CHARTING THE EVOLUTION AND TRANSFORMATIVE IMPACT OF THE PACIFIC SYMPOSIUM ON BIOCOMPUTING THROUGH A 30-YEAR RETROSPECTIVE ANALYSIS OF COLLABORATIVE NETWORKS AND THEMES USING MODERN COMPUTATIONAL TOOLS

Leah Zhang*
Thomas Jefferson High School for Science & Technology
Alexandria, VA, USA
Email: 2025lzhang@tjhsst.edu

Sameeksha Garg*,
Department of Computer Science, Carnegie Mellon University
Pittsburgh, PA USA
Email: sameeksg@andrew.cmu.edu

Edward Zhang*, Sean McOsker, Carly Bobak, Kristine Giffin, Brock Christensen
Dartmouth College Geisel School of Medicine
Hanover, NH USA
Email: edward.b.zhang.27@dartmouth.edu, Sean.A.McOsker.GR@dartmouth.edu , carly.a.bobak@dartmouth.edu , kristine.a.giffin@dartmouth.edu , brock.c.christensen@dartmouth.edu

Joshua Levy**
Department of Computational Biomedicine, Cedars Sinai Medical Center
Los Angeles, CA USA
Email: joshua.levy@cshs.org

Founded nearly 30 years ago, the Pacific Symposium on Biocomputing (PSB) has continually promoted collaborative research in computational biology, annually highlighting emergent themes that reflect the expanding interdisciplinary nature of the field. This study aimed to explore the collaborative and thematic dynamics at PSB using topic modeling and network analysis methods. We identified 14 central topics that have characterized the discourse at PSB over the past three decades. Our findings demonstrate significant trends in topic relevance, with a growing emphasis on machine learning and integrative analyses. We observed not only an expanding nexus of collaboration but also PSB's crucial role in fostering interdisciplinary collaborations. It remains unclear, however, whether the shift towards interdisciplinarity was driven by the conference itself, external academic trends, or broader societal shifts towards integrated research approaches. Future applications of next-generation analytical methods may offer deeper insights into these dynamics. Additionally, we have developed a web application that leverages retrieval augmented generation and large language models, enabling users to efficiently explore past PSB proceedings.

Keywords: natural language processing, network analysis, Pacific Symposium on Biocomputing, topic modeling, interdisciplinary collaboration

* Denotes equal contribution as co-first authors.
** To whom correspondence should be addressed. This work is supported by DoD grant PR220927 and NIH P30CA023108 support for JL.

1. Introduction

The Pacific Symposium on Biocomputing (PSB) was co-founded in 1996 by Dr. Teri Klein, Dr. Lawrence Hunter, and Sharon Surles, originating from the Biotechnology Computing Tracks at the Hawaiian International Conference on System Sciences [1–5]. Initially, PSB aimed to provide a platform for pioneering work in databases, algorithms, interfaces, visualization, modeling, and other computational methods applied to the challenges of molecular biology. As an annual multidisciplinary scientific conference held in Hawaii, it has continuously fostered international collaboration in computational biology.

Over the past 30 years, PSB has undergone significant evolution. Each year, the conference themes are curated to address emerging and critical issues in biocomputing, driven by proposals from leading researchers in new areas. This dynamic approach, unique among scientific gatherings, makes PSB an ideal subject for examining the progression of research themes over time, thereby reflecting the evolving landscape of computational biology.

Attending PSB offers significant opportunities for career advancement, professional development, and networking. These conferences are essential for discussing cutting-edge scientific themes and acquiring new knowledge [6,7]. Beyond immediate academic and professional benefits, attendees gain exposure to new technologies and methodologies that can be implemented in their own practices and institutions. Therefore, understanding the academic impact of such conferences is crucial for appreciating their role in advancing science and practice.

Over the past thirty years, PSB has witnessed transformative changes in biocomputing. This period has seen the rise of artificial intelligence in medicine [8], the sequencing of the human genome, and advancements in precision health. Innovations such as multimodal, single-cell [9], and spatial analyses, along with vast bioimaging datasets, have revolutionized our approach to biological data. Concurrent advancements in computing speed, storage capacity, GPUs, and internet connectivity have further enabled these scientific breakthroughs.

In 1996, PSB manuscripts and presentations focused on the foundational aspects of computational biology [4]. In contrast, by 2024, the focus has shifted towards leveraging complex integrations of multimodal data and advanced computational techniques [2]. The upcoming 30th anniversary of PSB presents a prime opportunity to reflect on the evolution of research themes, highlighting the growth in collaboration and scientific impact within the community.

To comprehensively understand these developments, we have conducted a quantitative retrospective analysis of the entire history of PSB proceedings. This study spans numerous articles and abstracts presented at PSB, providing insights into the dynamic nature of biocomputing as a field. By employing advanced computational tools for this meta-analysis, we aim to elucidate the intricate patterns of research evolution, collaboration networks, and thematic shifts over the conference's history. This analysis not only underscores the importance of PSB in shaping the field but also demonstrates the power of computational methods in understanding scientific progress.

2. Methods

2.1. *Overview*

Inspired by a similar work analyzing conference themes and impact over 30 years [10], our analysis utilizes topic modeling, large language models (LLM) and network analysis to map out:

1. **Topic Modeling:** The main themes of PSB, their prevalence and evolution over time.
2. **Evolving Co-Authorship Networks:** The personal impact of participation in the symposium, examining how PSB has spurred the formation of new, transdisciplinary collaborations.
3. **Citation Networks:** The scientific impact of PSB themes, as evidenced by citation metrics, broken down by topic and reported independently.
4. **Interactive Dashboard for Perusing Prior Proceedings:** The development of a Retrieval Augmented Generation (RAG) tool as an interactive research tool for rapid access of past proceedings.

Readers can find the code used for data preprocessing and analysis as well as instructions for deploying our interactive PSB dashboard at the following GitHub repositories: https://github.com/Leahie/PSBmodel

2.2. *Examining Evolving PSB Themes through Count-Based and Neural Topic Modeling*

2.2.1. *Extraction of Proceedings Text*

We used the Beautiful Soup package to web scrape PDFs of all PSB conference proceedings, available at https://psb.stanford.edu/psb-online/ [11,12]. Each year's proceedings included documents ranging from session introductions, short abstracts, workshops, and full peer reviewed papers. Only peer reviewed papers, from the years 1996-2024, with viable linked PDF files were downloaded and parsed. Due to inconsistencies in web formatting, separate web scrapers were developed for years 1996, 1997, 1998-2001, and 2002-2024. Document parsing for all proceedings led to the extraction of information such as the link of the pdf, title of authors, for each manuscript.

