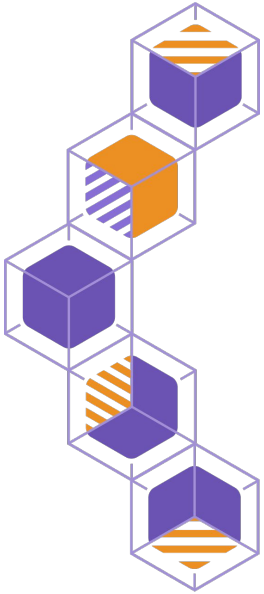
Biomedical Knowledge Graphs from public data

Shashank Jatav, Director Data Products

Elucidata



Agenda

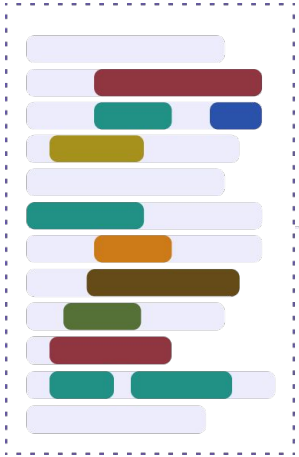
- Motivation
- Challenges
- Solution
- GraphOmix
- Data Driven Insights

Motivation

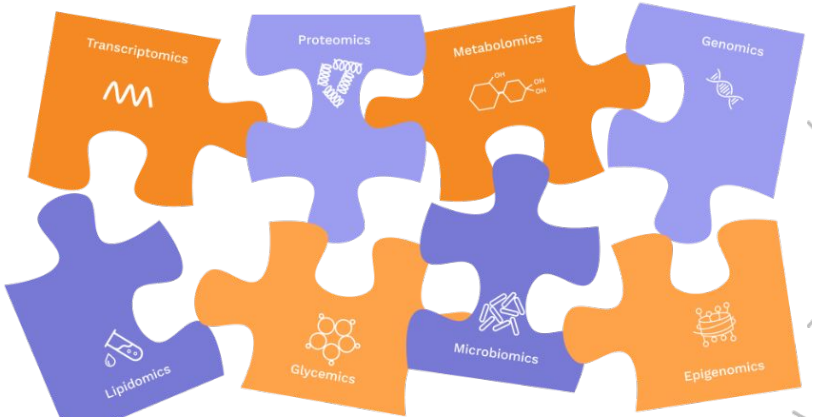
Knowledge Graphs for Data Integration

- Even a seemingly straightforward, single-omics experiment will consider **connected entities** like biomolecules, pathways, diseases, etc.
- The emergence of multi-omics makes biomedical data even more complex.
- Combining all this data together gives us the potential to unlock more valuable insights.
- Knowledge graphs (KGs) are a natural way for -
  - capturing and **integrating** large amounts of **connected** heterogeneous information
  - deriving useable **insights** and knowledge from that data (e.g., hypothesis generation)
  - comprehensibility, interpretability, and **explainability** of insights
- KGs are proving to be successful in several downstream tasks in drug discovery like drug target identification and drug repurposing, especially with the advent of Graph Machine Learning (GML).

90% of the data generated is not used



- Less than **5%** of the data generated is analyzed and presented in a publication
- Current Knowledge Graph approaches use **only text mining** to find relationships across datasets



- Biomedical data is inherently heterogeneous and more than **90%** of the data is not integratable
- More than **50%** of the data is missing annotations

Challenges

Challenges

- There still remain several challenges when it comes to building and maintaining a KG like -
  - Heterogeneous, multimodal data generated in biomedical domain (Unclean and unusable data)
    - Lack of a unified data schema and metadata harmonization
    - Lack of features
  - Lack of updates

Despite significant initiatives to “digitally transform” Novartis, their CEO, Vas Narsimhan, has remarked on the difficulty to clean and link their heterogeneous data.

