

# LSKBのフロンティア BERT (MNLI) 文献解析とChatGPTでわかること

2023/10/26

株式会社ワールドフュージョン

# Introduction

## 1. 薬剤開発の現状

1. 薬剤開発はハイリスクハイリターン：成功すれば大きな収益、しかし成功率は低い
2. それに伴い、多様な創薬アプローチとモダリティの拡大が進行中。結果として、疾患への適応範囲も拡大している。

## 2. AI技術の導入の意義

1. 創薬の各段階、特に初期のターゲット探索やリード化合物の探索においてAIが効率化をもたらす。
2. 情報収集の効率化：膨大な文献やデータベースからの情報取得が必要。そのため、AI技術による文献の自動分析やデータのクラス分類が重要に。

## 3. BERT (MNLI) の役割

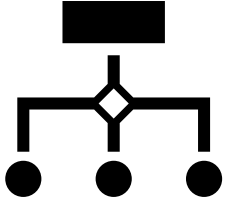
1. 文献の内容を効率よく分類・解析するためのツールとしてのBERTの導入。
2. LSKBにBERT (MNLI) を搭載し、より高度な情報処理が可能に。

## 4. ChatGPTの役割

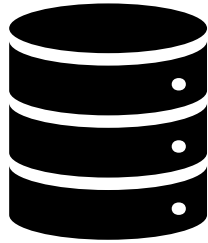
1. BERTの結果をさらに精緻化、または新しい視点からの解析を可能にするための補完的なツール。
2. 特に、文献の記述内容からの仮説生成や文献間の関連性の検証に活用。

# LSKB is scientific web

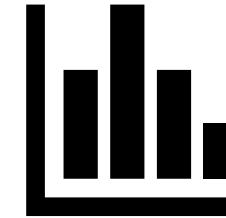
Ontology



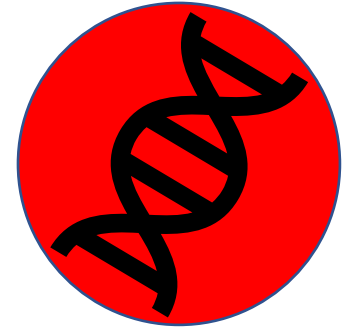
Contents



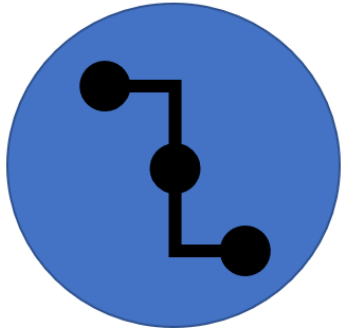
Analysis



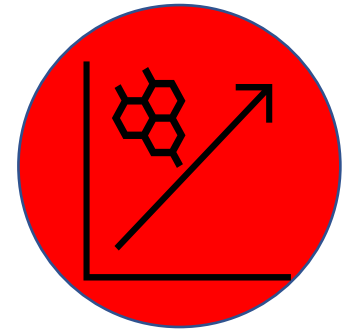
Bioinformatics



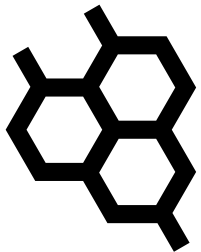
Interaction



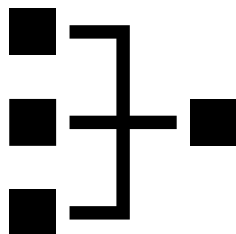
Cheminformatics



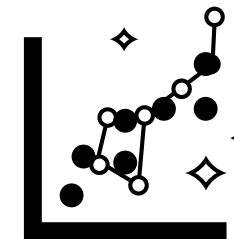
Structure



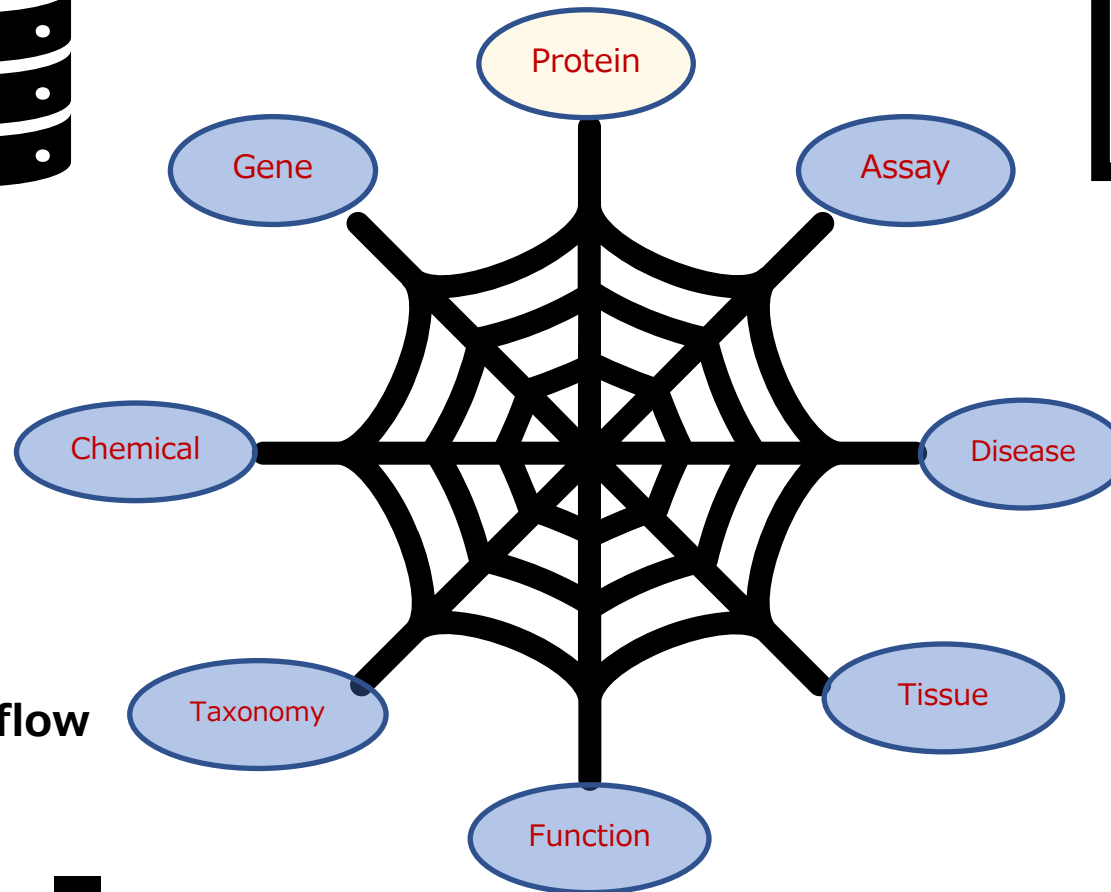
Workflow



Elpis Map

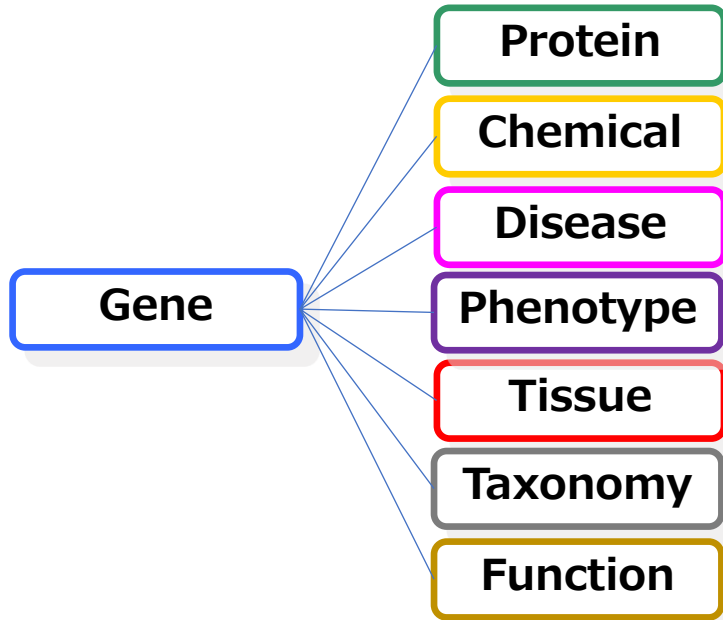
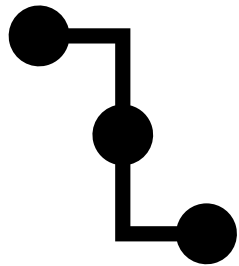


Exploration



# 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
- 20 years of PubMed literature
  - Disease & Toxicity associated Gene
  - Disease vs Phenotype

# BERTとMNLI

## • BERT

- BERT (Bidirectional Encoder Representations from Transformers) は、自然言語処理 (NLP) のための事前学習モデル。従来のモデルがテキストを片方向 (左から右またはその逆) だけで処理していたのに対し、BERTは両方向から情報を取得。これにより、文脈を考慮したより精密な単語の表現を得ることができる。

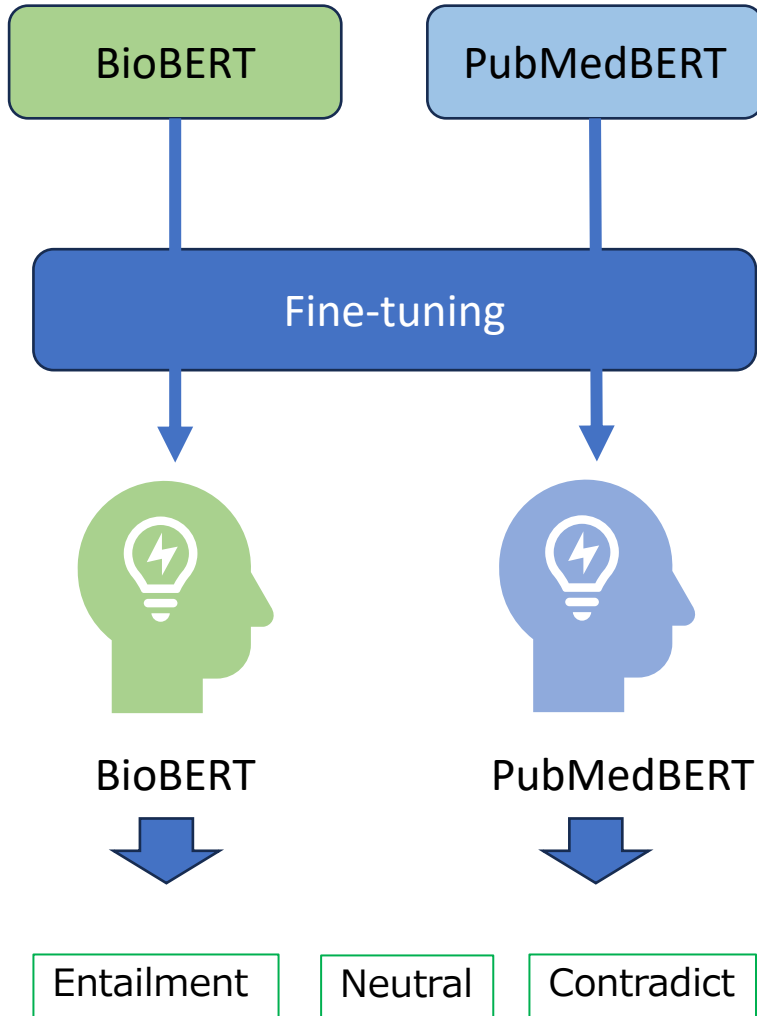
## • MNLI

- MNLI (MultiNLI) は、自然言語推論 (Natural Language Inference, NLI) のための大規模なデータセット。このデータセットは、文のペアとそれらの関係 (矛盾、中立、または包含) からなり、モデルの推論能力を評価するために使用される。

## • BioBERT & PubMedBERT

- BioBERTとPubMedBERTは、BERTの事前学習モデルのバリエーションで、特に生物医学の文献やテキストに特化しています。これらのモデルは、生物医学のコーパス (例: PubMed) で事前学習されており、生物医学分野のNLPタスクにおいて高いパフォーマンスを示す。

# BERT (MNLI) 実行プロセス



## データの前処理:

MNLIデータセットは、文章のペアとそれらの関係ラベル（矛盾、中立、または包含）から構成される。  
BERTやその派生モデルで、入力を受付ける処理可能な形式に前処理を行う。

## モデルのロード:

事前学習済みのモデル（例：BERTやBioBERT PubMedBERT）をロードする。

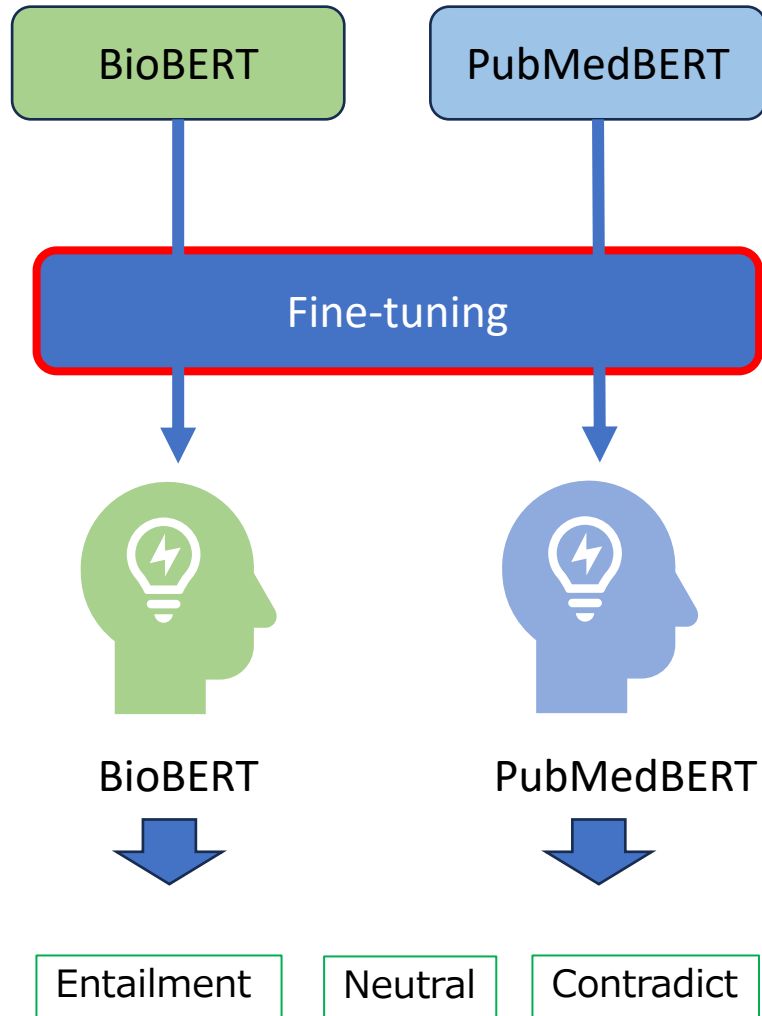
## ファインチューニング:

MNLIデータセットでモデルを訓練する。この際、モデルの出力層には3つのラベル（Contradiction：矛盾、Neutral：中立、Entailment：包含）に対応するクラス分類の層が追加され、この部分の重みも学習される。

## 評価:

MNLIデータセットの検証データやテストデータを用いて、ファインチューニングしたモデルの性能を評価。  
今回 サンプルングした例を **ChatGPT(GPT-4)**にて検証

# Fine-Tuning : Non-Domain Specific training dataset



## 1st Fine-tuning 用 トレーニングセット 例

Entailment	PREMISE:	How do you know? All this is their information again.
	Hypothesis:	This information belongs to them.
Neutral	PREMISE:	He turned and smiled at Vrenna.
	Hypothesis:	He smiled at Vrenna who was walking slowly behind him with her mother.
Contradiction	PREMISE:	but that takes too much planning
	Hypothesis:	It doesn't take much planning.

