Word Embeddings for Chemical Patent Natural Language Processing

Camilo Thorne Saber Akhondi

Elsevier
{c.thorne.1,s.akhondi}@elsevier.com

Latinx @ ICML 2020

Problem

- NLP methods are key for processing chemistry patent, given their large quantity
- Many biomedical embeddings have been proposed/learnt over very large biomedical corpora
  1. Word2Vec / SkipGram: [TDW+19, PGM+13, SSM15]
  2. Contextualized / ELMo: [JDCL19]
- More recently, embeddings learnt over chemistry patents have been proposed [ZNA+19]

ELSEVIER

Problem

- NLP methods are key for processing chemistry patent, given their large quantity
- Many biomedical embeddings have been proposed/learnt over very large biomedical corpora
  1. Word2Vec / SkipGram: [TDW+19, PGM+13, SSM15]
  2. Contextualized / ELMo: [JDCL19]
- More recently, embeddings learnt over chemistry patents have been proposed [ZNA+19]

Compare embeddings by **extrinsic** and **intrinsic evaluation**!

ELSEVIER

Embeddings

Embedding	Vocabulary	Dimensions	Source
Mat2Vec W2V	529,686	200	[TDW+19]
PubMed W2V	2,351,706	200	[PGM+13]
Drug W2V	553,195	420	[SSM15]
CheMU W2V	1,252,586	200	[ZNA+19]

Embedding	Dimensions	Source
PubMed ELMo	1,204	[JDCL19]
CheMU ELMo	1,204	[ZNA+19]

Extrinsic Evaluation – Chemical NER

Task:
- ▷ detect chemical compounds in patents

Datasets:
- ▷ Biosemantics chemical NER gold standard
- ▷ SCAI corpus

Methods:
- ▷ biGRU-CRF NER model
- ▷ compare F1 scores

Extrinsic Evaluation – Chemical NER

Task:
- ▷ detect chemical compounds in patents

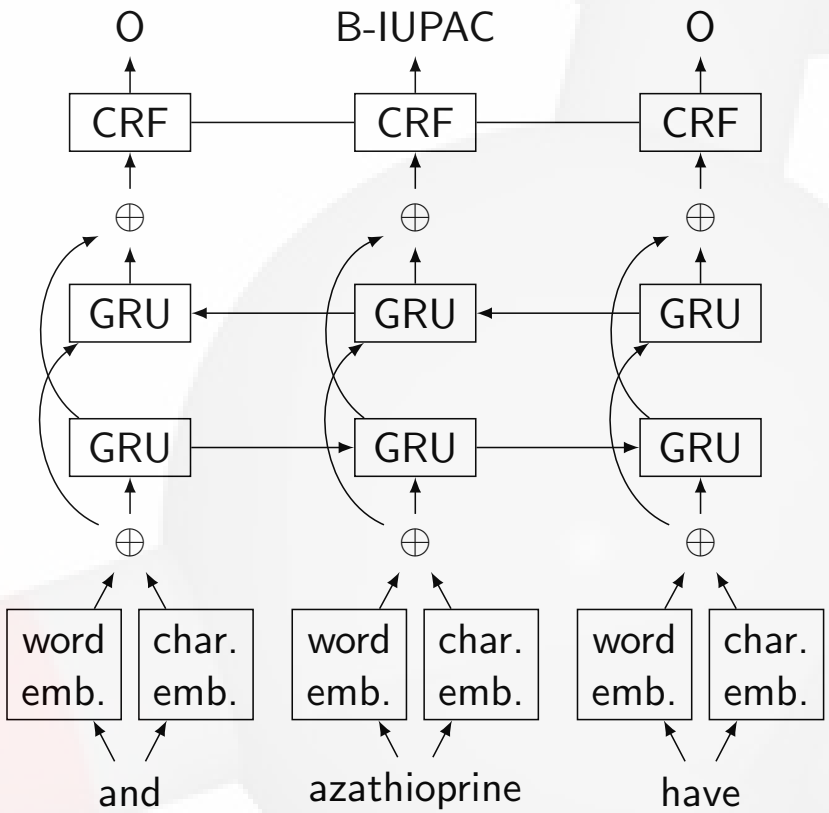
Datasets:
- ▷ Biosemantics chemical NER gold standard
- ▷ SCAI corpus

