The Diachronic Drift of 'Dark Matter': Theory Change, Semantic Stability, and Continuity in Scientific Practice

Simon Allzén*[1,2,3]

[1]Department of Philosophy, Stockholm University
[2]Institute of Physics, University of Amsterdam
[3]Vossius Center for the History of Humanities and Sciences

October 9, 2026

Abstract

Dark matter is well established as a phenomena at gravitational scales but remains underdetermined at the scale of microphysics. This combination makes it a revealing case of persistent underdetermination, one in which epistemic labor gets redistributed from canonical confirmation to practices that stabilize the target of inquiry. This paper asks what that redistribution implies for semantic stability. I examine two such practices: research histories that narrate continuity and scientific usage that drifts across time. The case study combines primary-source analysis of dark-matter usage from 1882 to the 1950s with computational analysis of post-1985 scientific literature using distributional semantics. The combined evidence reveals dual historical usage, modern ontological proliferation, and cumulative semantic displacement despite influential narratives of continuity. The result is diagnostic rather than debunking: in empirically scarce research programs, continuity may need to be narratively maintained because it is not yet semantically secured. Dark matter thereby clarifies how theory change, reference, and scientific practice interact when canonical confirmation remains unavailable.

"A rose by any other name would smell as sweet"
—William Shakespeare, *Romeo & Juliet*

1 Introduction

Dark matter combines robust gravitational support with persistent openness about constituent nature. That combination makes it an unusually revealing case for thinking about semantic stability in an active research program. When canonical confirmation remains unavailable, the sense that scientists are tracking one target rather than several adjacent ones cannot be taken for granted. It must be maintained, negotiated, and sometimes actively constructed. The governing question of this paper is simple: what does the dark matter case show about semantic stability in empirically scarce domains? Both critics and defenders of scientific realism often assume that the meaning and reference of theoretical terms can be straightforwardly read off of past theories. In the pessimistic meta-induction, philosophers treat 'phlogiston' or 'aether' as having determinate meanings that failed to refer. Realist responses like Psillos's divide-and-conquer strategy depend on identifying stable 'core descriptions' that preserve reference across

*simon.allzen@philosophy.su.se · ORCID: 0000-0002-6124-8152

theory change. But these debates rarely test whether the semantic stability they assume actually obtains in scientific practice, especially under persistent underdetermination. This paper does so by treating dark matter as a capstone case: a case in which the target is stable enough to organize research, but unstable enough in constituent ontology that historical and semantic work become epistemically significant in their own right.

Realist responses to the pessimistic meta-induction typically invoke referential continuity: if a term's reference is preserved across theory change, then past success supports present realism. For instance, to be realist about 'aether' just is to be realist about the electromagnetic field if they co-refer. But continuity cannot be inferred from naming conventions or functional roles alone — the same label may track different entities, or different labels may track the same entity. Here in Psillos' own words:

> From a realist point of view, an important constraint on the theory of reference — if not the raison d'être of a theory of reference — is the accommodation of linguistic practices to the causal structure of the world. Given that this accommodation is, as a matter of fact, an historical process, it is vital for a realist view of science that the theory of reference allows for convergence in the scientific image of the world. Theoretical terms are necessary for scientists to talk about posits not given in experience; but they should be taken to have transtheoretical robustness. Unless continuity in theory change and convergence are established, past failures of scientific theories will act as defeaters of the view that current science is on the right track. The causal descriptivist approach to the reference of theoretical terms is fit to strike a balance between, on the one hand, the theory dependence of theoretical terms and, on the other hand, viewing theoretical terms as relevantly transtheoretical. (Psillos, 2012, p. 234)

In order to probe 'linguistic practices' and how scientists 'talk about posits not given in experience', I leverage state-of-the-art methods in distributional semantics to detect semantic patterns in a comprehensive dark matter corpus drawn from the modern post-1985 literature. But before presenting computational evidence, we must first understand how continuity is constructed in the research histories that shape contemporary understanding. The argumentative sequence is relatively straightforward. I first isolate the target-stabilizing role of historical narrative. I then reconstruct the heterogeneous usage those narratives downgrade or ignore. Finally, I show that modern scientific usage does not simply inherit a single stabilized referent, but reorganizes it across a pluralized landscape of candidates, roles, and ontological commitments. The historical sections are therefore inputs to the semantic diagnosis, not a second independent history paper inside the capstone.

2 Target stabilization in research histories

In contemporary dark matter research, explicitly historical syntheses are not peripheral but central to disciplinary self-understanding. Review papers with titles like *History of Dark Matter* (Bertone and Hooper, 2018) achieve high citation rates across subfields and appear prominently in both technical literature and pedagogical contexts. These texts do not merely survey past results — they construct narratives that organize the research landscape, stabilize what counts as the target of inquiry, and articulate why sustained commitment remains rational despite persistent underdetermination. Understanding how such narratives function is essential for assessing whether the semantic stability assumed in realist discourse actually obtains in scientific practice.

2.1 Research histories as epistemic structures

I use the term *research history* for a genre of scientific synthesis that is simultaneously retrospective and programmatic: it narrates a past in order to orient present practice. Following Staley (2008, p. 297), research histories are:

> [...] explicitly historical but highly selective accounts of the emergence and implication of a theory, experiment, or discovery that are presented in major research or research review papers. (Staley, 2008, p. 297)

Unlike academic historiography, these accounts are written by practicing scientists for scientific audiences, and they serve distinct epistemic functions within ongoing research programs (Kragh, 2017). In resolved scientific episodes — when a theory has been decisively confirmed or abandoned — historical accounts typically trace either the path to success or the sources of error. But dark matter occupies an unusual epistemic position. It is simultaneously a stable commitment within cosmology and astrophysics (incorporated into the standard ΛCDM model)[1] and a radically open question regarding microphysical nature (with no agreed non-gravitational detection despite decades of experimental effort). In such contexts, research histories can establish conceptual *continuity*: by selecting turning points, defining conceptual ancestry, and identifying stable targets across theoretical change, they help determine what the community treats as the same object of inquiry over time (Kragh, 1987; Pestre, 1999).

This continuity operates through a three-stage mechanism. First, a research history *anchors* the subject matter by selecting which earlier episodes, actors, and uses of a term count as relevant antecedents of the present target. This is not a neutral act of recovery but a substantive interpretive choice about which past practices track what we now call 'dark matter'. Second, the selected anchors *stabilize* what the term is taken to pick out across time, typically by treating heterogeneous past uses as either partial glimpses of the same target or as 'prehistory' whose divergence is not conceptually constitutive. Third, once the target is stabilized diachronically, the resulting narrative can *rationalize* ongoing commitment so that the programme appears coherent, cumulative, and historically well-motivated rather than as a sequence of disconnected responses to anomalies. In this way, continuity narratives do not replace empirical work, but they can function as supporting structures that legitimate theoretical and experimental investment when canonical forms of confirmation remain elusive.

2.2 The constructed continuity of dark matter

A particularly influential example of this genre is the comprehensive review *History of Dark Matter* by Bertone and Hooper (2018). Published in *Reviews of Modern Physics*, this 73-page synthesis has accumulated over 2000 citations in just seven years, placing it in roughly the top 0.2% of cited papers within the dark matter literature (see Figure 1 from Allzén (2024, p. 17)).[2] The paper's uptake extends across traditional disciplinary boundaries and it is cited not only in astrophysics and cosmology but also in particle phenomenology, detector physics, and gravitational theory, indicating its function as a cross-community orienting text.

[1] The ΛCDM model is the standard cosmological model that describes the evolution of the universe. It includes a cosmological constant, Λ, representing dark energy and cold dark matter (CDM) as the dominant form of matter.

[2] Based on ADS bibliometric data for papers containing 'dark matter' ($N = 177,280$), Bertone and Hooper (2018) ranks at the 99.8th percentile for average citations per year.

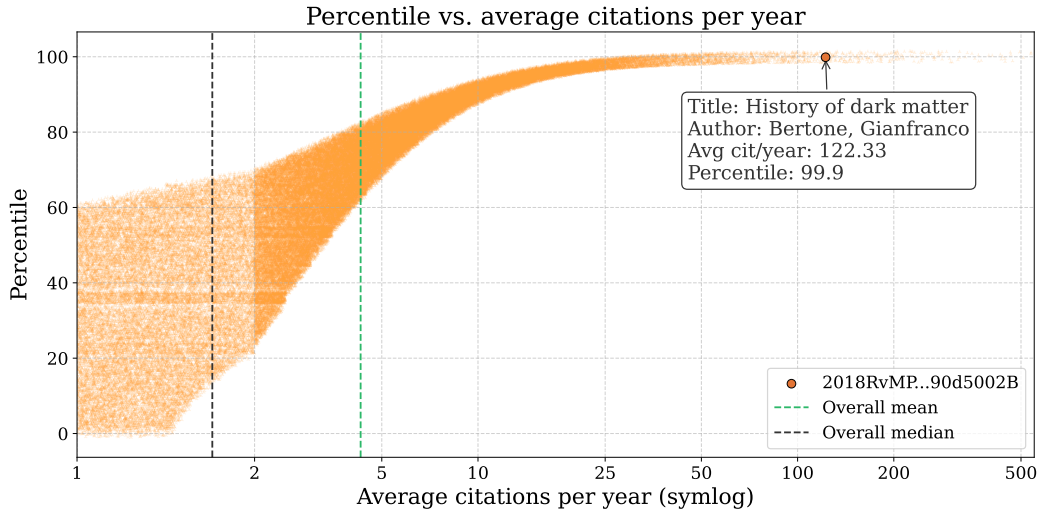


Figure 1: Rank of papers (percentile) based on the total sum of citations a paper has divided by number of years since its publication. The highlighted paper is among the top 0.2% of cited papers out of 177.280 papers mentioning 'dark matter'.

Bertone & Hooper construct continuity by anchoring 'dark matter' in a dynamical lineage. Their account "study[s] the emergence of the concept of dark matter in the late 19th century and identify[s] a series of articles and other sources that describe the first *dynamical estimates* for its abundance in the known Universe" (Bertone and Hooper, 2018, p. 4, emphasis added). The conceptual origins of dark matter are thereby explicitly tied to dynamics from the outset. Early twentieth-century figures — Kapteyn, Jeans, Lindblad, Öpik, and Oort — are presented as nodes connecting past to present precisely because their work involved 'dynamical evidence' for non-luminous mass. This establishes the first line of continuity: 'dark matter' is *univocally* a dynamical phenomenon from its earliest instantiations. The second line traces this characterization forward through Zwicky's cluster dynamics (1930s), galactic rotation curves (1970s), and ultimately to contemporary particle-motivated search strategies. Bertone & Hooper emphasize that "the pioneering work of Kapteyn, Jeans, Lindblad, Öpik and Oort opened the path toward modern determinations of the local dark matter density, a subject that remains of importance today, especially for experiments that seek to detect dark matter particles through their scattering with nuclei" (Bertone and Hooper, 2018, p. 12). The result is a narrative in which present-day direct-detection experiments inherit not merely a set of observational puzzles but a stable, gradually sharpening concept with continuous referential identity.

What this narrative systematically downplays or reclassifies is any usage of 'dark matter' that does not fit the dynamical template. Early applications to light-obscuring phenomena — matter that blocks or absorbs starlight, creating dark lanes in nebulae or obscured regions of the Milky Way — are relegated to a 'prehistory' chapter under headings like *Dark Stars, Dark Planets, Dark Clouds* (Bertone and Hooper, 2018, pp. 5–7). The implicit message is that these were proto-concepts or semantic false starts, distinct from the dynamical usage that constitutes the 'real' history of dark matter. Yet as we will see, this dual usage was not incidental, conflated, or contradictory but *integrated* in the practice of astronomers like Oort, Lundmark, and Pannekoek, who used the same term 'dark matter' to explain both gravitational anomalies and light-absorption effects — often within the same papers and sometimes within the same calculations.

2.3 Testing continuity empirically

Research histories implicitly rely on judgments about semantic stability. When Bertone & Hooper present dark matter as having continuous semantic identity from Kapteyn to contemporary WIMPs, they are making a claim about referential stability: that these diverse episodes involved tracking *the same* entity, even as theoretical understanding evolved. The narrative's persuasive force depends on this continuity being real rather than constructed. If the historical usage of 'dark matter' were radically heterogeneous — tracking incompatible explanatory targets or exhibiting conceptual ruptures rather than gradual refinement — then the appearance of continuous inquiry would be an artifact of selective historical reconstruction rather than a feature of scientific practice.

