USC Viterbi
Information Sciences Institute

Towards Evidence Extraction: Analysis of Scientific Figures from Studies of Molecular Interactions

Gully Burns[1], Xiangyang Shi[1], Yue Wu[1], Huaigu Cao[1], Prem Natarajan[1]

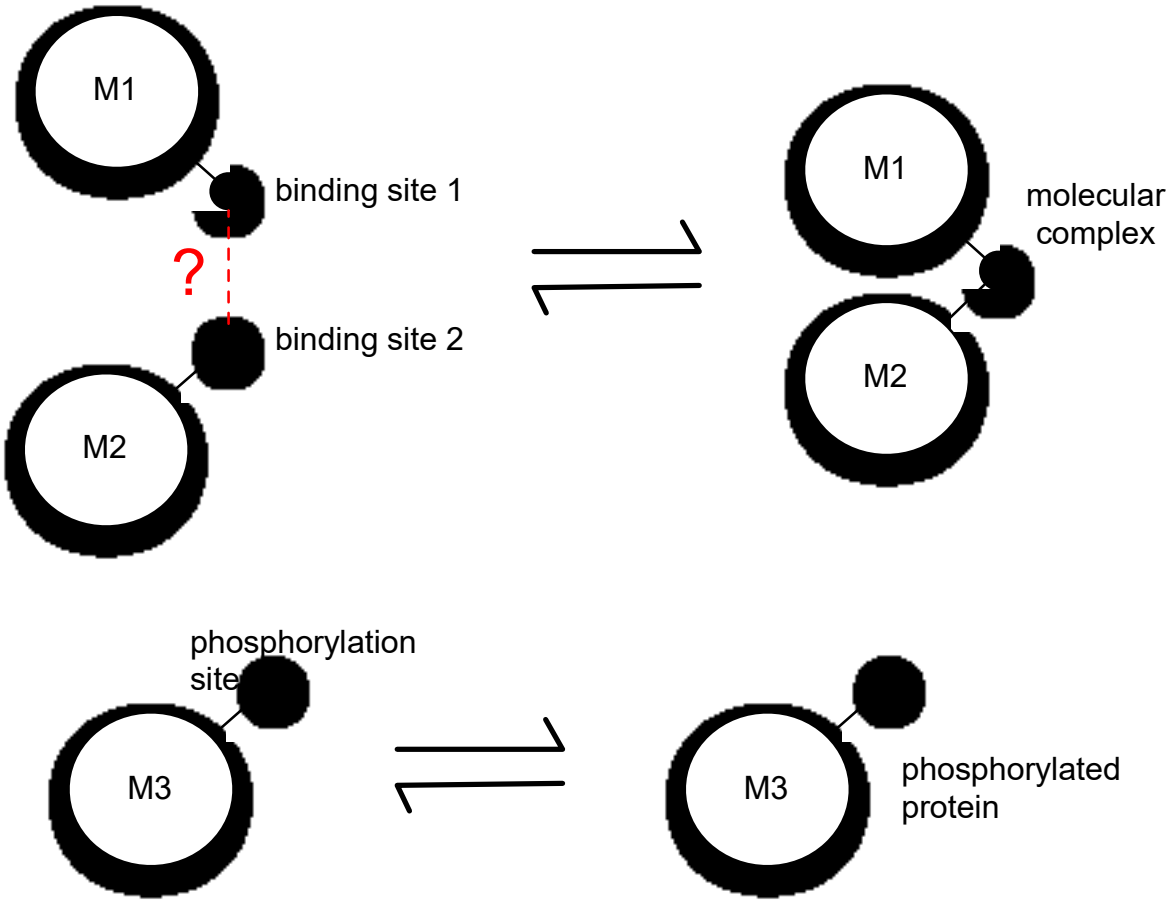
[1] Intelligent Systems Division, USC's Information Sciences Institute

https://sciknowengine.github.io/

Molecular Biology is both complex and beautiful

Inner Life of the Cell (https://youtu.be/wJyUtbn0O5Y)

Molecular Interactions

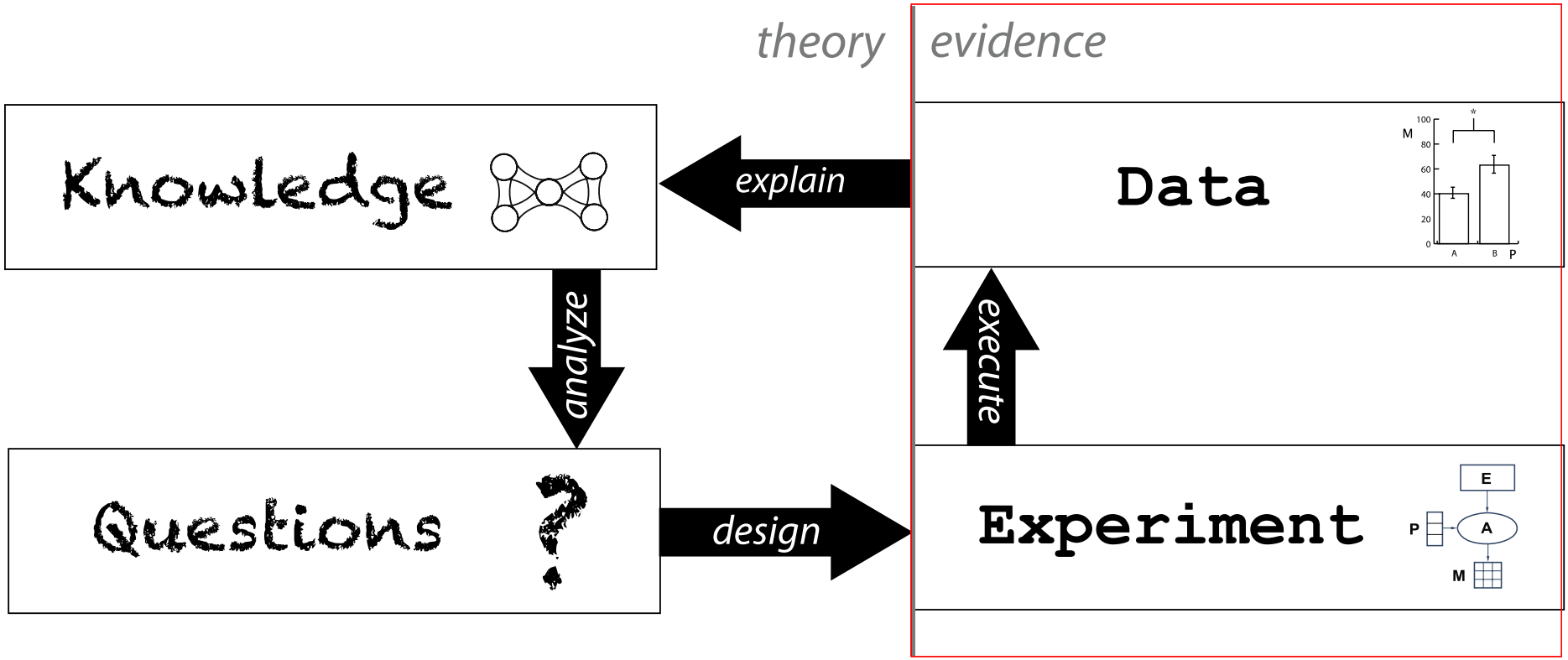


IntAct

https://www.ebi.ac.uk/intact/

- European Bioinformatics Institute @ Cambridge, England
- A large-scale, high-quality manually curated database
- Content
  - 20,065 curated papers in total
  - 2,254 open access papers
- Describes molecular interactions + detection methods linked to specific subfigures (*e.g.,* Fig. 1b, 3d, etc.)

