Fottea, Olomouc, 25(1): 1–11, 2025
DOI: 10.5507/fot.2024.009

Diatom studies three decades into the molecular age: a bibliometric analysis reveals genetic underexploration of diatoms compared to other taxa

Jan Kollár[1,*], Kateřina Kopalová[1,2] & Tyler J. Kohler[1]

[1]*Department of Ecology, Faculty of Science, Charles University, Viničná 7, CZ–12844, Prague, Czech Republic; *Corresponding author e–mail: jan.kollar.phd@gmail.com*
[2]*Department of Phycology, Institute of Botany, Czech Academy of Sciences, Dukelská 135, CZ–37901, Třeboň, Czech Republic*

Abstract: Diatoms are among the most diverse and environmentally significant protists on Earth. Like many other organismal groups, a large portion of their diversity appears to lie beyond the resolution of the traditional light microscopy–based methods routinely and sometimes exclusively utilized in their investigation. Although the technological and conceptual developments in the fields of molecular biology and bioinformatics unlocked a remarkable opportunity to study diatoms in a previously unimagined depth and breadth, molecular diatomists anecdotally claim that diatoms remain genetically understudied compared to other taxa. However, this claim has never been quantified and rigorously tested. Therefore, we performed a bibliometric analysis of over 42,000 WoS–indexed diatom documents published in the past 35 years, between 1988 and 2023. The claim is confirmed: only ~15% of the analyzed diatom literature incorporated molecular data, about half compared to other groups, including other algae, cyanobacteria, plants, fungi, and animals. Interestingly, research for all groups seems to asymptotically saturate with molecular methods once they are used in about one–third of the documents annually, an observation which has important implications. In addition, past trends in the use of molecular data in diatomology were explored and some future ones were predicted.

Keywords: diatoms, Bacillariophyta, barcoding, DNA, genetics, omics, molecular data, molecular phylogenetics, molecular taxonomy, RNA

Introduction

Diatoms (Bacillariophyta) are single–celled photoautotrophic eukaryotes (or algae) encased in silica cell walls called 'frustules'. These frustules are highly ornamented, and for over 300 years have served as the primary guide for exploring diatom diversity (Round et al. 1990). With an estimated 100,000 extant species (Mann & Droop 1996; Mann & Vanormelingen 2013), diatoms are among the most diverse, abundant, and environmentally significant protists on Earth, playing an important role in the biogeochemical cycles (e.g., silica, oxygen, carbon) and food webs of many aquatic and semiaquatic ecosystems (e.g., Falkowski et al. 1998). As such, they have immense potential not only in the fundamental research fields such as ecology, evolution, and paleolimnology (e.g., Lewis et al. 2008), but also in applied disciplines such as biomonitoring (e.g., Smol & Stoermer 2010), biotechnology (e.g., Hu et al. 2008) and nanotechnology (e.g., Drum & Gordon 2003). Many of these rely increasingly on molecular methods for the characterization and study of diatom strains, populations, and communities.

Over the last few decades, molecular analytical methods have effectively transformed our understanding of the natural world, in terms of quantifying and investigating the origins of Earth's diversity, as well as understanding how communities are assembled at both local and global scales. While there have been major milestones in molecular technologies over the last century (e.g., Watson & Crick 1953; Sanger et al. 1977; Mullis et al. 1986), the advent of next–generation sequencing (Mardis 2013) has arguably had among the greatest impacts on the widespread use of molecular methods in the life sciences due to an increased quality and quantity of generated data, as well as improved affordability. As a result, the number of molecular–based studies in the life sciences has greatly increased over the last decades, including within diatomology.

The first diatom DNA sequence (SSU rDNA of marine *Skeletonema costatum* (Greville) Cleve) was published in 1988 (Medlin et al. 1988), heralding the onset of the molecular age for diatomology. Nearly two decades later, the first (*Thalassiosira pseudonana*

Hasle et Heimdal; Armbrust et al. 2004) and second (*Phaeodactylum tricornutum* Bohlin; Bowler et al. 2008) diatom genomes were published, followed by the first multigene diatom phylogenies (Alverson et al. 2007; Theriot et al. 2010). Also, in the last couple of decades, identification through short DNA fragments extracted from both diatom strains (DNA barcoding; Hebert et al. 2003) and environmental samples (metabarcoding; e.g., Ficetola et al. 2008) became increasingly widespread, revolutionizing how diatom lineages and communities are characterized and studied.

Although the intricately ornamented silica frustules of diatoms exhibit immense morphological diversity, it does not match their enormous species diversity, as evidenced by numerous discoveries of cryptic species across the diatom tree of life (e.g., Mann & Evans 2008; Pinseel et al. 2020). Under such circumstances, and given the importance of diatoms in biomonitoring and other applications (Smol & Stoermer 2010), DNA (meta) barcoding shows particular promise thanks to its ability to identify organisms objectively, affordably, and (once the reference barcode databases are sufficiently developed) without a need for specialized knowledge of the local floras (Mann et al. 2010). As a result, broadscale biogeographical assessments are possible (e.g., Malviya et al. 2016; Chonova et al. 2023), as well as the study of even the most extensive cryptic species complexes (such as *Pinnularia borealis* Ehrenberg, with over 100 discovered species and ca 400 estimated; Pinseel et al. 2020) which are practically irresolvable using traditional microscopy–based methods.

Yet, a notion exists within the (molecular) diatomological community that these methods have not been used to their potential and, as a result, that diatoms are genetically understudied in comparison to other taxonomic groups. However, this claim remains anecdotal, as it has never been rigorously tested. Here, 35 years after the publication of the first diatom DNA sequence (Medlin et al. 1988), we put the claim to the test by performing a bibliometric analysis of diatomological research published between 1988 and 2023, along with research published for other taxonomic groups for comparison. Specifically, we asked: (1) Are molecular methods utilized less frequently in the field of diatomology compared to the other taxonomic fields? (2) What do we study with molecular methods in diatomology? (3) Can we identify some past trends in the use of molecular methods in diatomology, and can we make some predictions for its future?

Materials and Methods

Data collection. Only WoS–indexed documents published between 1988 and 2023 were targeted. For diatoms, three main bibliometric datasets were assembled: (1) a dataset containing all the diatom research (denoted as 'whole'), (2) a dataset containing only the molecular diatom research ('molecular'), and (3) a dataset containing only the non–molecular diatom research ('non–molecular'). Additionally, (4) a dataset containing diatom research utilizing electron microscopy was assembled ('em') for reasons discussed below (Results and Discussion).

The query for the 'whole' dataset consisted of a taxon–defining element and a timespan–defining element. Diatoms were defined by (diatom* NOT diatomite* NOT diatomous* NOT diatomic* NOT diatomaceous earth) OR (bacillariophy*) searched through titles, keywords, and abstracts using the Topic field (TS). The timespan was defined by PY=(1988–2023). The query for the 'molecular' dataset was identical except for an addition of a part broadly defining molecular data, i.e., molecular OR DNA OR rDNA OR RNA OR rRNA OR eDNA OR aDNA OR "gene" OR genetic OR DNA barcod* OR metabarcod* OR SSU OR 16S OR 18S OR LSU OR 28S OR 5S OR rbc* OR cox1 OR COI OR COXI OR psb* OR matK OR microsatellite* OR MSAT* OR AFLP OR RFLP OR *genom* OR *proteom* OR *transcript* OR *metabolom* OR amplicon OR NGS OR next–generation sequencing OR HTS OR high–throughput sequencing OR Sanger sequencing OR nucleotide OR SNP* OR PCR OR polymerase chain reaction OR *exom* OR exon* OR intron*, likewise searched through TS. The 'non–molecular' dataset was obtained by removing the documents of the 'molecular' dataset from the 'whole' dataset in R version 4.2.1 (R Core Team 2022) using the function anti_join() in the dplyr v. 1.0.10 R package (Wickham et al. 2023). The 'em' dataset was obtained as the 'whole' with the addition of electron microscopy defining part searched through TS: SEM OR "electron microscop*" OR TEM OR EM OR ultrastructur*.

