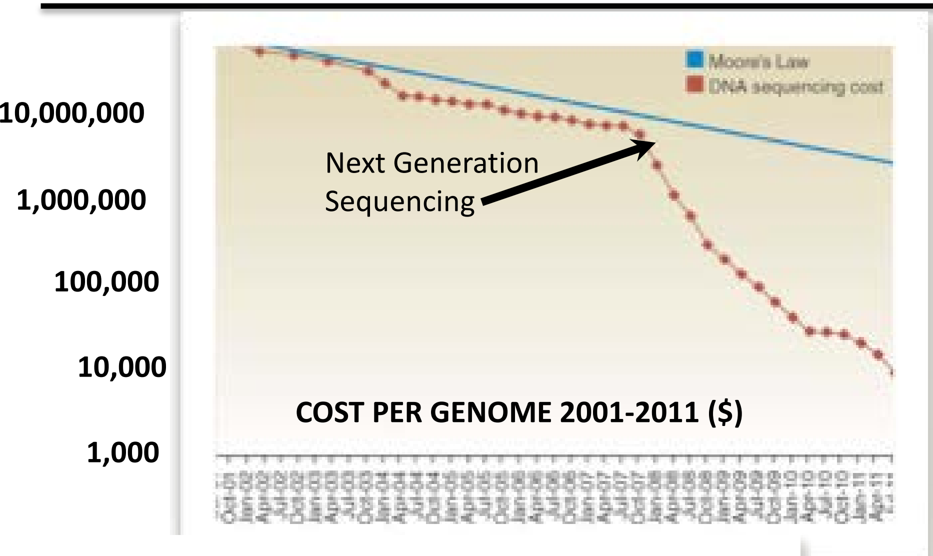
Automated Hypothesis Generation Based on Mining Scientific Literature

Scott Spangler[*,1], Angela D. Wilkins[*,3], Benjamin J. Bachman[3], Meena Nagarajan[1],
Tajhal Dayaram[3], Peter Haas[1], Sam Regenbogen[3], Curtis R. Pickering[2], Austin Comer[2],
Jeffrey N. Myers[2], Ioana Stanoi[1], Linda Kato[1], Ana Lelescu[1], Jacques J. Labrie[1],
Neha Parikh[3], Andreas Martin Lisewski[3], Lawrence Donehower[3], Ying Chen[1], Olivier Lichtarge[3]

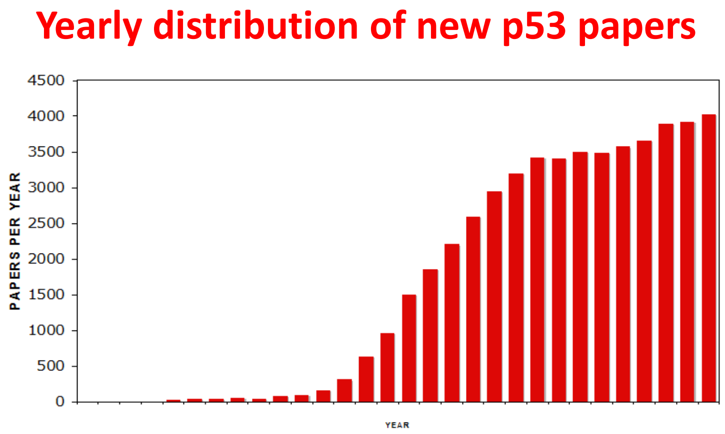
[1]IBM Research
San Jose, California
[2]The University of Texas MD Anderson Cancer Center
[3]Baylor College of Medicine
Houston, Texas

McNair Medical Institute, DARPA

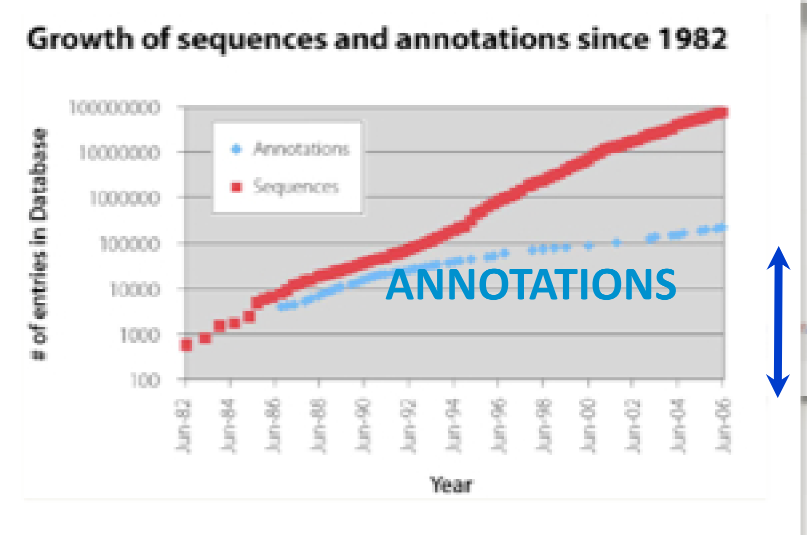
DATA OVERFLOW



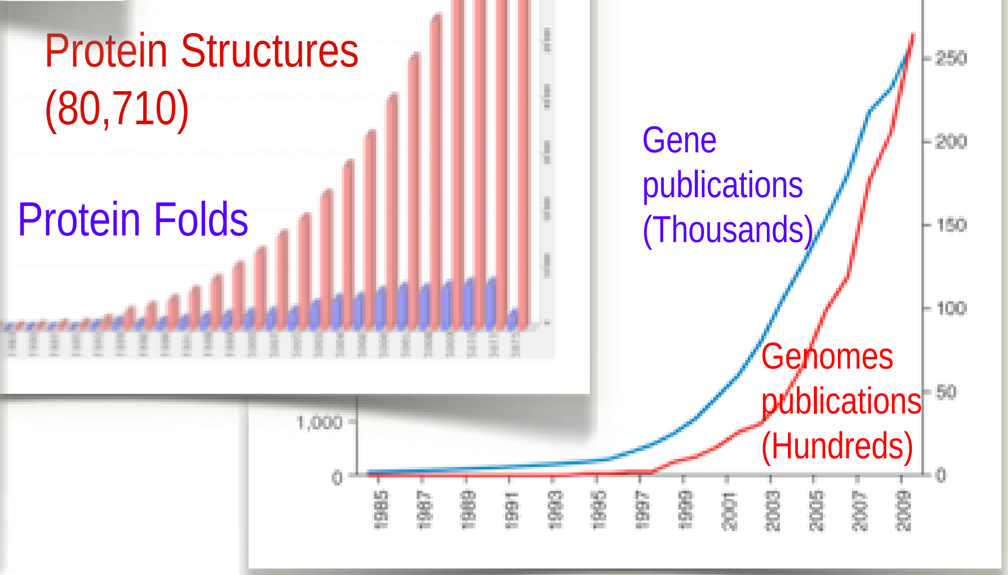
COST PER GENOME 2001-2011 ($)



Yearly distribution of new p53 papers



Growth of sequences and annotations since 1982



A mismatch between raw data and our analytic abilities

PUBLICATION OVERFLOW

Overall

- 50 million scientific papers
- 1 million more per year
- 2 new papers per minute

