NMR in Biomedicine

MRS-Sim: Open-Source Framework for Simulating In Vivo-like Magnetic Resonance Spectra

Journal:	*NMR in Biomedicine*
Manuscript ID	NBM-23-0200.R2
Wiley - Manuscript type:	Research Article
Date Submitted by the Author:	06-Aug-2024
Complete List of Authors:	LaMaster, John; Munich Institute of Biomedical Engineering Technical University of Munich; Technical University of Munich, School of Computation, Information, and Technology Oeltzschner, Georg; Johns Hopkins Medicine Russell H Morgan Department of Radiology and Radiological Science Li, Yan; University of California San Francisco Department of Radiology and Biomedical Imagery
Keywords:	
Keywords:	

Note: The following files were submitted by the author for peer review, but cannot be converted to PDF. You must view these files (e.g. movies) online.

MRS_Sim_Figure1_PM_stages_1.eps
MRS-Sim_Figure2_Spatial_localization.eps
MRS-Sim_Figure3_B0_effects.eps
MRS-Sim_Figure4_random_walk.eps
MRS-Sim_Figure5_phase_effects.eps
MRS-Sim_Figure6_ eddy_current_effects.eps
MRS-Sim_Figure7_transients.eps
MRS_Sim_Figure_8_Lorentzian_Broadening_samples.eps
MRS_Sim_Figure_9_Gaussian_Broadening_samples.eps
MRS_Sim_Figure_10_Voigt_Broadening_samples.eps
MRS_Sim_Figure_11_Frequency_Shifts_samples.eps
MRS_Sim_Figure_12_Zero-Order_Phase_samples.eps
MRS_Sim_Figure_13_First-Order_Phase_samples.eps
MRS_Sim_Figure_14_SNR_samples.eps
MRS_Sim_Figure_15_ECC_samples.eps
MRS_Sim_Figure_16_multicoil_transients_samples.eps
MRS_Sim_Figure_17_B0_inhomo_samples.eps
MRS_Sim_Figure_18_clean_samples.eps
MRS_Sim_Figure_19_.Random_samples_with_Artifacts.eps
tab:supp:condensed.tex
MRS-Sim_Supplement_rebuttal_2.zip

SCHOLARONE™
Manuscripts

Supplement: MRS-Sim

1 | B_0 FIELD SIMULATOR

FIGURE S1.1 B_0 field volume with labels corresponding to the mathematical model for simulating the heterogeneities.

As described in Li et al.[1], using the B_0 field map, measured during the acquisition, to modulate the basis functions prior to fitting showed significant improvement in metabolite quantification and the fit residual surrounding the metabolite peaks. The fact that this technique has a direct impact on metabolite quantification shows the importance of including it when simulating spectra. The first step is to define the spectral and imaging resolutions of the simulated acquisitions. The defaults, shown below, assume a cubic spectroscopy voxel, but any cuboidal shape is acceptable.

spectroscopy voxel [10 mm x 10 mm x 10 mm]

imaging voxel [0.5mm x 0.5mm x 0.5mm]

The dimensions of the volume being modeled are the same as the spectroscopy voxel. The number of points included in the simulation is the quotient of the spectroscopy resolution and the anatomical imaging resolution.

$$\mathbf{num_pts} = \frac{\text{spectroscopy resolution}}{\text{imaging resolution}} = \left[\frac{10\text{mm}}{0.5\text{mm}}_X, \frac{10\text{mm}}{0.5\text{mm}}_Y, \frac{10\text{mm}}{0.5\text{mm}}_Z\right] \tag{1}$$

Once the model volume has been defined, the actual B_0 field, $dB0$, can then be modeled. Magnetic fields do not have hard, disjointed heterogeneities, but smooth variations. This is true even in the presence of high susceptibility effects causing strong distortions as long as the voxels are reasonably small. This model therefore assumes linearly varying gradients across the volume of the voxel. This assumption allowed the problem to be reduced to four variables:

$\mathbf{dx}$ half of the mean change in the x-direction

$\mathbf{dy}$ half of the mean change in the y-direction

$\mathbf{dz}$ half of the mean change in the z-direction

μ the mean offset of the entire voxel

The next step defines the gradients in each direction as shown in Eqns. 2, 3, and 4. As mentioned above, this model assumes linearly varying gradients, but more complex gradients can be implemented manually at this stage.

$$\mathbf{x} = \text{linspace}(\textbf{-dx}, \textbf{dx}, \textbf{num_pts}_X) \tag{2}$$

$$\mathbf{y} = \text{linspace}(\textbf{-dy}, \textbf{dy}, \textbf{num_pts}_Y) \tag{3}$$

$$\mathbf{z} = \text{linspace}(\textbf{-dz}, \textbf{dz}, \textbf{num_pts}_Z) \tag{4}$$

Peer Review Only

Once the gradients have been defined, the field map can be modeled very easily, as shown in Eqn. 5. The primary challenge in this step is reshaping the tensors to get the correct n+3D output.

$$\mathbf{dB0} = \mathbf{x} * \mathbf{y} * \mathbf{z} + \mu, \text{ such that} \tag{5}$$

$$\mathbf{dB0}.size() = torch.Size([batchSize, \ldots, \mathbf{num_pts_X}, \mathbf{num_pts_Y}, \mathbf{num_pts_Z}]) \tag{6}$$

Modulation due to the heterogeneous magnetic field can be applied in both the time-domain and the frequency-domain. In the time-domain, the complex exponential is applied via multiplication with the basis functions, as is shown below in Eqn. 7. In the frequency-domain, the complex exponential needs to be convolved with the spectrum. In both mathematical and practical terms, it is simpler to apply the modulation in the time-domain. Therefore, this model has only implemented the time-domain modulation.

$$F(t) = \sum_{n}^{N} M_n * basisfcn_n * \underbrace{\sum_{r=1}^{R} e^{-i\Delta\omega_r t}}_{B_0\ inhomog.}, \text{ where } \Delta\omega = \mathbf{dB0}.flatten(dims = [-3:-1]) \text{ and } R = cumprod(\mathbf{num_pts}) \tag{7}$$

In this equation, the model **dB0** is flattened along the last three dimensions representing the **x**-, **y**-, and **z**-dimensions respectively. Therefore, **r** represents the linear indices from [1,R] corresponding the subscripts of the points within the 3D volume. The effective magnetic field at each of those points within the voxel has a direct effect on each moeity that resonates and contributes to the overall signal. Because of this, each basis function is modulated independently before summation.

Peer Review Only

2 | SIMULATION PARAMETERS

The following table contains spin definitions and recommended simulation parameter ranges for 35 brain metabolites and water. This table is based on literature values and does have some gaps. The recommended ranges represent the minimum and maximum values across physiological and pathological differences. The concentration ranges were scaled to represent $[0.9 * min, 1.1 * max]$. To compare a variety of sources and to find the most up-to-date values, please check the GitHub repository.

The data provided here is an excerpt of the database. So far, data has been compiled from textbooks[2,3] and commonly referenced literature sources[4,5] for defining the spin systems. The chemical shifts and coupling constants have also been reported from MARRS[6] since that is the recommended software for simulating basis sets. MARRS also compiled several sources, including Near et al.'s[7] work describing the GABA spin systems. T_2 values were collected from Wyss et al.[8] who reported moiety-specific values for selected metabolites and from Gudmundson et al.'s[9] recent meta-analysis. Concentration ranges were compiled from De Graaf's textbooks[2,3] and Gudmundson's meta-analysis[9]. Macromolecule and lipid nuisance signal characterizations come from Gudmundson et. al[9] and Hamilton et al.[10]. Finally, Wermter et al.'s[11] work describing moiety-level temperature-induced frequency shifts has also been included. The following citations were used to compile the current version of the database. As new information is added, the citation list in the repository will be updated:

Peer Review Only

Name	Moieties	Group	Chemical Shift (ppm)[2,3]	Interaction	Scalar coupling (Hz)[2,3]	T_2 [ms] Low	T_2 [ms] High	Conc. [mM] Low	Conc. [mM] High
Acetate	–	$^{2}CH_3$	1.904	–	–	–	–	–	–
Alanine	–	^{2}CH	3.775	2-3	7.230	100.00	250.00	0.090	1.650
		$^{3}CH_3$	1.467	3-3'	–	100.00	250.00		
		–	–	3-3"	–	100.00	250.00		
		–	–	3'-3"	–	100.00	250.00		
Ascorbic Acid	–	^{4}CH	4.492	4-5	2.07	73.00	250.00	0.367	2.072
		^{5}CH	4.002	5-6	6.00	73.00	250.00		
		$^{6}CH_2$	3.743	5-6'	7.60	73.00	250.00		
		–	3.716	6-6'	-11.50	73.00	250.00		
Aspartate	–	^{2}CH	3.891	2-3	3.647	36.00	204.550	0.900	7.668
		$^{3}CH_2$	2.804	2-3'	9.107	36.00	204.550		
		–	2.670	3-3'	-17.426	36.00	204.550		
2-Hydroxy-glutarate[12]	–	^{2}CH	4.022	2-3 / 2-3'	7.600 / 4.100	–	–	–	–
		$^{3}CH_2$	1.825	3-3'	-14.00	–	–		
		–	1.977	3-4 / 3-4'	5.300 / 10.400	–	–		
		$^{4}CH_2$	2.221	3'-4 / 3'-4'	10.600 / 6.00	–	–		
		–	2.272	4-4'	-15.00	–	–		
Choline	–	$N(CH_3)_3$	3.185	–	0.570	119.00	394.00	0.00	2.750
		$^{1}CH_2$	4.054	1-2 / 1'-2'	3.140 / 3.168	93.00	345.00		
		–	–	1-1'	2.572	93.00	345.00		
		$^{2}CH_2$	3.501	1'-2 / 1-2'	7.011 / 6.979	93.00	345.00		
		–	–	2-2'	2.681	93.00	345.00		
Creatine	–	$N(CH_3)$	3.027	–	–	104.00	242.700	0.325	16.505
		$^{2}CH_2$	3.913	–	–	76.00	213.800		
		NH	6.650	–	–	164.080	242.700		
Ethanol-amine	–	$^{1}CH_2$	3.818	1-2 / 1'-2'	3.897 / 3.798	–	–	0.00	1.650
		–	–	1-1' / 1-N	-10.640 / 0.657	–	–		
		$^{2}CH_2$	3.147	1'-2 / 1-2'	6.694 / 6.794	–	–		
		–	–	2-2' / 1'-N	-11.710 / 0.1420	–	–		
γ-Amino-butyric acid	–	$^{2}CH_2$	2.287	2-2'	-15.938	–	–	0.033	10.720
		–	–	2-3 / 2-3'	7.678 / 6.980	25.00	229.00		
		–	–	2'-3 / 2'-3'	6.980 / 7.678	25.00	229.00		
		$^{3}CH_2$	1.892	3-3'	-15.00	25.00	229.00		
		–	1.895	3-4 / 3-4'	8.510 / 6.503	25.00	229.00		
		$^{4}CH_2$	3.003	4-4'	-14.062	25.00	229.00		
		–	3.005	3'-4 / 3'-4'	6.503 / 8.510	25.00	229.00		
Glucose 36% α/64%β	GlcA	Refer to GlcA below				38.00	250.00	0.00	2.200
	GlcB	Refer to GlcB below				40.00	250.00	0.00	2.200
Glucose, $\alpha - anomer$ (36% = 0.36 1H)	–	^{1}CH	5.216	1-2	3.8	38.00	250.00	0.00	2.200
		^{2}CH	3.519	2-3	9.6	38.00	250.00		
		^{3}CH	3.698	3-4	9.4	38.00	250.00		
		^{4}CH	3.395	4-5	9.9	38.00	250.00		
		^{5}CH	3.822	5-6	1.5	38.00	250.00		
		^{6}CH	3.826	5-6'	6.0	38.00	250.00		
		$^{6'}CH$	3.749	6-6'	-12.1	38.00	250.00		
Glucose, $\beta - anomer$ (64% = 0.64 1H)	–	^{1}CH	4.630	1-2	8.0	40.00	250.00	0.00	2.200
		^{2}CH	3.230	2-3	9.1	40.00	250.00		
		^{3}CH	3.473	3-4	9.4	40.00	250.00		
		^{4}CH	3.387	4-5	8.9	40.00	250.00		
		^{5}CH	3.450	5-6	1.6	40.00	250.00		
		^{6}CH	3.882	5-6'	5.4	40.00	250.00		
		$^{6'}CH$	3.707	6-6'	-12.3	40.00	250.00		

(continued on the next page)

Peer Review Only

Name	Moieties	Group	Chemical Shift (ppm)[2,3]	Interaction	Scalar coupling (Hz)[2,3]	T_2 [ms]		Conc. [mM]	
						Low	High	Low	High
Glutamine	–	^{2}CH	3.767	2-3 / 2-3'	5.846 / 6.500	57.00	234.00	2.700	6.600
		$^{3}CH_2$	2.137	3-4 / 3-4'	9.165 / 6.324	57.00	234.00		
		–	2.121	3-3'	-14.504	57.00	234.00		
		$^{4}CH_2$	2.431	3'-4 / 3'-4'	6.347 / 9.209	57.00	234.00		
		–	2.458	4-4'	-15.371	57.00	234.00		
		$^{5}NH_2$	6.816	–	–	57.00	234.00		
		$^{2}NH_2$	7.529	–	–	57.00	234.00		
Glutamate	–	^{2}CH	3.748	2-3 / 2-3'	7.331 / 4.651	50.00	219.580	2.692	20.774
		$^{3}CH_2$	2.046	3-3'	-14.849	50.00	219.580		
		–	2.120	3-4 / 3-4'	6.413 / 8.406	50.00	219.580		
		$^{4}CH_2$	2.334	3'-4 / 3'-4'	8.478 / 6.875	50.00	219.580		
		–	2.352	4-4'	-15.915	50.00	219.580		
Glycine	–	$^{2}CH_2$	3.547	–	–	28.00	206.00	0.820	2.395
Glycerol	–	$^{1}CH_2$	3.552	1-2	4.427	–	–	–	–
		–	3.640	1-1'	-11.715	–	–		
		^{2}CH	3.770	1'-2 / 2-3'	6.485 / 6.485	–	–		
		$^{3}CH_2$	3.552	2-3	4.427	–	–		
		–	3.640	3-3'	-11.715	–	–		
Glycerophospho-choline	Glycerol moiety	$^{1}CH_2$	3.605	1-2 / 1'-2	5.7700 / 4.5300	101.00	371.00	0.033	10.720
		–	3.672	1-1'	-14.78	101.00	371.00		
		^{2}CH	3.903	2-3 / 2-3'	5.77 / 4.53	101.00	371.00		
		$^{3}CH_2$	3.871	3-3'	-14.78	101.00	371.00		
		–	3.946	3,3'-P	6.03 / 6.03	101.00	371.00		
	Phospho-choline moiety	$^{1}CH_2$	4.312	1-2 / 1-2'	3.10 / 5.90	124.00	396.00		
		–	–	1'-2 / 1'-2'	5.90 / 3.10	124.00	396.00		
		$^{2}CH_2$	3.659	1-1' / 2-2'	2.67 / 2.67	124.00	396.00		
		–	–	1,1'-P	6.03 / 6.03	124.00	396.00		
		$N(CH_3)_3$	3.212	–	–	124.00	396.00		
Glutathione	Glycine moiety	$^{2}CH_2$	3.769	–	–	38.00	188.360	0.118	3.804
		NH	7.154	–	–	38.00	188.360		
	Cysteine moiety	^{2}CH	4.561	2-3	7.09	33.00	188.360		
		$^{3}CH_2$	2.926	2-3'	4.71	33.00	188.360		
		–	2.975	3-3'	-14.06	33.00	188.360		
		NH	8.177	–	–	33.00	188.360		
	Glutamate moiety	^{2}CH	3.769	2-3 / 2-3'	6.34 / 6.36	34.00	225.00		
		$^{3}CH_2$	2.159	3-4 / 3-4'	6.7 / 7.6	34.00	225.00		
		–	2.146	3-3'	-15.48	34.00	225.00		
		$^{4}CH_2$	2.510	3'-4 / 3'-4'	7.6 / 6.7	34.00	225.00		
		–	2.560	4-4'	-15.92	34.00	225.00		
Homocarnosine	GABA moiety	$^{2}CH_2$	2.969	2-3 / 2-3'	8.00 / 8.00	–	–	0.00	0.440
		–	2.944	2-2'	-12.500	–	–		
		$^{3}CH_2$	1.896	3-3'	-13.900	–	–		
		–	1.881	2-3' / 2'-3	7.500 / 7.500	–	–		
		$^{4}CH_2$	2.378	3'-4 / 3'-4'	7.500 / 7.500	–	–		
		–	2.348	3-4 / 3-4'	7.500 / 7.500	–	–		
		–	–	4-4'	-15.200	–	–		
	Alanine moiety	^{2}CH	4.467	2-3	5.020	–	–		
		$^{3}CH_2$	3.191	3-3'	8.640	–	–		
		–	3.013	3-3'	-15.300	–	–		
	Imidazole moiety	^{5}CH	7.08s	4-NH	0.580	–	–		
		^{2}CH	8.08s	2-NH	1.130	–	–		

(continued on the next page)

Peer Review Only

Name	Moieties	Group	Chemical Shift (ppm)[2,3]	Interaction	Scalar coupling (Hz)[2,3]	T_2 [ms]		Conc. [mM]	
						Low	High	Low	High
Lactate	–	^{2}CH	4.097	2-3	6.930	61.00	226.520	0.00	15.824
		^{3}CH$_3$	1.313	–	–	61.00	226.520		
myo-Inositol	–	^{1}CH	3.522	1-2	2.889	19.00	439.00	1.380	23.393
		^{2}CH	4.054	2-3	3.006	19.00	439.00		
		^{3}CH	3.522	3-4	9.997	19.00	439.00		
		^{4}CH	3.614	4-5	9.485	19.00	439.00		
		^{5}CH	3.269	5-6	9.482	19.00	439.00		
		^{6}CH	3.614	1-6	9.998	19.00	439.00		
N-acetylaspartate	Acetyl moiety	^{2}CH$_3$	2.008	–	–	125.00	411.00	2.465	25.344
	Aspartate moiety	^{2}CH	4.382	2-3	3.861	109.00	376.00		
		^{3}CH$_2$	2.673	2-3'	9.821	109.00	376.00		
		–	2.486	3-3'	-15.592	109.00	376.00		
		NH	7.820	2-NH	6.400	242.700	320.170		
N-acetylaspartyl-glutamate	Glutamate moiety	^{2}CH	4.128	2-3 / 2-3'	7.330 / 4.650	36.00	216.110	0.119	3.182
		^{3}CH$_2$	1.881	3-3'	-14.850	36.00	216.110		
		–	2.049	3-4 / 3-4'	6.410 / 8.410	36.00	216.110		
		^{4}CH$_2$	2.190	3'-4 / 3'-4'	8.480 / 6.880	36.00	216.110		
		–	2.180	4-4'	-15.920	36.00	216.110		
		NH	7.950	2-NH	–	36.00	216.110		
	Acetyl moiety	^{2}CH$_3$	2.042	–	–	69.00	229.00		
	Aspartate moiety	^{2}CH	4.607	2-3 / 2-3'	4.412	27.00	234.00		
		^{3}CH$_2$	2.721	2-3'	9.515	27.00	234.00		
		–	2.519	3-3'	-15.910	27.00	234.00		
		NH	8.260	2-NH	–	27.00	234.00		
Phenylalanine	Alanine moiety	^{2}CH	3.975	2-3	5.209	–	–	–	–
		^{3}CH$_2$	3.273	2-3'	8.013	–	–		
		–	3.105	3-3'	-14.573	–	–		
	Phenyl moiety	^{2}CH	7.322	2-3 / 2-4	7.909 / 1.592	–	–		
		^{3}CH	7.420	2-3' / 2-2'	0.493 / 1.419	–	–		
		^{4}CH	7.369	3-4 / 3-3'	7.204 / 0.994	–	–		
		^{5}CH	7.420	2'-3 / 3'-4	0.462 / 7.534	–	–		
		^{6}CH	7.322	2'-4 / 2'-3'	0.970 / 7.350	–	–		
Phosphoryl-choline	–	^{1}CH$_2$	4.282	1-2 / 1-2'	2.284 / 7.231	93.00	345.00	0.007	4.288
		–	–	1'-2 / 1'-2'	7.326 / 2.235	93.00	345.00		
		^{2}CH$_2$	3.643	1,1'-N	2.680 / 2.772	93.00	345.00		
		–	–	1-1' / 2-2'	-14.89 / -14.19	93.00	345.00		
		–	–	1,1'-P	6.298 / 6.249	–	–		
		N(CH$_3$)$_3$	3.209	–	–	119.00	394.00		
Phospho-creatine	–	N(CH$_3$)	3.029	–	–	130.00	210.00	0.779	10.123
		^{2}CH$_2$	3.930	–	–	100.00	180.00		
		^{1}NH	6.58s	–	–	130.00	210.00		
		^{2}NH	7.30s	–	–	130.00	210.00		
Phosphoethanol-amine	–	^{1}CH$_2$	3.977	1-2 / 1-2'	3.182 / 6.716	34.00	250.00	0.00	2.200
		–	–	1'-2 / 1'-2'	7.204 / 2.980	34.00	250.00		
		^{2}CH$_2$	3.216	1-1' / 2-2'	-14.560 / -14.710	34.00	250.00		
		–	–	1,1'-P	7.288 / 7.088	34.00	250.00		
		–	–	1,1'-N	0.464 / 0.588	–	–		
Pyruvate	–	^{3}CH$_3$	2.358	–	–	100	250	0.090	0.550
scyllo-Inositol	–	$^{1-6}$CH	3.340	–	–	73.00	250.00	0.00	0.563

