Meaning at the Planck scale?

Contextualized word embeddings for doing history, philosophy, and sociology of science*

Arno Simons

1. Introduction

Scientific concepts are dynamic entities, continually shaped and reshaped by the sociohistorical contexts from which they emerge. Their meanings are influenced by the intellectual, cultural, and material conditions of their time, as well as by the disciplinary frameworks that govern their usage. This dynamic nature is evident in medical concepts such as "syphilis" (Fleck, 1979) and "atherosclerosis' (Mol, 2002), physical concepts like "electricity" (Steinle, 2016) and "temperature" (Chang, 2007), as well as more abstract methodological notions like "probability" (Hacking, 1975) and "objectivity" (Daston and Galison, 2007).

From the perspective of the history, philosophy, and sociology of science (HPSS), understanding how scientific concepts acquire and stabilize their meanings is crucial for several reasons. Changes in scientific language reflect shifts in how researchers conceptualize their subject matter, and as concepts evolve, they reshape the frameworks through which reality is interpreted. By examining these processes, HPSS highlights how scientific knowledge is not merely discovered but constructed through complex social and intellectual negotiations, emphasizing the contingent and dynamic nature of scientific understanding (Fleck, 1979; Foucault, 1970; Hacking, 1999; Kuhn, 1962; Latour, 1987; Pickering, 1995). As Thomas Kuhn (1962: 128) famously remarked:

> "the Copernicans who denied its traditional title 'planet' to the sun were not only learning what 'planet' meant or what the sun was. Instead, they were changing the meaning of 'planet' so that it could continue to make useful distinctions in a world where all celestial bodies, not just the sun, were seen differently from the way they had been seen before".

* This chapter is almost identical to an earlier preprint (Simons, 2024b).

Transformations in the meanings of scientific concepts can also have significant societal repercussions. When concepts shift, they alter not only the internal structures of scientific knowledge but also the broader societal and cultural frameworks in which that knowledge is embedded. For example, historical changes in the meanings of terms like "nature" (Merchant, 1980) or "the normal and the pathological" (Canguilhem, 1991) have profoundly influenced education, policy, and social hierarchy. These shifts have reinforced or challenged societal norms and power dynamics. Redefining scientific concepts creates new frameworks of understanding that affect how institutions are organized, how resources are allocated, and how humans and non-human entities are categorized and treated (Bowker and Star, 1999; Clarke et al., 2010; Hacking, 1999).

To study the evolving meanings of scientific concepts across different socio-historical contexts, HPSS scholars have primarily relied on qualitative methods such as "close reading". While effective, these methods are often time-intensive, challenging to replicate, and limited in scalability. To complement these approaches, HPSS scholars have increasingly turned to "distant reading" techniques, such as co-word analysis (Callon et al., 1983), latent semantic analysis (LSA) (Deerwester et al., 1990), latent Dirichlet allocation (LDA) (Blei et al., 2003), and word vectors (Mikolov et al., 2013). These methods enable researchers to explore conceptual changes across large corpora, uncovering patterns in the evolution of individual terms or entire fields (Boyack et al., 2011; Chen et al., 2018; Courtial and Law, 1989; Lean et al., 2023; Leydesdorff and Rafols, 2009; Laubichler et al., 2019; Malaterre and Léonard, 2024; Overton, 2013; Pence and Ramsey, 2018; Rip and Courtial, 1984; Venturini et al., 2014; Wevers and Koolen, 2020). However, these approaches often struggle to capture context-dependent meanings of words and sentences.

This is where contextualized word embeddings (CWEs) provide a significant leap forward. CWEs rely on patterns of usage in large text corpora to encode words as points in a continuous vector space. In this space, semantic and syntactic relationships between words are represented through relative distances, with similar contexts grouped closely together. Unlike earlier word embeddings, which assign static representations to words, CWEs dynamically adjust a word's representation depending on its context. This adaptability enables researchers to analyze how meanings differ and evolve across time, disciplines, or other settings, surpassing the static associations revealed by traditional methods (Zichert and Simons, 2026).

To demonstrate the potential of CWEs for HPSS, I use "Planck" as a test case, a term with multiple distinct meanings in physics—such as the physicist Max Planck, Planck units, and the Planck mission—all well-defined and widely recognized within the physics community. This diversity of known interpretations offers a clear "ground truth" for evaluating how effectively CWEs can capture and disambiguate these meanings, while striking a balance between simplicity and complexity. While some senses, like "Max Planck Society" or "Fokker-Planck equation", can be disambiguated through straightforward co-occurrence patterns, others demand a deeper contextual understanding. For example, domain-adapted BERT models can discern synonymous meanings in contexts such as "a Planck function at that temperature" and "the spectral distribution for Planck's law", while also interpreting nuanced cases like "Planck's experiment yields Ps $\times$ 10–9 for the power of scalar modes at the end of inflation" as referring to the Planck space mission rather than Max Planck the physicist. The analysis focuses on five BERT-based

models with varying domain specializations and training methodologies, including both general-purpose models and domain-specific models fine-tuned on scientific corpora. Among these is my own Astro-HEP-BERT (Simons, 2024a), a specialized model developed for conceptual analysis in astrophysics and high-energy physics (HEP), available on Hugging Face.[1]

As you may have already guessed, the title of this chapter is a play on words. In physics, the term "Planck scale" often refers to extremely small quantities of space and time, defined using natural units introduced by Max Planck in 1899. Here, "meaning at the Planck scale" serves as a metaphor for the linguistic task at hand: disambiguating the multiple senses of "Planck". Just as physics requires high resolution to measure reality at the smallest scales, understanding the contextual meanings of "Planck" requires CWEs to detect subtle semantic distinctions at the level of individual words and subwords.

The structure of this chapter is as follows. First, I introduce CWEs as a computational tool for HPSS, highlighting their advantages over traditional methods and the benefits of BERT-based architectures. Next, I describe the datasets and labels used, providing essential context for the analysis. This is followed by a detailed comparison of the five BERT-based models used in the empirical evaluation. The core of the paper consists of four empirical sections. The first evaluates CWEs' ability to disambiguate the term "Planck" through supervised word-sense prediction. The second assesses the quality of sense clusters formed in an unsupervised setting, measuring alignment with labeled data. The third examines the balance between cluster separation and internal cohesion. Finally, the fourth empirical section explores the diachronic evolution of "Planck", tracking shifts in the prominence of different senses over time. The paper concludes by synthesizing the findings and discussing their broader implications for HPSS research. It highlights how CWEs can illuminate the socio-historical evolution of scientific concepts, identifies their limitations, and proposes future directions for integrating CWEs into the study of scientific discourses.

2. CWEs as a new tool for HPSS

CWEs represent each word in a given sequence, such as a sentence or paragraph, as a unique numerical vector in a continuous embedding space. This space encodes relationships and patterns derived from large text corpora, where words with similar meanings or contexts are positioned closer together. By dynamically adjusting a word's representation based on its specific context, CWEs capture subtle differences in meaning, distinguishing between various senses of the same word or phrase. For example, the word "Planck" in the sentence "We used Planck's law to calculate the radiation emitted by a black body" will have a different vector representation than in "We used Planck's data to map the early universe", reflecting the distinct contexts and meanings of the word in each case.

The ability to generate context-aware embeddings is a hallmark of modern transformer-based architectures, particularly BERT (Bidirectional Encoder Representations

1 https://huggingface.co/arnosimons/astro-hep-bert

from Transformers) (Devlin et al., 2018). BERT's bidirectional design processes text by considering both the preceding and succeeding words in a sentence simultaneously. Through multiple layers of neural networks and attention mechanisms, it captures contextual relationships, making it especially well-suited for deep language understanding. This bidirectional approach gives BERT an edge over unidirectional models like GPT for analyzing complex linguistic contexts, though it is less effective for generative tasks.

CWEs lie at the heart of BERT's outstanding performance in natural language understanding tasks. As BERT is most often used for high-level applications like text classification, named entity recognition, or question answering, its embeddings normally remain "under the hood". However, CWEs can be accessed by extracting hidden states from BERT's intermediate or final layers. A common approach involves averaging the last four layers to create robust word embeddings (Devlin et al., 2018; Periti and Montanelli, 2024).

