We're trying to predict FUTURE discoveries by reading the ENTIRE history of astronomy papers! (No pressure...)

arXiv:2602.14335v1 [astro-ph.IM] 15 Feb 2026

PREDICTING NEW CONCEPT–OBJECT ASSOCIATIONS IN ASTRONOMY BY MINING THE LITERATURE

Jinchu Li
School of Computer Science,
Georgia Institute of Technology
Atlanta, GA 30332, USA
jinchu.li@gatech.edu

Yuan-Sen Ting
Department of Astronomy,
The Ohio State University
Columbus, OH 43210, USA
ting.748@osu.edu

Alberto Accomazzi

Tirthankar Ghosal
National Center for Computational Sciences,
Oak Ridge National Laboratory
Oak Ridge, TN 37831, USA

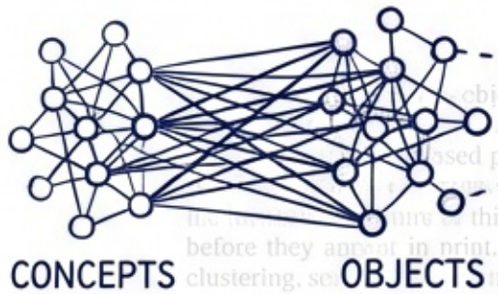
Argonne National Laboratory
Lemont, IL 60439, USA

Literature-based Crystal Ball

TL;DR SUMMARY

We built a GIANT "knowledge graph" linking astronomical OBJECTS (like stars, galaxies) to scientific CONCEPTS from tons of old papers.

Then we asked: Can the past connections help us guess NEW connections *before* they get published?



ABSTRACT

...object knowledge graph from the entire astro-ph corpus ...automated pipeline to extract named astrophysical objects...based papers...Solr...SIMBAD identifiers, and link ...annotated in ...years. We then ask whether the ...structure of this graph can forecast new concept–object associations before they appear in print. Because the underlying concepts are derived from clustering, ...ned concepts inevitably overlap; we address this by applying an ...framework ...rmly to all methods. Evaluating across four temporal cutoffs on a physically meaningful subset of concepts, we find that an implicit-feedback matrix factorization model (Alternating Least Squares: ALS) with concept smoothing outperforms the strongest neighborhood baseline (KNN based on text-embedding concept similarity) by 16.8% on NDCG@100 (0.144 vs. 0.123) and 19.8% on Recall@100 (0.175 vs. 0.146), and exceeds the best recency heuristic by 96% and 88%, respectively. These results indicate that historical literature encodes predictive structure not captured by either global heuristics or local neighborhood voting, suggesting a path toward tools that could help triage follow-up targets for costly telescope resources.

*Spoilers: **YES!** Our method is way better at guessing than just looking at what's similar.*

This could save astronomers a TON of time & money by telling them where to look first!

WHY BOTHER?

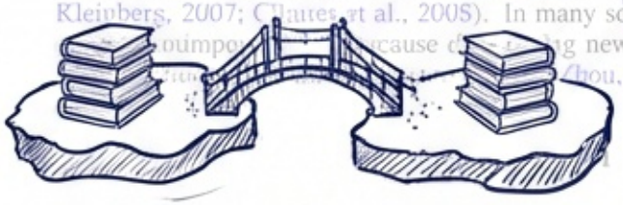
Because science is DROWNING in papers! Too many for any human to read.

→ difficulty of navigating a rapidly growing literature

→ undiscovered public knowledge

INTRODUCTION

...ot by a lack of ...ficulty of ...rature and identifying promising new connections. The idea of ...suggest such connections predates modern machine learning: classic literaiure-based discovery (LBD) frames "undiscovered public knowledge" as relationships implic... ...986; Henry & McInnes, 2017), while link prediction in evolving graphs formalizes which missing or ...e edges are likely to appear (Liben-Nowell & Kleinberg, 2007; Clauset et al., 2008). In many scientific settings, edges are missing not because ...because ...ng new interactions ...ceaffle; accurate link prediction ... Zhou, 2011).

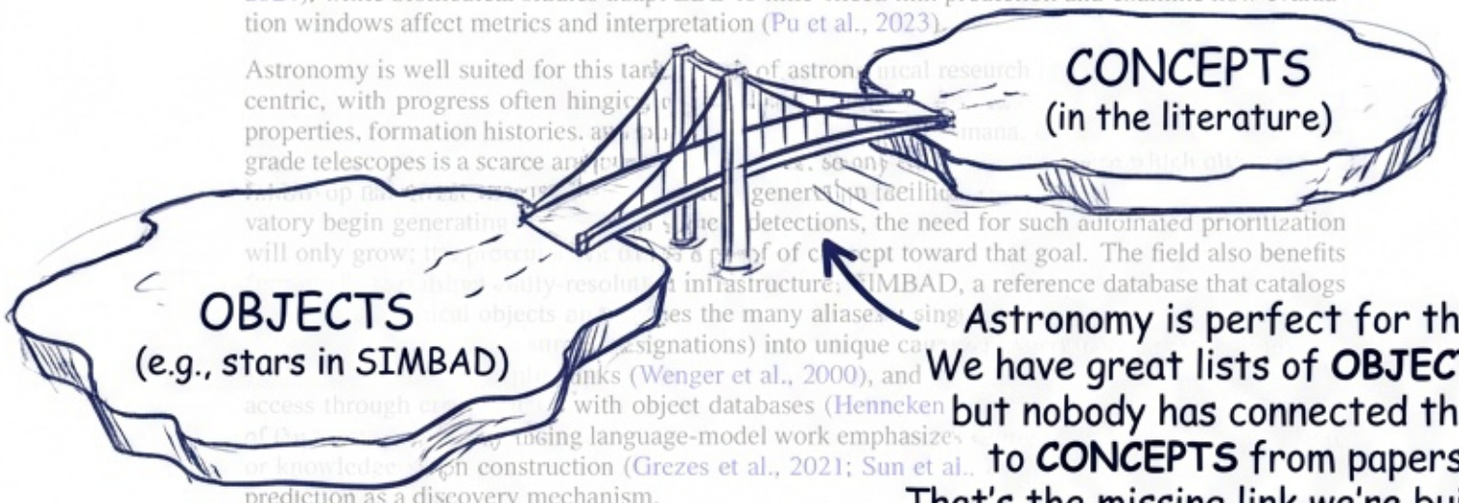


The answers might be hidden in plain sight, connecting two things nobody thought to put together!

So, what changed? **Better AI for reading!** Now computers can actually "understand" scientific text and build these complex knowledge graphs. (Before, it was just too hard!)

What has changed recently is the feasibility of building richer representations of scientific knowledge from text. Pretrained and large language models have improved extraction and representation for scientific corpora, enabling scalable pipelines for entity recognition, normalization, and relation mining (Devlin et al., 2019; Beltagy et al., 2019; Brown et al., 2020), and have been used directly for knowledge graph completion and inductive link prediction (Yao et al., 2019; Daza et al., 2021). These advances shift the bottleneck from whether we can construct literature-derived graphs to how we evaluate and use them for discovery.