Source: [1] External Link

Vas Narasimhan, CEO

NOVARTIS

Data Schemas are Inconsistent

Legacy multi-omics datasets require hours to days of cleaning-up

Expression matrix split into multiple files (GSE96075)

Supplementary file	Size	Download	File type/resource
GSM2532812_SM09_ReadsPerGene.out.tab.gz	364.4 Kb	(ftp)(http)	TAB

Supplementary file	Size	Download	File type/resource
GSM2532811_SM05_ReadsPerGene.out.tab.gz	357.9 Kb	(ftp)(http)	TAB

Supplementary file	Size	Download	File type/resource
GSM2532812_SM09_ReadsPerGene.out.tab.gz	364.4 Kb	(ftp)(http)	TAB

A scientist spends an hour to a day making a dataset analysis-ready. [1]

GPL570

Gene Symbol	ENTREZ_GENE_ID
DDR1 /// MIR4640	780 /// 100616237
RFC2	5982
HSPA6	3310
PAX8	7849
GUCA1A	2978
MIR5193 /// UBA7	7318 /// 100847079

GPL26227

ID	ENTREZ_GENE_ID
100009600_at	100009600
100009609_at	100009609
100009614_at	100009614
100012_at	100012
100017_at	100017

GPL27634

ID	SPOT_ID
ENSG00000000003.14_at	ENSG00000000003.14
ENSG00000000005.5_at	ENSG00000000005.5
ENSG00000000419.12_at	ENSG00000000419.12
ENSG00000000457.13_at	ENSG00000000457.13
ENSG00000000460.16_at	ENSG00000000460.16
ENSG00000000938.12_at	ENSG00000000938.12

Gene identifiers not consistent across different platforms

Only 3% of GEO datasets are machine-readable.[2]

Source: [1] NCBI; [2] Based on Elucidata case studies

Metadata is seldom standardized

Often the most relevant dataset is never identified, let alone used

NCBI GEO Gene Expression Omnibus

Samples (52) Less...	GSM2667747	hesc_cyto_rep1
	GSM2667748	hesc_nuc_rep1
	GSM2667749	hesc_monosome_rep1
	GSM2667750	hesc_poly_low_rep1
Samples (11) Less...	GSM2706433	control, biological replicate A
	GSM2706434	control, biological replicate B
	GSM2706435	control, biological replicate C
	GSM2706436	control, biological replicate D
Samples (24) Less...	GSM2671391	Lung_Non-infected_2
	GSM2671392	Lung_Non-infected_3
	GSM2671393	Lung_Non-infected_1
	GSM	

Different conventions for indicating biological/technical replicates in metadata

Use of controlled vocabularies or ontologies is rare at best, patchy when present

"find all type 2 diabetes studies where MC4R is differentially expressed" could require anywhere from days to weeks to months [1]

Source: [1] Based on Elucidata case studies

But this is easy right?

Or will just a few lines of tidyverse will do the job

100,000	x	50+	x	10+
new datasets per year		omics repositories		storage formats

Handling data velocity and variation requires automation and tech expertise

Allow data scientists to focus on extracting value from data, leave data-prep to Polly

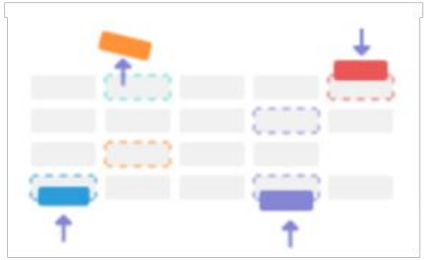
Solution

How does data become ML-ready?

Every dataset on Polly undergoes 2 key steps that make it machine readable and ML-ready

