LIMNOLOGY and OCEANOGRAPHY: METHODS

ASLO

Limnol. Oceanogr.: Methods 21, 2023, 220–241

doi: 10.1002/lom3.10541

phytoclass: A pigment-based chemotaxonomic method to determine the biomass of phytoplankton classes

Alexander Hayward [1,2]* **Matthew H. Pinkerton** [1] **Andres Gutierrez-Rodriguez** [1,3]

[1]National Institute of Water and Atmospheric Research, Wellington, New Zealand
[2]University Of Otago, Dunedin, New Zealand
[3]Instituto Español de Oceanografía, Centro Oceanográfico de Gijón, Gijón, Spain

Abstract

Pigment-based chemotaxonomy is a widely utilized tool to determine the biomass of phytoplankton classes from pigment biomarkers. The CHEMTAX approach is sensitive to the initial estimates of pigment-to-chlorophyll *a* (Chl *a*) ratios for the phytoplankton classes required, even though these are modified by the CHEMTAX process. We present an alternative chemotaxonomic method that utilizes simulated annealing with a steepest descent algorithm to derive class abundances and pigment-to-Chl *a* ratios. The simulated annealing algorithm is tested on two synthetic datasets of Southern Ocean phytoplankton communities. Each dataset is composed of 1000 inversion samples (set of phytoplankton class abundances, pigment ratios, and pigment profiles) with sizes ranging between 5 and 60 individual samples. We show that the new simulated annealing approach displays higher accuracy than two common configurations of the CHEMTAX method, with lower differences between true and estimated class abundances. Symmetric mean absolute percentage error were 4.8–11%, compared to 18–70% with CHEMTAX approaches. Proportions of variance explained (R^2) between true and estimated class abundances using the simulated annealing approach were 0.98–0.99 compared to 0.71–0.89 for CHEMTAX. Overall, this new methodology is capable of determining phytoplankton class abundances at higher accuracy than CHEMTAX without sensitivity to initial estimates of pigment-to-Chl *a* ratios.

Phytoplankton have high taxonomic diversity and a range of cell sizes covering several orders of magnitude (Quigg et al. 2003; Finkel et al. 2010; Estrada et al. 2016). Due to this diversity and size range, the assessment of phytoplankton community structure is not a trivial task (Kruk et al. 2011). Marine phytoplankton are intrinsically linked to global biogeochemical cycles and ecosystem dynamics (Falkowski 1994; Racault et al. 2012), yet the influence of phytoplankton as climate-active elements varies significantly between functional types, and so knowledge of the structure of phytoplankton communities is crucial to understanding the implications of a changing climate on marine ecosystems and the feedbacks between them (Henley et al. 2020; Pinkerton et al. 2021).

Traditionally, optical microscopy with morphometric identification using the Utermöhl method has been used to estimate the biomass of phytoplankton classes (Rott et al. 2007; Edler and Elbrächter 2010). Taxonomic identification using light microscopy provides high taxonomic resolution but it is typically limited to the larger size spectrum (> 5 μm), omitting classification of cryptic, smaller phytoplankton cells (Domingues et al. 2008). Preservation of water samples for optical microscopy also introduces artifacts such as cell shrinkage or enlargement and losses that skew estimates of biovolume and biomass (Menden-Deuer and Lessard 2000; Broglio et al. 2004; Zarauz and Irigoien 2008; Jakobsen and Carstensen 2011). In addition, with the requirement for extensive taxonomic training and relatively long analysis time per sample, optical microscopy cannot provide the sampling resolution required for assessing community structure at large spatial scales.

To resolve the challenge of quantifying phytoplankton at the class level (e.g., Cryptophyceae, Dinophyceae, Bacillariophyceae, Cyanophyceae), chemotaxonomic analysis of pigment data is frequently used to derive phytoplankton class abundances in measures of chlorophyll *a* (Chl *a*; Letelier et al. 1993; Mackey

*Correspondence: alexhayward1995@gmail.com

Author Contribution Statement: Project oversight: AH, MP, AGR. Experimental design: AH, MP. Testing and results: AH, MP. Writing: AH, MP, AGR.

Additional Supporting Information may be found in the online version of this article.

et al. 1996; Uitz et al. 2006; Latasa 2007; Van den Meersche et al. 2008; Higgins et al. 2011; Kramer and Siegel 2019). Phytoplankton contain many pigments that either utilize light energy for photosynthesis such as Chl *a*, and accessory or carotenoid pigments that harvest light (e.g., fucoxanthin), or dissipate light energy via photoprotective mechanisms (e.g., diadinoxanthin; Bidigare and Ondrusek 1996; Brunet et al. 2011). In addition, the pigment composition in a given class of phytoplankton varies with both light and nutrient status, which adds further complexity to phytoplankton classification, as pigment concentrations and their respective pigment-to-Chl *a* ratios (hereby referred to as "pigment ratios") are dynamic and change with environmental conditions (Schlüter et al. 2000; Henriksen et al. 2002).

With the advent of advanced high-performance liquid chromatography (HPLC) methods (Jeffrey 1997; Wright and Jeffrey 2006), novel phytoplankton pigments have been discovered, adding further power to chemotaxonomic identification (Higgins et al. 2011). Higher resolution HPLC measurements have also revealed that minor concentrations of many pigments are present in major algal classes, leaving few previously classified markers truly unambiguous (Roy et al. 2011). This adds further difficulty for chemotaxonomic approaches to delineate phytoplankton classes with shared diagnostic pigments.

The most common approach to chemotaxonomic pigment analysis has been the CHEMTAX program (see Mackey et al. 1996). CHEMTAX performs nonnegative matrix factorization to solve for phytoplankton class abundances based on HPLC pigment measurements. Groups of samples are assumed to share the same relative abundances of pigments in each phytoplankton class. CHEMTAX uses a steepest descent algorithm to modify each initial estimate of pigment ratios until the error between estimated and measured pigment concentrations is minimized, or an iteration limit is reached. Total Chl *a* biomass is then partitioned across the different phytoplankton classes. The CHEMTAX approach is sensitive to the starting pigment ratios, and this sensitivity to the initial guess can bias results (Latasa 2007; Pan et al. 2011; Swan et al. 2016). When tested on "synthetic" datasets (i.e., data generated from statistical relationships to simulate field measurements), pigment ratios frequently do not converge to their true values, resulting in erroneous biomass of phytoplankton classes (Latasa 2007). Furthermore, CHEMTAX v1.95 is run on a graphical user interface which means that it is problematic to apply at scale (to large datasets) and there have been no studies of its behavior on large synthetic datasets (Swan et al. 2016).

As an alternative to CHEMTAX, a Bayesian Compositional Estimator (BCE) has been developed that uses a Markov-Chain Monte-Carlo (MCMC) algorithm to determine probability intervals and point estimates for the biomass of phytoplankton classes (Van den Meersche et al. 2008). In their study, Van den Meersche et al. showed a good agreement between phytoplankton class abundances calculated by BCE and CHEMTAX. However, the BCE is computationally demanding and requires either prior knowledge of probability distributions for phytoplankton pigment ratios, or extensive prior predictive simulation for the Markov chain to converge. Due to the complex nature and time constraints of MCMC algorithms, the implementation of the BCE is typically limited to small sample sizes (Higgins et al. 2011). Limited studies have used the BCE for phytoplankton pigments alone, and the method has instead been used more in conjunction with fatty acid biomarkers (De Carvalho and Caramujo 2014; Strandberg et al. 2015).

To both increase the accuracy of estimating phytoplankton class abundances and reduce the requirement for a priori knowledge of pigment ratios, in this paper we explore alternative methodologies to CHEMTAX for partitioning Chl *a* between phytoplankton classes based on measurements of pigment concentrations. We use simulated annealing, a tool used to find the global minima of a function when the function has many local minima (Press et al. 2007) and couple this to a novel steepest descent algorithm.

We test the simulated annealing method alongside two "standard" CHEMTAX configurations on large synthetic datasets which were generated to span the scope of potential pigment measurements in the Southern Ocean. Drivers of phytoplankton community structure in the Southern Ocean include salinity, temperature, photosynthetically active irradiance, mixing, nutrients and trace elements like iron (Bathmann et al. 1997; Carter et al. 2008). Top-down controls such as mortality due to grazers has been shown to have a large control on chlorophyll biomass and the structure of phytoplankton communities in the Southern Ocean (Arteaga et al. 2020). This diversity leads to a great breadth of phytoplankton community structures (Deppeler and Davidson 2017) and pigment compositions and makes the Southern Ocean an ideal candidate for testing chemotaxonomic techniques.

The new methodology described in this paper will be available as a package *phytoclass* in the programming language R (R Core Team 2022) for wider scrutiny and use.

Materials and procedures

Definitions and notation

In this paper, when we refer to a single sample this consists of three parts: class abundances (**c**), pigment ratios ($\mathbf{F}_{\text{True}}$), and pigment profiles (**s**). The matrix $\mathbf{F}_{\text{True}}$ specifies the pigment-to-Chl *a* ratios for each phytoplankton class; it is comprised of m phytoplankton classes, n pigments, and has dimensions of m (rows) by n (columns), written ($m \times n$). We use the subscript "True" here to distinguish $\mathbf{F}_{\text{True}}$ from an estimate by an inversion method (*see* below). The vector **c** is the abundance of each phytoplankton class in units of Chl *a* biomass. The vector **s** is the concentration of each pigment in a sample and attained through matrix multiplication between and **c** and $\mathbf{F}_{\text{True}}$ (Eq. 1).

$$\mathbf{s} = \mathbf{c}\mathbf{F}_{\text{True}} \tag{1}$$

We then consider a set of p samples which share a common $\mathbf{F}_{\text{True}}$ and are grouped together for inversion. Using matrix

notation, $\mathbf{C}_{\text{True}}$ ($p \times m$) is the matrix of class abundances consisting of p sets of $\mathbf{c}$. Similarly, $\mathbf{S}_{\text{True}}$ ($p \times n$) is the matrix of pigment samples (p sets of $\mathbf{s}$ and n pigments). For example, $\mathbf{c}$ and $\mathbf{C}_{\text{True}}$ are constructed as:

$$\mathbf{c} = c_{i,1}, c_{i,2}, \ldots, c_{i,m}\ \mathbf{C}_{\text{True}} = \begin{bmatrix} c_{1,1} & c_{1,2} & \cdots & c_{1,m} \\ c_{2,1} & c_{2,2} & \cdots & c_{2,m} \\ \vdots & \vdots & \ddots & \vdots \\ c_{p,1} & c_{p,2} & \cdots & c_{p,m} \end{bmatrix} \quad (2)$$

Through matrix multiplication between $\mathbf{C}_{\text{True}}$ and $\mathbf{F}_{\text{True}}$, we attain $\mathbf{S}_{\text{True}}$ (3). We call this set of $\mathbf{C}_{\text{True}}$, $\mathbf{F}_{\text{True}}$, and $\mathbf{S}_{\text{True}}$ an "inversion sample."

$$\mathbf{S}_{\text{True}} = \mathbf{C}_{\text{True}}\mathbf{F}_{\text{True}} \quad (3)$$

$\mathbf{S}_{\text{True}}$ represents the data that would be produced via HPLC analysis of a set of water samples collected in the field. Note that each inversion sample (each row of $\mathbf{S}_{\text{True}}$ and $\mathbf{C}_{\text{True}}$) shares the same $\mathbf{F}_{\text{True}}$. Inversion methods such as CHEMTAX are then used to estimate $\mathbf{C}_{\text{True}}$ and $\mathbf{F}_{\text{True}}$ by inverting $\mathbf{S}_{\text{True}}$. From the inversion of $\mathbf{S}_{\text{True}}$ we attain $\mathbf{F}_{\text{Est}}$ and $\mathbf{C}_{\text{Est}}$. Likewise, matrix multiplication between $\mathbf{F}_{\text{Est}}$ and $\mathbf{C}_{\text{Est}}$ produces $\mathbf{S}_{\text{Est}}$ (Eq. 3).

Existing CHEMTAX methods

Given only $\mathbf{S}_{\text{True}}$, the inversion problem (Eq. 3) for $\mathbf{C}_{\text{Est}}$ and $\mathbf{F}_{\text{Est}}$ is highly underdetermined (fewer constraints than unknowns in $\mathbf{C}_{\text{Est}}$ and $\mathbf{F}_{\text{Est}}$). On the other hand, given $\mathbf{S}_{\text{True}}$ and $\mathbf{F}_{\text{True}}$ (or an initial estimate of $\mathbf{F}_{\text{True}}$), the inversion problem for $\mathbf{C}_{\text{Est}}$ is overdetermined (i.e., the number of constraints is greater than the number of free variables) so that an exact solution is not (usually) possible and the "closest" solution of $\mathbf{C}_{\text{Est}}$ is found (see Eq. 4 for definition of "closest").

The least-squares nonnegative matrix factorization approach of Lawson and Henson (Lawson and Hanson 1995; Eq. 4) is used by the original CHEMTAX method to solve the overdetermined least square problems with inequality and equality constraints.

$$\text{minimize}\|\mathbf{S}_{\text{True}} - \mathbf{C}_{\text{Est}}\mathbf{F}_{\text{Est}}\| \text{ subject to } [\mathbf{C}_{\text{Est}}] \geq 0 \text{ and } \sum[\mathbf{C}_{\text{Est}}] = 1 \quad (4)$$

where $||\ ||$ is the Frobenius norm of the matrix operation. The inequality constraint ensures that all rows and columns are nonnegative values while the equality constraint ensures that the sum of every matrix row is equal to 1. The root mean square error (RMSE) between $\mathbf{S}_{\text{True}}$ and $\mathbf{S}_{\text{Est}}$ ($= \mathbf{C}_{\text{Est}}\mathbf{F}_{\text{Est}}$) is calculated after every manipulation of $\mathbf{F}_{\text{Est}}$ using Eq. 5.

$$\varepsilon = \|\mathbf{S}_{\text{True}} - \mathbf{S}_{\text{Est}}\| \quad (5)$$

The CHEMTAX approach then iteratively varies $\mathbf{F}_{\text{Est}}$ to improve the fit between the $\mathbf{S}_{\text{Est}}$ and $\mathbf{S}_{\text{True}}$ (Eq. 5), and this process is repeated until a stable solution is reached (one where no further changes to $\mathbf{F}_{\text{Est}}$ and $\mathbf{C}_{\text{Est}}$ lead to improvements) or until an iteration limit is reached.

The original CHEMTAX program uses a steepest descent to reduce the RMSE between $\mathbf{S}_{\text{True}}$ and $\mathbf{S}_{\text{Est}}$ (Eq. 5). For each iteration (k) of the steepest descent algorithm, every nonzero element of $\mathbf{F}_{\text{Est}}$ is varied by a specified amount called the "step-ratio" (e.g., factor of 1/5). The error is then recalculated and the element causing the largest improvement is selected, with this repeated for all nonzero values in $\mathbf{F}_{\text{Est}}$. The element in $\mathbf{F}_{\text{Est}}$ that causes the largest decrease in error is retained, and the process is repeated from the resulting $\mathbf{F}_{\text{Est}}$. This creates a series of $\mathbf{F}_{\text{Est}}$ ($\mathbf{F}_{\text{Est}}(k) \rightarrow \mathbf{F}_{\text{Est}}(k+1)$) with corresponding $\mathbf{C}_{\text{Est}}$ ($\mathbf{C}_{\text{Est}}(k) \rightarrow \mathbf{C}_{\text{Est}}(k+1)$). The steepest descent algorithm reduces the size of the step after a number of iterations by increasing the step ratio. This has the effect of decreasing how much the element is varied as the iteration count increases, meaning the step between $\mathbf{F}_{\text{Est}}(k)$ and $\mathbf{F}_{\text{Est}}(k+1)$ becomes progressively smaller.

