SciBite

Phenotype Triangulation and Beyond with ChemAxon

Monday, March 25, 2019

Adam Brown, PhD
Technical Sales Manager, North America, SciBite
adam@scibite.com

THE LANGUAGE OF SCIENCE

SciBite

SciBite's Purpose

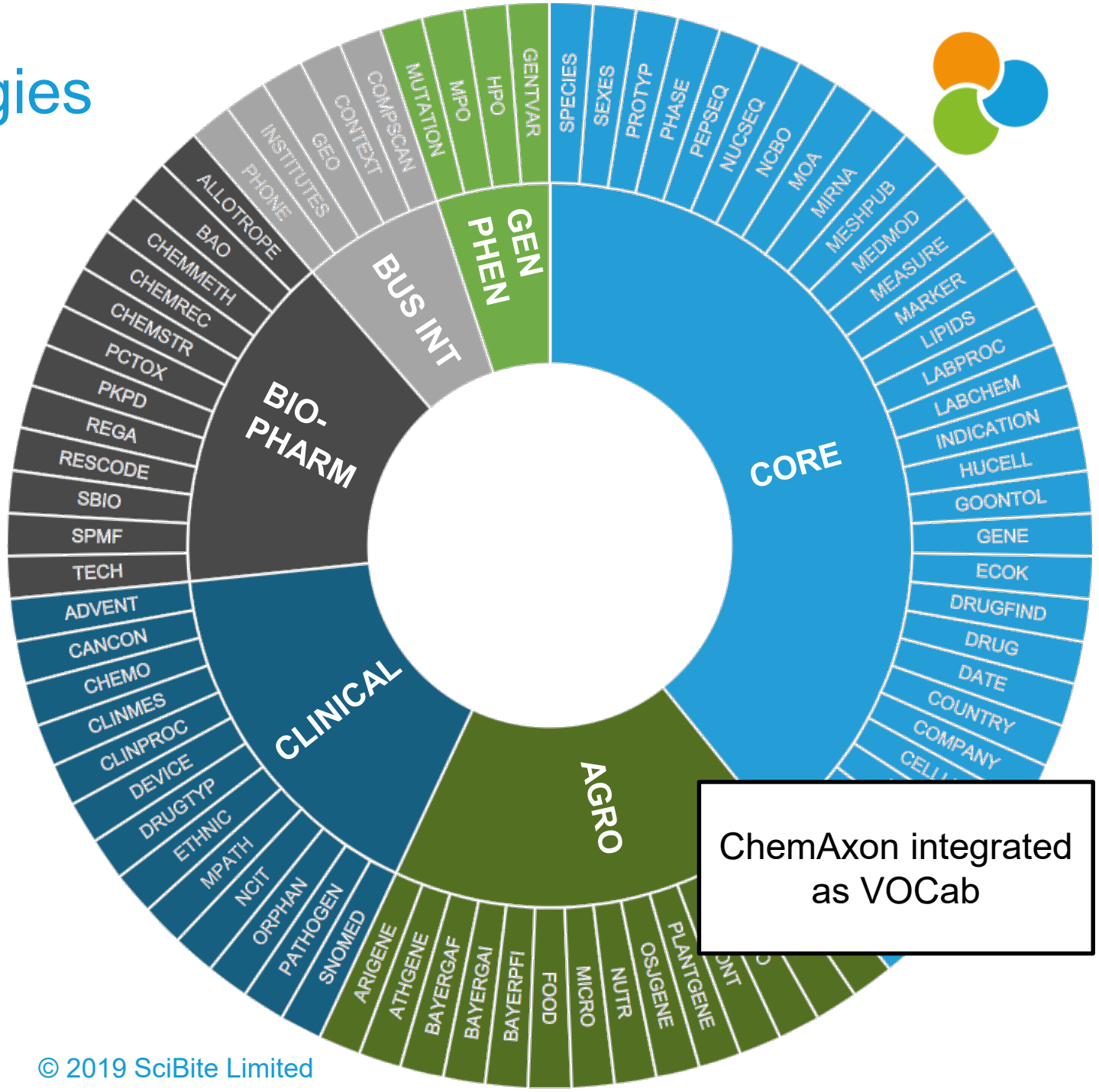
To enable scientists to use insights locked in **unstructured data** to power their decision and speed up innovation by:

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

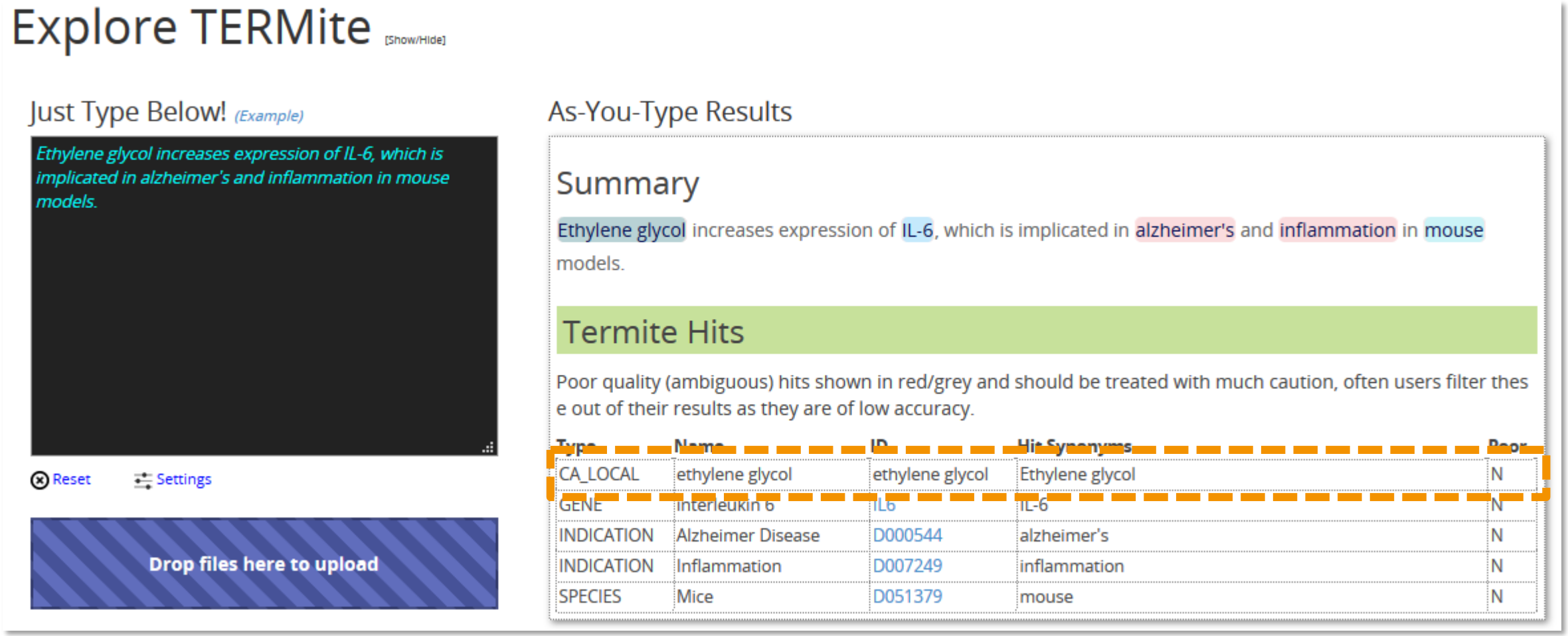
- Ontology content tuned for text mining
- Maintained by our dedicated team using expert curation and machine learning
- Comprehensive coverage
- Aligned to industry standards to maintain interoperability
- Enriched with synonyms and rules to manage the complexity of scientific language
- Customize,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



