nature machine intelligence

ARTICLES

https://doi.org/10.1038/s42256-022-00474-8

Prediction of robust scientific facts from literature

Alexander V. Belikov[1]✉, Andrey Rzhetsky[2,3] and James Evans[1,4]✉

The growth of published science in recent years has escalated the difficulty that human and algorithmic agents face in reasoning over prior knowledge to select the next experiment. This challenge is increased by uncertainty about the reproducibility of published findings. The availability of massive digital archives, machine reading, extraction tools and automated high-throughput experiments allows us to evaluate these challenges computationally at scale and identify novel opportunities to craft policies that accelerate scientific progress. Here we demonstrate a Bayesian calculus that enables positive prediction of robust scientific claims with findings extracted from published literature, weighted by scientific, social and institutional factors demonstrated to increase replicability. Illustrated with the case of gene regulatory interactions, our approach automatically estimates and counteracts sources of bias, revealing that scientifically focused but socially and institutionally diverse research activity is most likely to replicate. This results in updated certainty about the literature, which accurately predicts robust scientific facts on which new experiments should build. Our findings allow us to identify and evaluate policy recommendations for scientific institutions that may increase robust scientific knowledge, including sponsorship of increased diversity of and independence between investigations of any particular scientific phenomenon, and diversity of scientific phenomena investigated.

Millions of scientific papers are published globally each year[1,2], which makes it difficult for scientists, research institutes, technology companies and other audiences of science to decide what to trust. The same is true for machine reasoning agents in the rising age of algorithmically driven experimentation[3–6]. More certainty about robust findings would allow better selection of the next experiment and accelerate collective discovery. Here, we ask how to predict certainty about scientific claims and identify robust facts from research publications. We also consider the implications of our predictions for how scientific institutions might be reformed to amplify robust signals from science.

Subjective bias is an inevitable reality underlying published science. As scientists undertake research and publish findings, they must consider not only accuracy but also scientific influence and their own academic survival. Scientists consider what will attract attention and inspire other scientists to build on their work in future, what journal editors and reviewers will allow[7], what patrons will fund and what promotion committees will accept? Beyond motivation, scientists and their investigations are situated in particular positions with respect to their objects of study, which defy detached and universal notions of objectivity[8] and necessarily shape their assessments. Scientists hold assumptions acquired through disciplinary education and prior experience[9], and they rationally incorporate the beliefs of others they trust—respected mentors and colleagues—into their own scientific expectations and certainty[10]. As we show in this paper, ignoring these contextual forces distorts predictions about which findings will replicate and generalize.

Above the level of scientists, the scientific system promulgates predictable biases. Competition between journals makes it easier to publish positive findings than neutral or 'negative' ones[1,2], which become underrepresented in the published record[11]. Moreover, favourable conditions for the 'wisdom of crowds' phenomenon[12], where collectives produce systematically more accurate estimates than individuals, are widely violated in science. Crowds are wise when their members have access to independent data[13] or utilize independent methods[14] to derive their answers, but they falter when engaged in centralized communication[15,16] and share prior experience, knowledge and methods[17]. By contrast, modern science is characterized by intensive and repeated collaboration[18–20], increasingly large[21,22] and distributed teams[23], star scientists[24,25], canonical citations[26–28] and expensive shared equipment[18,20].

These forces have led to widespread concerns regarding the reliability and reproducibility of findings in fields ranging from pharmacology[5,29–31] and genetics[3,4,11,32–34] to psychology[35,36] with widespread implications for the accumulation of certainty in science. Some have even feared that distortions from publication and confirmation bias could lead to the canonization of false facts[7]. This is a problem for scientists, but also for future innovation. Prior work has attempted to identify the replicability of scientific literature through simulation[7,37,38], experiments[35,39] and meta-analysis of prior results[17,40]. Psychologists have called efforts to replicate the robust essence of an experiment with an alternative research design, measurements or methods a 'conceptual replication'[41–48]. Here, we translate that approach into an algorithmic research pipeline to predict robust scientific facts on which future science can build. We demonstrate this in the context of genomic science, following a five-step procedure outlined in Fig. 1 and formalized in Methods: (1) data production, where we (1a) extract gene–gene interaction claims from the biomedical literature, (1b) consolidate results from a massively replicated high-throughput gene–gene experiment and (1c) align claims and results for conceptual replication; (2) data selection, where we predict the existence of a genetic interaction in high-throughput experiments based on its position within the network of genetic interaction claims from the literature; (3) signal identification, where we predict the validity of a published genetic claim to isolate factors that increase or decrease the likelihood of replication; (4) knowledge resolution, where we follow Bayes' rule to infer the direction of all published claims based on genetic interactions weighted by the signals of reproducibility identified in step 3 to update our scientific understanding and certainty; (5) augmented discovery, where we discover and simulate policies that would redesign scientific institutions for

[1]Knowledge Lab and Department of Sociology, University of Chicago, Chicago, IL, USA. [2]Departments of Medicine and Human Genetics, University of Chicago, Chicago, IL, USA. [3]Institute for Genomic and Systems Biology, University of Chicago, Chicago, IL, USA. [4]Santa Fe Institute, Chicago, IL, USA. ✉e-mail: belikov@uchicago.edu; jevans@uchicago.edu

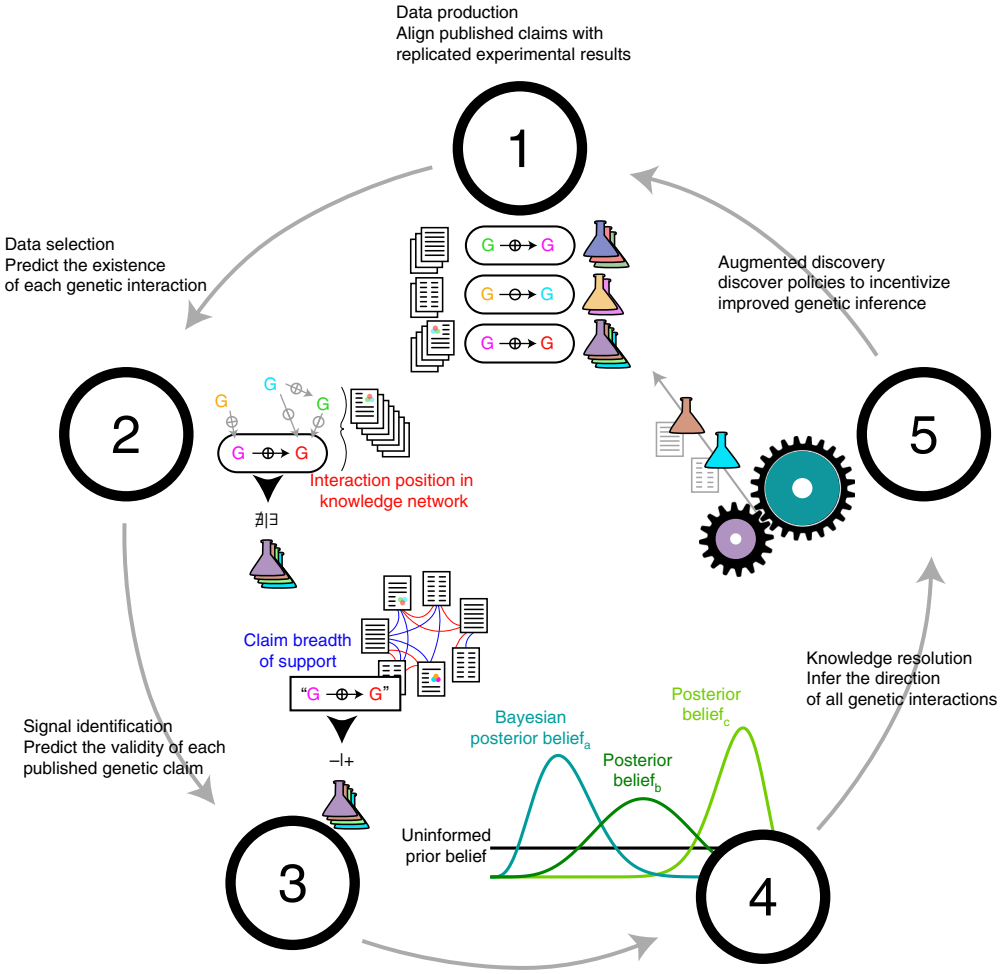


Fig. 1 | Analysis synopsis. In step 1, we prepare the data, aligning published claims and massively replicated experimental results by specific genetic interaction (+/ arrows indicate positive and negative interactions, respectively; coloured 'G's suggest different genes; stacked papers indicate the research literature mentioning each genetic interaction, aligned with the coloured flasks indicating different experiments performed on those same interactions). In step 2, we select genetic interactions for further analysis by predicting their existence based on position in the network of prior published knowledge (stacked papers represent articles on a given interaction and all nearby interactions). In step 3, we identify reproducible signals by predicting the validity of published genetic claims using both their position in prior knowledge and their breadth of support (the network of papers indicates articles all mentioning the same interaction, variously linked by shared authors and institutions (blue ties) and shared references (red ties)) for claims about interactions selected in step 2. In step 4, we use Bayes' rule to infer the direction of all genetic interactions on the basis of our evidence (genetic claims a, b and c from the literature are weighted by features predictive of reproducibility in step 3), which weakens our (posterior) confidence in some published interactions but strengthens it in others. In step 5, we propose and validate policies with data-driven simulations that could improve the publication of reproducible claims in future science (gears indicate the improved machinery of science generating new experiments and findings), suggested by the arrow linking steps 5 and 1.

accelerated advance. Because of the complex interconnection between each of these research steps, we reference them by number throughout the manuscript.

Automated validation pipelines

To predict robust facts from published science, we must first extract claims from published literature (step 1a). Here, we deploy two algorithmic approaches built on distinct architectures: GeneWays[7] and Literome[8]. The GeneWays (a portmanteau of 'genetics' and 'pathways') system semantically parses the biomedical literature. It identifies biological substances and processes (nouns and verbs) then parses them with a context-free grammar tuned to the sublanguage of biomedicine[49] that extracts relations yielding a graph with directed links from source to target. Proteins may bind, activate, inhibit or unleash other transformations (acetylate, methylate or phosphorylate) upon their targets[50]. We simplify these interactions into positive (for example, 'activate', 'enhance', 'increase', 'promote', 'stimulate', '[over]produce', 'upregulate' and so on) and negative (for example, 'inactivate', 'depress', 'limit', 'inhibit', 'constrain', 'hinder', 'downregulate' and so on). Literome inverts the GeneWays pipeline, parsing articles into dependent clauses[51], then extracts biological entities, including genes and proteins. From co-presence within parsed phrases, Literome identifies directed relationships and filters these for genes from the GENIA dataset[52]. 'Gene' in this context is used as a shorthand for 'gene or gene product'.

For both datasets, we limit the examination of claims to those extracted from article abstracts, to increase the likelihood that they were not merely reiterated from referenced papers, but empirically demonstrated within the associated article. The GeneWays and Literome algorithms were run on overlapping collections of articles present in MEDLINE (197k for GeneWays; 220k for Literome). The precision of GeneWays and Literome is evaluated to be 95%[44] and 25%[51], and their percentage of positive interactions as 77% and 96%, respectively. In summary, GeneWays is more accurate while Literome has wider scope: GeneWays and Literome yield directed genetic graphs with 5,141 genes involved in 23,405 interactions and

