Robust Moiety Model Selection Using Mass Spectrometry Measured Isotopologues

Huan Jin¹ and Hunter N.B. Moseley²,³,⁴,⁵,∗

¹Department of Toxicology and Cancer Biology
²Department of Molecular & Cellular Biochemistry
³Markey Cancer Center
⁴Resource Center for Stable Isotope Resolved Metabolomics
⁵Institute for Biomedical Informatics

University of Kentucky, Lexington KY, United States

∗Corresponding Author: hunter.moseley@uky.edu
Abstract: Stable isotope resolved metabolomics (SIRM) experiments use stable isotope tracers to provide superior mass spectroscopy (MS) and nuclear magnetic resonance (NMR) metabolomics datasets for metabolic flux analysis and metabolic modeling. Several software packages exist for metabolic flux analysis when provided a metabolic model and appropriate isotopomer and/or isotopologue datasets, mostly from $^{13}$C tracer time series experiments. However, assumptions of model correctness can seriously compromise interpretation of metabolic flux results generated from these packages. Therefore, we have developed a metabolic modeling software package specifically designed for moiety model comparison and selection based on the metabolomics data provided. This moiety modeling framework facilitates analysis of time-series SIRM MS isotopologue profiles using a set of plausible moiety models and data in a JSONized representation. The moiety_modeling Python package is available on GitHub and the Python Package Index and provides facilities for model parameter optimization, analysis of optimization results, and model selection. Furthermore, this package is capable of analyzing multi-tracer datasets. Here, we tested the effectiveness of this moiety modeling framework in model selection with two sets of time-series MS isotopologue datasets for uridine diphosphate N-acetyl-D-glucosamine (UDP-GlcNAc) generated from different MS platforms: direct infusion nanoelectrospray Fourier transform MS and liquid chromatography MS. Results generated from the analyses of these datasets demonstrate the robustness of our model selection methods by the successful selection of the optimal model from over 40 models provided. Also, the effects of specific optimization methods, degree of optimization, selection criteria, and specific objective functions on model selection are illustrated. Furthermore, different types of error can exist in the datasets, and proper selection of the objective function can help reduce the optimization side effects caused by the specific types of uncertainty in these datasets. Overall, these results indicate that over-optimization can lead to failure in model selection, but combining multiple datasets can help prevent this overfitting effect. The implication is that SIRM datasets in public repositories of reasonable quality can be combined with newly acquired datasets to improve
model selection. Furthermore, curation efforts of public metabolomics repositories to maintain high data quality could have huge impacts on future metabolic modeling efforts.

**Keywords:** stable isotope resolved metabolomics (SIRM), moiety modeling, model selection, isotopologue deconvolution, overfitting

**Introduction**

Metabolomics is a relatively new field of ‘omics’ technology involving the systematic characterization of metabolites being created and/or utilized in cells, tissues, organisms, and ecosystems. This combined consumption and biosynthesis of metabolites can be represented as flux through specific metabolic paths within cellular metabolism, reflecting specific physiological and pathological states in biomedically useful detail and in ways that are distinct and often more sensitive than other omics methods. It is increasingly recognized that metabolomics biomarkers have great utility in characterizing and monitoring diseases with significant metabolic reprogramming like cancer. Therefore, better regulatory understanding of specific metabolic flux phenotypes of metabolic diseases will aid in developing new therapeutic strategies.

Stable isotope resolved metabolomics (SIRM) experiments utilize stable isotopes from a labeling source to isotopically enrich detected metabolite analytical features, providing more complex but data-rich metabolomics datasets for metabolic flux analysis. Advances in mass spectroscopy (MS) and nuclear magnetic resonance (NMR) greatly contribute to the generation of high-quality SIRM datasets. However, computational methods are required to gain biologically meaningful interpretation from such complex datasets, especially in terms of metabolic flux through specific metabolic paths in cellular metabolism. Most current metabolic flux analysis methods heavily depend on a predetermined metabolic network and are mostly focused on the analysis of $^{13}$C tracer experiments. However, large numbers of ‘unknown’ metabolites in the
metabolomics datasets strongly indicate that current metabolic network are far from complete, especially for secondary metabolism and central metabolism of non-model organisms\textsuperscript{8,9,10}. Without an accurate and reasonably-defined metabolic network, it is not possible to conduct meaningful metabolic flux analyses. Even worse, assuming a metabolic model is accurate compromises the scientific rigor of the metabolic modeling and can lead to misinterpretation of results\textsuperscript{11}.

Our newly developed moiety deconvolution package called moiety\_modeling is a novel method for analyzing time series SIRM MS isotopologue profiles that can involve single or multiple isotope tracers\textsuperscript{12}. This package integrates facilities for moiety (i.e. biochemical functional group) model and data representation, model (parameter) optimization, analysis of optimization results, and model selection under a single moiety modeling framework. A typical data analysis workflow for this moiety modeling framework is shown in Figure 1. First, plausible and hypothetical moiety models of an interesting metabolite are provided by a user based on a relevant metabolic network. Then, each moiety model is optimized. Based on the optimized results of model parameters, the optimal model is selected for downstream metabolic flux analysis and interpretation. In this paper, we use this moiety modeling framework to investigate the effects of the optimization method, optimizing degree, objective function and selection criterion on model selection to identify modeling criteria that promote robust model selection. To our knowledge, this is the first attempt to investigate how all of these factors can affect model selection in metabolic modeling.

