CBI学会2023年大会
ランチョンセミナー

AI Knowledge Station LSKB

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

World Fusion

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。

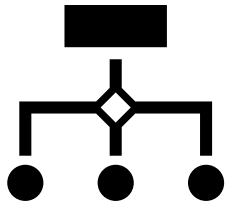
4.ChatGPTの役割

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。

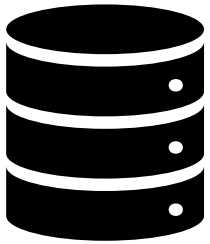
LSKB is scientific web

Knowledge Station LSKB

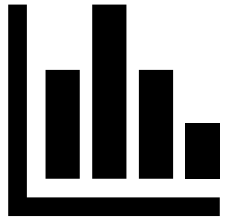
Ontology



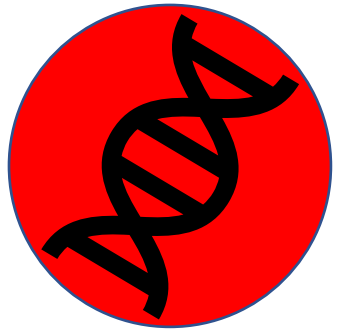
Contents



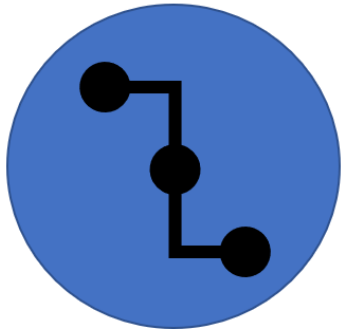
Analysis

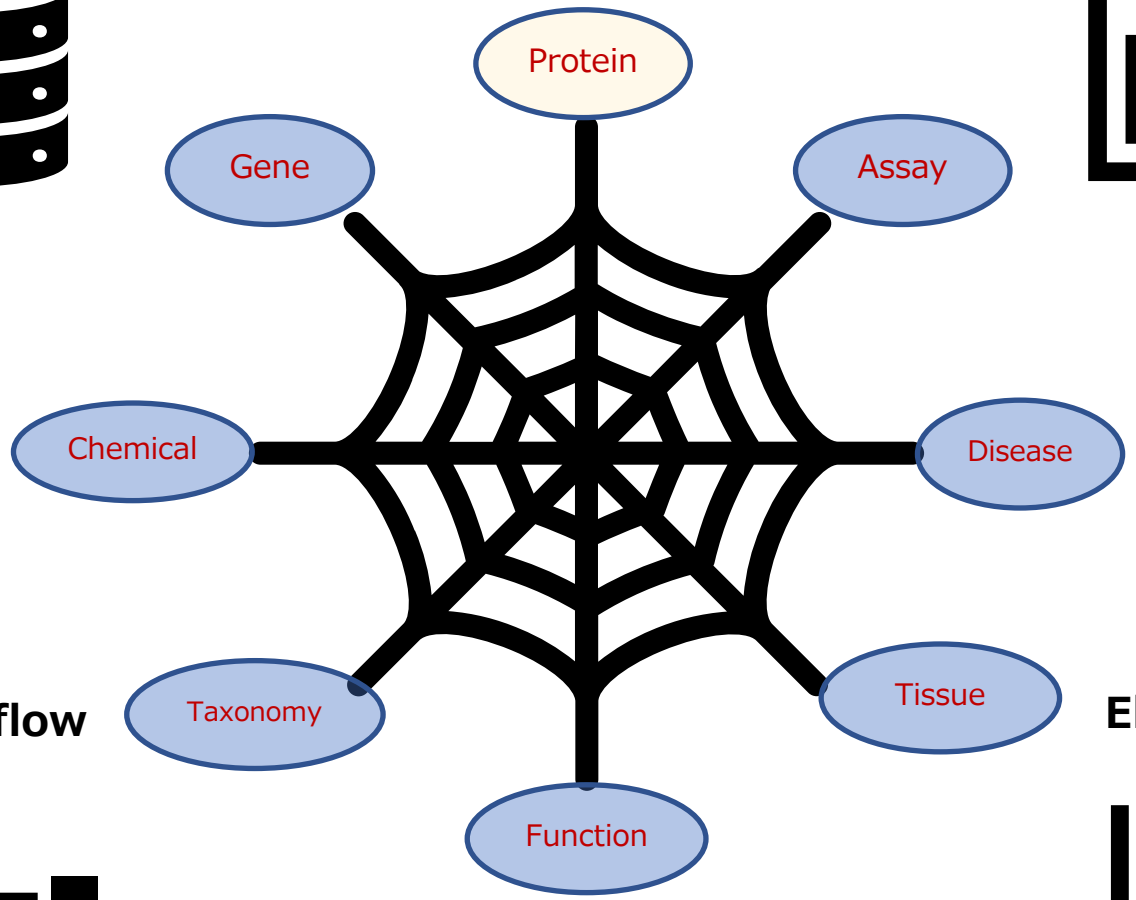


Bioinformatics

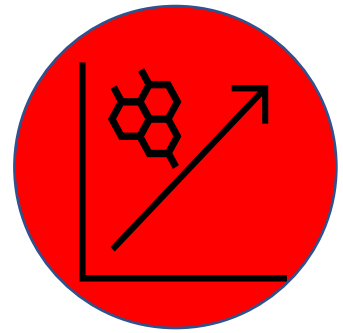


Interaction



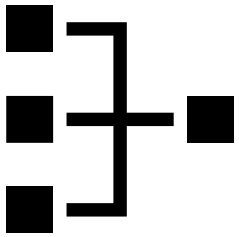


Cheminformatics

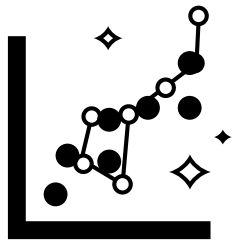


Structure

Workflow



Elpis Map



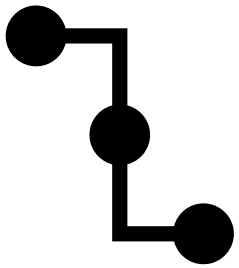
Exploration

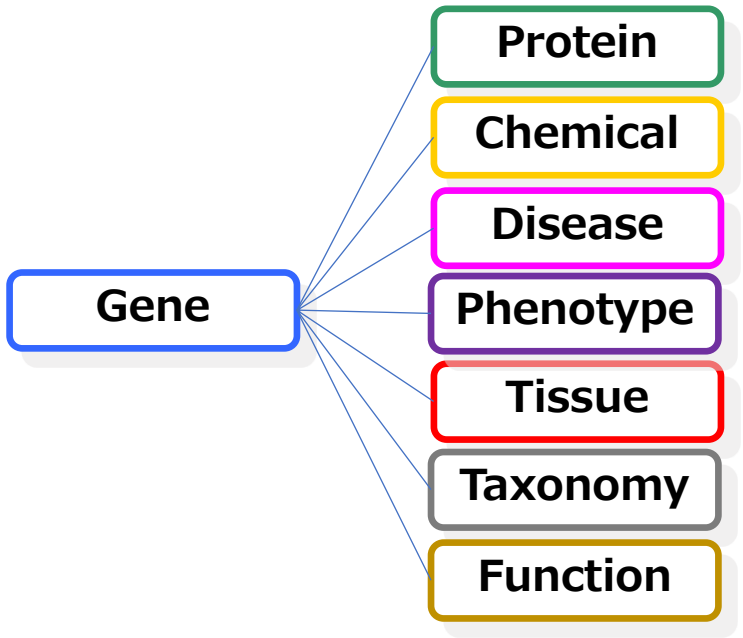
World Fusion

LSKB Interaction

Knowledge Station LSKB

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

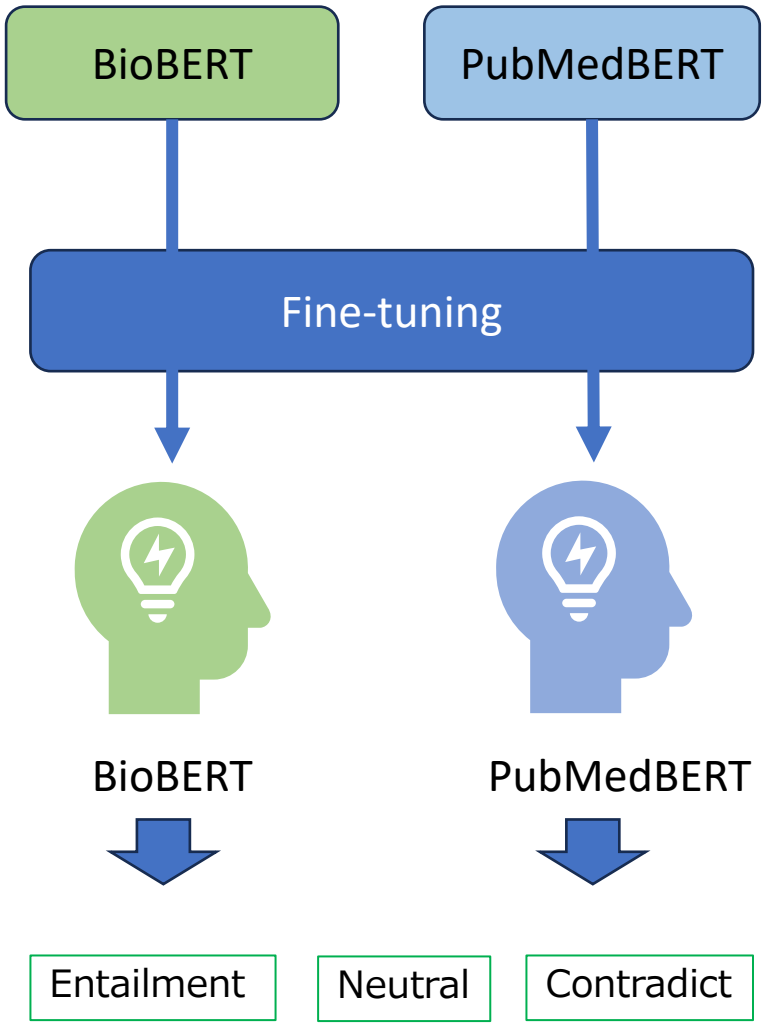
4) AI Curated Annotation
- Mechanism of Action
- Gene RIF
- 20 years of PubMed literature
  - Disease & Toxicity associated Gene
  - Disease vs Phenotype

World Fusion

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。

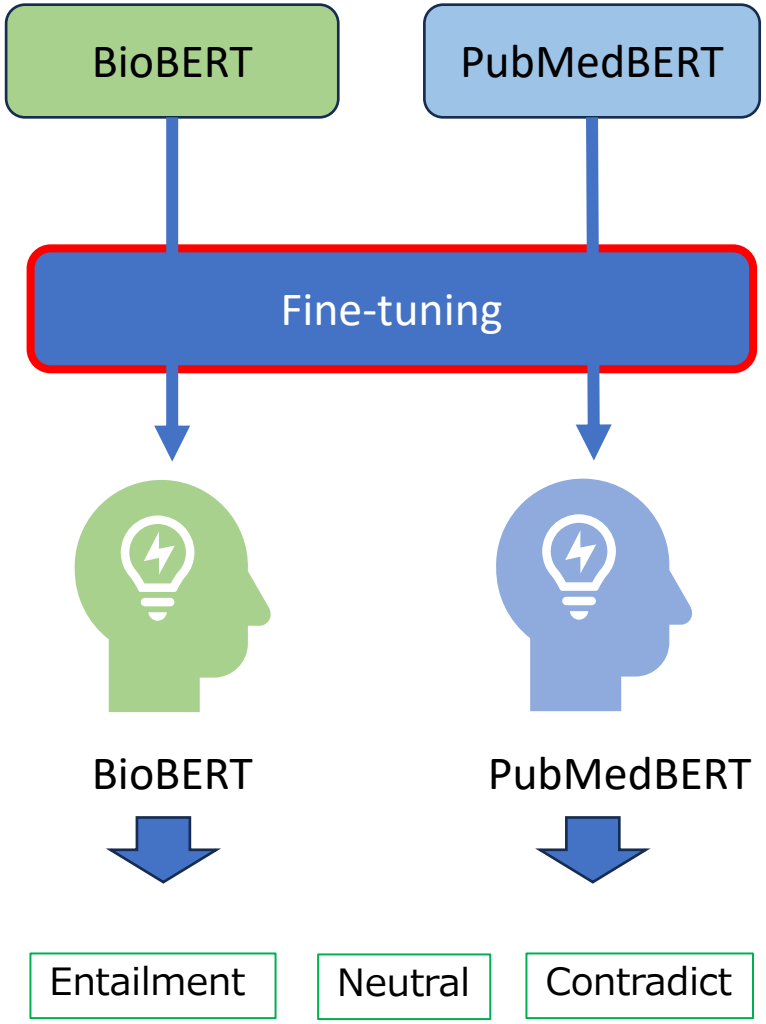
ファインチューニング:

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 サンプリングした例を **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.

