

# Supporting Information

## Identifying translational science through embeddings of controlled vocabularies

Qing Ke

*Center for Complex Network Research,  
Northeastern University, Boston, MA 02115, USA*

(Dated: September 29, 2018)

| Branch Name  | Terms | Occurrences |
|--|-------|-------------|
| <b>A</b> Anatomy   | 1,683 | 27,782,594  |
| <b>B</b> Organisms   | 3,656 | 34,348,641  |
| <b>C</b> Diseases  | 4,596 | 49,158,246  |
| <b>D</b> Chemicals and Drugs   | 9,160 | 82,773,807  |
| <b>E</b> Analytical, Diagnostic and Therapeutic Techniques and Equipment | 2,702 | 58,616,922  |
| F Psychiatry and Psychology  | 951   | 8,851,299   |
| <b>G</b> Phenomena and Processes   | 1,949 | 45,631,068  |
| H Disciplines and Occupations  | 375   | 5,552,349   |
| I Anthropology, Education, Sociology and Social Phenomena                | 542   | 4,785,093   |
| J Technology, Industry, Agriculture                                      | 510   | 3,030,164   |
| K Humanities   | 189   | 1,210,479   |
| L Information Science  | 407   | 5,595,960   |
| <b>M</b> Named Groups  | 214   | 17,095,141  |
| <b>N</b> Health Care   | 1,565 | 39,266,571  |
| V Publication Characteristics  | 151   | –           |
| Z Geographicals  | 386   | 5,241,664   |

TABLE S1. Total occurrences of all terms in each branch of the MeSH tree, counted based on all terms in all MEDLINE papers. Branches marked in red are included in our analysis. Terms in branch V (“Publication Characteristics”) are not used because they are only for annotating publication type.

| Category               | Unique ID               | MeSH Term           | Tree Number                             |
|------------------------|-------------------------|---------------------|---|
| Cell and molecular (C) | <a href="#">D002477</a> | Cells               | A11                                     |
|                        | <a href="#">D001105</a> | Archaea             | B02                                     |
|                        | <a href="#">D001419</a> | Bacteria            | B03                                     |
|                        | <a href="#">D014780</a> | Viruses             | B04                                     |
|                        | <a href="#">D015394</a> | Molecular Structure | G02.111.570                             |
|                        | <a href="#">D055599</a> | Chemical Processes  | G02.149                                 |
| Animal (A)             | <a href="#">D056890</a> | Eukaryota           | B01                                     |
| Human (H)              | <a href="#">D006801</a> | Humans              | B01.050.150.900.649.801.400.112.400.400 |
|                        | <a href="#">D009272</a> | Persons             | M01                                     |

TABLE S2. Root nodes used for defining “basic” and “applied” terms. We consider basic terms as those located in the subtrees rooted at the nodes in (1) the cell and molecular (C) and (2) animal (A) category. Applied terms are those in the subtrees rooted at the nodes in the human (H) category.

| Unique ID | MeSH term                 | Level score |
|-----------|---------------------------|-------------|
| D005614   | Freeze Fracturing         | -0.640      |
| D003572   | Cytochalasins             | -0.637      |
| D003571   | Cytochalasin B            | -0.628      |
| D011161   | Porifera                  | -0.627      |
| D012430   | Ruthenium Red             | -0.622      |
| D011952   | Receptors, Concanavalin A | -0.610      |
| D011554   | Pseudopodia               | -0.603      |
| D002462   | Cell Membrane             | -0.603      |
| D004705   | Endocytosis               | -0.601      |
| D002450   | Cell Communication        | -0.600      |

TABLE S3. Top 10 most basic MeSH terms in 1980.

| Unique ID | MeSH term                                    | Level score |
|-----------|--|-------------|
| D010043   | Outcome and Process Assessment (Health Care) | 0.877       |
| D010358   | Patient Participation                        | 0.876       |
| D011369   | Professional-Patient Relations               | 0.875       |
| D012017   | Referral and Consultation                    | 0.873       |
| D010817   | Physician-Patient Relations                  | 0.873       |
| D001291   | Attitude of Health Personnel                 | 0.871       |
| D012949   | Social Work, Psychiatric                     | 0.871       |
| D012657   | Self-Help Groups                             | 0.868       |
| D010821   | Physicians, Family                           | 0.867       |
| D010343   | Patient Admission                            | 0.865       |

TABLE S4. Top 10 most applied MeSH terms in 1980.

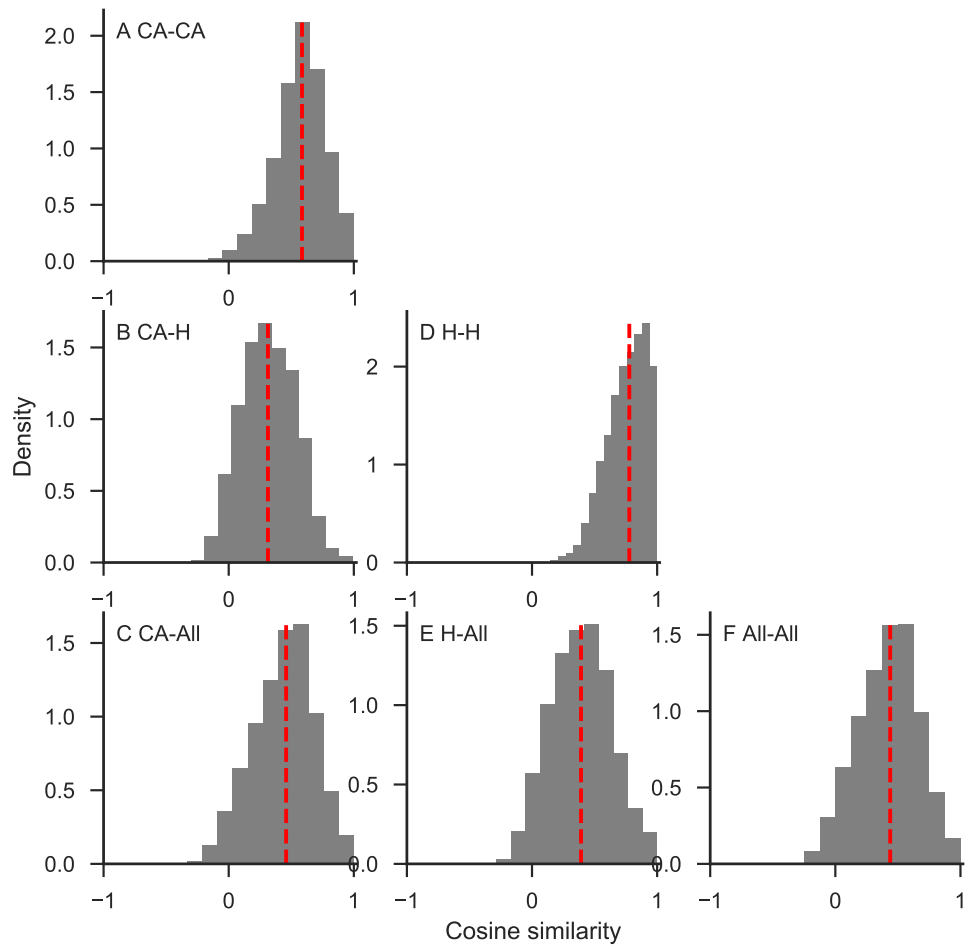


FIG. S1. All pairwise cosine similarity between MeSH terms at year 1980.

