Using Natural Language Processing (NLP) to Streamline Literature Selection for Meta-Analysis (MA)

MERCK
INVENTING FOR LIFE

Jenny Ding[1], Youfang Cao[2], Sean Hayes[1], Gregory Bryman[2], Kelly Yee[1]

[1]Quantative Pharmacology & Pharmacometrics, Merck & Co. Inc.

[2]Pharmacometrics, Eisai Co., Ltd.

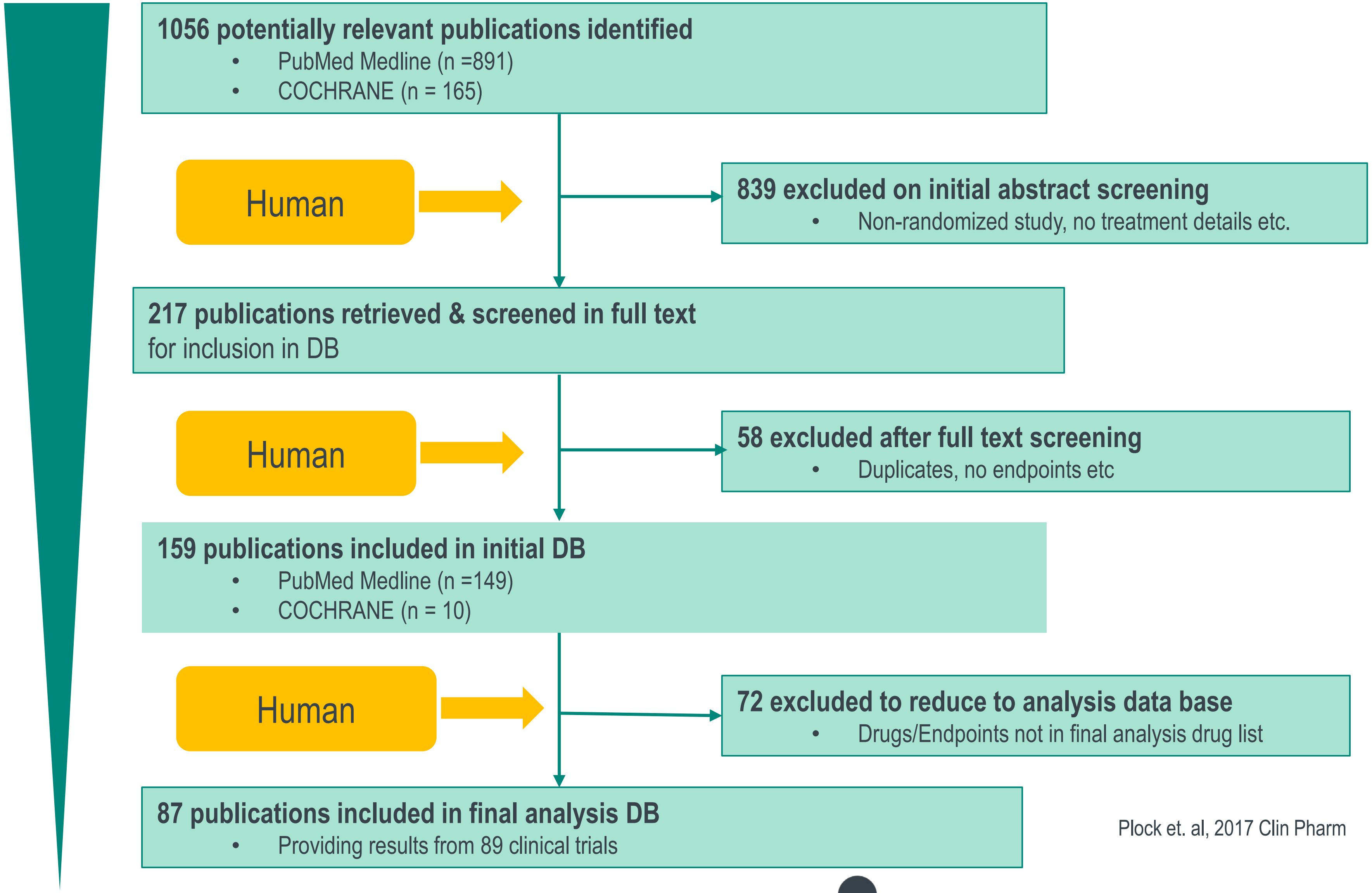
[3]Research & Development Sciences IT - Data Science & Scientific Informatics, Merck & Co. Inc.