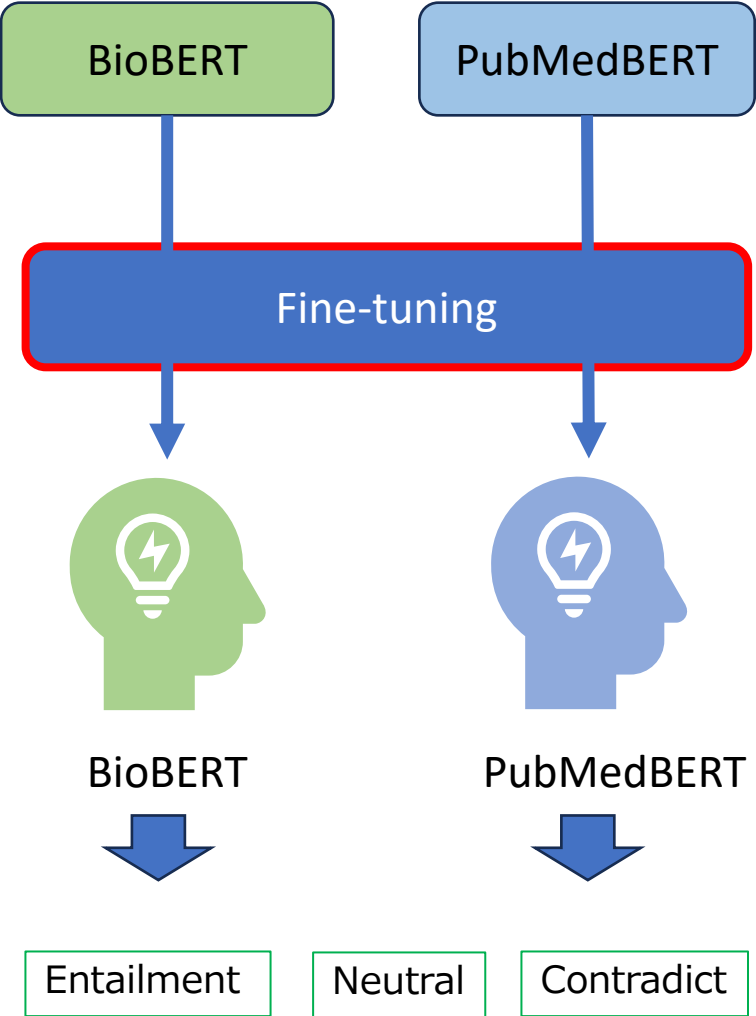
1st	Recall(再現率)	Precision(精度)	f1-score
contradiction	88.70%	75.00%	81.30%
entailment	83.10%	83.30%	83.20%
neutral	67.60%	82.50%	74.30%

BioBERT

1st	Recall(再現率)	Precision(精度)	f1-score
contradiction	85.15%	81.09%	83.07%
entailment	87.94%	79.33%	83.41%
neutral	70.07%	83.26%	76.10%

PubMedBERT

# Fine-Tuning : Domain Specific training dataset



2<sup>nd</sup> Fine-tuning 用 トレーニングセット 例

Entailment	PREMISE:	Regarding Asparagine degradation, It also has asparagine hydrolase activity.
	Hypothesis:	Asparagine hydrolase activity can be used to degrade Asparagine.
Neutral	PREMISE:	Regarding Sucralfate, Sucralfate is also used to treat mucositis.
	Hypothesis:	Marrow cells can be used to treat diseases.
Contradiction	PREMISE:	Regarding Oxymetholone, It is used mainly in the treatment of anemias.
	Hypothesis:	Obesity is a common symptom of Rubinstein-Taybi syndrome 2.

2 <sup>nd</sup>	Recall(再現率)	Precision(精度)	f1-score
contradiction	99.79%	98.83%	99.31%
entailment	99.15%	99.45%	99.30%
neutral	99.29%	99.99%	99.64%

BioBERT

2 <sup>nd</sup>	Recall(再現率)	Precision(精度)	f1-score
contradiction	99.79%	99.46%	99.63%
entailment	99.52%	99.58%	99.55%
neutral	99.71%	99.99%	99.85%

PubMedBERT



# PubMed Abstractから Disease Termの認識と仮説の生成



PMID	Abstract (抜粋)
23292288	<p>Androgen receptor (AR) knockout male mice display hepatic steatosis, suggesting that AR signaling may regulate hepatic fat. However, the effects of testosterone replacement on hepatic fat in men are unknown.</p> <p>アンドロゲン受容体 (AR) ノックアウト雄マウスは脂肪肝を示し、ARシグナル伝達が肝脂肪を調節している可能性があることを示唆しています。ただし、男性の肝脂肪に対するテストステロン補充の効果は不明です。</p>

Gene
Androgen receptor (AR)
Toxicity Term
hepatic steatosis

仮説 (自動生成)

テキストマイニングと組み合わせて  
10種以上の仮説を自動生成する

“AR” associates with “hepatic steatosis”.

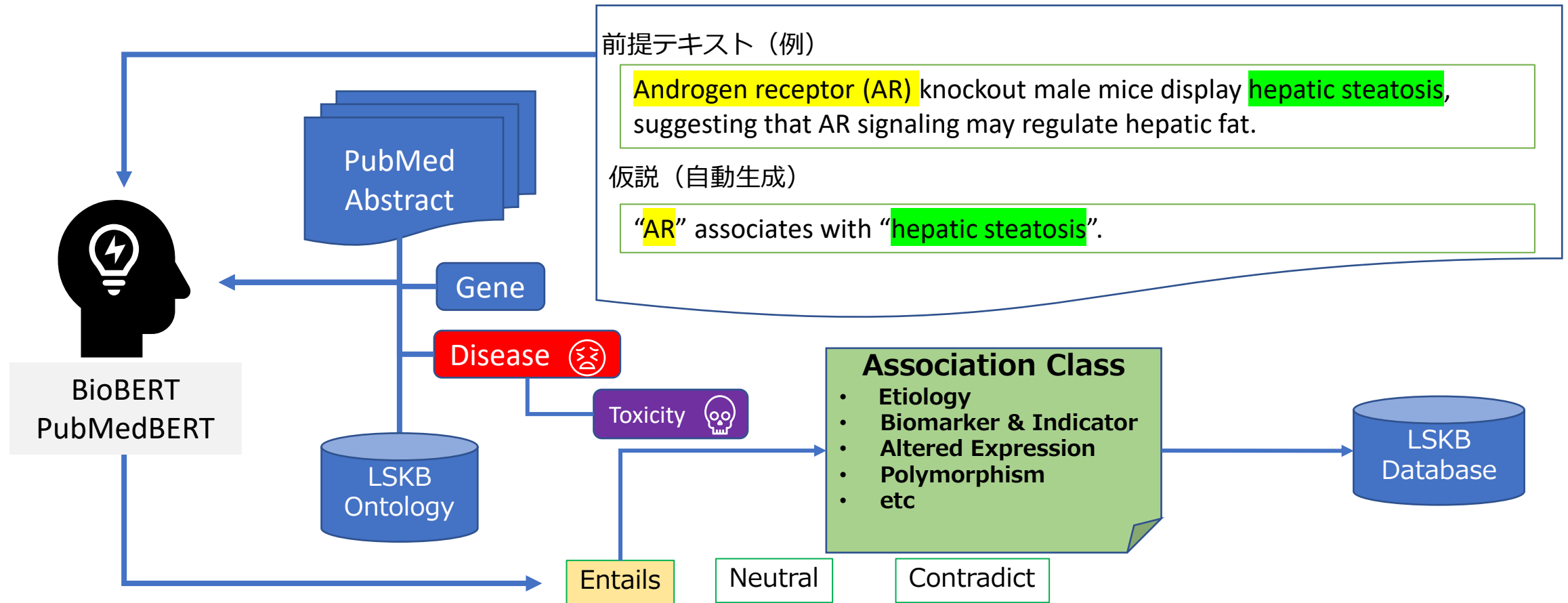
PubMed 文面

Androgen receptor (AR) knockout male mice display hepatic steatosis, suggesting that AR signaling may regulate hepatic fat. However, the effects of testosterone replacement on hepatic fat in men are unknown.

BERT MNLiにて判断



# BERTによる遺伝子vs疾患(毒性)の関係性の判別概要



2nd	Recall(再現率)	Precision(精度)	f1-score
contradiction	98.83%	99.79%	99.31%
entailment	99.45%	99.15%	99.30%
neutral	99.99%	99.29%	99.64%

# GeneRIF(Gene Reference Into Function)

GeneRIF  
搭載件数：1,589,973件



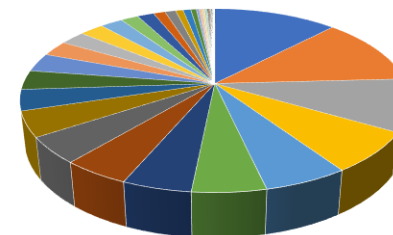
疾患 21,681との  
共起センテンス



自動生成仮説とMNLi実行

- GeneRIF
  - 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 搭載 108,378 遺伝子：ヒト (20,220 遺伝子)、マウス (14,305 遺伝子)、ラット (7,545 遺伝子)、植物 (9,781 遺伝子)
- 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

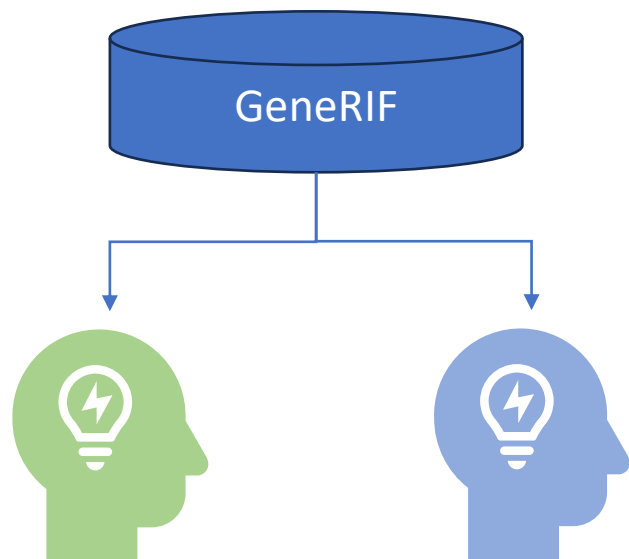
Disease Class in GeneRIF



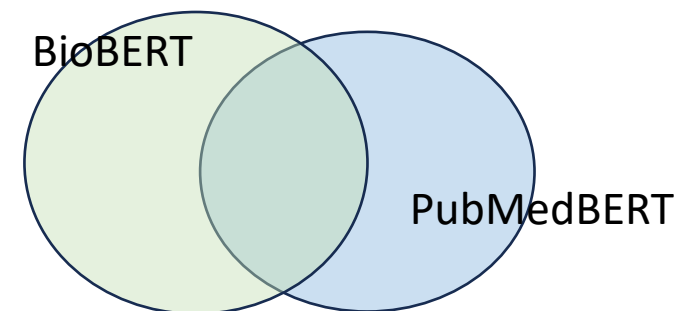
- |   |   |
|---|---|
| ■ Congenital, Hereditary, and Neonatal Diseases and Abnormalities | ■ Nervous System Diseases                     |
| ■ Neoplasms   | ■ Pathological Conditions, Signs and Symptoms |
| ■ Infections  | ■ Nutritional and Metabolic Diseases          |
| ■ Cardiovascular Diseases   | ■ Skin and Connective Tissue Diseases         |
| ■ Musculoskeletal Diseases  | ■ Urogenital Diseases                         |
| ■ Hemic and Lymphatic Diseases                                    | ■ Immune System Diseases                      |
| ■ Digestive System Diseases                                       | ■ Eye Diseases                                |

