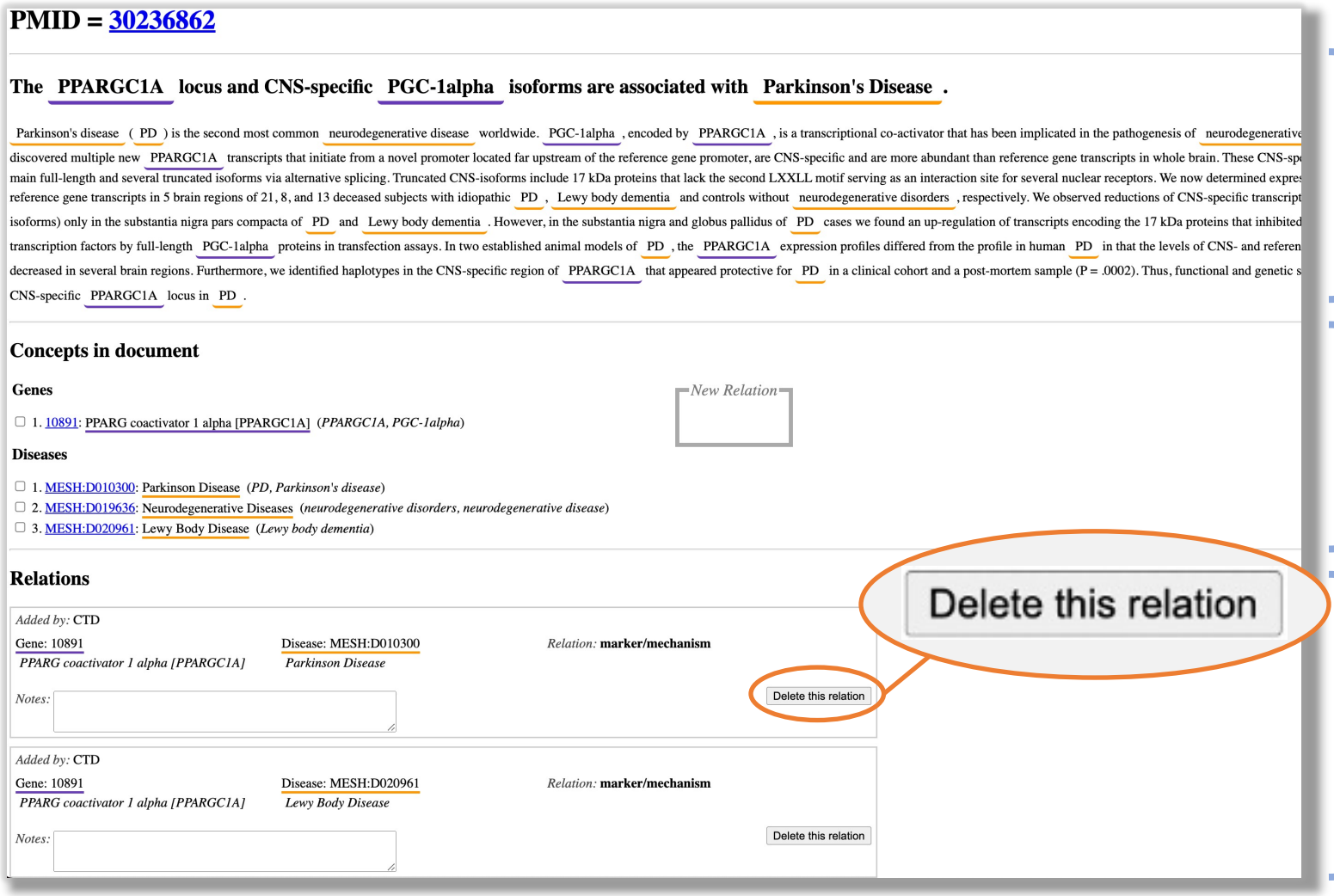
Record a relationship between Gene A and Disease X; but not between Gene A and Disease Z.

B: *Inferring a relationship across sentences.*

"We have previously identified a panel of fusion genes in aggressive prostate cancers. In this study, we showed that . . . CCNH-C5orf30 and TRMT11-GRIK2 gene fusions were found in breast cancer, colon cancer, . . . "

Record a 'Gene-Disease: marker/mechanism' *relationship between C5orf30 and prostate cancers.*