Methods:
- ▷ biGRU-CRF NER model
- ▷ compare F1 scores

Split	Entities	Tokens
Train	731 IUPAC, 212 Modifier, 73 Partiupac	33,457
Validation	240 IUPAC	4,654
Test	48 IUPAC, 2 Modifier	28,240

ELSEVIER

Chemical NER – biGRU-CRF Model



Chemical NER – F1 Scores

Word Embedding	F1	Δ (F1)
Mat2Vec W2V	26.89%	—
PubMed W2V	27.23%	+ 0.3%
Drug W2V	48.48%	+21.3%
CheMU W2V	53.24%	+ 4.8%
PubMed ELMo	70.15%	+16.9%
CheMU ELMo	72.41%	+ 2.3%

Intrinsic Evaluation – Similarity Queries

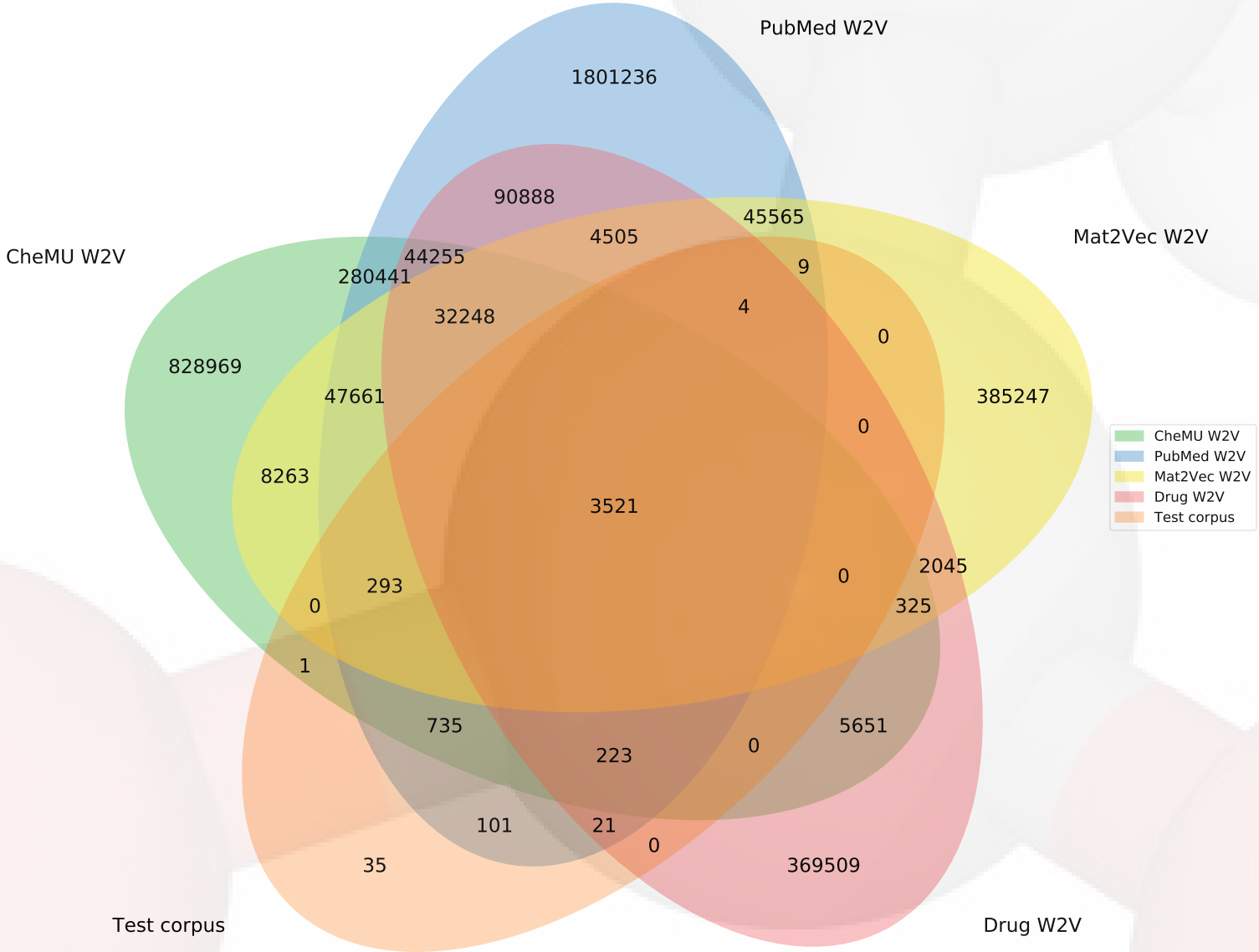
Task:
- ▷ compare most similar terms to a given chemical entity

