A data-driven algorithm to determine ^{1}H-MRS basis set composition

Christopher W. Davies-Jenkins[1,2], Helge J. Zöllner[1,2], Dunja Simicic[1,2], Seyma Alcicek[3,4,5] Richard A.E. Edden[1,2], and Georg Oeltzschner[1,2,*]

[1] *The Russell H. Morgan Department of Radiology and Radiological Science, Johns Hopkins University School of Medicine, Baltimore, MD, USA*
[2] *F.M. Kirby Research Center for Functional Brain Imaging, Kennedy Krieger Institute, Baltimore, MD, USA*
[3] *Institute of Neuroradiology, University Hospital Frankfurt, Goethe University, Frankfurt/Main, Germany*
[4] *University Cancer Center Frankfurt (UCT), Frankfurt/Main, Germany*
[5] *Frankfurt Cancer Institute (FCI), Frankfurt/Main, Germany*

*Corresponding author:
Georg Oeltzschner, Ph.D.
Division of MR Research, Russell H. Morgan Department of Radiology and Radiological Science
The Johns Hopkins University School of Medicine
600 N Wolfe St, Park 320
Baltimore, MD 21287
goeltzs1@jhmi.edu

Word count: 5000
Figure count: 8

Abstract

Purpose
Metabolite amplitude estimates derived from linear combination modeling of MR spectra depend upon the precise list of constituent metabolite basis functions used (the "basis set"). The absence of clear consensus on the "ideal" composition or objective criteria to determine the suitability of a particular basis set contributes to the poor reproducibility of MRS. In this proof-of-concept study, we demonstrate a novel, data-driven approach for deciding the basis-set composition using Bayesian information criteria (BIC).

Methods
We have developed an algorithm that iteratively adds metabolites to the basis set using iterative modeling, informed by BIC scores. We investigated two quantitative "stopping conditions", referred to as max-BIC and zero-amplitude, and whether to optimize the selection of basis set on a per-spectrum basis or at the group level. The algorithm was tested using two groups of synthetic in-vivo-like spectra representing healthy brain and tumor spectra, respectively, and the derived basis sets (and metabolite amplitude estimates) were compared to the ground truth.

Results
All derived basis sets correctly identified high-concentration metabolites and provided reasonable fits of the spectra. At the single-spectrum level, the two stopping conditions derived the underlying basis set with 77-87% accuracy. When optimizing across a group, basis set determination accuracy improved to 84-92%.

Conclusion
Data-driven determination of the basis set composition is feasible. With refinement, this approach could provide a valuable data-driven way to derive or refine basis sets, reducing the operator bias of MRS analyses, enhancing the objectivity of quantitative analyses, and increasing the clinical viability of MRS.

Keywords: magnetic resonance spectroscopy; basis set; information criteria; model selection; 2HG; Cystathionine

1. Introduction

In-vivo proton magnetic resonance spectroscopy (MRS) is a non-invasive method for estimating the concentrations of approximately 20 metabolites in the human brain. Levels of neurotransmitters and antioxidants, for example, may serve as biomarkers of function and pathology, but also general indicators of neuronal health and cell proliferation. Extracting metabolite concentrations from the in-vivo spectrum depends on reliably resolving the contributions of constituent signals. For this task, expert consensus recommends[1,2] linear-combination modeling (LCM), a well-established method that uses a weighted sum of simulated signals—*basis functions*—to approximate the measured spectrum.

Poor spectral dispersion at clinical field strengths causes overlap between metabolite signals, preventing them from being estimated independently. To achieve optimal precision and accuracy, the basis set included in LCM should be selected without bias. Including too few metabolites results in unmodeled or poorly modeled peaks. Most LCM implementations will attempt to reconcile the model by modulating the baseline and lineshape, or shifting metabolites to reduce the residual. Conversely, including too many basis functions in the model may result in overfitting the data—with noise or lineshape effects being incorrectly attributed to metabolite signals. Many compounds are present below the threshold of detection in the healthy brain, only reaching detectable levels in pathology (e.g., 2-hydroxyglutarate in primary brain tumors with isocitrate dehydrogenase mutations[3]).

In practice, choosing which metabolites to include in the basis set is delicate because *many different* compositions are *admissible*, i.e., are reasonable choices. Several fundamental challenges arise:

1. Wrongly in- or ex-cluding basis functions can lead to substantial biases of[4–8] and interactions between[9,10] metabolite amplitude estimates.
2. There is no consensus on the ideal basis set composition even for the healthy brain, resulting in significant analytic variability between research groups.
3. Defining basis-set composition currently requires a-priori knowledge (e.g., an external brain tumor diagnosis to adequately model a brain tumor spectrum), limiting the clinical application of MRS.
4. The number of metabolites that can be modeled with reasonable accuracy and precision depends strongly on the quality of the data. High-SNR, well-resolved spectra allow the modeling of metabolites that might otherwise have simply led to the overfitting of lower-quality data.
5. Finally, no objective criteria exist to assess whether a specific basis set will overfit, underfit, or appropriately model a spectrum.

To address these challenges, we investigated the feasibility of determining the appropriate basis set composition directly from the spectrum itself using model selection. Model selection is the process of determining the most suitable model from a list of potential candidates informed by a quantitative information criterion (IC). There are several specific definitions of ICs[11–14], but generally, they attempt to "score" candidate models, balancing goodness-of-fit and model complexity in a single value[15,16]. Higher IC scores are interpreted as a proxy measure of model

parsimony, i.e., the "cost-effectiveness" of particular candidate models; ICs have previously been used in MRS to derive optimal baseline model parameters[17].

In this study, we demonstrate an automated data-driven procedure—using iterative fitting of the spectra informed by IC scores—to determine the appropriate composition of the basis set. We developed two variations of this algorithm: the first determines the optimal basis set for a single spectrum, and the second for a group of spectra. We also designed two different stopping conditions: the first stops adding metabolites to the basis set after reaching the maximum IC score, and the second keeps adding metabolites beyond this point until they no longer contribute any signal to the fit. We tested the performance of the algorithm with two classes of in-vivo-like simulated spectra, representing healthy brain and low-grade glioma. The latter included two oncometabolites that were not present in the healthy brain data. We quantified how well the basis sets selected by the different algorithm implementations overlapped with the ground-truth basis sets from which spectra were generated.

2. Materials & methods

2.1. Synthetic in-vivo-like data generation

We first simulated realistic in-vivo-like spectra to establish a known ground truth. We generated two datasets of 100 spectra, each with distinct spectral characteristics: healthy-appearing brain spectra and those typically seen in low-grade glioma, respectively. We chose the latter case as these spectra commonly feature metabolites that are effectively absent from healthy brain tissue (cystathionine[18,19] and 2-hydroxyglutarate[20]) and exhibit markedly different amplitudes of major metabolites.

Synthetic spectra were derived from a "library set" of metabolite basis functions, simulated using the density-matrix formalism of a 2D-localized 101 x 101 spatial grid (field of view 50% larger than the voxel) implemented in MRSCloud[21], which is based on FID-A[22]. We synthesized 3T sLASER spectra to reflect a typical acquisition used to measure 2HG (TE = 97 ms with TE1/2 = 32/65 ms; 8192 complex points (downsampled to 2048 for the synthetic data); spectral width 4000 Hz).

We simulated 32 metabolite basis functions: Acetoacetate, AcAc; acetate, Ace; alanine, Ala; ascorbate, Asc; aspartate, Asp; citrate, Cit; creatine, Cr; creatine methylene, CrCH2; cystathionine, Cystat; ethanolamine, EA; ethanol, EtOH; γ-aminobutyric acid, GABA; Glycerophosphocholine, GPC; glutathione, GSH; glucose, Glc; glutamine, Gln; glutamate, Glu; glycine, Gly; myo-inositol, mI; lactate, Lac; N-acetylaspartate, NAA; N-acetylaspartylglutamate, NAAG; phosphocholine, PCh; phosphocreatine, PCr; phosphoethanolamine, PE; phenylalanine, Phenyl; scyllo-inositol, sI; serine, Ser; taurine, Tau; tyrosine, Tyros; β-hydroxybutyrate, bHB; and 2-hydroxyglutarate, 2HG. We also included basis functions for 5 macromolecular (MM09, MM12, MM14, MM17, MM20) and 3 lipid resonances (Lip09, Lip13, Lip20). Out of the full set of 40 basis functions, we selected 26 to assemble the synthetic healthy ground-truth spectra: Asc; Asp; Cr; CrCH2; GABA; GPC; GSH; Gln; Glu; mI; Lac; NAA; NAAG; PCh; PCr; PE; sI; Tau; MM09; MM12; MM14; MM17; MM20; Lip09; Lip13; and Lip20. The locations, widths, and

relative amplitudes of the parameterized lipid and MM resonances are reported in **Supplementary Table 1** and the full library is visualized in **Supplementary Figure 1**.

For the tumor spectra, we added additional contributions from 2HG and Cystat for a total of 28 basis functions in the tumor ground-truth basis set. The Osprey function "OspreyGenerateSpectra"[23] was used to combine the individual simulated profiles with metabolite-specific amplitudes, Gaussian and Lorentzian linebroadening terms, and white noise, taken from Gaussian distributions. Crucially, each group was defined with distinct model parameter ranges, which were informed by previously modeled in-vivo data[7,24]. Briefly, besides the additional 2HG and Cystat contributions, the tumor spectra also had higher tCho (100%), lower tNAA (60%), higher mI (30%), lower Glu (25%), higher Lac (1000%), higher lip09 (550%), and higher Lip20 (550%). The means and standard deviations of the parameters for each group are fully reported in **Supplementary Table 2** and the spectra are visualized in **Supplementary Figure 2.**

2.2. Basis set determination algorithm

Our proposed algorithm uses IC scores to build the basis set in an iterative process. Introducing more basis functions reduces model residuals but will add model parameters (i.e. more metabolite-specific amplitudes, frequency shifts, and lineshape parameters). Determining the IC scores for each potential basis set composition allows us to counterbalance these two competing aspects to arrive at a parsimonious compromise, i.e., maximize the goodness-of-fit without overfitting.

