MAY 1, 2021

DOUBLE ELECTRON-ELECTRON RESONANCE

ANALYSIS OF RESONANCE DATA FOR EXTRACTING DISTANCES IN BIOMOLECULES

ALAN CUTFORTH
SCHOOL OF PHYSICS AND ASTRONOMY, UNIVERSITY OF NOTTINGHAM
3rd Year Physics Project

CONTENTS

Abstract

This project examined DEER (double electron-electron resonance), a form of pulsed electron paramagnetic resonance that utilises electron spin echoes to measure distances in chemical structures. The usefulness of principal component analysis (PCA) by singular valued composition (SVD) on existing data taken from academic literature sources was assessed. The hypothesis that a higher quantity of principal components contained within the data is indicative of a more rigid chemical-biradical structure was tested. **65%** of the post-PCA recalculated distance values matched within 1σ or a 5% tolerance of the values stated in the source literature. The rigid chemical-biradical structures analysed were found to have an average principal component to DEER trace ratio of **0.604**, and the non-rigid chemical-biradical structures had an average ratio of **0.495**. Only limited data was analysed, and the fact that the magnetic field strength should affect the number of principal component curves was not accounted for when calculating these ratios. Despite this, the results appear to indicate a weak correlation between the number of principal component curves and the rigidity of the biradical molecular structures.

Introduction

Electron paramagnetic resonance (EPR) spectroscopy uses the magnetization of electron spins to provide information about chemical structure and dynamics[1]. It can be performed on systems with one or more unpaired electrons. It is currently widely used to measure distances at the 20-80 Angstrom range[2]. It is used especially frequently for measuring distances in biomolecules, such as proteins or nucleic acids, particularly in the study of their chemical structure[3]. This project deals with one of the most popular EPR experiments, double electron-electron resonance (DEER). DEER separates pairwise couplings between electron spins from other electron spin interactions[4].

DEER works by initially exciting observer electrons - flipping their spin vectors by an angle using pulses of electromagnetic radiation, producing measurable changes in their magnetic dipolar fields. Spin echoes were first discovered by Erwin Hahn in 1950, when he applied two electromagnetic pulses causing two successive $\frac{\pi}{2}$ radian changes in the spin of an electron. After this, he detected a signal when no pulse was applied – this was the spin echo[5]. It was further developed by using a π radian pulse for the second pulse[6]. DEER is used to calculate distances by analysing the intensity curves produced by measuring the resultant changes in the observer electrons' magnetic fields due to these spin echoes.

The first aim of this project was to use pre-existing DEER trace data from published literature sources to recreate these distance calculations, but by first performing principal component analysis (PCA) by singular valued decomposition (SVD) on the DEER traces. PCA is a technique for reducing the dimensionality of large datasets, increasing interpretability but at the same time minimizing information loss[7]. PCA by SVD works by effectively dismantling a signal into its component parts and taking only the principal components – the components by which the data can be effectively constructed. In doing this, little information should be lost. The intention is to streamline the DEER

distance calculation and trace analysation process by only considering meaningful components that can be used to accurately construct the entire set of DEER traces.

The second aim was to test the hypothesis that the more rigid the chemical-biradical structure being observed by DEER, the more principal components the data will contain per trace; potentially revealing another method of analysing the rigidity of molecules when observing them using EPR.

Theory

Double Electron-Electron Resonance

To produce a spin echo two electromagnetic pulses are used to excite a group of electrons and change the orientation of their spin vectors. Usually a $\frac{\pi}{2}$ radian pulse is followed by a π radian pulse some variable time later[6]. Between the two pulses, due to the local magnetic field variations, the signal decays, slowing the rotation of some of their spin vectors. When the second pulse is applied, the spin vectors which were slowed by the magnetic variations are now in front of the other spin vectors. The fast spin vectors catch up with the slower vectors, eventually producing a complete refocusing or "echoing" of the initial flipped orientation[4]. This produces a loss of phase coherence of electron spin magnetisation[4]. As this occurs, a change is produced in the dipolar magnetic field. Electromagnetic waves are then used to flip the other (pump) electron(s) spin vectors. The spin echoes from the observer electrons interacts with the pump electrons magnetic dipolar field, effecting the observer's spin echo at some time after the initial observer echo causing a change in its intensity. See Figure 1 for a diagram of this process[8].

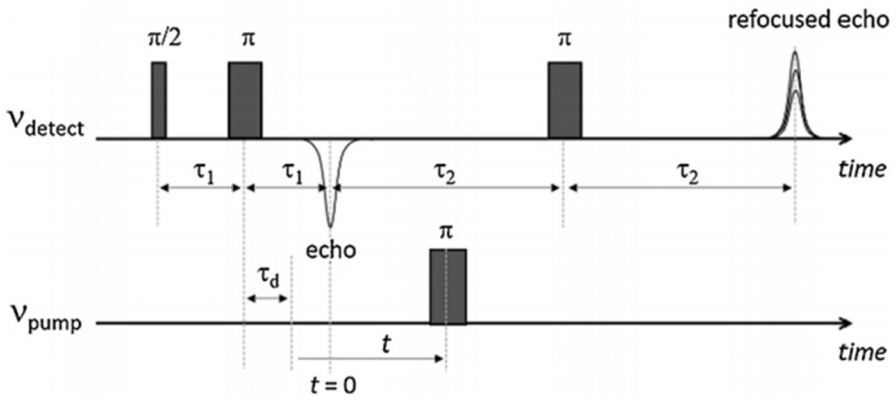

Figure 1: DEER sequence (four-pulse)[8]. The intensity of the refocused echo is measured over varied time *t*. In this diagram the observer electron is referred to by the subscript 'detect'.

The pump electrons are flipped at different times, producing an intensity distribution for the interaction between the pump and observer's magnetic dipolar fields. The peak of the Fourier transform of this intensity distribution gives the resonant frequency most useful for analysing the distances between pump and observer electrons[3].

For a point-dipole approximation, the frequency of the dipolar interaction depends on the distance between the electrons, and an angle determining the orientation of the magnetic field[3]. Assuming

exchange coupling between the electron spins is neglected and both spins are quantized along the external magnetic field[4], the equation for the dipolar interaction frequency is given by[3]:

$$\omega_{dd}(r_i, \theta_i) = \frac{\mu_0 g_i g_{observer} \beta_e}{4\pi\hbar} \frac{(1 - 3cos^2\theta_i)}{{r_i}^3} = \omega_{dd,0}(1 - 3cos^2\theta_i) \quad (1)$$

where ω_{dd} is the dipolar interaction frequency between the observer and pump electrons i, g_i and $g_{observer}$ are the gyromagnetic ratios of the pump and observer electrons respectively, and r_i and θ_i are the distance and angle between the observer and pump electron i [3]. Figure 2 shows a diagram of these quantities[4].

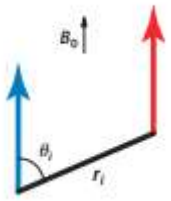

Figure 2: Diagram of observer electron (blue) and pump electron (red), with the net external magnetic field B_0 shown[4].

The distances between pump and observer electrons are assumed to be within the sensitive range of DEER and are hence measurable, as the distance will only be correctly measured by this process if the electrons are in the correct range for the magnetic field changes to resonate[4].

The DEER intensity signal, or DEER trace, $S(t)$ is caused by the excitation of spin pairs both within and outside the same molecule. It is given by[3]:

$$s_j(t) = cos[\omega_{dd}(r_j, \theta_j)] \quad (2)$$

$$S(t) = \prod_{j,all\ pairs} s_j(t) = S_{inter}(t) S_{intra}(t) \quad (3)$$

where $S_{inter}(t)$ is the term corresponding to the intensity within the molecular structure, and $S_{intra}(t)$ is the term corresponding to the external intensity. It can be shown that the summation over all pairs outside the molecule can be described by the integral[3]:

$$S_{intra}(t) = \frac{S(t)}{S_{inter}(t)} = 1 - \int_0^{2\pi} d\phi \int_0^{\pi} sin\theta d\theta \int_0^{\infty} \lambda(\theta, \phi)\, x\, [1 - cos(\omega_{dd}(r, \theta)t)] f(r) dr \quad (4)$$

where ϕ and θ are the angles relating to the orientation of the pump electrons with respect to the observers, r is the distance between the electrons, f(r) is the distance distribution, and $\lambda(\theta, \phi)$ is the pair excitation probability density function[3]. A set of DEER traces, along with their associated magnetic field strengths, can be seen in Figure 3.

The pair excitation probability density function can be expanded into spherical harmonics[3]:

$$\lambda(\theta, \phi) = \sum_{l=0}^{\infty} \sum_{m=-l}^{l} \lambda_{lm} Y_{lm}(\theta, \phi) \quad (5)$$

where $Y_{lm}(\theta, \phi)$ are the spherical harmonics, and l and m are the l and m quantum numbers. The modulation depth is defined as $\lambda = \lambda_{00}$[3].

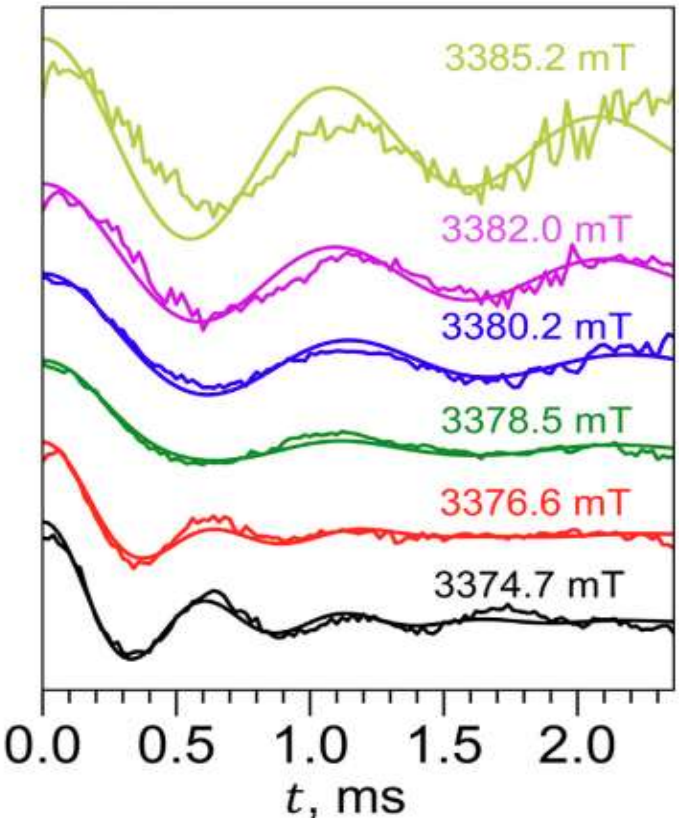

Figure 3: Double electron-electron resonance (DEER) traces with their associated magnetic field strengths[3].

