Science Advances
AAAS

advances.sciencemag.org/cgi/content/full/7/17/eabb9004/DC1

Supplementary Materials for

Neural embeddings of scholarly periodicals reveal complex disciplinary organizations

Hao Peng, Qing Ke, Ceren Budak, Daniel M. Romero, Yong-Yeol Ahn*

*Corresponding author. Email: yyahn@iu.edu

Published 23 April 2021, *Sci. Adv.* **7**, eabb9004 (2021)
DOI: 10.1126/sciadv.abb9004

This PDF file includes:

S1 Supplementary Materials

S1.1 Supplementary Tables

Discipline Category	Num. of Journals	Percentage
Biology	1057	8.27
Biotechnology	238	1.86
Brain Research	741	5.80
Chemical, Mechanical, & Civil Engineering	1023	8.00
Chemistry	644	5.04
Earth Sciences	490	3.83
Electrical Engineering & Computer Science	779	6.10
Health Professionals	1387	10.85
Humanities	654	5.12
Infectious Diseases	660	5.16
Math & Physics	738	5.77
Medical Specialties	1657	12.96
Social Sciences	2712	21.22
Interdiscipline	29	0.02

Table S1: The number of journals in 13 disciplines defined in the UCSD map of science. These 12,780 journals can be matched between the MAG data and the UCSD map data, and are covered in our embeddings. 29 journals belonging to multiple disciplines are labeled as "Interdiscipline". Throughout the paper, we abbreviate "Chemical, Mechanical, & Civil Engineering" as "Engineering", and "Electrical Engineering & Computer Science" as "EE & CS" to save space.

W	D	ΔMean(**sub**)	ΔMean(**dis**)	Mean(**rand**)
2	50	0.302	0.105	0.233
2	100	0.349	0.118	0.253
2	200	0.314	0.103	0.250
2	300	0.299	0.096	0.243
5	50	0.432	0.179	0.073
5	100	0.457	0.183	0.086
5	200	0.419	0.165	0.084
5	300	0.399	0.157	0.082
10	50	0.420	0.172	0.069
10	**100**	**0.469**	**0.192**	0.069
10	200	0.428	0.172	0.067
10	300	0.406	0.161	0.066

Table S2: **Hyperparameter tuning in the model training**. Each model is trained with the same 100 million periodical citation trails. The minimum frequency is set to 50 in all settings. W is the context window size, D is the number of embedding dimensions. Mean(**sub**), Mean(**dis**), and Mean(**rand**) are the mean cosine similarity of journal pairs in the same sub-discipline, journal pairs in the same discipline, and journal pairs in any discipline, respectively. Note that we randomly selected 100,000 journal pairs for each group. ΔMean(**sub**) $=$ Mean(**sub**) $-$ Mean(**rand**), ΔMean(**dis**) $=$ Mean(**dis**) $-$ Mean(**rand**)).

Discipline	**Num. of Participants**	**Num. of Selected Targets**
Social Sciences	50	318
EE & CS	39	224
Engineering	20	129
Math & Physics	3	21
Earth Sciences	2	10
Health Professionals	2	9
Biology	1	11
Brain Research	1	5
Biotechnology	1	4

Table S3: The number of qualified participants and the total number of target journals selected by them across different disciplines.

S1.2 Supplementary Figures

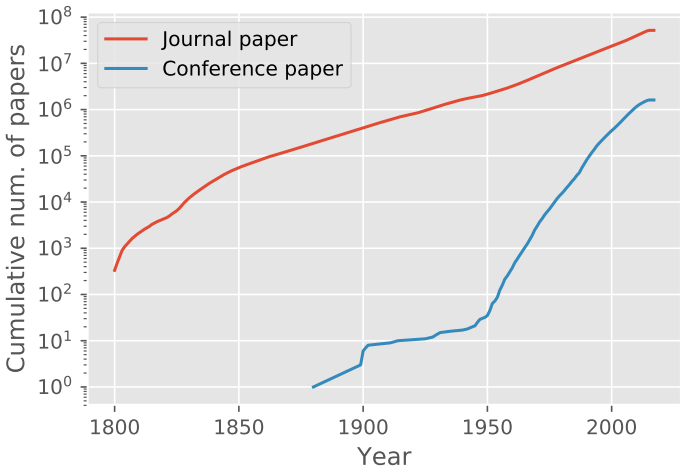

Figure S1: The cumulative number of journal and conference papers from 1800 to 2016. A total number of 53 million papers are used in this study.

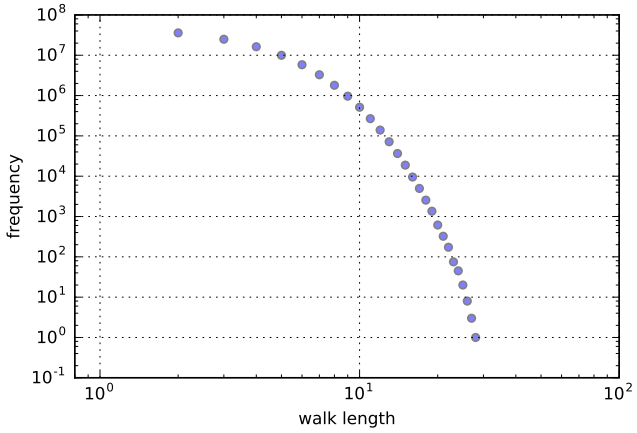

Figure S2: The length distribution of 100 million periodical trails. Note that length-one trails were discarded during the random walk process.

A

Top 10 venue similar to PNAS

Journal	Similarity
Nature	0.87
Science	0.82
BioEssays	0.80
Cell Rep.	0.80
EMBO J.	0.79
Nat. Commun.	0.78
Curr. Biol.	0.77
Annu. Rev. Biochem.	0.77
PLoS Biol.	0.77
BMC Biol.	0.75

B

Biochem. J. − Cell + PRL

Journal	Similarity
Phys. Rev. B	0.64
J. Phys. Chem.	0.59
Trans. Faraday Soc.	0.58
J. Electrochem. Soc.	0.57
Acta Chem. Scand.	0.57
Phys. Rev. A	0.56
Helv. Chim. Acta	0.55
Anal. Chem.	0.54
Phys. Rev.	0.54
Analyst	0.53

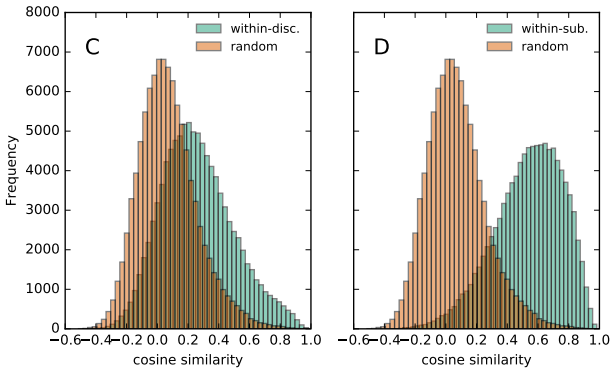

Figure S3: **Periodical recommendations.** (A) The 10 most similar periodicals to *PNAS*, a multi-disciplinary yet biomedical-dominated journal, based on the cosine similarities between periodical embeddings. Other multi-disciplinary journals, such as *Nature*, *Science*, *Nature Communications*, and some biological journals are among the top list. (B) The 10 most similar periodicals to the vector analogy: $\mathbf{v}(\textit{Biochemical Journal}) - \mathbf{v}(\textit{Cell}) + \mathbf{v}(\textit{Physical Review Letters})$. (C) The histogram of cosine similarities of 100,000 randomly selected journal pairs that are in the same discipline ("within-disc.") or in any discipline ("random"). (D) As in (C), but for journal pairs in the same sub-discipline ("within-sub.").

Figure S4: **The interface of the journal recommendation survey.** **a**, The survey interface where participants were first asked to choose a discipline to begin the task. **b**, Participants were asked about their familiarity with the 20 target journals in "Social Sciences". The survey continues only if at least 3 target journals were selected.

