Structural Bioinformatics and Crystallography Tools for Automated Protein Model Building and Validation

Dissertation with the aim of achieving a doctoral degree at the

Faculty of Mathematics, Informatics and Natural Sciences

Department of Chemistry of the University of Hamburg

Submitted by

JOANA PEREIRA

European Molecular Biology Laboratory, Hamburg Outstation

Hamburg

2016

To my boyfriend and my family

The work presented in this thesis was carried out between September 2012 and September 2016 under external supervision of

Dr. Victor S. Lamzin

at the European Molecular Biology Laboratory (EMBL) Hamburg Outstation. University support was provided by

Prof. Dr. Andrew E. Torda

from the research group for Biomolecular Modeling of the Centre for Bioinformatics from the Faculty of Mathematics, Informatics and Natural Sciences at the University of Hamburg.

The members of the Thesis Advisory Committee (TAC), mandatory by EMBL, in addition to Dr. Victor S. Lamzin and Prof. Dr. Andrew E. Torda, also included Dr. Thomas Schneider (Macromolecular Crystallography, EMBL Hamburg), Prof. Dr. Gerard Kleywegt (Structural Validation of Proteins, EMBL-EBI Hinxton) and Prof. Dr. Inari Kursula (Macromolecular Crystallography, CSSB-HZI Hamburg).

1. Evaluator: Prof. Dr. Andrew E. Torda

2. Evaluator: Prof. Dr. Henning Tidow

Chair of examination commission: Prof. Dr. Andrew E. Torda

Date of disputation: 09.12.2016

Approved for publication on: 13.12.2016

Summary

Deciphering the structures of biological macromolecules is essential to understand their function in cellular processes and their role in human diseases. Macromolecular X-ray crystallography is the most successful and widely used technique for such studies. However, large macromolecules and their assemblies usually do not provide crystallographic data at an atomic level of detail and the resulting electron density maps are insufficiently informative. This makes their automatic interpretation more difficult and less accurate. While methods for the extension of fragmented protein models have been developed in the past, more complete models are not necessarily more correct. It is therefore imperative to validate crystallographic models not only before depositing them to a databank but also during the model-building procedure.

In this thesis, this issue is addressed at two levels. The main one concerns the development of a validation tool that is not only useful at the later stages of the crystallographic experiment but also during protein model building. The most commonly used methods for the validation of protein backbone conformation are based on the two-dimensional distribution of its dihedral angles. Based on the premise that molecular conformation can be defined by the relative position of atoms in the three-dimensional space and by the chirality of asymmetric atomic groups, a three-dimensional space, DipSpace, was developed which allows the description of protein stereochemistry and highlights residues in an unusual conformation that may not be detectable with other approaches. It was implemented within a tool, DipCheck, which can be used for the general validation of protein models but is also used by ARP/wARP during automated protein model building. DipCheck evaluates any protein model, pointing to problematic residues and providing an overall score of protein backbone quality. Following a modification of the ARP/wARP protein model building protocol, the quality of protein models built at a resolution between 2.5 and 3.0 Å by ARP/wARP is now improved.

The second level concerns the study and the implementation of changes to the ARP/wARP protocol for protein automated model building at medium-to-low resolution, including the geometrisation of identified dipeptide units and the application of density shape descriptors for the identification of side-chains, but also the improvement in the automated building of bound ligands. In this last case, the inclusion of an energetic term during the ranking of possible ligands for a given binding site proved helpful for the validation of already deposited protein-ligand complexes and also for the correct identification of the ligand in a given density cluster.

Zusammenfassung

Die Kenntnis der Strukturen biologischer Makromoleküle ist von zentraler Bedeutung für das Verständnis ihrer Funktion in zellulären Prozessen und ihrer Rolle bei Erkrankungen des Menschen. Makromolekulare Röntgenkristallographie ist die erfolgreichste und am weitesten verbreitete Technik für solche Studien. Allerdings können insbesondere bei großen Makromolekülen und deren Komplexen oft keine hinreichend hoch aufgelösten kristallographischen Daten gemessen werden, und die daraus resultierenden Elektronendichtekarten sind nicht ausreichend informativ. Dies macht ihre automatische Interpretation schwieriger und weniger präzise. Während die in der Vergangenheit entwickelten Verfahren zur Erweiterung von fragmentierten Proteinmodellen zwar vermeintlich vollständigere Modelle ergeben, sind diese nicht notwendigerweise richtiger. Es ist daher zwingend notwendig, kristallographische Modelle nicht nur dann zu validieren, wenn diese in Datenbanken veröffentlicht werden, sondern bereits während das Model in die Elektronendichtekarten gebaut wird.

Die vorliegende Arbeit befasst sich mit diesem Problem auf zwei Ebenen. Die erste und wichtigere Ebene betrifft die Entwicklung eines Validierungswerkzeuges, das sowohl in den späteren Phasen des kristallographischen Experimentes, als auch während der Proteinmodellierung nützlich ist. Die am häufigsten verwendete Methode zur Validierung der Konformation des Protein-Rückgrats basiert auf der zweidimensionalen Verteilung der Torsionswinkel. Unter der Annahme, dass die molekulare Konformation durch die relative Position der Atome im dreidimensionalen Raum und die Chiralität von asymmetrischen Atomgruppen definiert ist, wurde DipSpace entwickelt. DipSpace ist ein dreidimensionaler Raum, der die Beschreibung der Protein-Stereochemie ermöglicht, wobei Aminosäuren in ungewöhnlicher Konformation auffallen, welche mit anderen Methoden nicht erkannt werden können. Das DipSpace-Konzept wurde als Software-Werkzeug implementiert – genannt DipCheck - welches für die allgemeine Validierung von Proteinmodellen verwendet werden kann, aber auch von ARP/wARP während der automatisierten Proteinmodellierung aufgerufen wird. Durch die entsprechende Anpassung von ARP/wARP, konnte die Qualität von Proteinmodellen mit einer Auflösung zwischen 2,5 und 3,0 Å deutlich verbessert werden.

Die zweite Ebene betrifft die Analyse und die Umsetzung von Änderungen an verschiedenen ARP/wARP Protokollen zum automatisierten Modellbau für Proteine bei mittlerer bis niedriger Auflösung. Dies schließt die Geometrisierung der identifizierten Dipeptideinheiten, die Anwendung von Deskriptoren für die dreidimensionale Form der Dichteverteilung zur

Identifizierung von Seitenketten, sowie die Verbesserung der automatisierten Modellierung von gebundenen Liganden ein.

Für den letzten Fall wurde die Bindungsenergie zwischen Protein und Ligand für die Bewertung von möglichen Liganden für eine bestimmte Bindungsstelle als zusätzlicher Term eingeführt, was sich als hilfreich für die Validierung bereits bekannter Protein-Ligand-Komplexe erwiesen hat aber auch die korrekte Identifizierung eines Liganden in einem Dichte-Cluster unterstützt.

Preface

Parts of this thesis (text and figures) have been (or will be) submitted to peer-reviewed journals and have been presented as posters and oral presentations at conferences and workshops.

Peer-Reviewed Publications

Pereira J, Lamzin V.S. A distance geometry-based description of protein main-chain conformational space. *Scientific Reports* (submitted)

Pereira J, Beshnova D, Chojnowski G, Oezugurel U, Heuser P, Lamzin V.S. ARP/wARP: helping crystallographers automatically build a macromolecular model since the early 90's. *Acta Crystallographica D* (in preparation)

Beshnova D, **Pereira J**, Lamzin V.S. Estimated energy of ligand binding for model building and validation. *Acta Crystallographica D* (submitted)

Oral Presentations at Conferences and Meetings

"DipSpace and DipCheck: a distance geometry-based description of protein main-chain conformation", ECM30: 30th European Crystallography Meeting, Basel, Switzerland, August 2016

"DipSpace and DipCheck: a distance geometry-based description of protein main-chain conformation", EMBL LabDay, Heidelberg, Germany, July 2016

"Automatic small molecule identification and ligand building with ARP/wARP", CCP4 Study Weekend, Nottingham, United Kingdom, January 2016

"New protein main-chain conformational descriptors on the validation and improvement of automatic protein model building", ECM29: 29th European Crystallography Meeting, Rovinj, Croatia, August 2015

"ARP/wARP", Software Fayre, 23rd Congress and General Assembly of the International Union of Crystallography, Montreal, Canada, August 2014

Oral presentations at Workshops

"EMBO Practical Course: Computational Structural Biology - From Data to Structure to Function", European Molecular Biology Laboratory, Hamburg, Germany, April 2013

"EMBO Practical Course: Computational Structural Biology - From Data to Structure to Function", European Bioinformatics Institute, Hinxton, United Kingdom, April 2014

"DLS-CCP4 Data Collection and Analysis Workshop", Diamond Light Source, Didcot, United Kingdom, December 2014-2015

"CCP4/APS Summer School in Macromolecular Crystallography: From Data Collection to Structure Refinement and Beyond", Argonne National Laboratory, Chicago, United States of America, June 2013-2015

"CCP4 Crystallography School and Workshop: From data processing to structure refinement and beyond", The Photon Factory, Tsukuba, Japan, November 2014

"CCP4/OIST School, Okinawa Institute of Science and Technology", Okinawa, Japan, November 2013 and 2015

"2nd Australian Advanced Methods in Crystallography Workshop", Australian Synchrotron, Clayton, Australia, June 2014

"Macromolecular Crystallography School: From data processing to structure refinement and beyond, Institut Pasteur, Montevideo, Uruguay, April 2015

"Macromolecular Crystallography School: From data processing to structure refinement and beyond, Instituto de Física de São Carlos, São Paulo, Brazil, April 2016

Poster Presentations at Conferences and Meetings

Pereira J, Lamzin V. New protein main-chain conformational descriptors on the validation and improvement of automatic protein model building. 3DSIG, Dublin, Ireland, July 2015

Pereira J, Wiegels T, Lamzin V. ValiFrag: are these proper folds? Evaluating fragment quality during automated protein model building. 23rd Congress and General Assembly of the International Union of Crystallography, Montreal, Canada, August 2014

Wiegels T, **Pereira J**, Vancea I, Carolan C, Beshnova D, Heuser P, Lamzin V. ARP/wARP for crystallographic model building and drug discovery. 23rd Congress and General Assembly of the International Union of Crystallography, Montreal, Canada, August 2014