**Materials and Methods**

**UDP-GlcNAc time course MS isotopologue datasets**
Two UDP-GlcNAc time course MS isotopologue datasets were used to test the robustness of model selection mechanism. The first is a direct infusion Fourier transform MS (FTMS) UDPGlcNAc $^{13}$C isotopologue dataset derived from LnCaP-LN3 human prostate cancer cells with [U-$^{13}$C]-glucose as isotope labeling source and collected on an Advion Nanomate nanoelectrospray inline connected to a Thermo 7T LTQ Fourier transform ion cyclotron resonance MS (FT-ICR-MS). This dataset includes 3 time points: 34h, 48h, and 72h$^{13}$. The second is a liquid chromatography-MS (LC-MS) UDP-GlcNAc $^{13}$C isotopologue dataset derived from human umbilical vein endothelial cells with [U-$^{13}$C]-glucose as the isotope labeling source and collected on a ThermoFisher Dionex UltiMate 3000 LC System in-line connected to a ThermoFisher Q-Exactive Orbitrap MS. This dataset has 5 time points: 0h, 6h, 12h, 24h and 36h$^{14}$.

**UDP-GlcNAc moiety models**

UDP-GlcNAc can be divided into four distinct moieties: glucose, ribose, acetyl, and uracil, in which isotopes incorporate through a metabolic network from an isotope labeling source. The expected (expert-derived) moiety model of $^{13}$C isotope incorporation from $^{13}$C-labeled glucose to UDP-GlcNAc (see Figure 2B) is built based on well-studied human central metabolism pathways that converge in UDP-GlcNAc biosynthesis, which is corroborated with NMR data$^{13}$. This expert-derived model is labeled as $6\_G1R1A1U3$, representing six optimizable parameters, one for the glucose moiety (G1), one for the ribose moiety (R1), one for the acetyl moiety (A1), and 3 for the uracil moiety (U3), for each moiety state equation representing the fractional $^{13}$C incorporation for each moiety. For example, the g6 state represents the incorporation of $^{13}$C$_6$ into the glucose moiety, whereas the g0 state represents no incorporation of $^{13}$C. Since both g0 and g6 must sum to 1, there is only one parameter that needs to be optimized for this moiety state equation. The set of isotopologue intensity equations are derived using the moiety model parameters and Equation 1, as illustrated for the expert-derived model in Figure 2B. Figure 2C shows an alternative hypothetical moiety model $7\_G0R3A1U3\_g3R2R3\_g6r5\_r4$ along with the isotopologue intensity equations generated from the model.
We also manually crafted 40 hypothetical moiety models to capture isotope flow from \([\text{U-}^{13}\text{C}]\)-glucose into each moiety. This set of models provides a mechanism for testing how robustly the expert-derived model can be selected from all the other models.

### Objective functions

We used four distinct forms of the objective function (Table 1) that compares the observed isotopologues and corresponding calculated isotopologues derived from model parameters obtained from model optimization. The first is a summation of absolute differences between observed and calculated isotopologues, which is generally expected to work well with data where the dominant type of error is additive. The second is a summation of the absolute differences.
between the log of observed and calculated isotopologues, which is generally expected to work well with data where the dominant type of error is proportional.

**Table 1.** Objective functions.

| Objective function                  | Equation                                      |
|-------------------------------------|-----------------------------------------------|
| Absolute difference                 | $\sum|I_{n,obs} - I_{n,calc}|$               |
| Absolute difference of logs         | $\sum|\log(I_{n,obs}) - \log(I_{n,calc})|$     |
| Square difference                   | $\sum(I_{n,obs} - I_{n,calc})^2$              |
| Difference of AIC                    | $2k + n\ln(RSS/n)$                            |

$k$ is the number of parameters.  
$n$ is the number of data points.  
RSS is the residual sum of squares: $RSS = \sum_{i=1}^{n}(I_{obs} - I_{calc})^2$.

**Optimization methods**

From a mathematics perspective, model optimization is actually a non-linear inverse problem. Several different optimization methods were used to solve this problem, including the SAGA-optimize method$^{12}$, and three other optimization methods (‘TNC’$^{15}$, ‘SLSQP’$^{16}$, and ‘L-BFGS-B’$^{17}$) available in the scipy.optimize Python module.

**Model selection criteria**

We used three different quality estimators (Table 2) in model selection: the Akaike Information Criterion (AIC)$^{18}$, the sample size corrected Akaike Information Criterion (AICc)$^{19}$, and the Bayesian Information Criterion (BIC)$^{20}$. The Akaike information criterion (AIC) is biased to select models with more parameters when the sample size is small, which can lead to overfitting$^{18}$. The sample size corrected AIC (AICc) was developed to handle this bias and prevent overfitting$^{19}$. The Bayesian information criterion (BIC) is another criterion commonly used in model selection$^{20}$.

**Table 2.** Model selection estimators.

| Selection Criterion               | Equation                                      |
|-----------------------------------|-----------------------------------------------|
| Akaike Information Criterion (AIC)| $2k + n\ln(RSS/n)$                            |
| Sample size corrected AIC (AICc)  | $AIC + (2k^2 + 2k)/(n - k - 1)$                |
| Bayesian Information Criterion (BIC) | $n\ln(RSS/n) + k\ln(n)$                     |

$k$ is the number of parameters.  
$n$ is the number of data points.  
RSS is the residual sum of squares: $RSS = \sum_{i=1}^{n}(I_{obs} - I_{calc})^2$.  

