Investigating β-sheet Nanocrystal Ordering and Correlation With Small-Angle X-ray Scattering

by

Qiushi Mou

A Dissertation Presented in Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy

Approved August 2015 by the
Graduate Supervisory Committee:

Jeffery L. Yarger, Chair
Chris Benmore
Gregory Holland
Robert Ros

ARIZONA STATE UNIVERSITY

December 2015

ABSTRACT

In disordered soft matter system, amorphous and crystalline components might be coexisted. The interaction between the two distinct structures and the correlation within the crystalline components are crucial to the macroscopic property of the such material. The spider dragline silk biopolymer, is one of such soft matter material that exhibits exceptional mechanical strength though its mass density is considerably small compare to structural metal. Through wide-angle X-ray scattering (WAXS), the research community learned that the silk fiber is mainly composed of amorphous backbone and β-sheet nano-crystals. However, the morphology of the crystalline system within the fiber is still not clear. Therefore, a combination of small-angle X-ray scattering experiments and stochastic simulation is designed here to reveal the nano-crystalline ordering in spider silk biopolymer. In addition, several density functional theory (DFT) calculations were performed to help understanding the interaction between amorphous backbone and the crystalline β-sheets.

By taking advantage of the prior information obtained from WAXS, a rather crude nano-crystalline model was initialized for further numerical reconstruction. Using Markov-Chain stochastic method, a hundreds of nanometer size β-sheet distribution model was reconstructed from experimental SAXS data, including silk fiber sampled from *Latrodectus hesperus*, *Nephila clavipes*, *Argiope aurantia* and *Araneus gemmoides*. The reconstruction method was implemented using MATLAB and C++ programming language and can be extended to study a broad range of disordered material systems.

To my parents and family

ACKNOWLEDGEMENTS

I would like to express my deepest gratitude to my academic advisor, Professor Jeffery L. Yarger for his support, guidance and encouragement during my PhD study. I have learned much from his energy, enthusiasm and professional insight on scientific research.

I would like to thank Dr. Chris Benmore for his guidance and patience on X-ray scattering research, Dr. Gregory Holland for his advices and help on NMR and DFT project. I like to thank Professor Robert Ross for serving on my committee.

I would like to express my deepest appreciation to Professor Klauss Schmidt-Rohr at Ames Laboratory and Professor Yang Jiao at ASU for their generous help and discussions during the difficult time of my research.

I would like to thank my group member, Chengchen Guo, J. Bennett Addison, Warner S Weber, Xiangyan Shi and Dian Xu for their kindest help on physical chemistry experiments and great collaborations.

I gratefully acknowledge the financial support by Department of Defense, AFOSR (FA9550-14-1-0014) and the US National Science Foundation (DMR-1264801).

TABLE OF CONTENTS

CHAPTER Page

LIST OF TABLES

LIST OF FIGURES

Figure Page

Figure Page

Chapter 1

X-ray scattering and the correlation function

1.1 Introduction

X-ray scattering is the primary method to measure crystalline structure with atomic level resolution. However, X-ray scattering is not limited to atomic level resolution and it's very versatile. Depending on the wavelength, X-ray can probe the system ranging from 1 to 100 nm. For people with different systems in interest, one can usually tune the wavelength of X-ray system and hence conduct X-ray scattering experiment on this system. By utilizing analytical method, one can rebuild the crystalline structure from X-ray scattering pattern with high precision.

X-ray scattering has been used extensively to solve protein structure and the technology has been pushed to the state of art level. The phase problem has been partially overcome by the invention of several algorithm. Therefore X-ray scattering is a very mature technique to resolve atomic level structure. At this resolution, people use Wide-angle X-ray scattering (WAXS) to refer to X-ray scattering at high-q range in reciprocal space.

For novel material such as biopolymer, one is usually interested in the morphology of the functional units in the material, such as the nano-crystallites embedded into the protein matrix. In this situation, small-angle X-ray scattering, which probe the

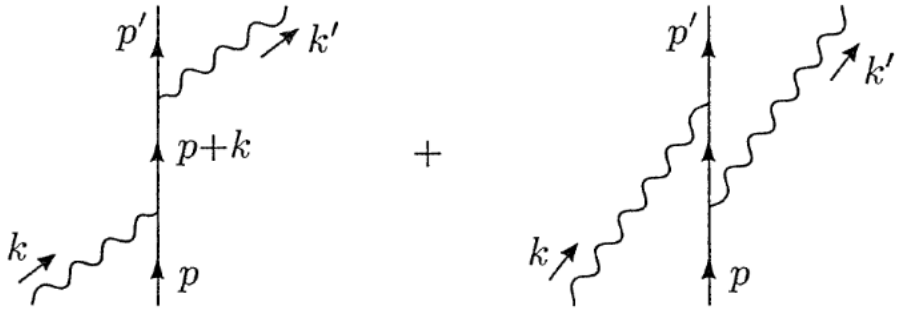

Figure 1.1: The Feynman diagrams of *Compton Scattering*. The left is s-channel scattering and the right is t-channel scattering.

system on a much larger length scale and produces low-q scattering intensities, is the best tool to study the system.

1.2 X-ray scattering by electrons

The interaction between X-ray photons and electrons is a fundamental scattering process and it has a complete analytical description in quantum electrodynamics (QED). The formal name to describe photon-electron scattering, or $e^-\gamma \to e^-\gamma$, is *Compton scattering* Peskin and Schroeder (1995).

The scattering process can be represented by Feynman diagrams as shown in Fig.1.1 where it's composed of two possible scatterings, the s-channel and the t-channel.

Using $\epsilon_\nu(k)$ and $\epsilon^*_\mu(k')$ to denote the polarization of the initial and final photons, we can write down the S-matrix as

$$iM = \bar{u}(p')(-ie\gamma^\mu)\epsilon^*_\mu(k')\frac{i(\not{p}+\not{k}+m)}{(p+k)^2-m^2}(-ie\gamma^\nu)\epsilon_\nu(k)u(p) \tag{1.1}$$

$$+\,\bar{u}(p')(-ie\gamma^\nu)\epsilon_\nu(k)\frac{i(\not{p}-\not{k}'+m)}{(p-k')^2-m^2}(-ie\gamma^\mu)\epsilon^*_\mu(k')u(p) \tag{1.2}$$

$$= -ie^2\epsilon^*_\mu(k')\epsilon_\nu(k)\bar{u}(p')\left[\frac{\gamma^\mu(\not{p}+\not{k}+m)\gamma^\nu}{(p+k)^2-m^2}+\frac{\gamma^\nu(\not{p}-\not{k}'+m)\gamma^\mu}{(p-k')^2-m^2}\right]u(p) \tag{1.3}$$

For electron and massless photon we have $p^2 = m^2$ and $k^2 = 0$, and thus the

denominators of the propagators become

$$(p+k)^2 - m^2 = 2p \cdot k \tag{1.4}$$

$$(p-k')^2 - m^2 = -2p \cdot k' \tag{1.5}$$

using Dirac algebra to simplify the numerators, we obtain the S-matrix of *compton scattering*

$$iM = -ie^2 \epsilon_\mu^*(k')\epsilon_\nu(k)\bar{u}(p')\left[\frac{\gamma^\mu \not{k} \gamma^\nu + 2\gamma^\mu p^\nu}{2p\cdot k} + \frac{-\gamma^\nu \not{k}' \gamma^\mu + 2\gamma^\nu p^\mu}{-2p\cdot k'}\right] u(p) \tag{1.6}$$

With equation 1.6, we can get the expression of scattering cross section by summing over the initial and final electron and photon polarizations

$$\frac{1}{4}\sum_{spins}|M|^2 = \frac{e^4}{4} g_{\mu\rho}g_{\nu\sigma} \cdot tr\Big\{(\not{p}' + m)\left[\frac{\gamma^\mu \not{k} \gamma^\nu + 2\gamma^\mu p^\nu}{2p\cdot k} + \frac{\gamma^\nu \not{k}' \gamma^\mu - 2\gamma^\nu p^\mu}{2p\cdot k'}\right] \tag{1.7}$$

$$\cdot (\not{p} + m)\left[\frac{\gamma^\sigma \not{k} \gamma^\rho + 2\gamma^\rho p^\sigma}{2p\cdot k} + \frac{\gamma^\rho \not{k}' \gamma^\sigma - 2\gamma^\sigma p^\rho}{2p\cdot k'}\right]\Big\} \tag{1.8}$$

After taking traces, the equation simplifies to

$$\frac{1}{4}\sum_{spins}|M|^2 = 2e^4\left[\frac{p\cdot k'}{p\cdot k} + \frac{p\cdot k}{p\cdot k'} + 2m^2\left(\frac{1}{p\cdot k} - \frac{1}{p\cdot k'}\right) + m^4\left(\frac{1}{p\cdot k} - \frac{1}{p\cdot k'}\right)^2\right] \tag{1.9}$$

To further simplify the equation, consider the compton scattering in lab reference frame, where the electron is initially at rest. The change of momentums and energy is illustrated in Fig.1.2. The general cross-section formula is

$$\frac{d\sigma}{dcos\theta} = \frac{1}{2\omega}\frac{1}{2m}\frac{1}{8\pi}\frac{(\omega')^2}{\omega m}\cdot\left(\frac{1}{4}\sum_{spins}|M|^2\right) \tag{1.10}$$

Using lab reference frame and substitute $p\cdot k = m\omega$ and $p\cdot k' = m\omega'$, the final cross-section is

$$\frac{d\sigma}{dcos\theta} = \frac{\pi\alpha^2}{m^2}\left(\frac{\omega'}{\omega}\right)^2\left[\frac{\omega'}{\omega} + \frac{\omega}{\omega'} - sin^2\theta\right] \tag{1.11}$$

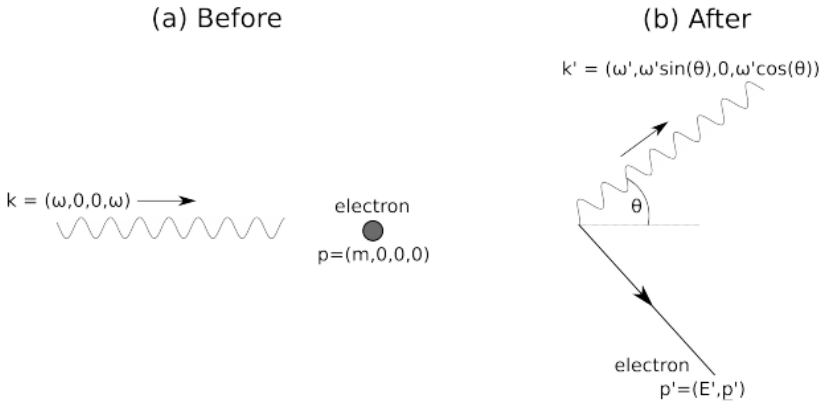

Figure 1.2: Compton scattering in lab reference frame, where the electron is initially at rest.

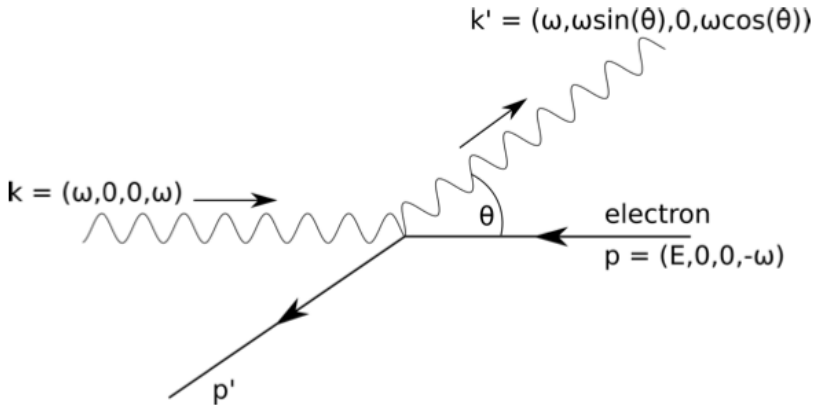

Figure 1.3: Compton scattering in center of mass frame.

where $\omega' = \frac{\omega}{1+\frac{\omega}{m}(1-cos\theta)}$ and this is the spin-averaged *Klein-Nishina formula*.

For high energy X-ray scattering, the equation 1.11 can be simplified in the center of mass frame, as shown in Fig.1.3. The scattering cross section of high energy photon is expressed as

$$\frac{d\sigma}{dcos\theta} = \frac{2\pi\alpha^2}{2m^2 + s(1+cos\theta)} \tag{1.12}$$

Integrate the formula over the solid angle, we obtain the total scattering cross section

$$\sigma_{total} = \frac{2\pi\alpha^2}{s} log\left(\frac{s}{m^2}\right) \tag{1.13}$$

where $s = E_{com}^2$. The last two equations can be used to calculate angle-dependent and total scattering cross-section of high energy X-ray.

The individual X-ray scattering event is governed by Compton scattering formula, which can be hard to generalize to the scattering of an ensemble of electrons. Fortunately, to express the scattering of a bulk system with a known electron density distribution, we can use general X-ray scattering theory, which is closely related to the *Fourier transformation* mathematically. This unique mathematical property enable us to effectively simulate the X-ray scattering of a bulk system. Next, we will look at how to express the scattering of an arbitrary electron density analytically.

1.3 X-ray scattering of arbitrary system and two-point correlation

Given an arbitrary electron density distribution $\rho(\mathbf{x})$, one can calculate the scattering amplitude

$$A(\mathbf{q}) = \int \rho(\mathbf{x}) exp(-2\pi i \mathbf{q} \cdot \mathbf{x}) \cdot d\mathbf{x} \tag{1.14}$$

where $\rho(\mathbf{x})$ can be continuous or discrete electron density distribution in 2 or 3 dimensional geometry and $\mathbf{q}$ is the scattering vector or momentum transfer in reciprocal space. The numerical value of $\mathbf{q}$ is calculated as $q = 2\sin\theta/\lambda$. The scattering intensity follows as

$$I(\mathbf{q}) = |A|^2 = A^* A \tag{1.15}$$

The scattering intensity profile $I(\mathbf{q})$ directly reflects the structural correlations on a wide range of length-scale. We can see this by expanding the Eq.1.15

$$I(\mathbf{q}) = A(\mathbf{q})^* A(\mathbf{q}) \tag{1.16}$$

$$= \int d\mathbf{x}' \cdot exp(2\pi i \mathbf{q} \cdot \mathbf{x}')\rho(\mathbf{x}') \int d\mathbf{x} \cdot exp(-2\pi i \mathbf{q} \cdot \mathbf{x})\rho(\mathbf{x}) \tag{1.17}$$

$$= \int \int d\mathbf{x}' d\mathbf{x} \cdot exp(-2\pi i \mathbf{q}(\mathbf{x} - \mathbf{x}'))\rho(\mathbf{x}')\rho(\mathbf{x}) \tag{1.18}$$

$$= \int d\mathbf{r} \cdot exp(-2\pi i \mathbf{q} \cdot \mathbf{r}) \int d\mathbf{x}' \rho(\mathbf{x}')\rho(\mathbf{x}' + \mathbf{r}) \tag{1.19}$$

Substituting the equal time correlation function Sethna (2006)

$$C_t(\mathbf{r}) = \langle \rho((\mathbf{x}, t)\rho((\mathbf{x} + \mathbf{r}, t) \rangle \tag{1.20}$$

$$= \frac{1}{V} \int d\mathbf{x} \rho(\mathbf{x})\rho(\mathbf{x} + \mathbf{r}) \tag{1.21}$$

One can write the scattering intensity in terms of correlation function

$$I(\mathbf{q}) = V \int d\mathbf{r} \cdot exp(-2\pi i \mathbf{q} \cdot \mathbf{r}) C(\mathbf{r}) \tag{1.22}$$

$$= V \tilde{C}(\mathbf{q}) \tag{1.23}$$

where V is the volume of the isotropic, homogeneous system, $\tilde{C}(\mathbf{q})$ is the Fourier transform of the two-point correlation function $C(\mathbf{r})$. Equation 1.22 has important physical significance as it tells us that the scattering intensity profile is just a Fourier transformation on the two-point correlation function of the system. Therefore any fluctuations in the $I(\mathbf{q})$ will be analytically related to the correlation characteristic of the scattering centers in the material. In biopolymer, the scattering centers are usually form by secondary structures in the macromolecules.

The correlation function is generally a up sloping curve with fluctuations. To taking the density of states into account, one has to scale it properly. For a discrete electron density distribution with point mass

$$\rho(\mathbf{x}) = \sum_i \delta(\mathbf{x} - \mathbf{x}_i) \tag{1.24}$$

the two-point correlation function can be calculated as

$$C(r) = \frac{1}{N}\sum_{i=1}\sum_{j\neq i}\frac{w_i w_j}{\langle w\rangle^2}\delta(r - r_{ij}) \tag{1.25}$$

where w_i is the weighting factor for the phase in interested. In a two-phase system, it will be set to 1 for crystalline phase and 0 for amorphous backbone phase. The correlation distance r runs through 0 to maximum length of the system and $r_{ij} = |\mathbf{x}_i - \mathbf{x}_j|$ will be calculated for each possible pair. On a length scale over 10 nm, we could assume the system in biopolymer is isotropic and homogeneous with an average density of ρ_0, and therefore we can obtain the pair correlation function in 3 dimensional space at intermediate length scale (10 - 500 nm)

$$P(r) = \frac{1}{4\pi r^2\rho_0}\frac{1}{N}\sum_{i=1}\sum_{j\neq i}\delta(r - r_{ij}) \tag{1.26}$$

For 2 dimensional system, the density of state scales proportional to $2\pi r\rho_0$ and the corresponding pair correlation function reads as

$$P(r) = \frac{1}{2\pi r\rho_0}\frac{1}{N}\sum_{i=1}\sum_{j\neq i}\delta(r - r_{ij}) \tag{1.27}$$

1.4 Numerical simulation of X-ray scattering

From the definition of correlation function $C(\mathbf{r})$, it's obvious that if we know the electron density distribution $\rho(\mathbf{r})$, then we can calculate the correlation function directly. In practice, one can define arbitrary distribution in space using a 2 or 3 dimensional matrix. For example, if we want to define a two-phase distribution in 2D space, then we construct a $N \times N$ matrix then initialize the matrix elements using the indicator function

$$I(\mathbf{x}) = \begin{cases} 1, & \mathbf{x} \in \Omega \\ 0, & \mathbf{x} \in \Omega^c \end{cases} \tag{1.28}$$

where the matrix element at $\mathbf{x} = (i, j)$ where $0 \leqslant i, j \leqslant N$ equals to 1 if it resides in the region Ω occupied by crystalline phase, and 0 otherwise.

Equation 1.24 describe the density distribution of point masses, which is not particularly useful for the study of real material. One can easily modify it to incorporate mass spread and geometry. For a local density distribution ρ_s located on each mass point, the total electron density distribution $\rho_{dt}(\mathbf{x})$ can be represented by

$$\rho_{dt}(\mathbf{x}) = \rho(\mathbf{x}) * \rho_s(\mathbf{x}') \tag{1.29}$$

$$= \int d\mathbf{x}' \cdot \rho_s(\mathbf{x}') \sum_i \delta(\mathbf{x}' - \mathbf{x} + \mathbf{x}_i) \tag{1.30}$$

$$= \sum_i \rho_s(\mathbf{x} - \mathbf{x}_i) \tag{1.31}$$

where $\mathbf{x}_i$ is coordinate of the ith scattering center in the system. To calculate the scattering pattern of the desired electron density distribution in Eq.1.29, one use Eq.1.15, which can be realized numerically by Fast Fourier transform (FFT).

Now the $\rho_{dt}(\mathbf{x})$ effectively describe the density distribution in the entire simulation region, with each scattering center has its unique shape ρ_s. However, in real numerical experiment, one can only represent the density map by a matrix, which by nature is discretized.

One can construct a continuous electron density map in 2-D plane or 3-D space using a discrete grid represented by a matrix and then convolute the grid using the continuous shape function of the elementary grid, which will be a square for 2-D matrix, or a cubic for 3-D matrix. The discretized version of $\rho_{dt}(\mathbf{x})$ is denoted as $\rho_d(\mathbf{x})$, and the shape of the elementary grid is denoted as $\rho_e(\mathbf{x})$, then we form a continuous density map

$$\rho_c(\mathbf{x}) = \rho_d(\mathbf{x}) * \rho_e(\mathbf{x}) \tag{1.32}$$

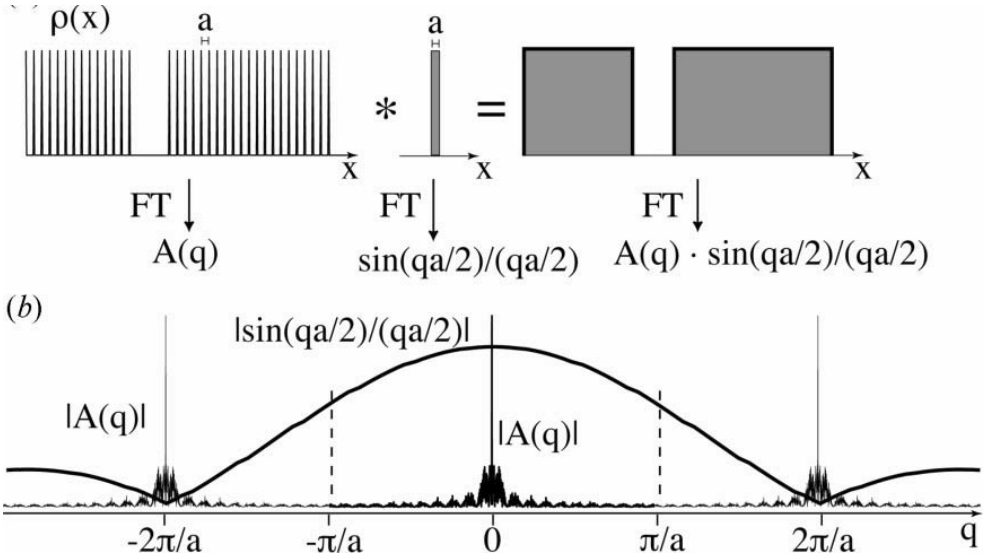

Figure 1.4: Illustration of the convolution. Constructing a continuous density distribution using a discrete matrix and the structure factor of the unit shape. Schmidt-Rohr (2007)

Now the scattering pattern should be calculated from $\rho_c(\mathbf{x})$

$$I(q) = \mathscr{F}[\rho_c(\mathbf{x})]^2 \tag{1.33}$$

$$= \mathscr{F}[\rho_d(\mathbf{x})]^2 \cdot \mathscr{F}[\rho_e(\mathbf{x}))]^2 \tag{1.34}$$

where we have used the *Convolution Theorem*. Eq.1.34 enables us to efficiently represent the electron density map in computer using a matrix, while preserves the ability to calculated the scattering pattern from a real continuous distribution when needed.

To illustrate the convolution construction of continuous density distribution using discrete matrix, we show a toy model in Fig.1.4. To represent a continuous block shape density distribution, we need to construct a 1-D vector in computer, where the '1' element represent the existence of density at that location, and '0' represents vacuum. Such description using a vector or matrix (2-D) is discrete in nature. If Fourier transformation is applied, the resulting spectrum will not be accurate. Hence we need to convolute the discrete vector or matrix with the shape function of the 'unit cell', which is a uniform distribution in 1-D and a square function in 2-D. According

to convolution theorem 1.34, the resulting structure factor $S(q)$ is the multiplication between the FFT of vector/matrix $|A(q)|^2$ and the structure factor of the 'unit cell' $(sin(qa/2)/(qa/2))^2$ for the case in Fig.1.4.

For a 2-D density distribution, suppose we store the desired density map in matrix $M(x, y)$, which is discrete in nature, then we can obtain the scattering structre factor by calculating

$$S(q) = \mathscr{F}[M(x,y)]^2 \cdot \left(\frac{sin(k_x a/2)}{k_x a/2} \frac{sin(k_y a/2)}{k_y a/2} \right)^2 \tag{1.35}$$

where k_x and k_y denote the reciprocal space unit vector along x and y dimension, and a is the length of the 'unit cell' in the density map. This formula can be easily generalized to 3-D space.

Chapter 2

Stochastic reconstruction with stimulated annealing

In this chapter we will discuss the basic concepts in stochastic process and Monte Carlo simulation. The fist key concept is Brownian motion, which is the formal description of the easy to apprehend idea of random walk. Then we will discuss the Markov-chain Monte-Carlo simulation method used in Quantum Chemistry and the implementation of the algorithm. This method form the ground of our stochastic optimization procedure for analyzing the X-ray scattering data. Lastly, we need to look at some important results in stochastic process and relate them to physical model we use.

2.1 Random Walk and Metropolis Sampling

2.1.1 Random walk and Brownian motion

Random walk is a path in real or state space that take successive steps in random directions. It's a common phenomenon in statistical physics and biology: trajectories of particles under collision, linked polymer morphology and pollen grains path in water, which is endorsed a famous name of Brownian motion. Random walk has several extraordinary properties. First, the morphology of random is self-similar under the observations in different length scale, therefore the physical structure is fractal. Second, the end point of the random walk is independent of the microscopic

details and can be describe by equation of classical motion, especially the *diffusion equation*. Both of these properties is crucial to understand the macroscopic behavior of the system that exhibits randomness.