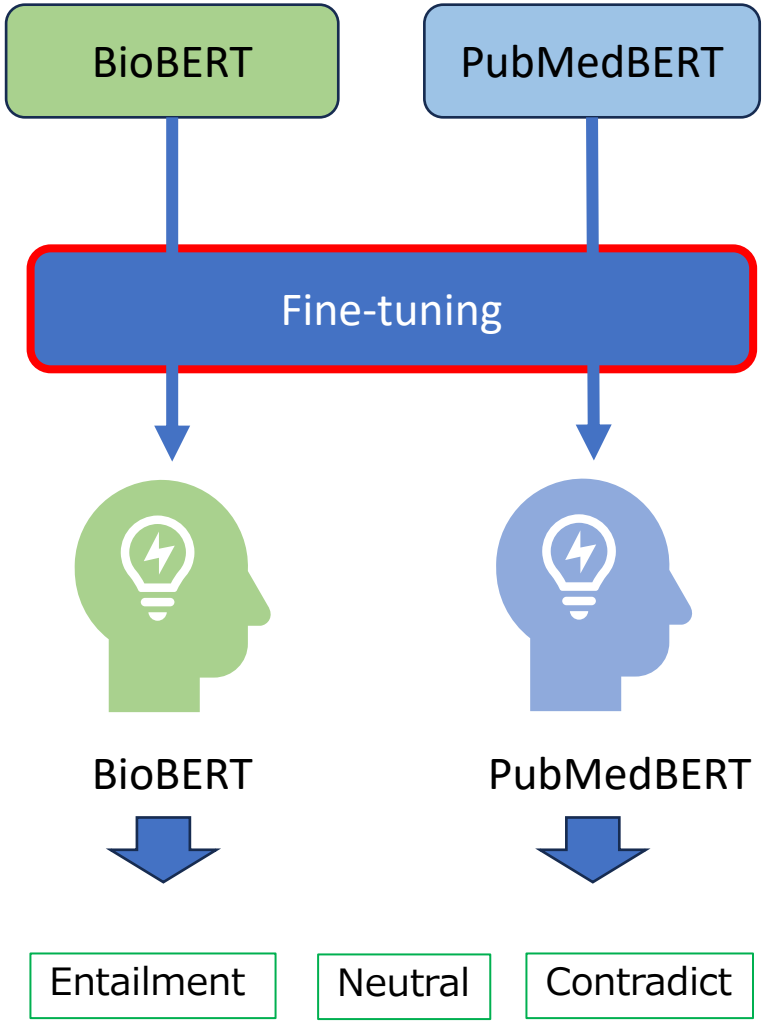
BioBERT

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%

PubMedBERT

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%

Fine-Tuning：Domain Specific training dataset



2nd 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.

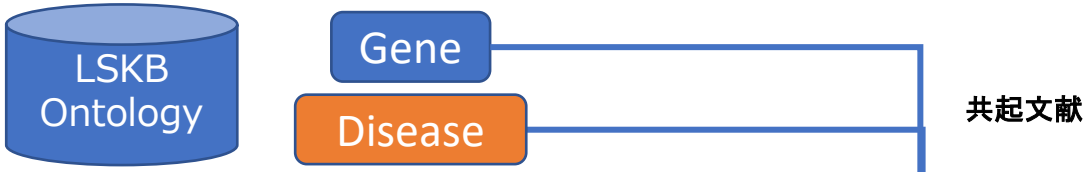
BioBERT

2nd	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%

PubMedBERT

2nd	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%

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



PMID	Abstract (抜粋)
23292288	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. 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。

Gene
Androgen receptor (AR)
Toxicity Term
hepatic steatosis

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

仮説（自動生成）

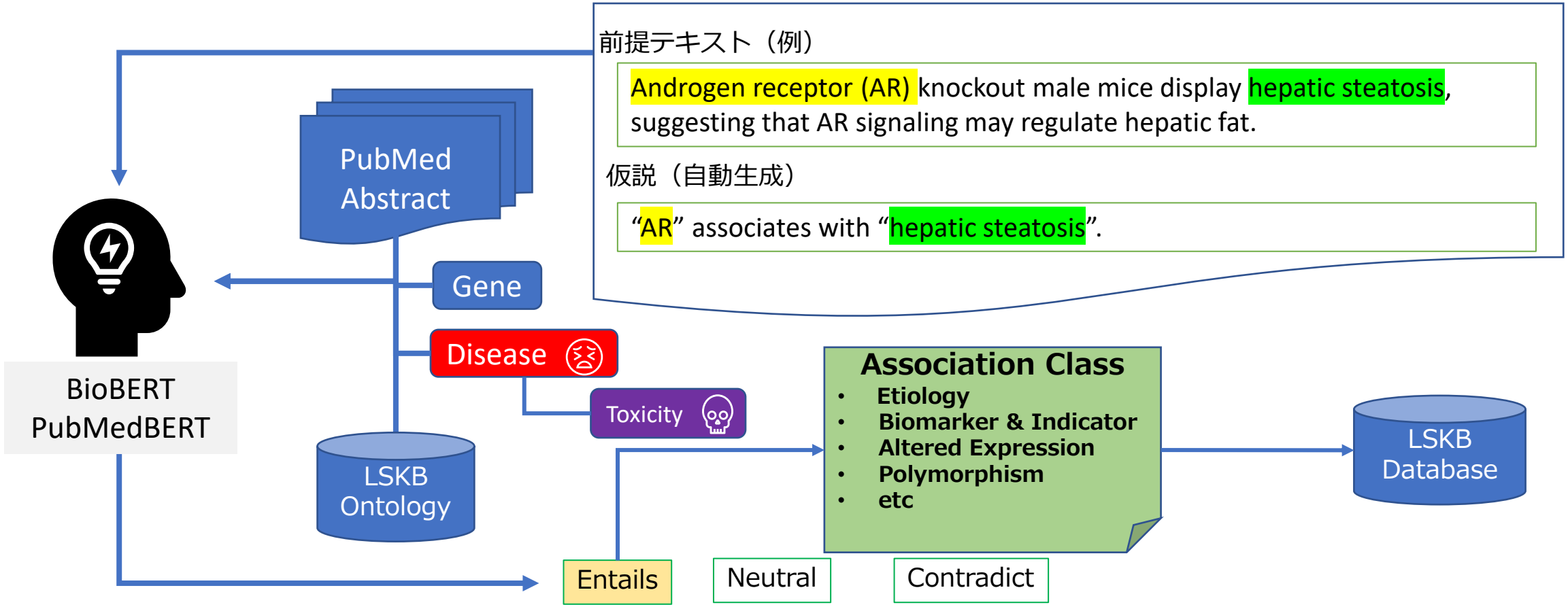
“AR” associates with “hepatic steatosis”.

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)

Knowledge Station LSKB

GeneRIF
搭載件数：1,589,973件

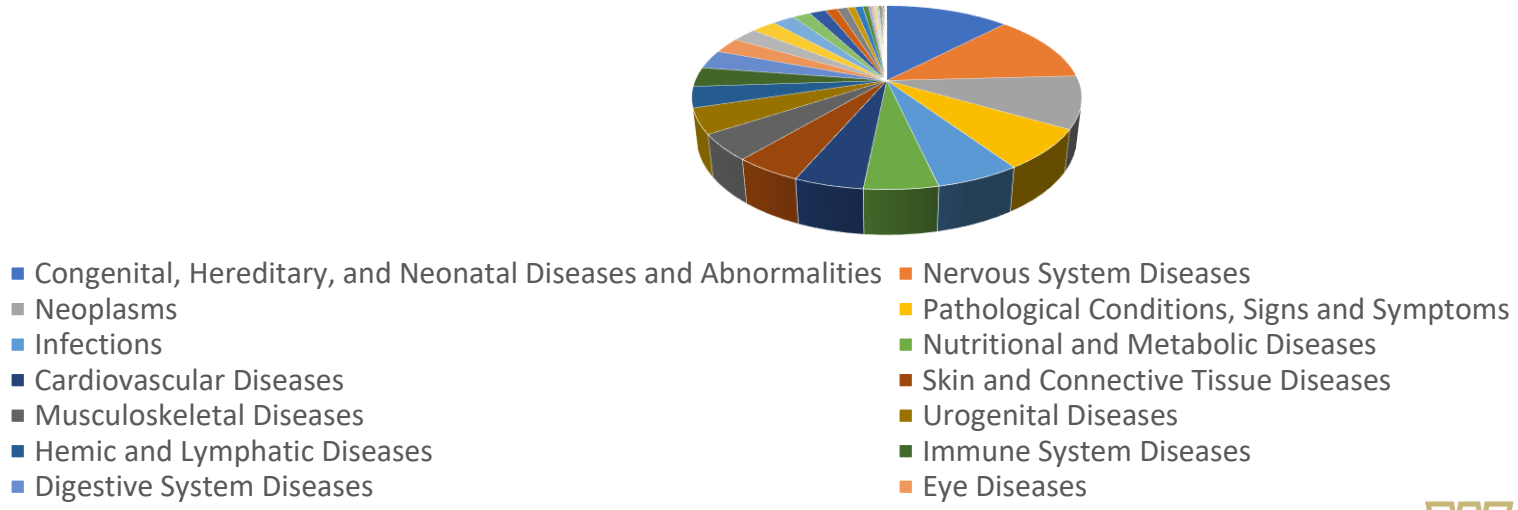
↓

疾患 21,681との
共起センテンス

↓

自動生成仮説とMNLI実行

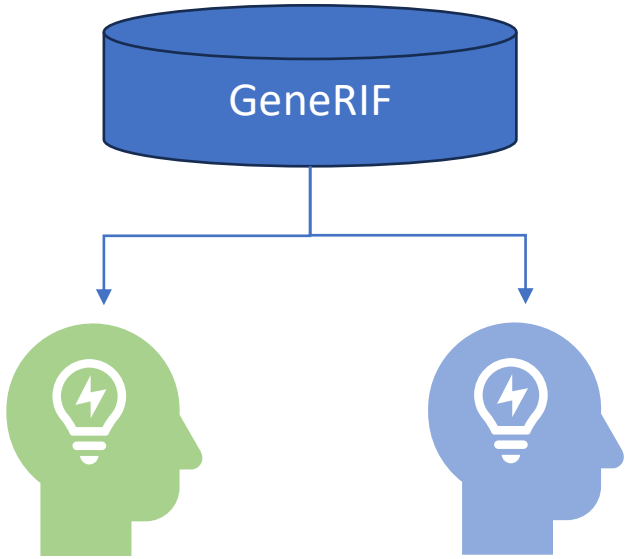
- GeneRIF
  - NCBIが提供する 遺伝子機能のアノテーション登録サイト
  - PubMedで引用されるPMIDと タイトル以外の簡潔なフレーズ(425文字以下)で１つまたは複数の機能を説明
- GeneRIF 搭載 108,378 遺伝子：ヒト (20,220 遺伝子)、マウス (14,305 遺伝子)、ラット (7,545 遺伝子)、植物 (9,781 遺伝子)
- 記載文章から LSKB Disease オントロジーを利用したテキストマイニングで疾患名を取得