# GeneRIF : 遺伝子vs疾患のクラス分類

## 2種のBERT(MNLI) と2回のFine-tuning結果の比較



- Association Class**
- Etiology
  - Biomarker & Indicator
  - Altered Expression
  - Polymorphism
  - etc



### Etiology Class : BioBERT・PubMedBERT 及び 2回のFine-tuning 結果

BioBERT	PubMedBERT	# of Entailment				Accuracy	
		1st Model		2nd Model		1st Model	2nd Model
<b>Both Entailment</b>		264,798	36.1%	19,297	20.5%	98.0%	90.0%
Entailment BioBERT only		152,029	20.7%	51,277	54.4%	12.0%	88.0%
Entailment PubMedBERT_only		316,383	43.2%	23,628	25.1%	30.0%	74.0%
Total		733,210		94,202		46.7%	84.0%
Total (+Both Non-Entailment)		2,020,529		606,419			

※ Accuracy: 各50件をランダムに選択し、ChatGPTで評価し Entailmentと一致した割合

# 2<sup>nd</sup> BERT Model による 疾患関連遺伝子分類

# Glaucoma: associated genes from PubMed

LSIB About Contact

keyword: "Glaucoma"

**Disease** Glaucoma

T047:Disease or Syndrome

Filter Condition(s) Send Gene ID Copy Gene ID

Show 25 entries

Gene ID	Gene Symbol	Gene Title	Organism	Association Class	# of Literature	# of PubMed Abstract (Associations)	GeneRank	GeneRank (log)	# of GeneRIF	# of GeneRIF (Associations)	SNPs(rsID)	Protein (Assayed)	Binds	Inhibit	Confidence Score	
1	4653	MYOC	myocilin	Homo sapiens	<ul style="list-style-type: none"> <li>Altered Expression (GR)</li> <li>Biomarker &amp; indicator (GR)</li> <li>Enhancement (GR)</li> <li>Etiology (GR)</li> <li>Inhibition (GR)</li> <li>Polymorphism (GR)</li> <li>Altered Expression (PM)</li> <li>Biomarker &amp; indicator (PM)</li> <li>Enhancement (PM)</li> <li>Etiology (PM)</li> <li>Inhibition (PM)</li> <li>Polymorphism (PM)</li> </ul>	513	512		1	119	83	79	MYOC_HUMAN Myocilin	2	0	0.499
2	4016	LOXL1	lysyl oxidase like 1	Homo sapiens	<ul style="list-style-type: none"> <li>Etiology (GR)</li> <li>Polymorphism (GR)</li> <li>Altered Expression (PM)</li> <li>Biomarker &amp; indicator (PM)</li> <li>Etiology (PM)</li> <li>Polymorphism (PM)</li> </ul>	94	90	0.67	1	51	28	47	LOXL1_HUMAN Lysyl oxidase homolog 1	0	1	0.473
3	1545	CYP1B1	cytochrome P450 family 1 subfamily B member 1	Homo sapiens	<ul style="list-style-type: none"> <li>Altered Expression (GR)</li> <li>Biomarker &amp; indicator (GR)</li> <li>Enhancement (GR)</li> <li>Etiology (GR)</li> <li>Inhibition (GR)</li> <li>Polymorphism (GR)</li> <li>Altered Expression (PM)</li> <li>Biomarker &amp; indicator (PM)</li> <li>Etiology (PM)</li> <li>Inhibition (PM)</li> <li>Polymorphism (PM)</li> </ul>	185	202	0.68	0	97	74	25	CP1B1_HUMAN Cytochrome P450 1B1	13	416	0.465
4	10133	OPTN	optineurin	Homo sapiens	<ul style="list-style-type: none"> <li>Etiology (GR)</li> <li>Polymorphism (GR)</li> <li>Altered Expression (PM)</li> <li>Biomarker &amp; indicator (PM)</li> <li>Enhancement (PM)</li> <li>Etiology (PM)</li> </ul>	166	142	0.65	0							

Showing 1 to 25 of 7,300 entries

**Association Class**

- Etiology
- Biomarker & Indicator
- Altered Expression
- Polymorphism
- etc

V8 Etiology Human Gene: 950件

↓

2nd BERT Model Human Gene: 1282件

V8 Polymorphism Human Gene: 381件

↓

2nd BERT Model Human Gene: 277件

# Newly discovered gene in Etiology Class by 2<sup>nd</sup> Fine-tuning Model

PMID	25620203	
Gene	<b>DDX58</b>	
Description	<p>To identify genetic causes of the disease, we performed exome sequencing in this family and identified a variant (c.1118A&gt;C [p.Glu373Ala]) of <b>DDX58</b>, whose protein product is also known as RIG-I. Further analysis of <b>DDX58</b> in 100 individuals with congenital <b>glaucoma</b> identified another variant (c.803G&gt;T [p.Cys268Phe]) in a family who harbored neither dental anomalies nor aortic calcification but who suffered from <b>glaucoma</b> and skeletal abnormalities. Cys268 and Glu373 residues of DDX58 belong to ATP-binding motifs I and II, respectively, and these residues are predicted to be located closer to the ADP and RNA molecules than other nonpathogenic missense variants by protein structure analysis.</p>	
	<p>この疾患の遺伝的原因を特定するために、このファミリーでエクソーム配列決定を実行し、そのタンパク質産物が RIG-I としても知られる <b>DDX58</b> の変異体 (c.1118A&gt;C [p.Glu373Ala]) を特定しました。先天性<b>緑内障</b>患者 100 人を対象とした <b>DDX58</b> のさらなる分析により、歯の異常も大動脈石灰化も持たないが、<b>緑内障</b>と骨格の異常を患っていた家族に別の変異型 (c.803G&gt;T [p.Cys268Phe]) が特定されました。DDX58 の Cys268 残基と Glu373 残基は、それぞれ ATP 結合モチーフ I と II に属しており、タンパク質構造解析により、これらの残基は他の非病原性ミスセンス変異体よりも ADP 分子と RNA 分子の近くに位置すると予測されています。</p>	
Etiology(PM)	"DDX58" contributes to the occurrence of "glaucoma"	Entailment
Polymorphism (PM)	Mutations in "DDX58" are associated with "glaucoma"	Entailment

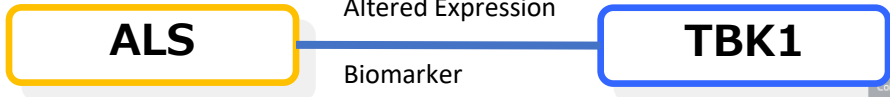
# MNLIによるクラス分類の詳細

LSKB About Contact logout

keyword: "als - amyotrophic lateral sclerosis"

**Disease** Amyotrophic Lateral Sclerosis Rare

T047:Disease or Syndrome



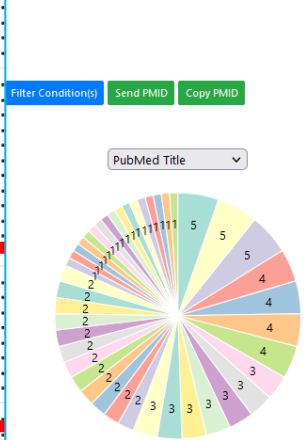
- Summary
- Annotation
- Protein [91]
- Classification [5]
- Drug [4836]
- GO Ranking [204]
- Gene [8025]
- SNPs(hg38) [1020]
- SNPs(hg19) [773]
- Gene Expression [55]
- Clinical Trial [318]
- GO-MoA

Filter Condition(s) Send Gene ID Copy Gene ID

Show 25 entries

Gene ID	Gene Symbol	Gene Title	Organism	Association Class	# of Literature	# of PubMed Abstract (Associations)	Gene Check (Ratio)	# of Homonyms	Pathway	HPO	UniProt Association	Max Phase Drug/Exp	Gene Expression	Gene Expression (Ortholog)	# of GeneRIF	# of GeneRIF (Associations)	SNPs(rsID)	Protein (Assayed)
1	23435	TARDBP	TAR DNA binding protein	Homo sapiens	<ul style="list-style-type: none"> <li>Altered Expression (GR)</li> <li>Biomarker &amp; indicator (GR)</li> <li>Etiology (GR)</li> <li>Inhibition (GR)</li> <li>Polymorphism (GR)</li> <li>Altered Expression (PM)</li> <li>Biomarker &amp; indicator (PM)</li> </ul>	884	1654	0.37	0	0	63	0	1	0	206	65	22	TAR DNA-binding protein 43

2	6647	SOD1	superoxide dismutase 1	Homo sapiens														
3	29110	TBK1	TANK binding kinase 1	Homo sapiens														
4	10233	OPTN	optineurin	Homo sapiens														
5	2521	FUS	FUS RNA binding protein	Homo sapiens														



**Gene** TBK1:TANK binding kinase 1

**Disease** Amyotrophic Lateral Sclerosis

Show 25 entries

PMID	PubMed Title	PubMed Abstract Sentence Number	Extracted PubMed Abstract Texts	Gene Check	Classification	Found Synonyms (1)	Found S
1	25700176	3	We performed whole-exome sequencing of 2869 ALS patients and 6405 controls. Several known ALS genes were found to be associated, and TBK1 (the gene encoding TANK-binding kinase 1) was identified as an ALS gene. TBK1 is known to bind to and phosphorylate a number of proteins involved in innate immunity and autophagy, including optineurin (OPTN) and p62 (SQSTM1/sequestosome), both of which have also been implicated in ALS.	YES	<ul style="list-style-type: none"> <li>Biomarker &amp; indicator</li> <li>Etiology</li> </ul>	<ul style="list-style-type: none"> <li>ALS</li> <li>TANK-binding kinase 1</li> <li>TBK1</li> </ul>	als
2	25700176	4	Several known ALS genes were found to be associated, and TBK1 (the gene encoding TANK-binding kinase 1) was identified as an ALS gene. TBK1 is known to bind to and phosphorylate a number of proteins involved in innate immunity and autophagy, including optineurin (OPTN) and p62 (SQSTM1/sequestosome), both of which have also been implicated in ALS. These observations reveal a key role of the autophagic pathway in ALS and suggest specific targets for therapeutic intervention.	YES	<ul style="list-style-type: none"> <li>Biomarker &amp; indicator</li> <li>Etiology</li> </ul>	<ul style="list-style-type: none"> <li>ALS</li> <li>OPTN</li> <li>SQSTM1</li> <li>TBK1</li> <li>optineurin</li> <li>p62</li> </ul>	als
3	25803835	5	Linkage analysis in four families gave an aggregate LOD score of 4.6. In vitro experiments confirmed the loss of expression of TBK1 LoF mutant alleles, or loss of interaction of the C-terminal TBK1 coiled-coil domain (CCD2) mutants with the TBK1 adaptor protein optineurin, which has been shown to be involved in ALS pathogenesis. We conclude that haploinsufficiency of TBK1 causes ALS and fronto-temporal dementia.	YES	<ul style="list-style-type: none"> <li>Altered Expression</li> <li>Etiology</li> <li>Polymorphism</li> </ul>	<ul style="list-style-type: none"> <li>ALS</li> <li>TBK1</li> <li>optineurin</li> </ul>	als
4	25803835	6	In vitro experiments confirmed the loss of expression of	YES	<ul style="list-style-type: none"> <li>Altered Expression</li> </ul>	<ul style="list-style-type: none"> <li>ALS</li> </ul>	als

# BioBERT と PubMedBERT の比較結果

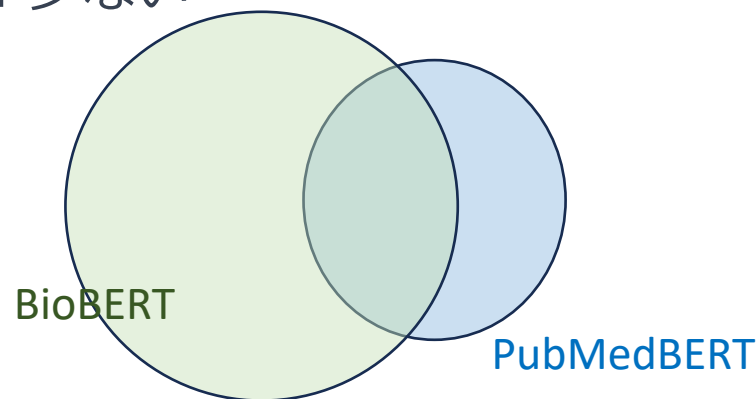
## • 遺伝子 vs 疾患の11仮説に基づいた比較

### 1. 関係性を示した遺伝子数:

- PubMedBERT: 優勢
- BioBERT: 劣勢

### 2. Entailmentと判定された文献数:

- PubMedBERT: 少ない
- BioBERT: 多い



### 3. 精度:

仮説によってばらつきがある

## 今後の方針

1. Fine-tuning 2ndで改善しないケースの対応  
仮説の内容を再評価・見直し  
Fine-tuningの手法やパラメータの再検討
2. 最終的な選択:  
BioBERT と PubMedBERT のどちらを採用するかを決定

# ChatGPTの利用

## <調査/開発の補助>

- NLP技術を用いた新機能の開発アイデアの意見聴取
- 略語の自動認識と同音異義語の処理技術のサポート
- データ解析用のPythonコードの提供
- MNLIデータセットを用いた仮説検証の方法と修正提案

## <データ作成>

- GPT-3.5 APIを使用したカスタムトレーニングデータの生成

## <検証>

- BERTを用いたMNLIデータセットの予測結果の評価

## <実利用>

- MNLIによるテキスト分類タスク用のChatGPTプロンプトの生成 (Ver.9)

# ChatGPTプロンプト生成

keyword: "als - amyotrophic lateral sclerosis"

**Disease** Amyotrophic Lateral Sclerosis Rare

T047:Disease or Syndrome

Filter Condition(s) Send Gene ID Copy Gene ID

Show 25 entries

Gene ID	Gene Symbol	Gene Title	Organism	Association Class	# of Literature	# of PubMed Abstract (Associations)	Gene Check (Ratio)	# of Homonyms	Pathway	HPO	UniProt Association	Max Phase Drug/Exp	Gene Expression	Gene Expression (Ortholog)	# of GeneRIF	# of GeneRIF (Associations)	SNPs(rsID)	Protein (Assayed)
1	23435	TARDBP	TAR DNA binding protein	Homo sapiens														
2	6647	SOD1	superoxide dismutase 1	Homo sapiens														
3	29110	TBK1	TANK binding kinase 1	Homo sapiens														
4	10433	OPTN	optineurin	Homo sapiens														
5	2521	FUS	FUS RNA binding protein	Homo sapiens														

Showing 1 to 25 of 8,025 entries (filtered from 17,685 total entries)

Gene TBK1:TANK binding kinase 1  
Disease Amyotrophic Lateral Sclerosis

Altered Expression Biomarker

ALS TBK1

Filter Condition(s) Send PMID Copy PMID

Show 25 entries

PubMed Title

PMID	PubMed Title
1	25700176 Exome sequencing in amyotrophic lateral sclerosis identifies risk genes and pathways
2	25700176 Exome sequencing in amyotrophic lateral sclerosis identifies risk genes and pathways
3	25803835 Haploinsufficiency of TBK1 causes familial ALS and fronto-temporal dementia.
4	25803835 Haploinsufficiency of TBK1 causes familial ALS and fronto-temporal dementia.

