Bayesian Probabilistic Analysis of DEER Spectroscopy Data

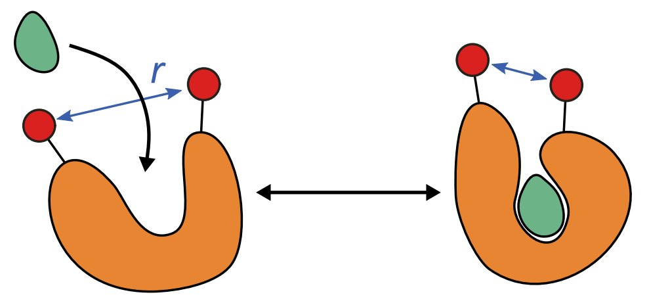
Stephan Pribitzer, Sarah Sweger, Stefan Stoll

Department of Chemistry, University of Washington, Seattle

Protein structure elucidation with EPR

Protein structure determination

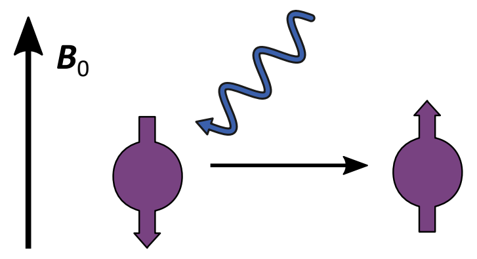
The function that a protein assumes depends to a large extent on its structure which makes protein structure determination of utmost importance for drug design. The human genome contains around 20 000 protein-encoding genes and yet, only the structures of a few of those proteins are known. Off particular interest is how proteins change their conformation when interacting with substrates.



A variety of methods exist to determine the atomic structure, one of which is electron paramagnetic resonance (EPR) spectroscopy, which studies unpaired electrons.

Principles of electron paramagnetic resonance

In the presence of an external magnetic field $\boldsymbol{B}_0$, an electron will align itself either parallel or antiparallel to the magnetic field. The orientation of the electron can be inverted by irradiating it with microwaves.

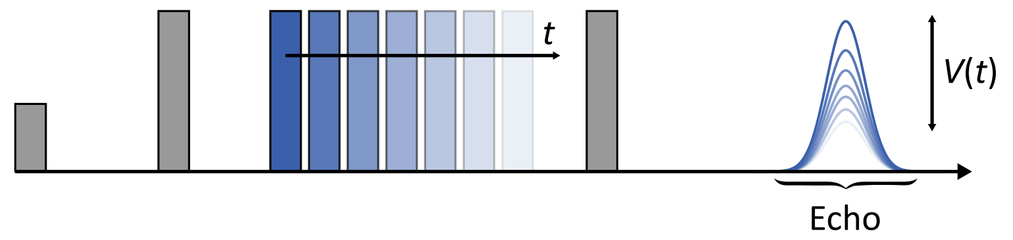


The microwave energy (the frequency) that is required to flip the electron depends on its chemical environment, in particular what nuclei and other electrons are nearby. After excitation the electron will slowly return to its initial state. This precession of the electron spin causes small fluctuations in the magnetic field that are picked up by coils that compose the detected signal.