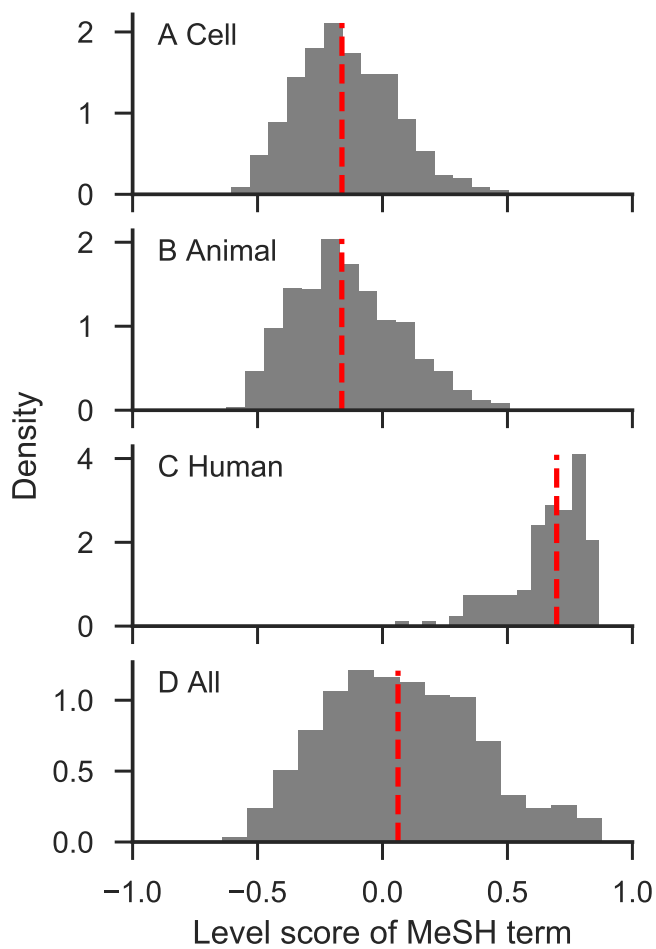


FIG. S2. Histogram of level score of MeSH terms at year 1980. Red dashed lines mark median values.

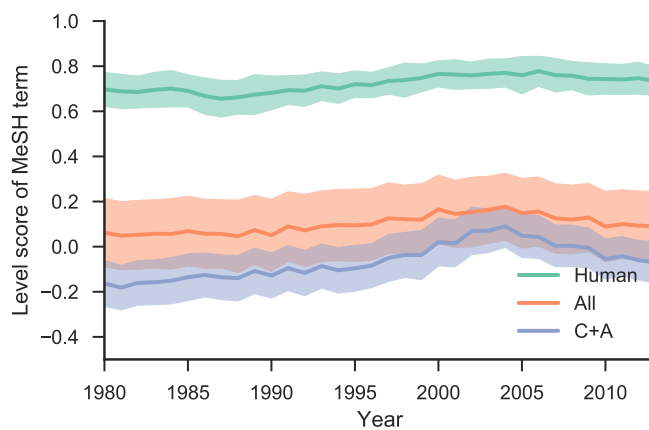


FIG. S3. Level score of MeSH terms over years. The lines show the median values, and the shaded regions cover one standard deviation.

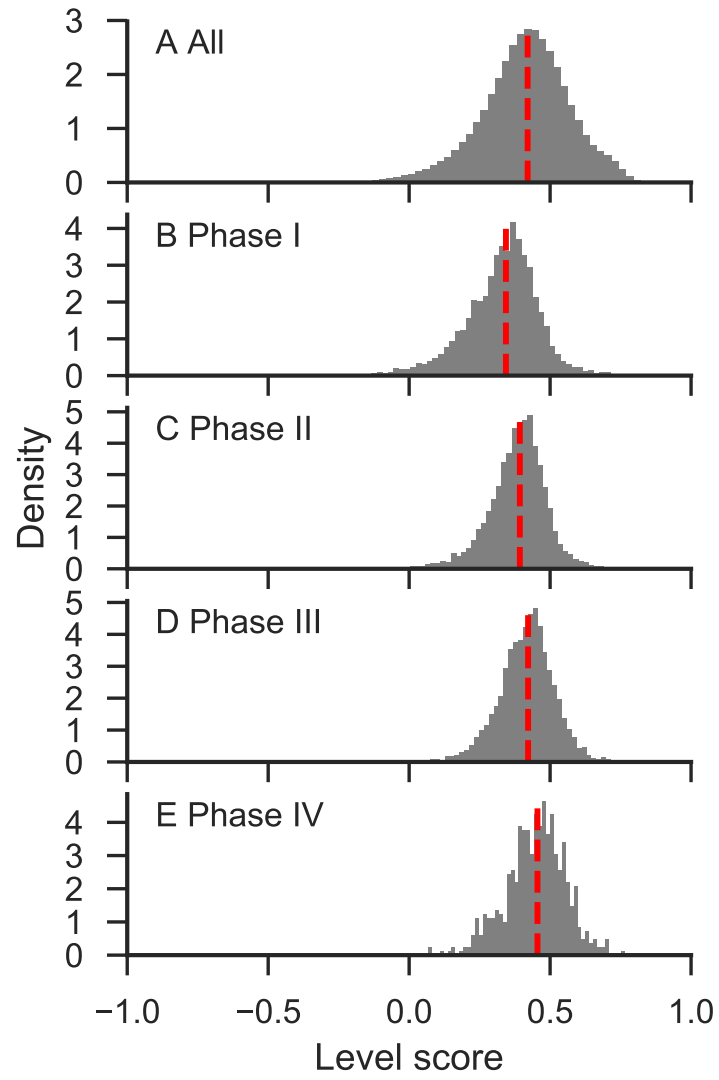


FIG. S4. Histogram of level score of clinical trial papers. Dash lines mark median values.