2.2.2. *Text Preprocessing*

After the PDFs were downloaded, pdfplumber was used to extract the text from each manuscript [13]. A custom text processor was developed to remove accents, special figures, numbers, stopwords, extra whitespace, and words less than 2 letters. After this step, contractions were expanded, and text was converted to lowercase.

Further text processing enhanced the readability of the documents. The appearance of section numbers and words such as "abstract", "introduction", "references" — words typically found in conference proceedings and part of the PSB manuscript template — helped filter PDFs that were poorly converted to text– these same subsections were used to divide the document into three components which were subsequently saved: 1) abstract, 2) main body, and 3) references. The main body of the document included all text between the "Introduction" and "Reference" headers.

2.2.3. *Topic Modeling with LDA, BERTopic, and DTM*

After preprocessing the text, we employed three primary methods to identify and model emerging themes: Latent Dirichlet Allocation (LDA), Dynamic Topic Modeling (DTM) and BERTopic [14–16]. These techniques focused exclusively on the main body of texts spanning from 1996 to 2024, allowing for a precise analysis of words and phrases that characterize the themes and topics of the documents. By applying these methods, we were able to ascertain the prevalence of each topic across various manuscripts and authors at specific time points. This approach facilitated a detailed

exploration of the evolving landscape of themes throughout the study period, offering insights into the dynamics of topic popularity and relevance over time.

LDA is a generative probabilistic model designed to identify latent topics within a corpus of text documents. LDA assumes that each document is a mixture of topics, and each topic is a distribution of words. By using Dirichlet distributions to guide the selection of topics for each document and words from topics, LDA can effectively capture the prevalence of topics across documents and the frequency of words within topics. The model achieves this by estimating three key components: (1) the topic distribution within each document, (2) the word distribution within each topic, and (3) the topic assignment for each word in the documents. We use the python package *Tomotopy* for our LDA implementation which uses Collapsed Gibbs Sampling, a Markov Chain Monte Carlo (MCMC) method which iteratively samples the conditional distributions of latent variables allowing the model to estimate the posterior distribution of topics within the corpus[17].

For LDA, words within each topic were initially ranked by the estimated Dirichlet parameters, $\{p_i \text{ for } i = 1,2, \dots, n | \sum p_i = 1\}$. The Dirichlet parameters in our topic modeling methods do not account for the ubiquity of words, which often leads to an oversaturation by commonly used terms. To address this issue, we implemented a re-ranking strategy for these words based on their saliency and relevance, both of which reweight the importance of words by considering their document frequency. Saliency quantifies a word's relative importance by measuring how significantly it contributes to the uniqueness of a topic. Relevance, on the other hand, evaluates a word based on both its probability within a topic and its frequency across documents. This dual metric ensures a balanced assessment that enhances topic interpretability.

Term frequencies were normalized, which served as a foundation for calculating saliency and relevance for each topic. The formulas for calculating Salience, Relevance, and Frequency Normalization are outlined below [18–20]:

$$S_{k,w} = P_{k,w} \log\left(\frac{P_{k,w}}{F'_w}\right)$$

$$R_{k,w} = \lambda \log\left(P_{k,w}\right) + (1-\lambda) \log\left(\frac{P_{k,w}}{F'_w}\right)$$

$$F'_w = \frac{F_w}{\sum_w F_w}$$

Dynamic topic models were utilized alongside standard LDA in our dataset. Unlike LDA, which assumes static topics, dynamic topic models incorporate changes over time by using topic priors from previous time steps to inform the topic distributions at subsequent time steps. This approach allows for the detection of emerging or evolving topics that might otherwise be overlooked by LDA's time-averaged methodology.

BERTopic, proposed by Maarten Grootendorst, is another topic modeling technique that integrates state-of-the-art transformer models such as Bidirectional Encoder Representations from Transformers (BERT). Our BERTopic implementation generates dense sentence-level embeddings which were aggregated across each manuscript to form document-level embeddings which were subsequently clustered to derive coherent topics across documents [16]. By using transformer models like BERT, these contrived embeddings encapsulate contextual relationships between words offering a rich semantic representation of the documents, addressing the limitations of traditional topic modeling which often approaches these texts as a bag of words.

Generated high dimensionality embedding produced by these transformer models are reduced in dimensionality with techniques such as Uniform Manifold Approximation and Projection (UMAP) [21] and subsequently clustered using Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN) [22] which identifies dense regions in the embeddings space and groups documents together without a need for a preset amount of clusters. After clustering the embeddings, BERTopic extracts the most representative words for each cluster by ranking them using the Class Based Term Frequency-Inverse Document Frequency (c-TF-IDF). C-TF-IDF is calculated by taking the logarithm of one plus the average number of words per class divided by the frequency of word across all classes. The term frequency emphasizes words that are more frequent and the inverse document frequency captures rarely used but still important words.

$$w_{x,c} = \|tf_{x,c}\| + \log\left(1 + \frac{A}{f_x}\right)$$

The optimal number of topics for each topic modeling method was determined using the coherence metric [23], which measures the semantic similarity between high scoring words within each topic. This metric helps ensure that the topics generated are meaningful and interpretable. We utilized the coherence scores to select the number of topics that provided the highest level of interpretability while maintaining a balance with model complexity.

2.2.4. *Characterizing Topic Prevalence over Time*

To streamline the interpretation process, we opted to restrict our analysis to LDA models which did not initially account for the temporal dynamics of each topic's evolution. This approach simplifies the initial modeling by focusing solely on prevalence of thematic content without the additional complexity of temporal variation in topic content. After training, we extracted document-topic distributions for each paper, which represent the proportion of each topic within each document. These distributions were then aligned with the corresponding dates of publication or timestamps. To capture temporal trends, we computed the average topic distribution for each defined time period.

To identify overarching patterns in the evolution of topic prevalence over time, we employed K-Means clustering via the *tslearn* python package [24]. This method utilized a dynamic time warping (DTW) distance matrix of the time series data [25]. DTW is particularly adept at capturing similarities in temporal sequences, even when there are shifts or timing differences among the sequences. By applying K-Means clustering to this DTW distance matrix [26], we were able to discern and illustrate the predominant trends and shifts in topic prevalence throughout the corpus.

2.3. *Evaluating the Influence of Collaborative Networks at PSB on Research Themes*

2.3.1. *Extraction and Fuzzy Matching of Author Names*

Author names for each manuscript were extracted from the proceedings website for each year, and a database of these titles and names was established. To ensure unique identification, we employed a combination of citation analysis and relied on the Scopus database of authors. Each paper was mapped to its unique DOI and PubMed ID using CrossRef's REST API (https://api.crossref.org/swagger-ui/index.html) and MetaPub (https://pypi.org/project/metapub/) [27,28]. Then, each identifier was looked up using Pybliometrics, a python-based wrapper for the Scopus API [29]. Using Pybliometrics, each paper was mapped to its authors and each author was mapped to their Scopus ID, a unique identifier assigned to them by Scopus. This approach allowed us to account for variations in spelling and other inconsistencies that commonly occur in author name listings. By using citation data, we were able to link each paper to a unique identifier and link variations of a name to a single author.