Results and Discussion

A simple comparison of two moiety models

Model optimization aims to minimize an objective function that compares calculated isotopologues based on moiety state parameters from the model to the directly observed, experimentally-derived isotopologues. Figure 3A shows the comparison of optimized model parameters between the expert-derived moiety model (6_G1R1A1U3) and the hypothetical moiety model (7_G0R3A1U3_g3r2r3_r4) for three time points of isotopologue intensity data, i.e. three sets of isotopologue intensities. In these model optimizations, the SAGA-optimize method and absolute difference objective function were used and DS0, DS1, and DS2 correspond to the 34h, 48h, and 72h time points in the FT-ICR-MS UDP-GlcNAc dataset. We can easily tell that the relative intensity of the corresponding model parameters between these two models are quite different, suggesting that the moiety-specific $^{13}$C isotopic incorporation derived from the same MS isotopologue profile varies from one model to another. Furthermore, experiment-derived and model parameter-calculated isotopologue profiles are shown in Figure 3B, illustrating how much better the expert-derived model vs an inaccurate model is able to reflect the observed data.

Effects of optimization method on model selection

The first question we were interested in was whether the optimization method could affect the model selection results. As in the previous analysis, we used 3 time points from the FT-ICR-MS dataset, the AICC criterion, and an absolute difference objective function in the initial trial. The optimization for each model was conducted 100 times, and we used the average of the 100 optimization results in the analysis (see Table 3). Most optimization methods can select the expert-derived model except for ‘SLSQP’. What interested us most was that the ‘SLSQP’ method
failed in model selection with the lowest loss value (value returned from objective function) and is

Figure 3. Optimized results for 6_G1R1A1U3 and 7_G0R3A1U3_g3r2r3_r4 models. Each model optimization was conducted 100 times. A) Comparison of mean of optimized model parameters with standard deviation. B) Reconstruction of the isotopologue distribution of UDP-GlcNAc from model parameters. Observed isotopologue data was compared with the mean of calculated isotopologue data with standard deviation from the optimized parameters for each model.
generally considered to be the fastest converging of the optimization methods we tested.

| Optimization method | Loss value | AICc   | Selected model          |
|---------------------|------------|--------|-------------------------|
| SAGA                | 0.469      | -401.760 | Expert-derived model    |
| SLSQP               | 0.320      | -408.341 | 7_G2R1A1U3_g5          |
| L-BFGS-B            | 0.763      | -342.164 | Expert-derived model    |
| TNC                 | 0.870      | -327.344 | Expert-derived model    |

Dataset: FT-ICR-MS (combined); Selection criterion: AICc; Objective function: Absolute difference.

Table 3. Comparison of optimization methods in model selection.

We repeated the experiment with the ‘SLSQP’ method 10 times and found that model selection fails when the loss value approaches 0.3 (Supplement Figure 1), suggesting strong instability of model selection at a critical point. Model optimization aims to minimize the objective function, which is actually a non-linear inverse problem, and one inherited issue in solving a non-linear inverse problem is overfitting (i.e. fitting to error in the data). Therefore, we developed the hypothesis that over-optimization of model parameters can lead to failure in model selection.

Over-optimization leads to failure in model selection

To test the above hypothesis, we first tried to increase the stop criterion of ‘SLSQP’ method to avoid over-optimization. The results are shown in Table 4. When optimization stops earlier, the expert-derived model can be selected, which supports our hypothesis.

| Optimization method | Loss value | AICc   | Selected model              | Stop criterion |
|---------------------|------------|--------|-----------------------------|----------------|
| SLSQP               | 0.320      | -408.341 | 7_G2R1A1U3_g5              | ‘ftol’: 1e-06  |
| SLSQP               | 0.514      | -393.934 | Expert-derived model        | ‘ftol’: 1e-05  |

Dataset: FT-ICR-MS (combined); Selection criterion: AICc; Objective function: Absolute difference.

Table 4. Over optimization experiments with ‘SLSQP’ method.

The SAGA-optimize method is more flexible in controlling the degree of optimization simply by adjusting the number of optimization steps. The more steps, the lower the average loss value reached by the optimization. Next, we performed a set of experiments using the SAGA-optimize method with increasing number of optimization steps to further validate the hypothesis. The results are summarized in the Table 5. We can see that the loss value decreases as optimization step increases. When the loss value reaches a certain critical point, the expert-
derived model cannot be selected, further supporting the hypothesis that over-optimization can lead to failure in model selection. Furthermore, the selected model can change with increasing degrees of over-optimization.

Table 5. Over optimization experiments with SAGA-optimize method.

| Optimization steps | Loss value | AICc   | Selected model          |
|--------------------|------------|--------|-------------------------|
| 500                | 2.070      | -219.488 | Expert-derived model     |
| 1000               | 1.754      | -235.728 | Expert-derived model     |
| 2000               | 1.377      | -260.654 | Expert-derived model     |
| 5000               | 0.941      | -305.651 | Expert-derived model     |
| 10000              | 0.664      | -375.192 | Expert-derived model     |
| 25000              | 0.469      | -401.760 | Expert-derived model     |
| 50000              | 0.408      | -414.737 | Expert-derived model     |
| 75000              | 0.328      | -418.228 | 7_G2R1A1U3_g5           |
| 100000             | 0.316      | -424.924 | 7_G1R2A1U3_r4           |

Dataset: FT-ICR-MS (combined); Selection criterion: AICc; Objective function: Absolute difference.

