

- ・ AI : 新しい発見
- ・ 自動化 : 研究の効率化
- ・ 視覚化 : 直感的な意思決定



# AIによるMoA取得と 読解エンジンとしての可能性



2021/12/20

櫻井 祐樹 緑川 淳

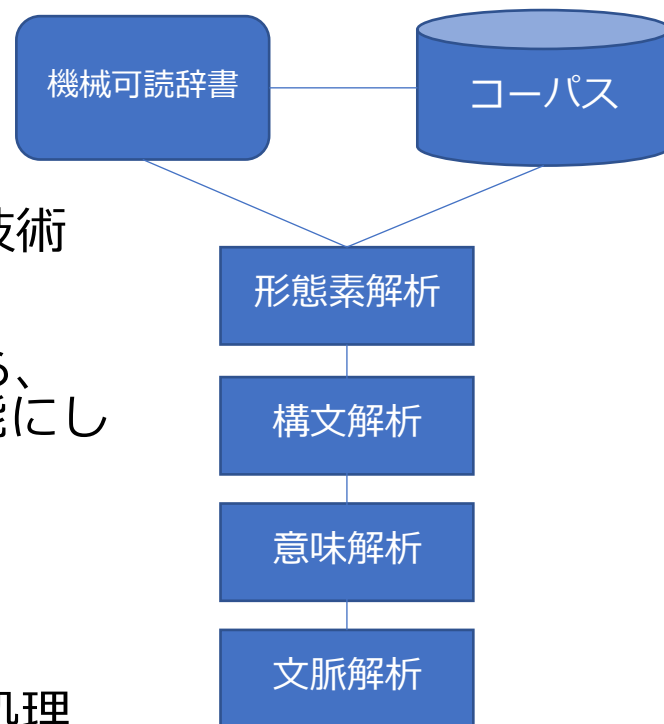
株式会社ワールドフュージョン

# AI (BERT) による 新しい MoA 抽出と 読解エンジンとしての活用の可能性

- 自然言語処理 と BERT
- BERTを使った新規MoA抽出プロジェクト実行環境
- BERTを使ったGeneRIFの遺伝子vs疾患の関係性の判別

# 自然言語処理 (Natural Language Processing)

- 自然言語処理 (Natural Language Processing) とは
  - 人間の言語 (自然言語) を機械で処理 (言葉が持つ意味を解析)
  - 曖昧性を含め、自然言語で書かれたテキストデータを処理
    - 話し言葉、書き言葉
  - 「かな変換」「Google 検索」「機械翻訳」など身近に利用されている技術
- Word2Vec
  - 言語を機械的に処理するためには、言語を単語や文字などの「記号」から、「数値ベクトル」に変換にする必要があり、Word2Vecはその変換を可能にした。
  - 従来の自然言語処理モデルでは、文章を単一方向からの処理に限定。(Word2Vec)
- BERT
  - BERT<sup>1)</sup>は 2018年10月 Jacob Devlin (Google) らが発表した自然言語処理モデル
    - BERTでは双方向のTransformerによる学習により 精度が向上。
  - スタンフォード大学の文章読解ベンチマークSQuAD 1.1<sup>2)</sup>での評価において、BERTは人工知能で初めて人間の平均の精度を超えた



1) <https://arxiv.org/abs/1810.04805>

2) <https://nlp.stanford.edu/pubs/rajpurkar2016squad.pdf>

# BERT (Bidirectional Encoder Representations from Transformers )

- BERTの特徴

- マスクされた言語モデル (Masked Language Model : MLM)
- 次文予測 (Next Sentence Prediction : NSP)
- 文脈から単語の意味を解釈できる。文章の語順を考慮できる (≠Word2Vec)
- 「Attention (注意機構)」の導入により離れた単語間の情報も適切に取り入れられる (≠RNN、LSTMベースモデル)

- BERTのタスク

- 固有表現抽出(Named Entity Recognition: NER)
  - テキストから一般名詞ではない 語句：固有表現(人名、組織名、etc...)を抽出
- MNLI
  - MultiNLIの略称で、自然言語推論のコーパスを意味している。元論文ではこのコーパスは約43万の前提と仮説のペアから構成される。
  - 前提と仮説の内容上の関係に応じて「含意 (entailment)」「矛盾 (contradiction)」「どちらとも言えない (neutral)」といったラベルが付けられる。

[https://ainow.ai/2019/05/21/167211/#Next\\_Sentence\\_PredictionNSP](https://ainow.ai/2019/05/21/167211/#Next_Sentence_PredictionNSP)

