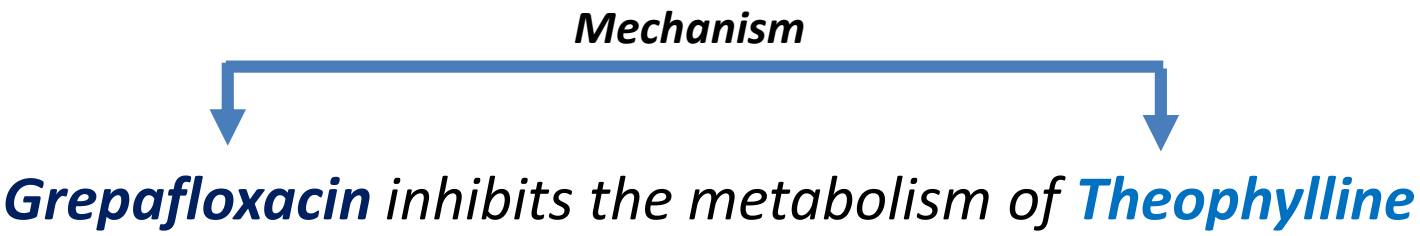
Enhancing Drug-Drug Interaction Extraction from Texts by Molecular Structure Information

Masaki Asada, Makoto Miwa, Yutaka Sasaki

Toyota Technological Institute, Japan

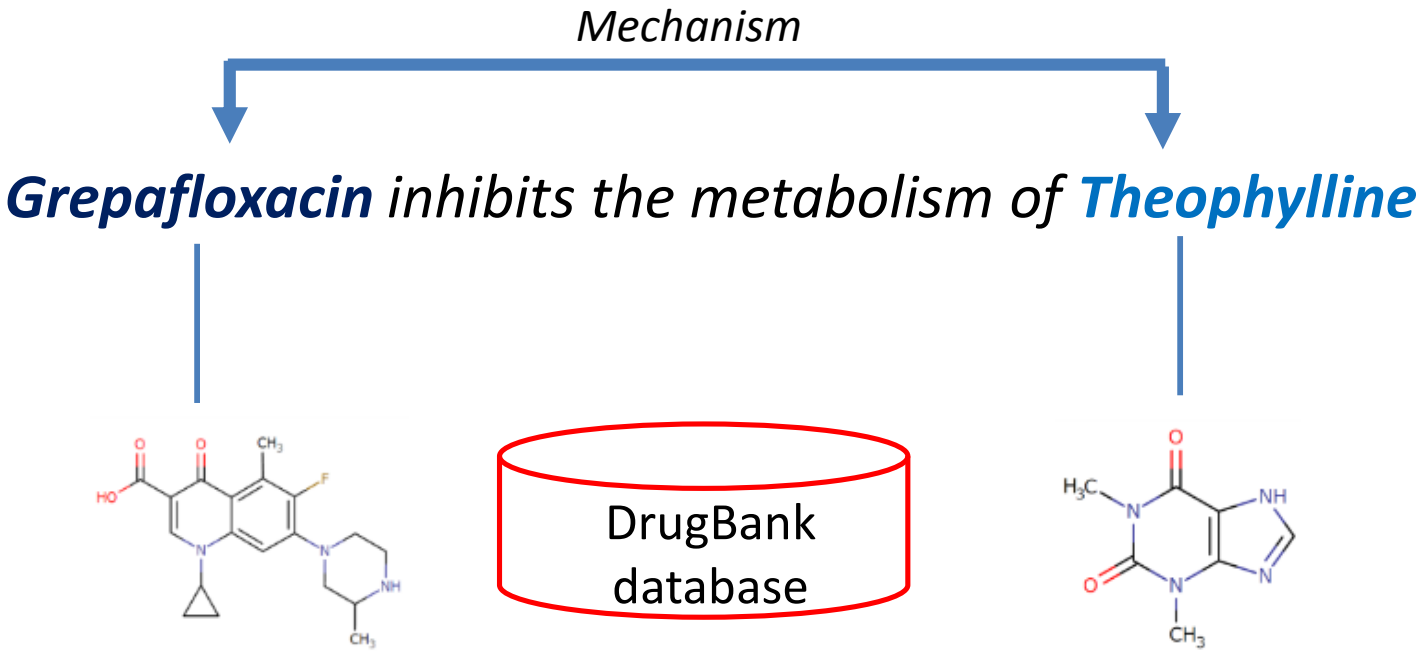
Introduction

- Our target problem is the extraction of drug-drug interactions (DDIs) from biomedical texts



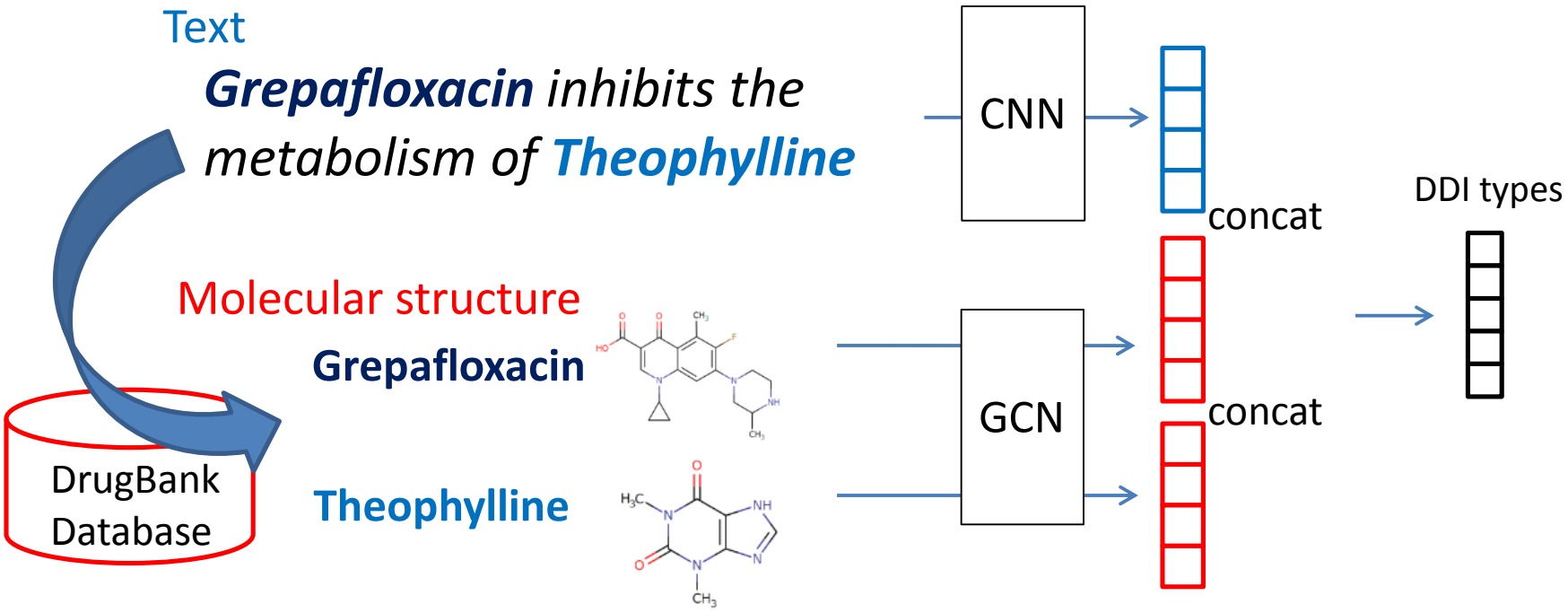
Introduction

- Our target problem is the extraction of drug-drug interactions (DDIs) from biomedical texts
- We investigate the use of external drug database (DrugBank) information in extracting DDIs from texts
- We especially focus on **molecular structure information**



