Introduction

Magnetic resonance spectroscopy (MRS) is a non-invasive imaging modality that provides in vivo information on the metabolic profile of tissues, enabling the evaluation of various pathologies. MRS generates spectra that can be analyzed to quantify metabolite concentrations, providing valuable insights into tissue composition and metabolic activity. In the brain, MRS has been extensively used to investigate a wide range of pathologies, including neurodevelopmental[1,2,3] and neurodegenerative diseases[4,5,6,7], inborn errors of metabolism 8,9,10, brain tumors[11,12,13,14], as well as age-related changes[15,16,17]. All of those studies are based on analyzing metabolite concentrations, or how the concentrations change. This is only possible with accurate post-processing and quantification steps.

For MRS to become a clinically relevant technique, it is important to assure the precision and accuracy of MRS data analysis across a wide range of data scenarios, e.g. acquisition parameters, data quality regimes, and disease-related metabolic patterns. Clinical MRS data varies drastically across these dimensions, but is usually not widely available for method development and evaluation due to data privacy restrictions. Furthermore, in vivo data lacks access to a ground truth. This makes it difficult to determine how reliable, sensitive, and robust the quantitative outcomes are from different MRS analysis procedures. Phantom data can be useful in developing and validating acquisition and analysis methods, but usually does not adequately reflect in vivo spectra. For example, phantom data does not contain broad signals from macromolecules and lipids, which are a major source of uncertainty during spectral modeling. In vivo MRS data is further affected by tissue and susceptibility heterogeneities, reflected in irregular lineshapes and artifacts.

Recent years have seen a dramatic increase in the use of synthetic MRS data, catalyzed by the advent of machine learning[18] (ML) and deep learning[19,20,21] (DL) quantification algorithms that require vast amounts of data for training and validation. These deep learning techniques require datasets comprising tens of thousands to hundreds of thousands of spectra to be effective, a volume of data that few centers globally possess. As a result, generating realistic, in vivo-like synthetic data is necessary to facilitate more research into these applications. Many software packages[22,23,24,25,26,27,28] can now accurately calculate the evolution of spin systems during any given pulse sequence, enabling researchers to generate metabolite basis sets. These can be assembled into arbitrarily large datasets that can approximate in vivo spectra with known ground truth values, offering a pathway to evaluate accuracy and precision of data analysis methods.

The main challenge of synthetic MRS data generation is to adequately incorporate all physical phenomena underlying in vivo data. One aspect is the adequate choice of model parameters to reflect different pathological conditions. For example, tumors have vastly different metabolic signatures than healthy tissue. More importantly, synthetic data needs to capture acquisition-induced artifacts and nuisance signals that are difficult to reproduce in a phantom: effects of susceptibility and field inhomogeneity, macromolecule, lipid, and residual water signals, and various other artifacts. While many research groups now routinely use synthetic data, the underlying software, crucial data generation models, and parameter distributions are rarely made publicly available. This not only hampers systematic comparison, but also the formation of consensus best practices for synthetic data generation.

Recent work has begun tocompare the impact of various spectral modeling components on

metabolite quantification as well as the performance of commonly used spectral fitting models.[29,30,31,32] Currently, these comparisons show worrisome agreement of metabolite quantities between the different linear combination fitting methods. Poor agreement between fitting methods prevents any sort of meaningful comparison between studies published by different institutions that employ different fitting protocols which limits the generalizability of published quantification parameters and any conclusions drawn from them. The rise in the use of synthetic MRS data without standardization will further exacerbate this problem. Literature shows that, already, synthetic data is simulated using physical models with a variety of complexity and spectral components.[20,33,34,35]

Growing interest in training ML and DL models with synthetic data will compound the reproducibility and generalizability problems already being experienced in the traditional MRS field. To address this challenge, it is essential to establish standards and best practices for simulating MRS data. This will require collaboration and consensus-building among researchers, as well as the development and adoption of open-source frameworks for data generation. To ensure widespread adoption, these frameworks will need to be applicable across a range of MRS applications. While a consensus-building effort is beyond the scope of this work, an open-source framework is not.