Step 1

Data Engineering

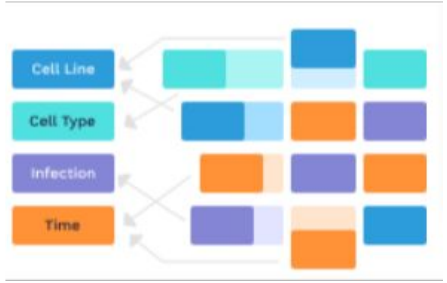


Standardized tabular data schema

Step 2

Metadata Harmonization

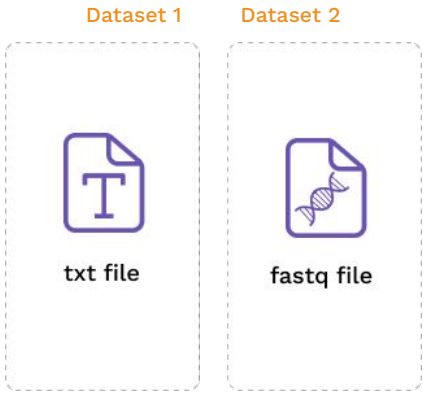
Standardized ontologies of Datasets



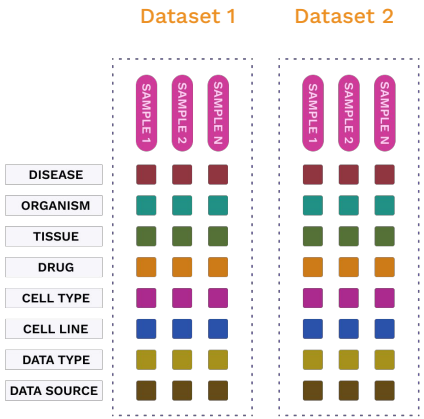
Standardized ontologies of Metadata

Polly Connectors: ETL pipelines to standardize data schema

BEFORE POLLY



AFTER POLLY



- Before Polly, No standard schema followed for each data types
- On Polly, Data streamlined in one consistent schema
- Over 1.4 million datasets are on Polly right now

Data available in different forms

- Matrix file
- raw file
- S4 object
- rds

Standard Data Schema on Polly:

- GCT
- H5ad or h5seurat

Unified Data Schema: Deep querying and flexible streaming

File Metadata

1,288
Number of Samples

Homo Sapiens
Organism

1,288
Number of Samples

gene_a.GCT
File name

Molecular Identifier

Mutations, Copy Number

Dosage

Molecular Properties & Relationships

Molecular Metadata

Sample Metadata

Data File

		Sample 1	Sample 2	Sample 3	Sample 4
		DMSO	BRD-K52313	BRD-K7585	BRD-K3567
		DMSO	Tacedinaline	Entinostat	Tacedinaline
		0 Hrs	12 Hrs	24 Hrs	36 Hrs
PSME1	5720	9.345	9.345	45.6765	67.36
CISD1	55789	33.56423	33.56423	12.7685	11.674
VDAC1	10174	1.435	1.435	23.69475	9.346
SORBS3	466	4.6578	4.6578	7.34643	34.6785
SPDEF	6676	78.324	78.324	9.43653	34.6896
ATF1	1870	7.342	7.342	4.898	78.9677
SPAG4	3480	9.342	9.342	2.8696	8.64

Data Matrix

15 fields with Ontologies

Disease

Drug

Organisms

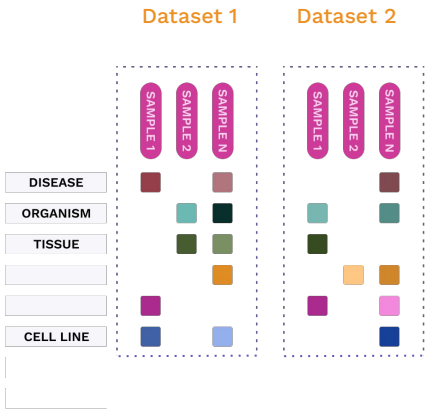
•
•
•
•

Cell Line

Our Unified Data Schema has **unified more than 1.6 million** datasets

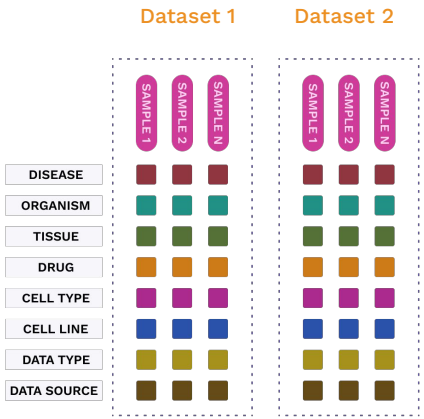
PollyBERT: NLP models for Metadata Harmonization