Annotator UI: Approve CTD-derived Reln

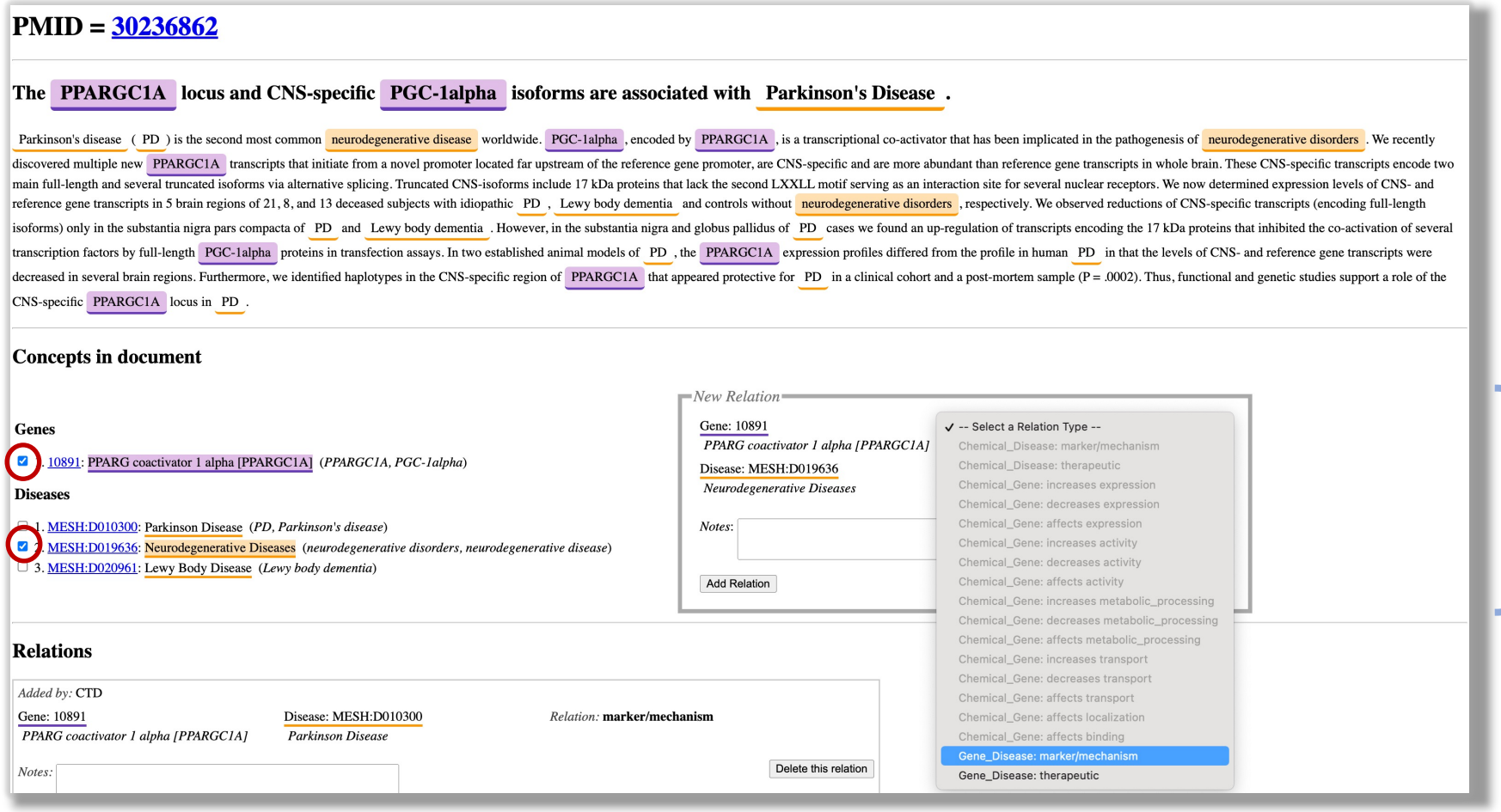


Abstract, with mentions underlined, Types color-coded

Entities linked in the abstract

Relationships derived from CTD

Annotator UI: Add missing ('New') Relationships



Selecting Entities will highlight mentions

Curator can select Relation Type

Annotator UI: Singleton Review

PMID = 15078100

Kinetic mechanism of quinone oxidoreductase 2 and its inhibition by the antimalarial quinolines .

Quinone oxidoreductase 2 (QR2) purified from human red blood cells was recently shown to be a potential target of the quinoline antimalarial compounds [Graves et al., (2002) Mol. Pharmacol. 62, 1364]. QR2 catalyzes the two-electron reduction of menadione via the oxidation of N-alkylated or N-ribosylated nicotinamides . To investigate the mechanism and consequences of inhibition of QR2 by the quinolines further, we have used steady-state and transient-state kinetics to define the mechanism of QR2 . Importantly, we have shown that QR2 when isolated from an overproducing strain of E. coli is kinetically equivalent to the enzyme from the native human red blood cell source. We observe ping-pong kinetics consistent with one substrate/inhibitor binding site that shows selectivity for the oxidation state of the FAD cofactor, suggesting that selective inhibition of the liver versus red blood cell forms of malaria may be possible. The reductant N-methyldihydronicotinamide and the inhibitor primaquine bind exclusively to the oxidized enzyme. In contrast, the inhibitors quinacrine and chloroquine bind exclusively to the reduced enzyme. The quinone substrate menadione , on the other hand, binds nonspecifically to both forms of the enzyme. Single-turnover kinetics of the reductive half-reaction are chemically and kinetically competent and confirm the inhibitor selectivity seen in the steady-state experiments. Our studies shed light on the possible in vivo potency of the quinolines and provide a foundation for future studies aimed at creating more potent QR2 inhibitors and at understanding the physiological significance of QR2 .

Approved Relations

Singleton Relations to Curate

Added by: Alison_Jee

Chemical: MESH:D024483 — *Vitamin K 3*

Gene: 4835 — *N-ribosyldihydronicotinamide:quinone reductase 2 [NQO2]*

Relation: **affects^binding**

Creator's Notes:

Notes: The quinone substrate menadione, on the other hand, binds nonspecifically to both forms of the enzyme.