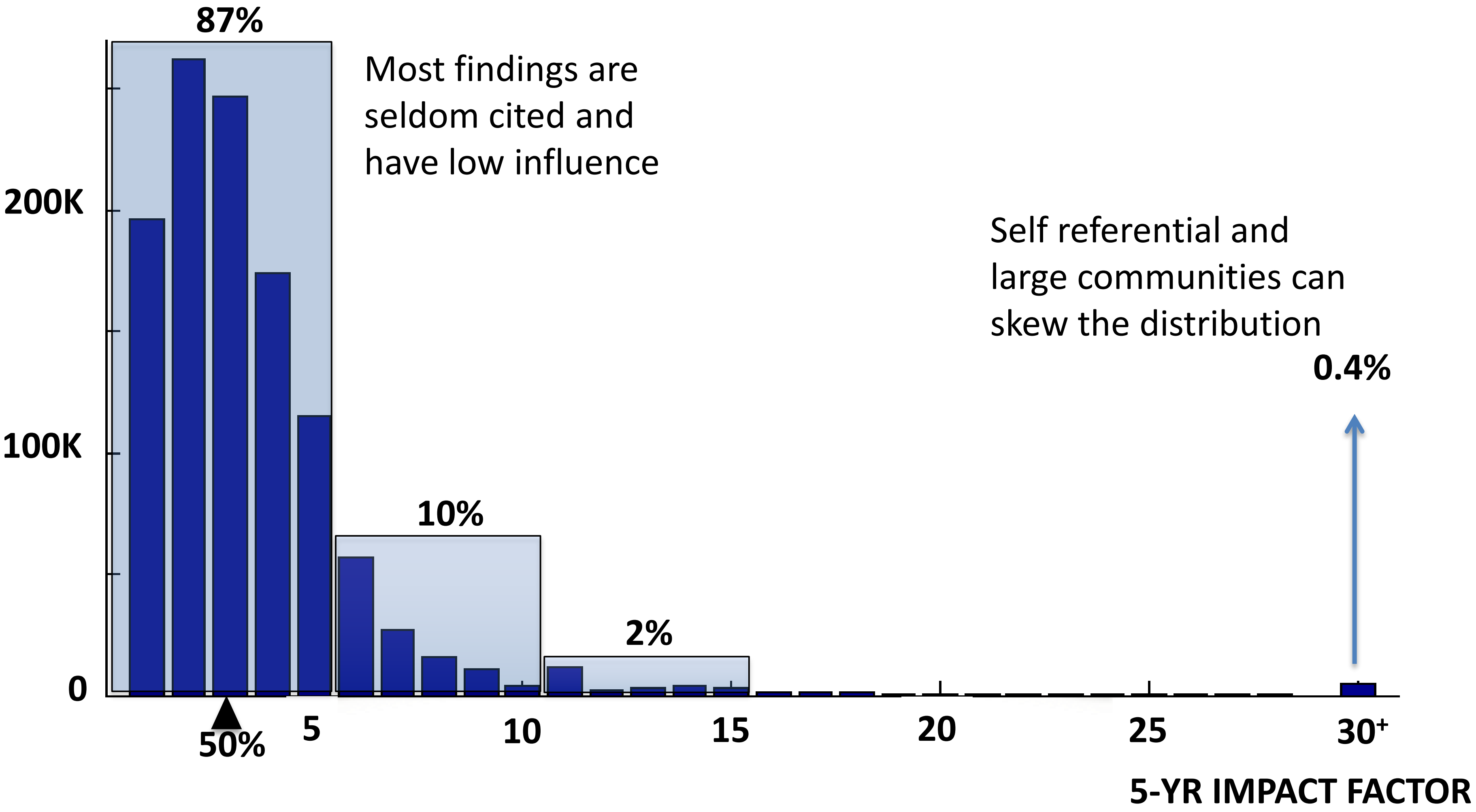
Biomedical research

- 10^3 to 10^5 papers per topic areas
- Over 70,000 papers on p53 (a tumor suppressor)

A fundamental bottleneck: we cannot keep up with discovery

BIG LITERATURE PROBLEM

5-YEAR IMPACT FACTOR OF ISI JOURNAL PUBLICATIONS IN 2012 (1.2 MILLION)

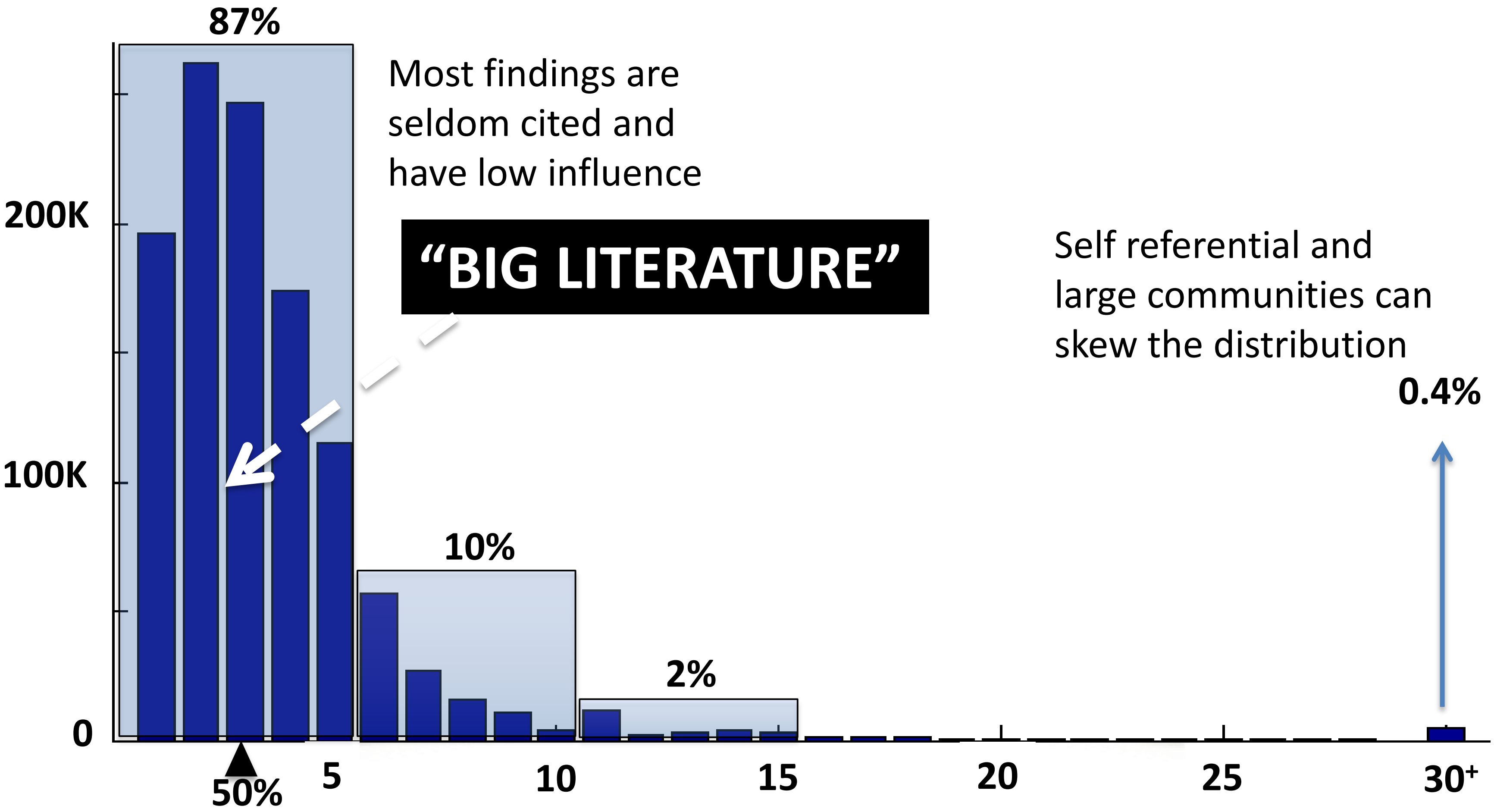


Most findings are seldom cited and have low influence

Self referential and large communities can skew the distribution

BIG LITERATURE PROBLEM

5-YEAR IMPACT FACTOR OF ISI JOURNAL PUBLICATIONS IN 2012 (1.2 MILLION)



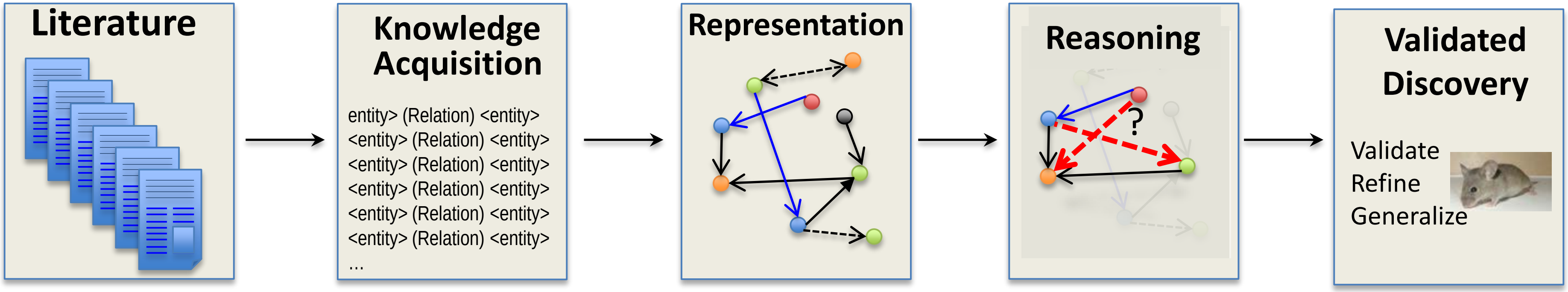
Most findings are seldom cited and have low influence