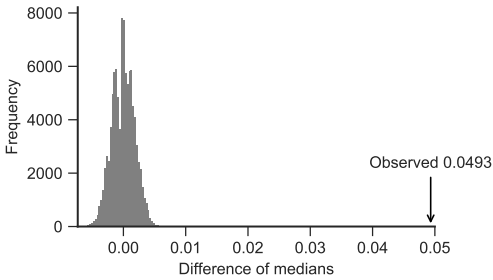
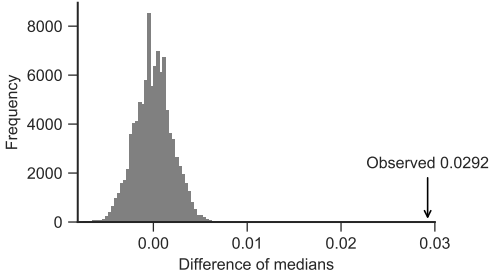
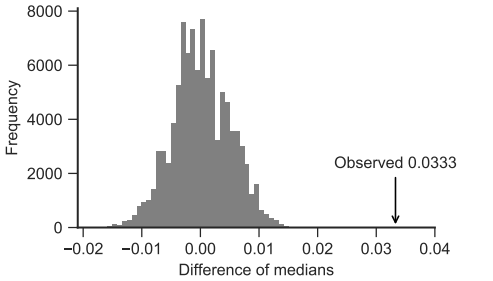
| Sample 1 median  | Sample 2 median  | Difference | Sig.  |
|------------------|------------------|------------|---|
| Phase II 0.3926  | Phase I 0.3433   | 0.0493     |   |
| Phase III 0.4219 | Phase II 0.3926  | 0.0292     |   |
| Phase IV 0.4552  | Phase III 0.4219 | 0.0333     |  |

TABLE S5. Statistical significance test of the difference between median level score of papers belonging to consecutive stages of clinical trials. We use permutation test, where  $10^5$  permutes are performed to obtain the null distribution showed in the last column. The  $p$ -values for all the three tests are  $< 10^{-5}$ .

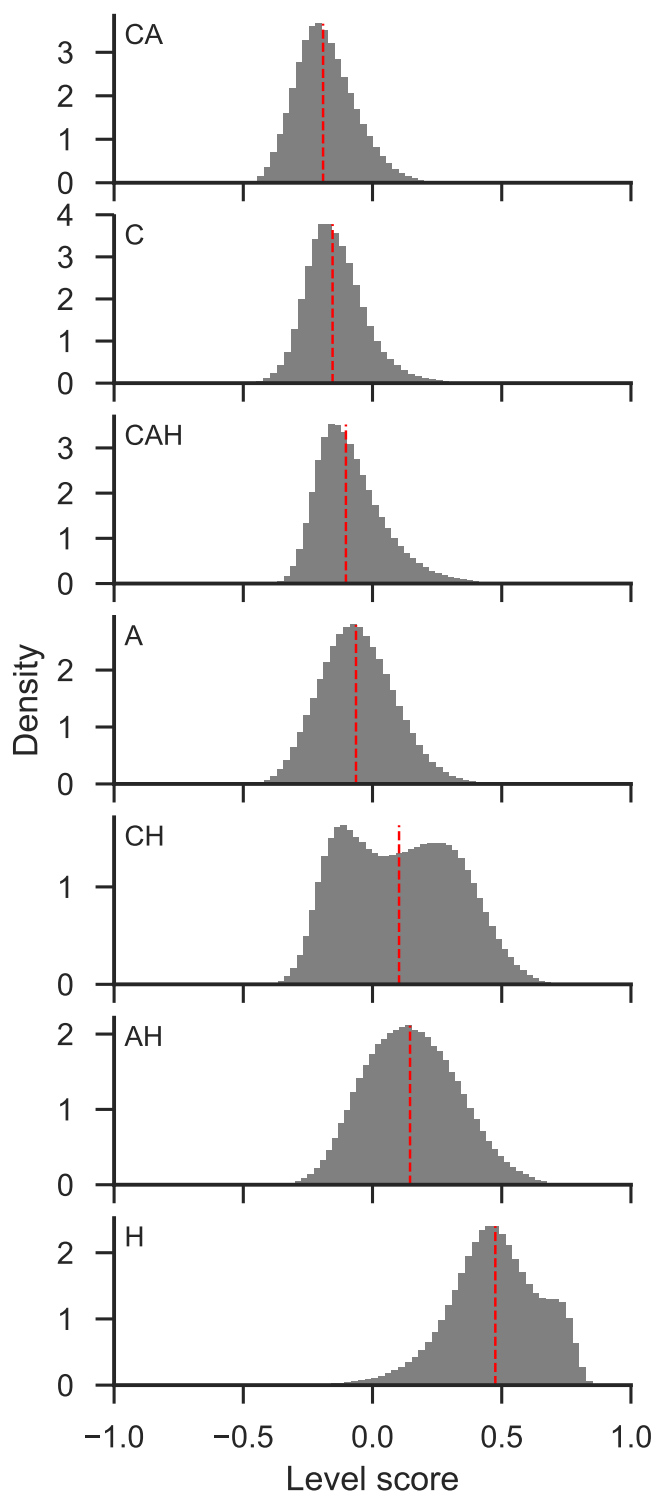


FIG. S5. Histogram of level score of papers in each category. The categorization is based on whether their MeSH terms contain the ones related to cell and molecular (C), animal (A), and human (H). Dash lines mark median values.



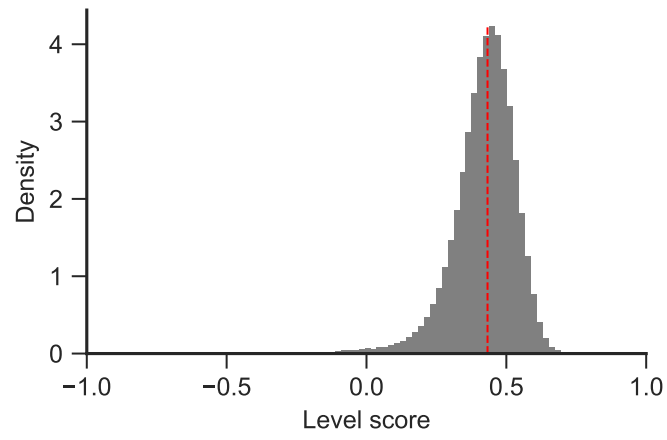


FIG. S6. Histogram of level score of papers that contain both “Humans” and “Magnetic Resonance Imaging” terms.

## S1. NULL MODEL

We observed from Fig. 4 in the main text that there is a “homophilous” pairing of direct citations—papers with similar level scores tend to cite each other. Can this observation be accounted for by the bimodal distribution of LS of papers? After all, a list of cited papers that are randomly selected would include some that have a score similar to the citing paper, especially if the LS of the citing paper is near the two peaks of the distribution. To address this question, we consider a null model where we randomize the citation network. In particular, we preserve the publication year and MeSH terms associated with each paper, therefore preserving their level scores, since the co-occurrence matrices  $M_t$  are preserved. However, we reshuffle citation pairs by randomizing the underlying citation network, preserving the yearly number of citations received by each paper through a series of link switches. Specifically, for each switch, we first randomly choose a pair of links where the two citing papers and the two cited papers were respectively published in the same year, and then switch the end-points (cited papers) of the two links. We perform  $50 \cdot E$  times of link switches to allow for good mixing, where  $E = 200,359,263$  is the number of links in the citation network, resulting in more than 10 billion switches. Based on this null model, the newly selected cited paper can be any other one that was published in the same year as the original cited paper, and thus can be of any level score. Through this way, we destroy the pairing of level scores of citing-cited paper pairs.