Tikhonov Regularisation

To produce the distance distribution in equation (4), *f(r)*, the parts of the equation which are angular dependent need to be removed. By taking a Fourier transform of the time-dependent portion of equation (4), it can be shown that a set of DEER traces can be represented using a linear combination of modified Pake pattern (MPP) components[3]:

$$\tilde{S}(\omega) = \sum_{k=0}^{\infty} W_k^{MPP} S_k^{MPP}(\omega) \quad (6)$$

where $W_k^{MPP} = \lambda_{2k,0}$ is an always positive k-th degree weight of the MPP component. This is useful as the actual MPP components, $S_k^{MPP}(\omega)$, depend only on the dipolar interaction and its underlying distance distribution, with the angular information contained within the weights W_k^{MPP}[3].

Using a system based on Tikhonov regularisation, an infinite series of kernels and the MPP components can be used to find the distance distribution, *f(r)*. The following calculation is repeated until convergence[3]:

$$\sum_{j,over\ all\ B_j}[\int_{r_{min}}^{r_{max}} \sum_{k=0}^{k_{max}} p_k(B_j) K_k(r,t) f(r) dr - S^j(t)]^2 + \eta[\sum_{j,over\ all\ B_j}[\sum_{k=0}^{k_{max}}(p_k(B_j))^2] \rightarrow min \quad (7)$$

where the coefficient $p_k(B_j) \rightarrow W_k^{MPP}(B_j)\ when\ k_{max} \rightarrow \infty$, B_j is the magnetic field strength of the recorded DEER trace *j*, η is a regularisation parameter which affects the scaling of solutions, and $K_k(r,t)$ are the kernels given by[3]:

$$K_k(r,t) = 2\pi \int_0^{\pi} cos(\omega(r,\theta)t) Y_{2k,0}(\theta,\phi) sin\theta d\theta \quad (8)$$

Using the iterative method described by equation (7), a distance distribution function $f(r)$ can be found with arbitrary amplitude values and hence plotted. However, as the peak of the distance distribution is what indicates the important quantity – the distance – it is not a problem that the amplitude values are arbitrary, and they can be normalised.

Principal Component Analysis

To use PCA by SVD in the context of this project, the data needed to be placed into a *m* x *n* matrix, where *m* is the number of signal traces and n is the number of points taken in these traces. This matrix has been called $\mathbf{X}$. SVD is achieved by decomposing $\mathbf{X}$ into the product of 3 other matrices[9].

As $\mathbf{X}$ is an *n* x *m* matrix, $\mathbf{X}^{\mathrm{T}}\mathbf{X}$ is a square, symmetric *m* x *m* matrix:

$$(\mathbf{X}^{\mathrm{T}}\mathbf{X})\hat{\boldsymbol{v}}_{\boldsymbol{i}} = \lambda_i \hat{\boldsymbol{v}}_{\boldsymbol{i}} \tag{9}$$

where $\hat{\boldsymbol{v}}_{\boldsymbol{i}}$ is the set of orthonormal *m* x 1 eigenvectors with associated eigenvalues, λ_i[9]. The singular values, σ_i, are defined such that $\sigma_i = \sqrt{\lambda_i}$. Now the *n* x *1* set of vectors $\hat{\boldsymbol{u}}_i$ can be defined such that:

$$\hat{\boldsymbol{u}}_i \sigma_i = \mathbf{X}\hat{\boldsymbol{v}}_i \tag{10}$$

This can be summarised in one matrix multiplication by constructing some new matrices. $\mathbf{\Sigma}$ is a diagonal *m* x *n* matrix, meaning its only non-zero values are along the leading diagonal. $\mathbf{\Sigma}$ contains the singular values σ_i in diagonally descending value order. The matrices **V** and **U** are unitary, singular and are constructed from the sets of orthonormal vectors $\hat{\boldsymbol{v}}_{\boldsymbol{i}}$ and $\hat{\boldsymbol{u}}_i$ appended with additional vectors such that all of their values are non-zero, to deal with degeneracy issues[9].

Hence $\mathbf{X}$ can be decomposed into the product of $\mathbf{U}$, $\mathbf{\Sigma}$, and the transpose of $\mathbf{V}$, as can be seen in equation (11)[7]:

$$\mathbf{X} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^{\mathrm{T}} \tag{11}$$

$\mathbf{U}$ contains fundamental values which are properties of $\mathbf{X}$, ordered in terms of their ability to describe the variance of the columns of $\mathbf{X}$. The columns of the transpose of **V** give the mixture of all $\mathbf{U}$ that makes up each corresponding column of $\mathbf{X}$. The singular values contained in $\mathbf{\Sigma}$ correspond to the respective trace in $\mathbf{X}$ – these describe how correlated $\mathbf{U}$ is with $\mathbf{X}$ and are coefficients[9].

Due to the diagonality of $\mathbf{\Sigma}$, expanding equation (11) by index notation gives:

$$\mathbf{X} = \sigma_1 U_1 {V_1}^T + \sigma_2 U_2 {V_2}^T + \ldots\ldots + \sigma_m U_m {V_m}^T \tag{12}$$

with all other components being 0.

If there are some σ_i such that $\sigma_i >> \sigma_j$ where both σ_i and σ_j are some singular values and $i \neq j$, σ_j can be considered negligible. In this case, the only remaining terms from equation (12) are the components with the most impact on the signal traces stored in $\mathbf{X}$ – these components are the principal components of $\mathbf{X}$[9].

Application of PCA on DEER traces will reveal which DEER traces contribute significantly to the dataset. This will provide information on how many MPP degrees are required to describe the traces, and demonstrates that higher degree MPP components have smaller contributions[3].

Rigidity and EPR Field Strength

For the DEER traces, it is possible that more rigid biradical structures will have a higher ratio of the number of principal components to the number of traces. This is because, with less rigid biradical structures, paths between electron spins will likely conform more to the same set of principal components.

DEER traces taken using high-field excitation should also produce more principal components, as high-field excitation produces a tighter excitation of orientation subsets.

It is also possible, in the case of datasets containing a high number of traces, there is more chance of having duplicate paths between electron spins. This could possibly skew some results. For instance a rigid biradical structure that would normally contain a high ratio of principal components to traces will, in the case of a high trace count, be more likely to contain duplicate or very similar traces. This causes them to conform more to the same principal components as each other, lowering the ratio.

Method

Scraping Data

No data was experimentally collected for the purposes of this project. DEER trace data has already been gathered for numerous academic papers; this project only required the re-analysation of this data. These papers do not contain raw data, only the trace graphs – to acquire numerical data to analyse, it had to be scraped from these graphs.

The software used for this data scraping process was the web-based application WebPlotDigitzer[10]. This program allows the user to upload an image containing a graph to be analysed. A box is drawn around the required plot, and a colour can be specified to assist in recognising which line is to be analysed. This program was not perfectly accurate, as with some graphs there were multiple overlapping traces where only one was relevant, or traces were dotted or dashed lines which made reading them with WebPlotDigitizer more difficult. In these cases, it was often necessary to use the manual point function supplied by the website to individually choose relevant points. The points were scraped with an accuracy of 1 point per 2 pixels. See Figure 4 and Figure 5.

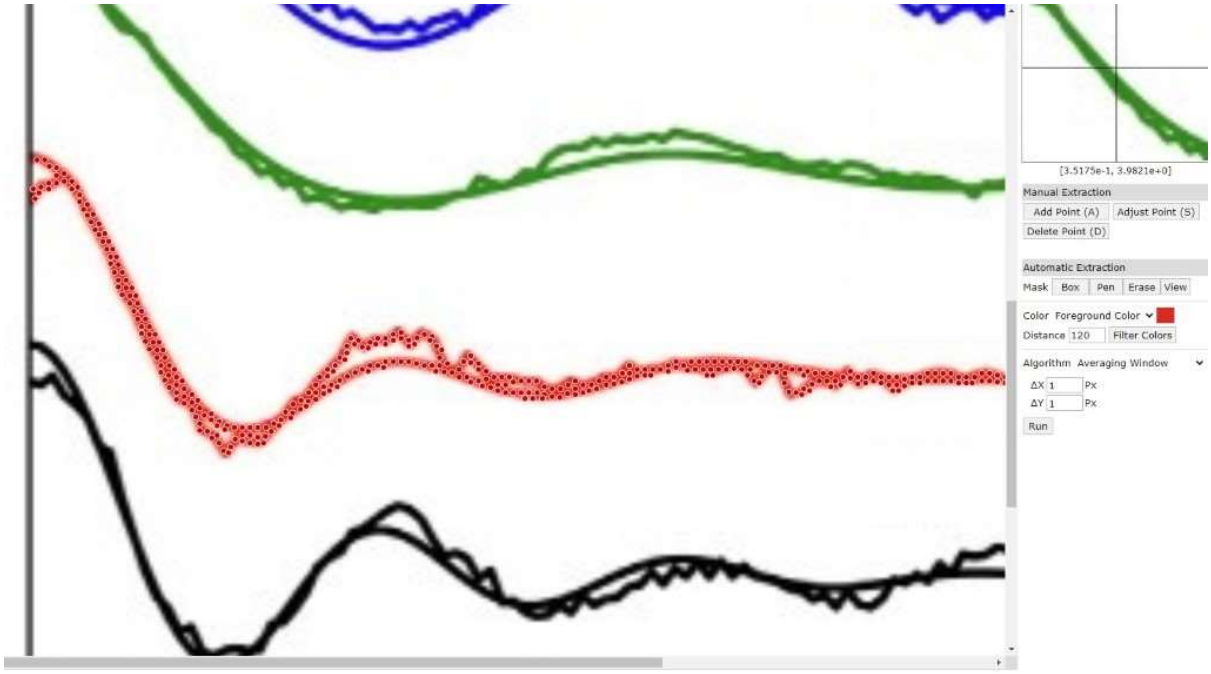

Figure 4: Screen capture of WebPlotDigitizer[10] showing the individual points after the red curve was isolated and selected. The individual points can be manually deleted or moved, allowing manual corrections to the curve detection algorithm for better data collection accuracy.

