

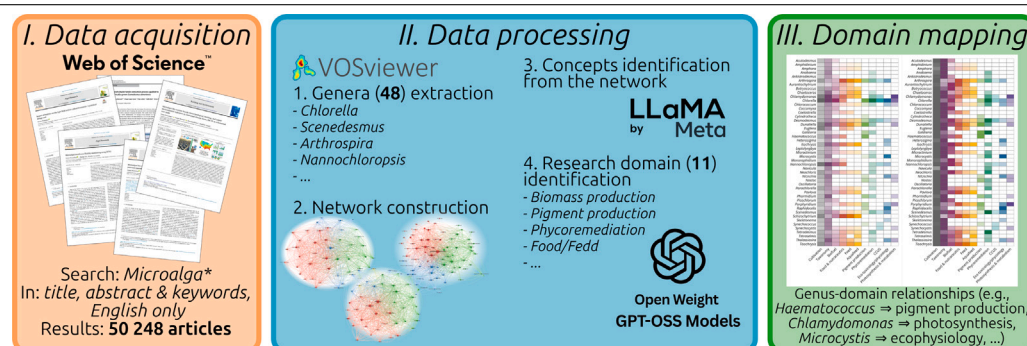


Why this genus? Mapping genus–domain relationships in microalgae research using bibliometrics and LLMs

Victor Pozzobon

Université Paris-Saclay, CentraleSupélec, Laboratoire de Génie des Procédés et Matériaux, Centre Européen de Biotechnologie et de Bioéconomie (CEBB), 3 rue des Rouges Terres 51110 Pomacle, France

GRAPHICAL ABSTRACT



ARTICLE INFO

Keywords:
Microalgae
Bibliometry
LLMs
Genus
Choice

ABSTRACT

This study presents a data-driven mapping of the relationships between major microalgal genera (48 in total) and their associated research domains. To do so, it combines bibliometrics (corpus of 50 248 articles featuring the keyword “microalga”) and Large Language Models (LLMs). An initial *semantic network* ⇒ *concepts identification* ⇒ *domains affectation* workflow yielded perfectible results. Nevertheless allowing to identify 11 research domains structuring microalgae literature (e.g., cultivation, taxonomy, biofuel, aquafeed, ...). In a second step, transitioning to a direct estimation of domain representativeness significantly improved result coherence. This refinement reduced misclassification and more effectively captured both dominant and secondary research domains for a given genus. Domain-wise normalization further clarified the relative contributions of each genus, mitigating biases arising from uneven literature coverage. Overall, the results demonstrate that, despite the inherent simplifications of bibliometric and semantic approaches, the proposed methodology reliably captures large-scale research domains and meaningful specialization patterns within the microalgae community (e.g., *Haematococcus* for pigment production, *Chlamydomonas* for photosynthesis, *Microcystis* for ecophysiology - algal bloom -, ...). From a technical standpoint, Llama 3.3 70B Instruct was the best-performing LLM for genus-specific concept identification, while GPT OSS 120B excels at semantic network-to-research domain mapping. Remaining limitations, such as residual noise in taxonomy-related concepts and sensitivity to data representation, highlight avenues for further improvement. Nonetheless, the framework provides a synthetic and interpretable overview of how microalgal diversity is functionally distributed across scientific and applied domains. From a practical standpoint, this mapping offers a clearer perspective on the implicit rationale underlying genus selection.

E-mail address: victor.pozzobon@centralesupelec.fr.

<https://doi.org/10.1016/j.biortech.2026.135343>

Received 28 April 2026; Received in revised form 22 June 2026; Accepted 6 July 2026

Available online 16 July 2026

0960-8524/© 2026 The Author. Published by Elsevier Ltd. This is an open access article under the CC BY-NC license (<http://creativecommons.org/licenses/by-nc/4.0/>).

1. Introduction

As with any research, microalgae research starts with a scientific question. The realm of possible questions is wide and ranges from eco-physiology (e.g., the doomed future of an ecological niche Williamson et al., 2019), or biology (e.g., the origin of multicellularity Boraas et al., 1998), to commodity-producing biotechnology (e.g., the feasibility of extensive biofuel production Davis et al., 2024). Fortunately for a researcher, while the explorable scientific questions are almost infinite, the prerogatives of the host intuition harboring her/his research are not. Therefore, the scope of what is to be researched is narrowed to a manageable extent. If not, then calls for proposals do take good care of that. With the general framing sorted out and, hopefully, funding acquired, the next step is to choose a model organism for the study (i.e., a strain, or at least a genus). The reasons behind this choice are the focus of this article. Indeed, this choice bears a tremendous amount of implicit knowledge, difficult to pass on, if at all. Usually, when asking why a particular strain was chosen for an investigation, the answer revolves around the fact that it is adequate for the task at hand (usually confirmed by a preliminary in-house screening or by analyzing other teams' results). Another reason that often comes up in informal discussions is the fact that the strain is already familiar to the laboratory. Indeed, any seasoned phycologist knows how difficult taming a new strain can be. Despite their simplicity, these two reasons have demonstrated merits. First, they allow limiting uncertainty and risk-taking. Second, they produce results reasonably fast. These two traits are, indeed, very desirable in the modern, ultra-competitive research landscape.

Yet, these two undeniably valid and relevant reasons are accompanied by a sizable amount of implicit knowledge, which we can describe as the researcher's expertise. The untold nature of this part of the rationale behind the choice of a model strain prevents embracing wider views, such as the overall scientific community dynamic associated with the chosen model strain, or the different subfields showing interest in the chosen model strain, and from which synergies could be drawn. Furthermore, they also restrict the horizon of possibilities to the investigator's prior knowledge and/or accessible information. Sometimes, they lead to wasting what is most valuable to a researcher: time. Indeed, including an unfit strain/genus (from a technical or regulatory perspective) in a study is a loss of laboratory workforce, financial means, and clarity in the resulting article. Sharing the rationale behind the choice could prevent it, or at least spark discussion on the matter.

A brief example of this can be found in research driving microalgae towards food applications. Authors sometimes restrict themselves to spirulina (*Arthrospira platensis*) and *Chlorella vulgaris* (Uribe-Wandurraga et al., 2021). Still, they leave the explanation of this choice to the reader's implicit knowledge. In this case, it is surely linked to the fact that those strains are approved as food by the US FDA, the EU EFSA, and the PRC National Health Commission (Ares (2022) 1668627 (European Commission, 2022), EU Novel Food Catalog (European Commission, 1997), GRN 00396 (GRAS Notices, 2012a), GRN 00417 (GRAS Notices, 2012b), announcements of August 2004 (NHC, 2004) and November 2012 (NHC, 2012)). Other authors of the same field chose to include a wider selection of species, such as *Phaeodactylum tricornutum* and *Tetraselmis suecica* (Batista et al., 2017). While this approach widens the scope of the studies, it does not expand their practical applicability.

Still, counterpoints to this example exist. The brightest one might be the numerous calls of Borowitzka to the biofuel production community (Borowitzka, 2013; Borowitzka and Vonshak, 2017), where he invited the investigators to carefully consider the choice of the strains for their studies. Indeed, as he underlined, choosing the right strains is the cornerstone of any microalgae endeavor, especially for biofuel. In this context, it means selecting a strain with a high lipid productivity (growth rate $\times$ lipid content), ease of harvest and extraction, as well as a tolerance to a broad range of temperatures and illuminations,

as demonstrated by large-scale projection (Davis et al., 2014, 2024), and systematic analysis (Nwokoagbara et al., 2015). Still, this level of awareness is not encountered in any other part of the microalgae community, where the reasons behind choosing a strain remain implicit, most of the time.

Therefore, to help navigate the landscape of the genera at play in the community, this article aims to scan the literature and extract the main domains of research tied to each genus. The process will combine bibliometric analysis and systematic processing of the data to try to get an interpretation as unbiased as possible. Bibliometrics belongs to the broader domain of scientometrics. It makes use of text processing and statistical methods to derive objective indicators and patterns out of a corpus of Refs. van Eck and Waltman (2010). Its applications are wide and range from the structure of cryptocurrency research in economics (Atree and Tripathy, 2025) to bibliometric analysis of the bibliometric analyses in pain-related research in medical science (Robert and Wilson, 2023), quite a *mise en abyme*. For a recent breakdown of the process with methodological guidelines, the reader is referred to the article of Donthu et al. (2021).

Focusing on the microalgae community, the applications of bibliometric analysis are nascent. Among them, the most prominent might be the one of Garrido-Cardenas et al. who analyzed the structure of the community (Garrido-Cardenas et al., 2018), i.e., top contributing countries (the US, China, and Spain), the top journals (Bioresour. Technology, Algal Research, and Journal of Applied Phycology), the top topics (biofuel, bioreactor - allegedly design of -, and lipids, at the time), and so on ... Very recently, two teams offered details analyzes of specific domains (EPA production and microalgae-born plant-based proteins) (Weng et al., 2026; He et al., 2026), where they deployed the same systematic approach with the same structure. Reassuringly, the authors were able to observe in their own fields the same structures as those reported by Garrido-Cardenas et al., e.g., US dominance followed by China's massive scientific production. They also dived further and analyzed the connections between entities and unveiled otherwise difficult to unearth findings. For example, their analyses demonstrated that the US leadership relies on international networks, while China's dynamic is supported autonomously, for most of it.