2.3.2. *Development of Collaborative Networks over Time*

Collaborative networks were constructed annually based on co-authorships (edges) within articles published that year [30]. The attributes of each node (representing an author) were defined by the average topic distribution from Latent Dirichlet Allocation (LDA), specifically averaged across the manuscripts the authors contributed to within PSB that year. Each network represented a cross-sectional snapshot at a specific point in time, typically characterized by sparse connections due to its annual limitation.

To gain a deeper understanding of the evolving collaborative landscape, we extended our analysis to include cumulative networks. In this approach, nodes and edges from previous years were incorporated into the current year's network. This method allowed us to observe not only isolated annual interactions but also the development and persistence of collaborative ties over time.

2.3.3. *Overall Measures of Interdisciplinarity and Collaboration*

In our study, we focused on characterizing authors' topical areas of interest by analyzing their cumulative topic distributions. These distributions were derived from the topic-document matrices of all their prior publications at PSB up to but not including the current evaluation point. We hypothesized that alignment in these topical areas might influence the likelihood of future collaborations, and that this influence could vary over time.

To empirically test this hypothesis, we calculated the cosine similarity between the topical distributions of two authors, each aggregated from prior years. Cosine similarity measures the cosine of the angle between two vectors in a multidimensional space, serving as an indicator of how aligned two authors are in their prior topics of interest. To assess the potential for these authors to form a collaborative connection (or 'edge'), we employed a logistic regression model that includes an interaction term with time, using R v4.3:

$$\text{logit}(p_{ij}) = \beta_0 + \beta_1 \times \text{similarity}(d_i, d_j) + \beta_2 \times t + \beta_3 \times (\text{similarity}(d_i, d_j) \times t)$$

where p_{ij} is the probability of forming an edge between authors *i* and *j*, $\text{similarity}(d_i, d_j)$ is the cosine similarity score between their prior topic distributions, and t represents the year of the

collaboration relative to the study period. This model not only quantifies the relationship between topical alignment and formation of collaborative links but also how this relationship evolves over time, permitting a dynamic analysis of factors influencing collaboration within the PSB community. Results were also stratified by the number of prior joint publications within a co-author dyad.

Furthermore, each author's ability to bridge across diverse topics was quantified using an entropy score, calculated at each timepoint, reflecting the variety and distribution of topics in their publications to that point. This score served as an indicator of an author's interdisciplinarity, suggesting their potential to contribute to and collaborate across various thematic areas.

Finally, an author's influence at each timepoint was quantified using various network centrality measures, including degree centrality, eigenvector centrality, and betweenness centrality [31]. Degree centrality measures the number of direct connections an author has, indicating their immediate influence within the network. Eigenvector centrality accounts for the influence of an author's connections, reflecting how connected they are to other highly connected authors. Betweenness centrality highlights authors who serve as bridges between different clusters or groups within the network, showcasing their role in facilitating information flow. Centrality measures were normalized based on the size of the connected component (subgraph) to which each node belongs.

As a descriptor of overall network dynamics, the final cumulative network for 2024 was analyzed using the Leiden algorithm [32]. This approach partitions the network into clusters based on the strength of the connections, ensuring that clusters are more connected internally than with other parts of the network. Each cluster was then labeled based on averaged topic distribution to that point, providing a thematic summary that reflects the predominant scholarly interests of each subgroup.

2.4. *Measuring Scholastic Impact through Citations*

Finally, the impact of PSB papers was characterized by analyzing the number of citations each paper received. For each topic identified by LDA analysis, now assigned to individual papers, we calculated the average number of citations both overall and across different time periods. This approach enabled us to determine which topics garnered the most attention and influence within the scholarly community, while accounting for the publication dates of the articles. Measures of interdisciplinarity and collaboration (*2.3.3*) were correlated with citation counts (independent variable) using linear regression modeling, adjusting for time as a covariate. The analysis was restricted to the 2005-2019 period to allow sufficient time for collaborations/topics to develop and to mitigate potential biases from lower citation counts associated with more recent publications.

2.5. *Developing Interactive Dashboard to Facilitate Review of Papers*

2.5.1. *Developing Retrieval Augmented Generation Approach*

Retrieval Augmented Generation (RAG) enhances the capabilities of large language models (LLMs) by incorporating a preliminary reference to a knowledge base before generating responses. This method is particularly beneficial when applying LLMs to specialized or highly specific domains that are not well-represented in the model's initial training data [33]. For efficient querying of PSB manuscripts, this involves augmenting the user query with a relevancy search within a vector database that contains embeddings of the knowledge base, addressing common issues such as inaccuracies or the generation of irrelevant content by the LLM.

Our RAG setup for analyzing PSB documents was implemented using LangChain [34]. Initially, papers were downloaded in PDF format and segmented into chunks of approximately 1000 words each. These segments were then transformed into vector embeddings using OpenAI's "text-embedding-3-small" model and stored within a vector database managed by Chroma. For each user query, the LangChain Merger Retriever searches this database to find and retrieve the most relevant embeddings, which are then provided as context to the LLM through the RunnablePassThrough function. This process ensures that the generated responses are both accurate and contextually relevant to the specific queries related to PSB documents.

2.5.2. *Web Application and Availability*

To facilitate user interaction with our RAG setup, we developed a web application using Streamlit [35]. This application provides a user-friendly interface for querying the PSB document database and viewing the augmented responses. The web application is accessible at https://psb-rag.streamlit.app, and the complete codebase for the RAG workflow and further reference to the application is available for public review and use on our GitHub repository. To utilize the site, users will need to provide an OpenAI API key.

