Ranking missing links by shared neighbours loses to popularity

A preregistered time-sliced test of Swanson's A-B-C ranking on research notes and two Stack Exchange sites, with rewired and permuted controls

Zain Dana Harper

Version 1.0, 9 November 2026. Finding F152 in the findings ledger. Honest null (tier D).

Abstract

Tools that promise to find undiscovered connections often rank unlinked pairs of concepts by the neighbours they share: if A goes with B and B goes with C, perhaps A goes with C. We tested that ranking against the plainest baseline there is, which pairs the most popular concepts. Each corpus was cut by date, and a ranker scored a hit when a pair it proposed from the early documents appeared together at least twice in the next stretch of documents. Ranking unlinked pairs by shared neighbours found fewer later links than ranking them by popularity on all three corpora: precision in the top 200 of 0.127 against 0.268 on our research notes, 0.018 against 0.088 on Biology Stack Exchange and 0.007 against 0.043 on Skeptics Stack Exchange. On degree-preserving rewired graphs the shared-neighbour score reached a 95th percentile of 0.164, above its real score of 0.127, so at the top of the list it added nothing we could detect beyond popularity. Among pairs of similar popularity the shared-neighbour score still separated later links from the rest: area under the curve 0.569 on the research notes, 0.757 on Biology and 0.696 on Skeptics. Three of four preregistered verdicts failed.

1 What this means for a reader

If a discovery tool shows you a ranked list of "missing links", ask how it does against a list that simply pairs the most common terms. In our test the common-terms list won on every corpus. The shared-neighbour idea is not empty: among pairs that are equally common, it does pick out the ones that later appear together more often than chance. Its weakness sits at the top of the list, which is where a reader looks first.

2 Introduction

Swanson [1] described knowledge that is public and still undiscovered: two bodies of literature that each hold half of an argument and never cite each other. His method, now called literature-based discovery, looks for a term C that is linked to a term B in one literature while B is linked to A in another, so that A and C become a candidate pair. Later work recast the problem as link prediction, the ranking of node pairs in a network by how likely they are to become joined. Liben-Nowell and Kleinberg [2] set out the problem on coauthorship networks, built rankers from measures of how close two nodes sit in the network, and found that subtle proximity measures can outperform direct ones. One such measure is the Adamic-Adar index [3]. Kastrin, Rindflesch and Hristovski [4] applied proximity measures, among them common neighbours, the Adamic-Adar index and preferential attachment, to a network of co-occurring MeSH terms with its history of link formation, and reported that supervised methods outperform the unsupervised ones.

Krenn and Zeilinger [5] built a semantic network of quantum physics concepts from 750,000 papers, linked two concepts when one article studies both, and trained a neural network to predict which concept pairs researchers will study in the next five years. Krenn and colleagues [6] later built a benchmark, Science4Cast, from about 143,000 papers on artificial intelligence and compared ten prediction methods on it; in the authors' public summary, the most effective methods used a curated set of network features.

This study asks a narrower question that sits under all of that work. Before a gap ranking is trusted to point at

undiscovered connections, does it beat a ranking that knows only how popular each concept is? The shared-neighbour score in our test discounts popular concepts on purpose, in the form used for time-sliced tests by Krenn and Zeilinger. The baseline is preferential attachment, the product of the two concepts' degrees.

3 Preregistration and design

The design was written and hash-locked on 31 October 2026 as a preregistration, before any score of the study was computed. Exploration before the lock read schemas, field names and date ranges only. The locked file names the corpora, the slicing, the vocabulary rules, the definition of a link and of a target, the ranker and its baseline, both controls, the measures, four verdicts, the forecasts and one kill condition. The preregistration's SHA-256 is given in the reproduction section.

Each verdict is a rule read off by code:

- Beats popularity: the bootstrap interval of the shared-neighbour precision minus the popularity precision lies above zero on the research notes.
- Beats controls: the shared-neighbour precision on the research notes lies above the 95th percentile of both controls.
- Holds with popularity fixed: the within-decile area under the curve on the research notes lies above 0.55.
- Replicates on Biology: the same interval as the first verdict lies above zero on Biology Stack Exchange.

The kill condition reads: if the ranking fails to beat popularity and also fails with popularity fixed, absent-link scores say nothing beyond how common a word is, and its top gaps get reported only as prompts for reading.