Method Overview

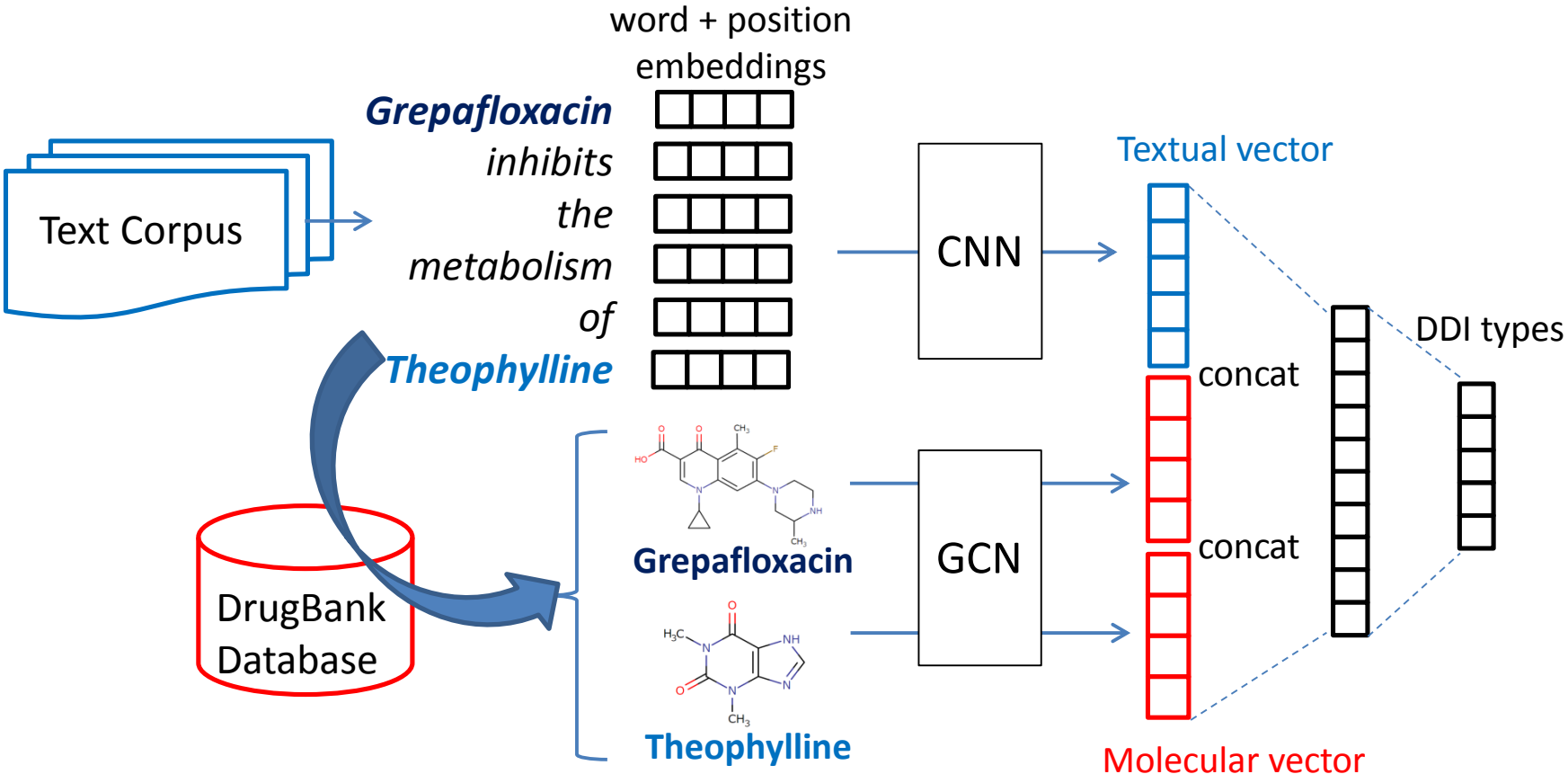
- We obtain the representations of textual drug pairs using convolutional neural networks (CNNs) and molecular drug pairs using graph convolutional networks (GCNs)
- We concatenate text-based and molecule-based vectors



Method

DDI extraction from texts using molecular structures

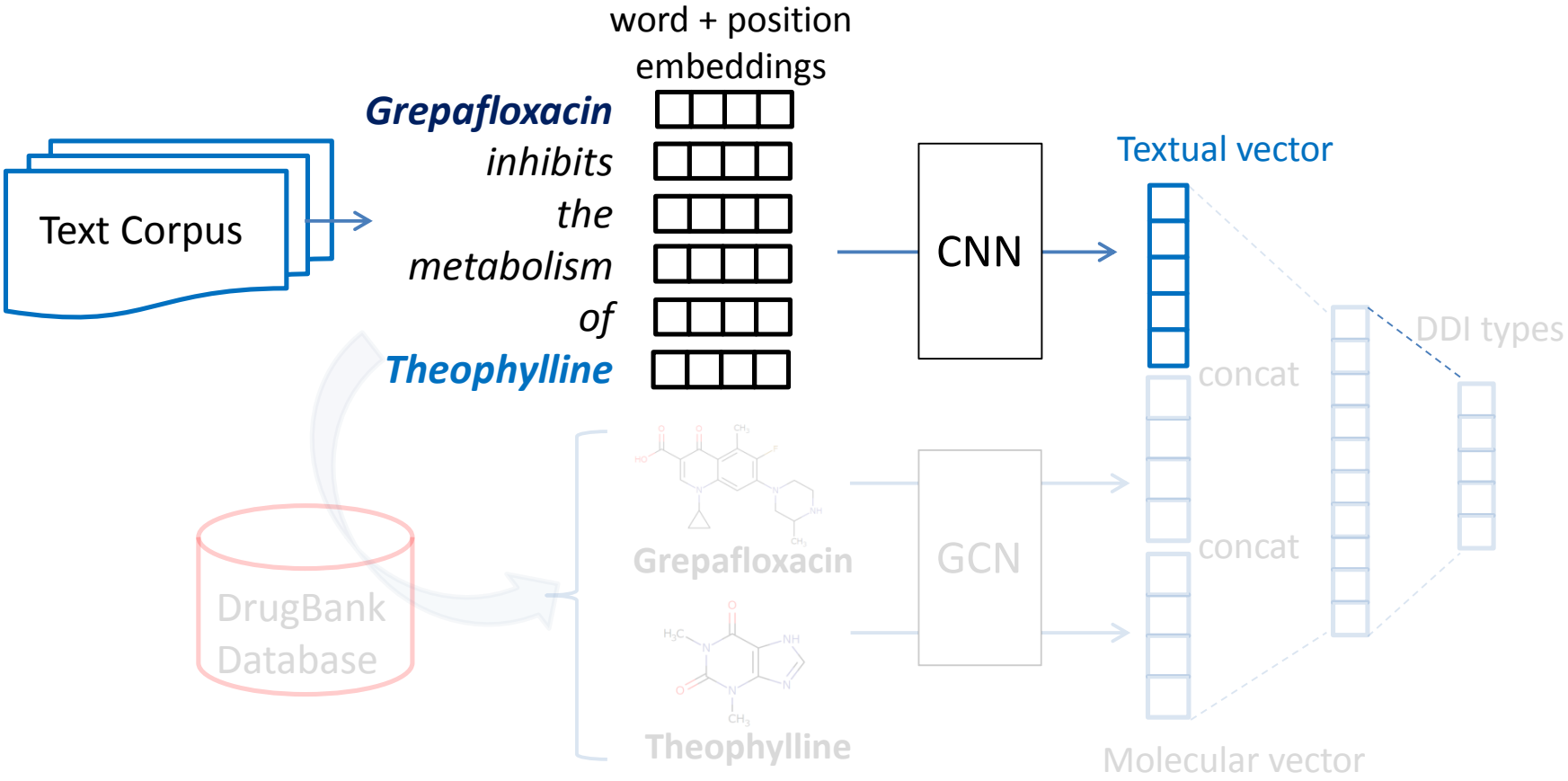
- Text-based DDI representation
- Molecular structure-based DDI representation