Settings within CHEMTAX allow for the program to stop after a predefined number of iterations, or if subsequent iterations do not cause a reduction in error.

Prior to analysis, both the $\mathbf{F}_{\text{Est}}$ and $\mathbf{S}_{\text{True}}$ matrices are normalized to unit row sum and $\mathbf{S}_{\text{True}}$ is weighted as the reciprocal of its column means, bound at a maximum weight of 30. This has the benefit of increasing the speed of inversion as all values are brought to a similar scale, while also promoting the accuracy in the derivation of pigments that have lower concentrations. A bounded weight at 30 ensures that pigments with a minority concentration are not overweighted which would result in the minority pigment concentrations being prioritized over pigment concentrations that are naturally higher.

CHEMTAX variants

There are two common and slightly different implementations of CHEMTAX described by Latasa (2007) and the CHEMTAX v1.95 release, 2017 (from here on referred to as CHEMTAX-1 and CHEMTAX-2, respectively). CHEMTAX-1 focuses on minimizing the sensitivity of CHEMTAX to the starting $\mathbf{F}_{\text{Est}}$ by increasing the iteration limit to 5000, with a step size set to 25 and a step ratio set at 2. Nine random (but reasonable) $\mathbf{F}_{\text{Est}}$ values are selected both exceeding and within values reported in the literature. The nine $\mathbf{F}_{\text{Est}}$ output from each run of CHEMTAX are then used as the input for each subsequent run. This process is repeated 10 times, with the aim of $\mathbf{F}_{\text{Est}}$ converging to its true values. After the 10^{th} iteration, the final $\mathbf{F}_{\text{Est}}$ are averaged and $\mathbf{S}_{\text{True}}$ is inverted to obtain the final $\mathbf{C}_{\text{Est}}$.

To account for the sensitivity of initial starting values, the CHEMTAX-2 approach selects an initial $\mathbf{F}_{\text{Est}}$ from pigment ratios in the literature. The literature-based $\mathbf{F}_{\text{Est}}$ is then randomized by a specified factor (0.7 is used in this study, as recommended by Wright, CHEMTAX 2017 release), to create

60 separate $\mathbf{F}_{Est}$. The CHEMTAX algorithm is run for 200 iterations on each $\mathbf{F}_{Est}$, using an initial step of 10, and a step ratio of 1.3. From the 60 $\mathbf{F}_{Est}$, the six with the lowest error are then averaged to produce the final $\mathbf{F}_{Est}$ and $\mathbf{C}_{Est}$.

CHEMTAX in R

We wanted to explore the performance of CHEMTAX on a large dataset (many inversion samples) but the latest CHEMTAX software (v1.95) cannot be applied in a "batch method" (processed automatically) to a large number of datasets (many datasets with varying $\mathbf{F}_{True}$). We note that CHEMTAX v.1.0 was built in MATLAB but is not publicly available. Hence, the steepest descent algorithm was built in the programming language "R" (henceforth called "R-CHEMTAX"), and the CHEMTAX-1 and CHEMTAX-2 methods were reproduced using R code (R Core Team 2022). We refer to these as R-CHEMTAX-1 and R-CHEMTAX-2, respectively. These R versions of CHEMTAX can be applied efficiently to large datasets to explore their performances, for example, on a synthetic set of test data. To improve the speed of the matrix factorization, we implemented the co-ordinate descent method of Franc et al. (2005) as an approximation to that used in the standard methods. Convergence times for large or small matrices of the R methods were low because they are generally sparse, and the solution is analytic.

As there is a randomized element to the CHEMTAX method, identical datasets were inverted using the original CHEMTAX method 20 times. Means and standard deviations were calculated from the output of 20 identical CHEMTAX and R-CHEMTAX inversions. We validated R-CHEMTAX against the CHEMTAX v1.95 release (Wright 2017) using regression analysis on the mean $\mathbf{F}_{Est}$ and $\mathbf{C}_{Est}$ for each approach. We computed the proportion of deviance explained (squared value of Pearson's correlation coefficient, R^2) to test the fit between each approach, and the *F*-statistic was calculated to determine if the R-CHEMTAX method fits the CHEMTAX method better than a model with no independent variables.

New optimization techniques

Two new optimization techniques for solving Eq. 4 as an alternative to CHEMTAX were developed and tested. An overview of the methodology is given in Fig. 1.

Method 1: Alternating least squares

We used alternating least squares (ALS) to find the minima of Eq. 4. Alternating least squares uses the output of the primary matrix factorization as an input for sequential factorization (Comon et al. 2009). We first solved for $\mathbf{C}_{Est}$ using Eq. 4; then using $\mathbf{C}_{Est}$ and $\mathbf{S}_{True}$, we solved for $\mathbf{F}_{Est}(k+1)$ using Eq. 6.

$$\mathbf{F}_{Est}(k+1) = \min \left\| \left(\mathbf{C}_{Est}^{\mathbf{T}}\mathbf{C}_{Est}\right)^{-1}\mathbf{C}_{Est}^{\mathbf{T}}\mathbf{S}_{True} \right\| \quad (6)$$

With $\mathbf{F}_{Est}(k+1)$, we then solved for $\mathbf{C}_{Est}(k+1)$ using Eq. 4. This alternating process continued for a set number of iterations or until further iterations did not reduce the error.

Method 2: Steepest descent algorithm

A novel steepest descent algorithm (SDA) was built to estimate $\mathbf{F}_{Est}$ and $\mathbf{C}_{Est}$. All nonzero elements of $\mathbf{F}_{Est}$ were randomized from a uniform distribution by a factor 5% above or below their initial values; the randomization of the SDA is limited to ± 5% as the tool is most effective at searching for a local minima, close to its initial starting value. To decrease convergence times, with each iteration of the SDA every element that reduced error replaced the previous value in $\mathbf{F}_{Est}$. This contrasts with the original CHEMTAX algorithm where only one element in $\mathbf{F}_{Est}$ is replaced after each iteration. If no elements reduced the error, the factor was reduced to 3%, and then to 1%.

To determine the effectiveness of the ALS and SDA for finding a global minimum, the methods were tested on an inversion sample of a synthetic dataset-1 (see "Synthetic Dataset-1" section). Each method was applied for 3000 iterations and implemented on random $\mathbf{F}_{Est}$, and an alternative $\mathbf{F}_{Est}$ where pigment ratios are randomized either side of their true values (within a ± 100% range). This process was repeated for six separate runs with different starting values for $\mathbf{F}_{Est}$.

Simulated annealing

Both the ALS and SDA were implemented within a simulated annealing framework. Simulated annealing is commonly used to find the global minimum of a function when the function has many local minima (Press et al. 2007). Given the sensitivity of CHEMTAX to the starting value of $\mathbf{F}_{Est}$, our hypothesis is that the solution space (Eq. 4) has a high number of local minima which makes the solution sensitive to the starting values of $\mathbf{F}_{Est}$. The new method ranges across these local minima to find the global minimum. This is achieved by including a random jump from the current best estimate to a new (random) starting position and then using ALS or SDA to find the local minimum close to this new starting point. In the simulated annealing approach, the size of the random jump decreases over time following the analogy of a controlled and gradual cooling of a metal into its lowest energy state. To ensure that the simulated annealing algorithm does not get stuck in a local minimum, at each iteration the algorithm can accept a worse solution using a probability-based acceptance criterion, allowing for the algorithm to "hill-climb" from a local minima.

Prior to analysis, a preliminary check of the matrices condition number is carried out to determine if the matrix is well conditioned (invertible), given the phytoplankton groups and pigments selected for analysis. $\mathbf{F}_{Est}$, is perturbed 100, 000 times, and multiplied by the transpose of $\mathbf{S}_{True}$, the condition number (κ) is then calculated as $\kappa = \left\|\mathbf{F}\mathbf{S}^{\mathbf{T}}\right\| \left\|\mathbf{F}\mathbf{S}^{\mathbf{T}^{-1}}\right\|$ (Eq. 7). A large condition number indicates that the matrix is singular, meaning that one or more of the columns are close to being linear combinations to the rest of the columns. When the condition number is exceptionally large (say $>10^{10}$), inversion

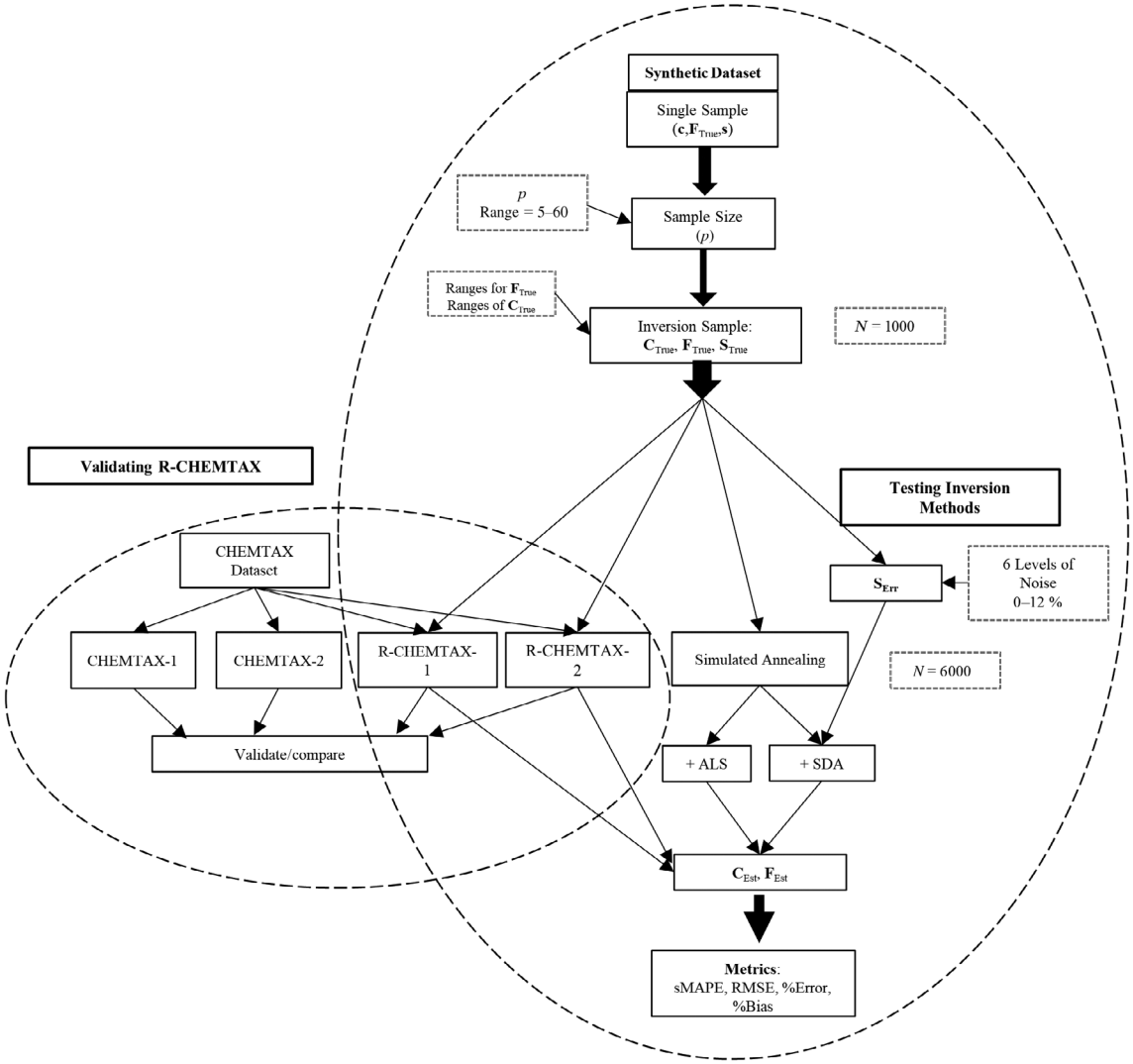

Fig. 1. Flow diagram depicting the various steps taken within the methodology of the present study, from the creation of synthetic datasets to the inversion and analysis of phytoplankton class abundances with various methodologies. **F** signifies the phytoplankton pigment to Chl *a* ratio; **S/s** signify the sample matrices/ vectors; and **C/c** signify the phytoplankton class abundances; and $\mathbf{S_{Err}}$ represents phytoplankton pigment samples with added noise.

of the matrix will be highly inaccurate and the method will not proceed. In this case, the problem must be restated with fewer groups.

To implement the simulated annealing algorithm, we imposed a bounded uniform randomization to pigment ratios in $\mathbf{F}_{\text{Est}}$, with limits set from 15% above and below values reported in the literature.

First, $\mathbf{F}_{\text{Est}}$ is randomized to give a candidate matrix, $\mathbf{F_C}$, and its associated error (Eq. 5) calculated ($\boldsymbol{\varepsilon}_\mathbf{C}$). A new jump is then considered by randomizing $\mathbf{F_C}$ to produce a new candidate $\mathbf{F}_{\text{Est}}$ ($\mathbf{F}'_\mathbf{C}$) and its associated error ($\boldsymbol{\varepsilon}'_\mathbf{C}$). If the error of $\boldsymbol{\varepsilon}'_\mathbf{C}$ is lower than $\boldsymbol{\varepsilon}_\mathbf{C}$, then the $\mathbf{F_C}$ matrix is updated with the values of $\mathbf{F}'_\mathbf{C}$. This process is repeated with the size of the jumps decreasing with each iteration, so that $\mathbf{F}'_\mathbf{C}$ stays closer and closer to $\mathbf{F_C}$ value (smaller and smaller random jumps). After a specified number of iterations, the $\mathbf{F_C}$ associated with the lowest error is returned, and the final $\mathbf{C}_{\text{Est}}$ is produced.

As in the standard CHEMTAX configurations, prior to analysis, both the $\mathbf{F}_{\text{Est}}$ and $\mathbf{S}_{\text{True}}$ matrices were normalized to unit row sum and $\mathbf{S}_{\text{True}}$ was weighted as the reciprocal of its column means, bound at a maximum weight of 30. To evaluate the performance of the simulated annealing method, we used an iteration length of 500 and a step of 0.009.

Synthetic dataset-1

To test the different chemotaxonomic approaches, we built synthetic dataset-1, of Southern Ocean phytoplankton class abundances and their associated pigment concentrations. Class

abundances consist of m (= 7) different classes of phytoplankton, namely: *Synechococcus*, prasinophytes, haptophyte-T3, haptophyte-T4, dinoflagellates, diatoms, and cryptophytes. Pigment profiles n (= 9) consist of: zeaxanthin (Zea), prasinoxanthin (Pras), 19′-butanoyloxyfucoxanthin (But), peridinin (Per), 19′-hexanoyloxyfucoxanthin (Hex), fucoxanthin (Fuco), alloxanthin (Allox), Chl *c*3, Chl *b*, and Chl *a*. We generated values for $\mathbf{C}_{\text{True}}$ based on phytoplankton classes biomass from several literature sources (Karl et al. 1991; Wright et al. 1996; Peeken 1997; Wright and van den Enden 2000; Ishikawa et al. 2002; Hashihama et al. 2008), and randomized from a log-normal distribution. Similarly, pigment ratios for $\mathbf{F}_{\text{True}}$ were selected by sampling from a log-normal distribution, bounded by limits reported in the literature (Wright unpubl.—data release, 2017). The initial $\mathbf{F}_{\text{True}}$ used to build the synthetic dataset was based on the Southern Ocean phytoplankton pigment ratios from Mackey et al. (Mackey et al. 1996; Table 1). Variations of these pigment ratios have frequently been used with the CHEMTAX program to derive class abundances in the Southern Ocean and test the effectiveness of different CHEMTAX configurations (Mackey et al. 1996; Wright et al. 1996; Latasa 2007).