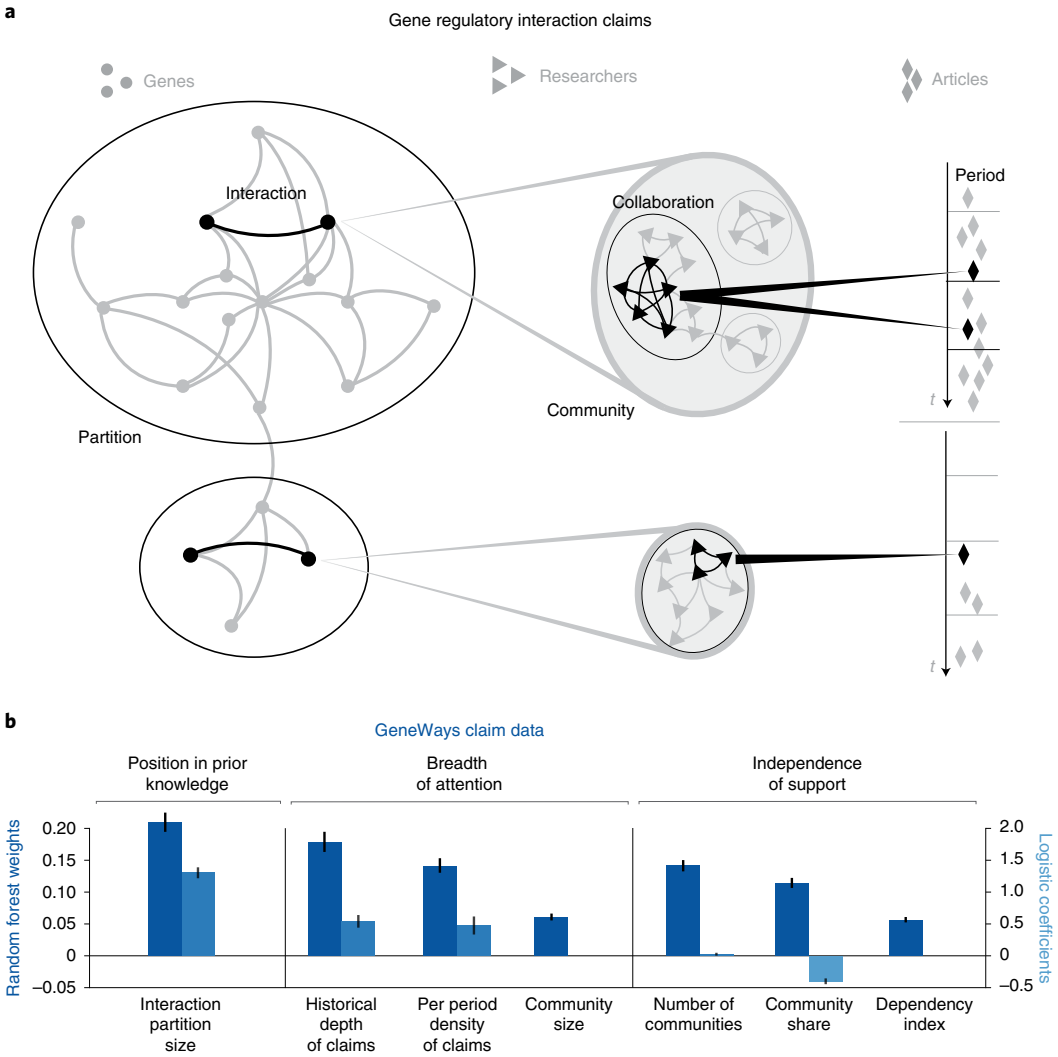


Fig. 2 | Feature visualization and estimates from claim-level prediction models. a, Illustrated relationship between genes engaged in regulatory interactions, the communities that research them and the articles in which this research is published. Interactions cluster into partitions, researchers cluster into communities and author teams publish articles within fixed periods. Together, these structures are used to assess the position of a claim within preexisting knowledge, the breadth of attention to a claim and the independence of support for that claim. **b**, Gini importance for features in the random forest models (left scale, darker colours) and coefficients from the logistic regression models (right scale, lighter colours) for GeneWays with 95% CI for the mean of the estimate (vertical bars). See Supplementary Information for details about how specific operationalizations of each of these variables were selected as model features. See Extended Data Fig. 9 for comparable coefficients for Literome.

10,703 genes engaged in 144,172 interactions, respectively. This yields an overlap of 4,516 genes, but only 6,516 overlapping interactions. We perform all of our analyses on both independently derived datasets, with GeneWays featured in the main figures and Literome in the Supplementary Information.

Next, we derive specific measures associated with how a claim fits into preexisting knowledge about genetic interactions and its breadth of support from article data and metadata, as illustrated in Fig. 2 and detailed in the Supplementary Information and Extended Data Fig. 1. We measure each genetic regulatory interaction's position in preexisting knowledge with (1) the centrality of its gene source and target in the network of publication claims, weighted by frequency of publication, as well as (2) the interaction community size of which it is a member, derived from the InfoMap algorithm[53] run on that same network. The centrality measure provides an estimate of the interactions' position within the global structure of knowledge (for example, *P53*, an important gene controlling cell division and death, is highly central), and the size of the gene community in which it is embedded represents its position within the local structure of knowledge. We measure the breadth of each interaction's support with (1) the density of articles published on it each year, (2) the number of years over which it has been investigated, (3) the number of research communities investigating it (derived from the same InfoMap algorithm used above, here run on the network of papers making each interaction claim, linked by coauthors, affiliations and shared references), (4) the absolute size of each research community and (5) the size of each community relative to others that investigate the same claim. We also derive measures capturing the reputation of the institution hosting the research and the journal publishing the claim, in addition to a number of related variables (detailed in Methods and Supplementary Information).

As the culmination of step 1 in the research, we aligned these findings with the massively high-throughput National Institutes of Health (NIH) Library of Integrated Network-Based Cellular Signatures (LINCS) L1000 experiment at the level of each genetic regulatory interaction. LINCS L1000 perturbed 77 distinct cell lines

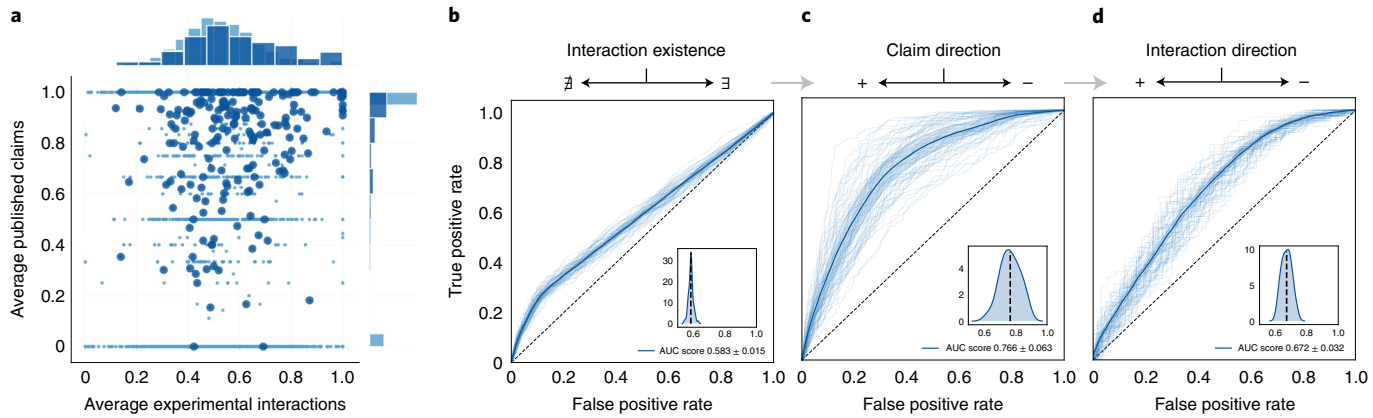


Fig. 3 | Positivity bias in published effects and prediction results. a, Joint plot of the mean experimental interaction strength (*x* axis) and the mean value of the published claim (*y* axis) for each genetic interaction. Darker hues of blue (and also greater marker size) correspond to interactions in GeneWays with ten or more claims per interaction; for lighter hues (and also smaller marker size), the cutoff is absent, representing the complete distribution (see Extended Data Fig. 10a for a comparable distribution for Literome). **b**, We first predicted the non-existence (∄) or existence (∃) of each published gene–gene regulatory interaction (GeneWays). **c**, Then, if the interaction was deemed existent (∃), we predicted whether each claim (of positivity or negativity) from literature was correct. **d**, Finally, using Bayesian inference, we estimated the sign (positive versus negative) of all genetic regulatory interactions. Mean ROCs (thick lines) are complemented by 95% CI contours, with fainter individual lines corresponding to ROC curves for 60 models corresponding to different training–validation samples. See Extended Data Fig. 10b–d for comparable AUCs for Literome.

with several different perturbation types, such as complementary DNA for overexpression of wild-type gene, for 5.8k genes[54]. Gene overexpression is a technique that utilizes expression vectors to force high levels of a gene's coding sequence. The resultant effect is high steady-state messenger RNA levels and high steady-state protein levels. This enables a vast gene–gene causal experiment: if one gene product is increased and consistently leads to an increase or reduction in another gene across cell lines, perturbagens, dosage and time, this suggests that the overexpressed gene has a consistent and likely causal association with the other. In step 1b of the research, we computed the mean *z* score across experiments for genetic regulatory interactions and transformed it to (0, 1) with the normal cumulative density function, denoting this as the average experimental strength of the interaction. In step 1c, we merged the literature-based claims (step 1a) and massively replicated high-throughput experimental findings (step 1b).

Bayesian certainty from science

We predict the scientific certainty of claims and when an interaction is neutral (non-existent), positive or negative by using a Bayesian calculus built atop established statistical and machine-learning methods (steps 2–4). These models incorporate features capturing position within preexisting knowledge and breadth of prior support, as described above. Some features occur at the level of the genetic regulatory interaction (for example, the position of source and target genes in the network of prior knowledge or the number of research communities publishing on the interaction). Others occur at the level of the published claim (for example, the size of the research community publishing a paper about the interaction). All features vary with time. We used these features to predict the robust aggregate results of LINCS L1000 experiments pertaining to the same source and target gene across distinct trials, dosages, tissues and durations.

Claims in the biomedical literature tend to be either positive or negative, with a strong positive bias. The distribution of experimental interactions does not share this bias, normally distributed and varying smoothly from negative to positive, peaked and centred at 0.5, indicating a 'neutral', inconsistent or non-existent interaction between those genes. The contrast between published findings and experimental data suggests the extent of the 'file drawer problem' in science where scientists euphemistically 'file' but do not publish negative or inconclusive results[55,56]. We note that the correlation between published and experimental results increases as we consider more popular interactions (Extended Data Fig. 2). These observations culminated in step 2 of the research (Fig. 1), binning experimental interactions into three categories: neutral, positive and negative, then predicting non-neutral interactions to select them for further analysis based on their position within the network of prior knowledge (Supplementary Information and Extended Data Fig. 3).

In step 3 of the research (Fig. 1), we built models to predict whether positive and negative claims correctly align with positive and negative experimental interactions based on a collection of features revealing both the position of claimed interactions within the network of prior knowledge and the depth and breadth of support associated with each claim. We separately built logistic regression and random forest models to estimate the influence of each feature on the neutrality of interactions (step 2) and the accuracy of claims (step 3). Logistic regressions provide us with interpretable directional estimates, but they assume that the log odds between positive and negative classes is a linear function of features. Random forests predict more variation and provide us with estimates of each feature's importance, but reduce interpretability by allowing nonlinear feature interactions. The most important features regarding how a claim fits into the fabric of prior knowledge and its breadth of prior support are defined in Fig. 2. For others, see the Supplementary Information and Extended Data Figs. 1 and 4. Because empirically neutral interactions are less likely to be mentioned in literature and never represented as neutral or non-existent (Fig. 3a), models predicting whether an interaction is non-neutral (step 2) cannot contain features associated with depth and breadth of support, but these are critical for models predicting the accuracy of published positive and negative claims (step 3). Features associated with the position of neutral claims within the network of prior knowledge were included in both models of interaction neutrality and published claim accuracy (steps 2 and 3).

Step 4 of the research (Fig. 1) involved a Bayesian calculus wherein we used estimates from the model of claim accuracy in step 3 to infer the direction of each genetic regulatory interaction using Bayes' formula, derived under the assumption of conditional independence between claims, and also independence between claim

correctness and interaction positivity (Methods). This allowed us to update what began as an uninformed, uniform prior about the direction of each non-neutral genetic interaction with the number of published claims, weighted by features demonstrated to predict reproducibility in step 3.

Predicting claim accuracy and interactions

We evaluated out-of-sample genetic predictions using receiver operator curves (ROCs). The distributions of the corresponding area under the curve (AUC) are presented in Fig. 3. Our model predicting whether a genetic regulatory interaction is non-neutral (∃), assessed across the wide range of LINCS experiments (step 2) and based on its position in the network of published interactions, betrays the difficulty of that task (with an average AUC of 0.58 from random forest models for GeneWays on 6.8k interactions and 0.54 for Literome on 25.4k interactions). The most important feature for identifying non-neutral interactions is notable: the degree of the source gene in the network of prior published knowledge. The more central the source gene, the more other genes it controls (for example, *P53* is central in that it regulates cascades of other genes in the complex processes of cell division and death). This finding suggests simply that in genetic discovery, the 'rich get richer'—claims about the influence of genes already known to be influential are more plausible than claims about the influence of peripheral genes with no prior signs of influence.

Our claim accuracy model predicting whether a published claim correctly identifies the direction of a genetic regulatory interaction, conditional on the interaction being non-neutral (step 3), is much more powerful (with an AUC of 0.77 for GeneWays on 580 interactions and 0.74 for Literome on 1,090 interactions), being strongly influenced by both the position of the claim within prior knowledge and its breadth of support (Fig. 2 and Extended Data Fig. 5). The feature manifesting most predictive power is the size of the partition of published genetic effects that surround the interaction in question. Its positive influence suggests that the structure and direction of nearby interactions may guide researchers to the correct conclusion[10]. When claims about genetic influence pathways are embedded within large, dense thickets of claims about related interactions, the structure and direction of those interactions logically and physically constrains the direction of the focal claim[10], which likely helps researchers robustly triangulate the correct claim. The presence of other researchers asking nearby questions may also socially discipline researchers to share their most robust results, as their work will receive scrutiny by those researching nearby.

The next most important class of influences are historical depth and social and institutional breadth of support. By contrast, empirical incorrectness is associated with higher relative community size and our indices tracing author, institutional and prior knowledge dependencies. Greater relative community size indicates that the majority of scientific activity in support of a genetic regulatory claim comes from just one or a few researcher communities, which weakens the independence of that support. Moreover, when the dependency index is high, support for a given genetic regulatory claim draws on overlapping authors, institutions and citations. This inflates the appearance of strong support for a claim (for example, publication in many research papers), without the independence required to justify that support (for example, papers are by the same authors and institutions, referencing the same prior work). This accords with recent research that shows how papers forwarding the same claim but sharing authors, institutions, methods and references decrease replicability by outsiders[17] and characterizations of dense scientific communities as echo chambers that drive out diverse perspectives and reproduce fragile findings[57]. In summary, a lack of social and theoretical independence between claims should reduce our confidence in them. Our analysis revealed that some features widely considered to be strong signs of support were in fact red herrings. The reputation of the journal publishing a claim (for example, *Science*) and the status of the institution hosting the underlying research (for example, Harvard) had no impact on our prediction and should not elevate our confidence in high-profile results.