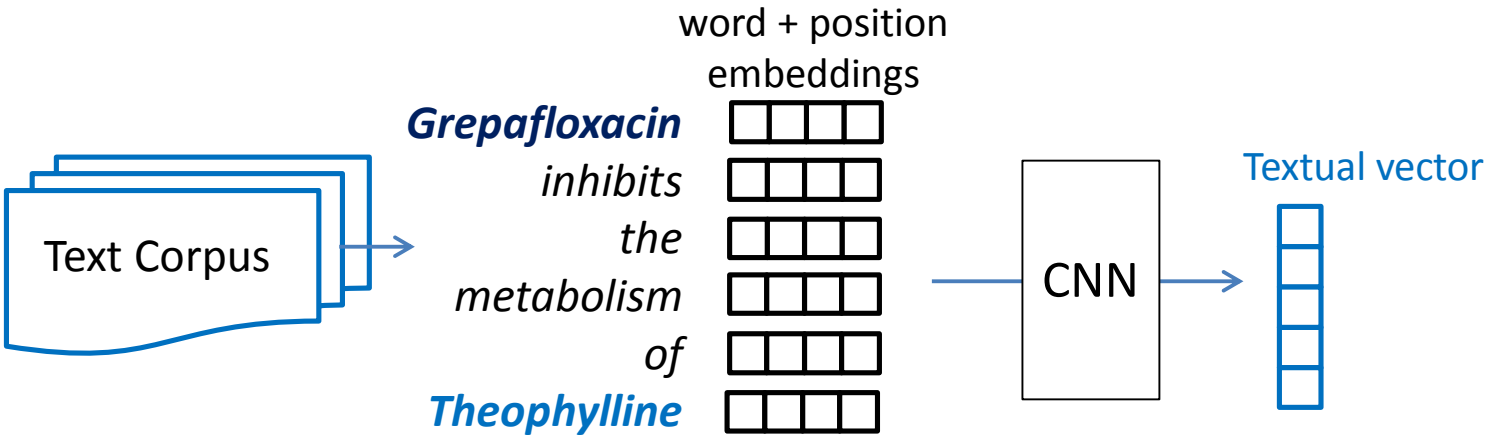
Method

DDI extraction from texts using molecular structures

- Text-based DDI representation
- Molecular structure-based DDI representation



Method: Text-based DDI Representation

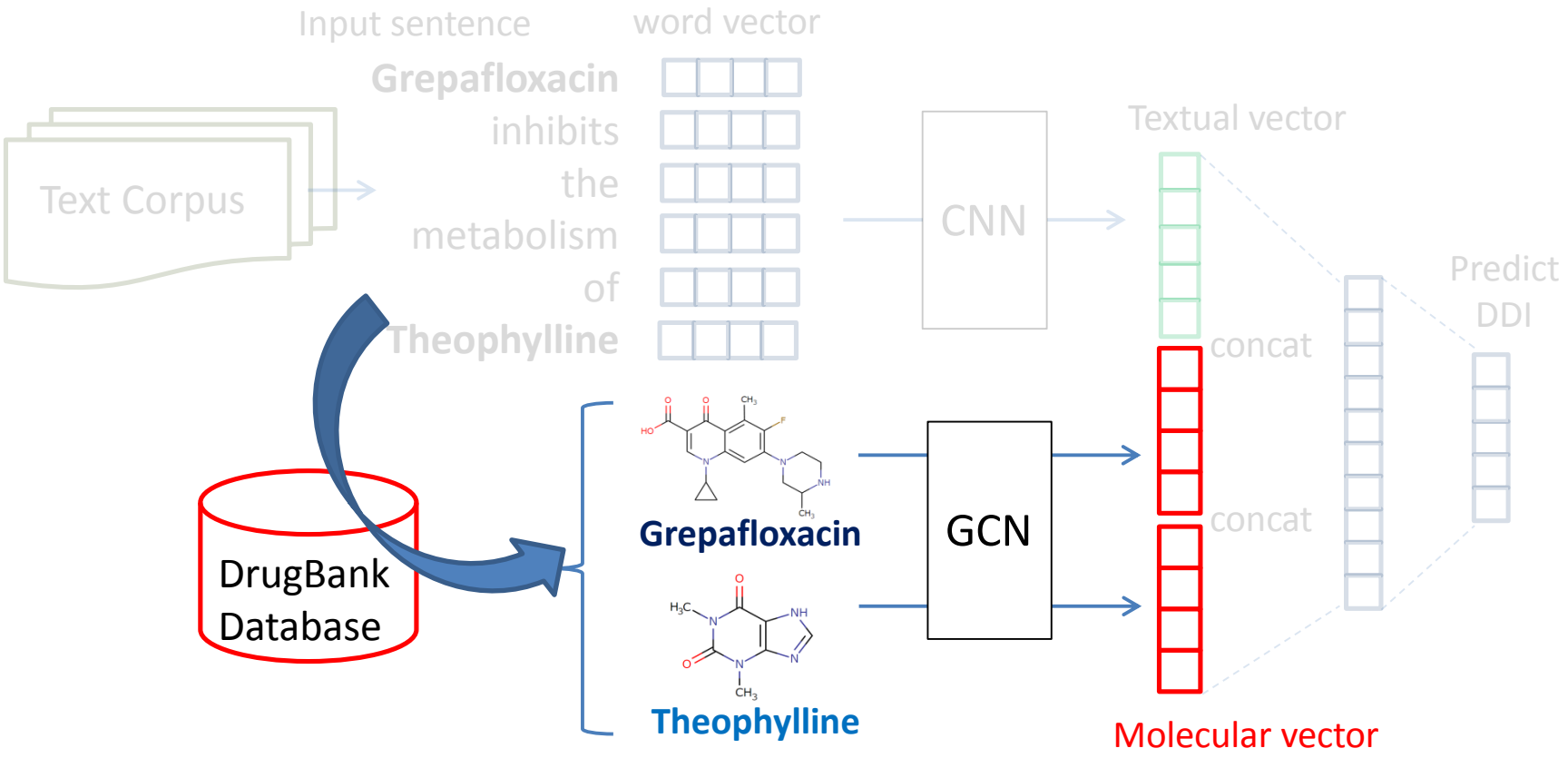


- Our model for representing textual DDIs is based on the CNN model by Zeng et al. (2014)
- We use word and position embeddings as the input to the convolution layer
- We convert the output of the convolution layer into a fixed-size textual vector

Method

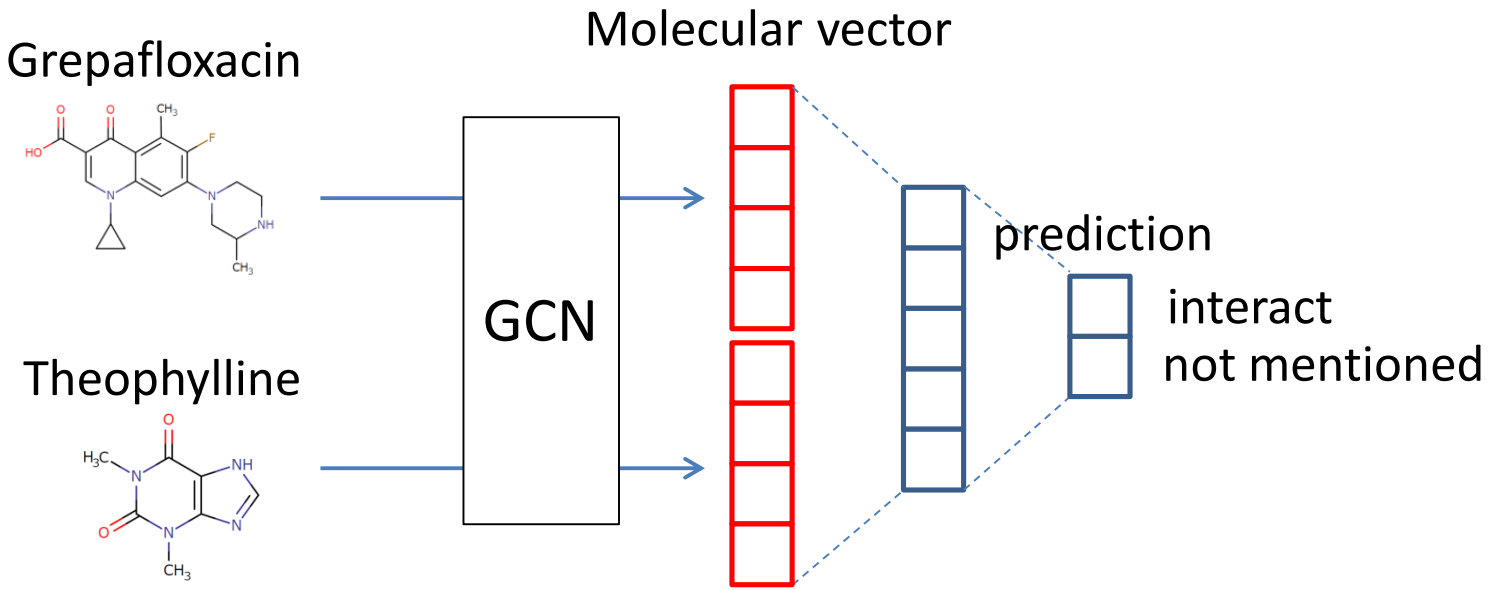
DDI extraction from texts using molecular structures

- Text-based DDI representation
- Molecular structure-based DDI representation



Method: Molecular Structure-based DDI Representation

