# Modeling Convergence Undercurrents in Brain Science via Statistics and Machine Learning

by
Mohammed Emtiaz Ahmed

A dissertation submitted to the Department of Computer Science,

College of Natural Sciences and Mathematics

in partial fulfillment of the requirements for the degree of

Doctor of Philosophy
in Computer Science

Chair of Committee: Ioannis Pavlidis

Committee Member: Lennart Johnsson

Committee Member: Christoph F. Eick

Committee Member: Christiane Spitzmueller

University of Houston
August 2020

## ACKNOWLEDGMENTS

First and foremost, I express my gratitude to my advisor Dr. Ioannis Pavlidis for his amazing guidance and support throughout my graduate study. This work would have not been possible without his insights and efforts. His valuable feedback pointed me in the right direction.

Besides my advisor, I would like to thank my committee members Dr. Lennart Johnsson, Dr. Christopher Eick, and Dr. Christiane Spitzmueller for serving on my committee and providing me valuable feedbacks.

My sincere thanks goes to Dr. Alexander Petersen, Dr. Brian Uzzi, and Dr. Dinesh Majeti for their help throughout my PhD.

I thank all my fellow labmates and alumni from Computational Physiology Lab for providing such a fun and friendly environment over the last few years.

Most of all, I owe special thanks to my parents, Tamanna Yesmin and Mohammed Ali Miah, my lovely wife Melia Mostafa, my sister Sazia Afrin, my father-in-law (Md. Golam Mostafa) and mother-in-law (Bilkis Jahan Belly) for their unconditional love and support during my entire life.

## ABSTRACT

From Leonardo da Vinci to the Nobel laureates, the science exemplar changed from the polymath to the dedicated specialist in response to evolutionary pressures. It is now believed that we are in the early stages of yet another evolutionary adaptation, where specialists from different disciplines need to collaborate to solve complex multidisciplinary problems. This idealized integration process, known as convergence, emerges as the new science exemplar. Despite the consequential nature of such a paradigm shift, the exact operationalization and efficacy of convergence remain unclear. To provide much needed answers to these two questions, we identified brain science as a unique convergence testbed to base our study. By jointly analyzing the disciplinary pedigree of the authors with the subject areas of nearly one million brain science publications between 1980 and 2019, we exposed a seeping convergence undercurrent: Science integration does not only neatly take place among researchers from different disciplines, but also awkwardly within researchers through expansive learning. Our models reveal three key findings: First, brain researchers tend to tackle subject areas beyond their core expertise, especially when these areas are epistemically close - a convergence shortcut. Second, this expansive learning behavior appears to precipitate and compete with true convergence, although it is clearly less impactful. In a finite science ecosystem, transnational epistemic moves by individual researchers effectively crowd out collaborations among distant disciplines, from where breakthroughs to grand challenges usually emanate. Third, major funding initiatives in brain science unknowingly promote shortcuts to convergence research.

Regarding the content of the research publications in our corpus, in addition to linear model analysis based on Medical Subject Headings (MeSH) keywords, we also applied Machine Learning (ML) analysis on the articles' abstracts. The ML methods yielded results that either validated or complemented those of the linear models. Furthermore, ML furnished insights regarding the timing and source of transformative developments in brain science that elucidate the abstract conclusions of the linear models. Such insights include the role and effect of Magnetic Resonance (MR) imaging and data analytic methods in brain science advancements.

# TABLE OF CONTENTS

# LIST OF TABLES

# LIST OF FIGURES

# 1 Introduction

It is widely acknowledged that scientific disciplines need to move toward convergence as a way of addressing complex problems [7]. The intellectual and institutional challenges of such an endeavor have preoccupied scholars and policy-makers for some time now [30]. At first glance, the scientific community's difficulty to come to grips with convergence seems paradoxical. Science was born out of convergence, and in a sense, it merely attempts to come full circle. Ancient philosophers were frequently indulging in multi-disciplinary pursuits, pondering with equal zest geometry, the make-up of the natural world, and human morality [15]. Even as far as the beginning of modern times, the polymath 'renaissance man' was the indisputable scholar model, famously exemplified in Leonardo da Vinci [17].

As science exploded with the advent of the industrial revolution, the original scholar model underwent a radical metamorphosis [4]. Gradually, scholars came to be identified with committed specialists. This modern scholarly ideal has dominated the scientific culture for more than two centuries and is personified in Nobel laureates [9].

In the present regressive drive to convergence, science of science researchers point to a significant difference with times past. It is teams of collaborating specialists [44], they argue, that serve as vehicles of contemporary convergence [3], rather than polymath scientists who mastered the broader scientific corpus. To put it differently, contemporary convergence is thought to take place among scientists rather than within scientists. The underlying behavioral factors for this type of integration may partly explain contemporary convergence's rocky path to culmination [42].

Starting in the middle of the 20th century, several attempts to convergence took place, some institutionally driven while others at the grassroots level. Successes were infrequent but transformative. The Manhattan Project (1940s), where physicists, chemists, and engineers successfully worked their way to control nuclear fission and produce the first atomic bomb, demonstrated contemporary convergence's powerful potential [20]. Several decades later (1990s-2000s), the Human Genome Project (HGP) forged a collaborative bond between biologists and computer scientists in

the context of consortium science [32]. In 10 short years, HGP led to the decoding of the human genome, ushering civilization into the genomics era.

In the 2010s, the research frontier moved yet again, traversing this time the brain science domain [37]. Brain science is naturally primed for convergence. The brain question is incredibly complex, it touches upon a rainbow of disciplines, and its solution will be a watershed moment for health and behavioral applications [11]. Brain science has also been supported by major funding programs, designed to encourage cross-disciplinary collaboration.

Unlike genomics, where most of the foundational research was centered around HGP, a U.S. centric program, the brain funding programs span over the world [16]. In late 2013, the United States launched the BRAIN Initiative® (Brain Research through Advancing Innovative Neurotechnologies), a public-private effort aimed at developing new experimental tools that will unlock the inner workings of brain circuits [23]. At the same time, the European Union launched the Human Brain Project (HBP), a 10 year funding program, based on exascale computers, that aims to build a collaborative infrastructure for advancing knowledge in the fields of neuroscience, brain medicine, and computing [2]. In 2014, Japan launched the Brain Mapping by Integrated Neurotechnologies for Disease Studies (Brain/MINDS), a program to develop innovative technologies for elucidating primate neural circuit functions [28]. China followed in 2016 with the China Brain Project (CBP), a 15 year program targeting the neural basis of human cognition [36]. Canada [21], South Korea [22], and Australia [5] followed suit, launching their own brain programs in the late 2010s.

The global character of the brain science community, its funding momentum, and the multi-disciplinary/multi-domain nature of its challenging goal, render brain research not only a science frontier but also a 'live experiment' in convergence evolution. Accordingly, we focus on analyzing the brain science ecosystem, as a way to capture the 'pulse' of convergence in our time. This effort follows on the heels of our work on formative patterns of convergence in genomics [32], thus offering rich opportunities for evolutionary comparisons.

Uniquely, we operationalize convergence by differentiating between the disciplinary composition

of a paper's author team and the epistemic domains it spans. We refer to the former as team cross-disciplinarity, while to the latter as cross-domain diversity. This dual consideration affords us an insightful look into the deeper mechanisms of contemporary convergence. Specifically, the said method allows us to investigate how the 'specialist' ideal holds up in the era of team science. Do specialists start naturally gravitating towards a more polymathic model? If yes, how does this affect their teaming behaviors? Are scholars from discipline A who acquire some expertise in discipline B still inclined to collaborate with experts in discipline B, or do they tend to go it alone? What are the effects of such trends in addressing grand scientific challenges, starting with the brain question?

Overall, our analysis operates on the unit of research production, that is, the paper, and unfolds along three principal dimensions:

**Convergence,** differentiating between author team cross-disciplinarity and cross-domain diversity.

**Temporal,** differentiating between the early part of 2010s, prior to the launch of the brain programs, and the latter part of 2010s (post-launch).

**Geographic,** differentiating among North America, Europe, and Australasia - the three epicenters of brain research activity.

The analytic tools brought to bear include co-occurrence matrices, network models, and econometric panel methods.

The results paint a revealing picture. Brain scientists are becoming more polymathic, especially in the context of mono-disciplinary teams, or cross-disciplinary teams from epistemically neighboring areas. In these cases, brain scientists tend to exhibit a more expansive approach to research, incorporating domains within and beyond their core expertise at the cost of decreasing impact. Team science has awkwardly met Renaissance science in what appears to be a grassroots movement, unintentionally reinforced by the brain funding initiatives.

## 2 Methods

We constructed a relatively comprehensive scholar-centric representation of the Human Brain science ecosystem by merging data from three publication indices – Web of Science, Scopus, and PubMed (see **Figure 1**).

### 2.1 Author Keystone via Web of Science (WOS)

In building a scholar-centric database, one needs a keystone for developing a list of authors who have published on a particular topic. For that, we chose WOS, using the topic field query 'Human Brain' (HB) to search its 'Core Collection' over the period 1955-2016. This search resulted in 224,201 records with distinct WOS article identifiers. Out of the said records, we extracted the full first and last names of ∼14,725 authors with $\geq 5$ publications in this sample, along with their affiliations.

### 2.2 Author Name Disambiguation

An important challenge in constructing a scholar-centric dataset is name disambiguation of authors. Here we overcome this challenge by using curated publication sets for each scholar obtained from Scopus via their profile-oriented API, which requires an author's full name and affiliation in order to identify their Scopus profile.

### 2.3 Brain Science data via Scopus and PubMed

Having amassed a comprehensive set of brain science researchers, the next step was to build a database that represents the totality of their work. Since brain science is multidisciplinary, these researchers come from different domains, publishing in diverse areas that feed brain science, thus creating an ecosystem prime for cross-domain analysis. We used the full name and affiliation-location of each WOS author to query the Scopus Author Profile repository. Among these profiles, 9,265 contained geographic and departmental affiliation information. In order to identify HB research articles, as opposed to other content such as comments and editorials and also non-biomedical

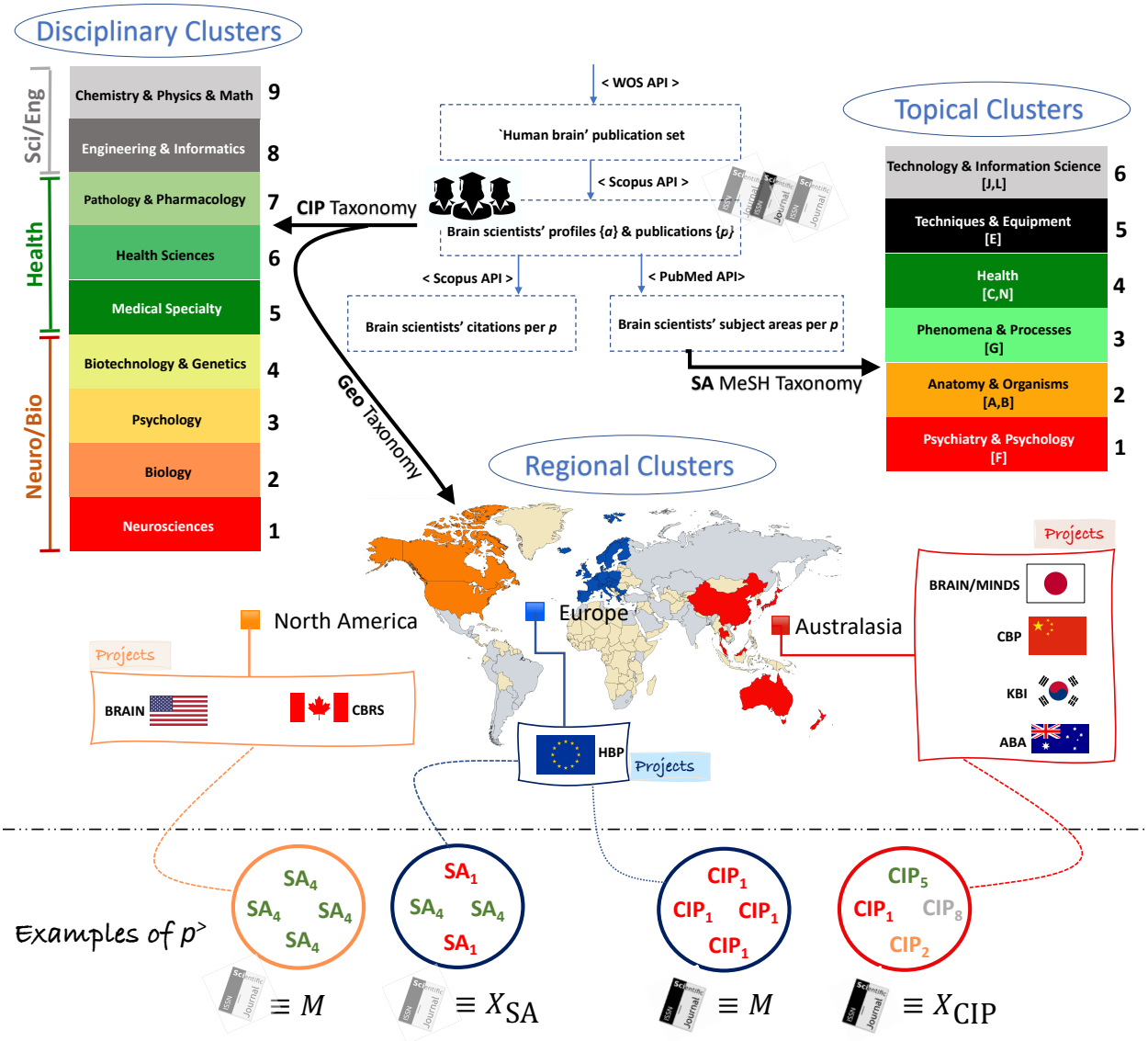



Figure 1: **Data collection and classification schemes.** The upper part of the figure shows the data generation mechanism along with the resulting topical (SA) and disciplinary (CIP) clusters. The middle part of the figure shows on the world map regional clusters pertaining to three large HB funding initiatives – North America (NA), Europe (EU), and Australasia (AA). The lower part of the figure shows an example of how all three categorizations are operationalized for analytic purposes. Circles represent four research articles with authorship from distinct regions. The articles feature different keyword (SA) or disciplinary (CIP) category mixtures assigned one of two diversity measures: mono- ($M$) and cross-domain ($X$).

research in the physical sciences, we searched for each article DOI in MEDLINE/PubMed. We only analyzed articles annotated with Medical Subject Heading (MeSH) keywords, which are indicators that this research-oriented content is biomedical in nature – resulting in a HB dataset with $0.98\times10^6$ articles over the period 1945-2018; see **Figure A2**(A) for $N(t)$. For each research article $p$ published in year $t$, we also obtained the number of Scopus citations $c_{p,t}$ tallied through the API download (census) date in November 2019.

### 2.4 Topical Keyword Classification using MeSH

Medical Subject Headings (MeSH) are a quasi-hierarchical biomedical ontology developed by the National Library of Medicine and implemented across articles indexed by PubMed by expert annotators to classify articles according to their topical and methodological contributions. With on average 12 MeSH per article, this ontology facilitates topic mapping and topic co-occurence analysis at multiple levels of specificity [34]. We restrict our analysis to only the 'Major Topic Heading' MeSH, which are indicated in PubMed by an asterisk, and account for roughly 1 in 3 MeSH descriptors. As such, we use these publication-level MeSH to determine the topical subject area of the research reported in each article. In total we encountered 14,212 distinct Major Topic MeSH.

### 2.5 Identifying Subject Area Clusters in the Brain Science Ecosystem

Each MeSH descriptor has a tree number that identifies its location within one of 16 broad categorical branches. We merged 9 of the science-oriented MeSH branches (A,B,C,E,F,G,J,L,N) into 6 Subject Area ($SA$) clusters (see **Figure 1**). **Figure 2** shows the 50 most prominent MeSH descriptors for each SA cluster. Hence, we take the set of MeSH for each $p$ denoted by $\vec{W}_p$, and map these MeSH to the corresponding MeSH branch (represented by the operator $O_{SA}$), yielding a count vector with six elements: $O_{SA}(\vec{W}_p) = \overrightarrow{SA}_p$. **Figure A2**(D) shows the distribution $P(N_{SA})$ of the number of $SA$ per publication: 72% of articles have two or more $SA$; the mean (median) $SA_p$ is 2.1 (2), with standard deviation 0.97, and maximum 6.

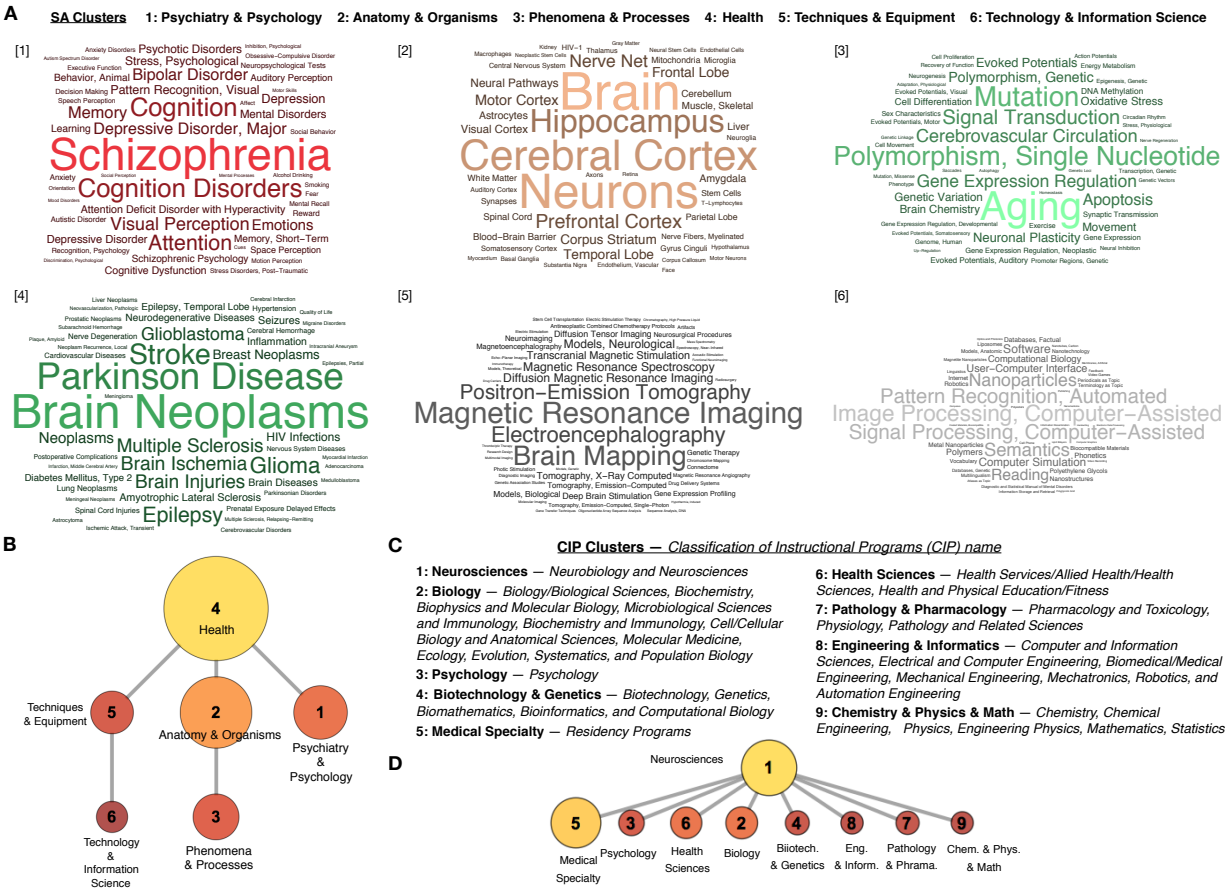



Figure 2: **Subject Area and Department clusters.** (A) Principal MeSH terms comprising 6 Subject Area (SA) clusters. (B) Minimum spanning tree representation of topical hierarchy based upon SA co-occurrence within articles; node size proportional to total number of articles featuring a particular SA. (C) Department CIP codes comprising 9 disciplinary clusters. (D) Minimum spanning tree representation of disciplinary hierarchy based upon CIP co-occurrence within articles; node size proportional to total number of articles featuring a particular CIP.

### 2.6 Geographic Regions

We obtained geographic location data from each scholar's Scopus Profile, associating each individual with one of 77 countries; the top five countries represented are the United States with 5030 scholars, Germany with 1192, UK with 1074, China with 1049, and Japan with 894. These coauthors associate each $p$ with a set of countries, which we cluster into four localized regions indexed by $R$: North America, corresponding to $R = 1$ (United States and Canada); Europe, $R = 2$ (33 European Union and non-European Union countries including Norway, Switzerland, Israel, Iceland, and Serbia); Australasia, $R = 3$ (Peoples Republic of China, Japan, South Korea, Australia, Taiwan, New Zealand, Singapore, Malaysia, and Thailand); and World, $R = 4$ (remaining countries including Brazil, India, Turkey, and South Africa, among others). 88% of the publications are covered by regional clusters $R = 1, 2, 3$.

### 2.7 Disciplinary Classification using CIP

We obtained host department information from each scholar's Scopus Profile. Based upon this information provided in the profile description, and in some cases using additional web search and data contained in the Scholar Plot web app [25], we manually annotated each scholar's home department name according to National Center for Education Statistics Classification of Instructional Program (CIP) codes. We then merged these CIP codes into 9 broad clusters and three super-clusters (Neuro/Biology, Health, and Science & Engineering, as indicated in **Figure 1**); for a list of constituent CIP codes for each cluster see **Figure 2**(C). Analogous to the notation for assigning $\overrightarrow{SA}_p$, we take the set of authors for each $p$ denoted by $\vec{A}_p$, and map their individual departmental affiliations to the corresponding CIP cluster (represented by the operator $O_{CIP}$), yielding a count vector with nine elements: $O_{CIP}(\vec{A}_p) = \overrightarrow{CIP}_p$.

### 2.8 Measuring Cross-domain Diversity with Categorical Co-occurrence

We calculate a measure of cross-domain co-occurrence using the vector $\vec{v}_p$ of category counts for a given publication $p$: for cross-disciplinary co-occurrence $\vec{v}_p \equiv \overrightarrow{CIP}_p$ and for cross-topic co-occurrence $\vec{v}_p \equiv \overrightarrow{SA}_p$. We measure article co-occurrence levels by way of the normalized outer-product

$$\mathbf{D}_p(\vec{v}_p) \equiv \frac{U(\vec{v}_p \otimes \vec{v}_p)}{||U(\vec{v}_p \otimes \vec{v}_p)||} \ , \tag{1}$$

where $\otimes$ is the outer tensor product, $U(\mathbf{G})$ is an operator yielding the upper-diagonal elements of the matrix $\mathbf{G}$ (i.e., representing the undirected co-occurrence network among the categorical elements), and $||...||$ indicates the matrix normalization implemented by summing all matrix elements. In essence, $\mathbf{D}_p(\vec{v}_p)$ captures a weighted combination of all category pairs. The objective of this normalization scheme is to control for the variation in $\vec{v}_p$ in a systematic way. As such, this co-occurrence is a article-level measure of diversity which controls for variations in the total number of categories and different count statistics for elements belonging to $\overrightarrow{CIP}_p$ and $\overrightarrow{SA}_p$. Consequently, totaling $\mathbf{D}_p(\vec{v}_p)$ across articles from a given publication year yields the total number of articles published in a given year, $\sum_{p|y_p \in t} ||\mathbf{D}_{p,t}|| = N(t)$.

We also define a categorical diversity measure for each article given by $f_{D,p} = 1 - \mathrm{Tr}(\mathbf{D}_p) \in [0, 1)$, which corresponds to the sum of the off-diagonal elements in $\mathbf{D}$. The average article diversity by publication year is denoted by $\langle f_D(t) \rangle$. In simple terms, articles featuring a single category have $f_{D,p} = 0$ whereas articles featuring multiple categories have $f_{D,p} > 0$.

### 2.9 Bi-partite Network between CIP and SA

We quantify the empirical association between CIP and SA categories by aggregating the information contained in $\overrightarrow{CIP}_p$ and $\overrightarrow{SA}_p$. We first applied this method to the subset of mono-domain articles comprised of $p$ with $O_{CIP}(\vec{F}_p) = O_{SA}(\vec{F}_p) = M$. By definition, each of these article features just a single CIP, making it possible to identify the SA that are most frequently associated

with mono-domain researchers from that CIP category. Formally, this amounts to calculating the bi-partite network between CIP and SA, operationalized by averaging the $\overrightarrow{SA}_p$ for mono-domain articles from each CIP category, given by $\langle\overrightarrow{SA}\rangle_{CIP} = \sum_{p\in CIP}(\overrightarrow{SA}_p/N_{SA,p})$; importantly, this definition accounts for variability in $N_{SA}$ by normalizing the sum of the SA counts contained in $\overrightarrow{SA}_p$ by $N_{SA,p}$ so that each article contributes equally to the average. Less prominent CIP-SA links are pruned from our Sankey chart visualization in order to emphasize the most meaningful CIP-SA relations. To this end, we remove the weakest links contained in $\langle\overrightarrow{SA}\rangle_{CIP}$, excluding those with value $< 0.5$ $Max[\langle\overrightarrow{SA}\rangle_{CIP}]$. Hence, the chart labelled $M_{CIP} \rightleftarrows M_{SA}$ in **Figure 3**(G) shows only the most prominent CIP-SA links.

For juxtaposition, we also calculated the bi-partite network using the non-overlapping subset of articles with $O_{SA\&CIP}(\vec{F}_p) = X_{SA\&CIP}$. Since these articles by construction have $N_{CIP,p} \geq 2$, we define the average association between CIP and SA as $\langle\overrightarrow{SA}\rangle_{CIP} = \sum_{p\in CIP}(\overrightarrow{SA}_p/N_{SA,p})/N_{CIP,p}$, where the vector $\overrightarrow{SA}_p/N_{SA,p}$ contributes to the average for all CIP present in $\overrightarrow{CIP}_p$. The bi-partite network labeled $X_{CIP} \rightleftarrows X_{SA}$ in **Figure 3**(G) also shows just the most prominent CIP-SA links, applying the same threshold that excludes links that have weight less than half of the most prominent weighted CIP-SA link for a given CIP.