Fig. S6A shows that four regions of high density emerge after randomization and LS of cited papers concentrates in the two regions corresponding to the two peaks. This is readily implied from the null model: For a given citation pair, the randomly picked cited paper can be any other one published in the same year as the original paper, therefore it is more likely to be from the two peaks, since there are more papers there.

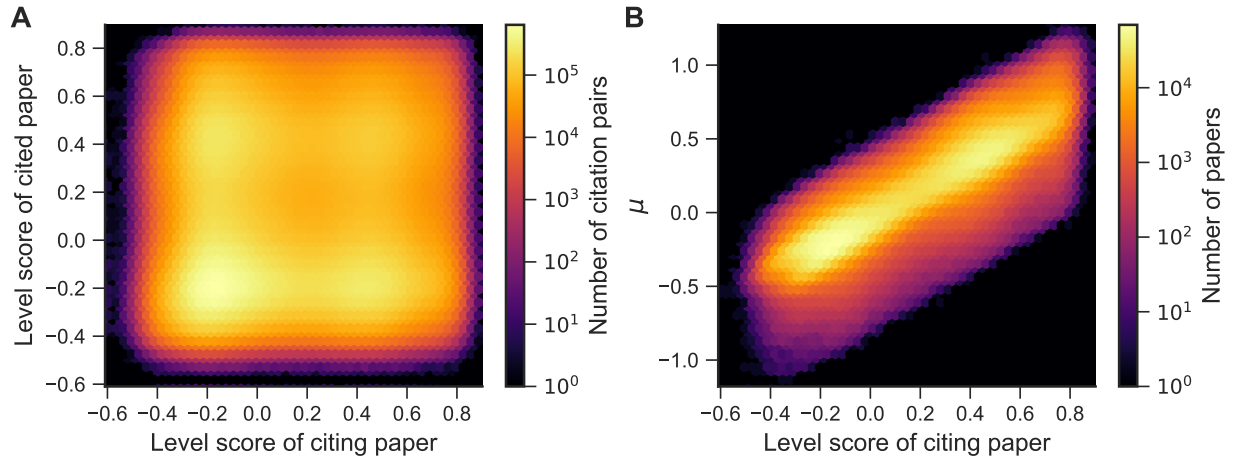


FIG. S7. The same as Fig. 4 in the main text, but the results are obtained from randomized citation network by preserving citation dynamics of each paper (Section S1). The results are averaged over 10 realizations.

## **S2. ALTERNATIVE CALCULATION OF LEVEL SCORE OF PAPERS**

In the main text, we calculated the level score of a paper as the average of the cosine similarities between the Translational Axis (TA) vector and the MeSH term vectors. Here we provide another way to do this. Specifically, we first obtain the TA vector as before, and get the vector of a paper by averaging the vectors of its MeSH terms. The LS of the paper is then the cosine similarity between the TA vector and the paper vector. Figs. [S8–S10](#) demonstrate that our results remain similar.

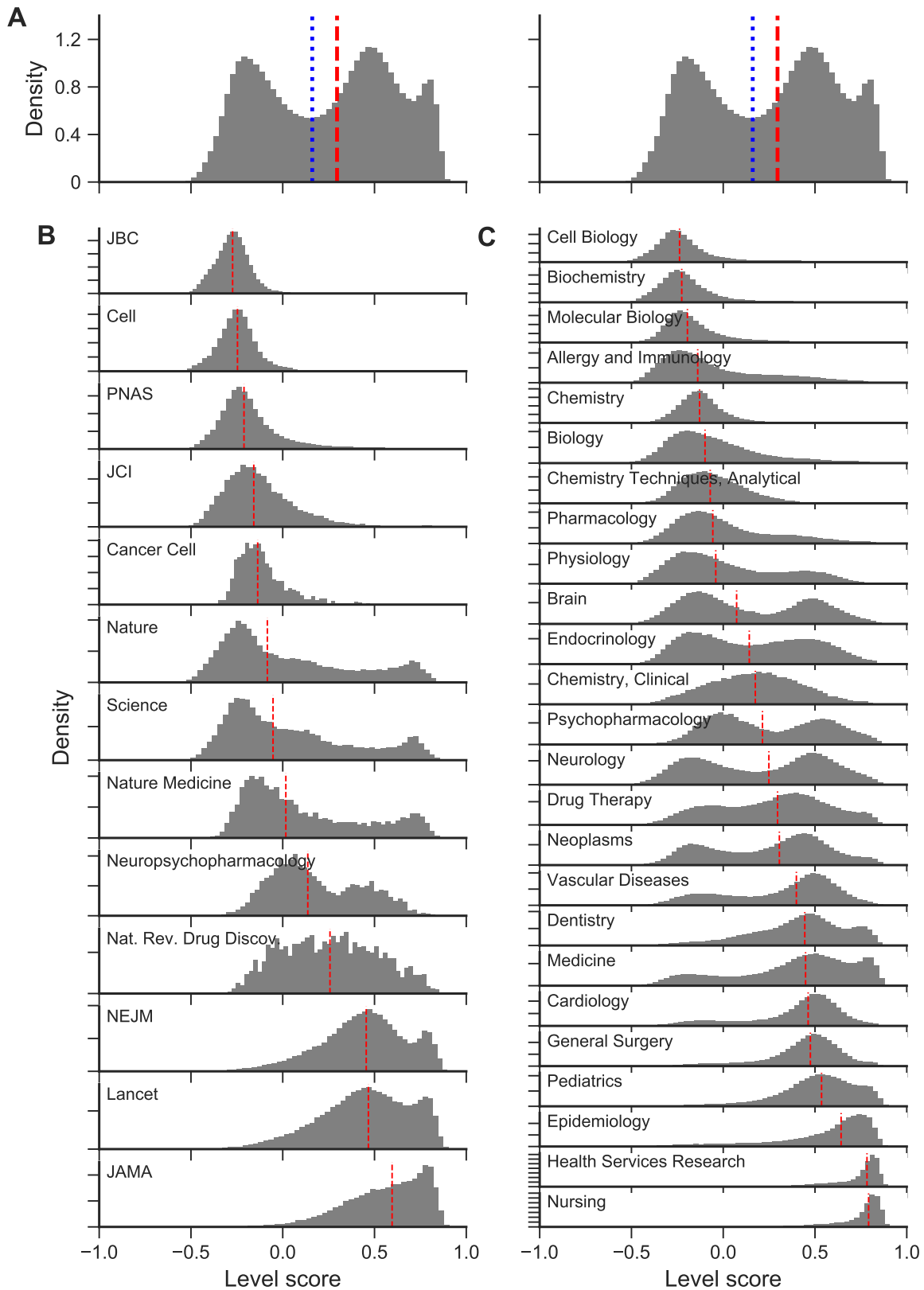


FIG. S8. The same as Fig. 2 in the main text, but the level score of a paper is the cosine similarity between the Translational Axis (TA) vector and the vector of the paper, obtained by averaging the vectors of its MeSH terms. Sec. S2 describes this in details. MeSH term vectors are obtained using LINE.

