Determination of Structural Ensembles of Flexible Molecules in Solution from NMR data Undergoing Spin Diffusion

Francesca Vasile[1] and Guido Tiana[2*]

[1]Department of Chemistry, Università degli Studi di Milano, I-20133 Milano, Italy

[2] Department of Physics and Center for Complexity and Biosystems, Università degli Studi di Milano and INFN, I-20133 Milano, Italy.

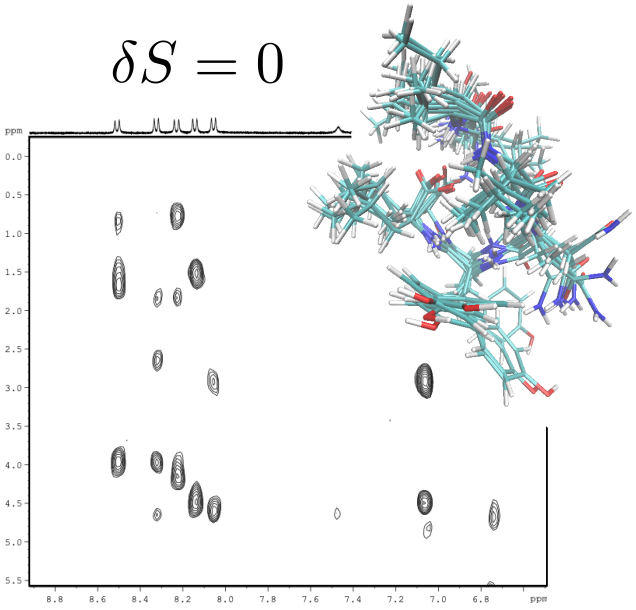
Corresponding Author

*Department of Physics and Center for Complexity and Biosystems, Università degli Studi di Milano and INFN, via Celoria 16, I-20133 Milano, Italy. Tel. +39-0250317221. Email: guido.tiana@unimi.it

ABSTRACT

Spin diffusion is a formidable problem when interpreting NMR data of chemical compounds. We developed a method to reconstruct the conformational ensemble of flexible molecules displaying spin diffusion, which minimizes the subjective bias in the interpretation of experimental data and which can be used routinely to obtain sets of structures with the correct thermodynamic weights. We showed in the case of a flexible molecule that the correct conformational ensemble is quite different from that obtained with standard methods.

TOC GRAPHICS



KEYWORDS

Molecular conformational ensembles, principle of maximum entropy, NMR relaxation matrix

Introduction

The determination of the structure of molecules from NMR data can be obtained by a simple restraint optimization only in the case they are rigid. Organic compounds often fluctuate in an ensemble of conformations displaying different properties and reactivities. For flexible small molecules or macrocycles, restraint optimization produces average structures that are usually not representative of any state populated in solution. To have a realistic picture of the system, one then needs an ensemble of relevant conformations and their correct statistical weights, given by the laws of equilibrium thermodynamics.

Several techniques have been developed to generate such ensembles of conformations compatible with the available NMR data, especially with nuclear Overhauser effect (NOE) intensities and, for large molecules like unstructured proteins, with data obtained by paramagnetic relaxation enhancement. These techniques include selection of pre-generated structures based on the experimental data[1,2], the sample-and-select strategy[3], replica-averaged molecular dynamics simulations[4,5], the on-the-fly[6] and the iterative[7,8] correction of force fields based on the experimental data of a specific system. In particular, the last three methods make explicit use of the principle of maximum entropy[9-11], to guarantee that the algorithm introduce the least amount of arbitrary extra-information that is not contained in the data.

For a large class of molecules, NOE, interpreted in terms of ensemble averages $\langle 1/d^6 \rangle$ of the distances d between the associated pairs of nuclei, is a powerful tool to obtain structural information of molecules and to generate ensembles of conformations. However, especially in the case of hydrophobic or of large molecules[12], the phenomenon of spin diffusion can invalidate this interpretation of NOEs, underestimating the distance between nuclei and resulting in conformations that appear more structured than what they really are[13].

Various methods were designed to keep into account correctly the process of spin diffusion[14-16], but again their purpose was to obtain spatial restraints to be optimized, and thus they are useful for rigid molecules. The goal of the present work is to develop an efficient method to generate ensembles of conformations compatible with NOE data undergoing spin diffusion and complying with the laws of equilibrium thermodynamics, without introducing uncontrolled, subjective assumptions.

The idea is to use NOE intensities to drive a molecular dynamics (MD) simulation to sample the ensemble of equilibrium conformations of the molecule through the principle of maximum entropy[11]. For each simulated conformation, the NOE intensity I_{ij} are from the complete solution of the relaxation equation

$$I_{ij} = \exp[-R_{ik}\tau_m]\, I^0_{kj}\,, \qquad (1)$$

where R_{ik} the multi-spin quantum relaxation matrix, τ_m is the mixing time of the experiment and I^0_{ij} are the diagonal intensities extrapolated at zero mixing time[17]. The map between each simulated conformation and its contribution to the predicted signal is called 'forward model'. The overall predicted signal, regarded as an average over ~10^{23} molecules, is obtained as thermal average $\langle I_{ij} \rangle$ of the forward model of Eq. (1). In this way, one describes correctly the diffusion of magnetization across spins, without relying on the small-τ_m approximation that is usually employed to decouple the relaxation of spin pairs and that results in $<1/d_{ij}^6>$.

We showed that the ensemble of conformations generated taking properly into account spin diffusion are markedly different from those generated in the usual small-τ_m approximation. The algorithm we propose can then be an efficient tool to interpret NMR data.

From the algorithmical point of view, we generated the representative ensemble of conformations with MD simulations starting from a standard Amber force field and modifying it

iteratively to match the experimental NOE within the framework of the principle of maximum entropy, to minimize the degree of arbitrary bias inserted in the model. The iterative MD simulations were performed to obtain an equilibrium-like distribution of conformations, and putative NOE intensities were calculated from these conformations using the full relaxation model of Eq. (1).

Further benefits of the method we propose is that it can keep into account unseen NOEs (uNOEs), namely undetected viable NOEs can be exploited as additional source of information to determine the conformational ensemble of the molecule, and it solves the problem of the ambiguous assignment of overlapping peaks, allowing the use of the overall NOE intensities between two groups of atoms to drive the MD simulation.

The method

The forward model. For any simulated conformation of the equilibrium ensemble of simulated molecules, the NOE intensities I_{ij} are calculated from the time propagation of the diagonal intensities[17] using Eq. (1). The relaxation matrix has the form

$$R_{ij} = \begin{cases} \rho_i & \text{if } i = j \\ \sigma_{ij} & \text{if } i \neq j \end{cases} \quad (2)$$

where

$$\rho_i = K \sum_j \frac{1}{d_{ij}^6} \left[\frac{1}{10} J(0) + \frac{3}{10} J(\omega) + \frac{6}{10} J(2\omega) \right] \quad (3)$$

and

$$\sigma_{ij} = K \frac{1}{d_{ij}^6} \left[-\frac{1}{10} J(0) + \frac{6}{10} J(2\omega) \right], \quad (4)$$

where $d_{ij} = |r_i - r_j|$ is the distance between the ith and the jth proton,

$$K = \frac{1}{2}\left(\frac{\mu_0 \hbar \gamma^2}{4\pi}\right)^2, \quad (5)$$

where μ_0 is magnetic permeability in vacuum, $\hbar$ is the reduced Planck constant and γ is the gyromagnetic ratio of the proton. For our spectrometer, it takes the value $K = 0.56$ nm^6/ms^2, and the spectral density is, under the assumption of isotropic tumbling

$$J(\omega) = \frac{\tau_c}{1+\omega^2\tau_c^2}, \quad (6)$$

where τ_c is the rotational correlation time.

Given an ensemble of conformations $\{r_i\}$ the predicted NOE intensities are calculated as the thermal average of the intensities of the single conformations

$$\langle I_{ij} \rangle = \sum_{\{r\}} \frac{\exp[-U(\{r_i\})/kT]}{Z} \exp[-R_{ik}(\{r_i\})\,\tau_m]\, I^0_{kj}. \quad (7)$$

where U is the energy of the system, T is the temperature, k is the Boltzmann constant, Z is the partition function and the sum is over the conformations of the ensemble. If the ensemble of conformations is the result of a MD simulation at constant temperature T, the conformations of the ensemble $\{r_i^t\}_b$ are automatically weighted with the correct Boltzmann weight, and the predicted NOEs are simply

$$\langle I_{ij} \rangle = \sum_{\{r\}_b} \exp[-R_{ik}(\{r_i^t\})\,\tau_m]\, I^0_{kj}. \quad (8)$$

Operatively, the matrix exponential is calculated exactly with the linear algebra functions of GNU Scientific Library v. 2.4.2, avoiding the standard small-τ_m expansion.

The principle of maximum entropy. The goal of the algorithm is to find a model for the molecule whose equilibrium properties match the experimental data and that is minimally biased, in the sense that the minimum amount of arbitrary hypotheses is used to build the model. This goal can be achieved with the principle of maximum entropy[6,18-20], which states that such optimal model can be found maximizing the entropy of the probability distribution $\mathrm{p}(\mathrm{r})$ of its conformations $\{\mathrm{r}\}$ under the constraints given by the experimental data. Analogously, if one has some prior information on the probability distribution, in the form of a distribution $\mathrm{p_0}(\mathrm{r})$ that is known to approximate $\mathrm{p}(\mathrm{r})$, the fairest model is the one that, while matching the experimental data, minimizes the Kullback-Leibler divergence.

It can be shown (cf. Section S1 in the SI) that if one assumes the system to be at equilibrium, the maximum-entropy model interacts with a potential that displays the same functional form as the forward model $\mathrm{I_{ij}}(\mathrm{r})$,

$$\mathrm{U}(\mathrm{r}) = \mathrm{U_0}(\mathrm{r}) + \sum_{\mathrm{i<j}} \lambda_{\mathrm{ij}}\, \mathrm{I_{ij}}(\mathrm{r}), \qquad (9)$$

where $U_0(r)$ is an approximated potential known a priori and λ_{ij} are the numerical energy parameters that define fully the potential.

The iterative MD scheme. The implementation of the principle of maximum entropy (or, equivalently, of Kullback-Leibler-divergence minimization) is obtained through an iterative MD scheme[7,21]. We start from a standard force field $U_0(r)$ and perform a MD simulation to sample the conformational space of the biomolecule. From the point of view of Eq. (S16), the initial simulation is performed setting all $\lambda_{ij} = 0$.

At the end of the sampling, the values of λ_{ij} are then adjusted to minimize the χ^2 between the NOE intensities calculated from the simulation by Eq. (8) and the experimental ones, defined as

$$\chi^2 = \frac{1}{n(n-1)} \sum_{i<j}^{n} \frac{\left(\langle I_{ij}\rangle - \mathrm{I}_{ij}^{exp}\right)^2}{\sigma^2}, \quad (10)$$

where σ is the experimental error, calculated from a triplicate experiment, and n is the number of hydrogens. Thus, the contribution of each NOE is weighted by the inverse error associated to it. To avoid a lengthy resampling after each step of the χ^2 minimization, we employed the reweighting scheme of ref. [22], which consists in calculating the averages of Eq. (8) after each energy change $U \rightarrow U'$ as

$$\langle I_{ij}\rangle = \frac{Z}{Z'} \sum_{\{r\}_s} I_{ij}(r)\, exp\left[-\frac{U'(r)-U(r)}{kT}\right], \quad (11)$$

where

$$Z' = \sum_{\{r\}_s} \exp\left[-U'(r)/kT\right] \quad (12)$$

and $\{r\}_s$ is the ensemble of conformations sampled with the original potential U.