Showing 1 to 25 of 93 entries

Prompt for ChatGPT

Read the following premises and answer the questions that follow with one of three choices: contradiction, neutral, or entail.

Title: Haploinsufficiency of TBK1 causes familial ALS and frontotemporal dementia.

Prerequisite Statement: Linkage analysis in four families resulted in a total LOD score of 4.6. In vitro experiments show that loss of expression of the TBK1 LoF mutant allele, or loss of interaction of the C-terminal TBK1 coiled-coil domain (CCD2) mutant with the TBK1 adapter protein optineurin, which has been shown to be involved in the development of ALS. confirmed. We conclude that haploinsufficiency of TBK1 causes her ALS and frontotemporal dementia. (Source: PMID 25803835)

Question 1: Positive expression of "TBK1" is the cause of "ALS"

Question 2: Positive expression of "optineurin" is the cause of "ALS"

Question 3: Overexpression of "TBK1" associates "ALS"

Question 4: Overexpression of "optineurin" associates "ALS"

Question 5: Underexpression of "TBK1" associates "ALS"

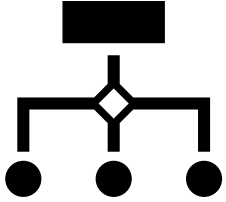
Question 6: Underexpression of "optineurin" associates "ALS"

Copyright World Fusion Co.,LTD

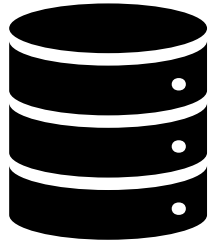
# Life Science Knowledge Bank(LSKB)

# LSKB is scientific web

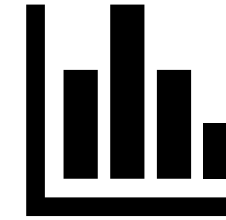
Ontology



Contents



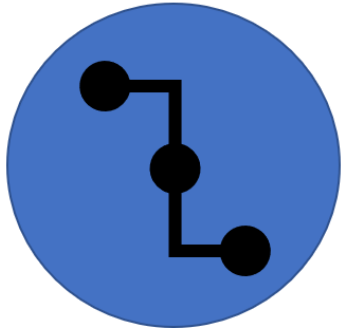
Analysis



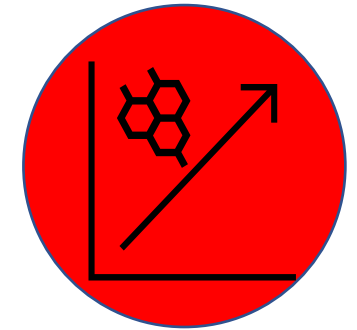
Bioinformatics



Interaction



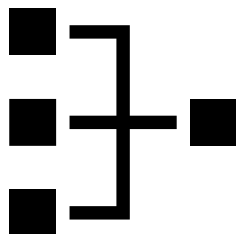
Cheminformatics



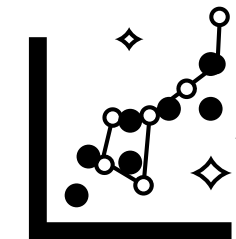
Structure



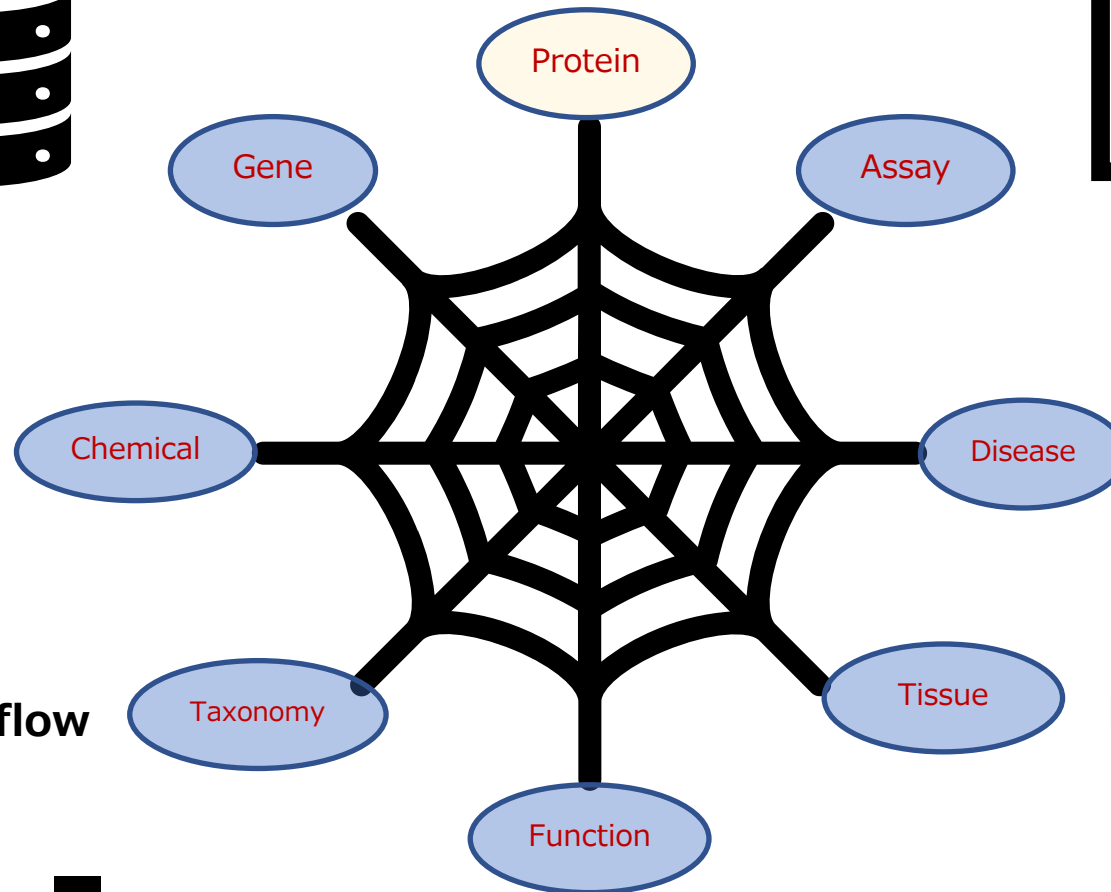
Workflow



Elpis Map

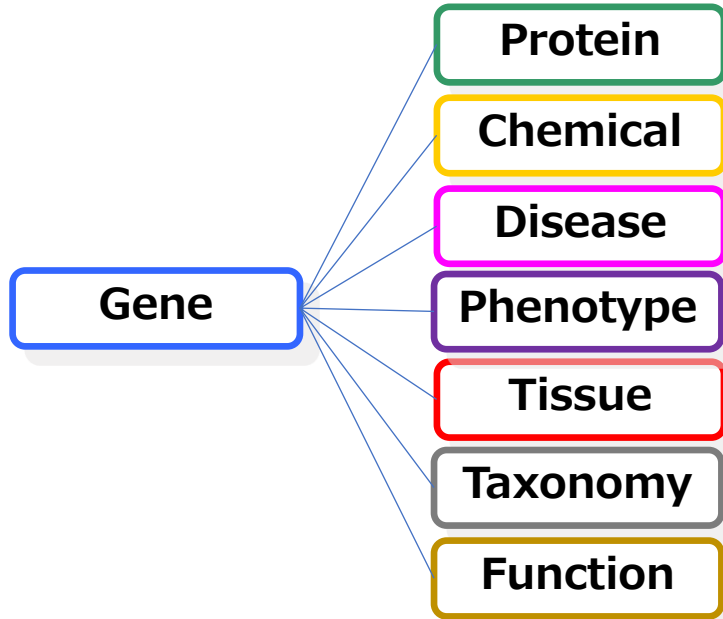
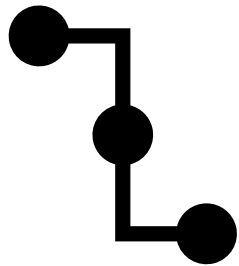


Exploration

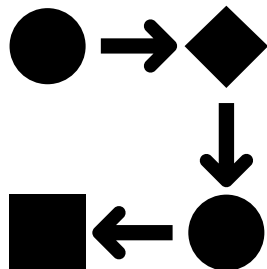


# LSKB Interaction

Interaction



GO-MoA database



## 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
- 20 years of PubMed literature

## 5) Misc

- GO-MoA database

# LSKB can help for these cases

**Target Exploration**

**Disease Exploration**

**Biomarker Exploration**

**Mechanism Exploration**

**Target Prediction**

**Novel Target Estimation  
by Protein Similarity**

**Drug Repurposing**

**Expression Analysis**

**Decision Support by  
Elpis Map**

**Data Capture by  
workflow**

**Toxicity Information**

# Target Explorer : Glaucoma related targets



## MeSH Disease Classification

- ▣ Nervous System Diseases [C10]
- ▣ Eye Diseases [C11]
  - Asthenopia [C11.093]
  - Cogan Syndrome [C11.180]
  - ▣ Conjunctival Diseases [C11.187]
  - ▣ Corneal Diseases [C11.204]
  - ▣ Eye Abnormalities [C11.250]
  - ▣ Eye Diseases, Hereditary [C11.270]
  - ▣ Eye Hemorrhage [C11.290]
  - ▣ Eye Infections [C11.294]
  - ▣ Eye Injuries [C11.297]
  - ▣ Eye Manifestations [C11.300]
  - ▣ Eye Neoplasms [C11.319]
  - ▣ Eyelid Diseases [C11.338]
  - ▣ Lacrimal Apparatus Diseases [C11.496]
  - ▣ Lens Diseases [C11.510]
  - ▣ Ocular Hypertension [C11.525]
    - ▣ Glaucoma [C11.525.381]
      - Glaucoma, Angle-Closure [C11.525.381.056]
      - Glaucoma, Neovascular [C11.525.381.348]
      - ▣ Glaucoma, Open-Angle [C11.525.381.407]
        - Low Tension Glaucoma [C11.525.381.703]

## Disease List

- Angle Closure Glaucoma
- Glaucoma, Open-Angle
- Hydrophthalmos
- Low Tension Glaucoma
- Neovascular Glaucoma
- Congenital Glaucoma
- Secondary open-angle glaucoma
- Primary Open Angle Glaucoma
- Glaucoma

Literature mining

AI-based Curation

OMIM/MedGen etc

Gene Expression

SNPs Information

Indication

Clinical Trial

Pathway

検索条件に入れた複数の疾患について  
様々なInteraction、文献  
解析情報などから  
ターゲットをランキング

# Target Explorer : Glaucoma related targets



Input :

- Angle Closure Glaucoma
- Glaucoma, Open-Angle
- Hydrophthalmos
- Low Tension Glaucoma

## 疾患と遺伝子の関係性の概要分類(BERT)

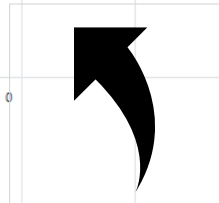
→ Filter Condition(s) Send Gene ID Copy Gene ID