BEFORE POLLY



Missing annotation : 50%
Harmonized : <2%
Missing fields

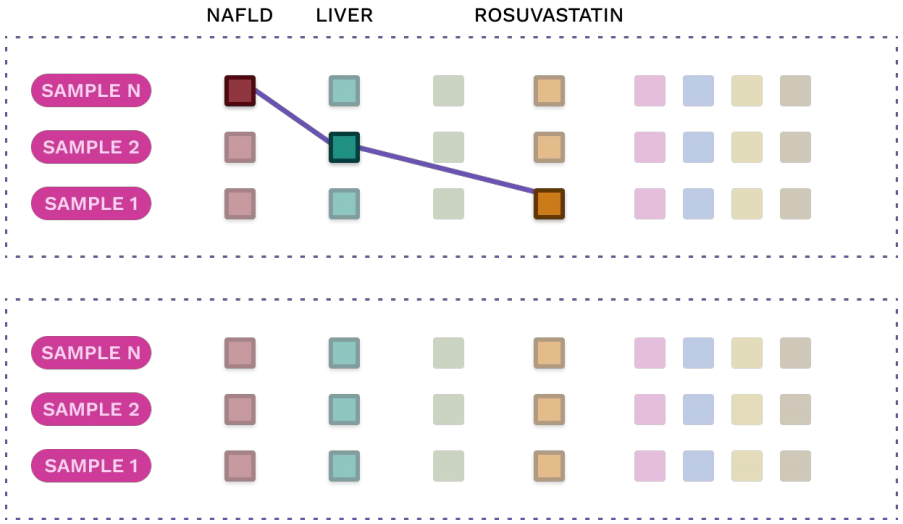
AFTER POLLY

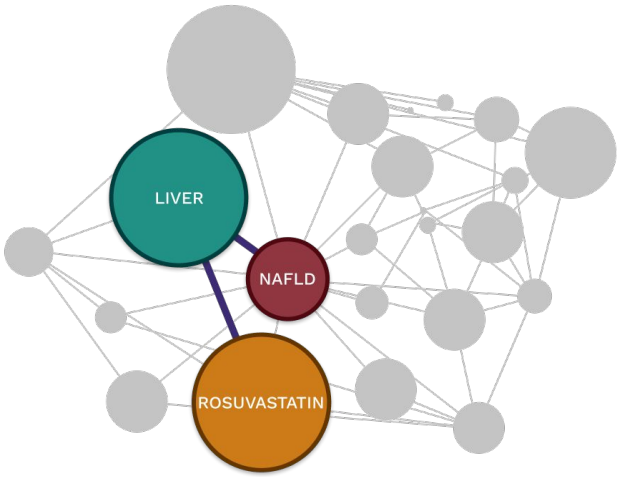


Missing annotation : <1%
Harmonized : 100%
# of new fields added: 4X

- Tag each sample with relevant information such as disease, tissue (source biomaterial), cell line etc.
- Tag each sample uniformly with the same vocabulary
- Process each dataset uniformly with same molecular identifiers

Knowledge Graph generation on Polly – GraphOmix

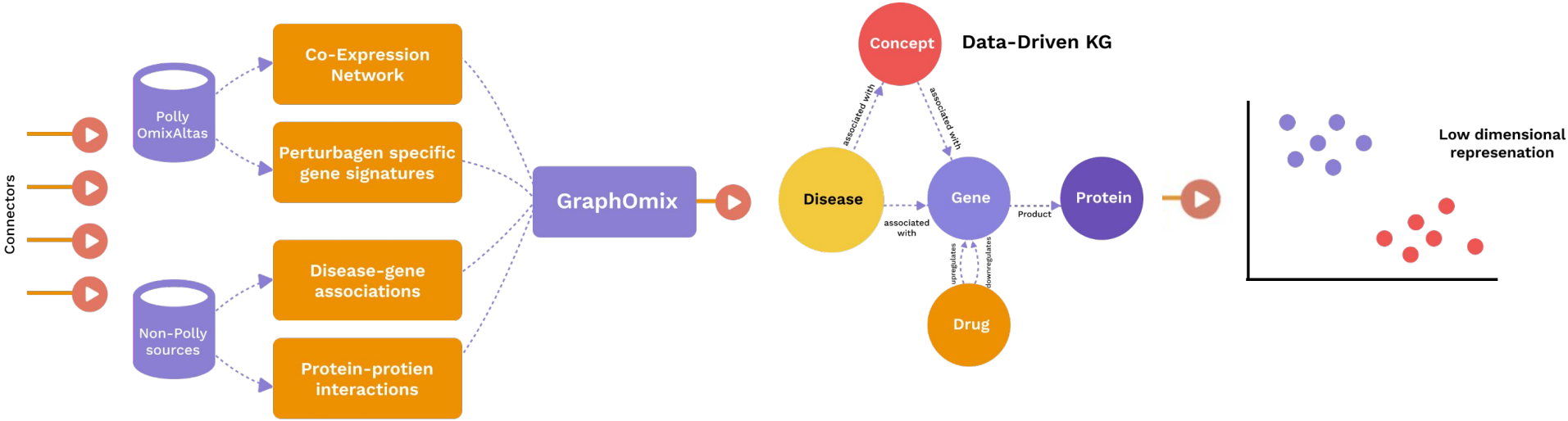




- Create **richer** knowledge graphs across **35 million** auto curated entities on Polly
- Use over **50 billion** data points to form relationships over curated metadata

