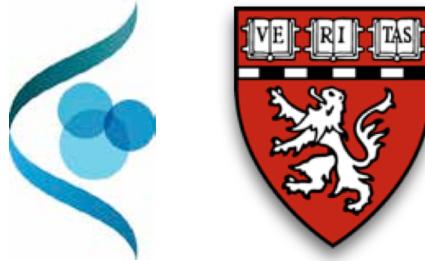


An automatically assembled knowledge graph from
literature-extracted biomedical knowledge
with human-machine dialogue to support discovery

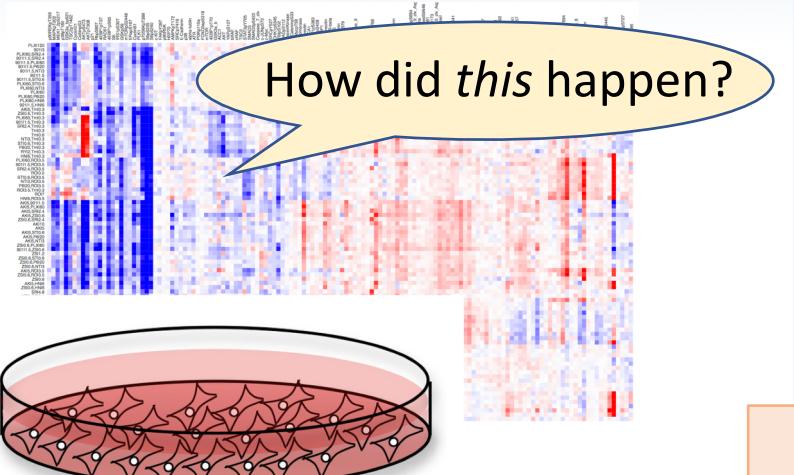


Benjamin M. Gyori
Laboratory of Systems Pharmacology
Harvard Medical School

KGC HCLS
5/2/2022

Human-machine collaboration for discovery and decision making