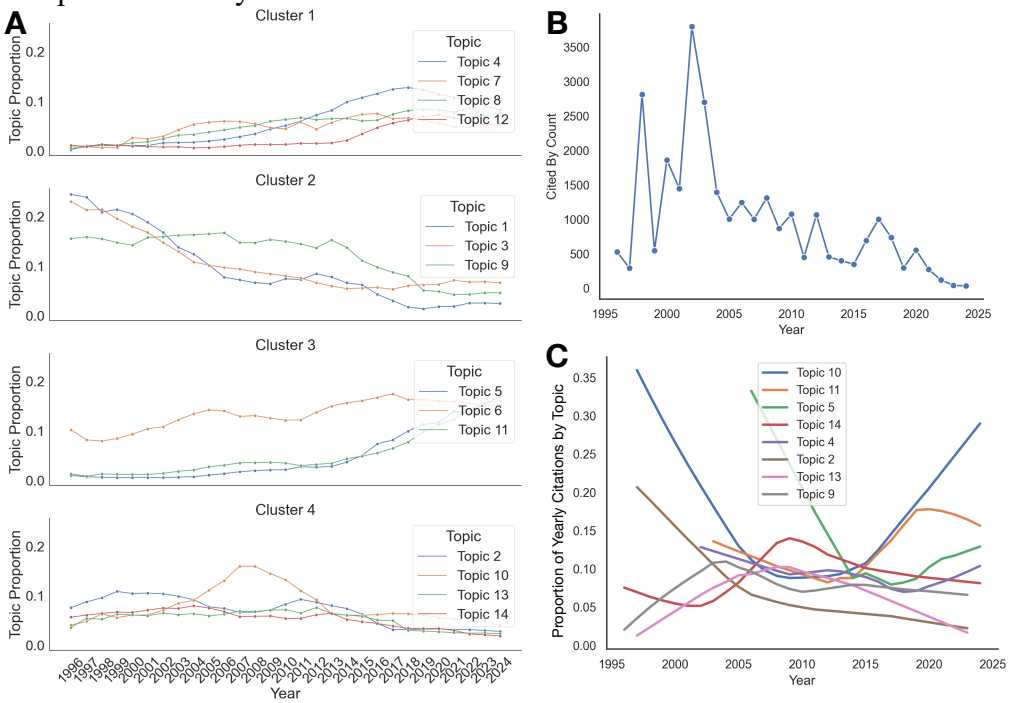

Figure 1: Topic and Citation Dynamics at PSB, 1996 to 2024: A) Prevalence of topics over time, highlighting evolving interests in specific research areas, grouped into four clusters to enhance readability. **B)** tracks the total cumulative citations of PSB publications broken down by uear, with a notable peak in the early 2000s. **C)** Proportion of yearly citations by select topics, indicating shifts in the impact of various research themes over three decades.

3. Results

3.1. *Topic Modeling Results*

Topic modeling was optimized using coherence metrics to ascertain the most interpretable number of topics for Latent Dirichlet Allocation (LDA), BERTopic, and dynamic topic models (DTM). This

approach identified 14 distinct topics using LDA, 13 with DTM, and 19 with BERTopic. Detailed topic-word distributions for all models are available in the supplementary materials hosted on our GitHub repository. These results (including LDA relevance metrics) are summarized in **Tables 1 and 2**, providing a direct comparison of the outputs from the topic modeling techniques, with complete parameters found in the supplementary. LDA topics were clustered based on their prevalence over time (**Figure 1**). While both LDA and BERTopic underwent thorough analysis, the LDA results demonstrated higher coherence, with less overlap between topics compared to BERTopic, where topics tended to show more redundancy. As a result, discussions in our main text have primarily focused on the LDA topics.

Table 1: Comparative Overview of Topic Keywords in LDA and BERTopic Models

LDA		BERTopic	
Topic	**Words**	**Topic**	**Words**
1	cancer cell tumor pathway samples cells survival pathways sub breast	**1**	protein proteins structure residues sequence structures binding set function two
2	drug drugs harm disease knowledge diseases relationships sources target meta	**2**	snps snp disease genetic population plo association gene genotype allele
3	reads peak rate sites posterior peaks read site likelihood mass	**3**	terms gene information ontology text system term used one database
4	sequences rna dna regions mutation mutations genome disordered base disorder	**4**	gene genes expression regulatory transcription network set binding motif time
5	interactions interaction proteins functional cluster clusters similarity clustering networks ppi	**5**	patient patients health clinical medical features set models using time
6	features performance learning training feature prediction classification fier trained models	**6**	cancer gene genes mutations sub tumor expression cell drug samples
7	snp snps plo genotype population allele variants populations locus genetic	**7**	network networks time graph that system state nodes are pathway
8	algorithm tree problem size optimal matrix probability proceedings trim let	**8**	tree trees taxa species number distance gene genomes two algorithm
9	text terms ontology query database relations system name language concepts	**9**	drug drugs target similarity compounds targets network based set chemical
10	this from which can each all have not our die	**10**	imaging brain features age subjects cognitive disease mci poe feature
11	residues binding structure structures energy residue motif amino motifs surface	**11**	sequence dna sequences coding length domain reads genome system gene
12	state reactions reaction activity compounds metabolic enzyme transcription molecules metabolites	**12**	cell cells immune spatial expression seq gene single crna tumor
13	clinical patients risk age health patient phenotypes causal cohort was	**13**	cancer features set mirna layer genes feature gene samples cluster
14	user software flow work visualization tools field file environment science	**14**	virus cov viral sequences protein human hiv sars proteins host
		15	drug harm gene relationships drugs xml genes heritable text ontology
		16	rac species rrna biome communities micro microbial diversity subgroup coa
		17	rna rnas sequence structure secondary sequences reads alignments base sci
		18	peptide mass spectra peak peaks peptides spectrum ion teo intensity
		19	alignment alignments sequences scoring score gap optimal length sequence path

Table 2: Evolution of Topics in Dynamic Topic Model: Progression of keywords for selected topics across three years—1996, 2010, and 2024—using DTM, illustrating the shift in research focus, such as Topic 1 evolving from basic molecular structures to complex cancer drug models. All years can be found in the supplementary material.

Topic	1996: Words	2010: Words	2024: Words
1	proton system proteins structure length molecular you	time algorithm sub interactions system class well	different performance samples sub models cancer drug
2	structure sequences molecular given site proteins solvent	problem algorithm shown networks different interactions function	table clustering cell patients disease samples values
3	time structure system molecular class structures points	graph size patterns algorithm different state are	values samples features patients models data disease
4	function sequences points course table system time	binding clustering sub different class algorithm rna	samples different use cell models values transcript
5	tree surface however sequences structure different proton	interactions clustering time structure table state algorithm	learning clinical data features predicted performance brain
6	would system surface proteins point sequences pair	possible however structure different nodes given time	data models drug age patients brain learning
7	structure system given information distance students tree	well clustering state base time algorithm harm	across patients use disease studies models graph
8	surface system sequences points structure given time	different class state size algorithm base function	data patients table features learning cell values
9	given system molecular grape die residues proteins	nodes different algorithm structure table state states	learning time drug transcript use data values
10	algorithm point database structure molecular system given	first drug use sub time class are	age effect features studies performance models brain
11	points structure time students system point sequences	interactions shown time drug size algorithm different	feature individuals performance ancestry samples spatial models
12	system would site second point die pair	nodes size class sub algorithm time samples	clinical table ancestry patients age across training
13	site system you tree value time could	drug hee different rna residues off rees	samples clinical table disease patients across clustering

