**Supplements**

**Supplement 1. Non-linear shrinkage (NLS)**

In this section, we present a new method for phase error correction called non-linear shrinkage (NLS). The NLS method uses power and magnitude spectra to estimate an absorption spectrum. The implementation of non-linear shrinkage is as follows:

1. If dispersion spectra are not available, estimate the theoretical dispersion spectra with the Hilbert transform.

The Hilbert transform, processed indirectly through a discrete Fourier transform, proved to be better than a direct Hilbert transform, as discussed by Ernst (1969). The procedure of the transformation in our current research was as follows (<https://www.mathworks.com/help/signal/ref/hilbert.html#bqijsn8-1>):

a) Use an inverse discrete Fourier transform to get the time domain data from a given absorption spectrum with *N* data points (<http://www2.math.ethz.ch/education/bachelor/seminars/fs2008/nas/woerner.pdf>):

Here, *Ak*is an observed absorption intensity at index *k*, *FIDk* is the corresponding time domain data point at index *k*.

b) Define *hk* as:

c) Calculate *FIDk* × *hk*, which transforms the entire time domain data FID back to the frequency domain through the discrete Fourier transformation. The transformed frequency domain data contains both absorption and dispersion spectra.

1. Calculate the corresponding magnitude spectrum with the formula:

represents the magnitude intensity for the *k*th observation; *Ak* and *Dk* are the absorption and dispersion intensities respectively at the *k*th index.

1. Detect the major peaks in a magnitude spectrum with continuous wavelet transform-based pattern matching (Du et al., 2006). Wavelet is a mathematical function to transform data from one domain to another in order to analyze data from a different perspective. The wavelet coefficients and wavelets can be defined as follows:

(Mexican hat wavelet)

*C(a,b)* is the matrix of the wavelet coefficients, which depends on parameters *a* and *b;* *M(x)* is our input magnitude spectrum; is the mother wavelet – we use Mexican hat as the mother wavelet; is the scaled and transformed wavelet; *a* is the scale – it is a positive number; *b* is the centre of a wavelet – it could be any real number, in fact, *b* is the index of the magnitude maximum of a peak; *x* represents the index of a spectrum.

1. Define valleys as the global minima between two major peaks. The indexes of two neighboring valleys are defined as a sub-range, which could contain one or more peaks.

is the index for the *l*th valley; *k* is the index of a magnitude spectrum; is the index of the *l*th peak maximum; is the index of the (*l+1)*th peak maximum; and *M(k)* is the magnitude intensity for the *k*th point in the spectrum. We also need to add 1 at the beginning of a valley index vector and add the length of a spectrum into the end of a valley index vector to ensure that a valley index vector can divide all points into sub-ranges.

1. Within each sub-range, apply the following shrinkage formula to the power sub-range to get an estimated absorption sub-range:
2. Combine estimated absorption intensities from all sub-ranges into a full spectrum.

# The Non-linear Shrinkage (NLS) method is a simple and effective approach for phase error correction. By using power and magnitude spectra, we can estimate absorption spectra without any systematic phase errors. The NLS method has been mathematically proven and shown to be intuitive through figures. This method can be applied to various spectroscopic techniques and has the potential to improve spectral analysis and interpretation.

## **Supplement 2. Phase error correction with multiple linear phase models (MPC)**

Instead of a single linear model for an entire spectrum, we propose using multiple models for phase error correction. The main idea is to apply a single linear model for each sub-ranges of a spectrum, and the procedure to get sub-ranges is the same as in non-linear shrinkage. The detailed procedure for multiple models is as follows:

1. Estimate the theoretical dispersion spectra with the Hilbert transform as described in Supplement 1 if dispersion spectra are not available.
2. Calculate the corresponding magnitude spectrum with the following formula:

represents the magnitude intensity for the *k*th observation, where Ak and Dk, which are the absorption and dispersion intensity, respectively, for the *k*th observation.

1. Detect the major peaks in a magnitude spectrum with continuous wavelet transform-based pattern matching (Du, Kibbe and Lin 2006). The wavelet coefficients and wavelets are defined as follows:

(Mexican hat wavelet)

The matrix of wavelet coefficients, *C(a,b)*, depends on the parameters *a* and *b*; *M(x)* is our input magnitude spectrum; is the mother wavelet – we use Mexican hat as the mother wavelet; is the scaled and transformed wavelet; *a* is the scale – it is a positive number; *b* is the centre of a wavelet – it could be any real number, in fact, *b* is the index of the magnitude maximum of a peak; *x* represents the index of a spectrum.

1. Define valleys as the global minima between two adjacent major peaks. The indexes of two neighboring valleys are defined as a sub-range, which could contain one or more peaks.

is the index for the *l*th valley; *k* is the index of a magnitude spectrum; is the index of the *l*th peak maximum; is the index of the (*l+1)*th peak maximum; and *M(k)* is the magnitude intensity for the *k*th point in the spectrum. Additionally, add 1 at the beginning and the length of the spectrum at the end of a valley index vector to ensure it divides all points into sub-ranges.

1. Apply a linear phase correction model to each sub-range, then combine the phase-corrected absorption intensities from all sub-ranges into a full spectrum.

The linear phase correction models for each sub-range can be written as follows:

Here, *l* represents the *l*th sub-range of a spectrum, assuming that we have *m* sub-ranges in total; is the index for the observation for the *l*th sub-range, which varies from 1 to the length of the sub-range; dependent variable represents the phase correction value at the th index; intercept is often called the zero-order parameter; slope is often called the first-order parameter; independent variable is just the scaled index within the *l*th sub-range; and is the total number of observations in the *l*th sub-range.

An optimization function is required to estimate the parameters and for each sub-range. Once we estimate the optimal parameters and *bl*, we then apply the linear phase correction model to determine the phase correction values for all points in a sub-range. The absorption intensity after phase error correction is calculated as follows:

Here, represents absorption intensity adjusted with the phase correction value at index ; is the observed absorption intensity at index ; and is the corresponding dispersion intensity at index . If the dispersion spectrum is not available, calculate a theoretical dispersion spectrum corresponding to the observed absorption spectrum using the Hilbert transform as described in Supplement 1.

Phase error correction with multiple models involves correcting the phase errors within each sub-range separately. Within each sub-range, the process is identical to using a single model.

## **Supplement 3. Optimization functions**

Phase error correction using either a single linear model (SPC: single linear phase correction model) or multiple models (MPC: multiple linear phase correction models) requires optimization to find the optimal numerical parameters. Different optimization functions produce different optimal parameters, and thus different phased spectra. In this section, we summarize four optimization functions used in our current research, including our new method and three commonly used ones, in the following table (Table S1).

Table S1. Optimization functions used in the current research

|  |  |
| --- | --- |
| Optimization function formula | Name and description of optimization function |
|  | 1) *Absolute area* m*inimization* (AAM)  This method is based on the idea that a phase error-free absorption spectrum should have the minimum integral of absolute intensity values among corresponding absorption spectra with all possible phase errors (de Brouwer, 2009; Džakula, 2000). |
|  | 2) *Entropy minimization with penalty* (EMP)  This method is used to obtain the optimal parameters that not only achieve minimal entropy but also penalize negative signals that should not be present in a phase-free absorption spectrum. Here, entropy is defined as the negative summation of absolute intensity times the logarithm of the absolute intensity (Binczyk et al., 2015), , and we use the sum of squared negative values as the penalty (de Brouwer, 2009). |
|  | 3) *Dispersion summation minimization* (DSM)  This method is based on the idea that the integral of the dispersion spectrum should be close to zero when there are no phase errors (Binczyk et al., 2015). |
|  | 4) *Delta absolute net minimization* (DANM)  This is our new method that considers both absolute and net areas. While minimizing absolute area, we also want to maximize net area for an absorption spectrum that in fact minimizes negative values. This is to minimize the difference between absolute area under a curve and net area under a curve. |

Among the four optimization functions, delta absolute net minimization (DANM) was developed by us. It is an optimization technique that aims to minimize the difference between the absolute area under a curve and the net area under the same curve.

The absolute area under a curve is the total area that lies above the x-axis, regardless of the direction of the curve, whereas the net area under a curve is the total area that is enclosed by the curve and the x-axis, taking into account the direction of the curve. The difference between these two areas can be used as a measure of the distortion or asymmetry of the curve, and minimizing this difference can help to obtain a more symmetrical or well-behaved curve.

The DANM technique can be applied in various fields, such as signal processing, image analysis, and data visualization, to optimize the shape or properties of curves or functions.

## **Supplement 4. Optimization process**

To estimate optimal parameters for single or multiple models, an optimization process is used based on the optimization function presented in Table S1. These parameters are then used for phase error correction with a single linear model or multiple linear phase models, respectively.

The optimization process for a single phase correction model is applied to the whole spectrum, while multiple linear phase correction models are applied to individual sub-ranges separately. However, the annotation in this section applies to both situations without specifying the sub-range index.

The optimization process setting is the same for any optimization function and is defined as:

(Rarely, a high order linear model could be used)

Here, is the phase correction value at index *k*; represents the absorption intensity adjusted with phase correction value at index *k*, while represents the dispersion intensity adjusted with phase correction value at index *k*. is the observed absorption intensity at index *k*, and is the corresponding observed dispersion intensity at index *k*. is an ordered scaled index in a spectrum or a sub-range. *a* is the intercept and *b* is the slope, which need to estimate. N is the total number of points within a spectrum or a sub-range.

The initial values for the intercept and slope are defined as and, respectively. Here, (*Am*, *Dm*) is the pair of absorption and dispersion values at the maximum magnitude value in a spectrum or a sub-range:

## **Supplement 5 Simulations buildup based on online real metabolite information**

To make our simulations as close as possible to real NMR data, we used real metabolite information available online to build up an NMR spectrum as our standard idealized spectrum, and then added different types of errors to it to be our simulated experimental spectra. The whole process is described in the following.

### 5.1 Simulating our idealized spectrum

5.1.1 Metabolite list

On advice from an NMR expert, we began with a list of 47 metabolites commonly detected by NMR.

5.1.2 Metabolite searching

We conducted a search for 1D 1H NMR data for these 47 metabolites originally available at the Madison Metabolomics Consortium Database (MMCD) (http://mmcd.nmrfam.wisc.edu/). Due to the unavailability of the original link (archived link: http://web.archive.org/web/20210130163136/http://mmcd.nmrfam.wisc.edu/), similar data can be accessed through other metabolomics resources, such as the Human Metabolome Database (HMDB) (https://hmdb.ca/). To ensure that the metabolite peak features were comparable, we specifically searched for data from 1D 1H NMR with matching pH and concentrations to the internal reference. Ultimately, we found comparable 1D 1H NMR data for 37 out of the 47 metabolites.

5.1.3 Features of matched metabolites

Initially, we intended to directly use the FID time domain data from the Madison Metabolomics Consortium Database (MMCD) for these 37 metabolites, but unfortunately, we did not achieve error-free standard spectra plotted at the website with or without preprocessing, even after trying several different software programs. Some of the spectra peaks were distorted in shape; and others had a large number of negative values, even though the absolute values of these negative intensities were not very large. Therefore, we decided to simulate these peaks with metabolite peak information that was available online.

We simulated Lorentzian functions based on the metabolite peak information from the website to get idealized metabolite spectra. We recorded the peak information for the 37 metabolites from the above website, including peak ppm location and height. Ideally, we would have also included peak area, but unfortunately, the peak area data were not available at the Madison Metabolomics Consortium Database (MMCD). The information for all peaks was recorded when a metabolite had multiple peaks.

5.1.4 Lorentzian functions

To depict each peak for these 37 metabolites, we simulated a Lorentzian function for each peak over a full ppm range. To determine the location and peak height, as found from the website, we first simulated a Lorentzian function with the desired ppm location and scale = 0.0035 to get an initial simulation for each peak. We then used the peak height from the website to rescale the simulation and arrive at the desired height.

The above formula was used to rescale each point (*i* = 1, 2…, N; where N was the total number of points in a simulation) from the initial simulated intensity value to the final simulated intensity value by the ratio of the desired peak height to the initial peak height , while the ppm values were kept the same.

We then compared our 37 metabolite simulations with the corresponding 37 standard metabolite 1D 1H NMR plots from the Madison Metabolomics Consortium Database (MMCD). We retained 10 of the 37 metabolite simulations that were closest to their standard NMR plots, resulting in a total of 75 peaks.

5.1.5 Idealized spectrum

The 10 metabolite simulations that were closest to their corresponding online standard NMR plots were: alanine, aspartate, beta-alanine, betaine, fumarate, glucose, lactate, malate, propionate, and succinate. We also simulated an internal reference DSS (4,4-dimethyl-4-silapentane-1-sulfonic acid) spectrum having an arbitrary peak height of 1 and location at ppm=0 as required. These 11 simulations were defined as the single metabolite idealized spectra.

Without loss of generality, we set all scaling/concentration factor values for these 11 metabolite spectra to 1; i.e., the multiple metabolite spectra summations/combinations were done without a concentration/scaling factor. These 11 single metabolite idealized spectra were combined to become our standard idealized spectrum, which contains 76 peaks including a DSS reference peak.

Our simulations did not contain solvent peaks for two reasons: first, the Madison Metabolomics Consortium Database (MMCD) did not provide solvent information; second, solvent peaks are often removed or ignored in practice.

### 5.2 Adding multiple types of errors

To simulate realistic spectra, we introduced noise, phase error, and baseline bias into the idealized single metabolite spectra or their combinations. The following subsections describe how we added each type of error.

5.2.1 Noise

We considered two types of noise: intensity-dependent and intensity-independent.

5.2.1.1 Intensity-dependent noises

We added intensity-dependent noise separately to each of the 11 idealized single metabolite spectra, and then combined them into a single spectrum with intensity-dependent noise. To add the noise, we assumed that the standard deviation for the noise in the *jth* metabolite spectrum, *σj*, followed a Beta distribution: ~ Beta (1, 100). We then randomly selected an intensity multiplier for the *jth* metabolite from a normal distribution: ). Next, we calculated the intensity of the *ith* point in the *jth* metabolite spectrum with the added noise as follows:

After adding the intensity-dependent noise to all 11 single metabolite spectra, we combined them into a single spectrum for all 11 metabolites with intensity-dependent noise. We set all scaling/concentration factor values for these 11 metabolite spectra with noise to 1, meaning that the multiple metabolite spectrum summation/combination was done without a concentration/scaling factor.

5.2.1.2 Intensity independent noise

Without loss of generality, we assumed that the variance of intensity-independent noise followed a Chi-square distribution with 1 degree of freedom: V ~ . We then sampled intensity-independent noise from a normal distribution with mean 0 and variance V for each simulation point of the combined spectra from Section S5.2.1.1, and added this noise to each intensity value.

5.2.2 Phase errors

We introduced both global and local phase errors into the spectra using a linear phase model. This model, typically employed for phase error correction in NMR data processing, was adapted here to simulate the phase errors instead. The model includes an intercept *a*, which corresponds to the global phase error, and a slope *b*, which corresponds to the local phase errors. The model is defined as follows:

Here, represents the phase error at index *k*, *C* is the index for the maximum absolute intensity of the spectrum, and *N* is the length of the spectrum. We assume that both *a* and *b* are normally distributed with their own means and variances.

To estimate the population means and variances for the intercept and slope in the phase error model, we extracted the phase error correction parameters from the raw NMR data for all 37 metabolites downloaded from the Madison Metabolomics Consortium Database (MMCD). Table S2 lists the intercept and slope values for the phase correction models for each of the 37 metabolites.

We derived the parameters for two normal distributions, and , using the data from the above Table S2. However, since our simulations required adding phase errors and Table S2 provided parameters for correcting phase errors, we used the means from the table with opposite signs to introduce phase errors into the simulations. Specifically, we estimated using -mean(P0) from Table S2 and using -mean(P1) from Table S2. We directly estimated and from Table S2 since variance is always positive.

Table S2. Phase error correction parameters for 37 downloaded metabolites in the phase correction models from the Madison Metabolomics Consortium Database (MMCD).

|  |  |  |
| --- | --- | --- |
| Metabolite | Intercept P0 (in degrees) | Slope P1 (in degrees) |
| acetate | 35.32 | -12.99 |
| alanine | 40.86 | -11.88 |
| AMP | -70.52 | -124.36 |
| arginine | -56.36 | -132.59 |
| aspartate | 42.97 | -10.49 |
| beta-alanine | -53.20 | -140.90 |
| betaine | 305.14 | -136.42 |
| choline | 298.53 | -124.83 |
| cystathionine (L-cystathionine) | 26.61 | 0.00 |
| formate | 36.24 | -13.72 |
| fumarate | -156.05 | -9.79 |
| glucose | -73.88 | -132.06 |
| glutamate | -72.65 | -124.74 |
| glutamine | 46.67 | -14.00 |
| glutathione | 226.63 | -91.93 |
| glycerol | -65.61 | -133.41 |
| glycine | -63.37 | -131.33 |
| GTP (dGTP) | 33.45 | -12.48 |
| histidine | -64.11 | -131.47 |
| inosine | 32.79 | -3.60 |
| isoleucine | 297.07 | -137.17 |
| lactate | 42.36 | -10.79 |
| leucine | 309.10 | -149.04 |
| lysine | -57.72 | -136.73 |
| malate | 295.08 | -126.38 |
| NAD+ | -60.01 | -131.33 |
| ornithine | 41.86 | -12.78 |
| phenylalanine | -57.83 | -137.85 |
| proline | 305.74 | -142.91 |
| propionate | 31.76 | -13.95 |
| serine | 22.33 | 3.39 |
| succinate | 296.75 | -132.37 |
| threonine | -63.40 | -135.33 |
| tryptophan | -63.49 | -135.26 |
| tyrosine | 248.95 | -107.39 |
| UDP-N-acetylglucosamine | 40.85 | -19.35 |
| valine | 178.54 | -65.38 |

After obtaining the estimated parameters for , and , we selected random parameters *a* and *b* and calculated the phase error with , for each spectrum index and for each of the 11 single metabolite spectra separately. We then summed the phase errors for each index across the 11 single metabolite spectra and transformed the phase errors from degrees to radius scale values to obtain the final phase error at index *k*. We used the following equation to develop a combined absorption spectrum containing phase errors:

Here, represents the intensity at index *k* with noise and phase errors, is the simulated intensity at index *k* with noise from the previous step, and is the corresponding dispersion intensity at index *k*, which is Hilbert transformed as described in Supplement 1 from the simulated absorption intensity at index *k* with noise from the previous step.