World Fusion

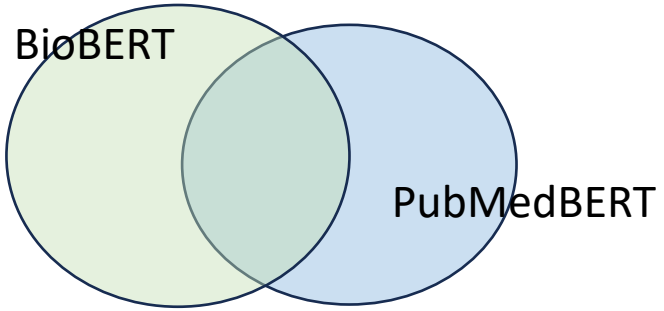
GeneRIF：遺伝子vs疾患のクラス分類

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**



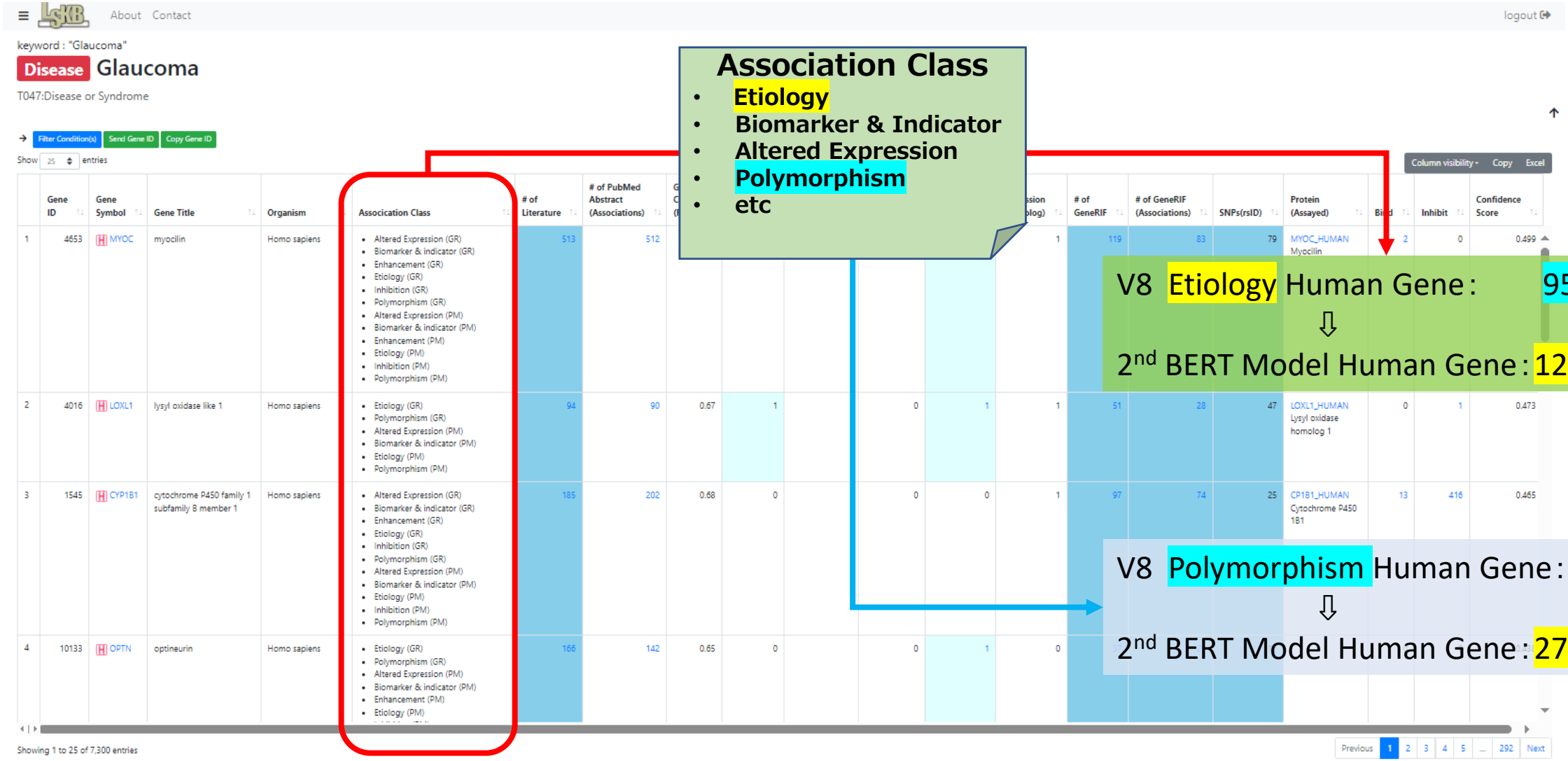
Etiology Class：BioBERT･PubMedBERT 及び 2回のFine-tuning 結果

BioBERT / PubMedBERT	# of Entailment				Accuracy	
	1st Model		2nd Model		1st Model	2nd Model
Both Entailment	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

2nd BERT Model による 疾患関連遺伝子分類

Glaucoma: associated genes from PubMed



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

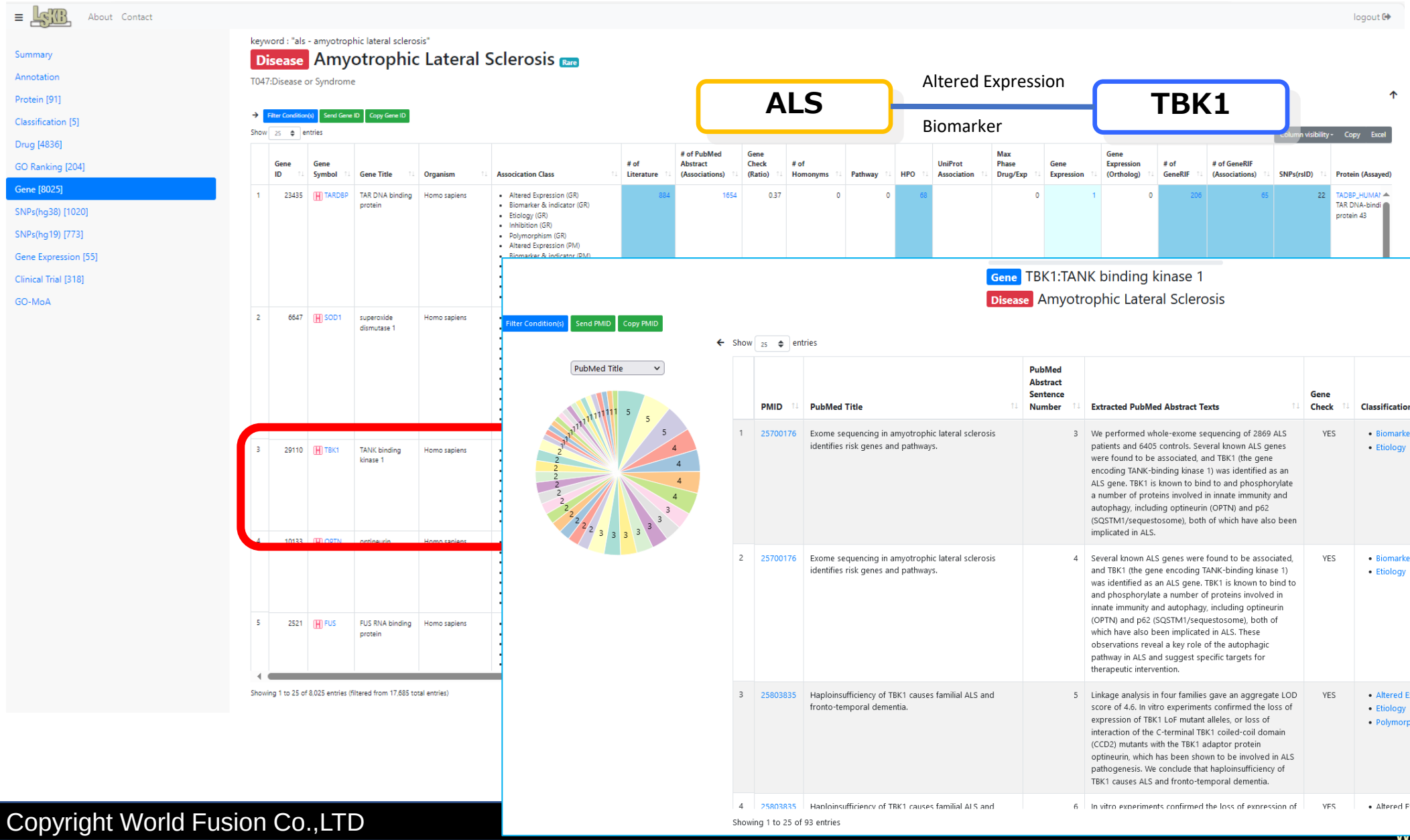
⇩

2nd BERT Model Human Gene : 277件

Newly discovered gene in Etiology Class by 2nd Fine-tuning Model

PMID	25620203	
Gene	**DDX58**	
Description	To identify genetic causes of the disease, we performed exome sequencing in this family and identified a variant (c.1118A>C [p.Glu373Ala]) of **DDX58**, whose protein product is also known as RIG-I. Further analysis of **DDX58** in 100 individuals with congenital **glaucoma** identified another variant (c.803G>T [p.Cys268Phe]) in a family who harbored neither dental anomalies nor aortic calcification but who suffered from **glaucoma** 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.	
	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 **DDX58** の変異体 (c.1118A>C [p.Glu373Ala]) を特定しました。 先天性**緑内障**患者 100 人を対象とした **DDX58** 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、**緑内障**と骨格の異常を患っていた家族に別の変異型 (c.803G>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。	
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