Link forecasting on literature-derived graphs has accordingly become an active research direction. SemNet constructs evolving concept networks from scientific text and tests whether historical snapshots predict future connections (Krenn & Zeilinger, 2020); Science4Cast standardizes the protocol of training up to a temporal cutoff and predicting which links appear afterward (Aghajohari et al., 2021). More recently, Impact4Cast scales forecasting to tens of millions of papers (Gu & Krenn, 2024), while biomedical studies adapt LBD to time-sliced link prediction and examine how evaluation windows affect metrics and interpretation (Pu et al., 2023).



Astronomy is well suited for this tar[illegible] of astron[illegible] centric, with progress often hingi[illegible] properties, formation histories, an[illegible] grade telescopes is a scarce ar[illegible] vatory begin generating [illegible] will only grow; [illegible] infrastructure. SIMBAD, a reference database that catalogs [illegible] designations) into unique ca[illegible] links (Wenger et al., 2000), and [illegible] with object databases (Henneken [illegible] or knowledge [illegible] prediction as a discovery mechanism.

Astronomy is perfect for this! We have great lists of **OBJECTS**... but nobody has connected them to **CONCEPTS** from papers. That's the missing link we're building!

Despite these databases, no systematically constructed concept–object knowledge graph exists in astronomy, to our knowledge. SIMBAD catalogs objects but does not map them to the scientific concepts they relate to; ADS links papers to objects but does not provide structured concept–object associations. Building such a graph requires bridging [illegible] extraction and [illegible] resolution across the full literature.

PIONEERING WORK

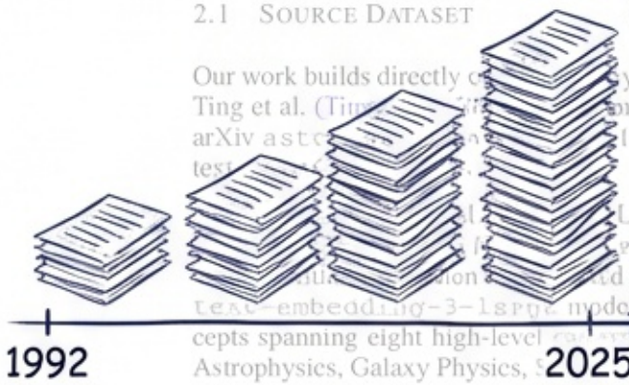
Our work addresses th[illegible] gap. We build an [illegible] to our knowledge, the first large-scale c[illegible] resolving object mentions to SIMBAD identifiers a[illegible] concepts across the [illegible] astro-ph corpus. As a proof of concept, we c[illegible] through a temporal [illegible] task: given all associations observed up to a cuto[illegible] year, we test whether [illegible] collaborative-filtering model can predict which objects will newly appe[illegible] with a given concept. Figure 1 provides an overview.

2 Dataset

To do this, we used A LOT of data. **Over 400,000 astronomy papers!** (from arXiv, spanning nearly 30 years)

2.1 Source Dataset

Our work builds directly [illegible] released, large-scale [illegible] Ting et al. (Tim[illegible] source corpus comprises 408,590 astrophysics papers from the arXiv `astro-ph` [illegible] 1992 through July 2025, each converted to machine-readable tex[illegible]



[illegible] LLM-based extraction pipeline to identify key scientific and [illegible] ver, yielding on average ∼10 raw concept mentions per paper. [illegible] ed via K-means clustering in embedding space (using OpenAI's `text-embedding-3-large` model), producing a controlled vocabulary of 9,999 unique concepts spanning eight high-level [illegible] (e.g., Cosmology & Nongalactic Physics, High Energy Astrophysics, Galaxy Physics, [illegible] & AI). The

We let an LLM loose on all these papers to automatically find the key concepts. (Because who has time to read 400k papers manually?!)

GPT-5-mini

Okay, we have the CONCEPTS. Now we need the OBJECTS. We ask the AI (GPT-5-mini) to read the papers and find them for us.

paper–concept associations, with each concept accompanied by a descriptive text definition and a fixed embedding vector.

We treat this concept vocabulary, its associated embeddings, and the paper–concept links as fixed inputs, and focus on augmenting the dataset with a new object-extr[illegible] object graph construction described below. We additionally use [illegible] for inference-time concept smoothing, which mitigates the s[illegible] derived concepts.

It also tells us *how* the object was used (was it the main subject? just a comparison?).

2.2 Object Extraction and SIMBAD Resolution

For each paper, we prompt GPT-5-mini with the title, abstract, and full OCR text to extract candidate astronomical object mentions intended for SIMBAD-style resolution. The extraction procedure is summarized in Appendix A; the full prompt and output schema are available in the Data and Code Availability section. For each extracted object, the model returns: (i) a single designation string, (ii) a semantic role describing how the object is used in the paper (e.g., primary subject, sample member, counterpart/host, calibration/reference), (iii) a study mode (new observations, archival/reanalysis, catalog compilation, theory/simulation, or incidental mention), and (iv) a short evidence span with its source (title/abstract/body).

Raw AI output

But AI can make mistakes. So we pass its answers through a "truth filter" (SIMBAD database) to make sure the objects are real and we have their correct names.

Raw outputs are cleaned and filtered to remove [illegible] bject entities (instruments, surveys, facilities) [illegible] o strings. We validate [illegible] against the [illegible] on constraint) and [illegible] fail validation. For the remaining object strings [illegible] each unique [illegible] canonical identifiers (Wenger et al. 2000); names that [illegible] mention instances removed. Across [illegible] 501 papers, the [illegible] tions (396,297 unique name strings) [illegible] after filtering. 1,621,480 mentions remain (376,692 unique names), of which 163,775 unique names (43.48%) successfully resolve, corresponding to 1,150,553 surviving mention instances (66.06% of raw mentions). After alias merging, we retain 100,560 unique SIMBAD objects across the corpus.

We ended up with ~100,000 unique objects!

2.3 Concept–Object Graph Construction

We combine the paper–concept associations from the source dataset with the paper–object associations extracted and resolved in the previous step to form a concept–object bipartite graph. An edge between concept c and object o exists if and only if at least one paper mentions o in connection with c. Each mention inherits the publication year of its source paper, enabling temporal analysis.

Strong link: Object was the MAIN subject of the paper.

CONCEPT (e.g., "star formation")

OBJECT (e.g., "M31")

Weak link: Object was just mentioned in passing.

[illegible] signal: an object that is the primary [illegible] observations is more informative than one mentioned only in [illegible] each concept–object edge a weight that reflects both how many papers sup[illegible] prominently each mention figures in its paper. As described in Section 2.2, [illegible] with a semantic role and a study mode. Role captures [illegible] (e.g., primary scientific target vs. [illegible] calibration/reference), while study mode captures the [illegible] of analysis [illegible] compilation, theory-only, or [illegible] flat [illegible] heuristically from domain intuition—for example, a paper [illegible] via new observations should receive more weight, compared to a mere [illegible] in a theoretical context.