Pereira J, Wiegels T, Lamzin V. On the Validation of Fragments During Automated Protein Model Building. EMBL PhD Retreat, Como, Italy, September 2013

Pereira J, Wiegels T, Lamzin V. On the Validation of Fragments During Automated Protein Model Building. 11th International Conference on Biology and Synchrotron Radiation (BSR), Hamburg, Germany, September 2013

Pereira J, Wiegels T, Lamzin V. ValiFrag: Are These Proper Folds? Fragment Validation During Protein Automated Model Building. ECM28: 28th European Crystallography Meeting, Warwick University, England, United Kingdom, August 2013

Contents

Abbreviations and Notation

General and crystallographic abbreviations

2D	Two-Dimensional
3D	Three-Dimensional
ACMI	Automatic Crystallographic Map Interpreter
ADP	Atomic Displacement Parameter
CCD	Charge-Coupled Device
CCP4	Collaborative Computational Project, Number 4
CRL	Compound Refractive Lenses
CSD	Cambridge Structural Database
DipCheck	Dipeptide Unit Check
DipScore	Dipeptide Unit Score
DipSpace	Dipeptide Unit Space
EDS	Uppsala Electron Density Server
EM	Electron Microscopy
EMBL	European Molecular Biology Laboratory
IR	Isomorphous Replacement
MAD	Multi-wavelength Anomalous Dispersion
MR	Molecular Replacement
MX	Macromolecular X-Ray Crystallography
NCS	Non-Crystallographic Symmetry
NMR	Nuclear Magnetic Resonance
pc	Principal Component
PCA	Principal Component Analysis
PDB	Protein Data Bank
PDBe	Protein Data Bank in Europe
RSCC	Real Space Correlation Coefficient
RSR	Real Space R-factor
SAD	Single wavelength Anomalous Dispersion
SAXS	Small Angle X-Ray Scattering
SecStr	Secondary Structure
VDW	Van der Waals
χ_{score}	Chi-score

Biochemistry

A3P	Adenosine-3’,5’-diphosphate
ACBP	Armadillo acyl-CoA-Binding Protein

ADP	Adenosine diphosphate
ATP	Adenosine triphosphate
BTN	Biotin
CMP	Adenosine-3',5'-cyclic-monophosphate
Cα	Amino acid alpha carbon
DNA	Deoxyribonucleic acid
HEPES	4-(2-hydroxyethyl)-1-piperazineethanesulfonic acid
MES	2-(N-morpholino)-ethanesulfonic acid
P_{II}	Polyproline II spirals
RNA	Ribonucleic acid
THP	Thymidine-3',5'-diphosphate
θ_i	Protein backbone (Cα_{i-1}-Cα_i-Cα_{i+1}) stretching angle
κ	Protein backbone (Cα_{i+2}-Cα_{i-1}-Cα_{i+1}) angle
φ	Protein backbone (C_{i-1}-N_i-Cα_i-C_i) dihedral angle
φ_d	Protein backbone (Cα_{i-1}-O_{i-1}-Cα_i-O_i) dihedral angle
ψ	Protein backbone (N_i-Cα_i-C_i-N_{i+1}) dihedral angle
ψ_d	Protein backbone (O_{i-1}-Cα_i-O_i-Cα_{i+1}) dihedral angle
ω	Protein backbone (Cα_i-C_i-N_{i+1}-C_{i+1}) dihedral angle
τ	Protein backbone (N_i-Cα_i-C_i) stretching or (Cα_{i-1}-Cα_i-Cα_{i+1}-Cα_{i+2}) dihedral angles
τ_d	Protein backbone (O_{i-1}-Cα_i-O_i) stretching angle

Crystallography

a_{hkl}	Phase angle at location (*h,k,l*) in reciprocal space
F_{calc}	Calculated structure factor
F_{obs}	Observed structure factor
F_{hkl}	Structure factor at location (*h,k,l*) in reciprocal space
I	Intensity of a diffracted reflection
R	Crystallographic R-factor
R_{free}	Crystallographic Free R-factor
λ	Radiation wavelength
ρ_{calc}	Calculated electron density
ρ_{obs}	Observed electron density
ρ_{xyz}	Electron density at location (*x,y,z*) in real space

Mathematics

CDF	Cumulative density function
CI	Chiral invariant
cov(·)	Sample covariance
G_O	Universal chiral index
HDR	Highest Density Region
H_0	Null hypothesis
H_1	Alternative hypothesis
I	Identity matrix
M_{ij}	Matrix element in row *i* and column *j*

MAD	Median Absolute Deviation
r.m.s.d.	Root-mean-square deviation
PDF	Probability density function
r	Linear correlation coefficient
R_g	Radius of gyration
SD	Sample standard deviation
$tr(A)$	Trace of matrix A
$\mathrm{var}(\cdot)$	Sample variance
V_C	Chiral volume
V_O	Oriented volume
$\bar{x}$	Average value
Å	Angstrom
γ_1	Population skewness
γ_2	Population kurtosis
λ	Eigenvalue
μ	Population mean
$\rho(A)$	Spectral radius of matrix A
σ	Population standard deviation
σ^2	Population variance
υ	Eigenvector
$\chi(G)$	Chromatic number of graph G
ν	Mixing proportion

Chapter 1

Introduction

> "Almost all aspects of life are engineered at the molecular level, and without understanding molecules we can only have a very sketchy understanding of life itself."
>
> — Francis Crick

Proteins are essential components of life [1]. Over more than 3.5 billion years, they evolved to the complex machines observed today [2], with their three-dimensional structure defining their function and building up most of the "cell machines" [3]. Their structural characterisation is important for understanding the molecular mechanisms that underlie biological function and evolution, which can be further used in biotechnology, pharmaceutical industry and medicine. In this chapter, the basis of protein three-dimensional structure and the methods to obtain structural information are outlined. Special focus is given to macromolecular X–ray crystallography (MX) and to the ARP/wARP software project (http://www.arp-warp.org), which provides an automated interpretation of experimental electron density maps derived from MX data and, in turn, an automated building of macromolecular structures.

1.1. The Three-Dimensional Structure of Proteins

Proteins are linear polymers of amino acids (Figure 1a) and, out of several hundred, there 20 common natural amino acids that are used for their construction [4], [5]. These have a common central alpha-carbon atom (Cα), an amino group (-NH_2), a carboxyl group (-COOH), and a side-chain (-R). These four groups are chemically distinct in all amino acids except glycine. Therefore, all non-glycine amino acids are chiral molecules that can exist in two different isomeric forms with different 'hands' (the L- and D-forms; exemplified in Figure 17) [4], [5]. Protein chains are formed by consecutive condensation reactions between amino acids (Figure 1a). Within a living cell, this process is governed by the ribosome, which has adapted to use only the L-form of the amino acids for protein synthesis [6].

The link between the carboxyl group of one amino acid and the amino group of another, the peptide bond, has a partial-double bond nature due to electronic resonance [7]. Therefore, there is hardly any rotation around it and can be found in two distinct isomerisation states: cis- or trans- (Figure 2). In the trans- form, the two Cα atoms are approximately 3.8 Å apart, while in the cis-form they are closer to each other (about 3.0 Å). Due to the close proximity between the groups attached to the Cα atoms (Figure 2b), the cis-form is energetically less favourable than the trans-configuration and occurs rarely, mainly preceding a proline [8]–[10]. Therefore, more than 99% of the peptide bonds are in the trans-configuration [8], [9], [11].

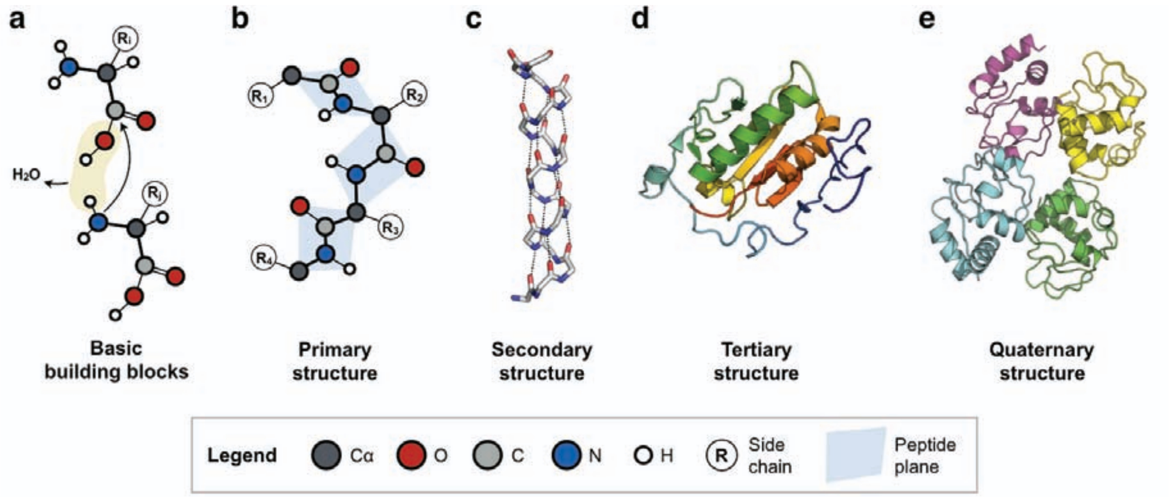


Figure 1 The different levels of protein structure. (a) The basic building blocks of proteins are amino acids. To form a protein chain, the amino acids condense with each other by forming a peptide bond. (b) The successive sequence of amino acids is the primary structure. (c) The local folding of small stretches forms the secondary structure and (c) the global orientation of these secondary structural elements makes the tertiary structure. (e) The arrangement of multiple protein polypeptide chains in a multi-subunit complex forms the quaternary structure.