Scientific Cycles of Investigation



Can we automatically classify types of scientific ~~evidence~~?

figures from articles?

Pipeline

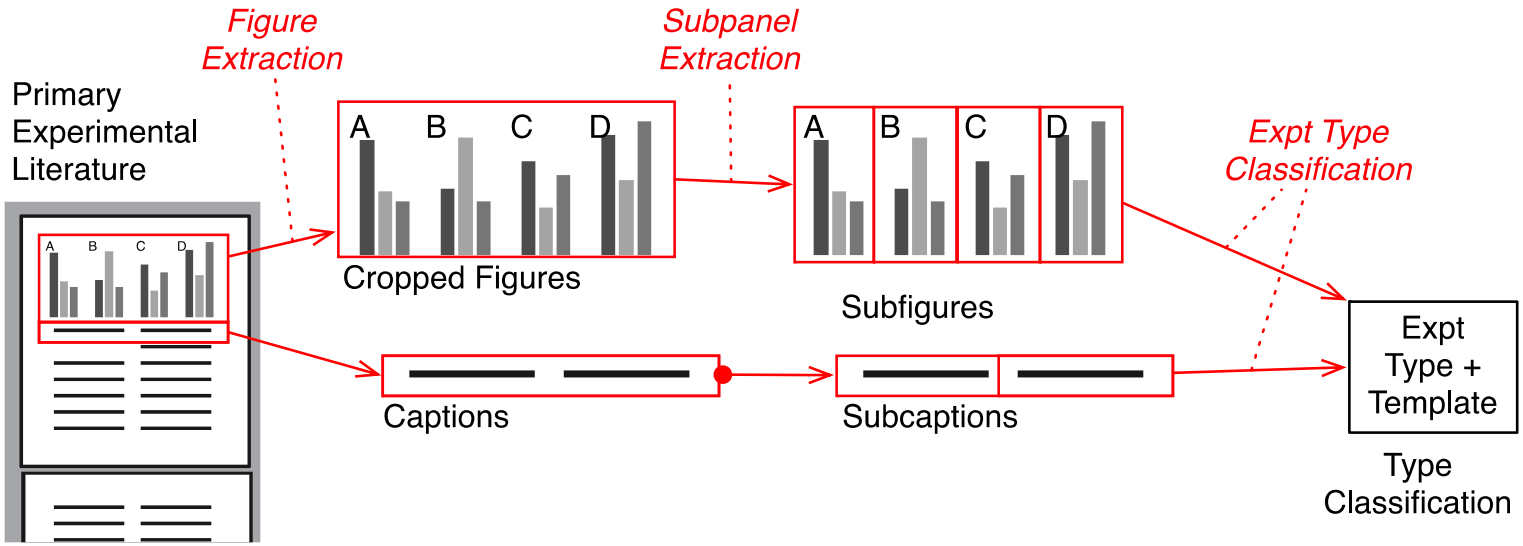


Figure Extraction
Subpanel Extraction
Expt Type Classification
Primary Experimental Literature
A B C D
Cropped Figures
A B C D
Subfigures
Captions
Subcaptions
Expt Type + Template
Type Classification
PMC Articles (.pdf + .nxml)

1. Extracting Whole Figures

https://github.com/BMKEG/lapdftext

1. Build a spatial index of tightly packed text blocks
2. Locate Caption Text blocks

   Regular Expression: ^Fig(ure|ure.|.)[0,1]\s*(\d+)
3. Find areas of low word density above, below and to the side of the caption
4. Remove caption text
5. Crop low word density space + caption region

indicate that the tumor progression to this most aggressive carcinoma stage in *Csf1*^{op}/*Csf1*^{op} PyMT mice was >10 wk delayed compared with +/*Csf1*^{op} littermates (Fig. 2 D).

Paucity of Tumor-associated Macrophages Is Correlated with Delayed Tumor Progression. CSF-1 is a macrophage growth factor, suggesting that the recruitment of macrophages to the tumor might be a factor that promotes the aggressive development of the +/*Csf1*^{op} tumors. Consequently, we examined the histology of the mammary tumors and the surrounding stroma from both +/*Csf1*^{op} and *Csf1*^{op}/*Csf1*^{op} PyMT mice at the period of the transition to carcinoma (~10 wk of age). Before the transition at 7 to 8 wk of age, a dramatic increase of leukocytic infiltration was found around the primary mammary tumors in +/*Csf1*^{op} mammary glands (Fig. 3 A, indicated by arrows). No such increase was observed at the tumor site in *Csf1*^{op}/*Csf1*^{op} mammary glands (Fig. 3 B), though the primary tumors in both *Csf1*^{op}/*Csf1*^{op} and +/*Csf1*^{op} PyMT mice had progressed to similar nonmalignant stages (adenoma). Immunohistochemical analysis, using a monoclonal antibody against the macrophage lineage–specific marker, F4/80, showed that a large percentage of infiltrated leukocytes in the +/*Csf1*^{op} mammary gland were F4/80 positive (Fig. 3 C) and that few such cells were found around *Csf1*^{op}/*Csf1*^{op} primary tumors (Fig. 3 D). In concert with the histopathological development of the primary tumors to the carcinoma stage, the infiltration of leukocytes became more intense and focal infiltration sites were often seen in the mammary glands of +/*Csf1*^{op} PyMT mice. Densely infiltrated cells with the morphology of granulocytes, mast cells, and monocyte-like cells were found in these sites and the tumor acini adjacent to them often displayed a disrupted boundary (Fig. 3 E, arrows). This suggests that the basement membrane of the acini at the infiltration site had lost its integrity, potentially allowing tumor cells to migrate into the adjacent connective tissue. These leukocytic infiltration sites were detected in +/*Csf1*^{op} PyMT mice as early as 9 wk of age but were absent in *Csf1*^{op}/*Csf1*^{op} mammary glands at the same age (Fig. 3 F). Furthermore, an intensive infiltration of F4/80^{+} cells was found in the vicinity of +/*Csf1*^{op} tumor that had developed to the carcinoma stage (Fig. 3, G and I), whereas the density of F4/80^{+} cells was still reduced in *Csf1*^{op}/*Csf1*^{op} tumors, even in those that had developed to the same histological stage (Fig. 3, H and J).