The amplitude axes from these papers are arbitrary, as only the location of the distance distribution peak is relevant. This meant the data was scraped using an arbitrary amplitude. The amplitudes were normalised such that 0 was the point at which the DEER traces converge by the end of their time sequence – see Figure 6.

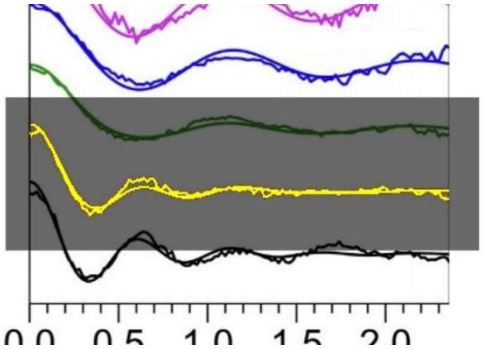

Figure 5: Screen capture of WebPlotDigitizer[10]. The image shows the colour yellow highlighted – this indicates that the data is being taken from the yellow curve.

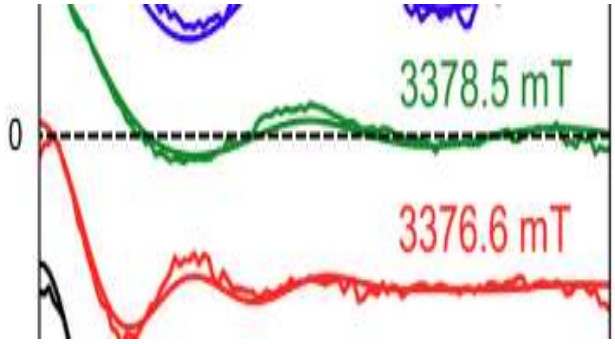

Figure 6: The DEER traces seen in Figure 3 with the zero-line drawn onto the green trace at the point where the trace converges. This was the amplitude that was taken as 0 for the purposes of scraping the data from the papers.

The resultant scraped data from this process, especially where manual point selection was necessary, produced data with non-uniform point distribution. To allow the data to be analysed systematically, it needed to be reorganised into a uniform distribution. This was accomplished using a cubic spline script written in Python. A cubic spline is a spline constructed using piecewise third-

order polynomials which pass through a certain number of control points[11] – detailed research into this process was unnecessary as Python has an inbuilt cubic spline method. The spline code produces data files for use in later analysis – it produces a point every 0.001 milliseconds, with the number of points and maximum/minimum time values from the scraping process varying in length depending on the paper. Later, this quantity of points was found to be unnecessary, so a function was written with the purpose of removing every 'i'-th row from the data matrix.

Performing Principal Component Analysis

Dr Alexey Potapov, the author of the paper in reference [3], provided the program used in producing that paper. The data scraped from this paper will be referred to as dataset 1. The program contained several algorithms that were used in this experiment, the first of which being the PCA section of the code. The code was initially configured to work with only the 6 traces produced in Figure 3. To allow this code to be used for other scraped datasets, modifications had to be made such that it accepted as many input traces as were required with each given dataset.

One change that had to be applied to the code was to remove the modulation depth, λ, values that were hardcoded into the algorithm. Each trace had an associated λ and λ-theoretical value, stored in an array of length 6 – the number of traces in the original DEER set for which this program was written to analyse. The algorithm then used these values, along with normalisation parameters, to produce the background-corrected set of DEER traces shown in Figure 3. However, the data scraped for this paper was scraped from the graph shown in Figure 3, hence this algorithm applied to the scraped data will re-apply the same normalisation and background corrections, causing the plot to be erroneous. For this reason, the λ values were omitted from the reworked program and taken to be 1. It was assumed that the data scraped from other papers was already in this normalised format, likely also corrected for background noise if applicable in the context.

The code also contained another constant, the regularisation parameter η, seen in equation (7). This parameter should be fine-tuned for each dataset to ensure correct reconstruction of the distance distributions; however, the literature does not state this value. For the purposes of this experiment, this value was taken to be 1 throughout.

After reworking the program, it was able to reproduce similar DEER trace plots to those given in the literature. Figure 7 shows the graph produced for dataset 1[3].

For dataset 1, as was the case for some of the papers that were analysed, one of the traces (shown in Figure 7 as the yellow trace) tended to infinity at the end of the trace. This is likely due to a limitation in the inbuilt Python cubic spline method. This is discussed in the error analysis subsection of the results section. Other than this deviation, by inspection Figure 7 appears to line up well with the initial data displayed in Figure 3, indicating that the data scraping method and cubic spline script appear to be working correctly.

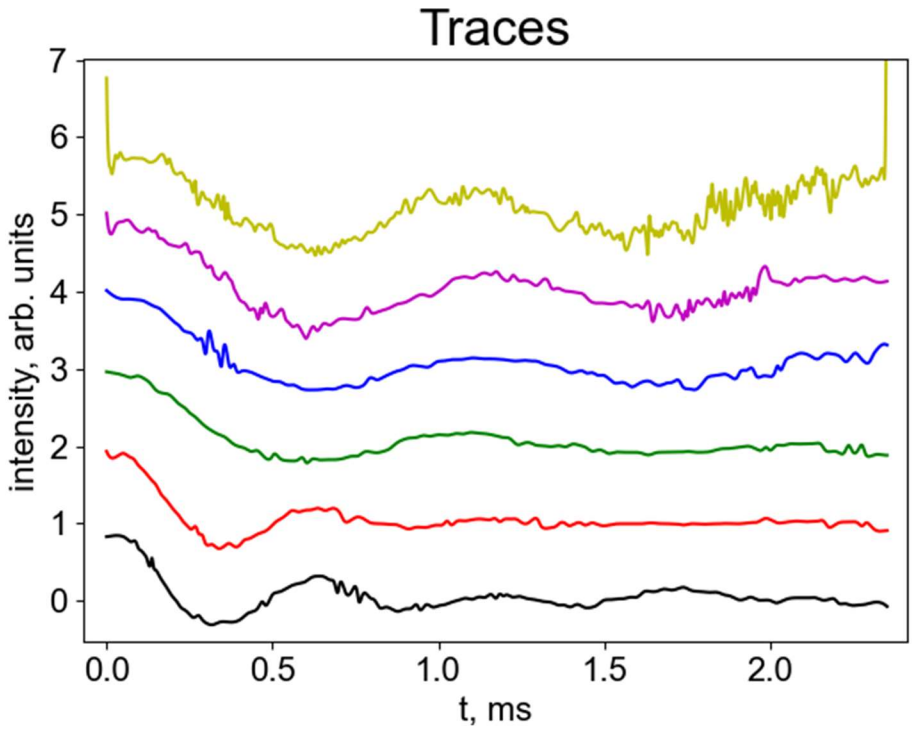

Figure 7: The scraped data in dataset 1, plotted after having been run through the cubic spline script. Parts of the graph tending to infinity can be seen in the yellow trace, but aside from this it does appear to match up to the original plot of these traces seen in Figure 3. For the purposes of producing this plot, the code was configured to colour the traces such that they match those from the source paper.

The next step was to perform the principal component analysis on the scraped data. The code functions using a trial-and-error method – the code takes an input (n) of a guessed number of principal component curves, and uses this number to compare the traces against the n-th most descriptive principal component curves that can be used to build up the traces. The summation of these principal curves is then plotted overlaid with the DEER traces to check how correlated these curves are to the data by eye, and to see how well they can make up this data. All principal component curves are also plotted by themselves, so that inspection can reveal which ones likely contribute the most to the build-up of the DEER traces. This process is then repeated for increasing n manually until the minimum number of principal components is found that appears to qualitatively represent the traces. A graph of these principal components for dataset 1 can be seen in Figure 8 (left), plotted top to bottom in increasing order of contribution to the DEER traces. A secondary section of the script was used to first centre the data by the mean amplitude value, shown in Figure 8 (right).

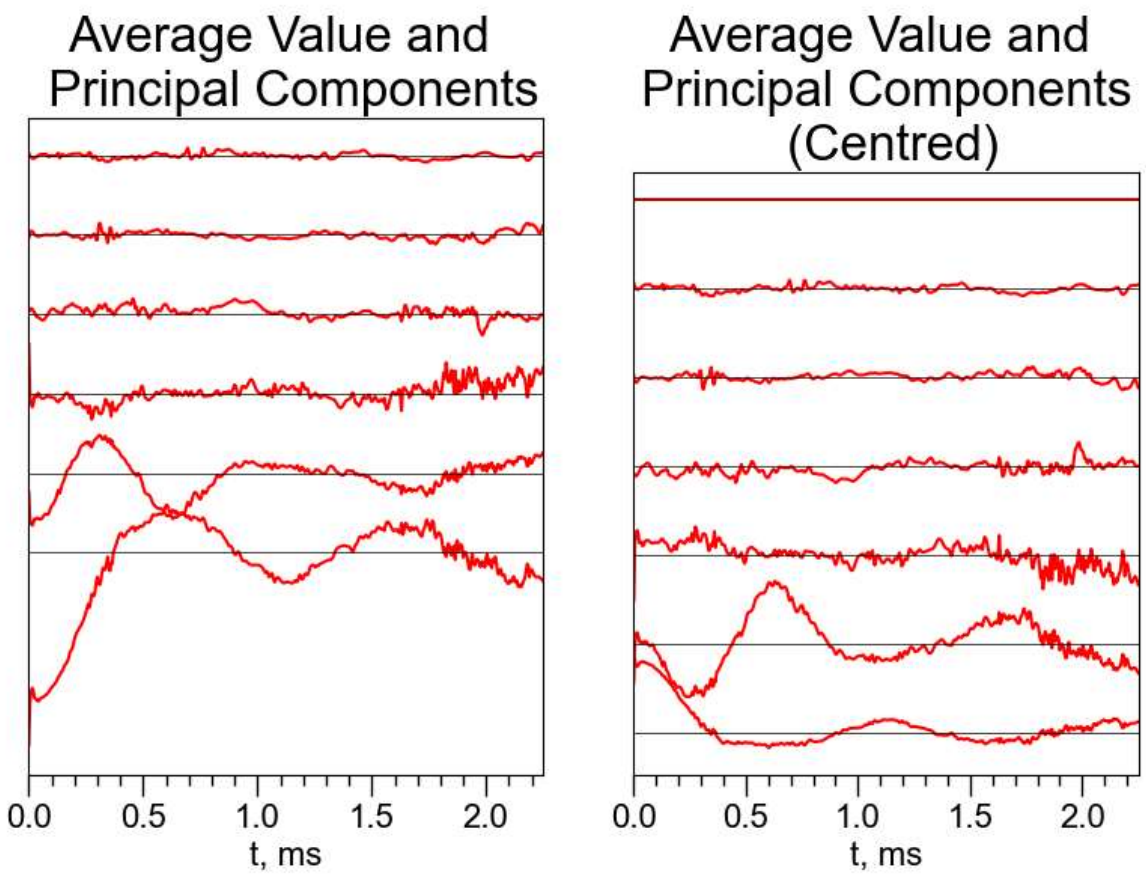

Figure 8: (Left) The average value (black lines) and principal components (red lines) plotted against time. (Right) The centred average values and principal components plotted against time. By eye, from these graphs, the bottom 3 principal components have the most effect on the DEER traces, having the most amplitude deviations.

