Supplementary File S1: Systematic literature review method description and study limitations.

This file is part of the publication:

From Shallow to Deep Bioprocess Hybrid Modeling: Advances and Future Perspectives

Roshanak Agharafeie (https://orcid.org/0000-0002-8450-7954); Nova Information Management School (NOVA IMS), NOVA University Lisbon, Campus de Campolide, 1070-312 Lisboa, Portugal.; D20200461@novaims.unl.pt

João R. C. Ramos (https://orcid.org/ 0000-0002-6832-6774); LAQV-REQUIMTE, Nova School of Science and Technology, NOVA University Lisbon, Campus de Caparica, 2829-516 Caparica, Portugal.; jr.ramos@campus.fct.unl.pt

Jorge M. Mendes (https://orcid.org/0000-0003-2251-3803); Nova Information Management School (NOVA IMS), NOVA University Lisbon, Campus de Campolide, 1070-312 Lisboa, Portugal.; NOVA Cairo at The Knowledge Hub Universities, New Administrative Capital, Cairo, Egypt 11835; jmm@isegi.unl.pt

Rui Oliveira (https://orcid.org/0000-0001-8077-4177); LAQV-REQUIMTE, Nova School of Science and Technology, NOVA University Lisbon, Campus de Caparica, 2829-516 Caparica, Portugal.; rmo@fct.unl.pt

Corresponding authors:

Roshanak Agharafeie; Nova Information Management School (NOVA IMS), NOVA University Lisbon, Campus de Campolide, 1070-312 Lisboa, Portugal.; D20200461@novaims.unl.pt;

Rui Oliveira; LAQV-REQUIMTE, Nova School of Science and Technology (NOVA-SST), NOVA University Lisbon, Campus da Caparica, 2829-516 Caparica, Portugal.; rfo@fct.unl.pt;

1. Systematic literature review method

This review focuses on the application of HNNs to bioprocesses. The preferred reporting items for systematic reviews and meta-analyses (PRISMA) methodology were adopted. PRISMA is an updated version of the QUOROM Statement (QUality Of Reporting Of Meta-analyses). PRISMA incorporates a checklist containing 27 items that help structure a systematic review [35]. For bibliometric analysis, the Mendeley application allowed the extraction of metadata and the elimination of duplicates. For network analysis, the VOSviewer software tool (V1.6.18) has been applied to visualize the dataset's extracted information and obtain quantitative and qualitative outcomes. The selection of articles obeyed to the following principles:

- Select articles from two databases, Scopus and Web of Science, based on the below described algorithms.
- Collect the documents of some well-known authors on this topic and refine them by keyword screening
- Select relevant articles cited by selected articles from databases and well-known authors (backward citation).

1.1. Algorithm for selection of articles from Scopus database

The paper selection algorithm from the Scopus database started with keyword screening in the “title, abstract, and keywords” of documents. Firstly, the advanced search performed by keywords; ("gray-box model*" OR "hybrid neural model*" OR "hybrid semiparametric model*" OR "hybrid semi-parametric" OR "hybrid neural network*" OR "hybrid mechanistic model" OR "hybrid white box model" OR "hybrid black box model" OR "hybrid parametric model" OR "hybrid nonparametric model" OR "Hybrid Artificial Neural Network" OR "Hybrid Process Model") AND (bioproc* OR biopharma* OR biofuel OR bioreact* OR ferment* OR biologic* OR biopolym* OR bioseparation* OR wastewater OR cell OR microorganism OR yeast OR bacteria OR

mammal* OR animal OR "systems biology" OR bioinformatics OR biotech* OR biomass OR "Escherichia Coli" OR "Recombinant Protein" OR "Recombinant Protein prod*" OR "e.coli" OR "microbial fuel" OR "biologic* wastewater treatment" OR bioethanol OR biodiesel) and retrieved 481 publications.

In the next step, some records were excluded based on the irrelevance of the subject areas (“Psychology”, “Economics, “Econometrics and Finance”, “Dentistry”, “Health Professions”, “Business, Management and Accounting”, “Social Sciences”, “Neuroscience”, “Physics and Astronomy”, “Earth and Planetary Sciences”) which resulted in 72 excluded documents and 409 publications.