2.2.1. Information criteria

The Akaike IC (AIC), corrected AIC (AICc), and Bayesian IC (BIC), only differ in how they regularize goodness-of-fit against model complexity and are linearly offset from one another for our purposes (**Supplementary Figure 3**). We elected to proceed using the BIC as the sole model performance metric in the algorithm, defined:

$$BIC = -2 \cdot n \cdot ln(\sigma) - ln(n) \cdot p$$

where n is the number of points, p is the total number of model parameters, and $\sigma = \sqrt[2]{\sum(residuals^2)}$, the root sum square of the fit residual.

2.2.2. Algorithm implementation

We designed an algorithm to determine which of the library set of 40 basis functions (32 metabolites plus 8 MM/Lips) merits inclusion in the "selected set". The selected set (which is initially empty) is built up one basis function at a time—the function added in each round of the algorithm is chosen from among the remaining candidates in the library set, based upon which model had the highest BIC. Thus, the algorithm consists of two nested loops: the outer loop (of rounds) to determine which is the next candidate to add to the selected set, and the inner loop (of steps) which consists of modeling the data with the current selected set plus one additional candidate.

Both loops fill the selected basis set iteratively until a certain stopping condition is triggered. The algorithm is illustrated in **Figure 1A**, and a **Supplementary Video**, with the specific steps as follows:

1. Initialize an empty selected set and a full library set.
2. Perform a preliminary fit using the full library set to estimate global lineshape and phase parameters which are then fixed during subsequent model calls.
3. Round 1:
    a. Step 1 consists of modeling the spectrum with a candidate set containing only the first library basis function.
    b. The remaining 39 steps of Round 1 consist of modeling the spectrum with a candidate set consisting of each library basis function. The BIC is calculated for each model.
    c. Round 1 concludes by moving the candidate function that resulted in the model with the highest BIC into the selected basis set.
4. The N-th Round then steps over the $(41 - N)$ remaining library functions, thus:
    a. Form the candidate basis set from the selected set and the nth basis function from the remaining library set.
    b. Model the data with the candidate set.
    c. Calculate the BIC and note the amplitude of the nth basis function within the model.
5. Check whether the stopping condition has been met (as described below).
6. If the stopping condition is not met, move the candidate basis function with the highest BIC score from the library set to the selected set and return to step 4.

We investigated two domains of selection, i.e. whether to select a basis set for each spectrum individually or to select one basis set at the group level for a set of spectra. For each domain, we also investigated two stopping conditions, i.e., when to stop adding basis functions to the selected set, see **Figure 1B**. The optimization domains are referred to as:

1. **"Single-spectrum"**: Executes the algorithm on each spectrum in isolation. Candidate basis functions are added based on the BIC scores derived from that spectrum alone, resulting in a spectrum-specific selected set.
2. **"Group-level":** Executes the algorithm as described, but aggregates BIC scores across a *group* of spectra. At each step, we fit *all* spectra within a group with the candidate set and then use the median BIC score across the group to decide which metabolite is selected. This procedure results in a single basis set for the entire group.

The two stopping conditions were:

1. **"Max-BIC":** The algorithm stops adding metabolites to the selected set once the BIC is decreased by the inclusion of the next basis function. For the group-level optimization, it stops once the *median* BIC across all datasets decreases.
2. **"Zero amplitude":** The algorithm keeps adding metabolites beyond the maximum BIC until the round when all remaining candidate functions are modeled with zero amplitude. For the group-level optimization, it stops once all potential candidate functions are estimated with a *median* amplitude of zero across all datasets.

To stabilize the procedure, we instructed the algorithm to "link" certain pairs of practically indistinguishable basis functions, i.e., add both of them to the temporary candidate basis set as a single step, and, if selected, to add both to the selected basis set. Specifically, we linked Cr and PCr (referred to as the tCr step) and PCho and GPC (the tCho step).

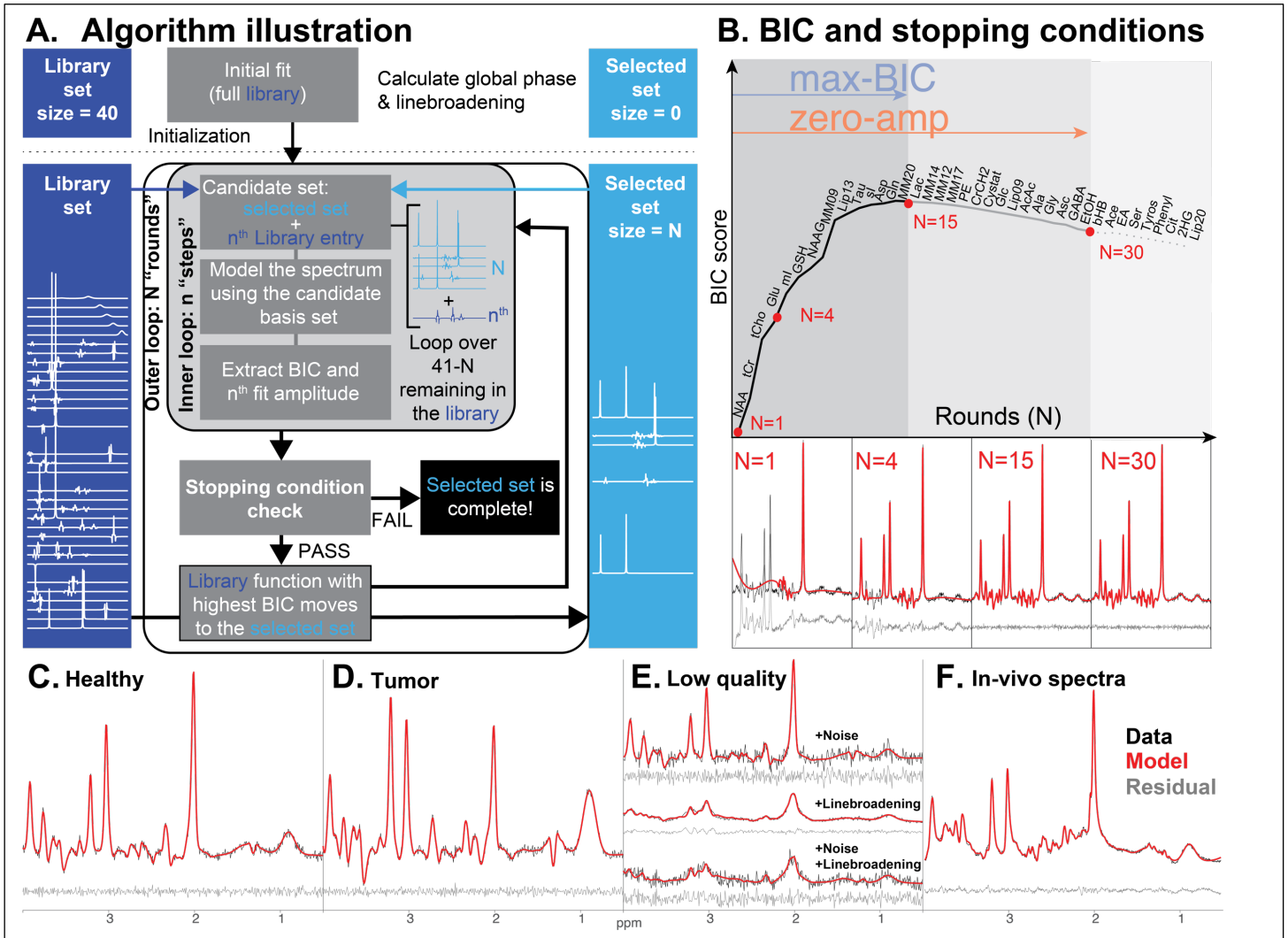

Figure 1. Basis set selection algorithm. **A.** A visualization of the algorithm. Metabolites are iteratively moved from a library set (dark blue) to a selected set (light blue) using BIC scores. The basis function that yields the best BIC score is moved to the selected set (unless a stopping condition is triggered). **B.** A visualization of the evolution of BIC scores as the basis set is optimized. The two stopping conditions are represented by the colored arrows and shaded regions. 4 example fits are shown in the panels below (corresponding to red points on the curve). **C–F.** Panels show representative spectra (black) and a corresponding model (red) derived from single-spectrum optimization and the zero-amplitude stopping condition. Model residuals are shown below (grey). **C.** A synthetic healthy-appearing spectrum **D.** A synthetic low-grade glioma spectrum **E.** Examples of synthetic lower-quality spectra. **F.** A representative real in-vivo spectrum.

2.2.3. MRS modeling

All MRS modeling steps were performed using Osprey's[24] generalized LCM framework, introduced in our recent work[23,25]. Aside from the Gaussian and zero-order phase parameters—which were fixed following the initial modeling step—the model was defined using default in-vivo parameters, naïve to the composition of the simulated spectra. We performed optimization over the range of 0.5–4.2 ppm and included a non-regularized spline baseline with knot spacing of 0.5 ppm. Metabolite-specific Lorentzian linewidth and frequency shift parameters were regularized (as described in the original LCModel publication[26]) with expectation and standard deviation values of 2.75 ± 1.5 and 0 ± 3 Hz, respectively.