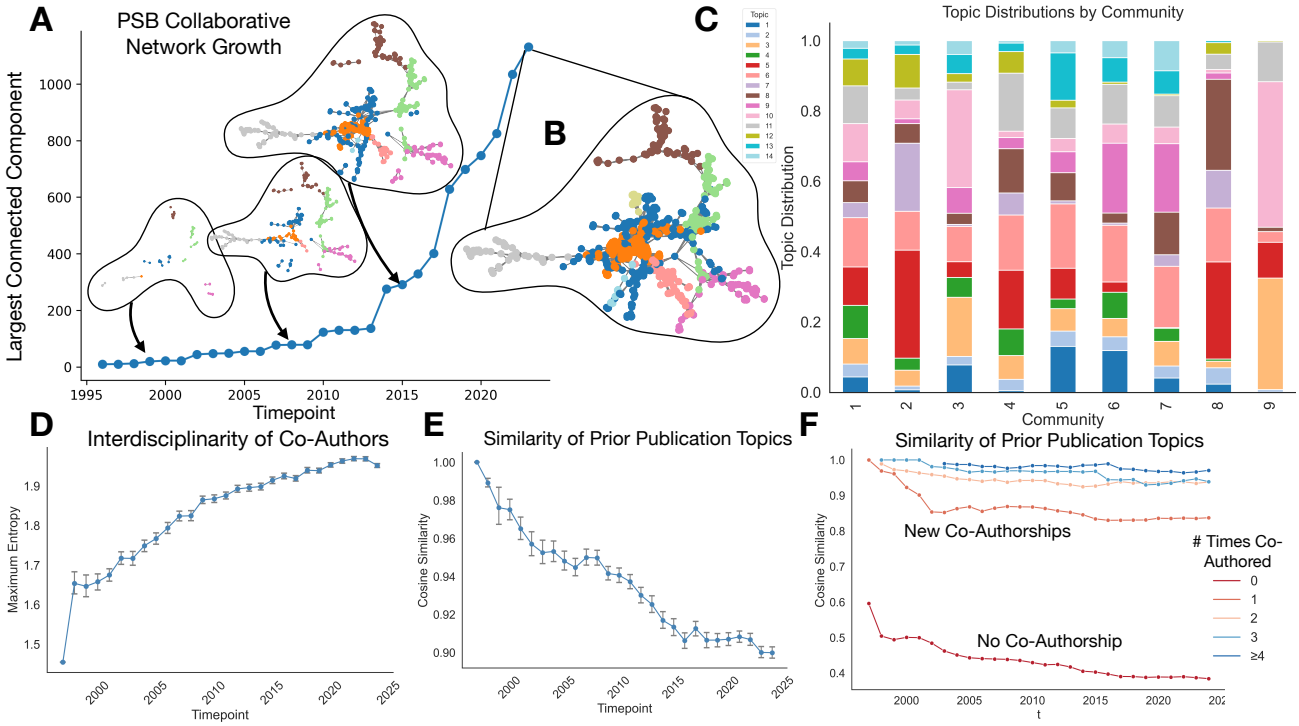

Figure 2: Analysis of Collaboration Dynamics: A) Growth of the largest connected component within the PSB collaboration network from 1996 to 2024, **B)** Visualization of the final 2024 collaborative network, with authors labeled by assigned community via the Leiden algorithm, **C)** Cumulative distribution of topics within each community, reflecting the thematic focus areas that have emerged among collaborators, **D)** Increasing interdisciplinary nature of co-authorships over time as denoted through maximal entropy of prior years' topic distribution within subsequent co-author dyads; **E)** Declining trend in cosine similarity of prior publication topics among subsequent co-author dyads; **F)** Analyzes the relationship between the frequency of co-authorship and topic similarity, showing that more frequent collaborators tend to share more similar research interests from prior years' topic similarity, while first-time collaborators often engage from more diverse thematic backgrounds with collaborators

3.2. *Collaboration Network Results*

Our analysis mapped the growth of the largest connected component in the collaboration network over time as an indicator of collaboration intensity (**Figure 2A**). Initially, in 1996, the largest group comprised of 9 co-authors. By 2003, this number had grown to 45. Significant growth occurred in 2011 and 2015, with the largest connected components increasing from 79 to 123 in 2011, and from 136 to 278 in 2015, respectively. By 2019, the component had expanded to 632 members, reaching 1147 by 2024—nearly one-third of the entire network size of 3932 PSB authors.

The resulting network was divided into 9 distinct communities, reflecting unique topical focuses as determined by average topic-document distributions among community members (**Figure 2B,C**).

Our analysis highlighted substantial shifts in the nature of collaborative ties within the PSB network. To quantify the diversity of topics present within collaborations, we calculated entropy

measurements for each co-author based on their topic distributions prior to the year of publication. These entropy values provided a numerical representation of the thematic diversity within each collaboration, illustrating the broadening scope of interdisciplinary interaction over time (**Figure 2D**). There was a gradual increase in the diversity of topics involved in collaborations, with entropy values rising steadily from the year 2000 onwards (β=0.01, $p<0.001$). This suggests that researchers are increasingly engaging in collaborations that cross traditional disciplinary boundaries.

Cosine similarity was used to assess the thematic alignment between collaborating authors over time based on prior years' aggregate topic distributions. Initially high similarity scores in the early years of the symposium have gradually decreased, suggesting that over time, collaborators are less likely to share a common research focus before co-authoring together (β=-2.8e-3, $p<0.001$). This trend is pronounced among new collaborations, where cosine similarity scores dropped by nearly 20% from 2000 to 2024, reflecting a broadening of interdisciplinary interaction (**Figure 2E,F**). Despite the decrease over time in topic alignment, prior years' topic alignment was positively associated with the likelihood of co-authorship (OR=1.8e6, $p<0.001$) and co-authors who continued to publish together maintained higher levels of topic alignment (β=0.03, $p<0.001$).

Centrality measures were computed yearly to identify key individuals within the final 2024 cumulative PSB collaboration network. These measures pinpointed those who were central in connecting larger subnetworks, reflecting their pivotal roles in fostering collaboration (**Table 3**).