IQ Consortium AI/ML Workshop

September 15, 2022

Meta-Analysis PRISMA Flowchart

- Meta-Analysis leverages published evidences to inform discovery and clinical decision making
- However, screening and selecting relevant literature from PubMed and other databases are resource/time-consuming



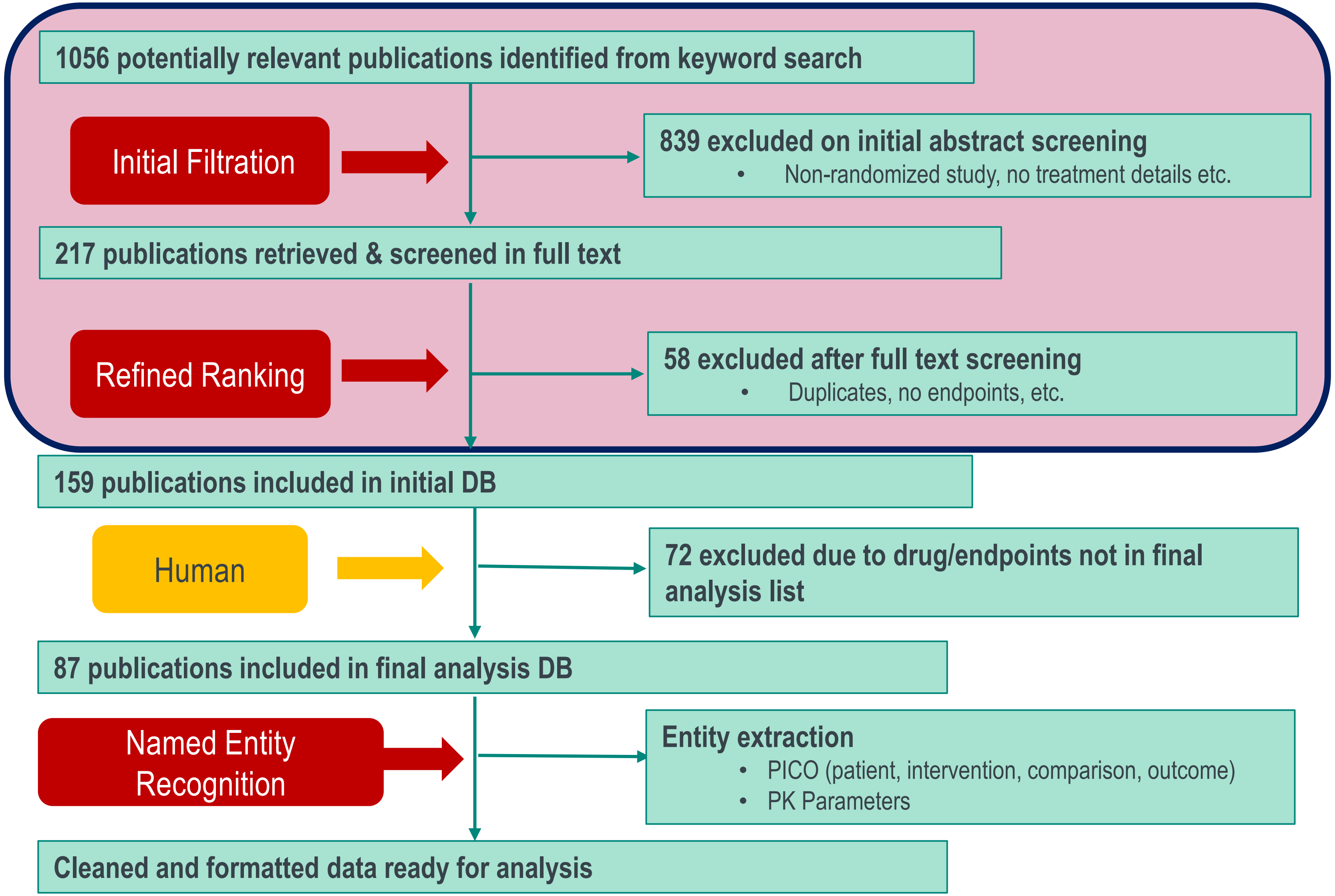
Plock et. al, 2017 Clin Pharm

PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses)