Afterwards, some records were excluded based on the irrelevance of the keywords (“Pattern Recognition”, “Blood”, “Photovoltaic Cells”, “Diagnosis”, “Sewage Pumping Plants”, “Nerve Cell Network”, “Neurons”, “Fuel Cells”, “Electrodes”, “Sewer”, “Forestry”, “Geometry”, "PID Controllers", "Paget Bone Disease", "Partial Discharges", "Plasmid", "Power Control", "Power Spectral Density", "Pressure Effects", "Pressure Filter", "Pressure Filters", "Program Processors", "Battery State Of Charge", "Behavior-finding", "Behavioral Research", "Behavior", "Blood Glucose", "Blood Pressure", "Blood Pressure (BP)", "Blood Pressure Estimation", "Blood Pressure Measurement", "Blood Pressure Monitoring", "Bone", “Photovoltaic Power”, "Attention", "Attention Mechanisms", "Battery Management Systems", "Biological Organs", "Biomedical Signal Processing", "Brain", "COVID-19"), "Charging (batteries)", "Classification (of Information)", "Cytology", "Digital Storage", "Diseases", "Electric Discharges", "Image Classification", "Image Enhancement", "Image Processing", "Lithium-ion Batteries", "Low Power Electronics", "Lung Cancer", "Medical Imaging", "Solar Power Generation", "Solid Oxide Fuel Cells (SOFC)"), and the outcome was 133 excluded documents and 276 eligible publications (At

this step, if we doubted whether the keyword was related to the topic or not, we would have reviewed the abstract of the articles containing the keyword). Then the resulting documents were refined by the document´s type ("Book chapter", "Review paper", "conference review paper", and "Letter") and 25 documents were excluded because of the document type, and 251 papers were remaining articles. Finally, 94 relevant cases were obtained by manually reviewing the abstracts and contents of eligible publications (157 were excluded).

1.2. Algorithm for selection of articles from Web of Science database

The paper selection algorithm from the Web of Science database also started with keyword screening in the Topic (title, abstract, and keywords) of documents and retrieved a total of 251 publications. Regarding the differences between Scopus and Web of Science, we refined the documents by the Web of Science Categories and excluded the irrelevant categories; "Telecommunication"," Computer Science Hardware Architecture", "Radiology Nuclear Medicine Medical Imaging", "Transportation Science Technology", "Oceanography", "Cardiac Cardiovascular Systems", "Engineering Civil", "Information Science Library Science", "Geography Physical", "Physics-Condensed Matter", "Optics", "Imaging Science Photographic Technology", "Forestry", "Robotics", "Engineering Electrical Electronic", "Genetics Heredity", "Marine Freshwater Biology", "Rehabilitation", "Toxicology" and "Water Resources", and resulted in 50 excluded documents and 201 publications for further analysis.

Then the resulting documents were refined by the document´s type ("Review Articles", "Meeting Abstracts", and "Letter") and 6 documents were excluded. Finally, 87 relevant cases were obtained by manually reviewing the abstracts and contents of eligible publications (108 publications were excluded).

1.3. Algorithm for selection of articles from well-known Authors' work

Authors whom we reviewed their works were "Carrondo, M.J.T.", "Simutis, R.", "Lübbert, A.", "Oliveira, R.", "Galvanauskas, V.", "von Stosch, M", "Teixeira, A.P.", "Peres, J.", "Gnoth, S.", "Sokolov, M.", "Feyo de Azevedo, S.", "Zhang, D." and 819 publications were extracted from their works. The documents were refined by keywords screening ("hybrid model*", "hybrid neural*", "hybrid artificial neural*", "hybrid gray box*", "hybrid semi-parametric*", "hybrid mechanistic*", "hybrid black box*", "hybrid white box*", "hybrid parametric*", "hybrid nonparametric*") in "title, abstract and keywords" and 168 were sought for retrieval. Then 33 documents were excluded because of the document type ("Book chapter", "Review paper", and "Short Survey") and 135 papers were remaining articles. Finally, 69 relevant cases were obtained by manually reviewing the abstracts and contents of eligible publications**.**

1.4. From the backward citation

During this systematic literature review, we also checked references of selected papers, and we found 74 articles interesting to add to our database, so we named "backward citation" this part of our dataset.

2. PRISMA output

Using the previously described selection algorithms and the PRISMA flow diagram instructions [34] the following outcome was obtained:

- Scopus: The algorithm initially retrieved 481 publications from the Scopus database and after screening 94 relevant cases were obtained.
- Web of Science (WoS): The algorithm initially retrieved 251 publications from the WoS database, and after screening 87 relevant cases were obtained.

-From the well-known authors' search, 819 publications were extracted, and after screening 69 relevant cases were obtained.

-From the backward citation, 74 relevant cases were obtained.

After merging the articles and deletion duplicates, 185 publications were selected for keyword analysis, covering journal and conference papers published before September 2023 (Figure S1). We put all articles in a list in Scopus to have the opportunity to use the analytical reports of Scopus.