Classical examples of random walk are of similar characteristic and have been extensively studies, such as the **Drunkard's walk** problem Sethna (2006). The random walk process has been abstracted to form the concepts of *Brownian motion*, which has a wide range of application in Physics, Chemistry and Economics. Therefore, we introduce several definitions that are essential to describe the Brownian motion. First is the definition of random variable from Chung and AitSahlia (2012).

Definition 2.1. Random Variable.

A numerically valued function X of ω with domain Ω:

$$\omega \in \Omega : \omega \to X(\omega) \tag{2.1}$$

is called a random variable on Ω.

The random variable *r.v.* is essentially a function that map a event or a subset of events in the sample space Ω to a numerically value in the range of [0,1]. With the definition of *r.v.*, we can proceed to define the Brownian motion.

Definition 2.2. Brownian motion.

A family of random variables $\{X(t)\}$, indexed by the time variable t ranging over $[0, \infty)$, is called the Brownian motion iff it satisfies the following conditions:

(i) $X(0) = 0$;

(ii) the increments $X(s_i + t_i) - X(s_i)$, over an arbitrary finite set of disjoint intervals $(s_i, s_i + t_i)$, are independent random variables;

(iii) for each $s \geq 0, t \geq 0, X(s + t) - X(s)$ has the normal distribution $N(0, t)$.

2.1.2 Random Number Generation

Generating random number is the first step in stochastic simulation. One common approach is to generate *pseudorandom* numbers Knuth (1981). Due to the advance in CPU hardware design, now it's possible to generate real random numbers sampled from the thermal fluctuations on Intel CPU using their Digital Random Number Generator (DRNG) API.

2.1.3 Markov chains

The random walk can be generalized to the *Markov chains*. The key property of Markov chain is the so-called no-memory, which means the state transition probabilities depend only on the current state of the system, not on how or when it got there. The transition from state j to k follows the transition probability p_{kj}, regardless the path it took from the initial state to j

$$P(X_k \leftarrow X_j) = p_{kj} \tag{2.2}$$

with X_k at time stamp t and X_j at time stamp $t+1$. The X_j state will transit to another state for certain, so $\sum_{k=1}^{N} p_{kj} = 1$.

Definition 2.3. Markov Chain.

A stochastic process $\{X_n, n \in N^0\}$ taking values in a countable set I is called a *homogeneous Markov Chain*, or *Markov chain with stationary transition probabilities*, iff equation 2.2 holds.

To describe multiple state transition during one time period, we can define the probability-space density at time i as column vector,

$$\mathbf{p}^{(i)} = \begin{bmatrix} p_1^{(i)} \\ \vdots \\ p_N^{(i)} \end{bmatrix}$$

$p_k^{(i)}$ is the probability that system is in state X_k at time i. The transition from time i to i+1 can be written in matrix notation

$$\mathbf{p}^{(i+1)} = \mathbf{P}\mathbf{p}^{(i)}$$

$\mathbf{P}$ is the $N \times N$ transition matrix. From initial state, after n steps walk, the state distribution is

$$\mathbf{p}^{(n)} = \mathbf{P}^n\mathbf{p}^{(0)}$$

which means the final state is independent of the 'trail' the system took during random walk process. From

$$\mathbf{p}^{(*)} = \mathbf{P}\mathbf{p}^{(*)}$$

we can get the equilibrium state distribution $\mathbf{p}^{(*)}$.

2.1.4 Metropolis Algorithm

It's been proved that a discrete dynamical system with a finite number of states can converge to an equilibrium distribution $\mathbf{p}^{(*)}$ if it's an ergodic(can reach every state and is acyclic) Markov chain and meanwhile satisfy *detailed balance* Gardiner (2009)

$$P_{kj}X_j = P_{jk}X_k \tag{2.3}$$

With equilibrium state distribution $\mathbf{p}^{(*)}$, we can find the acceptance probability A. If each walk is based on uniform density, then the acceptance probability is

$$A(k \leftarrow j) = min\left(1, \frac{p_k}{p_j}\right) \tag{2.4}$$

In real physical problem, the probability p can be *Boltzmann* distribution $p(\xi) = Z^{-1}e^{-\xi/k_BT}$, wave function $\psi(r)$, or any density distribution we need. If the energy landscape of the simulation space is not uniform, and probability of moving from state

j to state k is T_{kj} given by the potential function, then the acceptance probability is Metropolis *et al.* (1953)

$$A(k \leftarrow j) = min\left(1, \frac{T_{jk}p_k}{T_{kj}p_j}\right)$$

and we can prove that the above choice satisfy *detailed balance*. The generalized form of Metropolis acceptance probability can be written

$$A(y, x; \Delta t) = min\left(1, \frac{G(x, y; \Delta t)p(y)}{G(y, x; \Delta t)p(x)}\right) \tag{2.5}$$

The exact form of function $G(x, y; \Delta t)$ should be chosen according to potential landscape of the simulated system.

2.2 Diffusion Monte Carlo

2.2.1 Imaginary time Schrödinger equation

We simply consider a single particle move along the x-axis in potential $V(x)$, the wave function is $\psi(x, t)$ which is governed by Schrödinger equation (use atomic unit, $m = 1$, $\hbar = 1$)

$$i\frac{\psi}{t} = \hat{H}\psi$$

the Hamiltonian is

$$\hat{H} = -\frac{\partial^2}{\partial x^2} + V(x)$$

The potential goes infinite as $x \to \infty$, the particle would move in a finite space, then wave function can be expended with eigenfunctions

$$\psi(x, t) = \sum_{n=0}^{\infty} c_n \phi_n(x) e^{-iE_n t}$$

If we shift the energy scale in form of $V(x) \to V(x) - E_T$, and meanwhile introduce the imaginary time Kosztin *et al.* (1996) $\tau = it$, Schrödinger equation change to

$$\frac{\partial \psi}{\partial \tau} = -D\nabla^2\psi + (V(x) - E_T)\psi \tag{2.6}$$

where $D = 1/2$. Now wave function can be expended

$$\psi(x,t) = \sum_{n=0}^{\infty} c_n \phi_n(x) e^{-(E_n - E_T)\tau}$$

after passing a long time, i.e. $\tau \to \infty$, we have $\psi(x, \tau \to \infty) = c_0\phi_0$ if $E_T = E_0$.

2.2.2 Green function and short time approximation

The integral form of Eq.(1) isq

$$\psi(y, \tau + \Delta\tau) = \int G(y, x; \Delta\tau)\psi(x, \tau)dx$$

where $G(y,x) = \langle y|(\hat{H} - E_T)|x\rangle$, and we can prove that $G(y,x)$ also satisfy Schrödinger Equation. This function can be interpreted as imaginary time propagator, which also is transition probability. As stated above, as long as we choose exact ground energy, after long time the wave function would converge to the eigenfunction of ground state.

Split the green function into diffusion and branching part Hammond *et al.* (1994)

$$G = G_{diff}G_B = e^{-T\tau}e^{-(V - E_T)\tau}$$

where we separate the kinetic and potential energy part in Hamiltonian. Because the Green function satisfy Schrödinger equation, we can get the diffusion function

$$G_{diff}(y, x; \tau) = (4\pi D\tau)^{-3N/2} e^{-(y-x)^2/4D\tau} \tag{2.7}$$

and branching function

$$G_B(y, x; \tau) = e^{-(\frac{1}{2}(V(x)+V(y)) - E_T)\tau} \tag{2.8}$$

If we divide the long time into small part, and each $\Delta\tau$ would mean a single diffusion monte carlo step in computer stimulation. We propose a trail move base on the diffusion factor, use metropolis to accept/reject the move, if accepted calculate the branching factor to delete or add walkers.

2.2.3 Importance sampling

Rather than using $\psi(x)$ only, we will introduce a guiding function that based on the available knowledge of $\psi_0(x)$, to construct a new distribution which also satisfy Schrödinger equation

$$f(x,\tau) = \psi_G(x,\tau)\psi(x,\tau)$$

This would also introduce a new "quantum force" factor into the drifting process

$$\mathbf{F} = 2\nabla\psi_G/\psi_G$$

With the guiding function, the walkers would be initialized more close to the ground state distribution, and the local energy would also be more close to the trial energy which was get from Variational methods by guiding wave function. Hence the fluctuation of distribution f will be minimized. With importance sampling, the branching part is modified to

$$G_B(y,x;\delta\tau) = e^{-(\frac{1}{2}(E_L(x)+E_L(y))-E_T)\delta\tau} \tag{2.9}$$

and the diffusion part is

$$G_{diff}(y,x;\delta\tau) = (4\pi D\delta\tau)^{-3N/2} e^{-(x-y-D\delta\tau\mathbf{F}(y))^2/4D\delta\tau} \tag{2.10}$$

To make sure the system would converge and go to equilibrium, Metropolis acceptance probability is

$$A(y,x;\delta\tau) = min\left(1, \frac{|\psi_G(y)|^2 G(x,y;\delta\tau)}{|\psi_G(x)|^2 G(y,x;\delta\tau)}\right) \tag{2.11}$$

and such choice guarantee detailed balance. While the algorithm is Markovian and ergodic, distribution f would converges to $\psi_G\psi_0$. To evaluate the ground energy, we use *Reference Energy* E_R to substitute local energy and update reference energy after each complete diffusion process. From n^{th} to $(n+1)^{th}$ DMC, reference energy is updated Hammond *et al.* (1994)

$$E_R^{(n+1)} = E_R^{(n)} + \frac{1}{\delta\tau}\left(1 - \frac{N_{n+1}}{N_n}\right) \tag{2.12}$$

where N_n is the number of walkers. After sufficient long time Kalos and Ceperley (1979), the population of walkers don't change(with small fluctuation), then the reference energy is steady and can be used to evaluate the ground energy. Besides the diffusion methods, there are some refinements like *Bessel* Green's function methods, *Domain* Green's function methods and *Coulomb* Green's function methods.

2.2.4 DMC Algorithm

1. Initialize a number of N_0 walkers according to the probability distribution of guiding function $|\psi_G|^2$. Initialize the reference energy E_R with the local energy E_L from previous Variational monte carlo process.

2. Run a complete DMC

   (a) For every i^{th} walker in the ensemble, move it according to

   $$\mathbf{y}_i = \mathbf{x}_i + D\delta\tau\mathbf{F} + ran(Gaussian)$$

   where the random number generater is based on Gaussian distribution with zero mean and variance of $2D\delta\tau$ (Eq.(5))

   (b) Compute the Metropolis acceptance probability

   $$W(y,x) = \frac{|\psi_G(y)|^2 G(x,y;\delta\tau)}{|\psi_G(x)|^2 G(y,x;\delta\tau)}$$

(c) Accept or reject the trail move according to Metropolis algorithm

$$A = min(1, W(y, x))$$

and record acceptance ratio

$$Accept = Accept + \frac{1}{N_n}$$

(d) Compute the branching function $u = G_B$

(e) Add a number of $BW = int(u + ran(seed))$ new walkers to the ensemble.

(f) Update reference energy E_R according Eq.(7), and update time step

$$\delta\tau = \delta\tau \cdot Accept$$

3. Repeat the above DMC walk for desired magnitude.

4. Collect data like E_L when sufficient time have passed and the system's fluctuation is minimized.

2.3 Stimulated Annealing reconstruction

Stimulated annealing is inspired by the annealing process in the alloy forming process and thus can be best understood in the statistical physics framework. Statistical physics gives an analytical framework to describe the behavior and dynamics of an ensemble system consisting of many particles. In the ensemble, each configuration, taking the position r_i of an atom for example, is weighted by the Boltzmann distribution $exp(-E(r_i)/k_BT)$, where $E(r_i)$ is the energy of the configuration, k_B is the Boltzmann constant, and T is the temperature of the system.

So the final configuration of the system in question, whether is crystal or glass, is not determined by the low limit of the temperature but the annealing process,

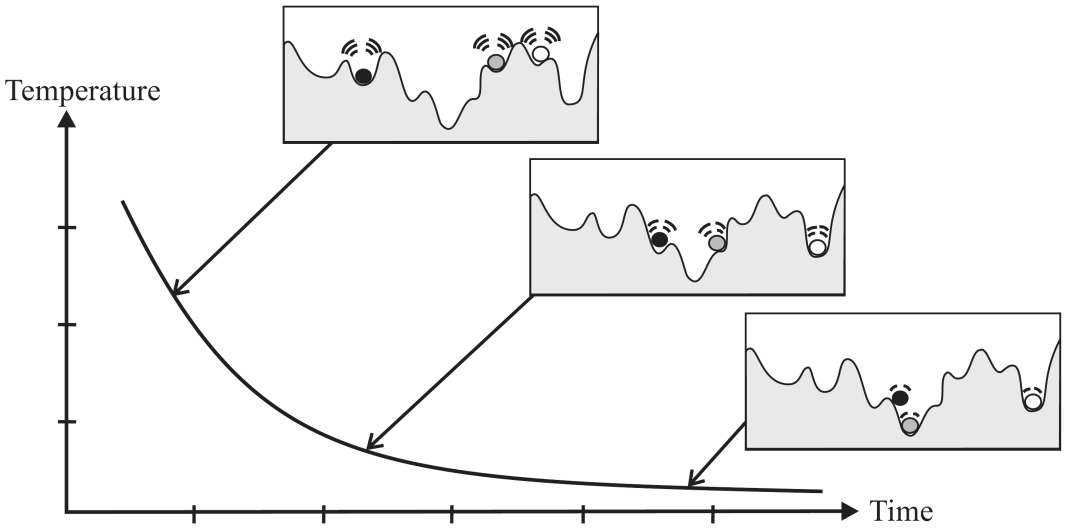

Figure 2.1: The stimulate annealing optimization process. At high temperature, the system has a finite probability of jumping out of the local minimum and thus achieving lower final energy.

found from experiments Kirkpatrick *et al.* (1983). The process is first to melt the material, then lowering the temperature slowly and spending a long time around the crystallization point. If the process is fast, then the material will get out of equilibrium state and forms glass, which is not in a globally stable ground state.

The stimulated annealing was developed from Metropolis-Hastings algorithm Metropolis *et al.* (1953), which is essentially a Markov Chain Monte Carlo (MCMC) method, as describe in previous section. In Metropolis' framework, each step one atom will be randomly moved and the change of the system energy ΔE, then the probability of accepting the move is calculated by

$$P(\Delta E) = e^{-\frac{\Delta E}{k_B T}} \tag{2.13}$$

Random number is then sampled from uniform distribution on $[0, 1]$. If it's smaller than $P(\Delta E)$, the new configuration is accepted, otherwise it's rejected.

For other ensemble system, the energy E is replaced by cost function of the system. In SAXS simulation, it is defined as the Euclidean distance from simulated curve and

experiment curve. The temperature T is now a psedudo physical parameter defined by the user. As large T, the system has a high probability of accepting a move that increase the system energy. As the annealing stage proceed, the temperature T is slowly reduced to near zero, at which point the system is close to 'freeze' state and the optimization is acting like a strict local search method. This annealing process is illustrated in Fig.2.1 (source:google). As a rule of thumb, the initial temperature can be set such that the acceptance ratio is around 0.5. In practice, one need to run several trail tests to find it. The cooling rate is usually chosen to be exponentially decreasing, such as $T_n = 0.9^n T_0$, where n is the number of stage. When there are certain number of accepted move, the algorithm can move to next stage, thus reduce the temperate by a predefined factor (0.9 in this example).

The final configuration of the system is in low energy state, but never was a true optimal one. As pointed by Kirkpatrick *et al.* (1983), such system usually have many degenerate lowest energy states. Even there exists a lower energy state, the difference would be so small to make a practical impact. In addition, systems need stimulated annealing solution are usually high dimensions. The high dimensionality causes the ground state to be highly degenerated. However, this pose no real problem to the statistical analysis. For the β-sheet crystal we studied here, the population in the model is around 5000. As this population, the ground energy state would be highly degenerated. In addition, we are looking the low-q range of structure factor, which mean only the long range ordering the crystals will affect the simulated structure factor curves. The resulting map for different trails all shows clustering effect and the characteristics are similar in terms of cluster size, correlation length. After calculating the pair distribution function, we will see that the correlation peaks are consistent without large discrepancy. Thus we can safely conclude that the stimulated annealing reachs a good ground state statistically.

Chapter 3

Self-similarity and mass fractal nano-crystal network in spider silk fiber

3.1 Introduction

Spider dragline silk fibers have excellent reversible extensibility and high tensile strength. Römer and Scheibel (2008); Lewis (2006); Vollrath (2000) Extensive studies on the molecular structure of spider silk have been performed over the years and it's widely believed that oriented anti-parallel β-sheet nano-crystals are the key contributor to spider silk's excellent mechanical properties. Keten *et al.* (2010) The molecular structure and chemical composition of spider silks have been studied by solid state nuclear magnetic resonance spectroscopy and these studies have shown that a large fraction of the amino acid sequences are poly-(Gly-Ala) and poly-Ala repeats, Jenkins *et al.* (2013); Asakura *et al.* (2013); Hayashi *et al.* (1999); Xu and Lewis (1990) which form rigid anti-parallel β-sheet nano-crystals through the periodic hydrogen bond assemblies. Keten *et al.* (2010); Keten and Buehler (2008, 2010) The less ordered amino acids are in the form of random-coil like helical secondary structures in which the β-sheet nano-crystals are embedded. The elastic and random-coil like α- and 3_{10}-helical structures are abundant and occupy a large volume of the fiber body, acting as the interconnections among the rigid crystals. Holland *et al.* (2013) The crystal

structure and physical size of individual β-sheet nano-crystal has also been resolved by X-ray diffraction studies. Past WAXS studies, as well as this work, have confirmed that typical β-sheet crystals have an orthorhombic unit cell with their physical sizes ranging from 2 to 4 nm, when produced at the natural extrusion speed. Riekel and Vollrath (2001); Sampath *et al.* (2012)

While the nano-scale dimensions of the β-sheet crystals have been well studied, their hierarchical structures and relation to the macroscopic mechanical properties are still not fully understood. Being the most rigid objects in the spider dragline silk, the β-sheet crystals play a crucial role in determining its mechanical properties. If we assume the β-sheet crystals to be the building blocks of the cylindrical-shape fiber, then its physical size, inter-crystal distances and long-range packing pattern will be the key parameters that define the macroscopic mechanical properties. Fortunately, several microscopy studies have gained insight on the crystallite structures of the spider silk fibers. Scanning electron microscopy (SEM) studies have shown that the texture of ion-etched silk fiber's is rather rough, as scattered crystalline-rich regions about the size of 20 nm to 50 nm have been observed across the silk fibril.Kitagawa and Kitayama (1997) Transmission electron microscopy (TEM) image shows relatively large crystallites on the scale of 70 to 120 nm are embedded in the amorphous matrix.Drummy *et al.* (2007); Trancik *et al.* (2001) To date, there is still no consistent model to describe the hierarchical structure of these large crystalline regions, as the dominant scattering centers, i.e. β-sheet crystals, only span several nanometers in all three dimensions. By analyzing of the SAXS structure factor using a stochastic reconstruction method, we provide evidence that the crystalline structure of spider dragline silk fiber is mass fractal, accompanied by dense clustered packing of the nano-crystals. The 'large crystals' that span up to 70 nm are composed of highly oriented and closely interlinked β-sheet crystals.

In addition, the strong 'lamellar' peaks observed in the low-q range of the structure factors is a distinct feature to the spider silks, which separating them from other types of fiber such as silkworm silk. This feature is a direct evidence of strong crystalline ordering on the ten to several hundreds of nanometers range scale in spider silk. While this feature is common in spider silks, we don't observe similar pattern in silkworm silk fiber, as shown Martel et. al. Martel *et al.* (2008). We will show that this feature combining other characteristics of the structure factor will lead to clustered nano-crystal morphology and mass fractal hierarchical structure in spider silk. The exceptional mechanical property of spider silk fibers is built upon this hierarchical structure and it makes spider silk fiber superior than other types of biopolymer fibers, such as the silkworm silk fiber that doesn't exhibit 'lamellar' peaks in SAXS structure factor.

3.2 X-ray experiments

Major ampullate dragline silks were collected by forced silking from living spiders anesthetized with carbon monoxide. The silks were reeled at 2.0 ± 0.1 cm·s^{-1} and were directly mounted across hollow cardboard holders driven by an electric motor. At these silking speeds, β-sheet crystals are supposed to be extremely well oriented. Du *et al.* (2006) The samples normally have 50 to 100 strands of fibers.

The X-ray experiment was carried out at Argonne National Laboratory Advanced Photon Source BioCars 14-ID-B beam line optimized with a SAXS setup, as shown in Fig.3.1. The incident beam energy was 9 KeV, corresponding to an X-ray wavelength of 1.38 . The scattering data were collected on a Mar165, a 3072 by 3072 pixel resolution CCD detector from Marresearch with a pixel size of 79 ţm. The detector to sample distance is fixed at 180 mm throughout the experiment. The sample was mounted to an xyz-translation goniometer head with helium gas chamber placed

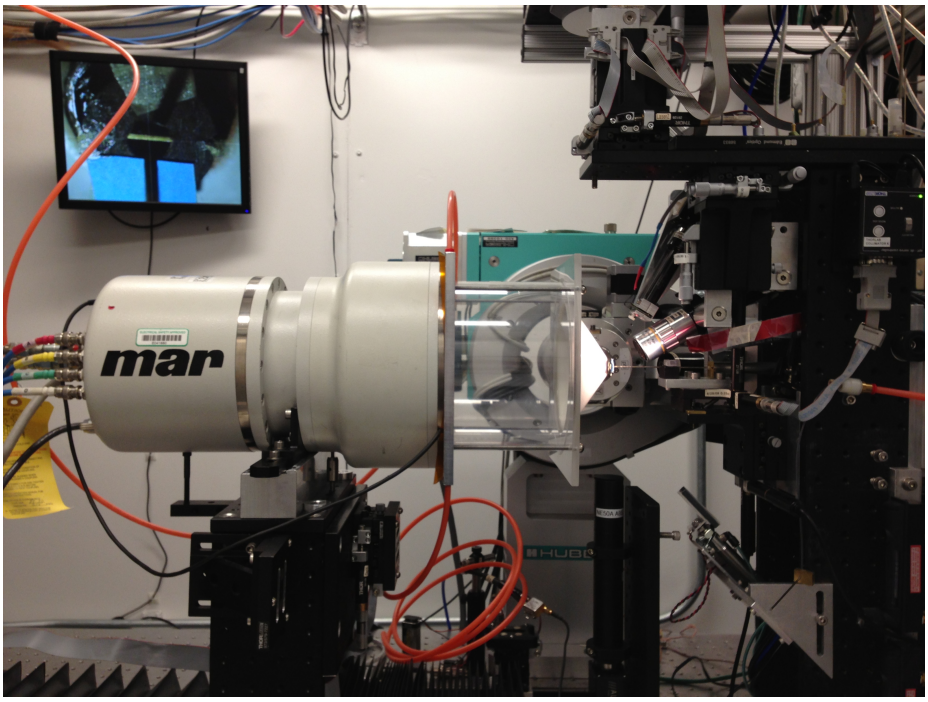

Figure 3.1: The detector is Mars165 CCD from Marresearch. The sample is loaded onto a hollow card board holder in front of a helium cone, which is used to reduce background noise signal.

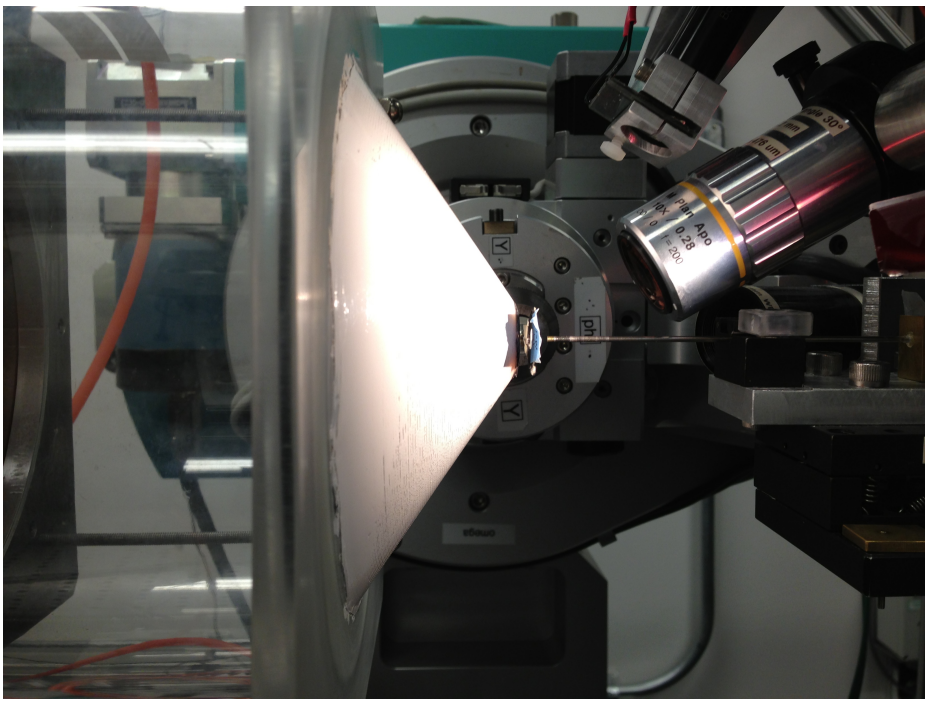

Figure 3.2: A closeup view of the sample holder, which is mounted on the XYZ goniometer head.

between the CCD detector and the sample to help reduce the background signal (Fig.3.2). The exposure time varied from 2 s to 6 s depending on the size of the silk bundle. For each sample ten exposures were taken. Backgrounds were measured after translating the silk bundle out of the X-ray beam.

