RICE UNIVERSITY

Advanced Computational Methods for Macromolecular Modeling and Structure Determination

by

Chong Zhang

A THESIS SUBMITTED
IN PARTIAL FULFILLMENT OF THE
REQUIREMENTS FOR THE DEGREE

Doctor of Philosophy

APPROVED, THESIS COMMITTEE

Dr. Jianpeng Ma, Chair, Professor of Bioengineering, Rice University and Lodwick T. Bolin Professor of Biochemistry, Baylor College of Medicine

Dr. Peter Nordlander, Professor of Physics & Astronomy, Professor of Electrical & Computer Engineering, Director of Rice Quantum Institute, Rice University

Dr. Ching-Hwa Kiang, Professor of Physics & Astronomy, Professor of Bioengineering, Rice University

Dr. Robert Raphael, Associate Professor of Bioengineering, Rice University

HOUSTON, TEXAS
OCTOBER 2013
ABSTRACT

Advanced Computational Methods for Macromolecular Modeling and Structure Determination

by

Chong Zhang

As volume and complexity of macromolecules increase, theories and algorithms that deal with structure determination at low X-ray resolution are of particular importance. With limited diffraction data in hand, experimentalists rely on advanced computational tools to extract and utilize useful information, seeking to determinate a three dimensional model that best fits the experiment data. Success of further studies on the property and function of a specific molecule – the key to practical applications – is therefore heavily dependent on the validity and accuracy of the solved structure.

In this thesis I propose Deformable Complex Network (DCN) and introduce Normal Mode Analysis (NMA), which are designed to model the average coordinates of atoms and associated fluctuations, respectively. Their applications on structure determination target two major branches – the positional refinement and temperature factor refinement. I demonstrate their remarkable performance in structure improvements based on several criteria, such as the free R value, overfitting effect and Ramachandran Statistics, with tests carried out across a broad range of real systems for generality and consistency.
Acknowledgments

I would like to express my gratitude to Applied Physics Program, which provided an ideal environment for learning fundamental physics and exploring at the interface between physical science and engineering, and to my advisor, who guided me in the past three years, encouraged me to think independently, and gave me sufficient freedom to design and perform research throughout my Ph.D. career.

I am grateful to the members of my committee, Dr. Peter Nordlander, Dr. Ching-Hwa Kiang and Dr. Robert Raphael, for their valuable suggestions and insightful comments on the improvement of this work. My thanks also go to all faculty and staff at Rice Quantum Institute who provided numerous research resources in both School of Natural Sciences and School of Engineering, and to people at Research Computing Support Group who offered strong hardware and technical support to the high performance computing community.

I enjoy discussing academic problems, exchanging opinions, and hanging out with my friends, group mates and fellow Applied Physics students. Times spent with Yunxiang, Dong, Zhongyu, Kit, Mingyang, Cheng, Brian and others make graduate life at Rice much more colorful and memorable.

I appreciate Dr. Guanghong Wei, Dr. Zuimin Jiang, Dr. Li Zhao and Ms. Jinglin Lu’s mentorship during my undergraduate physics study and support on my application for an advanced degree. Dr. Luwei Zhou and Dr. Xiaoyuan Hou’s personalities as a physicist constantly remind me to maintain the highest ethical standards in all situations. I am also thankful to Mr. Xiang Li, Ms. Zhilan Cui, Mr. Jianhua Jiang and Mr. Qiang Feng,
who brought me to the world of physics, awakened my curiosity, and grew my interest in science. All these mentors let me believe that physics is all about people communicating for truth—a communication that is free and equal, regardless of position, and a truth that is pure and consistent, invariant to standpoint.

Last but foremost, I would like to dedicate this thesis to my parents, Mr. Xinsheng Zhang and Ms. Guofen Xu, who have educated and supported me for more than twenty-six years. I am deeply indebted to them for their understanding of my faith, respect for the career I chose, and unconditional love.
Contents

Acknowledgments ............................................................................................................. ii

Contents ............................................................................................................................ iv

List of Figures ................................................................................................................... vi

List of Tables .................................................................................................................. viii

List of Equations .............................................................................................................. ix

Introduction to X-ray Diffraction and Structure Determination ................................. 1
  1.1. Fundamentals of X-ray diffraction ........................................................................... 2
  1.2. Application on macromolecular structure determination ......................................... 9
  1.3. Refinement ............................................................................................................. 11

Elements of Refinement .................................................................................................. 13
  2.1. Parameters .............................................................................................................. 13
  2.2. Target function ....................................................................................................... 21
  2.3. Optimization .......................................................................................................... 31
  2.4. Progress indicators and validation .......................................................................... 35

Deformable Complex Network ...................................................................................... 44
  3.1. Motivation and introduction ................................................................................... 44
  3.2. Method ................................................................................................................... 46
    3.2.1. Summary .......................................................................................................... 46
    3.2.2. Target function ................................................................................................. 47
    3.2.3. DAN and DCN approach ................................................................................. 48
      3.2.3.1. Introduction ............................................................................................... 48
      3.2.3.2. DAN/DCN modes ..................................................................................... 49
      3.2.3.3. DCN energy restraint equations ................................................................. 52
    3.2.4. Selection of reference model and \( (\gamma, w_{DCN}, \mu) \) parameter group .......... 53
    3.2.5. Input data preparation before refinement ........................................................ 55
    3.2.6. Refinement protocol ........................................................................................ 57
    3.2.7. Coding and program ........................................................................................ 58
  3.3. Results and analysis ............................................................................................... 58
    3.3.1. Automatic full refinement ................................................................................ 58
    3.3.2. Automatic re-refinements ................................................................................ 63
3.3.2.1. Results overview ....................................................................................... 64
3.3.2.2. Decrease in $R_{\text{free}}$ .................................................................................. 65
3.3.2.3. Decrease in $R_{\text{free}} - R_{\text{work}}$ .................................................................. 66
3.3.2.4. Increase in Ramachandran Statistics ......................................................... 67
3.3.2.5. Improvement in electron density map interpretation ................................ 68
3.3.2.6. Re-refinement with NCS and experimental phase .................................... 72
3.4. Supplementary information .............................................................................. 75
3.5. Discussion and implementation ......................................................................... 78

Normal Mode Analysis and its Integration with DCN ........................................... 79

4.1. Motivation and introduction ............................................................................. 79
4.2. Method .............................................................................................................. 80
4.2.1. Summary ..................................................................................................... 80
4.2.2. Target function ........................................................................................... 81
4.2.3. Normal Mode Analysis .............................................................................. 82
4.2.3.1. Introduction ............................................................................................ 82
4.2.3.2. Modeling anisotropic atomic motions with Normal Mode Analysis .......... 83
4.2.4. Integrating NMA with DCN ....................................................................... 88
4.2.5. Refinement protocol ............................................................................... 94
4.2.6. Coding and program ............................................................................. 96
4.3. Results and analysis ..................................................................................... 96
4.3.1. Case 1: PDB ID 1JL4 ............................................................................... 96
4.3.1.1. The $R_{\text{free}}$ value .................................................................................. 97
4.3.1.2. Overfitting effect ............................................................................... 98
4.3.1.3. Ramachandran Statistics ..................................................................... 99
4.3.2. Case 2: PDB ID 2YHJ ................................................................ .......... 101
4.3.2.1. The $R_{\text{free}}$ value ............................................................................... 101
4.3.2.2. Overfitting effect ............................................................................. 103
4.3.2.3. Ramachandran Statistics ................................................................... 104
4.3.3. Case 3: PDB ID 1YM7 .......................................................................... 105
4.3.3.1. The $R_{\text{free}}$ value ............................................................................. 105
4.3.3.2. Overfitting effect ............................................................................. 107
4.3.3.3. Ramachandran Statistics ................................................................... 108
4.4. Conclusion and discussion ............................................................................ 109

References ........................................................................................................ 112
List of Figures

Figure 1-1  Bragg Law and Diffraction............................................................................. 3

Figure 1-2  Workflow of crystallographic macromolecular structure determination
........................................................................................................................................... 10

Figure 1-3  Work flow of refinement............................................................................. 11

Figure 2-1  Isotropic oscillation shell — a sphere ........................................................ 18

Figure 2-2  Anisotropic oscillation shell – an ellipsoid whose principle axes do not
have to be x,y,z coordinate axes..................................................................................... 19

Figure 2-3  Dihedral angle of four sequential atoms ................................................... 28

Figure 2-4  Optimization of a target function with sophisticated energy landscape
using gradient minimization .......................................................................................... 32

Figure 2-5  Optimization of a target function with sophisticated energy landscape
using simulated annealing .............................................................................................. 33

Figure 2-6  Optimization of a target function with sophisticated energy landscape
using grid search ............................................................................................................. 34

Figure 2-7  Relationship between GDT (and MaxSub) score for random structure
pairs and length of proteins$^{19}$ ......................................................................................... 40

Figure 2-8  Relationship between TMscore (and rTMscore with $d_0 = constant$) for
random structure pairs and length of proteins$^{19}$ ........................................................... 41

Figure 2-9  An example of Ramachandran Plot$^1$ ........................................................... 42

Figure 3-1  Directional mode (D-mode) of DAN/DCN .................................................... 50

Figure 3-2  Arbitrary mode (A-mode) of DAN/DCN....................................................... 51

Figure 3-3  3D grid search for best parameter set $(\gamma, w_{DCN}, \mu)$ .................................. 55

Figure 3-4  $R_{free}$ vs Resolution for Conventional, DEN and DCN ............................... 60

Figure 3-5  RMSD vs Resolution for Conventional, DEN and DCN............................... 61
Figure 3-6  GDT(<1Å) vs Resolution for Conventional, DEN and DCN................. 61
Figure 3-7  TMscore vs Resolution for Conventional, DEN and DCN ............... 62
Figure 3-8  \( R_{\text{free}} \) of sixteen test systems for Conventional, DEN and DCN........... 66
Figure 3-9  \( R_{\text{free}} - R_{\text{work}} \) of sixteen test systems for Conventional, DEN and DCN... 67
Figure 3-10  Ramachandran Statistics of sixteen test systems for Conventional, DEN and DCN .......................................................................................................................... 68
Figure 3-11  View of backbone trace. PDB ID 1JL4 centered on A23-THR is shown. ........................................................................................................................................... 71
Figure 3-12  View of a remarkable branch deviation. PDB ID 2BF1 centered on A368-GLY is shown. ....................................................................................................... 72
Figure 3-13  \( R_{\text{free}} \) vs availability of NCS information ................................................. 74
Figure 3-14  \( R_{\text{free}} \) (and Ramachandran) vs availability of experiment phase........... 74
Figure 4-1  System in internal coordinates ............................................................................. 87
Figure 4-2  DCN and NMA refinements: comparison of free R value for 1JL4....... 97
Figure 4-3  DCN and NMA refinements: comparison of \( R_{\text{free}} - R_{\text{work}} \) for 1JL4........... 99
Figure 4-4  DCN and NMA refinements: Comparison of Ramachandran Statistics for 1JL4........................................................................................................................................ 100
Figure 4-5  DCN and NMA refinements: Comparison of free R value for 1JL4..... 102
Figure 4-6  DCN and NMA refinements: Comparison of \( R_{\text{free}} - R_{\text{work}} \) for 2YHJ.... 103
Figure 4-7  DCN and NMA refinements: Comparison of Ramachandran Statistics for 2YHJ ........................................................................................................................ 104
Figure 4-8  DCN and NMA refinements: Comparison of free R value for 1YM7.. 106
Figure 4-9  DCN and NMA refinements: Comparison of \( R_{\text{free}} - R_{\text{work}} \) for 1YM7 ... 107
Figure 4-10  DCN and NMA refinements: Comparison of Ramachandran Statistics for 1YM7........................................................................................................................................ 108
List of Tables

Table 3-1  Refinement of tobacco PR-5d protein (PDB ID 1AUN) ............................ 62
Table 3-2  Results of sixteen low-resolution re-refinement tasks. .......................... 64
Table 3-3  Refinement with and without NCS or experiment phase information .... 73
Table 3-4  A list of the structure property of all the re-refinement cases ................. 75
Table 3-5  A list of property of experiment data and reference model .................... 76
Table 3-6  Comparison of results between this work (ligands included) and previous work$^{13}$ (ligands excluded) with Conventional and DEN approach ......................... 77
Table 4-1 DCN and NMA refinements: free R value for 1JL4 ............................... 97
Table 4-2 DCN and NMA refinements: $R_{\text{free}} - R_{\text{work}}$ for 1JL4 ...................... 98
Table 4-3 DCN and NMA refinements: Ramachandran Statistics for 1JL4 .............. 100
Table 4-4 DCN and NMA refinements: free R value for 2YHJ ............................. 101
Table 4-5 DCN and NMA refinements: $R_{\text{free}} - R_{\text{work}}$ for 2YHJ .................. 103
Table 4-6 DCN and NMA refinements: Ramachandran Statistics for 2YHJ .......... 104
Table 4-7  DCN and NMA refinements: free R value for 1YM7 ......................... 105
Table 4-8  DCN and NMA refinements: $R_{\text{free}} - R_{\text{work}}$ for 1YM7 .................. 107
Table 4-9  DCN and NMA refinements: Ramachandran Statistics for 1YM7 ........ 108
List of Equations

Equation 1-1 Relationship between X-ray wave vector and wavelength.............. 2
Equation 1-2 Bragg’s Law.................................................................................. 3
Equation 1-3 Phase difference between two atoms in a crystal......................... 4
Equation 1-4 definition of wave vector difference ............................................ 4
Equation 1-5 Structure factor for a monoatomic crystal with a basis................. 5
Equation 1-6 Diffraction amplitude expression in terms of structure factor.......... 5
Equation 1-7 Intensity, interference coefficient and structure factor.................. 6
Equation 1-8 Reducing the interference coefficient.......................................... 6
Equation 1-9 Condition when \( I_c \) reaches maximum................................. 6
Equation 1-10 structure factor for molecular crystal..................................... 7
Equation 1-11 atomic form factor and electron density................................. 8
Equation 1-12 structure factor and electron density....................................... 8
Equation 1-13 electron density and structure factor....................................... 8
Equation 2-1 Instantaneous atom position....................................................... 14
Equation 2-2 Structure factor time averaged................................................. 14
Equation 2-3 Reducing \( \langle e^{ih \cdot r_i} \rangle \)................................................................. 15
Equation 2-4 U matrix....................................................................................... 15
Equation 2-5 U matrix definition................................................................. 16
Equation 2-6 Introducing B factor................................................................. 16
Equation 2-7 Definition of isotropic B factor................................................. 16
Equation 2-8 Alternative form of definition of isotropic B factor...................... 17
Equation 2-9  Debye-Waller factor in terms of B factor, incident angle and wavelength ................................................................................................................................................................................. 17

Equation 2-10  Total calculated structure factor with flat density bulk solvent...... 21

Equation 2-11  Refinement target, i.e., Total potential energy ................................................. 22

Equation 2-12  experiment potential with electron density map ........................................ 22

Equation 2-13  Least Squares target function .............................................................................. 23

Equation 2-14  Likelihood of observing a diffraction pattern given a model .......... 24

Equation 2-15  Maximum Likelihood target function definition ........................................ 25

Equation 2-16  Detailed expression of likelihood of one observation given a model 25

Equation 2-17  Data-to-parameter ratio ...................................................................................... 26

Equation 2-18  Data-to-parameter-ratio with constraints ......................................................... 26

Equation 2-19  Data-to-parameter ratio with restraints .......................................................... 27

Equation 2-20  Bond length restraint energy ................................................................................... 27

Equation 2-21  Bond angle restraint energy .................................................................................... 28

Equation 2-22  Dihedral angle energy restraint .................................................................................... 29

Equation 2-23  Planarity restraint energy ....................................................................................... 29

Equation 2-24  Chiral Volume of a C alpha atom........................................................................ 30

Equation 2-25  Chirality energy restraint ....................................................................................... 30

Equation 2-26  non-bonded restraint energy .................................................................................. 30

Equation 2-27  Force derived from the gradient........................................................................... 32

Equation 2-28  Definition of $R$ value ......................................................................................... 35

Equation 2-29  Definition of the work and free $R$ values ....................................................... 37

Equation 2-30  Definition of RMSD ......................................................................................... 38

Equation 2-31  Definition GDT Total_Score ................................................................................. 38
Equation 2-32  Definition of TMscore ................................................................. 39
Equation 2-33  $d_0$ as a function of $L_\infty$ in TMscore ........................................ 40
Equation 2-34  Calculation of Ramachandran Statistics ...................................... 42
Equation 3-1  Refinement Target of DCN .............................................................. 47
Equation 3-2  DCN restraint energy ....................................................................... 52
Equation 3-3  Deformability of DCN ..................................................................... 53
Equation 4-1  Anisotropic Debye-Waller factor ..................................................... 81
Equation 4-2  Isotropic Debye-Waller factor .......................................................... 81
Equation 4-3  Anisotropic U matrix ....................................................................... 82
Equation 4-4  atomic displacement via normal mode parameter .......................... 84
Equation 4-5  Eigenvector matrix .......................................................................... 84
Equation 4-6  Eigenvalue vector ............................................................................. 85
Equation 4-7  Hessian Matrix ................................................................................ 85
Equation 4-8  Hamiltonian for modified eNMA ....................................................... 88
Equation 4-9  Atomic form factor by 5-Gaussian approximation ............................ 89
Equation 4-10 Atomic electron density by 5-Gaussian approximation .................... 89
Equation 4-11  Isotropic atomic form factor by 5-Gaussian approximation ............ 90
Equation 4-12 Isotropic atomic electron density by 5-Gaussian approximation ...... 90
Equation 4-13 Anisotropic Atomic form factor by 5-Gaussian approximation ......... 91
Equation 4-14 Anisotropic atomic total U matrix .................................................... 91
Equation 4-15 Anisotropic atomic electron density ............................................... 92
Equation 4-16 Inverse of anisotropic atomic total U matrix .................................... 92
Equation 4-17 Positional derivatives of electron density ........................................ 93
Equation 4-18 Anisotropic U derivatives of electron density........................................... 94
Chapter 1

Introduction to X-ray Diffraction and Structure Determination

Studies on property of advanced materials and function of biomolecules require an accurate picture of three dimensional structure at atomic level. With increased complexity, flexibility and volume of a molecule, theories and algorithms designed to efficiently carry out automatic structure determination on high performance computing (HPC) clusters, especially those capable of deriving a structure with experimental data bearing low quality and completeness, are of particular interest and importance. Despite a wide range of applications in chemistry, biology and materials science, fundamental studies on advancement of structure determination, however, are mostly related to mathematics, statistics, physics and computation. This thesis will focus on X-ray diffraction based method – the most widely used one for macromolecular determination to date.
1.1. Fundamentals of X-ray diffraction

Experimental techniques used to explore the atomic world of macromolecules include X-ray diffraction, nuclear magnetic resonance (NMR), electron microscopy, etc. Here I only focus on introduction to principles of X-ray diffraction and algorithm that deals with experiment data collected by X-ray equipment.