GraphOmix

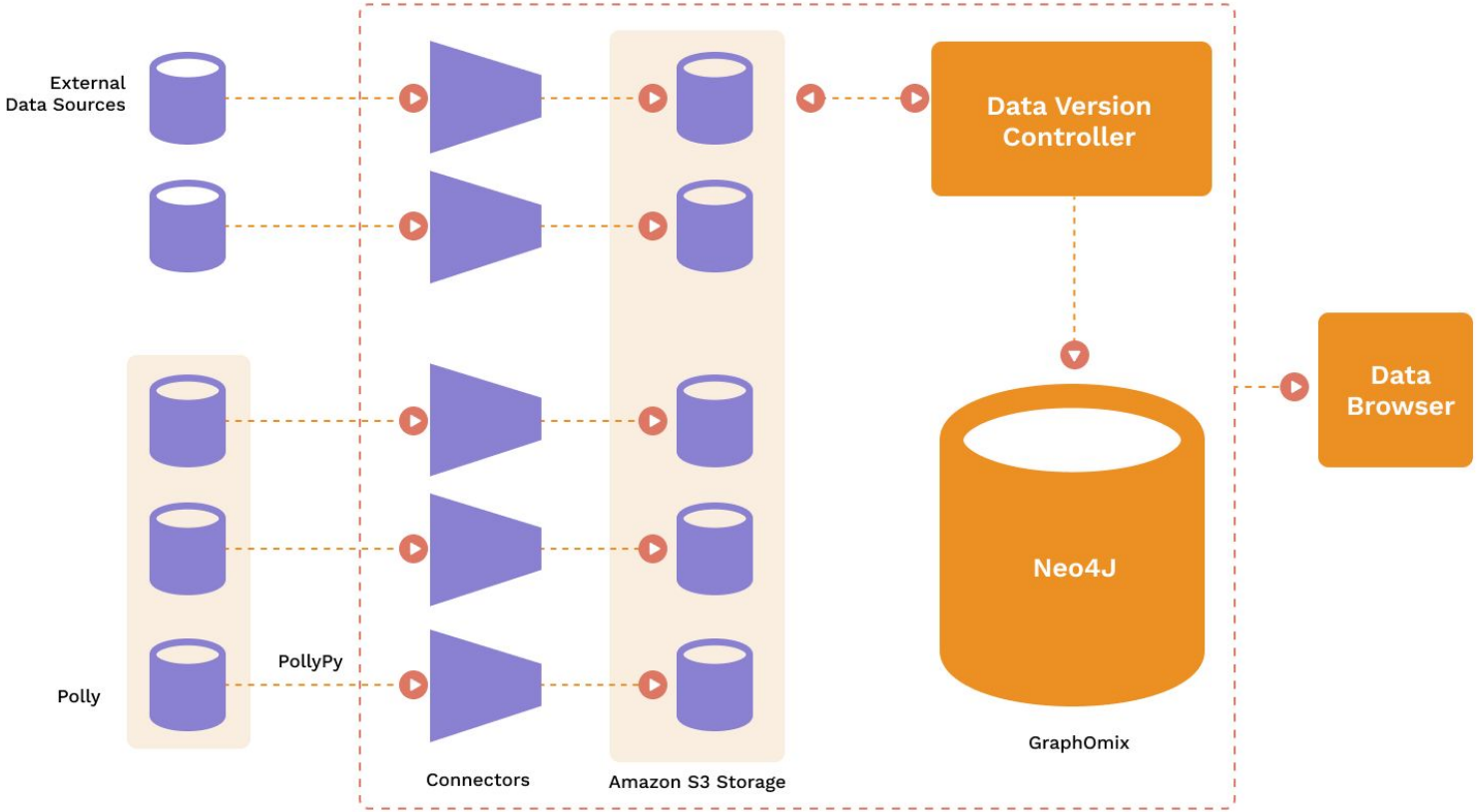
How did we construct the knowledge graph?