Using Bayesian inference (step 4), we applied these estimates to infer the direction of any given genetic interaction. We begin with a uniform prior that assumes nothing about the genetic regulatory interaction in advance. Then, we update that prior based on evidence from the number of published claims about the interaction, weighted by features shown in the prior model (step 3) to be predictive of reproducible research. Our out-of-sample predictions demonstrate substantially greater signal than random regarding the robust direction of a genetic interaction (with an AUC of 0.67 for GeneWays and 0.63 for Literome). These models are less predictive than our models predicting accurate research claims because of inequality in research attention, collectively focused on a few, popular interactions. While scientific certainty about any particular interaction might be satisfied with a moderate number of replications, the inequality of research attention and activity are more likely to furnish the 100th replication of a popular claim than the 2nd of an unpopular one, despite the drop in information this entails for science as a system[58]. All of our findings are robust to analysis of only those claims published in elite journals, and authored by scientists at elite universities (Supplementary Table 3).

Policies to optimize scientific certainty

Our computational pipeline not only enables a Bayesian update of certainty regarding the robustness of scientific literature (step 4). It also suggests policies to increase collective certainty across claims. In step 5 of the research, we design data-driven statistical experiments that simulate the impact of two policies on the accuracy of science's collective certainty. The first experiment manipulates the distribution of independent communities examining a research claim, and the second manipulates the distribution of attention traced by claims across interactions. For both, we examine their influence on the correctness of all scientific inferences that can be made about genetic regulatory interaction—the robustness of our collective understanding about the genetic world. These policies could be implemented science-wide by research funders such as the US NIH. For the NIH to influence the number of communities studying a specific topic, it could simply prefer to fund each new investigation on the basis of, in part, its social, institutional and intellectual independence. For the NIH to broaden the distribution of claims across interactions, it could prefer research on new topics over old. We demonstrate predictively that each policy would increase the correct identification of the direction of genetic regulatory interactions as measured by the AUC of our model. Both policies enacted together would enable science to know more about more on the basis of the same resources.

For the first statistical experiment, we divide interactions from the test sample into disjoint groups by the number of author communities that publish on them. For the subset of positive and negative interactions we (1) split the dataset into training and testing samples, (2) estimate the model of claim correctness, then using Bayesian inference (3) predict model certainty for groups of interactions having 1, 2, 3, 4 or more communities in the test sample and (4) repeat this procedure 30 times to estimate the metric distributions. Figure 4a,b shows that a greater number of communities has a profound positive effect on the distributions of AUCs for interaction positivity. The more communities studying any particular interaction, the better we can infer genetic facts from the resulting corpus of research.

For the second statistical experiment, we artificially shift the distribution of published claim numbers by sampling interactions according to the number of times each is published. Specifically,

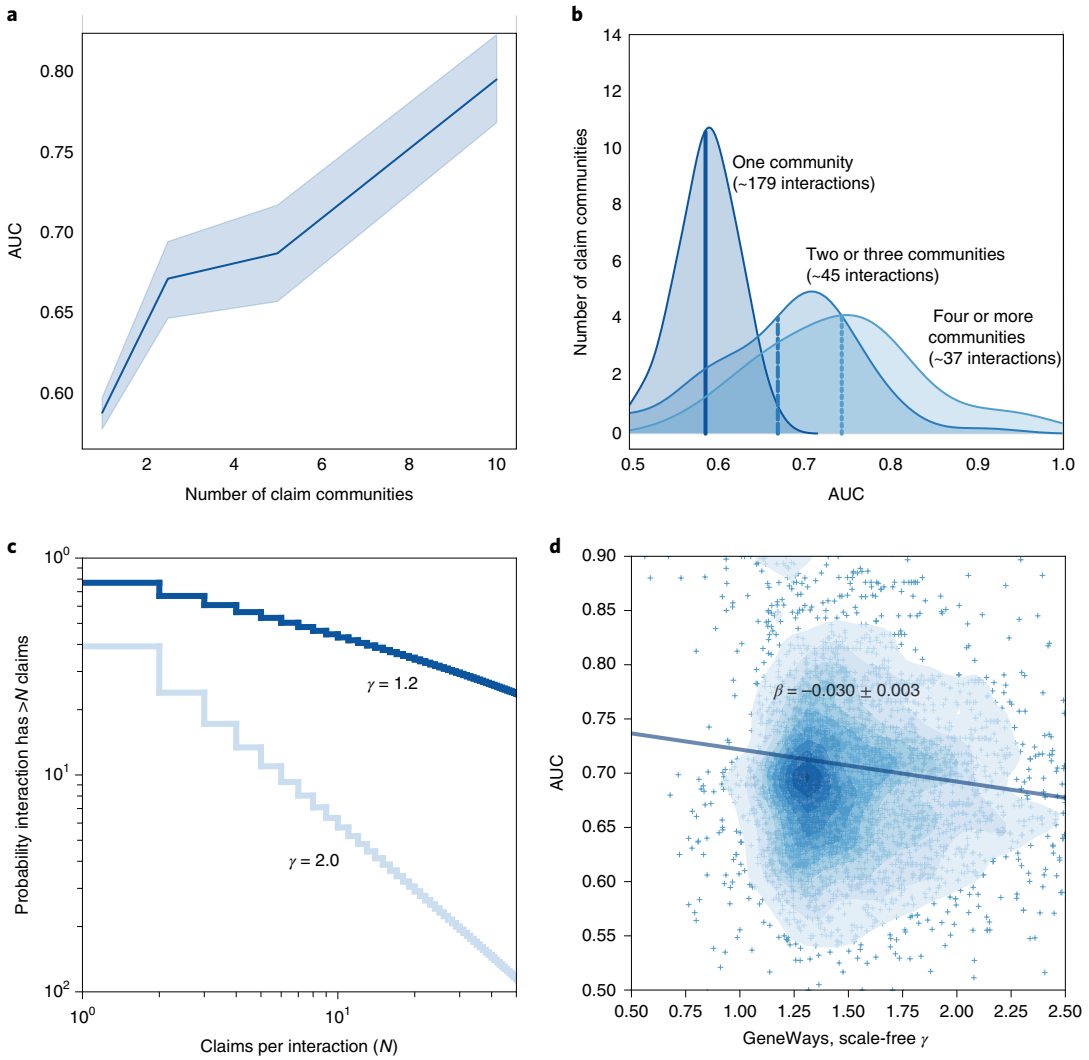


Fig. 4 | Science policy experiments revealing the relationship between community independence, collective attention and certainty about genetic regulatory interactions. a, Relationship between the number of communities studying a particular genetic regulatory interaction and the average AUC of out-of-sample predictions for positive interactions, with 95% CI (band). **b**, Distribution of AUC values for GeneWays extracted interactions with one, two or three, and four or more communities. **c**, Two synthetic examples of claim number distributions that can be approximated by power laws (that is, the probability of observing an interaction in the corpus is proportional to the number of claims about this interaction raised to the power of some exponent $\gamma > 1$, such that lower γ values correspond to flatter distributions). Here we render these as survival functions, which formally complement the cumulative distribution functions. For flatter claim number distributions, scientific attention spreads more widely across genetic regulatory interactions, which corresponds to significantly higher certainties (AUCs) of accurate interactions (Extended Data Fig. 7). **d**, Relationship between the shape of the distribution of number of claims per interaction (γ) on the AUC of out-of-sample predictions for positive interactions (plus signs are γ-AUC coordinates; darker contours represent higher densities). The shape is quantified by the slope (β) of the claim number per interaction distribution on the AUC for predicting all genetic regulatory interactions using GeneWays data. See Extended Data Fig. 9c for comparable simulations for Literome.

we (1) fix time t for all interactions in the sample and consider only claims published before t, (2) prepopulate the sample with ~20% of claims in chronological order then (3) repopulate the sample with claims (and interactions), year by year, until each interaction contains the complete history of observed claims. Figure 4c presents two synthetic examples of claim number distribution (see also Extended Data Fig. 6). The claim number distribution can be approximated by a power law, with the probability distribution function proportional to the number of interactions having a given number of claims about them raised to the power of an exponent (γ), where lower values correspond to flatter distributions. We demonstrate in Fig. 4d that flatter claim number distributions, where scientific attention is spread more widely across genetic regulatory interactions, correspond to significantly higher AUCs predicting accurate interactions (increases of 0.03 ± 0.003 for GeneWays and 0.007 ± 0.002 for Literome per unit of exponent of the claim number distribution γ).

Amplifying certainty in science

The deluge of published scientific information available to twenty-first century scientists has overwhelmed their capacity to account for all available signals from science in their efforts to innovate atop established facts. This project represents the first automated, machine-driven pipeline of which we are aware that reads scientific research papers, extracts information about scientific claims, aligns them with high-throughput experiments and provides a Bayesian update of that knowledge. This is precisely what individual scientific experts do when they critically evaluate

the literature, on the basis of deep personal understanding of the dynamics and reputations within their specific field, here scaled by machine to many areas of biomedicine beyond the scope possible for any single scientific reasoner. Moreover, our findings suggest that some common human heuristics for quality, such as trust in high-profile journals and affiliations, may be unwarranted.

In contrast to prior efforts that only detect overconfidence in science, describing the level at which science fails to replicate, here we also correct it, accounting for relevant factors to predict replicability and update uncertainty about particular facts. Our models predict replication based on two, separate collections of publications, read by different algorithms, demonstrating deep consistency regarding what predicts replication and what might be altered by science policy-makers to improve the state of scientific knowledge. These findings reveal an essential tension in the undertaking of science. Robust findings are predicted by many investigators devoted to studying topics central in the network of scientific claims and embedded within dense areas of investigation. Nevertheless, greater independence between investigators, communities, institutions and prior knowledge dramatically increases claim robustness. This leads to a paradox in developing a policy for the optimal scientific agent. When scientists flock together by studying the same phenomenon, it increases our collective understanding. When they flock together in their approach and collaborations, it decreases our collective understanding, increasing the illusion that we know more than we do. Our policy experiments suggest that if scientific institutions take this tension seriously, they could dramatically increase the robustness of collective understanding and accelerate innovation. When an important scientific problem merits attention, sponsoring diverse parallel approaches will pay dividends in robust replicable understanding.

Our approach has several natural limitations. Despite our replication of all findings with two different samples of research papers and claim extraction algorithms, our study nevertheless only explores claims about pairwise genetic regulatory interactions. Moreover, both information extraction algorithms (GeneWays and Literome) had limitations described above, but they balanced one another in terms of precision and recall. Finally, linkages across datasets that facilitate our mapping of research claims to LINCS L1000 experimental results necessarily excluded interactions from the literature not present in the experiment, and vice versa. Our analysis of observational data makes it impossible for us to make strong causal claims about the impact of optimal policies or reforming scientific institutions according to the suggestive patterns we document. Notwithstanding these limitations, all of which would have decreased the signal we might have expected to isolate from scientific literature, our system was able to predict the likelihood of conceptual replication far above random and identify consistent patterns of focus and independence that, if enhanced, could substantially increase the robustness of science. Our approach can immediately be factored into prediction engines that efficiently drive high-throughput research to accelerate robust replicable science.

Methods

Analysis synopsis. Here, we introduce terminology that allows us to formalize the structure of our analysis and outline the Bayesian calculus we use to update certainty estimates for gene–gene interactions.

Let α index genetic interactions (g_s, g_t), where g_s denotes the source gene that undergoes manipulation and g_t denotes the target gene whose expression is modified (or not). We denote by π_j^α the strength of the interaction α inferred from experiment j in the LINCS L1000 dataset (real valued) and by c_i^α the value of the claim α extracted from publication i drawn from academic literature. In what follows, we treat LINCS L1000 as ground truth in the framework of supervised learning. While we understand that the high-throughput LINCS L1000 experiment does not represent truth, its recency, transparency (through external audit), accuracy and replication strategy and the influence of its findings across biomedicine suggest that it represents our best proxy.

As stated in the narrative of the paper, our objectives in this analysis are, first, to decide whether a given claim c_i^α is correct, second, to infer the direction of the genetic regulatory interaction π_j^α from knowledge about the entire literature of claims on the topic $\{c_i^\alpha\}$ (data and metadata) and, finally, to suggest improved exploration strategies to guide investigations for improved collective knowledge regarding genetic interaction strength and direction.

Our exploration proceeded in the following steps, summarized in Fig. 1 and the narrative of the paper. We:

(1) Produced then aligned publication and experimental data, by (1a) aggregating claims from the GeneWays and Literome algorithms for each publication i and per interaction α, then projected them onto Boolean-valued c_i^α, which equals 1 for a positive interaction and 0 for a negative one (neutral interactions are not published). The next step involved (1b) aggregating data from a series of experiments j in LINCS L1000 with regards to interaction α into a single value π^α (Extended Data Fig. 3). This allowed us to (1c) align the real-valued interaction strength π^α (step 1b) and the c_i^α from the literature (step 1a).