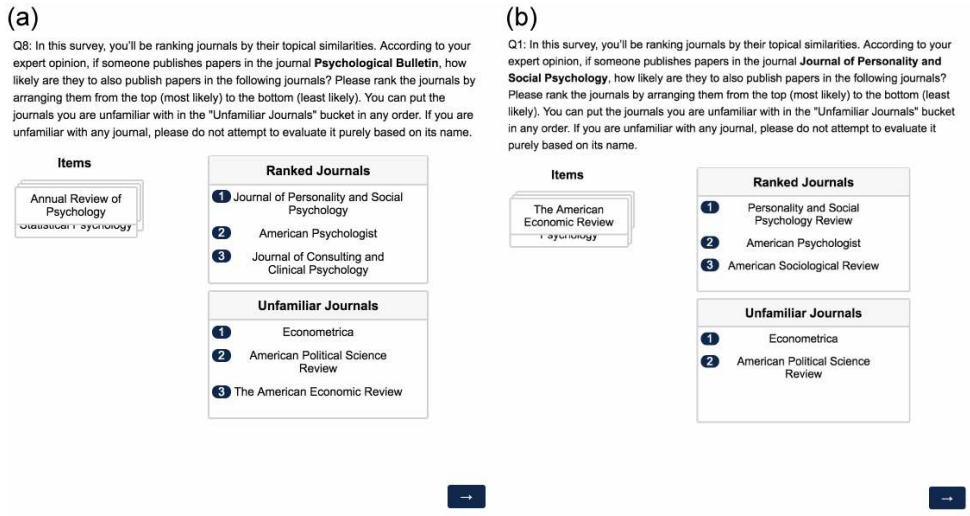

Figure S5: **Screenshots of the rank task interface for two exemplar target journals.** **a**, *Psychological Bulletin*. **b**, *Journal of Personality and Social Psychology*. The candidate journals on the left side are randomly stacked on top of each other. Participants can place unfamiliar candidates in the "Unfamiliar Journals" bucket in any order.

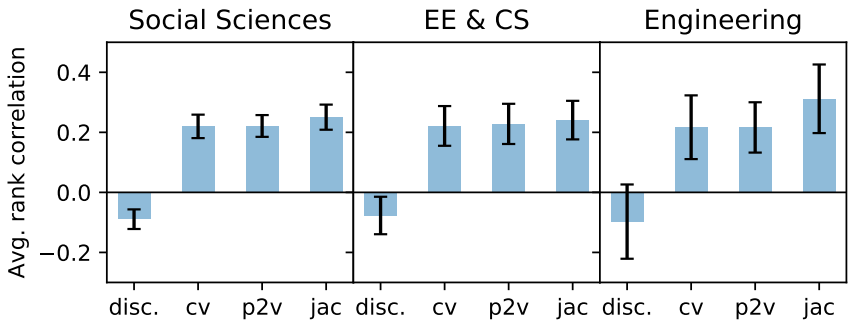

Figure S6: **The average Kendall's rank correlation coefficient between experts and four models in three disciplines.** Target journals with an average expert agreement above 0.2 are used in the evaluation. The four labels—*disc.*, *cv*, *jac*, and *p2v*—represent the first baseline method, the citation-based sparse vector-space model, the Jaccard similarity matrix, and our periodical embeddings.

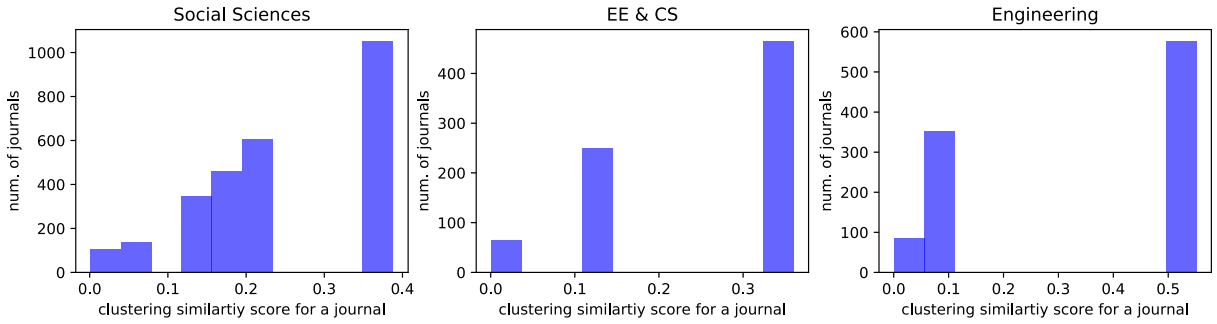

Figure S7: Histograms of the similarity (agreement) scores between two clusterings (UCSD journal categorizations vs. a clustering based on our periodical embeddings) for journals in three disciplines, which display a multimodal distribution. The discreteness comes from the fact that there are only 13 clusters and we are comparing two clusterings.

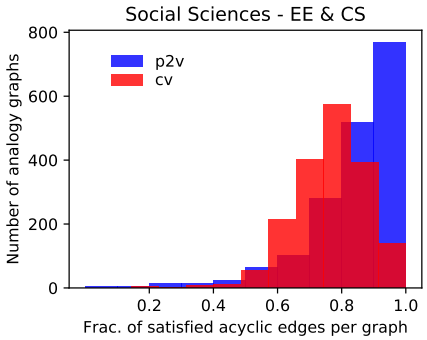

Figure S8: The distribution of the fraction of acyclic edges that satisfy the author overlap criterion across 1,800 analogy graphs in ("Social Sciences", "EE & CS"). For a periodical analogy "$A : B \sim C : D$", we verify whether $\frac{O(C,A)}{O(C,B)} > \frac{O(D,A)}{O(D,B)}$. Here $O(P_1, P_2)$ indicates the number of common authors who have ever published a paper in both periodicals. Our embedding method (*p2v*) is compared against the citation-based sparse encoding model (*cv*) for generating analogy graphs. The mean value for *p2v* is significantly higher than that for *cv* (0.84 vs 0.76).

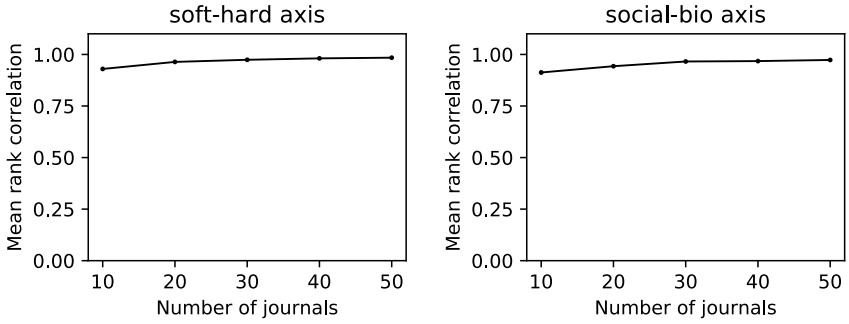

Figure S9: The average Spearman's rank correlation between the ordering of 20,835 periodicals (covered in our embedding model) and their ordering based on an axis built with a subset of randomly selected journals in the two broad disciplines. The x-axis represents the number of journals in the subset. Error bars indicate 95% confidence intervals.

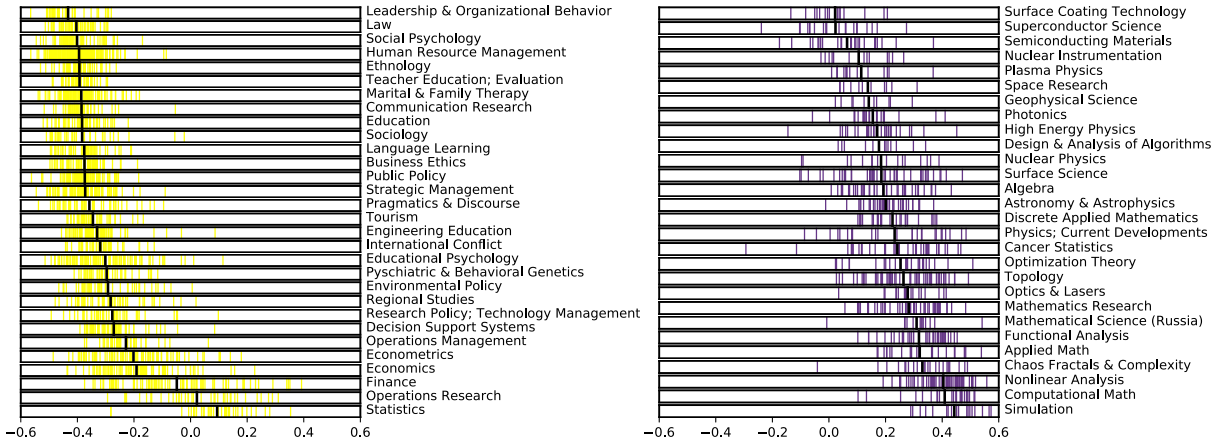

Figure S10: **The organization of sub-disciplines on the "soft-hard" sciences axis. Left,** The spectrum of sub-disciplines in "Social Sciences" on this axis. Each journal is represented by a vertical line inside the box. Sub-disciplines are ordered by their mean values (the black vertical line). There is a relatively clear separation of subfields on this axis. **Right,** The spectrum of sub-disciplines in "Math & Physics" on the "soft-hard" sciences axis, which does not exhibit a separation as clear as that in "Social Sciences". Note that, to save space, only the top 30 subfields (based on their number of journals) are shown in each discipline.