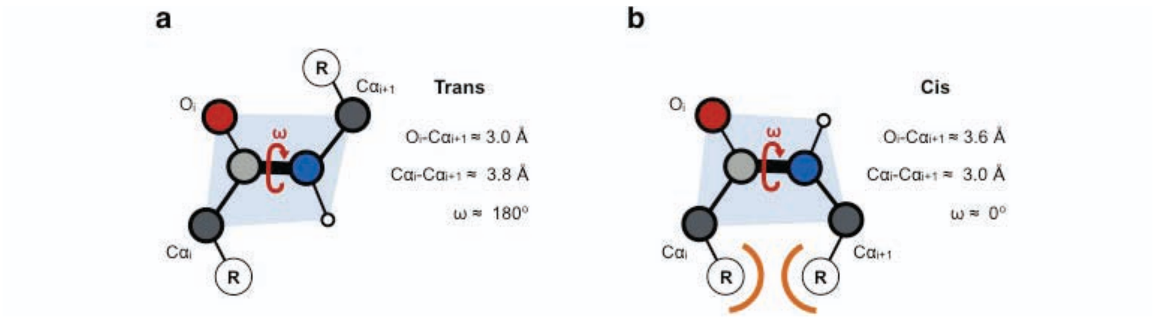

a
b
Trans
R
$C\alpha_{i+1}$
O_i
ω
O_i-$C\alpha_{i+1}$ ≈ 3.0 Å
$C\alpha_i$-$C\alpha_{i+1}$ ≈ 3.8 Å
ω ≈ 180°
$C\alpha_i$
R
Cis
O_i
ω
O_i-$C\alpha_{i+1}$ ≈ 3.6 Å
$C\alpha_i$-$C\alpha_{i+1}$ ≈ 3.0 Å
ω ≈ 0°
$C\alpha_i$
$C\alpha_{i+1}$
R
R

Figure 2 The peptide bond. (a) Trans-peptide bond/unit. (b) Cis-peptide bond/unit.

Since different amino acids have side-chains with different chemical properties, the sequence by which they are connected defines the *protein's primary structure* (Figure 1b). Geometrically, protein polypeptide chains can be represented by a sequence of planes, the peptide units, defined by the atoms involved in the formation of the peptide bond (Figure 1b and Figure 2). The interaction between main-chain and side-chain atoms allows the polypeptide chain to fold and to form regular structures. This makes the *protein secondary structure* (Figure 1c), the general three-dimensional form of local segments of the chain, formed by the establishment of hydrogen bonds between amine hydrogen and carbonyl oxygen atoms of adjacent residues in the protein main-chain.

There are two main types of secondary structural elements that can be observed in proteins: helices (Figure 1c) [7] and β-strands [12]. In the first, the protein main-chain follows a helical path (Figure 1c) [7]. The most common are α-helices, with a periodicity of 3.6 residues per turn [7], [13]. Variations are denoted as π- (4.1 residues per turn) [14], [15] and 3_{10}-helices (3 residues per turn) [16]. Only in α-helices the main-chain atoms are ideally packed to provide a stable structure, in π- and 3_{10}-helices they are packed either too loosely or too tightly. Residues fold into 3_{10}-helices in about 4% of the cases and typically are found at the N- or C-termini of α-helices [17]. π-Helices are rare [18] and usually appear as a result of an insertion of one residue in an α-helix [15]. On the other hand, β-Strands form an almost fully extended conformation with the side-chains of adjacent residues pointing to opposite directions [12]. They interact with each other in a parallel or an anti-parallel manner forming a β-sheet, through an extensive hydrogen bond network [12], [19].

Most protein structures are built up from the combination of secondary structural elements connected by loop regions of various lengths and shapes. Examples are the β-hairpins and turns, which are often three to five residues long and connect two adjacent anti-parallel secondary structural elements [20]. Longer loops are in general long coiled segments without any specific hydrogen bonding network [21]. Secondary structural elements arrange themselves in simple motifs by packing side-chains from adjacent helices and strands close to each other. Several of these motifs may be combined to form a compact three-dimensional structure (the *tertiary structure*; Figure 1d) intimately related to protein function [22]–[25]. Some proteins contain two or more polypeptide chains, which can be identical or different, forming a functional multi-subunit complex. This arrangement constitutes the fourth level of protein structure, the *quaternary structure* (Figure 1e). The association of several protein chains can serve various proposes: it allows the stabilisation of proteins that are not functional alone, the regulation of several protein functions as well as building up large complexes, for example virus capsids, transmembrane channels, or large macromolecular machines [1].

The fold space, occupied by all possible protein folds, is vast [24]–[26] and allows the classification of proteins into different classes [22], [23]. The sequence space is much larger [27], so that different sequences can adopt similar three-dimensional structures. At the same time, there are also many sequences for which no structure has yet been determined or correspond to disordered proteins [28]–[30].

1.2. Description of Protein Main-Chain Conformation

The mathematical description of protein main-chain conformation allows the understanding and classification of the overall space occupied by the atoms in the protein backbone, providing means for the validation and modelling of protein structures. The first mathematical representation of protein backbone was developed by Linus Pauling in 1951 [7] as a set of well-defined interatomic distances and angles. In 1953, Francis Crick [31] derived three equations for the computation of the Cartesian coordinates of the Cα atoms of proteins in a coiled-coil structure. Ten years later, Ramachandran and colleagues proposed the use of two main-chain torsion angles - φ (C_{i-1}-N_i-$C\alpha_i$-C_i) and ψ (N_i-$C\alpha_i$-C_i-N_{i+1}) [32]. The concept behind the Ramachandran plot and the joint use of torsion angles became the basis for the development of further tools for the description of protein conformational space (Figure 3).

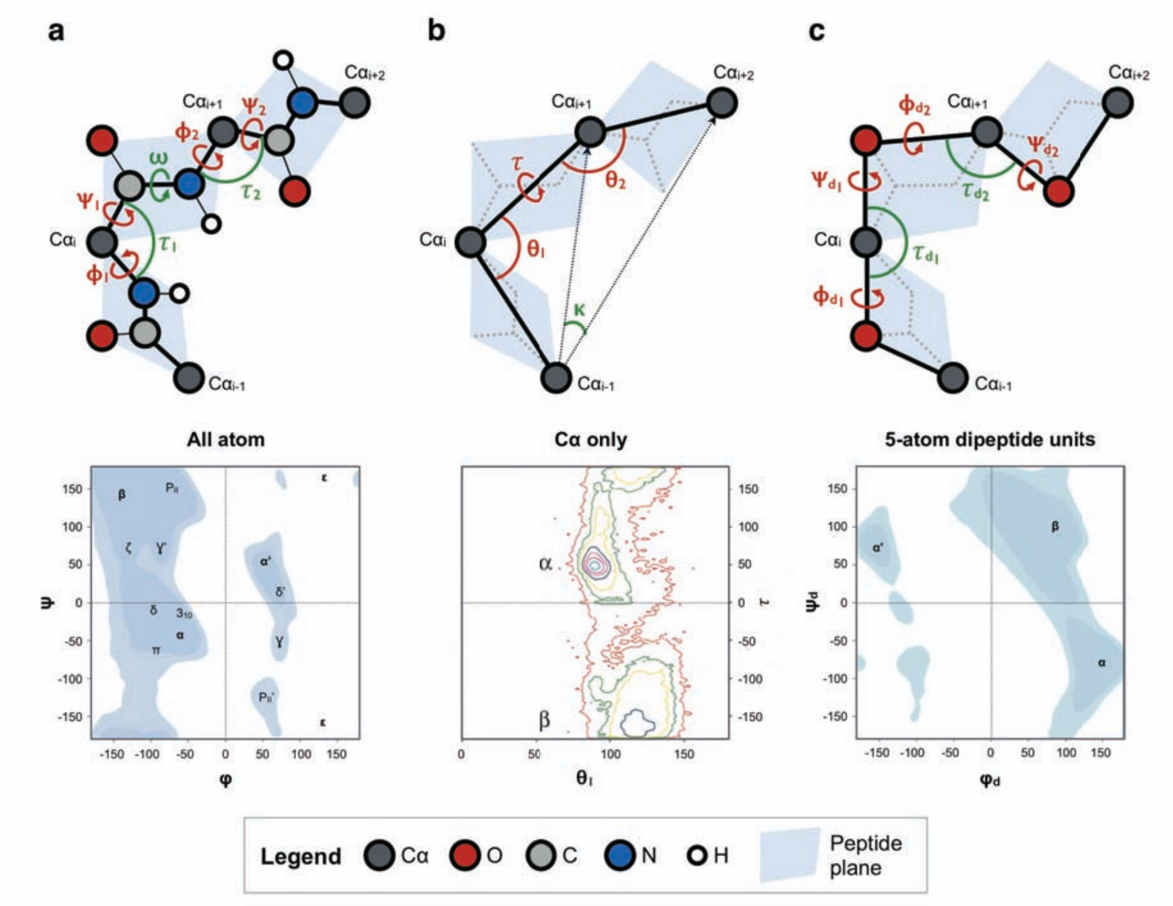


Figure 3 Representations of the protein polypeptide main-chain. (a) All-atom representation, described by the two Ramachandran φ and ψ angles (in red); the ω torsion and τ stretching angles are also shown (in green). (b) Cα representation described by one pseudo-torsion and two pseudo-stretching angles (in red). The joint distribution of the θ and τ angles was adapted from Kleywegt 1997 [33]. (c) 5-atom (double-plane) representation, described by the two Ramachandran-like φ_d and ψ_d dihedral angles (in red) and the τ_d stretching angle (in green). The general Ramachandran plot as well as the Ramachandran-like plot were calculated for the set of structures used in this thesis, with two different shades of blue representing the allowed (lighter) and favoured (darker) regions as defined by Lovell *et al.* [34]. The proposed nomenclature [35] for different Ramachandran regions are shown.

1.2.1. All-Atom Representation and The Ramachandran Angles

The Ramachandran approach (Figure 3a) is based on the assumption that while there are four covalent bonds in the protein main-chain, only two are relevant for its stereochemistry. The carbonyl C=O double bond cannot affect the conformation of the chain and the C-N bond can define only two different conformers (the cis- and the trans-forms; Figure 2). However, the rotation around the single bonds N-Cα and Cα-C changes protein backbone conformation, as described by the φ and ψ dihedral angles (Figure 3a). Ramachandran and colleagues explored the energetic landscape associated with these rotations and showed that only about one quarter of the φ/ψ space is energetically favoured [32].

The two-dimensional distribution of these two dihedral angles is denoted the *Ramachandran plot* and is usually divided into several areas: 'favoured', 'allowed', 'generously allowed' and 'disallowed' (Figure 3a). It was originally drawn only for amino acids other than glycine and proline: glycine (due to the absence of the side-chain) is much more flexible while proline (due to the presence of the heterocyclic ring) is too rigid. Therefore, the favoured stereochemistry of glycine and proline differ from that of the other 18 L-amino acids [32]. With the rapid increase of protein structures in the Protein Data Bank (PDB) [36] (about 113,000 models as of September 2016), the details of the Ramachandran plot have been refined, using for example the information derived from crystal structures determined at very high resolution [37], and several software approaches to compute these dihedral angles have been developed (e.g. Procheck [38] and MolProbity [39]). The Ramachandran plot has thus become one of the most important main-chain quality indicators for a protein model [34], [37], [40].