Let **A** (**B**) represent the matrix representation of $X_{CIP} \rightleftarrows X_{SA}$ ($M_{CIP} \rightleftarrows M_{SA}$) – after pruning less prominent CIP-SA links. We then compute the difference between the matrices, $\mathbf{\Delta_{XM}} \equiv \mathbf{C} = \mathbf{A} - \mathbf{B}$, such that positive (negative) elements of **C** indicate prominent links that are relatively over-represented in cross-domain (mono-domain) articles. The Sankey chart labeled $\Delta_{XM}$ in **Figure 3**(G) shows just the positive elements, which tend to be larger in magnitude than the (relatively few) negative elements.

### 2.10 Normalization of Citation Impact

We normalize the Scopus citation count $c_{p,t}$ for each publication by leveraging the well-known log-normal properties of citation distributions [38]. To be specific, we disaggregate the articles by publication year $y_p$, and apply a normalization method that removes the time-dependent trend in

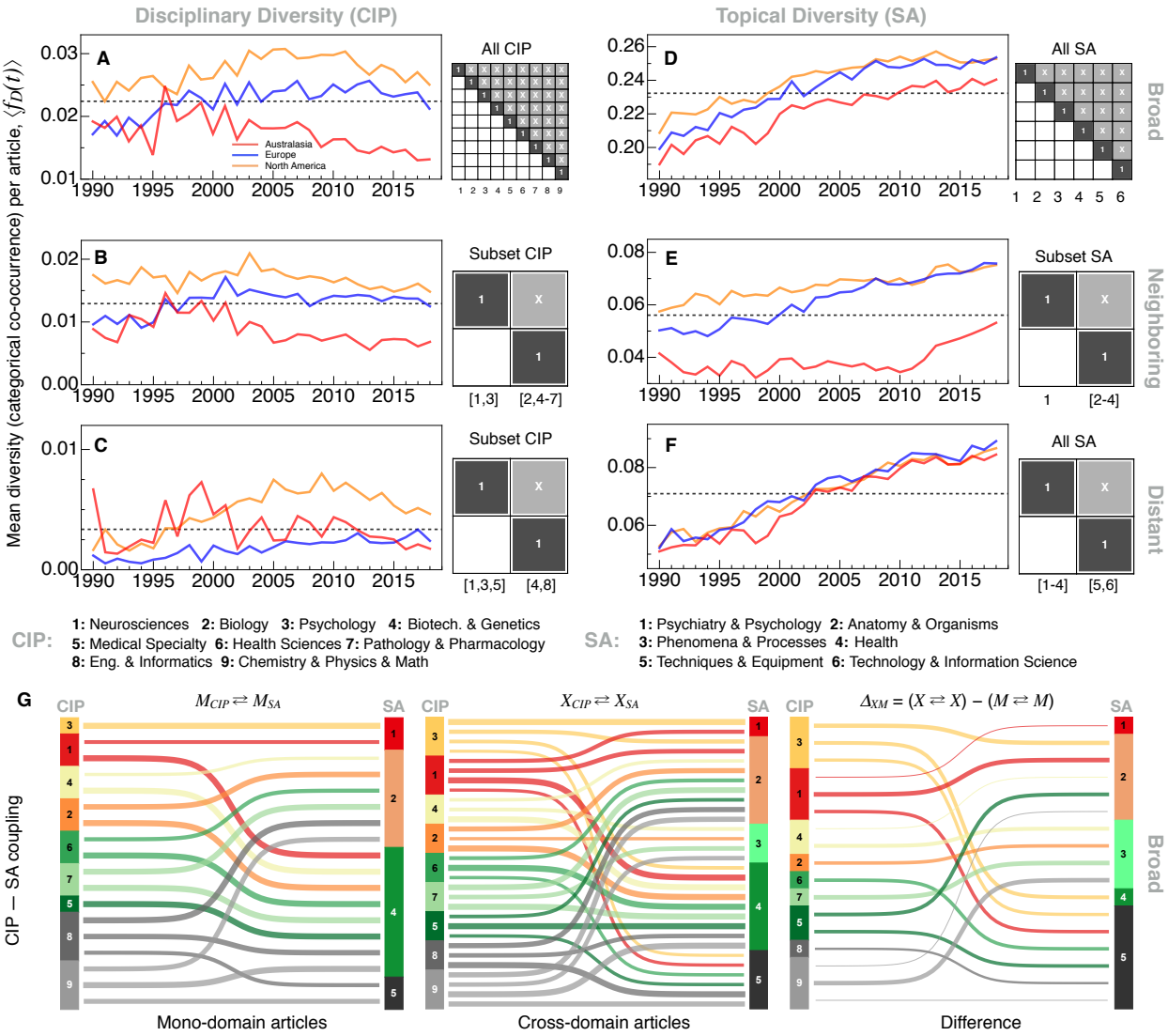



Figure 3: **Evolution of boundary-crossing research in human brain science.** (A-F) Each $\langle f_D(t) \rangle$ represents the the average article diversity measured as categorical co-occurrence, by geographic region: Australasia (red), Europe (blue), and North America (orange). Each matrix motif indicates the set of CIP or SA categories used to define $\mathbf{D}_p$ defined in Eq. (1); categories included in brackets are considered in union. For example, panel (A) calculates $\langle f_{D,CIP}(t) \rangle$ across all 9 CIP categories; instead, panel (B) is based upon counts for two super-groups, the first consisting of the union of CIP counts for categories 1 and 3, and the second comprised of categories 2, 4, 5, 6 and 7. (A,D) *Broad* diversity calculated using all categories considered as separate domains; (B,E) *Neighboring* represents shorter-distance convergence across the neurobiological ↔ bioengineering interface; (C,F) *Distant* represents longer-distance convergence across the neuro-psycho-medical ↔ techno-computational interface; (G) Empirical CIP-SA association networks calculated for non-overlapping sets of mono-domain ($M_{CIP} \rightleftarrows M_{SA}$) and cross-domain ($X_{CIP} \rightleftarrows X_{SA}$) articles, based upon the *Broad* configuration. The difference between these two bi-partite networks ($\Delta_{XM}$) indicates the research channels that are facilitated by simultaneous $X_{CIP}$ and $X_{SA}$.

the location and scale of the underlying log-normal citation distribution. The normalized citation value for a publication $p$ from year $y_p = t$ that has $c_{p,t}$ Scopus citations is given by

$$z_p = (\ln(c_{p,t} + 1) - \mu_t)/\sigma_t \ , \tag{2}$$

where $\mu_t \equiv \langle \ln(c_t + 1) \rangle$ is the mean and $\sigma_t \equiv \sigma[\ln(c_t + 1)]$ is the standard deviation of the citation distribution for a given $t$; we add 1 to $c_{p,t}$ to avoid the divergence of $\ln 0$ associated with uncited publications – a common method which does not alter the interpretation of results.

**Figure A2**(G) show the probability distribution $P(z_p)$ calculated across all $p$ within five-year non-overlapping time periods. The resulting normalized citation measure is well-fit by the Normal $N(0, 1)$ distribution, independent of $t$, and thus is a stationary measure across time. Publications with $z_p > 0$ are thus above the average log citation impact $\mu_t$, and since they are measured in units of standard deviation $\sigma_t$, standard intuition and statistics of $z$-scores apply. The annual $\sigma_t$ value is rather stable across time, with average and standard deviation $\langle \sigma \rangle \pm \mathrm{SD} = 1.24 \pm 0.09$ over the 49-year period 1970-2018.

### 2.11 Abstract Collection

We used the full name and affiliation-location of each WOS author to query in Scopus application programming interfaces (APIs). Among these profiles, 9,265 contained geographic and departmental affiliation information. Based on unique DOI (digital object identifier) article identifiers, these scholars produced $1.4 \times 10^6$ distinct articles. Parts of articles were removed if they dont have abstracts. Also, parts of abstracts(or full abstracts) that were not in english were removed using text search. Finally we got total $1.07 \times 10^6$ distinct articles.

### 2.12 Data Pre-processing for Abstracts

Initially we combined article titles and abstracts together. Then we did data cleaning process. In the first step, we removed control characters from the texts (For example: newline character, b, x0c

etc). Then we removed all the diction characters from the texts (For example: parenthesis, brackets, curly brackets, comma, semicolon etc). In the next step, we removed all numeric numbers, digits, and non-english words from texts. Then we did word tokenization from texts using the python nltk library. Then we removed all physical units. For getting the best output, we removed all the english stop words from texts (For example: I, we, my, our, your, her etc). To handle uppercase and lowercase words, we converted all the words into lower case words. Some abstracts contained metadata words and we removed them from our corpus. We also created a custom article stopwords list. Almost all the articles contained these kind of stopwords. **Table 1** contains all the custom stopwords. We removed these custom stop words from our tokens (see **Figure A10**). Finally, we converted word tokens into stem tokens using the popular Snowball stemmer. This stemmer reduced different forms of words to its core root. For example, the word report, reporting, and reports are from the same root word report. Using Snowball stemmer we minimize the number of word tokens from our corpus.

### 2.13 Handling Phrases

Generally, we provide word tokens into the corpus and corpus knows only words. So, for handling phrases which contain multiple words, we need to combine multiple words together. For combining multiple words together and create a single word, we used symbol underscore(_) between words. For example: we replaced gene expression, and magnetic resonance using gene_expression, and magnetic_resonance. In this way corpus considers phrases as single words.

### 2.14 Handling Plural Words

Handling plural words is a challenging task. We can use nltk stemmer tool to solve this problem. But this stemmer is not working properly with phrases. That's why we converted plural words into singular words using manual annotation. For example, we replaced neural_networks, and gene_expressions using neural_network, and gene_expression etc.

Table 1: Custom stop words

| Custom Stop Words | | | |
|---|---|---|---|
| abstract | title | method | result |
| discussion | conclusion | doi | preprint |
| copyright | peer | reviewed | org |
| https | et | al | author |
| figure | rights | reserved | permission |
| used | using | biorxiv | medrxiv |
| license | fig | fig. | al. |
| Elsevier | PMC | CZI | www |
| announcement | bookreview | erratum | editorialnotes |
| news | article | events | acknowledgement |
| foreword | prelude | commentary | workshop |
| conference | symposium | comment | retract |
| correction | memorial | report | case |
| year | present | associ | studi |
| follow | group | result | subject |
| relat | chang | xce | day |
| new | american | societi | elsevi |
| scienc | springer | busi | media |
| verlag | wiley | liss | berlin |
| heidelberg | john | son | oxford |
| univers | press' | univers | analysi |
| reveal | data | indic | suggest |
| decis | make | import | role |
| patient | underw | signific | decreas |
| differ | higher | increas | improv |
| lower | reduc | materi | method |
| william | wilkin | bbb | gmbh |
| kgaa | weinheim | lippincott | macmillan |
| publish | unrestrict | distribut | vch |
| wolter | kluwer | license | author |
| mdpi | biom | basel | inf |
| limit | trade | taylor | public |
| switzerland | vch | taken | togeth |
| mari | ann | liebert | |

### 2.15 Article Classification

In our analysis, we classify articles using pre-processed abstracts. We used K-means unsupervised machine learning algorithm to classify articles, where each document is described by a term frequency–inverse document frequency (TF–IDF) vector. For handling phrases we used 1-gram and 2-gram in TF-IDF and selected top 200 tokens based on TF-IDF score. Then we pass the vector into the principle component analysis (PCA) to reduce the dimensions. Based on PCA output we applied K-means algorithm to classify all the articles. Cluster outputs are represented using word cloud in the **Figure A7**.

### 2.16 Represent Brain Parts Based on Articles

We want to investigate how all the brain parts are located within articles. We used top 20 human brain parts for our analysis. **Table 2** contains the name of the top 20 human brain parts. First, we created a text corpus using our pre-processed abstracts. Second, we used skip-gram of Word2Vec model for word embedding.

Table 2: Brain parts

| Brain Parts | |
|---|---|
| Thalamus | Frontal Lobe |
| Motor Cortex | Cerebellum |
| Visual Cortex | Hippocampus |
| Cerebral Cortex | Auditory Cortex |
| Amygdala | Prefrontal Cortex |
| Parietal Lobe | Corpus Striatum |
| Somatosensory Cortex | Basal Ganglia |
| Substantia Nigra | Temporal Lobe |
| Gyrus Cinguli | Corpus Callosum |
| Hypothalamus | Pons |

We used the Word2vec implementation in gensim (https://radimrehurek.com/gensim/) with a few modifications. The vocabulary consisted of all words that occurred more than ten times. The rest of the hyperparameters were as follows: we used 300-dimensional embeddings, window size = 8, and 250 iterations for Word2Vec model. Finally, we used t-distributed stochastic neighbor

embedding (t-SNE) for showing the embedding of brain parts in a 2D representation. **Figure 4** represents the 2D plot of brain parts. We used the cosine distance between the embeddings as a metric, perplexity 40, and 250 iterations - with coordinates initialized using Principal Component Analysis (PCA).

### 2.17 Cosine Similarity Words of Gene Expression and Magnetic Resonance

For finding the top 10 cosine similar words of gene expression and magnetic resonance imaging we used temporal analysis using a skip-gram of Word2Vec model for word embedding. We used the year range from 1965 to 2018. For each year we crated corpus and applied the Word2Vec model. We used the Word2vec implementation in gensim (https://radimrehurek.com/gensim/) with a few modifications. We set the hyperparameters as follows: we used 300-dimensional embeddings, window size = 8, min-count = 1, and 250 iterations for Word2Vec model. Then we find the top 10 cosine similar words for each year (see **Figure A11** for analysis details). **Figure A8** and **Figure A9** represents the temporal word cloud for top cosine similar words of gene expression and magnetic resonance.

## 3 Results

### 3.1 Human Brain Science – Data Collection & Methods Summary

We constructed a comprehensive dataset on Human Brain Science to facilitate measuring factors relating to cross-domain diversity and shifts in this activity associated with four global HB flagship projects conceived circa 2013 – (i) the *European Commission Future and Emerging Technologies Flagship* Human Brain Project started officially in late 2013; (ii) and the BRAIN Initiative (Brain Research through Advancing Innovative Neurotechnologies) was announced in Spring 2013 and established a project timeline by 2014; (iii) Japanese the Brain Mapping by Integrated Neurotechnologies for Disease Studies project commenced in 2014; and (iv) the China Brain Project established in 2016; similar brain projects in South Korea, Australia, and Canada have since emerged.

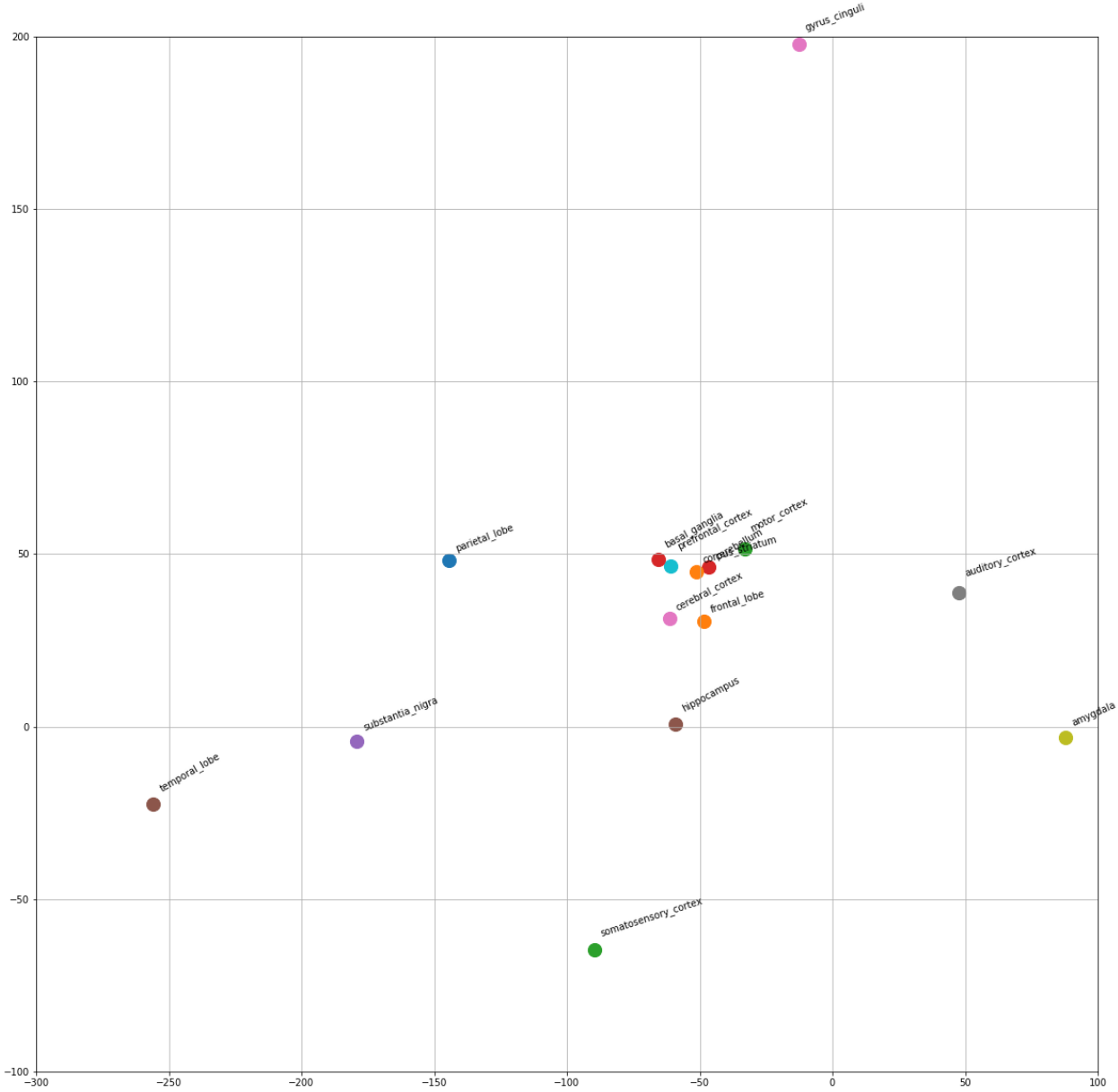



Figure 4: **Brain parts 2D representation.**

**Figure 1** shows the multiple sources combined in our study, which integrates publication and author data from Scopus, PubMed and the Scholar Plot web app [25] (see Methods for a detailed procedural description). In total our data sample consists of roughly $10^6$ articles from 1945-2018, to which we apply the following variable definitions and subscript conventions to capture both article- and scholar-level information. At the article level, subscript $p$ indicates publication-level information such as publication year, $y_p$; the number of coauthors, $k_p$; and the number of keywords, $w_p$. Regarding the temporal dimension, a superscript $>$ (respectively, $<$) indicates data belonging to the 5-year "post" period 2014-2018 (5-year "pre" period 2009-2013), while $N(t)$ represents the total number of articles published in year $t$. Regarding proxies for scientific impact, we obtained the number of citations $c_{p,t}$ from Scopus, which are counted through the data census year corresponds to late 2019; since nominal citation counts suffer from systematic temporal bias, we use a normalized citations measure, denote by $z_p$ (see Methods 'Normalization of Citation Impact' in section 2.10). Regarding author-level information, we use the index $a$, such as the scholar age measured in years since her first publication, $\tau_{a,p}$, which has both subscripts since it depends on $a$ and $y_p$.

**Figure 1** illustrates how each article is classified by three category systems indicative of topical, disciplinary and regional clusters. The first category system captures research topic clusters grouped into Subject Areas (SA); counts for each article are represented by a vector with 6 elements, $\overrightarrow{SA}_p$, each corresponding to top-level Medical Subject Heading (MeSH) categories implemented by PubMed [and indicated by the letters in brackets next to the category titles]: (1) Psychiatry & Psychology [F], (2) Anatomy & Organisms [A,B], (3) Phenomena & Processes [G], (4) Health [C,N], (5) Techniques & Equipment [E], and (6) Technology & Information Science [J,L]. The variable $N_{SA,p}$ counts the total number of SA categories present in a given article, with min value 1 and max value 6. The second taxonomy identifies disciplinary clusters determined by author departmental affiliation information categorized according to *Classification of Instructional Program* (CIP) codes. Article-level CIP category counts are represented by $\overrightarrow{CIP}_p$, with 9 elements pertaining to the following categories: (1) Neurosciences, (2) Biology, (3) Psychology, (4) Biotechnology & Genetics, (5) Medical Specialty, (6) Health Sciences, (7) Pathology & Pharmacology, (8) Engineering &

Informatics, and (9) Chemistry & Physics & Math. The variable $N_{CIP,p}$ counts the total number of CIP categories, with min value 1 and max value 9.

The third taxonomy captures the broad regional scope of each research article team determined by each Scopus author's university location, and represented by the vector $\overrightarrow{R}_p$ which has 4 elements representing Australasia, Europe, North America, and rest of World; the variable $N_{R,p}$ counts the total number of different regions represented by a given article, with minimum value 1 and maximum value 4. See **Figure 2** for the composition of CIP and SA clusters, and *Methods* for additional description of how these classification systems are constructed. **Figures 5** and **6** show the frequency of each CIP category (SA) and the pairwise frequency of all CIP-CIP (SA-SA) combinations over the 10-year period centered on 2014, along with their relative changes after 2014; See *Appendix A* for discussion of the relevant changes in SA and CIP categories after 2014.

We represent the collection of article features by $\vec{F}_p \equiv \{\overrightarrow{SA}_p, \overrightarrow{CIP}_p, \overrightarrow{R}_p\}$. As indicated in **Figure 1**, based upon the distribution of types occurring for each article that are represented as counts within each category vector, an article is either cross-domain – representing a diverse mixture of types denoted by $X$ – or mono-domain – denoted by $M$. We use a generic operator notation to specify how articles are classified as $X$ or $M$, The objective criteria of the feature operator $O$ is specified by its subscript: for example $O_{SA}(\vec{F}_p)$ yields one of two values – $X_{SA}$ or $M$; similarly, $O_{CIP}(\vec{F}_p) = X_{CIP}$ or $M$. Note that all scholars map onto a single CIP, and so solo-authored research articles are by definition classified by $O_{CIP}$ as $M$. We also classify articles featuring both $X_{SA}$ and $X_{CIP}$ as $O_{SA\&CIP}(\vec{F}_p) = X_{SA\&CIP}$ (and otherwise $M$).

### 3.2 Increasing Prevalence of Cross-domain Science

**Figure 7**(A) shows the frequencies of mono-domain ($M$) research articles versus cross-domain articles ($X$) in our HB sample. Articles were separated into above- and below-average citation impact ($z$) for each publication-year cohort ($t$), and within each of these two subsets we calculated the fraction $f_{\#}(t|z)$ of articles containing combinations across # =1,2,3 and 4 categories. The fraction of mono-domain articles is trending downward, which we observe for both research topics

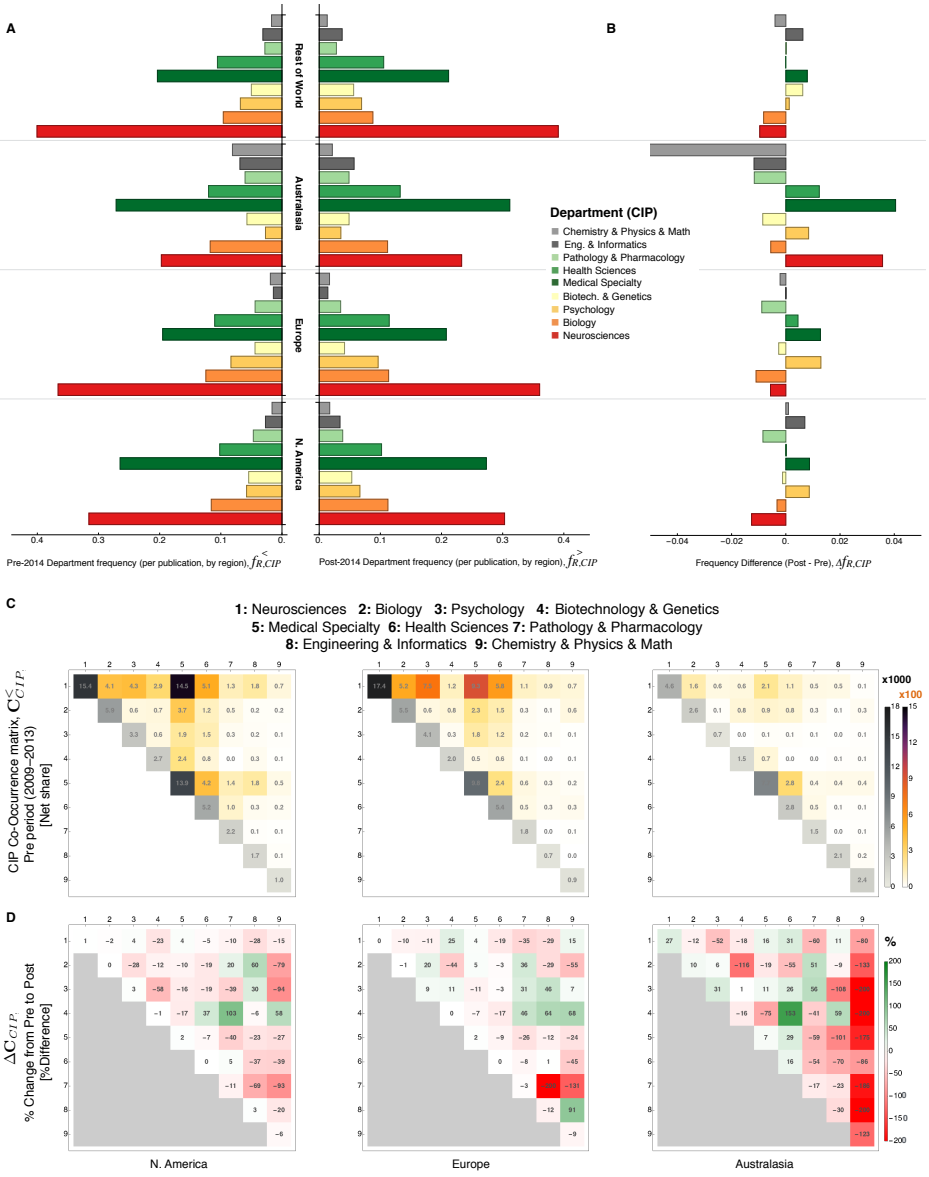



Figure 5: **Scholar departments (CIP) in human brain research in the 5-year period before and after 2014 – by geographic region.** (A) Relative frequency of department CIP clusters in the 5-year period before 2014 ($f^{<}_{R,CIP}$) and after 2014 ($f^{>}_{R,CIP}$); $f$ values are normalized to unity within region. (B) Shift in CIP cluster frequencies given by the difference $\Delta f_{R,CIP} = f^{>}_{R,CIP} - f^{<}_{R,CIP}$. (C) Each co-occurrence matrix $\mathbf{C}^{<}_{CIP}$ measures the frequency of a given CIP-CIP pair over the 5-year pre-period 2009-2013 ; see Eqn. (3) for its definition. Diagonal elements measure the frequency of publications featuring only a single $CIP$ category. Note the use of two legends, one for the mono-dimensional diagonal elements (gray-scale legend reported in units of 1000 publications) and one for off-diagonal elements (color-scale legend reported in units of 100 publications); as indicated by the legend scales, mono-CIP publications occur with significantly higher frequency than multi-CIP publications. (D) Relative change in the co-occurrence matrix: $\Delta C_{CIP,ij}$ measures the percent difference in the frequency of publications characterized by each $(CIP, CIP)$ pair; matrix elements $C^{>}_{CIP,ij}$ measure co-occurrences in the 5-year post-period 2014-2018.