<https://buildersbox.corp->

[sansan.com/entry/2021/09/21/120000#%E5%9B%BA%E6%9C%89%E8%A1%A8%E7%8F%BE%E6%8A%BD%E5%87%BANER%E3%81%A8%E3%81%AF](https://buildersbox.corp-sansan.com/entry/2021/09/21/120000#%E5%9B%BA%E6%9C%89%E8%A1%A8%E7%8F%BE%E6%8A%BD%E5%87%BANER%E3%81%A8%E3%81%AF)

# BERTを使った新規MoA抽出プロジェクト

- プロジェクト概要
- BERT 用 事前準備
  - 訓練処理用テキストデータ
  - 推論処理用テキストデータ
  - 実行環境
- 結果
- 今後の課題

# プロジェクト概要

- 過去のCBI学会で発表した GO-MoAは、疾患に有意な生物学的機能を提示することで、新たなターゲットや薬剤を提案する有用な方法である。GO-MoAの拡張において新しい Mechanism of Action (MoA) を早期に取得することが課題であった。
- この課題を解決する上でAI技術を調査した結果、深層学習の転移学習モデルである、「BERT」による「固有表現抽出 (NER : Named Entity Recognition) タスク」にて検討することとした。
- 今回 BERTの汎用的な固有表現の学習済みモデルとして、特にライフサイエンス分野の利用目的に特化した「BioBERT」モデルを学習済みモデルとして採用した。

「BERT」の事前訓練データ : Wikipedia (2.5B num of words) ,BookCorpus(0.8B)

「BioBERT」 : 上記 + PubMed Abst (4.5B) ,PMC Full-text(13.5B)

# BERT用事前準備

## 訓練処理用テキストデータ

下記データをSQL処理で組み合わせ、出力した論文テキストデータに対してMoAタームをラベリングすることにより、教師有りの訓練データとした。

1. ChEMBL-DBのDRUG\_MECHANISMテーブル  
( MoA、Target、Chemicalの組み合わせ)
2. LSKB のターゲット（遺伝子）、MoA、化合物オントロジー
3. 「2」のオントロジータームによる直近20年のPubMed論文テキストマイニングデータ

