



# SciBite

## Phenotype Triangulation and Beyond with ChemAxon

*Monday, March 25, 2019*

**Adam Brown, PhD**

Technical Sales Manager, North America, SciBite

[adam@scibite.com](mailto:adam@scibite.com)

THE LANGUAGE OF SCIENCE



# SciBite's Purpose

To enable scientists to use insights locked in **unstructured data** to power their decision and speed up innovation by:

Combining **world class ontologies & machine learning** to revolutionise the access to and utilisation of scientific information

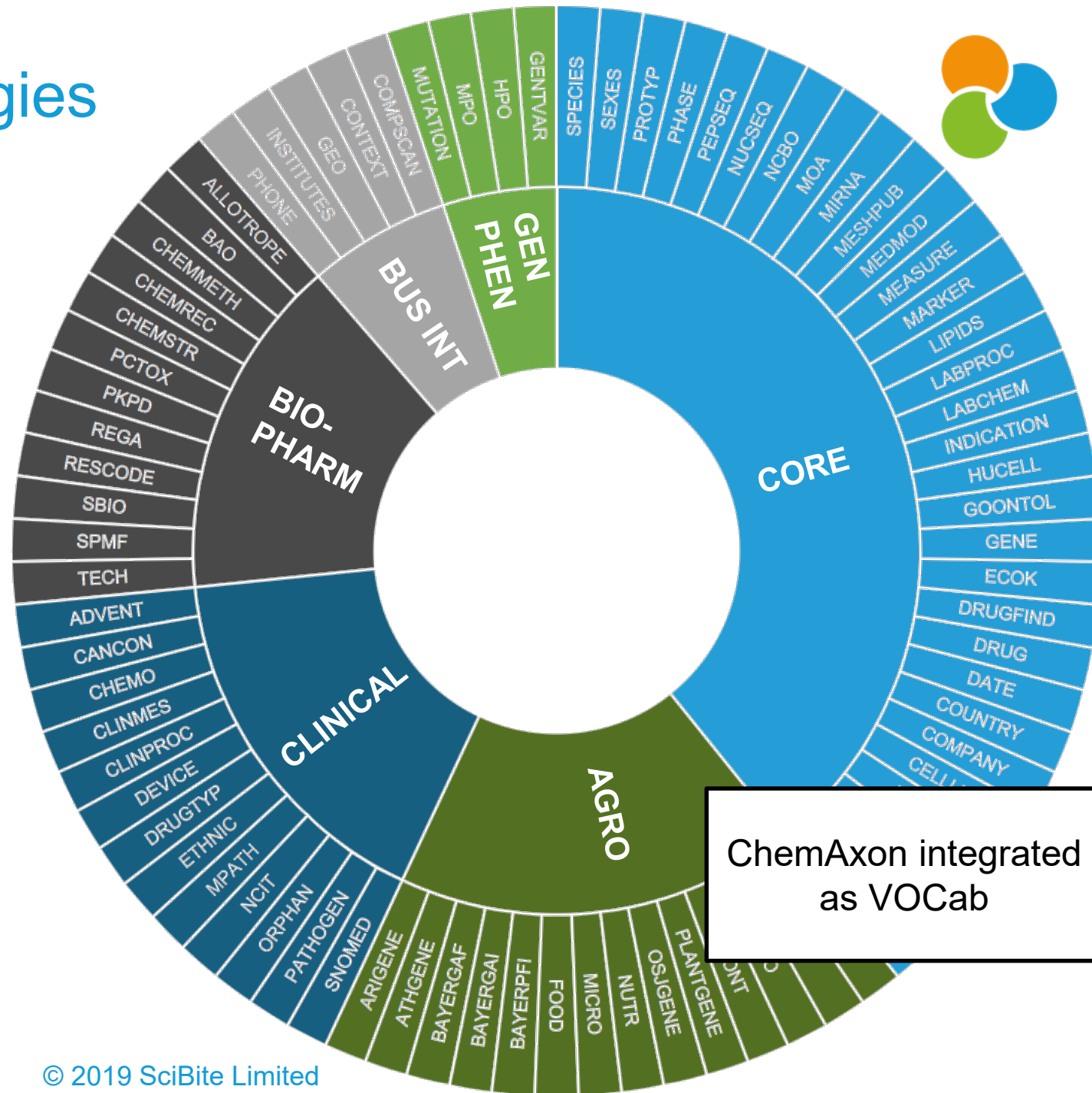
Transforming unstructured text into **contextualised, machine readable data** suitable for new discovery