These results are consistent with the hypothesis that CSF-1 acts through macrophages in its promotion of tumor progression. However, in human mammary tumor, CSF-1 receptor expression has been detected in tumor cells as well as infiltrated macrophages (8). To determine if macrophages are the only target cell at the tumor site in mice, we performed in situ hybridization for CSF-1R expression. CSF-1R–positive cells were found to be in the infiltrated cells surrounding the tumor in +/*Csf1*^{op} mammary glands but the tumor cells were consistently negative (Fig. 4 A, ii). The positively stained cells were mononuclear, dendritic cells with elongated cell bodies, and were in the same location as F4/80^{+} cells consistent with their identification as macrophages (Fig. 4 A, iii). Very few positive cells were found when the same *cfms* antisense probe was hybridized to tumor-bearing *Csf1*^{op}/*Csf1*^{op} mammary glands (data not shown). The sense *cfms* probe on adjacent +/*Csf1*^{op} mammary gland sections was consistently negative (Fig. 4 A, i). To obtain a quantitative comparison of macrophage infiltration in the mammary tumors of *Csf1*^{op}/*Csf1*^{op} and +/*Csf1*^{op} PyMT mice, the expression of *cfms* was determined by Northern analysis. Expression of *cfms* mRNA was detected in +/*Csf1*^{op} mammary glands and an approximately threefold increase of the mRNA was observed at 6 wk compared with 4 wk of age (Fig. 4 B). The *cfms* mRNA in *Csf1*^{op}/*Csf1*^{op} mammary gland was barely detectable until 12 wk, at which age the level of *cfms* mRNA was less than the level found in +/*Csf1*^{op} mammary glands at 4 wk of age (Fig. 4 B). As all mononuclear phagocytes express the CSF-1R, this data is consistent with both the F4/80^{+} im-

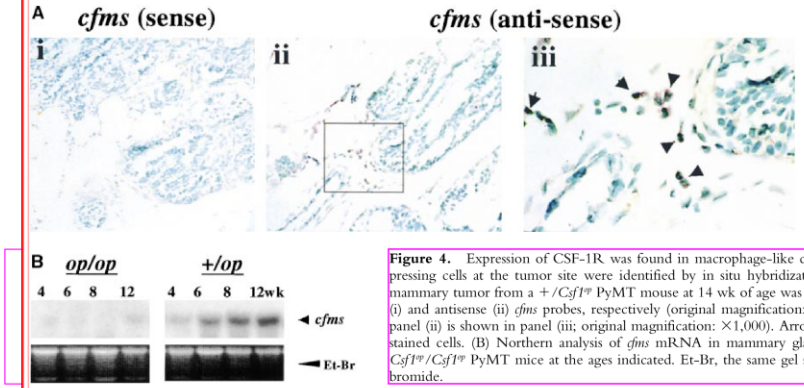
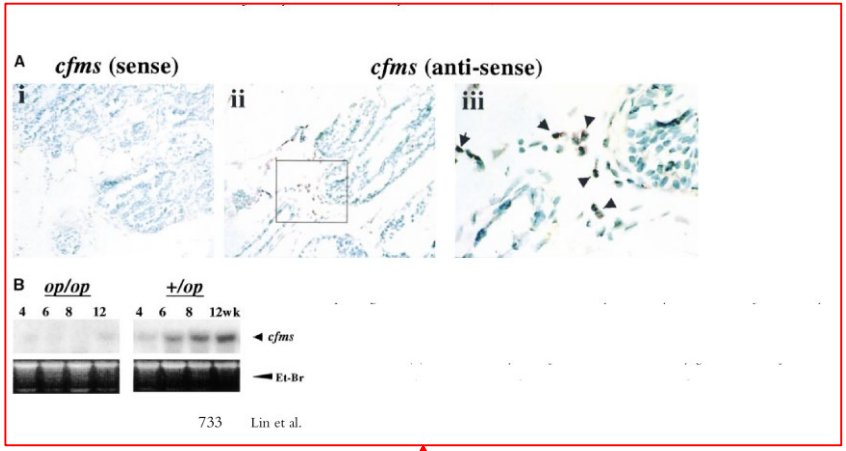
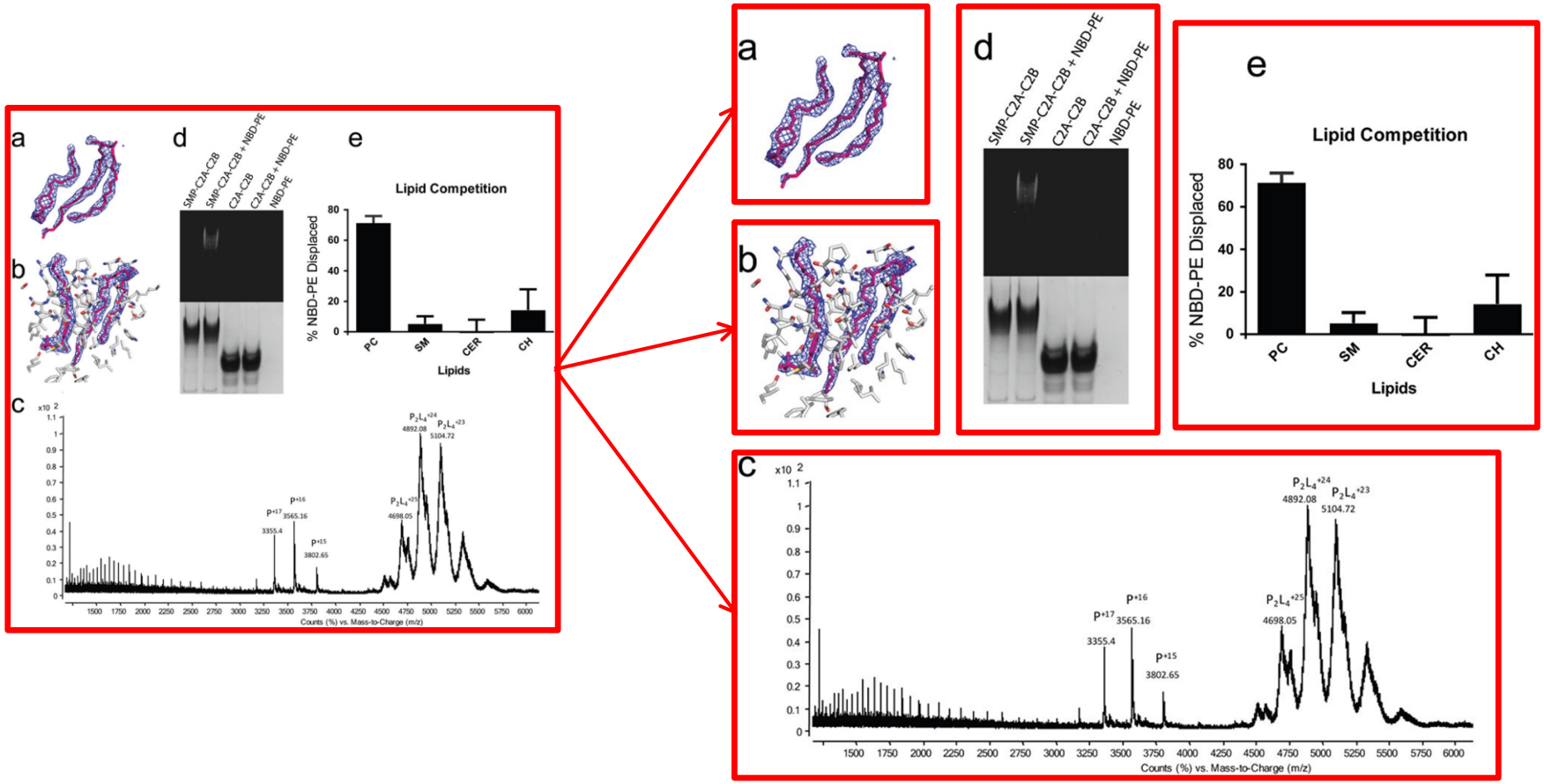


Figure 4. Expression of CSF-1R was found in macrophage-like cells. (A) CSF-1R expressing cells at the tumor site were identified by in situ hybridization for *cfms*. Primary mammary tumor from a +/*Csf1*^{op} PyMT mouse at 14 wk of age was hybridized with sense (i) and antisense (ii) *cfms* probes, respectively (original magnification: ×250). The inset in panel (ii) is shown in panel (iii; original magnification: ×1,000). Arrows indicate positively stained cells. (B) Northern analysis of *cfms* mRNA in mammary glands of +/*Csf1*^{op} and *Csf1*^{op}/*Csf1*^{op} PyMT mice at the ages indicated. Et-Br, the same gel stained with ethidium bromide.