(2) Selected relevant data for modelling by first partitioning interactions into three classes: neutral, positive and negative, then introducing Boolean-valued variables π_0^α, which equals 1 for neutral interactions and 0 otherwise, and π_+^α, which equals 1 for positive interactions and 0 for negative. Then, we trained a model to predict π_0^α, whether an interaction α exists ($\exists$, that is, non-neutral, has a z score over expression above or below certain thresholds) or does not exist ($\nexists$, that is, neutral) in LINCS L1000 experiments using features derived from the network of interlocking genetic regulatory claims across publications.

(3) Identified the correctness of published claims by training models to predict y_i^α (where $y_i^\alpha = 1 - |c_i^\alpha - \pi_+^\alpha|$, the agreement between published claims and high-throughput experiments) using publication features $P(y_i^\alpha|f_i)$. Such models are validated out of sample. Publication features included the position of the claim within other published genetic interactions, as above, but also the depth and breadth of prior support for the claim. Note that we do this only for interactions deemed positive or negative, that is, that exist ($\exists$) and are non-neutral, in step 2.

(4) Inferred genetic interaction strength π_+^α with our Bayesian calculus. We used Bayes' formula and results from the aforementioned model $P(y_i^\alpha|f_i)$ in step 3:

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

(5) In language, we updated our understanding of the strength for a given genetic regulatory interaction by taking into account features associated with more and less replicable claims as established with findings from the LINCS L1000 experiment and validated out of sample (step 3). We began with an uninformed, uniform prior about the interaction, then updated it based on claims from literature weighted by publication-level features demonstrated in step 3 to predict reproducibility.

(6) Proposed and validated strategies for augmented discovery through data-driven simulations of modifications to knowledge production that optimize the AUC for our models $P(\pi_+^\alpha|\{(c_i^\alpha, f_i^\alpha)\})$, which proxies for our overall certainty about genetic reality. The two policies we investigated, on the basis of findings from step 3, increase the number of (1) independent author communities studying a genetic interaction and (2) different genetic interactions studied.

Details regarding the steps in this process are presented below.

Information extraction algorithms (step 1a). *GeneWays.* This algorithm and associated database of automatically extracted claims[50,59] contains approximately 496k unique claims (after aggregating them so that there is a single claim per interaction per publication) and approximately 313k unique interactions (defined as a triplet including source gene, target gene and action, where the action is a verb that takes values including 'bind', 'interact', 'induce', 'associate', 'regulate' and so on) expressed in approximately 197k publications from MEDLINE. Approximately 32% of claims were extracted from abstracts. We found that claims in publications could either result from independent original research or simply reference a finding from a cited publication. The former were much more likely to be mentioned in the abstract, so in our research we considered only claims extracted from abstracts. This operation leaves us with ~172k unique publication claims and ~130k unique interactions from the abstracts of approximately ~109k unique publications.

A typical record in the GeneWays database has the form '**abg** *prevents* **tert**'. To simplify the representation of interactions, we identify all such verbs that can be interpreted as positive or negative directional actions. As positive, we encode 'activate', 'actuate', 'cause', 'control', 'direct', 'enhance', 'facilitate', 'force', 'increase', 'induce', 'lead', 'overproduce', 'promote', 'provoke', 'stimulate', 'transactivate', 'trigger', 'regulate', 'produce' and 'upregulate'. As negative, we encode 'constrain', 'degrade', 'destroy', 'downregulate', 'hinder', 'inactivate', 'inhibit', 'interrupt', 'limit', 'reduce', 'repress', 'shut' and 'suppress'. After projecting the interactions to positive or negative, we are left with ~36k unique interactions and ~68.6k unique claims from ~51k unique publications from PubMed.

For each attribute, GeneWays contains a flag indicating whether the claim is negative, where ~4% of claims are negative. According to logic, the negation of 'a increases b' is the union of both 'a decreases b' and 'a does not affect b'. Non-interactions are never recorded, so we assume that a positive interaction is the negation of a negative interaction, and vice versa. If we encounter claims with respect to the same interaction extracted from the same paper that negate one another, we discard them. We retain claims from publications present in our version of MEDLINE from 3k journals that we could identify using an available copy of the Web of Science database.

The final iteration has 23k unique interactions and 44k unique claims from 33k unique publications.

Literome. Literome[60] contains 144k unique interactions and 259k unique claims from 220k unique publications extracted from MEDLINE abstracts by means of distant supervision via Markov logic[51]. We only consider claims extracted from the abstracts and note that Literome has a strong bias towards positive interactions (~98%). For the set of final models described in the articles, we exclude claims with respect to gene *TP53* (Entrez id 7157) acting on *CDKN1A* (Entrez id 1026) because the 150 extracted claims on that interaction were all deemed incorrect or ambiguous as evaluated by a biomedical expert.

Genetic dataset from LINCS L1000 (step 1b). We use LINCS, which was compiled using the Luminex bead technology called L1000, as the ground truth with respect to gene–gene interactions derived within the same context[54]. The experimental technique of LINCS L1000 is based on tracking gene expression, the procedure by which information from genes chemically perturbed in the experiment causes the synthesis of functional gene products, such as proteins, resulting in an altered cellular phenotype. We use the GSE92742 level 5 version of LINCS L1000. The level 5 dataset contains signatures from aggregated replicates. The experiments are performed on 77 cell lines, using various perturbation types, durations and dosages. Multiple experiments are performed per combination of cell line, perturbation type, duration and dosage. The result of an experiment is a z score that quantifies the expression of a particular gene under the action of a perturbagen, relative to the baseline experiment.

We aggregate the z scores of experiments in the following manner: For a given cell line, perturbagen, dosage and duration, we compute the mean value. Then, across cell lines, perturbagens, dosages and durations, we take the maximum of the absolute value for a given interaction. The z score is then transformed using the normal cumulative density function (that takes values in (0, 1)). We denote this by $\hat{\pi}^{\alpha}$ and call it the experimental regulatory interaction strength.

For GeneWays, 40% of claims and 32% of interactions remain after merging with LINCS L1000, while for Literome, the corresponding fractions are 29% and 25%. After merging GeneWays and Literome onto aggregated LINCS L1000 data, we obtain 15.5k and 50.5k claims and 6.8k and 25.4k interactions, respectively. The overlap between the published claims (1) extracted by GeneWays, (2) extracted by Literome and also (3) present in (and merged with) LINCS L1000 is 2k interactions (31% of all interactions of GeneWays, 8% of all interactions of Literome) or 827 claims. The correlation of the intersection of GeneWays and Literome claims is 0.38 (representing 13% of GeneWays claims but only 4% of Literome claims). The number of overlapping interactions is greater than the number of overlapping claims because the majority of interactions are discussed in sets of publications that are disjoint between GeneWays and Literome. If we restrict the merged GeneWays–LINCS and Literome–LINCS datasets to the strongest positive and negative experimental regulatory interactions (intervals (0, 0.1] and [0.9. 1.0) on the interaction strength cumulative density function), the overlap between GeneWays and Literome is 81 claims (34 interactions) with a correlation on the claim variable of 0.57. We conclude that the GeneWays and Literome datasets, while being significantly different, are in moderate agreement where they overlap, suggesting that they are largely independent sources of genetic regulatory interaction claims. We note that the distribution of the number of claims per interaction follows Zipf's law (Extended Data Fig. 7). The correlation between regulatory interaction strength from LINCS L1000 and mean claim value per interaction from the literature is negligible, but increases as we introduce a threshold for the number of publications in which the claim appears (Supplementary Table 1 and Fig. 1).

Features predicting replicability (steps 2 and 3). Our models account for a range of scientific and social factors that could influence the likelihood that a claim is robust and generalizable. Two important classes of factors involve (1) how a claim fits into prior knowledge about nature and (2) its breadth of prior support. We measure how a claim fits with preexisting knowledge by assessing its position in the complex network of other scientific claims. We measure whether a claim is central or peripheral in the network, and whether the entire claim network is decentralized or hierarchical, controlled through a small number of central nodes. A claim's plausibility may also be affected by its position in the macro cluster structure of the network, that is, whether it lies in a large or small cluster of interactions. We examine a scientific claim's breadth of prior support by evaluating the distribution of researchers who have reiterated it and the depth of time over which a claim has been examined. We also include features measuring the authority of whether a finding was published in elite, high-impact journals, or was authored by scientists from elite schools with a strong reputation.

Specifically, we define communities associated with each genetic regulatory interaction for claims made within a variety of fixed time intervals: the past one, two, three and all years leading up to a given year. Each claim is made within a unique publication, which is produced in an institutional, social and knowledge context, reflected by the multiple authors, affiliations and citations referenced within it. We denote the set of affiliations (or authors, or references) by V and the set of publications by U, such that edges (u, v) between members of these two sets form a bipartite graph, which is reduced to a weighted graph defined on the set of publications U, with weights proportional to the number of common affiliations. In each such local weighted graph of publications defined over a given time period (that is, one, two, three and all prior years), we identify communities using the information theory-inspired InfoMap algorithm[61] and assign the number of communities, community size and community share for a given claim as derived features. All features are described in the Supplementary Information and listed in Supplementary Table 2, while the correlations of these features with claim correctness and interaction neutrality are presented in Extended Data Fig. 4.

To classify (1) the neutrality of a genetic regulatory interaction and (2) the correctness of a claim (the positivity or negativity of a regulatory interaction is derived from our correctness model), we used random forest and logistic regression models to enable both prediction and interpretation. While random forest allows us to reach near-maximal predictive performance, logistic regression enables the linear interpretation of features, rendering some effects positive and others negative. We choose models of optimal complexity and estimate metrics over the ensemble using procedures described in the Supplementary Information and Extended Data Fig. 8. In Fig. 2, features are presented pictorially with the highest Gini importance for the random forest model. The detailed methodology of the feature importance calculation is provided in the Supplementary Information. Figure 2 displays the Gini importance for each variable in the random forest model and associated coefficients for the logistic regression, plotted in decreasing importance for the GeneWays random forest.

Sampling procedure. For each model type, we randomly generated 20 threefold samples of interactions. Leaving 1 out of 3 for validation in each of these 20 samples yields 60 training–validation pairs. These threefold samples were constructed randomly per interaction (using the claim number distribution function). For each experiment, models were trained on training samples and metrics were evaluated on validation samples. The claim correctness model is trained and then validated on sets containing disjoint genetic interactions.

Sensitivity analyses. We performed sensitivity analyses of the pipeline used to derive model features and estimate model parameters to account for the heterogeneous quality of research across the scientific system. First, we restricted our sample of claims to a subset of GeneWays and Literome published in high-reputation journals, measured as being in the 90th percentile of the Article Influence Score[62]. This measure uses an eigenvalue-based centrality measure on citation networks to assess those journals that exert the most direct and indirect influence over other journals. Our entire sample of GeneWays and Literome claims is derived from thousands of journals and may raise concerns that lower-quality or fraudulent results may be more highly represented there and artificially inflate our findings (for example, if our models discounted findings from low- but not high-profile journals).

Second, we restricted our sample of claims to those authored by scientists at respected research universities. These scientists represent what Ioannidis and colleagues have called the 'Continuously Publishing Core in the Scientific Workforce'[63]. We performed this analysis on a subset of the GeneWays and Literome dataset restricted to publications by authors affiliated with the top 100 universities according to the 100 Quacquarelli Symonds (QS) World University Rankings of biological sciences in 2011. The QS World University Rankings is an annual publication of university rankings, previously known as the Times Higher Education–QS World University Rankings (through 2009). Our entire sample of claims is derived from both active, centrally positioned scientists (for example, at the University of Oxford) and those who publish intermittently at peripheral institutions (for example, at Friends College) and may raise the concerns that lower-quality results may be more highly represented and artificially inflate our findings (for example, if our models discounted findings from researchers at low- but not high-status institutions).

With these new subsets, we estimated AUC distributions (out of sample) for all three models described above. The high-profile journal filter leaves us with 46% and 51% of GeneWays and Literome, respectively. The high-profile affiliations filter leaves 21% and 22% of the GeneWays and Literome dataset, respectively. These results are summarized in Supplementary Table 3 and demonstrate the stability of our results to these sample restrictions. Results for the restricted high-profile affiliations and journal samples are comparable to those from the full sample. For the Literome dataset, the AUCs for high-profile affiliations and journals are slightly higher than the full sample for all models. For GeneWays, the high-profile samples are slightly higher than the full sample for the neutral model (step 2), lower on the claims model (step 3) and approximately the same for the

positive–negative interaction inference (step 4). In short, all reported findings are robust to these perturbations. These findings are also consistent with our result that claims published in higher-profile journals or affiliated with higher-profile institutions are not more likely to replicate.

Additional details with regard to all the steps are given in the Supplementary Information. We used Python and scikit-learn to develop the models. Large-scale computations were made possible thanks to the Cloud Kotta infrastructure[64].

Data availability

To illustrate our pipeline, we used the publicly available GeneWays and Literome datasets (available at https://github.com/KnowledgeLab/geneways and https://github.com/KnowledgeLab/literome), linked with Clarivate's Web of Science database of bibliographic information. While we cannot share the Web of Science, we share a linked file https://github.com/KnowledgeLab/nmi_robust_facts_supplemenary, which includes all claims of interest and citation metadata required to perform described analyses.

Code availability

Our code is publicly available at https://github.com/alexander-belikov/datahelpers and https://github.com/KnowledgeLab/bm_support.