Based on the above results, we conclude that it is not the optimization method but the degree of optimization that affects model selection, which is explained by overfitting to error in the data when solving a non-linear inverse problem. When optimization reaches a certain critical point, successful model selection cannot be guaranteed. Therefore, proper control of the degree of optimization is of great importance in model selection.

Effects of selection criterion on model selection

Next, we investigated whether selection criterion could affect model selection. We compared the model selection results using different model selection criteria (see Table 6). From these results, we can see that the rank of top models is quite consistent across different selection criteria, suggesting that these model selection criteria have little effect on robust model selection, at least under this model selection context. Since our previous experiments used AICc as the selection criterion, we will stick with AICc in the following experiments.

Table 6. Comparison of model rank based on different model selection criteria.

| Models                  | AICc    | rank | AIC     | rank | BIC     | rank |
|-------------------------|---------|------|---------|------|---------|------|
| Expert-derived model    | -401.7597 | 1    | -421.3026 | 1    | -385.5009 | 1    |
| 7_G1R1A2U3              | -384.3075 | 2    | -413.1825 | 2    | -371.4139 | 2    |
| 7_G2R1A1U3_g5           | -381.2868 | 3    | -410.1618 | 3    | -368.3932 | 3    |
Effects of the objective function on model selection

| Model        | Objective Value | Weight 1 | Weight 2 | Weight 3 | Weight 4 |
|--------------|-----------------|----------|----------|----------|----------|
| 7_G1R2A1U3_r3 | -379.2657       | 4        | -408.1407| 4        | -366.3720| 4        |
| 7_G1R2A1U3_r4 | -378.8969       | 5        | -407.7719| 5        | -366.0033| 5        |
| 7_G2R1A1U3_g4 | -375.9538       | 6        | -404.8288| 6        | -363.0601| 6        |
| 6_G1R1A1U3_g5 | -374.9694       | 7        | -394.5122| 10       | -358.7105| 8        |
| 6_G1R1A1U3_r4 | -374.1820       | 8        | -393.7249| 11       | -357.9231| 9        |
| 7_G1R1A1U4    | -373.4563       | 9        | -402.3313| 7        | -360.5626| 7        |
| 6_G1R1A1U3_u4 | -370.0716       | 10       | -389.6145| 13       | -353.8127| 11       |
| 7_G2R1A1U3_g1 | -367.8353       | 11       | -396.7103| 8        | -354.9416| 10       |
| 7_G2R1A1U3_g2 | -360.1668       | 12       | -389.0418| 14       | -347.2732| 13       |
| 7_G1R1A1U3C1  | -360.0296       | 13       | -388.9046| 15       | -347.1360| 14       |
| 7_G1R2A1U3_r1 | -354.8814       | 14       | -383.7564| 16       | -341.9878| 16       |
| 8_G1R2A2U3_r3 | -354.4480       | 15       | -395.8273| 9        | -348.0917| 12       |
| 8_G2R1A2U3_g4 | -351.9886       | 16       | -393.3679| 12       | -345.6323| 15       |
| 6_G0R2A1U3_g3r3r3_g6r5 | -345.1277 | 17 | -364.6706 | 21 | -328.8689 | 17 |
| 8_G2R1A2U3_g1 | -334.2882       | 18       | -375.6675| 17       | -327.9319| 18       |
| 7_G2R1A1U3_g3 | -332.9148       | 19       | -361.7898| 22       | -320.0211| 19       |
| 7_G1R2A1U3_r2 | -332.3262       | 20       | -361.2012| 23       | -319.4326| 21       |
| 8_G1R2A2U3_r1 | -325.9344       | 21       | -367.3137| 18       | -319.5781| 20       |
| 8_G1R1A2U3C1  | -324.5196       | 22       | -365.8989| 19       | -318.1633| 22       |
| 8_G2R1A2U3_g5 | -324.5004       | 23       | -365.8797| 20       | -318.1441| 23       |
| 7_G0R2A2U3_g3r3r3_g6r5 | -324.0749 | 24 | -352.9499 | 26 | -311.1813 | 25 |
| 7_G1R2A1U3_g3r3r3 | -324.0721 | 25 | -352.9471 | 27 | -311.1784 | 26 |
| 8_G2R1A2U3_g2 | -318.5771       | 26       | -359.9564| 24       | -312.2208| 24       |
| 6_G1R1A1U3_a1 | -318.2498       | 27       | -337.7927| 31       | -301.9910| 28       |
| 8_G1R2A2U3_r4 | -317.3169       | 28       | -358.6962| 25       | -310.9606| 27       |
| 8_G2R1A2U3_g3 | -302.7897       | 29       | -344.1690| 28       | -296.4334| 29       |
| 8_G1R2A2U3_g3r3r3_g6r5_g5 | -297.7429 | 30 | -339.1222 | 30 | -291.3866 | 30 |
| 8_G1R2A2U3_r3r3 | -295.0078       | 31       | -336.3871| 32       | -288.6515| 31       |
| 8_G1R2A2U3_r2 | -294.7900       | 32       | -336.1693| 33       | -288.4337| 32       |
| 8_G1R2A2U3_g3r3r3 | -292.7867 | 33 | -334.1660 | 34 | -286.4304 | 33 |
| 9_G2R2A2U3_r2r3_g6r5_g3_g5 | -281.8920 | 34 | -340.0458 | 29 | -286.3433 | 34 |
| 7_G0R3A1U3_g3r3r3_g6r5_g5r4 | -279.0349 | 35 | -307.9099 | 37 | -266.1412 | 36 |
| 9_G2R2A2U3_r2r3_g4 | -273.5807       | 36       | -331.7345| 35       | -278.0320| 35       |
| 9_G2R2A2U3_r2r3_g5 | -254.4087       | 37       | -312.5625| 36       | -258.8599| 37       |
| 9_G2R2A2U3_r2r3_g3 | -248.2277       | 38       | -306.3815| 38       | -252.6789| 38       |
| 9_G2R2A2U3_r2r3_g2 | -242.9984       | 39       | -301.1522| 39       | -247.4497| 39       |
| 9_G2R2A2U3_r2r3_g1 | -242.4110       | 40       | -300.5648| 40       | -246.8623| 40       |
| 7_G0R3A1U3_g3r3r3_g6r5_r4 | -226.7271       | 41       | -255.6021| 41       | -213.8334| 41       |