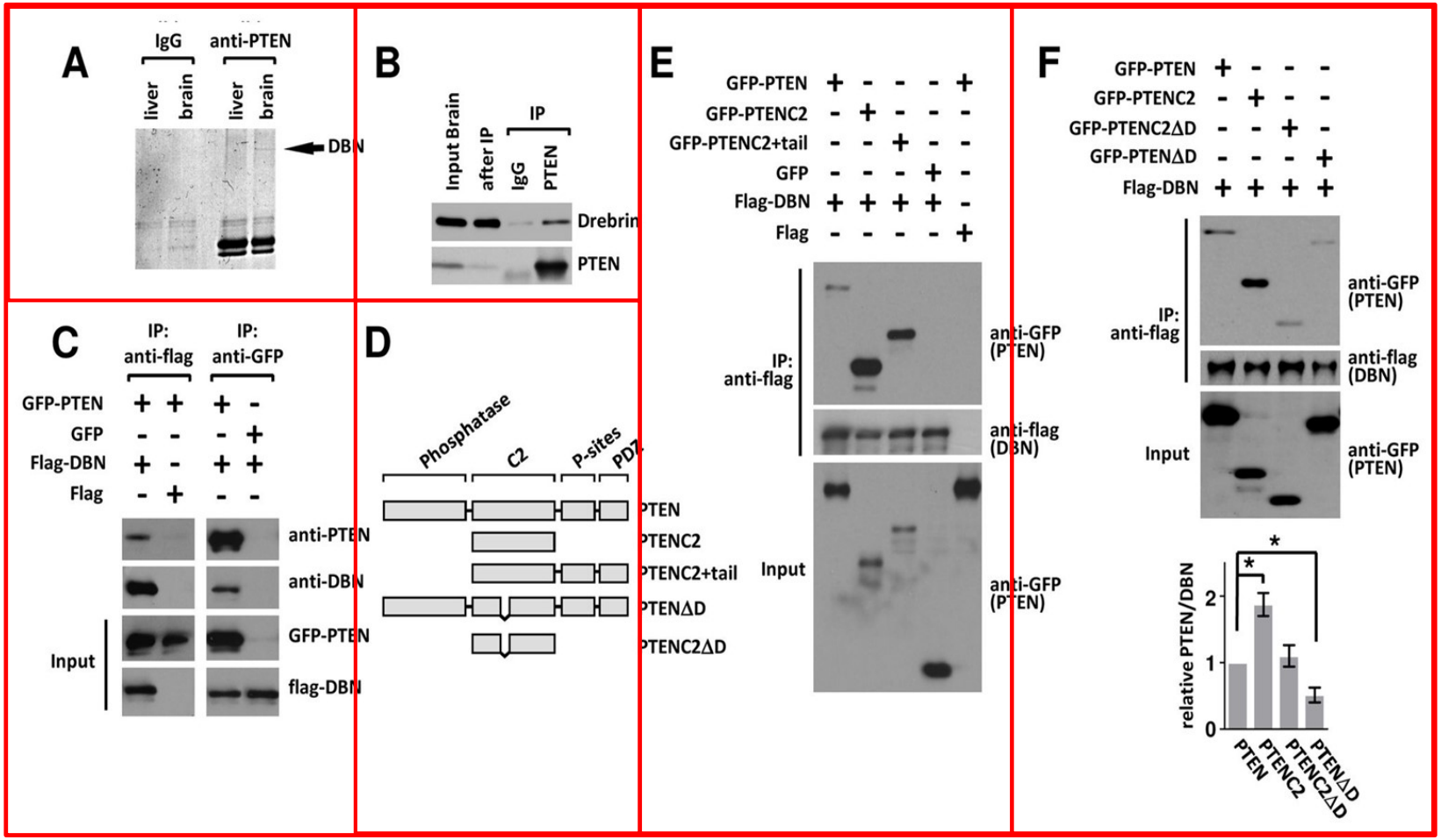
2. Delineating Subfigures



Heuristic Baseline Method

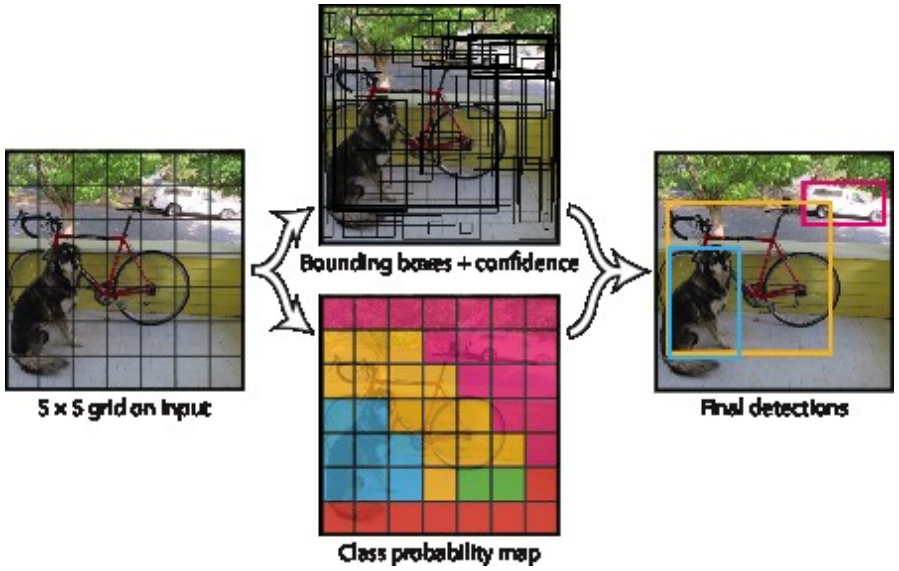
1. Detect letters that denote each subfigure (‘A’, ‘B’, etc.) using connected component analysis.
2. Use a greedy tiling mechanism that places the letter in the top left corner of panels to construct a rectangular layout for each panel in a figure.
3. A figure is cut into multiple sub-panels by straight lines that go along the top or left side of each detected letter.

