DeerAnalysis User Manual - Version 2018

G. Jeschke, ETH Zürich; gjeschke@ethz.ch
S. Stoll, University of Washington, Seattle; stst@uw.edu
http://www.epr.ethz.ch/software/index

July 20, 2018

1 Purpose of the program

The program DeerAnalysis can extract distance distributions from dead-time free pulse ELDOR data (constant-time and variable-time four-pulse DEER) [4, 5]. Furthermore, it can be used for direct comparison of primary data of similar samples [6]. With some caution [7] the program may also be applied to the analysis of dead-time free double-quantum coherence EPR experiments [8], SIFTER experiments [25], RIDME, [26, 27] experiments on spins $S = 1/2$. For RIDME experiments on spins $S > 1/2$, the companion program OvertoneAnalysis [28] should be employed. It should not be used for data from experiments that have a significant dead time,

$$t_d > 2r^3 \text{ns nm}^{-3}, \tag{1}$$

where r is the shortest distance in the distribution.

Except for fits of two user-defined models for three-spin systems (Section 7.6.1), the program assumes isolated spin pairs. If more than two spins are coupled (e.g. spin-labeled protein oligomers), distance distributions are only approximate with artifact contributions at both short and long distances [1]. DeerAnalysis allows for an approximate correction of these contributions by the 'ghost suppression' feature. The program also assumes that the spin-spin vectors of excited spin pairs are uniformly distributed with respect to the magnetic field (no orientation selection) and that all dipolar frequencies can be excited by the pump and observer pulses.

The underlying mathematical problem is (moderately) ill-posed, i.e., quality of the analyzed data is very crucial. Pre-processing tools are implemented to correct for experimental imperfections (phase errors, displacements of the time origin of the modulation) and to separate the intramolecular distances of interest from the intermolecular background contribution. Furthermore, the program provides several independent approaches for extracting the distance distribution, which helps to get a feeling for the reliability of the distribution. Characterization of the distance distribution in terms of its mean value $\langle r \rangle$ and width (standard deviation σ_r) is usually reliable [6] and is therefore a standard output. The performance of the different approaches for data analysis depends on the type of distance distribution (narrow or broad peaks or both) and was discussed in some detail in Ref. [6].

DeerAnalysis features a module for validation of distance distributions obtained by Tikhonov regularization. With this module a systematic error analysis can be performed that may consider experimental noise, uncertainties in background correction for a given dimensionality of the background and uncertainties in the dimensionality of the background. This analysis can provide error bars for the points in the distance distribution. Without such Monte Carlo validation, realistic error bars are not accessible. Display of error bars without validation that existed in earlier versions has been disabled.

With DEERNet (neural network) processing (Section 7.5), uncertainty estimates are obtained by using an ensemble of independently trained neural networks.

DeerAnalysis is based on experience with earlier programs DeerFit and DeerTrafo as well as with model-specific fitting of data [9, 11, 12]. It supersedes these earlier programs with respect to reliability and functionality. At the present time, DeerAnalysis is released only as source code that can be run within MATLAB (versions 2014b or newer) but not as a stand-alone application. In

contrast to DeerAnalysis2004-2015, DeerAnalysis2018 no longer relies on calling the external compiled Fortran program FTIKREG. With older Matlab versions, you can use previous versions of DeerAnalysis that are still available at www.epr.ethz.ch/software/deer-analysis-older-version.html.

2   Citations

If you use DeerAnalysis2018 in your research, please cite [2]:

- G. Jeschke, V. Chechik, P. Ionita, A. Godt, H. Zimmermann, J. Banham, C. R. Timmel, D. Hilger, H. Jung, *Appl. Magn. Reson.* **2006**, 30, 473–498.

If you use the feature of constrained fits, please read `README_DeerAnalysis2011con.pdf` and cite [3]:

- O. Dalmas, H. C. Hyde, R. E. Hulse, E. Perozo. *J. Am. Chem. Soc.* **2012**, 134, 16360–16369.

If you use ghost suppression for multiple-spin systems, please cite [13]:

- T. von Hagens, Y. Polyhach, M. Sajid, A. Godt, G. Jeschke. *Phys. Chem. Chem. Phys.* **15** (2013) 5854–5866.

If you use the AIC or GCV criterion to determine the regularization parameter, please cite [24]

- T. H. Edwards, S. Stoll. *J. Magn. Reson.* **2018**, 288, 58-68.

If you use DEERNet, the neural network module, please cite [29]

- S. G. Worswick, J. A. Spencer, G. Jeschke, I. Kuprov. *Sci. Adv.* **2018**, accepted

3   History of changes

3.1   Changes in DeerAnalysis2018 with respect to DeerAnalysis2016

DeerAnalysis2018 improves speed and implements several new methods for DEER data processing.

- DeerAnalysis2018 uses a faster implementation of constrained Tikhonov regularization using a dedicated implementation of the FNNLS algorithm. Also, it uses Fresnel integrals to calculate the kernel, with significant speedups compared to the previous versions, which used numerical powder averaging.

- DeerAnalysis2018 implements a new method for DEER data processing based on deep neural networks (DEERNet) that includes an uncertainty estimate.

- For the selection of the regularization parameter, DeerAnalysis2010 now offers three methods: the L-curve corner method (as in previous versions), the Akaike information criterion (AIC), and the generalized cross-validation (GCV). The L-curve corner method tends to oversmooth. AIC and GCV are less prone to oversmoothing, but tend to undersmooth and can fail when the tail of the DEER trace contains non-DEER signals is not properly truncated.

- Validation defaults are now related to data set length and include noise and background fit validation. If a DEERNet computation was performed before, the validation window is skipped and Tikhonov validation is performed with an ensemble of reconstructed background models from DEERNet.

- The graphical user interface window is scalable.

3.2   Changes in DeerAnalysis2016 with respect to DeerAnalysis2015

DeerAnalysis2016 uses a faster implementation of Tikhonov regularization, based on the `regtools` toolbox for Matlab by Per Christian Hansen for L curve computation and unconstrained Tikhonov regularization and on non-negative least-squares optimization for regularization under the constraint $P(r) > 0$. Except for the case where data sets with much less than 512 data points are processed in validation with large regularization parameters α, computation time savings are substantial without sacrificing quality. L curves are more finely graded and of better quality.

The statistics information provided by earlier versions of DeerAnalysis, if the statistics toolbox was available, is no longer provided due to a change in the used Tikhonov regularization routine. We suggest validation instead on relying on statistical measures, which are questionable for an ill-posed problem.

3.3   Changes in DeerAnalysis2015 with respect to DeerAnalysis2013.2

DeerAnalysis2015 is compatible with Matlab2014b and should be backwards compatible to at least Matlab2010 (not tested). Please note that older DeerAnalysis versions work only with Matlab2014a or older.

DeerAnalysis2015 by default disables the constrained fitting and statistical analysis of model fits of H. C. Hyde and E. Perozo. These features can be enabled by the ttfamily Statistics checkbox, if the statistics and optimization toolbox have been found on starting DeerAnalysis. Future extensions will only be included in DeerAnalysis if they work with the basic Matlab distribution without any toolboxes.

3.4   Changes in DeerAnalysis2013.2 with respect to DeerAnalysis2013

DeerAnalysis2013.2 includes constrained fitting and statistical analysis of model fits, courtesy of H. C. Hyde and E. Perozo. Please see `README_DeerAnalysis2011con.pdf` in the subdirectory `Manual` of the distribution. The current application of constrained fitting is for singly labeled homotetrameric and homopentameric proteins [3].

3.5   Changes with respect to DeerAnalysis2011

DeerAnalysis2013 is an upgrade of DeerAnalysis2011, which fixes a few glitches, can retain modulation depth information in user model fitting, can reload its own output files, and, last but not least, has a feature for suppression of ghost contributions to distance distributions that can arise in multi-spin systems [13].

3.6   Changes with respect to DeerAnalysis2009 and DeerAnalysis2010

DeerAnalysis2011 is a minor upgrade of DeerAnalysis2009 and DeerAnalysis2010, which fixes a few glitches, improves the `Validation` tool, and introduces Rice distribution models (courtesy M. Spitzbarth, S. Domingo Köhler, M. Drescher) [14]. Distance distributions have now a color-coded background to indicate reliability in different distance ranges (see Section 7.2). The patched version from March 2012 allows for modulation depth scaling comparison of DEER traces with different length and time resolution, independent of the sequence in which they were loaded and allows for switching off automatic modulation depth scaling between simulated and experimental form factors in the display.

3.7   Changes with respect to DeerAnalysis2008

DeerAnalysis2009 is a minor upgrade of DeerAnalysis2008, which fixes a few bugs and can be run on Mac (Mac executable for Tikhonov regularization courtesy Glenn Millhauser). The following list gives an overview of the changes

- automatic phase correction including offset of the imaginary part
- irregular cutoff behavior and occasional failures in data loading fixed
- user-defined models for three-spin systems with equilateral geometry
- improved automatic determination of the corner of the L curve
- Mac executables of the Tikhonov regularization modules

There is one Microsoft Windows Vista® bug that cannot be fixed. Sometimes the Tikhonov regularization executable crashes for unknown reasons. In this case the same computation has to be performed once again. There is usually no crash on second call. The bug seems to be fixed in Windows 7.

3.8   Changes with respect to DeerAnalysis2006

DeerAnalysis2008 is a major upgrade of DeerAnalysis2006, which fixes a few bugs and introduces a number of improvements in data analysis and interpretation. The following list gives an overview of the most important changes
- no limitations on minimum time increment and maximum
  number of data points in experimental data
- resolution of 1 ns (by interpolation) for zero-time setting
- display option for error estimates with Tikhonov regularization,
  including error estimates obtain by systematic analysis
  with respect to uncertainties in input data
- estimates for distance constraints are provided
- data reduction with use of the information from all data points
  for enhancing computation speed
- enhanced display update during L curve computation and
  progress bar window with estimate of computation time
  remaining; long computations can be interrupted
- possibility for semi-manual background correction (input of modulation
  depth) and decay time constant
- new model-based fit for random-coil conformations (unfolded proteins)
- user-defined models can have up to eight fit parameters
- test data sets for training are provided, and theoretical distance
  distributions are displayed for such data sets
- Linux executables of the Tikhonov regularization modules

Note that error estimates for Tikhonov regularization that can be displayed without using the new validation tool are *not* the true errors of the distance distribution. They are nevertheless included, as they provide some hint to problems in data analysis. To obtain better estimates of the error, please use the new validation tool.

To compute dipolar spectra with higher resolution you may edit the file `zero_filling.dat`. This file is an ASCII file that contains a single integer number n. Data are zero-filled to n times the length of the original time-domain data before Fourier transformation. The default value is 4. Note that zero-filling leads to a purely cosmetical sinc interpolation of data points in the spectrum. No new information is obtained, but the data look better. Purists may want to use a value of 1.

3.9   Changes with respect to DeerAnalysis2004

The user interface of the new version DeerAnalysis2006 was written from scratch, while many of the computational subroutines are well tested subroutines of DeerAnalysis2004. The choice of analysis techniques was narrowed down to the ones that we and others found most reliable. An optimum regularization parameter for Tikhonov regularization is now predicted from the L curve as suggested by Freed's group [16], while the stabilizing constraint of a purely positive distance distribution is maintained as in our previous approach. Recent results of Tsvetkov's group on the effects of finite microwave pulse amplitude [17, 18] were also taken into account. An optional

excitation bandwidth correction is now included that will be described in detail elsewhere. The following list gives an overview of the most important changes

- two data sets can be directly compared on screen (dual display)
- reasonably fast excitation bandwidth correction
- easy work with experimental background functions
- significance check for minor peaks in the distance distribution
- computation and display of L curves in Tikhonov regularization
- the total number n of coupled spins is now displayed, not $n - 1$
- results are no longer saved automatically
- user-defined models with up to six parameters can be fitted

4   Installation

DeerAnalysis2018 requires Matlab R2014b or higher and is platform independent. Use of the DEERNet module requires the Matlab Neural Network Toolbox. With earlier versions of Matlab back to R2008a (7.6), DeerAnalysis2013 is the recommended version. It was tested in Windows, Linux, and Mac environments (Mac test by Glenn Millhauser). If you have an even earlier Matlab version, please use DeerAnalysis2006. Most tests were performed under Matlab 2017a on Windows. Design of the user interface may not be optimum for Linux or Mac, but the scalable GUI should work there. The Windows package can be installed by unpacking the ZIP file DeerAnalysis2018.zip into a directory of your choice. Linux installation is the same.

Please note that the DeerAnalysis directory must not be write protected if the validation tool is used.

(*Tip:* It may be useful to add the path to DeerAnalysis to your default Matlab path in the Matlab startup script `startup.m` by `addpath(''c:/Program Files/DeerAnalysis2018'')`) or to include this directory in the Matlab path using `Set path...` in the Matlab File menu.

5   The user interface

To run the program, start Matlab, change to the directory where it is installed (unless this directory is already on the Matlab path) (e.g., by `cd(''c:/Program Files/DeerAnalysis2018'')`) and call it by typing `DeerAnalysis` at the Matlab prompt

The graphical user interface shown in Fig. 1 opens, of course first without a loaded data set.

The user interface has been programmed with the following ideas in mind

- no *unnecessary* complexity
- a minimum hidden functionality (no menus)
- default behavior should give reasonable results for most data
- experienced users can easily override default behavior.

Default behavior is to read Elexsys (Xepr) data files, assume that the last three quarters of the data can be used for the background fit, adjust the phase automatically, and correct for exponential background decay (homogeneous spatial distribution of nanoobjects). Initially, no points are cut off at the end of the data set, the distance distribution is obtained by approximate Pake transformation (APT) [7], and the mean distance $\langle r \rangle$ and standard deviation σ_r by moment analysis within the range from 1.5 to 8 nm using distance domain smmothing with a filter width of 0.2 nm. A suggestion for cutting off noisy or distorted data points at the end of the data set is made. All this happens automatically once you load a data set via the `Load` button.