4 Data and provenance

Three corpora were used. The first is our own research notes: entries of our findings ledger, the titles of saved readings, a catalogue of laws and a table of checks, each dated. They are internal records and are not public, so no one outside can reproduce that arm. Their dates span days of reading, not years, so a slice of them is a stretch of attention, not a stretch of a field's history.

The other two are the question posts of Biology Stack Exchange (28,798 questions) and Skeptics Stack Exchange (11,202 questions), 2011 to 2024, taken from the sites' public data dump. Each question's tags serve as its concepts. Stack Exchange content is licensed CC BY-SA [7]; this paper reports statistics only.

Each corpus was ordered by date and cut into five time slices of equal document count. The fifth slice was held back for a later study and was not read here. A vocabulary was built from the training slices only: at most 1,500 concepts, each appearing in at least four documents and in no more than 10 percent of them, with each document keeping at most 40 concepts by tf-idf weight. On the research notes the concepts are single words and word pairs; on the Stack Exchange sites they are tags.

A link is two concepts in one document. A target is a pair with no link in the training slices that appears together in at least two documents of the test slice.

5 Methods

Folds

Three folds trained on slice 1, slices 1 to 2 and slices 1 to 3, and each tested on the slice after its training block. Every number in the results is the mean over the three folds unless a fold is named.

The rankers

Let A be the training graph, d_i the degree of concept i, and E the number of edges. Each edge between i and k

gets a weight that falls as both ends become popular, floored at zero:

$$w_{ik} = \ln \frac{2E}{d_i d_k} \tag{1}$$

The A-B-C ranking scores an unlinked pair i and j by the cosine of their weighted neighbour rows, so that it is high when the two concepts share specific neighbours:

$$S_{ij} = \frac{\sum_k w_{ik} w_{jk}}{\sqrt{\sum_k w_{ik} w_{ik}} \sqrt{\sum_k w_{jk} w_{jk}}} \tag{2}$$

The popularity baseline is

$$P_{ij} = d_i d_j \tag{3}$$

The Adamic-Adar index was reported beside them as a secondary measure. It was not in the preregistration, and its result is read as a lead only:

$$AA_{ij} = \sum_k \frac{1}{\ln d_k} \tag{4}$$

where the sum runs over the neighbours k that i and j share. Ties in any score are broken by a fixed small noise term, the same in every run.

Measures

The main measure is precision in the top 200:

$$P_{200} = \frac{h}{200} \tag{5}$$

where h is the number of targets among a ranker's 200 highest-scored unlinked pairs. It is set beside the base rate, the share of all unlinked candidate pairs that are targets.

The interval for the difference in precision comes from a document bootstrap: 200 times, the test slice's documents are reweighted by a multinomial draw, the targets are recomputed, and both fixed top-200 lists are rescored. The difference is averaged over the three folds within each draw, and the 2.5th and 97.5th percentiles of the 200 averages form the 95 percent interval.

The second measure holds popularity fixed. Candidate pairs are split into tenths by the product of their degrees, and the area under the ROC curve is computed inside each tenth from 40,000 sampled pairs of a target and a non-target. The within-decile AUC is their weighted mean:

$$A_{strat} = \frac{\sum_b n_b A_b}{\sum_b n_b} \tag{6}$$

where A_b is the area in tenth b and n_b its number of targets.

6 Controls and planted checks

Two controls were run on the research notes, 20 draws each, scored against the true targets of each fold. Degree-preserving rewiring attempts two random double-edge swaps per edge on the training graph, skipping any swap that would duplicate an edge. Every concept keeps its popularity and loses its real neighbourhood. If the shared-neighbour score finds anything beyond popularity at the top of the list, it must beat this control. Node-label permutation keeps the graph's shape and breaks its tie to the real concepts. It is the weaker control: a score that loses to it would be worse than noise. The 95th percentile of each control's null distribution is the bar.

There was no planted check. No corpus with a known set of future links was built, so the study cannot say how often the instrument misses a real effect.

7 Results

Table 1 gives the main measure on each corpus and Figure 1 draws it. Ranking unlinked pairs by shared neighbours found fewer later links than ranking them by popularity on all three corpora: precision in the top 200 of 0.127 against 0.268 on our research notes, 0.018 against 0.088 on Biology Stack Exchange and 0.007 against 0.043 on Skeptics Stack Exchange. All three bootstrap intervals for the difference lie below zero.

Table 1. Precision in the top 200, mean of three folds, with the document-bootstrap interval for the shared-neighbour score minus popularity (200 draws). The last column is the within-decile area under the curve of the shared-neighbour score.

