Quantifying Scientific Discovery to Improve the Knowledge of Facts

LPI, Paris
15 December 2022

Alexander (Sasha) Belikov
Hello Watt, Paris, France

Vocabulary

Fact - a universal truth (reproducible given a specified context)

Context - the domain of applicability of a fact. Includes methods, conditions, quantifiers etc

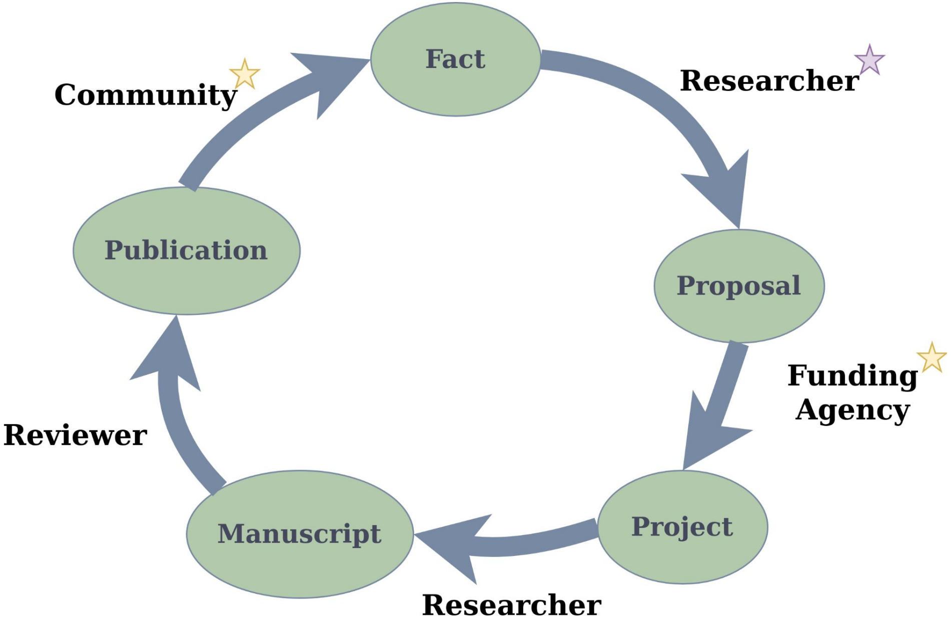
Proposal / Publication - a statement that is not universal. Publications contain **claims**.

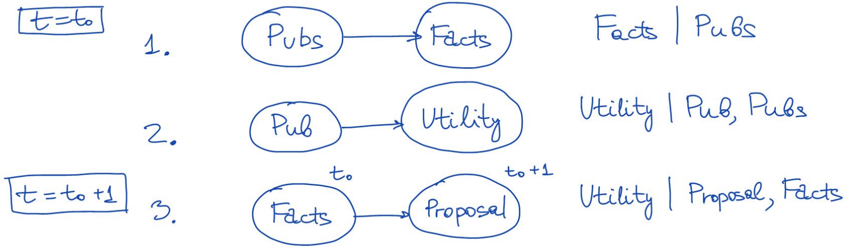
T(rue) / Utility / Value - scalar functions defined on **Facts** / **Proposals**, e.g. claim correctness, fact truthfulness.

Examples

- Probability distribution of **Fact** being true : P(**T =** true | **Fact**, **Context**).
- Utility function over **Facts, Context** : P (**Utility** | **Proposal**, **Context**). Surprise
-

The Cycle of Knowledge Accumulation





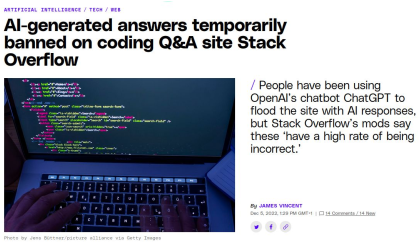
Publication to Fact Model

Traditionally we … wait, for empirical convergence

But empirical convergence to consensus is not sufficient.

Credit: Dilok Klaisataporn/iStock.

1. Consensus convergence is often not monotonic. Reproducibility crisis. Bias. Lack of transparency.
2. Given the amount of data, discovery of pertinent information is hindered.
3. Generative AI: an avalanche (ChatGPT, Galactica).



ARTIFICIAL INTELLIGENCE / TECH / WEB

AI-generated answers temporarily banned on coding Q&A site Stack Overflow

/ People have been using OpenAI's chatbot ChatGPT to flood the site with AI responses, but Stack Overflow's mods say these 'have a high rate of being incorrect.'

By JAMES VINCENT

Dec 5, 2022, 1:29 PM GMT+1 | 14 Comments / 14 New

Photo by Jens Büttner/picture alliance via Getty Images

Credit: Verge

* - "Nonreplicable publications are cited more than replicable ones." Serra-Garcia et al., Science Advances (2021).

Examples of “deviant” empirical convergence

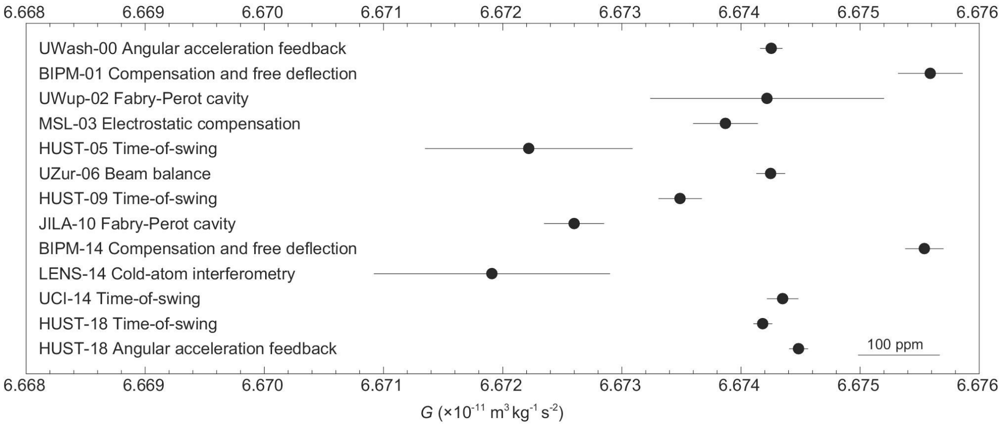
Social Sciences. Psychology.

An effect called ego depletion: willpower can be worn down over time was found by Baumestier et al., 1998 (more than 7K citations).

Hagger et al. (2016) tried to replicate these results in 24 labs. And failed.

Hard Sciences. Physics.

Gravitational constant remains the physical constant with the largest systematic error.



Xue et al, *National Science Review*, Vol. 7, Issue 12, 2020

From Publication to Fact: beyond Empirical Convergence

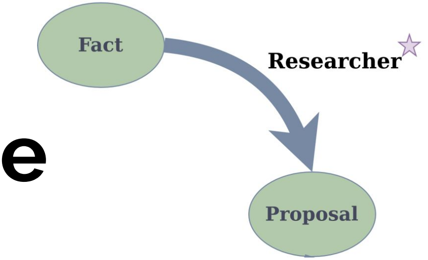
How to decide if **Fact** is true?

Claims $\boldsymbol{C_i}$ empirically converge to facts: **lim** $\boldsymbol{C_i}$ at $\boldsymbol{t > t_{thr}}$ exists, and equals to $\boldsymbol{F^*}$ which we then choose to use as the definition of **Fact**.

P(**Fact** | {($\boldsymbol{C_i}$, $\boldsymbol{Context_i}$)}) : where $\boldsymbol{C_i}$ is made in $\boldsymbol{Context_i}$

It would be nice to know if claim $\boldsymbol{C_j}$ is correct… P($\boldsymbol{C_j}$ | **Fact,** {($\boldsymbol{C_i}$, $\mathbf{Context_i}$)})

Define distance $\boldsymbol{D}$($\mathbf{C_i}$, **Fact**) which is a proxy for correctness.