Dataset: FT-ICR-MS (combined); Optimization method: SAGA-optimize (25000 steps); Objective function: Absolute difference.
Considering that the dominant type of error existing in metabolomics datasets may vary from dataset to dataset, different forms of objective function may affect model optimization and then influence the results of model selection. Here, we test the use of four objective functions in the context of model selection: absolute difference, absolute difference of logs, square difference, and difference of AIC. To speed up optimization, we first split the FT-ICR-MS dataset based on time point (34h, 48h, 72h) into separate model optimizations executed on their own CPU core, and then combine the optimization results for the model selection. This functionality is provided by the moiety_modeling package. We set a series of experiments for each objective function with SAGA-optimize method. The results are shown in Table 7 to Table 10. In comparing Table 7 to Table 5, the number of optimizations per time point provides roughly the same degree of optimization as three times the number of optimization steps used on a combined optimization.

From these tables, we can see that optimization with the absolute difference of logs objective function is less likely to fail (> 250000 steps) in the model selection compared to the other three objective functions (10000 - 20000 steps). One interpretation from these results is that the FT-ICR-MS dataset is dominated by proportional error instead of additive error. However, the AICc produced with the absolute difference of logs is significantly higher (less negative) than that produced by the other objective functions. Therefore, this objective function may simply be hindering efficient optimization, especially if the dataset is dominated by an error structure that is not as compatible with this objective function. From this alternative viewpoint, additive error may actually dominate this dataset. We used a graphical method to visualize errors in both FT-ICR-MS and LC-MS datasets (Figure 4 and 5). For the plots of FT-ICR-MS datasets, we used another dataset generated from the same procedure, which included two replicates at 0, 3h, 6h, 11h, 24h, 34h, and 48h time points. For two replicates with proportional error, a scatter plot of each replicate against the other will show an increasing spread of values with increasing signal, and the log-transformed data will collapse into a line. Plot of two replicates with additive error can be viewed as uniformly deviated from the line of identity, but once log-transformed will show an increasing
spread of values with decreasing signal. The original plots of raw data indicate existence of proportional error in both FT-ICR-MS and LC-MS datasets (Figure 4A and 5A). However, the original plots of normalized data almost collapsed to a straight line (Figure 4C and 5C), suggesting that normalization somehow removes the proportional error in the raw data. In addition, from the log-transformed plots, we can see that additive error does not exist in the normalized FT-ICR-MS datasets (Figure 4D), but does exist in the normalized LC-MS datasets (Figure 5D). The replicate plots of all time points (S Figure 1-2) show similar tendency with selected optimized datasets.

Based on the above results, the absolute difference of logs objective function can hinder efficient optimization in FT-ICR-MS datasets. We also compared four objective functions in the context of model selection with LC-MS datasets (S Table 1-4). From these tables, we can see that model selection fails earlier with absolute difference of logs objective function compared to other objective functions, also suggesting that additive error may dominate in the normalized LC-MS datasets. Based on the above results, the objective function clearly affects model selection and the selection of certain objective functions for model optimization is able to resist failure in model selection caused by over-optimization; however, this is likely due to less efficient model optimization caused by the selection of an objective function not appropriate for the type of error in the data.

| Optimization steps | Loss value | AICc       | Selected model          |
|--------------------|------------|------------|-------------------------|
| 500                | 1.045      | -293.540   | Expert-derived model    |
| 1000               | 0.819      | -330.411   | Expert-derived model    |
| 2000               | 0.651      | -361.038   | Expert-derived model    |
| 5000               | 0.459      | -408.167   | Expert-derived model    |
| 10000              | 0.392      | -422.516   | Expert-derived model    |
| 15000              | 0.359      | -431.276   | Expert-derived model    |
| **20000**          | **0.290**  | **-434.468**| **7_G1R1A2U3**          |
| 25000              | 0.285      | -436.909   | **7_G1R1A2U3**          |

Dataset: FT-ICR-MS (split); Selection criterion: AICc; Objective function: absolute difference.
Table 8. Model selection test with square difference objective function.

| Optimization steps | Loss value | AICc      | Selected model          |
|--------------------|------------|-----------|-------------------------|
| 500                | 0.085      | -298.516  | Expert-derived model    |
| 1000               | 0.047      | -330.096  | Expert-derived model    |
| 2000               | 0.023      | -367.279  | Expert-derived model    |
| 5000               | 0.011      | -404.509  | Expert-derived model    |
| 10000              | 0.007      | -425.695  | Expert-derived model    |
| **15000**          | **0.005**  | **-429.869** | **7_G2R1A1U3_g5**   |
| 20000              | 0.005      | -435.348  | **7_G1R2A1U3_r4**      |

Dataset: FT-ICR-MS (split); Selection criterion: AICc; Objective function: square difference.