This procedure is equivalent to adding phase errors to each of the 11 single metabolite spectra and then combining them together to create a combined spectrum containing 11 metabolites and phase errors, as phase is angle and can be added. Finally, we used the Hilbert transform as described in Supplement 1 to obtain the corresponding dispersion spectrum for the combined absorption spectrum with the 11 metabolites containing noise and phase errors.

5.2.3 Baseline bias

To add the baseline bias to a simulated spectrum with added noise and phase errors, we used the following model:

*Baseline bias* ~ | *a1* +*a2x* + *a3x*2 + *a4* *x*3 |

Here, *x* represents the index of a given point in the spectrum, divided by the total number of points in the spectrum. The coefficients *a1*, *a2*, *a3*, and *a4* were randomly sampled from a normal distribution with a mean equal to the intensity mean of the spectrum and a standard deviation equal to the intensity standard deviation of the same spectrum.

### 5.3 Simulation datasets

We created three sets of simulations based on our idealized spectrum:

1) Nset: 1,000 simulations containing noise only;

2) NPset: 1,000 simulations containing noise and phase errors;

3) NPBset: 1,000 simulations containing noise, phase errors, and baseline bias.

The details of the idealized spectrum and the methods used to introduce noise, phase errors, and baseline bias can be found in Sections 5.1 (Simulating our idealized spectrum) and 5.2 (Adding multiple types of errors).

We then evaluated these three simulation sets, along with eight different phase error correction methods, including raw data without any phase error correction. This resulted in a total of 24 sets of phased spectra, as shown in Table S3.

In summary, our simulation datasets allow us to evaluate phase error correction methods with simulations containing different combinations of error types.

Table S3. 24 sets of phased spectra (1000 spectra per cell)

|  |  |  |  |
| --- | --- | --- | --- |
| Methods | Simulation Sets | | |
| Nset | Npset | NPBset |
| NPC | Nset\_NPC | NPset\_NPC | NPBset\_NPC |
| SPC\_AAM | Nset\_SPC\_AAM | NPset\_SPC\_AAM | NPBset\_SPC\_AAM |
| SPC\_EMP | Nset\_SPC\_EMP | NPset\_SPC\_EMP | NPBset\_SPC\_EMP |
| SPC\_DSM | Nset\_SPC\_DSM | NPset\_SPC\_DSM | NPBset\_SPC\_DSM |
| SPC\_DANM | Nset\_SPC\_DANM | NPset\_SPC\_DANM | NPBset\_SPC\_DANM |
| MPC\_DANM | Nset\_MPC\_DANM | NPset\_MPC\_DANM | NPBset\_MPC\_DANM |
| MPC\_EMP | Nset\_MPC\_EMP | NPset\_MPC\_EMP | NPBset\_MPC\_EMP |
| NLS | Nset\_NLS | NPset\_NLS | NPBset\_NLS |

## 

## **Supplement 6. Evaluation of phase error correction methods with visualizations**

### In this section, we describe our methods to compare different phase error correction methods using an idealized spectrum and simulation datasets. We compared the methods based on peak height, peak range area, and point-to-point errors.

### 6.1 Peak height error comparisons

To compare peak height estimation errors, we used an idealized spectrum that had no noise, phase errors, or baseline bias. We measured peak height and accounted for possible peak location shifts due to remaining phase errors or any other reasons within a range of 11 indices, from peak index -5 to peak index +5. For each phase error correction method and simulation dataset, we generated separate boxplots of peak height errors for each of the 76 peaks.

To further evaluate the overall methods' performance visually, we combined peak height errors across different peaks and simulation data sets. We then used histograms to display the error distributions of peak heights, as well as empirical density plots with rectangular kernel and identity position, which are smoothed versions of the histograms that can easily overlap.

### 6.2 Peak range area error comparisons

Based on our idealized spectrum, we manually defined 17 peak ranges out of the 76 peaks. To compare the errors in peak range area estimation, we utilized the same idealized spectrum as for peak height error comparison, which was devoid of any noise, phase errors, or baseline bias. The peak range area was calculated by summing all absorption intensity values within the given peak range. For each phase error correction method and simulation dataset, we generated separate boxplots for peak range area errors for each of the 17 peak ranges.

To further evaluate the overall methods' performance visually, we combined peak range area errors across different peak ranges and simulation data sets. We then used histograms to display the error distributions of peak range area, as well as empirical density plots with rectangular kernel and identity position, which are smoothed versions of the histograms that can easily overlap.

### 6.3 Point-to-point comparison with the idealized spectrum

We compared a test spectrum after phase error correction to the idealized spectrum point by point, calculating the intensity error and the mean and 95% intensity error intervals for each point across spectra within each simulated dataset and a phase error correction method. We generated a line plot for the point-to-point errors.

Overall, these comparison methods provide a comprehensive evaluation of the performance of different phase error correction methods on simulated spectra, allowing for informed decisions on method selection for real-world applications.

## **Supplement 7****. Performance Comparison Using L1 Error, L2 Error, and F Test**

In our study, we evaluated the performance of different methods using various error metrics. One approach involved testing a mixed model to determine if there are significant differences among the methods in terms of estimation error. In addition, we directly compared overall estimation errors using L1-norm and L2-norm, also known as L1 and L2 loss functions. While these loss functions are commonly used for optimization, we used them here to compare different phase error correction methods, referring to them as L1 error and L2 error.

To calculate these error metrics, we used mean absolute error (MAE) and mean squared error (MSE) formats, which can be expressed as the sum of absolute error and sum of squared error, respectively:

Since our simulations all had the same sample size (m), we can ignore it in our calculations and focus on the essential parts of the formula. Therefore, we used the sum of absolute error and the sum of squared error as L1 and L2 loss values, respectively, and referred to them as L1 error and L2 error, which can be written as:

L1 error ~

L2 error ~

Here, we used the tilde symbol (~) to indicate the essential parts that are used in our comparisons.

We used the F test to determine if there are significant differences in L2 error among the different methods. Specifically, we compared the L2 error of any other method (*i*) to that of a reference method (*j*). The F statistic tests the equality of two variances, and we can simplify it due to the equal sample sizes of the methods being compared:

We then used the F distribution to calculate the right tail probability (Pr (>F)) by integrating the F distribution curve.

For peak height and peak range area, the degrees of freedom for both L2 errors were calculated based on the number of simulations and peaks or peak ranges in our data sets.

Overall, our study used a variety of error metrics to comprehensively compare the performance of different methods.

**Supplement 8. A case study**

To evaluate the effectiveness of phase error correction methods for NMR spectra, we conducted a case study using 50 spectra extracted from the ASICSdata R package. These spectra were already phased using ACD/1D NMR Manager 8.0.

Figure S1 displays the sub-ranges of spectra for the two glucose ranges: 4.63 - 4.67 and 5.22 - 5.26, with 25 diabetes and 25 control groups. Each of the eight phase error correction methods, including spectra directly from ASICSdata labelled as "original," are also displayed.

Using the combined peak range areas, we plotted boxplots to compare the performance of the different phase error correction methods.

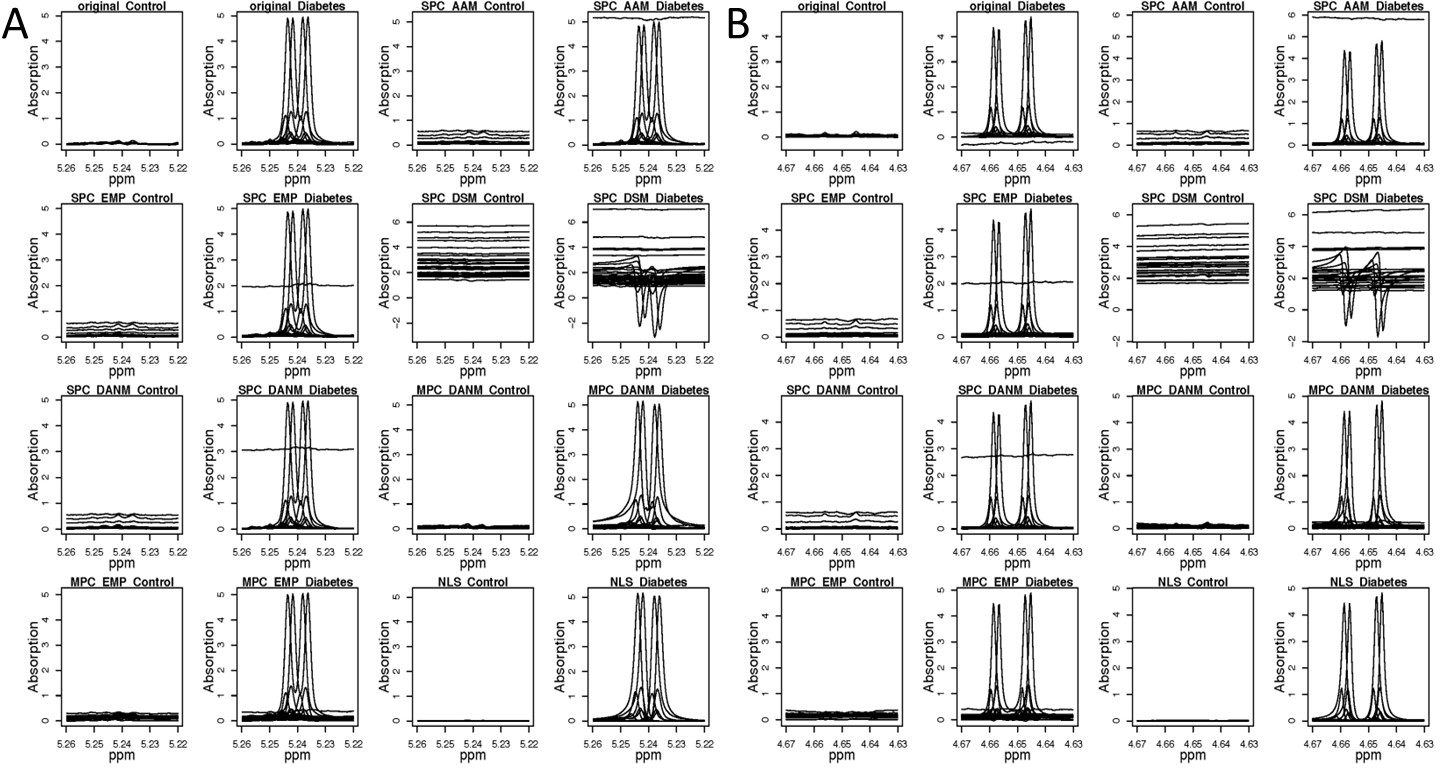


Figure S1. Glucose peak ranges for eight different phase error correction methods; 50 control samples and cases of diabetes are shown separately. (A) is for ppm range: 5.22 - 5.26, (B) is for ppm range: 4.63 - 4.67.

**Supplement 9 Design and analysis method in metabolite spike-in experiments**

In this study, we used three urine samples with identification numbers 14883, 17548, and 22324 from a commercial company and spiked them with eight metabolites at designed concentrations to evaluate the accuracy and sensitivity of our approach. The spike-in metabolites included arginine, caffeine, citric Acid, creatinine, glycine, histidine, leucine, and phenylalanine.

To evaluate the effectiveness of our approach, we performed two types of experiments: single spike-in experiments and multiple spike-in experiments. In the single spike-in experiments, we added one metabolite to each of the urine samples at varying concentrations. In contrast, in the multiple spike-in experiments, we added multiple metabolites to each of the urine samples with varying concentration combinations.

9.1. Single spike-in metabolite experiment

We conducted a single spike-in metabolite experiment in which one pre-determined spike-in metabolite was added to each urine sample for every spectrum, along with a non-spike-in spectrum as a reference.

To model the relationship between metabolite concentration and the area under the curve, we used twelve different spike-in concentrations of creatinine ranging from 0 to 15 mM (millimolar). The concentrations used were as follows: 0 mM, 0.002 mM, 0.01 mM, 0.1 mM, 0.5 mM, 1 mM, 2 mM, 4 mM, 5 mM, 8 mM, 10 mM, and 15 mM.

For the other seven metabolites, we spiked in six different concentrations ranging from 0.25 to 3 mM, with 0 mM as the reference level. The concentrations used for these metabolites were: 0 mM, 0.25 mM, 0.5 mM, 1 mM, 2 mM, and 3 mM.

9.2. Multiple spike-in metabolites experiment

To conduct the spike-in metabolites experiment, we used the three non-zero concentration levels of each metabolite, representing low, medium, and high concentrations. The actual concentrations used for creatinine were 0.5, 5, and 10 mM, and for the other seven metabolites, concentrations of 0.25, 1, and 3 mM were used.

Next, we randomly selected a concentration level among low, medium, and high for each of the eight metabolites, resulting in six distinct concentration combinations. We manually adjusted the concentrations to balance them as required. The details of the concentration combinations are presented in Table S4.

Table S4 Metabolite Spike-in concentration combinations (mM)

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Spike-in samples | 1 | 2 | 3 | 4 | 5 | 6 |
| creatinine | 0.5 | 10 | 0.5 | 5 | 10 | 5 |
| citric acid | 0.25 | 1 | 1 | 3 | 0.25 | 3 |
| glycine | 1 | 3 | 0.25 | 3 | 1 | 0.25 |
| phenylalanine | 0.25 | 1 | 1 | 0.25 | 3 | 3 |
| histidine | 1 | 3 | 0.25 | 3 | 1 | 0.25 |
| leucine | 3 | 1 | 3 | 0.25 | 0.25 | 1 |
| arginine | 1 | 0.25 | 3 | 1 | 0.25 | 3 |
| caffeine | 3 | 0.25 | 1 | 0.25 | 3 | 1 |

For each spectrum, we added multiple spike-in metabolites based on the designed concentration combinations in Table S4. We also included a non-spike-in spectrum as a reference.

Additionally, we measured three endogenous metabolites that commonly exist in urine samples for all spectra, namely alanine, formic Acid, and sucrose. These endogenous metabolites were not used in our spike-in data analysis but were included for quality control purposes only to ensure their detectability in our experiments.

9.3. Sample preparation for 1H NMR analysis

All urine samples remained stored at −80 °C until the day of analysis, when they were thawed at room temperature. Samples were prepared as a mixture of urine containing 10% D2O and 0.5 mM sodium trimethylsilyl propionate-[2,2,3,3-2H4] (TSP) final concentration. Samples were transferred into 3-mm outer diameter NMR tubes (Bruker, Rheinstetten, Germany) and then stored at 5°C inside the SampleJetTM automatic sample changer until measurement.

9.4. 1H NMR spectroscopy data acquisition and preprocessing

All data were acquired on a Bruker AVANCE III HD 600 NMR spectrometer operating at a frequency of 600.17 MH and equipped with a 5-mm triple resonance cryoprobe (CPQCI 1H-31P/13C/15N), and a SampleJetTM for refrigerated sample storage and preheating prior to automated data acquisition using Topspin 3.5pl7 and IconNMR v5.0.8 software (Bruker, Rheinstetten, Germany).

Experiments were carried out at 298K. The 1H spectra were acquired using the pulse sequence *noesygppr1d* (Bruker Biospin Ltd) to achieve good suppression of the water signal. 1H spectrum was acquired with 64 scans, a 1H 90° pulse length of 8.2 μs, a spectral width of 12 kHz, and a recycle delay of 4 s for a total of 66 K data points.

Raw time domain NMR data were Fourier transformed into the frequency domain.

9.5. Defining metabolite ranges and areas

We searched for ppm ranges of all metabolites in each spectrum to ensure consistency. However, the ppm range size for the same metabolite but in different experiments (single spike-in or combination of multiple spike-in) was not necessarily the same. This was because the presence of multiple spike-in metabolites made the process more complicated, and we might end up with narrower ranges to avoid overlap with other metabolites.

In the single spike-in experiment, we estimated creatinine's area using both of its peak ranges (around 3 ppm with 3 protons and around 4 ppm with 2 protons) to ensure higher accuracy for estimating the transformation coefficient from area to concentration. In the multiple spike-in experiment, however, we only used the peak around 3 ppm with 3 protons for consistency and fairness in comparing the results. For all other metabolites, we only used one peak to estimate their area. It's important to note that the range used for metabolite concentration estimation may contain a single peak or a multiplet, depending on the metabolites and the selected ranges.

After defining a range, we estimated the area of the range by calculating the absolute value of its area under a curve, which was the summation of all absolute intensities within the range. We used the absolute value to ensure that the concentration estimates were positive. It's worth noting that we only used net area for simulations and the glucose case study, as we did not need to obtain metabolite concentrations. Comparing the net area was fair for phase error correction methods, as negative net area indicates a large amount of phase errors.

9.6. Estimating coefficients from area to concentration

In this section, we describe our approach to estimating coefficients that convert metabolite areas to concentrations. The goal of this analysis is to obtain accurate and reliable estimates of metabolite concentrations from NMR data.

To estimate these coefficients, we used Creatinine as a reference metabolite in a single spike-in experiment. Specifically, we calculated the absolute area per proton of Creatinine (area/H) by dividing the absolute area by the number of protons in the given range. We then took the mean of the two absolute areas per proton from two measured ranges and used this as input to a linear regression model:

Here, *Y* represents the designed concentration of Creatinine in mM, *X* is the mean of the two ranges' absolute areas per H of Creatinine for a spike-in spectrum, *X0* is the mean of the two ranges' areas per H of Creatinine for a non-spike-in reference spectrum, β is the coefficient we need to estimate, and ε is the random error assumed to follow a normal distribution. Note that there is no intercept in this model.

We estimated a coefficient β for each phase error correction method separately in order to achieve the best fit. After estimating the coefficients, we used them to estimate the concentration of each metabolite for each method by plugging in the absolute peak area per H. Furthermore, we estimated concentration changes from a test spectrum compared to its corresponding reference spectrum for each method, which were then used as final concentration estimates.

Overall, this analysis allowed us to estimate coefficients that convert metabolite areas to concentrations, which is a critical step in metabolomics data analysis.