(●) Accept this relation
( ) Reject this relation

Reset review of this relation

✓

Added by: Parasvi_Patel

Chemical: MESH:C037219 — *quinoline*

Gene: 4835 — *N-ribosyldihydronicotinamide:quinone reductase 2 [NQO2]*

Relation: **affects^binding**

Creator's Notes:

Notes:

( ) Accept this relation
(●) Reject this relation

Reset review of this relation

✗

Reviewer can Accept or Reject the singleton Relationship

Curation: Annotation Agreement

	Curators				
Relationships	**A**	**B**	**C**	**D**	**E**
All	0.85	0.84	0.83	0.88	0.88
CTD-derived	0.99	0.96	0.97	0.99	0.96
New	0.76	0.77	0.69	0.82	0.83

Agreement F1 scores for the 5 annotators against the 'majority-approved' reference subset.

Agreement for CTD-derived relationships (prompted in curation UI) is higher than for New relationships (curator has to enter the new relationship).

Outline

- Task definition
- ChemDisGene: Automated Derivation
- ChemDisGene: Manual Curation Process
- Dataset Statistics
- Benchmark Models

Data Statistics: Highlights

- Distantly supervised *training* data:
  76,942 abstracts with entity linking from PubTator
  and 167k relation labels from CTD

- Manually curated *test* data: 523 abstracts (3,911 relationships).

- 18 relation types among Chemicals, Diseases, Genes/Gene products

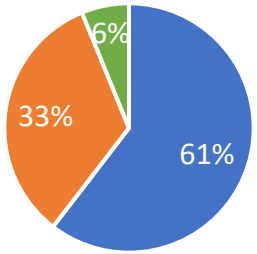
Corpus Statistics

	CTD-derived Corpus			Curated Corpus
	Train	Dev	Test	
Nbr. abstracts	76,942	1,521	1,939	523
Abstracts w/o relations	7,244	397	436	5
Nbr. relationships	167,005	3,290	5,116	3,911
Unique relationships	93,801	3,127	4,801	3,806
Total entity mentions	1,532,117	36,114	49,839	14,248
Chemicals	686,102	13,986	19,895	5,739
Diseases	478,397	8,962	11,750	2,931
Genes	367,618	13,166	18,194	5,578
Unique entities in relationships	14,991	1,894	2,345	1,875
Chemicals	7,187	759	999	670
Diseases	2,413	283	287	318
Genes	5,391	852	1,059	887

Distr. Of Relationships, CTD-derived Corpus

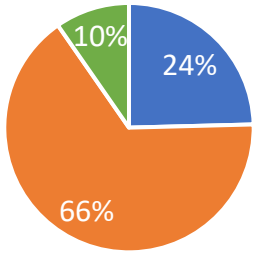
Pub. Years

Training



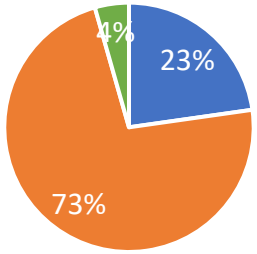
≤ 2017
(167k relns)