Owing to the maturity of the bibliometrics community, this article will rely upon its tools to analyze the published literature since the middle of the last century (date from which digital records start to be available in major databases). The goal will not be to analyze once again the structure of the community. Instead, this article aims at drawing from the systematicity of the bibliometric analysis techniques in order to identify the majority of the microalgae species encountered in publications (defined as mentioned in at least 0.1% of the number of documents) and pinpoint the topics they are associated with. To do so, on a per-genus basis, a semantic network will be drawn, and tentative clusters will be formed. Of course, this process leads to simplification of the concepts surrounding the genus, a phenomenon known as the loss of context (van Eck and Waltman, 2014). To alleviate this loss, the context is generally rebuilt afterward by experts capable of cross-fertilizing the implicit knowledge they possess with the objectivity of procedurally generated networks (Weng et al., 2026; He et al., 2026). Owing to the ambition of the task at hand (and the limited knowledge of the author), the reconstruction will also be handled automatically by state-of-the-art semantic analyzers that are Large Language Models (LLMs). This procedure aims to limit as much as possible potential author bias, while automating the analysis process (Litaina et al., 2024).

Taking a step back, this article is original in both its ambition and its methodology. First, to the best of the author's knowledge, it is the first study to systematically map genus-research domain relationships across the microalgae literature as a whole. Indeed, traditional bibliometric studies typically focus on describing the structure of a research community through elements such as countries, journals, authors, or keywords. In contrast, the proposed approach uses bibliometric methods to systematically identify and organize the scientific context surrounding individual genera, while LLMs are employed to reconstruct

the underlying concepts and research domains represented within these semantic networks. The idea of drawing synergy between LLMs and bibliometrics is an emerging field currently discussed today at various conferences among different communities. For example, LLMs can be used to process bibliometric data (e.g., affiliation) before supplying them to VOSviewer (a bibliometric analysis software, used hereinafter), for a co-authorship network analysis of teams attending a medicine conference (Bitaraf and Jafarpour, 2025) (reducing error rate from 33.9 to 2.7% during the affiliation extraction process). Still, as pointed out in other contexts, such as legal/historical documents analysis, even the most powerful commercial LLMs cannot handle an infinite amount of data (their contexts saturate around 7 pages of text). Furthermore, they perform better with English-written material. Finally, they show varying performances between models themselves (some having better context retention than others) (Litaina et al., 2024). These observations reinforce the rationale for the proposed method: bibliometric analysis (embracing all the data), keyword extraction (identifying the most relevant data to avoid LLM context dilution or, worst, saturation), concepts extraction (by an unbiased LLM, rendering two to ten key concepts to the human reader), reconstruction of a major domains of research from the concepts, and evaluation of the genus-domains relationships. This strategy allows the analysis of more than 50 000 publications while avoiding the context-length limitations that would prevent direct LLM analysis of the entire corpus. Finally, beyond providing a large-scale map of genus-domain associations, it illustrates how bibliometric analysis and LLMs can be combined to transform large volumes of scientific literature into interpretable knowledge structures that may support future data-driven research planning and genus selection.

2. Data acquisition & processing - Overall workflow

The data acquisition and processing workflow is made of five stages presented hereinafter. First is the acquisition and curation of the bibliographic corpus. Then comes the identification of the microalgal genera and the construction of genus-specific semantic networks. These networks are then used to extract the underlying concepts using LLMs, before being aggregated into broader research domains. The rationale and implementation of each stage are described in the following subsections and summarized in Fig. 1 for the data acquisition, curation, and genera identification ones.

2.1. Raw data

Data were obtained from Web of Science on September 25 2025. The research was led by topic (i.e., in the title, the abstract, and the article keywords) with only one keyword in the query “microalgae” (encompassing the singular form “microalga”). No time limit was set for the query, as the goal of the article is to encompass all the available literature. Yet, while ensuring exhaustivity, it blurs the line between current and past research trends. Using all the databases harbored by Web of Science, the search yielded 55 932 documents, out of which 5766 were reviews, and 50 248 were original research articles. The first comment is that Web of Science obviously features some overlap in the document type classification, as the sum of the two types of documents (i.e., 56 014) overpasses the total number of documents. Second, this review-to-article ratio (1.15:10, Table 1), while it may seem high, is in line with what can be encountered in other fields of physical and medical science (e.g., 1.03:10 for keyword *pyrolysis* - 271 372 documents -, 1.40:10 for keyword *cardiovascular* - 4 280 129 documents -), yet substantially above the rate in economic and social sciences (e.g., 0.19:10 for keyword *salary* - 79 620 documents -, 0.55:10 for keyword *education* - 4 773 370 documents -).

Once the articles were identified, all the bibliometric information possible was downloaded. The most important fields were title, abstract, years of publication, DOI, and accession number (unique identifier). This information was downloaded in batches of 1 000 records

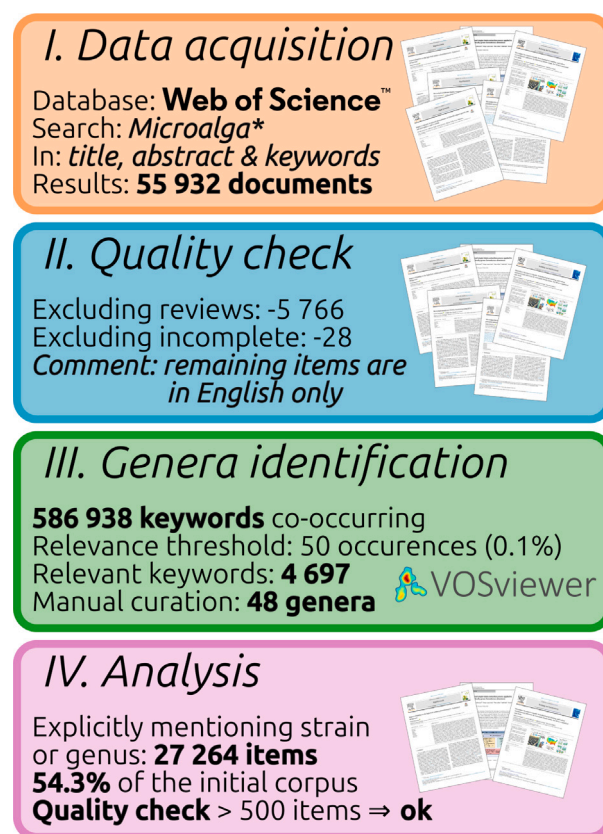


Fig. 1. Graphical summary of the data acquisition and preprocessing.

at a time (in tab-delimited format, as advised by VOSviewer, software used hereinafter). Data were then curated (removing secondary abstracts in languages other than English) and agglomerated. Restricting to English was deemed a minor loss of information as Garrido-Cardenas et al. showed that 95.2% of the microalgae literature was published in English (Garrido-Cardenas et al., 2018), while ensuring consistency of the initial corpus for the following analysis. Finally, as a safeguard procedure, potential duplicate records were searched based on their accession number. None were found. Ultimately, the agglomerated database includes 50 220 records, as 28 did not possess the essential features (title and English abstract) to be included in the study. A 0.056% rejection rate was deemed negligible.

2.2. Genera identification

The resulting corpus was analyzed thanks to VOSviewer (van Eck and Waltman, 2010). To do so, a co-occurrence analysis was carried out based on titles and abstracts. It yielded a total of 586 938 words co-occurring at least once with the term *microalgae*. In detail, binary counting was preferred to full counting as suggested by van Eck and Waltman (2014). This implies that if a word appears several times in a given title and abstract couple, it only counts as one occurrence. Hence, it avoids overestimating the occurrence of a certain term. Also, it is better suited for the purpose of the study, as the mention of a genus is a binary concept.

Despite the fact that VOSviewer prunes the most common words as well as those constituting the headers of structured abstracts, 586 938 words still represent a quantity too large to be analyzed and obviously contain a high level of noise. Therefore, a threshold at 0.1% of the number of documents (i.e., a minimum of 50 occurrences) was set in place. This measure allowed for the reduction of the number of words to be analyzed to 4 697. These words were then analyzed one by one

Table 1

Comparison of the review-to-article ratio as an indicator of the field structure and extraction procedure coherence.

Keyword	Reviews	Articles	Total documents	Fraction	Ratio for 10 article
Microalgae	5766	50 248	55 932	10.31%	1.15
Pyrolysis	25 781	251 127	271 372	9.50%	1.03
Cardiovascular	440 028	3 132 005	4 280 129	10.28%	1.40
Salary	1070	55 610	79 620	1.34%	0.19
Education	175 623	3 183 642	4 773 370	3.68%	0.55

to identify the microalgae strains and genera among them (indeed, contrary to what we can sometimes read, *Daphnia magna* is not a microalgae (Canniff and Hoang, 2018)). As a safeguard, this procedure was led twice, three months apart, and yielded identical results.