For the comparison with other taxonomic groups, analogical 'whole' and 'molecular' datasets (i.e., differing only in the taxonomic parts of the queries) were assembled for cyanobacteria, rhodophytes, chlorophytes sensu stricto (i.e., green algae excluding charophytes), charophytes, phaeophytes, chrysophytes s.s. (i.e., excluding synurophytes), synurophytes, eustigmatophytes, dinophytes, cryptophytes, euglenophytes, glaucophytes, embryophytes, fungi, and metazoans. These were selected to broadly represent major algal and/or multicellular groups with an aim to compare diatoms to (1) comparable groups (algae), i.e., composed of unicellular primary producers with similar research history, and (2) presumably the most explored groups (i.e., land plants, fungi, and animals; e.g., Pawlowski et al. 2012; Larsen et al. 2017). Cyanobacteria, formerly known as the blue–green algae, were included for historical reasons. Likewise, mostly multicellular phaeophytes and rhodophytes are included to cover all major lineages traditionally studied by the science of algology. The overall dataset analyzed in this study potentially exceeds four million documents (Tab. S1).

Prior to downstream analyses, datasets were subjected to additional filtering to remove residual documents that matched the queries but not necessarily the intended scope of the analyses. All datasets were downloaded on 1 January 2024. Nevertheless, some database lag is inevitable, meaning that some documents published in 2023, and possibly even in 2022, were not yet present in the WoS database by the download date. The exact search queries are summarized in the Supplementary Material (Note S1). The datasets are publicly available in BibTeX format in JK's repository (https://www.researchgate.net/profile/Jan–Kollar–6).

Data analysis. Data were analysed in R v. 4.2.1 (R Core Team 2022) using packages bibliometrix v. 0.3.0 (Aria & Cuccurullo 2017), dplyr v. 1.0.10 (Wickham et al. 2023), tidyverse v. 1.3.2 (Wickham et al. 2019), and ggraph v. 2.1.0 (Pedersen 2024). Besides basic descriptive statistics, some of the known mathematical principles of scientific bibliometrics

were explored, namely Lotka's Law (Lotka 1926) and Bradford's Law (Bradford 1934). For a given scientific field, Lotka's Law predicts the number of authors with a given number of publications. On the other hand, Bradford's Law identifies core journals for the given field, i.e., journals collectively publishing one–third of the field's scientific production. In addition to these quantitative insights into the assembled datasets, we were interested in what topics are studied with molecular methods in diatomology, suspecting that it might be important for interpreting the results. For practical reasons, this conceptual structure of the field was explored only for a subset of the diatom 'molecular' dataset.

Firstly, core molecular diatom research was identified by searching diatoms (taxonomic part of the query) and molecular data (methodological part; Note S1) only through document titles (TI) and author keywords (AK), instead of more inclusive topics (TS; thus omitting abstracts and WoS's Keyword Plus©). Secondly, to further reduce the dataset, only documents broadly focused on diatom diversity, ecology, and evolution (including areas such as diatom taxonomy, biogeography etc.) were targeted using the WoS categories: WC= ("Plant Sciences" OR "Evolutionary Biology" OR "Ecology" OR "Biodiversity Conservation" OR "Marine Freshwater Biology" OR "Oceanography" OR "Environmental Sciences" OR "Limnology" OR "Microbiology" OR "Multidisciplinary Sciences" OR "Palaeontology" OR "Biology") NOT WC= ("Biotechnology Applied Microbiology"). This reduced the dataset size to 20% of the original 'molecular' diatom dataset, and the result was denoted as the 'mol–eco–evo' dataset.

The conceptual structure of the field of molecular diatomology was explored using several kinds of analyses implemented in the bibliometrix R package. Firstly, a coupling analysis constructed a network of documents based on shared references. In other words, if both document A and document B reference document C, then documents A and B are considered similar, i.e., their connection within the network is strengthened (Aria & Cuccurullo 2017). Secondly, a co–citation analysis constructed a network of documents based on how often they are co–cited by other documents. In other words, if both document A and document B are referenced by document C, then documents A and B are considered similar. Finally, co–occurrence was constructed with Keywords Plus© as the unit of analysis instead of documents, contrary to the previous two. The logic of the analysis is as follows: Both keyword A and keyword B are found in the same article. Therefore, both keywords describe related topics and are thus clustered together.

Results and Discussion

Dataset properties

The 'whole' dataset of diatom research published between 1988 and 2023 contained 42,215 documents. These were published in 3,543 sources (journals, books, etc.) by 76,125 authors, with 32.9% based on an international collaboration. Only 32.4% of authors published ≥2 documents, in accordance with a prediction by Lotka's Law (Lotka 1926; Fig. S1).

The molecular diatom research of the past 35 years (the 'molecular' dataset) was encapsulated in 6,496 (15.4%) documents published in 1,162 sources. These were produced by 18,049 authors, with 37.7% resulting from international collaborations. An average document in the 'molecular' dataset was 10.4 years old, written by 5.1 co–authors, and received 36.6 citations. In comparison, the 'non–molecular' dataset encapsulated 35,719 (84.6%) documents published in 3,206 sources by 65,310 authors, and 32.0% of these resulted from international collaborations. An average 'non–molecular' document was 14.2 years old, written by 4.2 co–authors, and received 31.0 citations. Therefore, molecular diatom research seems to be (on average) conducted by larger, more international research teams, and receive more attention from the scientific community based on citations per document (Fig. 1). In all three datasets, the most frequent document type was research article (86.2 ± 0.7%) and the molecular diatom research included relatively more reviews than the non–molecular (7.7% vs 3.2%, respectively), possibly reflecting both the relative novelty of the molecular methods to the community (compared to the long–established methodologies of non–molecular diatom research) over the studied timespan and the dynamic nature of their development.

It was argued that for some research questions in some areas of diatom studies (e.g., diatom taxonomy, phylogenetics, and classification) sufficient resolution is achievable by supplementing the standard light microscopy observations with electron microscopy (EM), which greatly enhances the range of available characters that are potentially informative for diatom species delimitations, identifications, and/or phylogeny reconstructions (e.g., Witkowski et al. 2015). Therefore, to obtain some insight into the question of whether diatomists tend to substitute molecular data with EM, the dataset containing research utilizing EM was additionally assembled (the 'em' dataset). It contained 3,216 documents (7.6%) published in 739 sources by 7,161 authors. Around 42% of the documents resulted from an international collaboration. An average 'em' document was 11.8 years old, written by 4.0 co–authors, and received 21.8 citations. These results do not seem to support the idea of the substitution of molecular data with electron microscopy observations, and surely not at the community–wide scale. While it may be possible that EM–users tend to mention electron microscopy in titles and keywords less frequently compared to mentions of molecular data/techniques by molecular diatomists, this is unlikely to obscure the results as the search through Topic field (TS) includes other bibliographic data, such as abstracts.

Lastly, the subset of the molecular dataset used to analyze the conceptual structure of the field (the 'mol–eco–evo' dataset) contained 1,295 documents (20% of the 'molecular' dataset), published in 270 sources by 3,820 authors. Almost 44% of these documents originated through international collaboration. An average document in this dataset was 9.6 years old, written by 5.3 co–authors, and received 33.3 citations.