This paper introduces computational methods from distributional semantics to empirically test such continuity claims. Rather than relying on philosophical intuitions about meaning or selective quotation from historical sources, we can now track semantic patterns at scale across comprehensive corpora. If the narrated continuity is accurate — if 'dark matter' really did gradually sharpen as a concept anchored in dynamics — we should find corresponding patterns in computational analysis: stable semantic cores, increasing specificity around gravitational phenomena, and perhaps the gradual pruning of peripheral senses. Conversely, if there is a systematic gap between the narrative presented in research histories and the semantic practices revealed by distributional analysis, this tells us something important about how continuity gets constructed and maintained in scientific communities, especially in domains characterized by persistent empirical underdetermination. The following sections pursue both forms of evidence. Section 3 provides targeted historical reconstruction showing that early twentieth-century usage of 'dark matter' exhibited a duality that contemporary research histories either downplay or erase: the term functioned simultaneously in explanations of dynamics and light-obscuring phenomena, often explicitly integrated by key actors in the dynamical lineage itself. The later computational sections do not pretend to observe that early episode directly; rather, they test whether subsequent scientific usage in the modern literature actually converges toward the sort of semantic stability those retrospective narratives presuppose. Together, these lines of evidence reveal a more complex picture. Semantic stability is not simply present or absent, but operates at different scales and within different contexts, and the appearance of continuity can be as much a product of narrative construction as of linguistic practice.

3 Light and darkness

The preceding section showed how Bertone & Hooper construct continuity by anchoring 'dark matter' in a dynamical lineage while relegating other uses to 'prehistory'. This section reveals what that construction obscures, namely that the early 20th-century usage was not univocally dynamical but exhibited a persistent duality. 'Dark matter' functioned simultaneously in explanations of gravitational anomalies and light-obscuring phenomena — often explicitly integrated by the very actors (Oort, Lundmark, Pannekoek) whom Bertone & Hooper place in the dynamical lineage itself. Selectively presenting parts of history that support the dynamical characterization promotes a narrative that implies continuity in usage and understanding between past scientists and their present counterparts. Semantic continuity in turn implies that terms have picked out *the same entity* throughout history. In the context of dark matter, this strengthens the idea that, despite substantial theoretical evolution and failure of canonical empirical detection, our knowledge has increased cumulatively. However, if there's no such continuity — if 'dark matter' as used by Zwicky referred to an entity with properties incompatible with contemporary understanding — then the notion of a linear history looks more like a rereading of the past constructed to meet a current epistemic need.

3.1 The prevalence and impact of history in dark matter

The presence of a history of a scientific theory is usually preceded by a pivotal moment in which the theory is either confirmed or falsified. The benefit of hindsight enables historians to discern the patterns which led scientists astray, when a theory has been falsified, or trace the key insights pointing towards truth in case of its confirmation. The presence of accounts on the history of dark matter is therefore unexpected, considering that it is currently one of the biggest *open* questions in physics. Even more unexpected is that a majority of these historical accounts were written by physicists actively working on the dark matter problem at the time.[3] These histories of dark matter are not simply autobiographical memorabilia peripheral to the standing of dark matter in the broader community but are of great interest to a majority of the physicists currently working on dark matter. The high level of impact generated by a paper on the history of dark matter within the physics community tells us that it is an incredibly popular topic among its members. More importantly, however, it tells us that the content and epistemic narrative presented in a historical account may seriously influence the beliefs and credence of the community regarding the *current* dark matter hypothesis. Support for the latter claim can be found by looking in cosmology textbooks, popular science media, and encyclopedic entries. It's not uncommon to come across dense historical claims in these sources, often in some version of the following proposition:

> Originally known as the "missing mass", dark matter's existence was first inferred by Swiss American astronomer Fritz Zwicky, who in 1933 discovered that the mass of all the stars in the Coma cluster of galaxies provided only about 1 percent of the mass needed to keep the galaxies from escaping the cluster's gravitational pull. The reality of this missing mass remained in question for decades, until the 1970s when American astronomers Vera Rubin and W. Kent Ford confirmed its existence. (Weiss, 2021)

This nugget of dark-matter history is not to be faulted for its lack of depth, considering that it's an encyclopedia entry. However, the fact that a historical description even *is* included in an encyclopedia entry suggests its significance for the general perception of dark matter. The entry, though short and dense, manages to nonetheless represent the physicist perspective fairly accurately. Although the scope, depth, and details vary, the usual starting point is Fritz Zwicky's (1933; 1937) studies of the Coma cluster, in which he found that luminous mass alone could not fully explain the galactic movements in the cluster. He concluded that there must be some 'dunkle materie' in the cluster, the presence of which would explain the dynamics. It is popular to credit Zwicky for introducing the term 'dark matter', often followed by berating the contemporary scientific community for overlooking the magnitude of Zwicky's incredible discovery. According to these historical narratives, Zwicky's work represents the genesis of dark matter. The portrait of Zwicky as a *"pioneer"* (Bertone and Hooper, 2018, p. 13) or *"one of those rare unorthodox geniuses"* (Sanders, 2010, p. 12) is a curiously frequent aspect of the narrative.

This simplified history presumably reaches the widest audience, shaping general perceptions. More significant for our purposes is the *research history* from which it derives: detailed but selective accounts in major review papers that establish canonical narratives.

3.2 A history of dynamics

The review paper by Bertone and Hooper (2018) is a comprehensive and thorough account of many of the most important steps in the theoretical evolution of what is today known as

[3] A notable exception is de Swart (2022).

dark matter. The aim is to *"provide the reader with a broader historical perspective on the observational discoveries and the theoretical arguments that led the scientific community to adopt dark matter as an essential part of the standard cosmological model"*. In a sense, the aim is to provide the reader with the arguments for the existence of dark matter which convinced scientists to accept it. As such, the paper should be treated as an epistemological endeavor as much as an historical one, especially considering its high impact in the dark-matter community. Their account makes extensive use of primary historical sources and state-of-the-art knowledge of dark matter to provide a comprehensive overview of dark matter's *"interesting history and how it came to be accepted as the standard explanation for a wide variety of astrophysical observations"* (Bertone and Hooper, 2018, p. 1).

Kapteyn, Jeans, Lindblad, Öpik, and Oort are represented in the subsection *Dynamical Evidence*, succeeding *Dark Stars, Dark Planets, Dark Clouds* in the chapter *Prehistory*. Chronologically ordered, the sections provide the reader with sources referring to a kind of proto-concept of dark matter present up until the end of the 19th century, a concept which is then narrowed by the advent of dynamical evidence in the early 20th century.[4] In their study of the emergence of the concept of dark matter, and in opposition to the simplified history, the authors emphasize that Zwicky's use of the term 'dark matter' was in fact neither novel nor isolated, but instead argue that the term was ubiquitous among his contemporaries:

> We [...] discuss the pioneering work of Zwicky within the context of the scientific developments of the early 20th century. [...] we find that his use of the term 'dark matter' was in continuity with the contemporary scientific literature. (Bertone and Hooper, 2018, p. 4)

> Although [Zwicky] doesn't explicitly cite any article, it is obvious [...] that he was well aware of the work of Kapteyn, Oort and Jeans discussed in the previous chapter. His use of the term 'dark matter' is, therefore, in continuity with the community of astronomers that had been studying the dynamics of stars in the local Milky Way. (Bertone and Hooper, 2018, p. 14)

Bertone and Hooper (2018) rejects the narrative trope of Zwicky as a lone and misunderstood genius, and with it the notion that he was the first to "discover" that additional matter was needed to explain the observed phenomena. Instead, the authors stress that Zwicky's use of the phrase 'dark matter' conformed to the inferential and terminological practices established in the contemporary scientific community, represented by Kapteyn, Oort, and Jeans, i.e., the actors present in the section on dynamical evidence. Zwicky's results are aligned with Smith (1936), whose estimates of the mass of the Virgo Cluster appear in Hubble's *The Realm of Nebulae*. The account says little of the development in the years running up to the late 1960s, suggesting little or no significant work on the topic until then.

This narrative establishes two forms of continuity. First, *semantic* continuity in which 'dark matter' picks out dynamical phenomena from Kapteyn through Zwicky to the present (without which no claim about the persistence of a referent could get off the ground). On the basis of this semantic continuity, the narrative further implies ontological continuity (the implied ontology constituted by the properties necessary for explaining the dynamical evidence): the same entity has been the referent throughout, despite theoretical change. The conceptual continuity of dark matter is preserved by the persistence of its semantic nucleus. The result is that dark matter appears as a stable, self-identical target of inquiry — a concept

[4] Interestingly, Bertone and Hooper (2018, p. 7) include black holes in the concept, *"as an explicit example of a discussion of a class of invisible astrophysical objects, that populate the universe while residing beyond the reach of astronomical observations"*. Interesting because *invisible* and *non-luminous* are not co-extensive terms, a distinction we will revisit later.

the resolution of which gradually increases rather than a term with heterogeneous historical usage. I argue that the semantic continuity of the term 'dark matter' is artificially constructed by highlighting a path through history only visible through the lens of a modern definition of dark matter.

3.3 The duality of dark matter: dynamics and light

Focusing on dark matter in explanations of dynamics overlooks its use in explanations of light-obscuring phenomena. That dark areas of the night sky could be caused by matter obscuring light had, according to Wilson (1904), been proposed already in 1891:

> In this great star cloud too are seen the curious dark patches and lanes, to which attention was called on Plate XVIII. Two explanations are offered for these, the one that they are really, as they appear, vacant spaces void of stars, through which we see into empty space beyond, the other that they are produced by the intervention of enormous clouds of dark matter between us and the stars of the Milky Way, absorbing and cutting off their light. The latter view was strongly advocated some thirteen years ago by Mr. A. C. Ranyard, then Editor of *Knowledge*, but is not generally accepted among astronomers. (Wilson, 1904, p. 215)

In the 1920s, the Swedish astronomer Knut Lundmark (1921) uses dark matter to explain the observed 'dark lanes' found in photographs of the spiral nebulae Messier 33. He conjectures that the appearance of these dark lanes could result from:

> whorls of dark matter between whorls of nebular matter and much pictorial evidence is in favor of this idea. (Lundmark, 1921, p. 324)

Later, Lundmark used dark matter more confidently, as evidenced when he, after calculating the total mass of the visible stellar universe, added this disclaimer to his estimate: *"of course, dark stars and dark matter exist and increase that value"* (Lundmark, 1925). In Lundmark (1930), dark matter appears in estimates of the total mass in galactic systems, determined by the ratio between light and dark matter, a ratio derived using spectrographic data of the rotational velocities of galaxies. In Lundmark's body of work, we see how dark matter is used both as an explanation for light absorption *and* galaxy dynamics. Bertone and Hooper (2018) shares a glimpse of this in the section on galactic rotation curves, writing that *"Holmberg argued in 1937 that the large spread in mass-to-light ratios found by Lundmark was a consequence of the absorption of light produced by the dark matter"* (Bertone and Hooper, 2018, pp. 18–19). The 'dual' use of dark matter also finds support in other sources. Dutch astronomer Anton Pannekoek explicitly used this duality when attempting to account for the mass needed in the galactic nucleus, should the hypothesis that our galaxy is a rotating spiral system be true:

> Thus we come to the conclusion that visible stars of the galactic system cannot provide the required central attracting masses. [...] Cannot the dark matter itself act as a part of the attracting masses? In the visual aspect of the Milky Way the dark absorbing nebulae have the same importance as the luminous star clouds. (Pannekoek, 1927, p. 40)

Two years later, Pannekoek's idea was picked up in Harper's (1929) annual address for the Royal Astronomical Society of Canada. Under the rubric *Current progress in astronomy* we find a discussion on the *Rotation of the Galaxy*, centered on the work of Strömberg and Lindblad, who had hypothesized that our galaxy was a rotating spiral nebula. Harper refers to

a pair of papers published by Oort which are "tending to verify the soundness" of the rotation hypothesis. One implication of the hypothesis is that the mass "required at the centre to give the requisite rotational velocity" is estimated at around $8 \times 10^{10} M_{\odot}$, prompting Harper to question the evidence for the presence of such a mass:

> This central nucleus is situated in the plane of the galaxy and in galactic longitude approximately 325° in the general direction of the constellation *Sagittarius*. Here the Milky Way clouds are brighter than in other parts of the sky, and Pannekoek examines their total luminousity to see if they would yield the necessary mass. He is forced to admit that [...] the visible stars of the galactic system cannot provide the central mass required. The question arises, would not the dark matter in our system yield this attracting mass? Photographs of distant spirals seen edge-on suggest enormous masses of obscuring material. Similar evidence is afforded nearer home in the almost complete obliteration of faint stars in certain regions of the sky by occulting matter which must be extremely tenuous but nevertheless of great effect when summed up over large volumes of space. (Harper, 1929)

The idea of turning to dark matter when one has a deficit in the mass-density budget is evidently not new. The literature contains an abundance of examples that illustrate the ubiquity of using the term 'dark matter' in explanations of phenomena related to both light absorption and dynamics, even, as Harper notices, in Oort. For Oort, estimating the mass, velocity, and rotation of a stellar system depended on assessing the luminosity of that stellar system reliably, but the existence of light absorption due to dark matter introduced uncertainty in those assessments. In a public lecture entitled *Non-Light Emitting Matter in the Stellar System*, Oort (1927) cites Lundmark and uses dynamical calculations to determine the amount of dark matter, which in turn is used to determine the degree of light absorption.