(continued on the next page)

Peer Review Only

Name	Moieties	Group	Chemical Shift (ppm)[2,3]	Interaction	Scalar coupling (Hz)[2,3]	T_2 [ms]		Conc. [mM]	
						Low	High	Low	High
Serine	–	^{2}CH	3.835	2-3	5.979	–	–	0.00	2.200
		^{3}CH$_2$	3.937	2-3'	3.561	–	–		
		–	3.976	3-3'	-12.254	–	–		
Succinate	–	$^{2\text{-}3}$CH$_2$	2.394	–	–	–	–	0.090	0.550
Taurine	–	^{1}CH$_2$	3.420	1-2 / 1-2'	6.740 / 6.460	66.00	231.140	0.00	6.600
		–	–	1-1'	-12.438	66.00	231.140		
		^{2}CH$_2$	3.246	1'-2 / 1'-2'	6.400 / 6.790	66.00	231.140		
		–	–	2-2'	-12.930	66.00	231.140		
Threonine	–	^{2}CH	3.578	2-3	4.917	–	–	0.00	0.550
		^{3}CH	4.246	3-4	6.350	–	–		
		^{4}CH$_3$	1.316	–	–	–	–		
Tryptophan	Alanine moiety	^{2}CH	4.047	2-3	8.145	–	–	–	–
		^{3}CH$_2$	3.475	2-3'	4.851	–	–		
		–	3.290	3-3'	-15.368	–	–		
	Indole moiety	^{2}CH	7.312	–	–	–	–		
		^{4}CH	7.726	4-5 / 4-6	7.600 / 1.00	–	–		
		^{5}CH	7.278	4-7	0.945	–	–		
		^{6}CH	7.197	5-6 / 5-7	7.507 / 1.200	–	–		
		^{7}CH	7.536	6-7	7.677	–	–		
Tyrosine	Alanine moiety	^{2}CH	3.928	2-3	5.147	–	–	–	–
		^{3}CH$_2$	3.192	2-3'	7.877	–	–		
		–	3.037	3-3'	–	–	–		
	Phenol moiety	^{2}CH	7.186	2-2' / 2-3	2.538 / 7.981	–	–		
		^{3}CH	6.890	3-3'	2.445	–	–		
		^{5}CH	6.890	2-3'	0.311	–	–		
		^{6}CH	7.186	2'-3 / 2'-3'	0.460 / 8.649	–	–		
Valine	–	^{2}CH	3.595	2-3	4.405	–	–	–	–
		^{3}CH	2.259	3-4	6.971	–	–		
		^{4}CH$_3$	1.028	3-4'	7.071	–	–		
		$^{4'}$CH$_3$	0.977	–	–	–	–		
Water	–	H$_2$0	4.650	–	–	–	–	–	–

TABLE S2.1 Table of recommended T_2 values and concentration ranges covering physiological and pathological ranges for selected brain metabolites. The T_2 values can vary by moiety within a metabolite, so they were matched with their respective spin systems. When basis sets consist of individual spins instead of metabolite-level basis functions, it will be necessary to align the spins to correspond properly with the developed parameter sampling scheme.

Peer Review Only

3 | BASELINE AND RESIDUAL WATER SIMULATIONS

A single simulation protocol is used to generate the contributions for both baseline offsets and residual water contributions. The capability of generating signal contributions with such widely varying profiles is entirely dependent on the underlying parameters. In order to better understand the capabilities of the proposed generator and its behavior, the following section will explore the effects the various parameters have on the final signal contributions that are generated. For a detailed explanation of how these sampled parameters are converted into their final outputs, please refer to Algorithm 1 in Sec. 2.1.5 of the main manuscript.

This generator uses a smoothed, pseudo-random bounded walk algorithm for the simulations. For simplicity, this will henceforth be referred to as a *(random) bounded walk*, or simply a *walk*. By using different configuration dictionaries, it is possible to switch from very smooth, undulating baselines to rather erratic and rough residual water regions. The following sections highlight the effects different parameters have on the simulations.

3.1 | Variables

These simulations use at least 9 degrees of freedom with two additional optional inputs:

1. Starting height (start)
2. Ending height (end)
3. Standard deviation of walk (std)
4. Lower bound (lower)
5. Upper bound (upper)
6. Point density (pt_density)
7. PPM range (ppm_range)
8. Length of smoothing window (window)
9. Scale (scale)
10. Dropout probability (drop_prob)
11. *Prime (prime)

Floats or integers can be provided for the point density, prime, and drop_prob. Everything else should be provided as a list with either a single value or a range in the form of [min_range, max_range]. When a single value is provided, that variable will be fixed. When a range is specified, then values will be sampled from that range uniformly. The point density and the PPM range are used to calculate the length of the walks. This length coupled with std control the flexibility of the raw simulations. The length of the smoothing window, $window$, the magnitude of the std, and scaling factor, $scale$, of the baseline will affect the smoothness of the resulting walk. It is specified as a fraction of the length of the walk, so that it is independent of walk length, and can also be either fixed or sampled. $scale$ determines how prominent each offset will be when added to the simulated data. $drop_prob$ is used to randomly omit the offset from that percentage of the simulations. $prime$ [units: ppm] is used for generating the residual water regions. This allows the region to vary in length by sampling two values in the range of $[-prime, prime)$ that can expand, contract, or slightly offset the residual water region.

3.2 | Post-processing

Once the random walks have been generated, they are then smoothed with kernels of either fixed or varying width. At this point, trend lines are calculated between the starting and ending points which are then removed so that they start and end on the x-axis. Afterwards, the walks are normalized to [0,1] and then scaled down according to $scale$. A function called *sim2acquired* then uses zero padding and a nonuniform interpolator to add tails to both sides of the walks so that they match the PPM range of the basis set regarding the spectral width and carrier frequency. Before being added to the FIDs, they are multiplied by the maximum value of the FID in the frequency domain to scale them up to the correct order of magnitude. This makes $scale$ relative to the maximum height in the spectra. The offsets are added to the FIDs in the frequency domain before an inverse FFT returns the data to the time domain.

The parameters presented below control the length of the walks, their trend lines, flexibility, and smoothness. The default values presented in the configuration dictionaries were determined through preliminary experiments to closely approximate what was observed in a clinical dataset.

Peer Review Only

3.3 | Preparing for the simulations

The following section defines parameters for the spectroscopy scenario and the configuration dictionary for the simulations.

```
# Spectroscopy scenario
spectralwidth = 2000                    # Hz
Ns            = 2048                    # number of spectral points
B0            = 3.0                     # T
gamma_H       = 42.577478518            # MHz/T
ppm_ref       = 4.65                    # ppm

carrier_frequency = B0 * gamma_H        # MHz
ppm = torch.linspace(-0.5*spectralwidth,
                     0.5*spectralwidth,Ns) # Hz
ppm /= carrier_frequency                # ppm
ppm += ppm_ref
t = torch.linspace(0,Ns/spectralwidth,Ns)

num_samples = 10
```

3.4 | Understanding the Simulation Plots

The plots below contain either one or two lines that are blue or red. The default color scheme is as follows:

blue raw random walk simulation

red smoothed walk that is added to the simulated spectra

In Sections 3.5.4 and 3.6.5, which discuss how to incorporate the walks into the simulated spectra, only a single blue line is plotted. This line is the smoothed, real component of the walk. In Sections 3.5.5 and 3.6.6 which show the effect of applying the Hilbert transform, the blue and red lines represent the real and imaginary components, respectively. Similarly, Sections 3.5.6 and 3.6.7, which include examples of random simulations for each contribution type, also show the real and imaginary components using blue and red, respectively.

3.5 | Baseline Offsets

The following section will explore this generator using the baseline configuration dictionary defined below. The values selected for the plots in the following sections were tailored specifically to the baseline simulations.

```
baseline_cfg = {
        "start":          [   -1,     1],
        "end":            [   -1,     1],
        "upper":          [        1],
        "lower":          [       -1],
        "std":            [ 0.05,  0.20],
        "window":         [ 0.15,   0.3],
        "pt_density":        128,
        "ppm_range":      [ -1.6,   8.5],
        "scale":          [    0,   1],
        "drop_prob":         0.0
}
```

3.5.1 | Standard Deviation

This section shows the effect of different std values on the random walk. The std variable is used when sampling the noise for the walk. It directly controls the amount of variation between two consecutive points prior to the cumulative summation.

FIGURE S3.1 Standard Deviation :: STD = 0.05; Window length = random; Kernel_size = random; Point density = 128

FIGURE S3.2 Standard Deviation :: STD = 0.10; Window length = random; Kernel_size = random; Point density = 128

FIGURE S3.3 Standard Deviation :: STD = 0.15; Window length = random; Kernel_size = random; Point density = 128

FIGURE S3.4 Standard Deviation :: STD = 0.20; Window length = random; Kernel_size = random; Point density = 128

3.5.2 | Smoothing Kernel

In this section, the effects of different smoothing kernel lengths are explored given a fixed *std* value.

FIGURE S3.5 Smoothing Kernel :: STD = 0.10; Window Length = 0.10, Kernel_size = 60; Point density = 128

FIGURE S3.6 Smoothing Kernel :: STD = 0.10; Window Length = 0.20, Kernel_size = 120; Point density = 128

FIGURE S3.7 Smoothing Kernel :: STD = 0.10; Window Length = 0.30, Kernel_size = 180; Point density = 128

3.5.3 | Point Density

This section explores the effect of varying the point density of the random walk while keeping the *std* fixed. Results are presented using two different kernel sizes for the smoothing.

FIGURE S3.8 Point Density :: STD = 0.10, Window Length = [0.10, 0.30], Kernel_size = [3,9]; Point density = 64

FIGURE S3.9 Point Density :: STD = 0.10, Window Length = [0.10, 0.30], Kernel_size = [3,9]; Point density = 128

FIGURE S3.10 Point Density :: STD = 0.10, Window Length = [0.10, 0.30], Kernel_size = [3,9]; Point density = 256

3.5.4 | Incorporating the offsets into the spectra

To increase variability, the starting and ending heights are randomly selected. When considering the entire spectrum, however, there are two options to avoid unrealistic, hard transition points. These variables can be manually adjusted or the trend line can be removed, which is how the current code works.

Peer Review Only

FIGURE S3.11 With the original trend lines still included

STD = 0.10, K_Size = 2068, W_len = 0.10
STD = 0.10, K_Size = 6205, W_len = 0.30
STD = 0.10, K_Size = 2068, W_len = 0.10
STD = 0.10, K_Size = 6205, W_len = 0.30
STD = 0.10, K_Size = 2068, W_len = 0.10
STD = 0.10, K_Size = 6205, W_len = 0.30
STD = 0.10, K_Size = 2068, W_len = 0.10
STD = 0.10, K_Size = 6205, W_len = 0.30
STD = 0.10, K_Size = 2068, W_len = 0.10
STD = 0.10, K_Size = 6205, W_len = 0.30
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
5
4
3
2
1
0

FIGURE S3.12 With the original trend lines removed

3.5.5 | Effect of the Hilbert Transform

Simulating spectra requires complex spectral components, including the baseline and residual water. The Hilbert transform is used to generate those corresponding imaginary components.

FIGURE S3.13 Compiled samples with their Hilbert pair

3.5.6 | Compiled Generator

This shows the variety of baselines that can be generated when randomly sampling all variables.

FIGURE S3.14 Compilation of randomly generated samples with their Hilbert pairs

3.6 | Residual Water

The following section will explore this generator using the residual water dictionary defined below. While the sections are the same as in Sec. 3.5, the values displayed are tailored to the ranges for the residual water simulations.

3.6.1 | Configuration dictionaries

The following dictionary entries are the standard defaults in MRS-Sim for generating residual water contributions.

```
resWater_cfg = {
        "start":          [           0],
        "end":            [           0],
        "upper":          [     0,     1],
        "lower":          [     0,     1],
        "std":            [   0.2,  0.40],
        "window":         [  0.05,  0.15],
        "pt_density":       1204,
        "ppm_range":      [   4.4,   4.9],
        "prime":            0.15,
        "scale":          [   1.0,   1.0], # typically: [0.05, 0.20] - better visualization.
        "drop_prob":         0.0
}
```

Review Only

3.6.2 | Standard Deviation

This section shows the effect of different std values on the random walk. The std variable is used when sampling the noise for the walk. It directly controls the amount of variation between two consecutive points prior to the cumulative summation.

FIGURE S3.15 Standard Deviation :: STD = 0.20; Window length = random; Kernel_size = random; Point density = 1204

FIGURE S3.16 Standard Deviation :: STD = 0.30; Window length = random; Kernel_size = random; Point density = 1204

FIGURE S3.17 Standard Deviation :: STD = 0.40; Window length = random; Kernel_size = random; Point density = 1204

FIGURE S3.18 Standard Deviation :: STD = 0.50; Window length = random; Kernel_size = random; Point density = 1204

3.6.3 | Smoothing Kernel

In this section, the effects of different smoothing kernel lengths are explored given a fixed *std* value.

FIGURE S3.19 Smoothing Kernel :: STD = 0.10, Window Length = 0.05, Kernel_size = 30; Point density = 1204

FIGURE S3.20 Smoothing Kernel :: STD = 0.10, Window Length = 0.10, Kernel_size = 60; Point density = 1204

FIGURE S3.21 Smoothing Kernel :: STD = 0.10, Window Length = 0.15, Kernel_size = 90; Point density = 1204

3.6.4 | Point Density

This section explores the effect of varying the point density of the random walk while keeping the *std* fixed. Results are presented using two different kernel sizes for the smoothing.

FIGURE S3.22 Point Density :: STD = 0.10, Window Length = [0.05, 0.15], Kernel_size = [12,38]; Point density = 512

FIGURE S3.23 Point Density :: STD = 0.10, Window Length = [0.05, 0.15], Kernel_size = [12,38]; Point density = 1024

FIGURE S3.24 Point Density :: STD = 0.10, Window Length = [0.05, 0.15], Kernel_size = [12,38]; Point density = 2048

3.6.5 | Incorporating the offsets into the spectra

To increase variability, the starting and ending heights are randomly selected. When considering the entire spectrum, however, these must be adjusted to avoid unrealistic, hard transition points. As mentioned above, this is achieved by removing the trend lines.

FIGURE S3.25 With the original trend lines still included

FIGURE S3.26 With the original trend lines removed

3.6.6 | Effect of the Hilbert Transform

Simulating spectra requires complex spectral components, including the baseline and residual water. The Hilbert transform is used to generate those corresponding imaginary components.

FIGURE S3.27 Compiled samples with their Hilbert pair

3.6.7 | Compiled Generator

This shows the variety of residual water contributions that can be generated when randomly sampling all variables.

FIGURE S3.28 Compilation of randomly generated samples with their Hilbert pairs

References

1. Li N, An L, Shen J. Spectral fitting using basis set modified by measured B0 field distribution. *NMR in Biomedicine* 2015; 28(12): 1707–1715. doi: 10.1002/nbm.3430

2. Graaf dRA. *In Vivo NMR Spectroscopy: Principles and Techniques*. Wiley. 2nd ed. 2007

3. Graaf dRA. *In Vivo NMR Spectroscopy: Principles and Techniques*. Wiley. 3rd ed. 2019

4. Govindaraju V, Young K, Maudsley AA. Proton NMR chemical shifts and coupling constants for brain metabolites. 2000: 129–153.

5. Govindaraju V, Young K, Aa M, et al. Corrigendum : Proton NMR chemical shifts and coupling constants for brain metabolites .. 2015(May): 923–924. doi: 10.1002/nbm.3336

6. Landheer K, Swanberg KM, Juchem C. Magnetic resonance Spectrum simulator (MARSS), a novel software package for fast and computationally efficient basis set simulation. *NMR in Biomedicine* 2021; 34(5): 1–13. doi: 10.1002/nbm.4129

7. Near J, Andersson J, Maron E, et al. Unedited in vivo detection and quantification of gamma-aminobutyric acid in the occipital cortex using short-TE MRS at 3 T. 2013(March): 1353–1362. doi: 10.1002/nbm.2960

8. Wyss PO, Bianchini C, Scheidegger M, et al. In Vivo Estimation of Transverse Relaxation Time Constant (T2) of 17 Human Brain Metabolites at 3T. *Magnetic Resonance in Medicine* 2018; 80(2): 452–461. doi: 10.1002/mrm.27067

9. Gudmundson AT, Koo A, Virovka A, et al. Meta-analysis and open-source database for in vivo brain Magnetic Resonance spectroscopy in health and disease. *Analytical Biochemistry* 2023; 676(July): 115227. doi: 10.1016/j.ab.2023.115227

10. Hamilton G, Yokoo T, Bydder M, et al. In vivo characterization of the liver fat 1 H MR spectrum. 2011(December 2010): 784–790. doi: 10.1002/nbm.1622

11. Wermter FC, Mitschke N, Bock C, Dreher W. Temperature dependence of 1H NMR chemical shifts and its influence on estimated metabolite concentrations. *Magnetic Resonance Materials in Physics, Biology and Medicine* 2017; 30(6): 579–590. doi: 10.1007/s10334-017-0642-z

12. Choi C, Ganji SK, Deberardinis RJ, et al. 2-hydroxyglutarate detection by magnetic resonance spectroscopy in subjects with IDH -mutated gliomas. 2012; 18(4). doi: 10.1038/nm.2682

How to cite this article: LaMaster J., G. Oeltzschner, and Y. Li (2022), MRS-Sim: Open-Source Framework for Simulating Realistic In Vivo-like MR Spectroscopy Data, *NMR in Biomedicine*.