Once genera were identified, synonyms were grouped. The first part of this operation is straightforward; it consists of merging keywords such as *c vulagris*, *chlorella vulgaris*, and the likes. The second part of the process was to ensure that, despite varying names, actually identical genera were merged. This action was actually more challenging than it seems. Indeed, microalgae taxonomy has evolved significantly over time. While some authors adhere strictly to the most recent classifications, some remain faithful to the old ones, allegedly for consistency's sake. The key mergers were: *Haematococcus lacustris* and *pluvialis* (Nakada and Ota, 2016), *spirulina*, *Arthrospira*, and *Limnospira* (Sinetova et al., 2024), *Raphidocelis subcapitata*, *Selenastrum capricornutum*, and *Pseudokirchneriella subcapitata* (Machado and Soares, 2024), and *Chlorella*-related genera (e.g., *Parachlorella*), as the new taxonomy does not seem to have spread wide already (less than 200 mention of the alternate genera). Contrary to *Chlorella*-related genera, the division between *Desmodesmus* and *Scenedesmus* (Hegewald et al., 2005) is now firmly established with than 350 articles mentioning the new genus. Therefore, the two were not merged. In total, 48 genera, across 9 phyla, were identified.

Once genera were identified, the bibliographic entries were scanned once again and grouped according to the genus/genera they mention (an article may mention more than one genus and can therefore be included several times). The first comment is that only 54.3% of the bibliographic items explicitly mention at least one genus, the other simply mentioning “microalga” or “microalgae”. Still, this leaves a sizable corpus of 27 264 documents, far above the minimum of 500 items set as guidelines by Donthu et al. (2021). Before moving to the keyword identification *per se*, this preliminary processing allows for a first statistical analysis of the bibliographic landscape. Fig. 2 presents the genera and their distribution in the literature. It is to be read as “*Chlorella* genus is encountered in 35.8% of the articles”, and not “*Chlorella* genus represents in 35.8% of the global research”. Indeed, as an article can feature more than one genus, the total amount of occurrence for all the genera is 133.46%. Regarding the phyla distribution (pie chart in Fig. 2), *Chlorophytae* represents 61.5% of the total mentions. Such a dominance is not explained by the role of *Chlorella* genus alone, but reflects the contribution of several other genera, such as *Scenedesmus* or *Chlamydomonas*. In contrast, the second phylum, *Cyanobacteria*, is clearly driven by the *Arthrospira* genus, accounting for more than two-thirds of its mentions. As a token of the quality of the results as of now, these figures align well with the ones reported by Garrido-Cardenas et al. (2018).

2.3. Keywords and clusters identification

With 48 genera identified, the next step was to draw a semantic network for each of them. To do so, VOSviewer keywords identification and clustering was relied upon. Retention of a keyword by VOSviewer is based on two considerations. It should not be a common and uninformative words such as “conclusion”, “new method”, or “interesting results” (examples provided by the VOSviewer user guide). These types of words are excluded based on a relevance score computed

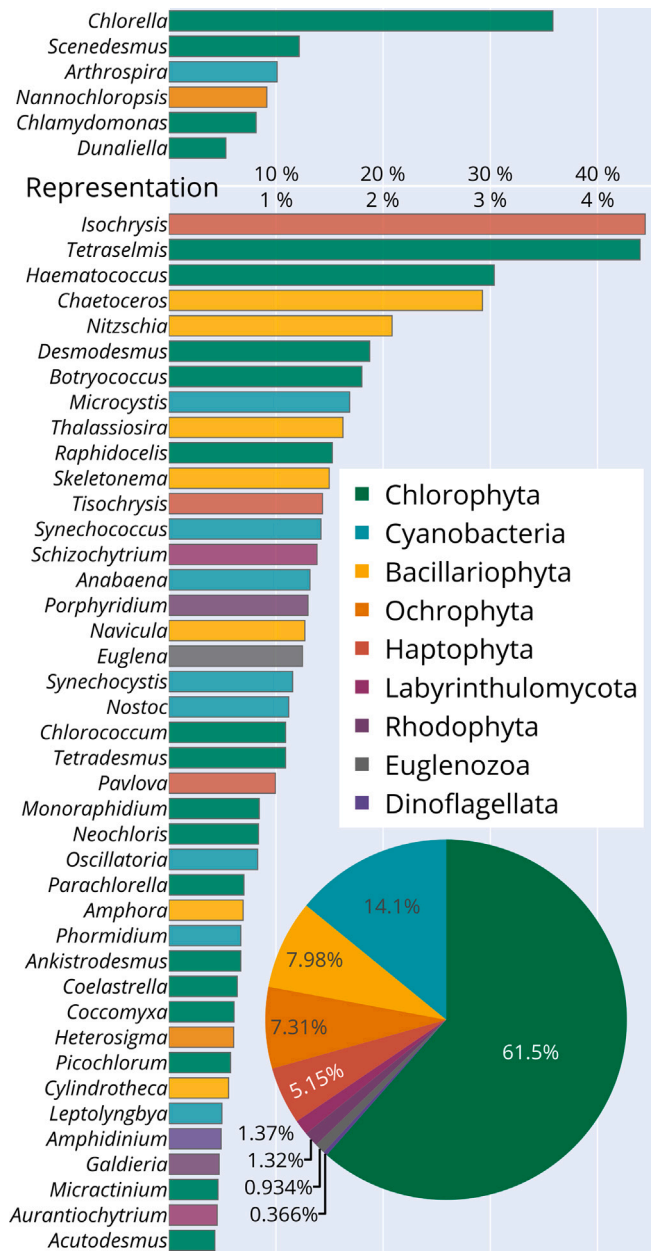


Fig. 2. Bar chart - genera and their distribution in the literature. Pie chart - phyla distribution in the literature.

by VOSviewer (based on binary counting, here). In addition, the words should appear a minimal number of times among the articles mentioning the targeted genus. Setting an adequate occurrence threshold, especially for such diverse subsets, is not trivial and is discussed further below (Section 3.1). Still, once the keywords have been identified, they are clustered based on their link strength. The clustering algorithm is