We generated 1000 synthetic inversion samples, with varying $\mathbf{F}_{\text{True}}$, p, $\mathbf{C}_{\text{True}}$, and $\mathbf{S}_{\text{True}}$. We sampled from a literature-bound log-normal distribution to generate a set of $\mathbf{F}_{\text{True}}$ (Wright unpubl.—data release, 2017). For each inversion sample, p was randomly selected as being between 5 and 60 with a uniform distribution. To create a set of synthetic $\mathbf{C}_{\text{True}}$ matrices with varying p, number of samples which share a common $\mathbf{F}_{\text{True}}$, we sampled class abundances from log-normal distributions based on literature values. A set of synthetic $\mathbf{S}_{\text{True}}$ were then computed by matrix multiplication between the $\mathbf{F}_{\text{True}}$ and $\mathbf{C}_{\text{True}}$ (Eq. 3). This synthetic dataset with 1000 inversion samples provides good coverage of the parameter space but is not too large for testing.

Sensitivity Analysis

To test robustness to errors in $\mathbf{S}_{\text{True}}$ (e.g., from measurement error), we conducted a sensitivity analysis by adding noise to every inversion sample in our dataset. This was done by randomizing each pigment concentration from a uniform distribution with replacement, at specified levels above and below the pigments' true concentration (Eq. 7).

$$\mathbf{s}_{\text{Err}}(i,j) = \mathbf{s}_{\text{True}}(i,j)U(-Z,+Z) \tag{7}$$

Here, $\mathbf{s}(i,j)$ is an element of $\mathbf{S}$ where i is the matrix row, and j is the matrix column. U(min,max) signifies a uniform distribution between min and max, and Z is the level of noise added. This method ensured that noise was dispersed randomly and evenly through each inversion sample. We created six levels of noise, between 0% and 12%, with each level representing a 2% increment. The 6000 inversion samples created with added noise are termed $\mathbf{S}_{\text{Err}}$.

We applied the simulated annealing method to the synthetic data ($\mathbf{S}_{\text{Err}}$) to estimate $\mathbf{F}_{\text{True}}$ and $\mathbf{C}_{\text{True}}$ with different levels of noise. The performance of inversion methods was then assessed by comparing the error between true and estimated matrices ($\mathbf{C}_{\text{True}}$ vs. $\mathbf{C}_{\text{Est}}$).

To further test the sensitivity of the method we created an additional test to determine if the program can accurately estimate a group with zero biomass. We systematically set the Chl *a* of *Synechococcus* to zero for every inversion sample. *Synechococcus* were chosen as they share the pigment zeaxanthin with prasinophytes and are often, but not always, present in the Southern Ocean.

To assess the sensitivity to phytoplankton groups with shared pigment makers, we added a third haptophyte group to our synthetic data (haptophyte-T8). In our dataset, the haptophyte-T8 group shares the same pigments as the haptophyte-T4 group, alongside similar pigments to haptophyte-T3 and diatom-1. To understand how the inversion process is affected by an additional haptophyte group, we compare the condition number of the matrices, both with and without the addition of haptophyte-T8.

Synthetic dataset-2 (increased complexity)

To further test the robustness of the chemotaxonomic approaches, we built a second set of inversion samples "synthetic dataset-2" from pigment ratios published in Wright et al. (Wright et al. 2010; Table 2), which gave higher taxonomic resolution (i.e., distinguishing between different

Table 1. Pigment-to-Chl *a* ratios used to formulate synthetic dataset-1 (Mackey et al. 1996; Latasa 2007).

Literatre pigment ratios (synthetic dataset-1)										
Groups	**Chl *c3***	**Per**	**But**	**Fuco**	**Hex**	**Pras**	**Allox**	**Zea**	**Chl *b***	**Chl *a***
Prasinophytes	0	0	0	0	0	0.315	0	0.01	0.945	1
Dinoflagellates	0	1.062	0	0	0	0	0	0	0	1
Cryptophytes	0	0	0	0	0	0	0.228	0	0	1
Haptophyte-T3	0.046	0	0	0	1.703	0	0	0	0	1
Haptophyte-T4	0.047	0	0.246	0.585	0.538	0	0	0	0	1
Synechococcus	0	0	0	0	0	0	0	0.348	0	1
Diatoms	0	0	0	0.754	0	0	0	0	0	1

Table 2. Pigment-to-Chl *a* ratios used to formulate synthetic dataset-2 (Wright et al. 2010).

Literatrure pigment ratios (synthetic dataset-2)												
	Chl *c3*	Chl *c1*	Per	Fuco	Neo	Pras	Violax	Hex	Allox	Lut	Chl *b*	Chl *a*
Prasinophytes	0	0	0	0	0.07	0.09	0.049	0	0	0.0066	0.55	1
Chlorophytes	0	0	0	0	0.071	0	0.032	0	0	0.23	0.15	1
Cryptophytes	0	0	0	0	0	0	0	0	0.21	0	0	1
Diatoms-1	0	0.21	0	1.04	0	0	0	0	0	0	0	1
Diatoms-2	0.016	0	0	0.83	0	0	0	0	0	0	0	1
Dinoflagellates-1	0	0	0.82	0	0	0	0	0	0	0	0	1
Haptophyte-H	0.34	0	0	0.13	0	0	0	0.43	0	0	0	1
Haptophyte-L	0.13	0	0	0.01	0	0	0	1.21	0	0	0	1

subgroups of diatoms) including two classes for diatoms, haptophytes, and green algae. These inversion samples reflect high-latitude Antarctic waters, omitting phytoplankton classes such as *Synechococcus* that are more commonly associated with lower latitudes (Zwirglmaier et al. 2008). An increased number of shared pigments is present between algal classes, which is likely to make the inversion to phytoplankton classes more ambiguous and we wanted to test this hypothesis.

To create 1000 new inversion samples, pigment ratios were sampled from a log normal distribution. Bounds were set on each pigment ratio from values reported in the literature, including experimental results from the response of *Phaeocystis antarctica* under high and low iron (Fe) concentration (DiTullio et al. 2007). To generate $\mathbf{C}_{\text{True}}$, standard deviations and means were obtained from Wright et al. (2010) and sampled from a log-normal distribution, with $\mathbf{S}_{\text{True}}$ then computed using Eq. 3.

R-CHEMTAX-1, R-CHEMTAX-2, and the simulated annealing algorithm were tested by solving these complex inversion samples.

Statistical analysis

To compare the effectiveness of each method, the bias, percent bias, RMSE, R^2, and symmetrical mean absolute percentage error (sMAPE; Armstrong 1985) were calculated between all true and estimated matrices, for each chemotaxonomic approach (Eqs. 8–10). Bias for each dataset was calculated by taking the average of the difference between the estimated and true values. The sMAPE was used in lieu of the MAPE; as the MAPE places a higher penalty on negative errors.

$$\text{Bias} = \frac{1}{N}\sum(\text{True} - \text{Est}) \tag{8}$$

$$\text{RMSE} = \sqrt{\frac{1}{N}\sum(\text{True} - \text{Est})^2} \tag{9}$$

$$\text{sMAPE} = \frac{1}{N}\sum\frac{|\text{True} - \text{Est}|}{(|\text{True}| + |\text{Est}|)/2} \tag{10}$$

where "True" represents $\mathbf{C}_{\text{True}}$ and $\mathbf{F}_{\text{True}}$, "Est" represents $\mathbf{C}_{\text{Est}}$ and $\mathbf{F}_{\text{Est}}$, and the sums are calculated over all N (values).

Assessment

Comparison between R-CHEMTAX and CHEMTAX

Due to the randomization of $\mathbf{F}_{\text{Est}}$, neither R-CHEMTAX, nor CHEMTAX consistently produced the same means and standard deviations, although they were similar (Supporting Information Tables S3, S4). The R^2 between the mean $\mathbf{F}_{\text{Est}}$ from each approach was 0.999 for CHEMTAX-1, and 0.997 for CHEMTAX-2 with RMSE of 0.007 and 0.0204, respectively.

R-CHEMTAX and CHEMTAX produced very similar $\mathbf{C}_{\text{Est}}$ when using both approaches. The R^2 was 0.999 for R-CHEMTAX-1, with a RMSE of 0.44 mg Chl *a* m^{-3}, and an *F*-statistic of 8.2×10^5. For R-CHEMTAX-2, the R^2 value is 0.991 with a RMSE of 2.4 mg Chl *a* m^{-3}, and an *F*-statistic of 3.24×10^4 (Fig. 2). The R^2 was high for each

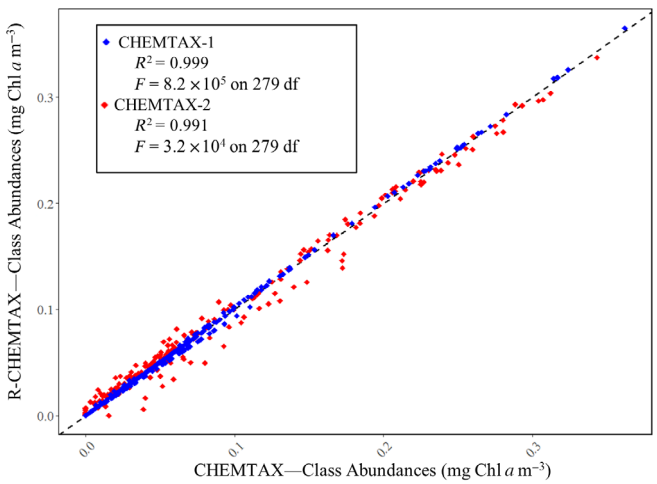

Fig. 2. Regression analysis between class abundances derived from CHEMTAX (1 and 2) and class abundances derived from R-CHEMTAX (1 and 2) for synthetic dataset-1.

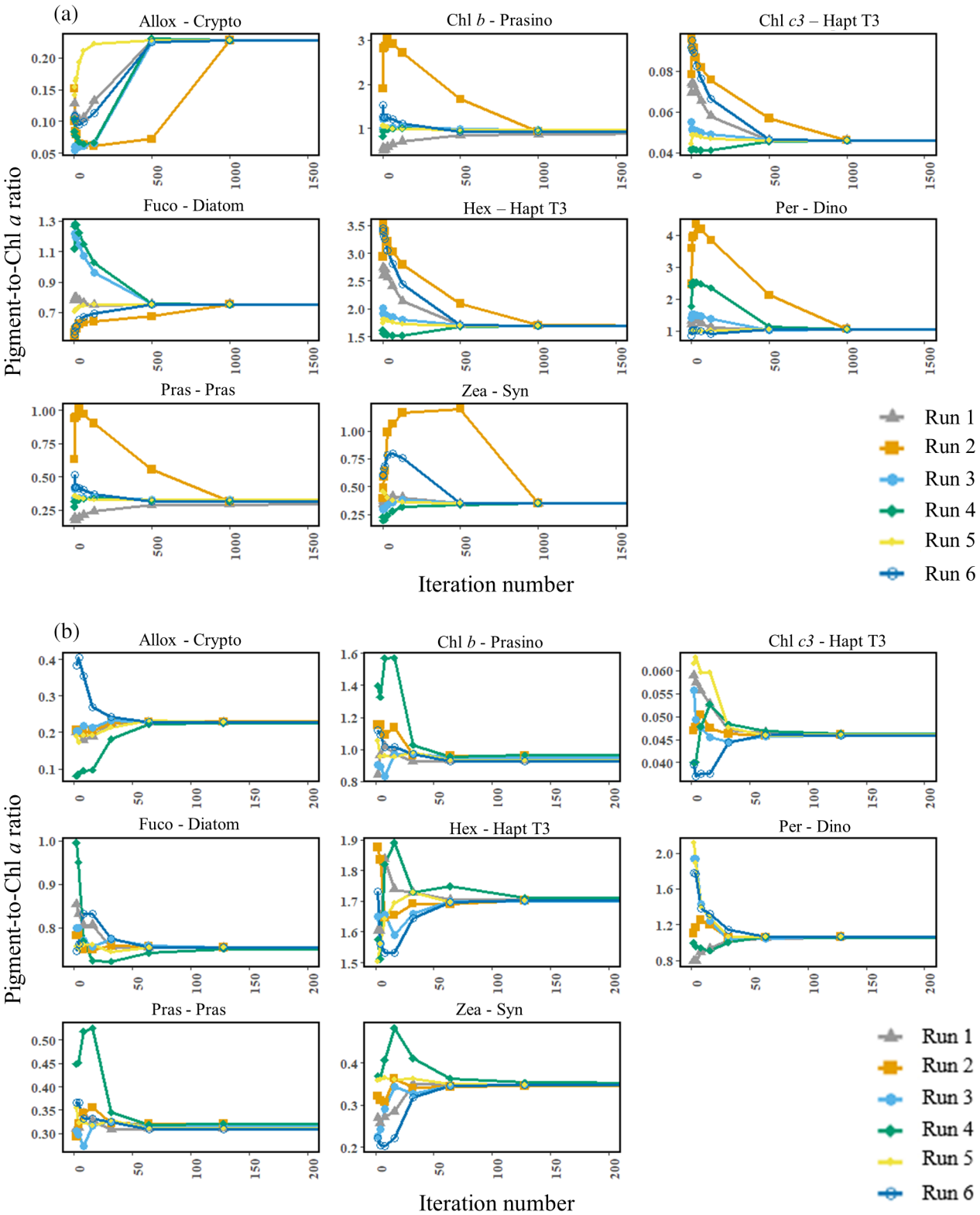

Fig. 3. Convergence of pigment ratios for phytoplankton classes for an inversion sample in synthetic dataset-1 with six different runs. (**a**) Pigment ratio convergence with the ALS method. (**b)** Pigment ratio convergence using the SDA method.