Fundamental hypothesis: Datasets with similar coexpression/differential expression must have similar biology

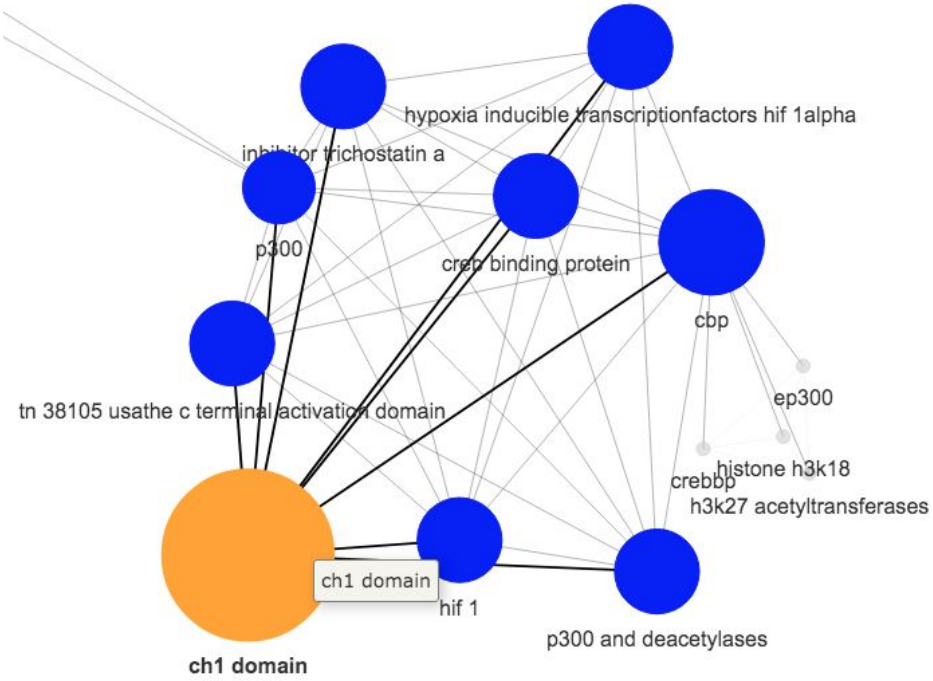
- Independent of any text mining interpretations of how various entities interact
- Any new knowledge graph can be easily constructed using existing curated data
- First knowledge graph which actually uses data at scale

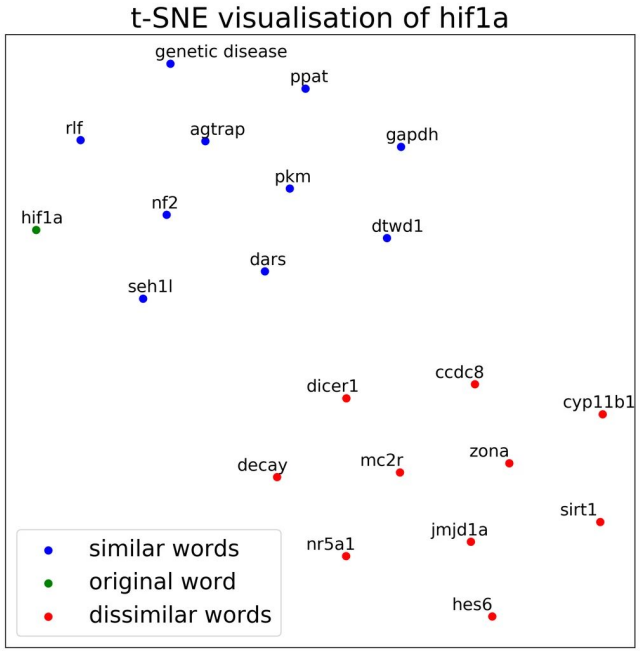
Architecture Diagram



Data Driven Insights

Knowledge Graph for Hypoxia





CH1 domain interacts with HIF1-alpha which is a binding site for CREB-binding protein. These links were evident with GraphOmix

A low dimensional representation cleanly separates out transcriptional factors known for opposite activity

Drug Repurposing for Skeletal Dysplasia

> ScientificWorldJournal. 2014 Jan 28;2014:619050. doi: 10.1155/2014/619050. eCollection 2014.