Finding Distance Distributions

Now that PCA could be performed on all scraped datasets, the distance distributions for all datasets needed to be plotted to find the distances between the coupled electron spins.

To achieve this, the number of relevant MPP degrees (corresponding to the number of principal components) had to be parsed through equations (5) to (7), finding the MPP weights and using these to build kernel functions for use in an iterative Tikhonov algorithm as laid out in the theory. This portion of the code was once again supplied by Dr Potapov in its specific 6-trace version and modified for any number of principal curve and trace inputs.

The distance function, $f(r)$, begins as a gaussian distribution of a specified distance, with this distance (the range of r) being the sensitive DEER radius. The algorithm is then run for 1000 iterations, effectively until the convergence of the Tikhonov regularisation given by equation (7). This code produces 3 informative plots. One is of the full kernel overlaid against the full trace (each DEER trace normalised and placed end-to-end) so that a comparison can be drawn between the two to verify the accuracy of the Tikhonov algorithm. The next is the plot of normalised r; the difference

between the experimental data and the fit calculated with a certain $f(r)$, against the number of iterations. Its convergence to some value is intended to verify that the algorithm is iterating correctly. The final plot is the distance distribution, the tool with which the distance between the electron spins will be found. These graphs, for dataset 1, can be seen in Figure 10, Figure 11 and Figure 12.

The final section of the MPP weight/distance distribution code is the distance calculation. This is performed by finding the distance which corresponds with the maximum peak of the distance distribution, $f(r)$.

Results

Verifying Functionality

An important test is the outcome of running the scripts on a dataset with known results to verify functionality. The most logical choice was dataset 1[3] - not only is this the only paper that utilises principal component analysis, but parts of the code used in the method of this project were modified versions of the script used to produce the graphs and results shown in that paper. According to the paper, these traces should be comprised primarily of 3 principal components[3]. This matches with the principal component curves displayed in Figure 8, indicating that the PCA section of the code works correctly. Figure 9 shows the resultant principal curve-DEER trace overlay graphs produced using dataset 1[3] with 1, 2, 3 and 4 principal components.

It can be seen from these graphs that the more principal components that are used to recreate the DEER traces, the more accurate the construction is. This is, until it reaches 4 components – at this point it appears that the improvement on the construction of the DEER traces is negligible. This indicates that the correct number of principal component curves has been found – 3 – which agrees with reference [3].

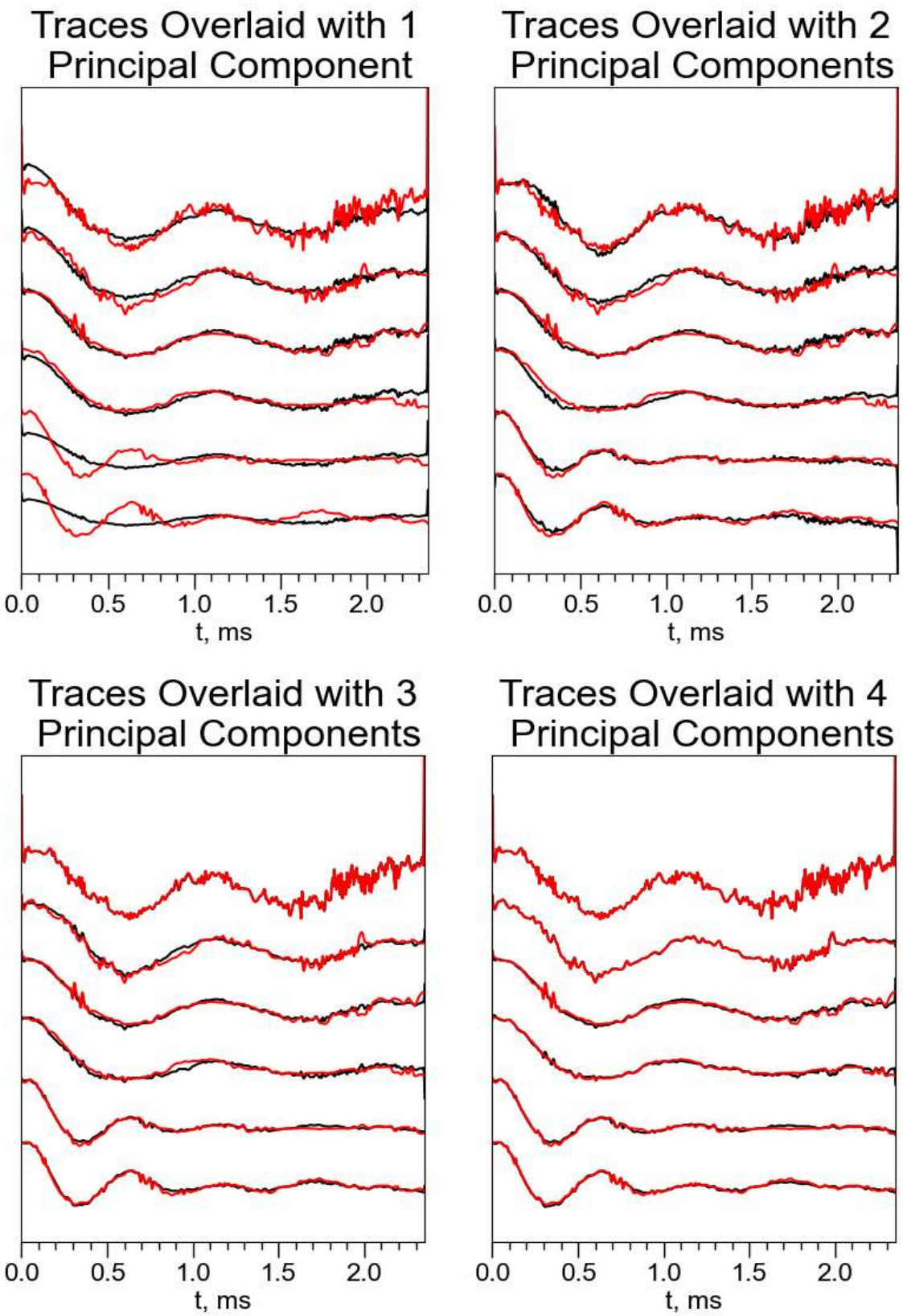

Figure 9: Varying numbers of principal component curves plotted overlaid with the combined DEER traces using dataset 1[3]. These graphs show that as the number of principal components increases, the principal curve combination better represents the data. The data can be seen to have only 3 principal curves, as the improvement towards the DEER traces between 3 components and 4 is negligible, hence it appears here that only 3 principal components are relevant.

Now that the PCA script appeared to function correctly, the MPP weight code used to calculate the distance distribution and hence the distance also needed to be tested. The plot of normalised *r*, created using dataset 1[3], seen in Figure 10 can be seen to converge to a value around 8.4 as the number of iterations moves towards its contextually defined value, 1000. This shows that the code is functioning iteratively as intended.

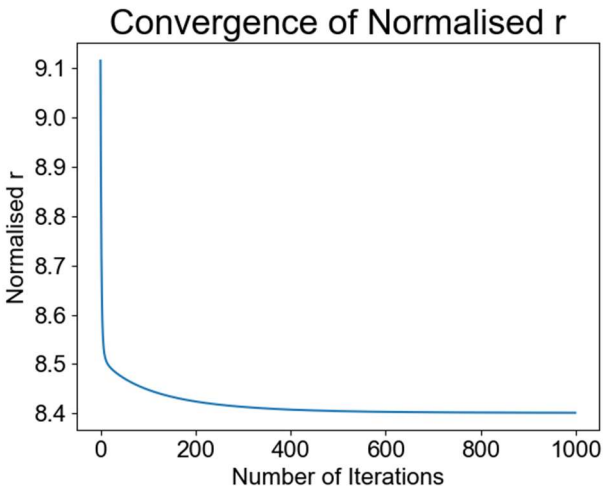

Figure 10: Convergence of the value of normalised *r*. The graph shows that this value converges to somewhere around 8.4, indicating the distance distribution code is functioning as intended.

The next plot produced is the full post-Tikhonov kernel overlaid onto the DEER traces, both placed end-to-end. This shows how closely the result of the iteration algorithm, accounting for the principal components, builds up the DEER traces. This graph for dataset 1[3] can be seen in Figure 11.