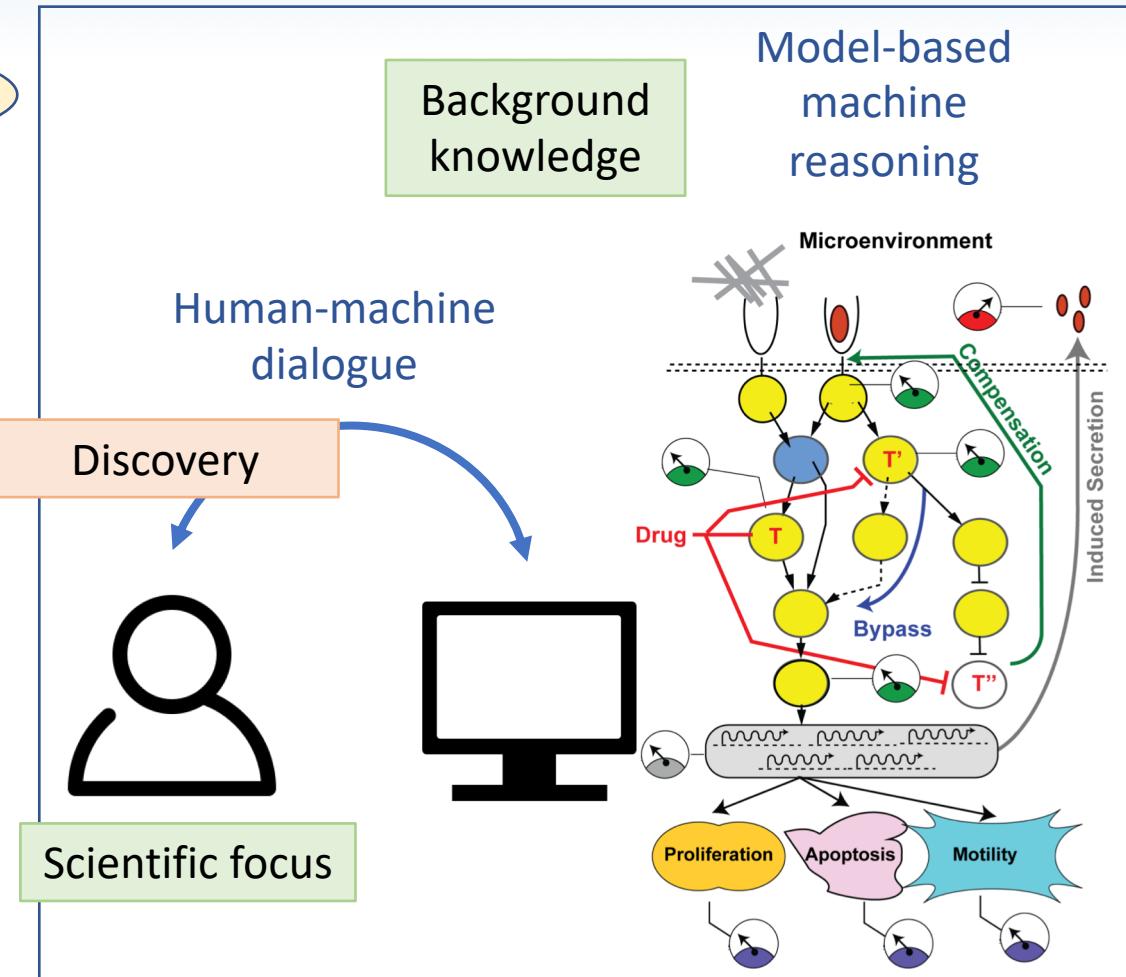
Scientific discovery



Clinical decision making

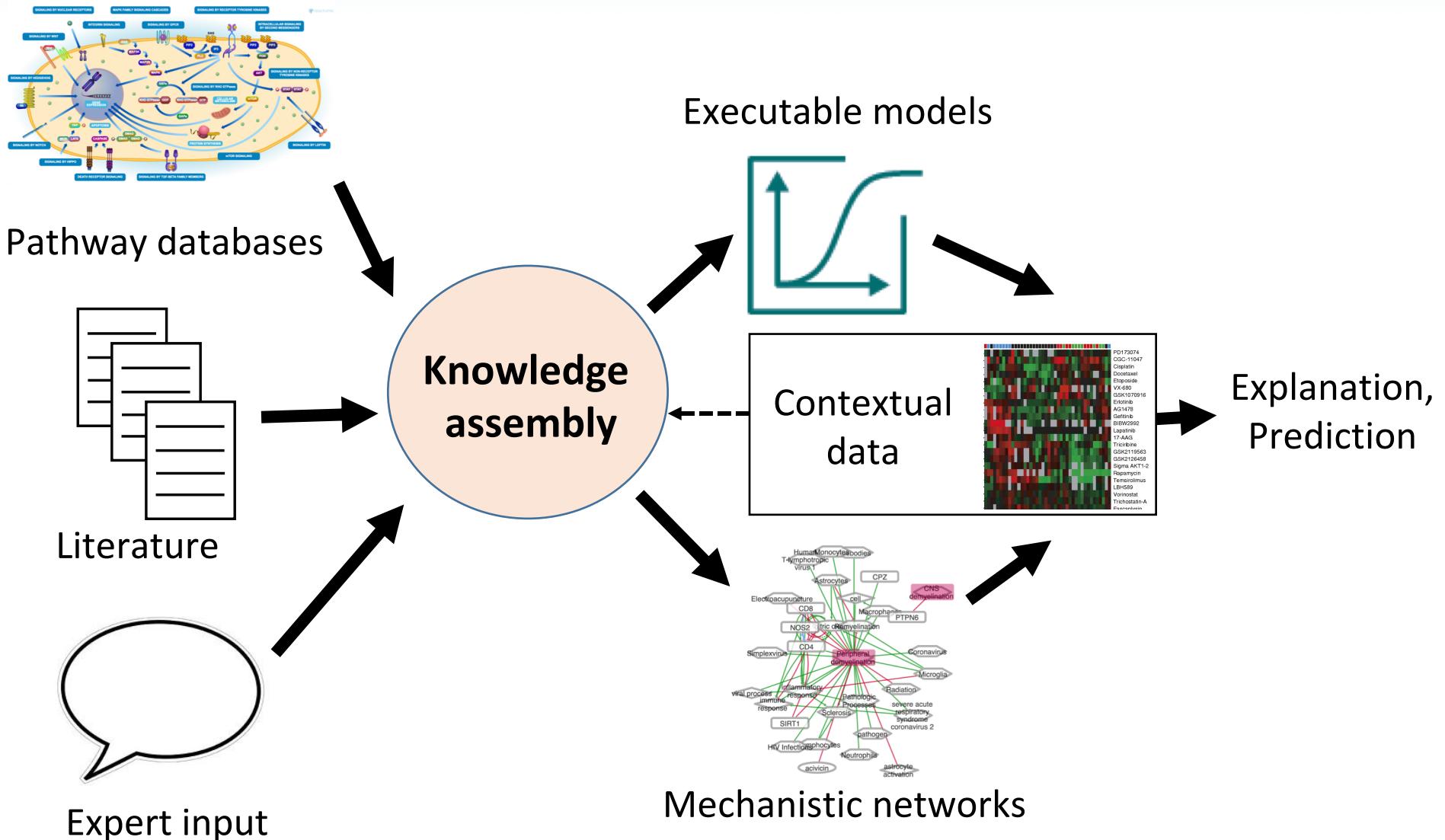


Is this treatment justified at the level of molecular mechanisms?

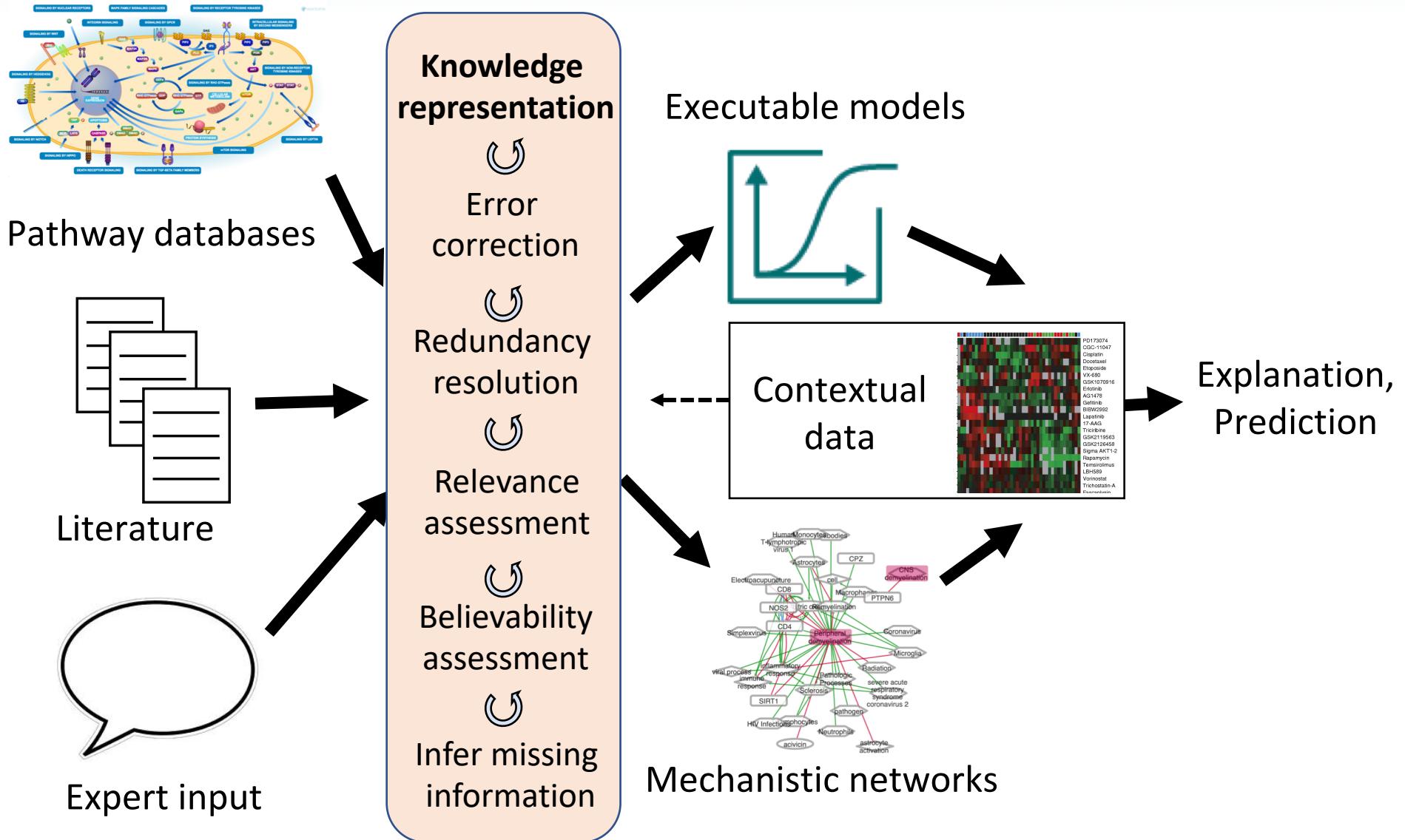


Collaborative problem solving

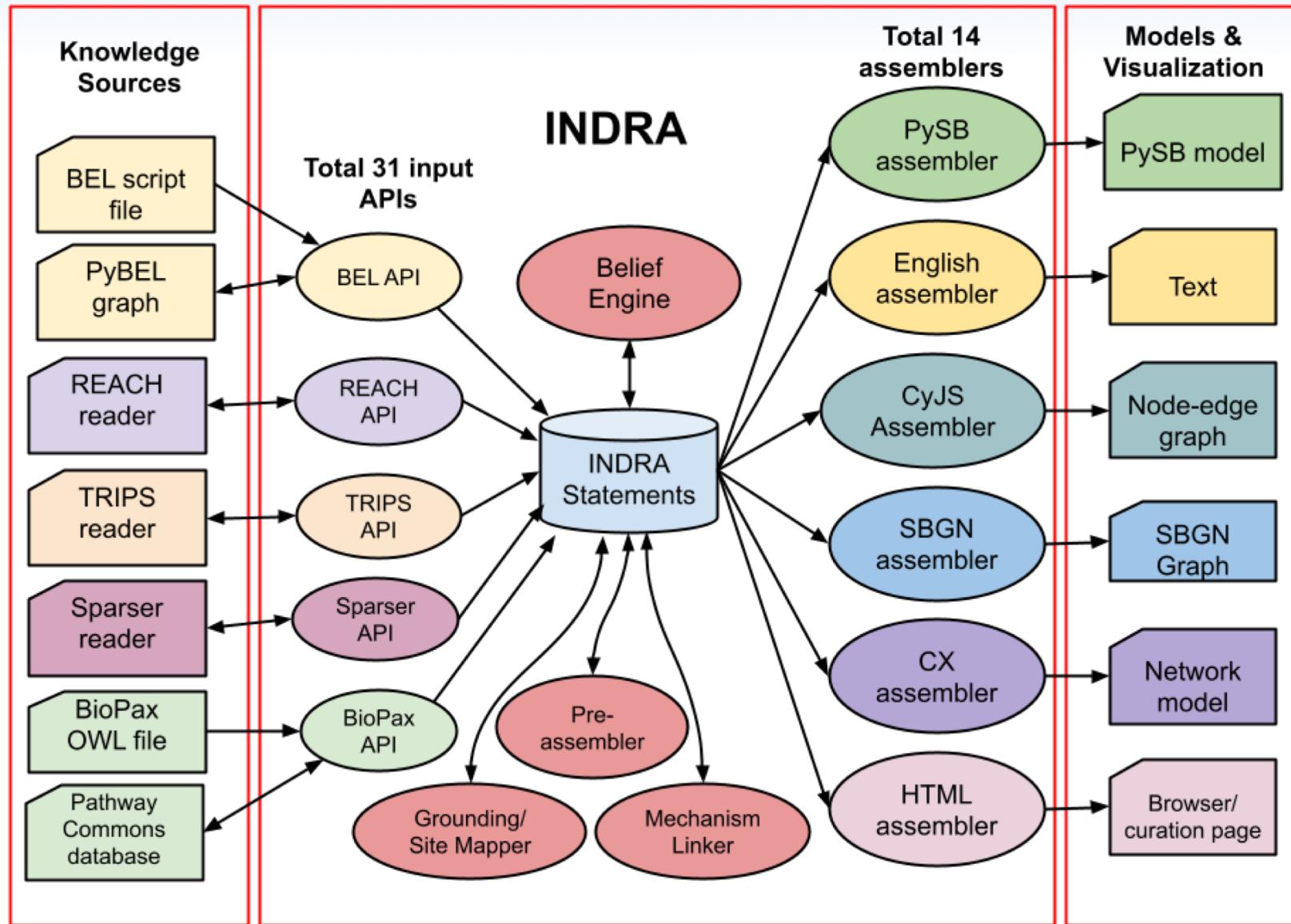
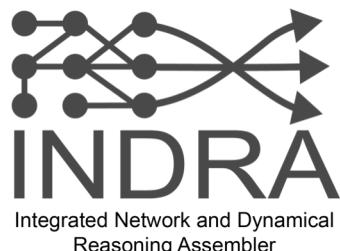
Conceptual overview of machine-assisted modeling