Formally, the weight of a concept–object edge is computed by aggregating per-[illegible] and applying a log transfor[illegible]

$$w(c,o) = \ln\left(1 + \sum_{m \in \mathcal{M}(c,o)} a(m)\right), \tag{1}$$

where $\mathcal{M}(c,o)$ is the set of paper-level mentions linking concept c to object o. Each mention's strength is the product of a role weight and a study-mode weight:

$$a(m) = \rho_{r(m)}\,\gamma_{o(m)}, \tag{2}$$

where $r(m)$ and $o(m)$ are the role and study mode of mention m, and ρ and γ are fixed lookup weights ($\rho \in [0.25, 3.0]$, $\gamma \in [0, 1.25]$ [illegible] roles and

This scary-looking math is just a way to calculate that link strength based on how important the object was in the paper.

PUTTING IT ALL TOGETHER: THE COMPLETE PIPELINE!

Step 1 & 2: Extracting & Cleaning (recap from last page!)

Raw text -> AI -> Truth Filter -> Clean Objects

Step 3: Building the Knowledge Graph

We want to predict these "future" links!

Step 4: The Math Magic (Simplifying the Graph)

The graph is huge, so we smush it down into two smaller, manageable matrices using ALS (a type of machine learning).

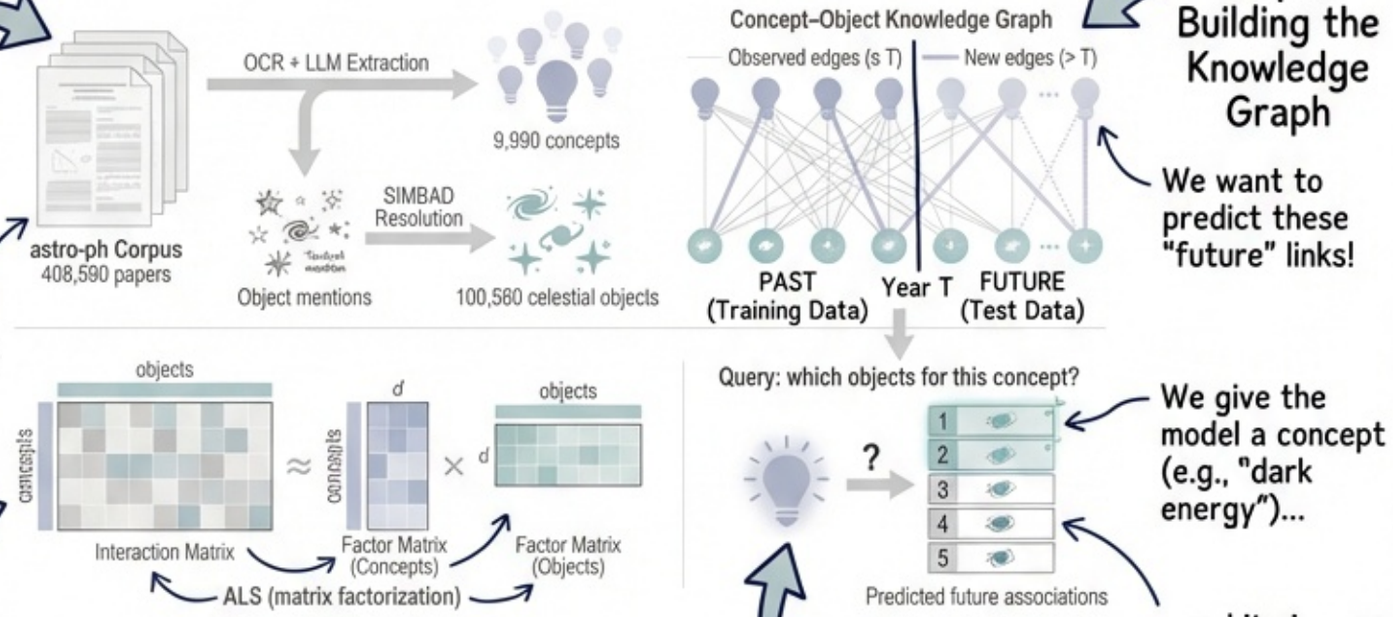


Figure 1: Pipeline overview. **Top left:** the astro-ph corpus (408,590 papers) is processed via OCR and LLM extraction to produce 9,999 concepts and raw object mentions, which are resolved through SIMBAD into 100,560 unique celestial objects. **Top right:** these are combined into a concept–object knowledge graph, split at a temporal cutoff T into observed edges ($\leq T$, gray) and new edges ($> T$, purple). **Bottom left:** ALS approximates the interaction matrix as a product of low-rank concept and object factor matrices. **Bottom right:** at inference, the model ranks candidate objects for a query concept to predict future associations.

We give the model a concept (e.g., "dark energy")...

...and it gives us a ranked list of objects it thinks will be linked to it in the "future".

Step 5: Making Predictions

The log transform compresses the dynamic range so that edges supported by many weak mentions do not dominate edges supported by a few strong ones. We additionally record a timestamp for each concept–object pair, defined as the earliest publication year among its supporting papers; this timestamp is what enables the temporal train/test split used in our evaluation (Section 3.1). Representative edges from the concept–object graph for two illustrative concepts, including aggregated edge weights and first-appearance years, are shown in Appendix C.

3 Methods

How do we know this works? We test its ability to "predict the future"!

The central question is whether the concept–object graph contains enough structure to anticipate future associations. We adopt a temporal holdout protocol: freeze the graph at a cutoff year T, train on all associations up to that point, and evaluate predictions of associations that first appear after T.

We train on the past...

PAST (Training Data) - $G_{\leq T}$ | Year T | FUTURE (Test Data) - $E_{>T}$

...and see if we can predict these future links!

3.1 Problem Definition

Let C denote the set of concepts and O the set of resolved astrophysical objects. From the graph $G = (C, O, E)$ constructed in Section 2, an edge $(c, o) \in E$ carries a positive weight $w_{c,o} > 0$ and a first-appearance year $y_{c,o}$.

Given a cutoff year T, we split the graph temporally: the training graph $G_{\leq T}$ contains only edges whose first-appearance year is at most T, with edge set $E_{\leq T}$, while the held-out set $E_{>T}$ contains edges appearing strictly after T. For each concept c, the task is to rank candidate objects $O_c^{\text{cand}} = \{o \in O \mid (c, o) \notin E_{\leq T}\}$—that is, objects not yet associated with c—so that the ones in $E_{>T}$ are ranked highly. For example, if the concept "High-Redshift Quasars" has been linked to a set of known quasars before T, the model should rank the quasars that will first appear in connection with this concept after T above the remaining candidates.

e.g., The model should know *beforehand* which quasars will become important for this concept.