BioBERT と PubMedBERT の比較結果

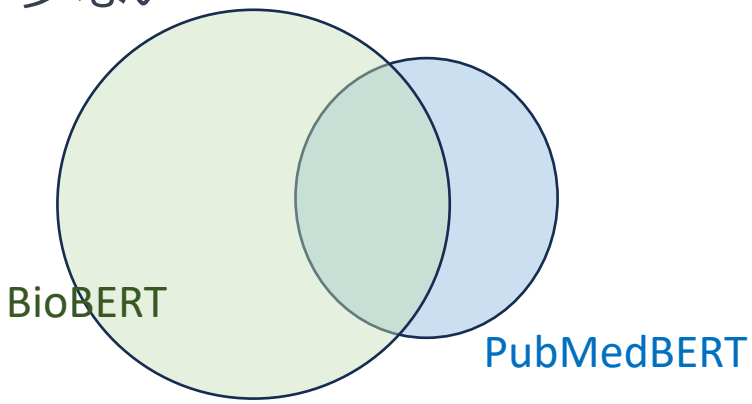
- **遺伝子 vs 疾患の11仮説に基づいた比較**

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

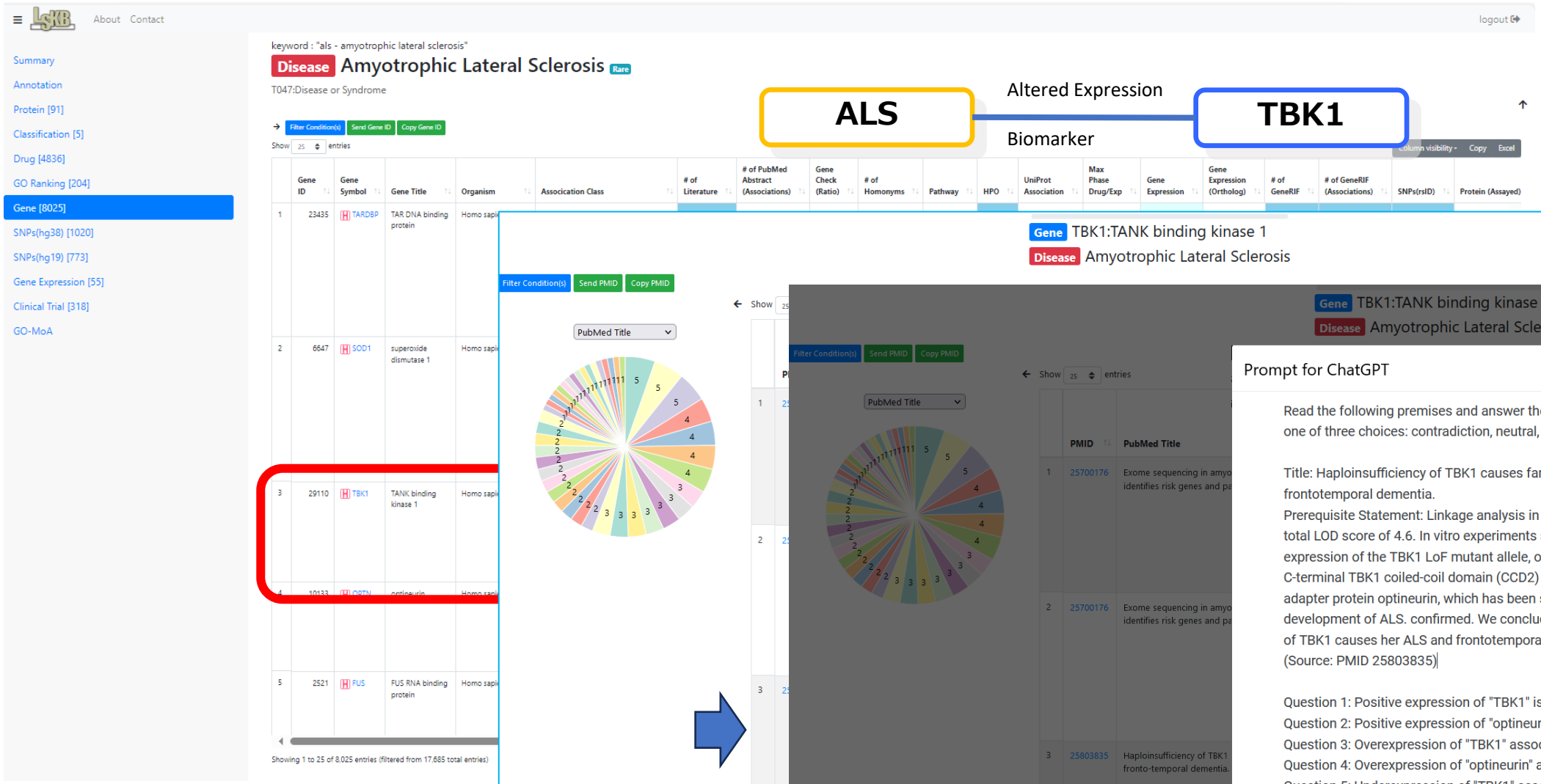
<検証>

- 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

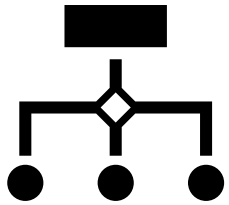


Life Science Knowledge Bank(LSKB)

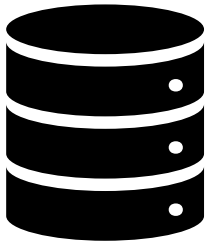
LSKB is scientific web

Knowledge Station LSKB

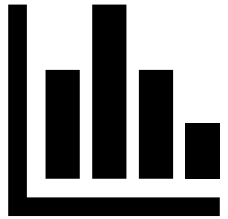
Ontology



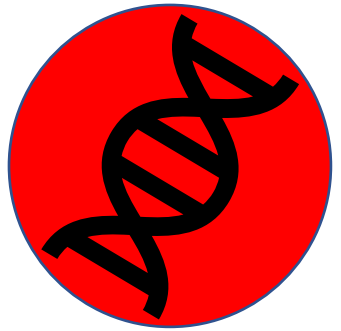
Contents



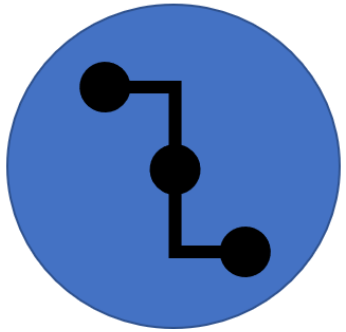
Analysis

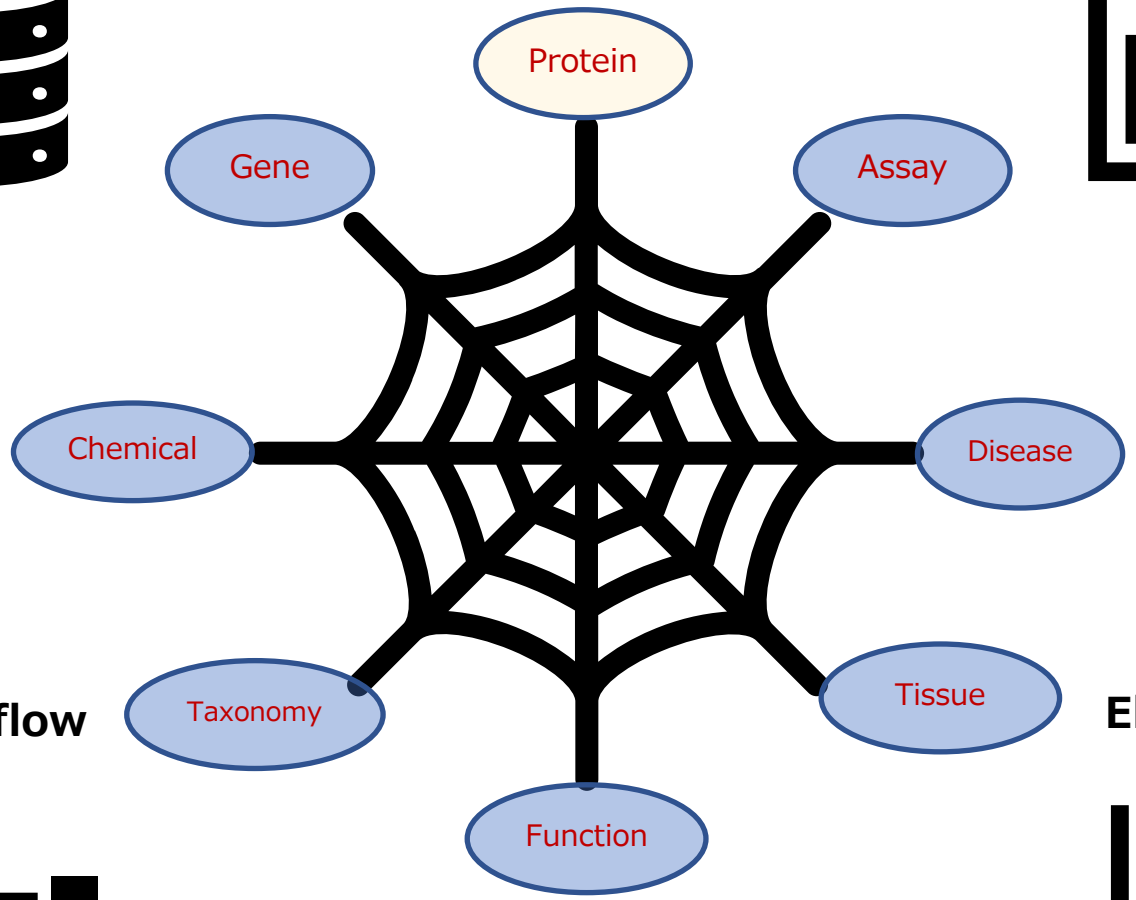


Bioinformatics

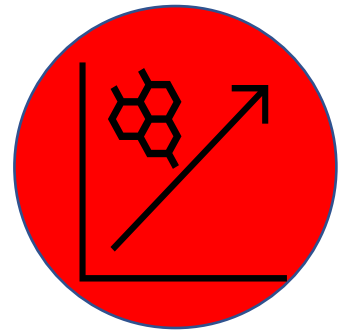


Interaction



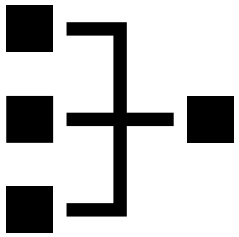


Cheminformatics

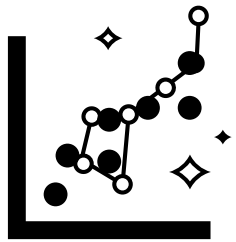


Structure

Workflow



Elpis Map



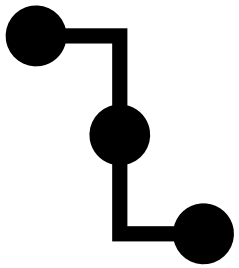
Exploration

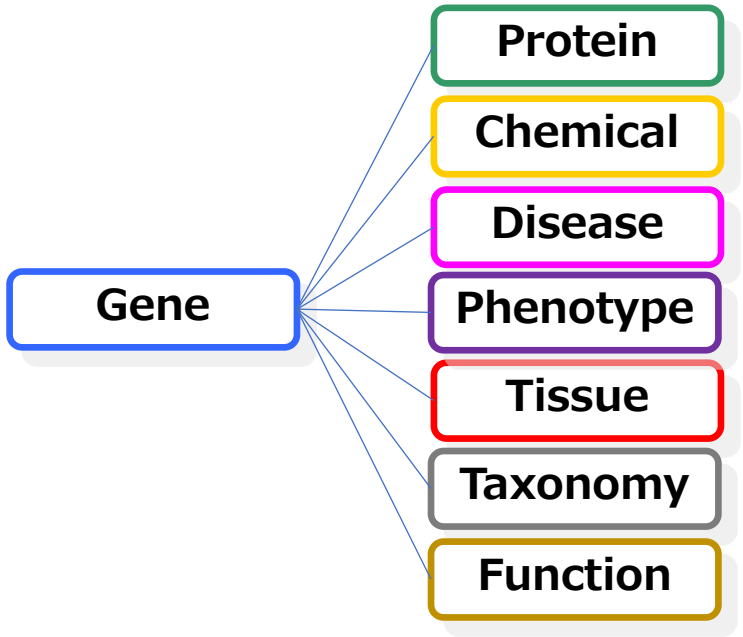
World Fusion

LSKB Interaction

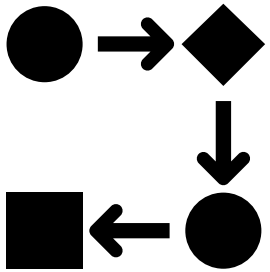
Knowledge Station LSKB

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

World Fusion

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

LSKB About Contact logout

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

Column visibility | Copy | Excel

	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	5	154	ADRB2	adrenoceptor beta 2	Homo sapiens	8	Etiology (GR) Polymorphism (GR) Biomarker & indicator (PM) Etiology (PM) Polymorphism (PM)	2	10	1.0	2	1	0	YES	0	0	0.03	0	4		1.86
6	6	348	APOE	apolipoprotein E	Homo sapiens	6	Altered Expression (GR) Etiology (GR) Polymorphism (GR) Biomarker & indicator (PM) Etiology (PM) Polymorphism (PM)	32	38	1.0	15	13	0		0	0	0.17	0	0		1.77
7	7	5737	PTGFR	prostaglandin F receptor	Homo sapiens	9	Etiology (GR) Polymorphism (GR) Biomarker & indicator (PM) Etiology (PM) Polymorphism (PM)	18	8	1.0	3	3	0	YES	0	0	0.17	0	4		1.76
8	8	4053	LTBP2	latent transforming growth factor beta binding protein 2	Homo sapiens	5	Etiology (GR) Polymorphism (GR) Biomarker & indicator (PM) Enhancement (PM) Etiology (PM) Polymorphism (PM)					6	2		0	0	2.02	0	0		1.69
9	9	134430	WDR36	WD repeat domain 36	Homo sapiens	7	Altered Expression (GR) Enhancement (GR) Etiology (GR) Inhibition (GR) Polymorphism (GR) Altered Expression (PM) Biomarker & indicator (PM) Enhancement (PM) Etiology (PM) Inhibition (PM) Polymorphism (PM)					2	2		0	0	4.7	0	0		1.68
10	10	4976	OPA1	OPA1 mitochondrial dynamin like GTPase	Homo sapiens	7	Altered Expression (GR) Biomarker & indicator (GR) Etiology (GR) Polymorphism (GR) Altered Expression (PM) Biomarker & indicator (PM) Enhancement (PM)	152	525	1.0	10	9	0		0	0	0.23	0	0		1.65

