The Application of Robust Analysis Methods on Sparse Data for Mass-Resolved Neutron Spectroscopy

Daniel Nixon (120263697)

May 8, 2016

Abstract

VESUVIO is a one of a kind neutron spectrometer and diffractometer situated at the ISIS facility on the Rutherford Appleton Laboratory (RAL) site in Oxfordshire. Recent development to the physical instrument have seen orders of magnitude improvements in the quality of data produced by the instrument.

This dissertation focuses on the analysis of data obtained from mass-resolved spectroscopic experiments, specifically extending the current analysis workflows to reduce the amount of human involvement required to analyse the data.

This has been attempted through the use of the FABADA minimiser recently implemented within the Manipulation and Analysis Toolkit for Instrument Data (MANTID) framework to create a Bayesian model selection algorithm based on a selection of models created through inspection of mass peaks present in an experimental spectrum.

Declaration

I declare that this dissertation represents my own work except where otherwise stated.

Acknowledgements

I would like to thank my supervisor Dr. Paolo Zuliani for his guidance throughout the project as well as Dr. Matthew Krzystyniak for his valuable scientific input.

Contents

1 Introduction

The subject of this dissertation is centred around the analysis of experimental neutron scattering data produced by the VESUVIO instrument (see 2.1).

The work undertaken is part of an initiative to improve the data reduction and analysis workflows for this unique instrument, which for the majority of the life of the instrument have existed as a collection of proprietary FORTRAN routines, only recently being ported into the widely used MANTID data analysis tooklit.

Parts of this work follow on from work that I undertook during my placement year at the ISIS facility in Oxfordshire which are described in detail in [10].

Recent improvements made to the VESUVIO instrument have allowed the duration of a single run to be significantly reduced, as such the volume of data produced by the instrument per operational cycle of the ISIS facility is growing.

This increased volume of data therefore increased the time taken to process the data as each experiment will require a new driver script to be created for the specific sample, which may then require testing several times in order to tune parameters, such as the intensity constraints, to obtain the best fit of the defined model to the experimental spectrum.

An ideal solution would be that which allows the analysis jobs to be automatically executed as the data is produced by the instrument, taking into account that a single experiment may involve multiple runs (data files) for example in the case where an experiment is performed over a range of sample temperatures.

This is a solution used by several other instruments at the ISIS facility using the recently released auto-reduction service, however currently this is not a feasible option for VESUVIO due to; the reliance on the scientist to define the fitting model, the need to verify the model and parameters and the computational cost of running the current analysis.

This project will aim to address only the first point; the need to manually define the fitting model. This ill be done through the use of Bayesian model selection in which the most appropriate fitting model is selected based on the probability of its hypothesis given the data.

As for the remaining two points; the computational cost can be addressed simply through optimisation of the current workflow, for instance currently only one thread is used to run the overall workflow (multiple threads are used during the more expensive correction algorithms) however as described in section 2.3 the workflow operates by fitting each spectrum in turn which is an obvious place to introduce multithreading due to the isolation of data treatment per spectrum.

The need to manually verify the fitting model is a difficult requirement to remove from any data analysis workflow, especially one in which the model is derived through an automatic model selection scheme.

1.1 Project Aim

The aim of the project is to improve the existing data analysis workflow for the VESUVIO instrument as implemented in the MANTID toolkit.

1.2 Objectives

The objectives of the project can be summarised in two main areas:

- Adding an additional fit function that allows fitting of the multivariate Gaussian Compton peak profile for anisotropic mass peaks, details of which are in section 3
- Providing functionality to allow the peaks in a sample to automatically be assigned a mass, detailed in section 4

2 Background

This section describes the general background to the project. Information specific to each feature implemented is provided in sections 3.1 and 4.1 respectively.

2.1 The VESUVIO instrument

VESUVIO is a one of a kind neutron spectrometer located in target station 1 of the ISIS facility on the RAL site, Oxfordshire.

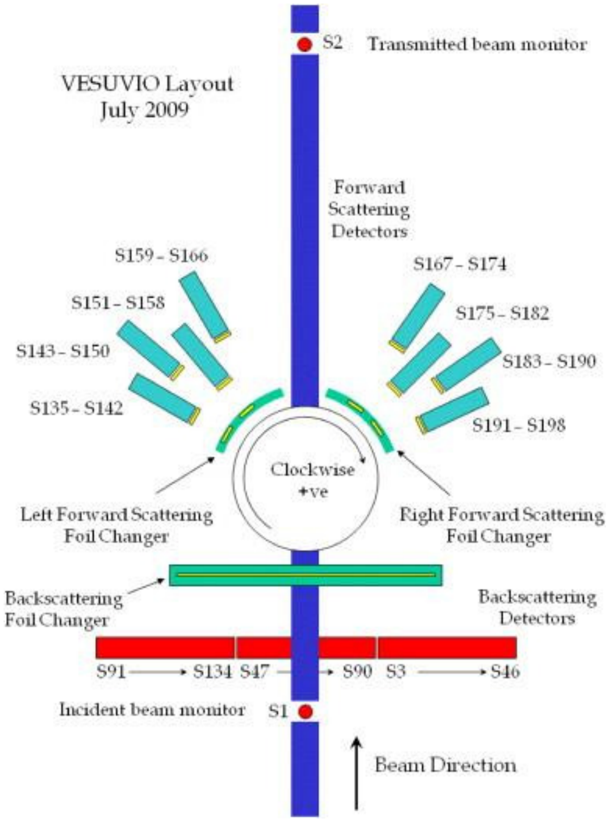

Figure 1: VESUVIO instrument layout [8]

The physical layout of the instrument is detailed by figure 1.

The image shows the sample tank at the centre of the instrument through which the beam is directed, two monitors are used to measure the beam intensity before (S1) and after (S2) it has passed through the sample.

The forward scattering detectors are composed of eight banks of eight Yttrium Aluminium Perovskite (YAP) detectors which detect Gamma rays emitted by the gold foils directly in front of the detectors when they absorb a neutron. Each bank is vertically arranged such that there are four detectors on either side of the horizontal plane [8].

The backscattering detectors are in three banks of 44 $(Li)^6$ doped glass bead neutron detectors.

A measurement mode known as the foil cycling technique [13] is employed on VESUVIO, whereby Gold foils are cycled in and out of the secondary beam path between the sample and analyser/detector. This technique is effectively a combination of the resonance detector and filter difference measurement techniques which can also be configured on VESUVIO.

This technique is used to give improved resolution by narrowing the Full Width Half Maxima (FWHM) of the energy transfer function. The final spectrum is given by subtracting the foil in spectrum from the foil out spectrum.

VESUVIO operates using a technique known as Neutron Compton scattering (NCS) in which momentum distributions are measured by means of inelastic neutron scattering. The technique is similar to traditional Compton scattering in which the momentum of electrons is measured by means of scattering high energy photons.

In principle VESUVIO operates as any other indirect inelastic neutron spectrometer does in that neutrons are scattered by the sample then have their momentum changed by an analyser bank (in the case of VESUVIO this is the gold foils) and then hit a detector.

The initial stages of analysis are therefore identical to an indirect inelastic spectrometer in that given the geometric and neutronic parameters of the instrument the count rate at a given time can be calculated.

The count rate (measured by the detector) is given the standard expression described by equation 2 for an indirect geometry time of flight spectrometer [17] for a system of N identical atoms scattering neurons into a detector at polar coordinates $d\Omega$ and θ.

$$E_0(E_1, t) = \frac{m}{2}\left(\frac{L_0 v}{v_1 t - L_1}\right)^2 \tag{1}$$

$$C(t) = 2\left(\frac{2}{m}\right)^{\frac{1}{2}} \frac{E_0^{\frac{3}{2}}}{L_0} I(E_0) D(E_R) N \frac{d^2\sigma}{d\Omega dE_1} d\Omega \tag{2}$$

where; m is the mass of a neutron, E_0 is the incident neutron energy, E_1 is the final energy of scattered neutrons, L_0 is the length of the incident flight path (from moderator to sample), L_1 is the final flight path (from sample to detector), v_1 is the final neutron velocity and $\frac{d^2\sigma}{d\Omega dE_1}$ is the partial differential scattering cross section of the atom causing scattering.

In VESUVIO a technique known as the Impulse Approximation (IA) forms the bases of the mass resolved data analysis.

Given that the count rate of an indirect spectrometer is given by equation 2, the IA states that at sufficiently heigh momentum transfer the count rate for NCS can be given by equation 3.

In this case given the high momentum transfer it can be assumed that all atoms cause incoherent scattering, hence the count rate can be rewritten as a combination of contributions form several unique atomic masses.

$$C(t) = 2\left(\frac{2}{m}\right)^{\frac{1}{2}} \frac{E_0^{\frac{3}{2}}}{L_0} I(E_0) D(E_R) \sum_M N_M \frac{d^2\sigma_M}{d\Omega dE_1} d\Omega \tag{3}$$

The neutron Compton profile $J_M(y_M, \hat{q})$ is described as the probability distribution of the momentum of atom with mass M along the direction $\hat{q}$. This is related to the count rate by equation 4 [6].

$$C(t) = \frac{E_0 I(E_0)}{q} \sum_M A_M M J_M(y_M) \tag{4}$$

where A_M is given by equation 5.

$$A_M = \frac{2}{L_0} D(E_R) \sqrt{\frac{2E_R}{m}} \Delta\Omega N_M b_M^2 \tag{5}$$

When operating on a single mass peak a transformation from raw detector counts to momentum (y) space is typically performed, this is described by equation 6 [2].

$$y = \frac{M}{\hbar^2 q}\left(\omega - \frac{\hbar^2 q^2}{2M}\right) \tag{6}$$

where; q is momentum transfer.

This transformation is commonly referred to as ”y-scaling” or ”West scaling” [16]. This implies that the momentum transfer, q and energy ω can together be used as a measure of the neutron Compton profile $J(y)$.

The full theory of data treatment for NCS data is described by Mayers [7].

2.2 MANTID

MANTID [4] is a cross platform data analysis suite centred around the reduction, analysis and visualisation of experimental neutron and muon scattering data. It is actively maintained by a collaboration of several facilities throughout the world.

As the existing data reduction workflow for the VESUVIO instrument is implemented as part of the MANTID toolkit, new features implemented as part of this dissertation were done so in such a way that they can be included with this software. This also has the advantage of allowing new features to leverage the existing functionality within the MANTID project.

MANTID provides a framework on which data manipulation algorithms and curve fitting functions can be implemented in such a way that allows use in various situations, i.e. running locally on a users computer or running as part of a batch analysis workflow.

The majority of the MANTID code base is C++ with support for Python scripting.

The core concepts of the framework relevant to this project are: workspaces, algorithms and curve fitting.

2.2.1 Workspaces

Workspaces are the storage type for all data within MANTID. There are several implementations for various types of data. The two that are used within the VESUVIO workflow are MatrixWorkspace and TableWorkspace.

MatrixWorkspaces are used for storing three dimensional data, in the X, Y and spectrum axis as well as error values for each X value. This is the standard format for data recorded on most instruments and typically remains in this format throughout the analysis.

MatrixWorkspaces can store data as either point data, in which the number of X values is equal to the number of Y values or more commonly as a histogram where there is one additional X value. It is a reasonable assumption that data derived from experimental data will be in histogram format, there are several exceptions to this however none of them apply to either the data generated by VESUVIO or the type of operations performed in the data analysis workflow.

TableWorkspaces are a simple two dimensional table that can store a range of data types (as opposed to MatrixWorkspaces which can only store double precision floating points). This workspace type is used to contain the final parameters after a fit algorithm has been executed.

2.2.2 Algorithms

All operations within MANTID are performed using operations known as Algorithms, such operations can have a range of input, output and input/output properties of various types (including workspaces).

It is possible to use algorithms inside other algorithms to form a workflow, such algorithms are known as Data Processing Algorithms. Many of the algorithms directly used in the VESUVIO workflow fall into this category and have been written specifically for this data analysis task.

2.2.3 Curve Fitting

MANTID provides an extensive curve fitting library that was used throughout this project, this is split into several key aspects:

Fitting Algorithms

Fitting algorithms are MANTID algorithms that are responsible for running the fitting, there are several implementations for specific situations, however the one that is used in almost all of the analysis workflows for VESUVIO is the generic `Fit` algorithm.

Functions

Functions are implementations of a model that can be fitted using the fitting framework, a function may declare a number of parameters (which are optimised as part of the fitting procedure) and attributes (which retain the same value throughout the fitting, similar to a property of an algorithm).

Minimizers

The minimizer is responsible for optimising the cost function by iteratively changing parameters.

The most commonly used implementation within MANTID is the Levenberg-Marquardt [3] algorithm. A recent alternative is the Fitting Algorithm for Bayesian Analysis of DAta (FABADA) [11] algorithm which was used extensively in the analysis routine implemented in section 4.

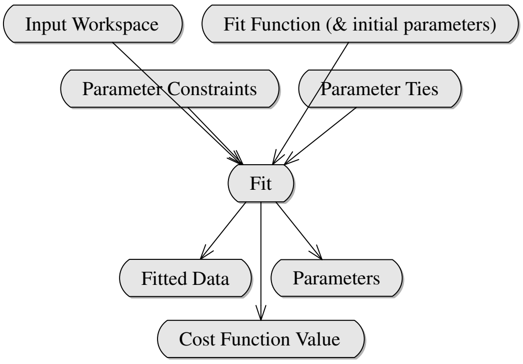

Figure 2: MANTID curve fitting

2.3 Existing data analysis workflow

The current data analysis workflow for VESUVIO has recently been ported to the MANTID data analysis toolkit as a collection of Python scripts after having been originally implemented in several routines in VMS FORTRAN.

A brief overview of the workflow is given in figure 3.

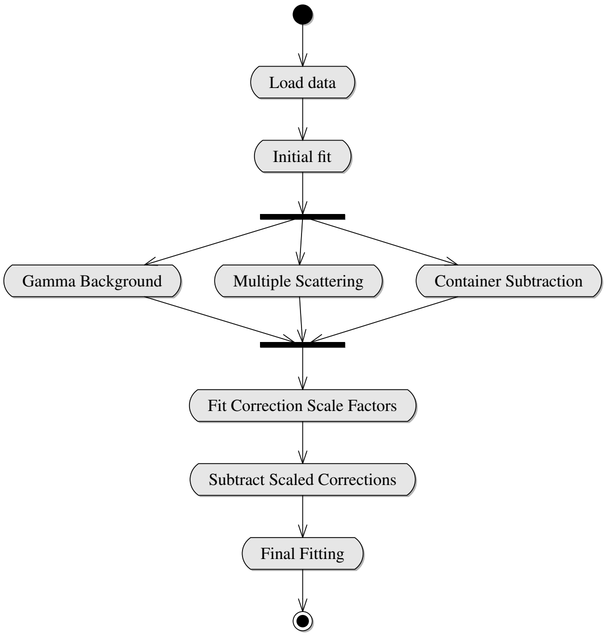

Figure 3: Existing data analysis workflow [10]

The workflow is initiated using a ”driver script” that allows the user to configure a given set of options for each stage of the workflow. An example of such a script is given in appendix A.

Initially all data is loaded, cropped and grouped to the user’s selection of spectra or detector banks and optionally rebinned, this is performed once per invocation of the workflow.

Each of the following stages are performed once per spectrum in the loaded data, a spectrum can either be representative of a single detector (in `spectra` mode) or a collection of detectors in a bank (in `bank`) mode.

An initial fit is first performed using the fitting model defined in the driver script, this fit is required to obtain a set of fitted parameters that are then used for the multiple scattering and gamma background corrections. Only the fitted parameters are produced form this fit.

A series of corrections can then be performed, specifics of each correction and whether it is calculated at all can be configured from the driver script.

