

# Integrating NOE and RDC using sum-of-squares relaxation for protein structure determination

Y. Khoo\*      A. Singer<sup>†</sup>      D. Cowburn<sup>‡</sup>

February 28, 2017

## Abstract

We revisit the problem of protein structure determination from geometrical restraints from NMR, using convex optimization. It is well-known that the NP-hard distance geometry problem of determining atomic positions from pairwise distance restraints can be relaxed into a convex semidefinite program (SDP). However, often the NOE distance restraints are too imprecise and sparse for accurate structure determination. Residual dipolar coupling (RDC) measurements provide additional geometric information on the angles between atom-pair directions and axes of the principal-axis-frame. The optimization problem involving RDC is highly non-convex and requires a good initialization even within the simulated annealing framework. In this paper, we model the protein backbone as an articulated structure composed of rigid units. Determining the rotation of each rigid unit gives the full protein structure. We propose solving the non-convex optimization problems using the *sum-of-squares* (SOS) hierarchy, a hierarchy of convex relaxations with increasing complexity and approximation power. Unlike classical global optimization approaches, SOS optimization returns a certificate of optimality if the global optimum is found. Based on the SOS method, we proposed two algorithms - RDC-SOS and RDC-NOE-SOS, that have polynomial time complexity in the number of amino-acid residues and run efficiently on a standard desktop. In many instances, the proposed methods exactly recover the solution to the original non-convex optimization problem. To the best of our knowledge this is the first time SOS relaxation is introduced to solve non-convex optimization problems in structural biology.

We further introduce a statistical tool, the Cramér-Rao bound (CRB), to provide an information theoretic bound on the highest resolution one can hope to achieve when determining protein structure from noisy measurements using any methodology. Our simulation results show that when the RDC measurements are corrupted by Gaussian noise of realistic variance, both SOS based algorithms attain the CRB.

We successfully apply our method in a divide-and-conquer fashion to determine the structure of ubiquitin from experimental NOE and RDC measurements ob-

---

\*Department of Physics, Princeton University, Princeton, NJ 08540, USA (ykhoo@princeton.edu).

<sup>†</sup>Department of Mathematics and PACM, Princeton University, Princeton, NJ 08544, USA (amits@math.princeton.edu).

<sup>‡</sup>Department of Biochemistry, Albert Einstein College of Medicine, Bronx, NY 10461, USA (david.cowburn@einstein.yu.edu).

tained in two alignment media, achieving more accurate and faster reconstructions compared to the current state of the art.

## 1 Introduction

The problem of positioning a set of points from geometrical constraints between them arises naturally when calculating the protein structure from Nuclear Magnetic Resonance (NMR) spectroscopy data. The best established structural calculation methods are based on the through-space dipole interaction of the Nuclear Overhauser Effect (NOE) [38, 80]. The NOE gives rise to qualitative distance constraints of the following form

$$d_{nm}^{\text{lower}} \leq \|\mathbf{x}_n - \mathbf{x}_m\|_2 \leq d_{nm}^{\text{upper}} \quad (1)$$

where  $\mathbf{x}_n, \mathbf{x}_m$  are the coordinate positions of atoms  $n$  and  $m$ , and  $d_{nm}^{\text{lower}}, d_{nm}^{\text{upper}}$  are lower and upper bounds, respectively, for the Euclidean distance between these atoms. Since the NOE interaction between a pair of atoms scales as  $r^{-6}$ , constraints for pairs of atoms that are more than 6 Å apart are too small and imprecise for use. For large molecules, the extraction of NOE restraints through resonance assignment is difficult and often leads to missing, ambiguous, or incorrect NOE distance measurements. Hence the inverse problem of positioning from distance constraints alone, also known as the distance geometry problem, can be challenging and even ill-posed [81]. While multiple ingenious and interesting methods are used to address these issues [48, 59], obtaining a fully automated structural determination software based on NOE alone remains challenging. As noted in [48], the process of filtering out the wrong NOE restraints may require manual intervention.

Residual dipolar coupling (RDC) measurements provide additional geometrical information involving pairs of atoms [71, 68]. RDC can be measured when the molecule ensemble in solution exhibits partial alignment with the magnetic field in an NMR experiment. The RDC measurements have relatively high precision due to the slower  $1/r^3$  decay of interaction, and provide alignment information involving pairs of atoms and the magnetic field. Under some technical assumptions, the RDC measurement  $r_{nm}$  for atoms  $n$  and  $m$  is related to their positions in the following way:

$$r_{nm} = \frac{(\mathbf{x}_n - \mathbf{x}_m)^T \mathbf{S} (\mathbf{x}_n - \mathbf{x}_m)}{d_{nm}^2}, \quad (2)$$

where  $d_{nm} = \|\mathbf{x}_n - \mathbf{x}_m\|_2$  is the distance between atoms  $n$  and  $m$ , and  $\mathbf{S}$  is a  $3 \times 3$  symmetric matrix with vanishing trace, known as the Saupe alignment tensor [57]. Roughly speaking, the eigenvectors of the Saupe tensor encode how the molecule aligns with respect to the magnetic field. Performing NMR experiments at different alignment conditions may lead to different Saupe tensors, and consequently different RDC measurements. While in principle both the Saupe tensor and the molecular structure are unknown, in this paper we assume that  $S$  can be estimated a-priori [46, 84] and our goal is to determine the atom positions given  $S$ . We primarily focus on protein backbone structure determination from RDC data. For a detailed exposition of RDC and the Saupe tensor, we refer readers to the appendix and to [44, 5, 70].

## 1.1 Existing Approaches

Most approaches to the structural determination problem apply a global optimization technique [49, 45, 60, 33, 32] to obtain the global minima of a non-convex “energy” function. The energy function includes pseudo-potential terms that restrain the pairwise interatomic distances (NOE), dihedral angles ( $J$ -coupling), packing (van der Waals interactions), and orientation with respect to a global magnetic field (RDC).

The mainstream approach to minimize the energy function is based on simulated annealing [34, 26, 15, 60]. In simulated annealing, the “tunneling” mechanism pushes the solution out of a local minimum with a certain probability and the procedure can be run for many iterations in order to increase the chances of escaping local minima. In principle, this gives simulated annealing the versatility to deal with arbitrary non-convex energy functions, in particular, one can consider the following non-convex RDC potential term:

$$\left( r_{nm} - \frac{(\mathbf{x}_n - \mathbf{x}_m)^T \mathbf{S} (\mathbf{x}_n - \mathbf{x}_m)}{d_{nm}^2} \right)^2 \quad (3)$$

This RDC potential term yields, however, a rugged energy landscape with sharp local minima that hinders the success of finding the correct conformation in the absence of a good initial structure [14, 3]. For example, [51] reports that direct minimization of the RDC potential using simulated annealing can yield structures that are as much as 20 Å away from the correct structure. A popular way to initialize simulated annealing for protein structure determination from RDC is by the molecular fragment replacement (MFR) approach [35]. MFR finds homologous short fragments of the protein in the Protein Data Bank with the aid of RDC and chemical shifts. The fragments are then merged together to form an initial structure to be locally refined by simulated annealing. However, using existing structures as initialization leads to model bias. Moreover, there is still no guarantee that the initialization is good enough to avoid trapping at a local minimum.

Besides stochastic optimization, a number of deterministic approaches based on branch and prune [82, 13] and dynamic programming [51] have been proposed more recently to find the globally optimal backbone structure. In particular, RDC-ANALYTIC [74, 75, 82] exploits that in the presence of two RDC measurements per amino-acid, the torsion angles that determine the orientation of an amino-acid have 16 possible value sets, and a solution tree with a total of  $16^M$  possible structures can be built sequentially for a protein with  $M$  amino-acids. The main advantage of branch and prune type methods is their ability to deal with sparse RDC datasets when used with an efficient adjunct pruning device such as the Ramachandran plot [56] and NOE. In addition, it can produce multiple low-energy solutions reflecting intrinsic flexibility [72]. Another approach with a similar flavor to the tree-searching based methods, REDCRAFT [10], performs Monte-Carlo sampling of the torsion angles of a protein based on the Ramachandran distribution. RDC measurements are then used to select the possible torsion angles. In general, the methods based on building the conformation space and pruning the unwanted conformations can lead to slow running times. Both REDCRAFT and RDC-ANALYTIC need an hour or two to solve for the structure of a typical size protein. Another approach with a different flavor is the dynamic programming approach in [51]. By casting the protein structuring problem from RDC as

a shortest path problem, a solution can be obtained optimally and efficiently. However, it cannot readily incorporate additional information such as distance restraints to improve the solution quality.

A separate line of research is based on convex relaxation, in which the non-convex domain of an optimization problem is replaced by a convex domain. When the global optimum of the convex surrogate problem lies in the original domain, we can be sure that the original problem is solved. Otherwise, a rounding scheme can be used to project the solution from the convex set back to the original domain. For the distance geometry problem, semidefinite programming (SDP) relaxations [64, 4, 19] have been proposed. Under certain conditions on the distance measurements, it is shown that the solution to the NP-hard [58] distance geometry problem can be computed in polynomial time [64]. Since the introduction of the SDP relaxation, numerous efforts have been made for its computational speedup using additional relaxation [76], divide-and-conquer procedures [41, 17], and facial reduction [1]. While these methods are highly accurate in the presence of abundant distance restraints and do not suffer from local minima issues, their performance is unsatisfactory when lacking sufficient NOE measurements (especially for large proteins due to spin diffusion [55]). In such cases, it is crucial to refine the solution obtained by SDP relaxation by minimizing the original non-convex energy using another method such as simulated annealing.

## 1.2 Overall approach

We limit our attention to the calculation of protein backbone structure, leveraging the RDC and NOE measurements for the backbone. Unlike previous convex relaxation approaches that focused solely on distance constraints, we propose convex relaxations for backbone structure determination that simultaneously incorporate both NOE and RDC measurements. An additional advantage of this combination method is that it can provide accurate solutions even when using RDC alone.

We believe our proposed algorithm provides a solution to the Open Problem posed in [20, Chapter 36]: *“Use SDP and the concept of distance geometry with angle restraints to model RDC-based structure determination.”* In some sense, the structural calculation problem from RDC measurements of the form (2) can be regarded as the distance geometry problem in an inner product space (corresponding to the Saupe tensor) different from the standard Euclidean space. Since the convex relaxations in [64, 4] proposed for the distance geometry problem only involve the Gram matrix (inner product matrix) [27] of the atom coordinates in the Euclidean space, these methods do not readily generalize to deal with RDC measurements that come from different inner product spaces. Such complication gives rise to the open problem in [20].

We deal with it by introducing every monomial of the atom coordinates to our optimization problems instead of just using the elements of the Gram matrix. Furthermore, in our approach we view the protein backbone as an articulated structure composed of rigid planes and bodies that are chained together via hinges [26], rather than just a loose set of points. The coordinates of the atoms can thus be determined by the orthogonal transformation of these rigid units. This has the advantage of lowering the number of variables, and facilitating the incorporation of chirality constraint for the rigid units via re-parameterizing the problem in unit quaternion. We remark that (unlike

existing optimization approaches that also model the protein as an articulated structure using torsion angle parameterization [26], with RDC measurements alone) the cost and the constraints in our formulation are separable in the optimization variables (the unit quaternions), i.e. each term in the cost and constraints only depends on a single unit quaternion. Such structure of cost and constraints yields a less nonlinear optimization problem, which is essential in obtaining a convex relaxation to it.

Since the cost function (9) is a 4-th order polynomial in the atomic coordinates (and in the rotations of the articulated structure as well), parameterizing the cost (9) in unit quaternions gives rise to an 8-th order polynomial minimization problem. We introduce the SOS hierarchy [40, 6, 39, 54] to convexly relax the problem. The global optimum of the SOS relaxation is then used as our solution. One of the main benefit of such optimization technique is that when a solution is returned, whether or not the solution is the global optimum can be checked easily by examining the rank of certain matrices. Moreover, if we increase the complexity of the convex relaxation in the SOS hierarchy, the solution provably converges to the optimum of the original problem [39]. Since its introduction, SOS relaxation has been applied successfully to solve many instances of non-convex optimization problems in sensor network localization [52], super-resolution [18], tensor decomposition [61], and control theory [28]. SOS optimization works by reformulating the polynomial optimization problem in terms of nonnegative polynomials. Since it is NP-hard to check whether a polynomial is nonnegative the sum-of-squares polynomials [40] are search in the SOS relaxations. Empirically, when there is sufficient number of measurements, our proposed methods recover the optimal solution exactly and efficiently when there is no noise in the RDC, and stably when noise is added to the RDC, using low complexity SOS relaxation.

The resulting algorithm RDC-SOS has running time of about an order of magnitude faster than existing toolboxes that use RDC for *de novo* calculation of the protein backbone [10, 82]. This is rather remarkable as the computational problem of determining the orientations has its domain on the product manifold of special orthogonal matrices, with a search space that is non-convex and exponential in size. Such fast and accurate determination of the initial structure could have potential applications in quick validation of backbone and NOE resonance assignment [25, 83] or refining Saupe tensor estimate through alternating minimization.

To include both RDC and NOE restraints to improve the solution quality, we propose a different method - RDC-NOE-SOS, at the expense of increasing the running time. The nature of the NOE restraint is very different from an RDC constrain in that it indicates distances, translatable to rotation about hinges of the rigid units, whereas the RDC refers to the orientation of the individual rigid unit in the Saupe tensor frames. Therefore, when dealing with NOE restraints, only the relative rotations between the rigid units, i.e. the Gram matrix of the rotations, are optimized over. To incorporate these two types of measurement, our proposed method uses the information from RDC to regulate the spectrum of the rotation Gram matrix through a linear matrix inequality [9]. The uses of rotation Gram matrix in RDC-NOE-SOS leads to a longer running time than RDC-SOS.

It is in general difficult to determine the backbone structure of an entire protein at once using an RDC-based algorithm, since along the chain of rigid units there are typically some sites having only a few or no RDC being measured. Therefore we divide up the protein backbone and run RDC-SOS or RDC-NOE-SOS on each of the fragments.

As a separate contribution, we propose an additional SDP that jointly solves for the relative translations of all fragments using inter-fragment NOE in order to form the global structure of the protein. In [82], a grid search is employed to find the translation that satisfies the NOE restraints between two fragments and the backbone is greedily and sequentially constructed based on the estimated pairwise translations. Our method, on the other hand, pieces all fragments at once rather than sequentially, and may therefore require fewer NOE measurements.

We tested the algorithms in calculating the structure of ubiquitin fragments from experimental RDC and NOE data deposited in the Protein Data Bank (PDB). We successfully computed the backbone structure for short fragments of ubiquitin (each consisting of 12 amino acids on average) up to 0.4 Å resolution, and the full backbone up to 0.86 Å resolution. This is competitive when comparing to the state of the art MFR method that gives structures with 0.56 Å and 0.87 Å RMSD for the fragments and full backbone of ubiquitin respectively. To further assess the quality of our structural calculation procedure, we introduce a classical statistical tool, the Cramér-Rao lower bound, which provides the minimum possible variance of the estimated atomic coordinates for a given noise model on the RDC and NOE. Both methods attain the CRB when aided by NOE restraints.

### 1.3 Broader contexts beyond structural biology

In a broader context, our solution to the protein structuring problem presents a general strategy for determining the pose of an articulated structure, a common problem that arises in robotics and computer vision [21, 2]. The way we model the articulated structure from rotation matrices results in a cost function and constraints that are separable in the rotations, which in turn facilitates subsequent optimization. More generally, the SOS techniques used in our method could be applied to optimization problems involving low degree polynomials in terms of rotation matrices.

### 1.4 Organization

The rest of the paper is organized as follows. In Section 2, we formulate the problem of backbone structure determination from RDC and NOE as a problem of finding the pose of an articulated structure. In Section 3, we describe a semidefinite program (SDP) that parallels the SDP proposed in [64] for protein structuring from NOE in terms of rotation matrices. In Section 4, we apply the SOS relaxation to solve a general optimization problem involving polynomials of rotation matrices, which includes the structure determination problem from RDC. In Section 5, we combine the two convex programs proposed in Section 3 and Section 4 to determine the pose of an articulated structure from both NOE and RDC. In Section 6, we propose an alternate SDP to piece together the fragments, when estimating the full protein structure directly is difficult. In Section 7, we present the numerical results with synthetic data and also for experimental data of ubiquitin (PDB ID: 1D3Z [16]). In the appendix, we give a brief description of the RDC, the SOS relaxation, and we introduce the Cramér-Rao lower bound for the structure determination problem from RDC.

## 1.5 Notation

We use  $\mathbf{I}_d$  to denote the identity matrix of size  $d \times d$ . We frequently use block matrices built from smaller matrices. For a block matrix  $\mathbf{A}$ , we use  $\mathbf{A}_{ij}$  to denote its  $(i, j)$ -th block,  $\mathbf{A}(p, q)$  to denote its  $(p, q)$ -th element, and  $\mathbf{A}_i$  to denote the  $i$ -th column of  $\mathbf{A}$ . The size of the blocks will be made clear from the context. We say that  $\mathbf{A}$  is positive semidefinite (PSD) if  $\mathbf{u}^T \mathbf{A} \mathbf{u} \geq 0$  for all  $\mathbf{u}$ , and use  $\mathbf{A} \succeq \mathbf{B}$  to denote that  $\mathbf{A} - \mathbf{B}$  is PSD [9], that is, . We use  $\mathbb{O}(d)$  to denote the group of  $d \times d$  orthogonal matrices. We use  $\|\mathbf{x}\|_2$  to denote the Euclidean norm of  $\mathbf{x} \in \mathbb{R}^n$  ( $n$  should be clear from the context). We use  $\text{vec}(\mathbf{A})$  to denote the vectorization of a matrix  $\mathbf{A}$ , and  $\text{mat}(\mathbf{a})$  to denote the inverse procedure. In this paper we only use the  $\text{mat}(\cdot)$  operation to form a  $3 \times 3$  matrix from a column vector in  $\mathbb{R}^9$ . We denote the trace of a square matrix  $\mathbf{A}$  by  $\text{Tr}(\mathbf{A})$ . The Kronecker product between matrices  $\mathbf{A}$  and  $\mathbf{B}$  is denoted by  $\mathbf{A} \otimes \mathbf{B}$ . The all-ones vector and all-zeros vector are denoted by  $\mathbf{1}$  and  $\mathbf{0}$  respectively (the dimension should be obvious from the context). The  $i$ -th canonical basis vector is denoted as  $\mathbf{e}_i$ .