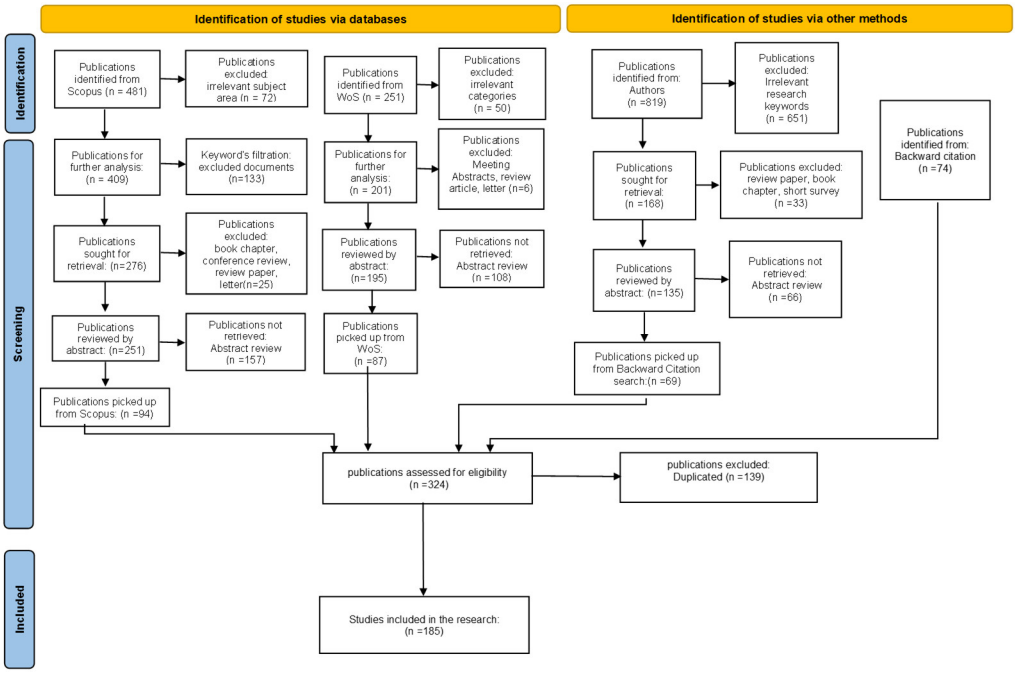

Figure S1-PRISMA flow diagram summarizes the selection of the articles based on the algorithm.

3. Statistical analysis of the PRISMA output

In this literature review, we analyzed 185 journal and conference papers published before September 2023. The first ten document sources that have published the highest number of articles, their ranking, publisher, and H-Index are summarized in Table S1.

Table S1- Specifications of the first ten document sources that have published the highest number of articles

No.	Source Title	Source Type	Documents	Country	Publisher	H-Index	Quartiles
1	Computers And Chemical Engineering	Journal	19	United Kingdom	Elsevier BV	152	Q1
2	Biotechnology And Bioengineering	Journal	12	Germany	Wiley-VCH Verlag	206	Q1

3	Computer-Aided Chemical Engineering	Book Series	11	Netherlands	Elsevier	30	Q4
4	Bioprocess And Biosystems Engineering	Journal	11	Germany	Springer Verlag	79	Q2
5	Journal Of Biotechnology	Journal	7	Netherlands	Elsevier	171	Q2
6	Biotechnology Progress	Journal	6	United States	Wiley-Blackwell	140	Q2
7	Brazilian Journal of Chemical Engineering	Journal	6	Brazil	Braz. Soc. Chem. Eng.	59	Q3
8	IFAC-Papers Online	Journal	5	Austria	IFAC Secretariat	86	Q3
9	AIChE Journal	Journal	5	United States	Wiley-Blackwell	182	Q1
10	Industrial And Engineering Chemistry Research	Journal	5	United States	American Chemical Society	245	Q1

It is also apparent in the subject area analytical report of Scopus that over the three decades, two subjects, “chemical engineering” and “Biochemistry, Genetics, and Molecular Biology” stand out with the highest number of publications. “Energy” and “Environmental Science” are two subjects that have attracted more attention since 2000. As expected, in the second decade (2001-2010), research on computer science subjects has grown significantly to develop methods in this field (Figure S2).