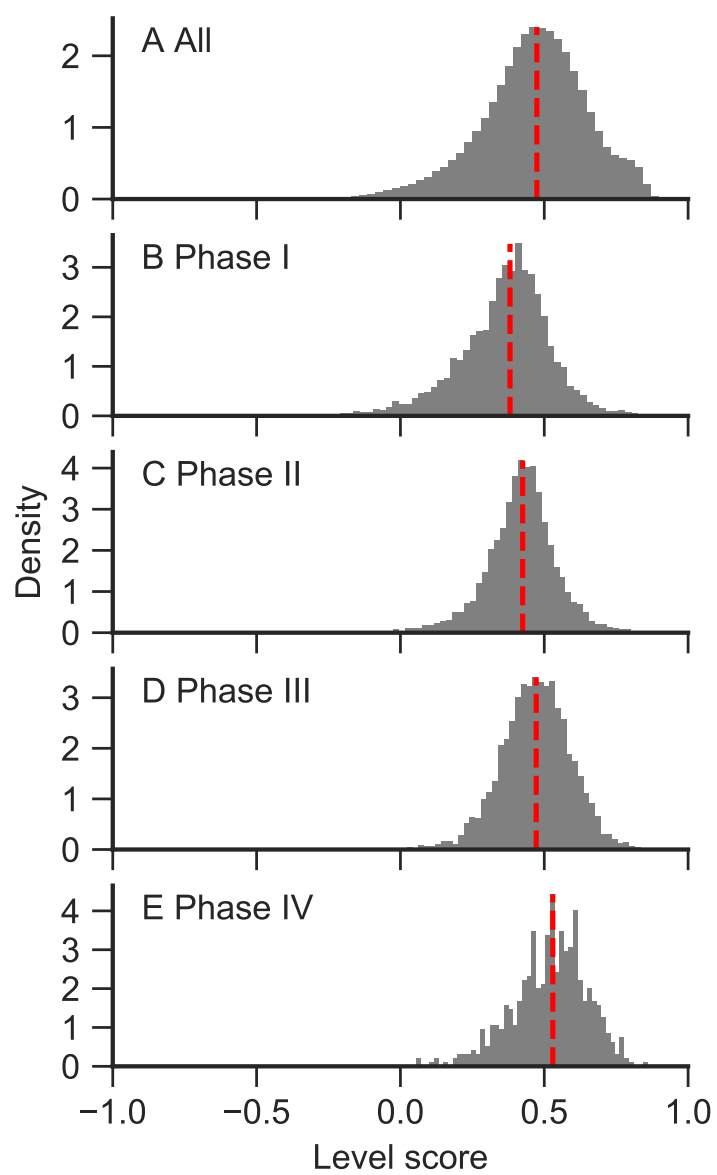


FIG. S9. The same as Fig. S4, but level score is obtained using the procedure described in Sec. S2.

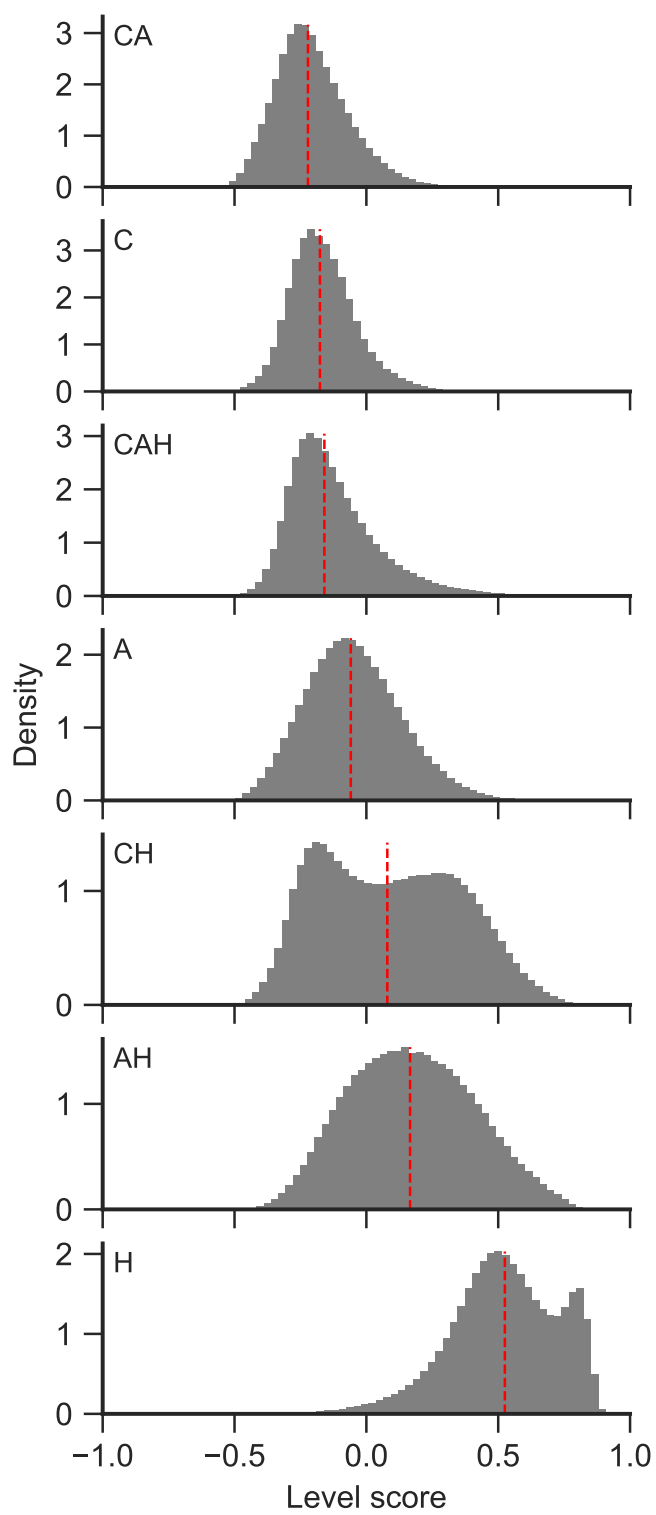


FIG. S10. The same as Fig. S5, but level score is obtained using the procedure described in Sec. S2.