Table 3: Key Influencers in the PSB Network Across Different Years, influence determined using weighted eigenvector, betweenness and degree centrality

Timepoint	Eigenvector	Betweenness	Degree
1999	Toshihisa Takagi	Subramanian Subbiah	Satoru Kuhara
	Satoru Kuhara	A. Keith Dunker	Toshihisa Takagi
	Emiko Furuichi	Satoru Kuhara	Adam Godzik
2004	Satoru Miyano	Satoru Miyano	Satoru Miyano
	David C. Kulp	Philip E. Bourne	Satoru Kuhara
	Conrad C. Huang	Adam Godzik	William Stafford Noble
2009	Bart L.R. de Moor	Satoru Miyano	Russ B. Altman
	Conrad C. Huang	Russ B. Altman	Philip E. Bourne
	Thomas E. Ferrin	Philip E. Bourne	William Stafford Noble
2014	Russ B. Altman	Marylyn D. Ritchie	Adam Godzik
	Philip E. Bourne	Russ B. Altman	Russ B. Altman
	Zoubin Ghahramani	Satoru Miyano	Philip E. Bourne
2019	Marylyn D. Ritchie	Marylyn D. Ritchie	Russ B. Altman
	Sarah A. Pendergrass	Sarah A. Pendergrass	Atul Janardhan Butte
	Shefali Setia Verma	Russ B. Altman	Jason H. Moore
2024	Marylyn D. Ritchie	Marylyn D. Ritchie	Russ B. Altman
	Shefali Setia Verma	Russ B. Altman	Lawrence E. Hunter
	Sarah A. Pendergrass	Shefali Setia Verma	Joel T. Dudley

3.3. *Citation Results*

The manuscripts published in the yearly PSB proceedings have significantly varied in their impact over time, with a notable peak in citations during the early 2000s. As illustrated in **Figure 1B**, the today's citation count for these papers shows a substantial rise around this period, followed by a gradual decline. This figure traces the number of current citations received by papers based on their publication year and does not normalize by passing time– manuscripts published earlier are more likely to have more citations. After adjusting for time, we found that articles with a higher entropy score (indicating interdisciplinarity; t=3.33, p=0.001) and lower cosine similarity (indicating formation of interdisciplinary relationship; t=-3.06, p=0.002) were associated with higher citation count. **Figure 1C** delineates the proportion of yearly citations attributable to specific topics,

assigning each manuscript the topic with the highest document-topic score. This analysis reveals that certain topics have gained or lost prominence in terms of citation impact over the years.

4. Discussion

4.1. *Topic Modeling Interpretation and Discussion*

The topics derived from BERTopic shared some commonalities with those from LDA, including areas such as pathway analysis, drug-drug polypharmacy interactions, CHIP-seq peak calling, SNPs, sequence alignment, protein-protein interactions, and biomedical ontologies. However, BERTopic covered a broader array of topics, including network analysis, COVID-19, microbiome analysis, brain imaging, spatial transcriptomics, and temporal features, showcasing its expansive thematic reach (**Table 1**). Conversely, LDA uniquely captured topics related to machine learning and residue binding, which were not present in the BERTopic set. Notably, the exclusion of rapidly emerging fields such as multimodal analysis in BERTopic was also observed, highlighting some limitations in its topic coverage. Dynamic topic models provided an evolutionary view of these topics, which were initially based on themes from 1996. Over time, these topics have notably shifted from focusing primarily on biomolecular structures and sequences to more complex areas such as clinical prediction models that integrate spatial data and RNA sequencing prediction models.

Cluster 3 highlights a marked increase in topics such as residue binding and machine learning (specifically topics 5, 6, and 7) (**Figure 1A**). The surge in these topics aligns with the rise of deep learning and sophisticated protein folding algorithms, which gained prominence nearly a decade ago [36]. This trend underscores the impact of technological advancements on driving research focus areas within bioinformatics, particularly those that leverage computational innovations.

In contrast, Cluster 2, which includes topics 1, 3, and 9, pertains to pathway analysis and biomedical ontologies. Notably, pathway analysis (topic 1) was a central theme in sessions as far back as 1996, with titles like “Genome, Pathway and Interaction Bioinformatics” and “Computation in Biological Pathways” in 1997 [37,38]. Despite their current popularity, these topics are long-established in the field rather than emerging areas. Over time, the prevalence of these foundational themes has seen a relative decrease, suggesting a shift in research focus toward newer computational techniques and applications.

4.2. *Collaborative Network Discussion and Interpretation*

The identified communities in the largest connected component from the 2024 network and their differing topic distributions highlight the symposium’s role in facilitating diverse interdisciplinary collaborations (**Figure 2**). Our results show a marked shift towards interdisciplinary collaboration at the PSB, as evidenced by increasing entropy in topic distributions and decreasing cosine similarity over time among collaborators. This evolving trend suggests that PSB participants are not only expanding their collaborative networks but are also engaging with a wider array of scientific disciplines than in previous years. The decrease in cosine similarity particularly highlights how the nature of these collaborations has evolved from close-knit, topic-specific interactions to more diverse, interdisciplinary exchanges. This shift may reflect broader changes in the field of bioinformatics, where cross-disciplinary approaches are becoming essential to tackle increasingly complex research questions [39–41].

The trend of decreasing topic similarity, especially notable among first-time collaborators, indicates that PSB is successfully fostering an environment where researchers feel encouraged to explore new collaborations outside their immediate expertise. This is crucial for driving innovation and adapting to the rapidly changing landscape of bioinformatics research. The data also suggest that while established collaborators continue to work within familiar thematic areas, there is a strong movement towards branching out into new topics.

Over time, the composition of influential members within the PSB network has evolved (**Table 3**), with recent years marking the rise of key figures, including three current editors/organizers. Their prominence might stem from consistent presence, increasing opportunities for co-authorship. While this could indicate a strategic integration of leadership roles, it might also reflect incidental outcomes of sustained participation. This observation underscores the complexities of interpreting the dynamics between leadership presence and collaborative patterns in academic networks.

4.3. *Citation Discussion and Interpretation*

It was not surprising that earlier PSB publications, especially those from around the year 2000, received more attention, as reflected by the number of cumulative citations. Our citation analysis also revealed a declining trend in the citation relevance of certain topics. For instance, LDA topic 2, which focuses on drug-drug interactions, and topic 9, covering ontologies, were highly cited in the early 2000s but have experienced a gradual decrease in citation percentage over the years. In contrast, topic 11 on protein folding has seen a noticeable increase in popularity.

The future trajectory of less frequently cited topics remains uncertain as the field evolves with new technologies. The process of these topics becoming mainstream could significantly alter their impact. Additionally, shifts in community focus—from established scholars to emerging researchers—may also influence citation patterns. The growing interdisciplinarity of the field presents another challenge, as works that span multiple disciplines sometimes struggle to connect with a well-defined audience, potentially diluting their impact [42]. Nevertheless, our citation analysis suggests that forming interdisciplinary ties, as fostered through this venue, was associated with greater scientific impact, even after adjusting for time.