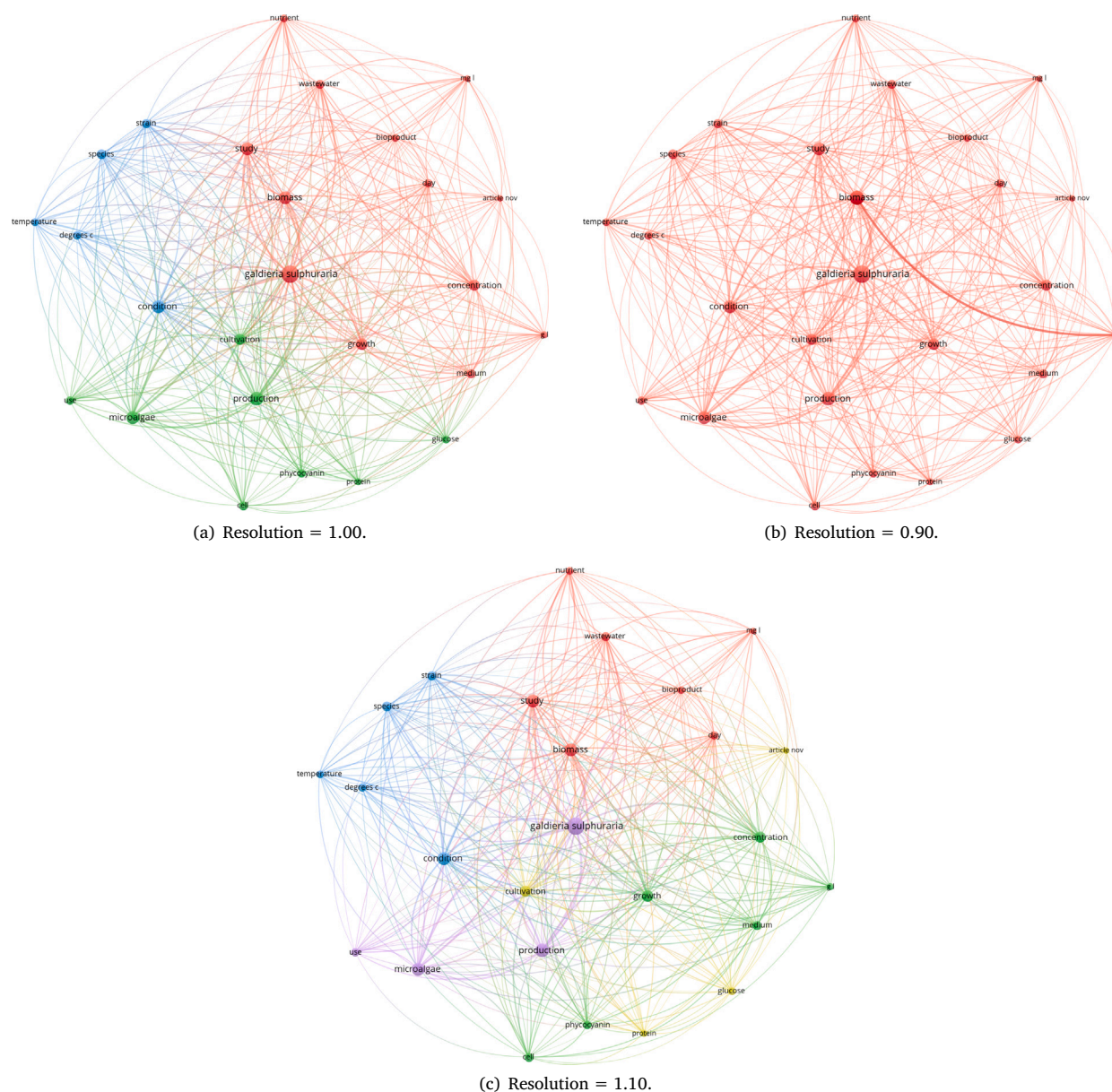


Fig. 3. Networks obtained with VOSviewer for different values of the resolution parameter in the clustering algorithm. Genus: *Galdieria*, number of article 128, minimum occurrence of keywords 15.

based on an optimization problem, described in detail by [Waltman et al. \(2010\)](#) (simultaneous mapping and clustering to be precise). As for the matter at hand, the tuning parameters were explored for both basic and advanced features. From this exploration, it was concluded that the default values already yielded stable results and were, therefore, retained. Furthermore, the most, and only, sensible parameter was the *resolution*, as underlined by the author of the algorithm ([Waltman et al., 2010](#)). A default value of 1.00 yields convincing results, increasing it to 1.10 led to cluster fragmentation, while decreasing its value to 0.90 led to network complexity reduction by cluster merging. [Fig. 3](#) illustrates these mechanisms in the case of the *Galdieria* genus. The default value identifies three distinct and coherent clusters. A value of 1.10 merges all the clusters together, while a value of 0.90 leads to the appearance of two additional clusters, of which underlying concepts are not evident.

2.4. Concept identification

Once keywords and their connections have been identified, the next step is to extract the underlying concepts. Several approaches exist. The first, and maybe best, one is to rely on the consensus of different experts coming from different scientific fields dealing with microalgae and different cultural horizons. Still, while conceptually interesting, this approach lay outside of the scope of what the author could mobilize. The second option could have been to rely on the author's expertise, which would have born numerous biases and shortcomings. An improvement of it is to rely on a group of authors, specialists of a specific topic (e.g., EPA production or microalgae-born plant-based proteins), and have them provide an analysis restricted to their area of expertise ([Weng et al., 2026](#); [He et al., 2026](#)).

Table 2

Tentative scaling for the minimum number of occurrence as a function of the number of articles for a given genus.

Scaling type	Formula	α values
Linear	$Min_{Occurrences} = \alpha \cdot Nb_{Articles}$	0.066, 0.10, 0.20
Square root	$Min_{Occurrences} = \alpha \cdot \sqrt{Nb_{Articles}}$	1, 2, 3
Logarithmic	$Min_{Occurrences} = \alpha \cdot \log_{10}(Nb_{Articles})$	10, 20, 30

Taking a step back, the question of identifying the underlying concepts supporting semantic choices is not new. In an effort to automate the process, the ConceptNet project, relying on word embedding, was developed (Speer et al., 2017). Still, the emergence of mainstream LLMs seems to have put this research on hold. Therefore, in this work, a LLM (Llama 3.3 70B Instruct, see Section 3.2) is used to analyze the network.

2.5. Domain identification

Leading the above process will make numerous concepts emerge (more than 200). These concepts may be close, feature slightly different wording for the same idea, or be each a single facet of a broader domain (e.g., β -carotene production, phycocyanin production, astaxanthin would all belong to the pigment production domain). In order to identify the domains serving as umbrellas for the concepts and render the results more legible, LLMs were asked to merge the wording belonging to the same concept and regroup the concepts into domains. It is GPT OSS 120B that performed the best at suggesting domains, which were then reviewed manually.

3. Adjusting hyperparameters

The proposed workflow relies on several numerical tools, each having a few hyperparameters to adjust. While some minor hyperparameters have already been discussed, this section focuses on the major ones. Indeed, it is important to ensure the robustness of the finding with respect to these parameters. The main parameters to adjust are the minimum occurrence of a keyword for VOSviewer to retain it, the choice of the LLM, and the prompt provided to the LLM. Finally, in addition to ascertaining key components' robustness, the reliability of the whole process is to be confirmed, using a bootstrapping method in this case.

3.1. Minimum number of occurrences

The first non-trivial parameter of importance is the number of occurrences for a keyword to be retained by VOSviewer. Indeed, it is very intuitive to feel that a one-size-fits-all value does not exist. To get a grasp of it, one only needs to consider the *Galdieria* genus, with 128 articles, and *Chlorella* genus with 9771 articles. Setting a fixed occurrence threshold, for example appearing in at least 50 articles, would lead to a too stringent selection for the first one, while being too broad for the second. This first comment allows to realize that an adequate occurrence level is surely a function of the number of articles featuring a given genus.

Consequently, several scaling methods were tested with several genera. Three types of scaling were opted for this preliminary screening: linear, square root, and logarithmic, with three levels of modulation (Table 2). In terms of genera, as discussed above, the number of articles featuring the genus matters. Therefore, five genera, covering the whole spectrum, were selected: *Chlorella* (9771 articles), *Scenedesmus* (3318 articles), *Dunaliella* (1451 articles), *Chaetoceros* (798 articles), and *Galdieria* (128 articles).

Unfortunately, none of the tested scaling led to satisfying network generation for all the genera (see Supplementary Materials for the generated networks). Still, despite having been obtained with different

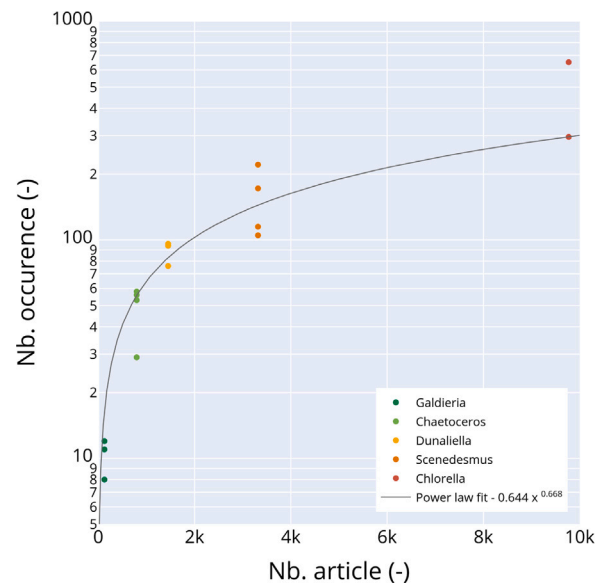


Fig. 4. Minimum keyword occurrence versus number of articles for a given genus. Power law parameters obtained using Barron's generalized loss function ($c = 0.10$ and $\alpha = -4$, with distance normed by the mean of the ordinates), with equal weight of each of the genera.

formulas, the networks deemed adequate (i.e., neither too few keywords, preventing the grasp of all the concepts, nor too many keywords, adding unnecessary noise to the concepts) build a dataset linking a number of articles and a relevant minimum occurrence threshold value. Using this dataset, it is possible to fit a phenomenological law (Fig. 4). With this procedure, a power law emerged as a good candidate. Using Barron's generalized loss function (Barron, 2019) ($c = 0.10$ and $\alpha = -4$, with distance normed by the mean of the ordinates), with equal weight of each of the genera, the following parameters were obtained: $\alpha = 0.644$ articles, and $\beta = 0.668$ (Eq. (1)).

$$Min_{Occurrences} = \alpha \cdot Nb_{Articles}^{\beta} \quad (1)$$

3.2. Choice of the LLM

Ten models were tested for this analysis (Table 3). These tested models were chosen as they are freely available (ensuring reproducibility) and easy to run (on conventional hardware, i.e., less than 20 GB GPU for a local computer, or single node 4×80 GB GPU - A100 or H100 - for a cluster, and conventional software, e.g., with Ollama or vLLM). They were also chosen for their diverse origins, increasing number of parameters (4 to 120 billion), and good performances on scientific benchmarks (provided enough parameters). Two benchmarks were targeted to be used as hallmarks GPQA:diamond and MMLU-Pro (Rein et al., 2024; Wang et al., 2024). GPQA:diamond consists of 198 questions deemed at a PhD level (elaborated by one PhD expert and validated by two independent PhD experts), initially intended to assess the performance of superhuman artificial intelligence. Each question has 4 possible answers among which the model is asked to choose. MMLU-Pro is an improved version of the original MMLU benchmark, with curated questions, focusing on reasoning (as opposed to pure knowledge-based questions), and increased distractors (choose one out of ten, instead of four). Still, evaluations on these two benchmarks were not available for all models. Therefore, variants of these benchmarks were used. While not being strictly comparable, these metrics allow for a first a priori ranking, suggesting that at least 30 billion parameters are required for a science-related task.