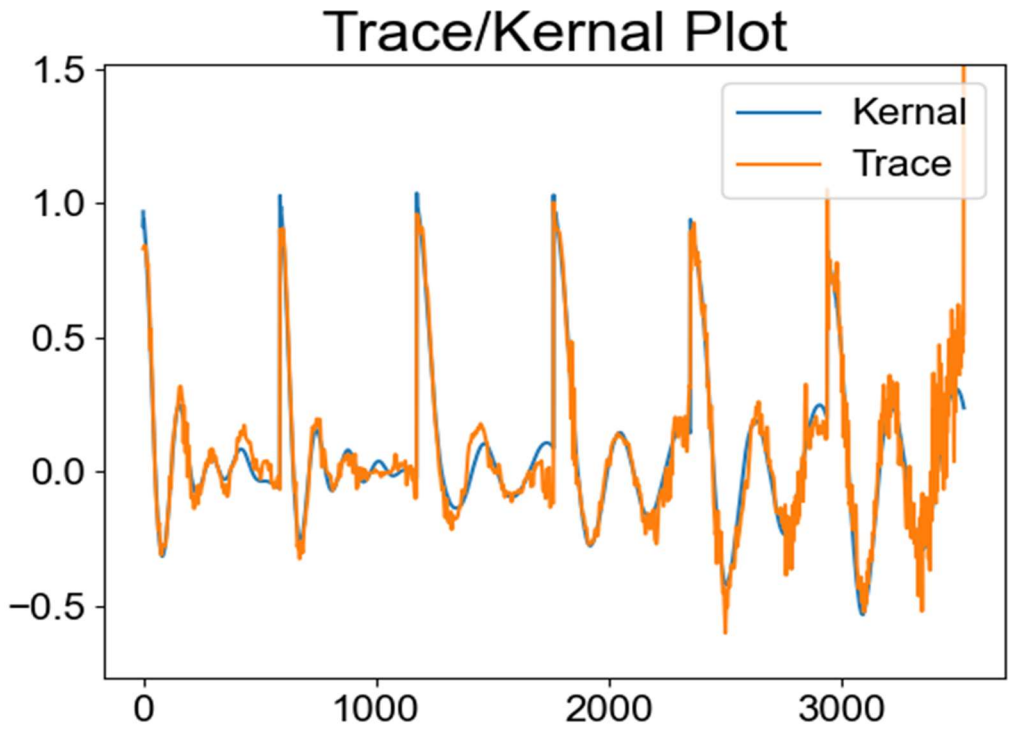

Figure 11: Plot of the full kernel overlaid on top of the DEER traces, both placed end-to-end. The kernel appears to match the trace closely, indicating the iterative Tikhonov regularisation function is building an accurate model of the traces from the meaningful principal component curves. The effect of one of the traces tending to infinity at its end and thus deviating from the kernel can be seen on the right-hand side of this graph – this is expected based on the trace plot seen in Figure 7.

The final plot produced is the distance distribution itself. This can be seen placed side-by-side with the distance distribution taken from the paper in reference [3], see Figure 12. It appears that the maximum peaks in the distance distribution function lie at similar values of r, with some other matching taller peaks also appearing. There are some small deviations between the distance distribution which could be potential noise – this will be discussed in the error analysis subsection.

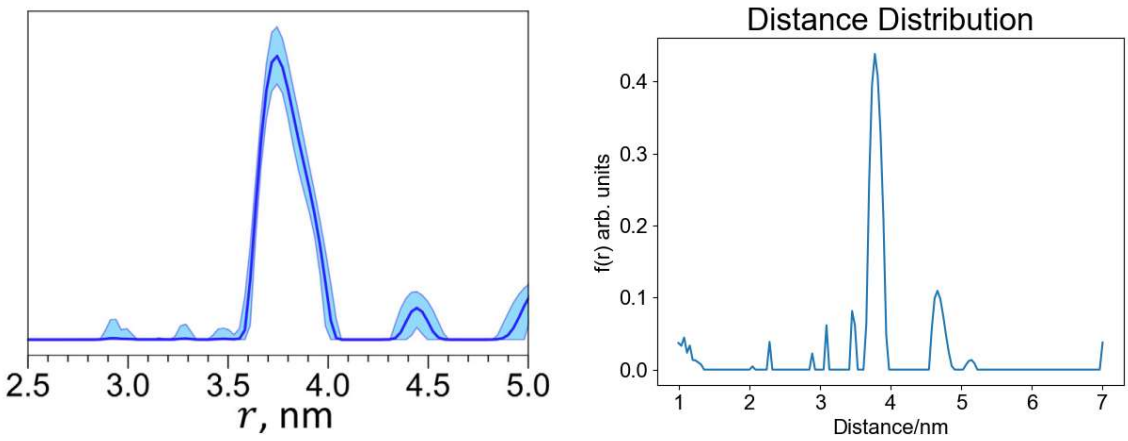

Figure 12: Comparison of distance distribution taken from the paper in reference [3] from *r* = 2.5nm to 5nm (left) against the distance distribution produced by the MPP weight script using data scraped from this paper from *r* = 1nm to 7nm (right). The two distance distributions appear similar, with the maximum peak appearing between *r* = 3.5nm and 4nm, an additional peak at between *r* = 4nm and 5nm and potentially another matching peak at around *r* = 5nm, but this is partially cut from the right-hand side of the plot taken from the reference paper.

These tests appear to conclude that the scripts written to process the data function correctly, with the principal component analysis, iterative Tikhonov regularisation algorithm, kernel/DEER trace comparison and distance distribution producing close to expected results for dataset 1[3].

Data Analysis Results

The results for the analysis of all the datasets can be seen in table 1. All scraped datasets were parsed through the same algorithms as dataset 1. The magnetic field frequencies/strengths were converted from the literature given values into bands using the table in reference [12].

Table 1:

Data Set Number	Data Source	Evaluated Distance (*r*)/nm	Source Literature Distance/nm	Number of DEER Traces	Number of Principal Curves	Biradical Species	Field Frequency Band
1	Reference [3], Figure 3c	3.77	3.75±0.13	6	3	Nitroxide (Non-rigid)	W
2	Reference [13], Figure 2a	3.28	3.25	13	9	Tyrosyl (Rigid)	W+ (very high-field, ~6T)
3	Reference [14], Figure 2a	2.45	3.5±0.1	5	4	$Copper^{2+}$ (Rigid)	X
4	Reference [14], Figure 3a	2.17	2.2±0.2	4	3	$Copper^{2+}$	X
5	Reference [15], Figure 3	2.88	2.2±0.3	4	4	$Copper^{2}$	X
6	Reference [16], Figure 3	3.57	3.62±0.05	5	3	$Copper^{2}$ (Rigid)	X
7	Reference [17], Figure 4	2.65	2.65±0.1	5	2	Nitroxide	X
8	Reference [18], Figure 4	2.40	2.55±0.5	17	5	Nitroxide (Rigid)	Q
9	Reference [19], Figure 4	2.45	1.93±0.3	6	5	Nitroxide (Non-rigid)	W
10	Reference [19], Figure 10	3.74	3.46±0.6	7	3	Nitroxide (Non-rigid)	X
11	Reference [20], Figure 4a (left)	3.05	3.1±0.3	11	9	Nitroxide (Rigid)	W (varied)
12	Reference [20], Figure 4a (right)	3.13	3.1±0.3	14	10	Nitroxide (Rigid)	W (fixed)
13	Reference [20], Figure 4b	3.44	2.8±0.2	10	4	Nitroxide (Non-rigid)	W
14	Reference [21], Figure 5	2.69	2.70	5	3	Nitroxide (Rigid)	X
15	Reference [8], Figure 9a	3.46	2.8±0.2	10	7	α-TOPP (Semi-rigid)	W (fixed)
16	Reference [8], Figure 9b	3.46	2.8±0.2	10	7	α-TOPP (Semi-rigid)	W (varied)
17	Reference [19], Figure 7	2.63	1.80±0.02	6	4	Nitroxide (Rigid)	W

Some papers did not contain experimental results for the distance values. However, excluding datasets 24 and 25, estimates for their values of *r* could be discerned using the distance distributions published in the papers. The data for these papers has been placed in table 2. The errors given in the source literature column are from the inaccuracies of reading the distance value from the graph and were chosen depending on the accuracy of discerning values from the plots in each case.

Table 2:

Data Set Number	Data Source	Evaluated Distance (*r*)/nm	Source Literature Distance/nm	Number of DEER Traces	Number of Principal Curves	Biradical Species	Field Frequency Band
18	Reference [22], Figure 3c (top)	2.1	2.3±0.2	4	2	$Copper^{2+}$ and Nitroxide (Non-rigid)	X
19	Reference [22], Figure 3c (bottom)	2.1	2.1±0.2	4	2	$Copper^{2+}$ and Nitroxide (Non-rigid)	X
20	Reference [23], Figure 4 (left)	3.3	3.3±0.2	5	2	Nitroxide (linear) (Non-rigid)	X
21	Reference [23], Figure 4 (right)	3.3	3.3±0.2	5	2	Nitroxide (bent) (Non-rigid)	X
22	Reference [24], Figure 6 (top left)	4.2	5.0±0.5	13	6	Nitroxide (Rigid)	W
23	Reference [24], Figure 6 (bottom right)	3.7	3.5±0.5	7	3	Nitroxide (Rigid)	W
24	Reference [25], Figure 4 (right)	3.4	N/A	6	3	Nitroxide (Rigid)	X
25	Reference [25], Figure 4 (left)	2.5	N/A	6	3	Nitroxide (Rigid)	X

Discussion

Error Analysis

A graphical representation of the results is shown in Figure 13. For the purposes of checking how closely our calculated values align with the literature values, a tolerance of 1σ or 5% (for the literature values that did not state an uncertainty or were not read from graphs by eye) has been applied to the data as seen in Figure 13. Datasets 24 and 25 have been omitted as their sources did not contain distance values.