Table 9. Model selection test with difference of AIC objective function.

| Optimization steps | Loss value | Selected model          |
|--------------------|------------|-------------------------|
| 500                | -345.559   | Expert-derived model    |
| 1000               | -371.852   | Expert-derived model    |
| 2000               | -398.570   | Expert-derived model    |
| 5000               | -436.582   | Expert-derived model    |
| **10000**          | **-458.064** | **7_G1R1A2U3**         |
| 15000              | -467.960   | **7_G2R1A1U3_g5**      |

Dataset: FT-ICR-MS (split); Selection criterion: AICc; Objective function: difference of AIC.

Table 10. Model selection test with absolute difference of logs objective function.

| Optimization steps | Loss value | AICc      | Selected model          |
|--------------------|------------|-----------|-------------------------|
| 500                | 31.647     | -221.501  | Expert-derived model    |
| 1000               | 29.628     | -223.363  | Expert-derived model    |
| 2000               | 28.164     | -224.330  | Expert-derived model    |
| 5000               | 27.096     | -225.911  | Expert-derived model    |
| 10000              | 26.631     | -227.499  | Expert-derived model    |
| 15000              | 25.865     | -229.926  | Expert-derived model    |
| 20000              | 25.777     | -230.232  | Expert-derived model    |
| 25000              | 26.271     | -228.178  | Expert-derived model    |
| 50000              | 26.126     | -228.892  | Expert-derived model    |
| 100000             | 25.949     | -228.926  | Expert-derived model    |
| 150000             | 25.865     | -229.926  | Expert-derived model    |
| 200000             | 25.777     | -230.232  | Expert-derived model    |

Dataset: FT-ICR-MS (split); Selection criterion: AICc; Objective function: absolute difference of logs.
**Figure 4.** Error analysis in FT-ICR-MS datasets. A and B are plots of raw data. C and D are plots of renormalized data after natural abundance correction. All these plots contain 3 time points (24 – 48h).
Effects of information quantity on model selection

From the above experiments, we found that over-optimization is a primary cause for failure in model selection and this is affected by the objective function used. The next question is whether the quantity of information affects model selection. One basic approach is to utilize more datasets in order to overcome the effects of over-optimization. In the following experiments, we repeated single model optimization 10 times in order to pragmatically finish these computational experiments. Every experiment was conducted 10 times using the AICc criterion and the absolute difference objective function. We used the SAGA-optimize method to test where model selection starts to fail.

Figure 5. Error analysis in LC-MS datasets. A and B are plots of raw data. C and D are plots of renormalized data after natural abundance correction. All these plots contain 3 time points (12 – 36h).
First, we used decreasing number of time points of the LC-MS dataset to test whether data quantity affects model selection (Figure 6A). However, model selection failed with few optimization steps when all five time points were included and when only one time point was included, with the most robust model selection occurring with 3 time points. Initially, these results were not expected, until we realized that the relative isotopologue intensity of the 0 and 6h time points is concentrated within the $^{13}$C₀ isotopologue with zero $^{13}$C tracer. Thus, these datasets are less informative with respect to capturing the isotope flow from labeling source to each moiety in the metabolite. When the 0 and 6h time points are removed, the selection results improved significantly. Likewise, when information-rich time points are removed, the model selection robustness decreases as well. Similar results were obtained when testing the FT-ICR-MS dataset (Figure 6B). Taken together, the addition of information-rich data contributes to successful model selection while the addition of information-poor data detracts from successful model selection. To further test this concept, we investigated whether combining FT-ICR-MS (34h, 48h, 72h) and LC-MS (12h, 24h, 36h) datasets can prevent failure in model selection (Figure 6C). From the comparison, we can see that combining information-rich FT-ICR-MS and LC-MS datasets is much more resistant to failure of model selection than just using the information-rich FT-ICR-MS or FT-ICR-MS dataset, strongly supporting our previous conclusions that utilizing more information-rich datasets can prevent failure in model selection. Similar results were obtained with absolute difference of logs objective function (S Figure 3). These datasets were collected at different times, on very different mass spectrometry platforms. One used chromatographic separation while the other utilized direct infusion. However, the really surprising part is that the datasets were derived from different human cell cultures: LnCaP-LN3 human prostate cancer cells and human umbilical vein endothelial cells.
Figure 6. Comparison of the log optimization steps where model selection with different datasets begins to fail. A) Test with LC-MS datasets. LC-MS_1 to LC-MS_5 represent LC-MS datasets with 36h, 24-36h, 12-36h, 6-36h and 0-36h. B) Test with FT-ICR-MS datasets. FT-ICR-MS_1 to FT-ICR-MS_3 represent FT-ICR-MS datasets with 48h, 48-72h and 34-72h. C) Test with combination of LC-MS and FT-ICR-MS datasets. The median values are indicated in the plots.
Conclusions
Here, we discussed the importance of model selection in isotopic flux analysis as a proxy for metabolic flux analysis and factors that affect robust model selection. We found that it is not the optimization method per se, but the degree of optimization that influences model selection, due to the effects of over-optimization, i.e. fitting of model parameters based on the error in the data. Overfitting is a known problem typically due to the ill-conditioning of the nonlinear inverse problem that is partially ill-posed. Moreover, the objective function in model optimization is also of great importance in model selection. Proper selection of an objective function can help increase resistance to failure in model selection. This may mean that different objective functions should be used for model selection versus parameter optimization for flux interpretation. Furthermore, we found that combining informative datasets can help prevent failure in model selection, which is of great practical significance. Most SIRM experimental datasets have few collected time points due to the cost and effort required to acquire these datasets. Our finding indicates that informative datasets in public metabolomics repositories can be combined to facilitate robust model selection. Moreover, these datasets do not need to come from identical biological systems, just biological systems that utilize the same part of metabolism being measured and modeled. The implication is that SIRM datasets in public repositories of reasonable quality can be combined with newly acquired datasets to improve model selection. Furthermore, curation efforts of public metabolomics repositories to maintain high data quality and provide metrics of measurement error could have a huge impact on future metabolic modeling efforts.