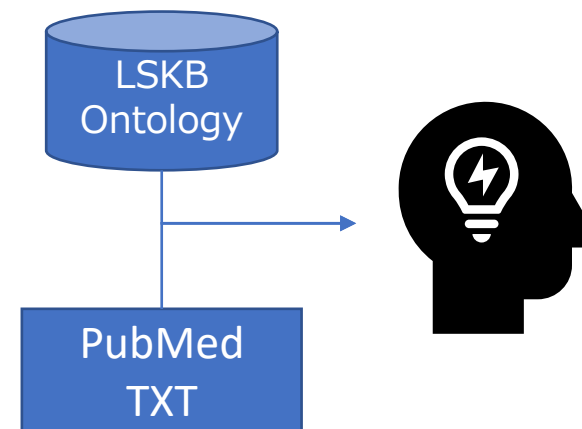
## 推論処理用テキストデータ

下記データをSQL処理で組み合わせ、出力した論文テキストデータを推論処理用データとした。

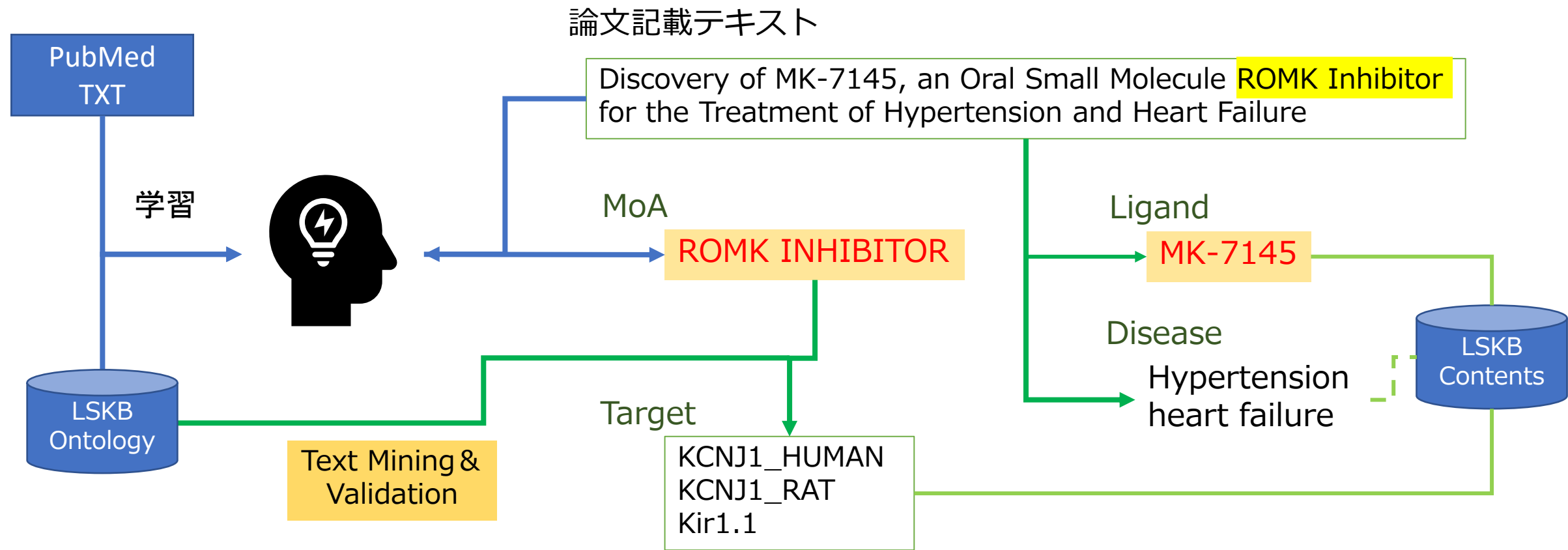
1. LSKB のターゲット（遺伝子）、MoA、化合物オントロジー
2. 「1」のオントロジータームによる直近20年のPubMed論文テキストマイニングデータ  
(具体的には「2」のマイニングデータの内、化合物タームとターゲットタームは共出現するが、MoAタームは検出されない論文データをテキスト化)

## 実行環境

- OS : Oracle Linux 7
- GPU : NVIDIA Tesla P100 16 GB
- Python およびGPU関連モジュール : OS提供Python : 3.6.8
- Tensorflow : 2.5.0



# 新規MoA 抽出と後処理プロセス概要

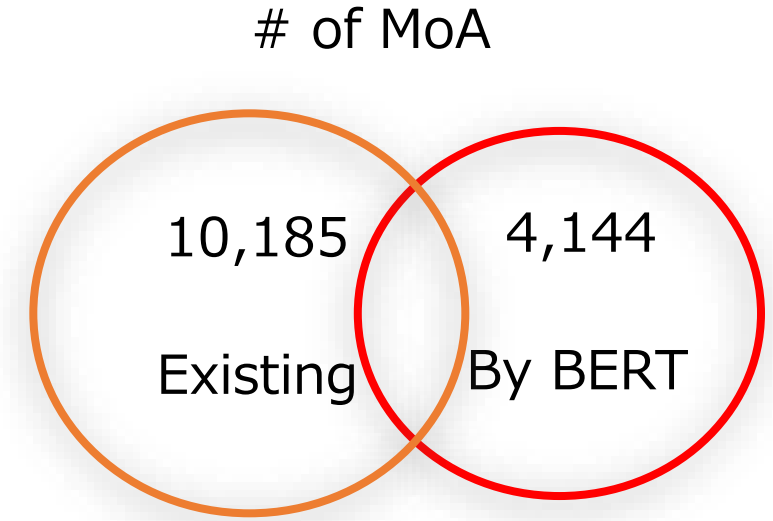


ラベル	Recall(再現率)	Precision(精度)	F1 Score
MoA	98.30%	95.30%	96.80%



# 結果

- Fine-tuningを実施したBERTモデルのNERタスクによるMoAターム抽出結果、新規MoAターム（候補）4,144件を取得した。
- MoAにアサインできたターゲットは888件であった。
- BERTの推論結果を利用して 抽出された MoAタームとエビデンスとして利用可能な論文情報と記載文章が取得できた。
- モデルの精度は次のとおり。