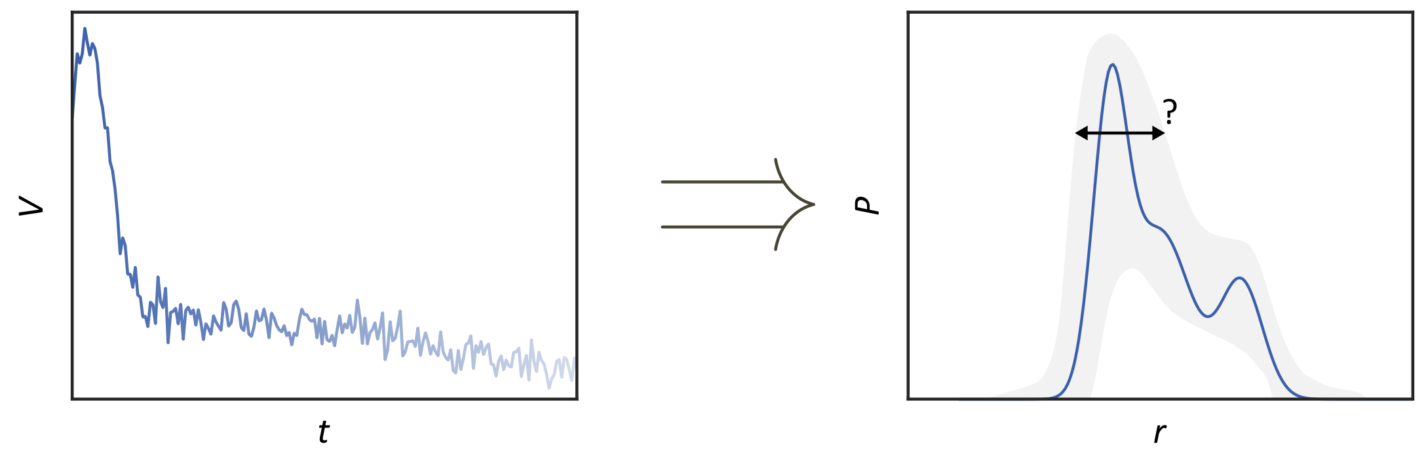
DEER spectroscopy

Double electron-electron resonance (DEER) is an EPR experiment that measures the distance r between two spin labels that are usually attached to the protein through biochemical methods. They contain unpaired electrons that are coupled via the dipolar

The strength of this coupling depends on the distance between the two electrons and is proportional to r^{-3}. In DEER spectroscopy a train of microwave pulses is used to excite the unpaired electrons of the spin labels; these pulses usually have lengths in the nanoseconds range. If set up correctly, a signal is detected that is usually reffered to as echo. The position of one of those pulses (the so-called pump pulse) is moved and the echo amplitude $V(t)$ is recorded for every position.



The recorded time domain trace exhibits modulations that depend on the dipolar couplings that are present. Similar to a Fourier transform, these oscillations can be transformed into the distance domain, yielding a distance distribution $P(r)$ of the interspin distance. This step constitutes an ill-posed, inverse problem that is usually solved with Tikhonov regularization .



DEER distance distributions can then be combined with other structure determination methods to refine atomic structure models of proteins.
Considering the role that DEER spectroscopy plays in structural biology, it is surprising that no method of reliably assessing uncertainty exists currently and most DEER data is indeed published without error bands. Here, we set out to quantify uncertainty in DEER spectrsocopy using a Bayesian approach.

> The ultimate goal of **DEER spectroscopy** is to obtain a **probability distribution** of the **spin-spin distance**, ideally including some sort of uncertainty assessment.

Markov chain Monte Carlo (MCMC) sampling

DEER theory

The noise-free DEER signal is

$$V_\mathrm{M}(t) = V_0 \cdot V_\mathrm{intra}(t) \cdot V_\mathrm{inter}(t)$$

where t is the position of the pump pulse and V_0 is the echo amplitude in absence of the pump pulse. $V_\mathrm{intra}(t)$ is the intramolecular modulation function given by

$$V_\mathrm{intra}(t) = (1-\lambda) + \lambda\int_0^\infty K(t,r)\,P(r)\,\mathrm{d}r$$

with the modulation depth λ and the normalized distribution $P(r)$ of the spin-spin distance r. $K(t,r)$ is called the dipolar kernel function and contains all the physics and quantum dynamics. $V_\mathrm{inter}(t)$ is the intermolecular modulation function