## 2 Problem Formulation

In this section, we formulate the protein structuring problem as a non-convex optimization problem in terms of rotation matrices. The protein is composed of small rigid units whose structure is known, and we express the Cartesian coordinates of the atoms as well as RDC and NOE restraints in terms of the unknown rotation matrices associated with the rigid units.

### 2.1 Articulated structure and protein backbone

An articulated structure is a chain of rigid units where one unit is ‘‘chained’’ together with the next unit with non-overlapping joints (Figure 1a). When there is a joint between two consecutive units, the relative translation is fixed but not the relative rotation. If there are two non-overlapping joints between two consecutive units, there is only one undetermined degree of freedom corresponding to a rotation around the axis defined by the two joints. This structure is also referred to as the *body-hinge* framework [79] in rigidity theory. Let an articulated structure be composed of  $K$  points residing in  $M$  rigid units. For such a structure, we define a set of points  $\{J_i\}_{i=1}^M$  as the joints between the units where  $J_i \in \{1, \dots, K\}$ . The  $i$ -th unit is joined to the  $(i - 1)$ -th unit at  $J_i$ . Since the coordinates in each unit are known a-priori up to a rigid transformation, we then use  $\mathbf{x}_k^{(i)}$  to denote the location of point  $k$  in the local coordinate system of the  $i$ -th rigid unit. Notice that due to the rigid motion ambiguity, a Euclidean transform needs to be applied to each of the local coordinates  $\mathbf{x}_k^{(i)}$  for each  $i$  in order to form the global structure.

Let  $\boldsymbol{\zeta}_k^{(i)}$  be the global coordinate of point  $k$  in the  $i$ -th unit. For an articulated structure, it is possible to represent the global coordinates  $\boldsymbol{\zeta}_k^{(i)}$  using the rotations  $\mathbf{R}_i, i = 1, \dots, M$  associated with the  $M$  rigid units. For  $i = 1$ , we let

$$\boldsymbol{\zeta}_k^{(1)} = \mathbf{R}_1(\mathbf{x}_k^{(1)} - \mathbf{x}_{J_1}^{(1)}) + \mathbf{t} \quad (4)$$

which amounts to orienting the first rigid unit with  $\mathbf{R}_1$  and adding a translation so that  $\zeta_{J_1}^{(1)}$  are placed at  $\mathbf{t} \in \mathbb{R}^3$ . The coordinates for the  $i = 2$  rigid unit can be obtained as

$$\zeta_k^{(2)} = \mathbf{R}_2(\mathbf{x}_k^{(2)} - \mathbf{x}_{J_2}^{(2)}) + \zeta_{J_2}^{(1)}. \quad (5)$$

The above operations ensure that the  $i = 2$  rigid unit is jointed to the  $i = 1$  rigid unit at joint  $J_2$ , since  $\zeta_{J_2}^{(2)} = \zeta_{J_2}^{(1)}$ . The same reasoning implies that in general

$$\zeta_k^{(i)} = \mathbf{R}_i(\mathbf{x}_k^{(i)} - \mathbf{x}_{J_i}^{(i)}) + \zeta_{J_i}^{(i-1)}. \quad (6)$$

Applying induction to (6) gives

$$\zeta_k^{(i)} = \mathbf{R}_i(\mathbf{x}_k^{(i)} - \mathbf{x}_{J_i}^{(i)}) + \sum_{s=1}^{i-1} \mathbf{R}_s(\mathbf{x}_{J_{s+1}}^{(s)} - \mathbf{x}_{J_s}^{(s)}) + \mathbf{t}. \quad (7)$$

The coordinate of each atom is thus expressed as a linear combination of the rotations  $\mathbf{R}_i$ 's and a global translation  $\mathbf{t}$ . As mentioned previously, when there are hinges in the articulated structure the rotations have fewer degrees of freedom. To incorporate the hinges, we define another set of joints  $\{H_i\}_{i=1}^M$  where  $\{H_i\}_{i=1}^M \cap \{J_i\}_{i=1}^M = \emptyset$ . Let  $\mathbf{v}_{kl}^{(i)}$  be the unit vector between the pair of points  $(k, l)$  pointing from atom  $l$  to atom  $k$  in the frame of the  $i$ -th rigid unit. To ensure two consecutive rigid bodies stay chained together by a hinge,  $\mathbf{R}_i$ 's should satisfy the hinge constraints

$$\mathbf{R}_i \mathbf{v}_{H_i J_i}^{(i)} = \mathbf{R}_{i-1} \mathbf{v}_{H_i J_i}^{(i-1)}, \quad i = 2, \dots, M. \quad (8)$$

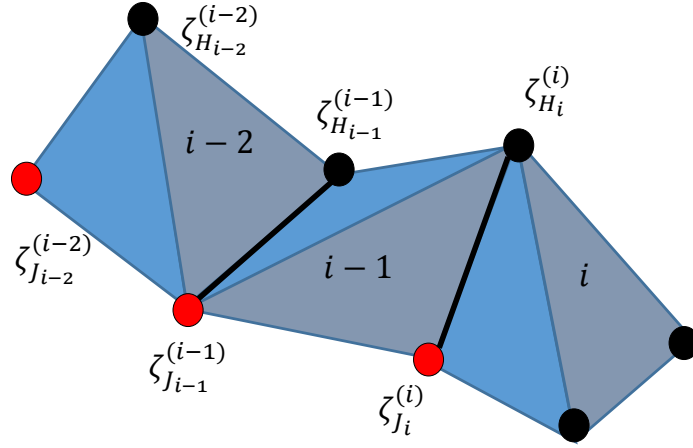
Using the above framework, we can reduce the problem of finding atomic coordinates of a protein backbone into a problem of finding the rotation matrices. This is because the protein backbone can be modeled as an articulated structure composed of peptide planes and CA-bodies. As depicted in Figure 1b, a peptide plane is a 2D rigid plane consisting of atoms from two consecutive amino acids: CA, C, O from one amino acid and H, N, CA from the next amino acid. The CA-body is a 3D rigid body consisting of five atoms CA, N, C, HA and CB all coming from one amino acid. The bonds (N, CA), (C, CA) act like hinges between the rigid units. This use of a rigid model requires no variation of the  $\omega$  backbone torsion angle.

## 2.2 RDC data

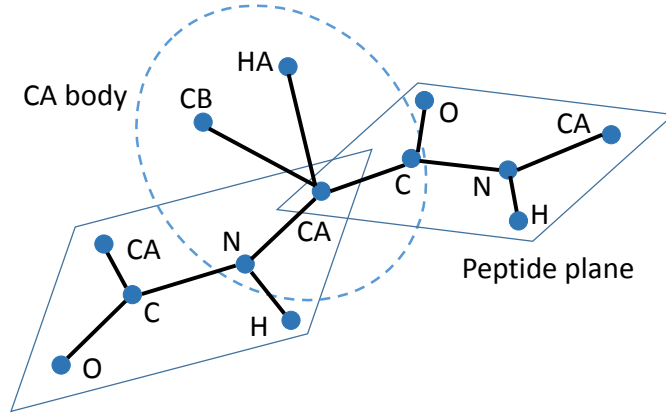
In the setting of calculating protein structure, the RDC measurements described in eq. 2 can be used to constrain the rotation of each rigid unit. Within each rigid unit, in principle all pairs of NMR active nuclei can give rise to an RDC, although in practice only a subset of these pairs have their RDC measured. Suppose  $N$  Saupe tensors for the protein in  $N$  different alignment media have been predetermined. In the  $j$ -th alignment media, the RDC measurements for the  $i$ -th rigid unit between the pair of atoms  $(n, m)$ , denoted  $r_{nm}^{(j)}$ , can be modeled in the following way:

$$r_{nm}^{(j)} = \mathbf{v}_{nm}^{(i)T} \mathbf{R}_i^T \mathbf{S}^{(j)} \mathbf{R}_i \mathbf{v}_{nm}^{(i)}, \quad (n, m) \in E_{\text{RDC}i},$$





(a)



(b)

Figure 1: (a): Example of an articulated structure with joints with indices  $J_i$ 's (Red dots) and  $H_i$ 's. The hinges are represented by black bars in the figure. (b): Protein backbone consists of peptide planes and CA bodies. These rigid units are chained together at the bonds (N, CA) and (C,CA).

$$i = 1, \dots, M, \quad j = 1, \dots, N. \quad (9)$$

The set  $E_{\text{RDC}_i}$  is the set of edges that give rise to RDC in the  $i$ -th rigid unit, and  $\mathcal{S}^{(j)}$  denotes the Saupe tensor in alignment media  $j$ . The orientation of the peptide planes and

CA-bodies can be obtained by solving equation (9) subject to the hinge constraint (8). Due to experimental errors in measuring the RDC, (9) is only satisfied approximately, and orientations can be estimated by minimizing the following cost

$$\sum_{i=1}^M \sum_{j=1}^N \sum_{(n,m) \in E_{\text{RDC}_i}} |\mathbf{v}_{nm}^{(i)T} \mathbf{R}_i^T \mathbf{S}^{(j)} \mathbf{R}_i \mathbf{v}_{nm}^{(i)} - r_{nm}^{(j)}|^2 \quad (10)$$

subject to (8). In the cost function (10) each bond is counted once, including bonds that lie in both the peptide plane and the CA-body (e.g., bond (C – CA)). The difficulty of minimizing the target function (10) lies in the non-convex nature of both the cost and domain. Therefore, RDC measurements are typically used when refining an existing, high quality structure derived from solving the distance geometry problem from NOE or from homology modeling [14].

### 2.3 NOE data

We now rewrite the distance constraints in (1) in terms of the rotations. Instead of working with bounds on distances, we use bounds on squared distances, for reasons that will become apparent later in Section 3. Assuming  $i > j$ , from (7) we have

$$\begin{aligned} \|\boldsymbol{\zeta}_m^{(i)} - \boldsymbol{\zeta}_n^{(j)}\|_2^2 &= \|\mathbf{R}_i(\mathbf{x}_m^{(i)} - \mathbf{x}_{j_i}^{(i)}) - \mathbf{R}_j(\mathbf{x}_n^{(j)} - \mathbf{x}_{j_j}^{(j)}) \\ &\quad + \sum_{s=j+1}^{i-1} \mathbf{R}_s(\mathbf{x}_{j_{s+1}}^{(s)} - \mathbf{x}_{j_s}^{(s)})\|_2^2. \end{aligned} \quad (11)$$

In this way, we write squared distances between two atoms as quadratic functions of  $\mathbf{R}_i$ 's. To satisfy the constraint (1), we can minimize

$$\begin{aligned} \max((d_{mn}^{\text{low}})^2 - \|\boldsymbol{\zeta}_m^{(i)} - \boldsymbol{\zeta}_n^{(j)}\|_2^2, 0)^p + \\ \max(\|\boldsymbol{\zeta}_m^{(i)} - \boldsymbol{\zeta}_n^{(j)}\|_2^2 - (d_{mn}^{\text{up}})^2, 0)^p \end{aligned} \quad (12)$$

where the choice of the parameter  $p$  depends on the specific noise model, and typical choices are  $p = 2$  (least squares) and  $p = 1$  (least unsquared deviations). In practice, the NOE measurements for the backbone amide hydrogens are more reliable and yield relatively hard constraints.

When we have both RDC and NOE data, we simply combine (10) and (12) into

$$\begin{aligned} \lambda \sum_{(m,n) \in E_{\text{NOE}}} [\max((d_{mn}^{\text{low}})^2 - \|\boldsymbol{\zeta}_m^{(i)} - \boldsymbol{\zeta}_n^{(j)}\|_2^2, 0)^p + \\ \max(\|\boldsymbol{\zeta}_m^{(i)} - \boldsymbol{\zeta}_n^{(j)}\|_2^2 - (d_{mn}^{\text{up}})^2, 0)^p] + \\ \sum_{i=1}^M \sum_{j=1}^N \sum_{(n,m) \in E_{\text{RDC}_i}} |\mathbf{v}_{nm}^{(i)T} \mathbf{R}_i^T \mathbf{S}^{(j)} \mathbf{R}_i \mathbf{v}_{nm}^{(i)} - r_{nm}^{(j)}|^2 \end{aligned} \quad (13)$$

where  $E_{\text{NOE}}$  denotes the set of atom pairs that have NOE measured. The choice of  $\lambda$  is typically around  $10^{-9}$ .

### 3 Convex relaxation with only NOE constraints

In this section, we describe a convex relaxation to solve the non-convex protein structuring problem from NOE. We purposefully choose not to present the convex relaxation for RDC first because the concepts involved there are more complicated, and presenting the simpler case with only NOE data can help readers to develop intuitions.

In order to deal with NOE restraints, we first write the problem in terms of the Gram matrix of rotations, i.e. the relative rotation  $\mathbf{R}_i^T \mathbf{R}_j$  for every pair of  $(i, j)$ ,  $i, j = 1, \dots, M$ . This is made possible since the Euclidean distances between different atoms only depend on the inner products of the atom coordinates. From (11), we get

$$\|\boldsymbol{\zeta}_m^{(i)} - \boldsymbol{\zeta}_n^{(j)}\|_2^2 = \text{Tr} \left( [\mathbf{R}_i, \mathbf{R}_{i+1}, \dots, \mathbf{R}_j]^T [\mathbf{R}_i, \mathbf{R}_{i+1}, \dots, \mathbf{R}_j] \begin{bmatrix} \mathbf{x}_n^{(j)} - \mathbf{x}_{J_j}^{(j)} \\ \mathbf{x}_{J_{j+1}}^{(j+1)} - \mathbf{x}_{J_j}^{(j+1)} \\ \vdots \\ \mathbf{x}_m^{(i)} - \mathbf{x}_{J_i}^{(i)} \end{bmatrix} \begin{bmatrix} \mathbf{x}_n^{(j)} - \mathbf{x}_{J_j}^{(j)} \\ \mathbf{x}_{J_{j+1}}^{(j+1)} - \mathbf{x}_{J_j}^{(j+1)} \\ \vdots \\ \mathbf{x}_m^{(i)} - \mathbf{x}_{J_i}^{(i)} \end{bmatrix}^T \right). \quad (14)$$

Introducing a new matrix variable

$$\mathbf{G} = [\mathbf{R}_1, \dots, \mathbf{R}_M]^T [\mathbf{R}_1, \dots, \mathbf{R}_M] \in \mathbb{R}^{3M \times 3M}, \quad (15)$$

the cost (12) can be written as

$$f_{\text{NOE}}(\mathbf{G}) = \sum_{(m,n) \in E_{\text{NOE}}} \max((d_{mn}^{\text{low}})^2 - \text{Tr}(\mathbf{A}_{mn} \mathbf{G}), 0)^p + \max(\text{Tr}(\mathbf{A}_{mn} \mathbf{G}) - (d_{mn}^{\text{up}})^2, 0)^p \quad (16)$$

using some coefficient matrices  $\mathbf{A}_{mn}$ 's. In terms of this new variable  $\mathbf{G}$ , we formulate the minimization problem

$$\begin{aligned} & \min_{\mathbf{G}} f_{\text{NOE}}(\mathbf{G}) & (17) \\ \text{s.t. } & \mathbf{G}_{ii} = \mathbf{I}_3, \\ & \mathbf{G} \succeq 0, \\ & \text{rank}(\mathbf{G}) = 3, \\ & \mathbf{v}_{J_i H_i}^{(i-1)} = \mathbf{G}_{(i-1)i} \mathbf{v}_{J_i H_i}^{(i)}, \quad i \in [2, M], \end{aligned}$$

where  $\mathbf{G}_{ii}$  denotes the  $3 \times 3$  blocks on the diagonal of the matrix  $\mathbf{G}$ . The first three constraints are equivalent to (15), in particular, the PSD-ness and rank-3 constraints of  $\mathbf{G}$  ensures the existence of a factorization in (15) and  $\mathbf{G}_{ii} = \mathbf{I}_3$  is equivalent to the orthogonality of  $\mathbf{R}_i$ 's. The last constraint comes from (8), by changing  $\mathbf{R}_{i-1}^T \mathbf{R}_i$  to  $\mathbf{G}_{(i-1)i}$ . Problem (17) is (almost) equivalent to the problem of finding the chain of rotations from NOE data, except it does not consider the chirality constraint  $\det(\mathbf{R}_i) > 0$ .

Observe that the cost in (17) is convex in the variable  $\mathbf{G}$ , whereas the domain is non-convex due to the rank-3 constraint. We therefore drop the rank-3 constraint in

order to derive a convex relaxation that is similar to the Max-Cut [22] and rotation synchronization [63, 17] SDP relaxations:

$$\begin{aligned}
& \min_{\mathbf{G}} f_{\text{NOE}}(\mathbf{G}) & (18) \\
& \text{s.t. } \mathbf{G}_{ii} = \mathbf{I}_3, \\
& \mathbf{G} \succeq 0, \\
& \mathbf{v}_{J_i H_i}^{(i-1)} = \mathbf{G}_{(i-1)i} \mathbf{v}_{J_i H_i}^{(i)}, \quad i \in [2, M].
\end{aligned}$$

This SDP can be easily solved using, e.g., one of the solvers implemented in CVX [24], a library of conic-programs solver available in Matlab and Python. There is a natural interpretation of the rank relaxed problem (18) that is similar to the SDP proposed in [64], in which the orthogonal transformation associated with each rigid unit is in a high dimensional ambient space  $\mathbb{R}^{3M}$  (instead of  $\mathbb{R}^3$ ). To see this, since  $\mathbf{G} \succeq 0$  in (18), it admits a Cholesky factorization

$$\mathbf{G} = [\mathbf{P}_1, \dots, \mathbf{P}_M]^T [\mathbf{P}_1, \dots, \mathbf{P}_M] \quad (19)$$

where  $\mathbf{P}_i \in \mathbb{R}^{3M \times 3}$ . Furthermore, since  $\mathbf{G}_{ii} = \mathbf{I}_3$ ,  $\mathbf{P}_i^T \mathbf{P}_i = \mathbf{I}_3$ . Therefore

$$\boldsymbol{\zeta}_k^{(i)} = \mathbf{P}_i \mathbf{x}_k^{(i)} \quad (20)$$

is a vector in  $\mathbb{R}^{3M}$ , obtained from rigidly transforming  $\mathbf{x}_k^{(i)} \in \mathbb{R}^3$  into  $\boldsymbol{\zeta}_k^{(i)} \in \mathbb{R}^{3M}$ . When applying (6) with  $\mathbf{P}_i$  instead of  $\mathbf{R}_i$ , we have a framework in  $\mathbb{R}^{3M}$  instead of  $\mathbb{R}^3$ , and the NOE constraint (12) is now a distance constraint placed on a body-hinge framework in  $\mathbb{R}^{3M}$ .