Association Class

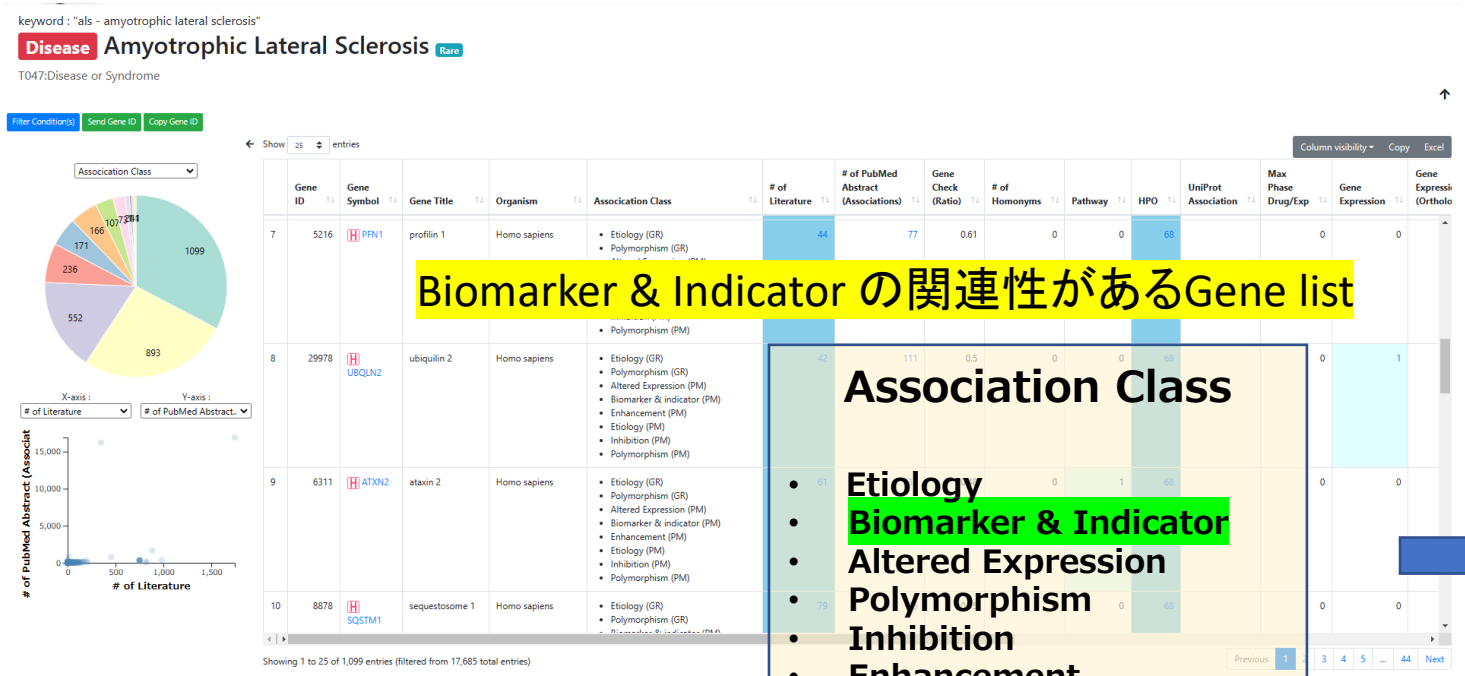
- Etiology
- Biomarker & Indicator
- Altered Expression
- Polymorphism
- Inhibition
- Enhancement

Showing 1 to 25 of 7,944 entries

Previous 1 2 3 4 5 ... 318 Next

Genetic Biomarker Exploration Example

Disease related Gene Information

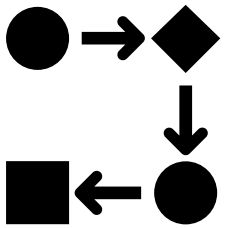


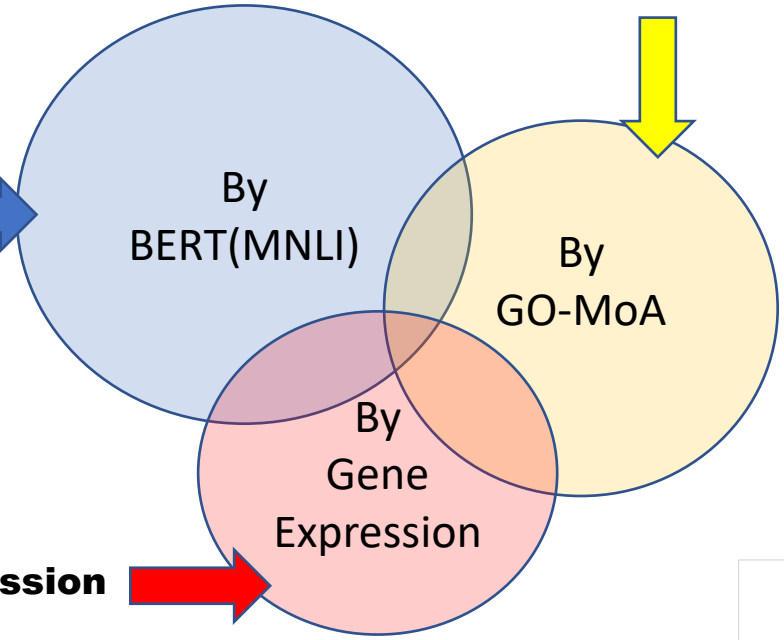
Biomarker & Indicator の関連性があるGene list

Association Class

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

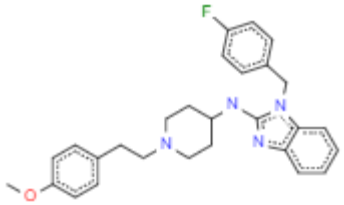
Genes from disease associated function By GO-MoA





Differential Gene Expression

Drug Repurposing Procedure

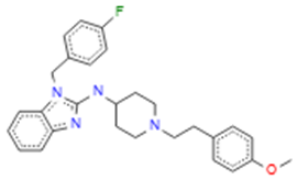


Activity Profile

Target Prediction

Expand Similar protein

Disease Explorer



Astemizole

Known/predicted targets

5HT2B_HUMAN	6.96
EGFR_HUMAN	7.09
ERBB2_HUMAN	7.13
FYN_HUMAN	6.12
KCNH2_HUMAN	5.85
LCK_HUMAN	7.13
ITA2B_HUMAN	6.14
ITAV_HUMAN	7.11
MK08_HUMAN	6.64
MK09_HUMAN	6.91
MK10_HUMAN	6.13
MK14_HUMAN	7.03
RAF1_HUMAN	6.93
NOS2_HUMAN	7.29
STAT3_HUMAN	5.35
CDK1_HUMAN	6.36
PAR1_HUMAN	6.41
PLD2_HUMAN	6.83
MRP1_HUMAN	6.7
JAK1_HUMAN	7.26

negative regulation of autophagy

Potent inducer of cellar autophagy

1)

Prostate cancer

Pathway

hsa05200	Pathways in cancer
WP2263	Prostate Cancer
hsa05223	Non-small cell lung cancer

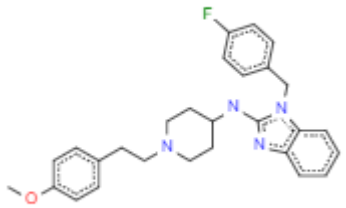
2)