When the modified potential has become too different from the one used for the sampling, the sampled conformations are no longer representative of the equilibrium state of the new potential and a new sapling is carried out with the new potential. This procedure is iterated till the χ^2 converges to a low value. The target of the iteration is to reach $\chi^2 = 1$, that is a model such that the difference between calculated and experimental NOE is comparable with the experimental error, avoiding an overfit. Since the experimenta errors σ weight the different terms of the χ^2, performing the NOESY experiment in triplicate is a necessary step of the present procedure.

The computational details are given in Sect. S2 of the SI. The averages are calculated with the relaxation model of Eq. (1); it is important to stress that this reweighting scheme is only compatible with an approximated calculation of the spectral density as that defined by Eq. (6). However, in the calculation we correct the two-body dispersion terms and the Ryckaert-Bellemans torsional terms of the potential, regarded as approximations of the many-body potential suggested by the

forward model (cf. Eq. (9)), and consequently we are satisfying only an approximated version of the principle of maximum entropy. The reason for this choice is that most NOEs are associated with pairs of atoms of type 1-4 (i.e., separated by 3 covalent bonds). In usual force fields, the relative positions of these atoms is controlled essentially from the torsional terms of the potential, rather than from Lennard-Jones terms. Thus, since we are not satisfying exactly the principle of maximum entropy, we cannot guarantee the unicity of the optimized parameters, and their dependence on the specific optimization run has to be checked a posteriori.

The back-calculated NOEs can be compared individually with the experimental ones in Eq. (10), or in the case that the multiple experimental peaks cannot be distinguished, can be summed together and compared with the overall height of the experimental peak, to avoid an arbitrary division of the peak into (usually equal) contributions. Moreover, one can extend the sum in Eq. (10) to unobserved NOEs (uNOEs), that is to bias pairs of protons that surely do not display a detectable intensity to display a $\langle I_{ij} \rangle$ equal to 0.

Calculation of the rotational correlation time. The calculation of the forward model (1) requires the knowledge of the rotational correlation time τ_c, under the assumption of isotropic tumbling. For this purpose, we performed both a NOESY and a ROESY experiment on the molecule with a short mixing time $\tau_m = 50$ ns. In this way, not all the peaks are detectable, but we are in a regime in which spin diffusion is negligible. Thus,

$$\mathrm{I}_{ij}^{NOE} = exp[-R_{ik}\tau_m]\, I_{kj}^0 \approx -R_{ik}\tau_m\, I_{kj}^0 \tag{13}$$

and

$$\mathrm{I}_{ij}^{ROE} = exp\left[-R_{ik}^{ROE}\tau_m\right] I_{kj}^0 \approx -R_{ik}^{ROE}\tau_m\, I_{kj}^0. \tag{14}$$

In the NOESY experiment, the off-diagonal relaxation rates are those of Eq. (4), while for the ROESY experiment they are[17]

$$\sigma_{ij}^{ROESY} = K\frac{1}{{d_{ij}}^6}\left[\frac{2}{10}J(0) + \frac{3}{10}J(\omega)\right]. \quad (15)$$

The ratio between the NOE and the ROE intensities can be obtained from Eqs. (13-15) and (4), and reads

$$\mathrm{r} = \frac{\mathrm{I}_{ij}^{NOE}}{\mathrm{I}_{ij}^{ROE}} = \frac{(1+\omega^2\tau_c^2)(5-4\omega^2\tau_c^2)}{(1+4\omega^2\tau_c^2)(5+2\omega^2\tau_c^2)}, \quad (16)$$

that is independent of the diagonal elements I_{ij}^0 and of the averages $\langle d^{-6}\rangle$ which characterize the conformational ensemble of the molecule. From the measured values of r one can obtain the relaxation times as

$$\tau_c = \frac{1}{\omega}\left(\frac{3\left(36r^2+4r+9\right)^{1/2} - 22r+1}{8(1+2r)}\right)^{1/2}. \quad (17)$$

Results

We applied the computational strategy described above to the calculation of the ensemble of equilibrium conformations of the peptide LIVNYL[23], as an example of flexible molecule, populating at equilibrium multiple conformations. A set of NOESY experiments are carried out at different mixing times τ_M as described in Sect. S3 of the SI. Each experiment is carried out in triplicate, to be able to assign to each signal a standard error.

As reported in Fig. 1, the NOE intensities of several hydrogen pairs are not linear functions of the mixing time, suggesting that the system displays spin diffusion beyond mixing times of 400 ms. For guiding the simulation, we used the spectrum acquired at 700 ms, because it allows the assignment and the measurement of the intensities of the largest set of cross peaks (see Table S1 in the SI).

While the standard procedure of turning NOE intensities into spatial restraints and minimizing them [24] gives a quite homogeneous ensemble of conformations, with an average root mean square

deviation (RMSD) from each other of 0.31 nm, there is no theoretical reason to believe that this ensemble represents the thermodynamic fluctuations of the peptide at equilibrium. Moreover, it is not able to reproduce the experimental data unless a treatment of quantum relaxation involving the time-propagation of magnetization, as in Eq. (1), is employed (cf. Fig. S2 in the SI). In fact, this calculation gives χ^2=73.2, indicating that the NOEs back-calculated from the minimized conformations keeping into account spin diffusion are different from the experimental values by very many standard errors.

A different strategy, that of performing unbiased MD simulations with standard force fields in water cannot reproduce to a quantitative grade the experimental intensities (cf. gray bars in Fig. 2) as well, resulting in a χ^2=56.3 between the simulated and the experimental NOEs.

To improve the results, we have applied the iterative MD approach, employing Eq. (1) to back-calculate the NOEs from the simulation, as described in the Methods. We performed some tens of iterations of the algorithm; in each iteration a 50ns replica-exchange simulation is carried out in implicit solvent, and afterwards the potential is adjusted to minimize the χ^2 between the back-calculated and the experimental NOEs. The trend of the χ^2 along the iterations is reported in Fig. S3 of the SI and, despite strong oscillations, decreases from ≈30 for the initial Amber force field to a minimum of 3.7. The comparison between the calculated and the experimental NOEs is reported in Fig. 2 (cf. blue and green bars). Of the 21 available NOEs, in 18 cases are within the error bars of the experimental values and in 3 cases they are just outside.

A question which requires investigation is the robustness of the results. We performed two independent iterations, both starting from the Amber force field, and reaching a χ^2 of 3.7 and 4.0, respectively. In both cases the agreement between the calculated and the experimental NOEs is good, but the sets of parameters optimized in the independent runs give poorly correlated results

(cf. Fig. S4 in the SI). This could be caused by the fact that the optimization is not fully at convergence in such a large parameter space, by the approximation of a Lorentzian spectral density, or because the principle of maximum entropy guarantees the unicity of the result if only the parameters of the potential scaling as $1/d^6$ like the forward model (see Sect. 2 in the SI), while we had to update also the torsional terms to allow the i-(i+4) NOEs to match the experimental data. Anyway, the conformational properties (discussed below) appear similar for the two sets, suggesting that they correspond two different ways of decomposing the potential between Lennard-Jones and torsional terms.

With the present algorithm is possible to include uNOEs in the simulation, that is two bias pairs of hydrogens to display zero intensity. Consequently, uNOEs are treated by the algorithm as any other NOE, setting their target intensity to zero. We implemented the 780 uNOEs corresponding to all the peaks not observed in the experimental spectrum, with a standard error set to the conventional value of 1000, corresponding to the order of magnitude of the error bars of observed NOEs. The χ^2 calculated only on the observed NOEs slightly increases to 4.1 from 3.7. The results are reported in Fig. S5 of the SI. The number of erroneously observed uNOE remains around 10, and their maximum intensity decreases from 6700 to 2900. The effect of uNOEs in the optimization of this molecules seems then marginal.

One should pay attention that, especially in the case of molecules more complex that this small peptide, there can be many reasons not related with the associated interatomic distance why a NOE is not observed, like e.g. a high exchange rate with the solvent. In the present case, we showed that the use of uNOE essentially does not affect the results at all; thus, including pairs that are absent from the spectrum for these distance-independent reasons has no consequences. For other molecules the situation could be different and one should be careful in selecting uNOEs.

From the optimized simulation one can study the equilibrium ensemble of conformations of the molecule. A cluster analysis based on the RMSD highlights three major clusters of conformations displaying different equilibrium probabilities and mutually exclusive interactions (cf. Fig. 3). The most populated cluster (45% of the conformations) displays a hydrogen bond between the OC of 6LEU and the HN of 3VAL and its aromatic ring is exposed; the least-populated (14%) one displays a hydrogen bond between the O of 1LEU and the HN of 4ASN; the middle one (31%) has no hydrogen bonds and the aromatic rings packs against the other hydrophobic residues.

The distribution of radii of gyration of the molecule is less broadened than that obtained from a simulation run with the uncorrected force field, as displayed in the lower panel of Fig. 3. The two independent simulations give results that are comparable with each other.

We also performed a maximum-entropy optimization disregarding spin diffusion and using as forward model the conformational average $<1/d^6>$ for each pair of hydrogens, as in refs. [7,8]. The agreement with the experimental data is worse than that obtained modelling spin diffusion (χ^2=16.4, see Fig. S6 in the SI). Not unexpectedly, in this case the equilibrium conformations are much more compact and structured (see cyan curve in Fig. 3), similarly to what is known to happen when interpreting NOEs as distance restrains in presence of spin diffusion[25].

From the calculation one can also calculate the free energy of the system as a function of any variable of interest. Based on the cluster analysis described above, we used the distance between the OC of 6LEU and the HN of 3VAL, that make a hydrogen bond in a fraction of the sampled conformations, and the radius of gyration. The result, displayed in Fig. 4, gives a perspective which is different from that of the cluster analysis. There are three thermodynamic states, two of them displaying low values of the distance between the hydrogen and the oxygen, and thus making the hydrogen bond that characterizes the most populated cluster, but with a different degree of

compactness. In the other state this hydrogen bond is not formed, and the radius of gyration is intermediate between those of the other two states. This latter state then seems to contain the two least-populated clusters discussed above.

This richness of structure is quite different from the case of the peptide optimized without taking into account spin diffusion (cf. Fig. S7 in the SI). In this case, there is essentially one state that is compact but does not display the hydrogen bonds between 3VAL and 6LEU.

Discussion and Conclusions

The principle of maximum entropy is implemented through an iterative MD scheme, that starts from a tentative force field in implicit solvent and corrects it so that the NOE spectrum calculated from the simulated ensemble through the forward model that accounts for spin diffusion and through a thermodynamic reweighting[22] matches the experimental one within the experimental error bars. Error bars are calculated as standard deviation of NOE intensities in a triplicate experiment. This scheme[7,11,21] has the advantage with respect to other algorithms[4,6,9,10,20] that still implement the principle of maximum entropy not to require the calculation of the forward model at each MD step, a calculation that would be lengthy in the case of the matrix operations of Eq. (1).