The multiple X-ray exposures were averaged and background subtracted by using software *Fit2D*. Hammersley *et al.* (1996) The integrated 1-D WAXS and SAXS profile were obtained by using azimuthal integration functionality in *Fit2D*. The automatic Gaussian peak fitting was performed using the *lsqcurvefit* optimization routine from MATLAB.

3.3 Simulation

The transformation from electron density map to scattering intensity adopts the fast Fourier transform (FFT) simulation method proposed by Klauss Schmidt-Rohr.Schmidt-Rohr (2007) The electron density map $\rho(\mathbf{r})$ was represented by an $N \times N$ matrix, where N is usually chosen to be a power of 2, and then converted to a reciprocal space scattering intensity $I(\mathbf{q})$ by the transformation algorithm. For each sample, we initialized the silk structure such that it conforms to the crystal sizes calculated from the WAXS data and maintained the closest approach distances, which is the d-spacing of SAXS lamellar peak (Tab.1). The scattering centers, which are the β-sheet crystals in this model, were represented by higher contrast rectangular shape sub-matrix and are uniformly oriented to be parallel to the silk fiber axis. The orientation of the crystals at >2.0 cm$\cdot$s^{-1} can be approximated by perfect alignment with very small error. van Beek *et al.* (2002); Du *et al.* (2006) A hard-shell exclusion geometry was used to maintain the closest approach distance.

We imposed a model with a lamellar modulation by generating 60 nm wide crystal-rich stripes separated by equal width empty spaces. The lamellar structure is essential

to physically represent the silk fibrils fine structure within the silk fiber. Kitagawa and Kitayama (1997); Du *et al.* (2006) The initial crystalline map was then fed to the stimulated annealing reconstruction routine. The reconstruction was proceeded by generating a electron density map such that the calculated structure factor $\hat{S}(q)$ matches the experimental $S(q)$ with acceptable error tolerance. This was achieved by minimizing the pseudo-energy

$$E_{pseudo} = \sum_{q} |S(q) - \hat{S}(q)|^2 \tag{3.1}$$

which measures the distance from simulated structure factor to the experimental value McGreevy and Pusztai (1988); Kaplow *et al.* (1968); Jiao *et al.* (2009). The optimization was realized by the stimulated annealing algorithm where each random walk is accepted or rejected by a probability of Kirkpatrick *et al.* (1983); Lenstra (2003); Rubinstein and Colby (2003)

$$P(s' \leftarrow s) = min(1, exp(-\frac{\Delta E_{s' \leftarrow s}}{k_B T})) \tag{3.2}$$

where s and s' are the states before and after one random walk, $\Delta E_{s' \leftarrow s}$ is the change of pseudo-energy after accepted state transition, k_B is the Boltzmann constant, T is the imaginary stimulated annealing temperature.

Table 3.1: Lattice parameters, nano-crystal sizes and inter-crystallite distance.

Lattice parameters[1]				From (200)[2]			From (120)			d-spacing[3]
Fiber	$a()$	$b()$	$c()$	σ	2θ	$\tau_1()$	σ	2θ	$\tau_2()$	$d()$
L. hesperus	10.5	9.6	6.9	1.102	15.18	21.3	0.824	18.24	26.4	77.1
N. clavipes	10.7	9.7	6.9	1.073	15.21	20.3	0.960	18.26	22.7	83.2
A. aurantia	10.5	9.7	6.9	1.107	15.06	19.6	0.933	18.13	23.4	108.7
A. gemmoides	10.6	9.7	7.1	1.023	15.09	21.3	0.902	18.10	23.7	106.6

[1] Based on orthonormal basis. The error is one unit in the last digit quoted.

[2] The physical dimension of β-sheet τ are calculated from the FWHM of WAXS reflections.

[3] Directly derived from SAXS structure factor.

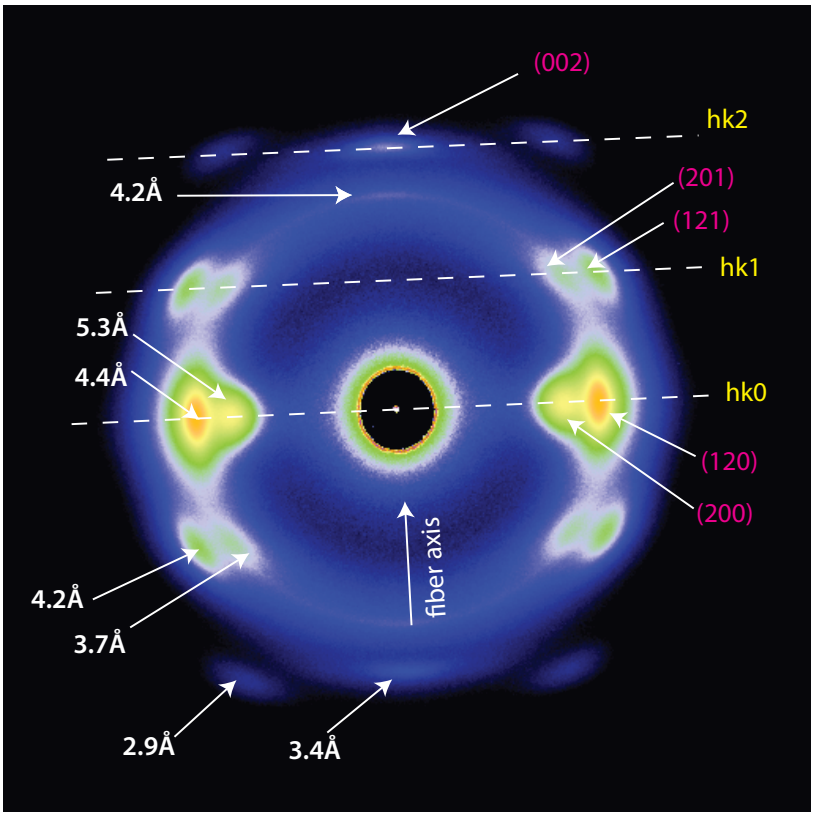

Figure 3.3: The scattering image of *L. hesperus* (Black Widow) major ampullate (dragline) silk.

3.4 WAXS results

We first performed a series of WAXS experiments on different species of spider silk samples to study the crystalline structure of the β-sheet unit. WAXS profile were analyzed by Fit2D software and the scattering pattern were integrated to obtain the WAXS pattern, as shown on Fig.3.4. From observing the WAXS pattern, we found that the micro-structure of β-sheets in different silk samples are very similar in terms of spot pattern and d-spacing. After obtaining the WAXS statistics, as shown in Table.3.1, we can say that the differences among the β-sheets are minimal and the different mechanical properties exhibited by these samples can't be explained by the β-sheet itself along. Therefore we conducted SAXS experiments subsequently on the same set of silk fiber samples.

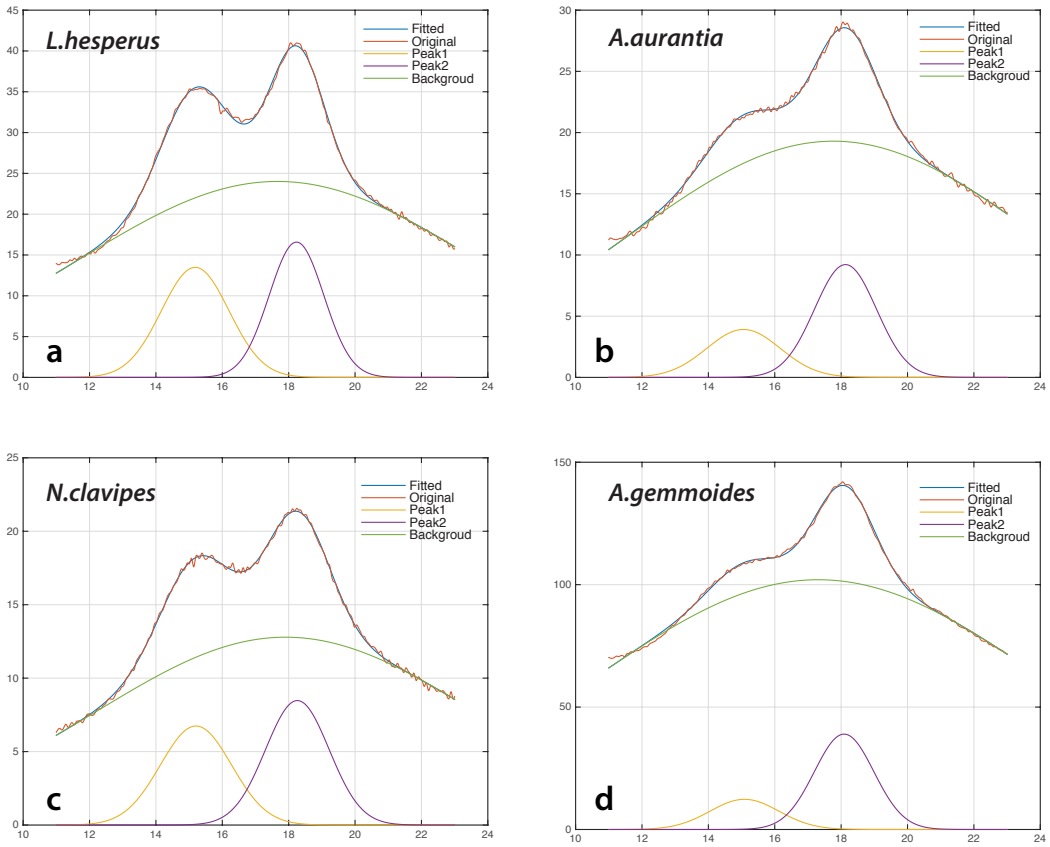

Figure 3.4: The 1-D profiles are obtained by azimuthal integrations. The integration region is chosen such that the dominant (120) and (200) reflections are completely enclosed. The 1-D WAXS structure factors were processed by multi-Gaussian curve fitting. The initial parameters were estimated from visual inspection and then supplied to MATLAB script for curve fitting. The routine used in the fitting procedure is lsqcurvefit.

Fig.3.6 shows the WAXS and azimuthal integrated SAXS profile of *L. hesperus* (Black Widow) dragline silk fiber. The diffraction pattern is divided into two distinct regions: the center small-angle region ($q<1.3$ nm^{-1}) and the outer wide-angle region. As shown in Fig.1a, the diffraction spots have been assigned with Miller indices from which we have calculated the unit cell parameters basing on an orthorhombic unit cell model. Sampath *et al.* (2012). As shown in Fig.3.3, the WAXS scattering image of sample *Black Widow* has been properly labeled with Miller indices.

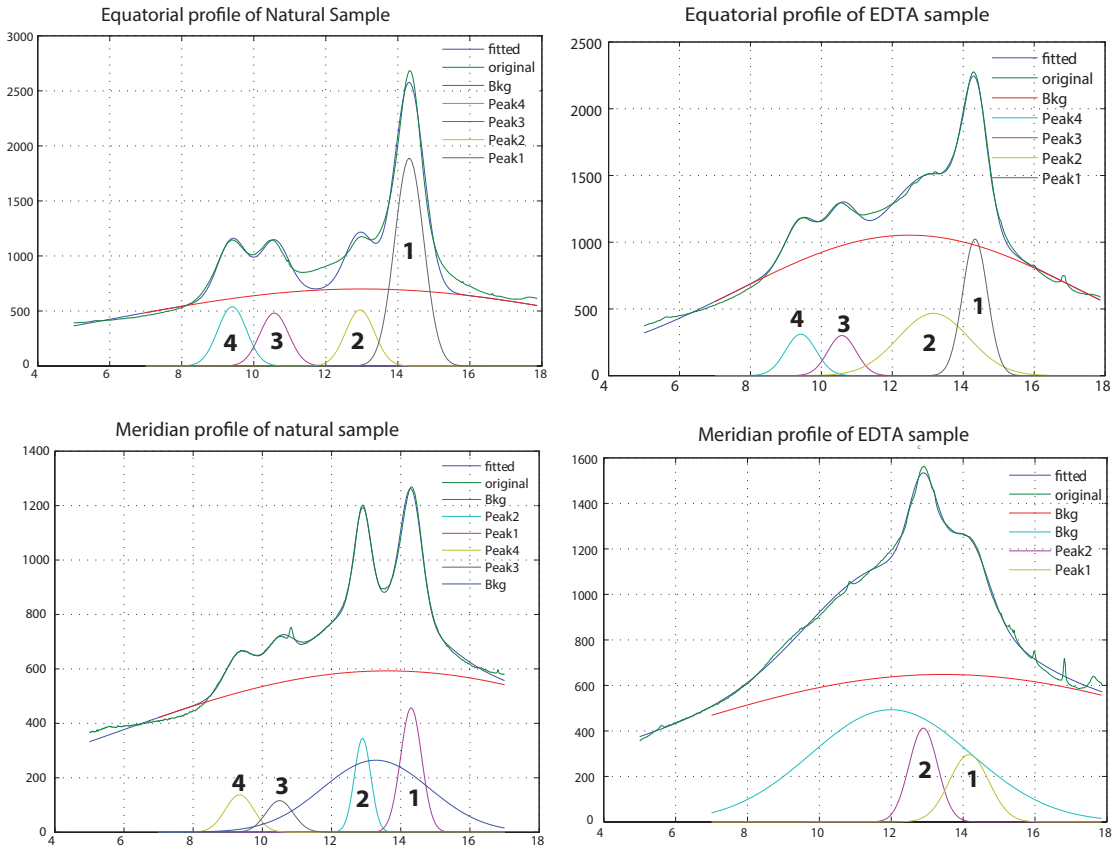

Figure 3.5: First of all, we notice the increased background after EDTA treatment of sample A. Second, Peak3 and Peak 4 intensity decreased by 40%. Peak1 intensity decreased by 50%. According to the model, the extraction of cation by EDTA possibly caused more disorder in Beta- sheet structure. Thus the diffraction from two type inter-sheet distance structure (peak 3 & peak 4) were weakened. Since the Phosphorus group became more mobile, the corresponding diffration (peak 1) was weakened too. The bottom 2 figures: Comparision of Meridian profile. The curve is Meridian wedge integration over 2-theta space.

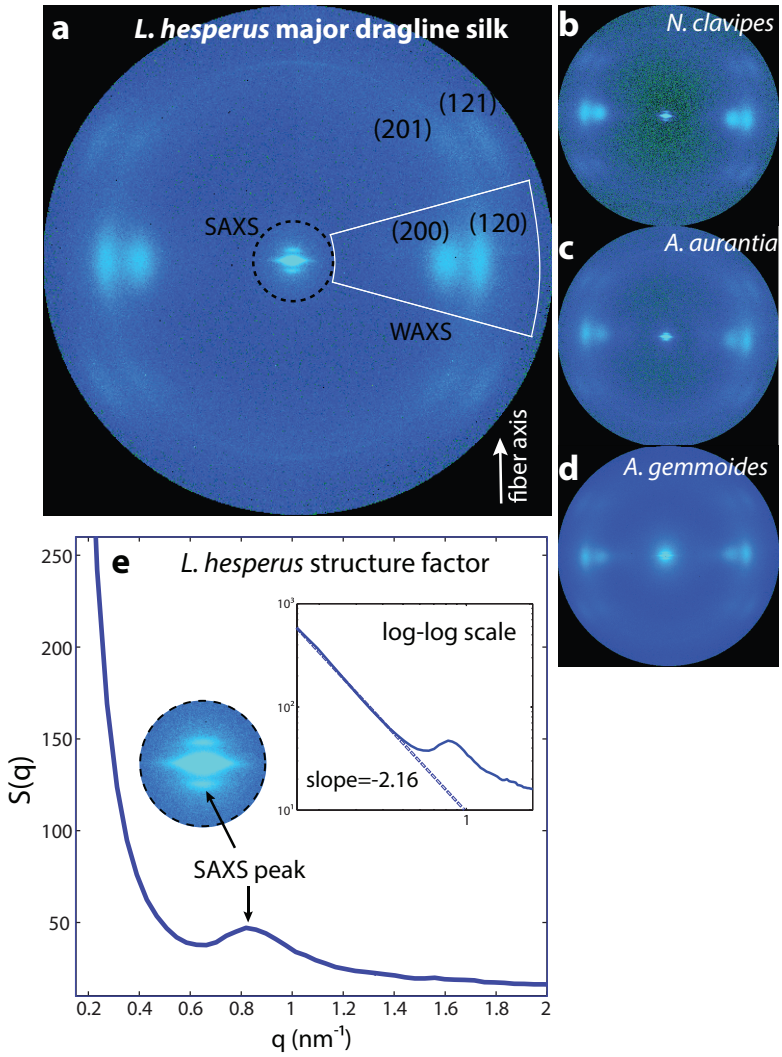

Figure 3.6: (**a**) WAXS pattern of *L. hesperus* (Black Widow) major ampullate (dragline) silk. The wide-angle diffraction spots have been assigned with Miller indices, from which the unit cell parameters have been calculated using an orthorhombic crystal model. Gaussian peak fitting was applied to the wedge shaped region (white line) and the Scherrer equation was used to determine the average nano-crystallite dimensions. The center region ($q<1.3$ nm^{-1}) corresponds to the SAXS scattering pattern. (**b-d**) Samples of *N. clavipes*, *A. aurantia* and *A. gemmoides* show similar wide-angle scattering patterns. (**e**) Azimuthal integration of scattering intensity from *L. hesperus* dragline silk fibers. The lamellar peak is located at $q=0.82 \pm 0.01$ nm^{-1}. The inset shows the SAXS structure factor $S(q)$ on log-log scale, where the 'matrix knee' represents the intermediate length scale (1 nm to 200 nm) and exhibits linearity with a slope of 2.16.

To retrieve the crystal sizes from X-ray data, we integrated the wedge shape equatorial region containing reflections (200) and (120) (Fig.3.6a), and then applied a 1-D Gaussian peak fitting procedure [see Fig.S1]. The full maximum at half width (FMHW) of the fitted Gaussian peaks are evaluated using the Scherrer equation Patterson (1939)

$$\tau = \frac{K\lambda}{\beta \cos\theta} \tag{3.3}$$

where $K = 0.9$, $\lambda = 1.38$, β is the FWHM of fitted peak and θ is the Bragg angle, to calculate the nano-crystal physical dimension τ. The orthorhombic unit cell and crystal sizes are summarized in Table.3.1. The crystal size in all three dimensions (a, b, c) shows very small variations for the four species of dragline silk examined here. The crystal sizes have a narrow distribution, namely from 19.6 to 21.3 for that calculated from (200) reflection and from 22.7 to 26.4 for the (120) reflection. Although different species of dragline silk shows significant differences in alanine content and mechanical properties, Jenkins *et al.* (2013) their building blocks are surprisingly similar in terms of physical appearance. Fig.1e shows the small-angle scattering profile of *L. hesperus* (Black Widow) major ampullate (dragline) silk fibers, where the characteristic lamellar peak is located at 0.82 nm^{-1}. The characteristic correlation peaks between 0.5 and 1.2 nm^{-1} have been observed frequently in polymer and copolymer materials. Yarusso and Cooper (1983) The correlation peaks manifest certain long-range ordering of the crystalline phases and such ordered state is crucial to the functional behavior of the material, such as water channels morphology observed in nafion polymers. Schmidt-Rohr and Chen (2008) Therefore, we propose that it's the intermediate length-scale morphology of the β-sheet crystals that determines the macroscopic properties such as mechanical strength, elasticity and thermal conductivity.

3.5 SAXS results

Since β-sheet crystals are identified as the only kind of strong scattering centers in spider dragline silks, it's natural to assume β-sheet crystals as the origin of the SAXS lamellar peaks. The packing pattern of the crystals has unique intermediate-range ordering, which leads to the appearance of strong lamellar peaks in the $q < 1$ nm^{-1} range of the SAXS structure factor. Schmidt-Rohr (2007); Pedersen (1994) For strong correlation peaks to appear, it's important that the scattering centers maintain a closest approach distance R_{CA} between each pair, as has been discussed by Yarusso and Copper. Yarusso and Cooper (1983) In real space, the closest approach distance R_{CA} limits the average spacing between each adjacent pair of scattering centers that reside in the amorphous backbone of biopolymer. Reflected in reciprocal space, the combination of closest approach distance and long range ordering will generate correlation peak in the small-angle scattering regime. For different species of spider dragline silks examined here, all of them exhibit correlation peaks in the range of 0.6 to 0.9 nm^{-1}. This indicates that the β-sheet crystals are not adjacent to each other statistically, but rather maintain an average closest distance among them, which can be quantified by the characteristic of SAXS lamellar peak. The SAXS signal exhibits an elliptical streak, elongated along the meridian direction. This indicates that the β-sheet nano-crystals have a rectangular shape with long axis parallel with fiber axis, as illustrated by Fig.3.7 in supplementary material. Combining these information, we constructed a 2-dimensional electron density map in which the beta-sheet crystals are initialized with constrained orthorhombic geometry. The d-spacings of the lamellar peaks were taken as the average inter-crystal distance in constructing the initial model. This in turns determines the density of the β-sheet crystals in each silk species. The linear correlation between the lamellar peak position and alanine

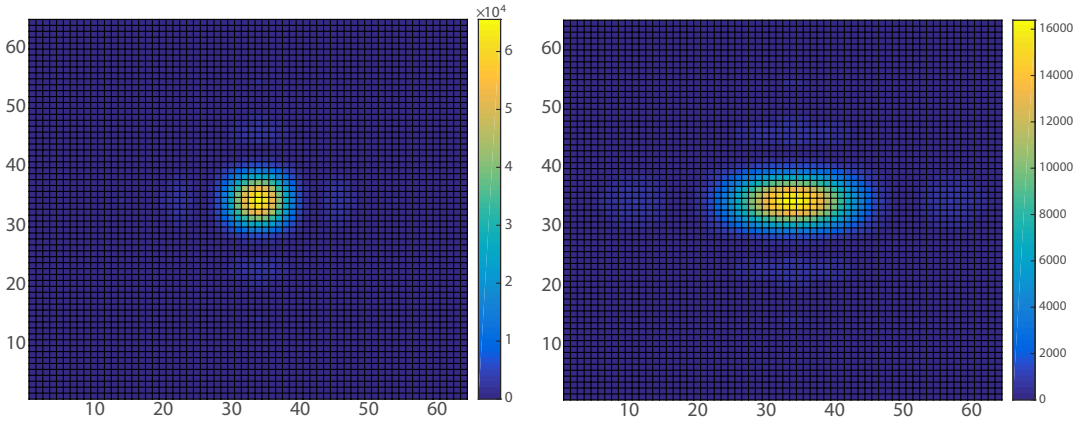

Figure 3.7: (Left) Cubic crystal geometry. (Right) Rectangular crystal geometry with long edge parallel with fiber axis. The elongated crystal shape is consistent with the SAXS pattern observed in spider dragline silk sample. The graphs were scaled to have the same unit length in 2D space.

content further strengthens this assumption.

Past solid state NMR experiments have shown that the fraction of alanine content, which primarily occur in the β-sheets, vary significantly among different species. Jenkins *et al.* (2013); Creager *et al.* (2010) By comparing the data from x-ray scattering and NMR, we have found the alanine content to be linearly correlated to lamellar peak's position in q-space, as shown in Fig.3.8. With a relatively small inter-crystal distance, such as 83.2 and 77.1 measured from *N. clavipes* and *L. hesperus* respectively, we find that the β-sheet crystals are more densely packed and thus leading to a higher fraction of beta-sheet crystals than the other two samples, namely *A. aurantia* and *A. gemmoides* (Fig.3.14).