Received: 21 November 2020; Accepted: 6 March 2022;
Published online: 28 April 2022

References

1. Hey, T. & Trefethen, A. in *Grid Computing: Making the Global Infrastructure a Reality* (eds Fox, G. C. & Hey, T.) 809–824 (Wiley, 2003).
2. Bell, G., Hey, T. & Szalay, A. Computer science. Beyond the data deluge. *Science* **323**, 1297–1298 (2009).
3. Burger, B. et al. A mobile robotic chemist. *Nature* **583**, 237–241 (2020).
4. King, R. D. et al. Functional genomic hypothesis generation and experimentation by a robot scientist. *Nature* **427**, 247–252 (2004).
5. Zhou, Q. et al. Learning atoms for materials discovery. *Proc. Natl Acad. Sci. USA* **115**, E6411–E6417 (2018).
6. Tshitoyan, V. et al. Unsupervised word embeddings capture latent knowledge from materials science literature. *Nature* **571**, 95–98 (2019).
7. Nissen, S. B., Magidson, T., Gross, K. & Bergstrom, C. T. Publication bias and the canonization of false facts. *eLife* **5**, e21451 (2016).
8. Daston, L. J. & Galison, P. *Objectivity* (Zone Books, 2007).
9. Foreman, P. Weimar culture, causality and quantum theory 1918–1927. *Hist. Stud. Phys. Biol. Sci.* **3**, 2–225 (1971).
10. Rzhetsky, A., Iossifov, I., Loh, J. M. & White, K. P. Microparadigms: chains of collective reasoning in publications about molecular interactions. *Proc. Natl Acad. Sci. USA* **103**, 4940–4945 (2006).
11. Ioannidis, J. P. A. Why most published research findings are false. *PLoS Med.* **2**, e124 (2005).
12. Surowiecki, J. *The Wisdom of Crowds: Why the Many Are Smarter Than the Few and How Collective Wisdom Shapes Business, Economies, Societies and Nations* (Doubleday, 2004).
13. Galton, F. Vox populi (the wisdom of crowds). *Nature* **75**, 450–451 (1907).
14. Hong, L. & Page, S. E. Groups of diverse problem solvers can outperform groups of high-ability problem solvers. *Proc. Natl Acad. Sci. USA* **101**, 16385–16389 (2004).
15. Becker, J., Brackbill, D. & Centola, D. Network dynamics of social influence in the wisdom of crowds. *Proc. Natl Acad. Sci. USA* **114**, E5070–E5076 (2017).
16. Lorenz, J., Rauhut, H., Schweitzer, F. & Helbing, D. How social influence can undermine the wisdom of crowd effect. *Proc. Natl Acad. Sci. USA* **108**, 9020–9025 (2011).
17. Danchev, V., Rzhetsky, A. & Evans, J. A. Centralized communities more likely generate non-replicable results. *eLife* **8**, e43094 (2019).
18. Hicks, D. M. & Katz, J. S. Where is science going? *Sci. Technol. Human Values* **21**, 379–406 (1996).
19. Guimerà, R., Uzzi, B., Spiro, J. & Amaral, L. A. N. Team assembly mechanisms determine collaboration network structure and team performance. *Science* **308**, 697–702 (2005).
20. Hand, E. 'Big science' spurs collaborative trend. *Nature* **463**, 282–282 (2010).
21. Wuchty, S., Jones, B. F. & Uzzi, B. The increasing dominance of teams in production of knowledge. *Science* **316**, 1036–1039 (2007).
22. Wu, L., Wang, D. & Evans, J. A. Large teams develop and small teams disrupt science and technology. *Nature* https://doi.org/10.1038/s41586-019-0941-9 (2019).
23. Jones, B. F., Wuchty, S. & Uzzi, B. Multi-university research teams: shifting impact, geography, and stratification in science. *Science* **322**, 1259–1262 (2008).
24. Merton, R. K. The Matthew effect in science: the reward and communication systems of science are considered. *Science* **159**, 56–63 (1968).
25. Azoulay, P., Stuart, T. & Wang, Y. Matthew: effect or fable? *Manage. Sci.* **60**, 92–109 (2014).
26. Evans, J. A. Electronic publication and the narrowing of science and scholarship. *Science* **321**, 395–399 (2008).
27. Simkin, M. V. & Roychowdhury, V. P. Do copied citations create renowned papers? *Ann. Improbable Res.* **11**, 24–27 (2005).
28. Chu, J. S. G. & Evans, J. A. Slowed canonical progress in large fields of science. *Proc. Natl. Acad. Sci. USA* **118**, e2021636118 (2021).
29. Mullard, A. Reliability of 'new drug target' claims called into question. *Nat. Rev. Drug Discov.* **10**, 643–644 (2011).
30. Prinz, F., Schlange, T. & Asadullah, K. Believe it or not: how much can we rely on published data on potential drug targets? *Nat. Rev. Drug Discov.* **10**, 712–712 (2011).
31. Freedman, L. P. & Gibson, M. C. The impact of preclinical irreproducibility on drug development. *Clin. Pharmacol. Ther.* **97**, 16–18 (2015).
32. Ioannidis, J. P., Ntzani, E. E., Trikalinos, T. A. & Contopoulos-Ioannidis, D. G. Replication validity of genetic association studies. *Nat. Genet.* **29**, 306–309 (2001).
33. Hirschhorn, J. N., Lohmueller, K., Byrne, E. & Hirschhorn, K. A comprehensive review of genetic association studies. *Genet. Med.* **4**, 45–61 (2002).
34. Lohmueller, K. E., Pearce, C. L., Pike, M., Lander, E. S. & Hirschhorn, J. N. Meta-analysis of genetic association studies supports a contribution of common variants to susceptibility to common disease. *Nat. Genet.* **33**, 177–182 (2003).
35. Open Science Collaboration. Estimating the reproducibility of psychological science. *Science* **349**, aac4716 (2015).
36. Van Bavel, J. J., Mende-Siedlecki, P., Brady, W. J. & Reinero, D. A. Contextual sensitivity in scientific reproducibility. *Proc. Natl Acad. Sci. USA* **113**, 6454–6459 (2016).
37. Zollman, K. J. S. The communication structure of epistemic communities. *Phil. Sci.* **74**, 574–587 (2007).
38. Payette, N. in *Models of Science Dynamics: Encounters between Complexity Theory and Information Sciences* (eds Scharnhorst, A., Börner, K. & van den Besselaar, P.) 127–157 (Springer, 2012).
39. Baker, M. Biotech giant publishes failures to confirm high-profile science. *Nature* **530**, 141 (2016).
40. Borenstein, M., Hedges, L. V., Higgins, J. P. T. & Rothstein, H. R. *Introduction to Meta-Analysis* (Wiley, 2011).
41. Nussbaum, D. The role of conceptual replication. *Psychologist* **25**, 350 (2012).
42. Barragan-Jason, G., Atance, C. M., Hopfensitz, A., Stieglitz, J. & Cauchoix, M. Commentary: Revisiting the marshmallow test: a conceptual replication investigating links between early delay of gratification and later outcomes. *Front. Psychol.* **9**, 2719 (2019).
43. MacLeod, C. & McLaughlin, K. Implicit and explicit memory bias in anxiety: a conceptual replication. *Behav. Res. Ther.* **33**, 1–14 (1995).
44. Hagemann, D., Naumann, E., Becker, G., Maier, S. & Bartussek, D. Frontal brain asymmetry and affective style: a conceptual replication. *Psychophysiology* **35**, 372–388 (1998).
45. Horselenberg, R., Merckelbach, H. & Josephs, S. Individual differences and false confessions: a conceptual replication of Kassin and Kiechel (1996). *Psychol. Crime Law* **9**, 1–8 (2003).
46. Belknap, P. & Leonard, W. M. A conceptual replication and extension of Erving Goffman's study of gender advertisements. *Sex Roles* **25**, 103–118 (1991).
47. Seyedghorban, Z., Tahernejad, H. & Matanda, M. J. Reinquiry into advertising avoidance on the internet: a conceptual replication and extension. *J. Advert.* **45**, 120–129 (2016).
48. Lu, Y., Ossmann, M. M., Leaf, D. E. & Factor, P. H. Patient visibility and ICU mortality: a conceptual replication. *HERD* **7**, 92–103 (2014).
49. Friedman, C., Kra, P. & Rzhetsky, A. Two biomedical sublanguages: a description based on the theories of Zellig Harris. *J. Biomed. Inform.* **35**, 222–235 (2002).
50. Rzhetsky, A. et al. GeneWays: a system for extracting, analyzing, visualizing, and integrating molecular pathway data. *J. Biomed. Inform.* **37**, 43–53 (2004).
51. Quirk, C. et al. MSR SPLAT, a language analysis toolkit. In *Proc. 2012 Conference of the North American Chapter of the Association for Computational Linguistics* (Association for Computational Linguistics, 2012).
52. Kim, J.-D., Ohta, T., Pyysalo, S., Kano, Y. & Tsujii, J. Overview of BioNLP'09 shared task on event extraction. In *Proc. BioNLP 2009 Workshop Companion Volume for Shared Task* (Association for Computational Linguistics, 2009).
53. Rosvall, M., Axelsson, D. & Bergstrom, C. T. The map equation. *Eur. Phys. J. Spec. Top.* **178**, 13–23 (2009).
54. Subramanian, A. et al. A next generation connectivity map: L1000 platform and the first 1,000,000 profiles. *Cell* **171**, 1437–1452 (2017).
55. Rosenthal, R. The file drawer problem and tolerance for null results. *Psychol. Bull.* **86**, 638 (1979).
56. Scargle, J. D. Publication bias (the 'file-drawer problem') in scientific inference. Preprint at https://arxiv.org/abs/physics/9909033 (1999).
57. Sunstein, C. R. *Republic.com* (Princeton Univ. Press, 2001).

58. Stoeger, T., Gerlach, M., Morimoto, R. I. & Nunes Amaral, L. A. Large-scale investigation of the reasons why potentially important genes are ignored. *PLoS Biol.* **16**, e2006643 (2018).
59. Rzhetsky, A. et al. GeneWays: a system for extracting, analyzing, visualizing, and integrating molecular pathway data. *J. Biomed. Inform.* **37**, 43–53 (2004).
60. Poon, H., Quirk, C., DeZiel, C. & Heckerman, D. Literome: PubMed-scale genomic knowledge base in the cloud. *Bioinformatics* **30**, 2840–2842 (2014).
61. Rosvall, M. & Bergstrom, C. T. Maps of random walks on complex networks reveal community structure. *Proc. Natl Acad. Sci. USA* **105**, 1118–1123 (2008).
62. Bergstrom, C. T., West, J. D. & Wiseman, M. A. The eigenfactor™ metrics. *J. Neurosci.* **28**, 11433–11434 (2008).
63. Ioannidis, J. P. A., Boyack, K. W. & Klavans, R. Estimates of the continuously publishing core in the scientific workforce. *PLoS ONE* **9**, e101698 (2014).
64. Babuji, Y. N., Chard K., Gerow, A. & Duede, E. Cloud Kotta: enabling secure and scalable data analytics in the cloud. In *IEEE International Conference on Big Data* 302–310 (IEEE, 2016).

Acknowledgements

We thank V. Sitnik, V. Danchev and P. Saleiro for fruitful discussions, Y. Babuji for technical help, H. Poon for suggestions regarding the formulation of the project and I. Mayzus, R. Melamed and O. Kel-Margoulis for help with the annotation and the interpretation of biological datasets. We are grateful for comments from participants of the MetaScience Conference at Stanford (2019), and for meetings associated with the Defense Advanced Research Projects Agency (DARPA) Big Mechanism programme. We acknowledge funding from DARPA (14145043, J.E. and A.V.B.; HR00111820006, J.E., A.V.B. and A.R.), the Air Force Office of Scientific Research (FA9550-19-1-0354, J.E.; FA9550-15-1-0162, J.E.), the National Science Foundation (SBE-1829366, J.E.; 1422902, J.E.; 1158803, J.E.) and the John Templeton Foundation to the 'Metaknowledge Network' (J.E. and A.R.).

Author contributions

A.V.B. proposed and implemented the methodology, validated the model, analysed the data and drafted the paper. J.E. was responsible for conception and funding of the project, contributed to the design of the methodology and drafted the paper. A.R. provided feedback on the experimental work and data interpretation, and participated in drafting the paper. All authors read and approved the final manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Extended data is available for this paper at https://doi.org/10.1038/s42256-022-00474-8.