While performing satisfactorily on a scientific benchmark prequalifies all the models for the task at hand, examining their performance

Table 3

Selected LLM performances (percentage of adequate answers) on GPQA and MMLU benchmarks families.

Sources: Model presentation Web pages.

Model	MMLU family		GPQA		
	MMLU-Pro	MMLU	Regular	Super	Diamond
Phi 3.5 MoE instruct	54.3	–	36.8	–	–
Qwen3 8B	56.8	76.9	44.4	31.6	–
Qwen3 30B A3B	78.4	–	70.4	–	–
Instruct 2507	–	–	–	–	–
Qwen3 30B A3B	78.5	–	65.8	51.8	–
Thinking 2507	–	–	–	–	–
Qwen3.5 27B	86.1	–	–	65.6	–
Gemma 2 27B	–	75.2 (5 shot)	–	–	–
Instruct	–	–	–	–	–
Gemma 4 31B	85.2	–	–	–	84.3
Mixtral 8 × 7B	–	70.6	29.2	–	–
Instruct	–	–	–	–	–
Llama 3.3 70B	68.9	88.6	–	–	49
Instruct	–	–	–	–	–
GPT OSS 120B	90	–	–	–	80.1

on a subset of the networks to be analyzed is more relevant. The three genera selected for this test were *Chlorella*, *Dunaliella*, and *Galdieria*, owing to their contrast in terms of available articles. To perform these tests, two parameters were adjusted: temperature and top-p. The temperature was set to 0 for one trial, and to 0.1 for 5 trials. The goal was to favor models' seriousness, while allowing for slight variation. The top-p parameter was set to 1 to enrich as much as possible the models' vocabulary. Finally, the regular prompt (see next Section and Supplementary Materials) was used to instruct the LLMs.

While detailed results (all the returned concepts for all the models arranged in tables) and analyses are available in Supplementary Materials to satisfy the kind reader's curiosity, the only main findings will be summarized here. Out of the ten tested models, two failed to deliver the instructed output (Mixtral 8x7B Instruct and Phi 3.5 MoE Instruct), five performed the tasks but missed critical features, such as the importance of β -carotene for *Dunaliella* genus and phycocyanin for *Galdieria* genus (Qwen3 8B, Qwen3 30B A3B Instruct 2507, Qwen3 30B A3B Thinking 2507, Gemma 2 27B Instruct, and Gemma 4 31B), two successfully achieved the identification and reporting of the key concepts (Qwen3.5 27B and GPT OSS 120B), but would require subsequent prompt engineering to streamline the results (e.g., limit the strict adherence they enforce to the initial number of clusters identified VOSviewer and the resulting lumping of the concepts). Finally, it is Llama 3.3 70B Instruct model that delivers the best results. Overall, the concepts identified by the model were systematically relevant and included both primary and secondary, i.e., marginal yet relevant, concepts. Its only flaw was tendency to multiply concepts for *Galdieria* genus, when temperature was set to 0.10.

3.3. Prompt engineering

Finally, as a last parameter to adjust, the sensitivity of the model to the prompting style was investigated. Three prompting styles were tested with different levels of intensity in the directives given to the model (full prompts available in Supplementary Materials using *Galdieria* genus as an example, owing to the limited length of the network passed to the model). In all cases, the prompts were designed following a clear structure: presentation of the task, instructions to carry out the tasks, and examples of the expected outputs (few-shot alike approach). Adopting a permissive prompt style led the model to struggle to adhere to the requested output (supplying extra explanations). In addition, the lower guidance resulted in the model missing important elements such as β -carotene for *Dunaliella* genus and phycocyanin for *Galdieria* genus. Furthermore, it reduced the proposed

Table 4

Comparison of the concept identified by Llama 3.3 70B Instruct model for the two most efficient prompting style (temperature of 0). Underlined: key differences.

Genus	Basic prompt	Directive prompt
<i>Chlorella</i>	Microalgae cultivation	Microalgae cultivation
	Biomass production	Biomass production
	Wastewater treatment	Biofuel production
	Lipid production	Wastewater treatment
	Biofuel production	Lipid production
	Biotechnology	Photosynthesis research
	Photosynthesis research	Biotechnology applications
	Carbon capture	Microalgal biomass production
	Nutrient removal	Nutrient removal
	Protein production	Carbon capture
<i>Dunaliella</i>	Microalgae cultivation	Microalgae cultivation
	Lipid production	Lipid production
	Protein production	Biomass production
	Biomass production	Growth rate optimization
	Wastewater treatment	Carotenoid production
	Aquaculture feed production	Wastewater treatment
	Carotenoid production	Protein production
	Bioproduct application	Marine microalgae cultivation
	Photosynthesis research	Bioproduct application
	Marine ecology	Photosynthesis research
<i>Galdieria</i>	Microalgae cultivation	Microalgae cultivation
	Bioproduct production	Bioproduct production
	Cell growth medium	Wastewater treatment
	Wastewater treatment	Microalgae growth
	Phycocyanin production	Phycocyanin production
	Biomass production	Nutrient management
	Nutrient condition	Biomass production
	<u>Glucose utilization</u>	Cell growth medium

concepts for *Galdieria* genus, leading to an oversimplification of the network. Therefore, this prompting style was ruled out. Opting for a more directive prompt redaction did not alter the output dramatically. Yet, some subtlety was lost (Table 4), and despite underlining that the model should pay attention to secondary concepts (such as the ones underlined in Table 4), it misses them. Consequently, the initial prompting style was retained for the rest of the study.

3.4. Bootstrap assessment - Overall workflow robustness

Beyond the sensitivity analyses presented above, the robustness of the workflow as a whole was evaluated using a bootstrapping process. The objective was to assess whether the workflow remained stable when only a fraction of the available literature was considered. To do so, three representative genera were selected: *Chlorella* (9 771 articles), *Dunaliella* (1 451 articles), and *Galdieria* (128 articles). These genera span nearly two orders of magnitude in corpus size and therefore provide a suitable stress test for evaluating the stability of the proposed methodology.

For each genus, 80% of the articles were randomly sampled without replacement, and the complete workflow was repeated, from keyword extraction and network generation to genus-domain relationship determination. This procedure was repeated 100 times. The cumulative average genus-domain relationship was monitored as a function of the number of bootstrap iterations. Despite the large differences in corpus size, the estimated domain representativeness rapidly converged towards stable values (see Supplementary Materials). For the three genera, the stabilization was achieved after approximately 20–30 bootstrap iterations. Furthermore, the converged values were highly consistent with the domain assignments obtained from the complete corpus, indicating that the results are not driven by a small subset of publications but rather reflect stable structures present throughout the literature. Consequently, results obtained with the refined workflow (see Section 4.2) were produced by averaging 30 bootstrap iterations (limited to this case because of the computational cost).

4. Results & discussion

4.1. Initial results

Once the workflow has been set and its metaparameters fixed, it was run for all the genera identified (48 in total, model temperature of 0 and top-p of 1). This led to the emergence of wording for concepts, of which many were redundant (e.g., microalgae production, biomass production, ...). In order to limit the noise, LLMs were asked to merge the wording belonging to the same concept and regroup the concepts into domains. It is GPT OSS 120B that performed the best on this task, and, with minimal human curation, eleven domains were identified (not ranked in a particular order):

1. Cultivation (encompassing microalgae production, growth, condition optimization, productivity, ...)
2. Taxonomy
3. Biofuel (including lipids for biofuel)
4. Food and nutraceuticals (encompassing protein and high-value compounds)
5. Feed
6. Aquaculture feed
7. Pigment production (carotenoids, phycocyanin, chlorophyll, ...)
8. Wastewater treatment/phycoremediation
9. Carbon Capture Utilization and Sequestration (a.k.a. CCUS)
10. Ecotoxicology and ecophysiology (encompassing harmful algal blooms)
11. Photosynthesis and metabolism

As rendering the detailed outcome of the analysis for all the genera would make the manuscript quite unpleasant to read, the concepts identified for each genus and the domain they belong to are listed in the Supplementary Materials. The main body of the manuscript will render them as a heat map (Fig. 5). The results are presented as the association between a specific domain and a genus. The intensity of the color is an indicator of the number of times this association is reported for a given couple. As one can see, the microalgae cultivation is common to all the genera and often the most represented, some exceptions being *Raphidocelis* (ecotoxicology as first domain), *Nannochloropsis* (biofuel as first domain), or *Tetraselmis* (aquafeed as first). This observation illustrates the fact that for most of the studies, producing the cells is an important and sometimes challenging stage. While it may seem obvious to any phycologist, it is worth taking a step back. Indeed, by comparing to another field the author is accustomed to (lignocellulosic biomass thermochemical conversion), it may appear as an oddity, as in this second field, researchers simply purchase the biomass on the market and refer to the wood species only, with minimal details on the biomass production, if at all.