Dev

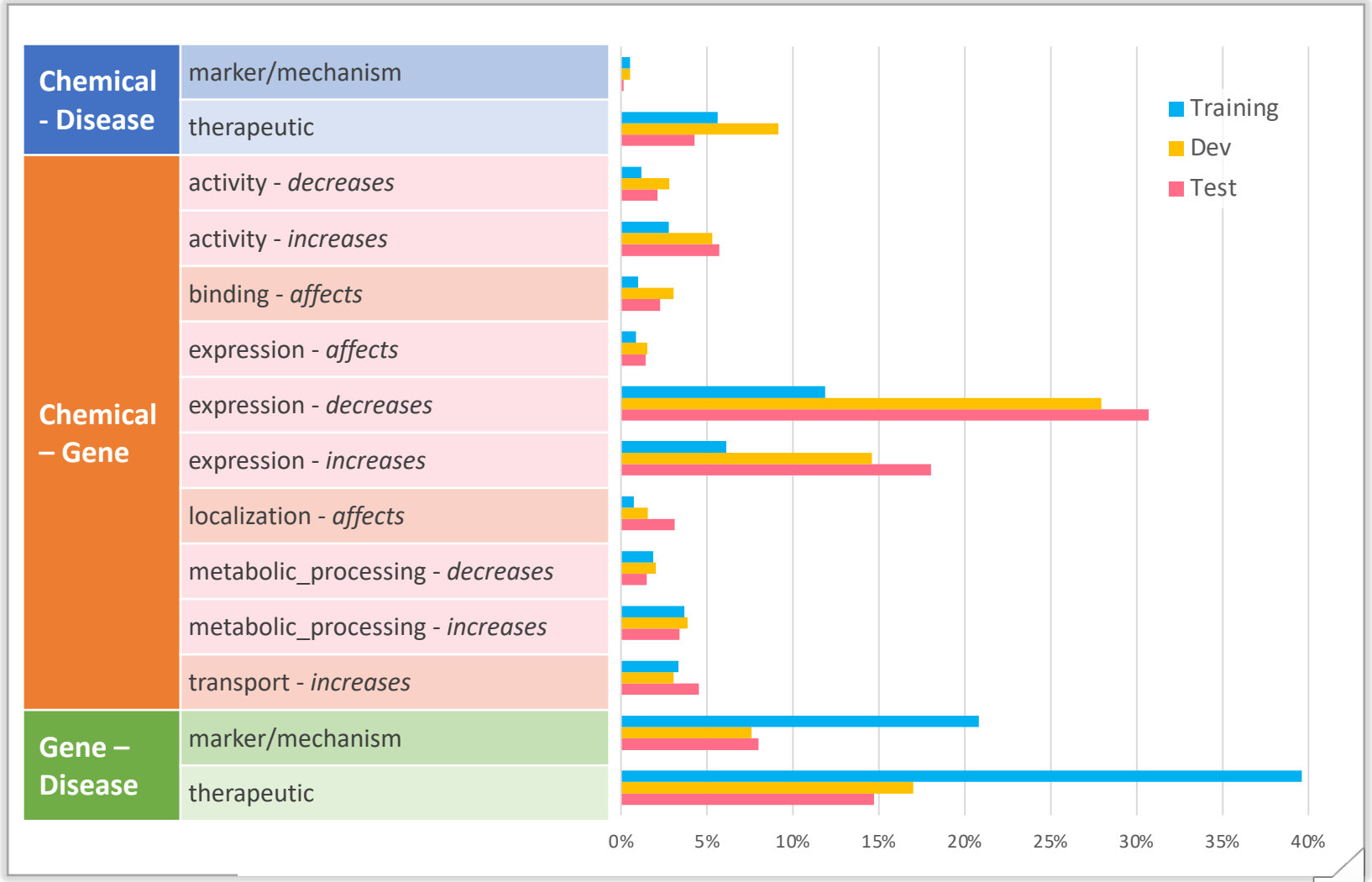


2018
(3,290)

Test



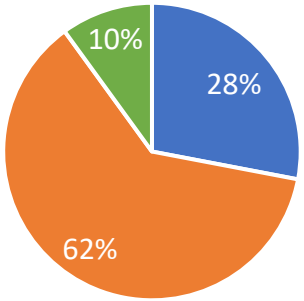
2019+
(5,116)