Baseline Example



‘You Only Look Once: Unified, Real-Time Object Detection

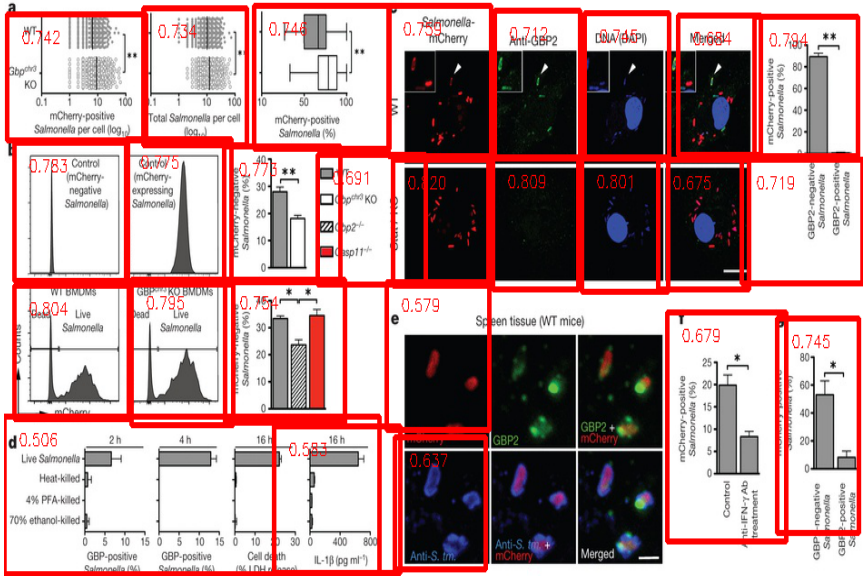
Redmon *et al.* (2015) *CoRR* abs/1506.02640