Procheck [38] regions of the Ramachandran plot were developed in 1992 [41] for a set of 121,870 residues from 463 protein structures based on the calculation of the number of residues in blocks ($10^o \times 10^o$ areas) of the conformational space. Procheck defines four main regions: the 'core' region which includes all blocks with more than 100 residues, the 'allowed' regions with more than eight residues, the 'generous' regions expanded out by 20^o all around the allowed region, and the 'outside' region as the space left. MolProbity [39] boundaries were defined a decade later by Lovell *et al.* [34] for a set of about 100 000 residues from 500 structures solved at resolution better than 1.8 Å. The 'favoured' region contains 98% of the data, the 'allowed' 99.95% and the 'disallowed' the remaining 0.05%.

According to the frequency of residues in protein structures with a given φ/ψ combination, several structural regions of the Ramachandran plot have been defined. The β-region occupies a large fraction of the -φ/+ψ quadrant (Figure 3a) and is divided into two separate zones: the β-area (occupied by residues that are found in β-strands) and the P_{II}-regions, occupied by residues that form polyproline II spirals, characterised by the absence of hydrogen bonds between the N-H residue of a residue and the C=O group of one of the following residues [35], [42]. From this region, two “peninsulas” are observed: the γ-region is relatively low populated but comprises residues in γ-turns (which have an O_i-NH_{i+2} hydrogen bond) [43] while the ζ-region is dominated by residues preceding prolines [44]. Right-handed α-helices cluster in a very narrow region around (-63,-43), with 3_{10}- and π-helices in a close vicinity [35] (Figure 3a). An area close to the α-region is referred as the δ-region or the “bridge sector” as it bridges α- and β-regions. It comprises residues found in a broad variety of turns and is characterised by an $NH_{i+1} \rightarrow N_i$ π-interaction and an opening up of the bond angle τ (N_i-$C\alpha_i$-C_i) [44]. At positive φ and extreme ψ (180 and -180) there is the ε-region, which is sparely populated mostly by glycines [35]

In addition to these regions, there are their mirror images, marked by adding a prime (Figure 3a). Any pair of points on the Ramachandran plot related by a point inversion around (0,0) (i.e, (φ,ψ) and (-φ,-ψ)) are mirror-imaged conformations. Given the preference for L-amino acids, protein backbone usually prefers negative values of the φ angle (Figure 3a). With the exception of the γ’-regions, which is more populated than the γ-region, all of the other mirror-image zones are little populated [35]. With the increase of protein structural information in the PDB it became possible to relate the protein backbone conformation to the conformation of the side-chains [45], [46] and hence to the protein sequence [44], [47]–[49] so that a specific Ramachandran plot can be computed for some residue types [35]. For example, residues preceding proline often populate the ζ-region while prolines themselves have a preference for the P_{II} and the α-helical regions. Glycines can acquire many more conformations, including the ε-regions [35].

Although the Ramachandran plot is one of the most important tools for the description and validation of protein backbone stereochemistry, it is a simplification of a multi-dimensional dependence. For example, the dependence of the φ/ψ angles on the bond angle τ (N_i-$C\alpha_i$-C_i) (Figure 3a) is not accounted for. For an *sp^3* C atom with a perfectly tetrahedral coordination the value of this angle should be 109.5^o [37], in proteins it generally ranges from 107.5^o to 114.0^o [50] and averages at about 110^o [37]. Similarly, the planarity of the peptide bond can be described by the value of the ω angle (Figure 2 and Figure 3). In the trans-configuration, this angle acquires a value close to 180^o while in the cis-configuration its value is about 0^o. However, surveys over known protein three-dimensional models and short polypeptides showed that deviations can be up to about 6^o for the trans-configuration and 20^o for the cis-configuration [9], [41], [51].

1.2.2. Cα-only Representation

While the local conformation of Cα atoms in most of the secondary structures found in proteins can be described by the Ramachandran dihedral angles, additional descriptors are required for secondary structural elements with longer-range interactions (e.g., the β-turns). In 1978, Rackovsky and Scheraga [52] introduced the differential-geometrical representation, considering four consecutive Cα atoms as the next level above the φ/ψ representation (Figure 3b), as a four-Cα unit is the smallest segment of polypeptide backbone that can be said to be folded [53]. Assuming only trans-peptide units (Figure 2a), they suggested the use of two angles to describe the conformation of a four-Cα unit (Figure 3b): κ ($C\alpha_{i+2}$-$C\alpha_{i-1}$-$C\alpha_{i+1}$), describing the way in which the chain changes its direction, and τ ($C\alpha_{i-1}$-$C\alpha_{i}$-$C\alpha_{i+1}$-$C\alpha_{i+2}$), describing the backbone twist [52], [53]

Other studies suggested the use of three angles [54], [55]: the defined above pseudo-torsion angle τ together with θ_1 ($C\alpha_{i-1}$-$C\alpha_{i}$-$C\alpha_{i+1}$) and θ_2 ($C\alpha_{i}$-$C\alpha_{i+1}$-$C\alpha_{i+2}$) angles (Figure 3b). The joint distribution of these three angles for a set of 83 structures collected from the PDB presents well-defined regions and shows that only about 30% of the total conformational space is occupied with two main peaks corresponding to the helical and the stranded conformations. A number of minor peaks represent different types of turns and transitions between the Ramachandran plot regions [55]. One advantage of this description is the possibility to describe β-turns, as they are defined as a sequence of four consecutive Cα where the distance between $C\alpha_{i}$ and $C\alpha_{i+3}$ is less than 7.0 Å [55]. Another application concerns the validation of Cα-only protein models (e.g., from low resolution experiments) [33].

1.2.3. Other Approaches

More recently, methods independent of angles have been proposed. Peng *et al.* [56] suggested a three-dimensional approach for the localisation of all backbone atoms from their relative position with respect to the Cα atoms. It is based on a miniature observer that travels through the Cα trace and "looks around" within a sphere for other atoms composing the main- and the side-chains. This sphere shows different clusters corresponding to different secondary structural elements and depends only on the Cα coordinates, providing purely geometric and direct visual information on the statistically expected all-atom structure in a given protein model [56].

The method proposed by Penner *et al.* [57] describes protein main-chain conformation around hydrogen bonds, which can be non-local along the backbone. It is based on the spatial rotation between hydrogen bonded peptide planes and the descriptor is a three-dimensional vector used to derive a position inside a sphere called the *rotational space*. It describes the geometry of the hydrogen bond and can be useful for the annotation of protein secondary and tertiary structure and the classification of protein folds [57].

1.3. Macromolecular X-Ray Crystallography

Several methods are available for obtaining spatial information about macromolecules in different states, including macromolecular X-ray crystallography – MX, nuclear magnetic resonance - NMR, electron microscopy – EM and small angle X-ray scattering – SAXS [58], [59].

MX is the most widely used technique for proteins and protein-ligand interactions as it is capable to deliver structural information at an atomic level of detail of macromolecules from a wide range of sizes [59] and has provided more than 90% of all entries in the PDB (Figure 4a), 98% of which are proteins or protein/nucleic acid complexes.

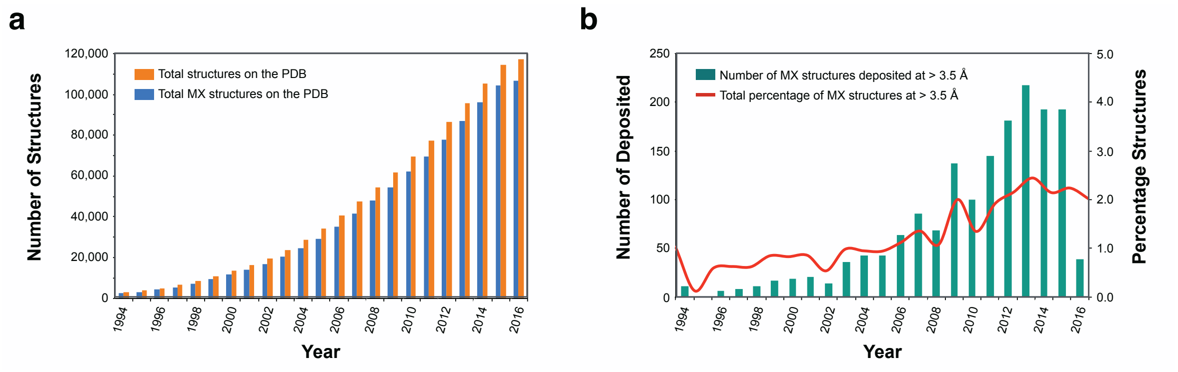


Figure 4 PDB statistics since 1994. (a) Annual growth of all structures and those determined by X-ray crystallography. (b) Total number and percentage of deposited crystallographic structures at a resolution worse than 3.5 Å, per yea.

X-rays have a very broad spectrum (0.01 to 50 nm) and can be used to resolve atoms. The method used for X-ray macromolecular structure determination is based on the interpretation of diffraction patterns that are produced when macromolecular crystals are irradiated with the beam [58] (Figure 5). X-rays are scattered in several directions by the electrons and the scattered beams are recorded on a detector (e.g., photographic films, imaging plates, charge-coupled devices (CCD) [60], or direct pixel detectors [61]). The overall shape, symmetry and detailed structure of the crystal define the directions of the diffracted beams while their intensities define the mutual locations of atoms in the macromolecule [62].

The ability of crystalline solids to produce patterns from reflected X-rays was first explained by William Lawrence Bragg and his father William Henry Bragg 100 years ago [63]. A crystal can be seen as many sets of discrete parallel planes separated by a constant distance *d* (Figure 5b). If the reflections of an incident X-ray beam interfere constructively, they produce intense spots at specific angles. This occurs when the path length differences of reflected X-ray beams equals to an integer multiple of the wavelength, referred as the *Bragg's Law* (eq. 1):

$$n\lambda = 2d\sin\theta \qquad (1)$$

where λ is the radiation wavelength, θ is the angle between the incident beam and the crystalline planes, *d* is the spacing between the planes and *n* is any integer. The reflection angle for a diffracted beam can be calculated from the distance *r* between the diffracted spot on the detector and the position where the incident beam hits the detector (Figure 5a). From equation 1, an increase of the θ angle is equivalent to a decrease of the spacing between the crystal planes (*d*).