The gamma background correction is applied only to the YAP detectors in forward scattering that are sensitive to gamma radiation, the correction aims to remove any spurious signal in the spectrum caused by gamma radiation emitted when stray neutrons collide with parts of the instrument and sample environment equipment.

Ideally the measured spectrum would only contain counts from neutrons that have been scattered at most once, this is to ensure a good quality measurement of the momentum distribution. It is possible for neutrons to be scattered multiple times inside the sample before exiting and hitting a detector. The multiple scattering correction aims to attenuate any signal that has been added as a result of neutrons that have undergone multiple scattering.

The container subtraction correction is a simple correction which involves subtracting a spectrum taken from the container without the sample material present from the spectrum with the sample present. This is a common correction and is very computationally cheap to apply.

Each correction generates a correction spectrum which are then linearly fitted to obtain a scale factor for each correction such that when scaled they best fit the data. Certain corrections can optionally have this scale factor fixed in the event that this fitting gives a bad scale factor.

The corrections are then scaled by the appropriate factor and subtracted from the input spectrum.

The final step is to perform the fitting once again, this time using the corrected spectrum and also outputting the fitted data.

Although not indicated in figure 3 there is an option to perform several iterations of this workflow, where the fitted parameters at the end of one iteration are used as the starting parameters of the next iteration. This can be continued until either the cost function converges on a predefined value or maximum number of iterations has been performed.

3 Multivariate Gaussian Fitting

The first objective concerns the addition of a new fitting model to be used in the data analysis workflow that uses a multivariate Gaussian model to fit an anisotropic mass peak, i.e. where the momentum of the stuck nucleus is not uniform in all directions, this is most commonly observed in experiments involving hydrogenous samples.

Previously two fitting models were used to model a mass peak on a VESUVIO spectrum, either the Gram-Charlier profile in the case where the mass has an anisotropic distribution (a good example of this is the mass peak generated by Hydrogen) or a simple Gaussian approximation in all other cases.

The Gram-Charlier profile has the advantage that is is physically speaking very good at fitting the peak and as it is an analytical method reasonably fast. However one drawback of this method is that it can be difficult to relate to theoretical data, most prominently data obtained through ab initio simulations.

For use in such situations a multivariate Gaussian model can be used in order to describe the peak through fitted parameters that are more easily related to simulated data, in the form of three Gaussian standard deviations: σ_x, σ_y and σ_z.

3.1 Background

The multivariate Gaussian model descries the momentum distribution of the mass peak as the product of three separate Gaussian functions where the fitted standard deviation of each individual function (σ_α) is representative of the momentum along each of the three Cartesian axes.

These momentum values can then be converted to kinetic energy using equation 7.

$$\langle E_K \rangle_\alpha = \frac{\hbar^2 \sigma_\alpha^2}{2M} \tag{7}$$

where; M is the atomic mass corresponding to the peak and α is an axis $\alpha \in (x, y, z)$.

The expression of the model function is given by Romanelli [12] where the multivariate Gaussian profile is described by equation 8.

$$J(y) = \frac{1}{\sqrt{2\pi}\sigma_x\sigma_y\sigma_z} \frac{2}{\pi} \int_0^1 d(cos\theta) \int_0^{\frac{\pi}{2}} d\phi S^2(\theta,\phi) exp\left(-\frac{y^2}{2S^2(\theta,\phi)}\right) \tag{8}$$

where; y is the y-space converted intensity and S^2 is given by equation 9.

$$\frac{1}{S^2(\theta,\phi)} = \frac{sin^2\theta cos^2\phi}{\sigma_x^2} + \frac{sin^2\theta sin^2\phi}{\sigma_y^2} + \frac{cos^2\theta}{\sigma_z^2} \tag{9}$$

This function defines the model in the ideal case where the momentum transfer is consistently high enough for the impulse approximation to hold, however as this can not always be guaranteed a proven correction for these so-called FSE must be applied.

A method of correcting for these effects is described by Sears [14] which describe the correction as the summation of a series of corrections in powers of $\frac{1}{q}$, as demonstrated in equation 10.

$$J(y,q) = J(y) + \sum_{n=3}^{\infty} (-1)^n A_n(q) \frac{d^n}{dy^n} J(y) \tag{10}$$

where $A_n(q)$ is given by equations 11 and 12 in the case of $n = 3$ and $n = 4$ respectively [12].

$$A_3 = \frac{\bar{\sigma}^4}{3q} \tag{11}$$

$$A_4 = \frac{\bar{\sigma}^6}{6q^2} \tag{12}$$

Typically only the A_3 and A_4 corrections are considered as beyond this the magnitude of the additive correction becomes dwarfed by the magnitude of the error bars of the sample data.

In the case of the multivariate Gaussian function it was decided through conversation with the VESUVIO instrument scientist that only the A_3 term should be considered for this function, this is due to a similar reason of the magnitude of further corrections becoming insignificant as well as a desire to reduce the computational cost of executing the function, this point will be elaborated on later.

Given that only the A_3 term is being considered equation 10 can be simplified to

$$J(y,q) = J(y) + -A_3(q)\frac{d^3}{dy^3}J(y) \tag{13}$$

The correction is then described by equation 14, the derivation of which is discussed by Romanelli [12].

$$-A_3(q)\frac{d^3}{dy^3}J(y) = \frac{\sigma_x^4 + \sigma_x^4 + \sigma_x^4}{9\sqrt{2\pi}\sigma_x\sigma_y\sigma_z q}\int_0^1 d(cos\theta)\int_0^{\frac{\pi}{2}} d\phi \\ \left[\frac{y^3}{S^2(\theta,\phi)^4} - 3\frac{y}{S^2(\theta,\phi)^2}\right] S^2(\theta,\phi)exp\left(-\frac{y^2}{2S^2(\theta,\phi)}\right) \tag{14}$$

Provisioning for both the FSE correction and a fitted intensity scaling parameter I the expression to be fitted by the new fit function is given by equation 15.

$$y' = I\left(J(y) + \left[-A_3(q)\frac{d^3}{dy^3}J(y)\right]\right) \tag{15}$$

where; y is the y-space transformation of the raw time of flight data and y' is the result of the evaluation with the current set of parameters.

3.2 Implementation

The function was implemented within the existing fitting framework within MANTID, this provides much of the core fitting functionality requiring only the evaluation of the function in y-space to be implemented.

Fit functions in MANTID are implemented as classes which inherit from a given abstract class depending on the nature of the function, for example the dimensionality of the data.

Within MANTID there is an abstract fit function already implemented that is designed to be inherited by fit functions that are designed to be used to fit models for neutron Compton scattering; `ComptonProfile`. This function subclasses the `ParamFunction` and `IFunction1D`, for dealing with fitted parameters and a function that is dependant on a single real value.

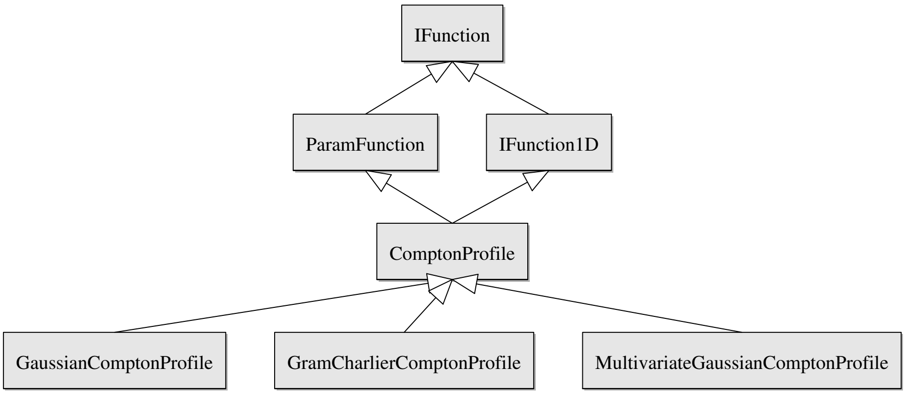

Figure 4: Structure of Compton profile fit functions

Fit functions implementing `ComptonProfile` are also able to be used in the `ComptonScatteringCountRate` composite function, this is a specialisation of the existing *CompositeFunction* which also handles the constraint matrix used for setting intensity constraints from the workflow script.

The `ComptonProfile` function only requires that child functions implement the evaluation of the mass peak function, in the case of the multivariate Gaussian this is given by equation 15. Data is provided having already undergone the y-space transform.

A cache of S^2 values are maintained in the class of the fit function, as these values are dependant on σ_x, σ_y and σ_z this cache must be rebuilt at the start of each fitting iteration and are calculated by a simple iteration over the integration domain.

The integration scheme employed in the evaluation is two dimensional Simpson's rule, whereby an integration over a two dimensional domain Ω can be expressed as

$$\int_\Omega f(x,y)\,d\Omega = \int_a^b \int_c^d f(x,y)\,dx\,dy \approx \frac{1}{9}nm \sum_{i=1}^{n} \sum_{j=1}^{m} A_{n,m} f(x_i, y_j) \tag{16}$$

where n and m are the number of steps in the x and y domains respectively, both of which must be an even number and $A_{n,m}$ is a matrix of identical dimensions to the domain Ω of the repeating format

$$A = \begin{bmatrix} 1 & 4 & 2 & \dots & 4 & 1 \\ 4 & 16 & 8 & \dots & 16 & 4 \\ 2 & 8 & 4 & \dots & 8 & 2 \\ 4 & 16 & 8 & \dots & 16 & 4 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 1 & 4 & 2 & \dots & 4 & 1 \end{bmatrix}$$

The calculation of the neutron Compton profile $J(y)$ and the FSE correction have been performed in two separate integration operations, initially this was to aid in testing by providing the option to output the FSE correction separately from the profile. However this could be modified to only require one integration operation per data bin.

As the integration is performed once per data bin per iteration of the fitting algorithm it is desirable to reduce the size of the integration domain as much as possible without too much reduction in the quality of the fit, as the theoretical accuracy of the multivariate Gaussian model is dependant on the accuracy of the integration. Romanelli recommends a domain of at least 35x35 [12], this is covered further in section 3.3.2.

3.3 Testing

3.3.1 Effects of FSE corrections

The function was used to fit a Hydrogen peak with and without the FSE corrections to ensure both that the neutron Compton profile and FSE correction were working as intended as it is normal for the peak to be slightly offset without the correction applied.

Figure 5 shows the difference between the model along (red line) and the model with the FSE corrections applied (green line) along with the original data (black line).

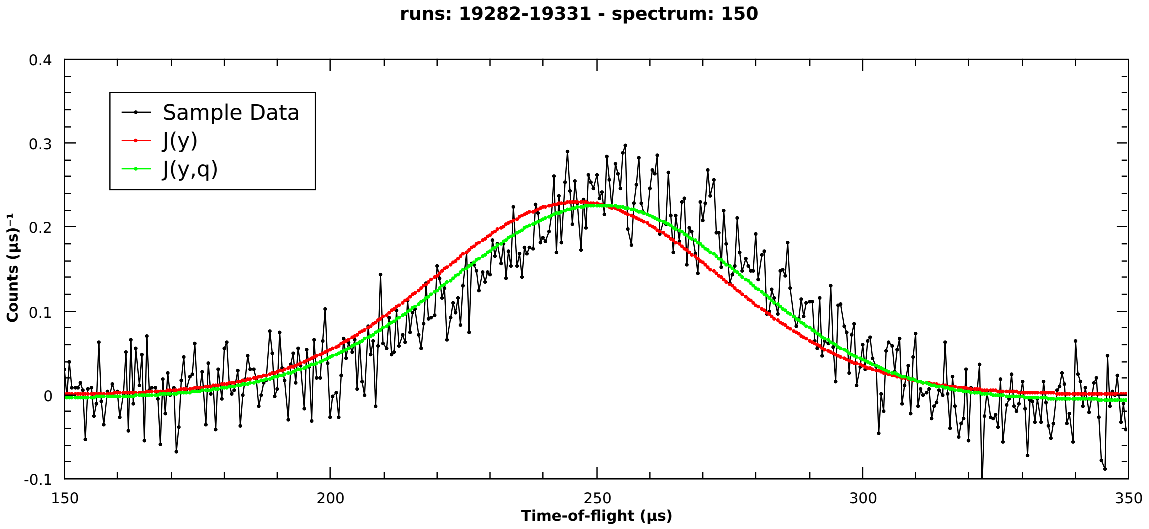

Figure 5: Comparison of fit with and without FSE correction

This shows a noticeable offset caused by the addition of the correction which shifts the fitted peak towards the centre of the experimentally measured peak. The centre of the fitted peak is still slightly towards the low side of the experimental peak however this offset is reasonable for the quality of the data and further corrections would only make the function slower to evaluate for only marginal improvement.

3.3.2 Effects of integration domain size

The effects of the size of the integration domain are briefly discussed by Romanelli in which it is summarised that the size of the domain must be at least 35x35, here several additional tests were performed in which the accuracy of the function relative to the Gram-Charlier profile and execution times are compared.

The final fitted parameters and execution times for both stages of model fitting are summarised in table 1, only the parameters relevant to the multivariate Gaussian function are listed (note that the mass parameter is fixed to 1.0079 for the Hydrogen peak).

Domain dimensions	8x8	32x32	64x4	256x256
SigmaX	2.49073	5.06038	8.25955	12.2212
SigmaY	11.7155	5.06598	2.51403	2.04771
SigmaZ	3.1926	4.3432	4.20634	4.16945
Intensity	1.20557	0.937984	0.945023	1.18128
Cost function	1.10373	1.20871	1.20523	1.07845
Initial fit time (s)	3.8	49.3	151.4	2456.4
Final fit time (s)	6.2	34.6	152	2591

Table 1: Comparison of final fitted parameters and execution times for different integration domain dimensions

Figure 6 shows a comparison between the Gram-Charlier profile (black line) and the four different integration domain dimensions: 8x8 (cyan line), 32x32 (green line), 64x64 (blue line) and 256x256 (red line).

This plot shows the majority of the fitted time of flight range used in VESUVIO model fitting, this emphasises the changing shape of the wings of the peak at $< 200\mu$s and $> 300\mu$s.

Note that the peak at approx. 370μs is from a combination of Carbon and Oxygen, both of which are fitted using the standard isotropic Gaussian hence will have the same peak shape in all fitted spectra.

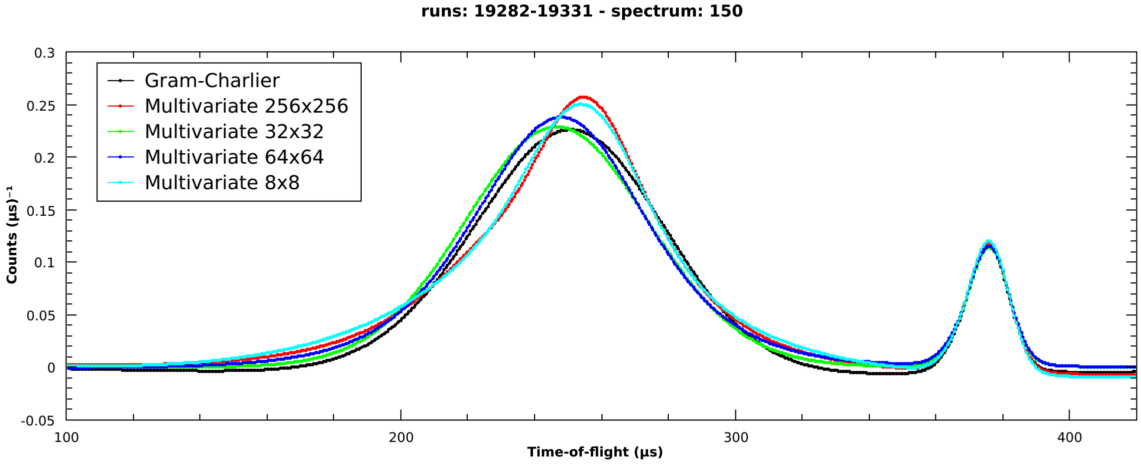

Figure 6: General comparison of different integration domain dimensions

Figure 7 shows the same fitted spectra as figure 6 with the plot area focused around the Hydrogen peak. This provides a good view of the quality of each fit in terms of its peak centre and shape with respect to the Gram-Charlier profile.