Used for Compound Figure Separation by Tsutsui & Crandall, (2017), ICDAR

YOLO is biased to fine-grained delineation

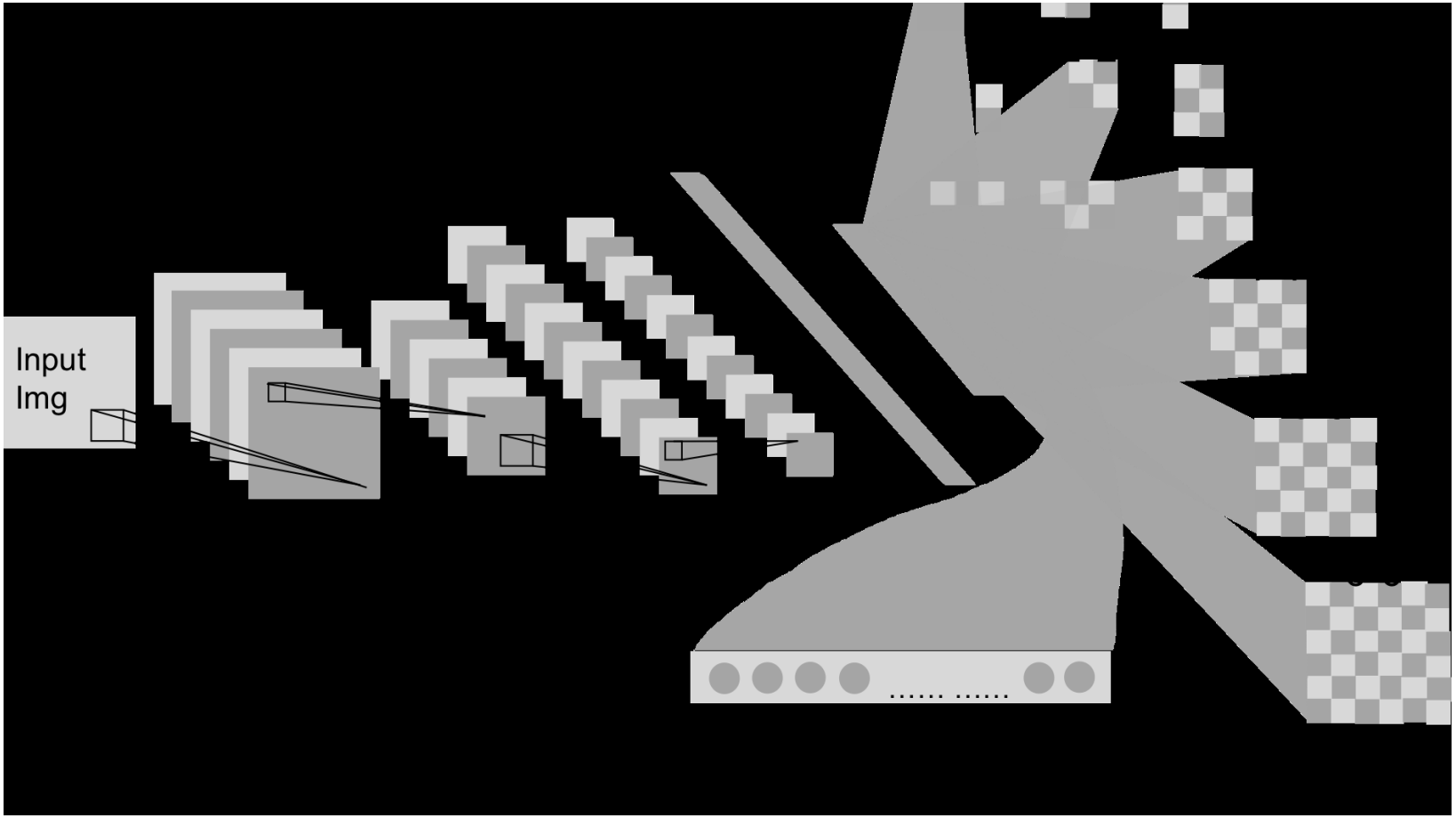
YOLO Grid

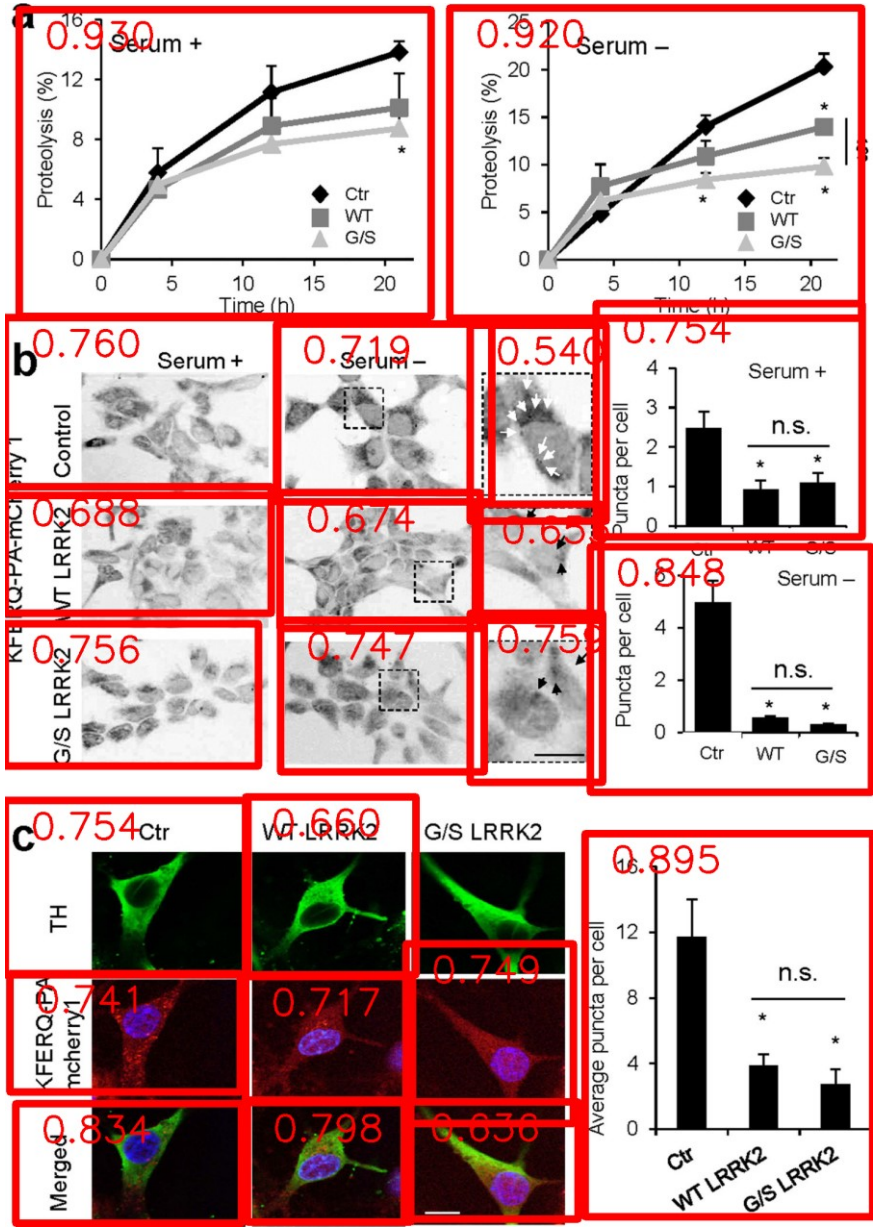


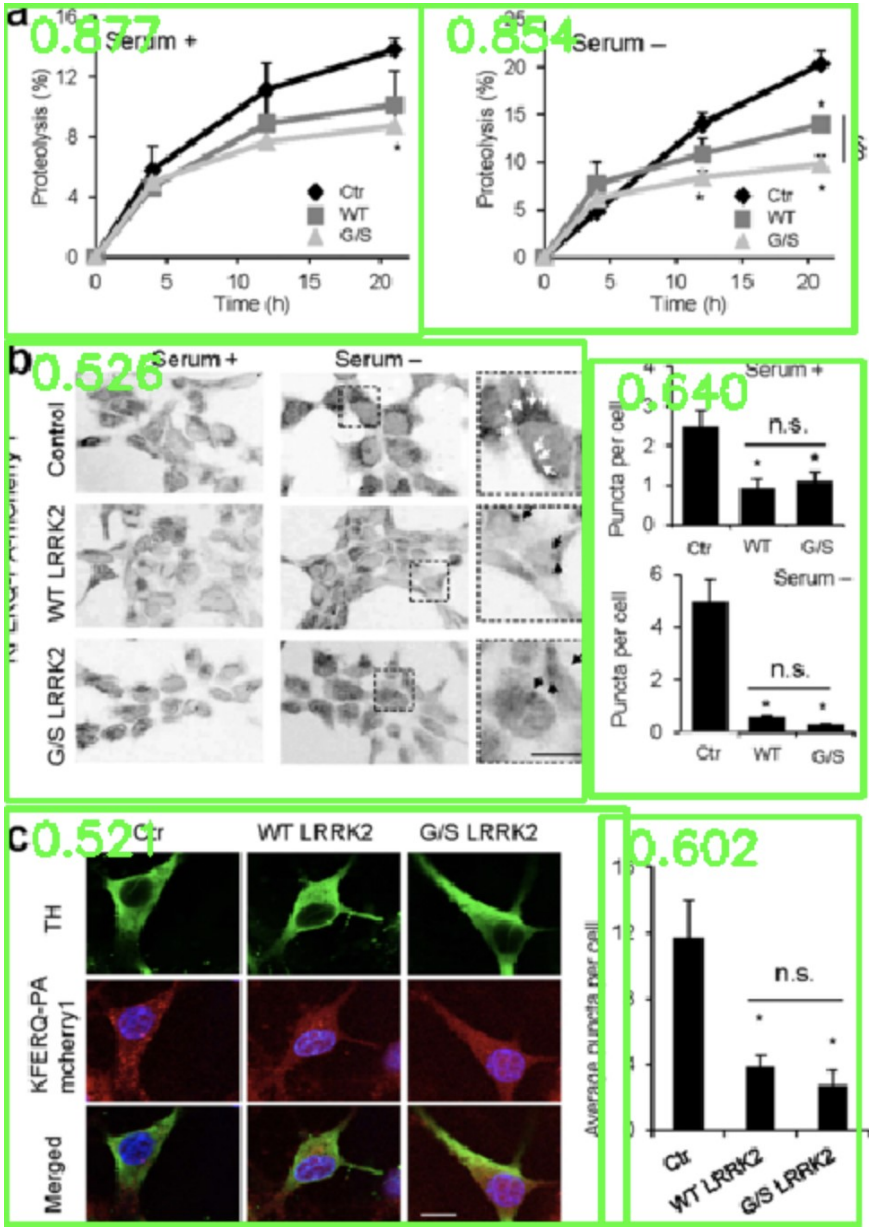
In standard YOLO, the hyper-parameter: **# of Grids** is set as 13*13,
Its too many for academic images, the network always favor tiny subfigures. (bias)

Different layout is sensible to Grid #.
We apply a model based on applying multiple sets of grids.

Layout Aware YOLO




a
Serum +
Serum –
Proteolysis (%)
Time (h)
Ctr
WT
G/S
b
Control
WT LRRK2
G/S LRRK2
Puncta per cell
n.s.
c
Ctr
WT LRRK2
G/S LRRK2
TH
Merged
Average puncta per cell


a
Serum +
Serum –
Proteolysis (%)
Time (h)
Ctr
WT
G/S
b
Control
WT LRRK2
G/S LRRK2
Puncta per cell
n.s.
c
Ctr
WT LRRK2
G/S LRRK2
TH
Merged
Average puncta per cell

Preliminary Analysis of data from INTACT 2017

Method	Accuracy
Heuristic Greedy Cut	0.78
YOLO	0.76
Layout Aware YOLO	**0.84**

3a. Classifying Subfigures

- LeNet image classification algorithm
- Manually annotated images from INTACT database

Table 1. Subfigure type detection performance.