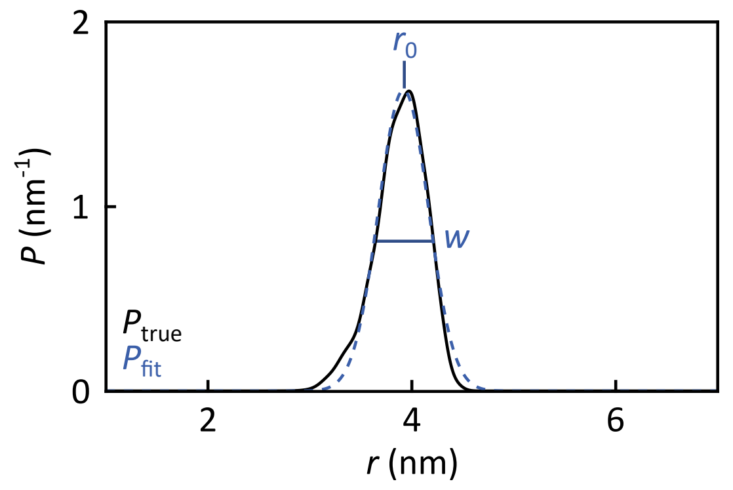
$$V_\mathrm{inter}(t) = \exp(-k|t|)$$

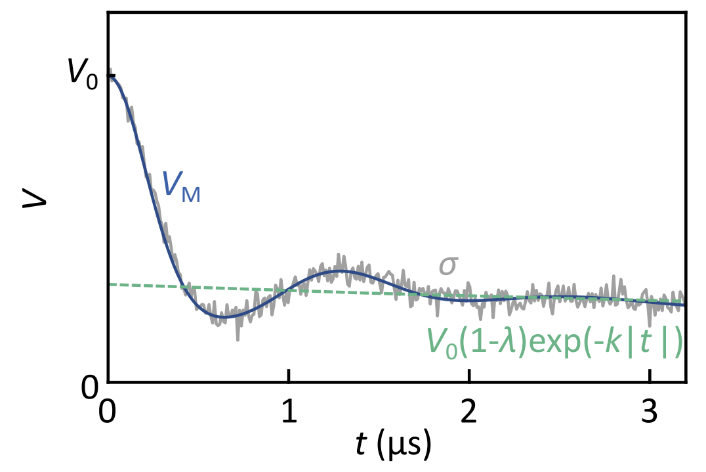
where k is the decay rate constant.
One way to approximate the spin-spin distance distribution $P(r)$ is a linear combination of normalized Gaussian basis functions:

$$P(r) = \sum_{i=1}^{m} a_i\,\mathrm{Gauss}(r;\, r_{0,i},\, w_i)$$

where m is the number of Gaussians, a_i the amplitudes, $r_{0,i}$ the centers of the Gaussians, and w_i are the full widths at half maximum.
In vector form, each measured data point can be written as a random sample from a Gaussian distribution with center $\boldsymbol{V}_\mathrm{M}$ and covariance matrix $\sigma^2\mathbf{1}$, where σ is the noise level:

$$\boldsymbol{V} \sim \mathrm{normal}(\boldsymbol{V}_\mathrm{M}(t_i),\, \sigma^2\mathbf{1})$$





Drawing from the posterior

We are interested in the posterior probability distribution of the parameter vector $\boldsymbol{\theta}$

$$\boldsymbol{\theta} = (\{r_{0,i}\}, \{w_i\}, \{a_i\}, k, \lambda, V_0, \sigma)$$

conditioned on the measured signal $\boldsymbol{V}$, model M and any additional information I.
The posterior can be expressed as

$$p(\boldsymbol{\theta}|\boldsymbol{V}, M, I) \propto p(\boldsymbol{V}|\boldsymbol{\theta}, M, I)\cdot p(\boldsymbol{\theta}|M, I)$$

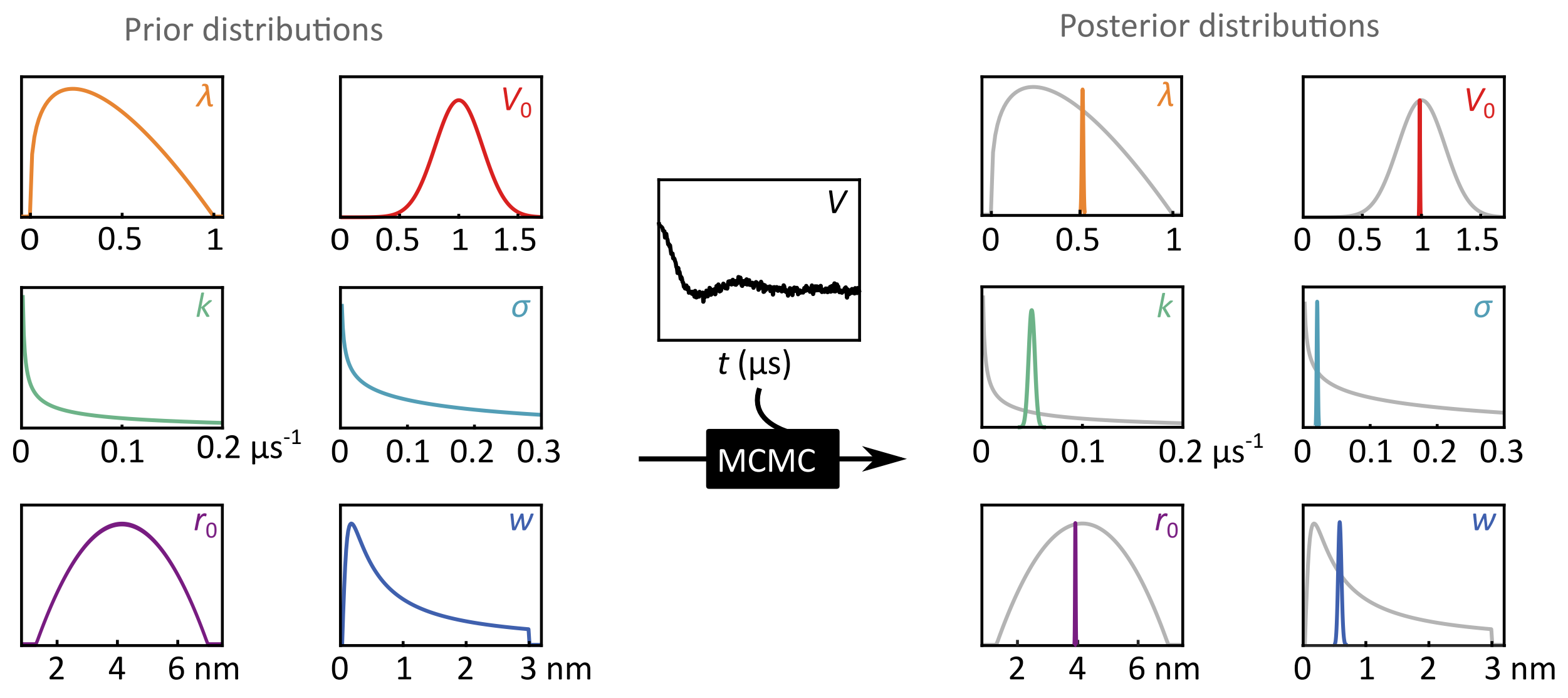
Here, $p(V|\boldsymbol{\theta}, M, I)$ is the likelihood