Natural Language Processing for Automated Literature Selection

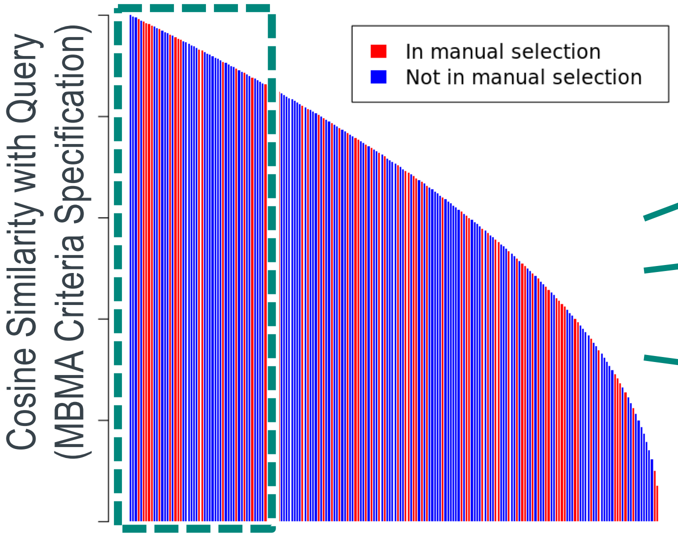
NLP Advantages

- A few **minutes** of run time vs months of manual curation
- **$5** computing cost vs 6 R3 FTE-months
- Streamlined process, less bias



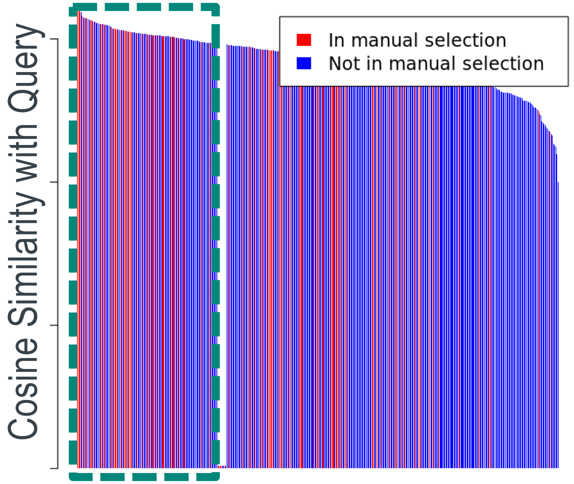
Exploration of NLP methods show Unsatisfactory Performance

Raw PubMed search: **32%** of abstracts selected with the cut are relevant

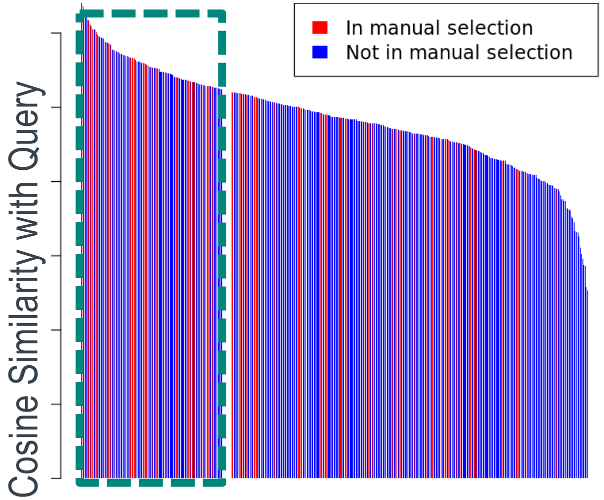


Total N=412 abstracts

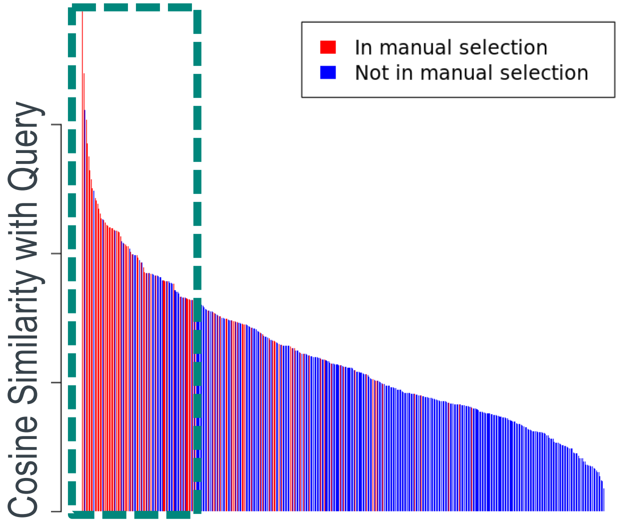
Ranked w/ Facebook **BioSentVec**: **38%** relevant abstracts



Ranked w/ Google **Universal Sentence Encoder**: **37%** relevant abstracts



Ranked w/ **TF/IDF** (term frequency–inverse document frequency): **55%** relevant abstracts



Improvements with domain-specific encoding

Use case is from a “NeuroPain” MBMA.