Supplementary Materials

Author Contributions
HJ and HNBM worked together on the design of the experiments and the analysis of the results. HJ wrote the manuscript and HNBM helped revise it. All authors have read and approved the manuscript.

Funding
The work was supported in part by grant NSF 1419282 (PI Moseley).

Acknowledgments

Availability of data and materials

All data used and the results generated in this manuscript are available on figshare:

https://figshare.com/articles/moiety_model_selection/10279688

Conflicts of Interest

The authors declare that they have no competing interests.

References

1. Wishart DS. Applications of Metabolomics in Drug Discovery and Development. *Drugs R D*. 2008;9(5):307-322. doi:10.2165/00126839-200809050-00002

2. DeBerardinis RJ, Thompson CB. Cellular Metabolism and Disease: What Do Metabolic Outliers Teach Us? *Cell*. 2012;148(6):1132-1144. doi:10.1016/j.cell.2012.02.032

3. Fan TW-M, Lorkiewicz PK, Sellers K, Moseley HNB, Higashi RM, Lane AN. Stable isotope-resolved metabolomics and applications for drug development. *Pharmacol Ther*. 2012;133(3):366-391. doi:10.1016/j.pharmthera.2011.12.007

4. Young JD. INCA: a computational platform for isotopically non-stationary metabolic flux analysis. *Bioinformatics*. 2014;30(9):1333-1335. doi:10.1093/bioinformatics/btu015

5. Sauer U. Metabolic networks in motion: 13 C-based flux analysis. *Mol Syst Biol*. 2006;2(1):62. doi:10.1038/msb4100109

6. Antoniewicz MR. Methods and advances in metabolic flux analysis: a mini-review. *J Ind Microbiol Biotechnol*. 2015;42(3):317-325. doi:10.1007/s10295-015-1585-x

7. Antoniewicz MR, Kelleher JK, Stephanopoulos G. Elementary metabolite units (EMU): A novel framework for modeling isotopic distributions. *Metab Eng*. 2007;9(1):68-86. doi:10.1016/j.ymgen.2006.09.001

8. Krumsiek J, Suhre K, Evans AM, et al. Mining the Unknown: A Systems Approach to Metabolite Identification Combining Genetic and Metabolic Information. McCarthy MI, ed.
9. Basu S, Duren W, Evans CR, Burant CF, Michailidis G, Karnovsky A. Sparse network modeling and Metscape-based visualization methods for the analysis of large-scale metabolomics data. Bioinformatics. January 2017:btx012. doi:10.1093/bioinformatics/btx012

10. Wilken SE, Swift CL, Podolsky IA, Lankiewicz TS, Seppälä S, O’Malley MA. Linking ‘omics’ to function unlocks the biotech potential of non-model fungi. Curr Opin Syst Biol. 2019;14:9-17. doi:10.1016/j.coisb.2019.02.001

11. Moseley HNB. ERROR ANALYSIS AND PROPAGATION IN METABOLIC DATA ANALYSIS. Comput Struct Biotechnol J. 2013;4(5):e201301006. doi:10.5936/csbj.201301006

12. Jin H, Moseley HNB. Moiety Modeling Framework for Deriving Moiety Abundances from Mass Spectrometry Measured Isotopologues. BMC Bioinformatics. 2019.

13. Moseley HN, Lane AN, Belshoff AC, Higashi RM, Fan TW. A novel deconvolution method for modeling UDP-N-acetyl-D-glucosamine biosynthetic pathways based on 13C mass isotopologue profiles under non-steady-state conditions. BMC Biol. 2011;9(1):37. doi:10.1186/1741-7007-9-37

14. Verdegem D, Moseley HNB, Vermaelen W, Sanchez AA, Ghesquière B. MAIMS: a software tool for sensitive metabolic tracer analysis through the deconvolution of 13C mass isotopologue profiles of large composite metabolites. Metabolomics. 2017;13(10):123. doi:10.1007/s11306-017-1250-7

15. Nash S. Newton-Type Minimization via the Lanczos Method. SIAM J Numer Anal. 1984;21(4):770-788. doi:10.1137/0721052

16. Boggs PT, Tolle JW. Sequential Quadratic Programming. Acta Numer. 1995;4:1. doi:10.1017/S0962492900002518

17. Zhu C, Byrd RH, Lu P, Nocedal J. Algorithm 778: L-BFGS-B: Fortran subroutines for
large-scale bound-constrained optimization. *ACM Trans Math Softw.* 1997;23(4):550-560. doi:10.1145/279232.279236