Corpus	Shared neighbours	Popularity	Adamic-Adar	Base rate	Difference interval	Within-decile AUC
Research notes	0.127	0.268	0.273	0.032	-0.182 to -0.125	0.569
Biology	0.018	0.088	0.122	0.0038	-0.105 to -0.062	0.757
Skeptics	0.007	0.043	0.055	0.0026	-0.077 to -0.042	0.696

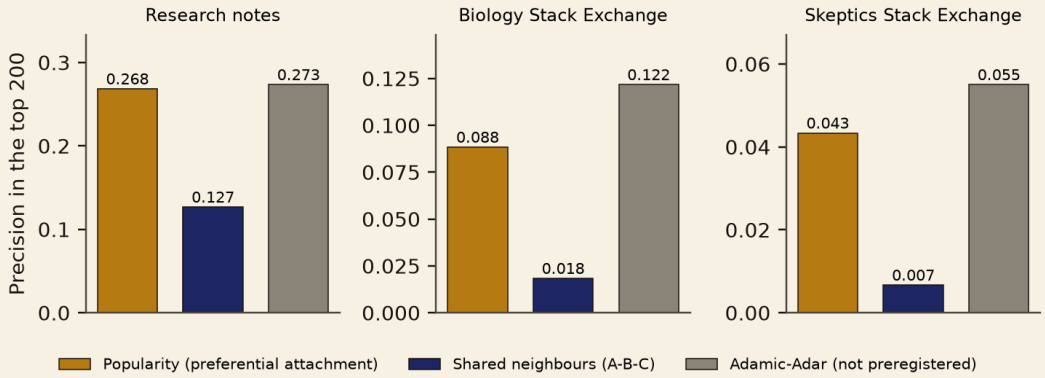


Figure 1. Precision in the top 200 for each ranker on each corpus, mean of three folds. Note the different vertical scales. Drawn from the run's result file.

Figure 2 sets the research-notes score against its controls. On degree-preserving rewired graphs the shared-neighbour score reached a 95th percentile of 0.164, above its real score of 0.127, so at the top of the list it added nothing we could detect beyond popularity. The rewired draws had a mean of 0.146, and 19 of 20 met or beat the real score (p value from draws 0.95). The score did beat the node-permuted graphs, whose 95th percentile was 0.048 (p 0.048, the smallest value 20 draws allow).

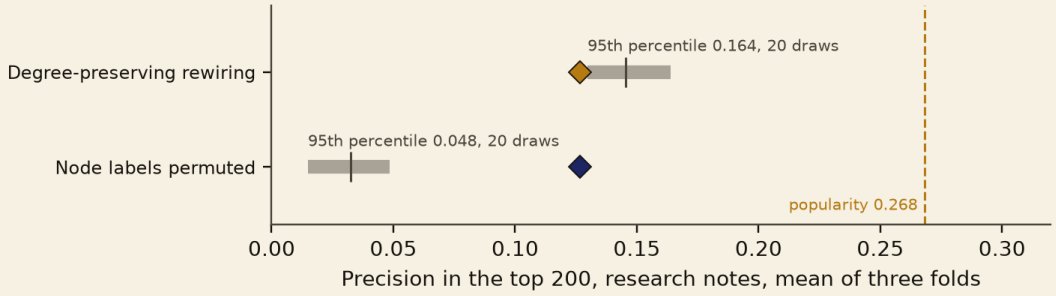


Figure 2. The shared-neighbour score on the research notes (diamonds) against the 5th to 95th percentile range (bars) and mean (ticks) of each control, 20 draws each. The dashed line is the popularity baseline.

Among pairs of similar popularity the shared-neighbour score still separated later links from the rest: area

under the curve 0.569 on the research notes, 0.757 on Biology and 0.696 on Skeptics. Over all candidate pairs, without holding popularity fixed, its area was 0.739 on the research notes against 0.782 for popularity, 0.810 against 0.815 on Biology, and 0.762 against 0.836 on Skeptics.

Adamic-Adar, not preregistered, matched or beat popularity: 0.273 against 0.268 on the research notes and 0.122 against 0.088 on Biology. No interval was computed for that comparison.

The folds explain much of the research-notes result. Table 2 shows that the first fold, with 285 concepts and a base rate of 0.094, holds almost all of the hits. The later folds have 1,500 concepts and base rates near 0.001, and the shared-neighbour score found no target in its top 200 in either of them.