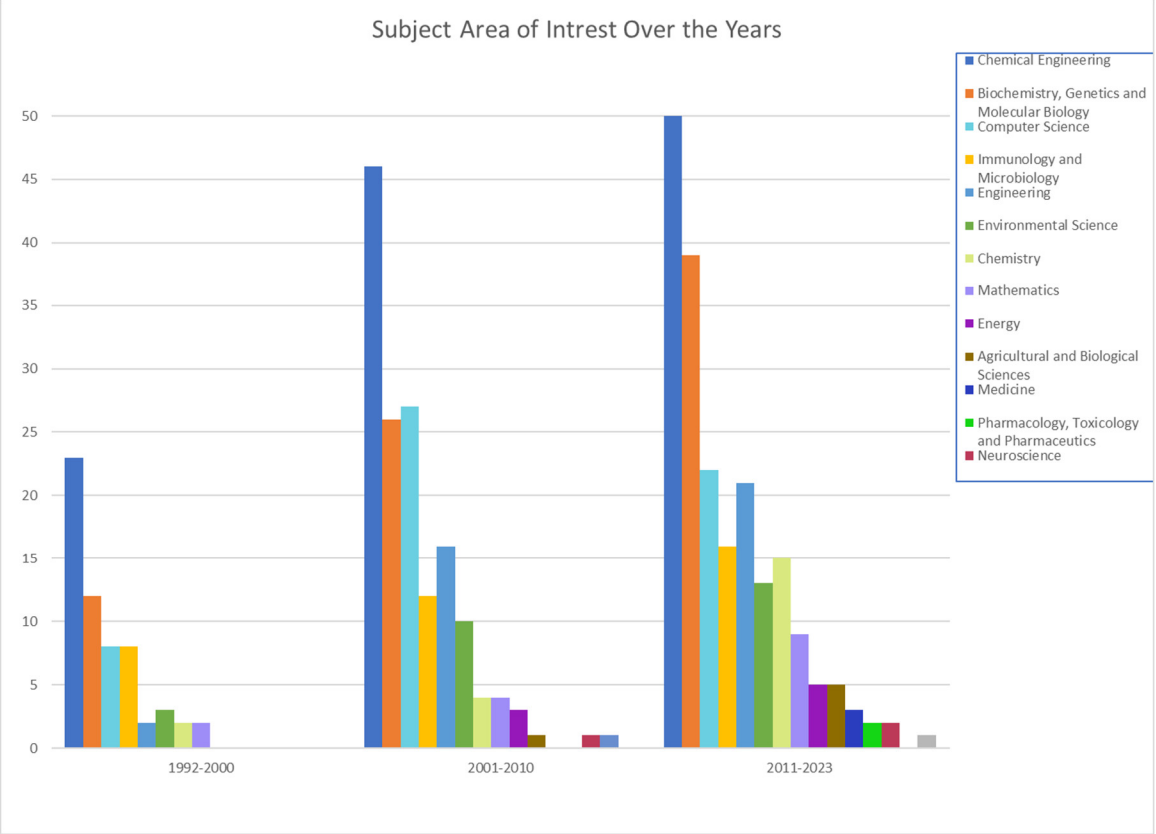

Figure S2-Subject areas of interest over the years based on the Scopus analytical reports

4. Keywords Analysis

The author´s keywords (included in the keyword section of the article) were analyzed at first. Additionally, indexed keywords (Indexed keywords are chosen by the database and are standardized to vocabularies derived from thesaurus) were also analyzed because some articles did not specify the author's keywords. Keyword analysis and visualization were performed with the help of VOSviewer. Firstly, the cooccurrence of authors´ keywords was analyzed with the full counting method and two times occurrences (in two different papers) as the minimum. Then, similar keywords were harmonized, and finally, uninformative keywords were omitted, such as

hybrid model, artificial neural network, and modeling. As a result, 42 author´s keywords were obtained (*Figure*)