# VOCabs : Beyond Ontologies



- Ontology content tuned for text mining
- Maintained by our dedicated team using expert curation and machine learning
- Comprehensive coverage
- Aligned to industry standards to maintain interoperability
- Enriched with synonyms and rules to manage the complexity of scientific language
- Customize, augment our existing or deploy your own vocabularies



# ChemAxon Integration - Example



ChemAxon lookup service seamlessly integrated into TERMite, SciBite's Biomedical Named Entity Recognition tool

## Explore TERMite [Show/Hide]

Just Type Below! *(Example)*

*Ethylene glycol increases expression of IL-6, which is implicated in alzheimer's and inflammation in mouse models.*

⊗ Reset    ⚙ Settings

Drop files here to upload

## As-You-Type Results

### Summary

Ethylene glycol increases expression of IL-6, which is implicated in alzheimer's and inflammation in mouse models.

### Termite Hits

Poor quality (ambiguous) hits shown in red/grey and should be treated with much caution, often users filter these out of their results as they are of low accuracy.

Type	Name	ID	Hit Synonyms	Poor
CA_LOCAL	ethylene glycol	ethylene glycol	Ethylene glycol	N
GENE	interleukin 6	IL6	IL-6	N
INDICATION	Alzheimer Disease	D000544	alzheimer's	N
INDICATION	Inflammation	D007249	inflammation	N
SPECIES	Mice	D051379	mouse	N

# TExpress: Pattern annotation powered by TERMite



- TExpress provides a UI and regex-like interface for finding patterns of scientific terms
- Examples include:
- Find sentences that include both a GENE and an INDICATION, separated by a BIOVERB

:(GENE) :{0,5} :(BIOVERB) :{0,5} :(INDICATION)

Dynamic pattern ('**pattern**'):  [\[entity lookup\]](#)  
Use the [Pattern Builder](#) to enter your pattern or type it manually above.

+

TExpress Pattern Builder [Save & Close](#) | [Test Pattern](#) | [Delete Last](#) | [Clear All](#)

Current Pattern (none yet)

Add: [Type](#) | [Individual\(s\)](#) | [Spacer](#) | [Word Bag](#) | [Taxonomy](#) | [Word List](#) | [Regex](#) | [Catcher](#) | [Open Any Order](#) | [Manual Entry](#)

[Help & manuals](#)

=

[Impaired expression of glutathione peroxidase-4 gene in peripheral blood mononuclear cells: a biomarker of increased breast cancer risk.](#)

Summary

Breast cancer aetiology is unclear despite comprising approximately 28% of female cancers. Several risk factors are known. Not all women exhibiting established risk factors will develop breast cancer but many without recognised risk factors will, indicating involvement of unknown risk factors. Impaired basal or oxidation-stimulated gene expression of redox enzymes, particularly [Glutathione Peroxidase 1 and 4 \(GPX1 and 4\)](#), resulting in increased oxidative stress, could be an "unknown" risk factor in breast cancer. We determined whether basal expression of [GPX1 and 4, two major redox enzymes, in Peripheral Blood Mononuclear Cells \(PBMC\) and/or their stimulated expression \(oxidative stress\) was impaired in women with breast cancer](#) who have no known markers of risk compared with control women without breast cancer. A significant 30% impairment (p< 0.01) in basal PBMC GPX4, but not GPX1, gene expression was observed in cancer patients. Oxidative stress stimulation in vitro did not increase [GPX4 expression significantly in cancer patients or control women whereas GPX1 expression was significantly increased \(30%, p< 0.05\) only in the cancer group](#). Attenuation of [GPX4 mRNA expression in PBMC suggests this could be a simple,early biomarker for future breast cancer risk](#) in the high proportion of women without known risk factors who eventually contract the disease.