2.3. Statistical analysis

The algorithm's performance was primarily judged by its ability to correctly identify the metabolite signals that were present in the synthetic spectra. This was assessed using three related metrics to compare each derived basis set to the ground truth set:

1. False positives: The number of basis functions selected for inclusion by the algorithm that were not present in the synthetic spectra (incorrectly included).
2. False negatives: The number of basis functions not selected for inclusion by the algorithm that were present in the synthetic spectra (incorrectly omitted).
3. The Sorensen-Dice coefficient (SDC): The SDC was used as a measure of overlap with the ground truth basis set, defined:
   $SDC = \frac{2 \cdot \text{TruePositive}}{2 \cdot \text{TruePositive} + \text{FalsePositive} + \text{FalseNegative}}$.

Finally, we investigated the effect that the derived basis set composition had on metabolite amplitude estimates. Amplitude deviations from the ground truth are reported as percentage changes relative to the known ground-truth value that entered the simulation. We also compared the amplitudes derived from each basis set to those derived from a fit using the *correct* ground truth basis set, i.e. modeling the data with only the metabolites we know to be present in the simulation. Distributions were compared using paired t-tests. All statistical analyses were conducted in Matlab 2022a.

2.4. Robustness to low SNR and high linewidth

To examine how our algorithm performed when spectral quality was reduced, we conducted a further analysis of the synthetic spectra. We augmented the healthy-brain dataset by introducing varying degrees of additional noise and Lorentzian linebroadening for a total of 9 data quality scenarios, each containing 100 spectra. These datasets were analyzed as described above, and the impact of linewidth and SNR was explored. Full implementation details are provided in the supplementary material.

2.5. Application to in-vivo data

To test how our algorithm performed in real spectra, we applied it to a freely available repository of 101 short-TE brain spectra[27]. In-vivo data were processed in Osprey[24], and a new appropriately-simulated library set was generated[21] which contained the same 40 metabolites as

the synthetic data analysis. We ran all procedural variations of our algorithm to derive basis sets for these spectra. Full implementation details are provided in the supplementary material.

The algorithm implementation, specific Osprey version, and other code used to generate the results of this manuscript are available online (DOI: 10.17605/OSF.IO/P2USJ).

3. Results

3.1. Algorithmic trends of BIC scores

As the algorithm added basis functions to the selected set, BIC scores followed a characteristic trajectory. High-concentration singlet resonances were initially prioritized by the algorithm, rapidly reducing model residuals and, consequently, increasing BIC scores. After this—as residual contributions reduced—the ascent of the BIC scores tapered off to a maximum (triggering the max-BIC stopping condition). According to the interpretation of BIC scores, this model is optimally parsimonious, and while additional basis functions *do* continue to reduce the residuals, they do not reduce them sufficiently to warrant the additional model parameters. This ascent of the BIC scores reliably identified the high-concentration metabolites that were present in the simulation.

Beyond the peak, BIC scores decreased as relatively smaller spectral contributions lowered the model's cost-effectiveness. Eventually, the candidate metabolites contribute an amplitude of zero when added to the model. In other words, it costs model parameters but provides no benefit to the residual (triggering the 2nd stopping condition: zero-amplitude). After this point, the remaining basis functions bring a continued decline. Importantly, ground-truth metabolites (those present in the synthetic spectra) tended to be included before the algorithm reached this final zone. Typical BIC score trajectories are illustrated in **Figure 2**.

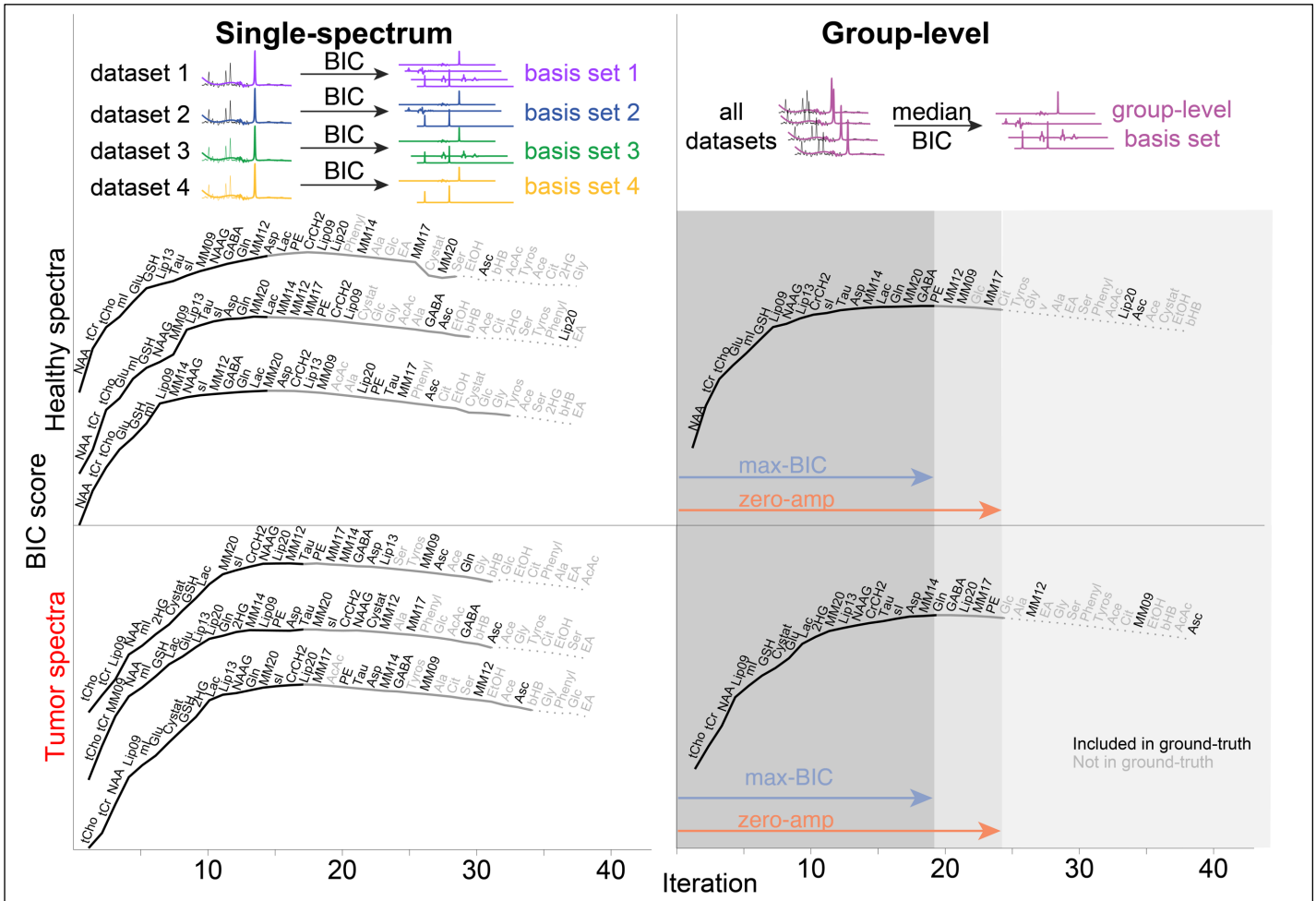

Figure 2. BIC scores as a function of algorithm round. The top row shows a conceptual visualization of the difference between the two optimization methods. The lower two rows show the results for the healthy spectra (middle row) and tumor spectra (bottom row). The left column shows 3 (out of 100) example BIC curves for the single-subject optimizations, and the right column shows the group-level optimization (median BIC). Text labels indicate the basis function added at each round, with label color reflecting the presence/absence (black/grey) of that metabolite in the ground-truth simulation. Line style is used to illustrate the extent of the derived basis sets in each stopping condition: included in both stopping conditions (solid black line), included only in the zero-amplitude condition (grey line), or included in neither (light-grey dotted line).

Figure 2 illustrates the between-spectrum variation of the basis set composition, and we have attempted to summarize the algorithmic trends of the single-spectrum optimization in **Figure 3** and **Supplementary Figure 4**. For healthy spectra, there is a clear prioritization of NAA, tCr, and tCho, in that order. The tumor spectra have a less clearly defined hierarchy of signals. The lower NAA and higher tCho amplitudes lead to a less unanimous prioritization of the singlets,

and the larger lipid signals are chosen as early as the 2nd round. Beyond the 5th round, both groups of spectra had a diverse set of metabolites selected.