The present approach is applied to an equilibrium ensemble of conformations, that displays relevant conformational fluctuations, and provides results that are compatible with the laws of equilibrium thermodynamics. The forward model is much more realistic than the standard $<1/d_{ij}^6>$ model at long mixing times, giving results that are quite different than those obtained with that model. The main assumption in this calculation is that of isotropic tumbling, whose correlation time is obtained comparing NOE with ROE intensities.

The possibility of comparing the summed contribution of different atom pairs with the height of a peak given by the superposition of different overlapping NOEs avoids the arbitrary splitting of such peaks in its contributions. The algorithm also allows one to take into account unobserved NOEs, although we showed that, at least with the molecule under study here, the results are rather insensitive to this further information.

The main approximation used in the present approach is that of using the expression of a rigid molecule which tumbles isotropically to calculate the spectral density in Eq. (6). This approximation allowed us to perform replica-exchange simulation instead of fixed-temperature MD simulations that would be required to calculate explicitly the spectral density.

To check the validity of this approximation, we performed two plain-MD simulations at constant temperature for 10 μs from two different initial conformations with the (same) optimized potential, and back-calculated the correlation functions responsible for the NOEs, as described in ref. 26 (for all the details see Sect. S5 of the SI). The tumbling of the molecule is found to be approximately isotropic, with a correlation time which is independent on the initial conformation; the overall correlation function C_{all}, whose Fourier transform gives the spectral density $J(\omega)$ that controls the transition rates of Eqs. (3-4), is approximately single-exponential, suggesting that the flexibility of the molecule does not necessarily invalidate Eq. (6); however, curves obtained from simulations starting from different initial conformations gives different results. This fact suggests that even long fixed-temperature simulations are not able to explore all the relevant conformational space available to this peptide. Consequently, the replica-exchange strategy is really necessary to guarantee an equilibrium sampling, even if this computational technique does not give realistic kinetic trajectories, preventing the calculation of the time correlation functions and obliging one to resort to the approximated Eq. (6).

Nonetheless, the NOE intensities calculated from the plain-MD simulation with and without the rigid-body approximation of the spectral density give comparable results, within the experimental error. These intensities are different from the experimental ones because the simulated rotational correlation time is smaller than the experimental one, probably due to the implicit character of the solvent used in the simulation (cf. Sect. S5 in the SI). These results, although stressing again that plain MD simulations cannot be used to calculated the NOEs for this molecule, suggest that the approximation of using a rigid-body spectral density for a system that is definitely not rigid is not dramatic if the correct time parameter is used.

Another approximation of the present treatment is that spin transfer and diffusion to the solvent is neglected. However, accounting for it would require both an explicit-solvent simulation and a much heavier forward model, features that would make the computational cost of the optimization extremely high.

Thus, the present strategy of modelling flexible molecules from NOEs undergoing spin diffusion is a fair trade-off between correctness of the model and computational cost. It allows one to go beyond the standard algorithms based on restraint minimizations, which can be misleading for flexible molecules, and to keep into account spin diffusion, whose neglection largely overestimate the degree of structure of the molecule.

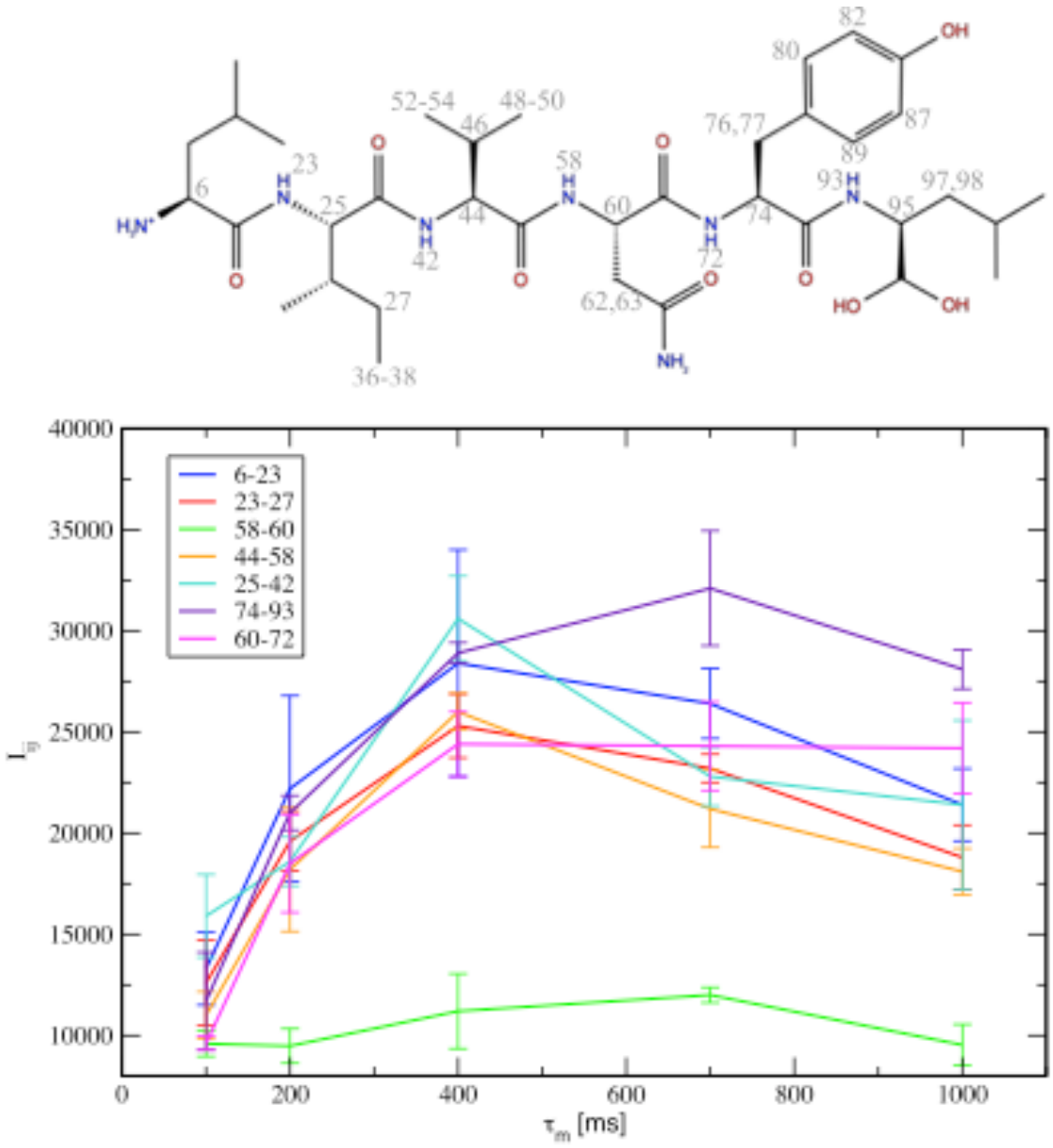


Figure 1: The structure of the peptide under study, where the numeration of the hydrogen atoms associated with the detected NOEs is indicated (above panel). The intensity of some NOEs, as a function of the mixing time τ_M (below panel).

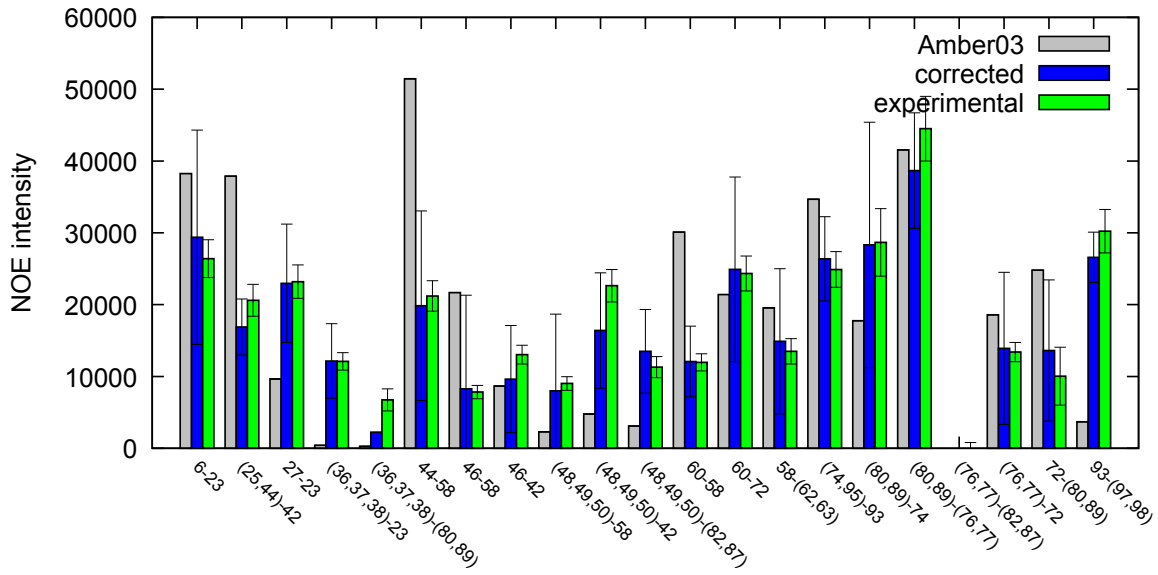


Figure 2: The NOE intensities obtained applying Eq. (1) to MD simulation in explicit water with the Amber03 force field, those obtained with the maximum-entropy correction and the experimental ones. The error bars in the experimental bars indicate the standard deviation of the triplicate experiment; those in the simulated bars indicate their fluctuations over the conformational ensemble.