$$p(\boldsymbol{V}|\boldsymbol{\theta}, M, I) = \mathrm{normal}(\boldsymbol{V};\, \boldsymbol{V}_\mathrm{M}(\boldsymbol{\theta}),\, \sigma^2\mathbf{1})$$

which quantifies the degree of fit between the data and the model.
For each parameter θ_i from the parameter vector we can write prior probability distributions that can be combined

$$p(\boldsymbol{\theta}|M, I) = \prod_i p(\theta_i|M,I)$$

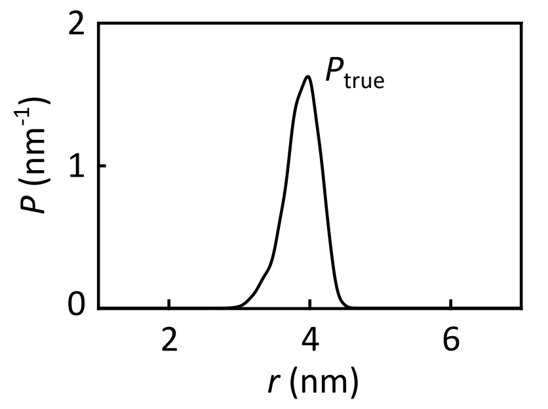
For the Markov chain Monte Carlo (MCMC) sampling from the posterior we use pymc3 [1] with a NUTS to generate a total of 8 MCMC chains. Each chain is initialized with different starting points and propagated for 5 000 steps to tune the sampler. The chains are then propagated for 20 000 - 80 000 steps. Convergence is assessed via the rank-normalized split $\hat{R}$ statistic [2].

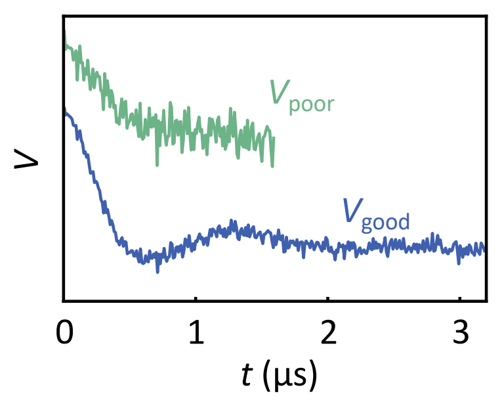


[1] J. Salvatier, T. V. Wiecki, C. Fonnesbeck, *PeerJ Comput. Sci.* **2016**, 2, e55.
[2] A. Vehtari, A. Gelman, D. Simpson, B. Carpenter, P.-C. Bürkner. *arXiv*, arXiv:1903.08008 [stat.CO] **2019**.