9.7. Phase correction method comparison for single metabolite spike-in experiment

Once we had estimated concentrations, we compared them to our spike-in concentrations to get estimation errors and evaluated performance of phase error correction methods across different metabolites and two experiments.

For the single metabolite spike-in experiment, we calculated error for each phase error correction method across different metabolites and different urine samples, investigated their error distribution patterns using empirical density plots with rectangular kernel and identity position, plotted correlation heatmap of seven methods, compared their L1 error (least absolute deviations), L2 error (least squares), and applied F test on L2 ratios.

Notice that a spectrum has only one spike-in metabolite for single metabolite spike-in experiment, so spectrum information will be ignored. We built up a fixed error model to detect if there is any significantly difference among phase error correction methods after adjusting for effects of metabolites, urine samples, and any two-way interactions, which can be represented in the following:

(S9.F1)

: concentration estimation error, which is estimated spike-in concentration - designed spike-in concentration

: a factor variable with seven different phase error correction methods, the best method with the smallest L2 error as the reference method

a factor variable with eight spike-in metabolites, Creatinine is used as the reference level since this is the metabolite for estimating converting coefficient from absolute area/H to concentration. To be consistent with other data, both original Creatinine peak ranges’ data are used directly (instead of mean of the two peak ranges) in the single spike-in experiment for modelling.

a factor variable with three urine samples, urine sample with ID 14883 is used as the reference level since this is the smallest ID although it does not matter which sample is used as reference any way.

, , and are the main effects of these three variables, and , , are the interaction terms between *X1* and *X2*, *X1* and *X3*, and *X2* and *X3* respectively.

: random error following normal distribution

9.8. Phase correction method comparison for multiple metabolite spike-in experiment

For the multiple metabolite spike-in experiment, we also calculated the error for each phase error correction method across different metabolites and different urine samples, investigated their error distribution patterns using empirical density plots with rectangular kernel and identity position, plotted a correlation heatmap of seven methods, compared their L1 error (least absolute deviations) and L2 error (least squares), and applied an F test on L2 ratios.

Considering the spatial correlation among metabolite signals within the same spectrum and within the same urine sample, we built up a mixed model to detect if there is any significant difference among phase error correction methods after adjusting for the effects of metabolites, urine samples, and any two-way interactions, which can be represented in the following equation:

(S9.F2)

Here, refers to a spectrum ID, and represents the random intercept effect of spectra (*X*4), *E* is random error, and all other terms are the same as in Equation (S9.F1).

Since the mixed model function lmer in the R package lme4 does not provide a p-value directly, we applied the Satterthwaite degrees of freedom approximation to achieve p-values of each effect using the R package lmerTest (Kuznetsova et al., 2017).

9.9. Phase correction method comparison for combination of two experiments

Combining both spike-in experiments, we again calculated the error for each phase error correction method across different metabolites and urine samples, investigated their error distribution patterns using empirical plots with rectangular kernel and identity position s, plotted a correlation heatmap of seven methods, compared their L1 error (least absolute deviations), L2 error (least squares), and applied an F test on L2 ratios.

Modelling the combination of two experiments is not straightforward, since the single spike-in experiment had only one spike-in metabolite per spectrum, while the multiple spike-in experiment had multiple spike-in metabolites measured in each spectrum. A mixed model would not work well for the single spike-in spectra, as there is no spatial correlation within a spectrum. Therefore, we used a fixed effect model to combine both experiments by ignoring the spectrum information. The model can be written as follows:

*ε*  (S9.F3)

Here, refers to the main effect of experiment, and , , and represent interaction effects between experiment and phase error correction method, between experiment and metabolite, and between experiment and urine sample, respectively. All other variables are the same as in Equations (S9.F1) and (S9.F2).

## 

## **Supplement 10. Performance comparison of phase error correction methods for peak height quantification in simulated spectra**

Peak height is used for both metabolite identification and quantification when a peak range area is not easily assessed. We compared the performance of all eight methods, including the one with no phase error correction (NPC), against our idealized spectrum across the three simulation sets. Boxplots of peak height error distributions for 76 peaks are shown in Figures S2-4 for simulation sets Nset, NPset, and NPBset, respectively.

### 10.1 Peak height error boxplots in Nset simulations

Figure S2 shows that the NLS (non-linear shrinkage) method performs the best among the seven phase error correction methods for the Nset simulation, which contains random noises only. Along with raw data with no phase error correction (NPC) and single model with absolute area minimization (SPC\_AAM), NLS is also one of the methods with the smallest boxplot range. Notably, NLS is the only method where the peak height error boxplots cross the horizontal 0 line. In contrast to other methods, the height errors of NLS within the boxes, which represent the IQR (interquartile range) from the 25th to 75th percentile, are the most symmetric to the horizontal 0 line. These results indicate that NLS is a highly effective method for correcting phase errors in simulation Nset.

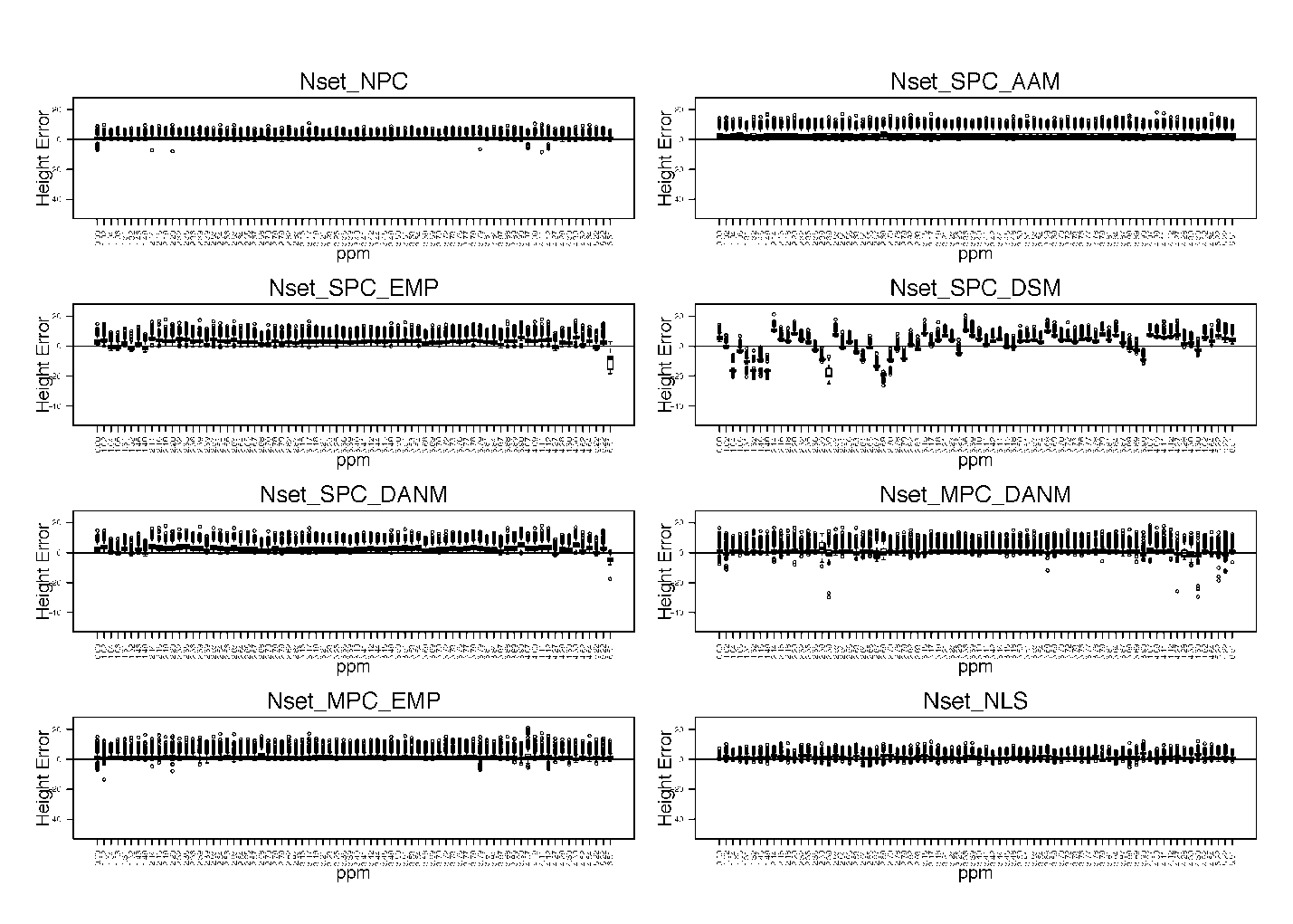


Figure S2. Peak height error distribution boxplots of 76 peaks for different methods in Nset with 1,000 simulations. Y-axis: peak height error = peak height in a simulation – peak height in the idealized spectrum.

### 

### 10.2 Peak height error boxplots in NPset simulations

When phase errors are present in addition to noises but baseline biases (NPset) are absent, Figure S3 demonstrates that the non-linear shrinkage (NLS) method performs the best overall. Specifically, NLS has the most desirable box locations, is the most symmetric to the horizontal 0 line, and has the overall smallest range of error values.

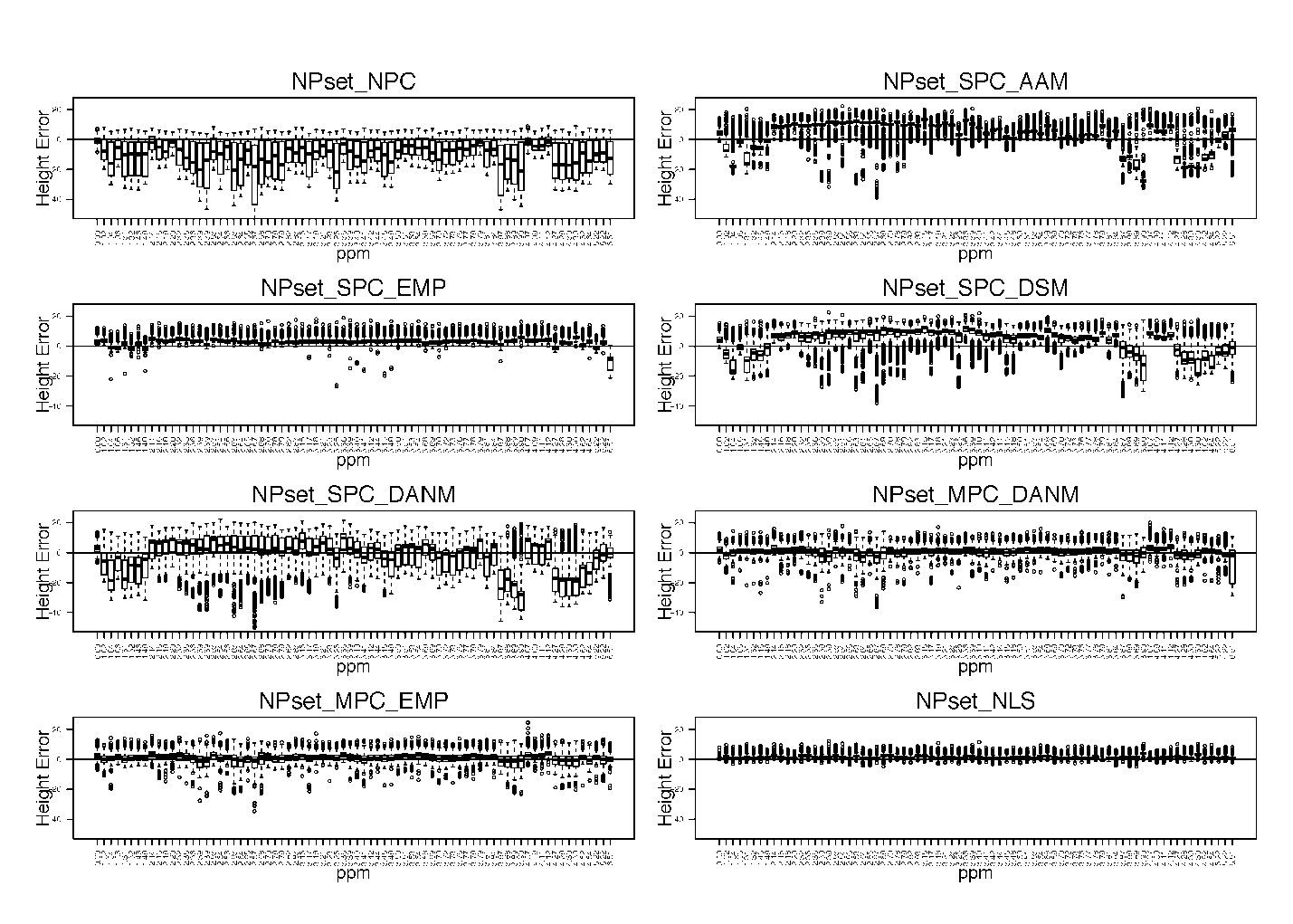


Figure S3. Peak height error distribution boxplots of 76 peaks for different methods in NPset with 1,000 simulations. Y-axis: peak height error = peak height in a simulation – peak height in the idealized spectrum.

### 10.3 Peak height error boxplots in NPBset simulations

Figure S4 shows that, even when phase errors are present along with noise and baseline biases (NPBset), the shrinkage method NLS continues to perform the best overall. Specifically, NLS exhibits the most desirable box locations, the highest degree of symmetry with the horizontal 0 line, and the overall smallest range of error values.

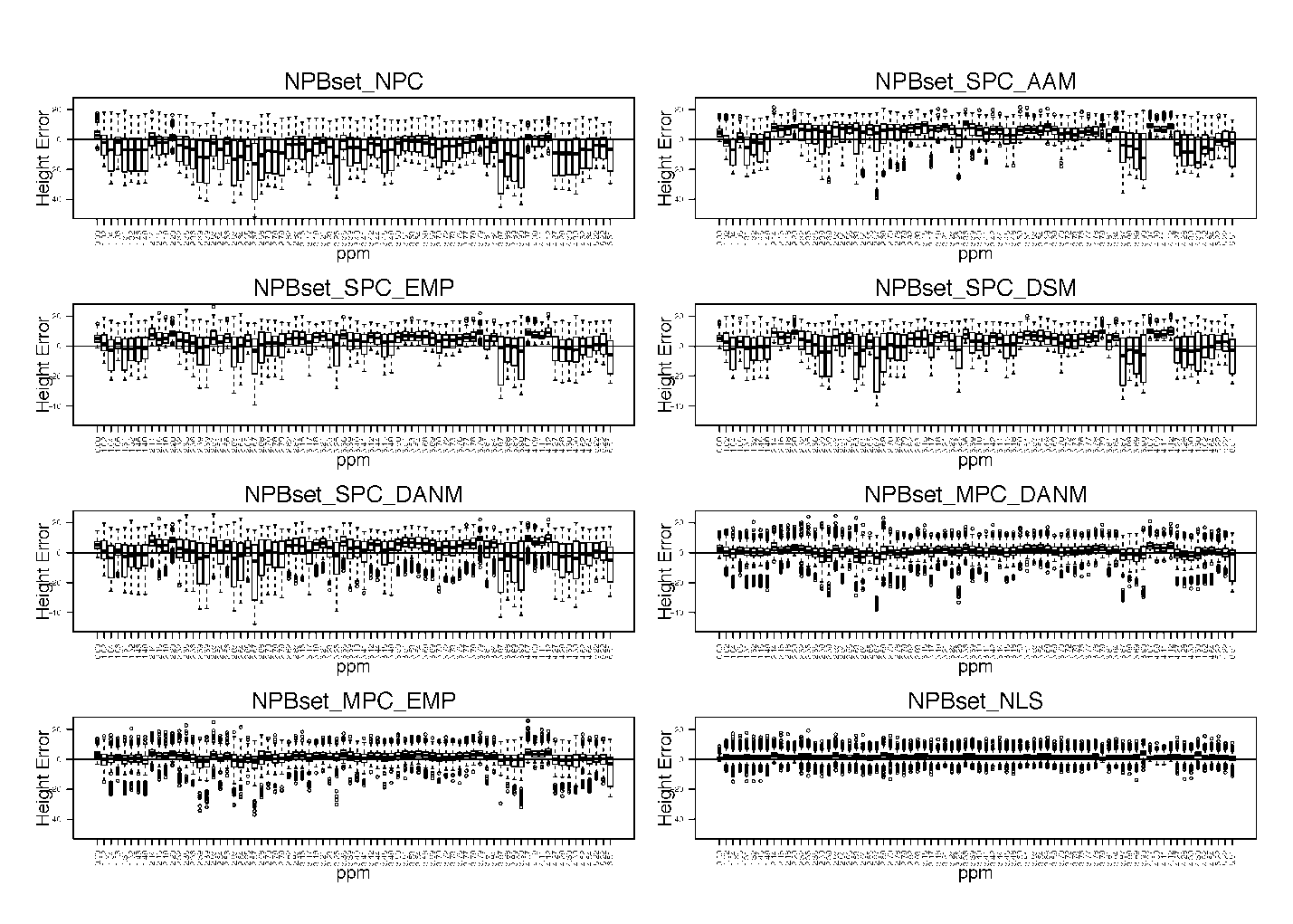


Figure S4. Peak height error distribution boxplots of 76 peaks for different methods in NPBset with 1,000 simulations. Y-axis: peak height error = peak height in a simulation – peak height in the idealized spectrum.

## **Supplement 11. Performance comparison of phase error correction methods for peak range area quantification in simulated spectra**

The area under the peak range is a critical feature for accurate metabolite identification and quantification. To assess the performance of various approaches, we compared eight different methods, including raw data without phase error correction (NPC), against an idealized spectrum using three sets of simulations. To facilitate comparison, we manually grouped 76 peaks into 17 ranges based on the idealized spectrum, ensuring each range contained at least one peak. Using boxplots, we compared the area error distribution across these different methods for simulation sets Nset, NPset, and NPBset, as illustrated in Figures S5, S6, and S7, respectively.