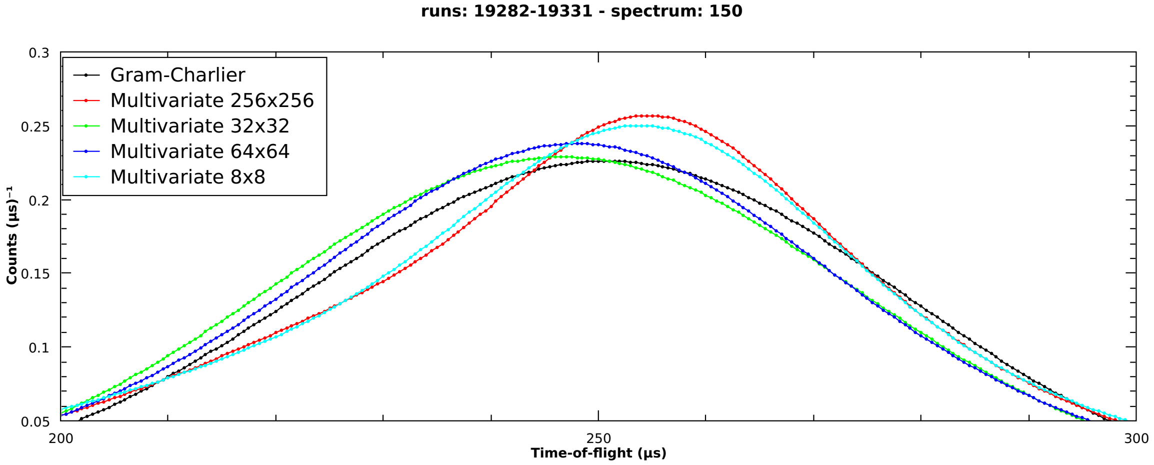

Figure 7: Comparison of different integration domain dimensions at Hydrogen peak

Given the lower cost function value is assigned to the domain of size 256x256 this fit is likely to be the best description of the data, however looking at the plot of each fitted spectrum around the Hydrogen peak in figure 7 it can be seen that the 256x256 domain (red line) is one of the furthest from the Gram-Charlier profile (black line). Ideally the multivariate Gaussian profile should be a close math to the Gram-Charlier profile. Having said this the final parameters produced by the fit using the 256x256 domain are somewhat reasonable given the data.

Furthermore the computational cost of this as reflected by the execution time for both fitting stages makes this improvement in fit quality infeasible for general use.

Comparatively the 8x8 domain which is expected to have a poor fit quality is fairly close to the much finder 256x256 sized domain.

The domains of size 32x32 and 64x64 both produce a similar and reasonable fit, however the parameters produced by the 64x64 domain are the best description of the data out of all that have been fitted as part of this test.

As such this was set as the default domain size and is the size used for the evaluation performed in section 5.1.

3.3.3 Automated Testing

The multivariate Gaussian fitting model is tested to the same extent as all other fitting models for VESUVIO data via a set of unit and system tests that are run as part of the MANTID continuous integration system.

These tests are mainly designed as simple logical tests to ensure that the fit function and associated profile in the workflow scripts behave as expected in terms of the workflow rather than the quality of the fitted data its self.

That being said there is a basic test that the function, when evaluated with a given set of parameters, gives the correct values. As well as a function that tests the generation of the S^2 matrix.

3.3.4 Informal Testing

Several informal tests were carried out as part of the development process; this included tests done by myself, the instrument scientist and the developer that approved the changes to MANTID.

This testing is a good source of a second opinion on the implemented function both from the scientific and software engineering perspectives.

Having the model tested by the scientist did highlight an issue that had previously gone unnoticed due to not being directly related to the work being carried out. As mentioned in section 2.3 an initial fit is performed to obtain a set of parameters to be used in the calculation of several corrections, one of which - the multiple scattering correction - when using parameters form the multivariate Gaussian profile was giving a largely underestimated correction for the area around the Hydrogen peak the function was fitting where there is known to be a large multiple scattering contribution to the measured signal.

Figure 8 shows the difference is multiple scattering correction caused by the differing parameters which are in turn provided in table 2.

In this table the function fitting the Hydrogen peak is the first function in the composite function `f0`. The values given for the width of the multivariate Gaussian function are obtained using the following equation which is used in the workflow script to convert the `Sigma[X,Y,Z]` parameters into a single width parameter $\bar{\sigma}$ for the multiple scattering correction algorithm.

$$\bar{\sigma} = \sqrt{\frac{\sigma_x^2 + \sigma_y^2 + \sigma_z^2}{3}} \tag{17}$$

	Gram-Charlier	Error	Multivariate	Error
f0.Mass	1.0079	0	1.0079	0
f0.Width	4.48946	0.122974	8.19068	-
f0.SigmaX	-	-	2.96842	1.53772
f0.SigmaY	-	-	13.4967	0.664743
f0.SigmaZ	-	-	3.20772	1.18797
f0.Intensity	69.1487	1.76213	0.967407	0.216778
f1.Mass	12.011	0	12.011	0
f1.Width	10	0	10	0
f1.Intensity	0.170138	1.07746	1.41289	1.29345
f2.Mass	15.9	0	15.9	0
f2.Width	13	0	13	0
f2.Intensity	5.78972	1.05313	5.53164	1.21508
f3.A0	0.0086126	0.00491099	0.0200821	0.0047604
f3.A1	-116.651	40.4553	-274.6	45.23
f3.A2	201.274	65400.6	462892	76824.8
Cost Function	1.0404	0	1.21149	0

Table 2: Fitted parameters used for multiple scattering calculation

As shown in table 2 the calculated with of the multivariate Gaussian profile is almost double that of the Gram-Charlier profile, however it is much more likely that the significantly lower intensity is the cause of the difference in multiple scattering calculation.

The reduced intensity is effectively removing the relative contribution of the Hydrogen peak to the multiple scattering in the sample, hence why the contributions of other masses are increased beyond their contributions when using the parameters from the Gram-Charlier profile.

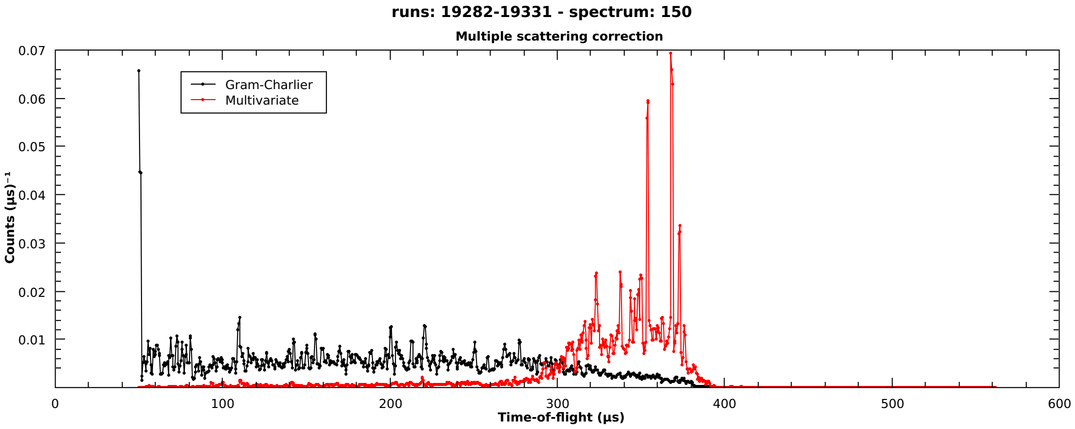

Figure 8: Comparison of multiple scattering corrections

Although not demonstrated the same effects are seen when comparing with the standard Gaussian profile.

4 Bayesian model selection

The second objective is to develop and integrate a model selection algorithm that given a set of raw time of flight data from the instrument can obtain a fitting model that best describes the data, this will be used to aid analysis in cases where the sample contains unexpected impurities that contribute to additional mass peaks or when the composition of the sample is unknown.

4.1 Background

The model selection algorithm will use Bayesian model selection to select the most likely model form a set of hypothesis, this has been chosen due to the wide use of this methodology in various other neutron scattering data analysis workflows.

Probability theory gives the two basic rules: the sum rule; given the probability of x being true we can derive the probability of x being false and the product rule; given the probability of x being true and the probability of y being true we can derive the probability of both x and y being true, given in equations 18 and 19 respectively.

$$P(x|I) + P(\bar{x}|I) = 1 \tag{18}$$

$$P(x, y|I) = P(x|y, I) \times P(y|I) \tag{19}$$

where; | denotes a ”given that” relationship between two events, $\bar{x}$ denotes the case where x is false and I is background information that could affect the probability of either x or y.

Given experimental data the measurable probability is that of the data given a hypothesis $P(D|H)$, however in data analysis the desirable probability to measure is the probability of a hypothesis given the experimental data $P(H|D)$.

$$P(x|y, I) = \frac{P(y|x, I) \times P(x|I)}{P(y|I)} \tag{20}$$

Using Bayes’ theorem given in equation 20 it is possible to effectively swap the dependency order of the given that relation between the hypothesis and data. [15]

Take for example two unique hypotheses A and B, in order to determine which best describes data D the ratio of their probabilities given the data can be calculated as

$$posterior = prior \times Bayes\,factor \tag{21}$$

$$\frac{P(A|D)}{P(B|D)} = \frac{P(A)}{P(B)} \times \frac{P(D|A)}{P(D|B)} \tag{22}$$

where the prior ratio is used to define the probability ratio of the two hypotheses based on prior knowledge, i.e. disregarding the data D.

This model works when both A and B are simple hypotheses that do not rely on fitted parameters, however assuming hypothesis B required an additional parameter to define the model the value of this parameter must be considered in the calculation of the probability ratio.

$$posterior = prior \times Bayes\,factor \times Ockhamfactor \tag{23}$$

$$\frac{P(A|D)}{P(B|D)} = \frac{P(A)}{P(B)} \times \frac{P(D|A)}{P(D|\lambda_0, B)} \times \frac{\lambda_{max} - \lambda_{min}}{\delta\lambda} \tag{24}$$

Equation 23 [15] shows how a fitted parameter λ can be integrated into the model selection. In this case the so called ”Ockham factor” is used to penalise hypothesis B for its additional fitted parameter, this follows the general desire for the best model to be the one which best describes the data with the fewest parameters.

Note that the parameters limits λ_{min} and λ_{max} must be chosen appropriately, an obvious general case may be to set them to infinity however this causes an infinite penalty to be applied for the addition of a parameter. Typically enough is known about the domain of the problem to allow a sensible selection of these limits.

Most curve fitting in MANTID uses the χ^2 cost function as defined by equation 25. Typically the χ^2 is divided by the number of free fitting parameters.

$$\chi^2 = \sum_{i=1}^{N} \left(\frac{D_i - f(D_i, M)}{\sigma_i} \right)^2 \tag{25}$$

where; N is the number of data points, D_i is the ith measured/sample data point, $f()$ is a function defining the model being fitted, M are parameters defining the model and σ_i is the standard error of the ith data point.

One issue with the χ^2 cost function is that it has a large dependence on the shape of the χ^2 distribution [9], for example if no local minima are found when searching for the global minimum.

The Fitting Algorithm for Bayesian Analysis of DAta (FABADA) provides a fitting methodology that does not depend on the shape of the cost function or relations between fitting parameters. This algorithm provides a probability distribution function for all fitted parameters as well as the cost function, which through inspection of can be used to create a Bayesian model selection scheme.

The FABADA algorithm is a Markov Chain Monte Carlo algorithm that operates by randomly changing a fitting parameter and accepting the change if it causes a decrease in the value of the cost function. There is also a chance that an increase in the cost function value will be allowed, the probability of which is given in the equation [9]

$$\frac{P(H(P_i^{new})|D)}{P(H(P_i^{old})|D)} = exp\left(-\frac{\chi^2_{new} - \chi^2_{old}}{2} \right) \tag{26}$$

where; $P(H|D)$ is the probability of the hypothesis being true given the measured data, $H(P_i)$ is the hypothesis described by parameter set P_i.

It is this property that allows the FABADA algorithm to "jump" over obstacles in the shape of the cost function that may cause traditional Levenberg-Marquardt methods to converge to a non-optimal fit.

One issue of this method is the number of steps in parameter value required to converge to the optimal solution, to address this the algorithm adds a bias to the changes applied to each parameter such that parameters who's changes are infrequently accepted are assigned larger steps in an attempt to invoke a larger step in the cost function. The opposite of this process happens for parameters with frequently accepted changes.

Fitting is complete when all parameters have converged, this is defined as the point at which the probabilities of each parameter have been accepted with equal probability.

4.2 Implementation

The model selection algorithm will be implemented as a new workflow algorithm within MANTID leveraging already implemented peak finding and fitting routines.

In this implementation the "model" is described as a composition of fitting functions that describe the data, typically this is one Compton profile per mass (where multiple masses can contribute to the same peak) and one background function which is usually a polynomial of order 2.

In MANTID a model is composed of:

- A fit function string which defines the functions being fit and their initial parameters
- A ties string which defines values which certain parameters are fixed to, this can be either a fixed value of another parameter.
- A constraints string which defines upper and/or lower constraints for he value of certain parameters

An example of a typical model for the VESUVIO data analysis workflow for a sample containing Hydrogen, Oxygen and Caesium in an Aluminium container is as follows:

Function

```
composite=CompositeFunction,NumDeriv=1;
name=GramCharlierComptonProfile,Mass=1.007900,HermiteCoeffs=1 0 0,Width=4.480264,
FSECoeff=0.528532,C_0=12.239281;
name=GaussianComptonProfile,Mass=16.000000,Width=10.000000,Intensity=2.829303;
name=GaussianComptonProfile,Mass=27.000000,Width=13.000000,Intensity=0.392174;
name=GaussianComptonProfile,Mass=133.000000,Width=30.000000,Intensity=0.707326;
name=Polynomial,n=2,A0=-0.003896,A1=5.387158,A2=1.003049
```

Ties

```
f0.Mass=1.007900,
f0.FSECoeff=f0.Width*sqrt(2)/12,
f1.Mass=16.000000,
f1.Width=10.000000,
f2.Mass=27.000000,
f2.Width=13.000000,
f3.Mass=133.000000,
f3.Width=30.000000
```

Constraints

```
2.000000 < f0.Width < 7.000000,
f0.C_0 > 0.0,
f1.Intensity > 0.0,
f2.Intensity > 0.0,
f3.Intensity > 0.0
```

The fitted results of this model are shown in figure 9, the Hydrogen peak can be seen around the $200\mu s$ to $300\mu s$ time of flight region and the single peak for all higher masses around $380\mu s$ to $400\mu s$.

