# A maximum entropy-based approach for the description of the conformational ensemble of calmodulin from paramagnetic NMR

**by: Francois Thelot**

**advisors: Mauro Maggioni and Bruce Donald**

Thesis submitted as partial requirement towards
graduation with distinction in Mathematics
Duke University, Durham N.C

# 1 Abstract

Characterizing protein dynamics is an essential step towards a better understanding of protein function. Experimentally, we can access information about protein dynamics from paramagnetic NMR data such as pseudocontact shifts, which integrate ensemble-averaged information about the motion of proteins. In this report, we recognize that the relative position of the two domains of calmodulin can be represented as the evolution of one of the domains in the space of Euclidean motions. From this perspective, we suggest a maximum entropy-based approach for finding a probability distribution on $SE(3)$ satisfying experimental NMR measurements. While sampling of $SE(3)$ is performed with the ensemble generator EOM, the proposed framework can be extended to uniform sampling of the space of Euclidean motions. At the end of this study, we find that the most represented protein conformations for calmodulin corresponds to conformations in which both protein domains are in close contact, despite being largely different from each other. Such a representation agrees with the random coil linker model, and sharply differs with the extended crystal structure of calmodulin.

# 2 Introduction

Far from being the motionless objects depicted on the protein data bank, proteins are flexible molecules, evolving under environmental thermodynamic fluctuations. In this view, we can conceive proteins as existing, at a given instant, as one of the many possible structures that are part of a perhaps infinitely large conformational ensemble. It is likely that a probabilistic conception of protein structure is a more realistic representation of nature, and will lead to a better insight in protein function.

In this study, we choose to focus on calcium-binding molecule calmodulin (CaM), which can be modelled by two terminal domains of similar-looking structure, tied by a seemingly alpha-helical linker region. It is the "model system" for many dynamic studies in protein dynamics, and has been used in several previous studies[1,2]. While the linker appears to be rigid on the crystal structure of the protein (pdb: 3CLN), we suspect that the alpha-helical

region must behave as a flexible coil, with bend and twist modes[3]. In addition, paramagnetic enhancement relaxation enhancement studies of the central part of the alpha-helical region (residues K77 to S81) suggest that the 5-residues central region of the linker can be modeled as a random coil, and that a significant proportion of the observed calmodulin conformations are characterized by a close contact between the two terminal domains, which would be impossible with a rigid linker[4]. Considering the functional role of calmodulin, such flexibility is hardly surprising: calmodulin is part of numerous signal transduction pathways, and its activation via calcium acquisition leads to binding to a large number of different protein targets and topologies.

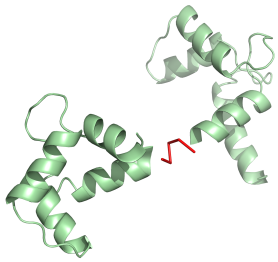

**Figure 1:** Crystal structure of calmodulin, showing the "extended" conformation. The central linker, marked in red, likely behaves as a random coil

We use paramagnetic NMR spectroscopy as a way to obtain dynamic information about our protein system. The introduction of a lanthanide ion attached to the protein leads to an interaction between the unpaired free electron of the metal with the nuclear spins of the system, affecting their procession and relaxation. The paramagnetic experiment allows for the recording of observables that reflect protein structure and dynamics, such as residual dipolar couplings (RDC), which integrate angular information between internuclear vectors and an applied magnetic field, and pseudocontact shifts (PCS), which provide both angular and translational information in a single measurement. We hope that the combined use of RDC and PCS will allow us to characterize the relative positioning of both calmodulin domains, in the space of Euclidean motions.

Confronted to a similar problem, Qi et al.[5] noticed that the relative angular positioning

of the two domains of a simplified model of protein A in staphylococcus aureus could be modeled by the rotation aligning the frame of one protein domain onto the domain, and is therefore characterized by a point in $SO(3)$ space. Qi et al. then used RDC data to fit the parameters of a Gaussian-like probability distribution on $SO(3)$, which provides a continuous and probabilistic insight in the relative orientation of both domains of protein A, according to experimental restraints. Rather than being an immediate application of the method introduced by Qi et al. on a new protein, our study introduces two improvements: 1) the use of PCS in addition to RDC should allow us to include a description of both rotations and translations of one of the protein domains in space, in a more accurate and intuitive depiction of reality; 2) the use of a maximum entropy approach for selecting a probability distribution over the conformational ensemble, therefore freeing ourselves from the constraints of imposing a Gaussian model, which is necessarily unimodal and subject to over-fitting due to the limited number of experimental restraints.

The final product of this approach is a probability distribution on an infinitely expandable ensemble of conformations of calmodulin. Such probabilistic description of protein dynamics is useful not only because it gives us an insight in the tumbling of the protein in a media, but also because it opens the door for the computation of thermodynamic quantities such as conformational entropy. In the case of calmodulin, Marlow et al[6] demonstrated that the conformational entropy of the protein represents a significant contribution component of its binding energy to a ligand. We could then envision a series of paramagnetic NMR experiments in which calmodulin is either free or bound to different types of ligand. The gain or loss of conformational entropy recovered from the probability distribution on $SE(3)$ might then be related to the affinity of calmodulin for several classes of ligands.