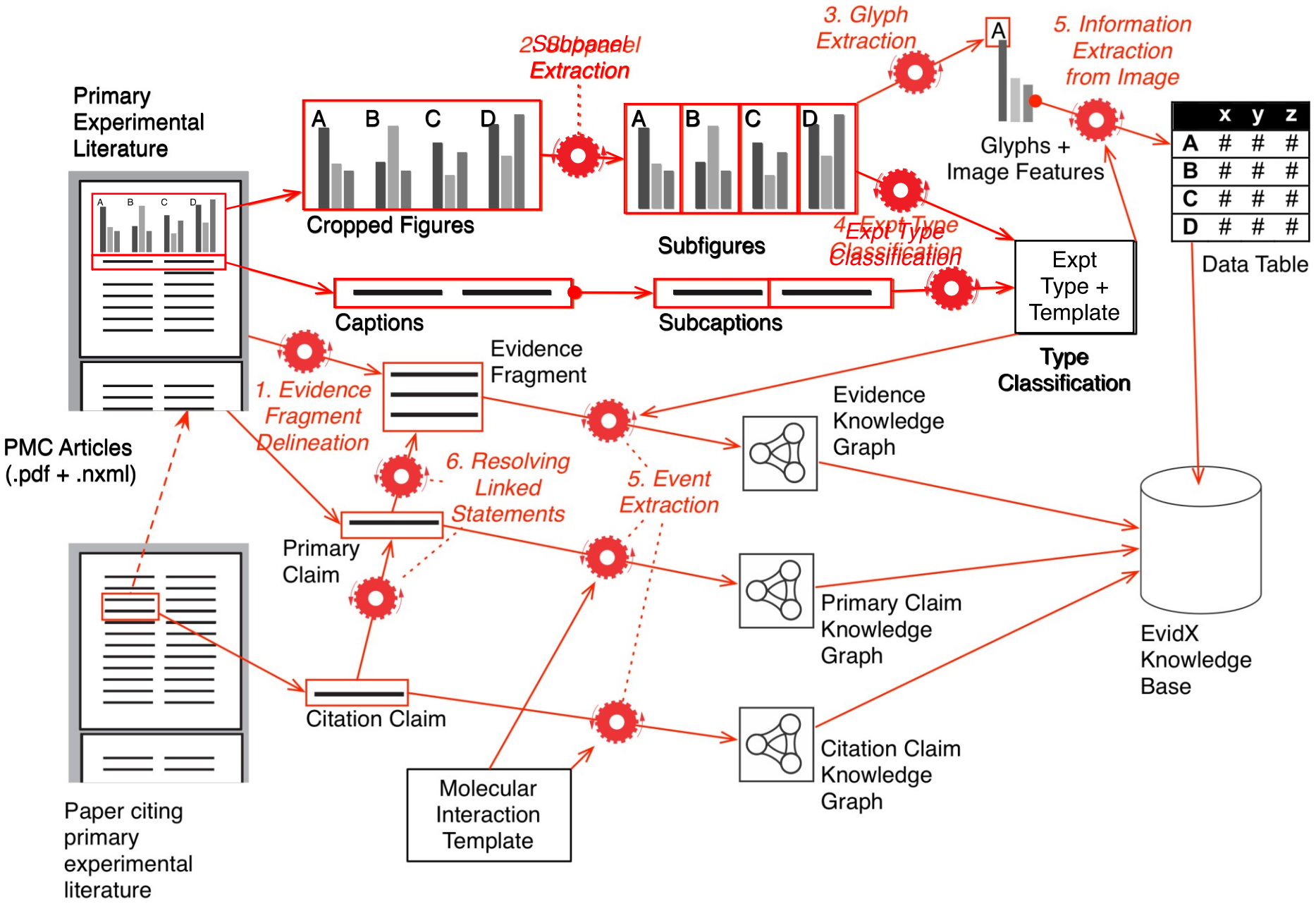
Figure Type	N(train)	N(test)	Tagging Accuracy
Chart	980	315	0.92
Diagram	819	197	0.40
Gel	1402	404	0.83
Histology	1299	398	0.97

3b. Classifying Evidence Text

- Train simple CNN + LSTM classifiers predict detection method codes from INTACT 'evidence fragments' and sub-captions
- 3,366 open access INTACT records

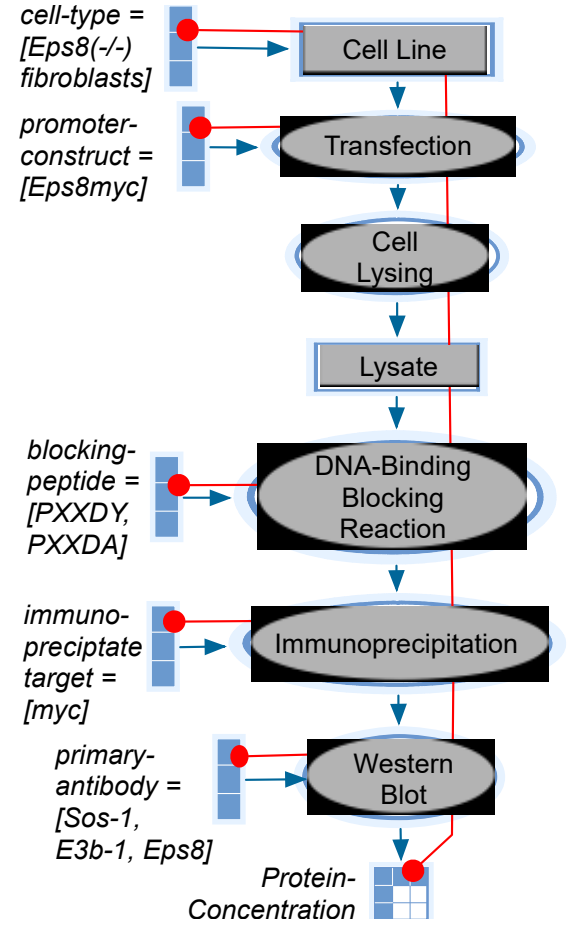
Table 2. Accuracy for experimental type classification from text.

	Evidence Fragment		Sub-Caption	
Detection Method	LSTM	CNN	LSTM	CNN
Participant (48 types)	0.37	0.48	0.48	0.59
Participant (6 types)	0.58	0.70	0.72	0.75
Interaction (122 types)	0.26	0.50	0.56	0.62
Interaction (18 types)	0.71	0.73	0.77	0.83
Interaction(Co-IP tagging)	0.79	0.84	0.87	0.90
Participant(WB tagging)	0.71	0.79	0.76	0.85

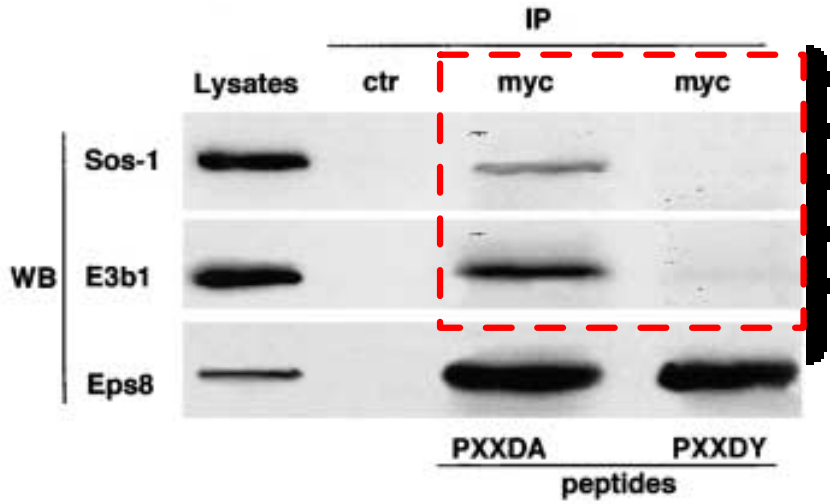

Primary Experimental Literature
A B C D
Cropped Figures
2. Subpanel Extraction
A B C D
Subfigures
3. Glyph Extraction
A
Glyphs + Image Features
5. Information Extraction from Image
x y z
A # # #
B # # #
C # # #
D # # #
Data Table
4. Expt Type Classification
Expt Type + Template
Type Classification
Captions
Subcaptions
1. Evidence Fragment Delineation
Evidence Fragment
Evidence Knowledge Graph
PMC Articles (.pdf + .nxml)
6. Resolving Linked Statements
5. Event Extraction
Primary Claim
Primary Claim Knowledge Graph
EvidX Knowledge Base
Citation Claim
Citation Claim Knowledge Graph
Molecular Interaction Template
Paper citing primary experimental literature