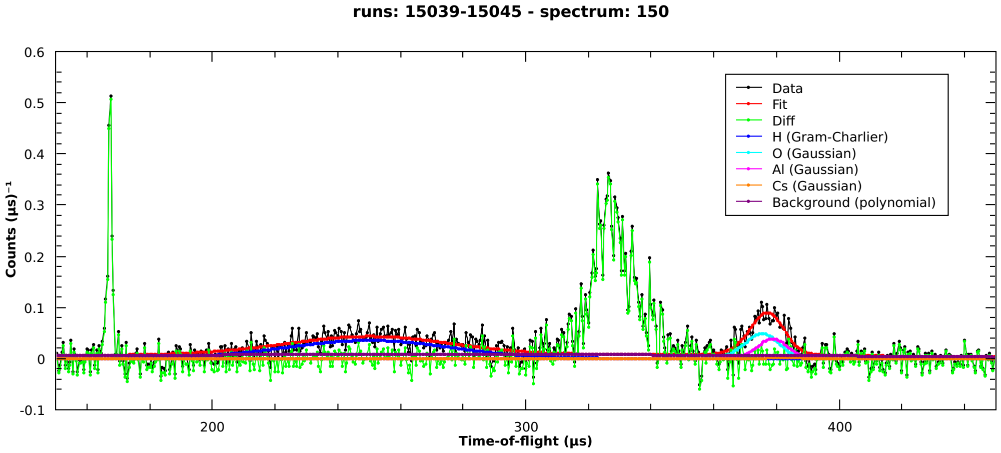

Figure 9: Example fitting model

The general workflow for the new algorithm is described by figure 10.

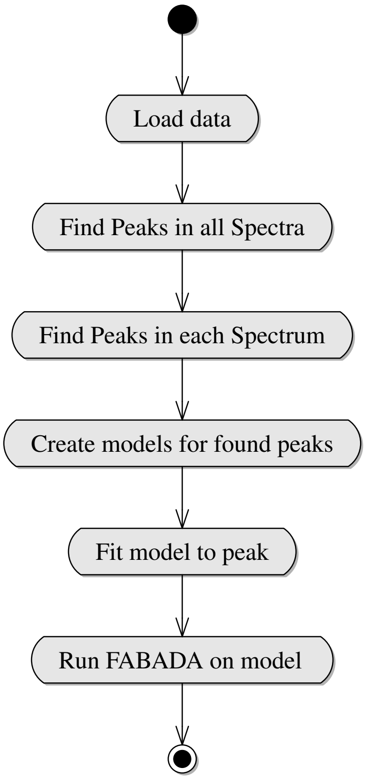
Load data
Find Peaks in all Spectra
Find Peaks in each Spectrum
Create models for found peaks
Fit model to peak
Run FABADA on model

Figure 10: Model selection algorithm workflow

In the case of this implementation a model as described above is a single hypothesis in the context of the Bayesian model selection.

The first step is to load the sample data, this uses the same implementation of the data loading step as is used in the existing VESUVIO workflow described in section 2.3.

The next step is to use the MANTID peak finding algorithm `FindPeaks` to search for peaks in all spectra of the loaded data, this provides an initial guess as to the locations of mass peaks in the measured data. This peak finding algorithm operates by fitting a Gaussian peak with a background to the specified input range [5], the search can be refined by providing additional information about the position and nature of the peak however at this stage no such information is provided.

This list of found peaks is then used as a peak position estimate in order to find a reasonable accurate peak position and shape for the relevant peak on each spectrum.

After this operation a list of reasonably accurate peaks for each spectrum will have been generated, these are then used to generate a list of models for the Bayesian model selection. This is done by assuming that each detected peak can have between 0 and 4 masses contribute to it (this is a fixed value at the moment but should be exposed to the user as a property for the algorithm) then creating a model with each permutation of number of masses for each peak.

Take, for example the following peaks having been found in a spectrum:

Position	Width
250μs	15
370μs	8

Table 3: Example found peaks

The following 25 models would be created:

Model	Peaks under $250\mu s$	Peaks under $370\mu s$
0	0	0
1	1	0
2	2	0
3	3	0
4	4	0
5	0	1
6	1	1
7	2	1
8	3	1
9	4	1
10	0	2
11	1	2
12	2	2
13	3	2
14	4	2
15	0	3
16	1	3
17	2	3
18	3	3
19	4	3
20	0	4
21	1	4
22	2	4
23	3	4
24	4	4

Table 4: Example models

In the models generated by the algorithm only the standard isotropic Gaussian function is used and has its mass parameter constrained based on the width of the peak found in the peak finding stage.

The limits of the mass constraint are obtained by taking the limits in time of flight which are calculated by taking the peak position plus/minus the Half Width Half Maxima (HWHM), these limits in time of flight are then converted into limits in atomic mass using equations 27 and 28.

$$tof = \frac{L_0 r_t + L_1}{v_0} \tag{27}$$

where; L_0 is the length of the primary flight path (form moderator to sample), L_1 is the length of the secondary flight path (from sample to detector), v_0 is the final neutron velocity and r_t is given by equation 28.

$$r_t = \frac{cos\theta + \sqrt{\frac{m}{m_n}^2 - sin^2\theta}}{\frac{m}{m_n} + 1} \tag{28}$$

where; θ is the scattering angle, m is the mass of the atomic mass of the stuck nucleus and m_n is the atomic mass of the neutron.

The following fitting is then performed per generated model per spectrum in the sample data.

Additional constrains on the model include the requirement for both the intensity and width parameters to be positive non zero values and that masses fitted by each Gaussian function must be unique to prevent the same mass being fitted by multiple functions.

These models are then fitted to each spectrum using the standard Levenberg-Marquardt algorithm in order to obtain a reasonable set of starting parameters for the fit using the FABADA minimiser which is known to require good parameters when a fit is performed over a large number of degrees of freedom.

This is also the reason parameters are constrained whenever possible in order to narrow the search space as much as it can be.

Once a set of starting parameters is derived the model is then fitted using the FABADA minimiser to obtain a refined set of parameters and their probability, the probabilities of each model are then compared and the most likely selected as the best model.

4.3 Testing

4.3.1 Automated Testing

As the model selection algorithms is yet to be functional to the full satisfaction of the instrument scientist no automated tests have been written for it as of yet.

4.3.2 Informal Testing

Throughout the implementation of the algorithm several stages of informal testing were performed to asses the accuracy and reliability of independent stages of the algorithm (which were in turn typically separate algorithms already implemented in MANTID).

One of the main examples of this was in the output generated by the `FindPeaks` algorithm which was used to perform the initial significant peak finding to search for peaks that could correspond to a mass peak.

The most significant issue raised by the use of this algorithm is the wide range of results given by the algorithm when executed with different data sets; some of which give a reasonable estimation of the location of peaks, some assign peaks to noise while missing larger more visually obvious peaks and some fail to find any peaks.

Several examples of common problems are shown with four samples; Iodobenzoic acid ($C_7H_5IO_2$), Polycrystalline zirconium-beryllium ($Zr_{40}Be_{20}$), Squaric acid ($C_4H_2O_4$) and Graphite (C). All samples are in Aluminium containers with the exception of Graphite which is in a Tin container.

Figure 11 shows the raw time of flight spectrum for a sample of Iodobenzoic acid in an Aluminium container, here true peaks are observed at around $200\mu s$ to $300\mu s$ for the Hydrogen peak and around $350\mu s$ to $420\mu s$ for heavier masses.

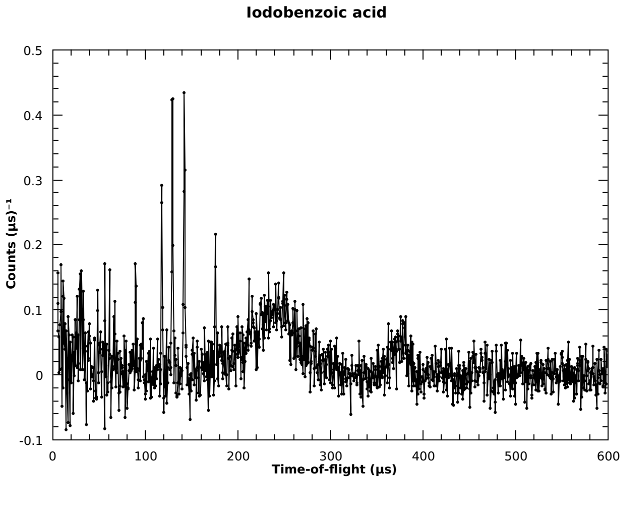

Figure 11: Raw spectrum of Iodobenzoic acid sample

Table 5 shows the results of the peak finding algorithm. In this case no true peaks were found by the algorithm, instead the two tallest peaks in the spectrum caused by noise were reported as the significant peaks in the sample.

Position (μs)	Width	Height
117.848	1.07304	0.304676
129.286	1.33195	0.453472

Table 5: Peak finding results for Iodobenzoic acid

The polycrystalline zirconium-beryllium sample (figure 12) is an example of where the results of peak finding contain both true and spurious peaks.

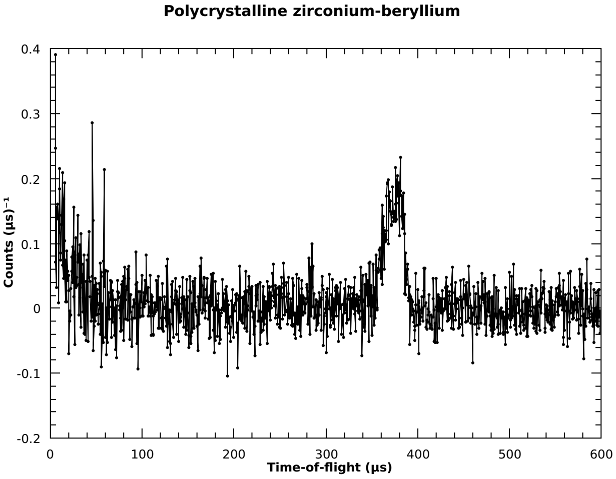

Figure 12: Raw spectrum of Polycrystalline zirconium-beryllium sample

In this case the peak caused by the masses known to be present in the sample is detected correctly (the peak at around $374\mu s$) however an additional peak has been found in noise at around $285\mu s$.

Position (μs)	Width	Height
285.225	1.63995	0.0685314
374.626	21.1158	0.129163

Table 6: Peak finding results for Polycrystalline zirconium-beryllium

In the case of the Squaric acid sample (figure 13) the issue is the lack of detection of a true peak, in this case the Hydrogen peak between $180\mu s$ and $310\mu s$. This is the most acceptable case of the peak finding algorithm failing to identify all peaks successfully and is partly to be expected by the significantly larger width of the Hydrogen peak.

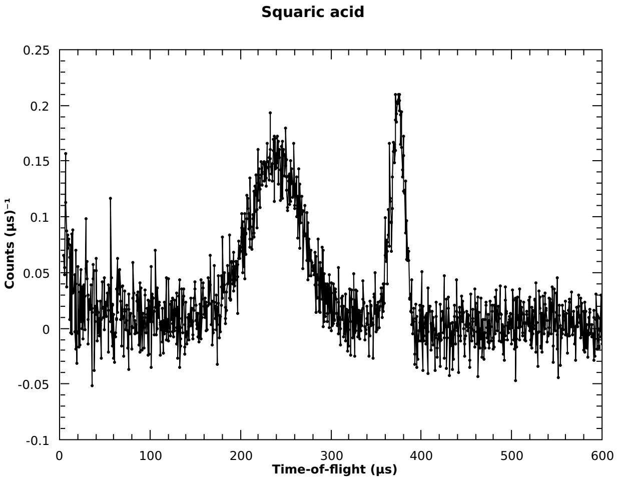

Figure 13: Raw spectrum of Squaric acid sample

The peak contributed to by heavier masses ($330\mu s$ to $400\mu s$) has been identified correctly in this case, as is shown in table 7.

Position (μs)	Width	Height
374.164	17.4555	0.191177

Table 7: Peak finding results for Squaric acid

The peak finding algorithm failed to find any significant peaks in the Graphite sample shown in figure 14. The two most likely causes of this is the relative intensity of the true peak to the background noise and the non-Gaussian distribution of the peak, both in the asymmetry and overly steep falling edge of the peak.

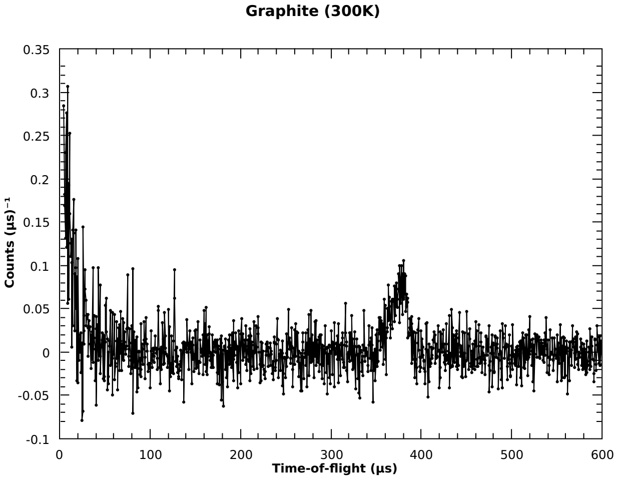

Figure 14: Raw spectrum of Graphite sample at 300K

Typically this issue can be resolved by changing the parameters to the `FindPeaks` algorithm, most notable the *FWHM* and *Tolerance* parameters.

The wide variation of peak widths can often mean fitting a single Gaussian width (determined by the `FWHM` parameter) can cause only a subset of true peaks to be found while possibly selecting false peaks in the noise, in the current peak finding solution this is a non-trivial problem to overcome without requiring a new peak finding implementation.

An additional problem is setting a reasonable tolerance for the required acceptance of peaks, this tolerance value is used in the implementation of the peak finding algorithm as described by Mariscotti [5] where a lower value gives rise to stricter peak selection.

Having this value incorrectly set can cause either true peaks to be missed if the value is too low or additional peaks being found in noise if the value is too high.

Another area that underwent ongoing manual testing was the quality of the models generated by an initial fit using the Levenberg-Marquardt algorithm. Typically this worked as expected where a model with a rough estimation of a peak position would have its mass fitted to be close to the mass that contributed to the majority of the peak.

However one significant issue with the initial fitting routine is that when fitting a model in which there are several profiles that are attempting to fit a single peak which is defined by less than the number of profiles (for instance attempting to fit 4 masses to a peak that was only contributed to by 2) the masses of multiple profiles tended to converge on the same mass.

This was a significant issue as it artificially increased the suitability of more complex models by obtaining parameters that give a value to the cost function χ^2 that is only marginally worse than that of a simpler model.

While the χ^2 of the initial fit is not inspected as the only purpose of this algorithm is to find a likely set of models this issue causes a reduction in the number of possible models due to lack of exploration of the full search space of the mass parameter, possibly excluding a correct mass from the Bayesian model selection that follows on from this initial fitting.

A solution to that above two issues can be to allow the user to modify the range of Gaussian widths and tolerances over which the `FindPeaks` algorithm is executed, however the need for the user to run the algorithm in order to look at the results and possibly have to change input parameters and execute the algorithm a further time until the results are reasonable somewhat negates the point of using this algorithm in the first place. Having said that the peak finding is a relatively computationally cheap operation compared to the full VESUVIO analysis routine so this process of manually modifying parameters to obtain good results at this stage is not as time consuming as doing the same with the full data analysis workflow.

5 Evaluation though Case Studies

Several case studies have been used to provide testing of the results of the new features implemented as part of this project.

This method of testing has the advantage that it is testing the routines in the same way in which they are intended to be used, this is not usually possible through unit testing and is due to build server time constraints usually limited to a small range of sample data when running as an automated system test.

5.1 Fitting Hydrogen peaks with a multivariate Gaussian

One of the most common use cases for a function that fits an anisotropic mass peak (i.e. the Gram-Charlier profile or multivariate Gaussian profile) is in fitting Hydrogen peaks.

These case studies will focus on the comparison of the existing Gram-Charlier profile and new multivariate Gaussian profile in the quality of the description of the sample data. The mean kinetic energy along each axis will also be calculated using equation 7 and compared with published data.