> Do we then have to give up hope to learn about the absorption in the interstellar gas? Not entirely. [...] I have told you above that *the observed velocities of the stars allow us to make an estimate of the mass of the non-light-emitting matter.* [...] It is not unreasonable to assume that apart from the light-emitting stars there could be a large quantity of dark bodies in the universe. If these bodies would be in the form of very faint stars of the same size and mass as the observed stars, then it is easy to show that there cannot be enough of these dark stars to provide an observable extinction. It becomes completely different, when the available mass would be in the form of small solid particles, distributed throughout the Galactic System. We then have just about enough to give rise to a significant absorption over distances of a thousand light-years and we can easily assume these particles to be spread out so densely over a large area of space that more distant stars are completely extinguished. (Oort, 1927, pp. 61–62)

Oort is clearly using his dynamic calculations of dark matter to estimate the amount of light absorption that one can expect from it. For dynamic calculations to be informative about the amount of light absorption, Oort would have to regard the two phenomena to be manifestations of the same cause. Estonian astronomer Ernst Öpik (1929) concludes that *"in some cases absorption by dark matter appears probable"* as an explanation of obscured regions in space. Dutch polymath de Sitter (1932, p. 91) writes that *"evidence has accumulated"* in support of the idea that *"clouds of dark matter* [...] *prevent us from seeing the stars"*.[5]

[5] In addition, see: Markov (1929, p. 342), Notes of the British Astronomical Association (1930, p. 49), and Holmberg (1937).

Zwicky's contemporary community clearly had a much wider understanding of the concept of dark matter than alluded to in previous historical accounts. This wider concept remained established in the community at least until the late 1950s, as illustrated by its presence in Wallenquist (1939, p. 48), Trumpler (1941, pp. 161–162), (Lindblad, 1948, p. [8]; 1951, p. [60]), Lindblad and Delhaye (1949, p. 20), Lindsay (1950, p. 8), Oort (1952, p. 235), and Holmberg (1952, p. 8). The characterization and longevity of this wider concept has two consequences. First, it undermines the notion that the core characteristic of the semantic continuity of the term 'dark matter' in the first half of the 20th-century can be understood only in the context of explanations for dynamical phenomena. Second, it raises doubts about the prospect of construing the hypothesis of dark matter as a single coherent historical framework. If we should consider the historical concept of dark matter as part of the history of dark matter, we have to accept that it is semantically discontinuous and in ontological contradiction with the modern concept of dark matter. Despite the increased historical resolution provided by Bertone and Hooper (2018), the narrative image remained the same: the history of dark matter is presented as coherent and continuous, evidence steadily accumulating toward the modern dark-matter hypothesis. However, the historical dots between which the line through history is drawn can be identified by their consistency with the current understanding of dark matter.

4 Semantic change, dataset, and methods

The preceding sections have shown that the historical usage of 'dark matter' was more heterogeneous than suggested by research histories, and that the narrative of continuity is constructed by selectively highlighting a path through history that aligns with contemporary understanding. The computational analyses that follow answer a narrower but still decisive question. They do not reconstruct the full history from 1882 onward in one continuous evidential stream. Instead, they test whether the modern scientific literature — precisely the literature to which continuity narratives often look for vindication — exhibits the kind of semantic consolidation that those narratives imply. If even this later discourse remains internally differentiated, drift-prone, and ontologically unsettled, then the case for straightforward referential continuity is weaker than the retrospective histories suggest.

4.1 Semantic change

The phenomenon of semantic change, or semantic shift, refers to the process by which the meaning of a word or phrase evolves over time. Ubiquitous examples include the change from "full of awe" to "terrible" for the word "awful" and "lighthearted", "joyous" to "homosexual person" for the word "gay". A word, phrase, or term can change or shift in various ways: it can widen so as to acquire more senses, as "mouse" by the advent of personal computers, or it can be narrowed or specialized, as in the case of "skyline" by the advent of skyscrapers.[6] The increasing availability of computational power, digitized large-scale corpora, and open-source language models in the last 10 years have facilitated reliable ways of detecting semantic change computationally. These methods draw on the core principles of *distributional semantics*, which holds that a term's meaning is reflected by the linguistic context in which it appears. This idea forms the basis for much of the work on diachronic semantic change. Fundamentally, my approach shares the spirit, if not the letter, with the framework proposed by Jatowt and Duh (2014), which measures semantic drift by comparing n-gram-based co-occurrence vectors across temporal slices:

[6] See Blank (1999) for an overview.

> Word representation lies at the heart of any lexical semantics work. We adopt here a distributional semantics view, where each word is represented by its usage context, that is, its neighboring words in the n-gram datasets. Distributional semantics is best portrayed by a well-known saying attributed to J. Firth: "You shall know a word by the company it keeps". In our implementation we compute a vector representation $d_i[w]$ for each word w in each decade i. Intuitively, if the similarity between the same word's vector at decade i and decade j is low (i.e. $sim(d_i[w], dj[w]) \rightarrow 0$), then a semantic change is likely to occur between these decades. (Jatowt and Duh, 2014, p. 3)

However, rather than relying on surface-level co-occurrence counts, language models like *sciBERT* enable a deeper analysis by providing *contextualized* phrase-level embeddings, so that the target phrase is represented together with its surrounding context. This allows the model to disambiguate between meanings of the same term on the basis of how it is used in context.[7] These methods have not previously been applied to questions of referential continuity in the discourse on scientific realism, but they do provide empirical grounding for debates about semantic stability in theory change. In what follows, I provide a brief overview of the methodology and data used in the ensuing analysis.

4.2 Data and method

The dataset used for the analyses in this paper was retrieved programmatically from the ADS using their native API and SOLR search engine. Only papers which explicitly contained the uncased phrase 'dark matter' were included. The level of information density in the data enables a wide variety of analyses, including the semantic drift detection analysis, which is based on the *abstract* variable that is complete to almost 93%. It should be noted that the use of abstracts, in contrast to full-text documents, diminishes the resolution of the semantic analysis somewhat, but given that abstracts represent the conceptual core of the full text, the trade-off between conceptual payload and computational feasibility is reasonable. The methods used in this paper should scale well to full-text analysis for anyone with the interest and computational resources to do so.

The computational workflow is deliberately staged. Starting from the ADS spine, I build a paper-local abstract frame, extract sentence-level occurrences of the phrase 'dark matter', and retain only mention contexts with enough local linguistic material to support contextual embedding. The resulting master mention table records bibcode, year, mention offsets, and provenance indices, and is then split into two corpora with distinct epistemic roles: an analysis corpus used for all downstream embedding and drift work, and a quarantined shadow corpus reserved for annotation and classifier development. The split is done in five-year bins, with at most two retained mentions per bibcode, year-balanced sampling within bins, and manifest files that record content hashes and final counts so later stages can verify alignment rather than silently drifting from the original corpus definition. Two features of the downstream pipeline warrant explicit defense against the obvious methodological worries they raise: the A/B/C classifier on which the dominant-label split depends, and the embedding representation into which all drift is measured.

On the label classifier. The classifier is a multi-label SciBERT head (Beltagy et al., 2019). Its seed labels are drawn from a shadow corpus separated from the analysis corpus at the paper level. The designated sample contains 1,800 mentions from six five-year bins spanning 1996–2025, capped at two mentions per bibcode and pre-stratified into five pools (A-leaning,

[7] See Periti et al. (2024), Montanelli and Periti (2023), Giulianelli et al. (2020), and Martinc et al. (2019) for methodological discussions about the use of contextualized embeddings for semantic drift detection. A diagram of the method can be found in appendix A.

B-leaning, C-leaning, mixed, and random) via lexical heuristics. GPT-5.4 annotated these mentions using a saved prompt (version 5) and a versioned model string. Excluding 104 mentions with no assigned label leaves 1,696 labeled examples, split into 1,356 training and 340 validation examples. Training and validation are split by mention, so papers can contribute to both partitions. The classifier uses class-weighted BCE loss, checkpoint selection by validation macro F1, and a per-label threshold sweep on the same validation set. At a common threshold of 0.5, the retained checkpoint has validation macro F1 of 0.738 and exact-label-set agreement of 0.471. These scores measure agreement with the GPT labels; they are not an independent assessment against human annotations. The seed sample does not cover the first two analysis bins, so classification of the earliest mentions extends beyond its temporal coverage.

An earlier repeat-labeling exercise on 200 mentions yielded exact agreement of 0.900, $\kappa = 0.856$, and mean Jaccard agreement of 0.948. That exercise used GPT-4.1 and prompt version 11. It establishes repeatability for that earlier protocol, but does not validate the later GPT-5.4 seed campaign. The results below should therefore be read as conditional on the annotation scheme and classifier, with their independent accuracy still to be established.

The classifier can assign zero, one, or several labels to a sentence. The dominant-label A/B analyses retain mentions assigned those dominant labels. The classifier figures accordingly report probability distributions and overlap structure as well as hard-label proportions.
On representation and temporal bias. Mentions are embedded at the phrase level in a centered context window (±550 characters around the "dark matter" span, re-snapped to whitespace), with the input truncated at 512 tokens and the mention vector defined as the mean of the token vectors corresponding to the span itself, followed by L2 normalization. The representation is thus anchored to the local linguistic environment of each mention. I compare five layer-aggregation strategies (top layer, last-four mean, middle-four mean, all-layer mean, and attention-weighted), five target dimensionalities (25, 50, 75, 100, 200), and five reducer families (PCA, Gaussian random projection, UMAP, autoencoder, variational autoencoder). The 120 retained configurations are scored on k-nearest-neighbor recall, trustworthiness, and drift correlation with the original 768-dimensional representation. The analyses use the saved 'PCA-100_all_mean' space, which averages the non-embedding transformer layers and reduces the resulting vectors to 100 dimensions. Its scores are 0.865 for neighbor recall, 0.99993 for trustworthiness, and 0.857 for drift correlation. This is to check how well the reduced space preserves the original geometry.[8] The collostructional analysis and lexical profiles then help interpret which patterns of scientific usage accompany the measured displacement.

5 Semantic roles

5.1 Sentence-level labeling

A central step in the pipeline is to distinguish different ways the phrase 'dark matter' functions in context. Each sentence-level mention is classified into three discourse roles: referential ('A'), role-oriented ('B'), and model-parameter or framework-oriented ('C'). Because these uses can overlap in practice, the classifier is trained and evaluated in a multi-label setting; the dominant-label split used in the drift analyses is therefore a downstream simplification rather than the native ontology of the annotation scheme. The classifier figures should be read in this light. They do not claim that every mention is univocal, but they do show that the three roles have distinct temporal profiles and a non-trivial overlap structure.

[8] All-layer averaging and comparison with the unreduced representation address choices within the embedding pipeline. They do not provide an independent benchmark of the model's treatment of scientific language across periods. A workflow overview is provided in Appendix A.