5. Conclusion

The Pacific Symposium on Biocomputing stands as a premier venue in bioinformatics, embodying the forefront of convergent thinking by bringing together individuals from diverse backgrounds to address complex problems that span multiple disciplines. Through our application of quantitative NLP and network analysis methods, we have effectively mapped the scope and nature of the various themes and collaborative ties that have formed at this venue over the past 30 years. These analyses reveal not only the evolving patterns of collaboration but also highlight the increasing diversity and interdisciplinarity of the research presented at PSB. Looking ahead, we anticipate that PSB will continue to foster groundbreaking interdisciplinary research, adapting to new scientific challenges and technologies. As the field grows, the symposium will likely play a crucial role in shaping future trends in bioinformatics and computational biology. We expect that continued innovations in analytical methods will further illuminate the dynamics of collaboration and influence within this community, enhancing our understanding of how interdisciplinary interactions drive scientific progress.

References

1. Pacific Symposium on Biocomputing [Internet]. Wikipedia. 2022 [cited 2024 Jul 31]. Available from: https://en.wikipedia.org/w/index.php?title=Pacific_Symposium_on_Biocomputing&oldid=1123233788
2. Altman RB, Hunter L, Ritchie MD, Murray T, Klein TE. Pacific Symposium on Biocomputing 2024. Biocomputing 2024. Kohala Coast, Hawaii, USA: WORLD SCIENTIFIC; 2023.
3. Altman RB, Dunker AK, Hunter L, Ritchie MD, Murray T, Klein TE. Biocomputing 2020: Proceedings of the Pacific Symposium [Internet]. WORLD SCIENTIFIC; 2020 [cited 2019 Nov 28]. Available from: https://www.worldscientific.com/worldscibooks/10.1142/11698
4. Hunter L, Klein TE. Pacific Symposium on Biocomputing'96: Hawaii, USA, 3-6 January, 1996 [Internet]. World Scientific; 1995 [cited 2024 Jul 31]. Available from: https://books.google.com/books?hl=en&lr=&id=20soDwAAQBAJ&oi=fnd&pg=PR7&dq=pacific+symposium+on+biocomputing&ots=EjZk6sRoiC&sig=JkZyQ7UgrWd3vZ6j0xJkegwG2xM
5. Altman RB, Hunter L, Klein TE, Murray T, Dunker AK, Ritchie MD. Biocomputing 2021:Proceedings of the Pacific Symposium [Internet]. 2020 [cited 2024 Jul 31]. Available from: https://directory.doabooks.org/handle/20.500.12854/42151
6. Hauss K. What are the social and scientific benefits of participating at academic conferences? Insights from a survey among doctoral students and postdocs in Germany. Res Eval. 2020 Aug 27;rvaa018. PMCID: PMC7499794
7. Augustine EF, Steele SJ, McIntosh S, Sugarwala L, White RJ, Yousefi-Nooraie R, Zand MS, Ossip DJ. Utilizing the Un-Meeting model to advance innovative translational and team science. J Clin Transl Sci. 7(1):e176. PMCID: PMC10514683
8. Daneshjou R, Brenner SE, Chen JH, Crawford DC, Finlayson SG, Kidziński Ł, Bulyk ML. Precision Medicine: Using Artificial Intelligence to Improve Diagnostics and Healthcare. Biocomputing 2022 [Internet]. Kohala Coast, Hawaii, USA: WORLD SCIENTIFIC; 2021 [cited 2024 Jul 31]. p. 223–230. Available from: https://www.worldscientific.com/doi/abs/10.1142/9789811250477_0021
9. Garmire LX, Yuan GC, Fan R, Yeo GW, Quackenbush J. SINGLE CELL ANALYSIS, WHAT IS IN THE FUTURE? Biocomputing 2019 [Internet]. Kohala Coast, Hawaii, USA: WORLD SCIENTIFIC; 2018 [cited 2024 Jul 31]. p. 332–337. Available from: https://www.worldscientific.com/doi/abs/10.1142/9789813279827_0030
10. Peek N, Combi C, Marin R, Bellazzi R. Thirty years of artificial intelligence in medicine (AIME) conferences: A review of research themes. Artificial Intelligence in Medicine. 2015 Sep 1;65(1):61–73.
11. Hajba GL. Website Scraping with Python: Using BeautifulSoup and Scrapy [Internet]. Berkeley, CA: Apress; 2018 [cited 2024 Jul 31]. Available from: http://link.springer.com/10.1007/978-1-4842-3925-4
12. Patel JM. Web Scraping in Python Using Beautiful Soup Library. Getting Structured Data from the Internet [Internet]. Berkeley, CA: Apress; 2020 [cited 2024 Jul 31]. p. 31–84. Available from: http://link.springer.com/10.1007/978-1-4842-6576-5_2