Conceptual overview of machine-assisted modeling

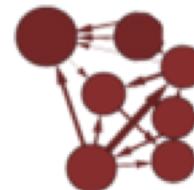
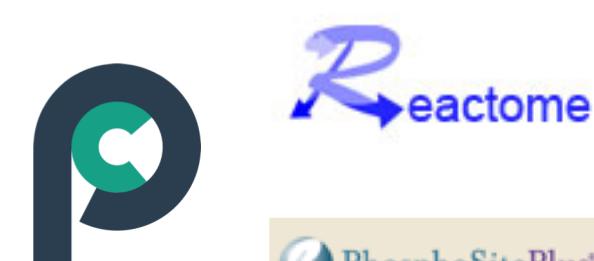


INDRA: Integrated Network and Dynamical Reasoning Assembler



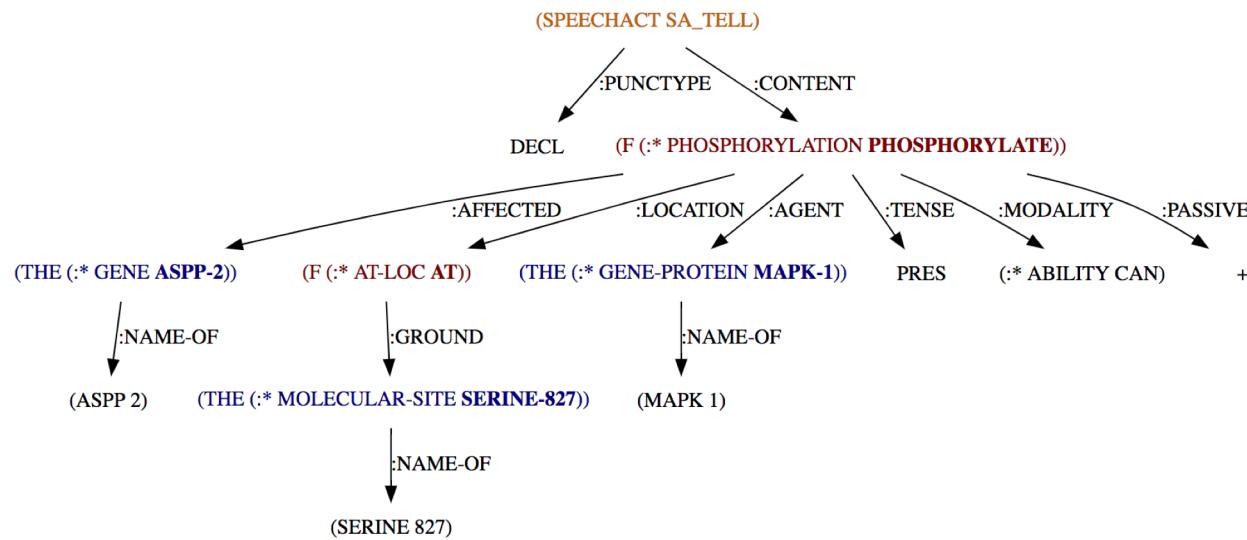
INDRA integrates pathway databases and other structured sources

Database / Exchange format	Module	Reference
PathwayCommons / BioPax	indra.sources.biopax	http://pathwaycommons.org/ http://www.biopax.org/
Large Corpus / BEL	indra.sources.bel	https://github.com/pybel/pybel https://github.com/OpenBEL
Signor	indra.sources.signor	https://signor.uniroma2.it/
BioGRID	indra.sources.biogrid	https://thebiogrid.org/
Target Affinity Spectrum	indra.sources.tas	https://doi.org/10.1101/358978
HPRD	indra.sources.hprd	http://www.hprd.org
TRRUST	indra.sources.trrust	https://www.grnpedia.org/trrust/
Phospho.ELM	indra.sources.phosphoelm	http://phospho.elm.eu.org/
VirHostNet	indra.sources.virhostnet	http://virhostnet.prabi.fr/
CTD	indra.sources.ctd	http://ctdbase.org
DrugBank	indra.sources.drugbank	https://www.drugbank.ca/
OmniPath	indra.sources.omnipath	https://omnipathdb.org/
DGI	indra.sources.dgi	https://www.dgib.org/
CRoG	indra.sources.crog	https://github.com/chemical-roles/chemical-roles
CREEDS	indra.sources.creeds	https://maayanlab.cloud/CREEDS/
UbiBrowser	indra.sources.ubibrowser	http://ubibrowser.ncpsb.org.cn/



INDRA is integrated with multiple machine reading systems

ASPP2 can be phosphorylated at serine 827 by MAPK1.