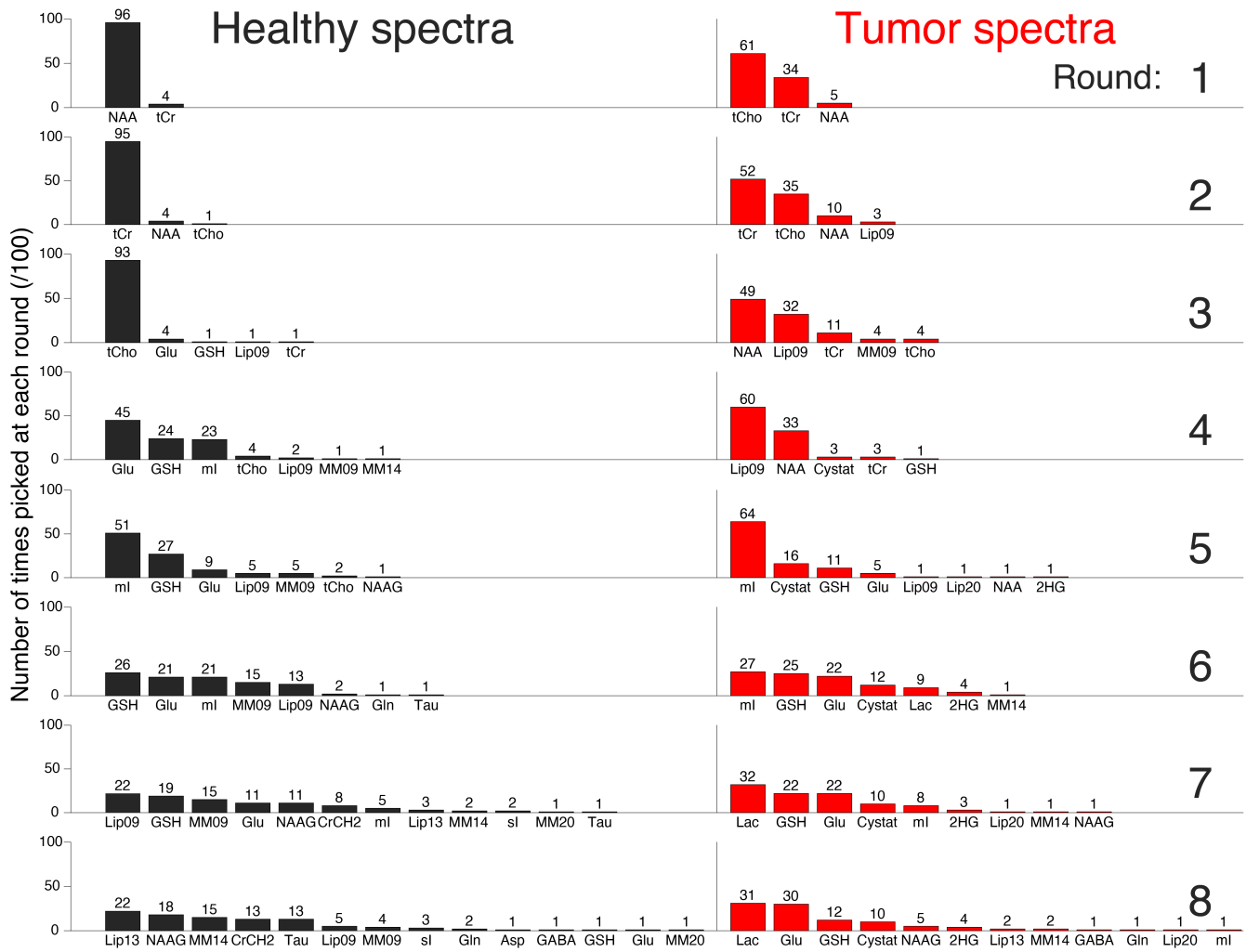

Figure 3. Histograms of the frequency with which particular metabolites were picked at each round of the algorithm. Rows represent the different rounds (increasing from top-to-bottom), and the two columns represent the two groups of spectra.

3.2. Overlap with the ground-truth basis set

For the single-spectrum optimization, the max-BIC stopping condition tended to be too conservative (few false positives, many false negatives) whereas the zero-amplitude condition tended to be too aggressive (few false negatives, many false positives). Bar plots showing the false positives, false negatives, and SDC are shown in **Figure 4**. Overall, the zero-amplitude condition exhibited a better overlap with the ground-truth set, as evidenced by the SDC for both healthy-appearing spectra (mean difference = 7%, $p < 0.005$) and tumor spectra (mean difference = 8%, $p < 0.005$).

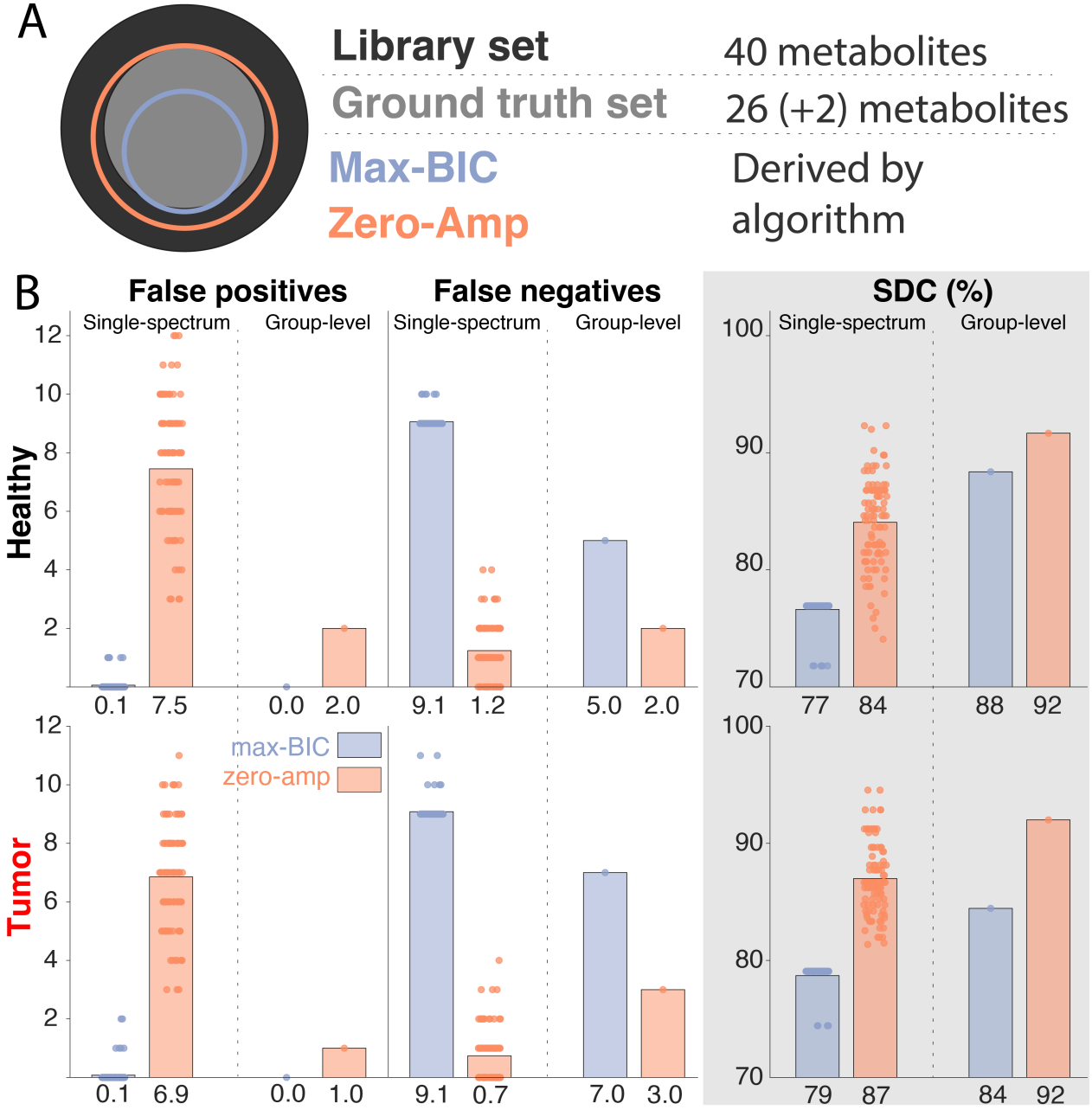

Figure 4. A. Conceptual Venn diagram showing the ground truth, library, and both derived basis sets. **B.** Overlap of the derived basis sets compared to the ground truth for the healthy spectra (top) and tumor spectra (bottom). From left to right, columns show the number of false positives, the number of false negatives, and the Sorensen-Dice coefficient (SDC) reported as a percentage. Within each panel, the left half shows the 100 single-spectrum results, and the right half shows the sole result of the group-wise optimization. Stopping conditions are color-coded: Max-BIC (purple) and zero-Amp (orange). Below each bar, the mean value is reported.

A similar trend was present for group-wise optimization but with greater fidelity to the ground truth. Across both datasets and both stopping conditions, the group-level optimization outperformed the single-spectrum optimization in all metrics with one exception (zero-amplitude stopping condition, false negatives for the tumor spectra). This better performance of the group-level approach is summarized in the SDC plots. The ground-truth overlap increased by 11% for the max-BIC (from 77% to 88%) and 8% for the zero-amplitude stopping conditions (from 84% to 92%) in the healthy spectra, and both increased by 5% for the tumor spectra. However, when linewidth and noise were increased, the single-spectrum optimization was more robust, maintaining >73% overlap with the ground truth basis set for even the worst-quality data (**Figure 5**).