### 

### 11.1 Peak range area error boxplots in Nset simulations

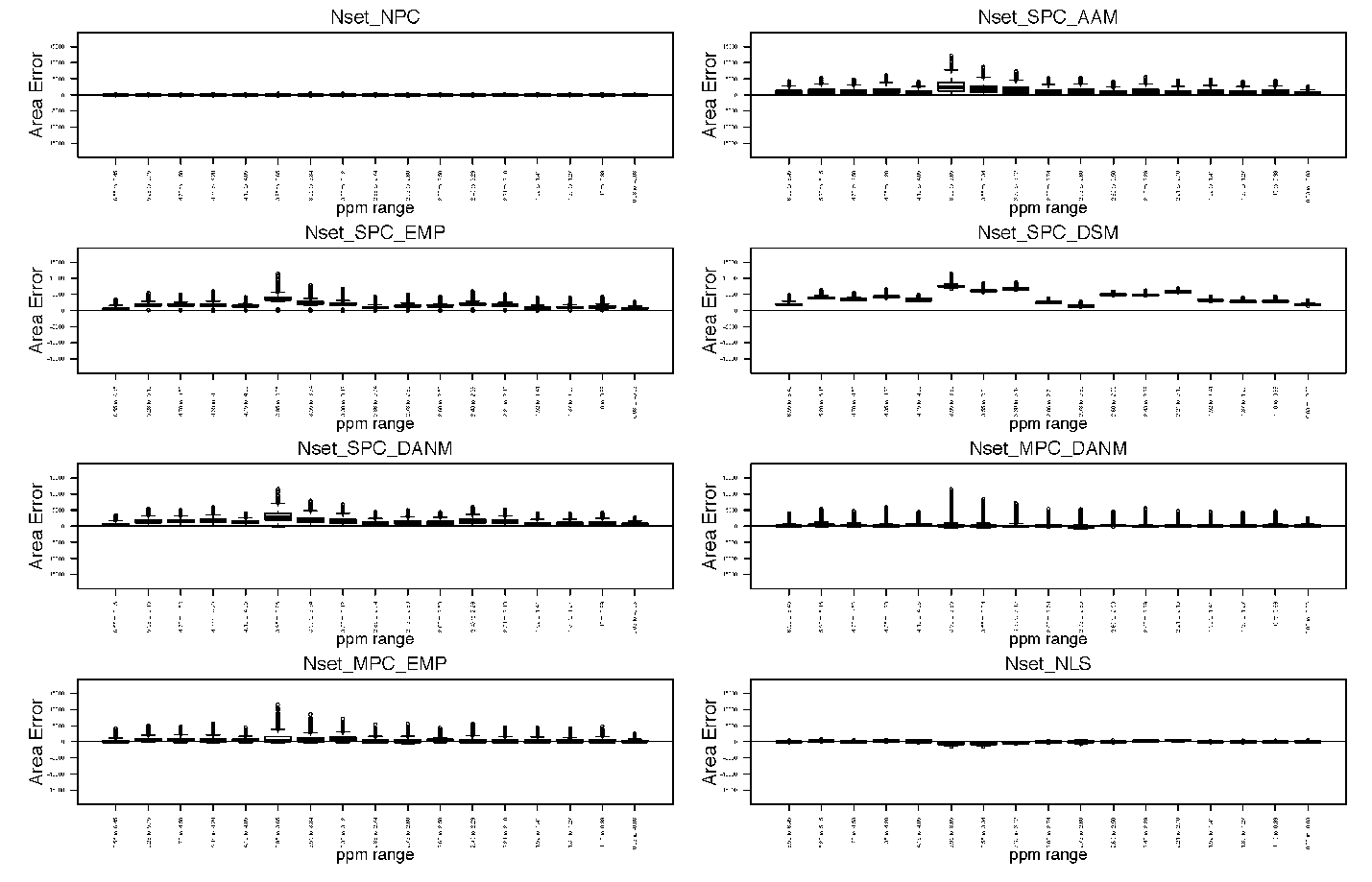


Figure S5. Peak range area error boxplots of 17 peak ranges for different methods in simulation Nset with 1,000 noise only simulations. Y-axis: peak range area in a simulation – peak range area in the idealized spectrum.

Figure S5 clearly demonstrates the superior performance of the NLS (shrinkage) method among all seven phase error correction approaches in simulation Nset, which contains only noise and no phase error or baseline bias. In this scenario, raw data NPC exhibits the best performance, since it only contains normally distributed noise with the standard idealized spectrum. However, NLS is the only method comparable to raw data NPC in terms of peak range area error centering and boxplot ranges (Figure S5).

### 11.2 Peak range area error boxplots in NPset simulations

In simulation NPset, where only phase errors and noise are present but no baseline bias, NLS (shrinkage) outperforms all other phase error correction methods, including raw data NPC, as shown in Figure S6. Although raw data NPC performs well in the absence of phase error, it is not suitable for correcting phase errors and exhibits inferior performance in the presence of phase errors. In contrast, NLS not only outperforms all other phase error correction methods but also performs better than raw data NPC in NPset.

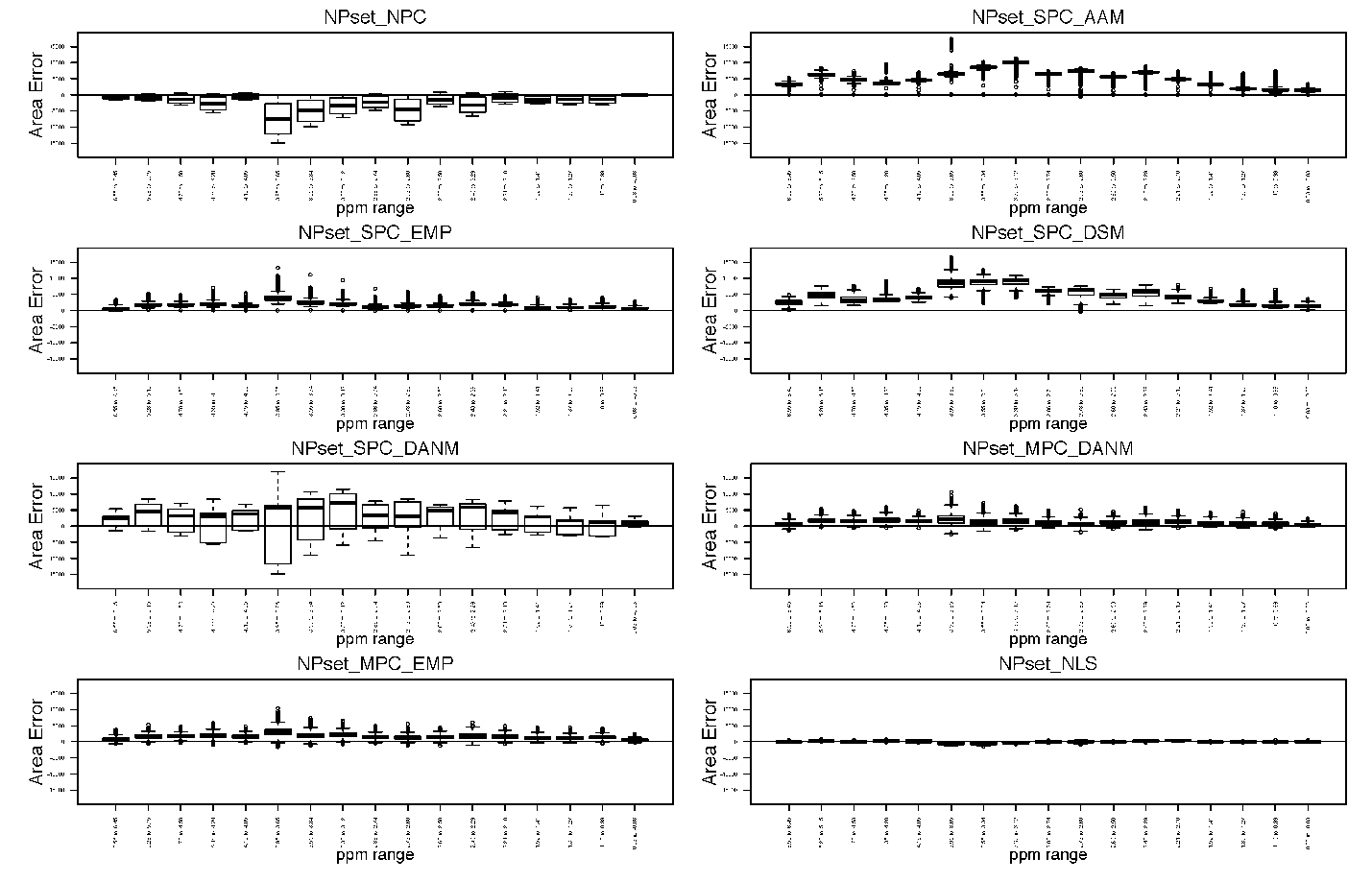


Figure S6. Peak range area error boxplots of 17 peak ranges for different methods in simulation NPset with 1,000 noise only simulations. Y-axis: peak range area in a simulation – peak range area in the idealized spectrum.

### 11.3 Peak range area error boxplots in NPBset simulations

When phase errors, noise, and baseline bias are present in simulation NPBset, Figure S7 shows that the shrinkage method NLS still performs the best. It is better than all other phase error correction methods and raw data NPC.

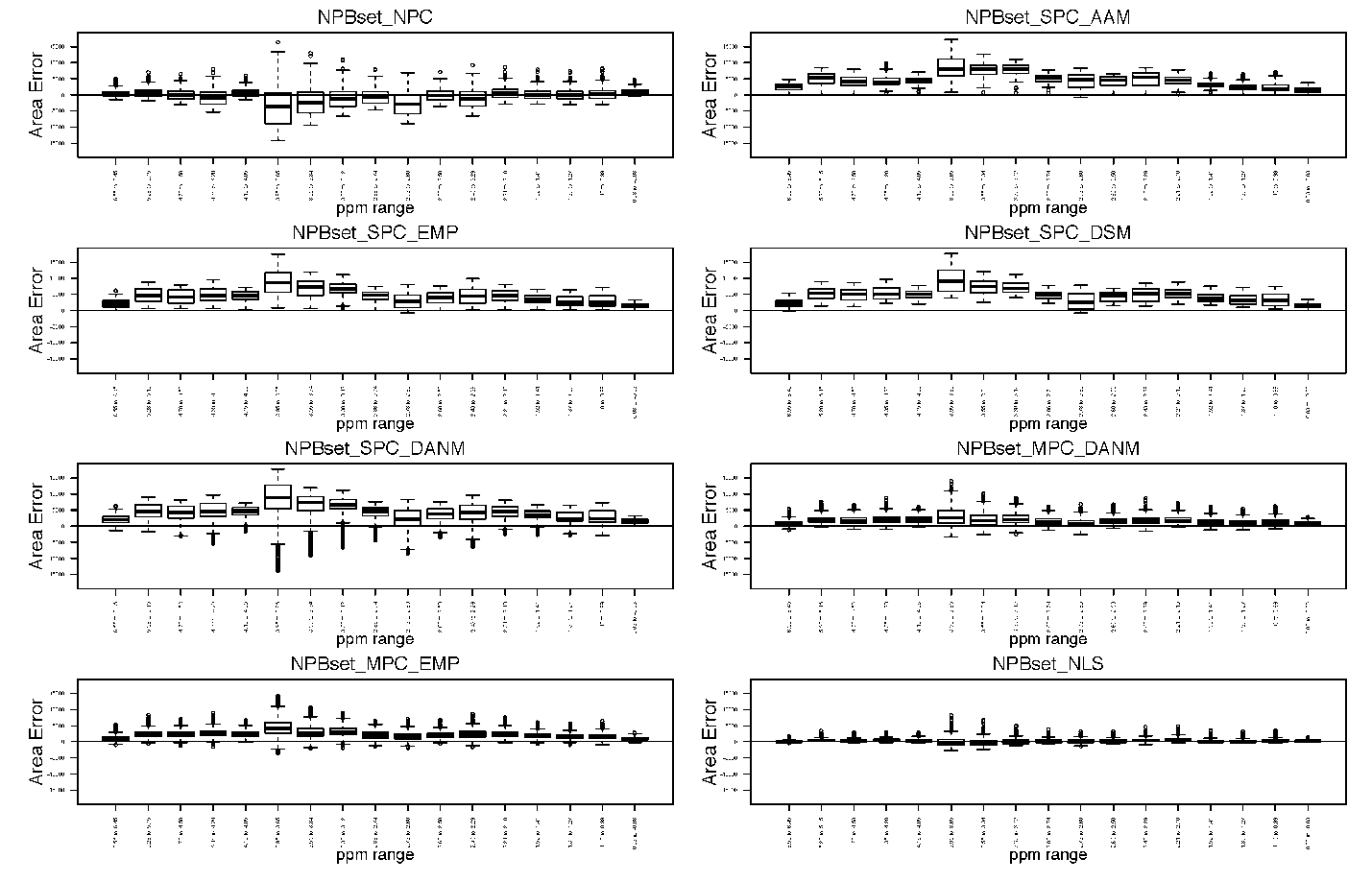


Figure S7. Peak range area error boxplots of 17 peak ranges for different methods in simulation NPBset with 1,000 noise only simulations. Y-axis: peak range area in a simulation – peak range area in the idealized spectrum.

## 

## **Supplement 12. Comparing phase error correction methods for accurate point-to-point intensity quantification**

While peak height and peak range area are commonly used for metabolite identification and quantification, a more comprehensive view of method performance can be obtained by comparing each data point. Thus, we compared the overall performance of eight approaches, including raw data NPC, against an idealized spectrum across three simulation sets. For each phase error correction method, we plotted the mean, 2.5%, and 97.5% of the intensity error away from the idealized spectrum across 1,000 repetitions in a simulated dataset for the entire spectrum range. Figures S8, S9, and S10 illustrate the performance of each method for simulation Nset, NPset, and NPBset, respectively.

### 

### 12.1 Point-to-point intensity error 95 percentile interval plots in Nset simulations

When only random noise is present and no phase error or baseline bias exists, raw data NPC is the best performer, as shown in Figure S8. For the seven phase error correction methods, single model with absolute area minimization (SPC\_AAM) has a narrow 95th percentile intensity error interval, but it is clearly biased since its 95th percentile error plot is towards the positive side. On the other hand, the shrinkage method NLS has balanced and narrow 95th percentile intensity error intervals, making it a reliable choice among the phase error correction methods.

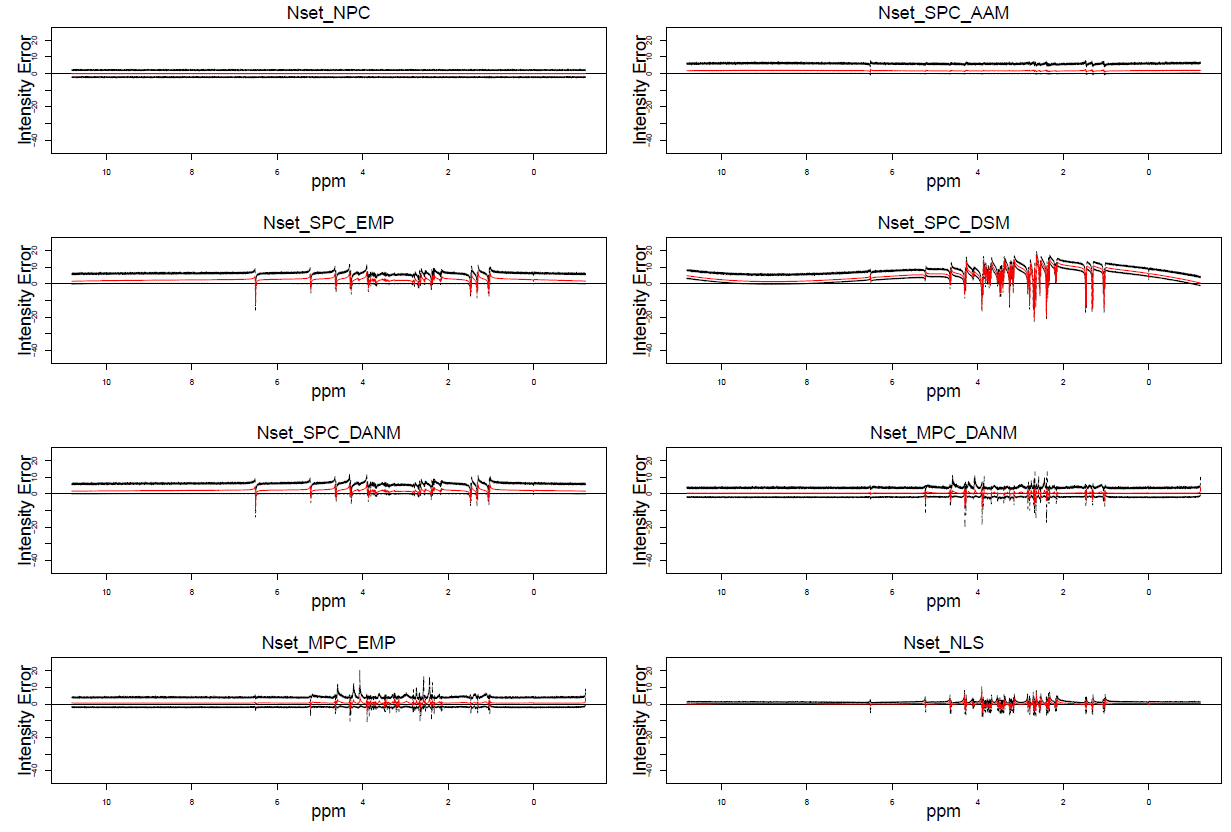


Figure S8. Point-to-point intensity error 95 percentile plots for simulation Nset: 1,000 noise only simulations. Y-axis: intensity in a simulation – intensity in the idealized spectrum. Red solid line: mean intensity error away from the idealized spectrum; Black dashed lines: 2.5% and 97.5% of intensity error away from the idealized spectrum.

### 12.2 Point-to-point intensity error 95 percentile interval plots in NPset simulations

In reality, real NMR data always contains phase errors in addition to random noise. When phase error and noise are present but baseline bias is absent in simulation NPset, Figure S9 clearly shows that the shrinkage method NLS performs the best. Notice that error patterns of NLS and single model minimizing entropy with penalty (SPC\_EMP) in Figure S9 are very close to their patterns in Figure S8, while all other methods, including raw data NPC, perform much worse in Figure S9 than in Figure S8. This is because NLS derives an absorption spectrum based on the magnitude and power spectra, which are independent of phase and phase error. SPC\_EMP is a good method to deal with phase errors when baseline bias is absent, but it is not as effective as NLS.