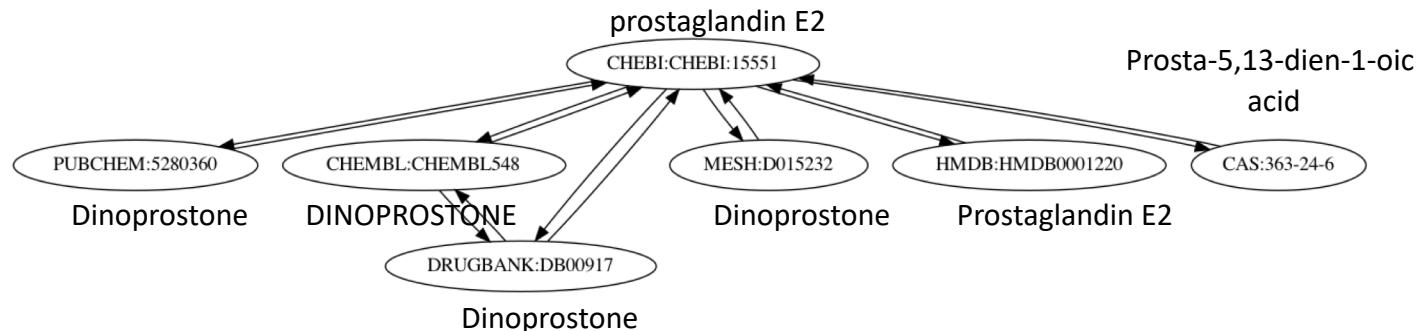


(Allen et al., 2015)

Reader	Module	Reference
TRIPS/DRUM	indra.sources.trips	http://trips.ihmc.us/parser/cgi/drumparser
REACH	indra.sources.reach	https://github.com/clulab/reach
Sparser	indra.sources.sparger	https://github.com/ddmcdonald/sparger
Eidos	indra.sources.eidos	https://github.com/clulab/eidos
TEES	indra.sources.tees	https://github.com/jbjorne/TEES
MedScan	indra.sources.medscan	https://doi.org/10.1093/bioinformatics/btg207
RLIMS-P	indra.sources.rlimsp	https://research.bioinformatics.udel.edu/rlimsp
ISI/AMR	indra.sources.isi	https://github.com/sgarg87/big_mech_isi_gg
Geneways	indra.sources.geneways	https://www.ncbi.nlm.nih.gov/pubmed/15016385
GNBR	indra.sources.gnbr	https://zenodo.org/record/3459420

INDRA assembly standardizes concepts and finds relationships between causal statements

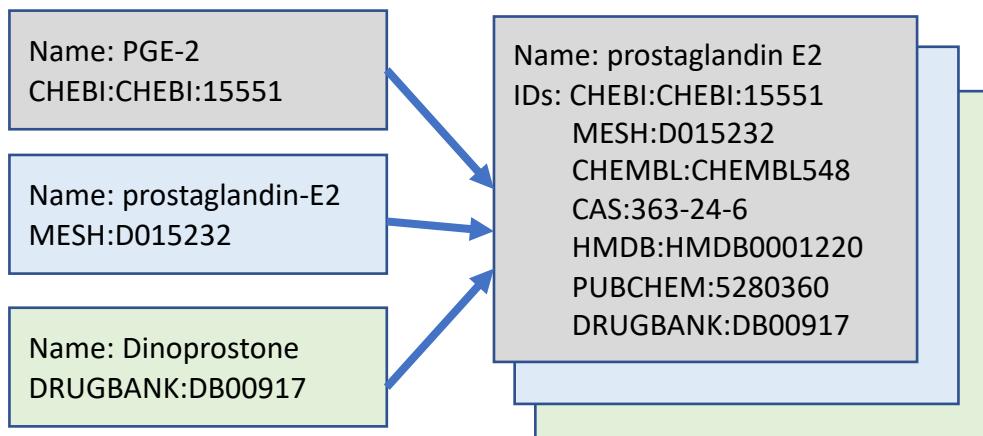
INDRA ontology graph cross-references



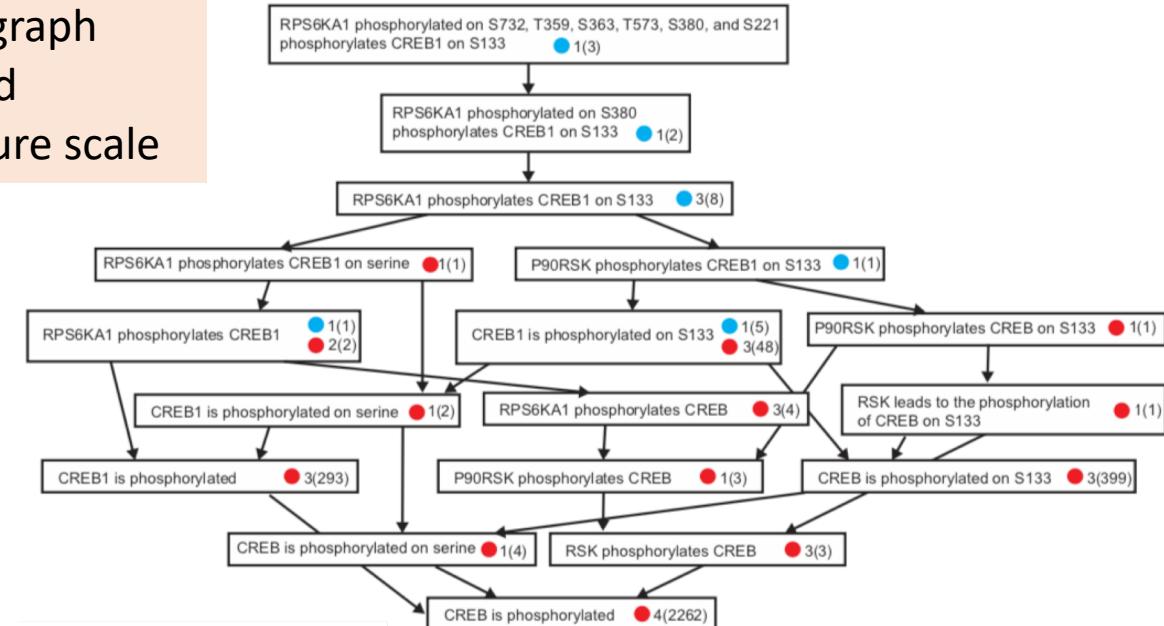
INDRA ontology family relationships induce relationships between statements

	SUBJECT	PREDICATE	OBJECT
S1:	ibuprofen CHEBI:5855	inhibits	nociceptive pain MESH:D059226
S2:	COX inhibitors CHEBI:35544	inhibits	pain MESH:D010146

Concept normalization from diverse sources based on cross-references



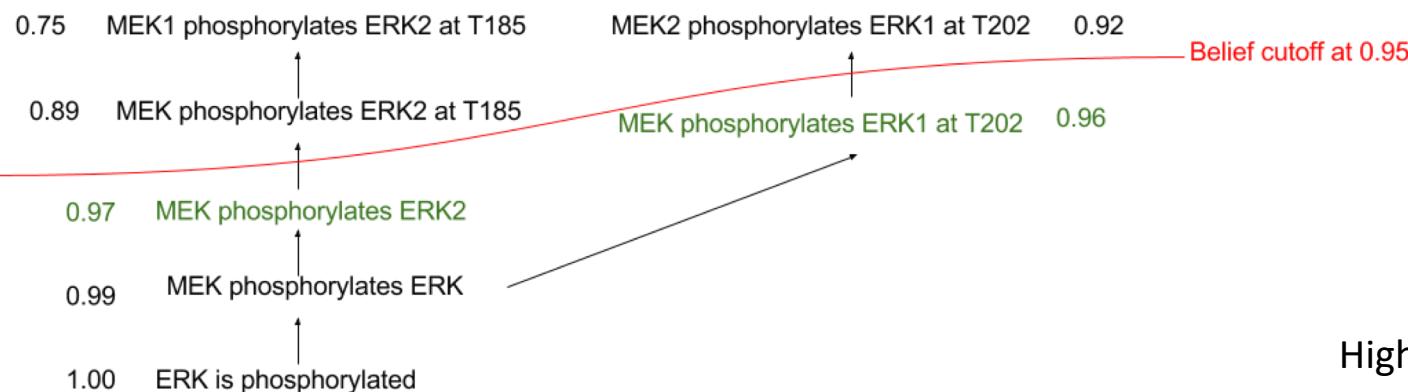
Statement support graph recovered at literature scale



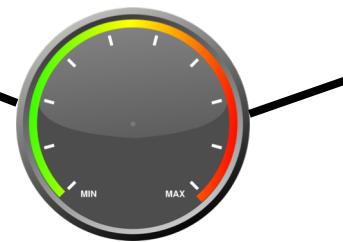
INDRA uses a probability model to determine a belief score