The experimental and numerically simulated SAXS structure factors $S(q)$ are shown in Fig.3.9a. The lamellar peaks are present in all fiber samples and they range from 0.6 to 0.9 nm^{-1} with variations in the peak intensity. The lamellar peaks of *A. aurantia* and *A. gemmoides* are relatively close in q-space with a minuscule difference Δq=0.01 nm^{-1}, while *L. hesperus* and *N. clavipes* are slightly further apart with a

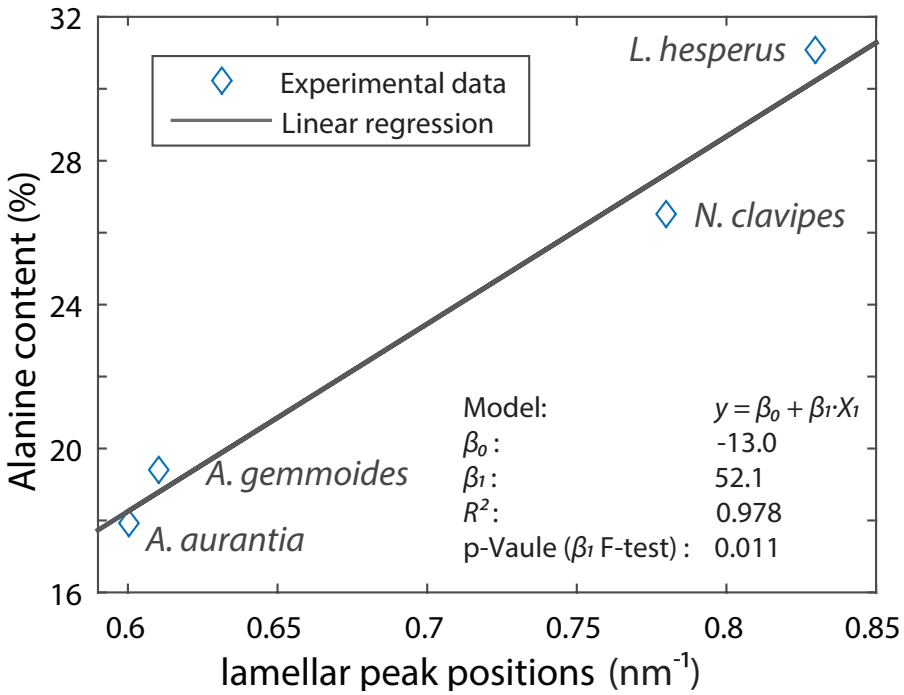

Figure 3.8: Alanine in spider silks is shown to primarily occur in β-sheet nanocrystals and an increase in alanine content in the spider silk protein is a contributing factor to increased crystallinity associated with the nanoscale clusters which give rise to the lamellar peaks. The alanine content data was previously retrieved from solid-state nuclear magnetic resonance experiments.

difference of Δq=0.05 nm^{-1}. The structure factor curves are very well reconstructed across the entire collected q-range by the stimulated annealing reconstruction method. The lower bound of the $S(q)$ is limited by the detector coverage on 14-ID-B beam line (Argonne National Laboratory, APS) while the upper bound is cut at q=1.2 nm^{-1}, beyond which point the structure factor $S(q)$ begins to exhibit WAXS feature. Fig.3.9b shows the pair correlation functions $P(r)$ calculated from the reconstructed electron density maps. The inter-molecular β-sheet pair correlation function $P(r)$ Lei *et al.* (2009); Proffen and Billinge (1999) is defined as,

$$P(r) = \frac{1}{2\pi r \rho_0} \frac{1}{N} \sum_{i=1} \sum_{j \neq i} \frac{w_i w_j}{\langle w \rangle^2} \delta(r - r_{ij}) \tag{3.4}$$

where $2\pi r \rho_0$ is the density of states in the 2D isotropic, homogeneous system, w_i is the weighting factor for the corresponding scattering center, and N is the β-

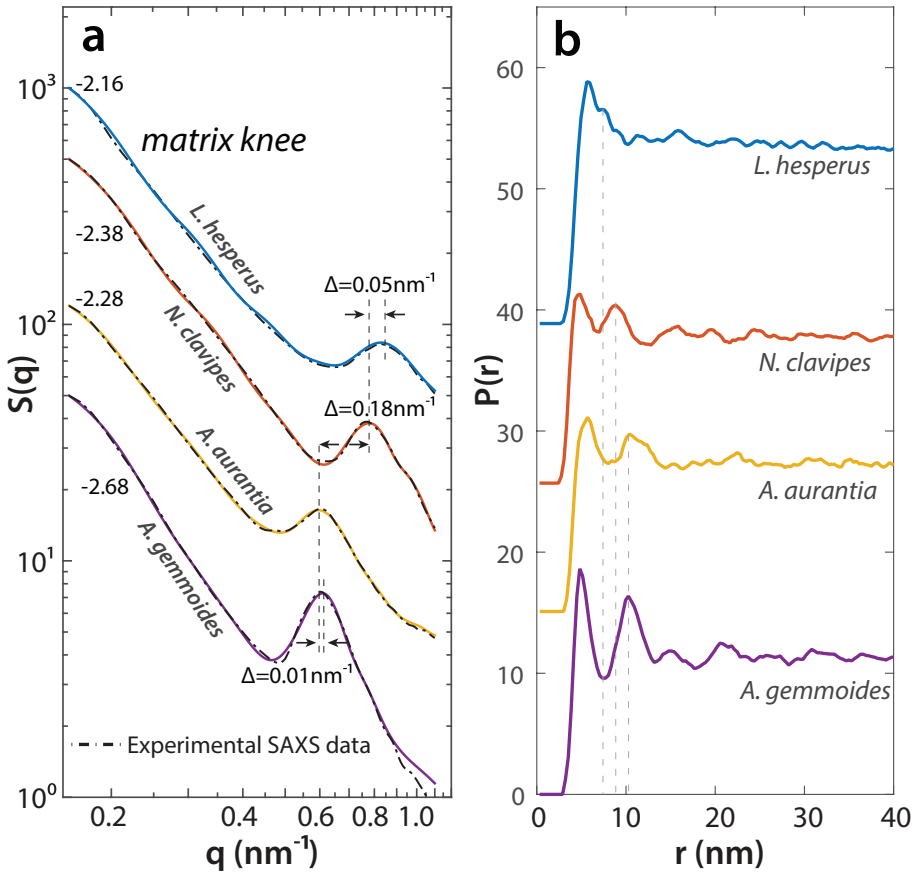

Figure 3.9: (**a**) The numerically simulated structure factor $S(q)$ (—) fits the experimental SAXS $S(q)$ (---). The small-angle lamellar peaks' positions, from *L. hesperus* (top) to *A. gemmoides* (bottom), are correspondingly located at 0.83, 0.78, 0.60 and 0.61 nm^{-1} with errors of approximately $\pm$0.01nm^{-1}. The low-q region of the structure factor $S(q)$, i.e. the matrix knees, exhibit linearity in all cases and the slopes range from -2.1 to -2.7 on the log-log scale. (**b**) Pair correlation function $P(r)$ calculated from the reconstructed electron density maps. The correlation function $P(r)$ were calculated from a population of around 5000 crystals and the curves were numerically smoothed. The intermediate range crystalline ordering is reflected as the multiple correlation peaks observed on the $P(r)$ curves in the range of 7 to 40 nm.

sheet crystal population. The intermediate range crystalline ordering is reflected as the multiple correlation peaks observed on the $P(r)$ curves in the range of 7 to 40 nm. The first correlation peaks appear near 6 nm and these peaks arise from the dominant closest pair-pair interaction of the β-sheet crystals. The second peaks are in the range of 7 to 10 nm, which arise from the intra-cluster interaction. As shown in Fig.3.9b, the second peaks is shifting right top-down, which follows the density and inter-crystal spacing constrain, though the exclusion region was reduced during the stochastic reconstruction to allow higher mobility (Fig.3.11). The intensity and sharpness of the lamellar peak reflects strength of pair correlations directly. While the *A. gemmoides* and *N. clavipes* have the stronger lamellar peak (Fig.3.9a), they also exhibit more correlation peaks in the > 10 nm range of their $P(r)$ function. This indicates that a larger fraction of the β-sheets are in long range ordered state for these two silk species.

While the geometry of scattering centers play an important role in determining the shape of SAXS pattern, we tested and visualized the SAXS pattern from different basic geometries. Fig.3.7 shows the comparison between cubic shape geometry and elongated rectangular geometry. The elongated shape results an elliptical streak pattern in the center SAXS region, which is more consistent with the experimental SAXS pattern observed from spider dragline silk samples. In this elementary model, the long edge of the rectangular crystal is parallel to the imaginary fiber axis, thus indicating that the actual beta-sheet crystals are arranged in this biased configuration. With a single crystal configured in the elementary model, we can already observe the presence of a weak lamellar peak on both meridian and equatorial direction. Intuitively, the strong meridian lamellar peaks observed from SAXS experiments result from the phase-dependent superposition of a large population of beta-sheet crystals in the dragline silk fibers, therefore rendering the Monte Carlo reconstruction simulation

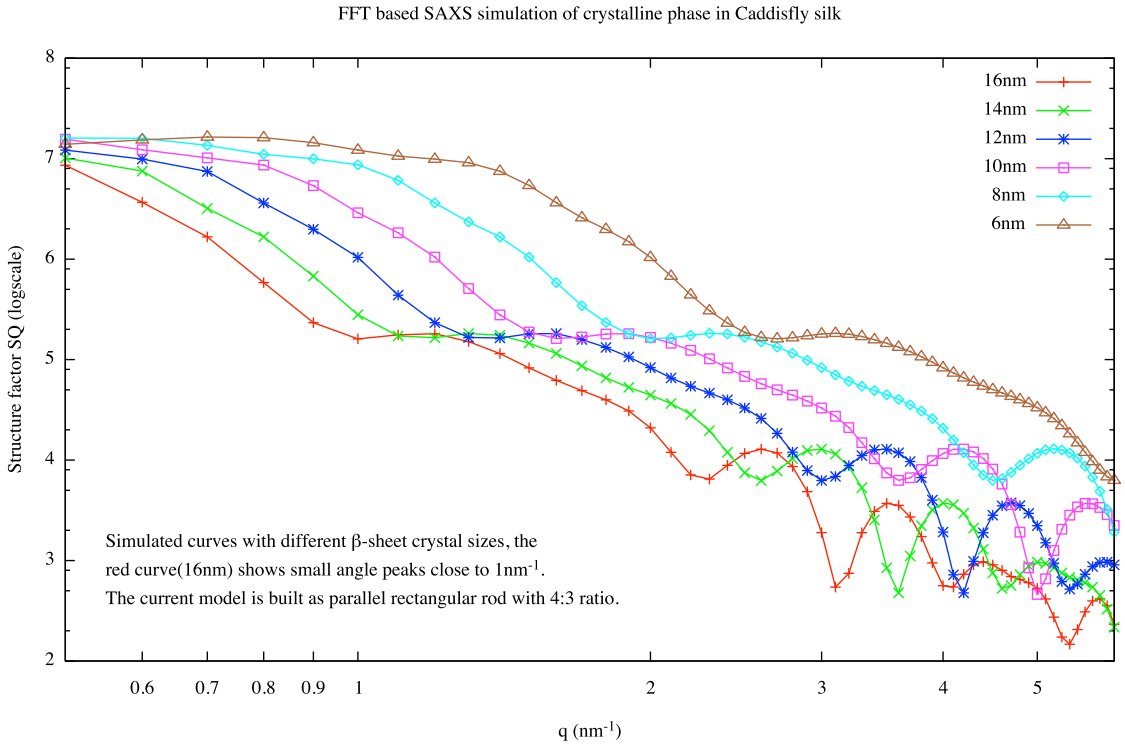

Figure 3.10: The location of SAXS lamellar peak depend on the average size of the nano-crystal. Each curve is calculated from randomly generated nano-crystals with similar size.

a necessary step to retrieve crucial structural information from the SAXS data.

In addition, the average size the β-sheet crystals do affect the position of the SAXS peaks in q-space. As shown in Fig.3.10, we conducted a series simulation with increasing average size of the β-sheet crystals. The crystal sizes range from 6 nm to 16 nm. As the size increases, the plateau part of the curve shift to low-q range. Note that for this simulation, all the crystals were initialized randomly and there was no stimulated annealing optimization followed. So the curve here doesn't resemble the ones from SAXS data.

Fig.3.12 shows the evolution of the stimulated structure factor and the pair-correlation function calculated from the reconstructed crystalline model. The correlation function shows no feature in the early stage of the reconstruction, mainly

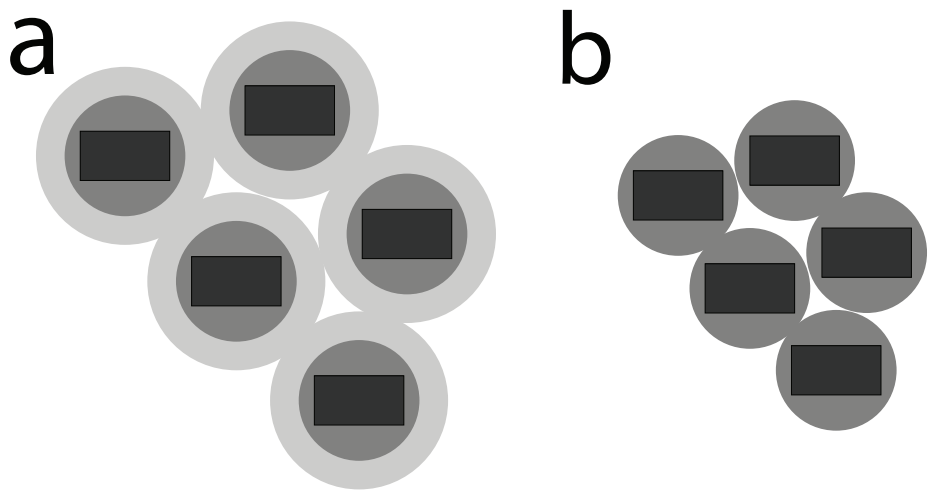

Figure 3.11: The black rectangles are beta-sheet crystals. The light shade circular region is the exclusion region used in building initial model. The exclusion region during stimulated annealing reconstruction process is reduced (darker shade circular region) to allow higher mobility of random walk, thus significantly improves the efficiency of the algorithm.

due to the randomized landscape of the crystals and the relatively small population, which varies between 4000 to 7000. As the structure factor converges to the experimental value, the correlation peaks gradually build up, notably for the peaks at 10, 21, 31 and 42 nm, indicated by the dashed line on Fig.3.12b. This dynamic proves that one can reconstruct the pair correlation function through reciprocal-space reconstruction and it supports the belief that the SAXS lamellar peaks, i.e. the correlation peaks between the q range of 0.6 to 0.8 nm^{-1}, arise from the intermediate length scale ordering of the β-sheet crystals.

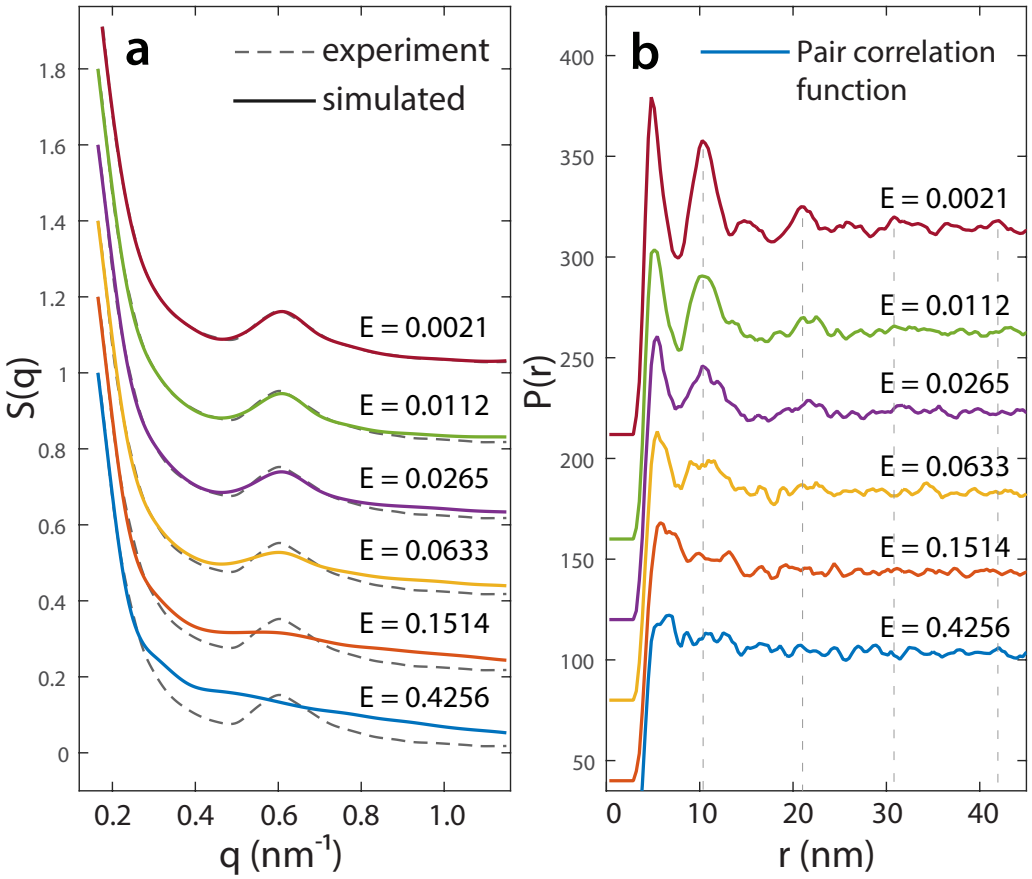

Figure 3.12: (**a**) From bottom to top, as the stimulated annealing temperature T drops, the calculated structure factor (—) converges to the experimental structure factor (---), reducing the pseudo-energy E as defined in Eq.**??**eq:engy. (**b**) Initially at $E = 0.4256$, the $P(r)$ function is absent from any correlation peak. As the stimulated annealing algorithm proceeded, the model build up intermediate range crystalline ordering which was reflected by the correlation peaks at 12, 15, 21, 26, 32, 37 and 42 nm on the top curve ($E = 0.0021$) . The correlation function $P(r)$ is sampled from a model contains 4721 β-sheet crystals.

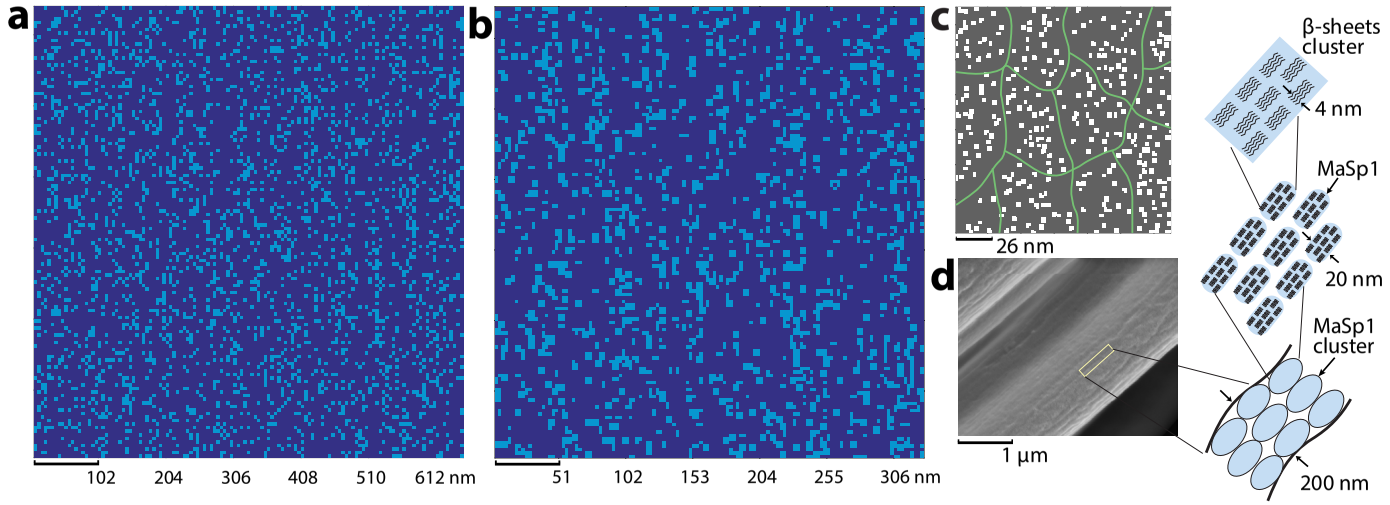

Figure 3.13: (**a-c**) Coarse grained electron density map from high to low degree of coarse graining. Each map shows the magnified upper right quater of the previous. The texture on these three maps with drastically different length scales show self-similarity property. The clustering effect of β-sheets is visually emphasized by the boundaries in (c). (**d**) The schematic illustration of the hierarchical mass fractal structure of the β-sheet clusters in silk fiber, from SEM image to the basic β-sheets network. The micro-fibril ($>$ 200 nm) is composed of functional mechanical units, and each unit (50 - 100 nm) is composed of multiple crystallites-rich MaSp1 proteins ($>$20 nm). Within a MaSp1, it contains multiple β-sheet nano-crystals (2-4 nm). Such fractal structure is mechanical robust and exhibits non-linear force-extension behavior.

The coarse-grained electron density maps are shown in Fig.3.13(a-c) for sample*A. gemmoides*. The electron density maps for the other samples are shown in Fig.3.14. The lighter area indicates the presence of a higher density of β-sheet crystals in that region whereas the darker area represents the amorphous backbone, which represents diffuse X-ray scattering. The β-sheet crystal distributions are initialized with a lamellar modulation when building the initial model. The modulation, which typically has a strip size of $N_{lamellar} = 128$ equaling to a physical size of 30 - 50 nm, is designated to represent the micro-fibril structure observed from both SEM Kitagawa and Kitayama (1997) and AFM experiments. Du *et al.* (2006) The existence of the parallel lamellar nano-fibrils could explain the discrepancy between the spider silks' axial and radial sound velocities reported by Koski et. al. Koski *et al.* (2013) Along the axial direction of fiber, these parallel lamellar fibrils contain a high density of β-sheet crystals with crystalline ordering up to 30 nm, which can be observed on Fig.3.13b. The highly ordered and continuous lamellar backbones act like phonon highways and thus support fast propagation of both longitudinal and transversal acoustic waves in the spider dragline silks, as observed in the recent Brillouin scattering measurements. Koski *et al.* (2013) On the other hand, due to the modulation of the lamellar structure, the crystalline ordering vanishes and the β-sheet crystal structure has a zero-density discontinuity along the radial direction. Consequently, the scattering of phonons is stronger and the measured velocities of the acoustic wave are much lower than that along the axial direction.

The other prominent structural feature is the clustered packing of the β-sheet crystals. During the reconstruction process, β-sheet crystals self assemble to form high density crystalline-rich islands. The clustering effect exist in all length scales examined here and the cluster sizes change from 30 nm in Fig.3.13c to 50 nm in Fig.3.13b, and to 100 nm in Fig.3.13a, increasing exponentially in accordance with coarse grain-

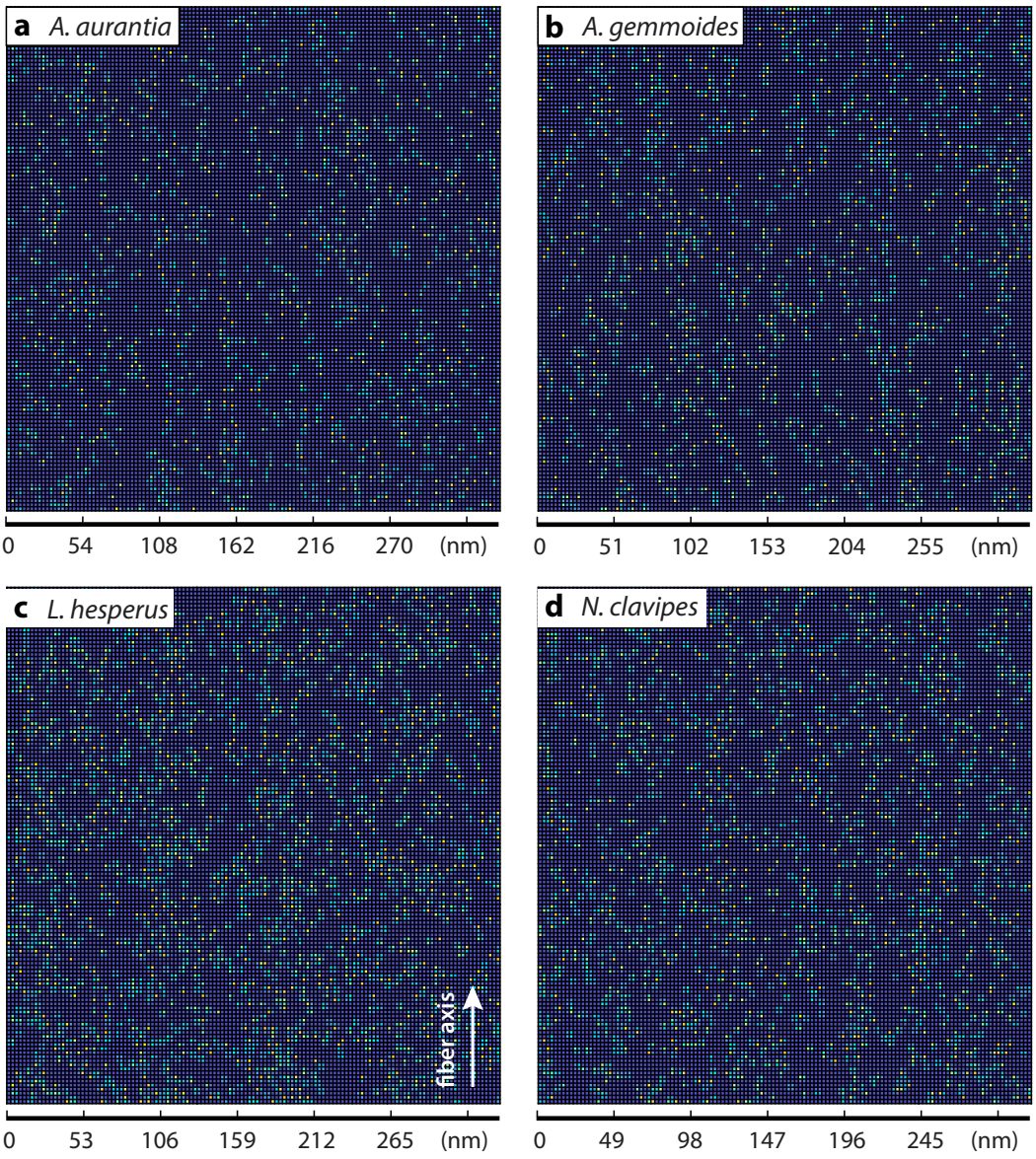

Figure 3.14: The density of the crystals is determined by the average inter-crystal spacing, which is calculated from the location of lamellar peaks in the SAXS structure factors. All maps show clustering of the beta-sheet crystals, which result from the power-law and the lamellar peak observed in SAXS structure factor.

ing. The reconstructed electron density map shows remarkable self-similarity property and is scale invariance statistically. The formation of these clusters is driven by both the characteristic of the matrix knees and the presence of the lamellar peaks. The SAXS structure factor has power law relation with respect to the scattering vector q

$$S(q) \propto q^{-r} \tag{3.5}$$

and $-r$ is the slope of the matrix knee in the log-log plot of $S(q)$ (Fig.3.6e). The parameter r is the fractal dimension in system exhibiting self-similarity. Pedersen (1994); Teixeira (1988); Martin and Hurd (1987) For the silk fibers examined in this study, the r is between 2.16 and 2.68, which means the crystalline structure has a mass fractal ($2 < r < 3$) Schaefer (1989); Stanley (1984) Crystal clustering is the dominant form of mass fractal and therefore the clustering effect can be observed at drastically different length scales. Now looking back at the pair correlation function (Fig.3.9b), the correlation peaks beyond 10 nm scale should arise from the inter-cluster interaction. No matter how strong the lamellar peak is, the mass fractal accompanied by clustering is an universal property in spider dragline silk, manifested by the power law of structure factor and the reconstructed electron density map (see Fig.3.14).