For Peer Review Only

Received: Added at production | Revised: Added at production | Accepted: Added at production

DOI: xxx/xxxx

ARTICLE TYPE

MRS-Sim: Open-Source Framework for Simulating In Vivo-like Magnetic Resonance Spectra

John LaMaster*[1,2] | Georg Oeltzschner[3] | Yan Li[4]

[1]Munich Institute of Biomedical Engineering, Technical University of Munich, Bavaria, Germany
[2]School of Computation, Information, and Technology, Technical University of Munich, Bavaria, Germany
[3]Russel H. Morgan Department of Radiology and Radiological Sciences, The Johns Hopkins' University School of Medicine, Maryland, USA
[4]Department of Radiology and Biomedical Imaging, University of California, San Francisco, California, USA

Correspondence
John LaMaster.
Email: john.t.lamaster@gmail.com
Twitter: @JohnTLaMaster

Present Address
Munich Institute of Biomedical Engineering, School of Computation, Information, and Technology, Technical University of Munich in the City of Garching bei München, Boltzmannstr 11. 85748 Garching bei München, Bavaria, Germany.

Summary

MRS-Sim is a powerful, open-source framework for simulating in vivo-like magnetic resonance spectroscopy data. Its modularity allows it to model the complexities of MRS data for a variety of clinical scenarios. The underlying physical equations include spectral components commonly found in linear-combination fitting routines plus some novel components. The first is a 3D B_0 field map simulator to model B_0 field inhomogeneities, ranging from slight variations to severe distortions. The second is a novel semi-parametric generator that mimics signals from poorly characterized residual water regions and baseline offset contributions. This framework can simulate everything from raw multi-coil transients to preprocessed coil-combined data.

A repository of information has been compiled to help non-expert users with general simulations of MRS data. Simulating realistic clinical-like datasets requires appropriate ranges and distributions for the model parameters, which are best determined by analyzing the parameter estimates from fitting real clinical data. Therefore, accompanying tools within MRS-Sim allow for analyzing the range and distributions of those parameters from in vivo datasets fitted with Osprey. This allows simulations to be tailored to specific clinical datasets.

The modularity of this framework allows for easy customization of simulations to suit a variety of clinical scenarios and encourages continued community development. Using a single framework for various scenarios overcomes the inconsistent protocols in current literature. By simulating in vivo-like data, this framework can aid in many tasks in MRS, including verifying spectral fitting protocols and reproducibility analyses. The accessibility of readily available, synthetic data will also assist deep learning research for MRS, particularly in cases where clinical data is insufficient or unavailable for training and, in general, will help address reproducibility issues and make MRS research more accessible to a wider audience.

Word count (abstract): 295 -> 297 -> 282 words

Word count (*paper): 4,206 -> 5,125 -> 6,314 words

*Limited to intro, methods, results, and conclusions. Including auxiliary text (author info, etc.): 4,807 -> 5,757 -> 6,805 words

KEYWORDS:
MRS, spectroscopy, in vivo, data simulation, synthetic data, open-source

1 | INTRODUCTION

Magnetic resonance spectroscopy (MRS) is a non-invasive imaging modality that provides in vivo information on the metabolic profile of tissues. It allows for evaluating various pathologies throughout the entire body by analyzing the spectra to quantify metabolite concentrations. 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].

In order for MRS to be a widely used clinical imaging modality, it is important to ensure the precision and accuracy of MRS data processing and analysis. However, it can be challenging due to variations across scenarios, such as pulse sequence, acquisition parameters, data quality regimes, and biological differences (adult versus pediatric subjects [18] and disease related metabolic patterns [19]). Such comprehensive collections of data are usually not widely available for method development and evaluation due to data privacy restrictions. Furthermore, in vivo data do not have access to ground truth values, making it difficult to determine the reliability, sensitivity, and robustness of different MRS analysis protocols. Phantom data can be useful in developing and validating acquisition and analysis methods, but it usually does not adequately reflect the complexities of in vivo spectra.

A logical solution to these challenges is to generate synthetic data for a variety of experimental scenarios. However, generating synthetic MRS data can be challenging as well, as it requires adequately incorporating all physical phenomena underlying in vivo data. One of the significant aspects is to choose adequate model parameters that reflect different physiological and pathological conditions. Arguably more important is the need to capture acquisition-induced artifacts and nuisance signals commonly found in in vivo data that are difficult to reproduce in a phantom, such as effects of susceptibility and field inhomogeneity, macromolecule, lipid, and residual water signals, and various other artifacts. Many research groups now routinely use synthetic data, but the data generation models and parameter distributions are rarely made publicly available. This lack of transparency hampers systematic comparison and the formation of consensus best practices for synthetic data generation.

It is evident from the literature that there is no standard set of best practices for simulating synthetic data. Simulations have been carried out using physical models with a range of complexity and spectral components [20,21,22,23]. Most of these methods begin with simulating basis functions, assuming appropriate pulse sequence parameters for their specific scenario. These functions represent metabolites that can be collected and are then modulated by scaling factors representing their underlying concentrations. A simple Lorentzian lineshape is generally applied [20,21,22] with optional phase offsets [22,23] and frequency shifts [20]. Finally, some type of broad baseline is typically a[illegible] These models are simple and do not fully capture the complexity of in vivo data. Additionally, they often contain non-public components such as baselines, macromolecules, and lipid signals that are either simulated in-house or extracted from private datasets, thus making detailed analysis of their models diff[illegible]

There are several software packages that claim to simulate in vivo spectra, such as FID-A [24], GAVA [25], and VESPA [26]. However, the fundamental difference between these methods and the current work is the definition of "in vivo spectra". Previously, this term referred to simulating idealized spectra to evaluate parameter settings for developing new in vivo pulse sequences. In this work, "in vivo spectra" refers to the data that comes off a clinical scanner ~~before~~ prior to any post-processing ~~and analysisbegins~~or analysis. This new definition is meant to reframe the task of simulating synthetic data, to challenge what are assumed to be acceptable simplifications, and to advance the use of synthetic data for MRS applications. While pulse sequences are necessary for generating basis sets, they are left to the many software options already available. The most relevant previous work is the recent update for the MARSS [27] software which has added a new functionality called synMARSS that simulates synthetic spectra and incorporates many spectral artifacts. They provide an extensive user manual with equations describing their implementation~~. However~~, however, their source code is not available, ~~nor can it be~~ meaning it cannot be evaluated or modified like M[illegible]sim.

Over the past few years, there has been a significant rise in the use of synthetic MRS data. This increase is primarily due to the emergence of machine learning [28] (ML) and deep learning [29,20,30] (DL) analysis algorithms that require vast amounts of data for training and validation that few centers worldwide possess. Therefore, generating realistic, in vivo-like synthetic data is necessary to facilitate more research into these applications. Synthetic data has known ground truth values that are essential to evaluate the accuracy and precision of the analysis methods being developed. However, 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. [31,32,33,34] To address this challenge, it is essential to establish standards and best practices for simulating synthetic MRS data. This will require researchers to collaborate and build consensus, as well as develop and adopt open-source frameworks for data generation.

This work presents MRS-Sim: a framework for a modular synthetic data generation model to address some of these challenges. The basic simulation model uses 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 B_0 map generator to simulate in vivo-like field heterogeneity conditions, and uses ~~parameterized~~ semi-parameterized models to describe residual water and smooth background signal contributions. ~~The software~~ MRS-Sim 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 compare new data processing methods, not just modeling algorithms. ~~The framework~~

~~And~~ the implemented code base is highly flexible and can simulate custom-tailored datasets for a large variety of clinical scenarios. This framework includes a detailed compilation of simulation parameters as well as tools to analyze existing in vivo datasets and extract parameter distributions, which is useful 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~~framework'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

The proposed physics model was developed ~~to mirror the actual data acquisition sequence by reverse engineering the spectral fitting process. This process dictated which steps to include and order of operations, which are meant to ensure that any experimental fitting parameters will match the simulation parameters. In order to maximize generalizability and usefulness, the data simulator needs to comprehensively model known spectral components, which were identified through a review of~~ after a thorough review of current state-of-the-art fitting ~~techniques and currently available fitting algorithms. These spectral components allow the model to account~~ algorithms (LCModel[35], Osprey[36], FSL-MRS[37], etc.). Commonly modeled spectral features and components were identified for inclusion, allowing for a large variety of scenarios and artifacts. The ~~proposed physics model is described by the following set of equations~~order of application was further specified for different operations and along with appropriate ranges and distributions for the underlying model parameters. The model was parameterized as follows:

$$f(t) = \sum_{n}^{N} \underbrace{M_n * basisfcn_n}_{amplitude\ scaling} \underbrace{e^{-d_n t - g_n t^2}}_{lineshape} \underbrace{e^{-i\Delta f_n t}}_{fshift} \underbrace{e^{i(\phi_0+\phi_1)}}_{phase} \underbrace{e^{-iA_0 * e^{-t/tc_0} t * 2\pi}}_{eddyCurrents} * \underbrace{\sum_{r=1}^{R} e^{-i\Delta\omega_r t}}_{B_0\ inhomog.} + \mathcal{H}\left(\underbrace{baseline + resH_2O}_{nuisance\ signals}\right) + \underbrace{\mathcal{N}(0, snr_0, l)}_{noise} \tag{1a}$$

$$f(t) = \mathcal{C}oil_0^C\left(f(t), \underbrace{\overbrace{snr_c}^{noise}, \overbrace{sens_c}^{sensitivity}, \overbrace{\Delta f_c, \Delta\phi_c}^{drifts}}_{coil\ artifacts}\right) \tag{1b}$$

$$F(\omega) = \mathcal{F}\left(zeroFill\left(f(t) * \underbrace{e^{-t*T_L}}_{apodize}\right), len\right) \tag{1c}$$

where N is the whole set of metabolites being modeled, M_n is the scaling factor for metabolite n, the Lorentzian variable d_n and the Gaussian variable g_n combine to define a Voigt lineshape, t is time, and Δf_n is the metabolite-specific frequency shift. Global zero- and first-order phase offsets are added using ϕ_0 and ϕ_1 while eddy current effects are described using two variables as a function of time t: the amplitude, A_0, and the time constant, tc_0. In lieu of the Gaussian term, imperfect shimming and severe lineshape distortions associated with large susceptibility effects can be applied using $\Delta\omega_r$, which is the modeled B_0 at location r inside the voxel of interest. snr_0 is the desired SNR of the spectrum, while $baseline$ and $resH_2O$ 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 $\mathcal{C}oil$ generates C coil transients and applies a distribution of SNR values and coil weights using snr_c and $sens_c$. Frequency drifts and phase drifts are then added using Δf_c and $\Delta\phi_c$, respectively. When necessary, Eqn. 1c can apply apodization using T_L in Hz and the FIDs can be zero-filled to length len. Then the Fourier transform $\mathcal{F}$ can convert the FIDs to the frequency domain. The following sections are presented in order according to their implementation in the physics model and discuss the terms above in greater detail. A step-by-step visualization of this model is illustrated in Fig. 1.

2.1.1 | Basis Functions

MRI, and its derivatives, are spatially resolved imaging modalities. MR pixels and voxels represent 3D volumes with a spatial distribution, as shown in Fig. ??, and ~~needs~~ need to be accounted for in simulations. The importance of spatial localization led to selecting Landheer *et al.*'s Magnetic Resonance Spectrum Simulator (MARSS)[27] software package for simulating the default basis functions provided with this simulator. MARSS produces

Peer Review Only

FIGURE 1 This is a step-by-step progression through the physics model. The real and imaginary components are depicted in black and gray, respectively. The red line includes only the metabolites and the offsets from the preceding steps. ?? shows the final spectrum with all artifacts applied. ?? is the pre-processed spectrum with the phase and frequency shifts removed.

high-fidelity outputs by simulating 128 points in each direction, accurately capturing the spatial nature of the RF pulses and slice-selective gradients. Using vendor-specific pulse sequences, MARSS can simulate individual or summed spins for a large number of common brain metabolites including macromolecules and lipids.

2.1.2 | Amplitude Modulation

For this framework, comprehensive ranges including physiological and pathological values were adapted from the meta-analysis by Gudmundson *et al.*[38] and are presented in the appendix. Near[39] and Larsen[40] discuss how non-trivial the choice of reference metabolite is with respect to quantification and interpretability. Therefore, the default ranges are in units of ~~milimolar~~ millimolar (mM) which allows any reference to be used. More specific information can be found in the MARSS user manual.

2.1.3 | Lineshape Profiles

The Voigt lineshape profile is the most commonly used in spectral fitting as it most closely matches in vivo data. It is a combination of a Lorentzian and a Gaussian and is used in various fitting packages, such as LCModel[35], TARQUIN[41], jMRUI[42], and Osprey[36]. For completeness, it is also possible to specify either a purely Lorentzian or a purely Gaussian lineshape. Individual Lorentzian and Gaussian values can be assigned to each moiety or metabolite separately, depending on the desired level of granularity. At the most basic level, each metabolite is assigned a Lorentzian value while a single Gaussian value is applied to the set of metabolites and a second value can be applied to the macromolecules and lipids. Standard practice from the aforementioned software packages assign a single Lorentzian value to each metabolite, instead of each moiety. It is also possible to specify moiety-level ~~T2~~ T_2 relaxation values for selected metabolites from Wyss *et al.*[43]. ~~Information is~~ Information is supplemented by the meta-analysis from Gudmundson *et al.*[38]. As more metabolites are characterized, new information can be added to the model. A detailed table is provided in the appendix with the current information; however, an up-to-date digital version will be maintained in the repository.

2.1.4 | ~~B_0~~ B_0 Inhomogeneities

Small B_0 inhomogeneities are, in general, sufficiently modeled by the Gaussian term of the Voigt lineshape. However, to simulate more severe distortions from large inhomogeneities, a B_0 field volume needs to be modeled and applied to the basis functions. Fig. 2 illustrates the difference between small and severe B_0 inhomogeneities. Fig. ?? shows the normal, subtle B_0 changes across the volume of a spectroscopy voxel while Fig. ?? illustrates these high susceptibility effects. In such cases, spectra from these regions exhibit significant lineshape distortions that cannot be adequately characterized using idealized lineshape profiles. Fig. 3 shows the result of high susceptibility effects on lineshapes.

In general, this approach mirrors Li *et al.*[44] in which the basis functions are modulated by the B_0 field map, but in the implementation the B_0 field map is simulated rather than acquired. As with MARSS, Li *et al.*[44] suggest 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. Any cuboidal shape, rectangular or otherwise, can be modeled. The B_0 field is defined by four variables: $\pm dx$, $\pm dy$, $\pm dz$, and μ. dx, dy, and dz describe half of the change in B_0 in their respective axis from the voxel's center and μ is the mean of the entire voxel. A linear B_0 gradient is assumed, but ~~other shapes~~ different gradient profiles can be implemented ~~. The first section in the supplement~~ by adding new functions that manipulate how the $x-$, $y-$, and $z-$components are calculated when generating the B_0 field map. Section 1 in the Supporting Information provides a more detailed explanation of how the B_0 is simulated.

~~Intravoxel localization~~ ~~Intravoxel spatial distribution of B_0 with minor distortions~~ ~~Spatial distribution of B_0 with major distortions~~

(a) Intravoxel localization **(b)** Intravoxel spatial distribution of B 0 **(c)** Spatial distribution of intravoxel magnetic susceptibility effects

FIGURE 2 Voxels are often treated as singularities, meaning the entire volume is considered to be a single point like the red box in the center. Accounting for these intravoxel distributions leads to more accurate and more realistic simulations. ?? highlights the spatial nature of voxels while ?? and ?? show different levels of intravoxel B_0 inhomogeneity.

~~Spectral peaks without B_0 inhomogeneities ($\mu = 0$Hz)~~ ~~Spectral peaks with moderate B_0 inhomogeneities ($\mu = 75$Hz)~~ ~~Spectral peaks with severe B_0 inhomogeneities ($\mu = 175$Hz)~~

(a) Spectral peaks without B 0 inhomogeneities (μ = 0Hz) **(b)** Spectral peaks with moderate B 0 inhomogeneities (μ = 75Hz) **(c)** Spectral peaks with severe B 0 inhomogeneities (μ = 175Hz)

FIGURE 3 These samples show the effects that can be modeled using the 3D B_0 field simulator. In ??, the spectral peaks have a purely Lorentzian lineshape. In ??, the lineshape is now Voigtian because the Gaussian term has been added back by the 3D B_0 field simulator. In ??, severe heterogeneities are modeled which produce extremely broad line widths. All three plots use the same x- and y-axes. The observed offsets are caused by the line broadening.

FIGURE 4 Simulated samples of spectral baselines and residual water regions using the pseudo-random bounded walk generator. The black and red lines are the real and imaginary components of the smoothed offsets that are then returned and applied to the simulated spectra. ~~Simulated samples of spectral baselines and residual water regions using the pseudo-random bounded walk generator. The blue lines are the raw simulations. The red lines are the smoothed versions that are then returned and applied to the simulated spectra.~~

2.1.5 | Baseline and Residual Water

~~Currently, the underlying physical phenomena that induce spectral baseline offsets are~~ Currently, there is no physics-based model for simulating these offsets and it is unknown what a true baseline actually looks like. Theory and protocols regarding baseline modeling have been developed extensively, but the physical phenomena underlying the spectral baseline signal are still poorly understood. ~~In fact,~~ Therefore, different modeling

Algorithm 1 Smoothed Bounded Pseudo-Random Walk

Require: $start$: ppm start value, end: ppm end value, std: standard deviation of the random noise, $lower_bound$: lower bound of the walk, $upper_bound$: upper bound of the walk, $length$: length of the walk

Function SmoothedBoundedPseudoRandomWalk($start, end, std, lower_bound, upper_bound, length, window_size$)

$bounds = upper_bound - lower_bound$

Generate raw random walk

$rand =$ generate_random_array$(size = length, std = std)$

$rand =$ cumsum$(rand, dim = -1)$

Calculate the trend line

$rand_trend_lines =$ generate_trend_array$(start, end, length)$

Calculate the difference between the random steps and the trend line

$rand_deltas = rand - rand_trend_lines$

Normalize the delta array

$delta_range = bounds \div (rand_deltas.max() - rand_deltas.min())$

$rand_deltas = rand_deltas * delta_range$

Adjust delta array if it exceeds *bounds*

$rand_deltas = conditional_adjustment(rand_deltas)$

Calculate the final random walk

$random_walk = trend_lines + rand_deltas$

Smooth using a uniform smoothing kernel

$random_walk =$ average_smoothing$(random_walk, kernel_size = window_size * length)$

return $random_walk$

protocols can produce different baselines. Bazgir et al.[45] interpolate between selected minima in the spectra, Wilson et al.[46] use penalized B-splines to overparametize the problem space which is solved using an optimized smoothing constraint, while Osprey[36] uses unregularized B-splines spread every 0.4 ppm to enforce smoothness. Even though all of these techniques have been proved useful, in vivo data does not have known ground truths meaning there is no way to identify which method is the most accurate. Without a physics-based model~~for simulating these offsets.~~ , other methods of baseline simulation are necessary.