Example: Discovery of Ornithine Cycle

Urea (Ornithine) Cycle by Krebs (1932)

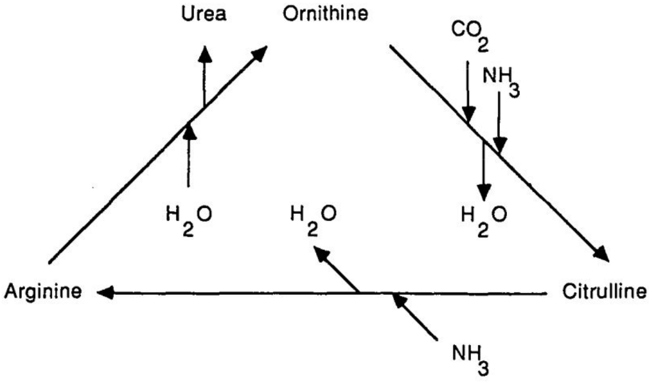
Ammonia (**NH_3**) is the amino acid reaction product that has to be converted to a less toxic substance **urea ($(NH_2)_2CO$)** for safe removal. **Urea** had been synthesized in early 19th century but what about *in vivo?*

The knowledge of its composition and synthesis paths lead to several possibilities.
Hypothesis: **ammonium salts**, **leucine**, **tyrosine** and **aspartic acid** increase the formatio of **urea**. Urea produced from amino acids and ammonia?
Method: perfusion left the question of the actual mechanism undecided.

1. Used **ornithine** (less common), positive effect

2. Narrowed the scope by looking at derivatives of **ornithine**, negative results.

3. New apparatus let him measure the quantities of **urea** produced and **ammonia** consumed. Thought that the (known) **arginine** reaction, by which **arginine** is converted to **ornithine** and **urea**, might be related to the **ornithine** effect.

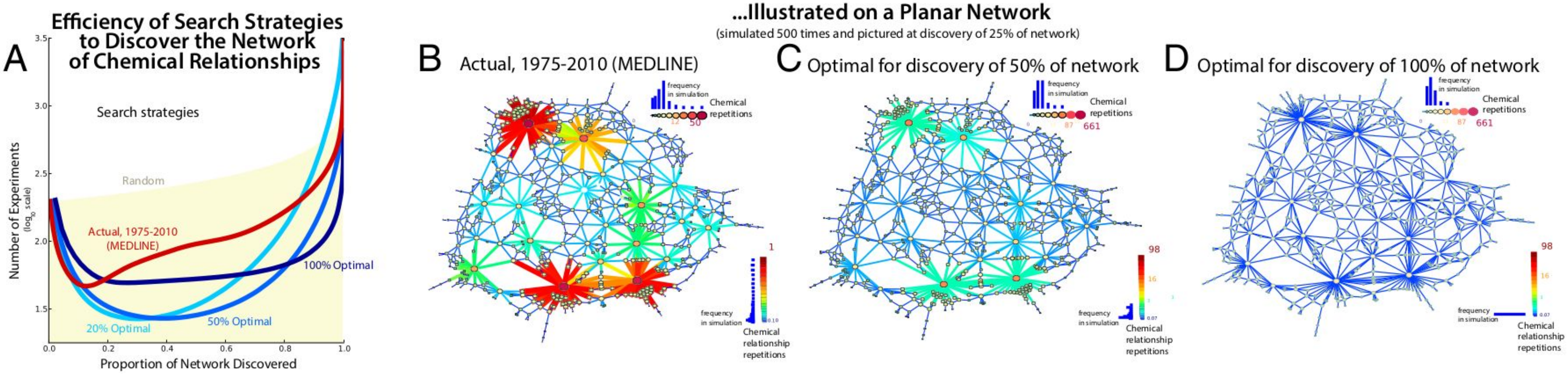


KEKADA,
by Kulkarni et al., 1988

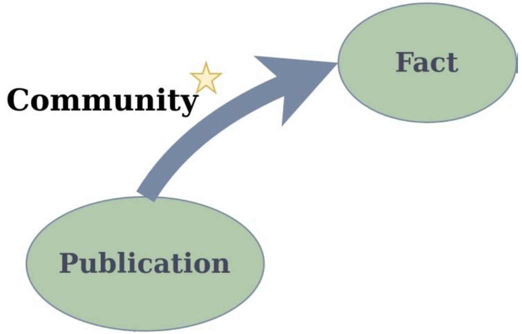
Discovery of Facts

Domain: Biomedical chemistry

Data: Medline and US Patents



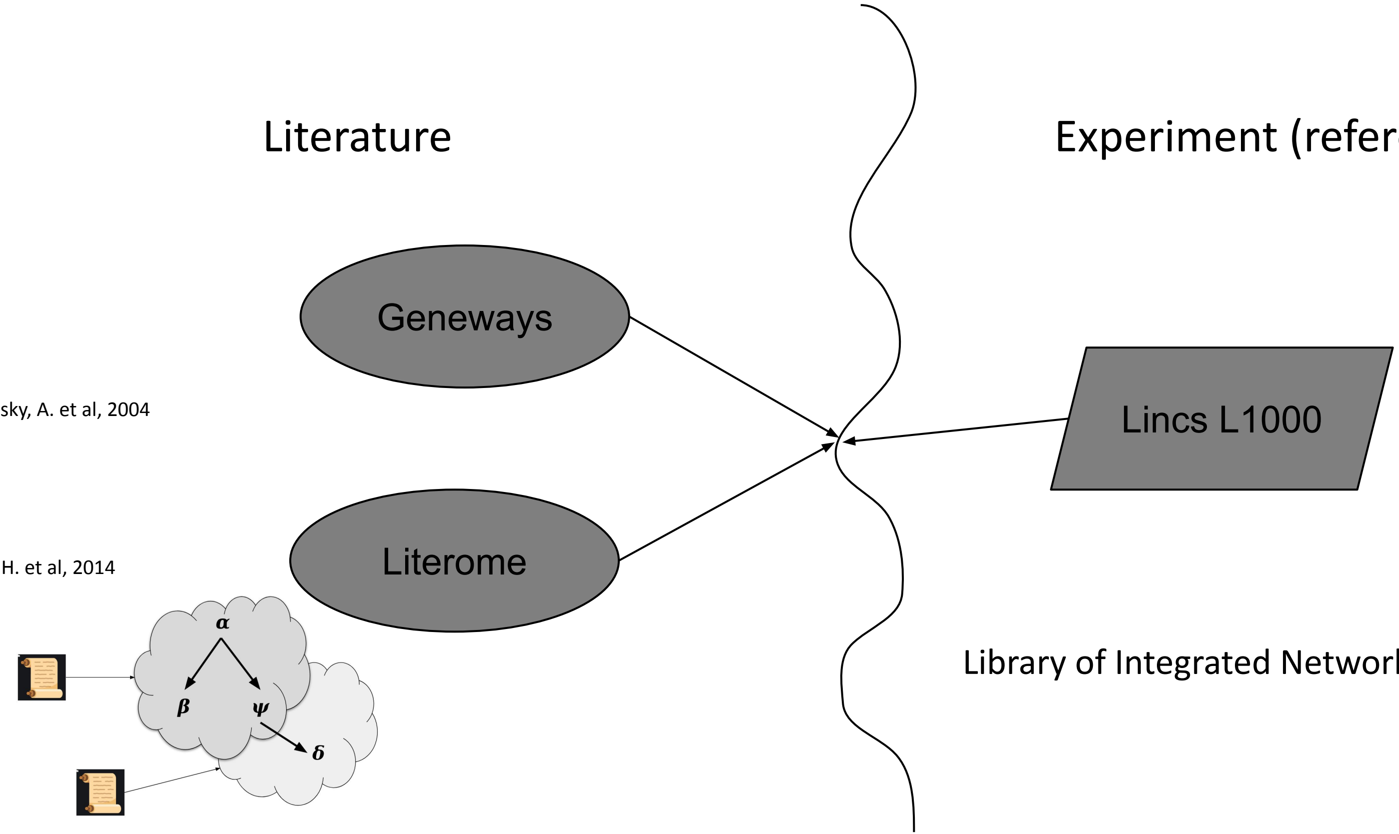
Rzhetsky et al, PNAS, 2015



Prediction of robust scientific facts

Belikov, A.V., Rzhetsky, A. & Evans, J. Prediction of robust scientific facts from literature. *Nat Mach Intell* **4**, 445–454 (2022)

Experimental setup. Datasets



Literature

Experiment (reference)