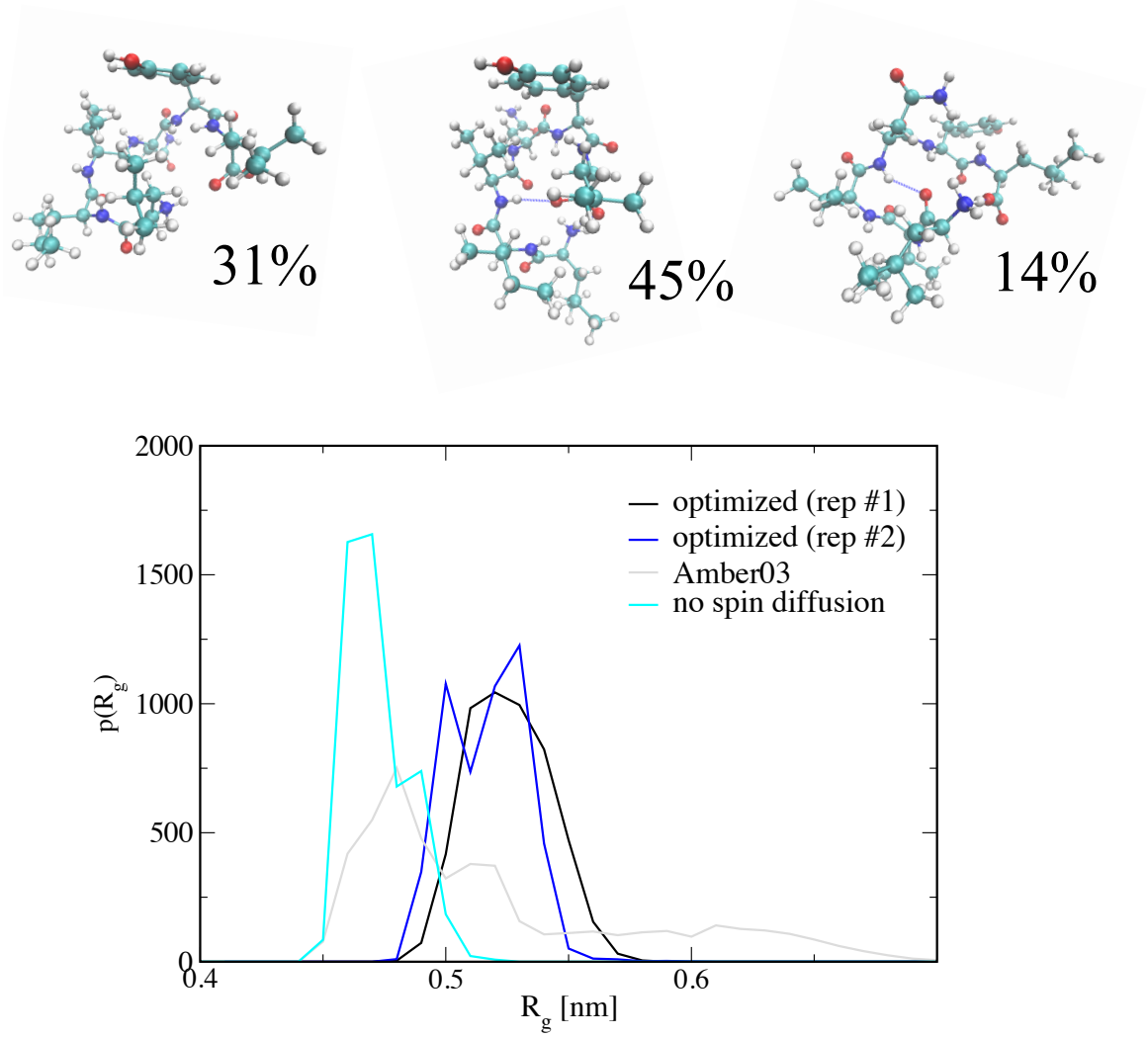


Figure 3: The representative structures of the three more populated clusters of molecule conformations with their equilibrium probabilities (above). The distribution of radius of gyration obtained from two independent optimizations, from the original Amber03 potential and for the optimization performed disregarding spin diffusion (below).

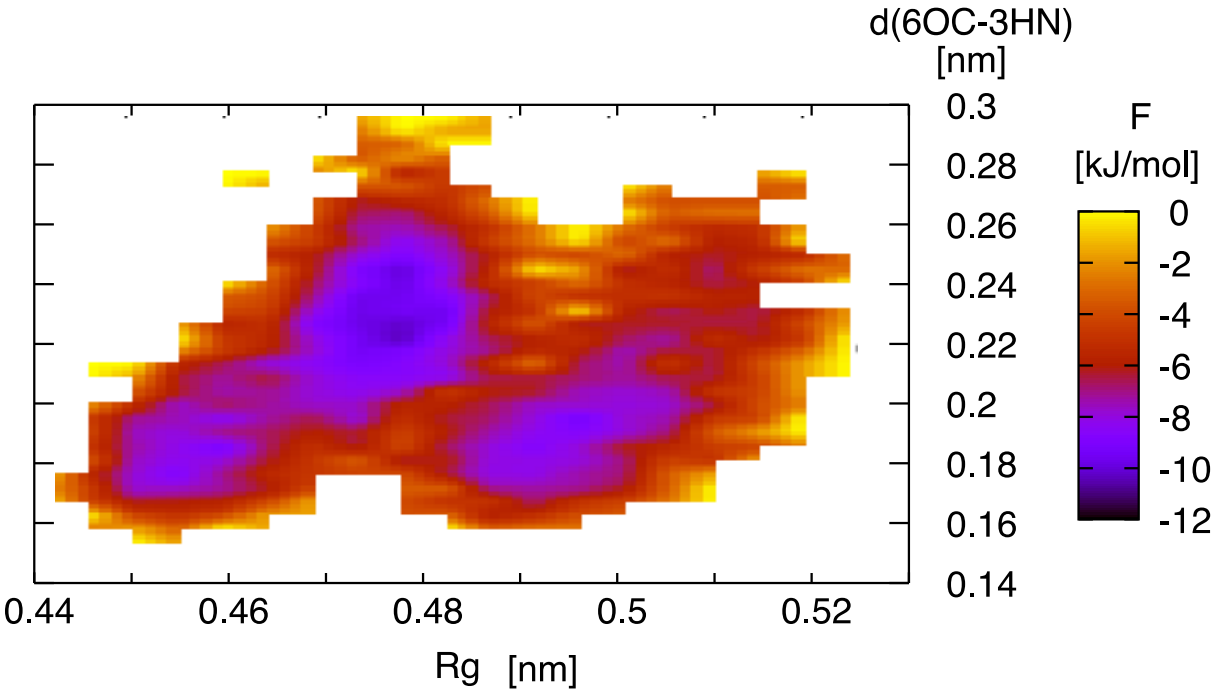


Figure 4: The free energy of the peptide, calculated as a function of the distance between the OC of 6LEU and the HN of 3VAL and the gyration radius.

ASSOCIATED CONTENT

Supporting Information. Supporting information contains the detailed calculations, the technical description of the algorithm and supporting figures.

The following files are available free of charge.

SuppInfo.pdf

AUTHOR INFORMATION

The authors declare no competing financial interests.

REFERENCES

(1) Salmon, L.; Nodet, G.; Ozenne, V.; Yin, G.; Jensen, M. R.; Zweckstetter, M.; Blackledge, M. NMR Characterization of Long-Range Order in Intrinsically Disordered Proteins. J Am Chem Soc 2010, 132, 8407–8418.

(2) Ozenne, V.; Bauer, F.; Salmon, L.; Huang, J.-R.; Jensen, M. R.; Segard, S.; Bernadó, P.; Charavay, C.; Blackledge, M. Flexible-Meccano: a Tool for the Generation of Explicit Ensemble Descriptions of Intrinsically Disordered Proteins and Their Associated Experimental Observables. Bioinformatics 2012, 28, 1463–1470.

(3) Chen, Y.; Campbell, S. L.; Dokholyan, N. V. Deciphering Protein Dynamics From NMR Data Using Explicit Structure Sampling and Selection. Biophys J 2007, 93, 2300–2306.

(4) Best, R. B.; Vendruscolo, M. Determination of Protein Structures Consistent with NMR Order Parameters. J Am Chem Soc 2004, 126, 8090–8091.

(5) Camilloni, C.; Vendruscolo, M. Statistical Mechanics of the Denatured State of a Protein Using Replica-Averaged Metadynamics. J Am Chem Soc 2014, 136, 8982–8991.

(6) Cesari, A.; Gil-Ley, A.; Bussi, G. Combining Simulations and Solution Experiments as a Paradigm for RNA Force Field Refinement. J. Chem. Theo. Comp. 2016, 12, 6192–6200.

(7) Vasile, F.; Civera, M.; Belvisi, L.; Potenza, D.; Tiana, G. Thermodynamically-Weighted Conformational Ensemble of Cyclic RGD Peptidomimetics From NOE Data. J Phys Chem B 2016, 120, 7098–7107.

(8) Vasile, F.; Panigada, M.; Siccardi, A.; Potenza, D.; Tiana, G. A Combined NMR-Computational Study of the Interaction Between Influenza Virus Hemagglutinin and Sialic Derivatives From Human and Avian Receptors on the Surface of Transfected Cells. IJMS 2018, 19, 1267–13.

(9) Cavalli, A.; Camilloni, C.; Vendruscolo, M. Molecular Dynamics Simulations with Replica-Averaged Structural Restraints Generate Structural Ensembles According to the Maximum Entropy Principle. J. Chem. Phys. 2013, 138, 094112.

(10) Roux, B.; Weare, J. On the Statistical Equivalence of Restrained-Ensemble Simulations with the Maximum Entropy Method. J. Chem. Phys. 2013, 138, 084107–084109.

(11) Tiana, G.; Giorgetti, L. Integrating Experiment, Theory and Simulation to Determine the Structure and Dynamics of Mammalian Chromosomes. Curr. Opin. Struct. Biol. 2018, 49, 11–17.

(12) Cavanagh, J.; Fairbrother, W. J.; Palmer, A. G.; Rance, M.; Skelton, N. J. Protein NMR Spectroscopy; Academic Press, 2007.

(13) Keepers, J. W.; James, T. L. A Theoretical Study of Distance Determinations From NMR. Two-Dimensional Nuclear Overhauser Effect Spectra. J. Magn. Res. 1984, 57, 404–426.
(14) Olejniczak, E. T.; Gampe, R. T., Jr; Fesik, S. W. Accounting for Spin Diffusion in the Analysis of 2D NOE Data. J. Magn. Res. 1986, 67, 28–41.
(15) Linge, J. P.; Habeck, M.; Rieping, W.; Nilges, M. Correction of Spin Diffusion During Iterative Automated NOE Assignment. J. Magn. Res. 2004, 167, 334–342.
(16) Orts, J.; Vögeli, B.; Riek, R. Relaxation Matrix Analysis of Spin Diffusion for the NMR Structure Calculation with eNOEs. J. Chem. Theo. Comp. 2012, 8, 3483–3492.
(17) Neuhaus, D.; WIlliamson, M. P. The Nuclear Overhauser Effect in Structural and Conformational Analysis; Wiley, 2000.
(18) Jaynes, E. T. Information Theory and Statistical Mechanics. Physical Review 1957, 106, 620–630.
(19) Pitera, J. W.; Chodera, J. D. On the Use of Experimental Observations to Bias Simulated Ensembles. J. Chem. Theo. Comp. 2012, 8, 3445–3451.
(20) White, A. D.; Voth, G. A. Efficient and Minimal Method to Bias Molecular Simulations with Experimental Data. J. Chem. Theo. Comp. 2014, 10, 3023–3030.
(21) Giorgetti, L.; Galupa, R.; Nora, E. P.; Piolot, T.; Lam, F.; Dekker, J.; Tiana, G.; Heard, E. Predictive Polymer Modeling Reveals Coupled Fluctuations in Chromosome Conformation and Transcription. Cell 2014, 157, 950–963.
(22) Norgaard, A. B.; Ferkinghoff-Borg, J.; Lindorff-Larsen, K. Experimental Parameterization of an Energy Function for the Simulation of Unfolded Proteins. Biophys J 2008, 94, 182–192.
(23) Vasile, F.; Rossi, D.; Collina, S.; Potenza, D. Diffusion-Ordered Spectroscopy and Saturation Transfer Difference NMR Spectroscopy Studies of Selective Interactions Between ELAV Protein Fragments and an mRNA Target. Eur. J. Org. Chem. 2014, 2014, 6399–6404.
(24) Güntert, P.; Mumenthaler, C.; Wuthrich, K. Torsion Angle Dynamics for NMR Structure Calculation with the New Program DYANA. J Mol Biol 1997, 273, 283–298.
(25) Kominos, D.; Suri, A. K.; Kitchen, D. B.; Bassolino, D.; Levy, R. M. Simulating the Effect of the Two-Spin Approximation on the Generation of Protein Structures From NOE Data. J. Magn. Res. 1992, 97, 398–410.
(26) Peter, C.; Daura, X.; van Gunsteren, W. F. Calculation of NMR-Relaxation Parameters for Flexible Molecules From Molecular Dynamics Simulations. J Biomol NMR 2001, 20, 297–310.