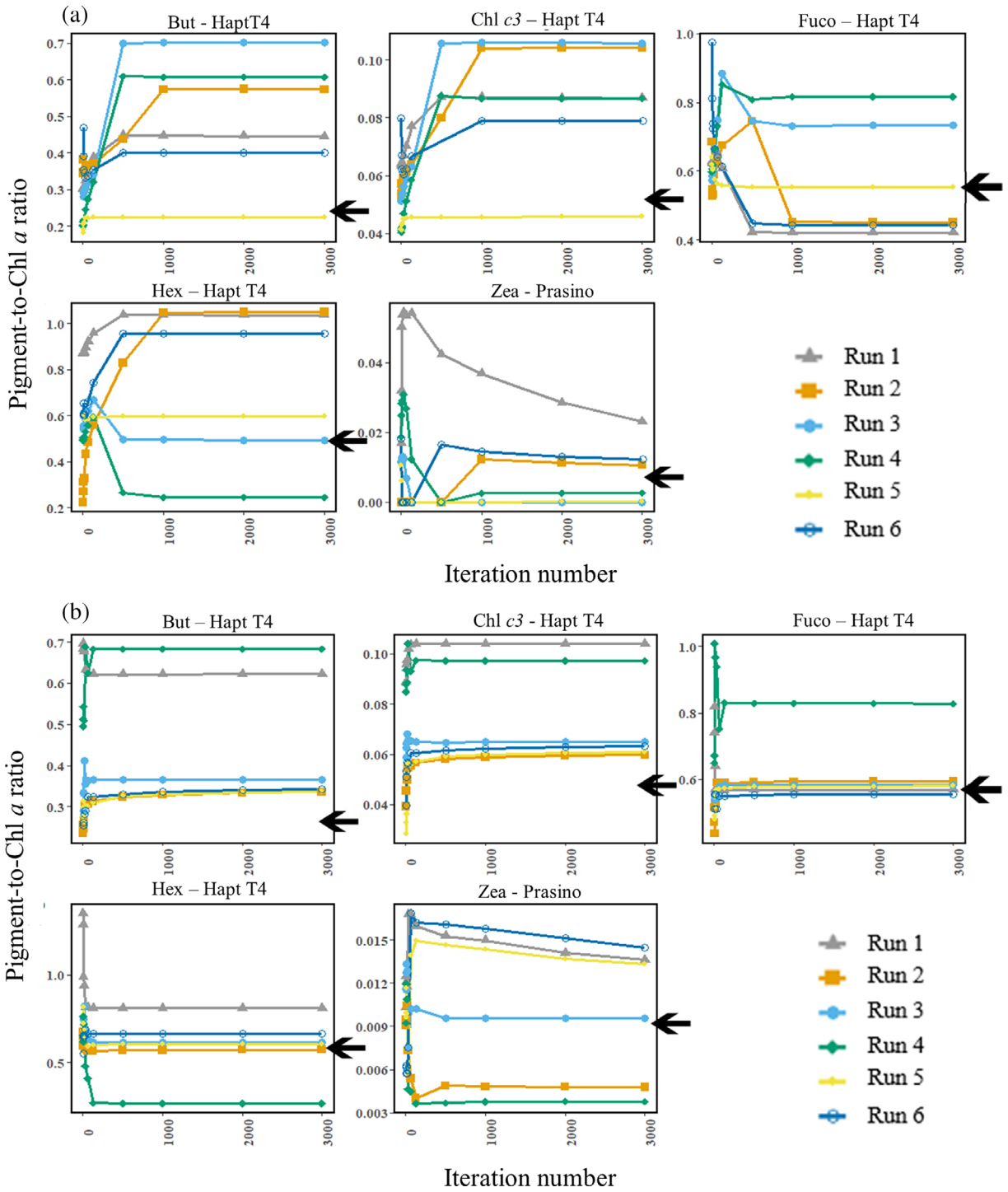

Fig. 4. Pigment ratios that failed to converge for an inversion sample in synthetic dataset-1 with six different runs. (**a**) Pigment ratios determined from the ALS algorithm. (**b**) Pigment ratios determined from the SDA method. The black arrows signify the true value of the pigment ratio.

approach; coupled with a low RMSE. This indicated that the R versions of CHEMTAX replicate the original program well. CHEMTAX-2 displayed a higher RMSE than CHEMTAX-1. As the CHEMTAX-2 configuration has a randomization component coupled with a short iteration length, there is a greater likelihood that $\mathbf{C}_{\mathrm{Est}}$ will reach slightly different minima, as the method is more sensitive to the starting $\mathbf{F}_{\mathrm{Est}}$.

Table 3. Summary statistics for pigments (S_{True}) in synthetic dataset-1 in units of mg m^{-3}.

Summary statistics for synthetic dataset-1 (mg m^{-3})										
Pigment/summary	**Chl *c3***	**Per**	**Fuco**	**Zea**	**Pras**	**But**	**Hex**	**Allox**	**Chl *b***	**Chl *a***
Mean	0.018	0.034	0.541	0.009	0.042	0.074	0.299	0.006	0.125	1.134
SD	0.007	0.017	0.154	0.004	0.067	0.037	0.100	0.002	0.202	0.303
Median	0.016	0.030	0.520	0.008	0.021	0.066	0.283	0.006	0.064	1.087
Max	0.079	0.245	1.850	0.078	1.768	0.395	1.037	0.017	7.368	7.518
Min	0.004	0.003	0.149	0.001	0.000	0.008	0.086	0.002	0.001	0.420

Optimization techniques

The ALS and SDA were analyzed independently of the simulated annealing method and tested on an inversion sample from synthetic dataset-1. When starting from random $\mathbf{F}_{Est}$, neither the ALS nor SDA methods were able to successfully determine $\mathbf{F}_{True}$. However, both methods proved effective in determining pigment ratios when the starting $\mathbf{F}_{Est}$ were within a $\pm$ 100% range of their true values. This highlights that both methods are suitable for finding local minima (Fig. 3). Accurate values for pigment ratios were obtained for 8 of the 13 pigment markers for both approaches. Using the ALS approach, convergence occurred within 1000 iterations. When using the SDA, convergence typically occurred within the first 100 iterations, making it a more computationally effective tool for finding the local minima, with computation times of $\sim$ 0.3 of ALS (Fig. 3). For this reason, the subsequent analysis of results focusses on the SDA method. Pigment markers that did not converge to their true pigment-to-Chl *a* ratios belong to

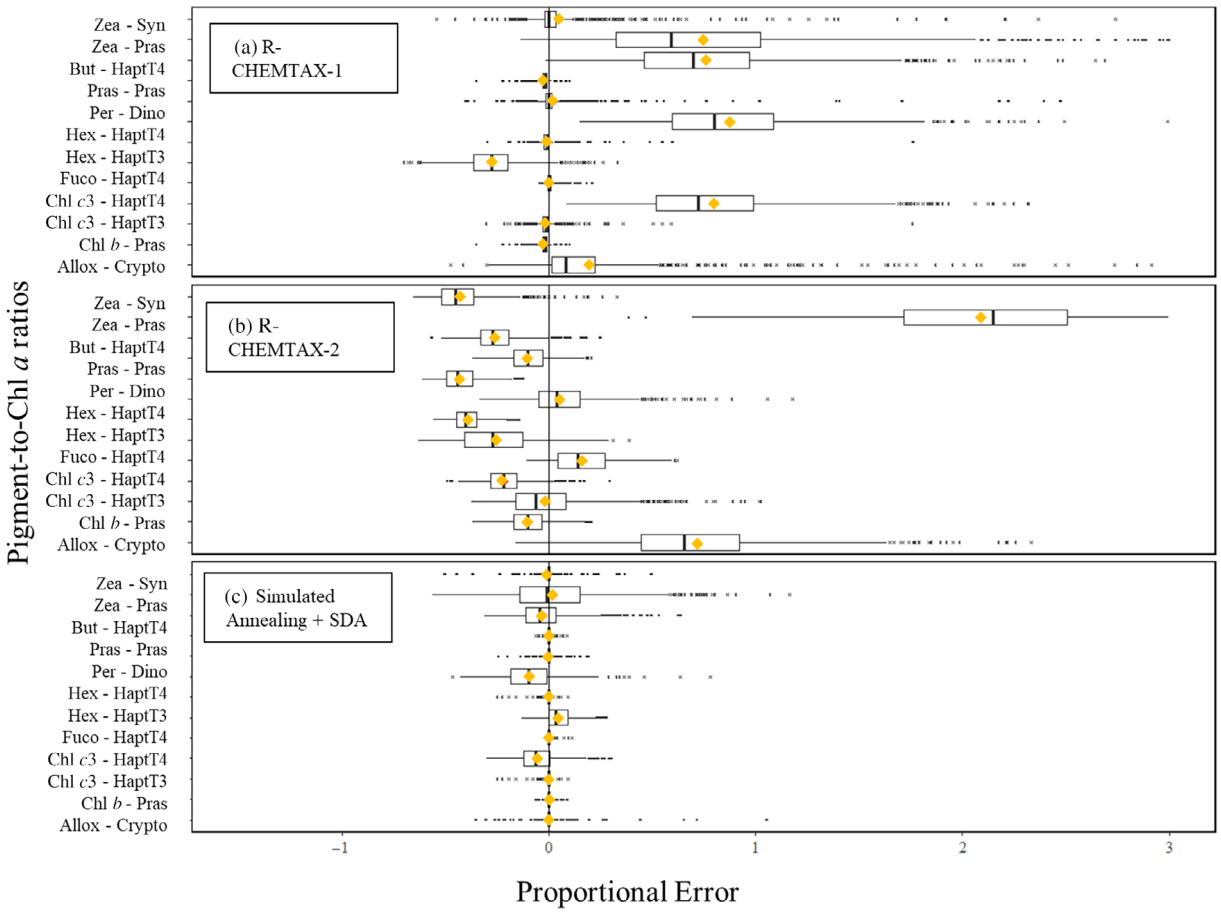

Fig. 5. Distribution of error in pigment-to-chlorophyll ratios for synthetic dataset-1 using the R-CHEMTAX-1 (**a**), R-CHEMTAX-2 (**b**), and simulated annealing + SDA (**c**) approaches. The gold diamond indicates the %bias for each class and the black vertical line is the median proportional error.

the haptophyte-T4 class, which shares markers with the haptophyte-T3 class/diatoms, and the prasinophyte class that shares Zea with *Synechococcus* (Fig. 4).

Synthetic dataset-1 inversion analysis

Within synthetic dataset-1, concentrations of total Chl *a* ranged between 0.4 and 7.4 mg m^{-3}, with Hex and Fuco being the most abundant pigments. For Fuco, a major pigment shared between both haptophyte and diatom classes, concentrations ranged between 0.149 and 1.85 mg m^{-3}, whereas for Hex, the primary diagnostic pigment for haptophytes, concentrations ranged between 0.086 and 1.037 mg m^{-3}. Other pigment concentrations average below 0.1 mg m^{-3}, except for Chl *b* at 0.125 mg m^{-3} (Table 3).

Community biomass was typically co-dominated by haptophyte-T4 and diatoms; biomass for *Synechococcus*, cryptophytes, and dinoflagellates was low, with mean biomass values for these classes < 0.1 mg Chl *a* m^{-3}. To account for occasional prasinophyte blooms in the Southern Ocean, their biomass could vary up to 7 mg Chl *a* m^{-3}, although < 1% of prasinophyte samples had concentrations higher than 0.1 mg Chl *a* m^{-3} (Karl et al. 1991; Peeken 1997; Litchman 2006).

R-CHEMTAX-1 typically resolved pigment ratios with higher accuracy as median values for pigment ratios in the diatoms, haptophyte-T3, prasinophytes and dinoflagellates classes were within 10% of their true values (Fig. 5). Pigment ratios that are shared between multiple classes were poorly resolved by R-CHEMTAX-1, with median proportional errors typically greater than 50%. Proportional error and bias associated with pigment ratios for the R-CHEMTAX-2 method were frequently negative, showing that this approach tends to underestimate pigment ratios. Cryptophytes, haptophyte-T3, *Synechococcus*, and diatoms were underestimated by the R-CHEMTAX-2 approach, whereas prasinophytes, haptophyte-T4 and dinoflagellates pigment ratios were frequently overestimated (Fig. 4). Median proportional errors associated with pigment ratios for CHEMTAX-2 were typically < 50%, with the exception of Zea for prasinophytes and Allox for cryptophytes.

Both the R-CHEMTAX-1 and R-CHEMTAX-2 approaches show moderate to high R^2 between $\mathbf{C}_{Est}$ and $\mathbf{C}_{True}$ for

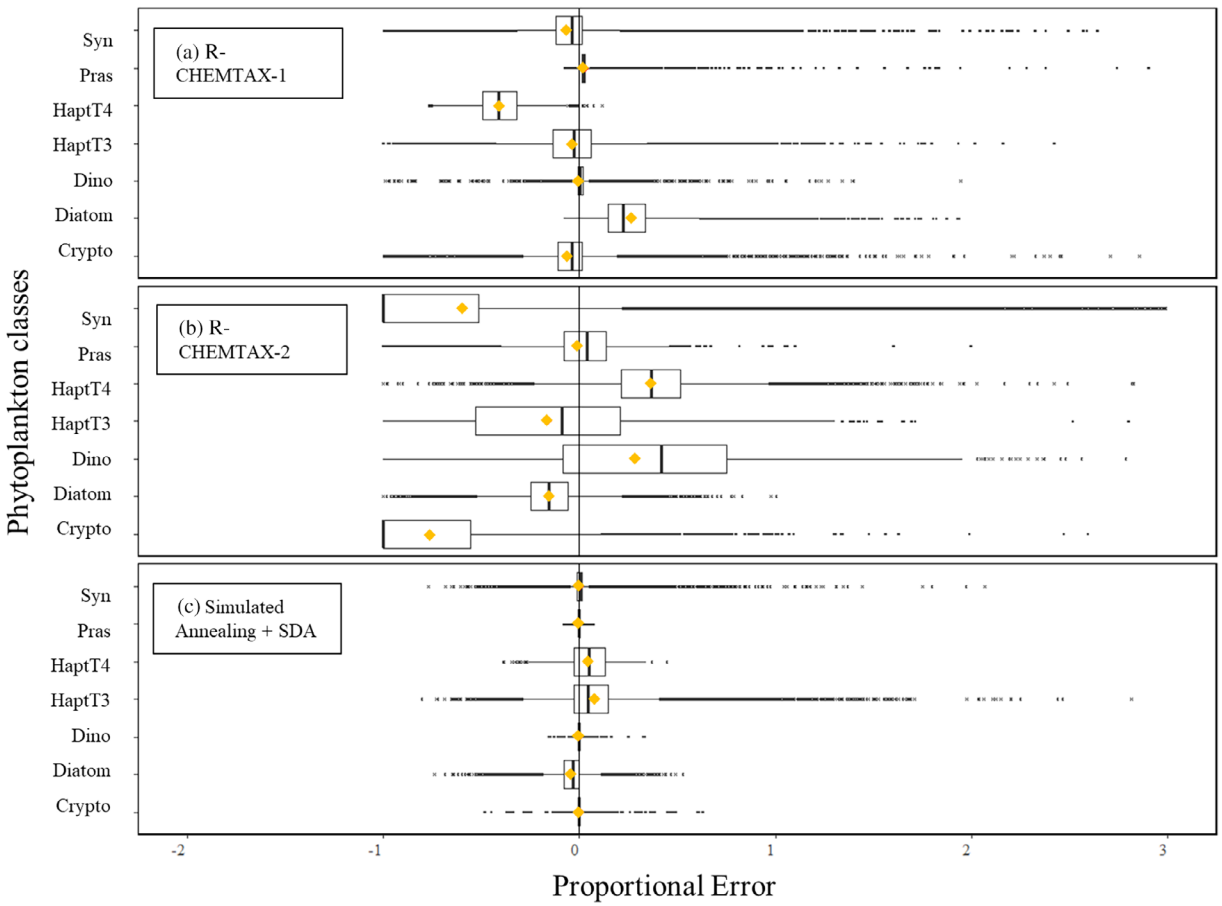

Fig. 6. Distribution of proportional error for phytoplankton classes within synthetic dataset-1, using the R-CHEMTAX-1 (**a**), R-CHEMTAX-2 (**b**), and simulated annealing + SDA (**c**) approaches. The gold diamond indicates the percent bias for each class and the black vertical line is the median proportional error.

Table 4. Summary statistics for the performance of inversion methods to deriving class abundances in synthetic dataset-1.