Geneways

Rzhetsky, A. et al, 2004

Literome

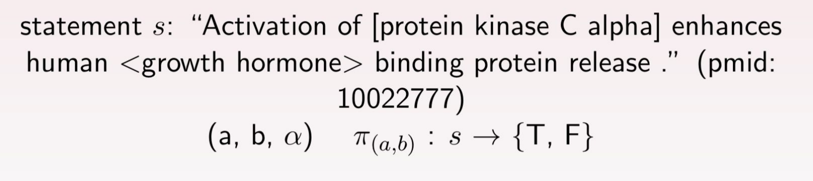
Poon,H. et al, 2014

Lincs L1000

Subramanian, A. et al, et al, 2017

Library of Integrated Network-Based Cellular Signatures

Literature Datasets

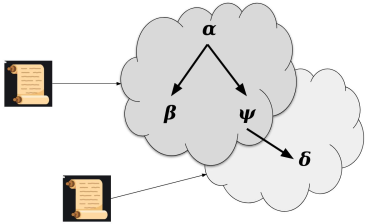


statement s: “Activation of [protein kinase C alpha] enhances human <growth hormone> binding protein release .” (pmid: 10022777)

(a, b, α) $\pi_{(a,b)} : s \to \{$T, F$\}$

	GeneWays	Literome
# claims	612K	409K
# publications	197K	220K
# genes	5,141	10,703
# gene-gene interactions	23,405	144,172
# positive claims	77%	96%
estimated precision	95%	25%

Literature: directed graph of interactions



claims c_i: "Activation of [protein kinase C alpha] enhances human <growth hormone> binding protein release ." (pmid: 10022777)

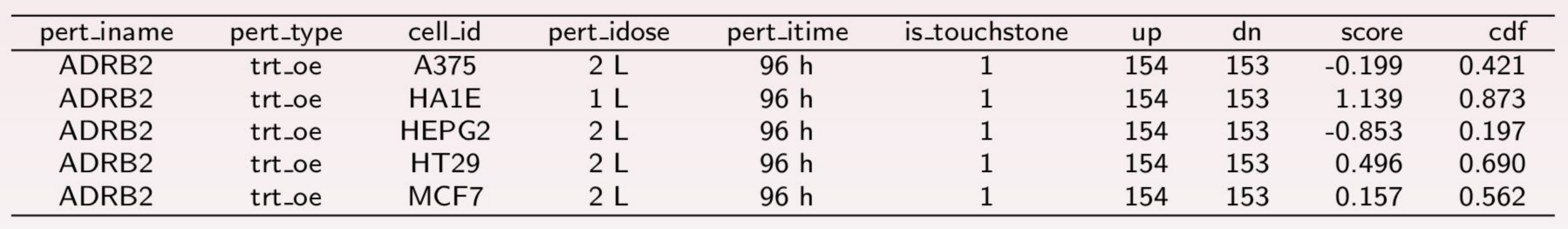
$$(\alpha,\ \beta,\ \mathrm{r}) \quad \pi_{(\alpha,\beta)} : c_i \rightarrow \{\mathrm{T},\ \mathrm{F}\}$$

Lincs L1000: directed graph of interactions

measures genome-wide mRNA

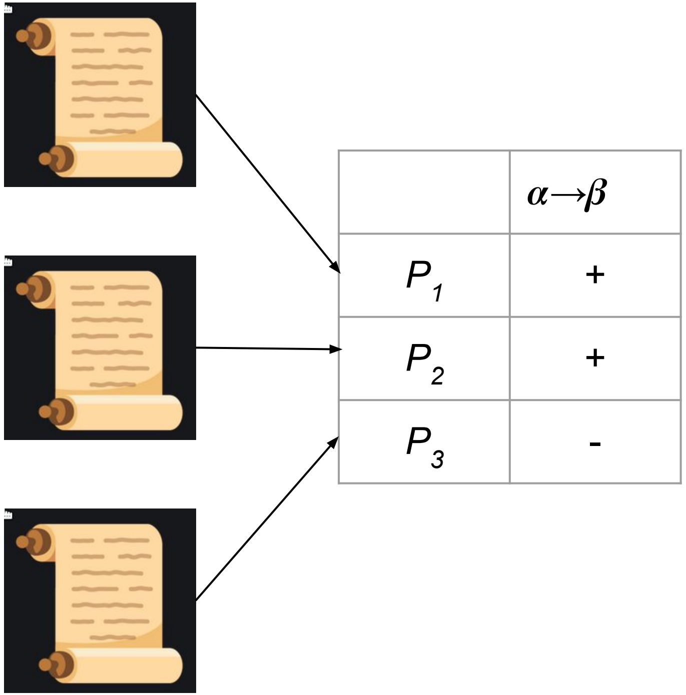
1.3M gene profiles, for a total of 474K gene signatures

71 cell lines, from 19 primary sites



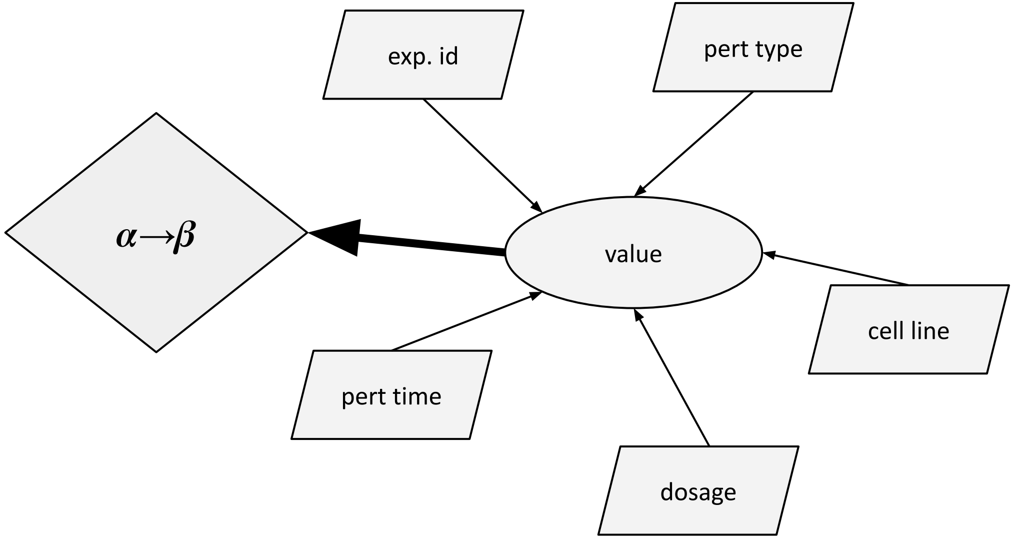
pert_iname	pert_type	cell_id	pert_idose	pert_itime	is_touchstone	up	dn	score	cdf
ADRB2	trt_oe	A375	2 L	96 h	1	154	153	-0.199	0.421
ADRB2	trt_oe	HA1E	1 L	96 h	1	154	153	1.139	0.873
ADRB2	trt_oe	HEPG2	2 L	96 h	1	154	153	-0.853	0.197
ADRB2	trt_oe	HT29	2 L	96 h	1	154	153	0.496	0.690
ADRB2	trt_oe	MCF7	2 L	96 h	1	154	153	0.157	0.562