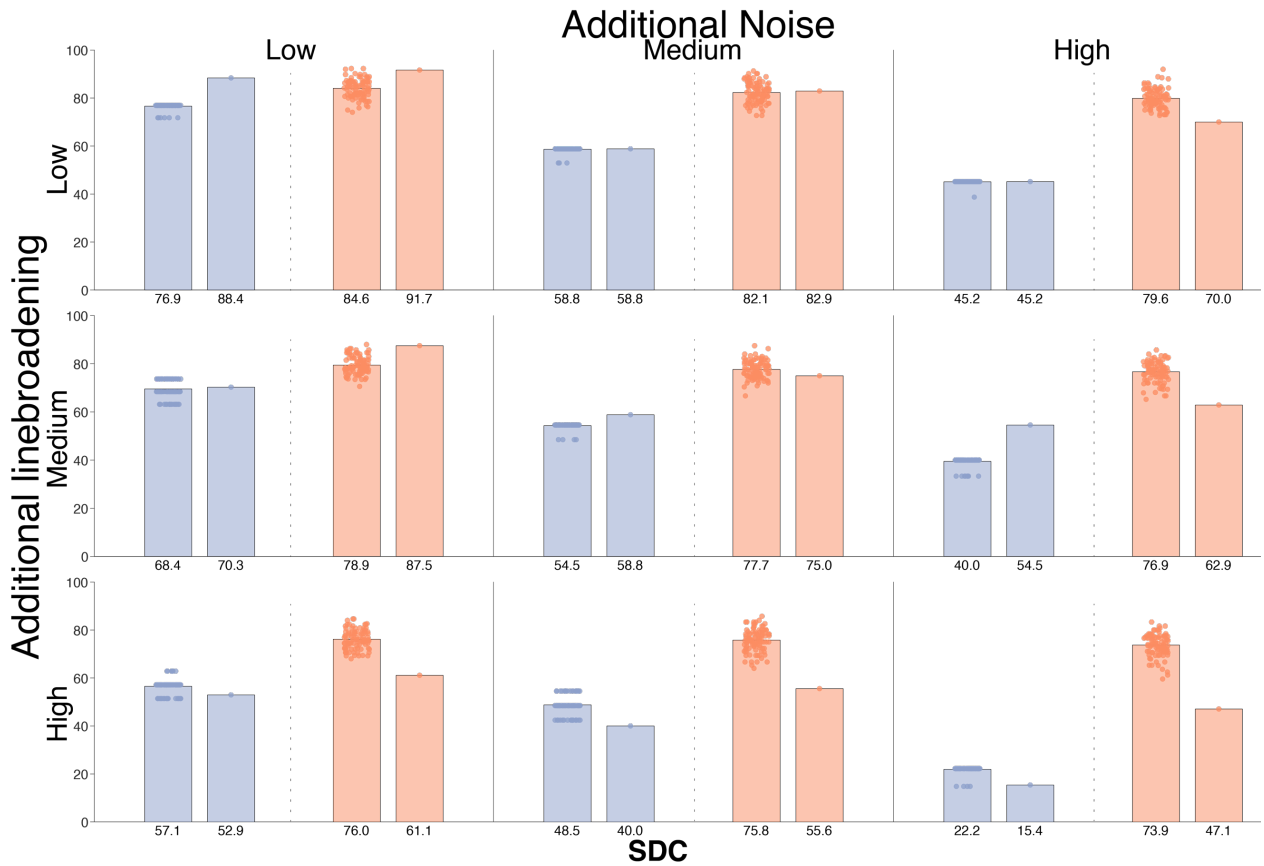

Figure 5. Sorenson-Dice coefficient (SDC; percentage) for each of the 9 data-quality scenarios. Bars are colored by stopping condition—Max-BIC (purple) and zero-Amp (orange)—with the single-spectrum optimization on the left (multiple data points) and group-wise optimization on the right (single data point). Labels on the x-axes indicate the median of the distributions.

3.3. Inclusion of 2HG and Cystat

For the single-spectrum optimization, the inclusion or omission of Cystat and 2HG depended on the stopping condition used, with neither providing perfect results. The zero-amplitude condition

included the oncometabolites in 97% of the tumor spectra (n_{Cystat} = 100; n_{2HG} = 94) but with many false positives for these same metabolites in healthy spectra (n_{Cystat} = 54; n_{2HG} = 42). The situation is somewhat reversed for the max-BIC stopping condition. It does better at omitting Cystat and 2HG from healthy spectra (n_{Cystat} = 0; n_{2HG} = 2) but includes them for fewer (87%) tumor spectra than zero-amplitude (n_{Cystat} = 89; n_{2HG} = 85). Notably, all of the excluded "false negatives" had very low ground-truth amplitude, suggesting a (fairly low) amplitude threshold above which the algorithm almost always correctly included Cystat and 2HG (**Figure 6B**).

For group-level optimization, the results are encouraging and definitive. Cystat and 2HG were correctly omitted from the basis set of the healthy spectra and correctly included in the tumor spectra for both stopping conditions.

3.4. Effect on metabolite amplitudes

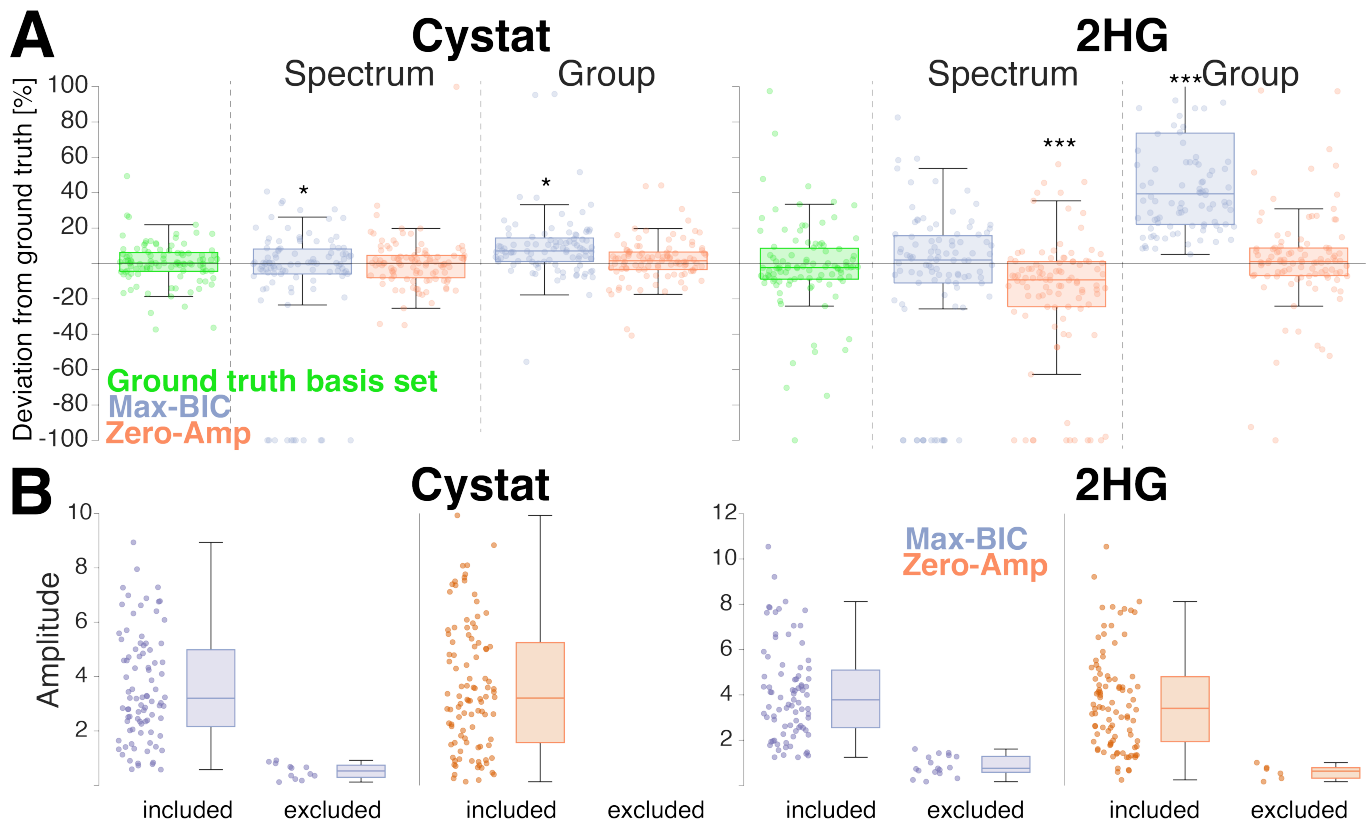

Figure 6. A. Boxplots of the percentage deviation of amplitude estimates from the simulated ground truth for Cystat and 2HG in the tumor spectra. The results using the ground truth basis set (i.e. including only the metabolites present in the simulation) are shown in green. The two stopping conditions—max-BIC (purple) and zero-Amp (orange)—are shown in each panel, for the single-spectrum optimization (left) and group-level optimization (right). Significant deviations from the estimates of the ground-truth basis set are marked with asterisks: "*" ($0.005 \leq p < 0.05$), "**" ($0.0005 \leq p < 0.005$), and "***" ($p < 0.0005$). Note that an estimated amplitude of zero is assumed for single-spectrum-optimized basis sets which omitted that particular metabolite. This resulted in a cluster of data points at –100%. However, 4/10 of these points for the single-spectrum, zero-amplitude basis sets were true zero-amplitude estimates of 2HG when it was present in the derived basis set.
B. Boxplots showing the ground-truth amplitude distributions of Cystat (left panel) and 2HG (right panel) for the synthetic tumor spectra. Boxplots are separated by the respective inclusion/exclusion (left/right) of the Cystat or 2HG in the derived basis sets derived for single-spectrum optimization with the zero-amplitude (orange) and max-BIC (purple) stopping conditions.

Metabolite amplitude errors depended on the ground truth amplitude, with larger relative error for low-concentration metabolites. **Figure 7** shows the percentage deviation of metabolite amplitude estimates from the ground truth for 5 metabolite measures commonly reported in the

MRS literature: tNAA (NAA + NAAG), tCr (Cr + PCr), tCho (GPC + PCho), mI, and Glx (Glu + Gln).