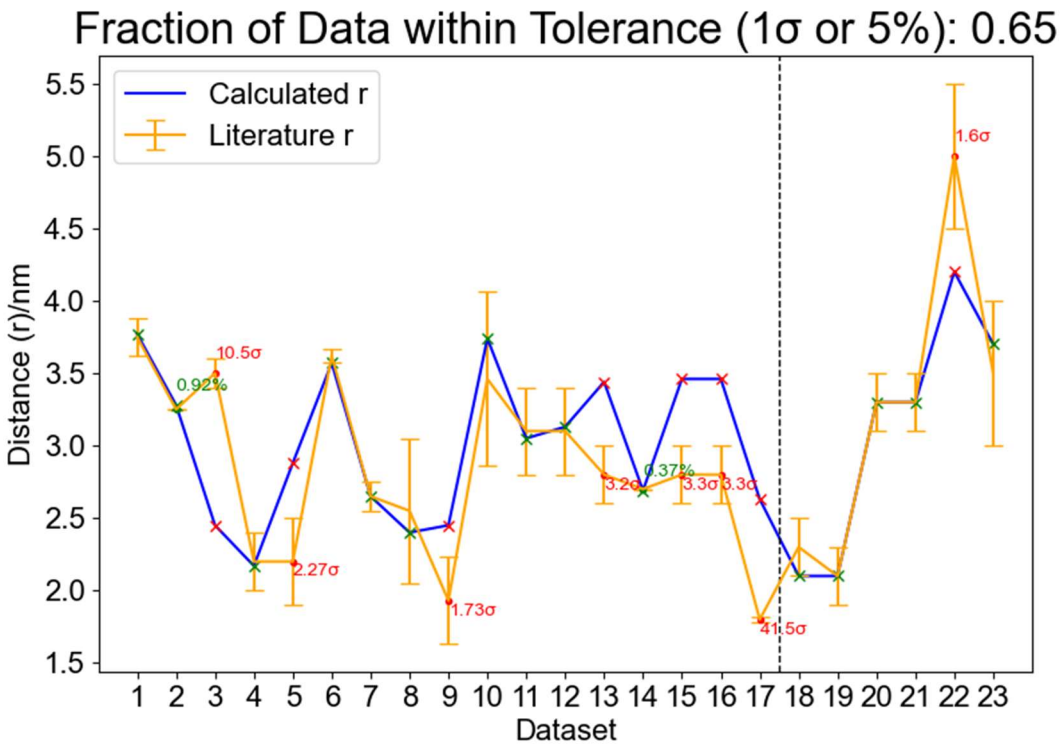

Figure 13: A comparison of the distance values taken from the referenced literature and those produced by the Tikhonov algorithm. The black line shows the separation between the data from table 1 and table 2. The green 'x' points and labels represent distance that lies within the chosen tolerance of 1σ or 5%. The red 'x' points show distance that does not lie within this tolerance. The *r*-axis in this case is effectively arbitrary and merely shows the typical distance range that DEER is used to measure.

The cubic splined DEER traces tending to infinity or negative infinity at the ends of the traces (present in the yellow trace in Figure 7) seemed to be a factor that caused issues during the production of the distance distributions. This was mitigated by removing some points from the beginning and/or end of these traces, attempting to leave the form of the traces intact but removing the rogue elements. The number of points removed was between 20 and 65 and changed depending on the number of points in that dataset. The effects of this process can be seen in Figure 14 which contains a comparison of the traces and distance distributions before and after this process for dataset 3[14].

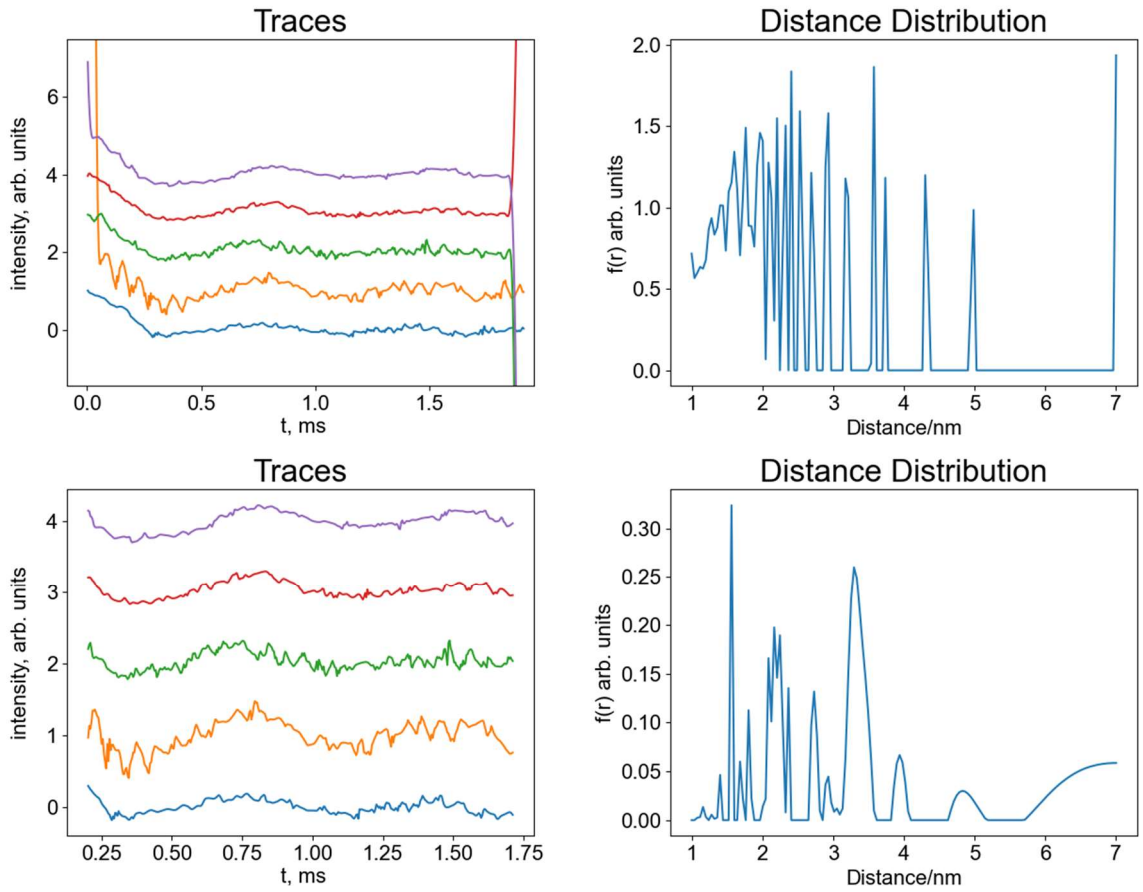

Figure 14: Comparison of distance distributions and traces for dataset 3[14] for data post-cubic spline (top) and data post-cubic spline with 50 data points cut from both ends of the traces (bottom). It is observed from these distance distributions that the sharp peaks likely associated with noise are minimised in the bottom distribution in comparison with the top distribution. This indicates, in this case, that the traces tending off to infinity likely produces an erroneous distance value. This dataset, however, does appear to be noise-dominated regardless, as the highest peak in the bottom distribution is contained within the erratic 1-3nm range containing multiple sharp spike-like peaks likely indicative of noise.

Datasets that did not require data cut-offs still produced noise in most cases. This could have come from the lack of a finely tuned regularisation parameter η (see equation (7)) not present in any of the literature data sources. This noise appears to cause erratic, spike-like peaks in the distance distributions, sometimes skewing the distance calculation as they may be higher than the correct distance peak. Inaccuracies in the scraping process could also be a factor in this noise, however the scraping process seems to produce accurate traces, implying that the noise-producing error is somewhere in the Tikhonov regularisation and formation of the distance distributions.

In the distance distribution for dataset 1[3], a small amount of what appears to be noise can be seen around the ~1nm distance range (see Figure 12). Here, the noise does not get anywhere close to influencing the highest peak and hence the calculated distance. Some datasets, such as dataset

17[19], produced distance distributions with a large amount of noise that did not appear to affect the distance result – see Figure 15. As this high noise peak is present, however, whether the maximum peak corresponds to the true distance is put into question. In the case of dataset 17[19], the measured value does not match the literature, indicating an error due to noise is a possibility.

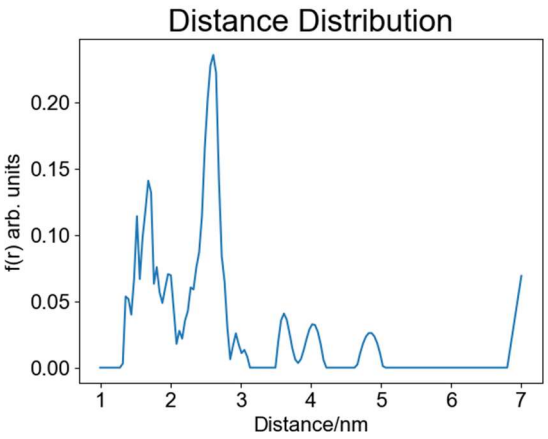

Figure 15: Distance distribution for dataset 17[19]. The graph shows two peaks, the lower of which is surrounded by sharp spike-like potential noise.

Some of the data, such as dataset 3[14], appear very noise-dominated, as can be seen from Figure 14. In this case, the tallest peak is very sharp and hence likely noise, putting the legitimacy of the other peaks into question. As can be seen in the figure, cutting some data from the beginning and/or end of the traces can change how the noise appears in the distance distribution.

Dataset 3[14] was tested with three different data cut-offs: no cut-off, 50 points from both ends, and 60 points from both ends. These tests produced three distinctly different distance calculations: 7.00nm (indicating outside the axis range), 1.56nm and 2.97nm respectively. This same test was applied to a dataset that required no cut-offs, dataset 7[17] – all producing a distance of 2.65nm, hence the cut-off produced no change in the traces or distance value. The distance distribution for dataset 7[17] can be found in Figure 16 and appears to contain only a small amount of spike-like noise. As the values for dataset 3[14] vary widely, it is likely that this result is error dominated and incorrect.

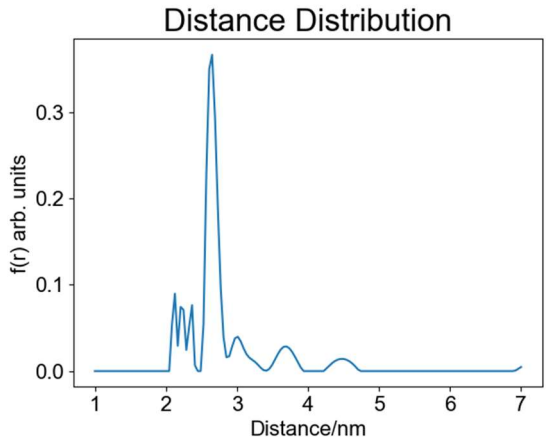

Figure 16: Distance distribution for dataset 7[17]. A small amount of sharp potential noise can be observed in the 2-2.5nm range, but aside from this the distribution appears relatively noise-free.

Some of the datasets do not match the literature values despite the fact the noise in their distance distributions appears to be minimal. This is the case for dataset 15[8], where the calculated distance is just over 3σ away from the literature quoted distance. Dataset 15's[8] distance distribution appears to have only a small amount of sharp noise, with the maximum peak appearing to look noise-free and the noise peaks not reaching an amplitude anywhere close to the maximum. No data cut-offs were applied to dataset 15[8] when analysing it but adding a 20-point beginning and end cut-off to the data appears to massively amplify the small sharp noise peaks that were present. The distance distribution comparison for cut and uncut data for this dataset can be seen in Figure 17.