While the global optimum of (18) can be efficiently obtained in polynomial time through standard convex optimization methods, its solution will only resemble the solution of (17) closely if there is a sufficient number of NOE constraints restricting the body-hinge framework to have a unique configuration in a low dimensional subspace close to  $\mathbb{R}^3$  [65]. Without sufficient distance measurements, the quality of the embedding can deteriorate quickly. In the next section, we show how one can use RDC measurements to further improve the quality of the solution using RDC.

## 4 Convex relaxation with only RDC constraints

To determine the backbone structure from RDC, the cost (10) needs to be minimized. Unlike the case of NOE (12), each term in (10) is a 4-th order polynomial function involving a single rotation. In Section 4.1, we first examine the case of optimization over a single rotation. We parameterize the rotations using unit quaternions and apply the SOS hierarchy for polynomial optimization problem over the set of unit quaternions. In Section 4.1.1, we further discuss a few techniques to reduce the size of the proposed convex relaxations. Finally in Section 4.2, based on the SOS formulation for optimization over a single rotation, we propose a convex relaxation to jointly optimize multiple rotation matrices that are coupled through the hinge constraints (8). This leads to the algorithm - RDC-SOS, for protein structuring from RDC data.

## 4.1 Optimization over a single rotation

In order to handle RDC, as mentioned earlier we need to be able to optimize a fourth order polynomial in terms of rotations. We first derive a convex relaxation method for solving the non-convex optimization problem of the form

$$\min_{\mathbf{R} \in \text{SO}(3)} f(\mathbf{R}). \quad (21)$$

where  $f$  is a polynomial function. By the Euler-Rodrigues formula, a rotation matrix can be derived from a unit quaternion  $\mathbf{q} \in \mathbb{R}^4$  via

$$\mathbf{R} = \phi(\mathbf{q}\mathbf{q}^T) := \begin{bmatrix} 1-2\mathbf{q}(3)^2-2\mathbf{q}(4)^2 & 2(\mathbf{q}(2)\mathbf{q}(3)-\mathbf{q}(4)\mathbf{q}(1)) & 2(\mathbf{q}(2)\mathbf{q}(4)+\mathbf{q}(3)\mathbf{q}(1)) \\ 2(\mathbf{q}(2)\mathbf{q}(3)+\mathbf{q}(4)\mathbf{q}(1)) & 1-2\mathbf{q}(2)^2-2\mathbf{q}(4)^2 & 2(\mathbf{q}(3)\mathbf{q}(4)-\mathbf{q}(2)\mathbf{q}(1)) \\ 2(\mathbf{q}(2)\mathbf{q}(4)-\mathbf{q}(3)\mathbf{q}(1)) & 2(\mathbf{q}(3)\mathbf{q}(4)+\mathbf{q}(2)\mathbf{q}(1)) & 1-2\mathbf{q}(2)^2-2\mathbf{q}(3)^2 \end{bmatrix} \quad (22)$$

where  $\mathbf{q}^T \mathbf{q} = 1$ . Therefore, we can equivalently consider solving an optimization problem of the form:

$$\min_{\mathbf{q}: \mathbf{q}^T \mathbf{q} = 1} f(\phi(\mathbf{q}\mathbf{q}^T)). \quad (23)$$

The choice of such parametrization using unit quaternion is motivated by the following fact. Let us consider the easier problem of optimizing a linear function over the set of rotation matrices. For any matrix  $\mathbf{C} \in \mathbb{R}^{3 \times 3}$ ,

$$\begin{aligned} \min_{\mathbf{R} \in \text{SO}(3)} \text{Tr}(\mathbf{C}\mathbf{R}) &= \min_{\mathbf{q} \in \mathbb{R}^4: \mathbf{q}^T \mathbf{q} = 1} \text{Tr}(\mathbf{C}\phi(\mathbf{q}\mathbf{q}^T)) \\ &= \min_{\mathbf{Q} \in \mathbb{R}^4: \text{Tr}(\mathbf{Q}) = 1, \mathbf{Q} \succeq 0} \text{Tr}(\mathbf{C}\phi(\mathbf{Q})). \end{aligned} \quad (24)$$

Denoting the convex hull of a set  $\mathcal{S}$  as  $\text{conv}(\mathcal{S})$ , the last equality comes from the fact that

$$\begin{aligned} \text{conv}(\{\mathbf{q}\mathbf{q}^T \mid \mathbf{q}^T \mathbf{q} = 1, \mathbf{q} \in \mathbb{R}^4\}) \\ = \{\mathbf{Q} \in \mathbb{R}^{4 \times 4} \mid \text{Tr}(\mathbf{Q}) = 1, \mathbf{Q} \succeq 0\}. \end{aligned} \quad (25)$$

By relaxing the domain of the set of matrices  $\mathbf{q}\mathbf{q}^T$  to its convex hull, optimizing a linear function over the special orthogonal group can be done exactly. Motivated by this observation, we may hope a suitable convex relaxation on the set of monomials of the unit quaternion to yield a tight convex relaxation to problem (21) where  $f$  is a higher-degree polynomial of the rotation matrices. We therefore consider solving (23) in place of (21).

In order to obtain a convex relaxation to (23), a general strategy-the SOS relaxation, can be used to derive hierarchies of convex relaxations with increasing complexity to solve problem (21). We give an intuitive presentation of the procedure here and refer interested reader to the appendix and excellent texts such as [40, 6] for a more in depth exposition. In the rest of the paper, for any variable  $\mathbf{x} \in \mathbb{R}^n$  and  $\boldsymbol{\alpha} \in \mathbb{N}^n$ ,  $\mathbf{x}^\alpha$  is defined as

$$\mathbf{x}^\alpha := \mathbf{x}(1)^{\alpha(1)} \mathbf{x}(2)^{\alpha(2)} \dots \mathbf{x}(n)^{\alpha(n)} \quad (26)$$

The notations  $[\mathbf{x}]_{\leq d}$  and  $[\mathbf{x}]_d$  are used to denote vectors containing the monomials of  $\mathbf{x}$  up to and with degree  $d$  respectively. We remind the reader that for a  $n$ -dimensional variable, there are  $\binom{n+d}{d}$  distinct monomials up to degree  $d$ , and  $\binom{n+d-1}{d}$  degree  $d$  monomials. We often use vector  $\mathbf{x}$  with the size of  $[\mathbf{q}]_{\leq d}$  ( $[\mathbf{q}]_d$ ) or matrix  $\mathbf{X}$  with the size of  $[\mathbf{q}]_{\leq d}[\mathbf{q}]_{\leq d}^T$  ( $[\mathbf{q}]_d[\mathbf{q}]_d^T$ ). In this case, we use  $\mathbf{x}_\alpha$  or  $\mathbf{X}_{\alpha\beta}$  to denote the entries associated with  $\mathbf{q}^\alpha$  or  $\mathbf{q}^\alpha\mathbf{q}^\beta$  respectively. For a vector  $\alpha \in \mathbb{N}^n$  of natural numbers, we define  $|\alpha| := \sum_{i=1}^n \alpha(i)$ . Finally, we say a polynomial  $g(\mathbf{x})$  is t-SOS if  $g(\mathbf{x}) = \sum_i h_i(\mathbf{x})^2$  where  $h_i(\mathbf{x})$ 's are some polynomials with the highest degree being  $t$ .

Now, introducing matrix variables

$$\mathcal{M}_{\leq 2d} := [\mathbf{q}]_{\leq d}[\mathbf{q}]_{\leq d}^T, \quad (27)$$

and

$$\mathcal{M}_{2d} := [\mathbf{q}]_d[\mathbf{q}]_d^T, \quad (28)$$

and assuming  $f$  is a polynomial of rotation  $\mathbf{R}$  with degree at most  $t$ , problem (23) can be written equivalently as

$$\min_{\mathcal{M}_{\leq 2t}, [\mathbf{q}]_{\leq t} | \alpha|_{\leq 2t}} \sum_{\alpha, \text{even}} f_{\alpha} \mathbf{q}^\alpha \quad (29)$$

$$\text{s.t. } \begin{aligned} \mathcal{M}_{\leq 2t} &= [\mathbf{q}]_{\leq t}[\mathbf{q}]_{\leq t}^T, \\ \mathcal{M}_{\leq 2t} \alpha\beta (\mathbf{q}^T \mathbf{q} - 1) &= 0, \text{ if } |\alpha + \beta| \leq 2t - 2. \end{aligned} \quad (30)$$

Here

$$f(\phi(\mathbf{q}\mathbf{q}^T)) := \sum_{\alpha} f_{\alpha, \text{even}} \mathbf{q}^\alpha \quad (31)$$

where the polynomial only involves even degree monomials. Notice that the coefficients  $f_{\alpha, \text{even}}$  that appears in (29) corresponds to  $f(\phi(\mathbf{q}\mathbf{q}^T))$  in (23) rather than  $f(\mathbf{R})$  in (21). We have inserted a set of equality constraints (30) that seems redundant, as they are all implied by  $\mathbf{q}^T \mathbf{q} - 1 = 0$ . However, these are crucial if we want to obtain a convex relaxation of the non-convex problem (29). Observe that these equalities are linear in the variable  $\mathcal{M}_{\leq 2t}$ . At this point, the obstacle of having a convex problem is due to the nonlinear equality constraint

$$\mathcal{M}_{\leq 2t} = [\mathbf{q}]_{\leq t}[\mathbf{q}]_{\leq t}^T, \quad (32)$$

which is equivalent to

$$\mathcal{M}_{\leq 2t} \succeq 0, \mathcal{M}_{\leq 2t} \alpha\beta = \mathbf{y}_{\alpha+\beta}, \mathbf{y}_{[0,0,0,0]} = 1, \quad (33)$$

$$\sum_{i=1}^4 \mathbf{y}_{\alpha+2\mathbf{e}_i} - \mathbf{y}_\alpha = 0 \text{ if } |\alpha| \leq 2t - 2, \text{rank}(\mathcal{M}_{\leq 2t}) = 1, \quad (34)$$

for some  $\mathbf{y} \in \mathbb{R}^{\binom{n+2t}{2t}}$ . If the non-convex rank constraint is removed, the convex problem

$$\min_{\mathcal{M}_{\leq 2t}, \mathbf{y}} \sum_{|\alpha| \leq 2t} f_{\alpha, \text{even}} \mathbf{y}_{\alpha} \quad (35)$$

$$\text{s.t. } \mathcal{M}_{\leq 2t} \succeq 0, \quad (36)$$

$$\mathbf{y}_{[0,0,0,0]} = 1, \quad (37)$$

$$\mathcal{M}_{\leq 2t} \boldsymbol{\alpha} \boldsymbol{\beta} = \mathbf{y}_{\boldsymbol{\alpha} + \boldsymbol{\beta}}, \quad (38)$$

$$\sum_{i=1}^4 \mathbf{y}_{\boldsymbol{\alpha} + 2\mathbf{e}_i} - \mathbf{y}_{\boldsymbol{\alpha}} = 0, \text{ for all } \boldsymbol{\alpha} \in \mathbb{N}^4, |\boldsymbol{\alpha}| \leq 2t - 2. \quad (39)$$

is obtained. Here  $\mathbf{e}_i$ 's are canonical basis vectors in  $\mathbb{R}^4$ . The purpose of having the redundant constraints (30) is now clear. When the rank-1 constraint of  $\mathcal{M}_{\leq 2t}$  is removed, it is easy to verify that the entries of  $\mathcal{M}_{\leq 2t}$  need not be high order monomials of  $\mathbf{q}$ . More precisely,  $\mathbf{y}_{\boldsymbol{\alpha} + \boldsymbol{\beta}} \neq \mathbf{y}_{\boldsymbol{\alpha}} \mathbf{y}_{\boldsymbol{\beta}}$  generally, which means  $\sum_{i=1}^4 \mathbf{y}_{2\mathbf{e}_i} - 1 = 0$  does not imply  $\sum_{i=1}^4 \mathbf{y}_{\boldsymbol{\alpha} + 2\mathbf{e}_i} - \mathbf{y}_{\boldsymbol{\alpha}} = 0$ . Although (35) is a surrogate convex problem of (29), in [39] it is shown that as  $t$  tends to infinity, the optimum of problem (35) converges to problem (29). We observe in our numerical study that in many instances, problem (35) converges to (29) already for small and finite values  $t$ , as in many other applications of SOS relaxation [53].

#### 4.1.1 Variable size reduction

In problem (35), the matrix  $\mathcal{M}_{\leq 2t}$  is of size  $\binom{4+t}{t} \times \binom{4+t}{t}$ . Here we exploit the special structure of the problem to reduce the size of the variable to  $\binom{3+t}{t} \times \binom{3+t}{t}$ , which is the size of the highest degree block, i.e.  $\mathcal{M}_{2t}$  in the matrix  $\mathcal{M}_{\leq 2t}$ . For example, for the typical value  $t = 4$  the matrix size is reduced from  $70 \times 70$  to  $35 \times 35$ .

The first observation is that since the set of unit quaternions forms a double cover of  $\mathbb{S}\mathbb{O}(3)$ , i.e. both  $\mathbf{q}$  and  $-\mathbf{q}$  define the same rotation matrix, it is not necessary to consider odd degree monomials in  $\mathcal{M}_{\leq 2t}$ . More precisely, we arrange  $\mathcal{M}_{\leq 2t}$  such that

$$\mathcal{M}_{\leq 2t} = \begin{bmatrix} \mathcal{M}_{\leq 2t, \text{even}} & \mathcal{M}_{\leq 2t, \text{odd}} \\ \mathcal{M}_{\leq 2t, \text{odd}}^T & \mathcal{M}_{\leq 2(t-1), \text{even}} \end{bmatrix} \quad (40)$$

where

$$\mathcal{M}_{\leq 2t, \text{even}} = \begin{bmatrix} [\mathbf{q}]_0 \\ [\mathbf{q}]_2 \\ \vdots \\ [\mathbf{q}]_t \end{bmatrix} \begin{bmatrix} [\mathbf{q}]_0^T & [\mathbf{q}]_2^T & \cdots & [\mathbf{q}]_t^T \end{bmatrix}, \quad (41)$$

$$\mathcal{M}_{\leq 2t, \text{odd}} = \begin{bmatrix} [\mathbf{q}]_0 \\ [\mathbf{q}]_2 \\ \vdots \\ [\mathbf{q}]_t \end{bmatrix} \begin{bmatrix} [\mathbf{q}]_1^T & [\mathbf{q}]_3^T & \cdots & [\mathbf{q}]_{t-1}^T \end{bmatrix}, \quad (42)$$

and

$$\mathcal{M}_{\leq 2(t-1), \text{even}} = \begin{bmatrix} [\mathbf{q}]_1 \\ [\mathbf{q}]_3 \\ \vdots \\ [\mathbf{q}]_{t-1} \end{bmatrix} \begin{bmatrix} [\mathbf{q}]_1^T & [\mathbf{q}]_3^T & \cdots & [\mathbf{q}]_{t-1}^T \end{bmatrix}. \quad (43)$$

After the convex relaxation in (35), we may assume  $\mathcal{M}_{\leq 2t, \text{odd}} = 0$ . Due to the quadratic dependence of a rotation matrix on the unit quaternion, the cost of the optimization problem (35) only involves  $\mathcal{M}_{\leq 2t, \text{even}}$  and  $\mathcal{M}_{\leq 2(t-1), \text{even}}$ , hence both matrices

$$\begin{bmatrix} \mathcal{M}_{\leq 2t, \text{even}} & \mathcal{M}_{\leq 2t, \text{odd}} \\ \mathcal{M}_{\leq 2t, \text{odd}}^T & \mathcal{M}_{\leq 2(t-1), \text{even}} \end{bmatrix}, \begin{bmatrix} \mathcal{M}_{\leq 2t, \text{even}} & -\mathcal{M}_{\leq 2t, \text{odd}} \\ -\mathcal{M}_{\leq 2t, \text{odd}}^T & \mathcal{M}_{\leq 2(t-1), \text{even}} \end{bmatrix} \quad (44)$$

give the same cost. Therefore without lost of generality, we may assume the solution we seek is the average of the matrices in (44)

$$\mathcal{M}_{\leq 2t} = \begin{bmatrix} \mathcal{M}_{\leq 2t, \text{even}} & \mathbf{0} \\ \mathbf{0} & \mathcal{M}_{\leq 2(t-1), \text{even}} \end{bmatrix} \succeq 0. \quad (45)$$

Now, notice that the entries of  $\mathcal{M}_{\leq 2(t-1), \text{even}}$  are a subset of the entries of  $\mathcal{M}_{\leq 2t, \text{even}}$ , it seems plausible that the constraint  $\mathcal{M}_{\leq 2(t-1), \text{even}} \succeq 0$  in (44) can be dropped altogether and problem (35) can be written solely in terms of  $\mathcal{M}_{\leq 2t, \text{even}}$ . Indeed, the constraint  $\mathcal{M}_{\leq 2(t-1), \text{even}} \succeq 0$  is already implied by  $\mathcal{M}_{\leq 2t, \text{even}} \succeq 0$ , since for odd  $\boldsymbol{\alpha}, \boldsymbol{\beta}$

$$\mathcal{M}_{\leq 2(t-1), \text{even}} \boldsymbol{\alpha} \boldsymbol{\beta} = \sum_{k=1}^4 \mathbf{y}_{\boldsymbol{\alpha} + \mathbf{e}_k + \boldsymbol{\beta} + \mathbf{e}_k} = \sum_{k=1}^4 \mathcal{M}_{\leq 2t, \text{even}}(\boldsymbol{\alpha} + \mathbf{e}_k)(\boldsymbol{\beta} + \mathbf{e}_k) \quad (46)$$

where the first equality follows from (39) (in other words,  $\mathcal{M}_{\leq 2(t-1), \text{even}} = \mathbf{A} \mathcal{M}_{\leq 2t, \text{even}} \mathbf{A}^T$  for some matrix  $\mathbf{A}$ ). Let

$$p(d) := \sum_{i=0, i \text{ even}}^d \binom{3+i}{i}, \quad (47)$$

where we note that the size of  $\mathcal{M}_{\leq 2t, \text{even}}$  is  $p(t)$ .