ChemAxon integrated as VOCab

ChemAxon Integration - Example

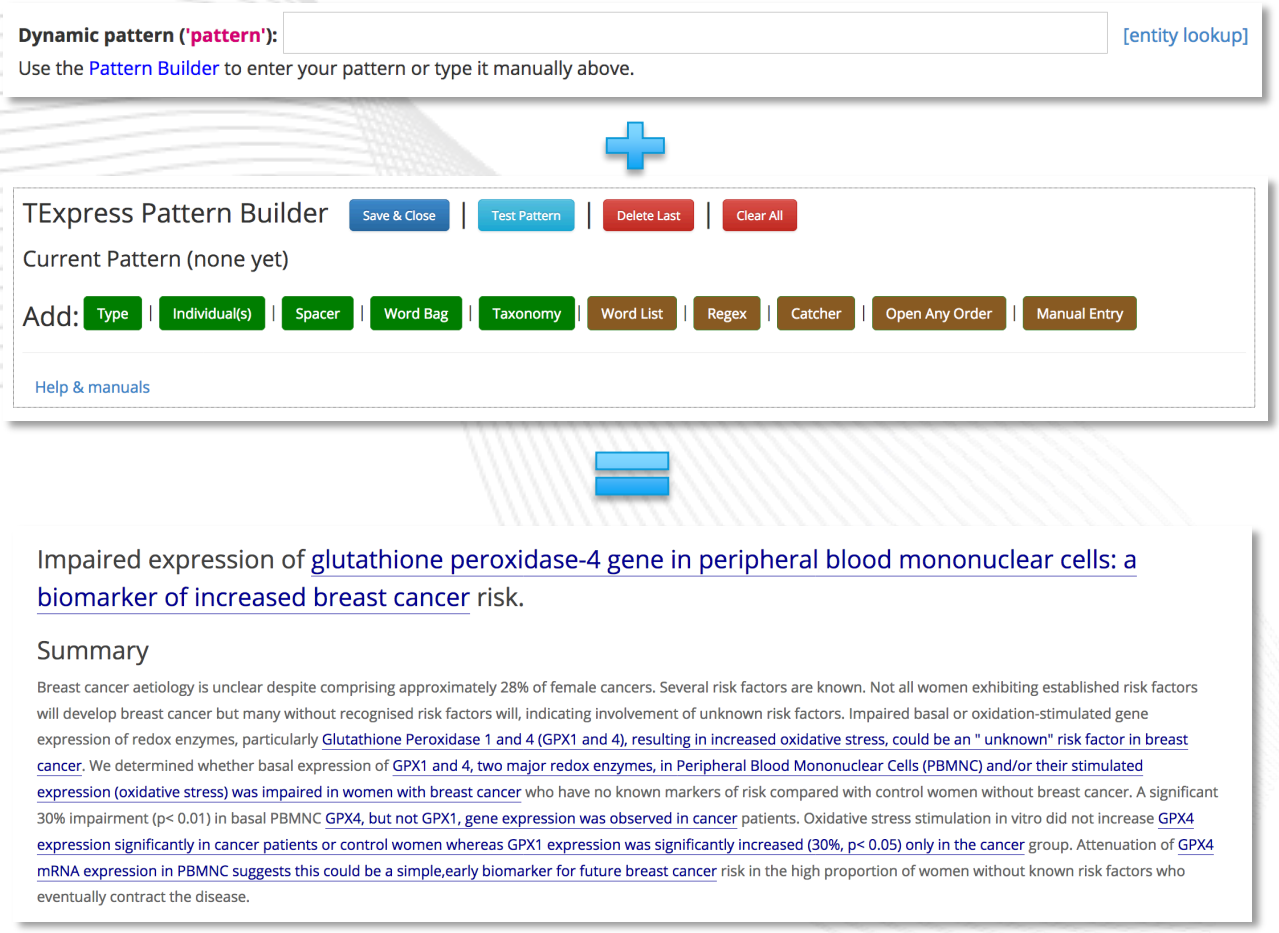
ChemAxon lookup service seamlessly integrated into TERMite, SciBite's Biomedical Named Entity Recognition tool



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)