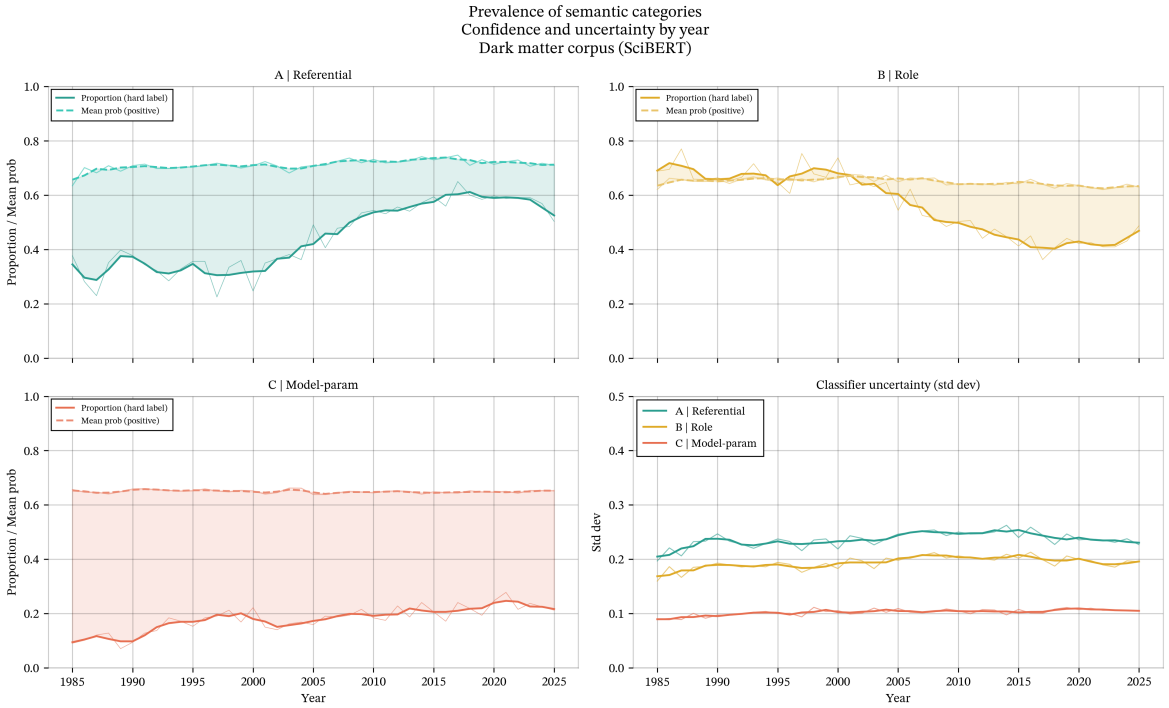


Figure 2: Sentence-level classifier outputs for referential (A), role-oriented (B), and model-parameter (C) usage. Panels show annual label prevalence, mean positive probabilities, and classifier uncertainty. Assigned probabilities are model outputs, not independently calibrated estimates of accuracy.

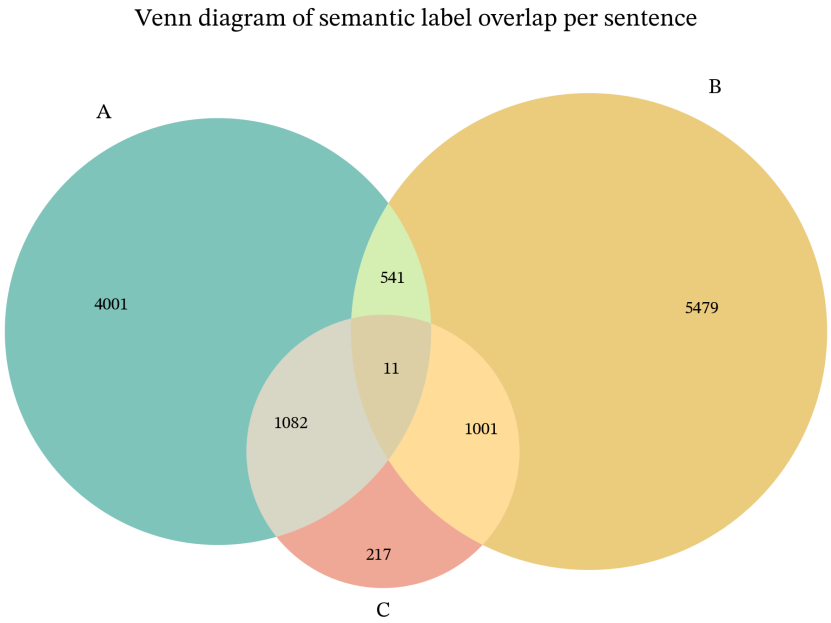


Figure 3: Sentence-level overlap among the three semantic labels. The classifier is multi-label, so the dominant-label analyses simplify an underlying overlap structure rather than recover mutually exclusive senses.

Figure 2 shows that the dominant discourse balance changes substantially over time. Referential ('A') usage becomes more prevalent from the mid-2000s onward, while role-oriented ('B') usage declines from its earlier dominance, and model-parameter discourse ('C') remains smaller but

persistent throughout. The mean positive probabilities remain relatively stable. This does not establish that those probabilities are calibrated or that classification accuracy is constant across periods. Figure 3 shows that the labels are not mutually exclusive, but neither are they massively conflated. Most 'A' and 'B' mentions are single-label, the A/B overlap is real but minority-sized, and the most notable asymmetry is that model-parameter discourse ('C') occurs more often in overlap with 'A' or 'B' than on its own.



Figure 4: Annual dominant-label composition of the analysis corpus together with the distribution of maximum assigned classifier probabilities across broad periods.

The composition results are the ones most directly relevant for the inferential branch that follows. The dominant-label approximation underwrites the 'A'/'B' split in the drift tests, and the binned proportions used to construct the analysis corpus show why that split is philosophically useful. Early bins are weighted toward role discourse, but from roughly 2006 onward the balance tilts toward referential discourse. The empirical question is therefore not simply whether 'dark matter' drifts, but whether these two recurrent modes of usage drift in the same way. The drift results below show that they do not: referential discourse is more rupture-prone locally, while role discourse undergoes a larger cumulative displacement.

5.2 Embedding-space drift estimation

The inferential drift analyses are run on contextualized, phrase-centered embeddings of 'dark matter' mentions in the shared PCA-100_all_mean space, with bootstrap uncertainty and grouped permutation tests applied to the dominant-label A/B split. For adjacent five-year transitions, I estimate centroid cosine distance, MMD, sliced Wasserstein distance, energy distance, and subspace drift. I also track cumulative cosine displacement from the earliest retained bin. The descriptive figure reports bootstrap confidence intervals, while the significance figure reports grouped permutation tests on bibcode-balanced subsets. Grouping at the bibcode level matters here because single papers can contribute multiple retained mentions; the permutation design therefore shuffles grouped bibcode blocks rather than individual rows, so the tests do not treat within-paper repetition as if it were independent evidence. In the manuscript proper, the emphasis is on the dominant-label referential ('A') and role-oriented ('B') split, because that is the contrast most directly motivated by the classifier figures above and most relevant to the philosophical question about role-level versus kind-level stability.

Figure 5 is the main descriptive drift figure. Across cosine, MMD, sliced Wasserstein, and energy-distance measures, both categories exhibit non-trivial adjacent change, but not in the same way. Referential discourse ('A') shows the sharpest local disruption, with a conspicuous peak at the transition from 2001–2005 to 2006–2010. Role-oriented discourse ('B') changes less punctually from one adjacent bin to the next, but its cumulative cosine drift grows more

strongly across the series. The most defensible description is therefore not that one category simply "drifts more" than the other, but that they exhibit different *shapes* of drift: rupture-like local movement in 'A', and larger long-run cumulative reorganization in 'B'.

Figure 6 strengthens that interpretation by showing which transitions are doing the inferential work. The mid-2000s break remains the clearest local discontinuity: in 'A', the transition from 2001–2005 to 2006–2010 is the strongest adjacent effect across multiple metrics. At the same time, cumulative cosine displacement increases overall despite reversals in both categories, and ends higher for 'B'. Several earlier adjacent transitions are also significant, so the story is not one of a single rupture after long stasis. It is a history of ongoing semantic change with one especially strong break and a pronounced divergence between local and cumulative forms of reorganization.

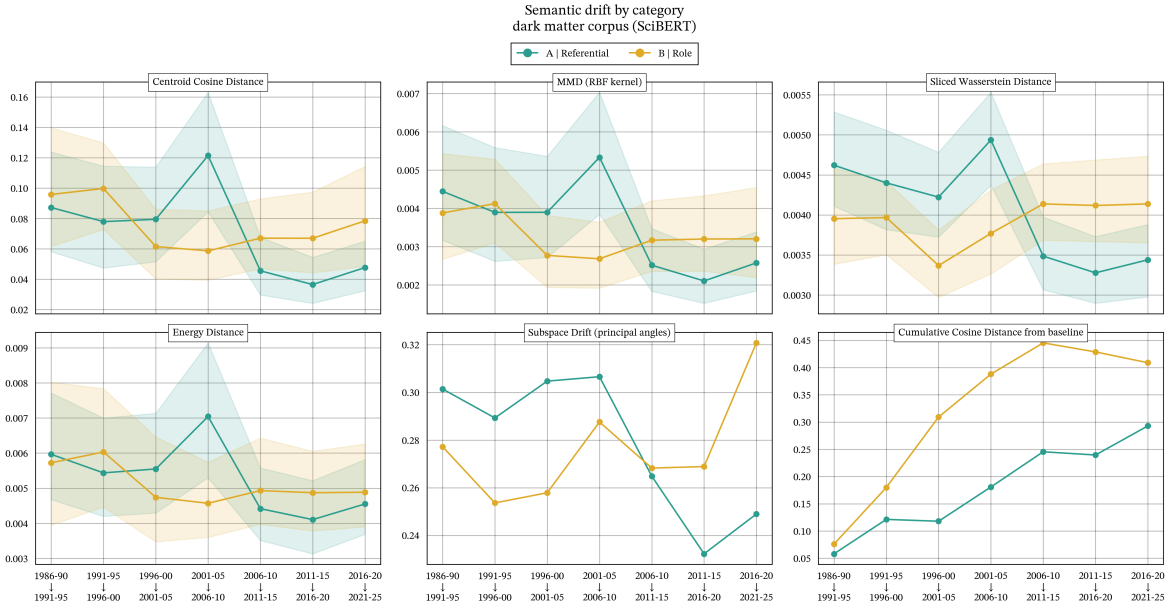


Figure 5: Adjacent and cumulative displacement for referential (A) and role-oriented (B) discourse in the saved PCA-100 all-layer-mean space. Shading shows 95% bootstrap intervals for adjacent comparisons; the cumulative curve gives displacement from the earliest bin. Uncertainty is conditional on the fixed representation and classifier labels.

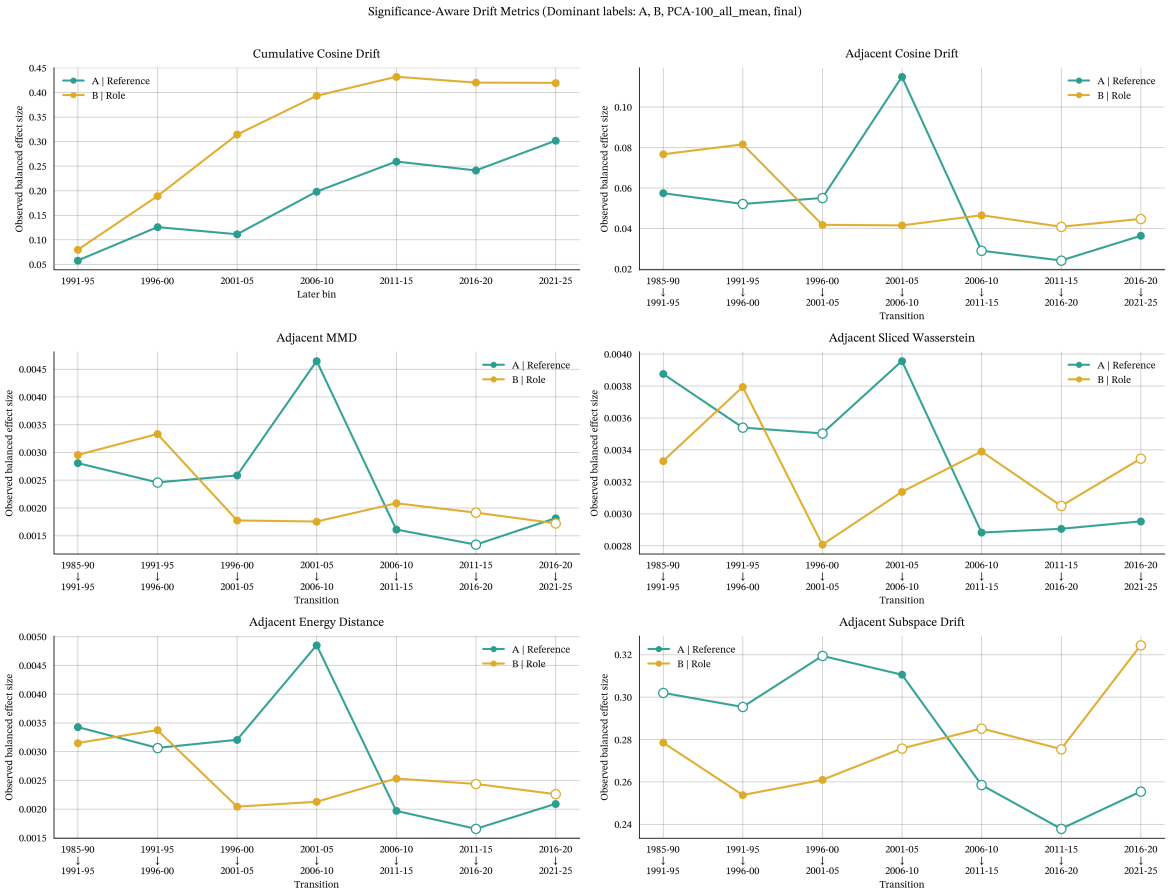


Figure 6: Drift estimates and grouped permutation tests for referential (A) and role-oriented (B) discourse. Filled markers denote q $<$ 0.05; open markers denote q $>=$ 0.05 after Benjamini–Hochberg adjustment. The tests use 999 permutations with paper-level grouping. The cumulative panel compares each later bin with the earliest bin.