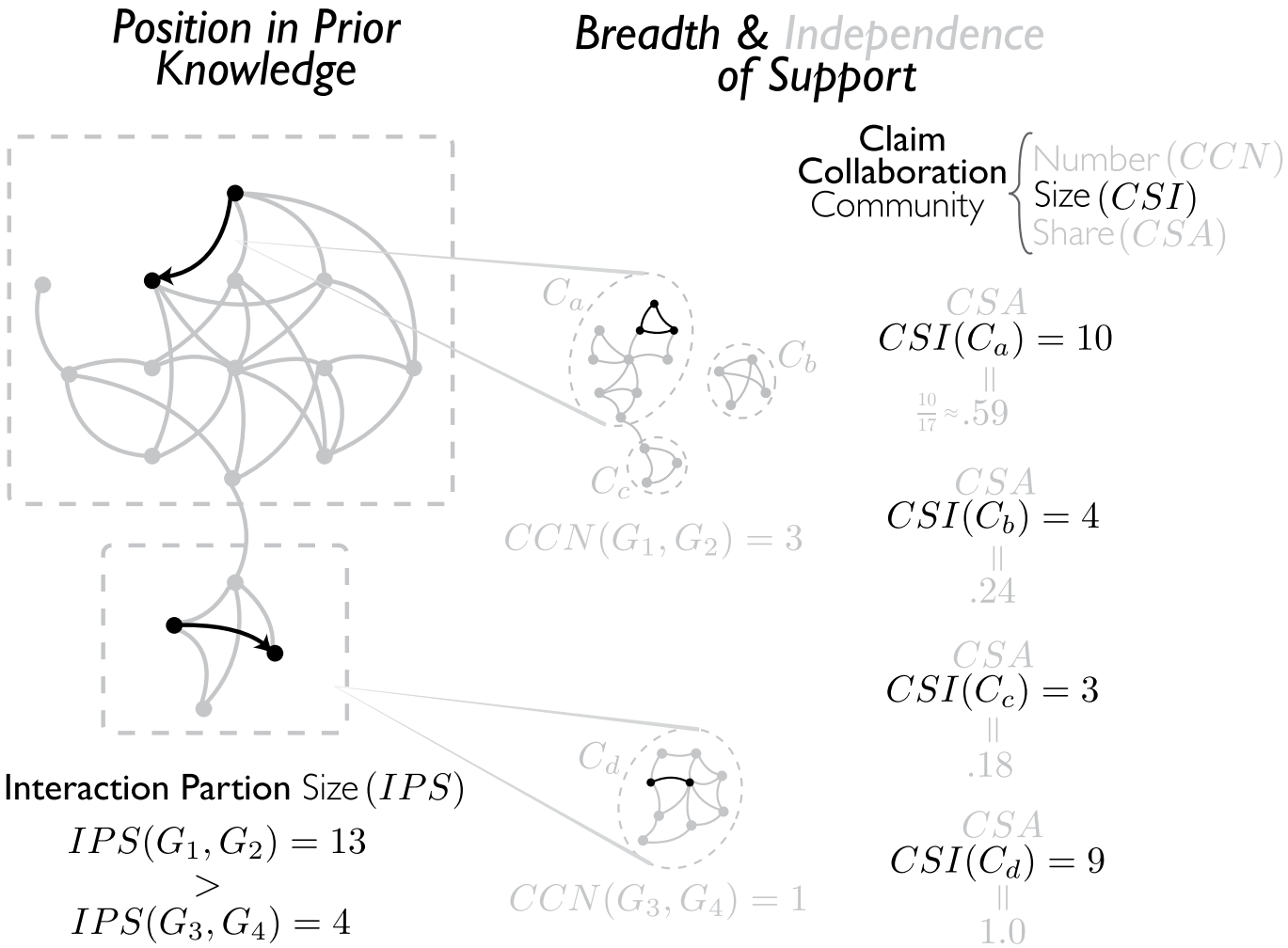
Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s42256-022-00474-8.

Correspondence and requests for materials should be addressed to Alexander V. Belikov or James Evans.

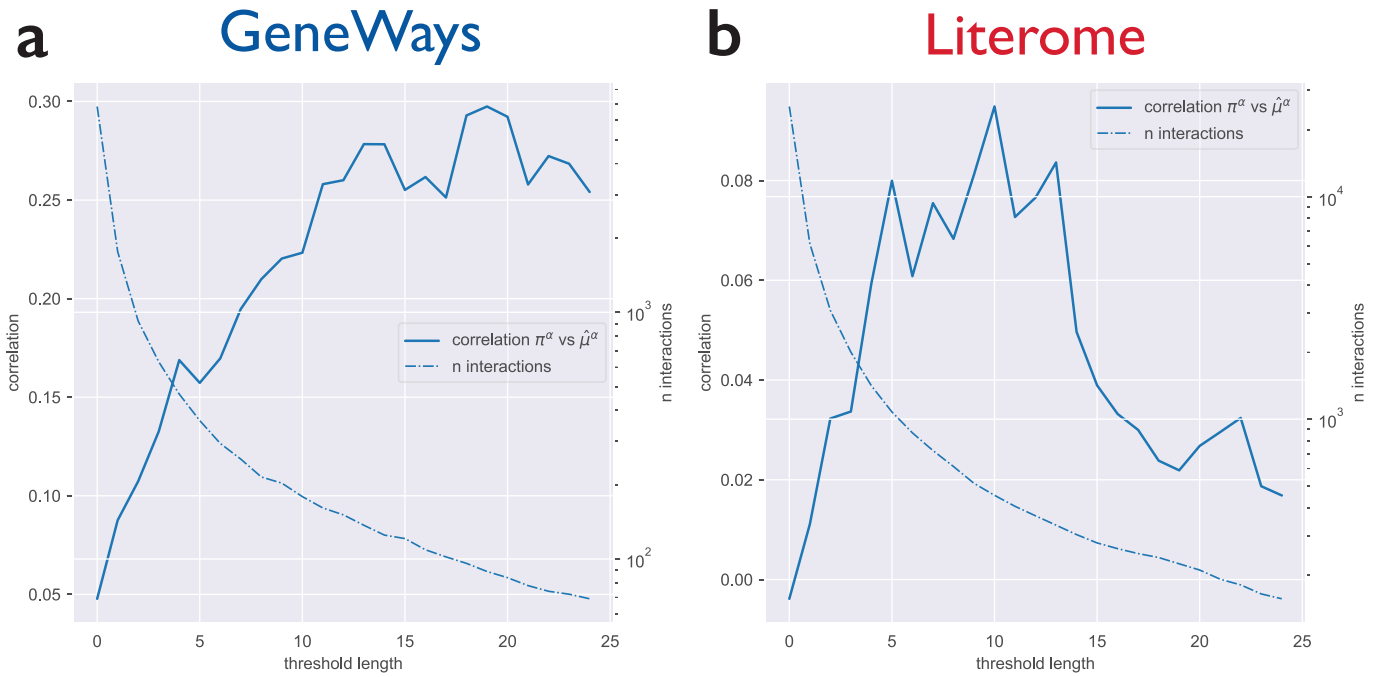
Peer review information *Nature Machine Intelligence* thanks Luis Amaral and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

Reprints and permissions information is available at www.nature.com/reprints.

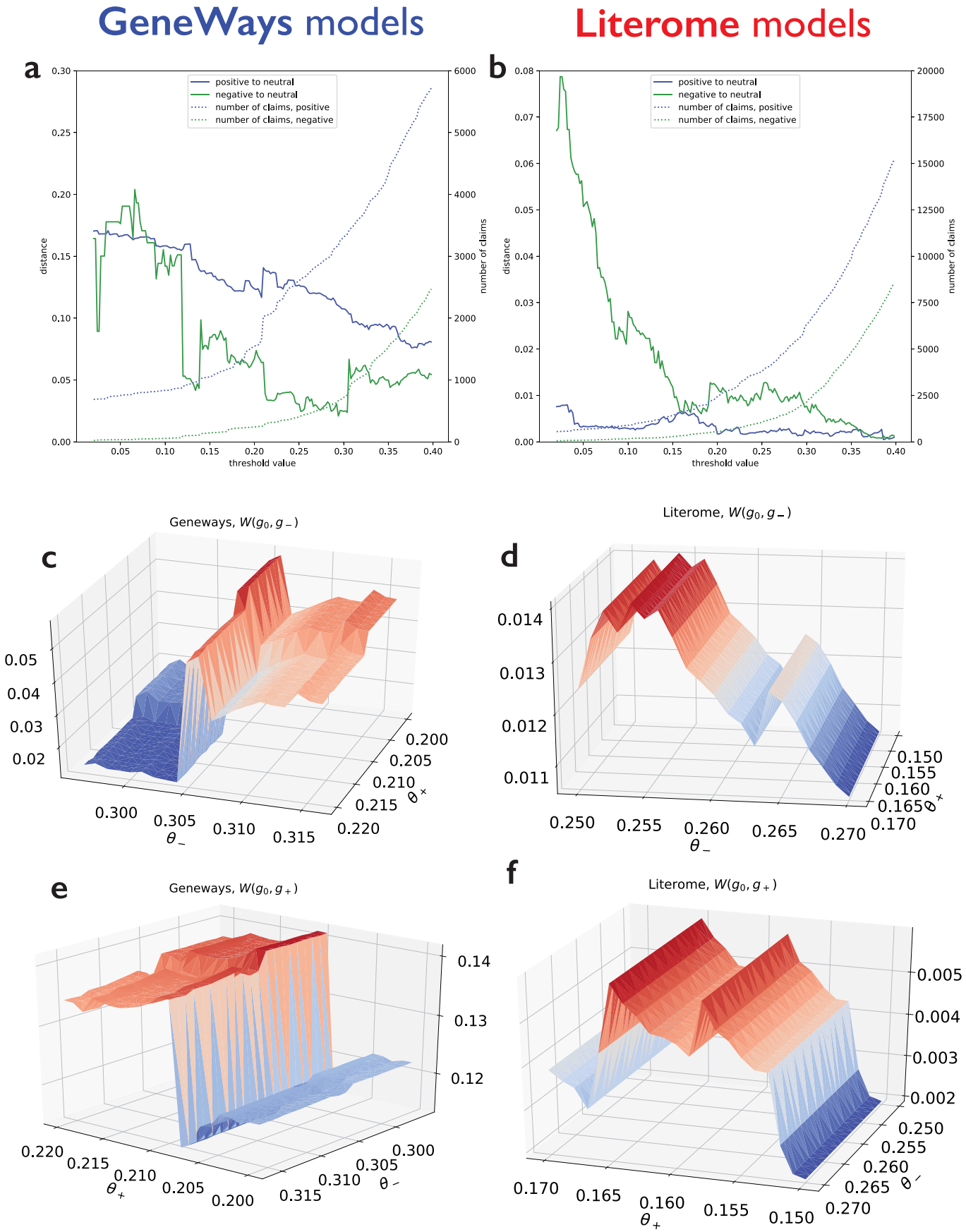
Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Extended Data Fig. 1 | Illustration of core interaction and claim variables. Directed regulatory interactions between genes constitute communities of researchers who study them. Features regarding the position of a claim within prior knowledge are derived from its relationship to other genetic regulatory interactions. Features regarding the breadth and independence of support are derived from the connection between publications making claims about the same interaction.

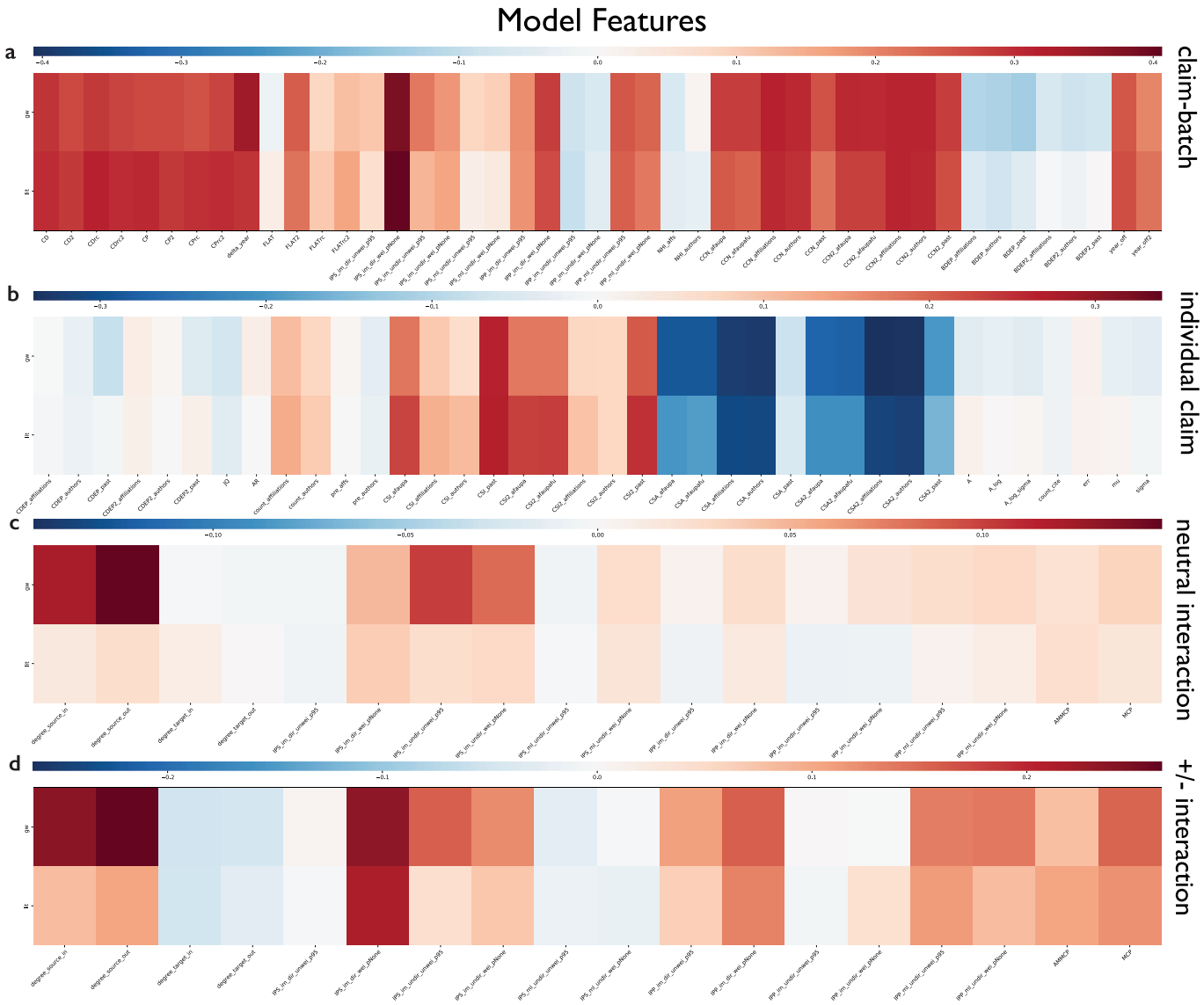


Extended Data Fig. 2 | Correlation between claim value and experimental strength across the claim frequency distribution. Correlation of mean claim value $\hat{\mu}^\alpha$ and interaction strength $\hat{\pi}^\alpha$ from LINCS L1000 as a function of threshold on minimum claim sequence length per interaction for GeneWays (**a**) and Literome (**b**).