### S3. ROBUSTNESS TEST

Figs. S11–S18 show the same set of results we showed before, but using the GloVe embedding method [1].

---

- [1] J. Pennington, R. Socher, and C. D. Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, pages 1532–1543, 2014.



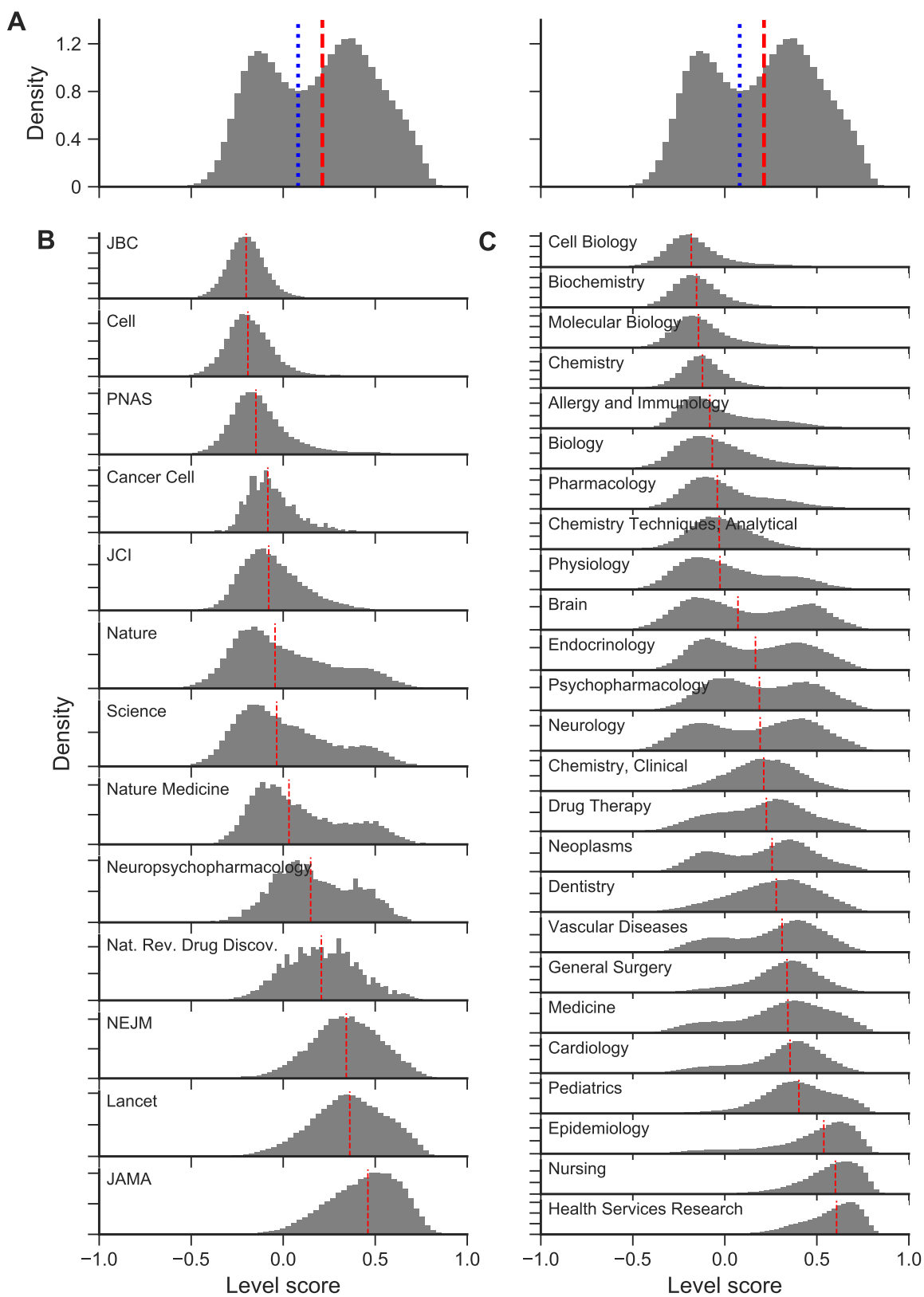


FIG. S11. The same as Fig. 2 in the main text, but level score is obtained based on the GloVe embedding method.

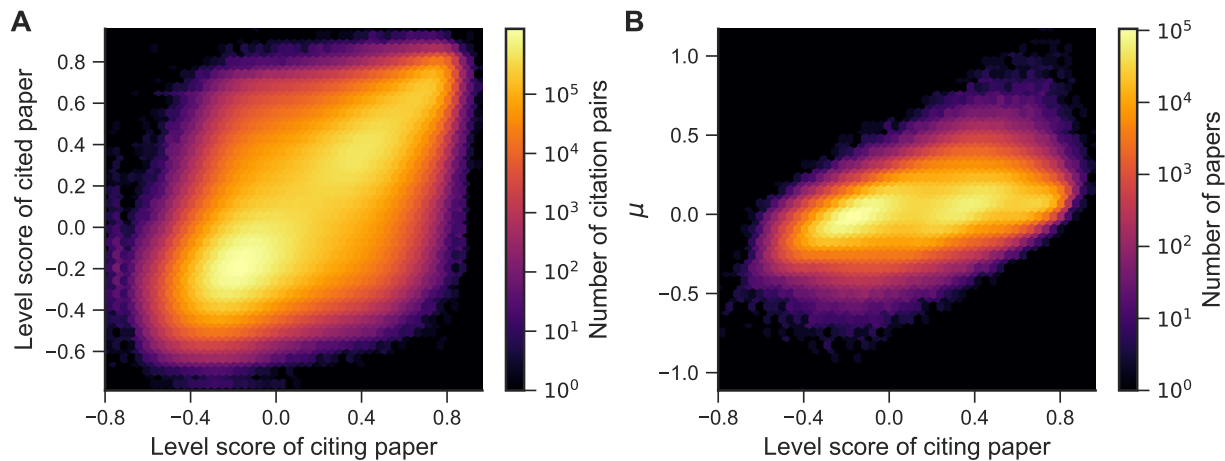


FIG. S12. The same as Fig. 4 in the main text, but based on the GloVe embedding method.