- We represent drug pairs in molecular graph structures using GCNs
- We **pre-train** GCNs using interacting (positive) pairs mentioned in the DrugBank and not mentioned (**pseudo negative**) pairs in the DrugBank

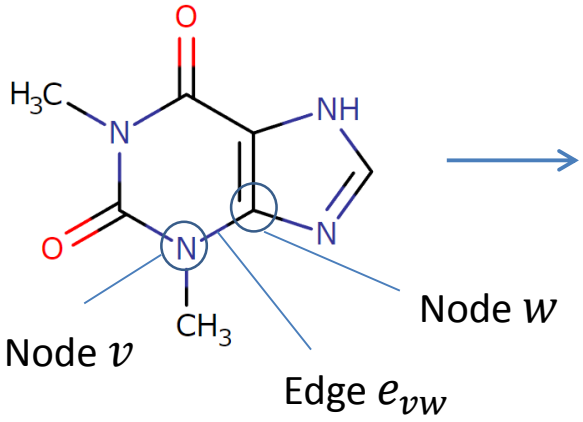


Method: Molecular Structure-based DDI Representation

Graph Convolutional Network (GCN) [Li et al. 2016]

We use GCNs to convert a drug molecule graph into a fixed size vector by aggregating node vectors $\boldsymbol{h}_v^T$

graph structure



GCN

molecular vector

$\boldsymbol{g}$

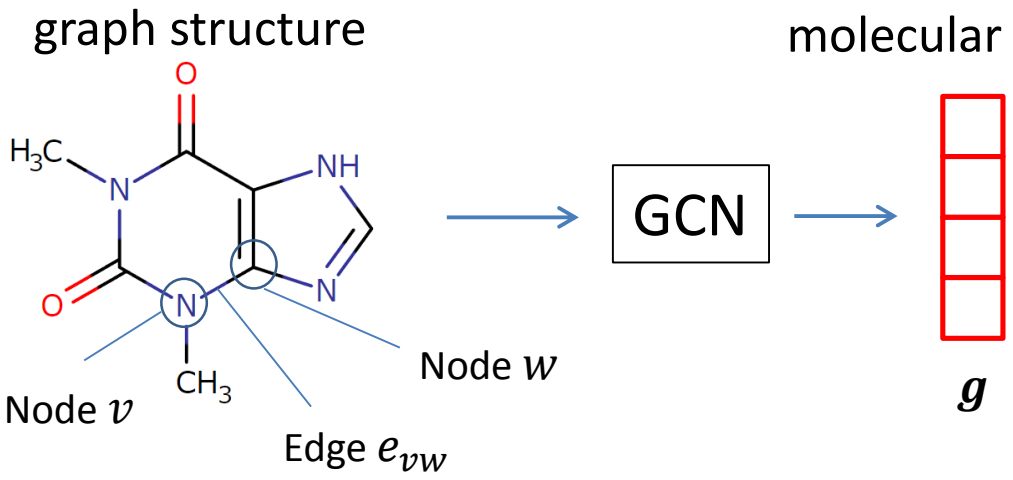
$$\boldsymbol{m}_v^{t+1} = \sum_{w \in N(v)} \boldsymbol{A}_{e_{vw}} \boldsymbol{h}_w^t$$
$$\boldsymbol{h}_v^{t+1} = \mathrm{GRU}([\boldsymbol{h}_v^t; \boldsymbol{m}_v^{t+1}])$$