# Phenotype Triangulation

Identifying new research candidates based on  
mechanistic similarities



# Phenotypic Triangulation



## The Problem

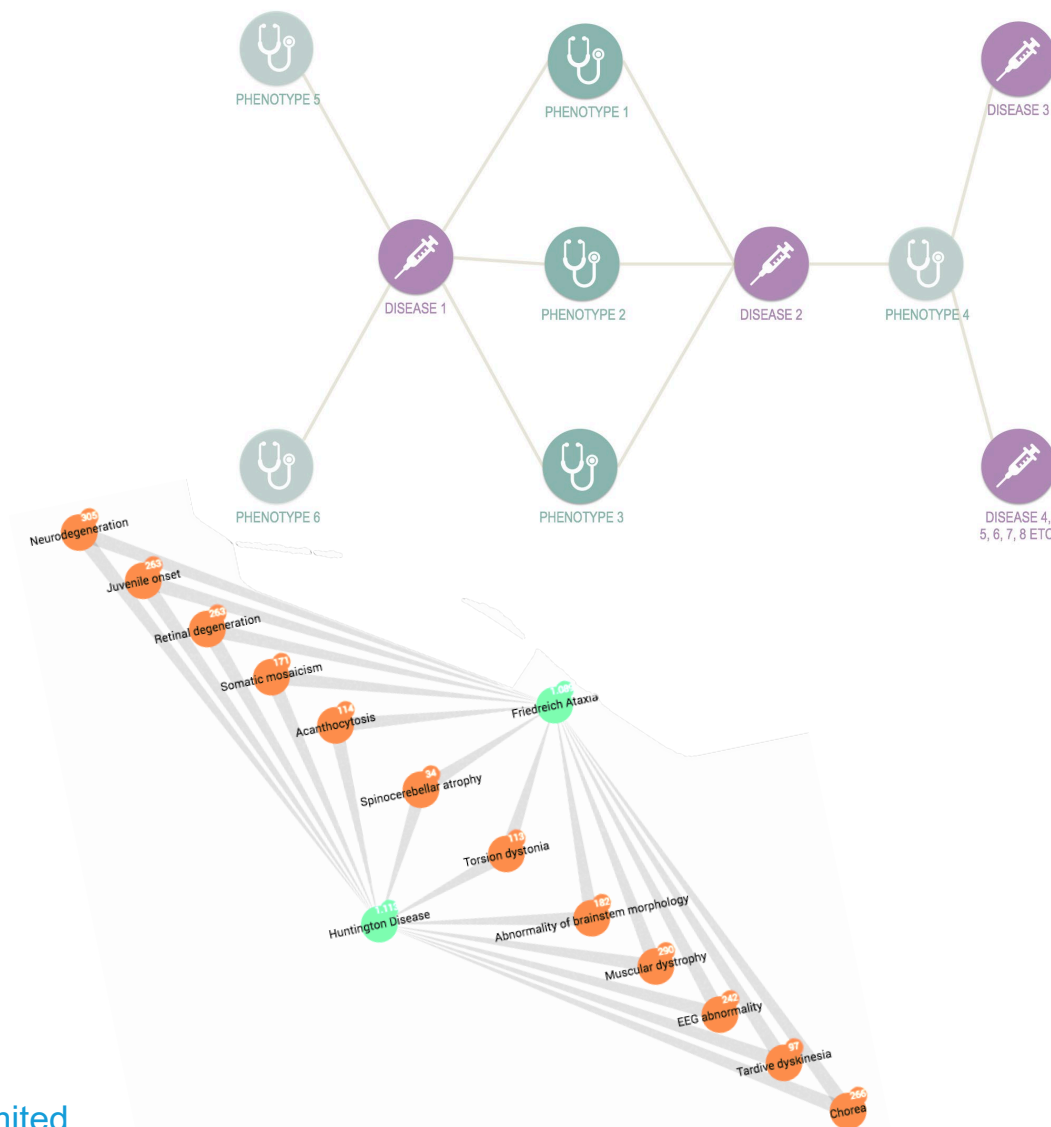
- Many diseases are understudied and lack clear molecular mechanisms
- Some entities (e.g. Phenotypes) are highly synonymous and difficult to standardise
- Scraping, standardising, and analysing research is time-consuming