Distr. Of Relationships, Curated Corpus

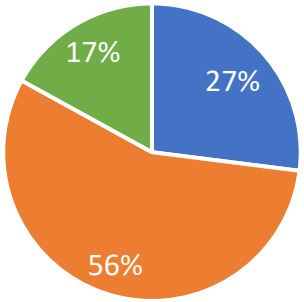
Nbr. Relationships

CTD-derived



1,279

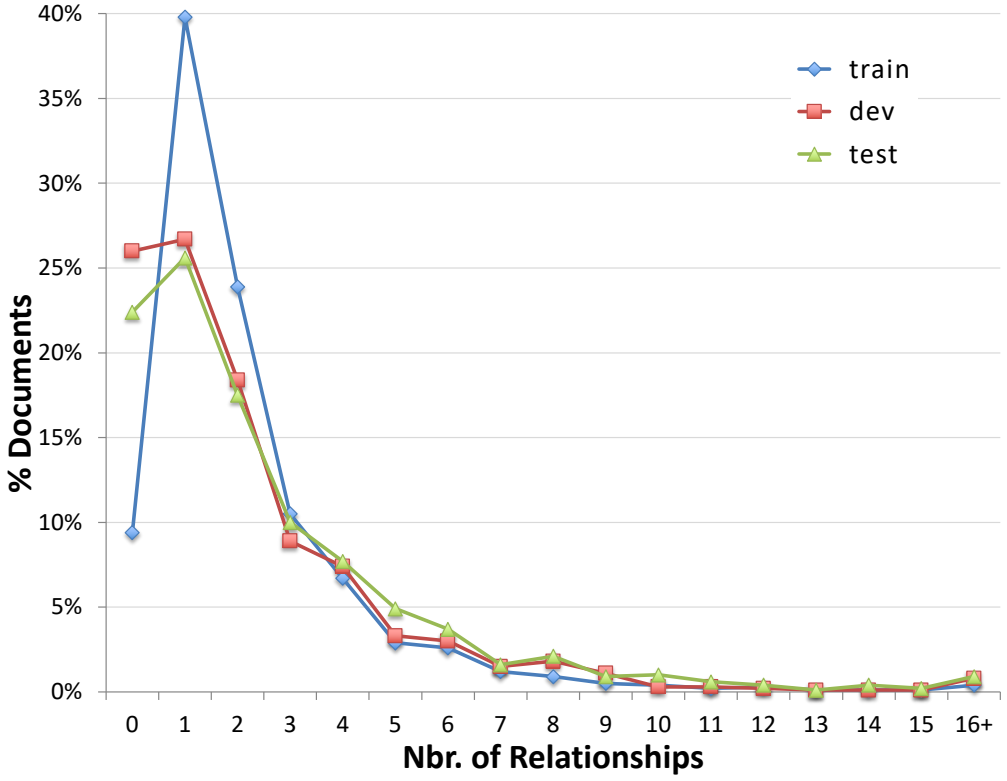
New



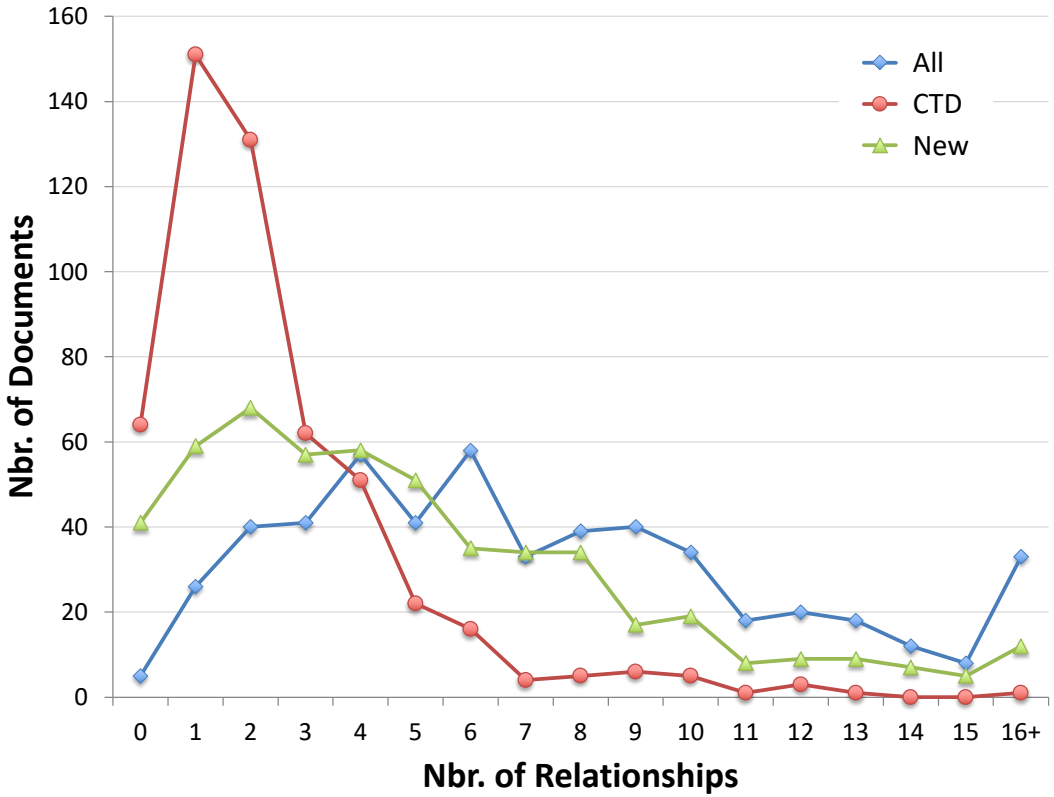
2,632

Chemical-Disease	CTD-... (%)	New (%)
marker/mechanism	16.4	16.6
therapeutic	12.0	10.4
Chemical-Gene	**CTD-derived**	**New**
activity - *affects*		1.2
activity - *decreases*	7.3	8.3
activity - *increases*	7.8	8.7
binding - *affects*	6.7	4.3
expression - *affects*	0.6	2.8
expression - *decreases*	13.1	10.4
expression - *increases*	18.4	11.8
localization - *affects*	1.5	0.8
metabolic_processing - *affects*		0.8
metabolic_processing - *decreases*	1.5	1.7
metabolic_processing - *increases*	4.0	3.0
transport - *affects*		0.8
transport - *decreases*		0.6
transport - *increases*	0.9	1.1
Gene-Disease	**CTD-derived**	**New**
marker/mechanism	9.3	14.1
therapeutic	0.5	2.9