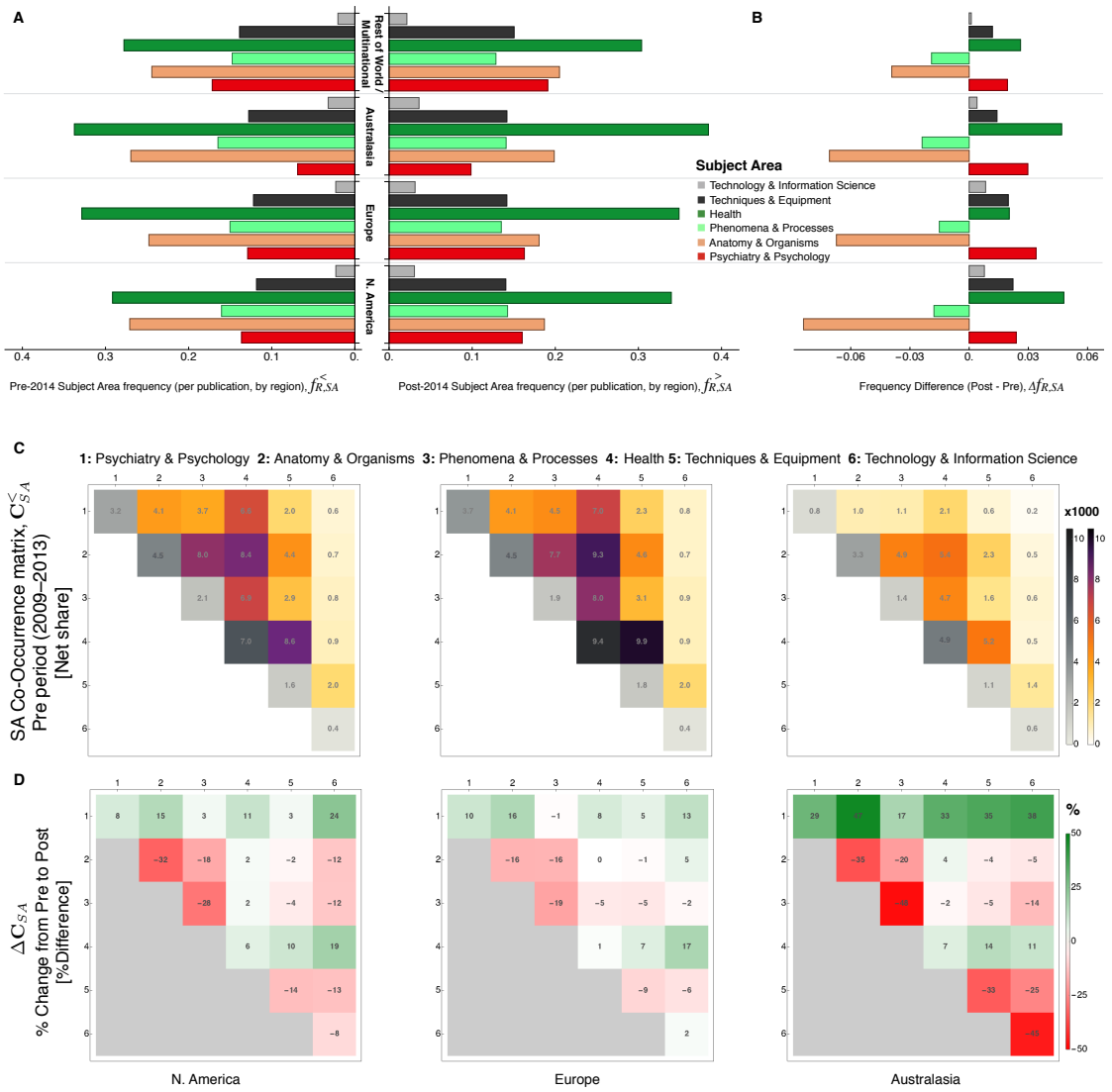



Figure 6: **Subject Areas (SA) in human brain research in the 5-year period before and after 2014 – by geographic region.** (A) Relative frequency of topical SA clusters in the 5-year period before 2014 ($f^{<}_{R,SA}$) and after 2014 ($f^{>}_{R,SA}$); $f$ values are normalized to unity within region. (B) Shift in SA cluster frequencies given by the difference $\Delta f_{R,SA} = f^{>}_{R,SA} - f^{<}_{R,SA}$. (C,D) Topical (SA-SA) co-occurrence in human brain science – by region. (C) Each co-occurrence matrix $\mathbf{C}^{<}_{SA}$ measures the frequency of a given SA-SA pair over the 5-year pre-period 2009-2013 based upon publications associated with one of three broad geographic regions; see Eqn. (3) for its definition. By construction, matrix element values $C^{<}_{SA,ij}$ are proportional to the net share of publications featuring the indicated pair. Diagonal elements measure the frequency of publications featuring only a single $SA$ category. Note the use of two legends, one for the mono-dimensional diagonal elements (gray-scale legend) and one for off-diagonal elements (color-scale legend), both of which are reported in units of 1000 publications. (D) Dynamic co-occurrence matrix, $\Delta C_{SA,ij}$, measuring the percent difference in the frequency of publications characterized by each $(SA, SA)$ pair; matrix elements $C^{>}_{SA,ij}$ measure co-occurrences in the 5-year post-period 2014-2018.

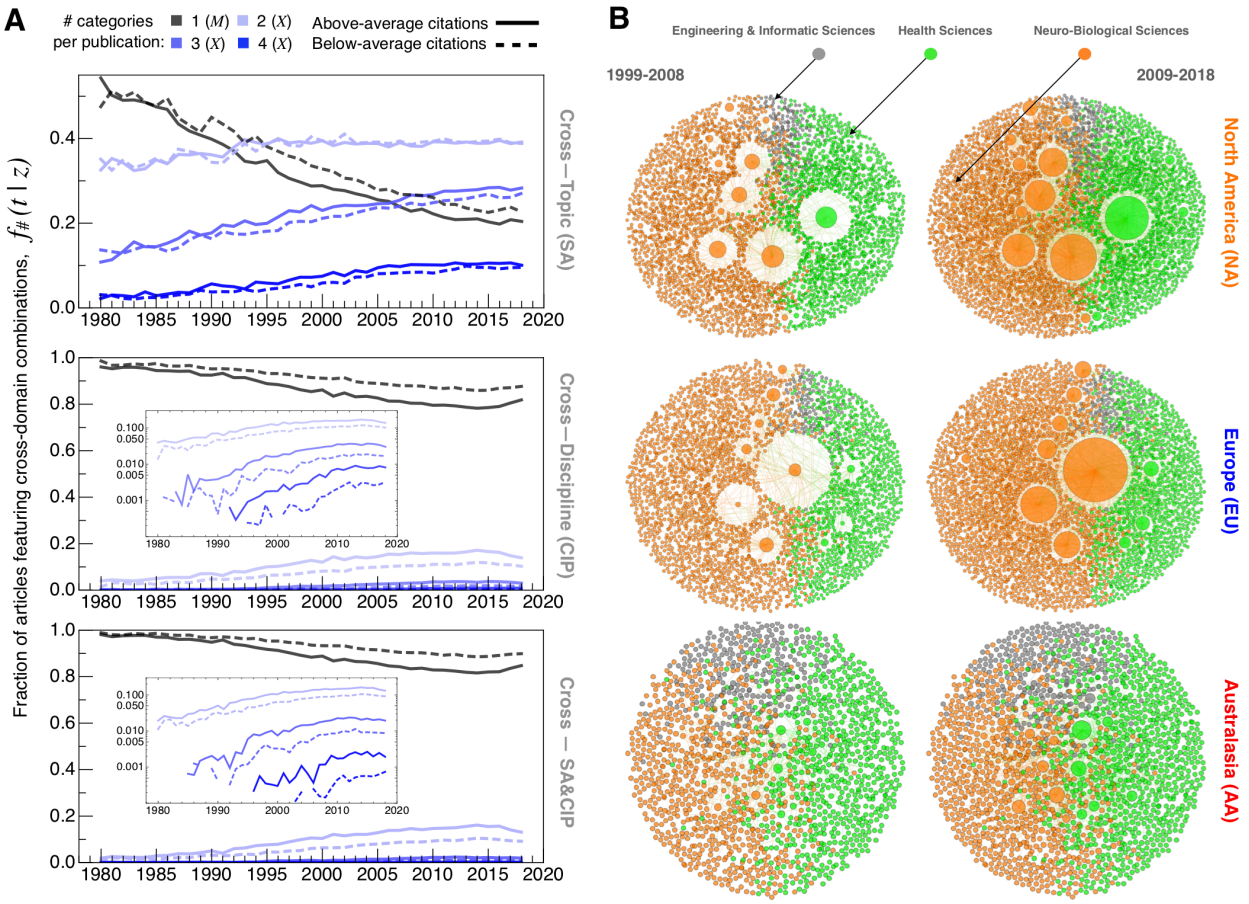



Figure 7: **Trends in cross-domain scholarship in Human Brain Science.** (A) Fraction $f_{\#}(t|z)$ of articles published each year $t$ that feature a particular number (#) of categories. They are split into an above-average citation subset ($z_p > 0$) and below-average citation subset ($z_p < 0$). Upper panel: Said categorization associated with SA. Middle panel: Said categorization associated with CIP; subpanel shows data on logarithmic y-axis; Lower panel: Said categorization associated with both SA and CIP simultaneously. Distinguishing frequencies by citation group indicates higher levels of cross-domain combinations among research articles with higher scientific impact – for both SA and CIP. However, levels of cross-domain activity are visibly higher for SA than for CIP, indicating higher barriers to boundary-crossing arising from mixing different scholar expertise. (B) Snapshots of the collaboration network at 10-year intervals indicating researcher population sizes by region, and the densification of convergence science at cross-disciplinary interfaces.

(SA) and authors' disciplinary affiliations (CIP). The decline, however, is much more precipitous in the SA realm than the CIP realm. Correspondingly, cross-domain articles have become increasingly prevalent, in particular for SA. For both SA and CIP the two-category mixtures dominate the three- and four-category mixtures in frequency in a sequence, and so in the sections that follow we do not distinguish between cross-domain articles with different #.

As a first indication of the comparative advantage associated with $X$, we observe a robust inequality $f_{\#}(t|z > 0) > f_{\#}(t|z < 0)$ for cross-domain research ($\# \geq 2$), meaning a higher frequency of cross-domain combinations observed among articles with higher impact. Contrariwise, in the case of mono-domain research the opposite phenomenon occurs, $f_1(t|z > 0) < f_1(t|z < 0)$. Taking into consideration temporal trends, these robust patterns indicate a faster depletion of impactful mono-domain articles, coincident with an increased prevalence of impactful research drawing upon cross-domain approaches.

### 3.3 Convergent Integration at Cross-disciplinary Interfaces

**Figure 7**(B) shows the population of HB researchers by region, represented as collaboration networks aggregated over 10-year intervals. Each node represents a researcher, colored according to the three broad departmental CIP groups: (i) neuro-biological sciences (corresponding to CIP 1-4), (ii) health sciences (CIP 5-7), and (iii) engineering & informatic sciences (CIP 8-9). Variable node size represents each researcher's link degree, counting the number of collaborators within the network in a particular time window. The location of nodes are fixed across each temporal snapshot, to facilitate the visual representation of the densification of the networks over time. Because the layout is determined by the underlying network structure, there is a high degree of clustering by node color, emphasizing not only the relative sizes of the subpopulations that are well-balanced across region and time, but also the convergent interfaces where cross-disciplinary collaboration is likely to catalyze. Consequently, links that span boundaries are fundamental conduits across which scientist's strategic affinity for interdisciplinary exploration [14] brings "together distinctive components of two or more disciplines" [27].

More efficient long-range exploration derived from multi-disciplinary teams of experts is a defining ingredient of convergence science [7], and contributes to the increased likelihood of high-impact research associated with team science [44]. As such, the emergence and densification of the interfaces between these three CIP groups capture the potential for recombinant innovation [12]. Borrowing from a triple-helix model of medical innovation [34], recombinant innovation in the present context of HB science can be conceptualized as sampling expertise from three interacting dimensions of supply, demand and technological capabilities: (i) the fundamental biology domain that supplies a theoretical understanding of the anatomical structure-function relation, (ii) the health domain that addresses the demand for effective science-based solutions, and (iii) the technological domain which develops scalable products, processes and services to facilitate (i) and (ii).

In order to overcome the challenges of selecting new strategies from the vast number of possible combinations, innovators are more likely to succeed when exploiting their own local expertise [12]. Extending upon this argument, exploration at unchartered interdisciplinary interfaces is likely to be more successful when integrating knowledge across a team of experts from different domains [13], thereby hedging against recombinant uncertainty underlying the exploration process. As such, these communities of expertise conjure the image of a Pólya urn, whereby successful combinations reinforce future combinations of similar configurations. A complementary argument for convergent problem solving draws on the advantage of diversity in harnessing collective intelligence to identify more successful hybrid strategies [29]. Recent work provides additional empirical support for the competitive advantage of diversity derived from cross-disciplinary collaboration [33] and cross-border mobility [31], where the latter study leverages the social capital disruption associated with researcher migration events to identify the positive role of research topic and collaborator diversity. Geographic diversity may also play an important role, as indicated by recent work reporting a marginal advantage associated with international diversity among collaborators [19], which is supported with analysis showing a reduced likelihood of novelty in research involving international collaboration [43].

### 3.4 Anatomy and Trends in Cross-domain Activity

In what follows we focus primarily on disciplinary and topical diversity, and use geographic variation to facilitate comparisons across regions. We explored the anatomy of cross-domain activity by tallying all CIP-CIP (SA-SA) category pairs present within each article to asses their relative frequencies.

The most notable results for CIP are the consistent strong coupling between Neuroscience [CIP 1] and other non-Sci/Eng. departments [2-6]. In particular we note a clique between Neurosciences, Medical Specialty and Health Sciences [CIP 1,5,6], indicative of health care being a leading sources of demand for HB science. Contrariwise, we observe relatively weak coupling across CIP 7-9. Comparing between regions, North America (NA) has relatively strong coupling between Neurosciences (Medical Specialty) and Biotechnology & Genetics [1,4] ([4,5]); and also between Medical Specialty and Engineering & Informations [5,8]. Europe (EU) has relatively strong coupling between Neurosciences and Psychology [1,3]. And Australasia (AA) has a relatively strong coupling between Neurosciences and Medical Specialty [1,5].

In the case of SA, we observe a consistent clique between topics at the interface of core biology and health [SA 2,3,4] in each region. Other SA exhibiting strong coupling are Psychiatry & Psychology and Health [1,4]; Health and Techniques & Equipment [4,5]; Anatomy & Organisms and Techniques & Equipment [2,5]; and Techniques & Equipment [2,5] with Technology & Information Science [5,6]. As with CIP, the technologically-oriented SA 5 and 6 represent the most weakly coupled HB topic domains.

All together, **Figures 5-6**(C) show the entire co-occurrence matrices $\mathbf{C}^{<}_{CIP}$ and $\mathbf{C}^{<}_{SA}$ over the 5-year pre-period 2009-2013. We focus on this period, indicated by superscript $<$, as it corresponds to the 5 years prior to the ramp-up of HB flagships across the globe. Each matrix element, e.g. $C^{<}_{SA,ij}$, represents the net share of all pairwise combinations corresponding to categories $i$ and $j$ using an appropriate normalization scheme to account for variation in the number of categories across articles that is increasing over time (see *Appendix B* and **Figure A1** for additional details).

In order to highlight shifts in research orientations that are likely to be associated with the

announcement of HB flagship funding initiatives, we also calculated co-occurrences for the 5-year post-period (2014-2018). The most prominent CIP changes coincident for NA and EU are increases in Biotech. & Genetics with Pathology & Pharmacology [CIP 4,7]; between Psychology and Engineering & Informatics [3,8]; and between Biology and Pathology & Pharmacology [2,7]. Contrariwise, we note consistent decreases between Neurosciences and Engineering & Informatics [1,8]; and between Medical Specialty [5] and Sci./Eng. CIP [8,9]. See **Figures 5-6**(D) for the percent differences across all matrix elements $\Delta\mathbf{C}_{CIP,ij}$ and $\Delta\mathbf{C}_{SA,ij}$, respectively.

This systematic approach also facilitates exploiting the different framings of the grand scientific Brain challenges embodied by each regional HB flagship project, as it is reasonable to expect that different project framings manifest in different recombinant strategies at particular disciplinary interfaces. As such, we explore three types of cross-domain ($X$) configurations – *Broad*, *Neighboring* and *Distant* – each defined accordingly by a particular combination of SA and CIP categories. Based upon stated mission and vision statements, the BRAIN initiative (NA) aligns with *Neighboring* and the Human Brain Project (EU) aligns more closely with *Distant*.

*Broad* is the first and most generic $X$ configuration we explored, based upon combinations of any two or more SA categories (or CIP categories), and represented by our operator notation as $O_{SA}(\vec{F}_p) = X_{SA}$ (and $O_{CIP}(\vec{F}_p) = X_{CIP}$, respectively). The second configuration captures the *Neighboring* neurobiological $\leftrightarrow$ bioengineering interface representing articles that combine: SA [1] with one or more SA from among [2-4]; or either CIP [1,3] with any CIP among [2,4-7]. Articles featuring these configurations are represented using our operator notation as $O_{\text{Neighboring},SA}(\vec{F}_p) = X_{\text{Neighboring},SA}$, $O_{\text{Neighboring},CIP}(\vec{F}_p) = X_{\text{Neighboring},CIP}$, or $O_{\text{Neighboring},SA\&CIP}(\vec{F}_p) = X_{\text{Neighboring},SA\&CIP}$; alternatively, articles not containing the appropriate category combinations are represented by the counterfactual mono-domain state $M$.

And finally, the third configuration captures the more *Distant* neuro-psycho-medical $\leftrightarrow$ techno-computational interface. The specific set of category combinations representing this interface are SA [1-4] $\times$ [5,6] and CIP [1,3,5] $\times$ [4,8]; as above, articles featuring (or not featuring) these configurations are represented as $X_{\text{Distant},SA}\ (M)$, $X_{\text{Distant},CIP}\ (M)$, $X_{\text{Distant},SA\&CIP}\ (M)$. By way of

example, **Figure 7**(A) illustrates an article combining SA 1 and 3, which is thereby classified as both $X_{SA}$ and $X_{\text{Neighboring},SA}$; and an article featuring CIP 1,2,6,8, which is thereby both $X_{CIP}$ and $X_{\text{Distant},CIP}$.

We developed an article-level method to measure cross-domain diversity that is suitable for temporal analysis (see *Methods* for additional definition details). By way of example, consider the vector $\overrightarrow{SA}_p$ which tallies the SA counts for a given article $p$ published in year $t$. We apply the outer tensor product $\overrightarrow{SA}_p \otimes \overrightarrow{SA}_p$ to represent all pairwise co-occurrences in a weighted matrix $\mathbf{D}_p(\vec{v}_p)$ (See *Appendix C* for examples of the outer tensor product). The sum of elements in this co-occurrence matrix are normalized to unity so that each $\mathbf{D}_p(\vec{v}_p)$ contributes equally to averages computed across all articles from a given year. Since the off-diagonal elements represent cross-domain combinations, their relative weight given by $f_{D,p} = 1 - \text{Tr}(\mathbf{D}_p) \in [0, 1)$ is a straightforward measure of categorical diversity.

**Figure 3** shows the trends in mean diversity $\langle f_D(t) \rangle$ at each *Broad*, *Neighboring* and *Distant* interface. Each interface reflects a particular category configuration. Thus, for each configuration we provide a schematic motif illustrating the category combinations captured by $\mathbf{D}_p(\vec{v}_p)$, with diagonal components representing mono-domain articles (indicated by 1 on the matrix diagonal) and upper-diagonal elements capturing cross-domain combinations (indicated by $X$). Comparing SA and CIP overall, there are higher diversity levels for SA following from higher baseline levels of categorical mixing within MeSH, but also a more prominent upward trend over time; data are also disaggregated by region to facilitate comparison. In terms of CIP, **Figure 3**(A) indicates a decline in *Broad* diversity in recent years, with North America (NA) showing higher levels than Europe (EU) and Australasia (AA); these general patterns also evident for *Neighboring* diversity, see **Figure 3**(B). *Distant* CIP diversity shown in **Figure 3**(C) indicates a recent decline for AA and NA, with NA peaking around 2009; contrariwise, EU shows a steady increases consistent with the framing of the Human Brain Project.

In contradistinction, all three regions show steady increase irrespective of configuration in the

case of SA diversity, suggesting that scholars first explore the integration of novel topical combinations before attempting to integrate scholarly expertise. For both *Broad* and *Neighboring* configurations, NA and EU show remarkably similar levels of SA diversity, above AA; however, in the case of *Neighboring*, AA appears to be catching up quickly since 2010, see **Figure 3**(D,E). And in the case of *Distant*, all regions are showing steady increase for the entire period that appears to be in lockstep. See **Figures A3-A4** and *Appendix D* for trends in CIP and SA diversity across additional relevant interfaces.

### 3.5 Integration of CIP and SA in Cross-domain Research

Up to this point we have treated CIP and SA categories as representations of distinct domains, yet in reality scholars are mixing and matching scholarly expertise to address the demands of a particular research problem. Inasmuch as mono-domain articles identify the topical boundary closely associated with individual disciplines, cross-domain articles are useful for identifying otherwise obscured boundaries that call for both $X_{CIP}$ and $X_{SA}$ in combination.

We identified these CIP-SA relations, representing the topical exploration frontier, by collecting articles in the focal period 2009-2018 that are purely mono-domain for both CIP and SA (i.e., those with $O_{CIP}(\vec{F}_p) = O_{SA}(\vec{F}_p) = M$) and a complementary non-overlapping subset of articles that are simultaneously cross-domain for both CIP and SA (i.e., $O_{SA\&CIP}(\vec{F}_p) = X_{SA\&CIP}$). Starting with the mono-domain articles, which by definition feature just a single CIP, we identified the SA that are most frequently associated with each CIP category. Formally, this amounts to calculating the bi-partite network between CIP and SA, denoted by $M_{CIP} \rightleftarrows M_{SA}$. These CIP-SA associations are calculated by averaging the $\overrightarrow{SA}_p$ for mono-domain articles from each CIP category, given by $\langle\overrightarrow{SA}\rangle_{CIP}$. **Figure 3**(G) highlights only the most prominent CIP-SA links (see *Methods 2.9 – Bi-partite Network between CIP and SA* for more details). Likewise, we then calculated the bi-partite network for the subset of $X_{SA\&CIP}$ articles. Not surprisingly, this second network denoted $X_{CIP} \rightleftarrows X_{SA}$ has a higher density of links.

To identify the cross-domain frontier, we calculated the network difference $\Delta_{XM} \equiv X_{CIP} \rightleftarrows$

$X_{SA} - M_{CIP} \rightleftarrows M_{SA}$, and plot just the links with positive values. Hence, these are the CIP-SA links that are over-represented in $X_{CIP} \rightleftarrows X_{SA}$ relative to $M_{CIP} \rightleftarrows M_{SA}$. The results for the *Broad* configuration show the wide array of SA that are reached by way of cross-disciplinary teams. In particular, SA 3 (Phenomena & Processes), representing topics related to the structure-function problem, is integrated by theory-oriented (CIP 9, Chemistry & Physics & Math), problem-oriented (CIP 7, Pathology & Pharmacology), and solution and design-oriented (CIP 4, Biotech. & Genetics) disciplines. In **Figure 8** we show the results for both the *Neighboring* and *Distant* configurations, which provide cross-validation for the choice of CIP and SA categories they represent.

Comparison of $\Delta_{XM}$ across configurations facilitates a qualitative difference-in-difference for identifying emergent cross-domain coupling particular of a given configuration. Regarding the *Neighboring* configuration relative to the *Broad* configuration, we observe over-representation of the links between CIP [2,4-9] and the core neurosciences domain (SA 1), and to a lesser degree Eng. & Informatics (CIP 8) and Biology (SA 2). Likewise, in the case of the *Distant* configuration, the over-represented links are CIP [2,4,7] with the Techniques & Equipment (SA 5). Hence, the hallmark of convergence science within the HB domain is the combination of biotechnology experts developing novel methods.

