Explanation of drug effects using a mechanistic model automatically assembled from natural language, databases, and literature

John A. Bachman*
Benjamin M. Gyori*
Peter K. Sorger
Laboratory of Systems Pharmacology
Harvard Medical School

DARPA Big Mechanism
DARPA Communicating with Computers

The “unmet need” for mechanistic explanation of large-scale data

Joe Cornish, https://www.biostars.org/p/119918/

On “big data” hype in bioinformatics:

“The rate limiting factor has never been the computational power, and is more infrequently a result of not having enough data, the problem has and still is that no matter how much data is generated or how much cleaner/precise/etc the data is, I still can't do a whole lot of anything with it **because the ability to turn these piles of data into information is feeble at best.** Whether it be the latest Illumina tech or the hottest MS approach, **all you get is a list of p-values that you dump into your pathway enrichment tool of choice, [generate] a few heatmaps and clustering diagrams and call it a day.”**

“Unmet need” for mechanistic explanation of large-scale data

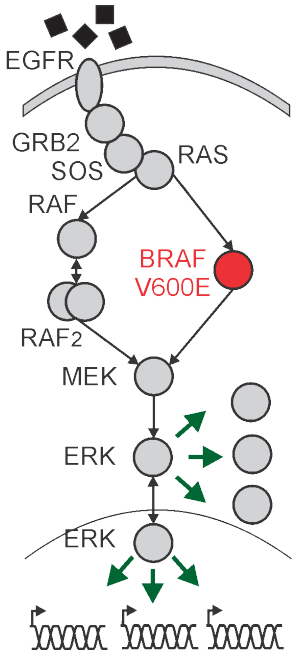
“Bottom-up”

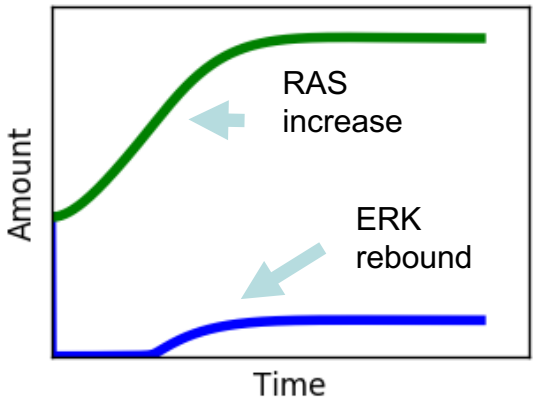
Detailed mechanistic studies

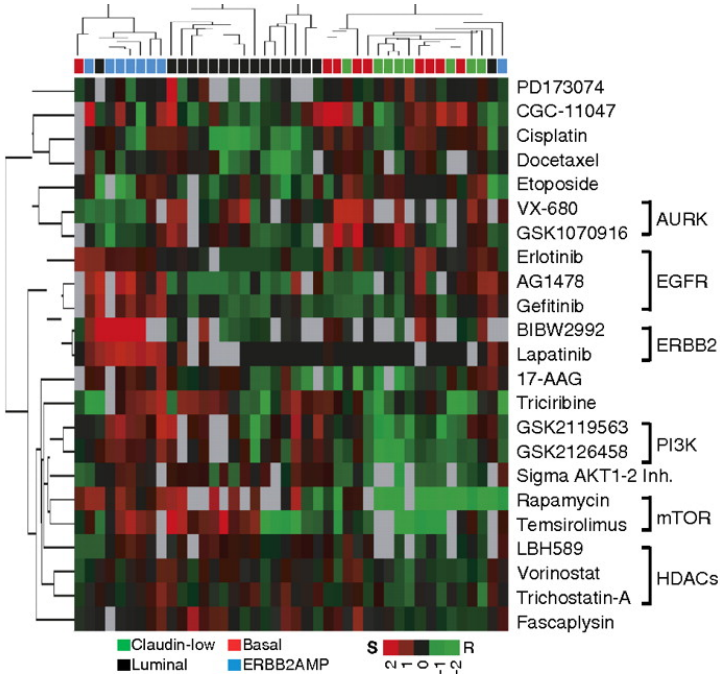
“Top-down”

Interpretation of large datasets

Heiser et al., 2011

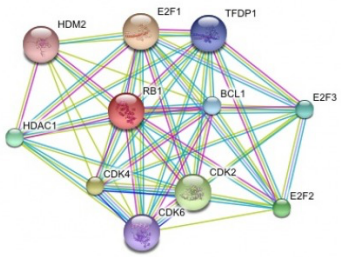






Approach: Machine assembly of detailed models at large scale.

Conceptual overview of the modeling process



Pathway databases

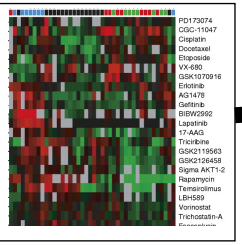
Literature

Expert input

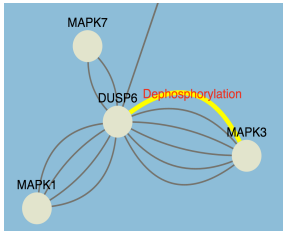
Knowledge assembly

Executable models

Contextual data

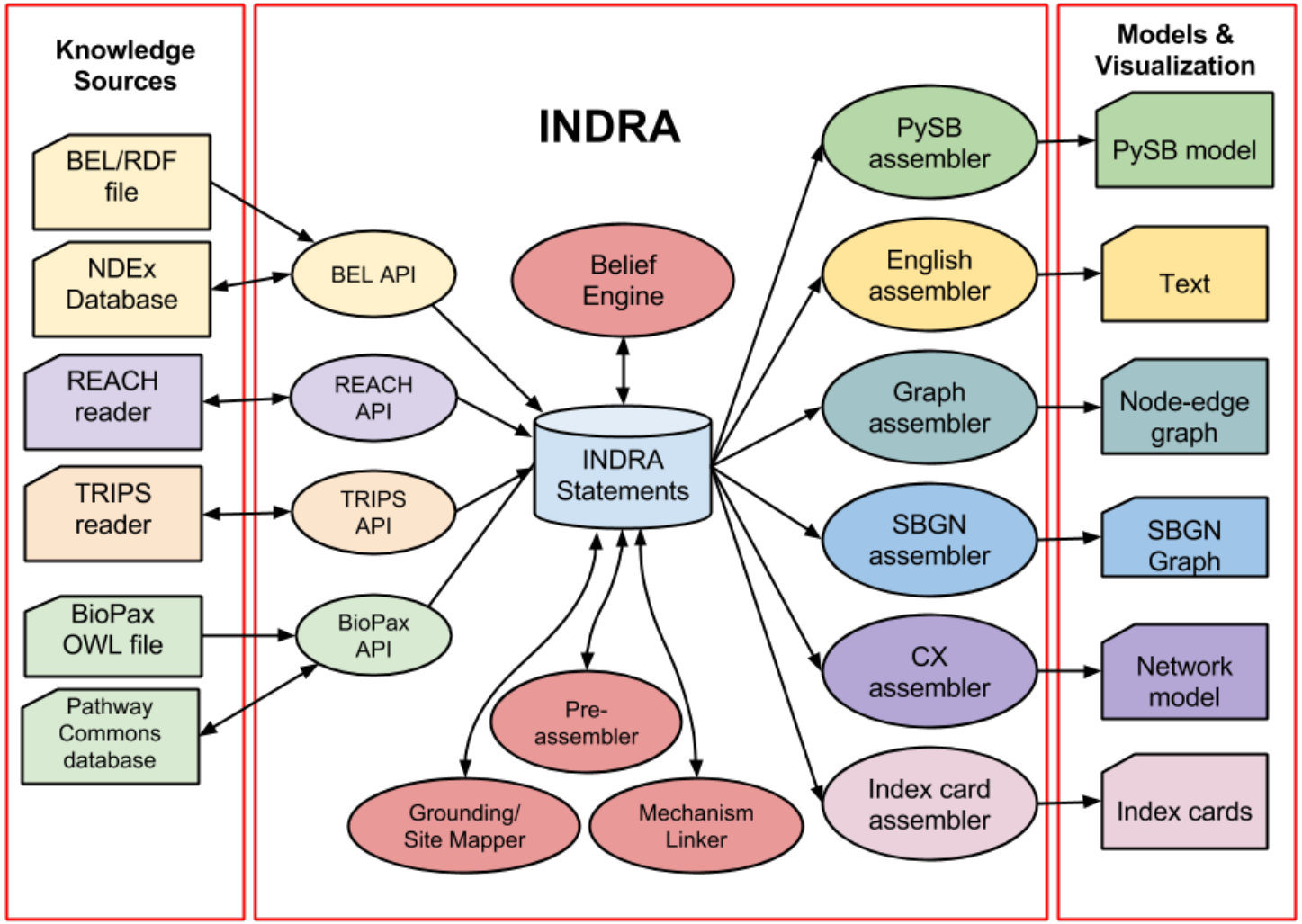


Explanation, Prediction



Mechanistic networks

INDRA: Integrated Network and Dynamical Reasoning Assembler



www.indra.bio

Sources and formats of mechanistic information: pathway databases

BioPAX: Pathway Commons (22 databases)

- Mechanistic information expressed as biochemical reactions
  - NCI Pathway Interaction Database
  - Reactome
  - BioGRID
  - PhosphoSitePlus
  - KEGG
  - et al.

Biological Expression Language

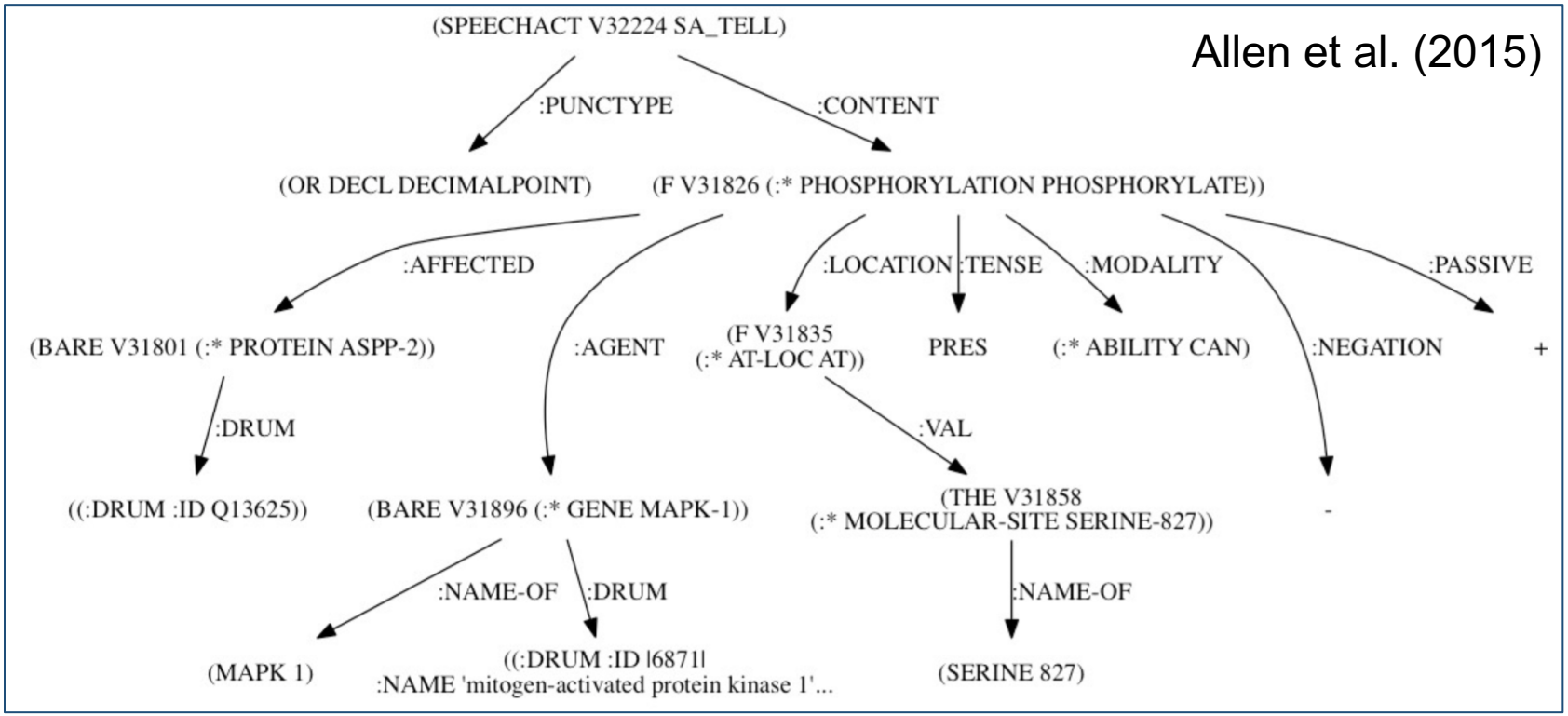
- BEL Large Corpus
  - BEL expresses causal relations
  - ~80,000 assertions
  - Both *mechanistic* and *observational* assertions

Extracting mechanisms from natural language (1)

TRIPS system: general purpose, deep, semantic reading with domain-specific ontologies

ihmc

"ASPP2 can be phosphorylated at serine 827 by MAPK1."