Different models for the background can be selected in the `Background models` panel (center of bottom half of Fig. 1) as described in more detail below. Similarly, Tikhonov regularization or fitting of the data by a model distance distribution can be selected in the `Distance analysis` panel. As these approaches are time-consuming, fitting is not started automatically but only after clicking on the corresponding `Fit` button. Adjustable parameters can be edited directly (the most common errors, such as non-digit input or values out of range, are corrected automatically) or incremented or decremented by + and − buttons, respectively. Several parameters can be adjusted or reset by automatic procedures (described below). This is done with the `!` buttons.
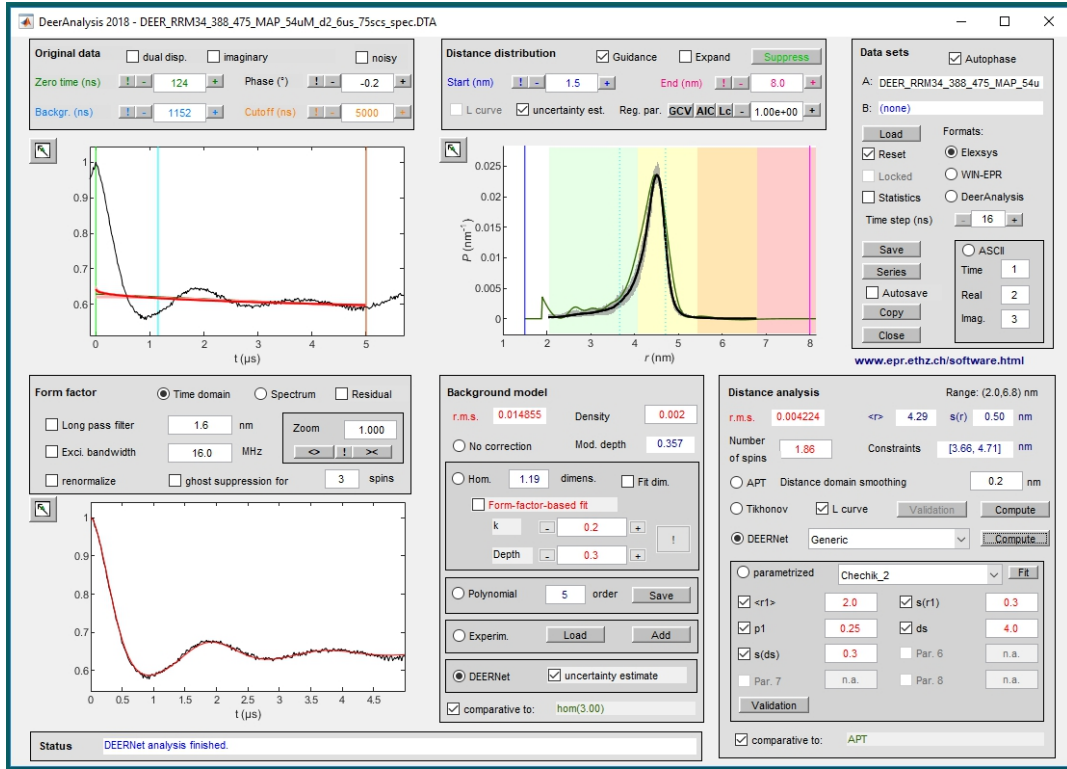
Figure 1: Graphical user interface of DeerAnalysis2018.

Display in each of the plot windows can be toggled. In the plot below the `Original data` panel display of the imaginary part (magenta trace) can be switched on or off by clicking on the `imaginary` checkbox. If two data sets have been loaded, the real part of the previous set (data set B) can be displayed as a blue trace by clicking on the `dual display` checkbox. This automatically suppresses display of the imaginary part of the active data set (data set A) and changes the `imaginary` checkbox into a `mod. depth scaling` checkbox. Dual display and modulation depth scaling also effect the other two plots, where results corresponding to data set B are also displayed as blue traces.

The plot below the `Dipolar evolution` panel can be toggled between time-domain display and display of the dipolar spectrum by checking the corresponding radiobuttons. Finally, the plot below the `Distance distribution` panel can alternate between display of the distance distribution and the L curve, after a Tikhonov regularization with L curve computation has been performed.

There is no Help function, but the controls are provided with short explanations that will show up when you move the cursor above them.

6 Pre-processing

6.1 Loading data

Data input and output is initiated by buttons in the `Data sets` panel. By default the program expects Bruker Elexsys data (binary format). It recognizes automatically if the data are complex (quadrature detection) or real (single-channel detection, discouraged). If the data set is one-dimensional, it is interpreted as output of a (classical) constant-time DEER experiment [4], see also Fig. 2. If the data set is two-dimensional with exactly two traces, it is interpreted as a variable-time DEER experiment [5] with the first trace being the reference trace and the second trace being the recoupled trace. For any other size of experimental data, program response is

Pulse sequence

Nitroxide spectrum

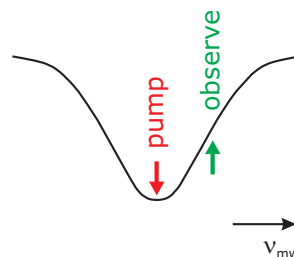$\Delta\nu$ = 65-70 MHz
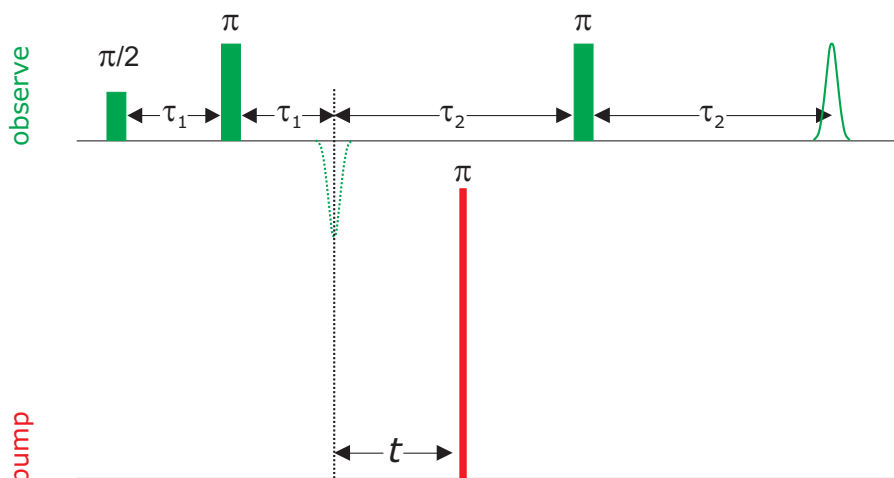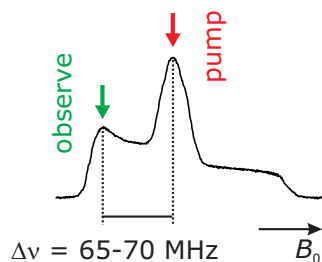
Microwave mode

Figure 2: Pulse sequence and positions of the observer and pump frequency with respect to the nitroxide spectrum and to the microwave mode for the four-pulse DEER experiment.

undefined. If you unintentionally load a data set of some other experiment, it is advisable to close the program and restart it.

Mainly as a support for ESP 380 machines, the program has the capability to read data in WIN-EPR binary format (select by radio button in the `Formats` column of the). As the binary number format of the ESP 380 is somewhat obscure, this mode requires that the data are first read into WIN-EPR on a PC and saved again from WIN-EPR. This mode is less well tested than the Elexsys mode and completely untested for two-dimensional data. Alternatively you can convert ESP 380 data to ASCII data (also possible in WIN-EPR with command sequence `1D processing/Parameters/List data file.../Save`).

From an ASCII file, only one-dimensional data can be read. If there are any header lines before the numerical data, they must start with a percentage character (%). By default, the program expects the time axis (in nanoseconds) in the first column, the real part of the data in the second column, and the imaginary part (if present) in the third column. These assignments can be adapted in the edit fields below the `ASCII` radio button. For ASCII data exported from WIN-EPR, the proper settings are 2, 3, and 4 instead of 1, 2, and 3. The first six lines (header lines) have to be deleted or commented out by a % character. The program automatically recognizes if there is no imaginary part. After successfully loading data, the `Status` panel shows a short characterization of the data set (const-time/variable-time DEER, complex/real, number of data points). The filename is included in the title of the DeerAnalysis main window and is also shown in line `A:` of the `Data sets` panel.

From version 2013 on, data saved by DeerAnalysis can be reloaded, when the `DeerAnalysis` radiobutton in the `Data sets` panel is activated. By default this activates the `Locked` checkbox

in the same panel. Any of the DeerAnalysis output files (e.g. `..._bckg.dat`) can be selected, data from all files are loaded. The same feature allows for loading output data from the DEER window of MMM, provided that a comparison of experimental data with a simulated distance distribution was performed in MMM. In `Locked` mode any automatic processing is switched off, data and results are displayed in the form found in the saved files. The `Locked` checkbox can be deactivated to allow for data processing. Note however that not the full information from the primary data set is available after such reloading. Only the real part of the primary data, starting at the zero time determined by DeerAnalysis or by the user is saved and can be reloaded. In other words, phase correction and zero time determination should not be changed in such data sets.

6.2 Determining zero time

The time origin of the dipolar evolution function corresponds to the position of the *pi* pump pulse coinciding with the unobserved first echo. (see Fig. 2). Because pulse lengths are finite, the relation between this condition and actual delays in the pulse sequence may not be trivial. We therefore suggest determination of the time origin (zero time) from experimental data with a good signal-to-noise ratio (SNR) for the pulse lengths and τ_1 delay that you actually use. To obtain a precise value, a standard sample with a short distance should be used. If you later measure on the same spectrometer with the *same* pulse lengths and τ_1, you can use the same value. Knowing this value is important for data with poor SNR where automatic determination is likely to fail. Automatic determination of zero time t_0 is based on the expectation that the real part of the signal should be symmetric about the time origin. For the proper choice of t_0, the first moment of the signal in a range symmetric about t_0 should thus be zero. In a first step, the program approximates zero time by the time $t_{\max}$ at which the real part is maximum. Then the first moment is determined in a window $t_x < t_{\max}/2$, where t_x is shifted through the whole data set. The optimum value of t_0 is the time t_x where the first moment is minimum. This procedure is performed with a time resolution of 1 ns, obtained by interpolation of the experimental data. Such enhanced time resolution improves results for very short distances, where it may be important that the true zero time may fall in between two experimental data points. Zero time is influenced by pulse lengths [17, 18]. This algorithm should work well for good SNR and distances up to ≈ 5 nm. If it fails under such conditions, τ_1 is too short (expected symmetry of the data is spoiled by interference between adjacent microwave pulses). The algorithm may fail for very long distances where data close to the maximum are pretty flat. For such long distances small mis-settings have only minor influence on the distance distribution.

You may correct the automatically determined zero time by the + and − buttons right and left from the value or by direct input of a new value in the edit field (fit by the eyes). A wrong choice may be easier to detect when you switch the `Dipolar evolution` plot to frequency domain (`spectrum`).

6.3 Phase correction

In a properly adjusted DEER experiment, the signal should be entirely in the real part of the data set. If receiver offsets are canceled by [(+x)-(-x)] phase cycling of the first pulse, as we strongly suggest, the imaginary part is zero. It is therefore tempting to acquire and process only the real part. We discourage this. For very weak signals, as you occasionally encounter with membrane proteins, it is difficult to adjust signal phase exactly during setup. Consequently, part of the signal will be in the imaginary part. Furthermore, depending on stability of your spectrometer, there may be small phase drifts during the experiment. It is better to correct for these drifts than to ignore them. Finally, unexpected artifact signals are likely to manifest in the imaginary part (see Fig. 3). If the imaginary part *after* phase correction strongly deviates from zero at early times, it is advisable to acquire data with a longer τ_1 value (see Fig. 2).

Automatic phase correction can be based on the expectation that the imaginary part should be zero at sufficiently long times. By default, the program determines the corresponding phase correction directly after loading complex data by minimizing the root mean square deviation of the imaginary part for the last three quarters of the data (part between the blue and orange cursors). The phase correction in degree is displayed in the `Original data` panel.
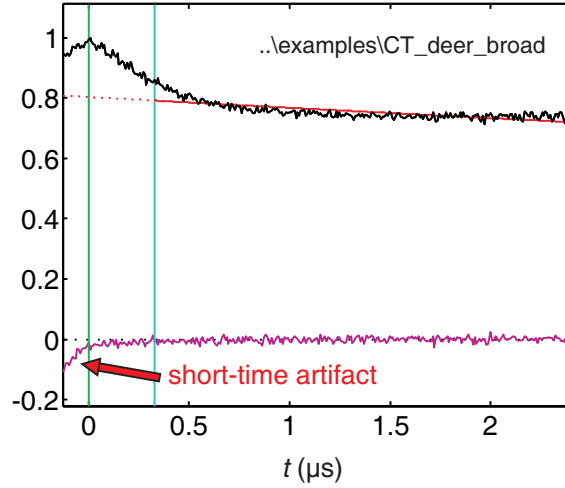
Figure 3: Imaginary-part artifact at early times (see red arrow) due to mw pulse interference. Interpulse delay τ_1 should be long enough for the artifact to have almost completely decayed at $t = 0$ (green vertical line).

You may correct phase manually by using the + and − buttons right and left from the value or by direct input of a new value in the edit field. Phase is automatically restricted to the range $(-180, +180)°$. If you did not phase cycle and do have a receiver offset, you may aim to flatten the imaginary part and put all modulation into the real part. Note however, that in this case you are likely to have a receiver offset in the real part, too. This will be detrimental to data analysis. Automatic phase correction can be reactivated by the ! button left from the value. It will always relate to the part of the data between the blue and orange cursors. If you move any of these cursors, the result may differ from the result that you got directly after loading.

Automatic phase correction after loading can be deactivated by unselecting the check box `Autophase` in the `Data sets` panel.