Supporting Information

Determination of Structural Ensembles of Flexible Molecules in Solution from NMR Data Undergoing Spin Diffusion

Francesca Vasile[1] and Guido Tiana[2*]

[1]Department of Chemistry, Università degli Studi di Milano, via Golgi 19, 20133 Milano, Italy

[2]Center for Complexity and Biosystems and Department of Physics, Università degli Studi di Milano and INFN, via Celoria 16, 20133 Milano, Italy.

*guido.tiana@unimi.it

S1. Obtaining a force field from the principle of maximum entropy

If one already has a *a priori* model for the system which provides a probability distribution $p_0(r)$, a way of defining the minimum-biased model is that of minimizing the Kullback-Leibler divergence between the two distributions

$$D_{KL}[p|p_0] = \sum_{\{r\}} p(r) \log \frac{p(r)}{p_0(r)} \quad . \tag{S1}$$

that can be constrained so that the averages performed with the model $\langle I_{ij} \rangle = \sum_{\{r\}} I_{ij}(r) p(r)$ match the experimental data I_{ij}^{exp} with the method of Lagrange multipliers. One has then to minimize the function

$$\sum_{\{r\}} p(r) \log \frac{p(r)}{p_0(r)} - \sum_{ij} \lambda_{ij} \left(\sum_{\{r\}} p(r) I_{ij}(r) - \mathrm{I}_{ij}^{exp} \right) \tag{S2}$$

obtaining

$$p(r) = p_0(r) \frac{exp[-\sum_{ij} \lambda_{ij} I_{ij}(r)]}{Z'}, \tag{S3}$$

where *Z'* is the normalizing partition function. If p and p_0 are equilibrium distributions following Boltzmann statistics

$$p(r) = \frac{1}{Z} exp\left[-\frac{U(r)}{kT} \right] \tag{S4}$$

and

$$p_0(r) = \frac{1}{Z_0} exp\left[-\frac{U_0(r)}{kT} \right], \tag{S5}$$

with $Z = \sum_{\{r\}} \exp\left[-U(r)/kT\right]$, then Eq. (S11) becomes

$$U(r) = U_0(r) + kT \sum_{ij} \lambda_{ij} I_{ij}(r). \tag{S6}$$

Consequently, one can correct any potential U_0 to match the experimental data in a minimally-biased way, with a correction that has the same functional dependence on the conformation of the system as that of the forward model. Here two problems arise. First, the correcting potential obtained substituting Eq. (1) in Eq. (S6) is not two-body, but the interaction between pairs of atoms depend on the position of all other atoms, like in polarizable force fields. Moreover, the relation

$$\frac{\partial \log Z}{\partial \lambda_{ij}} = \mathrm{I}_{ij}^{exp} \tag{S7}$$

that relates the Lagrange multipliers, which now act as parameters of the potential, to the experimental NOEs is an implicit equation involving the partition function, and thus is of little practical use.

The former problem is solved assuming a two-body functional form, as usually done by standard force fields (but still calculating the forward model, as described in the next Section, with the many-body Eq. (1)), that is using

$$U(r) = U_0(r) + \sum_{ij} \frac{\lambda'_{ij}}{(r_i - r_j)^6}, \tag{S8}$$

where λ'_{ij} includes the Lagrange multipliers, the thermal contribution kT and the contribution of the forward model, assumed constant. The numerical values of λ'_{ij} are calculated minimizing the χ^2 between the NOE intensities calculated from the simulation and the experimental ones, as described in the next Section, thus avoiding the use of Eq. (S7). The use of a correction to the potential that has the same functional form of the forward model guarantees that we are searching among parametrizations that minimize the Kullback-Leibler divergence; the minimization of the χ^2guarantees that the Lagrange multipliers are satisfying the constraints on the thermodynamic averages. Since the expression in Eq. (S2) is convex, it displays a unique solution; thus, if one is able

to find a distribution $p(r)$ that minimizes the Kullback-Leibler divergence and matches the experimental averages, this will be the only possible solution.

S2. Computational details

The initial potential $U_0(r)$ is the Amber 03 force field in implicit solvent, modelled with GBSA. At each iteration, replica-exchange MD simulations at four temperatures (T=300K, 330K, 370K and 420K) are performed to sample the conformational space for 50 ns each replica, recording 5000 conformations at 300K. Calculations are carried out with a tailor-made code calling Gromacs 4.5.5 for the replica-exchange part.
At the end of the sampling, the parameters of the interactions between the pairs of hydrogen atoms and of heavy atoms bound to the hydrogen atoms and the involved in the NOE signals undergo 500 random updates, accepting only the changes that decrease the χ^2 between the calculated and the experimental NOE intensities. The same kind of update is applied to the parameters of the Ryckaert-Bellemans torsional potential associated with the observed i-(i+4) NOE, that were added to the potential with all parameters set to zero at the beginning of the simulation. A new MD simulation is then started with the new potential, starting from the last conformation of the previous run.
The simulation is explicit solvent used as comparison are performed in 990 TIP/3P water molecules. The code is freely available at https://github.com/guidotiana/ffoptim .

S3. NOESY experiments

All NMR spectra were registered on Bruker Avance III 400 MHz using a solution 3.5 mM of the peptide in water (with 10% D2O). The water suppression war carried out by excitation sculpting. The assignment was performed through one- and two-dimensional ^{1}H-NMR spectra by standard method. The assignment of the molecule is reported in Table S3. For the conformational analysis three independent NOESY spectra (with 32 scans and 256 increments) were collected using a mixing time of 700 ms (Figure S1) and the intensities of cross peaks were measured. To obtain the NOE build up curves, NOESY spectra with 100, 200, 400 700 and 1000 ms were used (Figure 1). The cross-peaks intensities were calculated and plotted as function of mixing time. Diagonal intensities are extrapolated at zero mixing time with a least-square cubic fit.

S4. Calculation of the rotational correlation time

In Table S1 we report the intensities obtained for six crosspeaks of the peptide. The value of τ_c, although being of the same order of magnitude, is different for each pair, because it is affected not only by the rotational motion of the molecule as a whole, but also the internal motion of the single groups, since the molecule is flexible. Not being able to find the value of τ_c for each pair, because not all crosspeaks can be detected at such a low mixing time, we used for the forward model the average of the values reported in Table S2, that is $\tau_c = 135$ ps.

S5. Approximation of the spectral density function

The main approximation used in the present approach is that of using the expression of a rigid molecule which tumbles isotropically to calculate the spectral density in Eq. (6). This approximation allowed us to perform replica-exchange simulation instead of fixed-temperature MD simulations that would be required to calculate explicitly the spectral density.

To check the validity of this approximation, we performed two independent 10 μs simulations from two different initial conformations with the (same) optimized potential, and back-calculated the correlation functions responsible for the NOEs, as described in ref. 26.

The first result is that the tumbling of the molecule is found to be approximately isotropic, with a correlation time which is independent on the initial conformation and has the same order of magnitude of the experimental one (see Fig. S8 in the SI). The rotational correlation time $\tau_c \approx 25$ ps is smaller than those obtained experimentally (cf. Sect. S4 above). This is not completely unexpected due to the fact that simulations are done in implicit solvent.

Moreover, we calculated the NOE intensity without approximating the spectral density as in Eq. (6) but calculating it explicitly from

$$C_{all}(\tau) = \langle \frac{P_2(\cos \alpha_{t,t+\tau})}{r^3(t) r^3(t+\tau)} \rangle, \qquad \text{(S9)}$$

Where $P_2(x) = \frac{3}{2}x^2 - \frac{1}{2}$ is the second-order Legendre polynomial and $\alpha_{t,t+\tau}$ is the angle between the inter-spin vectors at the two times. The spectral function J(ω) is the calculated as Fourier transform of Eq. (S9). It is important to note that this approach requires a real, fixed-temperature MD simulation to calculate the time correlation functions, while cannot be used with trajectories obtained from a replica-exchange simulation, as those used in the main part of our work.

The correlation functions $C_{all}(\tau)$ for the two contacts 60-74 and 48-82, taken respectively as representative of short-range and long-range contacts, are displayed in Fig. S9. Also, the value of the NOEs calculated by Eq. (S9) is indicated in the figure. The overall correlation function $C_{all}(\tau)$ (cf. Fig. S9 in the SI), whose Fourier transform gives the spectral density J(ω) that controls the transition rates of Eqs. (3-4), is approximately single-exponential, suggesting that the flexibility of the molecule does not necessarily invalidate Eq. (6)

The two correlation functions calculated for each contact in the two independent simulations appear quite different from each other. This difference can be better appreciated noticing that the associated NOE intensities (estimated in the small-τ_m limit, neglecting spin diffusion) are markedly different from each other, more than typical error bars. This difference suggests that plain MD simulations at fixed temperature, at variance with replica-exchange simulations, cannot reach thermodynamic equilibrium and thus are affected by a strong dependence on the initial conditions. In other words, although the calculation of NOEs from a plain MD simulation using Eq. (S9) is in principle more correct than using the approximation of Eq. (6) with a replica-exchange simulation, in practice a plain MD simulation is not able to calculate the thermodynamic average that appears in Eq. (S9).

Nonetheless, we tested the approximation of the spectral density on the plain-MD data, at least to verify internal consistency. For this purpose, we postulated that the plain-MD trajectories described the thermodynamic equilibrium of the system (something which is clearly not true, see above), and calculated the NOE intensities both with the true spectral density of Eq. (S9) and with the approximated one of Eq. (6), using in the latter case the rotational correlation time obtained from Fig. S8. The calculated values are listed in Table S4. The difference between the NOE intensities calculated from the exact and the approximated spectral densities is of the order of 10%, comparable with the experimental error bars.

The NOEs calculated in this way are approximately one order of magnitude smaller than the experimental ones. The reason for that lies in the low value of the simulated rotational correlation time. In fact, calculating the NOEs from the same plain-MD simulations with the spectral density of Eq. (6), but using now the experimental value of the rotational correlation time gives values (cf. Table S4) that display the same order of magnitude of the experimental ones (still being different in the two independent simulations and different from the experimental values, for the reason explained above).

Supporting Tables

atom *i*	atom *j*	average I_{ij}	σ_{ij}
6	23	26400	1734
25,44	42	20600	2224
27	23	23200	692
36-38	23	12100	1209
36-38	80,89	6732	1536
44	58	21200	1907
46	58	7833	929
46	42	13033	1069
48-50	58	9021	944
48-50	42	22633	2253
48-50	82,87	11298	1476
60	58	11966	379
60	72	24333	2190
58	62,63	13500	1768
74,95	93	24900	1650
80,89	74	28666	4700
80,89	76,77	44500	4500
76,77	82,87	0	1000
76,77	72	13400	2300
72	80,89	10046	4030
93	97,98	30232	2400

Table S1: The average and standard deviation over three replicate experiments of the NOE intensities recorded at a mixing time of 700 ms. The overlapping spins are indicated in an aggregated way.