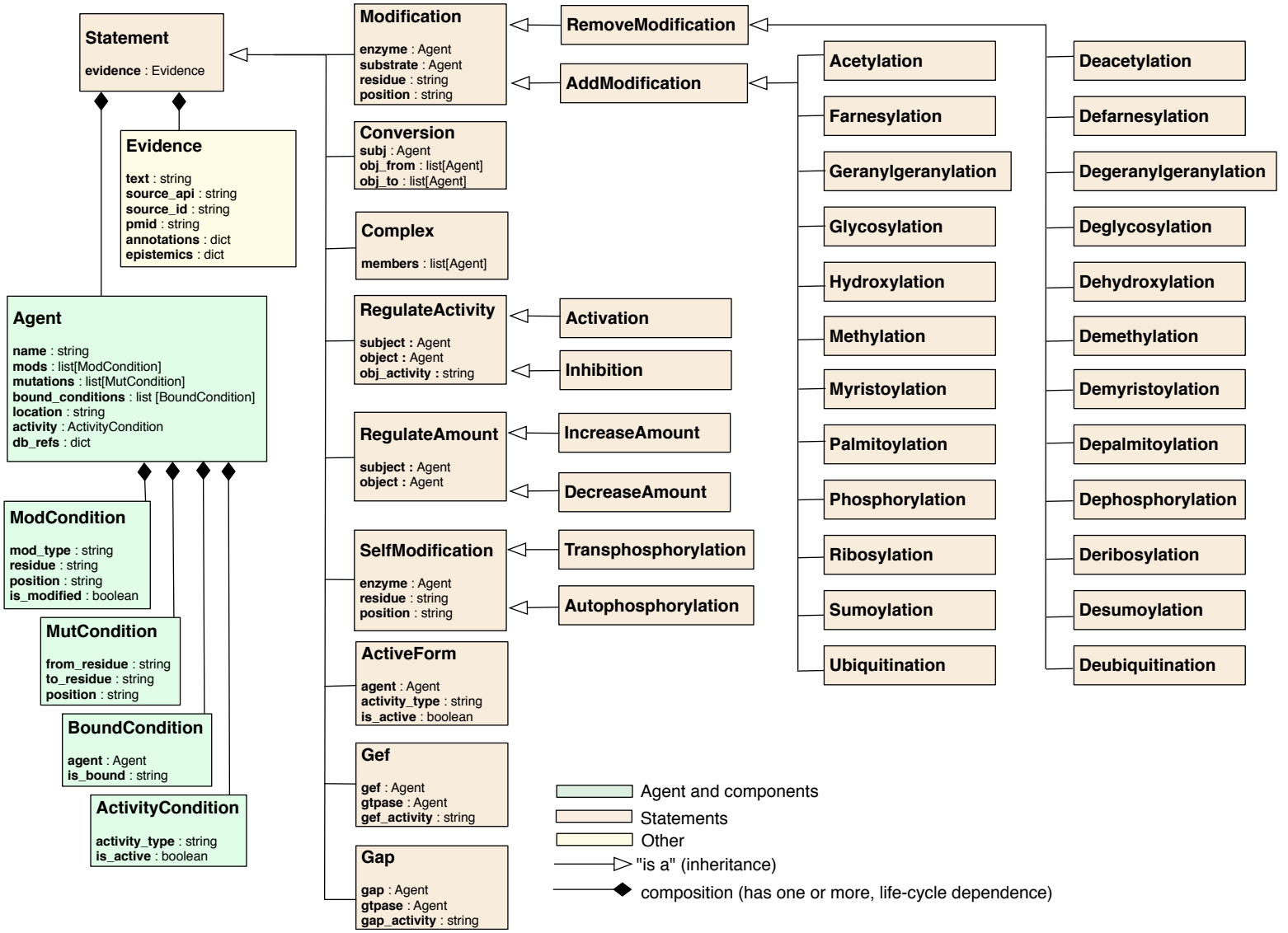
Allen et al. (2015)

Extracting mechanisms from natural language (2)

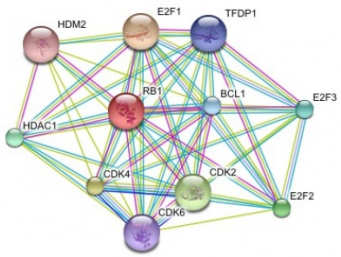
REACH system: domain-specific set of patterns used to identify mechanisms in text

```
- name: Positive_${ ruleType }_syntax_8_verb
  priority: ${ priority }
  example: "We found that prolonged expression of active Ras resulted in up-regulation of the MKP3 gene"
  label: ${ label }
  action: ${ actionFlow }
  pattern: |
    trigger = [lemma=result] in [word=/(?i)^(${ triggers })/]
    controlled:${ controlledType } = prep_of nn?
    controller:${ controllerType } = nsubj /appos|nn|prep_of|amod|conj_|cc/{,2}
```

INDRA represents detailed biochemical entities and their interactions



From knowledge to model-based explanations: Assembly



Pathway databases

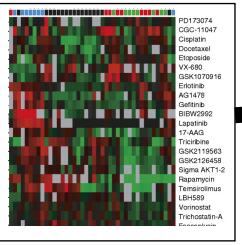
Expert input

Literature

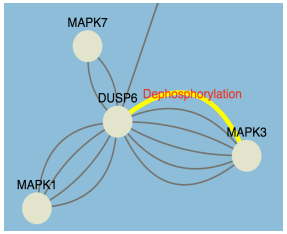
Knowledge assembly

Executable models (PK/PD)