X-ray can be produced through an X-ray tube, which utilizes high voltage to accelerate electrons. These electrons hit a metal target at very high velocities, releasing X-rays due to energy level transition. The X-ray photon is also a form of electromagnetic wave, whose wavelength approximately falls between 0.01nm – 10nm. As this wavelength is comparable to crystal lattice, incident beam can be easily diffracted, leading to it being an ideal tool for detecting atomic structures, Conventionally, we denote the incident wave vector and wavelength with $\mathbf{k}$ and $\lambda$, respectively. The direction of $\mathbf{k}$ indicates the propagation of the X-ray, while the magnitude is defined to satisfy the following condition

$$|\mathbf{k}| = \frac{2\pi}{\lambda}$$

Equation 1-1  Relationship between X-ray wave vector and wavelength

When crystals with periodic arrangement of atomic structures are bombarded by X-ray, reflected X-ray beams can be observed at certain incidence angles. Figure 1-1 illustrates the situation in which two rays in parallel and in phase bombard two layers of a crystal. The dots within the crystal may represent atoms, icons and even molecules
(including macromolecules). $d$ denotes the distance between the two layers and $\theta$ the angle between the incident beam and crystal layer plane. We let $\vec{k}'$ be the scattered X-ray wave vector.

![Figure 1-1 Bragg Law and Diffraction](image)

Suppose that the scattering is elastic. Therefore, the X-ray photon energy – or the wavelength in terms of wave – conserves before and after being diffracted by the lattice grids. In order to produce a constructive interference, the two outbound rays have to be in phase as well. It implies an identity stating that the optical path difference of these two rays is an integer multiple of their wavelength.

$$2d \sin \theta = n\lambda \quad n = 1,2,3...$$

**Equation 1-2 Bragg’s Law**
Bragg’s Law describes the simplest case of diffraction. There exist other statements such as Lauer condition, which could be proven equivalent. The equation shows that by rotating the ray/crystal and changing the incident angle, diffractions patterns of higher orders, referred to as Bragg peaks, are to be observed. This idea forms the principle of modern X-ray crystallography, where lattice grids become larger and more complex. By studying experiment patterns diffracted by target crystal during continuous orientation change, researchers are able to decipher the exact 3D-structure of a macromolecule, with the aid of powerful computation resources, advanced mathematical algorithm and sometimes other a priori information.

Suppose that we have a monoatomic crystal with a basis in hand. Two atoms within the same cell positioning at $R_1$ and $R_2$ will scatter the X-ray with a phase difference that equals

$$\Delta \phi_{1,2} = (R_1 - R_2) \cdot h$$

**Equation 1-3 Phase difference between two atoms in a crystal**

where $h$ is a Bragg peak and defined as the wave vector difference between the scattered and incident X-rays

$$h \equiv k' - k$$

**Equation 1-4 definition of wave vector difference**
As Equation 1-3 holds for all atoms within the crystal, the amplitude of the wave scattered by atom \( i \) and \( j \) will differ by a factor \( e^{i\mathbf{h} \cdot (\mathbf{R}_i - \mathbf{R}_j)} \). Thus, scattered X-ray at each position \( \mathbf{R}_j \) should be proportional to \( e^{i\mathbf{h} \cdot \mathbf{R}_j} \). The final amplitude scattered by a primitive cell would be the total of individual ones

\[
S(\mathbf{h}) = \sum_{i=1}^{\text{cell}} e^{i\mathbf{h} \cdot \mathbf{R}_i}
\]

**Equation 1-5 Structure factor for a monoatomic crystal with a basis**

where the summation is taken over all atoms belong to the same primitive cell. \( S(\mathbf{h}) \) is usually called the structure factor, which depends on the internal 3D arrangement of atoms for a lattice grid. The overall diffraction amplitude shall be accounted for by atoms within the entire crystal, and reduced using the structure factor as

\[
F(\mathbf{h}) \propto \sum_{k=1}^{\text{crystal}} e^{i\mathbf{h} \cdot \mathbf{r}_k} = \sum_{n_1,n_2,n_3} e^{i\mathbf{h} \cdot (n_1\mathbf{a} + n_2\mathbf{b} + n_3\mathbf{c} + \mathbf{R}_i)}
\]

\[
= \sum_{n_1} e^{i\mathbf{h} \cdot \mathbf{a}} \sum_{n_2} e^{i\mathbf{h} \cdot \mathbf{b}} \sum_{n_3} e^{i\mathbf{h} \cdot \mathbf{c}} \sum_{i} e^{i\mathbf{h} \cdot \mathbf{R}_i}
\]

\[
= \mathbf{M}_F \cdot S(\mathbf{h})
\]

**Equation 1-6 Diffraction amplitude expression in terms of structure factor**

Here, \( \mathbf{r}_k \) is the atom position with respect to crystal origin, while \( \mathbf{R}_i \) to lattice origin. \( \mathbf{a}, \mathbf{b}, \mathbf{c} \) are three crystal lattice parameters. Assuming the crystal has \( N_1, N_2, N_3 \) lattice grids along \( \mathbf{a}, \mathbf{b}, \mathbf{c} \) directions, respectively. Since the intensity is the square of the amplitude, we can write down
\[ I(h) = F(h)^2 \propto M_F^2 \cdot S(h)^2 \]

**Equation 1-7** Intensity, interference coefficient and structure factor

\( M_F^2 \) is sometimes referred to as the interference coefficient, which can be reduced since those three summations are actually geometric sequences

\[
M_F^2 = \left( \sum_{n_x=0}^{N_x-1} e^{iN_x h a} \sum_{n_y=0}^{N_y-1} e^{iN_y h b} \sum_{n_z=0}^{N_z-1} e^{iN_z h c} \right)^2
\]

\[
= \left( \frac{e^{iN_x h a } - 1}{e^{i h a } - 1} \frac{e^{iN_y h b } - 1}{e^{i h b } - 1} \frac{e^{iN_z h c } - 1}{e^{i h c } - 1} \right)^2
\]

\[
= \frac{\sin^2 \left( \frac{N_x h \cdot a}{2} \right)}{\sin^2 \left( \frac{h \cdot a}{2} \right)} \cdot \frac{\sin^2 \left( \frac{N_y h \cdot b}{2} \right)}{\sin^2 \left( \frac{h \cdot b}{2} \right)} \cdot \frac{\sin^2 \left( \frac{N_z h \cdot c}{2} \right)}{\sin^2 \left( \frac{h \cdot c}{2} \right)}
\]

**Equation 1-8** Reducing the interference coefficient

The characteristics of the function is that, the maximum value \( N_x^2 N_y^2 N_z^2 \) is reached whenever the following condition is met

\[
\begin{align*}
a \cdot h &= m_1 \cdot 2\pi \\
b \cdot h &= m_2 \cdot 2\pi \\
c \cdot h &= m_3 \cdot 2\pi
\end{align*}
\]

**Equation 1-9** Condition when \( I_F \) reaches maximum
It is easy to show that Equation 1-9 is equivalent to Lauer condition and Bragg Law Equation 1-2 along three different crystal axes. Note, the interference coefficient also maximizes when $m_1, m_2, m_3$ equal zero, in which case the incident X-ray is parallel to one of the crystal planes from the view of Figure 1-1.

$M_F^2$ also has multiple local maxima for other $m_1, m_2, m_3$ values with lower peaks. However, at other $m_1, m_2, m_3$ configurations, $M_F^2$ quickly vanishes. The larger $N_1, N_2, N_3$ are, the more quickly $M_F^2$ vanishes between those maxima, the closer they are apart from each other, and the larger their peak height difference will be. This makes the actual $h$ dependence of $M_F^2(h)$ less sensitive and distinctive than $S(h)$. The profile of structure factors is derived simply after some scaling operations from that of the experiment amplitudes $F(h)$.

As for molecular crystals (or polyatomic crystals) whose primitive cell contains multiple species, the diffraction ability of each atom has to be distinguished and appropriately weighted. Therefore, we take the structure factor

$$
S(h) = \sum_{i=1}^{cell} f_i(h)e^{ih\cdot R},
$$

Equation 1-10 structure factor for molecular crystal

where $f_i(h)$ is the atomic form factor. It depends on the wave vector difference and, more importantly, the electron density distribution $\rho(r)$ of an atom, since X-rays
bombarded on to the crystal are actually interacting with and diffracted by the electrons of each atom.

\[ f_i(h) = \int \rho(r') e^{ihr'} dr' \]

**Equation 1-11 atomic form factor and electron density**

\( r' \) is the position of electron density within a specified atom. In fact, taking elementary treatment, we can combine the summation over all atoms in a cell and the integral over electrons in an atom and transform them into an expanded integral, in which we count the diffraction contribution directly from electrons but integrate across the entire unit cell space for the structure factor.

\[ S(h) = \int \rho(r) e^{ihr} dr \]

**Equation 1-12 structure factor and electron density**

It is straightforward to obtain the 3D electron density information, i.e., the molecular structure, via a Fourier Transform of Equation 1-10

\[ \rho(r) = \frac{1}{(2\pi)^3} \int_{\text{diffractions}} S(h) e^{-ihr} dh \]

**Equation 1-13 electron density and structure factor**
Experiment data collected on a screen are X-ray intensities. They are proportional to the square of the X-ray amplitudes, therefore also to the square of the structure factors. Since structure factors are complex, by squaring we are in essence multiplying them by their conjugates. This leads to a major data loss – all phase information is wiped out during this process, making it impossible to solve structure problem by immediate adoption of Equation 1-10.

1.2. Application on macromolecular structure determination

The nature of measurement inhibits the detection of diffraction phases. What we can do, however, is to try to estimate an ‘initial phase’ (herein after ‘phasing’) that is believed to be close to, or share some commonalities with, that of the target structure through various experimental or theoretical techniques. We then build a rough model based on the resultant electron density map, which is derived from the initial phases and experiment intensities (amplitudes). With atomic coordinates determined, by going back to couple the original experiment amplitudes, it is easy to calculate a more reliable density map. The newly updated density map can be used to adjust the current coordinate again. With continuous modification of the coordinates and other parameters of atoms (referred to as ‘refinement’) for the purpose of better fitting the self-produced map (or experiment data) and optimizing a ‘target function’, molecular structure is gradually improved. This procedure proceeds until the agreement between the structure and data converges to a satisfactory level, generally indicated by the R value².

As to determine a specific category of real objects, here the macromolecules, it is obvious that some experimental procedures have to be involved before data collection.
The following chart illustrates the complete workflow, from purifying studied sample to publishing the determined structure and depositing into certain commonly accessible online database (such as the Protein Data Bank).

Figure 1-2 Workflow of crystallographic macromolecular structure determination

Should results after refinement be deemed unsatisfactory by several validation criteria, the current structure is sent back to be refined again. The process repeats until the final structure is satisfactory for publishing and deposition.
1.3. Refinement

Refinement is an iterated process where the molecular parameters are constantly changed in order to optimize a target function that takes experiment data, stereochemistry and other factors into consideration. Due to the complexity of the energy landscape profile of a macromolecule with a large number of parameters, one refinement task typically consists of many rounds (pre-defined) of micro-cycles. When refinement proceeds, structure of the molecule is improved and able to better describe the diffraction data, ultimately yielding an accurate model.

![Figure 1-3 Work flow of refinement](image)

Majority of computing resources for structure determination is spent on refinement. To boost performance and reduce wall time, algorithms are usually tested and implemented on high performance computing (HPC) clusters. Thus, following chapters will be focused on theories and computational methods on refinement, and most work is
done using the HPC provided by the Research Computing Support Group (RCSG) under NSF, NIH and other grants.

Besides solving unknown structures, refinement is also useful for improving quality of molecules that is already deposited. In addition to feasibility, there is also a motivation to do so. The more precisely a molecular structure is determined, the more smoothly its function shall be researched, and the more probably its applications in industry can be realized.
Chapter 2

Elements of Refinement

2.1. Parameters

The aim of refinement is to modify model parameters to improve the overall structure. What are those parameters? Clearly, the coordinate of each atom within a molecule is the most important. Other parameters that can be refined include the temperature factor and atom occupancy. These are due to the local fluctuation of atoms, disorder in crystals and other sorts of errors or molecular motions.

1) Atomic coordinates

The positions of atoms are described via a set of three dimensional coordinates, typically in the Cartesian coordination system. The main task of refinement is to adjust the coordinates of every atom, guide the system through a wide range of conformation change and energy landscape, and finally improve the structure that can best explain the experiment data. The number of parameters per atom needed to fully express the

13
coordinates is three – \((x, y, z)\). In accordance with the PDB format, the three coordinates are listed at columns 31 to 54 in a PDB file\(^3\).

2) Temperature factors (B-factors)

Temperature factor accounts for the local mobility of an atom around its equilibrium position. Historically, temperature factor was also called the B-factor (especially in case of isotropic), and is related to Debye-Waller factor as explained below.

Assume the instantaneous position of \(j\) th atom in a unit cell \(\mathbf{r}_j\)

\[
\mathbf{r}_j = \langle \mathbf{r}_j \rangle + \Delta \mathbf{r}_j
\]

**Equation 2-1 Instantaneous atom position**

where \(\langle \mathbf{r}_j \rangle\) is the equilibrium position of atom \(j\) and \(\Delta \mathbf{r}_j\) the deviation from that position.

Experiment measurement is always a time-averaged ensemble of numerous instantaneous conformations of the molecule. Therefore, the structure factor should be an average of definition given by Equation 1-5.

\[
S(\mathbf{h}) = \left\langle \sum_{j=1}^{\text{cell}} f_j(\mathbf{h}) e^{ih\mathbf{r}_j} \right\rangle = \sum_{j=1}^{\text{cell}} f_j(\mathbf{h}) \left\langle e^{ih\mathbf{r}_j} \right\rangle
\]

**Equation 2-2 Structure factor time averaged**

Insert Equation 2-1 into the last factor of Equation 2-2,
\[ \langle e^{i h r_j} \rangle = \left\langle e^{i h \left( \langle r_j \rangle + \Delta r_j \right)} \right\rangle = e^{i h \langle r_j \rangle} \cdot \left\langle e^{i h \Delta r_j} \right\rangle \]
\[ = e^{i h \langle r_j \rangle} \cdot \left( 1 + i h \cdot \Delta r_j + \frac{1}{2!} \left( i h \cdot \Delta r_j \right)^2 + \cdots \right) \]
\[ = e^{i h \langle r_j \rangle} \cdot \left( 1 + i h \cdot \langle \Delta r_j \rangle - \frac{1}{2} \left( h \cdot \Delta r_j \right)^2 + \cdots \right) \]
\[ = e^{i h \langle r_j \rangle} \cdot \left( 1 - \frac{1}{2} \left( h \cdot \Delta r_j \right)^2 + \cdots \right) \]
\[ \approx e^{i h \langle r_j \rangle} \cdot e^{-\frac{1}{2} \left( h \cdot \Delta r_j \right)^2} \]

**Equation 2-3 Reducing \langle e^{i h r_j} \rangle**

Here we used \( \langle \Delta r_j \rangle = 0 \) due to \( \langle r_j \rangle = \langle \langle r_j \rangle + \Delta r_j \rangle = \langle r_j \rangle + \langle \Delta r_j \rangle \). The last step is a Gaussian Approximation.

We can write the exponential factor \( -\frac{1}{2} \left( h \cdot \Delta r_j \right)^2 \) in the form of matrix \( U_j \) and row vector \( h \).

\[-\frac{1}{2} \left( h \cdot \Delta r_j \right)^2 \]
\[ = -\frac{1}{2} \left( h \cdot \Delta r_j \cdot (\Delta r_j \cdot h) \right) \]
\[ = -\frac{1}{2} \left( h (\Delta r_j \times \Delta r_j) h^T \right) \]
\[ = -\frac{1}{2} h \left( \Delta r_j \times \Delta r_j \right) h^T \]
\[ = -\frac{1}{2} h U_j h^T \]

**Equation 2-4 U matrix**
with

\[
U_j \equiv \langle \Delta r_j \times \Delta r_j \rangle = \begin{pmatrix}
\langle \Delta x_j \Delta x_j \rangle & \langle \Delta x_j \Delta y_j \rangle & \langle \Delta x_j \Delta z_j \rangle \\
\langle \Delta y_j \Delta x_j \rangle & \langle \Delta y_j \Delta y_j \rangle & \langle \Delta y_j \Delta z_j \rangle \\
\langle \Delta z_j \Delta x_j \rangle & \langle \Delta z_j \Delta y_j \rangle & \langle \Delta z_j \Delta z_j \rangle
\end{pmatrix}
\]

**Equation 2-5 U matrix definition**

\(\Delta x_j, \Delta y_j, \Delta z_j\) are the three components of \(\Delta r_j\). If the vibration is approximated to be harmonic and isotropic, then

\[
U_j = \begin{pmatrix}
\frac{\Delta r_j^2}{3} \\
\frac{\Delta r_j^2}{3} \\
\frac{\Delta r_j^2}{3}
\end{pmatrix}
\]

\[-\frac{1}{2} \left\langle \left( \mathbf{h} \cdot \Delta r_j \right)^2 \right\rangle = -\frac{1}{6} \mathbf{h}^2 \cdot \Delta r_j^2 \equiv -\frac{1}{16\pi^2} \mathbf{h}^2 \cdot B_j\]

**Equation 2-6 Introducing B factor**

Here shows the isotropic B-factor definition

\[B_j \equiv \frac{8\pi^2}{3} \Delta r_j^2\]

**Equation 2-7 Definition of isotropic B factor**
a quantity that is directly proportional to the fluctuation of the atom. $\Delta r_j^2$ is the mean square of total displacement with respect to the equilibrium, not that of a specific direction. There exist other expressions where the diagonal term of $U_j$ is denoted by $u_j^2$, which lead to a slightly different form to Equation 2-7

$$B_j = 8\pi^2 u_j^2$$

**Equation 2-8 Alternative form of definition of isotropic B factor**

The coefficient introduced in Equation 2-6 is for the purpose of simplifying the temperature term – also called the Debye-Waller factor – of

$$\left\langle e^{i\mathbf{h} \cdot \Delta \mathbf{r}} \right\rangle = e^{-\frac{1}{16\pi^2 \hbar^2} B_j} = e^{-\frac{B_j (\sin \theta)^2}{\lambda}}$$

**Equation 2-9 Debye-Waller factor in terms of B factor, incident angle and wavelength**

with the aid from the Lauer Condition

$$h = 2 \cdot \sin \theta \cdot \frac{2\pi}{\lambda}$$

The isotropic B-factor value is recorded at columns 61-66 in a PDB file.

In refinement, these B-factors for each atom are treated like independent parameters, though restraints on them are imposed from time to time.
When anisotropy is considered, a total of 6 independent parameters, which correspond to the 3 diagonal terms and 3 non-diagonal terms of matrix $U_j$ are used for each atom. In practice, a separate line starting with ‘ANISOU’ under each ‘ATOM’ entry is deployed. Columns 29-70 in each ANISOU entry are reserved for six anisotropic terms in the order of $B_{11}, B_{22}, B_{33}, B_{13}, B_{13}, B_{23}$, each occupies seven columns.

Use of anisotropic U factor will lead to a six-fold parameter increase and is subject to the completeness and quality of experimental diffraction data.
3) Occupancy

Known as the static crystal disorder, all molecular copies from different regions of the crystal are not exactly in identical conformation, due to the local flexibility of the protein side chains or even main chains. Those alternative conformations can be stable and distinguishable by the electron density map (usually with high resolution data), exhibiting a significant population compare with the main conformation. In order to account for this, a parameter called atom occupancy is introduced to describe the possibility an atom is in a certain conformation state (position), or, from the statistical point of view, the ratio between the number of molecule copies where the atom is in the particular conformation and the total number of molecule copies within a crystal.