$\boldsymbol{h}_v^t$: node vector
$N(v)$: neighbors of v
GRU : gated Recurrent Unit
i, j : linear layer
$\odot$: element-wise product
$[\ldots;\ldots]$: concatenation
$\boldsymbol{A}$: learned weight

Method: Molecular Structure-based DDI Representation

Graph Convolutional Network (GCN) [Li et al. 2016]

We use GCNs to convert a drug molecule graph into a fixed size vector by aggregating node vectors $\boldsymbol{h}_v^T$



$$\boldsymbol{m}_v^{t+1} = \sum_{w \in N(v)} \boldsymbol{A}_{e_{vw}} \boldsymbol{h}_w^t$$

$$\boldsymbol{h}_v^{t+1} = \mathrm{GRU}([\boldsymbol{h}_v^t; \boldsymbol{m}_v^{t+1}])$$

$$\boldsymbol{g} = \sum_v \sigma(i([\boldsymbol{h}_v^T; \boldsymbol{h}_v^0])) \odot (j([\boldsymbol{h}_v^T; \boldsymbol{h}_v^0]))$$

$\boldsymbol{h}_v^t$: node vector

$N(v)$: neighbors of v

GRU : gated Recurrent Unit

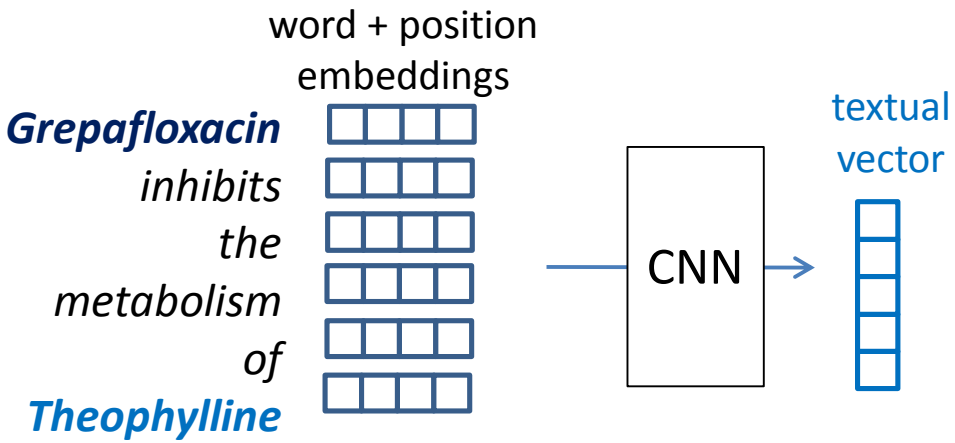
i, j : linear layer

$\odot$: element-wise product

$[\ldots; \ldots]$: concatenation

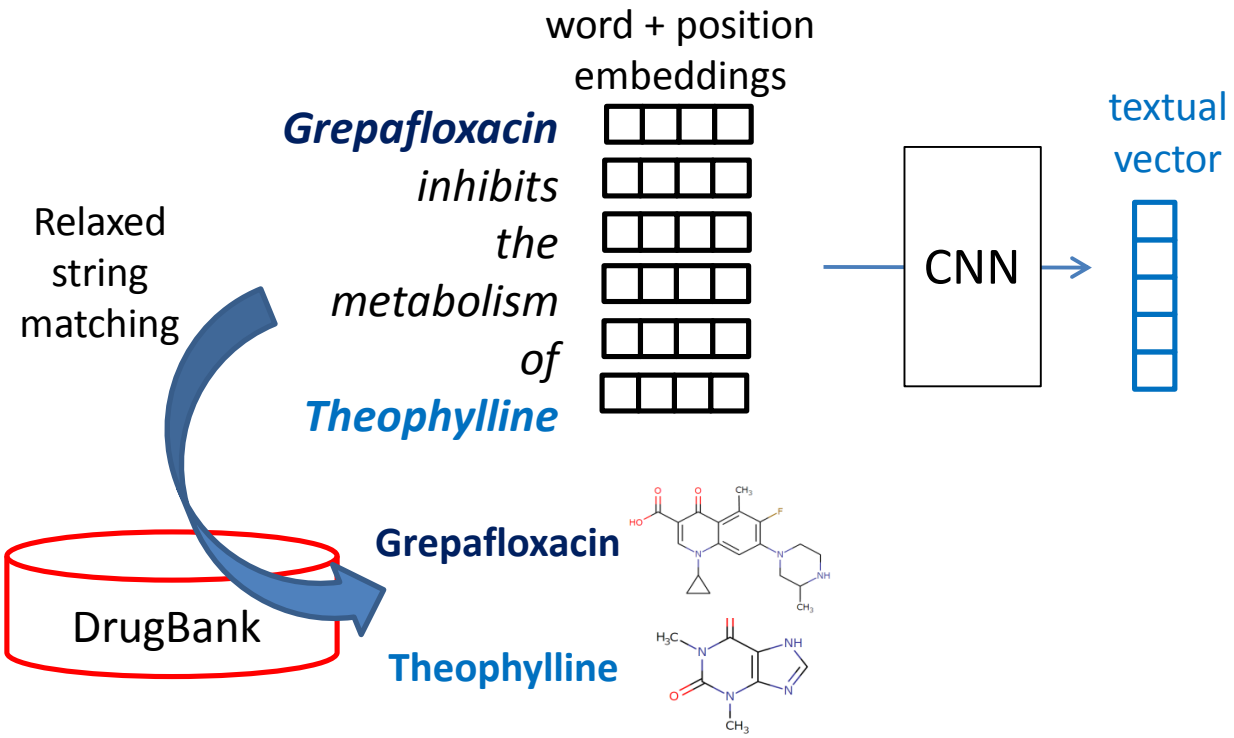
$\boldsymbol{A}$: learned weight

Method: DDI Extraction from Texts Using Molecular Structures



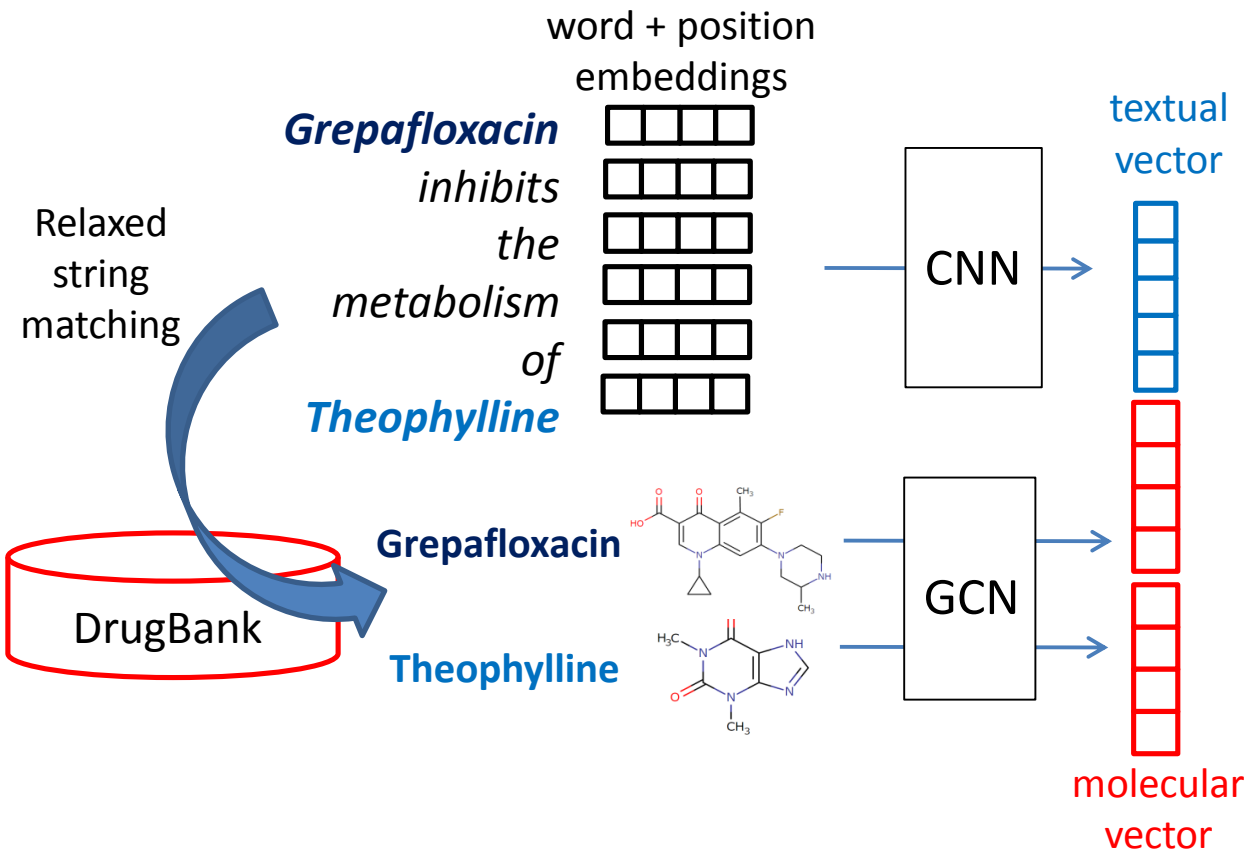
Method: DDI Extraction from Texts Using Molecular Structures

- Link mentions in text corpus to drug database entries by relaxed string matching



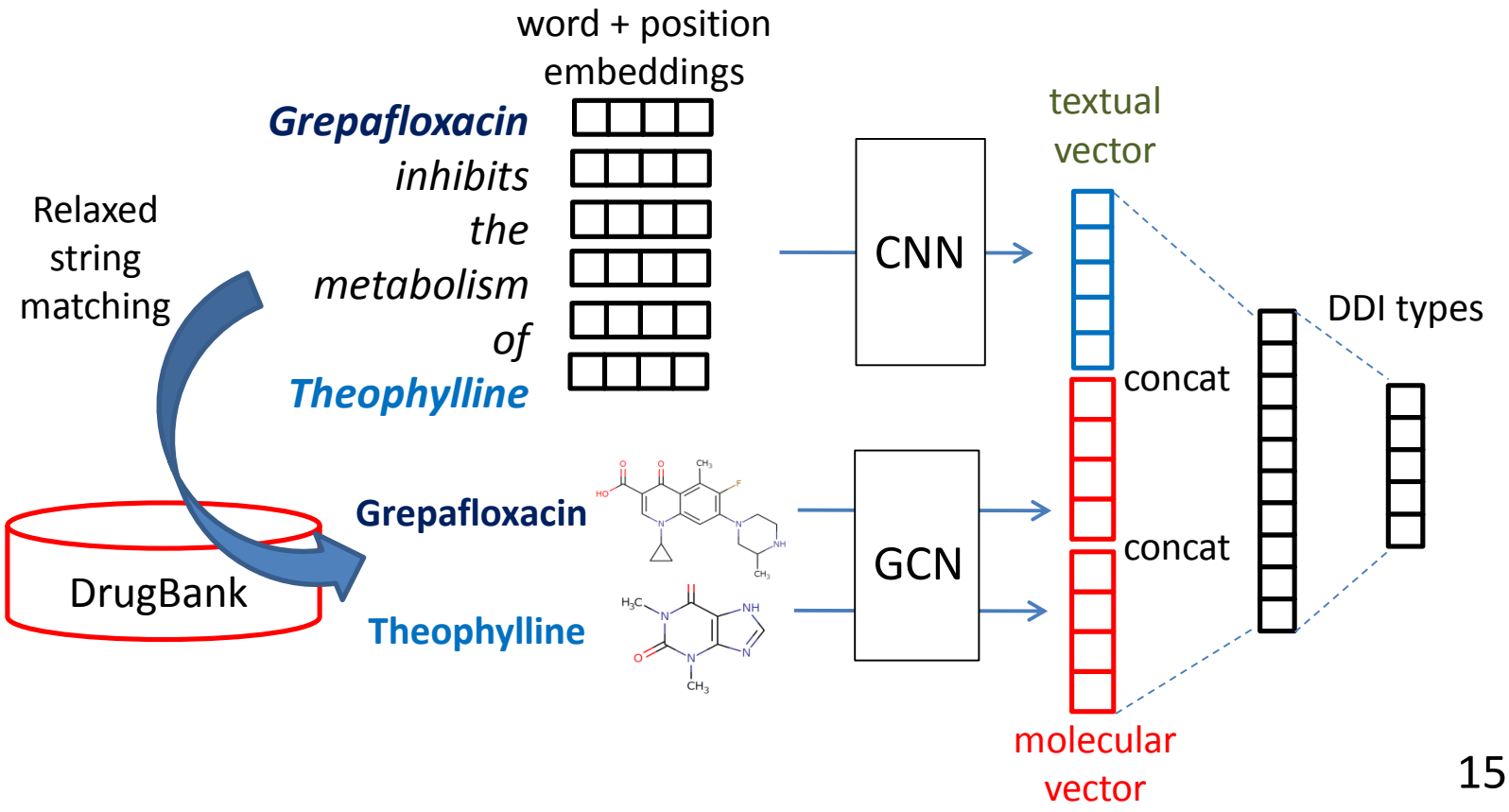
Method: DDI Extraction from Texts Using Molecular Structures

- Link mentions in text corpus to drug database entries by relaxed string matching
- Obtain molecular vectors via GCNs with fixed parameters