Show 25 entries

Rank	Gene ID	Gene Symbol	Gene Title	Organism	# of Disease	Association Class	# of Literature	# of PubMed Abstract (Associations)	Gene Check (Ratio)	# of GeneRIF	# of GeneRIF (Associations)	HPO	UniProt Association	Gene Expression	Gene Expression (Ortholog)	SNPs(rsID)	Pathway	Max Phase Drug/Exp	Gene in Disease Summary	SUM of Confidence Score
5	154	ADRB2	adrenoceptor beta 2	Homo sapiens	8	<ul style="list-style-type: none"> <li>Etiology (GR)</li> <li>Polymorphism (GR)</li> <li>Biomarker &amp; indicator (PM)</li> <li>Etiology (PM)</li> <li>Polymorphism (PM)</li> </ul>	2	10	1.0	2	1	0	YES	0	0	0.03	0	4		1.86
6	348	APOE	apolipoprotein E	Homo sapiens	6	<ul style="list-style-type: none"> <li>Altered Expression (GR)</li> <li>Etiology (GR)</li> <li>Polymorphism (GR)</li> <li>Biomarker &amp; indicator (PM)</li> <li>Etiology (PM)</li> <li>Polymorphism (PM)</li> </ul>	32	38	1.0	15	13	0		0	0	0.17	0	0		1.77
7	5737	PTGFR	prostaglandin F receptor	Homo sapiens	9	<ul style="list-style-type: none"> <li>Etiology (GR)</li> <li>Polymorphism (GR)</li> <li>Biomarker &amp; indicator (PM)</li> <li>Etiology (PM)</li> <li>Polymorphism (PM)</li> </ul>	18	8	1.0	3	3	0	YES	0	0	0.17	0	4		1.76
8	4053	LTBP2	latent transforming growth factor beta binding protein 2	Homo sapiens	5	<ul style="list-style-type: none"> <li>Etiology (GR)</li> <li>Polymorphism (GR)</li> <li>Biomarker &amp; indicator (PM)</li> <li>Enhancement (PM)</li> <li>Etiology (PM)</li> <li>Polymorphism (PM)</li> </ul>	6	2				6	2		0	0	2.02	0	0	1.69
9	134430	WDR36	WD repeat domain 36	Homo sapiens	7	<ul style="list-style-type: none"> <li>Altered Expression (GR)</li> <li>Enhancement (GR)</li> <li>Etiology (GR)</li> <li>Inhibition (GR)</li> <li>Polymorphism (GR)</li> <li>Altered Expression (PM)</li> <li>Biomarker &amp; indicator (PM)</li> <li>Enhancement (PM)</li> <li>Etiology (PM)</li> <li>Inhibition (PM)</li> <li>Polymorphism (PM)</li> </ul>	2	2				2	2		0	0	4.7	0	0	1.68
10	4976	OPA1	OPA1 mitochondrial dynamin like GTPase	Homo sapiens	7	<ul style="list-style-type: none"> <li>Altered Expression (GR)</li> <li>Biomarker &amp; indicator (GR)</li> <li>Etiology (GR)</li> <li>Polymorphism (GR)</li> <li>Altered Expression (PM)</li> <li>Biomarker &amp; indicator (PM)</li> <li>Enhancement (PM)</li> </ul>	152	525	1.0	10	9	0		0	0	0.23	0	0	0	1.65

**Association Class**

- Etiology
- Biomarker & Indicator
- Altered Expression
- Polymorphism
- Inhibition
- Enhancement



# Genetic Biomarker Exploration Example



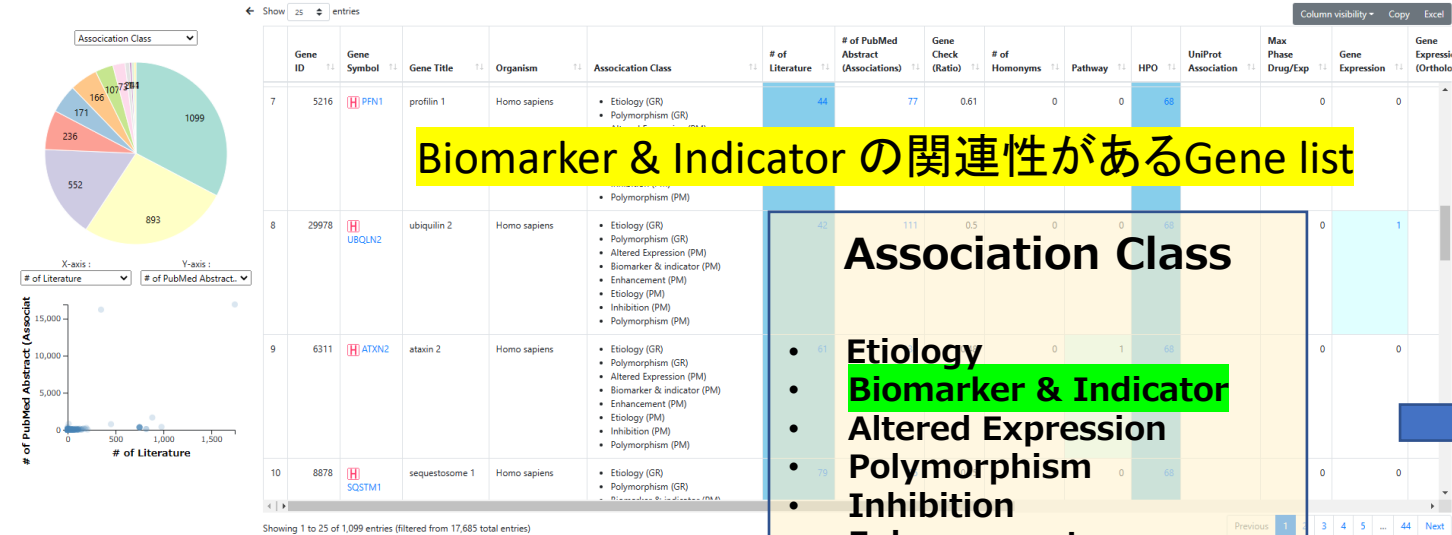
## Disease related Gene Information

keyword : "als - amyotrophic lateral sclerosis"

**Disease** Amyotrophic Lateral Sclerosis Rare

T047:Disease or Syndrome

Filter Conditions Send Gene ID Copy Gene ID

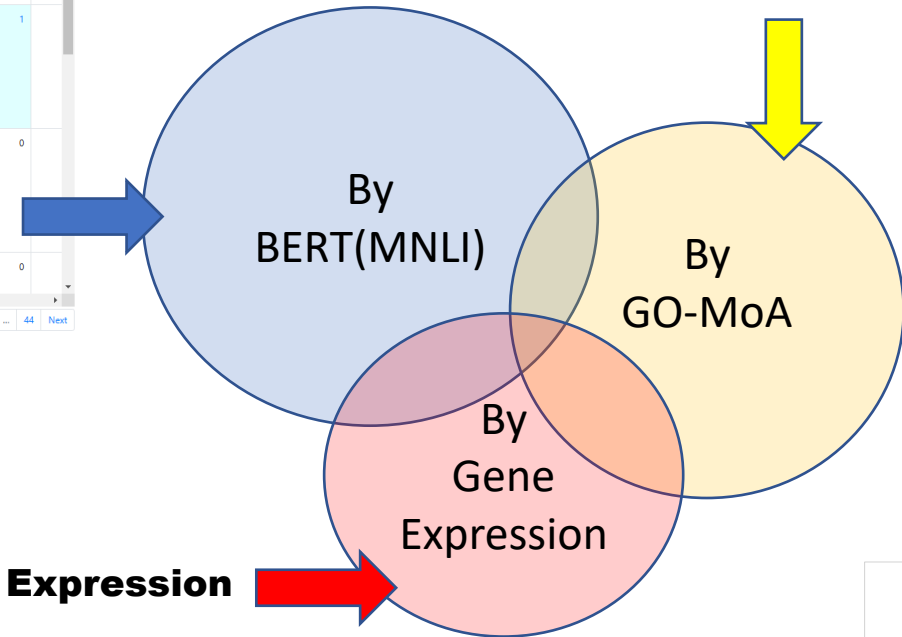
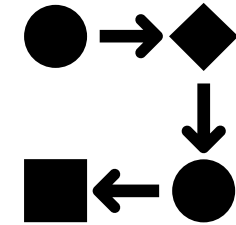


Biomarker & Indicator の関連性があるGene list

### Association Class

- Etiology
- Biomarker & Indicator
- Altered Expression
- Polymorphism
- Inhibition
- Enhancement

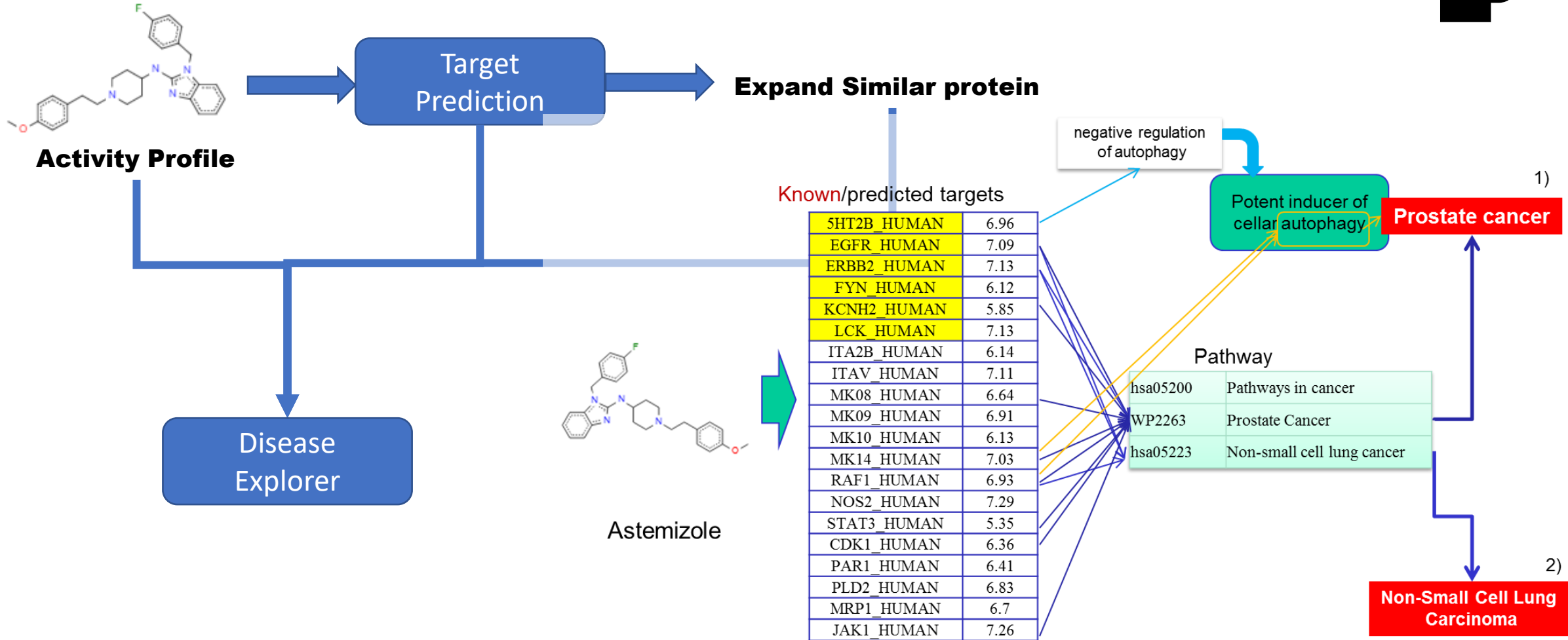
## Genes from disease associated function By GO-MoA



Differential Gene Expression



# Drug Repurposing Procedure



Target Prediction & map the targets to pathway by LSKB.

1) Drug repurposing from an academic perspective

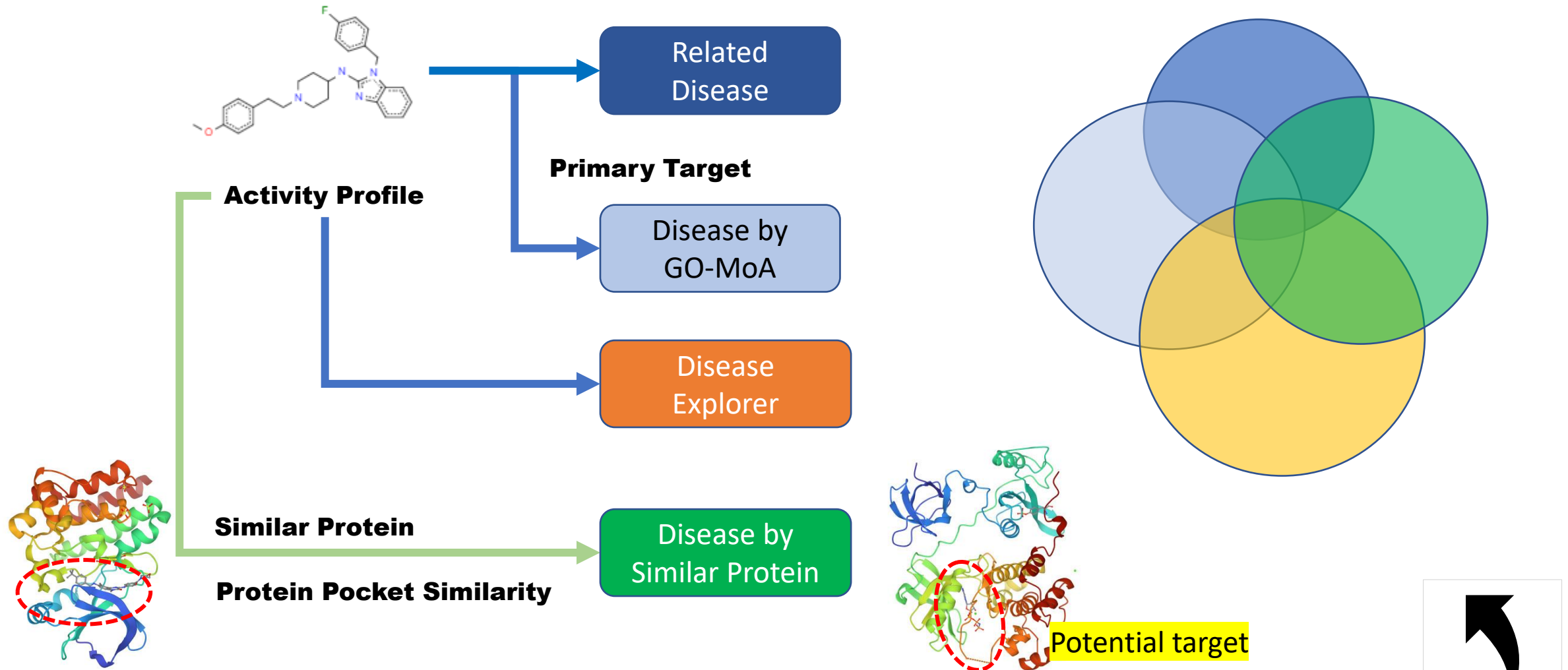
DOI: 10.1016/j.ddstr.2011.10.002

2) Repurposing Cationic Amphiphilic Antihistamines for Cancer Treatment

<http://dx.doi.org/10.1016/j.ebiom.2016.06.013>



# Drug Repurposingやターゲット探索へのパイプライン構築などに応用



# Toxicity Information

## Toxicity associated Targets

### Toxicity Hepatotoxicity

(beta version)