"BIG LITERATURE"

Self referential and large communities can skew the distribution

Too little time to read and to learn

A KNOWLEDGE INTEGRATION TOOLKIT: *KnIT*


Literature
Knowledge Acquisition
entity> (Relation) <entity>
<entity> (Relation) <entity>
<entity> (Relation) <entity>
<entity> (Relation) <entity>
<entity> (Relation) <entity>
<entity> (Relation) <entity>
...
Representation
Reasoning
?
Validated Discovery
Validate
Refine
Generalize

KNOWLEDGE ACQUISITION

- Survey relevant text
- Extracts relevant entities (human proteins called kinases)
- Model each entities as a points in feature space: these features are coordinates that form an aggregate “text signature” of the entity

KNOWLEDGE REPRESENTATION

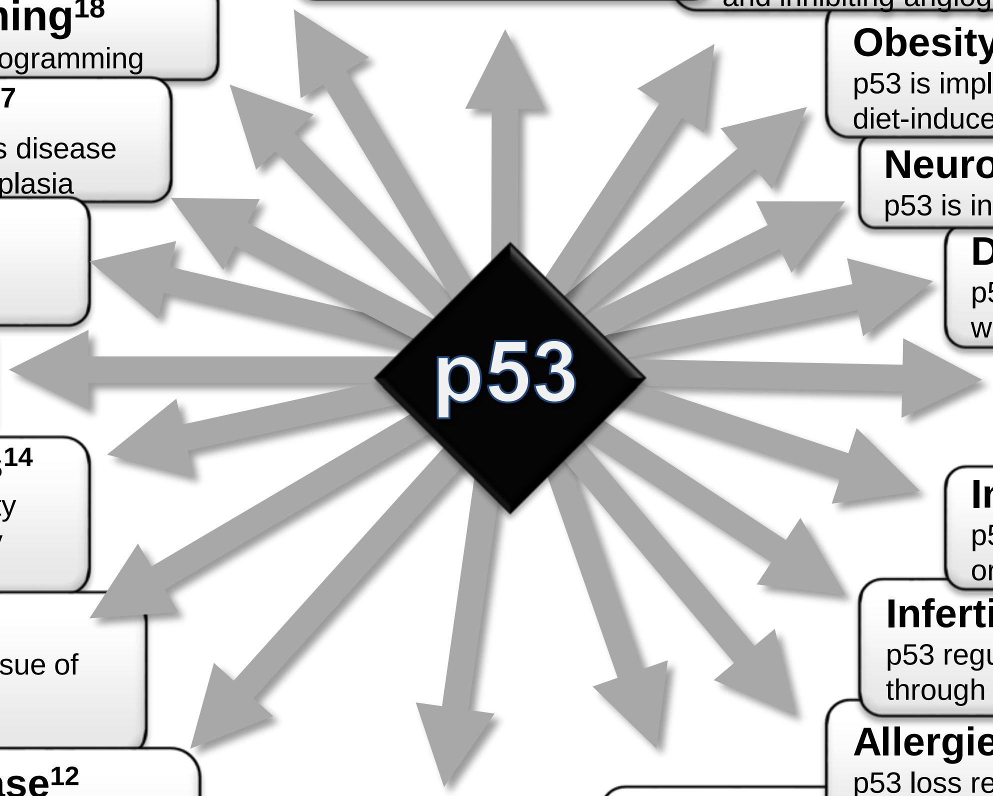
- A graph represents similarity relationship among entities.
- Helps visualize hidden literature connections between entities.
- Coloring may reveal sub-graphs of clusters of interest.
- Critically, a sub-graph may contain unexpected entities
- Hypothesis: perhaps these unexpected entities share their neighbors’ functional properties

REASONING

- Diffuses information among linked entities
- Rank order candidates for further experimentation of novel annotation predictions.
- The domain expert can evaluate the rankings, the supporting evidence, and choose which to pursue experimentally.

To accelerate scientific progress
by integrating mining, visualization, and analytics

PROOF OF PRINCIPLE
p53 THE “GUARDIAN OF THE GENOME”



Brain Cancer[1]
p53 is inactivated in more than 50% of human cancers

Heart failure[2]
P53 may exacerbate hearth failure by inducing apoptosis and inhibiting angiogenesis.

Obesity[3]
p53 is implicated in protection against diet-induced obesity

Neurodegeneration[4]
p53 is involved in **Alzheimer’s** disease

Diabetes[5]
p53 polymorphism is associated with insulin resistance

Retinal disorders[6]
p53 loss is linked to PVR, a blinding disease

Intrinsic aging[7]
p53 has a role in regulating organismal ageing

Infertility[8]
p53 regulates maternal reproduction through its interaction with LIF

Allergies/Asthma[9]
p53 loss results in enhanced allergic responses

Stroke[10]
p53-CypD complex is implicated in stroke pathology

Atherosclerosis[11]
p53 inactivation may be an important factor in the development of atherosclerosis

Parkinson’s disease[12]
Repression of p53 is associated with autosomal recessive juvenile Parkinson's disease

Lung diseases[13]
p53 is over-expressed in lung tissue of patients with chronic obstructive pulmonary disease

Autoimmune diseases[14]
p53 could reduce arthritis severity by down-regulating inflammatory responses

Kidney diseases[15]
p53 is renoprotective after ischemic kidney injury

Liver diseases[16]
Inhibition of p53 attenuates steatosis and liver injury

Digestive diseases[17]
p53 over expression in Crohn’s disease patients is associated with dysplasia

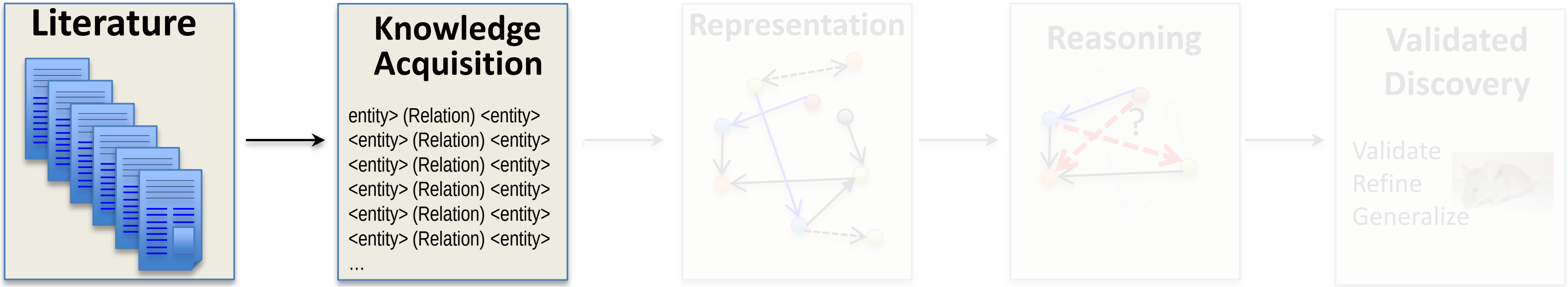
Cell reprogramming[18]
p53 antagonizes cell reprogramming

Infectious diseases[19]
P53 serves as an Influenza A virus host antiviral factor

1. PMID: 22239435
2. PMID: 21747195
3. PMID: 23412343
4. PMID: 22387179
5. PMID: 23269546
6. PMID: 22901751
7. PMID: 11780111
8. PMID: 23412905
9. PMID: 21966524
10. PMID: 22726440
11. PMID: 10086392
12. PMID: 19801972
13. PMID: 20423464
14. PMID: 18341615
15. PMID: 23222126
16. PMID: 23211317
17. PMID: 17676397
18. PMID: 19668186
19. PMID: 22105999

p53 is the “first responder” to cellular stress, as a result it is linked to a plethora of brain diseases

A KNOWLEDGE INTEGRATION TOOLKIT: *KnIT*



Literature

Knowledge Acquisition

entity> (Relation) <entity>
<entity> (Relation) <entity>
<entity> (Relation) <entity>
<entity> (Relation) <entity>
<entity> (Relation) <entity>
<entity> (Relation) <entity>
...