ラベル	Recall(再現率)	Precision(精度)	F1 Score
MoA	98.3%	95.3%	96.8%

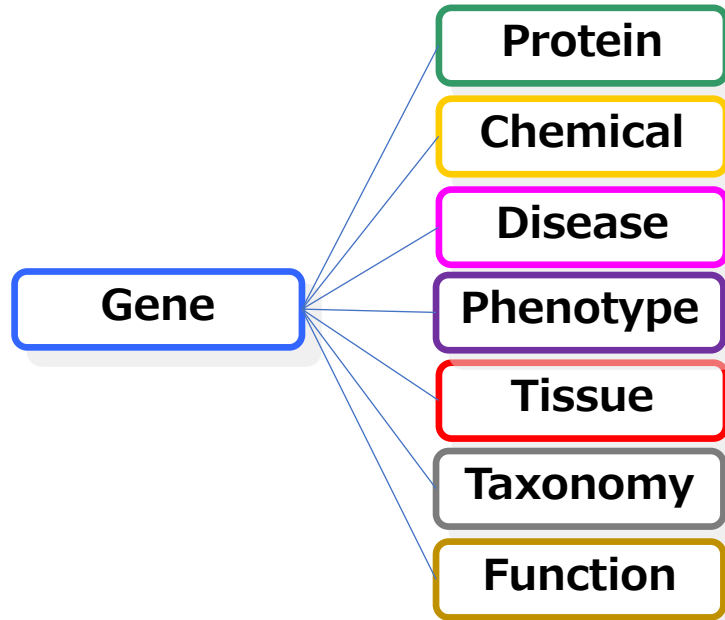
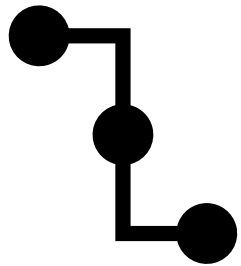
Recall(再現率)：テストデータに存在するとあるラベルのうちモデルが正しく判定した割合

Precision(精度)：とあるラベルと予測したうち実際にそのラベルが当たっていた割合

F1 Score：RecallとPrecisionの調和平均

# LSKB Interaction

Interaction



## 1) Text-Mining

- 20 years of PubMed literature (3 levels)
- Clinical Trial
- Assay Description

## 2) Assay Data

- Target Gene/Protein - Chemical
- Activities Endpoint
- Mode : Inhibition agonism/antagonism
- Expression: up/down regulation
- GWAS

## 3) Curated Annotation

- Disease Target
- Gene Ontology
- Pathway

## 4) AI Curated Annotation

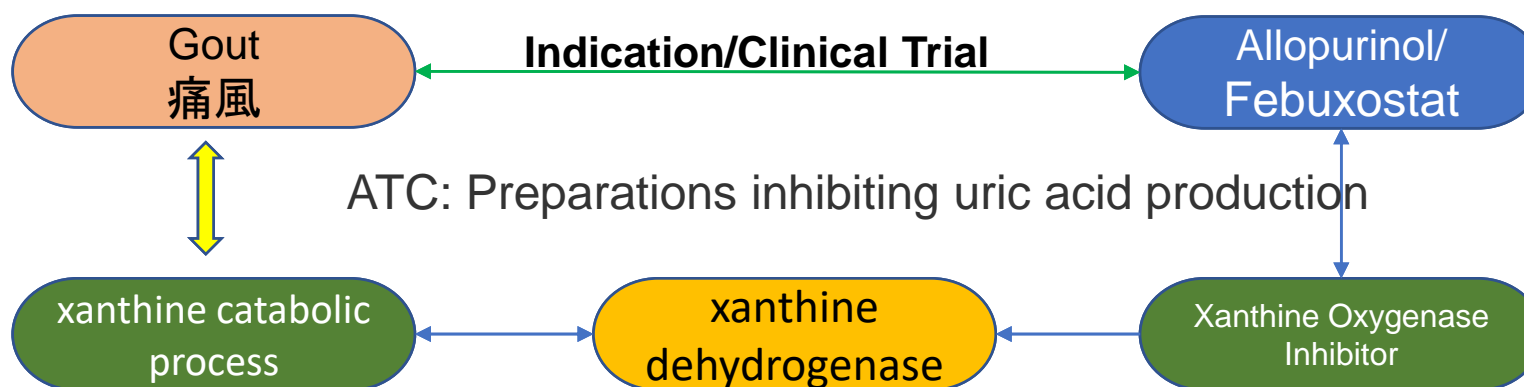
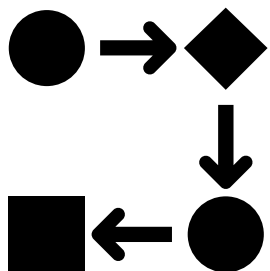
- Mechanism of Action

## 5) Misc

- GO-MoA database

**Disease related biological function database constructed by relationship of mechanism of action and the disease.**

### GO-MoA database



- **12,973 diseases related biological function (11,264 : v6.00 )**
- **99,610 combination of GO-MoA (32,212 : v6.00 )**
- **12,731 drugs (6,256 : v6.00 ) and 2,164 targets are included.**
- **2,042 considerable human targets**
  - **functionally similar but unknown target to the disease**
- **4,019 considerable drugs**
  - **for drug repurposing**
- **750k active compounds to the targets**
  - In case of p-value <0.0001