‘Towards’ Evidence Extraction Templates

A



B



C

cell-type	promoter-construct	blocking-peptide	immuno-precipitate	primary-antibody	protein-concentration
Eps8(-/-) fibroblasts	Eps8myc	PXXDY	myc	Sos-1	none
Eps8(-/-) fibroblasts	Eps8myc	PXXDY	myc	E3b-1	none
Eps8(-/-) fibroblasts	Eps8myc	PXXDA, control	myc	Sos-1	low
Eps8(-/-) fibroblasts	Eps8myc	PXXDA, control	myc	E3b-1	low

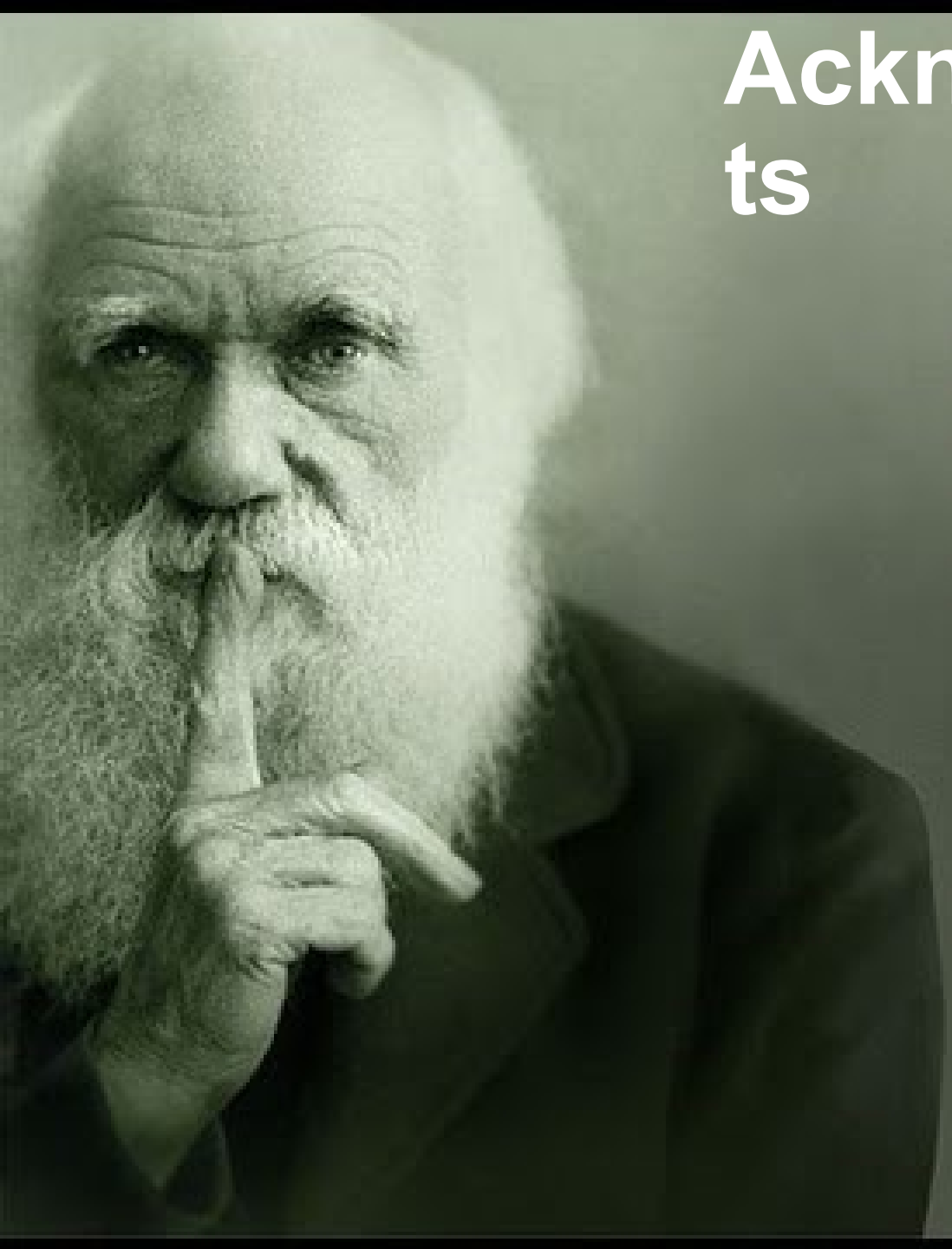
Linked Data as a Research Object http://purl.org/ske/ro/semsci18

Tools:

- https://github.com/SciKnowEngine/evidX/releases/tag/v0.1.0
  TensorFlow Classifiers used to classify text of subfigure captions by method type.
- https://github.com/SciKnowEngine/UimaBioC
  Text preprocessing pipelines
- https://github.com/SciKnowEngine/lapdftext
  PDF image and text extraction tools

Data

- https://doi.org/10.5281/zenodo.1315036
  'Molecular Biology Open Access Pubmed Word and Sentence Representations'
- https://doi.org/10.5281/zenodo.13150211
  'Method Classification of Open Access INTACT Molecular Interaction data.'
- https://doi.org/10.5281/zenodo.1319198
  'Partitioned Image Data for Machine Learning Analysis of Molecular Biology Figures'



Acknowledgements

Nanjun Peng
Xiangci Li
Pradeep Dasigi
Ed Hovy
Anita De Waard

NIH 5R01LM012592-02

DARPA Big Mechanism under ARO contract W911NF-14-1-0436


CAN A NINJA CATCH AN ARROW? ON THIS EPISODE, WE'LL FIND OUT!
MMM, SCIENCE.
HEY. MYTHBUSTERS IS ENTERTAINING, BUT IT'S NOT SCIENCE.
BRAAAAAI!INNS...
ZOMBIE FEYNMAN!
YOU GOT A PROBLEM WITH MYTHBUSTERS?
THEY FAIL AT BASIC RIGOR!
"IDEAS ARE TESTED BY EXPERIMENT." THAT IS THE CORE OF SCIENCE. EVERYTHING ELSE IS BOOKKEEPING.
BY TEACHING PEOPLE TO HOLD THEIR BELIEFS UP TO EXPERIMENT, MYTHBUSTERS IS DOING MORE TO DRAG HUMANITY OUT OF THE UNSCIENTIFIC DARKNESS THAN A THOUSAND LESSONS IN RIGOR.
SHOW THEM SOME LOVE.
ANYWAY, BACK TO ZOMBIE STUFF. I HUNGER FOR BRAAAAAI!INNS!
UH, TRY THE PHYSICS LAB NEXT DOOR.
I SAID BRAINS. ALL THEY'VE GOT ARE STRING THEORISTS.