## The Solution

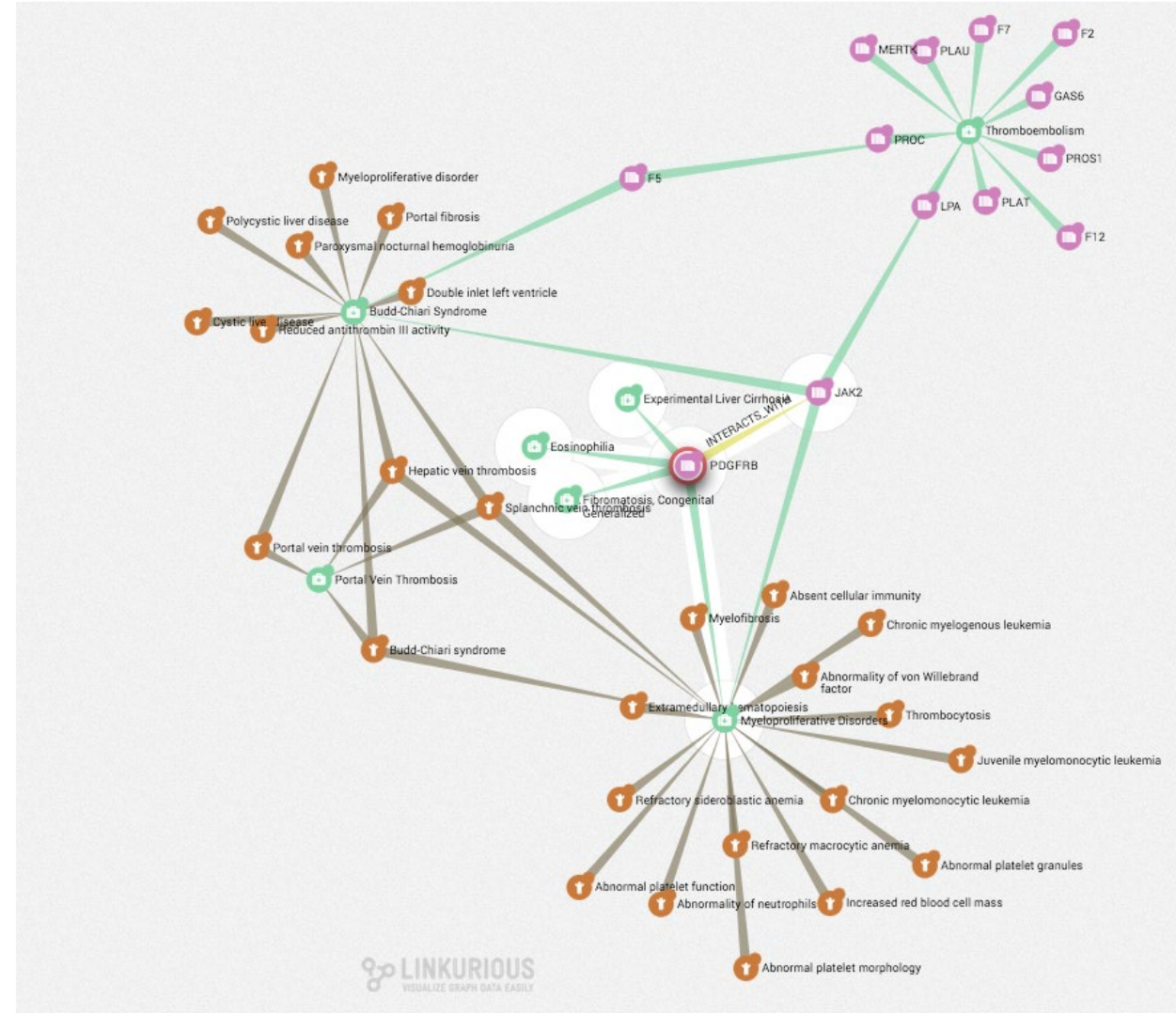
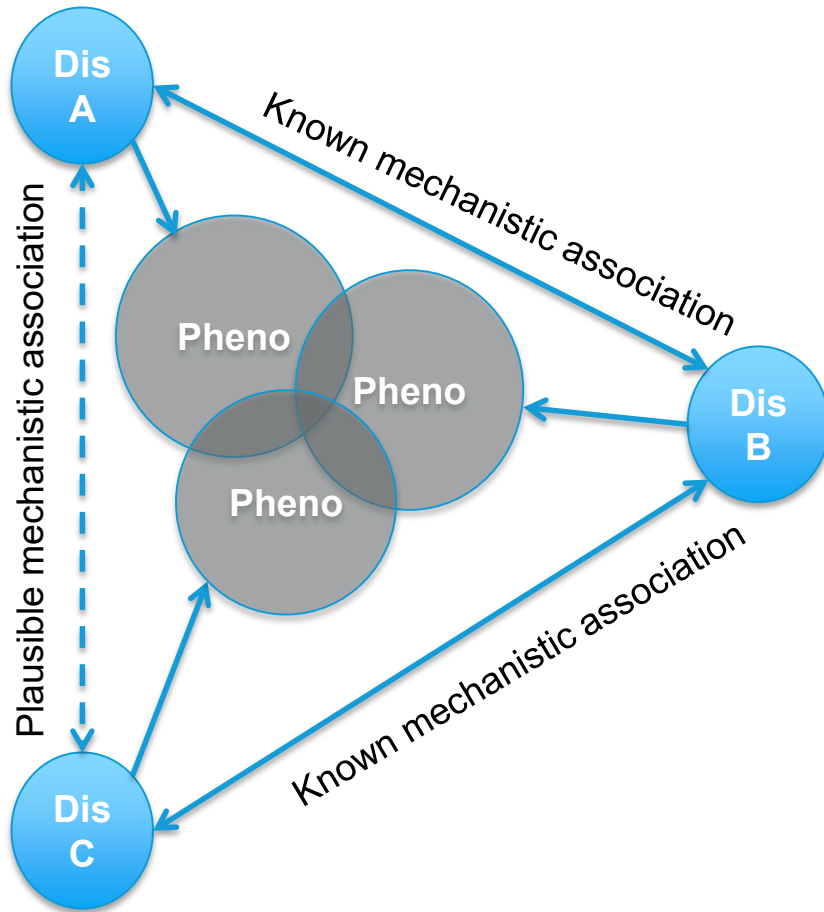
- **Standardise terminology** using SciBite VOCabularies
- Transform unstructured text into interoperable **machine-readable data** compatible with downstream applications
- Build network views of disease-phenotype mappings to identify common mechanistic pathways and shared knowledge
- **Align Chemaxon capabilities** to map compounds to the analysis