Similarly, the residual suppressed water region is also poorly characterized. ~~Therefore~~Because these phenomena are poorly understood, a naive random model can be used in conjunction with observed constraints to approximate what is expected 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 using fixed 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. ?? shows very broad, smooth lines while Fig. ?? 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~~Supporting Information Section 3 Baseline and Residual Water Simulations.

2.1.6 | 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.

2.1.7 | Phase Offsets

Zero-Order Phase

FIDs and spectra are complex data types consisting of real and imaginary components. 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~~ anti-symmetrical peaks which are both positive and negative. As shown in Fig. 5, non-zero degree offsets result in a mixture of absorption and dispersion spectra. In its current form, a phase offset is selected randomly and independently during parameter sampling and applied in the time domain using the following complex exponential:

$$f(t) = f(t) * e^{-i\phi_0 * (\pi/180)} \tag{2}$$

First-Order Phase

First-order phase, i.e. linear phase, is a frequency-dependent linear offset that emanates from a reference point, i.e. the center frequency referred to as ppm_ref, which is typically the water peak at 4.65ppm. A linear phase offset creates asymmetrical line shapes that grow larger as one moves away from the reference point. This is illustrated in Fig. 5 by comparing ?? and ??. Currently, the ppm_ref term is fixed to 4.65ppm while the ϕ_1 term is randomly sampled and then applied in the frequency domain using the following complex exponential:

$$F(\omega) = F(\omega) * e^{-i\phi_1 * (\pi/180) * (ppm - ppm_ref} \tag{3}$$

~~Spectrum with no phase offsets~~ ~~Spectrum with zero-order phase offset ($\phi_0 = 45°$)~~

2.1.8 | Frequency Shifts

During data acquisition, the FID experiences a global frequency shift. However, some functional groups can experience individual frequency shifts which are attributed to effects such as temperature and pH. Similar to Sec. 2.1.3, the bare minimum requires separate global frequency shifts for the metabolites and nuisance signals. However, this model also allows each basis function to have an independent frequency shift in addition to the global shift which is in line with common spectral fitting protocols. For more in vivo-like, realistic spectra, values can be used from the work by Wermter *et al.*[47] which characterized the temperature-induced frequency shift of several 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. Currently available data will be included in the table in the ~~appendix.~~ Supplemental Information Section 2.

Peer Review Only

~~Spectrum with first-order phase offset ($\phi_1 = 20°$)~~

(a) Spectrum with no phase o sets **(b)** Spectrum with zero-order phase o set ($_0$ = 45) **(c)** Spectrum with rst-order phase o set ($_1$ = 20)

FIGURE 5 These samples show how identical spectra are affected by zero- and first-order phase offsets. In ??, the real component (black) is in absorption mode exhibiting narrow line widths and is fully positive. In ??, a zero-order phase offset is applied. As the spectrum shifts from absorption to dispersion mode, the peaks uniformly lose their symmetry and negative values from the imaginary component are transferred to the real component. In ??, a first-order phase shift is applied. This is evident because the asymmetry increases across the spectrum and emanates from the water peak.

~~Eddy current amplitude = 1.0 Eddy current amplitude = 3.0 Eddy current amplitude = 5.0~~

(a) Eddy current amplitude = 1.0 **(b)** Eddy current amplitude = 3.0 **(c)** Eddy current amplitude = 5.0

FIGURE 6 These 3 samples show the effect of eddy currents on MRS spectra to various degrees. The strength of the eddy currents increases from ?? to ??. If the time constant, tc, is set too long, the eddy current artifact will appear as a global frequency shift. In these examples, it can be seen that only some frequencies are affected.

2.1.9 | Eddy Currents

Eddy current correction techniques, such as the Klose[48], 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.

2.1.10 | Multi-Coil Transients

In this work, the term *multi-coil transients*, or simply *transients*, refers to the signal from multiple individual RF coils measured from a phased-array receiver coil for a single average prior to coil combination. The terminology consensus paper by Kreis et al.[49] provides many clarifications, but does not specify how to differentiate between data before and after coil-combination nor specifically when a transient becomes a spectrum. For clarity and consistency, *spectrum* and *FID* will refer to data that has been coil-combined. When simulating transients, additional considerations need to be included. During acquisition, transients experience additional artifacts including zero-order phase and frequency drifts, scaling due to coil sensitivity, and decreased SNR values. Each of these are included in the proposed framework. To allow for maximum variation in the simulations, each parameter can be sampled from distributions and is discussed below.

FIGURE 7 Examples of 8 simulated coil transients for a 3T GE PRESS sequence with TE=30ms. ?? shows transients with various SNRs and coil sensitivities along with zero-order phase and frequency offsets. ?? shows the transients after phase alignment. ?? shows the transients after frequency alignment. After ??, the transients can be averaged together and the coil-combined spectrum can then be fitted.

Noise

Averaging multi-coil acquisitions leads 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. These offsets and alignments are shown in Fig. 7.~~

~~Raw spectral transients Phase alignment Frequency alignment Examples of 8 simulated coil transients for a 3T GE PRESS sequence with TE=30ms. ?? shows transients with various SNRs and coil sensitivities along with zero-order phase and frequency offsets. ?? shows the transients after phase alignment. ?? shows the transients after frequency alignment. After ??, the transients can be averaged together and the coil-combined spectrum can then be fitted.~~

Coil Sensitivity

A variety of coil combination techniques can be used to ~~successfully combine~~ combine sets in vivo of multi-coil ~~spectra~~transients into coil-combined spectra. Hall et al.[50] present an overview of some of these methods, of which the following use weighting schemes: water signal, water SNR, signal to peak area, metabolite SNR, time domain combination, and PCA noise whitening with weighted summation. These methods can be simulated in the same way by defining a parametric distribution that, when sampled, produces groups of weighting factors similar to what is normally observed with in vivo acquisitions. The current default implementation samples scaling factors from a Gaussian distribution with $\mu = 1.0$ and $std = 0.5$ that is subsequently clamped to the range [0,2]. ~~While these techniques differ in how they calculate the weights, all of them use weights to scale the transients before averaging. Assigning context~~

Since these weights are defined in the parameter sampling scheme, which is rewritten for every dataset, the settings used for sampling can be easily redefined. Assigning these weights a context that corresponds to an in vivo protocol, 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~~ when planning the simulations. A statistical analysis of the coil sensitivity values from a given in vivo dataset from a specific scanner will provide simulations that more closely approximate experimentally acquired data. This approach is in line with ~~a given acquisition protocol~~the preprocessing consensus paper that recommends always averaging transients[39] instead of summing them. However, if one wishes to sum the transients, then the scaling factors need to be normalized by their sum so as to maintain the overall amplitude. Future work can implement more complex combination techniques.

2.1.11 | Frequency Drift and Phase Drift

Frequency drifts and phase drifts are phenomena observed in transients from multiple acquisitions in which each transient has an independent offset. These offsets and alignments are shown in Fig. 7.

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 the use of the simulated spectra in various software packages, they are also exported in the NIfTI-MRS format [51].

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. Customization of these parameters depends on the intended use and application of the dataset. For example, when creating deep learning-based quantification models, it is beneficial to use independent, uniform distributions that include all values the model will be expected to encounter. On the other hand, 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 create a dataset that accurately mimics in vivo conditions, it is crucial to have precise descriptions of clinical fitting parameters. This work does not provide scenario-specific parameter recommendations, as it is beyond the scope of this work. However, some tools are provided to assist in identifying ranges and distributions to match an existing dataset. In collaboration with the developers of Osprey [36], new functionality was added to their software to export the spectral fitting parameters after quantification. The tools in this framework can load those exported files and prepare the data for further analysis. Currently, this framework uses the Python library Fitter [52] to identify the best-fitting probability distribution for each parameter. *A priori* knowledge, either from prior knowledge or data exploration, can narrow down the search range and accelerate the analysis. The results for each parameter include evaluation metrics for the best performing distributions as well as a numerical characterization of the best fitting distribution.

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.

Metabolite and model parameter recommendations are provided as default values and are included in the ~~appendix and the~~ Supporting Information Section 2 and the online repository. Model parameter values generally come from the default values of spectral fitting programs [36,53,24] while the parameters describing the metabolites come from literature [54,55,38,27,47,43]. For the most current information, please refer to the repository.

2.4.1 | Data Evaluation

Validating synthetic data is challenging because there is no gold standard and the ~~realness~~ "realness" of the data depends on both the simulation methods, i.e. equations and components, and the simulation parameter values. This makes directly evaluating the simulated spectra ~~directly~~ difficult. The scope of this work is focused on the simulation framework itself instead of recommending in vivo parameter values. ~~Therefore, the most appropriate approach for evaluating this protocol is to evaluate the simulation methods described above.~~ Even though there is no quantitative metric describing the degree of in vivo realness, it is still important to evaluate the model outputs qualitatively. Therefore, the most appropriate approach for evaluating this protocol is to evaluate the simulation methods described above and then visually assess the outputs.

The selection of simulations presented in Sec. 3 ~~focus~~ focuses on short echo (TE=30ms) 3T PRESS spectra with a spectral width of 2000Hz and randomly sampled parameters. The concentrations were sampled with respect to total creatine (tCr). The choice of short echo spectra is motivated by their ability to capture a broader range of metabolite peaks compared to long echo spectra. The simulations presented include metabolites sampled from a comprehensive set of common brain metabolites including ~~Asc, Asp, Ch, Cr, GABA, Gln, Glu, GPC, GSH, Lac, mI, NAA, NAAG, PCh, PCr, PE, sI, Tau~~ascorbate (Asc), aspartate (Asp), choline (Ch), creatine (Cr), γ-aminobutyric acid (GABA), glutamine (Gln), glucose (Glu), glycerylphosphocholine (GPC), glutathione (GSH), lactate (Lac), myo-inositol (mI), N-ascetylaspartate (NAA), N-acetylaspartylglutamate (NAAG),

phosphorylcholine (PCh), phosphocreatine (PCr), phosphoethanolamine (PE), scyllo-inositol (sI), taurine (Tau), and a variety of macromolecular and lipid basis functions. For consistency, the SNR was fixed to ~~15dB~~ 15 and the chemical shift reference point was set to 4.65ppm. The metabolite concentrations and ~~lineshapes~~ T_2^* values, adapted from Wyss et al.[43], were fixed to further emphasize the effects of the simulated artifacts.

~~Two individual spectra were simulated as templates to showcase the framework's capabilities~~Except for Fig. 18 and Fig. 19, spectral baselines were also omitted for continuity and visual clarity. The samples shown in Figs. 8 to 15 show standard spectral components and artifacts. Fig. 16 then shows examples of simulating multi-coil transients. The samples in Fig. 17 show how the B_0 simulator affects a spectrum with varying mean inhomogeneity offsets. The samples in Fig. 18 show a single spectrum with various sampled baseline ~~offests~~ offsets and residual water signals. ~~The~~ And finally, the samples in Fig. 19 demonstrate the complete model. Those samples were generated by sampling and randomly applying all spectral artifacts, except zero-order phase because of how much it obfuscates spectral details.

It is important to note that while pulse sequence implications are integral to the overall simulation process, their detailed exploration is beyond the scope of this work. Therefore, ~~in this study, the simulations are specifically tailored to the PRESS sequence~~ the simulations in Sec. 3 are restricted to a single pulse sequence and echo time. The main objective of this work is to manipulate pre-simulated basis sets until they approximate in vivo spectra~~. Therefore,~~ , meaning the primary machinations of MRS-Sim are the same regardless of the basis functions used. Overall, the presented simulations using the PRESS sequence provide a valuable foundation for understanding the performance and capabilities of the MRS-Sim framework.

3 | RESULTS~~& DISCUSSION~~

3.1 | Standard Spectral Components

The following collection of figures show the effects of several standard spectral model components when their parameters are varied. When spectral components are not being studied, their values are fixed. Fig. 8 shows the effect of lorentzian line broadening in the absence of any gaussian broadening. In the top row, the height of the y-axis drops from left-to-right showing the decrease in signal intensity. On the bottom row the differences are not as severe, so a single y-axis is used. Here it is more clearly shown that the spectral peaks are both losing intensity and widening. Fig. 9 uses the same display protocol to show the effects of gaussian broadening without lorentzian broadening whereas Fig. 10 shows several combinations of lorentzian and gaussian broadening that, when combined, give the Voigt lineshape. Fig. 10 breaks down the effects of the lorentzian and gaussian broadening on the Voigt lineshape in the top and bottom rows. On top, the gaussian value is fixed to 20Hz while scaling the T_2^* values using the scaling factors indicated in the captions. On the bottom, the lorentzian broadening is fixed and the gaussian values vary.

Fig. 11 shows two applications of frequency shifts, with all other parameters held constant. The top row shows how global frequency shifts translate spectra upfield and downfield. The bottom row shows the effects of giving each metabolite basis function a minor, randomly sampled frequency shift, which corresponds to fitting protocols. In this case, it is more appropriate to use informed frequency shift values for each metabolite rather than randomly sampling, like the global shifts, because this can increase undesirable spectral overlap.

Next, Fig. 12 and 13 show the effects of zero- and first-order phase offsets, respectively. The amount of phase added to each sample is specified in the caption underneath each plot. Then, Fig. 14 shows identical spectra with varying levels of spectral SNR. Most simulations should include noise, but this becomes especially important when simulating multi-coil transients and multiple acquisitions or editing sequences. The final standard spectral artifact is presented in Fig. 15, eddy currents. As described in Sec. 2.1.9, this model uses two variables: an amplitude and a time constant. The top row of Fig. 15 shows the effects of increasing the amplitude given a fixed time constant. The bottom then shows the effects of varying time constants given a fixed amplitude.

Instead of a spectral artifact, Fig. 16 shows one of the unique features of MRS-Sim: the ability to simulate groups of transients. These examples represent spectra acquired with an 8-channel receiver coil. The coil-combined spectra are shown in blue at the bottom. From left to right the spectral SNR of the coil-combined spectrum increases from 7.5 to 15. The first example was simulated with standard artifacts, but no phase or frequency drift. Global phase offsets are still visible, but the spectral SNR shows significant improvement when the transients are averaged. In the examples in the middle and the right, randomly sampled zero-order phase and frequency drifts were applied. The impact of these effects can be seen in their respective] coil-combined spectra which show improved noise profiles but diminished spectral intensity.

Peer Review Only

FIGURE 8 T_2* values were fixed and then scaled to represent varying magnitudes of lorentzian broadening. The spectral SNR = 15. No gaussian broadening, baselines, phase offsets, or eddy currents were included.

FIGURE 9 Gaussian broadening values were sampled from [0,70.17] in increasing order, as displayed beneath each plot. The lorentzian broadening was left unscaled (scaling coeffient = 1.0 corresponding to Fig. 8). The spectral SNR = 15 and phase offsets and eddy currents were omitted.

(a) T_2 × 0.10 and G = 20Hz (b) T_2 × 1.00 and G = 20Hz (c) T_2 × 2.00 and G = 20Hz

(d) T_2 × 1.00 and G = 10Hz (e) T_2 × 1.00 and G = 20Hz (f) T_2 × 1.00 and G = 30Hz

FIGURE 10 Theses samples explore the effects of lorentzian and gaussian broadening on the Voigt lineshape. The spectral SNR = 15. No phase offsets, eddy currents, or baselines were included. The top row varies the lorentzian broadening with a fixed gaussian value where the bottom row fixes the lorentzian broadening values and varies the gaussian.

(a) F_{shift} =0.2 ppm =160 Hz (b) F_{shift} =0.0 ppm =0 Hz (c) F_{shift} =-0.3 ppm =-240 Hz

(d) F_{shift} [0.10, 0.10] ppm =± 80Hz (e) F_{shift} [0.10, 0.10] ppm =± 80Hz (f) F_{shift} [0.10, 0.10] ppm =± 80Hz

FIGURE 11 Two types of frequency shifts are shown in these examples. The top row shows global frequency shifts. The bottom row shows the effects of using randomly sampled metabolite-level frequency shifts. The spectral SNR = 15 with fixed lorentzian and gaussian broadening. No baselines, phase offsets, or eddy currents were included.

FIGURE 12 Identical spectra are displayed with varying zero-order phase offsets as indicated below each plot. The lorentzian broadening was left unscaled, gaussian broadening = 20 Hz, spectral SNR = 15. First-order phase and eddy currents were omitted.

FIGURE 13 Identical spectra are displayed with varying first-order phase offsets as indicated below each plot. The lorentzian broadening was left unscaled. The gaussian broadening = 20 Hz and the spectral SNR = 15. Zero-order phase, and eddy currents were omitted.

FIGURE 14 Identical spectra are displayed with varying spectral SNRs as indicated below each plot. The lorentzian broadening was left unscaled and the gaussian broadening = 20 Hz. All phase offsets and eddy currents were omitted.

FIGURE 15 Eddy currents are simulated with two variables: an amplitude, A, and a time constant, t_c. In the top row, t_c is fixed to 0.15 seconds and the amplitudes are set to 1Hz, 5.5Hz, and 10Hz. In the bottom row, the amplitude is fixed to 5Hz and the time constants are set to 0.02 sec, 0.16 sec, and 0.30 sec. The values listed correspond to the plots from left to right.

(a) Coil-Combined SNR=7.5, drift aligned **(b)** Coil-Combined SNR=10, drift unaligned **(c)** Coil-Combined SNR=15, drift unaligned

FIGURE 16 These are examples of 8-channel simulated acquisitions. Their respective coil-combined spectra are shown at the bottom in blue. **(A)** shows low-SNR transients that have been phase- and frequency-drift aligned. **(B)** and **(C)** show transients with higher SNRs that have not been aligned, resulting in the coil-combined spectra having lower SNRs than otherwise expected.

3.2 | B_0 Field Inhomogeneity Simulator

Fig. 17 shows a short echo (TE=30ms) spectrum that is corrupted by varying levels of B_0 inhomogeneities. The first example is the reference and has Lorentzian lineshapes with a mean B_0 offset of 1 Hz. The center plot in the top row shows that even a 20 Hz mean offset has a large impact on the spectral intensity. The B_0 inhomogeneity increases from left-to-right and top-to-bottom from 1 Hz to 270 Hz with an interval size of 19 Hz. The total change in each dimension is between 2-10 Hz centered about the mean offset. The red spectra represent the real component from the previous example to highlight the effects of the increasing inhomogeneities. This figure shows that even moderate inhomogeneities noticeably impact the spectral resolution, signal intensity and frequency offset.

3.3 | Baseline and Residual Water Generator

Fig. 18 illustrates a single clean short echo (TE=30ms) PRESS spectrum showcasing various combinations of residual water and baseline contributions. The metabolite concentrations and lineshapes were fixed in addition to the settings described above. These samples omit spectral artifacts to more clearly highlight the variety that can be achieved by randomly sampling parameters for the baseline and residual water generator.

~~Currently, it is unknown what a true baseline actually looks like. Theory and protocols regarding baseline modeling have been developed extensively, but the physical phenomena underlying the baseline signal is still poorly understood. Different modeling protocols can produce different baselines. Bazgir et al.[45] interpolate between selected minima in the spectra, Wilson et al.[46] uses penalized B-splines to overparametize the problem space which is solved using an optimized smoothing constraint, while Osprey[36] uses unregularized B-splines spread every 0.4 ppm to enforce smoothness. Even though all of these techniques have been proved useful, in vivo data does not have known ground truths meaning there is no way to identify which method is the most accurate. Without a physics-based model, other methods of baseline simulation are necessary.~~

Fig. 18 shows a variety of baseline profiles that can be simulated with the presented framework. Instead of claiming to produce in vivo baselines, the proposed generator produces a wide variety of baseline offsets that approximate the baselines extracted by traditional methods. The motivation behind this approach is that if the generator cannot be limited to strictly in vivo-like baselines, then it should incorporate a variety of appropriate baseline profiles such that true baselines are included. Because this generator is not based on a baseline fitting method, future work could use it to evaluate the baseline modeling performance of various fitting protocols given different baseline profiles.