Proprietary

Transformer Models are Revolutionizing Biomedicine

Attention Is All You Need

2017: Transformers

Ashish Vaswani* Google Brain avaswani@google.com
Noam Shazeer* Google Brain noam@google.com
Niki Parmar* Google Research nikip@google.com
Jakob Uszkoreit* Google Research usz@google.com
Llion Jones* Google Research llion@google.com
Aidan N. Gomez* † University of Toronto aidan@cs.toronto.edu
Łukasz Kaiser* Google Brain lukaszkaiser@google.com
Illia Polosukhin* ‡ illia.polosukhin@gmail.com

BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding

2018: BERT

Jacob Devlin Ming-Wei Chang Kenton Lee Kristina Toutanova
Google AI Language
{jacobdevlin,mingweichang,kentonl,kristout}@google.com

Domain-Specific Language Model Pretraining for Biomedical Natural Language Processing

2020: PubMedBERT

YU GU*, ROBERT TINN*, HAO CHENG*, MICHAEL LUCAS, NAOTO USUYAMA, XIAODONG LIU, TRISTAN NAUMANN, JIANFENG GAO, and HOIFUNG POON, Microsoft Research

2021: AlphaFold (BERT-based model)
revolutionized protein 3D structure prediction

nature

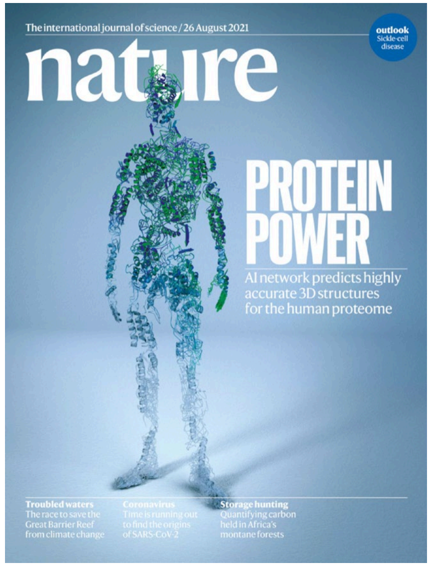
Search Login

Content Journal info Publish

NEWS | 30 November 2020

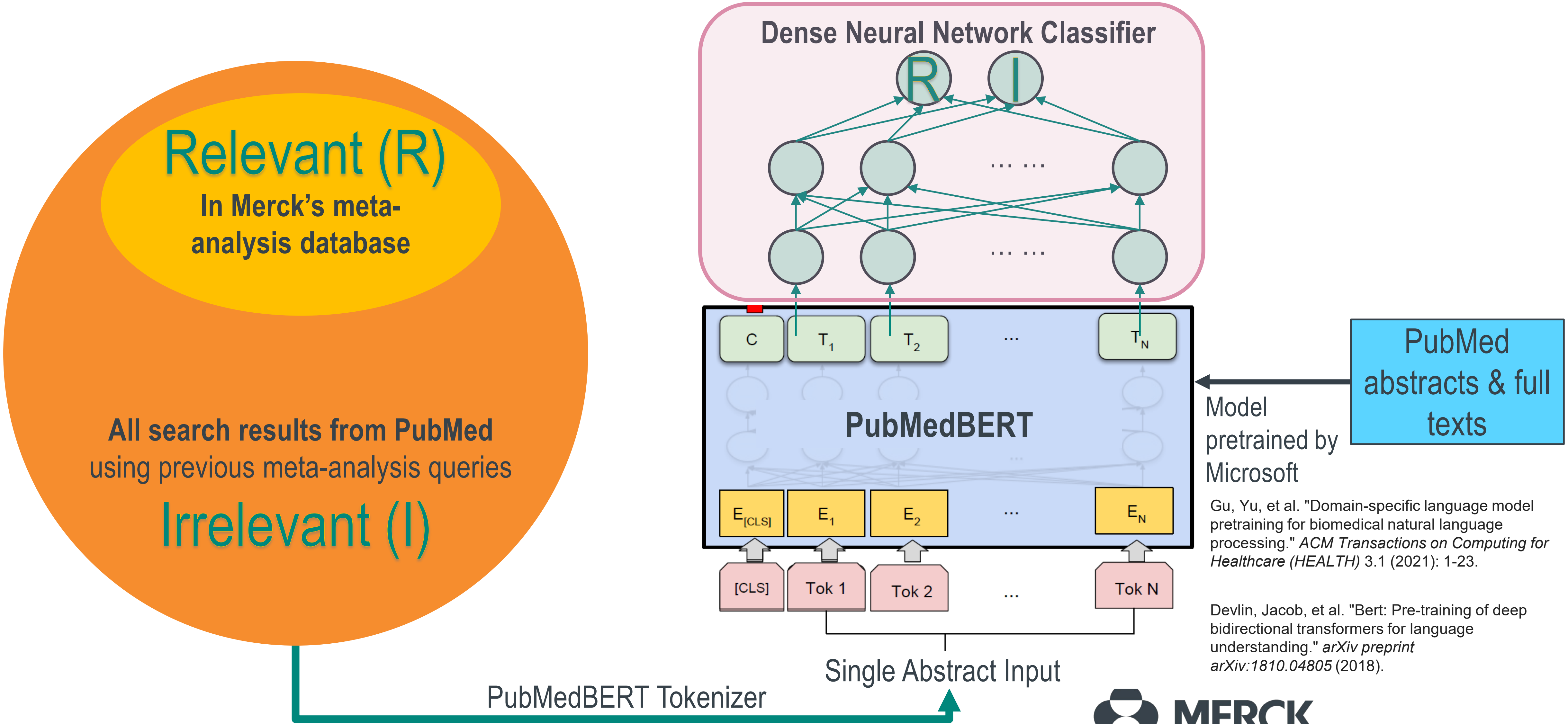
'It will change everything': DeepMind's AI makes gigantic leap in solving protein structures