Table 2. Each fold separately. Concepts and targets are counts; the last three columns are precision in the top 200, with A-B-C the shared-neighbour score.

Corpus and fold	Concepts	Targets	Candidates	Base rate	A-B-C	Popularity	Adamic-Adar
Research notes, 1	285	2,889	30,841	0.094	0.380	0.750	0.775
Research notes, 2	1,500	1,054	966,650	0.0011	0.000	0.040	0.030
Research notes, 3	1,500	1,378	979,245	0.0014	0.000	0.015	0.015
Biology, 1	378	487	66,577	0.0073	0.040	0.150	0.205
Biology, 2	540	371	136,117	0.0027	0.010	0.095	0.125
Biology, 3	609	218	172,390	0.0013	0.005	0.020	0.035
Skeptics, 1	252	158	29,894	0.0053	0.010	0.065	0.100
Skeptics, 2	391	101	72,530	0.0014	0.010	0.030	0.030
Skeptics, 3	476	123	107,553	0.0011	0.000	0.035	0.035

Table 3 sets the forecasts written before the run against the outcomes.

Table 3. Forecasts fixed in the preregistration against what the run gave.

Quantity	Forecast	Outcome
Research notes, popularity precision	0.25	0.268
Research notes, shared-neighbour precision	0.15	0.127
Research notes, base rate	0.02	0.032
Within-decile AUC, research notes	0.62	0.569
Biology, popularity precision	0.03	0.088
Biology, shared-neighbour precision	0.03	0.018
Beats popularity, research notes	probability 0.25	no
Beats popularity, Biology	probability 0.3	no
Beats both controls	probability 0.85	no

8 Discussion

Kastrin and colleagues [4] carry preferential attachment among their unsupervised measures, and our test adds a narrow fact to that line of work: a popularity-discounted shared-neighbour score falls below it at the top of the list on these three corpora. The rewired control sharpens that. A graph that keeps every concept's popularity and scrambles its neighbours gives the shared-neighbour score a higher top-200 precision than the real graph does. On the research notes, then, the neighbourhoods the score reads add nothing at the top that popularity alone does not carry.

The within-decile result points the other way, and both can hold at once. Inside a band of similar popularity, shared neighbours do separate future pairs from the rest. The discount in eq. 1 pushes popular pairs down the list, and popular pairs are where most of the later links are. A gap tool built on shared neighbours may therefore be useful as a re-ranker within a popularity band, which this study did not test.

The Adamic-Adar result is the lead worth a new test. It weights shared neighbours by their rarity without discounting the pair itself, and on two corpora it met or passed the popularity baseline. That was not preregistered

and carries no interval here.

Krenn and Zeilinger [5] and Krenn and colleagues [6] predict concept pairs with richer features and learned models. Our result does not bear on those methods. It bears on any tool that shows shared-neighbour gaps to a reader as candidate discoveries without first showing how a popularity list would do.

9 Limitations

- A later co-occurrence is an attention event. A hit means two concepts were later used together; it does not show that a link between them is true or new to science.
- On the research notes almost all hits come from the first fold, with 285 concepts. Later folds ran at base rates near 0.001, where precision in the top 200 rests on a handful of pairs.
- The research notes span days of reading rather than years, come from a few reading batches, and are not public. The two Stack Exchange corpora carry the reproducible part of the result.
- Both controls were run on the research notes only. The rewired control was not run on the public corpora.
- One tokenizer was used on the research notes, with no synonym handling; tags on the Stack Exchange sites are chosen by users and moderators and merge synonyms by their own rules.
- The study does not show that Adamic-Adar is the better method. It was not preregistered.
- There was no planted check, so the study has no measure of how often the instrument misses a real effect.
- Scoring was done by our own code and was not blind.

10 Verified

Verified: on 9 November 2026 every precision value, base rate, fold count, bootstrap interval, area under the curve, control mean, percentile and p value in this paper was re-read from the run's result file. The headline values match the findings ledger. The preregistration's SHA-256 was recomputed from the locked file and matches the hash recorded in the result file. The question counts for the two Stack Exchange sites were counted from the extracted question file the run read. Both figures were drawn by a script that reads only the result file. The study was not re-run for this paper. The literature in the introduction was found by searches run on 9 November 2026 and read at the level of abstracts and publisher pages; the full texts of the cited papers were not read.