Comparison of diatoms and other taxa

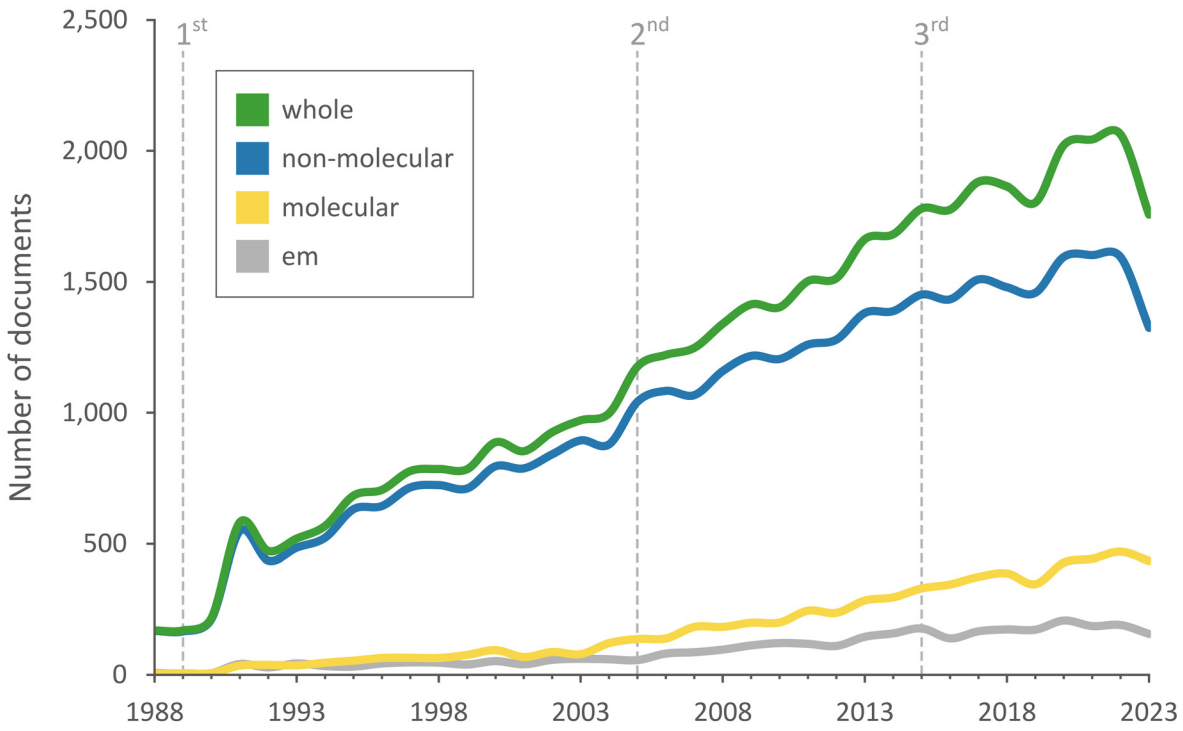

Fig.1. Annual scientific production in diatomology between 1988 and 2023: green = 'whole' (n = 42,215 documents); blue = 'non–molecular' (35,719); yellow = 'molecular' (6,496); grey = 'em' (electron microscopy; 3,216); dashed lines = major milestones in the development of molecular technology (1st = first commercial PCR thermocycler; 2nd = second–generation sequencing platforms developed; 3rd = Oxford Nanopore third–generation sequencing technology commercialized). Trends of annual production (both 'whole' and 'molecular') observed in different parts of the timespan (e.g., exponential increase after 1989, or sudden drop in 2022–2023) are universal across taxa (Fig. S3). The drop in 2022 and 2023 was most likely caused by database lag, although some influence of the COVID–19 pandemic cannot be excluded.

The proportion of research incorporating molecular methods to study diatoms accounted for 15.4%, around half of the average proportion found in the other taxonomic groups (30.1 ± 7.4%; Fig. 2). Similarly, the proportion of molecular research in other algae averaged at 29.2 ± 8.1%, or with the inclusion of cyanobacteria which were historically treated as blue–green algae, 29.8 ± 8.0%. The closest proportions to diatoms were found in euglenophytes (17.9%), and golden algae (i.e., chrysophytes and synurophytes with 21.1 ± 4.9%), whose taxonomy is largely based on the morphology of solid shell structures (such as loricae or silica scales; e.g., KRISTIANSEN 2005), similarly to diatoms. These were followed by groups composed largely of macroscopic photoautotrophs, such as embryophytes, charophytes, and phaeophytes (26.7 ± 1.1%). Thus, the anecdotal claim, namely that molecular methods are underutilized in the field of diatomology compared to other taxonomic groups, seems valid. Note that while designing the query for assembling the dataset, extensive testing was performed to find a reasonable balance between data quantity and quality. Nevertheless, this main result, namely that diatoms are the least explored by molecular means, and that it is roughly by half compared to the other taxa, was consistent across designs. For example, an application of a stricter query (searching only through document titles and keywords, in comparison to the final query, which searched also through abstracts) resulted in a far smaller dataset of ca 18.9 thousand diatom documents, with 1.6 thousand of them utilizing molecular methods. These 8.5% were again roughly half of the average found for other taxa (15.8%) when applying the same stricter query design.

Surprisingly, the proportion of research utilizing molecular methods over time was discovered to be approximately asymptotic in most of the taxonomic groups, with a notable exception of the diatoms (see Fig. 3, eventually Fig. S2, and Fig. S3). While the trend was unclear for a subset of taxonomic groups due to strongly oscillating curves in their plots (i.e., eustigmatophytes, glaucophytes, and synurophytes), upon closer examination of the dataset, it is clear that these patterns are due to the relatively small size of the respective datasets (with the 'whole' research of 35 years encapsulated in 543, 439, and 367 documents, respectively). However, this same ambiguity cannot be claimed for the diatoms, whose dataset was the fifth largest among the compared groups (with >42,000 documents), surpassed only by embryophytes (ca 1.9 million documents), metazoans (1.6 million), fungi (0.3 million), and cyanobacteria (49,000; Tab. S1). Furthermore, even cryptophytes, with the 'whole' dataset composed of a relatively modest 3,158 documents, exhibited an approximately asymptotic relationship between time and the proportion of molecular research (Fig. S3).

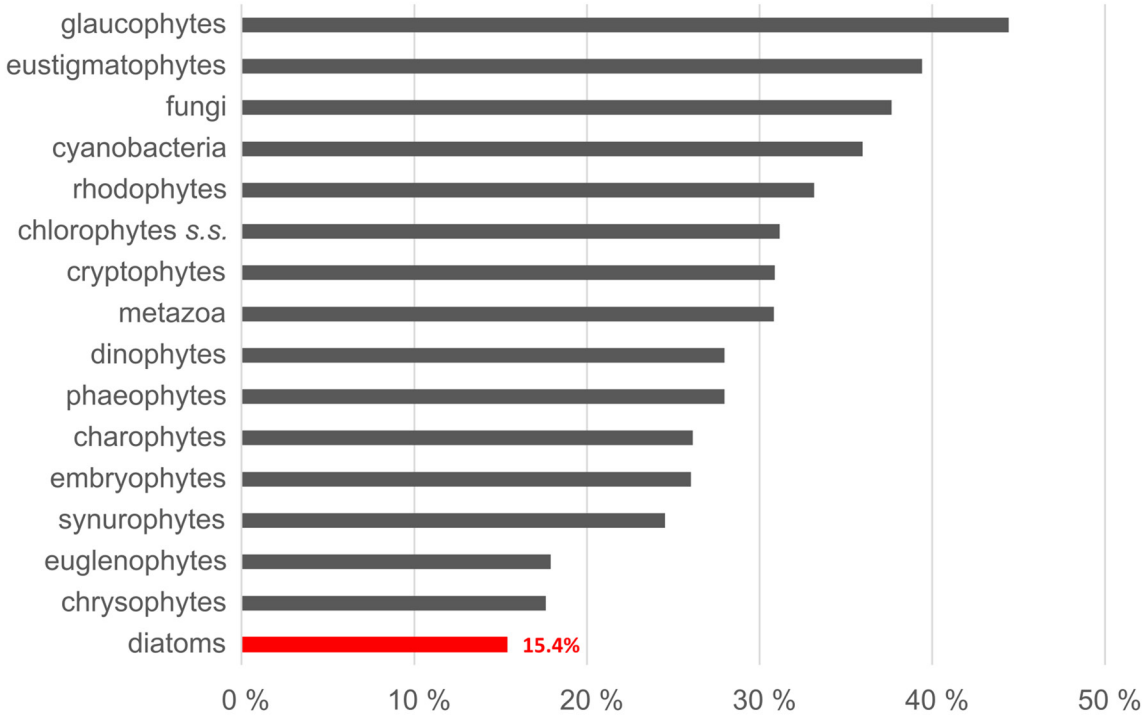

Fig. 2. An overall proportion of 1988–2023 research incorporating molecular methods in a study of different taxonomic groups: the proportion in diatoms is roughly half compared to the averages based on the rest of the taxonomic groups (30.1%), the rest of algae (29.2%), and the major macroscopic groups (metazoans, fungi, and embryophytes; 31.5%). The algal average is not dramatically inflated by the top–scoring small groups, namely glaucophytes and eustigmatophytes (algal median is 28.0%).