## The Outcome

- Uncovering novel relationships in disease biology not previously evident in the source data
- Identify plausible new candidates for screening
- Scalable, structured analysis mappable to public ontologies with the flexibility to integrate additional sources over time



# Phenotype triangulation



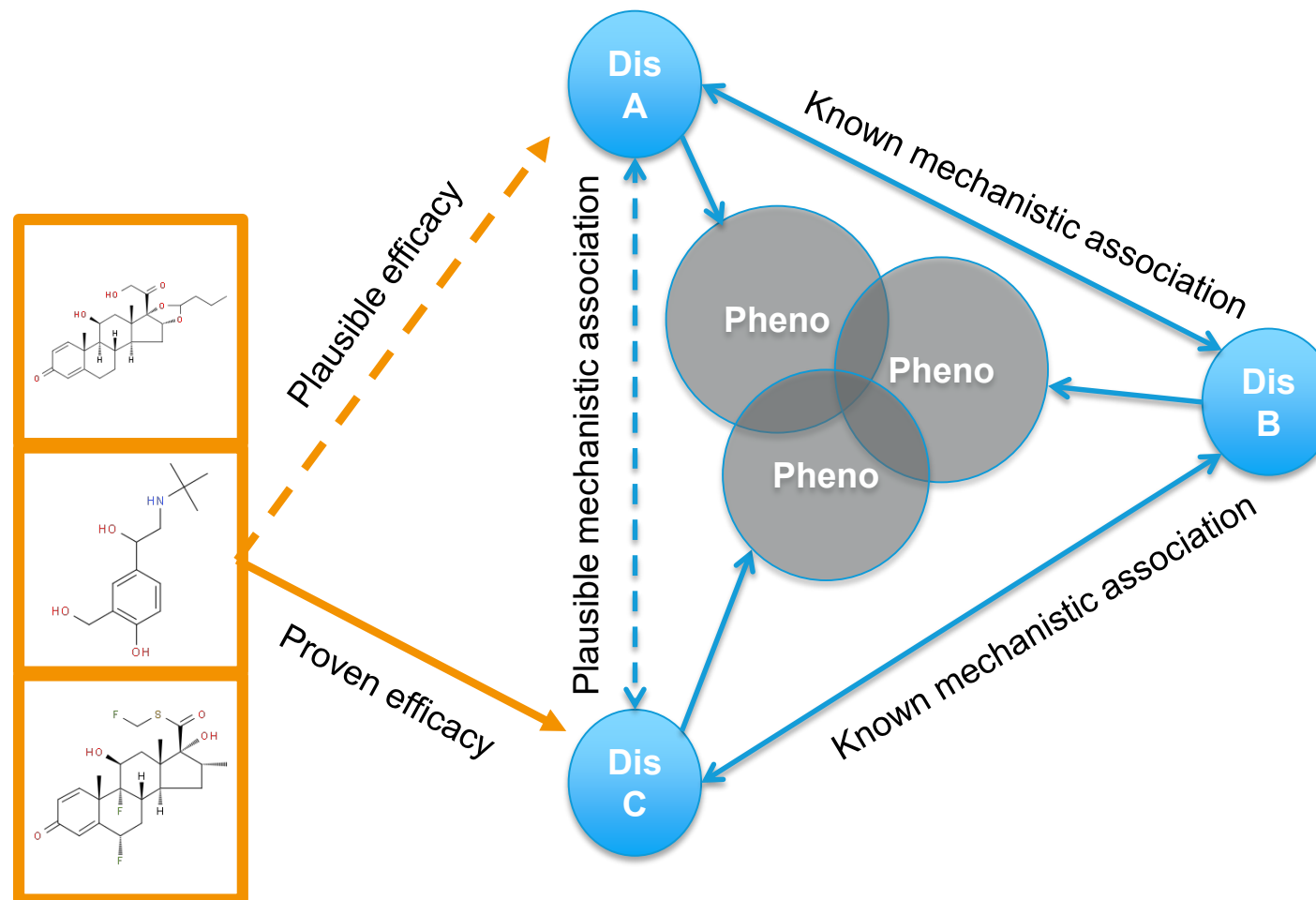
Building network views of disease-phenotype mappings



# Adding Best-in-Class Chemistry Enables Repurposing



By layering in ChemAxon's Named Chemical Entity Extraction, we expand the use of Phenotype Triangulation to Drug Repurposing at-scale





# Thanks for Listening!

## Questions?

**Adam Brown, PhD**

Technical Sales Manager, North America, SciBite

[adam@scibite.com](mailto:adam@scibite.com)





# SciBite

THE LANGUAGE OF SCIENCE



SciBite

# Supplemental Slides



# Many Data Sources Available



## **MEDLINE**

>25M scientific abstracts mined for proximity co-occurrence relationships between Indications and Phenotypes (TExpress)

## **DisGeNET & Orphanet**

High confidence structured data linking indications and rare disorders to associated Genes