# 3 Theoretical background

## 3.1 NMR restraints

The first type of NMR restraint we are exploiting are Residual Dipolar Couplings (RDC), which integrate angular information about the angle between NH internuclear vectors in the

protein and the applied magnetic field. In the notation introduced by Donald[7], the recorded observable can be expressed as: $D = \frac{K}{2}\langle v^T S v \rangle$, in which $K$ is a physical constant, $v$ is the normalized NH vector for which the RDC is recorded, and $S$ is the alignment tensor for the protein domain considered: a 3x3 symmetric and traceless matrix characterizing the anisotropic tumbling of a protein molecule in solution. In this notation, the angle brackets are used to denote that the recorded value is averaged over time, and therefore over conformational ensemble. Assuming that both protein domains behave as strictly rigid bodies, we use experimental RDC data to obtain the alignment tensor of each domain, according to Prestegard's method[8]. Once we have obtained an alignment tensor for each protein domain, we use Qi's framework[5], which relates the alignment tensors of both domains by a matrix of rotation-averaged terms. Qi's relationship is: $\mathbf{s}_{II} = \mathbf{E}[Q]\mathbf{s}_I$, where $\mathbf{s}_I$ and $\mathbf{s}_{II}$ are vectors containing the 5 independent parameters of the alignment tensor of each domain, and $\mathbf{E}[Q]$ is a 5x5 matrix, in which every entry is a function of the parameters of a rotation, averaged over $SO(3)$. Each entry of matrix $\mathbf{E}[Q]$ takes the form: $\mathbf{E}[Q]_{ij} = \int_{SO(3)} Q_{ij}(q)p(q)dq$ where $q$ represents one rotation in $SO(3)$, $p(q)$ is a probability distribution over $SO(3)$, and $Q_{ij}(q)$ is the function of rotation corresponding to $\mathbf{E}[Q]_{ij}$. One remarkable property of Qi's relationship is that $\mathbf{s}_I$, $\mathbf{s}_{II}$ and $Q_{ij}$ can be defined such that all values of $Q$ are orthogonal to each other; that is: $\int_{SO(3)} Q_{ij}(q)Q_{kl}(q)p(q)dq = 0$ except if $i = k, j = l$. The orthogonality of the entries of $Q$ are of prime importance in a maximum entropy framework, and guarantees both the convexity of the Lagrange dual and the existence of a unique set of Lagrange parameters satisfying the constraints[9].

The second type of NMR restraints we are exploiting are pseudocontact shifts (PCS), which integrate angle and distance information about the internuclear vector ranging from the lanthanide ion to another nucleus in the sample (in this study, to amide hydrogens). While PCS are often expressed as: $\delta = \frac{K'}{\|r\|^3}(3\cos^2(\theta) - 1)$, for $K'$ a physical constant, $r$ the internuclear vector between lanthanide and amide hydrogen, and $\theta$ the angle between $r$ and the applied magnetic field $B$. We follow the RDC proof in Donald's *ASMB*[7] and derive instead:

$$\delta = \frac{K'}{\|r\|^3}\left(3\frac{(B.r)^2}{\|r\|^2} - \frac{r^T I r}{\|r\|^2}\right) = \frac{K'}{\|r\|^3}\left(3\frac{r^T B B^T r}{\|r\|^2} - \frac{r^T I r}{\|r\|^2}\right) = \frac{K' r^T (3BB^T - I) r}{\|r\|^5}$$

Recognizing that the alignment tensor is defined as $S = (3BB^T - I)$, we have found the more compact form of the PCS equation: $\delta = K'\langle\frac{r^T S r}{\|r\|^5}\rangle$, in which the brackets account for averaging over the conformational ensemble of the protein. While the PCS equation looks very similar to the definition of RDCs, we must emphasise that $r$ in the PCS equation represents the internuclear vector between lanthanide and some amide hydrogen anywhere in the molecule, and is therefore different from the $v$ vector from the RDC equation. While $v$ must necessarily relate two atoms of the same protein domain, the two extremities of $r$ can lie on different domains. This point is absolutely critical, and explains why PCS directly integrate distance information about relative protein domain motion. A perhaps more subtle point is that the PCS equation only factors in the alignment tensor of the domain on which the lanthanide is attached. While the relative tumbling of both protein domains certainly impacts the value of the recorded chemical shift, the tumbling of the domain on which no lanthanide is attached is unnecessary because it is entirely characterized by the interdomain $r$ vector, averaged over conformational ensemble.