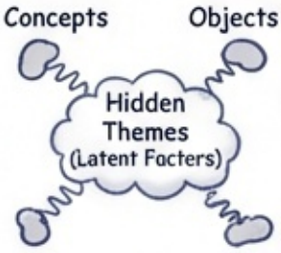
THE CORE ENGINE: MAKING RECOMMENDATIONS

3.2 Model and Baselines

The ranking problem defined above can be viewed as a recommendation task: given the historical concept–object associations, recommend which objects a concept is likely to be associated with next. We treat this as an implicit-feedback problem, meaning we observe which co… co-occurred (positive signal) but never observe explicit "negative" pairs—… to an object may simply reflect an association that has not yet appeared in p… is irrelevant. To work with standard matrix methods, we represent the graph … $W \in \mathbb{R}_{\geq 0}^{|C| \times |O|}$, where $W_{c,o} = w_{v,o}$ for observed edges and $W_{v,o} = 0$ for p… papers. The goal is to learn a scoring function that assigns high scores to zero… to become nonzero in the future.

Think of this like Netflix recommending your next binge-watch. We have patterns of past connections (positive signal) and want to predict *new* hits that haven't happened yet.

Matrix Factorization. Our primary model is implicit Alternating Least Squares (ALS) (Hu et al., 2008), a matrix factorization method. The core idea is that the interaction matrix W, despite being very large (0,999 concepts × 100,360 objects), has low-rank structure: the patterns of which concepts associate with which objects are governed by a much smaller number of latent factors (e.g., shared physical properties, observational techniques, or research themes). We approximate W by representing each concept c as a vector $p_c \in \mathbb{R}^d$ and each object o as a vector $q_o \in \mathbb{R}^d$ (where $d \ll |C|, |O|$ is the latent dimension), such that their dot product $p_c^\top q_o$ estimates the strength of their association. Because concepts that share similar objects are pushed toward similar vectors during training, the model can generalize to predict associations for concept–object pairs not seen in the training data.



Back to the "Math Magic" (ALS) from the last page! The assumption: deep down, there are hidden structures connecting things.

THE MATH BEHIND THE MAGIC (The Objective Function)

Concretely, the model minimizes … objective with ℓ_2 regularization:

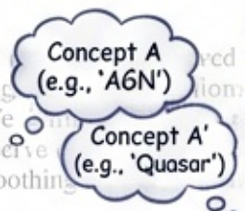
$$\min_{\{p_c\},\{q_o\}} \sum_{c \in C} \sum_{o \in O} \left(1 + \alpha w_{c,o}\right) \left(\mathbb{I}[w_{c,o} > 0] - p_c^\top q_o\right)^2 + \lambda \left(\sum_c \|p_c\|_2^2 + \sum_o \|q_o\|_2^2 \right), \qquad (3)$$

Try to match the links we've actually seen (weighted by confidence)…

…but keep the model simple so it doesn't just memorize the data ("overfitting").

where $\mathbb{I}[w_{c,o} > 0]$ is a binary indicator of whether an association has been observed, $\alpha = 10$ is a hyperparameter controlling how much edge weight amplifies confidence, and $(1 + \alpha w_{c,o})$ scales the confidence … that even unobserv… a bas… ively push the model tow… by default: … with confidence $1 + \alpha w_{c,o} > 1$ … edge weigh… … regularization strength $\lambda = 0.05$ pr… the latent dimension to $d = 128$ and run 30 ALS iterations. The predicted relevance score for a concept–object pair is then $s_{\text{ALS}}(c, o) = p_c^\top q_o$.

Concept-embedding similarity weights. Because the 9,999 … clustering, semantically related concepts inevitably overlap, frag… tifically coherent (see Appendix D for detailed discussion). We … ights that account for this overlap; these weights serve … a concept-aware KNN baseline and an inference-time smoothing …



HANDLING SIMILAR IDEAS: Sometimes different words mean almost the same thing. This helps the model share info between them.

HANDLING SIMILAR IDEAS: Like soarnar based on vconpireer aind the same relatives to describe other. Sometimes different words mean almost the same thing. This helps the model share info between them.

… the fixed concept embeddings $\{c_c\}$ from the source dataset, we form a k-nearest-neighbor … for each concept c and define nonnegative neighbor weights by clipping negative cosine … and ℓ_1-normalizing:

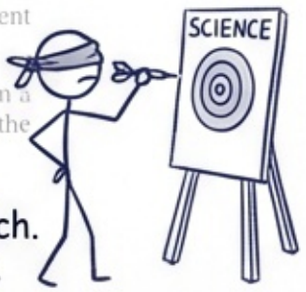
$$S_{c,c'} = \frac{\max\left(\cos(c_c, c_{c'}),\, 0\right)}{\sum_{c'' \in \mathcal{N}_k(c)} \max\left(\cos(c_c, c_{c''}),\, 0\right)}, \qquad (4)$$

for $c' \in \mathcal{N}_k(c)$, and $S_{c,c'} = 0$ otherwise. If all similarities in $\mathcal{N}_k(c)$ are non-positive, we fall back to uniform weights over $\mathcal{N}_k(c)$.

THE COMPETITION (Things we have to beat)

Baselines. We compare ALS against several non-parametric baselines, each testing a different hypothesis about what drives future associations.

Random ranks candidate objects uniformly at random, assigning each object a score drawn from a uniform distribution: $s_{\text{rand}}(o) \sim \text{Uniform}(0, 1)$. This establishes a lower bound and verifies that the evaluation protocol behaves as expected.



1. RANDOM GUESSING: The "monkey throwing darts" approach. Sets the absolute bottom bar.

The Contenders & The Rules of the Game

The Lineup of Me're comparing

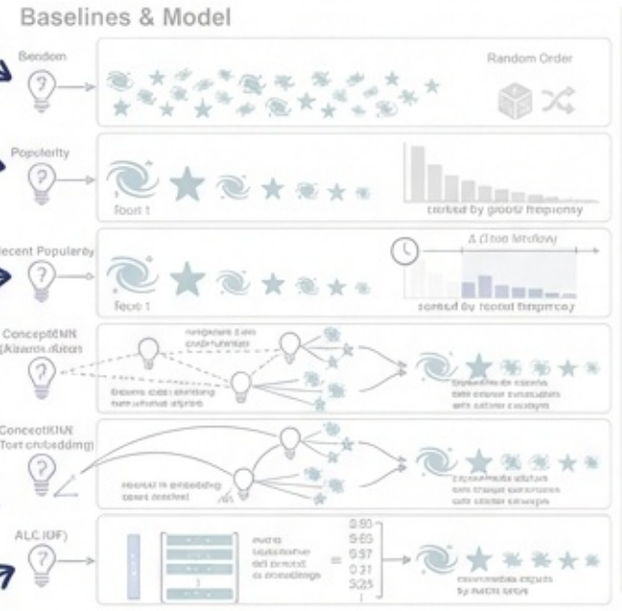
The 'monkey' baseline (see prev. page)

The 'Bandwagon' approach: just recommend what's already famous.