T037:Injury or Poisoning

Filter Condition(s) Copy Gene ID

Show 25 entries

Gene ID	Gene Symbol	Gene Title	Organism	Association Class	# of Literature	# of PubMed Abstract (Associations)	Gene Check (Ratio)	Pathway	UniProt Association	Gene Expression	Gene Expression (Ortholog)	# of GeneRIF
4	1576	CYP3A4 cytochrome P450 family 3 subfamily A member 4	Homo sapiens	<ul style="list-style-type: none"> <li>Etiology (GR)</li> <li>Altered Expression (PM)</li> <li>Biomarker &amp; indicator (PM)</li> <li>Enhancement (PM)</li> <li>Etiology (PM)</li> <li>Inhibition (PM)</li> </ul>	59	46	0.22	5		1	0	
5	8856	1 group 1 member 2		<ul style="list-style-type: none"> <li>Enhancement (GR)</li> <li>Etiology (GR)</li> <li>Inhibition (GR)</li> <li>Polymorphism (GR)</li> <li>Biomarker &amp; indicator (PM)</li> <li>Enhancement (PM)</li> <li>Etiology (PM)</li> <li>Inhibition (PM)</li> <li>Polymorphism (PM)</li> </ul>						0	1	
6	2944	GSTM1 glutathione S-transferase mu 1	Homo sapiens	<ul style="list-style-type: none"> <li>Etiology (GR)</li> <li>Polymorphism (GR)</li> <li>Biomarker &amp; indicator (PM)</li> <li>Enhancement (PM)</li> <li>Etiology (PM)</li> <li>Inhibition (PM)</li> <li>Polymorphism (PM)</li> </ul>			0.48	3		0	0	
7	54657	UGT1A4 UDP glucuronosyltransferase	Homo sapiens	<ul style="list-style-type: none"> <li>Altered Expression (PM)</li> <li>Biomarker &amp; indicator (PM)</li> </ul>						1	1	

疾患と遺伝子の関係性の概要分類(BERT)

### Association Class

- Etiology
- Biomarker & Indicator
- Altered Expression
- Polymorphism
- Inhibition
- Enhancement

## Drugs observed as side effects

LsKB About Contact

Summary

Annotation

Classification [2]

Drug [20]

Gene [5855]

SNPs(hg38) [593]

SNPs(hg19) [300]

Gene Expression [56]

### Toxicity Hepatic steatosis

(beta version)

T047:Disease or Syndrome

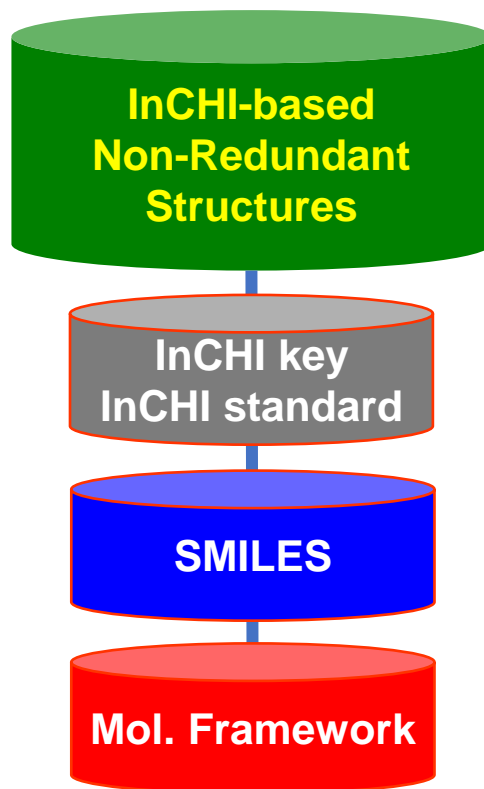
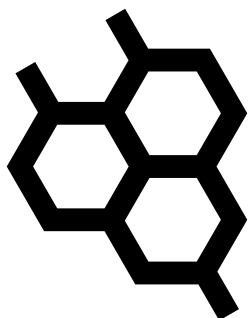
Filter Condition(s) Copy Chem ID Download SD File hide Chemical Structure

Show 25 entries

LSKB_CHEM_ID	Chemical	MW	Molecule Type	Withdrawn	Reported as Side Effect	ATC Class
3	1000401435 Etretinate	354.49	Small molecule	<ul style="list-style-type: none"> <li>Congenital Abnormality : Risk for birth defects</li> </ul>	0	<ul style="list-style-type: none"> <li>D05B801</li> <li>D05B8 Retinoids for treatment of psoriasis</li> </ul>
4	1000902543 Cianidanol	290.271	Small molecule	<ul style="list-style-type: none"> <li>Hemolytic Anemia : Hemolytic Anemia</li> </ul>	0	
5	1000507334 Rimonabant	463.796	Small molecule	<ul style="list-style-type: none"> <li>Mental Depression : Psychiatric effects, especially depression</li> <li>Depressed Mood : Psychiatric effects, especially depression</li> <li>Cancer patients and suicide and depression : Psychiatric effects, especially depression</li> <li>Depressivity :</li> </ul>	0	<ul style="list-style-type: none"> <li>A08AX01</li> <li>A08AX Other antiobesity drugs</li> </ul>

Showing 1 to 20 of 20 entries (filtered from 9,760 total entries)

## Structure



### 1) Huge Structure Collection

- InCHI-based Non-redundant
- 110Million
- Molecular Framework

### 2) Structure Search

- Exact Match
- Similarity Search
- Substructure Search
- Molecular Framework search

### 3) Prediction/Analysis

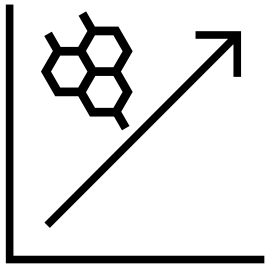
- Target Prediction
- Target Confirmation
- Mechanism Explorer

### 4) Misc

- Multiple structures search
- Automatic saving of search result



## Cheminformatics



Target Prediction

Target Confirmation

Chemical Properties

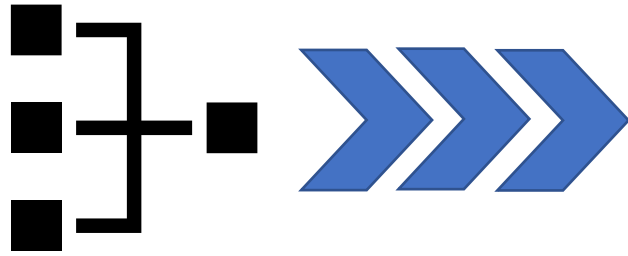
Molecular Framework

Go To MOE

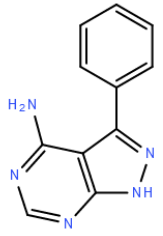
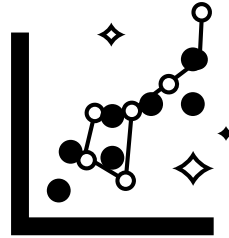
Workflow

# 2. PIK3A/ PIK3B/ PIK3G/ PIK3Dの活性データ

## Workflow



## Elpis Map



注目している部分構造を持つ化合物の検索

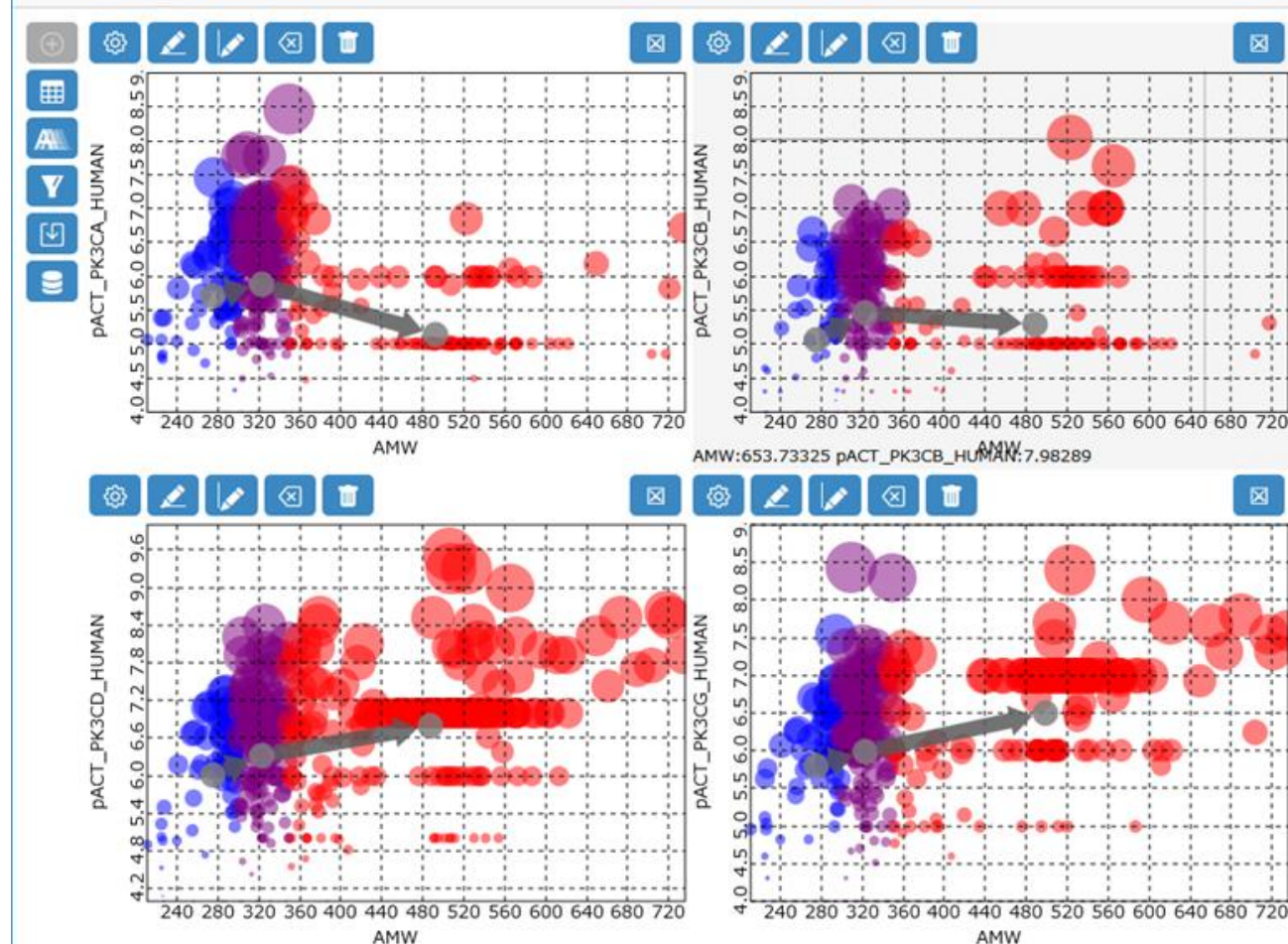
ターゲットたんぱく質の指定

ターゲットたんぱく質と化合物双方を考慮したアッセイデータ検索

アッセイデータ中の化合物の収集

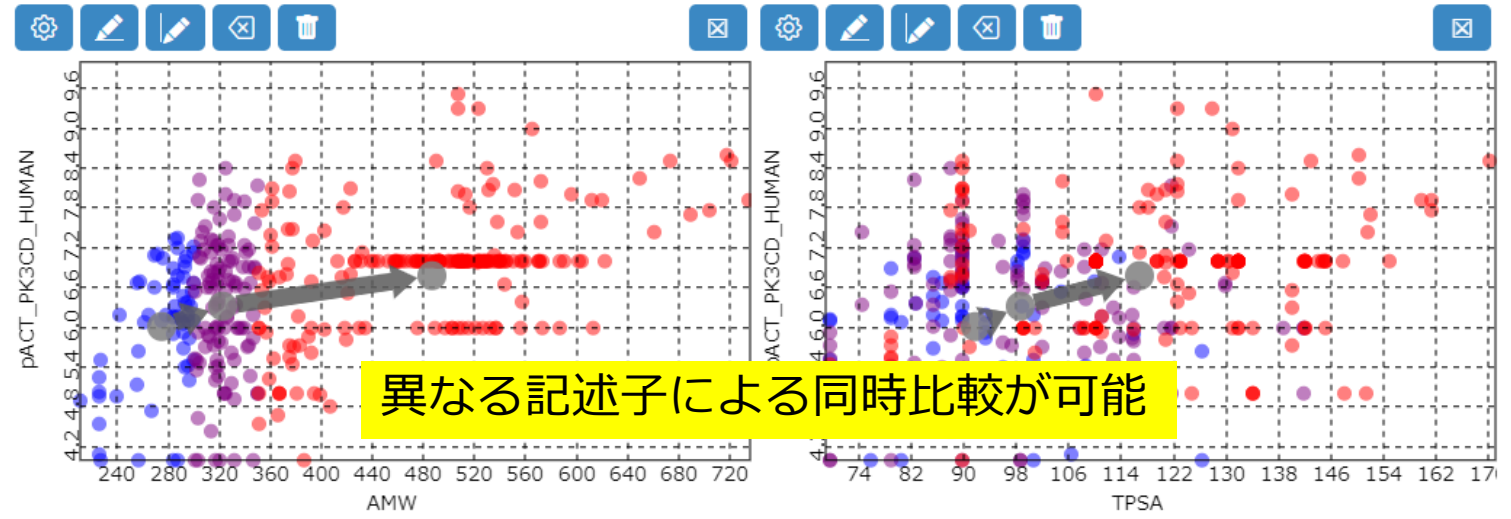
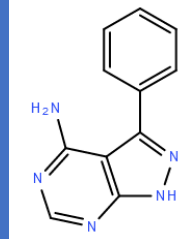
化合物の特徴量の計算・検索