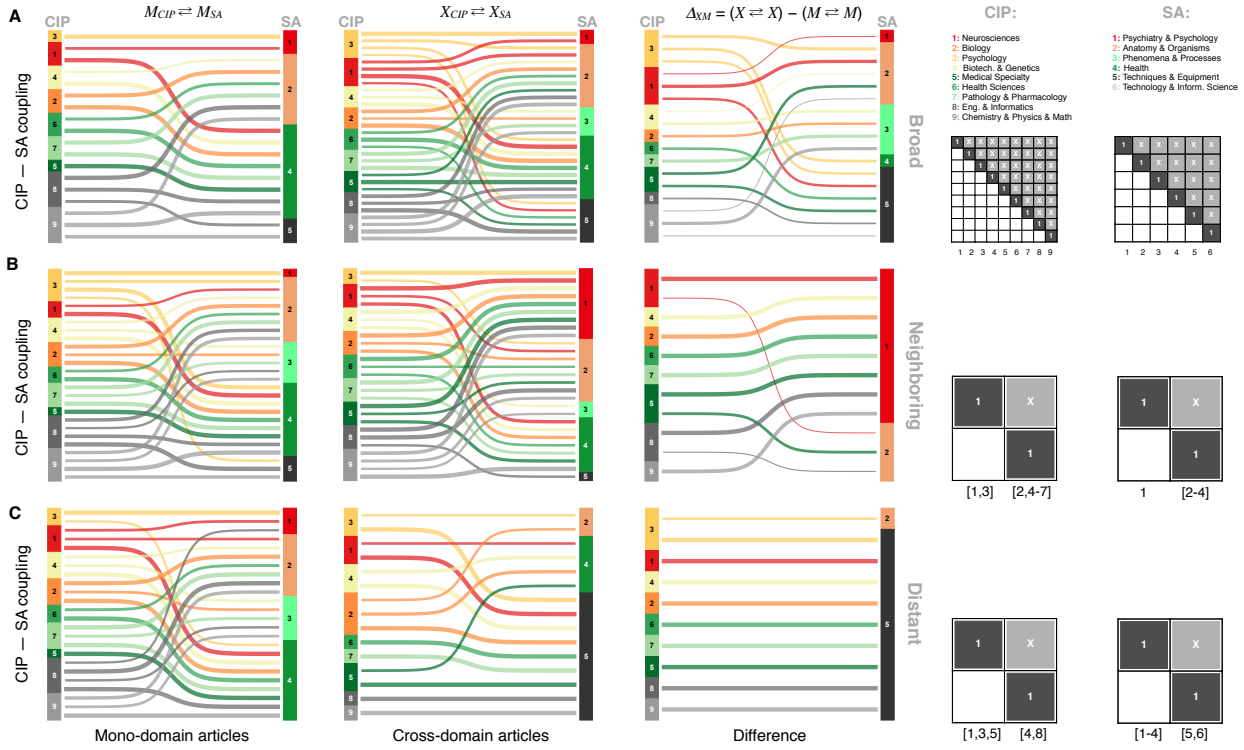



Figure 8: **Cross-domain CIP-SA coupling.** (A) *Broad* configuration. (B) *Neighboring*. (C) *Distant*. The first column illustrates mono-mono coupling, calculated using only the mono-domain articles ($M$). For this case, the bi-partite CIP-SA networks are rather consistent across each configuration, indicating a common baseline for comparison across configurations. The second column shows the CIP-SA coupling network calculated using only the cross-domain articles ($X_{SA\&CIP}$). The third column shows the difference between the corresponding mono- and cross-domain networks in each row. As such, comparison across any two networks in the third column corresponds to a difference-in-difference.

### 3.6 Panel regression – Modeling the Prevalence and Impact of $X$

We constructed article-level and author-level panel data to facilitate measuring factors relating to SA and CIP diversity and shifts related to the ramp-up of HB flagship projects circa 2013 around the globe. The most important control variables recorded for each article are the publication year $y_p$; the total number of coauthors, $k_p$; and the total number of MeSH terms, $w_p$. We also include the total number of international regions associated with the authors' affiliations $N_{R,p}$, and also the total number of distinct categories featured by the article, $N_{SA,p}$ and $N_{CIP,p}$, two measures of categorical breadth. **Figure A2** shows the distribution of these article-level features.

We then focus on two models upon two dependent variables: the first is the propensity for cross-domain research – indicated generically by $X$, but corresponding in different models to $X_{SA}$, $X_{CIP}$ or $X_{SA\&CIP}$ – using a Logit specification to model the likelihood $P(X)$. In the second model the dependent variable is the article's scientific impact, proxied by $c_p$. Building on previous efforts [33, 31], we apply a logarithmic transform to $c_p$ that facilitates removing the time-dependent trend in the location and scale of the underlying log-normal citation distribution [38] (see *Methods 2.10 – Normalization of Citation Impact*). **Figure A5** shows the covariation matrix between the principle variables of interest.

#### 3.6.1 Article-level Model A: Quantifying the Propensity for $X$ and the Role of Flagship HB Projects

We operationalized $O(\vec{F}_p) = X$ or $M$ as a two-state outcome variable, i.e., corresponding to complementary likelihoods $P(X) + P(M) = 1$. Thus, we apply logistic regression to model the odds $Q \equiv \frac{P(X)}{P(M)}$, which measures the propensity to pursue cross-domain research approaches. Since each coauthor is associated with a single CIP category, we exclude solo-authored research papers (i.e., those with $k_p = 1$) from this analysis since the likelihood for those articles is predetermined (i.e., $P(M) = 1$); for the same reason, we also exclude articles with a single Major MeSH category (i.e., those with $w_p = 1$). In addition to the covariates mentioned above, we also include the mean journal citation impact, $\overline{z}_j = \langle z_p | \text{ journal j} \rangle$, calculated as the average $z_p$ for articles from journal

$j$.

We start by estimating the annual growth in $P(X)$, controlling for confounding sources of variation, in particular increasing $k_p$ associated with the growth of team science [44]. In short form, we model the odds as $\log(Q_p) = \beta_0 + \beta_y y_p + \vec{\beta} \cdot \vec{x}$, where $\vec{x}$ represents the additional control variables. See *Appendix E*, in particular Eqns. (4)-(6), for the full model specifications; and **Tables 3-5** for parameter estimates.

**Figure 9**(A) indicates a roughly 3% annual growth in $P(X_{SA})$, consistent with temporal trends in empirical frequencies (**Figure 7**), while also controlling for potential confounders such as increasing team size [44]. In the case of $P(X_{SA\&CIP})$, growth rates are higher for *Broad* and *Distant*. However, for $P(X_{CIP})$ growth rates are generally smaller, indicative of the additional barriers to integrating individual expertise as opposed to just combining research interests.

A timely question we is how HB projects have altered the propensity for $X$. Hence, we added an indicator variable $I_{2014+}$ which takes the value 1 for articles with $y_p \geq 2014$ and 0 otherwise. **Figure 9**(B) indicates significant decline in $P(X)$ for $X_{CIP}$ and $X_{SA\&CIP}$ for each configurational interface, on the order of -30%, consistent with the recent increase in $f_1(t|z)$ visible in **Figure 7**(B).

To further explore how this downturn in $X$ relates to the ramp-up of Flagship HB projects, we use a difference-in-difference (DiD) approach to test for this shift for each of three focal regions, relative to the rest of the world - see **Figure A6**(D-F). AA shows an increasing propensity for HB research at the *Neighboring* interface, and a decline at the *Distant* interface primarily for $X$ involving CIP diversity; EU shows a significant increase in $P(X_{CIP})$ attributable to the HBP, whereas shifts in NA are relatively marginal.

Table 3: Modeling the prevalence of cross-domain activity at the article level. Article-level analysis implemented using the logit model. The dependent variable is a binary indicator variable taking the value 1 if the article features cross-domain combinations (represented by $X_{SA,p}$ or $X_{CIP,p}$ or $X_{SA\&CIP,p}$) and 0 otherwise. Publication data included: articles published in period $y_p \in [1970, 2018]$ with $k_p \geq 2$ and $w_p \geq 2$. Robust standard errors are shown in parenthesis below each point estimate. Reported are odds ratios, $\exp(\beta)$.

| | (1) | (2) | (3) | (4) | (5) | (6) |
|---|---|---|---|---|---|---|
| | $X_{SA}$ | $X_{CIP}$ | $X_{SA\&CIP}$ | $X_{SA}$ | $X_{CIP}$ | $X_{SA\&CIP}$ |
| $y$ | 1.032*** | 1.009*** | 1.046*** | 1.033*** | 1.021*** | 1.061*** |
| | (0.000866) | (0.00242) | (0.00339) | (0.00118) | (0.00315) | (0.00394) |
| $\overline{z}_j$ | 0.997 | 1.282*** | 1.415*** | 0.978 | 1.223* | 1.308** |
| | (0.0298) | (0.0911) | (0.0805) | (0.0320) | (0.103) | (0.126) |
| $\ln k$ | 0.885*** | 1.753*** | 1.562*** | 0.897*** | 1.821*** | 1.639*** |
| | (0.0184) | (0.109) | (0.124) | (0.0142) | (0.117) | (0.140) |
| $\ln w$ | 4.655*** | 0.933* | 4.858*** | 4.678*** | 0.929* | 4.896*** |
| | (0.144) | (0.0285) | (0.217) | (0.150) | (0.0286) | (0.215) |
| $N_R$ | 1.324*** | 7.810*** | 12.00*** | 1.211** | 3.028*** | 4.569*** |
| | (0.0807) | (0.928) | (2.096) | (0.0755) | (0.789) | (0.898) |
| $N_{CIP}$ | 1.307*** | | | 1.294*** | | |
| | (0.0828) | | | (0.0818) | | |
| $N_{SA}$ | | 1.216*** | | | 1.206*** | |
| | | (0.0476) | | | (0.0487) | |
| $I_{2014+}$ | | | | 0.949 | 0.754*** | 0.716*** |
| | | | | (0.0348) | (0.0420) | (0.0367) |
| $I_{R_{NA}}$ | | | | 0.913 | 0.380*** | 0.397*** |
| | | | | (0.0680) | (0.103) | (0.0755) |
| $I_{R_{EU}}$ | | | | 0.942 | 0.313*** | 0.345*** |
| | | | | (0.0712) | (0.0849) | (0.0669) |
| $I_{R_{AA}}$ | | | | 0.746*** | 0.229*** | 0.191*** |
| | | | | (0.0567) | (0.0651) | (0.0386) |
| $I_{R_{NA}} \times I_{2014+}$ | | | | 1.074* | 1.007 | 1.012 |
| | | | | (0.0352) | (0.0186) | (0.0218) |
| $I_{R_{EU}} \times I_{2014+}$ | | | | 0.955 | 1.038* | 0.938** |
| | | | | (0.0313) | (0.0186) | (0.0192) |
| $I_{R_{AA}} \times I_{2014+}$ | | | | 1.111** | 0.898*** | 0.950* |
| $N$ | 602599 | 602599 | 207281 | 602599 | 602599 | 207281 |
| Pseudo $R^2$ | 0.0725 | 0.1945 | 0.3081 | 0.0735 | 0.2045 | 0.3191 |

Exponentiated coefficients; Standard errors in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 4: Conditional definition of $X_p$ – identifying "Neighboring" or shorter-distance cross-domain combinations. Article-level analysis implemented using the logit model. The dependent variable is a binary indicator variable taking the value 1 if the article features cross-domain combinations (represented by $X_{\text{Neighboring},SA,p}$ or $X_{\text{Neighboring},CIP,p}$ or $X_{\text{Neighboring},SA\&CIP,p}$) and 0 otherwise. Publication data included: articles published in period $y_p \in [1970, 2018]$ with $k_p \geq 2$ and $w_p \geq 2$. Robust standard errors are shown in parenthesis below each point estimate. Reported are odds ratios, $\exp(\beta)$.

| | (1) | (2) | (3) | (4) | (5) | (6) |
|---|---|---|---|---|---|---|
| | $X_{\text{Neighboring},SA}$ | $X_{\text{Neighboring},CIP}$ | $X_{\text{Neighboring},SA\&CIP}$ | $X_{\text{Neighboring},SA}$ | $X_{\text{Neighboring},CIP}$ | $X_{\text{Neighboring},SA\&CIP}$ |
| $y$ | 1.030*** | 1.002 | 1.025*** | 1.028*** | 1.012** | 1.036*** |
| | (0.00243) | (0.00267) | (0.00284) | (0.00294) | (0.00378) | (0.00513) |
| $\overline{z}_j$ | 1.488*** | 1.344*** | 1.765*** | 1.428*** | 1.266*** | 1.646*** |
| | (0.0855) | (0.0929) | (0.122) | (0.0746) | (0.0757) | (0.0954) |
| $\ln k$ | 0.530*** | 1.755*** | 1.132 | 0.543*** | 1.832*** | 1.188 |
| | (0.0362) | (0.131) | (0.145) | (0.0346) | (0.133) | (0.154) |
| $\ln w$ | 1.756*** | 0.889*** | 1.816*** | 1.788*** | 0.889*** | 1.799*** |
| | (0.0992) | (0.0304) | (0.106) | (0.0828) | (0.0237) | (0.0888) |
| $N_R$ | 1.763*** | 6.297*** | 8.424*** | 1.853*** | 2.617** | 4.071*** |
| | (0.181) | (0.845) | (1.485) | (0.268) | (0.867) | (1.679) |
| $N_{CIP}$ | 1.429** | | | 1.415** | | |
| | (0.172) | | | (0.171) | | |
| $N_{SA}$ | | 1.230*** | | | 1.215*** | |
| | | (0.0526) | | | (0.0529) | |
| $I_{2014+}$ | | | | 1.029 | 0.770*** | 0.795** |
| | | | | (0.0560) | (0.0478) | (0.0653) |
| $I_{R_{NA}}$ | | | | 1.122 | 0.405* | 0.511 |
| | | | | (0.182) | (0.151) | (0.225) |
| $I_{R_{EU}}$ | | | | 1.192 | 0.327** | 0.404* |
| | | | | (0.202) | (0.121) | (0.179) |
| $I_{R_{AA}}$ | | | | 0.626** | 0.172*** | 0.156*** |
| | | | | (0.110) | (0.0651) | (0.0700) |
| $I_{R_{NA}} \times I_{2014+}$ | | | | 1.053 | 1.040* | 1.009 |
| | | | | (0.0481) | (0.0187) | (0.0418) |
| $I_{R_{EU}} \times I_{2014+}$ | | | | 1.044 | 1.114*** | 1.035 |
| | | | | (0.0481) | (0.0163) | (0.0429) |
| $I_{R_{AA}} \times I_{2014+}$ | | | | 1.274*** | 1.081*** | 1.210*** |
| | | | | (0.0578) | (0.0187) | (0.0475) |
| $N$ | 602599 | 602599 | 430801 | 602599 | 602599 | 430801 |
| Pseudo $R^2$ | 0.0496 | 0.1716 | 0.1919 | 0.0554 | 0.1837 | 0.2041 |

Exponentiated coefficients; Standard errors in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 5: Conditional definition of $X_p$ – identifying "Distant" or longer-distance cross-domain combinations. Article-level analysis implemented using the logit model. The dependent variable is a binary indicator variable taking the value 1 if the article features cross-domain combinations (represented by $X_{\text{Distant},SA,p}$ or $X_{\text{Distant},CIP,p}$ or $X_{\text{Distant},SA\&CIP,p}$) and 0 otherwise. Publication data included: articles published in period $y_p \in [1970, 2018]$ with $k_p \geq 2$ and $w_p \geq 2$. Robust standard errors are shown in parenthesis below each point estimate. Reported are odds ratios, $\exp(\beta)$.

| | (1) | (2) | (3) | (4) | (5) | (6) |
|---|---|---|---|---|---|---|
| | $X_{\text{Distant},SA}$ | $X_{\text{Distant},CIP}$ | $X_{\text{Distant},SA\&CIP}$ | $X_{\text{Distant},SA}$ | $X_{\text{Distant},CIP}$ | $X_{\text{Distant},SA\&CIP}$ |
| $y$ | 1.033*** | 1.017*** | 1.043*** | 1.036*** | 1.035*** | 1.071*** |
| | (0.000769) | (0.00445) | (0.00451) | (0.00174) | (0.0101) | (0.00955) |
| $\overline{z}_j$ | 0.635*** | 1.210 | 0.838 | 0.624*** | 1.127 | 0.750** |
| | (0.0253) | (0.141) | (0.0837) | (0.0202) | (0.119) | (0.0659) |
| $\ln k$ | 0.867*** | 1.740*** | 1.289* | 0.879*** | 1.861*** | 1.403* |
| | (0.0331) | (0.166) | (0.166) | (0.0318) | (0.178) | (0.185) |
| $\ln w$ | 2.258*** | 0.918 | 2.584*** | 2.261*** | 0.894 | 2.496*** |
| | (0.114) | (0.111) | (0.211) | (0.115) | (0.105) | (0.167) |
| $N_R$ | 1.107 | 4.594*** | 5.094*** | 0.986 | 1.652* | 1.924** |
| | (0.0627) | (1.235) | (1.476) | (0.0549) | (0.340) | (0.432) |
| $N_{CIP}$ | 1.181*** | | | 1.169*** | | |
| | (0.0204) | | | (0.0204) | | |
| $N_{SA}$ | | 1.183 | | | 1.171 | |
| | | (0.114) | | | (0.115) | |
| $I_{2014+}$ | | | | 0.871*** | 0.735* | 0.648** |
| | | | | (0.0130) | (0.105) | (0.0991) |
| $I_{R_{NA}}$ | | | | 0.872* | 0.378*** | 0.450*** |
| | | | | (0.0494) | (0.0957) | (0.100) |
| $I_{R_{EU}}$ | | | | 0.894 | 0.123*** | 0.132*** |
| | | | | (0.0537) | (0.0306) | (0.0298) |
| $I_{R_{AA}}$ | | | | 0.725*** | 0.188*** | 0.130*** |
| | | | | (0.0460) | (0.0505) | (0.0305) |
| $I_{R_{NA}} \times I_{2014+}$ | | | | 1.063*** | 0.860 | 0.842 |
| | | | | (0.0106) | (0.0900) | (0.120) |
| $I_{R_{EU}} \times I_{2014+}$ | | | | 1.031** | 1.228* | 1.039 |
| | | | | (0.0104) | (0.125) | (0.144) |
| $I_{R_{AA}} \times I_{2014+}$ | | | | 1.118*** | 0.646*** | 0.711* |
| | | | | (0.0106) | (0.0643) | (0.0968) |
| $N$ | 602599 | 602599 | 396471 | 602599 | 602599 | 396471 |
| Pseudo $R^2$ | 0.0375 | 0.1492 | 0.1474 | 0.0496 | 0.1716 | 0.1919 |

Exponentiated coefficients; Standard errors in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

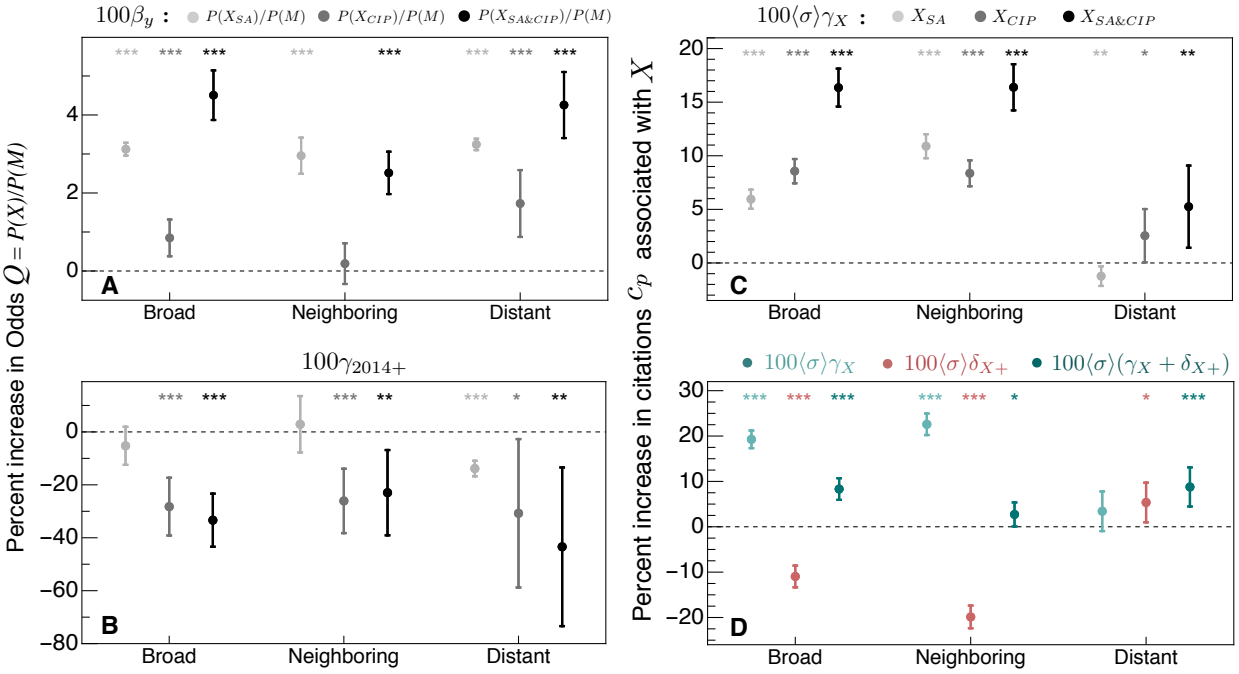



Figure 9: **Propensity for $X$ and citation impact attributable to cross-domain activity at the article level.** (A) Annual growth rate in the likelihood $P(X)$ of research having cross-domain attributes represented generically by $X$. (B) Decreased likelihood $P(X)$ after 2014. (C) Citation premium estimated as the percent increase in $c_p$ attributable to cross-domain mixture $X$, measured relative to mono-domain (M) research articles representing the counterfactual baseline. Calculated using a researcher fixed-effect model specification which accounts for time independent individual-specific factors; see Tables 6-7 for full model estimates. (D) Difference-in-Difference ($\delta_{X+}$) estimate of the "Flagship project effect" on the citation impact of cross-domain research. Shown are point estimates with 95% confidence interval. Asterisks above each estimate indicate the associated $p-$value level: $^{*}$ $p < 0.05$, $^{**}$ $p < 0.01$, $^{***}$ $p < 0.001$.

### 3.6.2 Author-level Model B: Quantifying Effect of $X$ on Scientific Impact

To test for a citation premium attributable to $X$, we model the normalized citation impact $z_p = \alpha_a + \gamma_{X_{SA}} I_{X_{SA,p}} + \gamma_{X_{CIP}} I_{X_{CIP,p}} + \vec{\beta} \cdot \vec{x}$, where $\vec{x}$ represents the additional control variables and $\alpha_a$ represents an author fixed-effect to account for unobserved time-invariant factors specific to each researcher. The primary test variables are $I_{X_{SA,p}}$ and $I_{X_{CIP,p}}$, two binary factor variables with $I_{X_{CIP,p}}$= 1 if $O_{CIP}(\vec{F}_p) = X_{CIP}$ and 0 if $O_{CIP}(\vec{F}_p) = M$, and defined similarly for SA; to facilitate distinguishing estimates by configuration, in the case of *Neighboring* we use the notation $I_{X_{\text{Neighboring},CIP}}$ and $I_{X_{\text{Neighboring},SA}}$, and similarly for *Distant*. Full model estimates are shown in **Tables 6 - 7**.

**Figure 9**(C) summarizes the model estimates – $\gamma_{X_{SA}}$, $\gamma_{X_{SA}}$ and $\gamma_{X_{SA\&CIP}}$ – quantifying the citation premium attributable to $X$. To translate the effect on $z_p$ into the associated citation premium in $c_p$, we readily calculate the percent change $100\Delta c_p/c_p$ associated with a shift in $I_{X,p}$ from 0 to 1, which due to the property of logs is given by $100\Delta c_p/c_p \approx 100\langle\sigma\rangle\gamma_X$, which follows because $\sigma_t \approx \langle\sigma\rangle = 1.24$ is approximately constant over the period 1970-2018 (see *Appendix E: Author-level Model*).

Our results indicate a robust statistically significant positive relationship between cross-disciplinarity ($X_{CIP}$) and citation impact, consistent with a different case study of the genomics revolution [33]. To be specific, we calculate a 8.6% citation premium for the Broad configuration ($\gamma_{X_{CIP}} = 0.07$; $p < 0.001$), meaning that the average cross-disciplinary publication is more highly cited than the average mono-disciplinary publication. We calculate a smaller 5.9% citation premium associated with $X_{SA}$ ($\gamma_{X_{SA}} = 0.05$; $p < 0.001$). Yet the effect associated with articles featuring $X_{CIP}$ and $X_{SA}$ simultaneously is considerably larger (16% citation premium; $\gamma_{X_{SA\&CIP}} = 0.13$; $p < 0.001$).