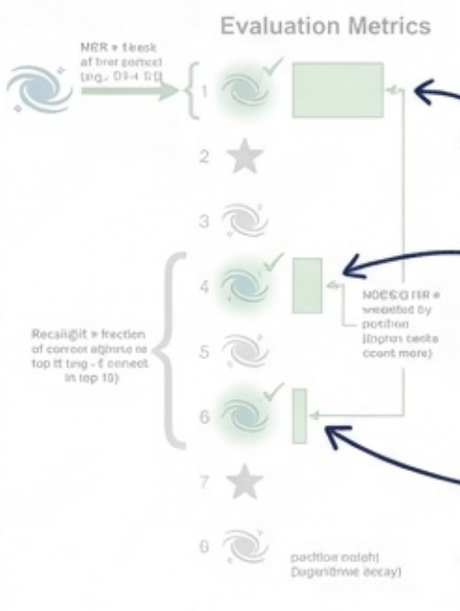
The 'Trendy' approach: what's famous *lately*.

The 'Copycat' approaches: find similar concepts and borrow their associations.

Our 'Matrix Factorization' star (from prev. page).



How we keep score



Did you get the right answer *quickly*? (Higher is better).

Did you find *all* the right answers in your top K picks? (Coverage).

The gold standard: Cares about finding the right answers AND putting them at the top.

Figure 2: Visual overview of baselines and evaluation metrics. **Left:** how each method ranks candidate objects for a query concept—from top to bottom: Random (shuffled ordering), Popularity (global frequency), RecentPopularity (time-windowed frequency), ConceptKNN-AA (Adamic–Adar neighbor aggregation), ConceptKNN-TextEmb (embedding-based neighbor aggregation), and ALS (dot product of learned latent factors). **Right:** how the three evaluation metrics score a ranked list; green checkmarks denote correct held-out associations. MRR rewards placing the first correct object high, Recall@K measures coverage in the top K, and NDCG@100 assigns position-discounted credit.

Lazy! Gives the same recommendations to everyone, regardless of the concept (concept-agnostic).

Popularity tests whether globally prominent objects are more likely to appear in future associations regardless of the query concept, scoring each object by its total training edge weight $s_{\text{pop}}(o) = \sum_{c \in C} w_{c,o}$. The resulting ranking is concept-agnostic.

RecentPopularity ($\Delta \in \{3, 5\}$ years) restricts the sum to papers published in $(T-\Delta, T]$, testing whether short-term trends carry more signal than lifetime popularity. Like Popularity, it is concept-agnostic.

Smarter! Tailors recommendations based on the specific concept (concept-aware).

ConceptKNN family. We define a concept-aware non-parametric family of baselines that borrow evidence from similar concepts. For a query concept c, we form a neighbor set $\mathcal{N}_b(c)$ under a chosen concept–concept similarity and score candidate objects by

$$s_{\text{kNN}}(c, o) = \sum_{c' \in \mathcal{N}_b(c)} \tilde{s}(c, c')\, w_{c',o}, \tag{5}$$

where $\tilde{s}(c, c')$ are nonnegative neighbor weights normalized over $\mathcal{N}_b(c)$. We treat k as a hyperparameter and report results across multiple values. We use two instantiations:

Uses shared connections in the data.

(i) **ConceptKNN_AA (graph-based).** We set similarity using the Adamic–Adar index (Adamic & Adar, 2005):

$$s_{\text{AA}}(c, c') = \sum_{o \in \mathcal{O}(c) \cap \mathcal{O}(c')} \frac{1}{\log |\mathcal{C}(o)|}, \tag{6}$$

and define $\tilde{s}(c, c')$ by normalizing $s_{\text{AA}}(c, c')$ over $c' \in \mathcal{N}_b^{\text{AA}}(c)$.

Uses similar meanings from text descriptions.

(ii) **ConceptKNN_Emb (text-embedding-based).** We set $\tilde{s}(c, c') = S_{c,c'}$, the concept-embedding similarity weights defined in equation 4. This baseline scores objects *entirely* by aggregating the associations of semantically similar concepts.

Training and inference. For each cutoff T, we train all methods on the time-filtered graph $G_{\leq T}$, keeping the concept and object vocabularies fixed across cutoffs. The ALS model is fit on the

A quick note on fair play: we train on past data and test on future data. Standard ML practice.

Final "Rules of the Game" & The Official Scoring Rulebook

A few more "ground rules" to keep the fight fair...

Fine-tuning our ALS model's "engine" before the race. We used a small grid search to find the best settings.

corresponding interaction matrix to produce latent vectors for all concepts and objects; baselines compute their scores directly from training-graph statistics. At inference time, for each concept c we rank all objects in the candidate set O_c^{cand} by their predicted score $s(c, o)$. The same candidate masking—excluding objects already associated with c in $G_{\leq \mathcal{T}}$—is applied to all methods, ensuring a fair comparison. We select ALS hyperparameters (d, α, λ) via a small grid search on a temporally valid development setting (using cutoff $T{=}2021$ for selection), and then fix $d{=}125$, $\alpha{=}10$, and $\lambda{=}0.05$ for all reported cutoffs. For ConceptKNN baselines, we report results across $k \in \{3, 10, 25, 50, 100, 150, 200\}$. Additionally, the concept-embedding weights enable smoothing any method's predictions at inference time. Given a base score $s(c, o)$, we blend it with neighbor-aggregated scores:

$$s_{\text{smooth}}(c, o) = (1 - \beta)\, s(c, o) + \beta \sum_{c' \in \mathcal{N}_k(c)} S_{c,c'}\, s(c', o), \tag{7}$$

Giving the "Copycat" (ConceptKNN) a helping hand: This "smoothing" blends its score with its neighbors, making it a stronger competitor.

where β controls the mixing strength. This preserves each method's learned structure while compensating for concept-boundary artifacts. We apply smoothing uniformly to all methods, tuning (k, β) at $T{=}2021$ and fixing $k{=}100$, $\beta{=}0.5$ for all reported experiments. Results without smoothing are in Appendix D.

The Official Scoring Rulebook (The Math-y version of the previous page!)

3.3 Evaluation Protocol

Evaluation is restricted to concepts with at least 10 training associations and at least one held-out test edge in $E_{>\mathcal{T}}$. The first condition ensures that concepts have enough history for the model to learn from; the second ensures a meaningful prediction target.

You must have *some* experience to enter this competition. No newbies!

We assess ranking quality with four standard metrics, where C_{eval} is the $E_{>\mathcal{T}}(c)$ the held-out objects for concept c:

Mean Reciprocal Rank (MRR): the reciprocal rank of the first correct held-out object, measuring how quickly a correct result appears:

$$\text{MRR} = \frac{1}{|C_{\text{eval}}|} \sum_{c \in C_{\text{eval}}} \frac{1}{\min_{o \in E_{>\mathcal{T}}(c)} \text{rank}_v(o)}. \tag{8}$$

Remember: Speed is key! 1st place is WAY better than 2nd.