Furthermore, as the curve in diatoms does not show any signs of levelling off (Fig. 3), we hypothesize that the plateau of the asymptote remains distant.

Universally across taxonomic groups, the plots of annual scientific production show certain trends, or changes in publication rates (Fig. 1, Fig. S3). Firstly, a period of exponential increase in annual scientific production between 1989 and 1991 may be partially connected to the introduction of the 1st commercial PCR thermocycler to the market, but an analogical increase in 'non–molecular' research suggests other influences, such as the spread of information technologies, or even drastic geopolitical changes induced by the dissolution of the Soviet Union. Secondly, production rate changes over the period between ca 2002 and 2015, corresponding with the development and commercialization of the second–generation sequencing technologies such as 454 pyrosequencing and Solexa/Illumina–sequencing (Satam et al. 2023). Thirdly, a further increase in production rate between 2016 and 2021 correlates with the advent of a third–generation sequencing technology Oxford Nanopore (Wang et al. 2021). Finally, a sudden drop in 2022–2023 is most likely caused by database lag, meaning that some documents published in 2023, and possibly even in 2022, were not yet uploaded to the WoS by the download date. Nevertheless, some influence of the COVID–19 pandemic cannot be excluded. More certainty on this matter can be achieved by repeating the analysis in a few years, once the database lag can be excluded.

These above–mentioned correlations do not necessarily imply causation, which are beyond the intended scope of this contribution. Nevertheless, increases in both 'molecular' and 'non–molecular' research suggest that advances in sequencing technology are not the (sole) cause of increases in scientific productivity, although they can affect the field indirectly (e.g., the knowledge gained from the sequencing of the representative genomes can stimulate both molecular and non–molecular research performed on given taxa, by raising new questions and opening previously unknown research routes; e.g., Theissinger et al. 2023). In addition, both the development of a sequencing technology and scientific productivity are surely influenced by increasing computational capacities, as well as by generally growing numbers of scientists worldwide (Fortunato et al. 2018).

Molecular saturation of the research fields

Perhaps more importantly, the asymptotic relationship between time and proportion of molecular research was found universally across taxa, and it seems to always plateau between roughly 25 and 45% (Fig. 3 and Fig. S2), i.e., when around one–third of the research incorporates molecular data, suggesting saturation of the research fields by molecular methods. The universal existence of this saturation level for molecular methods is intriguing and has relevant implications.

Firstly, although diatoms are demonstrably and considerably

understudied with molecular methods in comparison to other taxa (Fig. 2), no 'call to (molecular) arms' is needed because the proportion of diatom molecular research is increasing (Fig. 3), and there seems to be an endpoint (i.e., the saturation level of ca 1/3). Thus, both trajectory and target exist, and therefore, it is only a matter of time before diatomology matches the other fields in the use of molecular methods. Based on the trends depicted in Fig. 3, we predict that diatomology might reach the threshold of ~1/3 molecular research, and thus possibly a saturation of the field by molecular methods, somewhere between 2033 and 2043.

Secondly, concerns exist that molecular methods may overrun taxonomic fields, leading to a loss of expertise in areas of organismal morphology and anatomy (e.g., EBACH et al. 2011; LÖBL et al. 2023). However, the existence of the saturation level would imply that individual research fields cannot be easily overrun by molecular methods. On the contrary, our results suggest that most of the research (possibly around 2/3 annually) does not, and likely will not, include any molecular methods at all. While molecular methods allow for the study of organisms in a previously unimagined depth (e.g., via whole genome sequencing) and breadth (e.g., via environmental metabarcoding), there always will be a lot to study without them, although this is more true for some research aims and scales than others. For example, exploration of diversity in taxonomic groups with a widespread occurrence of cryptic species, such as diatoms, benefits greatly from the resolution and practicality of molecular data, although alternative approaches are not impossible (e.g., geometric morphometrics, crossbreeding culture experiments, physiological culture experiments; MANN & EVANS 2008). Likewise, large–scale surveys of

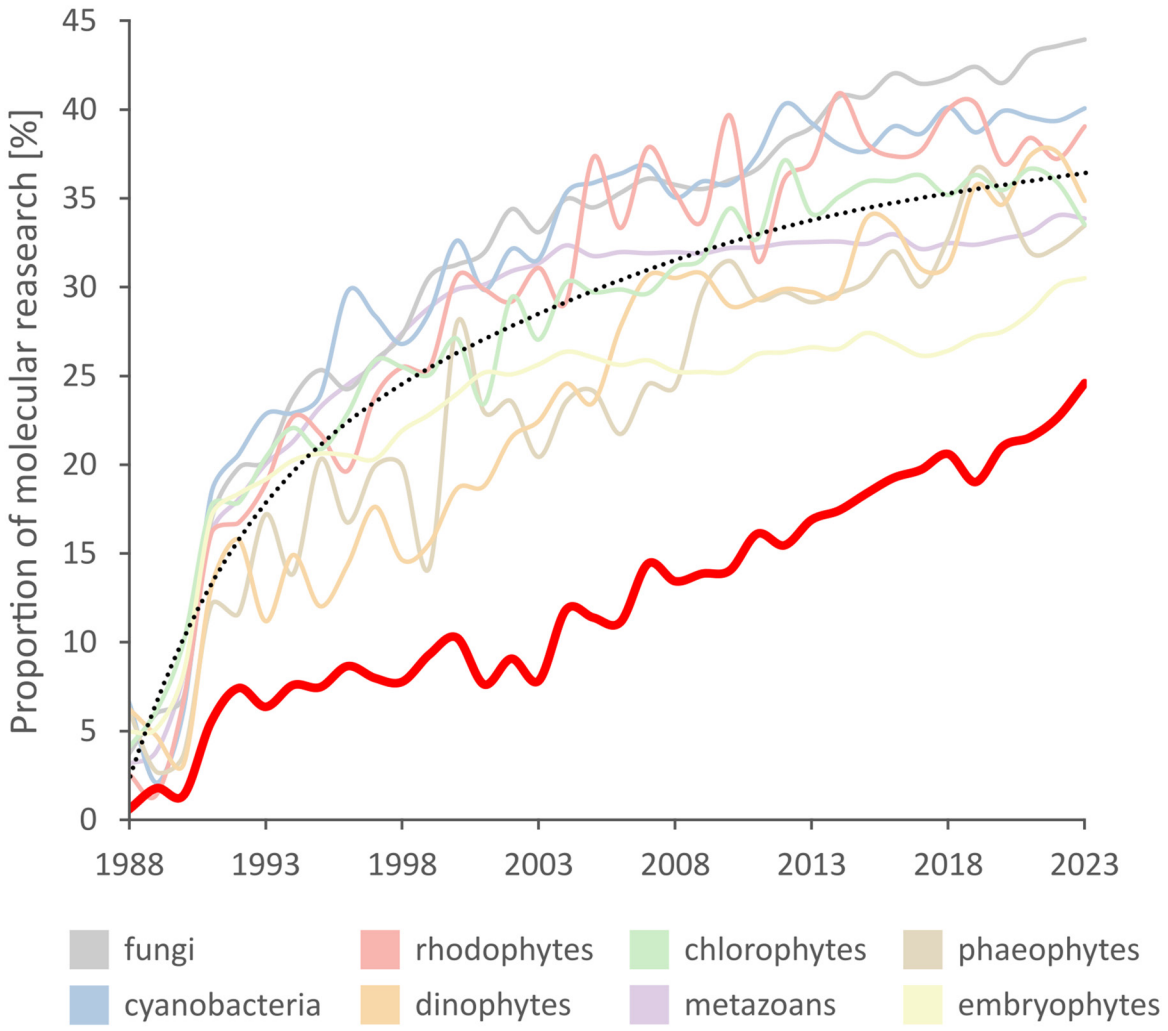

Fig. 3. Development of a proportion of molecular research over time between 1988 and 2023: a comparison of diatoms (red) and other taxa (pastel colours in the background); black dotted line = an average for the other–than–diatom taxa depicted. Only taxa with datasets larger than 10,000 documents are shown (an analogical plot including groups with datasets larger than 2,000 documents may be explored in Supplementary Material, Fig. S2). Except for diatoms, development of the annual proportion of molecular research in a vast majority of groups seemed approximately asymptotic (Fig. S3), suggesting saturation of the fields (e.g., zoology, mycology, botany) by molecular methods.

biodiversity, especially microorganismal, would be either unfeasible or indeed impossible (e.g., in the case of bacteria) without an employment of modern environmental DNA metabarcoding techniques (Malviya et al. 2016).