11 Corrections

- 9 November 2026: our earlier write-up called Adamic-Adar's 0.122 on Biology a win over popularity. It is a higher point value with no interval, and this paper says so.
- 9 November 2026: first published version.

12 Reproduction

- Data: the public Stack Exchange data dump for the Biology and Skeptics sites (CC BY-SA). From each site's posts file keep question posts only, with their id, creation date and tags.
- Order each site's questions by date and cut them into five slices of equal count. Build the vocabulary from the training slices: tags in at least four documents and at most 10 percent of them, the 1,500 most frequent kept.
- For folds training on slice 1, slices 1 to 2 and slices 1 to 3, build the training co-occurrence graph and list as targets the unlinked pairs that co-occur in at least two documents of the next slice.
- Score candidates with eqs. 2, 3 and 4, compute eq. 5 and eq. 6, and take the document-bootstrap interval of the difference with 200 draws.

- Build the rewired control with two double-edge swaps per edge and score the shared-neighbour ranking on it against the true targets, 20 draws.
- Preregistration SHA-256: 9ccb8658ff6af3c02a9e7d3e6185577c4ba835892b50ed43fd9f4c6baf263b92. Plan SHA-256: bc0c19899c7bb9f51879f64bdd951f523953362060ecb0a2b5b84b39cc7d15c2. Run 20261031T065109Z, seed 20261031.
- Cost: the run's timing record gives 29 seconds for the analysis and 112 seconds for the two controls, on one desktop CPU.

13 Appendix A. The preregistered verdicts

Each verdict is a set of rules written before the run. A verdict passes only when every rule holds. Values are read from the run's result file.

Table 4. Preregistered verdicts and the rule that decided each one.

Verdict	Deciding rule	Value	Holds
Beats popularity	lower end of the research-notes difference interval above 0	-0.182	no
Beats controls	research-notes precision above the rewired 95th percentile (0.164)	0.127	no
Beats controls	research-notes precision above the permuted 95th percentile (0.048)	0.127	yes
Holds with popularity fixed	research-notes within-decile AUC above 0.55	0.569	yes
Replicates on Biology	lower end of the Biology difference interval above 0	-0.105	no
Kill condition	fails to beat popularity and fails with popularity fixed	second part holds	not met

The verdict "Beats controls" needs both of its rules and so fails. The kill condition did not fire, because the within-decile result passed.

14 Glossary

Each term is defined where it is first used in the text. The same definitions, with the field each comes from and what each excludes, are collected here.

- **literature-based discovery**. The search for connections that no single document states but that follow from linking what separate documents say. Field: information science, after Swanson. Excludes finding links that some document already states, which is ordinary retrieval.
- **link prediction**. Ranking pairs of nodes that are not yet joined in a network by how likely they are to become joined later. Field: network science. Excludes recovering links that exist but were hidden at random from the same snapshot.
- **preferential attachment**. A ranking of unlinked pairs by the product of the two concepts' degrees, so the most popular pairs come first. Field: network science. Excludes any use of which neighbours the two concepts share.
- **preregistration**. A dated written plan for a study, fixed and hash-locked before any result is seen: the question, the data, the statistic, the controls, the forecasts and the kill conditions. Field: research method (metascience). Excludes a plan revised after looking at the outcome, and a protocol written only for the record after the run.
- **control**. A version of the data with the structure under test removed (for example shuffled or pooled), scored the same way as the real data, so the real score has something to beat. Field: statistics and experimental design. Excludes a comparison with a different dataset or with a published number.
- **verdict**. The pass or fail of one preregistered rule set against the run's numbers, read off by code. Field:

research method (our usage). Excludes an interpretation written after reading the numbers.