Google's deep-learning program for determining the 3D shapes of proteins stands to transform biology, say scientists.



MERCK
INVENTING FOR LIFE

Transformer-based NLP Framework for MBMA Abstract Ranking



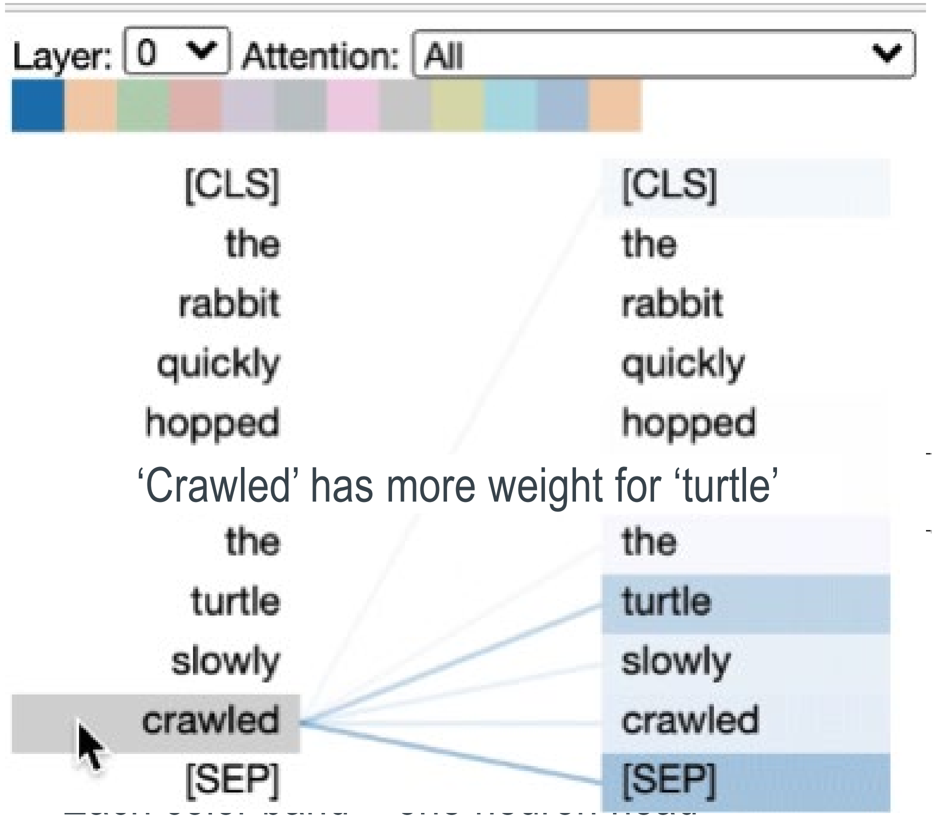
Gu, Yu, et al. "Domain-specific language model pretraining for biomedical natural language processing." *ACM Transactions on Computing for Healthcare (HEALTH)* 3.1 (2021): 1-23.

Devlin, Jacob, et al. "Bert: Pre-training of deep bidirectional transformers for language understanding." *arXiv preprint arXiv:1810.04805* (2018).