SciBite

Phenotype Triangulation

Identifying new research candidates based on mechanistic similarities

Phenotypic Triangulation

The Problem

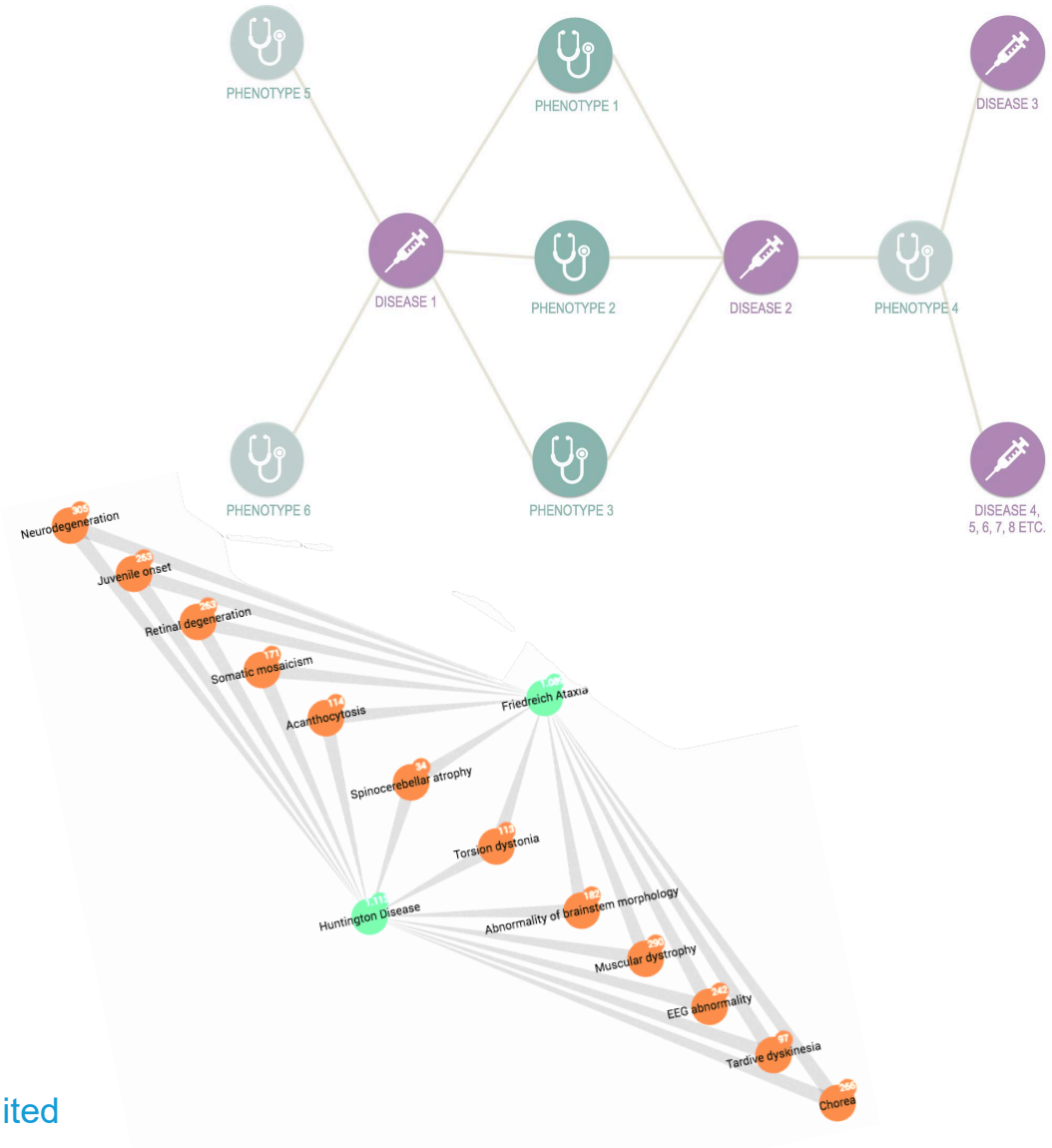
- Many diseases are understudied and lack clear molecular mechanisms
- Some entities (e.g. Phenotypes) are highly synonymous and difficult to standardise
- Scraping, standardising, and analysing research is time-consuming