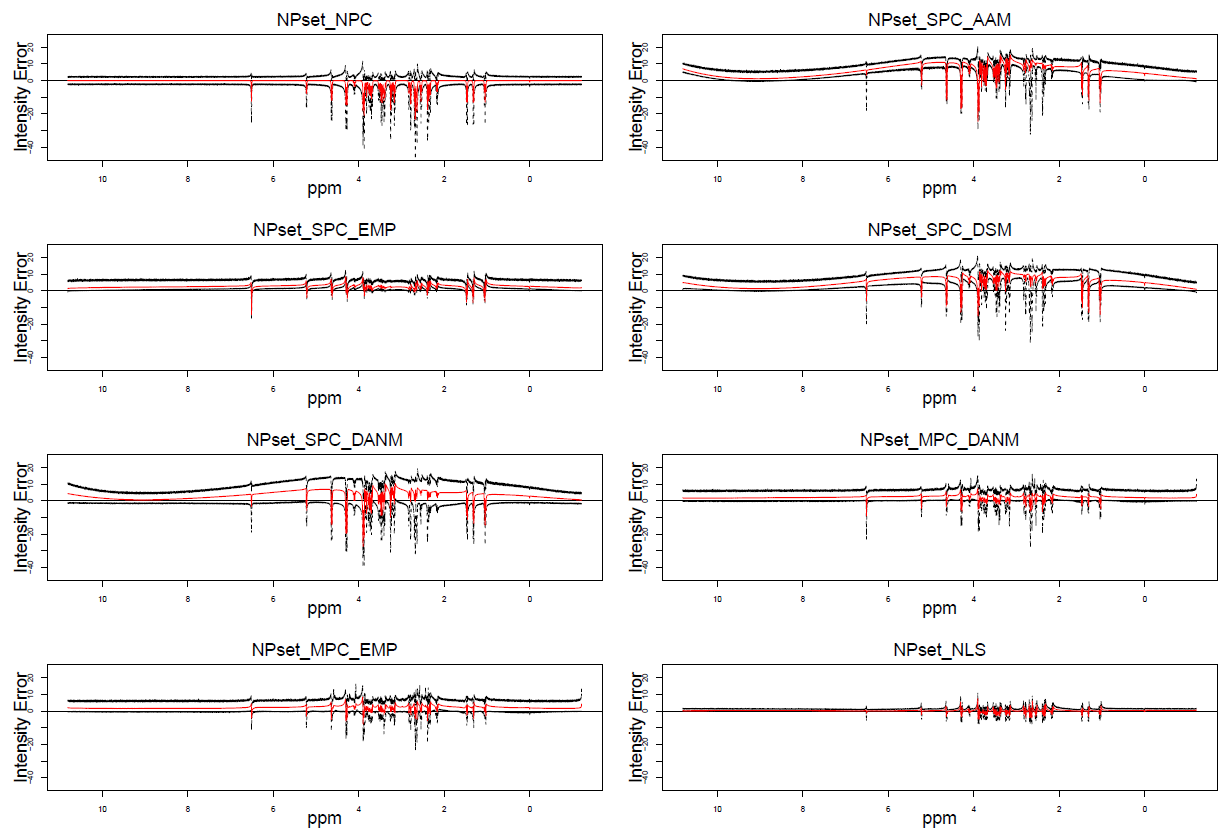


Figure S9. Point-to-point intensity error 95 percentile plots for simulation NPset: 1,000 noise plus phase error simulations. Y-axis: intensity in a simulation – intensity in the idealized spectrum. Red solid line: mean intensity error away from the idealized spectrum; Black dashed lines: 2.5% and 97.5% of intensity error away from the idealized spectrum.

### 12.3 Point-to-point intensity error 95 percentile interval plots in NPBset simulations

When phase error, noise, and baseline bias are present in simulation NPBset, Figure S10 shows that the shrinkage method NLS performs the best. Although the error pattern for NLS in Figure S10 is slightly worse than its patterns in Figure S9 and Figure S8, it is still the top performer. On the other hand, single model minimizing entropy with penalty (SPC\_EMP) performs much worse in Figure S10 than in Figure S9 and Figure S8, indicating that it is sensitive to baseline bias. In contrast, NLS is not only phase and phase error-free but also robust to baseline bias. The two multiple model approaches with either delta absolute net minimization: MPC\_DANM, or entropy minimization with penalty: MPC\_EMP, are the next best options after NLS for dealing with phase error and baseline bias data.

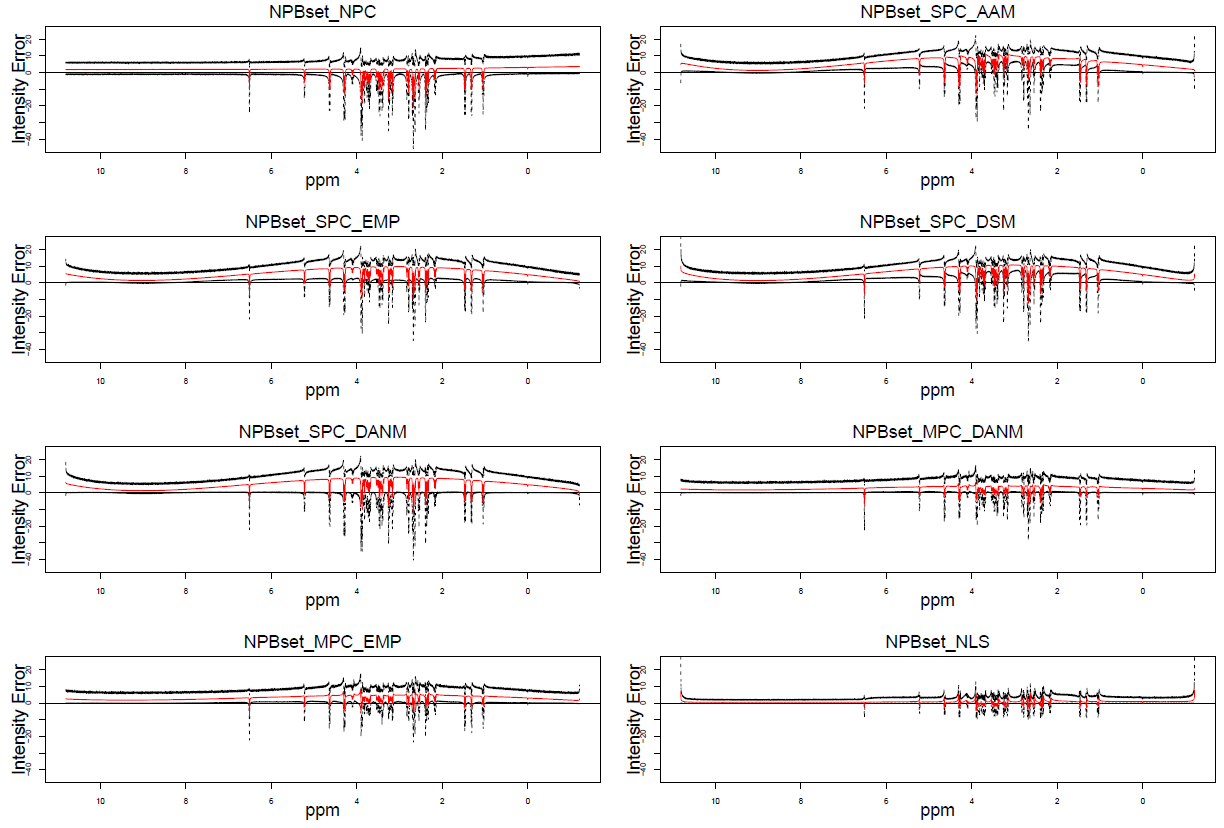


Figure S10. Point-to-point intensity error 95 percentile plots for simulation NPBset: 1,000 noise plus phase error and baseline bias simulations. Y-axis: intensity in a simulation – intensity in the idealized spectrum. Red solid line: mean intensity error away from the idealized spectrum; Black dashed lines: 2.5% and 97.5% of intensity error away from the idealized spectrum.

In conclusion, NLS is the most effective method for correcting phase errors in NMR data. Even in the most challenging scenario where both phase error and baseline bias exist, NLS still outperforms other methods, including raw data. The two MPC approaches are also useful alternatives for correcting phase error and baseline bias data.

## 

## **Supplement 13. Performance comparison of phase error correction methods for error distribution in simulated spectra**

## In this section, we investigate the distribution of peak height and peak range area errors for each method across all 76 peaks and 17 peak ranges in 3 different data simulation datasets.

## 13.1. Peak height error histograms

We compared each method on their peak height error distribution and presented the results in histograms (Figure S11), which show the distribution of peak height error for different methods across all 76 peaks and all 3,000 simulations from 3 simulation datasets. The narrowest distribution of peak height error was observed for non-linear shrinkage (NLS), indicating that it outperformed all other methods.

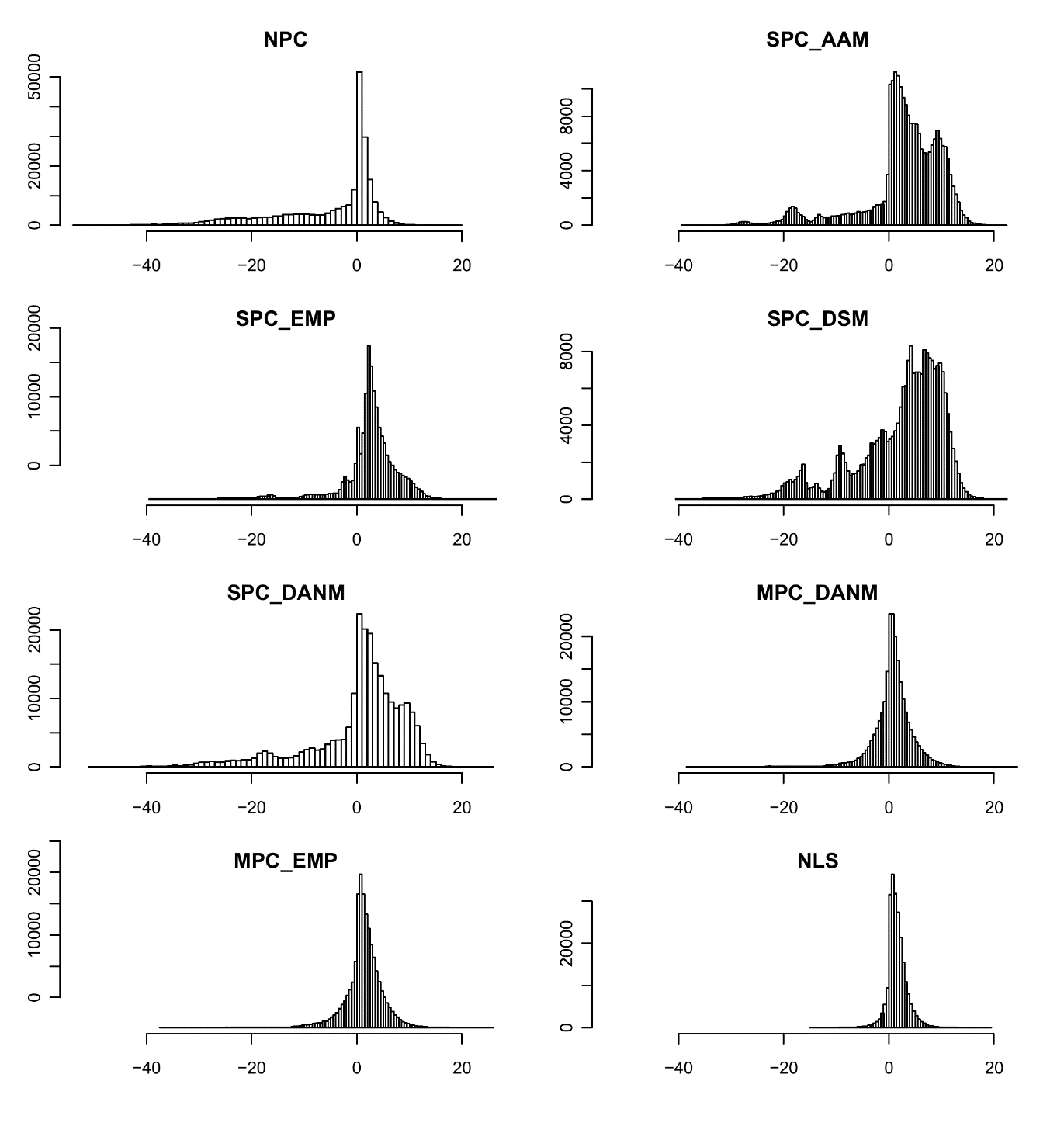


Figure S11. Distribution of peak height error for different methods. All distributions are across all 76 peaks and all 3,000 simulations in 3 different simulation datasets: Nset, NPset, and NPBset.

## 13.2. Peak range area error histograms

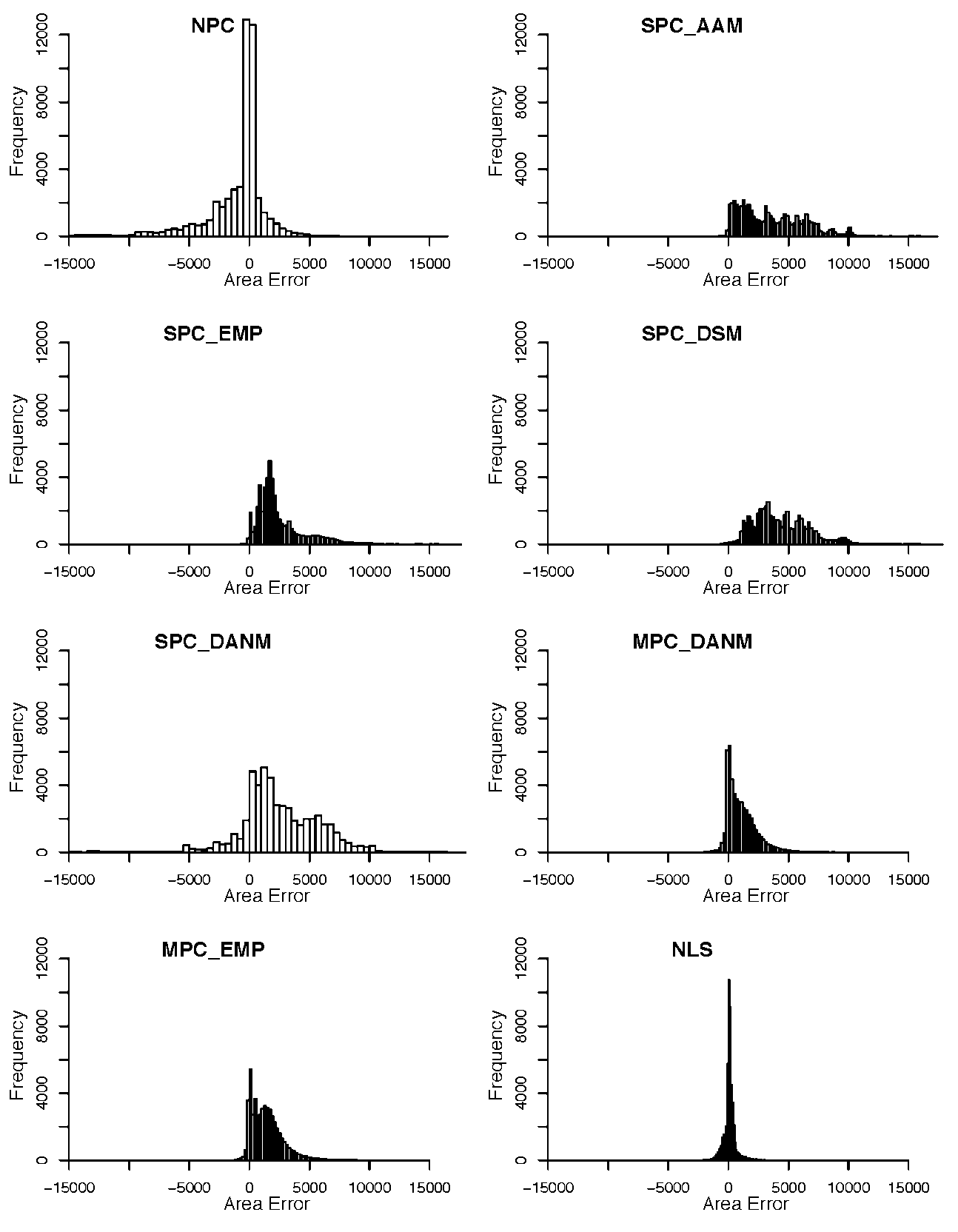
**

Figure S12. Distribution of peak range area error for different methods. All distributions are across all 17 peak ranges and all 3,000 simulations in 3 different simulation datasets: Nset, NPset, and NPBset.

Similarly, we examined the distribution of peak range area error for each method across all 17 peak ranges and 3 different data simulation datasets in comparison to NPC (Figure S12). The results show that NLS has the narrowest distribution of peak range area error, indicating that it performs the best among all methods.

## **Supplement 14. L1, L2, and F test for simulations**

To assess the performance of different methods for phase error correction, we used L1, L2, and F tests. The L1 and L2 tests measure the absolute and squared errors between the estimated peak heights or areas and the true values in the simulated spectra. The F test is a statistical test that compares the variances of the error distributions of different methods.

14.1 Peak height error results

Table S5 Summary statistics of peak height error for different methods. The results are across 3 different simulation datasets, 1000 simulated spectra per set, and 76 different peaks per simulation. Peak heights in the idealized spectrum are used as truth, the degree of freedom for each method is 3 ×1000 ×76 – 1 = 227999.

Table S5. The summary statistics of peak height errors for different methods across three simulation datasets.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Method | Sum of absolute height error (L1 error) | Sum of squares of height error (L2 error) | F = | Pr (> F) |
| NPC | 1590806.73 | 29401973.68 | 22.3140 | <0.0001 |
| SPC\_AAM | 1445672.93 | 15183557.96 | 11.5232 | <0.0001 |
| SPC\_EMP | 1043891.93 | 8603857.03 | 6.5297 | <0.0001 |
| SPC\_DSM | 1655957.91 | 17515128.56 | 13.2927 | <0.0001 |
| SPC\_DANM | 1491982.43 | 19045434.78 | 14.4541 | <0.0001 |
| MPC\_DANM | 608576.1 | 3578253.34 | 2.7156 | <0.0001 |
| MPC\_EMP | 624007.89 | 3338358.88 | 2.5336 | <0.0001 |
| NLS | 416320.66 | 1317646.78 |  |  |

14. 2 Peak Range Area Error Results

Table S6 Summary statistics of peak range area error for different methods. The results are across 3 different simulation datasets, 1000 simulated spectra per set, and 17 different peak ranges per simulation. Peak range areas in the idealized spectrum are used as truth, the degree of freedom for each method is: 3 ×1000 ×17 – 1 = 50999.