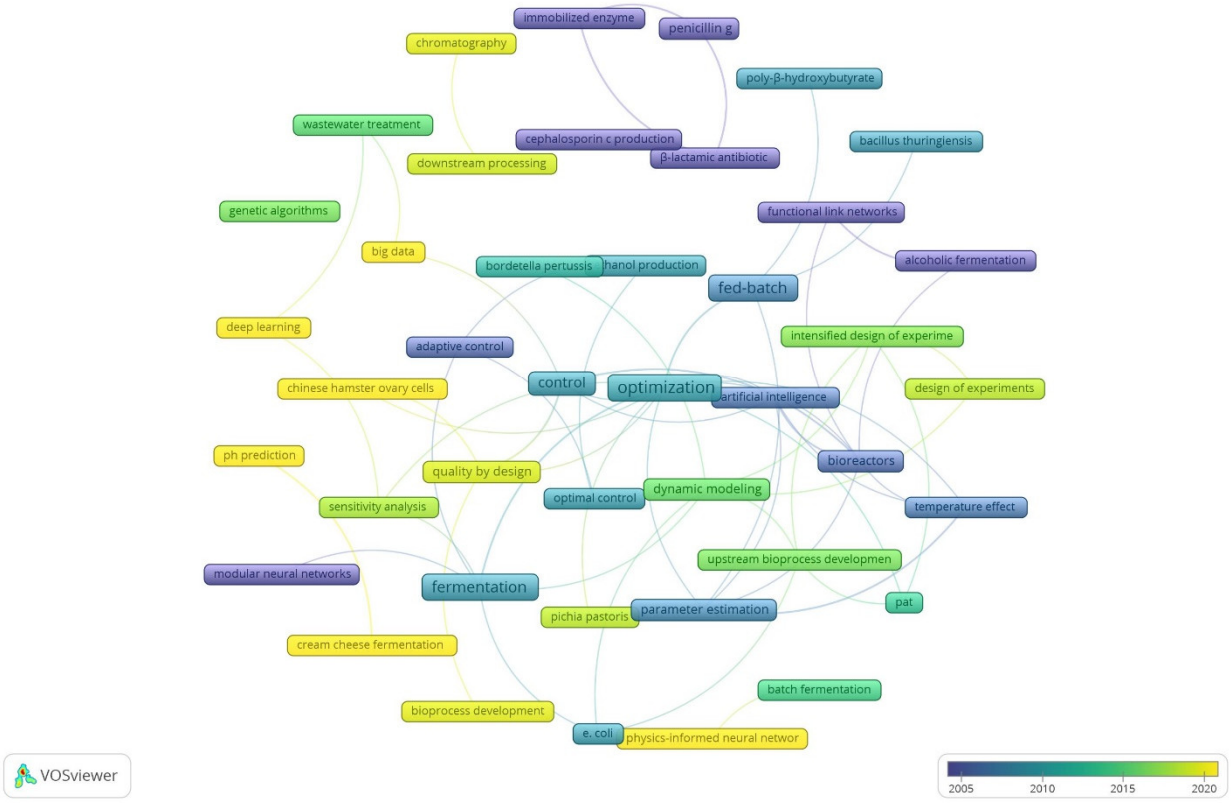

Figure S3- Author´s keywords occurrence analysis by year overlay visualization

The visualization showed that hybrid models were first applied for process control and optimization and parameter estimation in upstream steps. Subsequently, it was applied to downstream steps and in association with other techniques such as Design of Experiments (DoE), Process Analysis Technology (PAT) and Quality by Design (QbD). Recently, topics such as big data, deep learning, and physics informed neural networks have emerged.

4.1. Keywords Occurrence Over Publication Year

The co-occurrence of all 493 keywords (author´s keywords and indexed keywords with two times occurrence) was analyzed over time. Figure summarizes all keywords' occurrence over the publication year. Due to the large number of keywords, three time periods were considered.