The mass fractal property is significant to the mechanical strength of spider silks. The nano-crystal clusters coming with different sizes are the basic functional mechanical units in spider silk fiber, as illustrated by Fig.3.13d. More importantly, these units are interlinked by random-coil like helical secondary structures, forming self-similar and robust crystalline network at different length scales. When the silk experiencing an external force, the largest cluster of the size 100 nm will deform to respond to the external kick. Then the deformation is propagated down to the smaller clusters which will deform in response. The force-induced deformation is propagated down

in this manner all the way to the β-sheets level, causing an exponentially increasing number of mechanical perturbations as the length scale shrinks. This scheme is highly consistent with the model proposed by Zhou & Zhang Zhou and Zhang (2005), which effectively explained the exponential force-extension property of spider silks.Becker *et al.* (2003) For the questions imposed by the large crystalline regions observed in the SEM and TEM images, we suggest that these 20 - 50 nm granules are not single crystal structures but rather high crystallinity regions, which contains densely packed and strongly interconnected β-sheet crystals of the sizes ranging only from 2 to 4 nm.

3.6 Conclusion

A combined WAXS and SAXS study of spider dragline silk fibers is presented and interpreted here. The analytical results show that the intermediate length scale ordering directly leads to the lamellar peaks observed in SAXS. The power law observed in SAXS structure factor indicates that the β-Sheet nano-crystallites in spider silks are mass fractal. The reconstructed electron density maps provide a direct visualization of the clustering and self-similarity effect in the crystalline structure. The axial lamellar backbone structure observed in the electron density maps helps to explain the large discrepancy between the axial and radial sound velocity in spider silks measured by Brillouin spectroscopy.Koski *et al.* (2013) The mass fractal and nanocrystal clustering property is fundamental to understand the exceptional mechanical performance of spider silks. In practice, it would be challenging to synthesize such complicate yet elegant nano structure, but we beleive it's achievable through molecular self-assembly and thermal dynamic control. This study provides useful data and insightful analysis to guide future engineering of high performance silk.

Chapter 4

Extract inter-molecular structure factor through numerical optimization

4.1 Introduction

Total X-ray scattering experiments on molecular liquids and amorphous solids measure a structure factor containing contributions from overlapping intra-molecular and inter-molecular interactions. The total X-ray scattering structure factor of a molecular liquid can be written as Narten and Levy (1971):

$$S(q) = \frac{1}{N} \frac{\sum_{i,j=k}^{N} \sum_{\alpha,\beta=1}^{m} f_{i,\alpha}(q) f_{j,\beta}(q) \frac{sin(q \cdot r_{\alpha\beta})}{q \cdot r_{\alpha\beta}}}{(\sum_{\alpha=1}^{m} f_{\alpha}(q))^2} \tag{4.1}$$

For molecular liquids, we can separate equation 4.1 into the contributions arising from intramolecular and intermolecular (molecule-molecule interactions) scattering as follows Narten (1977); Narten and Habenschuss (1984):

$$S(q) = S_{intra}(q) + S_{inter}(q) \tag{4.2}$$

A multi-parameter data fitting is required to separate out the intra-molecular scattering contribution from the measured x-ray structure factor, S(q). This has previously been achieved by iterative methods for simple molecules containing only a few atoms i.e. <10 Narten and Habenschuss (1984). The problem lies in extending this fitting process to approximate the intra-molecular scattering from larger, high

molecular weight molecules that contain several tens of atoms and retain a structural conformation that is chemically realistic. For a molecule with n atoms, the maximum dimension of the model fitting problem is $\frac{n(n+1)}{2}$. The multiple fitting parameters are related to the structural conformation of the molecule, which is defined by an x-ray weighted average distribution of atom-atom distances and their root-mean-square vibrational amplitudes.

The solution is based on the Sine Fourier transform relation between the structure factor and the atom atom pair distribution function. XISF (X-ray Intermolecular Structure Factor) calculates the intra-molecular scattering for a series of atom-atom pairs, based on the atomic positions of a single molecule (from a given crystal structure or calculation) using a zeroth order Bessel function j_0

$$S_{intra}(q) = \frac{\sum_\alpha \sum_{\beta \neq \alpha} \{f_\alpha(q) \cdot f_\beta(q) \cdot j_0(r_{\alpha\beta} \cdot q) \cdot exp(-\frac{1}{2} l_{\alpha\beta}^2 q^2)\}}{(\sum_\alpha f_\alpha(q))^2} \tag{4.3}$$

where f_α and f_β are x-ray form factors for atom types α and β respectively Waasmaier and Kirfel (1995), and $l_{\alpha\beta}$ is the root-mean-square value of atom-atom distance $r_{\alpha\beta}$ Narten (1972); Nasr *et al.* (1999). Since the x-ray scattering structure factor is dominated by intra-molecular interactions of the heavier elements (with higher Z), the weakly scattering hydrogen atoms are neglected in the multi-dimensional fitting. This approximation greatly reduces the number of interactions. The values for the root mean square deviations lab are obtained by fitting the medium-to-high q ranges, which were empirically found to work best for values of q>4 $^{-1}$) of the structure factor S(q) using a trust-region optimization Nocedal and Wright (2006) routine lsqcurvefit and constraining the peak positions and widths to be physically reasonable Benmore *et al.* (2013). For large molecules, the minimum q of intramolelcuar S(q) fitting procedure can be set to lower value but the result should be carefully examined to avoid potential overlap with intermolecular interactions. The q-dependent atomic form fac-

tors are taken from Waasmaier & Kirfel (1995). Based on this model the calculated intra-molecular structure factor is subtracted from the total scattering data across the entire q-range to yield the inter-molecular structure factor.

The uniqueness of modeling any structure factor based on diffuse scattering data alone represents a known problem, and has been discussed at length for similar fitting methods McGreevy (2001); Soper (2007). The same difficulties apply here. Namely, two quite structurally different models may fit the diffraction data equally well. The procedure will however enable some models to be ruled out. A distinct advantage with our program is the knowledge of the intra-molecular shape from other methods can be easily incorporated. We have previously applied this procedure with prior knowledge of the molecular conformation based on NMR spectroscopy Benmore *et al.* (2013); Weber *et al.* (2013) which goes a long way to addressing this issue.

4.2 Optimization methods

The optimization problem involves an objective function that depends on some variables

$$\min_x f(x) \tag{4.4}$$

where $x \in R^n$ is a real vector with $n \geqslant 1$ elements and f is a smooth function on R. The iterative line search algorithm can solve this problem, though may have may restrictions and imperfections. The *Gradient Decent* method, for example, will move along its gradient $p_k = -\nabla f_k / \|\nabla f_k\|$ for each step, therefore the algorithm update the variable x according to

$$x_{k+1} = x_k + \alpha \cdot p_k \tag{4.5}$$

where α is the step length. This method is also called steepest decent because the move direction is always orthogonal to the contours of the objective function. The

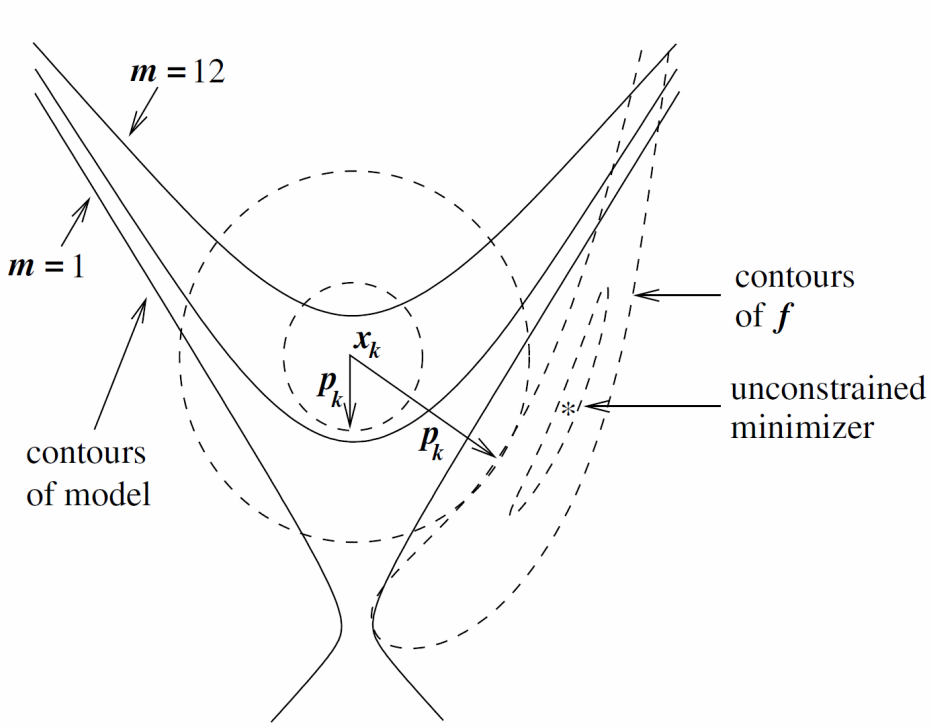

Figure 4.1: The trust region construction Nocedal and Wright (2006). Two possible choices of trust regions are shown here.

advantage of gradient decent is that second derivatives are not needed. However the algorithm will be slow on difficult objective functions, such as poorly scaled function.

The primary method we use for structure factor optimization is *trust region*. In this method, information is gathered on f and then is used to construct a model function m_k whose behavior near the current point x_k is similar to the actual function. In addition, the algorithm restrict the search for the minimum of approximation function m_k to a region around x_k. The problem can be formulated as

$$\min_p m_k(x_k + p) \quad subject\ to \|p\|_2 \leqslant \Delta_k \tag{4.6}$$

where Δ_k is the radius of the trust region and the model function m_k is defined to ba a quadratic function

$$m_k(x_k + p) = f_k + p^T \nabla f_k + \frac{1}{2} p^T B_k p \tag{4.7}$$

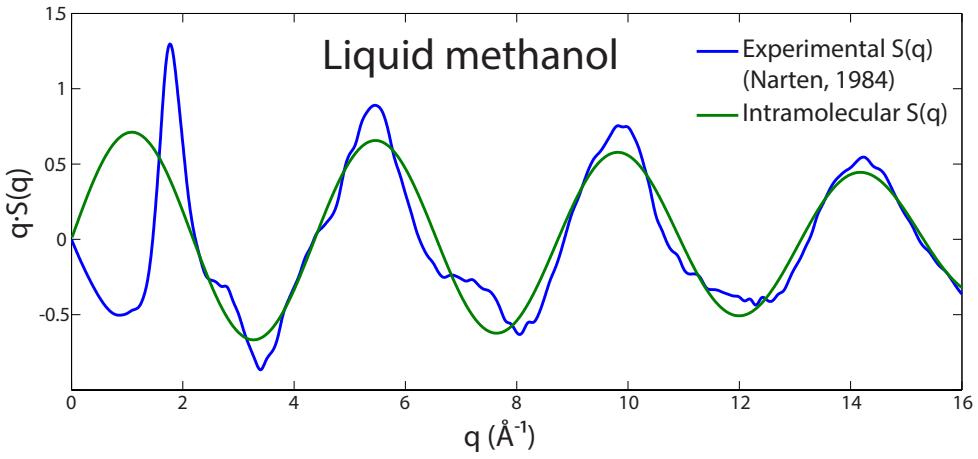

Figure 4.2: Calculated intramolecular structure factor of liquid methanol. The fitting range is from 4 to 16 $^{-1}$).

The search direction p doesn't necessarily need to match the gradient and $x_k + p$ should fall inside the trust region, as illustrated in Fig.4.1. The ∇f_k is used here to ensure that model function agrees with the true objective function to the first order. The matrix B_k is usually an approximation to the Hessian $\nabla^2 f_k$ in practice.

4.3 Results

Fig.4.2 shows the calculated intramolecular structure factor of liquid methanol based on the experimental X-ray data taken from Narten and Habenschuss (1984). The best fitted rms- deviation of C-O bond (1.437 Å) is 0.069 Å, which matches Narten's estimation of 0.064 Å with 8% discrepancy. The difference is due to the use of molecular form factors in Narten's work and the chosen optimization algorithm used in XISF. For a sanity check, we have plotted the atomic structure factors for set of atoms using the XISF C++ code 4.4.

To demonstrate the programs ability to separate out the intra and inter-molecular contributions to S(q) test the program's ability and potentially differentiate between

different molecular structures, we supplied the program with two different intramolecular conformations of the Probucol molecule. In Fig.4.3, the intermolecular S(q)'s are shown as the red curves, for both Probucol form I and form II (inset). The q range of 4.5 $^{-1}$) to 16 $^{-1}$) was chosen to calculate the intramolecular structure factor S(q). The bent conformation provides a superior fit (notably around q=8 to 14 $^{-1}$) and has been shown by NMR to be the dominant molecular shape adopted in the glass Weber *et al.* (2013). However, it should be pointed out that differentiating structures with similar intermolecular S(q) curves is likely to be problematic.

We have utilized XISF to study the glass transition in amorphous drugs, as show in Fig.4.5 Benmore *et al.* (2013). The X-ray structure factors and chemical structures for glassy carbamazepine, cinnazirine, miconazole, clotrimazole, and probucol are shown in Figure 2. The X-ray curves represent the average of several measurements taken on different samples. Most of the molecules studied here contained 30-60 atoms, allowing for some variation in the fitting parameters $\Delta r^2 \alpha\beta$. However, the largest molecule is probucol with m=83 distinct atoms in the molecule, corresponding to 3403 total interactions. Of the total 83 atoms in probucol, 48 are hydrogen atoms, which scatter X-rays weakly, reducing the number of dominant intermolecular interactions to 595, allowing reasonable q-space data fitting to be achieved. Both the measured and calculated curves were then Fourier transformed over the same momentum transfer range using the same Lorch modification function Lorch (1969) to give the intramolecular $D^X_{intra}(r)$ and intermolecular $D^X_{inter}(r)$ differential pair distribution functions. The differential distribution $D(r)$ is defined using the Hannon-Howells-Soper nomenclature as Egelstaff *et al.* (1971); Walford and Dore (1977); Keen (2001)

$$D(r) = 4\pi\rho r[G(r) - 1] \tag{4.8}$$

$$= \frac{2\rho}{\pi} \int_{q_{min}}^{q_{max}} q[S_X(q) - 1] sin(q \cdot r) dq \tag{4.9}$$

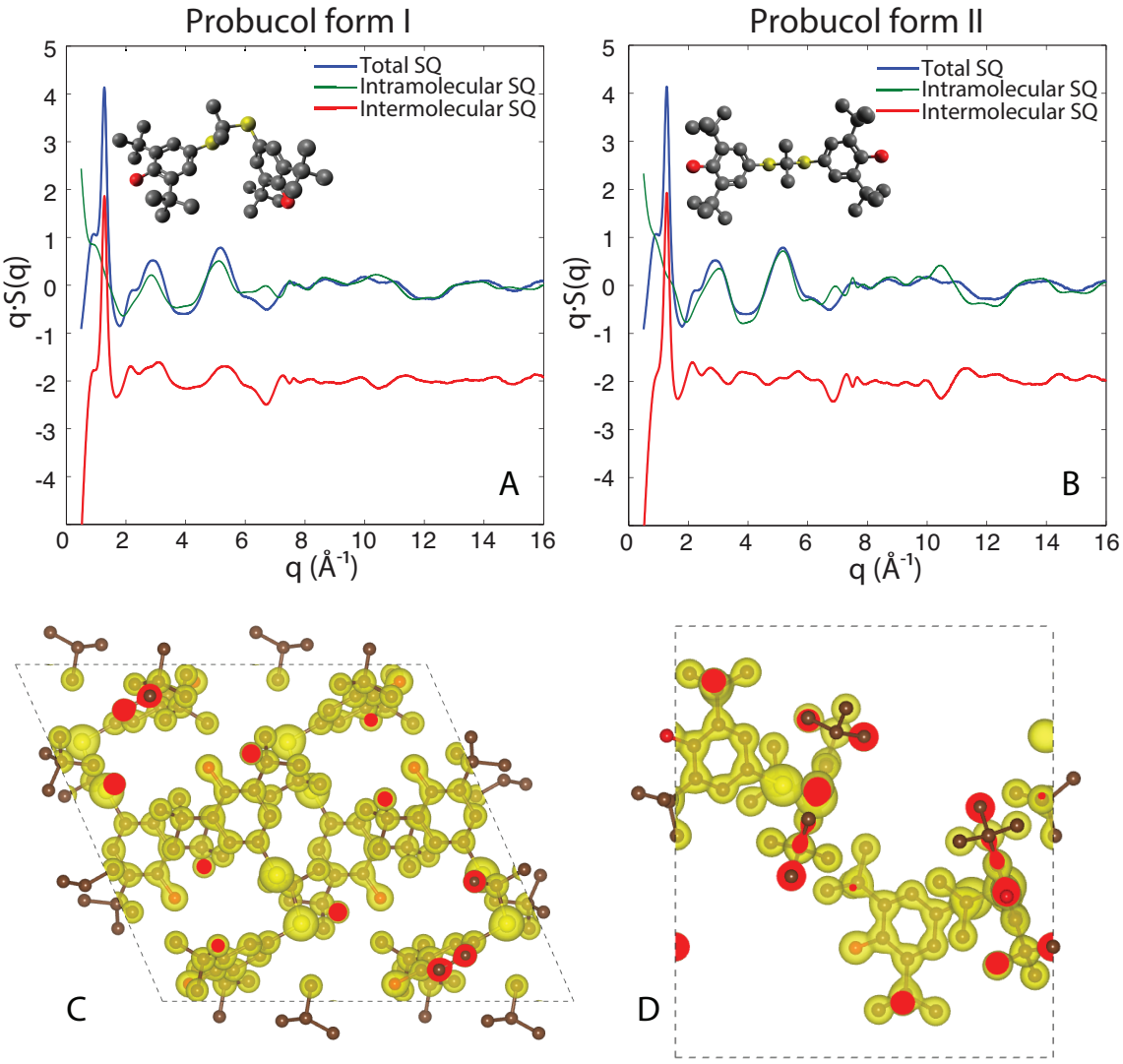

Figure 4.3: Bent (left) and linear (right), fitted to the same q-weighted experimental x-ray data. The extracted intermolecular S(q)'s are shown as red curves. The molecular conformations with representative rms-deviations of the atoms are shown as inserts. The electron density maps (bird view along b-axis) calculated from the extracted rms-deviations are shown for bent form crystal (left, space group P21/c) and linear form crystal (right, space group P21).

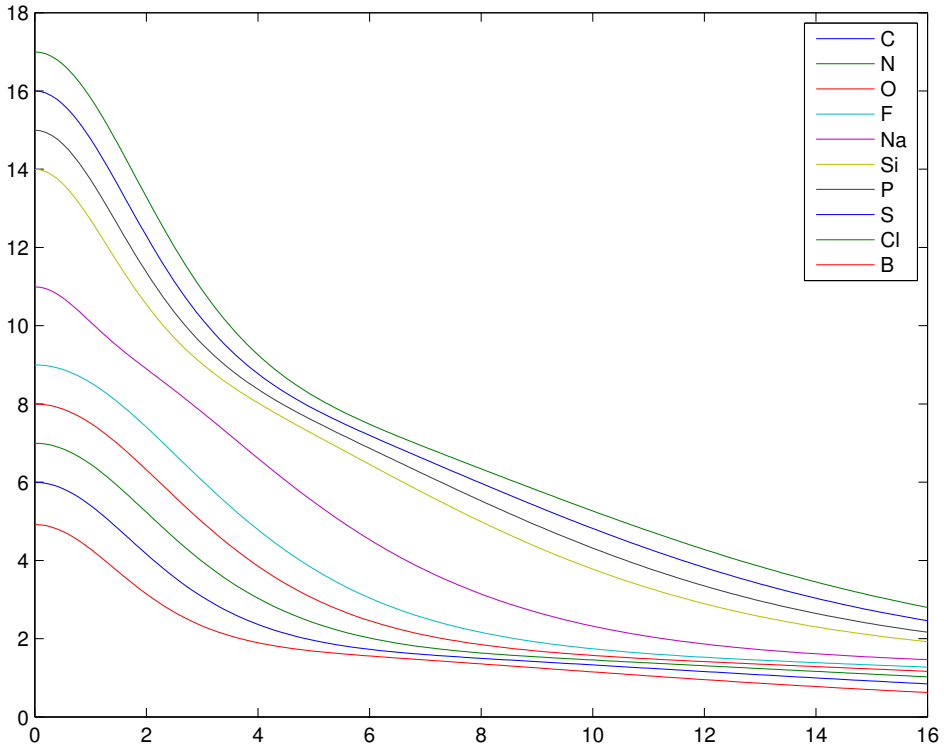

Figure 4.4: Atomic form factors calculated by C++ code using MATLAB MEX API.

where $G(r)$ is the total pair distribution funtion, which oscillates about 1 at high r. The $D(r)$ function removes the bulk density to emphasize the ordered structural peaks and dips, especially at higher r.

These functions have been shown multiplied by r in Fig.4.6 to highlight the medium-range interactions in the samples. Given the variability in accurately fitting $\Delta r^2 \alpha\beta$, any small oscillations in $rD^X_{inter}(r)$ should be treated with some caution. However, significant peak shifts and large oscillations can be reasonably be considered as real effects in the glassy states. The resulting intermolecular pair distribution functions revealed broad nearest and next-nearest neighbor molecule-molecule correlations.