Datasets:
- ▷ all embeddings
- ▷ test corpus (SCAI corpus)

Methods:
- ▷ restrict results to the test corpus (SCAI)
- ▷ compare and analyze similarity rankings
- ▷ compare rankings to molecular fingerprint similarity

ELSEVIER

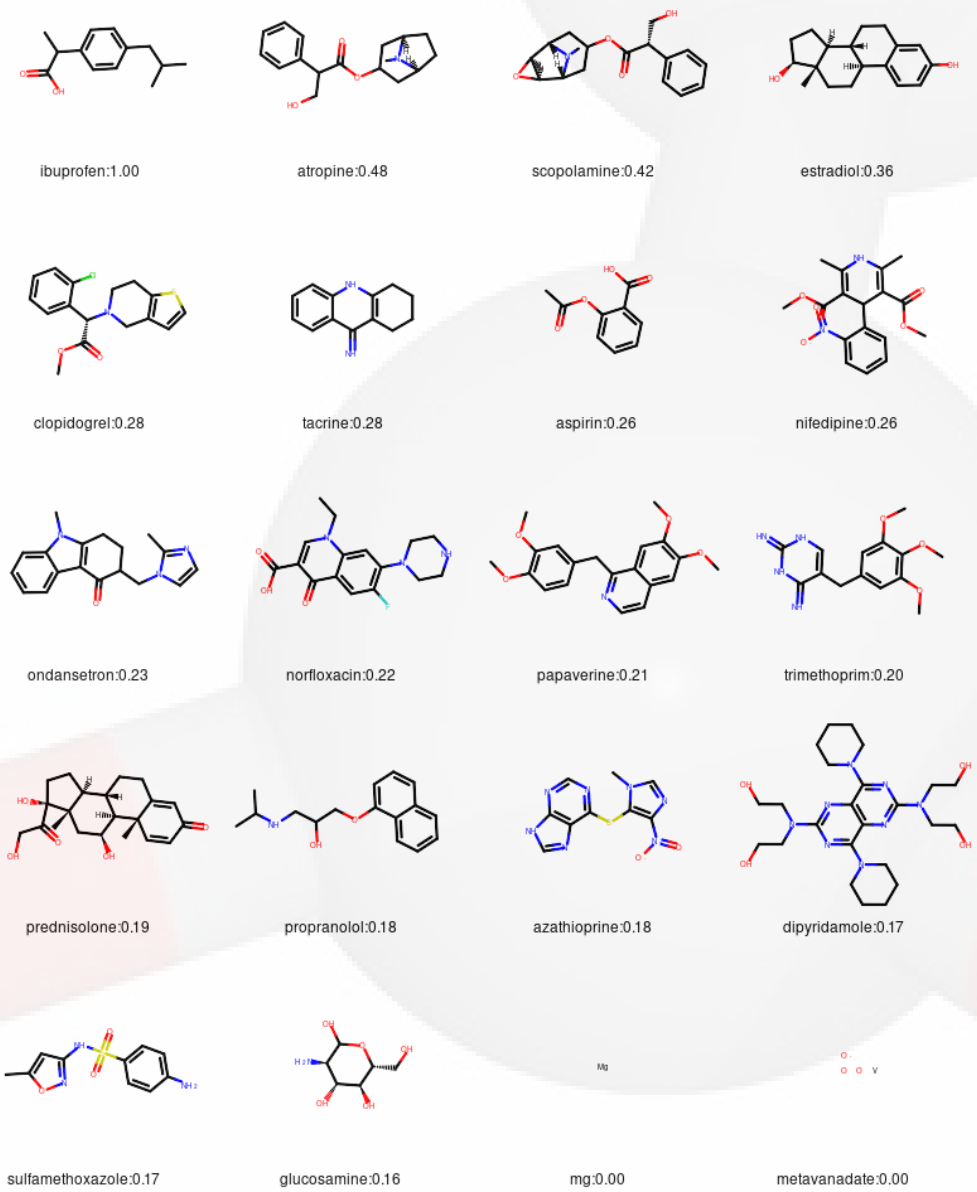
Similarity Queries – Vocabulary Overlaps