Non-Small Cell Lung Carcinoma

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



Related Disease

Primary Target

Activity Profile

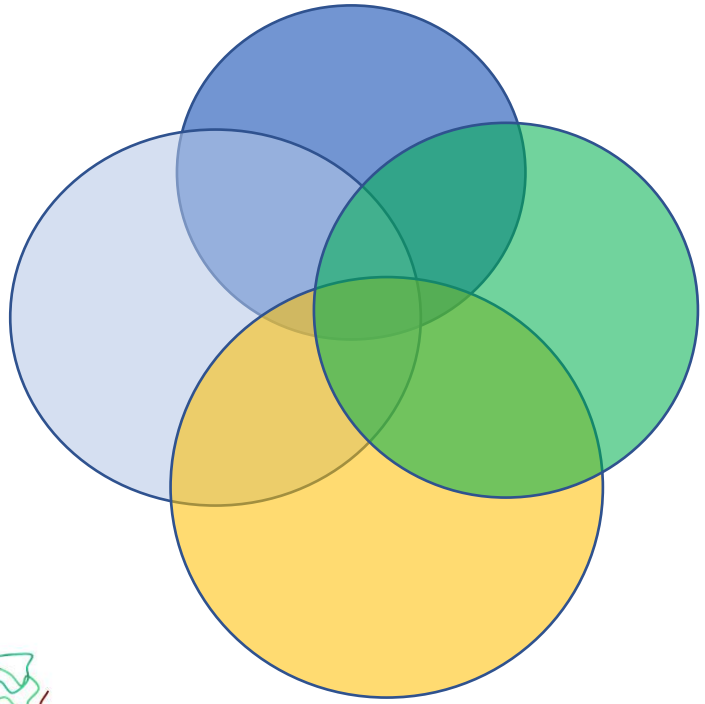
Disease by GO-MoA

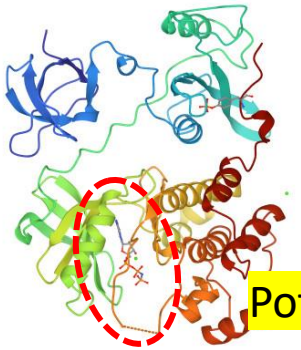
Disease Explorer

Similar Protein

Protein Pocket Similarity

Disease by Similar Protein

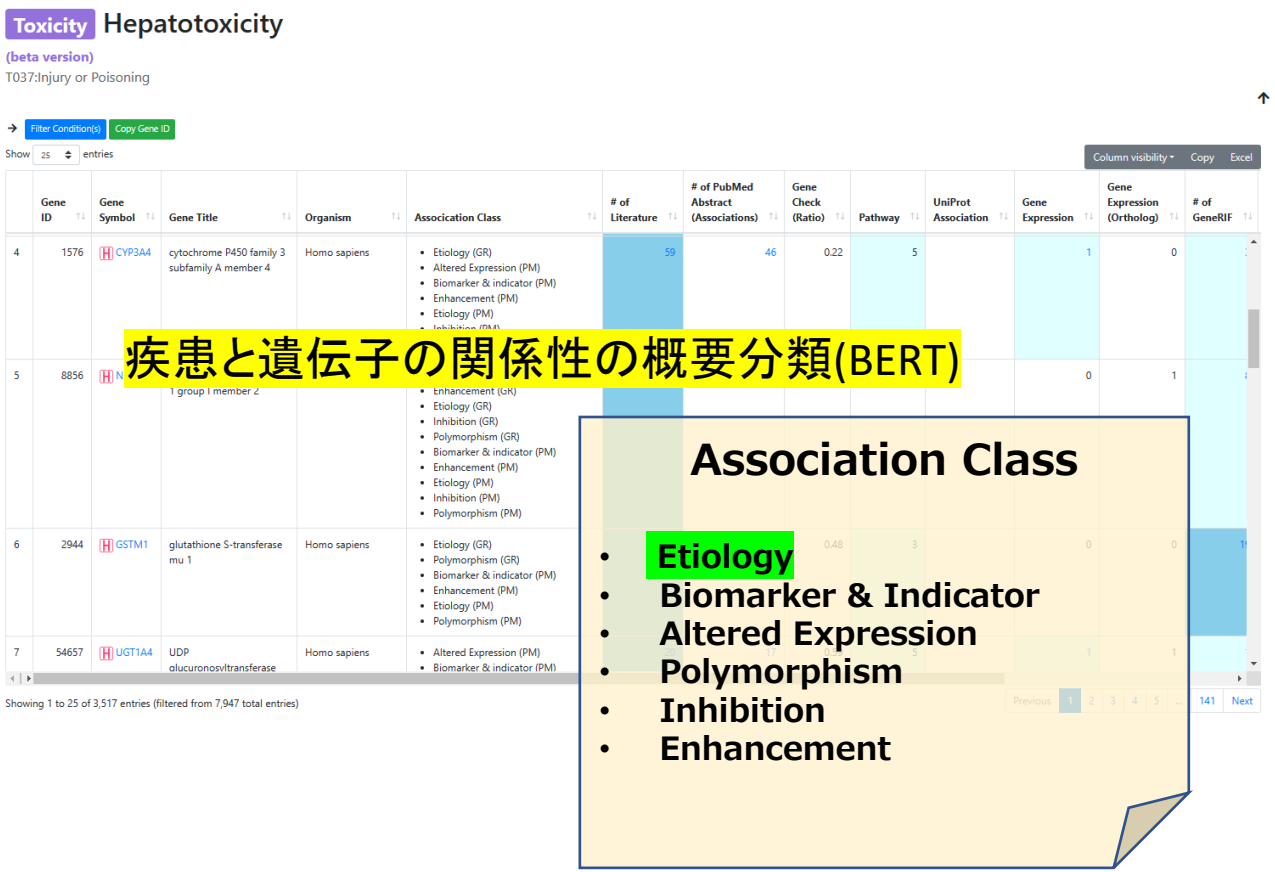




Potential target

Toxicity Information

Toxicity associated Targets

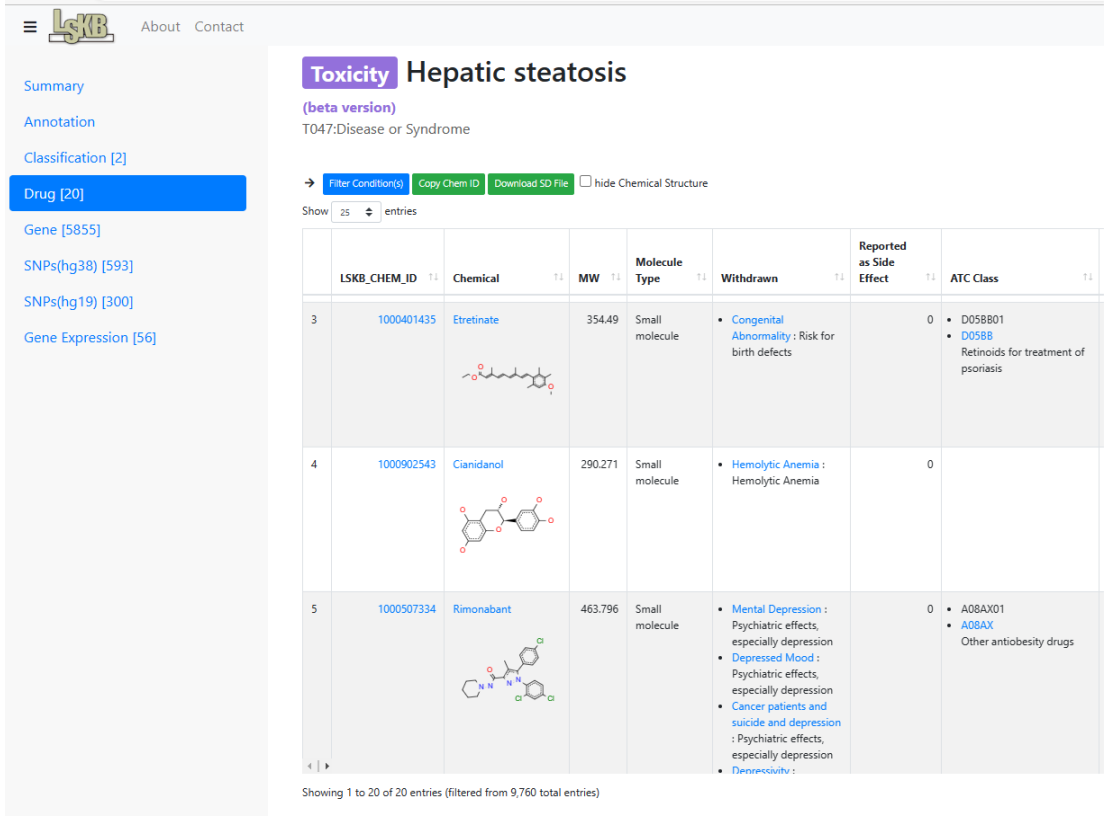


疾患と遺伝子の関係性の概要分類(BERT)

Association Class

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

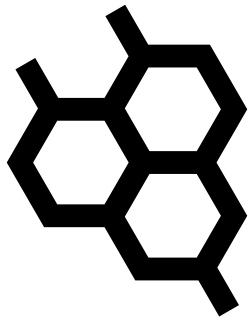
Drugs observed as side effects

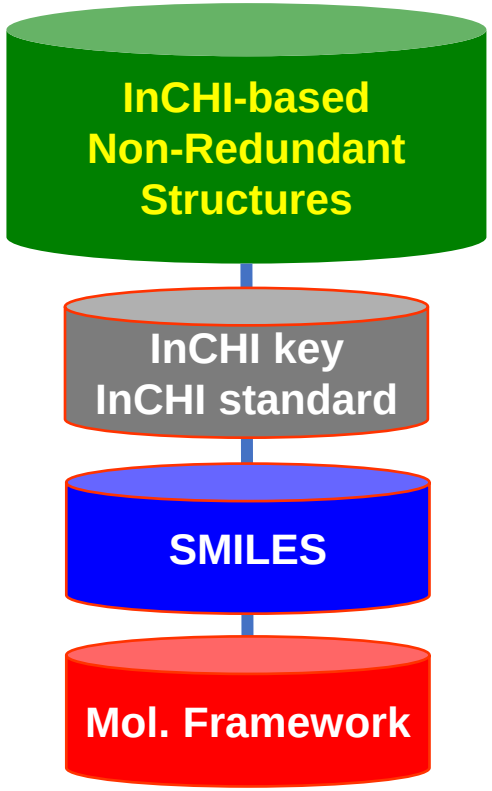


Chemical Structure

Knowledge Station LSKB

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

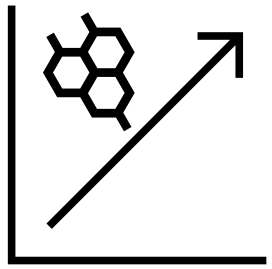
- Target Prediction
- Target Confirmation
- Mechanism Explorer