# 今後の展開

- 抽出した新規MoA は、ターゲット、化合物、関連疾患をアサインし検証のプロセスを行った。
- 結果をDB化するまでの過程において多くの属人的キュレーション作業が必要としている。今後はLSKB内の2次DBを利用すること等により、自動化をすすめる。
- 新規MoAを利用して、GO-MoAを更新した。
- 新規MoA の抽出課程については、新しいデータソースを訓練データに追加し、効率のよいモデル作成を検討する。

# BERT (MNLI) を使った GeneRIFの遺伝子vs疾患の関係性の判別

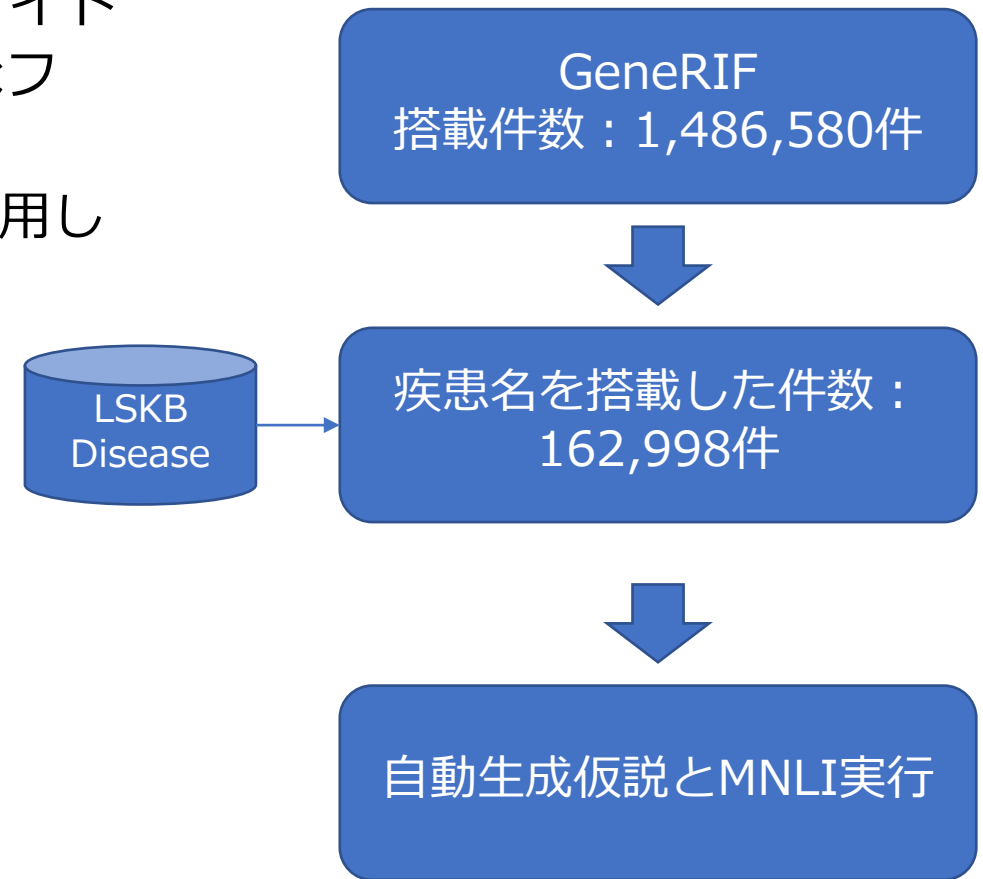
- プロジェクト概要
- GeneRIFとは
- 訓練処理と推論処理の実行環境と利用データ
- 実行結果・成果物
- 今後の課題・展望

# プロジェクト概要

- LSKBでは、疾患とターゲット遺伝子の関連情報を元に、疾患からターゲットを探す Target Explorer、ターゲットから関連疾患を探す Disease Explorerを搭載している。
- この有用なソースの一つにGeneRIFがあり、遺伝子に関するAuthorのコメントが記載されているが、その記載は Biomarkerや Variantにおける関連、疾患の原因など 多岐にわたっており、その関係性を判断できれば、更に有用性が向上する。
- この課題を解決する上でAI技術を調査した結果、深層学習の転移学習モデルである、「BERT」による「MNLIタスク」が有望であることが判明した。
  - MNLIでは文章と仮説を比較し、「含意 (entailment) 」 「矛盾 (contradiction) 」 「どちらとも言えない (neutral) 」 と判断
  - 比較に用いる仮説はGeneRIFに含まれるキーワードから 仮説を機械的に生成した。