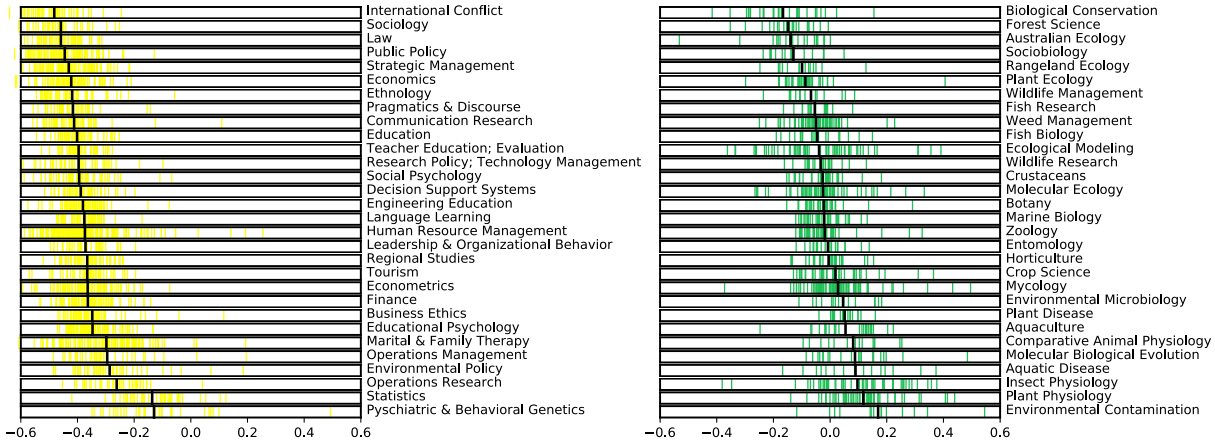

Figure S11: **The organization of sub-disciplines on the "social-bio" science axis. Left**, The spectrum of sub-disciplines in "Social Sciences" on this axis. Each journal is represented by a vertical line inside the box. Sub-disciplines are ordered by their mean values (the black vertical line). There is a relatively clear separation of subfields on this axis. **Right**, The spectrum of sub-disciplines in "Biology" on the "social-bio" science axis, which does not exhibit a separation as clear as that in "Social Sciences". Note that only the top 30 subfields (based on their number of journals) are shown in each discipline.

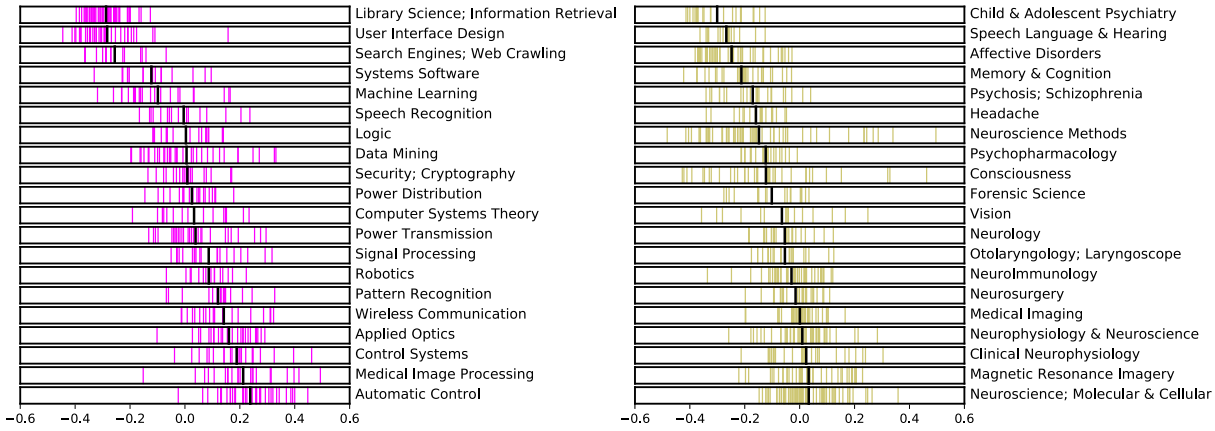

Figure S12: **The ordering of sub-disciplines on the "soft-hard" sciences axis. Left,** The spectrum of journals in sub-disciplines of "**EE & CS**" on the "soft-hard" sciences axis. Each journal is represented by a vertical line inside the box. The color represents the discipline category in the UCSD map of science. We focus on the top 20 sub-disciplines based on the number of journals, and ordered each category by their mean projection values (the black vertical line). Research domains such as "Library Science", "Information Retrieval", "User Interface Design", "Machine Learning", and "Data Mining" are "softer" than domains such as "Signal Processing", "Robotics", "Wireless Communication", and "Controls Systems". **Right,** The spectrum of journals in "**Brain Research**" on the "soft-hard" sciences axis. Sub-disciplines such as "Neurology", "Medical Imaging", and "Magnetic Resonance Imagery" are "harder" than "Psychiatry", "Speech", "Hearing", "Headache", and "Consciousness".

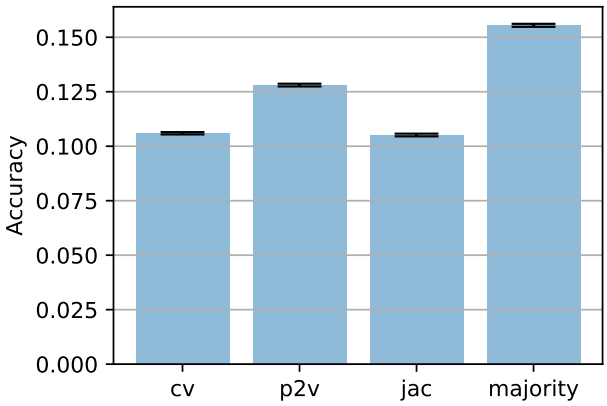

Figure S13: **The accuracy of four different methods in predicting papers' publication venue based on their cited periodicals.** We use a random sample of 10,000 papers published after year 2000 (as younger papers are likely to have more complete reference information in the MAG dataset). The experiment is repeated 100 times to estimate the error. Our dense periodical embedding model (*p2v*) is compared with the citation vector model (*cv*), the Jaccard similarity matrix (*jac*), and another baseline method (*majority*) that predicts a paper's publication venue to be the most frequently cited periodical in its reference list. The three vector-space models predict the periodical that is the closest to the average vector of cited periodicals based on cosine similarity.

S1.3 Annotated maps of journals in each discipline

The 2-*d* projection of the embeddings of 12,780 journals provides an overview of the organizational structure of major academic disciplines. Here we further investigate the interdisciplinary nature of many academic journals in each discipline and show that many of them cannot be properly categorized into a unique discipline by an existing journal classification system—the UCSD map of science. Figs. S14–S26 highlight all journals in a given discipline with journals in all other disciplines blurred in the background. The discipline category comes from the UCSD catalog. In the map of each discipline, we annotated some exemplar interdisciplinary journals and micro-clusters that are located near disciplinary boundaries or are far away from their main discipline clusters. When annotating individual journals, we also provide its cosine distance to the centroid of all journal vectors in that discipline. Journals that are far away from its discipline center are likely to be misclassified in the UCSD map of science.