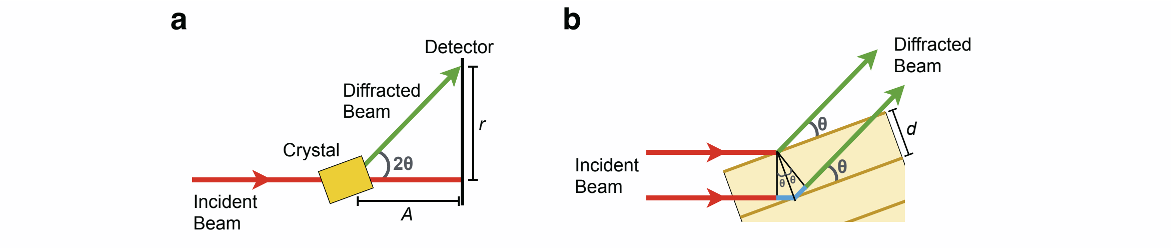


Figure 5 Diffraction of X-rays by a crystal. (a) Relationship between the angle θ and the position of the reflection spot on the detector. *A* is the sample-to-detector distance and *r* the distance between the spot and the beam centre. (b) Representation of a crystal as a set of parallel planes illustrating the Bragg's law. The crystal is in yellow, the incident X-ray beam in red and the diffracted beam in green. The path length difference between the waves is coloured in blue. *d* is the distance between the crystal planes.

What is measured in the diffraction experiment are the intensities of the waves diffracted from the crystal planes at the coordinates of the reciprocal space denoted as *hkl*, where the amplitude of the wave $|F_{hkl}|$ is proportional to the square root of the diffracted intensity. In a unit cell volume *V*, the electron density ρ_{xyz} at location (*x,y,z*) corresponds to the summation of the amplitudes of the structure factors F_{hkl} and the associated phase angle a_{hkl} in reciprocal space [64] (eq. 2):

$$\rho_{xyz} = \frac{1}{V}\sum_{h}\sum_{k}\sum_{l}|F_{hkl}|\cos 2\pi(hx + ky + lz - a_{hkl}) \qquad (2)$$

The electron density in real space and the diffraction pattern in the reciprocal space are then related through a *three-dimensional Fourier Transform*. While $|F_{hkl}|$ can be derived from the intensity of the diffraction spot, the phase cannot be obtained directly. This is known as the *phase problem* and presents a significant challenge in structure determination [65]. To obtain an electron density map, one needs to recover the phase information, as discussed in the next section. The obtained crystallographic models are the interpretation of the time- and space-averages of the electron density over the whole crystal [59].

The minimum distance to resolve two point atoms in three-dimensional electron density maps, the *resolution limit* of the data D_L, relates to d_{min} by a factor of ~0.9 (eq. 3) [66]:

$$D_L = 0.917 d_{min} \qquad (3)$$

Therefore, the higher the angle θ, the smaller the minimum distance and the higher the resolution limit. The effective resolution of the data is a measure of the extent, quality and completeness of the X-ray diffraction data. In MX, peaks of density are also affected by thermal motion and disorder and, therefore, the effective resolution can be lower than the estimate obtained with equation 3 [67].

The first globular protein structure obtained using MX was that of sperm whale myoglobin at 5.0 Å resolution by John Kendrew and colleagues in 1958 [68], [69]. Since then, the number of protein crystal structures has been increasing exponentially, from 10 known structures in 1973 to 102 166 in September 2016 (Figure 4a). This is a result of the emerging sophisticated techniques developed over the last 20-30 years: robotics, advanced sample handling as well as synchrotron beamlines. Coupled with the continuously improved software for data interpretation, modelling and validation, MX makes it nowadays possible to acquire data for cases where no structural information could have been obtained before [58], [59], [62].

1.3.1. The Crystallographic Experiment and Map Calculation

A macromolecular crystallographic experiment starts in the wet-lab, by the overexpression and purification of a sufficient amount of the target macromolecule (Figure 6). The next step is the identification of conditions in which crystals diffracting to a resolution sufficient to answer the scientific question posed are formed. Although the parameters governing the process of protein crystallisation are now better understood, it is still impossible to predict the conditions under which a particular protein will crystallise [59], [70]. These steps are labour-intensive and time-consuming and can take months [59]. After a well-diffracting crystal is obtained, it is irradiated with a beam of X-rays and rotated in order to obtain a set of diffraction patterns. Initially, X-rays were generated in various types of vacuum tubes, where highly accelerated electrons bombarded anode targets made of metals, leading to the emission of characteristic X-rays with wavelengths dependent on the anode material [70]. In mid-70s, these tubes were superseded by synchrotron radiation, which increased the attainable fluxes of X-rays by many orders of magnitude and allowed the selection of any wavelength in the range within 0.5 - 3.0 Å [70].

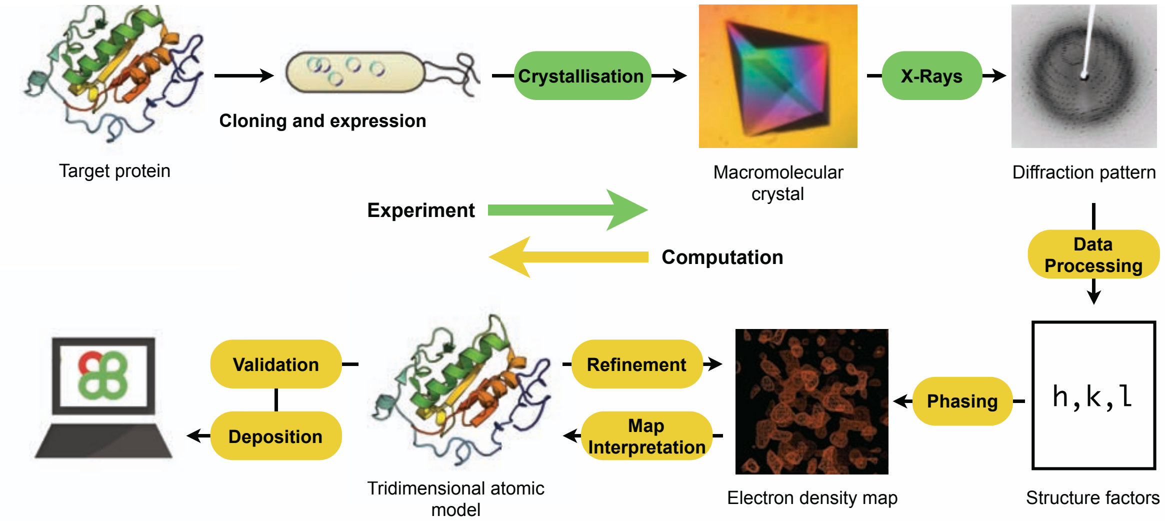


Figure 6 General workflow of a macromolecular crystallography experiment.

Data processing leads from diffraction patterns to a set of structure factors and comprises three main steps. The first is indexing, where the crystal geometry and its orientation is determined, allowing the spots in the diffraction pattern to be assigned Miller Indices (*h,k,l*). The second is the integration of the data, where the intensities of the diffracted reflections, *I*, are determined. Finally, the intensities of the same reflections measured at different images are combined into one scaled set, from which structure factor amplitudes and their associated standard deviations, $\sigma(F)$, are derived.

In order to calculate the electron density [64] one needs to obtain the missing phase information. The approaches include direct methods, isomorphous replacement (IR), multiple and single anomalous dispersion (MAD and SAD) and molecular replacement (MR) [71]. Direct methods allow the *ab initio* determination of phase information from the measured amplitudes alone. They are based on phase relationships between the normalised structure factors, implying that once the phases of some reflections are known, or can be given a variety of starting values, the phases of other reflections can be determined [65]. Their application is limited by the model's size and the data's resolution, being suited for smaller systems than for macromolecules [64]. IR, MAD and SAD use the positions of a few (heavy) atoms to derive the initial phase information. IR exploits additional data collected from the same structure but with one or a few electron-rich atoms added. These give rise to measurable intensity changes, which are then used to compute the positions of these heavy atoms [72]. In MAD and SAD the signal of anomalously scattering atoms present in the macromolecules (e.g., sulphur and phosphorous) or added to them (halides or heavy atoms) is used to derive their positions [73], [74].

MR is an alternative to the experimental methods [75]. Here, the approximate phase information is obtained from a homologous macromolecule, or its fragment, oriented and positioned so that its calculated structure factors fit the observed data. The initial phase estimates are taken from these calculated structure factors. With the increase of the number of protein structures deposited in the PDB, MR has become one of the most used phasing techniques [76]. A number of programs and pipelines have been developed that allow obtaining a starting model and the calculation of an initial electron density map. These use different types of models, ranging from the complete models taken from the PDB to small decoys obtained by homology or *ab initio* modelling. Example software approaches include AMoRe [77], MOLREP [78], Phaser [79], MrBUMP [80], BALBES [81], AMPLE [82] and ARCIMBOLDO [83], [84].

Typically the experimentally determined phases are insufficiently accurate to give a fully interpretable electron density map and further phase improvement is needed. Phases can be improved by density modification, including solvent flattening, histogram matching and non-crystallographic symmetry averaging [65], [85]–[87]. It is often realised in an iterative manner, involving back-transformation of the modified electron-density map to produce modified phases, their recombination with the experimental phases and calculation of a new map for the next round of density modification. This continues until convergence [65].

1.3.2. Map Interpretation and Model Building

The ultimate result of a macromolecular X-ray crystallography experiment is the electron density map and it is its *interpretation* that allows the building of the macromolecular model (Figure 6) [62], [70]. The electron density is interpreted in terms of atoms and bonds based on the *a priori* knowledge of the chemical nature of the molecule. As the full protein main chain can be determined to a high degree of accuracy by the positions of the Cα atoms alone [88] and side-chain placement is dependent on the main chain, protein model building is often seen as a problem of locating the Cα positions, a pattern recognition problem that becomes more difficult with the decrease in the information that can be deduced from the electron density map.