We therefore have two type of experimental restraints. The first type originates from Qi et al.'s RDC framework, in which we must find a probability distribution $p_R(q)$ over $SO(3)$ such that, for $F_i$ a function of the domain alignment tensors and $G_i$ a function of Qi's $Q$ matrix :

$$F_i(\mathbf{s}_I, \mathbf{s}_{II}) = \int_{SO(3)} G_i\big(Q(q)\big)p(q)dq$$

The second type of experimental restraint comes from the PCS equation. With $q$ a rotation in $SO(3)$, $t$ a translation in $\mathbb{R}^3$ and $m(q,t)$ a combined rotation-translation in $SE(3)$, we want to find a probability distribution $p(m)$ over $SE(3)$ such that:

$$\delta_i = K' \int_{SE(3)} \frac{r_i^T S r_i}{\|r_i\|^5} p(m)dm$$

If there exists infinitely many distributions satisfying the constraints above, we wish to pick the distribution that is as close to uniform, and therefore makes the least assumptions about

the protein system. There are several philosophical reasons for justifying this choice, many of which are discussed Dudìk[10], and can essentially be understood as: without prior knowledge of the different states of a system, it is reasonable to assign equal probabilities to each state. The uniformity of a distribution over a space $D$ is named "information entropy" and represented by $-\int_D p(x)\log p(x)$ for all $x \in D$. Maximizing the uncertainty in our model is then equivalent to maximizing this entropy term. In the discrete case, we can then rephrase our problem of identifying maximally uniform probability distributions over $SO(3)$ or $SE(3)$ as:

maximize:

$$-\sum_x p(x)\log(p(x))$$

subject to:

$$\sum_x r_i(x)p(x) = m_i$$

where $i = 1..n$ are used to number the experimental constraints $m_i$ and functions $r_i$ corresponding to these observables. This is the classic maximum entropy problem.

## 3.2 Maximum entropy distributions

The general solution to this problem is derived from setting up the Lagrangian $L$ and solving for $\frac{\partial L}{\partial p} = 0$. For a vector $\lambda$ of $n$ Lagrange multipliers, the solution has form:

$$p_\lambda(x) = \frac{\exp\big(\sum_{i=1}^n \lambda_i r_i(x)\big)}{Z(\lambda)}$$

with:

$$Z(\lambda) = \sum_x \exp\big(\sum_{i=1}^n \lambda_i r_i(x)\big)$$

It can be shown that the expression $\log\big(Z(\lambda)\big) - \sum_{i=1}^n \lambda_i m_i$ is an upper bound for the entropy of the distribution (by exploiting the fact that Kullback-Leibler divergence is always non-negative), and that we can find the for the set of Lagrange multipliers satisfying the moment constraints by minimizing this potential over $\lambda$. When the moment functions are either uncorrelated or linearly correlated[9], we can show that the Hessian matrix of $\log Z(\lambda)$

is positive semi-definite, and therefore that $\log Z(\lambda)$ is convex in $\lambda$ parameter space. In practice, the function $\log\big(Z(\lambda)\big) - \sum_{i=1}^{n} \lambda_i m_i$ can often be minimized by gradient descent. The problem of finding a set of Lagrange multipliers satisfying experimental constraints can be summarized as:

$$\min_{\lambda} \log\big(Z(\lambda)\big) - \sum_{i=1}^{n} \lambda_i m_i$$

If moment functions are correlated in a non-linear manner, there is no guarantee that the search occurs on a convex space. While this is not an issue when solving for the maximum entropy distribution satisfying RDC constraints(Qi et al.[5] showed these can be orthogonalized), it is problematic in the study of PCS.

### 3.3 Maximum Entropy with Prior

We want to see whether we can take advantage of our capacity to reliably determine the Lagrange multipliers in the RDC problem, in order to improve our results in the PCS problem. We can do so by approaching the PCS problem in a different way: instead of maximizing the entropy of the distribution, we want to minimize the relative entropy (KL divergence) between the final distribution and a prior distribution. For a prior probability distribution $p_0$, that is minimizing:

$$D\big(p|p_0\big) = \sum_{x} p(x) \log\big(\frac{p(x)}{p_0(x)}\big)$$

The new form of the distribution is:

$$p_\lambda\big(x\big) = \frac{p_0(x)\exp\big(\sum_{i=1}^{n} \lambda_i r_i(x)\big)}{Z\big(\lambda, p_0\big)}$$

where:

$$Z\big(\lambda, p_0\big) = \sum_{x} p_0(x)\exp\big(\sum_{i=1}^{n} \lambda_i r_i(x)\big)$$

We then minimize:

$$\min_{\lambda} \log\big(Z\big(\lambda, p_0\big) - \sum_{i=1}^{n} \lambda_i m_i$$

If we choose the prior distribution to be uniform, this approach is exactly equivalent to searching for the maximum entropy distribution satisfying the constraints. Instead, we note that $p_0(q,t) = p_0(q)p_0(t|q)$, and that $p_0(q)$ can be reliably determined, since the moment functions in the RDC case are orthogonal. We then define $p_0(t|q)$ as a uniform function, and use $p_0$ as a prior. There are two advantages to this choice: we improve our chances of finding a satisfying set of Lagrange multipliers in the PCS problem, and we factor in an additional dataset (the RDC set) increasing the overall number of experimental restraints.