Moving to the analysis of the other largely represented domains, one can note the sizable presence of biofuel, the focus of the microalgal biotechnology community over the 2008–2014 period (He et al., 2026), and food (including nutraceuticals), the current focus of the community. Then comes wastewater treatment/phycoremediation, where the proposed approach underlines the roles of *Chlorella* genus, and ecotoxicology/ecophysiology, where the role of *Raphidocelis* is pinpointed.

While detailing the successes achieved by the proposed workflow, it may be more interesting to pay close attention to where it falls short. For example, the *Chlorella* genus is not associated with the food domain, despite its high relevance to this domain being underlined in the introduction. Also, *Microcystis* genus is flagged as having phycoremediation as primary domain, while ecophysiology would have been more relevant, as this microalgae is studied for its involvement in toxic algal blooms. As a third example underlining the need for improvement in the workflow, *Haematococcus* genus is not represented as a top genus for pigment production. Another comment is that the

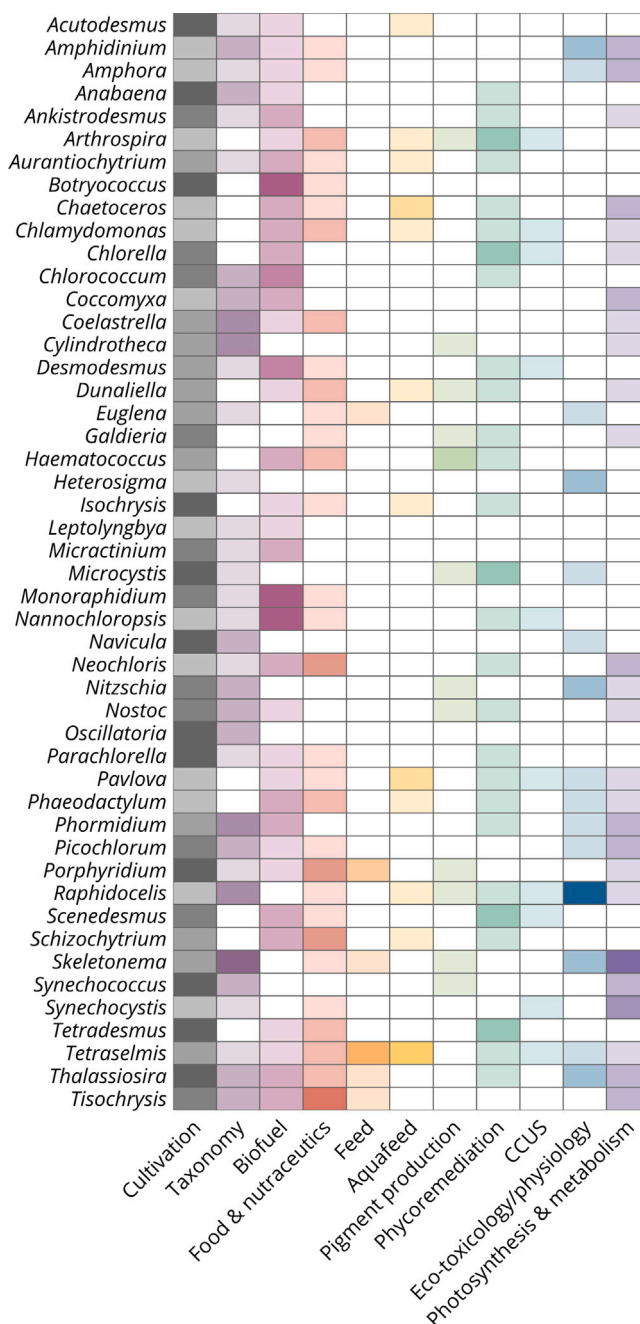


Fig. 5. Association between a specific domain and a genus. Opacity computed as the minimum of 1 and (number of association)/5.

taxonomy domain seems overrepresented. By analyzing the detailed results (Supplementary Materials), one can realize that when a genus or strain name appears, the LLM classifies the concept as taxonomy. While the mechanism is understandable, finding another genus in a semantic network underlines the co-occurrence with the analyzed genus (e.g., synergy/competition in the case of wastewater treatment) and should not lead to the classification of the concept as taxonomy. This column, therefore, suffers from a high false positive rate. Finally, while the eleven identified research domains can be deemed to cover most of the literature, the most recent and niche topics, such as bio-electricity (Yi et al., 2025), may pose some challenges.

Finally, the way the outcomes are presented also calls for criticism. The scaling applied in Fig. 5 allows for a quick overview of the

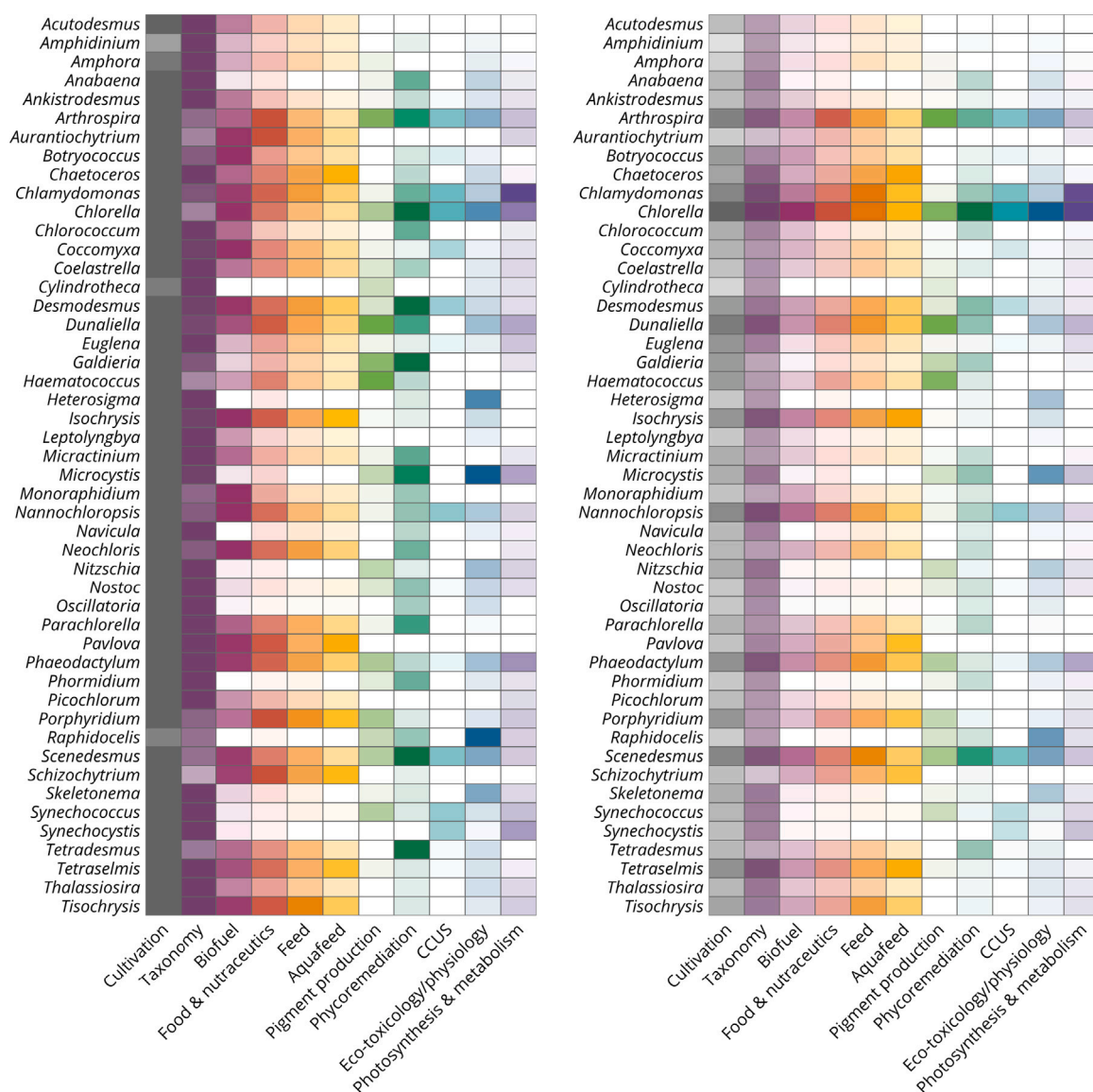


Fig. 6. Association between a specific domain and a genus. Left - genus-wise scaling (excluding cultivation and taxonomy domains). Right - domain-wise scaling. Opacity scaled following a root square transformation. Linear scaling available in Supplementary Materials.

outcomes, but does not provide much guidance, genus-wise or domain-wise. Revising it to norm by domain entry would allow for faster identification of the genera contributing to it, while norming by genus entry would allow for a better grasp of the domain it is represented in.