In case multiple conformations are detected and recorded, each occupancy is less than 1 and subject to refinement. One constraint is that the occupancy must add up to 1 for multiple conformations, except for those non-protein atoms, such as metal atom and
other ligands, whose occupancy is allowed to be fractional owing to imperfection in co-crystallization under certain experiment conditions. Occupancy of each atom is recorded at columns 55-60 in a PDB file.

**Bulk solvent correction and** $k_{sol}, B_{sol}$ **optimization**

There are two additional parameters, $K_{sol}, B_{sol}$, that are related to the bulk solvent. Bulk solvent is the region of the unit cell other than the protein molecule. The mask that separates molecule and the bulk solvent is called the solvent mask. Since solvent makes contribution to X-ray diffraction as well, it is important to account for it for a more accurate calculated structure factor to fit the experiment data.

One model to describe the bulk solvent is based on Babinet’s Principle, stating that a 180 (half wavelength) shift exists between the Fourier Transform of the solvent mask and the protein mask. The implication that the electron density of the solvent is proportional to that of the protein with opposite phases does not hold at resolution higher than 15 Å$^4$, and therefore not recommended.

The other model is the flat density model$^5$, in which no assumption of structure factor relationship between the molecule and bulk solvent is made. This model needs to determine a relatively accurate mask by placing the molecule on a grid and distinguishing the grid points falling in and out of the molecular region, followed by further refining the boundary with two parameters SOLRAD and SHRINK, which are normally set as constant. $k_{sol}$ represents the flat density of the solvent and another parameter $B_{sol}$ is introduced to smooth sharp edge effects arising from the Fourier transform of grids between solvent and solvent-excluded regions. The total structure is
\[ S_{\text{sol}}(h) = S(h)_{\text{mol}} + k_{\text{sol}} e^{-\frac{h^2}{16\pi^2\beta_{\text{sol}}}} S(h)_{\text{sol}} \]

or

\[ S_{\text{sol}}(h) = S(h)_{\text{mol}} + k_{\text{sol}} e^{-\frac{\sin^2\theta}{\lambda^2}} S(h)_{\text{sol}} \]

**Equation 2-10** Total calculated structure factor with flat density bulk solvent

These two parameters are usually optimized immediately after the commencement of a new refinement macro-cycle, before other parameters begin refined. It is done by minimizing R value in lowest resolution shell without significantly increasing that in the high resolution, keeping parameters fixed one after the other in an iterative style in 2D space.

### 2.2. Target function

Refinement is a process that by optimizing a target function, parameters of a model are continuously changed to better explain the experiment data, at the same time without generating irrational results judged by stereochemistry and *a priori* knowledge.

In a physics style, we treat the refinement target as a total potential energy term. The total energy should therefore involve experiment-related energy, stereochemistry energy and *a priori* knowledge based restraints potential. Our goal is to minimize an overall function as a linear combination of them all, across the function’s complicated energy landscape.
\[ E_{\text{target}} = f(E_{\text{experiment}}, E_{\text{stereo}}, E_{a\ priori\ knowledge}) \]

**Equation 2-11  Refinement target, i.e., Total potential energy**

1) **Experimental energy**

The ultimate goal of refinement is to predict an accurate model to best interpret the experiment data in hand. The extent of agreement reached between model and data can be quantified as an energy term, and depends on the way experiment data are incorporated.

**Fitting electron density – real space refinement**

The most original idea of refinement is to calculate experiment electron density map using experiment amplitude and model phase determined in the previous step, followed by fitting a current model density map to the experiment density, generating a new model and a new experiment density map, and repeat\(^6\).

This way, the experiment-related energy term is defined as

\[ E_{\text{experiment}} = \sum_{\text{grid points}} (\rho_{\text{best}} - k \rho_{\text{calc}})^2 \]

**Equation 2-12  experiment potential with electron density map**

where \( \rho_{\text{best}} \) is the best available map (an experiment map or \( 2mF_c - DF_c \) map), \( \rho_{\text{calc}} \) is model calculated map. Though this method was widely used before 1941\(^7\), it is seldom used in modern refinement algorithms due to the dependence of the calculated density
map on the quality of the diffraction data. Density map corresponding to identical structure but derived from data with different resolutions can differ quite much. This undesired feature makes real space refinement unreliable and gradually abandoned, especially after the emergence of refinement targets in reciprocal space.

**Least Squares – reciprocal space refinement**

Hughes proposed a refinement target function in reciprocal space called Least Squares energy.

$$E_{\text{experiment}} = \sum_{\mathbf{h}} w(\mathbf{h}) (F_{\text{obs}}(\mathbf{h}) - kF_{\text{cal}}(\mathbf{h}))^2$$

**Equation 2-13 Least Squares target function**

$k$ is a scaling coefficient, $F_{\text{obs}}$ and $F_{\text{cal}}$ are observed and calculated diffraction amplitude, respectively. $w$ is a weight to account for the importance of each diffraction’s contribution to the total target function. The summation is taken over all diffractions. The least squares energy essentially describes the agreement between the model calculated structure factors and their counterpart from diffraction data. It was widely used in small molecule refinement. The limitation of least squares target is that it is unable to smartly adjust the weight of contribution according to different quality of measurement of diffraction entries. Moreover, in cases where there are missing atoms or chemical groups in a model, least squares target fails to correctly interpret this situation and will consequently guides the refinement towards unfavorable directions.

**Maximum Likelihood**
There is another way of expressing the agreement between data and model, with ability to take into consideration the incompleteness and error of the model\textsuperscript{8}, a desired feature for refining macromolecules.

The target function is defined as the likelihood of observing a data set (in this case the experiment data set in hand) given a known model. Our goal is to maximize this likelihood by modifying the model parameters, or minimize its opposite number (treated as a form of potential energy for a consistent style with other target function terms).

Assuming the conditional probability of an observation amplitude given the model is $P(F_{\text{obs}}(h) | F_{\text{cal}}(h))$ and different diffraction entries are independent of each other. In order to observe a particular pattern of diffraction, all diffractions this pattern is composed of should be observed simultaneously. Therefore, the associated likelihood $L$ is the multiplication of the single diffraction’s conditional probability.

$$L = \prod_h P(F_{\text{obs}}(h) | F_{\text{cal}}(h))$$

\textbf{Equation 2-14  Likelihood of observing a diffraction pattern given a model}

It is a common practice in statistics to compute the minus logarithm of the likelihood rather than itself. The log-likelihood form also results in computational convenience by having transformed $\prod$ into $\sum$. ‘1’ is subsequently multiplied to achieve the purpose of maximizing Equation 2-14 by minimizing the following energy function:
\[ E_{\text{experiment}} = -\ln L = -\sum_{h} \ln P(F_{\text{obs}}(h) | F_{\text{cal}}(h)) \]

**Equation 2-15 Maximum Likelihood target function definition**

Consider model error \( \sigma_{\Delta} \), measurement error \( P_{\text{meas}}(F_{\text{obs}} | F) \), *a priori* phase distribution \( P_{\text{phase}} \), and re-write \( F(h) \) with \( F(h) \cdot e^{i\phi(h)} \) (same for \( F_{\text{obs}}(h) \) and \( F_{\text{cal}}(h) \)),

\[
P(F_{\text{obs}}(h) | F_{\text{cal}}(h)) = \frac{1}{\pi \sigma_{\Delta}^2} \int \int F(h) \cdot P_{\text{meas}}(F_{\text{obs}}(h) | F(h)) \cdot P_{\text{phase}}(\phi(h)) \cdot e^{-\frac{[F(h) - D_{\text{obs}}(h)]^2}{\sigma_{\Delta}^2}} d\phi(h) dF(h)
\]

**Equation 2-16 Detailed expression of likelihood of one observation given a model**

Inserting Equation 2-16 to Equation 2-15 yields the maximum likelihood target energy. According to the data type, there are totally three variants of Equation 2-16.

- **MLF** target function, for data expressed by diffraction amplitudes

- **MLI** target function, for data expressed by diffraction intensities

- **MLHL** target function, for data with experimental phase information

For macromolecular crystallography, Maximum Likelihood target function is superior over Least Squares.

**2) Stereochemistry energy**
Structure geometry, e.g., bond length or bond angle of a chemical group, often follows a standard value\textsuperscript{11}. Explicitly introducing stereochemistry energy to the total target function is indispensable for refinement with low resolution data, as those data are not informative enough for maintaining accurate geometry during the refinement process.

Despite the rigor of many chemical bonds, angles, dihedrals, etc., stereochemistry energy is typically added in the form of restraint instead of constraints. They are different in definition, computational treatment, and the way data-to-parameter ratio is influenced.

**Restraints and Constraints**

Restraints and constraints both imply a trend of which a parameter always converges to an ‘ideal value’. A constraint imposes a ‘hard’ equality that their difference must strictly vanish at all times. This introduces correlation between “independent” refinable parameters and equivalently decreases the number of them. Suppose the number of experiment data, refinable parameters and constraints $N_D$, $N_p$ and $N_C$, respectively. The data-to-parameter ratio $\gamma$ is increased by considering constraints.

$$\gamma_{\text{original}} = \frac{N_D}{N_p}$$

*Equation 2-17 Data-to-parameter ratio*

$$\gamma_C = \frac{N_D}{N_p - N_C}$$

*Equation 2-18 Data-to-parameter-ratio with constraints*
A restraint, on the other hand, is ‘soft’ and usually takes the expression of a harmonic energy in which the ideal value is the equilibrium value, and the parameter is allowed to deviate from equilibrium, incurring a restoring force that drags it back. The degree of softness depends on the potential coefficient and will decrease with increasing coefficient. These restraints provide additional information between independent refinement parameters and equivalently serve as extra data.

\[
\gamma_R = \frac{N_D + N_R}{N_P}
\]

**Equation 2-19**  Data-to-parameter ratio with restraints

**Restraint for bond lengths**

Atom pair interacting with each other via a chemical bond should be close to the standard value of that chemical bond type. Total stereochemistry bond energy should be the summation of all bond length restraints. \( w_{\text{bond}} \) is associated weight assigned to each term. A similar weight is present for other restraint classes.

\[
E_{\text{bond}} = \sum_{\text{bonds}} w_{\text{bond}} (d_{\text{model}} - d_{\text{ideal}})^2
\]

**Equation 2-20**  Bond length restraint energy

**Restraint for bond angles**

Similar to Equation 2-20, bond angle restraint is defined as
\[ E_{\text{angle}} = \sum_{\text{angles}} w_{\text{angle}} (\omega_{\text{model}} - \omega_{\text{ideal}})^2 \]

**Equation 2-21** Bond angle restraint energy

**Restraint for dihedral angles**

Dihedral angle is the angle between two planes defined by four atoms. Suppose that atoms are sequentially labeled as \( A, B, C, D \). Dihedral angle \( \theta \) is the angle between plane \( ABC \) and plane \( BCD \).

![Dihedral angle of four sequential atoms](image)

**Figure 2-3** Dihedral angle of four sequential atoms

The dihedral angle energy is
\[ E_{\text{dihedral}} = \sum_{\text{dihedrals}} w_{\text{dihedral}} (\theta_{\text{model}} - \theta_{\text{ideal}})^2 \]

**Equation 2-22  Dihedral angle energy restraint**

**Restraint for planarity**

Planarity can be maintained by defining related improper angles (a dihedral angle with torsion axis not a chemical bond) and fix them to 0 or 180 degree. This could be inefficient when many atoms are involved. Another planarity restraint energy is defined to penalize out of plane conformation atoms\(^{12}\).

\[ E_{\text{planarity}} = \sum_{\text{groups}} w_{\text{plane}} \sum_i g_i^2 \]

**Equation 2-23  Planarity restraint energy**

\( g_i \) is the orthogonal distance of \( i \)th atom within a group from the plane defined by all atoms of that group. Double summations are taken over all atoms within a group and all planar groups.

**Restraint for chirality**

An asymmetric carbon atom leads to chirality in a molecule which is non-superposable with its mirror image. A quantity called chiral volume is defined to describe the chirality of an atom (e.g. \( C_\alpha \)).
\[ V_{\text{model}} = (\mathbf{r}_N - \mathbf{r}_{\text{C}_\alpha}) \cdot \left[ (\mathbf{r}_C - \mathbf{r}_{\text{C}_\alpha}) \times (\mathbf{r}_{\text{C}_p} - \mathbf{r}_{\text{C}_\alpha}) \right] \]

**Equation 2-24  Chiral Volume of a C alpha atom**

with the chirality energy restraint

\[ E_{\text{chiral}} = \sum_{\text{chiral}} w_{\text{chiral}} (V_{\text{model}} - V_{\text{ideal}})^2 \]

**Equation 2-25  Chirality energy restraint**

**Non-bonded restraint energy**

Van der Waals interaction and electrostatics interaction are merged into a single non-bonded energy term

\[ E_{\text{non-bonded}} = \sum_{\text{nonbonded-pair}} \left( \frac{A}{r_{ij}^{12}} - \frac{B}{r_{ij}^6} + \frac{Cq_i q_j}{r_{ij}} \right) \]

**Equation 2-26  non-bonded restraint energy**

The Van der Waals energy accounts for both a repulsive term \( \frac{A}{r_{ij}^{12}} \) and an attractive term \( \frac{B}{r_{ij}^6} \).

3) *a priori* knowledge based restraints
A priori knowledge can be made use of to restrain appropriate properties of a model structure, especially for refinements at low resolution, for the purpose of ensure important features (e.g. the secondary structure) that are otherwise difficult to reveal from the experiment data alone.

**Reference model.** By analyzing a high resolution homologous model, select features can be analogous and used for refining the target model with low resolution data. An example is the DEN refinement\(^1\).

**Secondary structure restraints.** H-bond restraints are imposed to maintain alpha helices, beta sheets and DNA/RNA base pairs.

**Ramachandran restraints.** Steer outliers in a Ramachandran plot\(^1\) graph to a favored region to fix incorrect dihedral angle pairs \((\phi,\psi)\).

**NCS restraints.** Non-crystallographic symmetric copies of a molecule should ideally have identical structures. Geometry difference between each copy can be restrained.

### 2.3. Optimization

Refinement target is an energy potential with complicated landscape, numerous local minima and maxima across the conformation space. By optimizing this target, we are essentially trying to explore the global minimum of this unknown landscape.

**Gradient minimization**

Gradient minimization is a method that by calculating a set of ‘effective force’ from each atom’s local gradient of potential environment, the conformation is driven
under this force and thus moves along the downhill of the landscape until arriving at the closest local minimum. Force on the $i$th atom is

$$F_i = -\nabla_i E(r_1, r_2, \ldots, r_i, \ldots)$$

**Equation 2-27 Force derived from the gradient**

This minimization strategy reveals the nearest path towards a local minimum, however, it is unable to overcome an energy barrier, has no access to the global minimum and therefore is not used for minimizing a sophisticated target potential.

![Figure 2-4 Optimization of a target function with sophisticated energy landscape using gradient minimization](image)

**Simulated Annealing**
Simulated annealing optimization is, in its name, a simulation to the annealing process. The latter is known as first melt a solid to liquid phase, then apply a long process of gradually cooling so that all particles are arranged in the lowest energy state.

In computation implementation, simulated annealing is usually controlled by two temperature parameters – the starting temperature and the cooling rate (assuming that we are decreasing the temperature with a constant rate). Generally an initial temperature high enough would assign particles with large velocities that are necessary to climb over high barriers, but may also lead to system ‘blowout’. On the other hand, a slow cooling process would make the conformation search for global minimum finer, but may dramatically increase CPU time.

![Simulated Annealing](image)

**Figure 2-5** Optimization of a target function with sophisticated energy landscape using simulated annealing
Grid search in conformation space

As expected, radius of convergence for grid search should be the largest among all the three techniques, as the high energy barriers make no difference to any other point of the landscape, because of the fact that the conformation change is realized directly from switching parameters on grid points across a pre-defined sample range. The obvious limitation is that it becomes computationally intractable and impractical for systems with a large number of parameters (N-dimensional search), or when one or more of those parameters requires a very broad sample range.

Figure 2-6 Optimization of a target function with sophisticated energy landscape using grid search

As an application with less than three parameters, bulk solvent correction mask model uses the grid search technique to determine a parameter pair \( k_{\text{sol}} \) and \( B_{\text{sol}} \).
Similarly, the Deformable Complex Network approach (proposed later) uses a 3D grid search to determine a DCN parameter combination that delivers the lowest R free value\(^2\).

### 2.4. Progress indicators and validation

1) The R value and overfitting problem

**The R value**

In crystallography, \( R \) factor is the common quantity defined to evaluate progress of refinement and quality of a structure\(^1\).

\[
R = \frac{\sum_h \left| F_{\text{obs}}(h) - kF_{\text{cal}}(h) \right|}{\sum_h F_{\text{obs}}(h)}
\]

**Equation 2-28 Definition of \( R \) value**

Here \( k \) is a scaling coefficient. Smaller \( R \) factor indicates a better agreement between observed and calculated amplitude profiles, thus a better structure in the sense of model/data consistency.

**Overfitting problem and \( R_{\text{work}}, R_{\text{free}} \)**

For macromolecular refinements, number of refinable parameters may exceed that of the diffraction entries. When this happens, a structure model can be ‘overfitted’ during an intensive optimization of the parameters.
The root cause of overfitting is the undetermined nature of the equation group that arises when independent parameters are more than applicable conditions. Thus, the $R$ value can in theory be made arbitrarily small for a refinement with poor observation-to-parameter ratio.

Cosmetic decrease in $R$ value does not necessarily mean an improvement in structure. Artificiality in $R$ value invalidates its objectiveness and makes it unsuitable as a refinement indicator. To address this issue, the idea of cross validation from statistics is introduced\(^2\) for assessment of structure quality.

In practice of cross validation, the entire diffraction data are divided into two sets. One is the working set, which takes up 90%-95% of the data and actually serves as the experiment data used in refinement. The other is the free set generated by a random selection of 5%-10% diffractions from the data pool. This set does not participate in the refinement process and is usually recorded by a ‘free flag’ for identification in an experiment data file. Therefore, summations over entire data set in all reciprocal space target functions previously discussed are now actually carried out only over the working subset data.

Accordingly, two variants of Equation 2-28, labeled as $R_{\text{work}}$ and $R_{\text{free}}$ are defined by simply doing the summation over the working and free data sets.
\[
R_{\text{work}} = \sum_{\mathbf{h} \in \text{working set}} \frac{|F_{\text{obs}}(\mathbf{h}) - kF_{\text{calc}}(\mathbf{h})|}{\sum_{\mathbf{h} \in \text{working set}} F_{\text{obs}}(\mathbf{h})}
\]

\[
R_{\text{free}} = \sum_{\mathbf{h} \in \text{free set}} \frac{|F_{\text{obs}}(\mathbf{h}) - kF_{\text{calc}}(\mathbf{h})|}{\sum_{\mathbf{h} \in \text{free set}} F_{\text{obs}}(\mathbf{h})}
\]

**Equation 2-29** Definition of the work and free R values

High correlation is expected and observed between the Least Squares target function in Equation 2-13 and \( R_{\text{work}} \) in Equation 2-29. As refinement proceeds, Least Square target function is continuously minimized while \( R_{\text{work}} \) drops down at the same time. Similar correlation exists between the Maximum Likelihood target and the \( R_{\text{work}} \) as well. \( R_{\text{free}} \), on the other hand, is not biased by the model or the refinement procedure, and used as a primary indicator of structure improvement in modern refinements.