### 3.4 Maximum entropy with hard margins

We are given an important number of inter-domain pcs constraints, which means that the gradient descent search is in a very high dimensional space. This is a problem if the measurements are imprecise, because the search might be over constrained,leading to the practical impossibility of finding a set of Lagrange multipliers satisfying the observables. Instead of imposing strict equalities between observables and expected moment constraints, we define small margins around the experimental values such that the search is subject to inequality rather than equality constraints. This modification should account for the experimental inaccuracy associated with collecting physical measurements, and should limit the natural tendency of maximum entropy methods to over-fit data. The equations for a maximum entropy method with hard margins are provided in Kazama and Tsujii:[11]

Maximize

$$-\sum_{x} p(x) \log(x)$$

subject to:

$$\sum_{x} r_i(x)p(x) - m_i - A_i \leq 0$$

$$m_i - \sum_{x} r_i(x)p(x) - B_i \leq 0$$

where $A_i$ and $B_i$ are vectors defining the width of the margin for each experimental constraint. It is necessary to introduce twice the number of Lagrange multipliers (say two vectors $\alpha$ and

$\beta$ of dimension $n$) as we did before, and to restrict the search for non-negative multipliers. In this case, the new form of the distribution is:

$$p_{\alpha,\beta}(x) = \frac{\exp\big(\sum_{i=1}^{n}(\alpha_i - \beta_i)r_i(x)\big)}{Z(\alpha,\beta)}$$

where:

$$Z(\alpha,\beta) = \sum_x \exp\big(\sum_{i=1}^{n}(\alpha_i - \beta_i)r_i(x)\big)$$

This is solved by minimizing :

$$\min_{\alpha,\beta} \log\big(Z(\alpha,\beta)\big) - \sum_{i=1}^{n}(\alpha_i - \beta_i)m_i - \sum_{i=1}^{n}\alpha_i A_i - \sum_{i=1}^{n}\beta i B_i$$

### 3.5 maxEnt with soft margins

Kazama and Tsujii[11] also introduce a maximum entropy variant with soft margin, in which the constraint violation is weighted by an arbitrary penalty factor, and added to the objective function. This reminds of SVM with soft margins. Naturally, the choice of penalty constants $C_1$ and $C_2$ becomes a whole new problem.

We maximize

$$-\sum_x p(x)\log(x) - \sum_i C_1\delta_i - \sum_i C_2\gamma_i$$

subject to:

$$\sum_x r_i(x)p(x) - m_i - A_i \leq \lambda_i$$

$$m_i - \sum_x r_i(x)p(x) - B_i \leq \gamma_i$$

Here the Lagrange dual becomes:

$$\min_{\alpha,\beta} \log\big(Z(\alpha,\beta)\big) - \sum_i(\alpha_i - \beta_i)m_i - \sum_i\big(\alpha_i A_i + \frac{\alpha_i^2}{4C_1}\big) - \sum_i\big(\beta i B_i + \frac{\beta i^2}{4C_2}\big)$$

Just like MaxEnt with hard margins, soft margins can easily be applied to our problem.

## 3.6 Sampling

In an ideal world, we want to sample the space of Euclidean motions as uniformly as possible, allowing for a faster convergence to a maximum entropy distribution in Lagrange parameter space. The uniform grid on the space of Euclidean motion would be generated by limiting sampling to a bounding box around calmodulin in $\mathbb{R}^3$, and using the software developed by Yershova et al.[12] for creating a uniform grid on $SO(3)$. The advantage of such a sampling protocol is that we easily control the precision of the grid, and make sure that no region of the space remains unexplored.

In practice, however, uniform sampling of a bounding box around calmodulin means that a very large portion of the sample corresponds to physically inaccessible regions of the conformation space. This is highly problematic: a "realistic" grid in the 6-dimensional space of rotations and translations in a large box around the protein would require a tremendously large number of points, making the computation intractable due to insufficient computing power. In addition, in case the experimental observables are insufficiently constraining, the very nature of maximum entropy frameworks on such a grid will lead to relatively uniform distributions within the bounding box, attributing relatively high probabilities to absurd conformations.

While there might exist smarter ways to sample the space than assigning a "bounding box" and while it might be possible to come up with uniform samplings that are limited to physically reachable regions of $SE(3)$, we choose to take an imperfect route and use the Ensemble Optimization Method (EOM) software[13,14]. This software makes very weak assumptions about the physics of the protein systems, and models the flexible linker of calmodulin as a random coil, supposedly only sampling feasible regions of the conformational space. The disadvantage of this method is that we are unable to guarantee either uniform or complete sampling of the space of realistically attainable conformations. The obvious advantage of this method is that it significantly reduces the number of points sampled in $SE(3)$, from at least several millions using the bounding box approach, down to about 10000. While the bias introduced by EOM might be reduced by sampling more conformations, this framework will never provide the desired sampling guarantees and is nothing more than a quick and

easy means to an end. Further effort should focus on improving the bounding box sampling method, reducing the number of sampled points while maintaining complete and uniform sampling of realistic conformations.

There exists one compromise between systematic, uniforme sampling and EOM sampling. We first generate an EOM conformation sample, then apply the Kabsch algorithm to recover a point in $\mathbb{R}^3$ and in $SO(3)$. We then use these point clouds in order to make an EOM-based complex hull, which delimitates a smaller region than the original bounding box or of the four-dimensional sphere. Within the convex hulls, we then perform uniform sampling, using a small resolution grid in $\mathbb{R}^3$ and the Yershova et al.[12] sampling in $SO(3)$. One way to implement this method consists of generating a grid over the whole space ($\mathbb{R}^3$ box or $SO(3)$), then remove all points that fall outside the convex hull. While this approach theoretically reduces the number of points required for uniform sampling, the total number of transformations required to obtain a 'realistic' grid resolution still far exceed the capabilities of my laptop computer. As a result, we conducted this study solely based on EOM types of sampling. Convex hull bounding however illustrates how we believe the sampling method should be improved in the future.