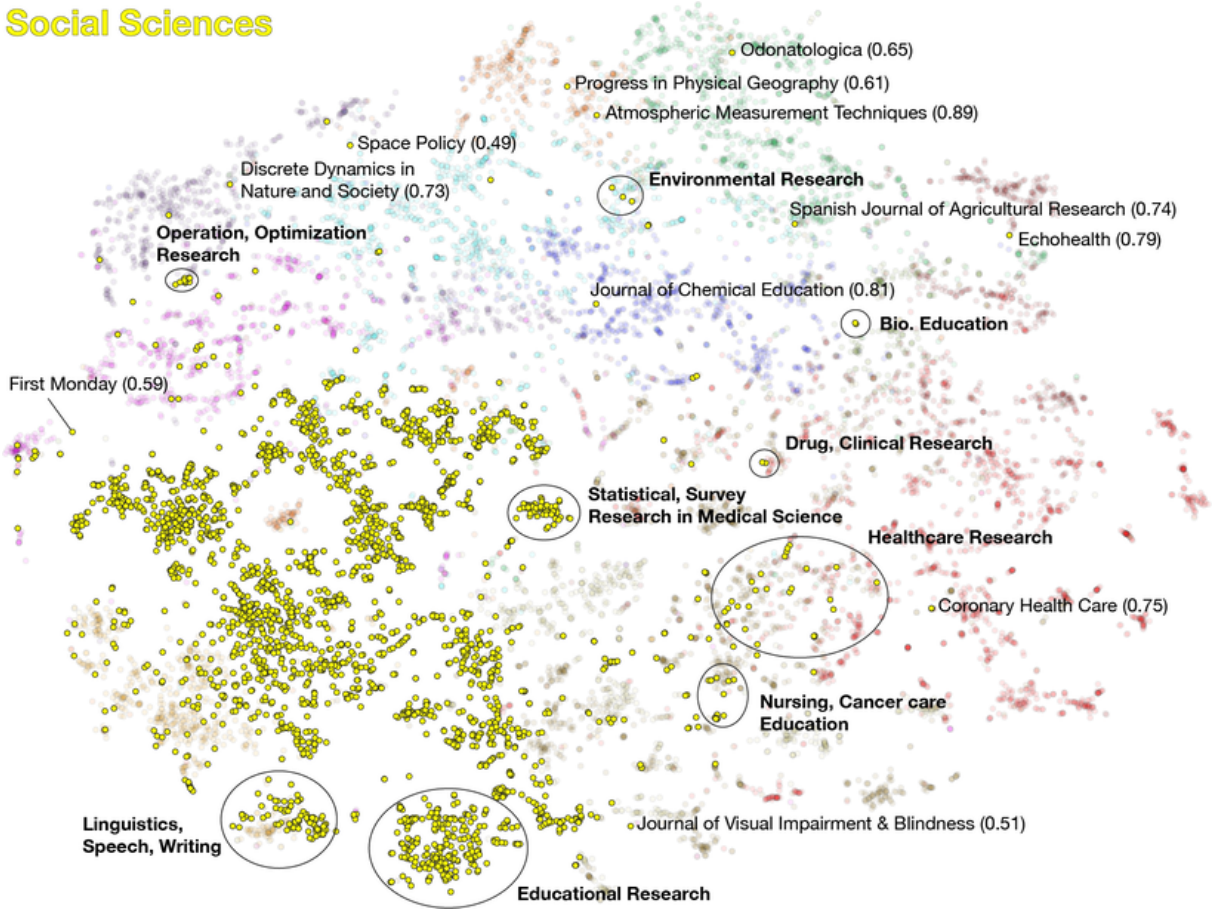

Figure S14: The realm of "Social Sciences" journals in the embedding space.

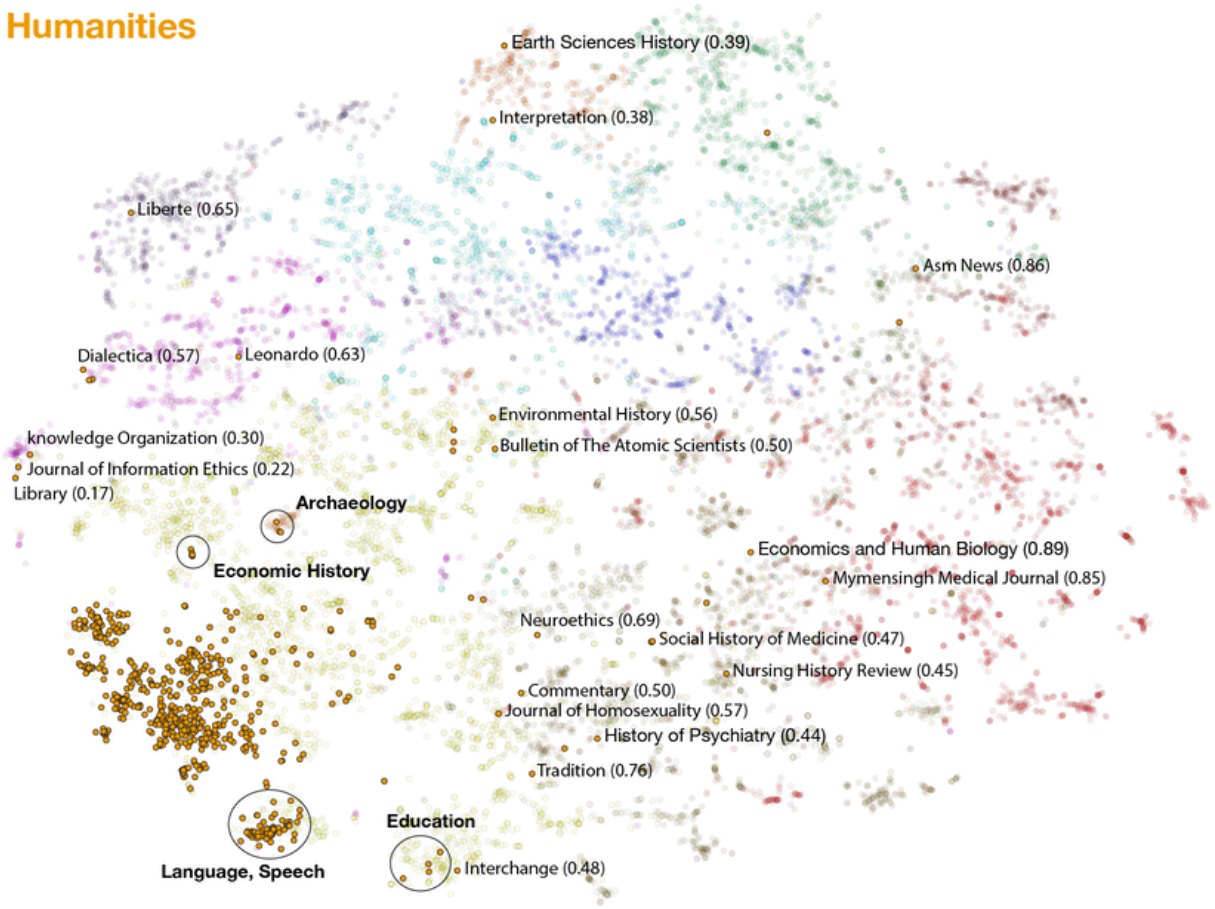

Figure S15: The region of "Humanities" journals in the embedding space.

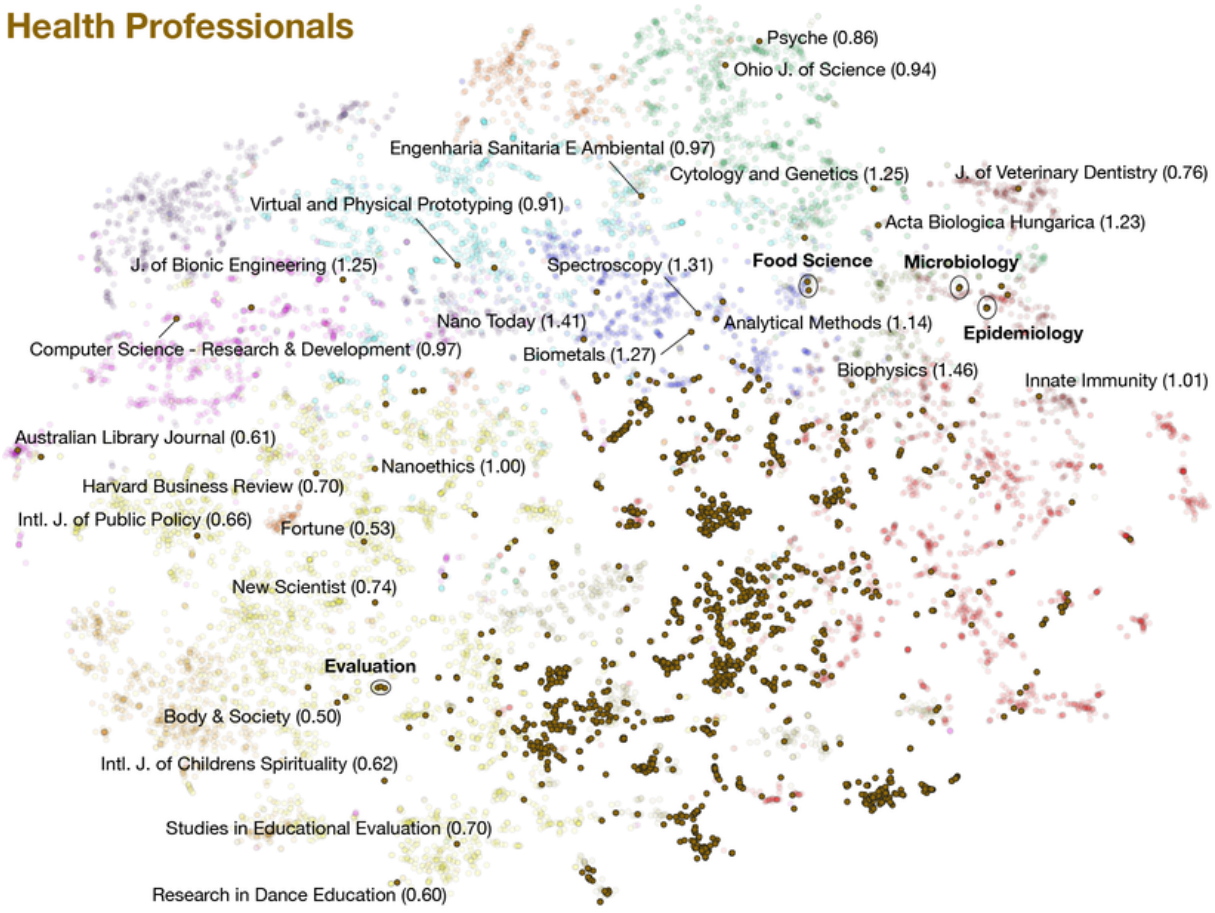

Figure S16: The colony of "Health Professionals" journals in the embedding space.