Alignment: publications to experiments



	$\alpha \to \beta$
P_1	+
P_2	+
P_3	-

Each claim matters



Average over repeating experiments, take max over setup parameters

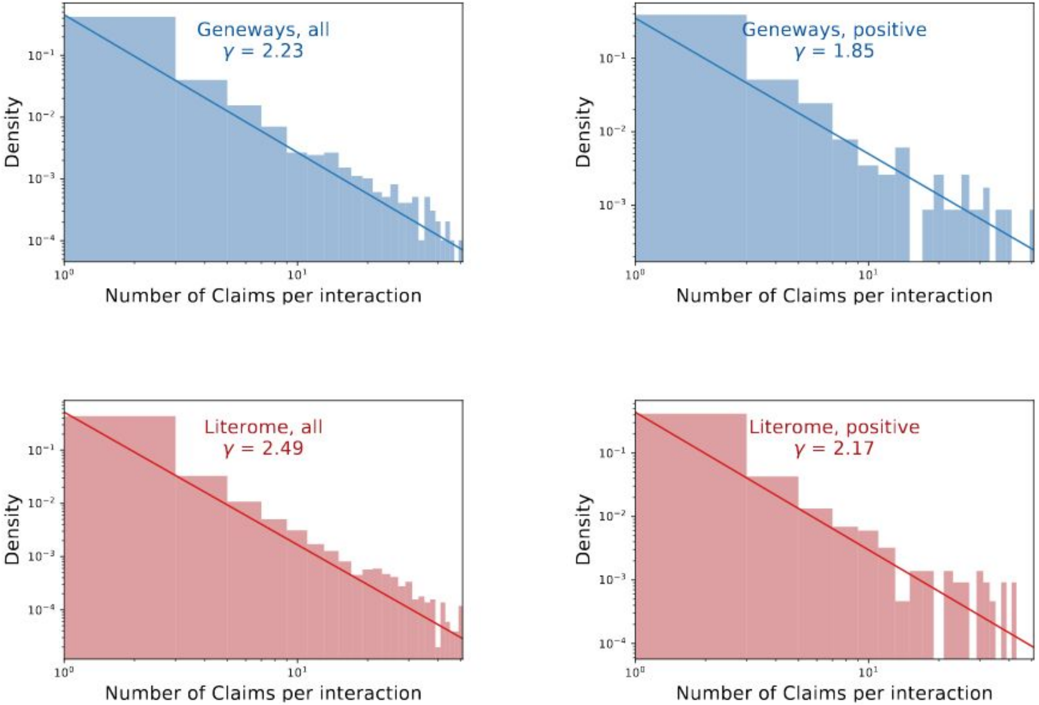
Dataset

1. Aggregate claims per publication
2. Take only claims from abstracts
3. Keep claims for which features can be derived.
4. Keep interactions mappable to LINCS L1000

Overlap between Geneways/Literome: 2K interactions or 827 claims with correlation ~ 0.38.

	GeneWays (claims/int)	Literome (claims/int)
# bi-projected	68.6K/36K	259K/144K
# feature merge	44K/23K	
# LINCS merged	15.5K/6.8K	50.5K/25.4K

Claim number distribution

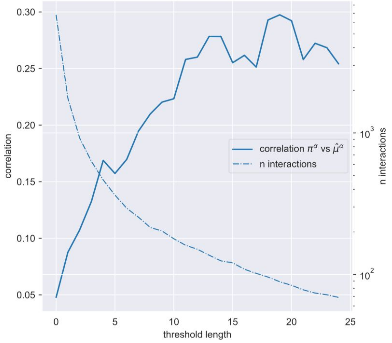


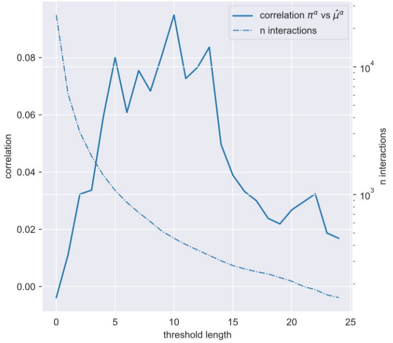
Peculiar distribution of published claims

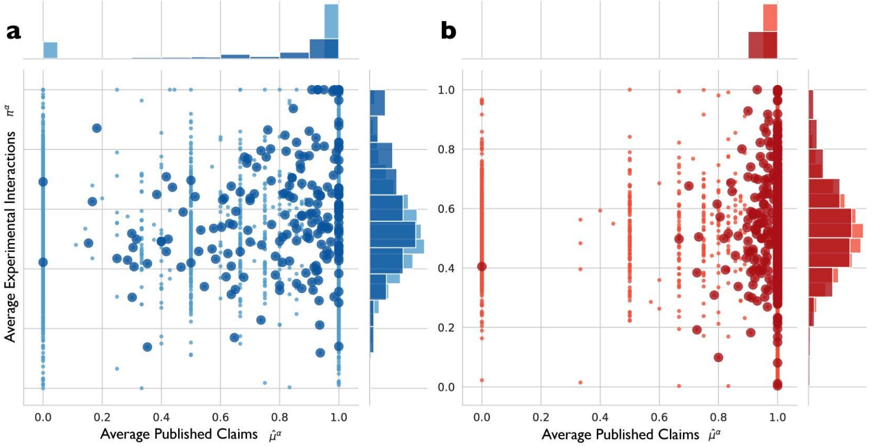
CDF of experimental strength does not correlate well with the mean of claims’ value, unless we start looking at more popular claims.

$(\alpha, \beta) : \{(c_i, f_i)\} ; \boldsymbol{\pi}$

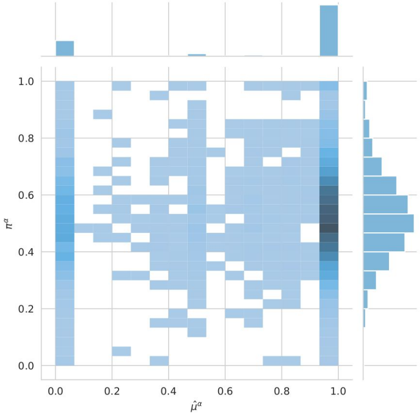
$$\mu = \frac{1}{n} \sum c_i$$







Preliminary conclusions

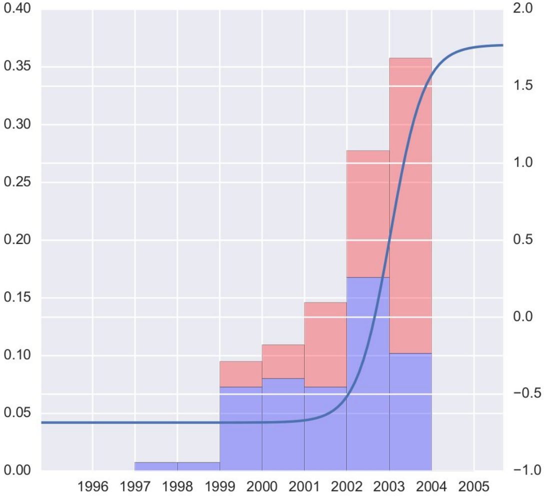


- Literature contains two types of claims: positive and negative.
- The distribution of the claims in the claims corpora has a strong bias towards positive claims.
- The distribution of the experimental strength (of in the experiment does not have a strong bias.
- The positive claim bias varies between Geneways and Literome.
- The interaction strength can be discretized into at least 3 categories: neutral, positive and negative.
- The correlation between interaction strength and the mean claim increases as we consider more popular interactions (defined as having more claims per interaction).

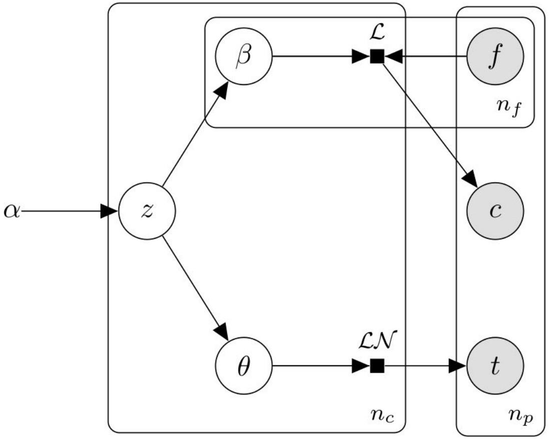
Graphical model point of view

Bayesian approach, graphical models (pymc, pyro).

Example on an ambiguous interaction (claims change sign):



Latent hyper-parameters α generate latent states $\boldsymbol{\theta}$ and β, which generate observable publications at time ***t***, features ***f*** and claim ***c***.



Partition of interactions

Partition interactions into positive, neutral and negative: using Wasserstein distance between naive *Beta* posteriors derived from corresponding claims.