Inversion analysis summary statistics for class abundances in synthetic dataset-1								
Metric	**Prasinophytes**	**Dinoflagellates**	**Cryptophytes**	**Haptophyte-T3**	**Haptophyte-T4**	***Synechococcus***	**Diatoms**	**Total**
Sample size	31,779	31,779	31,779	31,779	31,779	31,779	31,779	222,465
R-CHEMTAX-1								
R^2	1.000	0.972	0.402	0.867	0.792	0.716	0.852	0.898
%Bias	2.715	0.270	−6.154	−3.922	−40.850	−6.941	26.399	−4.069
RMSE	0.005	0.003	0.006	0.017	0.141	0.005	0.143	0.076
sMAPE (%)	2.5	3.1	14.6	16.1	53.0	16.7	22.0	18.0
R-CHEMTAX-2								
R^2	0.988	0.601	0.039	0.663	0.849	0.043	0.789	0.894
%Bias	−1.378	28.305	−76.431	−16.230	36.849	−51.848	−15.377	−13.730
RMSE	0.032	0.024	0.023	0.048	0.133	0.023	0.111	0.071
SMAPE	17.5	60.5	78.0	41.7	38.2	96.0	18.0	70.0
Simulated annealing								
R^2	1.000	1.000	0.988	0.884	0.930	0.961	0.955	0.987
%Bias	−0.041	−0.004	0.058	8.027	4.581	0.531	−4.293	1.265
RMSE	0.002	0.000	0.001	0.017	0.043	0.002	0.042	0.024
sMAPE (%)	< 1	< 1	< 1	12.6	10.1	3.5	6.7	4.7

phytoplankton classes of high biomass (Fig. 6). For example, diatoms and haptophyte-T3 have $R^2 > 0.85$ using R-CHEMTAX-1 and > 0.65 for CHEMTAX-2. Low biomass classes, such as *Synechococcus* and cryptophytes, presented a challenge for both approaches, with R^2 values between 0.4 and 0.72 for R-CHEMTAX-1 and no correlative relationships for R-CHEMTAX-2 (Table 4). The sMAPE between true and estimated phytoplankton classes ranged between 2.5% and 53% for R-CHEMTAX-1, with the highest values for haptophyte-T4 and diatoms, whereas for the R-CHEMTAX-2 approach the sMAPE ranged between 18% and 96%.

When using simulated annealing with the SDA, the R^2 between estimated and true phytoplankton classes was greater than 0.9, except for haptophyte-T3 (0.89). In addition, the median proportional error for the Chl *a* assigned to each phytoplankton class was below 10%, indicating higher accuracy than the CHEMTAX methods in resolving class abundances. Eight of the 13 pigment ratios converged to their true values (Fig. 5), and for those pigments and classes that did not (pigments within haptophyte-T4 class, and Zea for prasinophytes), median values of pigment ratios were within 15% of their true values, indicating an improvement from R-CHEMTAX methods.

The simulated annealing approach outperformed traditional chemotaxonomic methods at calculating phytoplankton class abundances and resolving pigment ratios for classes with shared pigment ratios. The R^2 between $\mathbf{C}_{Est}$ and $\mathbf{C}_{True}$ is 0.99 with a RMSE of 0.024 mg Chl *a* m^{-3} and an sMAPE of 4.8%, over 222,465 data points (Fig. 7).

The sMAPE between $\mathbf{C}_{Est}$ and $\mathbf{C}_{True}$ increased by ~ 5% with each noise increment added, indicating that the method was sensitive to sampling error (Fig. 8). When sample size was low, sMAPE was high; however, this stabilizes with a sample size of 12 for 0–2% noise added, a sample size of 20 for 4–6% noise added and at larger sample sizes when > 6% noise was added. This indicates that datasets with a larger number of samples will be more resistant to noise introduced from measurement or sampling error.

When the biomass of *Synechococcus* was set to zero, the simulated annealing approach worked well at resolving the biomass for the class. As shown from Fig. 9, the distribution of *Synechococcus* biomass was very low (median = 7.0×10^{-4} mg Chl *a* m^{-3}). Although the median biomass is not zero, it made up a very minor proportion (usually < 0.001%) of total biomass for all inversion samples.

The condition number for synthetic dataset-1 was very low with a median value of 1663. As shown by Fig. 10, with the addition of the haptophyte-T8 to synthetic dataset-1, the matrix becomes singular (uninvertible), and the condition number increases to 1.9×10^{16}. When inverted, the sMAPE associated with the class abundances increases from 5% to 95%. This indicates that multiple phytoplankton classes with very similar pigment profiles cannot be used.

Synthetic dataset-2 inversion analysis

The median condition number for the more complex, synthetic dataset-2 was 60,879 (*see* Supporting Information Fig. S1). As synthetic dataset-2 has many groups that share pigment markers, the condition number was higher than synthetic dataset-1. Despite this, the condition number was substantially lower than when a third haptophyte subgroup was added to synthetic dataset-1 (i.e., 1.9×10^{16}).

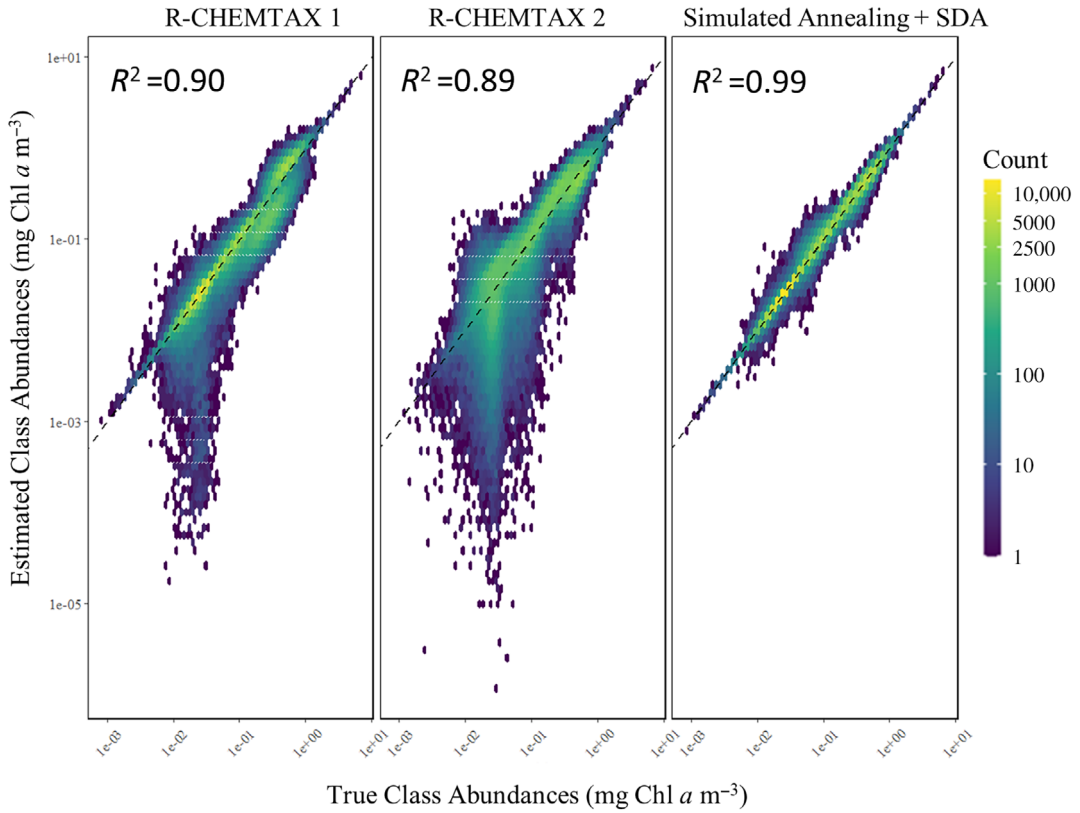

Fig. 7. Density plot of true class abundances and predicted class abundances for R-CHEMTAX-1, R-CHEMTAX-2, and simulated annealing + SDA approaches in synthetic dataset-1.

For synthetic dataset-2, biomass ranged from moderate (0.56 mg Chl *a* m^{-3}), to high, typical of eutrophic conditions (~ 5 mg Chl *a* m^{-3}). Mean values for biomass represented Antarctic coastal waters with a value of 1.13 mg Chl *a* m^{-3}. Concentrations for Fuco ranged between 0.057 and 3.87 mg m^{-3}, with an average value of 0.51 mg m^{-3}, whereas

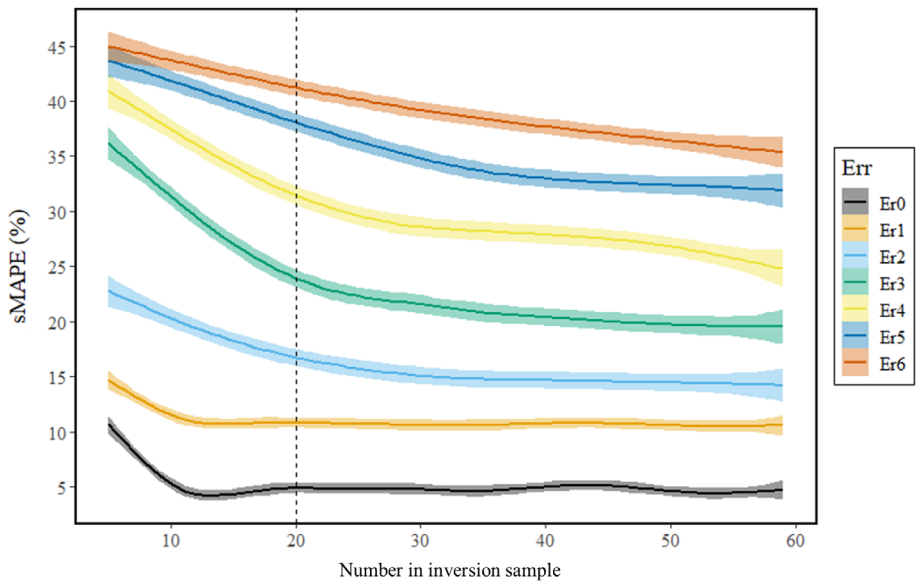

Fig. 8. The sMAPE for class abundances for varying levels of error addition for different sample sizes, tested on synthetic dataset-1. Er0 indicated the sMAPE with no additional noise, whereas Er1 is 2% added error, Er2 is 4% added error, Err3 is 6% added error, Err4 is 8% added error, Err5 is 10% added error, and Er6 is 12% added error. The shading indicates the standard error around the mean value for each number in inversion sample.

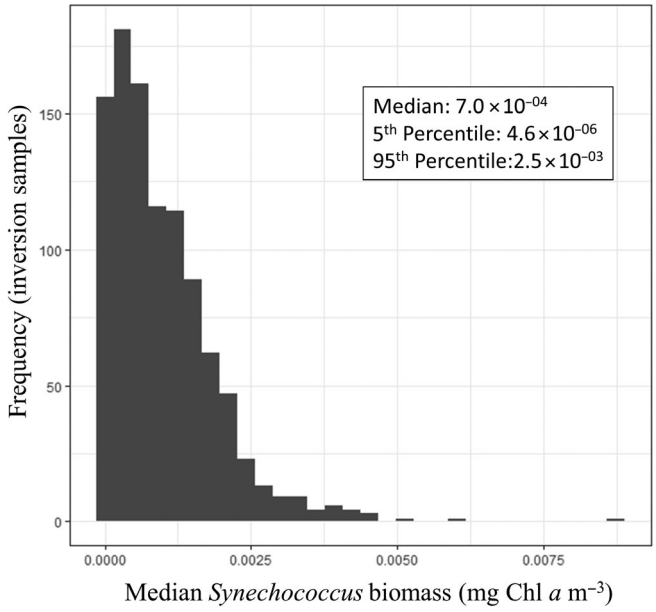

Fig. 9. The median biomass for *Synechococcus* for each inversion sample, when the real biomass value is zero.

Hex concentrations ranged between 0.021 and 2.06 mg m^{-3} (Table 5). The remaining pigment concentrations were low, with mean concentrations < 0.1 mg m^{-3}.

When using R-CHEMTAX-1 and R-CHEMTAX-2, pigment ratios generally did not converge to their true values (Fig. 11). Compared to synthetic dataset-1, a higher proportion of pigment ratios did not converge when the pigment was shared between multiple phytoplankton classes. Similar to synthetic dataset-1, pigment ratios for the cryptophyte and dinoflagellate classes were well resolved by R-CHEMTAX-1 as these classes have the unambiguous marker pigments Per and Allox. Although proportional error for pigment ratios can be high using these approaches, it is important to note that true pigment ratios can be small for many phytoplankton classes (i.e., Chl *c3* for diatom-1, mean = 0.016). A small absolute deviation from the true value then causes a large increase in the proportional error.

For synthetic dataset-2, neither the R-CHEMTAX-1 or R-CHEMTAX-2 approaches successfully resolved class abundances for the green algae lineages ($R^2 < 0.26$; Table 5). Proportional error for classes with low biomass such as diatom-1 was high, with a maximum of 202% for R-CHEMTAX-1 and 558% for CHEMTAX-2, indicating difficulty in delineating diatom classes that share common pigments. Similarly, both approaches struggled to distinguish between the green algae classes, with an overestimation of chlorophytes and underestimation of prasinophytes. Phytoplankton classes with unambiguous pigment markers are better represented by R-CHEMTAX-1, with median class abundances within 10% of their true values (Fig. 12). The sMAPE value between $\mathbf{C_{Est}}$ and $\mathbf{C_{True}}$ was 49% for R-CHEMTAX-1 and 63% for R-CHEMTAX-2. These higher sMAPE values compared to those in synthetic dataset-1 indicate that the CHEMTAX methods struggled to resolve the more complex inversion samples in synthetic dataset-2 (Table 6).

Simulated annealing was effective at resolving phytoplankton class abundances for synthetic dataset-2. Proportional errors associated with both haptophyte classes were low, with median errors < 10%. The proportional error for

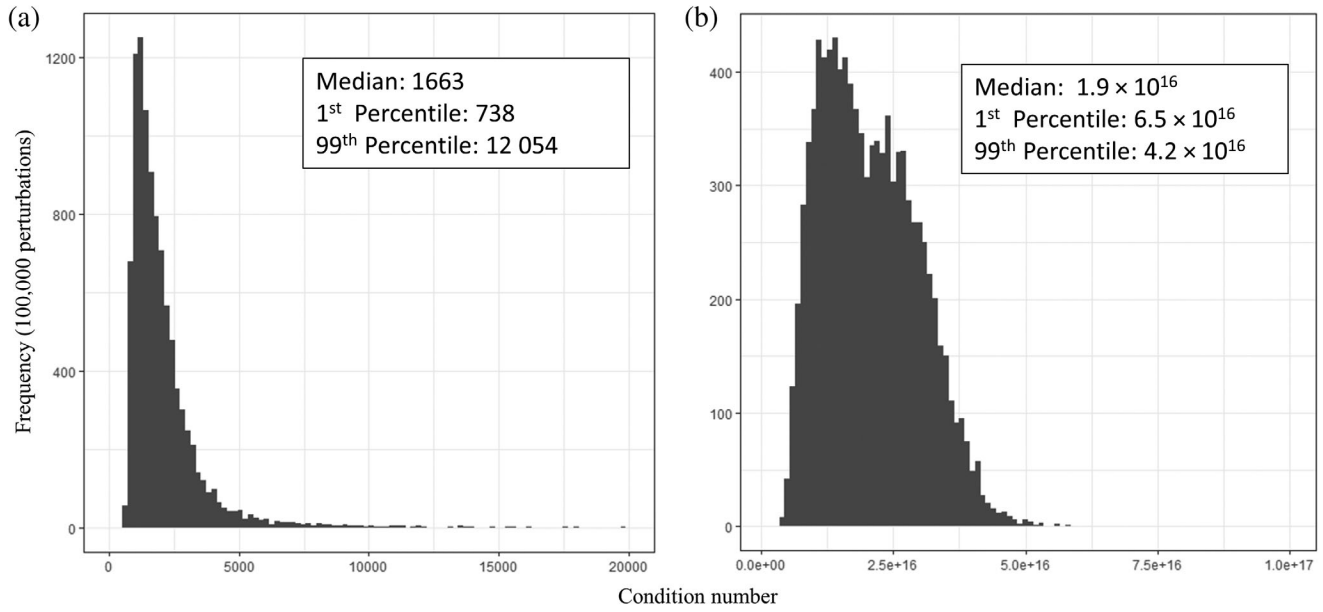

Fig. 10. The distribution of condition numbers for 100,000 perturbations using synthetic dataset-1. (**a**) Without the addition of the haptophyte-T8 group. (**b**) With the inclusion of the haptophyte-T8 group.