Contextual data



Explanation, Prediction



Mechanistic networks

Knowledge assembly is like genome assembly

MEK phosphorylates ERK

MEK phosphorylates ERK

Methyl Ethyl Ketone phosphorylates ERK

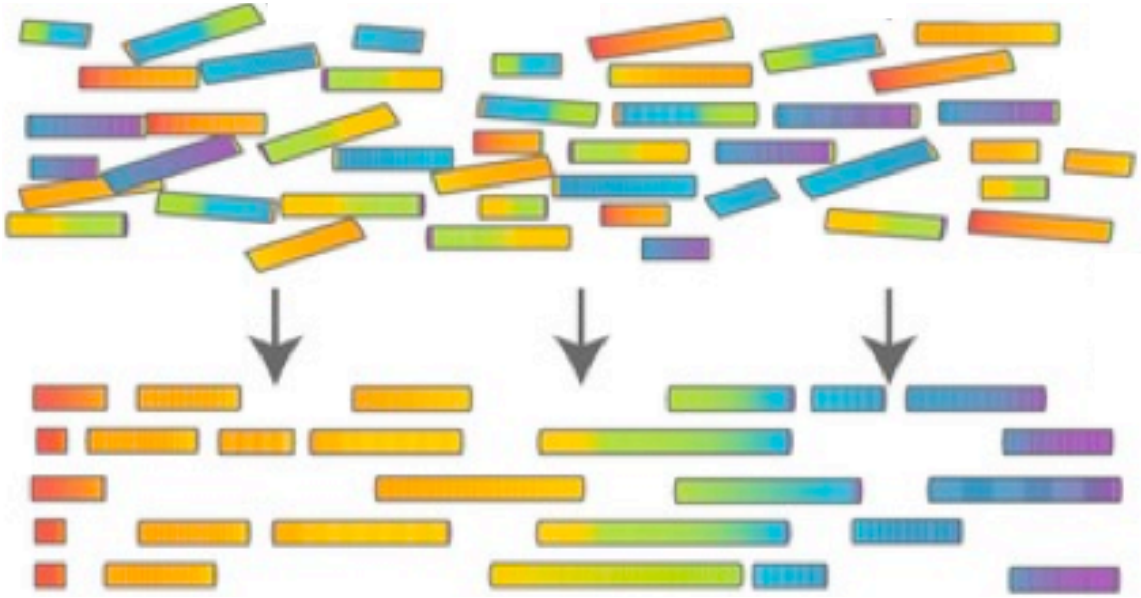
MEK phosphorylates ERK

MEK1 phosphorylates ERK2 at T185

ERK phosphorylates MEK

MEK1p218p222 phosphorylates ERK2 at T184

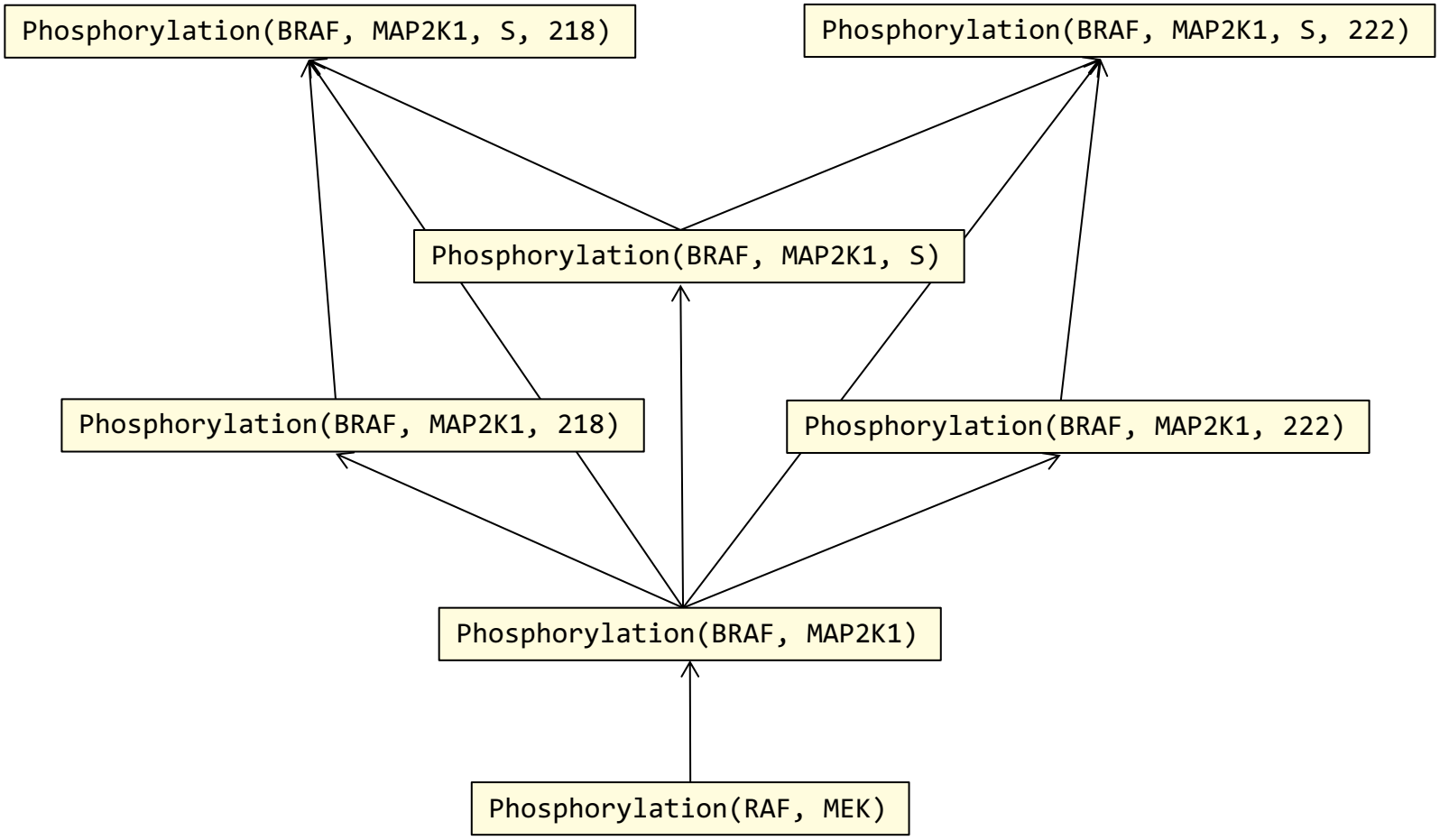
"Raw" mechanisms



MEK1p218p222 phosphorylates ERK2 at T185.

Assembled mechanisms

INDRA assembly resolves hierarchical redundancies



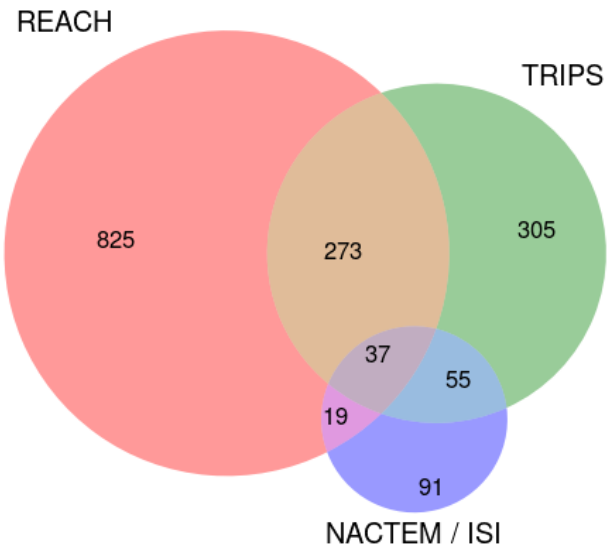
Combine entity, modification, location and activity hierarchies

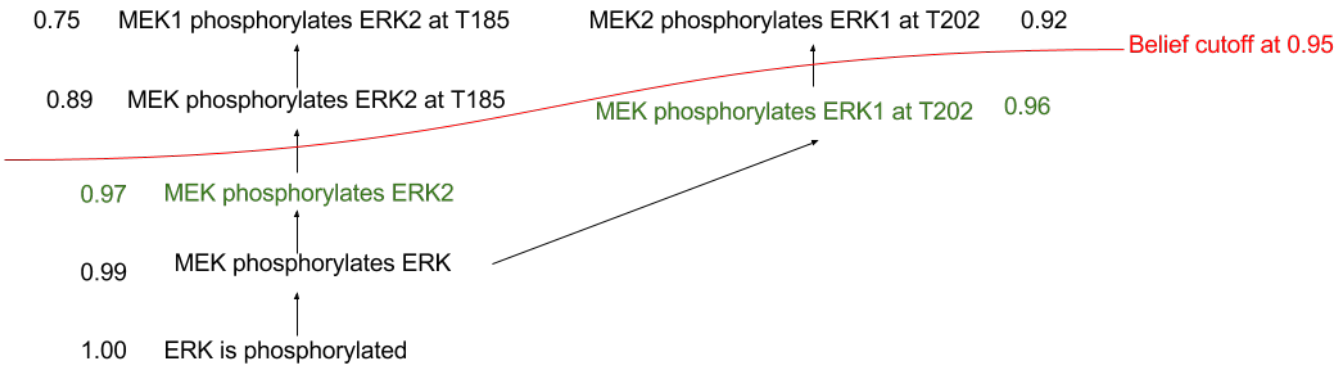
INDRA uses belief propagation to determine probability of correctness

Estimate reliability of Statements probabilistically by:

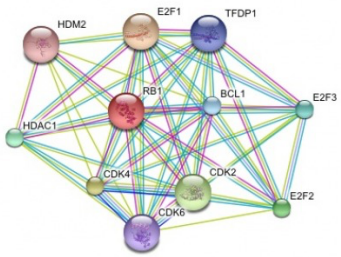
- Calculating joint probability of an incorrect statement given repeated extractions from different sentences
- Combining results from different readers
- Propagating error estimates through the network of related statements

Overlap between readers





From knowledge to model-based explanations: Assembly



Pathway databases

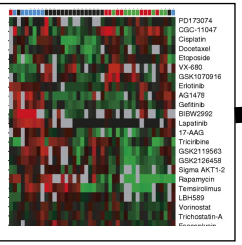
Expert input

Literature

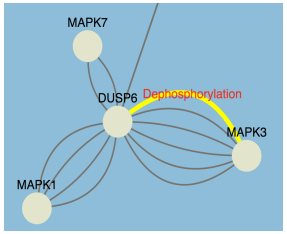
Knowledge assembly

Executable models (PK/PD)

Contextual data



Explanation, Prediction

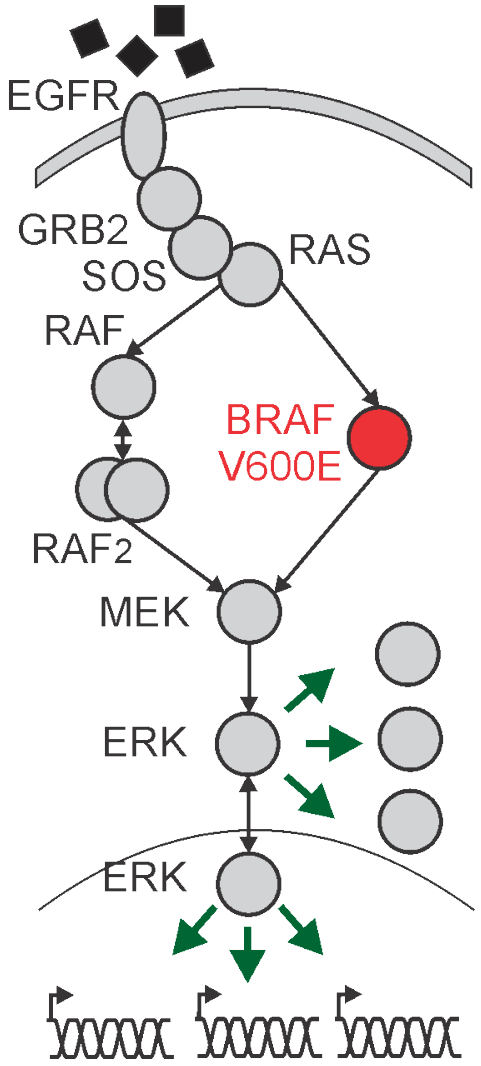


Mechanistic networks

Application: “Natural language modeling”

Word Model

EGFR binds the growth factor ligand EGF.
The EGFR-EGF complex binds another EGFR-EGF complex.
The EGFR-EGFR complex binds GRB2.
EGFR-bound GRB2 binds SOS1 that is not phosphorylated.
GRB2-bound SOS1 that is not phosphorylated binds NRAS that is not bound to BRAF.
SOS1-bound NRAS binds GTP.
GTP-bound NRAS that is not bound to SOS1 binds BRAF.
NRAS-bound BRAF binds NRAS-bound BRAF.
Vemurafenib binds BRAF that is not bound to BRAF.
Vemurafenib binds BRAF-bound BRAF.
BRAF V600E that is not bound to Vemurafenib phosphorylates MAP2K1.
PP2A-alpha dephosphorylates MAP2K1 that is not bound to ERK2.
Phosphorylated MAP2K1 is activated.
Active MAP2K1 that is not bound to PP2A-alpha phosphorylates ERK2.
Phosphorylated ERK2 is activated.
DUSP6 dephosphorylates ERK2 that is not bound to SOS1.
Active ERK2 that is not bound to DUSP6 phosphorylates SOS1 that is not bound to NRAS.
A phosphatase dephosphorylates SOS1.

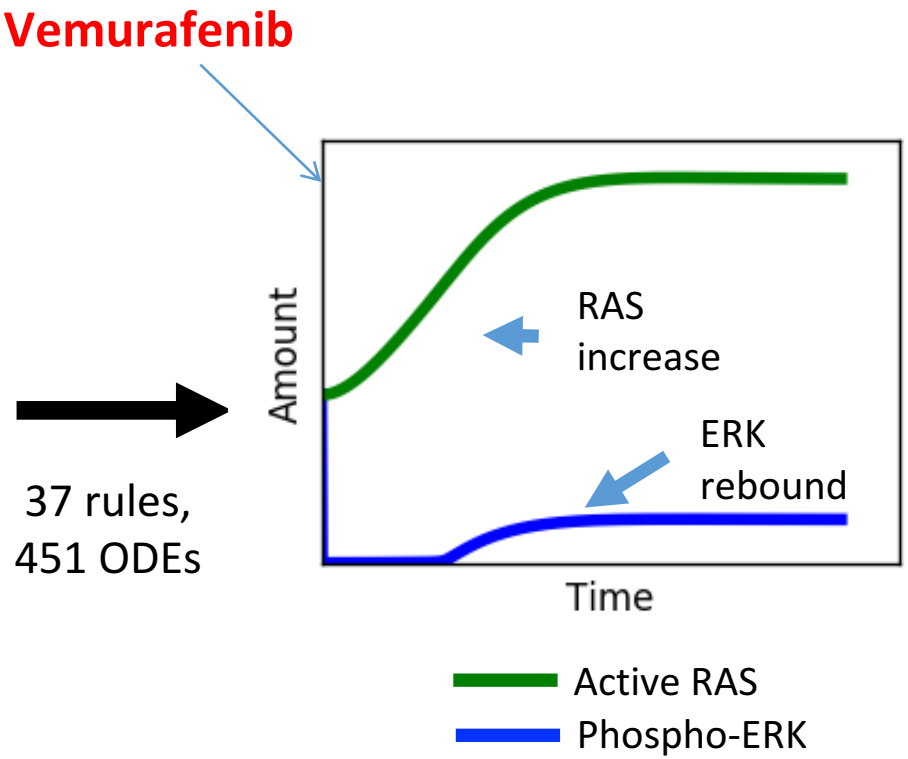


(Luca Gerosa)

Application: “Natural language modeling”