Table 6: Career-level analysis using panel model with individual researcher fixed effects. Publication data included: articles published in period $y_p \in [1970, 2018]$ with $k_p \geq 2$ and $w_p \geq 2$; only includes researchers with $N_a \geq 10$ articles satisfying these criteria. Robust standard errors are shown in parenthesis below each point estimate. Y indicates additional fixed effects included in the regression model.

| | (1) | (2) | (3) | (4) | (5) | (6) |
|---|---|---|---|---|---|---|
| | $z_p$ | $z_p$ | $z_p$ | $z_p$ | $z_p$ | $z_p$ |
| $\ln k$ | 0.415*** | 0.416*** | 0.421*** | 0.438*** | 0.424*** | 0.406*** |
| | (0.00470) | (0.00468) | (0.00469) | (0.00551) | (0.00547) | (0.00527) |
| $\ln w$ | 0.0320*** | 0.0376*** | 0.0379*** | 0.0244*** | 0.0544*** | 0.0385*** |
| | (0.00468) | (0.00465) | (0.00466) | (0.00615) | (0.00528) | (0.00536) |
| $\tau$ | -0.0107*** | -0.0106*** | -0.0105*** | -0.0115*** | -0.00987*** | -0.0101*** |
| | (0.00148) | (0.00148) | (0.00149) | (0.00229) | (0.00170) | (0.00157) |
| $I_{X_{SA}}$ | 0.0480*** | | | | | |
| | (0.00367) | | | | | |
| $I_{X_{CIP}}$ | 0.0691*** | | | | | |
| | (0.00466) | | | | | |
| $I_{X_{\text{Neighboring},SA}}$ | | 0.0878*** | | | | |
| | | (0.00461) | | | | |
| $I_{X_{\text{Neighboring},CIP}}$ | | 0.0675*** | | | | |
| | | (0.00496) | | | | |
| $I_{X_{\text{Distant},SA}}$ | | | -0.00993** | | | |
| | | | (0.00376) | | | |
| $I_{X_{\text{Distant},CIP}}$ | | | 0.0205* | | | |
| | | | (0.0102) | | | |
| $I_{X_{SA\&CIP}}$ | | | | 0.132*** | | |
| | | | | (0.00728) | | |
| $I_{X_{\text{Neighboring},SA\&CIP}}$ | | | | | 0.132*** | |
| | | | | | (0.00886) | |
| $I_{X_{\text{Distant},SA\&CIP}}$ | | | | | | 0.0424** |
| | | | | | | (0.0158) |
| constant | -0.719*** | -0.719*** | -0.736*** | -0.677*** | -0.781*** | -0.711*** |
| | (0.0541) | (0.0540) | (0.0543) | (0.0770) | (0.0547) | (0.0583) |
| Year ($y$) dummy | Y | Y | Y | Y | Y | Y |
| Topic category ($\overrightarrow{SA}$) dummy | Y | Y | Y | Y | Y | Y |
| Department category ($\overrightarrow{CIP}$) dummy | Y | Y | Y | Y | Y | Y |
| Region ($\overrightarrow{R}$) dummy | Y | Y | Y | Y | Y | Y |
| $N$ | 825147 | 825147 | 825147 | 358237 | 552254 | 527347 |
| adj. $R^2$ | 0.102 | 0.102 | 0.101 | 0.131 | 0.090 | 0.093 |
| F | 265.7 | 262.0 | 251.4 | 231.3 | 198.5 | 193.5 |
| # researcher profiles ($df_r + 1$) | 8448 | 8448 | 8448 | 8422 | 8435 | 8441 |

Standard errors in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 7: Flagship Project effect: Career-level analysis using panel model with researcher fixed effects. Publication data included: articles published in period $y_p \in [1970, 2018]$ with $k_p \geq 2$ and $w_p \geq 2$; only includes researchers with $N_a \geq 10$ articles satisfying these criteria. Robust standard errors are shown in parenthesis below each point estimate. Y indicates additional fixed effects included in the regression model.

| | (1) | (2) | (3) |
|---|---|---|---|
| | $z_p$ | $z_p$ | $z_p$ |
| $\ln k$ | 0.438*** | 0.425*** | 0.406*** |
| | (0.00551) | (0.00547) | (0.00528) |
| $\ln w$ | 0.0250*** | 0.0546*** | 0.0385*** |
| | (0.00616) | (0.00528) | (0.00536) |
| $\tau$ | -0.0108*** | -0.00965*** | -0.00878*** |
| | (0.00255) | (0.00189) | (0.00175) |
| $I_{2014+}$ | 0.0137 | 0.00818 | -0.0571*** |
| | (0.0160) | (0.0133) | (0.0111) |
| $I_{X_{SA\&CIP}}$ | 0.155*** | | |
| | (0.00796) | | |
| $I_{X_{SA\&CIP}} \times I_{2014+}$ | -0.0884*** | | |
| | (0.00985) | | |
| $I_{X_{\text{Neighboring},SA\&CIP}}$ | | 0.182*** | |
| | | (0.00973) | |
| $I_{X_{\text{Neighboring},SA\&CIP}} \times I_{2014+}$ | | -0.160*** | |
| | | (0.0103) | |
| $I_{X_{\text{Distant},SA\&CIP}}$ | | | 0.0276 |
| | | | (0.0180) |
| $I_{X_{\text{Distant},SA\&CIP}} \times I_{2014+}$ | | | 0.0432* |
| | | | (0.0180) |
| constant | -0.666*** | -0.777*** | -0.690*** |
| | (0.0729) | (0.0515) | (0.0559) |
| Year ($y$) dummy | Y | Y | Y |
| Topic category ($\overrightarrow{SA}$) dummy | Y | Y | Y |
| Department category ($\overrightarrow{CIP}$) dummy | Y | Y | Y |
| Region ($\overrightarrow{R}$) dummy | Y | Y | Y |
| $N$ | 358237 | 552254 | 527347 |
| adj. $R^2$ | 0.131 | 0.091 | 0.093 |
| F | 229.0 | 198.6 | 191.3 |
| # researcher profiles ($df_r + 1$) | 8422 | 8435 | 8441 |

Standard errors in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

We observe similar citation premium values corresponding to the *Neighboring* configuration. In particular, for articles performing research at the the neurobiological $\leftrightarrow$ bioengineering interface, where differences in disciplinary expertise are relatively small, the citation premium is relatively larger for $SA$ than its *Broad* counterpart (11% citation premium; $\gamma_{X_{\text{Neighboring},SA}} = 0.088$; $p < 0.001$), whereas there is little difference between *Neighboring* and *Broad* for $X_{CIP}$ and $X_{SA\&CIP}$. Regarding estimates for the *Distant* configuration, capturing the relatively larger disciplinary differences that research endeavors must overcome in order to succeed, we observe smaller effect sizes. However, among the values, the largest value corresponds to when research combines both $X_{SA}$ and $X_{CIP}$ simultaneously (5.2% citation premium; $\gamma_{X_{\text{Distant},SA\&CIP}} = 0.04$; $p < 0.001$); to be clear, this estimate is calculated excluding articles with $X_{SA}$ or $X_{CIP}$ so that the counterfactual to $X_{\text{Distant},SA\&CIP}$ is also articles classified as $M$. The reduction in $\gamma_{X_{\text{Distant},SA\&CIP}}$ likely reflects the challenges bridging communication, methodological and theoretical gaps across the *Distant* techno-computational interface.

As in the Article-level model, we also tested for shifts in the citation premium attributable to the advent of Flagship HB projects using a similar DiD approach. **Figure 9**(D) shows the citation premium $\gamma_{X_{SA\&CIP}}$ attributable for articles published prior to 2014, and the difference $\delta_{X+}$ corresponding to the added effect for articles published after 2014. For *Broad* and *Distant* we consistently observe $\delta_{X+} < 0$, indicating a reduced citation premium for post-2014 research. By way of example for the *Broad* configuration: whereas cross-domain articles published prior to 2014 show a 19% citation premium ($\gamma_{X_{SA\&CIP}} = 0.15$; $p < 0.001$), those published after 2014 have just a 19%-11% = 8% citation premium ($\delta_{X_{SA\&CIP}+} = -0.09$; $p < 0.001$). The reduction of the citation premium is even larger for *Neighboring* ($\delta_{\text{Neighboring},X_{SA\&CIP}+} = -0.16$; $p < 0.001$). Yet for *Distant*, we observe a completely different trend – research combining both $X_{SA}$ and $X_{CIP}$ simultaneously has advantage over those with just $X_{CIP}$ or $X_{SA}$ or $M$, in that order ($\delta_{\text{Dist.},X_{SA\&CIP}+} = 0.04$; $p = 0.016$; 95% CI = [.01, .08]).

We briefly summarize the explanatory variables coefficients that are consistent across all models. Consistent with prior research on team-science [44] and the role of cross-disciplinarity [33], we

observe a positive relationship between team-size and citation impact ($\beta_k = 0.415$; $p < 0.001$), which translates to a $\langle\sigma\rangle\beta_k \approx 0.5\%$ increase in citations associated with a 1% increase in team size (since $k_p$ enters in log in our specification). We also observe a positive relationship for topical breadth ($\beta_w = 0.03$; $p < 0.001$), which translates to a much smaller $\langle\sigma\rangle\beta_w \approx 0.04\%$ increase in citations associated with a 1% increase in the number of major MeSH headings used to describe the research content. And finally, regarding the career life-cycle, we observe a negative relationship with increasing career age ($\beta_\tau = -0.011$; $p < 0.001$) consistent with prior studies [33], translating to a $100\langle\sigma\rangle\beta_\tau \approx$ -1.3% decrease in $c_p$ associated with every additional career year. See **Tables 6-7** for the full set of model parameter estimates.

### 3.7 Article Level Relation between Brain Parts

In our analysis, we identified the relation between top 20 brain parts according to the articles. It also indicates how brain science researchers worked for their research and what portion of the brain parts they used for their analysis. See *Methods 2.16: Represent Brain Parts Based on Articles* for detail of the methods for this representation. **Figure 4** represents the 2D plot of brain parts based on articles. From this figure we can see that the 2D representation of brain parts mostly captures the original locations of brain parts. That means when researchers worked for any brain part then most of the time they used few more neighbor parts for their analysis. The interesting point is this NLP analysis beautifully captures the research syntax.

### 3.8 Temporal Analysis of Gene Expression and Magnetic Resonance

Gene expression and magnetic resonance are two key research areas in brain science field. We investigated the evolution of gene expression and magnetic resonance using temporal analysis. **Figure 10** shows the temporal analysis time series output. According to this figure, gene expression starts growing from 1990 when actually the genomics project was started. After genomics project the brain science project started and its end in 2013. We can also see that the gene expression research is going downward after 2013. We know that the magnetic resonance imaging (MRI) is

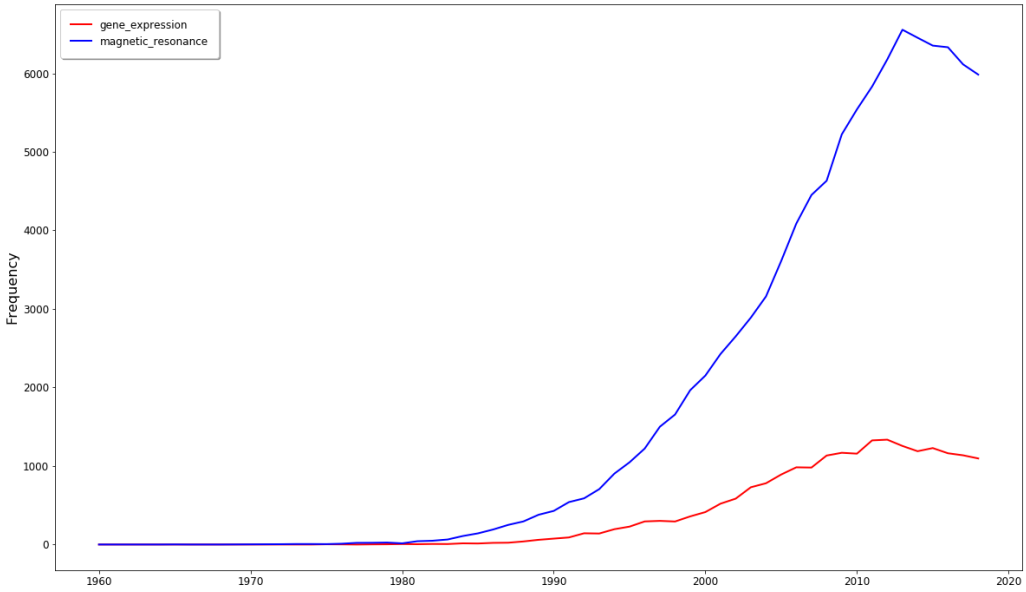



Figure 10: **Time series representation of Gene Expression and Magnetic Resonance.**

one of the key breakthroughs for brain science. If we see the history, MRI was invented on 1977 and fMRI was invented on 1990, which is very important for brain research. From the figure we can also see this thing that, from 1990 the magnetic resonance line takes off which is actually its knee. So, this NLP based temporal analysis beautifully captures the trend of gene expression and magnetic resonance from articles.

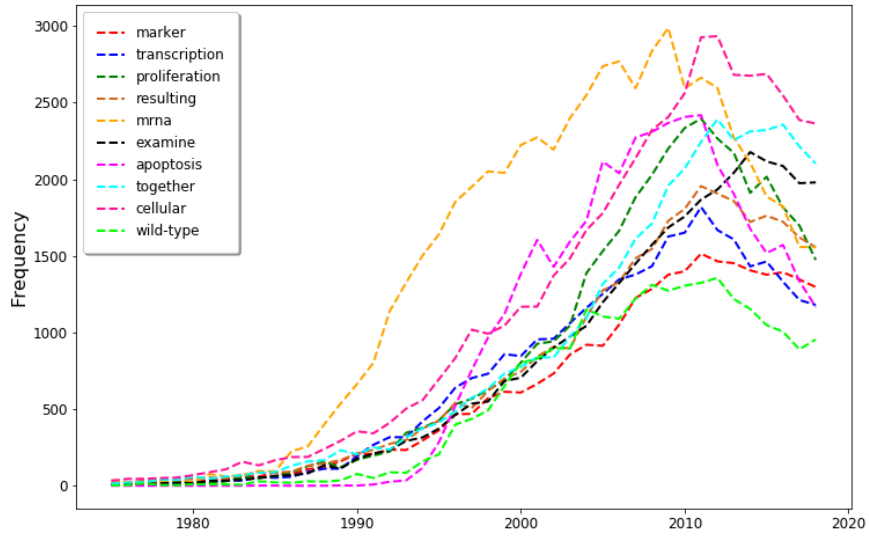
## 3.9 Temporal Analysis of Cosine Similar Topics of Gene Expression and Magnetic Resonance

After our investigation of the evolution of gene expression and magnetic resonance we want to go more deep and want to see the evolution of few similar areas related to gene expression and magnetic resonance. We selected top 10 cosine similar words for both gene expression and magnetic resonance. See *Methods 2.17* for the detail methods used for this process. We used the same temporal analysis for all top 10 cosine similar terms of gene expression and magnetic resonance.

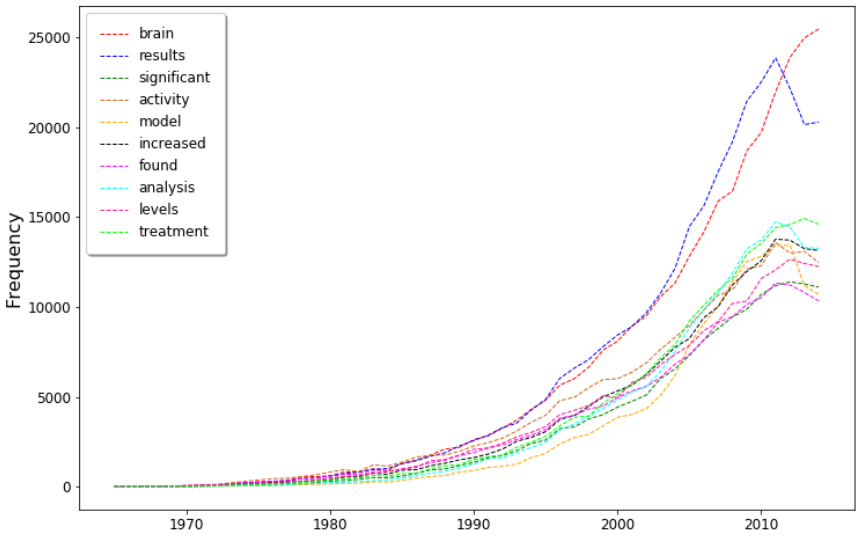
**Figure 11**(a) shows the temporal analysis time series output for all top 10 terms related to gene expression, and **Figure 11**(b) shows the temporal analysis time series output for all top 10 terms related to magnetic resonance. If we see the temporal plot, all the terms related to gene expression basically started going upward after 1990 and keep this upward trend until 2013 which follows the same pattern like gene expression. As we said the magnetic resonance used for brain science research and from this temporal analysis we can see that the brain is the top cosine similar topic of magnetic resonance and it started growing upward after 2000's.

### 3.10 Predicting Breakthroughs

In brain science, we found two important breakthroughs which are Magnetic Resonance Imaging (MRI) and Genomics. Statistics can find the factors, but it can not find the breakthroughs from text. We tried to find these breakthroughs and pattern using machine learning and NLP process. We track these terms in terms of frequency and in terms of associative words. For Magnetic Resonance Imaging (MRI) and Genomics, the term frequency tracks the impacts successfully (see **Figure 10**). For MRI, we found that associative terms like brain and results are closely related to MRI (see **Figure 11**). So, the increasing term frequency accompanied by associative terms could serve as predictive features for breakthroughs.




(a)




(b)

Figure 11: **Time series representation of top 10 cosine similar words for Gene Expression and Magnetic Resonance.** (a) represents the top 10 cosine similar words of Gene Expression (b) represents the top 10 cosine similar words of Magnetic Resonance.

# 4 Behind the Numbers

The paper's analytic models demonstrate that publications made out of an epistemically distant mix - combining cross-domain topics with cross-disciplinary author pedigree (i.e., $X_{\text{Distant},SA\&CIP}$) - trend well and exhibit persistent impact (**Figure 9**). Looking beyond the modeling results, and into the actual publications, one identifies four driving patterns behind this converging channel of HBS:

**Magnetic Resonance (MR) imaging.** MR imaging has been instrumental in identifying functional networks in the brain. For this reason, MR imaging has reshaped brain research since the 1990s and remained a strong presence through the 2000s and 2010s. As a method that involves both sophisticated technology and core brain expertise - two aspects that cannot be easily bridged within a mono-disciplinary team - MR imaging has been a focal point for $X_{\text{Distant},SA\&CIP}$ scholarship. As an impactful instance of such scholarship, we cite from our corpus the paper by Van Dijk et al. [41] that addresses the problem of motion effects in MR imaging. Motion is a pernicious confounding factor that can invalidate any MR brain study. Hence, the said paper serves as an excellent example of how an existential threat to a line of research acts as an attractor of distant cross-disciplinary collaborations with an all-encompassing theme. Specifically, article [41] includes authors from CIP 5 (medical specialists) and CIP 8 (engineers and computer scientists), while thematically runs the gamut of brain subject areas including SA 2 (Anatomy & Organisms), SA 3 (Phenomena & Processes), SA 5 (Techniques & Equipment), and SA 6 (Technology & Information Science).

**Genomics.** Much like MR imaging has been acting as a catalyzer of $X_{\text{Distant},SA\&CIP}$ scholarship, genomics has been playing a similar role. Following the completion of the Human Genome Project (HGP) in the early 2000s, genomics and biotechnology methods in general, have established a foothold in brain research, which is supported by cross-disciplinary teams featuring a mix of medical and genomics researchers. This up and coming line of convergent research made headway in solving long-standing morbidity riddles and formulating novel therapies.

An example of the former is a deeper understanding of the genetic basis of developmental delay by Cooper et al. [6]. An example of the latter is the treatment of glioblastoma with recombinant poliovirus [8]. In more detail, both articles include authors from CIP 4 (biotechnologists/geneticists) and CIP 5 (medical specialists). Thematically, the articles cast a wide net on brain subject areas with [6] covering SA 1 (Psychiatry & Psychology), SA 3 (Phenomena & Processes), SA 4 (Health), and SA 5 (Techniques & Equipment), and [8] covering SA 2 (Anatomy & Organisms), SA 4 (Health), and SA 5 (Techniques & Equipment).

**Robotics.** In the early 2010s neurally controlled robotic prosthesis started coming of age thanks to collaboration between neuroscientists (CIP 1) and biotechnologists (CIP 4). A prime example of this is the work by Hochberg et al. [18] on robotic arms for tetraplegics, which covers every single subject area of HBS from SA 1 (Psychiatry & Psychology) all the way to SA 6 (Technology & Information Science).

**Artificial Intelligence (AI) and Big Data.** From the mid 2010s, following explosive developments in machine learning (ML), deep AI methods were brought to bear on MR data, pushing decidedly brain imaging towards more quantitative, accurate, and automated diagnostic methods. The work by Kamnitsas et al. [24] on brain legion segmentation using Convolutional Neural Networks (CNN) is an apt example of this trend. It is the product of collaboration between medical specialists (CIP 5) and engineers (CIP 8), bringing together nearly all subject areas in brain science (SA 2-4 and SA 6). Simultaneously, massive brain datasets to feed the new voracious AI engines made their appearance along with methods to control noise and ensure their validity. The work by Alfaro-Almagro et al. [1] exemplifies this line of research, being the collaborative product of neuroscientists (CIP 1), health scientists (CIP 6), and engineers (CIP 8), with an all encompassing content (SA 2-6).

All in all, $X_{\text{Distant},SA\&CIP}$ products are characterized by total SA coverage, typically including 3-4 non-technical SA plus 1-2 technical SA.

The results of the present study also demonstrate the increasingly prominent role of scholarship

from an epistemically neighboring mix, combining cross-domain features with cross-disciplinary pedigree (i.e., $X_{\text{Neighboring},SA\&CIP}$). While in the case of Distant $X_{SA\&CIP}$, publications incorporate content that runs the gamut of subject areas in brain science, in the case of Neighboring $X_{SA\&CIP}$, publication content is restricted to neuro-psycho-biological areas. Given this restriction, the SA coverage (3-4 non-technical SA) exceeds the disciplinary bounds implied by the CIP set of the authors (typically two non-technical CIP).

A prime example of $X_{\text{Neighboring},SA\&CIP}$ scholarship is the neuroscience review on the attention system of the human brain [35]. The authorship features a psychologist (CIP 3) and a medical specialist (CIP 5), while the content of the paper covers SA 1 (Psychiatry & Psychology), SA 2 (Anatomy & Organisms), and SA 5 (Techniques & Equipment). In the paper's introduction, the authors themselves make an unequivocal case about the breadth of the research by stating: "The framework presented in the original article has helped to integrate behavioral, systems, cellular, and molecular approaches to common problems in attention research."

Antipodal to mixed scholarship from cross-disciplinary teams, stands focused scholarship from mono-disciplinary teams. This type of scholarship seems to draw its impact from long-standing problems that continue to receive conventional treatment. An illustrating example is the stroke reports published yearly by a group of the American Heart Association (AHA) - see for example, the 2015 report in the journal *Circulation* [26]. The authorship of these reports is invariably a group of health scientists (CIP 6) with a theme squarely placed in the health domain (SA 6).

# 5 Conclusions

If the state of HBS is any indication, convergence in science appears to be in a state of flux, developing in ways that are not immediately apparent and often in tension with the prevalent view that multi-disciplinary teams have been gaining unstoppable momentum. Although it is true that the team size has grown over time, the present study suggests that the team's disciplinary dimensions tend to be a subset of the topical dimensions of the problem under investigation. In other words, teams tend to economize in disciplinary expertise, adopting an expansive approach where convergence takes place in part (or in whole) within polymathic researchers rather than between specialists. This phenomenon is particularly intense in research publications involving topics that are epistemically close (i.e., $X_{\text{Neighboring},SA}$ configuration).

Arguably, a certain degree of expansiveness is needed in multi-disciplinary teams to operate in harmony. For example, in the case of a psychologist collaborating with a medical specialist, it would be ideal if each one knew a little bit about the other's field, so that they establish an effective knowledge bridge. After all, this is what transforms a multi-disciplinary team to a cross-disciplinary team, where convergence becomes operative. However, this is different than what appears to happen here, where using our example, the medical specialist or specialists choose not to partner at all with psychologists in the prosecution of bi-domain research. In essence, a meaningful strategy of partial redundancy is abandoned in favor of a risky strategy of total replacement.

One could point out that in times past polymathic scientists and engineers (e.g., Leonardo da Vinci) worked wonders - why not now? The results clearly suggest that over-expansive polymathic approaches do not work very well nowadays. The impact of publications from polymathic teams ($X_{\text{Neighboring},SA}$ and $X_{\text{Distant},SA}$) is significantly inferior to the impact of publications from more balanced disciplinary teams ($X_{\text{Neighboring},SA\&CIP}$ and $X_{\text{Distant},SA\&CIP}$) - **Figure 9**(C). There are various possible explanations for this. Certainly, science has progressed and grown by leaps and bounds since Renaissance and thus, it is much more challenging for researchers to master effectively multiple domains. It is also true that convergence is a relatively young trend and if our higher

education systems become more attuned to it, they may produce more adept polymathic scientists in the future.

For the moment, it is somewhat unnerving that this semi-mature polymathic trend proliferates and competes with the gold standard, that is, configurations featuring balanced cross-disciplinary teams and topics. Disturbingly, Flagship HBS projects appear to aid and abet expansive team behaviors - **Figure 9**(B). One could form reasonable hypotheses about how funding initiatives unwittingly amplify such suboptimal scholarly behaviors. Take for instance the US BRAIN Initiative, with the expressed aim to support multi-disciplinary research at the meso-level, that is, the investigation of brain networks. This research goal points to Neighboring topics, where scientists with polymathic tendencies may feel more emboldened to short-circuit expertise. To add to that, there are practical pressures associated with proposal calls. For instance, it is easier and faster for researchers to find collaborators from their own discipline, thus forming teams on time to meet proposal deadlines. Not to mention that funding levels are not unlimited and bringing additional specialists into the team is a financial consideration.

Since the polymathic trend pre-existed the Flagship HBS projects, however, it must have deeper roots. Our working hypothesis is that it represents an emergent scholarly behavior in the context of globalized and Internetized science, which can be explained by the theory of expansive learning [10]. Indeed, many of the activity signals brought to the fore in this study bear the hallmarks of expansive learning. Perhaps the most telling such signal is the propensity towards topically diverse publications - **Figure 3**(D-F), which largely stems from horizontal movements in the research focus of individual scientists rather than vertical integration among experts from different disciplines - **Figure 3**(A-C). The entire system is also highly interconnected, as it is evident from the collaboration networks (**Figure 7**(B))and its behavioral interlocking (**Figure 3**(F)). These are conditions that are conducive to boundary crossing, especially with respect to topics, which in learning terms can act as objects facilitating "minimum energy" expansions [40].

Consistent also with other studies in expansive learning, actions taken by participants do not necessarily correspond to the intentions by the interventionists [39]. The participants are brain

scientists in this case, and the interventionists are the funding agencies and the scientific establishment at large. While the latter aim to promote research powered by true multi-disciplinary teams, the former significantly water down this ideal.