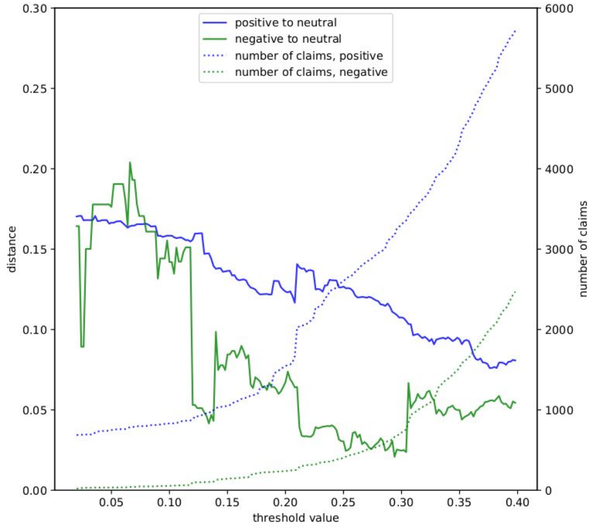
$$g_x(\mu) = Beta(a_0 + \sum_{\alpha \in \mathcal{C}_x} \sum_{i=1}^{n_\alpha} y_i^\alpha, \quad b_0 + \sum_{\alpha \in \mathcal{C}_x} \left(n_\alpha - \sum_{i=1}^{n_\alpha} y_i^\alpha \right))$$

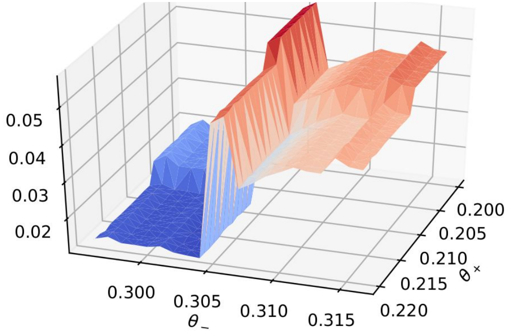
$$W(g_+, g_0) = \inf_{\gamma \in \Gamma(g_+, g_0)} \int d(x,y) d\gamma(x,y)$$

$$\theta_-^* = \arg\min_{\theta_-} \delta^L W(g_-, g_0, \theta_-, \theta_+)$$

$$\theta_+^* = \arg\min_{\theta_+} \delta^R W(g_+, g_0, \theta_-, \theta_+)$$

GeneWays	0.305	0.218
Literome	0.256	0.157





Target variables

interaction (α,β) : π_0 - interaction neutral? interaction neutrality

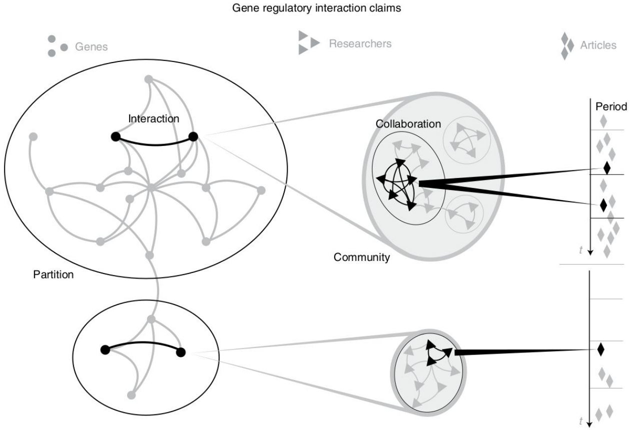
interaction (α,β) : π_+ - non-neutral interaction positive? interaction positivity

interaction (α,β), claim C_i : y_i - is this a correct claim? claim validity

Types of features

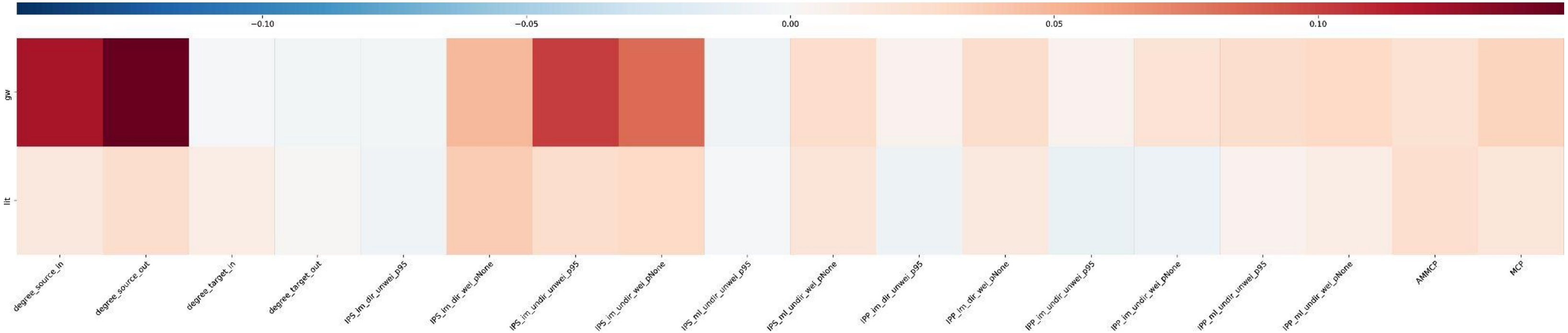
claim level, batch level, interaction level

- features are defined with respect to an time interval
- infomap used for community detection