2) **Root Mean Square Deviation (RMSD)**

RMSD between two related structures is defined to quantitatively assess the overall difference between two coordinates set, or the ‘deviation’ of one coordinate set from the other, usually after a global alignment.

Consider two molecule structures with identical number of atoms. The RMSD\(^{16,17}\) for equivalent atoms is
\[
RMSD(1,2) = \sqrt{\frac{\sum_{j=1}^{n}(r_{ij} - r_{2j})^2}{n}}
\]

**Equation 2-30 Definition of RMSD**

In cases when \((r_{i1}, r_{i2}, ..., r_{in})\) and \((r_{21}, r_{22}, ..., r_{2n})\) denote structures of only main chain atoms instead of all atoms, the quantity is referred to as main chain (or backbone) RMSD.

3) **Global Distance Test (GDT) score**

The disadvantage of RMSD roots from the equality of weights for all atom pairs. Certain regions of a molecule, for instance, the loop region, may be extremely flexible with various allowed and favorable conformations. Deviation between different conformations can be significant, making considerable contribution to the RMSD, while the structure may in essence be well acceptable compared to the impression a hefty RMSD gives.

To address this issue, a GDT score is calculated as the proportion \(P\) of equivalent alpha carbons in two structures with distances smaller than a defined cutoff (1Å, 2Å, 4Å, 8Å).

\[
GDT_{TS} = \frac{P(<1\text{Å}) + P(<2\text{Å}) + P(<4\text{Å}) + P(<8\text{Å})}{4}
\]

**Equation 2-31 Definition GDT Total_Score**
Distance cutoff is usually directly labeled with the score name. For example, the GDT(<1Å) score used in the Deformable Complex Network approach below is a GDT score with a cutoff of 1Å.

\[ GDT(<1\text{Å}) = P(<1\text{Å}) \]

Generally, GDT score increases with larger cutoff radius.

4) TMscore

Zhang et.al. proposed a template/model score (TMscore)\textsuperscript{19} to eliminate the system size dependence of GDT scores for random structure pairs. GDT scores of 3656 protein pairs with sequence identity less than 30% were computed and a power law between GDT score and protein length was observed.

TMscore, another quantity measuring structure similarity between two proteins, is a variant of the Levitt-Gerstein score\textsuperscript{20},

\begin{equation}
TMscore = \text{Max} \left[ \frac{1}{L_N} \sum_{i=1}^{L_T} \frac{1}{1 + \left( \frac{d_i}{d_0} \right)^2} \right]
\end{equation}

Equation 2-32  Definition of TMscore

with \( L_N \) and \( L_T \) the length of native structure and aligned residues to the template, respectively. \( d_i \) is the distance of \( i \) th pair of aligned residues. \( d_0 \) is a normalized term,
which takes the following form rather than a constant as in other approaches\textsuperscript{20-22} in order to eliminate protein size dependence.

\[ d_0 = 1.24 \sqrt[3]{L_N} - 15 - 1.8 \]

**Equation 2-33** \( d_0 \) as a function of \( L_N \) in TMscore

![Figure 2-7](image.png)  
**Figure 2-7** Relationship between GDT (and MaxSub) score for random structure pairs and length of proteins\textsuperscript{19}
5) Ramachandran Statistics

Ramachandran Statistics calculates the percentage of dihedral angle pairs \((\phi, \psi)\) falling in the favorable regions according to Ramachandran plot\(^{14}\). \(\phi\) is the dihedral angle spanned by two planes formed by atom \(C_{i-1} - N - C^\alpha\) and \(N - C^\alpha - C_i\). Similarly, for \(\psi\) it is spanned by \(N_i - C^\alpha - C_i\) and \(C^\alpha - C_i - N_{i+1}\). Pairs for all residues are plotted on a two-dimensional map and pairs falling within several favored regions (determined by the characteristics of alpha helices and beta sheets) of the map are counted. The percentage is defined as the Ramachandran Statistics and is used as an indicator of secondary structure quality.
As shown in Figure 2-9, black dots represent $(\phi, \psi)$ pairs, red circles outline the favored region, while orange lines sketch the allowed region. Ramachandran Statistics is defined as

$$\text{Ramachandran Statistics} = \frac{N_{\text{pairs in favored regions}}}{N_{\text{total pairs}}}$$

**Equation 2-34 Calculation of Ramachandran Statistics**

The higher the Ramachandran Statistics is, the better the secondary structure of a model has.
6) Electron Density Map

Electron density map is calculated based on Equation 1-13

$$\rho(\mathbf{r}) = \frac{1}{(2\pi)^3} \int_{\text{diffractions}} S(\mathbf{h}) \cdot e^{-i\mathbf{h} \cdot \mathbf{r}} \, d\mathbf{h} = \frac{1}{(2\pi)^3} \int_{\text{diffractions}} S(\mathbf{h}) e^{i\theta} \cdot e^{-i\mathbf{h} \cdot \mathbf{r}} \, d\mathbf{h}$$

Several density maps can be obtained depend on what data are used for substituting

- **F_C** map – Electron density calculated with model structure factor and phase, which is a map solely related to the current model

- **F_O** map – Electron density calculated with experiment amplitude and model phase, which shows the observed electron density

- **F_O-F_C** – Difference between the observed and model density map, which tends to be zero when model is correct, moderate non-zero when incorrect atom type is modeled, large positive when an atom is missing from the model and large negative when model contains an atom but supposedly not.

- **2F_O-F_C** – Summation of observed and difference density map, widely used for model building and structure validation.
3.1. Motivation and introduction

It is often a challenge to atomically determine the structure of large macromolecular assemblies due to their weak diffraction of X-ray. Effective number of diffractions available for structure determination looks small, especially when we search for the optimum conformation in the conventional coordinate space. Moreover, interpretation of the experiment data to predict the structure is often hindered by the limited agreement between them. There has always been a need for low resolution structures to be determined at higher accuracies in order to allow valid and intensive studies on their functions and properties.

This work combines the torsion-angle protocol with the deformable complex network (referred to as DCN) approach, to further derive and make use of useful information from a pre-determined homologous or comparative protein model. It can be
shown that, by merging the information independently fetched from the deformable angular network (DAN) and the DEN\textsuperscript{13,23} thus generating a DCN model, there is still room for additional improvement to macromolecular refinements over the existing DEN method\textsuperscript{13}, by a boost from 13% to 264% as assessed by the free R value\textsuperscript{2}.

Firstly, in order to objectively evaluate the quality of the refined structure, we performed a full refinement against an experiment data set (without experimental phase information) that already has a high resolution (1.8Å) structure deposited into the Protein Data Bank. The data set was then truncated to three different limits to serve as three lower resolution data sets. We used those data sets for subsequent refinement and compared the results with the existing high-resolution structure (‘true structure’). Improvements are observed across multiple criteria, from the R\textsubscript{free} value, to the all-atom Root Mean Square Deviation(RMSD), the GDT (<1Å) score\textsuperscript{18} and the TM score\textsuperscript{19}. Further improvement is expected with the availability of the non-crystallographic symmetry (NCS) information, as well as phase information from experimental methods such as heavy atom isomorphous replacement\textsuperscript{24}. Secondly, to ensure generality, we also randomly selected sixteen low resolution structures from the Protein Data Bank and performed re-refinements with published experiment data. Consistent improvements by DCN have been seen over conventional refinement and standalone DEN refinement, as indicated by the R\textsubscript{free} value, the Ramachandran Statistics\textsuperscript{25} as well as the calculated phase combined electron density map.
3.2. Method

3.2.1. Summary

Starting from a given protein’s sequence (target sequence) information, we first individually performed a FASTA search to each chain of the molecule. Templates that shared higher sequence identity with the target sequence and possessed higher resolution would be preferable. Five homology structure candidates were subsequently constructed to a chosen template and one with the lowest DOPE score was picked and served as a reference model for a certain chain of the molecule.

After reference structures for all chains (if applicable) of a molecule had been built, these structures were merged together and served as the only reference model for the whole molecule. DCN excludes all sorts of inter-chain atoms’ interactions when deformable angular and elastic network models are defined. As a result, rather than in the target molecule, different chains in the reference ‘molecule’ are independent and can take whatever relative positions and orientations. The DCN model and corresponding restraints were automatically generated according to a pre-set criteria for angular network triplets and elastic network pairs. These restraints contribute to the $E_{\text{DCN}}$ term in the total energy function (target function, see below). Simulated annealing$^{26}$ was used as the refinement protocol, with a starting temperature of 3,000$^\circ$K and a cooling rate of 50$^\circ$K per step. The torsion-angle dynamics$^{27}$ was performed as the Molecular Dynamics. Refinement with each parameter group was repeated ten times with different random seeds for initial velocities assignments and DCN restraint selections.
3.2.2. Target function

The target function takes the following form

\[ E_{\text{target}} = E_{\text{stereo}} + w_a \cdot E_{\text{experiment}} + w_{\text{DCN}} E_{\text{DCN}} \]

Equation 3-1 Refinement Target of DCN

\( E_{\text{stereo}} \) is the usual stereochemistry energy which regularizes bond lengths, bond angles and others to pre-defined, well-accepted standard values. \( E_{\text{experiment}} \) is the experimental term that incorporates the reflection data, and its weight \( w_a \) is determined automatically and adjusted frequently to ensure the force derived from the experiment term and that of other terms are approximately of the same order of magnitude in this equation. Typically the amplitude-based maximum likelihood function (MLF) would be used instead of the conventional crystallographic residual (least square). In case phase information is obtained through experimental techniques, a refinement is executed with the MLHL target function where experiment phase contributes in the form of Hendrickson Lattman coefficients\(^{28}\). \( E_{\text{DCN}} \) is the potential energy that roots from the deviation of selected atom pairs and triplets in the target molecule from their corresponding equilibrium values; these values are derived from both the reference models as well as the target’s current structure itself.
3.2.3. DAN and DCN approach

3.2.3.1. Introduction

The deformable angular network (DAN) model is a model composed of a series of angles, each spanned by two bonds within an atom triplet which is to be found by referring to the same chain ID, residue number and atom name in both the reference and target structures. Generally, when a reference structure is obtained via homology modeling\textsuperscript{29,30}, all of the atoms that have equivalent atoms in the target structure are picked as candidate elements of DAN triplets. Further filtering of the triplets arises from the requirement that:

(1) vertex atom of the triplets has interactions with both tail atoms with a user-defined search radius or one equal to the upper distance cutoff of DEN restraints.

(2) vertex atom and each tail atom should be no more that ten residues apart.

After the first two rounds of preliminary selection, the remaining eligible triplets are subject to a third one based on the vertex angles that spanned by the two “line” connecting the vertex and each tail. We choose those angles that have a value between 60 and 120 in degree. The final angular restraints for later refinement purpose are randomly selected from the triplets pool. This random behavior is controled by a seed pre-defined before refinement. The total number of restraint entries is determined by the product of the number of atoms that participate in the generation of the DAN model and a multiplicity factor (typically 1). Since interactions between different chains are excluded, DAN is usually made chain by chain and then merged into a single reference model. The selection of atoms for DAN generation can also be constrained to any groups of atoms or
their combinations within the same chain. Deformable complex network (DCN) is established if both DAN and deformable elastic network (DEN) are present. In this case, two models are built independent, even from different results of homology modeling. However, they work together to lead the direction of conformation search throughout the refinement process. The restraints of DAN and DEN contribute in a reciprocal way to improve the refinement and final structure. These restraints are considered as unified DCN restraints and added to the total refinement target function.

### 3.2.3.2. DAN/DCN modes

Due to the way the DAN triplets pool is initially constructed, there are two modes for DAN (therefore for DCN as well). The first is the directional mode. After all atoms have been serialized (i.e., the atom serial number is assigned to each atom in PDB), the vertex atom seeks interaction only with atoms of higher atom serial numbers. This mode results in a fact that in each triplet, vertex atom has the lowest atom serial number (Figure 3-1). Hence, the directional mode is only capable of covering a fraction of selections from the entire candidates pool. In the other hand, this mode also comes with a feature that explicitly eliminates an unfavorable situation, in which each of the three atoms within a single triplet is picked as the vertex one after another, and three resultant restraints actually correspond to three interior angles of the same triangle. This kind of restraints is considered too strong, especially when the reference model has a low quality and is not reliable. The second is called the arbitrary mode(Figure 3-2). In this case, direction in which the vertex points to a tail atom is not restricted at all. This allows a hundred percent coverage of all angles for possible selection from the triplets pool. The definition of arbitrary mode ensures that no useful information from reference model is
discarded from the beginning, therefore, chances are increased that a list of better atom triplets will be established and serves as the DAN restraints. However, because vertex atom can “point back”, arbitrary mode typically involves fewer different atoms than directional mode. In a DAN reference file, for both modes, each restraint entry is listed in the order of ‘vertex atom – first tail atom – second tail atom – angle value’. The atom serial number of the first tail atom is constantly lower than that of the second to prevent restraints duplications. Typically the arbitrary mode is chosen as the default mode. However, it is wise to sometimes perform the DCN refinement in directional mode to achieve a possibly lower $R_{\text{free}}$.

![Figure 3-1 Directional mode (D-mode) of DAN/DCN](image)

**Figure 3-1 Directional mode (D-mode) of DAN/DCN**

Figure 3-1: The directional mode (D-mode) of DAN/DCN. Atom serial number of the vertex is always lower than that of the first and second tail atom. For example, for triplet 3-4-5, in directional mode, the only angle that can be selected is $\angle 435$, where atom 3 is the vertex. Therefore, no more than one angle will be restrained for a given
atom triplet and the directional mode tends to ‘spread’ over the entire structure and lead to more different atoms being included in the final DAN restraint list.

![Diagram of DAN/DCN](image)

**Figure 3-2 Arbitrary mode (A-mode) of DAN/DCN**

Figure 3-2: The arbitrary mode (A-mode) of DAN/DCN. No restriction is placed for angle selection when a certain atom triplet has been picked. For triplet 3-4-5, in addition to \( \angle 435 \) that can also be shot by directional mode, arbitrary mode as well allows angles such as \( \angle 354 \) (shown), where atom 5 is the vertex, and \( \angle 345 \) (not shown), where atom 4 is the vertex. The arbitrary mode will include all possible angles present in a structure. If the cutoff criterion for DAN is flexible enough, two or even all three angles within the same triplet may be eligible for candidacy for final restraints selection. As a result, arbitrary mode is possible to target angles that have been excluded by directional mode at the first place, but may create less atom diversities than the restraint list extracted from directional mode DAN pool. The generated DAN restraint file lists the triplet in the
order of vertex, first tail and second tail. For both modes, the atom serial number of the first tail is by definition lower than the second to avoid selection duplication.

3.2.3.3. DCN energy restraint equations

The DCN potential is the sum of the harmonic bending energy of DAN and stretching energy of DEN.

\[
E_{DCN} = k \cdot E_{DAN} + E_{DEN}
\]

\[
E_{DAN} = \sum_{i,j,k} (\theta_{ijk} - \theta_{ijk}^0(\mu,n))^2
\]

\[
E_{DEN} = \sum_{l,m} (d_{lm} - d_{lm}^0(\gamma,n))^2
\]

Equation 3-2 DCN restraint energy

The summations are taken over all angle triplets for DAN and all distance pairs for DEN. \(\theta_{ijk}\) and \(d_{lm}\) are the instantaneous angle for an atom triplet and distance for an atom pair at a conformation state during the refinement, respectively. \(\theta_{ijk}^0(\mu,n)\) and \(d_{lm}^0(\gamma,n)\) are the corresponding equilibrium angle and distance at a specific (\(n\)th, see blow) refinement step. We set the coefficient \(k\) to 0.01 as the angles are in degree. \(\theta_{ijk}(\mu,n)\) and \(d_{lm}^0(\gamma,n)\) are updated every six MD steps (when the temperature also drops 50K) according to the following equations,
\[
\begin{align*}
\theta_{ijk}^0(\mu, n + 1) &= (1 - \phi) \cdot \theta_{ijk}^0(\mu, n) + \phi \cdot \left[ \mu \theta_{ijk} + (1 - \mu) \theta_{ijk}^{\text{ref}} \right] \\
d_{lm}^0(\gamma, n + 1) &= (1 - \kappa) \cdot d_{lm}^0(\gamma, n) + \kappa \cdot \left[ \gamma d_{lm} + (1 - \gamma) d_{lm}^{\text{ref}} \right]
\end{align*}
\]

**Equation 3-3 Deformability of DCN**

The angle and distance’s next equilibrium values \((\theta_{ijk}^0(\mu, n + 1), d_{lm}^0(\gamma, n + 1))\) are functions of their current equilibrium values \((\theta_{ijk}^0(\mu, n), d_{lm}^0(\gamma, n))\), their actual instantaneous values \((\theta_{ijk}, d_{lm})\), as well as values of equivalent triplet and pair in the reference model \((\theta_{ijk}^{\text{ref}}, d_{lm}^{\text{ref}})\). For numeric stability, typically, the initial equilibrium values of the atom triplet \(\theta_{ijk}^0(\mu, 0)\) and pair \(d_{lm}^0(\gamma, 0)\) are set to be equal to those values in the starting structure. \(\phi\) and \(\kappa\) control the transition between consecutive equilibrium values. For initial relaxation, \(\phi\) and \(\kappa\) are set to 0 (hence terms including \(\mu\) and \(\gamma\) vanish) during the first three macro-cycles. After that, \(\phi\) and \(\kappa\) are set to a fixed value of 0.1. \(\mu\) and \(\gamma\) are optimized together with \(w_{\text{DCN}}\), the weight of DCN potential, via a 3D grid search. \(w_{\text{DCN}}\) is set to 0 during the last two cycles to reduce the bias effect of the target energy minimum.