Finally, although the saturation level for molecular research seems to be currently around one third, it is not a physical law, and the proportion may change with future advances in our technology and knowledge. For example, an invention of a new sequencing technology (or technology in general, e.g., artificial intelligence), or a continuous increase in the affordability of established ones, may shift the saturation level. Similarly, our knowledge of some taxonomic groups may theoretically reach a point when further deployment of molecular methods will yield only diminishing results, thus potentially leading to a decrease of the saturation level for a given group.

Molecular diatomists and where to find them

Based on the corresponding author's affiliation, most of the diatom research, both molecular and non–molecular, originates from G7 countries (USA, Germany, United Kingdom, France, Canada, Japan, Italy) and China, collectively producing 56.9% of all diatom research, and 68.8% of molecular diatom research. This overall pattern is universal across science (e.g., Oliveira et al. 2022). Among countries producing at least 50 molecular diatom documents over the past 35 years, the top 10 in terms of proportion included five of the G7 countries (i.e., Italy with 24.6% of diatom research incorporating molecular methods, Germany with 22.3%, Japan with 20.4%, France with 19.7%, and USA with 17.8%), South Korea (22.6%), China (21.3%), Czech Republic (20.2%), Sweden (15.9%), and Denmark (13.9%; Fig. 4).

Bradford's law (Bradford 1934) identified 21 journals as core for molecular diatomology, publishing one third of all the diatomological documents incorporating molecular data (Fig. S4). The top five are Journal of Phycology, PLOS One, Frontiers in Microbiology, Phycologia, and Scientific Reports, collectively publishing 13% of all molecular diatom research of the past 35 years. For the subset of molecular diatomology broadly focused on diatom diversity, ecology, and evolution (the 'mol–eco–evo' dataset), Bradford's law identified 8 core journals, adding European Journal of Phycology, Protist, and Diatom Research to the five mentioned above, collectively publishing 34.6% of such research. In other words, a researcher who would follow these eight journals alone would cover one third of molecular research focused on

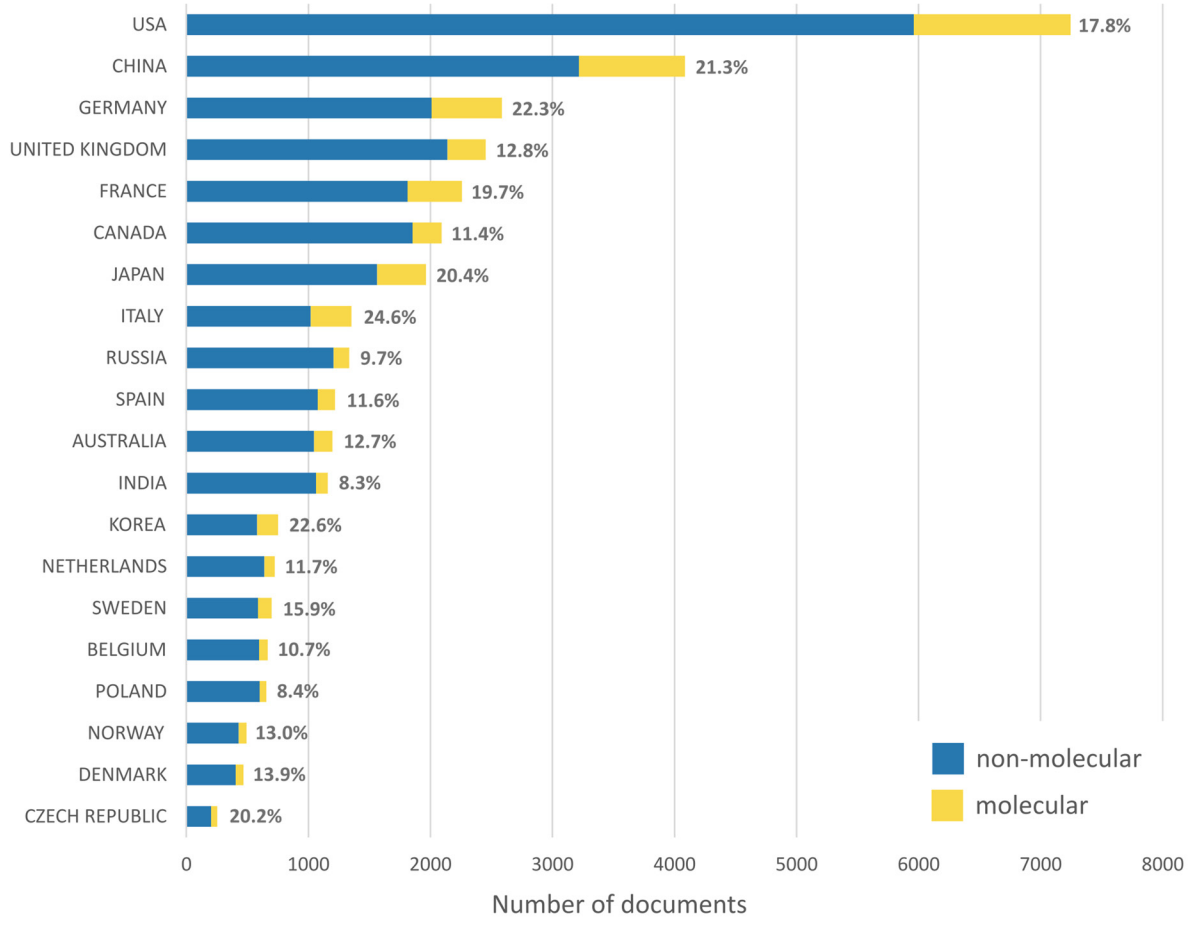

Fig. 4. Geography of the diatom research based on corresponding author's affiliation: proportions of molecular research are given as percentages; only countries producing at least 50 molecular diatom research documents over the past 35 years (1988–2023) were included.

diatom diversity, ecology, and evolution.

Conceptual structure of molecular diatomology

The following qualitative part of our study is inevitably more dependent on interpretation than the more quantitative section above, and thus conclusions should be made with caution. Therefore, we will outline only some of the broadest results below. For practical reasons, these are based on a subset of the 'molecular' dataset focused broadly on diatom diversity, ecology, and evolution (i.e., the 'mol–eco–evo' dataset, containing 20% of the 'molecular' dataset). Nevertheless, an interested reader is encouraged to download the datasets, explore the data, and make interpretations of her/his own.

Consistently across analyses (Fig. S5–7), this dataset was divided into two major clusters, representing potential subfields of molecular diatomology. Based on both specific documents (e.g., see Fig. S6) and keywords frequently found within these clusters (e.g., see Fig. S7), we interpreted and labelled them as (gen)omics and molecular taxonomy in a broad sense (i.e., including DNA barcoding, biogeography etc.). The consistent separation of these clusters suggests that the two subfields may be relatively independent of each other, with only modest overlap among researchers, and a relatively low frequency of referencing across the clusters. This seems also apparent from available conference programs of International Diatom Symposia and European Diatom Meetings (which seem to attract the molecular taxonomy community), and Molecular Life of Diatoms (assembling the genomics community).

In some of the analyses (e.g., coupling network; Fig. S5), the third cluster separated and, based on both individual documents and frequent keywords such as 'metabarcoding', 'biomonitoring', 'environmental DNA', 'rbcL' and 'high–throughput sequencing', the cluster was labelled metabarcoding. Its relatively high Mean Normalized Global Citation Score (Bornmann & Haunschild 2016) compared to the other two clusters (Fig. S5) means that it generally includes documents with above–average citation rates compared to other documents of the same age, suggesting a relatively big impact of metabarcoding (as far as citations can be taken as a proxy for an impact) and making it a 'rising star' of the field.