The second size reduction comes from the equality constraints in line (39). Essentially (39) implies

$$\begin{aligned} \sum_{i=1}^4 \mathcal{M}_{\leq 2t, \text{even}} \boldsymbol{\alpha} + \boldsymbol{\beta} + 2\mathbf{e}_i - \mathcal{M}_{\leq 2t, \text{even}} \boldsymbol{\alpha} + \boldsymbol{\beta} \\ = 0, \text{ if } |\boldsymbol{\alpha} + \boldsymbol{\beta}| \leq 2t - 2, \end{aligned} \quad (48)$$

or equivalently,

$$\begin{aligned} (\mathbf{e}_{[2,0,0,0]}^T \boldsymbol{\alpha} + \mathbf{e}_{[0,2,0,0]}^T \boldsymbol{\alpha} + \mathbf{e}_{[0,0,2,0]}^T \boldsymbol{\alpha} + \mathbf{e}_{[0,0,0,2]}^T \boldsymbol{\alpha} - \mathbf{e}_{[0,0,0,0]}^T \boldsymbol{\alpha}) \cdot \mathcal{M}_{\leq 2t, \text{even}} \\ = 0 \quad \forall \boldsymbol{\alpha} \text{ such that } |\boldsymbol{\alpha}| \leq 2t - 2. \end{aligned} \quad (49)$$

where  $\mathbf{e}_{\boldsymbol{\alpha}}$  for  $\boldsymbol{\alpha} \in \mathbb{N}^4$  are canonical basis vectors with the size of  $[\mathbf{q}]_{\leq t}$ . Notice that there are  $p(t-1)$  equality constraints on  $\mathcal{M}_{\leq 2t} \leq 2t$  in (49). Let  $\mathbf{U}_{\leq 2t} \in \mathbb{R}^{p(t) \times (p(t) - p(t-1))}$  be a matrix whose columns form a basis to  $\text{Range}(\mathcal{M}_{\leq 2t, \text{even}})$ ,

$$\mathcal{M}_{\leq 2t, \text{even}} = \mathbf{U}_{\leq 2t} \tilde{\mathcal{M}}_{\leq 2t} \mathbf{U}_{\leq 2t}^T \quad (50)$$



where  $\tilde{\mathcal{M}}_{\leq 2t}$  has size  $(p(t) - p(t-1)) \times (p(t) - p(t-1)) = \binom{3+t}{t} \times \binom{3+t}{t}$  and is PSD. Hereafter, whenever we optimize over the variable  $\mathcal{M}_{\leq 2t}$ , we can remove all the equality constraints in (39) and simply work with  $\tilde{\mathcal{M}}_{\leq 2t}$ . We note that this procedure of removing the nullspace of a PSD variable is called facial reduction [37] in the convex optimization community. Semidefinite facial reduction is not only important for reducing the size of the variable, but also necessary to ensure the numerical stability of semidefinite programs.

## 4.2 Optimization over multiple rotations - RDC-SOS

When minimizing the RDC cost in (10) over multiple rotations, the hinge constraints (8) have to be included. In terms of unit quaternions, constraints (8) are of the form

$$\mathbf{l}_{1,i}(\phi(\mathbf{q}_i \mathbf{q}_i^T)) = \mathbf{l}_{2,i}(\phi(\mathbf{q}_{i-1} \mathbf{q}_{i-1}^T)), \quad i = 2, \dots, M. \quad (51)$$

where the linear functions  $\mathbf{l}_{1,i}, \mathbf{l}_{2,i} : \mathbb{R}^{3 \times 3} \rightarrow \mathbb{R}^3$ . Again, we have to include other valid equalities

$$\mathbf{l}_{1,i}(\phi(\mathbf{q}_i \mathbf{q}_i^T))^\gamma = \mathbf{l}_{2,i}(\phi(\mathbf{q}_{i-1} \mathbf{q}_{i-1}^T))^\gamma, \quad i = 2, \dots, M, \quad (52)$$

as in Section 4.1 since we are going to apply SOS relaxation to the matrices  $\mathcal{M}_{\leq 2t,i} = [\mathbf{q}]_{\leq t,i} [\mathbf{q}]_{\leq t,i}^T$ . Let

$$\begin{aligned} \mathbf{l}_{1,i}(\phi(\mathbf{q}_i \mathbf{q}_i^T))^\gamma &:= \sum_{\alpha} \mathbf{l}_{1,i,\alpha}^\gamma \mathbf{q}_i^\alpha, \\ \mathbf{l}_{2,i}(\phi(\mathbf{q}_i \mathbf{q}_i^T))^\gamma &:= \sum_{\alpha} \mathbf{l}_{2,i,\alpha}^\gamma \mathbf{q}_i^\alpha, \\ i &= 2, \dots, M, \end{aligned} \quad (53)$$

Also, we define for each term of the cost (10) its polynomial expansion in terms of its corresponding quaternion

$$\begin{aligned} \sum_{\alpha} f_{\text{RDC},\alpha}^{i,j,(n,m)} \mathbf{q}^\alpha &:= (\mathbf{v}_{nm}^{(i)T} \phi(\mathbf{q}_i \mathbf{q}_i^T)^T \mathbf{S}^{(j)} \phi(\mathbf{q}_i \mathbf{q}_i^T) \mathbf{v}_{nm}^{(i)} - r_{nm}^{(j)})^2, \\ i &= 1, \dots, M, \quad j = 1, \dots, N, \quad (n,m) \in E_{\text{RDC}i}. \end{aligned} \quad (54)$$

We arrive at the following convex program

$$\begin{aligned} \min_{\{\tilde{\mathcal{M}}_{\leq 2t,i}, \mathbf{y}_i\}_{i \in [1,M]}} \quad & \sum_{i=1}^M \sum_{j=1}^N \sum_{(n,m) \in E_{\text{RDC}i}} \sum_{\alpha} f_{\text{RDC},\alpha}^{i,j,(n,m)} \mathbf{y}_\alpha \quad (55) \\ \text{s.t.} \quad & \tilde{\mathcal{M}}_{\leq 2t,i} \succeq 0, \quad i \in [1, M], \\ & \mathbf{y}_{i[0,0,0]} = 1, \quad i \in [1, M], \\ & (\mathbf{U}_{\leq 2t} \tilde{\mathcal{M}}_{\leq 2t,i} \mathbf{U}_{\leq 2t}^T)_{\alpha\beta} = \mathbf{y}_{i\alpha+\beta}, \quad i \in [1, M], \\ & \sum_{\alpha} \mathbf{l}_{1,i,\alpha}^\gamma \mathbf{y}_\alpha = \sum_{\alpha} \mathbf{l}_{2,i,\alpha}^\gamma \mathbf{y}_\alpha, \quad \gamma \in \mathbb{N}^3, |\gamma| \leq t, \quad i \in [2, M] \end{aligned} \quad (56)$$

for protein structuring from RDC, which is an SDP with  $M$  PSD variables of size  $\binom{3+t}{3} \times \binom{3+t}{3}$ . Again, we typically use  $t = 4$ , where each PSD variable is of size  $35 \times 35$ .

## 5 Convex relaxation with both NOE and RDC constraints

In the absence of sufficient RDC information, it is important to include NOE restraints along with RDC in order to determine the protein structure. In this section, we describe how to combine RDC-SOS and (18) to incorporate both types of measurements. Previously in (18), we relaxed the non-convex equality constraint

$$\mathbf{G} = [\mathbf{R}_1, \dots, \mathbf{R}_M]^T [\mathbf{R}_1, \dots, \mathbf{R}_M] \quad (57)$$

into  $G \succeq 0$  and forgo using the factors  $[\mathbf{R}_1, \dots, \mathbf{R}_M]$  in the factorization of  $G$ . In this case, only the relaxed version of the relative rotations  $\mathbf{R}_i^T \mathbf{R}_j$ 's are being optimized over, making the formulation in (18) oblivious to the absolute orientation of each individual rigid unit with respect to the Saupe tensor frame. Therefore in order to incorporate RDC-SOS into (18), we need to be able to work with each individual rotation matrices directly in (18). To this end, we use the following different convex relaxation of  $\mathbf{G}$ :

$$\mathbf{G} \succeq [\mathbf{R}_1, \dots, \mathbf{R}_M]^T [\mathbf{R}_1, \dots, \mathbf{R}_M], \quad (58)$$

or equivalently

$$\begin{bmatrix} & & & \mathbf{R}_1^T \\ & \mathbf{G} & & \vdots \\ & & & \mathbf{R}_M^T \\ \mathbf{R}_1 & \dots & \mathbf{R}_M & \mathbf{I}_3 \end{bmatrix} \succeq 0 \quad (59)$$

in order to include the variables  $\mathbf{R}_1, \dots, \mathbf{R}_M$  explicitly. Then, the  $t = 2$  blocks in RDC-SOS can be used to fix the matrices  $\mathbf{R}_i$ 's in (59). The extra information from RDC-SOS on  $[\mathbf{R}_1, \dots, \mathbf{R}_M] \in \mathbb{R}^{3 \times 3M}$ , along with the constraint  $\mathbf{G} \succeq [\mathbf{R}_1, \dots, \mathbf{R}_M]^T [\mathbf{R}_1, \dots, \mathbf{R}_M]$  help to concentrate the spectrum of  $\mathbf{G}$  into three prominent eigenvalues. In this case, instead of solving the distance geometry problem in a high dimensional space as in (18), the matrix  $\mathbf{G}$  in RDC-NOE-SOS has much lower rank. This tightens the convex relaxation in (18) by getting us closer to solving the distance geometry problem in  $\mathbb{R}^3$ . Here is the resulting convex program RDC-NOE-SOS that incorporates both RDC and NOE:

$$\min_{\mathbf{G}, \{\tilde{\mathcal{M}}_{\leq 2t, i}, \mathbf{y}_i, \mathbf{R}_i\}_{i \in [1, M]}} \lambda f_{\text{NOE}}(\mathbf{G}) + \sum_{i=1}^M \sum_{j=1}^N \sum_{(n, m) \in E_{\text{RDC}_i}} \sum_{\alpha} f_{\text{RDC}, \alpha}^{i, j, (n, m)} \mathbf{y}_{\alpha} \quad (60)$$

$$\text{s.t. } \mathbf{G}_{ii} = \mathbf{I}_3, \quad i \in [1, M], \quad (61)$$

$$\mathbf{G} \succeq [\mathbf{R}_1, \dots, \mathbf{R}_M]^T [\mathbf{R}_1, \dots, \mathbf{R}_M],$$

$$\mathbf{R}_i = \phi \left( \begin{bmatrix} \mathbf{y}_{i[2,0,0,0]} & \mathbf{y}_{i[1,1,0,0]} & \mathbf{y}_{i[1,0,1,0]} & \mathbf{y}_{i[1,0,0,1]} \\ \mathbf{y}_{i[1,1,0,0]} & \mathbf{y}_{i[0,2,0,0]} & \mathbf{y}_{i[0,1,1,0]} & \mathbf{y}_{i[0,1,0,1]} \\ \mathbf{y}_{i[1,0,1,0]} & \mathbf{y}_{i[0,1,1,0]} & \mathbf{y}_{i[0,0,2,0]} & \mathbf{y}_{i[0,0,1,1]} \\ \mathbf{y}_{i[1,0,0,1]} & \mathbf{y}_{i[0,1,0,1]} & \mathbf{y}_{i[0,0,1,1]} & \mathbf{y}_{i[0,0,0,2]} \end{bmatrix} \right), \quad i \in [1, M],$$

$$\tilde{\mathcal{M}}_{\leq 2t, i} \succeq 0, \quad i \in [1, M],$$

$$\mathbf{y}_{i[0,0,0,0]} = 1, \quad i \in [1, M],$$

$$(\mathbf{U}_{\leq 2t} \tilde{\mathcal{M}}_{\leq 2t, i} \mathbf{U}_{\leq 2t}^T)_{\alpha\beta} = \mathbf{y}_{i\alpha+\beta}, \quad i \in [1, M],$$

$$\sum_{\alpha} l_{1, i}^{\gamma} \mathbf{y}_{i\alpha} = \sum_{\alpha} l_{2, i}^{\gamma} \mathbf{y}_{i\alpha}, \quad \gamma \in \mathbb{N}^3, \quad |\gamma| \leq t, \quad i \in [2, M]. \quad (62)$$

The soft penalty  $\lambda f_{\text{NOE}}(\mathbf{G})$  is used to enforce the NOE restraints, where the typical choice of  $\lambda$  is  $10^{-9}$  in order to balance the costs associated with RDC and NOE.

Alternatively, since the NOE restraints on the backbone are quite reliable, one can also include them as hard upper and lower bound in RDC-NOE-SOS. We note that for RDC-NOE-SOS, in addition to the  $M$  PSD variables  $\mathcal{M}_{\leq 2t,i}$  of size  $\binom{t+3}{3} \times \binom{t+3}{3}$ , we have another PSD variable of size  $3(M+1) \times 3(M+1)$  for the constraint (61). In order to take into account the RDC cost which is an 8-th order polynomial in terms of the unit quaternions, we have to choose  $t \geq 4$ . In our numerical experiments, we always fix  $t = 4$ , because this choice already gives high quality solutions. We note that when using primal-dual interior point method SDP solvers [69, 67, 50] to solve an SDP with  $n \times n$  PSD matrix, the typical complexity per iteration is  $O(n^3)$  [36]. In RDC-NOE-SDP, there are  $M$  PSD variables of size  $\binom{3+t}{3} \times \binom{3+t}{3}$  and a  $3(M+1) \times 3(M+1)$  PSD variable that hosts  $G$ . This gives  $O(M^3 + M(t^3)^3)$  per iteration complexity.

## 5.1 Rounding

In this section, we describe a *rounding* scheme to extract rotations from the solutions of RDC-SOS and RDC-NOE-SOS. When the matrices  $\mathcal{M}_{\leq 2t,i}$ 's from the solution to RDC-SOS and RDC-NOE-SOS are rank 1, they correspond to the monomials of a quaternion. If not, two things can happen: (1) RDC-SOS and RDC-NOE-SOS only manage to find approximate solutions, (2) RDC-SOS and RDC-NOE-SOS exactly recover multiple global minimizers successfully. Denote the solution to RDC-SOS and RDC-NOE-SOS as  $\mathcal{M}_{\leq 2t,i}^*$ . The *flat extension theorem* [29] provides a sufficient condition to check if we fall in the second case. When applied to RDC-SOS and RDC-NOE-SOS, the flat extension theorem says that if

$$\text{rank}(\mathcal{M}_{\leq 2t,i}^*) = \text{rank}(\mathcal{M}_{\leq 2t-2,i}^*), \quad (63)$$

the global minimizers can be obtained from the Cholesky factorization

$$\mathcal{M}_{\leq 2t,i}^* = \mathbf{V}_{\leq t,i} \mathbf{V}_{\leq t,i}^T \quad (64)$$

of  $\mathcal{M}_{\leq 2t,i}^*$ . The implementation of the solution extraction algorithm from  $\mathbf{V}_{\leq t,i}$  is described in detail in [29]. We remind the reader again that  $\mathcal{M}_{2t,i}^*$  can be reconstructed from the variable  $\mathcal{M}_{\leq 2t,\text{even},i}^*$  in RDC-SOS and RDC-NOE-SOS, as mentioned in section 4.1.1.

If the condition of the flat extension theorem is not satisfied, there is no guarantee that RDC-SOS and RDC-NOE-SOS return the solution to the un-relaxed problem. However, it is possible to extract an approximate solution based on a heuristic presented below. First, let the rank-1 approximation to  $\mathcal{M}_{4,i}^* \in \mathbb{R}^{10 \times 10}$  be

$$\mathcal{M}_{4,i}^* \approx [\mathbf{q}]_{2,i}^* [\mathbf{q}]_{2,i}^{*T}. \quad (65)$$

The vector  $[\mathbf{q}]_2^* \in \mathbb{R}^{10}$  can be seen as an approximation to the degree two monomials of the unit quaternion. After forming a  $4 \times 4$  matrix  $\mathbf{Q}$  such that

$$\mathbf{Q}_{i\alpha\beta}^* = [\mathbf{q}]_{2,i\alpha+\beta}^*, \quad |\alpha|, |\beta| \leq 1, \quad (66)$$

the top eigenvector  $\mathbf{q}_i^*$  of  $\mathbf{Q}_i^*$  is used to find the best rank-1 approximation to  $\mathbf{Q}_i^*$  and its corresponding rotation  $\mathbf{R}_i^* = \phi(\mathbf{q}_i^* \mathbf{q}_i^{*T})$ . We note that there is a sign ambiguity when

computing  $[\mathbf{q}]_{2,i}^*$ , and we choose the sign such that the largest eigenvalue in magnitude of  $\mathbf{Q}_i^*$  is positive (recall that  $\mathbf{Q}_i$  needs to satisfy  $\text{rank}(\mathbf{Q}_i) = 1$ ,  $\text{Tr}(\mathbf{Q}) = 1$ . In particular its only non-zero eigenvalue should be positive).

For the case when RDC-SOS and RDC-NOE-SOS do not give solution that satisfies the condition of the flat extension theorem, the non-convex problem of finding the rotations of the rigid units is not solved exactly. After rounding there is no guarantee that  $\mathbf{R}_i^*$  orient the rigid units optimally such that the combination of the costs (10) and (12) is minimized. In this case, since the pose recovery problem for an articulated structure is an optimization problem on the product of  $\mathbb{S}\mathbb{O}(3)$  manifolds, we use the manifold optimization toolbox Manopt [8] to refine  $\mathbf{R}_i^*$  further in order to obtain a solution with a lower cost. However, since ManOpt only handles unconstrained optimization problems on a Riemannian manifold, we have to use the penalty method to handle the hinge constraint (8) of the type  $h(\mathbf{R}_i) = 0$  by adding a penalty  $(\mu/2)\|h(\mathbf{R}_i)\|_2^2$  with increasing  $\mu$ . We note that without a good initialization, manifold optimization can easily get stuck in a local minima as it is essentially a gradient descent based approach that descends along the geodesics of a manifold.