In some measurements of weak samples with some spectrometers we found that automatic phase correction did not work and that after a manual phase correction that brought all modulation into the real part, a flat non-zero imaginary part remained. Apparently the problem comes from overload of the videoamplifier during the pulse. Small residual constant offsets of the imaginary part may also result if the phase drifted significantly during the measurement. Such data are less reliable than data with zero offset in the imaginary part, in particular with respect to background correction, as there may also be an offset in the real part. However, if the hardware problem cannot be fixed, it may still be warranted to process the data and interpret them (with caution), in particular, as for weak samples the background may be flat, so that an offset in the real part does not pose a problem.

6.4 Cutting data

For several reasons, you may want to exclude points at the end of your data set from analysis. First, some people prefer to acquire data up to delays t, where the pump pulse starts to interfer with the last observer pulse or even overlaps with it. In this case, the last data points are spoiled. Second, if at maximum t the signal has decayed to a very small value (say 0.1 times maximum intensity), the dipolar evolution function after background correction will be rather noisy, as correction involves division by the background decay. Third, and most important, overlap of pump and observer excitation bands cause a "2+1" artefact at the end of the trace.

By default no data points are cut off at the end, but a suggestion for cutoff is displayed as an orange vertical cursor in the `Dipolar evolution` plot (see Fig. 4). This suggestion is derived from the difference D between the experimental dipolar evolution function and its fit by the APT result. The mean square deviation M_k of eleven consecutive points $D_{k-5} \ldots D_{k+5}$ around the kth

data point is computed for all indices k. The minimum of M is a measure for the noise level. An acceptable noise level of $6\min(M)$ is assumed. The program then searches for a range of consecutive points at the end of the data set that *all* fulfil the condition $M(k) > 6\min(m)$. If such a range of points exists, the program suggests to cut it off. Otherwise the orange cutoff cursor is set to the end (right border) of the trace.
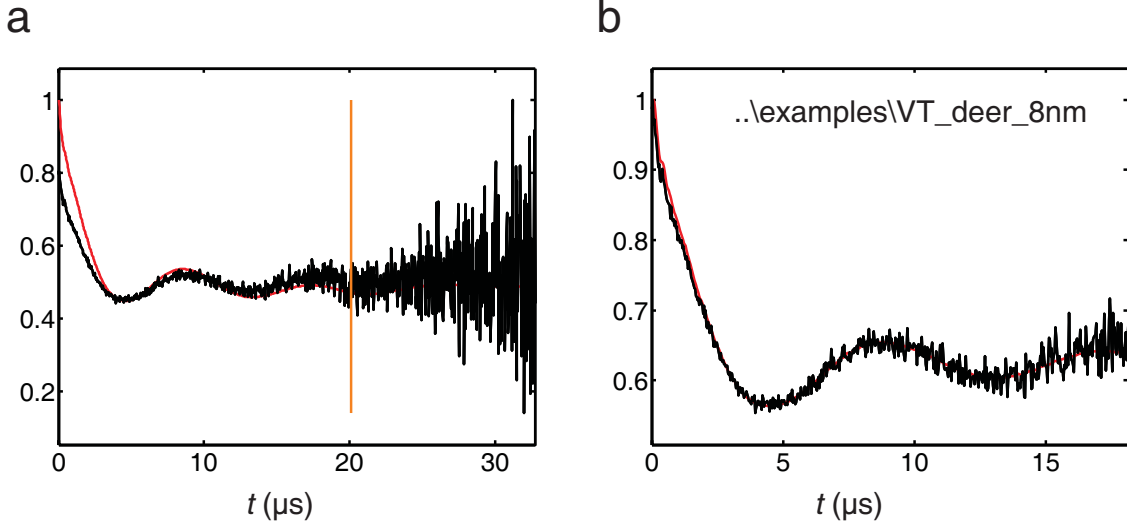
a

b

..\examples\VT_deer_8nm

t (µs)

t (µs)

Figure 4: Cutting off the noisy part at the end of variable-time DEER data. (a) `Dipolar evolution` plot for the whole data set. The orange cursor shows the suggested cutoff time. (b) Dipolar evolution function (black) and fit (red) by a distance distribution obtained with APT after cutting the data at the suggested time.

The suggestion can be accepted by clicking on the ! button of the `Cutoff` controls in the `Original data` panel. Note that this may in turn improve the fit, thus leading to a smaller value $\min(M)$ and a new cutoff suggestion. Therefore, it is advisable to click on the ! button several times to iteratively approach the optimum cutoff. Furthermore, the cutoff suggestion depends on correct settings of other parameters (zero time, phase). For instance, for the variable-time DEER data set shown in Fig. 4 the zero time must be zero, while the program automatically determines 96 ns. If this is not corrected, not good fit is obtained and unnecessarily many data points are cut off. Note also that this data set with relatively poor SNR was intentionally selected for explanation of data cutoff. For many data sets, no cutoff at all may be required and DeerAnalysis immediately sets the cutoff cursor to the right border.

Generally, cutting off a significant amount of data will suppress noise but will also cause a suppression of long distances by background correction. Proper background correction may become more difficult.

6.5   Background correction

In most cases, EPR distance measurements are performed to elucidate the structure of a nanoscopic object. Only distances within this object are of interest. The contribution of distances to neighboring objects should be suppressed. If you think about a biradical or bilabelled protein molecule, you want to measure the intramolecular distance and suppress contributions from intermolecular distances.

Such a separation of the signal $V(t) = \{1 - [1 - \Delta D(t)]\}B(t)$ into a dipolar evolution function $D(t)$ for the nanoobject itself and a background decay $B(t)$ due to neighboring objects requires a criterion for distinguishing the two contributions. Furthermore, the functional form of the background decay has to be known. This functional form is related to the spatial distribution of the nanoobjects. A separation can only be successful if distances within the object are typically shorter than distances to neighboring objects. The wanted contribution is then confined to the

earlier part of the time domain data, while later parts are dominated by the background decay. The decay can only be fitted properly if the maximum time t in the pulse sequence (Fig. 2) is significantly longer than the time at which the dipolar modulation has decayed. A more detailed discussion can be found in Ref. [6].

Separation into the two contributions is simple and reliable if the distance distribution is dominated by distances shorter than 4 nm. In protein samples, it becomes challenging for distances between 4 and 6 nm, and near to impossible for distances longer than 6 nm, unless protons around the spin labels can be strongly diluted by deuteration [5]. Note that one can still get a quite reliable estimate of a distance of closest approach if separation fails. However, the width and shape of the distance distribution should not be discussed in such a situation (see also Fig. 6).

In simple cases (short distances and homogeneous distribution of the nanoobjects in three dimensions), separation depends only weakly on the choice of parameters. Default behavior of the program should then be sufficiently good. By default, an exponential background decay corresponding to a homogeneous three-dimensional distribution is fit to the last three quarters of the data. The fit parameter is the decay time constant, which is proportional to the concentration of nanoobjects. With proper calibration such fits can be used to determine local concentrations (see Section 6.6).

Generally, the background is shown as a red line in the `Original data` plot. A continuous line is plotted in the range where the background was fitted (between the blue and orange cursors), a dotted line is plotted where the fit was extrapolated. The r.m.s. value of the background fit is displayed in the `Background model` panel.
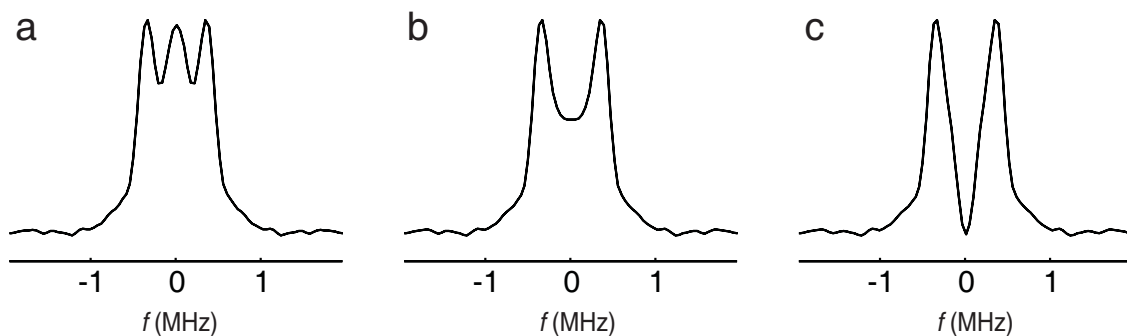

Figure 5: Manifestation of different background fits in the dipolar spectrum (example data set CT_DEER_5nm). (a) Part of the background is attributed to the biradical. (b) Good separation of intra- and intermolecular contributions, as obtained with automatic correction (! button). (c) Part of the biradical contribution is attributed to the background.

For distances of $\approx$ 4 nm and longer, choice of the time range for background fitting may decide whether you obtain artifacts in the distance distribution at long distances. Unlike the other problems in determining a distance distribution, this problem is most severe for narrow distributions of distances. In this case the modulation decays more slowly and thus interferes more strongly with the background fit. Our automatic determination of the optimum fit range is based on the assumption that the longest detectable distance exceeds the largest distance within the nanoobject. If this condition is met, the distance distribution after correct background correction is zero at the maximum detectable distance. This can be checked by approximate Pake transformation (APT, see below). APT is sufficiently fast to be applied at all possible choices of the starting time for the background fit. For any selected background model, this search for the optimum starting time can be initiated by clicking on the blue ! button near `Backgr. (ns)` in the `Original data` panel. Depending on the length of your data set and the speed of your computer, this optimization can take up to a few minutes.

The starting time for background fitting can also be adjusted manually with the blue + and – buttons or by direct input into the edit field. The consequences can best be judged when switching the bottom left plot below to frequency domain. For a narrow distance distribution, the black trace should look like a Pake pattern. Deviations are best seen at zero frequency. There should

be neither a positive spike nor an obvious hole in the center of the Pake pattern (see Fig. 5).

Background correction can be switched off completely by selecting the `No correction` radiobutton in the `Background model` panel. In this modus the input data are interpreted as a dipolar evolution function which is already separated from background. The modus is intended for compatibility with external pre-processing programs, for polynomial fitting of single-label data to derive an experimental background function (see below), or for fitting by a user model that explicitly contains the background contribution. User models consisting of a single Gaussian peak with 3D homogeneous background (`Gaussian_hom`) or of two Gaussian peaks with 3D homogeneous background (`Two_Gaussians_hom`) are already included in DeerAnalysis. However, we strongly discourage fitting background and distance distribution simultaneously, as such fits are very likely to end up in local minima of the error hypersurface, especially for those cases, where previous background correction is difficult. Whenever a separation of the background contribution from the contribution of the nanoobject can be performed with some confidence, it should be done before analysis of the distance distribution.

For long distances, the intramolecular contribution may be significant throughout the whole time range. In such cases, any fit of a background function to the data is biased. If independent information on expected modulation depth and concentration (density) is available, e.g. from other double mutants of the same protein, it may be better to directly input these estimates instead of fitting. This can be done by selecting the option `Form factor based fit`. This option is displayed red when the program estimates from the preliminary distance distribution that fitting would be more reliable and is displayed green if the program expects problems with background fitting.

It is also possible to systematically vary modulation depth and density and search for the values that lead to the best fit of the form factor. The rational behind this algorithm is that the background contribution cannot be fitted by a distribution of distances with an upper limit. Incomplete background correction or overcorrection will thus lead to a deterioration of the fit of the form factor. This algorithm works well for data with very high signal-to-noise ratio, but is easily mislead by noise. It should thus be used with caution. The search for optimum modulation depth and density can be started by selecting the option `Form factor based fit` and then clicking on the ! button in the `Form factor based fit` subpanel.

To test for uncertainties in parameters of background correction, the result of Tikhonov regularization should always be checked by using the `Validation` module.

DEERNet (Section 7.5) analyzes background and distance distribution simultaneously without fitting background explicitly in only a limited range of the data. This appears to work better in cases where the trace is not sufficiently long. DEERNet uncertainty estimates include background uncertainty.

By default, the previous background fit is displayed in the `Original data` panel in green color if background model or parameters were changed. The mode of the previous background fit is displayed at the bottom of the `Background model` panel, where you also find a checkbox that can be deactivated to suppress display of the previous fit.

In the following we shortly discuss the possible choices for the spatial distribution of nanoobjects. They can be selected by checking the corresponding radiobutton in the `Background model` panel.

6.5.1 Homogeneous

This model is strongly suggested for all cases where you do not have experimental background functions from singly labelled molecules. The general background function in this model is

$$B\left(t\right) = \exp\left(-(kt)^{d/3}\right) \qquad (2)$$

where k depends on the density of the spins and the pump pulse inversion efficiency, and d is the dimensionality of the homogeneous distribution. Unless there is a confinement on length scales below 10 nm, the distribution is homogeneous in $d = 3$ dimensions. This case applies to most solutions. Membrane proteins in a liposome may be confined to $d = 2$ dimensions. If possible, such confinement should be established by control measurements on singly labelled proteins, for which

$d = 2$ is expected give a better fit than $d = 3$. For labels attached to a stretched polymer chain, $d = 1$ may be appropriate. Note also that a choice of $d = 6$ corresponds to a Gaussian background decay, as it has been observed with the single-frequency SIFTER experiment [10]. The dimension is not necessarily an integer number- if experimental data of a singly labelled sample can be nicely fitted with a fractal dimension, it is advisable to use the same fractal dimension for background correction of the corresponding doubly labelled sample.

When the `Fit dimensionality` checkbox is selected, both k and d are fitted. This mode is suggested only for determining the fractal dimension of purely homogeneous (singly-labelled) samples or for high-quality data where a sufficiently long part of the trace is dominated by background. For fitting purely homogeneous background of a single label, the `Bckg.` control in the `Original data` panel should be set to zero (green and blue cursors coincide), as the early decay of the data is most sensitive to the parameter d.

6.5.2 Polynomial

Short distances are underrepresented in the intermolecular distance distribution, ff the spin labels are attached to nanoobjects that cannot penetrate each other. As a result, the intermolecular contribution decays more slowly at early times than would be expected for a homogeneous distribution. If singly labelled objects are available, the intermolecular part can be measured separately and an experimental background function can be derived. Directly using the noisy experimental data set of the singly labelled sample would introduce significant statistical errors. It is therefore prudent to use a smooth fit function for that purpose.