Recall@K (with $K \in \{10, 100\}$): the fraction of held-out objects appearing in the top K predictions:

$$\text{Recall@}K = \frac{1}{|C_{\text{eval}}|} \sum_{c \in C_{\text{eval}}} \frac{|\{o \in E_{>\mathcal{T}}(c) : \text{rank}_v(o) \leq K\}|}{|E_{>\mathcal{T}}(c)|}. \tag{9}$$

Did your net catch the right fish? (How many good answers are in your top K?)

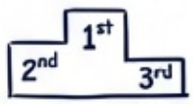
NDCG@100: a ranking-quality measure that awards more credit for correct objects placed near the top (Järvelin & Kekäläinen, 2002):

$$\text{NDCG@100} = \frac{1}{|C_{\text{eval}}|} \sum_{c \in C_{\text{eval}}} \frac{\text{DCG@100}(c)}{\text{IDCG@100}(c)}, \quad \text{DCG@100}(c) = \sum_{i=1}^{100} \frac{\mathbb{I}[\mathcal{R}_v(i) \in E_{>\mathcal{T}}(c)]}{\log_2(i+1)}, \tag{10}$$

where

$$\text{IDCG@100}(c) = \sum_{i=1}^{\min(|E_{>\mathcal{T}}(c)|, 100)} \frac{1}{\log_2(i+1)} \tag{11}$$

is the ideal DCG obtained by ranking all held-out objects first. All metrics are macro-averaged over concepts; no test-edge information is used during training.

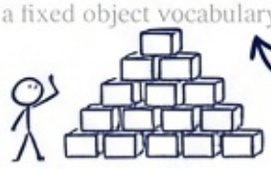


Are the BEST answers at the very TOP? Ranking order *really* matters here.

4 Results

We evaluate concept–object edge prediction across four cutoff years $T \in \{2017, 2019, 2021, 2023\}$. Depending on the cutoff, the training graph contains 2.11M–2.90M concept–object edges, and the held-out set contains 0.26M–1.0SM edges. Evaluation is restricted to 7.2k–7.7k eligible concepts per cutoff (each with at least 10 training associations and at least one held-out test association), with a fixed object vocabulary of 100,560 SIMBAD-resolved objects.

Running the race at different points in time!



Setting the stage for the MAIN EVENT! We're testing on a MASSIVE scale across different years to be thorough.

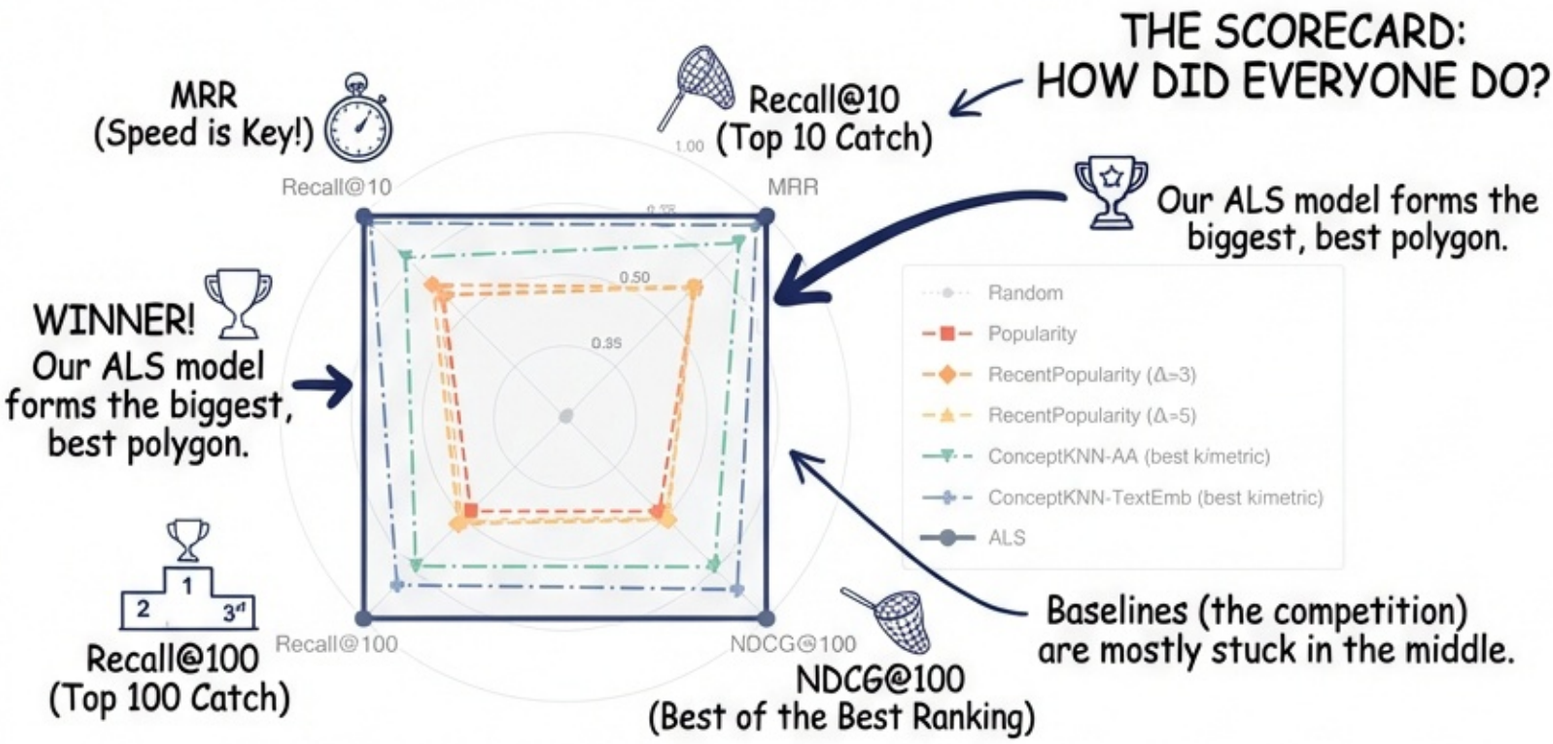


Figure 3: Radar plot comparing all methods on the physical concept subset with concept smoothing. The four axes correspond to MRR, Recall@10, Recall@100, and NDCG@100, each normalized so the best method equals 1.0. Methods shown: Random (gray dotted), Popularity (red dashed), RecentPopularity with $\Delta \in \{3, 5\}$ (orange/yellow dashed), ConceptKNN-AA (teal dash-dot), ConceptKNN-TextEmb (blue dash-dot), and ALS (dark solid). ALS forms the outermost polygon, leading on all four metrics.

THE DETAILED SCORE SHEET
(The Nitty Gritty)

THE VICTORY MARGIN
(Just HOW much better?)