In this work, we therefore present a framework for a modular synthetic data generation model. The basic model applies well-defined distributions of physical parameters commonly used in linear-combination modeling software, corresponding to amplitudes, lineshapes, phases, and frequency shifts. Furthermore, it incorporates a realistic B0 map generator to simulate in vivo-like field heterogeneity conditions, and uses parameterized models to describe residual water and smooth background signal contributions.

The software is designed to simulate spectra at any stage of acquisition: from individual coil elements to unaligned transients to fully processed spectra. These features allow researchers to benchmark new data processing methods, not just modeling algorithms, against known ground truth parameters. The framework's flexibility allows users to simulate custom-tailored datasets for a large variety of clinical scenarios. Accompanying this simulator are tools to analyze existing in vivo clinical datasets and extract parameter distributions, which then allows for augmenting in vivo data with similar synthetic data.

The open-source code base allows for seamless incorporation of future additions to expand the software's capabilities to include more types of spectra, spectral model components, and parameter distributions reflecting pathological metabolic signatures. This work is intended to be a community resource to provide researchers, trainees, and experts alike with access to high-quality and comparable synthetic data as a source of continuity in the field.

2 METHODS

2.1 Physics Model Algorithm

In literature, various MRS physics models have been proposed to simulate brain spectra. They begin with simulated basis functions that are assumed to have been simulated using appropriate pulse sequence parameters for the scenario of interest. These metabolite basis functions are then modulated by scaling factors that indicate their underlying concentrations. Most models then apply a simple Lorentzian lineshape.[20,33,34] Phase offsets[34,35] and frequency shifts[20] can optionally be applied. Finally, some type of broad baseline is typically added. These models are simple and do not capture the full complexity of clinical data. Additionally, they often include non-public components such as baselines, macromolecules, and lipid signals that are extracted from private datasets.

To maximize generalizability and usefulness, the data simulator should comprehensively model known spectral components, which were identified through a review of state-of-the-art fitting techniques and currently available fitting algorithms. These spectral components allow the model to account for a large variety of scenarios and artifacts. The physics model proposed in this work is described by the following set of equations:
where N is the whole set of metabolites being modeled, Mn is the scaling factor for metabolite n, the Lorentzian variable dn and the Gaussian variable gn combine to define a Voigt lineshape, t is time, and Δfn is the metabolite-specific frequency shift. Global zero- and first-order phase offsets are added using φ0 and φ1 while eddy current effects are described using two variables as a function of time t: the amplitude, A0, and the time constant, tc0. In lieu of the Gaussian term, imperfect shimming and severe lineshape distortions associated with large susceptibility effects can be applied using Δωr , which is the modeled B0 at location r inside the voxel of interest. snr0 is the desired SNR of the spectrum, while baseline and resH2O are semi-parameterized signals that account for the broad baseline offset and the poorly defined residual water contributions. If coil-combined FIDs are required, then the simulation can stop after Eqn. 1a. Multi-coil acquisitions are simulated in Eqn. 1b in which the operator Coil generates C coil transients and applies a distribution of SNR values and coil weights using snrc and sensc. Frequency drifts and phase drifts are then added using Δfc and Δφc, respectively. When necessary, Eqn. 1c can apply apodization using TL in Hz and the FIDs can be zero-filled to length len. Then the Fourier transform F can convert the FIDs to the frequency domain. Each term is discussed in more detail below.

2.1.1 Overview

The proposed physics model was developed to mirror the actual data acquisition sequence and spectral fitting process. This reverse engineering informed both the steps to include and the order of operations, which are meant to ensure that any experimental fitting parameters will match the simulation parameters. The following sections are presented in order according to their implementation in the physics model. A step-by-step visualization of this model is illustrated in Fig. 1.

2.1.2 Basis Functions

MRI, and its derivatives, are spatially resolved imaging modalities. Even singular pixels in MRI images represent a 3D volume with a spatial distribution, as shown in Fig. 2a. Addressing this spatial component is important when working with quantitative MR modalities like spectroscopy. Inaccurately simulated basis functions cause errors in metabolite quantification when fitting in vivo data. With simulated data, such basis functions negatively

impact the realism of the simulations, limiting their usefulness, especially for validating quantitative methods.