# 4 Framework

A flowchart representing the different steps of the suggested framework is shown in Figure 2. From RDC data, we compute the alignment tensors for each domains of calmodulin. In parallel, we run EOM and obtain a conformational ensemble, and compute the points in $SE(3)$ corresponding to this conformational ensemble, such that we recover the non-uniform sampling on $SO(3)$ generated by EOM. We then apply the method introduced by Qi et al. and obtain a maximally-uniform discrete distribution on $SO(3)$, for the points corresponding to the non-uniform EOM grid. We then use the variant of maximum entropy with a prior distribution, and with both hard and soft margins. The prior distribution is set to be $p_0(q)p_0(t|q)$, where $p_0(q)$ is obtained from RDC data, and where $p_0(t|q)$ is a uniform distribution. The margin size (in the case of hard margins) and the penalty weight (in the case of soft margins) is varied throughout several trial, and allows for a more informed choice

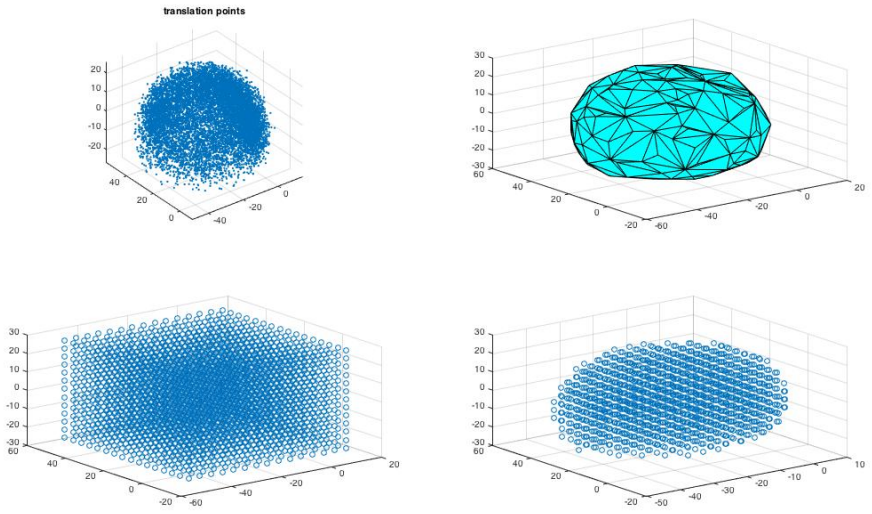


**Figure 2:** Example of convex hull bounding. The point cloud on the top left-hand figure is produced from EOM sampling using the Kabsch algorithm. From this point cloud, we make a convex hull(top right-hand). We then accept all points of a uniform grid that fall within the hull (bottom right-hand)

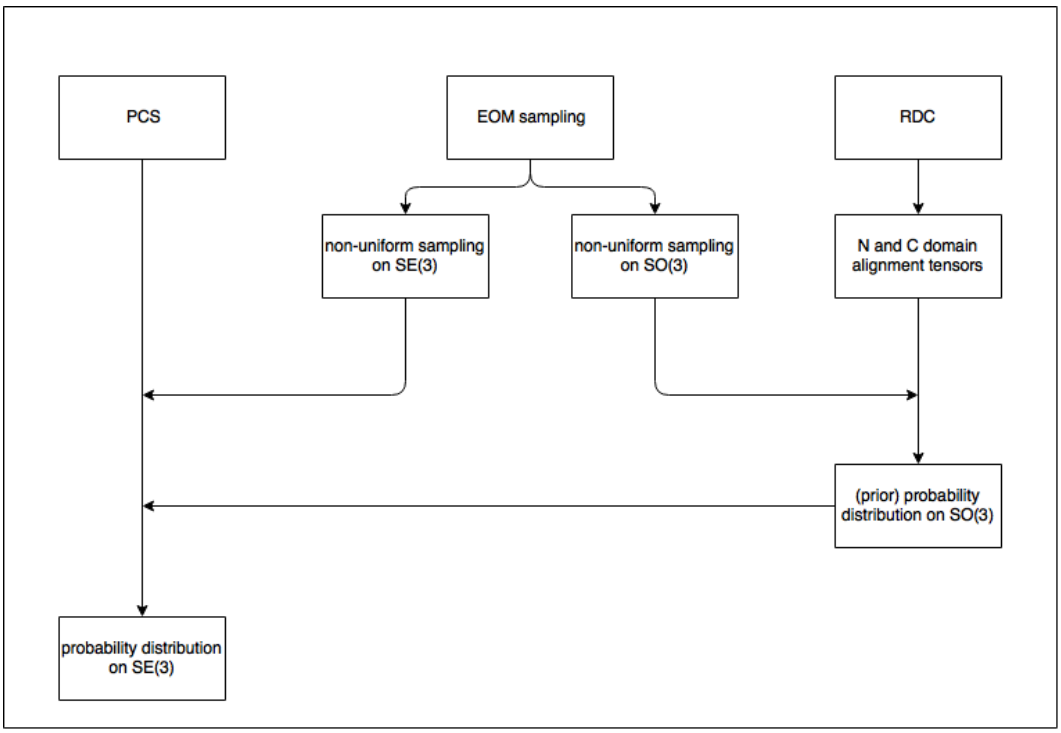


**Figure 3:** Suggested framework

of margin size and penalty. The integration space corresponds to the non-uniform grid on $SE(3)$ generated by EOM.