Table 5. Summary statistics for pigments (S_{True}) in synthetic dataset-2 in units of mg m^{-3}.

Summary statistics for synthetic dataset-2 (mg $^{-3}$)												
Pigment/summary	**Chl *c3***	**Chl *c1***	**Per**	**Fuco**	**Neo**	**Pras**	**Violax**	**Hex**	**Allox**	**Lut**	**Chl *b***	**Chl *a***
Mean	0.160	0.013	0.011	0.513	0.003	0.004	0.002	0.420	0.031	0.001	0.024	1.406
SD	0.135	0.013	0.007	0.402	0.001	0.001	0.000	0.375	0.013	0.001	0.005	0.346
Median	0.112	0.006	0.009	0.323	0.003	0.004	0.002	0.340	0.028	0.001	0.024	1.360
Min	0.013	0.000	0.001	0.057	0.002	0.002	0.001	0.021	0.005	0.000	0.010	0.564
Max	0.908	0.107	0.081	3.874	0.007	0.010	0.005	2.060	0.165	0.007	0.048	4.988

the diatom-1 class was the largest, with an overestimation of 34%. As median biomass was 0.026 mg Chl *a* m^{-3} for the diatom-1 class, an error of 34% equates to a deviation of only 0.009 mg Chl *a* m^{-3} and so the absolute error was low. Green algae classes were well represented by the simulated annealing approach, prasinophytes converged to their true values, and chlorophytes displayed a slight underestimation.

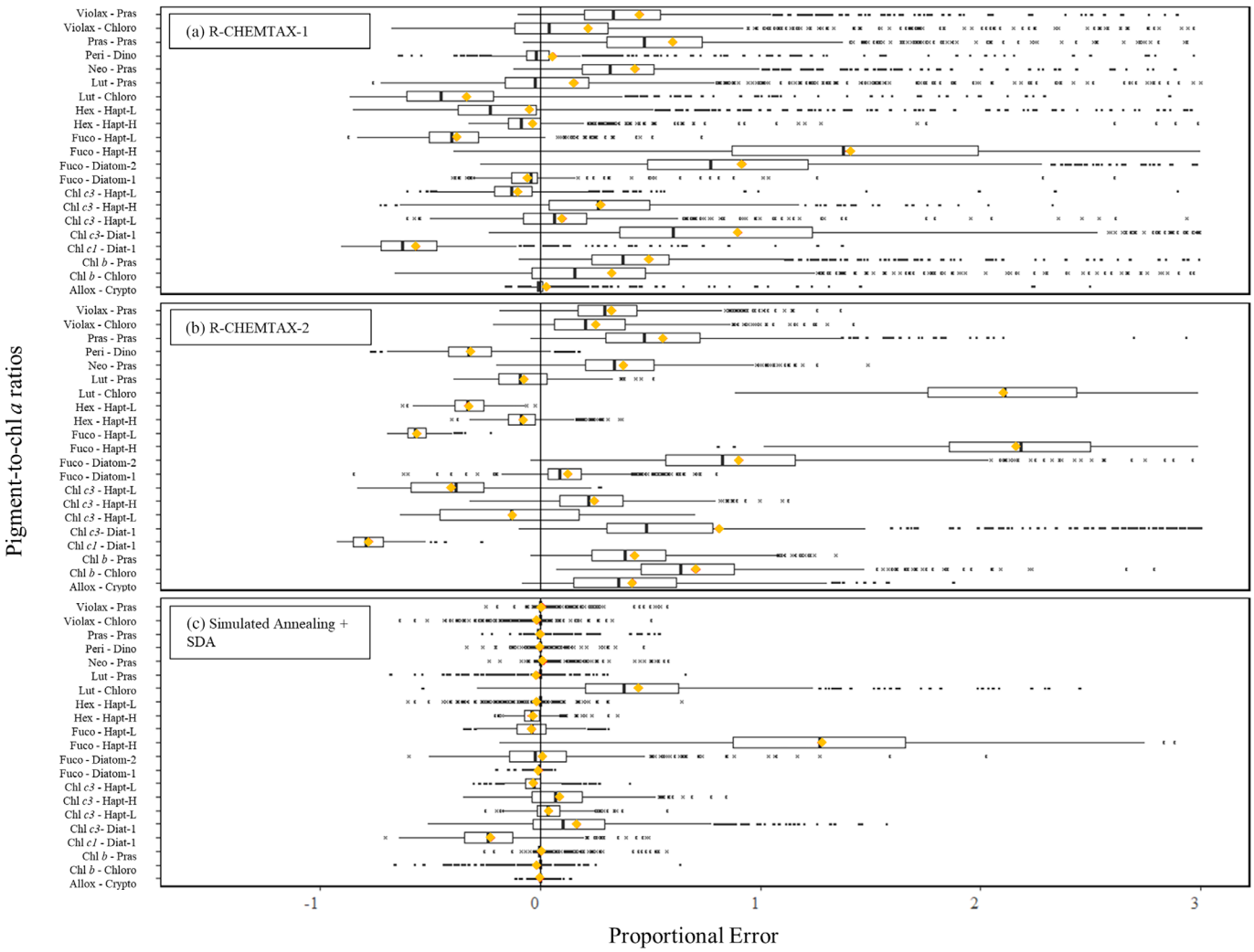

Fig. 11. Distribution of proportional error for phytoplankton pigment ratios within synthetic dataset-2, using the R-CHEMTAX-1 (**a**), R-CHEMTAX-2 (**b**), and simulated annealing + SDA (**c**) approaches. The gold diamond indicates the percent bias for each class and the black vertical line is the median proportional error.

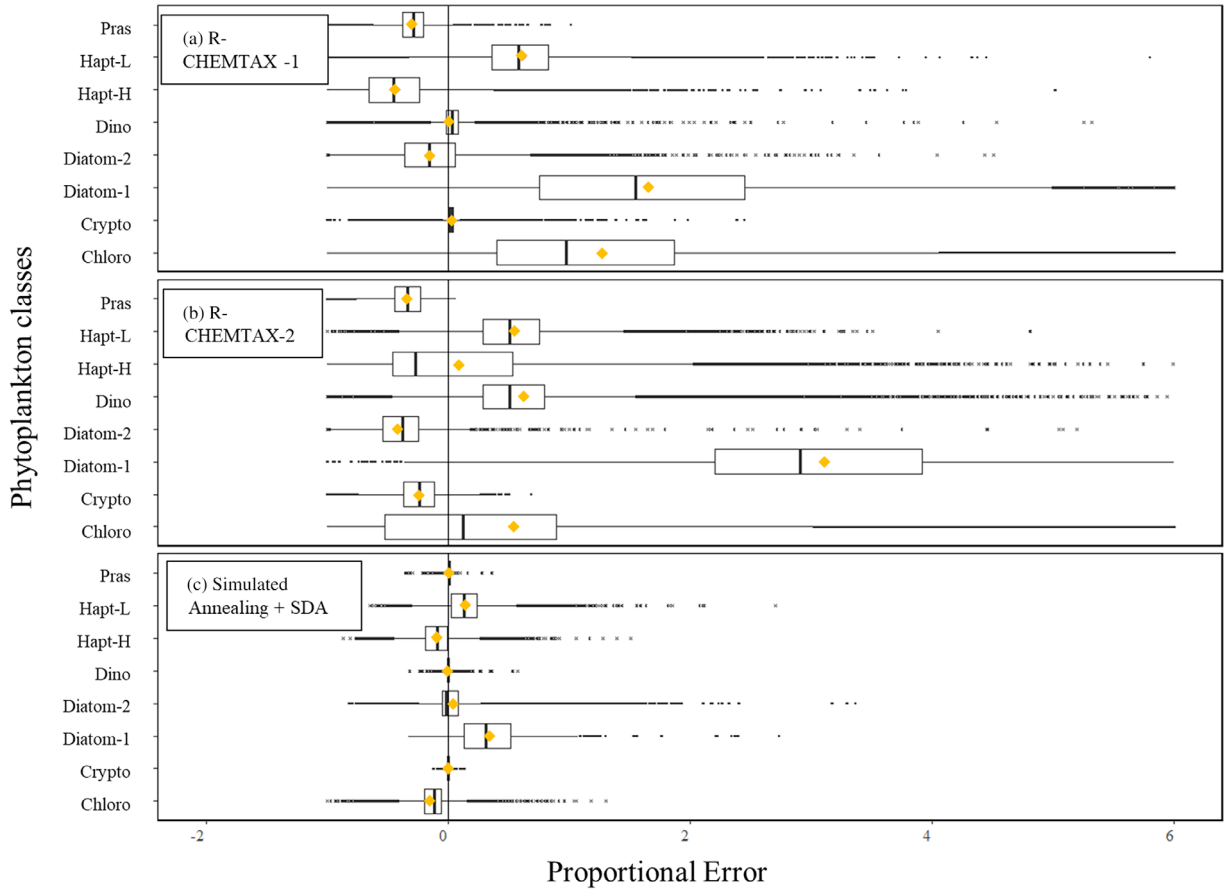

Fig. 12. Distribution of proportional error for phytoplankton classes within synthetic dataset-2, using R-CHEMTAX-1 (**a**), R-CHEMTAX-2 (**b**), and simulated annealing + SDA (**c**) approaches. The gold diamond indicates the percent bias for each class and the black vertical line is the median proportional error.

Table 6. Summary statistics for the performance of inversion methods at deriving class abundances in synthetic dataset-2.

Inversion analysis summary statistics for class abundances in synthetic dataset-2

	Prasinophyte	Chlorophyte	Cryptophyte	Diatom-1	Diatom-2	Dinoflagellates	Haptophyte-H	Haptophyte-L	Total
Sample size	31,814	31,814	31,814	31,814	31,814	31,814	31,814	31,814	254,526
R-CHEMTAX-1									
R^2	0.253	0.046	0.950	0.774	0.897	0.756	0.782	0.931	0.711
%Bias	−30.629	185.554	3.388	202.792	−14.700	1.213	−42.996	61.724	45.793
RMSE	0.016	0.012	0.012	0.216	0.214	0.004	0.248	0.235	0.162
sMAPE (%)	38.6	79.7	5.0	90.5	41.0	14.5	74.0	46.0	48.0
R-CHEMTAX-2									
R^2	0.250	0.038	0.766	0.837	0.946	0.682	0.819	0.919	0.717
%Bias	−34.121	168.587	−24.945	558.141	−40.879	63.047	9.409	54.500	94.217
RMSE	0.018	0.013	0.042	0.211	0.220	0.012	0.231	0.229	0.158
sMAPE (%)	44.0	94.0	31.0	132.0	58.0	45.0	51.0	42.0	63.0
Simulated annealing									
R^2	0.973	0.965	1.000	0.924	0.996	0.999	0.966	0.962	0.977
%Bias	−0.744	−13.732	0.028	34.014	4.681	0.022	−9.335	14.520	3.868
RMSE	1.1 E-03	5.8 E-04	6.3 E-04	0.034	0.039	2.5 E-04	0.088	0.077	0.045
sMAPE (%)	1.2	17.0	< 1	28.6	11.0	< 1	15.7	16.0	11.2

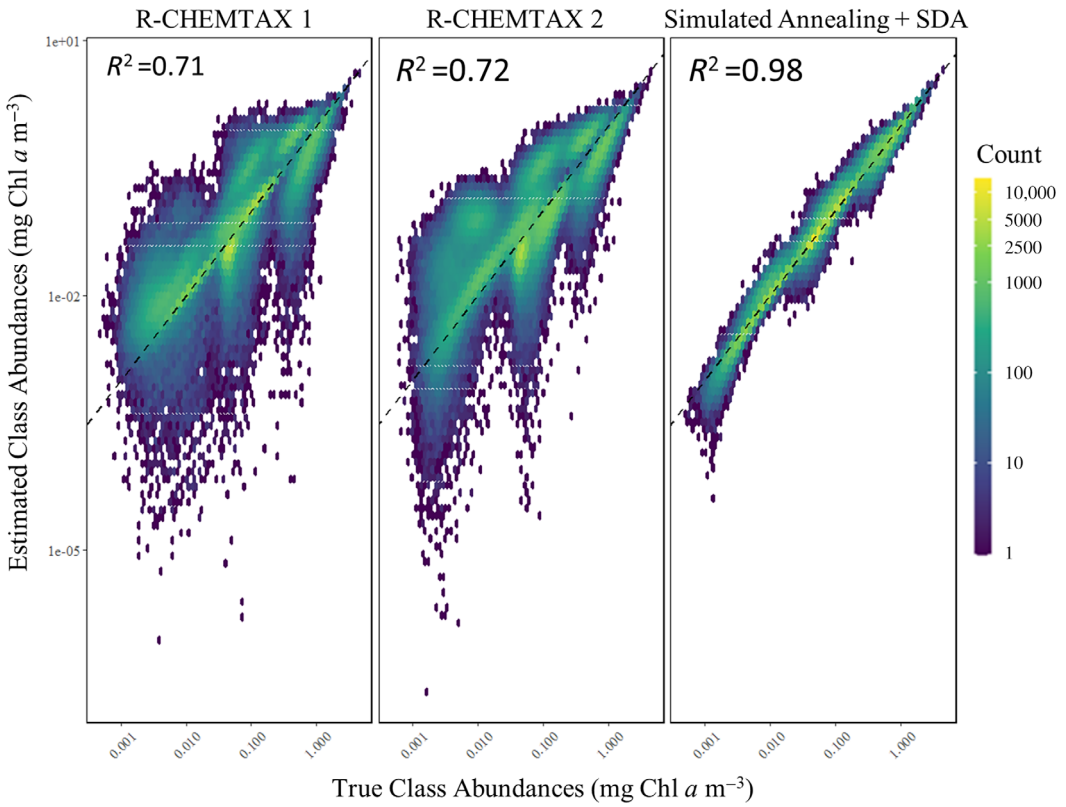

Fig. 13. Density plots of true class abundances and predicted class abundances for R-CHEMTAX-1, R-CHEMTAX-2, and simulated annealing + SDA approaches using synthetic datasets in synthetic dataset-2.

The R^2 between $\mathbf{C_{Est}}$ and $\mathbf{C_{True}}$ was 0.98 for synthetic dataset-2; however, the RMSE and sMAPE were marginally higher than the previous synthetic dataset at 0.045 mg Chl *a* m^{-3} and 11%, respectively (Table 6). This is unsurprising given the larger condition number for synthetic dataset-2. The accuracy of simulated annealing at estimating $\mathbf{C_{True}}$ and $\mathbf{F_{True}}$ for synthetic dataset-2 show a marked improvement from the CHEMTAX method, which had R^2 values that ranged between 0.71 and 0.72, RMSE values of 0.148–0.162 mg Chl *a* m^{-3} and sMAPE values of 48–63% (Fig. 13; Table 6).