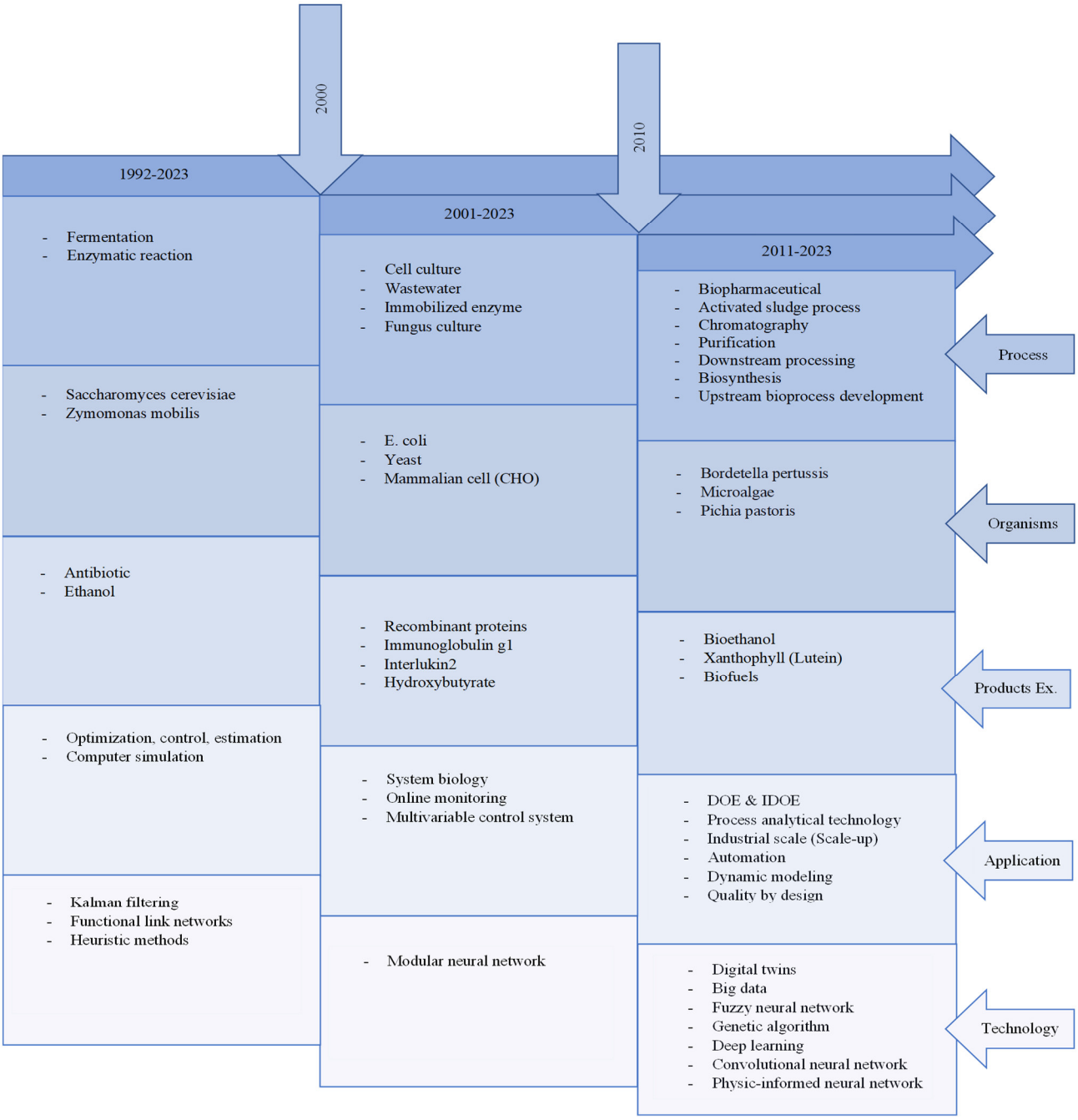

Figure S4- All Keywords (author´s keywords and indexed keywords) Occurrence Over the Years

4.1.1. Keyword occurrence from 1992 until 2000

In this period, 157 keywords were identified, which were reduced to 22 by the previously described keyword analysis algorithm. Based on the information output of keywords occurrence by year overlay visualization (Figure), the HANN modeling was mainly applied to fermentation and enzymatic reaction in this period. Antibiotics and ethanol were the most frequent keywords referring to products. Saccharomyces cerevisiae and Zymomonas Mobilis were two microorganisms that appeared in this period. Computer simulation, Kalman filtering, functional link networks, and heuristic methods were combined/compared with the HNNs to control the process or to optimize and/or to estimate process parameters.