In the early years of MX, crystallographers printed slices of electron density maps onto plexiglass sheets, which were glued to wooden frames and stacked in order to build a three-dimensional representation of sections of the density. Brass or balsa wood were then used to build models that followed the main-chain path [68], [89]. Forests of rods were used in the 1960s to build the models of myoglobin [90] and haemoglobin [91], where coloured clips attached to more than 1,000 vertical 2 meters-long rods marked the electron density and guided the model building. Around the same time, Frederic Richards invented the Richard's box [92], used to build the medium-resolution model of RNAse S by projecting an electron density map upon the model with semi-transparent mirrors, reducing the size of the model and improving its adjustment, movement and analysis. With the advances in computer graphics, plexiglass and rulers were replaced by molecular graphics programs such as FRODO [93], O [94], XtalView [95] and COOT [96], [97]. Although molecular graphics stimulated the use of the manual interpretation of electron density, it has been still a labour-intensive and subjective process. The need to speed up macromolecular structure determination and to provide some objectivity into the model building process gave rise to automated model building procedures (Figure 7).

Perhaps the first step towards automation was the skeletonisation of the electron density map, developed by Jonathan Greer [98] (Figure 7a). It is based on placing points at density peaks and connecting them following the density paths, reducing the 3D electron density to a set of connected line segments. The obtained skeletal representation can then be used to derive potential Cα positions, and QUANTA [99], [100] and CAPRA [101], [102] are examples of programs that use this approach. QUANTA identifies regions that correspond to α-helices and β-strands by principal component analysis (PCA) of the skeleton representation. The identified segments are used to define plausible Cα positions in order to build the protein backbone. CAPRA predicts Cα positions by using a range of rotation-invariant electron density numerical features and connects them into chains by a heuristic search method. This has been implemented in the TEXTAL software, by coupling it with the sequence alignment and real space refinement [103]–[105].

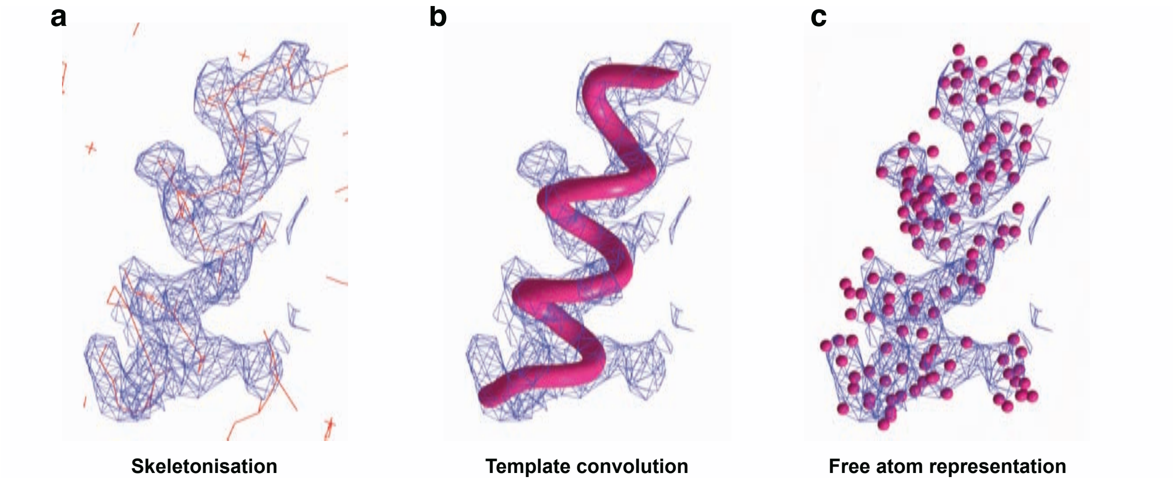


Figure 7 Methods for the automated interpretation of MX electron density maps. (a) Skeletonisation as introduced by Jonathan Greer [98], (b) template convolution, by a search for secondary structural elements, as introduced by Kleywegt and Jones [106] and (c) the representation of the electron density map as a set of free atoms without any chemical identity, as introduced by Isaacs and Agarwal [107].

Another method is based on a search for known small motifs within the electron density using template convolution. ESSENS [106] was developed by Kleywegt and Jones in 1997, and used penta-alanine templates in an ideal α-helical and β-stranded conformations (Figure 7b). FFFEAR [108] follows the same concept by computing the target function for nine-residue long search fragments in reciprocal Fourier space to reduce the computation time. This search function is used by BUCCANEER [109] to locate possible Cα atoms which are subsequently refined before extension into chains. BUCCANEER uses then an exhaustive search over the Ramachandran plot and assigns a probability for each amino acid type. The software *phenix.resolve* [110]–[112], a part of the PHENIX project [87], [113], [114], employs a search function similar to FFFEAR.

The next approach is based on the representation of the electron density as a set of ‘free’ atoms without chemical identity (Figure 7c) as introduced by Isaacs and Agarwal in 1985 [107] and implemented for protein model building in ARP/wARP [115]–[117] (described in detail in section 1.4). Depending on data quality, resolution limits and accuracy of the phase estimates, the locations of free atoms may be quite close to corresponding positions in the final structure [118]. The free-atoms set is used to search for peptide planes and dipeptide units using the distance and density information and extended to build longer protein backbone fragments, which are subsequently decorated with side-chains and connected by short loops.

The ACMI (Automatic Crystallographic Map Interpreter) method employs probabilistic inference to compute a probability distribution of the coordinates of each amino acid given the electron density map [119] and constructs all-atom models by stepwise extension of incomplete models drawn from this distribution using a statistical method called particle filtering [120].

All these methods use related techniques, mimicking the steps a crystallographer would take when building the model manually. They differ in the search patterns or density shapes and often result in models built to a different extent of completeness for the same crystallographic data. Therefore, a model built in an automated way should not be regarded as truly final but better be inspected visually together with the electron density.

1.3.3. Model Refinement

To produce an accurate model, crystallographic refinement and model rebuilding is carried out in an iterative manner, gradually approaching the final model. Refinement can be accomplished in real or reciprocal space aiming at an optimisation of the agreement between the model and the observed experimental data.

The agreement in real space can be measured by the real space R-value (RSR) (eq. 4) [121]:

$$RSR = \frac{\sum\left|\rho_{obs} - \rho_{calc}\right|}{\sum\left|\rho_{obs} + \rho_{calc}\right|} \tag{4}$$

It is computed for a group of atoms, where the observed (ρ_{obs}) and calculated (ρ_{calc}) electron densities are sampled on a grid that covers them. Currently, several versions of the RSR exist, and different programs can output different RSR for the same data. Another measure is the real space correlation coefficient (RSCC) (eq. 5) [121]:

$$RSCC = \frac{\operatorname{cov}(\rho_{obs}, \rho_{calc})}{\sqrt{\operatorname{var}(\rho_{obs}) \cdot \operatorname{var}(\rho_{calc})}} \tag{5}$$

which is a standard linear sample correlation coefficient where var(·) is the sample variance and cov(·) the sample covariance.

While RSR varies from 0 (‘good’) to 1 (‘bad’), the RSCC is good at values close to 1 and bad close to 0. Both represent a quantitative measure of how well a residue (or any other group of atoms) fits its local density and sum over all map-grid points that are near this group. While these can be used to improve the model built, they also have important application for model validation [121]. Calculated for each residue in a protein model, they may highlight problematic regions that do not agree with the density. However, by using the information about the electron density, they account for both structure factor amplitudes and phases, and are, therefore, affected by the quality of both [62].

Reciprocal space refinement aims at optimising the agreement between the structure factors calculated from the model parameters (F_{calc}) and the experimental data (F_{obs}). It is typically followed by monitoring the R-factor (eq. 6):

$$R = \frac{\sum\left||F_{obs}| - |F_{calc}|\right|}{\sum|F_{obs}|} \qquad (6)$$

Model parameters that are subject to the optimisation include atomic positional coordinates complemented with individual, group or overall atomic displacement parameters (ADPs), overall scale factors, bulk solvent, twin fractions, etc. [122]. In cases when the number of parameters exceeds the number of experimental observations (usually at lower resolution, as discussed in 1.5) additional information is needed in a form of restrains or constrains. These include *a priori* structural knowledge about protein stereochemistry, chirality and planarity of atomic groups, NCS between molecular fragments or substructures, etc. [123], [124]. Well-refined macromolecular models are expected to have R < 20%, subject to the resolution of the data. When R is above 30%, the model should be regarded with a high degree of caution because at least some of its parts may be incorrect [62].

Despite the use of stereochemical restraints, it is possible to overfit the model. Thus, an unbiased factor similar to the R-factor was introduced to control the accuracy of the models, the R_{free} [125], [126]. It gives a less biased quality index as it is computed from a small subset of structure factors, usually 5% of the data, that is not used during refinement and model building [125]. Therefore, only changes to the model that lead to a better explanation of the experimental data will improve R_{free}. It is always higher that the R-factor but should not exceed it by more than about 7% [62].

The refinement of a macromolecular model is a complex optimisation problem. In many refinement methods each atom was described by a coordinate in a 3D space and an atomic displacement parameter indicating how far the atom moves around its equilibrium position [127]. The use of structural information was introduced later, by rotating the atomic groups around rotatable bonds while keeping the stereochemistry fixed [128]. In both cases, least-squares procedures in real and reciprocal space were applied to minimise the residual between the observed and the calculated data [129], [130]. The need to account for the uncertainty in model parameters led to the use of maximum likelihood methods. REFMAC5 [122], [131] is an example of a refinement software that efficiently employs this approach to maximise the probability of observing the current model given the set of observations, together with the additional knowledge of protein stereochemistry.

Likelihood target functions may differ and depend on the input diffraction data and their resolution. The target function (f_{total}) has two main components: the one utilising geometry (or prior knowledge; f_{geom}) and a component describing the experimental X-ray data ($f_{\text{x-ray}}$) (eq. 7):

$$f_{total} = f_{geom} + w \cdot f_{x-ray} \qquad (7)$$

The optimum weight w defining the relative contributions of these two components can be selected automatically on-the-fly. From a probabilistic point of view, these functions are described by eq. 8, 9 and 10:

$$f_{total} = -\,log\left[P_{posterior}(model; obs)\right] \tag{8}$$

$$f_{geom} = -\,log\left[P_{prior}(model)\right] \tag{9}$$

$$f_{x-ray} = -\,log\left[P_{likelihood}(obs; model)\right] \tag{10}$$

Different refinement programs differ in their target functions and optimisation techniques and, therefore, may lead to different results and values of R-factors. *Phenix.refine* [124], from the PHENIX project, for example, has two target functions, one for the coordinates (T_{xyz}, eq. 11) and another for ADPs (T_{adp}, eq. 12):

$$T_{xyz} = wxc_{scale} \times wxc \times T_{exp} + wc \times T_{xyz\_restraints} \tag{11}$$

$$T_{adp} = wxu_{scale} \times wxu \times T_{exp} + wu \times T_{adp\_restraints} \tag{12}$$

where T_{exp} is the crystallographic term that relates the experimental data to the model structure factors and can be a least-squares target, an amplitude-based maximum likelihood target or a phased maximum-likelihood target. It can also be defined in real space for the refinement of coordinates against the given density map. $T_{\text{xyz\_restraints}}$ and $T_{\text{adp\_restraints}}$ are restraints terms that introduce the *a priori* knowledge. Weights wxc_{scale}, wxc, wc, wxu_{scale}, wxu and wu balance the relative contributions of the experimental and restraints terms and are defined automatically [124].