# Gene Reference Into Function (GeneRIF)

- GeneRIFとは
  - NCBIが提供する 遺伝子機能のアノテーション登録サイト
  - PubMedで引用されるPMIDと タイトル以外の簡潔なフレーズ(425文字以下)で 1 つまたは複数の機能を説明
- GeneRIF記載文章から LSKB Disease オントロジーを利用したテキストマイニングで疾患名を取得
- GeneRIFの文章から次の**仮説を自動生成**
  - 遺伝子と疾患との関連
  - 遺伝子発現の変動と疾患との関連
    - Up/Down regulation
  - SNPが与える疾患への影響
  - 疾患のバイオマーカーとなり得る遺伝子情報



# Gene RIF から Diseaseの認識と 仮説の生成

GENE	GENERIF	DISEASE
slc41a1	Knockdown of slc41a1 expression in zebrafish resulted in ventral body curvature, hydrocephalus, and cystic kidneys, similar to the effects of knocking down other <b>nephronophthisis</b> genes.	nephronophthisis
	ゼブラフィッシュにおけるslc41a1発現のノックダウンは、他の <b>髄質性嚢胞腎</b> 遺伝子をノックダウンする効果と同様に、腹側体の湾曲、水頭症、および嚢胞性腎をもたらしました。	
ALDH1	loss of ALDH1A1 expression is suggested to promote <b>carcinogenesis</b> especially in the smoking-related lung adenocarcinomas	carcinogenesis
	ALDH1A1発現の喪失は、特に喫煙関連の肺腺癌において <b>発癌</b> を促進することが示唆されています。	

仮説（自動生成）



Negative expression of "slc41a1" associates with "nephronophthisis".

Gene RIF 文面

Knockdown of **slc41a1** expression in zebrafish resulted in ventral body curvature, hydrocephalus, and cystic kidneys, similar to the effects of knocking down other **nephronophthisis** genes

BERT MNLIにて判断



# BERT用事前準備

## 訓練処理用テキストデータ

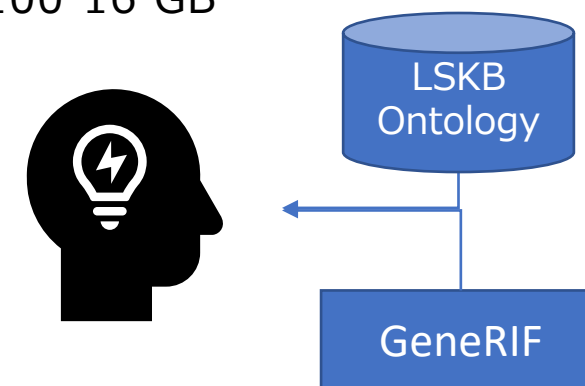
- TensorFlow Datasetsより、entailment, neutral, contradictionのラベル付けをされた、前提文と仮説の組み合わせレコード計391,829のテキストデータを教師有りの訓練データとした。