Managing a highly dynamic process, such as the evolving behaviors in a science ecosystem, is a challenging undertaking. Policy makers and the scientific commons cannot aspire to effectively manage something like that if they do not even understand it. In this context, the present work is an important contribution, because it lays bare the two competing undercurrents in the making of brain science - polymathic vs. cross-disciplinary. Importantly, it provides the conceptual and methodological framework to dissect the formative processes in any other science frontier.

## Bibliography


[1] Alfaro-Almagro, F., Jenkinson, M., Bangerter, N. K., Andersson, J. L., Griffanti, L., Douaud, G., Sotiropoulos, S. N., Jbabdi, S., Hernandez-Fernandez, M., Vallee, E., et al. Image processing and Quality Control for the first 10,000 brain imaging datasets from UK Biobank. *Neuroimage 166* (2018), 400–424.

[2] Amunts, K., Ebell, C., Muller, J., Telefont, M., Knoll, A., and Lippert, T. The Human Brain Project: Creating a European research infrastructure to decode the human brain. *Neuron 92*, 3 (2016), 574–581.

[3] Begg, M. D., Crumley, G., Fair, A. M., Martina, C. A., McCormack, W. T., Merchant, C., Patino-Sutton, C. M., and Umans, J. G. Approaches to preparing young scholars for careers in interdisciplinary team science. *Journal of Investigative Medicine 62*, 1 (2014), 14–25.

[4] Bowler, P. J., and Morus, I. R. *Making modern science: A historical survey*. University of Chicago Press, 2010.

[5] Committee, A. B. A. S., et al. Australian Brain Alliance. *Neuron 92*, 3 (2016), 597–600.

[6] Cooper, G. M., Coe, B. P., Girirajan, S., Rosenfeld, J. A., Vu, T. H., Baker, C., Williams, C., Stalker, H., Hamid, R., Hannig, V., et al. A copy number variation morbidity map of developmental delay. *Nature Genetics 43*, 9 (2011), 838.

[7] Council, N. R. *Convergence: Facilitating transdisciplinary integration of life sciences, physical sciences, engineering, and beyond*. National Academies Press, Washington, D.C., 2014.

[8] Desjardins, A., Gromeier, M., Herndon, J. E., Beaubier, N., Bolognesi, D. P., Friedman, A. H., Friedman, H. S., McSherry, F., Muscat, A. M., Nair, S., et al. Recurrent glioblastoma treated with recombinant poliovirus. *New England Journal of Medicine 379*, 2 (2018), 150–161.

[9] Doherty, P. *The beginner's guide to winning the nobel prize: advice for young scientists*. Columbia University Press, 2006.

[10] Engeström, Y., and Sannino, A. Studies of expansive learning: Foundations, findings and future challenges. *Educational Research Review 5*, 1 (2010), 1–24.

[11] Eyre, H. A., Lavretsky, H., Forbes, M., Raji, C., Small, G., McGorry, P., Baune, B. T., and Reynolds, C. Convergence science arrives: how does it relate to psychiatry? *Academic Psychiatry 41*, 1 (2017), 91–99.

[12] Fleming, L. Recombinant uncertainty in technological search. *Management Science 47*, 1 (2001), 117–132.

[13] Fleming, L. Perfecting cross-pollination. *Harvard Business Review 82*, 9 (2004), 22–24.

[14] Foster, J. G., Rzhetsky, A., and Evans, J. A. Tradition and innovation in scientists' research strategies. *American Sociological Review 80*, 5 (2015), 875–908.

[15] Gerson, L. P., and Gerson, L. P. *Ancient epistemology*, vol. 1. Cambridge University Press, 2009.

[16] Grillner, S., Ip, N., Koch, C., Koroshetz, W., Okano, H., Polachek, M., Poo, M.-m., and Sejnowski, T. J. Worldwide initiatives to advance brain research. *Nature Neuroscience 19*, 9 (2016), 1118–1122.

[17] Heller, A. *Renaissance man*. Routledge, 2015.

[18] Hochberg, L. R., Bacher, D., Jarosiewicz, B., Masse, N. Y., Simeral, J. D., Vogel, J., Haddadin, S., Liu, J., Cash, S. S., Van Der Smagt, P., et al. Reach and grasp by people with tetraplegia using a neurally controlled robotic arm. *Nature 485*, 7398 (2012), 372–375.

[19] Hsiehchen, D., Espinoza, M., and Hsieh, A. Multinational teams and diseconomies of scale in collaborative research. *Science Advances 1*, 8 (2015), e1500211.

[20] Hughes, J. A., and Hughes, J. *The Manhattan project: Big science and the atom bomb*. Columbia University Press, 2003.

[21] Jabalpurwala, I. Brain Canada: one brain one community. *Neuron 92*, 3 (2016), 601–606.

[22] Jeong, S.-J., Lee, H., Hur, E.-M., Choe, Y., Koo, J. W., Rah, J.-C., Lee, K. J., Lim, H.-H., Sun, W., Moon, C., et al. Korea Brain Initiative: integration and control of brain functions. *Neuron 92*, 3 (2016), 607–611.

[23] Jorgenson, L. A., Newsome, W. T., Anderson, D. J., Bargmann, C. I., Brown, E. N., Deisseroth, K., Donoghue, J. P., Hudson, K. L., Ling, G. S., MacLeish, P. R., et al. The brain initiative: developing technology to catalyse neuroscience discovery. *Philosophical Transactions of the Royal Society B: Biological Sciences 370*, 1668 (2015), 20140164.

[24] Kamnitsas, K., Ledig, C., Newcombe, V. F., Simpson, J. P., Kane, A. D., Menon, D. K., Rueckert, D., and Glocker, B. Efficient multi-scale 3D CNN with fully connected CRF for accurate brain lesion segmentation. *Medical Image Analysis 36* (2017), 61–78.

[25] Majeti, D., Akleman, E., Ahmed, M. E., Petersen, A. M., Uzzi, B., and Pavlidis, I. Scholar plot: Design and evaluation of an information interface for faculty research performance. *Frontiers in Research Metrics and Analytics 4* (2020), 6.

[26] Mozaffarian, D., Benjamin, E. J., Go, A. S., Arnett, D. K., Blaha, M. J., Cushman, M., De Ferranti, S., Després, J.-P., Fullerton, H. J., Howard, V. J., et al. Executive summary: Heart disease and stroke statistics—2015 update: A report from the American Heart Association. *Circulation 131*, 4 (2015), 434–441.

[27] Nissani, M. Fruits, salads, and smoothies: A working definition of interdisciplinarity. *The Journal of Educational Thought 29* (1995), 119–126.

[28] Okano, H., Miyawaki, A., and Kasai, K. Brain/MINDS: brain-mapping project in Japan. *Philosophical Transactions of the Royal Society B: Biological Sciences 370*, 1668 (2015), 20140310.

[29] Page, S. E. *The difference: how the power of diversity creates better groups, firms, schools, and societies.* Princeton University Press, 2008.

[30] Pavlidis, I., Petersen, A. M., and Semendeferi, I. Together we stand. *Nature Physics 10*, 10 (2014), 700.

[31] Petersen, A. M. Multiscale impact of researcher mobility. *Journal of The Royal Society Interface 15*, 146 (2018), 20180580.

[32] Petersen, A. M., Majeti, D., Kwon, K., Ahmed, M. E., and Pavlidis, I. Cross-disciplinary evolution of the genomics revolution. *Science Advances 4*, 8 (2018), eaat4211.

[33] Petersen, A. M., Majeti, D., Kwon, K., Ahmed, M. E., and Pavlidis, I. Cross-disciplinary evolution of the genomics revolution. *Science Advances 4*, 8 (2018), eaat4211.

[34] Petersen, A. M., Rotolo, D., and Leydesdorff, L. A triple helix model of medical innovation: supply, demand, and technological capabilities in terms of medical subject headings. *Research Policy 45*, 3 (2016), 666–681.

[35] Petersen, S. E., and Posner, M. I. The attention system of the human brain: 20 years after. *Annual Review of Neuroscience 35* (2012), 73–89.

[36] Poo, M.-m., Du, J.-l., Ip, N. Y., Xiong, Z.-Q., Xu, B., and Tan, T. China Brain Project: Basic neuroscience, brain diseases, and brain-inspired computing. *Neuron 92*, 3 (2016), 591–596.

[37] Quaglio, G., Corbetta, M., Karapiperis, T., Amunts, K., Koroshetz, W., Yamamori, T., and Draghia-Akli, R. Understanding the brain through large, multidisciplinary research initiatives. *The Lancet Neurology 16*, 3 (2017), 183–184.

[38] Radicchi, F., Fortunato, S., and Castellano, C. Universality of citation distributions: Toward an objective measure of scientific impact. *Proceedings of the National Academy of Sciences 105* (2008), 17268–17272.

[39] Rasmussen, I., and Ludvigsen, S. The hedgehog and the fox: A discussion of the approaches to the analysis of ICT reforms in teacher education of Larry Cuban and Yrjö Engeström. *Mind, Culture, and Activity 16*, 1 (2009), 83–104.

[40] Toiviainen, H. Inter-organizational learning across levels: An object-oriented approach. *The Journal of Workplace Learning 19*, 6 (2007), 343–358.

[41] Van Dijk, K. R., Sabuncu, M. R., and Buckner, R. L. The influence of head motion on intrinsic functional connectivity MRI. *Neuroimage 59*, 1 (2012), 431–438.

[42] Van Rijnsoever, F. J., and Hessels, L. K. Factors associated with disciplinary and interdisciplinary research collaboration. *Research Policy 40*, 3 (2011), 463–472.

[43] Wagner, C. S., Whetsell, T. A., and Mukherjee, S. International research collaboration: Novelty, conventionality, and atypicality in knowledge recombination. *Research Policy 48*, 5 (2019), 1260–1270.

[44] Wuchty, S., Jones, B. F., and Uzzi, B. The increasing dominance of teams in production of knowledge. *Science 316*, 5827 (2007), 1036–1039.

# Appendices

## A Shifts in SA and CIP Portfolios in the Decade of Multi-national Human Brain Flagship Projects

**Figure 5**(A) shows the relative frequency $f^{<}_{R,CIP}$ ($f^{>}_{R,CIP}$) by region, calculated in the 5-year period before (<) and after (>) the HB flagship project ramp-up year 2014. Each $f_{R,CIP}$ value represents the average $\overrightarrow{CIP}_p$ vector calculated across all articles belonging to a particular region, and normalized to unity to facilitate comparison, i.e., $\sum_{CIP=1}^{9} f_{R,CIP} = 1$.

In both the pre-2014 period [2009-2013] and post-2014 period [2014-2018], the most prominent disciplines are Neurosciences [CIP 1] and Medical Specialty [5] in the North American (NA) and European (EU) regions. The Australasian (AA) region shows higher levels of scholars from disciplines in Engineering & Informatics [8] and Chemistry & Physics & Math [9] than their NA and EU counterparts in the pre-2014 period. However, after 2014 we observe a realignment of AA with the remarkably similar NA and EU profiles. This realignment is achieved by decreases in Engineering & Informatics [8] and Chemistry & Physics & Math [9], and increases in Neurosciences [1] & Medical Specialty [5]. **Figure 5**(B) shows these relative shifts calculated as the difference $\Delta f_{R,CIP} = f^{>}_{R,CIP} - f^{<}_{R,CIP}$. Overall, there appears to be a remarkable synchrony in the direction and magnitude of $\Delta f_{R,CIP}$ for the NA and EU regions, primarily associated with decreases in Neurosciences [1] and Pathology & Pharmacology [7] and increases in Psychology [3] and Medical Specialty [5]. NA is the only region showing increase in both Science & Engineering domains [CIP 8&9].

Similarly, **Figure 6**(A) shows the analog frequencies $f^{<}_{R,SA}$ ($f^{>}_{R,SA}$) for each SA by region. In the pre-2014 period, the most prominent SA categories are Anatomy & Organisms [SA 2] and Health [4], with all regions showing similar distribution profiles. The most prominent distinction in AA is a reduced prominence of Psychiatry & Psychology [1]. By and large, the profiles remain consistent in the post-2014 period, with AA and NA experiencing prominent increases in Health

[4], and AA showing a modest increase in Psychiatry & Psychology [1], which nevertheless does not fully compensate for the initial deficit in this category with respect to both NA and EU.

**Figure 6**(B) indicates that all regions experienced a consistent decline in research involving the *structure*-oriented topics associated with Anatomy & Organisms [2], as well as the *function*-oriented topics associated with Phenomena & Processes [3]. The most prominent distinction between regions is for NA and EU, which both feature increases in Technology & Information Science [6] that are relatively larger than observed for AA and World, likely reflecting the technological capacity related to the tech hubs in these regions; another distinction relates to the Psychiatry & Psychology [SA 1] which increases in EU and AA more than for NA and World; and also for Health [4] which increases in NA and AA more than for EU and World.

# B Levels and Changes in SA and CIP Co-occurrence Before and After 2014

We also seek to identify which category pairs are frequently combined in articles, and to assess their frequency shifts after 2014. To this end, we introduce a tensor-product method to readily measure SA an CIP co-occurrence statistics for the purpose of identifying particular cross-domain orientations observed in cross-domain HB science.

In order to juxtapose the relative frequencies of mono-category articles separately from multi-category articles, we define a modified outer-product matrix designed purely for visualization purposes:

$$\tilde{\mathbf{D}}_p(\vec{v}_p) \equiv \begin{cases} \frac{U(\Upsilon)}{||U(\Upsilon)||}, & \text{if } \vec{v}_p \text{ contains 2 or more categories.} \\ \\ \frac{U(\vec{v}_p \otimes \vec{v}_p)}{||U(\vec{v}_p \otimes \vec{v}_p)||} = \text{DiagonalMatrix}(\text{Sign}[\vec{v}_p]), & \text{otherwise.} \end{cases} \tag{3}$$

where $\otimes$ is the outer tensor product and $\circ$ indicates the element-wise or Hadamard product. Note

that this definition is slightly different than $\mathbf{D}_p(\vec{v}_p)$ defined in Eq.(1). The difference occurs in the first case, for which the matrix $\boldsymbol{\Upsilon} \equiv \vec{v}_p \otimes \vec{v}_p - \vec{v}_p \circ \mathbb{1} \circ \vec{v}_p$ for which the diagonal elements are eliminated via subtraction, i.e., $\mathrm{Tr}(\boldsymbol{\Upsilon}) = 0$.

Simply stated, when $\vec{v}_p$ (representing $\overrightarrow{SA}_p$ or $\overrightarrow{CIP}_p$) has 2 or more non-zero elements then we primarily count the off-diagonal elements of the outer-product matrix and are not concerned with the relative frequencies of the on-diagonal elements. Contrariwise, in the case that there is just one category present – e.g. $\vec{v}_p = \{0, 3, 0, 0, 0, 0\}$ – then we track only the diagonal element, which counts the occurrence of the single category. In this second case, the resulting matrix $\tilde{\mathbf{D}}_p(\vec{v}_p) = \mathrm{DiagonalMatrix}(\mathrm{Sign}[\vec{v}_p])$ has only one non-zero element, which occurs for the diagonal value $\tilde{D}_{22} = 1$; and all other matrix elements = 0. Note that in either case the total sum of all elements are normalized to unity, $||\tilde{\mathbf{D}}_p(\vec{v}_p)|| = 1$. This normalization implies that totaling $\tilde{\mathbf{D}}_p(\vec{v}_p)$ across articles from a given publication year yields the total number of articles, $N(t)$.

We then calculated the aggregate co-occurrence matrix, denoted by $\mathbf{C}^{<} = \sum_{y_p \in [2009-2013]} \tilde{\mathbf{D}}_p$, using all articles published in the pre-period. It then follows from our normalization procedure that the total across all matrix elements is proportional to the total number of articles published in a given period, i.e., $||\mathbf{C}^{<}|| = \sum_{t=2009}^{2013} N(t) = N_{[2009-2013]}$. **Figure 5**(C) and **Figure 6**(C) show $\mathbf{C}^{<}_{CIP}$ and $\mathbf{C}^{<}_{SA}$, respectively.

To measure relative changes, we then calculated the percent difference in each matrix element $\Delta C_{ij} = 200(C^{>}_{ij} - \theta C^{<}_{ij})/(C^{>}_{ij} + \theta C^{<}_{ij})$, where $\theta = N_{[2014-2018]}/N_{[2009-2013]}$ corrects for bias associated with differences in the number of articles published each of the pre- and post-periods. To illustrate why this correction is important, we randomized the counts contained in $\vec{v}_p = \overrightarrow{SA}_p$ and plot the resulting $\mathbf{C}^{<}_{\mathrm{rand.},SA}$ and $\boldsymbol{\Delta}\mathbf{C}_{\mathrm{rand.},SA}$ matrices in **Figure A1**. As anticipated, this randomization scheme eliminates the variation among on-diagonal elements and off-diagonal elements in panel (A); Moreover, in panel (B) the off-diagonal elements all show percent change values that are in the range of ±3%, thereby indicative of the threshold for distinguishing statistically significant percent changes in the real data.

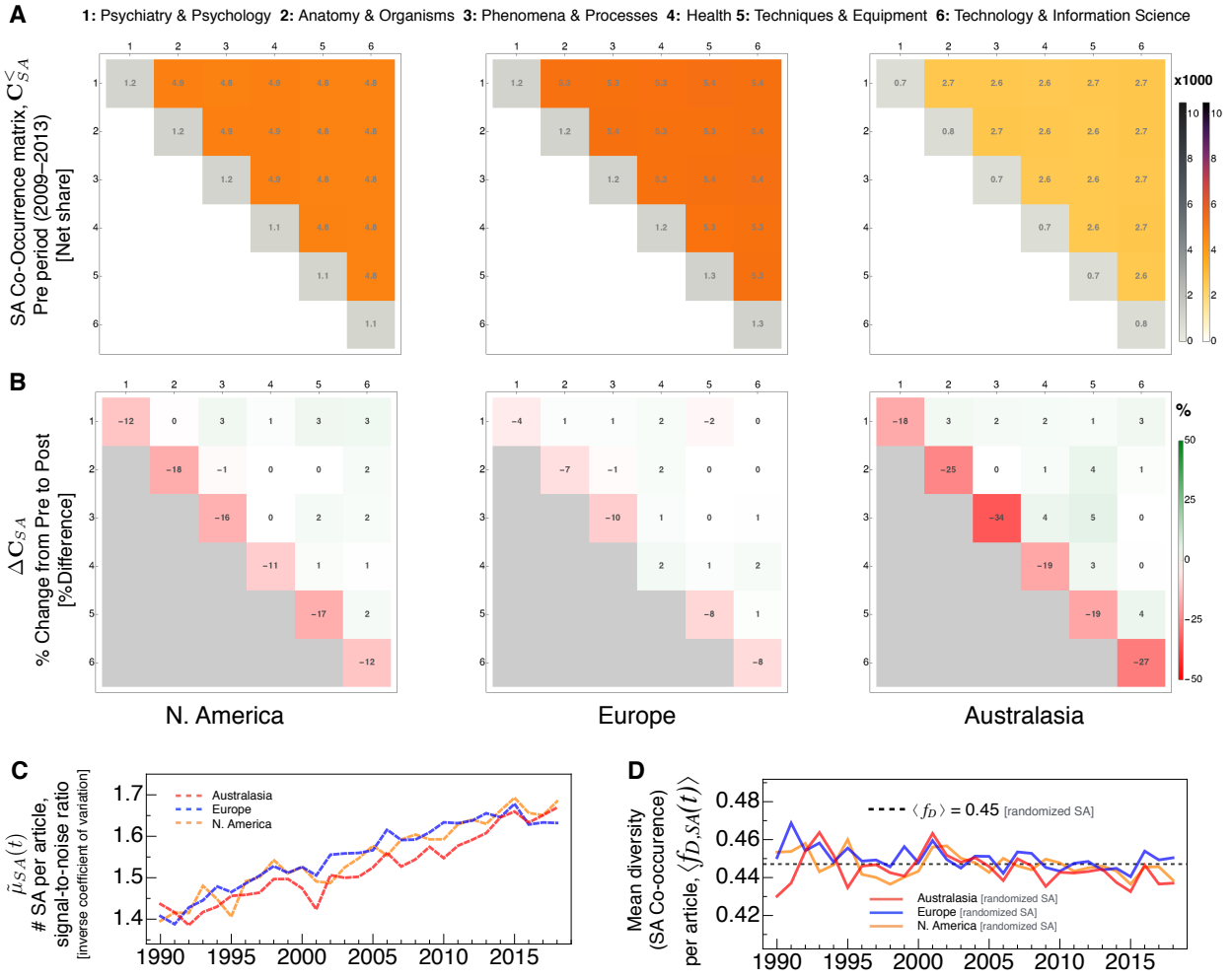



Figure A1: **Co-occurrence metrics appropriately account for secular growth in research output and MeSH annotation.** Demonstration of consistent co-occurrence metrics calculated using randomized category vectors, $\vec{v}_{p,\text{rand.}}$, with vector elements shuffled such that the total counts $|\vec{v}_p|$ for each $p$ is conserved. (A) Elimination of variation among the diagonal and off-diagonal elements of the co-occurrence matrix $\mathbf{C}^{<}_{SA,\text{rand.}}$ indicate that no other significant statistical biases underly this co-occurrence calculation. (B) Reduction of the relative change between two shuffled co-occurence matrices $\mathbf{C}^{<}_{SA,\text{rand.}}$ and $\mathbf{C}^{>}_{SA,\text{rand.}}$ to the level of noise; The largest off-diagonal value observed is 3%, representing a threshold for classifying significant shifts in the corresponding real data shown in Figs. 5(D) and 6(D). (C) Increase in the signal-to-noise ratio $\tilde{\mu}_{SA}(t) = \mu_{SA}(t)/\sigma_{SA}(t)$, measuring the average number of SA per article ($\mu_{SA}$), normalized by the standard deviation ($\sigma_{SA}$). Increase in the number of SA (and MeSH) per article is a source of secular growth that could introduce temporal bias challenging the interpretation of the results. (D) Accounting for this secular growth of $\tilde{\mu}_{SA}(t)$ yields a mean diversity per article $\langle f_{D,SA,\text{rand.}}$ which is approximately constant over time, consistent with the expected results for randomized category vectors, $\vec{v}_{p,\text{rand.}}$.

Returning to the real data and the calculation of $\mathbf{C}^{<}_{CIP}$, the most notable results of this visualization are the consistently strong couplings between CIP category [1] and all other categories [2,3,4,5,6]; between categories [1,2,3] and [5]; and also between categories [1,2,5] and [6]. Also of note is the higher-order clique among [1,5,6] where each CIP is strongly coupled to each other. Contrariwise, we observe relatively weak coupling between [7,8,9] and most all other CIP.

Other prominent CIP that couple by region: NA shows relatively higher coupling between [1,4] and [4,5] and [5,8] compared to other regions; and EU shows relatively higher coupling between [1,9] and [2,9]. Regarding the shifts from the pre- to post-2014 captured by $\Delta\mathbf{C}_{CIP}$, NA and EU regions show consistent increase in CIP pairs [4,7] and [4,9], [3,8] and [2,7]; and consistent decrease between [1,8] and [2,9] and [6,9] and all combinations between 5 and [7,8,9]. Notably, AA exhibits higher % change levels, following from the fact that several elements in $\mathbf{C}^{<}_{CIP}$ that are nearly 0.

**Figure 6**(C) shows $\mathbf{C}^{<}_{SA}$ calculated by region. For all regions, the matrix elements corresponding to SA pairs [2,3] and [2,4] and [3,4] are relatively strong, thus forming another clique among these SA representing traditional branches of biology. Other strongly coupings are SA [4] with both [1,5]; and [5] with [2,4,6]. As with $\mathbf{C}^{<}_{CIP}$, the technologically-oriented SA are the most weakly coupled categories.

Regarding the shifts from the pre-2014 to post-2014 captured by $\Delta\mathbf{C}_{SA}$, the most consistent increases are between SA [1] and each of [2,4,6]; and between 4 and both [5,6]. Contrariwise, the most consistent decreasing coupling is between [3] and both [2,6], and between [5,6]. The matrices for NA and EU are rather similar, with the most prominent distinction between [2,6] – showing a -12% change for NA and a +5% change for EU; and also between [3,6] – showing also a -12% decrease for NA but no significant change for EU. This latter disparity is an example of where EU may be taking the lead in in-silico-oriented approaches to HB science, consistent with the framing of the Human Brain Project.

The most notable distinction for AA relative to NA and EU is in the larger magnitude of shifts, representing a period of international convergence for all couplings involving SA [1], and in particular between [1,2] and between [1,6]; contrariwise for AA, there is a prominent decoupling

between SA [5,6] which is consistent with the relative shifts away from these two SA to compensate for the prominent redirection towards [1] and [4], as also indicated by **Figure 6**(B).

# C Calculation of Cross-domain Co-occurrence: An Illustrative Example of the Tensor Product

Take for example an article $p$ with 4 metadata entities belonging to 3 categories, $\vec{v}_p = \{1, 2, 0, 0, 1, 0\}$. Calculation of the co-occurrence matrix $\mathbf{D}_p(\vec{v}_p)$ using the normalized outer-product defined in Eq.(1) yields

$$\mathbf{D}_p(\vec{v}_p) = \begin{pmatrix} \frac{1}{11} & \frac{2}{11} & 0 & 0 & \frac{1}{11} & 0 \\ 0 & \frac{4}{11} & 0 & 0 & \frac{2}{11} & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \frac{1}{11} & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

with $||U(\vec{v}_p \otimes \vec{v}_p)|| = 11$. The categorical diversity is calculated as the total across off-diagonal elements, $f_{D,p} = 1 - \mathrm{Tr}(\mathbf{D}_p) = 5/11$.