13. S.V J. pdfplumber (Version 0.8.0) [Internet]. 2020. Available from: https://github.com/jsvine/pdfplumber
14. Blei DM, Ng AY, Jordan MI. Latent Dirichlet Allocation. Journal of Machine Learning Research. 2003;3(Jan):993–1022. PMID: 36346659
15. Blei DM, Lafferty JD. Dynamic topic models. Proceedings of the 23rd international conference on Machine learning - ICML '06 [Internet]. Pittsburgh, Pennsylvania: ACM Press; 2006 [cited 2024 Jul 31]. p. 113–120. Available from: http://portal.acm.org/citation.cfm?doid=1143844.1143859
16. Grootendorst M. BERTopic: Neural topic modeling with a class-based TF-IDF procedure [Internet]. arXiv; 2022 [cited 2024 Jul 31]. Available from: http://arxiv.org/abs/2203.05794
17. Chauhan U, Shah A. Topic Modeling Using Latent Dirichlet allocation: A Survey. ACM Comput Surv. 2021 Sep 17;54(7):145:1-145:35.
18. Sievert C, Shirley K. LDAvis: A method for visualizing and interpreting topics. Proceedings of the workshop on interactive language learning, visualization, and interfaces [Internet]. 2014 [cited 2024 Jul 31]. p. 63–70. Available from: https://aclanthology.org/W14-3110.pdf
19. Chuang J, Manning CD, Heer J. Termite: visualization techniques for assessing textual topic models. Proceedings of the International Working Conference on Advanced Visual Interfaces [Internet]. Capri Island Italy: ACM; 2012 [cited 2024 Jul 31]. p. 74–77. Available from: https://dl.acm.org/doi/10.1145/2254556.2254572
20. Fan A, Doshi-Velez F, Miratrix L. Assessing topic model relevance: Evaluation and informative priors. Statistical Analysis. 2019 Jun;12(3):210–222.
21. McInnes L, Healy J, Saul N, Großberger L. UMAP: Uniform Manifold Approximation and Projection. Journal of Open Source Software. 2018 Sep 2;3(29):861. PMID: 33588368
22. Campello RJGB, Moulavi D, Sander J. Density-Based Clustering Based on Hierarchical Density Estimates. In: Pei J, Tseng VS, Cao L, Motoda H, Xu G, editors. Advances in Knowledge Discovery and Data Mining. Springer Berlin Heidelberg; 2013. p. 160–172.
23. Röder M, Both A, Hinneburg A. Exploring the Space of Topic Coherence Measures. Proceedings of the Eighth ACM International Conference on Web Search and Data Mining [Internet]. Shanghai China: ACM; 2015 [cited 2024 Jul 31]. p. 399–408. Available from: https://dl.acm.org/doi/10.1145/2684822.2685324
24. Tavenard R, Faouzi J, Vandewiele G, Divo F, Androz G, Holtz C, Payne M, Yurchak R, Rußwurm M, Kolar K. Tslearn, a machine learning toolkit for time series data. Journal of machine learning research. 2020;21(118):1–6.
25. Dynamic Time Warping. Information Retrieval for Music and Motion [Internet]. Berlin, Heidelberg: Springer Berlin Heidelberg; 2007 [cited 2024 Jul 31]. p. 69–84. Available from: http://link.springer.com/10.1007/978-3-540-74048-3_4
26. MacQueen J. Some methods for classification and analysis of multivariate observations. Proceedings of the fifth Berkeley symposium on mathematical statistics and probability [Internet]. Oakland, CA, USA; 1967 [cited 2024 Jul 31]. p. 281–297. Available from: https://www.google.com/books?hl=en&lr=&id=IC4Ku_7dBFUC&oi=fnd&pg=PA281&dq=+Some+Methods+for+classification+and+Analysis+of+Multivariate+Observations&ots=nQUdG-L8oP&sig=Z6DngAT2EpGdYnq-cXK49tVTDPI
27. Hendricks G, Tkaczyk D, Lin J, Feeney P. Crossref: The sustainable source of community-owned scholarly metadata. Quantitative Science Studies. MIT Press One Rogers Street, Cambridge, MA 02142-1209, USA journals-info …; 2020;1(1):414–427.

28. Metapub [Internet]. [cited 2024 Jul 31]. Available from: https://metapub.org/
29. Rose ME, Kitchin JR. pybliometrics: Scriptable bibliometrics using a Python interface to Scopus. SoftwareX. Elsevier; 2019;10:100263.
30. Liu X, Bollen J, Nelson ML, Van de Sompel H. Co-authorship networks in the digital library research community. Information processing & management. Elsevier; 2005;41(6):1462–1480.
31. Borgatti SP. Centrality and network flow. Social networks. Elsevier; 2005;27(1):55–71.
32. Traag VA, Waltman L, Van Eck NJ. From Louvain to Leiden: guaranteeing well-connected communities. Scientific reports. Nature Publishing Group; 2019;9(1):1–12.
33. Lewis P, Perez E, Piktus A, Petroni F, Karpukhin V, Goyal N, Küttler H, Lewis M, Yih W tau, Rocktäschel T. Retrieval-augmented generation for knowledge-intensive nlp tasks. Advances in Neural Information Processing Systems. 2020;33:9459–9474.
34. Topsakal O, Akinci TC. Creating large language model applications utilizing langchain: A primer on developing llm apps fast. International Conference on Applied Engineering and Natural Sciences [Internet]. 2023 [cited 2024 Jul 31]. p. 1050–1056. Available from: https://www.researchgate.net/profile/Oguzhan-Topsakal/publication/372669736_Creating_Large_Language_Model_Applications_Utilizing_LangChain_A_Primer_on_Developing_LLM_Apps_Fast/links/64d114a840a524707ba4a419/Creating-Large-Language-Model-Applications-Utilizing-LangChain-A-Primer-on-Developing-LLM-Apps-Fast.pdf
35. Khorasani M, Abdou M, Hernández Fernández J. Web Application Development with Streamlit: Develop and Deploy Secure and Scalable Web Applications to the Cloud Using a Pure Python Framework [Internet]. Berkeley, CA: Apress; 2022 [cited 2024 Jul 31]. Available from: https://link.springer.com/10.1007/978-1-4842-8111-6
36. Fidelis K, Grudinin S. Session introduction: AI-driven Advances in Modeling of Protein Structure. Biocomputing 2022 [Internet]. Kohala Coast, Hawaii, USA: WORLD SCIENTIFIC; 2021 [cited 2024 Jul 31]. p. 1–9. Available from: https://www.worldscientific.com/doi/abs/10.1142/9789811250477_0001
37. Karp P, Romero PR, Neumann E. GENOME, PATHWAY AND INTERACTIONS BIOINFORMATICS. Pacific Symposium on Biocomputing [Internet]. World Scientific; 2002 [cited 2024 Jul 31]. p. 398–399. Available from: http://psb.stanford.edu/psb-online/proceedings/psb03/intro-path.doc
38. Karp PD, Riley M. Session on Computation in Biological Pathways. Biocomputing'97-Proceedings Of The Pacific Symposium [Internet]. World Scientific; 1996 [cited 2024 Jul 31]. p. 18. Available from: https://books.google.com/books?hl=en&lr=&id=bEBPDwAAQBAJ&oi=fnd&pg=PA18&dq=%22Computation+in+Biological+Pathways%22&ots=Z30I-4GwvN&sig=8fE0E70A2sb7FN9qTQjxjCW5oSY
39. Powell WW, White DR, Koput KW, Owen-Smith J. Network Dynamics and Field Evolution: The Growth of Interorganizational Collaboration in the Life Sciences. American Journal of Sociology. The University of Chicago Press; 2005 Jan;110(4):1132–1205.
40. Romano P, Giugno R, Pulvirenti A. Tools and collaborative environments for bioinformatics research. Briefings in Bioinformatics. 2011 Nov 1;12(6):549–561.

41. Exploratory Analysis of Topic Interests and Their Evolution in Bioinformatics Research Using Semantic Text Mining and Probabilistic Topic Modeling | IEEE Journals & Magazine | IEEE Xplore [Internet]. [cited 2024 Jul 31]. Available from: https://ieeexplore.ieee.org/document/9738599
42. Yegros-Yegros A, Rafols I, D'Este P. Does Interdisciplinary Research Lead to Higher Citation Impact? The Different Effect of Proximal and Distal Interdisciplinarity. [cited 2024 Jul 31]; Available from: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0135095