The Solution

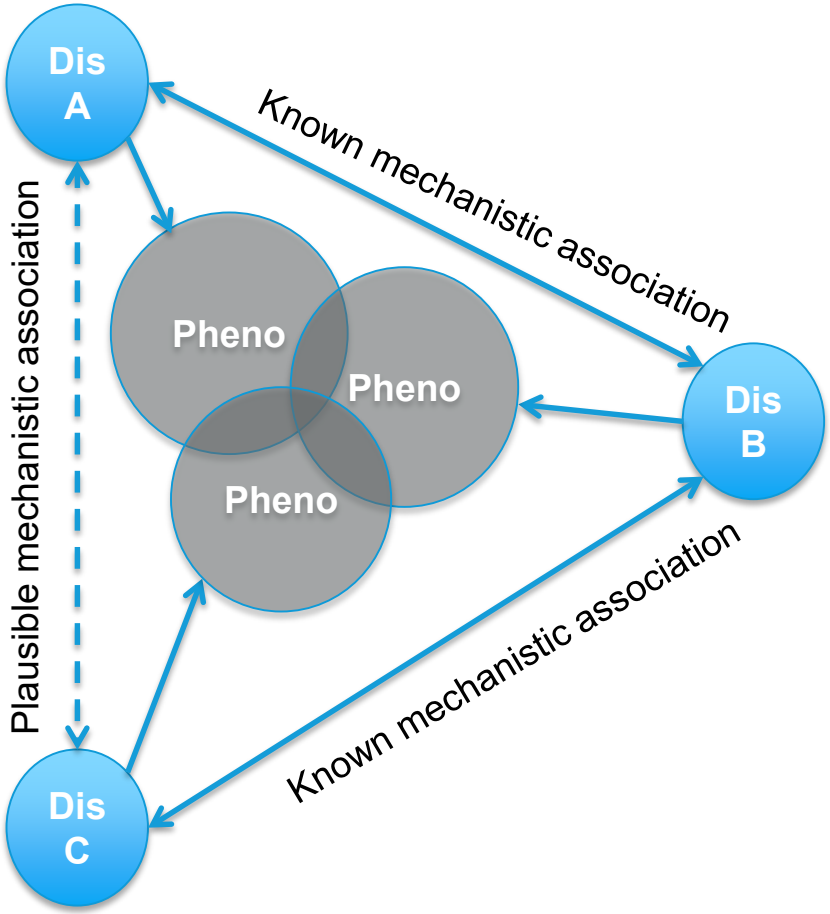
- **Standardise terminology** using SciBite VOCabularies
- Transform unstructured text into interoperable **machine-readable data** compatible with downstream applications
- Build network views of disease-phenotype mappings to identify common mechanistic pathways and shared knowledge
- **Align Chemaxon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