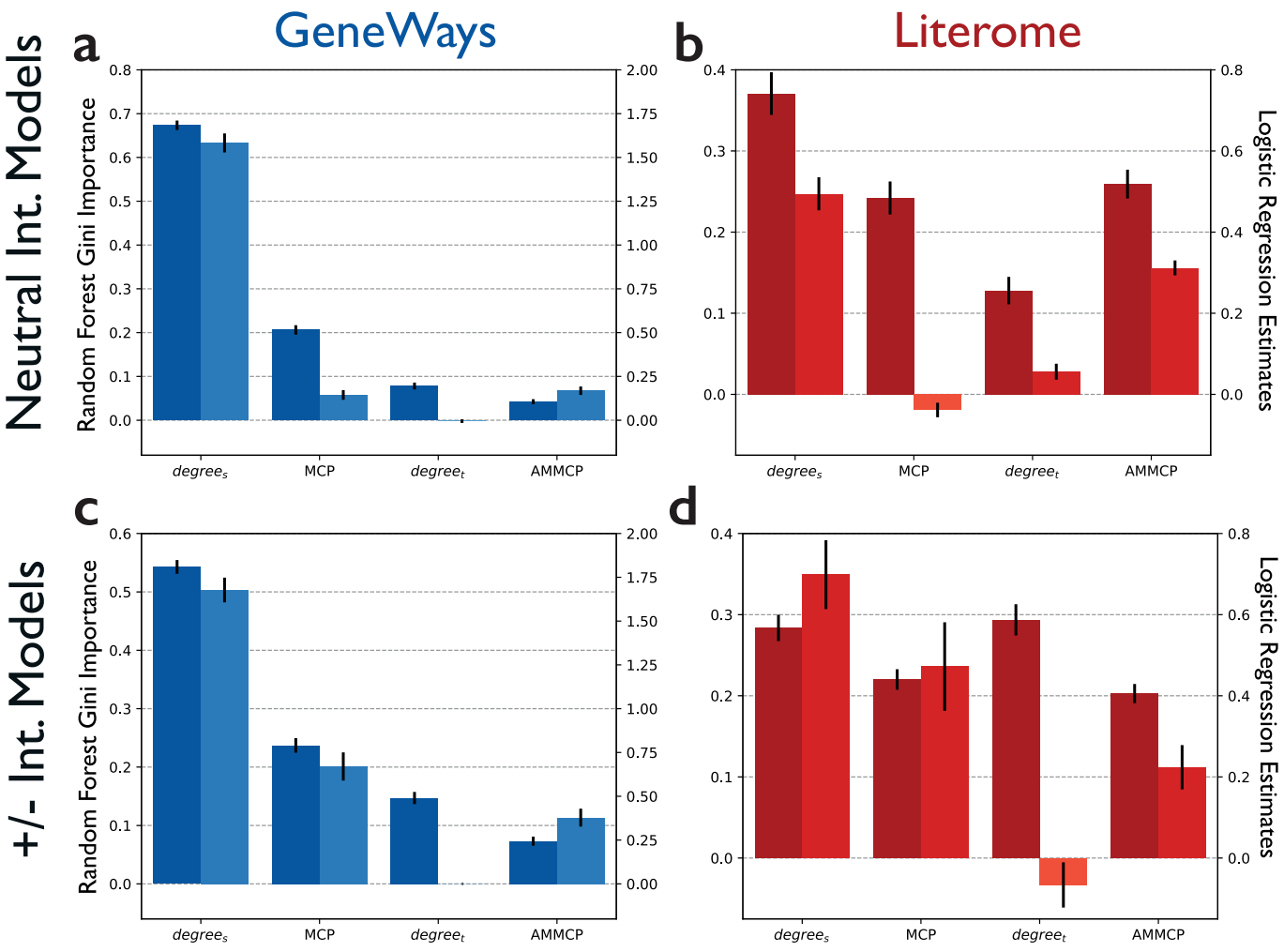


Extended Data Fig. 3 | See next page for caption.

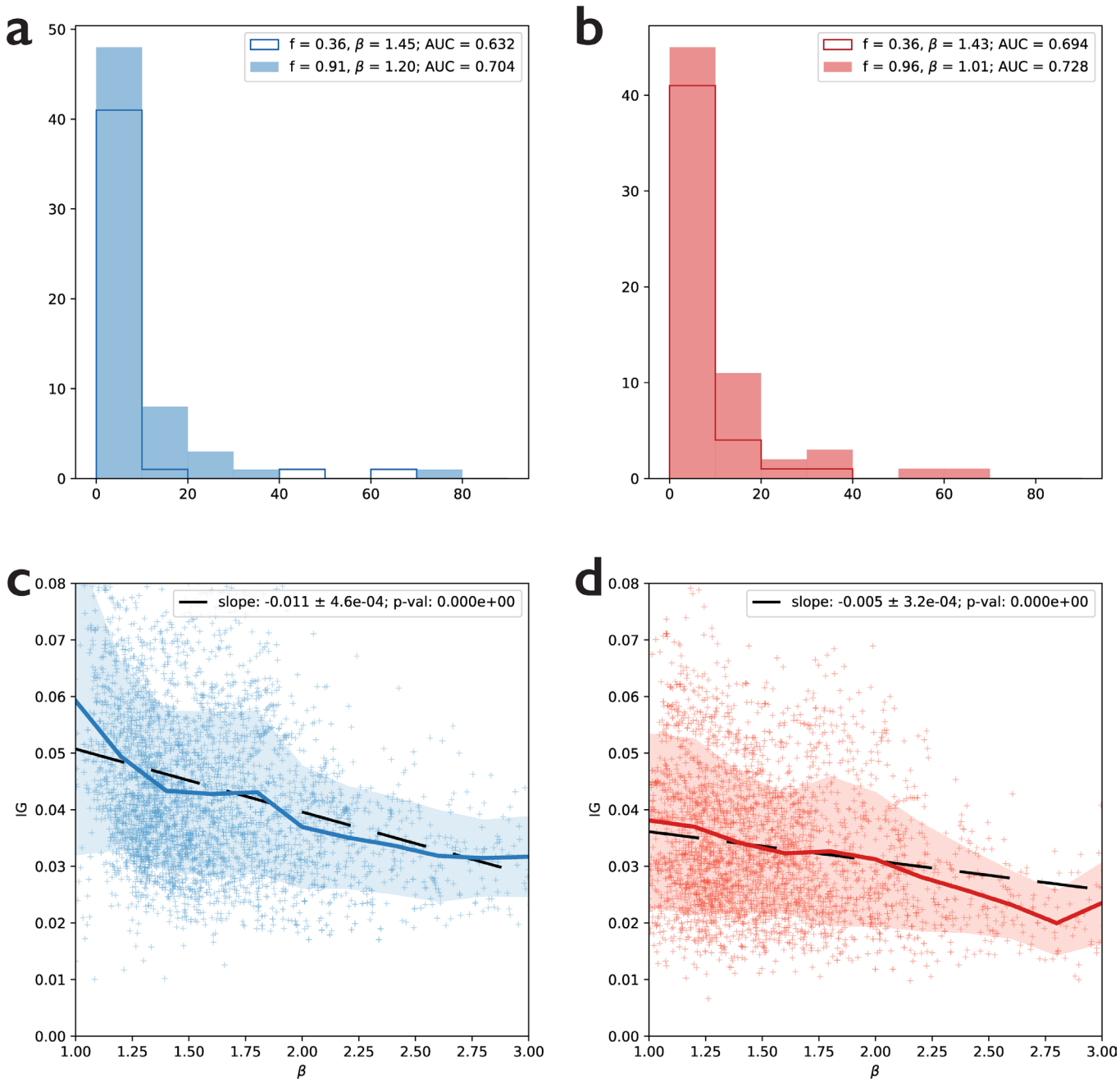
Extended Data Fig. 3 | Data-driven thresholds to partition interactions into neutral, negative and positive interactions for analysis. C_0, C_- and C_+ correspond to classes of neutral, negative and positive genetic regulatory interactions. Distance between $C_0$and C_- ($W(g_0, g_-, \theta_-, \theta_+)$, solid green), and C_0 and C_+ ($W(g_0, g_+, \theta_-, \theta_+)$, solid blue), number of claims in C_-, dotted green, number of claims in C_+, dotted blue) for GeneWays (**a**) and Literome (**b**). Distance between C_0 and C_- ($W(g_0, g_-, \theta_-, \theta_+)$) in GeneWays (**c**) and Literome (**d**); Distance between C_0 and C_+ ($W(g_0, g_+, \theta_-, \theta_+)$) in GeneWays (**e**) and Literome (**f**).



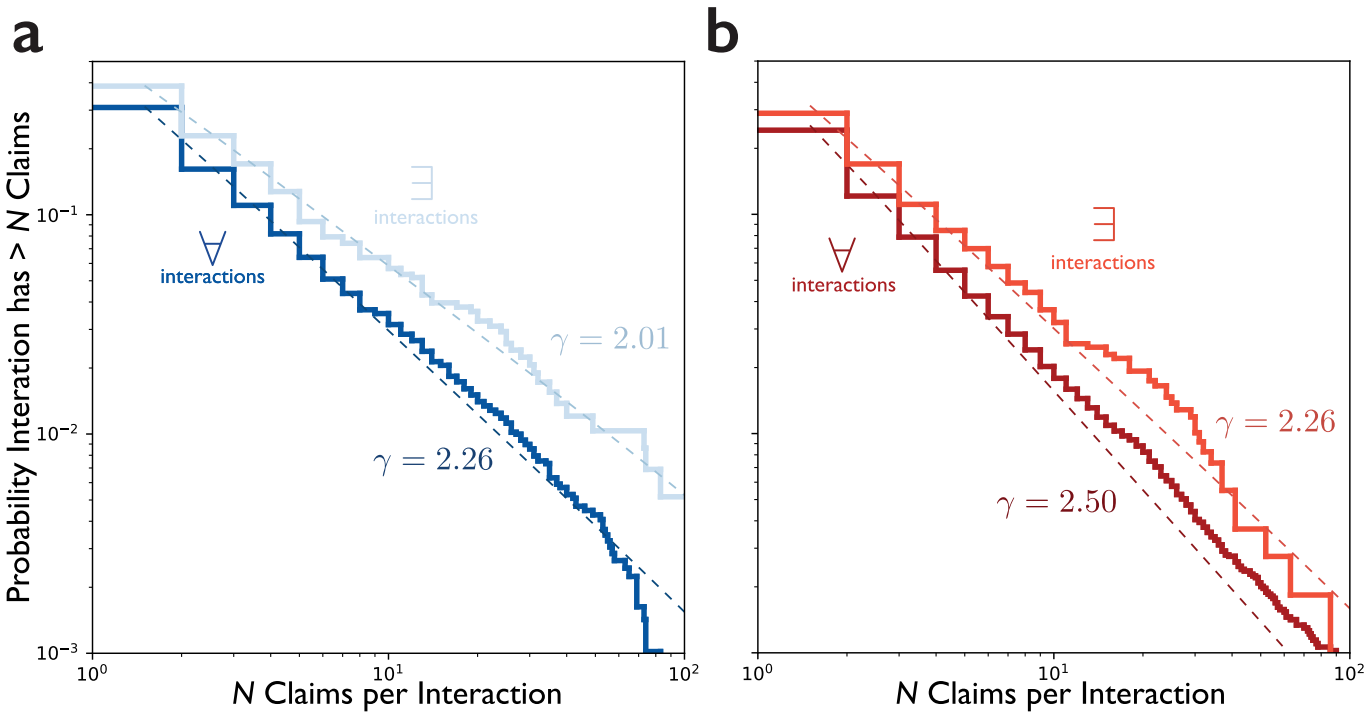
Extended Data Fig. 4 | Pearson correlation between core analysis variables in both Geneways and Literome datasets. Heat map indicating correlation between: (**a**) claim correctness y_i^α and batch-level features for GeneWays (top row) and Literome (bottom row); (**b**) claim correctness y_i^α and claim-level features for GeneWays (top row) and Literome (bottom row); (**c**) interaction non-neutrality π_0^α and interaction-level features for GeneWays (top row) and Literome (bottom row); (**d**) interaction positivity π_+^α and interaction-level features for GeneWays (top row) and Literome (bottom row).