Though originally designed for higher-level tasks, CWEs are particularly well-suited for exploring the varying and evolving meanings of scientific concepts. Advances in word sense disambiguation (WSD) (Bevilacqua et al., 2021; Loureiro et al., 2020), word sense induction (WSI) (Periti and Montanelli, 2024; Sun and Platos, 2023), and lexical semantic change (LSC) analysis (Periti and Montanelli, 2024; Tahmasebi and Dubossarsky, 2023) suggest their capacity to uncover distinct meanings that scientific terms acquire across disciplines and historical contexts (Ahmadi, 2026; Kleymann et al., 2022; Zichert et al., 2025). WSD identifies the intended sense of a word in context, WSI clusters word occurrences to infer senses unsupervised, and LSC analyzes shifts in word meanings over time. LSC can be considered a temporal extension of both WSD and WSI when it uses information about a word's senses to measure semantic change, for example by tracking how changes in the relative prominence of these senses reflect changes in the word's overall meaning (Periti and Montanelli, 2024).

WSD, WSI, and LSC all use BERT's ability to map semantically similar words to similar vectors, enabling the modeling of semantic relationships. These tasks largely rely on three core operations:

Measuring distances between embeddings: This operation quantifies semantic similarity or divergence between terms, typically using cosine similarity or Euclidean distance. It can be used to compare either the same term in different contexts (as in WSD and WSI) or different terms within the same context. Additionally, average pairwise similarity or distance across a set of embeddings reveals broader properties, such as the degree of polysemy, by reflecting the dispersion of meanings.

Prototypical embedding creation: By averaging the embeddings of multiple instances of a term, researchers can derive a prototypical representation of its central or typical meaning. Depending on how instances are grouped (e.g. by predefined sense labels for WSD, discovered clusters for WSI, or temporal periods for LSC), prototypical embeddings can reveal patterns specific to these contexts.

Clustering embeddings: Clustering groups semantically similar terms or usages, making it valuable for WSI to uncover known or emerging senses, for WSD to generate prototypes and predict senses, and for LSC to track how senses evolve over time by grouping embeddings from different periods or contexts.

Effectively using CWEs for HPSS requires understanding domain adaptation, which customizes BERT models for specialized contexts. BERT's pretraining relies on masked language modeling, in which certain words in a sentence are hidden and predicted based on surrounding context. This process fine-tunes the model's internal parameters through extensive training on large, diverse corpora. However, while pretrained BERT models are robust general-purpose tools, they have found to underperform in the sciences, where specialized language dominates (Beltagy et al., 2019; Lee et al., 2020). Two main strategies address this limitation:

Training from scratch: A model is trained entirely on domain-specific data. This approach allows for a custom vocabulary tailored to the target domain, enhancing the model's ability to handle specialized terms. However, it requires substantial domain-specific data and significant computational resources (Grezes et al., 2022; Gu et al., 2021; Hellert et al., 2024).

Re-using weights from pretrained models: Pretrained models such as BERT-base are further trained on domain-specific corpora. This method efficiently builds on the general linguistic knowledge captured during pretraining, adapting the model to domain-specific terminology and contexts. Although this approach is resource-efficient, it typically reuses the original vocabulary, which may limit the model's ability to fully represent highly specialized terms. Subword tokenization helps mitigate this issue but cannot entirely address it (Devlin et al., 2018).

The choice between these approaches depends on the availability of data and computational resources. Additional training of a pretrained model is often sufficient for many applications—as demonstrated in this paper by the comparative performance of Astro-HEP-BERT against models like PhysBERT and astroBERT, which were trained from scratch—while training from scratch may be necessary for highly specialized tasks that demand more precise domain knowledge (Grezes et al., 2022, 2021; Hellert et al., 2024).

3. Datasets and labels

For this study, I use two custom-compiled datasets.

The Astro-HEP Corpus[2] consists of 21.84 million paragraphs extracted from 600,000 scientific articles (approximately 4.2 billion tokens) published on arXiv between 1986 and 2022, spanning the domains of astrophysics and high-energy physics (HEP). The para-

2 https://huggingface.co/datasets/arnosimons/astro-hep-corpus

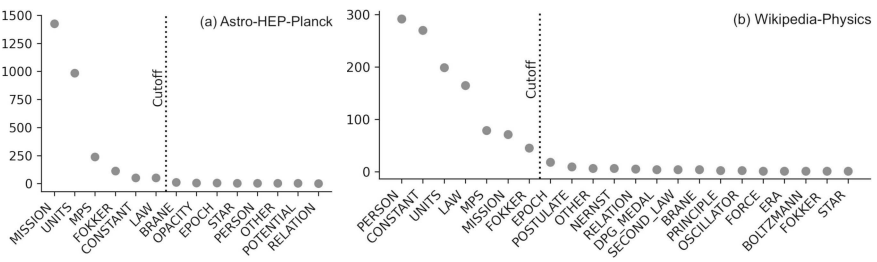
graphs were extracted in plain text format from the original LaTeX files using Pandoc[3]. To ensure data quality, I replaced literature references with "[CIT]" and multiline formulas with "FORMULA". Additionally, paragraphs that were either unusually short or exhibited an abnormal frequency of whitespace characters were filtered out based on character frequency analysis.

The Wikipedia-Physics Corpus[4] contains 102,409 paragraphs extracted from 6,642 key physics-related Wikipedia articles. These articles were selected using the PetScan[5] tool, which generated a list of all pages categorized under "physics" or its immediate subcategories. Markup was removed and minimal cleaning applied to produce plain text paragraphs, and while references were removed, all formulas were retained.

To establish a ground truth for the WSD and LSC analyses of "Planck" (Sections 5–8), I annotated occurrences of the term in both corpora using 21 predefined labels, determined by theoretical insights, expert knowledge, and a qualitative review of the contexts. In the Astro-HEP Corpus, I manually labeled 2,932 occurrences from a random sample of 1,500 paragraphs, and this subset is referred to as the Astro-HEP-Planck Corpus[6]. In the Wikipedia-Physics Corpus, I labeled all 1,186 occurrences across 885 paragraphs. I identified occurrences using a case-insensitive regular expression to capture variations like "PLANCK" and "Planck(2015)" while excluding irrelevant forms like "planckian".

For further analysis, I focused on occurrences corresponding to the seven most common labels (see Table 1 and Figure 1). For each corpus, I created six subsets containing the two, three, four, five, six, and seven most common labels within the corpus, respectively. For each subset, I extracted all CWEs by averaging the model's final four hidden layers (Devlin et al., 2018; Periti and Montanelli, 2024).

*Figure 1: Distribution of labels and their cutoff thresholds for the word sense disambiguation and induction tasks in the Astro-HEP-Planck (**a**) and Wikipedia-Physics (**b**) corpora. The x-axis shows the labels ranked by frequency, while the y-axis represents the number of occurrences. Cutoffs indicate the number of labels included in each subset for subsequent analyses.*



3 https://pandoc.org

4 https://huggingface.co/datasets/arnosimons/wikipedia-physics-corpus

5 https://petscan.wmcloud.org

6 https://huggingface.co/datasets/arnosimons/astro-hep-planck-corpus

Table 1: Description of the seven most common labels (in alphabetical order) used for word sense disambiguation of the term "Planck". Columns AHC and WPC depict the number of labeled occurrences in the Astro-HEP-Planck Corpus and the Wikipedia-Physics Corpus, respectively.

Label	AHC	WPC	Description
CONSTANT	52	270	Planck's constant h: A fundamental physical constant of foundational importance in quantum mechanics. A photon's energy is equal to its frequency multiplied by the Planck constant, and the wavelength of a matter wave equals the Planck constant divided by the associated particle momentum.
FOKKER	112	45	Fokker—Planck equation: A partial differential equation that describes the time evolution of the probability density function of the velocity of a particle under the influence of drag forces and random forces, as in Brownian motion.
LAW	52	165	Planck's law: A formula that describes the spectral density of electromagnetic radiation emitted by a black body in thermal equilibrium at a given temperature, when there is no net flow of matter or energy between the body and its environment.
MISSION	1426	71	Planck mission: A space observatory operated by the European Space Agency from 2009 to 2013. The project that mapped the anisotropies of the cosmic microwave background (CMB) at microwave and infrared frequencies, with high sensitivity and angular resolution.
MPS	239	79	The Max Planck Society for the Advancement of Science: A formally independent non-governmental and non-profit association of German research institutes
PERSON	4	292	Max Planck: The famous German theoretical physicist whose discovery of energy quanta won him the Nobel Prize in Physics in 1918.
UNITS	986	199	Planck units: A system of units of measurement defined exclusively in terms of four universal physical constants: $c, G, \hbar,$ and kB. Expressing one of these physical constants in terms of Planck units yields a numerical value of 1.