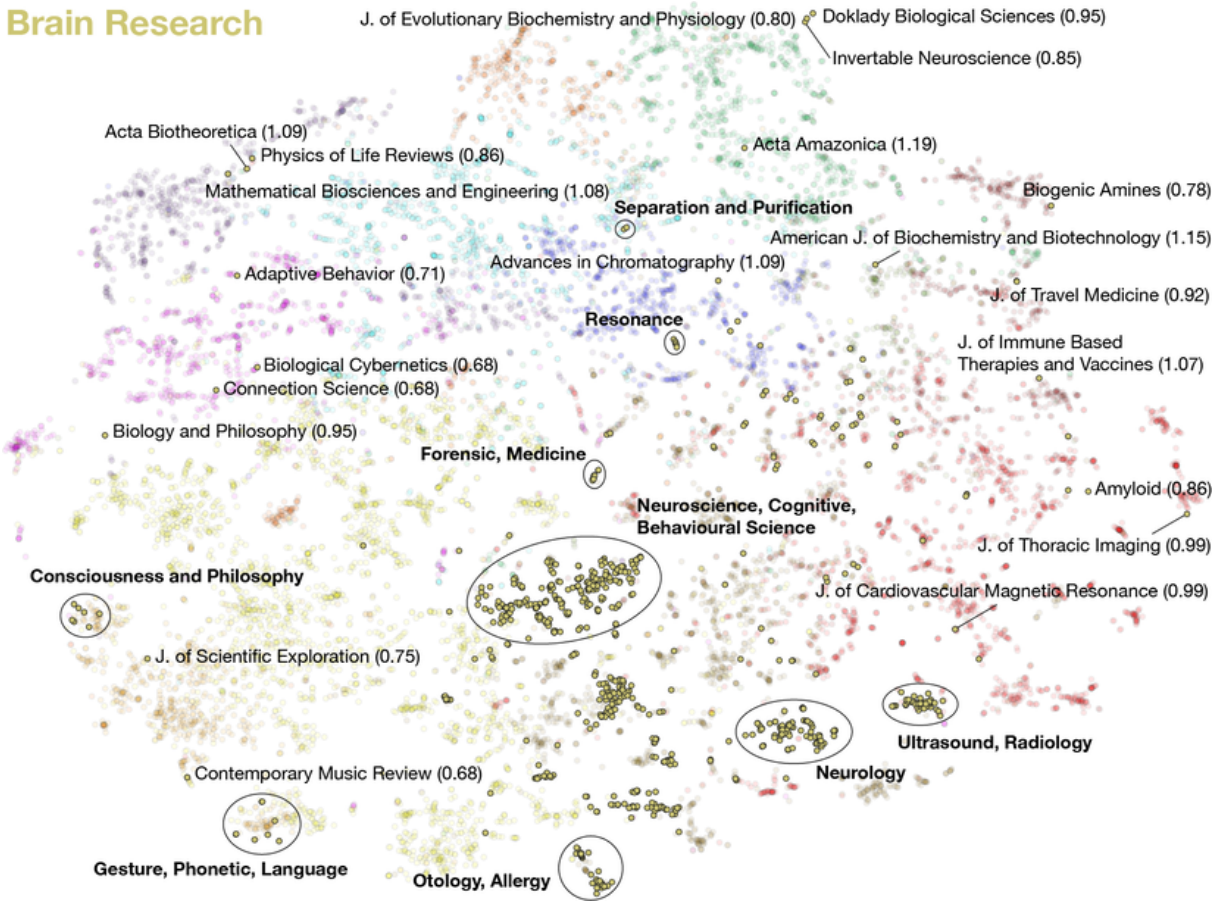

Figure S17: The realm of "Brain Research" journals in the embedding space.

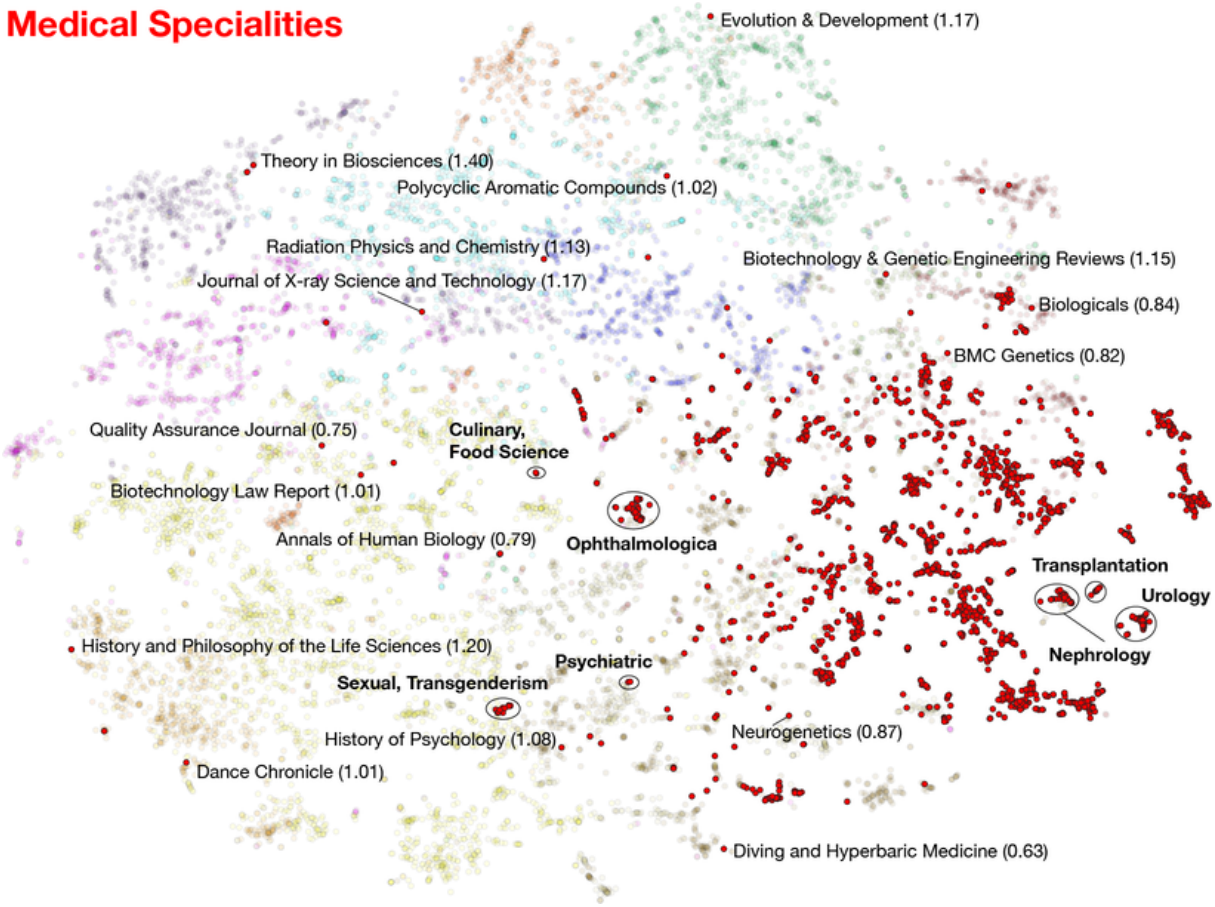

Figure S18: The region of "Medical Specialties" journals in the embedding space.

Figure S19: The colony of "Biology" journals in the embedding space.