Estimate reliability of Statements probabilistically by:

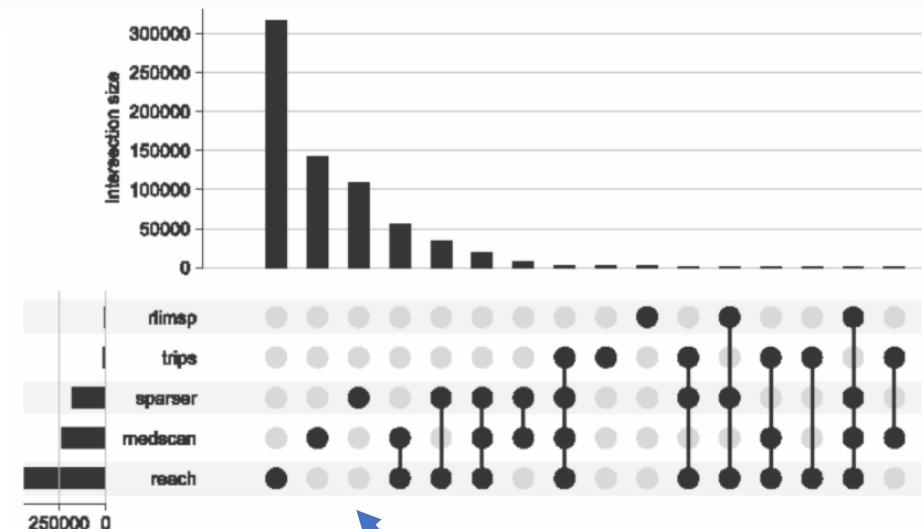
- Calculating joint probability of an incorrect statement given repeated extractions from different sentences
- Combining results from different readers
- Propagating error estimates through the network of related statements



Low cutoff: allows uncorroborated results into model



High cutoff: the mechanism needs evidence from multiple sentences / systems / with database backup to reach high belief level



Machine-learning approaches based on curation data

Large-scale continuous assembly of causal mechanisms with INDRA

Bioinformatics
literature
text mining

Expert
curated
database

Results

I found statements that are from papers with MeSH ID C000657245, are not only from medscan, and have an agent where HGNC=11876 with role=OBJECT

ERG affects TMPRSS2 — 295 8

ACE2 affects TMPRSS2 — 1 72 17

camostat affects TMPRSS2 3 6 41

Camostat inhibits TMPRSS2. 3 6 37

Camostat inhibits TMPRSS2. 3 6 37

reach However, the mechanism of inhibition of **TMPRSS2** by **Camostat** mesylate, per se, has not been clearly elucidated. 32469279

eidos Next , using the established assay in 384-well format , we tested the inhibition of **TMPRSS2** by **camostat** , nafamostat and gabexate (Figure 3) . 32596694

reach **Camostat** mesylate (a serine protease inhibitor), well documented for the treatment of cancer, pancreatitis and liver fibrosis, has recently been shown to inhibit **TMPRSS2** and entry of SARS-CoV-2 virus into the cells (Hoffmann et al., 2020). 32469279

eidos It was reported the ten-fold reduction of SARS-CoV titers in the Calu-3 cells with the inhibition of **TMPRSS2** by **camostat** (Shulla et al. 2011) . 32546018

Select error type... ▾ Optional description (240 chars)

Prior Curations ↻

7/4/2020, 5:18:20 PM ben.gyori@gmail.com correct No text given. EMMAA

Our Sources:

Databases

- Causal Bionet
- PathwayCommons Signor
- CTD BioGRID DrugBank
- VirHostNet TRRUST
- Phosphosite Plus BEL
- HPRD
- Target Affinity Spectrum
- Phospho.ELM

Reading

- ISI/AMR RLIMS-P Eidos
- TRIPS/DRUM MedScan
- Sparser REACH

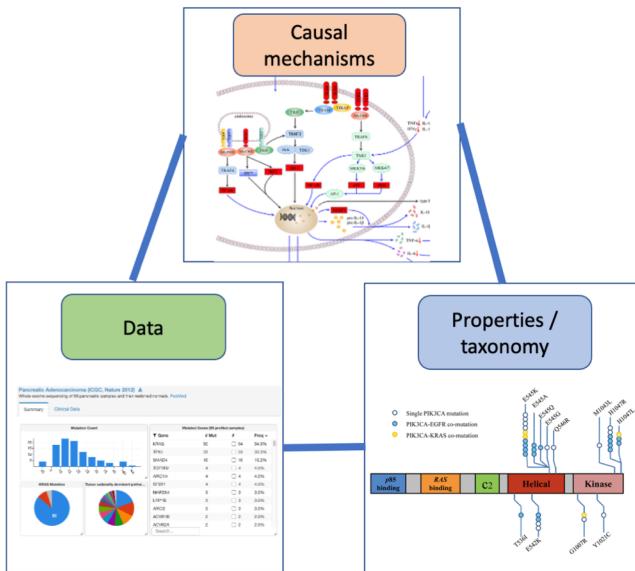
psp cbn pc bel_c signor biogrid tas hprd trrust

ctd vhn pe drugbank omnipath conib crog dgi minerva creeds

knowledge assembly

INDRA Context Graph Extension (CoGEx) – knowledge graph

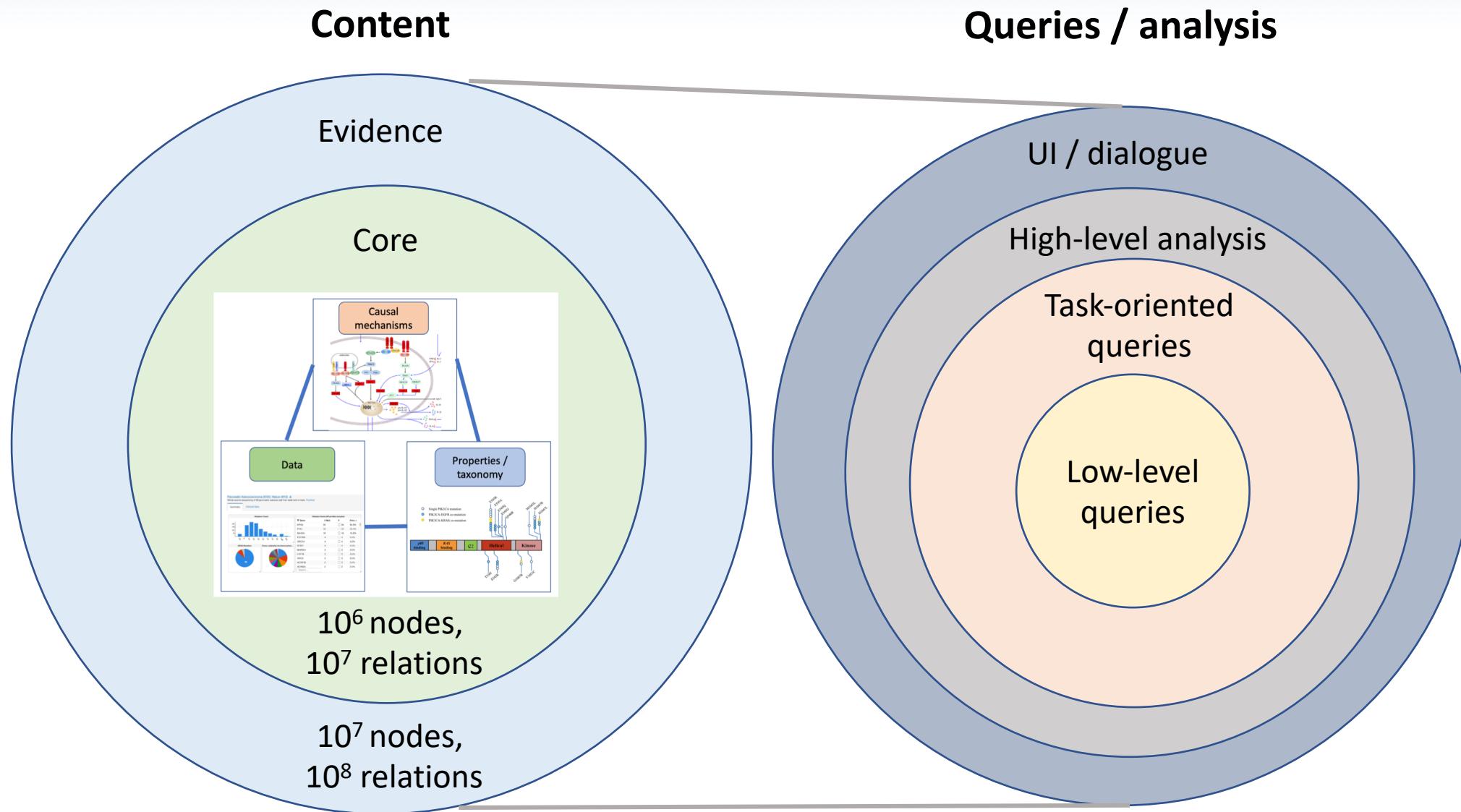
Idea: combine causal mechanisms assembled by INDRA with relations representing data and properties/taxonomy in a knowledge graph