4. Models

I compare the performance of five BERT-based models: The first model, BERT-base or simply BERT (Devlin et al., 2018), is a general-purpose model trained from scratch over 40 epochs on a dataset of Wikipedia articles and free books, totaling 3.3 billion tokens. It uses a 30,000-token uncased vocabulary and serves as the baseline model for this analysis. The second model, SciBERT-scivocab-uncased or simply SciBERT (Beltagy et al., 2019), was developed specifically for scientific text and trained from scratch on 1.14 million biomedical and computer science papers, with its uncased vocabulary learned directly from the corpus. The third model, astroBERT (Grezes et al., 2022), was trained on 400,000 astrophysics publications from the SAO/NASA Astrophysics Data System (ADS), specifically designed to identify, disambiguate, and tag entities in astrophysical litera-

ture. Its vocabulary is case-sensitive and was learned from the corpus. The fourth model, PhysBERT-uncased or simply PhysBERT (Hellert et al., 2024), was trained from scratch over 10 epochs on 1.2 million arXiv physics papers, with its uncased vocabulary learned directly from the corpus. The model was also fine-tuned to produce domain-adapted sentence embeddings.

The fifth model, Astro-HEP-BERT, is my own (Simons, 2024a). Instead of training it from scratch, as was done for the other four models, I reused BERT's learned linguistic patterns and vocabulary and retrained the model for three additional epochs using on my Astro-HEP Corpus. To enhance learning, I trained the model solely on entire paragraphs rather than packing in as many sentences as possible, as often suggested in BERT tutorials. This "full-paragraphs format" preserves sentences within their original context, which is especially meaningful in academic writing where paragraphs focus on one idea. Previous research shows that using multiple sentences from the same document improves results (Liu et al., 2019), so I expected this approach to help the model capture more semantic information.

5. Word sense prediction using sense prototypes

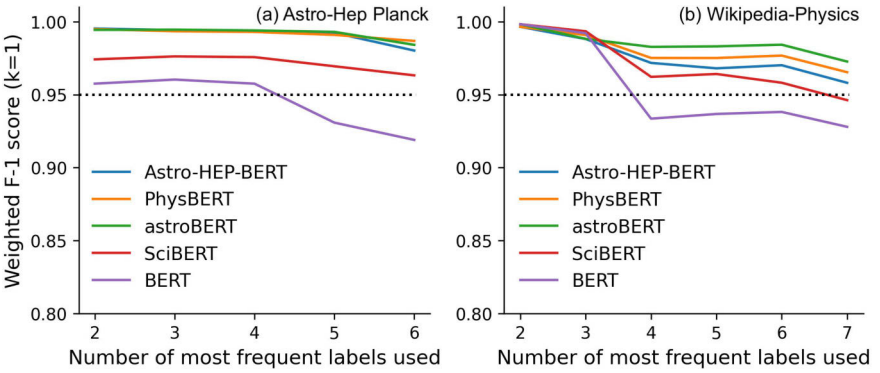
As a first WSD test I use a simple supervised 1-nearest neighbor (1NN) approach as described in (Loureiro et al., 2020). This test compares the models' ability to predict the correct sense of "Planck" occurrences by computing the cosine similarity between their CWEs and precomputed sense prototypes. The closest prototype is then used to assign the appropriate sense to the target occurrence. Failure to classify an occurrence correctly suggests that the CWEs for the senses are not well-separated. For each corpus, the sense prototypes are generated by averaging the CWEs of all labeled "Planck" occurrences corresponding to each sense.

Figure 2 compares the models' performance in predicting different senses of "Planck" using word sense prototypes across the two corpora. Each model was evaluated using six separate 1NN classifiers, with each classifier trained on a subset containing between two and seven of the most common labels in the corpus. Figures 2a and 2b present the weighted F-1 scores for the Astro-HEP-Planck and Wikipedia-Physics corpora, respectively. The x-axis represents the number of labels in the subset, while the y-axis shows the F-1 score for each model. This setup enables a direct comparison of how effectively the models distinguish between the senses of "Planck" as the classification task becomes more complex with the inclusion of additional senses (labels).

In both figures, we observe that the classifiers generally achieve higher F-1 scores when they have to predict fewer labels, with performance declining as the number of labels—and thus ambiguity—increases. This in itself should not be surprising because fewer labels make the classification task simpler, reducing ambiguity, whereas more labels increase complexity and the likelihood of misclassification. A minor exception occurs in the Wikipedia-Physics Corpus, where all models show a slight performance rebound after an initial decline when predicting the four most common labels. The initial decline in the four-label case is primarily due to misclassifying CWEs labeled as LAW—the newly introduced fourth most frequent label—as either CONSTANT or

PERSON, and vice versa. As the labels introduced in the five-, six-, and seven-label settings are less ambiguous, the overall F-1 scores improve slightly.

*Figure 2: Comparison of model performance in disambiguating the word "Planck" using 1-nearest neighbor (1NN) classifiers across subsets of labels from the Astro-HEP-Planck (**a**) and Wikipedia-Physics (**b**) corpora. The x-axis represents the number of labels in each subset, while the y-axis shows the weighted F-1 scores. Each model was evaluated on classifiers trained on CWEs extracted for two to six labels in the Astro-HEP-Planck Corpus and two to seven labels in the Wikipedia-Physics Corpus.*



The results also highlight differences in how well the models handle the challenge of dealing with increasing numbers of labels and ambiguity: some maintain relatively high performance even as the number of labels grows, while others experience a steeper decline. Overall, PhysBERT, astroBERT, and Astro-HEP-BERT perform exceptionally well on the Astro-HEP-Planck Corpus, consistently achieving F-1 scores above 0.98, with most scores exceeding 0.99. SciBERT follows, with F-1 scores generally around 0.97. The weakest performer in the Astro-HEP-Planck Corpus is BERT, starting with an F-1 score near 0.96 and dropping to 0.92 in the six-label case. In the Wikipedia-Physics Corpus, astroBERT outperforms the other models, maintaining F-1 scores above 0.97. PhysBERT and Astro-HEP-BERT follow closely, with scores above 0.96 and 0.95, respectively. Although BERT starts strong with F-1 scores near 0.99 in the two-and three-label cases, its performance declines sharply, dropping to around 0.93 in the more complex label settings.

While BERT shows the weakest performance across both corpora, its F-1 score never falls below 0.92, reflecting the strong baseline performance of BERT models in word sense disambiguation (Wiedemann et al., 2019). In contrast, the stronger performance of astroBERT, PhysBERT, and Astro-HEP-BERT in disambiguating the term "Planck" may suggest the advantages of domain-specific pretraining.

The results of the word sense prediction task showed that domain-specific models outperform general-purpose ones in distinguishing between distinct senses of "Planck". However, sense disambiguation often requires unsupervised approaches, especially

when sense labels are unknown. The next section evaluates the models' capacity to cluster senses effectively without supervision.

6. Cluster purity

While the previous section evaluated the models' ability to classify "Planck" occurrences into predefined senses, a classical WSD task, this section evaluates the models' ability to perform WSI by forming meaningful word sense clusters in an unsupervised manner employing the K-means clustering (cf. Periti and Montanelli, 2024; Sun and Platos, 2023). This algorithm partitions data into a predefined number of k clusters by initializing centroids (usually randomly) and iteratively refining them to minimize variance within each cluster, based on the average distance between data points and their nearest centroid. For each corpus, model, and subset of labeled CWEs, k is set to the number of labels l in the subset. For each value of k, the best clustering solution out of 100 random seed initializations is selected based on inertia—the sum of squared distances between CWEs and their nearest cluster center.

To evaluate how effectively a clustering solution s disambiguated the predefined senses of "Planck", I propose a purity indicator, using a coefficient of variation (CV) approach, where I compare the weighted average CV for the clusters against a theoretical maximum that can occur if clusters are perfectly pure. The CV is a standardized measure of dispersion of frequency distribution, defined as the ratio of the standard deviation σ to the mean μ. Building on this definition, my **purity indicator** is computed as:

$$\text{purity}_s = \frac{\text{waCV}_s}{\text{TM}_l}$$

where *waCV*$_s$is the weighted average CV for the clusters in solution s, and *TM*$_l$ the theoretical maximum value that any empirical waCV$_s$ can attain, given l available labels. The waCV$_s$ is computed as:

$$\text{waCV}_s = \sum_{i=1}^{k} w_i' \text{CV}_i$$

with k as the number of clusters, w_i' the normalized size of cluster i, and CV$_i$ the CV of the label distribution in cluster i, defined as:

$$\text{CV}_i = \frac{\sigma_i}{\mu_i}$$

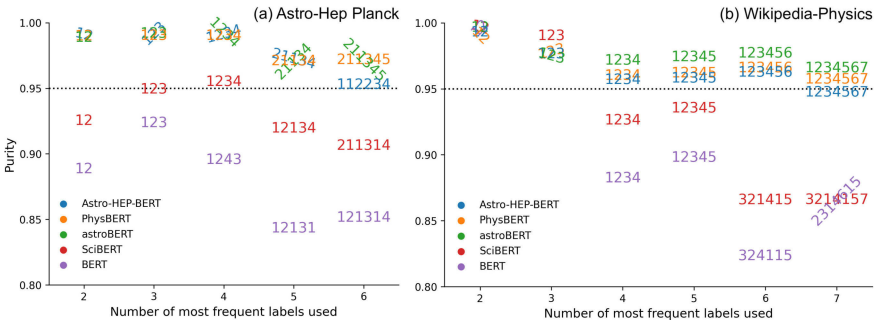
where σ_i is the standard deviation and μ_i the mean of the label distribution in cluster i. The theoretical maximum TM$_l$ for waCV$_s$, given l labels, occurs when all clusters are "pure", meaning they contain only one type of label. Thus, we can simplify to a single-cluster-single-label scenario:

$$\text{TM}_l = \frac{\sigma_f((1,0^{l-1}))}{\mu_f((1,0^{l-1}))}$$

A high purity score indicates that clusters are mostly composed of instances carrying the same label, whereas a lower score suggests mixing of labels, indicating less effective dis-

ambiguation. Figures 3a and 3b display the purity scores of cluster solutions across models and subsets for the Astro-HEP-Planck and Wikipedia-Physics corpora. The x-axis represents the subset label count, while the y-axis shows the purity score. Each cluster solution is annotated with a permutation of dominant labels, reflecting their frequency in the corpus. For instance, "1234" represents clusters dominated by the top four labels in descending order of frequency. By contrast, "1243" (seen in BERT's four-label solution for the Astro-HEP-Planck Corpus) indicates that the third-largest cluster is dominated by the fourth label rather than the third, suggesting an artificial size increase due to "pollution" from other labels.

*Figure 3: Purity scores of clustering solutions across models and subsets of CWEs for the Astro-HEP-Planck (**a**) and Wikipedia-Physics corpora (**b**). The x-axis represents the number of predefined word sense labels in each subset, and the y-axis shows the purity scores for each model. Clustering solutions are annotated with permutations of dominant labels, reflecting the frequency distribution of word senses in the datasets. Detailed label mappings can be inferred from figures 1a and 1b for the Astro-HEP-Planck and Wikipedia-Physics corpora, respectively.*



Results indicate distinct model performance patterns. In the Astro-HEP-Planck Corpus, Astro-HEP-BERT, PhysBERT, and astroBERT achieve the highest clustering quality, consistently above 0.95, and even exceed 0.99 with fewer labels. By contrast, SciBERT and BERT peak at 0.96 and 0.92, respectively, with performance declining faster as label complexity increases. Up to four labels, most models maintain the corpus's frequency order, with BERT as an exception. In the five- and six-label cases, however, all models deviate from frequency order, indicating challenges as complexity rises. Here, Astro-HEP-BERT, PhysBERT, and astroBERT tend to form multiple clusters dominated by labels like MISSION or UNITS, indicating potential context-specific subgroups within dominant senses. SciBERT and BERT also show overrepresentation, producing up to three MISSION-dominant clusters in the five- and six-label cases.

These trends may stem from domain-specific pretraining in Astro-HEP-BERT, PhysBERT, and astroBERT, enhancing their sensitivity to physics terms like "Planck". Conversely, SciBERT and BERT lack this focus and may thus merge distinct senses. Additionally, the K-Means algorithm, designed to minimize variance around centroids, may contribute to overrepresenting dominant senses by favoring larger clusters around common labels, often at the expense of rarer ones. Thus, beyond model-specific factors, the

clustering method itself may introduce a bias toward frequent senses, influencing observed patterns

In the Wikipedia-Physics Corpus, astroBERT consistently outperforms, maintaining purity scores above 0.97, with PhysBERT and Astro-HEP-BERT closely following. SciBERT achieves the highest purity (0.99) in the three-label scenario, but its performance drops in six- and seven-label cases, while BERT shows inconsistent results. Models perform well in two- to five-label cases, aligning closely with corpus frequency order. Beyond five labels, deviations appear, particularly for SciBERT and BERT, which form multiple clusters dominated by the PERSON label in the six-and seven-label scenarios, highlighting challenges with less frequent senses.

Overall, astroBERT, PhysBERT, and Astro-HEP-BERT demonstrate b clustering quality, while BERT and SciBERT fall behind. Although the top three models maintain high purity scores across both corpora, they more accurately reproduce the label distribution in the Wikipedia-Physics Corpus, likely due to its more balanced label frequencies (see Figure 1), more conventional language, and clearer semantic boundaries. In contrast, the specialized Astro-HEP-Planck Corpus, with its technical language and niche contexts, seems to present challenges for all models and for K-Means in accurately distinguishing less frequent senses.

7. Cluster separation and cohesion

In addition to purity, I evaluate the balance between internal cluster cohesion and inter-cluster separation to better assess the quality of clustering solutions across different models, particularly in cases where purity scores were similar. To measure inter-cluster separation, I calculate the **average pairwise similarity (APS)** between CWEs from different clusters:

$$\mathrm{APS}(E_1,E_2) = \frac{1}{|E_1||E_2|} \sum_{e_{1,i}\in E_1, e_{2,j}\in E_2} \mathrm{CS}(e_{1,i},e_{2,j})$$

where E_1 and E_2 are the sets of CWEs in two different clusters, and CS is the cosine similarity function quantifying similarity between pairs of CWEs from different clusters.

For internal cluster cohesion, I compute the **average inner similarity (AIS)** within each cluster:

$$\mathrm{AIS}(E,E) = \frac{2}{|E|(|E|-1)} \sum_{e_i,e_j\in E, e_i\neq e_j} \mathrm{CS}(e_i,e_j)$$

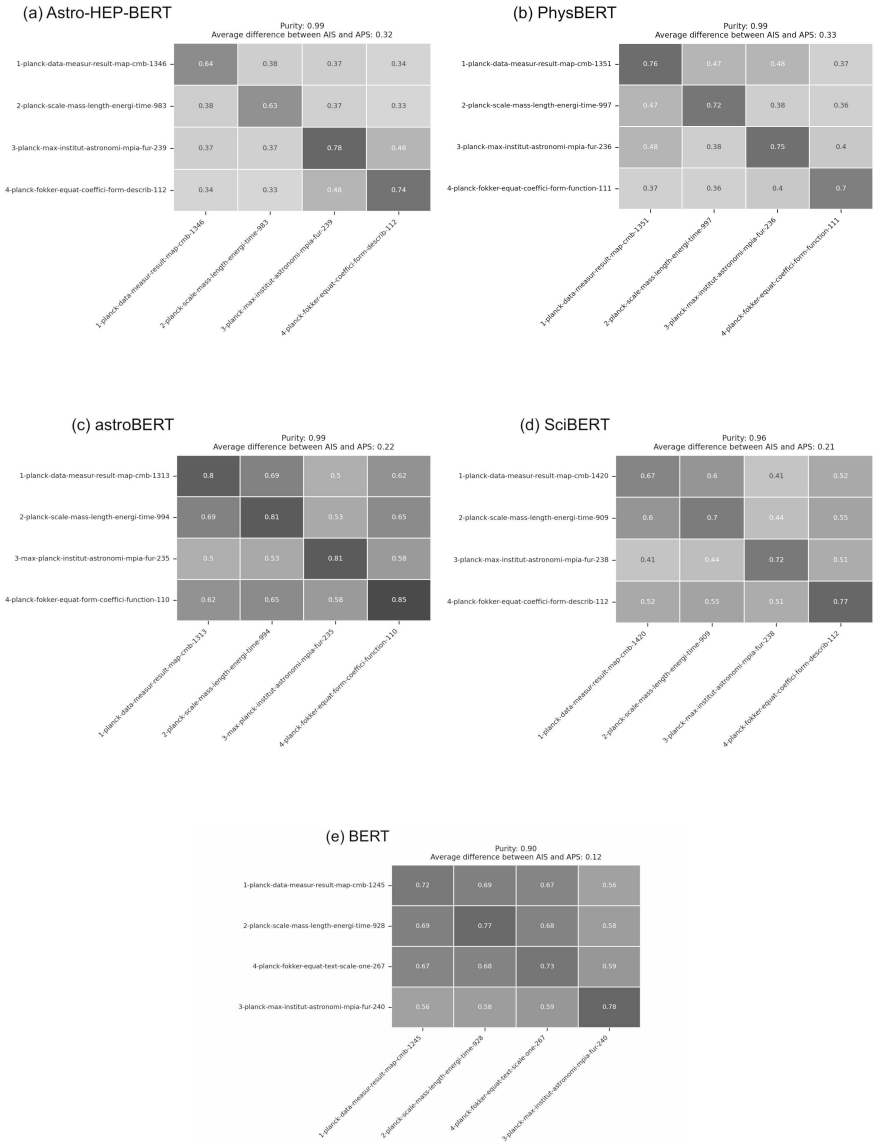
where E represents the set of CWEs within a single cluster.