Similarity Queries – "Ibuprofen"

CheMU ELMo	PubMed ELMo	CheMU W2V	PubMed W2V	Drug W2V	Mat2Vec W2V
tacrine	atropine	aspirin	aspirin	pronounced	drug
ondansetron	ondansetron	clopidogrel	ondansetron	ultrastructure	drugs
aspirin	sulfamethoxazole	prednisolone	clopidogrel	mimics	aspirin
clopidogrel	aspirin	azathioprine	propranolol	surgical	sulfamethoxazole
dipyridamole	tacrine	atropine	placebo	favorable	propranolol
atropine	trimethoprim	nifedipine	tacrine	intestine	trimethoprim
prednisolone	propranolol	sulfamethoxazole	nifedipine	trained	norfloxacin
propranolol	prednisolone	dipyridamole	prednisolone	extinct	estradiol
trimethoprim	clopidogrel	propranolol	mg	slightly	antibiotics
nifedipine	papaverine	papaverine	topical	combination	nifedipine

Similarity Queries – Molecular Fingerprint Similarity



Takeaway

1. We studied the quality of embeddings trained over chemical patents against biomedical embeddings
2. Patent specific embeddings outperform larger-scale but generic embeddings on NER
3. They seem to provide a better understanding of the chemistry domain
4. **But:** Large scale generic embeddings provide OK performance

ELSEVIER

Thank you!

References I

[JDCL19] Qiao Jin, Bhuwan Dhingra, William W Cohen, and Xinghua Lu. Probing biomedical embeddings from language models. *CoRR*, abs/1904.02181, 2019.

[PGM+13] S. Pyysalo, F. Ginter, H. Moen, T. Salakoski, and S. Ananiadou. Distributional semantics resources for biomedical text processing. In *Proceedings of LBM 2013*, 2013.

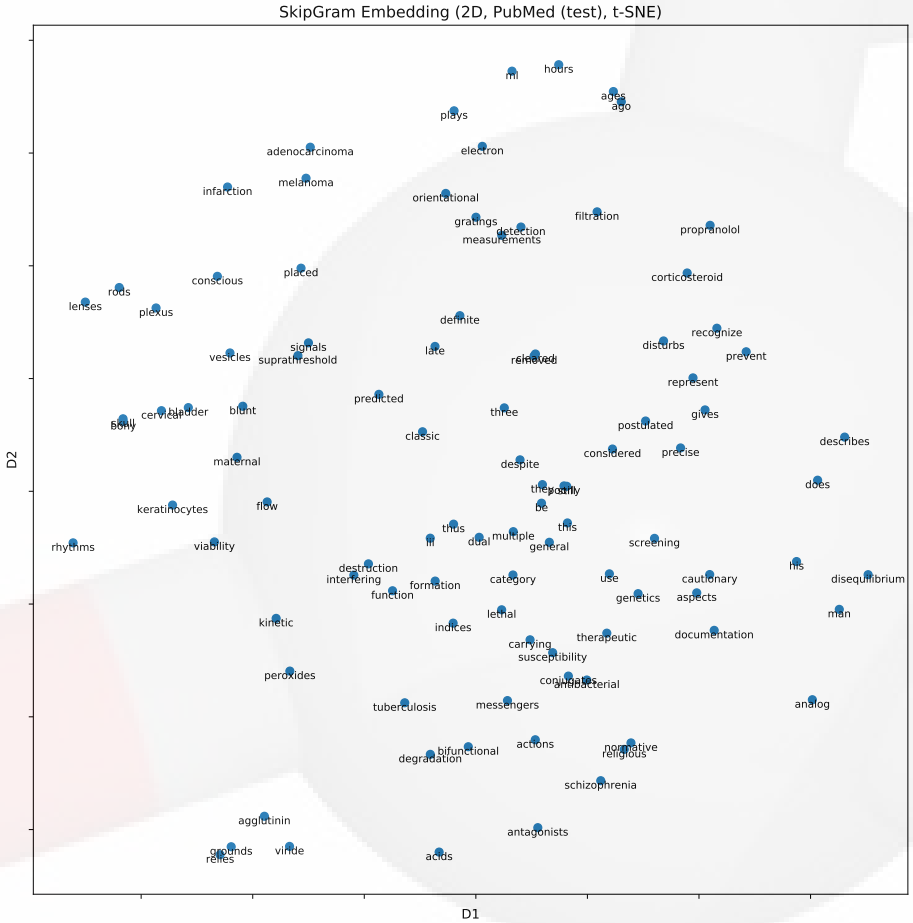
[SSM15] Isabel Segura-Bedmar, Víctor Suárez-Paniagua, and Paloma Martínez. Exploring word embedding for drug name recognition. In *Proceedings of the Sixth International Workshop on Health Text Mining and Information Analysis, Louhi@EMNLP 2015, Lisbon, Portugal, September 17, 2015*, pages 64–72, 2015.