Discussion

Although there will likely be differences between R-CHEMTAX and CHEMTAX, our results show that R-CHEMTAX replicates CHEMTAX v1.95 very closely, suggesting that it was a suitable proxy to use in-place of the CHEMTAX graphical user interface. The creation of R-CHEMTAX provided the capability for us to test the algorithm on many inversion samples without the requirement of manual data input.

When analyzing different optimization techniques, we found that the SDA was more effective, with faster convergence times and more accurate results for "unconverged" pigments (did not reach true solution), than ALS. Extending the iteration count for both the ALS and SDA did not improve estimates for the pigments that did not converge; after a local minimum was determined, pigments would remain static for the remaining iterations (Fig. 4). Pigments that failed to converge shared three characteristics: they belonged to the same algal class, the algal class shared pigments with another class, and either the pigment ratio or pigment concentration was low. Pigments that did not converge to their true ratios were attributed to the haptophyte-T4 class that share marker pigments with haptophyte-T3, and diatoms (Fig. 4). For the SDA, pigment ratios for the haptophyte-T4 class either converged at the wrong values (i.e., Chl *c3*), or at values close to their true ratios such as Fuco (Fig. 4). The pigment Zea did not converge to its true values for the prasinophyte class; this pigment was shared with *Synechococcus* and is a minority pigment for most prasinophytes, with a true pigment ratio of 0.01 in the synthetic datasets used here.

The simulated annealing algorithm was effective at finding global minima with no sensitivity to initial estimates of phytoplankton pigment ratios. Simulated annealing displayed a marked increase in accuracy when compared to the R-CHEMTAX methods, suggesting that it is better suited to determine phytoplankton class abundances from pigment data than R-CHEMTAX. Increased accuracy in deriving class

abundances coupled with the higher accuracy between phytoplankton subgroups will provide better quantification and parameterization of phytoplankton class abundances for biogeochemical models and will aid in further understanding biogeochemical cycles, and ecosystem functions.

Synthetic datasets proved to be a useful tool for testing different chemotaxonomic methods. Literature values for pigment ratios and biomass distributions were used to mimic a natural phytoplankton community from field samples, while accounting for some of the variance between pigments for phytoplankton classes (Zapata et al. 2004). We, however, note that our synthetic datasets do not consider all correlations between different classes; that is, ratios are likely to be correlated due to a particular set of environmental conditions. Cross-correlation between pigment ratios and class abundances would reduce variation in the synthetic datasets because only a subset of the random combinations is possible. Hence, our synthetic datasets likely represent a harder test of the inversion methods than a more realistic dataset that includes co-variance between pigment ratios.

The biomass of cryptophytes and dinoflagellates are typically low in synthetic dataset-1 though we recognize that they can be an important component of the Southern Ocean phytoplankton community (Garibotti et al. 2005; McLeod et al. 2012). Despite the low biomass of these classes, accuracy between $\mathbf{C}_{\text{True}}$ and $\mathbf{C}_{\text{Est}}$ for these classes were high when using simulated annealing due to their unique pigment markers (Allox and Per). We recognize that dinoflagellates present a challenge in chemotaxonomic analysis as they include groups that both include and lack Per, contain pigments commonly associated with haptophyte/diatom classes, and/or utilize photosynthetic apparatus of other phytoplankton classes via kleptoplasty (Tangen and Björnland 1981; Gast et al. 2007; Kang 2010). Dinoflagellates represented in this dataset are only those that contain Per so that further testing of the effect of dinoflagellates should be further considered in the future.

The condition number of the matrices indicates if a feasible solution is available given the phytoplankton groups and pigments selected. By adding a third haptophyte subgroup to synthetic dataset-1, we show that when classes share very similar pigment profiles and are together in the same analysis, the inversion becomes singular (noninvertible). We suggest that the user of *phytoclass* avoids selecting many classes (or subgroups) of phytoplankton with overlapping pigment profiles. The user should make an assessment on their choice of groups/pigments based on the condition number associated with their selection. Prior to analysis, a high condition number will be flagged to the user, and the program will abort if the number is very large. If the condition number of the given $\mathbf{F}_{\text{Est}}$ and $\mathbf{S}_{\text{True}}$ is large, the user should reevaluate the selection of phytoplankton groups in $\mathbf{F}_{\text{Est}}$. Due to potential co-linearity between matrix columns, we advise that pigment inversion methods should only be used to determine the biomass of phytoplankton classes at coarse resolution, without the selection of many subgroups.

Both synthetic datasets omit the presence of pelagophytes, a phytoplankton class that is common in water samples collected in the Southern Ocean (Schlüter et al. 2011). Thus, if pelagophytes were present within a water sample but omitted from the chemotaxonomic analysis, their biomass would be misallocated to another phytoplankton class that share a similar pigment profile. The inversion of synthetic data was approached with the advantage of knowing what algal classes were present. When encountering field-based pigment samples, the phytoplankton classes present will not always be known, and it is important to scientists to consult the literature on the biogeography of phytoplankton classes expected within their water sample, and base their selection on the presence or absence of certain pigment markers (e.g., only including *Synechococcus* if Zea is present). If in doubt, ambiguous phytoplankton classes should be included. Inspection of the condition number test will reveal if the problem has become noninvertible (excessively ambiguous) when it will be necessary to reduce the classes.

When following the R-CHEMTAX-1 method, multiple iterations would cause unambiguous pigment ratios to converge to their true values when starting from random points. However, initial pigment ratios too distant from their true values would remain in local minima outside the reasonable solution space and skew the final estimates. The R-CHEMTAX-2 approach would cause pigment ratios to converge to a local minima close to their initial ratios, indicating that this approach was more sensitive than R-CHEMTAX-1 to the initial ratios used. To overcome this sensitivity, the approach starts from 60 unique $\mathbf{F}_{\text{Est}}$, based on literature values; however, if the solution space has many local minima, 60 random matrices may not be sufficient to cover the feasible solution space.

R-CHEMTAX methods were not able to determine the correct abundances or pigment ratios for low biomass classes such as *Synechococcus* and cryptophytes for synthetic dataset-1, or the green algae classes for synthetic dataset-2. This indicates a natural bias within R-CHEMTAX to resolve phytoplankton classes with higher biomass more accurately. We have discussed two common configurations to using R-CHEMTAX; it is important to highlight that CHEMTAX v1.95 has many settings, and adjustment of these will affect the derivation of $\mathbf{C}_{\text{Est}}$ and $\mathbf{F}_{\text{Est}}$.

Simulated annealing performed well at resolving phytoplankton biomass, with overall accuracy ranging between 89% and 95%. Although the more complex synthetic dataset-2 had not performed as well as synthetic dataset-1, it showed the greatest improvement from the R-CHEMTAX methods at delineating phytoplankton classes with shared pigment biomarkers. This was demonstrated by increased accuracy at resolving haptophyte classes partitioned by their Fe preference when compared to the R-CHEMTAX methods.

We note that the relative errors associated with pigment ratios are frequently higher than the error associated with the derivation of phytoplankton class abundances (Figs. 5, 6, 11, 12). Exceptions to this rule do however occur and are exampled in synthetic dataset-2 (R-CHEMTAX approaches) where the relative error for the abundances of diatoms-1 are higher than their dominant marker pigment, Fuco. This likely occurred as in this dataset, Fuco is divided between four phytoplankton classes, and error associated with the Fuco in any of these classes will compound and affect the accuracy in estimates of class abundances that share the pigment.

The addition of noise in $\mathbf{S}_{\mathrm{Err}}$ widened the gap between $\mathbf{C}_{\mathrm{True}}$ and $\mathbf{C}_{\mathrm{Est}}$, and the sMAPE between each noise level would typically increase by $\sim 5\%$. This highlights the importance of accurately sampling and analyzing phytoplankton pigments through HPLC laboratory protocols. Uncertainty for HPLC methods is typically low for both instrument calibration and the preparation of samples but can be higher for replicate samples (Tomić et al. 2012). As error was greatest at a lower sample size, we recommend sample sizes of at least 12 samples when using the simulated annealing approach, whereas if higher uncertainty is expected, we recommend a sample size of 20 similar samples, and at the very minimum a sample size of 12. Higher sample size will typically offer more leverage to accurately determine class abundances, especially if high uncertainty is anticipated with sampling or analytical procedures.

As pigment ratios change with environmental conditions, such as light levels or nutrient regimes (Schlüter et al. 2000; Henriksen et al. 2002), pigment samples should be clustered prior to analysis to ensure that all samples within $\mathbf{S}_{\mathrm{True}}$ will share similar values for $\mathbf{F}_{\mathrm{True}}$. Following the work of Nunes et al. (2018) and Vaillancourt et al. (2018), we recommend data are always clustered based on environmental conditions or through established hierarchal clustering methods (such as the Ward method; Punj and Stewart 1983) on pigment concentrations that have been normalized to Chl *a* and transformed using the Box–Cox method (Murtagh and Legendre 2014). Samples collected from the field will have varying pigment ratios, as such, prior to inversion we strongly emphasize the importance of correctly clustering samples so that during inversion, samples with similar pigment ratios will be analyzed together. Further testing of clustering methods in large datasets where pigment ratios differ greatly is required.

The effectiveness of the simulated annealing algorithm is dependent upon the iteration limit and step used. With a step of 0.009 and a liberal iteration limit of 500, pigment ratios would converge to, or close to their true values. Convergence may occur at lower (or higher) iteration limits; this will, however, depend upon the complexity of the dataset and pigment concentrations used. If the RMSE associated with the inversion is > 0.1, we suggest that the user increases both the step and iteration limit of the annealing algorithm, or recluster their data. The minimum and maximum values for the simulated annealing randomization were 15% greater than those reported in the literature to account for environmental conditions that may alter the pigment ratios to value lower or higher than those reported in the literature. Although this is a reasonable range, it will not cover outliers created by extreme environmental conditions that would favor the synthesis of accessory pigments over Chl *a*, thereby increasing pigment-to-Chl *a* ratios.

This paper will be followed by an open-source R package *phytoclass* that implements the simulated annealing technique discussed. We have implemented the flexibility for the user to predefine maximum and minimum values for pigment ratios, alongside iteration limits and numerous other setting within the *phytoclass* package while also providing default values. We anticipate that with global open-source pigment data available (Peloquin et al. 2013), the *phytoclass* program will prove to be a useful tool for the validation of biogeochemical modeling exercises. As *phytoclass* is built within the programming language "R," it can easily be implemented into the workflow of remote sensing scientists, modelers, and field ecologists. We are interested to see how chemotaxonomic approaches such as *phytoclass* perform with freshwater ecosystems, the sea ice microbial community (Pinkerton and Hayward 2021), and with different biomarkers such as lipid concentrations or phycobiliproteins.

References

Arteaga, L. A., E. Boss, M. J. Behrenfeld, T. K. Westberry, and J. L. Sarmiento. 2020. Seasonal modulation of phytoplankton biomass in the Southern Ocean. Nat. Commun. **11**. doi:10.1038/s41467-020-19157-2

Armstrong, J. S. 1985. Long range forecasting, p. 533–535. *In* From crystal ball to computer. J. Operat. Res. Soc.. doi:10.1057/jors.1986.91.;348

Bathmann, U. V., R. Scharek, C. Klaas, C. D. Dubischar, and V. Smetacek. 1997. Spring development of phytoplankton biomass and composition in major water masses of the Atlantic sector of the Southern Ocean. Deep-Sea Res. II Top. Stud. Oceanogr. **44**: 51–67. doi:10.1016/S0967-0645(96)00063-X

Bidigare, R. R., and M. E. Ondrusek. 1996. Spatial and temporal variability of phytoplankton pigment distributions in the central equatorial Pacific Ocean. Deep-Sea Res. II Top. Stud. Oceanogr. **43**: 809–833. doi:10.1016/0967-0645(96)00019-7

Broglio, E., E. Saiz, A. Calbet, I. Trepat, and M. Alcaraz. 2004. Trophic impact and prey selection by crustacean zooplankton on the microbial communities of an oligotrophic coastal area (NW Mediterranean Sea). Aquat. Microb. Ecol. **35**: 65–78. doi:10.3354/ame035065

Brunet, C., G. Johnsen, J. Lavaud, and S. Roy. 2011. Pigments and photoacclimation processes, p. 445–471. *In* Phytoplankton pigments: Characterization, chemotaxonomy and applications in oceanography. Cambridge Environmental Chemistry Series. doi:10.1017/CBO9780511732263.017

Carter, L., I. N. McCave, and M. J. M. Williams. 2008. Chapter 4 circulation and water masses of the southern ocean: A review, p. 85–114. *In* Developments in earth and environmental sciences. doi:10.1016/s1571-9197(08)00004-9

Comon, P., X. Luciani, and A. L. De Almeida. 2009. Tensor decompositions, alternating least squares and other tales. J. Chemom. Soc. **7-8**: 393–405. doi:10.1002/cem.1236

De Carvalho, C. C., and M. J. Caramujo. 2014. Fatty acids as a tool to understand microbial diversity and their role in food webs of Mediterranean temporary ponds. Molecules **5**: 5570–5598. doi:10.3390/molecules19055570

Deppeler, S. L., and A. T. Davidson. 2017. Southern Ocean phytoplankton in a changing climate. Front. Mar. Sci. **4**: 40. doi:10.3389/fmars.2017.00040

DiTullio, G. R., N. Garcia, S. F. Riseman, and P. N. Sedwick. 2007. Effects of iron concentration on pigment composition in Phaeocystis Antarctica grown at low irradiance, p. 71–78. *In* Phaeocystis, major link in the biogeochemical cycling of climate-relevant elements, v. **83**. Springer. doi:10.1007/s10533-007-9080-8

Domingues, R. B., A. Barbosa, and H. Galvão. 2008. Constraints on the use of phytoplankton as a biological quality element within the Water Framework Directive in Portuguese waters. Mar. Pollut. Bull. **56**: 1389–1395. doi:10.1016/j.marpolbul.2008.05.006

Edler, L., and M. Elbrächter. 2010. The Utermöhl method for quantitative phytoplankton analysis, p. 13–20. *In* Microscopic and molecular methods for quantitative phytoplankton analysis, v. **110**. UNESCO. doi:10.25607/OBP-1371

Estrada, M., M. Delgado, D. Blasco, M. Latasa, A. M. Cabello, V. Benítez-Barrios, E. Fraile-Nuez, P. Mozetič, and M. Vidal. 2016. Phytoplankton across tropical and subtropical regions of the Atlantic, Indian and Pacific oceans. PLoS One **11**: e0151699. doi:10.1371/journal.pone.0151699

Falkowski, P. G. 1994. The role of phytoplankton photosynthesis in global biogeochemical cycles, p. 235–258. *In* Photosynthesis research. Springer. doi:10.1007/BF00014586

Finkel, Z. V., J. Beardall, K. J. Flynn, A. Quigg, T. A. Rees, and J. A. Raven. 2010. Phytoplankton in a changing world: Cell size and elemental stoichiometry. J. Plankton Res. **32**: 119–137. doi:10.1093/plankt/fbp098

Franc, V., V. Hlaváč, and M. Navara. 2005. Sequential coordinate-wise algorithm for the non-negative least squares problem, p. 407–414. *In* International Conference on Computer Analysis of Images and Patterns. Springer. doi:10.0007/11556121_50