18. Akaike H. Information Theory and an Extension of the Maximum Likelihood Principle. In: ; 1998:199-213. doi:10.1007/978-1-4612-1694-0_15

19. Cavanaugh JE. Unifying the derivations for the Akaike and corrected Akaike information criteria. *Stat Probab Lett.* 1997;33(2):201-208. doi:10.1016/S0167-7152(96)00128-9

20. Schwarz G. Estimating the Dimension of a Model. *Ann Stat.* 1978;6(2):461-464. doi:10.1214/aos/1176344136
**S Table 1.** Model selection test with absolute difference objective function.

| Optimization steps | Loss value | AICc  | Selected model         |
|--------------------|------------|-------|------------------------|
| 500                | 0.840      | -344.734 | Expert-derived model   |
| 1000               | 0.682      | -368.696 | Expert-derived model   |
| 2000               | 0.580      | -386.000 | Expert-derived model   |
| 5000               | 0.492      | -398.243 | Expert-derived model   |
| 10000              | 0.447      | -402.611 | Expert-derived model   |
| 15000              | 0.430      | -405.722 | Expert-derived model   |
| 25000              | 0.458      | -407.414 | 6_G1R1A1U3             |

Dataset: LC-MS (split); Selection criterion: AICc

**S Table 2.** Model selection test with square difference objective function.

| Optimization steps | Loss value | AICc  | Selected model         |
|--------------------|------------|-------|------------------------|
| 500                | 0.031      | -348.250 | Expert-derived model   |
| 1000               | 0.021      | -368.818 | Expert-derived model   |
| 2000               | 0.015      | -387.196 | Expert-derived model   |
| 5000               | 0.011      | -404.563 | Expert-derived model   |
| 10000              | 0.010      | -411.177 | Expert-derived model   |
| 15000              | 0.010      | -413.499 | Expert-derived model   |
| 25000              | 0.009      | -415.498 | Expert-derived model   |

Dataset: LC-MS (split); Selection criterion: AICc

**S Table 3.** Model selection test with absolute difference of logs objective function.

| Optimization steps | Loss value | AICc  | Selected model         |
|--------------------|------------|-------|------------------------|
| 500                | 50.595     | -315.616 | Expert-derived model   |
| 1000               | 47.213     | -319.026 | Expert-derived model   |
| 2000               | 44.137     | -323.619 | Expert-derived model   |
| 5000               | 40.811     | -320.949 | Expert-derived model   |
| 10000              | 39.474     | -328.551 | Expert-derived model   |
| 15000              | 57.856     | -331.700 | 6_G1R1A1U3             |

Dataset: LC-MS (split); Selection criterion: AICc

**S Table 4.** Model selection test with difference of AIC objective function.

| Optimization steps | Loss value | Selected model         |
|--------------------|------------|------------------------|
| 500                | -365.957   | Expert-derived model   |
| 1000               | -389.987   | Expert-derived model   |
| 2000               | -409.322   | Expert-derived model   |
| 5000               | -427.064   | Expert-derived model   |
| 10000              | -435.618   | Expert-derived model   |
| 15000              | -437.970   | Expert-derived model   |
| 25000              | -439.533   | Expert-derived model   |

Dataset: LC-MS (split); Selection criterion: AICc
S Figure 1. Error analysis in FT-ICR-MS datasets. A and B are plots of raw data. C and D are plots of renormalized data after natural abundance correction. All these plots contain all time points.
**Figure 2.** Error analysis in LC-MS datasets. A and B are plots of raw data. C and D are plots of renormalized data after natural abundance correction. All these plots contain all time points.
S Table 5. Inclusion of less informative dataset can lead to failure in model selection.

| Optimization step | 5 time points (0-36h) | 4 time points (6-36h) | 3 time points (12-36h) | 2 time points (24-36h) | 1 time point (36h) |
|-------------------|-----------------------|-----------------------|------------------------|------------------------|-------------------|
| 500               | ED model              | ED model              | ED model               | ED model               | ED model          |
| 1000              | ED model              | ED model              | ED model               | ED model               | ED model          |
| 2000              | ED model              | ED model              | ED model               | ED model               | ED model          |
| 5000              | 6_G1R1A1U3_u4         | 6_G1R1A1U3_u4         | ED model               | ED model               | 7_G1R2A1U3_r1     |
| 10000             | 6_G1R1A1U3_u4         | 6_G1R1A1U3_u4         | ED model               | 7_G1R2A1U3_r1          | 7_G1R2A1U3_r1     |
| 15000             | 6_G1R1A1U3_u4         | 6_G1R1A1U3_u4         | 6_G1R1A1U3_u4          | 7_G1R2A1U3_r1          | 7_G1R2A1U3_r1     |
| 25000             | 6_G1R1A1U3_u4         | 6_G1R1A1U3_u4         | 6_G1R1A1U3_u4          | 7_G1R2A1U3_r1          | 7_G1R2A1U3_r1     |

Dataset: LC-MS (split); Objective function: log difference; Selection criterion: AICc; Optimization method: SAGA-optimize.

S Figure 3. Comparison of the log of optimization steps where model selection with different datasets begins to fail with absolute difference of logs objective function.