Table S6. The summary statistics of peak range area errors for different methods across three simulation datasets.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Method | Sum of absolute area error  (L1 error) | Sum of squares  of area error  (L2 error) | F = | Pr (> F) |
| NPC | 74529306.4 | 3.67E+11 | 27.8030 | < 0.0001 |
| SPC\_AAM | 184298673.4 | 1.05E+12 | 79.5455 | < 0.0001 |
| SPC\_EMP | 131685227.4 | 5.80E+11 | 43.9394 | < 0.0001 |
| SPC\_DSM | 227327316.2 | 1.30E+12 | 98.4849 | < 0.0001 |
| SPC\_DANM | 164541074.4 | 9.18E+11 | 69.5455 | < 0.0001 |
| MPC\_DANM | 59210816.7 | 1.48E+11 | 11.2121 | < 0.0001 |
| MPC\_EMP | 75397629.9 | 2.06E+11 | 15.6061 | < 0.0001 |
| NLS | 15808192.2 | 1.32E+10 |  |  |

## **Supplement 15. Mixed model results for simulations**

To further assess the performance of different methods for phase error correction, we used mixed models to analyse the peak height error and peak range area error. Mixed models allow for the incorporation of both fixed and random effects, making them a powerful tool for analysing complex data structures like those found in our simulations.

15. 1 Peak height error mixed model results

In our analysis of peak height errors, we modelled method and peak as fixed effects in a peak height error mixed-effects model (see Table S7). We also included simulation data set and the spectrum within each simulation data set as random effects (see Table S8). Additionally, we performed pairwise comparisons among phase error correction methods (see Table S9).

Table S7. Fixed effects in the peak height error mixed model

|  | Estimate | Std Error | t value |
| --- | --- | --- | --- |
| (Intercept) | 3.63099 | 0.43471 | 8.353 |
| MPC\_DANM\_vs\_NLS | -0.84356 | 0.07312 | -11.537 |
| MPC\_EMP\_vs\_NLS | -0.32163 | 0.07312 | -4.399 |
| SPC\_AAM\_vs\_NLS | 1.16227 | 0.07312 | 15.895 |
| SPC\_DANM\_vs\_NLS | -1.12028 | 0.07312 | -15.321 |
| SPC\_DSM\_vs\_NLS | 0.43521 | 0.07312 | 5.952 |
| SPC\_EMP\_vs\_NLS | 0.90097 | 0.07312 | 12.322 |
| NPC\_vs\_NLS | -6.83568 | 0.07312 | -93.485 |
| peakppm1.0236 | -2.84988 | 0.05288 | -53.896 |
| peakppm1.0427 | -6.68058 | 0.05288 | -126.34 |
| peakppm1.0619 | -2.57595 | 0.05288 | -48.715 |
| peakppm1.3065 | -5.69573 | 0.05288 | -107.715 |
| peakppm1.3239 | -5.36724 | 0.05288 | -101.503 |
| peakppm1.4572 | -4.72597 | 0.05288 | -89.376 |
| peakppm1.4753 | -4.91795 | 0.05288 | -93.006 |
| peakppm2.1394 | 1.60783 | 0.05288 | 30.407 |
| peakppm2.1588 | -0.23113 | 0.05288 | -4.371 |
| peakppm2.1777 | -0.12073 | 0.05288 | -2.283 |
| peakppm2.197 | 1.38248 | 0.05288 | 26.145 |
| peakppm2.3231 | -0.33155 | 0.05288 | -6.27 |
| peakppm2.3486 | -0.89106 | 0.05288 | -16.851 |
| peakppm2.3614 | -1.99333 | 0.05288 | -37.697 |
| peakppm2.388 | -3.97008 | 0.05288 | -75.081 |
| peakppm2.3909 | -4.62332 | 0.05288 | -87.434 |
| peakppm2.5266 | 0.65779 | 0.05288 | 12.44 |
| peakppm2.5433 | -1.67116 | 0.05288 | -31.604 |
| peakppm2.5598 | -0.06552 | 0.05288 | -1.239 |
| peakppm2.6348 | -3.57283 | 0.05288 | -67.568 |
| peakppm2.6419 | -3.80185 | 0.05288 | -71.899 |
| peakppm2.6527 | -1.07163 | 0.05288 | -20.266 |
| peakppm2.6744 | -5.82717 | 0.05288 | -110.201 |
| peakppm2.6808 | -3.42774 | 0.05288 | -64.824 |
| peakppm2.6964 | -2.20614 | 0.05288 | -41.722 |
| peakppm2.7767 | -2.17834 | 0.05288 | -41.196 |
| peakppm2.7855 | -2.82626 | 0.05288 | -53.449 |
| peakppm2.8203 | -0.54179 | 0.05288 | -10.246 |
| peakppm2.8291 | -0.61809 | 0.05288 | -11.689 |
| peakppm3.1487 | 0.44538 | 0.05288 | 8.423 |
| peakppm3.1654 | -1.94494 | 0.05288 | -36.782 |
| peakppm3.1822 | -0.12482 | 0.05288 | -2.361 |
| peakppm3.2095 | 0.58336 | 0.05288 | 11.032 |
| peakppm3.2314 | -0.86513 | 0.05288 | -16.361 |
| peakppm3.2535 | -3.95568 | 0.05288 | -74.808 |
| peakppm3.3628 | 1.5416 | 0.05288 | 29.154 |
| peakppm3.3863 | -0.67546 | 0.05288 | -12.774 |
| peakppm3.3973 | -1.13432 | 0.05288 | -21.452 |
| peakppm3.4088 | -2.43292 | 0.05288 | -46.01 |
| peakppm3.4204 | -0.50812 | 0.05288 | -9.609 |
| peakppm3.4376 | -0.63047 | 0.05288 | -11.923 |
| peakppm3.4541 | -2.90621 | 0.05288 | -54.961 |
| peakppm3.4774 | -2.63955 | 0.05288 | -49.918 |
| peakppm3.5016 | -0.55135 | 0.05288 | -10.427 |
| peakppm3.5126 | 0.15508 | 0.05288 | 2.933 |
| peakppm3.5282 | -0.10176 | 0.05288 | -1.924 |
| peakppm3.5372 | 0.01979 | 0.05288 | 0.374 |
| peakppm3.6783 | 0.25907 | 0.05288 | 4.899 |
| peakppm3.6855 | -0.71344 | 0.05288 | -13.492 |
| peakppm3.7007 | -2.33417 | 0.05288 | -44.143 |
| peakppm3.7168 | -1.65706 | 0.05288 | -31.338 |
| peakppm3.7302 | -1.82107 | 0.05288 | -34.439 |
| peakppm3.7574 | -1.17978 | 0.05288 | -22.311 |
| peakppm3.7709 | -0.57544 | 0.05288 | -10.882 |
| peakppm3.7766 | -0.20459 | 0.05288 | -3.869 |
| peakppm3.7946 | 1.23308 | 0.05288 | 23.32 |
| peakppm3.8149 | -2.01477 | 0.05288 | -38.103 |
| peakppm3.8415 | -0.16706 | 0.05288 | -3.159 |
| peakppm3.8699 | -6.87922 | 0.05288 | -130.097 |
| peakppm3.8787 | -5.26697 | 0.05288 | -99.607 |
| peakppm3.8916 | -6.27583 | 0.05288 | -118.686 |
| peakppm3.9004 | -7.43382 | 0.05288 | -140.585 |
| peakppm4.0723 | 1.65887 | 0.05288 | 31.372 |
| peakppm4.0896 | 0.46885 | 0.05288 | 8.867 |
| peakppm4.107 | 0.4971 | 0.05288 | 9.401 |
| peakppm4.1242 | 1.7119 | 0.05288 | 32.375 |
| peakppm4.2705 | -5.15483 | 0.05288 | -97.486 |
| peakppm4.2771 | -6.02052 | 0.05288 | -113.858 |
| peakppm4.2958 | -5.9865 | 0.05288 | -113.214 |
| peakppm4.3022 | -5.8119 | 0.05288 | -109.912 |
| peakppm4.6227 | -4.27944 | 0.05288 | -80.931 |
| peakppm4.6425 | -4.3539 | 0.05288 | -82.339 |
| peakppm5.2153 | -2.71371 | 0.05288 | -51.321 |
| peakppm5.2237 | -2.41287 | 0.05288 | -45.631 |
| peakppm6.5084 | -6.03863 | 0.05288 | -114.2 |

Table S8. Random effects of peak height error mixed model\*

|  |  |  |  |
| --- | --- | --- | --- |
| Groups | Name | Variance | Std.Dev. |
| Spectrum within simulation set | (Intercept) | 7.5785 | 2.7529 |
| simulation set | (Intercept) | 0.5548 | 0.7448 |
| Residual |  | 33.5525 | 5.7925 |

\*1,824,000 observations, 24,000 spectra within simulation sets, and 3 simulation sets

Table S9. Pairwise comparisons of peak height error among methods

based on the full mixed model\*

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| contrast | estimate | SE | df | z.ratio | p.value |
| NLS - MPC\_DANM | 0.84356 | 0.07312 | Inf | 11.537 | < 0.0001 |
| NLS - MPC\_EMP | 0.32163 | 0.07312 | Inf | 4.399 | 0.0003 |
| NLS - SPC\_AAM | -1.1623 | 0.07312 | Inf | -15.895 | < 0.0001 |
| NLS - SPC\_DANM | 1.12028 | 0.07312 | Inf | 15.321 | < 0.0001 |
| NLS - SPC\_DSM | -0.4352 | 0.07312 | Inf | -5.952 | < 0.0001 |
| NLS - SPC\_EMP | -0.901 | 0.07312 | Inf | -12.322 | < 0.0001 |
| NLS – NPC | 6.83568 | 0.07312 | Inf | 93.485 | < 0.0001 |
| MPC\_DANM - MPC\_EMP | -0.5219 | 0.07312 | Inf | -7.138 | < 0.0001 |
| MPC\_DANM - SPC\_AAM | -2.0058 | 0.07312 | Inf | -27.432 | < 0.0001 |
| MPC\_DANM - SPC\_DANM | 0.27672 | 0.07312 | Inf | 3.784 | 0.0038 |
| MPC\_DANM - SPC\_DSM | -1.2788 | 0.07312 | Inf | -17.488 | < 0.0001 |
| MPC\_DANM - SPC\_EMP | -1.7445 | 0.07312 | Inf | -23.858 | < 0.0001 |
| MPC\_DANM – NPC | 5.99212 | 0.07312 | Inf | 81.948 | < 0.0001 |
| MPC\_EMP - SPC\_AAM | -1.4839 | 0.07312 | Inf | -20.294 | < 0.0001 |
| MPC\_EMP - SPC\_DANM | 0.79866 | 0.07312 | Inf | 10.922 | < 0.0001 |
| MPC\_EMP - SPC\_DSM | -0.7568 | 0.07312 | Inf | -10.351 | < 0.0001 |
| MPC\_EMP - SPC\_EMP | -1.2226 | 0.07312 | Inf | -16.72 | < 0.0001 |
| MPC\_EMP – NPC | 6.51406 | 0.07312 | Inf | 89.086 | < 0.0001 |
| SPC\_AAM - SPC\_DANM | 2.28255 | 0.07312 | Inf | 31.216 | < 0.0001 |
| SPC\_AAM - SPC\_DSM | 0.72706 | 0.07312 | Inf | 9.943 | < 0.0001 |
| SPC\_AAM - SPC\_EMP | 0.2613 | 0.07312 | Inf | 3.574 | 0.0084 |
| SPC\_AAM – NPC | 7.99795 | 0.07312 | Inf | 109.38 | < 0.0001 |
| SPC\_DANM - SPC\_DSM | -1.5555 | 0.07312 | Inf | -21.273 | < 0.0001 |
| SPC\_DANM - SPC\_EMP | -2.0213 | 0.07312 | Inf | -27.643 | < 0.0001 |
| SPC\_DANM – NPC | 5.7154 | 0.07312 | Inf | 78.164 | < 0.0001 |
| SPC\_DSM - SPC\_EMP | -0.4658 | 0.07312 | Inf | -6.37 | < 0.0001 |
| SPC\_DSM – NPC | 7.27089 | 0.07312 | Inf | 99.437 | < 0.0001 |
| SPC\_EMP – NPC | 7.73666 | 0.07312 | Inf | 105.806 | < 0.0001 |

\*The results are averaged over all peaks, p value adjustment with Tukey method for comparing a family of eight estimates for eight methods; df is Inf since the sample size is too big to be considered (3 simulation datasets, 1,000 simulated spectra per simulation set, and 76 peaks per spectrum).

## 

## 15.2. Peak range area error mixed model

Similar to our analysis of peak height errors, we also modelled method and peak range area as fixed effects in a peak range area error mixed-effects model (see Table S10). We also included simulation data set and the spectrum within each simulation data set as random effects (see Table S11). Additionally, we performed pairwise comparisons among phase error correction methods (see Table S12).

Table S10. Fixed effects of peak range area error mixed model

|  |  |  |  |
| --- | --- | --- | --- |
|  | Estimate | Std. Error | t value |
| (Intercept) | -957.51 | 374.87 | -2.554 |
| MPC\_DANM\_vs\_NLS | 989.53 | 33.36 | 29.66 |
| MPC\_EMP\_vs\_NLS | 1345.99 | 33.36 | 40.344 |
| SPC\_AAM\_vs\_NLS | 3509.23 | 33.36 | 105.184 |
| SPC\_DANM\_vs\_NLS | 2440.43 | 33.36 | 73.148 |
| SPC\_DSM\_vs\_NLS | 4353.27 | 33.36 | 130.482 |
| SPC\_EMP\_vs\_NLS | 2477.96 | 33.36 | 74.273 |
| NPC\_vs\_NLS | -1041.13 | 33.36 | -31.206 |
| ppmRange1.1\_0.99 | 370.06 | 14.74 | 25.099 |
| ppmRange1.3699\_1.27 | 334.03 | 14.74 | 22.655 |
| ppmRange1.5199\_1.4101 | 610.27 | 14.74 | 41.391 |
| ppmRange2.21\_2.1 | 1354.16 | 14.74 | 91.844 |
| ppmRange2.43\_2.2901 | 1301.64 | 14.74 | 88.282 |
| ppmRange2.5999\_2.5 | 1111.78 | 14.74 | 75.405 |
| ppmRange2.7299\_2.6001 | 524.27 | 14.74 | 35.558 |
| ppmRange2.8599\_2.74 | 1052.97 | 14.74 | 71.416 |
| ppmRange3.2999\_3.1201 | 2143.36 | 14.74 | 145.371 |
| ppmRange3.55\_3.3401 | 1852.22 | 14.74 | 125.625 |
| ppmRange3.95\_3.6501 | 2247.78 | 14.74 | 152.453 |
| ppmRange4.1499\_4.05 | 1186.6 | 14.74 | 80.48 |
| ppmRange4.3499\_4.2001 | 1130.72 | 14.74 | 76.69 |
| ppmRange4.6999\_4.58 | 1103 | 14.74 | 74.809 |
| ppmRange5.28\_5.15 | 1441.74 | 14.74 | 97.784 |
| ppmRange6.55\_6.4501 | 250.93 | 14.74 | 17.019 |

Table S11. Random effects of peak range area error mixed model \*

|  |  |  |  |
| --- | --- | --- | --- |
| Groups | Name | Variance | Std.Dev. |
| Spectrum within simulation set | (Intercept) | 1516178 | 1231.3 |
| simulation set | (Intercept) | 419604 | 647.8 |
| Residual |  | 2608653 | 1615.1 |

\* 408,000 observations, 24,000 spectra within simulation sets, and 3 simulation sets

Table S12. Pairwise comparisons of peak range area error among methods

based on the full mixed model\*

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| contrast | estimate | SE | df | z.ratio | p.value |
| MPC\_DANM - MPC\_EMP | -989.53 | 33.3629 | Inf | -29.66 | <.0001 |
| MPC\_DANM – NLS | -1346 | 33.3629 | Inf | -40.344 | <.0001 |
| MPC\_DANM – NPC | -3509.2 | 33.3629 | Inf | -105.18 | <.0001 |
| MPC\_DANM - SPC\_AAM | -2440.4 | 33.3629 | Inf | -73.148 | <.0001 |
| MPC\_DANM - SPC\_DANM | -4353.3 | 33.3629 | Inf | -130.48 | <.0001 |
| MPC\_DANM - SPC\_DSM | -2478 | 33.3629 | Inf | -74.273 | <.0001 |
| MPC\_DANM - SPC\_EMP | 1041.13 | 33.3629 | Inf | 31.206 | <.0001 |
| MPC\_EMP – NLS | -356.46 | 33.3629 | Inf | -10.684 | <.0001 |
| MPC\_EMP – NPC | -2519.7 | 33.3629 | Inf | -75.524 | <.0001 |
| MPC\_EMP - SPC\_AAM | -1450.9 | 33.3629 | Inf | -43.488 | <.0001 |
| MPC\_EMP - SPC\_DANM | -3363.7 | 33.3629 | Inf | -100.82 | <.0001 |
| MPC\_EMP - SPC\_DSM | -1488.4 | 33.3629 | Inf | -44.613 | <.0001 |
| MPC\_EMP - SPC\_EMP | 2030.66 | 33.3629 | Inf | 60.866 | <.0001 |
| NLS – NPC | -2163.2 | 33.3629 | Inf | -64.84 | <.0001 |
| NLS - SPC\_AAM | -1094.4 | 33.3629 | Inf | -32.804 | <.0001 |
| NLS - SPC\_DANM | -3007.3 | 33.3629 | Inf | -90.138 | <.0001 |
| NLS - SPC\_DSM | -1132 | 33.3629 | Inf | -33.929 | <.0001 |
| NLS - SPC\_EMP | 2387.12 | 33.3629 | Inf | 71.55 | <.0001 |
| NPC - SPC\_AAM | 1068.8 | 33.3629 | Inf | 32.036 | <.0001 |
| NPC - SPC\_DANM | -844.04 | 33.3629 | Inf | -25.299 | <.0001 |
| NPC - SPC\_DSM | 1031.27 | 33.3629 | Inf | 30.911 | <.0001 |
| NPC - SPC\_EMP | 4550.36 | 33.3629 | Inf | 136.39 | <.0001 |
| SPC\_AAM - SPC\_DANM | -1912.8 | 33.3629 | Inf | -57.334 | <.0001 |
| SPC\_AAM - SPC\_DSM | -37.533 | 33.3629 | Inf | -1.125 | 0.9515 |
| SPC\_AAM - SPC\_EMP | 3481.56 | 33.3629 | Inf | 104.354 | <.0001 |
| SPC\_DANM - SPC\_DSM | 1875.31 | 33.3629 | Inf | 56.209 | <.0001 |
| SPC\_DANM - SPC\_EMP | 5394.4 | 33.3629 | Inf | 161.688 | <.0001 |
| SPC\_DSM - SPC\_EMP | 3519.09 | 33.3629 | Inf | 105.479 | <.0001 |

\*The results are averaged over all peak ranges, P value adjustment with Tukey method for comparing a family of eight estimates for eight methods; df is Inf since the sample size is too big to be considered (3 simulation datasets, 1,000 simulated spectra per simulation set, and 17 peak ranges per spectrum).