Almost any intermolecular decay can be reproduced by fitting a polynomial to the *logarithm* of the original data. DeerAnalysis allows for polynomials with an order of up to 15, but note that the lowest order should be selected that still gives a good fit (flat trace in the `Dipolar evolution` plot. Polynomial fits are mainly intended for deriving and afterwards saving experimental background functions from singly labelled samples, not for direct background correction.

6.5.3 Experimental

Once experimental background functions have been derived from singly labelled samples, they can be used for correcting the background in corresponding doubly labelled samples. In this mode, the relative magnitudes of the polynomial coefficients are kept fixed. The background model is given by

$$B\left(t\right) = \exp\left(-k\sum_{n=0}^{o} c_n t^n\right) \tag{3}$$

where k is the density (concentration) parameter, o the order of the polynomial, and the c_n are the polynomial coefficients determined previously on the singly labelled samples. The only fit parameter is k.

In principle, background data should be individually measured for both label positions in a doubly labelled sample, as the supression of short distances depends on how deep the label is buried in the nanoobject. The weighted sum of both background functions is a better approximation for the actual background in the doubly labelled sample than each individual background function. Several background polynomials can be added using the `Add` button in the `Background model` panel. A weighting factor can be specified in a dialog box that opens after clicking on this button. Note that the different labeling efficiencies at the two positions are already accounted for with weighting factor 1.0 if both singly labelled samples were measured with the same protein concentration.

6.6 Determining local concentrations

The parameters of the background fit are related to the number of coupled spins within the nanoobject (modulation depth after background correction) and to the density of nanoobjects (parameter k). For calculation of the number of spins and of absolute densities, the modulation depth parameter λ has to be known, which depends strongly on the excitation position, length, and flip angle of the pump pulse and weakly on line broadening in the nitroxide spectrum and

shape of the resonator mode. Reliable quantification therefore requires a calibration with known samples and proper adjustment of the flip angle of the pump pulse (see Section 12). The calibration should be repeated if the resonator or the length of the pump pulse is changed. Protonated and deuterated nitroxide spin labels also require separate calibrations. Determination of the number of coupled spins is more reliable when based on Tikhonov regularization or a fit of the data by a model distribution and is therefore discussed later on (Section 9.2).

For a 3D homogeneous distribution of objects, the density is proportional to the local concentration. The term local refers to the length scale of the DEER experiment, which extends to approximately 10-20 nm for the background. Measurements of local concentrations can be calibrated with a solution of an appropriate spin label (e.g., protonated or deuterated TEMPOL) in toluene. An example data set from our own calibration (CT_DEER_tempol_2500uM) is provided. This data set was acquired with a 2 mM TEMPOL solution in toluene, which corresponds to a concentration of 2.5 mM at 80 K, as toluene shrinks to approximately 80% of its room temperature volume when freeze-quenched in liquid nitrogen.

To calibrate 3D background fitting for determination of concentrations, select `Homogeneous` as the background model, set `dimensions` to 3, and load a data set for a sample with known concentration. Adjust zero time and phase, if necessary. Now input the concentration (in the units you prefer) into the edit field `Density`. The color of the density value then changes to green. When you now load other experimental data sets that have been measured with the same resonator and experimental settings and use the same background model, you can directly read off concentrations from the edit field `Density`. Note that the program looses calibration on restart.

6.7 Long-pass filtering

The major artifact contribution to DEER time-domain signals is usually nuclear modulation due to matrix protons. At X-band frequencies, such proton modulation corresponds to a distance of approximately 1.5 nm. By restricting the distance range for analysis to (1.75, 8) nm, contributions by nuclear modulation can be suppressed. However, as computation of distance distribution is an ill-posed problem, an out-of-range artifact may still influence the result within the range of interest. Very strong proton modulations, as they are sometimes encountered for membrane proteins in liposomes or detergent micelles, should thus be eliminated by filtering. Note, however, that it is preferable to eliminate such nuclear modulation experimentally by averaging traces with different τ_1 values.

This can be achieved by completely eliminating contributions above a certain maximum frequency, which roughly corresponds to suppressing distances below a certain minimum distance. Such complete suppression was described in Ref. [5]. For broad distance distributions with contributions both below and above 1.75 nm, complete suppression may introduce an artificial hole at $t = 0$ into the time-domain data and may thus replace the nuclear modulation artifact with a suppression artifact. To avoid this, filtering in DeerAnalysis is performed by fitting a third-order polynomials to the real and imaginary parts of the frequency-domain data between the cut-off frequency and the Nyquist frequency. The frequency-domain data in this range are then replaced by the polynomial. This suppresses the sharp nuclear modulation peak as well as high-frequency noise, while keeping the high frequency contributions of broad distance distributions intact.

Filtering is enabled by selecting the `Long pass filter` checkbox in the `Dipolar evolution` panel. The cut-off distance (lower limit, default 1.6 nm) can be changed in the edit field right from this check box. When working with broad distributions of short distances, the default value is often a good compromise between residual proton modulation and partial suppression of short distances.

7 Extracting distance distributions

7.1 General remarks

The computation of a distance distribution $P(r)$ from a dipolar evolution function $V(t)$ is an ill-posed problem. For such problems, small variations in the input data (e.g., noise) can cause large variations in the output data. In other words, significantly different distance distributions may

correspond to very similar dipolar evolution functions. Data analysis therefore depends strongly on striking a good compromise between improving resolution and decreasing the influence of experimental noise. First and foremost, data should be acquired with as good as possible SNR. Reproducing results for a given sample is usually a good idea. Second, ill-posedness must be taken into account in data analysis. There are several ways of doing this, which all have one feature in common: one tries to find a resolution in distance domain at which a good fit of the experimental data is obtained without introducing strong noise artifacts into the distance distribution.

In order to facilitate comparison of different approaches to computing the distance distribution, the result of the previous computation is shown in green color in the `Distance distribution` panel. The type of previous analysis is displayed at the bottom of the `Distance analysis` panel, where you also find a checkbox that can be deactivated in order to suppress display of the previous distribution.

7.2   Reliability of distance distributions

The reliability of distance distributions depends strongly on the maximum dipolar evolution time. A rule of thumb was derived by reanalyzing data that were simulated from known distance distributions. At a maximum dipolar evolution time $t_{max} = 2 \ \mu s$ the shape of the distance distribution is reliable up to a distance of about 3 nm (reliable distribution limit). The mean distance $\langle r \rangle$ and width σ_r are reliable up to a distance of 4 nm (reliable width limit), whereas the mean distance, but not the width is reliable between 4 and 5 nm (reliable meand distance limit). Beyond 5 nm no reliable mean value can be determined, although the presence of a long distance (distinct from background) can be recognized up to 6 nm (distance regonition limit). All these limits scale with the cubic root of t_{max}.
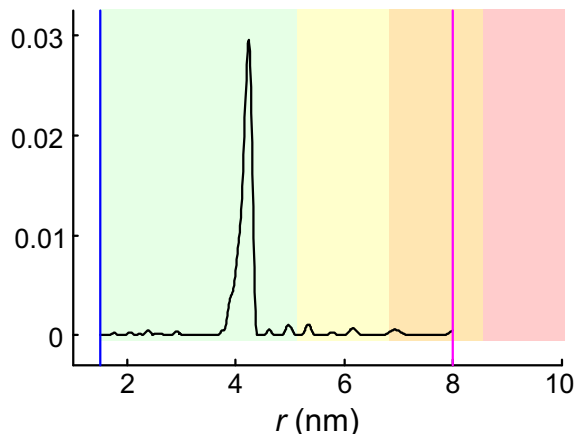

Figure 6: Color coding for reliability ranges. Pale green: Shape of distance distribution is reliable. Pale yellow: Mean distance and width are reliable. Pale orange: Mean distance is reliable. Pale red: Long-range distance contributions may be detectable, but cannot be quantified. The example data were cut off at a maximum dipolar evolution time of 10 μs. In this particular case the shape of the distribution can safely be interpreted.

DeerAnalysis displays reliability ranges by color-coding the background of the distance distribution plot. The pale green range corresponds to a reliable shape of the distribution, the pale yellow range to reliable mean distance and width, the pale orange range to reliable mean distance (but not width), and the red range to recognition of a long-distance contribution that cannot be quantified. The computed distribution is displayed only up to the distance recognition limit.

This feature is intended to caution users against overinterpretation of data. For preparing figures of distance distributions for papers it may be required to suppress these background colors. This can be achieved by deactivating the checkbox `Guidance`.

7.3 Approximate Pake Transformation (APT)

A very fast algorithm relies on an approximate integral transformation to dipolar frequency domain, subsequent correction of cross-talk artifacts, and mapping to distance domain (APT) [7]. Ill-posedness is moderated by proper discretization in dipolar frequency domain. If SNR is too small, the distance distribution may still be influenced by strong noise artifacts. A better compromise between reliability of the distribution and resolution can then be achieved by distance-domain smoothing, i.e., by giving up resolution in favor for a smoother distribution. As APT is very fast, it can also be used to generate starting values for fit procedures. The disadvantage of APT with respect to other techniques is that it cannot incorporate the constraint $P(r) > 0$ (for all r). This disadvantage, however, is significant, as the constraint strongly stabilizes the solution. For this reason, several other approaches for data analysis are incorporated into DeerAnalysis.

7.4 Tikhonov regularization

Other approaches rely on computation of a simulated time-domain signal $S(t)$ from a given distance distribution $P(r)$ by

$$S(t) = K(t, r) P(r), \tag{4}$$

where K is the kernel function. For the DEER experiment with ideal pulses, the kernel function is known analytically

$$K(t, r) = \int_0^1 \cos\left[\left(3x^2 - 1\right) \omega_{\mathrm{dd}} t\right] \mathrm{d}x, \tag{5}$$

with

$$\omega_{\mathrm{dd}}(r) = \frac{2\pi \cdot 52.04 \ \mathrm{MHz \ nm}^{-3}}{r^3}. \tag{6}$$

The case of non-ideal pulses is discussed in Section 7.7.

The most widely used approach to ill-posedness is Tikhonov regularization. In this approach, the compromise between smoothness (artifact suppression) and resolution of the distance distribution is quantified by a regularization parameter α. The optimum distance distribution $P(r)$ is found by minimizing the objective function

$$G_\alpha(P) = \|S(t) - D(t)\|^2 + \alpha \left\| \frac{\mathrm{d}^2}{\mathrm{d}r^2} P(r) \right\|^2 \tag{7}$$

for a given α. The first term on the right hand side of eqn (7) is the mean square deviation between the simulated and experimental dipolar evolution function while the second term is the regularization-parameter weighted square norm of the second derivative of $P(r)$, which is a measure for the smoothness of $P(r)$. The larger α the less noise artifacts are introduced. However, a larger α also causes a stronger broadening of peaks in the distance distribution. Therefore, small α are required for samples with well defined distances (narrow peaks) and large α for very broad distributions, which otherwise disintegrate into many narrow peaks. Unfortunately, the correct width of the peaks is often not known in advance.

There are different ways for mathematically defining an optimum regularization parameter. The past version DeerAnalysis2004 used the self-consistency criterion [20, 21]. However, determination of an optimum α is itself influenced by noise [6], and the self-consistency criterion appears to be more sensitive to noise distortions than the L curve criterion [16]. The L curve is a plot of $\log \eta(\alpha)$ versus $\log \rho(\alpha)$, where

$$\rho(\alpha) = \|S(t) - D(t)\|_\alpha^2 \tag{8}$$

quantifies the means square deviation and

$$\eta(\alpha) = \left\| \frac{\mathrm{d}^2}{\mathrm{d}r^2} P(r) \right\|_\alpha^2 \tag{9}$$

the smoothness. For well behaved data (good signal-to-noise ratio, relatively narrow peaks in the distribution), this plot is L-shaped as is illustrated in Fig. 7a. In the range of small regularization

parameters α (left of the corner, undersmoothing) the slope is steep and negative, as increasing α and thus the smoothing strongly decreases the norm of the second derivative of $P(r)$ without strongly affecting the mean square deviation. In contrast, right of the corner (oversmoothing) the mean square deviation increases strongly with increasing α as the simulation is no longer a good fit of the data. At the same time, η decreases only gradually as noise-related spikes in $P(r)$ are already smoothed out. If the SNR is worse and the peaks in the distance distribution are broader, the corner of the L curve is somewhat less pronounced (Fig. 7b).
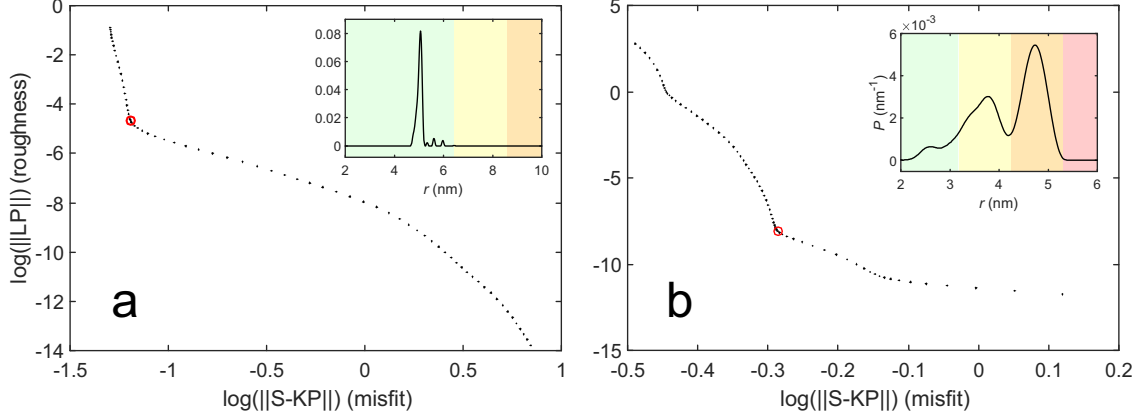

Figure 7: Tikhonov L curves. The red data points correspond to the optimum regularization parameter selected by GCV. The insets show the distance distribution obtained with this parameter. a) Data set dOTP_5nm, $\alpha = 7.94$. b) Data set CT_DEER_broad, $\alpha = 398$.