These metrics provide insight into how well-separated the clusters are and how cohesive each cluster is internally. Low APS scores suggest clear distinctions between clusters, while high AIS scores indicate strong internal coherence within clusters.

Figure 4 presents similarity heatmaps for the best clustering solutions (based on inertia) of Astro-HEP-BERT, astroBERT, PhysBERT, and BERT in the four-label case of the Astro-HEP Corpus. This configuration is ideal for a focused analysis of separation and cohesion across models because it combines balanced complexity—with clear, inter-

pretable heatmaps that highlight trends across models—and similarly high purity scores of 0.99 for astroBERT, Astro-HEP-BERT, and PhysBERT, while SciBERT and BERT perform noticeably lower.

*Figure 4: Heatmaps showing cluster cohesion and separation for each model's **(a-e)** best four-label clustering solutions on the Astro-HEP-Planck Corpus. The diagonal cells display average inner similarity (AIS) scores for each cluster, reflecting internal cohesion, while the off-diagonal cells show average pairwise similarity (APS) between clusters, indicating separation. Cluster labels include the dominant label index, the six most frequent neighboring words within 10 tokens of "Planck", and the total number of embeddings in the cluster.*



Each subfigure shows the mutual APS scores between the four clusters, with the AIS scores on the diagonal. For each clustering solution, the purity score as well as the average difference between AIS and APS scores are displayed at the top of the heatmap. Cluster names start with the index of the dominant label (as in Figure 3), followed by the six most frequent words within 10 tokens on either side of the "Planck" occurrences (stop words excluded), and the number of embeddings in the cluster. The heatmaps are sorted by cluster size, with the largest clusters positioned at the top and left.

Astro-HEP-BERT, PhysBERT, and astroBERT all achieve peak and matching purity scores, followed by SciBERT and then BERT, which performs significantly weaker. A comparison of the heatmaps reveals notable variations in APS and AIS scores across models, reflecting how well they balance internal cluster cohesion and separation. To evaluate this balance, we can examine the difference between a model's AIS and APS scores.

In the four-label case, PhysBERT and Astro-HEP-BERT achieve the best balance, with average differences of 0.33 and 0.32, respectively. PhysBERT demonstrates stronger coherence within its first two clusters, dominated by the MISSION and UNITS labels, though these clusters are less well-separated. In contrast, Astro-HEP-BERT has weaker coherence in its first two clusters but better separation, suggesting that they are more semantically diverse but also more distinct from each other. astroBERT achieves the highest AIS values overall (0.8 to 0.85), reflecting strong internal cohesion, but also has higher APS values, indicating less separation between clusters. For instance, clusters 1, 2, and 4 (recall that these indices reflect the actual frequency distribution of labels in the corpus rather than in the clustering solution) are much less distinct from one another in astroBERT compared to the other models. SciBERT shows very similar patterns, but performs worse than astroBERT. The weakest performer is clearly BERT, with smaller differences between AIS and APS, averaging 0.12. This indicates relatively poor separation between clusters, aligning with its lower purity scores and overall weaker performance in disambiguation.

As further detailed in the Supplement, the improved balance between AIS and APS observed in PhysBERT and Astro-HEP-BERT, compared to astroBERT and the other two models, may only partially result from differences in isotropy—the degree to which embeddings are evenly distributed within the vector space. While Astro-HEP-BERT demonstrates the highest global isotropy among all models, astroBERT lies at the opposite end, exhibiting the most anisotropic embedding space. Nevertheless, astroBERT outperforms both SciBERT and BERT on our WSD and WSI tasks, despite their relatively higher isotropy.

Having established the models' ability to distinguish the different senses of "Planck" effectively, we now turn to their application in a diachronic setting, tracking how the distribution of these senses evolves over time. This analysis bridges sense disambiguation with the broader task of understanding conceptual shifts in scientific language.

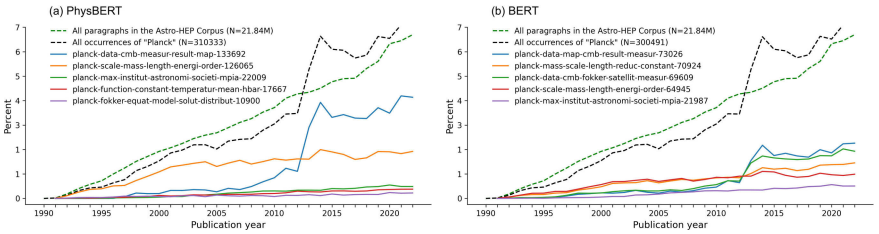
8. Lexical semantic change

Building on the models' tested performance in WSD and WSI from the previous sections, we can now explore their ability to measure semantic change for the term "Planck"

in a real-world, large-corpus setting. By tracking the evolving distributions of CWEs across sense clusters, we examine changes over three decades in the unlabeled Astro-HEP Corpus, which includes over 300,000 instances of "Planck". A K-Means five-cluster solution, selected from 100 random seed initializations based on inertia, was applied to each model.

Figures 5a and 5b illustrate the relative prominence of these clusters (colored lines) over time for PhysBERT and BERT, respectively. The figures also show the overall relative frequency of "Planck" occurrences (dashed black line) and corpus growth (dashed green line). As before, cluster names were derived from the six most frequent words found in a 10-token window around "Planck" (excluding stop words), accompanied by the number of embeddings per cluster. The focus on PhysBERT (a domain-specific model) and BERT (a general-purpose model) reflects their contrasting design and the relative quality of their clustering solutions.

*Figure 5: Evolution of the relative frequency of "Planck" occurrences across five clusters (colored lines) and over time, as modeled using PhysBERT (**a**) and BERT (**b**). The x-axis shows the years from 1990 to 2022, and the y-axis represents the normalized frequencies of occurrences per year. The dashed black line indicates the overall relative frequency of "Planck", while the dashed green line represents overall corpus growth. Each cluster label includes the six most frequent neighboring words and the total number of embeddings assigned to the cluster.*



PhysBERT (5a) produced the most interpretable five-cluster breakdown, with clusters corresponding to prominent senses of "Planck": the Planck mission (blue), Planck units (orange), the Max Planck Society (green), Planck's law and constant (red), and the Fokker-Planck equation (purple). Notably, PhysBERT avoids splitting the mission and units senses into multiple clusters, yielding clearer and more distinct sense groupings. From 1996 to 1998, the (blue) mission cluster shows an initial rise, surpassing all clusters except the larger (orange) units cluster. Between 2007 and 2011, the mission cluster grows rapidly, overtaking the units cluster from 2012 to 2013. This growth peaks in 2014, after which the mission cluster stabilizes at a frequency roughly double that of the units cluster, maintaining this level through 2022.

In contrast, BERT (Figure 5b) disperses the mission and units senses across multiple clusters (blue/green for mission and orange/red for units), obscuring their relative growth. The model also blends the mission and Fokker-Planck equation senses in the green cluster, resulting in the least interpretable five-cluster breakdown. Still, the clus-

ters referencing the mission sense (blue and green) surpass all others between 2012 and 2013.

These results align with milestones in the history of the Planck mission.[7] Originally proposed as the COBRAS/SAMBA mission in the early 1990s, it was approved by the European Space Agency (ESA) in 1996 and renamed the Planck mission in honor of Max Planck. Around this time, references to the mission began appearing in scientific literature.

Later, from 2011 to 2018, ESA released key data from the mission, including the 2013 first all-sky map of the cosmic microwave background (CMB), which may have contributed to the prominence of the mission-related sense during this period.

To quantify changes in the sense distribution of "Planck" and potentially the term's overall meaning, I applied two complementary indicators described in Periti and Montanelli (2024). The first, **Jensen-Shannon divergence (JSD)**, is a "sense-based" metric that captures shifts in the relative prominence of distinct senses over time by comparing their normalized frequency distributions. It is particularly sensitive to changes in the balance of senses, even if the dominant meaning remains stable. JSD is calculated as:

$$\mathrm{JSD}(d_1,d_2) = \frac{1}{2}(\mathrm{KL}(d_1,\mathrm{M}) + \mathrm{KL}(d_2,\mathrm{M})),$$

where d_1 and d_2 are the frequency distributions of clusters in two consecutive years, KL is the Kullback-Leibler divergence, and M is the average of the two distributions.