### 3.2.4. Selection of reference model and \((\gamma, w_{\text{DCN}}, \mu)\) parameter group

Having the target protein’s primary structure, with *Modeller*\(^{30}\) one could perform a FASTA search chain by chain against an existing protein database so that several homologous protein chains sharing various sequence identity would be listed out as the templates for the reference model. Typically, candidates with higher identity, higher
resolution as well as longer length would be preferable as they generally ensure higher quality. We used the automatic procedure of *Modeller* program throughout the whole modeling process, including the sequence alignment as well as the template building. We chose to build five final models for each template picked, and the one with the lowest DOPE score was designated as the reference model for a chain of the target structure. After all chains (or part of them that interest) had their reference models built, these models were merged into one PDB file and served as the unique reference model for subsequent use. The parameter group \((\gamma, w_{DCN}, \mu)\) is optimized via a 3D grid search (Figure 3-3) through 180 grid points: \((0, 0.2, 0.4, 0.6, 0.8, 1)\) for \(\gamma\), \((3, 10, 30, 100, 300)\) for \(w_{DCN}\) and \((0, 0.2, 0.4, 0.6, 0.8, 1)\) for \(\mu\). At each point, ten refinements with different random seeds but otherwise identical were carried out and the result with the lowest \(R_{\text{free}}\) would represent the final refined structure at that grid point. The seed controls the assignment of initial velocities for atoms as well as the selection of DCN restraints from the pair and triplet pool. It should be noted that, due to the relatively stochastic nature of the effect of the refinement, particular practice like using a parameter group with value falling between the closest search grid points, carrying out more refinement repeats (e.g. 20) or simply picking a distinct random integer as the seed (e.g. 18593) had a chance to give considerably better results for select systems (data not shown). However, to ensure consistency, generality and valid comparison, we stuck to the same grid search strategy and used the exact integers from 1 to 10 as the ten random seeds throughout this work.
3.2.5. Input data preparation before refinement

Many proteins possess non-standard ligands and modified residues, which are usually listed as hetero-atom entries (HETATM) in the PDB files. Currently many of them are not included in the CNS database and a straight refinement with their presence in the initial structure would cause the task to cease. Previous work\textsuperscript{13} used an automated import method thus only residues and ligands recognized by CNS were involved. Here, prior to performing the refinement, we fetched the topology and parameter files of those ligands and modified residues from the Hetero-compound Information Center Uppsala.
(HIC-Up) and subsequently imported their coordinates for refinement like other compounds of a molecule. In that those ligands diffract X-ray as well, improved agreement with experiment data was expected as manifested by $R_{\text{free}}$. (supplementary Table 3).

For the case of tobacco PR-5d protein (PDB ID 1AUN), the high resolution data set obtained from the Protein Data Bank was truncated using CCP4 software into three lower resolution sets at 3.5 Å, 4.0 Å and 4.5 Å, respectively. These three synthetic sets served as the original experiment data for subsequent refinement and analysis. The starting structure for the refinement could be half-refined, theoretically predicted or manual built, but should be reasonably close to the target structure. In this work, for the full refinement of 1AUN, its reference model (PDB ID 1PCV) was selected as the starting structure, whose positions and orientations were determined by molecular replacement with Phaser$^{31}$ against each of the three low resolution data sets. Straightforwardly, as for other re-refinement tasks, the starting structure was just the known low resolution structure to be refined. In this work, the lower cutoff value for DEN and DAN selection was set to 3Å and 60°, while the upper 15Å and 120°, respectively. For both DEN and DAN, the sequence separation range was chosen to be 0 to 10 residues, and the restraints/atoms ratio was set to 1. The search probe radius for atom interactions in both DEN and DAN structures was set to be equal to DEN’s upper distance cutoff value (15Å) for all cases, except for PDB ID 2VKZ, where the value was set to 13Å in DAN.
3.2.6. Refinement protocol

Torsion angle molecular dynamics\textsuperscript{27} (TAMD) with reduced degree of freedom, combined with traditional simulated annealing\textsuperscript{26} was used as the main refinement protocol\textsuperscript{13}. The time of each MD step was 4\textit{fs}. For the annealing process, the initial temperature was set to 3,000\textdegree K, with a decreasing rate of 50\textdegree K per 6 TAMD steps. Every 6 TAMD steps could be defined as a ‘micro-cycle’, which determined the adjustment frequency of both the temperature and each DCN restraint’s equilibrium. The period the temperature dropped from 3,000\textdegree K to 0\textdegree K consisted of a ‘macro-cycle’. Each refinement task in this work, including conventional refinement, DEN refinement and DCN refinement used eight such macro-cycles. During the first three of them, $\phi$ and $\kappa$ were set to zero rather than 0.1 to allow initial relaxation. The van der Waals radii had been shrunk to 75\% of the original value during several initial macro-cycles, together with a reduced van der Waals force constant to facilitate sampling, and were thereafter fully restored in the last two cycles. Moreover, DCN restraint weight was also set to zero at the last two macro-cycles to reduce the bias of the target’s global minimum. Anisotropic overall B-factor correction and bulk solvent correction were applied for all refinements and no positional minimization used. For the sixteen re-refinement tasks, 50 steps of group B-factor minimization with a ten-fold increase for target sigma values of B-factor main/side chain bonds/angles restraints were performed and the initial values of B factors were reset to 50\AA$^2$. Ligands not by default recognized by CNS were explicitly defined as groups for group B-factor minimization. As what had been done with DCN parameters, for appropriate comparison purpose, all these refinement parameter settings were also kept identical across all test systems in this work, even though a different value of a
parameter, for instance, the multiplicity of target sigma value for group B factor minimization (data not shown), the initial temperature or the cooling rate (as one of the most important parameters whenever simulated annealing protocol is introduced), would no doubt be possible to be further optimized for a lower $R_{\text{free}}$. Upon completion of a refinement, all refined structures were sorted according to the $R_{\text{free}}$ and one with the lowest value was then picked for subsequent analysis, remodeling or other purposes.

### 3.2.7. Coding and program

Algorithms of the DCN approach, packed into several source code files and header files, were written fully compatible with version 1.3 of the Crystallography and NMR System (CNS)\textsuperscript{32,33} and compiled with Intel Fortran v10.1.015 in this work. The computation was carried out on the Shared University Grid at Rice (SUG@R) cluster platform of the Shared Computing Resources (ShareCoRe). Each refinement task was done on a single core of an Intel Xeon processor running at 2.83GHz. VMD\textsuperscript{34} and Coot\textsuperscript{35} were used for drawing purpose. TMscore\textsuperscript{19} program was used for calculations of GDT(<1Å) score and TMscore. Molprobity\textsuperscript{25} was used to evaluate Ramachandran Statistics.

### 3.3. Results and analysis

#### 3.3.1. Automatic full refinement

In our test case, we used the crystal structure of tobacco PR-5d protein (PDB ID 1AUN) and its experimental data truncated to 3.5Å, 4.0Å and 4.5Å. To allow proper assessment of the DCN approach, we repeated the refinement under exactly the same
protocol, but with two different target functions. One of them is the conventional target, which only combines the stereochemistry potential¹¹ and the experiment data (in the form of Maximum-likelihood energy⁸). The other is the conventional target plus the DEN potential. Thanks to the presence of the ‘true structure’ (to the reliability of 1.8Å), in addition to the R_free, which measures the fit of the structure to the experiment, we were able to assess the quality of the refined structure by showing the all-atom RMSD from the true structure, the global distance test (GDT) score, and the TM-score. It is shown that, among all three approaches, DCN delivers the most favorable GDT (Figure 3-6) and TM scores (Figure 3-7) among all three refinement approaches, and more accurate (i.e. lower RMSD) structure coordinates than DEN (Figure 3-5). The DCN R_free is also significantly improved over DEN and conventional refinement, except at 4.5Å, where DCN has a slightly (∼2×10⁻³) higher R-free value than DEN (Figure 1-1). It is noticed that, even in these cases where the R_free improvement by DCN is not very remarkable, the actual quality of the DCN refined structure is better than that of DEN (Figure 3-5, Figure 3-6, Figure 3-7). It is noted that at 3.5Å, conventional refinement has the lowest RMSD compared with DCN and DEN. Therefore, DCN is expected to have the best performance for refinement tasks with X-ray resolution data lower than 4Å (Table 1).
Figure 3-4 $R_{\text{free}}$ vs Resolution for Conventional, DEN and DCN
Figure 3-5  RMSD vs Resolution for Conventional, DEN and DCN

Figure 3-6  GDT(<1Å) vs Resolution for Conventional, DEN and DCN
Figure 3-7  TMscore vs Resolution for Conventional, DEN and DCN

<table>
<thead>
<tr>
<th>Resolution (Å)</th>
<th>Refinement Approach</th>
<th>R_{free}</th>
<th>All-atom RMSD</th>
<th>GDT(&lt;1Å) score</th>
<th>TMscore</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.5</td>
<td>Conventional</td>
<td>0.3267</td>
<td>2.968</td>
<td>0.9567</td>
<td>0.9887</td>
</tr>
<tr>
<td></td>
<td>DCN</td>
<td>0.3046</td>
<td>2.981</td>
<td>0.9615</td>
<td>0.9888</td>
</tr>
<tr>
<td>4</td>
<td>Conventional</td>
<td>0.3679</td>
<td>3.031</td>
<td>0.8413</td>
<td>0.9774</td>
</tr>
<tr>
<td></td>
<td>DCN</td>
<td>0.2994</td>
<td>3.010</td>
<td>0.9183</td>
<td>0.9826</td>
</tr>
<tr>
<td>4.5</td>
<td>Conventional</td>
<td>0.4724</td>
<td>3.372</td>
<td>0.3269</td>
<td>0.9241</td>
</tr>
<tr>
<td></td>
<td>DCN</td>
<td>0.3384</td>
<td>3.201</td>
<td>0.6250</td>
<td>0.9565</td>
</tr>
<tr>
<td></td>
<td>DCN</td>
<td>0.3408</td>
<td>3.113</td>
<td>0.7067</td>
<td>0.9662</td>
</tr>
</tbody>
</table>

Table 3-1  Refinement of tobacco PR-5d protein (PDB ID 1AUN) based on a homology model of a plat antifungal protein osmotin (PDB ID 1PCV) with a sequence identity of (79.51%) and an initial all-atom RMSD of 3.156Å to the 'true structure' of 1AUN.
Table 3-1: The conventional, DEN and DCN refinements were carried out at three resolution limits, truncated from the high resolution experiment data set. The best refined structures with lowest $R_{\text{free}}$ value by each approach were subsequently subject to three additional validations as all-atom RMSD, GDT(<1Å) score and TMscore to further assess the quality of the structures after each refinement. For each of the total twelve controls (four kinds of score × three resolutions), most favorable results, i.e., lowest $R_{\text{free}}$ and all-atom RMSD, and highest GDT (<1Å) and TMscore, are highlighted with bold font, whereas least with italic font. DCN gives ten out of twelve best results, and no worst, while DEN delivers two best results (one of them shared with DCN) but also two worst. Conventional refinement concedes all other ten worst values with only one best (RMSD) at the high resolution.

3.3.2. Automatic re-refinements

We also randomly selected sixteen low-resolution structures (4.0Å – 4.51Å) from the Protein Data Bank and performed re-refinements with the aid of respective high-resolution homology models. All the structures are required to have an all-atom (that is, including side chain atoms) coordinates. For certain structures, the topology and parameter files of non-standard ligands, ions, and modified residues, which are indispensable for refinement to be executed, were obtained from the Hetero-compound Information Center – Uppsala (HIC-Up). To test the performance of the DCN approach, we automatically carried out the re-refinements without any manual inspections, interruptions or manipulations throughout the refinement. In order to minimize bias, we reset the DCN potential to zero at the last two of totally eight refinement macro-cycles. As a control, identical protocol and settings were used for DEN and conventional
refinements, for each of the sixteen re-refinement tasks. These re-refinements enabled a wider and more general comparison between all three methods across the PDB database.

3.3.2.1. Results overview

Table 3-2: Results of sixteen low-resolution re-refinement tasks.

Table 3-2: $R_{\text{free}}$ and its improvement, $R_{\text{free}}-R_{\text{work}}$ as well as Ramachandran Statistics are shown. Properties of structure, experiment data and reference model of each test system are listed in Supplementary Table 1 and Supplementary Table 2. Out of a total of sixteen test systems, DCN outperforms DEN in sixteen (100%) in $R_{\text{free}}$, sixteen (100%) in $R_{\text{free}}-R_{\text{work}}$, and thirteen (81.25%) in Ramachandran Statistics. When compared with conventional, these ratios come to 100%, 87.5% and 93.75%, respectively. Moreover, 87.5% cases achieve an $R_{\text{free}}$ improvement over 0.010 by DCN with the largest one of 0.0675 (PDB ID 1XXI).
3.3.2.2. Decrease in $R_{\text{free}}$

Cross-validated free R value (termed $R_{\text{free}}$) was introduced to address the overfitting problem in macromolecular crystallography\textsuperscript{26}, and acts as an indicator of the fit between the experimental data and the refined structure, without the influence and bias of the refinement target itself throughout the process. Among our tests, all the final free R values obtained by DCN have been substantially improved over the conventional method, with a range from 0.0041 to 0.0675 (Table 3-2, Figure 3-8). Fourteen out of sixteen (87.5\%) structures have been refined with an $R_{\text{free}}$ improvement over 0.01. When compared with standalone DEN method, DCN achieves a 1.1x to 3.6x performance boost.

We illustrate the information for $R_{\text{free}}$ ($R_{\text{free}}$-$R_{\text{work}}$, Ramachandran Statistics) in Figure 3-8 (Figure 3-9, Figure 3-10). For convenient comparison purpose, values for Conventional refinements are scaled to unity (other scaled accordingly), and values for DEN refinements are used to sort the PDB IDs on the horizontal axis. Pre-scaled data are taken from Table 3-2.
3.3.2.3. Decrease in $R_{\text{free}}$ - $R_{\text{work}}$

Degree of overfitting could be assessed from the absolute value of difference between the free R and working R (termed $R_{\text{work}}$, the factor correlated with the maximum-likelihood scoring function and calculated using the reflections that are actually involved in the refinement process). Typically, $R_{\text{work}}$ should be smaller than $R_{\text{free}}$ due to the continuous optimization of maximum-likelihood function and high correlation between the function and $R_{\text{work}}$. In most of our test cases, DCN consistently delivers the smallest $R_{\text{free}} - R_{\text{work}}$ among all three methods (Table 3-2, Figure 3-9), thus minimizes the bias inherent in fitting the structure to the working set of reflection data throughout the
refinement process. The most favorable $R_{\text{free}} - R_{\text{work}}$ value for DCN is to the order of $10^{-3}$, which almost eliminates the overfitting effect, whereas for DEN and conventional, the best is to the $10^{-2}$.

Figure 3-9  $R_{\text{free}} - R_{\text{work}}$ of sixteen test systems for Conventional, DEN and DCN

3.3.2.4. Increase in Ramachandran Statistics

To further evaluate the quality of the refined structures without the availability of a high-resolution model, we carried out the Molprobity structure validation\textsuperscript{25}. The Ramachandran Statistics was calculated to assess the quality of the secondary structures. Thirteen out of sixteen DCN-refined structures exhibit a larger percentage of residues that fall in the favored regions, resulting in a higher Ramachandran Statistics, compared with
DEN-refined structures (Table 3-2, Figure 3-10). The restraints imposed by DCN add more geometry information from a high-resolution reference model, which usually possesses considerably accurate details on those secondary structures detected by the high-resolution diffraction data.

![Scaled Ramachandran Statistics](image)

**Figure 3-10** Ramachandran Statistics of sixteen test systems for Conventional, DEN and DCN

3.3.2.5. Improvement in electron density map interpretation

Along with the final structures, the phase combined sigma weighted 2F_o-F_c electron density maps, derived from the experiment amplitudes and calculated model phases after the refinements with conventional (red), DEN (blue) and DCN (green)
approach, have been shown for different features. Case one (PDB ID 1JL4, Figure 3-11) illustrates the feature of DCN than improves the map and enhances the backbone interpretation. It is observed that, among all three maps, DCN (green) is the only one showing continuous main chain density. Clear breaks are, however, observed in both DEN and conventional maps. Case two (PDB ID 2BF1, Figure 3-12) shows the structure auto-correction feature of DCN. With a 0.04 improvement over conventional $R_{free}$, DEN-refined structure allows remarkable residue shifts in several places on the main chain from the structure determined by conventional refinement. Nevertheless, it has been indicated by the corresponding density maps (blue mesh) that, selected shifts in DEN (e.g., C alpha atom in 368 GLY, chain A, shifts by 11.55Å and 9.83Å from itself in conventional and DCN, respectively) are poor-defined thus not reliable. DCN-refined structure is observed to be closer (2.24Å apart for C alpha) to the conventional-refined structure than DEN (9.83Å), and produces a self-consistent density map as well. Therefore, in some situations, DEN refined structure with significant branch deviations from conventional structure is not necessarily superior. DCN automatically reduces the remarkable disagreement and corrects the coordinates by favoring the better structure, and simultaneously fitting well to the density map generated by DCN itself. Even though appearing quite close to conventional structure, the DCN structure actually has a much higher quality with an improvement of 0.060 (Table 3-2) in $R_{free}$—that is more than 14% of its own value.
Figure 3-11 View of backbone trace. PDB ID 1JL4 centered on A23-THR is shown.
3.3.2.6. Re-refinement with NCS and experimental phase

DCN could easily incorporate other information to further facilitate the refinement process and improve the results (Table 3-3, Figure 3-13, Figure 3-14). We carried out refinements with Non-Crystallographic Symmetry (NCS) when related information was explicitly provided in the header of the PDB files. We then tested the effect of NCS information by intentionally turning NCS off before repeating the
refinement. It is observed that the NCS information has improved the structures for all three methods. Among them, DCN-refined structure is the most accurate in $R_{\text{free}}$ no matter NCS is used or not. Therefore when applicable, it is encouraged to add NCS information during DCN refinement for seeking the best structure with the lowest $R_{\text{free}}$. In cases where experimental phase is obtained, for example, by single or multiple isomorphous replacement, using the MLHL\textsuperscript{10} target function (with experimental phase) would produce better result than the MLF target (without experimental phase) for DCN in terms of Ramachandran Statistics. Whereas conventional and DEN are also expected to benefit from the phase information, it is noted that, once again, improvement by DCN over the conventional and DEN refinements persists regardless of the availability of experiment phase data.

<table>
<thead>
<tr>
<th>PDB ID</th>
<th>Target</th>
<th>NCS</th>
<th>$R_{\text{free}}$</th>
<th>DCN Improvement</th>
<th>Net gain fraction over DEN improvement</th>
<th>Ramachandran Statistics</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>DCN</td>
<td>Conventional</td>
<td>DEN</td>
</tr>
<tr>
<td>1YM7</td>
<td>MLF</td>
<td>No</td>
<td>0.3398</td>
<td>0.3169</td>
<td>0.0229</td>
<td>103%</td>
</tr>
<tr>
<td></td>
<td>MLF</td>
<td>Yes</td>
<td>0.2764</td>
<td>0.2723</td>
<td>0.0041</td>
<td>64%</td>
</tr>
<tr>
<td>3FUS</td>
<td>MLF</td>
<td>No</td>
<td>0.4584</td>
<td>0.4001</td>
<td>0.0583</td>
<td>32%</td>
</tr>
<tr>
<td></td>
<td>MLHL</td>
<td>No</td>
<td>0.4187</td>
<td>0.4007</td>
<td>0.0159</td>
<td>38%</td>
</tr>
</tbody>
</table>

Table 3-3  Refinement with and without NCS or experiment phase information
Figure 3-13 $R_{\text{free}}$ vs availability of NCS information

Figure 3-14 $R_{\text{free}}$ (and Ramachandran) vs availability of experiment phase.
3.4. Supplementary information

Table 3-4  A list of the structure property of all the re-refinement cases.