For completeness, consider the representation of a mono-disciplinary article with the same number of metadata entities that all fall into the second category, $\vec{v}_p = \{0, 4, 0, 0, 0, 0\}$. Then

$$\mathbf{D}_p(\vec{v}_p) = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix} \text{ with corresponding } f_{D,p} = 1 - \mathrm{Tr}(\mathbf{D}_p) = 1 - 1 = 0.$$

# D Historical Trends in SA & CIP Diversity: 2000-2018

We investigate historical trends in SA & CIP diversity using the matrix $\mathbf{D}_p$ defined in Eq. (1), which simultaneously measures mono-dimensional and multi-dimensional features of each article. More specifically, we define $f_{D,p} = 1 - \mathrm{Tr}(\mathbf{D}_p)$ as the fraction of the article's co-occurrence matrix capturing combinatorial diversity. Hence, in the limiting case that the article features just a single category, then $f_{D,p} = 0$; and when all categories are present in equal quantities then $f_{D,p} = (d-1)/(d+1)$, where $d$ is the dimension of the categorical vector $\vec{v}_p$. As $d$ increases then $f_{D,p}$ approaches 1. Hence, for sufficiently large $d$ then $0 \leq f_{D,p} \lesssim 1$. **Figures A2**(D,E) show the unconditional distributions, $P(N_{SA})$ and $P(N_{CIP})$, with observed values spanning across the full range $d = 6$ and $d = 9$, respectively.

As a bounded quantity, the average article-level diversity $\overline{f}_D(t) = N(t)^{-1} \sum_{p \in N(t)} f_{D,p}$ is an appropriate measure of a characteristic article, where $N(t)$ is the number of articles being considered from year $t$. However, $\overline{f}_D(t)$ is nevertheless sensitive to bias associated with a systematic increase over time in $\langle N_{CIP,t} \rangle$ and $\langle N_{SA,t} \rangle$, the average number of categories present per article per year. We

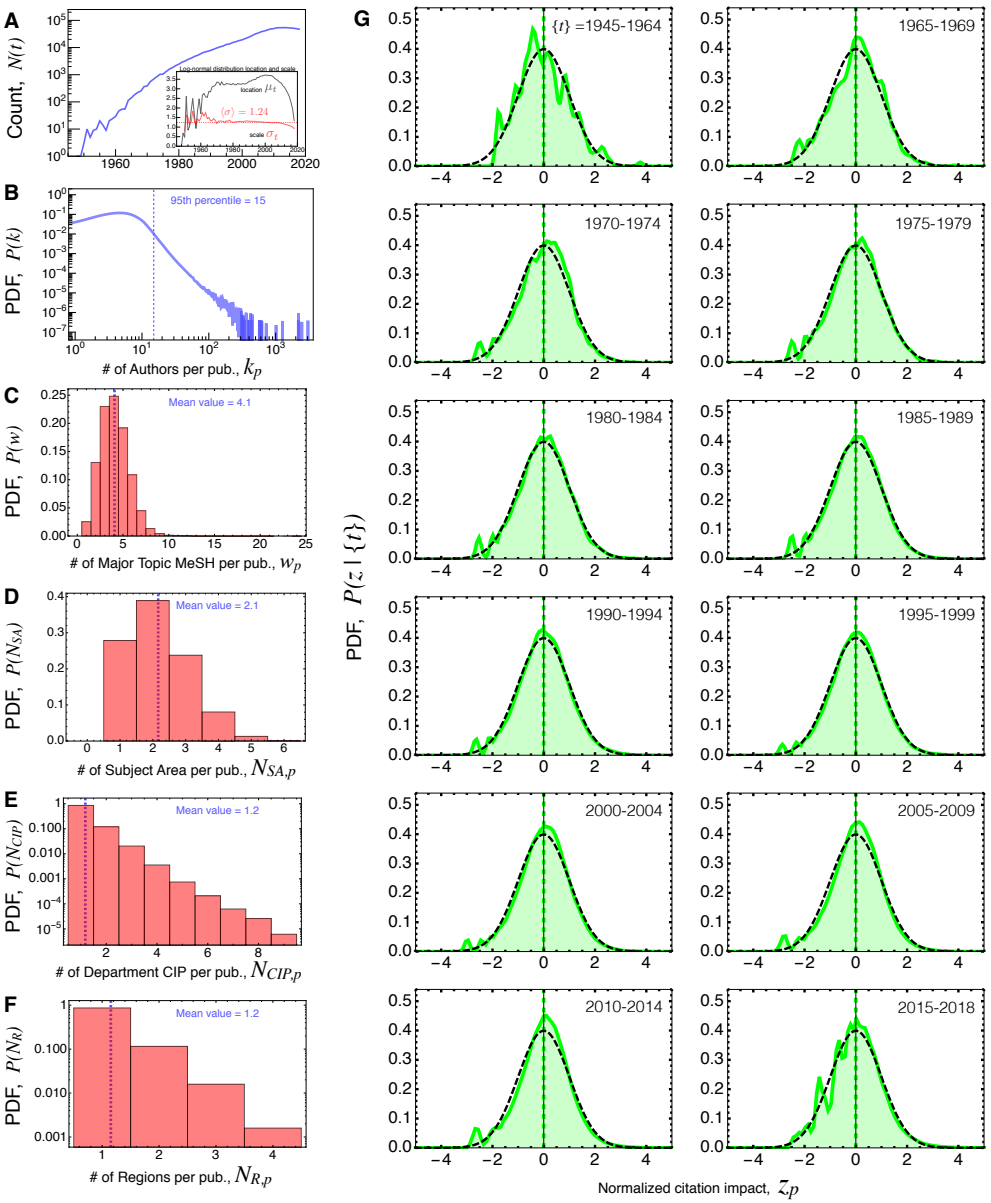



Figure A2: **Distributions of article-level variables.** (A) $N(t)$ is the number of HB articles by year. (B) $P(k)$ is the probability distribution (PDF) of the number of coauthors per article. (C) $P(w)$ is the PDF of the number of Major Topic MeSH "keywords" per publication, denoted by $w_p$. (D) Each MeSH keyword maps onto one of the 6 SA clusters. Shown is the PDF of the number of distinct SA categories per publication, $N_{SA,p}$. (E) Each departmental affiliation maps onto one of the 9 CIP clusters. Shown is the PDF of the number of distinct CIP categories per publication, $N_{CIP,p}$. (F) Each Scopus Author's affiliation maps onto one of 4 regions: Australasia, Europe, North America, and (rest of) World. Shown is the PDF of the number of region categories per publication, $N_{R,p}$. (G) Probability distribution (PDF) of $z_p$ disaggregated by publication cohort $\{t\}$; each green curve represents the smoothed kernel density estimate of the $P(z)$, calculated with kernel bandwith = 0.1. Data are split into 5-year periods from 1965-2018, with the first panel including data from 1945-1964.

address this issue by applying a temporal deflator which adjusts the annual averages to account for systematic shifts in the underlying data generating process. To be specific, we define $\langle f_{D,SA}(t)\rangle = \overline{f}_{D,SA}(t) \times [\langle\tilde{\mu}_{SA}\rangle/\tilde{\mu}_{SA}(t)]$, where $\tilde{\mu}_{SA}(t) = \langle N_{SA,t}\rangle/\sigma_{N_{SA,t}}$ is the inverse coefficient of variation (also called the signal-to-noise ratio) with respect to the number of SA per article, represented by $N_{SA,p}$; and $\langle\tilde{\mu}_{SA}\rangle$ is the average value calculated across the roughly 3 decades of analysis. **Figure A1**(C) shows that $\tilde{\mu}_{SA}(t)$ is increasing steadily with time. Hence, adjusting for this secular growth is essential so that observed increases are not simply artifacts of the underlying growth in $N_{SA,p}$ or $N_{CIP,p}$. We apply the same method to adjust for systematic shifts in $\langle N_{CIP,t}\rangle$.

To illustrate the utility of this deflator method, we randomized the SA for all articles (by randomly shuffling the counts in each $\overrightarrow{CIP}_p$ or $\overrightarrow{SA}_p$). **Figure A1**(D) demonstrates that there is no trend in the corresponding $\langle f_{D,SA}(t)\rangle$, indicating that this method removes the underlying bias.

Returning to the empirical data, **Figure A3** shows the evolution of disciplinary diversity captured by coauthors' departmental affiliations. Each panel shows $\langle f_{D,CIP}(t)\rangle$ calculated for a specified combination of categories contained in each $\overrightarrow{CIP}_p$ vector, as indicated by the schematic motif provided alongside each panel. For example, **Figure A3**(A) calculates $\langle f_{D,CIP}(t)\rangle$ from all 9 CIP categories considered independently, whereas **Figure A3**(B) collects the counts associated with the combined categories [1-4] and [5-9] and calculates the diversity based upon the fraction of $\mathbf{D}_p$ belonging to the single off-diagonal element $D_{12,p}$, which records the disciplinary mixing between these two supergroups.

**Figure A3**(A) is calculated using the *Broad* configuration, and exhibits a slow increase in CIP diversity from 1990 to the mid 2000s in North America (NA) and European (EU) regions, which stalled thereafter, and even declined in the last decade for NA and AA, but not for EU. **Figure A3**(B) shows relatively lower levels and trends in the diversity at the intersection of supercategories [1-4] (representing traditional neuro/biology departments) and [5-9] (representing all other CIP jointly). By way of comparison, this trend indicates that the decline in panel (A) is not derived from the intersection explored in panel (B). Instead, **Figures A3**(C,D) indicate that the decline in (A) is attributable to declines at the individual intersections between all permutations of

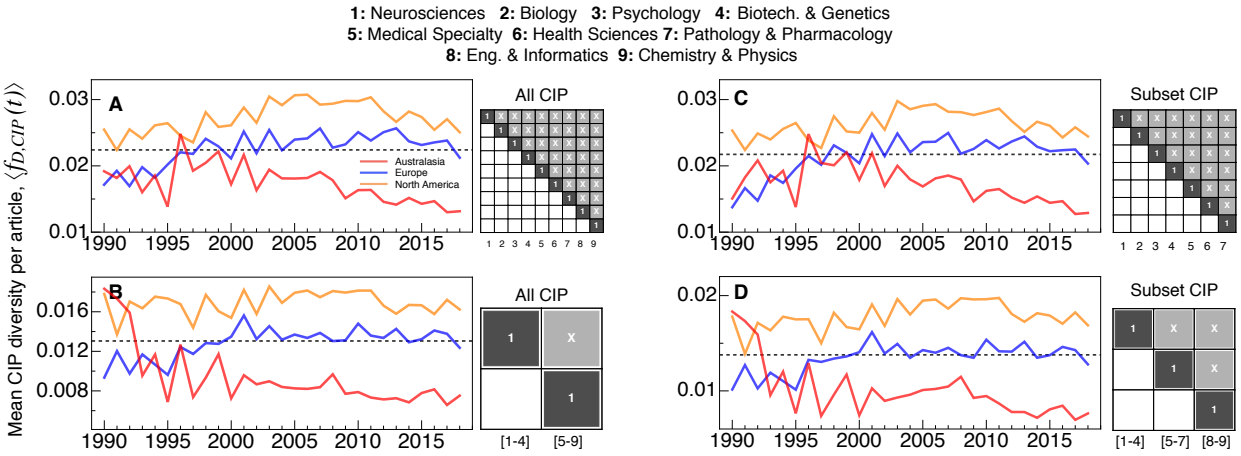



Figure A3: **Trends in cross-disciplinary (CIP) scholarship in human brain science.** Each curve corresponds to $\langle f_{D,CIP}(t) \rangle$, representing the average article diversity measured as categorical CIP co-occurrence in the off-diagonal matrix elements of $\mathbf{D}_{CIP,p}$, see Eq. (1); each curve is calculated for articles belonging to a given geographic region, as determined by the coauthors' regional affiliations: Australasia (red), Europe (blue), and North America (orange). For each panel we provide a matrix motif indicating the set of focal CIP categories; counts for categories included in brackets are considered in union. For example, whereas panel (A) calculates $\langle f_{D,CIP}(t) \rangle$ across all 9 CIP categories (each category considered separately); instead, panel (B) calculates each $\mathbf{D}_p$ by considering just two super-groups, the first consisting of the union of CIP counts for categories [1-4], and the second comprised of categories [5-9].

CIP categories 1-7, and to a lesser extent between the three disciplinary subdomains: neuro/biology [1-5], health [5-7] and science and engineering [8-9]. Overall, we also observe higher levels of CIP diversity in NA, followed by EU, and then followed by AA.

Likewise, **Figure A4** shows the evolution of research topic diversity captured by SA counts in each $\overrightarrow{SA}_p$. We observe much stronger trends for SA, suggesting that scholars tend to also cross disciplines as mono-disciplinary teams rather than via cross-disciplinary collaboration. **Figure A4**(A) shows $\langle f_{D,SA}(t) \rangle$ calculated for the *Broad* configuration which includes all SA categories. The diversity trend is increasing since 1990 for all regions, but with reduced pace since the early 2010s. Similar to our findings for CIP, we observe AA lagging the other two regions; however, in this case of SA we do observe more similar levels of diversity between EU and NA. **Figure A4**(B) indicates that much of the increase in SA diversity is attributable to research combining Health [SA 4] and the other categories – in other words, the domain of health science appears to be a persistent driving force behind convergence trends. Supporting evidence for this observation is also captured in the hierarchical clustering of SA represented by the minimum spanning tree (MST) representation of the aggregate SA co-occurrence matrix $\tilde{\mathbf{D}}_{SA,p}$ – see **Figure 2**(B). By way of comparison, the analog MST representation of $\tilde{\mathbf{D}}_{CIP,p}$ in **Figure 2**(D) features a less prominent hierarchy across the CIP categories.

We analyzed several additional SA category subsets and super-category combinations to more deeply explore the anatomy of research topic diversity. **Figure A4**(C,D) show that increasing diversity associated with Health [4] is largely captured via the incorporation of technology- and informatics-oriented capabilities [5,6] – as opposed to integrating more traditional biological SA representing research domains associated with questions relating to how Anatomy & Organisms (structure) [2] and Phenomena & Processes (function) [3] relate to complex human behavior addressed by Psychiatry & Psychology [1] – as illustrated in **Figure A4**(E).

Similarly, a significant component of the increasing diversity captured between SA [4,5,6] derives from the increase between research that is centered around Techniques & Equipment [5] and Technology & Information Science [6]; although this contribution shown in **Figure A4**(F) only

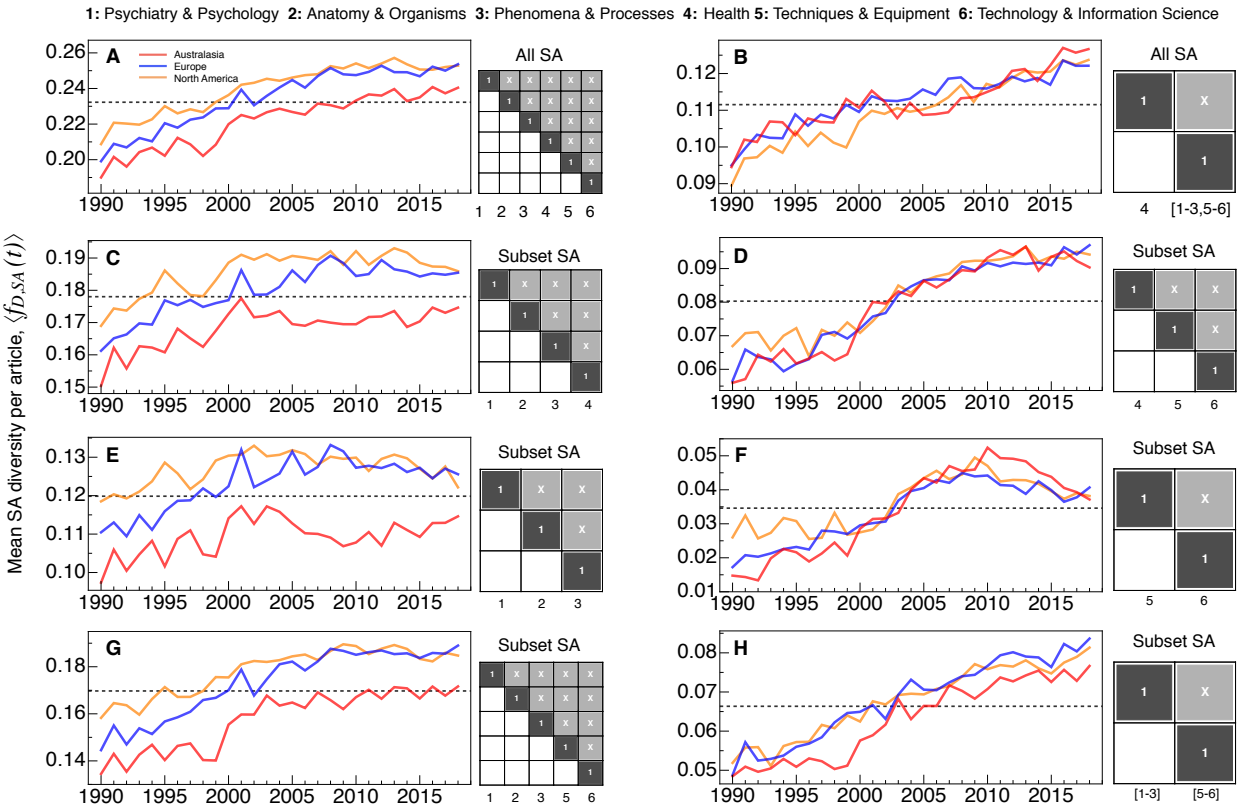



Figure A4: **Trends in cross-topical (SA) scholarship in human brain science.** Each curve corresponds to $\langle f_{D,SA}(t) \rangle$, representing the average article diversity measured as categorical SA co-occurrence in the off-diagonal matrix elements of $\mathbf{D}_{SA,p}$, see Eq. (1); each curve is calculated for articles belonging to a given geographic region, as determined by the coauthors' regional affiliations: Australasia (red), Europe (blue), and North America (orange). For each panel we provide a matrix motif indicating the set of focal SA categories; counts for categories included in brackets are considered in union. For example, whereas panel (A) calculates $\langle f_{D,SA}(t) \rangle$ across all 6 SA categories (each category considered separately); instead, panel (C) calculates each $\mathbf{D}_{SA,p}$ by considering a subset of four SA categories 1-4.

contributes to increases in diversity until 2010, after which there is a prominent decline. Interestingly, this is a configuration which emphasizes the leading role of AA since 2010 in combining these two areas. To further emphasize the role of Health, we exclude this category [4] from the diversity measures shown in **Figure A4**(G), indicating that combinations of SA across the traditional domains of biology and the technology-oriented domains have also saturated around 2010, and their contribution to SA diversity primarily appears when considered the biology [1-3] and technology-oriented [5,6] as super-clusters illustrated in **Figure A4**(H).

# E Panel Regression: Model Specification

We constructed article-level and author-level panel data to facilitate measuring factors relating to SA and CIP diversity and shifts related to the ramp-up of three regional HB flagship projects circa 2013, and several others thereafter. **Figure A2** shows the distribution of various article-level features; and **Figure A5** shows the covariation matrix between the principle variables of interest.

We use the following operator notation to specify how we classify articles as being cross-domain ($X$) or mono-domain ($M$). Starting with the feature vector $\vec{F}_p \equiv \{\overrightarrow{SA}_p, \overrightarrow{CIP}_p, \overrightarrow{R}_p\}$, we obtain a binary diversity classification for each article denoted by $X$ and $M$. We specify the objective criteria of the feature operator $O$ by its subscript. For example, $O_{SA}(\vec{F}_p) = X_{SA}$ if two or more SA categories are present, otherwise the value is $M$; and by analogy, $O_{CIP}(\vec{F}_p) = X_{CIP}$ if two or more categories are present, and otherwise $O_{CIP}(\vec{F}_p) = M$. In the case of models oriented around articles featuring $X_{SA}$ and $X_{CIP}$ simultaneously (represented by $O_{SA\&CIP}(\vec{F}_p) = X_{SA\&CIP}$), we exclude the set of articles classified as $X_{SA}$ but not $X_{CIP}$ and those classified as $X_{CIP}$ but not $X_{SA}$. Hence, in what follows, the counterfactual baseline group for $X_{SA\&CIP}$ articles are also the subset of mono-domain articles, which facilitates comparison of effect sizes across models oriented around $X_{CIP}$, $X_{SA}$ and $X_{SA\&CIP}$.

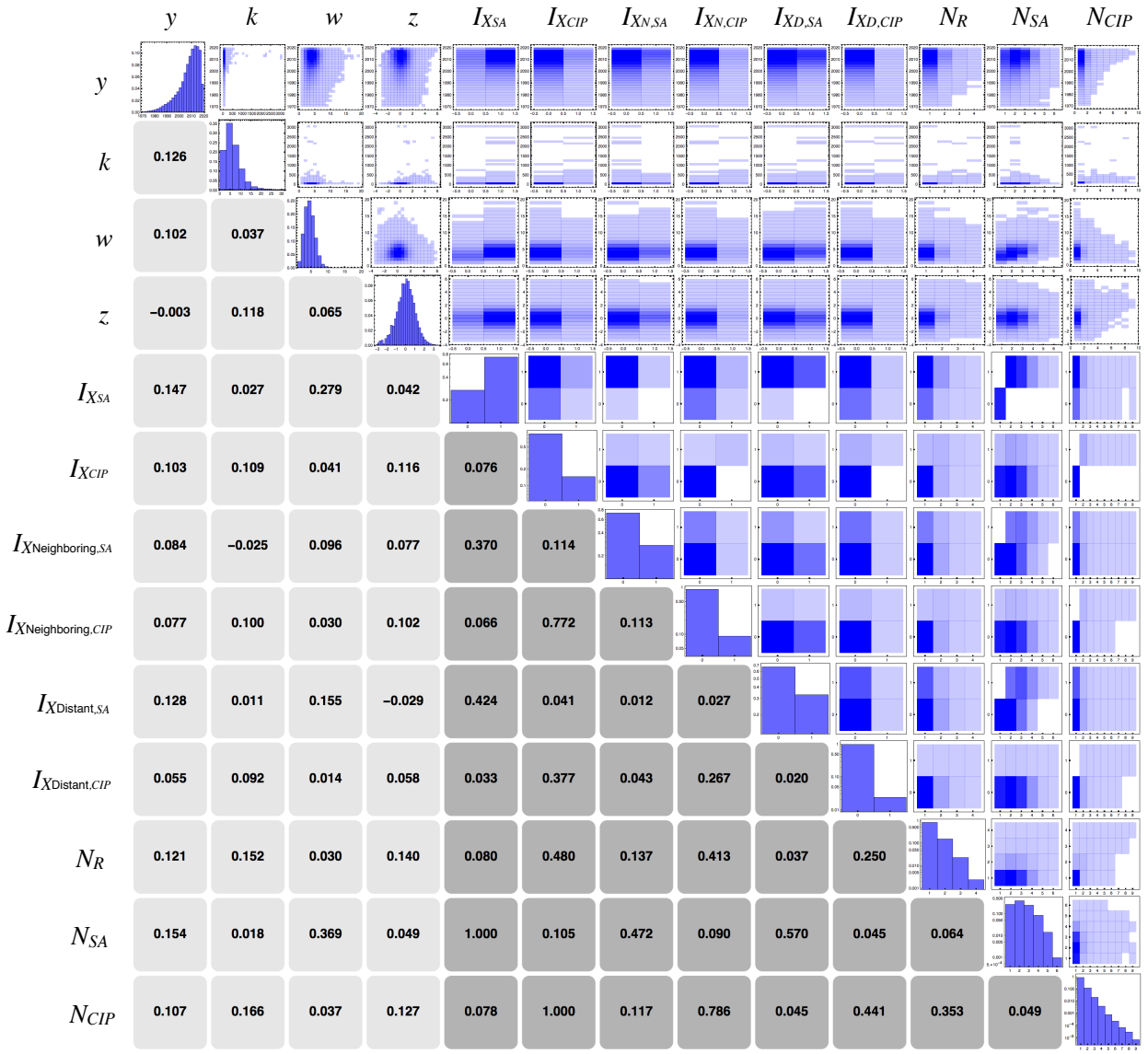



Figure A5: **Cross-correlation and descriptive statistics for regression model variables.** Upper-diagonal elements: bivariate histogram between row and column variables. Diagonal elements: histogram for variable indicated by the row/column labels. Lower-diagonal elements: bivariate cross-correlation coefficient: light-shaded squares indicate the Pearson's correlation coefficient between two variables that are both continuous measures; dark-shaded squares indicate the Cramer's V associate between two variables that are both nominal (categorical).

## E.1 Article-level Model

### E.1.1 Quantifying Factors Associated with Propensity for CIP and SA Diversity

In the first model, we seek to better understand the factors associated with the prevalence of CIP and SA diversity as they evolve over time, and in particular their relation to the launching of the HB flagship programs. In order to model the article-level factors (indicated by $p$) associated with cross-domain research activity we define the binary indicator variable generically denoted as $I_{X,p}$.

By way of example, if we are considering SA diversity, then the indicator variable $I_{X_{SA,p}}$ takes the value 1 if $O_{SA}(\vec{F}_p) = X_{SA}$ and 0 if $O_{SA}(\vec{F}_p) = M$. We then model the 2-state odds $Q \equiv \frac{P(O_{SA}(\vec{F}_p)=X_{SA})}{P(O_{SA}(\vec{F}_p)=M)} = \frac{P(X_{SA})}{P(M)}$, which represents the propensity for cross-domain research, where $P(X_{SA}) + P(M) = 1$. Likewise, in the case of CIP diversity we model the odds as $Q \equiv \frac{P(O_{CIP}(\vec{F}_p)=X_{CIP})}{P(O_{CIP}(\vec{F}_p)=M)}$; and finally, we also consider the likelihood of research featuring both types of cross-domain activity, for SA & CIP, represented as $Q \equiv \frac{P(O_{SA\&CIP}(\vec{F}_p)=X_{SA\&CIP})}{P(O_{SA\&CIP}(\vec{F}_p)=M)}$. Because all Scopus scholars map onto a single CIP, and since this model is primarily concerned with identifying factors associated with orientation towards cross-domain research, we exclude solo-authored research papers (i.e., those with $k_p = 1$) from this analysis since the likelihood for those articles is predetermined (i.e., $P(M) = 1$); for the same reason, we also exclude articles with a single Major MeSH category (i.e., those with $w_p = 1$).