MERCK
INVENTING FOR LIFE

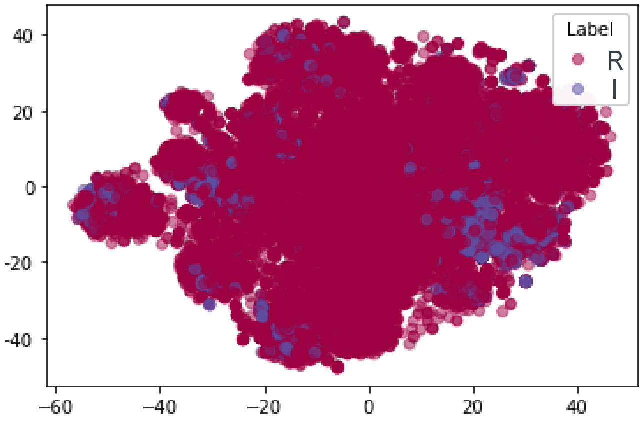
Transformers can understand Language Context

- Transformers (like BERT) have **attention** mechanisms that can learn **semantics** instead of only word frequency (TF-IDF), which is insufficient to capture long-term dependencies in sequences

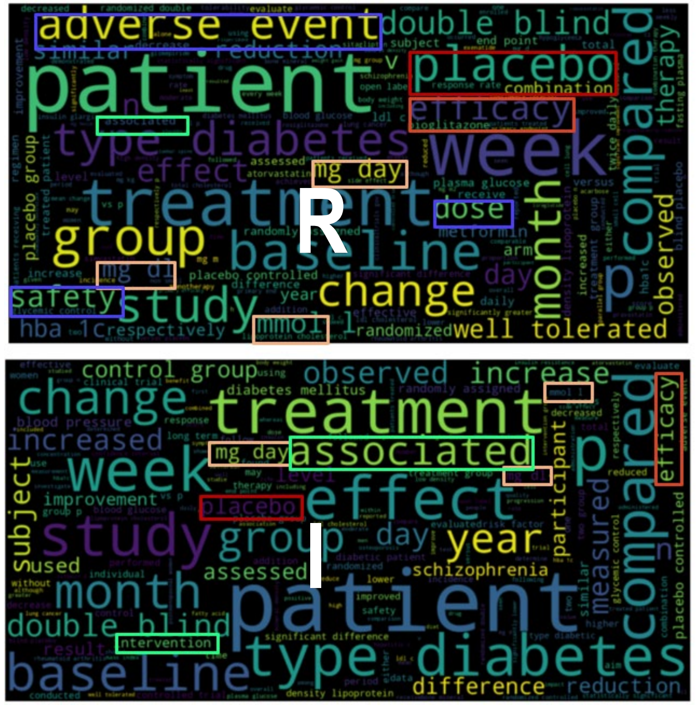


Intensity = magnitude, hue = Sign

Example of attention as shown from BertViz



T-SNE visualization of tokens selected (R) and not selected (I) shows high overlap.



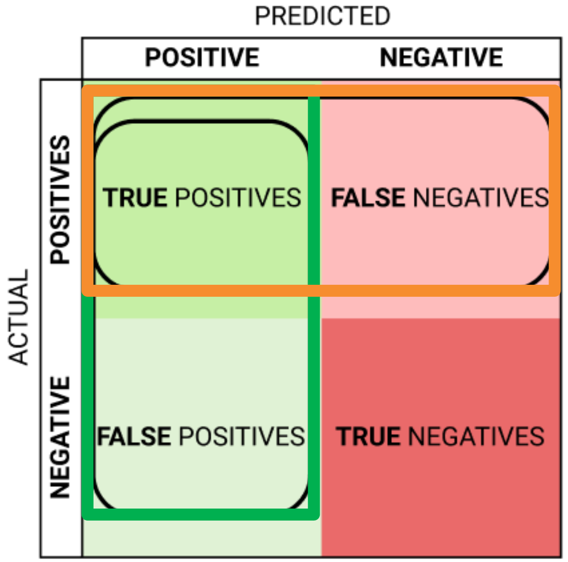
World cloud of abstracts selected (R) and not selected (I) for MBMA is hard to differentiate.