## 実行環境

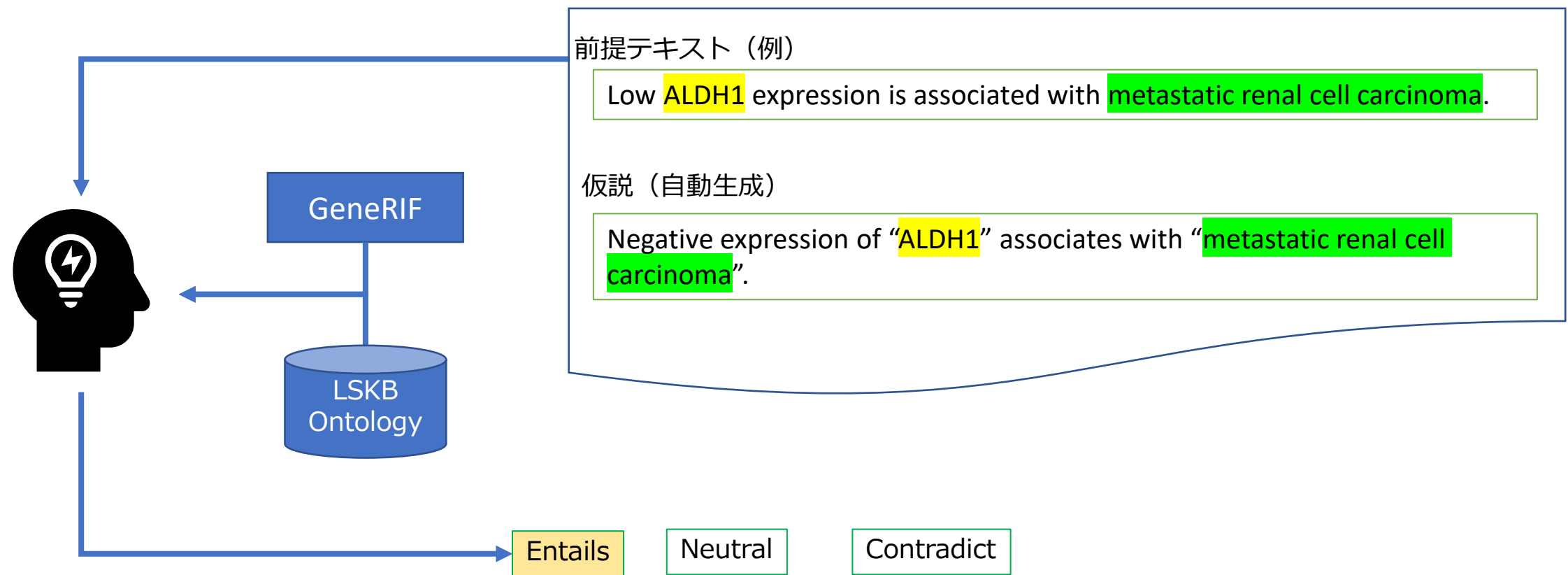
- OS : Oracle Linux 7
- Python およびGPU関連モジュール : OS提供Python : 3.6.8
- Tensorflow : 2.5.0

## 推論処理用テキストデータ

- GeneRIFのテキストデータにLSKBの疾患オントロジーでテキストマイニングを実施し、ヒットしたテキストデータを仮説検証の対象とした。
- 検証用の仮説はGeneRIFのテキストデータに対する特定用語（Activity, Enhance, Inhibitionなど）でのテキストマイニングにヒットした結果を用いて機械的に生成した。



# BERTによるGeneRIFの遺伝子vs疾患の関係性の判別概要



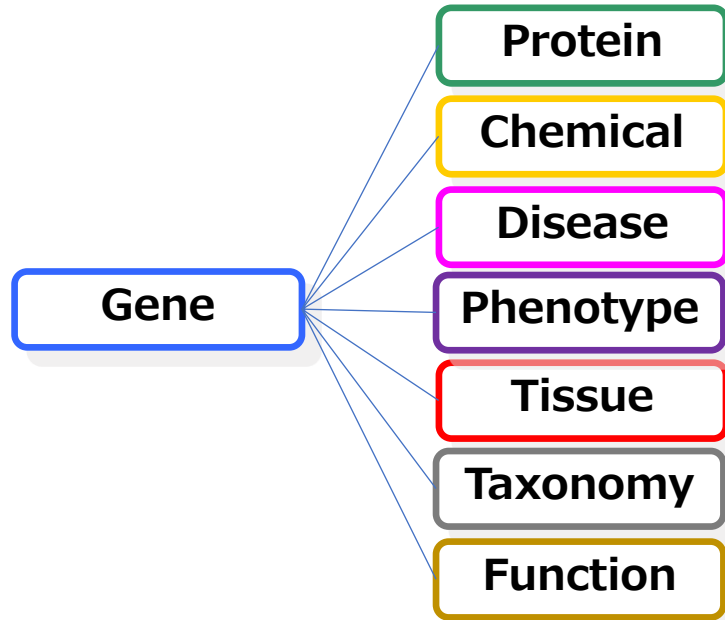
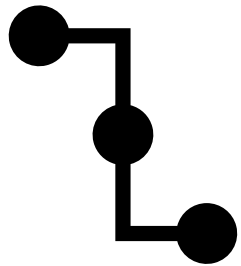
	Recall(再現率)	Precision(精度)	f1-score
contradiction	88.68%	75.03%	81.29%
entailment	83.11%	83.29%	83.20%
neutral	67.60%	82.54%	74.33%

# 結果

- GeneRIFの搭載文章 1,486,580件のうち、162,988件（Gene 19190件）をBERTのMNLI タスクを実行した。
  - 遺伝子の生物種の内訳は Human (10523)、Mouse (5323)、Rat (2671)、Zebrafish (273) であった。
- 9 種類の仮説を自動生成（計 1,499,079 件）した。その内訳は
  - “遺伝子と疾患との関連” 仮説で取得された 疾患は 10502件、遺伝子は18135件であった。
  - “遺伝子発現の負の変動と疾患との関連性” 仮説で取得された 疾患は 3714件、遺伝子は 5673件であった。
- 各仮説の判定結果（Entailment）をランダムに20-25件 前提文章と比較検証したところ、精度は 80-100%であった。
- 実際の利用場面（Entailment）の MNLIのタスクの精度は F1-Score : 0.832であった。