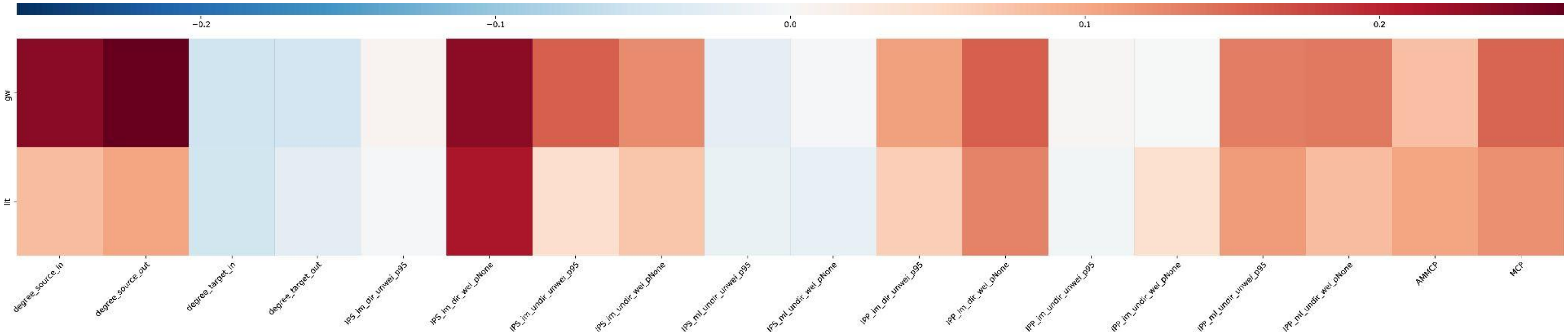


Interaction level correlations

interaction neutrality correlations

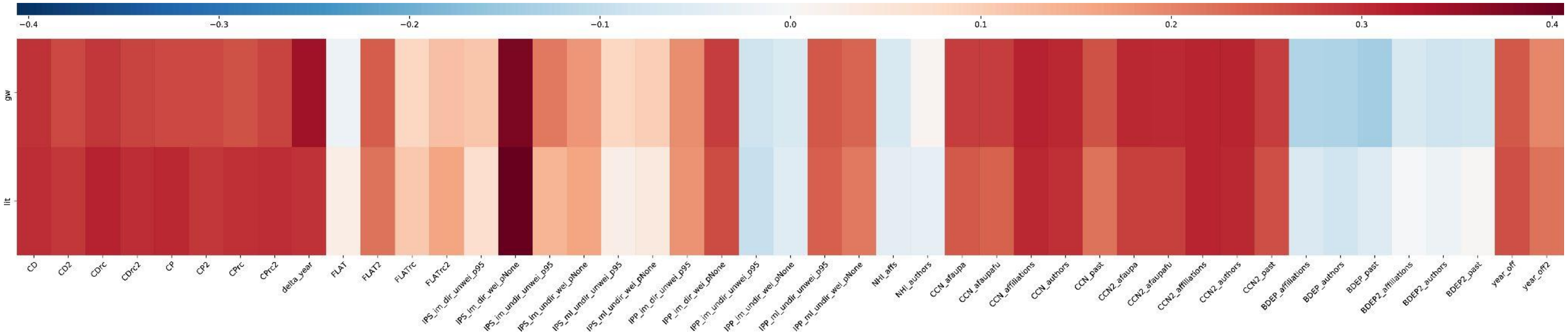


interaction positivity correlations

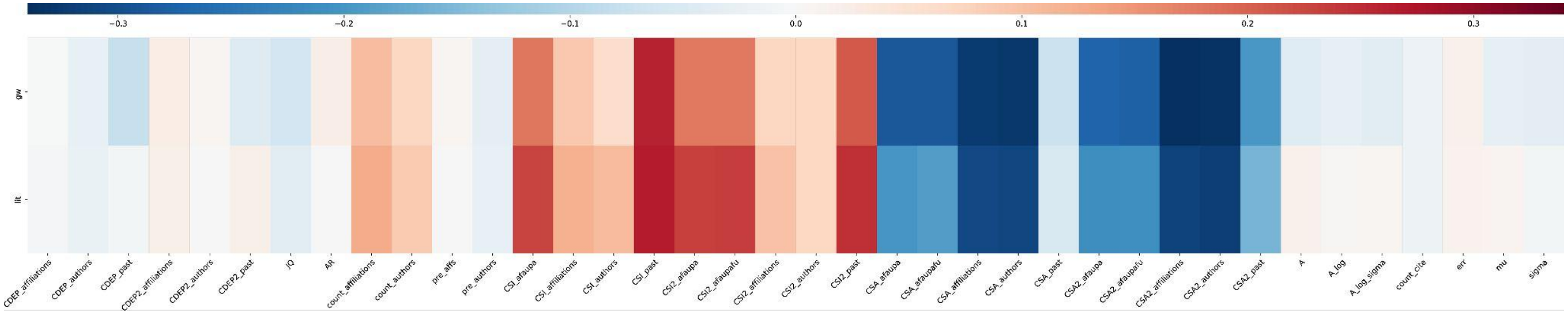


Claim validity correlations

batch level features



claim level features

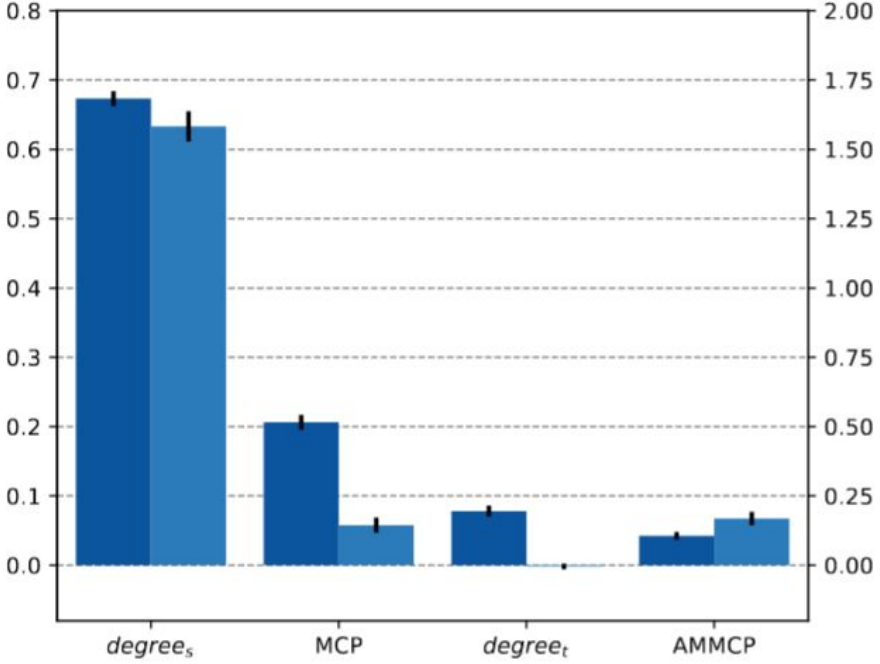


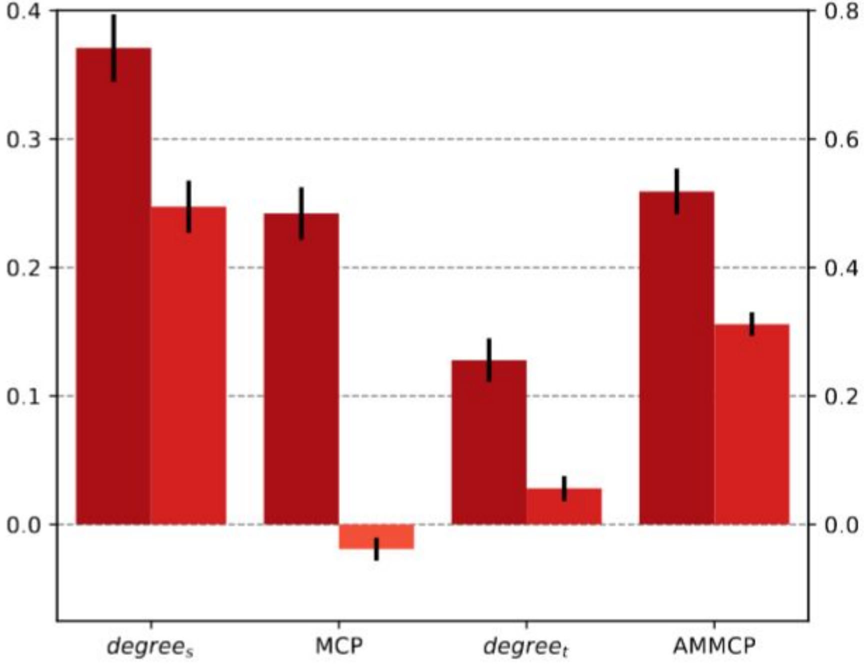
Modeling

Linear Regression and **Random Forest** methods were used, chosen for interpretability and robustness.

- 20 threefold samples of interactions. 1 out of 3 for validation : 60 training–validation pairs.
- Samples are drawn randomly per interaction (using the claim number distribution function). Claim correctness model is trained and then validated on sets containing disjoint genetic interactions.
- Tests for over-fitting revealed the regime of high variance and therefore it is desirable to use models of low complexity

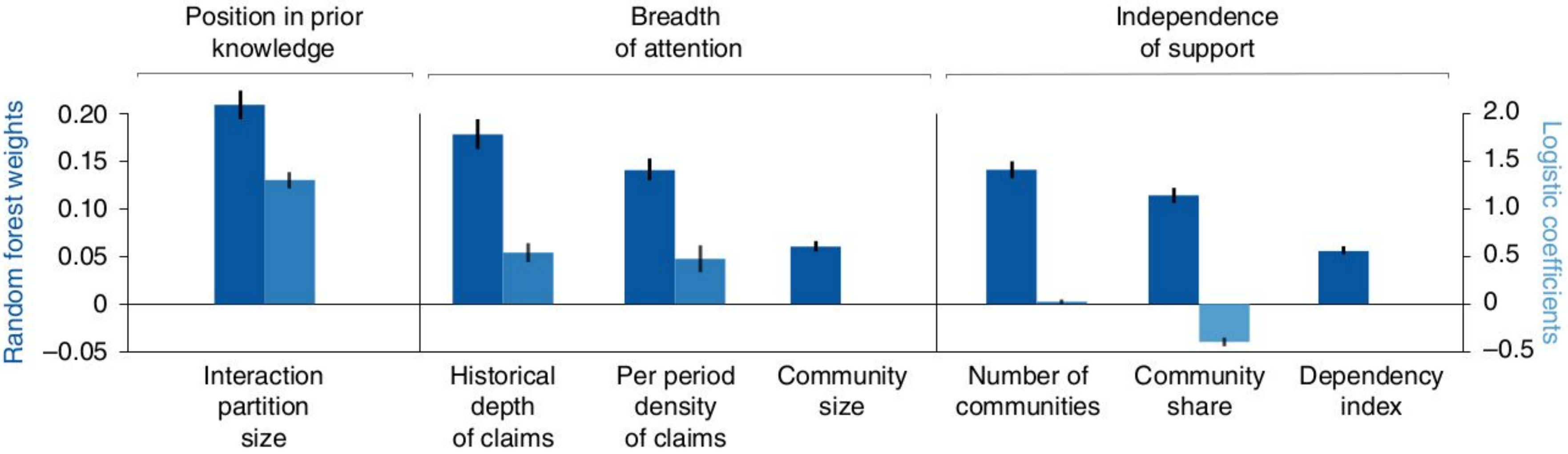
Feature importances (neutral)