Nbr. of Relationships per Document



CTD-derived Corpus

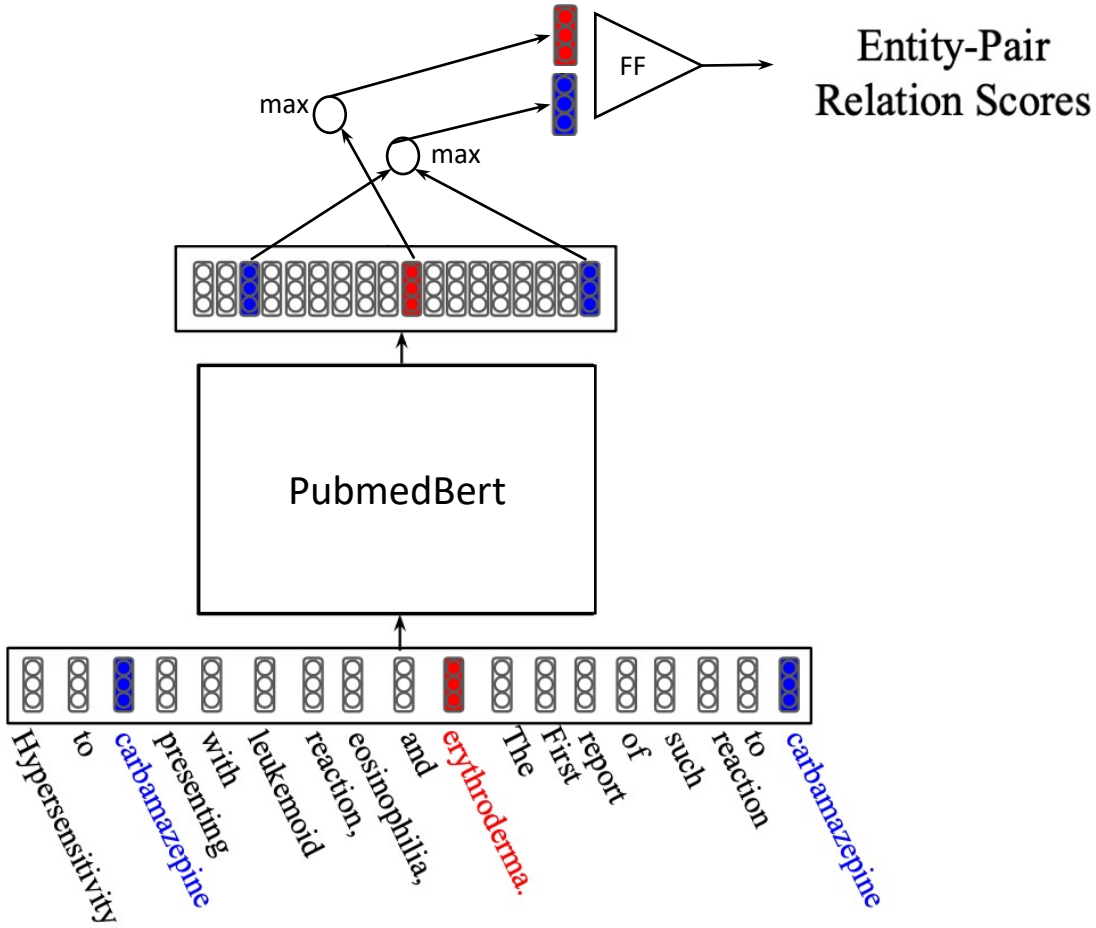


Curated Corpus

Outline

- Task definition
- ChemDisGene: Automated Derivation
- ChemDisGene: Manual Curation Process
- Dataset Statistics
- Benchmark Models