Diatom genomics, the second world

Since the field was consistently divided into molecular taxonomy sensu lato and genomics, we decided to explore whether the relative genetic underexploration of diatoms, the main finding of this study, is equally valid for both major subfields. Concerning the number of whole annotated genomes, thirteen for thirteen diatom species were listed in GenBank by the end of the studied timespan (https://www.ncbi.nlm.nih.gov/datasets/genome; accessed on the 1st of January 2024). When all nuclear genomes are searched (meaning also unannotated, unpublished, Metagenome Assembled Genomes or MAGs etc.), sixty–six diatom genomes are retrieved from the database (Fig. S8).

Using the same data source for the other taxonomic groups revealed that diatoms rank relatively high among algae, surpassed only by two other algal groups in number of annotated genomes (chlorophytes s.s. and dinophytes with 58 and 17, respectively; Fig. S8A), and by one in their number of all nuclear genomes (chlorophytes s.s. with 257; Fig. S8C). Thus, while diatoms are considerably genetically understudied in the subfield of molecular taxonomy, the same may not be true for the subfield of diatom genomics where, on the contrary, they might be among the best–explored algae. In addition, the development of a cumulative number of genomes over time seems exponential (especially when all nuclear genomes are considered; Fig. S8D), meaning that we may expect far more diatom genomes in the very near future (Fig. S8B and D).

This rough estimation is based on a simple metric of the number of genomes alone. However, the overall picture drawn from the subfield of genomics would likely change once the numbers of genomes (or, better, the number of species with reference genomes sequenced) are weighted by the expected species richness of the groups. Then, diatoms with their relatively high species richness estimates compared to the other taxonomic groups (Andersen 1992; Pawlowski et al. 2012) might easily fall in rank behind many of the less species–rich algal groups, although this may be prevented by ambitious projects such as 100 Diatom Genomes (https://jgi.doe.gov/csp–2021–100–diatom–genomes).

Frustules, a blessing and a curse in diatomology?

Despite the industrial applications and ecological importance of diatoms (both in serving as the base of local foodwebs and their role in global elemental cycles), it might seem shocking that this group would lag behind others in terms of molecular studies. On the contrary, one would believe that, given their importance, molecular inquiry would create a research niche that needs to be filled, and would be filled, potentially quicker than other less diverse, abundant, and functionally relevant groups. However, our work shows that this is not the case. While causation was not explicitly the point of our exercise, the question seems inevitable: Why are molecular methods utilized less frequently in the study of diatom diversity? In our opinion, several factors may be involved.

Firstly, culturing necessary for the generation of pure genetic lineages may be more challenging for benthic diatoms compared to most other algae due to their ability to adhere strongly to a substrate (typically glass or plastic labware) through the specialized structures such as a so–called raphe or mucilaginous stalks. Especially in the case of the former (i.e., the raphid diatoms, which compose a bulk of the known diatom diversity; Nakov et al. 2018), standard single–cell isolations require considerable skill to detach the targeted individual from the substrate (typically with a laboratory needle) without

breaking its relatively fragile frustule, and capturing it (typically with an automatic pipette or glass micropipette held in the other hand, sometimes even with mouth–controlling the influx) within a few seconds before it attaches again. In addition, even when the single–cell isolation is successful, many diatom species have proven unculturable under standard culturing conditions and/or in standard culturing media (Mann & Chepurnov 2004). Although this would make sense given the relatively large species diversity of diatoms compared to other algae (Andersen 1992; Pawlowski et al. 2012), with species adapted to many different environmental conditions, this has never been quantified for diatoms to the best of our knowledge (unlike for bacteria, nicknamed the Great Plate Count Anomaly; e.g., Harwani 2012), making it yet another anecdotal claim worth checking.

Secondly, it may be harder to successfully process diatom cultures in a molecular laboratory and obtain diatom DNA sequences than it is for other groups. While this should also be tested, the major difference between diatoms and other algae is the possession of a solid silica frustule encasing cells, which may complicate the release of DNA. This might be evidenced by our 'in–house' lab experiments, which suggest that the incorporation of a mechanical homogenization step (through so–called beadbeating of cultures with zirconium beads) into the DNA extraction protocol increases DNA yields by as much as an order of magnitude. Whether this is true for other algae groups as well shall be tested in the future. On the other hand, recent studies proved successful in genetic transformations in several diatom species, possibly suggesting that frustules might not be an impenetrable barrier to the movement of (at least some) DNA elements (e.g., Naser et al. 2022; Okada et al. 2023). A final consideration is that the large species diversity of diatoms (reflected also in their genetic diversity) produces challenges for selecting DNA primers needed for amplification through Polymerase Chain Reaction (PCR; Mullis et al. 1986). Nucleotide substitutions occurring in standard primer binding sites were reported even in groups of closely related diatom lineages (e.g., Kollár et al. 2019), and this is further epitomized in the diatomists' struggle to find a single DNA barcode marker for diatoms as a group, resulting in a dual barcode system that diatomology effectively utilizes today (Mann et al. 2010).

Finally, we suggest that the diatomological community is relatively reluctant overall to incorporate molecular methods for historical reasons. Prior to the molecular age, diatomists had a great advantage compared to (most of the) other algologists – diatoms have frustules. These intricately ornamented silica shells provide a wealth of taxonomically relevant characters which were, compared to the morphologies of most other algae (in contrast sometimes called soft algae), solid, relatively stable, and later explorable in great detail by scanning electron microscopes. It thus comes as no surprise that the classification system, built on such a solid foundation for almost three centuries, is being supplemented with molecular methods only reluctantly. This is in stark contrast to the morphologically more challenging algal groups (e.g., chlorophytes notorious for a huge variety of so–called 'green balls'), whose specialists might have been more eager to embrace the increased resolution provided by molecular data.

In addition, we would not underestimate an influence of human emotions (and natural attraction towards visual regularities and beauty) in shaping scientific fields. Charles Darwin famously wrote in the 4th edition of On the Origin of Species (Darwin 1866): "Few objects are more beautiful than the minute siliceous cases of the diatomaceae". Such feelings might potentially add to the reservedness of the community towards molecular methods, with the perception that adopting them would move the field away from the physical beauty that attracted many researchers to diatoms in the first place, and provided the field with such a rich history for centuries.

Conclusion

Over the past decades, the technological and conceptual developments in the fields of molecular biology and bioinformatics have unlocked a remarkable opportunity to study diatoms in a previously unimagined depth and breadth. Despite this, diatomology was found lagging in use of molecular methods behind fields occupied with the study of other taxonomic groups such as other algae, plants, fungi, and animals. Nevertheless, even in diatomology, the use of molecular methods is on the rise, which should sooner or later lead to the molecular saturation observed in other taxonomic fields. In particular, the metabarcoding wave seems to be rising, and we can thus expect to see more community–level explorations of diatom diversity, biogeography, and ecology in the near future. Similarly, we can expect far more genomes (especially MAGs, and organellar) and transcriptomes, promising deeper understanding of diatom evolution, ecology, and environmental roles. Let us conclude with the word of clarification: Despite using the term 'molecular diatomology' frequently above for convenience, we do not really think that it should be distinguished as a separate scientific field. Rather than by methods, we think we should distinguish ourselves based on the foci of our study, being it diatom diversity, ecology, bioindication, physiology or biotechnology. These, then, should be studied with whatever methods are the most appropriate.

Acknowledgements
This work was funded by the Charles University project PRIMUS/22/SCI/001. In addition, JK has been supported by Charles University Research Centre program No. UNCE/24/SCI/006, and KK by RVO 67985939 and L'Oréal-UNESCO For Women in Science Programme.

References

Alverson, J.A.; Jansen, R.K. & Theriot, E.C. (2007): Bridging the Rubicon: phylogenetic analysis reveals repeated colonizations of marine and fresh waters by thalassiosiroid diatoms. – Molecular Phylogenetics and Evolution 45: 193–210.