Taken together, the two drift figures point to repeated semantic reweighting rather than smooth monotonic change. The key empirical contrast is not that one discourse mode simply changes "more," but that the two modes reorganize differently: referential discourse is more locally rupture-prone, whereas role-oriented discourse accumulates more long-run displacement.

5.3 Collostructional lexical structure

The embedding-space results establish that drift is present; the collostructional analyses show what that drift is moving through. Using the same balanced 'analysis_corpus', I computed category-conditioned collostructional profiles for the dominant-label 'A' and 'B' corpora across the five-year bins. Context tokens are taken from a ±550-character window around the mention, filtered through a small blocklist of trivial self-repetitions and generic filler items, and then scored by collostruction strength before being summarized as heatmaps and timelines. This lexical layer makes the 'A/B' distinction interpretable through the words and phrases that cluster around 'dark matter'. Because the profiles use the same classifier-defined partitions, they do not independently validate the classification.

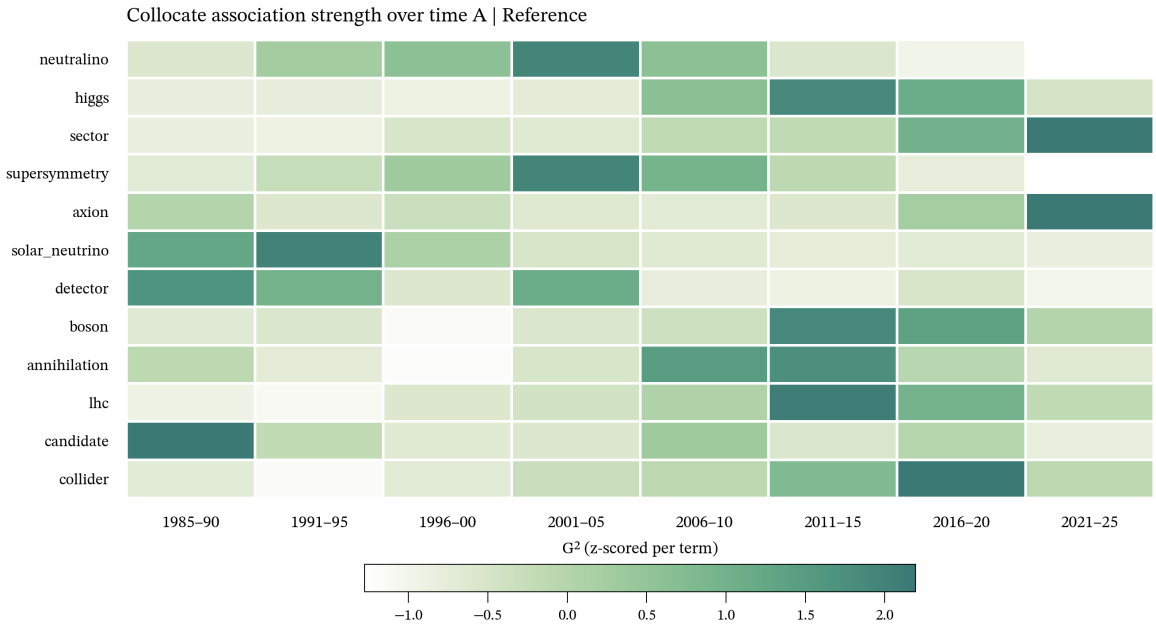


Figure 7: Collostructional association strength over time for dominant-label referential ('A') discourse. Colors are z-scored within term, so darker cells indicate periods of relatively stronger attraction for that term.

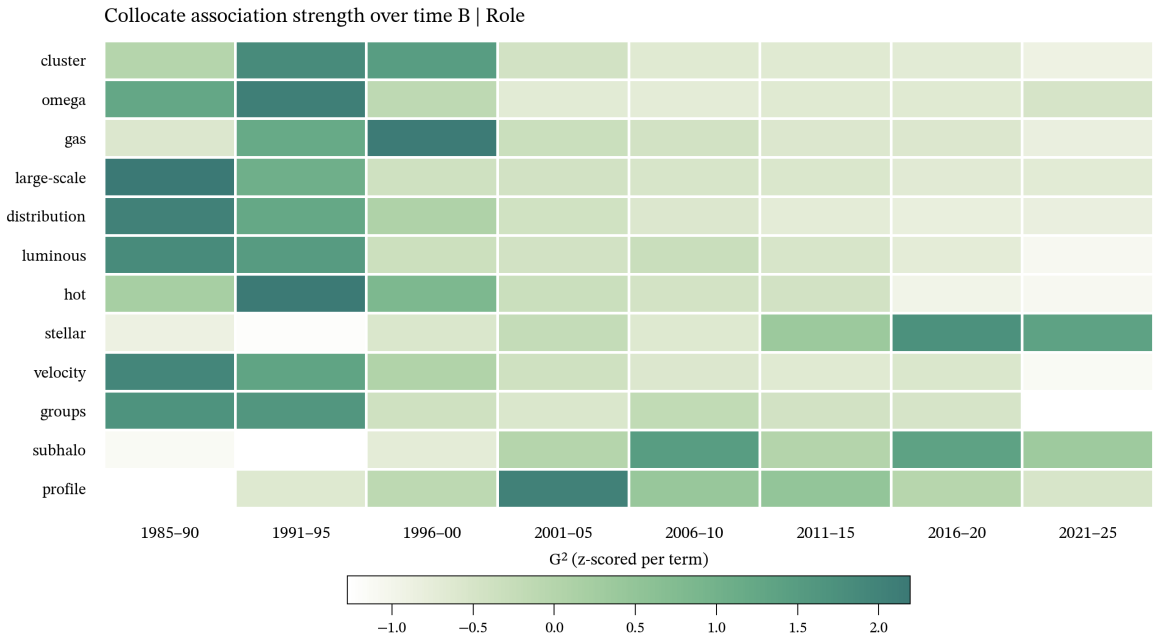


Figure 8: Collostructional association strength over time for dominant-label role-oriented ('B') discourse. Colors are z-scored within term, so darker cells indicate periods of relatively stronger attraction for that term.

The collostruction heatmaps recover a sharp lexical division. The referential side ('A') is organized around candidate kinds, particle-theoretic frameworks, and detection channels: 'neutralino', 'supersymmetry', 'axion', 'detector', 'annihilation', 'lhc', and 'collider'. The role-oriented side ('B') is instead organized around cosmological structure, gravitational phenomenology, and observational evidence: 'cluster', 'omega', 'gas', 'large-scale', 'distribution',

'stellar', 'velocity', 'subhalo', and 'profile'. The lexical attractions are consistent with the role/reference distinction and show what the classifier-defined categories contain. They do not by themselves rule out a classifier artifact.

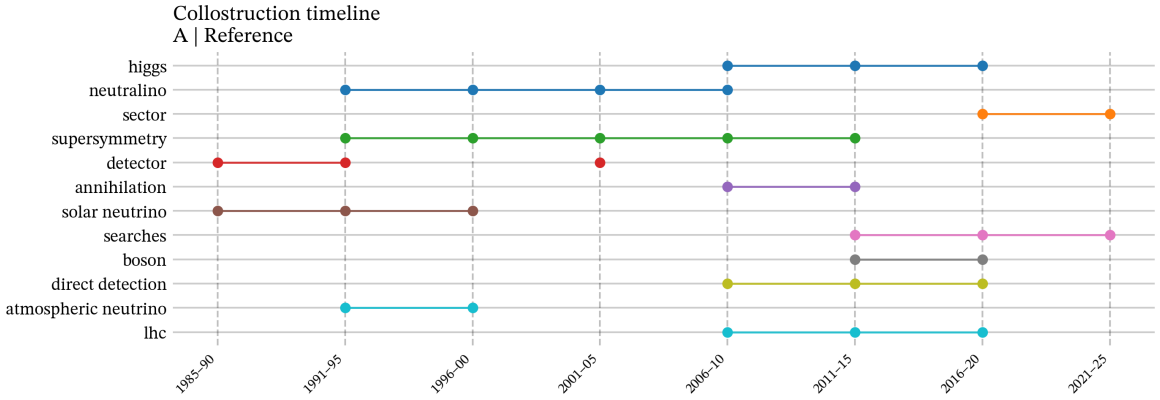


Figure 9: Collostruction timeline for dominant-label referential ('A') discourse, showing when the most diagnostic terms enter and persist.

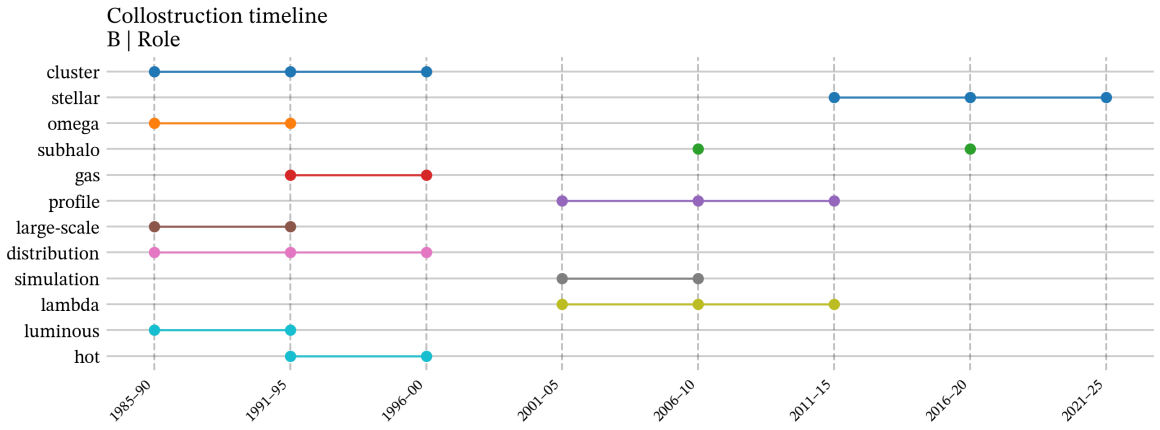


Figure 10: Collostruction timeline for dominant-label role-oriented ('B') discourse, showing when the most diagnostic terms enter and persist.