**Supplement 16. Histogram-based analysis of concentration estimation errors in metabolite spike-in experiments**

To further compare the performance of different phase error correction methods, we generated concentration estimation error histograms based on metabolite spike-in experiments. Specifically, we analysed the errors in the peak height and peak range area measurements using the single metabolite spike-in experiment (Figure S13), multiple metabolite spike-in experiment (Figure S14), and a combination of both experiments (Figure S15).

Although it is hard to tell which method performs the best from histograms of concentration estimation errors for all phase correction methods, Figures S13-16 clearly show that SPC\_DSM is the worst method since its error histogram has the widest spread, which indicates that it produces much more error than any other method.

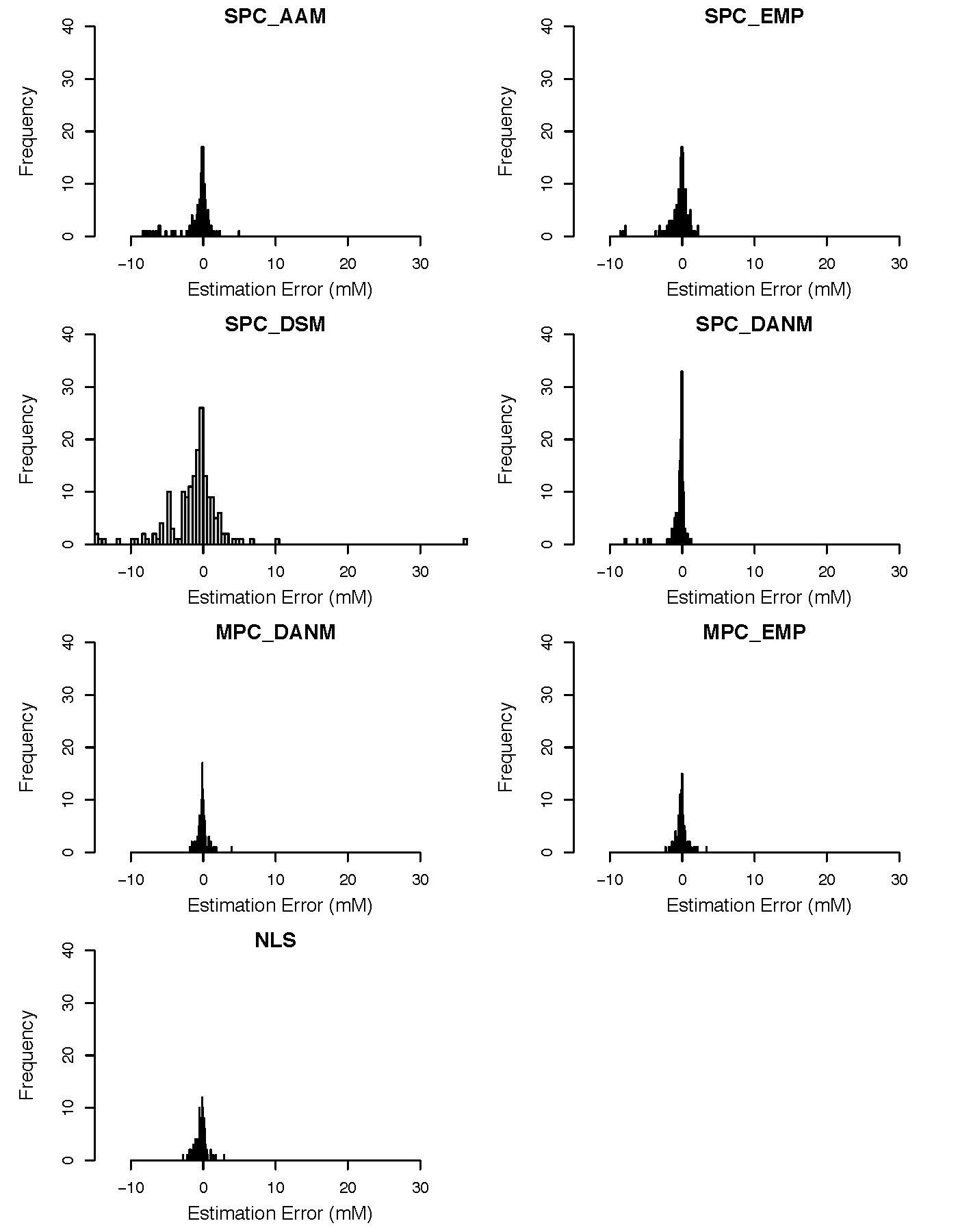


Figure S13 Histograms of estimated errors of seven phase error correction methods within the single metabolite spike-in experiment.

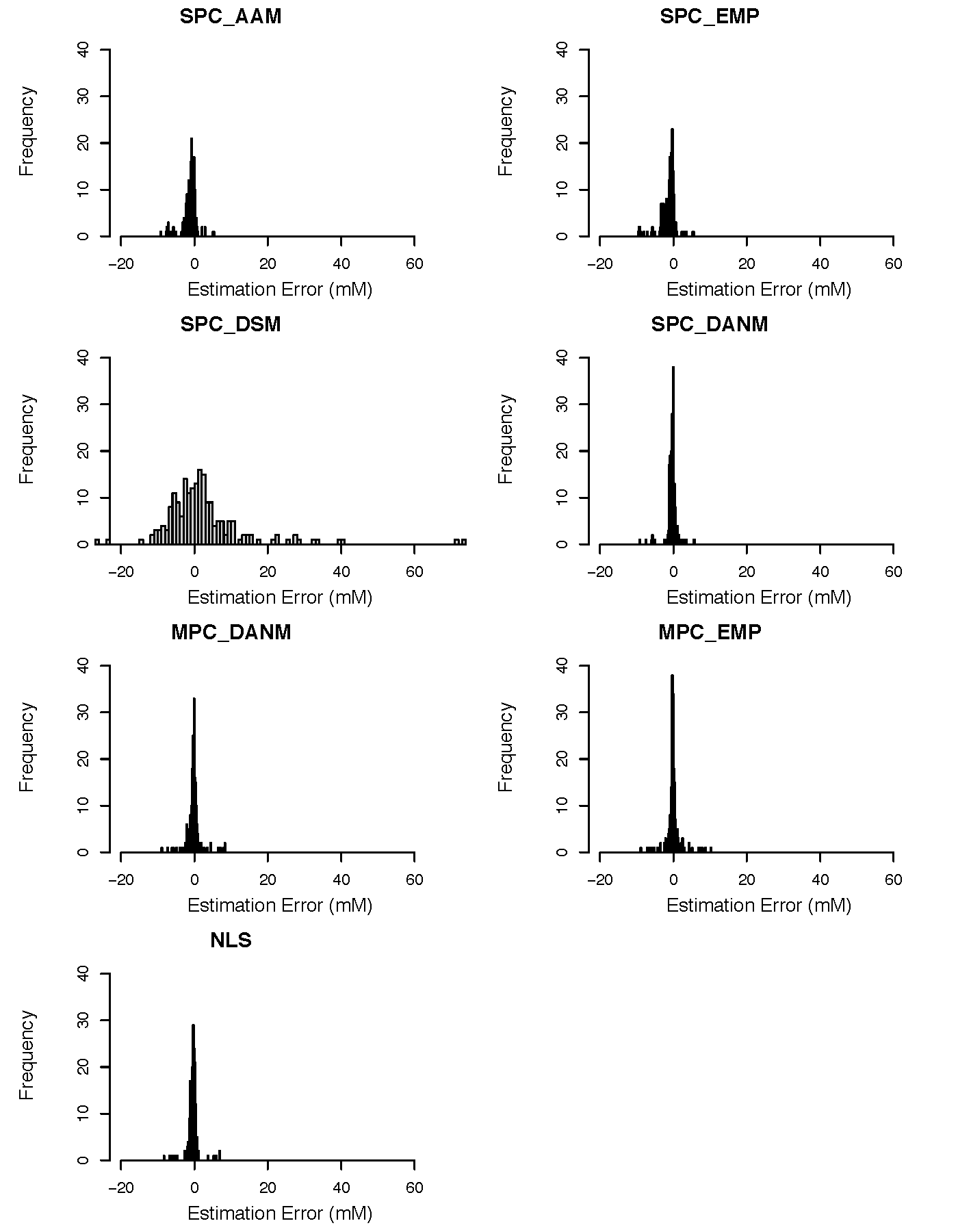


Figure S14 Histograms of estimated errors of seven phase error correction methods within the multiple metabolite spike-in experiment.

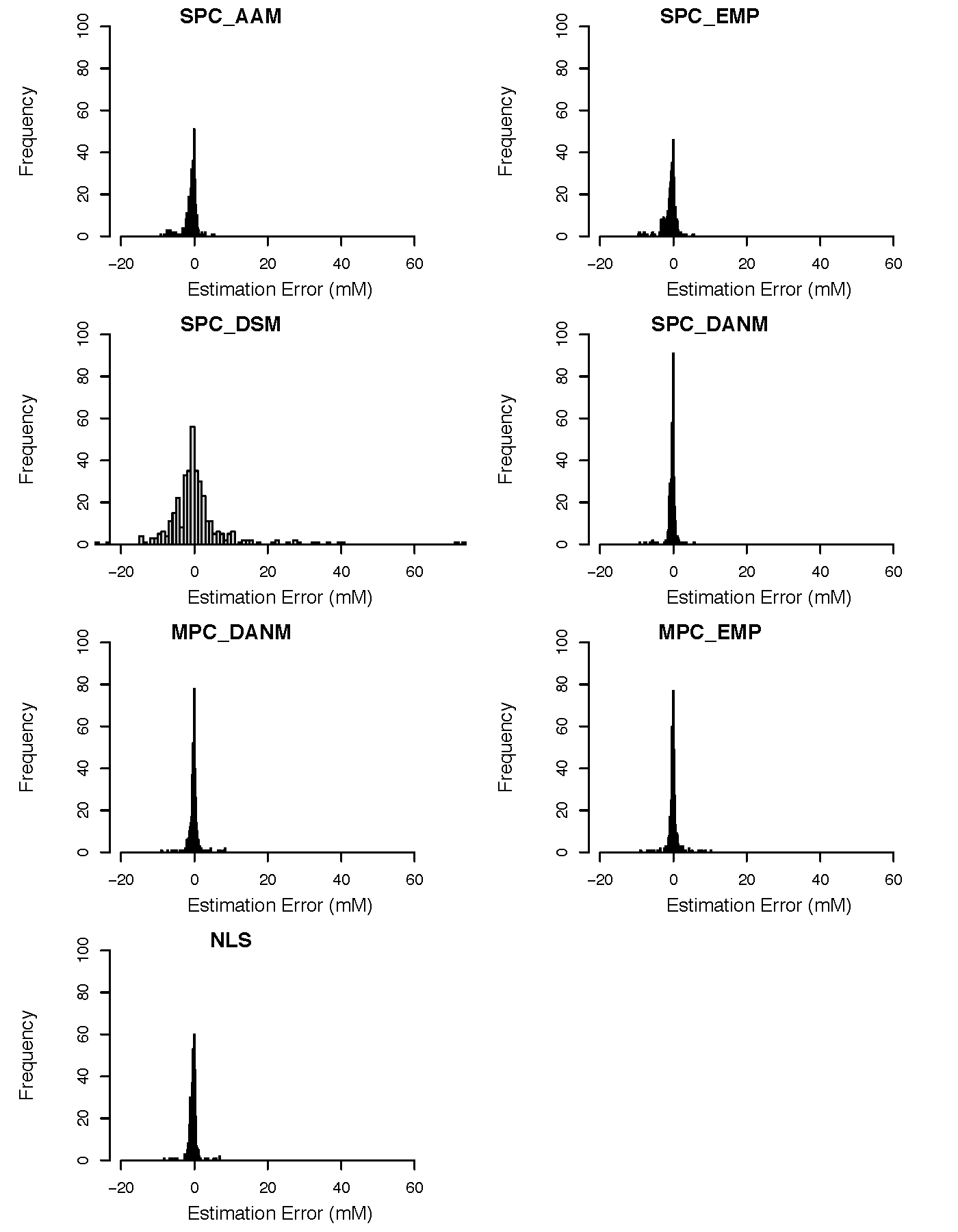


Figure S15 Histograms of estimated errors of seven phase error correction methods in combination of both the single and multiple metabolite spike-in experiments.

**Supplement 17. Correlation analysis of phase error correction methods using metabolite concentration estimation error heatmaps**

In this section, we explore the correlation among different phase error correction methods in metabolomics experiments using metabolite concentration estimation error heatmaps. Three heatmaps (S16-18) were generated to display the correlation among phase error correction methods based on metabolite concentration estimation error for single metabolite spike-in experiments, multiple metabolite spike-in experiments, and a combination analysis of these two experiment types.

Each heatmap displays the correlation matrix of the phase error correction methods and the corresponding metabolite concentration estimation errors. The rows and columns represent the different phase error correction methods, and the values in each cell represent the correlation coefficient between the two methods. Positive values indicate a positive correlation, while negative values indicate a negative correlation.

The heatmap S16 displays the correlation among phase error correction methods for single metabolite spike-in experiments. The heatmap S17 displays the correlation among phase error correction methods for multiple metabolite spike-in experiments. The heatmap S18 displays the correlation among phase error correction methods for the combination analysis of these two experiment types.

The heatmaps reveal that some phase error correction methods are highly correlated with each other, while others are negatively correlated or uncorrelated. These findings provide valuable insights into the selection and optimization of phase error correction methods in metabolomics experiments.

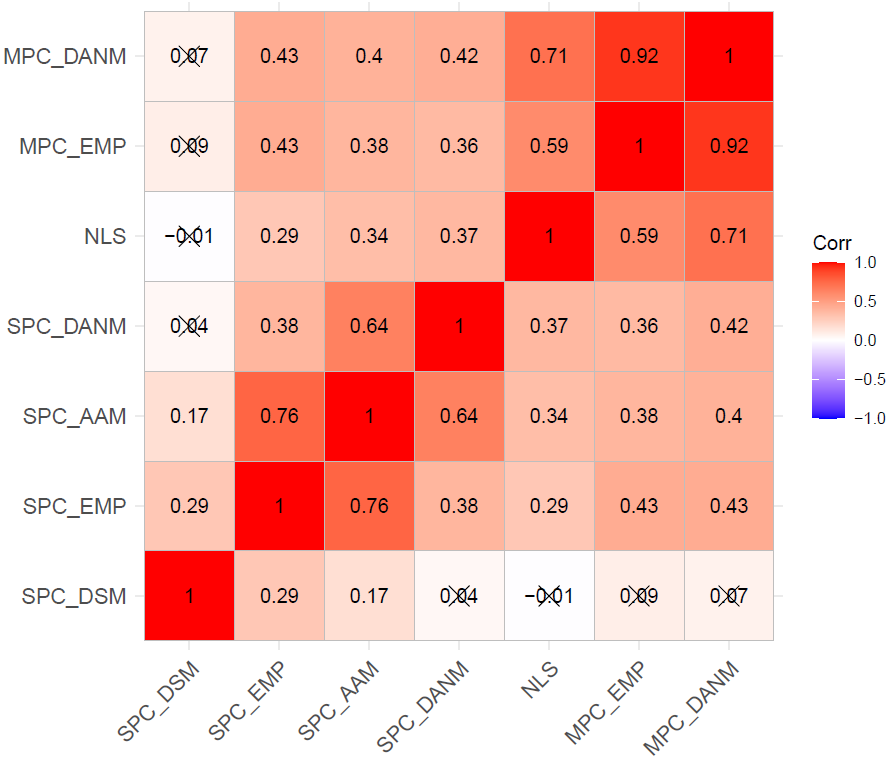


Figure S16 Heatmap of Pearson correlation coefficients of concentration errors among phase error correction methods in the single metabolite spike-in experiment. The corresponding correlation coefficients with p-values greater than 0.05 are crossed out.