Refinement procedures were first used during the final stages of MX structure determination. Currently, they are often used to improve partial models and to obtain better electron density maps for subsequent rounds of model building. Therefore, model building and refinement programs are used hand-in-hand in MX structure solution, as for example in ARP/wARP [117] or *phenix.autobuild* [132].

1.3.4. Ligand Building and Identification

Even after multiple successful rounds of protein model building and refinement, some of the density may remain uninterpretable due to a manifold of reasons. For example, proteins may crystallise in the presence of small molecules or nucleic acids to which they can bind. Small molecules can be known (e.g., in drug design projects) or unknown (e.g., buffer components). While for protein and nucleic acids the model building procedure depends on the known sequence of the macromolecule and the completeness and quality of the data itself, ligand building presents several challenges: (1) the universe of compounds that interact with macromolecules is vast, accommodating different complexities, shapes and topologies (as of September 2016, there are more than 20 000 ligand entries in the PDB); (2) ligands can be partially disordered; and (3) can even ‘cooperate’ with other ligands. Therefore, the automated modelling of ligands has always been less advanced than that for proteins. The increasing interest in structure-based drug design has promoted an increase in the number of methods and software packages for the automated building of small molecules in electron density maps.

ARP/wARP [133], PHENIX [87], [113], [114] and COOT [96], [134] are examples of academic software packages widely used for ligand fitting into MX electron density maps. PRIVATEER [135], from the developers of BUCCANEER [109], is specifically designed for the fitting and validation of carbohydrate molecules. All of these packages use different methods and approaches that maximise the fit of the ligand to a density cluster. ARP/wARP (described in detail in section 1.4.3) identifies atomic features in the given density cluster and interprets them in terms of connectivity [136] and possible conformational space [137]. PHENIX searches for rigid parts of the ligand and then attempts their extension following the density shape [138]. COOT works by identifying the density that fits a predefined conformation of the ligand and then adjusts the ligand conformation in order to maximize the fit [96].

These programs also provide tools for the identification of possible binding sites (when a ligand is known but the correspondent density cluster is not) and the guessing of possible ligands fitting into a known density cluster when the identity of the ligand is unknown. ARP/wARP ligand guess and the identification of the binding site are described in detail in section 1.4.3. PHENIX methods find all possible binding sites by identifying contiguous regions of density and matching them to the most likely ligand (currently, from a database of 200 small molecules) by fitting each ligand into the density clusters and choosing the one with the best fit and complementarity to the atoms surrounding the binding site [114], [139]. COOT can screen a cocktail of ligands [134].

1.3.5. Model Validation

After building and refinement, the complete macromolecular model may still contain errors. These can be a result of incorrect tracing of chain fragments or flexible loops, presence of peptide flips, incorrect side-chain conformation, etc. [140], [141]. Therefore, it is important to validate the model and to ensure that it makes sense from a biochemical point of view [142], [143]. During refinement, only geometrical restraints, derived from the analysis of the X-ray structures of amino acids, peptides and small molecules in the Cambridge Structural Database (CSD) [123], [144], are used. Neither the conformational attributes of the macromolecule, nor energetic terms are taken into account [51].

The most widely used protein backbone validation method is the Ramachandran plot (Figure 3a), and several validation tools exist that use it for the identification of conformational problems in protein models. Procheck [38], MolProbity [39] and WHAT_CHECK [145] are three main examples. Procheck provides a detailed check of the protein stereochemistry by computing the Ramachandran plot for several residue types and a number of plots for side-chain conformational parameters, main-chain and side-chain bond lengths and main-chain angles [38]. MolProbity is similar to Procheck but also includes clash analysis from interatomic contacts, providing evaluation independent of refinement targets [39], [146]. While it can be used as a stand-alone tool, PHENIX makes use of MolProbity for the validation of automatically built protein models [147].

WHAT_CHECK was developed as part of the WHAT_IF package and provides a series of Z-scores for several conformational features of proteins models, including packing, Ramachandran plot, side-chain rotamers, backbone conformation and bond lengths and angles [145], [148]. For example, the Ramachandran appearance Z-score tells how many standard deviations the overall distribution of φ/ψ angles for a given protein model deviates from the distribution observed for a set of reference structures. To calculate it, WHAT_CHECK divides the Ramachandran plot into several $10^{o}\times10^{o}$ bins and counts the number of residues that fall in each bin for a set of protein models. The higher the population of the bin, the higher the likelihood of the conformation to be correct. This is carried out for a set of 3×20 Ramachandran plots, corresponding to three main types of secondary structural elements and to each residue type. A Z-score is then calculated for each residue in the protein model and an overall protein model score *C* calculated as the average of all these computed Z-scores. A Z-score is further calculated for *C* by calculating its deviation from the mean <*C*> calculated for a set of good quality models [149].

Concerns have been raised regarding the quality of the protein models deposited in the PDB [142], [143]. Validation reports are now generated automatically during the deposition of a macromolecular model [40]. The mandatory deposition of experimental data allows the validation of the deposited MX models, and such reports can be found for each PDB entry. A further step was taken by Robbie Joosten with the PDB_REDO project, making use of the deposited X-ray data to automatically re-refine and re-build all models from the PDB according to the current standards and software [150], [151]. Such re-refinement shows that the geometric validation scores can be improved for many PDB entries [151].

More recently, efforts are being put into the improvement and validation of the deposited protein-ligand complexes too, as shown by the recent CCP4 study weekend held in January 2016 fully dedicated to protein-ligand building, refinement and validation. This was raised by the identification of several deposited small molecule models with incorrect ligand geometry or lack of supporting density [141], [152]–[154]. In MX, ligand validation has always been performed in real space by using RSCC or RSR. From a streochemical side, several tools allow the verification of the proposed conformation and binding mode [40]. COOT [134] incorporates tools for the 2D representation of the binding mode, ligand binding pocket layout and scoring of ligand conformation with the CCDC program Mogul [155]. PHENIX uses MolProbity and the complementarity of the ligand atoms with the binding pocket [39], [147]. Ligand geometry can be optimised with PRODRG [156], but also compared with another instances of the same ligand in the PDB with ValLigURL [157]..

1.4. The ARP/wARP Project

ARP/wARP (www.arp-warp.org) has been a collaborative development led by the Lamzin group since the early 90’s at the EMBL in Hamburg [118], [133], with the aim of automatically build complete and accurate protein [116], [117], [158]–[163], nucleotide [164] and small molecule [136], [137], [165], [166] models from the automatic interpretation of an MX electron density map. It is based on the idea of combining the interpretation of an electron density map with the iterative model rebuilding and refinement of the atomic parameters [115], [118] and on the application of the ‘free atom’ concept [107] for the identification of likely atomic positions (Figure 8a).

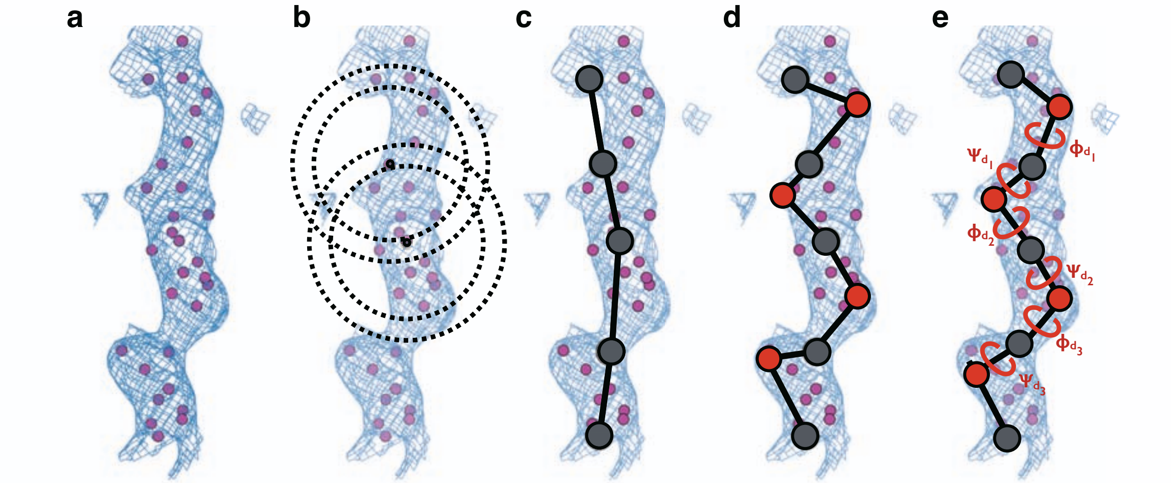


Figure 8 Main-chain tracing by ARP/wARP. (a) Parameterisation of the electron density as a set of free atoms, (b) identification of putative Cα-Cα pairs (in grey), (c) search for potential dipeptide units and (d) calculation of the positions of oxygen atoms (in red). After geometrisation, (e) the dipeptide conformation is checked against the Ramachandran-like plot.

1.4.1. Main-chain tracing

To find the best subset of free atoms that looks like a protein, a list of possible connections is generated by evaluating the distances between all free atoms, followed by a peptide-shape density analysis. If two free atoms are 3.8 ± 1.0 Å (Figure 8b and Figure 9a) apart and there is reasonable density around the peptide plane, these two free atoms are marked as a putative pair of Cα atoms. At this stage, putative peptide units are composed of two Cα atoms only and, given the presence of supporting density between them, the position of the oxygen atom can be estimated (Figure 8d).