4) Misc

- Multiple structures search
- Automatic saving of search result

Knowledge Station LSKB

Cheminformatics



Target Prediction

Target Confirmation

Chemical Properties

Molecular Framework

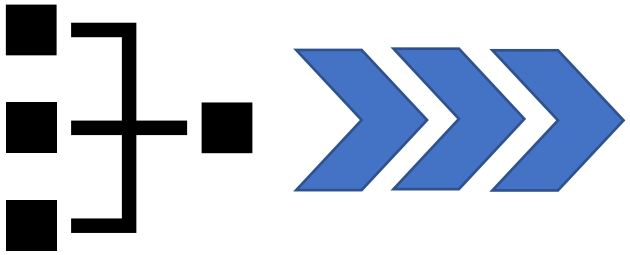
Go To MOE

Workflow

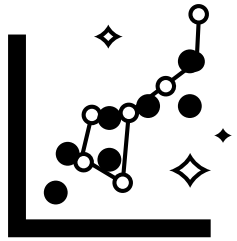
2. PIK3A/ PIK3B/ PIK3G/ PIK3Dの活性データ

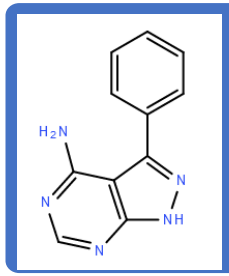
Knowledge Station LSKB

Workflow



Elpis Map





注目している部分構造を持つ化合物の検索

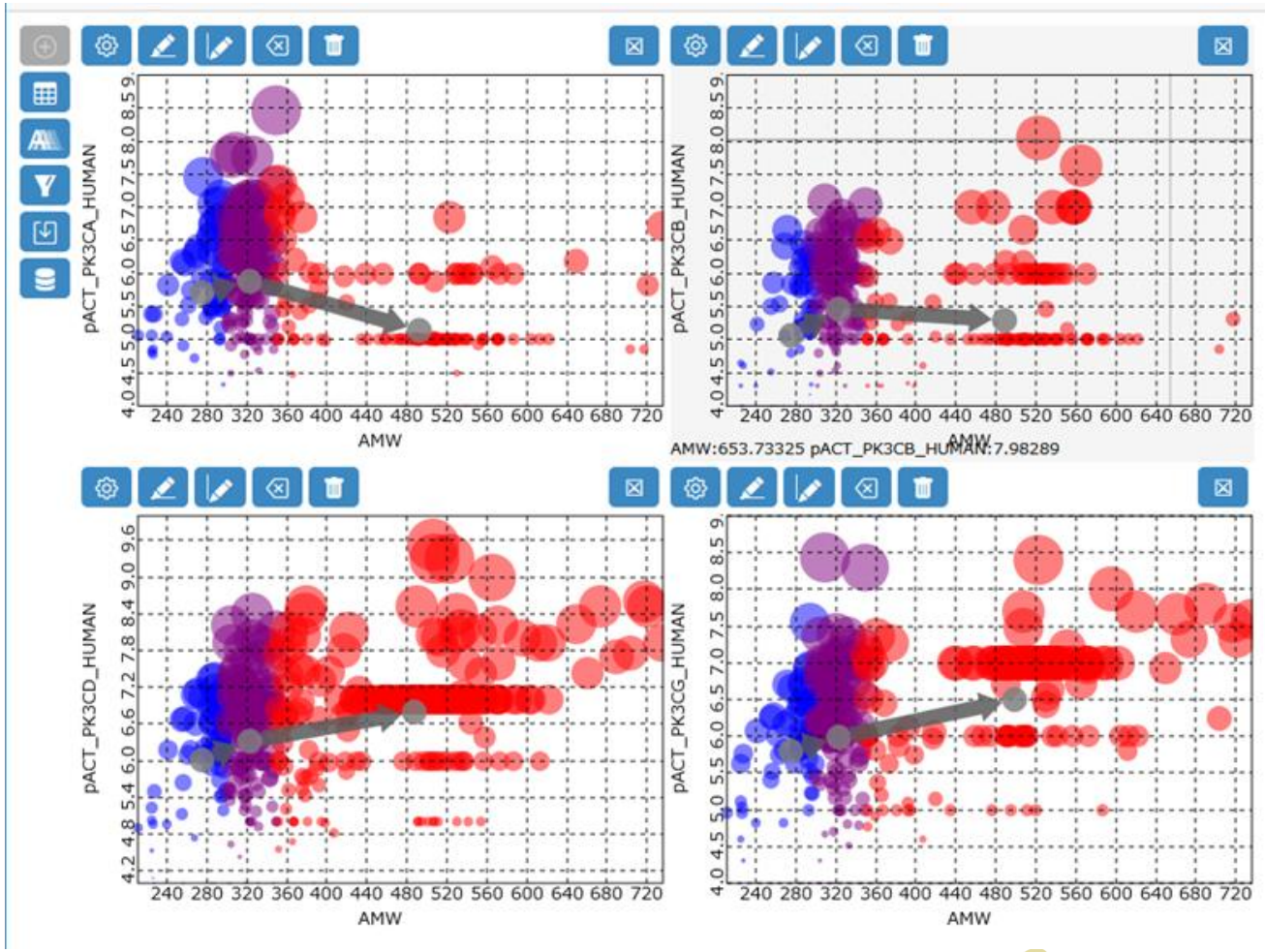
ターゲットたんぱく質の指定

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

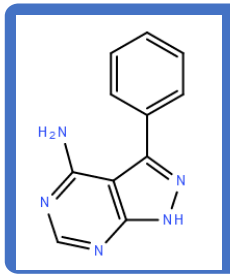
表形式ファイル、およびプロット図作成を指示



World Fusion

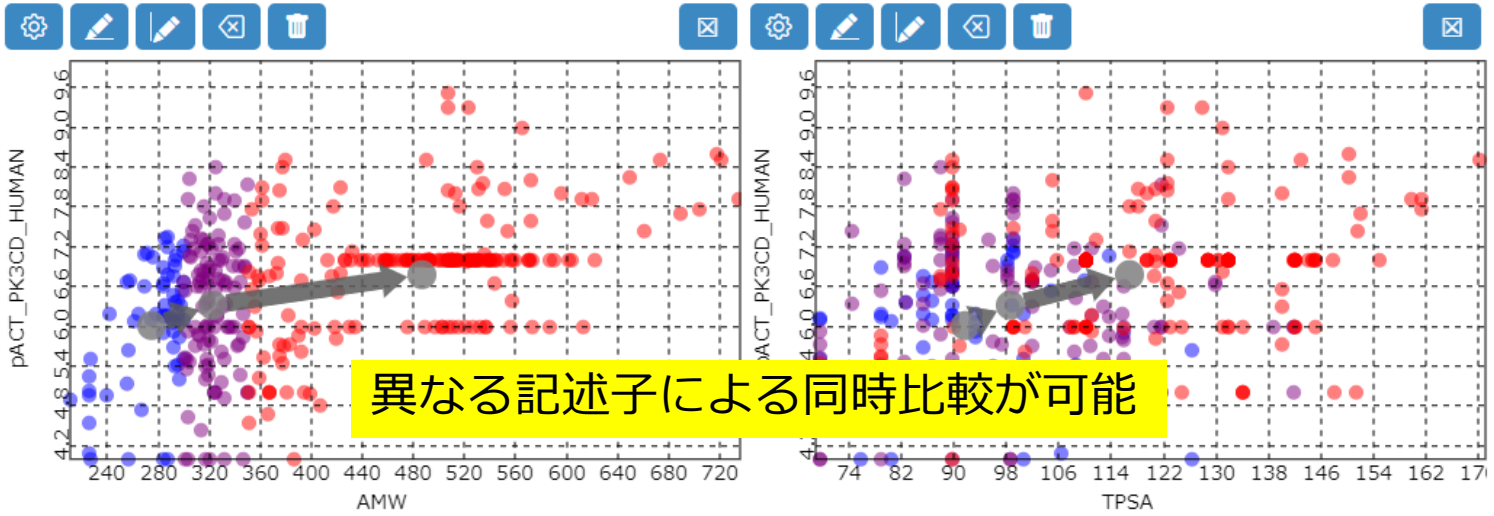
2. PIK3A/ PIK3B/ PIK3G/ PIK3Dの活性データ

Knowledge Station LSKB

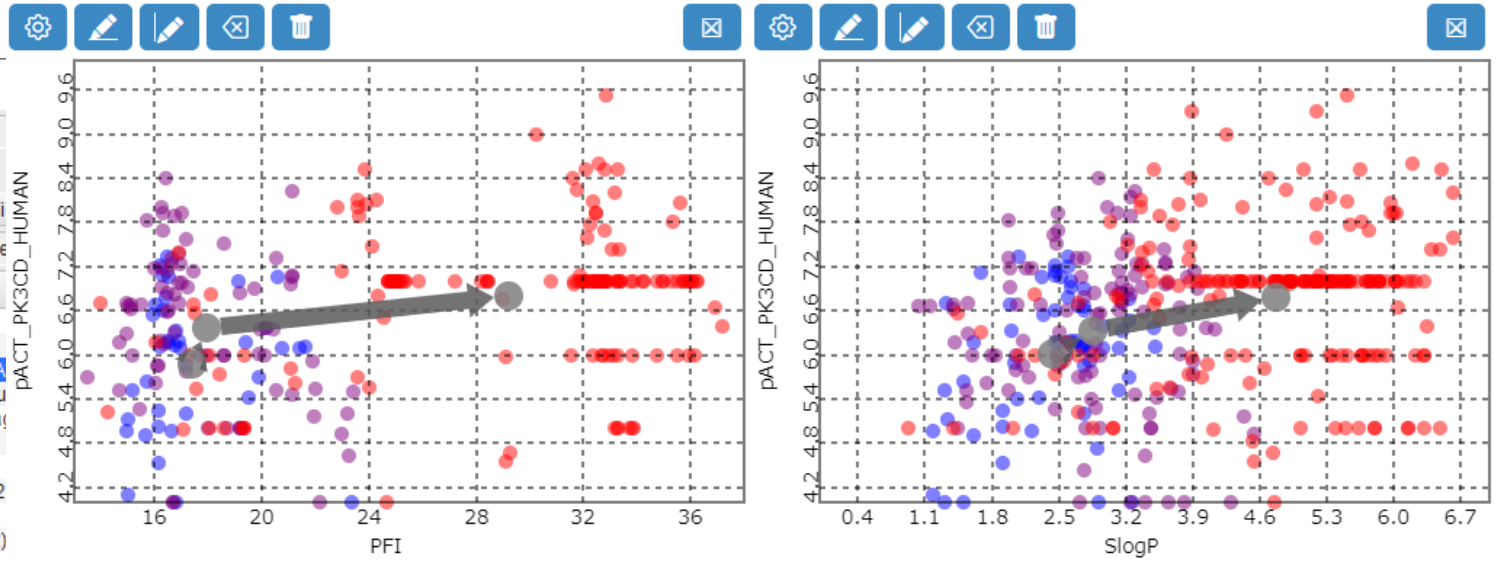


部分構造を持つ活性化合物

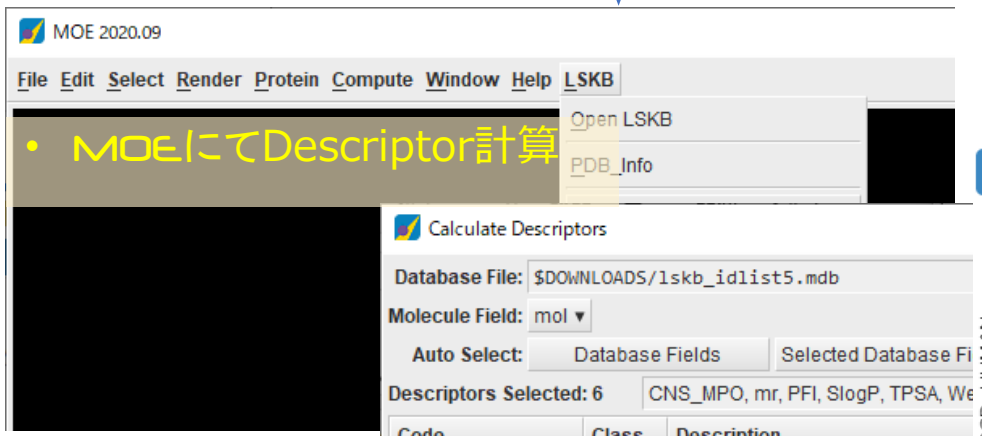
PIK3A/ PIK3B/ PIK3D/ PIK3Gの活性値

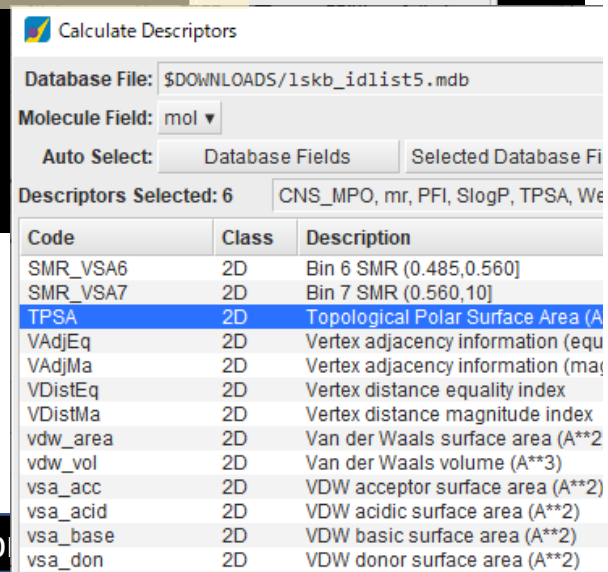


異なる記述子による同時比較が可能



- MOEにてDescriptor計算

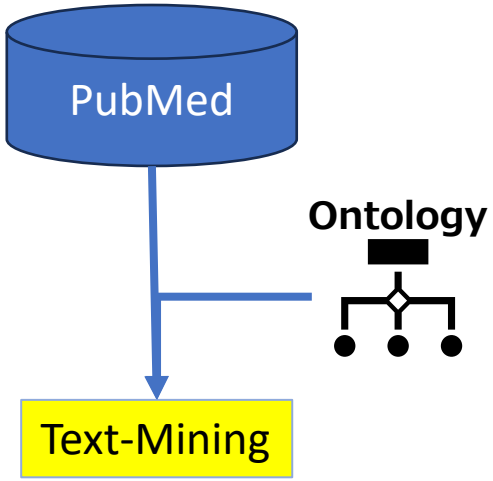




Code	Class	Description
SMR_VSA6	2D	Bin 6 SMR (0.485,0.560]
SMR_VSA7	2D	Bin 7 SMR (0.560,10]
TPSA	2D	Topological Polar Surface Area (A
VAdjEq	2D	Vertex adjacency information (equ
VAdjMa	2D	Vertex adjacency information (ma
VDistEq	2D	Vertex distance equality index
VDistMa	2D	Vertex distance magnitude index
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

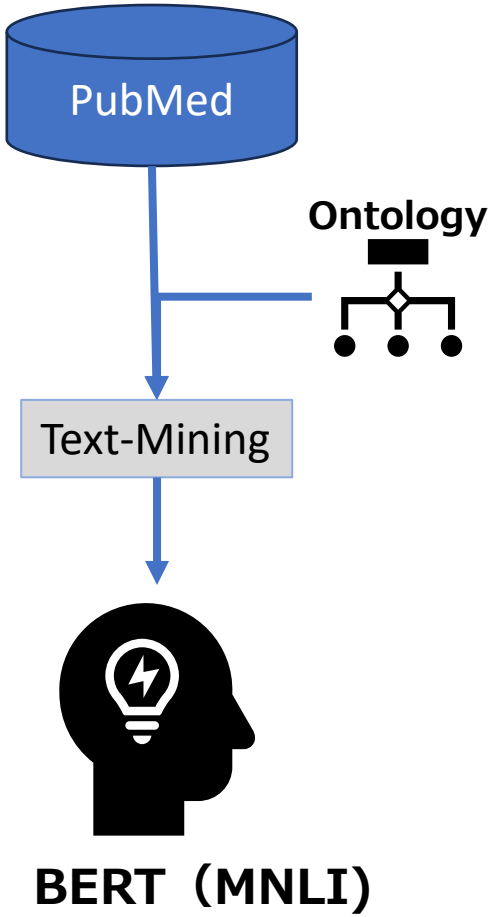
文献情報解析



- 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 sigma non-opioid intracellular receptor 1 gene (SIGMAR1) has been linked to autosomal recessive dHMN with pyramidal signs in several families. This phenotype can mimic amyotrophic lateral sclerosis (ALS).
Lv. 1	遺伝子名 と 疾患名 が 同一センテンスに記載
	UNC13A variant rs12608932 is associated with increased risk of amyotrophic lateral sclerosis 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
	Bortezomib induces a potent cytocidal effect with caspase-3 activation.