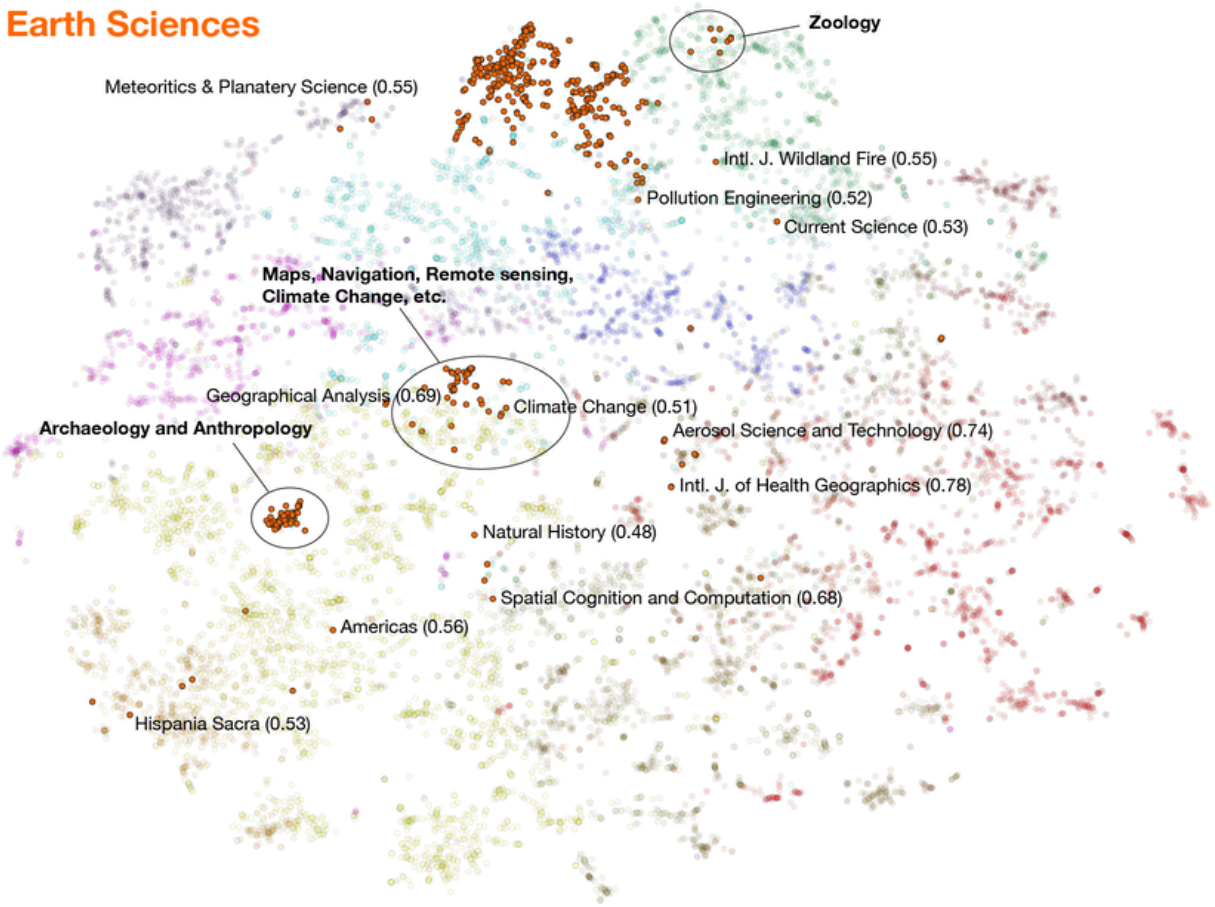

Figure S20: The realm of "Earth Sciences" journals in the embedding space.

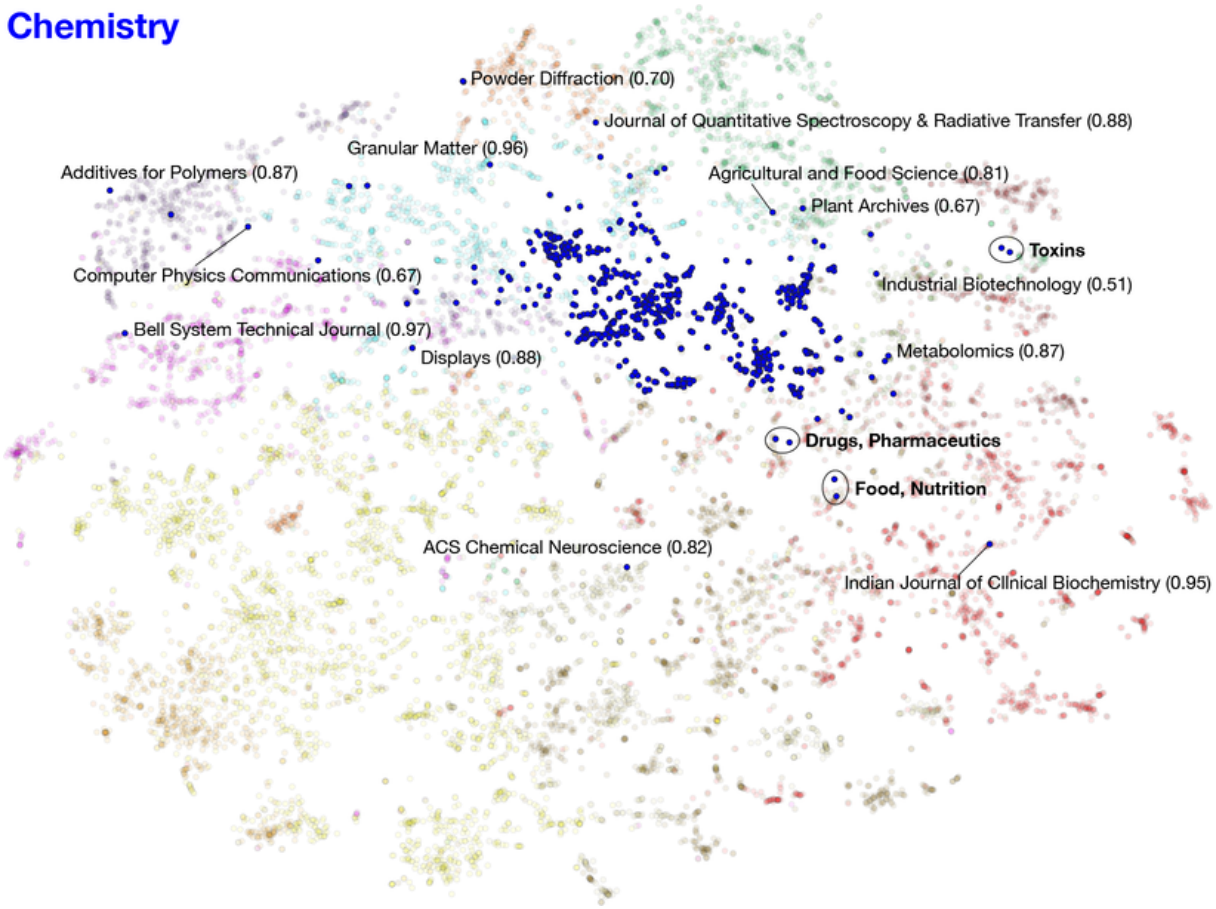

Figure S21: The region of "Chemistry" journals in the embedding space.