References II

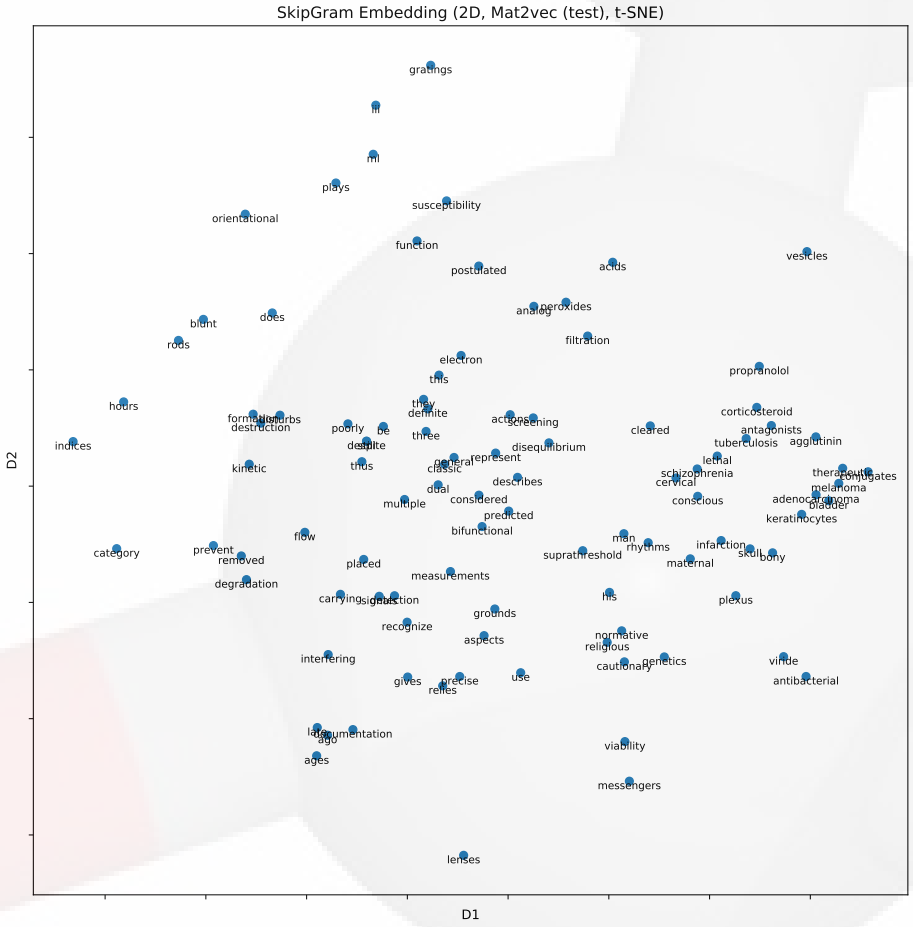
[TDW+19] Vahe Tshitoyan, John Dagdelen, Leigh Weston, Alexander Dunn, Ziqin Rong, Olga Kononova, Kristin A Persson, Gerbrand Ceder, and Anubhav Jain. Unsupervised word embeddings capture latent knowledge from materials science literature. *Nature*, 571(7763):95–98, 2019.

