PROCEEDINGS OF SCIENCE

Revealing Connections in QCD with Machine Learning

Patrick L.S. Connor[a,*] **and Antonin Sulc**[b]

[a] *Center for Data and Computing in Natural Sciences, Universität Hamburg, Germany, Albert-Einstein-Ring 10, D-22761 Hamburg*

[b] *Helmholtz-Zentrum Berlin fuer Materialien und Energie, Albert-Einstein-Str. 15, 12489 Berlin, Germany*

E-mail: patrick.connor@desy.de, antonin.sulc@helmholtz-berlin.de

This work utilises text analysis techniques to uncover connections and trends in quantum chromodynamics (QCD) research over time. Through embedding-based analysis, we are able to draw conceptual connections between disparate works across QCD subfields. Examining topic clustering and trajectories over time provides insights into new phenomena gaining momentum and experimental approaches coming to prominence in the QCD research area. Furthermore, we construct citation graphs between influential papers to reveal impactful contributions and relationships, compare them with respect to their topic, and propose intertopical and citation-related recommendations.

42nd International Conference on High Energy Physics (ICHEP2024)
18-24 July 2024
Prague, Czech Republic

*Speaker

PoS(ICHEP2024)1054

1. Introduction

The literature covering QCD poses a challenge for those seeking to learn about it. While mainstream search engines offer the easiest way to find information, they do not provide a complete picture of the QCD literature landscape. The growing complexity and specialization of QCD literature make it increasingly difficult to understand QCD across subfields as a whole. However, the availability of metadata from databases like InspireHEP or arXiv has opened up new ways to interpret complex domains like QCD. We analyze massive corpora formed from the QCD literature of the past couple of decades and attempt to build meaningful visualizations using state-of-the-art text modeling tools.

2. Data & Method

We extract a list of peer-reviewed publications from the InspireHEP database via their API. All articles must have either `QCD` or `quantum chromodynamics` in their list of keywords. We exclude proceedings and theses, as well as documents written in languages other than English. The classification is based solely on the content of the abstracts. After an initial run, we identified a list of problematic keywords such as `review`, `d0note` to exclude some metadata that spoil the analysis.

We modeled our corpus using BERTopic [1], a topic modeling algorithm that uses embeddings to create coherent topics from documents. The algorithm can be summarized in the following steps: (1) Document embedding [2], which encodes texts as high-dimensional vectors where the angle between abstracts represents their semantic similarity. (2) UMAP [3] dimensionality reduction: the original dimensionality of the embedding is reduced (in our case to 5D). (3) Hierarchical clustering: HDBSCAN clusters the reduced 5D embeddings. (4) Topic creation: Extracting representative words for each cluster using c-TF-IDF.

A citation graph is obtained by matching the reference list in the metadata of most publications to their respective titles. To visualize each publication represented as a node, we used node2vec embedding [4], which performs biased random walks on a graph to sample node neighborhoods and then uses these walks as input for a skip-gram model to learn vector representations of nodes. Finally, UMAP [3] is used to reduce dimensionality to 2D for visualization purposes.

3. Results

The results are displayed in Fig. 1 (left). The keywords shown in the figure were identified by the model as being the most representative ones from the given category.

We find that the map provides a reasonable description of QCD. For instance, `1,jet,production` and `19,shower,generator` are found to be close to one another, as expected from the non-Abelian nature of QCD. Similarly, `18,gravity,dark`, `27,hole,black`, and `16,axion,dark` are also found close by. With no surprise, the keyword "color" appears several times (23,26,38) throughout the whole map. Instead, a few outliers appear, such as `7,star,matter` and `31,quantum,qubit`: these may have ended up in the list for various reasons, such as the use of a similar acronym or wrongly assigned keywords. The results may be compared with a type of citation map on Fig. 1 (right). The same points are taken and their colour is the same as on the first map. The structure of

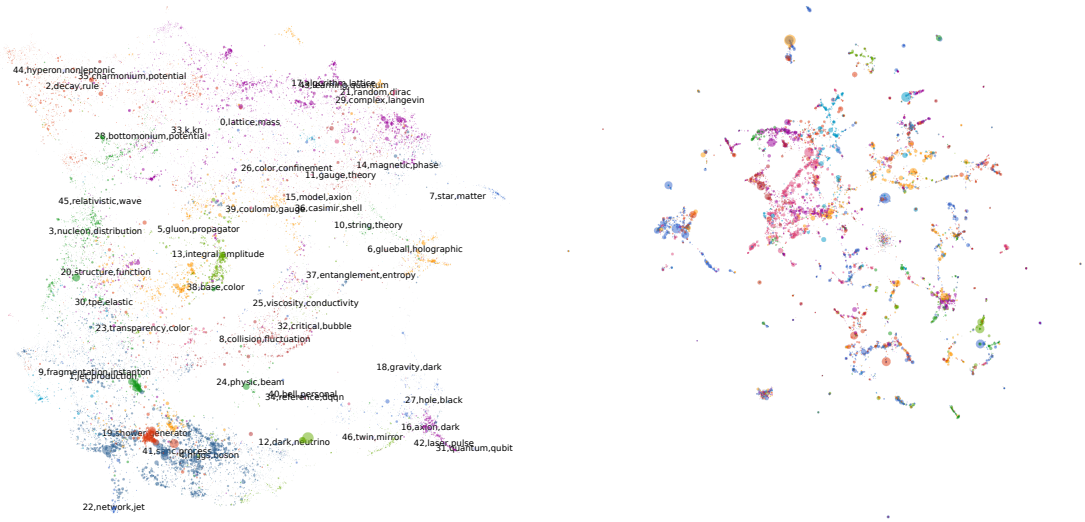

Figure 1: Left: map of publications with "'QCD" in abstract obtained after processing the QCD literature; right: Citation graph. Each point corresponds to a publication. The colours correspond to a topic. The size of a point corresponds to its number of citations. Each category is denoted by a colour, a number, and a couple of representative keywords.

the literature appears differently, highlighting the potential of language models to scan the literature and establish connections.

4. Conclusions & prospects

In summary, we find the language models have shown promise in terms of reviewing articles of vast fields of research. Their use should be further explored to summarise categories, perhaps even to suggest new ideas of research.

References

[1] M. Grootendorst, *Bertopic: Neural topic modeling with a class-based tf-idf procedure*, *arXiv preprint arXiv:2203.05794* (2022) .

[2] N. Reimers and I. Gurevych, *Sentence-bert: Sentence embeddings using siamese bert-networks*, in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*, Association for Computational Linguistics, 11, 2019, https://arxiv.org/abs/1908.10084.

[3] L. McInnes, J. Healy and J. Melville, *Umap: Uniform manifold approximation and projection for dimension reduction*, *arXiv preprint arXiv:1802.03426* (2018) .

[4] A. Grover and J. Leskovec, *node2vec: Scalable feature learning for networks*, in *Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 855–864, 2016.