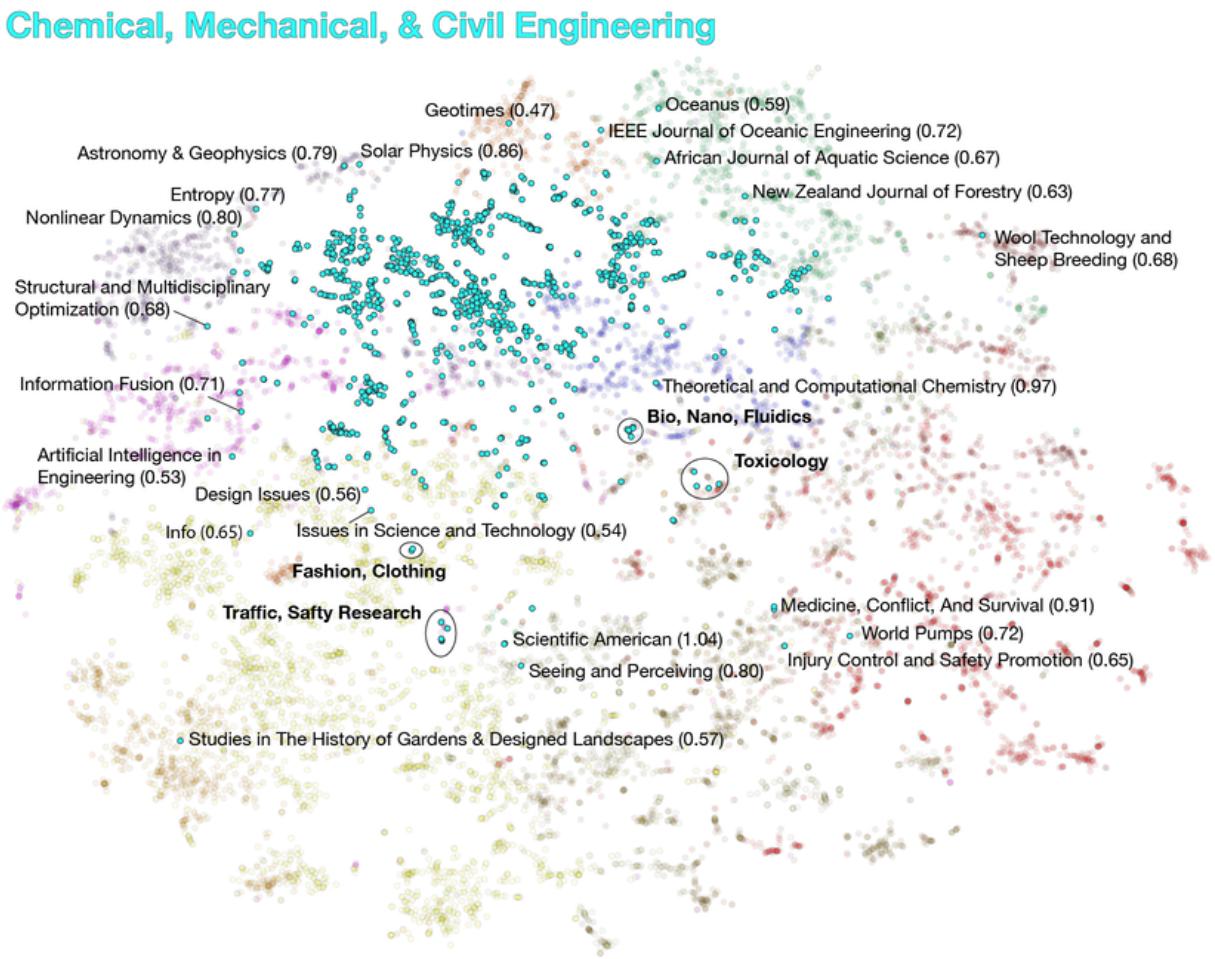

Figure S22: The colony of "Chemical, Mechanical, & Civil Engineering" journals.

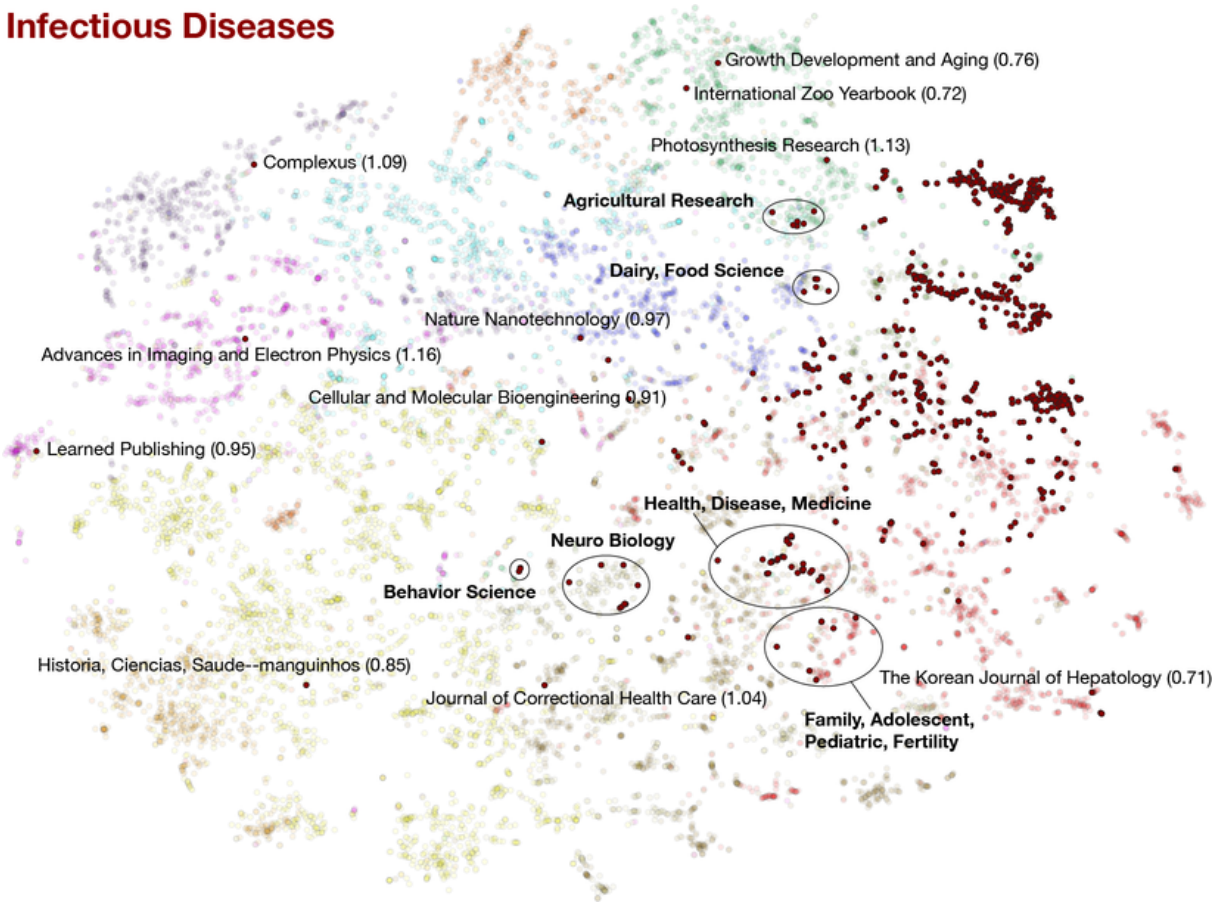

Figure S23: The realm of "Infectious Diseases" journals in the embedding space.

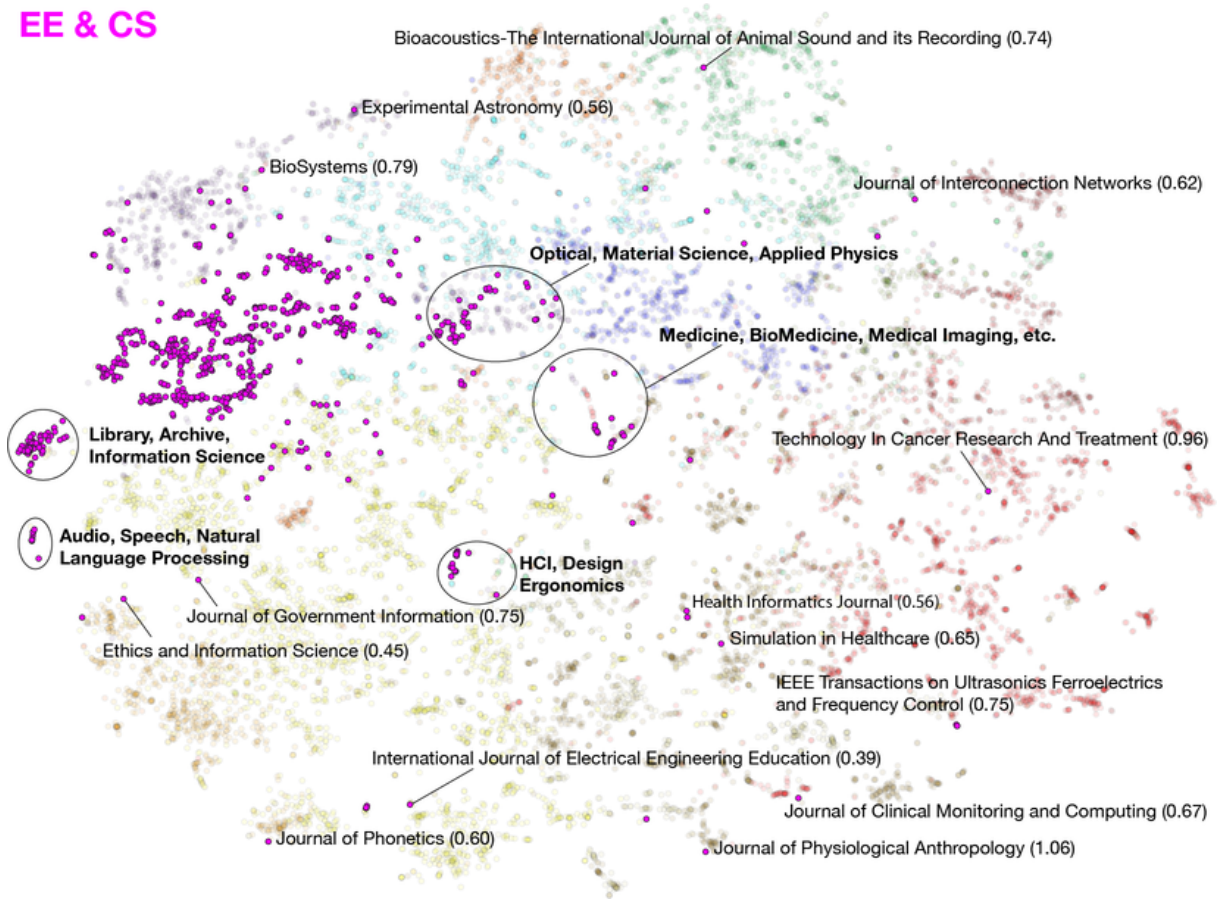

Figure S24: The region of "Electrical Engineering & Computer Science" journals.