The importance of considering spatial localization led to Landheer et al.'s MARSS[23] software package being selected for the default basis functions provided with this simulator. MARSS produces high-fidelity outputs by simulating 128 points in each direction. This very accurately captures the spatial nature of MR imaging. MARSS has a large number of common brain metabolites already defined in their template files. Vendor-specific basis functions for these metabolites can be simulated with PRESS and STEAM sequences. Custom basis functions can also be simulated with various metabolites, T1 and T2 relaxations, and specialized pulse sequences, e.g. editing sequences, (semi-)LASER, etc.

Macromolecules and Lipids

Current spectral fitting methods for modeling MM and lipid signals are based on creating a group of curves that resemble clinical data but are not informed by any underlying physical phenomenon. Each fitting package contains their own set of basis functions for modeling these regions. Until this knowledge gap is filled and better simulation methods are developed, this work will use pre-generated basis functions from Osprey[36] that were resampled to match the simulated basis functions from MARSS.

2.1.3 Amplitude Modulation

Metabolite quantities produced during spectral fitting are of an arbitrary scale. Comparing these quantities with a standard reference puts them into context. In vivo proton scans generally use an internal reference metabolite for relative quantification. Creatine is the default metabolite because its concentration is relatively stable. As a result, concentrations maps are generally reported as ratios with respect to creatine and all amplitude values in this model are defined wrt creatine as the default. For this framework, physiological values were derived from work by Das et al.[18,34]. These ranges were then expanded to include values observed in in vivo scans from a private glioma dataset. In keeping in line with the LCModel, the expected concentrations ranges represent the scaling factors needed for spectral fitting instead of the ratios of peak integrals or peak heights

2.1.4 Lineshape Profiles

In spectral fitting, the Voigt lineshape profile is the most commonly used as it most closely matches clinical data. It is a combination of a Lorentzian and a Gaussian and is used in various fitting packages, such as LCModel[37], TARQUIN[38], jMRUI[27], and Osprey[36]. Each peak is characterized by an individual Lorentzian value while a single Gaussian value is applied to the metabolites, while a second value can be applied to the macromolecules and lipids. Standard practice from the aforementioned software packages assigns a single Lorentzian value to each metabolite, instead of each moiety. However, experimental results from Wyss et al.[39] can be selected which characterized T2 relaxation values at the moiety-level for various brain metabolites at 3T in three different regions in the brain. As more metabolites are characterized, new information can be added to the model. For completeness, it is also possible to specify either a purely Lorentzian or a purely Gaussian lineshape.

2.1.5 B0 Inhomogeneities

Lower and higher order shimming procedures homogenize the magnetic field in the target

volume to different degrees. However, certain regions of the brain, such as the prefrontal cortex or deep brain structures like the thalamus or basal ganglia, are more difficult to shim and therefore suffer from large magnetic susceptibility effects, resulting in significant lineshape distortions. Fig. 2b shows the normal, subtle B0 changes across the volume of a spectroscopy voxel while Fig. 2c illustrates these high susceptibility effects. In such cases, spectra from these regions exhibit significant lineshape distortions which cannot be adequately characterized using idealized lineshape profiles. Fig. 3 shows the result of high susceptibility effects on lineshapes.

Small B0 inhomoegeneities are, in general, sufficiently modeled by the Gaussian term of the Voigt lineshape. However, to simulate more severe distortions, a B0 field volume needs to be modeled and applied to the basis functions. In general, this approach mirrors Li et al.[40], but the B0 field map is simulated rather than acquired. As with MARSS, Li et al. suggests using multiple points in each direction instead of a single value per voxel. The exact number of points used in each direction is described by the size of the spectroscopy voxel divided by the size of an anatomical imaging voxel. The default values assume sizes of $10cm^3$ and $0.5cm^3$ respectively, which results in 20^3 simulation points. However, any cuboidal shape, rectangular or otherwise, can be modeled. The B0 field is defined by four variables, all of which are mean offsets: ±dx, ±dy, ±dz, and μ. dx, dy, and dz describe half of the change in B0 in their respective direction from the voxel's center and μ is the mean of the entire voxel.