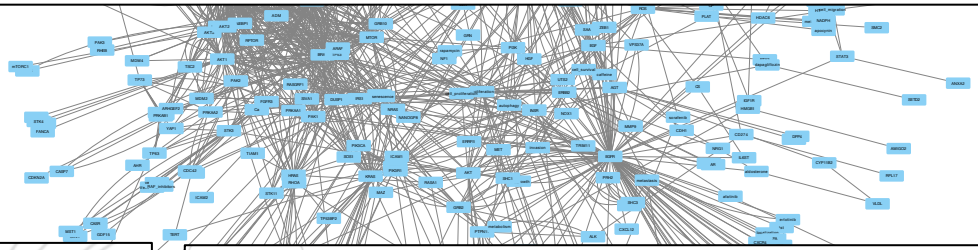
Word Model

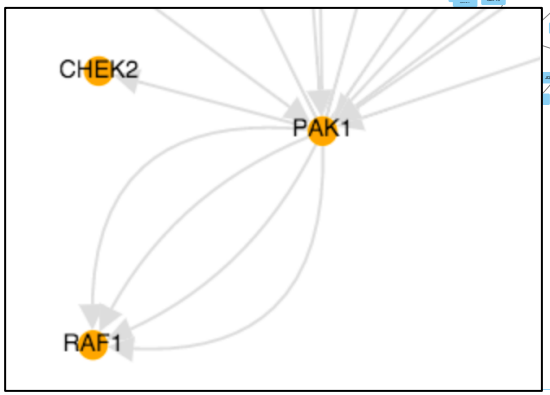
EGFR binds the growth factor ligand EGF.
The EGFR-EGF complex binds another EGFR-EGF complex.
The EGFR-EGFR complex binds GRB2.
EGFR-bound GRB2 binds SOS1 that is not phosphorylated.
GRB2-bound SOS1 that is not phosphorylated binds NRAS that is not bound to BRAF.
SOS1-bound NRAS binds GTP.
GTP-bound NRAS that is not bound to SOS1 binds BRAF.
NRAS-bound BRAF binds NRAS-bound BRAF.
Vemurafenib binds BRAF that is not bound to BRAF.
Vemurafenib binds BRAF-bound BRAF.
BRAF V600E that is not bound to Vemurafenib phosphorylates MAP2K1.
PP2A-alpha dephosphorylates MAP2K1 that is not bound to ERK2.
Phosphorylated MAP2K1 is activated.
Active MAP2K1 that is not bound to PP2A-alpha phosphorylates ERK2.
Phosphorylated ERK2 is activated.
DUSP6 dephosphorylates ERK2 that is not bound to SOS1.
Active ERK2 that is not bound to DUSP6 phosphorylates SOS1 that is not bound to NRAS.
A phosphatase dephosphorylates SOS1.



Preprint:
“From word models to executable models of signaling networks using automated assembly”
http://www.biorxiv.org/content/early/2017/03/24/119834

Application: @TheRasMachine, a self-updating network model of Ras



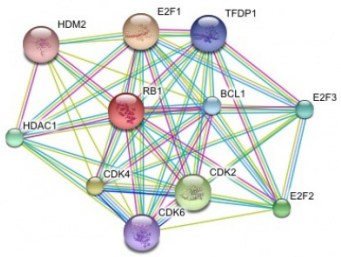


Edges | Nodes | Provenance

Subject	Predicate	Object	Citations	INDRA statement
PAK1				
PAK1	ActivityActivity	RAF1	3	ActivityActivity(PAK1(), activity, increas...
PAK1	Complex	RAF1	1	Complex(PAK1(), RAF1())
PAK1	Phosphorylation	RAF1	1	Phosphorylation(PAK1(), RAF1(), S, 338)
PAK1	Phosphorylation	RAF1		Phosphorylation(PAK1(), RAF1(), S, 339)

http://ndexbio.org/#/network/50e3dff7-133e-11e6-a039-06603eb7f303

From knowledge to model-based explanations: Explanation



Pathway databases

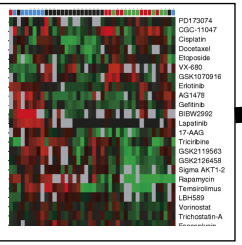
Expert input

Literature

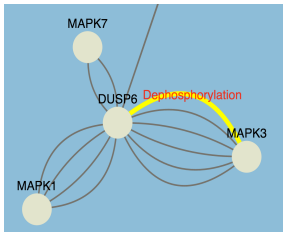
Knowledge assembly

Executable models (PK/PD)

Contextual data

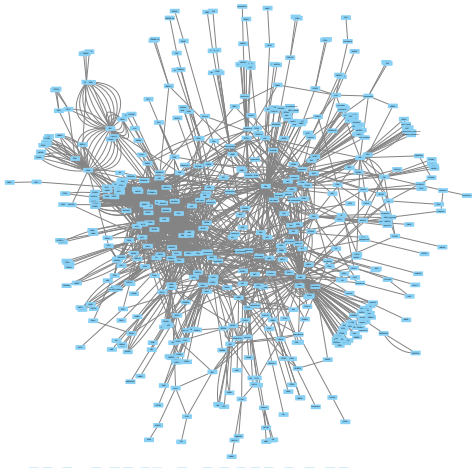


Explanation, Prediction

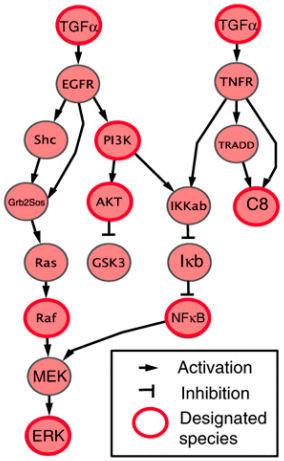


Mechanistic networks

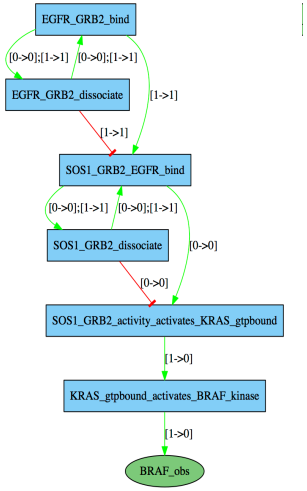
Model representations for statically identifying causal paths



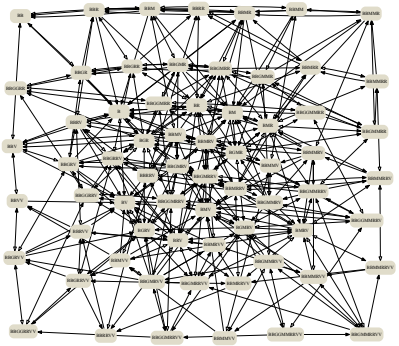
Directed protein interaction graph



Logical network



Kappa rule influence map



Chemical reaction network

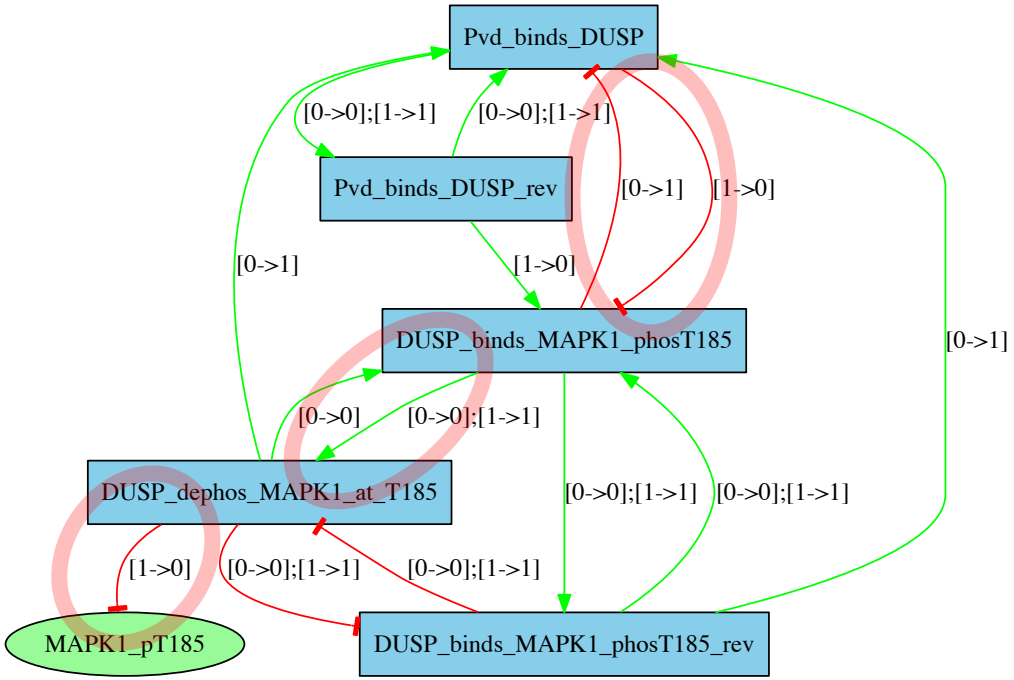
Mechanistic detail/causal context

More false positive paths
(less stringent context)

More false negative paths
(more stringent context)

Causal analysis of the *influence map* of a rule-based model (Kappa)

Explain: "Pervanadate increases MAPK1 phosphorylation."



Pros: Meaningful explanations; fewer false positives; link to simulation

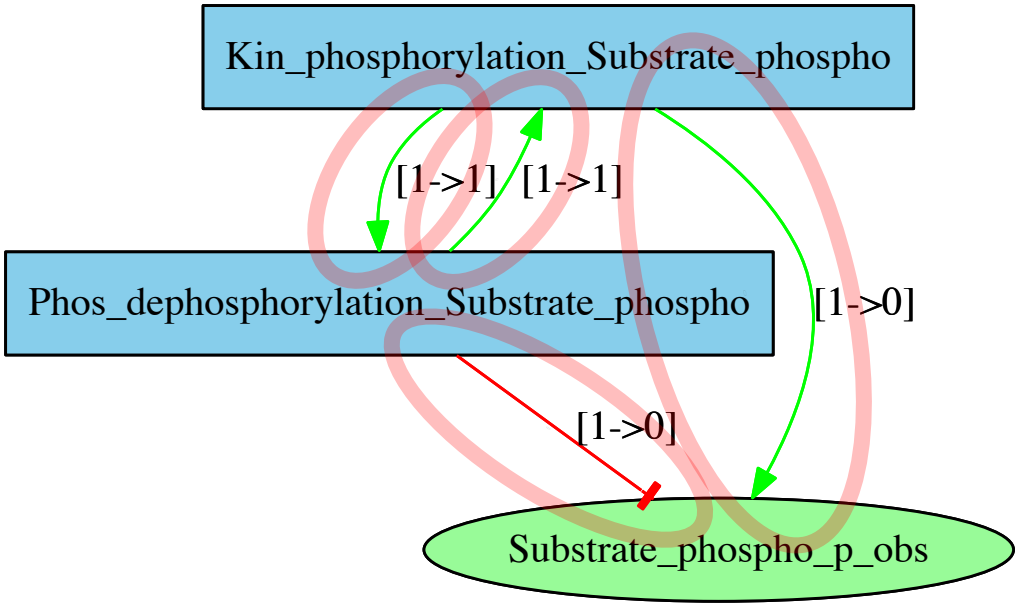
*Cons: False positives possible depending on model—**influence is not necessarily transitive***

Ioana Cristescu, Pierre Boutillier, Jerome Feret, Walter Fontana et al.