Andersen, R.A. (1992): Diversity of eukaryotic algae. – Biodiversity Conservation 1: 267–292.

Aria, M. & Cuccurullo, C. (2017): bibliometrix: An R–tool for comprehensive science mapping analysis. – Journal of Informetrics 11: 959–975.

Armbrust, E.V.; Berges, J.A.; Bowler, C.; Green, B.R.; Martinez, D.; Putnam, N.H.; Zhou, S.; Allen, A.E.; Apt, K.E.; Bechner, M.; Brzezinski, M.A.; Chaal, B.K.; Chiovitti, A.; Davis, A.K.; Demarest, M.S.; Detter, J.C.; Glavina, T.; Goodstein, D.; Hadi, M.Z.; Hellsten, U.; Hildebrand, M.; Jenkins, B.D.; Jurka, J.; Kapitonov, V.V.; Kröger, N.; Lau, W.W.Y.; Lane, T.W.; Larimer, F.W.; Lippmeier, J.C.; Lucas, S.; Medina, M.; Montsant, A.; Obornik, M.; Schnitzler Parker, M.; Palenik, B.; Pazour, G.J.; Richardson, P.M., Rynearson, T.A.; Saito, M.A.; Schwartz, D.C.; Thamatrakoln, K.; Valentin, K.; Vardi, A.; Wilkerson, F.P. & Rokhsar, D.S. (2004): The genome of the diatom *Thalassiosira pseudonana*: Ecology, evolution, and metabolism. – Science 306: 79–86.

Bornmann, L. & Haunschild, R. (2016): Citation score normalized by cited references (CSNCR): The introduction of a new citation impact indicator. – Journal of Informetrics 10: 875–887.

Bowler, C.; Allen, A.E.; Badger, J.H.; Grimwood, J.; Jabbari, K.; Kuo, A.; Maheswari, U.; Martens, C.; Maumus, F., Otillar, R.P.; Rayko, E.; Salamov, A.; Vandepoele, K.; Beszteri, B.; Gruber, A.; Heijde, M.; Katinka, M.; Mock, T., Valentin, K.; Verret, F.; Berges, J.A.; Brownlee, C.; Cadoret, J.–P.; Chiovitti, A; Choi, C.J.; Coesel, S.; De Martino, A.; Detter, J.C.; Durkin, C., Falciatore, A., Fournet, J.; Haruta, M., Huysman, M.J.J.; Jenkins, B.D.; Jiroutova, K.; Jorgensen, R.E., Joubert, Y.; Kaplan, A.; Kröger, N.; Kroth, P.G.; La Roche, J.; Lindquist, E.; Lommer, M.; Martin–Jézéquel, V.; Lopez, P.J.; Lucas, S.; Mangogna, M.; McGinnis, K.; Medlin, L.K.; Montsant, A.; Oudot–Le Secq, M.–P.; Napoli, C.; Obornik, M.; Schnitzler Parker, M.; Petit, J.–L.; Porcel, B.M.; Poulsen, N.; Robison, M.; Rychlewski, L.; Rynearson, T.A.; Schmutz, J.; Shapiro, H.; Siaut, M.; Stanley, M.; Sussman, M.R.; Taylor, A.R.; Vardi, A.; von Dassow, P.; Vyverman, W.; Willis, A.; Wyrwicz, L.S.; Rokhsar, D.S.; Weissenbach, J.; Armbrust, E.V.; Green, B.R.; Van de Peer, Y. & Grigoriev, I.V. (2008): The *Phaeodactylum* genome reveals the evolutionary history of diatom genomes. – Nature 456: 239–244.

Bradford, S. (1934): Sources of information on specific subjects. – Engineering 26: 85–86.

Callon, M.; Courtial, J.P. & Laville, F. (1991): Co–word analysis as a tool for describing the network of interactions between basic and technological research: The case of polymer chemistry. – Scientometrics 22: 155–205.

Chonova, T.; Rimet, F.; Bouchez, A.; Kahlert, M.; Schneider, S.C.; Bailet, B.; Eulin–Garrigue, A.; Gassiole, G.; Monnier, O.; Ouattara, A.; Rey, S.; Rhoné, M. & Keck, F. (2023): Revisiting global diversity and biogeography of freshwater diatoms: New insights from molecular data. – Environmental DNA 5: 1505–1515.

Darwin, C.R. (1866): On the origin of species by means of natural selection, or the preservation of favoured races in the struggle for life (4th ed.). – 593 pp., John Murray, London.

Drum, R.W. & Gordon, R. (2003): Star Trek replicators and diatom nanotechnology. – Trends in Biotechnology 21: 325–328.

Ebach, M.C.; Valdecasas, A.G. & Wheeler, Q.D. (2011): Impediments to taxonomy and users of taxonomy: accessibility and impact evaluation. – Cladistics 27: 550–557.

Falkowski, P.G.; Barber, R.T. & Smetacek, V. (1998): Biogeochemical controls and feedbacks on ocean primary production. – Science 281: 200–206.

Ficetola, G.; Miaud, C.; Pompanon, F. & Taberlet, P. (2008): Species detection using environmental DNA from water samples. – Biological Letters 4: 423–425.

Fortunato, S.; Bergstrom, C.T.; Börner, K.; Evans, J.A.; Helbig, D.; Milojević, S.; Petersen, A.M., Radicchi, F.; Sinatra, R.; Uzzi, B.; Vespignani, A.; Waltman, L.; Wang, D. & Barabási A.–L. (2018): Science of science. – Science 359: 1–7.

Harwani, D. (2012): The Great Plate Count Anomaly and the unculturable Bacteria. – International Journal of Scientific Research 2: 350–351.

Hebert, P.D.N.; Cywinska, A.; Ball, S.L. & deWaard, J.R. (2003): Biological identifications through DNA barcodes. – Proceedings of the Royal Society B: Biological Sciences 270: 313–321.

Hu, Q.; Sommerfeld, M.; Jarvis, E.; Ghirardi, M.; Posewitz, M.; Seibert, M. & Darzins A. (2008): Microalgal triacylglycerols as feedstocks for biofuel production: Perspectives and advances. – The Plant Journal 54: 621–639.

Kollár, J.; Pinseel, E.; Vanormelingen, P.; Poulíčková, A.; Souffreau, C.; Dvořák, P. & Vyverman, W. (2019): A Polyphasic approach to the delimitation of diatom species: a case study for the genus *Pinnularia* (Bacillariophyta). – Journal of Phycology 55: 365–379.

Kristiansen, J. (2005): Golden algae: a biology of chrysophytes. – 167 pp., A.R.G. Gantner Verlag, Königstein.

Larsen, B.; Elizabeth, C.; Matthew, K. & Wiens, J. (2017): Inordinate fondness multiplied and redistributed: the number of species on Earth and the new pie of life. – The Quarternary Review of Biology 92: 229–265.

Lewis, A.R.; Marchant, D.R.; Ashworth, A.C.; Hedenäs, L.; Hemming, S.R.; Johnson, J.V.; Leng, M.J.; Machlus, M.L.; Newton, A.E.; Raine, J.I.; Willenbring, J.K.; Williams, M. & Wolfe, A.P. (2008): Mid–Miocene cooling and the extinction of tundra in continental Antarctica. – PNAS 105: 10676–10680.

Löbl, I.; Klausnitzer, B.; Hartmann, M. & Krell, F.T. (2023): The silent extinction of species and taxonomists – an appeal to science policymakers and legislators. – Diversity 15: 1–17.

Lotka, A.J. (1926): The frequency distribution of scientific productivity. – Journal of Washington Academy of Science 16: 317–323.

Malviya, S.; Scalco, E.; Audic, S.; Vincent, F.; Veluchamy, A.; Poulain, J.; Wincker, P.; Iudicone, D.; de Vargas, C.; Bittner, L.; Zingone, A. & Bowler, C. (2016): Insights into global diatom distribution and diversity in the world's ocean. – PNAS 113: 1516–1525.

Mann, D.G. & Chepurnov, V.A. (2004): What have the Romans ever done for us? The past and future contribution of culture studies to diatom systematics. – Nova Hedwigia 79: 237–291.