文献情報解析

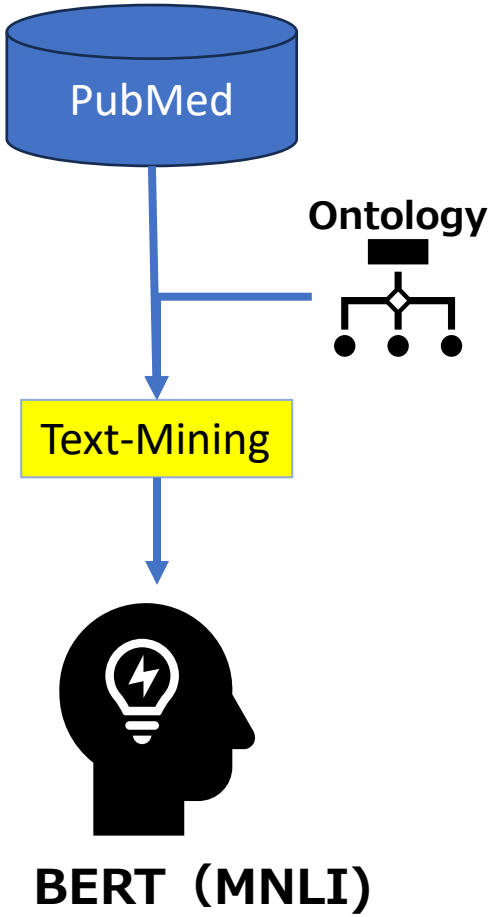


- BERT（MNLI）による記載内容の分類

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

Gene	Disease	
UNC13A	amyotrophic lateral sclerosis	
		Association Class
		Polymorphism
		Association(Etiology)

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

- 1st Generation Hypothesis
- Altered Expression Class
  - positive expression of "gene" is the risk of "disease“
  - negative expression of "gene" is the risk of "disease"

Association Class

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

Disease vs Gene 11 種の仮説

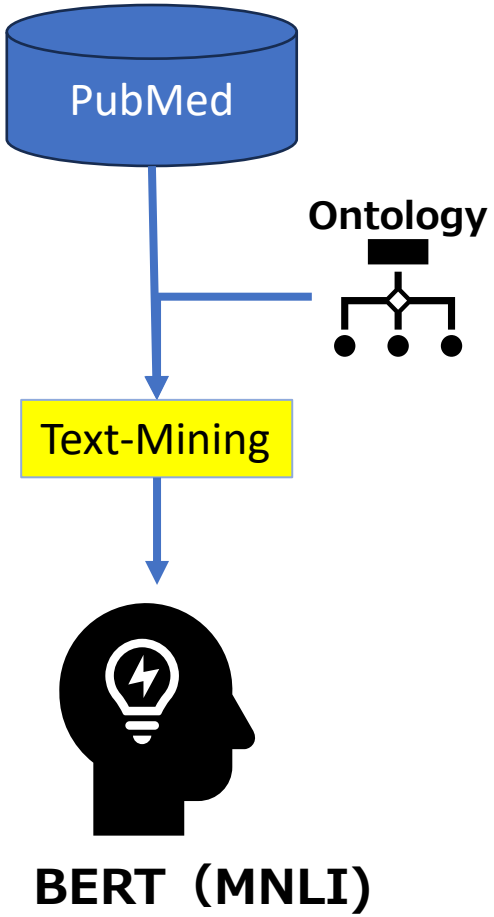
- 2nd Generation Hypothesis
- Altered Expression Class
  - Overexpression of "gene" associates "disease“
  - Underexpression of "gene" associates "disease"

Disease vs Gene
20 種の仮説

Association Class

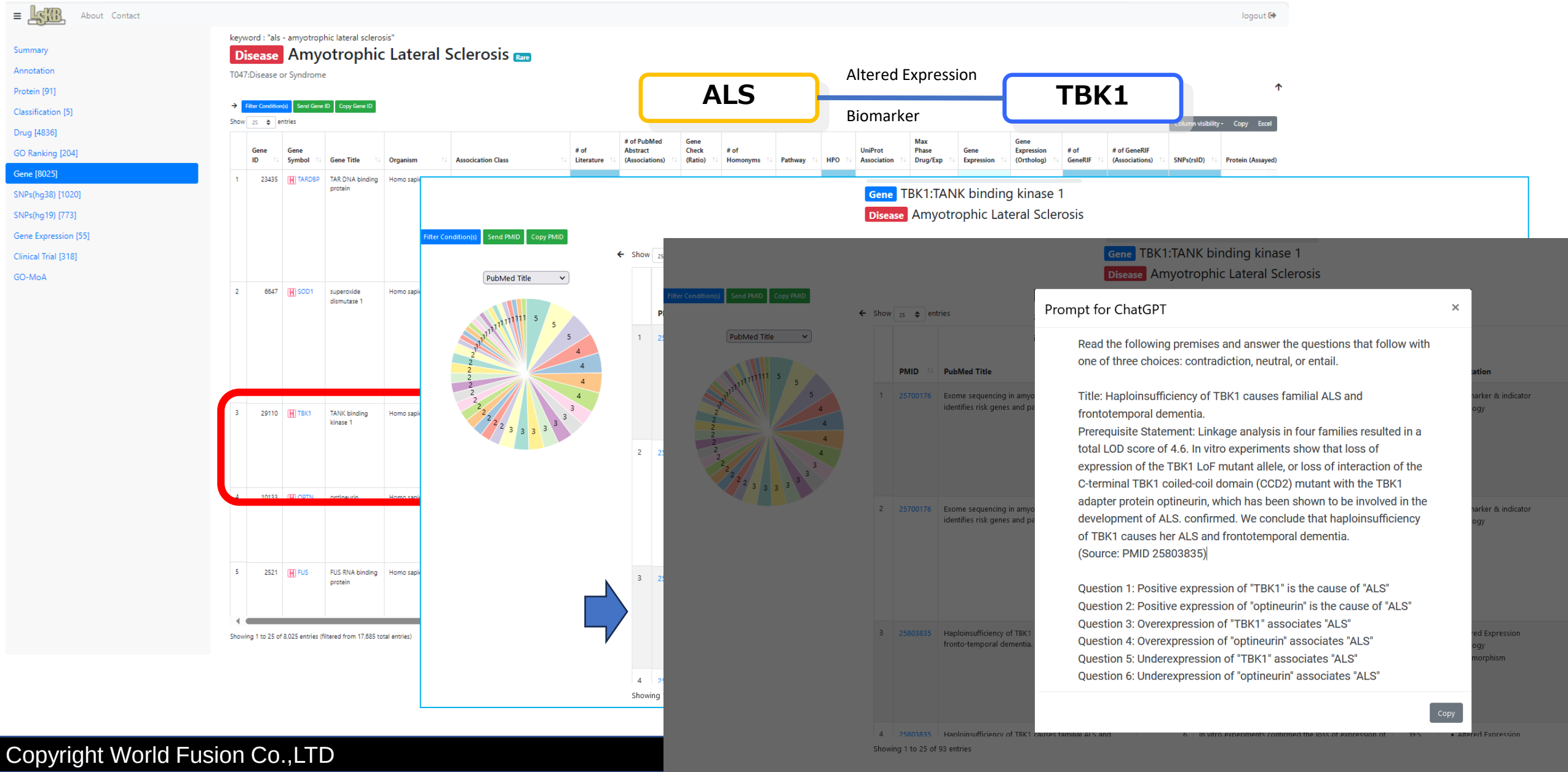
- **Etiology**
- **Biomarker**
  - **Diagnostic Biomarker**
  - **Prognostic Biomarker**
  - **Safety Biomarker**
- **Altered Expression**
  - **Overexpression**
  - **Updexpression**
- **Polymorphism**
- **Overactivation**
- **Underactivation**
- **Genetic Regulation**

文献情報解析：アップデート

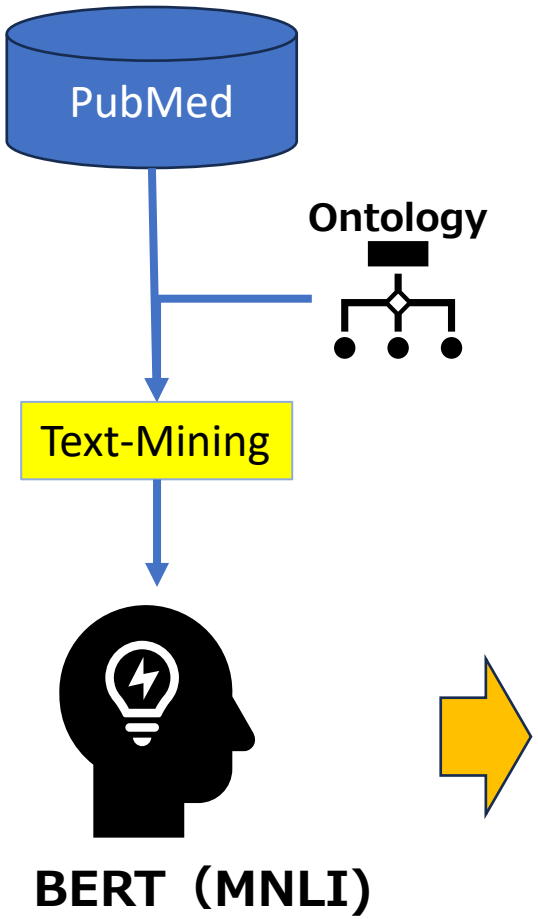


- 文献のテキストマイニングのウィークリーアップデート
- MNLIによるクラス分類のマンスリーアップデート
  - 遺伝子 vs 疾患（毒性）
- MNLIの Fine-tuning アップデートモデルによるデータ更新
- MNLIの仮説の増強と更新
- BioBERT/PubMedBERTでの検証結果の搭載
- ChatGPTの検証クエリー生成

ChatGPTの検証クエリー生成



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。

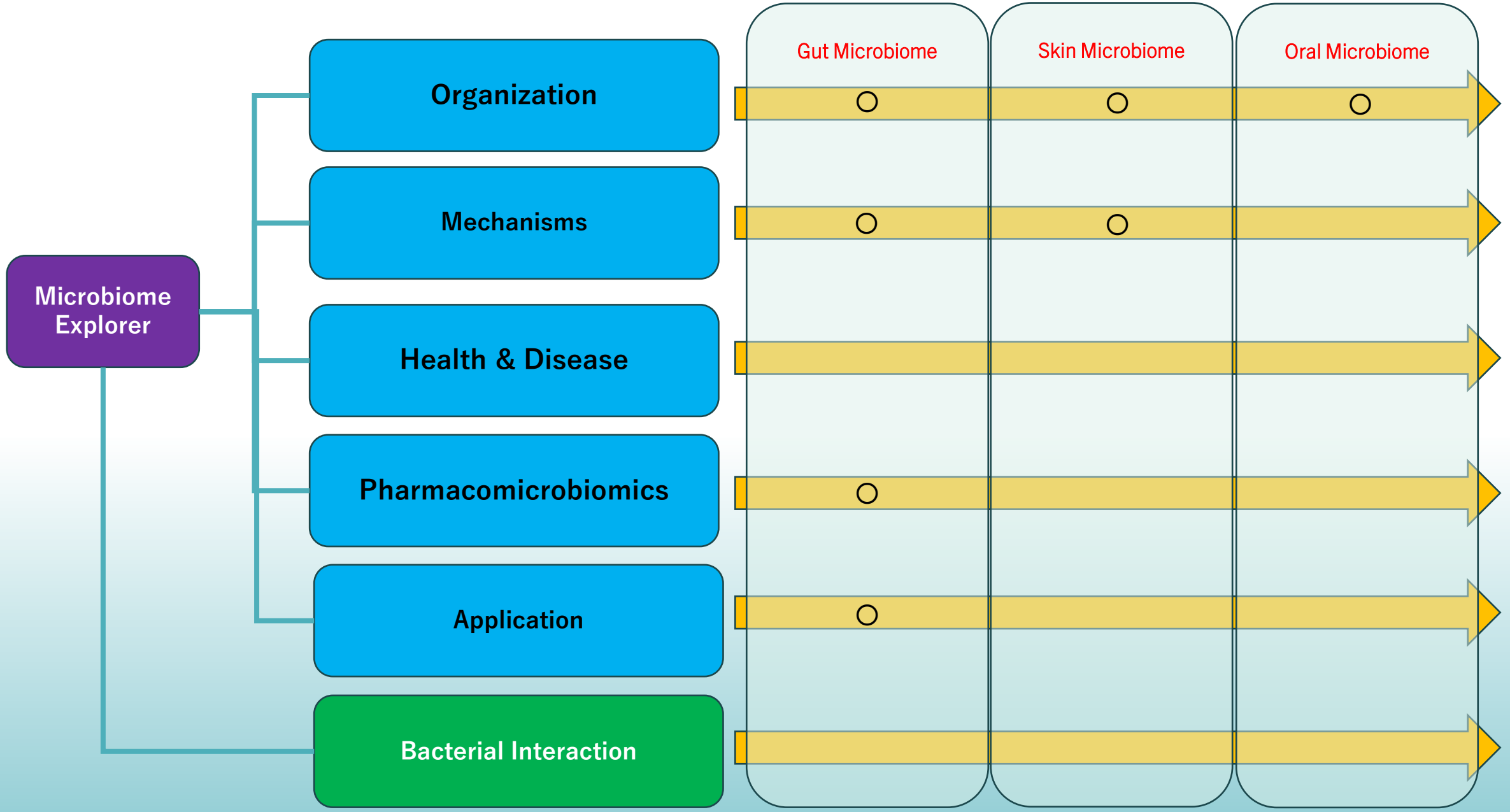
おすすめ論文のPick-Up

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

World Fusion

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
までお願いいたします。

https://www.lskb.jp/