<table>
<thead>
<tr>
<th>PDB ID</th>
<th>Resolution (Å)</th>
<th>Number of Chains</th>
<th>Sequence Length</th>
<th>No. of Observed Protein Residues</th>
<th>No. of All Observed Residues*</th>
<th>Ramachandran Statistics of Deposited Structure</th>
<th>Deposited R_free</th>
<th>Deposited R_work</th>
<th>Re-calculated R_work</th>
<th>Difference in R_work</th>
</tr>
</thead>
<tbody>
<tr>
<td>1ISR</td>
<td>4.00</td>
<td>1</td>
<td>490</td>
<td>448</td>
<td>451</td>
<td>0.948</td>
<td>0.259</td>
<td>0.237</td>
<td>0.2414</td>
<td>-0.004</td>
</tr>
<tr>
<td>1L14</td>
<td>4.30</td>
<td>4</td>
<td>557</td>
<td>597</td>
<td>597</td>
<td>0.922</td>
<td>0.431</td>
<td>0.400</td>
<td>0.3680</td>
<td>0.052</td>
</tr>
<tr>
<td>1RSU</td>
<td>4.50</td>
<td>11</td>
<td>4259</td>
<td>3517</td>
<td>3527</td>
<td>0.805</td>
<td>0.373</td>
<td>0.345</td>
<td>0.2531</td>
<td>0.092</td>
</tr>
<tr>
<td>1X0E</td>
<td>4.10</td>
<td>10</td>
<td>3562</td>
<td>3532</td>
<td>3544</td>
<td>0.937</td>
<td>0.369</td>
<td>0.366</td>
<td>0.3929</td>
<td>-0.027</td>
</tr>
<tr>
<td>1YF1</td>
<td>4.50</td>
<td>4</td>
<td>574</td>
<td>574</td>
<td>772</td>
<td>0.968</td>
<td>0.343</td>
<td>0.295</td>
<td>0.2949</td>
<td>0.000</td>
</tr>
<tr>
<td>1YF7</td>
<td>4.50</td>
<td>4</td>
<td>2756</td>
<td>2422</td>
<td>2422</td>
<td>0.899</td>
<td>0.279</td>
<td>0.224</td>
<td>0.2011</td>
<td>0.033</td>
</tr>
<tr>
<td>2A62</td>
<td>4.50</td>
<td>1</td>
<td>322</td>
<td>319</td>
<td>325</td>
<td>0.749</td>
<td>0.346</td>
<td>0.271</td>
<td>0.2690</td>
<td>0.002</td>
</tr>
<tr>
<td>2BF1</td>
<td>4.00</td>
<td>1</td>
<td>316</td>
<td>304</td>
<td>354</td>
<td>0.680</td>
<td>0.388</td>
<td>0.385</td>
<td>0.3920</td>
<td>-0.007</td>
</tr>
<tr>
<td>2G67</td>
<td>4.15</td>
<td>3</td>
<td>1044**</td>
<td>954</td>
<td>975</td>
<td>0.896</td>
<td>0.382</td>
<td>0.377</td>
<td>0.3332</td>
<td>0.054</td>
</tr>
<tr>
<td>2G7N</td>
<td>4.00</td>
<td>4</td>
<td>1336</td>
<td>1320</td>
<td>1365</td>
<td>0.793</td>
<td>0.287</td>
<td>0.237</td>
<td>0.2793</td>
<td>-0.042</td>
</tr>
<tr>
<td>2QAQ</td>
<td>4.00</td>
<td>3</td>
<td>1206</td>
<td>702</td>
<td>705</td>
<td>0.895</td>
<td>0.392</td>
<td>0.376</td>
<td>0.3652</td>
<td>0.011</td>
</tr>
<tr>
<td>2KZ6</td>
<td>4.00</td>
<td>6</td>
<td>11814</td>
<td>10941</td>
<td>10947</td>
<td>0.935</td>
<td>0.268</td>
<td>0.268</td>
<td>0.2305</td>
<td>0.037</td>
</tr>
<tr>
<td>2YH1</td>
<td>4.00</td>
<td>2</td>
<td>638</td>
<td>570</td>
<td>570</td>
<td>0.977</td>
<td>0.300</td>
<td>0.247</td>
<td>0.3150</td>
<td>-0.068</td>
</tr>
<tr>
<td>3ALZ</td>
<td>4.51</td>
<td>2</td>
<td>630</td>
<td>526</td>
<td>528</td>
<td>0.812</td>
<td>0.338</td>
<td>0.326</td>
<td>0.2276</td>
<td>0.098</td>
</tr>
<tr>
<td>3FUS</td>
<td>4.00</td>
<td>1</td>
<td>316</td>
<td>304</td>
<td>359</td>
<td>0.700</td>
<td>0.354</td>
<td>0.346</td>
<td>0.3775</td>
<td>-0.032</td>
</tr>
<tr>
<td>3V3Z</td>
<td>4.20</td>
<td>14</td>
<td>1624</td>
<td>1520</td>
<td>1584</td>
<td>0.888</td>
<td>0.334</td>
<td>0.296</td>
<td>0.3851</td>
<td>-0.059</td>
</tr>
</tbody>
</table>

Average 4.20 4.4 1965 1781 1812 0.863 0.342 0.315 0.3072 0.008
Maximum 4.51 11 11814 10941 10947 0.977 0.453 0.420 0.3929 0.098
Minimum 4.00 1 316 304 325 0.680 0.268 0.224 0.2011 -0.068

*All observed residues denote the sum of residue entries of protein, nucleic, heterogen, solvent that are observed and used in the refinement.

**Sequence length of 2137 in PDB website is, however, recorded as 1047. This is because three modified residues of ACE were categorized as heterogen entries in the PDB structure file, but denoted as ‘X’ and included in the FASTA sequence file. These residues did not take part in the homology modeling process, did not have a corresponding residue in the reference structure, and as a result were not counted into sequence length or protein backbone residues.

***Chain M of 1RSU consists of unknown residues (UNK) and were excluded before refinement. Therefore number of protein residues is smaller than sequence length minus number of missing residues.

The sequence length varies from 316 to 11814 and represents a broad range of proteins with various sizes. Deposited values of R_free and R_work were directly taken from the PDB header. Re-calculated R_work denotes the value determined by CNS at the initial stage of refinement. Exact values could fluctuate because of multiple factors including but not limited to bulk solvent and B-factor correction, software and hardware environment of computing. The most and least favorable values in difference of R_work are 0.098 and -0.068, with an average difference of 0.008 (a good overall reproduction).
Moreover, the proportion of cases that results in a lower reproduced $R_{\text{work}}$ is 9 out of 16 (56.25%). It is a substantial improvement over previous work\textsuperscript{13} where these values are 0.053 and -0.109, with an average of -0.025 and only 4 out of 19 (21.05%) cases with lower reproduced $R_{\text{work}}$, respectively. This is due to the inclusion of all observed residues in the refinement together with other reasons stated above.

Table 3-5 A list of property of experiment data and reference model.

<table>
<thead>
<tr>
<th>PDB ID</th>
<th>Resolution (Å)</th>
<th>Total No. of Diffractions</th>
<th>No. of Diffractions per Residue</th>
<th>Reference Model</th>
<th>Resolution (Å)</th>
<th>Sequence Identity</th>
</tr>
</thead>
<tbody>
<tr>
<td>1ISR</td>
<td>4.00</td>
<td>6528</td>
<td>14.70</td>
<td>2.20</td>
<td>552</td>
<td>99.8%</td>
</tr>
<tr>
<td>2BF1</td>
<td>4.00</td>
<td>5842</td>
<td>16.50</td>
<td>0.79</td>
<td>280</td>
<td>97.8%</td>
</tr>
<tr>
<td>2Q7N</td>
<td>4.00</td>
<td>35237</td>
<td>25.81</td>
<td>1.36</td>
<td>1852</td>
<td>74.1%</td>
</tr>
<tr>
<td>2YHU</td>
<td>4.00</td>
<td>11151</td>
<td>19.56</td>
<td>0.97</td>
<td>555</td>
<td>99.8%</td>
</tr>
<tr>
<td>2QAG</td>
<td>4.00</td>
<td>40338</td>
<td>10.11</td>
<td>1.13</td>
<td>2124</td>
<td>69.5%</td>
</tr>
<tr>
<td>3PUS</td>
<td>4.00</td>
<td>5841</td>
<td>16.27</td>
<td>0.78</td>
<td>279</td>
<td>39.1%</td>
</tr>
<tr>
<td>2VKZ</td>
<td>4.00</td>
<td>160231</td>
<td>14.64</td>
<td>0.78</td>
<td>8547</td>
<td>96.0%</td>
</tr>
<tr>
<td>1XI1</td>
<td>4.10</td>
<td>35818</td>
<td>10.11</td>
<td>1.13</td>
<td>4020</td>
<td>99.8%</td>
</tr>
<tr>
<td>2137</td>
<td>4.15</td>
<td>11807</td>
<td>12.11</td>
<td>0.66</td>
<td>645</td>
<td>99.8%</td>
</tr>
<tr>
<td>3US2</td>
<td>4.20</td>
<td>14037</td>
<td>8.86</td>
<td>0.47</td>
<td>744</td>
<td>81.3%</td>
</tr>
<tr>
<td>1JL4</td>
<td>4.30</td>
<td>5880</td>
<td>10.56</td>
<td>1.16</td>
<td>645</td>
<td>94.7%</td>
</tr>
<tr>
<td>1YE1</td>
<td>4.50</td>
<td>3442</td>
<td>4.46</td>
<td>0.46</td>
<td>354</td>
<td>99.0%</td>
</tr>
<tr>
<td>2A62</td>
<td>4.50</td>
<td>3929</td>
<td>12.09</td>
<td>0.99</td>
<td>323</td>
<td>100.0%</td>
</tr>
<tr>
<td>1HEU</td>
<td>4.50</td>
<td>56023</td>
<td>15.88</td>
<td>0.49</td>
<td>1721</td>
<td>79.9%</td>
</tr>
<tr>
<td>1YM7</td>
<td>4.50</td>
<td>23110</td>
<td>9.54</td>
<td>0.50</td>
<td>1210</td>
<td>100.0%</td>
</tr>
<tr>
<td>3ALZ</td>
<td>4.51</td>
<td>13413</td>
<td>25.40</td>
<td>1.33</td>
<td>783</td>
<td>92.2%</td>
</tr>
<tr>
<td>Average</td>
<td>4.20</td>
<td>27047</td>
<td>17.11</td>
<td>1.01</td>
<td>1535</td>
<td>91.3%</td>
</tr>
<tr>
<td>Minimum</td>
<td>4.00</td>
<td>3442</td>
<td>4.46</td>
<td>0.46</td>
<td>279</td>
<td>37.0%</td>
</tr>
<tr>
<td>Maximum</td>
<td>4.51</td>
<td>160231</td>
<td>8547</td>
<td>57.25</td>
<td>8547</td>
<td>100.0%</td>
</tr>
</tbody>
</table>

Diffraction data was fetched from the Protein Data Bank and converted into CNS recognized hkl file with no other modification. For data set without explicit $R_{\text{free}}$ flag, a free data set was generated with the number of 5% of total diffractions by CCP4, a default value setting in the software. For several test systems (e.g. 1JL4), there exist diffraction entries with resolution higher than that given in the PDB header. Those entries were excluded before refinement and we only used the portion of data that agrees with the published resolution. Reference models were chosen according to several preferences
stated in ‘Method’. The sequence identity and resolution values in the table are linearly averaged by chain length according to the sequence, with information of resolution and identity of each chain’s corresponding template.

Table 3-6 Comparison of results between this work (ligands included) and previous work\textsuperscript{13} (ligands excluded) with Conventional and DEN approach

<table>
<thead>
<tr>
<th>PDB ID</th>
<th>Ligands not defined in CNS</th>
<th>Approach</th>
<th>$R_{free}$ (this work)</th>
<th>$R_{free}$ (previous work)</th>
<th>Improvement</th>
</tr>
</thead>
<tbody>
<tr>
<td>1ISR</td>
<td>GD</td>
<td>Conventional</td>
<td>0.223</td>
<td>0.237</td>
<td>0.014</td>
</tr>
<tr>
<td></td>
<td></td>
<td>DEN</td>
<td>0.216</td>
<td>0.233</td>
<td>0.017</td>
</tr>
<tr>
<td>2BF1</td>
<td>BMA,NDG</td>
<td>Conventional</td>
<td>0.487</td>
<td>0.492</td>
<td>0.005</td>
</tr>
<tr>
<td></td>
<td></td>
<td>DEN</td>
<td>0.443</td>
<td>0.479</td>
<td>0.036</td>
</tr>
<tr>
<td>1XXI</td>
<td>ADP</td>
<td>Conventional</td>
<td>0.382</td>
<td>0.465</td>
<td>0.083</td>
</tr>
<tr>
<td></td>
<td></td>
<td>DEN</td>
<td>0.322</td>
<td>0.407</td>
<td>0.085</td>
</tr>
<tr>
<td>1YE1</td>
<td>HEM</td>
<td>Conventional</td>
<td>0.338</td>
<td>0.350</td>
<td>0.012</td>
</tr>
<tr>
<td></td>
<td></td>
<td>DEN</td>
<td>0.332</td>
<td>0.412</td>
<td>0.085</td>
</tr>
<tr>
<td>2QAG</td>
<td>GTP,GDP</td>
<td>Conventional</td>
<td>0.405</td>
<td>0.401</td>
<td>-0.004</td>
</tr>
<tr>
<td></td>
<td></td>
<td>DEN</td>
<td>0.388</td>
<td>0.392</td>
<td>0.004</td>
</tr>
<tr>
<td>2VKZ</td>
<td>CER,FMN</td>
<td>Conventional</td>
<td>0.312</td>
<td>0.337</td>
<td>0.025</td>
</tr>
<tr>
<td></td>
<td></td>
<td>DEN</td>
<td>0.299</td>
<td>0.327</td>
<td>0.028</td>
</tr>
<tr>
<td>Average</td>
<td></td>
<td></td>
<td>0.343</td>
<td>0.369</td>
<td>0.026</td>
</tr>
</tbody>
</table>

Results of Conventional and DEN have been substantially improved in this work due to the inclusion of ligands during the refinement as well as factors such as software and hardware computing environment. This helps set up higher-standard controls in the first place. In most cases, improvement by DEN over Conventional is also larger in this work. DCN performance over DEN was calculated and compared with these ‘better’ DEN results.
3.5. Discussion and implementation

More elaborate tailoring of DCN settings, such as carefully adjusting the DCN model angle criteria, and selecting certain regions of molecule that have more reliable reference structures present, is expected to further enhance DCN’s performance for structure predication, improvement and molecular-replacement phasing\textsuperscript{36}. When best homology model found in the database does not have satisfactory sequence identity or resolution, it is possible to assign two distinct homology models to identical chain of a molecule for two parts of DCN, such that DAN and DEN information could be simultaneously absorbed from independent sources, to avoid the refinement from being guided under a single reference structure in an unfavorable or unreliable direction. Also, deformation of angular network and distance network do not need to be synchronous. A more robust conformation sampling may emerge when uneven frequencies or interleaved phases of the deformation period for the two networks are deployed. Moreover, DCN could be effortlessly implemented in grid computing servers with an online GUI\textsuperscript{37}, allowing interested users to carry it out via a web portal with ease.
Chapter 4

Normal Mode Analysis and its Integration with DCN

4.1. Motivation and introduction

The success of Deformable Complex Network has been able to push the structure accuracy to a new limit. All these improvements come from a dedicated method that aids the positional refinement process; temperature factors that account for the dynamics of the systems in DCN, on the other hand, are refined via isotropic ‘group B’, i.e., two B factors for each residue, one for main chain atoms and the other for side chain. Although the primary purpose of refinement is to get the averaged positions of all the atoms, which are visualized by various analysis tools such as VMD\textsuperscript{34} and PyMOL\textsuperscript{38}; temperature factors that describes the atom’s fluctuation around its mean plays an important role as well, and sometimes even reveals important functionality of a macromolecule\textsuperscript{39}. We are interested in how DCN and a better temperature factor refinement method are mutually beneficial, and whether their improvements stack, delivering structures of best accuracy.
This is where the Normal Mode Analysis (NMA)\textsuperscript{40} kicks in, more specifically, a coarse grained elastic NMA that treat each residue as a mass point whose position is assigned from the coordinate of C\textsuperscript{α} atom. Generally speaking, atomic motions within a molecule are neither independent nor isotropic. From the standpoint of modeling, NMA is able to describe the anisotropy and dependence of atoms, which is closer to the real case. From the perspective of refinement – a very demanding application, NMA is capable of representing the large collective motions with very few low frequency modes, overcoming a tough issue that increased macromolecule size typically means too many refinement parameters for diffraction data to afford. In addition, the anisotropy introduced here explores a whole new conformation space that no isotropic refinement can reach, thus offers the opportunity for better agreement between a structure and diffraction data.

4.2. Method

4.2.1. Summary

Similar to how we carried out Deformable Complex Network refinement, we need to have in hand the target structure PDB file, diffraction data, and a reference model as described in 3.2.1. We turned off the isotropic B-factor refinement in CNS, and supersede it with NMA based refinement, followed by a macro-cycle of positional refinement with created or updated ANISOU record in the structure file. New structure with relocated atoms is then subject to the next NMA refinement. These two procedures alternate for a total of eight cycles, and that defines one complete refinement task. Because initial velocity assignments for simulated annealing vary with random seed, with all others
being equal, we change the random number from integer 1 to 10 and repeat the refinement task 10 times. Parameters group \((\gamma, W_{DCN}, \mu)\) for DCN is subject to a grid search as usual.

### 4.2.2. Target function

The target function does not change externally because anisotropic U only influences the calculation of \(E_{\text{experiment}}\):

\[
E_{\text{target}} = E_{\text{stereo}} + w_a E_{\text{experiment}} + w_{DCN} E_{DCN}
\]

where this time \(E_{\text{experiment}}\) still takes the form of amplitude-based maximum likelihood function. Within that function the calculated structure factor has new Debye-Waller factor

\[
e^{-\frac{1}{2}(q^T U q)}
\]

Equation 4-1 Anisotropic Debye-Waller factor

instead of an isotropic one

\[
e^{-\frac{1}{16\pi^2}(|h|^2 g)}
\]

Equation 4-2 Isotropic Debye-Waller factor
where $U_j$ is a $3 \times 3$ symmetric matrix representing the anisotropic motion from three dimensions

$$
U_j = \langle \Delta r_j \times \Delta r_j \rangle = \begin{pmatrix}
\langle \Delta r_x^2 \rangle_j & \langle \Delta r_x \Delta r_y \rangle_j & \langle \Delta r_x \Delta r_z \rangle_j \\
\langle \Delta r_y \Delta r_x \rangle_j & \langle \Delta r_y^2 \rangle_j & \langle \Delta r_y \Delta r_z \rangle_j \\
\langle \Delta r_z \Delta r_x \rangle_j & \langle \Delta r_z \Delta r_y \rangle_j & \langle \Delta r_z^2 \rangle_j
\end{pmatrix}
$$

Equation 4-3 Anisotropic $U$ matrix

### 4.2.3. Normal Mode Analysis

#### 4.2.3.1. Introduction

Normal Mode Analysis is a method to decompose and then describe the motion of a harmonic system by a linear combination of normal modes. In physics, it involves identification of the degree of freedom (dof) for all the objects in the system, compute the Hamiltonian of the whole system, and calculate the 2$^{\text{nd}}$ derivative of the Hamiltonian with respect to all dof pairs thus construct a Hessian Matrix. What’s next is pure linear algebra, i.e., to solve the eigenvalues and eigenvectors of Hessian Matrix, by well-developed computational algorithms. In our case, since the molecule is coarse grained to the level of residue, these ‘objects’ are C alpha atoms standing for individual residues. It is assumed, as the elastic network$^{41}$ does, that the current structure is already at its energy minimum. It enables us to do the NMA without having to go through an initial positional minimization, which almost always distorts the geometry too detrimentally to use in
subsequent refinement. Arising from the elastic network, a so called “tip effect” could also prevents the success of application in refinement because the eigenvector profile sometimes exhibits irrationally large values for select residues. This issue was somehow addressed by doing the elastic normal mode procedure in internal coordinate with a modified potential function\textsuperscript{42}.