# 5 Methods

The ensemble of conformations was generated with the Ensemble Optimization Method (EOM) software , published by the Biological Small Angle Scattering Group at European Molecular Biology Laboratory in Hamburg[13,14]. We provided the software with structure files of the two terminal domains of calmodulin determined by NMR[15] (pdb: 1J70 and 1J7P). EOM is set up to keep the two terminal domains of calmodulin rigid, and to allow the 5-residues linker (K77 to S81) to behave as a random coil between the two termini. 10000 structures were generated in this manner.

Experimental data was generously provided by Dr. Giacomo Parigi from the Magnetic Resonance Center (CERM) at the University of Florence. It consists of RDC and PCS data for both domains of calmodulin, obtained with 3 different lanthanides (Dysprosium, Terbium, Thulium). Assuming that both termini are perfectly rigid bodies tumbling in solution, the RDC measurements were used to find the alignment tensor of the N and C terminal domains of Calmodulin in all three alignments, using singular value decomposition on the linear system introduced by Prestegard[8]. These alignment tensors were used both as a starting point for finding the prior distribution on $SO(3)$, and as components of the PCS moment functions. The search for a maximum entropy distribution was constrained by 132 PCS constraints, in 3 different alignments. With the hard margins protocol, we relaxed experimental constraints by considering constraints to be "satisfied" if they fall within margins of width 0, 0.004, 0.008, 0.012, 0.016, 0.020 ppm. With the soft margin protocol, the penalty weight was set to be the same on both sides of the inequality constraint, and was set at 1, 10, 100, 1000, 10000, 100000. All scripts were written either in python or Matlab and are available upon request.

# 6 Results

We followed the suggested framework for several hard margin widths, and obtained corresponding probability distributions over EOM ensemble. Each time, we are constraining the search with 132 PCS values. The distributions we obtained are characterized in the plots of Fig. 4.

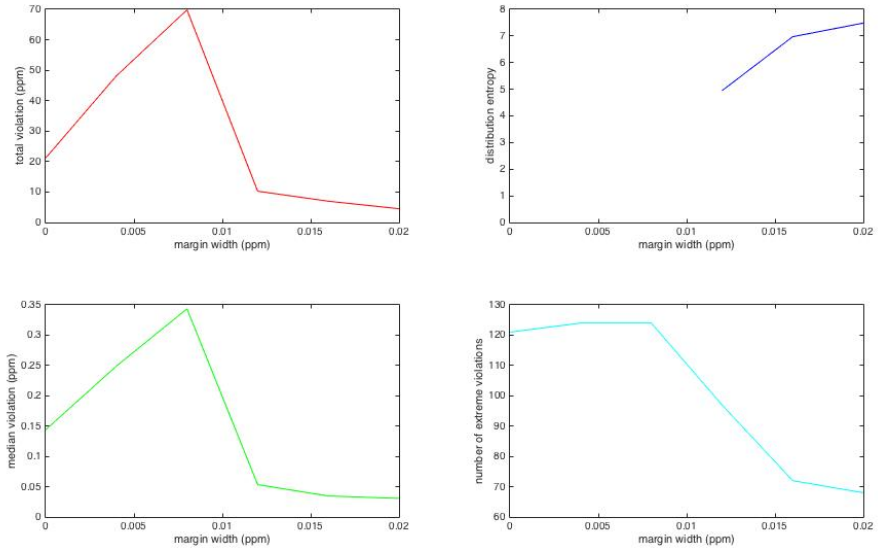


**Figure 4:** Hard margin plots. The -x axis is always hard margin width, while the -y axis represents successively total violation (ppm), distribution entropy, median violation (ppm), and number of violations above 0.03ppm

Considering the results shown in Fig. 4, we immediately note that small margins lead to worse distributions (sometimes worse constraint violations, and lower entropy) than wider margins. The missing entropy values in the entropy plot, which correspond to entropy values of zero, are characteristic of over-constrained systems, in which all the probability weight is put on a single or very few points in space. The relatively high constraint violation might at first be surprising considering the small margin width, but can most likely be explained by observing that the incredible penalty associated with a very negative entropy would lead the algorithm to deviate from regions of the parameter space with good restraint satisfaction. In contrast, the region with highest margin ($A_i = B_i = 0.008$ppm and more for all $i$) gave

both the best restraint satisfaction, and the most uniform distribution, showing the benefits of relaxing experimental constraints when using the maximum entropy framework. While we did not increase the margin width to higher values, we predict that total entropy would keep on increasing, while restraint satisfaction would deteriorate. From this perspective, the problem of maximizing restraint satisfaction and entropy is that of finding an optimal margin width value. In the future, with more computational resources, we should be able to conduct the framework for more margin values, and attempt to identify optimal margin widths for our problem.