Atom *i*	Id	Atom *j*	Id	I_{ij}^{NOE}	I_{ij}^{ROE}	r	τ_c *[ps]*
HA1	6	HN2	23	8100	29764	0.27	181
HA3	44	HN4	58	5566	29814	0.18	212
HN3	42	HB3	46	5845	8253	0.7	83
HN2	23	HB2	27	8261	12984	0.63	96
HB5	76,77	HB7	97,98	12264	30044	0.41	143
HN6	93	HB7	97,98	8261	12984	0.63	96

Table S2: The crosspeak intensities for selected pairs in the NOESY and in the ROESY experiments at 50 ns. Their ratio *r* allows one to calculate the rotational time τ_c through Eq. (S24).

	NH	***Hα***	***Hβ***	***Hγ***	***Hδ***	***Other***
L		3.93*(6)*	1.59	1.48	0.82-077	
I	8.50 *(23)*	4.13 *(25)*	1.69	1.4-1.09 *(27)*	0.75 *(36-38)*	
V	8.22 *(42)*	3.95 *(44)*	1.81 *(46)*	0.66 *(48-50 and 52-54)*		
N	8.32 *(58)*	4.57 *(60)*	2.59 *(62,63)*			NH_2 6.73-7.46
Y	8.04 *(72)*	4.43 *(74)*	2.79-2.94 *(76,77)*			7.04 *(80,89)* 6.74 *(82,87)*
L	8.12 *(95)*	4.24 *(95)*	1.53 *(97,98)*	1.42	0.78-0.81	

Table S3: NMR assignment of peptide LIVNYL in water. The number reported in bracket correspond to the numeric id of some of the hydrogens used to discuss the results.

	I, true $J(\omega)$ from MD	*I,* rigid $J(\omega)$ from MD	*I,* rigid $J(\omega)$ from MD with experimental τ_c	*I,* rigid $J(\omega)$ from rep-ex
60-72 (replicate 1)	2.83×10^3	3.01×10^3	1.82×10^4	2.52×10^4
60-72 (replicate 2)	5.82×10^3	6.13×10^3	4.21×10^4	
48-82 (replicate 1)	9.42×10^2	9.96×10^2	5.74×10^3	2.57×10^3
48-82 (replicate 2)	2.14×10^3	2.84×10^3	1.84×10^4	

Table S4: In the second column, the NOE intensities calculated from the plain MD simulations for a short-range pair of protons (60-72) and for a long range one (48-82) using the exact spectral density and Eq. (S9); in the third column, the NOEs calculated with the rigid-body spectral density of Eq. (6) and the isotropic rotational correlation time obtained from the simulation (cf. Fig. S8); in the fifth column, the NOEs calculated from the rigid-body spectral density using the experimental correlation time (cf. Sect. S4); in the fifth column, the NOEs calculated from the replica-exchange simulations with the algorithm suggested in the main text.

Supporting Figures

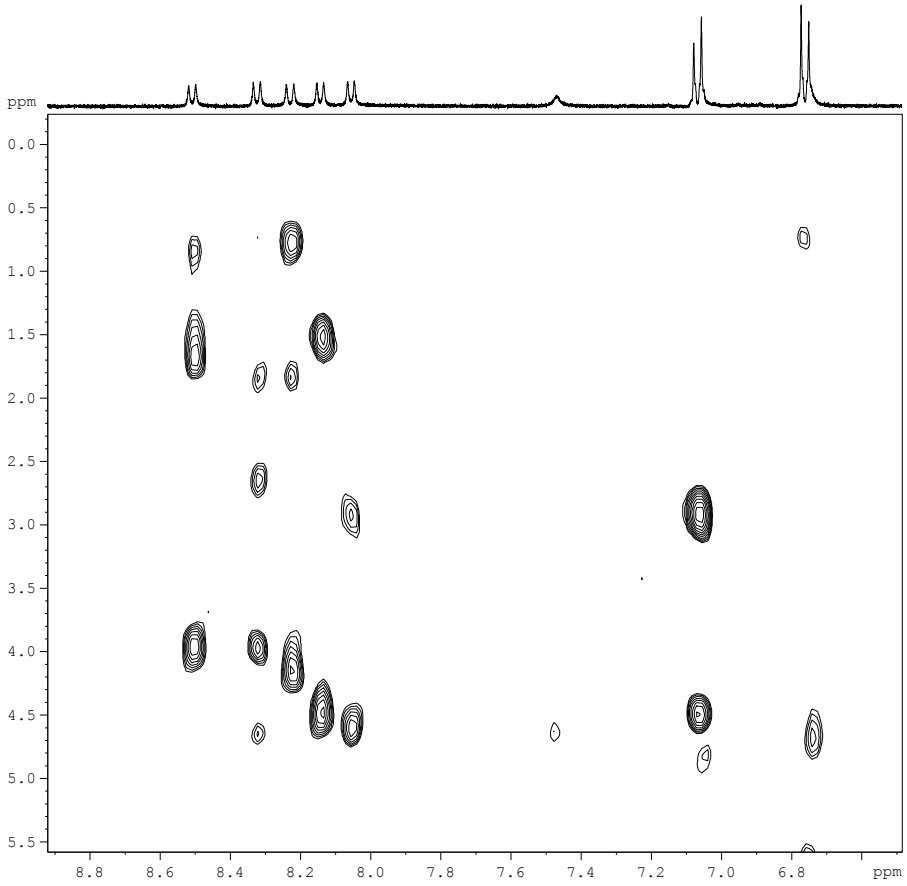


Figure S1: Finger print region of NOESY spectrum (mixing time = 700ms) of peptide LIVNYL in water. The spectrum is recorded on a Bruker Avance III operating at 400 MHz at 298K.

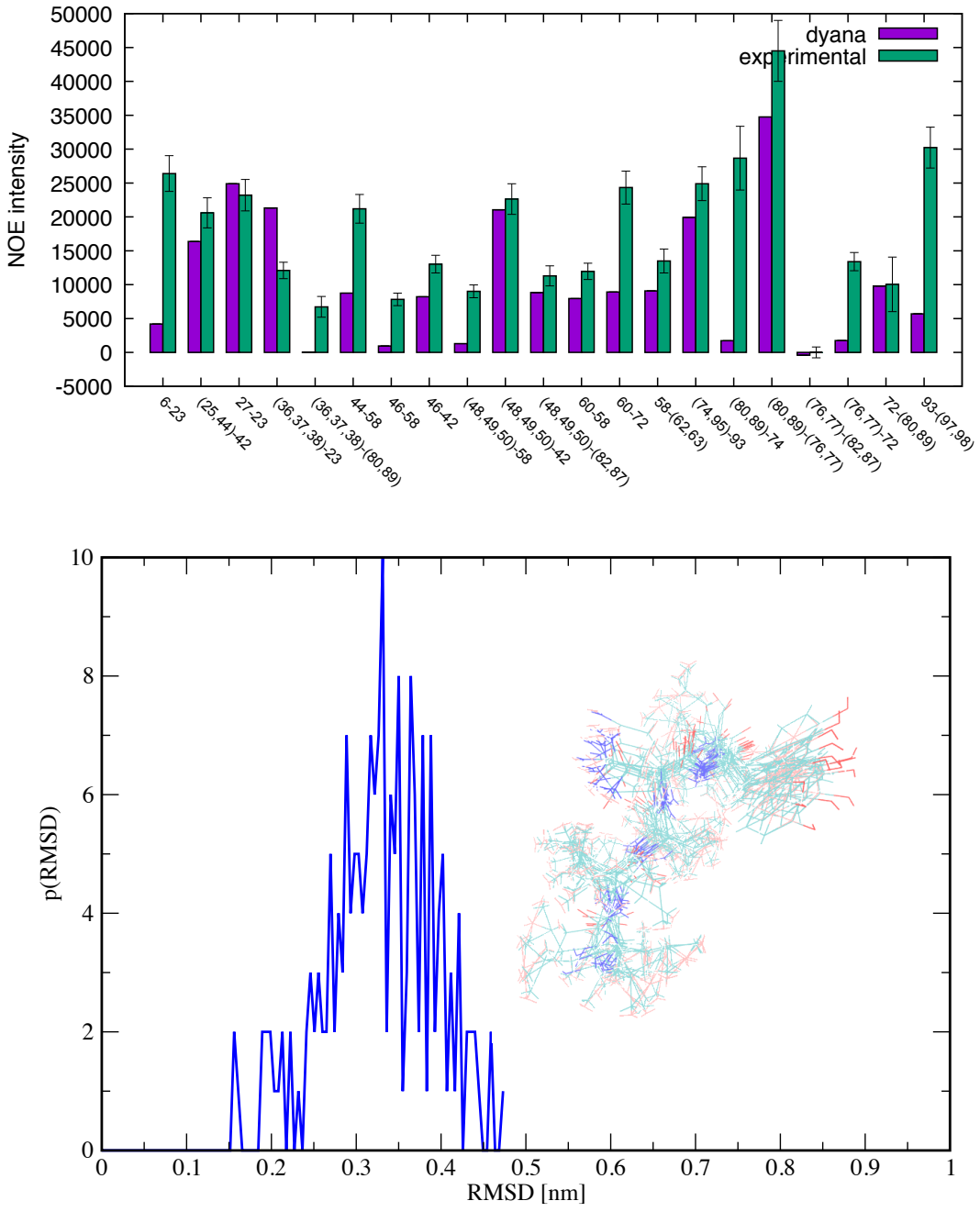


Figure S2: (**above**) The comparison between the NOE intensities calculated with Eq. (1) from the ensemble of conformation obtained by restraint minimization with Dyana, and the experimental ones (giving a χ^2=73.2). (**Below**) The distribution of RMSD between each pair of conformations among the 20 that minimize restraint violations (conformations shown in the inset).