Representation

Reasoning

Validated Discovery

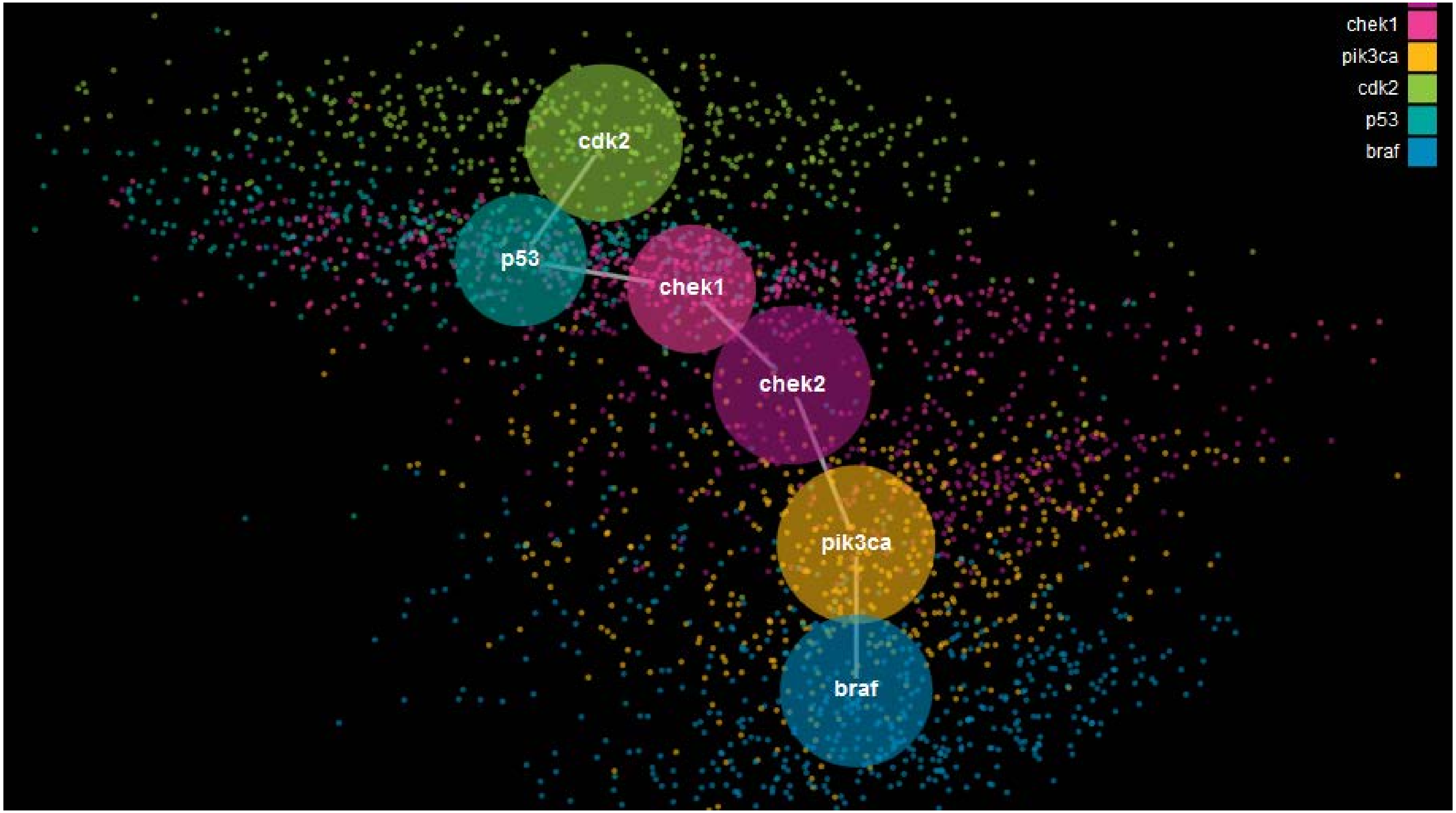
Validate
Refine
Generalize

KNOWLEDGE ACQUISITION

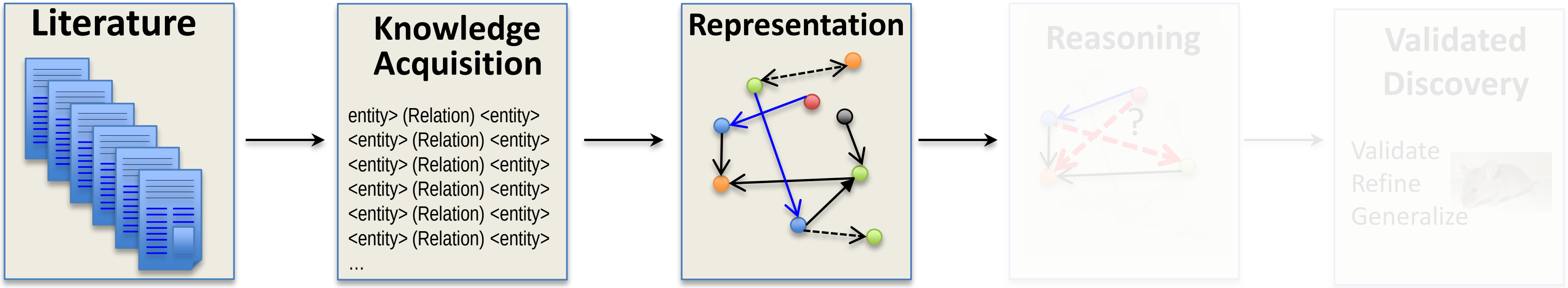
- Survey relevant text
- Extracts relevant entities (human proteins called kinases)
- Model each entities as a points in feature space: these features are coordinates that form an aggregate “text signature” of the entity

To accelerate scientific progress
by integrating mining, visualization, and analytics

Biological Entity Similarity Network



A KNOWLEDGE INTEGRATION TOOLKIT: *KnIT*



Literature

Knowledge Acquisition

entity> (Relation) <entity>
<entity> (Relation) <entity>
<entity> (Relation) <entity>
<entity> (Relation) <entity>
<entity> (Relation) <entity>
<entity> (Relation) <entity>
…