In order to further look into the problem of optimizing margin width, we followed the soft margin approach described in the methods section and in Kazama and Tsujii.[11] The plots characterizing the distributions we recovered are shown in Fig. 5.

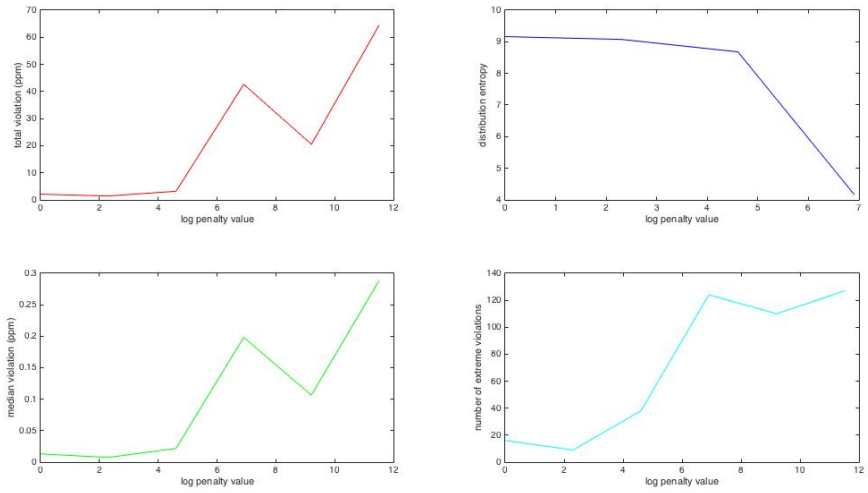


**Figure 5:** Soft margin plots. The -x axis is always the natural log of the penalty weight, while the -y axis represents successively total violation (ppm), distribution entropy, median violation (ppm), and number of violations above 0.03ppm

From Fig. 5, we immediately see that increasing penalty weight past a certain threshold deteriorates the quality of the recovered probability distribution. This finding is consistent with the hard margin results: when distributions are only slightly penalized for small disagreements with experimental constraints, the recovered probability distribution has both high entropy and low restraint violation. In addition, we note that the best solutions from

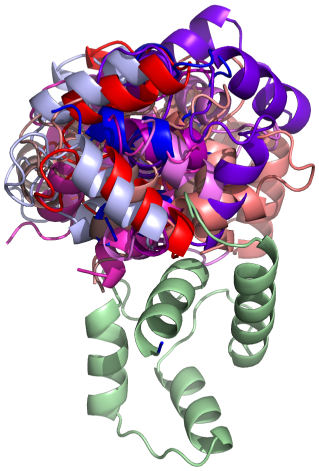

**Figure 6:** 10 most likely conformers according to one distribution derived from soft margins. The common domain is shown in green; all C-domain conformations are colored from red (higher probability) to blue (lower probability)

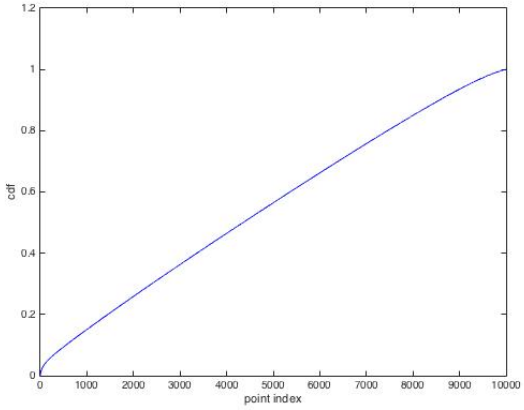


**Figure 7:** cumulative density function of the distribution derived from soft margin. Higher probability points are summed first

the soft margin protocols (penalty constraint of 1 or 10) have higher entropy and lower restraint violation than the best hard margin distributions: soft margins perform better than hard margins in the inter-domain dynamics problem. We take the distribution from the soft margin trial corresponding to a penalty weight of 1, and superimpose the structure of the corresponding 10 most likely conformers. The 10 most likely conformers are represented on Fig. 6.

According to the best distribution we have found (best restraint satisfaction and highest entropy), the 10 most likely conformations appear to be restricted to one region of the space. Perhaps surprisingly, the most likely conformations do not correspond to the extended, alpha-helical linker model, and favor a compact representation rather than an elongated one. Moreover, the cumulative density function of the probability distribution shows that the distribution we have found is quasi-uniform (its entropy is 9.16; that of a uniform distribution would be 9.21), suggesting that the C-terminal domain in fact bounces all over the EOM-

space, as if the linker behaved as a random coil. This result is in agreement with the conclusions of Emberly et al.[3], and is at odds with the rigid crystal model.

While the recovered distribution is nearly uniform, which supports our intuition that the C-terminal domain is indeed experiencing random tumbling, our distribution performs better than a uniform distribution in terms of restraint satisfaction. Although our distribution does not perfectly satisfy experimental restraints (median violation is around 0.008ppm, which corresponds to 4-5Hz in most spectrometers), we account for such a discrepancy by the non-uniform nature of EOM sampling.

# 7 Discussion

We have introduced and conducted a maximum-entropy based framework for finding a probability distribution over a set of computer-generated conformational ensemble of calmodulin, restricted by PCS and RDC constraints. In the end, we find that the most likely conformers define a 'compact', folded calmodulin conformation. All regions of the EOM conformation space however remain largely accessible, validating the random-coil model of the calmodulin linker and suggesting a quasi isotropic tumbling of the C-terminal domain.