3.4 | Complete Model

In contrast, Fig. 19 presents a simulated PRESS spectrum (TE=30ms) with randomly sampled artifacts and offsets. This figure highlights the variety a single spectrum can assume just by randomly sampling the artifacts. Sophisticated simulations, such as these, are useful for developing and validating data processing techniques such as artifact removal and new spectral fitting protocols. Spectra can be simulated to be as clean or as artifact-laden as desired, all while maintaining known ground truth values.

Upon closer inspection of Fig. 19, the baselines and residual water regions sometimes appear to be uncorrelated with the depicted spectra, but they are in fact paired. These discrepancies are due to the additional artifacts that are applied to the simulations after all of the spectral components are combined. The plotted baseline and residual water regions are the ground truth signals and do not have ~~other~~ those additional artifacts applied to them.

Incorporating a robust collection of artifacts and phenomena commonly encountered with in vivo data into these simulations, makes resulting datasets more closely approximate raw, in vivo data. Even though post-processing techniques are highly accurate, they have limitations and biases, leaving some residue of the corrected artifacts. Including these artifacts in the simulations and removing them with the users' own post-processing and fitting protocols ensures consistency between simulated and in vivo data. These artifacts can also be scaled down in the simulations to emulate residual artifacts. However, this is not ideal because it cannot be guaranteed that scaling down the artifacts will accurately reflect the biases and residues found in the in vivo data.

Deep learning techniques also benefit from robust and consistent synthetic data. Ensuring ~~consistency~~ congruence of the synthetic training data with the in vivo data promotes better model performance. While including fine-grained detail, such as sampling moiety-level ~~T2~~ T_2 values, may not be essential for every use case, in a deep learning setting, it acts as a form of data augmentation. ~~Data~~ Such augmentation techniques maintain the underlying data but manipulate its appearance so that models learn a broader range of features instead of overfitting a single pattern. ~~Such~~ These techniques are standard practice in deep learning and are an inherent feature of this framework.

4 | ~~CONCLUSIONS~~DISCUSSION

~~In recent years, the~~ This work presents a flexible, open-source framework to help researchers generate complex, in vivo-like MRS data. Spectral components from state-of-the-art spectral fitting models were combined to create a comprehensive model. Flexibility and modularity are key components that make this framework a viable option for simulating in vivo-like data representing a multitude of clinical scenarios.

The use of synthetic MRS data has become increasingly popular in recent years due to the need for large datasets with known spectral components. ~~This is useful for both traditional spectroscopy and deep learning applications.~~ Using synthetic data allows researchers to create unlimited spectra with known ground truth values. This ~~has many~~ is useful for both traditional spectroscopy and deep learning applications, including evaluating the accuracy and precision of data analysis methods. However, there is a lack of standardization in synthetic data generation, which poses challenges in terms of reproducibility and generalizability.

~~As discussed in the introduction, researchers~~ Researchers currently use a variety of simulation frameworks to generate synthetic data for their particular needs. These frameworks often involve simplifying assumptions ~~for~~ regarding the spectral components~~to exclude some spectral~~, which excludes some components that are deemed non-essential for ~~their~~ that particular task. These assumptions make synthetic data less generalizable, similar to phantom data which struggles to approximate in vivo data. In general, simulating data has two primary challenges: which spectral components and artifacts to include and what parameter values to use for the included components. These choices can be difficult for ~~expert users~~experts, let alone non-expert users. Moreover, the heterogeneity in the simulation methods used makes reproducibility challenging.

~~This work has presented~~ MRS-Sim addresses these challenges as an open-source ~~framework for a modular~~ solution for a synthetic data generation model that includes a comprehensive list of spectral components and acquisition-induced artifacts. Several novel contributions ~~were made~~ are presented to include experimental artifacts that are usually ignored during spectral modeling. These modules are designed to make ~~these simulations~~ the simulations more comprehensive and in vivo-like. The first contribution is a B_0 ~~magnetic~~ field simulator. It is capable of introducing distortions due to imperfect shimming and high susceptibility effects into the simulations. ~~Additionally, a novel generator was~~ Then, a semi-parametric generator is proposed that can simulate both broad, undulating baseline offsets and highly irregular residual water contributions. The implementation of the presented framework is flexible and can be updated with new information as research progresses to improve the realness and accuracy of the simulations. ~~The appendix and repository include a table that provides up-to-date information about moiety- and metabolite-level characterizations of things like spin-systems, temperature-induced artifacts, T2 values, and metabolite concentration ranges. This information can be continually updated to provide the community with comprehensive and state-of-the-art information about brain metabolites at a glance. Additionally, these suggested values provide a good starting point for simulating datasets. Cumulatively, this framework offers a pathway to generate realistic, in vivo-like synthetic data that captures acquisition-induced artifacts and nuisance signals and can be applied across a range of MRS applications.~~

The modularity of the MRS-Sim ~~framework~~ implementation is one of its key features, allowing for flexible and customizable use of the framework. Firstly, the functions that add artifacts and spectral components are modular and independent, which makes it very easy to understand the code and its purpose. Secondly, it allows for the easy generation of highly customized datasets for various clinical scenarios, as the simulation step itself is also modular. ~~This feature also~~ Consequently, this feature enables easy community contributions to develop new functionalities, thus preventing the need for entirely new frameworks. New scenarios, such as J-difference edited spectra, diffusion spectra, or ~~, eventually,~~ 2D spectra, can be developed and seamlessly integrated.

The appendix and digital repository include a table that provides up-to-date information about moiety- and metabolite-level characterizations of things like spin-systems, temperature-induced artifacts, T_2 values, and metabolite concentration ranges. This information can be continually updated to provide the community with comprehensive and state-of-the-art information about brain metabolites at a glance. Additionally, these suggested values provide a good starting point for simulating datasets. Cumulatively, this framework offers a pathway to generate realistic, in vivo-like synthetic data that captures acquisition-induced artifacts and nuisance signals and can be applied across a range of MRS applications.

5 | CONCLUSIONS

The development of open-source frameworks for data generation is crucial to ensure that synthetic MRS data is widely adopted ~~,~~ and to improve the generalizability of synthetic ~~MRS~~ data and the reproducibility of MRS research. By making this work available to a larger community of researchers, the authors hope to further democratize MRS research by providing easy access to high-quality ~~information and simulations~~ simulations and supporting information. Improvements in the quality of synthetic data will ultimately lead to better generalizability to in vivo data, directly improving the overall applicability of synthetic data in MRS. Moving forward, collaboration and consensus-building among researchers will be essential to establish standards and best practices for simulating MRS data. Future development can focus on adding additional clinical scenarios to the framework's repertoire, increasing its applicability to more fields within MRS. By addressing the challenges associated with synthetic MRS data generation and promoting standardization, MRS can continue to advance in every aspect of the field.

For Peer Review Only

ABBREVIATIONS

Asc ascorbate.

Asp aspartate.

Ch choline.

Cr creatine.

DL deep learning.

GABA gamma-aminobutyric acide.

Gln glutamine.

Glu glutamate.

GPC glycerophosphorylcholine.

GSH glutathione.

Lac lactate.

LASER localized adiabatic selective refocusing.

MARSS Magnetic Resonance Spectrum Simulator.

mI myo-inositol.

ML machine learning.

NAA N-acetylaspartate.

NAAG N-acetylaspartylglutamate.

PCh phosphocholine.

PCr phosphocreatine.

PE phosphatidylethanolamine.

PRESS point resolved spectroscopy.

sI scyllo-inositol.

sLASER semi-localized adiabatic selective refocusing.

STEAM stimulated echo acquisition method.

Tau taurine.

tCr total Creatine.

ACKNOWLEDGMENTS

~~I~~ The authors would like to ~~give sincere thanks to~~ thank Dr. Tobias Lasser for his ~~kind support and mentorship . I would also like to thank~~ mentorship during this work; Dr. Dhritiman Das for ~~our insightful conversations and the help he provided with understanding the simulation methods from his own work. Finally, I would like to thank Leah Harper and Ted Loewenberg, from Harper House , who provided great support during the development of this work while I was visiting Dr. Yan Li's lab~~ discussions about his experience with simulating MRS SVS data; the Harper House for their support while visiting UCSF to further develop the physics model; and the DAAD and Technical University of Munich's Graduate Center of Computation, Information and Technology for helping to finance the stay at UCSF in San Francisco, USA.

~~Financial disclosure~~Funding Information

This work was ~~partially funded~~ supported in part by a doctoral ~~stipend~~ fellowship from the German Academic Exchange Service (DAAD) and the following grants from the National Institutes of Health: R00 AG062230, R01 CA262630, and R21 EB033516.

Conflict of interest

The authors declare no potential conflict of ~~interests~~interest.

References

1. Augustine EM, Spielman DM, Barnes PD, et al. Can magnetic resonance spectroscopy predict neurodevelopmental outcome in very low birth weight preterm infants?. *Journal of Perinatology* 2008; 28(9): 611–618. doi: 10.1038/jp.2008.66

2. Laccetta G, De Nardo MC, Cellitti R, Angeloni U, Terrin G. 1H-magnetic resonance spectroscopy and its role in predicting neurodevelopmental impairment in preterm neonates: A systematic review. *Neuroradiology Journal* 2022; 35(6): 667–677. doi: 10.1177/19714009221102454

3. Tomiyasu M, Shibasaki J, Kawaguchi H, et al. Altered brain metabolite concentration and delayed neurodevelopment in preterm neonates. *Pediatric Research* 2022; 91(1): 197–203. doi: 10.1038/s41390-021-01398-6

4. Gao F, Barker PB. Various MRS application tools for Alzheimer disease and mild cognitive impairment. *American Journal of Neuroradiology* 2014; 35(6 SUPPL.): 1–17. doi: 10.3174/ajnr.A3944

Peer Review Only

5. Martin WR. MR spectroscopy in neurodegenerative disease. *Molecular Imaging and Biology* 2007; 9(4): 196–203. doi: 10.1007/s11307-007-0087-2

6. McKiernan E, Su L, O'Brien J. MRS in neurodegenerative dementias, prodromal syndromes and at-risk states: A systematic review of the literature. *NMR in Biomedicine* 2023(January): 1–13. doi: 10.1002/nbm.4896

7. Öz G., ed.*Magnetic Resonance Spectroscopy of Degenerative Brain Diseases*. Contemporary Clinical NeuroscienceCham: Springer International Publishing . 2016

8. Cecil KM. MR Spectroscopy of Metabolic Disorders. *Neuroimaging Clinics of North America* 2006; 16(1): 87–116. doi: 10.1016/J.NIC.2005.10.004

9. Gropman AL, Anderson A. Novel imaging technologies for genetic diagnoses in the inborn errors of metabolism. *Journal of Translational Genetics and Genomics* 2020: 429–445. doi: 10.20517/jtgg.2020.09

10. Lai LM, Gropman AL, Whitehead MT. MR Neuroimaging in Pediatric Inborn Errors of Metabolism. *Diagnostics* 2022; 12(4). doi: 10.3390/diagnostics12040861

11. Calvar JA, Meli FJ, Romero C, et al. Characterization of brain tumors by MRS, DWI and Ki-67 labeling index. *Journal of Neuro-Oncology* 2005; 72(3): 273–280. doi: 10.1007/s11060-004-3342-2

12. Lukas L, Devos A, Suykens JA, et al. Brain tumor classification based on long echo proton MRS signals. *Artificial Intelligence in Medicine* 2004; 31(1): 73–89. doi: 10.1016/J.ARTMED.2004.01.001

13. Padelli F, Mazzi F, Erbetta A, et al. In vivo brain MR spectroscopy in gliomas: clinical and pre-clinical chances. *Clinical and Translational Imaging* 2022; 10(5): 495–515. doi: 10.1007/s40336-022-00502-y

14. Nelson SJ. Multivoxel Magnetic Resonance Spectroscopy of Brain Tumors1. *Molecular Cancer Therapeutics* 2003; 2(5): 497–507.

15. Forester BP, Berlow YA, Harper DG, et al. Age-related changes in brain energetics and phospholipid metabolism. *NMR in Biomedicine* 2010; 23(3): 242–250. doi: 10.1002/nbm.1444

16. Inglese M, Ge Y. Quantitative MRI. *Topics in Magnetic Resonance Imaging* 2004; 15(6): 355–363. doi: 10.1097/01.rmr.0000168069.12985.15

17. Reyngoudt H, Claeys T, Vlerick L, et al. Age-related differences in metabolites in the posterior cingulate cortex and hippocampus of normal ageing brain: A 1H-MRS study. *European Journal of Radiology* 2012; 81(3): e223–e231. doi: 10.1016/j.ejrad.2011.01.106

18. Dezortova M, Hajek M. 1H MR spectroscopy in pediatrics. *European Journal of Radiology* 2008; 67(2): 240–249. doi: 10.1016/j.ejrad.2008.02.035

19. Li Y, Srinivasan R, Ratiney H, Lu Y, Chang SM, Nelson SJ. Comparison of T1 and T2 metabolite relaxation times in glioma and normal brain at 3T. *Journal of Magnetic Resonance Imaging* 2008; 28(2): 342–350. doi: 10.1002/jmri.21453

20. Hatami N, Sdika M, Ratiney H. Magnetic resonance spectroscopy quantification using deep learning. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* 2018; 11070 LNCS: 467–475. doi: 10.1007/978-3-030-00928-1$_{53}$

21. Das D, Davies ME, Chataway J, Chandran S, Menze BH, Marchall I. Metabolite Quantification of 1H-MRSI spectra in Multiple Sclerosis: A Machine Learning Approach. *ISMRM* 2018.

22. Das D, Thrippleton MJ, Semple SIK, et al. MRSI Brain Temperature Mapping Using Machine Learning. *ISMRM* 2018: 2–6.

23. Iqbal Z, Nguyen D, Thomas MA, Jiang S. Deep learning can accelerate and quantify simulated localized correlated spectroscopy. *Scientific Reports* 2021: 1–17. doi: 10.1038/s41598-021-88158-y

24. 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. *Magnetic Resonance in Medicine* 2017; 77(1): 23–33. doi: 10.1002/mrm.26091

25. Soher BJ, Young K, Bernstein A, Aygula Z, Maudsley AA. GAVA : Spectral simulation for in vivo MRS applications. *Journal of Magnetic Resonance* 2007; 185: 291–299. doi: 10.1016/j.jmr.2007.01.005

Peer Review Only

26. Soher BJ, Semanchuk P, Todd D, Ji X. VeSPA: Integrated applications for RF pulse design, spectral simulation and MRS data analysis. *Magnetic Resonance in Medicine* 2023; 90: 823–838. doi: 10.1002/mrm.29686

27. Landheer K, Swanberg KM, Juchem C. Magnetic resonance Spectrum simulator (MARSS), a novel software package for fast and computationally efficient basis set simulation. *NMR in Biomedicine* 2021; 34(5): 1–13. doi: 10.1002/nbm.4129

28. Das D, Coello E, Schulte RF, Menze BH. Quantification of metabolites in magnetic resonance spectroscopic imaging using machine learning. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* 2017; 10435 LNCS: 462–470. doi: 10.1007/978-3-319-66179-7$_53$

29. Gurbani SS, Sheriff S, Maudsley AA, Shim H, Cooper LA. Incorporation of a spectral model in a convolutional neural network for accelerated spectral fitting. *Magnetic Resonance in Medicine* 2019; 81(5): 3346–3357. doi: 10.1002/mrm.27641

30. Lee HH, Kim H. Intact metabolite spectrum mining by deep learning in proton magnetic resonance spectroscopy of the brain. *Magnetic Resonance in Medicine* 2019; 82(1): 33–48. doi: 10.1002/mrm.27727

31. Craven AR, Bhattacharyya PK, Clarke WT, et al. Comparison of seven modelling algorithms for γ-aminobutyric acid–edited proton magnetic resonance spectroscopy. *NMR in Biomedicine* 2022; 35(7): 1–17. doi: 10.1002/nbm.4702

32. Marjańska M, Terpstra M. Influence of fitting approaches in LCModel on MRS quantification focusing on age-specific macromolecules and the spline baseline. *NMR in Biomedicine* 2021; 34(5): 1–9. doi: 10.1002/nbm.4197

33. Marjańska M, Deelchand DK, Kreis R, et al. Results and interpretation of a fitting challenge for MR spectroscopy set up by the MRS study group of ISMRM. *Magnetic Resonance in Medicine* 2022; 87(1): 11–32. doi: 10.1002/mrm.28942

34. Zöllner HJ, Považan M, Hui SC, Tapper S, Edden RA, Oeltzschner G. Comparison of different linear-combination modeling algorithms for short-TE proton spectra. *NMR in Biomedicine* 2021; 34(4): 1–17. doi: 10.1002/nbm.4482

35. Provencher SW. Automatic quantitation of localized in vivo 1H spectra with LCModel. *NMR in Biomedicine* 2001; 14(4): 260–264. doi: 10.1002/nbm.698

36. Oeltzschner G, Zöllner HJ, Hui SC, et al. Osprey: Open-source processing, reconstruction estimation of magnetic resonance spectroscopy data. *Journal of Neuroscience Methods* 2020; 343: 1–27. doi: 10.1016/j.jneumeth.2020.108827

37. Clarke WT, Stagg CJ. FSL-MRS : An end-to-end spectroscopy analysis package. *Magnetic Resonance in Medicine* 2021; 85: 2950–2964. doi: 10.1002/mrm.28630

38. Gudmundson AT, Koo A, Virovka A, et al. Meta-analysis and open-source database for in vivo brain Magnetic Resonance spectroscopy in health and disease. *Analytical Biochemistry* 2023; 676(July): 115227. doi: 10.1016/j.ab.2023.115227

39. Near J, Harris AD, Juchem C, et al. Preprocessing, analysis and quantification in single-voxel magnetic resonance spectroscopy: experts' consensus recommendations. *NMR in Biomedicine* 2021; 34(5): 1–23. doi: 10.1002/nbm.4257

40. Larsen RJ, Gagoski B, Litt J, Morton SU, Grant PE, Sutton BP. Quantification of magnetic resonance spectroscopy data using a combined reference : Application in typically developing infants. *NMR in Biomedicine* 2021: 1–16. doi: 10.1002/nbm.4520

41. Wilson M, Reynolds G, Kauppinen RA, Arvanitis TN, Peet AC. A constrained least-squares approach to the automated quantitation of in vivo 1H magnetic resonance spectroscopy data.. *Magnetic resonance in medicine : official journal of the Society of Magnetic Resonance in Medicine / Society of Magnetic Resonance in Medicine* 2011; 65(1): 1–12. doi: 10.1002/mrm.22579

42. Stefan D, Cesare FD, Andrasescu A, et al. Quantitation of magnetic resonance spectroscopy signals: The jMRUI software package. *Measurement Science and Technology* 2009; 20(10). doi: 10.1088/0957-0233/20/10/104035

43. Wyss PO, Bianchini C, Scheidegger M, et al. In Vivo Estimation of Transverse Relaxation Time Constant (T2) of 17 Human Brain Metabolites at 3T. *Magnetic Resonance in Medicine* 2018; 80(2): 452–461. doi: 10.1002/mrm.27067

44. Li N, An L, Shen J. Spectral fitting using basis set modified by measured B0 field distribution. *NMR in Biomedicine* 2015; 28(12): 1707–1715. doi: 10.1002/nbm.3430

For Peer Review Only

45. Bazgir O, Mitra S, Nutter B, Walden E. Fully automatic baseline correction in magnetic resonance spectroscopy. *Proceedings of the IEEE Southwest Symposium on Image Analysis and Interpretation* 2018; 2018-April: 97–100. doi: 10.1109/SSIAI.2018.8470319

46. Wilson M. Adaptive baseline fitting for 1H MR spectroscopy analysis. *Magnetic Resonance in Medicine* 2021; 85(1): 13–29. doi: 10.1002/mrm.28385

47. Wermter FC, Mitschke N, Bock C, Dreher W. Temperature dependence of 1H NMR chemical shifts and its influence on estimated metabolite concentrations. *Magnetic Resonance Materials in Physics, Biology and Medicine* 2017; 30(6): 579–590. doi: 10.1007/s10334-017-0642-z

48. Klose U. In vivo proton spectroscopy in presence of eddy currents. *Magnetic Resonance in Medicine* 1990; 14(1): 26–30. doi: 10.1002/mrm.1910140104

49. Kreis R, Boer V, Choi IY, et al. *Terminology and concepts for the characterization of in vivo MR spectroscopy methods and MR spectra: Background and experts' consensus recommendations*. No. May 2020John Wiley Sons Ltd . 2020

50. Hall EL, Stephenson MC, Price D, Morris PG. Methodology for improved detection of low concentration metabolites in MRS: Optimised combination of signals from multi-element coil arrays. *NeuroImage* 2014; 86: 35–42. doi: 10.1016/j.neuroimage.2013.04.077

51. Clarke WT, Bell TK, Emir UE, et al. NIfTI-MRS: A standard data format for magnetic resonance spectroscopy. *Magnetic Resonance in Medicine* 2022; 88(6): 2358–2370. doi: 10.1002/mrm.29418

52. Cokelaer T. FITTER documentation — fitter 1.5.2 documentation. 2019.

53. Provencher SW. Estimation of metabolite concentrations from localized in vivo proton NMR spectra. *Magnetic Resonance in Medicine* 1993; 30(6): 672–679. doi: 10.1002/mrm.1910300604

54. Graaf dRA. *In Vivo NMR Spectroscopy: Principles and Techniques*. Wiley. 2nd ed. 2007

55. Graaf dRA. *In Vivo NMR Spectroscopy: Principles and Techniques*. Wiley. 3rd ed. 2019

How to cite this article: LaMaster J., G. Oeltzschner, and Y. Li (2022), MRS-Sim: Open-Source Framework for Simulating Realistic In Vivo-like MR Spectroscopy Data, *NMR in Biomedicine*.