In contrast, the second indicator, **cosine distance between prototypes (CDPT)**, is a "form-based" metric that tracks changes in the overall meaning of a term by comparing the averaged word embeddings (prototypes) from consecutive years. CDPT is less affected by fluctuations in sense distributions and instead highlights broader shifts in dominant meaning. It is defined as:

$$\mathrm{CDPT}(\mathrm{PT}_1,\mathrm{PT}_2) = 1 - \mathrm{CS}(\mathrm{PT}_1,\mathrm{PT}_2),$$

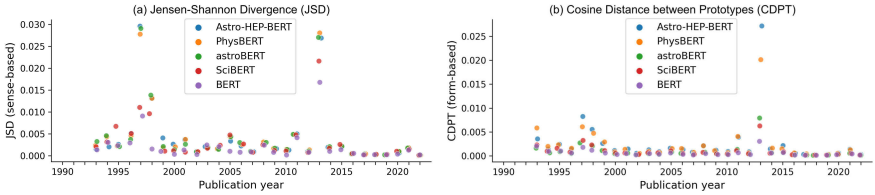
where CS is the cosine similarity and PT_1 and PT_2 are the year-specific prototypes.

Figures 6a and 6b illustrate the results of JSD and CDPT, revealing significant semantic shifts in 1997 and 2013 that align with historical milestones of the Planck mission. In 1997, the renaming of the COBRAS/SAMBA mission to "Planck" coincides with a sharp spike in JSD, particularly in domain-specific models like Astro-HEP-BERT and PhysBERT, capturing the rise of the mission sense even as the dominant meaning remained tied to units. CDPT showed a smaller spike, suggesting limited impact on the overall dominant meaning. Similarly, in 2013, the release of key data from the Planck mission aligns with sharp increases in both indicators, reflecting a more cohesive shift in the dominant meaning of "Planck".

The complementary nature of JSD and CDPT is evident here: JSD is more sensitive to shifts in the balance of senses, such as the emergence of the mission sense in 1997, while CDPT highlights broader contextual changes, as seen in the major semantic shift in 2013. Post-2013, both indicators stabilize, but JSD captures ongoing smaller fluctuations in sense distributions, whereas CDPT reflects the settling of the dominant meaning.

7 https://www.esa.int/Science_Exploration/Space_Science/Planck_overview and https://www.cosmos.esa.int/web/planck/mission-history; https://en.wikipedia.org/wiki/Planck_(spacecraft)

Figure 6: Semantic change of the term "Planck" over time, measured using two indicators: Jensen-Shannon divergence (JSD) and cosine distance between prototypes (CDPT). JSD (a) captures shifts in the distributions of clusters, representing different senses, while CDPT (b) measures global shifts in the aggregated meaning of the term. Both plots track changes across the years from 1990 to 2022 for each model.



These findings demonstrate the value of combining sense-based and form-based LSC indicators to capture both fine-grained sense dynamics and broader shifts in dominant meaning, providing a comprehensive view of lexical semantic change for scientific concepts over time. Among the models, domain-specific ones like Astro-HEP-BERT and PhysBERT display greater sensitivity to both JSD and CDPT changes, reflecting their superior performance in WSD and WSI tasks.

9. Conclusion and discussion

This study demonstrates the value of contextualized word embeddings (CWEs) for the history, philosophy, and sociology of science (HPSS), providing a new tool for analyzing the meanings of scientific concepts. Using the term "Planck" as a test case, I evaluated five BERT-based models across tasks in word sense disambiguation, sense induction, clustering quality, and lexical semantic change. The findings show that CWEs effectively disambiguate scientific terms, uncover distinct meanings, and trace diachronic trends, revealing how concepts like "Planck" evolve alongside developments in science, technology, and society.

The study highlights the promise of domain-adapted CWEs for HPSS research, particularly when models are trained on specialized corpora tailored to specific scientific fields or historical periods. This approach maximizes contextual relevance, and while critics may fear that such specialization risks overfitting, we can instead view it as a form of context-sensitive tailoring that aligns a model's interpretive focus with the unique conceptual landscape of its target domain. By allowing models to learn the distinct characteristics of specific scientific discourses, this approach enhances performance in these contexts, even at the expense of broader generalizability. In this regard, the success of Astro-HEP-BERT serves as a proof-of-concept for developing specialized models for HPSS on a limited budget.

By reusing pretrained general-purpose weights, Astro-HEP-BERT achieved performance comparable to, or better than, more resource-intensive domain-adapted models like PhysBERT and astroBERT, despite using fewer training examples.

Beyond tracing "conceptual histories" (Wevers and Koolen, 2020; Zichert et al., 2024; Zichert and Simons, 2026), domain-adapted CWEs and their basic operations—such

as measuring similarity, clustering, and generating prototypes—may offer broader applications for exploring other HPSS phenomena related to semantic similarity and divergence. These could include investigating lexical codification within research fields (Amahdi, 2026; Gläser et al., 2018; Teich et al., 2021) or addressing linguistic aspects of translating scientific concepts across paradigms (Kuhn, 2022) and within "trading zones" (Galison, 1997). For instance, semantic distance measures could reveal patterns of conceptual consistency, while clustering CWEs could uncover emergent terminologies or hybrid concepts bridging disciplines. Prototypical embeddings for terms from different paradigms could further illuminate semantic alignment or divergence, offering new and complementary methods for studying the evolution of scientific language. These possibilities remain speculative and sketchy, outlining potential approaches that must be explored further and specified through targeted studies and practical applications.

CWEs trained on scientific corpora can also enhance downstream tasks and advanced computational workflows, offering new opportunities for specialized applications in HPSS. For instance, they can improve domain-specific named entity recognition (NER) or the generation of specialized sentence embeddings, as demonstrated by recent studies (Hellert et al., 2024; Grezes et al., 2022). Beyond these tasks, domain-adapted CWEs could play a pivotal role in advanced computational workflows. A compelling example is retrieval-augmented generation (RAG), where embedding domain-specific queries and documents into a shared semantic space enhances retrieval accuracy and contextual relevance. This integration could enable more precise information retrieval, facilitate robust knowledge synthesis, and support dynamic interactions with complex scientific corpora.

For future research, scholars should consider developing HPSS-specific evaluation datasets tailored to a variety of computational tasks. Foundational tasks like WSD, WSI, LSC, and Word-in-Context (WiC) (Pilehvar and Camacho-Collados, 2019) remain essential. WiC, which tests whether a word maintains the same meaning across two different contexts, can also be extended to diachronic settings (Loureiro et al., 2022), providing insights into the evolution of scientific concepts over time. Beyond these, the scope of evaluation datasets could be broadened to include tasks such as domain-specific NER and the generation of sentence embeddings using models like SentenceBERT (Reimers and Gurevych, 2019). By creating tailored evaluation datasets for these and other tasks, researchers can address the unique challenges posed by specialized scientific languages, fostering more precise and context-aware computational analyses. My own two labeled datasets for the term "Planck," which provide sense annotations based on predefined categories and, in one case—the Astro-HEP-Planck Corpus—include the publication year, exemplify how HPSS-specific annotations can support the classical WSD task and prepare for both synchronic and diachronic analyses.

A key limitation is that BERT-based CWEs rely entirely on textual data, which restricts analyses to patterns and contexts within the analyzed corpus. Non-textual aspects of scientific practice—such as experimental settings, interpersonal communication, and institutional or cultural contexts—are often inaccessible or only indirectly reflected in texts. To address this, HPSS researchers can combine CWEs with qualitative methods and explore multimodal pipelines. Qualitative methods provide the interpretive depth needed to analyze complex senses and socio-historical factors, ensuring CWE-

based analyses remain grounded in broader historical and disciplinary contexts. Multimodal pipelines, by integrating non-textual sources like figures, diagrams, metadata, and citation networks, might help bridge the gap between textual representations and the material and social dimensions of scientific activity. However, the feasibility of such approaches remains to be fully explored (Yin et al., 2024).

Staying informed about advancements in adjacent fields, such as computational humanities, computational linguistics, and machine learning, is essential for researchers employing CWEs in HPSS (cf. Simons et al., 2026). By engaging with developments in model architectures, semantic tools, and advanced pipelines, HPSS scholars can enhance the precision and interpretability of CWE-based analyses. This interdisciplinary approach will not only deepen our understanding of scientific concepts and their evolution but also open new avenues for studying the dynamics of scientific language and its broader cultural implications. By embracing CWEs and their extensions, HPSS researchers have an opportunity to lead a transformative shift in the computational analysis of scientific discourse.[8]

Acknowledgements

I am grateful to my colleagues Adrian Wüthrich and Michael Zichert for their insightful feedback on this draft, and for our collaborative exchange of ideas regarding the use of computational methods in HPSS. I also want to acknowledge funding by the European Union (ERC Consolidator Grant, Project No. 101044932). Views and opinions expressed are however those of the author only and do not necessarily reflect those of the European Union or the European Research Council. Neither the European Union nor the granting authority can be held responsible for them.