While conducting this study, we were surprised to find similarities with the December 2015 work of Camilloni and Vendruscolo[2], who used both RDC and PCS constraints as a potential energy term in molecular dynamics simulations of calmodulin. Their approach is designed such that the structural ensemble generated is in agreement with recorded experimental measurements, and relies on the proof of equivalence between Jaynes' maximum entropy approach and restrained-ensemble molecular dynamics simulation, in the large number of replicas of the protein system[16,17]

Although the authors describe their method as a "maximum entropy framework", a great distinction with our approach is that the Camilloni et al. method does not use molecular dynamics as a starting point for calculations, but instead implements paramagnetic observations as a component of molecular dynamics. As a result, the method is inherently dependent on force field imperfections, numerical approximations, and incomplete sampling of the simulation. The framework we are suggesting is not subject to such limitations, since the EOM

step can easily be substituted for alternative sampling methods.

Even though the suggested framework could be improved in several minor ways (such as attempting more margin and penalty weight values), we believe that the most important task ahead consists of generating a uniform, infinitely expandable set of conformers, in the manner of the convex hull sampling method. This will allow for a better control on the convergence rate of the model and to a less biased depiction of reality.

# 8 Acknowledgements

I would like to thank Drs. Mauro Maggioni, Bruce Donald and Terry Oas for their advice, help and attention in my time at Duke. In addition, I would like to thank (soon to be) Dr. Yang Qi for the invaluable mentorship he has provided me for the last two years, as well as all members of the Donald and Oas labs.

# References


[1] Ivano Bertini, Yogesh Gupta, Claudio Luchinat, Giacomo Parigi, Massimiliano Peana, Luca Sgheri and Jing Yuan. Paramagnetism-Based NMR Restraints Provide Maximum Allowed Probabilities for the Different Conformations of Partially Independent Protein Domains. *JACS*, 2007

[2] Carlo Camilloni and Michele Vendruscolo. Using Pseudocontact Shifts and Residual Dipolar Couplings as Exact NMR Restraints for the Determination of Protein Structural Ensembles. *Biochemistry*, 2015

[3] Eldon Emberly, Ranjan Mukhopadhyay, Ned Wingreen and Chao Tang. Flexibility of -helices: Results of a statistical analysis of database protein structures. *Journal of molecular biology*, 2003

[4] Nicholas Anthis and Marius Clore. The Length of the Calmodulin Linker Determines the Extent of Transient Interdomain Association and Target Affinity. *JACS*, 2013

[5] Yang Qi, Jeff Martin, Adam Barb, Francois Thelot, Anthony Yan, Bruce Donald, Terry Oas. Continuous distributions of interdomain orientations in staphylococcal protein A reveal conformational components of binding thermodynamics, 2016

[6] Michael Marlow, Jakob Dogan, Kendra Frederick, Kathleen Valentine and Joshua Wand. the role of conformational entropy in molecular recognition by calmodulin. *Nature chemical biology*, 2010

[7] Bruce Donald. Algorithms in Structural Molecular Biology. *MIT Press*, 2011

[8] Judit Losonczi, Michael Andrec, Mark Fischer and James Prestegard. Order Matrix Analysis of Residual Dipolar Couplings Using Singular Value Decomposition. *Journal of Magnetic Resonance*, 1999

[9] Y. Alhassid, N. Agmon, R.D. Levine. An upper bound for the entropy and its applications to the maximal entropy problem. *Chemical Physics Letters*, 1978

[10] Miroslav Dudik. Maximum entropy density estimation and modeling geographic distributions of species, Doctoral thesis, Princeton University, 2007

[11] Jun'ichi Kazama and Jun'ichi Tsujii. Evaluation and Extension of Maximum Entropy Models with Inequality Constraints. *Proceedings of the 2003 conference on Empirical methods in natural language processing*, 2003

[12] Anna Yershova, Swati Jain, Steven LaValle and Julie Mitchell. Generating Uniform Incremental Grids on SO(3) Using the Hopf Fibration. *International Journal of Robotics Research*, 2009

[13] Giancarlo Tria, Haydyn Mertens, Michael Kachala and Dmitri Svergun. Advanced ensemble modelling of flexible macromolecules using X-ray solution scattering. *IUCrJ*, 2015

[14] Pau Bernado, Efstratios Mylonas, Maxim Petoukhov, Martin Blackledge and Dmitri Svergun. Structural Characterization of Flexible Proteins Using Small-Angle X-ray Scattering. *JACS*, 2007

[15] James Chou, Shipeng Li, Claude Klee and Ad Bax. Solution structure of Ca2+calmodulin reveals flexible hand-like properties of its domains. *Nature Structural Biology*, 2001

[16] Jed Pitera and John Chodera. On the Use of Experimental Observations to Bias Simulated Ensembles. *Journal of Chemical Theory and Computation*, 2012

[17] Benoît Roux and Jonathan Weare. On the statistical equivalence of restrained-ensemble simulations with the maximum entropy method. *Journal of Chemical Physics*, 2013