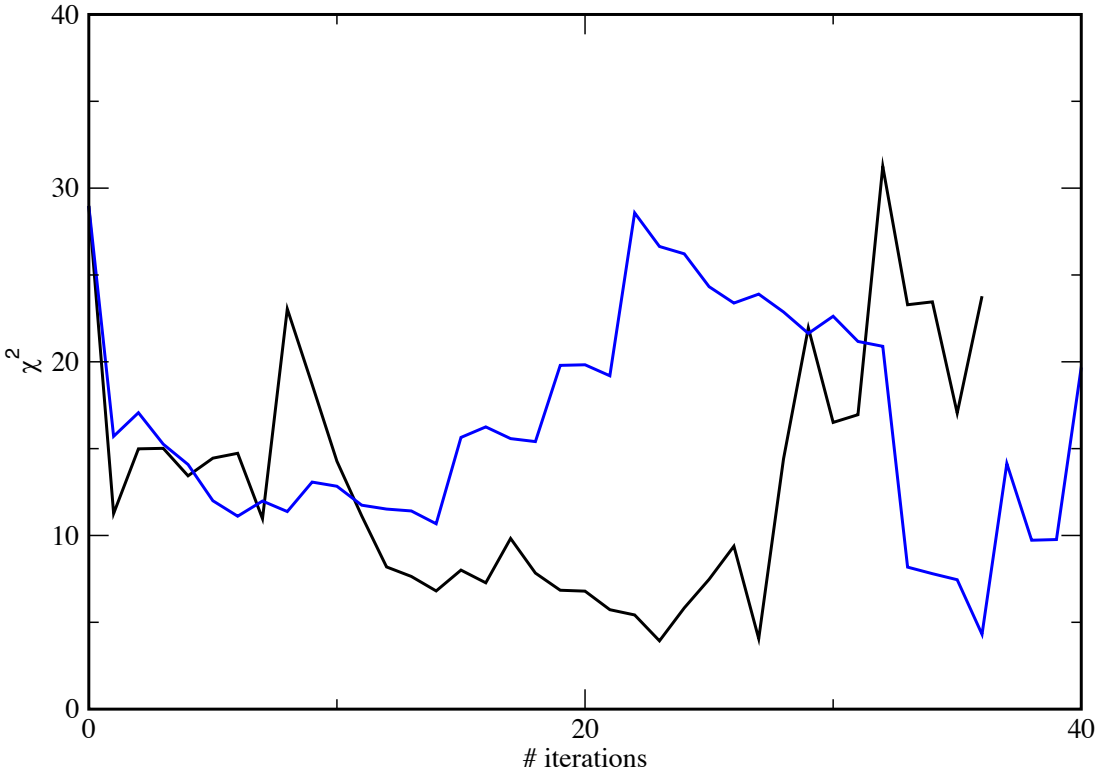


Figure S3: The χ^2 between the calculated and the experimental NOEs as a function of the number of iterations of the optimization algorithms for two independent runs, starting from the Amber03 force field in implicit solvent.

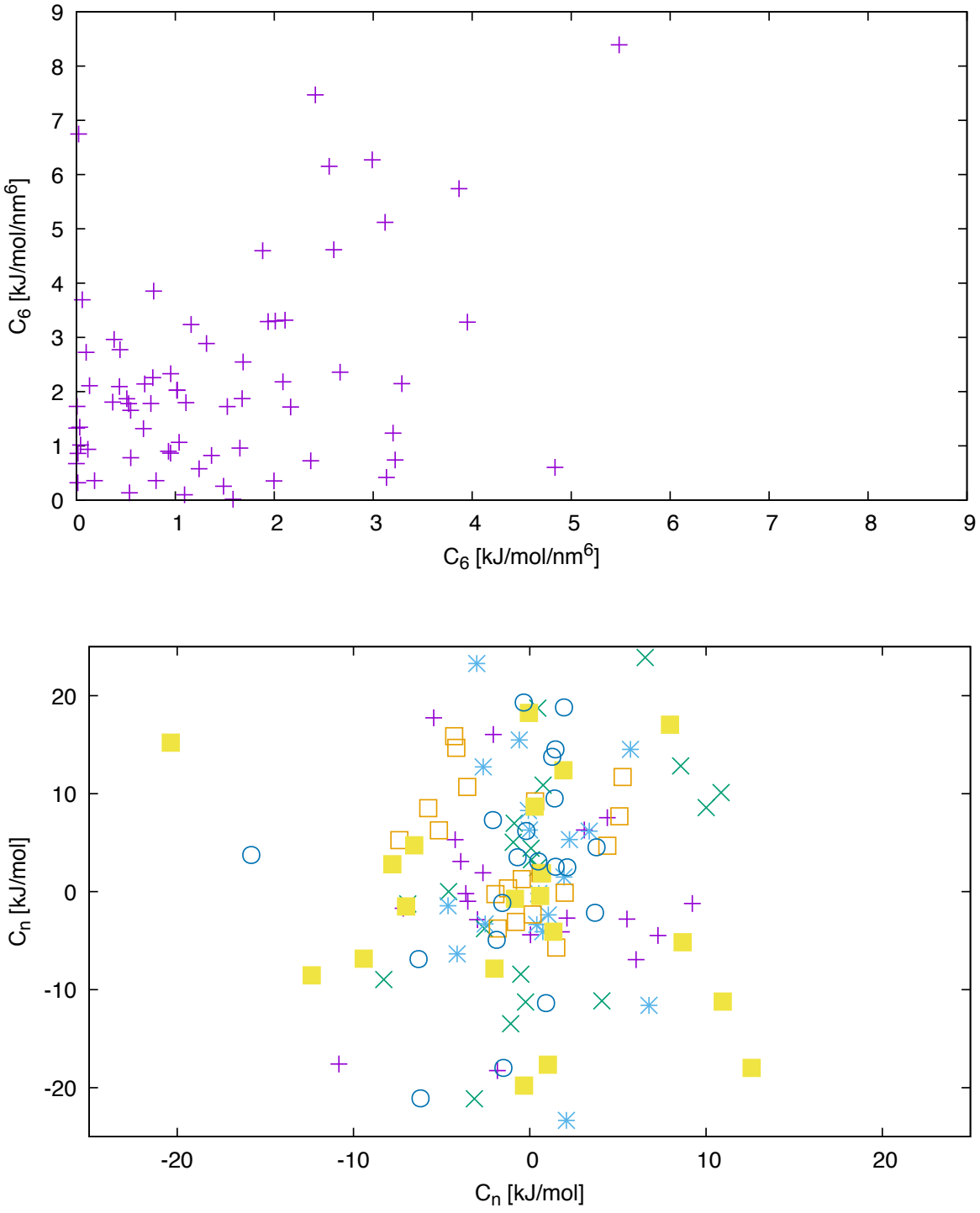


Figure S4: Scatter plots of the C_6 parameters defining the Lennard-Jones potential (**above**) and of the parameters C_n defining the Ryckaert-Bellemans torsional potentials (**below)** obtained in two different optimizations. The correlation coefficients are 0.37 and 0.18, respectively.

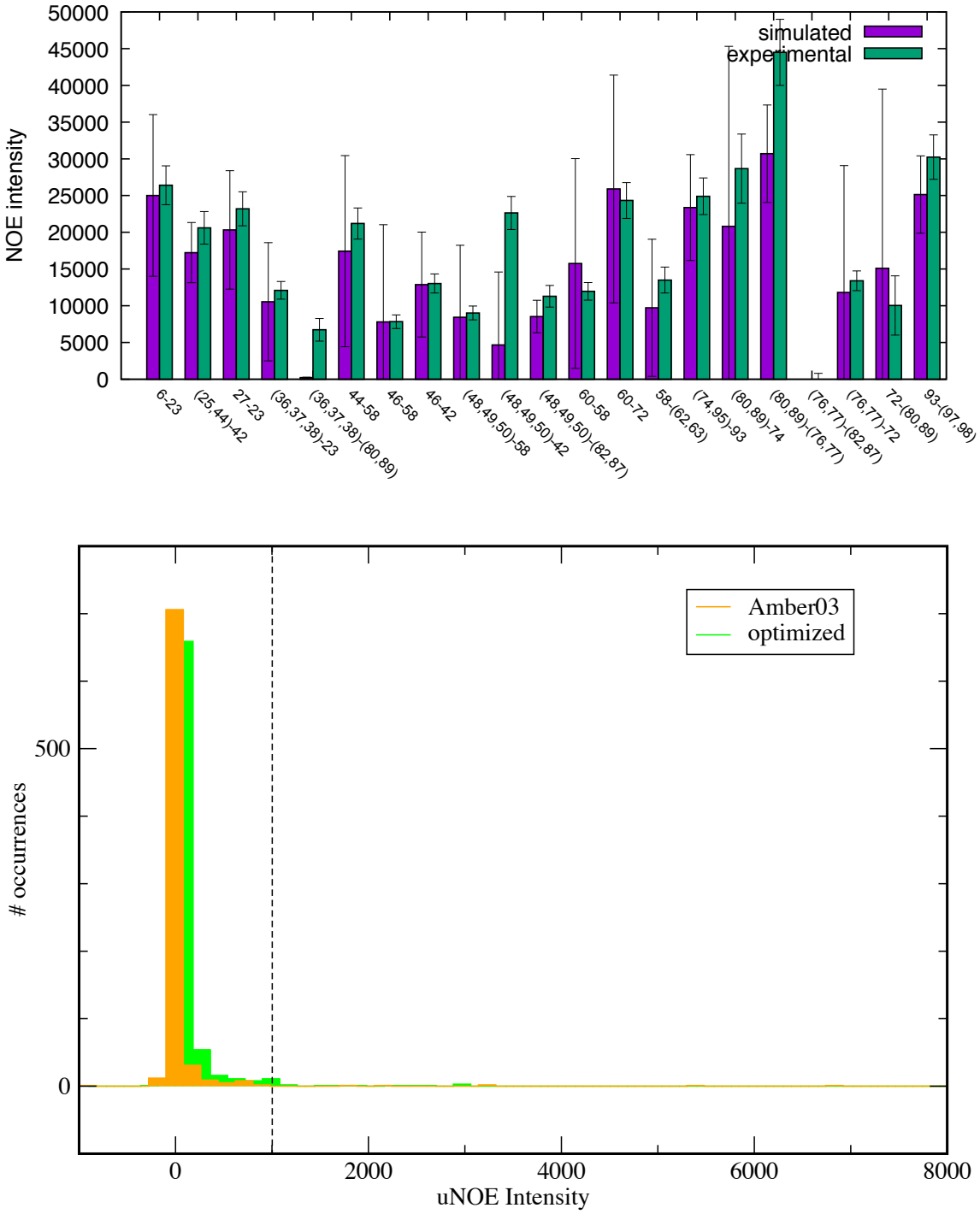


Figure S5: (**above**) Comparison between calculated and experimental NOE intensity for a simulation in which, in addition to the optimization of the observed NOEs, uNOE are optimized to be zero, with error bars set to a conventional value of 1000. Only observed NOEs are plotted here. The overall χ^2 is 0.26, while that restricted to observed NOEs is 4.1. (**Below**) The distribution of uNOE obtained by the simulation with standard Amber in explicit solvent and for the optimized potential. The dashed line indicates the width of the error bar. There are 8 erroneously observed (i.e., above 1000) uNOE in the Amber03 simulation and 10 in the optimized simulation.