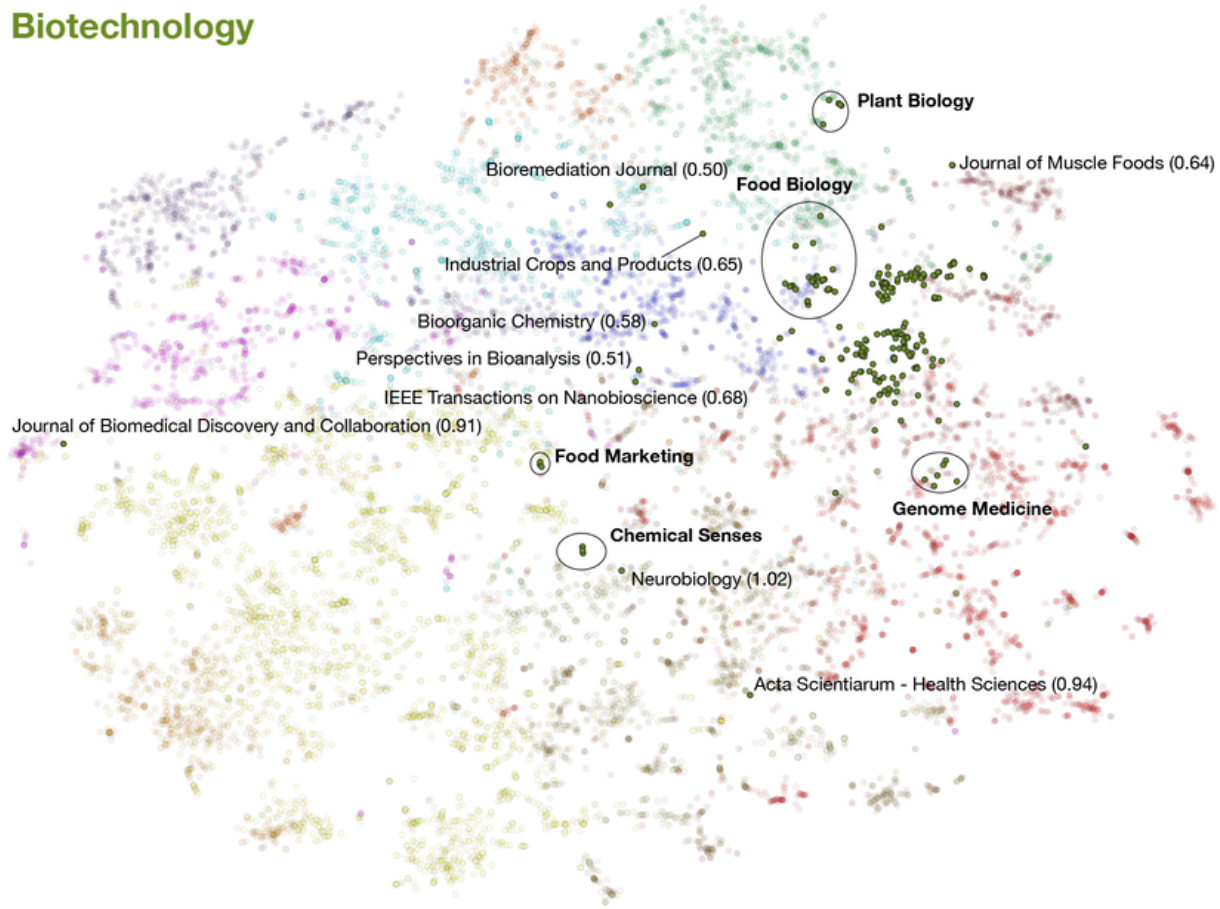

Figure S25: The colony of "Biotechnology" journals in the embedding space.

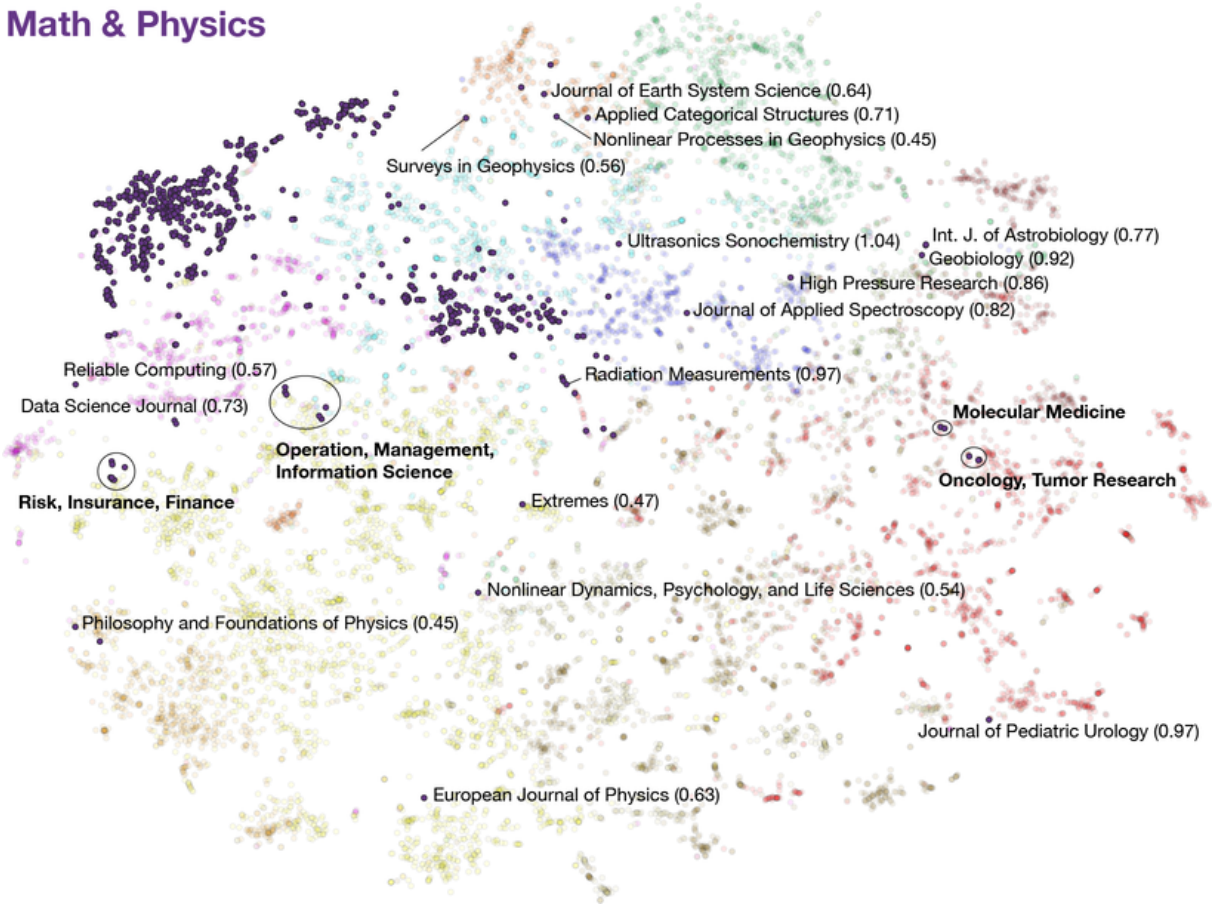

Figure S26: The territory of "Math & Physics" journals in the embedding space.