In DeerAnalysis, the L curve criterion is implemented for unconstrained Tikhonov regularization, which allows for very fast computation of the L curve. Initially, constrained Tikhonov regularization is computed only at the automatically detected corner of the L curve (red circles in Fig. 7). The user can click on other points in the curve and constrained Tikhonov regularization is then recomputed for the corresponding regularization parameter. For regularization parameters that do not correspond to the automatically detected corner, a blue circle is shown.

Tikhonov regularization is performed by clicking on the corresponding `Fit` button in the `Distance analysis` panel. By default, L curve computation is enabled, since the shape of the L curve contains information about the quality of the data, and because the extended α is needed for the automatic selection methods (L-curve corner, AIC, GCV - see below) to work. If it is too time consuming for a particular dataset, it can be switched off, which is however not recommended. The regularization parameter (default: 1) can be changed in the corresponding edit field in the `Distance distribution` panel. The distance range for Tikhonov regularization is determined by the blue and magenta start and end values in the `Distance distribution` panel, which can also be edited. Computation of the L curve can be requested by clicking on the `L curve` checkbox in the `Distance analysis` panel and subsequently clicking on the `Compute` button. The status line indicates the progress of the L-curve calculation. After the computation, the L curve is automatically displayed instead of the `Distance distribution` plot with the automatically derived selection of the L-curve corner highlighted in red and the corresponding regularization parmeter shown in the `Reg. par.` control. The selection of the regularization parameter can be shifted with the `+` and `-` buttons of the `Reg. par.` controls. Such changes update the fit in the `Dipolar evolution` panel and the r.m.s. value in the `Distance analysis` panel. The distance distributions for different regularization parameters can be inspected in the same way after unselecting the `L curve` checkbox in the `Distance distribution` panel.

Automatic L curve corner recognition selects the regularization parameter that has the shortest distance from the lower left corner in a log-log plot of square norm of the second derivative of $P(r)$ vs. mean square deviation of the simulated from the experimental form factor (new algorithm since DeerAnalysis2009). The lower left corner is defined by the minimum square norm of the second derivative of $P(r)$ and the minimum mean square deviation among all regularization parameters

for which Tikhonov regularization was performed. Note that L curves can be misshaped if there are problems with background correction, if data are very noisy, or if the distribution cotains both very broad and very narrow features. In such cases the automatically determined choice of the regularization parameter may not be reasonable. If you have any information on the expected width of the distribution (or of the most narrow features in the distribution) from data other than the one being analyzed, it is usually best to select the regularization parameter manually. The optimum choice is the one that just does not cause undue broadening of expected narrow features.

Beside the L-cruve corner, DeerAnalysis2018 offers two additional methods for selecting an appropriate value of the regularization parameter α, the Akaike information criterion (AIC) and the generalized cross-validaton (GCV)[24]. Both can be obtained by clicking on the respective buttons labeled `AIC` and `GCV` next to the edit box for the regularization parameter. The L-curve corner value can be obtained again by pressing the button labeled `Lc`. The AIC method treats the solutions determined with different values for α as different models and determines the value that corresponds to the most parsimonious model. The GCV method drops one datapoint from the dataset at a time and determines α that generates a fit from the remaining datapoints that best recovers this dropped datapoint. This procedure is repeated for every datapoint in the DEER time trace. Whereas the L-curve corner method tends to over-broaden the distance distributions, the AIC and GCV typically result in much sharper distributions, with a tendency to underbroaden. They also can fail if the tail of the DEER trace is not properly cropped.

After a Tikhonov regularization has been performed, the `Validation` button becomes accessible. An error analysis with respect to noise and uncertainties in background correction can now be performed (see Section 8).

7.5 DEERNet (neural network ensemble

DEERNet provides a solution to the ill-posed inverse problem that does not rely on a single regularization parameter and that can thus recover both narrow and broad features at the same time. Furthermore, this approach, based on neural network ensembles trained on a large number of simulated data sets, does not rely on a background separation that requires fitting of the background model to part of the data. Rather, the neural networks have learned the different patterns of dipolar evolution functions of isolated spin pairs and of homogeneous background. This makes DEERNet an alternative approach that works better for some complicated cases [29].

Since the neural network ensembles are large, they are not included in the downloadable version of DeerAnalysis. Upon first use, DeerAnalysis connects to the internet and downloads the networks. This can take several minutes, but happens only once. DEERNet requires the Neural Network Toolbox of Matlab and uses parallel computing for speedup.

Four network ensembles are available, which can be selected in the popup menu in the `Distance analysis` panel. If unsure, use the generic one, which is also the default. Special networks are available for distributions with only narrow or only broad peaks and for cases where a fixed exchange coupling is observed on top of a distance distribution.

Computation is started by clicking on the `Compute` button and returns an ensemble of solutions, from which DeerAnalysis derives a mean and an uncertainty estimate. DEERNet does not *per se* return a background function, a fit to the original data, or an estimate of modulation depth. This information is reconstructed by simulating the form factor from the distance distribution and fitting modulation depth, fractal dimension, and decay constant for a stretched exponential background model. There is no strict mathematical reason for this being valid, but for the problem at hand, the approach works, because DEERNet was trained with stretched exponential background functions and appears to implicitly model background by such functions most of the time. Failure of this reconstruction would be apparent by the simulation in the `Form factor` panel not fitting the experimental data.

After a DEERNet fit, uncertainty of the reconstructed background is visualized in the `Original data` panel by plotting the mean background in red colour and the background functions reconstructed for the individual neural networks in the ensemble in pink colour. Furthermore, the mean DEERNet reconstructed background can be used for further analysis by other approaches and the background ensemble is used in validation of Tikhonov data, unless a new background analysis is performed in between.

7.6 User models

Generally, the solution of an ill-posed problem can be stabilized by introducing additional constraints. A distance distribution $P(r)$ that conforms to a simple model with only a few parameters, for example a distribution consisting of one or two Gaussian peaks, is strongly constrained. Fitting of the data by a model distribution can thus improve reliability of the analysis. Furthermore, by comparing the parameters for a series of related samples trends can be easily recognized. This approach is offered in DeerAnalysis by an interface for fitting pre-processed data by user-defined models for the distance distribution $P(r)$. Model functions with one and two Gaussian peaks are already implemented. The model library can be extended by the user as described below.

In applying this approach one should be aware that a model can impose constraints that do not apply to the true distance distribution and may thus suppress information contained in the original data. For instance, the example data set dOTP_5nm can be fitted relatively well by a distance distribution consisting of a single Gaussian peak, but this imposes a symmetry on the peak that is not a feature of the true distribution. The true distribution decays more steeply towards high distances than towards low distances as seen in the inset in Fig. 7a (and the reason for this asymmetry is well understood). It is thus advisable to perform a model-independent analysis by Tikhonov regularization first. From a set of distance distributions for the same class of samples, it is then often possible to derive a model function that does not impose undue constraints but does make use of additional information on the sample that comes from other characterization techniques.

7.6.1 Fitting with existing models

When DeerAnalysis starts, the program checks the subdirectory models for existing Matlab scripts (extension .m). The current distribution contains the scripts

- Chechik2.m
  Biradicals and monoradicals on a sphere surface. Single Gaussian peak with mean distance $\langle r \rangle$ and standard deviation $\sigma(r)$ superimposed on the distribution of objects on the surface of a sphere (triangular) with a Gaussian distribution of sphere diameters ds with standard deviation $s(ds)$. Parameter $p1$ quantifies the relative distribution of the single Gaussian peak.

- Gaussian.m
  Single Gaussian peak with mean distance $\langle r \rangle$ and standard deviation $\sigma(r)$.

- Gaussian_hom.m
  Single Gaussian peak with mean distance $\langle r \rangle$ and standard deviation $\sigma(r)$. and homogeneous 3D background. To be used with **Background model** No correction and only with great care.

- random_coil.m.
  Random-coil model for a polymer chain or an unfolded protein or an unfolded domain of a protein. N is the number of amino acid residues between the two labels including both labeled residues. ν is the scaling exponent (0.602 for good solvents, expected for soluble proteins in water, 0.5 for θ-solvent, less than 0.5 for poor solvents).

- Rice3d.m
  Single Rice peak [14] with mean distance $\langle \nu \rangle$ and standard deviation σ in three dimensions. Note that for the Gaussian limit of the Rice distribution, the standard deviation σ is by a factor $\sqrt{2}$ smaller than the value of σ in the Gaussian distribution as implemented in DeerAnalysis.

- Sphere_Surface.m.
  Homogeneous distribution of spin labels on the surface of a sphere. The sphere diameter d_s has a Gaussian distribution with standard deviation $\sigma(d_s)$.

- `Three_Gaussians.m`
  Three Gaussian peaks with mean distances $\langle r_1 \rangle$, $\langle r_2 \rangle$, and $\langle r_3 \rangle$ and standard deviations $\sigma(r_1)$, $\sigma(r_2)$, and $\sigma(r_3)$. The populations of the first and second peak (integrals) are p_1 and p_2, the one of the third peak is $1 - p_1 - p2$. Tends to be unstable.

- `Triangle_DGauss.m`
  Assumes a three-spin system (equilateral triangle) with double Gaussian distribution of the center-vertex distance (two Gaussian peaks). Useful for homotrimers with two distinct conformations or significantly non-Gaussian distance distributions. Pair- and three-spin contributions to the form factor are considered, based on [1] with a correction in the fraction of two-spin contributions. The total modulation depth Δ (`Delta`) and the number of Monte Carlo trials (`nmc`) are fixed parameters, whereas Δ should be set to the modulation depth obtained with background fitting.

- `Triangle_Gauss.m`
  Assumes a three-spin system (equilateral triangle) with Gaussian distribution of the distance between the center and the vertices. Useful for homotrimers. Pair- and three-spin contributions to the form factor are considered, based on [1] with a correction in the fraction of two-spin contributions. The total modulation depth Δ (`Delta`) and the number of Monte Carlo trials (`nmc`) are fixed parameters, whereas Δ should be set to the modulation depth obtained with background fitting.

- `Two_Gaussians.m`
  Two Gaussian peaks with mean distances $\langle r_1 \rangle$ and $\langle r_2 \rangle$ and standard deviations $\sigma(r_1)$ and $\sigma(r_2)$. The population of the first peak (integral) is p_1, the one of the second peak $1 - p_1$.

- `Two_Gaussians_hom.m`
  Two Gaussian peaks with mean distances $\langle r_1 \rangle$ and $\langle r_2 \rangle$ and standard deviations $\sigma(r_1)$ and $\sigma(r_2)$. The population of the first peak (integral) is p_1, the one of the second peak $1 - p_1$. The concentration for a homogeneous 3D background is c. To be used with **Background model** `No correction` and only with utmost care (do this only if you are desperate and interpret results with great caution).

- `Two_Rice3d.m`
  Two Rice peaks [14] with mean distances $\langle \nu_1 \rangle$, $\langle \nu_2 \rangle$ and standard deviations σ_1 and σ_2. The population of the first peak (integral) is p_1, the one of the second peak $1 - p_1$. Note that for the Gaussian limit of the Rice distribution, the standard deviation σ is by a factor $\sqrt{2}$ smaller than the value of σ in the Gaussian distribution as implemented in DeerAnalysis.

- `WLC_rigid.m`.
  Worm-like chain model [12] for a semi-rigid polymer (or DNA) with label-to-label distance L and persistence length L_p.

- `WLC_rigid_Gauss.m`.
  Worm-like chain model convoluted with a Gaussian distribution with standard deviation $\sigma(r)$ that accounts for conformational distribution of the label.

These models, and any models implemented by the user, are included in the `model fit` popupmenu of the `Distance analysis` panel. On selecting an entry of this menu, the parameter definitions, default values and limits of the corresponding model are read and the parameter controls in the `model fit` subpanel are updated. A model can have up to eight parameters. If it has less, superfluous parameter controls are disabled.

Models with Gaussian peaks that also include homogeneous background, are strongly discouraged. Their fitting behaviour was found to be rather unstable.

Before fitting, select the `model fit` radiobutton in the `Data analysis` panel. The `Distance distribution` plot now shows the APT result as a black dotted narrow line and the distance distribution corresponding to the current model and parameter values as a red dotted bold line. The `Dipolar evolution` plot displays the experimental data (black line) and the data simulated with the current model (red dotted line). You may now edit the starting values of the fit parameters

in the `model fit` subpanel until you obtain a reasonable agreement between experimental and simulated data. Of course, this step can be skipped and fitting can be started immediately, but by first improving your starting values you decrease the probability to get stuck in a local minimum of the error hypersurface. Before fitting you can also decide whether you want to fit all parameters (default behavior) or whether you want to keep some parameters fixed at their starting values. To fix a parameter, unselect the corresponding checkbox.

Fitting is started by clicking on the `Fit` button in the `model fit` subpanel. During fitting, the `Status` panel displays the current r.m.s. value. Note that fitting can be rather slow if the excitation bandwidth correction (see Section 7.7) is switched on. After the fit is completed, the parameter values are updated, the `Distance distribution` plot shows the fitted distance distribution as a black bold line, and the `Dipolar evolution` plot displays the experimental data (black line) and the fit (red line).

Model fitting considers the distance distribution in the range between 1 and 10 nm. For data sets extending to times longer than 4 μs, an upper limit of 10 nm may be too short if the homogeneous background is also fitted. As mentioned earlier, we strongly suggest to remove the background contribution before fitting.

7.6.2 Implementing a new model

The interface between DeerAnalysis2018 and the model scripts was designed to allow for writing model scripts without knowledge on the inner working of the main program. A model script has two input variables, a vector of distances `r0` at which values of the distance distribution have to be computed and a vector of parameters `par`. The only output parameter is the distance distribution, which is a vector of the same length as `r0`.