(a) Absolute performance (with inference-time concept smoothing)

Method	MRR	Recall@10	Recall@100	NDCG@100
Random	0.0061±0.0020	0.0002±0.0001	0.0010±0.0001	0.0011±0.0003
Popularity	0.2031	0.0278	0.0817	0.0678
RecentPopularity (Δ=3)	0.2031	0.0303	0.0827	0.0734
RecentPopularity (Δ=5)	0.2024	0.0282	0.0902	0.0719
ConceptKNN-AA (best k per metric)	0.2724	0.0365	0.1294	0.1070
ConceptKNN-TextEmb (best k per metric)	0.3040	0.0445	0.1457	0.1230
ALS	**0.3150±0.0010**	**0.0460±0.0001**	**0.1746±0.0002**	**0.1436±0.0001**

The "dart-throwing monkey" (Random) is predictably terrible.

Our Champion (ALS) dominates on EVERY single metric!

(b) Relative improvement of ALS over baselines (%)

Baseline	MRR	Recall@10	Recall@100	NDCG@100
Random	5054.48	20694.20	18176.33	12756.58
Popularity	55.13	65.43	113.75	112.66
RecentPopularity (Δ=3)	55.11	51.81	88.31	95.63
RecentPopularity (Δ=5)	55.62	62.90	93.60	99.55
ConceptKNN-AA (best k per metric)	15.62	26.05	34.94	34.20
ConceptKNN-TextEmb (best k per metric)	5.01	3.39	19.80	16.78

This table shows the improvement of ALS over the others. It's often HUNDREDS or THOUSANDS of percent better than the basic baselines! Even the "tougher competitors" (ConceptKNN) get left behind by 15-20%.

Table 1: Link prediction performance on the physical concept subset with concept smoothing, averaged across four cutoffs. ALS and Random are stochastic and reported as mean±std over seeds; all other baselines are deterministic. KNN baselines: best k per metric. Panel (b): relative improvement of ALS over each baseline (%).

Evaluation concept subset. Not all concept categories are equally informative for forecasting astrophysical associations. Concepts tied to instrument usage or broadly applicable methodologies yield associations driven more by scheduling or cross-field applicability than by astrophysical structure, diluting evaluation. We therefore report headline results on a *Physical* subset that excludes Statistics & AI, Numerical Simulation, and Instrumental Design concepts (details in Appendix B).

Focusing on the REAL science. We excluded concepts like "telescope" or "method" to see how well we predict actual astrophysical ideas. This makes the test harder and more meaningful!

THE TEST OF TIME:
Does it hold up year after year?

2017 → 2023

Method	2017	2019	2021	2023
Random	0.0010±0.0002	0.0009±0.0002	0.0008±0.0003	0.0012±0.0004
Popularity	0.0764	0.0812	0.0865	0.0826
RecentPopularity (Δ=3)	0.0841	0.0846	0.1001	0.1021
RecentPopularity (Δ=5)	0.0816	0.0845	0.0964	0.0982
ConceptKNN-AA (best k)	0.1201	0.1317	0.1412	0.1430
ConceptKNN-TextEmb (best k)	0.1324	0.1429	0.1547	0.1594
ALS	0.1463±0.0004	0.1596±0.0004	0.1766±0.0005	0.1851±0.0005

(b) NDCG@100

Method	2017	2019	2021	2023
Random	0.0015±0.0003	0.0012±0.0003	0.0009±0.0002	0.0008±0.0002
Popularity	0.0832	0.0744	0.0647	0.0478
RecentPopularity (Δ=3)	0.0909	0.0771	0.0704	0.0352
RecentPopularity (Δ=5)	0.0882	0.0775	0.0682	0.0336
ConceptKNN-AA (best k)	0.1369	0.1226	0.1063	0.0848
ConceptKNN-TextEmb (best k)	0.1363	0.1349	0.1193	0.0954
ALS	0.1617±0.0002	0.1459±0.0003	0.1306±0.0003	0.1070±0.0003

SPOILER: YES. ALS wins every single year.

Table 2: Per-cutoff performance on the physical concept subset without smoothing. ALS leads at every cutoff on both metrics. ALS and Random: mean±std over seeds; other baselines are deterministic. KNN baselines: best k per cutoff.

Table 1 and Figure 3 report performance averaged across all four cutoffs with concept smoothing applied uniformly to all methods. ALS is the strongest method on all four metrics, exceeding the best text-embedding KNN baseline by 5.0% on MRR, 3.4% on Recall@10, 19.8% on Recall@100, and 16.8% on NDCG@10). The gains over recency heuristics are substantially larger: ALS improves **margin is widest on long-horizon retrieval metrics** 00 and 96% on NDCG@100. The margin is widest on long-horizon retrieval metrics (Recall@100, NDCG@100)—precisely the regime most relevant to practical triaging, where a researcher scans a moderately sized candidate list. On early-rank metrics the advantage is smaller but still positive, indicating that matrix factorization captures complementary structure beyond local neighborhood voting. Results without smoothing (Appendix D) preserve the qualitative ordering but show a narrower ALS advantage on early-rank metrics; smoothing most strongly benefits ALS, consistent with the idea that learned latent factors propagate information more effectively once cluster-boundary artifacts are mitigated.

Why this matters: We're best at the "long game"—finding distant, non-obvious connections useful for "practical triaging".

Because aggregate improv **without smoothing, for a cleaner comparison of base methods.** Table 2 breaks down Recall@100 and NDCG@100 by cutoff year (without smoothing, for a cleaner comparison of base methods; smoothed per-cutoff trends are in Appendix F). ALS outperforms all baselin at every cutoff on both metrics, confirming that the gains are consistent across time. RecentPopularity becomes more competitive at later cutoffs but does not close the gap; KNN baselines improv steadily with T but remain behind ALS.

We took the "smoothing" filter off here to show the RAW performance. ALS still dominates.

THE BOTTOM LINE: Consistent wins, especially on the hard stuff (real physics concepts & long-horizon search).

5 Discussion, Implications

Summary of findings. With inference-time concept smoothing to mitigate the semantic overlap inherent in clustering-derived concepts, ALS consistently outperforms all baselines across every metric and temporal cutoff, with the largest gains on long-horizon retrieval measures most relevant to practical triaging. These improvements are concentrated in astrophysically meaningful concepts, indicating that the literature-derived graph captures predictive structure beyond local similarity or short-term trends.

Implications for literature-driven discovery. The br ing the published literature can surface laten ucture as SIMBAD and ADS—not by cataloging objects or relationship between scientific ideas and the objects

BIG IDEA: The literature isn't just a pile of papers. It's a giant, evolving MAP of knowledge. Our model helps draw that map.

A FINAL WORD OF CAUTION:
Our tool finds *expected* connections based on the past. But the BIGGEST scientific leaps often come from the **UNEXPECTED**! This is a tool to *help* your judgment, not replace it.

building block for tools that help astr... plored objects for a given resea... ch qu...

Limits of prediction and scientific judgment. At the ... tool based on historical patterns carries an inherent tension. While our results show that future concept–object associations are statistically forecastable from past literature, some of the most consequential discoveries in astronomy arise precisely from unexpected associations—objects that no one predicted would be relevant to a given research question. A triaging system that ranks objects by expected relevance risks reinforcing existing research trajectories at the expense of serendipitous findings. Balancing systematic prioritization with openness to the unexpected is not a problem that any algorithm can resolve on its own; it is ultimately a matter of scientific judgment and community practice. We view the graph presented here as a tool to inform that judgment, not to replace it.