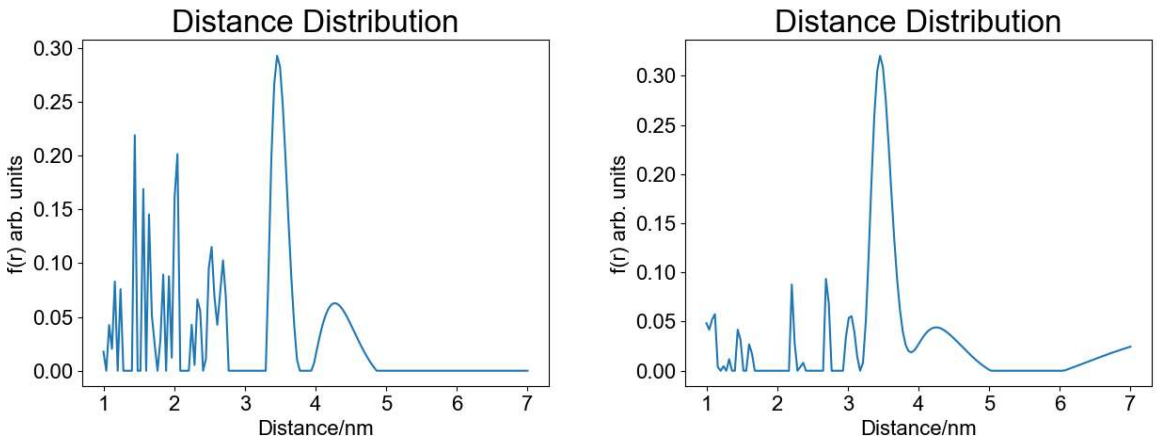

Figure 17: Distance distribution for dataset 15[8] with no data cut-offs (right) compared with the distance distribution with a cut-off of 20 points from the beginning and end of the traces (left). The 20-point cut off appears to amplify the small amount of spike-like noise present in the uncut data distance distribution.

The distances calculations for both the uncut and cut dataset 15[8] produced a value of 3.46nm for the distance despite the amplification of noise observed between 1 and 3nm. This agrees with the negligible change observed when adding a point cut-off to dataset 7[17] as discussed above, containing seemingly minimal noise by observation. The largest maxima for dataset 3[14], seen in Figure 14, shifted when the cut-off was applied, however in that case the maxima did appear to be sharp peaks suggesting noise. This could indicate that the tallest maxima and hence distance for dataset 15[8] is not influenced by noise, and the deviation from the literature value is caused by another unknown error source - from another by-product of a lack of an accurate regularisation parameter, somewhere in the analysation or scraping process, or in the literature source itself.

Analysing Rigidity and EPR Field Strength

To test the hypotheses laid out in the theory, it is first necessary to check the correlation between the magnetic field strength applied in the EPR experiments against the number of principal component curves contained within the scraped data. Figure 18 shows the relationship between the magnetic field strength applied during the experiment and the ratio of the number of principal component curves found to the number of traces. There appears to be a weak correlation between the ratio of principal component curves and the strength of the magnetic field. However, this does not give a full picture of the situation as the rigidity of the biradical molecule being measured should also have an impact if the hypothesis is correct. Higher numbers of traces should also have a greater chance of representing duplicate paths

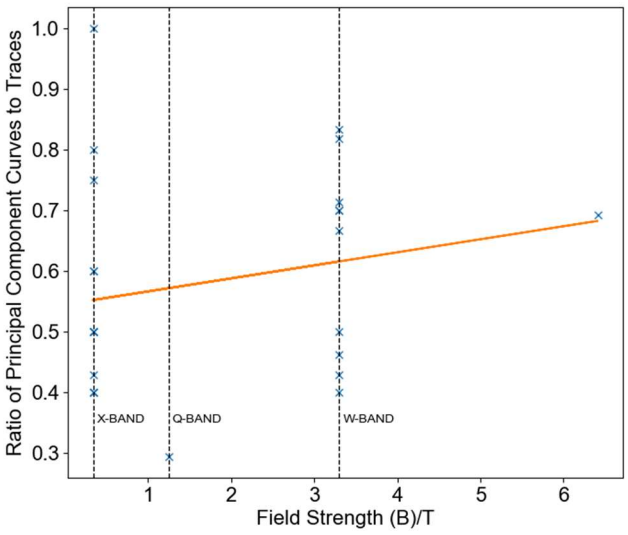

Figure 18: Graph of the ratio of principal component curves to traces against the magnetic field strength used in the EPR experiment. It can be seen from this graph that there appears to be a weak correlation.

between electron spins, meaning the traces could conform more to the same set of principal components, causing an inaccuracy in the correlation presented in the graph on Figure 18.

The next factor to consider was the correlation between the ratio of the number of principal components found to the number of traces. Some papers contained statements about the rigidity of the biradical structures, some contained structural diagrams from which the rigidity could be discerned, and some provided neither. Whether or not the biradical appears to be rigid has been stated in tables 1 and 2 – if there is no rigidity mentioned, none could be discerned from the respective literature. Figure 19 shows a comparison between two such biradical structural diagrams, pertaining to datasets 9 and 17[19]. Those datasets do not appear to agree with the hypothesis, having a principal component to trace ratio of 0.83 for the less rigid nitroxide biradical and 0.71 for the more rigid one.

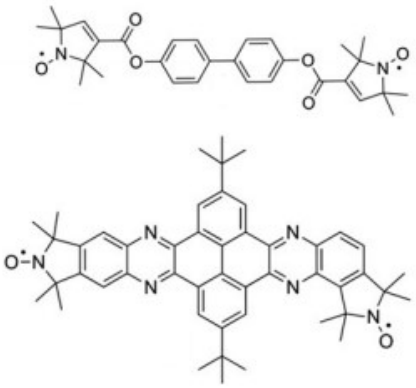

Figure 19: Example of a non-rigid nitroxide biradical (top) and a rigid nitroxide biradical (bottom)[19] corresponding to datasets 9 and 17 respectively[19].

One paper, containing datasets 11, 12 and 13[20], provided distances and data for two biradicals: a rigid biradical in an RNA duplex and a less rigid biradical in an α-helical peptide. Datasets 11 and 12 correspond to the biradical attached to the RNA duplex, and dataset 13 the peptide. The difference between datasets 11 and 12 is that 11's magnetic field was varied whereas 12's was not[20]. Diagrams showing these structures can be seen in Figure 20.

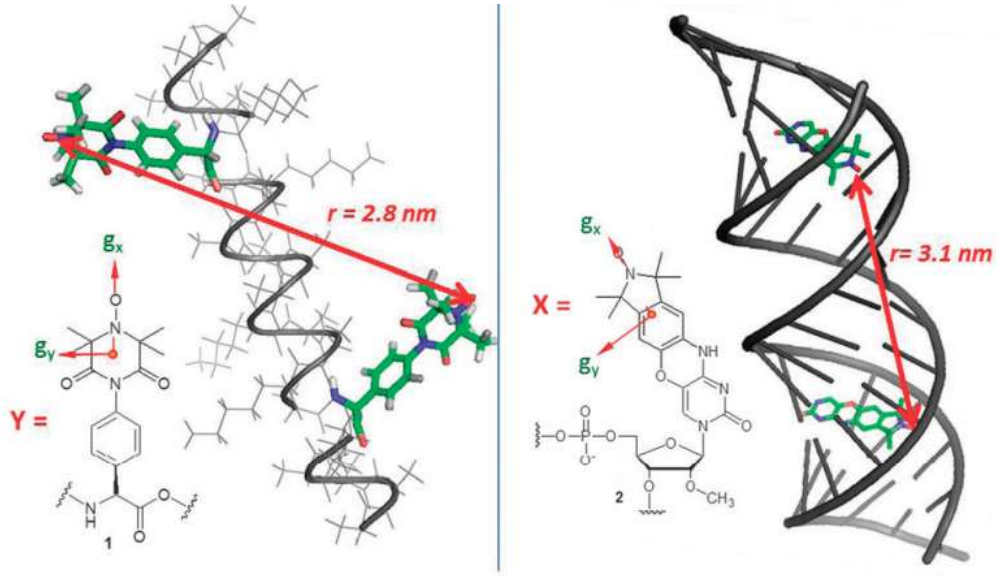

Figure 20: Structural diagrams and distances for the RNA duplex (right) corresponding to datasets 11 and 12 and the α-helical peptide (left) corresponding to dataset 13[20]. The rigidity of the two biradical structures can be seen in the small diagrams to the lower left of the two images, with the peptide's biradical having more points of potential movement and hence less rigidity than the RNA duplex's biradical.

The calculated distances for datasets 11 and 12[20] were both within 1σ of the literature value, however the distance calculated for dataset 13[20] was quite far from the literature value. All three distance distributions contained some noise between 1 and 3nm, so the distance calculated for

dataset 13[20] could be erroneous. If the noise has indeed come from the lack of an accurate regularisation parameter, however, the errors in the distance distribution should not be carried through to the traces themselves and hence the principal components. This means this noise should not have affected the ratios. The distance distributions for these three datasets can be found in Figure 21. The ratio of principal components to traces for these datasets (0.82, 0.71 and 0.40 for datasets 11, 12 and 13[20] respectively) follows the hypothesis presented in the theory.

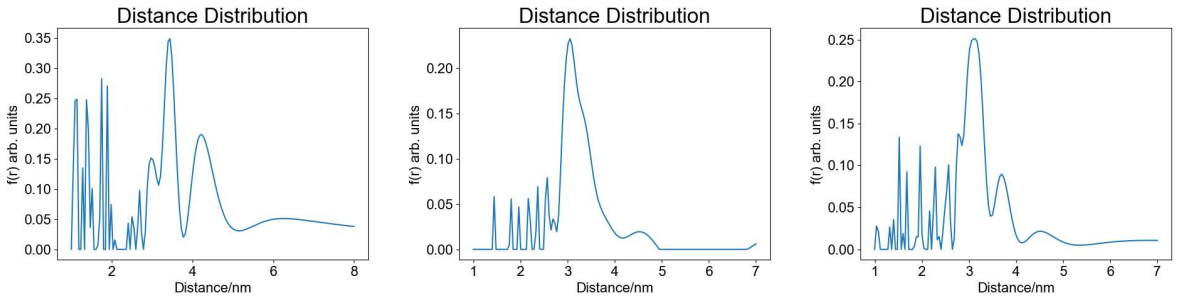

Figure 21: Comparison of the distance distributions for dataset 13's α-helical peptide's biradical (left), and 11 (centre) and 12 (right)'s RNA duplex's biradical[20]. A large amount of sharp potential noise can be found in the 1 to 3 nm range in all three cases.