References

Ahmadi ES (2026) Exploring disciplinary differences in semantic uniformity. A computational approach to codification. In: Simons A, Wüthrich A, Zichert M, et al. (eds) *Understanding Science with Large Language Models? Potentials for the History, Philosophy, and Sociology of Science*. Bielefeld: transcript, part-3.

Ait-Saada M and Nadif M (2023) Is Anisotropy Truly Harmful? A Case Study on Text Clustering. In: *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)* (eds A Rogers, J Boyd-Graber, and N Okazaki), Toronto, Canada, July 2023, pp. 1194–1203. Association for Computational Linguistics. Available at: https://aclanthology.org/2023.acl-short.103

8 This chapter was written with support from large language models (LLMs). All model-generated text was reviewed and, where necessary, rewritten by the authors, who remain fully responsible for the final version. For details on the use of LLMs in this volume, see the statement in the volume's introduction.

Beltagy I, Lo K and Cohan A (2019) SciBERT: A Pretrained Language Model for Scientific Text. http://arxiv.org/abs/1903.10676.

Bevilacqua M, Pasini T, Raganato A, et al. (2021) Recent Trends in Word Sense Disambiguation: A Survey. In: *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence*, Montreal, Canada, August 2021, pp. 4330–4338. International Joint Conferences on Artificial Intelligence Organization. Available at: https://www.ijcai.org/proceedings/2021/593

Biś D, Podkorytov M and Liu X (2021) Too Much in Common: Shifting of Embeddings in Transformer Language Models and its Implications. In: *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (eds K Toutanova, A Rumshisky, L Zettlemoyer, et al.), Online, June 2021, pp. 5117–5130. Association for Computational Linguistics. Available at: https://aclanthology.org/2021.naacl-main.403.

Blei DM, Ng AY and Jordan MI (2003) Latent Dirichlet allocation. *Journal of machine Learning research* 3(Jan): 993–1022.

Bowker GC and Star SL (1999) *Sorting Things out: Classification and Its Consequences*. The MIT Press.

Boyack KW, Newman D, Duhon RJ, et al. (2011) Clustering more than two million biomedical publications: Comparing the accuracies of nine text-based similarity approaches. *PloS one* 6(3): e18029.

Cai X, Huang J, Bian Y, et al. (2020) Isotropy in the contextual embedding space: Clusters and manifolds. In: *International conference on learning representations*, 2020. Available at: https://openreview.net/forum?id=xYGNO86OWDH.

Callon M, Courtial J-P, Turner WA, et al. (1983) From translations to problematic networks: An introduction to co-word analysis. *Social Science Information* 22(2): 191–235.

Canguilhem G (1991) *The Normal and the Pathological* (trans. CR Fawcett). Revised edition. New York: MIT PR.

Chang H (2007) *Inventing Temperature: Measurement and Scientific Progress*. Oxford: Oxford University Press.

Clarke AE, Mamo L, Fosket JR, et al. (eds) (2010) *Biomedicalization: Technoscience, Health, and Illness in the U.S.* Duke University Press.

Courtial J-P and Law J (1989) A Co-Word Study of Artificial Intelligence. *Social Studies of Science* 19(2): 301–311.

Daston L and Galison P (2007) *Objectivity*. New York, NY: Zone Books.

Deerwester S, Dumals ST, Furnas GW, et al. (1990) Indexing by Latent Semantic Analysis. *Journal of the American Society for Information Science* 41(6): 391–407.

Devlin J, Chang M-W, Lee K, et al. (2018) BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. http://arxiv.org/abs/1810.04805.

Ding Y, Martinkus K, Pascual D, et al. (2022) On Isotropy Calibration of Transformer Models. In: *Proceedings of the Third Workshop on Insights from Negative Results in NLP* (eds S Tafreshi, J Sedoc, A Rogers, et al.), Dublin, Ireland, May 2022, pp. 1–9. Association for Computational Linguistics. Available at: https://aclanthology.org/2022.insights-1.1.

Ethayarajh K (2019) How Contextual are Contextualized Word Representations? Comparing the Geometry of BERT, ELMo, and GPT-2 Embeddings. http://arxiv.org/abs/1909.00512.

Fleck L (1979) *Genesis and Development of a Scientific Fact*. University of Chicago Press.

Foucault M (1970) *The Order of Things: An Archaeology of Human Sciences*. Random House.

Galison P (1997) *Image and Logic: A Material Culture of Microphysics*. illustrated edition. Univ of Chicago Pr.

Gläser J, Laudel G, Grieser C, et al. (2018) Scientific fields as epistemic regimes: new opportunities for comparative science studies. DEU. Epub ahead of print 2018.

Godey N, de la Clergerie É and Sagot B (2023) Is Anisotropy Inherent to Transformers? h ttp://arxiv.org/abs/2306.07656.

Grezes F, Blanco-Cuaresma S, Accomazzi A, et al. (2021) Building astroBERT, a language model for Astronomy & Astrophysics. http://arxiv.org/abs/2112.00590.

Grezes F, Allen T, Blanco-Cuaresma S, et al. (2022) Improving astroBERT using Semantic Textual Similarity. http://arxiv.org/abs/2212.00744.

Gu Y, Tinn R, Cheng H, et al. (2021) Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare (HEALTH)* 3(1). ACM New York, NY: 1–23.

Hacking I (1975) *The Emergence of Probability: A Philosophical Study of Early Ideas about Probability, Induction and Statistical Inference*. Cambridge University Press.

Hacking I (1999) *The Social Construction of What?* Harvard University Press.

Hellert T, Montenegro J and Pollastro A (2024) PhysBERT: A Text Embedding Model for Physics Scientific Literature. http://arxiv.org/abs/2408.09574.

Kleymann R, Niekler A and Burghardt M (2022) Conceptual Forays: A Corpus-based Study of "Theory" in Digital Humanities Journals. *Journal of Cultural Analytics* 7(4).

Kuhn TS (1962) *The Structure of Scientific Revolutions*. Chicago: University of Chicago Press.

Kuhn TS (2022) *The Last Writings of Thomas S. Kuhn: Incommensurability in Science* (ed. B Mladenovic). First Edition. Chicago: University of Chicago Press.

Latour B (1987) *Science in Action: How to Follow Scientists and Engineers through Society*. Harvard University Press.

Laubichler MD, Maienschein J and Renn J (2019) Computational History of Knowledge: Challenges and Opportunities. *Isis* 110(3): 502–512.

Lean OM, Rivelli L and Pence CH (2023) Digital Literature Analysis for Empirical Philosophy of Science. *The British Journal for the Philosophy of Science* 74(4): 875–898.

Lee J, Yoon W, Kim Sungdong, et al. (2020) BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics* 36(4): 1234–1240.

Leydesdorff L and Rafols I (2009) A global map of science based on the ISI subject categories. *Journal of the American Society for Information Science and Technology* 60(2): 348–362.

Liu Y, Ott M, Goyal N, et al. (2019) RoBERTa: A Robustly Optimized BERT Pretraining Approach. http://arxiv.org/abs/1907.11692.

Loureiro D, Rezaee K, Pilehvar MT, et al. (2020) Language Models and Word Sense Disambiguation: An Overview and Analysis. Available at: https://www.semanticscholar .org/reader/a7cb7c67a181c35f2e2dd2767e21defc0af9c14a.

Loureiro D, D'Souza A, Muhajab AN, et al. (2022) TempoWiC: An Evaluation Benchmark for Detecting Meaning Shift in Social Media. http://arxiv.org/abs/2209.07216.

Malaterre C and Léonard M (2024) Epistemic Markers in the Scientific Discourse. *Philosophy of Science* 91(1): 151–174.

Merchant C (1980) *The Death of Nature: Women, Ecology, and the Scientific Revolution*. New York: Harper & Row.

Mickus T, Grönroos S-A and Attieh J (2024) Isotropy, Clusters, and Classifiers. http://arxiv.org/abs/2402.03191.

Mikolov T, Chen K, Corrado G, et al. (2013) Efficient Estimation of Word Representations in Vector Space. http://arxiv.org/abs/1301.3781.

Mol A (2002) *The Body Multiple: Ontology in Medical Practice*. Duke University Press.

Mu J, Bhat S and Viswanath P (2018) All-but-the-Top: Simple and Effective Postprocessing for Word Representations. http://arxiv.org/abs/1702.01417.

Overton JA (2013) "Explain" in scientific discourse. *Synthese* 190(8): 1383–1405.