Note that the integral of the distance distribution can be arbitrary, as DeerAnalysis internally renormalizes the distribution to an integral of 0.01 for simulations and later computes the number of coupled spins from the modulation depth of the experimental data. This means that no amplitude parameter is needed. Only if the distribution corresponds of more than one contribution (for instance two Gaussian peaks), a parameter for the relative amplitude of an additional component with respect to the first component has to be defined. Consequently, a Gaussian distribution is defined by only two parameters, the mean distance $\langle r \rangle$ and the width (standard deviation) σ_r. A distribution consisting of two Gaussian peaks thus has the parameters $\langle r(1), \rangle$, $\sigma_{r(1)}$, the relative contribution of the first peak $p(1)$, and $\langle r(2), \rangle$, $\sigma_{r(2)}$. It is convenient to define the relative contributions so that they relate to the integral of the peaks (number of spins) and that $p(1) + p(2) = 1$. The model script `Two_Gaussians.m` is written this way.

A model script needs to declare its parameters to DeerAnalysis and provide default values as well as lower and upper limits for them. This is done in a comment section. As an example consider the script `Gaussian.m`:

```
function distr=Gaussian(r0,par),
%
% Model library of DeerAnalysis2011:  Gaussian
%
% single Gaussian peak with mean distance <r> and width (standard
% deviation) s(r)
% (c) G. Jeschke, 2006
%
% PARAMETERS
% name symbol default lower bound upper bound
% par(1) <r> 3.5 1.5 10
% par(2) s(r) 0.5 0.05 5
gauss0=(r0-par(1)*ones(size(r0)))/par(2); distr=exp(-gauss0.^2);
```

The first line is the function declaration, which is the same for all user models except for the function name (here `Gaussian`). The following lines, which start with the % character, are all comment lines, as far as Matlab is concerned. However, when the model is selected, DeerAnalysis2018 scans these comment lines in the source file for parameter declarations. A parameter

declaration line begins with the % character, followed by at least one space and the parameter name. Valid parameter names are `par(1)`, `par(2)`, `par(3)`, `par(4)`, `par(5)`, `par(6)`, `par(7)`, and `par(8)`. Only as many parameters have to be declared as are needed for the model (here 2). The parameter name is followed by at least one space and then by the parameter symbol. The symbol consists of at least one non-space character. It is shown as identification of the parameter control in the `model fit` subpanel. A symbol of up to five non-space characters can always be displayed, longer symbols are completely displayed only if some of the characters are narrow. The symbol is followed by at least one space and then the default value of this parameter. The default value is displayed in the edit field of this parameter and is the starting value for the fit if the user does not make any input before clicking on the `Fit` button. A good set of starting values provide for a distribution that is mainly confined between 1.5 and 8 nm and that clearly exhibits all relevant features of the model. The default value is followed by at least one space and the lower limit. No input smaller than this value is accepted by the edit field. Likewise, the value is used as a lower boundary in parameter fitting. The lower limit is followed by at least one space and the upper limit, which is analogous to the lower limit. Note that definition of the default values and limits is *mandatory*. Program response is undefined if the parameter line is incomplete.

Extended models can return both the distance distribution and the background corrected DEER trace including the constant offset part to DeerAnalysis. This is required when accounting for multi-spin effects. Declaration of such functions requires the keyword `#extended#` in the comment section. In both extended and standard user functions it is possible to limit the number of parameters that are fitted by default by using the keyword `#enable#` in the comment section. In multi-spin fits it may be necessary to switch off automatic modulation depth scaling between experimental and simulated traces in the display of DeerAnalysis. This feature has been disabled as it confused users of simpler models. If you need a private version that includes it, please contact gjeschke@ethz.ch.

```
function [deer,distr]=Triangle_Gauss(r0,t0,par)
%
% Model library of DeerAnalysis2008:  triangle with Gaussian distribution of
% vertex positions
%
% single Gaussian peak with mean distance <r> and width (standard
% deviation) s(r)
% (c) G. Jeschke, 2009
%
% #extended# denotes a model that provides both distribution and deer trace
% #enable# 2 only the first two parameters are fitted by default
%
% PARAMETERS
% name symbol default lower bound upper bound
% par(1) <rv> 2.5 0.5 10 mean distance from C3 axis
% par(2) s(v) 0.5 0.02 5 std.  dev.  of vertex position
% par(3) Delta 1 0.1 1 total modulation depth
% par(4) nmc 5000 1000 100000 number of Monte Carlo trials
```

7.7 Accounting for limited excitation bandwidth

Analysis of DEER distance measurements is usually based on analytical expressions, such as eqn (5), that assume ideal pulses. Past versions of our analysis programs accounted for this by suggestion a lower limit of 1.75 nm for the reliability of the distribution. Maryasov and Tsvetkov [17] first suggested to use corrected expressions to get more reliable results for short distances. Their approach considered the full Hamiltonian during the pulse, except for the pseudosecular contribution of the dipole-dipole coupling. They still assumed that the observed spins are not excited by the pump pulse and the pumped spins are not excited by the observer pulse. With these remaining assumptions, which are however not very well fulfilled, they could still obtain analytical expressions for the three-pulse DEER experiment. Based on these expressions, the effect of finite pulse lengths on determining distance distributions was assessed in a later contribution by Milov et al. [18].

To relax the remaining assumptions and extend the approach to four-pulse DEER, we examined the dependence of the modulation depth λ on the dipolar frequency ω_{dd} for typical lengths of the observer and pump pulses. Numerical density matrix computations of the full pulse sequence were performed for this purpose. Details will be published elsewhere. The dependence of λ on ω_{dd} can be aproximated quite nicely by a Gaussian function

$$\lambda(\omega_{dd}) = \exp\left(-\frac{\omega_{dd}^2}{\Delta\omega^2}\right) , \tag{10}$$

where $\Delta\omega$ is an effective excitation bandwidth with respect to dipolar frequencies. For a four-pulse DEER experiments with a pulse length of 24 ns for all pump and observer pulses and for an experiment with a 12 ns pump pulse and 32 ns observer pulses, we find the same excitation bandwidth of 16 MHz. For a four-pulse DEER experiments with a pulse length of 24 ns for all pump and observer pulses the excitation bandwidth is 12 MHz. The expression in eqn (10) can be used as a correction of the kernel function, eqn (5):

$$K(t, r; \Delta\omega) = \int_0^1 \exp\left(-\frac{\omega_{dd}^2}{\Delta\omega^2}\right) \cos\left[(3x^2 - 1)\,\omega_{dd}t\right] dx , \tag{11}$$

so that effects of finite pulses length can be accounted for without much additional computational effort if the kernel is anyway computed during fitting, as DeerAnalysis2018 does it during Tikhonov regularization.
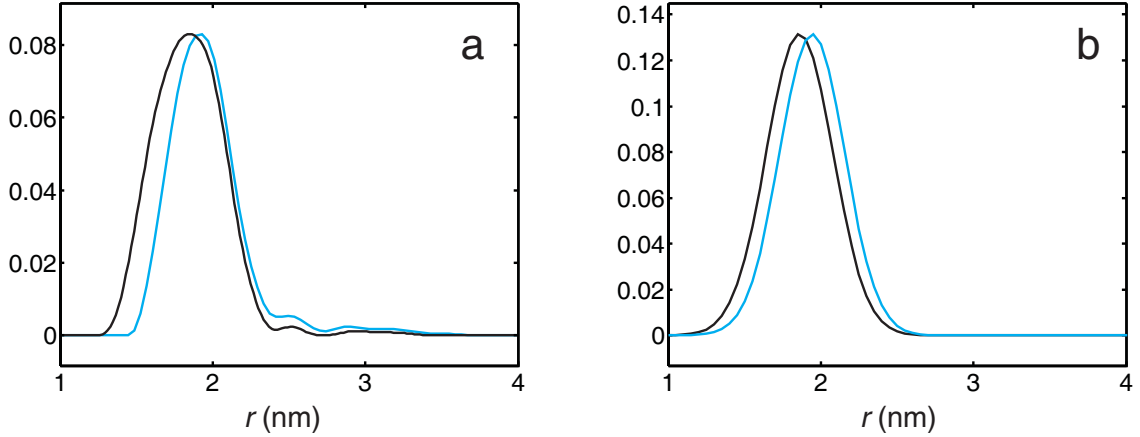

Figure 8: Excitation bandwidth correction. Blue distance distributions were obtained without, black ones with correction. a) Tikhonov regularization with optimum regularization parameter $\alpha = 1$. b) Fit by a single Gaussian peak.

However, simulations of the dipolar evolution function from a distance distribution, as they are required in model fits or at the end of Tikhonov regularization, can be performed with a pre-computed kernel for the expression given by eqn (5, while the kernel must be computed "on-the-fly" for the expression given by eqn (11. This is because the latter expression depends on an additional variable parameter $\Delta\omega$ and, furthermore, does not allow for scaling. In the former expression, a scaling of the t axis by a factor x can be compensated by scaling of the distance axis by a factor $x^{1/3}$. Without bandwidth correction, DeerAnalysis uses fast computations with a pre-computed ideal kernel. Therefore, bandwidth correction considerably slows down simulations and model fits and is thus not selected as default behavior of the program. It can be activated by selecting the `Exci. bandwidth` checkbox in the `Dipolar evolution` panel.

The effect of excitation bandwidth correction is illustrated in Fig. 8 for data set `deer_bi_oligo_n8_50K` from the `calibdepth` subdirectory. Data were cut off at 1504 ns to improve the background fit. Without correction (blue distributions) distances below 1.75 nm are strongly suppressed. With correction they are recovered. The r.m.s. deviation improves from 0.000320 without correction

to 0.000286 with correction in Tikhonov regularization and from 0.000396 without correction to 0.000335 with correction for a Gaussian fit. An improvement in the r.m.s. value may not always be found. The mean distance obtained with Tikhonov regularization changes from 1.97 to 1.85 nm. For a slightly longer flexible biradical (data set `deer_bi_oligo_n10_50K`), the correction is somewhat smaller, as the mean distance changes from 2.07 to 1.98 nm (data not shown). Note also that the Gaussian fits do not account very nicely for the true shape of the distribution in this case.

7.8 Ghost suppression

If more than two spin labels are present in a nanoobject, sum and difference combinations appear in the DEER signal [23]. This leads to ghost peaks in distance distributions that can be suppressed by power scaling of the form factor with an exponent $1/(N-1)$, where N is the number of spins per nanoobject, for instance the number of protomers in a symmetric protein homooligomer [13]. Such ghost suppression can be performed by activating the checkbox `ghost suppression` in the `Form factor` panel and inserting the number of spins in the edit field on the right of this checbox. It is advisable to perform data processing both with and without such suppression and to compare the distance distributions obtained. For detailed discussion of performance and limitations, see [13].

7.9 Test data sets

DeerAnalysis has a subdirectory `simulated` with Matlab scripts for generating test data with artificial noise and artificial background. The directory also contains a number of such test data sets: Gaussian distributions, distributions with two Gaussian peaks, a boxcar distribution, and a sawtooth distribution. These data sets are intended for test and training purposes.

Test data sets are loaded by selecting the `ASCII` subpanel of the `Data sets` panel. The default column assignment (Time 1, Real 2, Imaginary 3) applies. These data sets can be processed in the same way as experimental data sets. In the `Distance distribution` plot of the main window and of the validation tool the distance distribution used for generating these data sets is displayed in cyan color.

Precomputed test data sets are a series of Gaussian distributions with fixed ratio of ten between mean distance and standard deviation, a boxbar from 2 to 4 nm, a sawtooth from 2 to 4 nm, Gaussian distributions at 3 and 4 nm with two-dimensional background, and a few distributions consisting of two Gaussian peaks. Further test data sets can be computed by using the Matlab scripts `create_test_data.m` (single Gaussian, 3D background), `create_test_data_2D.m` (single Gaussian, 2D background), `create_test_data_nb.m` (two Gaussians), `create_test_data_special.m` (boxcar), `create_test_data_special_2.m` (sawtooth), and `create_test_data_special_3.m`.

8 Error analysis (validation) of distance distributions

For an ill-posed problem, the relation between noise in the input data and uncertainty of the output data is difficult to predict. Furthermore, background deconvolution is not exact for experimental data, which also introduces an error in the form factor. Often this error due to imperfect background correction even dominates the error in distance distribution.

While error propagation in Tikhonov regularization cannot be predicted analytically, there is an obvious numerical approach for such a prediction. Assume that the uncertainties of background correction can be modelled by variation of the starting time for the background fitting within certain bounds and of the dimensionality of the spatial distribution within certain bounds. Background correction can now be performed for a sufficiently large number n_{trials} of parameter sets within these bounds and the form factors obtained can be subjected to Tikhonov regularization. This provides n_{trials} distance distributions that can be statistically analyzed. Thus, a lower and upper limit, a mean value, and a standard deviation are obtained for each point in the distance distribution.

Alternatively, uncertainty of background parameters can be given in terms of lower and upper bounds for the density (proportional to concentration), the modulation depth, and the background

dimensionality. This approach was followed until DeerAnalysis2010 and is still available for backward compatibility. This old approach does not consider that the uncertainties in the background parameters may be correlated. Thus it tends to overestimate the error. It is more advisable to vary the starting time for the background fit and, if necessary, background dimensionality. Default behaviour is validation with respect to background fit range (start at 0.2 to 0.6 times the maximum length) and to white noise (adding 50% of the original noise level) with 110 individual runs. If you also vary dimensionality, you may want to reduce the number of noise trials from 10 to 2 or 3.

If dimensionality is also varied, some combinations of dimensionality and fit starting time may result to poor fits of the form factor. A poor fit can be defined as a fit whose root mean square deviation from the experimental data exceeds the root mean square deviation (r.m.s.d.) of the best fit by a factor L_{prune} (prune level). DeerAnalysis suggests a prune level of 1.15, but expert users can define this value according to their own experience. Parameter combinations that lead to fits with r.m.s.d. $> prune\ level \cdot \min(\text{r.m.s.d.})$ are excluded from statistical analysis.