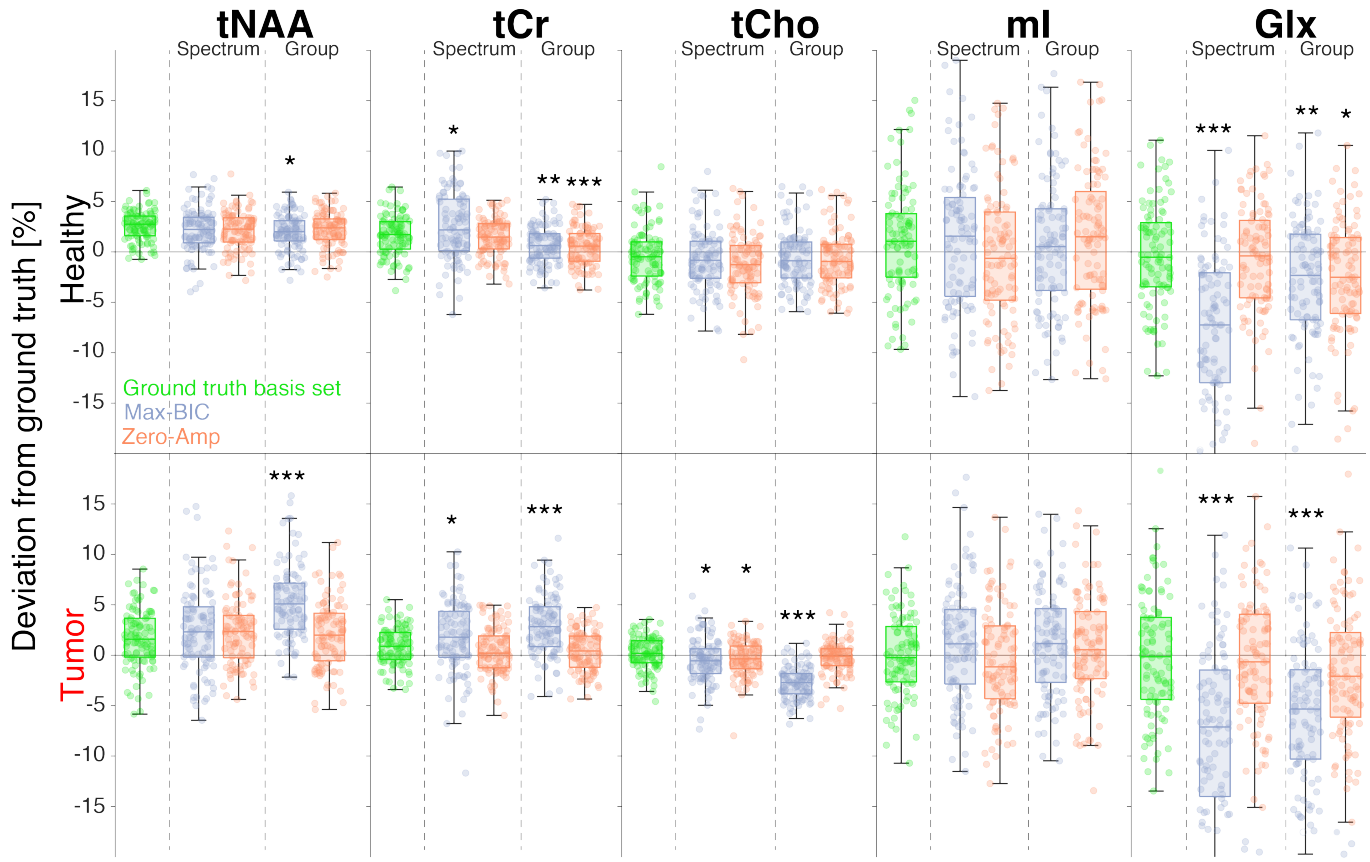

Figure 7. Boxplots of the percentage deviation of amplitude estimates from the simulated ground truth for 5 commonly reported metabolites (tNAA, tCr, tCho, mI, and Glx). Results for the healthy spectra are shown in the top row and tumor spectra in the bottom. The results using the ground truth basis set (i.e. including only the metabolites present in the simulation) are shown in green. The two stopping conditions—max-BIC (purple) and zero-Amp (orange)—are shown in each panel, for the single-spectrum optimization (left) and group-level optimization (right). Significant deviations from the estimates derived using the ground-truth basis set are marked with asterisks: "*" ($0.005 \leq p < 0.05$), "**" ($0.0005 \leq p < 0.005$), and "***" ($p < 0.0005$). Note that the plot limits exclude Glx outliers for comparability.

Absolute amplitude errors were typically below 15% for these 5 metabolites, with some negative Glx outliers as large as –37%, namely for max-BIC-derived basis sets. For the major singlet resonances (tNAA, tCr, and tCho), absolute amplitude errors were < 16% overall, and ≤ 11 % for the healthy-appearing spectra, specifically.

The sparser max-BIC basis sets produced larger mean amplitude errors (4.6%, single-spectrum; 3.9%, group-wise) than the zero-amplitude stopping condition (3.2%, single-spectrum & group-wise). The best-performing optimization method (as measured by SDC; group-wise & zero-amplitude) only deviated from the ground truth amplitude estimates in healthy spectra for Glx

(mean difference = 2.1%, $p = 0.0071$) and tCr (mean difference = 1.2%, $p < 0.0005$). When spectral quality was reduced, amplitude estimates were robust to noise, but significant deviations were observed with increased linewidth, even for the correct basis set (**Supplementary Figure 10**).

The amplitude results for Cystat and 2HG are shown in **Figure 6A** for the tumor spectra. Except for the group-based max-BIC condition, all conditions achieved median ground-truth errors of less than 10% for both 2HG and Cystat, with small (but significant) deviations from the ground-truth basis set estimates. The group-based zero-amp was the only condition that estimated 2HG and Cystat without significant differences to the ground-truth basis set.

3.5. In-vivo data

Both single-spectrum and group-based flavors of our algorithm selected most metabolites that are commonly encountered in in-vivo basis sets (**Figure 8**). Notable omissions from the group-wise-optimized basis set were MM09, MM12, and Lip20, likely because they are sufficiently modeled by similar-appearing broad signals (Lip09, Lip13, MM20). Similarly, EtOH and Cit were surprisingly included, likely because they can resemble aspartyl signals around 2.5 ppm (Cit) and, if sufficiently broadened, MM/Lip signals around 1.2 ppm (EtOH).

4. Discussion

A key challenge for in-vivo MRS is that a 'true' model of the spectrum cannot be known, since the processes that shape it—the interplay of concentrations, microstructure, microscopic particle motion, macroscopic subject motion, temperature, pH, etc.—are too complex. All conventional ^{1}H-MRS modeling methods therefore focus on the most relevant macroscopic quantities—i.e., signal amplitudes for key metabolites—and simplify other aspects. This epistemic uncertainty has provided a fertile environment in which a multitude of model functions, algorithms, software tools, and fitting strategies have grown over the last thirty years. Unsurprisingly, the different models and strategies do not agree very well with each other[28–30], contributing to the poor reproducibility and comparability of metabolite estimates in the MRS literature[31].

"If we have several options to model our data, which should we pick?" is a common question across the sciences. If there are multiple suitable ways to model in-vivo MRS data, we posit that modeling procedures should not only traverse the parameter space *within* a single model definition (via least-squares optimization) but also search *across* model spaces, i.e., explore the breadth of reasonably admissible models. Quantitative model selection methods provide a formal, rigorous framework to do this—exemplified by the ABfit algorithm, which selects from many admissible baseline spline models, affording just enough model flexibility to approximate the data, but not more[17].

The composition of the basis set is another prime example of the range of reasonably admissible models, having been at the sole discretion of individual researchers since the inception of LCM for in-vivo MRS. The surprisingly limited number of studies that have investigated this modeling decision highlights the potential for operator bias, with changes to the basis set composition

substantially affecting metabolite estimates[4–10]. In this proof-of-concept study, we therefore demonstrate the use of model selection for determining basis set composition directly from the spectra.

Across all procedural variations, the algorithm derived reasonable basis sets and consistently produced fits with flat residuals. High-concentration metabolites were correctly included in the algorithm-derived basis sets, without exception. The algorithm's ability to correctly identify lower-concentration metabolites depended on the precise procedural variation. We found that group-level basis set estimation with the 'zero-amplitude' stopping condition minimized the number of false positives and negatives, correctly recognized 2HG and Cystat in the tumor group, and provided metabolite estimates close to those estimated with the ground-truth basis set. This is an encouraging starting point, although further investigation is required.

4.1. Choice of Stopping condition

We initially designed the algorithm to select the basis set that simply maximized the BIC—certainly the most intuitive choice, as it is purportedly the optimal balance of model complexity and fidelity. In practice, the "max-BIC" condition was too conservative: the least-squares model was flexible enough to mimic the signal of metabolites (particularly J-coupled, low-concentration ones) through some combination of baseline bumps, lineshape distortions, and overlapping metabolites already present in the basis set. This—according to raw BIC scores—is preferable to including another basis function, which adds model parameters.

Adding further metabolites (beyond the maximum BIC) until all remaining basis functions are fit with zero amplitude alleviated these problems, albeit at the expense of including more false positives. Despite this, quantitative overlap with the ground truth and amplitude estimation was still better than for the max-BIC stopping condition. Together, the two stopping conditions delineated a "Goldilocks" region that reliably bounded the ground truth basis set.

4.2. Group-wise or single-spectrum

Group-wise optimization improved the identification of lower-concentration metabolites. The aggregation of BIC scores across many spectra reduced the impact of noise and individual amplitude variations, allowing the algorithm to better prioritize signals that are inconsistently modeled in a single spectrum. However, single-spectrum optimization performed better for low-SNR and high-linewidth spectra, suggesting this benefit of group-wise optimization is limited to consistently high-quality datasets (**Figure 5** & **Supplementary Figures 5–10**). Group-wise optimization also removes our ability to adapt to individual outlier spectra that may be present in heterogeneous in-vivo samples. To overcome this limitation, 'soft inclusion' methods (machine learning or traditional regularization strategies) may harness the probability of the presence of metabolites from the full group and then include specific metabolites in single-spectrum fits if sufficiently justified by the data, rather than 'hard exclusion' of a metabolite (see Section 4.4 *Perspectives*).