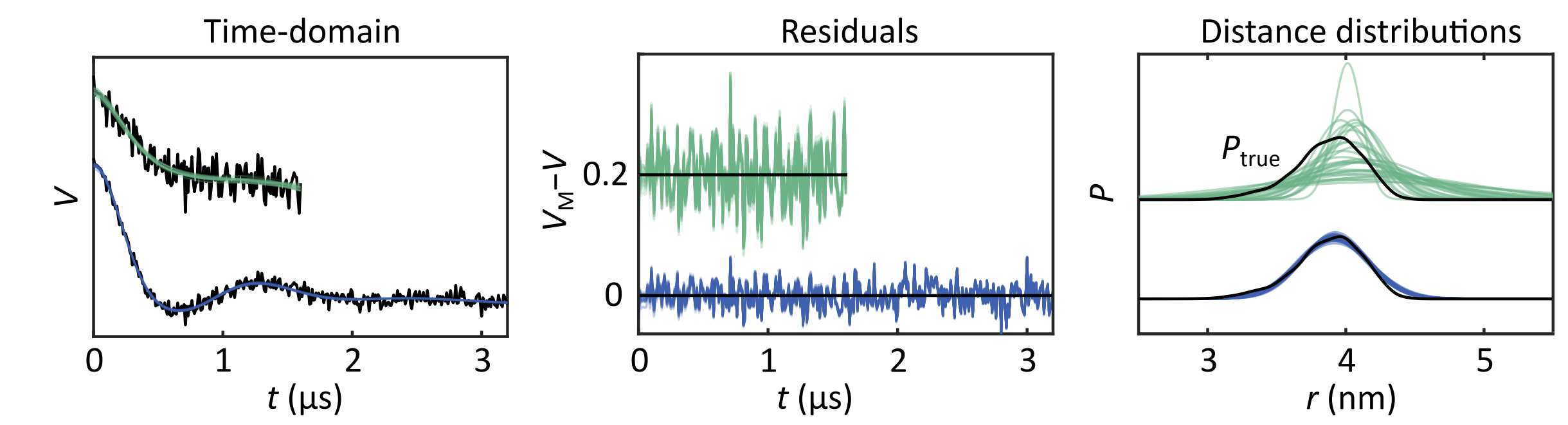
Single Gaussian analysis

The probabalistic analysis method is first shown for a synthetic data set with low complexity. We used a synthetic distance probability distrbution that is taken from the large simulated T4 lysozyme (T4L) test data [1], which resembles a single Gaussian.