The influence of noise on the distance distribution can be estimated in a similar way as the influence of uncertainties in background parameters. In this case noise is artificially enhanced by adding pseudorandom numbers so that the noise level is increased by a certain factor L_{noise}. Errors are somewhat overestimated when taking $L_{\mathrm{noise}} = 2$. A value $L_{\mathrm{noise}} = 1.5$ is suggested and is also the default. Tests for noise and background influence are by default combined to obtain a total error estimate.
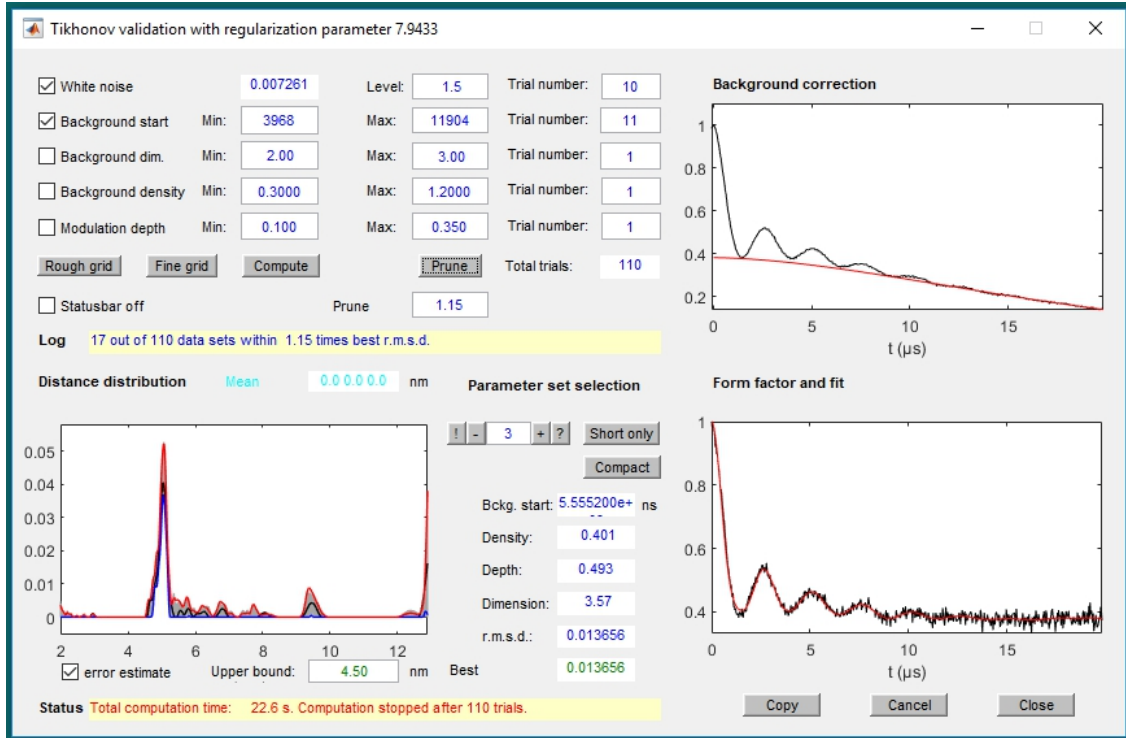

Figure 9: Screenshot of the validation window.

The validation tool is started by clicking on the `Validation` button in the `Tikhonov regularization` subpanel of the `Data analysis` panel. This button is set inactive when a Tikhonov regularization has not yet been performed. Validation uses the regularization parameter that is selected in the main window when the button is clicked. This regularization parameter is displayed in the title line of the validation window (Figure 9)

The user can now select what parameters have to be treated as uncertain (default: Background starting time), what lower and upper bounds are estimated for these parameters and how many values within these bounds should be tested. The trial values are distributed uniformly between

the bounds. Each allowed value of each variable parameter is combined with each allowed value of each other parameter. One should be aware that this may lead to large numbers of trials and that for each trial a Tikhonov regularization has to be performed. Computation times may be substantial. The `Rough grid` and `Fine grid` buttons make suggestions for the bounds and number of trials that are based on the result of the background correction performed in the main window. These suggestions should only be used if there is no independent information on the bounds.

After starting the computation by clicking on the `Compute` button, a progress bar will appear as soon as the first Tikhonov regularization has been performed. At that point an estimate of the remaining computation time is displayed. If necessary the computation can be interrupted by closing the progress bar. For that the × button in the upper right corner of the progress bar has to be clicked. After the next Tikhonov regularization is completed, a window appears that allows for interrupt or continuation of the computation.

After all trials are computed the `Distance distribution` plot displays the distance distribution with the best r.m.s.d. as bold green line, grey error bars that indicate the full variation of the probability of a given distance over all trials, a lower error estimate corresponding to the mean value of the probability minus two times its standrad deviation, and an upper error estimate corresponding to the mean value plus two times the standard deviation (red dotted lines). The mean distance is indicated by a vertical cyan dotted line. For test data sets, the distance distribution used in simulating the data is displayed as a cyan solid line. By using the - and + buttons left and right from the data set number display, all computed distance distributions (and the corresponding background functions and fits iof the form factor) can be inspected in turn. The `Short only` button selects the data set with the smallest contribution of long distances. The `Compact` button selects the set with the most narrow distribution. The `?` button selects the data set with largest r.m.s.d. of the fit from the experimental form factor and the `!` button the data set with the lowest r.m.s.d.

Poor fits can now be excluded from statistical analysis by selecting the prune level and clicking on the `Prune` button. Note that this is irreversible. The excluded data sets are lost and can only be recovered by repeating the computation.

Separate Matlab figures of all plots (for copying, saving or printing) are obtained with the `Copy` button. The validation tool can be left either with the `Cancel` button, leaving the original state of the main window intact, or with the `Close` button, transferring the mean distance distribution and the error bounds to the main window. A report of the validation is stored automatically in the latter case.

After returning to the main window, the mean distance distribution from all trials is displayed. If the `error est.` option in the `Distance distribution` panel is selected, the lower and upper error bounds (two times standard deviation) are displayed as grey error bars.

If a DEERNet computation was performed before Tikhonov regularization and the background model was not changed in between, behaviour of the `Validation` button changes. The validation window does not open anymore. Instead, the DEERNet reconstructed background functions are combined with 5 noise realizations each. This mode is also available for validating fits of parametrized models (user models), except for the ones that also fit homogeneous background.

9   Post-processing

For many cases, one wants to quantify the distance distribution in terms of a few numbers, i.e., mean distance and width of the whole distribution or of individual peaks. For oligomers of membrane proteins and self-assembled supramolecular systems, it may also be of interest to derive the number of spins within an individual nanoobject. All these values can be obtained by post-processing.

9.1   Moment analysis and peak picking

Analysis of a number of simulated and experimental DEER data sets suggested that the first moment (mean distance) and second moment (variance, square of the standard deviation) of the

distance distribution are stable parameters. In other words, these values are only very slightly influenced by noise-induced artificial splittings in the distance distribution. This applies in particular to the results of those techniques that incorporate the constraint $P(r) > 0$ (Tikhonov regularization and model fitting). Moment analysis of the distance distribution in the range of interest (default: 1.5-8 nm) is therefore performed automatically. The mean distance ($\langle r \rangle$) and standard deviation (s(r)) are displayed in the `Distance analysis` panel. To exclude obvious artifacts at the short or long end of the distance range (due to nuclear modulations or errors in background correction), you may change the range for analysis using the `+` and `-` buttons for the blue and magenta cursor in the `Distance distribution` panel or direct input into the corresponding edit fields. This option can also be used for extending the distance range if very long distances have been measured or for selecting only a single peak in a multimodal distance distribution and determining its mean distance and width. When the `Expand` checkbox is selected, the distance distribution is displayed only between the cursors.

9.2 Number of coupled spins

The number of spins within a nanoobject can be derived from the (calibrated) modulation depth if decay due to spins in other nanoobjects can be neglected, as was shown early on by the Novosibirsk group [22]. The same applies for the modulation depth in the dipolar evolution function after appropriate correction of the background decay [6]. The total modulation depth is given by

$$\Delta = 1 - \exp\left[\lambda\left(\langle n \rangle - 1\right)\right] , \tag{12}$$

where $\langle n \rangle$ is the average number of spins in the observed nanoobjects.

To use this information, DeerAnalysis therefore retains information on the modulation depth in the dipolar evolution function. Quantification requires knowledge of the modulation depth parameter λ, which depends strongly on the excitation position, length, and flip angle of the pump pulse and weakly on line broadening in the nitroxide spectrum and shape of the resonator mode. Reliable quantification therefore requires a calibration with known samples and proper adjustment of the flip angle of the pump pulse (see Section 12). Spectra from our own series of calibration samples (six biradicals and one triradical) and typical X-band 4-pulse DEER settings are provided in the folder `calibdepth`. They correspond to 12 ns π pump pulses irradiated at the maximum of the nitroxide spectrum (see Fig. 2) using a Bruker 3mm split-ring resonator. Note that not all example spectra in other folders were measured under the same conditions. To calibrate modulation depths for your own applications, you should measure at least one genuine biradical with close to 100% degree of spin-labeling under your measurement conditions. Analyse the data for this biradical, preferably with Tikhonov regularization and change the number of spins in the corresponding edit field of the `Distance analyis` panel to 2. The number is then displayed in green instead of red color. If another data set, measured under the same conditions, is loaded and processed, the displayed number of spins should correspond to the true average number $\langle n \rangle$ of spins in the nanoobject.

Note that this calibration is lost on restarting DeerAnalysis and that it is unreliable when using excitation bandwidth correction. Also consult Section 6.6.

9.3 Comparing data sets (dual display)

To compare two data sets of the same sample or of similar samples first load one of the data sets and process it as usual. To keep the same processing parameters for the second data set, you may then want to uncheck the `Reset` checkbox below the `Load` button in the `Data sets` panel. After loading the second data set, its file name is shown in line `A:` of the `Data sets` panel. This is the active data set. The file name of the previous data set is shown in line `B:`. The original data and processing results can now be compared by selecting the `Dual display` checkbox in the `Original data` panel. Traces corresponding to the previous data set are now shown in blue in all plots. In the `Dipolar evolution` plot, only experimental data, but no fits are shown for the previous data set.

If the two data sets differ considerably in their modulation depth, but have similar distance distribution, the samples may just differ in the extent of spin labelling or the measurement condi-
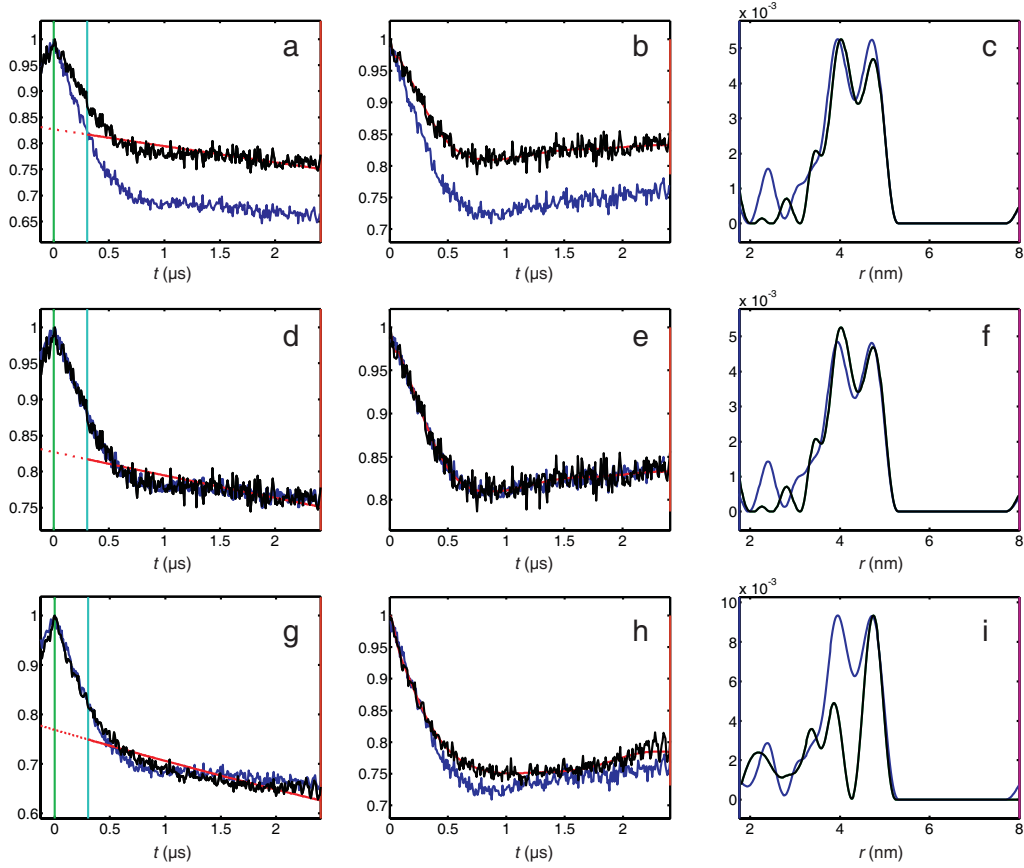
Figure 10: Dual display for comparison of two data sets. All data sets are from the subdirectory `examples\series`. Left column: Original data. Middle column: Dipolar evolution functions after background correction. Right column: Distance distributions. a-c) Comparison of data sets `series2` (active set A, black traces) and `series1` (set B) without modulation depth scaling. d-f) Comparison of data sets `series2` (active set A, black traces) and `series1` (set B) with modulation depth scaling. g-i) Comparison of data sets `series8` (active set A, black traces) and `series1` (set B) with modulation depth scaling.

tions (flip angles, resonator, pulse lengths) may have been slightly different. To check for this, use modulation depth scaling [6] by selecting the `mod. depth scaling` checkbox in the `Original data` panel. Differences in the distance distribution are noise-related if the original data are not significantly different after such modulation depth scaling.