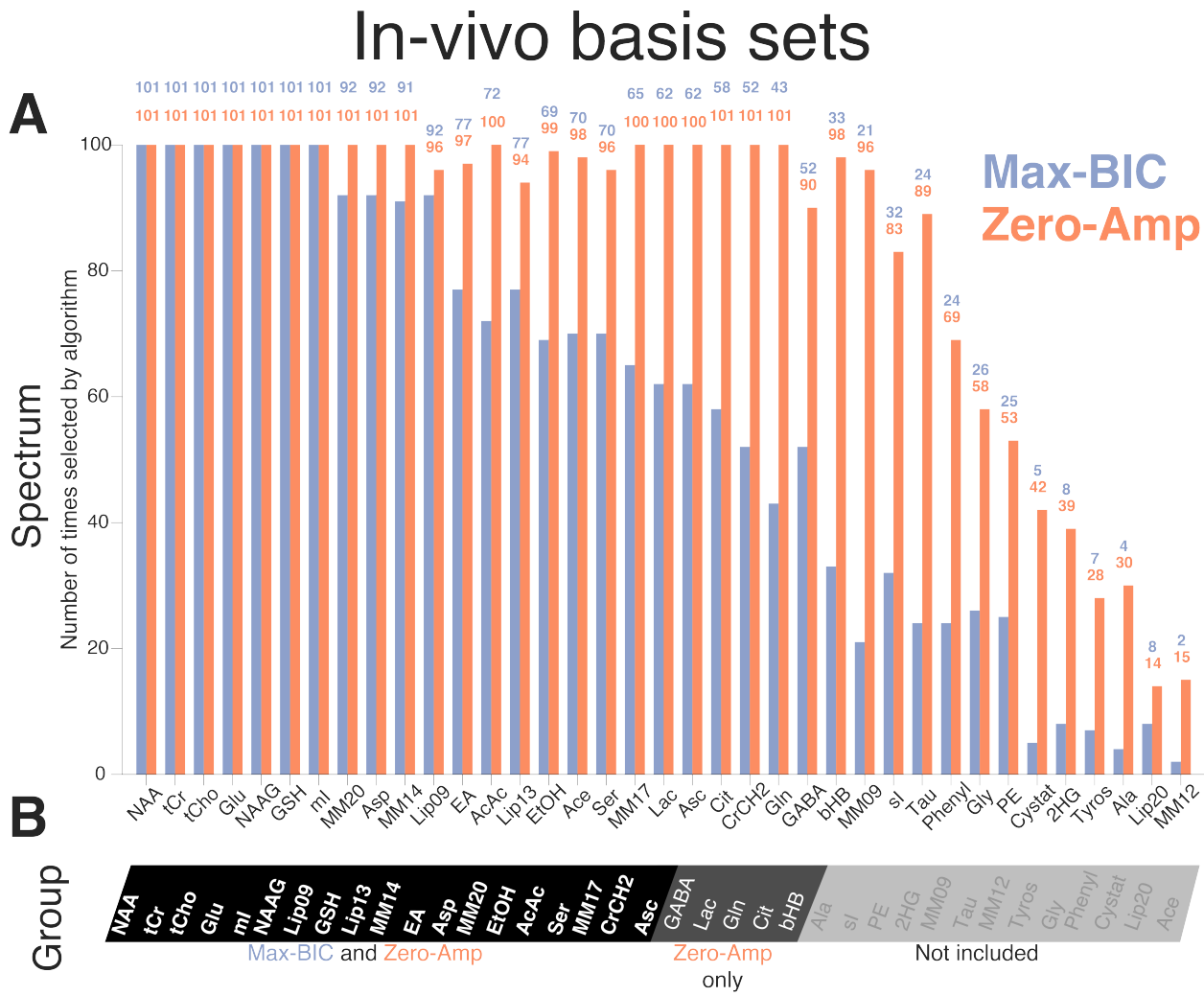

Figure 8. Basis set composition derived from short-TE in-vivo data. A. Bar chart showing the number of times each metabolite was selected for inclusion in the single-spectrum-optimized basis sets for the in-vivo data. Colors represent the two stopping conditions. Metabolites are listed in descending order according to the total number of inclusions across both stopping conditions, i.e., more likely included (left), less likely (right).
B. A representation of the metabolites selected from the same data as above using group-wise optimization. The black panel indicates metabolites included in both stopping conditions, the dark grey panel shows metabolites included only in the zero-amplitude stopping condition, and the light grey panel indicates metabolites excluded by both stopping conditions.

Future work could incorporate additional information into the decision-making process to tackle issues of metabolite overlap. These may include Cramér-Rao lower bounds or metabolite amplitude correlation matrices derived from the Fisher information matrix. or it may be beneficial to modify the IC itself, changing the relative weighting of the terms representing the model residual and number of parameters. Optimization of this balance might improve the max-IC stopping condition, but further work is needed to explore this.

The correct inclusion of 2HG and Cystat in the tumor cohort (and simultaneous exclusion in the healthy cohort) illustrates that the selection process might even deem two different models to be most appropriate if two cohorts exhibit markedly different spectral characteristics. This is, in general, of course, desirable! For example, healthy tissue does not produce detectable levels of 2HG[32(p2)], and model selection for a set of healthy spectra *should* select a model that does not contain a 2HG basis function because fewer degrees of freedom will reduce the variance in estimating signals like Glu/Gln that overlap with 2HG. However, the interaction between these signals is acquisition-dependent, and so the inclusion of 2HG—even when it is present—will depend on the degree to which we can independently model its signal.

4.3. Limitations

Our synthetic data omitted some common characteristics of in-vivo MRS data. For example, we did not introduce phase alterations, frequency shifts, correlated noise, or baseline components. Broad resonances underlying the metabolite signals, in particular, significantly impact the accuracy of MRS modeling[8,29]. Baseline modeling would be further complicated by the frequent omission of MM and lipid basis functions. The algorithm tended to capture some of these—admittedly small—signals using the baseline, reducing the number of model parameters, but compromising the validity of the baseline model. This is a particular concern for short-TE spectra—such as our in-vivo data—where larger lipid and MM contributions are expected. Future work could include some algorithm penalty for baseline amplitudes or experimentally acquired MMs, as recommended by expert consensus[33] and our previous investigations[29,34].

Furthermore, careful considerations need to be made for signal artifacts. Out-of-voxel "ghost" echoes introduce unpredictable signals that cannot be accurately captured by the basis set. Large, broad signals might also be introduced by extraneous lipids or residual water signals, which get incorporated into the model's spline baseline or existing lipid basis functions. Our approach could misattribute metabolite basis functions to these signals, so specific measures would need to be implemented to avoid this, preferably during data acquisition or, if need be, with specific post-processing techniques[35,36] before applying our algorithm.

Lastly, the repeated modeling required for this method is far more computationally expensive than the traditional single-basis-set approach. For a library set of size N, our "greedy" algorithm requires $\frac{1}{2}N(N+1)$ model calls to complete the full IC curves, like those shown in **Figure 2**. While this is still significantly faster than an exhaustive search of all possible basis sets (requiring 2^N model calls—a factor of 1.4×10^9 slower for N = 40), there is also a possibility of not establishing the global IC optimum with our approach. To improve the computational efficiency of our method, one approach could be to use this algorithm to refine some expert-defined basis set, adding (or subtracting) case-specific metabolites like 2HG or Cystat.

4.4. Perspectives

Model selection can, in principle, be applied to any aspect of modeling. Information criteria may be useful to determine optimal lineshape representations, e.g., deciding whether simple Voigtian parametrization suffices, or whether a more generalized convolution kernel (with more model parameters) is required to adequately fit lineshape irregularities resulting from B_0

inhomogeneities. They may also help decide whether metabolite-specific frequency shifts are warranted by the data or whether a single, global, shift parameter suffices.

Parameter regularization may provide an alternative to model selection. This usually involves additional terms in the model expression that penalize the deviation of model parameters from expectation values or impose constraints on, e.g., smoothness. This often-overlooked injection of prior knowledge is widely considered to be the "secret ingredient" of LCModel and is not well studied. Future work should, therefore, explore whether this approach can complement model selection. For example, instead of excluding certain metabolites like 2HG altogether, it might be wiser to simply set their amplitude expectation values to zero. This would incentivize the model to only assign signal to 2HG when it is strongly warranted by the data, akin to approaches such as "LASSO"[37]. Of course, choosing the nature and strength of the regularization terms is, yet again, a form of operator bias.

Finally, it is important to note the influence of the model, itself. A modification to—e.g. regularization of the metabolite amplitudes or baseline—will inevitably alter the IC scores, and, eventually, the derived basis set. Indeed, the best solution to the epistemic uncertainty problem may be to not even try to select a *single* model but to synthesize the evidence from *multiple* models. So-called multiverse analyses[38–42] have been pioneered to integrate multiple statistical models in psychology and have recently been applied to resting-state fMRI analysis. It has further been proposed to weigh different models by their Akaike information criterion scores[43–45] to emphasize the evidence from more parsimonious models, conceptually similar to CRLB weighting to improve statistical inference across multiple spectra[46].

Conclusion

Data-driven determination of the basis set composition is feasible with single-spectrum and group-level optimization. With refinement, this method could provide a valuable data-driven way to derive or refine basis sets, reduce the operator bias of MRS analyses, enhance the objectivity of quantitative analyses, and increase the clinical viability of MRS.

Consent for publication

All authors consent to the publication of this study.

Data availability

The synthetic data, the means to generate it, the algorithm, and all supplementary code are freely available online: DOI 10.17605/OSF.IO/P2USJ.

Funding

This work was supported by National Institutes of Health Grants R01 EB035529, R21 EB033516, R00 AG062230, R01 EB016089, R01 EB023963, K99 AG080084, and P41

EB031771. SA was funded by the Mildred Scheel Career Center Frankfurt (Deutsche Krebshilfe).