MERCK
INVENTING FOR LIFE

Generalization: PubMedBERT can Predict Diseases not in Training Dataset

	Disease	Total PubMed results	Human selected papers	Recall	Precision
Test data	Asthma	60	33	100%	82%
Train data	HCV	1343	164		
	Psoriasis	856	125		
	RA	2062	208		
	Neuro pain	412	99		
	T2 diabetes	8994	921		
	T1 diabetes	2319	148		
	Osteoporosis	2475	171		
	Schizophrenia	3161	239		
	NASH	1647	117		
	Lung cancer	1577	374		
	Dyslipidemia	3189	384		
	Grass pollen allergy	398	59		
	Endometriosis	486	117		
	Total/Mean	**28979**	**3159**		

- Task 1: Leave-1-disease-out cross validation
- Train a model on 13 diseases and test model on the left-out disease
  - E.g., train on HCV to Endometriosis, test on Asthma)



Recall: %Captured by Model out of All True Positives

Precision: %True Positives out of All Predicted to be Positives

Generalization: PubMedBERT can Predict Diseases not in Training Dataset

	Disease	Total PubMed results	Human selected papers	Recall	Precision
Train	Asthma	60	33		
Test	HCV	1343	164	93%	28%
Train	Psoriasis	856	125		
	RA	2062	208		
	Neuro pain	412	99		
	T2 diabetes	8994	921		
	T1 diabetes	2319	148		
	Osteoporosis	2475	171		
	Schizophrenia	3161	239		
	NASH	1647	117		
	Lung cancer	1577	374		
	Dyslipidemia	3189	384		
	Grass pollen allergy	398	59		
	Endometriosis	486	117		
	Total/Mean	**28979**	**3159**		

- Task 1: Leave-1-disease-out cross validation
- Train a model on 13 diseases and test model on the left-out disease
  - E.g., train on Asthma to Endometriosis EXCLUDING HCV, then test on HCV)
- Repeat (retrain 12 other models), so each disease has a chance to be the test set

MERCK
INVENTING FOR LIFE

Generalization: PubMedBERT can Predict Diseases not in Training Dataset

Disease	Total PubMed results	Human selected papers	Recall	Precision
Asthma	60	33	100%	82%
HCV	1343	164	93%	28%
Psoriasis	856	125	84%	31%
RA	2062	208	93%	24%
Neuro pain	412	99	78%	47%
T2 diabetes	8994	921	94%	24%
T1 diabetes	2319	148	96%	15%
Osteoporosis	2475	171	88%	15%
Schizophrenia	3161	239	92%	17%
NASH	1647	117	83%	15%
Lung cancer	1577	374	75%	45%
Dyslipidemia	3189	384	75%	23%
Grass pollen allergy	398	59	66%	25%
Endometriosis	486	117	74%	44%
Total/Mean	**28979**	**3159**	**85%**	**31%**

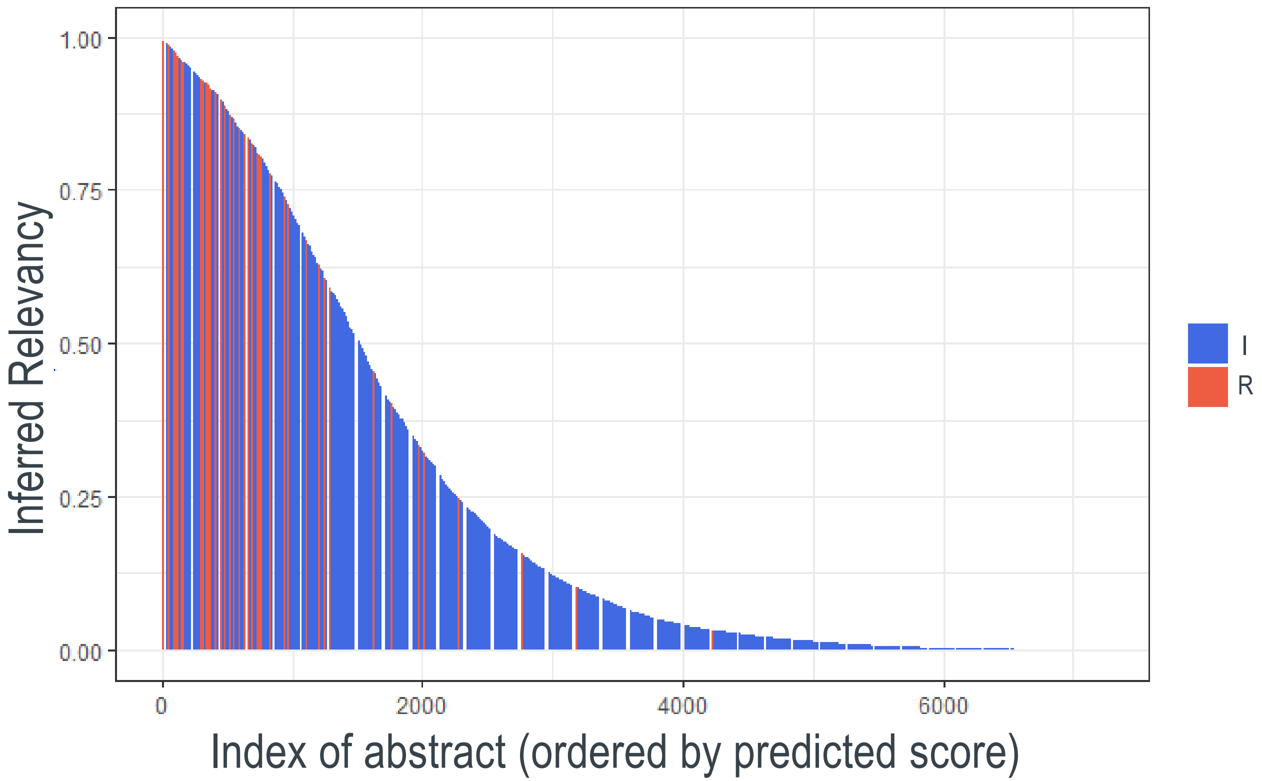
All leave-one-disease-out cross validation results