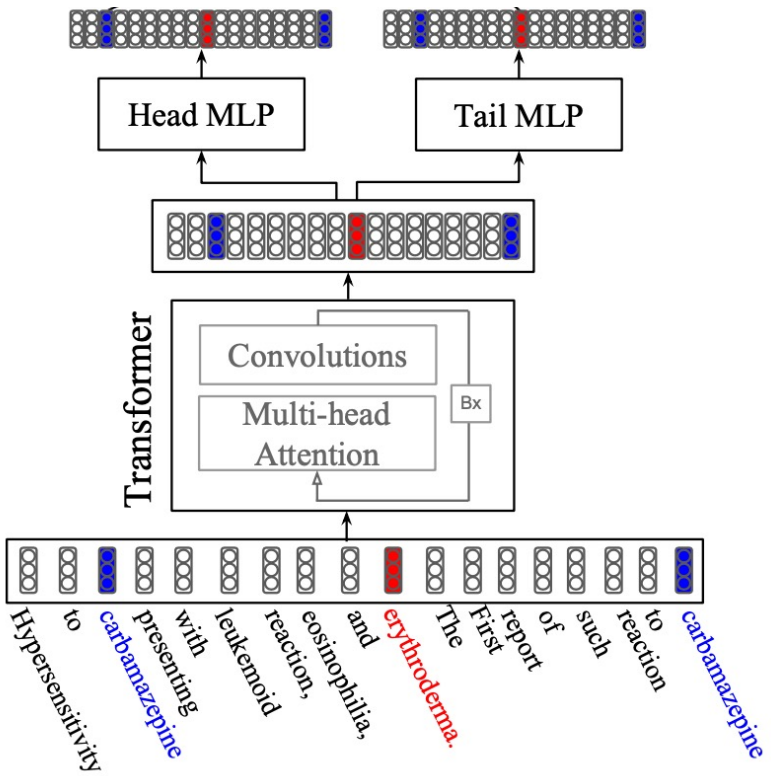
PubmedBert



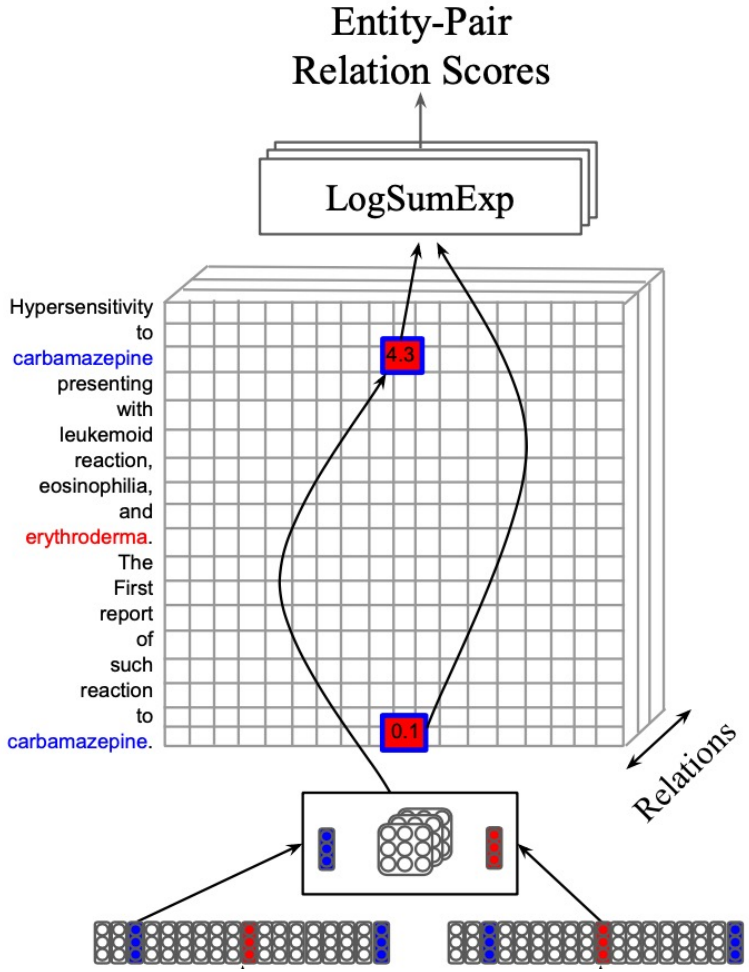
… (Gu et al, 2021)

Bi-Affine Relation Attention Networks (BRAN)

Transformer-based Text Encoder



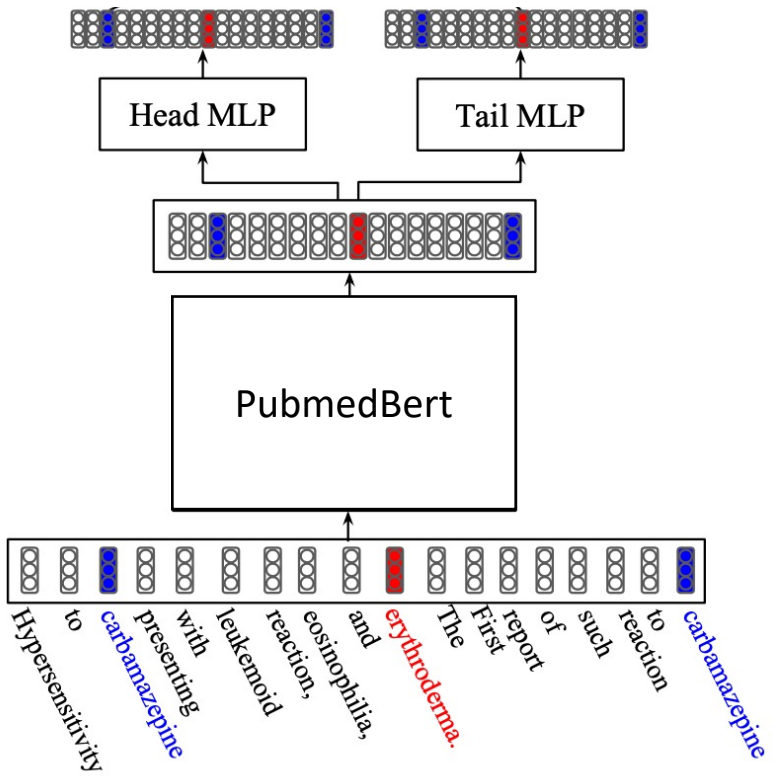
Bilinear Relation Scorer



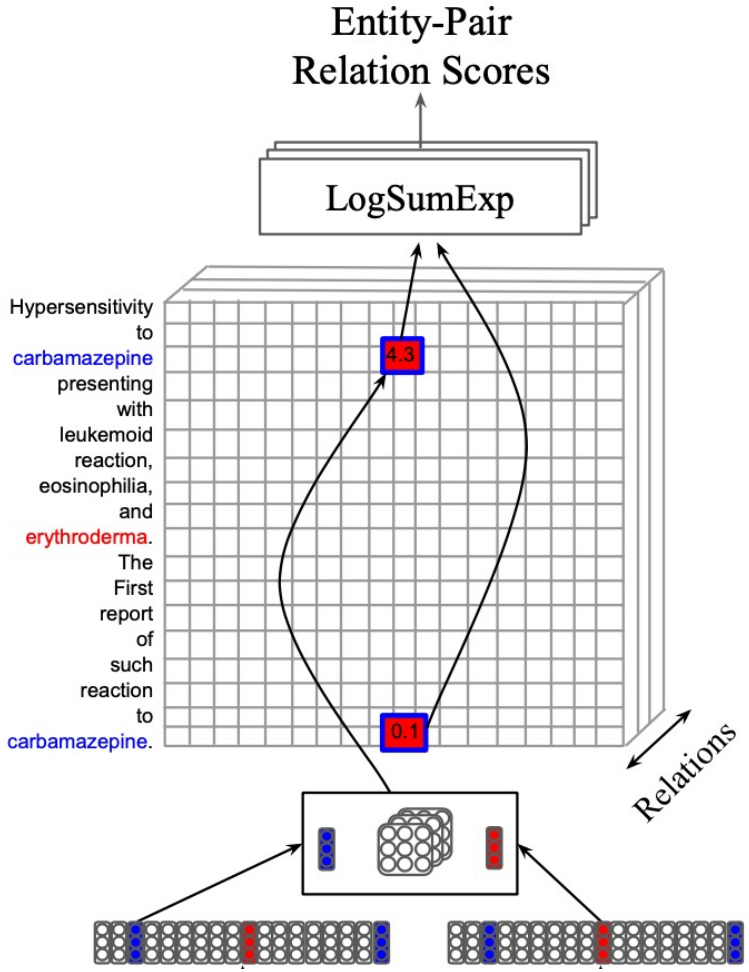
… (Verga et al, 2018)