Method: DDI Extraction from Texts Using Molecular Structures

- Link mentions in text corpus to drug database entries by relaxed string matching
- Obtain molecular vectors via GCNs with fixed parameters
- Predict DDIs from concatenated textual and molecular vectors



Task Settings

SemEval2013 shared task 9.2

The data set is composed of documents annotated with drug mentions and their 4 types of interactions (*Mechanism*, *Effect*, *Advice* and *Interaction*) or no interaction

	DDI type	Train	Test
Positive	*Mechanism*	1,319	302
	Effect	1,687	360
	Advice	826	221
	Int	189	96
	Total	4,021	979
Negative		23,771	4,737
Total		27,792	5,716

Statistics of the DDI SemEval2013 shared task

Data for Pre-training GCNs

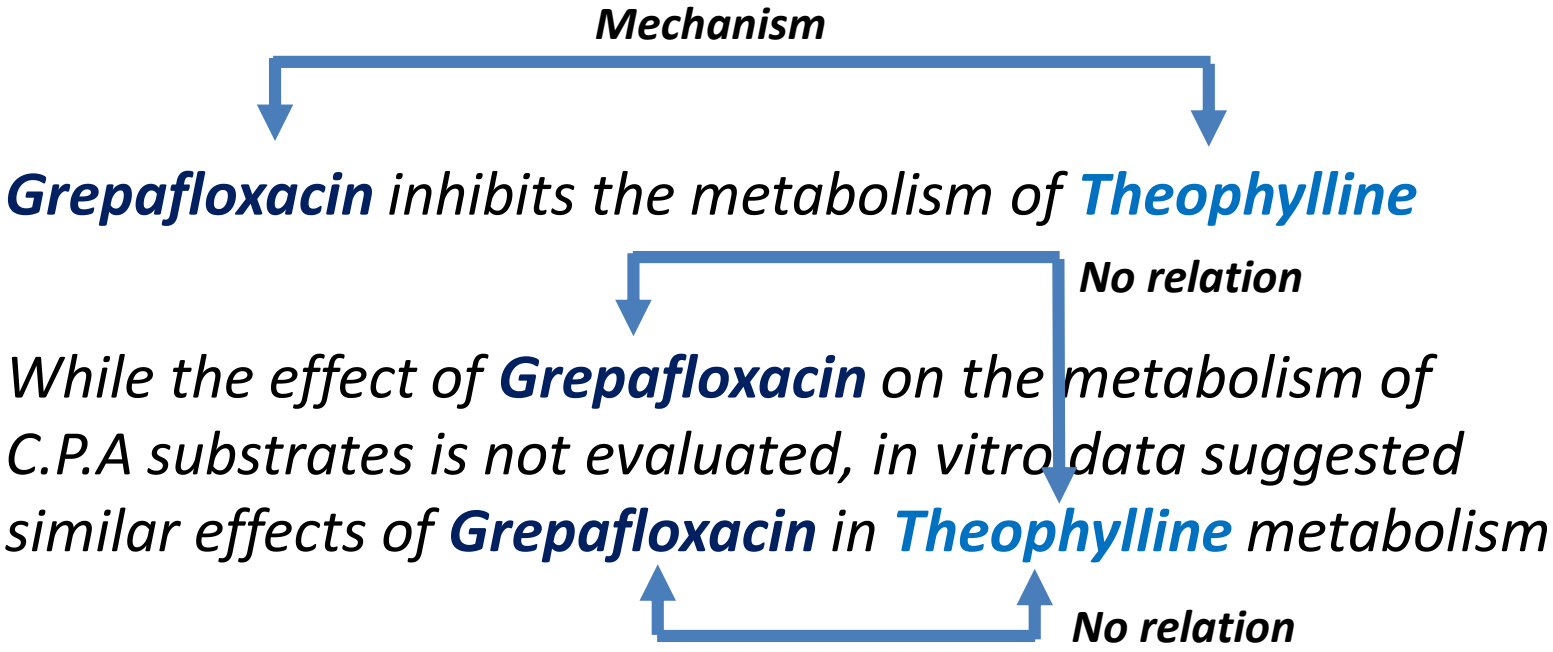
- We extracted 255,229 interacting (positive) pairs from DrugBank and generated the same number of ***pseudo negative pairs*** by randomly pairing DrugBank drugs
- We deleted drug pairs mentioned in the test set of the text corpus

Molecular Structure Features

- To obtain the graph of a drug molecule, we took as input the SMILES string encoding of the molecule from DrugBank and then converted it into the 2D graph structure using RDKit
- For the initial atom (node) vectors, we used randomly embedded vectors for atoms, i.e., *C*, *O*, *N*, …
- We also used 4 bond (edge) types: *single*, *double*, *triple*, and *aromatic*

Differences of Labels in Text and Database Tasks

- Interacting drug pairs in database may not appear as positive instances in the text task
- Text task define 4 detailed types, while database task has one positive type.



Training Settings

- Mini-batch training using the Adam optimizer with L2 regularization
- Word embeddings trained by the word2vec tool on the 2014 MEDLINE/PubMed baseline distribution
  - Skip-gram
  - Vocabulary size: 215k

Training Settings

Hyper-parameters

Parameter	Value
Word embedding size	200
Word position embedding size	20
Convolution window size	[3, 5, 7]
Convolution filter size	100
Hidden layer size	500
Initial learning rate	0.001
Mini-batch size	50
L2 regularization parameter	0.0001

Hyper-parameters for text-based model

Parameter	Value
Molecular vector size	50
Number of steps	4
Hidden layer size	1,000
Initial learning rate	0.001
Mini-batch size	100
Hidden layer size of NFP	50
GRU unit size of GGNN	50