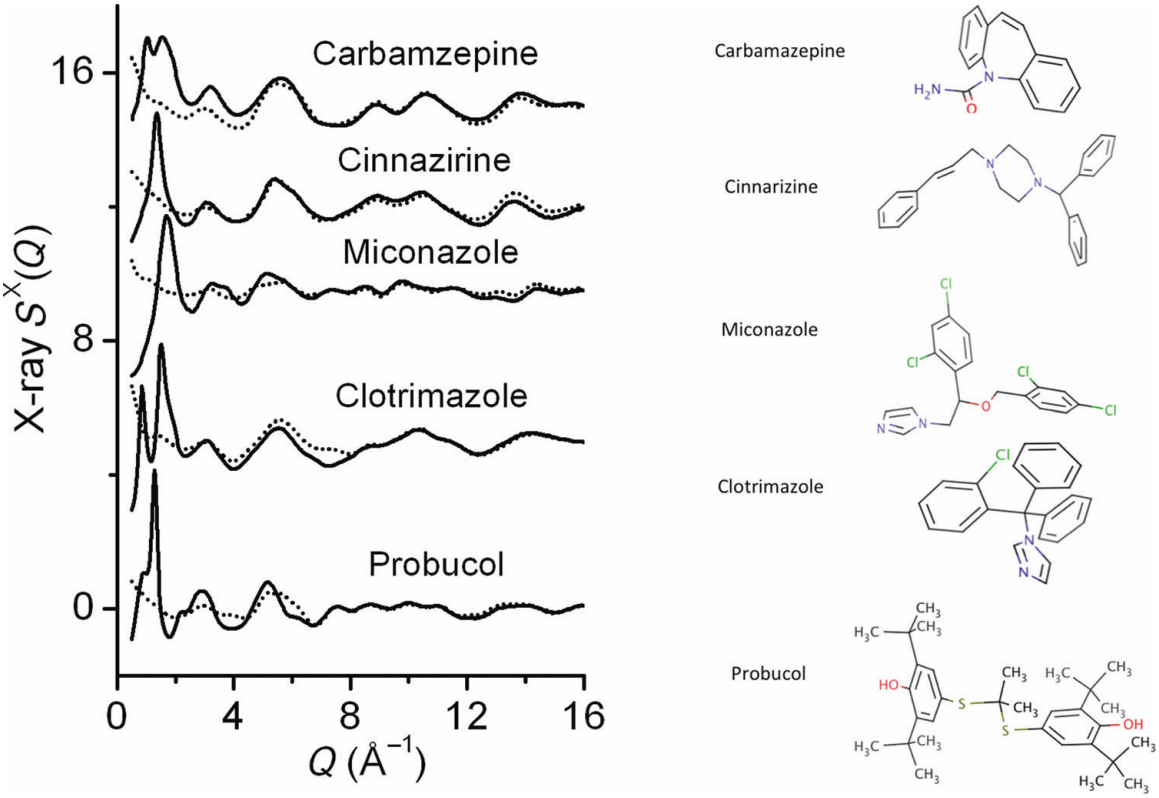

Figure 4.5: (Left) The measured total X-ray structure factors for glassy drugs (solid lines). With the high X-ray flux and detector sensitivity used, Bragg peaks would be clearly observed if more than 1% crystalline materials were present in the sample. The absence of any Bragg peaks confirms the amorphous nature of the drugs produced by melting acoustically levitated samples. The dotted lines represent the calculated intra-molecular curves $S_{intra}(Q)$ based on the crystal structures. (Right) Typical molecular structures of the drugs studied taken from the crystal structures.

4.4 Software environment

The program is primarily written in MATLAB programming language with core cost function written in C++. The program employs MATLAB-C++ (MEX) programming paradigm to achieve fast computation of cost function. We provide a pre-compiled binary for Windows 64-bit operating system. The software depends on 64-bit Visual C++ redistributable and Matlab Compiler Runtime which can be downloaded free of charge. The source code is hosted on Github repository https://github.com/xrayapp/XISF, with the input data stored under the folder named

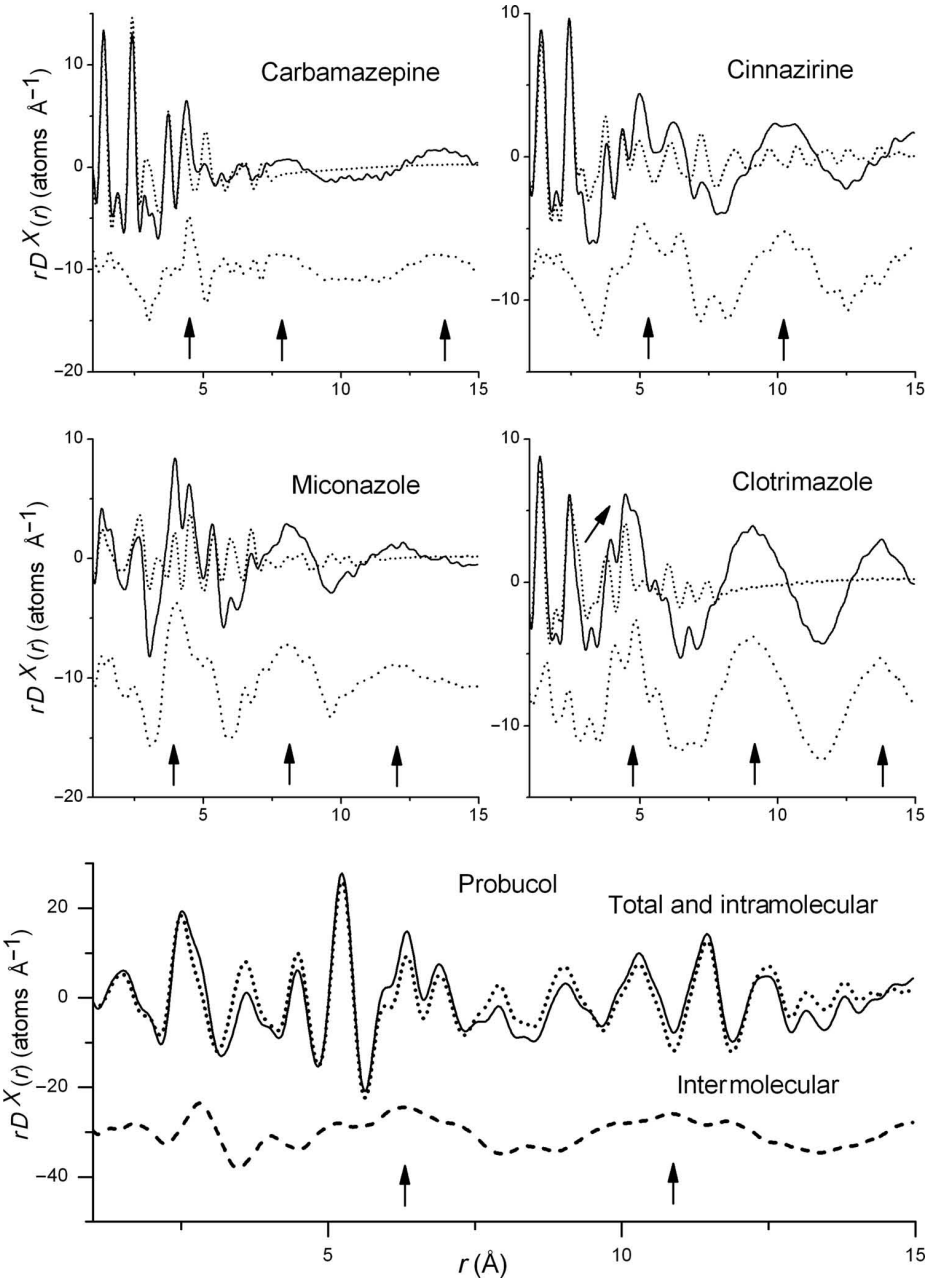

Figure 4.6: The measured total X-ray differential pair distribution function multiplied by r to highlight the medium-range interactions $rD^X(r)$ (solid lines) compared with the intramolecular curves $rD^X_{intra}(r)$ calculated from the crystal structures (dotted lines, these curves represent the Fourier transforms of the dotted lines in Fig.4.5). The difference between the solid and dotted lines corresponds to the intermolecular pair distribution function $rD^X_{inter}(r)$ (dashed line below). The peaks corresponding to the marked arrows are discussed in the text.

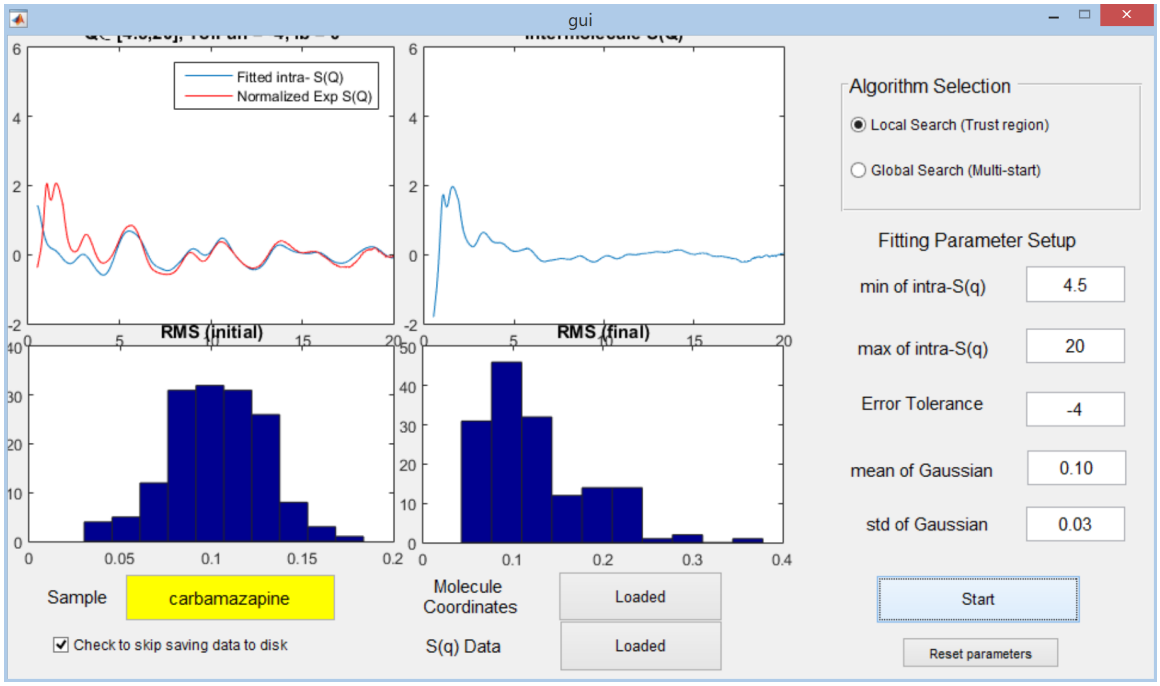

Figure 4.7: The GUI interface of XISF. Here is a trail output on drug sample Carbamazapine.

input. The website has a README guide on how to execute the program in MATLAB.

The basic system requirement is identical to MATLAB, which can be viewed on official website. For XISF, the memory required at run time depends on the input data size, and it generally scales as $O(n^2)$where n is the number of (non-hydrogen) atoms in the molecule. For the test case Probucol (35 non-hydrogen atoms), XISF takes up to 1.0 GB memory on Windows 8.1 64-bit OS.

XISF has a GUI designed to be easy to use and informative. Fig.4.7 and Fig.4.8 shows the output of an optimization trail on drug sample Carbamazapine. The program will show the fitted intramolecular structure factor and extracted intermolecular curve, along with the initial and final distribution of the r.m.s values. The program offers two optimization method, the local search and global search, both are based

```
Fitting range:                [4.5, 20.0]
rms population:               153
Fitting accuracy:             1e-4
Adjusted data population:     200
Adjusted Fitting range:       [4.5, 20.0]

rms initialization parameter (Random Gaussian Distribution)
 Mean:            0.1
 Std:             0.03

                                                  Norm of      First-order
 Iteration  Func-count     f(x)                    step          optimality
     0        156       0.0679583                                    1.18
     1        312       0.0452915                  1.27889          0.609
     2        468       0.0403444                 0.462446          0.519
     3        624       0.0392768                 0.118053          0.496
     4        780       0.0383199                 0.115586          0.475
     5        936       0.0382707              0.000696544          0.129
```

Figure 4.8: The console interface of XISF. Here is a trail output on drug sample Carbamazapine.

on trust region method. In the future, we will implemented stochastic stimulated annealing optimization, which usually offer better fitting accuracy for difficult objective function, though the runtime will be much longer.

4.5 Program specification

The program takes the q-weighted experimental structure factor as the first input, and a file containing the molecule coordinates as the second one. The calculated q-weighted intermolecular structure factor will be automatically saved in ASCII format. The program is designed for X-ray scattering data of molecular liquids and amorphous solids, and has been tested on the amorphous drugs probucol, cinnazirine, carbamazepine, miconazole nitrate and clotrimazole (Benmore et al., 2013). On a 3.4GHz Intel based Windows 8.1 64-bit machine, the runtime of carbamazepine (18 atoms) is around 10 minutes, and the runtime of probucol (35 atoms) is about 1 hours, all

measured with fitting accuracy of 10^{-4} using single-thread version of the program. XISF has about 1200 lines of code total, with half of them written in C++.

Chapter 5

Density functional theory study of the secondary structures in spider silk fibers

5.1 Introduction

Spider dragline silk produced from the major ampullate gland is one of the toughest biopolymers known. Lewis (2006) It is comprised almost entirely of two proteins, major ampullate spidroin 1 and 2 (MaSp1 and MaSp2). Xu and Lewis (1990); Hinman and Lewis (1992) The unique mechanical properties of spider silk are thought to originate from the secondary and tertiary structure of these silk proteins. The structure of spider silk proteins within the fiber has been studied extensively with a number of techniques including X-ray diffraction (XRD), solid-state nuclear magnetic resonance (NMR), infrared (IR) and Raman spectroscopy. This combination of structural characterization techniques has illustrated that poly(Ala) and poly(Gly-Ala) repeat units form nano-crystalline β-sheet structures while, the Gly-Gly-X and Gly-Pro-Gly-X-X motifs take on disordered 3_{10}-helical Kümmerlen *et al.* (1996) and elastin-like type II β-turn structures respectively.

At the core of the secondary and tertiary structure of proteins is hydrogen-bonding. These hydrogen-bonding interactions can either occur within a given protein strand (intra-strand) or between strands (inter-strand). Solution-state NMR has

played a critical role in understanding hydrogen-bonding distances for proteins in aqueous solutions from amide proton chemical shifts. Proton (1H) combined rotation with multiple pulse spectroscopy (CRAMPS) solid-state NMR techniques have been used for decades to improve resolution and determine hydrogen-bond strengths from the proton chemical shifts of small molecules. Recently, the advent of very fast (35-40 kHz) and ultra-fast (Z60 kHz) magic angle spinning (MAS) NMR probes has allowed improved proton resolution in rigid solids by averaging the strong 1H-1H dipolar interactions with rapid MAS rates particularly when applied at high magnetic fields (Z600 MHz). Very fast and ultra-fast MAS proton NMR has been used to characterize hydrogen-bonding in benzoxazine dimers, pharmaceutical solids, and phosphonic acids. The experimental results from fast and ultra-fast MAS proton NMR has been used with theoretical chemical shift calculations to determine hydrogen-bonding lengths from the amide proton chemical shift for silk-like model peptides.

In the present contribution, we characterize the hydrogen-bonding interactions in Nephila clavipes spider dragline silk (major ampullate silk fibers) with two-dimensional (2D) 1H-13C HETCOR solid-state NMR at a very fast MAS rate of 40 kHz. The greater chemical shift dispersion in the 13C dimension of the 2D 1H-13C HETCOR MAS experiment is necessary to assess the hydrogen-bonding for individual amino acid residues. The projection in the 1H dimension illustrates the typical resolution in a one-dimensional 1H experiment of spider silk with 40 kHz MAS where distinct amide chemical shifts are not resolved (see Fig. 1). The 2D 1H-13C HETCOR MAS experiment can be used to extract the amide proton chemical shift in an amino acid specific manner by slicing through the 13C dimension and comparisons can be made to DFT proton chemical shift calculations to determine the hydrogen-bond strength for Gly and Ala in β-sheet and 310-helical structures.

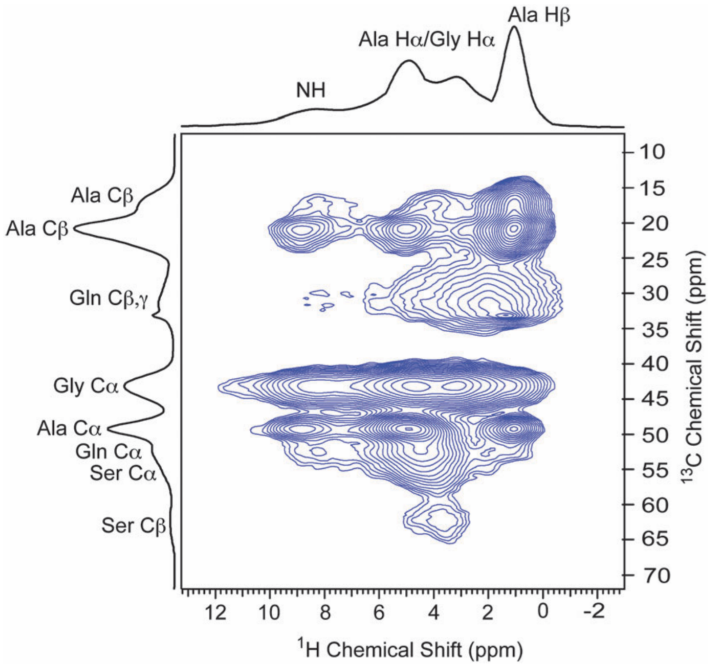

Fig. 1 The 2D ^{1}H–^{13}C HETCOR solid-state NMR spectrum of ^{13}C-labeled *N. clavipes* spider dragline silk. The spectrum was collected at 800 MHz with a 1 ms CP contact time and 40 kHz MAS.

5.2 DFT Proton Chemical Shift Calculations

Proton NMR chemical shift calculations were performed using B3LYP and the 6-31++G(2d,2p) basis set in Gaussian09 similar to previously described approaches. Yamauchi *et al.* (2000); Blanchard *et al.* (2012) The Gly-Gly β-sheet model (Fig. S2, a) was built manually in GaussView 5. First, a single strand was constructed followed by geometry optimization in Gaussian 09 using the GIAO method with the above stated basis set. Stability was checked with the same basis set following geometry optimization. Based on the optimized single strand structure, the Gly-Gly β-sheet model with inter-strand hydrogen- bonding was constructed by duplicating the single strand model. For the β-sheet hydrogen-bonding trend, the inter-strand NH-OC hydrogen-bond length was varied from 1.7 to 2.7 and the corresponding NMR chemical shifts were calculated. The calculated chemical shift was calibrated

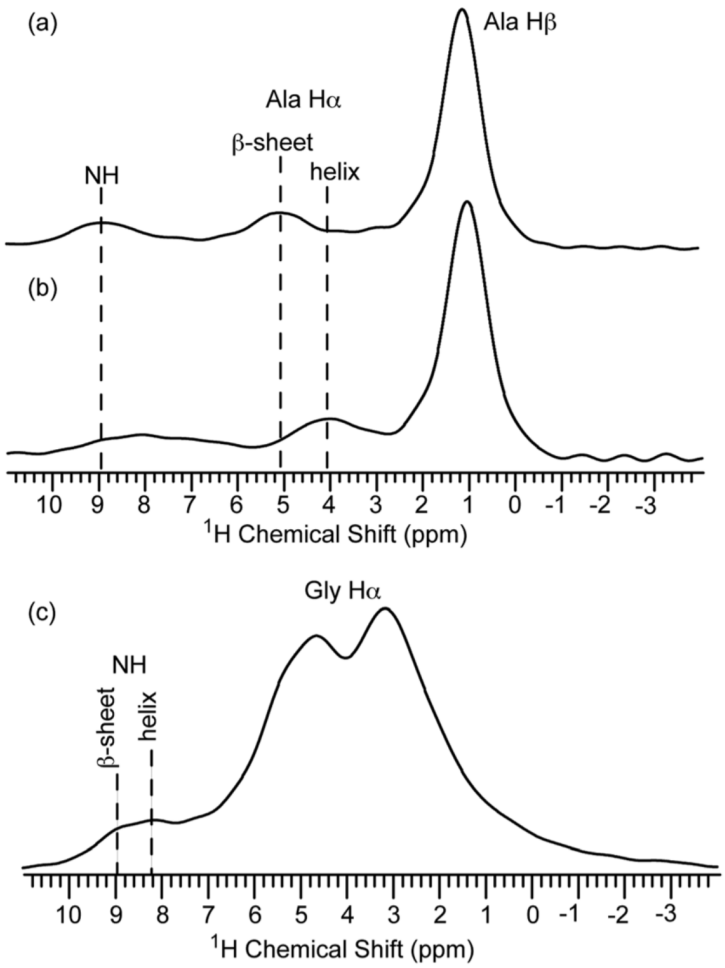

Fig. 2 ^{1}H slices extracted from the 2D ^{1}H–^{13}C HETCOR MAS spectrum of ^{13}C-labeled *N. clavipes* dragline silk. Slices obtained at the two Ala Cβ components that correspond to (a) ordered β-sheet (21.0 ppm) and (b) disordered 3_{10}-helical (17.4 ppm) conformation and (c) the Gly Cα (43.3 ppm).

with the TMS proton magnetic shielding that is built into the Gaussian 09 database.

The 310-helix and α-helix models were obtained from the Lorieau Research Group website in the Department of Chemistry at The University of Illinois at Chicago. These helical models were generated with the Hydrogen-Bond Database Grishaev *et al.* (2004) and XPLOR-NIH. Original helix structures were truncated to smaller helical models. Optimization and stability check were carried out in a similar fashion to the β-sheet model discussed above. For the inter-helix calculation, two small 310-

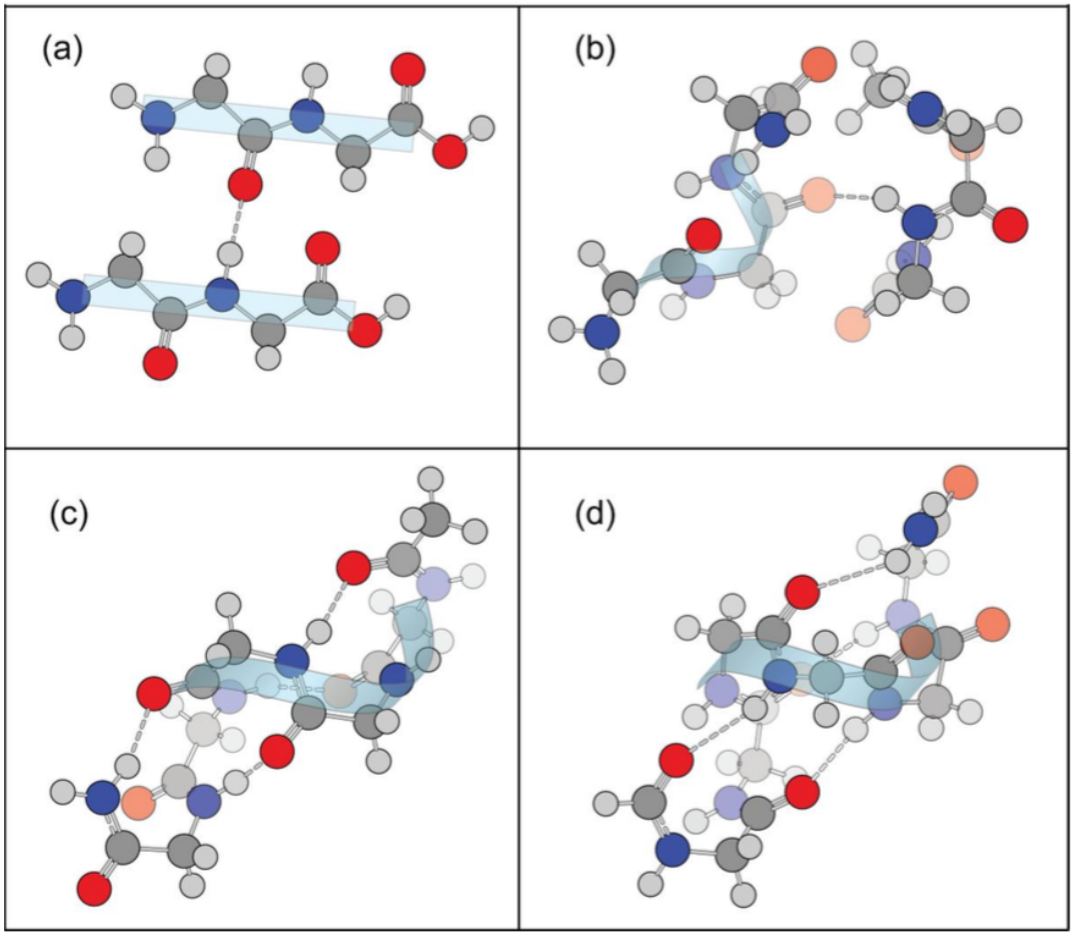

Fig. S2 The protein backbone models used for DFT proton chemical shift calculations, (a) Gly-Gly β-sheet model with inter-strand hydrogen bond, (b) 3_{10}-helical model with inter-strand hydrogen-bond, (c) 3_{10}-helical model with intra-strand hydrogen-bonding and (d) α-helix with intra-strand hydrogen-bonding.

helical strands measuring approximately two repeat units were constructed with one inter-helix hydrogen-bond (Fig. 2S, b). This inter-strand hydrogen-bond is similar to that first observed by Crick and Rich in the 310-helical structure of polyglycine II. CRICK and RICH (1955) The NH-OC length was varied from 1.7 to 2.3 and corresponding NMR chemical shifts were calculated. For the 310-helix (Fig. 2S, c) and α-helix (Fig. 2S, d) intra-bond calculation, larger structures with approximately five repeat units were used. A two-stage geometry optimization was carried out with HF 3-21G and B3LYP 6-31G++ basis sets. NMR chemical shift calculations were then carried out as described above. The NH-OC intra-strand hydrogen-bond lengths varied between 2.0 to 2.4 for the 310-helix and α-helix depending on the environment. This variability is believed to occur because of the intrinsically different hydrogen-bonding sites in the two helical conformations. The amide N-H bond length remained very close to the theoretical value of 1.00 and varied by less than 0.3% for all structures following geometry optimization.