This process of comparing rigidity with the ratio of principal components to traces was performed for the remaining datasets with discernible rigidity. The semi-rigid biradical structures were considered rigid in this case. The average ratio for the non-rigid datasets was 0.495, and the average ratio for the rigid datasets was 0.604. This shows a loose correlation between the ratio of principal component curves to traces and the rigidity of the biradical structure. This is not the only factor as discussed, with the magnetic field strength and potentially duplicate paths between electron spins also influencing the ratio. This looks very possible in the case of dataset 8[18], which contains a high number of traces in comparison to the other datasets with 17 traces and a very low ratio of principal components to traces of 0.29, meaning it could be an outlier. Dataset 22 is another possible outlier, containing a high number of traces at 13 and a low (0.46) principal component to trace ratio. There is no way of knowing for sure, however, that these results are outliers with the data analysation techniques used.

Potential Improvements

To improve this analysis and minimise error, more datasets should be analysed. It would be ideal to compare the ratios of principal component curves to traces for each field frequency band for a better comparison. However, if this were done with these datasets there would not be enough data for each band to reach a meaningful conclusion. More research could also be done into the structure of the biradicals to better consider potential duplicate paths between electron spins, potentially allowing the proper identification of suspected outlier results.

To improve the accuracy of the individual distance distributions and distance calculations, the regularisation parameter η (equation (7)) should be fine-tuned to each dataset. This was not possible by looking at the source literature, as it is not stated in any of the papers analysed. A thorough error analysis could also be performed, as if additional error bars were present on Figure

13 on the blue (calculated) line they may have overlapped the orange (literature) error bars, allowing what appears to be incorrect data to fit within a reasonable error tolerance.

The parts of the traces that tend to infinity in the cubic spline script could also have caused errors, forcing the data points to be cut off from the beginning and/or end to produce analysable data – further research into the workings of the inbuilt Python cubic spline method could potentially reveal a solution to this issue. Another option could be the use of a different, more sophisticated interpolative algorithm for this purpose.

The data-scraping process should contain minimal errors as it produces a visibly similar version of the graph scraped from the literature, however it could be potentially fine-tuned if an automatic image processing algorithm were purpose-written to minimise any potential human error.

Conclusion

The calculated distance values produced using the Python program appeared to roughly match the values outlined in the source papers with 65% of the calculated data agreeing (see Figure 13). It is possible that the regularisation parameter η being approximated to 1 caused this percentage to be lower than it could have been - given this, the percentage seems a reasonable match for the expected values.

The fact the scraped traces appear to be accurate to the source literature traces also implies that, the number of principal component curves found should be also be correct. This is under the assumption that the PCA by SVD Python algorithm functions correctly, which it appears to by the tests run during this experiment. For this reason, and looking at the correlation shown in Figure 18, the number of principal components that are significant in constructing the traces as a whole is proportional to the strength of the magnetic field; agreeing with the hypothesis laid out in the theory. The average ratios of principal component curves to traces for rigid and non-rigid biradical-chemical structures of 0.495 and 0.604 respectively also agrees with the hypothesis to the extent of the data analysed. This implies that there is a correlation between the quantity of significant principal components and the rigidity of the biradical molecules being observed. In both cases a stronger correlation would be ideal to better prove the hypothesis. To improve the investigation more datasets should be analysed, and the two correlations combined into a definitive, rigorous result agreeing with both parts of the hypothesis. The lack of an accurate regularisation parameter should not have produced any errors in this part of the project.

This is a promising result, revealing another method of measuring the rigidity of molecules using EPR. Use of PCA in future DEER experiments could prove useful, in removing non-meaningful MPP degrees and improving measurement accuracy. It could also improve processing speed for more in-depth analyses, especially where large amounts of DEER traces are present or data needs to be parsed through long, complex algorithms.

Notes

Some of the data was re-analysed when writing this report, and so some of the values in tables 1 and 2 differ slightly from the values calculated in the project notebook. This is due to changes in the plotting parameters and cut-offs applied to the data – the scraped data remains the same as that which was used in the notebook.

References

[1] Takahashi, S., Brunel, LC., Edwards, D. et al. "Pulsed electron paramagnetic resonance spectroscopy powered by a free-electron laser", 2012, https://doi.org/10.1038/nature11437

[2] Indra D. Sahu, Gary A. Lorigan. Encyclopedia of Analytical Science (Third Edition), 2019.

[3] Alexey Potapov, "Application of spherical harmonics for DEER data analysis in systems with a conformational distribution", 2020, https://doi.org/10.1016/j.jmr.2020.106769.

[4] Gunnar Jeschke, "DEER Distance Measurements on Proteins", 2012, https://doi.org/10.1146/annurev-physchem-032511-143716

[5] E. L. Hahn, Spin echoes, 1950, 10.1103/PhysRev.80.580

[6] H. Y. Carr, E. M. Purcell, Effects of Diffusion on Free Precession in Nuclear Magnetic Resonance Experiments, 1954, 10.1103/PhysRev.94.630

[7] Ian T. Jolliffe and Jorge Cadima, "Principal component analysis: a review and recent developments", 2016, http://doi.org/10.1098/rsta.2015.0202

[8] I. Tkach, U. Diederichsen, M. Bennati, "Studies of transmembrane peptides by pulse dipolar spectroscopy with semi-rigid TOPP spin labels". *Eur Biophys* J 50, 143–157 (2021). https://doi.org/10.1007/s00249-021-01508-6

[9] A Tutorial on Principal Component Analysis, Jonathon Shlens, *Google Research*, April 7, 2014; Version 3.02

[10] WebPlotDigitizer, https://automeris.io/WebPlotDigitizer/, Accessed 19th February 2021.

[11] Wolfram MathWorld: Cubic Spline, https://mathworld.wolfram.com/CubicSpline.html#:~:text=A%20cubic%20spline%20is%20a,equations, Accessed 4th of May 2020

[12] EPR: Theory. (2020, August 15). Retrieved May 4, 2021, from https://chem.libretexts.org/@go/page/1794

[13] Vasyl P. Denysenkov, Daniele Biglino, Wolfgang Lubitz, Thomas F. Prisner, Marina Bennati, "Structure of the Tyrosyl Biradical in Mouse R2 Ribonucleotide from High-Field PELDOR", https://doi.org/10.1002/anie.200703753

[14] Zhongyu Yang, Michael R. Kurpiewski, Ming Ji, Jacque E. Townsend, Preeti Mehta, Linda Jen-Jacobson, Sunil Saxena, "ESR spectroscopy identifies inhibitory Cu^{2+} sites in a DNA-modifying enzyme to reveal determinants of catalytic specificity", Proceedings of the National Academy of Sciences, 2012, 10.1073/pnas.1200733109

[15] Yang Z, Becker J, Saxena S. “On Cu(II)-Cu(II) distance measurements using pulsed electron electron double resonance”. J Magn Reason, 2007, doi: 10.1016/j.jmr.2007.08.006

[16] Alice M. Bowen, Michael W. Jones, Janet E. Lovett, Thembanikosi G. Gayle, Michael J. McPherson, Jonathan R. Dilworth, Christiane R. Timmel, Jeffrey R. Harmer, “Exploiting orientation-selective DEER: determining molecular structure in systems containing Cu(II) centres”, 2016

[17] Christoph Abé, Daniel Klose, Franziska Dietrich, Wolfgang H. Ziegler, Yevhen Polyhach, Gunnar Jeschke, Heinz-Jürgen Steinhoff, “Orientation selective DEER measurements on vinculin tail at X-band frequencies reveal spin label orientations”, https://doi.org/10.1016/j.jmr.2011.12.024

[18] Austin Gamble Jarvi, Kalina Ranguelova, Shreya Ghosh, Ralph T. Weber, Sunil Saxena, “On the Use of Q-Band Double Electron-Electron Resonance To Resolve the Relative Orientations of Two Double Histidine-Bound Cu2+ Ions in a Protein”, doi: 10.1021/acs.jpcb.8b07727

[19] Dinar Abdullin, Gregor Hagelueken, Robert I. Hunter, Graham M. Smith, Olav Schiemann, “Geometric model-based fitting algorithm for orientation-selective PELDOR data”, Molecular Physics, 2015, doi: 10.1080/00268976.2014.960494

[20] Igor Tkach, Soraya Pornsuwan, Claudia Höbartner, Falk Wachowius, Snorri Th. Sigurdsson, Tatiana Y. Baranova, Ulf Diederichsen, Giuseppe Sicoli, Marina Bennati, “Orientation selection in distance measurements between nitroxide spin labels at 94 GHz EPR with variable dual frequency irradiation”, 2013, http://dx.doi.org/10.1039/C3CP44415E

[21] Andriy Marko, Thomas F. Prisner, “An algorithm to analyze PELDOR data of rigid spin label pairs”, 2013, http://dx.doi.org/10.1039/C2CP42942J

[22] D. Abdullin, G. Hagelueken, O. Schiemann, “Determination of nitroxide spin label conformations via PELDOR any X-ray crystallography”, 2016

[23] A. Marko, D. Margraf, H. Yu, Y. Mu, G. Stock, and T. Prisner , "Molecular orientation studies by pulsed electron-electron double resonance experiments", 2009, https://doi.org/10.1063/1.3073040

[24] M. A. Stevens, J. E. McKay, J. L. S. Robinson, H. EL Mkami, G. M. Smith, D. G. Norman, “The use of the Rx spin label in orientation measurement on proteins by EPR”, 2016, http://dx.doi.org/10.1039/C5CP04753F

[25] Andriy Marko, Dominik Margraf, Pavol Cekan, Snorri Th. Sigurdsson, Olav Schiemann, Thomas F. Prisner, “Analytical method to determine the orientation of rigid spin labels in DNA”, 2010, 10.1103/PhysRevE.81.021911