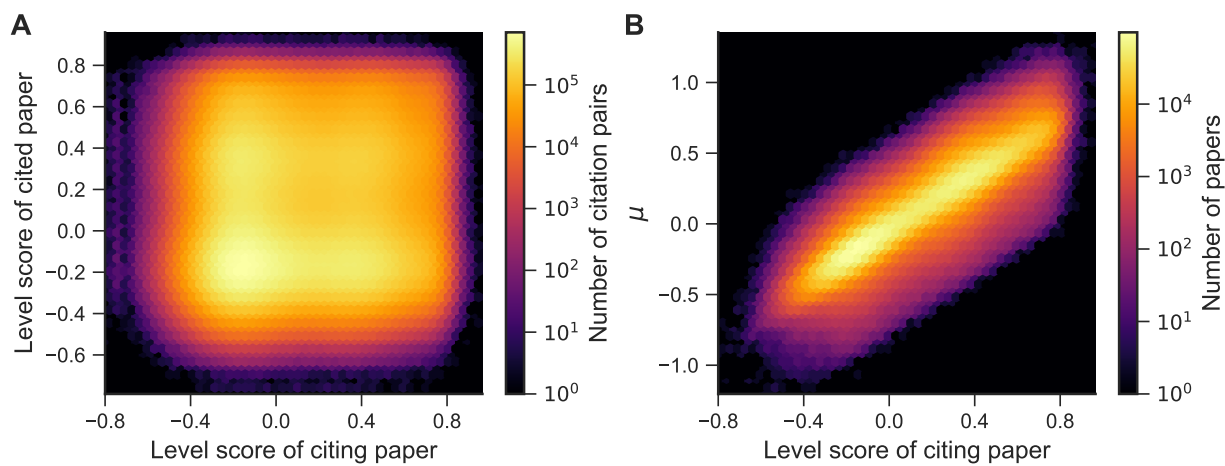


FIG. S13. The same as Fig. S7, but based on the GloVe embedding method.

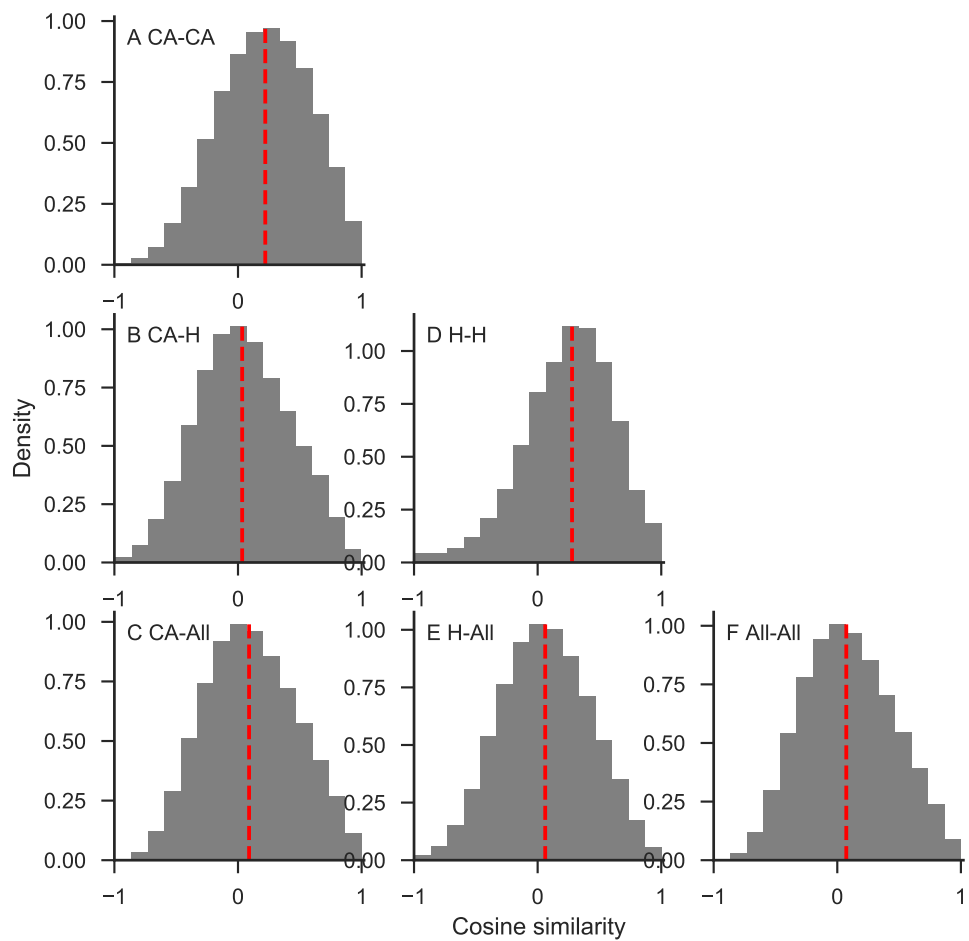


FIG. S14. The same as Fig. S1, but based on the GloVe embedding method.

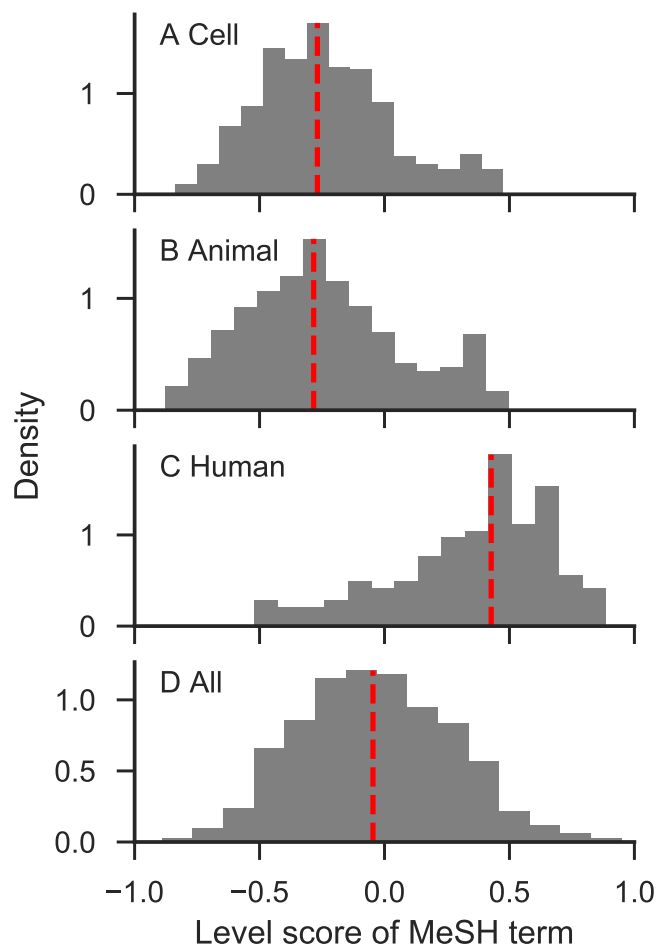


FIG. S15. The same as Fig. S2, but based on the GloVe embedding method.