References

1. Near J, Harris AD, Juchem C, et al. Preprocessing, analysis and quantification in single-voxel magnetic resonance spectroscopy: experts' consensus recommendations. *NMR Biomed*. 2021;34(5):e4257. doi:10.1002/nbm.4257

2. Wilson M, Andronesi O, Barker PB, et al. Methodological consensus on clinical proton MRS of the brain: Review and recommendations. *Magn Reson Med*. 2019;82(2):527-550. doi:10.1002/mrm.27742

3. Dang L, White DW, Gross S, et al. Cancer-associated IDH1 mutations produce 2-hydroxyglutarate. *Nature*. 2009;462(7274):739-744. doi:10.1038/nature08617

4. Demler VF, Sterner EF, Wilson M, Zimmer C, Knolle F. The impact of spectral basis set composition on estimated levels of cingulate glutamate and its associations with different personality traits. *BMC Psychiatry*. 2024;24(1):320. doi:10.1186/s12888-024-05646-x

5. Branzoli F, Deelchand DK, Liserre R, et al. The influence of cystathionine on neurochemical quantification in brain tumor in vivo MR spectroscopy. *Magn Reson Med*. 2022;88(2):537-545. doi:10.1002/mrm.29252

6. Horská A, Barker PB. Imaging of Brain Tumors: MR Spectroscopy and Metabolic Imaging. *Neuroimaging Clin*. 2010;20(3):293-310. doi:10.1016/j.nic.2010.04.003

7. Považan M, Hangel G, Strasser B, et al. Mapping of brain macromolecules and their use for spectral processing of 1H-MRSI data with an ultra-short acquisition delay at 7T. *NeuroImage*. 2015;121:126-135. doi:10.1016/j.neuroimage.2015.07.042

8. Bhogal AA, Schür RR, Houtepen LC, et al. 1H–MRS processing parameters affect metabolite quantification: The urgent need for uniform and transparent standardization. *NMR Biomed*. 2017;30(11):e3804. doi:10.1002/nbm.3804

9. Hong S, Shen J. Neurochemical correlations in short echo time proton magnetic resonance spectroscopy. *NMR Biomed*. 2023;36(7):e4910. doi:10.1002/nbm.4910

10. Hong S, An L, Shen J. Monte Carlo study of metabolite correlations originating from spectral overlap. *J Magn Reson*. 2022;341:107257. doi:10.1016/j.jmr.2022.107257

11. Akaike H. A new look at the statistical model identification. *IEEE Trans Autom Control*. 1974;19(6):716-723. doi:10.1109/TAC.1974.1100705

12. Sugiura N. Further analysis of the data by Akaike's information criterion and the finite corrections: Further analysis of the data by akaike' s. *Commun Stat - Theory Methods*. 1978;7(1):13-26. doi:10.1080/03610927808827599

13. Hurvich CM, Tsai CL. Model Selection for Extended Quasi-Likelihood Models in Small Samples. *Biometrics*. 1995;51(3):1077-1084. doi:10.2307/2533006

14. Schwarz G. Estimating the Dimension of a Model. *Ann Stat*. 1978;6(2):461-464.

15. Vonta I. Model selection and model averaging. *J Appl Stat*. 2010;37(8):1419-1420. doi:10.1080/02664760902899774

16. Burnham KP, Anderson DR, eds. *Model Selection and Multimodel Inference*. New York, NY: Springer; 2004. doi:10.1007/b97636

17. Wilson M. Adaptive baseline fitting for MR spectroscopy analysis. *Magn Reson Med*. 2021;85(1):13-29. doi:10.1002/mrm.28385

18. Branzoli F, Deelchand DK, Sanson M, Lehéricy S, Marjańska M. In vivo 1H MRS detection of cystathionine in human brain tumors. *Magn Reson Med*. 2019;82(4):1259-1265. doi:10.1002/mrm.27810

19. Branzoli F, Pontoizeau C, Tchara L, et al. Cystathionine as a marker for 1p/19q codeleted gliomas by in vivo magnetic resonance spectroscopy. *Neuro-Oncol*. 2019;21(6):765-774. doi:10.1093/neuonc/noz031

20. Choi C, Ganji SK, DeBerardinis RJ, et al. 2-hydroxyglutarate detection by magnetic resonance spectroscopy in IDH-mutated patients with gliomas. *Nat Med*. 2012;18(4):624-629. doi:10.1038/nm.2682

21. Hui SCN, Saleh MG, Zöllner HJ, et al. MRSCloud: A cloud-based MRS tool for basis set simulation. *Magn Reson Med*. 2022;88(5):1994-2004. doi:10.1002/mrm.29370

22. Simpson R, Devenyi GA, Jezzard P, Hennessy TJ, Near J. Advanced processing and simulation of MRS data using the FID appliance (FID-A)—An open source, MATLAB-based toolkit. *Magn Reson Med*. 2017;77(1):23-33. doi:10.1002/mrm.26091

23. Zöllner HJ, Davies-Jenkins C, Simicic D, Tal A, Sulam J, Oeltzschner G. Simultaneous multi-transient linear-combination modeling of MRS data improves uncertainty estimation. *Magn Reson Med*. 2024;92(3):916-925. doi:10.1002/mrm.30110

24. Oeltzschner G, Zöllner HJ, Hui SCN, et al. Osprey: Open-source processing, reconstruction & estimation of magnetic resonance spectroscopy data. *J Neurosci Methods*. 2020;343:108827. doi:10.1016/j.jneumeth.2020.108827

25. Simicic D, Zöllner HJ, Davies-Jenkins CW, Hupfeld KE, Edden RAE, Oeltzschner G. Model-based frequency-and-phase correction of 1H MRS data with 2D linear-combination modeling. *bioRxiv*. March 2024:2024.03.26.586804. doi:10.1101/2024.03.26.586804

26. Provencher SW. Estimation of metabolite concentrations from localized in vivo proton NMR spectra. *Magn Reson Med*. 1993;30(6):672-679. doi:10.1002/mrm.1910300604

27. Gong T, Hui SCN, Zöllner HJ, et al. Neurometabolic timecourse of healthy aging. *NeuroImage*. 2022;264:119740. doi:10.1016/j.neuroimage.2022.119740

28. Craven AR, Bhattacharyya PK, Clarke WT, et al. Comparison of seven modelling algorithms for γ-aminobutyric acid–edited proton magnetic resonance spectroscopy. *NMR Biomed*. 2022;35(7):e4702. doi:10.1002/nbm.4702

29. Zöllner HJ, Považan M, Hui SCN, Tapper S, Edden RAE, Oeltzschner G. Comparison of different linear-combination modeling algorithms for short-TE proton spectra. *NMR Biomed*. 2021;34(4):e4482. doi:10.1002/nbm.4482

30. Marjańska M, Deelchand DK, Kreis R, Team the 2016 IMSGFC. Results and interpretation of a fitting challenge for MR spectroscopy set up by the MRS study group of ISMRM. *Magn Reson Med*. 2022;87(1):11-32. doi:10.1002/mrm.28942

31. Lin A, Andronesi O, Bogner W, et al. Minimum Reporting Standards for in vivo Magnetic Resonance Spectroscopy (MRSinMRS): Experts' consensus recommendations. *NMR Biomed*. 2021;34(5):e4484. doi:10.1002/nbm.4484

32. Du X, Hu H. The Roles of 2-Hydroxyglutarate. *Front Cell Dev Biol*. 2021;9. doi:10.3389/fcell.2021.651317

33. Cudalbu C, Behar KL, Bhattacharyya PK, et al. Contribution of macromolecules to brain 1H MR spectra: Experts' consensus recommendations. *NMR Biomed*. 2021;34(5):e4393. doi:10.1002/nbm.4393

34. Zöllner HJ, Davies-Jenkins CW, Murali-Manohar S, et al. Feasibility and implications of using subject-specific macromolecular spectra to model short echo time magnetic resonance spectroscopy data. *NMR Biomed*. 2023;36(3):e4854. doi:10.1002/nbm.4854

35. Gudmundson AT, Davies-Jenkins CW, Özdemir İ, et al. Application of a 1H brain MRS benchmark dataset to deep learning for out-of-voxel artifacts. *Imaging Neurosci*. 2023;1:1-15.

36. Lin L, Považan M, Berrington A, Chen Z, Barker PB. Water removal in MR spectroscopic imaging with L2 regularization. *Magn Reson Med*. 2019;82(4):1278-1287. doi:10.1002/mrm.27824

37. Tibshirani R. Regression Shrinkage and Selection Via the Lasso. *J R Stat Soc Ser B Methodol*. 1996;58(1):267-288. doi:10.1111/j.2517-6161.1996.tb02080.x

38. Dafflon J, F. Da Costa P, Váša F, et al. A guided multiverse study of neuroimaging analyses. *Nat Commun*. 2022;13(1):3758. doi:10.1038/s41467-022-31347-8

39. Steegen S, Tuerlinckx F, Gelman A, Vanpaemel W. Increasing Transparency Through a Multiverse Analysis. *Perspect Psychol Sci*. 2016;11(5):702-712. doi:10.1177/1745691616658637

40. Del Giudice M, Gangestad SW. A Traveler's Guide to the Multiverse: Promises, Pitfalls, and a Framework for the Evaluation of Analytic Decisions. *Adv Methods Pract Psychol Sci*. 2021;4(1):2515245920954925. doi:10.1177/2515245920954925

41. Simonsohn U, Simmons JP, Nelson LD. Specification curve analysis. *Nat Hum Behav*. 2020;4(11):1208-1214. doi:10.1038/s41562-020-0912-z

42. Carp J. On the Plurality of (Methodological) Worlds: Estimating the Analytic Flexibility of fMRI Experiments. *Front Neurosci*. 2012;6. doi:10.3389/fnins.2012.00149

43. Wagenmakers EJ, Farrell S. AIC model selection using Akaike weights. *Psychon Bull Rev*. 2004;11(1):192-196. doi:10.3758/BF03206482

44. Akaike H. On the Likelihood of a Time Series Model. *J R Stat Soc Ser Stat*. 1978;27(3-4):217-235. doi:10.2307/2988185

45. Bozdogan H. Model selection and Akaike's Information Criterion (AIC): The general theory and its analytical extensions. *Psychometrika*. 1987;52(3):345-370. doi:10.1007/BF02294361

46. Miller JJ, Cochlin L, Clarke K, Tyler DJ. Weighted averaging in spectroscopic studies improves statistical power. *Magn Reson Med*. 2017;78(6):2082-2094. doi:10.1002/mrm.26615