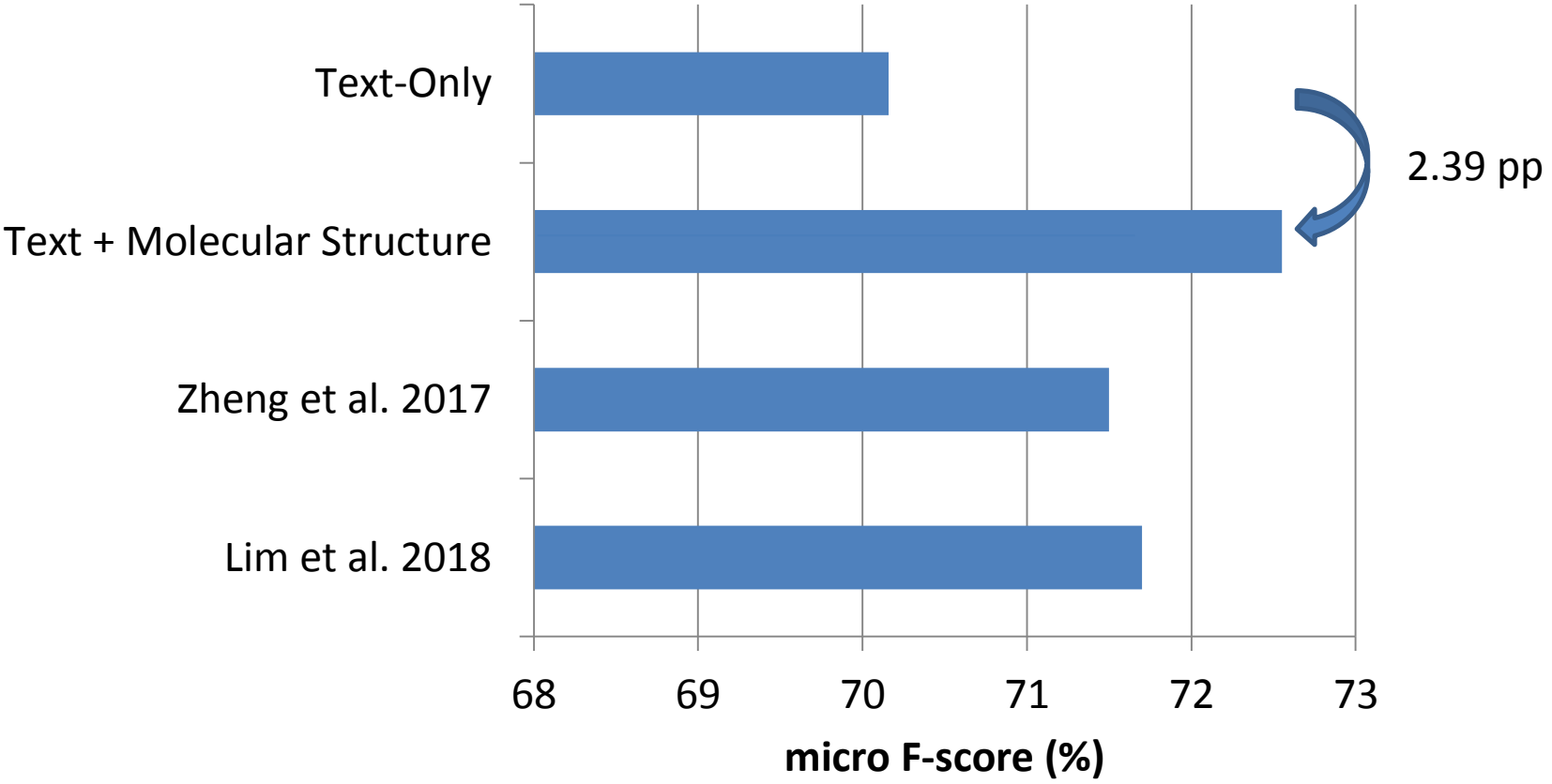
Hyper-parameters for molecule-based model

Evaluation on Relaxed String Matching

- How much of drug mentions in texts are linked to DrugBank entries by relaxed string matching?
  - We lowercased the mentions and the names in the entries and chose the entries with the most overlaps
  - As a result, 92.15% and 93.09% of drug mentions in train and test SemEval2013 data set matched the DrugBank entries

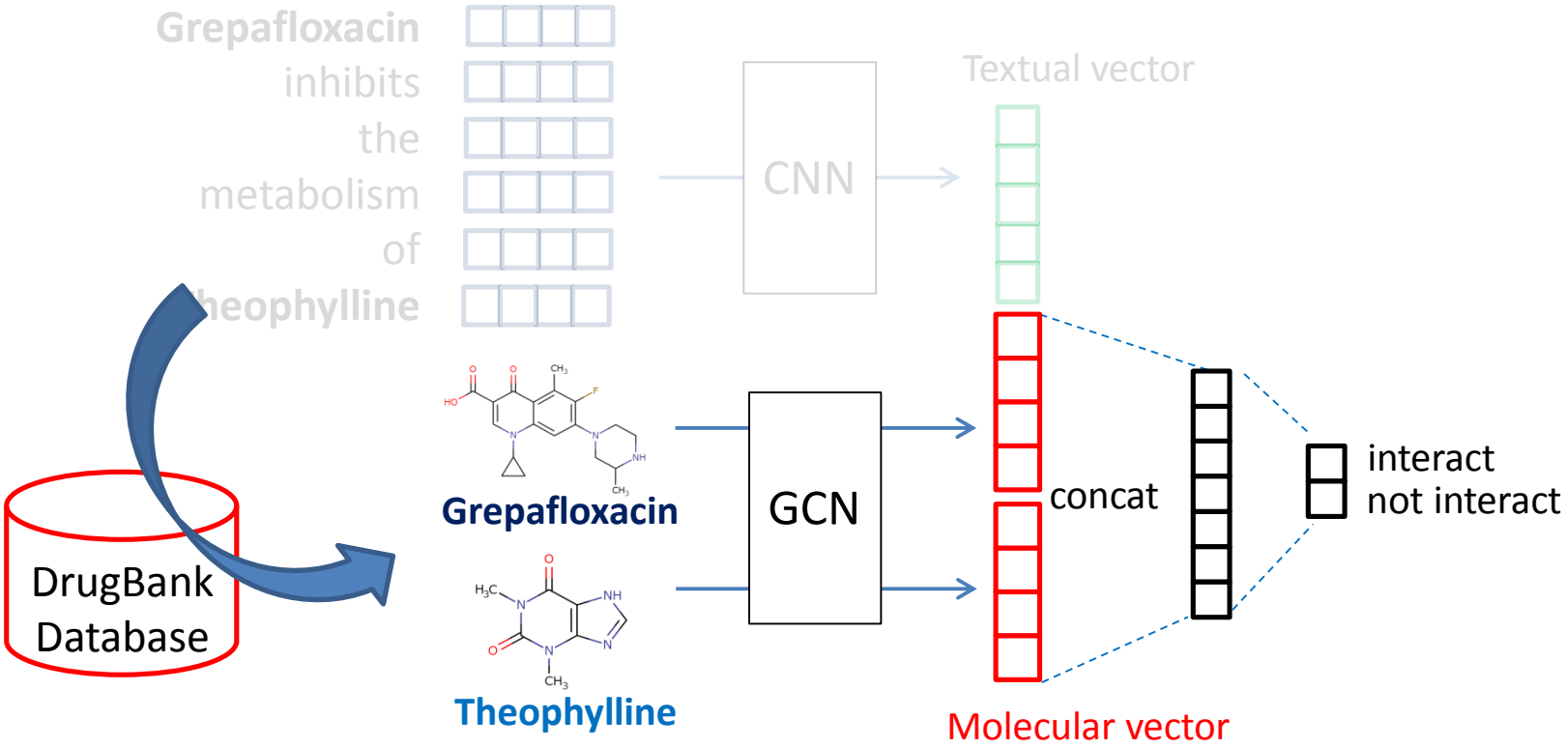
Evaluation on DDI Extraction from Texts (SemEval2013 Shared Task)

- We observe the increase of micro F-score by using molecular structures



Analysis

Can molecular structures alone represent DDIs in texts ?



- Low F-score (23.90%)
- This might be because the drug pairs that interact can appear in the textual context that does not describe their interactions

Conclusions

- We proposed a novel neural method for DDI extraction using both textual and molecular information
- The molecular information has improved DDI extraction performance
- As future work, we will investigate the use of other information in DrugBank