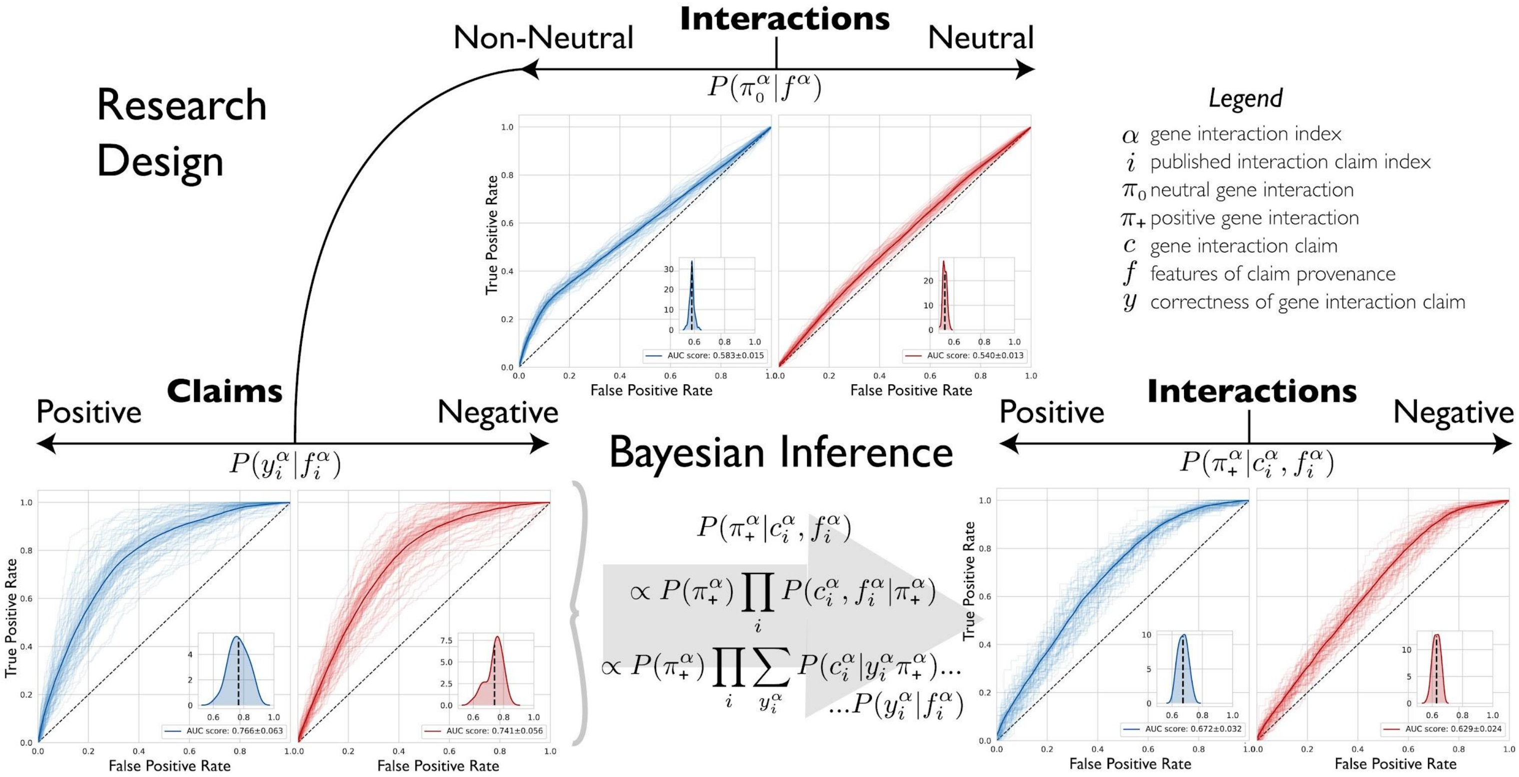


MCP - mean claim percentile
AMMCP - absolute value of the median mean claim percentile

Feature importances for claim validity



Predicting neutral interactions



$$P(\pi_+^\alpha | c_i^\alpha, f_i^\alpha)$$

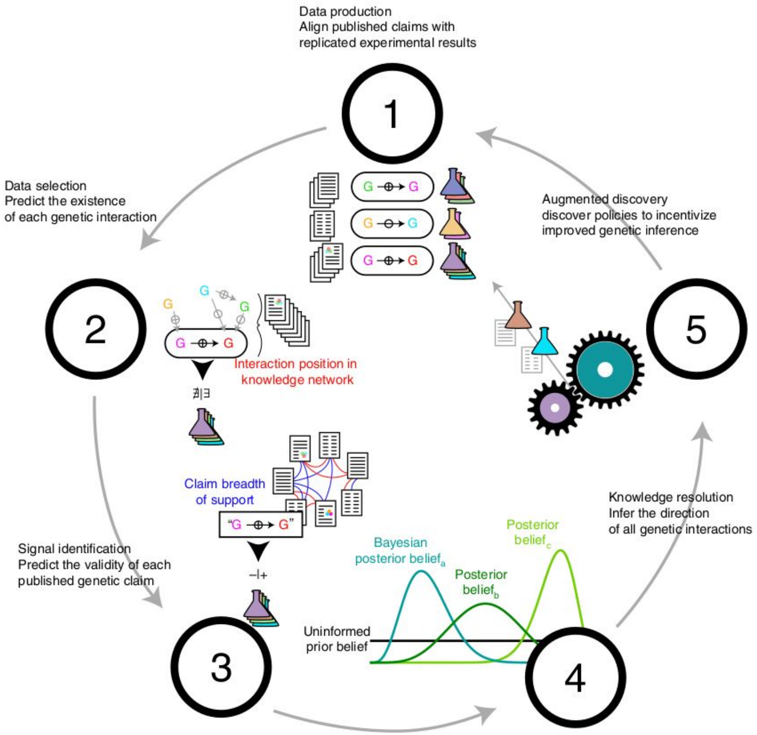
$$\propto P(\pi_+^\alpha) \prod_i P(c_i^\alpha, f_i^\alpha | \pi_+^\alpha)$$

$$\propto P(\pi_+^\alpha) \prod_i \sum_{y_i^\alpha} P(c_i^\alpha | y_i^\alpha \pi_+^\alpha) \ldots$$
$$\ldots P(y_i^\alpha | f_i^\alpha)$$

Legend

α gene interaction index
i published interaction claim index
π_0 neutral gene interaction
π_+ positive gene interaction
c gene interaction claim
f features of claim provenance
y correctness of gene interaction claim

Augmented discovery



Data production
Align published claims with
replicated experimental results

Data selection
Predict the existence
of each genetic interaction

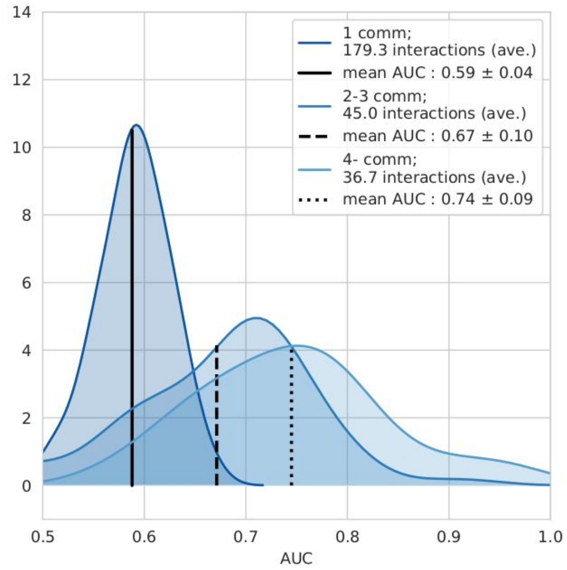
Signal identification
Predict the validity of each
published genetic claim