## 6 Translation Estimation

In the presence of RDC measurements, the backbone conformation of the full protein can be determined from the calculated  $\mathbf{R}_i$ 's, up to a global translation. However, it is usually the case that some of the amino-acid residues contain very few or no RDC's being measured. While RDC-SOS will certainly fail in these situations, using RDC-NOE-SOS is also undesirable. The convex relaxation in (60) is typically not tight if some parts of the protein are solely constrained by the NOE. In this case we need to break up the protein and calculate the conformations for selected fragments of the protein backbone. Then we figure out the relative translation between the fragments in order to combine the backbone segments coherently. Using such divide-and-conquer scheme can also speed up the structural calculation process. In this section, we propose a semidefinite relaxation that *jointly* uses NOE restraints between all fragments to piece them together. Let there be  $F$  fragments. We denote the coordinate of the  $k$ -th atom in the  $i$ -th fragment as  $\mathbf{z}_k^{(i)}$ . We note that in this section, the superscript “ $(i)$ ” is no longer used as the index for rigid peptide plane or CA-body, but as the index of a fragment composed of multiple amino acid residues. The goal is to find  $t_1, \dots, t_F \in \mathbb{R}^3$  such that

$$(d_{kl}^{\text{low}})^2 \leq \|\mathbf{z}_k^{(i)} + \mathbf{t}_i - (\mathbf{z}_l^{(j)} + \mathbf{t}_j)\|_2^2 \leq (d_{kl}^{\text{up}})^2, \quad (67)$$

where  $(k, l) \in E_{\text{NOE}}$ . It should be understood that in this context,  $E_{\text{NOE}}$  only contains the NOE distance restraints between the fragments. The squaring of the constraint is important to obtain a semidefinite relaxation to solve for the pairwise translations. Now let

$$\mathbf{T} = \begin{bmatrix} \mathbf{t}_1^T \\ \vdots \\ \mathbf{t}_F^T \\ \mathbf{I}_3 \end{bmatrix} \begin{bmatrix} \mathbf{t}_1 & \cdots & \mathbf{t}_F & \mathbf{I}_3 \end{bmatrix}$$

$$= \begin{bmatrix} \mathbf{t}_1^T \mathbf{t}_1 & \dots & \mathbf{t}_1^T \mathbf{t}_F & \mathbf{t}_1^T \\ \vdots & \ddots & \vdots & \vdots \\ \mathbf{t}_F^T \mathbf{t}_1 & \dots & \mathbf{t}_F^T \mathbf{t}_F & \mathbf{t}_F^T \\ \mathbf{t}_1 & \dots & \mathbf{t}_F & \mathbf{I}_3 \end{bmatrix} \in \mathbb{R}^{(3+F) \times (3+F)} \quad (68)$$

where  $\mathbf{T}$  is rank 3 and positive semidefinite. Again, by writing (67) in terms of  $\mathbf{T}$  and by relaxing the rank 3 constraint for  $\mathbf{T}$  we can solve for the pairwise translations through the following semidefinite program

$$\begin{aligned} & \min_{\substack{\mathbf{T} \succeq 0, \\ e_{kl}^{\text{up}} \geq 0, e_{kl}^{\text{low}} \geq 0}} \sum_{(k,l) \in E_{\text{NOE}}} e_{kl}^{\text{up}} + e_{kl}^{\text{low}} - \gamma \text{Tr}(\mathbf{T}) & (69) \\ \text{s.t. } & 2(\mathbf{T}(F+1:F+3,i) - \mathbf{T}(F+1:F+3,j))^T (\mathbf{z}_k^{(i)} - \mathbf{z}_l^{(j)}) \\ & + \mathbf{T}(i,i) + \mathbf{T}(j,j) - 2\mathbf{T}(i,j) + \|\mathbf{z}_k^{(i)} - \mathbf{z}_l^{(j)}\|_2^2 \\ & \leq (d_{kl}^{\text{up}})^2 + e_{kl}^{\text{up}}, \quad (k,l) \in E_{\text{up}}, \\ & 2(\mathbf{T}(F+1:F+3,i) - \mathbf{T}(F+1:F+3,j))^T (\mathbf{z}_k^{(i)} - \mathbf{z}_l^{(j)}) \\ & + \mathbf{T}(i,i) + \mathbf{T}(j,j) - 2\mathbf{T}(i,j) + \|\mathbf{z}_k^{(i)} - \mathbf{z}_l^{(j)}\|_2^2 \\ & \geq (d_{kl}^{\text{low}})^2 - e_{kl}^{\text{low}}, \quad (k,l) \in E_{\text{low}}, \\ & \mathbf{T}(F+1:F+3, F+1:F+3) = \mathbf{I}_3 \\ & \mathbf{T}(1:F, 1:F) \mathbf{1} = 0. \end{aligned}$$

The last constraint is there simply to remove the global translation ambiguity. Instead of using (67) as hard constraints to find pairwise translations that satisfy them, we penalize the violation of such bounds through the cost in (69). This is necessary because errors in estimating individual fragment coordinates and also ambiguous NOE assignments may cause violations of (67). The additional maximum variance unfolding [78] type regularization  $-\gamma \text{Tr}(\mathbf{T})$  prevents the fragments from clustering too tightly by maximizing the spread of the translations [4]. Empirically, we find that a small regularization parameter  $\gamma$  between  $10^{-3}$  and  $10^{-2}$  works well. After obtaining the solution  $\mathbf{T}^*$ , we simply use  $\mathbf{T}^*(F+1:F+3, 1:F)$  as the translations for the fragments.

We conclude this section with a toy example that demonstrates the superiority of joint translation estimation using SDP. For the convenience of illustration, we provide the example in 2D. In order to sequentially assemble the fragments from pairwise distances, it is necessary that there is a pair of fragments where there are at least two distance measurements between them. This is needed to fix the relative translation between the two fragments with two degrees of freedom. In the toy example in Figure 2, this necessary condition for greedy sequential methods is not satisfied, but even so by solving (69) we are able to recover the correct positions of the fragments. This property is quite important, since in practice there are typically only a few NOE restraints between secondary elements of the protein backbone (with the exception of  $\beta$  strands) [51].

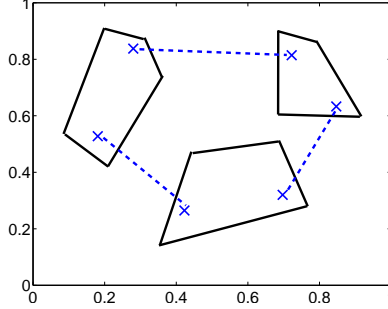


Figure 2: Three fragments in 2D positioned by solving (69) using the distance measurements (Blue dotted lines). While it is impossible to determine the translations sequentially with the distance measurement pattern shown here, with the convex program (69) the three fragments can be assembled jointly.

## 7 Numerical experiments

### 7.1 Synthetic data

In this section, we present the results of numerical simulations with synthetic data for RDC-SOS and RDC-NOE-SOS. All numerical experiments are run on a Samsung NP940X3G laptop with a Intel(R) Core(TM) i5-4200 2.3GHz CPU and 4 GB of memory. We first describe the noise model in our simulations. Let  $\boldsymbol{\zeta} = [\boldsymbol{\zeta}_1, \dots, \boldsymbol{\zeta}_K] \in \mathbb{R}^{3 \times K}$  be the ground truth coordinates. We drop the superscript “ $i$ ” when denoting the atom coordinate since the membership of an atom to a rigid unit is immaterial here. Now let  $E_{\text{RDC}}$  be the set of atom pairs with RDC measured, and assume that the RDC measurements are generated through

$$r_{nm}^{(j)} = \mathbf{v}_{nm}^T \mathbf{S}^{(j)} \mathbf{v}_{nm} + \sigma \boldsymbol{\varepsilon}_{nm}^{(j)}, \quad (n, m) \in E_{\text{RDC}}, j = 1, 2, \quad (70)$$

where the bond direction  $\mathbf{v}_{nm}$  is related to the coordinates  $\boldsymbol{\zeta}_n, \boldsymbol{\zeta}_m$  through

$$\mathbf{v}_{nm} = \frac{\boldsymbol{\zeta}_n - \boldsymbol{\zeta}_m}{\|\boldsymbol{\zeta}_n - \boldsymbol{\zeta}_m\|_2}. \quad (71)$$

We assume  $\boldsymbol{\varepsilon}_{nm}^{(j)} \sim \mathcal{N}(0, 1)$  where  $\mathcal{N}(0, 1)$  is the standard normal distribution. While it is quite common for different types of atomic pairs with RDC measured at different levels of uncertainty, in this section we assume  $r_{nm}$ ’s are all corrupted by i.i.d. Gaussian noise of same variance  $\sigma^2$ .

In this simulation study, we use the  $\alpha$  helix of the protein ubiquitin (residue 24 - residue 33) to generate synthetic RDC data. The data file for the PDB entry 1D3Z contains RDC datasets measured in two alignment media. From the known PDB structure, we determine the two Saupe tensors  $\mathbf{S}^{(1)}, \mathbf{S}^{(2)}$  in these alignment media and use them for simulation purposes. We simulate synthetic RDC data using the noise

model (70) where atom pair directions are obtained from the ground truth PDB model. For this simulation we use the pairs (N, H), (C, CA), (C, N) from the peptide plane, and (CA, HA) from the CA-body to generate RDCs, as the RDCs associated with these pairs are commonly measured. In addition to RDC measurements, we also run the simulation with the aid of 16 NOE restraints on the backbone for residues 24-33. The form of NOE restraints is in terms of hard upper and lower bounds (instead of using them as a penalty term in the cost in RDC-NOE-SOS). To measure the quality of a coordinate estimator  $\hat{\zeta}$ , we use the Root-Mean-Square-Distance (RMSD)

$$\text{RMSD} = \sqrt{\frac{\|\hat{\zeta} - \zeta\|_F^2}{K}} \quad (72)$$

where  $\zeta$  is the starting PDB model. We evaluate the RMSD for the atoms CA, CB, C, N, H, O and HA in all amino acids.

We present the simulation results in Figure 3. We simulate RDC noise with  $\sigma \in [0, 5e-5]$ . Every data point is averaged over 40 noise realizations of RDC. When there is no noise, RDC-SOS and RDC-NOE-SOS exactly recover the rotations with  $t = 4$ . This is a property that simulated annealing based methods do not enjoy, as even without noise these methods can still suffer from local minima. In this simulation,  $\mathcal{M}_{2t, \text{even}, i}^*$  returned with  $t = 4$  are rank-1 to  $10^{-2}$  precision most of the time, therefore we do not use manifold optimization to further refine the solution. We further compare our results against the Cramér-Rao lower bound. The CRB (formula given in Section 10.3.1) provides an information-theoretic lower bound for the least possible variance that can be achieved by any unbiased coordinate estimator. With RDC-SOS and RDC-NOE-SOS we are able to obtain RMSD lower than the CRB. Since RDC-SOS and RDC-NOE-SOS can produce biased coordinate estimators, their error can be lower than the CRB. Here we remark that we slightly abused terminology by referring to the normalized RDC as RDC, where the un-normalized RDC is defined in (76). We emphasize that when  $\sigma = 5e-5$ , the magnitude of noise on the un-normalized RDC is rather large. For example, since the dipolar coupling constant for the N-H RDC is about 23 kHz, when  $\sigma = 5e-5$  the actual noise is 1.15 Hz. This is larger than the typical experimental uncertainty of N-H RDC ( $<0.5$  Hz) [30].

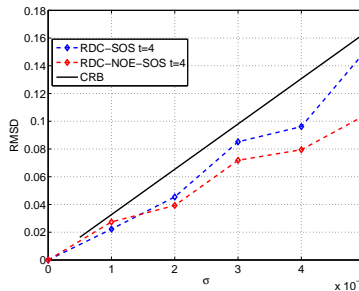


Figure 3: Comparison between running RDC-SOS and RDC-NOE-SOS with  $t = 4$ . Here we do not use manifold optimization to further refine the solution.

## 7.2 Experimental data

In this section, we present results on the analysis of experimental RDC data obtained in two alignment media for ubiquitin. We also provide a comparison of our methods with the molecular fragment replacement (MFR) method proposed in [3] using the full ubiquitin sequence with 76 amino acids and about 500 backbone atoms. We first give a brief introduction to the MFR method. MFR is an RDC-based method that determines the structure of a protein through finding homologous structures in the PDB for short fragments of the protein. For a short fragment, candidate structures from the PDB are used to construct the coordinates in (2). Then a least-squares procedure detailed in the appendix is used to obtain the Saupe tensor based on the experimentally measured RDC and the candidate structure. If a PDB candidate structure gives a low residual in the least-squares fitting, it will be deemed a structure similar to the protein fragment under inspection. Other experimental information such as chemical shifts can also be compared to the information recorded in the database to find a similar structure. The homologous structures for short fragments of the protein are then merged and simulated annealing is applied to further refine the structure based on the RDC measurements. In this numerical study, we start simulated annealing with temperature of 600 K and cool down to 0 K in 30000 steps. For a fair comparison between MFR and our proposed methods, we do not use chemical shift information for the MFR procedure but only RDC and NOE.

We only consider the peptide planes and CA-bodies coming from the first 70 amino acids since the last 6 residues are highly flexible and do not contribute to rigid constraints. We solve for the structure using RDC measurements from the bonds (C, N), (N, H), (CA, HA) in two alignment media. Here we do not use (C, CA) RDCs as in the previous subsection in order to demonstrate the usefulness of the proposed methods when there are less data. We also examine the situation when we are supplemented with 187 experimentally reported backbone NOE's. We use RDC-SOS and RDC-NOE-SOS with  $t = 4$  to solve the backbone structure of six ubiquitin fragments, each containing 12-13 residues on average. We split the fragments at amino-acid sites where there are too few or no RDC measurements. The results are summarized in Table 1. When using only RDC, it is more difficult to determine the backbone structure near the starting and end point of a fragment since RDC measurements are generally sparser in those regions. In this situation, having extra distance constraints may help.

In terms of accuracy, due to the additional distance restraints, RDC-NOE-SOS outperforms RDC-SOS, except for fragment 2. The lack of constraints on residue 9 and 18 causes RDC-SOS and RDC-NOE-SOS to give different solutions with the same cost. The average RMSD of the fragments are 0.47 Å and 0.39 Å for RDC-SOS and RDC-NOE-SOS respectively when comparing with the X-ray structure 1UBQ [73]. To provide a different perspective, we also compare the results from our method with the high resolution NMR structure 1D3Z [16]. When we combine the fragments using (69), the conformation errors of the whole protein backbone obtained from fragments determined by RDC-SOS and RDC-NOE-SOS are 1.05 (1.00) Å and 0.86 (0.80) Å RMSD respectively when comparing to 1UBQ (1D3Z). Figure 4 further compares the backbone traces obtained from our proposed methods and the X-ray structure. Our results are competitive when comparing to MFR, which gives average fragment RMSD



of 0.54 Å and overall RMSD of 0.87 Å. Since RDC-SOS only involves  $M$  PSD variables of size  $35 \times 35$  when  $t = 4$ , whereas RDC-NOE-SOS involves another PSD variable of size  $3M \times 3M$ , the running time of RDC-SOS is faster than RDC-NOE-SOS.

Fragment No.		1	2	3	4	5
Residue No.		1-7	9-18	22-36	37-53	54-70
RMSD (Å) 1UBQ	RDC-SOS	0.36	0.34	0.51	0.56	0.57
	RDC-NOE-SOS	0.37	0.51	0.31	0.51	0.25
	MFR	0.42	0.51	0.45	0.78	0.52
RMSD (Å) 1D3Z	RDC-SOS	0.33	0.25	0.46	0.51	0.54
	RDC-NOE-SOS	0.26	0.49	0.20	0.50	0.17
	MFR	0.40	0.46	0.42	0.71	0.44
Time (s)	RDC-SOS	14	18	30	35	33
	RDC-NOE-SOS	20	22	80	111	114
	MFR	1560 (all 5 fragments)				

Table 1: Results of computing the structure of five ubiquitin fragments using RDC-SDP, RDC-NOE-SDP and MFR from experimental data. We compare with both the X-ray structure 1UBQ and the high resolution NMR structure 1D3Z. For MFR we only report the total running time for calculating the entire backbone.

## 8 Conclusion

We presented two novel convex relaxations RDC-SOS and RDC-NOE-SOS to calculate the protein backbone conformation from both RDC and NOE measurements. In simulations, our methods exactly recover the protein structure when there is no noise, whereas simulated annealing based methods can still get trapped at local minima even when the data is clean. We demonstrate the robustness of our algorithm in simulation by showing that in the presence of noise, the error of our solution attains the CRB. We further demonstrated the success of our methods by obtaining a backbone structure of 1 Å resolution for ubiquitin using experimental data. Both proposed methods are fast in practice, in particular RDC-SOS can determine a protein fragment of typical size in less than half a minute. In comparison, the running time of current methods such as MFR, RDC-Analytics and REDCRAFT range from tens of minutes to two hours. This property of our algorithm can be useful when iterating between estimating resonance or NOE assignments and structural calculation [25].

There are a few remaining problems we would like to address in future works. In this paper, we only use the SOS hierarchy with  $t = 4$  which is sufficient for the numerical experiments considered. We would like to investigate the utility of  $t > 4$  in the SOS hierarchy for sparse RDC data, especially with only one alignment media.