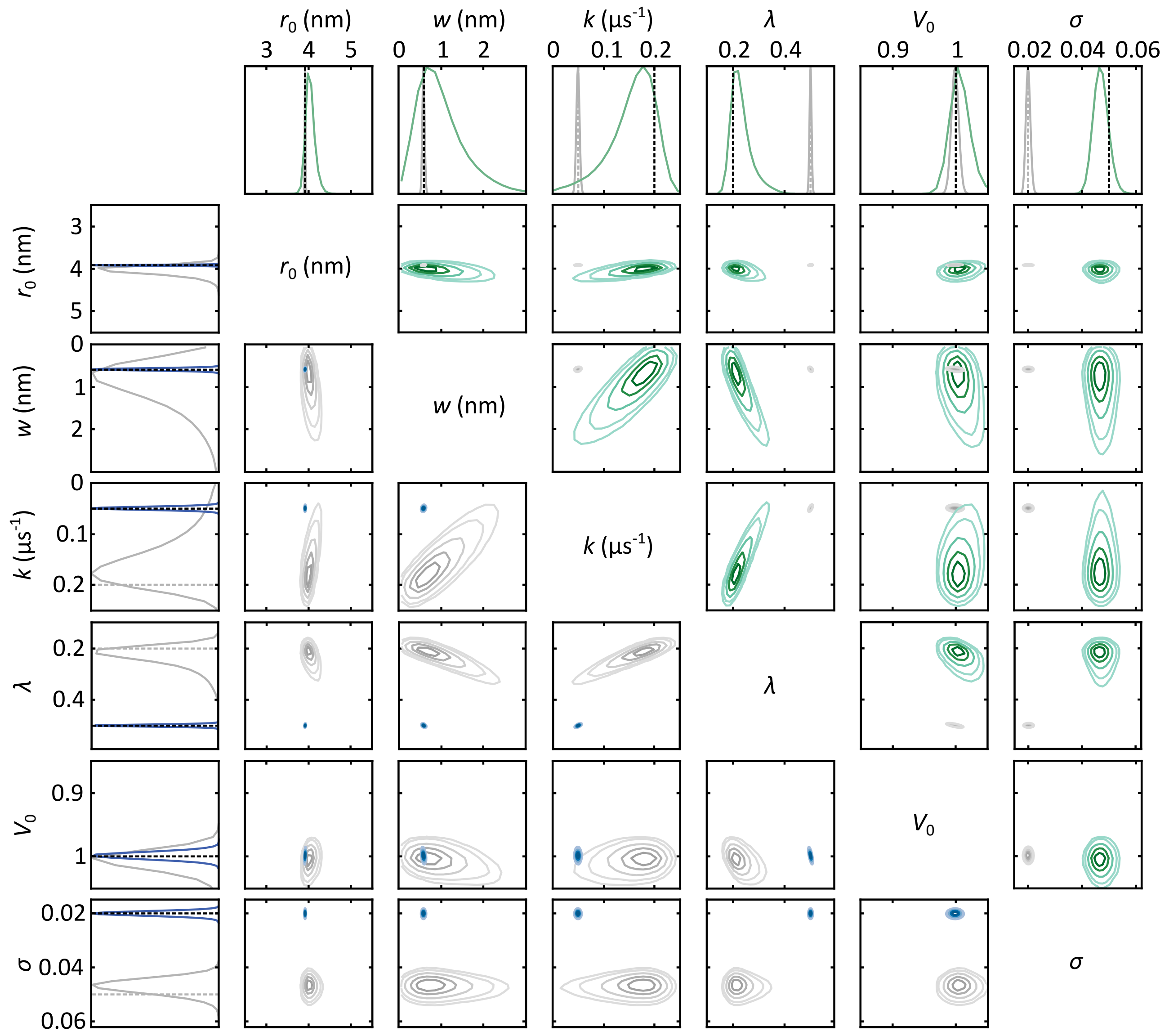




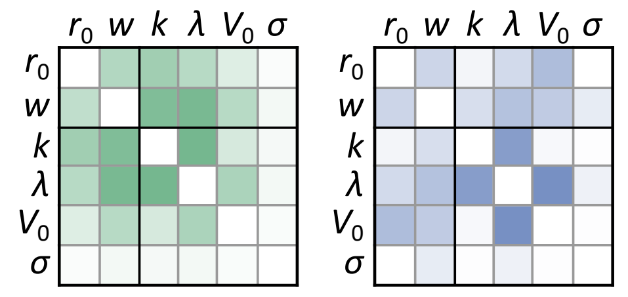
From this $P(r)$ we generated two noisy signal traces, one with favorable values V_good and one with less ideal values Vpoor for the modulation depth, background decay rate, trace length and noise level.
The figure on the right shows the results of the Bayesian analysis for both cases. The top row depicts the marginalized posteriors for each parameter for the poor case. The first column shows the same for the good case. Dashed lines represents the ground truths. Though both analysis yield distributions with modes close to the ground truth, the shorter and noisier trace clearly produces broader distributions. The rest of the plot shows the marginalized posteriors of all parameter pairs for both cases.
While parameter posteriors show the outcome of the Bayesian analysis most directly, the most desired quantities are the distance distributions and the model fit. To show these we draw a small set of random parameter vectors $\boldsymbol{\theta}^{(i)}$ from the pooled MCMC chain samples and calculate distance distributions and noise-free time domain signals.



> The interpretation of **posterior-based ensembles** of distributions is **essential** for truly visualizing the information extraced by the Bayesian analysis.



The pairwise marginalized posteriors present an overwhelming amount of information. To help with interpretation it is advantegeous to condense them to matrices of pairwise Peason correlation coefficients.