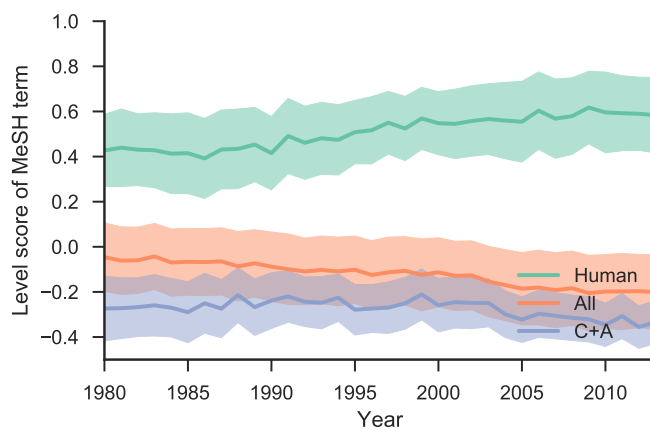


FIG. S16. The same as Fig. S3, but based on the GloVe embedding method.

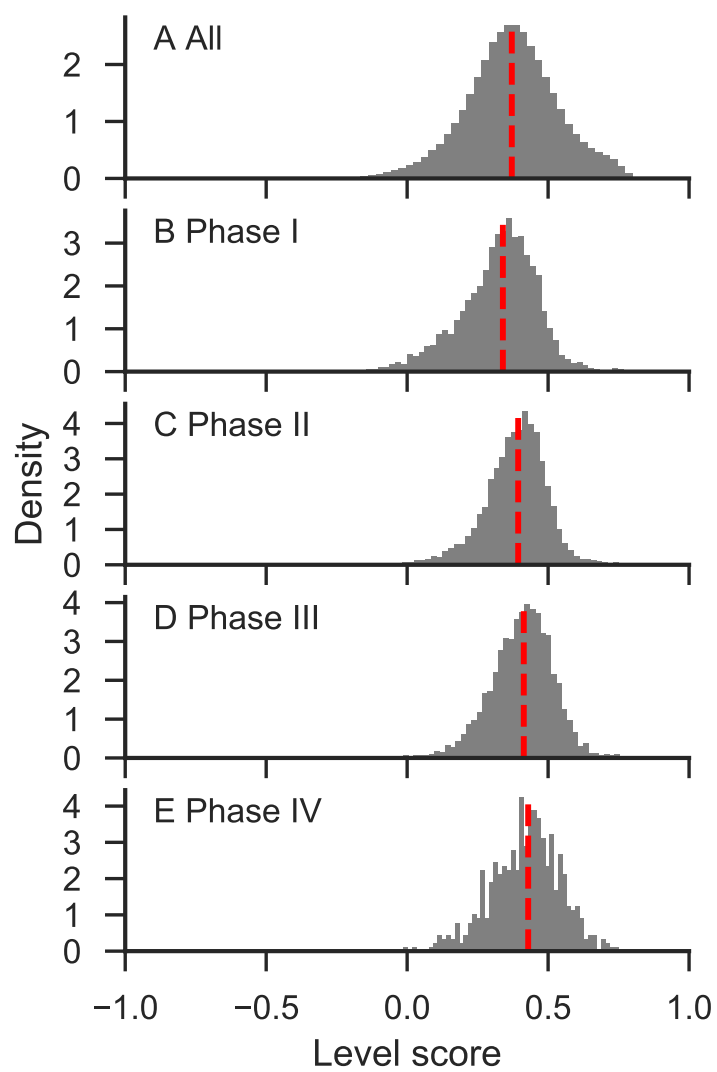


FIG. S17. The same as Fig. S4, but based on the GloVe embedding method.

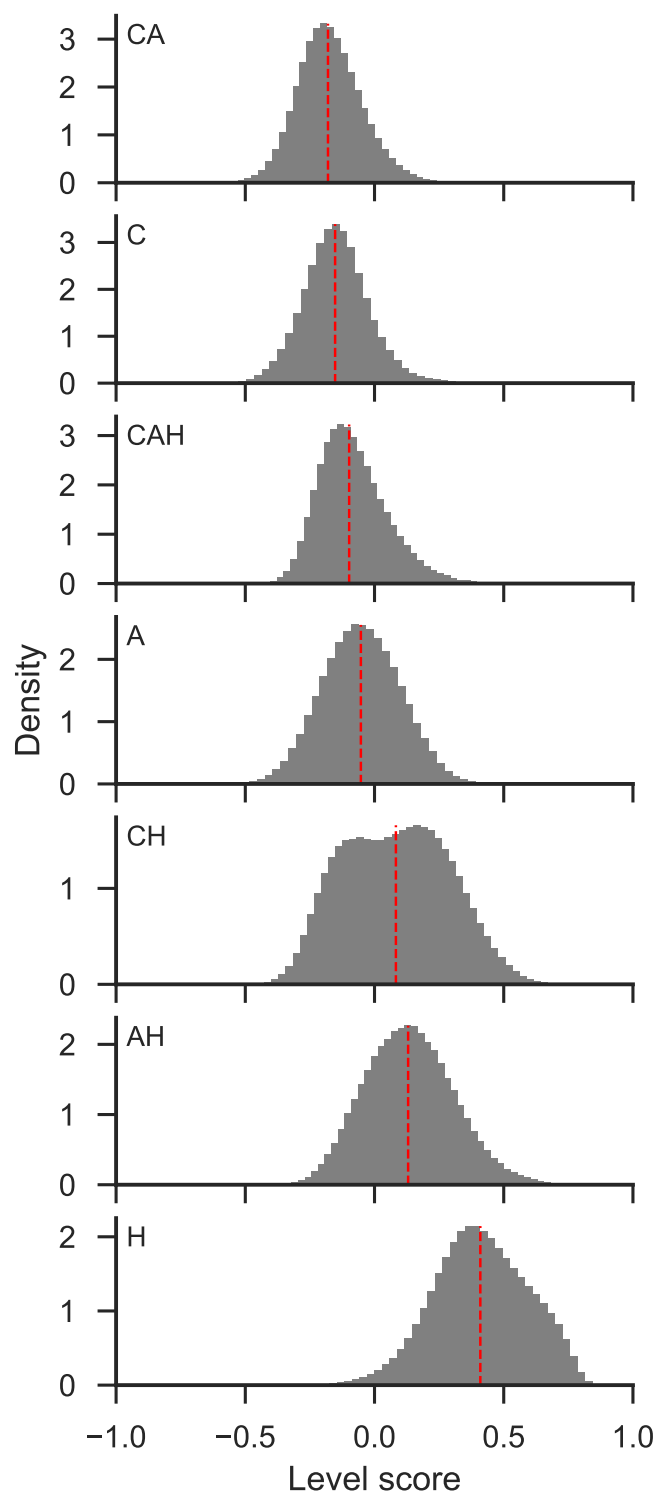


FIG. S18. The same as Fig. S5, but based on the GloVe embedding method.