表形式ファイル、およびプロット図作成を指示



# 2. PIK3A/ PIK3B/ PIK3G/ PIK3Dの活性データ

部分構造を持つ活性化合物  
PIK3A/ PIK3B/ **PIK3D**/ PIK3Gの  
活性値



MOE 2020.09

File Edit Select Render Protein Compute Window Help **LSKB**

- MOEにてDescriptor計算

Open LSKB  
PDB\_Info

Calculate Descriptors

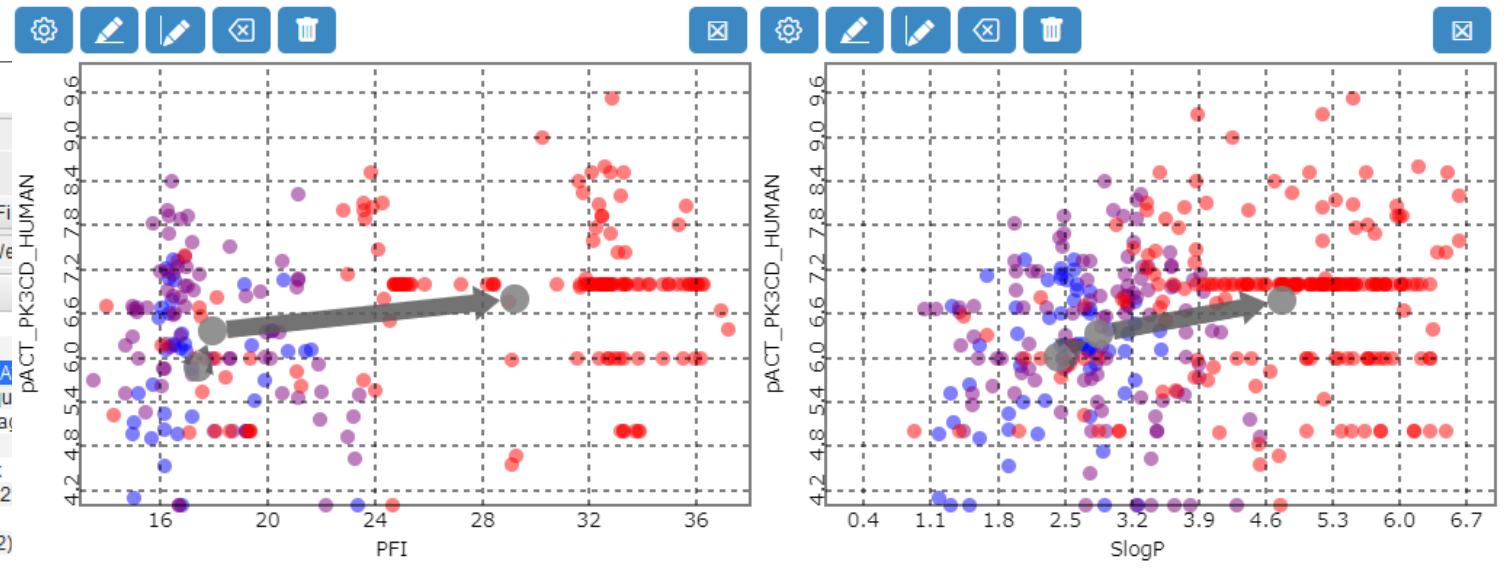
Database File: \$DOWNLOADS/lskb\_idlist5.mdb

Molecule Field: mol

Auto Select: Database Fields Selected Database Fields

Descriptors Selected: 6 CNS\_MPO, mr, PFI, SlogP, TPSA, W...

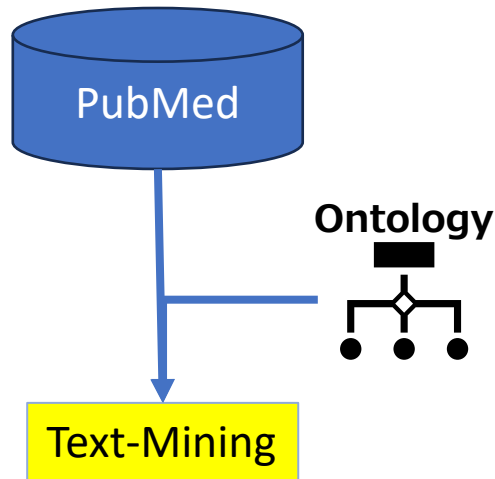
Code	Class	Description
SMR_VSA6	2D	Bin 6 SMR (0.485,0.560]
SMR_VSA7	2D	Bin 7 SMR (0.560,10]
TPSA	2D	Topological Polar Surface Area (A
VAdjEq	2D	Vertex adjacency information (equ
VAdjMa	2D	Vertex adjacency information (mag
VDistEq	2D	Vertex distance equality index
VDistMa	2D	Vertex distance magnitude index
vdw_area	2D	Van der Waals surface area (A**2
vdw_vol	2D	Van der Waals volume (A**3)
vsa_acc	2D	VDW acceptor surface area (A**2)
vsa_acid	2D	VDW acidic surface area (A**2)
vsa_base	2D	VDW basic surface area (A**2)
vsa_don	2D	VDW donor surface area (A**2)





# LSKB ver. 9 アップデート戦略

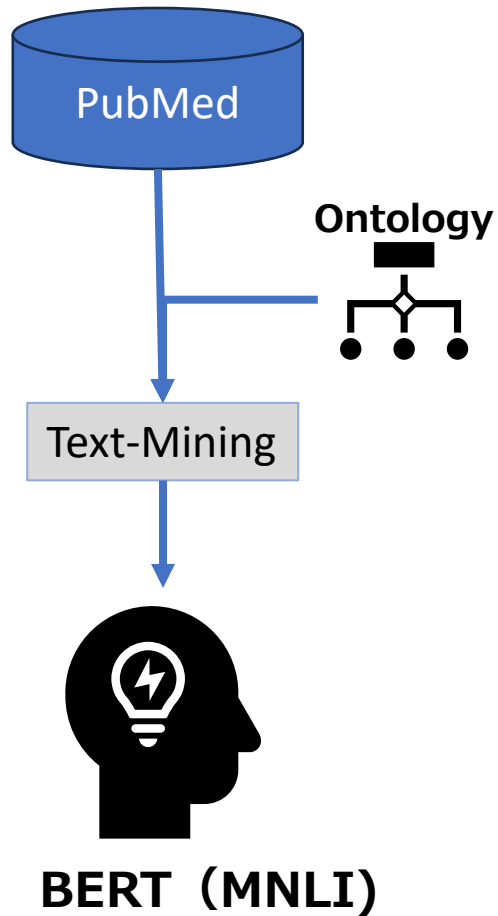
# 文献情報解析



- LSKBオントロジーによるテキストマイニング
- 二項関係の共起例（遺伝子-疾患 化合物-遺伝子）

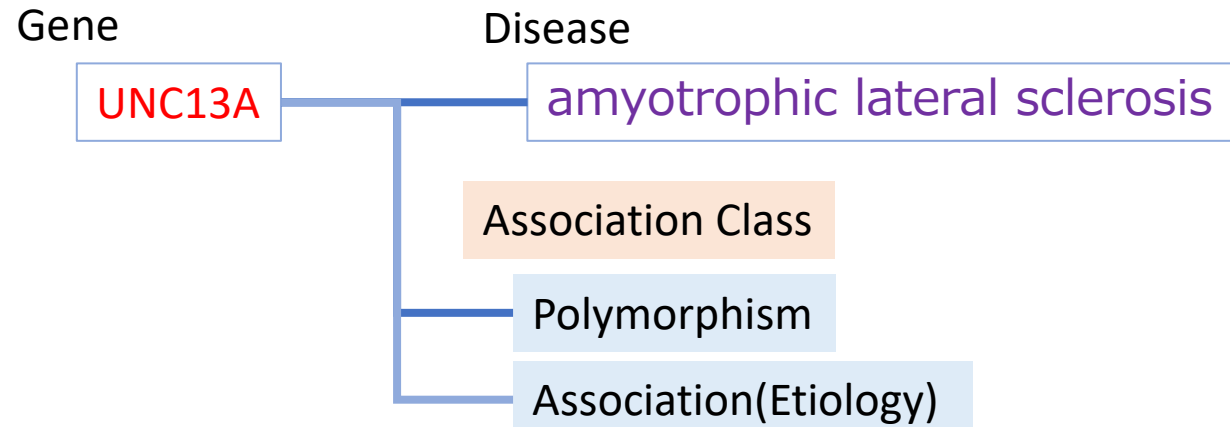
Mining Level	説明/ 例文
Lv.0	遺伝子名 と 疾患名 が 同一論文だが、別々のセンテンスに記載
	Mutations in the <b>sigma non-opioid intracellular receptor 1</b> gene ( <b>SIGMAR1</b> ) has been linked to autosomal recessive dHMN with pyramidal signs in several families. This phenotype can mimic <b>amyotrophic lateral sclerosis (ALS)</b> .
Lv. 1	遺伝子名 と 疾患名 が 同一センテンスに記載
	<b>UNC13A</b> variant rs12608932 is associated with increased risk of <b>amyotrophic lateral sclerosis</b> and reduced patient survival: a meta-analysis.
Lv. 2	化合物名、遺伝子名および 機能用語が 同一センテンスに記載
	<b>Bortezomib</b> induces a potent cytotoxic effect with <b>caspase-3</b> activation.

# 文献情報解析

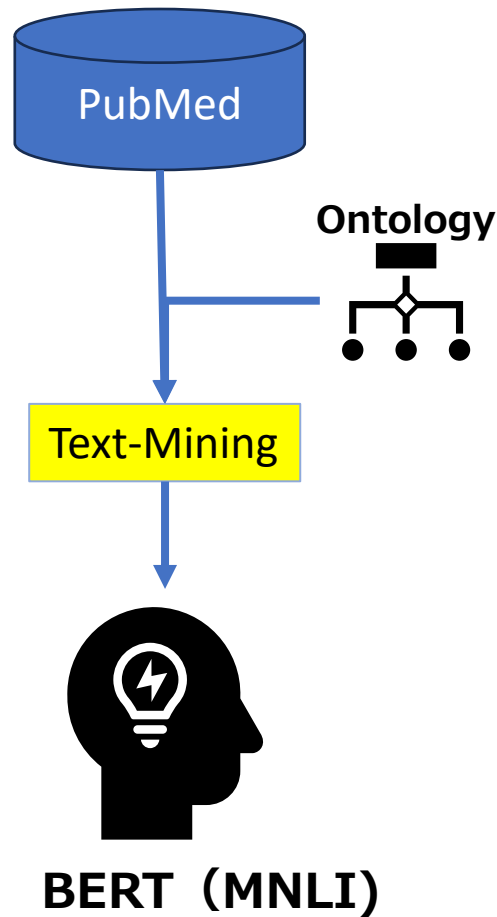


- BERT (MNLi) による記載内容の分類

UNC13A variant rs12608932 is associated with increased risk of amyotrophic lateral sclerosis and reduced patient survival: a meta-analysis.



# 文献情報解析：アップデート



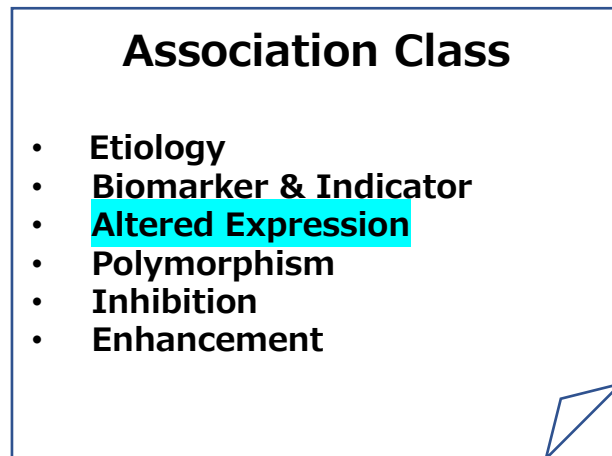
- 文献のテキストマイニングのウィークリーアップデート
- MNLIによるクラス分類のマンスリーアップデート
  - 遺伝子 vs 疾患（毒性）
- MNLIの Fine-tuning アップデートモデルによるデータ更新
- MNLIの仮説の増強と更新
- BioBERT/PubMedBERTでの検証
- ChatGPTへの検証クエリー生成

# クラス分類用MNLI仮説の更新と追加

- 1<sup>st</sup> Generation Hypothesis

- Altered Expression Class

- positive expression of "gene" is the risk of "disease"
- negative expression of "gene" is the risk of "disease"