Representation

Reasoning

Validated Discovery

Validate
Refine
Generalize

KNOWLEDGE ACQUISITION

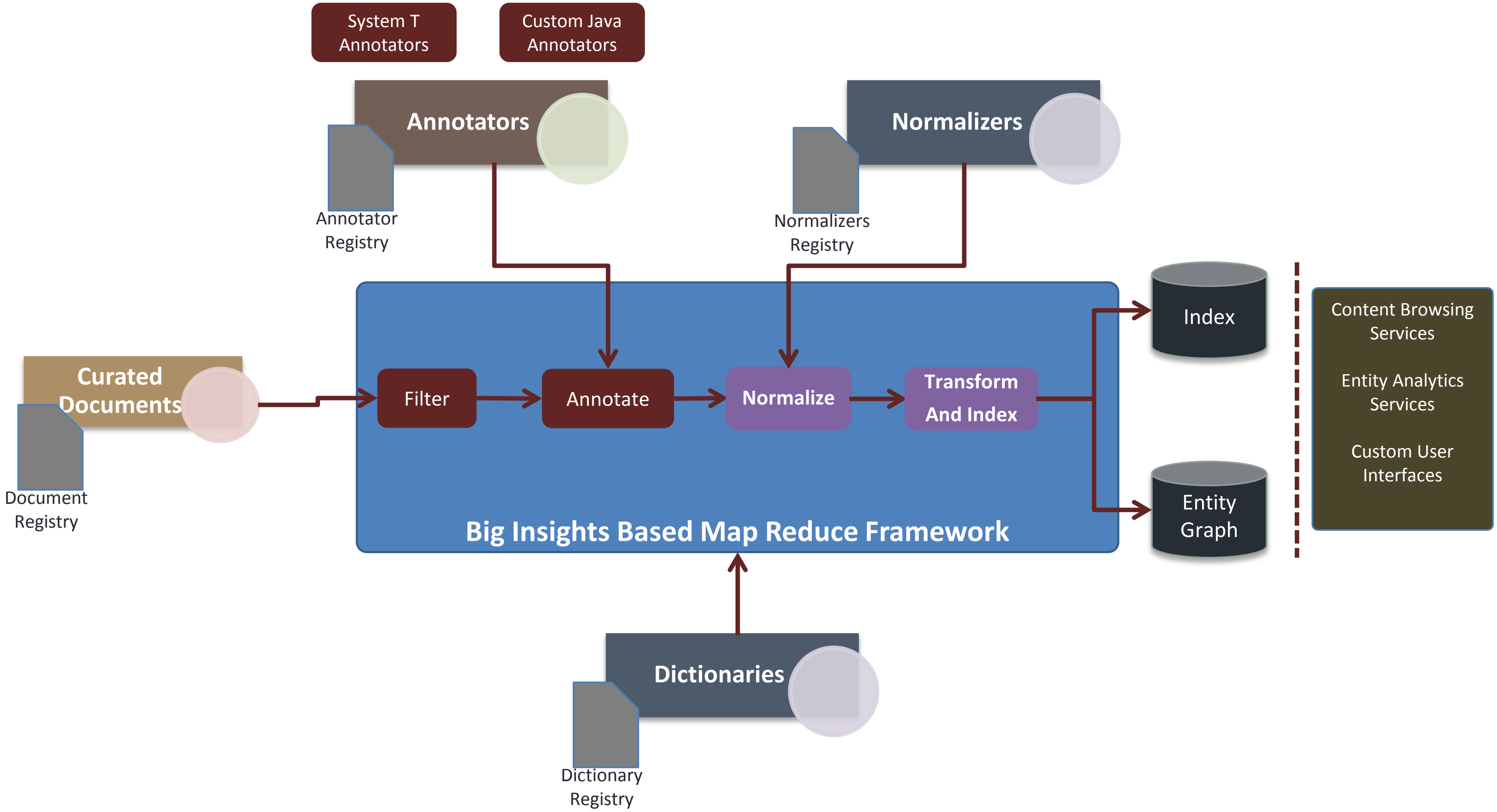
- Survey relevant text
- Extracts relevant entities (human proteins called kinases)
- Model each entities as a points in feature space: these features are coordinates that form an aggregate "text signature" of the entity

KNOWLEDGE REPRESENTATION

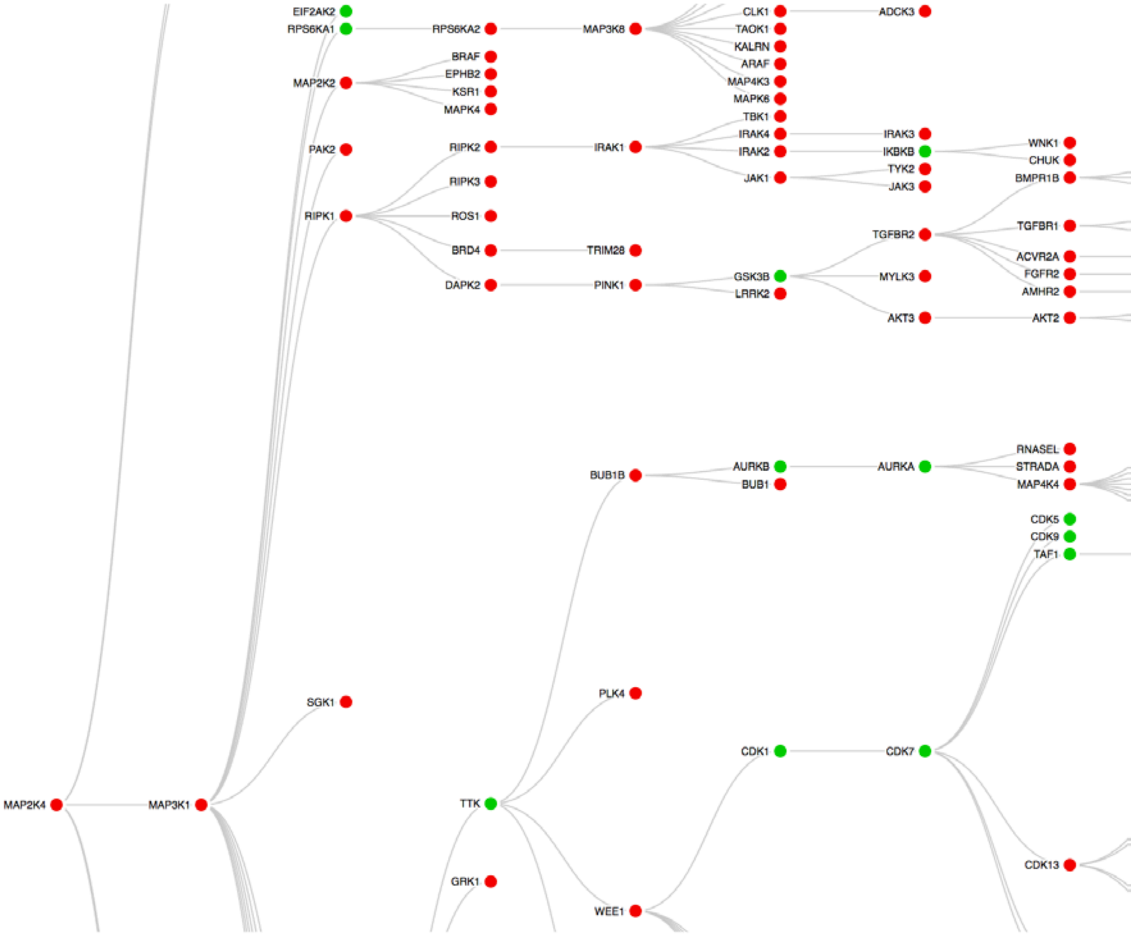
- A graph represents similarity relationship among entities.
- Helps visualize hidden literature connections between entities.
- Coloring may reveal sub-graphs of clusters of interest.
- Critically, a sub-graph may contain unexpected entities
- Hypothesis: perhaps these unexpected entities share their neighbors' functional properties

To accelerate scientific progress
by integrating mining, visualization, and analytics