Pence CH and Ramsey G (2018) How to do digital philosophy of science. *Philosophy of Science* 85(5). Cambridge University Press: 930–941.

Periti F and Montanelli S (2024) Lexical Semantic Change through Large Language Models: a Survey. *ACM Computing Surveys* 56(11): 282:1-282:38.

Pickering A (1995) *The Mangle of Practice: Time, Agency, and Science*. Chicago: University of Chicago Press.

Pilehvar MT and Camacho-Collados J (2019) WiC: the Word-in-Context Dataset for Evaluating Context-Sensitive Meaning Representations. In: *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)* (eds J Burstein, C Doran, and T Solorio), Minneapolis, Minnesota, 2019, pp. 1267–1273. Association for Computational Linguistics. Available at: https://aclanthology.org/N19-1128.

Reimers N and Gurevych I (2019) Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. http://arxiv.org/abs/1908.10084.

Rip A and Courtial JP (1984) Co-word maps of biotechnology: An example of cognitive scientometrics. *Scientometrics* 6(6): 381–400.

Simons A (2024a) Astro-HEP-BERT: A bidirectional language model for studying the meanings of concepts in astrophysics and high energy physics. http://arxiv.org/abs/2411.14877.

Simons A (2024b) Meaning at the Planck scale? Contextualized word embeddings for doing history, philosophy, and sociology of science http://arxiv.org/abs/2411.14073.

Simons A, Zichert M and Wüthrich A (2026) Large language models for history, philosophy, and sociology of science: Interpretive uses, methodological challenges, and critical perspectives. *Studies in History and Philosophy of Science* 117: 102151. https://doi.org/10.1016/j.shpsa.2026.102151.

Steinle F (2016) *Exploratory Experiments: Ampère, Faraday, and the Origins of Electrodynamics*. University of Pittsburgh Press.

Sun Y and Platoš J (2023) A method for constructing word sense embeddings based on word sense induction. *Scientific Reports* 13(1): 12945.

Tahmasebi N and Dubossarsky H (2023) Computational modeling of semantic change. https://arxiv.org/abs/2304.06337.

Teich E, Fankhauser P, Degaetano-Ortlieb S, et al. (2021) Less is More/More Diverse: On The Communicative Utility of Linguistic Conventionalization. *Frontiers in Communication* 5.

Venturini T, Baya Laffite N, Cointet J-P, et al. (2014) Three maps and three misunderstandings: A digital mapping of climate diplomacy. *Big Data & Society* 1(2): 2053951714543804.

Wang L, Huang J, Huang K, et al. (2020) Improving Neural Language Generation with Spectrum Control. In: *International Conference on Learning Representations*, 2020. Available at: https://openreview.net/forum?id=ByxY8CNtvr.

Wevers M and Koolen M (2020) Digital begriffsgeschichte: Tracing semantic change using word embeddings. *Historical Methods: A Journal of Quantitative and Interdisciplinary History* 53(4): 226–243.

Wiedemann G, Remus S, Chawla A, et al. (2019) Does BERT Make Any Sense? Interpretable Word Sense Disambiguation with Contextualized Embeddings. http://arxiv.org/abs/1909.10430.

Zichert M and Simons A (2026) From early digital methods to LLMs. Computational conceptual history of scientific concepts. In: Simons A, Wüthrich A, Zichert M, et al. (eds) *Understanding Science with Large Language Models? Potentials for the History, Philosophy, and Sociology of Science*. Bielefeld: transcript, part-3.

Zichert M, Simons A and Wüthrich A (2025) Expanding Conceptual Histories: Using Contextualized Word Embeddings for the History and Philosophy of the Virtual Particle Concept. *Computational Humanities Research*: 1–22. Available at: https://doi.org/10.1017/chr.2025.10013

Supplement: isotropy analysis

In Section 7 I hypothesized that the improved balance between AIS and APS observed in PhysBERT and Astro-HEP-BERT, compared to astroBERT and the other two models may be partly due to a more isotropic em- bedding space, where embeddings are more evenly distributed, allowing for better cluster separation. To investigate this further, this section begins with a brief overview of the ongoing debate in the literature regarding the roles of anisotropy and isotropy in BERT models and their impact on model performance.

The discussion around anisotropy in BERT models originated with the surprising discovery by Ethayarajh (2019) that CWEs produced by BERT exhibit significant anisotropy—an uneven distribution of CWEs in the high-dimensional space. This finding contrasted with earlier research by Mu et al. (2018), which demonstrated that anisotropy in static word embeddings could lead to performance degradation. However, since BERT's contextualized representations have shown remarkable performance across numerous tasks, researchers began questioning whether anisotropy had the same negative effect on contextualized embeddings as it did on static ones.

Subsequent studies produced mixed results regarding the impact of reducing anisotropy in BERT and other Transformer models. Some researchers reported performance improvements by mitigating anisotropy, using methods such as post-processing techniques (Biś et al., 2021; Wang et al., 2020) or modifications to pretraining objectives

(Godey et al., 2023). However, others found no significant improvements or, in some cases, even performance declines when attempting to enforce isotropy (Ait-Saada and Nadif, 2023; Ding et al., 2022). This has led to varying opinions on whether reducing anisotropy consistently enhances the performance of Transformer models across different tasks.

A related discussion has emerged around the distinction between global anisotropy and local isotropy. Several authors found that while Transformer models, including BERT, display global anisotropy (an uneven embedding distribution across the entire space), they show local isotropy within specific subspaces or clusters (Ait-Saada and Nadif, 2023; Cai et al., 2020; Ding et al., 2022; Mickus et al., 2024). This local isotropy may explain BERT's high performance, as it allows for expressive representations in particular regions of the embedding space, despite the global anisotropy. Thus, the relationship between (an)isotropy and model performance, particularly for contextualized embeddings, remains complex and multifaceted.

The aim of this supplementary section is not to make a direct contribution to this broader debate, but rather to assess how our five models compare in terms of their global isotropy, which I calculated using the average cosine similarity (ACS) between the contextualized word embeddings (CWEs) of randomly sampled words—a standard approach in this research area. For random vectors with independent, identically distributed components, expected cosine similarity approaches zero as dimensionality increases. Thus, ACS values closer to zero suggest a more isotropic (and thus more balanced) embedding space.

For each model and corpus, 200,000 tokens were sampled, paired, and cosine similarity was calculated for 100,000 random pairs, with the resulting ACS providing a measure of global isotropy.

Results indicate that Astro-HEP-BERT has the most isotropic embedding space, with an ACS just below 0.1 in both corpora, followed by BERT and PhysBERT, which show ACS values slightly above 0.2 in the Astro-HEP Corpus. In contrast, PhysBERT shows a slightly higher ACS in the Wikipedia-Physics Corpus, nearing 0.3. SciBERT and astroBERT demonstrate higher anisotropy, with astroBERT reaching an ACS near 0.5 in the Wikipedia-Physics Corpus.

Certain models, like PhysBERT and SciBERT, display multiple peaks in their cosine similarity distributions, while Astro-HEP-BERT and BERT exhibit a more singular peak. This difference suggests that models with multiple peaks might have more compartmentalized embedding spaces, indicating anisotropic regions or clustering tendencies, whereas Astro-HEP-BERT's single, concentrated peak reflects a more isotropic, evenly distributed space.

These findings highlight two key observations: Astro-HEP-BERT's increased isotropy relative to models like PhysBERT and astroBERT may stem from its adaptation of BERT's pretrained embedding space rather than starting from scratch. Refining BERT's embeddings with astrophysics and HEP-specific data likely preserved and enhanced isotropy by organizing domain-specific terms within an existing, well-structured framework. In contrast, models trained from scratch on physics texts appear to have developed more clustered or anisotropic embedding spaces, potentially due to early overfitting on domain-specific patterns without a generalized framework.

Figures S1c and S1d show how isotropy evolved in Astro-HEP-BERT over three pretraining epochs. Notably, after the first epoch, ACS decreased by approximately 0.1, with smaller reductions in subsequent epochs, indicating progressive isotropy improvement as embeddings spread more uniformly.

Figure S1: Global isotropy analysis across models and corpora using average cosine similarity (ACS) and density distributions. (a) ACS and cosine similarity distribution for CWEs in the Astro-HEP Corpus, with each model's ACS represented by a colored vertical line. (b) ACS and cosine similarity distribution in the Wikipedia-Physics Corpus. (c, d) Evolution of ACS for Astro-HEP-BERT over three pretraining epochs, evaluated on the Astro-HEP (c) and Wikipedia-Physics (d) corpora, showing progressive improvement in isotropy. Lower ACS values indicate a more isotropic and balanced embedding space.