4.2.3.2. Modeling anisotropic atomic motions with Normal Mode Analysis

Atomic motions within a molecule should be anisotropic, though sometimes describing them with isotropic B factor tends to be a good approximation. In principle, if there are enough diffraction data to afford an anisotropic refinement (typically at ultra high resolution), one is expected to do so.

However, for data collected at low resolution, we can tell from Table 3-5 that on average there are only 17 working diffractions and 1 free diffraction per residue. A full anisotropic refinement requires 6 parameters per atom which is too demanding to utilize in low resolution refinement. Normal Mode refinement, on the other hand, can be performed with user-defined number of parameters, based on that normal modes are collective variables and a few low frequency normal modes can effectively model the anisotropy of the entire structure\textsuperscript{43}. This feature enables the possibility to do anisotropic refinement on all atoms without the danger of overfitting.

The essence of Normal Mode Analysis is to reconstruct the anisotropic $U$ of atoms with normal mode parameters. Recall that $U$ is defined as
Anisotropic U sometimes is also referred to as ADP (Atomic Displacement Parameter). Here, instead of simply refining the 6 independent elements of U matrix, we express $\Delta \mathbf{r}_j$ as

$$\Delta \mathbf{r}_j = \mathbf{E}_j \sigma$$

**Equation 4-4 atomic displacement via normal mode parameter**

where $\mathbf{E}_j$ is a $3 \times M$ sub-matrix of a $3N \times M$ matrix $\mathbf{E}$ that consists of $M$ columns of normalized eigenvectors

$$\mathbf{E} = \begin{pmatrix}
\mathbf{E}_{1x}^1 & \mathbf{E}_{1x}^2 & \cdots & \mathbf{E}_{1x}^i & \cdots & \mathbf{E}_{1x}^M \\
\mathbf{E}_{1y}^1 & \mathbf{E}_{1y}^2 & \cdots & \mathbf{E}_{1y}^i & \cdots & \mathbf{E}_{1y}^M \\
\mathbf{E}_{1z}^1 & \mathbf{E}_{1z}^2 & \cdots & \mathbf{E}_{1z}^i & \cdots & \mathbf{E}_{1z}^M \\
\vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\
\mathbf{E}_{jx}^1 & \mathbf{E}_{jx}^2 & \cdots & \mathbf{E}_{jx}^i & \cdots & \mathbf{E}_{jx}^M \\
\mathbf{E}_{jy}^1 & \mathbf{E}_{jy}^2 & \cdots & \mathbf{E}_{jy}^i & \cdots & \mathbf{E}_{jy}^M \\
\mathbf{E}_{jz}^1 & \mathbf{E}_{jz}^2 & \cdots & \mathbf{E}_{jz}^i & \cdots & \mathbf{E}_{jz}^M \\
\vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\
\mathbf{E}_{N x}^1 & \mathbf{E}_{N x}^2 & \cdots & \mathbf{E}_{N x}^i & \cdots & \mathbf{E}_{N x}^M \\
\mathbf{E}_{N y}^1 & \mathbf{E}_{N y}^2 & \cdots & \mathbf{E}_{N y}^i & \cdots & \mathbf{E}_{N y}^M \\
\mathbf{E}_{N z}^1 & \mathbf{E}_{N z}^2 & \cdots & \mathbf{E}_{N z}^i & \cdots & \mathbf{E}_{N z}^M 
\end{pmatrix}$$

**Equation 4-5 Eigenvector matrix**
$N$ is the number of nodes (atoms or, in coarse grained mode, the residue) and $M$ is the total number of lowest eigenvector we want to use. $E_j$ is therefore the $(3j-2)$th to $3j$th rows. $\sigma$ is a vector composed of all $M$ eigenvalues, arranged in a column.

$$\sigma = \begin{pmatrix} \sigma_1 \\ \sigma_2 \\ \vdots \\ \sigma_M \end{pmatrix}$$

**Equation 4-6 Eigenvalue vector**

Both $E$ and $\sigma$ are obtained by solving the system’s $3N \times 3N$ Hessian matrix

$$H = \begin{pmatrix} \frac{\partial^2 V}{\partial x_1^2} & \frac{\partial^2 V}{\partial x_1 \partial y_1} & \cdots & \frac{\partial^2 V}{\partial x_1 \partial z_N} \\ \frac{\partial^2 V}{\partial y_1 \partial x_1} & \frac{\partial^2 V}{\partial y_1^2} & \cdots & \frac{\partial^2 V}{\partial y_1 \partial z_N} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 V}{\partial z_N \partial x_1} & \frac{\partial^2 V}{\partial z_N \partial y_1} & \cdots & \frac{\partial^2 V}{\partial z_N^2} \end{pmatrix}$$

**Equation 4-7 Hessian Matrix**

Now we can re-write $U$ matrix
\[ U_j = \langle \Delta r_j \times \Delta r_j \rangle = \langle E_j \sigma \times E_j \sigma \rangle = E_j \langle \sigma \times \sigma \rangle E_j^T \]

\[
= E_j \begin{pmatrix}
\langle \sigma_1^2 \rangle & \langle \sigma_1 \sigma_2 \rangle & \cdots & \langle \sigma_1 \sigma_M \rangle \\
\langle \sigma_2 \sigma_1 \rangle & \langle \sigma_2^2 \rangle & \cdots & \langle \sigma_2 \sigma_M \rangle \\
\vdots & \vdots & \ddots & \vdots \\
\langle \sigma_M \sigma_1 \rangle & \langle \sigma_M \sigma_2 \rangle & \cdots & \langle \sigma_M^2 \rangle 
\end{pmatrix} E_j^T
\]

where \( M \) diagonal terms and \( M(M-1)/2 \) off-diagonal terms. In normal mode refinement, we assume \( E \) matrix to be fixed and optimize \( \langle \sigma_i \sigma_j \rangle \) against experiment data. In this regard, the fluctuation of atoms has been expressed by a linear combination of deterministic normal mode and adjustable amplitude (weight) for each mode. Choice of \( M \) is completely user-controlled. However, it is worth noting that a too small \( M \) may result in the loss of fluctuation details by discarding high frequency modes, while a large \( M \) may easily cause overfitting due to an increase in parameter-to-data ratio, as the number of parameter scales with \( M^2 \).

The NMA variant we use in this work is an improved version of elastic NMA. There are several features. First, it is C\(^\alpha\) based. Solving a large-scale linear algebra problem is computational expensive, which could take from minutes to days depending on the size of full-atom system. Diagonalization of large matrices alone can exhaust computer memories. By coarse-graining to C\(^\alpha\) trace for Hamiltonian derivation and normal mode calculation, and then extrapolating C\(^\alpha\)’s ADP to all other atoms within the same residue for subsequent refinement, we have been able to extend NMA to much larger system with reasonable computing resources, and retain sufficient fluctuation details. Second, it is an elastic network, which assumes that the current structure is at the
valley of the energy landscape. Conventional Normal Mode Analysis must be performed with respect to ‘equilibrium’ distances, requiring an initial energy minimization beforehand. This positional minimization could easily distort structures with good geometry, particularly at low resolution. Whereas for eNMA, the system is already at energy minimum defined by the harmonic network potential since we did not use any other physics based force field such as AMBER\textsuperscript{44} or CHARMM\textsuperscript{45}, availing immediately the construction and subsequent diagonalization of Hessian matrix. Third, to prevent ‘tip effect’ – a specific issue arising from the use of elastic network – that in some circumstances unrealistic eigenvector profiles could be produced, as manifested by extremely large fluctuations for select/unpredictable nodes and (due to normalization) vanishing fluctuations for others, the NMA is carried out in internal coordinates(IC) with modified Hamiltonian. It was shown\textsuperscript{42} that tip effect is suppressed and quality of mode has been improved for possibility of meaningful applications such as refinement\textsuperscript{39,46}.

![Figure 4-1 System in internal coordinates](image-url)
\[
V = \frac{\gamma}{2} \sum_i \sum_j h_{ij} \left( |r_{ij}| - |r^0_{ij}| \right)^2 + \frac{\omega}{2} \sum_{\alpha} (\phi_{\alpha} - \phi^0_{\alpha})^2
\]

\[
h_{ij} = \begin{cases} 
1 & |r^0_{ij}| \leq r_c \\ \omega = \zeta \min(H^0_{\alpha\alpha}) \\
0 & |r^0_{ij}| > r_c
\end{cases}
\]

Equation 4-8 Hamiltonian for modified eNMA

where \( r_c \) is a cutoff radius, \( \phi_{\alpha} \) stands for N-2 both bond angles \( \theta \) and N-3 pseudodihedral angles \( \varphi \) in IC, provided the system has N C\(^a\)s. \( \gamma \) is relative to \( \omega \) and can be set to 1.0 for simplification. \( H^0_{\alpha\alpha} \) are diagonal elements of Hessian matrix of conventional eNMA potential in IC. \( \zeta \) is the adjustable stiffness of the angular term. It is wise to assign it with a larger value when system tends to be more flexible, typically within a range between 3 and 30.

4.2.4. Integrating NMA with DCN

In order to take advantage of the power from both DCN and NMA, we need to integrate them into a standalone refinement task. We realize this purpose by a Linux script that runs elastic NMA, normal mode refinement and DCN positional refinement sequentially and repeatedly. The elastic NMA is a fairly independent program\(^{39}\) as it only asks for the input structure’s C\(^a\) trace and nothing else. The Normal Mode refinement then absorbs the freshly calculated eigenvalues and eigenvectors, and performs the fitting with experimental data. This process also needs libraries such as topology and atomic scattering factor, and is done using a program extended from REFMAC 5\(^{47}\). DCN refinement is currently implemented in CNS 1.3. CNS can only carry out isotropic
refinement because to date it does not accept ANISOU records from a PDB file. In this work, even though we do not use CNS to do anisotropic refinement, we need it to recognize ANISOU and be able to calculate structure factors as well as derivatives in an anisotropic fashion.

In real computation structure factors are calculated using Fourier Transform with respect to atomic electron densities. To obtain an angular dependent density $\rho(r)$, we begin from the five Gaussian approximation for atomic form vector\textsuperscript{48}

$$f_{\text{atom}}(s) = \sum_{k=1}^{5} a_k \exp\left(-\frac{b_k s^2}{4}\right)$$

\textbf{Equation 4-9 Atomic form factor by 5-Gaussian approximation}

where $b_5 = 0$, $s = \frac{1}{2\pi} \mathbf{q}$, $a_i$ and $b_i$ are atomic scattering parameters. Equation 4-9’s FT gives atomic electron density

$$\rho_{\text{atom}}(r) = (4\pi)^\frac{3}{2} \sum_{k=1}^{5} \frac{a_i}{b_k^{3/2}} \exp\left(-\frac{4\pi^2 r^2}{b_k}\right)$$

\textbf{Equation 4-10 Atomic electron density by 5-Gaussian approximation}

In case of isotropic Debye-Waller factor

$$f_{\text{Debye-Waller}}^{\text{iso}} = \exp\left(-\frac{B s^2}{4}\right)$$
Isotropic atomic form factor becomes

\[
\begin{align*}
\mathbf{f}_{\text{iso}}(s) &= \mathbf{f}_{\text{atom}}(s)\mathbf{f}_{\text{Debye-Waller}}^\text{iso} \\
&= \sum_{k=1}^{5} a_k \exp \left(-\frac{b_k s^2}{4}\right) \exp \left(-\frac{B s^2}{4}\right) \\
&= \sum_{k=1}^{5} a_k \exp \left(-\frac{(b_k + B) s^2}{4}\right) \\
&= \sum_{k=1}^{5} a_k \exp \left(-\frac{(B_{k}^\text{total}) s^2}{4}\right)
\end{align*}
\]

Equation 4-11 Isotropic atomic form factor by 5-Gaussian approximation

where

\[
B_{k}^\text{total} = b_k + B + B_{\text{off}}
\]

\(B_{\text{off}}\) is an artificial offset to minimize aliasing during FFT computation. It is straightforward that by applying inverse FT to Equation 4-11 we get an isotropic atomic electron density

\[
\rho_{\text{iso}}(r) = \left(4\pi\right)^{\frac{3}{2}} \sum_{k=1}^{5} \frac{a_i}{B_{k}^\text{total}^{\frac{3}{2}}} \exp \left(-\frac{4\pi^2 r^2}{B_{k}^\text{total}}\right)
\]

Equation 4-12 Isotropic atomic electron density by 5-Gaussian approximation

It is equivalent to think in a way that isotropic motion of the atom effectively ‘disperses’ its scattering impact. Using similar procedure to introduce angular dependence from atom’s anisotropic displacement, recall the anisotropic Debye-Waller factor
Therefore the anisotropic atomic form factor is

\[
f_{\text{aniso}}(\mathbf{s}) = f_{\text{atom}}(\mathbf{s}) f_{\text{Debye-Waller}}^{\text{aniso}} = \sum_{k=1}^{s} a_k \exp \left( \frac{-b_k \mathbf{s}^\top \mathbf{s}}{4} \right) \exp \left( -2\pi^2 \mathbf{s}^\top \mathbf{U}s \right)
\]

\[
= \sum_{k=1}^{s} a_k \exp \left( -2\pi^2 \mathbf{s}^\top \left( \mathbf{U} + \frac{b_k}{8\pi^2} \right) \mathbf{s} \right)
\]

\[
= \sum_{k=1}^{s} a_k \exp \left( -2\pi^2 \mathbf{s}^\top \mathbf{U}_k^{\text{total}} \mathbf{s} \right)
\]

**Equation 4-13 Anisotropic Atomic form factor by 5-Gaussian approximation**

Here we denote

\[
\mathbf{U}_k^{\text{total}} = \mathbf{U} + \frac{b_k}{8\pi^2} \times \mathbf{I} + \frac{B_{\text{off}}}{8\pi^2} \times \mathbf{I}
\]

\[
= \begin{pmatrix}
U_{11} + \frac{b_k + B_{\text{off}}}{8\pi^2} & U_{12} & U_{13} \\
U_{21} & U_{22} + \frac{b_k + B_{\text{off}}}{8\pi^2} & U_{23} \\
U_{31} & U_{32} & U_{33} + \frac{b_k + B_{\text{off}}}{8\pi^2}
\end{pmatrix}
\]

**Equation 4-14 Anisotropic atomic total \( U \) matrix**

Apply inverse FT to Equation 4-13 we get the anisotropic atomic electron density
\[
\rho_{\text{aniso}}(\mathbf{r}) = (2\pi)^{-\frac{3}{2}} \sum_{k=1}^{5} \frac{a_i}{\sqrt{\det U_k^{\text{total}}}} \exp\left(-\frac{1}{2} \mathbf{r}^T (U_k^{\text{total}})^{-1} \mathbf{r}\right)
\]

Equation 4-15 Anisotropic atomic electron density

Note that the ‘core’ of the atom is still assumed to be isotropic or spherical. Angular dependence sources from the anisotropic motion. Due to symmetry of matrix \(U_k^{\text{total}}\), its inverse is readily calculated via

\[
\left[\left(U_k^{\text{total}}\right)^{-1}\right]_{ij} = \frac{\left(U_k^{\text{total}}\right)^{m}_{ij}}{\det U_k^{\text{total}}}
\]

Equation 4-16 Inverse of anisotropic atomic total \(U\) matrix

\(\left(U_k^{\text{total}}\right)^{m}_{ij}\) is the minor of matrix \(U_k^{\text{total}}\) with respect to element \(\left(U_k^{\text{total}}\right)_{ij}\). More concretely,

\[
\det U_k^{\text{total}} = \left|U_k^{\text{total}}\right| = \left(U_k^{\text{total}}\right)_{11} \left(U_k^{\text{total}}\right)_{22} \left(U_k^{\text{total}}\right)_{33} - \left(U_k^{\text{total}}\right)_{11} \left(U_k^{\text{total}}\right)_{23}^2 - \left(U_k^{\text{total}}\right)_{22} \left(U_k^{\text{total}}\right)_{13}^2 - \left(U_k^{\text{total}}\right)_{33} \left(U_k^{\text{total}}\right)_{12}^2 + 2 \left(U_k^{\text{total}}\right)_{12} \left(U_k^{\text{total}}\right)_{13} \left(U_k^{\text{total}}\right)_{23}
\]
For derivatives, let

\[ \mathbf{V}_k \equiv (U_{k}^\text{total})^{-1} \]

\[
\frac{\partial \rho_{\text{aniso}}(\mathbf{r})}{\partial x} = \sum_{k=1}^{5} \left[ -(2\pi)^{3/2} \frac{a_i}{\sqrt{\det U_k^\text{total}}} \exp\left(-\frac{1}{2} r^T \mathbf{V}_k \mathbf{r}\right) \left(x V_{k11} + y V_{k12} + z V_{k13} \right) \right]
\]

\[
\frac{\partial \rho_{\text{aniso}}(\mathbf{r})}{\partial y} = \sum_{k=1}^{5} \left[ -(2\pi)^{3/2} \frac{a_i}{\sqrt{\det U_k^\text{total}}} \exp\left(-\frac{1}{2} r^T \mathbf{V}_k \mathbf{r}\right) \left(x V_{k12} + y V_{k22} + z V_{k23} \right) \right]
\]

\[
\frac{\partial \rho_{\text{aniso}}(\mathbf{r})}{\partial z} = \sum_{k=1}^{5} \left[ -(2\pi)^{3/2} \frac{a_i}{\sqrt{\det U_k^\text{total}}} \exp\left(-\frac{1}{2} r^T \mathbf{V}_k \mathbf{r}\right) \left(x V_{k13} + y V_{k23} + z V_{k33} \right) \right]
\]

\textbf{Equation 4-17 Positional derivatives of electron density}
\[
\frac{\partial \rho_{\text{aniso}}(\mathbf{r})}{\partial (U^\text{total}_k)_{ij}} = \sum_{k=1}^{5} \left\{ -(2\pi)^{-\frac{3}{2}} \frac{a_i}{2\sqrt{\det U^\text{total}_k}} \exp \left( -\frac{1}{2} \mathbf{r}^T \mathbf{V}_k \mathbf{r} \right) \left[ \frac{\partial (\det U^\text{total}_k)}{\partial (U^\text{total}_k)_{ij}} \frac{1}{\det U^\text{total}_k} + c^i_{i,j} \right] \right\}
\]

with
\[
c^i_{i,j} \equiv x^2 \frac{\partial (V_k)_{11}}{\partial (U^\text{total}_k)_{ij}} + y^2 \frac{\partial (V_k)_{22}}{\partial (U^\text{total}_k)_{ij}} + z^2 \frac{\partial (V_k)_{33}}{\partial (U^\text{total}_k)_{ij}} + 2xy \frac{\partial (V_k)_{12}}{\partial (U^\text{total}_k)_{ij}} + 2xz \frac{\partial (V_k)_{13}}{\partial (U^\text{total}_k)_{ij}} + 2yz \frac{\partial (V_k)_{23}}{\partial (U^\text{total}_k)_{ij}}
\]

**Equation 4-18 Anisotropic U derivatives of electron density**

In this work, we use CNS to perform DCN positional refinement and \(C^\alpha\) based eNMA for anisotropic U refinement. Therefore essentially Equation 4-17 is implemented at this time while Equation 4-18 is not.