## **ChEMBL**

Drug-target interactions; SciBite-generated target druggability scores

## **DailyMed**

Launched Drugs linked to Companies, Indications and Adverse Events (entities extracted via TERMite)

## **BioGPS & Illumina Body Map**

Tissue/cell-specific protein expression

# SciBite's Semantic Glue



To enable data integration, identifiers from external data sources were extracted and mapped to SciBite identifiers (SciBite VOCabs in upper case)

## Direct id mappings

DisGeNET gene symbol (HGNC) -> GENE

Orphanet gene ID (HGNC) -> GENE

Orphanet indication ID -> ORPHAN

ChEMBL molecule id -> DRUG

ChEMBL target ids -> GENE

OMIM approved gene symbol (HGNC) -> GENE

## Entities extracted from field label text

Illumina & BioGPS tissue/cell names in column headers -> ANAT & HUCCELL

## Entities extracted from semi-structured text

DisGeNET disease name -> INDICATION

DailyMed XML company, drug, indications & adverse events fields -> COMPANY, DRUG INDICATION, ORPHAN & ADVENTMED

## Entities extracted from unstructured text

MEDLINE titles and abstracts -> INDICATION, ORPHAN & HPO

# Populating the graph DB

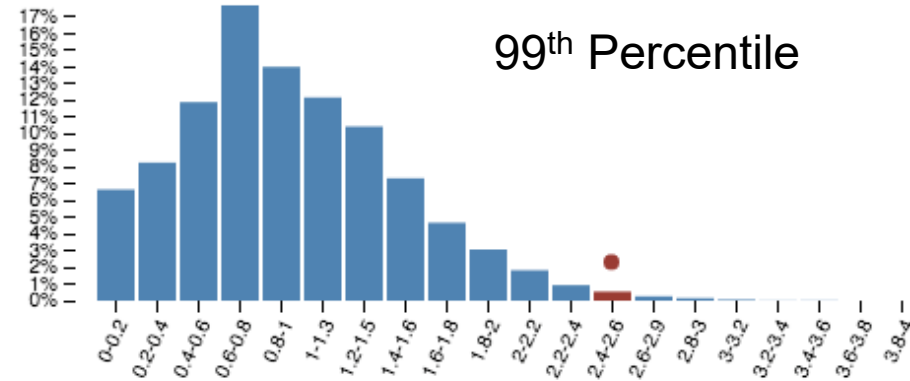


- A graph database consists of nodes (the things in the graph) and edges (the relationships between the things)
- Nodes are be described by:
  - Type (e.g INDICATION, PHENOTYPE, GENE, DRUG)
  - Properties (e.g. id, name, data source, druggability score)
- In this case, the nodes were generated using all entities from the relevant SciBite VOCabs
- Relationships between the nodes were then added using tabular versions of the data from the previous slide
- Relationship examples:
  - (INDICATION {name: "Chron Disease"})-[HAS\_PHENOTYPE]->(HPO)
  - (DRUG)-[HAS\_TARGET]->(GENE)

# Calculating significance of a relationship



- For INDICATION → HPO we used TExpress to pull out sentence co-occurrence relationships from MEDLINE
- Mutual information scoring was applied to filter out the noise from the set of co-occurrence relationships
- E.g. Some text might report vomiting in the context of a broad range of indications. So the fact that an individual indication is linked to vomiting is not very interesting
- The result of the analysis is a set of linked HPO entities per INDICATION, with each link having a score that can be plotted on the distribution of all the scores for that INDICATION
- We can then rank and filter the results by percentile e.g.





# Similarity scoring



- Once we have a set of indications linked to phenotypes, we can identify similar indications based on their phenotypic profiles
- E.g. Insulin Resistance (IR) vs Alzheimers Disease (AD):

Shared Phenotype	IR percentile rank	AD percentile rank
Hyperleucinemia	88	89
Abnormality of the pineal gland	80	83
Abnormal mitochondrial number	76	79
Abnormal homeostasis	84	68
Abnormality of mitochondrial metabolism	64	69
Etc ....		

Cosine similarity = 0.07; percentile rank = 92