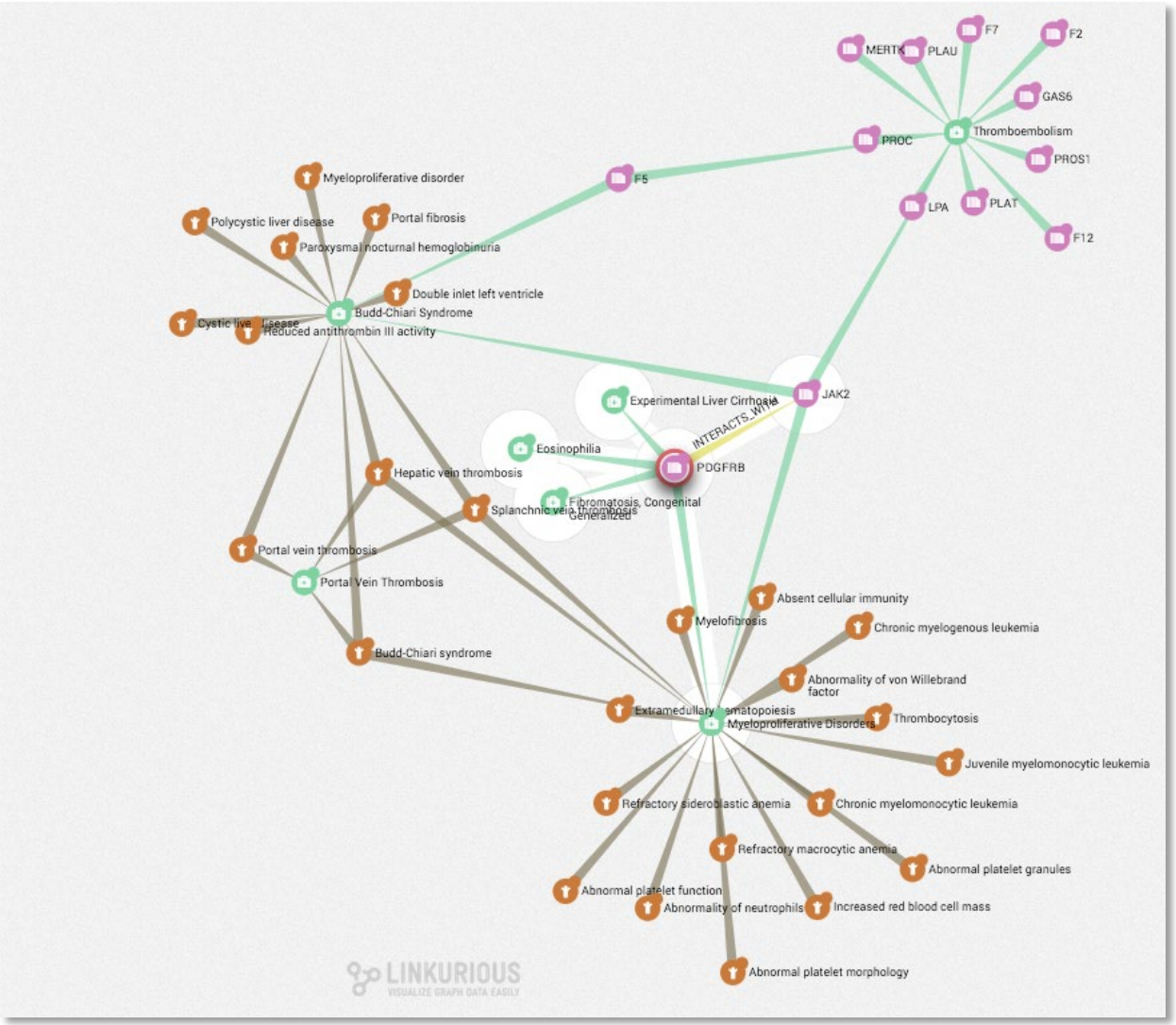
The Outcome

- 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

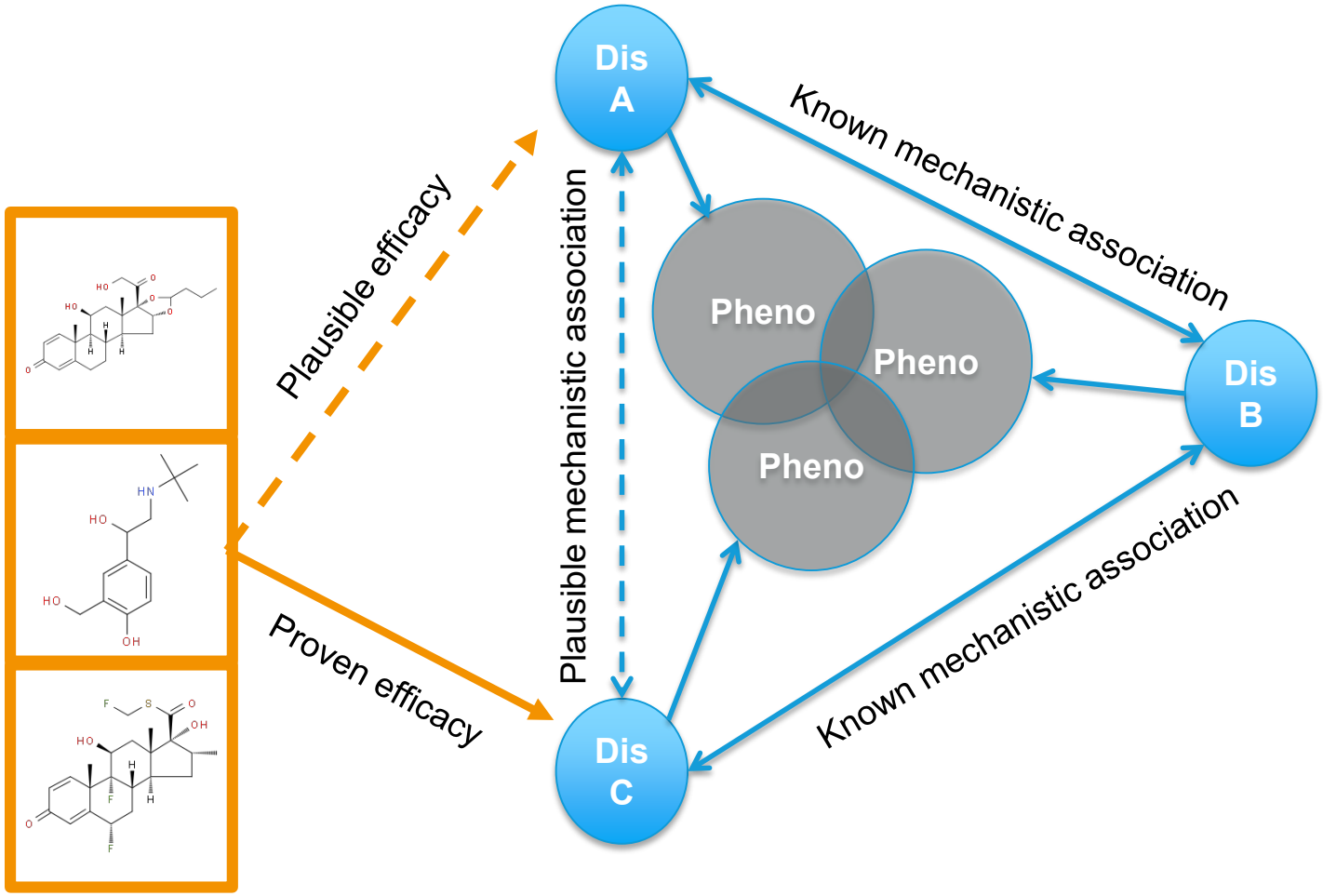




Building network views of disease-phenotype mappings

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



SciBite

Thanks for Listening!

Questions?

Adam Brown, PhD