2M		365K		34M		22M	
Biological Entities		Clinical Trials		Publications		Evidences	
310M	21M	7M	4M	1M	1M	1M	619K
MeSH Annotations	Citations	Causal Relations	Gene Expressions	CNVs	Mutations	Xrefs	Part Of
536K	535K	429K	309K	254K	159K	69K	46K
Disease Trials	Subclasses	Has Part	Side Effects	Drug Trials	GO Annotations	Sensitivities	Drug Indications

Knowledge source	Relation	Description
INDRA	indra_rel	The source regulates or interacts with the target according to an INDRA Statement.
INDRA Ontology	isa / partof	The source node is a subclass or part of the target node.
Gene Ontology	associated_with	The gene represented by the source is associated with the GO term represented by the target.
BGee	expressed_in	The gene represented by the source is expressed in the tissue/cell type represented by the target.
CCLE	mutated_in / copy_number_altered_in	The gene represented by the source is mutated or its copy number is altered in the cancer cell line represented by the target.
CCLE	sensitive_to	The cancer cell line represented by the source is sensitive to the drug represented by the target.
ClinicalTrials.gov	tested_in	The drug represented by the source is tested in the clinical trial represented by the target.
ClinicalTrials.gov	has_trial	The disease/condition represented by the source has a clinical trial represented by the target.
ChEMBL	has_indication	The chemical represented by the source has been studied for use against the indication represented by the target.
SIDER	has_side_effect	The chemical represented by the source has a side effect represented by the target.
Reactome	haspart	The pathway represented by the source node contains the gene represented by the target node.
WikiPathways	haspart	The pathway represented by the source node contains the gene represented by the target node.

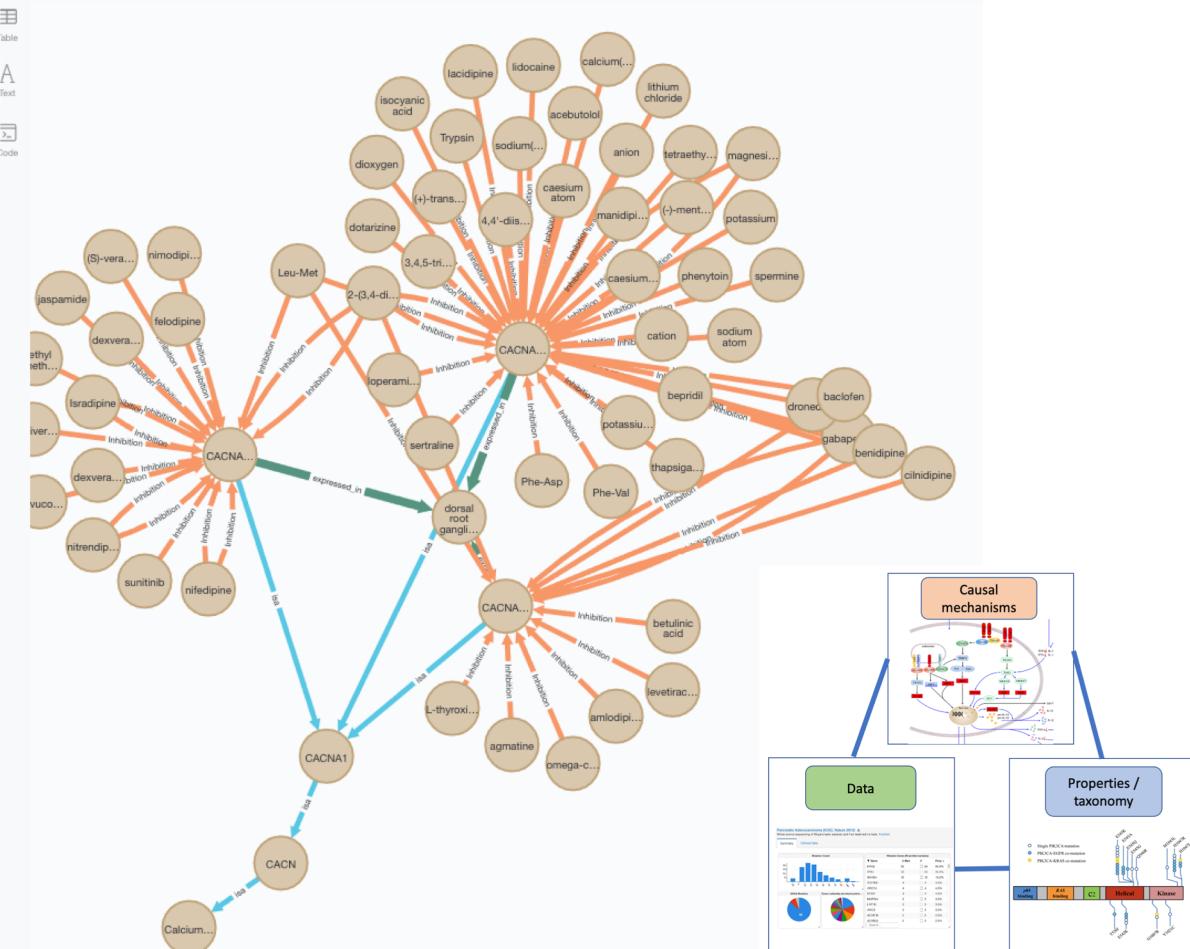
INDRA CoGEx Integrated analysis approach



Querying the INDRA CoGEx KG programmatically

```
1 MATCH p1=(n)-[:isa*1..]→({id: 'fplx:Calcium_channels'})  
2 MATCH p2=(n)-[:expressed_in]→({id: 'uberon:0000044'})  
3 MATCH p3=(d)-[:stmt:Inhibition]→(n)  
4 MATCH p4=(d)-[:isa*1..]→({id: 'chebi:24431'})  
5 RETURN p1, p2, p3  
6 LIMIT 5000
```

graph TD; subgraph "INDRA CoGEx Knowledge Graph"; direction TB; CACNA1 -- "expressed_in" --> dorsal_root_ganglion; dorsal_root_ganglion -- "Inhibition" --> CACNA1; end;



“What drugs target calcium channels that are expressed in dorsal root ganglion?”

discovery.indra.bio/apidocs

INDRA CoGEx Query API ^{1.0}

[Base URL: /]
[/swagger.json](#)