4.2. Refined results

The first step to refine the preliminary results is to take a step back and acknowledge what worked and what did not. On the one hand, the identification of eleven domains can be deemed reliable. Indeed, as they are very large-scale features, identified over the whole corpus, there is little doubt that they actually structure the research. On the other hand, the affection of concept to domain may sometimes be criticized. Even the reduction of the network to an intermediary concept is to be questioned. Indeed, the *network* $\mapsto$ *concepts identification* $\mapsto$ *domains affectation* workflow may be deemed to introduce an unnecessary restriction of the complexity with its intermediary step. Especially now that a pool of domains has been identified. Consequently, a reduced version of the workflow *network* $\mapsto$ *domains representation estimation* was set up.

In a nutshell, instead of identifying concepts from the semantic network to then assign them to a domain, for each domain, the LLM was asked to evaluate the percentage of presence of the domain in the network. The obtained percentages of representativeness were then scaled, genus-wise (Fig. 6 - Left) and domain-wise (Fig. 6 - Right). One should note that on both subfigures, a square root scaling of the opacity was applied to enhance the visualization of the results. In addition, the domain-wise scaling was weighted by the representativeness of the genus in the corpus (square root of the percentage identified in Fig. 2) to avoid a very focused genus to outweigh the others (e.g., *Chlamydomonas* genus for the photosynthesis and metabolism domain), while also preventing the *Chlorella* genus from overshadowing the others.

This new approach calls for several comments. First of all, it is able to alleviate the shortcomings of the initial one. Namely, *Chlorella* genus plays an important role in the food domain, the reasons behind its relevance being already detailed in the introduction (regulatory approval by the three leading agencies worldwide). *Microcystis* genus primary domain (ecophysiology) is now adequately reported, both genus-wise and domain-wise. Indeed, this genus and its main representative, *Microcystis aeruginosa*, have been under scrutiny for years owing to their deleterious harmful algal blooms (de Figueiredo et al., 2004).

Haematococcus genus role in the investigation of pigment production is now clearly identifiable at first glance. Here again, any phycologist knows that the interest in this genus is driven by its long documented ability to produce astaxanthin (Higuera-Ciupara et al., 2006).

Secondly, this approach underlines the importance of the microalgal production stage for all the studies. As this domain was overbearing in the analysis, it was excluded from the scaling to enhance legibility. The same was done for the taxonomy. Indeed, despite an attempt to reduce it, the false positive in this domain remains relatively high compared to what is reasonable to expect. This last observation remains a shortcoming of a process, which, otherwise, shows great merit.

5. Conclusion

In this work, a data-driven framework combining bibliometrics and Large Language Models (LLMs) was developed to map the relationships between major microalgal genera (appearing in more than 0.1% of a corpus exceeding 50 248 articles) and their associated research domains. The initial *network* $\rightarrow$ *concepts identification* $\rightarrow$ *domains affectation* workflow enabled the identification of eleven robust domains structuring microalgae research.

Building on this foundation, a refined approach based on the direct estimation of domain representativeness proved more effective, improving coherence, reducing misclassification, and capturing both dominant and secondary research themes. Combined with appropriate normalization strategies, this approach provides a clearer, more balanced view of genus-domain associations, thereby limiting biases induced by uneven literature coverage.

Overall, the results show that the proposed methodology can reliably capture large-scale research domains and reveal meaningful specialization patterns within the microalgae community, despite the inherent simplifications of bibliometric and semantic analyses. While limitations remain, particularly regarding residual noise in taxonomy-related concepts and sensitivity to data representation, the framework establishes a solid basis for further methodological improvements.

Beyond its methodological contribution, this study offers a synthetic and interpretable overview of how microalgal diversity is functionally distributed across scientific and applied domains. By making explicit the collective research landscape associated with each genus, it provides a complementary perspective to traditional experience-driven genus selection. Overall, it provides a more informed and transparent rationale for selecting model organisms in microalgae research and could open the way to AI-assisted genus selection.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

Communauté urbaine du Grand Reims, Département de la Marne, Région Grand Est and European Union (FEDER Grand Est 2021–2027) are acknowledged for their financial support to the Chair of Biotechnology of CentraleSupélec and the Centre Européen de Biotechnologie et de Bioéconomie (CEBB). This work was performed using computational resources from the “Mésocentre” computing center of Université Paris-Saclay, CentraleSupélec and École Normale Supérieure Paris-Saclay supported by CNRS and Région Île-de-France (<https://mesocentre.universite-paris-saclay.fr>). ILaaS initiative support was invaluable in leading this research.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at <https://doi.org/10.1016/j.biortech.2026.135343>.

Data availability

Data will be made available on request.

References

- Atree, M.K., Tripathy, N., 2025. Cryptocurrency research: Bibliometric review and content analysis. *Int. Rev. Econ. Financ.* 98, 103940. <http://dx.doi.org/10.1016/j.iref.2025.103940>, URL <https://www.sciencedirect.com/science/article/pii/S1059056025001030>.
- Barron, J.T., 2019. A general and adaptive robust loss function. pp. 4331–4339, URL https://openaccess.thecvf.com/content_CVPR_2019/html/Barron_A_General_and_Adaptive_Robust_Loss_Function_CVPR_2019_paper.html.
- Batista, A.P., Nicolai, A., Fradinho, P., Fragoso, S., Bursic, I., Rodolfi, L., Biondi, N., Tredici, M.R., Sousa, I., Raymundo, A., 2017. Microalgae biomass as an alternative ingredient in cookies: Sensory, physical and chemical properties, antioxidant activity and in vitro digestibility. *Algal Res.* 26, 161–171. <http://dx.doi.org/10.1016/j.algal.2017.07.017>, URL <https://www.sciencedirect.com/science/article/pii/S2211926417302953>.
- Bitaraf, E., Jafarpour, M., 2025. AI-powered affiliation insights: LLM-based bibliometric study of European medical informatics conferences. In: *Intelligent Health Systems – from Technology to Data and Knowledge*. IOS Press, pp. 823–827. <http://dx.doi.org/10.3233/SHTI250474>, URL <https://ebooks.iospress.nl/doi/10.3233/SHTI250474>.
- Boraas, M.E., Seale, D.B., Boxhorn, J.E., 1998. Phagotrophy by a flagellate selects for colonial prey: A possible origin of multicellularity. *Evol. Ecol.* 12 (2), 153–164. <http://dx.doi.org/10.1023/A:1006527528063>.
- Borowitzka, M.A., 2013. Species and Strain Selection. In: Borowitzka, M.A., Moheimani, N.R. (Eds.), *Algae for Biofuels and Energy*. Springer Netherlands, Dordrecht, pp. 77–89. http://dx.doi.org/10.1007/978-94-007-5479-9_4.
- Borowitzka, M.A., Vonshak, A., 2017. Scaling up microalgal cultures to commercial scale. *Eur. J. Phycol.* 52 (4), 407–418. <http://dx.doi.org/10.1080/09670262.2017.1365177>, eprint: <https://doi.org/10.1080/09670262.2017.1365177>.
- Canniff, P.M., Hoang, T.C., 2018. Microplastic ingestion by *Daphnia magna* and its enhancement on algal growth. *Sci. Total Environ.* 633, 500–507. <http://dx.doi.org/10.1016/j.scitotenv.2018.03.176>, URL <https://www.sciencedirect.com/science/article/pii/S0048969718309355>.
- Davis, R.E., Fishman, D.B., Frank, E.D., Johnson, M.C., Jones, S.B., Kinchin, C.M., Skaggs, R.L., Venteris, E.R., Wigmosta, M.S., 2014. Integrated evaluation of cost, emissions, and resource potential for algal biofuels at the national scale. *Environ. Sci. Technol.* 48 (10), 6035–6042. <http://dx.doi.org/10.1021/es4055719>.
- Davis, R., Hawkins, T.R., Coleman, A., Gao, S., Klein, B., Wiatrowski, M., Zhu, Y., Xu, Y., Snowden-Swan, L., Valdez, P., Zhang, J., Singh, U., Ou, L., 2024. Economic, Greenhouse Gas, and Resource Assessment for Fuel and Protein Production from Microalgae: 2022 Algae Harmonization Update. Tech. Rep. NREL/TP-5100-87099, National Renewable Energy Laboratory (NREL), Golden, CO (United States). <http://dx.doi.org/10.2172/2318964>, URL <https://www.osti.gov/biblio/2318964>.
- de Figueiredo, D.R., Azeiteiro, U.M., Esteves, S.M., Gonçalves, F.J.M., Pereira, M.J., 2004. Microcystin-producing blooms—a serious global public health issue. *Ecotoxicol. Environ. Safety* 59 (2), 151–163. <http://dx.doi.org/10.1016/j.ecoenv.2004.04.006>, URL <https://www.sciencedirect.com/science/article/pii/S0147651304000740>.
- Donthu, N., Kumar, S., Mukherjee, D., Pandey, N., Lim, W.M., 2021. How to conduct a bibliometric analysis: An overview and guidelines. *J. Bus. Res.* 133, 285–296. <http://dx.doi.org/10.1016/j.jbusres.2021.04.070>, URL <https://www.sciencedirect.com/science/article/pii/S0148296321003155>.
- European Commission, 1997. EU Novel Food Status Catalogue – Entry for *Arthrospira Platensis* Gomont 1892. Database / Regulatory Guidance Tool, European Commission, Brussels, URL <https://ec.europa.eu/food/food-feed-portal/screen/novel-food-catalogue/search>.
- European Commission, 2022. Consultation Request to Determine the Novel Food Status of the *Chlorella* Sp. Report / Consultation Document Ares (2022) 1668627, European Commission, Brussels, URL https://food.ec.europa.eu/system/files/2022-03/novel-food_consult-status_chlorella-sp.pdf.
- Garrido-Cardenas, J.A., Manzano-Agugliaro, F., Acien-Fernandez, F.G., Molina-Grima, E., 2018. Microalgae research worldwide. *Algal Res.* 35, 50–60. <http://dx.doi.org/10.1016/j.algal.2018.08.005>, URL <https://www.sciencedirect.com/science/article/pii/S2211926418303436>.
- GRAS Notices, 2012a. GRN No. 396 *Chlorella Vulgaris* Strain Beij. GRAS Notice (Generally Recognized as Safe Notification), U.S. Food and Drug Administration, United States, URL <https://hfpappexternal.fda.gov/scripts/fdcc/index.cfm?set=GRASNotices&id=396>.
- GRAS Notices, 2012b. GRN No. 417 Dried Biomass of *Arthrospira Platensis* Also Known as *Spirulina Platensis* (Spirulina). GRAS Notice (Generally Recognized as Safe Notification), U.S. Food and Drug Administration, United States, URL <https://hfpappexternal.fda.gov/scripts/fdcc/index.cfm?set=grasnotices&id=417>.