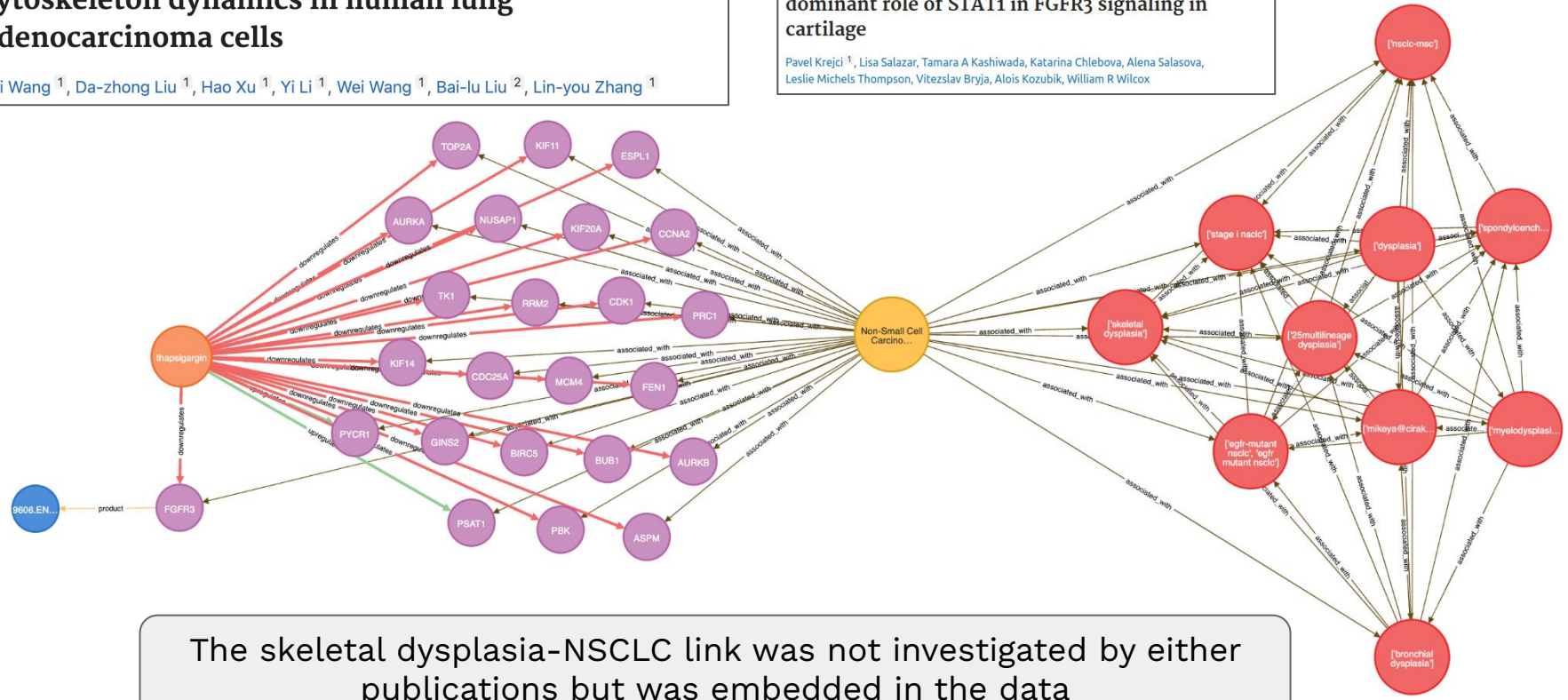
Thapsigargin induces apoptosis by impairing cytoskeleton dynamics in human lung adenocarcinoma cells

Fei Wang [1], Da-zhong Liu [1], Hao Xu [1], Yi Li [1], Wei Wang [1], Bai-lu Liu [2], Lin-you Zhang [1]

> PLoS One. 2008;3(12):e3961. doi: 10.1371/journal.pone.0003961. Epub 2008 Dec 17.

Analysis of STAT1 activation by six FGFR3 mutants associated with skeletal dysplasia undermines dominant role of STAT1 in FGFR3 signaling in cartilage

Pavel Krejci [1], Lisa Salazar, Tamara A Kashiwada, Katarina Chlebova, Alena Salasova, Leslie Michels Thompson, Vitezslav Bryja, Alois Kozubik, William R Wilcox



The skeletal dysplasia-NSCLC link was not investigated by either publications but was embedded in the data

Thanks!

Any questions?

You can find us at www.elucidata.io & shashank.jatav@elucidata.io