10 Output

10.1 Saving data

Unlike some of its predecessor versions, DeerAnalysis does not automatically save results, except as an option during processing series of data sets (see Section 11). On attempt to close the program after time-consuming fits (Tikhonov regularization, model fits) without saving results, the user is reminded. The whole set of data including background correction, experimental and fitted dipolar evolution function and spectrum, distance distribution, processing parameters, results of moment analysis and fitted parameters and L curve (if available) are saved together with the same basis file name, but into different ASCII files.

After clicking on the `Save` button, the user is asked for the file name. The last extension and, if present, a suffix `_res` are removed to derive the basis name *basname* (this is useful for overwriting

old results by selecting their name in the diplayed file list). The following files are then saved:

- *basname*_res.txt
  a summary of the program settings and the results
- *basname*_bckg.dat
  the phase-corrected original data and background fit
  1st column: time axis (in μs),
  2nd column: real part of original data,
  3rd column: background fit
  4th column: imaginary part of original data (if present)
- *basname*_fit.dat
  the dipolar evolution function and its fit
  1st column: time axis (in μs),
  2nd column: dipolar evolution function after background correction,
  3rd column: fit of the dipolar evolution function
- *basname*_spc.dat
  the dipolar spectrum and its fit
  1st column: frequency axis (in MHz),
  2nd column: experimental dipolar spectrum,
  3rd column: fit of the dipolar spectrum
- *basname*_distr.dat
  the distance distribution
  1st column: distance axis (in nm),
  2nd column: distance distribution $P(r)$
- *basname*_Lcurve.dat
  the L curve of Tikhonov regularization (only if computed)
  1st column: $\log(\rho)$,
  2nd column: $\log(\eta)$,
  3rd column: corresponding regularization parameters α

The results file ***basname*_res.txt** protocols all relevant program settings, the mean distance, width of the distance distribution, and third moment, and for Tikhonov regularization, the regularization parameter. For model fits, the values of all fit parameters are also saved here.

10.2 Copying or printing individual plots

The three current plots of DeerAnalysis2018 can be copied into individual Matlab figures by clicking on the `Copy` button in the `Data sets` panel. Using the figure menu, the plots can then be rescaled, edited, annotated, printed, exported as different graphics formats or copied into the Windows clipboard (item `Copy figure` in the `Edit` menu). Matlab has a good help system that explains these possibilities.

11 Processing a series of similar data sets

This feature has not been tested in combination with recently introduced features of DeerAnalysis. Be careful if you use it and apply consistency checks.

A global analysis of several data sets is useful when measurements on the same sample have been reproduced or when samples have been prepared under slightly different conditions and one wants to check whether structural changes have occurred (see also Section 9.3). The first case requires computation of an average distance distribution that takes into account the signal-to-noise ratio of the individual data sets. In the second case the comparison should be performed for modulation-depth normalized primary data rather than for distance distributions as it is difficult to estimate what degree of change in the distance distribution is significant [6]. For both tasks a text file *listname*.txt has to be prepared that contains a list of filenames (without extension) of all the data sets that are to be processed together (for an example, see the file `series.txt` in the subdirectory `example\series`).

List processing starts with analysis of a pilot data set, which should ideally be the data set with the best signal-to-noise ratio. This data set with best signal-to-noise ratio should also be the first set in the list, as the first data set is used as a reference for modulation depth scaling. After loading the pilot data set it is processed as usual. Series processing is then initiated by the `Series` button in the `Data sets` panel. Progress is reported in the `Status` panel and line `A:` of the `Data set` panel. Plots are also updated (with a slight delay) during series processing. The program will return after the last data set has been processed. This data set is now the active data set.

The average distance distribution and average dipolar evolution function after series processing as well as average results of moment analysis are not displayed on screen, but are saved automatically. These files have the following formats:

- *listname*_res.txt
  a summary of the program settings and the results for the average
  of all data sets
- *listname*_mean.dat
  the mean dipolar evolution function
  1st column: time axis (in μs),
  2nd column: mean dipolar evolution function after background
  correction,
- *listname*_cmp.dat
  modulation-depth normalized primary data (without background
  correction)
  1st column: time axis (in μs),
  n remaining columns: primary data (real part) for data sets $1 \cdots n$,
- *listname*_diff.dat
  $n \times n$ matrix quantifying the difference between data sets
  large values in element (k, j) indicate that data sets k and j differ
  significantly

Primary data sets and distance distributions are averaged with a weighting factor that is inversely proportional to the mean square deviation of the fit of the dipolar evolution function. This corresponds to a maximum likelihood estimate of the average.

By default results for the individual data sets are not automatically saved. Automatic saving can be initiated by selecting the `Autosave` checkbox below the `Series` button. Note that even with this checkbox selected, automatic saving takes place only during series processing, not when processing individual data sets via the `Load` button.

12 Hints for Data Acquisition

Conversion of a dipolar evolution function as measured by a magnetic resonance experiment to a distance distribution is an ill-posed mathematical problem [7]. This means that even small deviations from the theoretical function (noise, phase problems, an intensity offset) can cause significant distortions in the distance distribution. Thus, it is of utmost importance to acquire experimental data with the *best* quality possible within a reasonable measurement time.

The choice of a number of experimental parameters has been discussed earlier [15]. From our own experience we suggest to perform measurements in the following way. A temperature of 80 K is a good compromise for most samples, but sensitivity is often somewhat better at 50 K. For critical samples such as membrane proteins, cooling to 50 K is often worth the effort. Unless the sample really has a strong signal, one should plan for measuring two samples in 24 hours, one during the day and one over night. Spectrometers tend to be stable enough over a period of several hours and the quality of the distance distribution tends to be limited by the signal-to-noise ratio except for synthetic model compounds with very narrow distance distributions. The observer and pump frequencies should be stable within 1 MHz during the measurement time, and this should be checked. It is good practice to acquire data with quadrature detection and to adjust the detector phase properly at the beginning. That way instability of the spectrometer can be recognized by the appearance of a significant imaginary part of the signal. Note that a small phase drift

(corrections up to $20°$ for a measurement extending over several hours) is no cause for alarm.

For four-pulse DEER on pairs of nitroxides at X-band frequencies we suggest that the pump pulse has a length of 12 ns. This can be achieved with a Bruker 3-mm-split-ring resonator. We also suggest that all the observer pulses have the same length of 32 ns. These conditions cannot be met at all spectrometers and with all probeheads. Using a length of 32 ns for all pulses, or a length of 16 ns for the $\pi/2$ pulses and a length of 32 ns for the π pulses also provides good results. If your pump π pulse has the same length as the π observer pulses, you may want to set the observer frequency to the center of the resonator mode and the pump frequency into the flank. Note however, that the opposite setting as suggested by Fig. 2 allows for a shorter pump pulse and hence larger modulation depth.

The power of the pump pulse should be adjusted for optimum flip angle (optimum echo inversion) using an inversion recovery sequence $\pi_{\mathrm{pump}} - T - \pi/2_{\mathrm{obs}} - \tau - \pi_{\mathrm{obs}} - \tau - echo$. This has to be done with coinciding pump and observer frequency at the position in the microwave mode where the pump pulse is applied. After this step the pump frequency should not be changed anymore. If this procedure is not followed, modulation depths are ill-defined and should not be compared between samples. The step is also an absolute requirement if concentrations are to be determined. We suggest that the pump pulse is applied at the maximum of the nitroxide spectrum, which maximizes modulation depth. This minimizes artifacts due to nuclear modulations, phase noise, and spectrometer imperfections.

The observer pulses are then applied at the low-field local maximum which corresponds to increasing the observer frequency (spectrometer frequency) by approximately 65 MHz. You may measure the field difference ΔB_0 between the two maxima and multiply it by 2.8 to obtain the exact frequency difference for your particular nitroxide. A phase cycle $(+x) - (-x)$ should be applied to the first observer pulse to eliminate offsets in the detector channels. If this phase cycling is omitted, any phase correction of the primary data will not be exact and hence background correction by program DeerAnalysis2018 will not be exact. Furthermore, modulation depth information is not reliable. In principle, the problems could be solved by introducing the offset as an additional parameter in background correction, but we strongly discourage such an approach, as it complicates separation of the dipolar evolution function from the background which may be difficult anyway for long distances.

For the interpulse delays in the four-pulse DEER experiment $\pi/2(\nu_{\mathrm{obs}})$-$\tau_1$-$\pi(\nu_{\mathrm{obs}})$-$t'$-$\pi(\nu_{\mathrm{pump}})$-$(\tau_1+\tau_2-t')$-$\pi(\nu_{\mathrm{obs}})$-$\tau_2$-$echo$ we suggest $\tau_1 = 200$ ns for protonated solvents/matrices and $\tau_1 = 400$ ns for deuterated solvents/matrices (at X band). To suppress proton modulations it is advantageous to perform the experiment at eight different values of τ_1 spaced by 8 ns and starting with the values given above. The signals of the eight experiments are added. In variable-time DEER [5] we suggest initial values $\tau_{2,0} = 300$ ns for protonated and $\tau_{2,0} = 500$ ns for deuterated samples. In constant-time DEER, $\tau_2 = 800$ ns is usually convenient for setup (adjustment of the detector phase). For the actual measurement, the choice of τ_2 depends on transverse relaxation, signal strength, and on the longest distances that have to be measured. It is difficult to give general suggestions, but the problem has been discussed in some detail in Ref. [5].

The integration gate should match the width of the observer echo, which is similar to the width of the longest observer pulses. The gate should be centered at the echo maximum. If you can save data in Xepr (Elexsys) format, DeerAnalysis 2004 can directly import binary data. For ESP 380 data we suggest importing them into WIN-EPR and saving them in the *binary* WIN-EPR format (this step converts the coding of binary float numbers to a format legible by MATLAB). If you use another data acquisition system, you should save your data in an ASCII representation.

Acknowledgments

References

[1] G. Jeschke, M. Sajid, M. Schulte, A. Godt. *Phys. Chem. Chem. Phys.* **11** (2009) 6580–6591.

[2] G. Jeschke, V. Chechik, P. Ionita, A. Godt, H. Zimmermann, J. Banham, C. R. Timmel, D. Hilger, H. Jung. *Appl. Magn. Reson.* **30** (2006) 473–498.

[3] O. Dalmas, H. C. Hyde, R. E. Hulse, E. Perozo. *J. Am. Chem. Soc.* **134** (2012) 16360-16369.

[4] M. Pannier, S. Veit, A. Godt, G. Jeschke, H. W. Spiess, *J. Magn. Reson.* **142** (2000) 331–340.

[5] G. Jeschke, A. Bender, H. Paulsen, H. Zimmermann, A. Godt, *J. Magn. Reson.* **169** (2004) 1–12.

[6] G. Jeschke, G. Panek, A. Godt, A. Bender, H. Paulsen, *Appl. Magn. Reson.* **26** (2004) 223-244.

[7] G. Jeschke, A. Koch, U. Jonas, A. Godt, *J. Magn. Reson.* **155** (2001) 72–82.

[8] P.P. Borbat, J.H. Freed, *Chem. Phys. Lett.* **313** (1999) 145–154.

[9] G. Jeschke, A. Godt, *ChemPhysChem* **4** (2003) 1328–1334.

[10] G. Jeschke, M. Pannier, A. Godt, H. W. Spiess, *Chem. Phys. Lett.* **331** (2000) 243–252.

[11] D. Hinderberger, H. W. Spiess, G. Jeschke, *J. Phys. Chem.* **108** (2004) 3698–3704.

[12] A. Godt, M. Schulte, H. Zimmermann, G. Jeschke, *Angew. Chem. Int. Ed.* **45** (2006) 7560–7564.

[13] T. von Hagens, Y. Polyhach, M. Sajid, A. Godt, G. Jeschke. *Phys. Chem. Chem. Phys.* **15** (2013) 5854–5866.

[14] S. Domingo Köhler, M. Spitzbarth, K. Diederichs, T. E. Exner, M. Drescher, *J. Magn. Reson.*, **208** (2011) 167–170.

[15] G. Jeschke, *ChemPhysChem* **3** (2002) 927–932.

[16] Y. W. Chiang, P. P. Borbat, J. H. Freed, *J. Magn. Reson.* **172** (2005) 279–295.

[17] A. G. Maryasov, Y. D. Tsvetkov, *Appl. Magn. Reson.* **18** (2000) 583–605.

[18] A. D. Milov, B. D. Naumov, Y. D. Tsvetkov, *Appl. Magn. Reson.* **26** (2004) 587–599.

[19] A. N. Tikhonov, V. Y. Arsenin, *Solutions of Ill-Posed Problems* (Wiley, New York, 1977).

[20] J. Honerkamp, J. Weese, *Cont. Mech. Therm.* **2** (1990) 17.

[21] J. Weese, *Comput. Phys. Commun.* **69** (1992) 99–111 (1992).

[22] A. D. Milov, A. B. Ponomarev, Y. D. Tsvetkov, *Chem. Phys. Lett.* **110** (1984) 67–72.

[23] G. Jeschke, M. Sajid, M. Schulte, A. Godt, *Phys. Chem. Chem. Phys.* **11** (2009) 6580–6591.

[24] T. H. Edwards, S. Stoll *J. Magn. Reson.* **288** (2018) 58–68.

[25] G. Jeschke, M. Pannier, A. Godt, H. W.Spiess *Chem. Phys. Lett.* **331** (2000) 243–52.

[26] L. V. Kulik, S. A. Dzuba, I. A. Grigoryev, Y. D. Tsvetkov *Chem. Phys. Lett.* **343** (2001) 315-24.

[27] S. Milikisyants, F. Scarpelli, M. G. Finiguerra, M. Ubbink, M. Huber *J. Magn. Reson.* **201** (2009) 48-56

[28] K. Keller, V. Mertens, M. Qi, A. I. Nalepa, A. Godt, A. Savitsky, G. Jeschke, M. Yulikov *Phys. Chem. Chem. Phys.* **19** (2017) 17856–17876

[29] S. G. Worswick, J. A. Spencer, G. Jeschke, I. Kuprov *Sci. Adv. accepted*