Peer Review Only

FIGURE 17 Sample spectrum simulated for a PRESS sequence with TE=30ms that highlights the effect of B_0 inhomogeneities on the spectra. The mean inhomogeneities range from [1, 270] Hz and monotonically increase by 19Hz. As the inhomogeneities increase, the intensity of the spectra decreases, shown by the y-axis range, and causes a global frequency shift.

FIGURE 18 Sample spectra simulated for a PRESS sequence with TE=30ms that highlight the effect of the baseline and residual water contributions.

FIGURE 19 Sample spectra, similar to Fig. 18 [PRESS, TE=30ms], with randomly sampled, uncorrected artifacts to approximate raw data.

Received: Added at production | Revised: Added at production | Accepted: Added at production

DOI: xxx/xxxx

ARTICLE TYPE

MRS-Sim: Open-Source Framework for Simulating In Vivo-like Magnetic Resonance Spectra

John LaMaster*[1,2] | Georg Oeltzschner[3] | Yan Li[4]

[1]Munich Institute of Biomedical Engineering, Technical University of Munich, Bavaria, Germany
[2]School of Computation, Information, and Technology, Technical University of Munich, Bavaria, Germany
[3]Russel H. Morgan Department of Radiology and Radiological Sciences, The Johns Hopkins' University School of Medicine, Maryland, USA
[4]Department of Radiology and Biomedical Imaging, University of California, San Francisco, California, USA

Correspondence
John LaMaster.
Email: john.t.lamaster@gmail.com
Twitter: @JohnTLaMaster

Present Address
Munich Institute of Biomedical Engineering, School of Computation, Information, and Technology, Technical University of Munich in the City of Garching bei München, Boltzmannstr 11. 85748 Garching bei München, Bavaria, Germany.

Summary

MRS-Sim is a powerful, open-source framework for simulating in vivo-like magnetic resonance spectroscopy data. Its modularity allows it to model the complexities of MRS data for a variety of clinical scenarios. The underlying physical equations include spectral components commonly found in linear-combination fitting routines plus some novel components. The first is a 3D B_0 field map simulator to model B_0 field inhomogeneities, ranging from slight variations to severe distortions. The second is a novel semi-parametric generator that mimics signals from poorly characterized residual water regions and baseline offset contributions. This framework can simulate everything from raw multi-coil transients to preprocessed coil-combined data.

A repository of information has been compiled to help non-expert users with general simulations of MRS data. Simulating realistic clinical-like datasets requires appropriate ranges and distributions for the model parameters, which are best determined by analyzing the parameter estimates from fitting real clinical data. Therefore, accompanying tools within MRS-Sim allow for analyzing the range and distributions of those parameters from in vivo datasets fitted with Osprey. This allows simulations to be tailored to specific clinical datasets.

The modularity of this framework allows for easy customization of simulations to suit a variety of clinical scenarios and encourages continued community development. Using a single framework for various scenarios overcomes the inconsistent protocols in current literature. By simulating in vivo-like data, this framework can aid in many tasks in MRS, including verifying spectral fitting protocols and reproducibility analyses. The accessibility of readily available, synthetic data will also assist deep learning research for MRS, particularly in cases where clinical data is insufficient or unavailable for training and, in general, will help address reproducibility issues and make MRS research more accessible to a wider audience.

Word count (abstract): 295 -> 297 -> 282 words

Word count (*paper): 4,206 -> 5,125 -> 6,314 words

*Limited to intro, methods, results, and conclusions. Including auxiliary text (author info, etc.): 4,807 -> 5,757 -> 6,805 words

KEYWORDS:
MRS, spectroscopy, in vivo, data simulation, synthetic data, open-source

1 | INTRODUCTION

Magnetic resonance spectroscopy (MRS) is a non-invasive imaging modality that provides in vivo information on the metabolic profile of tissues. It allows for evaluating various pathologies throughout the entire body by analyzing the spectra to quantify metabolite concentrations. 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].

In order for MRS to be a widely used clinical imaging modality, it is important to ensure the precision and accuracy of MRS data processing and analysis. However, it can be challenging due to variations across scenarios, such as pulse sequence, acquisition parameters, data quality regimes, and biological differences (adult versus pediatric subjects [18] and disease related metabolic patterns [19]). Such comprehensive collections of data are usually not widely available for method development and evaluation due to data privacy restrictions. Furthermore, in vivo data do not have access to ground truth values, making it difficult to determine the reliability, sensitivity, and robustness of different MRS analysis protocols. Phantom data can be useful in developing and validating acquisition and analysis methods, but it usually does not adequately reflect the complexities of in vivo spectra.

A logical solution to these challenges is to generate synthetic data for a variety of experimental scenarios. However, generating synthetic MRS data can be challenging as well, as it requires adequately incorporating all physical phenomena underlying in vivo data. One of the significant aspects is to choose adequate model parameters that reflect different physiological and pathological conditions. Arguably more important is the need to capture acquisition-induced artifacts and nuisance signals commonly found in in vivo data that are difficult to reproduce in a phantom, such as effects of susceptibility and field inhomogeneity, macromolecule, lipid, and residual water signals, and various other artifacts. Many research groups now routinely use synthetic data, but the data generation models and parameter distributions are rarely made publicly available. This lack of transparency hampers systematic comparison and the formation of consensus best practices for synthetic data generation.

It is evident from the literature that there is no standard set of best practices for simulating synthetic data. Simulations have been carried out using physical models with a range of complexity and spectral components [20,21,22,23]. Most of these methods begin with simulating basis functions, assuming appropriate pulse sequence parameters for their specific scenario. These functions represent metabolites that can be collected and are then modulated by scaling factors representing their underlying concentrations. A simple Lorentzian lineshape is generally applied [20,21,22] with optional phase offsets [22,23] and frequency shifts [20]. Finally, some type of broad baseline is typically added. These models are simple and do not fully capture the complexity of in vivo data. Additionally, they often contain non-public components such as baselines, macromolecules, and lipid signals that are either simulated in-house or extracted from private datasets, thus making detailed analysis of their models difficult.

There are several software packages that claim to simulate in vivo spectra, such as FID-A [24], GAVA [25], and VESPA [26]. However, the fundamental difference between these methods and the current work is the definition of "in vivo spectra". Previously, this term referred to simulating idealized spectra to evaluate parameter settings for developing new in vivo pulse sequences. In this work, "in vivo spectra" refers to the data that comes off a clinical scanner prior to any post-processing or analysis. This new definition is meant to reframe the task of simulating synthetic data, to challenge what are assumed to be acceptable simplifications, and to advance the use of synthetic data for MRS applications. While pulse sequences are necessary for generating basis sets, they are left to the many software options already available. The most relevant previous work is the recent update for the MARSS [27] software which has added a new functionality called synMARSS that simulates synthetic spectra and incorporates many spectral artifacts. They provide an extensive user manual with equations describing their implementation, however, their source code is not available, meaning it cannot be evaluated or modified like MRS-Sim.

Over the past few years, there has been a significant rise in the use of synthetic MRS data. This increase is primarily due to the emergence of machine learning [28] (ML) and deep learning [29,20,30] (DL) analysis algorithms that require vast amounts of data for training and validation that few centers worldwide possess. Therefore, generating realistic, in vivo-like synthetic data is necessary to facilitate more research into these applications. Synthetic data has known ground truth values that are essential to evaluate the accuracy and precision of the analysis methods being developed. However, 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. [31,32,33,34] To address this challenge, it is essential to establish standards and best practices for simulating synthetic MRS data. This will require researchers to collaborate and build consensus, as well as develop and adopt open-source frameworks for data generation.

This work presents MRS-Sim: a framework for a modular synthetic data generation model to address some of these challenges. The basic simulation model uses 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 B_0 map generator to simulate in vivo-like field heterogeneity conditions, and uses semi-parameterized models to describe residual water and smooth background signal contributions. MRS-Sim 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 compare new data processing methods, not just modeling algorithms. And the implemented code base is highly flexible

For Peer Review Only

and can simulate custom-tailored datasets for a large variety of clinical scenarios. This framework includes a detailed compilation of simulation parameters as well as tools to analyze existing in vivo datasets and extract parameter distributions, which is useful for augmenting in vivo data with similar synthetic data. The open-source code base allows for seamless incorporation of future additions to expand the framework'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

The proposed physics model was developed after a thorough review of current state-of-the-art fitting algorithms (LCModel[35], Osprey[36], FSL-MRS[37], etc.). Commonly modeled spectral features and components were identified for inclusion, allowing for a large variety of scenarios and artifacts. The order of application was further specified for different operations and along with appropriate ranges and distributions for the underlying model parameters. The model was parameterized as follows:

$$f(t) = \sum_{n}^{N} \underbrace{M_n * basisfcn_n}_{amplitude\ scaling} \underbrace{e^{-d_n t - g_n t^2}}_{lineshape} \underbrace{e^{-i\Delta f_n t}}_{fshift} \underbrace{e^{i(\phi_0+\phi_1)}}_{phase} \underbrace{e^{-iA_0 * e^{-t/tc_0} t * 2\pi}}_{eddyCurrents} * \underbrace{\sum_{r=1}^{R} e^{-i\Delta\omega_r t}}_{B_0\ inhomog.} + \mathcal{H}\left(\underbrace{baseline + resH_2O}_{nuisance\ signals}\right) + \underbrace{\mathcal{N}(0, snr_0, l)}_{noise} \quad (1a)$$

$$f(t) = \mathcal{C}oil_0^C\left(f(t), \underbrace{\overbrace{snr_c}^{noise}, \overbrace{sens_c}^{sensitivity}, \overbrace{\Delta f_c, \Delta\phi_c}^{drifts}}_{coil\ artifacts}\right) \quad (1b)$$

$$F(\omega) = \mathcal{F}\left(zeroFill\left(f(t) * \underbrace{e^{-t*T_L}}_{apodize}\right), len\right) \quad (1c)$$

where N is the whole set of metabolites being modeled, M_n is the scaling factor for metabolite n, the Lorentzian variable d_n and the Gaussian variable g_n combine to define a Voigt lineshape, t is time, and Δf_n is the metabolite-specific frequency shift. Global zero- and first-order phase offsets are added using ϕ_0 and ϕ_1 while eddy current effects are described using two variables as a function of time t: the amplitude, A_0, and the time constant, tc_0. In lieu of the Gaussian term, imperfect shimming and severe lineshape distortions associated with large susceptibility effects can be applied using $\Delta\omega_r$, which is the modeled B_0 at location r inside the voxel of interest. snr_0 is the desired SNR of the spectrum, while $baseline$ and $resH_2O$ 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 $\mathcal{C}oil$ generates C coil transients and applies a distribution of SNR values and coil weights using snr_c and $sens_c$. Frequency drifts and phase drifts are then added using Δf_c and $\Delta\phi_c$, respectively. When necessary, Eqn. 1c can apply apodization using T_L in Hz and the FIDs can be zero-filled to length len. Then the Fourier transform $\mathcal{F}$ can convert the FIDs to the frequency domain. The following sections are presented in order according to their implementation in the physics model and discuss the terms above in greater detail. A step-by-step visualization of this model is illustrated in Fig. 1.

2.1.1 | Basis Functions

MRI, and its derivatives, are spatially resolved imaging modalities. MR pixels and voxels represent 3D volumes with a spatial distribution, as shown in Fig. 2a, and need to be accounted for in simulations. The importance of spatial localization led to selecting Landheer *et al.*'s Magnetic Resonance Spectrum Simulator (MARSS)[27] software package for simulating the default basis functions provided with this simulator. MARSS produces high-fidelity outputs by simulating 128 points in each direction, accurately capturing the spatial nature of the RF pulses and slice-selective gradients. Using vendor-specific pulse sequences, MARSS can simulate individual or summed spins for a large number of common brain metabolites including macromolecules and lipids.

FIGURE 1 This is a step-by-step progression through the physics model. The real and imaginary components are depicted in black and gray, respectively. The red line includes only the metabolites and the offsets from the preceding steps. 1k shows the final spectrum with all artifacts applied. 1l is the pre-processed spectrum with the phase and frequency shifts removed.

2.1.2 | Amplitude Modulation

For this framework, comprehensive ranges including physiological and pathological values were adapted from the meta-analysis by Gudmundson *et al.*[38] and are presented in the appendix. Near[39] and Larsen[40] discuss how non-trivial the choice of reference metabolite is with respect to quantification and interpretability. Therefore, the default ranges are in units of millimolar (mM) which allows any reference to be used. More specific information can be found in the MARSS user manual.

(a) Intravoxel localization (b) Intravoxel spatial distribution of B_0 with minor distortions (c) Spatial distribution of B_0 with major distortions

FIGURE 2 Voxels are often treated as singularities, meaning the entire volume is considered to be a single point like the red box in the center. Accounting for these intravoxel distributions leads to more accurate and more realistic simulations. 2a highlights the spatial nature of voxels while 2b and 2c show different levels of intravoxel B_0 inhomogeneity.

(a) Spectral peaks without B_0 inhomogeneities ($\mu = 0$Hz) (b) Spectral peaks with moderate B_0 inhomogeneities ($\mu = 75$Hz) (c) Spectral peaks with severe B_0 inhomogeneities ($\mu = 175$Hz)

FIGURE 3 These samples show the effects that can be modeled using the 3D B_0 field simulator. In 3a, the spectral peaks have a purely Lorentzian lineshape. In 3b, the lineshape is now Voigtian because the Gaussian term has been added back by the 3D B_0 field simulator. In 3c, severe heterogeneities are modeled which produce extremely broad line widths. All three plots use the same x- and y-axes. The observed offsets are caused by the line broadening.

2.1.3 | Lineshape Profiles

The Voigt lineshape profile is the most commonly used in spectral fitting as it most closely matches in vivo data. It is a combination of a Lorentzian and a Gaussian and is used in various fitting packages, such as LCModel[35], TARQUIN[41], jMRUI[42], and Osprey[36]. For completeness, it is also possible to specify either a purely Lorentzian or a purely Gaussian lineshape. Individual Lorentzian and Gaussian values can be assigned to each moiety or metabolite separately, depending on the desired level of granularity. At the most basic level, each metabolite is assigned a Lorentzian value while a single Gaussian value is applied to the set of metabolites and a second value can be applied to the macromolecules and lipids. Standard practice from the aforementioned software packages assign a single Lorentzian value to each metabolite, instead of each moiety. It is also possible to specify moiety-level T_2 relaxation values for selected metabolites from Wyss *et al.*[43]. Information is supplemented by the meta-analysis from Gudmundson *et al.*[38]. As more metabolites are characterized, new information can be added to the model. A detailed table is provided in the appendix with the current information; however, an up-to-date digital version will be maintained in the repository.

2.1.4 | B_0 Inhomogeneities

Small B_0 inhomogeneities are, in general, sufficiently modeled by the Gaussian term of the Voigt lineshape. However, to simulate more severe distortions from large inhomogeneities, a B_0 field volume needs to be modeled and applied to the basis functions. Fig. 2 illustrates the difference between small and severe B_0 inhomogeneities. Fig. 2b shows the normal, subtle B_0 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 that cannot be adequately characterized using idealized lineshape profiles. Fig. 3 shows the result of high susceptibility effects on lineshapes.

In general, this approach mirrors Li *et al.*[44] in which the basis functions are modulated by the B_0 field map, but in the implementation the B_0 field map is simulated rather than acquired. As with MARSS, Li *et al.*[44] suggest 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. Any cuboidal shape, rectangular or otherwise, can be modeled. The B_0 field is defined by four variables: $\pm dx$, $\pm dy$, $\pm dz$, and μ. dx, dy, and dz describe half of the change in B_0 in their respective axis from the voxel's center and μ is the mean of the entire voxel. A linear B_0 gradient is assumed, but different gradient profiles can be implemented by adding new functions that manipulate how the $x-$, $y-$, and $z-$components are calculated when generating the B_0 field map. Section 1 in the Supporting Information provides a more detailed explanation of how the B_0 is simulated.

2.1.5 | Baseline and Residual Water

Currently, there is no physics-based model for simulating these offsets and it is unknown what a true baseline actually looks like. Theory and protocols regarding baseline modeling have been developed extensively, but the physical phenomena underlying the spectral baseline signal are still poorly understood. Therefore, different modeling protocols can produce different baselines. Bazgir et al.[45] interpolate between selected minima in the spectra, Wilson et al.[46] use penalized B-splines to overparametize the problem space which is solved using an optimized smoothing constraint, while Osprey[36] uses unregularized B-splines spread every 0.4 ppm to enforce smoothness. Even though all of these techniques have been proved

FIGURE 4 Simulated samples of spectral baselines and residual water regions using the pseudo-random bounded walk generator. The black and red lines are the real and imaginary components of the smoothed offsets that are then returned and applied to the simulated spectra.

useful, in vivo data does not have known ground truths meaning there is no way to identify which method is the most accurate. Without a physics-based model, other methods of baseline simulation are necessary.