Two well studied hydrogenous samples will be used; water at 300K and ice at 260K.

5.1.1 Water

The fitted curves for for both the Gram-Charlier profile and multivariate Gaussian profile are given in figures 15 and 16 respectively.

The Gram-Charlier fit is visually a very good fit of the data, shown by the mass profile for the Hydrogen peak (blue line) closely following that of the sample data (black line).

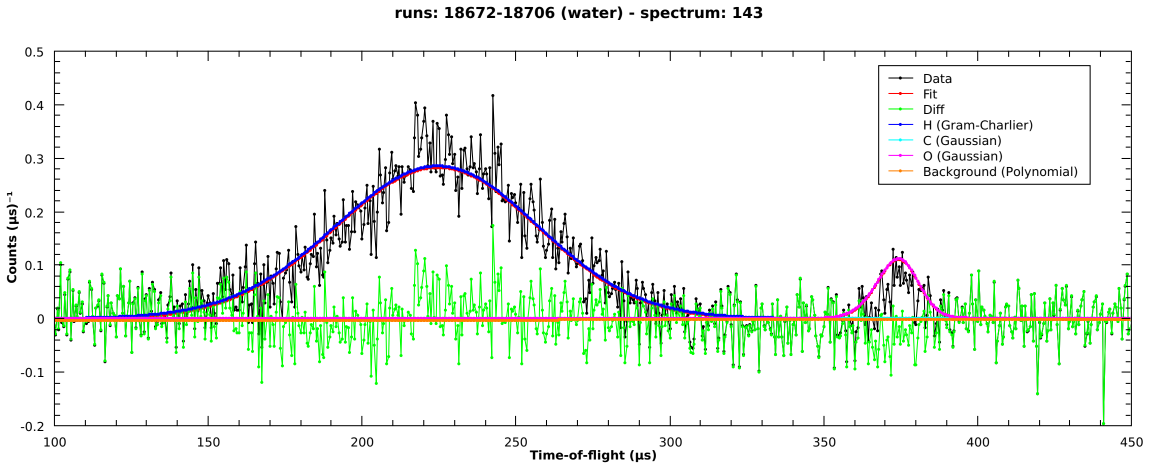

Figure 15: Fitting results for water sample using Gram-Charlier

In the case of the fit using the multivariate Gaussian the fit is of a noticeably lower quality. The peak centre can be seen to be shifted to the left slightly giving a "kink" to the greed difference curve, this can be indicative of an understated final state effects correction.

Another possibility is the lack of exploration of the full parameter space. This can be an issue with fit functions with a high degree of freedom using the Levenberg-Marquardt optimisation function in that a premature minima may be reached in the cost function which is not the global minimum.

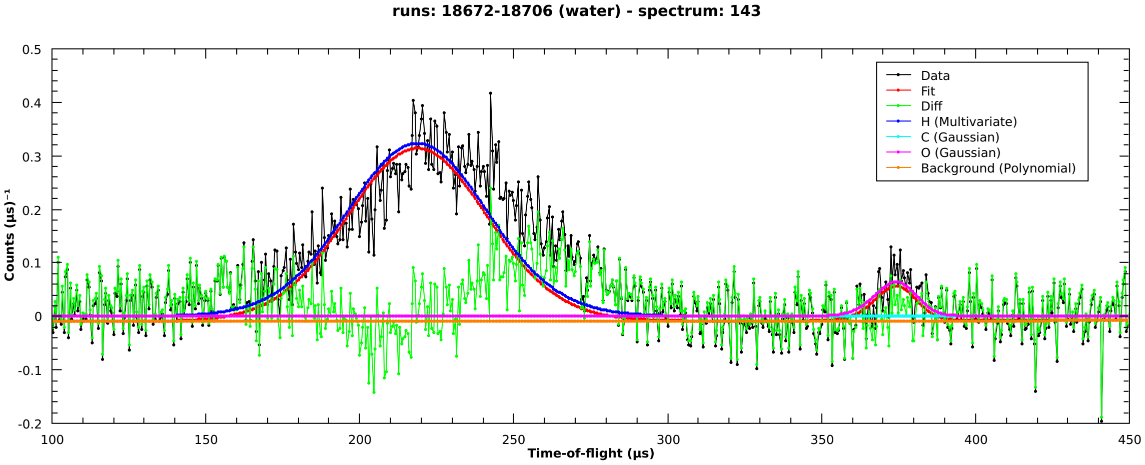

Figure 16: Fitting results for water sample using multivariate Gaussian

Parameters for both functions are given by tables 8 and 9 respectively.

Parameter	Value	Error
f0.Mass	1.0079	0
f0.Width	5	0.126964
f0.FSECoeff	0.5892556	0
f0.C_0	109.027	2.74958
f1.Mass	12.011	0
f1.Width	10	0
f1.Intensity	0	1.26891
f2.Mass	15.9	0
f2.Width	13	0
f2.Intensity	5.82717	1.24073
f3.A0	-0.004548	0.00589503
f3.A1	8.367988	48.387
f3.A2	1.004511	79719.6
Cost Function	1.149798	0

Table 8: Fitted parameters for water sample using Gram-Charlier

Note that in the case of the fit using the Gram-Charlier profile the parameter errors are relatively low given the value of each parameter, this is reflective of the good quality of the fit.

Parameter	Value	Error
f0.Mass	1.0079	0
f0.SigmaX	3.0069848	9.52070
f0.SigmaY	3.5111716	16.91488
f0.SigmaZ	4.1116870	3.414547
f0.Intensity	0.8536424	0.83102
f1.Mass	12.011	0
f1.Width	10	0
f1.Intensity	0.0216973	1.278684
f2.Mass	15.9	0
f2.Width	13	0
f2.Intensity	3.35907	1.24844
f3.A0	-0.01038	0.005909
f3.A1	4.224282	46.1621
f3.A2	2595.102	75832.87
Cost Function	1.68549	0

Table 9: Fitted parameters for water sample using multivariate Gaussian

In the case of the multivariate Gaussian the parameter errors are noticeably larger, particularly for the σ_α parameters of the multivariate Gaussian profile.

Table 10 shows a companion between the mean kinetic energies calculated for the sample analysed above and a similar water sample (at 285K) analysed by Romanelli.

Parameter	Multivariate Gaussian	Romanelli [12]
$\langle E_K \rangle_x$	46.25	18.3
$\langle E_K \rangle_y$	63.05	51.8
$\langle E_K \rangle_z$	86.46	83.8

Table 10: Comparison of mean kinetic energies (meV)

The results of my implementation give similar results to that of Romanelli's previous analysis, there are obvious deviations from is results however given the errors on the parameters produced by my results this is certainly expected.

5.1.2 Ice

Fitted curves for the same profile fitting data from a sample of ice are shown for the Gram-Charlier profile and multivariate Gaussian profile in figures 17 and 18 respectively.

As before the Gram-Charlier profile is a very good fit given the noise in the sample data. Some small fluctuations in the difference curve (shown in green) can be seen under the Hydrogen peak.

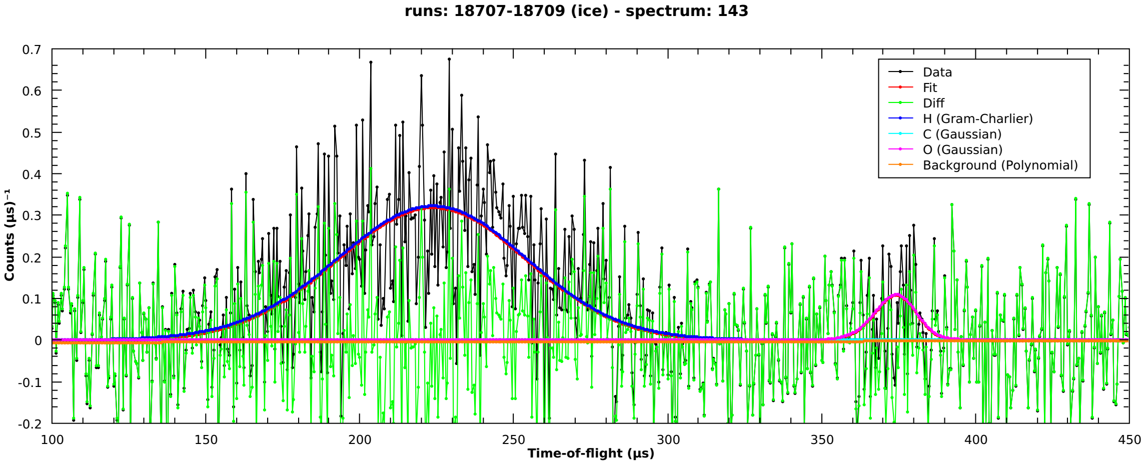

Figure 17: Fitting results for ice sample using Gram-Charlier

Similarly in the fit using the multivariate Gaussian profile, the Hydrogen peak is fitted reasonably well. There are two noticeable peaks in the difference curve at either side of the Hydrogen peak, this is due to convergence on bad fitting parameters.

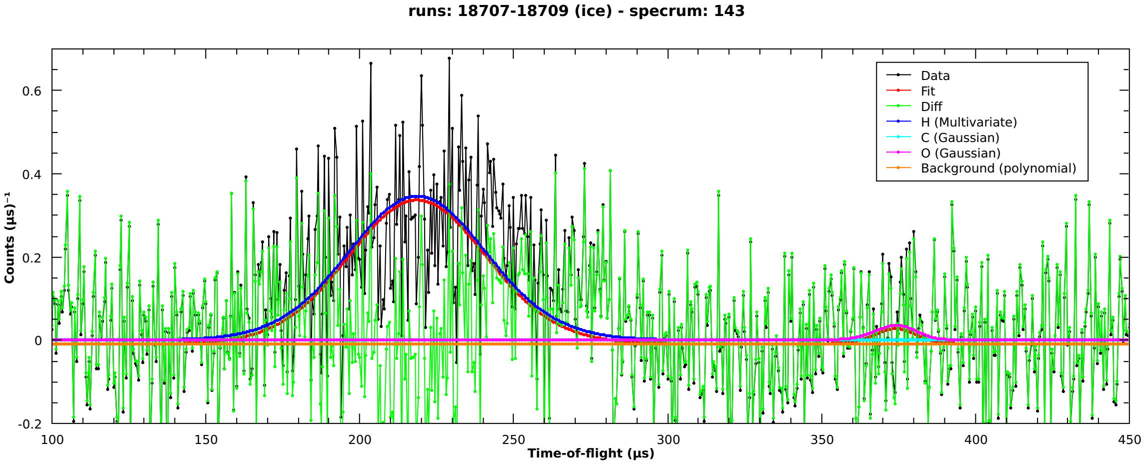

Figure 18: Fitting results for ice sample using multivariate Gaussian

The parameters for both fitting models are shown in tables 11, for the Gram-Charlier profile and 12 for the multivariate Gaussian.

Parameter	Value	Error
f0.Mass	1.0079	0
f0.Width	4.61202317	0.36917
f0.FSECoeff	0.543532	0
f0.C_0	115.3064	8.69815
f1.Mass	12.011	0
f1.Width	10	0
f1.Intensity	-7.05674e-07	0.0223604
f2.Mass	15.9	0
f2.Width	13	0
f2.Intensity	5.689151	1.56584
f3.A0	-0.0068479	0.0199176
f3.A1	10.8744496	160.853
f3.A2	1.00203	263695
Cost Function	1.0332	0

Table 11: Fitted parameters for ice sample using Gram-Charlier

As before the parameters and errors for the fit using the Gram-Charlier profile are reflective of a very good fit.

Parameter	Value	Error
f0.Mass	1.0079	0
f0.SigmaX	3.02910	101.9631
f0.SigmaY	3.18033	117.5588
f0.SigmaZ	4.17846	7.002974
f0.Intensity	0.873875	1.87307
f1.Mass	12.011	0
f1.Width	10	0
f1.Intensity	0.044426	4.503
f2.Mass	15.9	0
f2.Width	13	0
f2.Intensity	1.86952	4.373681
f3.A0	-0.010624	0.020005
f3.A1	3.79177	158.3326
f3.A2	2638.56	260539
Cost Function	1.0942	0

Table 12: Fitted parameters for ice sample using multivariate Gaussian

The parameters for the fit using the multivariate Gaussian show a large uncertainty in the σ_α parameters. This is expected given the observed peaks in the difference curve under the Hydrogen peak.

Table 10 shows a companion between the mean kinetic energies calculated for the ice sample analysed above and a similar sample of ice (at 271K) analysed by Romanelli.

Parameter	Multivariate Gaussian	Romanelli [12]
$\langle E_K \rangle_x$	46.93	28.9
$\langle E_K \rangle_y$	51.73	38.1
$\langle E_K \rangle_z$	89.29	86.7

Table 13: Comparison of mean kinetic energies for ice sample (meV)

As with the water sample my results for the kinetic energy of the ice sample deviate greatly from that of Romanelli, however the change in parameter values between the two samples does still follow the same trend (i.e. the decrease of $\langle E_K \rangle_y$).

5.2 Model Selection

The model selection algorithm has been tested against a series of known (and some only partially known) samples in order to provide a reliable indication as to its effectiveness.

These sample have been summarised in table 14.

	Runs	Sample [container]
Iodobenzoic acid (20K)	19387-19436	$C_7H_5IO_2$ [Al]
Benzoic acid (20K)	19282-19331	$C_7H_6O_2$ [Al]
Lithium hydride (100K)	21303-21342	LiH [Al]
Lithium deuteride (100K)	21143-21182	LiD [Al]
Squaric acid (scan)	16929-16948	$C_4H_2O_4$ [Al]
Boron Nitride (4K)	16648-16655	BN [Sn]
Boron Nitride (300K)	16656-16661	BN [Sn]
Graphite (4K)	16674-16679	C [Sn]
Graphite (300K)	16719-16725	C [Sn]
Super Proton Conductor (scan)	14917-14928	Rb_3HSO_4 [Al]
Deuterated Ammonium Palladium Hexachloride (scan)	14515-14529	$(ND_4)2PdCl_6$ [Al]
Glassy zirconium-beryllium (300K)	22542-22575	$Zr_{40}Be_{60}$ [Al]
Polycrystalline zirconium-beryllium (300K)	22576-22608	$Zr_{40}Be_{60}$ [Al]

Table 14: Model selection case studies

The results of each sample that was successfully predicted a model are analysed in detail in the following sections.

Lithium hydride, lithium deuteride and graphite (300K) failed to find a significant model. In all cases this was due to the failure of the initial peak finding algorithm to find significant peaks in the sample data.

When a sample was successfully assigned a model the best model was selected as per the output of the algorithm and the best fitted spectrum selected manually based on the masses identified. All samples were fitted within the 142 to 148 spectra range.

Note that only the fitting parameters relevant to the mass peaks are shown in the parameter tables (i.e. not the background function).

The mass of Deuterium (D) will only be considered in sample known to contain it, where its atomic mass is 2.014102.

5.2.1 Iodobenzoic acid

Figure 19 shows the best model found for the sample of iodobenzoic acid. As with all hydrogenous samples there is a large contribution from Hydrogen in the form of a wide peak around $200\mu s$ and a cluster of heavier masses around $360\mu s$.