At this point, both RDC-SOS and RDC-NOE-SOS can only compute the structure of the protein backbone but not the protein side-chains. RDC measurements on side chains are complicated by the existence of rotamer states and only a few recent analyses are able to address this issue [42]. On the other hand, a major obstacle of obtaining

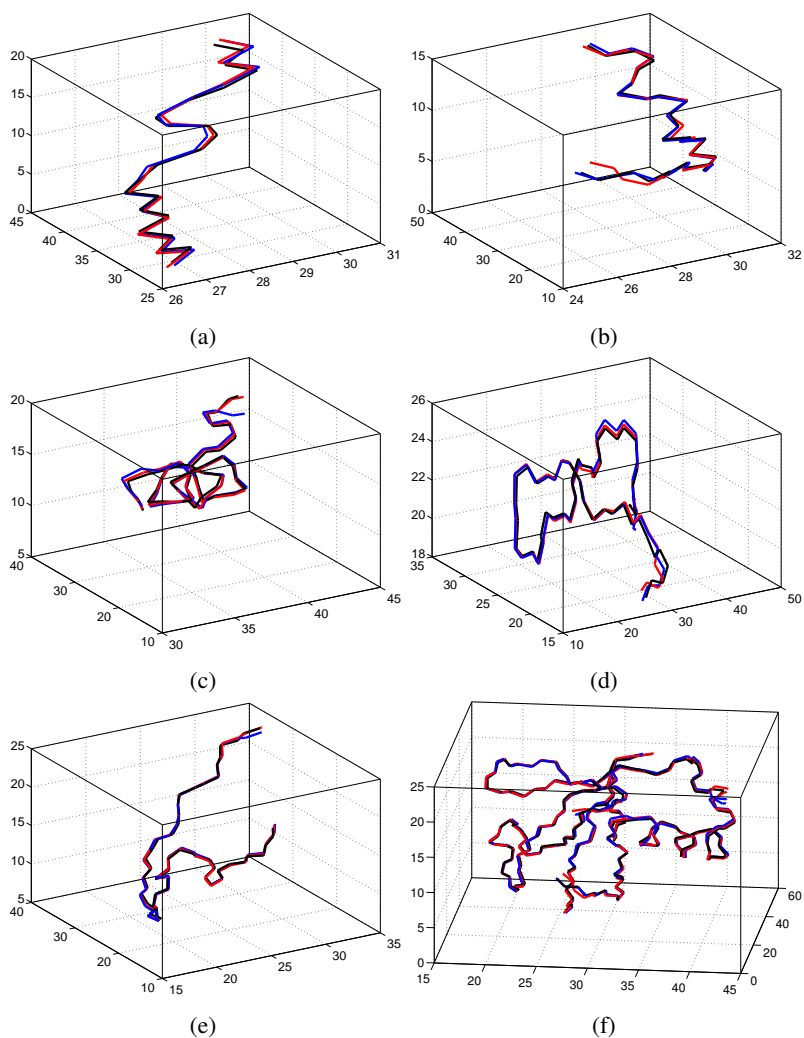


Figure 4: The trace of protein backbone drawn using N, CA and C. (a),(b),(c),(d),(e): Fragments 1,2,3,4 and 5 of ubiquitin defined in Table 1. The black, blue and red curves come from the X-ray model 1UBQ, RDC-SOS solution and RDC-NOE-SOS respectively. (f): Full backbone structure obtained from assembling the five ubiquitin fragments.

complete NOE restraints for the protein side-chains is the ambiguity in NOE assignment, especially for larger systems. We hope to extend our proposed methods to help detecting the correct NOE assignments for the side-chains, through providing a high quality backbone conformation for assignment validation.

Currently, our method requires sufficient NOE restraints between the fragments when combining the fragments together using the convex program (69). However,

noted in [82], there can be very few NOE restraints between the secondary structural elements. We observed such a situation when applying our algorithm to the protein DinI (PDB ID: 1GHH). While all the fragments in DinI can be determined by our proposed method to within 1 Å resolution, our method failed to assemble the fragments together due to the lack of inter-fragments NOE. We hope to solve this issue in the future by including database derived restraints. For example, torsion angle restraints can be derived from chemical shifts of backbone atoms using TALOS [62]. Furthermore, side-chain rotamer library [47] can be used to model protein side-chains, which can in turn provide additional NOE restraints arising from the side-chains.

## 9 Acknowledgements

The authors would like to thank James Saunderson for discussions related to unit quaternion parameterization for optimization problems on  $\mathbb{S}\mathbb{O}(3)$ . The authors are grateful to João M. Pereira, Roy R. Lederman and Yutong Chen for discussions regarding this problem, to Nicolas Boumal for the discussion on manifold optimization and proof-reading an earlier version of this manuscript. The authors also want to thank Richard Harris and Roberto Tejero for assisting with interpreting and reading NMR restraint files. The research of AS was partially supported by award R01GM090200 from the NIGMS, by awards FA9550-12-1-0317 from AFOSR, by the Simons Foundation investigator award and the Simons Foundation Collaboration on Algorithms and Geometry, and the Moore Foundation Data Driven Discovery Investigator award.

## 10 Appendix

### 10.1 The residual dipolar coupling and Saupe tensor

We give here a brief introduction to RDC and the Saupe tensor, while a detailed exposition can be found in [70] for example. Let  $\mathbf{v}_{nm}$  be the unit vector denoting the direction of the bond between nuclei  $n$  and  $m$ . Let  $\mathbf{b}$  be the unit vector denoting the direction of the magnetic field. The RDC  $D_{nm}$  due to the interaction between nuclei  $n$  and  $m$  is

$$D_{nm} = D_{nm}^{\max} \left\langle \frac{3(\mathbf{b}^T \mathbf{v}_{nm})^2 - 1}{2} \right\rangle_{t,e}. \quad (73)$$

$D_{nm}^{\max}$  is a constant depending on the gyromagnetic ratios  $\gamma_n, \gamma_m$  of the two nuclei, the bond length  $r_{nm}$ , and the Planck's constant  $h$  as

$$D_{nm}^{\max} = -\frac{\gamma_n \gamma_m h}{2\pi^2 r_{nm}^3}, \quad (74)$$

and  $\langle \cdot \rangle_{t,e}$  denotes the ensemble and time averaging operator. As presented, RDC depends on the relative angle between the magnetic field and the bond. By extension of terminology, we refer to the normalized RDC

$$r_{nm} = D_{nm} / D_{nm}^{\max} \quad (75)$$

as simply the RDC.

It is conventional to interpret the RDC measurement in the molecular frame. More precisely, we treat the molecule as being static in some coordinate system, and the magnetic field direction being a time and sample varying vector. In this case the RDC becomes

$$D_{nm} = D_{nm}^{\max} \mathbf{v}_{nm}^T \mathbf{S} \mathbf{v}_{nm}, \quad (76)$$

where the Saupe tensor  $\mathbf{S}$  is defined as

$$\mathbf{S} = \frac{1}{2}(3\mathbf{B} - \mathbf{I}_3), \quad \mathbf{B} = \left\langle \mathbf{b}\mathbf{b}^T \right\rangle_{t,e}. \quad (77)$$

We note that  $\mathbf{S}$  is symmetric and  $\text{Tr}(\mathbf{S}) = 0$ . In order to use RDC for structural refinement of a protein,  $\mathbf{S}$  is usually first determined from a known structure (known  $\mathbf{v}_{nm}$ ) that is similar to the protein.

We now detail a classical way of obtaining the Saupe tensor from a known template structure [46]. Using the fact that  $\mathbf{S}$  is symmetric and  $\text{Tr}(\mathbf{S}) = 0$ , eq. (76) can be rewritten as

$$\begin{aligned} r_{nm} = & (\mathbf{v}_{nm2}^2 - \mathbf{v}_{nm1}^2)\mathbf{S}(2,2) + (\mathbf{v}_{nm3}^2 - \mathbf{v}_{nm1}^2)\mathbf{S}(3,3) \\ & + 2\mathbf{v}_{nm1}\mathbf{v}_{nm2}\mathbf{S}(1,2) + 2\mathbf{v}_{nm1}\mathbf{v}_{nm3}\mathbf{S}(1,3) \\ & + 2\mathbf{v}_{nm2}\mathbf{v}_{nm3}\mathbf{S}(2,3) \end{aligned} \quad (78)$$

where  $\mathbf{v}_{nmi}$ ,  $i = x, y, z$  are the different components of  $\mathbf{v}_{nm}$  in the molecular frame. When there are  $L$  RDC measurements, eq. (78) results in  $L$  linear equations in five unknowns ( $\mathbf{S}(2,2), \mathbf{S}(3,3), \mathbf{S}(1,2), \mathbf{S}(1,3)$  and  $\mathbf{S}(2,3)$ ), that can be written in matrix form as

$$\mathbf{A} \mathbf{s} = \mathbf{r}, \quad \mathbf{s} = \begin{bmatrix} \mathbf{S}(2,2) \\ \mathbf{S}(3,3) \\ \mathbf{S}(1,2) \\ \mathbf{S}(1,3) \\ \mathbf{S}(2,3) \end{bmatrix} \in \mathbb{R}^5, \quad \mathbf{r} = \begin{bmatrix} r_{n_1 m_1} \\ \vdots \\ r_{n_L m_L} \end{bmatrix} \in \mathbb{R}^M \quad (79)$$

and  $\mathbf{A} \in \mathbb{R}^{L \times 5}$ . An ordinary least squares procedure can be used to estimate  $\mathbf{s}$  if  $\mathbf{A}$  has full rank. This is also referred to as the SVD procedure in [46].

## 10.2 Sum-of-squares relaxation

In this section, we explain why the convex relaxation presented in Section 4 is coined SOS. The polynomial optimization problem

$$p_1 = \min_{\mathbf{x} \in \mathbb{R}^n} f(\mathbf{x}) \quad \text{s.t. } h(\mathbf{x}) = 0, \quad (80)$$

where  $f(\mathbf{x}), h(\mathbf{x})$  are polynomial functions, can be expressed equivalently as

$$\max_d \quad \text{s.t. } f(\mathbf{x}) - d \geq 0 \quad \text{on } h(\mathbf{x}) = 0. \quad (81)$$

This is equivalent to

$$d_1 = \max_{d, t_{\alpha}} d \quad \text{s.t. } f(\mathbf{x}) - d + \left(\sum_{\alpha} t_{\alpha} \mathbf{x}^{\alpha}\right) h(\mathbf{x}) \geq 0 \quad \forall \mathbf{x} \quad (82)$$

[6, Chapter 3], which is actually the dual problem to (80). However, due to the NP-hardness in testing the non-negativity of a polynomial [6], we further restrict the search space from the set of non-negative polynomials to the set of SOS polynomials:

$$d_2 = \max_{d, t_{\alpha}} d \quad \text{s.t. } f(\mathbf{x}) - d + \left(\sum_{\alpha} t_{\alpha} \mathbf{x}^{\alpha}\right) h(\mathbf{x}) \text{ is SOS.} \quad (83)$$

This results in a standard SDP

$$\max_{d, \mathbf{P} \succeq 0, t_{\alpha}} d \quad \text{s.t. } f(\mathbf{x}) - d + \left(\sum_{\alpha} t_{\alpha} \mathbf{x}^{\alpha}\right) h(\mathbf{x}) = [\mathbf{x}]_t^T \mathbf{P} [\mathbf{x}]_t \quad (84)$$

for some specific choices of  $t$ . Since  $p_1 = d_1 \geq d_2$ , solving (84) provides a lower bound to (80). Indeed, the dual of (84) is exactly the type of convex relaxations presented in Section 4 for optimization problems of the form (80).

### 10.3 Cramér-Rao lower bound

In this section, we introduce a classical tool from statistics, the Cramér-Rao bound (CRB) [12], to give perspective on the lowest possible error any unbiased estimator can achieve when estimating coordinates from noisy RDC measurements. We first describe the CRB for general point estimators. Let  $\boldsymbol{\theta} \in \mathbb{R}^n$  be a multidimensional parameter which is to be estimated from measurements  $\mathbf{x} \in \mathbb{R}^m$ . Suppose  $\mathbf{x}$  is generated from the distribution  $p(\mathbf{x}|\boldsymbol{\theta})$ . The Fisher information matrix (FIM) is defined as the  $n \times n$  matrix

$$\mathbf{I}(\boldsymbol{\theta}) = \mathbb{E}[(\nabla_{\boldsymbol{\theta}} \ln p(\mathbf{x}|\boldsymbol{\theta}))(\nabla_{\boldsymbol{\theta}} \ln p(\mathbf{x}|\boldsymbol{\theta}))^T] \quad (85)$$

where expectation is taken with respect to the distribution  $p(\mathbf{x}|\boldsymbol{\theta})$  and the gradient  $\nabla_{\boldsymbol{\theta}}$  is taken with respect to  $\boldsymbol{\theta}$ . For any unbiased estimator  $\hat{\boldsymbol{\theta}}$  of  $\boldsymbol{\theta}$ , that is  $\mathbb{E}(\hat{\boldsymbol{\theta}}) = \boldsymbol{\theta}$ , the following relationship holds:

$$\mathbb{E}[(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})^T] \succeq \mathbf{I}(\boldsymbol{\theta})^{-1} \quad (86)$$

if  $\mathbf{I}(\boldsymbol{\theta})$  is invertible. Therefore the total variance of the estimator  $\hat{\boldsymbol{\theta}}$  is lower bounded by  $\text{Tr}(\mathbf{I}(\boldsymbol{\theta})^{-1})$ . We remark that for an unbiased estimator, its variance and the mean-squared error are the same, therefore we often use these terms interchangeably.

We also introduce the CRB in the case when  $\boldsymbol{\theta}$  and  $\hat{\boldsymbol{\theta}}$  are constrained to be in the set  $\{\boldsymbol{\theta} | f(\boldsymbol{\theta}) = 0\}$  where  $f: \mathbb{R}^n \rightarrow \mathbb{R}^k$  [66]. Let  $\mathbf{Df}(\boldsymbol{\theta}) \in \mathbb{R}^{k \times n}$  be the gradient matrix of  $f$  at  $\boldsymbol{\theta}$  with full row rank, and  $\mathbf{Q} \in \mathbb{R}^{n \times (n-k)}$  be a set of orthonormal vectors satisfying

$$\mathbf{Df}(\boldsymbol{\theta})\mathbf{Q} = 0 \quad (87)$$

i.e.  $\mathbf{Q}$  is an orthonormal basis of the null space of  $\mathbf{Df}(\boldsymbol{\theta})$ . In this case, for any unbiased estimator  $\hat{\boldsymbol{\theta}}$  satisfying  $f(\hat{\boldsymbol{\theta}}) = 0$ , the CRB is then

$$\mathbb{E}[(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})^T] \succeq \mathbf{Q}\mathbf{Q}^T \mathbf{I}(\boldsymbol{\theta})\mathbf{Q}^{-1} \mathbf{Q}^T \quad (88)$$

if  $\mathbf{Q}^T \mathbf{I}(\boldsymbol{\theta})\mathbf{Q}$  is invertible.

### 10.3.1 CRB for the variance of coordinate estimator

We are now ready to investigate the CRB for estimating atomic positions from RDC data. Let  $\boldsymbol{\zeta} = [\boldsymbol{\zeta}_1, \dots, \boldsymbol{\zeta}_K] \in \mathbb{R}^{3 \times K}$  be the coordinates of the atoms we want to estimate. We aim to derive a lower bound  $\text{Tr}(\mathbf{Q}(\mathbf{Q}^T \mathbf{I}(\boldsymbol{\zeta}) \mathbf{Q})^{-1} \mathbf{Q}^T)$  of  $\mathbb{E}[\text{Tr}((\hat{\boldsymbol{\zeta}} - \boldsymbol{\zeta})^T (\hat{\boldsymbol{\zeta}} - \boldsymbol{\zeta}))]$  for any unbiased estimator  $\hat{\boldsymbol{\zeta}}$  of  $\boldsymbol{\zeta}$ , and compare

$$\sqrt{\frac{\text{Tr}(\mathbf{Q}(\mathbf{Q}^T \mathbf{I}(\boldsymbol{\zeta}) \mathbf{Q})^{-1} \mathbf{Q}^T)}{K}} \quad (89)$$

with the RMSD of the solutions from RDC-SOS and RDC-NOE-SOS in Fig. 3.

We assume that the RDC measurements are generated through the noise model in (70). This noise model is used to get an expression for  $\mathbf{I}(\boldsymbol{\theta})$ . There are several sets of equality constraints that need to be considered when deriving  $\mathbf{Q}$ . We assume that within each rigid unit, the distance between any pair of atoms is fixed. We therefore have a set of equality constraints

$$d_{nm}^2 = \|\boldsymbol{\zeta}_n - \boldsymbol{\zeta}_m\|_2^2, \quad (n, m) \in E_{\text{fixed}} \quad (90)$$

where  $E_{\text{fixed}}$  consists of all atom pairs within each and every rigid unit. Without loss of generality, we also consider the constraint

$$\boldsymbol{\zeta} \mathbf{1} = 0 \quad (91)$$

which implies the points  $\boldsymbol{\zeta}_1, \dots, \boldsymbol{\zeta}_K$  are centered at zero. This is due to the fact that

$$\begin{aligned} & \text{Tr}((\hat{\boldsymbol{\zeta}} - \boldsymbol{\zeta})^T (\hat{\boldsymbol{\zeta}} - \boldsymbol{\zeta})) \\ &= \text{Tr}((\hat{\boldsymbol{\zeta}}_c - \boldsymbol{\zeta}_c - t \mathbf{1}^T)^T (\hat{\boldsymbol{\zeta}}_c - \boldsymbol{\zeta}_c - t \mathbf{1}^T)) \\ &= \text{Tr}((\hat{\boldsymbol{\zeta}}_c - \boldsymbol{\zeta}_c)^T (\hat{\boldsymbol{\zeta}}_c - \boldsymbol{\zeta}_c)) + (1/K) \|t\|_2^2 \\ & \quad - 2 \text{Tr}((\hat{\boldsymbol{\zeta}}_c - \boldsymbol{\zeta}_c)^T t \mathbf{1}^T) \\ &= \text{Tr}((\hat{\boldsymbol{\zeta}}_c - \boldsymbol{\zeta}_c)^T (\hat{\boldsymbol{\zeta}}_c - \boldsymbol{\zeta}_c)) + (1/K) \|t\|_2^2 \\ &\geq \text{Tr}((\hat{\boldsymbol{\zeta}}_c - \boldsymbol{\zeta}_c)^T (\hat{\boldsymbol{\zeta}}_c - \boldsymbol{\zeta}_c)) \end{aligned} \quad (92)$$

where  $\boldsymbol{\zeta}_c$  and  $\hat{\boldsymbol{\zeta}}_c$  denote the zero centered coordinates and coordinate estimators, and  $t$  is the relative translation between  $\boldsymbol{\zeta}$  and  $\hat{\boldsymbol{\zeta}}$ . Eq. (92) implies that deriving a lower bound for  $\mathbb{E}[\text{Tr}((\hat{\boldsymbol{\zeta}}_c - \boldsymbol{\zeta}_c)^T (\hat{\boldsymbol{\zeta}}_c - \boldsymbol{\zeta}_c))]$  is sufficient for obtaining a lower bound for  $\mathbb{E}[\text{Tr}((\hat{\boldsymbol{\zeta}} - \boldsymbol{\zeta})^T (\hat{\boldsymbol{\zeta}} - \boldsymbol{\zeta}))]$ . When there are atoms that are constrained to lie on the same plane, we need to add the constraint that any three vectors in the plane span a space with zero volume, i.e.

$$\det([\boldsymbol{\zeta}_i - \boldsymbol{\zeta}_j, \boldsymbol{\zeta}_k - \boldsymbol{\zeta}_l, \boldsymbol{\zeta}_m - \boldsymbol{\zeta}_n]) = 0 \quad (93)$$

for atoms  $i, j, k, l, m, n$  in the same plane.