- He, J., Xu, H., Lu, K., Guo, C., Du, M., Jia, R., 2026. Research trends and hotspots of microalgae protein as a functional ingredient in plant-based foods: A bibliometric analysis based on the web of science. *Algal Res.* 94, 104541. <http://dx.doi.org/10.1016/j.algal.2026.104541>, URL <https://linkinghub.elsevier.com/retrieve/pii/S2211926426000263>.
- Hegewald, E., Schmidt, A., Braband, A., Tsarenko, P., 2005. Revision of the Desmodismus (Sphaeropleales, Scenedesmaceae) species with lateral spines. 2. The multi-spined to spineless taxa. *Algological Stud.* 116 (1), 1–38.
- Higuera-Ciajara, I., Félix-Valenzuela, L., Goycoolea, F.M., 2006. Astaxanthin: A review of its chemistry and applications. *Crit. Rev. Food Sci. Nutr.* 46 (2), 185–196. <http://dx.doi.org/10.1080/10408690590957188>, eprint: <https://doi.org/10.1080/10408690590957188>.
- Litaina, T., Soularidis, A., Bouchouras, G., Kotis, K., Kavakli, E., 2024. Towards LLM-based semantic analysis of historical legal documents.
- Machado, M.D., Soares, E.V., 2024. Features of the microalga *Raphidocelis subcapitata*: physiology and applications. *Appl. Microbiol. Biotechnol.* 108 (1), 219. <http://dx.doi.org/10.1007/s00253-024-13038-0>.
- Nakada, T., Ota, S., 2016. What is the correct name for the type of *Haematococcus* Flot. (Volvocales, Chlorophyceae)? *Taxon* 65 (2), 343–348. <http://dx.doi.org/10.12705/652.11>, URL <https://onlinelibrary.wiley.com/doi/abs/10.12705/652.11>, eprint: <https://onlinelibrary.wiley.com/doi/pdf/10.12705/652.11>.
- NHC, 2004. Reply of the Ministry of Health on the Use of Certain Raw Materials in General Foods. Official Reply / Regulatory Notice, PRC National Health Commission, Beijing, URL <https://www.nhc.gov.cn/sps/c100088/200408/bf86e4274bdb4f64b0b227b6b1db2c68.shtml>.
- NHC, 2012. Announcement on the Approval of Protein-Pyrenoid Chlorella and Three Other New Resource Foods (Announcement No. 19 of 2012). Government Notice / Regulatory Announcement, PRC National Health Commission, Beijing, URL <https://www.nhc.gov.cn/sps/c100088/201212/836e39de1b214a8abe5fde06ac824236.shtml>.
- Nwokoagbara, E., Olaleye, A.K., Wang, M., 2015. Biodiesel from microalgae: The use of multi-criteria decision analysis for strain selection. *Fuel* 159, 241–249. <http://dx.doi.org/10.1016/j.fuel.2015.06.074>, URL <https://www.sciencedirect.com/science/article/pii/S0016236115006547>.
- Rein, D., Hou, B.L., Stickland, A.C., Petty, J., Pang, R.Y., Dirani, J., Michael, J., Bowman, S.R., 2024. Gpqa: A graduate-level google-proof q&a benchmark. In: *First Conference on Language Modeling*.
- Robert, C., Wilson, C.S., 2023. Thirty-year survey of bibliometrics used in the research literature of pain: Analysis, evolution, and pitfalls. *Front. Pain Res.* 4, <http://dx.doi.org/10.3389/fpain.2023.1071453>, URL <https://www.frontiersin.org/journals/pain-research/articles/10.3389/fpain.2023.1071453/full>.
- Sinetova, M.A., Kupriyanova, E.V., Los, D.A., 2024. Spirulina/Arthrospira/Limnospira—Three names of the single organism. *Foods* 13 (17), 2762. <http://dx.doi.org/10.3390/foods13172762>, URL <https://www.mdpi.com/2304-8158/13/17/2762>. Number: 17.
- Speer, R., Chin, J., Havasi, C., 2017. ConceptNet 5.5: An open multilingual graph of general knowledge. *Proc. AAAI Conf. Artif. Intell.* 31 (1), <http://dx.doi.org/10.1609/aaai.v31i1.11164>, URL <https://ojs.aaai.org/index.php/AAAI/article/view/11164>.
- Uribe-Wandurraga, Z.N., Igual, M., Reino-Moyón, J., García-Segovia, P., Martínez-Monzó, J., 2021. Effect of microalgae (*arthrospira platensis* and *chlorella vulgaris*) addition on 3D printed cookies. *Food Biophys.* 16 (1), 27–39. <http://dx.doi.org/10.1007/s11483-020-09642-y>.
- van Eck, N.J., Waltman, L., 2010. Software survey: VOSviewer, a computer program for bibliometric mapping. *Scientometrics* 84 (2), 523–538. <http://dx.doi.org/10.1007/s11192-009-0146-3>.
- van Eck, N.J., Waltman, L., 2014. Visualizing bibliometric networks. In: Ding, Y., Rousseau, R., Wolfram, D. (Eds.), *Measuring Scholarly Impact: Methods and Practice*. Springer International Publishing, Cham, pp. 285–320. http://dx.doi.org/10.1007/978-3-319-10377-8_13.
- Waltman, L., van Eck, N.J., Noyons, E.C.M., 2010. A unified approach to mapping and clustering of bibliometric networks. *J. Inf.* 4 (4), 629–635. <http://dx.doi.org/10.1016/j.joi.2010.07.002>, URL <https://www.sciencedirect.com/science/article/pii/S1751157710000660>.
- Wang, Y., Ma, X., Zhang, G., Ni, Y., Chandra, A., Guo, S., Ren, W., Arulraj, A., He, X., Jiang, Z., 2024. Mmlu-pro: A more robust and challenging multi-task language understanding benchmark. *Adv. Neural Inf. Process. Syst.* 37, 95266–95290.
- Weng, Q., Lin, X., Luo, L., Tan, L., Liu, Q., Lu, S., Li, G., Zhang, Z., 2026. Research hotspots and evolutionary trends in microalgae-derived EPA production: A bibliometric review. *Algal Res.* 94, 104505. <http://dx.doi.org/10.1016/j.algal.2025.104505>, URL <https://www.sciencedirect.com/science/article/pii/S2211926425006162>.
- Williamson, C.J., Cameron, K.A., Cook, J.M., Zarsky, J.D., Stibal, M., Edwards, A., 2019. Glacier algae: A dark past and a darker future. *Front. Microbiol.* 10, URL <https://www.frontiersin.org/articles/10.3389/fmicb.2019.00524>.
- Yi, F., Yang, L., Zhang, L., Peng, B.-y., Tan, X., Hong, Y., You, X., Zhou, X., Zhang, Y., 2025. Enhanced photosynthetic carbon fixation in microalgae through tandem synergies in electron transport flows. *Chem. Eng. J.* 512, 162505. <http://dx.doi.org/10.1016/j.cej.2025.162505>, URL <https://www.sciencedirect.com/science/article/pii/S1385894725033315>.