The timelines make the temporal succession clearer. In 'A', early diagnostic terms such as 'solar neutrino' and 'detector' give way to later, more explicit candidate- and experiment-facing vocabulary such as 'higgs', 'lhc', 'direct detection', and 'searches'. In 'B', early prominence is concentrated in broad observational and structural terms such as 'cluster', 'omega', 'distribution', and 'large-scale', while later bins retain more specialized structure terms like 'profile', 'lambda', 'subhalo', and 'stellar'. The lexical history is therefore not a mere accumulation of synonyms. It is a reorganization among recurring semantic fields. The historical reconstruction established early heterogeneity, the classifier-defined corpus showed structured role differentiation, the drift tests documented significant diachronic movement, and the collostructional profiles showed that the lexical field reorganizes alongside that movement. The recurring gravitational explanatory role remains identifiable even as its associated vocabulary changes while discourse about constitution, candidates, and evidential realization is repeatedly reorganized in practice.[9]

[9]Polysemy, in this sense, refers to the coexistence of multiple related but distinct meanings under the same lexical form, while specialization (narrowing) describes the process by which particular usages acquire more

6 Semantic stability under empirical scarcity

The preceding analyses document significant semantic drift, ontological proliferation, and usage heterogeneity in the term 'dark matter' across 140 years of scientific discourse. Taken together, they suggest a broader diagnosis: in empirically scarce domains, semantic stability is not simply inherited from successful theory, but partly sustained by target-stabilizing practices. This closing section draws out three implications of that diagnosis. First, the semantic specificity required for non-trivial referential success is harder to extract from practice than realist arguments often assume. Second, the gap between research histories and distributional evidence reveals how continuity can be constructed rather than merely discovered. Third, these pressures may be structural features of persistent underdetermination rather than idiosyncrasies of the dark matter case.

6.1 Specificity and referential success

Realist responses to the pessimistic meta-induction include a *divide-and-conquer* strategy which holds that even when theories are abandoned, their theoretical terms may preserve reference by picking out stable entities or properties (Chakravartty, 2007; Psillos, 1999). The strategy depends on identifying a referentially stable core — what Psillos (2012) calls a "core causal description" — that survives theory change and grounds transtheoretic convergence.[10] One might think the incremental drift analysis already provides the realist with exactly such a core: the recurring role-oriented patterns recovered above preserve a stable explanatory anchor in large-scale structure, galactic dynamics, and related gravitational phenomena. But stability of functional role is not sufficient for the causal-descriptivist strategy. For that strategy to do non-trivial philosophical work, the core description must be specific enough to discriminate between candidate referents — to pick out a *kind*, not merely constrain a *role*. A description compatible with many distinct entities trivializes referential success: if 'dark matter' could refer to WIMPs, axions, primordial black holes, or modified gravity effects indifferently, claiming it successfully refers becomes vacuous — whichever turns out to exist was somehow what we meant all along.[11]

My results reveal precisely this problem. First, the embedding-space tests show statistically reliable drift between multiple adjacent periods, with the strongest local break occurring in the transition from 2001–2005 to 2006–2010. The dominant-label analyses refine this result: referential ('A') discourse shows the sharpest local rupture in the mid-2000s across adjacent cosine, MMD, sliced Wasserstein, and energy-distance measures, whereas role-oriented ('B') discourse shows the larger cumulative cosine displacement across the series. Second, cumulative cosine displacement ends above its early values despite reversals, showing that successive changes do not return the measured discourse to its initial position. Third, the collostructional profiles show that these shifts are lexically substantive: what changes is not just the position

precise, context-dependent meanings within the broader discourse. The present results support a constrained version of that picture: not indefinite sense proliferation, but patterned semantic drift together with continued specialization inside a historically persistent explanatory frame. (Aitchison, 2001; Blank, 1999; Geeraerts, 2009; Traugott and Dasher, 2001)

[10]The alternative — grounding reference via a Kripkean causal-historical chain rather than descriptive content — is not available to the scientific realist. Causal-historical theories sever the connection between reference and epistemic access that realist arguments require: if reference is secured by the initial baptism and causal chain alone, independent of any description the scientific community associates with the term, the realist cannot use referential continuity to argue that past and present theories are tracking the same entity and that accumulated knowledge is therefore cumulative. Psillos's causal descriptivism is explicitly designed to preserve this epistemic connection, which is why the present analysis targets it.

[11]See Martens (2022) and Allzén (2026) for arguments that the referential success of 'dark matter' is trivialized by its compatibility with a wide range of candidates, and that this undermines the semantic route to scientific realism in this case.

of the centroid, but the balance among recurring candidate vocabularies, detection language, and cosmological explanatory frames. These findings document changes in scientific usage. Interpreting them as evidence of unresolved ontological commitments requires the further philosophical argument about specificity developed here; a changing linguistic context alone does not establish a change of reference. The different ontological objects play the same explanatory role, but are not ontologically interchangeable. The proliferation is not merely at the level of theoretical models — different ways of characterizing the same entity — but at the level of *kinds*: particle vs. macroscopic object, fundamental vs. composite, hot vs. cold vs. warm.

A realist might respond that this simply reflects scientific uncertainty about which specific entity 'dark matter' picks out, consistent with referential stability at a more abstract level (e.g., the non-baryonic gravitationally-interacting component responsible for the dark phenomena). But this response faces a dilemma.[12] If the core description is abstract enough to encompass the full range of candidates, it becomes too coarse-grained to do non-trivial reference-fixing work — it constrains a role rather than picking out a kind. If the core description is instead specific enough to discriminate between candidates, it risks excluding live alternatives and making a hostage to fortune about which models will survive empirical scrutiny. I have argued this point at length elsewhere (Allzén, 2026): formalizing the ΛCDM-motivated core as a bundle of kind-constitutive properties yields both a failure of necessity (live dark-sector models satisfying the explanatory role but violating the core — self-interacting, dissipative, mirror, and superfluid dark matter) and a failure of sufficiency (Standard-Model neutrinos satisfy the proposed core but do not realize the causal-explanatory role). Tightening the core to discriminate between realizers presupposes precisely the kind of canonical detection and parameter determination the semantic strategy was meant to secure under empirical scarcity. Baerdemaeker (2025) formulates a compatible distinction by separating dark matter's functional role from its ontological constitution, and the empirical findings reported here recover recurring gravitational explanatory roles alongside persistent and unresolved candidate proliferation. This is not an in-principle obstacle to scientific realism, but it is an *empirical obstacle* to the semantic route that prominent realist arguments take. The specificity required for the causal-descriptivist strategy depends on a degree of empirical resolution that dark matter research has not yet achieved — and may not achieve without the kind of canonical detection that would make semantic arguments unnecessary in the first place.

6.2 The construction of continuity

The gap between research histories and distributional evidence illuminates a second implication. Section 2 showed how Bertone and Hooper's influential history constructs continuity by anchoring 'dark matter' in a dynamical lineage (Kapteyn, Jeans, Oort, Zwicky) and relegating other uses to "prehistory". Section 3 demonstrated that this narrative systematically downplays or erases the dual usage prevalent through the 1950s, in which 'dark matter' explained both

[12] A related response draws on Kitcher's (1993) distinction between *working posits* — the entities a theory is genuinely committed to in its explanatory practice — and *presuppositional posits* — background assumptions whose referential status is not under scrutiny. One might argue that 'dark matter' functions as a working posit at the cosmological level (where its gravitational effects ground stable reference) while remaining a presuppositional posit at the microphysical level. This move faces two difficulties. First, empirically: the ontological proliferation documented here occurs precisely at the level of the working posit. As competing candidates, neutralinos, WIMPs, axions, and PBHs are each advanced as the entity that actually does the explanatory work, not as background assumptions. Second, theoretically: Psillos (1997) argues that Kitcher's account does not offer a principled way to determine when historical actors were involved in different modes of reference when producing different tokens of the same expression-type, and that allowing tokens of expression-types to refer systematically to different things makes conceptual progress too easy to be philosophically interesting. Applied here, this means Kitcher's framework cannot adjudicate between the proliferating candidate claims documented in the corpus — which is precisely the problem that needs solving.

gravitational anomalies and light-absorption phenomena — often explicitly integrated by the same actors in the dynamical lineage itself (Oort 1927, Lundmark 1930, Pannekoek 1927). The modern computational findings concern a later period. They show a pronounced mid-2000s discontinuity in referential discourse together with continued candidate proliferation rather than convergence on one semantic core. These patterns are consistent with repeated conceptual renegotiation as explanatory demands changed, although they do not decide whether reference remained stable. Together with the historical reconstruction, they give reason to examine the continuity claim instead of taking its semantic basis for granted. This matters for realist arguments because it suggests that referential continuity can be as much a product of *narrative construction* as of *linguistic practice*. Research histories are epistemic: they stabilize what the community treats as the same target of inquiry by selecting which historical uses count as tracking the current referent and which count as false starts. When Bertone and Hooper (2018) present dark matter as having continuous identity from Kapteyn to contemporary WIMPs, they are not merely describing a pre-existing fact about reference — they are actively constructing an interpretive frame that makes present-day research appear as the coherent continuation of past successes.

This is not a debunking claim. Scientists may be *justified* in constructing such narratives when they facilitate productive research. The point is rather that *philosophers* should not treat the apparent continuity in those narratives as transparent evidence for referential stability. The narrative's persuasive force — its ability to present dark matter as a unified research programme — may depend precisely on backgrounding the semantic heterogeneity that distributional analysis reveals. What looks like gradual convergence from inside the narrative may look like conceptual renegotiation from outside.

6.3 Semantic stability and empirical scarcity

The third implication concerns generalizability. Is dark matter an exceptional case, or does it exemplify a broader pattern in how theoretical terms evolve under sustained underdetermination? I suggest the latter, for structural reasons. Realist semantics of the Psillos sort assume that scientific practice can furnish sufficiently specific causal descriptions to ground non-trivial reference-fixing. This assumption is plausible in domains where *canonical forms of confirmation* are available: detection via distinct causal signatures, controlled experimental isolation, or convergent measurement across independent methods. In such contexts, empirical success constrains semantic content by narrowing the space of entities compatible with observed effects. The classic examples that anchor divide-and-conquer arguments — the electron, the gene, caloric/heat — are precisely cases where such confirmation eventually obtained, allowing philosophers to retrospectively identify what succeeded in referring. But dark matter occupies a different epistemic situation. Its gravitational effects are robustly detected, yet these effects are compatible with a wide range of microphysical realizations. Non-gravitational detection has remained elusive despite decades of experimental effort. The result is that empirical constraints operate primarily at the level of *aggregate properties* (total mass, spatial distribution, equation of state) rather than *constituent nature* (particle type, interaction channels, production mechanism). I suggest that this combination of robust phenomenological evidence and persistent microphysical underdetermination helps explain the changes in scientific usage documented here. Crucially, this pattern may not be unique to dark matter. Contemporary physics increasingly confronts regimes where phenomena are well-established but constituent ontology remains multiply realizable: dark energy, quantum gravity, high-energy unification, fine-tuning problems. In such domains, theoretical terms may function more like *placeholders* for whatever entity eventually explains the phenomena than like *names* for entities whose nature is progressively clarified. The proliferation of models, the persistence of conceptually incompatible frameworks, and the absence of decisive empirical

discrimination are structural features of these research programmes, not temporary obstacles to be overcome.

If this diagnosis is correct, then the challenge dark matter poses for realist semantics may be symptomatic rather than anomalous. The semantic specificity required by causal-descriptivist strategies depends on empirical achievements — canonical detection, controlled isolation, convergent measurement — that may be unavailable not merely *yet* but *in principle* for phenomena at the current frontiers of physics. In such contexts, continuity may have to be constructed narratively because it cannot be grounded linguistically.

6.4 What would semantic stability require?

These implications do not entail anti-realism about dark matter, nor do they rationally recommend suspending belief in its existence. Gravitational evidence is overwhelming, and the explanatory indispensability of some non-luminous component is difficult to deny. The point is rather that *justifying realism via semantic arguments* faces empirical obstacles in this case — and likely in similar cases — that philosophical analysis alone cannot overcome. What would genuine semantic stability require, given our findings? At minimum: (i) convergence in ontological kind across theoretical models, not merely functional role; (ii) a drift profile that stabilizes rather than repeatedly reorganizes; (iii) usage patterns reflecting shared referential commitments rather than heterogeneous explanatory functions; (iv) referential claims tracking a narrowing candidate space rather than proliferating alternatives.

The analyses show that continuity in explanatory role coexists with changing usage and a plurality of candidate kinds. They do not by themselves determine whether reference has remained stable. This does not prove the term fails to refer — reference may be secured by causal mechanisms that outrun semantic transparency. But it does show that the *semantic route to realism* cannot simply be read off of scientific practice or constructed from selective historical narratives. It requires a level of empirical determination that dark matter has not yet achieved — and that realist philosophers should not presuppose it possesses.