We first start with deriving an expression for the Fisher information matrix when RDC data are generated through (70). From (70) and (71), the likelihood function for

the coordinates is

$$\begin{aligned}
p(\{r_{nm}\}_{(n,m) \in E_{\text{RDC}}} | \boldsymbol{\zeta}_1, \dots, \boldsymbol{\zeta}_K) &= \\
&\prod_{\substack{(n,m) \in E_{\text{RDC}} \\ j=1,2}} \frac{1}{\sqrt{2\pi\sigma^2}} \\
&\exp\left(-\frac{\left((\boldsymbol{\zeta}_n - \boldsymbol{\zeta}_m)^T \mathbf{S}^{(j)} (\boldsymbol{\zeta}_n - \boldsymbol{\zeta}_m) - r_{nm}^{(j)} d_{nm}^2\right)^2}{2d_{nm}^4 \sigma^2}\right) \quad (94)
\end{aligned}$$

and the log-likelihood is (up to an additive constant)

$$\begin{aligned}
&l(\{r_{nm}\}_{(n,m) \in E_{\text{RDC}}} | \boldsymbol{\zeta}_1, \dots, \boldsymbol{\zeta}_K) \\
:= &\ln p(\{r_{nm}\}_{(n,m) \in E_{\text{RDC}}} | \boldsymbol{\zeta}_1, \dots, \boldsymbol{\zeta}_K) \\
= &\sum_{\substack{(n,m) \in E_{\text{RDC}} \\ j=1,2}} \frac{-\left((\boldsymbol{\zeta}_n - \boldsymbol{\zeta}_m)^T \mathbf{S}^{(j)} (\boldsymbol{\zeta}_n - \boldsymbol{\zeta}_m) - r_{nm}^{(j)} d_{nm}^2\right)^2}{2d_{nm}^4 \sigma^2} \\
= &-\sum_{\substack{(n,m) \in E_{\text{RDC}} \\ j=1,2}} \frac{(\mathbf{e}_{nm}^T \boldsymbol{\zeta}^T \mathbf{S}^{(j)} \boldsymbol{\zeta} \mathbf{e}_{nm} - r_{nm}^{(j)} d_{nm}^2)^2}{2d_{nm}^4 \sigma^2} \quad (95)
\end{aligned}$$

where  $\mathbf{e}_{nm} = \mathbf{e}_n - \mathbf{e}_m$ . The derivative of  $l$  with respect to  $\text{vec}(\boldsymbol{\zeta})$  is then

$$\begin{aligned}
\nabla_{\text{vec}(\boldsymbol{\zeta})} l &= \\
&-\sum_{\substack{(n,m) \in E_{\text{RDC}} \\ j=1,2}} \frac{2(\mathbf{e}_{nm}^T \boldsymbol{\zeta}^T \mathbf{S}^{(j)} \boldsymbol{\zeta} \mathbf{e}_{nm} - r_{nm}^{(j)} d_{nm}^2)}{d_{nm}^4 \sigma^2} \\
&\quad (\mathbf{e}_{nm} \mathbf{e}_{nm}^T \otimes \mathbf{S}^{(j)}) \text{vec}(\boldsymbol{\zeta}). \quad (96)
\end{aligned}$$

It follows from the noise model (70) and the independence of  $\boldsymbol{\epsilon}_{nm}^{(j)}$ 's that the Fisher information matrix

$$\begin{aligned}
\mathbf{I}(\boldsymbol{\zeta}) &= \mathbb{E}((\nabla_{\text{vec}(\boldsymbol{\zeta})} l)(\nabla_{\text{vec}(\boldsymbol{\zeta})} l)^T) = \\
&4 \sum_{\substack{(n,m) \in E_{\text{RDC}} \\ j=1,2}} \\
&\frac{(\mathbf{e}_{nm} \mathbf{e}_{nm}^T \otimes \mathbf{S}^{(j)}) \text{vec}(\boldsymbol{\zeta}) \text{vec}(\boldsymbol{\zeta})^T (\mathbf{e}_{nm} \mathbf{e}_{nm}^T \otimes \mathbf{S}^{(j)})}{\sigma^2 d_{nm}^4} \quad (97)
\end{aligned}$$

Having the Fisher information matrix, we now incorporate the constraints in (90) and (91) in order to obtain a bound as in (88). Stacking the equality constraints (90) into a  $|E_{\text{fixed}}| \times 1$  matrix, we get

$$f(\text{vec}(\boldsymbol{\zeta})) := \left[ \mathbf{e}_{nm}^T \boldsymbol{\zeta}^T \boldsymbol{\zeta} \mathbf{e}_{nm} - d_{nm}^2 \right]_{(n,m) \in E_{\text{fixed}}} = 0 \quad (98)$$

The gradient matrix is thus

$$\mathbf{Df}(\text{vec}(\boldsymbol{\zeta})) = \text{vec}(\boldsymbol{\zeta})^T [(\mathbf{e}_{nm}\mathbf{e}_{nm}^T \otimes \mathbf{I}_3)]_{(n,m) \in E_{\text{fixed}}} \quad (99)$$

where  $\mathbf{Df}(\text{vec}(\boldsymbol{\zeta})) \in \mathbb{R}^{|E_{\text{fixed}}| \times 3K}$ . We note that  $\mathbf{Df}(\text{vec}(\boldsymbol{\zeta}))$  is known as the *rigidity matrix* [31], and the vectors in its null space indicate the direction of infinitesimal motion the atoms can take without violating (90). Even in the case when all pairwise distances between the atoms are known, there is still a 6-dimensional null space for  $\mathbf{Df}(\text{vec}(\boldsymbol{\zeta}))$ , corresponding to an infinitesimal global rotation and translation to the coordinates  $\boldsymbol{\zeta}$  that preserves all pairwise distances. We now augment  $f(\text{vec}(\boldsymbol{\zeta})) = 0$  with the centering constraint  $\boldsymbol{\zeta} \mathbf{1} = 0$ , and this augments  $\mathbf{Df}(\text{vec}(\boldsymbol{\zeta}))$  with three rows  $\mathbf{1}^T \otimes \mathbf{I}_3$ , i.e.

$$\mathbf{Df}(\text{vec}(\boldsymbol{\zeta})) = \begin{bmatrix} \text{vec}(\boldsymbol{\zeta})^T [(\mathbf{e}_{nm}\mathbf{e}_{nm}^T \otimes \mathbf{I}_3)]_{(n,m) \in E_{\text{fixed}}} \\ \mathbf{1}^T \otimes \mathbf{I}_3 \end{bmatrix} \quad (100)$$

The inclusion of such centering constraint eliminates the three dimensional subspace in the kernel of the rigidity matrix that corresponds to the translational degree of freedom. Let  $\mathbf{Q}$  be an orthonormal basis that spans the null space of  $\mathbf{Df}(\text{vec}(\boldsymbol{\zeta}))$ . Together with (97) and (88) we obtain the desired CRB. We omit detailing the derivative for constraint (93) but simply note that the inclusion of such constraints eliminates the out of plane infinitesimal motion for atoms lying on rigid planar unit.

### 10.3.2 Inclusion of NOE constraints

We have so far neglected the use of NOE measurements when deriving the CRB. Unlike RDC, the NOE restraints remain more qualitative, with imprecise upper and lower bound [7] due to the  $r^{-6}$  scaling of the interaction. In protein structural calculation, it is customary to include a flat potential well-like penalty (e.g. (12)) in addition to the RDC log-likelihood function derived from RDC, or treat the backbone NOE as inequality constraints on the distances. In any of these cases, when the coordinates  $\boldsymbol{\zeta}$  strictly satisfy both upper and lower bounds on the distances, the CRB is exactly the same as the CRB derived in Section 10.3.1 [23] since the CRB only depends on the local curvature of the log-likelihood function around  $\boldsymbol{\zeta}$ . Therefore when the noise on RDC is large and the NOE restraints are active in determining a coordinate estimator  $\hat{\boldsymbol{\zeta}}$ , the CRB may no longer serve as a lower bound for the mean squared error of  $\hat{\boldsymbol{\zeta}}$ . In particular, it is possible for  $\hat{\boldsymbol{\zeta}}$  to have a mean squared error lower than the CRB due to the bias introduced by the NOE (by favoring solutions that satisfies the distance bounds), as observed in Fig. 3. A fundamental results in statistical estimation theory-the bias-variance trade-off [77], states that the mean squared error of an estimator can be obtained from the summation of the variance and squared bias of the estimator. It is possible that with the expense of having some bias, the variance of an estimator can be greatly reduced, resulting a mean squared error that is lower than the CRB [77, Chapter 7].

### 10.3.3 Observed Fisher information matrix and protein variability

We remark that since the LHS of (86) (or (88)) is the covariance matrix for the estimator  $\hat{\boldsymbol{\theta}}$ , the leading eigenvectors of  $\mathbf{I}(\boldsymbol{\theta})^{-1}$  give the direction of the greatest variations of



the protein based on the observed data, whereas the corresponding eigenvalues give the variance (amplitude) of variations. When deriving the CRB, we use the FIM (97) which is obtained from averaging over the distribution of the data. An estimator  $\widehat{\mathbf{I}}(\widehat{\boldsymbol{\theta}})$  of the FIM can be obtained from the observed data, by replacing  $\boldsymbol{\theta}$  in FIM by its maximum-likelihood estimator  $\widehat{\boldsymbol{\theta}}$  and plugging in the observed data (in our case the observed RDC  $r_{nm}$ ) instead of taking expectation over the distribution of the data. The direction for which an estimator  $\widehat{\boldsymbol{\theta}}$  has the greatest variance can be estimated by the top eigenvector of  $\widehat{\mathbf{I}}(\widehat{\boldsymbol{\theta}})^{-1}$ . In the constrained case, we compute the top eigenvector of  $\widehat{\mathbf{Q}}(\widehat{\mathbf{Q}}^T \widehat{\mathbf{I}}(\widehat{\boldsymbol{\theta}})^{-1} \widehat{\mathbf{Q}})^{-1} \widehat{\mathbf{Q}}^T$  where  $\widehat{\mathbf{Q}}$  are computed based on  $\widehat{\boldsymbol{\theta}}$  instead of  $\boldsymbol{\theta}$ . In Fig. 5a, we demonstrate the variation of the ubiquitin fragment for residue 1-7 (with 159 atoms) using the eigenvector of estimated FIM. In Fig. 5b we show the largest 10 eigenvalues of the inverse of the estimated FIM. As we see, there is one prominent mode of variation for this protein fragment. We note that this procedure of determining the modes of protein variation bear resemblance to normal mode analysis [11]. In such analysis, the Hessian for the pseudo energy function of a protein near a local minimum is first determined. Then the normal modes are determined by the eigenvectors of the Hessian matrix. If we treat the log-likelihood function as some pseudo energy function, our FIM-based analysis of the modes of atomic displacement corresponds to the classical normal modes analysis.

### 10.3.4 Infinitesimal rigidity and invertibility of the Fisher information matrix

In this subsection, we study the infinitesimal rigidity [43] of the protein structure given RDC and distance measurements and how it guarantees invertibility of the Fisher information matrix. Let a framework with coordinates  $\boldsymbol{\zeta} \in \mathbb{R}^{3 \times K}$  be constrained by

$$(\boldsymbol{\zeta}_n - \boldsymbol{\zeta}_m)^T (\boldsymbol{\zeta}_n - \boldsymbol{\zeta}_m) = d_{nm}^2, \quad (n, m) \in E_{\text{fixed}}, \quad (101)$$

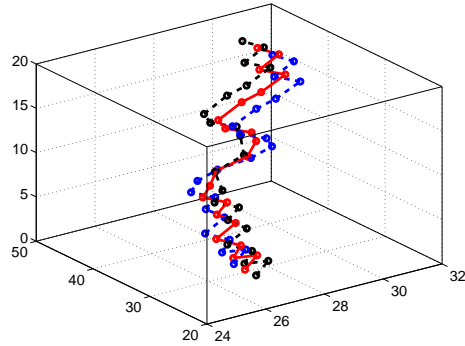
and

$$\begin{aligned} (\boldsymbol{\zeta}_n - \boldsymbol{\zeta}_m)^T \mathbf{S}^{(j)} (\boldsymbol{\zeta}_n - \boldsymbol{\zeta}_m) &= r_{nm}^{(j)}, \\ j &= 1, \dots, N, \quad (n, m) \in E_{\text{RDC}}. \end{aligned} \quad (102)$$

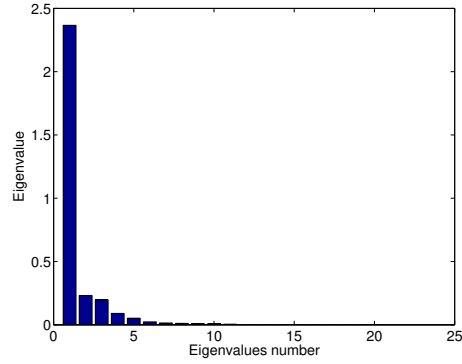
In order to derive a condition for infinitesimal rigidity, we first let  $\text{vec}(\boldsymbol{\zeta}(s))$  be a curve in dimension  $\mathbb{R}^{3K}$  parameterized by  $s$ , where  $\boldsymbol{\zeta}(0)$  satisfies (101) and (102). Taking derivative of the constraints in (101) and (102) with respect to  $s$  at  $s = 0$ , we have

$$\begin{aligned} \left[ \begin{array}{c} \text{vec}(\boldsymbol{\zeta}(0))^T [\mathbf{e}_{nm} \mathbf{e}_{nm}^T \otimes \mathbf{I}_3]_{(n,m) \in E_{\text{fixed}}} \\ \text{vec}(\boldsymbol{\zeta}(0))^T [\mathbf{e}_{nm} \mathbf{e}_{nm}^T \otimes \mathbf{S}^{(j)}]_{(n,m) \in E_{\text{RDC}}, j \in [1, N]} \end{array} \right] \frac{d}{ds} \text{vec}(\boldsymbol{\zeta}(0)) \\ = \mathbf{R}(\boldsymbol{\zeta}(0)) \frac{d}{ds} \text{vec}(\boldsymbol{\zeta}(0)) = 0. \end{aligned} \quad (103)$$

The null space of the generalized rigidity matrix  $\mathbf{R}(\boldsymbol{\zeta}(0))$  with dimension  $(|E_{\text{fixed}}| + |E_{\text{RDC}}|) \times 3K$  represents the direction of infinitesimal motion such that  $\boldsymbol{\zeta}(s)$  satisfies the constraints (101), (102) for infinitesimally small  $s$ . If  $\mathbf{R}(\boldsymbol{\zeta}(0))$  only has a three dimensional nullspace, i.e. the global translations in  $x, y, z$ -directions, we say the framework  $\boldsymbol{\zeta}(0)$  along with the constraints (101) and (102) is infinitesimally rigid.



(a)



(b)

Figure 5: (a) Variation of the structure of a ubiquitin fragment (residue 1-7). We compare the RDC-NOE-SOS solution  $\hat{\zeta}$  (solid) with two other structures (dashed) obtained from adding the top eigenvector of  $\hat{Q}(\hat{Q}^T \hat{I}(\hat{\theta})^{-1} \hat{Q})\hat{Q}^T$  multiplied by small scalars to  $\hat{\zeta}$ . (b) The largest 10 eigenvalues of the inverse of FIM.

Now we verify that the constrained Fisher information matrix is invertible if  $\mathbf{R}(\zeta(0))$  has a three dimensional null space corresponds to global translation of the points. We define  $\ker(\mathbf{A})$  to be the kernel of a matrix  $\mathbf{A}$  and  $\text{range}(\mathbf{A})$  to be the column space of  $\mathbf{A}$ . Let  $\mathbf{Q}$  again be the basis of the nullspace of  $\mathbf{Df}(\text{vec}(\zeta))$  defined in (100) such that  $\mathbf{Df}(\text{vec}(\zeta))\mathbf{Q} = 0$ . Let  $\mathbf{v}$  satisfies

$$\mathbf{Q}^T \mathbf{I}(\zeta) \mathbf{Q} \mathbf{v} = 0$$

$\mathbf{Q}^T \mathbf{I}(\zeta) \mathbf{Q} \mathbf{v} = 0$  if and only if  $\mathbf{v} \in \ker(\mathbf{Q})$  or  $\mathbf{Q} \mathbf{v} \in \ker(\mathbf{I})$ . Since the columns of  $\mathbf{Q}$  are linearly independent,  $\mathbf{Q} \mathbf{v} \neq 0$  unless  $\mathbf{v} = 0$ . This means  $\mathbf{Q}^T \mathbf{I}(\zeta) \mathbf{Q} \mathbf{v} = 0$  if and only if  $\mathbf{v} = 0$  or  $\mathbf{Q} \mathbf{v} \in \ker(\mathbf{I}) \cap \text{range}(\mathbf{Q}) = \ker(\mathbf{I}) \cap \ker(\mathbf{Df}(\text{vec}(\zeta)))$ . Therefore if

$$\ker(\mathbf{I}) \cap \ker(\mathbf{Df}(\text{vec}(\zeta))) = \emptyset,$$

or in other words

$$\text{range}(\mathbf{I}) \cup \text{range}(\mathbf{Df}(\text{vec}(\boldsymbol{\zeta}))) = \mathbb{R}^{3K} \quad (104)$$

then  $\mathbf{Q}^T \mathbf{I}(\boldsymbol{\zeta}) \mathbf{Q}$  is invertible. From the form of the (97), it is easy to show that the range condition (104) is satisfied if and only if the range of

$$\begin{bmatrix} \mathbf{1}^T \otimes \mathbf{I}_3 \\ \text{vec}(\boldsymbol{\zeta}(0))^T [\mathbf{e}_{nm} \mathbf{e}_{nm}^T \otimes \mathbf{I}_3]_{(n,m) \in E_{\text{fixed}}} \\ \text{vec}(\boldsymbol{\zeta}(0))^T [\mathbf{e}_{nm} \mathbf{e}_{nm}^T \otimes \mathcal{S}^{(j)}]_{(n,m) \in E_{\text{RDC}}, j \in [1, N]} \end{bmatrix} = \begin{bmatrix} \mathbf{1}^T \otimes \mathbf{I}_3 \\ \mathbf{R}(\boldsymbol{\zeta}(0)) \end{bmatrix} \quad (105)$$

is  $\mathbb{R}^{3K}$ . Then we arrive at the conclusion that if the framework  $\boldsymbol{\zeta}$  is infinitesimally rigid with the null space of  $\mathbf{R}(\boldsymbol{\zeta})$  being the global translations, the constrained Fisher information matrix defined as  $\mathbf{Q}^T \mathbf{I}(\boldsymbol{\zeta}) \mathbf{Q}$  is invertible.