PubmedBert + BRAN

BERT-based Text Encoder



Bilinear Relation Scorer



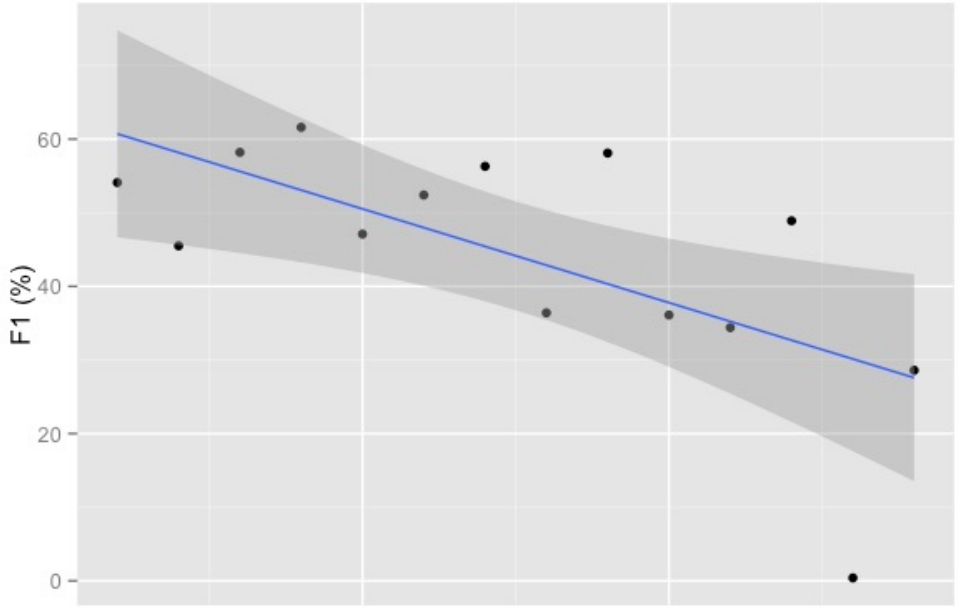
Benchmark RE Model Results

Model	Micro P	Micro R	Micro F1	Micro Avg. P	Macro P	Macro R	Macro F1
CTD-derived corpus: 'dev' split / 'test' split							
BRAN	32.1 / 31.7	46.3 / 44.2	37.9 / 36.9	28.4 / 27.9	25.9 / 23.6	32.3 / 30.1	28.2 / 26.0
PubmedBert	50.3 / 49.6	59.3 / 56.1	54.5 / 52.6	50.3 / 50.1	43.6 / 39.0	50.3 / 48.4	44.9 / 41.7
PubmedBert + BRAN	53.9 / 53.9	61.0 / 57.3	57.3 / 55.6	54.0 / 54.3	45.0 / 42.7	54.1 / 50.4	48.7 / 44.4
Curated corpus: CTD-derived relationships only / All relationships							
BRAN	24.4 / 41.8	45.8 / 26.6	31.8 / 32.5	28.1 / 33.5	20.3 / 37.2	35.7 / 22.5	24.5 / 25.8
PubmedBert	43.0 / 64.3	61.7 / 31.3	50.7 / 42.1	50.7 / 46.9	34.7 / 53.7	53.4 / 32.0	39.6 / 37.0
PubmedBert + BRAN	46.5 / 70.9	61.1 / 31.6	52.8 / 43.8	53.0 / 50.6	45.8 / 69.8	59.0 / 32.5	47.0 / 40.5

Model Performance per Relation Type

Metrics for PubmedBert + BRAN on the Curated Corpus,

sorted on decreasing relation frequency in Training data.



Relation Type	F1
Chemical-Disease : marker/mechanism	54.1
Chemical-Disease : therapeutic	45.5
Chemical-Gene : expression - increases	58.2
Chemical-Gene : expression - decreases	61.6
Gene-Disease : marker/mechanism	47.1
Chemical-Gene : activity - increases	52.4
Chemical-Gene : activity - decreases	56.3
Chemical-Gene : metabolic_processing - increases	36.4
Chemical-Gene : binding - affects	58.1
Chemical-Gene : transport - increases	36.1
Chemical-Gene : metabolic_processing - decreases	34.4
Chemical-Gene : localization - affects	48.9
Chemical-Gene : expression - affects	0.4
Gene-Disease : therapeutic	28.6

Thankyou!

dataset:
https://github.com/chanzuckerberg/ChemDisGene