[ZNA+19] Zenan Zhai, Dat Quoc Nguyen, Saber Akhondi, Camilo Thorne, Christian Druckenbrodt, Trevor Cohn, Michelle Gregory, and Karin Verspoor. Improving Chemical Named Entity Recognition in Patents with Contextualized Word Embeddings. In *Proceedings of the 18th BioNLP Workshop*, pages 328–338, 2019.

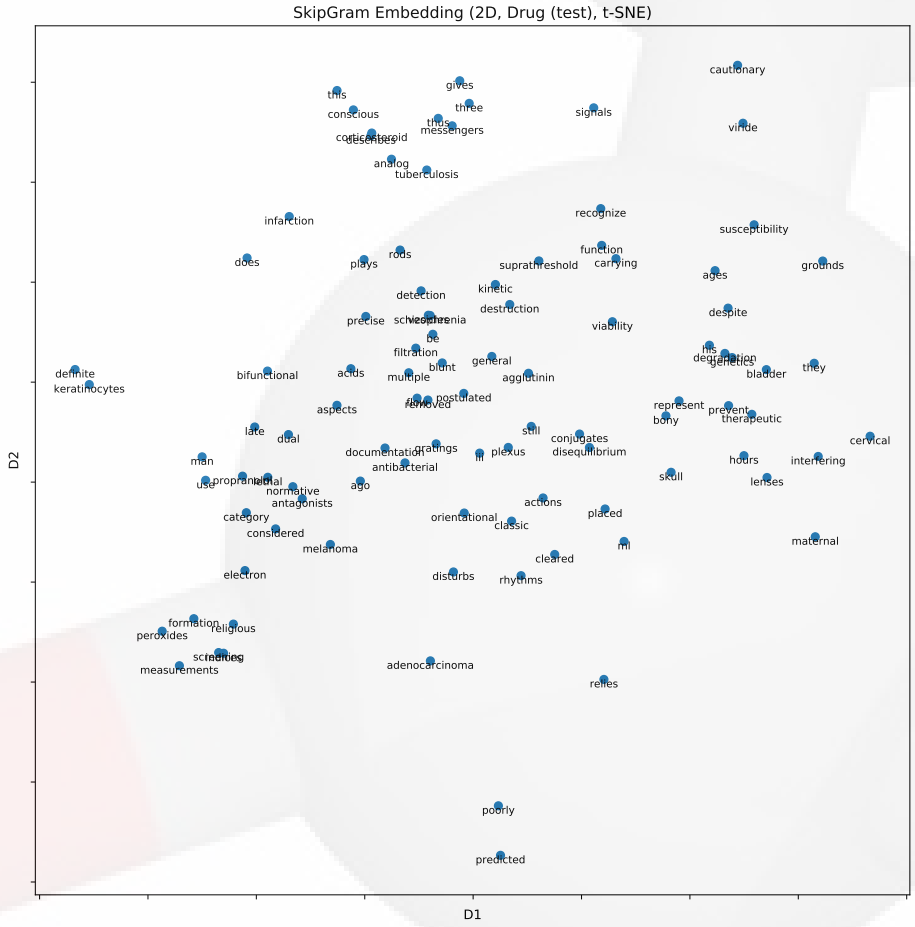
Appendix – Visualization I



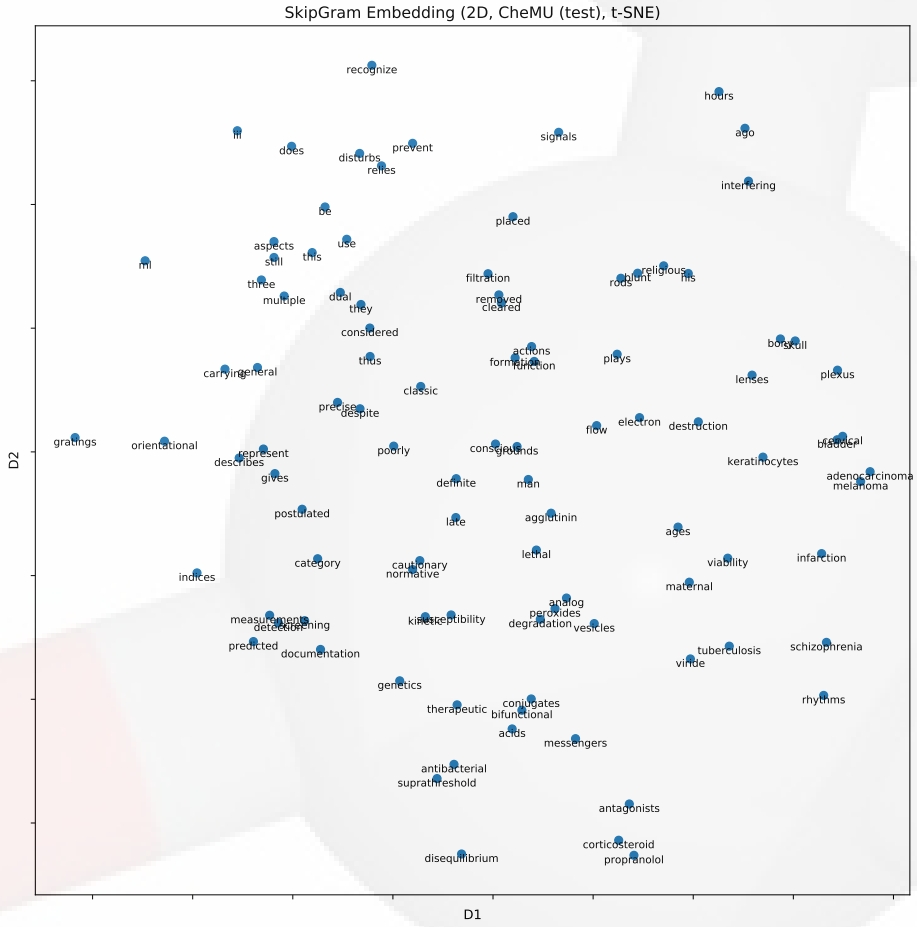
Appendix – Visualization II



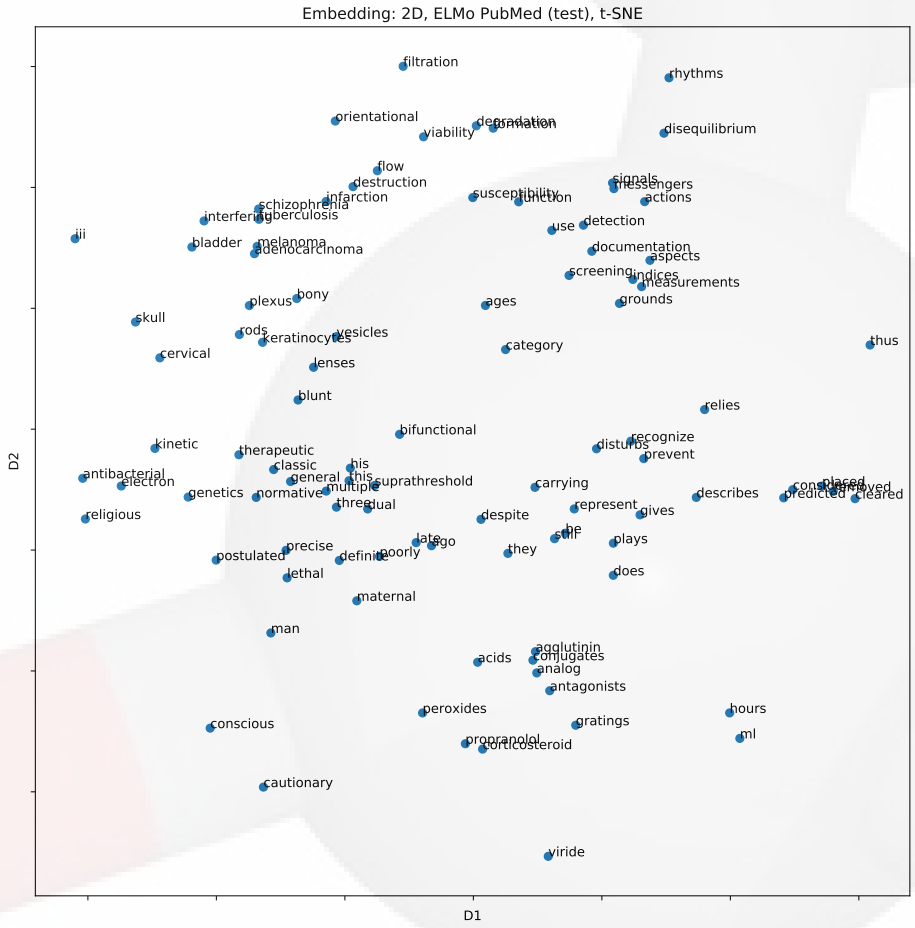
Appendix – Visualization III



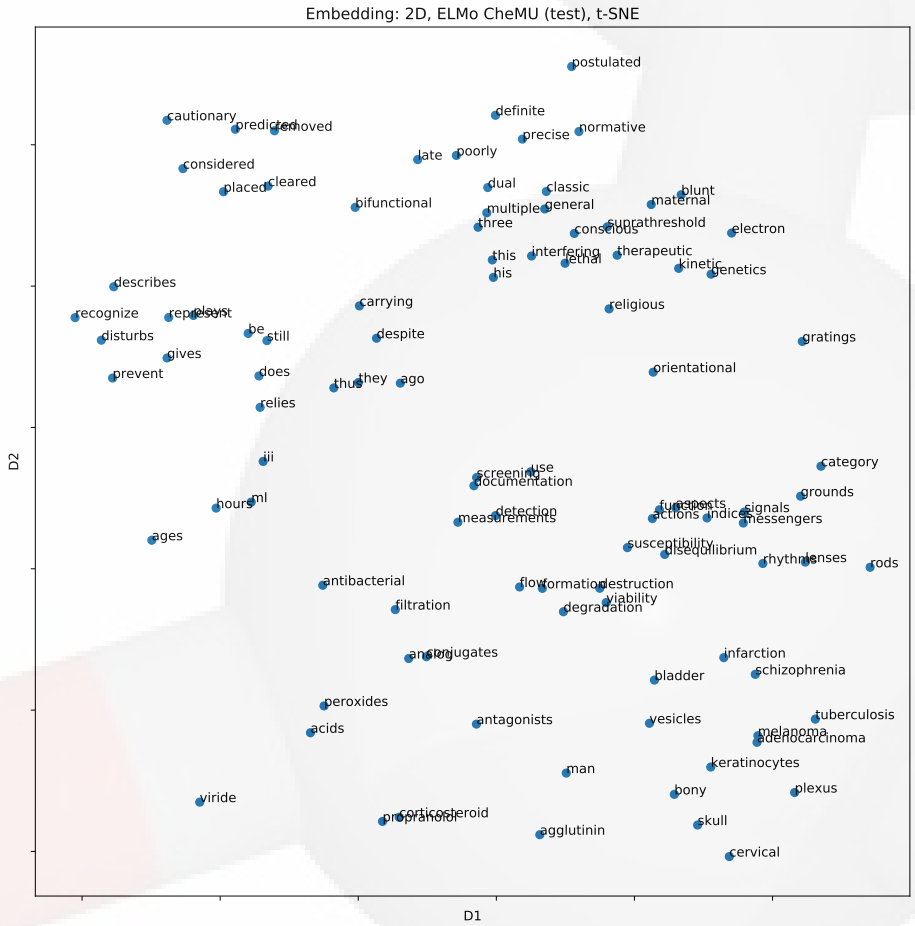
Appendix – Visualization IV



Appendix – Visualization V



Appendix – Visualization VI



Appendix – Similarity Overlap

	PubMed W2V	Drug W2V	Mat2Vec W2V	CheMU ELMo	PubMed ELMo
W2V CheMU	0.33	0.25	0.25	0.43	0.54
W2V PubMed	—	0.25	0.18	0.43	0.54
W2V Drug	—	—	0.18	0.25	0.18
W2V Mat2Vec	—	—	—	0.11	0.33
ELMo CheMU	—	—	—	—	0.54