In [82], it is shown that if there exists RDC measurements for a bond in the peptide plane and a bond in the CA-body in a single alignment media, the solutions of the protein structure form a discrete set. Therefore under this condition, there is no infinitesimal motion other than global translation such that the protein framework satisfies the RDC and NOE constraints. We can thus compute the CRB safely under such condition.

## References

- [1] Babak Alipanahi, Nathan Krislock, Ali Ghodsi, Henry Wolkowicz, Logan Donaldson, and Ming Li, *Determining protein structures from NOESY distance constraints by semidefinite programming*, Journal of Computational Biology **20** (2013), no. 4, 296–310.
- [2] Mykhaylo Andriluka, Stefan Roth, and Bernt Schiele, *Pictorial structures revisited: People detection and articulated pose estimation*, Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on, IEEE, 2009, pp. 1014–1021.
- [3] Ad Bax, Georg Kontaxis, and Nico Tjandra, *Dipolar couplings in macromolecular structure determination.*, Methods in enzymology **339** (2001), 127.
- [4] Pratik Biswas, Tzu-Chen Liang, Kim-Chuan Toh, Yinyu Ye, and Ta-Chung Wang, *Semidefinite programming approaches for network localization with noisy distance measurements*, Automation Science and Engineering, IEEE Transactions on **3** (2006), no. 4, 360–371.
- [5] Martin Blackledge, *Recent progress in the study of biomolecular structure and dynamics in solution from residual dipolar couplings*, Progress in Nuclear Magnetic Resonance Spectroscopy **46** (2005), no. 1, 23–61.
- [6] Grigoriy Blekherman, Pablo Parrilo, and Rekha Thomas, *Semidefinite optimization and convex algebraic geometry*, 2011.
- [7] Alexandre MJJ Bonvin and Axel T Brünger, *Do NOE distances contain enough information to assess the relative populations of multi-conformer structures?*, Journal of biomolecular NMR **7** (1996), no. 1, 72–76.

- [8] Nicolas Boumal, Bamdev Mishra, P-A Absil, and Rodolphe Sepulchre, *Manopt, a matlab toolbox for optimization on manifolds*, The Journal of Machine Learning Research **15** (2014), no. 1, 1455–1459.
- [9] Stephen Boyd and Lieven Vandenberghe, *Convex optimization*, Cambridge university press, 2004.
- [10] Michael Bryson, Fang Tian, James H Prestegard, and Homayoun Valafar, *RED-CRAFT: a tool for simultaneous characterization of protein backbone structure and motion from RDC data*, Journal of Magnetic Resonance **191** (2008), no. 2, 322–334.
- [11] David A Case, *Normal mode analysis of protein dynamics*, Current Opinion in Structural Biology **4** (1994), no. 2, 285–290.
- [12] George Casella and Roger L Berger, *Statistical inference*, vol. 2, Duxbury Pacific Grove, CA, 2002.
- [13] Andrea Cassioli, Benjamin Bardiaux, Guillaume Bouvier, Antonio Mucherino, Rafael Alves, Leo Liberti, Michael Nilges, Carlile Lavor, and Thérèse E Malliavin, *An algorithm to enumerate all possible protein conformations verifying a set of distance constraints*, BMC bioinformatics **16** (2015), no. 1, 23.
- [14] Kang Chen and Nico Tjandra, *The use of residual dipolar coupling in studying proteins by NMR*, NMR of Proteins and Small Biomolecules, Springer, 2012, pp. 47–67.
- [15] G Marius Clore, Angela M Gronenborn, and Nico Tjandra, *Direct structure refinement against residual dipolar couplings in the presence of rhombicity of unknown magnitude*, Journal of Magnetic Resonance **131** (1998), no. 1, 159–162.
- [16] Gabriel Cornilescu, John L Marquardt, Marcel Ottiger, and Ad Bax, *Validation of protein structure from anisotropic carbonyl chemical shifts in a dilute liquid crystalline phase*, Journal of the American Chemical Society **120** (1998), no. 27, 6836–6837.
- [17] Mihai Cucuringu, Amit Singer, and David Cowburn, *Eigenvector synchronization, graph rigidity and the molecule problem*, Information and Inference **1** (2012), no. 1, 21–67.
- [18] Yohann De Castro, F Gamboa, Didier Henrion, and J-B Lasserre, *Exact solutions to super resolution on semi-algebraic domains in higher dimensions*, IEEE Transactions on Information Theory **63** (2017), no. 1, 621–630.
- [19] Yichuan Ding, Nathan Krislock, Jiawei Qian, and Henry Wolkowicz, *Sensor network localization, Euclidean distance matrix completions, and graph realization*, Optimization and Engineering **11** (2010), no. 1, 45–66.
- [20] Bruce R Donald, *Algorithms in structural molecular biology*, MIT Press Cambridge, MA., 2011.

- [21] Darius M Gavrila, *The visual analysis of human movement: A survey*, Computer vision and image understanding **73** (1999), no. 1, 82–98.
- [22] Michel X Goemans and David P Williamson, *Improved approximation algorithms for maximum cut and satisfiability problems using semidefinite programming*, Journal of the ACM (JACM) **42** (1995), no. 6, 1115–1145.
- [23] John D Gorman and Alfred O Hero, *Lower bounds for parametric estimation with constraints*, Information Theory, IEEE Transactions on **36** (1990), no. 6, 1285–1301.
- [24] Michael Grant and Stephen Boyd, *CVX: Matlab software for disciplined convex programming, version 2.1*, <http://cvxr.com/cvx>, March 2014.
- [25] Peter Güntert, *Automated NMR structure calculation with CYANA*, Protein NMR Techniques, Springer, 2004, pp. 353–378.
- [26] Peter Güntert, Ch Mumenthaler, and Kurt Wüthrich, *Torsion angle dynamics for NMR structure calculation with the new program DYANA*, Journal of molecular biology **273** (1997), no. 1, 283–298.
- [27] Timothy F Havel, *Distance geometry: Theory, algorithms, and chemical applications*, Encyclopedia of Computational Chemistry **120** (1998), 723–742.
- [28] Didier Henrion and Andrea Garulli, *Positive polynomials in control*, vol. 312, Springer Science & Business Media, 2005.
- [29] Didier Henrion and Jean-Bernard Lasserre, *Detecting global optimality and extracting solutions in GloptiPoly*, Positive polynomials in control, Springer, 2005, pp. 293–310.
- [30] Weidong Hu and Lincong Wang, *Residual dipolar couplings: Measurements and applications to biomolecular studies*, Annual Reports on NMR Spectroscopy **58** (2006), 231–303.
- [31] Bill Jackson, *Notes on the rigidity of graphs*, Levico Conference Notes, vol. 4, Citeseer, 2007.
- [32] Keehyoung Joo, InSuk Joung, Qianyi Cheng, Sung Jong Lee, and Jooyoung Lee, *Contact-assisted protein structure modeling by global optimization in CASP11*, Proteins: Structure, Function, and Bioinformatics (2016).
- [33] Keehyoung Joo, InSuk Joung, Jinhyuk Lee, Jinwoo Lee, Weontae Lee, Bernard Brooks, Sung Jong Lee, and Jooyoung Lee, *Protein structure determination by conformational space annealing using NMR geometric restraints*, Proteins: Structure, Function, and Bioinformatics **83** (2015), no. 12, 2251–2262.
- [34] Scott Kirkpatrick, C Daniel Gelatt, Mario P Vecchi, et al., *Optimization by simulated annealing*, Science **220** (1983), no. 4598, 671–680.

- [35] Georg Kontaxis, Frank Delaglio, and Ad Bax, *Molecular fragment replacement approach to protein structure determination by chemical shift and dipolar homology database mining*, *Methods in Enzymology* **394** (2005), 42–78.
- [36] Kartik Krishnan and Tamás Terlaky, *Interior point and semidefinite approaches in combinatorial optimization*, *Graph theory and combinatorial optimization*, Springer, 2005, pp. 101–157.
- [37] Nathan Krislock, *Semidefinite facial reduction for low-rank euclidean distance matrix completion*, Ph.D. thesis, University of Waterloo, 2010.
- [38] Anil Kumar, RR Ernst, and K Wüthrich, *A two-dimensional nuclear Overhauser enhancement (2d NOE) experiment for the elucidation of complete proton-proton cross-relaxation networks in biological macromolecules*, *Biochemical and biophysical research communications* **95** (1980), no. 1, 1–6.
- [39] Jean B Lasserre, *Global optimization with polynomials and the problem of moments*, *SIAM Journal on Optimization* **11** (2001), no. 3, 796–817.
- [40] Jean Bernard Lasserre, *An introduction to polynomial and semi-algebraic optimization*, no. 52, Cambridge University Press, 2015.
- [41] Ngai-Hang Z Leung and Kim-Chuan Toh, *An SDP-based divide-and-conquer algorithm for large-scale noisy anchor-free graph realization*, *SIAM Journal on Scientific Computing* **31** (2009), no. 6, 4351–4372.
- [42] Fang Li, Alexander Grishaev, Jinfa Ying, and Ad Bax, *Side chain conformational distributions of a small protein derived from model-free analysis of a large set of residual dipolar couplings*, *Journal of the American Chemical Society* **137** (2015), no. 46, 14798–14811.
- [43] Leo Liberti, Carlile Lavor, Nelson Maculan, and Antonio Mucherino, *Euclidean distance geometry and applications*, *SIAM Review* **56** (2014), no. 1, 3–69.
- [44] Rebecca S Lipsitz and Nico Tjandra, *Residual dipolar couplings in NMR structure analysis*, *Annu. Rev. Biophys. Biomol. Struct.* **33** (2004), 387–413.
- [45] Adam Liwo, Jooyoung Lee, Daniel R Ripoll, Jaroslaw Pillardy, and Harold A Scheraga, *Protein structure prediction by global optimization of a potential energy function*, *Proceedings of the National Academy of Sciences* **96** (1999), no. 10, 5482–5485.
- [46] Judit A Losonczi, Michael Andrec, Mark WF Fischer, and James H Prestegard, *Order matrix analysis of residual dipolar couplings using singular value decomposition*, *Journal of Magnetic Resonance* **138** (1999), no. 2, 334–342.
- [47] Simon C Lovell, J Michael Word, Jane S Richardson, and David C Richardson, *The penultimate rotamer library*, *Proteins: Structure, Function, and Bioinformatics* **40** (2000), no. 3, 389–408.

- [48] Fabien Mareuil, Thérèse E Malliavin, Michael Nilges, and Benjamin Bardiaux, *Improved reliability, accuracy and quality in automated NMR structure calculation with ARIA*, Journal of biomolecular NMR **62** (2015), no. 4, 425–438.
- [49] Jorge J Moré and Zhijun Wu, *Distance geometry optimization for protein structures*, Journal of Global Optimization **15** (1999), no. 3, 219–234.
- [50] APS Mosek, *The MOSEK optimization software*, (2010).
- [51] Rishi Mukhopadhyay, Stephanie Irausquin, Christopher Schmidt, and Homayoun Valafar, *DYNAFOLD: A dynamic programming approach to protein backbone structure determination from minimal sets of residual dipolar couplings*, Journal of bioinformatics and computational biology **12** (2014), no. 01, 1450002.
- [52] Jiawang Nie, *Sum of squares method for sensor network localization*, Computational Optimization and Applications **43** (2009), no. 2, 151–179.
- [53] ———, *Optimality conditions and finite convergence of Lasserres hierarchy*, Mathematical programming **146** (2014), no. 1-2, 97–121.
- [54] Pablo A Parrilo, *Semidefinite programming relaxations for semialgebraic problems*, Mathematical Programming **96** (2003), no. 2, 293–320.
- [55] James H Prestegard, David A Agard, Kelley W Moremen, Laura A Lavery, Laura C Morris, and Kari Pederson, *Sparse labeling of proteins: Structural characterization from long range constraints*, Journal of Magnetic Resonance **241** (2014), 32–40.
- [56] GN Ramachandran, CT Ramakrishnan, and V Sasisekharan, *Stereochemistry of polypeptide chain configurations*, Journal of molecular biology **7** (1963), no. 1, 95–99.
- [57] A Saupe and G Englert, *High-resolution nuclear magnetic resonance spectra of orientated molecules*, Physical Review Letters **11** (1963), no. 10, 462.
- [58] James B Saxe, *Embeddability of weighted graphs in k-space is strongly np-hard*, Carnegie-Mellon University, Department of Computer Science, 1980.
- [59] Elena Schmidt and Peter Guntert, *A new algorithm for reliable and general NMR resonance assignment*, Journal of the American Chemical Society **134** (2012), no. 30, 12817–12829.
- [60] Charles D Schwieters, John J Kuszewski, Nico Tjandra, and G Marius Clore, *The Xplor-NIH NMR molecular structure determination package*, Journal of Magnetic Resonance **160** (2003), no. 1, 65–73.
- [61] Parikshit Shah and WISC EDU, *Guaranteed tensor decomposition: A moment approach*, (2015).
- [62] Yang Shen, Frank Delaglio, Gabriel Cornilescu, and Ad Bax, *Talos+: a hybrid method for predicting protein backbone torsion angles from NMR chemical shifts*, Journal of biomolecular NMR **44** (2009), no. 4, 213–223.

- [63] Amit Singer, *Angular synchronization by eigenvectors and semidefinite programming*, Applied and computational harmonic analysis **30** (2011), no. 1, 20–36.
- [64] Anthony Man-Cho So and Yinyu Ye, *Theory of semidefinite programming for sensor network localization*, Mathematical Programming **109** (2007), no. 2-3, 367–384.
- [65] ———, *Theory of semidefinite programming for sensor network localization*, Mathematical Programming **109** (2007), no. 2-3, 367–384.
- [66] Petre Stoica and Boon Chong Ng, *On the Cramér-Rao bound under parametric constraints*, Signal Processing Letters, IEEE **5** (1998), no. 7, 177–179.
- [67] Jos F Sturm, *Using SeDuMi 1.02, a MATLAB toolbox for optimization over symmetric cones*, Optimization methods and software **11** (1999), no. 1-4, 625–653.
- [68] Nico Tjandra and Ad Bax, *Direct measurement of distances and angles in biomolecules by NMR in a dilute liquid crystalline medium*, Science **278** (1997), no. 5340, 1111–1114.
- [69] Kim-Chuan Toh, Michael J Todd, and Reha H Tütüncü, *Sdpt3a matlab software package for semidefinite programming, version 1.3*, Optimization methods and software **11** (1999), no. 1-4, 545–581.
- [70] Joel R Tolman and Ke Ruan, *NMR residual dipolar couplings as probes of biomolecular dynamics*, Chemical Reviews **106** (2006), no. 5, 1720–1736.
- [71] JR Tolman, JM Flanagan, M Am Kennedy, and JH Prestegard, *Nuclear magnetic dipole interactions in field-oriented proteins: information for structure determination in solution*, Proceedings of the National Academy of Sciences **92** (1995), no. 20, 9279–9283.
- [72] Chittaranjan Tripathy, Jianyang Zeng, Pei Zhou, and Bruce Randall Donald, *Protein loop closure using orientational restraints from NMR data*, Proteins: Structure, Function, and Bioinformatics **80** (2012), no. 2, 433–453.
- [73] Senadhi Vijay-Kumar, Charles E Bugg, and William J Cook, *Structure of ubiquitin refined at 1.8Å resolution*, Journal of Molecular Biology **194** (1987), no. 3, 531–544.
- [74] Lincong Wang and Bruce Randall Donald, *Exact solutions for internuclear vectors and backbone dihedral angles from NH residual dipolar couplings in two media, and their application in a systematic search algorithm for determining protein backbone structure*, Journal of Biomolecular NMR **29** (2004), no. 3, 223–242.
- [75] Lincong Wang, Ramgopal R Mettu, and Bruce Randall Donald, *A polynomial-time algorithm for de novo protein backbone structure determination from nuclear magnetic resonance data*, Journal of Computational Biology **13** (2006), no. 7, 1267–1288.



- [76] Zizhuo Wang, Song Zheng, Yinyu Ye, and Stephen Boyd, *Further relaxations of the semidefinite programming approach to sensor network localization*, SIAM Journal on Optimization **19** (2008), no. 2, 655–673.
- [77] Larry Wasserman, *All of statistics: a concise course in statistical inference*, Springer Science & Business Media, 2013.
- [78] Kilian Q Weinberger and Lawrence K Saul, *An introduction to nonlinear dimensionality reduction by maximum variance unfolding*, AAAI, vol. 6, 2006, pp. 1683–1686.
- [79] Walter Whiteley, *Counting out to the flexibility of molecules*, Physical Biology **2** (2005), no. 4, S116.
- [80] Kurt Wüthrich, *NMR studies of structure and function of biological macromolecules (Nobel lecture)*, Angewandte Chemie International Edition **42** (2003), no. 29, 3340–3363.
- [81] Yingqi Xu, Yu Zheng, Jing-Song Fan, and Daiwen Yang, *A new strategy for structure determination of large proteins in solution without deuteration*, Nature methods **3** (2006), no. 11, 931–937.
- [82] Anna Yershova, Chittaranjan Tripathy, Pei Zhou, and Bruce Randall Donald, *Algorithms and analytic solutions using sparse residual dipolar couplings for high-resolution automated protein backbone structure determination by NMR*, Algorithmic Foundations of Robotics IX, Springer, 2011, pp. 355–372.
- [83] Jianyang Zeng, Jeffrey Boyles, Chittaranjan Tripathy, Lincong Wang, Anthony Yan, Pei Zhou, and Bruce Randall Donald, *High-resolution protein structure determination starting with a global fold calculated from exact solutions to the RDC equations*, Journal of biomolecular NMR **45** (2009), no. 3, 265–281.
- [84] Markus Zweckstetter, *NMR: prediction of molecular alignment from structure using the PALES software*, Nature protocols **3** (2008), no. 4, 679–690.