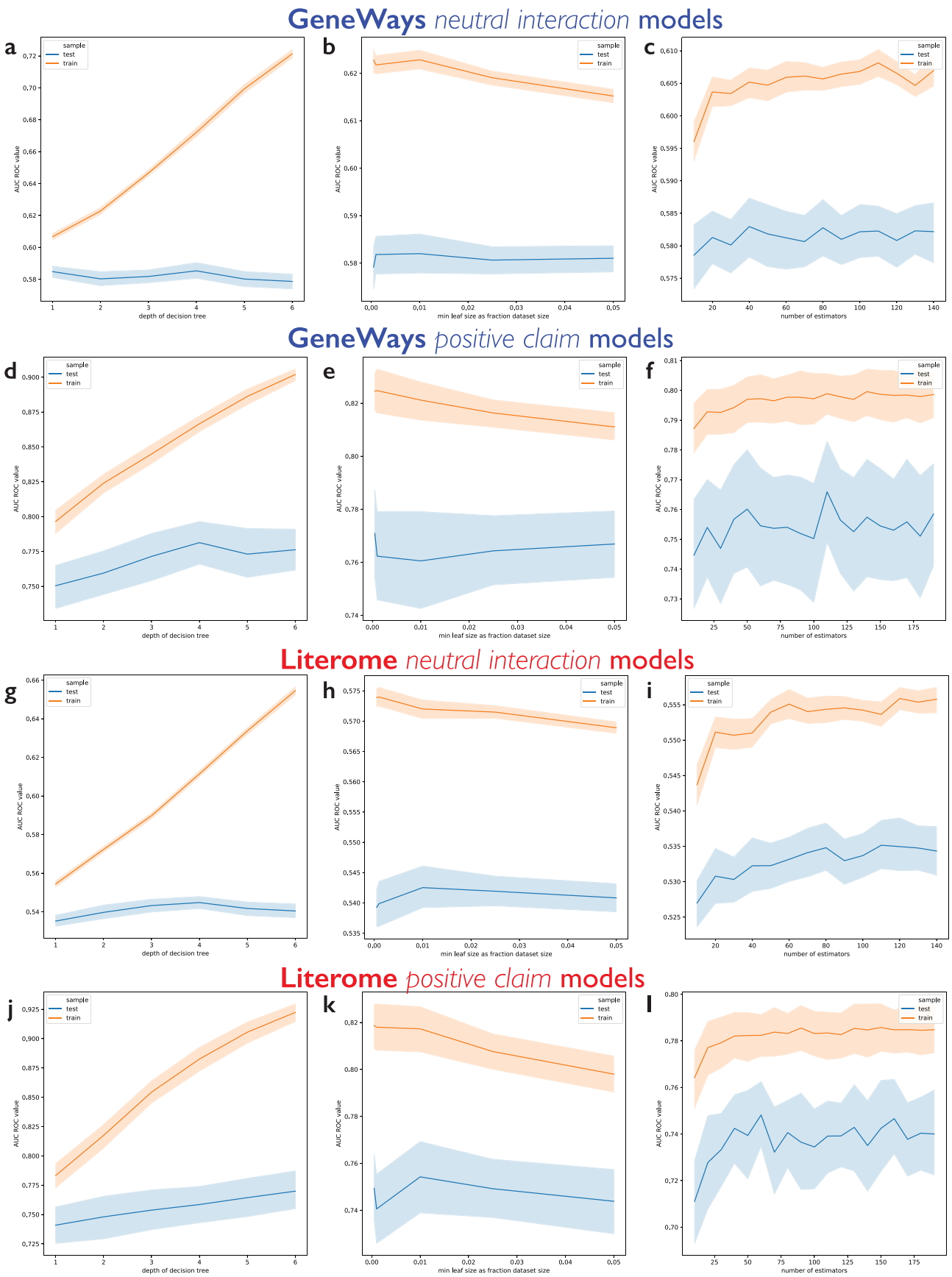
Extended Data Fig. 5 | Variable importance and significance in models of the non-neutrality and positivity of genetic regulatory interactions. Family importances of random forest model (left, darker shade) and logistic regression coefficients (right, lighter shade) for the model of classification of neutral interactions (top) and positive interactions (bottom) for GeneWays (left) and Literome (right). Vertical centered lines show 95% confidence level on the mean of the corresponding importance/coefficient.



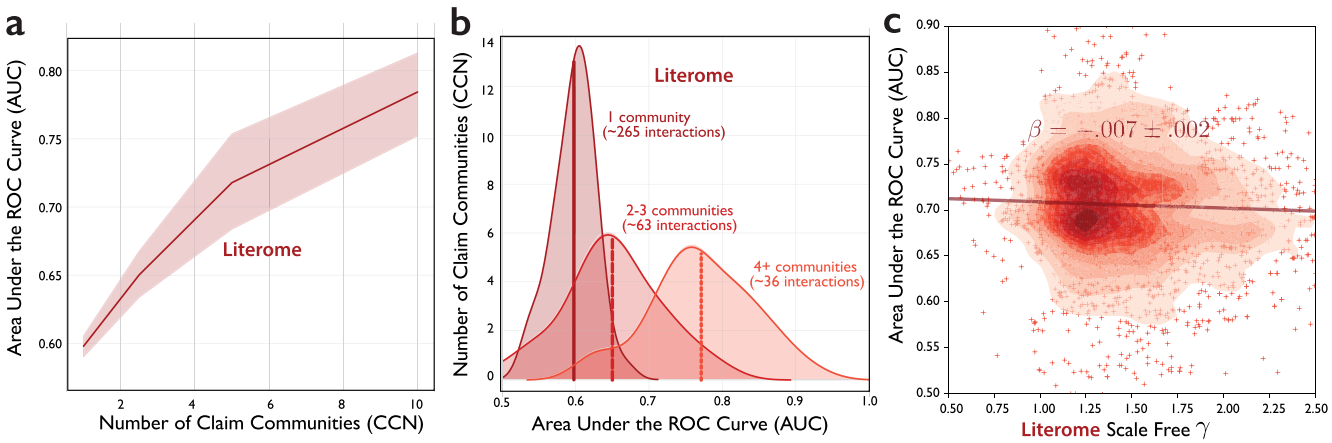
Extended Data Fig. 6 | Analysis of the relationship between the distribution of claims per interaction and overall certainty about those interactions. Examples of claim number distribution $\rho(n_\alpha)$ per interaction for test subsamples from GeneWays (**a**) and Literome (**b**). Information gain as a function of the slope of claim number distribution β. Solid lines correspond to binned averages and shaded regions denote one standard deviation of the data confidence interval for GeneWays (**c**) and Literome (**d**).



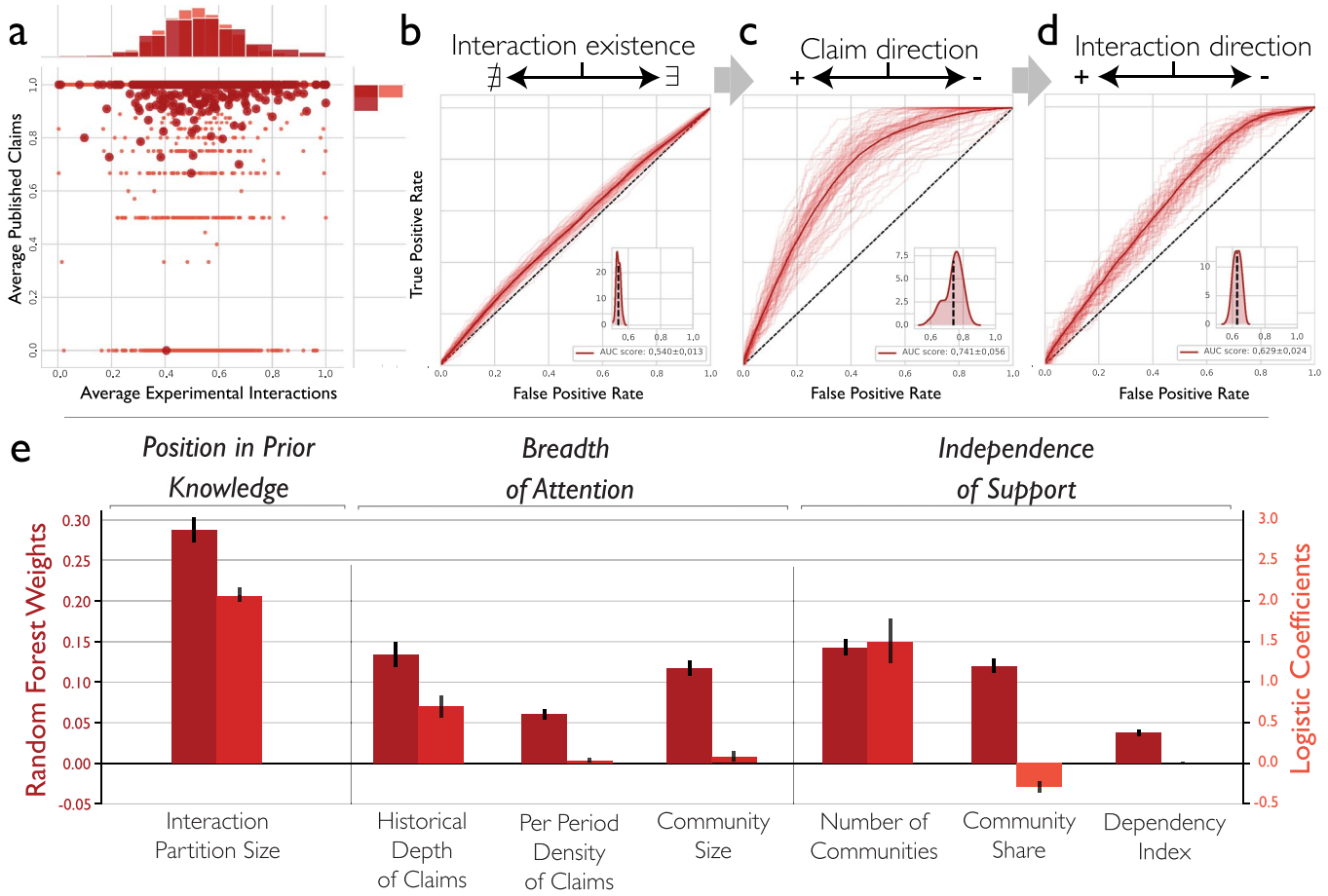
Extended Data Fig. 7 | Survival functions (complements of the cumulative distribution functions) of claim number per interaction. Survival functions for GeneWays (a) and Literome (b); for all interactions (∀) and nonzero (∃) interactions, where the probability distribution function is modeled as $\rho \propto n^{\gamma}$. Exponents γ equal 2.26 and 2.01 for Geneways for all and non-neutral interactions, respectively; and equal 2.5 and 2.26 for Literome for all and non-neutral interactions. The exponents were obtained by Maximum Likelihood Estimation.



Extended Data Fig. 8 | Model selection using ROC AUC values for all models. Neutral interaction models (**a-c,g-i**) and positive interaction models (**d-f,j-l**). Left: the distribution of ROC AUC as a function of random forest depth (**a,d,g,j**). Center: the distribution of ROC AUC as a function of minimum number of samples in a decision tree leaf (**b,e,h,k**). Right: the distribution of ROC AUC as a function of the number of trees in a random forest (**c,f,i,l**).



Extended Data Fig. 9 | Science policy experiments revealing the relationship between community independence, collective attention, and certainty about genetic regulatory interactions (complement to Fig. 4). a, Relationship between the number of communities studying a particular genetic regulatory interaction and the average AUC of out-of-sample predictions for positive interactions. **b**, Distribution of the average AUC curves for Literome for interactions with 1, 2-3 and greater than 4 communities. **c**, Relationship between the shape of the distribution of number of claims per interaction on the AUC of out-of-sample predictions for positive interactions. β represents the slope of the claim number per interaction distribution for Literome. (Complement to main Fig. 4).



Extended Data Fig. 10 | Positivity bias in published effects and prediction results for Literome (complement to Fig. 3); random forest Gini Importance scores and logistic regression coefficients for features from Literome (complement to Fig. 2b). a, Joint plot of the mean experimental interaction strength (x-axis) and mean value of the published claim (y-axis) for each genetic interaction. More intense hues of the red (and also greater marker size) correspond to the interactions in Literome with 10 or more claims per interaction; for less intense hues (and also smaller marker size) the cutoff is absent, representing the complete distribution. (See Fig. 3a for comparable Geneways distribution). **b**, We first predicted the nonexistentence (∄) or existence (∃) of each published gene-gene regulatory interaction (Literome). **c**, Then, if the interaction was deemed existent (∃), we predicted whether each claim (of positivity or negativity) from literature was correct. **d**, Using Bayesian inference, we estimated the sign (positive vs negative) of all genetic regulatory interactions. Mean ROC curves in bold are complemented by a 95% c.i. contours, with fainter individual lines corresponding to ROC curves for 60 models corresponding to different training/validation samples. (Complement to Fig. 3 in the main manuscript). **e**, Gini Importance or Mean Decrease in Impurity for features in the random forest models (left vertical scale, bold colors), and coefficients from the logistic regression models (right vertical scale, fainter colors) for Literome. Vertical bars represent 95% c.i. for the mean value of the estimate.