To illustrate the amide proton chemical shift trend as a function of hydrogen-bond distance, the calculated proton amide chemical shift data was fit to an equation of the form

$$\delta_{NH} = ad^{-3} + b \tag{5.1}$$

where δ_{NH} is the amide proton chemical shift and d is the hydrogen-bond distance. The calculated amide proton chemical shift data is included in Fig. 3 along with the fits and follow the expected trend.

The backbone Hα chemical shifts are known to depend on conformation and were also tabulated from the DFT proton chemical shifts calculations for the different structures. The average Hα proton chemical shift was 4.7, 4.1 and 3.7 ppm for the β-sheet, 310-helix and α-helix secondary structures. This is in agreement with experimentally

determined H chemical shifts for various solid polypeptides with ^{1}H combined rotation and multiple pulse spectroscopy (CRAMPS) NMR where α-helical structures gave 3.9-4.0 ppm shifts while, β-sheet forms were 5.1-5.5 ppm. Interestingly, results from our calculations indicate that it may be possible to distinguish between the 310 and α-helix since nearly all the Hα shifts for the 310-helix were centered at 4.1 ppm with essentially no deviation and the α-helix ranged from 3.4-3.8 ppm with an average of 3.7 ppm.

5.3 DFT and NMR analysis

The 2D 1H-13C HETCOR MAS NMR spectrum of 13C-labeled Nephila clavipes spider dragline silk is shown in Fig. 1. The assignment of the 13C resonances is based on our previous 2D 13C homonuclear through-bond and through space solid-state NMR correlation experiments. Holland *et al.* (2008a,b) The data was collected with a 1 ms CP contact time. At this intermediate contact time length, all protons within a given amino acid are observed for Ala and Gly thus, Ha, Hb and amide proton chemical shifts can be extracted for a given amino acid.

In order to extract the proton chemical shifts for Ala and Gly, slices were extracted from the 2D 1H-13C HETCOR MAS spectra at specific 13C chemical shifts. The Ala Cb resonance has been shown in previous studies to be heterogeneous with a minimum of two-components at 17.4 and 21.0 ppm that can be assigned to Ala present in disordered 310-helical and ordered β-sheet structures, respectively. The Ala in 310-helical structures have been ascribed to Ala located in the repitive Gly-Gly-X motif while, the Ala in β-sheet structures are located in the poly(Ala) and flanking poly(Gly-Ala) motifs in the primary amino acid sequence. Creager *et al.* (2010) Interestingly, the 1H spectrum for these two Ala environments display differing Ha and amide proton chemical shifts (see Fig. 2a and b). For the Ala Ha, the observed

proton chemical shifts are 4.2 and 5.1 ppm for the slices extracted at a 13C chemical shift of 17.4 and 21.0 ppm, respectively. These observed proton Ha chemical shifts agree with previous conformational studies on polypeptides where helical structures had shifts of 3.9-4.0 ppm while, β-sheet conformations had 5.1-5.5 ppm shifts. Shoji *et al.* (1996) These results also agree with our DFT calculated Ha chemical shifts for various conformations (see ESI†). In addition, the DFT calculations indicate an average Ha chemical shift of 4.1 ppm for the model 310-helix and a smaller average Ha shift of 3.7 ppm for the a-helix. This indicates that the Ha shift can potentially be used to distinguish different helices with the helix observed in spider silk more closely matching the 310-helix compared to the a-helix. The 1H spectrum for Gly was extracted at the Gly Ca chemical shift and shown in Fig. 2c. The observed Ha has two components that could represent two distinct conformational environments however, the more likely explanation is that the two Ha protons are diastereotopic as is observed for crystalline glycine.

Two amide NH proton environments are observed for both Ala and Gly in spider silk (see Fig. 2a-c). For Ala, there are two environments that can be extracted from the two Cb components. The 310-helical component (Fig. 2b) has a broad amide NH proton resonance positioned at 8.2 ppm and the component (Fig. 2a) has a sharper resonance at 9.0 ppm. The Gly 1H slice extracted at 43.3 ppm in the 13C dimension is asymmetric with two components at 8.2 and 9.0 ppm. Thus, both Ala and Gly exhibit two distinct amide NH environments. It is known that the amide NH shift can be dependent on both the backbone conformation and the hydrogen-bonding distance. Kimura *et al.* (1998, 2000) In order to discern the two contributions to the chemical shift, a series of DFT proton NMR chemical shift calculations were conducted on poly(Gly) model systems in , 310-helical and a-helical conformations. The calculated amide proton chemical trends for the various secondary structures

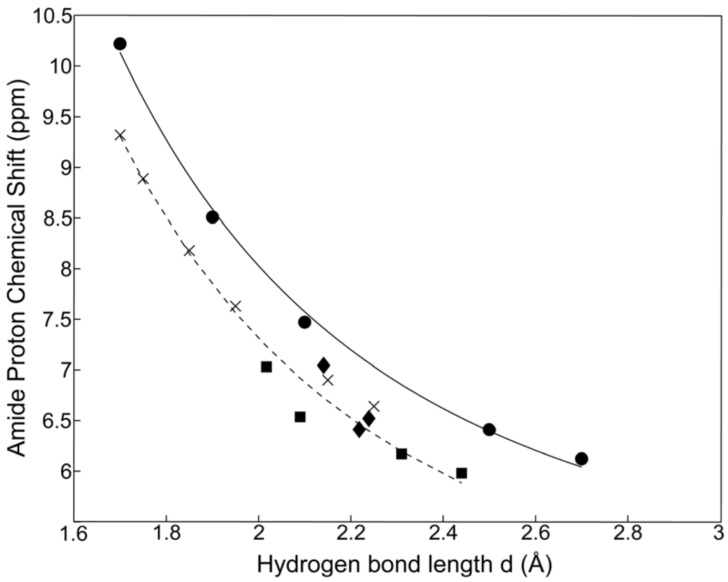

Fig. 3 Plot of the calculated amide proton chemical shift (ppm) as a function of NH···OC hydrogen-bond length, *d* (Å). The amide proton chemical shifts were calculated for (●) β-sheet and helical poly(Gly) models. Various helical structures were used including the (◆) intra-helical and (×) inter-helical hydrogen bonding in a 3_{10}-helix and the (■) intra-helical hydrogen bonding in an α-helix. The calculated data was fit to extract eqn (1) and (2) for determining NH···OC hydrogen-bond lengths for β-sheet (—) and helical (---) structures (see ESI†).

were fit to extract the equations 5.3 for determining hydrogen-bond lengths in and helical conformations, where δ_{NH} is the amide 1H chemical shift and d is the NH-OC hydrogen-bond distance (see Fig. 3). The trend is very close to previously report trends in model peptides and the helical trend is on average shifted to lower ppm. If one uses the y-intercept as a measure of the impact of secondary structure, there is a 0.6 ppm lower amide shift for helical structures. This is in reasonable agreement with the experimentally determined 0.2-0.6 ppm lower average amide proton chemical shifts determined for helical structures compared to with protein solution-state NMR.

$$\delta_{NH} = 26.8d^{-3} + 4.7, \beta - sheet \tag{5.2}$$

$$\delta_{NH} = 25.3d^{-3} + 4.1, helix \tag{5.3}$$

The amide proton chemical shift as a function of hydrogen-bond length can be used to determine the hydrogen-bond strength for Ala and Gly in and 310-helical structures for spider dragline silk. From equation 5.3 it is determined that the hydrogen-bond length is 1.84 and 1.83 for the two structures, respectively. Upon first inspection, the smaller amide chemical shifts observed for the 310-helical structures would indicate weaker hydrogen-bonding however, when the impact of secondary structure is accounted for the hydrogen-bond strength is identical for both environments. The hydrogen-bond strength is also notably stronger for spider silk compared to alanine tripeptide silk mimics. In addition, the strong hydrogen bonding is consistent with inter-strand hydrogen-bonding as weaker hydrogen-bonds would be expected for intra-strand hydrogen-bonds.

The combination of 1H-13C HETCOR solid-state NMR with fast MAS and DFT proton chemical shift calculations have been used to determine the hydrogen-bonding strength for and 310-helical structures in spider silk. The hydrogen-bond strength was found to be identical for both structures and quite strong compared to model silk peptide mimics. The strong hydrogen-bonding is indicative of inter-strand interactions and provide some of the first evidence for inter-molecular interactions in spider silk. The hydrogen-bonding trends reported here should be useful for researchers determining hydrogen-bond strength in both and helical conformations from amide proton chemical shifts.

5.4 ^{13}C NMR calculation

While we successfully established the non-linear relation between h-bond length and ^{1}H chemical shift, we are curious to see if such relationship exists for ^{13}C. We used both *Quantum Espresso* and *Gaussian* 09 to test on alanine gas phase model. For *Quantum Espresso*, the gas phase and crystalline phase of alanine were tested. The

model is shown in Fig.5.1 and Fig.5.2 respectively. To construct gas phase model with plane-wave based DFT package, one need to construct a relatively large unit cell which contains only one gas molecule. The gas molecules is still periodic on the lattice but the nearest distance between molecules is large enough to approximate gas phase interaction.

For ^{13}C model, the simulated electron density is different from the actual molecule environment where the atoms are located. This is mainly due to the complex electronic orbitals in carbon atom. As a result, the chemical shift calculated from DFT is vastly different from the actual experimental data that use TMS as the baseline. The method to compare calculation from different trails is to establish linear correlation between them, as shown in fig.5.3. Cross-comparison between different trails is difficult because the linear fitting parameters will change for every new calculation.

The calculated chemical shifts (from Gaussian 09) are shown in Fig.5.4 with the experimental values showed side by side. The three secondary structures explored here are 3_{10} helix, α-helix and β-sheet. For each structure, the DFT result and experimental value follows near perfect linear correlation, such as the one shown in Fig.5.3. However, the coefficients are slightly different. Since the difference of experimental chemical shift is also quite small, therefore we can't use one set of linear coefficient to convert the DFT result for all structures. In addition, we see that the chemical shift of α-C in 3_{10} helix and α-helix are 138.58 and 138.59 ppm respectively, almost in-distinguishable. For future work, we will continue to explore the capability of the DFT method and try to distinguish the helical structures with the simulated NMR result.

Figure 5.1: The gas phase model for alanine molecule. For each unit cell, only one alanine molecule is included. The resulting structure will be used to approximate gas phase of the substance.

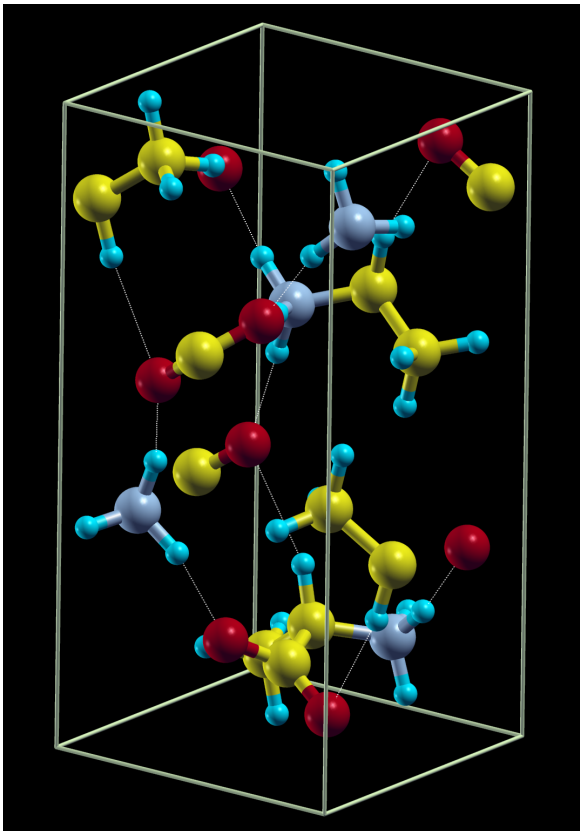

Figure 5.2: The crystalline phase model for alanine. The unit cell is orthorhombic.

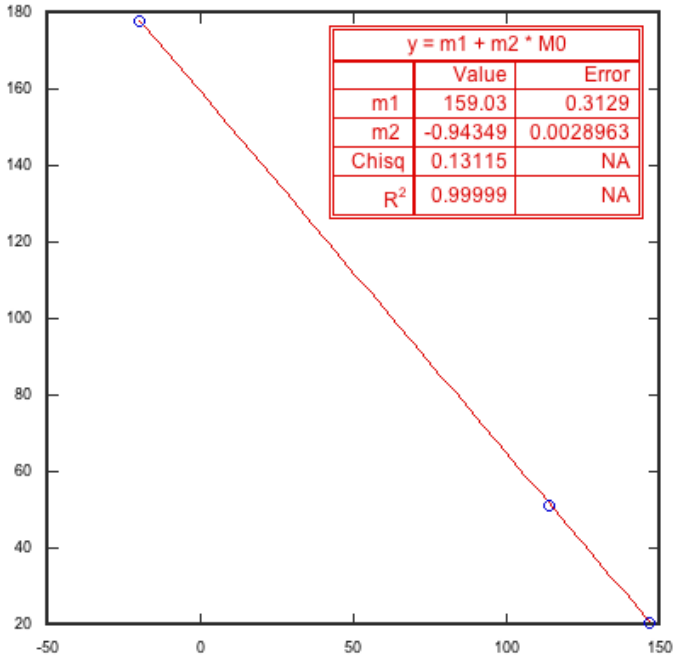

Figure 5.3: The linear correlation between calculated alanine model and the experimental values. The y-axis is the experimental value, and the x-axis is the calculated chemical shift from Quantum Espresso.

Figure 5.4: The ^{13}C chemical shift calculated for three different secondary structures.

Asakura, T., Y. Suzuki, Y. Nakazawa, G. P. Holland and J. L. Yarger, "Elucidating silk structure using solid-state nmr", Soft Matter **9**, 11440–11450 (2013).

Becker, N., E. Oroudjev, S. Mutz, J. P. Cleveland, P. K. Hansma, C. Y. Hayashi, D. E. Makarov and H. G. Hansma, "Molecular nanosprings in spider capture-silk threads", Nat Mater **2**, 4, 278–283 (2003).

Benmore, C. J., J. K. R. Weber, A. N. Tailor, B. R. Cherry, J. L. Yarger, Q. Mou, W. Weber, J. Neuefeind and S. R. Byrn, "Structural characterization and aging of glassy pharmaceuticals made using acoustic levitation", Journal of Pharmaceutical Sciences **102**, 4, 1290–1300 (2013).

Blanchard, J. W., T. L. Groy, J. L. Yarger and G. P. Holland, "Investigating hydrogen-bonded phosphonic acids with proton ultrafast mas nmr and dft calculations", The Journal of Physical Chemistry C **116**, 35, 18824–18830 (2012).

Chung, K. L. and F. AitSahlia, *Elementary probability theory: with stochastic processes and an introduction to mathematical finance* (Springer Science & Business Media, 2012).

Creager, M. S., J. E. Jenkins, L. A. Thagard-Yeaman, A. E. Brooks, J. A. Jones, R. V. Lewis, G. P. Holland and J. L. Yarger, "Solid-state nmr comparison of various spiders' dragline silk fiber", Biomacromolecules **11**, 8, 2039–2043 (2010).

CRICK, F. H. C. and A. RICH, "Structure of polyglycine ii", Nature **176**, 4486, 780–781 (1955).

Drummy, L. F., B. L. Farmer and R. R. Naik, "Correlation of the [small beta]-sheet crystal size in silk fibers with the protein amino acid sequence", Soft Matter **3**, 877–882 (2007).

Du, N., X. Y. Liu, J. Narayanan, L. Li, M. L. M. Lim and D. Li, "Design of superior spider silk: From nanostructure to mechanical properties", Biophysical Journal **91**, 12, 4528 – 4535 (2006).

Egelstaff, P., D. Page and J. Powles, "Orientational correlations in molecular liquids by neutron scattering carbon tetrachloride and germanium tetrabromide", Molecular Physics **20**, 5, 881–894 (1971).

Gardiner, C., *Stochastic Methods, A Handbook for the Natural and Social Sciences*, vol. 13 of *Springer Series in Synergetics* (Springer, 2009), 4th edn.

Grishaev, A., and A. Bax, "An empirical backbone−backbone hydrogen-bonding potential in proteins and its applications to nmr structure refinement and validation", Journal of the American Chemical Society **126**, 23, 7281–7292, pMID: 15186165 (2004).

Hammersley, A. P., S. O. Svensson, M. Hanfland, A. N. Fitch and D. Hausermann, "Two-dimensional detector software: From real detector to idealised image or two-theta scan", High Pressure Research **14**, 4-6, 235–248 (1996).

Hammond, B., W. Lester and P. Reynolds, *Monte Carlo Methods in Ab Initio Quantum Chemistry* (World Scientific, 1994).

Hayashi, C. Y., N. H. Shipley and R. V. Lewis, "Hypotheses that correlate the sequence, structure, and mechanical properties of spider silk proteins", International Journal of Biological Macromolecules **24**, 2–3, 271 – 275 (1999).

Hinman, M. B. and R. V. Lewis, "Isolation of a clone encoding a second dragline silk fibroin. nephila clavipes dragline silk is a two-protein fiber.", Journal of Biological Chemistry **267**, 27, 19320–4 (1992).

Holland, G. P., M. S. Creager, J. E. Jenkins, R. V. Lewis and J. L. Yarger, "Determining secondary structure in spider dragline silk by carbon−carbon correlation solid-state nmr spectroscopy", Journal of the American Chemical Society **130**, 30, 9871–9877, pMID: 18593157 (2008a).

Holland, G. P., J. E. Jenkins, M. S. Creager, R. V. Lewis and J. L. Yarger, "Quantifying the fraction of glycine and alanine in [small beta]-sheet and helical conformations in spider dragline silk using solid-state nmr", Chem. Commun. pp. 5568–5570 (2008b).

Holland, G. P., Q. Mou and J. L. Yarger, "Determining hydrogen-bond interactions in spider silk with 1h-13c hetcor fast mas solid-state nmr and dft proton chemical shift calculations", Chem. Commun. **49**, 6680–6682 (2013).

Jenkins, J. E., S. Sampath, E. Butler, J. Kim, R. W. Henning, G. P. Holland and J. L. Yarger, "Characterizing the secondary protein structure of black widow dragline silk using solid-state nmr and x-ray diffraction", Biomacromolecules **14**, 10, 3472–3483 (2013).

Jiao, Y., F. H. Stillinger and S. Torquato, "A superior descriptor of random textures and its predictive capacity", Proceedings of the National Academy of Sciences **106**, 42, 17634–17639 (2009).

Kalos, M. and D. Ceperley, *Monte Carlo Methods in Statistical Physics* (Springer, 1979).

Kaplow, R., T. A. Rowe and B. L. Averbach, "Atomic arrangement in vitreous selenium", Phys. Rev. **168**, 1068–1079 (1968).

Keen, D. A., "A comparison of various commonly used correlation functions for describing total scattering", Journal of applied crystallography **34**, 2, 172–177 (2001).

Keten, S. and M. J. Buehler, "Asymptotic strength limit of hydrogen-bond assemblies in proteins at vanishing pulling rates", Phys. Rev. Lett. **100**, 198301 (2008).

Keten, S. and M. J. Buehler, "Nanostructure and molecular mechanics of spider dragline silk protein assemblies", Journal of the Royal Society Interface **7**, 53, 1709–1721 (2010).

Keten, S., Z. Xu, B. Ihle and M. J. Buehler, "Nanoconfinement controls stiffness, strength and mechanical toughness of [beta]-sheet crystals in silk", Nat Mater **9**, 4, 359–367 (2010).

Kimura, H., S. Kishi, A. Shoji, H. Sugisawa, and K. Deguchi, "Characteristic 1h chemical shifts of silk fibroins determined by 1h cramps nmr", Macromolecules **33**, 26, 9682–9687 (2000).

Kimura, H., T. Ozaki, H. Sugisawa, K. Deguchi, and A. Shoji, "Conformational study of solid polypeptides by 1h combined rotation and multiple pulse spectroscopy nmr. 2. amide proton chemical shift", Macromolecules **31**, 21, 7398–7403 (1998).

Kirkpatrick, S., C. D. Gelatt and M. P. Vecchi, "Optimization by simulated annealing", Science **220**, 4598, 671–680 (1983).

Kitagawa, M. and T. Kitayama, "Mechanical properties of dragline and capture thread for the spider nephila clavata", Journal of Materials Science **32**, 8, 2005–2012 (1997).

Knuth, D., *Seminumerical Algorithms, The Art of Computer Programming*, vol. 2 (Addison-Wesley, 1981), 2nd edn.

Koski, K. J., P. Akhenblit, K. McKiernan and J. L. Yarger, "Non-invasive determination of the complete elastic moduli of spider silks", Nat Mater **12**, 3, 262–267 (2013).

Kosztin, I., B. Faber and K. Schulten, "Introduction to the diffusion monte carlo method", American Journal of Physics **64**, 5, 633–644 (1996).

Kümmerlen, J., J. D. van Beek, F. Vollrath, and B. H. Meier, "Local structure in spider dragline silk investigated by two-dimensional spin-diffusion nuclear magnetic resonance", Macromolecules **29**, 8, 2920–2928 (1996).

Lei, M., A. M. R. de Graff, M. F. Thorpe, S. A. Wells and A. Sartbaeva, "Uncovering the intrinsic geometry from the atomic pair distribution function of nanomaterials", Phys. Rev. B **80**, 024118 (2009).

Lenstra, J. K., *Local search in combinatorial optimization* (Princeton University Press, 2003).

Lewis, R. V., "Spider silk: Ancient ideas for new biomaterials", Chemical Reviews **106**, 9, 3762–3774 (2006).

Lorch, E., "Neutron diffraction by germania, silica and radiation-damaged silica glasses", Journal of Physics C: Solid State Physics **2**, 2, 229 (1969).

Martel, A., M. Burghammer, R. J. Davies, E. D. Cola, C. Vendrely and C. Riekel, "Silk fiber assembly studied by synchrotron radiation saxs/waxs and raman spectroscopy", Journal of the American Chemical Society **130**, 50, 17070–17074, pMID: 19053481 (2008).

Martin, J. E. and A. J. Hurd, "Scattering from fractals", Journal of Applied Crystallography **20**, 2, 61–78 (1987).

McGreevy, R. L., "Reverse monte carlo modelling", Journal of Physics: Condensed Matter **13**, 46, R877 (2001).

McGreevy, R. L. and L. Pusztai, "Reverse monte carlo simulation: A new technique for the determination of disordered structures", Molecular Simulation **1**, 6, 359–367 (1988).

Metropolis, N., A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller and E. Teller, The Journal of Chemical Physics **21**, 6, 1087–1092 (1953).

Narten, A. H., "Diffraction pattern and structure of noncrystalline bef2 and sio2 at 25c", The Journal of Chemical Physics **56**, 5, 1905–1909 (1972).

Narten, A. H., "X-ray diffraction pattern and models of liquid benzene", The Journal of Chemical Physics **67**, 5, 2102–2108 (1977).

Narten, A. H. and A. Habenschuss, "Hydrogen bonding in liquid methanol and ethanol determined by x-ray diffraction", The Journal of Chemical Physics **80**, 7, 3387–3391 (1984).

Narten, A. H. and H. A. Levy, "Liquid water: Molecular correlation functions from x-ray diffraction", The Journal of Chemical Physics **55**, 5, 2263–2269 (1971).

Nasr, S., M.-C. Bellissent-Funel and R. Cortes, "X-ray and neutron scattering studies of liquid formic acid dcood at various temperatures and under pressure", The Journal of Chemical Physics **110**, 22, 10945–10952 (1999).

Nocedal, J. and S. Wright, "Numerical optimization, series in operations research and financial engineering", Springer, New York, USA (2006).

Patterson, A. L., "The scherrer formula for x-ray particle size determination", Phys. Rev. **56**, 978–982 (1939).

Pedersen, J. S., "Determination of size distribution from small-angle scattering data for systems with effective hard-sphere interactions", Journal of Applied Crystallography **27**, 4, 595–608 (1994).

Peskin, M. E. and D. V. Schroeder, *An introduction to quantum field theory* (Westview, 1995).

Proffen, T. and S. J. L. Billinge, "*PDFFIT*, a program for full profile structural refinement of the atomic pair distribution function", Journal of Applied Crystallography **32**, 3, 572–575 (1999).

Riekel, C. and F. Vollrath, "Spider silk fibre extrusion: combined wide- and small-angle x-ray microdiffraction experiments", International Journal of Biological Macromolecules **29**, 3, 203 – 210 (2001).