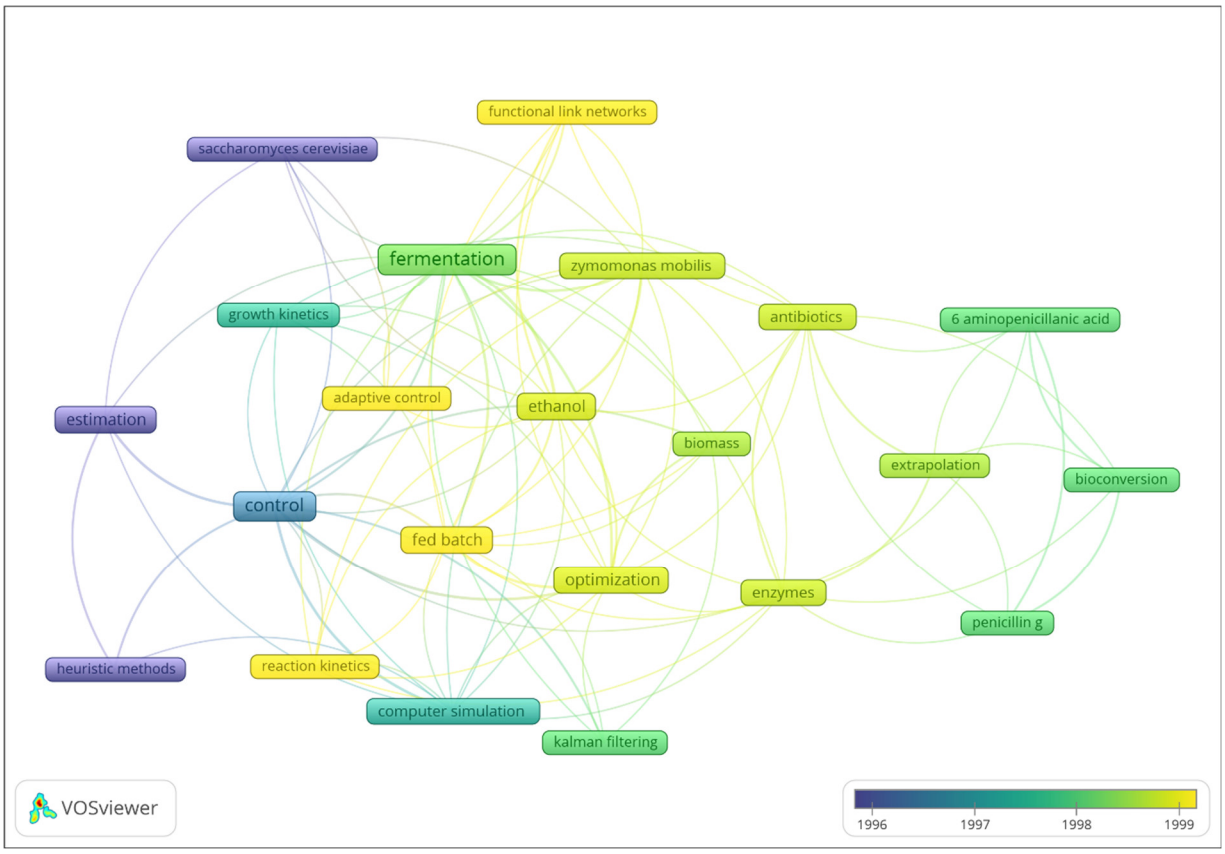

Figure S5- All keywords occurrence from 1992 until 2000 by year overlay visualization

4.1.2. Keyword occurrence from 2001 until 2010

In this period, 728 keywords were identified and then reduced to 63 by keyword analysis. Some new keywords appeared such as recombinant proteins, immunoglobulin g1, interlukin2, hydroxybutyrate, mammalian cell, CHO, cell culture, and system biology. Wastewater, immobilized enzyme, and fungal culture are other keywords that appeared during this period (Figure S6). "E. coli", "Saccharomyces cerevisiae", and "yeast" are subject microorganisms that appeared in this period.

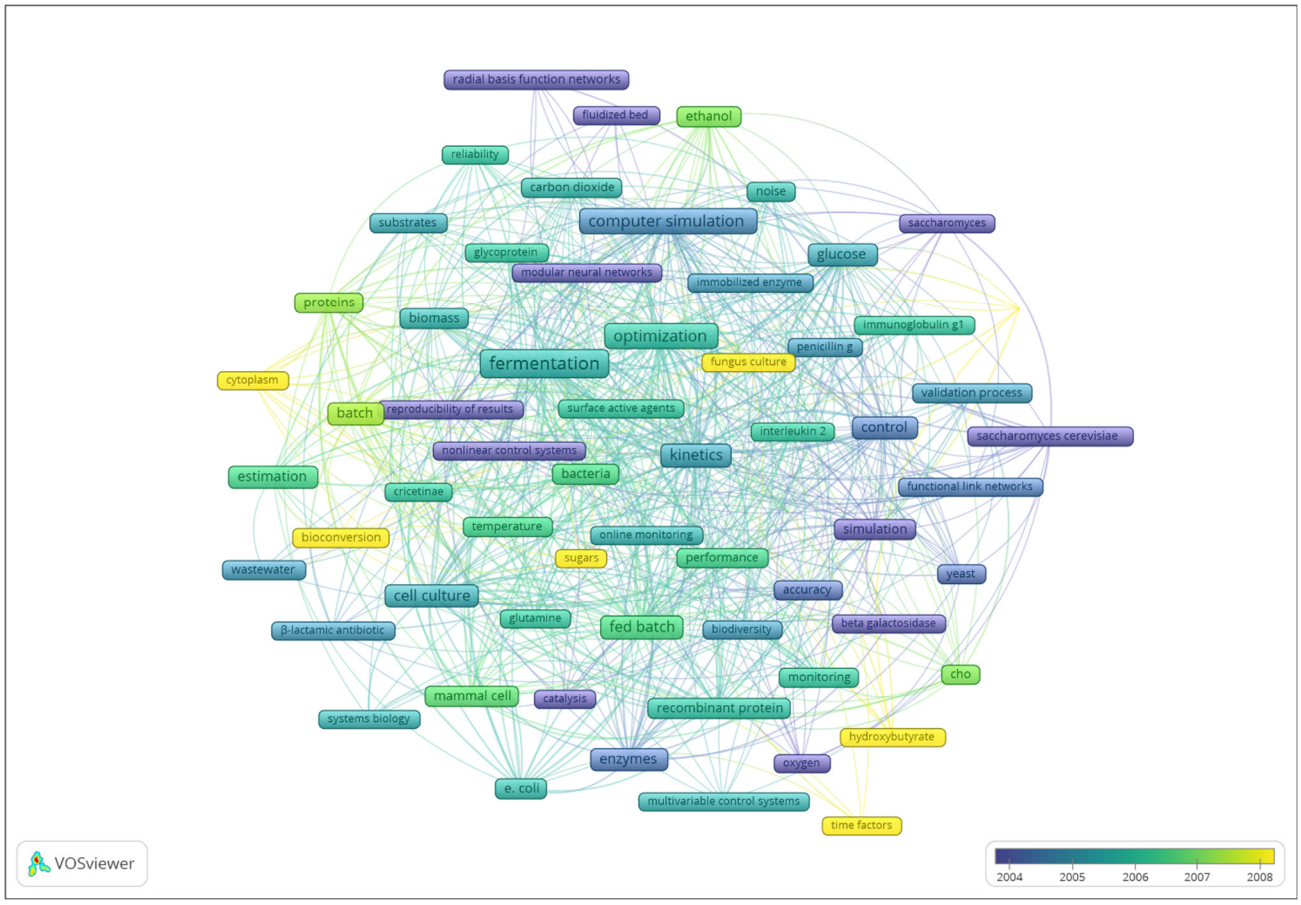

Figure S6- All keywords occurrence from 2001 until 2010 by year overlay visualization