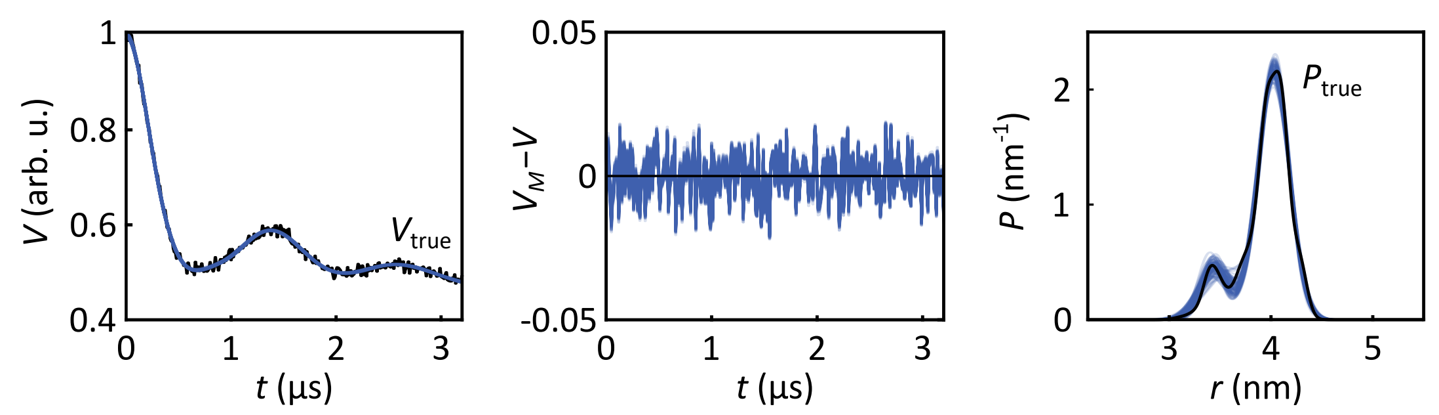


[1] T. H. Edwards, S. Stoll, *Journal of Magnetic Resonance* **2018**, 288, 58–68.

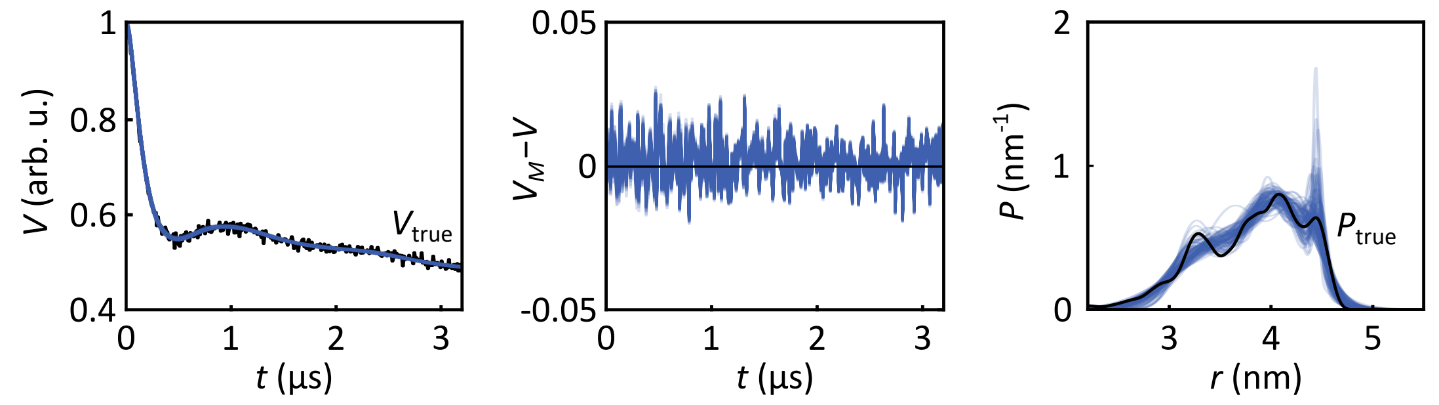
Multimodal distributions and model selection

Synthetic data

Most spin−spin distance distributions encountered in DEER spectroscopy of proteins are asymmetric and multimodal, and therefore poorly approximated by a single Gaussian. We analyzed a noisy time trace generated from a bimodal distribution from the T4L test set, with two distinct modes, one significantly weaker than the other.



The other example shown here is a challenging, broad distribution with several poorly resolved modes of similar intensities. The Bayesian analysis was conducted using a three-Gaussian distribution model.