- **kill condition**. A rule written in the preregistration that, when met, decides against the claim with no room for judgement after the run. Field: research method (our usage, after Popper's falsification). Excludes reasons found after the run for setting a result aside.
- **findings ledger**. The append-only, dated list of every result we have recorded, numbered F1, F2 and on; corrections are added as new dated lines and old entries are never rewritten. Field: our records. Excludes a curated list of successes.
- **time slice**. One of five blocks of equal document count cut from a corpus ordered by date; a fold trains on the early slices and tests on the next one. Field: forecast evaluation. Excludes random splits, which let later documents inform the training graph.
- **link**. Two concepts appearing in the same document; a concept here is a question tag on the Stack Exchange sites and a word or word pair on the research notes. Field: our usage for this study. Excludes any judgement that the two concepts are related in fact.
- **target**. A pair of concepts never linked in the training slices that appears together in at least two documents of the next slice. Field: our usage for this study. Excludes pairs that appear together only once in the test slice, and any check that the later pairing is true.
- **degree**. The number of other concepts a concept is linked to in the training graph, our measure of how popular it is. Field: network science. Excludes how many documents mention the concept.
- **A-B-C ranking**. Scoring an unlinked pair A and C by the intermediate concepts B that each is linked to; here a cosine over shared neighbours with each edge weighted down when both of its ends are popular. Field: literature-based discovery, after Swanson. Excludes methods trained on known outcomes, and the expert reading that Swanson applied to each candidate.
- **Adamic-Adar index**. A count of the neighbours two nodes share in which each shared neighbour is weighted by one over the logarithm of its own degree. Field: network science. Excludes the popularity discount on edges that the A-B-C score uses.
- **precision in the top 200**. The share of a ranker's 200 highest-scored unlinked pairs that turn out to be targets. Field: information retrieval. Excludes how the ranker orders pairs below the top 200.
- **base rate**. The share of all unlinked candidate pairs that are targets, the precision a random ranking would reach on average. Field: statistics. Excludes the precision of any particular ranker.
- **document bootstrap**. Redrawing the test slice's documents with replacement, recomputing which pairs are targets, and rescoring the fixed top-200 lists each time, so the interval shows how much the result depends on which documents happened to arrive. Field: statistics (resampling). Excludes uncertainty from the training graph, which is held fixed.
- **95% interval**. The range that holds the middle 95 percent of the bootstrap or control draws for a number; we report it beside the number. Field: statistics. Excludes a statement that the true value lies inside with 95 percent probability.
- **area under the ROC curve**. The chance that a randomly chosen target scores above a randomly chosen non-target; 0.5 is chance and 1 is perfect. Field: statistics and machine learning. Excludes how the top of the list performs, which precision in the top 200 measures.
- **within-decile AUC**. The area under the ROC curve computed separately inside each tenth of the candidate pairs ranked by degree product, then averaged with weights equal to each tenth's number of targets. Field: our usage for this study. Excludes any advantage that comes from ranking popular pairs above unpopular ones.
- **degree-preserving rewiring**. Repeatedly swapping the ends of two randomly chosen edges, so every concept keeps its degree while who is linked to whom is scrambled. Field: network science, after Maslov and Sneppen. Excludes changes to any concept's popularity.

- **node-label permutation**. Shuffling which concept sits at which node of the training graph, so the graph's shape stays and its link to the real concepts is broken. Field: statistics (permutation test). Excludes any change to the shape of the graph.
- **null distribution**. The spread of scores a control gives over many random draws; its 95th percentile is the bar a real score must clear. Field: statistics. Excludes a theoretical distribution assumed rather than drawn from the data.
- **p value (from draws)**. The share of control draws scoring at or above the real score, with one added to the count and to the number of draws, so it is never zero. Field: statistics. Excludes a p value from a textbook formula that assumes a distribution.

15 References

[1] Swanson, D. R. (1986). Undiscovered public knowledge. The Library Quarterly 56(2), 103-118. doi:10.1086/601720

[2] Liben-Nowell, D. and Kleinberg, J. (2007). The link-prediction problem for social networks. Journal of the American Society for Information Science and Technology 58(7), 1019-1031. doi:10.1002/asi.20591

[3] Adamic, L. A. and Adar, E. (2003). Friends and neighbors on the web. Social Networks 25, 211-230.

[4] Kastrin, A., Rindflesch, T. C. and Hristovski, D. (2016). Link prediction on a network of co-occurring MeSH terms: towards literature-based discovery. Methods of Information in Medicine 55(4), 340-346. doi:10.3414/ME15-01-0108

[5] Krenn, M. and Zeilinger, A. (2020). Predicting research trends with semantic and neural networks with an application in quantum physics. Proceedings of the National Academy of Sciences 117(4), 1910-1916. doi:10.1073/pnas.1914370116

[6] Krenn, M., Buffoni, L., Coutinho, B. and others (2023). Forecasting the future of artificial intelligence with machine learning-based link prediction in an exponentially growing knowledge network. Nature Machine Intelligence 5, 1326-1335. doi:10.1038/s42256-023-00735-0

[7] Stack Exchange, Inc. Stack Exchange data dump, Biology and Skeptics sites, question posts 2011 to 2024. Licensed CC BY-SA.