Römer, L. and T. Scheibel, "The elaborate structure of spider silk: Structure and function of a natural high performance fiber", Prion **2**, 4, 154–161 (2008).

Rubinstein, M. and R. H. Colby (Oxford University Press, 2003).

Sampath, S., T. Isdebski, J. E. Jenkins, J. V. Ayon, R. W. Henning, J. P. R. O. Orgel, O. Antipoa and J. L. Yarger, "X-ray diffraction study of nanocrystalline and amorphous structure within major and minor ampullate dragline spider silks", Soft Matter **8**, 6713–6722 (2012).

Schaefer, D. W., "Polymers, fractals, and ceramic materials", Science **243**, 4894, pp. 1023–1027 (1989).

Schmidt-Rohr, K., "Simulation of small-angle scattering curves by numerical Fourier transformation", Journal of Applied Crystallography **40**, 1, 16–25 (2007).

Schmidt-Rohr, K. and Q. Chen, "Parallel cylindrical water nanochannels in nafion fuel-cell membranes", Nat Mater **7**, 1, 75–83 (2008).

Sethna, J. P., *Entropy, Order Parameters, and Complexity* (Oxford University Press, 2006).

Shoji, A., H. Kimura, T. Ozaki, H. Sugisawa, and K. Deguchi, "Conformational study of solid polypeptides by 1h combined rotation and multiple pulse spectroscopy nmr", Journal of the American Chemical Society **118**, 32, 7604–7607 (1996).

Soper, A. K., "On the uniqueness of structure extracted from diffraction experiments on liquids and glasses", Journal of Physics: Condensed Matter **19**, 41, 415108 (2007).

Stanley, H., "Application of fractal concepts to polymer statistics and to anomalous transport in randomly porous media", Journal of Statistical Physics **36**, 5-6, 843–860 (1984).

Teixeira, J., "Small-angle scattering by fractal systems", Journal of Applied Crystallography **21**, 6, 781–785 (1988).

Trancik, J. E., J. T. Czernuszka, C. Merriman and C. Viney, "A simple method for orienting silk and other flexible fibres in transmission electron microscopy specimens", Journal of Microscopy **203**, 3, 235–238 (2001).

van Beek, J. D., S. Hess, F. Vollrath and B. H. Meier, "The molecular structure of spider dragline silk: Folding and orientation of the protein backbone", Proceedings of the National Academy of Sciences **99**, 16, 10266–10271 (2002).

Vollrath, F., "Strength and structure of spiders' silks", Reviews in Molecular Biotechnology **74**, 2, 67 – 83 (2000).

Waasmaier, D. and A. Kirfel, "New analytical scattering-factor functions for free atoms and ions", Acta Crystallographica Section A **51**, 3, 416–431 (1995).

Walford, G. and J. Dore, "Neutron-diffraction studies of the structure of water: Ii. temperature variation effects for heavy water", Molecular Physics **34**, 1, 21–32 (1977).

Weber, J., C. Benmore, A. Tailor, S. Tumber, J. Neuefeind, B. Cherry, J. Yarger, Q. Mou, W. Weber and S. Byrn, "A neutron-x-ray, {NMR} and calorimetric study of glassy probucol synthesized using containerless techniques", Chemical Physics **424**, 0, 89 – 92, neutron Scattering Highlights on Water and Biological Systems (2013).

Xu, M. and R. V. Lewis, "Structure of a protein superfiber: spider dragline silk.", Proceedings of the National Academy of Sciences **87**, 18, 7120–7124 (1990).

Yamauchi, K., S. Kuroki, K. Fujii and I. Ando, "The amide proton {NMR} chemical shift and hydrogen-bonded structure of peptides and polypeptides in the solid state as studied by high-frequency solid-state 1h {NMR}", Chemical Physics Letters **324**, 5–6, 435 – 439 (2000).

Yarusso, D. J. and S. L. Cooper, "Microstructure of ionomers: interpretation of small-angle x-ray scattering data", Macromolecules **16**, 12, 1871–1880 (1983).

Zhou, H. and Y. Zhang, "Hierarchical chain model of spider capture silk elasticity", Phys. Rev. Lett. **94**, 028104 (2005).

Appendix A

The SAXS reconstruction code

The Stimulated Annealing function

```
function [engyf] = mc_move(nx, xsed1, xsed2)

global mtx_xtal occ_bkb
global xtal_pos xtal_sidex xtal_sidey xtal_bkbexcl scl
global n_acpt kT t_acpt egyf
engyf = egyf;
divx = 16;
divy = 16;

pos1x = xtal_pos(xsed1,1);
pos1y = xtal_pos(xsed1,2);
dx = ceil((2*rand(1)—1)/2*nx/divx);     %dx > dy
dy = ceil((2*rand(1)—1)/2*nx/divy);     %rand between —0.5 +0.5

pos1x_new = modnozero(double(pos1x+dx),nx);
pos1y_new = modnozero(double(pos1y+dy),nx);

lmtx = xtal_sidex(xsed1); %sidex = half of actual side length
lmty = xtal_sidey(xsed1);

pos2x = xtal_pos(xsed2,1);
pos2y = xtal_pos(xsed2,2);
dx = ceil((2*rand—1)/2*nx/divx);        %dx > dy
dy = ceil((2*rand—1)/2*nx/divy);        %rand between —0.5 +0.5

pos2x_new = modnozero(double(pos2x+dx),nx);
pos2y_new = modnozero(double(pos2y+dy),nx);

lmtxp = xtal_sidex(xsed2); %sidex = half of actual side length
lmtyp = xtal_sidey(xsed2);

r = sqrt(double((pos1x_new—pos2x_new)^2+(pos1y_new—pos2y_new)^2));

if (occ_bkb(pos1x_new, pos1y_new) == 0) && (occ_bkb(pos2x_new, pos2y_new) == 0) .../
&& (occ_bkb(modnozero(pos1x_new—lmtx,nx),modnozero(pos1y_new—lmty,nx)) == 0) .../
&& (occ_bkb(modnozero(pos1x_new+lmtx,nx),modnozero(pos1y_new+lmty,nx)) == 0) .../
&& (occ_bkb(modnozero(pos1x_new—lmtx,nx),modnozero(pos1y_new+lmty,nx)) == 0)  .../
&& (occ_bkb(modnozero(pos1x_new+lmtx,nx),modnozero(pos1y_new—lmty,nx)) == 0) .../
&& (occ_bkb(modnozero(pos2x_new+lmtxp,nx),modnozero(pos2y_new—lmtyp,nx)) == 0) .../
&& (occ_bkb(modnozero(pos2x_new+lmtxp,nx),modnozero(pos2y_new+lmtyp,nx)) == 0) .../
&& (occ_bkb(modnozero(pos2x_new—lmtxp,nx),modnozero(pos2y_new—lmtyp,nx)) == 0) .../
&& (occ_bkb(modnozero(pos2x_new—lmtxp,nx),modnozero(pos2y_new+lmtyp,nx)) == 0) .../
&& (r > 2.5*max(lmtx,lmtxp))

ttxtal_mtx = mtx_xtal;  %temp matrix to hold structure

for k = —lmty:lmty      %paint crystal1
kyplusk=modnozero(pos1y_new+k,nx);
for j = —lmtx:lmtx
jxplusj = modnozero(pos1x_new+j,nx);
ttxtal_mtx(jxplusj,kyplusk)=0;
end
```

```
end

for j = —lmtx:lmtx      %unpaint crystal1
jxplusj = modnozero(pos1x+j,nx);
for k = —lmty:lmty
kyplusk=modnozero(pos1y+k,nx);
ttxtal_mtx(jxplusj,kyplusk)=1;
end
end

for j = —lmtxp:lmtxp      %paint crystal2
jxplusj = modnozero(pos2x_new+j,nx);
for k = —lmtyp:lmtyp
kyplusk=modnozero(pos2y_new+k,nx);
ttxtal_mtx(jxplusj,kyplusk)=0;
end
end

for j = —lmtxp:lmtxp      %unpaint crystal2
jxplusj = modnozero(pos2x+j,nx);
for k = —lmtyp:lmtyp
kyplusk=modnozero(pos2y+k,nx);
ttxtal_mtx(jxplusj,kyplusk)=1;
end
end

% MCMC
eng_i = egyf;
eng_f = post_p_test(ttxtal_mtx, scl, 1.85);

%       rto = eng_f/eng_i; %Q(x—>x') is symetrical
%       acpt = min(1,1/rto);   %acpt < 1 if the move increase energy
%       a_slot(modnozero(kk,200)) = rto;

dE = eng_f—eng_i;
p_accpt = exp(—dE/kT);

p_test = rand;

if p_test < p_accpt
engyf = eng_f;
n_acpt = n_acpt + 1;
t_acpt = t_acpt + 1;
mtx_xtal = ttxtal_mtx;

exlcu1x = xtal_bkbexcl(xsed1,1);
exlcu1y = xtal_bkbexcl(xsed1,2);
exlcu2x = xtal_bkbexcl(xsed2,1);
exlcu2y = xtal_bkbexcl(xsed2,2);

for j = —exlcu1x:exlcu1x      %unpaint crystal1 exclu
jxplusj = modnozero(pos1x+j,nx);
for k = —exlcu1y:exlcu1y
```

```
kyplusk=modnozero(pos1y+k,nx);
occ_bkb(jxplusj,kyplusk)=occ_bkb(jxplusj,kyplusk)—1;
end
end

for j = —exlcu1x:exlcu1x      %paint crystal1 exclu
jxplusj = modnozero(pos1x_new+j,nx);
for k = —exlcu1y:exlcu1y
kyplusk=modnozero(pos1y_new+k,nx);
occ_bkb(jxplusj,kyplusk)=occ_bkb(jxplusj,kyplusk)+1;
end
end

for j = —exlcu2x:exlcu2x      %unpaint crystal1 exclu
jxplusj = modnozero(pos2x+j,nx);
for k = —exlcu2y:exlcu2y
kyplusk=modnozero(pos2y+k,nx);
occ_bkb(jxplusj,kyplusk)=occ_bkb(jxplusj,kyplusk)—1;
end
end

for j = —exlcu2x:exlcu2x      %paint crystal1 exclu
jxplusj = modnozero(pos2x_new+j,nx);
for k = —exlcu2y:exlcu2y
kyplusk=modnozero(pos2y_new+k,nx);
occ_bkb(jxplusj,kyplusk)=occ_bkb(jxplusj,kyplusk)+1;
end
end

xtal_pos(xsed1,1) = pos1x_new;
xtal_pos(xsed1,2) = pos1y_new;
xtal_pos(xsed2,1) = pos2x_new;
xtal_pos(xsed2,2) = pos2y_new;
end
end

end
```

The FFT transfomation function

```
function [In, Iq] = SAXSrodcylformfac(rho)
% take density map rho as imput
global nx

nxf = nx;
% can be smaller than in SAXSpreprodcyl, to speed up calculation
nxfd2=nxf/2;
iqcent=nxfd2+1;

Ampl=fft2(rho);
Ampl=fftshift(Ampl);  %center of grid (nxd2+1) is at q=0
```

```
% ................. I(q)=|A(q)|^2 .....................
Iq=abs(Ampl).^2;

iq1=(1:nx)-iqcent;
%elementary square
q1ad2=pi*iq1/nx+0.00000001;   % q1*a/2 = iq1c*2pi/(nx*a) *a/2
formf=(sin(q1ad2)./q1ad2).^2;

Iscatt = zeros(nxf,1);

iqmin=iqcent-nxfd2;
iqmax=iqcent+nxfd2-1;

parfor iq1=iqmin:iqmax  %
temp1 = zeros(nxf,1);
%temp2 = zeros(nxf,1);
formx = formf(iq1);
for iq2=iqmin:iqmax  %
formy = formf(iq2);
for iq3=iqcent:iqcent  %thin sheet along q3

qabs=sqrt((iq1-iqcent)^2+(iq2-iqcent)^2)+1; % +1: prevent Iscatt(iqabs=0)
iqabs=round(qabs);
ishar=qabs-iqabs;

share=abs(ishar);       % how much to share
isignshare=sign(ishar);  % +1 if q > iq; 0 if exactly on

addI=Iq(iq1,iq2)*formx*formy;  %sincsqr: effect of elementary square

temp1(iqabs)=addI*(1-share);
temp1(iqabs+isignshare)=addI*share; % 0 if exactly on
end
end
Iscatt = Iscatt + temp1;
end
%-----------------------------------------
% calculate I(q)
%-----------------------------------------
Isum=0;
parfor iq=2:nxfd2
Isum=Isum+Iscatt(iq);
In(iq-1)=Iscatt(iq)/(iq-1)^2;  %iq=0 really q=0
%In2D(iq-1)=Iscatt(iq)/(iq-1);  %in-plane averaging
end
end
```

The function that calculates RDF from the crystal population.

```
function [pdf] = rdf(xtal_pos)

n = length(xtal_pos);
```

```
nmax = int16(ceil(sqrt(2*(2048—1)^2)));

pdf = zeros(nmax,1);

for i = 1:(n—1)
for j = i+1 : n

r_ix = double(xtal_pos(i,1));
r_iy = double(xtal_pos(i,2));
r_jx = double(xtal_pos(j,1));
r_jy = double(xtal_pos(j,2));

r_ij = sqrt((r_ix—r_jx)^2+(r_iy—r_jy)^2);

if r_ij < 2
disp(sprintf('r=(%d, %d), id=(%d, %d)', r_ix, r_iy, i, j));
end
irabs = round(r_ij);
ishar = r_ij — irabs;
shar = abs(ishar);
isignshar = sign(ishar);

pdf(irabs+1) = pdf(irabs+1) + (1 — shar);
pdf(irabs+isignshar+1) = pdf(irabs+isignshar+1) + shar;

end
end

end
```

Appendix B

Structure factor calculation code

C++ code that calculates the structure factor equation 4.3.

```
#include "mex.h"
#include <matrix.h>
#include <string.h>
#include <assert.h>
#include <math.h>
#include <omp.h>

// nx is r, nxq is xdata, strlN is atom_lst length respectively, nyf is atomff
    ↪ population
static mwSize nx, nxq, strlN, nxf, nyf;

inline void clear(double *lst) {
for (int i=0; i<nxq; i++) {
lst[i] = 0;
}
}

inline void J(double *reslt, double r, double *xdata) {
for (int i=0; i<nxq; i++) {
double t1;
t1 = r*xdata[i];
reslt[i] = sin(t1)/t1;
}
}

inline void addconst(double *reslt, double *x, const double y) {
for (int i=0; i<nxq; i++) {
reslt[i] = x[i]+y;
}
}

inline void dotp(double *reslt, double *x, double *y) {
for (int i=0; i<nxq; i++) {
reslt[i] = x[i]*y[i];
}
}

inline void vtime(double *reslt, double x, double *y) {
for (int i=0; i<nxq; i++) {
reslt[i] = x*y[i];
}
}


inline void add(double *reslt, double *x, double *y) {
for (int i=0; i<nxq; i++) {
reslt[i] = x[i]+y[i];
}
}

inline void expX(double *reslt, const double pre, const double x, double *xdata) {
// calculate pre*exp(x*xdata.^2)
```

```
dotp(reslt, xdata, xdata);  //reslt hold xdata.^2
for (int i=0; i<nxq; i++) {
reslt[i] = pre*exp(x*reslt[i]);
}
}

//nxf should be 11, formfactor elements for each atom, nyf should be 10: the current
    ↪ list length
void F(double *reslt, char *a, double *xdata, const double *formf, char**atomff_idx) {
int flag = 0;
int i, k;
double *fct, *tmp, *tmp1, *xsq;
fct = (double *)mxMalloc(nxf*sizeof(double));

tmp = (double *)mxMalloc(nxq*sizeof(double));   //hold tmp vector of xdata size
tmp1 = (double *)mxMalloc(nxq*sizeof(double));
xsq = (double *)mxMalloc(nxq*sizeof(double));

clear(tmp);
clear(tmp1);
clear(xsq);

for (k=0; k<nyf; ++k) {
if (strcmp(a, atomff_idx[k]) == 0) {
//mexPrintf("found atom %s\n", a);
for (i=11; i--;) {          //subsitute 11 with nxf if nxf is not 11 anymore (list
    ↪ change)
fct[i] = formf[k*11 + i];
}
flag = 1;
}
}
if (flag == 0)
mexErrMsgIdAndTxt( "MATLAB:sqfactor:F",
"Can't find atom in the form factor list.");

//mxAssert(flag, "mxAssert:F atom not found in form factor list, please update the
    ↪ list manually.");
vtime(xsq, 0.0795775, xdata); //s = q/(4*pi)

expX(tmp, fct[0], -fct[1], xsq);
add(tmp1, tmp1, tmp);
expX(tmp, fct[2], -fct[3], xsq);
add(tmp1, tmp1, tmp);
expX(tmp, fct[4], -fct[5], xsq);
add(tmp1, tmp1, tmp);
expX(tmp, fct[6], -fct[7], xsq);
add(tmp1, tmp1, tmp);
expX(tmp, fct[8], -fct[9], xsq);
add(tmp1, tmp1, tmp);
clear(reslt);
addconst(reslt, tmp1, fct[10]);
mxFree(fct);
```

```
mxFree(tmp);
mxFree(tmp1);
mxFree(xsq);
}

void sqfactor(double *x, double *xdata, double *res, const double *r, const double *
    ↪ formf, char **atom_lst, char **atomff_idx, char **plst1, char **plst2) {
double *f1, *f2, *ft, *tmp1, *norm, *v_private;
f1 = (double *)mxMalloc(nxq*sizeof(double));
f2 = (double *)mxMalloc(nxq*sizeof(double));
ft = (double *)mxMalloc(nxq*sizeof(double));
//zero vectors
tmp1 = (double *)mxMalloc(nxq*sizeof(double));
norm = (double *)mxMalloc(nxq*sizeof(double));

clear(tmp1);
clear(norm);

//extract string name of atom pair
//fc = fc + f(a1,xdata).*f(a2,xdata).*J(r(k)*xdata).*exp(-0.5*x(k)^2*xdata.^2);
#pragma omp parallel
{
const int nthreads = omp_get_num_threads();
const int ithread = omp_get_thread_num();
#pragma omp master
{
v_private = (double *)mxMalloc(nthreads*nxq*sizeof(double));
for(int i=0; i<(nxq*nthreads); i++)
v_private[i] = 0;
}
#pragma omp for
for (int i=0; i<nx; i++) {
F(f1, plst1[i], xdata, formf, atomff_idx);
F(f2, plst2[i], xdata, formf, atomff_idx);
dotp(ft, f1, f2);            //f hold f1*f2
J(f1, r[i], xdata);         //f1 hold J(r.*xdata)
expX(f2, 1.0, -0.5*x[i]*x[i], xdata);  //f2 hold exp(...)
dotp(f1, f1, f2);       //f1 hold J()*exp()
dotp(f2, ft, f1);            //f2 hold f1()*f2()*J()*exp()
for (int k=0; k<nxq; k++) {
v_private[ithread*nxq+k] += f2[k];
}
}
for (int i=0; i<nthreads; i++) {
for (int k=0; k<nxq; k++) {
tmp1[k] += v_private[i*nxq+k];
}
}
for (int i=0; i<strlN; i++) {
F(f1, atom_lst[i], xdata, formf, atomff_idx);    //f1 hold f()
add(norm, norm, f1);
}
dotp(ft, norm, norm);           //ft = norm.^2
```

```
#pragma omp for
for (int i=0; i<nxq; i++) {
double t = tmp1[i];
tmp1[i] = t/ft[i];       //tmp1 hold f./normf the normalized
}
#pragma omp for
for (int i=0; i<nxq; i++) {
double m1, m2;
res[i] = x[nx]*tmp1[i] + x[nx+1]/xdata[i];
}
}   //pragma omp parallel
mxFree(f1);
mxFree(f2);
mxFree(ft);
mxFree(tmp1);
mxFree(norm);
mxFree(v_private);
}

void mexFunction( int nlhs, mxArray *plhs[],
int nrhs, const mxArray *prhs[])
{

//load all required global parameters
//-----------------------------------------------------------------------------
mxArray *array_r;
mxArray *matx_formf;
const char *rname = "r";
const char *ffname = "formf";
const char *atype = "atom_type";
const char *affindex = "atomff_index";
const char *pname1 = "pinfo1";
const char *pname2 = "pinfo2";

//2D char matrix hold global variable
int status;
char **atom_lst, **atomff_idx;
char **plst1, **plst2;

//-----------------------------------------------------------------------------
// load global r
array_r = mexGetVariable("global", rname);
nx = mxGetM(array_r);   //r length
const double *r = mxGetPr(array_r);

// load global formf matrix
matx_formf = mexGetVariable("global", ffname);  //should be 11x10 matrix, after
    ↪ reading should be 10x11(original size)

nxf = mxGetM(matx_formf);     //x is col order, vertical
nyf = mxGetN(matx_formf);     //y is row order, horizontal

const double *formf = mxGetPr(matx_formf);      //formf is row-major order now
```

```
//-------------------------------------------------------------------------------------
// load atom_lst
mxArray *atom_idx;
char *buf;

atom_idx = mexGetVariable("global", atype);
buf = mxArrayToString(atom_idx);
strlN = mxGetN(atom_idx);   //atom list length, atom population
mwSize strsN = mxGetM(atom_idx);


atom_lst = (char**)mxMalloc(strlN*sizeof(char*));
for (int i=0; i<strlN; ++i) {
atom_lst[i] = (char *)mxMalloc((strsN+1)*sizeof(char));
}
for (int i=0; i<strlN; ++i) {
int j=2*i;
strncpy(atom_lst[i], buf+j, 2);
atom_lst[i][strsN] = '\0';
//mexPrintf("%s", atom_lst[i]);
}

// load atomff_index
mxArray *atom_idx1;
char *buf1;

atom_idx1 = mexGetVariable("global", affindex);
buf1 = mxArrayToString(atom_idx1);
mwSize strlN1 = mxGetN(atom_idx1);
mwSize strsN1 = mxGetM(atom_idx1);
atomff_idx = (char**)mxMalloc(strlN1*sizeof(char*));
for (int i=0; i<strlN1; ++i) {
atomff_idx[i] = (char *)mxMalloc((strsN1+1)*sizeof(char));
}
for (int i=0; i<strlN1; ++i) {
int j=2*i;
strncpy(atomff_idx[i], buf1+j, 2);
atomff_idx[i][strsN1] = '\0';
//mexPrintf("%s", atomff_idx[i]);
}

// load pinfo1
mxArray *atom_idx2;
char *buf2;

atom_idx2 = mexGetVariable("global", pname1);
buf2 = mxArrayToString(atom_idx2);
mwSize strlN2 = mxGetN(atom_idx2);
mwSize strsN2 = mxGetM(atom_idx2);
plst1 = (char**)mxMalloc(strlN2*sizeof(char*));
for (int i=0; i<strlN2; ++i) {
plst1[i] = (char *)mxMalloc((strsN2+1)*sizeof(char));
}
```

```
for (int i=0; i<strlN2; ++i) {
int j=2*i;
strncpy(plst1[i], buf2+j, 2);
plst1[i][strsN2] = '\0';
//mexPrintf("%s", plst1[i]);
}

// load pinfo2, use the same buffer parameters, since they are same size
atom_idx2 = mexGetVariable("global", pname2);
buf2 = mxArrayToString(atom_idx2);
strlN2 = mxGetN(atom_idx2);
strsN2 = mxGetM(atom_idx2);
plst2 = (char**)mxMalloc(strlN2*sizeof(char*));
for (int i=0; i<strlN2; ++i) {
plst2[i] = (char *)mxMalloc((strsN2+1)*sizeof(char));
}
for (int i=0; i<strlN2; ++i) {
int j=2*i;
strncpy(plst2[i], buf2+j, 2);
plst2[i][strsN2] = '\0';
//mexPrintf("%s", plst2[i]);
}

//load all required local input parameters
//-------------------------------------------------------------------------------------
double *x;
double *xdata;
double *res;
nxq = mxGetM(prhs[1]);      //xdata length;

x = mxGetPr(prhs[0]);
xdata = mxGetPr(prhs[1]);

plhs[0] = mxCreateDoubleMatrix(nxq, 1, mxREAL);
res = mxGetPr(plhs[0]);
//call working function
sqfactor(x, xdata, res, r, formf, atom_lst, atomff_idx, plst1, plst2);

// clean mxArray
mxDestroyArray(array_r);
mxDestroyArray(atom_idx);
mxDestroyArray(atom_idx1);
mxDestroyArray(atom_idx2);
mxDestroyArray(matx_formf);

// free dynamically allocated memory
for (int i=0; i< strlN; ++i) {
mxFree(atom_lst[i]);
}
mxFree(atom_lst);
for (int i=0; i< strlN1; ++i) {
mxFree(atomff_idx[i]);
}
```

```
mxFree(atomff_idx);
for (int i=0; i< strlN2; ++i) {
mxFree(plst1[i]);
}
mxFree(plst1);
for (int i=0; i< strlN2; ++i) {
mxFree(plst2[i]);
}
mxFree(plst2);
}
```