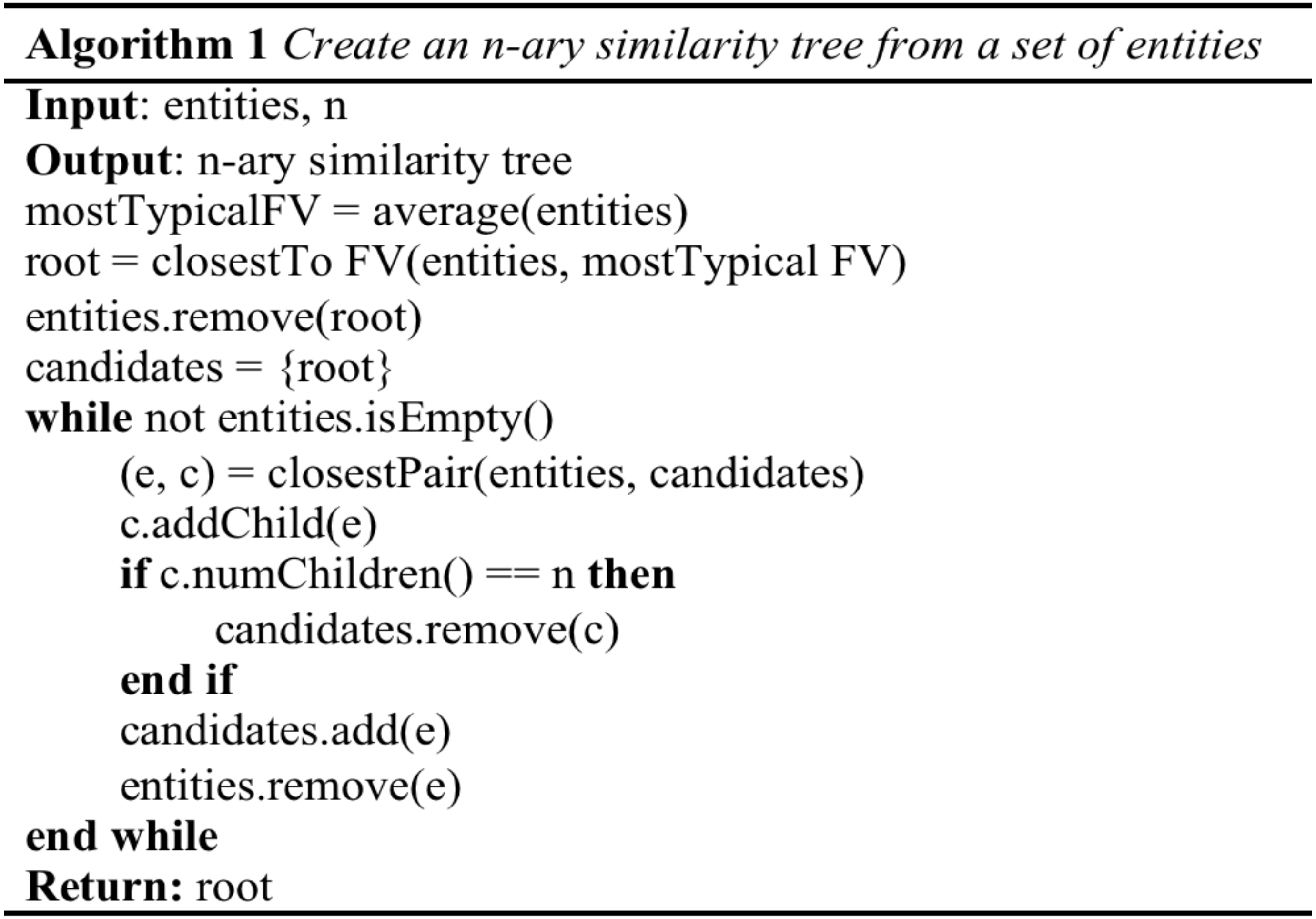
DISCOVERY END TO END WORKFLOW



Kinase Similarity Tree



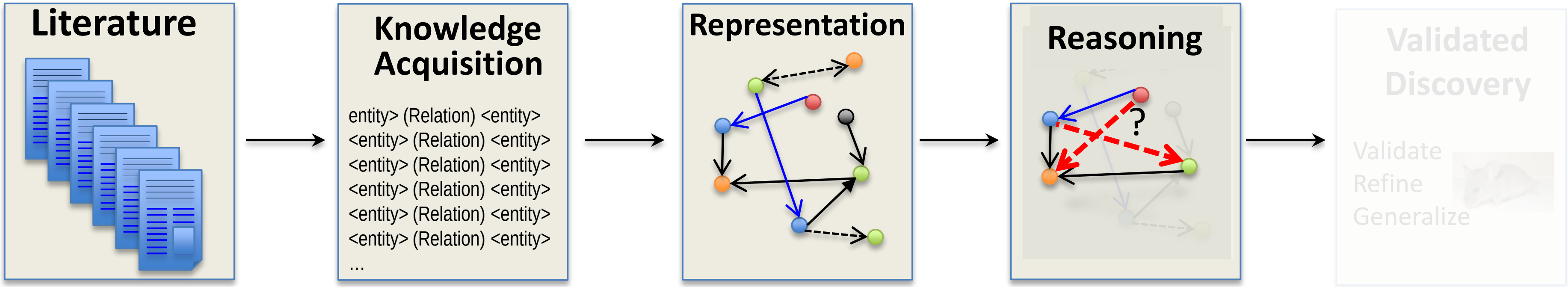
Algorithm



Algorithm 1 *Create an n-ary similarity tree from a set of entities*

```
Input: entities, n
Output: n-ary similarity tree
mostTypicalFV = average(entities)
root = closestTo FV(entities, mostTypical FV)
entities.remove(root)
candidates = {root}
while not entities.isEmpty()
    (e, c) = closestPair(entities, candidates)
    c.addChild(e)
    if c.numChildren() == n then
        candidates.remove(c)
    end if
    candidates.add(e)
    entities.remove(e)
end while
Return: root
```

A KNOWLEDGE INTEGRATION TOOLKIT: *KnIT*



Literature

Knowledge Acquisition

entity> (Relation) <entity>
<entity> (Relation) <entity>
<entity> (Relation) <entity>
<entity> (Relation) <entity>
<entity> (Relation) <entity>
<entity> (Relation) <entity>
...

Representation

Reasoning

Validated Discovery

Validate
Refine
Generalize

KNOWLEDGE ACQUISITION

- Survey relevant text
- Extracts relevant entities (human proteins called kinases)
- Model each entities as a points in feature space: these features are coordinates that form an aggregate "text signature" of the entity

KNOWLEDGE REPRESENTATION

- A graph represents similarity relationship among entities.
- Helps visualize hidden literature connections between entities.
- Coloring may reveal sub-graphs of clusters of interest.
- Critically, a sub-graph may contain unexpected entities
- Hypothesis: perhaps these unexpected entities share their neighbors' functional properties

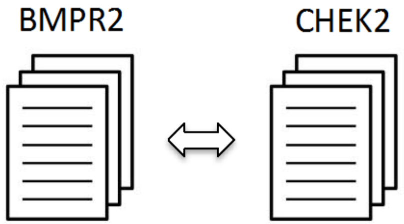
REASONING

- Diffuses information among linked entities
- Rank order candidates for further experimentation of novel annotation predictions.
- The domain expert can evaluate the rankings, the supporting evidence, and choose which to pursue experimentally.