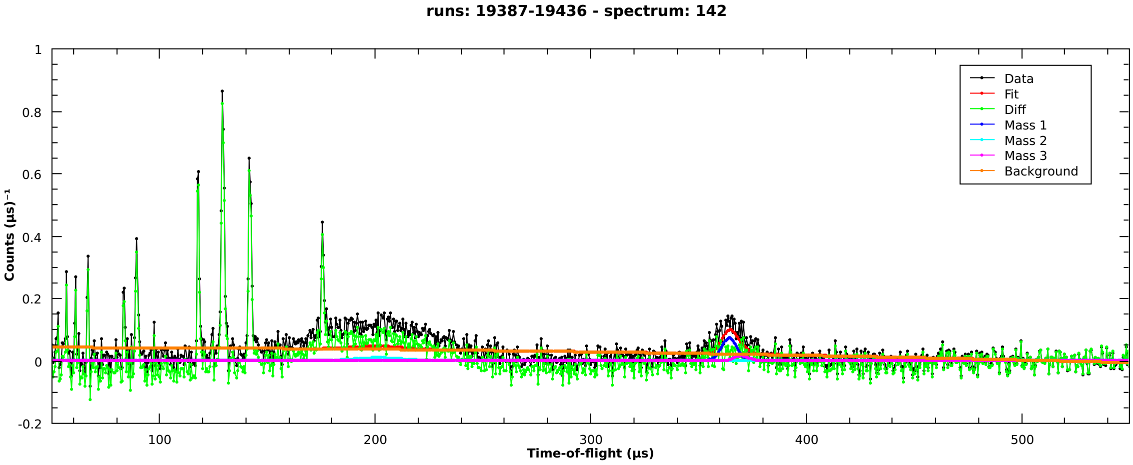

Figure 19: Fitted peaks for Iodobenzoic acid

The parameters describing the best model for iodobenzoic acid are shown in table 15.

Out of the masses detected this was a relatively good fit. The only misidentified mass was Oxygen which was identified as Nitrogen, however given the close masses of the two elements (14.007 vs 15.999) this is a reasonable error.

There are two masses missing from the identification; Carbon and Iodine. It is possible that the peak identified as Nitrogen was actually a combination of the contributions for both Carbon and Oxygen given the closeness of their atomic masses.

Parameter	Value	Error	Closest Element	Likely Element
f0.Mass	14.9833	8.86602	N	O
f0.Width	0.0666095	614.831	-	-
f0.Intensity	2.04385	5.25302	-	-
f1.Mass	1.04072	0.0262829	H	H
f1.Width	0.038286	57.0594	-	-
f1.Intensity	1.57235	0.804694	-	-
f2.Mass	27.4341	118.367	Al	Al
f2.Width	0.419496	1065.94	-	-
f2.Intensity	0.458271	5.24503	-	-
Cost function	2.59649	0	-	-

Table 15: Masses predicted for Iodobenzoic acid

One issue notable both here and throughout the rest of the samples is the low quality of the fitted parameters of the selected model. In this sample the lowest error of a fitted mass is approx. 40% (in the case of `f1.Mass`) and in the case of `f2.Mass` the error far exceeds the value of the parameter.

This does not give good reliability to the model generated, however as demonstrated in this case it is sufficient to determine masses in the sample which can then be used in the existing workflow to obtain a higher quality fit.

5.2.2 Benzoic acid

The best model for benzoic acid is shown by figure 20. This shows the prominent (yet missed by the selection algorithm) Hydrogen peak and a relatively wide peak contributed to by the heavier masses.

There is also a significant overestimation of the background function (shown by the dark purple line). As the peak of this can be seen to be around the $100\mu s$ to $300\mu s$ region the most likely cause of this is the fact that the large Hydrogen peak has not been fitted. Therefore the background function sees this as noise to be removed.

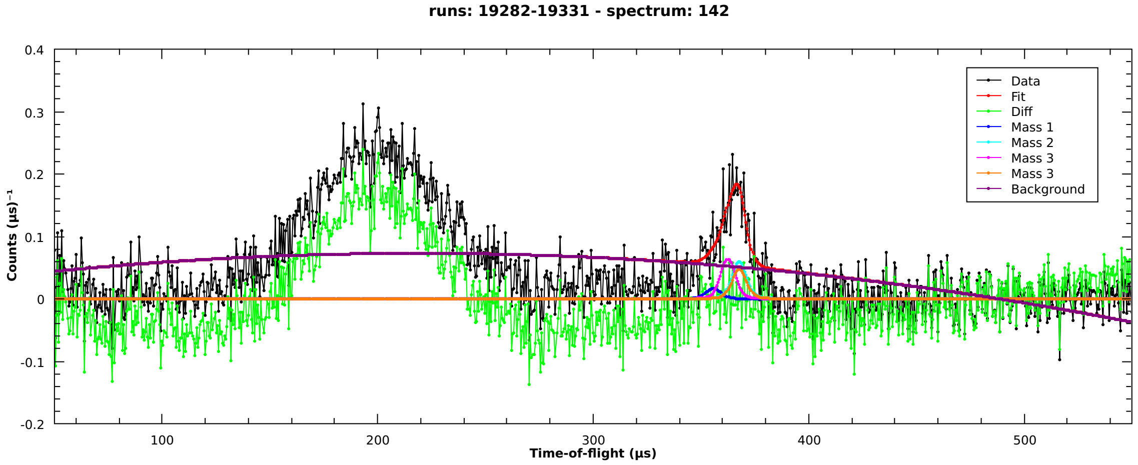

Figure 20: Fitted peaks for Benzoic acid

The parameters describing the bets model are given in table 16.

This shows that multiple peaks have been fitted to what is most likely a contribution form the same mass. This is most obvious in `f1.Mass` and `f3.Mass` (which correspond to the cyan and orange lines in figure 20 respectively).

Only two of the four elements present in this sample could be attributed to the generated fitting parameters and only one of them (Carbon) could be inferred without knowledge of the sample.

Parameter	Value	Error	Closest Element	Likely Element
f0.Mass	8.72683	0.851248	Be	C
f0.Width	4.95629e-07	0	-	-
f0.Intensity	0.493435	3.84378	-	-
f1.Mass	21.0068	0.553725	Ne	Al
f1.Width	1.34619e-06	0	-	-
f1.Intensity	1.62224	4.10071	-	-
f2.Mass	13.0533	1.48582	C	C
f2.Width	1.52466e-06	0	-	-
f2.Intensity	1.78076	3.96747	-	-
f3.Mass	21.0065	0.441612	Ne	Al
f3.Width	8.65973e-06	5.96076e-07	-	-
f3.Intensity	1.29374	4.1007	-	-
Cost function	4.62681	0	-	-

Table 16: Masses predicted for Benzoic acid

The low quality of this selection is likely due to two issues; the low quality of the fit alluded to by the high cost function value and erroneous width parameter values and the fact that multiple peaks have been fitted for the same mass.

5.2.3 Squaric acid

The sample of squaric acid shown in figure 21 is another example of a hydrogenous sample with two distinct peaks; the broad Hydrogen peak and a peak around the cluster of heavy masses.

In this case despite the large cost function value shown in table 17 the fit is reasonable with the exception of the underestimated Hydrogen peak and overestimated background function.

As with the boron nitride sample, the overestimated background is likely partly due to the underestimation of the Hydrogen peak.

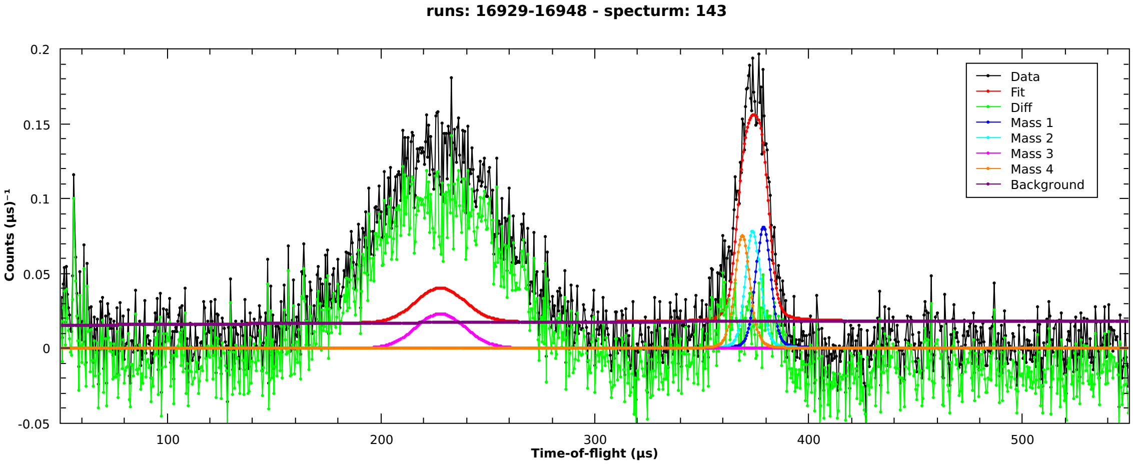

Figure 21: Fitted peaks for Squaric acid

The parameters of this model are given in table 17.

Generally there is good agreement between the fitted masses and the mass the peak is most likely to correspond to, keeping in mind the proximity of Boron and Carbon in terms of atomic mass.

Parameter	Value	Error	Closest Element	Likely Element
f0.Mass	34.5203	146.887	Cl	Al
f0.Width	0.0137715	25682.8	-	-
f0.Intensity	2.213	22.8747	-	-
f1.Mass	17.031	29.7924	O	O
f1.Width	0.0153673	24686.8	-	-
f1.Intensity	2.19405	40.8207	-	-
f2.Mass	1.0489	0.00928642	H	H
f2.Width	-1.35839e-05	0.0223607	-	-
f2.Intensity	3.14565	0.383372	-	-
f3.Mass	11.2096	16.754	B	C
f4.Width	0.0155789	3558.19	-	-
f5.Intensity	2.19305	20.6385	-	-
Cost function	5.80133	0	-	-

Table 17: Masses predicted for Squaric acid

Again with these fitted results there is consistently higher than desired error values on every parameter, however this is partially expected for a fit where obvious features have not been correctly defined (i.e. the Hydrogen peak).

5.2.4 Boron Nitride (4K)

The fit of the best model for boron nitride as shown by figure 22 is one of the highest quality fits of the samples used in these case studies.

This is characterised by the relatively low (yet still higher then desired) parameter errors and low cost function value in table 18 and the good description of the sample data by the fit curve (displayed in red).

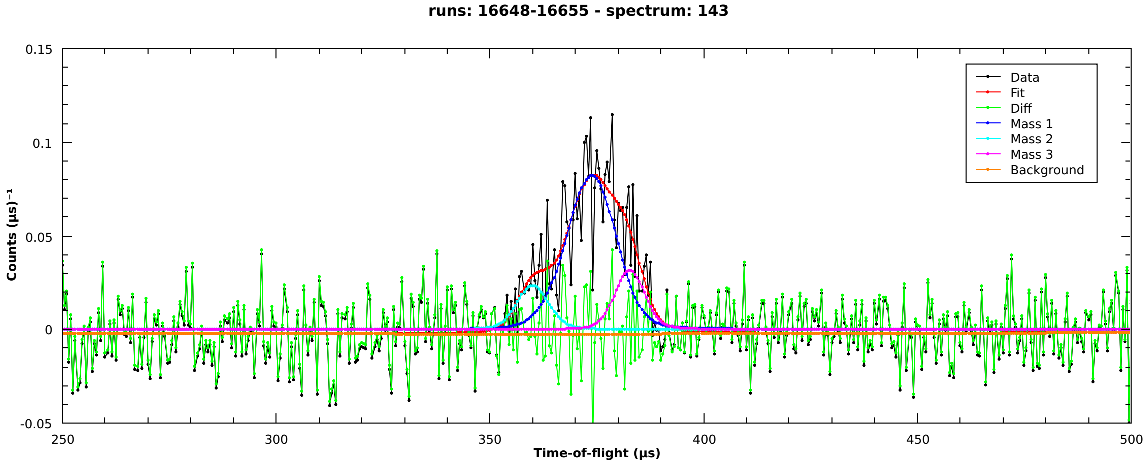

Figure 22: Fitted peaks for Boron Nitride at 4K

Despite the quality of the fit each of the predicted masses have been misassigned, however in the case of Boron and Nitrogen this can be expected given the proximity of the misassigned elements.

The misassignment of Tin is most likely due to noise in the sample data making the true peak centre difficult to observe, therefore a shift in the peak centre would make little difference to the fit quality.

Parameter	Value	Error	Closest Element	Likely Element
f0.Mass	15.9064	1.59037	O	N
f0.Width	10.3502	1.02807	-	-
f0.Intensity	3.71348	5.53052	-	-
f1.Mass	6.8638	2.95772	Li	B
f1.Width	2.99084e-07	0	-	-
f1.Intensity	0.744579	6.42603	-	-
f2.Mass	177.747	0.00661947	Hf	Sn
f2.Width	1.90087e-07	0	-	-
f2.Intensity	0.852565	7.47332	-	-
Cost function	0.964627	0	-	-

Table 18: Masses predicted for Boron Nitride at 4K

5.2.5 Boron Nitride (300K)

In the case of boron nitride at 300K as shown in figure 23 similar mass missassignments have been made, however the model selection preferred a model with 4 masses which slightly changed the nature of the missassignment.

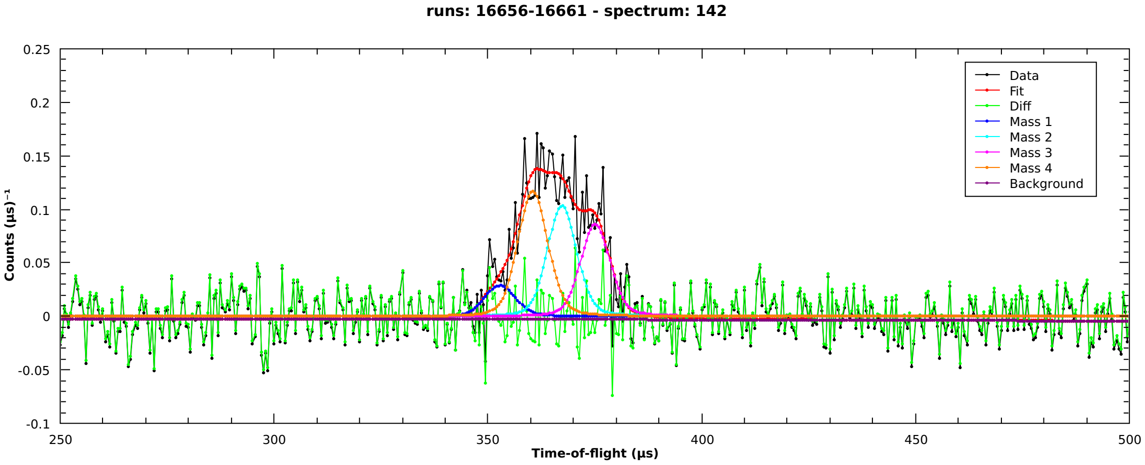

Figure 23: Fitted peaks for Boron Nitride at 300K

The parameters for this mode are given in table 19.

The peak misassignments for this sample are typically further from the correct mass than the same sample at 4K. The reason for this is likely due to the addition of an extra peak in the model, this causes peaks to be more spread out across the observed peak in the sample data rather than converging on more reasonable peak centres.

Parameter	Value	Error	Closest Element	Likely Element
f0.Mass	7.63575	2.62601	Li	B
f0.Width	-6.48345e-07	44.7214	-	-
f0.Intensity	0.892253	5.13656	-	-
f1.Mass	19.9917	1.39316	Ne	N
f1.Width	9.52517e-07	0	-	-
f1.Intensity	2.82771	5.35195	-	-
f2.Mass	130.97	0.0289522	Xe	Sn
f2.Width	9.80083e-07	0	-	-
f2.Intensity	2.27163	5.8403	-	-
f3.Mass	11.3929	4.7515	B	B
f3.Width	1.02277e-06	0	-	-
f3.Intensity	3.35674	5.19103	-	-
Cost function	1.01451	0	-	-

Table 19: Masses predicted for Boron Nitride at 300K

This is still a relatively good fit given the low parameter errors and cost function value, the failure of the model selection in both cases of the boron nitride sample is most likely due to noise in the sample data.