4.1.3. Keyword occurrence from 2011 until September 2023

In this period, 1190 keywords were identified and then reduced to 63 by keyword analysis. New keywords appeared such as the design of experiments, process analytical technology, digital twins, deep learning, big data, and physic-informed neural network. These are now hot topics for the application of hybrid models to bioprocesses. Moreover, fuzzy neural network, genetic algorithm, gaussian process models, and intensified design of experiment are keywords that appeared in this period. This result suggests a growing interlink between the areas of machine learning and hybrid modeling (Figure). The “E. coli”, “Bordetella pertussis”, “yeast”, “microalgae”, and “Saccharomyces cerevisiae” are subject microorganisms that appeared in this period.

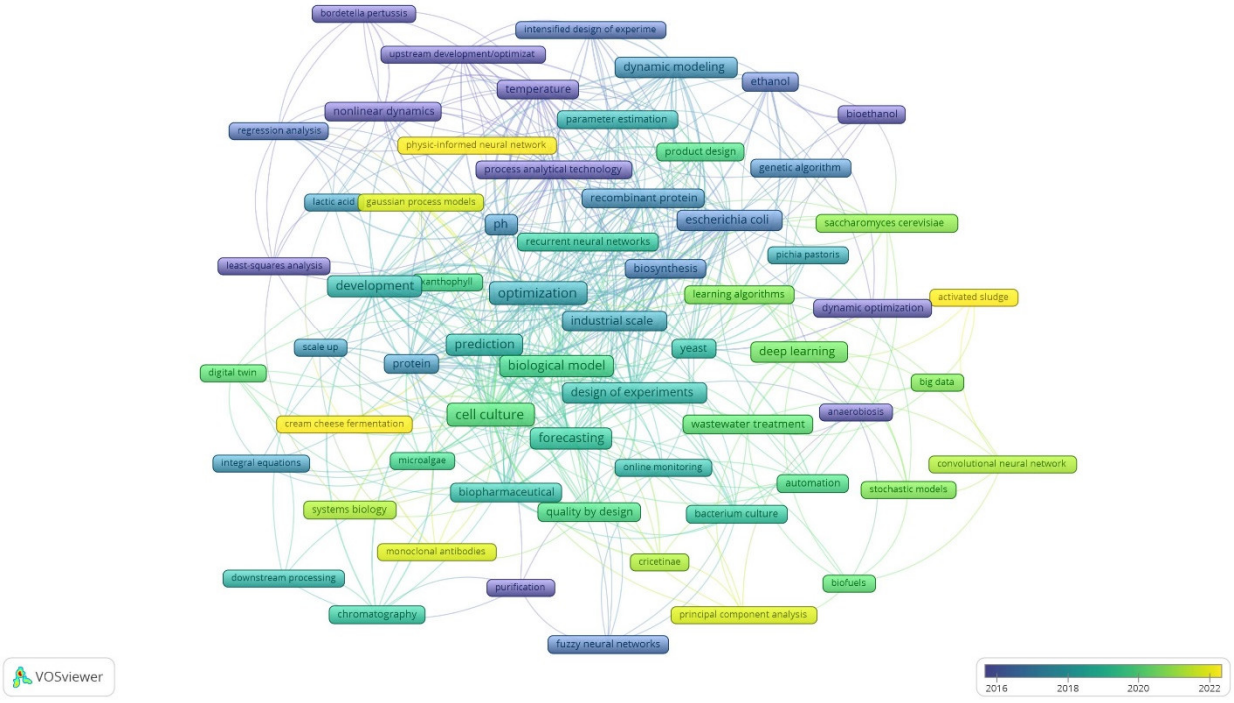

Figure S7- All keywords occurrence from 2011 until September 2023 by year overlay visualization

5. Study limitation

This systematic literature review is focused on two databases (Scopus and Web of Science). To choose the synonyms of “hybrid model” as a search keyword, many common synonyms were chosen that however do not prevent missing records. To mitigate the possibility of missing relevant publications the study was complemented with well-known authors’ search by their names. The well-known authors’ publication records were then added to the repository of relevant cases. Two additional problems were identified; 1) Some of the articles did not have the author's keywords (even recent articles). 2) Although some articles had keywords, the databases' search engines could not find them. It seems that doing a systematic literature review and automatically choosing keywords may have some bugs in categorization.