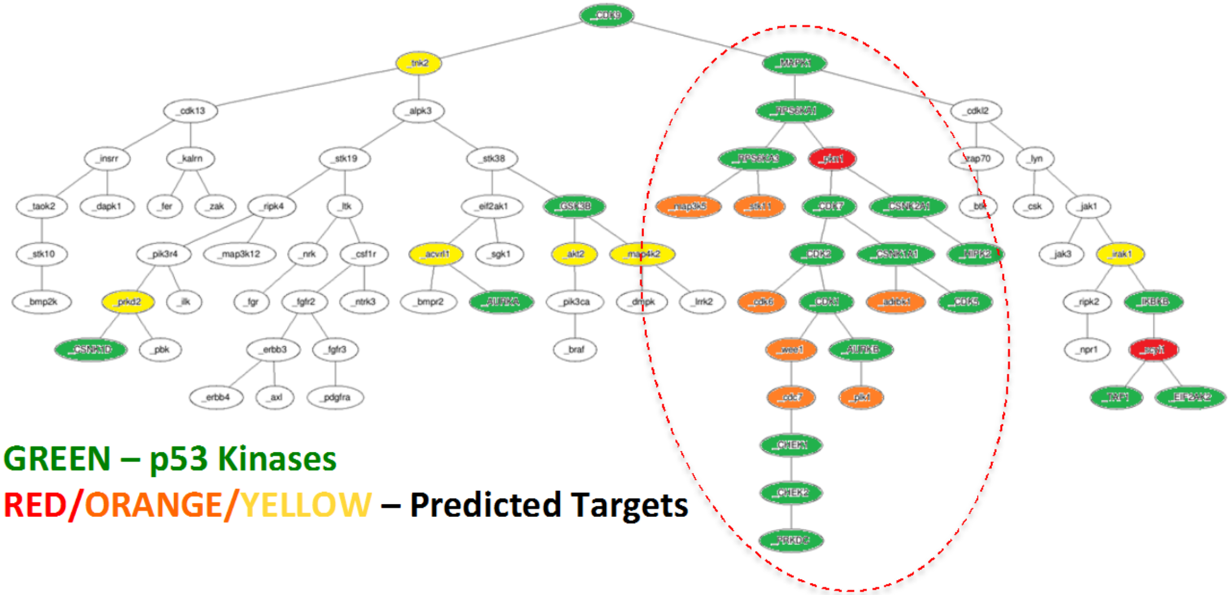
To accelerate scientific progress

by *integrating mining, visualization, and* analytics

ENTITY SIMILARITY TREE



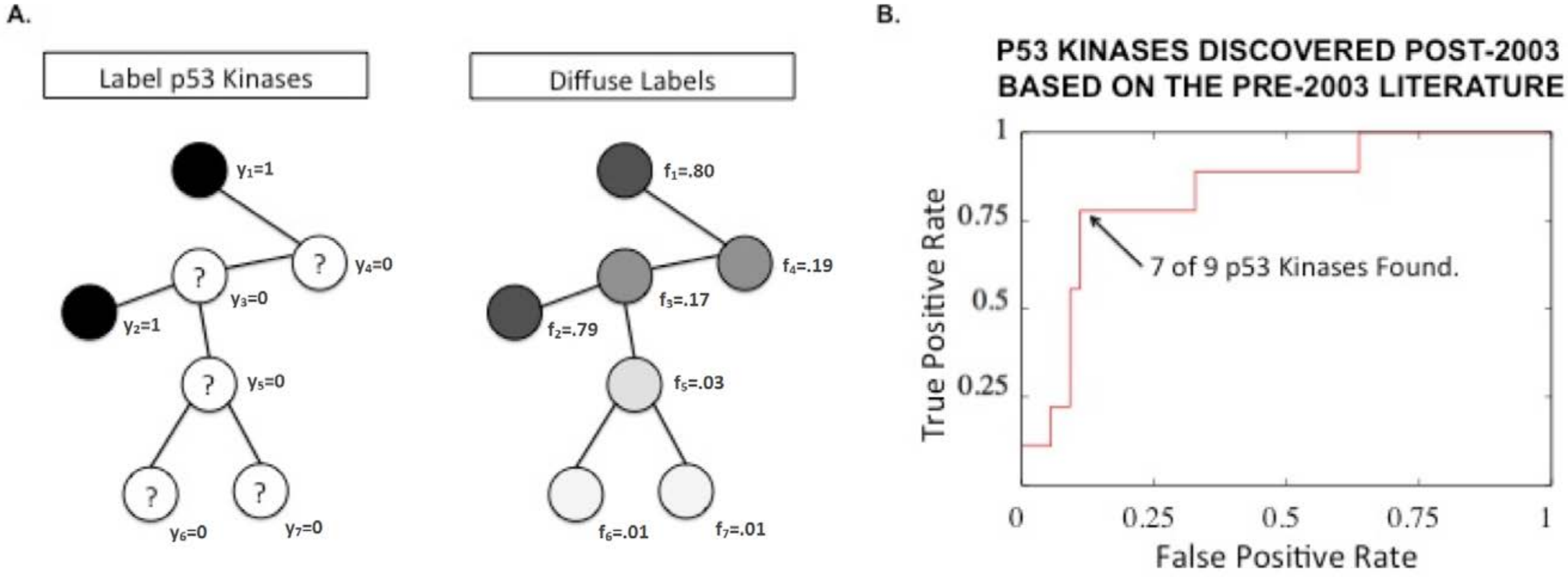
$$F = \begin{bmatrix} f_{11} & f_{12} & \cdots & \cdots & f_{1n} \\ \vdots & f_{22} & & & \vdots \\ \vdots & & \ddots & & \vdots \\ \vdots & & & \ddots & \vdots \\ f_{n1} & f_{n2} & \cdots & \cdots & f_{nn} \end{bmatrix}$$



GREEN – p53 Kinases

RED/ORANGE/YELLOW – Predicted Targets

RETROSPECTIVE CONTROL

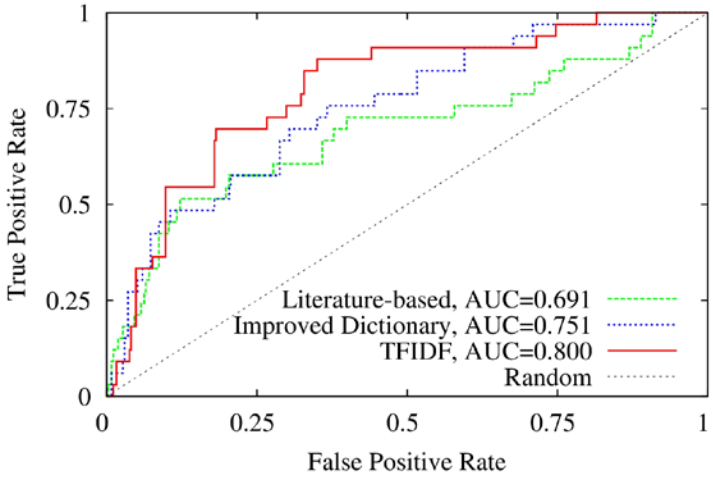


GRAPH INFORMATION DIFFUSION OF P53 KINASE LABELS KNOWN PRIOR TO 2003 RECOVERS P53 KINASES DISCOVERED AFTER 2003

Example of this type of diffusion: Lisewski et al ***CELL*** (2014)

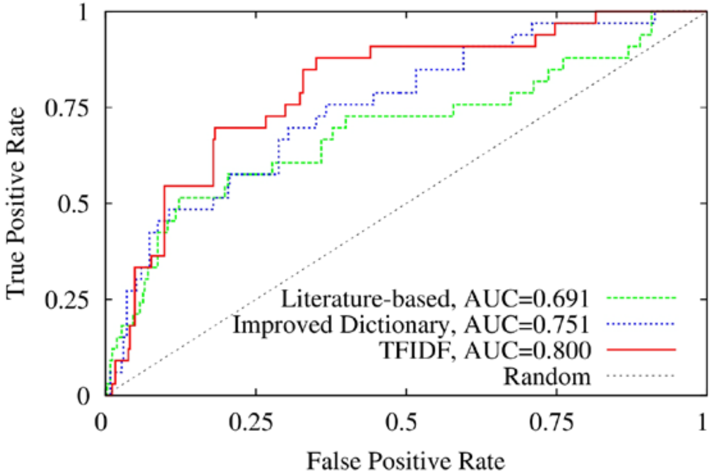
LEAVE-ONE-OUT EXPERIMENTS ALSO SUGGEST THAT THIS APPROACH CAN PREDICT WHICH KINASES TARGET A GIVEN PROTEIN