The problem of “transitive triangles” in the influence map

Model:

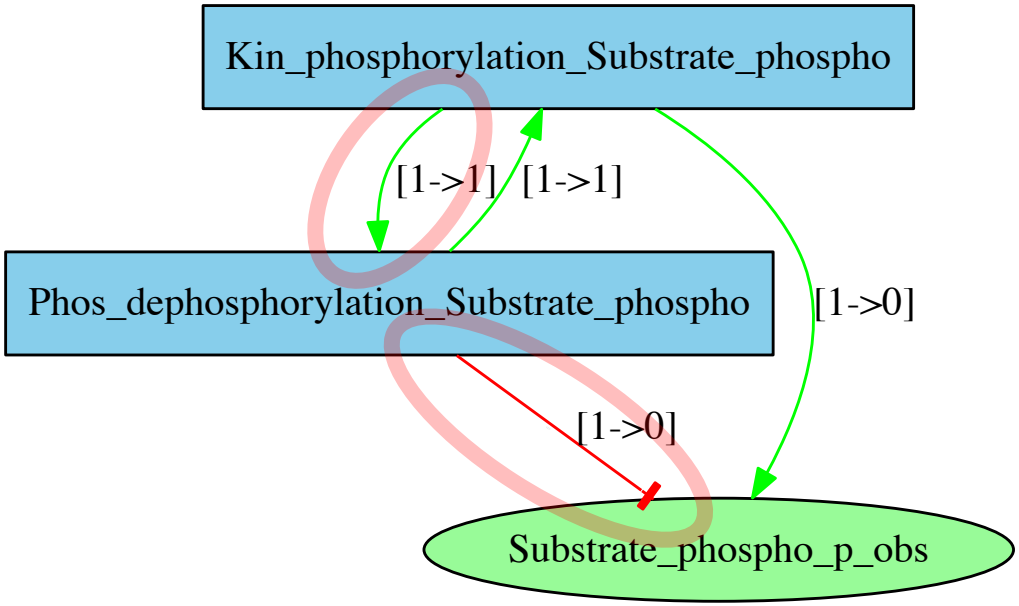
```
Kinase phosphorylates Substrate
Phosphatase dephosphorylates Substrate
```



The problem of “transitive triangles” in the influence map

Model:

```
Kinase phosphorylates Substrate
Phosphatase dephosphorylates Substrate
```

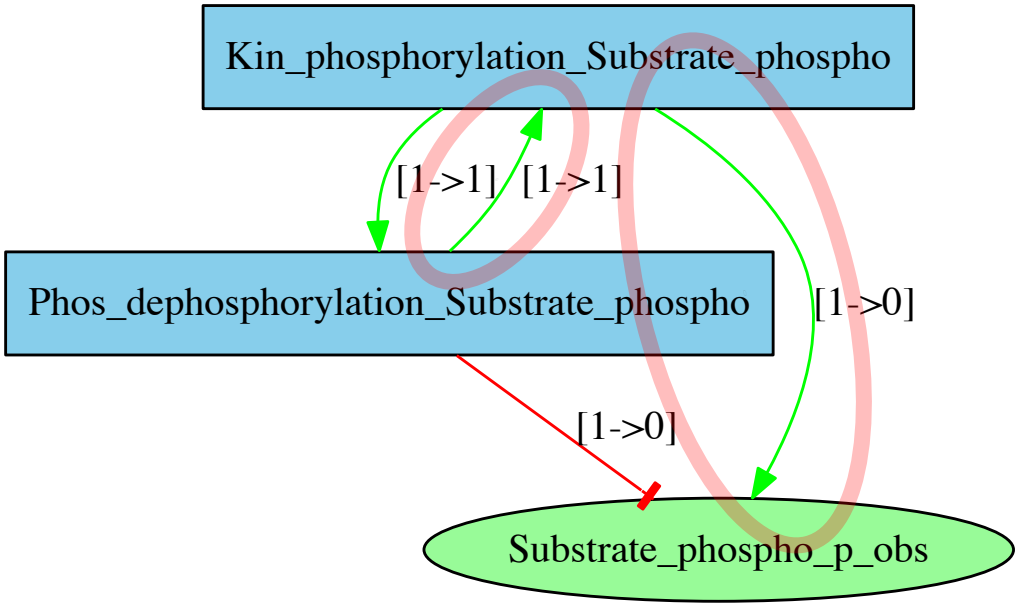


Kinase decreases phospho-Substrate (!)

The problem of “transitive triangles” in the influence map

Model:

```
Kinase phosphorylates Substrate
Phosphatase dephosphorylates Substrate
```

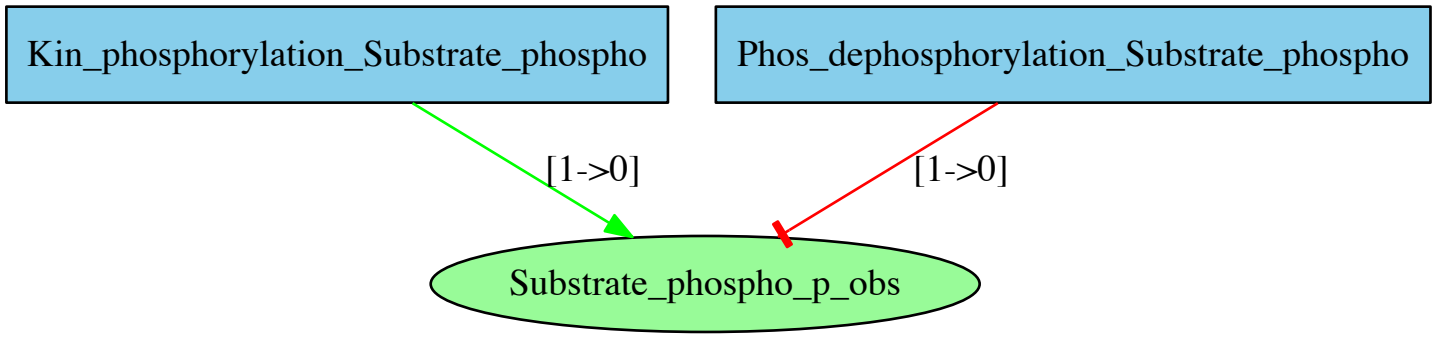


Phosphatase increases phospho-Substrate (!)

After pruning out links in “transitive triangles”

Model:

```
Kinase phosphorylates Substrate
Phosphatase dephosphorylates Substrate
```



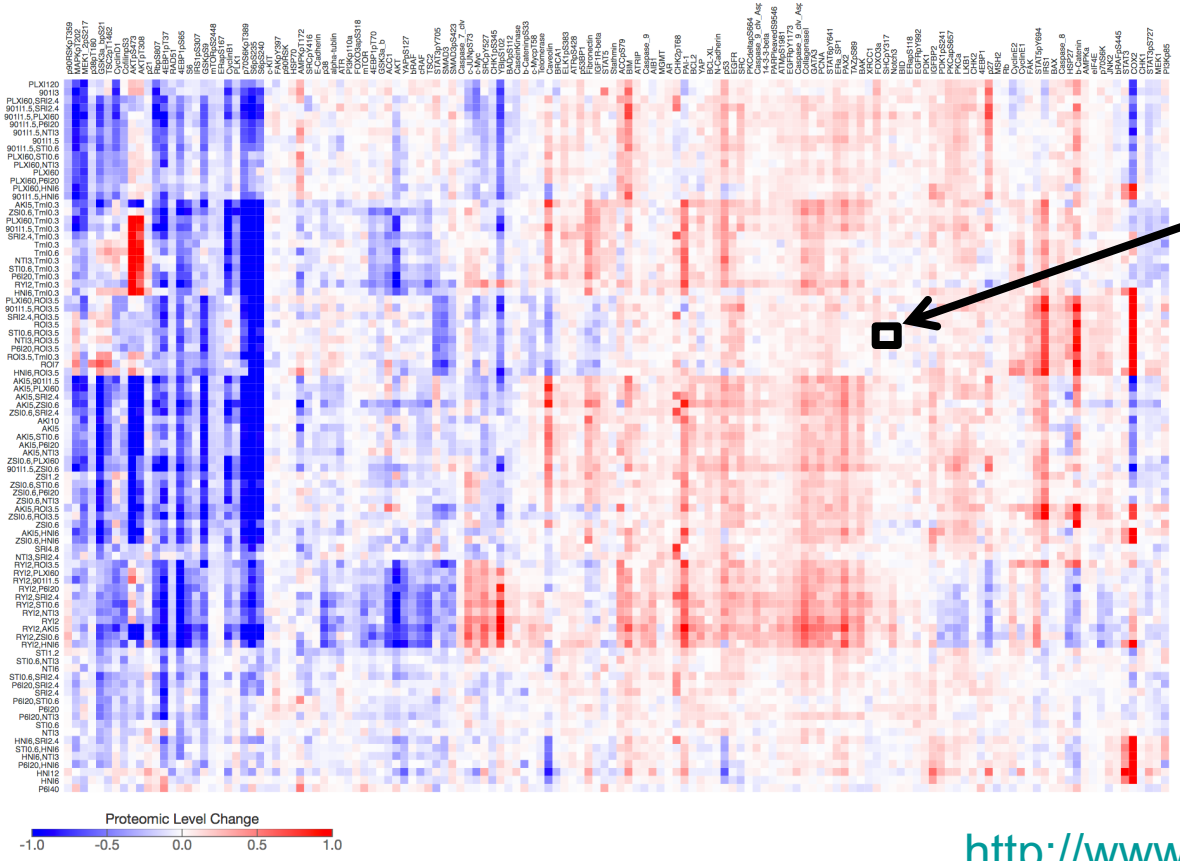
Systematically removing links between rules that share downstream targets eliminates these paths.

Use case for explanation: interpreting phosphoproteomic data

- Previously published phosphoproteomic dataset (Korkut et al.)
- Melanoma cell line treated with different drug combinations
- Protein and phospho-protein abundances measured at 24 hrs

RPPA measurements

Drug combinations



http://www.sanderlab.org/pertbio/

What we did: Model construction

- Reading
  - Read ~95,000 papers covering relevant genes with three NLP systems
  - Retrieved mechanisms from Pathway Commons and the BEL Large Corpus
- Assembly
  - Fixed grounding and sequence errors
  - Expanded statements involving protein families and complexes
  - Identified duplicates and refinements
  - Identified activations/inhibitions superseded by detailed mechanisms (Mechanism Linker)
  - Filtered out low-probability statements
  - Filtered out statements with no causal relevance to the observables of interest
  - Assembled a rule-based model (*221 proteins, 1451 rules*)

Paths obtained for largest effects (>50%)

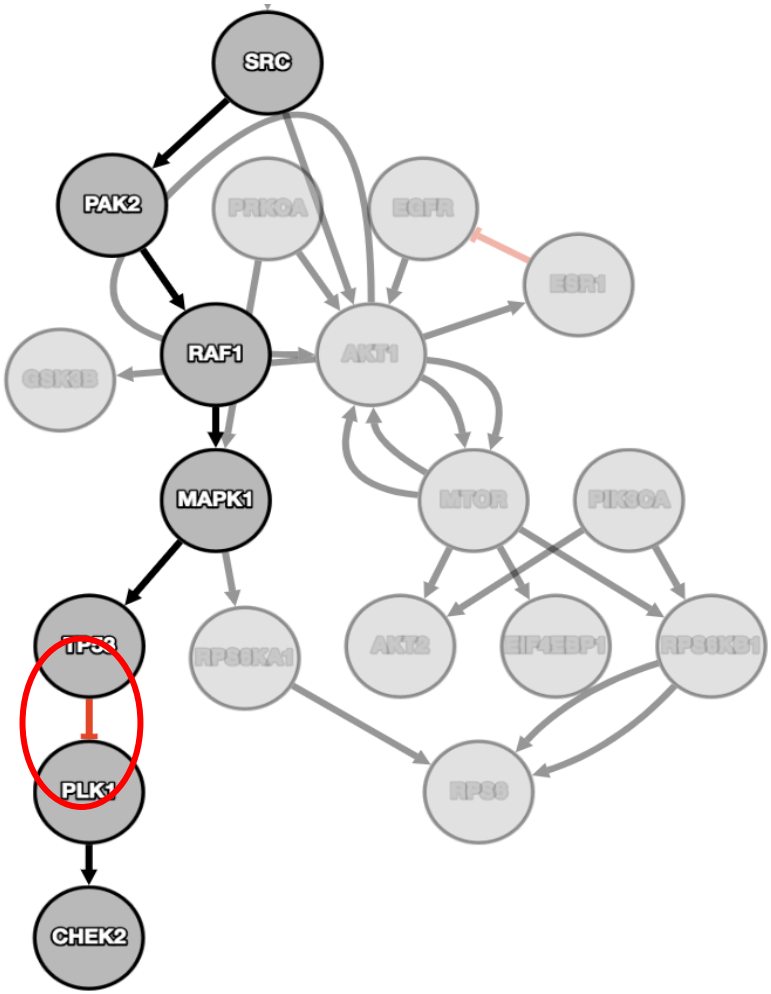
Explanations obtained for 20 out of the 22 strongest drug effects (91%)