Garibotti, I. A., M. Vernet, and M. E. Ferrario. 2005. Annually recurrent phytoplanktonic assemblages during summer in the seasonal ice zone west of the Antarctic Peninsula (Southern Ocean). Deep Sea Res. I Oceanogr. Res. Pap. **52**: 1823–1841. doi:10.1016/j.dsr.2005.05.003

Gast, R. J., D. M. Moran, M. R. Dennett, and D. A. Caron. 2007. Kleptoplasty in an Antarctic dinoflagellate: Caught in evolutionary transition? Environ. Microbiol. **9**: 39–45. doi:10.1111/j.1462-2920.2006.01109.x

Hashihama, F., T. Hirawake, S. Kudoh, J. Kanda, K. Furuya, Y. Yamaguchi, and T. Ishimaru. 2008. Size fraction and class composition of phytoplankton in the Antarctic marginal ice zone along the 140° E meridian during February–March 2003. Pol. Sci. **2**: 109–120. doi:10.1016/j.polar.2008.05.001

Henley, S. F., and others. 2020. Changing biogeochemistry of the Southern Ocean and its ecosystem implications. Front. Mar. Sci. **581**: 7. doi:10.3389/fmars.2020.00581

Henriksen, P., B. Riemann, H. Kaas, H. M. Sørensen, and H. L. Sørensen. 2002. Effects of nutrient-limitation and irradiance on marine phytoplankton pigments. J. Plankton Res. **24**: 835–858. doi:10.1093/plankt/24.9.835

Higgins, H., S. Wright, and L. Schlüter. 2011. Quantitative interpretation of chemotaxonomic pigment data, p. 257–313. *In* Phytoplankton pigments: Characterization, chemotaxonomy and applications in oceanography (Cambridge Environmental Chemistry Series). Cambridge Univ. Press. doi:10.1017/CBO9780511732263.010

Ishikawa, A., S. W. Wright, R. van den Enden, A. T. Davidson, and H. J. Marchant. 2002. Abundance, size structure and community composition of phytoplankton in the Southern Ocean in the austral summer 1999/2000. Pol. Biosci. **15**: 11–26. doi:10.1007/s00300-014-1542-6

Jakobsen, H. H., and J. Carstensen. 2011. FlowCAM: Sizing cells and understanding the impact of size distributions on biovolume of planktonic community structure. Aquat. Microb. Ecol. **65**: 75–87. doi:10.3354/ame01539

Jeffrey, S. W. 1997. Application of pigment methods to oceanography, p. 127–166. *In* S. W. Jeffrey, R. F. C. Mantoura, and S. W. Wright [eds.], Phytoplankton pigments in oceanography: Guidelines to modern methods. UNESCO. doi:10.1023/A:1007168802525

Kang, N. S., and others. 2010. Description of a new planktonic mixotrophic dinoflagellate *Paragymnodinium shiwhaense* n. gen., n. sp. from the coastal waters off western Korea: Morphology, pigments, and ribosomal DNA gene sequence. J. Eukaryot. Microbiol. **57**: 121–144. doi:10.1111/j.1550-7408.2009.00462.x

Karl, D. M., O. Holm-Hansen, G. T. Taylor, G. Tien, and D. F. Bird. 1991. Microbial biomass and productivity in the western Bransfield Strait, Antarctica during the 1986–87 austral summer. Deep Sea Res. A Oceanogr. Res. Pap. **38**: 1029–1055. doi:10.1016/0198-0149(91)90095-W

Kramer, S. J., and D. A. Siegel. 2019. How can phytoplankton pigments be best used to characterize surface ocean phytoplankton groups for ocean color remote sensing algorithms? J. Geophys. Res. Oceans **124**: 7557–7574. doi:10.1029/2019jc015604

Kruk, C., E. T. Peeters, E. H. Van Nes, V. D. Huszar, L. S. Costa, and M. Scheffer. 2011. Phytoplankton community composition can be predicted best in terms of morphological classes. Limnol. Oceanogr. **56**: 110–118. doi:10.4319/lo.2011.56.1.0110

Latasa, M. 2007. Improving estimations of phytoplankton class abundances using CHEMTAX. Mar. Ecol. Prog. Ser. **329**: 13–21. doi:10.3354/meps329013

Lawson, C. L., and R. J. Hanson. 1995. Solving least squares problems. Society for Industrial and Applied Mathematics. SIAM. doi:10.2307/2286501

Letelier, R. M., R. R. Bidigare, D. V. Hebel, M. Ondrusek, C. D. Winn, and D. M. Karl. 1993. Temporal variability of phytoplankton community structure based on pigment analysis. Limnol. Oceanogr. **38**: 1420–1437. doi:10.4319/lo.1993.38.7.1420

Litchman, E., C. A. Klausmeier, J. R. Miller, O. M. Schofield, and P. G. Falkowski. 2006. Multi-nutrient, multi-group model of present and future oceanic phytoplankton communities. Biogeosciences **3**: 585–606. doi:10.5194/bg-3-585-2006

Mackey, M. D., D. J. Mackey, H. W. Higgins, and S. W. Wright. 1996. CHEMTAX-a program for estimating class abundances from chemical markers: Application to HPLC measurements of phytoplankton. Mar. Ecol. Prog. Ser. **144**: 265–283. doi:10.3354/meps144265

McLeod, D. J., G. M. Hallegraeff, G. W. Hosie, and A. J. Richardson. 2012. Climate-driven range expansion of the red-tide dinoflagellate *Noctiluca scintillans* into the Southern Ocean. J. Plankton Res. **34**: 332–337. doi:10.1093/plankt/fbr112

Menden-Deuer, S., and E. J. Lessard. 2000. Carbon to volume relationships for dinoflagellates, diatoms, and other protist plankton. Limnol. Oceanogr. **45**: 569–579. doi:10.4319/lo.2000.45.3.0569

Murtagh, F., and P. Legendre. 2014. Ward's hierarchical agglomerative clustering method: Which algorithms implement Ward's criterion? J. Classif. **3**: 274–295. doi:10.1007/s00357-014-9161-z

Nunes, S., M. Latasa, J. Gasol, and M. Estrada. 2018. Seasonal and interannual variability of phytoplankton community structure in a Mediterranean coastal site. Mar. Ecol. Prog. Ser. **592**: 57–75. doi:10.3354/meps12493

Pan, X., A. Mannino, H. G. Marshall, K. C. Filippino, M. R. Mulholland, and M. R. 2011. Remote sensing of phytoplankton community composition along the northeast coast of the United States. Remote Sens. Environ. **115**: 3731–3747. doi:10.1016/j.rse.2011.09.011

Peeken, I. 1997. Photosynthetic pigment fingerprints as indicators of phytoplankton biomass and development in different water masses of the Southern Ocean during austral spring. Deep-Sea Res. II Top. Stud. Oceanogr. **44**: 261–282. doi:10.1016/S0967-0645(96)00077-X

Peloquin, J., and others. 2013. The MAREDAT global database of high performance liquid chromatography marine pigment measurements. Earth Syst. Sci. Data **5**: 109–123. doi:10.5194/essd-5-109-2013

Pinkerton, M. H., P. W. Boyd, S. Deppeler, A. Hayward, J. Höfer, and S. Moreau. 2021. Evidence for the impact of climate change on primary producers in the Southern Ocean. Front. Ecol. Evol. **9**: 134. doi:10.3389/fevo.2021.592027

Pinkerton, M. H., and A. Hayward. 2021. Estimating variability and long-term change in sea ice primary productivity using a satellite-based light penetration index. J. Mar. Syst. **221**: 103576. doi:10.1016/j.jmarsys.2021.103576

Press, W. H., S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery. 2007. Numerical recipes, 3rd Edition. Cambridge Univ. Press. doi:10.1142/S021896799000199

Punj, G., and D. W. Stewart. 1983. Cluster analysis in marketing research: Review and suggestions for application. J. Market. Res. **20**: 134–148. doi:10.1177/002224378302000204

Quigg, A., and others. 2003. The evolutionary inheritance of elemental stoichiometry in marine phytoplankton. Nature **425**: 291–294. doi:10.1038/nature01953

Racault, M. F., C. Le Quéré, E. Buitenhuis, S. Sathyendranath, and T. Platt. 2012. Phytoplankton phenology in the global ocean. Ecol. Indic. **14**: 152–163. doi:10.1016/j.ecolind.2011.07.010

Rott, E., N. Salmaso, and E. Hoehn. 2007. Quality control of Utermöhl-based phytoplankton counting and biovolume estimates—An easy task or a Gordian knot? Hydrobiologia **578**: 141–146. doi:10.1007/s10750-006-0440-5

Roy, S., C. A. Llewellyn, E. S. Egeland, and G. Johnsen [eds.]. 2011. Phytoplankton pigments: Characterization, chemotaxonomy and applications in oceanography. Cambridge Univ. Press. doi:10.1017/CB09780511732263

Schlüter, L., F. Møhlenberg, H. Havskum, and S. Larsen. 2000. The use of phytoplankton pigments for identifying and quantifying phytoplankton groups in coastal areas: Testing the influence of light and nutrients on pigment/chlorophyll *a* ratios. Mar. Ecol. Prog. Ser. **192**: 49–63. doi:10.3354/meps192049

Schlüter, L., P. Henriksen, T. G. Nielsen, and H. H. Jakobsen. 2011. Phytoplankton composition and biomass across the southern Indian Ocean. Deep-Sea Res. I Oceanogr. Res. Pap. **58**: 546–556. doi:10.1016/j.dsr.2011.02.007

Strandberg, U., S. J. Taipale, M. Hiltunen, A. W. Galloway, M. T. Brett, and P. Kankaala. 2015. Inferring phytoplankton community composition with a fatty acid mixing model. Ecosphere **1**: 1–8. doi:10.1890/ES14-00382.1

Swan, C. M., M. Vogt, N. Gruber, and C. Laufkoetter. 2016. A global seasonal surface ocean climatology of phytoplankton types based on CHEMTAX analysis of HPLC pigments. Deep Sea Res. Part I Oceanogr. Res. Pap. **109**: 137–156. doi:10.1016/j.dsr.2015.12.002

Tangen, K., and T. Björnland. 1981. Observations on pigments and morphology of *Gyrodinium aureolum* Hulburt, a marine dinoflagellate containing 19′-hexanoyloxyfucoxanthin as

the main carotenoid. J. Plankton Res. **3**: 389–401. doi:10.1093/plankt/3.3.389

R Core Team. 2022. R: A language and environment for statistical computing.

Tomić, T., N. Uzorinac-Nasipak, and S. Babić. 2012. Estimating measurement uncertainty in high-performance liquid chromatography methods. Accreditation Qual. Assur. **17**: 291–300. doi:10.1007/s00769-011-0872-0

Uitz, J., H. Claustre, A. Morel, and S. B. Hooker. 2006. Vertical distribution of phytoplankton communities in open ocean: An assessment based on surface chlorophyll. J. Geophys. Res. Oceans **111**: C08005. doi:10.1029/2005JC003207

Vaillancourt, R. D., V. P. Lance, and J. F. Marra. 2018. Phytoplankton chemotaxonomy within contiguous optical layers across the western North Atlantic Ocean and its relationship to environmental parameters. Deep Sea Res. Part I Oceanogr. Res. Pap. **139**: 14–26. doi:10.1016/j.dsr.2018.05.007

Van den Meersche, K., K. Soetaert, and J. J. Middelburg. 2008. A Bayesian compositional estimator for microbial taxonomy based on biomarkers. Limnol. Oceanogr. Methods **6**: 190–199. doi:10.4319/lom.2008.6.190

Wright, S. 2017. Chemtax version 1.95 for calculating the taxonomic composition of phytoplankton populations [Data set]. Australian Antarctic Data Centre. doi:10.4225/15/59FFF1C5EA8FC

Wright, S. W., D. P. Thomas, H. J. Marchant, H. W. Higgins, M. D. Mackey, and D. J. Mackey. 1996. Analysis of phytoplankton of the Australian sector of the Southern Ocean: Comparisons of microscopy and size frequency data with interpretations of pigment HPLC data using the "CHEMTAX" matrix factorisation program. Mar. Ecol. Prog. Ser. **144**: 285–298. doi:10.3354/meps144285

Wright, S. W., and R. L. van den Enden. 2000. Phytoplankton community structure and stocks in the East Antarctic marginal ice zone (BROKE survey, January–March 1996) determined by CHEMTAX analysis of HPLC pigment signatures. Deep-Sea Res. II Top. Stud. Oceanogr. **47**: 2363–2400. doi:10.1016/S0967-0645(00)00029-1

Wright, S. W., and S. W. Jeffrey. 2006. Pigment markers for phytoplankton production, p. 71–104. *In* Marine organic matter: Biomarkers, isotopes and DNA. Springer.

Wright, S. W., R. L. van den Enden, I. Pearce, A. T. Davidson, F. J. Scott, and K. J. Westwood. 2010. Phytoplankton community structure and stocks in the Southern Ocean (30–80 E) determined by CHEMTAX analysis of HPLC pigment signatures. Deep-Sea Res. II Top. Stud. Oceanogr. **57**: 758–778. doi:10.1016/S0967-0645(00)00029-1

Zapata, M., S. W. Jeffrey, S. W. Wright, F. Rodríguez, J. L. Garrido, and L. Clementson. 2004. Photosynthetic pigments in 37 species (65 strains) of Haptophyta: Implications for oceanography and chemotaxonomy. Mar. Ecol. Prog. Ser. **270**: 83–102. doi:10.3354/meps270083

Zarauz, L., and X. Irigoien. 2008. Effects of Lugol's fixation on the size structure of natural nano–microplankton samples, analyzed by means of an automatic counting method. J. Plankton Res. **30**: 1297–1303. doi:10.1093/PLANKT/FBN084

Zwirglmaier, K., L. Jardillier, M. Ostrowski, S. Mazard, L. Garczarek, D. Vaulot, F. Not, R. Massana, O. Ulloa, and D. J. Scanlan. 2008. Global phylogeography of marine *Synechococcus* and *Prochlorococcus* reveals a distinct partitioning of lineages among oceanic biomes. Environ. Microbiol. **10**: 147–161. doi:10.1111/j.1462-2920.2007.01440.x

Acknowledgments

The authors acknowledge NIWA for providing funding through the NIWA PhD scholarship (CDPS2001) and the University of Otago (Department of Marine Science) for their support throughout the PhD program. The authors are also grateful to the New Zealand MBIE Endeavor Program C01X1710 (Ross-RAMP), Antarctic Science Platform, Project 3 (MBIE contract ANTA1801), MBIE NIWA SSIF ("Structure and function of marine ecosystems") for their support to M.H.P. and A.G.-R. The authors thank Professor Cliff Law for advice and insight. The authors thank Dr. Maxime Rio for his thorough insights on data science and numerical methods. The authors acknowledge the contribution to this work by the R open-source collective and R Studio. The authors greatly thank Dr. Simon Wright for his valuable contributions to the field of phytoplankton chemotaxonomy and data made publicly available on phytoplankton pigment ratios. The authors thank Dr. Mikel Latasa for his contributions to the field of pigment research, previous research on synthetic pigment datasets, and past guidance on the treatment of pigment data. The authors thank Dr. Karen Westwood for insights into phytoplankton pigments and providing a forum for discussion on chemotaxonomic methods. The authors thank Dr Andy McKenzie (NIWA, Wellington) for help with R-coding. Open access publishing facilitated by National Institute of Water and Atmospheric Research, as part of the Wiley - National Institute of Water and Atmospheric Research agreement via the Council of Australian University Librarians.

Submitted 11 August 2022
Revised 18 December 2022
Accepted 13 February 2023

Associate editor: David J Suggett