For each article we also include several covariates of $I_{X,p}$: the article publication year $y_p$; the mean journal citation impact, calculated as the average $z_p$ for articles from journal $j$, denoted by $\overline{z}_j = \langle z_p | \text{ journal j} \rangle$; the natural logarithm of the total number of coauthors, $\ln k_p$; and the natural logarithm of the total number of Major MeSH terms, $\ln w_p$. As additional controls, we also include the total number of international regions associated with the authors' affiliations $N_{R,p}$, and also the total number of categories featured by the article, $N_{SA,p}$ and $N_{CIP,p}$.

We then model the odds $Q$ by way of a Logit regression model, specified in the case of $X_{SA}$ as

$$\text{Logit}\Big(P(X_{SA})\Big) = \log\Big(\frac{P(X_{SA})}{P(M)}\Big) = \beta_0 + \beta_y y_p + \beta_{\overline{z}}\overline{z}_p + \beta_k \ln k_p + \beta_w \ln w_p + \beta_{N_R} N_{R,p} + \beta_{N_{CIP}} N_{CIP,p} + \epsilon \; ; \quad (4)$$

in the case of $X_{CIP}$ as,

$$\text{Logit}\Big(P(X_{CIP})\Big) = \log\Big(\frac{P(X_{CIP})}{P(M)}\Big) = \beta_0 + \beta_y y_p + \beta_{\overline{z}}\overline{z}_p + \beta_k \ln k_p + \beta_w \ln w_p + \beta_{N_R} N_{R,p} + \beta_{N_{SA}} N_{SA,p} + \epsilon \; ; \quad (5)$$

and in the case of $X_{SA\&CIP}$ as,

$$\text{Logit}\Big(P(X_{SA\&CIP})\Big) = \log\Big(\frac{P(X_{SA\&CIP})}{P(M)}\Big) = \beta_0 + \beta_y y_p + \beta_{\overline{z}}\overline{z}_p + \beta_k \ln k_p + \beta_w \ln w_p + \beta_{N_R} N_{R,p} + \epsilon \; . \quad (6)$$

To account for errors that are geographically correlated over time, we estimated the model using robust standard errors clustered on a regional categorical variable. The full set of parameter results are tabulated in models (1)-(3) in **Tables 3-5**, which report the exponentiated coefficients. To be specific, the exponentiated coefficient $\exp(\beta)$ is the odds ratio, representing the factor by which $Q$ changes for each 1-unit increase in the corresponding independent variable, i.e., $Q_{+1}/Q = \exp(\beta)$. In real terms, $100\beta \approx 100(\exp(\beta) - 1)$ represents the percent change in $Q$ corresponding to a 1-unit increase in the corresponding independent variable (where the approximation holds for small $\beta$ values). As a result, $\exp(\beta)$ values that are less than (greater than) unity indicate variables that negatively (positively) correlate with the likelihood $P(X)$.

### E.1.2 Quantifying Shifts in Propensity for CIP and SA Diversity Associated with the Announcement of Global Flagship HB Projects Circa 2013

In order to identify shifts in the 5-year period after the 2013 ramp-up of HB projects worldwide, we incorporated an interaction between the pre-/post periods – indicated by $I_{2014+,p}$, which takes the value 1 for $y_p \geq 2014$ and 0 otherwise – and a categorical variable specifying the region, represented by $I_{R,p}$. We use the *Rest of World* region category (indicated by countries colored gray in **Figure 1**) as the baseline for regional comparison since these regions did not feature flagship HB programs on the scale of those announced in Australia, Canada, China, Japan, Europe, South Korea, and the United States.

By way of example, in the case of modeling the likelihood $P(X_{SA})$, the interaction term is added in the second row,

$$\text{Logit}\Big(P(X_{SA})\Big) = \log\Big(\frac{P(X_{SA})}{P(M)}\Big) = \beta_0 + \beta_y y_p + \beta_{\overline{z}}\overline{z}_p + \beta_k \ln k_p + \beta_w \ln w_p + \beta_{N_R} N_{R,p} + \beta_{N_{CIP}} N_{CIP,p} + \gamma_R I_{R,p} + \gamma_{2014+} I_{2014+,p} + \delta_{R+}(I_{R,p} \times I_{2014+,p}) + \epsilon \ . \quad (7)$$

To differentiate different types of model variables, $\beta$ is used to identify coefficients associated with continuous variables, $\gamma$ is used for indicator variables, and $\delta$ is used to indicate interactions between indicator variables. In particular, the coefficient $\delta_{R+}$ measures the Difference-in-Difference (DiD) estimate of the effect of HB projects on the propensity for research teams to pursue $X_{SA}$ approaches. **Figure A3** and **Figure A4** demonstrate that historical trends in the prevalence of cross-domain diversity satisfy the parallel trend assumption for both CIP and SA, respectively. The full set of parameter results are tabulated in models (4)-(6) in **Tables 3-5**, and the point estimates for principal test variables are visually summarized in **Figure A6**.

## E.2 Author-level Model

In the second model, we seek to measure the relation between the two different types of article diversity – CIP and SA – and the article's scientific impact, proxied by $c_p$. Our approach leverages

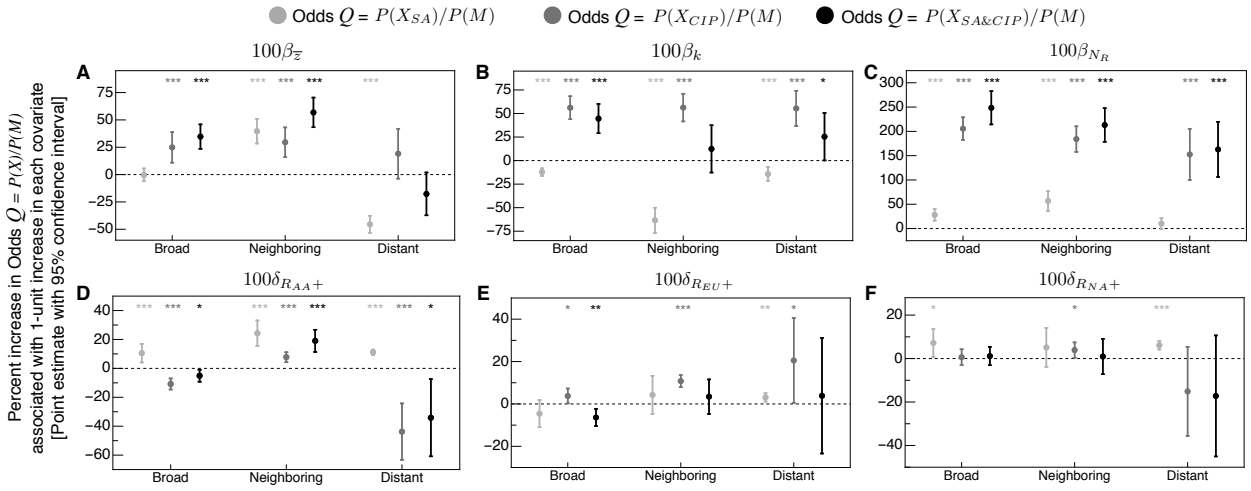



Figure A6: **Summary of Logit model parameter estimates.** (A-C) Reported are $100\beta$ for the main covariates of interest reported in Tables 3-5, quantifying the percent increase in the odds $Q \equiv P(X)/P(M)$ associated with a one-unit increases in: (A) mean journal citation impact $\overline{z}_{j,p}$; (C) $\ln k$; (B) number of coauthors, $k_p$; (C) number of major MeSH terms (keywords), $w_p$; (D-F) difference-in-difference estimates ($100\delta_{R+}$) capturing the effect of Flagship project ramp-ups after 2013 on rates of cross-domain research – at three levels of specificity regarding the diversity range captured by $X$. The *Broad* configuration correspond to unconstrained combinations of SA and CIP (represented by $X_{SA}$, $X_{CIP}$, $X_{SA\&CIP}$). The *Neighboring* configuration corresponds to specific set of category combinations capturing the neurobiological -vs- bioengineering interface, represented by SA [1] $\times$ [2-4] and CIP [1,3] $\times$ [2,4-7] (and represented by $X_{\text{Neighboring},SA}$, $X_{\text{Neighboring},CIP}$, $X_{\text{Neighboring},SA\&CIP}$). And *Distant* also identifies a specific set of category combinations capturing the neuro-psycho-medical -vs- techno- computational interface, represented by SA [1-4] $\times$ [5,6] and CIP [1,3,5] $\times$ [4,8] ($X_{\text{Distant},SA}$, $X_{\text{Distant},CIP}$, $X_{\text{Distant},SA\&CIP}$). Reported are percent increase in $Q$, a ratio representing the propensity for cross-domain research relative to mono-domain research, directly associated with the ramp-up of Brain projects in: (D) Australasia; (E) Europe; (F) North America. Shown are point estimates with 95% confidence interval. Standard errors clustered by region to account for residuals that are correlated within regions over time. Asterisks above each estimate indicate the associated $p-$value level: $^{*}$ $p < 0.05$, $^{**}$ $p < 0.01$, $^{***}$ $p < 0.001$.

the hierarchical features of the article-level data grouped into author-specific subgroups representing HB researcher publication portfolios. As a result, model coefficients represent estimates net of author-specific time-independent factors. In other words, this fixed-effect specification yields parameter estimates that are net of the author-specific baseline $\alpha_a = \langle z_a \rangle$, where $a$ is an author index. This specification identifies a clear counterfactual framework for identifying the different outcomes associated with $X$ and $M$ that are relevant to researcher problem identification and team-assembly strategies.

First, in order to measure relative differences in citation impact within and across publication cohorts, we apply a logarithmic transform that facilitates removing the time-dependent trend in the location and scale of the underlying log-normal citation distribution [38]. As such, the normalized citation impact defined in Eq. (2) is

$$z_p \equiv \frac{\ln(1 + c_{p,t}) - \mu_t}{\sigma_t} \ ,$$

where $\mu_t \equiv \overline{\ln(1 + c_t)}$ is the mean and $\sigma_t \equiv \sigma[\ln(1 + c_t)]$ is the standard deviation of the log-citation distribution for articles grouped by publication year. We uniformly add 1 to each $c_{p,t}$ count to avoid the divergence $\ln 0$ associated with uncited publications, a common method that does not alter the interpretation of our results. Importantly, the standard deviation $\sigma_t \approx \langle \sigma \rangle = 1.24$ is approximately constant over the focal period of our analysis. Consequently, we are able to transform the relation between $z_p$ and a given covariates into a percent change in $c_{p,t}$ associated with the same covariate.

More specifically, building on previous work [33, 31] we define the citation premium as the percent change $100\Delta c_p/c_p$ associated with shift in the independent variable $v$. For sake of simplicity, consider the basic linear model $Y(c) = z_p = \beta_0 + \beta_v v$ with the decomposition of differentials, $\partial Y(c)/\partial v = (\partial Y/\partial c)(\partial c/\partial v) = \beta_v$; it follows from the property of logarithms that $\partial Y/\partial c = \frac{1}{\sigma_t(1+c)}$. Hence, calculating the percent change $100\Delta c_p/c_p$ follows from rearranging the differential relations above, yielding $\frac{dc_p}{\sigma_t(1+c_p)dv} = \beta_v$. Hence, when the independent variable $\beta_v$ is a binary indicator variable, then the shift from value 0 to 1 corresponds to $dv = 1$, and so the percent change

$100\Delta c_p/c_p \approx 100 dc_p/c_p \approx 100 \times \sigma_t \times \beta_v \approx 100 \times \langle\sigma\rangle \times \beta_v$. By extension, when the independent variable is a scalar quantity then the percent change in $c_p$ associated with a 1-unit increase $dv$ is also given by $100 \times \langle\sigma\rangle \times \beta_v$. And in the case that the scalar quantity enters in logarithm (e.g. $\ln k_p$), then a 1% increase in $v$ corresponds to a $\langle\sigma\rangle \times \beta_v$ percent increase in $c_p$.

### E.2.1 Quantifying the Effect of Cross-domain Diversity on Scientific Impact

While previous work aimed to identify the role of $X_{CIP}$ in the ecosystem of biology and computing researchers that championed the genomics revolution [33], here we seek to simultaneously identify the relative impact of $X_{CIP}$ and $X_{SA}$ in the emerging ecosystem of HB science. In this way, we are able to compare research strategies that leverage combinations of diverse researcher expertise – i.e., cross-disciplinary collaboration – to those that do not, in the ultimate pursuit of interdisciplinary knowledge and research [27].

To this end, we model the relation between $z_p$ and $X_{CIP}$ & $X_{SA}$ by applying ordinary least-squares (OLS) regression to estimate the coefficients of the panel regression model implemented with researcher profile fixed effects:

$$z_{a,p} = \alpha_a + \beta_k \ln k_p + \beta_w \ln w_p + \beta_\tau \tau_{a,p} + \gamma_{X_{SA}} I_{X_{SA}} + \gamma_{X_{CIP}} I_{X_{CIP}} + \gamma(y_p, \overrightarrow{SA}_p, \overrightarrow{CIP}_p, \overrightarrow{R}_p) + \epsilon_{a,p} \ , \tag{8}$$

where the model parameters are estimated using Huber-White robust standard errors, which account for heteroscedasticity and serial correlation within the publication set of each scholar, indexed by $a$.

The control variables include $\ln k_p$, measuring the natural logarithm of the total number of coauthors; $\ln w_p$ is the natural logarithm of the total number of Major MeSH terms; the career age variable $\tau_{a,p}$, measuring the number of years since the researcher's first publication, capturing variation attributable to the career life cycle; and we also include factor variables controlling for

publication year and other article-level features measured by $\overrightarrow{SA}_p$, $\overrightarrow{CIP}_p$, $\overrightarrow{R}_p$.We exclude solo-authored research papers (i.e., those with $k_p = 1$) along with articles with a single Major MeSH category (i.e., those with $w_p = 1$).

**Table 6** shows the full parameter estimates for six similar models that differ primarily in the type of cross-domain diversity included as the principle test variable, represented generically by $I_X$. In models (1)-(3) we vary the specification of the types of SA and CIP being tested. To be specific, in model (1) we include indicators $I_{X_{SA}}$ and $I_{X_{CIP}}$, where $I_{X_{CIP}}$ takes the value 1 if $O_{CIP}(\vec{F}_p) = X_{CIP}$ and 0 if $O_{CIP}(\vec{F}_p) = M$, and similarly for $I_{X_{SA}}$. These definitions of $X$ correspond to the *Broad* configuration, calculated using all CIP and SA categories, as indicated in **Figures 3**(A,D). According to this definition, articles combining SA (CIP) from any two or more categories are classified as $X_{SA}$ ($X_{CIP}$).

In model (2) we use $X$ indicators defined according to the *Neighboring* configuration representing shorter-distance cross-domain activity – here capturing the neurobiological -vs- bioengineering interface. In our model specification, $X$ is represented by the binary indicator variables $I_{X_{\text{Neighboring},SA}}$ and $I_{X_{\text{Neighboring},CIP}}$. In the case of $X_{\text{Neighboring},SA}$, this interface corresponds to articles combining at least one MeSH mapping onto SA 1 (Psychiatry & Psychology) and at least one MeSH mapping onto SA [2] (Anatomy & Organisms), [3] (Phenomena & Processes) or [4] (Health). In the case of $X_{\text{Neighboring},CIP}$, this interface corresponds to articles combining at least one coauthor whose department maps onto CIP [1] (Neurosciences) or [3] (Psychology) and at least one coauthor whose department maps onto CIP [2] (Biology), [4] (Biotechnology & Genetics) or [5] (Medical Specialty) or [6] (Health Sciences) or [7] (Pathology & Pharmacology). Note that all Scopus scholars map onto a single CIP, and so solo-authored research articles are by definition mono-disciplinary.

In model (3) we use $X$ indicators defined according to the Distant configuration, representing longer-distance or "Convergent" cross-domain activity – here capturing the neuro-psycho-medical -vs- techno-computational interface. In our model specification, $X$ is represented by the binary indicator variables $I_{X_{\text{Distant},SA}}$ and $I_{X_{\text{Distant},CIP}}$. In the case of $X_{\text{Distant},SA}$, this interface corresponds to articles combining SAs [1-4] (corresponding to Psychiatry & Psychology (mind), Anatomy &

Organisms (structure), Phenomena & Processes (function) and Health, respectively) and at least one MeSH mapping onto SAs [5,6] (Techniques & Equipment and Technology & Information Science, respectively). In the case of $X_{\text{Neighboring},CIP}$, this interface corresponds to articles combining at least one coauthor whose department maps onto CIPs [1,3,5] (Neurosciences, Psychology and Medical Specialty, respectively) and at least one coauthor whose department maps onto CIPs [4,8] (Biotechnology & Genetics and Engineering & Informatics, respectively).

Likewise, Models (4-6) instead focus on $X_{SA\&CIP}$ (represented by $I_{X_{SA\&CIP}}$); each model corresponds to the either the *Broad*, *Neighboring* or *Distant* configurations defining $X_{SA}$ and $X_{CIP}$. As such, these models test the citation premium associated with articles featuring cross-domain diversity in combination. Because we exclude the confounding subsets of articles featuring $X_{SA}$ but not $X_{CIP}$, or vice versa, then the counterfactual to $X_{SA\&CIP}$ in are articles that are mono-dimensional in both categories. Thus, since the counterfactual groups are similar, the the citation premium estimated by $\gamma_{X_{SA\&CIP}}$ are comparable with the $\gamma_{X_{SA}}$ and $\gamma_{X_{CIP}}$ estimated in models (1-3). The full set of parameter results are reported in **Table 7**, and the transformed point estimates measuring the percent increase in $c_p$ associated with each $X$ definition are visually summarized in **Figure 9**(C).

### E.2.2 Quantifying Shifts in the Effect of Cross-domain Diversity Associated with the Announcement of Global Flagship HB Projects Circa 2013

We test for shifts in the citation premium attributable to the advent of global Flagship HB projects by introducing an interaction between $I_{2014+,p}$ and $I_{X_{SA\&CIP}}$, as indicated by the addition of two terms into the second row,

$$\begin{aligned} z_{a,p} = \alpha_a + \beta_k \ln k_p + \beta_w \ln w_p + \beta_\tau \tau_{a,p} + \gamma_{X_{SA\&CIP}} I_{X_{SA\&CIP}} + \\ + \gamma_{2014+} I_{2014+} + \delta_{X_{SA\&CIP}+} (I_{X_{SA\&CIP}} \times I_{2014\,+}) + \\ \gamma(y_p, \overrightarrow{SA}_p, \overrightarrow{CIP}_p, \overrightarrow{R}_p) + \epsilon_{a,p}\ , \end{aligned} \tag{9}$$

As before, this Difference-in-Difference approach is based upon the counterfactual comparison of articles featuring $X_{SA\&CIP}$ to those featuring $M$, integrating an additional comparison between articles published after 2014 to those published before 2014. As in the previous citation model, the model parameter $\gamma_{X_{SA\&CIP}}$ represents the citation premium attributable to research endeavors simultaneously featuring cross-domain combinations of both SA and CIP. However, in this specification $\gamma_{X_{SA\&CIP}}$ applies to articles published before 2014. The analog estimate of the relative citation premium for articles published after 2014 is $\gamma_{X_{SA\&CIP}} + \delta_{X_{SA\&CIP+}}$. In other words, if all other covariates are held at the average values, then the citation premium difference is given by $\delta_{X_{SA\&CIP+}}$, with positive (negative) values indicating an increase (decrease) in the citation premium after 2014. The principle test variables $\gamma_{X_{SA\&CIP}}$, $\delta_{X_{SA\&CIP+}}$ and their sum are visually summarized in **Figure 9**(D).

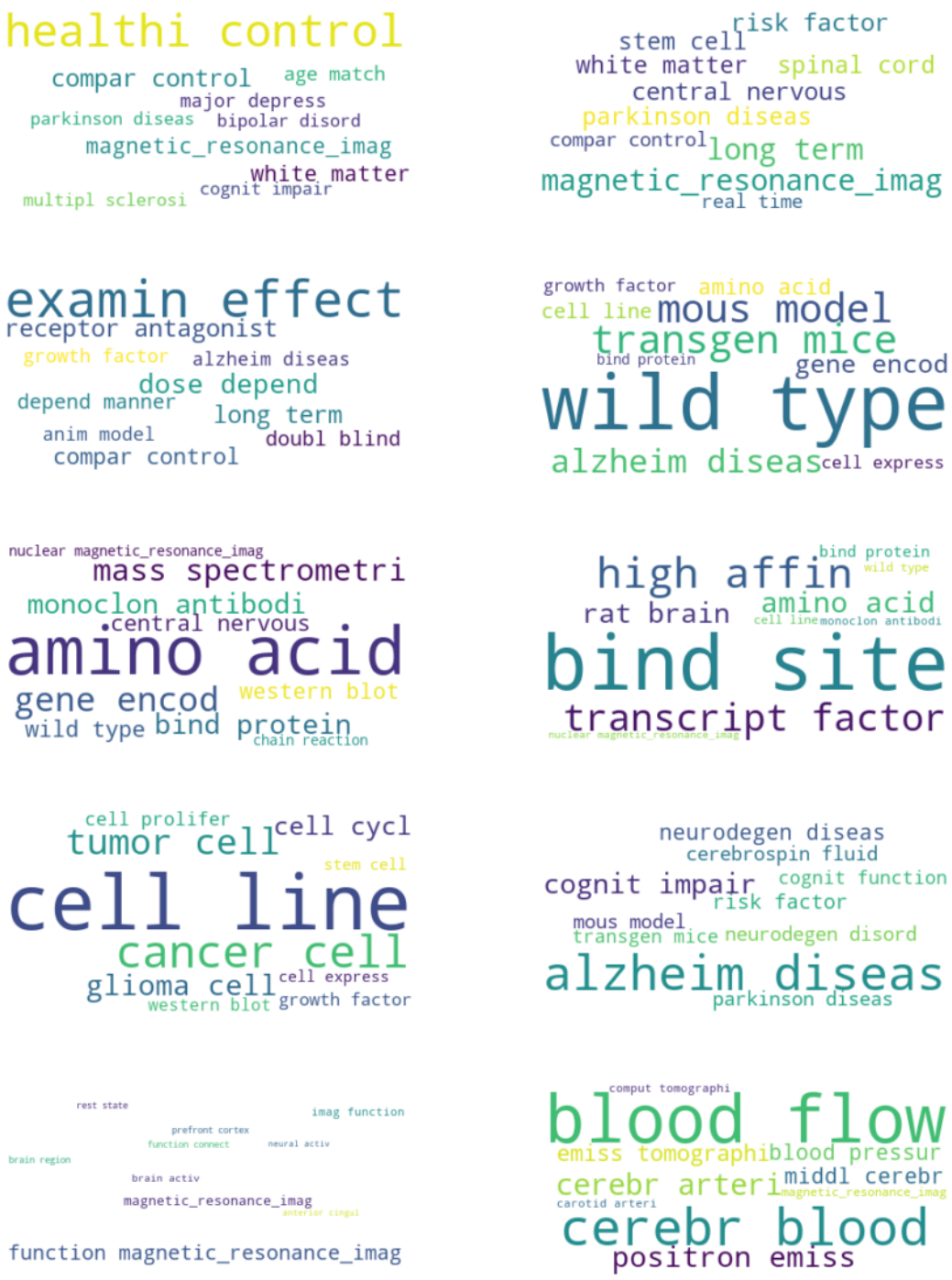



Figure A7: **Word cloud representation of article clustering using K-means algorithm.**


Figure A8: **Word cloud representation of cosine similar words of gene expression from 1965-2018.**


Figure A9: **Word cloud representation of cosine similar words of magnetic resonance from 1965-2018.**

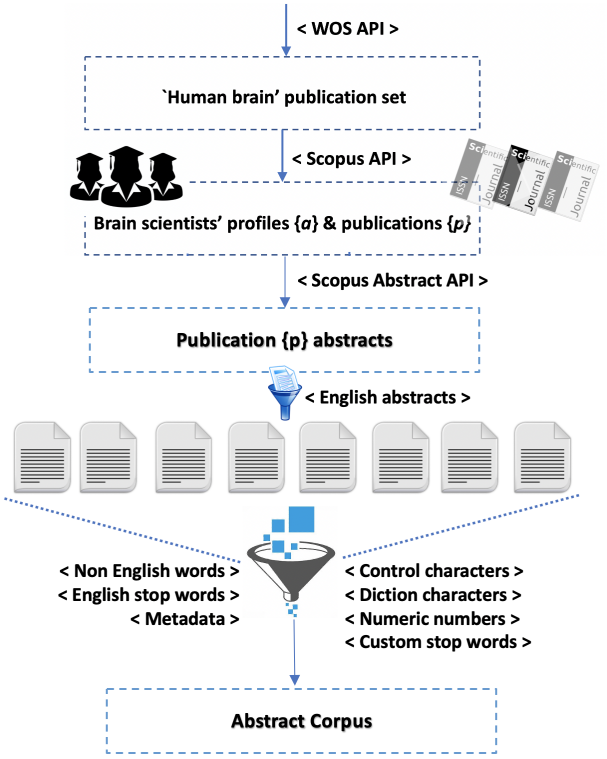



Figure A10: **Article pre-processing steps.**

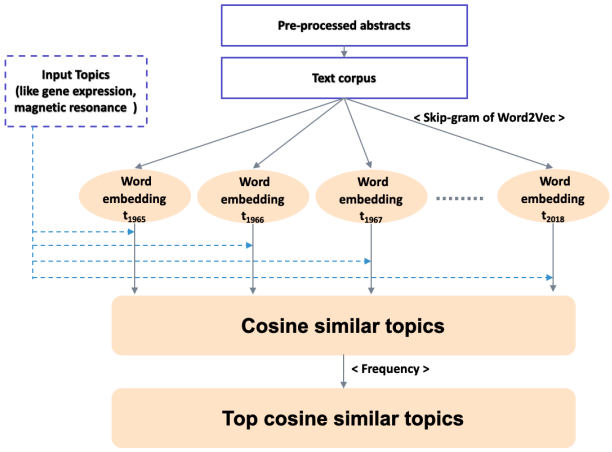



Figure A11: **Temporal analysis of cosine similar topics.**