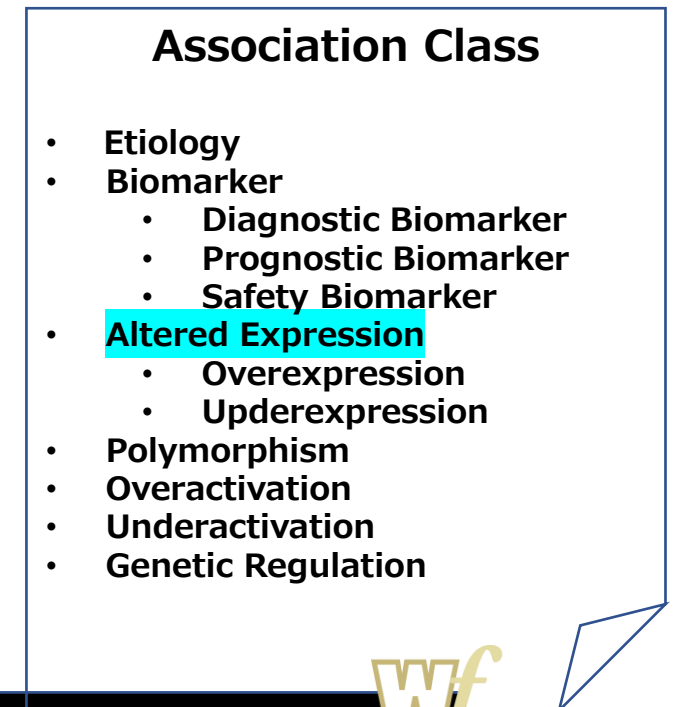


Disease vs Gene 11 種の仮説

- 2<sup>nd</sup> Generation Hypothesis

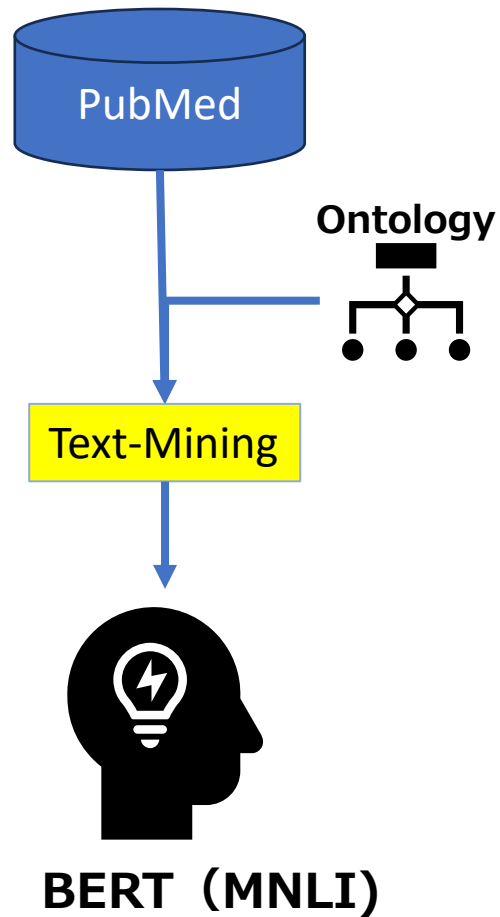
- Altered Expression Class

- Overexpression of "gene" associates "disease"
- Underexpression of "gene" associates "disease"



Disease vs Gene 20 種の仮説

# 文献情報解析：アップデート



- 文献のテキストマイニングのウィークリーアップデート
- MNLIによるクラス分類のマンスリーアップデート
  - 遺伝子 vs 疾患（毒性）
- MNLIの Fine-tuning アップデートモデルによるデータ更新
- MNLIの仮説の増強と更新
- BioBERT/PubMedBERTでの検証結果の搭載
- ChatGPTの検証クエリー生成

# ChatGPTの検証クエリー生成

keyword: "als - amyotrophic lateral sclerosis"

**Disease** Amyotrophic Lateral Sclerosis Rare

T047:Disease or Syndrome

Filter Condition(s) Send Gene ID Copy Gene ID

Show 25 entries

Gene ID	Gene Symbol	Gene Title	Organism	Association Class	# of Literature	# of PubMed Abstract (Associations)	Gene Check (Ratio)	# of Homonyms	Pathway	HPO	UniProt Association	Max Phase Drug/Exp	Gene Expression	Gene Expression (Ortholog)	# of GeneRIF	# of GeneRIF (Associations)	SNPs(rsID)	Protein (Assayed)
1	23435	TARDBP	TAR DNA binding protein	Homo sapiens														
2	6647	SOD1	superoxide dismutase 1	Homo sapiens														
3	29110	TBK1	TANK binding kinase 1	Homo sapiens														
4	10433	OPTN	optineurin	Homo sapiens														
5	2521	FUS	FUS RNA binding protein	Homo sapiens														

Showing 1 to 25 of 8,025 entries (filtered from 17,685 total entries)

Gene TBK1:TANK binding kinase 1  
Disease Amyotrophic Lateral Sclerosis

Filter Condition(s) Send PMID Copy PMID

Show 25 entries

PubMed Title

PMID	PubMed Title
1	25700176 Exome sequencing in amyotrophic lateral sclerosis identifies risk genes and pathways
2	25700176 Exome sequencing in amyotrophic lateral sclerosis identifies risk genes and pathways
3	25803835 Haploinsufficiency of TBK1 causes familial ALS and frontotemporal dementia.
4	25803835 Haploinsufficiency of TBK1 causes familial ALS and frontotemporal dementia.

Showing 1 to 25 of 93 entries

Prompt for ChatGPT

Read the following premises and answer the questions that follow with one of three choices: contradiction, neutral, or entail.

Title: Haploinsufficiency of TBK1 causes familial ALS and frontotemporal dementia.

Prerequisite Statement: Linkage analysis in four families resulted in a total LOD score of 4.6. In vitro experiments show that loss of expression of the TBK1 LoF mutant allele, or loss of interaction of the C-terminal TBK1 coiled-coil domain (CCD2) mutant with the TBK1 adapter protein optineurin, which has been shown to be involved in the development of ALS. confirmed. We conclude that haploinsufficiency of TBK1 causes her ALS and frontotemporal dementia. (Source: PMID 25803835)

Question 1: Positive expression of "TBK1" is the cause of "ALS"

Question 2: Positive expression of "optineurin" is the cause of "ALS"

Question 3: Overexpression of "TBK1" associates "ALS"

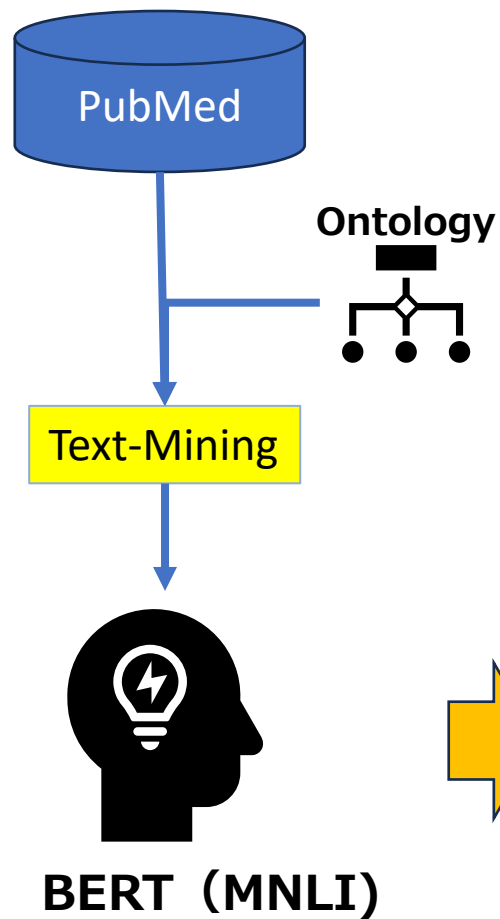
Question 4: Overexpression of "optineurin" associates "ALS"

Question 5: Underexpression of "TBK1" associates "ALS"

Question 6: Underexpression of "optineurin" associates "ALS"

Copy

# 文献情報解析アプローチ



- 文献のテキストマイニング << Weekly Update
  - 2項関係での早期解析
- MNLiによるクラス分類 << Monthly Update
  - 記載内容の分類と情報の幅の拡大
- ChatGPTの検証クエリー生成
  - 検証

欲しい文献の提案

登録論文の解析 >> 類似論文を提案



次世代PubMed検索システム

# PubMedify

## キーワード登録レスのPubMed検索WEBブラウザで利用可能

### PubMed PMIDや文章から自動的にキーワードを抽出

PMIDや文章からキーワードを抽出し、毎週PubMed論文Pick-Upします。



システムが膨大なオントロジーを利用してキーワード抽出し、毎週論文を提案。

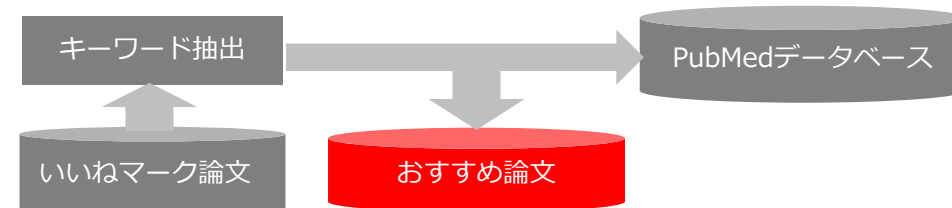
複数の論文や文章を登録することで、PubMedifyは検索精度が高くなります。



AIIC利用する	オントロジーターム	オントロジータイプ	TP*	IDF	TF-IDF*
<input type="checkbox"/>	Periodontitis	Phenotype	0.013369	9.12549	0.121998
<input type="checkbox"/>	Periodontitis	Disease	0.013369	8.9554	0.119725
<input type="checkbox"/>	Subgingival plaque	Disease	0.00802139	14.0617	0.112795
<input type="checkbox"/>	Prevotella intermedia	Taxonomy	0.00802139	13.9345	0.111775
<input type="checkbox"/>	Periodontitis, Aggressive, 1	Disease	0.00802139	13.4445	0.107845
<input type="checkbox"/>	Acute Periodontitis	Disease	0.00802139	13.4372	0.107785
<input type="checkbox"/>	Aggressive Periodontitis	Disease	0.00802139	13.3998	0.107485

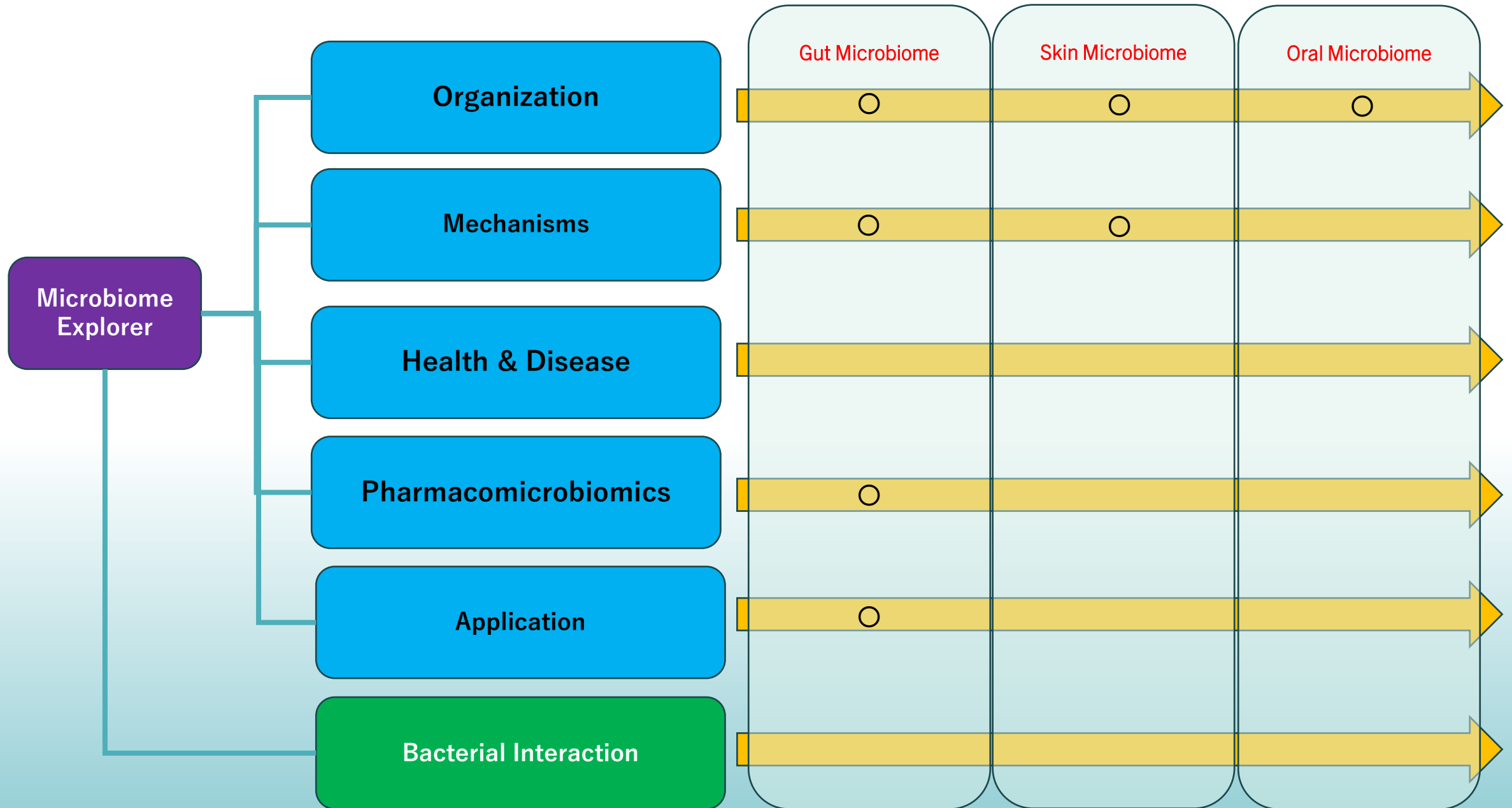
### おすすめ論文のPick-Up

登録文章やいいねマーク論文から抽出されたキーワードをもとに、当社内のPubMed Abstractデータベースをテキストマイニングして、週2回おすすめ論文を作成します。



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

# MICROBIOME EXPLORER



# Conclusion

- LSKBにおける文献解析の研究において、BERT（特にBioBERTとPubMedBERT）の利用が重要な役割を果たしました。
- 具体的には、BioBERTとPubMedBERTを使用して2種類のFine-tuningを行い、モデルの作成を実施しました。この過程で、GeneRIFデータにおけるDiseaseとGeneの関連性を解析し、両モデルがDomain SpecificなFine-tuningデータを使用することで精度が向上することを確認しました。
- さらに、BioBERTとPubMedBERTの性能比較を行った結果、特定のタスクやデータセットにおいて優劣の傾向は見られました。この研究を通じて、ChatGPTの潜在的な役割やその有用性についても考察しました。
- この文献解析の精度の向上はLSKBの多様な探索機能を強化します。
- 最終的に、LSKBの機能を強化するLSKB v9のアップデート戦略を策定しました。



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/>