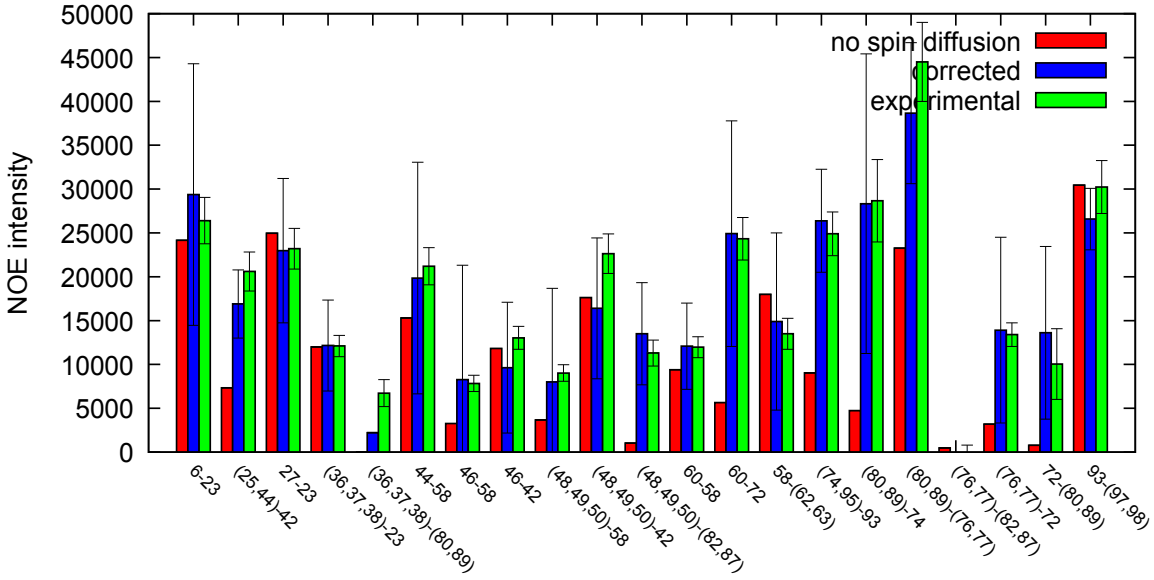


Figure S6: Comparison of the NOE intensities of Fig. 1 with those obtained disregarding spin diffusion, that is using as forward model for each pair of spin the quantity $\langle 1/d^6 \rangle$, where d is the interatomic distance. The χ^2 between the calculated and the experimental NOEs is 16.4.

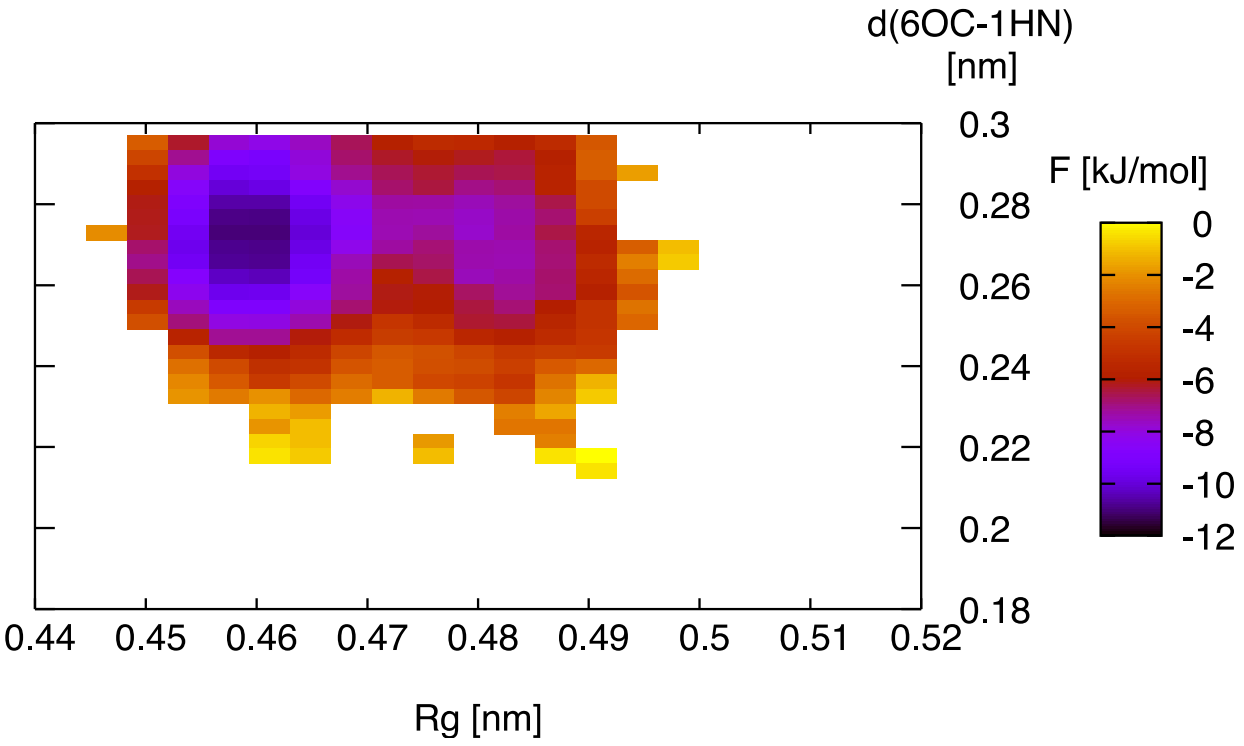


Figure S7: The free energy of the system interacting with the potential optimized without taking into account spin diffusion.

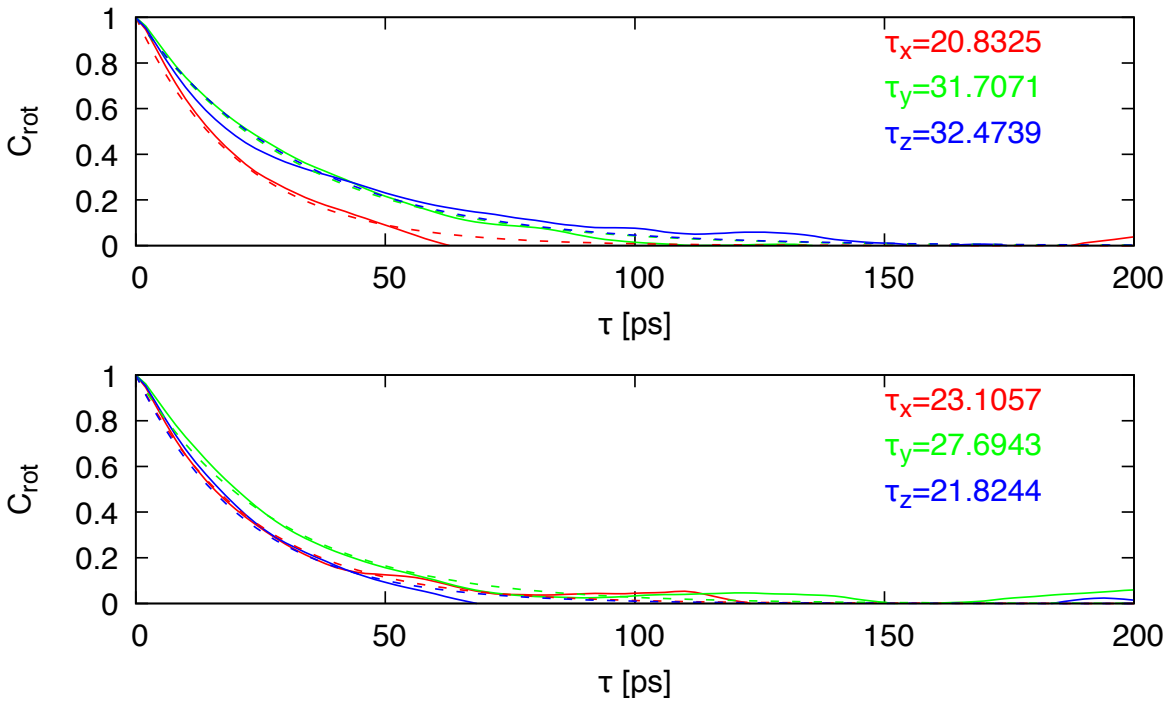


Figure S8: The rotational correlation times along the three Cartesian axes (solid colored lines) calculated from two fixed-temperature MD simulations starting from two initial conditions (upper and lower panel, respectively) with the (same) optimal potential found by the iterative algorithm. Dashed lines indicate the single-exponential fits used to obtain the correlation times indicated in the plots.

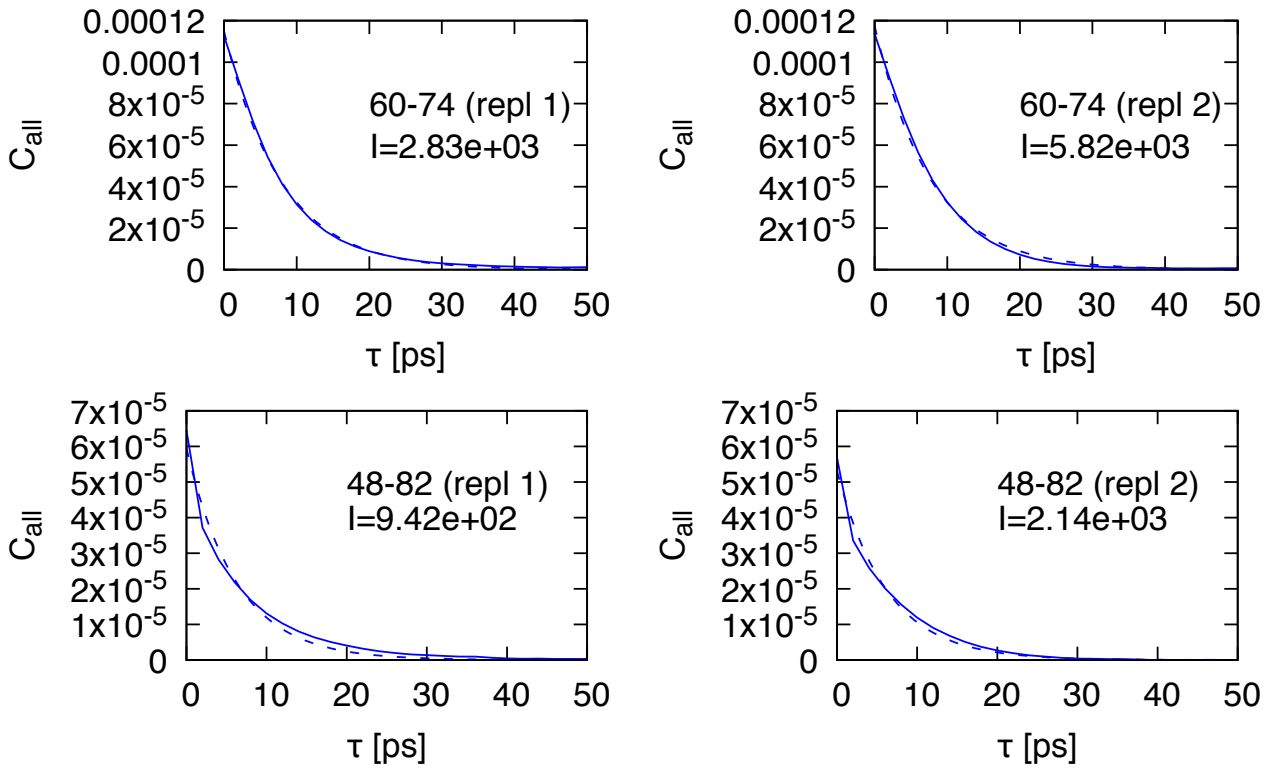


Figure S9: The overall correlation time (solid curve) calculated with Eq. (26) of ref. 25 in the two replicated fixed-temperature MD simulations for contact 60-74 and 48-82. The dashed lines indicate the single-exponential fit. To appreciate the difference between the various behaviors, in each plot is indicated the NOE intensity *I* that would be obtained averaging the full correlation function on the two trajectories.