Smaller dataset show higher variability in outcomes

MERCK
INVENTING FOR LIFE

Generalization: Model trained on Historical Data can classify New Publications

- Task 2: Train a single14-disease model on previous 3-year data of each disease and test on most recent 3-year
  - E.g., train on 2002-2006 data for asthma and on 2003-2010 data for HCV and …(12 other diseases)
  - Then, test on 2007-2010 asthma and on 2011-2014 HCV abstracts and …(12 other diseases)



Result	Pred Pos	Pred Neg
True Pos (R)	5570	1096
True Neg (I)	127	422

Recall = 77%

Precision = 28%

Pilot on Endemic SARS-CoV-2 Show Promising Results

Initial pilot on endemic SARS-CoV-2 vaccine

- Both clinical and non-clinical
- Outcomes include viral replication/titer, antibodies, hospitalization, etc.
- Purpose is to build a model that can use animal data to predict vaccine success

Time

779 literature from keyword search → 346 literature selected by rule-based filter → 16 literature selected by Merck experts (Rate limiting step)

Comparison with NLP model

Result	Pred Pos	Pred Neg
True Pos (R)	14	2
True Neg (I)	60	703

Recall = 87.5%

Precision = 19%

- Used a **non-stringent cutoff** of 0.01 to catch more potentially relevant papers
- 74/799 abstracts selected, **90.5% reduction** versus 44.4% by quick scan
- Need to reduce **false negative ratio (12.5%)**

MERCK
INVENTING FOR LIFE

Next Step: Entity Extraction for more interpretable features and enhanced flexibility

- More **inclusive and streamlined** alternative to manual curation
- Example of an abstract **ranked highly by algorithm but missed/left out by manual selection**

Patient

Comparison

To compare the efficacy and safety of liraglutide versus sitagliptin as add-on to metformin after 26 weeks of treatment in Chinese patients with type 2 diabetes mellitus (T2DM). This 26-week open-label, active comparator trial (NCT02008682) randomized patients (aged 18-80 years) with T2DM inadequately controlled with metformin [glycated haemoglobin (HbA1c) 7.0-10.0% (53-86 mmol/mol)] 1 : 1 to once-daily subcutaneously administered liraglutide 1.8 mg (n = 184)…The primary endpoint was change in HbA1c from baseline to week 26. Liraglutide was superior to sitagliptin in reducing HbA1c from baseline [8.1% (65 mmol/mol)] to 26 weeks, as evidenced by estimated mean HbA1c change of -1.65% (-18.07 mmol/mol) versus -0.98% (-10.72 mmol/mol)…More patients receiving liraglutide (76.5%) than sitagliptin (52.6%) achieved the HbA1c target….

Outcome

Treatment

Conclusions

- **Summary:**
  - BERT-based NLP methods outperform traditional NLP methods (e.g., TF-IDF)
  - Potentially a cheaper and quicker alternative

- **Leveraged state-of-the-art biomedical-specific NLP model:**
  - Fine-tuned a neural network classifier on top of PubMedBERT model using internal MBMA data
- **In test set used to date, generalized to unseen disease and unseen (non-training set) abstracts**
  - 85% Recall in capturing top-ranked abstracts of unseen diseases
  - 77% Recall in predicting abstracts published downstream of training data

- **Future efforts:**
  - Conduct more pilot testing over different therapeutic areas
  - Reduce false negative ratio by further fine-tuning
  - Expand functionality based on literature curation needs; e.g., entity recognition