# LSKB Interaction

Interaction



## 1) Text-Mining

- 20 years of PubMed literature (3 levels)
- Clinical Trial
- Assay Description

## 2) Assay Data

- Target Gene/Protein - Chemical
- Activities Endpoint
- Mode : Inhibition agonism/antagonism
- Expression: up/down regulation
- GWAS

## 3) Curated Annotation

- Disease Target
- Gene Ontology
- Pathway

## 4) AI Curated Annotation

- Mechanism of Action
- Gene RIF

Disease

Huntington Disease

Rare

T047:Disease or Syndrome

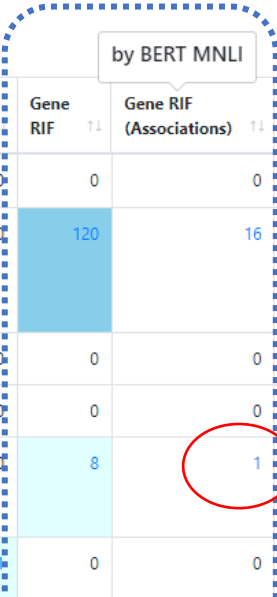
BioAssayOntology ExperimentalFactorOntology Guide to Pharmacology Medgen Orphanet UMLS UniProt

Link - - - - -

Filter Condition(s) Send Gene ID

Show 25 entries

	Gene ID	Gene Symbol	Gene Title	Organism	# of Literature	# of Literature LV1	# of Literature LV2	MedGen	OMIM	HPO	Gene Expression	Gene RIF	Gene RIF (Associations)	Protein (Assayed)	Bind	Inhibit
1	100303750	HD	Huntington disease	Homo sapiens	7246	1794	1373			0	0	0	0		0	0
2	3064	HTT	huntingtin	Homo sapiens	2948	1592	1461	C0020179 > MedGen	143100 > OMIM 613004 > OMIM	140		120	16	HD_HUMAN Huntingtin	0	15
3	4397	MS	multiple sclerosis	Homo sapiens	360	273	268			0	0	0	0		0	0
4	4908	NTF3	neurotrophin 3	Homo sapiens	243	39	38			0	0	0	0		0	0
5	627	BDNF	brain derived neurotrophic factor	Homo sapiens	226	70	70			0	0	8	1	BDNF_HUMAN Brain-derived neurotrophic factor	3	0
6	351	APP	amyloid beta precursor protein	Homo sapiens	61	10	10			0	1	0	0	A4_HUMAN Amyloid-beta	49	1122



LSKB About Contact

Gene Entry BDNF brain derived neurotrophic factor  
Disease Entry Huntington Disease

PMID	PubMed Article	GeneRIF	Hypo-1	Hypo-2	Hypo-3	Hypo-4	Hypo-5
22209255 > PubMed	Aberrant heart rate and brainstem brain-derived neurotrophic factor (BDNF) signaling in a mouse model of Huntington's disease.	A link was established between diminished BDNF expression in brainstem cardiovascular nuclei and abnormal heart rates in Huntington's disease mice.	*				

- negative expression of "BDNF" is the cause of "Huntington's disease"
- negative expression of "BRAIN DERIVED NEUROTROPHIC FACTOR" is the cause of "Huntington's disease"
- negative expression of "BRAIN-DERIVED NEUROTROPHIC FACTOR" is the cause of "Huntington's disease"
- negative expression of "BULN2" is the cause of "Huntington's disease"
- negative expression of "NEUROTROPHIN" is the cause of "Huntington's disease"

Found Synonyms
Huntington's Disease

# Huntington病の 関連遺伝子情報

# 今後の展開

- 仮説の種類を統合し、現在のGeneRIF に追加アノテーションとして搭載。
  - 遺伝子と疾患との関連性
  - 遺伝子の発現変動と疾患の影響
  - 遺伝子のSNPと疾患の関連性
- 検証プロセスの確立
  - 初期検討では ネイティブのスタッフと検証
- Fine-tuning用の教師あり学習データや 新しい方法など継続検討を行う。
- 他のデータへの拡張を検討する。



ご清聴ありがとうございました。

ご質問は  
[support@lskb.jp](mailto:support@lskb.jp)  
までお願いいたします。

<https://www.lskb.jp/>