REST API for INDRA CoGEx queries

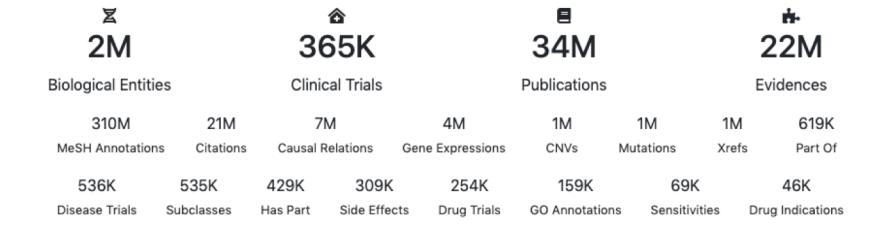
CoGEx Queries Queries for INDRA CoGEx

- POST** [`/api/get_diseases_for_trial`](#) Return the diseases for the given trial
- POST** [`/api/get_drugs_for_side_effect`](#) Return the drugs for the given side effect
- POST** [`/api/get_drugs_for_target`](#) Return the drugs targeting the given protein
- POST** [`/api/get_drugs_for_trial`](#) Return the drugs for the given trial
- POST** [`/api/get_evidences_for_mesh`](#) Return the evidence objects for the given MESH term
- POST** [`/api/get_evidences_for_stmt_hash`](#) Return the matching evidence objects for the given statement ID
- POST** [`/api/get_evidences_for_stmt_hashes`](#) Return the matching evidence objects for the given statement IDs
- POST** [`/api/get_genes_for_go_term`](#) Return the genes associated with the given GO term
- POST** [`/api/get_genes_for_pathway`](#) Return the genes for the given pathway
- POST** [`/api/get_genes_in_tissue`](#) Return the genes in the given tissue

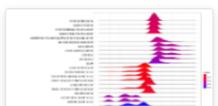
INDRA Biomedical Discovery Engine using the CoGEx KG

INDRA Biomedical Discovery Engine

The INDRA Biomedical Discovery Engine is built on INDRA CoGEx, a graph database integrating causal relations, ontological relations, properties, and data, assembled at scale automatically from the scientific literature and structured sources.



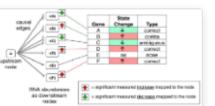
Apps and Services using INDRA CoGEx



Discrete Gene List Analysis

Performs gene set enrichment analysis using INDRA CoGEx.

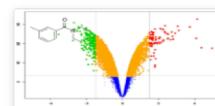
[Click to Start](#)



Signed Gene List Analysis

Perform signed gene set enrichment analysis using INDRA CoGEx and the Reverse Causal Reasoning algorithm.

[Click to Start](#)



Continuous Gene Set Enrichment Analysis

Perform gene set enrichment analysis on continuous data using INDRA CoGEx.

[Click to Start](#)

Discrete Metabolite List Analysis

Perform metabolite set enrichment analysis using INDRA CoGEx.

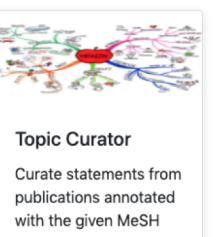
[Click to Start](#)

Directed Curation



GO Pathway Curator

Curate networks induced by genes associated with a given



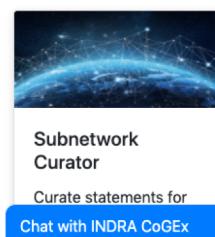
Topic Curator

Curate statements from publications annotated with the given MeSH term



Publication Curator

Curate statements from a given publication.



Subnetwork Curator

Curate statements for

[Chat with INDRA CoGEx](#)

Gene List Analysis

Query

HGNC:11133 (SNAP29) HGNC:4910 (HIF1A) HGNC:2734 (DDX1) HGNC:1116 (BSG) HGNC:11432 (STX17) HGNC:18348 (TICAM1)
HGNC:23785 (PIKFYVE) HGNC:4931 (HLA-A) HGNC:1697 (CD74) HGNC:16404 (IL17F) HGNC:5432 (IFNAR1) HGNC:29517 (DHX58)
HGNC:11878 (TMPRSS2) HGNC:9788 (RAB7A) HGNC:28948 (TMEM41B) HGNC:4983 (HMGB1) HGNC:6727 (LY6E)
HGNC:20593 (VPS39) HGNC:10819 (SOTA) HGNC:8912 (PHB) HGNC:6122 (IRF7) HGNC:613 (APOE) HGNC:9253 (PPIA)
HGNC:5985 (IL17RA) HGNC:6118 (IRF5) HGNC:6378 (KLRL1) HGNC:11120 (SMPD1) HGNC:6374 (KLRC1) HGNC:10488 (S100A8)
HGNC:11878 (TMPRSS4) HGNC:4962 (HLA-E) HGNC:8568 (FURIN) HGNC:8120 (IRF5) HGNC:6021 (IL6ST) HGNC:11584 (TBK1)
HGNC:12113 (VPS41) HGNC:3036 (PHB2) HGNC:6769 (SMAD3) HGNC:20820 (TPCN2) HGNC:6019 (IL6R) HGNC:18873 (IFIH1)
HGNC:4932 (HLA-B) HGNC:8004 (NRP1) HGNC:2537 (CTSL) HGNC:1119 (BST2) HGNC:18358 (IL17RC) HGNC:6148 (ITGA6)
HGNC:7087 (CIITA) HGNC:5433 (IFNAR2) HGNC:11849 (TLR3) HGNC:5981 (IL17A) HGNC:6018 (IL6)

Over-Representation Analysis

[GO](#) [Reactome](#) [WikiPathways](#) [INDRA Downstream](#) [INDRA Upstream](#)

These results are acquired by running over-representation analysis using Fisher's exact test and correcting using `fdr_bh` and $\alpha=0.05$ on the genes annotated to terms in the [Gene Ontology](#) via the [Gene Ontology Annotations Database](#).

Show 10 entries

Search:

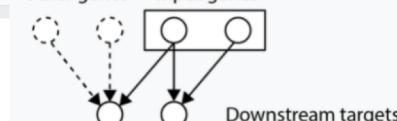
CURIE Name

go:0005515 protein bind
go:0032755 positive reg

Over-Representation Analysis

[GO](#) [Reactome](#) [WikiPathways](#) [INDRA Downstream](#) [INDRA Upstream](#)

Other genes Input genes



These results are acquired by running over-representation analysis using Fisher's exact test and correcting using `fdr_bh` and $\alpha=0.05$ on genes causally downstream in one step from all entities in the INDRA Database.

Show 10 entries

Search:

CURIE	Name	p-value	q-value
go:0006955	immune response	4.90e-41	1.47e-36
go:0001816	cytokine production	1.44e-39	2.16e-35
mesh:D007239	Infections	1.80e-37	1.80e-33
go:0006914	autophagy	1.25e-36	9.41e-33
hgnc:5438	IFNG	2.53e-36	1.52e-32



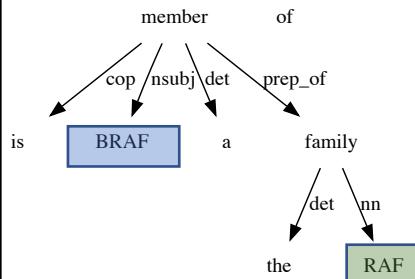
Human-machine dialogue

Example: Is BRAF a member of the RAF family?