Mann, D.G. & Droop, S.J.M. (1996): Biodiversity, biogeography and conservation of diatoms. – Hydrobiologia 336: 19–32.

Mann, D.G. & Evans, K.M. (2008): The species concept and cryptic diversity. – Proceedings of the 12th International Conference on Harmful Algae: 262–268.

Mann, D.G.; Sato, S.; Trobajo, R.; Vanormelingen, P. & Souffreau, C. (2010): DNA barcoding for species identification and discovery in diatoms. – Cryptogamie Algologie 31: 557–577.

Mann, D.G. & Vanormelingen, P. (2013): An inordinate fondness? The number, distributions, and origins of diatom species. – Journal of Eukaryotic Microbiology 60: 414–420.

Mardis, E. (2013): Next–generation sequencing platforms. – Annual Reviews in Analytical Chemistry 6: 287–303.

Medlin, L.; Elwood, H.J.; Stickel, S. & Sogin, M.L. (1988): The characterization of enzymatically amplified eukaryotic 16S–like rRNA–coding regions. – Gene 71: 491–499.

Mullis, K.; Faloona, F.; Scharf, S.; Saiki, R.; Horn, G. & Erlich, H. (1986): Specific enzymatic amplification of DNA *in vitro*: the polymerase chain reaction. – Cold Spring Harbor Symposia Quantitative Biology 51: 263–273.

Nakov, T.; Beaulieu, J.M. & Alverson, A.J. (2018): Accelerated diversification is related to life history and locomotion in a hyperdiverse lineage of microbial eukaryotes (Diatoms, Bacillariophyta). – New Phytologist 219: 462–473.

Naser, I.; Yabu, Y.; Maeda, Y. & Tanaka, T. (2022): Highly efficient genetic transformation methods for the marine oleaginous diatom *Fistulifera solaris*. – Marine Biotechnology 25: 657–665.

Okada, K.; Morimoto, Y.; Shiraishi, Y.; Tamura, T.; Mayama, S.; Kadono, T.; Adachi, M.; Ifuku, K. & Nemoto, M. (2023): Nuclear transformation of the marine pennate diatom *Nitzschia* sp. strain NIES–4635 by multi–pulse electroporation. – Marine Biotechnology 25: 1208–1219.

Oliveira, E.A.; Oliveira, M.C.L.; Colosimo, E.A.; Martelli, D.B.; Silva, L.R.; Silva, A.C.S.E. & Martelli–Júnior, H. (2022): Global scientific production in the pre–Covid–19 Era: An analysis of 53 countries for 22 years. – Anais da Academia Brasileira Ciencias 94: 1–11.

Pawlowski, J.; Audic, S.; Adl, S.; Bass, D.; Belbahri, L.; Berney, C.; Bowser, S.S.; Cepicka, I.; Decelle, J.; Dunthorn, M.; Fiore–Donno, A.M.; Gile, G.H.; Holzmann, M.; Jahn, R.; Jirků, M.; Keeling, P.J.; Kostka, M.; Kudryavtsev, A.; Lara, E.; Lukeš, J.; Mann, D.G.; Mitchell, E.A.; Nitsche, F.; Romeralo, M.; Saunders, G.W.; Simpson, A.G.; Smirnov, A.V.; Spouge, J.L.; Stern, R.F.; Stoeck, T.; Zimmermann, J.; Schindel, D. & de Vargas, C. (2012): CBOL protist working group: barcoding Eukaryotic richness beyond the animal, plant, and fungal kingdoms. – PLoS Biology 10: e1001419.

Pedersen, T. (2024): *ggraph*: An implementation of grammar of graphics for graphs and networks. – R package.

Pinseel, E.; Janssens, S.B.; Verleyen, E.; Vanormelingen, P.; Kohler, T.J.; Biersma, E.M.; Sabbe, K.; van de Vijver, B. & Vyverman, W. (2020): Global radiation in a rare biosphere soil diatom. – Nature Communications 11: 1–19.

R Core Team (2022): R: A language and environment for statistical computing. – R Foundation for Statistical Computing Vienna, Austria.

Round, F.E.; Crawford, R.M. & Mann, D.G. (1990): The diatoms – biology and morphology of the genera. – 747 pp., Cambridge University Press, Cambridge.

Sanger, F.; Niclen, S. & Coulson, A. (1977): DNA sequencing with chain–terminating inhibitors. – PNAS 74: 5463–5467.

Satam, H.; Joshi, K.; Mangrolia, U.; Waghoo, S.; Zaidi, G.; Rawool, S.; Thakare, R.P.; Banday, S.; Mishra, A.K.; Das, G. & Malonia, S.K. (2023): Next–Generation Sequencing technology: current trends and advancements. – Biology 12: 997.

Smol, J.P. & Stoermer, E.F. (2010): The diatoms: applications for the environmental and Earth sciences (2nd ed.). – 667 pp., Cambridge University Press, Cambridge.

Theissinger, K.; Fernandes, C.; Formenti, G.; Bista, I.; Berg, P.R.; Bleidorn, C.; Bombarely, A.; Crottini, A.; Gallo, G.R.; Godoy, J.; Jentoft, S.; Malukiewicz, J.; Mouton, A.; Oomen, R.A.; Paez, S.; Palsbøll, P.J.; Pampoulie, C.; Ruiz-López, M.J.; Secomandi, S.; Svardal, H.; Theofanopoulou, C.; de Vries, J.; Waldvogel, A.M.; Zhang, G.; Jarvis, E.D.; Bálint, M.; Ciofi, C.; Waterhouse, R.M.; Mazzoni, C.J. & Höglund, J. (2023): How genomics can help biodiversity conservation. – Trends in Genetics 39: 545–559.

Theriot, E.C.; Ashworth, M.; Ruck, E.; Nakov, T. & Jansen, R.K. (2010): A preliminary multi gene phylogeny of the diatoms (Bacillariophyta): challenges for future research. – Plant Ecology and Evolution 143: 278–296.

Wang, Y.; Zhao, Y.; Bollas, A.; Wang, Y. & Au, K.F. (2021): Nanopore sequencing technology, bioinformatics and applications. – Nature Biotechnoogy 39: 1348–1365.

Watson, J. & Crick, F. (1953): Molecular structure of nucleic acids: a structure for deoxyribose nucleic acid. – Nature 171: 737–738.

Wickham, H.; Averick, M.; Bryan, J.; Chang, W.; McGowan, L.; François, R.; Grolemund, G.; Hayes, A.; Henry, L.; Hester, J.; Kuhn, M.; Lin Pedersen, T.; Miller, E.; Milton Bache, S.; Müller, K.; Ooms, J.; Robinson, D.; Seidel, D.P.; Spinu, V.; Takahashi, K.; Vaughan, D.; Wilke, C.; Woo, K. & Yutani, H. (2019): Welcome to the Tidyverse. – Journal of Open Source Software 4: 1686.

Wickham, H.; François, R.; Henry, L.; Müller, K. & Vaughan, D. (2023): *dplyr*: a grammar of data manipulation. – R package.

Witkowski, J.; Williams, D. & Kociolek, J.P. (2015): Diatoms and the continuing relevance of morphology to studies on taxonomy, systematics and biogeography. – Nova Hedwigia. 144: 228.

Supplementary material

The following supplementary material is available for this article:

Note S1. Search queries for the datasets.

Table S1. Summary of the dataset sizes.

Fig. S1. Frequency of publication by authors (compared to Lotka's Law).

Fig. S2. Proportion of molecular research for taxa with > 2,000 docs.

Fig. S3. Saturation plots for all taxa.

Fig. S4. Core journals of molecular diatomology (based on Bradford's Law).

Fig. S5. Conceptual structure of the field through coupling network analysis.

Fig. S6. Conceptual structure of the field through co-citation network analysis.

Fig. S7. Conceptual structure of the field through co-occurrence network analysis.

Fig. S8. A quantitative comparison of algal genomics.

This material is available as part of the online article (http://fottea.czechphycology.cz/contents)

Received June 18, 2024
Revised July 30, 2024
Accepted September 24, 2024
Prepublished online January 24, 2025