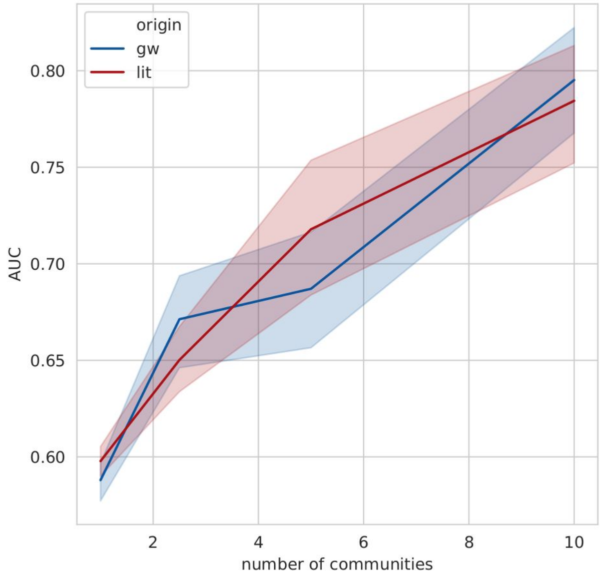
Knowledge resolution
Infer the direction
of all genetic interactions

Augmented discovery
discover policies to incentivize
improved genetic inference

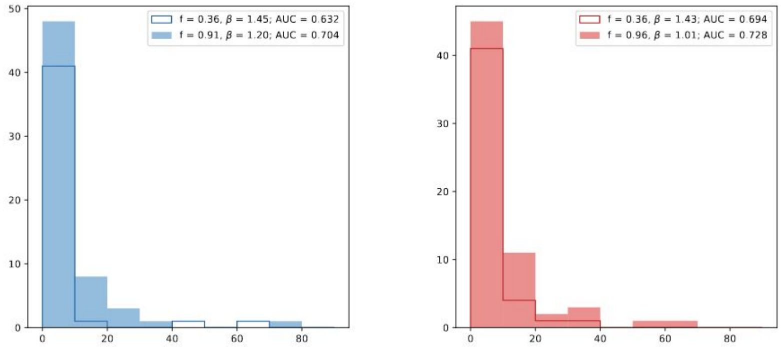
Policy A: promote independence

Selecting subsamples with more communities improves AUCs of interaction prediction model





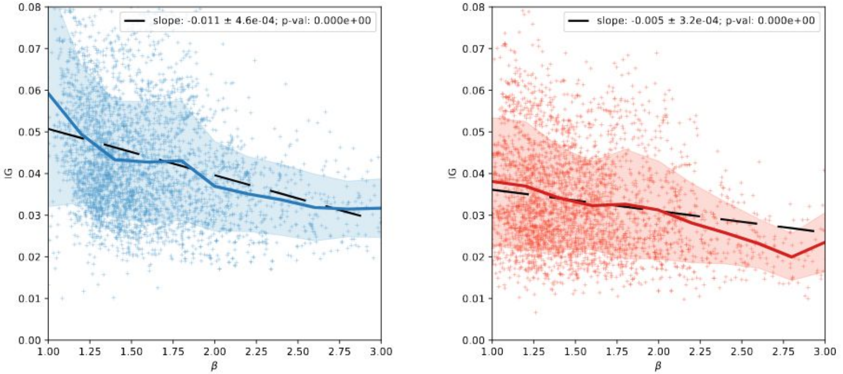
Policy B: altering the attention



$$IG = \mathrm{ent}(p^{(0)}) - \frac{1}{k}\sum_{\alpha=1}^{\alpha=k} \mathrm{ent}(p^{\alpha})$$

$$\mathrm{ent}(p^{\alpha}) = -\sum p_i^{\alpha} \log p_i^{\alpha}$$

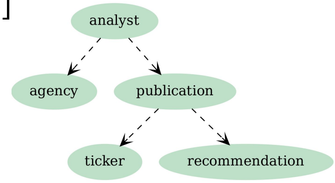
Flatter distributions result in higher information gain



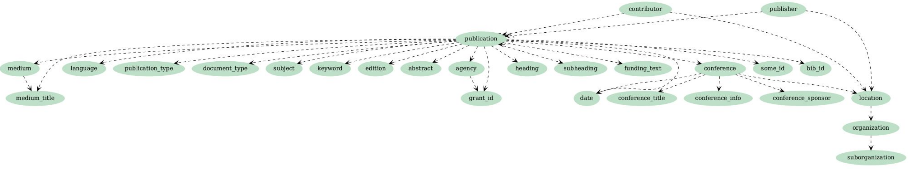
Discussion

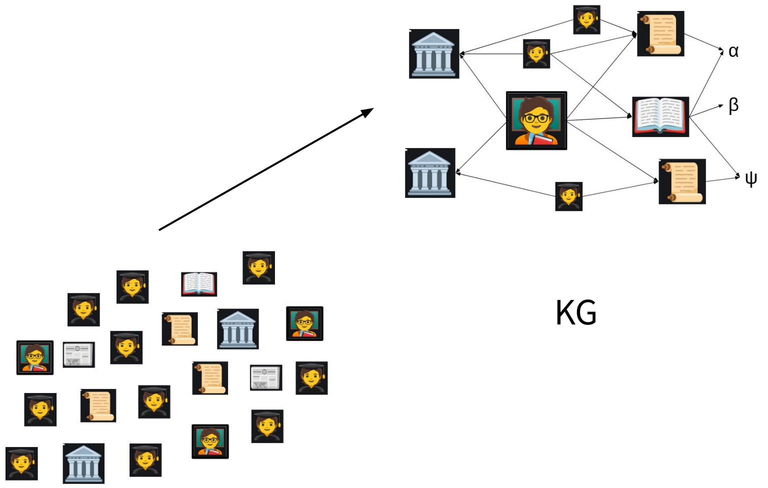
- The representation of data is simplified.
- Raw data does not contain robust signals.
- Publication biases are apparent.
- Features, thought to be important traditionally, were found to be irrelevant.
- Similar performance of **predictive** models and similar feature importances for two (almost) independent datasets. Correct model validation is very important.
- Simulations reveal that the overall knowledge can be improved by policy modifications.
- Independence measures should be tested in other domains (science, finance).

Communities of analysts [IBES]



Outlook : Data

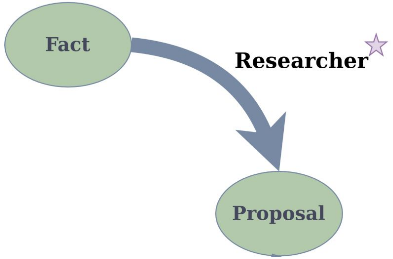




- Unification of data resources: Pubmed, WoS, OpenAlex.
- Improving Data Management: Graph Databases. ArangoDB, neo4j, TigerGraph.
- Improved Relation Extraction and Entity Linking. Use of Knowledge Graphs.
- Capture context and include context in modeling.

Better ML methods

- Embedding methods to simplify the representation of logical formulae.
- Graph Neural Networks (MPNN, GCN etc) are highly generalizable and may take advantage of network structure without explicit feature derivation, see Davies, A., *et al.* Advancing mathematics by guiding human intuition with AI. *Nature* **600**, 70–74 (2021). First examples of actual discoveries
- Reinforcement Learning on graphs is the path to automated discovery, Policy-GNN, ReWatt, Care-GNN.

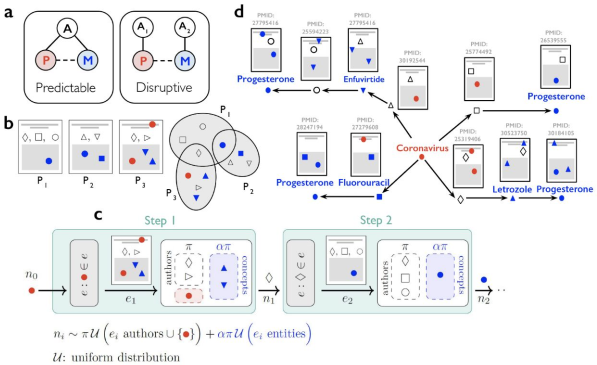


Hypothesis Generation

Accelerating science with human versus alien artificial intelligences, J.Sourati et al., arXiv:2104.05188.

Complementary artificial intelligence designed to augment human discovery, J.Sourati et al., arXiv:2207.00902.

Random walk on a hypergraph : <expert, material, property>.



Thank you!

Questions ?

Talk will be available at alexander-belikov.github.io

References

- Belikov, A.V., Rzhetsky, A. & Evans, J. Prediction of robust scientific facts from literature. *Nat Mach Intell* **4**, 445–454 (2022)
- Sourati, J., Belikov, A., & Evans, J. (2022). Data on How Science Is Made Can Make Science Better. Harvard Data Science Review, 4(2)