5.2.6 Graphite (4K)

Graphite is another example of a sample that generates a good quality fit from its best model, as shown in figure 24.

In this case the model selection correctly predicted the number of masses and the position of the most significant mass (Carbon). Tin being the container material so is almost always known before the experiment.

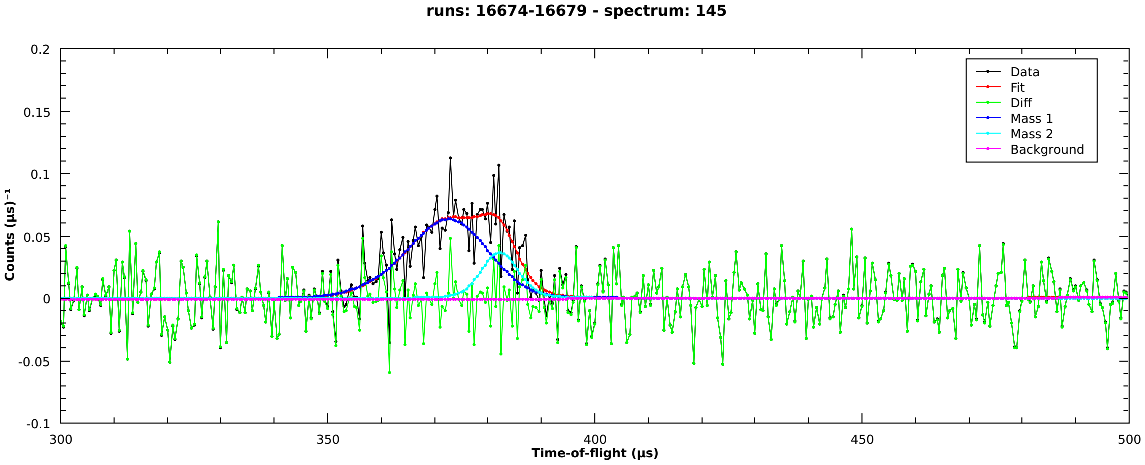

Figure 24: Fitted peaks for Graphite 4K

The parameters in table 20 show reasonable confidence in the prediction of the Carbon contribution given the low error on the `f0.Mass` parameter. Whereas the prediction of Tellurium (which in this sample must be the Tin container) is questionable given the consistently high error on all parameters.

Parameter	Value	Error	Closest Element	Likely Element
f0.Mass	11.8535	1.9326	C	C
f0.Width	12.4429	4.91039	-	-
f0.Intensity	3.75095	1.00662	-	-
f1.Mass	128.033	136.74	Te	Sn
f1.Width	0.000175556	9.99318e+06	-	-
f1.Intensity	0.98068	0.908481	-	-
Cost function	1.11158	0	-	-

Table 20: Masses predicted for Graphite at 4K

The quality of this fit is generally very good, as demonstrated by the low cost function value and visually good description of the sample data. Given this has also been seen with both boron nitride samples it is a reasonable assumption that the complexity of the sample plays a role in the accuracy of the model selection algorithm.

5.2.7 Super Proton Conductor

The best model generated for a super proton conductor (Rb_3HSO_4) is shown in figure 25. This is another example where the quality of the fit of the best model has been reasonably good visually.

The number of masses in the sample has been underestimated, only 3 have been identified as opposed to the 5 present in the sample and container.

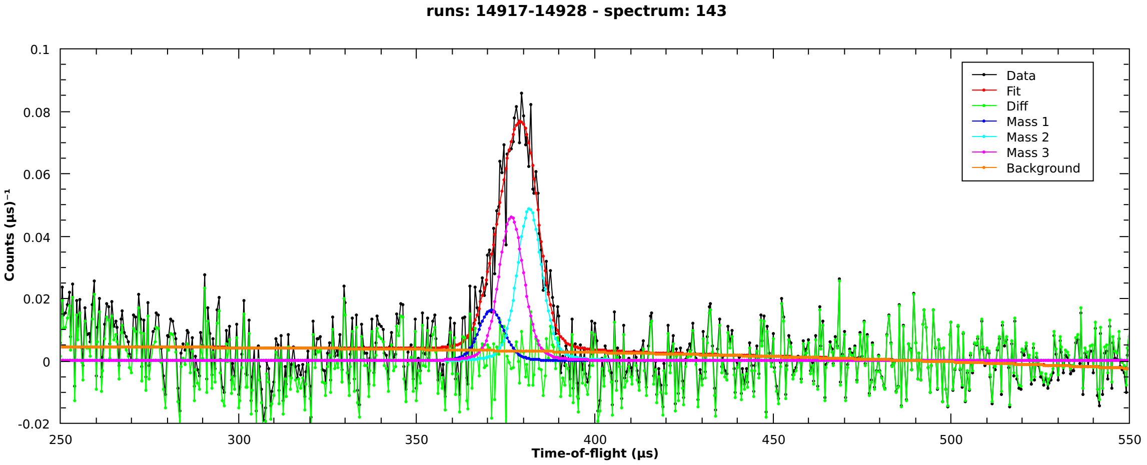

Figure 25: Fitted peaks for Super Proton Conductor

The parameters describing the model are shown in table 21.

This shows that all detected peaks have been misassigned, this is almost certainly due to the oversimplification of the selected model. In this model multiple masses will be contributing to a single peak, therefore shifting the observed peak centre that is fitted by the model selection algorithm.

Parameter	Value	Error	Closest Element	Likely Element
f0.Mass	11.5321	5.0778	C	O
f0.Width	-7.26446e-07	0.0223607	-	-
f0.Intensity	0.46858	1.28711	-	-
f1.Mass	87.932	581.984	Sr	Rb
f1.Width	0.000565619	2.87817e+06	-	-
f1.Intensity	1.30862	6.45269	-	-
f2.Mass	21.2588	35.5052	Ne	Al
f2.Width	0.0920639	2736.79	-	-
f2.Intensity	1.26992	7.53864	-	-
Cost function	1.13736	0	-	-

Table 21: Masses predicted for Super Proton Conductor

Despite the good visual fit and cost function value, the quality of the fit in terms of parameter errors is less than desirable. Particularly with the high errors of the mass parameters which often exceed the value its self.

5.2.8 Deuterated Ammonium Palladium Hexachloride

The sample of deuterated ammonium palladium hexachloride shown in figure 26 is another example of where oversimplification of the best model causes peaks to be misassigned.

In this case only 2 of the 5 masses in the sample and container were identified.

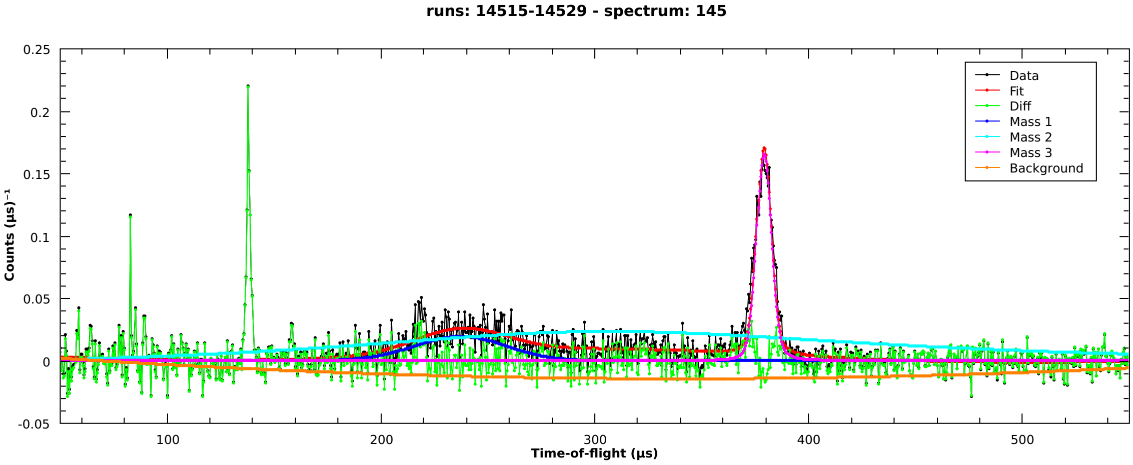

Figure 26: Fitted peaks for Deuterated Ammonium Palladium Hexachloride

The parameters for this model as described in table 22 further show the poor quality of the fit given the erroneous values assigned to certain parameters (e.g `f2.Width`) and high parameter errors.

The fit its self contains several erroneous features, two of the most prominent are the shape of the background which forms a negative parabola centred around 310μs and the fit of the second mass peak (cyan on figure 26) which appears to be a very wide peak.

These two features seem to be cancelling each other out as the difference curve (green line) appears to remain centred around $y = 0$ between the two between the 200μs to 500μs region.

Parameter	Value	Error	Closest Element	Likely Element
f0.Mass	0.984636	128.52	H	D
f0.Width	2.83965	3.8558	-	-
f0.Intensity	4.48147	4.20866	-	-
f1.Mass	1.00794	74.0645	H	D
f1.Width	20.1118	2.93947	-	-
f1.Intensity	24.1379	2.2865	-	-
f2.Mass	34.3805	1.55398	Cl	Cl
f2.Width	9.04439e-07	0	-	-
f2.Intensity	4.48988	14.3249	-	-
Cost function	1.32818	0	-	-

Table 22: Masses predicted for Deuterated Ammonium Palladium Hexachloride

5.2.9 Glassy zirconium-beryllium

A sample of glass like zirconium-beryllium is show in in figure 27.

The fit of this data is generally good despite there being an additional predicted mass (shown by the cyan line), ideally this mass would not have been predicted and the intensities of the other two masses increased to maintain the good fit.

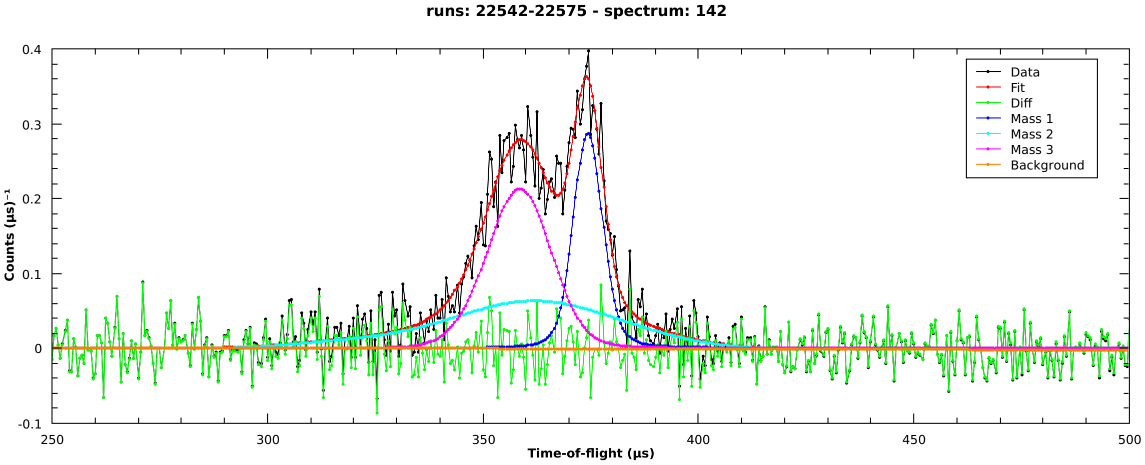

Figure 27: Fitted peaks for glassy zirconium-beryllium

The parameters for this fit are shown in table 23.

There is a single erroneous parameter; `f0.Width`, which is not reflected in the data as the peak seems to match the data well. The mass of this peak is also fitted with a value lower than it should be, however looking at the experimental data it would appear that the fitted value was in fact too high.

Both of these issues could be an artefact of the additional mass being fitted, however if the erroneous peak were to be removed one would expect the position of the first peak (shown by the blue line) to shift down to match the experimental data.

Parameter	Value	Error	Closest Element	Likely Element
f0.Mass	85.2165	0.146364	Rb	Zr
f0.Width	5.71076e-06	0	-	-
f0.Intensity	7.59551	3.80516	-	-
f1.Mass	7.89529	1.84715	Li	Be
f1.Width	24.6869	0.647833	-	-
f1.Intensity	10.2838	1.59485	-	-
f2.Mass	9.66407	5.9531	Be	Be
f2.Width	8.79777	2.36345	-	-
f2.Intensity	12.7169	2.57865	-	-
Cost function	0.966989	0	-	-

Table 23: Masses predicted for glassy zirconium-beryllium

Again this fit seems to be very good based on the majority of parameter errors and the cost function value. It seems less likely that the failure to identify all masses in this case is the result of the sample data given the lack of noise.

5.2.10 Polycrystalline zirconium-beryllium

A sample of polycrystalline zirconium-beryllium (shown in figure 28) was provided as an example of a typical use case in that it contains an unexpected contribution which could be from an additional mass.

This additional contribution is visible around the $325\mu s$ area between the mass peaks for Beryllium and Zirconium which cannot be seen on the glass like zirconium-beryllium sample.

The selected model also contains an erroneous peak `f3` shown as an orange line on the plot. This is made obvious when looking at the parameters in table 24 which show a negative mass for this peak.

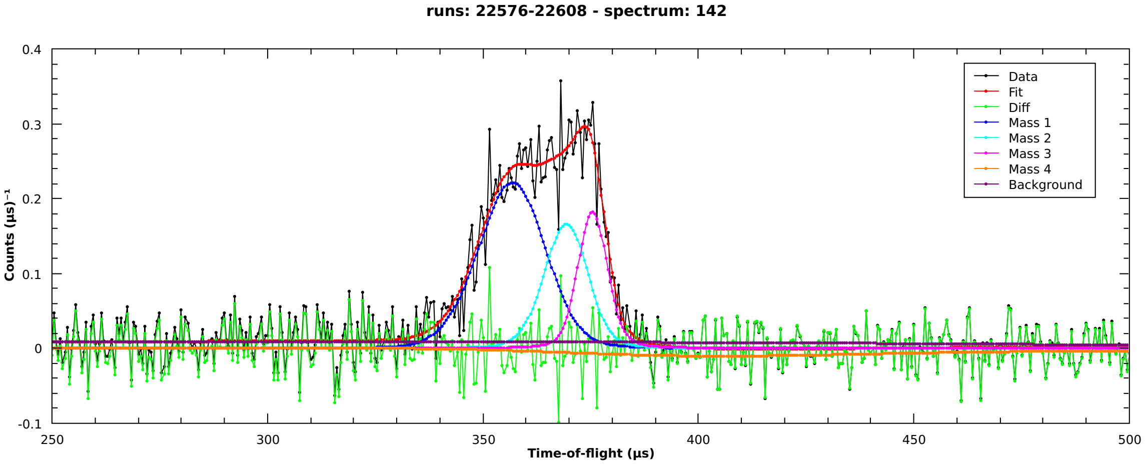

Figure 28: Fitted peaks for polycrystalline zirconium-beryllium

As shown in table 24 the additional contribution has been associated to the Aluminium container. It is unlikely that this is the sole contribution to this given that the glass like zirconium-beryllium is also in an Aluminium container and does not show this additional contribution.