Technical Sales Manager, North America, SciBite
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 & HUCELL

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

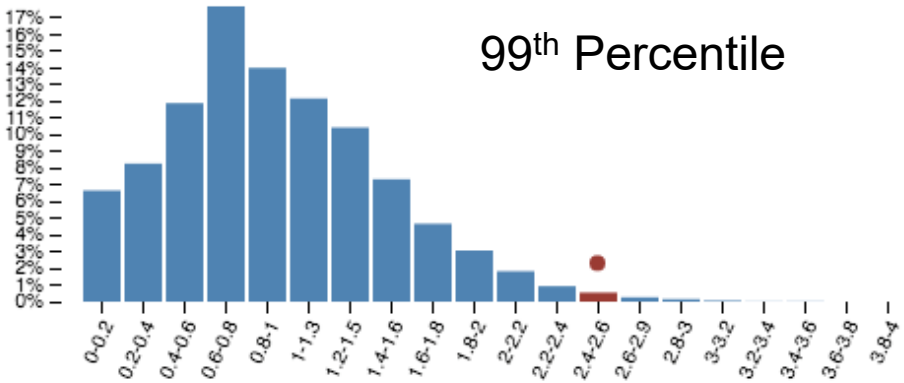
- MEDLINE titles and abstracts -> INDICATION, ORPHAN & HPO

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