Similarly, the residual suppressed water region is also poorly characterized. Because these phenomena are poorly understood, a naive random model can be used in conjunction with observed constraints to approximate what is expected 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

Algorithm 1 Smoothed Bounded Pseudo-Random Walk

Require: $start$: ppm start value, end: ppm end value, std: standard deviation of the random noise, $lower_bound$: lower bound of the walk, $upper_bound$: upper bound of the walk, $length$: length of the walk

Function SmoothedBoundedPseudoRandomWalk($start, end, std, lower_bound, upper_bound, length, window_size$)

$bounds = upper_bound - lower_bound$

Generate raw random walk

$rand =$ generate_random_array$(size = length, std = std)$

$rand =$ cumsum$(rand, dim = -1)$

Calculate the trend line

$rand_trend_lines =$ generate_trend_array$(start, end, length)$

Calculate the difference between the random steps and the trend line

$rand_deltas = rand - rand_trend_lines$

Normalize the delta array

$delta_range = bounds \div (rand_deltas.max() - rand_deltas.min())$

$rand_deltas = rand_deltas * delta_range$

Adjust delta array if it exceeds *bounds*

$rand_deltas = conditional_adjustment(rand_deltas)$

Calculate the final random walk

$random_walk = trend_lines + rand_deltas$

Smooth using a uniform smoothing kernel

$random_walk =$ average_smoothing$(random_walk, kernel_size = window_size * length)$

return $random_walk$

(a) Spectrum with no phase offsets (b) Spectrum with zero-order phase offset ($\phi_0 = 45^\circ$) (c) Spectrum with first-order phase offset ($\phi_1 = 20^\circ$)

FIGURE 5 These samples show how identical spectra are affected by zero- and first-order phase offsets. In 5a, the real component (black) is in absorption mode exhibiting narrow line widths and is fully positive. In 5b, a zero-order phase offset is applied. As the spectrum shifts from absorption to dispersion mode, the peaks uniformly lose their symmetry and negative values from the imaginary component are transferred to the real component. In 5c, a first-order phase shift is applied. This is evident because the asymmetry increases across the spectrum and emanates from the water peak.

outputs can be achieved by randomly sampling the parameters from distributions instead of using fixed 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 Supporting Information Section 3 Baseline and Residual Water Simulations.

2.1.6 | 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.

2.1.7 | Phase Offsets

Zero-Order Phase

FIDs and spectra are complex data types consisting of real and imaginary components. 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 anti-symmetrical peaks which are both positive and negative. As shown in Fig. 5, non-zero degree offsets result in a mixture of absorption and dispersion spectra. In its current form, a phase offset is selected randomly and independently during parameter sampling and applied in the time domain using the following complex exponential:

$$f(t) = f(t) * e^{-i\phi_0 * (\pi/180)} \tag{2}$$

First-Order Phase

First-order phase, i.e. linear phase, is a frequency-dependent linear offset that emanates from a reference point, i.e. the center frequency referred to as ppm_ref, which is typically the water peak at 4.65ppm. A linear phase offset creates asymmetrical line shapes that grow larger as one moves away from the reference point. This is illustrated in Fig. 5 by comparing 5a and 5c. Currently, the ppm_ref term is fixed to 4.65ppm while the ϕ_1 term is randomly sampled and then applied in the frequency domain using the following complex exponential:

$$F(\omega) = F(\omega) * e^{-i\phi_1 * (\pi/180) * (ppm - ppm_ref} \tag{3}$$

(a) Eddy current amplitude = 1.0 (b) Eddy current amplitude = 3.0 (c) Eddy current amplitude = 5.0

FIGURE 6 These 3 samples show the effect of eddy currents on MRS spectra to various degrees. The strength of the eddy currents increases from 6a to 6c. If the time constant, tc, is set too long, the eddy current artifact will appear as a global frequency shift. In these examples, it can be seen that only some frequencies are affected.

2.1.8 | Frequency Shifts

During data acquisition, the FID experiences a global frequency shift. However, some functional groups can experience individual frequency shifts which are attributed to effects such as temperature and pH. Similar to Sec. 2.1.3, the bare minimum requires separate global frequency shifts for the metabolites and nuisance signals. However, this model also allows each basis function to have an independent frequency shift in addition to the global shift which is in line with common spectral fitting protocols. For more in vivo-like, realistic spectra, values can be used from the work by Wermter *et al.*[47] which characterized the temperature-induced frequency shift of several 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. Currently available data will be included in the table in the Supplemental Information Section 2.

2.1.9 | Eddy Currents

Eddy current correction techniques, such as the Klose[48], 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.

2.1.10 | Multi-Coil Transients

In this work, the term *multi-coil transients*, or simply *transients*, refers to the signal from multiple individual RF coils measured from a phased-array receiver coil for a single average prior to coil combination. The terminology consensus paper by Kreis et al.[49] provides many clarifications, but does not specify how to differentiate between data before and after coil-combination nor specifically when a transient becomes a spectrum. For clarity and consistency, *spectrum* and *FID* will refer to data that has been coil-combined. When simulating transients, additional considerations need to be included. During acquisition, transients experience additional artifacts including zero-order phase and frequency drifts, scaling due to coil sensitivity, and decreased SNR values. Each of these are included in the proposed framework. To allow for maximum variation in the simulations, each parameter can be sampled from distributions and is discussed below.

Noise

Averaging multi-coil acquisitions leads 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.

Coil Sensitivity

A variety of coil combination techniques can be used to combine sets in vivo of multi-coil transients into coil-combined spectra. Hall et al.[50] present an overview of some of these methods, of which the following use weighting schemes: water signal, water SNR, signal to peak area, metabolite SNR, time domain combination, and PCA noise whitening with weighted summation. These methods can be simulated in the same way by defining

Peer Review Only

FIGURE 7 Examples of 8 simulated coil transients for a 3T GE PRESS sequence with TE=30ms. 7a shows transients with various SNRs and coil sensitivities along with zero-order phase and frequency offsets. 7b shows the transients after phase alignment. 7c shows the transients after frequency alignment. After 7c, the transients can be averaged together and the coil-combined spectrum can then be fitted.

a parametric distribution that, when sampled, produces groups of weighting factors similar to what is normally observed with in vivo acquisitions. The current default implementation samples scaling factors from a Gaussian distribution with $\mu = 1.0$ and $std = 0.5$ that is subsequently clamped to the range [0,2].

Since these weights are defined in the parameter sampling scheme, which is rewritten for every dataset, the settings used for sampling can be easily redefined. Assigning these weights a context that corresponds to an in vivo protocol, such as water peak height or coil sensitivity maps, can help define the necessary parameter ranges and distributions when planning the simulations. A statistical analysis of the coil sensitivity values from a given in vivo dataset from a specific scanner will provide simulations that more closely approximate experimentally acquired data. This approach is in line with the preprocessing consensus paper that recommends always averaging transients[39] instead of summing them. However, if one wishes to sum the transients, then the scaling factors need to be normalized by their sum so as to maintain the overall amplitude. Future work can implement more complex combination techniques.

2.1.11 | Frequency Drift and Phase Drift

Frequency drifts and phase drifts are phenomena observed in transients from multiple acquisitions in which each transient has an independent offset. These offsets and alignments are shown in Fig. 7.

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 the use of the simulated spectra in various software packages, they are also exported in the NIfTI-MRS format[51].

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. Customization of these parameters depends on the intended use and application of the dataset. For example, when creating deep learning-based quantification models, it is beneficial to use independent, uniform distributions that include all values the model will be expected to encounter. On the other hand, 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.

Peer Review Only

To create a dataset that accurately mimics in vivo conditions, it is crucial to have precise descriptions of clinical fitting parameters. This work does not provide scenario-specific parameter recommendations, as it is beyond the scope of this work. However, some tools are provided to assist in identifying ranges and distributions to match an existing dataset. In collaboration with the developers of Osprey[36], new functionality was added to their software to export the spectral fitting parameters after quantification. The tools in this framework can load those exported files and prepare the data for further analysis. Currently, this framework uses the Python library Fitter[52] to identify the best-fitting probability distribution for each parameter. *A priori* knowledge, either from prior knowledge or data exploration, can narrow down the search range and accelerate the analysis. The results for each parameter include evaluation metrics for the best performing distributions as well as a numerical characterization of the best fitting distribution.

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.

Metabolite and model parameter recommendations are provided as default values and are included in the Supporting Information Section 2 and the online repository. Model parameter values generally come from the default values of spectral fitting programs[36,53,24] while the parameters describing the metabolites come from literature[54,55,38,27,47,43]. For the most current information, please refer to the repository.

2.4.1 | Data Evaluation

Validating synthetic data is challenging because there is no gold standard and the "realness" of the data depends on both the simulation methods, i.e. equations and components, and the simulation parameter values. This makes directly evaluating the simulated spectra difficult. The scope of this work is focused on the simulation framework itself instead of recommending in vivo parameter values. Even though there is no quantitative metric describing the degree of in vivo realness, it is still important to evaluate the model outputs qualitatively. Therefore, the most appropriate approach for evaluating this protocol is to evaluate the simulation methods described above and then visually assess the outputs.

The selection of simulations presented in Sec. 3 focuses on short echo (TE=30ms) 3T PRESS spectra with a spectral width of 2000Hz and randomly sampled parameters. The concentrations were sampled with respect to total creatine (tCr). The choice of short echo spectra is motivated by their ability to capture a broader range of metabolite peaks compared to long echo spectra. The simulations presented include metabolites sampled from a comprehensive set of common brain metabolites including ascorbate (Asc), aspartate (Asp), choline (Ch), creatine (Cr), γ-aminobutyric acid (GABA), glutamine (Gln), glucose (Glu), glycerylphosphocholine (GPC), glutathione (GSH), lactate (Lac), myo-inositol (mI), N-ascetylaspartate (NAA), N-acetylaspartylglutamate (NAAG), phosphorylcholine (PCh), phosphocreatine (PCr), phosphoethanolamine (PE), scyllo-inositol (sI), taurine (Tau), and a variety of macromolecular and lipid basis functions. For consistency, the SNR was fixed to 15 and the chemical shift reference point was set to 4.65ppm. The metabolite concentrations and T_2^* values, adapted from Wyss et al.[43], were fixed to further emphasize the effects of the simulated artifacts. Except for Fig. 18 and Fig. 19, spectral baselines were also omitted for continuity and visual clarity. The samples shown in Figs. 8 to 15 show standard spectral components and artifacts. Fig. 16 then shows examples of simulating multi-coil transients. The samples in Fig. 17 show how the B_0 simulator affects a spectrum with varying mean inhomogeneity offsets. The samples in Fig. 18 show a single spectrum with various sampled baseline offsets and residual water signals. And finally, the samples in Fig. 19 demonstrate the complete model. Those samples were generated by sampling and randomly applying all spectral artifacts, except zero-order phase because of how much it obfuscates spectral details.

It is important to note that while pulse sequence implications are integral to the overall simulation process, their detailed exploration is beyond the scope of this work. Therefore, the simulations in Sec. 3 are restricted to a single pulse sequence and echo time. The main objective of this work is to manipulate pre-simulated basis sets until they approximate in vivo spectra, meaning the primary machinations of MRS-Sim are the same regardless of the basis functions used. Overall, the presented simulations using the PRESS sequence provide a valuable foundation for understanding the performance and capabilities of the MRS-Sim framework.

3 | RESULTS

3.1 | Standard Spectral Components

The following collection of figures show the effects of several standard spectral model components when their parameters are varied. When spectral components are not being studied, their values are fixed. Fig. 8 shows the effect of lorentzian line broadening in the absence of any

Peer Review Only

gaussian broadening. In the top row, the height of the y-axis drops from left-to-right showing the decrease in signal intensity. On the bottom row the differences are not as severe, so a single y-axis is used. Here it is more clearly shown that the spectral peaks are both losing intensity and widening. Fig. 9 uses the same display protocol to show the effects of gaussian broadening without lorentzian broadening whereas Fig. 10 shows several combinations of lorentzian and gaussian broadening that, when combined, give the Voigt lineshape. Fig. 10 breaks down the effects of the lorentzian and gaussian broadening on the Voigt lineshape in the top and bottom rows. On top, the gaussian value is fixed to 20Hz while scaling the T_2^* values using the scaling factors indicated in the captions. On the bottom, the lorentzian broadening is fixed and the gaussian values vary.

Fig. 11 shows two applications of frequency shifts, with all other parameters held constant. The top row shows how global frequency shifts translate spectra upfield and downfield. The bottom row shows the effects of giving each metabolite basis function a minor, randomly sampled frequency shift, which corresponds to fitting protocols. In this case, it is more appropriate to use informed frequency shift values for each metabolite rather than randomly sampling, like the global shifts, because this can increase undesirable spectral overlap.

Next, Fig. 12 and 13 show the effects of zero- and first-order phase offsets, respectively. The amount of phase added to each sample is specified in the caption underneath each plot. Then, Fig. 14 shows identical spectra with varying levels of spectral SNR. Most simulations should include noise, but this becomes especially important when simulating multi-coil transients and multiple acquisitions or editing sequences. The final standard spectral artifact is presented in Fig. 15, eddy currents. As described in Sec. 2.1.9, this model uses two variables: an amplitude and a time constant. The top row of Fig. 15 shows the effects of increasing the amplitude given a fixed time constant. The bottom then shows the effects of varying time constants given a fixed amplitude.

Instead of a spectral artifact, Fig. 16 shows one of the unique features of MRS-Sim: the ability to simulate groups of transients. These examples represent spectra acquired with an 8-channel receiver coil. The coil-combined spectra are shown in blue at the bottom. From left to right the spectral SNR of the coil-combined spectrum increases from 7.5 to 15. The first example was simulated with standard artifacts, but no phase or frequency drift. Global phase offsets are still visible, but the spectral SNR shows significant improvement when the transients are averaged. In the examples in the middle and the right, randomly sampled zero-order phase and frequency drifts were applied. The impact of these effects can be seen in their respective] coil-combined spectra which show improved noise profiles but diminished spectral intensity.

FIGURE 8 $T_2{}^*$ values were fixed and then scaled to represent varying magnitudes of lorentzian broadening. The spectral SNR = 15. No gaussian broadening, baselines, phase offsets, or eddy currents were included.

FIGURE 9 Gaussian broadening values were sampled from [0,70.17] in increasing order, as displayed beneath each plot. The lorentzian broadening was left unscaled (scaling coeffient = 1.0 corresponding to Fig. 8). The spectral SNR = 15 and phase offsets and eddy currents were omitted.

FIGURE 10 Theses samples explore the effects of lorentzian and gaussian broadening on the Voigt lineshape. The spectral SNR = 15. No phase offsets, eddy currents, or baselines were included. The top row varies the lorentzian broadening with a fixed gaussian value where the bottom row fixes the lorentzian broadening values and varies the gaussian.

FIGURE 11 Two types of frequency shifts are shown in these examples. The top row shows global frequency shifts. The bottom row shows the effects of using randomly sampled metabolite-level frequency shifts. The spectral SNR = 15 with fixed lorentzian and gaussian broadening. No baselines, phase offsets, or eddy currents were included.

FIGURE 12 Identical spectra are displayed with varying zero-order phase offsets as indicated below each plot. The lorentzian broadening was left unscaled, gaussian broadening = 20 Hz, spectral SNR = 15. First-order phase and eddy currents were omitted.

FIGURE 13 Identical spectra are displayed with varying first-order phase offsets as indicated below each plot. The lorentzian broadening was left unscaled. The gaussian broadening = 20 Hz and the spectral SNR = 15. Zero-order phase, and eddy currents were omitted.

FIGURE 14 Identical spectra are displayed with varying spectral SNRs as indicated below each plot. The lorentzian broadening was left unscaled and the gaussian broadening = 20 Hz. All phase offsets and eddy currents were omitted.

FIGURE 15 Eddy currents are simulated with two variables: an amplitude, A, and a time constant, t_c. In the top row, t_c is fixed to 0.15 seconds and the amplitudes are set to 1Hz, 5.5Hz, and 10Hz. In the bottom row, the amplitude is fixed to 5Hz and the time constants are set to 0.02 sec, 0.16 sec, and 0.30 sec. The values listed correspond to the plots from left to right.

FIGURE 16 These are examples of 8-channel simulated acquisitions. Their respective coil-combined spectra are shown at the bottom in blue. **(A)** shows low-SNR transients that have been phase- and frequency-drift aligned. **(B)** and **(C)** show transients with higher SNRs that have not been aligned, resulting in the coil-combined spectra having lower SNRs than otherwise expected.

3.2 | B_0 Field Inhomogeneity Simulator

Fig. 17 shows a short echo (TE=30ms) spectrum that is corrupted by varying levels of B_0 inhomogeneities. The first example is the reference and has Lorentzian lineshapes with a mean B_0 offset of 1 Hz. The center plot in the top row shows that even a 20 Hz mean offset has a large impact on the spectral intensity. The B_0 inhomogeneity increases from left-to-right and top-to-bottom from 1 Hz to 270 Hz with an interval size of 19 Hz. The total change in each dimension is between 2-10 Hz centered about the mean offset. The red spectra represent the real component from the previous example to highlight the effects of the increasing inhomogeneities. This figure shows that even moderate inhomogeneities noticeably impact the spectral resolution, signal intensity and frequency offset.

3.3 | Baseline and Residual Water Generator

Fig. 18 illustrates a single clean short echo (TE=30ms) PRESS spectrum showcasing various combinations of residual water and baseline contributions. The metabolite concentrations and lineshapes were fixed in addition to the settings described above. These samples omit spectral artifacts to more clearly highlight the variety that can be achieved by randomly sampling parameters for the baseline and residual water generator.

Fig. 18 shows a variety of baseline profiles that can be simulated with the presented framework. Instead of claiming to produce in vivo baselines, the proposed generator produces a wide variety of baseline offsets that approximate the baselines extracted by traditional methods. The motivation behind this approach is that if the generator cannot be limited to strictly in vivo-like baselines, then it should incorporate a variety of appropriate baseline profiles such that true baselines are included. Because this generator is not based on a baseline fitting method, future work could use it to evaluate the baseline modeling performance of various fitting protocols given different baseline profiles.

3.4 | Complete Model

In contrast, Fig. 19 presents a simulated PRESS spectrum (TE=30ms) with randomly sampled artifacts and offsets. This figure highlights the variety a single spectrum can assume just by randomly sampling the artifacts. Sophisticated simulations, such as these, are useful for developing and validating data processing techniques such as artifact removal and new spectral fitting protocols. Spectra can be simulated to be as clean or as artifact-laden as desired, all while maintaining known ground truth values.

Upon closer inspection of Fig. 19, the baselines and residual water regions sometimes appear to be uncorrelated with the depicted spectra, but they are in fact paired. These discrepancies are due to the additional artifacts that are applied to the simulations after all of the spectral components are combined. The plotted baseline and residual water regions are the ground truth signals and do not have those additional artifacts applied to them.

Incorporating a robust collection of artifacts and phenomena commonly encountered with in vivo data into these simulations, makes resulting datasets more closely approximate raw, in vivo data. Even though post-processing techniques are highly accurate, they have limitations and biases, leaving some residue of the corrected artifacts. Including these artifacts in the simulations and removing them with the users' own post-processing and fitting protocols ensures consistency between simulated and in vivo data. These artifacts can also be scaled down in the simulations to emulate residual artifacts. However, this is not ideal because it cannot be guaranteed that scaling down the artifacts will accurately reflect the biases and residues found in the in vivo data.