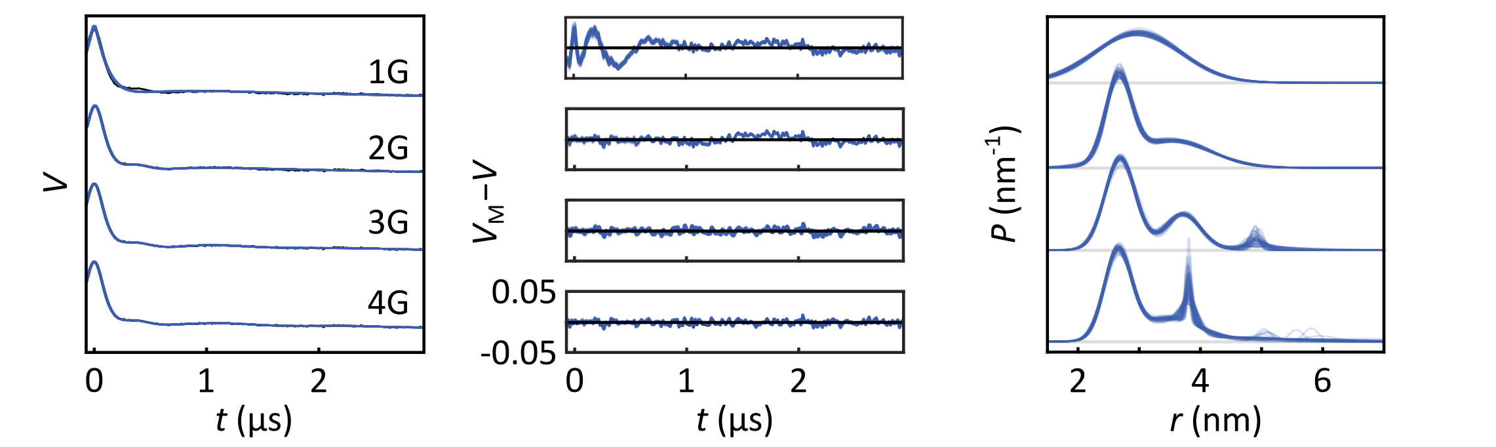


Challenges

Mixture models suffer from a phenomenon known as label switching. For example, switching the labels of the two Gaussians in a two-Gauss distance distribution changes the location in parameter space ($\boldsymbol{\theta}_1 \neq \boldsymbol{\theta}_2$), but does not affect the distance distribution ($P(\boldsymbol{\theta}_1) = P(\boldsymbol{\theta}_2)$) nor the likelihood or the posterior. This renders the posterior multimodal, complicating both the sampling and the analysis of the posterior. We take an approach similar to online relabeling [2] and enforce the constraints $r_{0,1} \leq r_{0,2} \leq \cdots \leq r_{0,\mathrm{m}}$ after every sample to restrict the parameter space. Occasionally, due to theimposed constraints, chains get stuck in regions with $r_{0,i} \approx r_{0,j}$, corresponding to the coalescence of two basis functions.

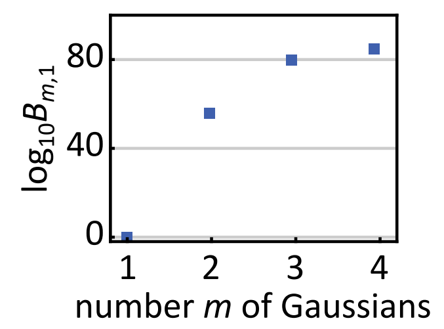
Experimental data

When the ground-truth distribution is not known, it is important to compare the quality of models with different numbers of Gaussians. We analyzed a DEER trace that was collected experimentally to determine intersubunit distances in SthK, a tetrameric bacterial cyclic nucleotide-gated (CNG) ion channel, with 1 to 4 Gaussians.



Both the 1G and 2G model show systematic deviations, even though they provide an apparently precise $P(r)$. The time-domain signal is well-described by both the 3G and 4G models.
Though not ideal, yet useful in this situation, a formal approach for model comparison and to identify overfitting is the calculation of the Bayes factors [3]. The Bayes factor is the ratio of the posterior odds of two models:

$$B_{i,j} = \frac{p(\boldsymbol{V}|M_i, I)}{p(\boldsymbol{V}|M_j, I)}$$



A $\log_{10}B_{i,j} > 8$ can be seen as a relatively strong indication for M_i over M_j. The Bayes factors $B_{\mathrm{m},1}$ for all m-Gauss models relative to the 1G model, give preference for 3G and 4G. However, the relatively small Bayes factor $\log_{10}B_{4,3} \approx 5$ indicates that the 4G is starting to overfit the data.

> Both, **residuals** and **Bayes factors** are required for **complete analysis**. While the former diagnoses systematic misfitting, the latter can identify overfitting.

Acknowledgments

This work was supported by the National Science Foundation with Grant CHE-1452967 (S.S.) and by the National Institutes of Health with Grants R01 GM125753 and R01 GM127325 (S.S.).

[1] E. G. B. Evans, J. L. W. Morgan, F. DiMaio, W. N. Zagotta, S. Stoll, *Proc Natl Acad Sci USA* **2020**, 117, 10839–10847
[2] M. Stephens, *J Royal Statistical Soc B* **2000**, 62, 795–809.
[3] R. E. Kass, A. E. Raftery, *Journal of the American Statistical Association* **1995**, 90, 773–795.