7 Conclusion

The dark matter case combines stable phenomenological commitment with persistent openness about constituent nature. The central claim of this paper has been that this combination places pressure on ordinary assumptions about semantic stability. When canonical confirmation is unavailable, epistemic labor is redistributed. Evidence, historical narrative, and semantic practice no longer simply report a settled target; they participate in stabilizing what the target is taken to be. The dark matter case makes this visible from two sides at once. Historically, early usage of 'dark matter' integrated dynamical and light-obscuring functions that later continuity narratives demoted to prehistory. In the modern corpus, the term does not converge cleanly on one ontological kind, but exhibits cumulative displacement, changes in candidate-oriented discourse, and repeated lexical reorganization across a pluralized landscape in which gravitational explanatory roles recur while proposed constituents change. Read together, these results show why continuity can appear secure from within a research narrative while remaining semantically unsettled in actual usage.

This is a diagnostic result, not a debunking one. It does not show that dark matter fails to refer, nor that commitment to dark matter is irrational. It shows instead that semantic stability in active, empirically scarce research programs may be more constructed than discovered. That matters for scientific realism, but it also matters more broadly for understanding theory change in contemporary physics. Dark matter is therefore useful not only because it is philosophically controversial, but because it reveals how scientific communities maintain the sense of a stable

target when constituent nature remains radically open.

References

J. Aitchison. *Language change: Progress or decay?* Cambridge university press, 2001.

S. Allzén. Dark Matter: Explanatory Unification and Historical Continuity, 2024. URL http://arxiv.org/abs/2412.13404. arXiv:2412.13404 [physics].

S. Allzén. Dark matter realism: How referential semantics restricts realism in contemporary fundamental physics. *Synthese*, 207(6):229, 2026. doi:10.1007/s11229-026-05589-8.

I. Beltagy, K. Lo, and A. Cohan. SciBERT: A Pretrained Language Model for Scientific Text. In *EMNLP*. Association for Computational Linguistics, 2019. URL https://www.aclweb.org/anthology/D19-1371.

G. Bertone and D. Hooper. History of dark matter. *Reviews of Modern Physics*, 90(4):045002, 2018.

A. Blank. Why do new meanings occur? a cognitive typology of the motivations for lexical semantic change. In A. Blank and P. Koch, editors, *Historical Semantics and Cognition*, pages 61–90. De Gruyter Mouton, Berlin, 1999. doi:10.1515/9783110804195.61.

A. Chakravartty. *A Metaphysics for Scientific Realism: Knowing the Unobservable*. Cambridge University Press, 2007.

S. De Baerdemaeker. A Tale of Two Dark Matter Concepts. Forthcoming, 2025.

W. de Sitter. The Size of the Universe. *Publications of the ASP*, 44(258):89, 1932. doi:10.1086/124203.

J. G. de Swart. *How dark matter came to matter: A history of missing mass, 1930-1974*. PhD thesis, University of Amsterdam, 2022. URL https://dare.uva.nl/search?identifier=38ae4a9f-a5f9-4d5a-bae3-2a033f567223.

D. Geeraerts. *Theories of lexical semantics*. OUP Oxford, 2009.

M. Giulianelli, M. Del Tredici, and R. Fernández. Analysing lexical semantic change with contextualised word representations. *arXiv preprint arXiv:2004.14118*, 2020.

W. E. Harper. Current Progress in Astronomy; President's Annual Address. *Journal of the Royal Astronomical Society of Canada*, 23:119, 1929.

E. Holmberg. A Study of Double and Multiple Galaxies Together with Inquiries into some General Metagalactic Problems. *Annals of the Observatory of Lund*, 6:1–173, 1937.

E. Holmberg. On the masses and luminosities of extra-galactic nebulae. *Meddelanden fran Lunds Astronomiska Observatorium Serie I*, 180:1–13, 1952.

A. Jatowt and K. Duh. A framework for analyzing semantic change of words across time. In *IEEE/ACM joint conference on digital libraries*, pages 229–238. IEEE, 2014.

P. Kitcher. *The advancement of science: Science without legend, objectivity without illusions*. Oxford University Press, 1993.

H. Kragh. Fundamental Theories and Epistemic Shifts: Can History of Science Serve as a Guide?, 2017. Publication Title: arXiv preprint arXiv:1702.05648.

H. S. Kragh. Anachronical and diachronical history of science. In *An Introduction to the Historiography of Science*, pages 89–107. Cambridge University Press, 1987. doi:10.1017/CBO9780511622434.010.

B. Lindblad. On the dynamical theory of spiral structure: Comparison between theory and observation. *Stockholms Observatoriums Annaler*, 15:4.1–4.58, 1948.

B. Lindblad. Dyamical Interpretation of Velocity Distribution and Structural Details in the Galaxy. *Publications of Michigan Observatory*, 10:59, 1951.

B. Lindblad and J. Delhaye. On the distribution of light-intensity and colour in the spiral nebula Messier 63. *Stockholms Observatoriums Annaler*, 15:9.1–9.28, 1949.

E. M. Lindsay. Exploring Between the Stars. *Irish Astronomical Journal*, 1:8, 1950.

K. Lundmark. The Spiral Nebula Messier 33. *Publications of the Astronomical Society of the Pacific*, 33(196):324–327, 1921. doi:10.1086/123126.

K. Lundmark. The Motions and the Distances of Spiral Nebulæ. *Monthly Notices of the Royal Astronomical Society*, 85:865, 1925. doi:10.1093/mnras/85.8.865.

K. Lundmark. Über die Bestimmung der Entfernungen, Dimensionen, Massen und Dichtigkeit fur die nächstgelegenen anagalacktischen Sternsysteme. *Meddelanden fran Lunds Astronomiska Observatorium Serie I*, 125:1–13, 1930.

A. Markov. On the Nature of Spiral and Gaseous Nebulae. *Astronomische Nachrichten*, 234 (17):329, 1929. doi:10.1002/asna.19282341702.

N. C. Martens. Dark matter realism. *Foundations of Physics*, 52(1):16, 2022.

M. Martinc, P. K. Novak, and S. Pollak. Leveraging contextual embeddings for detecting diachronic semantic shift. *arXiv preprint arXiv:1912.01072*, 2019.

S. Montanelli and F. Periti. A survey on contextualised semantic shift detection. *arXiv preprint arXiv:2304.01666*, 2023.

Notes of the British Astronomical Association. Notes of the Association. *Journal of the British Astronomical Association*, 41:45–52, 1930.

J. H. Oort. Niet-lichtgevende materie in het Sterrenstelsel. *Hemel en Dampkring*, 25:13–21 and 60–70, 1927.

J. H. Oort. Problems of Galactic Structure. *Astrophysical Journal*, 116:233, 1952. doi:10.1086/145609.

E. Öpik. Stellar Distribution and the Law of Chance. *Publikationen der Kaiserlichen Universitaets-Sternwarte Jurjew*, 27:F1–F28, 1929.

A. Pannekoek. On the possible existence of large attracting masses in the centre of the galactic system. *Bull. Astron. Inst. Netherlands*, 4:39, 1927.

F. Periti, S. Picascia, S. Montanelli, A. Ferrara, and N. Tahmasebi. Studying word meaning evolution through incremental semantic shift detection. *Language Resources and Evaluation*, pages 1–37, 2024.

D. Pestre. Commemorative practices at CERN: Between physicists' memories and historians' narratives. *Osiris*, 14:203–216, 1999.

S. Psillos. Kitcher on reference. *International Studies in the Philosophy of Science*, 11(3): 259–272, 1997. doi:10.1080/02698599708573570.

S. Psillos. *Scientific realism: How science tracks truth*. Routledge, 1999.

S. Psillos. Causal descriptivism and the reference of theoretical terms. In A. Raftopoulos and P. Machamer, editors, *Perception, Realism, and the Problem of Reference*, pages 212–238. Cambridge University Press, Cambridge, 2012. doi:10.1017/CBO9780511979279.010. URL https://www.cambridge.org/core/product/53B6500CEB74D14D3A1707C3E2026072.

R. H. Sanders. *The dark matter problem: a historical perspective*. Cambridge University Press, Cambridge, 2010. Publication Title: The dark matter problem: a historical perspective.

S. Smith. No. 532. The mass of the Virgo cluster. *Contributions from the Mount Wilson Observatory / Carnegie Institution of Washington*, 532:1–8, 1936.

R. Staley. *Einstein's generation: The origins of the relativity revolution*. University of Chicago Press, 2008.

E. C. Traugott and R. B. Dasher. *Regularity in Semantic Change*. Cambridge Studies in Linguistics. Cambridge University Press, 2001.

R. J. Trumpler. The Milky Way. *Publications of the ASP*, 53(313):155, 1941. doi:10.1086/125294.

Å. Wallenquist. A study of the distribution of the stars in the Sagittarius and Ophiuchius regions of the Milky Way. *Annals of the Bosscha Observatory Lembang (Java) Indonesia*, 5:E1–E138, 1939.

A. Weiss. Dark Matter. Encyclopaedia Britannica, 2021. URL https://www.britannica.com/science/dark-matter.

H. C. Wilson. Celestial Photography at a High Altitude. *Popular Astronomy*, 12:509–515, 1904. Final page printed 215 [515].

F. Zwicky. Die Rotverschiebung von extragalaktischen Nebeln. *Helvetica Physica Acta*, 6: 110–127, 1933. doi:10.5169/seals-110267. URL https://articles.adsabs.harvard.edu/pdf/1933AcHPh...6..110Z.

F. Zwicky. On the Masses of Nebulae and of Clusters of Nebulae. *The Astrophysical Journal*, 86:217, 1937.

A Appendix

A.1 Workflow overview

The corpus was assembled from ADS records retrieved by query, then cleaned, deduplicated, and merged by bibcode before downstream annotation and embedding. The main text already reports the corpus properties that matter analytically: temporal class balance, label overlap, classifier confidence, drift, and collostructional structure. The appendix provides a compact account of the workflow and the exploratory RDF branch used to inspect explicit referential claims.

Table 1: Overview of the main analytical stages, retained settings, and paper outputs.

Stage	**Active input**	**Retained setting**	**Main output in paper**
Corpus assembly	ADS abstract spine with metadata and bibcodes	Mention extraction from abstracts; provenance indices; manifest with content hashes and bin counts	Local prepared corpus and reproducible mention table
Corpus split	Master mention table	Five-year bins; year-balanced sampling; max two mentions per bibcode; quarantined analysis/shadow corpora	Analysis corpus and shadow annotation set
Multi-label classification	Balanced analysis corpus mentions	Three labels (A, B, C), with dominant-label collapse used only for downstream drift and collostructional branches	Classifier trend, overlap, and composition figures
Embedding and substrate selection	Phrase-centered mentions from the analysis corpus	Contextual SciBERT embeddings; five layer modes; reducer search over PCA, GRP, UMAP, autoencoder, and VAE; retain PCA-100 all-mean	Shared representation used by later inferential stages
Embedding-space drift	Phrase-centered SciBERT embeddings of dark matter mentions	Shared PCA-100 all mean space; adjacent and cumulative drift; bootstrap CI; grouped permutation tests on dominant A/B split	Drift and significance figures
Collostructional analysis	Same balanced analysis corpus	Category-conditioned lexical attraction over five-year bins for dominant A and B	Heatmaps and lexical timelines
RDF extraction	Abstract-level triplet extraction and normalization	Prompt-guided extraction, ontology cleaning, and reproducible consolidation	Alluvial and time-resolved referent figures