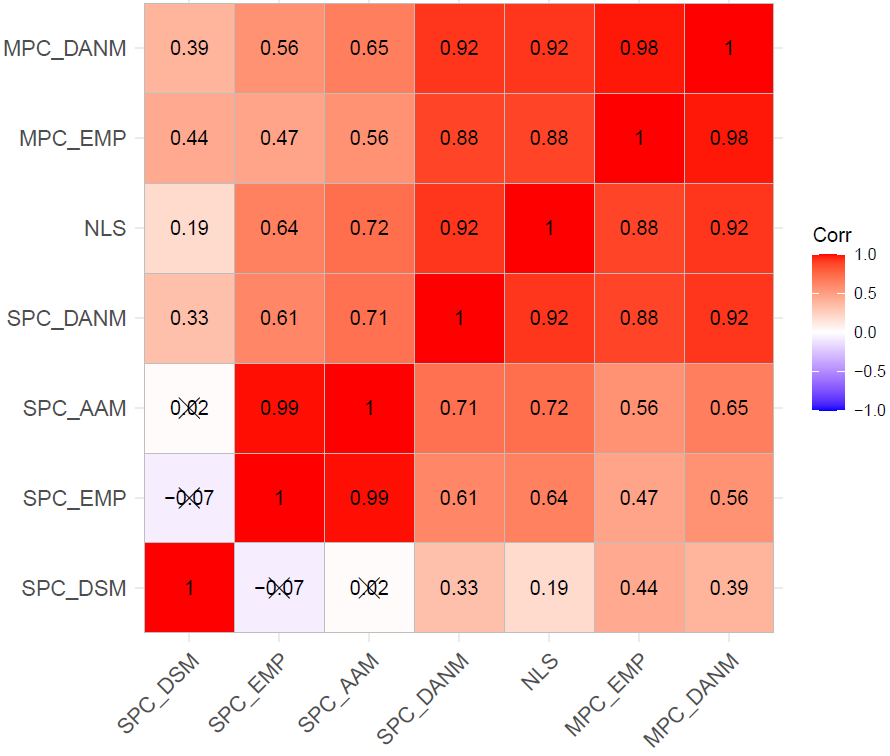


Figure S17. Heatmap of Pearson correlation coefficients of concentration errors among phase error correction methods in the multiple metabolite spike-in experiment. The corresponding correlation coefficients with p-values greater than 0.05 are crossed out.

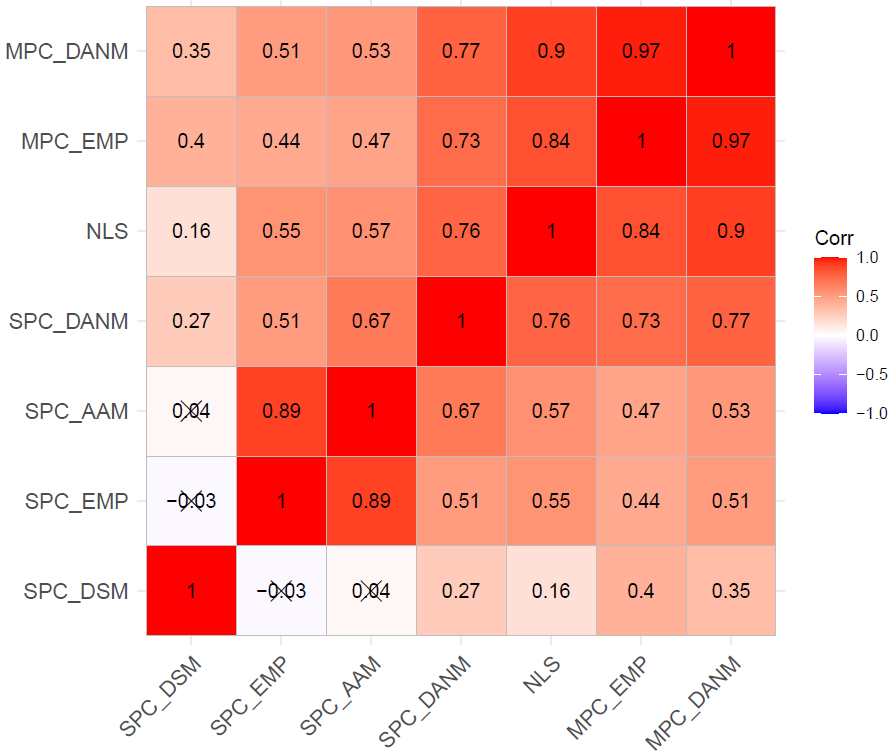


Figure S18. Heatmap of Pearson correlation coefficients of concentration errors among phase error correction methods in combination of both the single and multiple metabolite spike-in experiments. The corresponding correlation coefficients with p-values greater than 0.05 are crossed out.

**Supplement 18. Model results for metabolite spike-in experiments**

In this section, we present the results of our modelling analysis for metabolite spike-in experiments. We used fixed and mixed models to estimate the metabolite concentration estimation errors, depending on the nature of the experiment.

First, we present the fixed model results for the single metabolite spike-in experiment in Figure S19. Since there is only one metabolite per spectrum in this experiment, the mixed model was not used. The fixed model provides accurate estimates of the metabolite concentration estimation errors, which can be used to assess the performance of different methods for correcting phase errors.

Next, we present the mixed model results for the multiple metabolite spike-in experiment in Figure S20. Since there are multiple metabolites per spectrum in this experiment, these metabolites in the same spectrum are correlated, and the mixed model was used to estimate their concentrations. The mixed model takes into account the correlation structure among metabolites and provides more accurate estimates than the fixed model.

Finally, we present the fixed model results for the combination analysis with both the single and multiple metabolite spike-in experiments in Figure S21. The mixed model was not used since it does not work for single metabolite spike-in experiments. The fixed model provides estimates of the metabolite concentration estimation errors in both experiments, which can be compared and analysed to identify common trends and patterns.

Overall, our modelling analysis provides valuable insights into the performance of different phase error correction methods and their impact on metabolite concentration estimation in spike-in experiments. The fixed and mixed models offer flexible and robust approaches for analysing metabolomics data, depending on the nature of the experiment and the research question at hand.

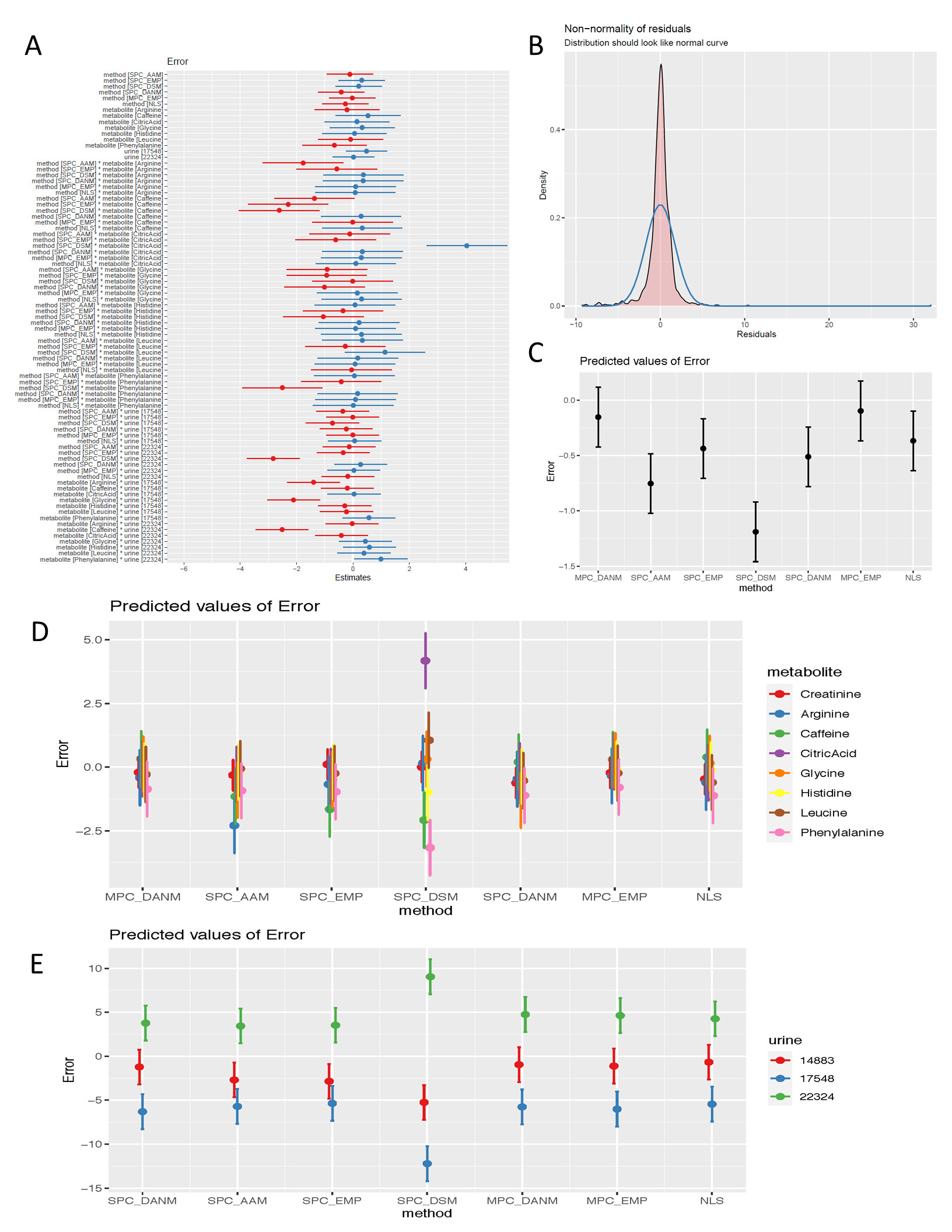


Figure S19. Plots of the fixed effect model for the single metabolite spike-in experiment. (A) Effect forest plot for all terms. (B) Residual plot. (C) Main effect forest plot for phase error correction methods. (D) Plot of interaction between methods and metabolites. (E) Plot of interaction between methods and urine samples.

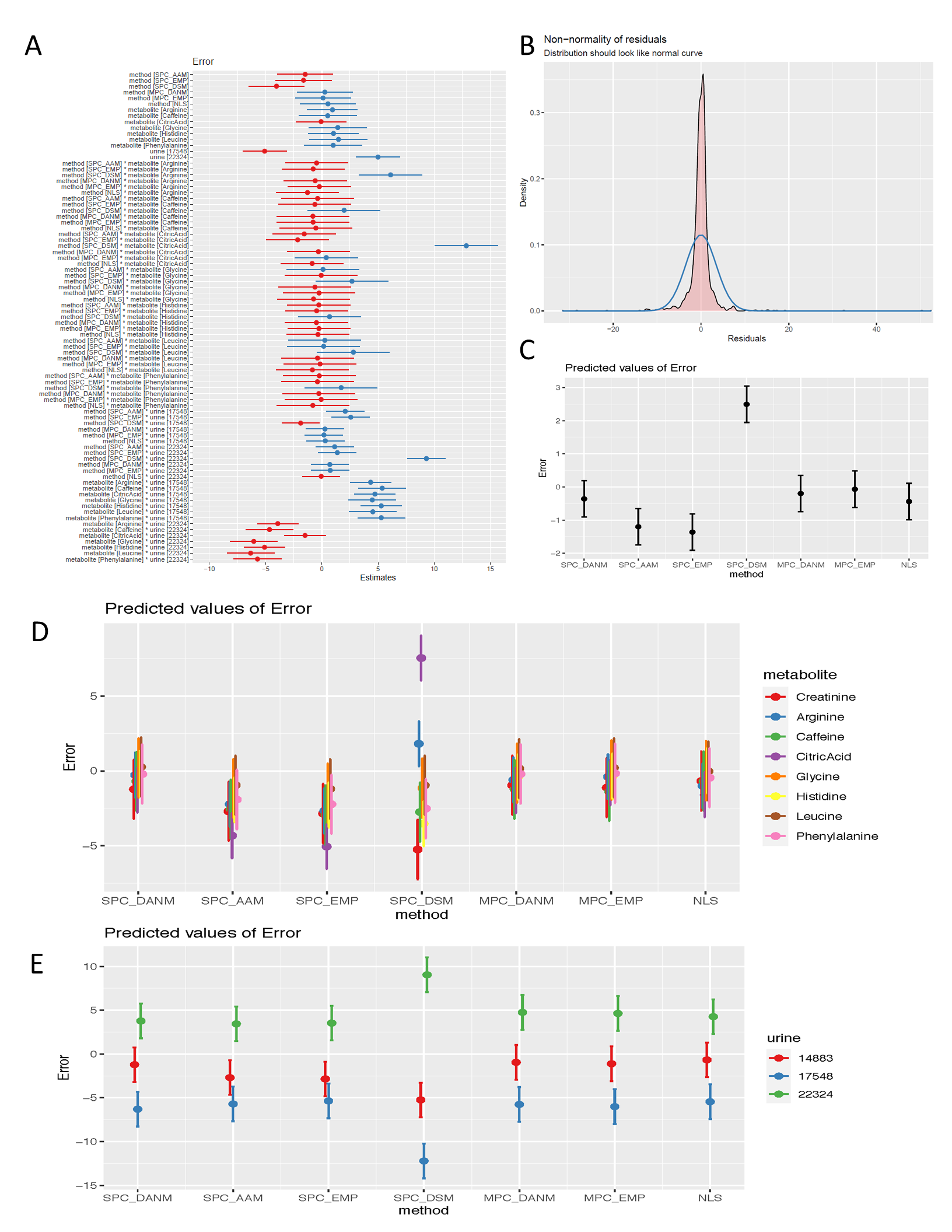


Figure S20 Plots of the fixed effects in the mixed effect model for the multiple metabolite spike-in experiment. (A) Effect forest plot for all fixed terms. (B) Residual plot. (C) Main fixed effect forest plot for phase error correction methods. (D) Plot of fixed interaction between methods and metabolites. (E) Plot of fixed interaction between methods and urine samples.

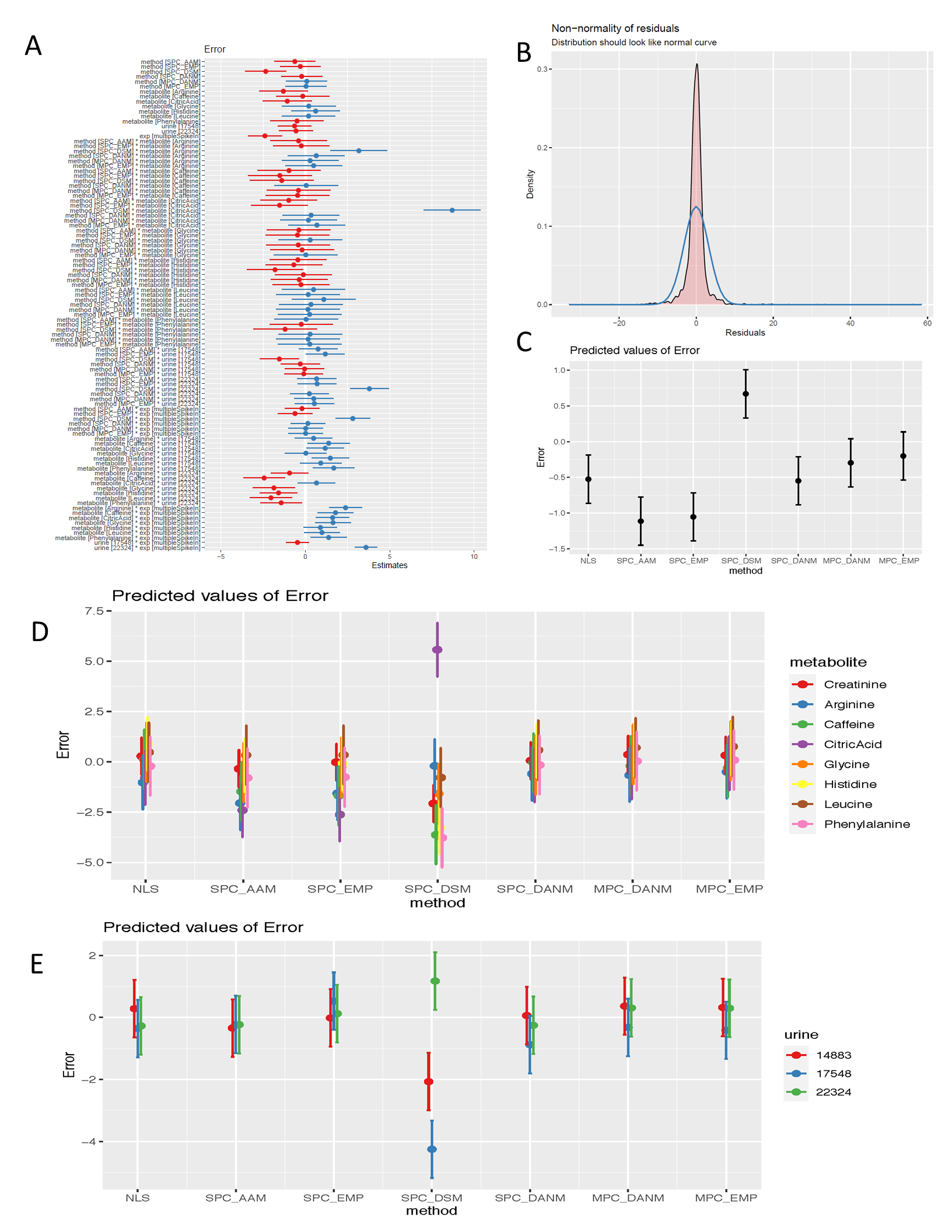


Figure S21 Plots of the fixed effect model for the combined data of both the single and multiple metabolite spike-in experiments. (A) Effect forest plot for all terms. (B) Residual plot. (C) Main effect forest plot for phase error correction methods. (D) Plot of interaction between methods and metabolites. (E) Plot of interaction between methods and urine samples.

**References:**

Binczyk, F., Tarnawski, R., & Polanska, J. (2015). Strategies for optimizing the phase correction algorithms in Nuclear Magnetic Resonance spectroscopy. *Biomedical Engineering Online*, *14 Suppl 2*(Suppl 2), S5. https://doi.org/10.1186/1475-925X-14-S2-S5

de Brouwer, H. (2009). Evaluation of algorithms for automated phase correction of NMR spectra. *Journal of Magnetic Resonance*, *201*(2), 230–238. https://doi.org/10.1016/j.jmr.2009.09.017

Du, P., Kibbe, W. A., & Lin, S. M. (2006). Improved peak detection in mass spectrum by incorporating continuous wavelet transform-based pattern matching. *Bioinformatics (Oxford, England)*, *22*(17), 2059–2065. https://doi.org/10.1093/bioinformatics/btl355

Džakula, Ž. (2000). Phase Angle Measurement from Peak Areas (PAMPAS). *Journal of Magnetic Resonance*, *146*(1), 20–32. https://doi.org/10.1006/jmre.2000.2123

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). lmerTest Package: Tests in Linear Mixed Effects Models. *Journal of Statistical Software; Vol 1, Issue 13 (2017)*. https://doi.org/10.18637/jss.v082.i13