Drug Target	Antibody	Fold-change	Path ?
MEK	MAPK pT202	0.47	
SRC	CHK2 pT68	1.75	
SRC	4EBP1 pT37	0.44	
AKT	AKT pT308	0.25	
AKT	GSK3A/B pS21	0.44	
AKT	AKT pS473	0.17	
AKT	S6 pS235	0.36	
CDK4	4EBP1 pS65	0.44	
CDK4	YBI pS102	2.13	
MTOR	AKT pT308	2.19	
MTOR	S6 pS240	0.05	
MTOR	AKT pS473	3.19	
MTOR	p70S6K pT389	0.33	
MTOR	S6 pS235	0.06	
PKC	GSK3A/B pS21	1.59	
PKC	S6 pS240	0.47	
PKC	S6 pS235	0.3	
PI3K	p70S6K pT389	0.5	
PI3K	S6 pS240	0.44	
PI3K	AKT pS473	0.2	
PI3K	S6 pS235	0.27	

An example explanation

Example explanation: *How does Src inhibition increase CHK2 pT68?*

```
SRC phosphorylated on Y418
phosphorylates PAK2 on S20. PAK2
phosphorylated on S20
phosphorylates RAF1 on S338. RAF1
phosphorylated on S338, T269 and
S471 phosphorylates MAPK1 on T185.
MAPK1 phosphorylated on T185 and
Y187 phosphorylates TP53 on S15.
TP53 phosphorylated on S20 and S15
decreases the amount of PLK1. PLK1
phosphorylates CHEK2 on T68, which
is measured by CHK2_pT68.
```



Every step in the path is auditable

Evidence for “TP53 decreases PLK1”

Pathway Commons URI:

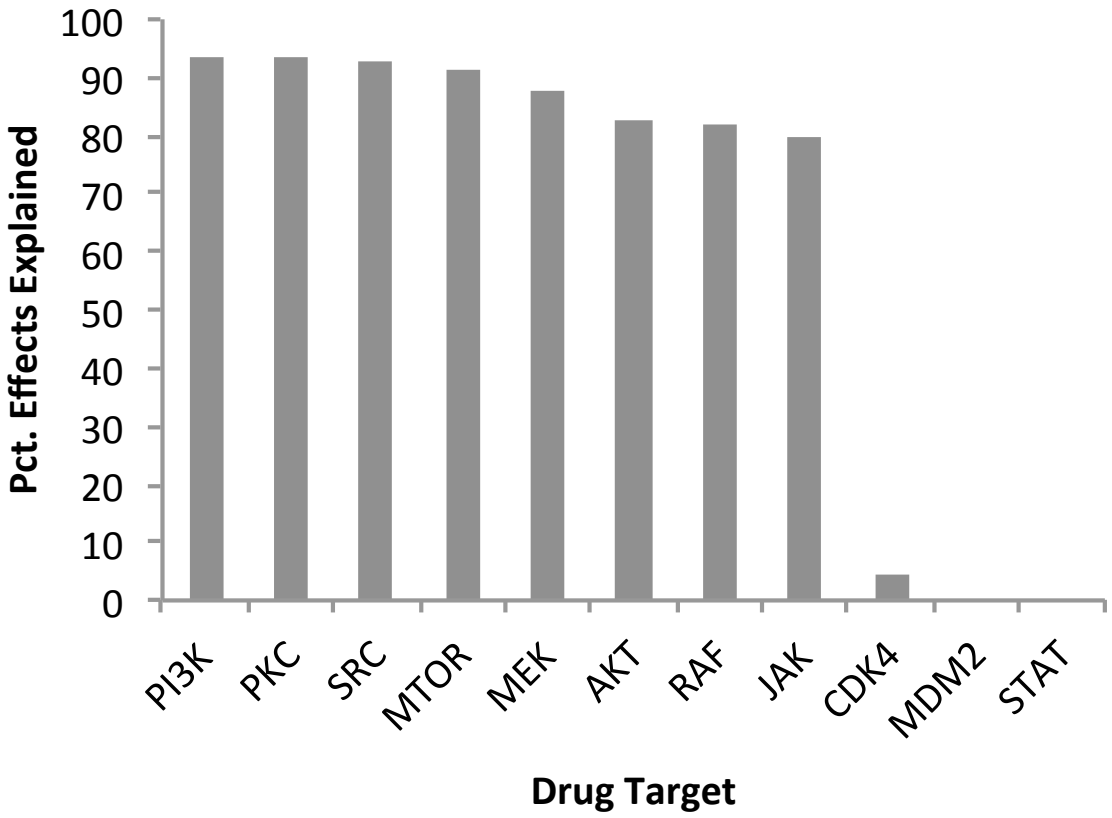
http://pathwaycommons.org/pc2/Control_6cd5f2d5cd2d3a5e33e33154560eb3e6

Text extracted by REACH:

PMID	Text
21726628	In addition, activation of **p53 has been shown to suppress the transcription of Plk1** directly or via the p21 dependent mechanism.
24407240	Our finding that p21 mediates the DNA damage induced **p53 dependent suppression of PLK1** does not exclude the possibility of direct suppression of PLK1 transcription by p53. We have previously shown in H1299 cells stably transfected with a temperature sensitive p53 mutant (tsp53) that the induction of functional **p53 decreases PLK1** protein levels in a p21 dependent manner.
26595675	Recent evidence from a knockout mouse model suggests that p21 is required for **p53 dependent repression of Plk1 expression**
22405092	Mechanistically, this is mediated by **p53 which represses PLK1** expression through chromatin remodelling. **PLK1 is down-regulated by p53** as part of the G2/M cell cycle checkpoint
24152729	Restoring **p53 by depletion of E6 also reduced the level of active Plk1** on chromatin (T210P level
24076372	Together, these data indicate that **p53 negatively regulates PLK1 expression**, while E2F1 positively regulates PLK1 expression. Consistently, over-expression of **p53 and p21 down-regulates PLK1** gene transcription in anaplastic thyroid carcinoma cells.
20962589	**Downregulation of PLK1 expression by p53** is relieved by the histone deacetylase inhibitor, trichostatin A, and involves recruitment of histone deacetylase to the vicinity of p53RE2, further supporting a transcriptional repression mechanism. Additionally, wild type, but not mutant, **p53 represses expression of the PLK1** promoter when fused upstream of a reporter gene.

Evaluation including smaller effects (> 20%)

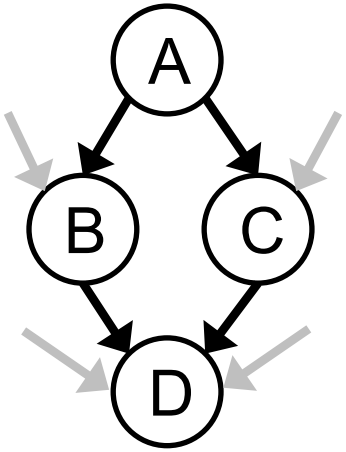
Overall performance: 95/135 paths found (70%)



Few explanations for effects of CDK4, MDM2, and STAT inhibition

Using the experimental data to rank causal paths

Model:

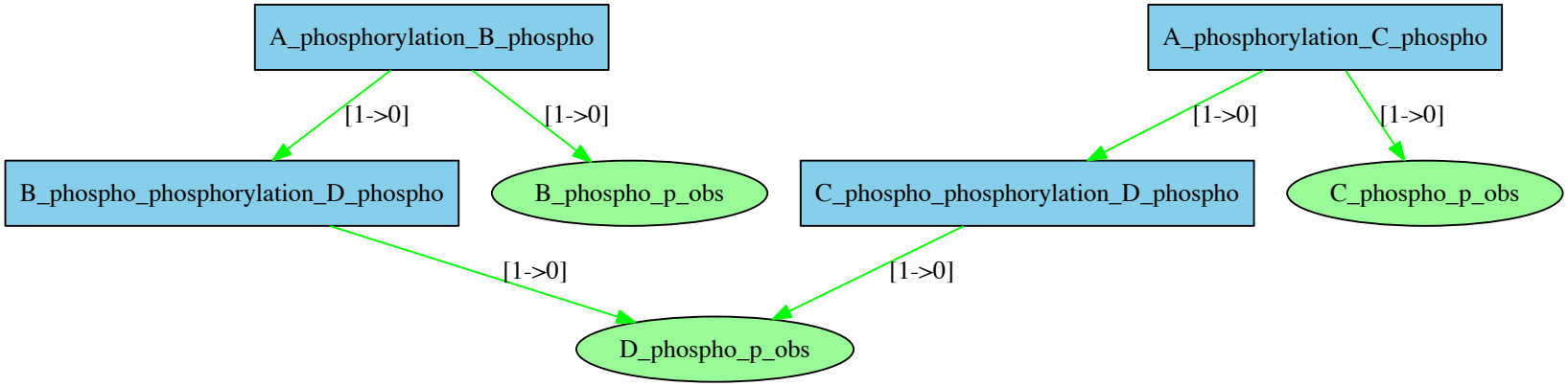


Observation: Stimulation of A increases phospho-D

Data:

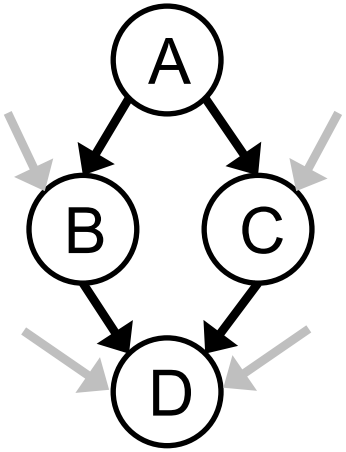
Phospho-protein	Log(Fold-change)
B	
C	
D	1

Influence map:



If B and C are unmeasured, both paths are **equally likely**

Model:

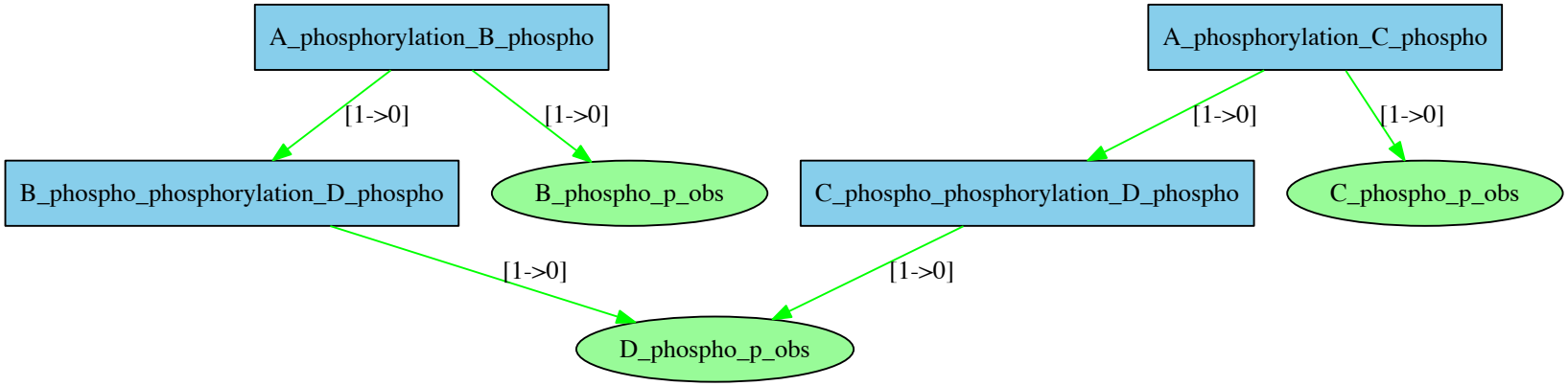


Observation: Stimulation of A increases phospho-D

Data:

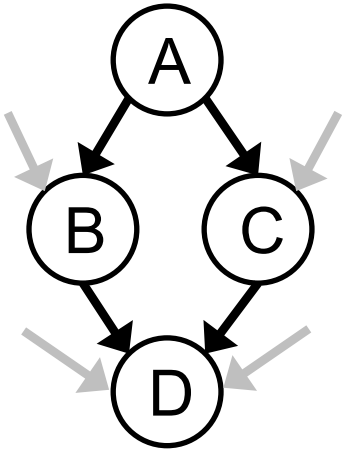
Phospho-protein	Log(Fold-change)
B	(unmeasured)
C	(unmeasured)
D	1

Influence map:



If B and C are both unchanged, both paths are **equally likely**

Model:

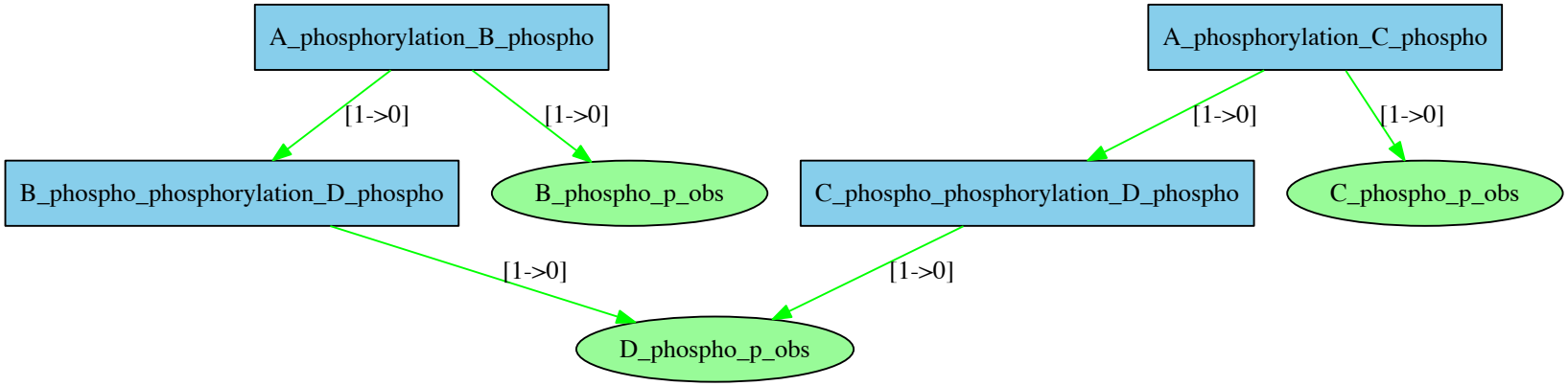


Observation: Stimulation of A increases phospho-D

Data:

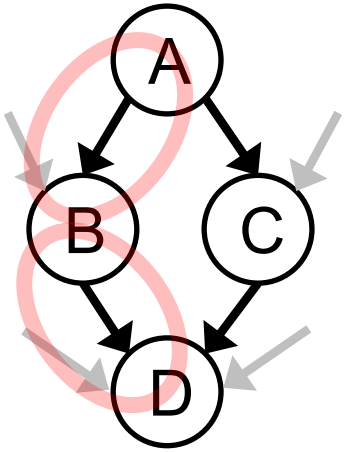
Phospho-protein	Log(Fold-change)
B	0
C	0
D	1

Influence map:



B increases but C goes down or is unchanged: A-B-D is more likely

Model:

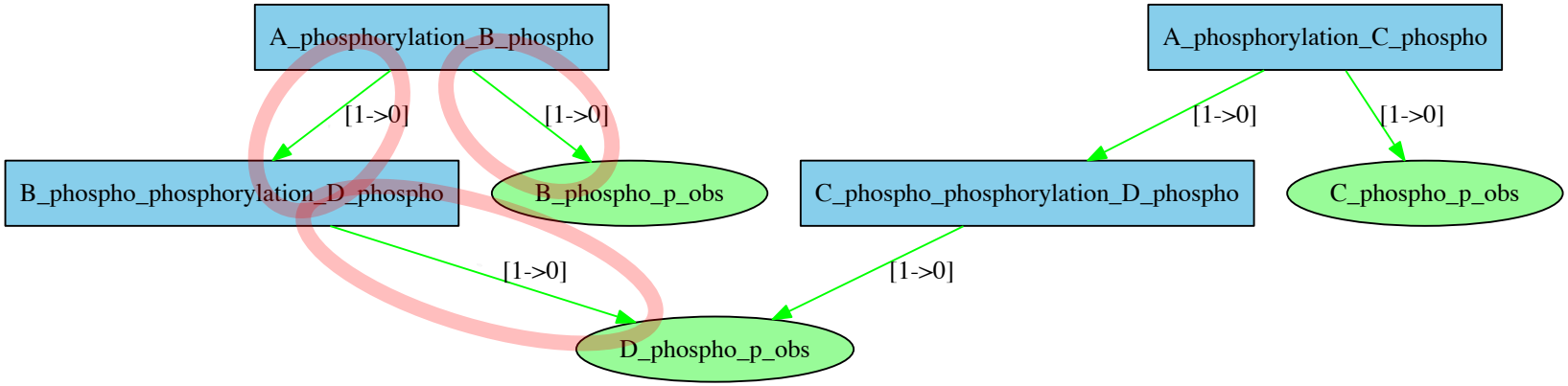


Observation: Stimulation of A increases phospho-D

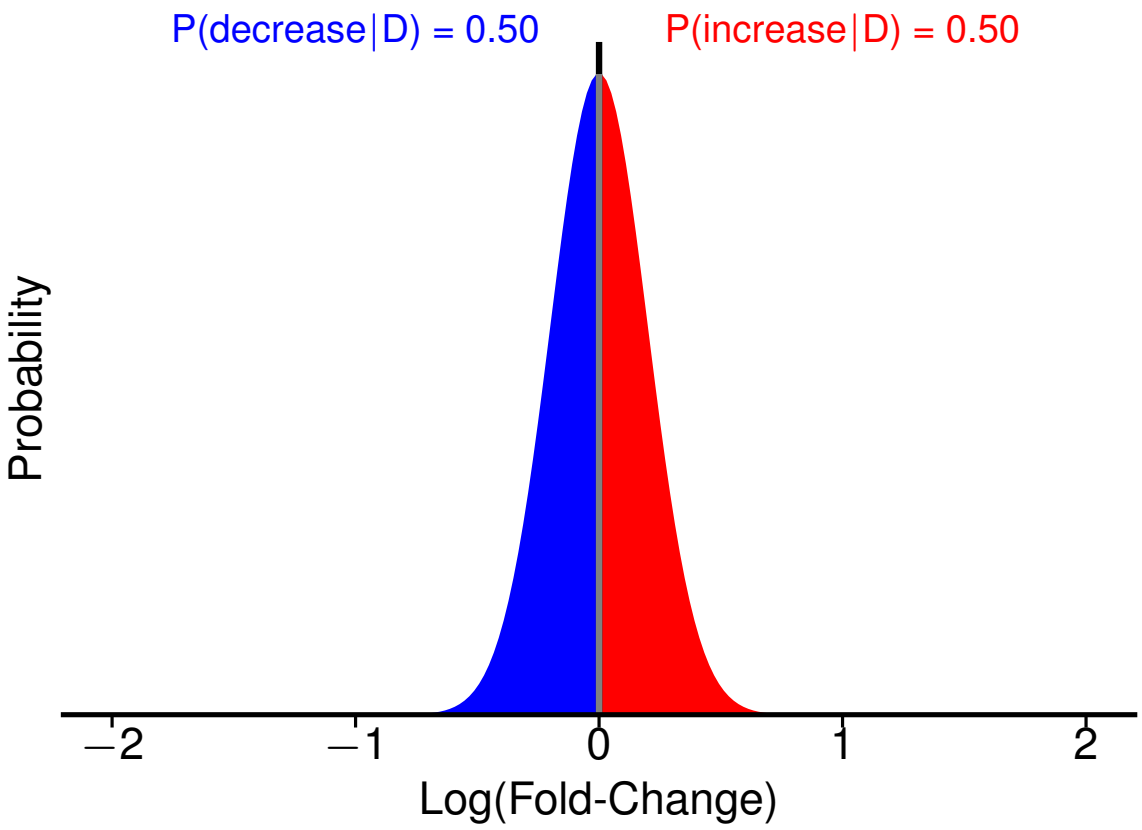
Data:

Phospho-protein	Log(Fold-change)
B	1
C	-1
D	1

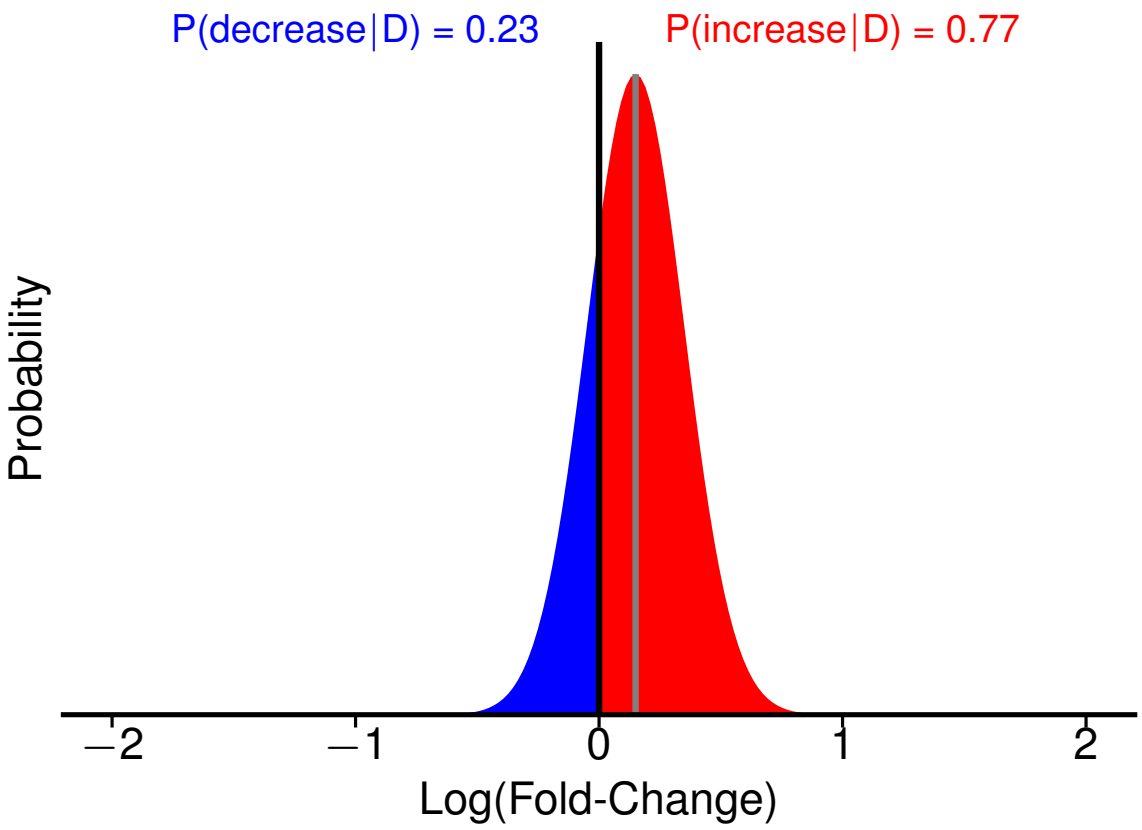
Influence map:



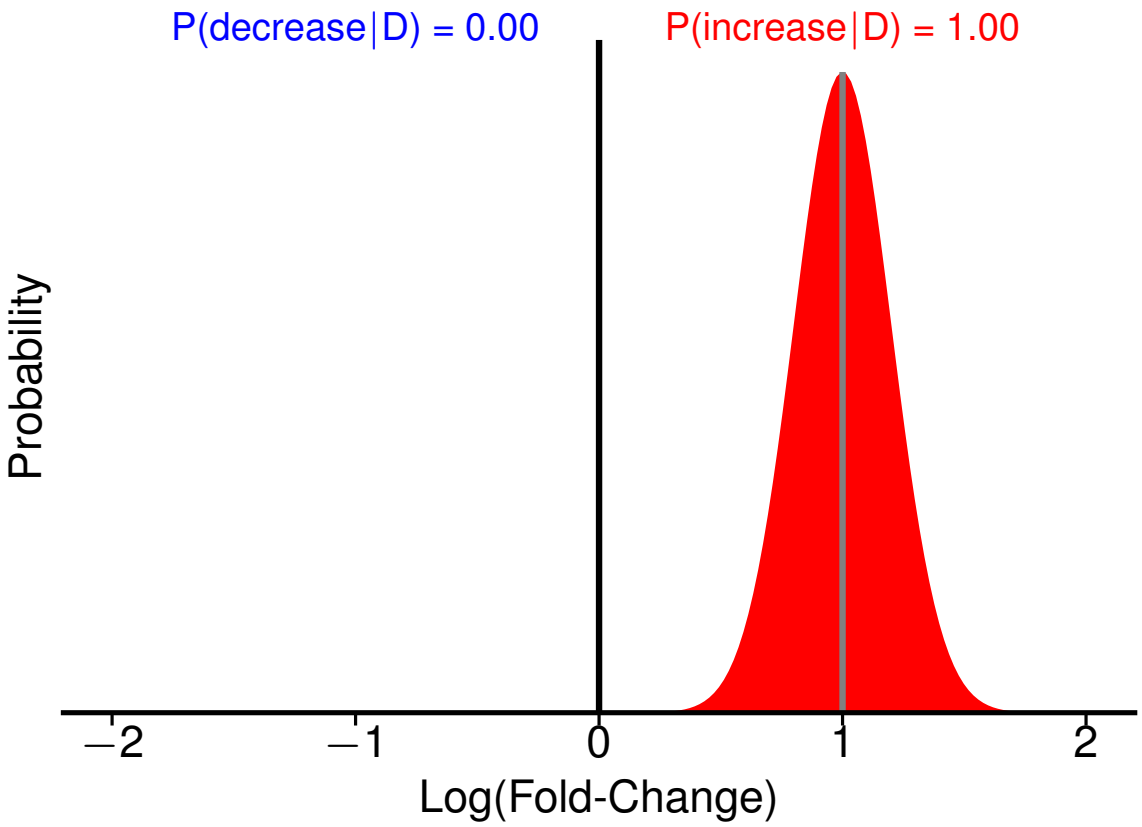
Probability model to rank likelihood of different paths



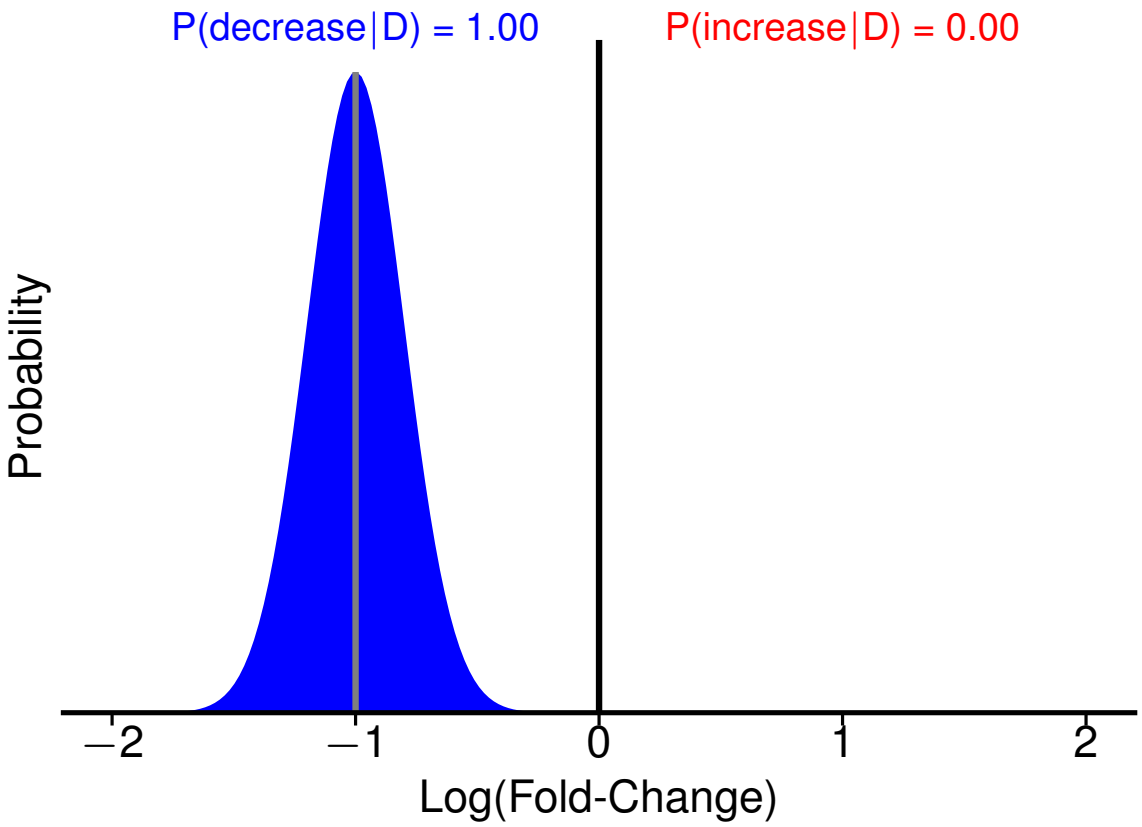
Probability model to rank likelihood of different paths



Probability model to rank likelihood of different paths



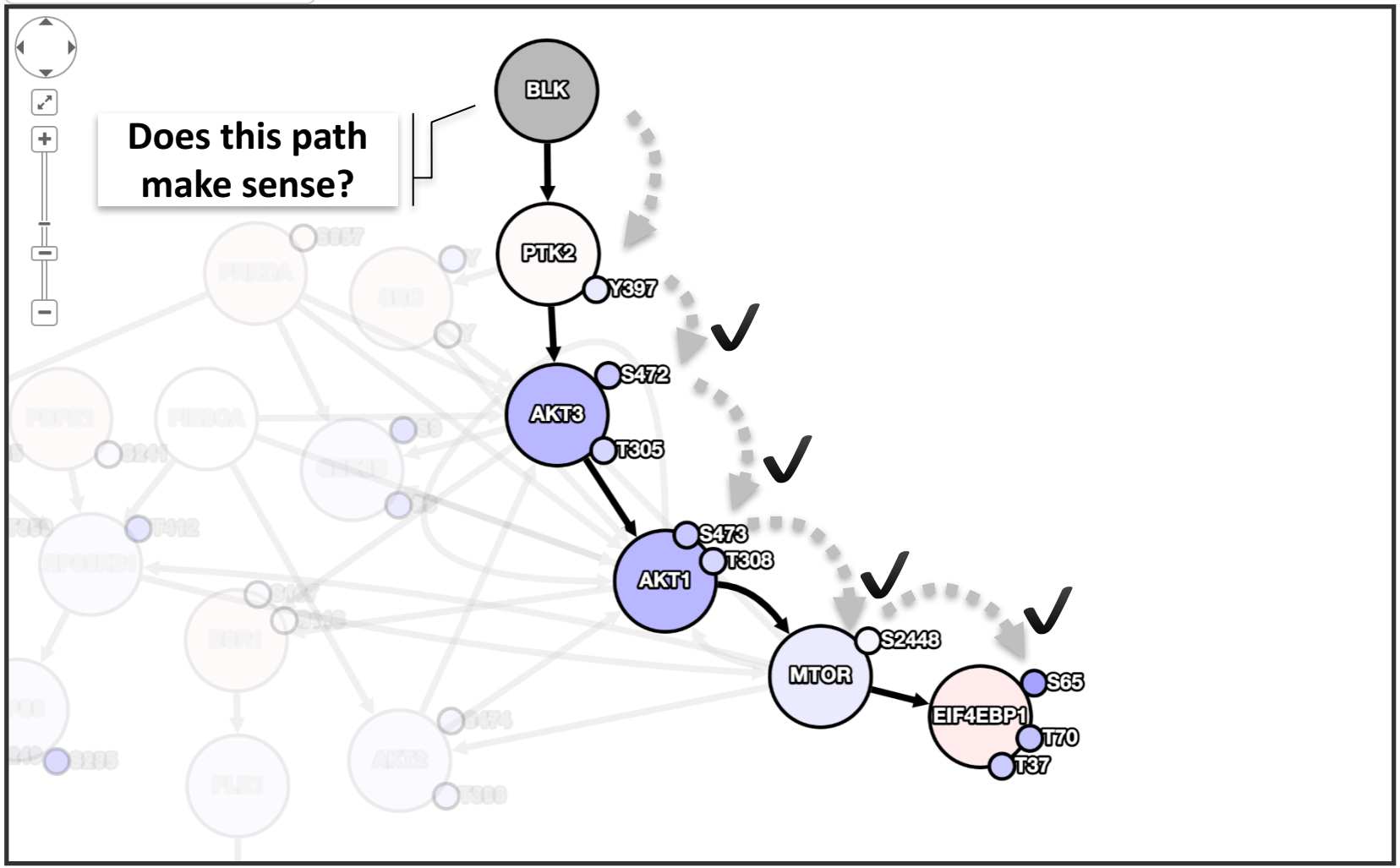
Probability model to rank likelihood of different paths



INDRA

Integrated Network and Dynamical Reasoning Assembler

☰ MODEL OPTIONS



Petar Todorov

Summary

- INDRA is a system that builds many types of mechanistic models and networks, from many sources including the literature
- Assembly involves correcting, merging, and filtering large numbers of mechanistic fragments
- Large models can be extracted from the literature and used to explain effects in large perturbation datasets
- The Kappa influence map serves as a useful tool for identifying causal paths
- Experimental data can be used to rank rule influence paths probabilistically
- New analytical methods will be needed to make best use of causal mechanistic models that are both large and detailed.

www.indra.bio

Acknowledgments

Petar Todorov
Kartik Subramanian
Jeremy Muhlich
Robert Sheehan
Lily Chylek
Isabel Latorre
Peter Sorger

Lucian Galescu
Choh-Man Teng
James Allen

Mihai Surdeanu

Jeff Rye
Rusty Bobrow
Scott Friedman
Mark Burstein

Trey Ideker
Dexter Pratt
Daniel Carlin

Funda Durupinar
Emek Demir

Anil Korkut

Paul Cohen