Deep learning techniques also benefit from robust and consistent synthetic data. Ensuring congruence of the synthetic training data with the in vivo data promotes better model performance. While including fine-grained detail, such as sampling moiety-level T_2 values, may not be essential for every use case, in a deep learning setting, it acts as a form of data augmentation. Such augmentation techniques maintain the underlying data but manipulate its appearance so that models learn a broader range of features instead of overfitting a single pattern. These techniques are standard practice in deep learning and are an inherent feature of this framework.

4 | DISCUSSION

This work presents a flexible, open-source framework to help researchers generate complex, in vivo-like MRS data. Spectral components from state-of-the-art spectral fitting models were combined to create a comprehensive model. Flexibility and modularity are key components that make this framework a viable option for simulating in vivo-like data representing a multitude of clinical scenarios.

The use of synthetic MRS data has become increasingly popular in recent years due to the need for large datasets with known spectral components. Using synthetic data allows researchers to create unlimited spectra with known ground truth values. This is useful for both traditional spectroscopy and deep learning applications, including evaluating the accuracy and precision of data analysis methods. However, there is a lack of standardization in synthetic data generation, which poses challenges in terms of reproducibility and generalizability.

Peer Review Only

FIGURE 17 Sample spectrum simulated for a PRESS sequence with TE=30ms that highlights the effect of B_0 inhomogeneities on the spectra. The mean inhomogeneities range from [1, 270] Hz and monotonically increase by 19Hz. As the inhomogeneities increase, the intensity of the spectra decreases, shown by the y-axis range, and causes a global frequency shift.

FIGURE 18 Sample spectra simulated for a PRESS sequence with TE=30ms that highlight the effect of the baseline and residual water contributions.

FIGURE 19 Sample spectra, similar to Fig. 18 [PRESS, TE=30ms], with randomly sampled, uncorrected artifacts to approximate raw data.

Researchers currently use a variety of simulation frameworks to generate synthetic data for their particular needs. These frameworks often involve simplifying assumptions regarding the spectral components, which excludes some components that are deemed non-essential for that particular task. These assumptions make synthetic data less generalizable, similar to phantom data which struggles to approximate in vivo data. In general, simulating data has two primary challenges: which spectral components and artifacts to include and what parameter values to use for the included components. These choices can be difficult for experts, let alone non-expert users. Moreover, the heterogeneity in the simulation methods used makes reproducibility challenging.

MRS-Sim addresses these challenges as an open-source solution for a synthetic data generation model that includes a comprehensive list of spectral components and acquisition-induced artifacts. Several novel contributions are presented to include experimental artifacts that are usually ignored during spectral modeling. These modules are designed to make the simulations more comprehensive and in vivo-like. The first contribution is a B_0 field simulator. It is capable of introducing distortions due to imperfect shimming and high susceptibility effects into the simulations. Then, a semi-parametric generator is proposed that can simulate both broad, undulating baseline offsets and highly irregular residual water contributions. The implementation of the presented framework is flexible and can be updated with new information as research progresses to improve the realness and accuracy of the simulations.

The modularity of the MRS-Sim implementation is one of its key features, allowing for flexible and customizable use of the framework. Firstly, the functions that add artifacts and spectral components are modular and independent, which makes it very easy to understand the code and its purpose. Secondly, it allows for the easy generation of highly customized datasets for various clinical scenarios, as the simulation step itself is also modular. Consequently, this feature enables easy community contributions to develop new functionalities, thus preventing the need for entirely new frameworks. New scenarios, such as J-difference edited spectra, diffusion spectra, or 2D spectra, can be developed and seamlessly integrated.

The appendix and digital repository include a table that provides up-to-date information about moiety- and metabolite-level characterizations of things like spin-systems, temperature-induced artifacts, T_2 values, and metabolite concentration ranges. This information can be continually updated to provide the community with comprehensive and state-of-the-art information about brain metabolites at a glance. Additionally, these suggested values provide a good starting point for simulating datasets. Cumulatively, this framework offers a pathway to generate realistic, in vivo-like synthetic data that captures acquisition-induced artifacts and nuisance signals and can be applied across a range of MRS applications.

5 | CONCLUSIONS

The development of open-source frameworks for data generation is crucial to ensure that synthetic MRS data is widely adopted and to improve the generalizability of synthetic data and the reproducibility of MRS research. By making this work available to a larger community of researchers, the authors hope to further democratize MRS research by providing easy access to high-quality simulations and supporting information. Improvements in the quality of synthetic data will ultimately lead to better generalizability to in vivo data, directly improving the overall applicability of synthetic data in MRS. Moving forward, collaboration and consensus-building among researchers will be essential to establish standards and best practices for simulating MRS data. Future development can focus on adding additional clinical scenarios to the framework's repertoire, increasing its applicability to more fields within MRS. By addressing the challenges associated with synthetic MRS data generation and promoting standardization, MRS can continue to advance in every aspect of the field.

ABBREVIATIONS

Asc ascorbate.

Asp aspartate.

Ch choline.

Cr creatine.

DL deep learning.

GABA gamma-aminobutyric acide.

Gln glutamine.

Glu glutamate.

GPC glycerophosphorylcholine.

GSH glutathione.

Lac lactate.

LASER localized adiabatic selective refocusing.

MARSS Magnetic Resonance Spectrum Simulator.

mI myo-inositol.

ML machine learning.

NAA N-acetylaspartate.

NAAG N-acetylaspartylglutamate.

PCh phosphocholine.

PCr phosphocreatine.

PE phosphatidylethanolamine.

For Peer Review Only

PRESS point resolved spectroscopy.

sI scyllo-inositol.

sLASER semi-localized adiabatic selective refocusing.

STEAM stimulated echo acquisition method.

Tau taurine.

tCr total Creatine.

ACKNOWLEDGMENTS

The authors would like to thank Dr. Tobias Lasser for his mentorship during this work; Dr. Dhritiman Das for discussions about his experience with simulating MRS SVS data; the Harper House for their support while visiting UCSF to further develop the physics model; and the DAAD and Technical University of Munich's Graduate Center of Computation, Information and Technology for helping to finance the stay at UCSF in San Francisco, USA.

Funding Information

This work was supported in part by a doctoral fellowship from the German Academic Exchange Service (DAAD) and the following grants from the National Institutes of Health: R00 AG062230, R01 CA262630, and R21 EB033516.

Conflict of interest

The authors declare no potential conflict of interest.

References

1. Augustine EM, Spielman DM, Barnes PD, et al. Can magnetic resonance spectroscopy predict neurodevelopmental outcome in very low birth weight preterm infants?. *Journal of Perinatology* 2008; 28(9): 611–618. doi: 10.1038/jp.2008.66
2. Laccetta G, De Nardo MC, Cellitti R, Angeloni U, Terrin G. 1H-magnetic resonance spectroscopy and its role in predicting neurodevelopmental impairment in preterm neonates: A systematic review. *Neuroradiology Journal* 2022; 35(6): 667–677. doi: 10.1177/19714009221102454
3. Tomiyasu M, Shibasaki J, Kawaguchi H, et al. Altered brain metabolite concentration and delayed neurodevelopment in preterm neonates. *Pediatric Research* 2022; 91(1): 197–203. doi: 10.1038/s41390-021-01398-6
4. Gao F, Barker PB. Various MRS application tools for Alzheimer disease and mild cognitive impairment. *American Journal of Neuroradiology* 2014; 35(6 SUPPL.): 1–17. doi: 10.3174/ajnr.A3944
5. Martin WR. MR spectroscopy in neurodegenerative disease. *Molecular Imaging and Biology* 2007; 9(4): 196–203. doi: 10.1007/s11307-007-0087-2
6. McKiernan E, Su L, O'Brien J. MRS in neurodegenerative dementias, prodromal syndromes and at-risk states: A systematic review of the literature. *NMR in Biomedicine* 2023(January): 1–13. doi: 10.1002/nbm.4896
7. Öz G., ed.*Magnetic Resonance Spectroscopy of Degenerative Brain Diseases*. Contemporary Clinical NeuroscienceCham: Springer International Publishing . 2016
8. Cecil KM. MR Spectroscopy of Metabolic Disorders. *Neuroimaging Clinics of North America* 2006; 16(1): 87–116. doi: 10.1016/J.NIC.2005.10.004
9. Gropman AL, Anderson A. Novel imaging technologies for genetic diagnoses in the inborn errors of metabolism. *Journal of Translational Genetics and Genomics* 2020: 429–445. doi: 10.20517/jtgg.2020.09
10. Lai LM, Gropman AL, Whitehead MT. MR Neuroimaging in Pediatric Inborn Errors of Metabolism. *Diagnostics* 2022; 12(4). doi: 10.3390/diagnostics12040861

Peer Review Only

11. Calvar JA, Meli FJ, Romero C, et al. Characterization of brain tumors by MRS, DWI and Ki-67 labeling index. *Journal of Neuro-Oncology* 2005; 72(3): 273–280. doi: 10.1007/s11060-004-3342-2

12. Lukas L, Devos A, Suykens JA, et al. Brain tumor classification based on long echo proton MRS signals. *Artificial Intelligence in Medicine* 2004; 31(1): 73–89. doi: 10.1016/J.ARTMED.2004.01.001

13. Padelli F, Mazzi F, Erbetta A, et al. In vivo brain MR spectroscopy in gliomas: clinical and pre-clinical chances. *Clinical and Translational Imaging* 2022; 10(5): 495–515. doi: 10.1007/s40336-022-00502-y

14. Nelson SJ. Multivoxel Magnetic Resonance Spectroscopy of Brain Tumors1. *Molecular Cancer Therapeutics* 2003; 2(5): 497–507.

15. Forester BP, Berlow YA, Harper DG, et al. Age-related changes in brain energetics and phospholipid metabolism. *NMR in Biomedicine* 2010; 23(3): 242–250. doi: 10.1002/nbm.1444

16. Inglese M, Ge Y. Quantitative MRI. *Topics in Magnetic Resonance Imaging* 2004; 15(6): 355–363. doi: 10.1097/01.rmr.0000168069.12985.15

17. Reyngoudt H, Claeys T, Vlerick L, et al. Age-related differences in metabolites in the posterior cingulate cortex and hippocampus of normal ageing brain: A 1H-MRS study. *European Journal of Radiology* 2012; 81(3): e223–e231. doi: 10.1016/j.ejrad.2011.01.106

18. Dezortova M, Hajek M. 1H MR spectroscopy in pediatrics. *European Journal of Radiology* 2008; 67(2): 240–249. doi: 10.1016/j.ejrad.2008.02.035

19. Li Y, Srinivasan R, Ratiney H, Lu Y, Chang SM, Nelson SJ. Comparison of T1 and T2 metabolite relaxation times in glioma and normal brain at 3T. *Journal of Magnetic Resonance Imaging* 2008; 28(2): 342–350. doi: 10.1002/jmri.21453

20. Hatami N, Sdika M, Ratiney H. Magnetic resonance spectroscopy quantification using deep learning. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* 2018; 11070 LNCS: 467–475. doi: 10.1007/978-3-030-00928-1$_53$

21. Das D, Davies ME, Chataway J, Chandran S, Menze BH, Marchall I. Metabolite Quantification of 1H-MRSI spectra in Multiple Sclerosis: A Machine Learning Approach. *ISMRM* 2018.

22. Das D, Thrippleton MJ, Semple SIK, et al. MRSI Brain Temperature Mapping Using Machine Learning. *ISMRM* 2018: 2–6.

23. Iqbal Z, Nguyen D, Thomas MA, Jiang S. Deep learning can accelerate and quantify simulated localized correlated spectroscopy. *Scientific Reports* 2021: 1–17. doi: 10.1038/s41598-021-88158-y

24. 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. *Magnetic Resonance in Medicine* 2017; 77(1): 23–33. doi: 10.1002/mrm.26091

25. Soher BJ, Young K, Bernstein A, Aygula Z, Maudsley AA. GAVA : Spectral simulation for in vivo MRS applications. *Journal of Magnetic Resonance* 2007; 185: 291–299. doi: 10.1016/j.jmr.2007.01.005

26. Soher BJ, Semanchuk P, Todd D, Ji X. VeSPA: Integrated applications for RF pulse design, spectral simulation and MRS data analysis. *Magnetic Resonance in Medicine* 2023; 90: 823–838. doi: 10.1002/mrm.29686

27. Landheer K, Swanberg KM, Juchem C. Magnetic resonance Spectrum simulator (MARSS), a novel software package for fast and computationally efficient basis set simulation. *NMR in Biomedicine* 2021; 34(5): 1–13. doi: 10.1002/nbm.4129

28. Das D, Coello E, Schulte RF, Menze BH. Quantification of metabolites in magnetic resonance spectroscopic imaging using machine learning. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* 2017; 10435 LNCS: 462–470. doi: 10.1007/978-3-319-66179-7$_53$

29. Gurbani SS, Sheriff S, Maudsley AA, Shim H, Cooper LA. Incorporation of a spectral model in a convolutional neural network for accelerated spectral fitting. *Magnetic Resonance in Medicine* 2019; 81(5): 3346–3357. doi: 10.1002/mrm.27641

30. Lee HH, Kim H. Intact metabolite spectrum mining by deep learning in proton magnetic resonance spectroscopy of the brain. *Magnetic Resonance in Medicine* 2019; 82(1): 33–48. doi: 10.1002/mrm.27727

Peer Review Only

31. Craven AR, Bhattacharyya PK, Clarke WT, et al. Comparison of seven modelling algorithms for γ-aminobutyric acid–edited proton magnetic resonance spectroscopy. *NMR in Biomedicine* 2022; 35(7): 1–17. doi: 10.1002/nbm.4702

32. Marjańska M, Terpstra M. Influence of fitting approaches in LCModel on MRS quantification focusing on age-specific macromolecules and the spline baseline. *NMR in Biomedicine* 2021; 34(5): 1–9. doi: 10.1002/nbm.4197

33. Marjańska M, Deelchand DK, Kreis R, et al. Results and interpretation of a fitting challenge for MR spectroscopy set up by the MRS study group of ISMRM. *Magnetic Resonance in Medicine* 2022; 87(1): 11–32. doi: 10.1002/mrm.28942

34. Zöllner HJ, Považan M, Hui SC, Tapper S, Edden RA, Oeltzschner G. Comparison of different linear-combination modeling algorithms for short-TE proton spectra. *NMR in Biomedicine* 2021; 34(4): 1–17. doi: 10.1002/nbm.4482

35. Provencher SW. Automatic quantitation of localized in vivo 1H spectra with LCModel. *NMR in Biomedicine* 2001; 14(4): 260–264. doi: 10.1002/nbm.698

36. Oeltzschner G, Zöllner HJ, Hui SC, et al. Osprey: Open-source processing, reconstruction estimation of magnetic resonance spectroscopy data. *Journal of Neuroscience Methods* 2020; 343: 1–27. doi: 10.1016/j.jneumeth.2020.108827

37. Clarke WT, Stagg CJ. FSL-MRS : An end-to-end spectroscopy analysis package. *Magnetic Resonance in Medicine* 2021; 85: 2950–2964. doi: 10.1002/mrm.28630

38. Gudmundson AT, Koo A, Virovka A, et al. Meta-analysis and open-source database for in vivo brain Magnetic Resonance spectroscopy in health and disease. *Analytical Biochemistry* 2023; 676(July): 115227. doi: 10.1016/j.ab.2023.115227

39. Near J, Harris AD, Juchem C, et al. Preprocessing, analysis and quantification in single-voxel magnetic resonance spectroscopy: experts' consensus recommendations. *NMR in Biomedicine* 2021; 34(5): 1–23. doi: 10.1002/nbm.4257

40. Larsen RJ, Gagoski B, Litt J, Morton SU, Grant PE, Sutton BP. Quantification of magnetic resonance spectroscopy data using a combined reference : Application in typically developing infants. *NMR in Biomedicine* 2021: 1–16. doi: 10.1002/nbm.4520

41. Wilson M, Reynolds G, Kauppinen RA, Arvanitis TN, Peet AC. A constrained least-squares approach to the automated quantitation of in vivo 1H magnetic resonance spectroscopy data.. *Magnetic resonance in medicine : official journal of the Society of Magnetic Resonance in Medicine / Society of Magnetic Resonance in Medicine* 2011; 65(1): 1–12. doi: 10.1002/mrm.22579

42. Stefan D, Cesare FD, Andrasescu A, et al. Quantitation of magnetic resonance spectroscopy signals: The jMRUI software package. *Measurement Science and Technology* 2009; 20(10). doi: 10.1088/0957-0233/20/10/104035

43. Wyss PO, Bianchini C, Scheidegger M, et al. In Vivo Estimation of Transverse Relaxation Time Constant (T2) of 17 Human Brain Metabolites at 3T. *Magnetic Resonance in Medicine* 2018; 80(2): 452–461. doi: 10.1002/mrm.27067

44. Li N, An L, Shen J. Spectral fitting using basis set modified by measured B0 field distribution. *NMR in Biomedicine* 2015; 28(12): 1707–1715. doi: 10.1002/nbm.3430

45. Bazgir O, Mitra S, Nutter B, Walden E. Fully automatic baseline correction in magnetic resonance spectroscopy. *Proceedings of the IEEE Southwest Symposium on Image Analysis and Interpretation* 2018; 2018-April: 97–100. doi: 10.1109/SSIAI.2018.8470319

46. Wilson M. Adaptive baseline fitting for 1H MR spectroscopy analysis. *Magnetic Resonance in Medicine* 2021; 85(1): 13–29. doi: 10.1002/mrm.28385

47. Wermter FC, Mitschke N, Bock C, Dreher W. Temperature dependence of 1H NMR chemical shifts and its influence on estimated metabolite concentrations. *Magnetic Resonance Materials in Physics, Biology and Medicine* 2017; 30(6): 579–590. doi: 10.1007/s10334-017-0642-z

48. Klose U. In vivo proton spectroscopy in presence of eddy currents. *Magnetic Resonance in Medicine* 1990; 14(1): 26–30. doi: 10.1002/mrm.1910140104

49. Kreis R, Boer V, Choi IY, et al. *Terminology and concepts for the characterization of in vivo MR spectroscopy methods and MR spectra: Background and experts' consensus recommendations*. No. May 2020John Wiley Sons Ltd . 2020

50. Hall EL, Stephenson MC, Price D, Morris PG. Methodology for improved detection of low concentration metabolites in MRS: Optimised combination of signals from multi-element coil arrays. *NeuroImage* 2014; 86: 35–42. doi: 10.1016/j.neuroimage.2013.04.077

Peer Review Only

51. Clarke WT, Bell TK, Emir UE, et al. NIfTI-MRS: A standard data format for magnetic resonance spectroscopy. *Magnetic Resonance in Medicine* 2022; 88(6): 2358–2370. doi: 10.1002/mrm.29418

52. Cokelaer T. FITTER documentation — fitter 1.5.2 documentation. 2019.

53. Provencher SW. Estimation of metabolite concentrations from localized in vivo proton NMR spectra. *Magnetic Resonance in Medicine* 1993; 30(6): 672–679. doi: 10.1002/mrm.1910300604

54. Graaf dRA. *In Vivo NMR Spectroscopy: Principles and Techniques*. Wiley. 2nd ed. 2007

55. Graaf dRA. *In Vivo NMR Spectroscopy: Principles and Techniques*. Wiley. 3rd ed. 2019

How to cite this article: LaMaster J., G. Oeltzschner, and Y. Li (2022), MRS-Sim: Open-Source Framework for Simulating Realistic In Vivo-like MR Spectroscopy Data, *NMR in Biomedicine*.

Peer Review Only