RECOVERY OF KNOW P53 KINASES

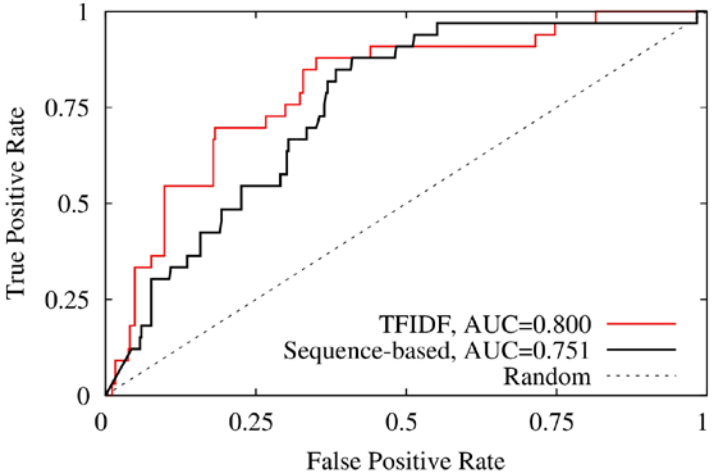


LEAVE-ONE-OUT EXPERIMENTS ALSO SUGGEST THAT THIS APPROACH CAN PREDICT WHICH KINASES TARGET A GIVEN PROTEIN

RECOVERY OF KNOW P53 KINASES

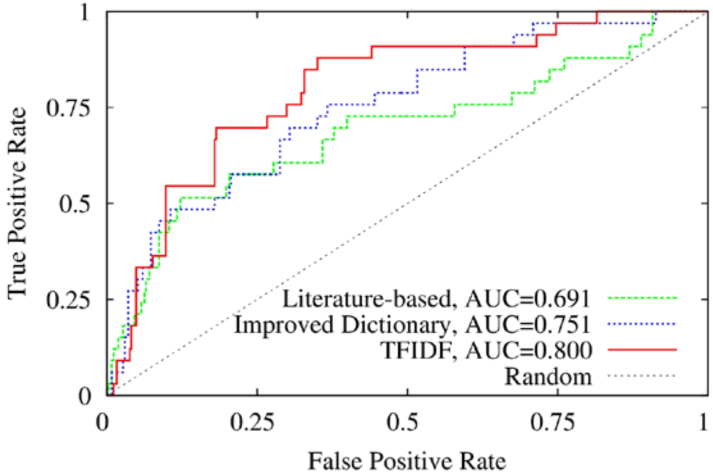


COMPARISON TO SEQUENCE ANALYSIS

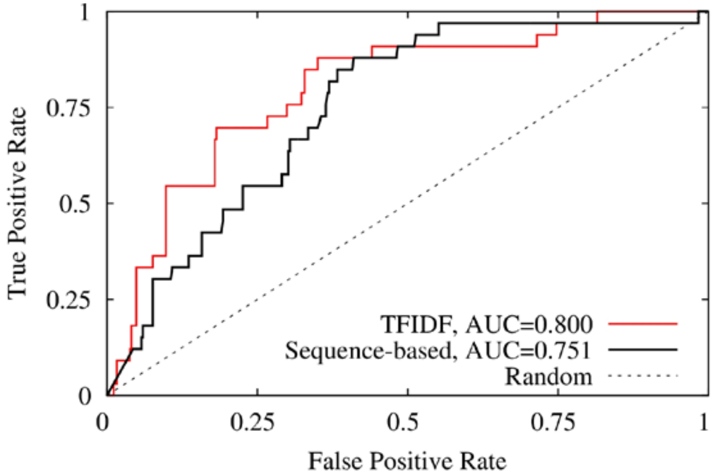


LEAVE-ONE-OUT EXPERIMENTS ALSO SUGGEST THAT THIS APPROACH CAN PREDICT WHICH KINASES TARGET A GIVEN PROTEIN

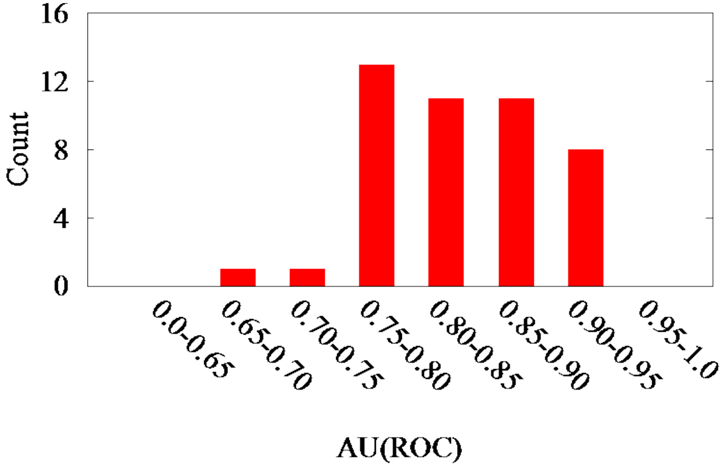
RECOVERY OF KNOW P53 KINASES



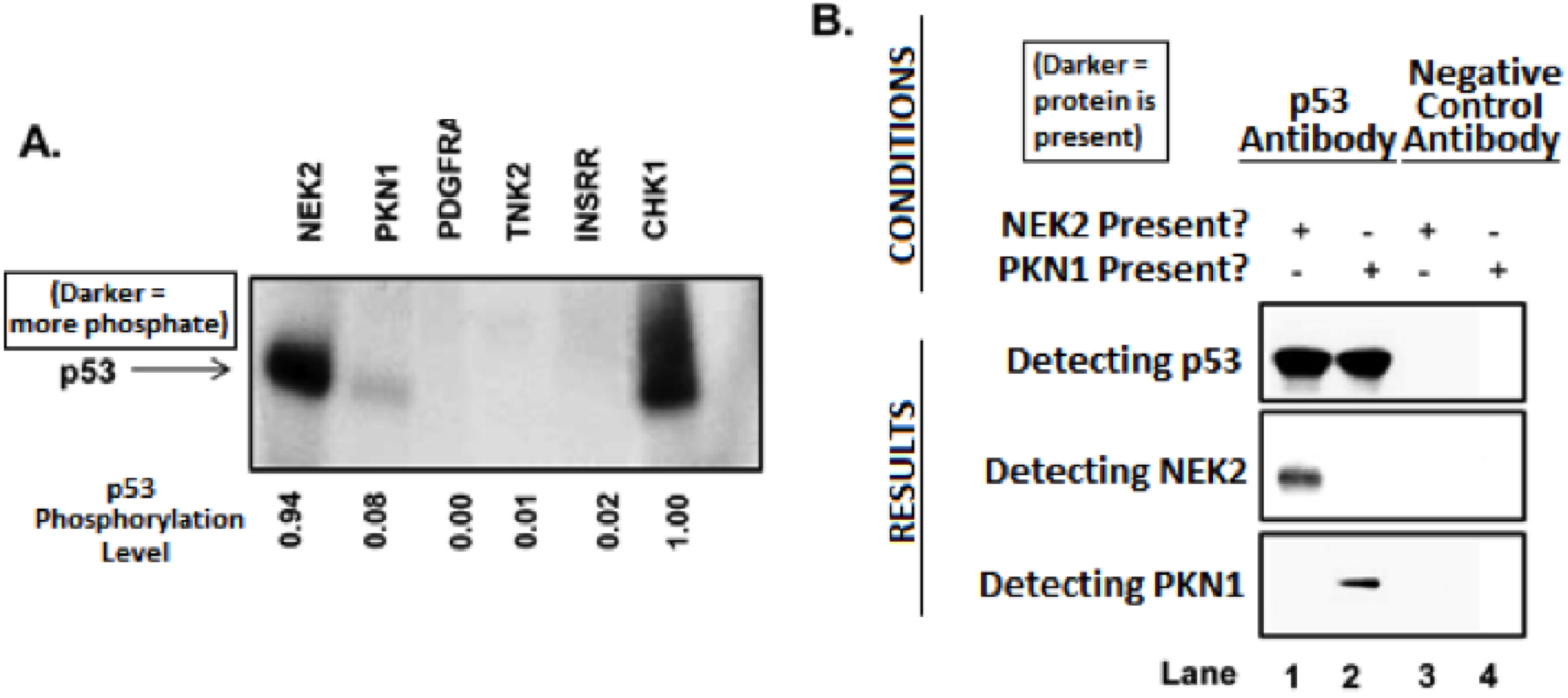
COMPARISON TO SEQUENCE ANALYSIS



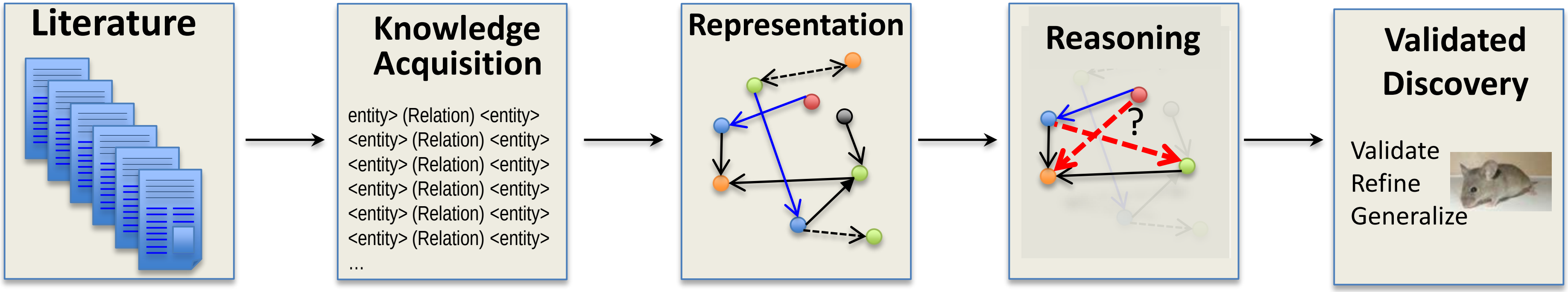
KINASE PREDICTIONS IN 45 OTHER PROTEINS



BONA FIDE EXPERIMENTAL VALIDATION



A KNOWLEDGE INTEGRATION TOOLKIT: *KnIT*



entity> (Relation) <entity>
<entity> (Relation) <entity>
<entity> (Relation) <entity>
<entity> (Relation) <entity>
<entity> (Relation) <entity>
<entity> (Relation) <entity>
…

Validate
Refine
Generalize

- Laboratory support for p53 kinases predicted from text mining
- Proof of principle for a strategy to predict some unknown fact from the scientific literature
- A first step to predict new connections based on everything else that is known.
- Future: more work needed to
  - Broaden the scope of proteins and functions
  - Comprehensive networks of interactions
  - To gather a more complete understanding of the mechanisms behind disease
  - Translate this into clinical impact.
  - Test this approach of mining literature to identify hidden relationships beyond cancer and beyond biology to other areas of human though where text is a bottleneck.
- Such acceleration of discovery is not only desirable, but also indispensable for human flourishing.