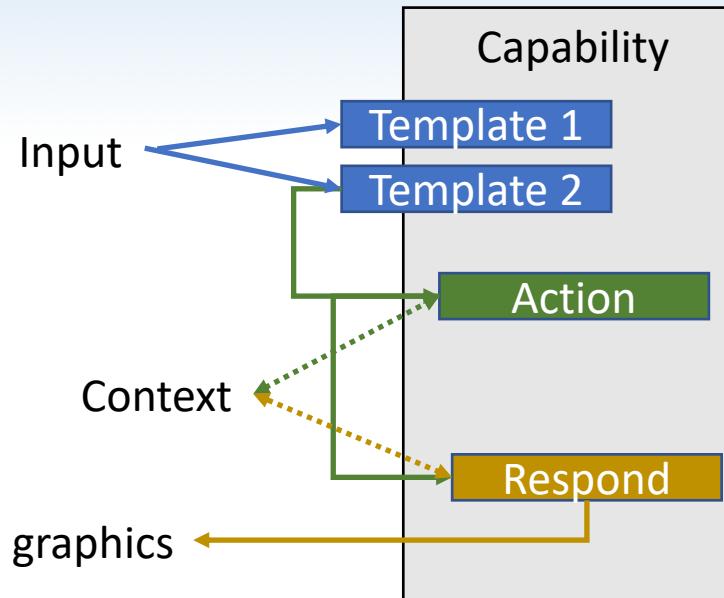
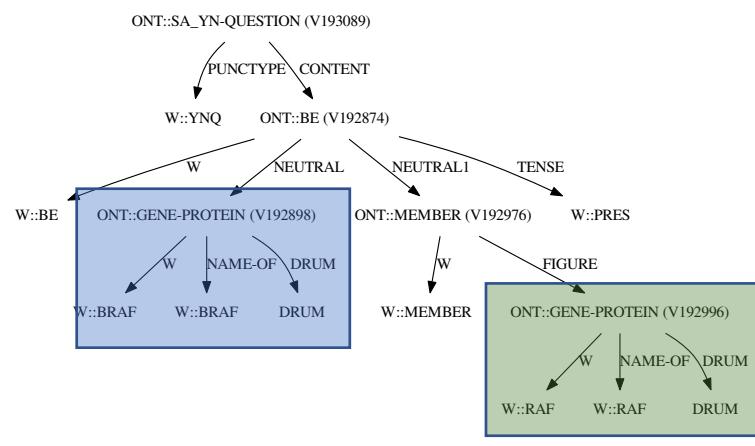
Regular expression templates

is **BRAF** a member of the **RAF** family

Syntactic dependency graph



Semantic interpretation graph



"BRAF" -> HGNC:1097

"RAF" -> FPLX:RAF

Grounding

isa_or_partof

Python API call

MATCH gene:BioEntity ...

Cypher query

Capability matching

IsMember

Coreference resolution

INDRA CoGEx graph

Human-machine dialogue

Private thread

Private thread

Channel

Session Manager

Dialogue Manager context

ben.gyori 1:51 PM
what are the members of calcium channels?

clare-test APP 1:52 PM
I know about 41 members of Calcium_channels: CACNA1H, CACNA1I, CACNA1S, CACNA2D1, CACNG2, CACNG3, CACNG4, CACNG5, CACNG6, CACNG7, CACNG8, CATSPERD, CATSPERG, ITPR1, ITPR2, ITPR3, RYR1, RYR2.

ben.gyori 1:52 PM
are any of those expressed in dorsal root ganglion

clare-test APP 1:53 PM
Out of those, CACNA1A, CACNA1B, CACNA1C, CACNG3, CACNG6, CACNG8, CATSPER2, ITPR1.

ben.gyori 1:53 PM
are there any drugs targeting any of those?

clare-test APP 1:53 PM
I found the following drugs for these targets:

- CACNA1C: nifedipine, nitrendipine, Verapamil
- CACNA2D1: gabapentin
- RYR1: EPIRYANODINE
- RYR2: ALADORIAN

COVID-19 Disease Map curators using dialogue

Curator	Number of Questions
1	150
2	75
3	72
4	48
5	42
6	30
7	28
8	20
9	20
10	18
11	16
12	10
13	10
14	5
15	3
16	2
17	2

Discovery Engine

a graph database integrating causal scale automatically from the scientific

4M Publications

22M Evidences

1M NVs

1M Mutations

1M Xrefs

619K Part Of

159K Annotations

69K 46K

Minimize Chat Window

Ben what does sorafenib inhibit?

CLARE Overall, I know that sorafenib can inhibit, for instance, cell population proliferation, angiogenesis, RAF, ERK, apoptotic process, Carcinoma, Hepatocellular, VEGFR, BRAF, RAF1, and VEGF. (View statements)

Ben which of those are kinases?

CLARE Out of those, BRAF, RAF1, KDR, FLT3, RET, PDGFRB, KIT, FLT4, MTOR, FLT1, EGFR, MAPK14, PDGFRA, ABL1, MAP3K20, CSF1R, DDR2, FGFR1,

Create Your Own Slack Bot

slack

using INDRA CoGEx.

using INDRA CoGEx and the Reverse Causal

enrichment analysis continuous data using

Enter Message

Send

Sequential exploration of causal mechanisms, properties and data through dialogue



bgyori 6:06 PM

what proteins might lead to the development of melanoma?



clare APP 6:06 PM

It looks like BRAF has a mutation frequency of 52% in melanoma so it could be a useful target.



bgyori 6:06 PM

is it a kinase?



clare APP 6:06 PM

Yes, BRAF is a kinase.



bgyori 6:06 PM

what does it phosphorylate?



clare APP 6:06 PM

Overall, I know that BRAF can phosphorylate, for instance, MEK, MAP2K1, ERK, MAP2K2, BRAF, MAPK7, PAX3, Pax3, BAD, and RAF1.

You can browse the full set of 177 statements here:

<https://s3.amazonaws.com/indrabot-results/49c73701-b9a8-4c12-97ad-9718f70df344.html>

Data

Properties / taxonomy

Causal mechanisms



bgyori 8:40 PM

do any of those activate EMT?



clare APP 8:41 PM

Out of those, I know that for instance, AKT, NFkappaB, ERK, STAT3, EGFR, SRC, MAPK, NANOG, MET, and TERT can activate epithelial to mesenchymal transition.

You can browse the full set of 72 statements here:

<https://s3.amazonaws.com/indrabot-results/2d48594c-ac79-4067-93bc-64e471962989.html>

Causal mechanisms



bgyori 8:44 PM

are there any drugs that target BRAF?



clare APP 8:44 PM

I know that ALW-II-38-3, AZ-628, dabrafenib, Encorafenib, GDC-0879, GSK-718429A, GW-5074, L-779450, LY3009120, PD0325901, PD173955, PLX-4720, regorafenib, SB590885, sorafenib, staurosporine, TAK-632, and vemurafenib target BRAF.

Causal mechanisms



bgyori 8:45 PM

are there any clinical trials attempting sorafenib?



clare APP 8:45 PM

I found the following clinical trials for sorafenib:

- preoperative trial of sorafenib in combination with cisplatin followed by paclitaxel for early stage breast cancer - terminated
- sorafenib and vinorelbine in treating women with stage iv breast cancer - completed
- phase i/ii trial of letrozole and sorafenib in post-menopausal hormone-receptor + breast cancer - active, not recruiting
- sorafenib and paclitaxel in treating patients with metastatic breast cancer - terminated

Data

Acknowledgments

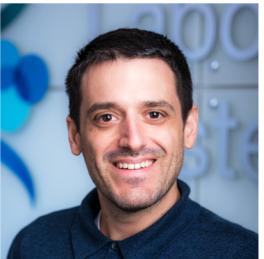


Peter Sorger
Laura Maliszewski
Catherine Luria

Charlie Hoyt
Klas Karis
Diana Kolusheva
Samuel Bunga



John Bachman
Albert Steppi
Patrick Greene



Website: indralab.github.io
Twitter: @IndraSysBio

Funding
DARPA Young Faculty Award
(W911NF2010255)