2.1.6 Baseline and Residual Water

Currently, the underlying physical phenomena that induce spectral baseline offsets are poorly understood. In fact, there is no physics-based model for simulating these offsets. Similarly, the residual water region is also poorly characterized. Therefore, a naive random model can be used in conjunction with clinically observed constraints to approximate what is observed in vivo. This work proposes a smoothed, pseudo-random, bounded walk generator for both the broad spectral baseline and the more irregular residual water region. The approach is elaborated on in Algorithm 1. Customizable profiles were developed for each artifact to more closely approximate what is expected in vivo. Immense variety of outputs can be achieved by randomly sampling the parameters from distributions instead of fixing them to set values. Once simulated, they are resampled to match the ppm range of acquired data and the order of magnitude is matched to the spectra. The Hilbert transform is then used to generate the corresponding complex component before being added to the FID. As shown in Fig. 4, this generator produces very different outputs depending on the specified configurations. Fig. 4a shows very broad, smooth lines while Fig. 4b shows highly irregular lines that closely resemble residual water regions. All outputs are then scaled to modulate the impact on the final spectra. A more detailed exploration of this algorithm and the effects of each parameter are presented in the supplement for baseline and residual water simulations.

2.1.7 Noise

The noise in this model assumes a Gaussian distribution. The input SNR is first converted from decibels to a linear SNR. Then the standard deviation for this distribution is calculated using the maximum height of a metabolite of choice in the real spectrum and the desired SNR. The real and imaginary components of the noise can be correlated using the Hilbert transform. If they are assumed to be uncorrelated, then separate noise vectors are sampled for each component.

2.1.8 Phase Offsets

Zero-Order Phase

FIDs and spectra are complex data types which consist of a real and imaginary component. A 0◦ zero-order phase offset results in absorption and dispersion spectra in these components, respectively. An absorption spectrum exhibits peaks with idealized lineshapes that are purely positive or purely negative, while dispersion spectra exhibit peaks that are both positive and negative. As shown in Fig. 5, non-zero degree offsets result in a mixture of absorption and dispersion spectra. In absorption mode, spectral peaks directly reflect the number of hydrogens of that species and the concentration of that molecule. This relationship means that the phase has a direct impact on metabolite quantification.

First-Order Phase

First-order phase, i.e. linear phase, is a frequency-dependent offset that emanates from a reference point, i.e. the center frequency which is typically the water peak at 4.65ppm but can be modified when necessary. It carries the unit degrees per ppm. A linear phase offset creates asymmetrical line shapes that grow larger as one moves away from the reference point.

2.1.9 Frequency Shifts

Frequency shifts observed in a spectrum result from complex interactions with a variety of factors. During the acquisition, the FID experiences a global frequency shift. However, individual moieties from metabolites and nuisance signals from macromolecules, lipids, and fats such as diglycerides and triglycerides, can experience individual frequency shifts which are attributed to temperature and pH effects.

The current implementation allows each metabolite, or moeity depending on the simulation, to have a minor, independent frequency shift in addition to the global shift which is in line with common spectral fitting protocols. For more clinically realistic spectra, values can be used from the work by Wermter et al.[41] which characterized the temperature-induced frequency shift of brain metabolite moieties with temperature sensitivity. As more metabolites are characterized for their temperature- and pH-sensitivities, this information can be added to simulate more realistic spectra.

2.1.10 Eddy Currents

Eddy currents are common artifacts in MRI acquisitions that are induced by changes in the magnetic field, typically caused by the imaging gradients and present as time-dependent resonant frequency shifts. Correction techniques, such as the Klose[42], tend to be non-parameterized, making it difficult to model the exact effect of each approach. Near et al. in FID-A[24], however, provide a parameterized equation for simulating first-order eddy currents. These artifacts are applied as a function of amplitude, A, time constant, tc, and time, t. The time constant must be short enough that it occurs entirely within the recorded echo, otherwise it will appear as a simple, global frequency shift. The effects of eddy currents can be seen in Fig. 6.

2.1.11 Multi-Coil Transients

A transient copy is made for each coil in the simulated scenario. These transients will experience additional artifacts including zero-order phase and frequency drifts, scaling due to coil sensitivity, and decreased SNR values. To allow for maximum variation in the

simulations, each parameter can be sampled from distributions and is discussed below.