Parameter	Value	Error	Closest Element	Likely Element
f0.Mass	8.84957	7.51495	Be	Be
f0.Width	8.46154	2.76017	-	-
f0.Intensity	13.7909	2.61793	-	-
f1.Mass	24.4406	0.922142	Mg	Al
f1.Width	11.951	0.808455	-	-
f1.Intensity	6.50716	3.35977	-	-
f2.Mass	172.097	0.0261974	Hf	Zr
f2.Width	1.53296e-06	0	-	-
f2.Intensity	4.79642	4.31215	-	-
f3.Mass	-4.49654	0.861339	n/a	n/a
f3.Width	23.4736	0.167577	-	-
f3.Intensity	2.57311	1.85932	-	-
Cost function	1.01689	0	-	-

Table 24: Masses predicted for polycrystalline zirconium-beryllium

The results of this fit (and several others) highlight another issue in the fitting performed by the Bayesian model selection in that the constraints of the fit function do not appear to be respected. An example of which is visible here by the fitted value of the `f3.Mass` parameter being negative.

6 Further Development

This section describes additional work that could be carried out to improve on the additional features implemented as part of this dissertation and how they assist is better meeting the objectives set out at the start of the project.

6.1 Multivariate Gaussian multiple scattering correction

Further work is required to allow a correctly scaled multiple scattering correction when fitting with the multivariate Gaussian function, as already mentioned in section 3.3.4.

While the true cause of this is yet to be found it is clear that the problem originates from the difference in relative intensities between the fitted Hydrogen peaks between the multivariate Gaussian model and other models.

The main issue with the lack of correct multiple scattering corrections is the lack of consistency between results of the entire workflow using the Gram-Charlier profile and when using the multivariate Gaussian profile. The difference being brought about due to the difference in the data spectrum at the start of the final fitting stage.

However if the calculated multiple scattering correction is too unmatched to the sample data then the correction scale factor fit will most likely attenuate the correction to a greater degree before it is applied to the sample data.

6.2 Optimise calculation of multivariate Gaussian mass profile

A simple improvement that could be undertaken with respect to the multivariate Gaussian is to optimise the calculation of the Compton profile and A_3 FSE correction to be performed in the same integration, this would greatly reduce the computational complexity introduced by the nested loop under which the integration is performed.

Under this new method the fitted function would become:

$$y' = I \cdot J(y,q) = \int_0^1 d(cos\theta) \int_0^{\frac{\pi}{2}} d\phi \, (J - A_3(q)) \, exp\left(-\frac{y^2}{2S^2(\theta,\phi)}\right) \tag{29}$$

where; J and $A_3(q)$ are pre calculated factors defined by equations 30 and 31 respectively.

$$J = \frac{1}{\sqrt{2\pi}\sigma_x\sigma_y\sigma_z}\frac{2}{\pi}S^2(\theta,\phi) \tag{30}$$

$$A_3(q) = \frac{\sigma_x^4 + \sigma_x^4 + \sigma_x^4}{9\sqrt{2\pi}\sigma_x\sigma_y\sigma_z q}\left[\frac{y^3}{S^2(\theta,\phi)^4} - 3\frac{y}{S^2(\theta,\phi)^2}\right] \tag{31}$$

6.3 Optimise, constrain and cull generated models

Ideally the number of models created after the initial peak finding stage should be as low as possible to decrease the time and chance of a bad model being selected in the Bayesian model selection stage and models should be as simple (i.e. the lowest number of degrees of freedom) as possible in order to prevent a good fit being generated for a bad model.

6.3.1 Degrees of freedom reduction

One possible way to both decrease the time taken for a fit in the Bayesian model selection stage is to reduce the degrees of freedom (number of freely fitted parameters) by fixing certain parameters.

One possible example is the width of the peaks associated with heavy masses (> 4amu). This will reduce the chance that such peaks could be fitted to a spike in background noise (with a very small width) or fitted to a curve in the background (with a very large width).

6.3.2 Using diffraction data

Through analysis of a diffraction experiment on VESUVIO (or any other powder diffractometer for that matter) it is possible to build a crystal structure from which the atomic vibrations within the sample can be modelled.

The parameters describing this atomic vibration can be related back to momentum distribution through thermal parameters U and B (in the case of isotropy) using the Debye [1] model.

Being able to obtain the momentum distribution would allow stricter constraints to be placed on the models used in the Bayesian model selection, reducing the chance of a bad model being selected.

This additional functionality would require further investigation into the state of powder diffraction reduction in MANTID, at the time of writing there is ongoing work to standardise the reduction of powder diffraction data which may provide the majority of the additional functionality required to implement these additional constraints. Otherwise this would require the implantation of a new diffraction reduction algorithm for VESUVIO.

6.4 Improved robustness of initial peak finding

As already described in section 4.3.2 a significant issue in the full automation of the model selection workflow was the inability to reliably obtain an initial prediction of peak positions.

There are multiple ways to address this problem; require the user to supply information to improve the quality of the existing peak finding method, run the existing peak finding method over a series of parameters, improve the quality of the sample data before initial peak finding is performed or implement a new peak finding algorithm.

Parameterise peak finding

The easiest option is to provide the option to set the *Tolerance* and *FWHM* parameters of the `FindPeaks` to the user when they run the model selection routine. This would still require the same running, manual inspection and possible re-running until the results were as expected which defeats the purpose of this algorithm.

Run existing algorithm over multiple parameter sets

An alternative to the previous option would be to create a list of parameters that would ensure all true peaks would always be found and execute the `FindPeaks` algorithm over each set of parameters. The found peaks would then be inspected and peaks that fall outside of the range possible for a mass peak would be removed, duplicate peaks would be identified with the peak with the widest width kept and others removed.

Two issues arise from this method:

1 The wide range of parameters will almost certainly introduce more spurious peaks which if not removed can cause issues in the Bayesian model selection stage. Both in terms of execution speed as this will cause more complex models to be created which take longer to optimise and depending on the relative quality of true mass peaks to false peaks there is potential for the most suitable model selected by the routine to contain false peaks.

2 The initial peak finding will take longer than it currently does, however this is a relatively fast process in the context of the entire model selection routine.

Improve signal to noise ratio in sample data

The signal to noise ratio of the sample data could potentially be improved by first fitting a background function to the sample data to obtain a background correction. The function used would most likely be a polynomial of order 2 in order to match that used by default in the standard analysis workflow.

This would assist in reducing the number of spurious peak identifications, however would not address the issue of missing true peaks. A realistic solution could be to employ this solution alongside the first or second method mentioned previously.

Implement new peak finding algorithm

Another solution is to implement a new peak finding algorithm to use as an alternative to the current `FindPeaks` algorithm. This would require further investigation to find an appropriate algorithm.

7 Conclusions

In this section I will review the project and how well it has met the original objectives as well as identify the causes of any problems hindering the project from meeting those objectives.

7.1 Objectives

7.1.1 Multivariate Gaussian fitting

I am confident the implementation of the multivariate Gaussian peak profile is correct and that if given a set of parameters that are known to describe an anisotropic mass peak, it would do so correctly.

The issues faced in section 5.1 are more likely to be caused by a combination of the higher dimensionality of the multivariate Gaussian fit function and the lack of constrains placed upon the parameters, other than basic constrains such as intensity being greater then zero, etc.

During the ongoing testing while implementing the new fit function I also had access to a different data set which had considerably less noise, as such the possibility that the issues observed are related somehow to noisy data is not zero, however it is unlikely.

If the multivariate Gaussian profile were to undergo the improvements mentioned in section 6 and investigation into a suitable set of constrains then I am confident this could become a usable part of the VESUVIO analysis workflow.

That being said, the implemented fit function and its integration into the analysis workflow are already part of MANTID and will be available to users as of the next release of the software.

7.1.2 Model selection

Whilst the model selection algorithm is not at a stage at which it could be widely used it is a reasonable indication of an unknown sample composition, especially for samples that doe not have a very complex composition (those with less than four masses typically perform best).

Out of the predictions made over all samples analysed as part of the case studies (section 5.2) 62% of all masses in the sample were identified as being present by the algorithm. This includes masses that were incorrectly identified (i.e. the closest atomic mass was not that of the expected element) but still within the error bars of the fitted mass parameter.

34% of all masses identified by the algorithm were identified correctly, a correct identification is where the element with the closest atomic mass to the mass reported by the model selection algorithm is a contributor to the peak being fitted. This is seen in the case studies as the *Closest Element* and *Likely Element* columns containing the same element for a given fitted peak.

This level of accuracy is considerably less that desired for the model selection routine in its current form to become a standard part of the VESUVIO analysis workflow. Given the amount of time taken for the current workflow a failure of this routine to correctly identify masses could end up costing more time than manual examination of the data and deduction of the mass based on knowledge of the sample would.

Several ways in which this accuracy could be increased have already been discussed in section 6, these ideas should be used to adapt the current routine into a truly robust model selection routine.

7.2 Summary

Overall I feel that this project would have progressed more if the original objectives had focused more on a single are rather than the two distinct areas of fitting and model selection. Having to work on two separate areas of the analysis workflow put too much of a strain on the time available for this project.

In the original objectives I state that at the end of the project what I aim to have are a set of additions to the existing VESUVIO workflow that are ready to be used by scientists to analyse their data, in this respect the project was partly unsuccessful. However what has been produced (especially in the case of the model selection algorithm) is a basis for further work on these features which will result in a more polished and ready to use solution to the problems this project aims to solve.

My time working on MANTID both in the past and during this project has provided me with enough knowledge and motivation to continue work in these areas of VESUVIO data analysis and produce what I set out to at the start of this project.

Appendices

A Example Workflow Script

```python
from vesuvio.workflow import fit_tof

## Standard flags to modify processing

runs = "15039-15045"

flags = dict()

flags[’fit_mode’] = ’spectra’
flags[’spectra’] = ’143-150’
flags[’bin_parameters’] = None

mass1 = {’value’: 1.0079, ’function’: ’GramCharlier’, ’width’: [2, 5, 7],
         ’hermite_coeffs’: [1,0,0], ’k_free’: 0, ’sears_flag’: 1}
mass2 = {’value’: 16.0, ’function’: ’Gaussian’, ’width’: 10}
mass3 = {’value’: 27.0, ’function’: ’Gaussian’, ’width’: 13}
mass4 = {’value’: 133.0, ’function’: ’Gaussian’, ’width’: 30}
flags[’masses’] = [mass1, mass2, mass3, mass4]

flags[’intensity_constraints’] = list([0, 1, 0, -4])

flags[’background’] = {’function’: ’Polynomial’, ’order’: 2}

## Corrections flags

flags[’output_verbose_corrections’] = True

flags[’container_runs’] = None
flags[’fixed_container_scaling’] = None

flags[’gamma_correct’] = True
flags[’fixed_gamma_scaling’] = None

flags[’ms_flags’] = dict()
flags[’ms_flags’][’SampleWidth’] = 10.0
flags[’ms_flags’][’SampleHeight’] = 10.0
flags[’ms_flags’][’SampleDepth’] = 0.5
flags[’ms_flags’][’SampleDensity’] = 241

# Optional parameters (default values are given)
# flags[’ms_flags’][’Seed’] = 123456789
# flags[’ms_flags’][’NumScatters’] = 3
# flags[’ms_flags’][’NumRuns’] = 10
# flags[’ms_flags’][’NumEvents’] = 50000
# flags[’ms_flags’][’SmoothNeighbours’] = 3
# flags[’ms_flags’][’BeamRadius’] = 2.5

## Advanced flags

flags[’ip_file’] = ’IP0004_10.par’
flags[’diff_mode’] = ’single’

flags[’max_fit_iterations’] = 5000
flags[’fit_minimizer’] = ’Levenberg-Marquardt,AbsError=1e-08,RelError=1e-08’

flags[’iterations’] = 1
flags[’convergence’] = None

## Run fit
fit_tof(runs, flags, flags[’iterations’], flags[’convergence’])
```

Listing 1: Example workflow driver script

B List of contributions to the MANTID project

The following is a list of code contributions to the MANTID project made as a result of work carried out on this project.

Allow fitting atomic mass of peak

Modifies the `ComptonProfile` fit function to allow the atomic mass of the peak being fitted by the function to be a fitted parameter rather than a static attribute.

`https://github.com/mantidproject/mantid/pull/15675`

Multivariate Gaussian profile

Adds the fit function for the multivariate Gaussian profile and integrates it into the existing VESUVIO analysis workflow.

`https://github.com/mantidproject/mantid/pull/15773`

Model selection algorithm (work in progress)

Initial development of the model selection algorithm.

`https://github.com/mantidproject/mantid/tree/14943_vesuvio_model_selection`

References

[1] P. Debye. “Zur Theorie der spezifischen Wärmen”. In: *Ann. Phys.* 344.14 (1912), pp. 789–839. DOI: 10.1002/andp.19123441404.

[2] S Jackson et al. “VESUVIO Data Analysis Goes MANTID”. In: *J. Phys.: Conf. Ser.* 571 (Dec. 2014), p. 012009. DOI: 10.1088/1742-6596/571/1/012009.

[3] Kenneth Levenberg. “A method for the solution of certain non-linear problems in least squares”. In: *Quarterly Journal of Applied Mathmatics* II.2 (1944), pp. 164–168.

[4] *Mantid: Manipulation and Analysis Toolkit for Instrument Data.* DOI: 10.5286/Software/Mantid.

[5] M.A. Mariscotti. “A method for automatic identification of peaks in the presence of background and its application to spectrum analysis”. In: *Nuclear Instruments and Methods* 50.2 (May 1967), pp. 309–320. DOI: 10.1016/0029-554x(67)90058-4.

[6] J. Mayers. *User guide to VESUVIO data analysis programs for powders and liquids.* http://www.isis.stfc.ac.uk/instruments/vesuvio/documents/vesuvio-data-analysis-manual11089.pdf. accessed 01/05/2016. Oct. 2010.

[7] J Mayers and T Abdul-Redah. “The measurement of anomalous neutron inelastic cross-sections at electronvolt energy transfers”. In: *Journal of Physics: Condensed Matter* 16.28 (July 2004), pp. 4811–4832. DOI: 10.1088/0953-8984/16/28/005.

[8] J. Mayers and M.A. Adams. “Calibration of an electron volt neutron spectrometer”. In: *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment* 625.1 (Jan. 2011), pp. 47–56. DOI: 10.1016/j.nima.2010.09.079.

[9] D Monserrat et al. “FABADA Goes MANTID to Answer an Old Question: How Many Lines Are There?” In: *J. Phys.: Conf. Ser.* 663 (Nov. 2015), p. 012009. DOI: 10.1088/1742-6596/663/1/012009.

[10] D Nixon. “Developments in MANTID relating to indirect inelastic spectroscopy between July 2014 - July 2015”. In: *RAL Technical Reports RAL-TR-2015-007* (2015).

[11] L C Pardo et al. “FABADA: a Fitting Algorithm for Bayesian Analysis of DAta”. In: *Journal of Physics: Conference Series* 325.1 (2011), p. 012006.

[12] G Romanelli. “On the quantum contributions to phase transitions in Water probed by inelastic neutron scattering”. In: (2015).

[13] E. M. Schooneveld et al. “Foil cycling technique for the VESUVIO spectrometer operating in the resonance detector configuration”. In: *Rev. Sci. Instrum.* 77.9 (2006), p. 095103. DOI: 10.1063/1.2349598.

[14] V. F. Sears. “Scaling and final-state interactions in deep-inelastic neutron scattering”. In: *Phys. Rev. B* 30 (1 July 1984), pp. 44–51. DOI: 10.1103/PhysRevB.30.44.

[15] D.S. Sivia et al. “An introduction to Bayesian model selection”. In: *Physica D: Nonlinear Phenomena* 66.1 (1993), pp. 234–242. ISSN: 0167-2789. DOI: 10.1016/0167-2789(93)90241-R.

[16] Geoffrey B. West. “Electron scattering from atoms, nuclei and nucleons”. In: *Physics Reports* 18.5 (1975), pp. 263–323. ISSN: 0370-1573. DOI: 10.1016/0370-1573(75)90035-6.

[17] C. Windsor. *Pulsed Neutron Scattering.* London: Taylor and Francis, 1821.