Methodological limitations and design choices. The concept–object graph depends on OCR quality, LLM-based extraction accuracy, and SIMBAD ... In addition, our evaluation targets are defined by ... ground-truth notion of physical discovery, so pred... and publication dynamics.

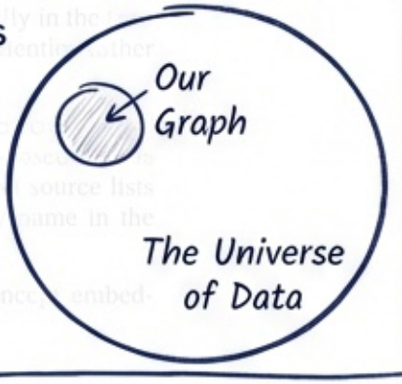
THE MESSY REALITY OF DATA:
- OCR is imperfect,
- LLMs make mistakes,
- "Discovery" is hard to define.

Because the 9,999 concepts are obtained by clust... overlap is unavoidable: closely related or near-syn... scientifically coherent. Our inference-time concept ... to mitigate this, but more principled approaches—such as ... assignment, or jointly learned concept embeddings—could potentially address concept overlap more directly and are a promising direction for future work.

Scope and future directions. The object vo... by name in the literature, a small fraction ... captures only objects that individual papers ... SDSS list millions to billions of sources. ... The resulting graph therefore reflects objects ... than the full censuses of known sources.

WHAT'S NEXT? (Because science is never done),

Our graph only includes objects *mentioned* in papers. That's a tiny fraction of what's out there!
Future work could be ***MUCH BIGGER***.



Future work could extend this foundation ... that evolve with the literature, richer temporal ... over graph snapshots), stronger graph-based ... from large surveys to broaden the object ve... literature.

The concept–object knowledge graph, inclu... and concept embeddings, are available in the Data and Code Av...

Data and Code Availability

All code to reproduce the object extraction, entity resolution, training and evaluation pipeline, along with (i) the SIMBAD-resolved object identifier mapping and object metadata used in this work, (ii) paper-level extracted object mentions with rule and study-mode annotations, is available at: https://github.com/Jinchul120/... o-link-forecasting.

Want to try it yourself?
GET THE CODE HERE!
(Open science ftw!)

Acknowledgments

This work was supported in part by the U.S. National Science Foundation under Grant AST-3406728 and by a Humboldt Research Award from the Alexander von Humboldt Foundation (YST). This work also used GPT-5-mini API access provided through the NSF National Artificial Intelligence Research Resource (NAIRR) Pilot.

Work at Argonne National Laboratory was supported by the U.S. Department of Energy, Office of High Energy Physics. Argonne, a U.S. Department of Energy Office of Science Laboratory, is operated by UChicago Argonne LLC under Contract No. DE-AC02-06CH11357. NR is supported by Laboratory Directed Research and Development (LDRD) funding from Argonne National Labora-

Thanks for reading! Now go find some cool, unexpected connections!
(And maybe cite us if you do 😉).

THE "GIANT'S SHOULDERS" WE STOOD ON (Recipe for Science!)

No. DEAC02-06CH11357.

This is the foundation! Science is a relay race. These are the people who ran before us.

References

Lada Adamic and Eytan Adar. How to search a social network. *Social Networks*, 27(3):187–203, 2005. ISSN 0378-8733. doi: https://doi.org/10.1016/j.socnet.2005.01.007. URL https://www.sciencedirect.com/science/article/pii/S0378873305000109.

Milad Aghajohari, Mohammad Sadegh Akhondzadeh, Saleh Ashkboos, and Kamran Chitsaz. Degree-based feature is all you need: Science4cast report. In *2021 IEEE International Conference on Big Data (Big Data)*, pp. 5791–5794, 2021. doi: 10.1109/BigData52589.2021.9671530.

Iz Beltagy, Kyle Lo, and Arman Cohan. SciBERT: A pretrained language model for scientific text. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan (eds.), *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 3615–3620, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1371. URL https://aclanthology.org/D19-1371/.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners, 2020. URL https://arxiv.org/abs/2005.14165.

Aaron Clauset, Cristopher Moore, and M. E. J. Newman. Hierarchical structure and the prediction of missing links in networks. *Nature*, 453(7191):98–101, May 2008. ISSN 1476-4687. doi: 10.1038/nature06830. URL https://doi.org/10.1038/nature06830.

Daniel Daza, Michael Cochez, and Paul Groth. Inductive entity representations from text via link prediction. In *Proceedings of the Web Conference 2021*, WWW '21, pp. 798–808, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450383127. doi: 10.1145/3442381.3450141. URL https://doi.org/10.1145/3442381.3450141.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In Jill Burstein, Christy Doran, and Thamar Solorio (eds.), *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1423. URL https://aclanthology.org/N19-1423/.

G. Eichhorn. Linking in the astrophysics data system. In U. Grothkopf, H. Andernach, and S. Stevens-Rayburn (eds.), *Library and Information Services in Astronomy III*, volume 153 of *Astronomical Society of the Pacific Conference Series*. Astronomical Society of the Pacific, 1998. URL https://www.stsci.edu/stsci/meetings/lisa3/eichhorn3.html.

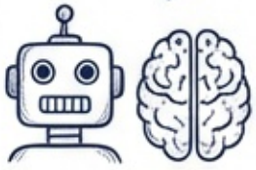
Felix Grezes, Sergi Blanco-Cuaresma, Alberto Accomazzi, Michael J. Kur… Edwin Henneken, Carolyn S. Grant, Donna M. Thompson, Roman Ch… Timothy W. Hostetler, Matthew R. Templeton, Kelly E. Lockhart, Na… Chen, Chris Tanner, and Pavlos Protopapas. Building astroBERT, a lan… & astrophysics, 2021. URL https://arxiv.org/abs/2112.…

Xuemei Gu and Mario Krenn. Forecasting high-impact research topics… evolving knowledge graphs, 2024. URL https://arxiv.org/abs… preprint. (Journal reference listed on arXiv: Mach. Learn.: Sci. Technol. 6 02…

Edwin Henneken. The ADS object search: How to find astronomical objects in the literature. NASA Astrophysics Data System Blog, September 2022. URL https://ui.adsabs.harvard.edu/blog/ads-object-search.

CLASSIC PAPERS & CORE CONCEPTS

Network theory, citation analysis... the fundamental ideas.



AI & LANGUAGE MODELS (The 'Brains')

These are the tools we used to 'read' the papers (like BERT, SciBERT).

ASTRONOMY DATA & TOOLS (The 'Eyes')

Where the actual science lives (ADS, arXiv).

This is your treasure map! Go explore these to start your OWN project. The adventure continues... *OVER TO YOU!*

Thanks for reading!

Thanks for coming on this journey with us! Now go make some science! 🚀