Noise
Multi-coil acquisitions lead to an SNR improvement of the final spectrum by a factor of the square root of the number of non-zero weighted transients. To vary the SNR among the transients, this model scales the target linear SNR according to the number of coils and then samples scaling factors from a narrow normal distribution to maintain the mean target SNR.

Frequency Drift and Phase Drift
Frequency drifts and phase drifts are phenomena observed in multi-coil acquisitions in which each coil transient has an independent offset. The lower SNR values of the transients make it harder to accurately correct these inter-transient variations. Therefore, drifts are typically minimized between the transients, called alignment. Proper alignment will preserve the underlying spectral features once the transients are combined. These offsets and alignments are shown in Fig. 7.

Coil Sensitivity
A variety of coil combination techniques can be used to successfully combine multi-coil spectra. While these techniques differ in how they calculate the weights, all of them use weights to scale the transients before averaging. Assigning context, such as water peak height or coil sensitivity maps, to these weights when planning the simulations can help define the necessary parameter ranges and distributions to be in line with a given clinical protocol.

2.1.12 Final Steps
The desired use case will determine if a FID or a spectrum is necessary. If a FID is required, the simulation is finished and the data will be exported. If a spectrum is required, the Fourier transform will recover the spectrum at which point it can be cropped and resampled to a desired ppm range and spectral length. The default interpolation technique in this framework is a cubic Hermite modified Akima interpolator with momentum.

2.2 Exporting Data
The default export file format is .mat. These files include the data, spectral fits, simulation parameters, baseline offsets, and quantification results. To facilitate using the simulated spectra in various software packages, they are also exported in the NIfTI-MRS format[43].

2.3 Fitting Parameter Analysis
The process of simulating a new dataset requires careful consideration of various factors, including the selection of appropriate parameter ranges and distributions. The optimal customization of these parameters depends on the intended use and application of the dataset. For instance, deep learning-based quantification models benefit from independent, uniform distributions that include all values the model will be expected to encounter. When validating a traditional spectral fitting model that includes soft constraints, it is crucial to incorporate those constraints when defining the parameter distributions. This ensures that the simulated dataset accurately reflects the underlying distribution of the target dataset.

To mimic an in vivo dataset, accurate descriptions of clinical fitting parameters are crucially important. In collaboration with the developers of Osprey[36], their software is now capable of exporting the spectral fitting parameters after quantification. Tools in this framework can then

load those exported files and prepare the data for further analysis. Currently, this framework uses the python library Fitter[44] to identify the best fitting probability distribution for every parameter. A priori knowledge, either from prior knowledge or data exploration, can narrow down the search range and speed up the analysis. The outputs for each parameter include evaluation metrics for the best performing distributions as well as a numerical characterization of the best fitting distribution.

2.3.1 Recommendations

The authors generally recommend that simulations include all relevant artifacts unless there is a specific reason to exclude them. A simulated dataset should include all phenomena that are expected to be encountered when the final work is deployed. Even highly accurate post-processing techniques have limitations and biases and leave some residue of the corrected artifacts. To ensure consistency between the simulated and clinical data, the artifacts should be included in the simulations and removed via the users' own fitting protocols.

Although not recommended, residual artifacts and post-processing techniques can be included in the simulations. Phase and frequency corrections can be simulated by applying a minimal offset during the initial simulation, which can be implemented in the parameter sampling protocol. Similarly, eddy currents can be scaled down by minimizing the sampled amplitudes. While not part of the acquisition protocol, apodization and zero filling are also possible. Apodization improves the SNR by multiplying the FID by a filter function, typically an exponential decay function or a Lorenztian-to-Gaussian transform. This framework implements an exponential decay as a function of time, t, and TL which defines the amount of apodization in Hz. Zero filling simply pads the FID with zeros to a defined length before the Fourier transform.

2.4 Code

This repository was written in PyTorch 1.11.0 and Python 3.9.7. Since this framework generates batches of spectra instead of individual spectra sequentially, a simulation batch size needs to be specified which will be affected by the spectral length and complexity of the simulations. As long as the batch size is set appropriately given the users' amount of RAM, this framework can be employed on standard computers without any special hardware. After publication, the repository will be available on GitHub, at https://github.com/JohnLaMaster/MRS-Sim, and MRSHub.