### 4.2.5. Refinement protocol

The protocol of this method involves alternative refinement between Normal Mode based temperature parameter refinement and DCN based positional refinement. Similar to original DCN protocol, for coordinates we performed torsion angle molecular dynamics with the aid of simulated annealing. The temperature was cooled from 3000K to 0K, decreasing by 50K over each six TAMD. The deformable network deforms with the same frequency. Initial relaxation was still applied by setting \(\phi\) and \(\kappa\) to 0 at the beginning and van der Waals radii reduced for enhanced sampling. Here CNS’s group B refinement methods are turned off. We supersede it with normal mode refinement and take a few lowest modes for modeling anisotropic ADP. To avoid overfitting, the number
of modes is chosen to equal the number of NMA refinement parameters to that of group B:

$$
21 + n_{\text{mode}} \frac{n_{\text{mode}} + 1}{2} = 2 \times N_{\text{residue}}
$$

$$
n_{\text{mode}} \approx \sqrt{2 \times (2 \times N_{\text{residue}} - 21)}
$$

here the number ‘21’ comes from the usage of an overall TLS refinement\(^49\) embedded with normal mode refinement\(^39\). The Shannon parameter for NMA refinement is carefully set to be consistent with the FFT_gridding_factor in CNS, in order to ensure that even in low resolution the absolute FFT grid is as fine as 1Å.

$$
\text{absolute grid} = 0.5 \times \text{high_res}/\text{Shannon\_factor} \text{ (in NM\_ref)}
$$

$$
\text{absolute grid} = \text{FFT\_gridding\_factor} \times \text{high\_res} \text{ (in DCN)}
$$

$$
\text{FFT\_gridding\_factor} = \min (0.3333, 1/\text{high\_res})
$$

$$
\Rightarrow \text{Shannon\_factor} = \max (1.5, 0.5 \times \text{high\_res})
$$

One macro-cycle in previous chapter is defined as a positional refinement with temperature declined from 3000K to 0K plus 50 steps of group B refinement. And eight such macro-cycles constitute a complete refinement task. In this work, such a macro-cycle is defined to be 50 steps of NMA refinement followed by the same annealing assisted positional refinement, while the number of macro-cycles used keeps unchanged. Refined structures with the lowest R\(_{\text{frees}}\) are kept for further analysis.
4.2.6. Coding and program

Codes that enable DCN positional refinement with the presence of anisotropic ADPs have been deployed in CNS. A relatively standalone program\textsuperscript{39} is used to calculate the eigensystem of a macromolecule’s C\textalpha trace and perform the normal mode based anisotropic ADP refinement. For proper cross-comparison with previous results, the computation was performed on identical hardware – the Shared University Grid at Rice (SUG@R) cluster running at 2.83GHz, and identical software environment such as compiler version and optimization level. Ramachandran Statistics was obtained using \textit{Molprobity} \textsuperscript{325}.

4.3. Results and analysis

We selected a total of three systems and performed four kinds of refinements on each of them: the conventional refinement (with group B), conventional plus normal mode, DCN refinement (with group B) and DCN plus normal mode. Normal mode refinement allows the possibility to manually set ‘interacting radius’ and ‘stiffness’ for better results. Here in this work, to ensure consistency we do not manipulate these parameters and fix them to 13.0 and 3.0, respectively. We show case by case how structures have been improved from both DCN and normal mode, and to what degree.

4.3.1. Case 1: PDB ID 1JL4

1JL4 possesses a total of 557 residues. Using equation in 4.2.5, we select the first 46 lowest normal modes in refinement so that two temperature refinement methods – the
group B and normal mode analysis – have nearly identical number of parameters. For a 4.3 Å structure, we set the Shannon_factor to 2.15.

4.3.1.1. The $R_{\text{free}}$ value

The most important quantity that indicates whether the structure is improved in term of a better fit with experiment is the $R_{\text{free}}$ value. We subsequently performed conventional, conventional+NMA, DCN and DCN +NMA refinements. $R_{\text{free}}$ in each case was recorded in the table below and visualized in Figure 4-2.

<table>
<thead>
<tr>
<th>1JL4</th>
<th>$R_{\text{free}}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Conventional</td>
<td>0.3700</td>
</tr>
<tr>
<td>NMA</td>
<td>0.3561</td>
</tr>
<tr>
<td>DCN</td>
<td>0.3525</td>
</tr>
<tr>
<td>DCN + NMA</td>
<td>0.3330</td>
</tr>
</tbody>
</table>

Table 4-1 DCN and NMA refinements: free R value for 1JL4

![Figure 4-2 DCN and NMA refinements: comparision of free R value for 1JL4](image)
This table gives us the following information:

1) DCN + NMA provides the best structure with a 0.0370 improvement in $R_{\text{free}}$ in comparison to conventional refinement. For DCN the improvement is 0.0175 while for NMA is 0.0139, both quite substantial.

2) For conventional task, adding NMA as the temperature refinement method leads to an improvement of 0.0139; in case of DCN, using NMA gives a decrease in $R_{\text{free}}$ by 0.0231. DCN helps NMA perform better.

3) For conventional task, adding DCN as the positional refinement method results in an improvement of 0.0175; while for NMA, employing DCN benefits refinement by 0.0195 in $R_{\text{free}}$. Existence of anisotropic profile brought by NMA gets DCN better improvement.

4) DCN delivers better result than NMA.

4.3.1.2. Overfitting effect

$R_{\text{free}} - R_{\text{work}}$ is monitored to check whether final results suffer from significant overfitting effect, as manifested by inflation of this value.

<table>
<thead>
<tr>
<th>1JL4</th>
<th>$R_{\text{free}} - R_{\text{work}}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Conventional</td>
<td>0.109</td>
</tr>
<tr>
<td>NMA</td>
<td>0.102</td>
</tr>
<tr>
<td>DCN</td>
<td>0.111</td>
</tr>
<tr>
<td>DCN + NMA</td>
<td>0.083</td>
</tr>
</tbody>
</table>

Table 4-2 DCN and NMA refinements: $R_{\text{free}} - R_{\text{work}}$ for 1JL4
In previous chapters, we showed that in general DCN reduces overfitting with a decline in $R_{\text{free}}$ minus $R_{\text{work}}$, with few exceptions (2 out of 16). 1JL4 was one of them (a slight increase of 0.002 from conventional). However, with the assistance of NMA, DCN + NMA achieves a substantially lower $R_{\text{free}} - R_{\text{work}}$ (0.026 than conventional and 0.028 than DCN alone). On the other hand, although DCN did not mitigate overfitting with conventional refinement in 1JL4, it does when NMA is used, decreasing the value from 0.102 to 0.083.

4.3.1.3. Ramachandran Statistics

As a secondary indicator, Ramachandran Statistics is a valuable indicator in that it measures the quality of the secondary structure of a refined model with no dependence on the experiment data.
From the table, we see that NMA does not improve secondary structure quite much (by 0.014 compared to conventional), while DCN, which absorbs useful information from reference structure, delivers a much higher Ramachandran Statistics (by 0.151). This behavior persists when we choose DCN and NMA as basis. For DCN, refining with NMA produces a result with higher Ramachandran Statistics (by 0.027); while for NMA, using DCN gets a much better secondary structure (by 0.164).
From the case of 1JL4, we conclude that refinement with DCN + NMA will always be the best choice, as indicated by the lowest $R_{\text{free}}$, $R_{\text{free}}-R_{\text{work}}$, and the highest Ramachandran Statistics; DCN and NMA are mutually beneficial to each other, as using one of them often boosts the performance of the other; NMA is more effective in improving the agreement between model and data than in the quality of secondary structure, and vice versa for DCN.

4.3.2. Case 2: PDB ID 2YHJ

The above analysis is repeated for 2YHJ. Similar to the reasoning, we set the number of lowest modes to 47 and Shannon_factor to 2.0.

4.3.2.1. The $R_{\text{free}}$ value

<table>
<thead>
<tr>
<th></th>
<th>$R_{\text{free}}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>2YHJ</td>
<td></td>
</tr>
<tr>
<td>Conventional</td>
<td>0.3734</td>
</tr>
<tr>
<td>NMA</td>
<td>0.3501</td>
</tr>
<tr>
<td>DCN</td>
<td>0.3442</td>
</tr>
<tr>
<td>DCN + NMA</td>
<td>0.3305</td>
</tr>
</tbody>
</table>

Table 4-4 DCN and NMA refinements: free R value for 2YHJ
We again observed substantial improvement brought by DCN+NMA. On top of conventional refinement, DCN alone drops $R_{\text{free}}$ by 0.0292, NMA does that by 0.0233, while DCN+NMA obtains a 0.0429 improvement.

For NMA refinement, adding DCN will allow an additional improvement of 0.0196. For DCN, NMA helps get $R_{\text{free}}$ lower by 0.0137. Their combination did not further boost their individual performance, but the absolute improvement is still very significant, especially given the fact that standalone DCN or NMA already found a much more accurate structures than conventional refinement and difficulties in further bringing $R_{\text{free}}$ down exponentially increase when $R_{\text{free}}$ of the target structure already resides in lower regions.
4.3.2.2. Overfitting effect

<table>
<thead>
<tr>
<th>2YHJ</th>
<th>$R_{\text{free}} - R_{\text{work}}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Conventional</td>
<td>0.095</td>
</tr>
<tr>
<td>NMA</td>
<td>0.070</td>
</tr>
<tr>
<td>DCN</td>
<td>0.082</td>
</tr>
<tr>
<td>DCN + NMA</td>
<td>0.059</td>
</tr>
</tbody>
</table>

Table 4-5 DCN and NMA refinements: $R_{\text{free}} - R_{\text{work}}$ for 2YHJ

There is a 0.036 of decline in $R_{\text{free}} - R_{\text{work}}$ for DCN+NMA refinement from conventional. NMA is responsible for majority of the improvement as it drops this value by 0.025 on top of conventional refinement and 0.023 on DCN, a very close performance. DCN on the other hand, contributes less as it delivers lower $R_{\text{free}} - R_{\text{work}}$ by 0.013 compared to conventional while 0.011 to DCN, a consistent behavior too.
4.3.2.3. Ramachandran Statistics

<table>
<thead>
<tr>
<th></th>
<th>Ramachandran</th>
</tr>
</thead>
<tbody>
<tr>
<td>Conventional</td>
<td>0.728</td>
</tr>
<tr>
<td>NMA</td>
<td>0.730</td>
</tr>
<tr>
<td>DCN</td>
<td>0.836</td>
</tr>
<tr>
<td>DCN + NMA</td>
<td>0.818</td>
</tr>
</tbody>
</table>

Table 4-6 DCN and NMA refinements: Ramachandran Statistics for 2YHJ

![Ramachandran Statistics]

Figure 4-7  DCN and NMA refinements: Comparison of Ramachandran Statistics for 2YHJ

Here we see something interesting. Although structure DCN + NMA produced did not give the highest Ramachandran Statistics, it was very close (a difference of 0.012) to the highest one by standalone DCN. In other words, when added to DCN, NMA delivered a much better structure in terms of $R_{free}$, but slightly worse in Ramachandran. Since $R_{free}$ is treated as the primary indicator, this structure is still considered to be
improved. What did the role NMA play in influencing Ramachandran with conventional refinement? We found that NMA gave a very marginal (0.002) improvement, if any, to conventional. This suggested that for the case of 2YHJ NMA is not quite beneficial for the quality of secondary structure. Coincided with the previous case, DCN seemed to contribute a lot to the secondary structure, as it utilized high quality model as a reference. It raised Ramachandran by 0.108 to conventional refinement and 0.088 to NMA. They are 54 and 44 times larger than NMA’s improvement over conventional, respectively.

4.3.3. Case 3: PDB ID 1YM7

For 1YM7 we set the number of modes to 98 and Shannon_factor 2.25. This is a case where isotropic refinement seems to be ineffective. As we can see from Table 3-2, conventional refinement delivered an $R_{\text{free}}$ of 0.2764 and DEN gave 0.2739. DCN was able to provide an additional 64% improvement, bringing it to 0.2723; however, the absolute improvement in $R_{\text{free}}$ was limited to the third and fourth decimals, and under isotropic refinement scheme a fairly optimal structure was already achieved. We show that how anisotropy of Debye-Waller factor that Normal Mode Analysis introduces leads to breakthroughs in this case.

4.3.3.1. The $R_{\text{free}}$ value

<table>
<thead>
<tr>
<th>1YM7</th>
<th>$R_{\text{free}}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Conventional</td>
<td>0.2764</td>
</tr>
<tr>
<td>NMA</td>
<td>0.2670</td>
</tr>
<tr>
<td>DCN</td>
<td>0.2723</td>
</tr>
<tr>
<td>DCN + NMA</td>
<td>0.2562</td>
</tr>
</tbody>
</table>

Table 4-7 DCN and NMA refinements: free R value for 1YM7
For 1YM7, the maximum improvement isotropic refinement obtained was 0.0041. NMA allowed anisotropic refinement with the same number of temperature parameters as isotropic group B, but resulted in a structure with an $R_{\text{free}}$ under 0.27, more precisely, a 0.0094 gain. That was more than 2 times better than isotropic refinement. In addition, DCN + NMA touched a new level of $R_{\text{free}}$ that neither standalone NMA nor DCN could ever reach. The final refined structure had an $R_{\text{free}}$ as low as 0.2562 – that was a 0.02+ improvement from conventional refinement, roughly 5 times of DCN’s performance, and also 2 times of NMA’s.
4.3.3.2. Overfitting effect

<table>
<thead>
<tr>
<th>Method</th>
<th>$R_{\text{free}} - R_{\text{work}}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Conventional</td>
<td>0.030</td>
</tr>
<tr>
<td>NMA</td>
<td>0.034</td>
</tr>
<tr>
<td>DCN</td>
<td>0.033</td>
</tr>
<tr>
<td>DCN + NMA</td>
<td>0.027</td>
</tr>
</tbody>
</table>

Table 4-8 DCN and NMA refinements: $R_{\text{free}} - R_{\text{work}}$ for 1YM7

This is the other exception (2 out of 16) that $R_{\text{free}} - R_{\text{work}}$ for DCN was slightly higher than conventional. Replacing group B with NMA did not mitigate the overfitting effect, rendering a value that is 0.001 higher than DCN. As what was demonstrated in $R_{\text{free}}$, the unification of DCN and NMA successfully produced favorable structure with
minimal overfitting – an $R_{\text{free}} - R_{\text{work}}$ as small as 0.027 – better than all three other refinement approaches.

4.3.3.3. Ramachandran Statistics

### Table 4-9  DCN and NMA refinements: Ramachandran Statistics for 1YM7

<table>
<thead>
<tr>
<th>Method</th>
<th>Ramachandran</th>
</tr>
</thead>
<tbody>
<tr>
<td>Conventional</td>
<td>0.703</td>
</tr>
<tr>
<td>NMA</td>
<td>0.764</td>
</tr>
<tr>
<td>DCN</td>
<td>0.751</td>
</tr>
<tr>
<td>DCN + NMA</td>
<td>0.829</td>
</tr>
</tbody>
</table>

**Figure 4-10  DCN and NMA refinements: Comparison of Ramachandran Statistics for 1YM7**

In this case, both DCN and NMA delivered a model with considerably better secondary structure than conventional refinement did. Generally speaking, because DCN took reference structure into refinement process, it was expected to perform better in
improving Ramachandran Statistics than NMA did. However, the ineffectiveness of isotropic refinement exhibited on case 1YM7 suggested otherwise, which is true. DCN increased Ramachandran from 0.703 to 0.751, while NMA lifted it even higher to 0.764. Consistent with $R_{\text{free}}$ and overfitting, DCN + NMA outperformed all three other refinements, with the highest Ramachandran of 0.829. The margin of 0.126 was even larger than the combination of what standalone DCN and NMA (0.048 and 0.061, respectively) provided.

Case 3 shows that in circumstances where isotropic positional refinement seems to converge to an optimal solution, bringing anisotropy in and refining with DCN + NMA could obtain substantially better results.

**4.4. Conclusion and discussion**

Here we proposed a new method that first unified advanced and unconventional refinement tools for both positional and temperature factor parameters. For positional refinement, we utilized Deformable Complex Network whose performance was throughout demonstrated in Chapter 3. In original DCN, the approach for refining temperature parameters was limited to grouped B – an isotropic scheme that was employed due to small number of diffraction entries to prevent structure from being overfitted. On the other hand, Normal Mode Analysis was an effective tool for modeling atomic anisotropic displacement, with successful application of macromolecular refinement on several systems\textsuperscript{39,46}. However, positional refinements used in those applications were limited to conventional method, and in the middle of that process there involved intensive manual adjustments – an effort tried to by hand place outlier atoms –
those falling outside of the calculated electron density map – back to it. Therefore a map needs computing explicitly at the end of every macro-cycle. It is worth noting that, in the traditional normal mode analysis, lower frequency mode is considered to vibrate with larger amplitude, given the assumption of the vibrational energy being proportional to $kT$ (Boltzmann constant multiplied by the temperature). Here, when NMA is applied to the area of refinement, eigenvectors of different modes undergoes a linearly combination before being able to describe the system’s dynamics. In this sense, amplitudes of different modes are essentially magnified by a factor of various “coefficients” in this linear expression, making each mode’s contribution to the overall fluctuation also dependent on the refined normal mode variables. Nevertheless, low frequency modes still possess a long-range profile while higher ones tend to be more localized. It is appropriate to use low frequency modes to capture the characteristics of the system’s fluctuation as a whole.

By simultaneous use of DCN and NMA, the potential for extra improvements in low-resolution X-ray macromolecular crystallography is unlocked. From out tests, it is shown that these two methods are beneficial to each other. Their individual effects in improvement could accumulate; sometimes, one even gets enlarged with the existence of the other. One reason for that is, for DCN, the introduction of anisotropy of atom displacement by NMA opens up new conformation space compared to an isotropic profile with fixed diffraction data. This anisotropy, however, could also come from a full atom NMA (rather than eNMA) that can better handle systems with ligands and heterogeneous atoms, or from TLS$^{49}$ – another anisotropic refinement method that is widely used. While for NMA, DCN incorporates geometry information from reference structures and ensures better coordinates. This allows a more favorable coarse grained
backbone and provides better eigenvector profiles from normal mode analysis. Those eigenvectors become more effective in anisotropic refinement with the delivery of accurate anisotropic U matrices, and a set of calculated structure factors that were not previously obtained from only isotropic B factors come in. Under their influence, positional refinement in real space is in turn enhanced, and improved structures are eventually found as manifested by a better agreement with diffraction data, a minimized overfitting effect, and a higher quality secondary structure.
References

1 Figures obtained from Wikimedia Commons, a freely licensed media file repository. commons.wikimedia.org.


http://www.csb.yale.edu/userguides/datamanip/xplor/xplorman/node183.html.

*Xplor Online Manual.*


38 The PyMOL Molecular Graphics System, Version 1.2r3pre, Schrödinger, LLC.