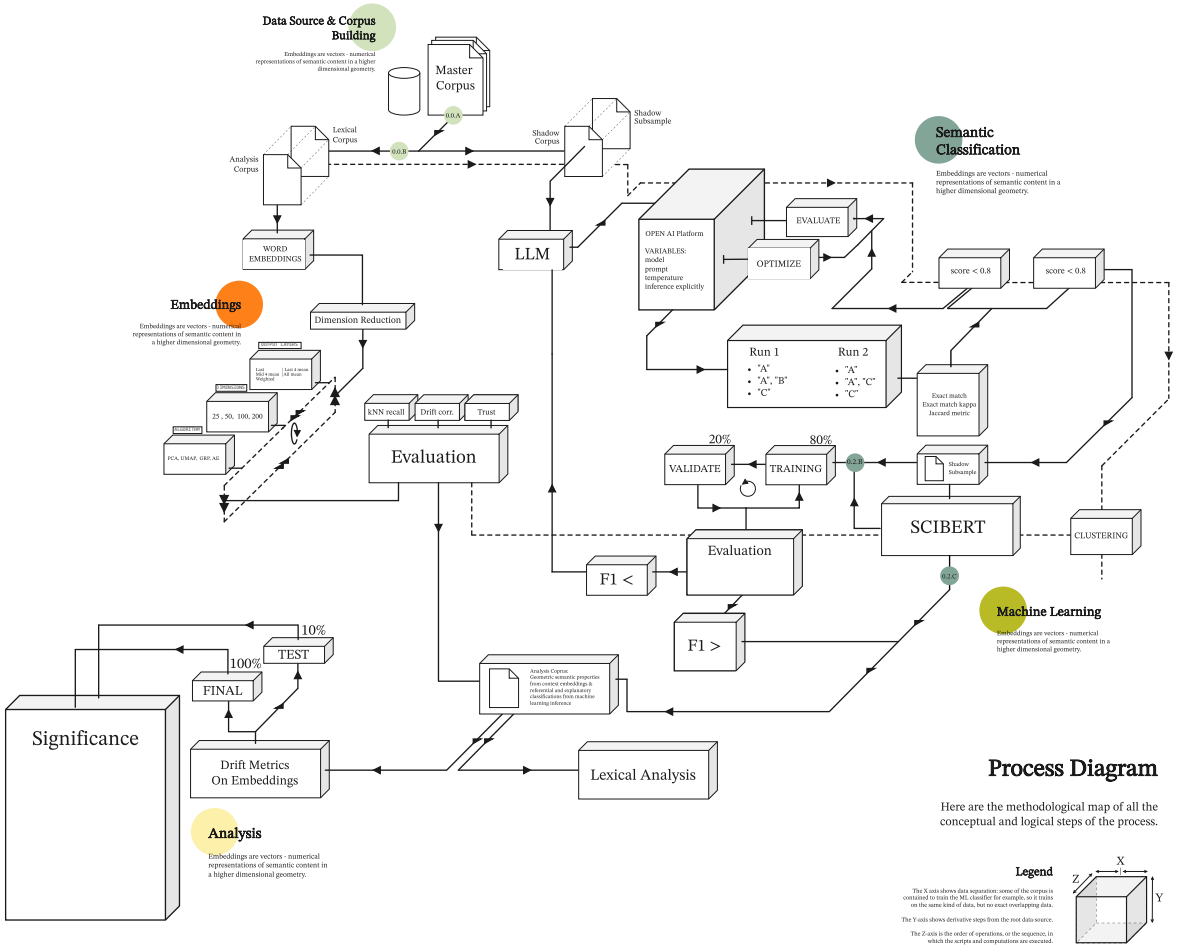


Figure 11: Workflow development, including the earlier repeat-label check and exploratory clustering. The retained analysis uses the later GPT-5.4 seed campaign and PCA-100 all-layer-mean representation; the preceding table specifies its stages and settings.

A.2 RDF extraction

The RDF branch was developed as a supplementary literalization of explicit ontological and candidate claims, not as a second main inferential pipeline. Revisions to the extraction prompt and normalization rules were checked on held-out batches until the output was regular enough for downstream aggregation. A retained 10,000-abstract supplementary sample from the post-1990 literature was then submitted to an OpenAI model through the REST API to extract compact subject-predicate-object triplets bearing on the nature, constitution, or candidate identity of dark matter. Predicate types were constrained to a small curated set, and the returned triplets were normalized through a shared ontology layer that harmonized abbreviations, candidate labels, and a limited set of predicate ambiguities. Earlier manual OpenRefine passes were retained as an audit trail for consolidation decisions, and the retained workflow is notebook-driven and reproducible. The resulting supplementary dataset contains just over 5,200 cleaned triplets expressing explicit propositions about what dark matter is taken to be.

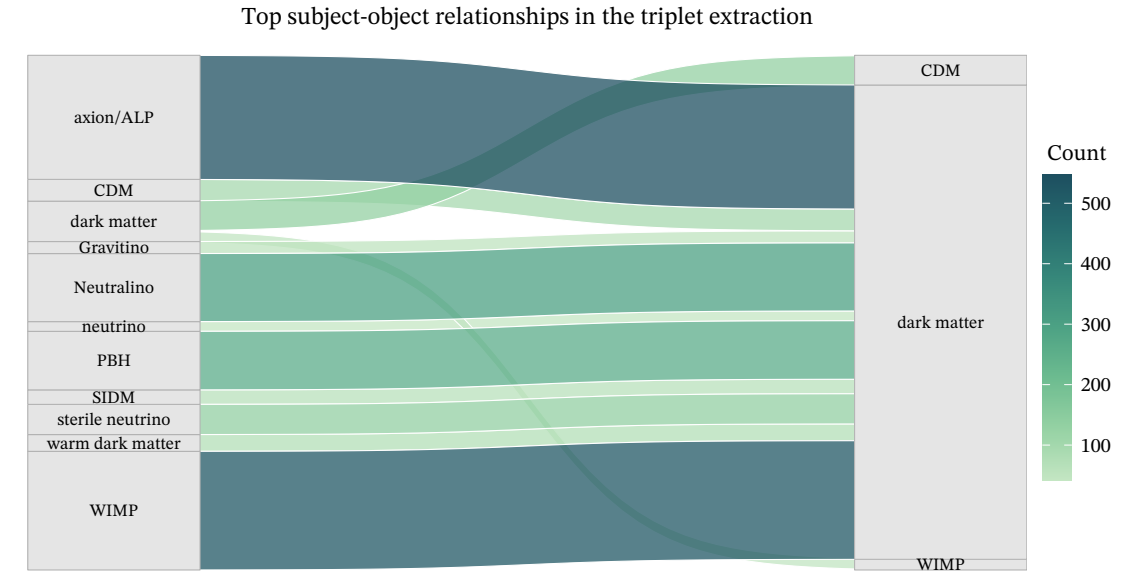


Figure 12: Appendix figure. Alluvial diagram showing the most frequently used models as referents for 'dark matter', omitting time.

The relation format is a compact subject → predicate → object representation, for example *dark matter* → **is** → *weakly interacting massive particle* or *axions* → **constitutes** → *dark matter*. Since constitutive predicates are not always structured as *X is Y* propositions, 'dark matter' appears as both subject and object. The following two figures are therefore best read as descriptive supplements to the main drift and collostructional results: they literalize the familiar candidate landscape rather than adding an independent layer of semantic evidence.

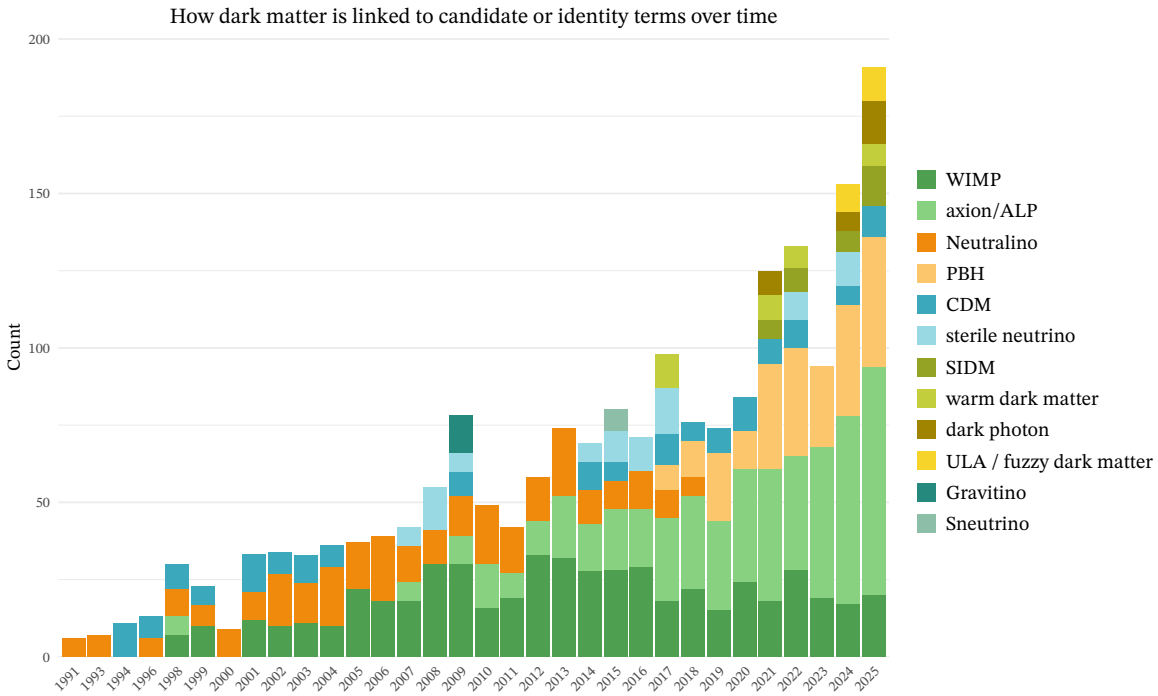


Figure 13: Appendix figure. Plot showing the most frequently used models as referents for 'dark matter' over time.

Read conservatively, these appendix figures support the same ontological-proliferation diagnosis developed in the main text, but they do not extend it very far beyond what the stronger pipeline already shows. The alluvial view is not dominated by one stabilized constitutive mapping, but by several incompatible candidate attributions that recur with appreciable weight. The time-resolved stack likewise suggests succession and coexistence rather than progressive narrowing toward a single canonical referent. The RDF branch is a descriptive supplement to the drift and collostructional analyses.

A.3 Exploratory concept network

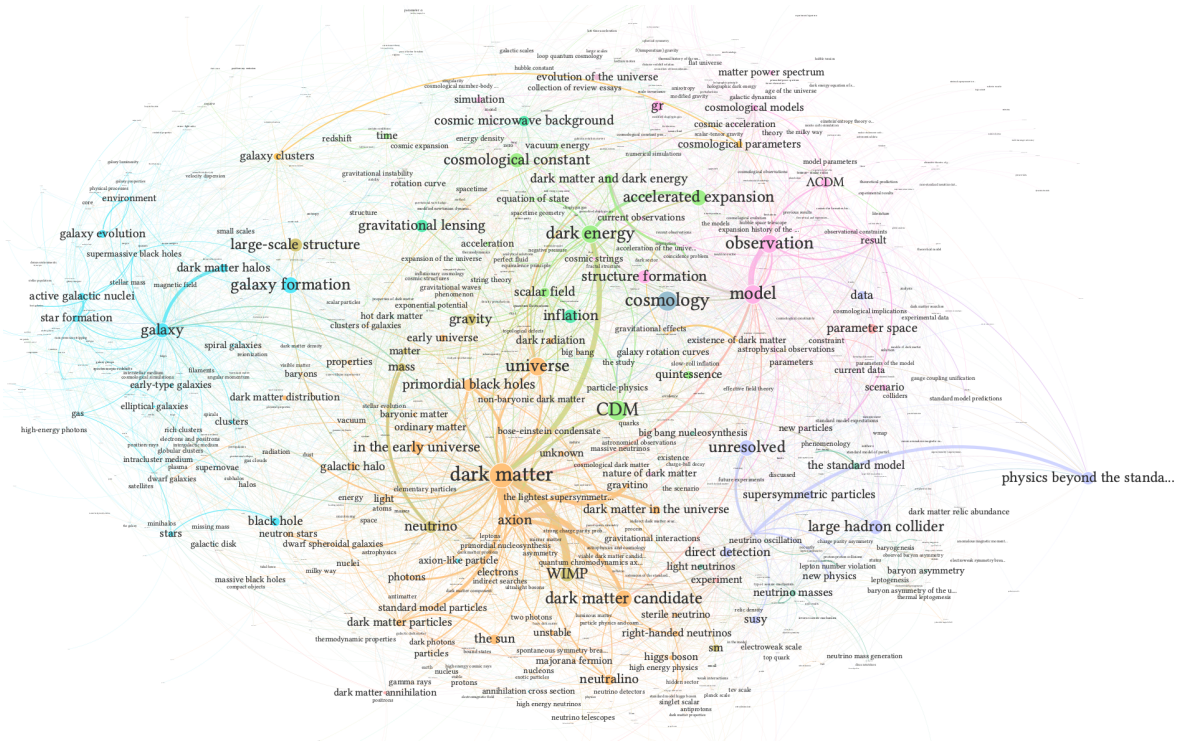


Figure 14: Exploratory concept network derived from the wider semantic extraction branch.

In addition to the narrower constitutive RDF extraction, I ran a wider conceptual extraction and visualized it as a network in Gephi. Nodes were filtered by degree to remove sparse concepts, scaled by eigenvector centrality, and grouped with modularity-based community detection. The layout was generated with Force Atlas 2. This figure is exploratory rather than inferential: its purpose is to show broad conceptual neighborhoods and bridges in the extracted relation space.