Peptide units that share one Cα atom and have the same direction can be connected to form a dipeptide unit (Figure 8c). These are described by five atoms, $C\alpha_{i-1}$-O_{i-1}-$C\alpha_i$-O_i-$C\alpha_{i+1}$, and their conformational space is described by the Ramachandran-like plot (Figure 3c and Figure 8e). The Cα atoms in putative peptide units sampled by ARP/wARP may have a deviation up to 1.0 Å from the standard value of 3.8 Å. Therefore, before evaluating the conformation of resulting dipeptide units, they have to be geometrised so that the inter-atomic distances within the peptide plane are closer to the expected geometry (Figure 9c) [118]. In one round, three $C\alpha_i$-O_i-$C\alpha_{i+1}$ atoms from an ideal peptide are least-squares-superimposed on each of the two peptide units; the resulting two positions of the middle $C\alpha_i$ atom are then averaged.

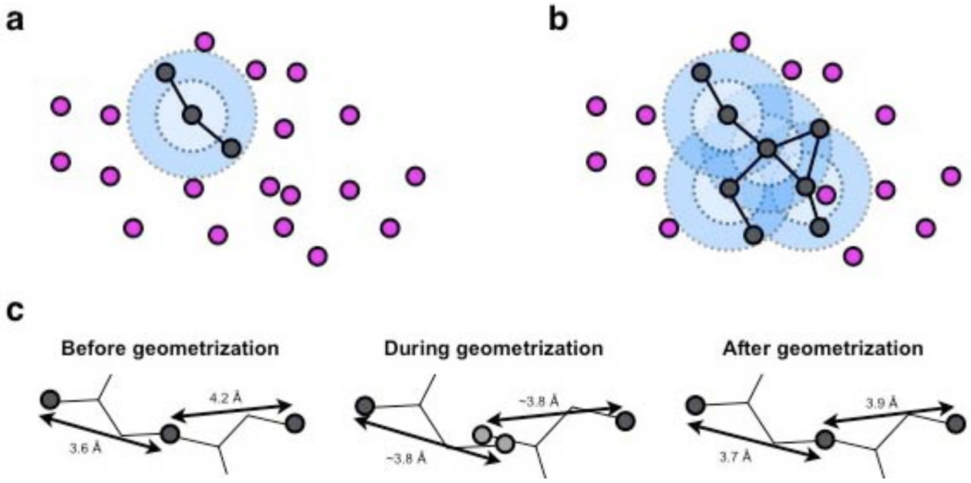


Figure 9 Chain path selection and geometrisation during main-chain tracing. (a) Pairs of atoms separated by 3.8 ± 1.0 Å (the ones inside the blue ring) are identified. (b) The connectivity of the built peptides is searched; often peptide units can have more than one possible incoming or outgoing connection. (c) Peptide units are geometrised so that the peptide plane approximates the geometry of perfect trans-peptide planes. If two peptide units share the same Cα atom, this geometrisation will move the common Cα atom to two different positions, which coordinates are then averaged.

Dipeptide units with valid conformations can be connected to build up the polypeptide chains. When the longest possible fragments are found, dipeptide units that give rise to steric clashes are removed. Iteratively, every next-longest chain is considered until no more chains longer than five Cα atoms can be found [118], [159]. If there is sufficient density support, four types of side-chains (glycine, alanine, serine and valine) are built. The chain fragments are then geometrised and real-space fit to the density. Since only a part of the free atoms is recognised as a set of polypeptide chains, the result is a 'hybrid model', incorporating chemical information from the partially built model and the free atoms, which continue to interpret the electron density in areas where no model is built. The restrained refinement of the chemically assigned parts helps improve the phases, allowing the building of more chemically assigned fragments in a better electron density map. Therefore, ARP/wARP combines model building and refinement in a scheme of restraints and 'free atoms' that are iteratively updated and result in a hybrid model that converges towards the final model [117], [159], [160].

1.4.2. Side-Chain Building and Loop Fitting

After main chain tracing, side-chains are built according to the available protein sequence [117], [160], [161], [163]. A feature vector is used that represents possible connectivity between the free atoms in the vicinity of each Cα (including the atoms of the guessed four types of side-chains during the main-chain tracing), and then compared to all full-chain connectivity vectors of the 20 possible residues. The sequence can be assigned by matching the assigned connectivity vectors to a matrix of connectivity vectors known for the 20 proteinogenic amino acids (Figure 10a) [160], [167]. After the sequence is docked, the best rotamers from a rotamer database are built and refined [160].

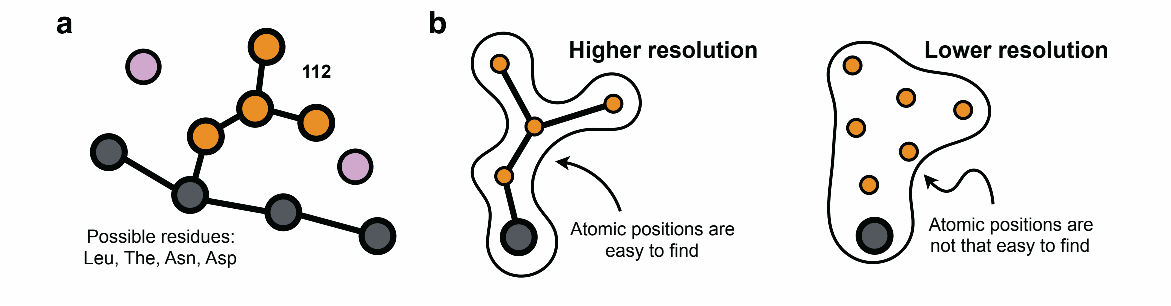


Figure 10 Sequence docking using connectivity vectors. (a) A feature vector is used to represent possible connectivity between the free atoms in the vicinity of each Cα atom and is compared with the connectivity vectors of the 20 residue types. (b) At lower resolution, the atomic positions are not so easy to find in the electron density; therefore, the connectivity may be different.

At the completion of automated model building, some loop regions may still be missing. ARP/wARP attempts to build them, up to 14-residues long, in most likely conformation by using structural and electron-density information [168]. By comparing the sequence of the fragments to the protein sequence, ARP/wARP identifies the fragments to be connected, fits Cα atoms of several template penta-peptides to the fragments' termini and extends the peptide segment iteratively. Subsequently, backbone conformations are constructed and the electron density correlation used to select the best loop. In the presence of non-crystallographic symmetry (NCS), ARP/wARP detects NCS-relations between the modelled fragments and uses them for chain extension [169].

1.4.3. Building and Fitting of Bound Ligands

ARP/wARP also allows the building and identification of ligands and ligand-binding sites, helping the crystallographer in several possible scenarios. The simplest case is when both the search ligand and the binding site are known. ARP/wARP represents the density region by a mesh of free atoms and tries to match the ligand topology to it [136]. As protein chain tracing, ligand fitting in ARP/wARP is organised as a pipeline of core modules for specific sub-tasks. It starts by preparing the ligand topology and a sparse grid representation of the binding site density (Figure 11a and b) and then constructs an ensemble of ligand models in plausible conformation to fit the sparsed grid. To address the different ligand sizes and complexities encountered, ARP/wARP uses two different ligand construction methods [137].

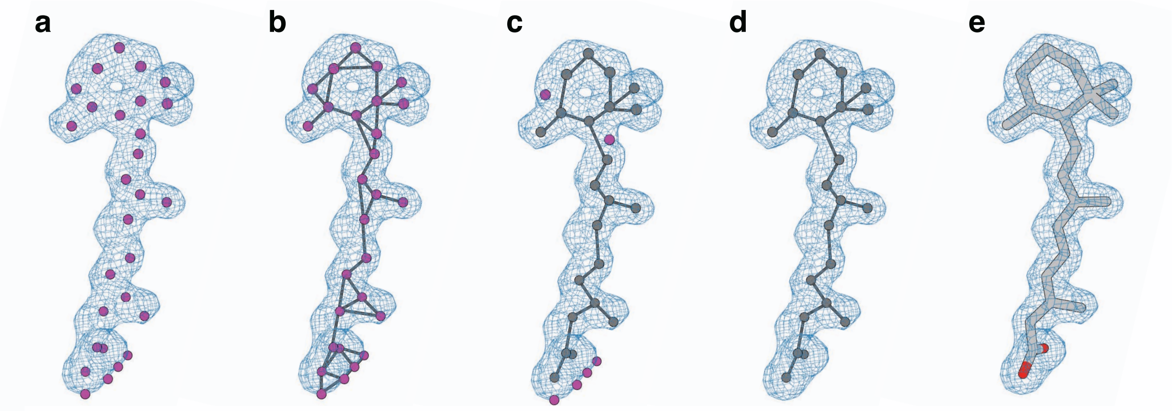


Figure 11 Retinoic acid (PDB ID: 1cbs) fitting with ARP/wARP. (a) Density representation as a set of free atoms, which (b) make a three-dimensional mesh with some connectivity. (c) The mesh is searched to find the mesh points that allow the full extension of the ligand. (d) The identified free atoms are used to fit the ligand. (e) The optimisation of ligand geometry and its fit to the density.

The first, *label swapping* [136], is an exhaustive graph search where the ligand is expanded on the sparse density, preserving its connectivity and not allowing steric clashes, trying every point of the sparse grid in turn as a starting point. All possible models are scored by their fit to the density and their expected stereochemistry. The models with the best fit and the longest expansion on the sparse grid are selected (Figure 11c and d). In parallel, a *metropolis search* on the ligands' conformational space is performed [137]. Here, the ligand rigid groups are rotated in order to maximise the fit to the density, while keeping all the rigid groups intact and penalising clashes. A combination of this method with label swapping provides better results than any of them alone [137]. Finally, real space refinement is employed to optimise the fit to the density and the ligand geometry (Figure 11e).

If the binding site is not known, ARP/wARP uses the *fragmentation-tree* method [137], which captures the dependence of the volume of the difference density blob on the change of the contour level. Contiguous regions of electron density higher than the contouring level represent density clusters and upon increase of the density-contouring threshold, the clusters of bound compounds reduce in their volume so that the ligand density areas are recognised from characteristic, approximately linear stretches [137]. Several clusters can be identified as possible binding sites. In order to decide which one is more likely, ARP/wARP uses *shape matching* to compare an electron density map calculated from the ligand to be fit to each potential density cluster (Figure 12a). This is accomplished using seven shape features that provide concise description of an object and the highest-scoring match taken for further ligand fitting [137].

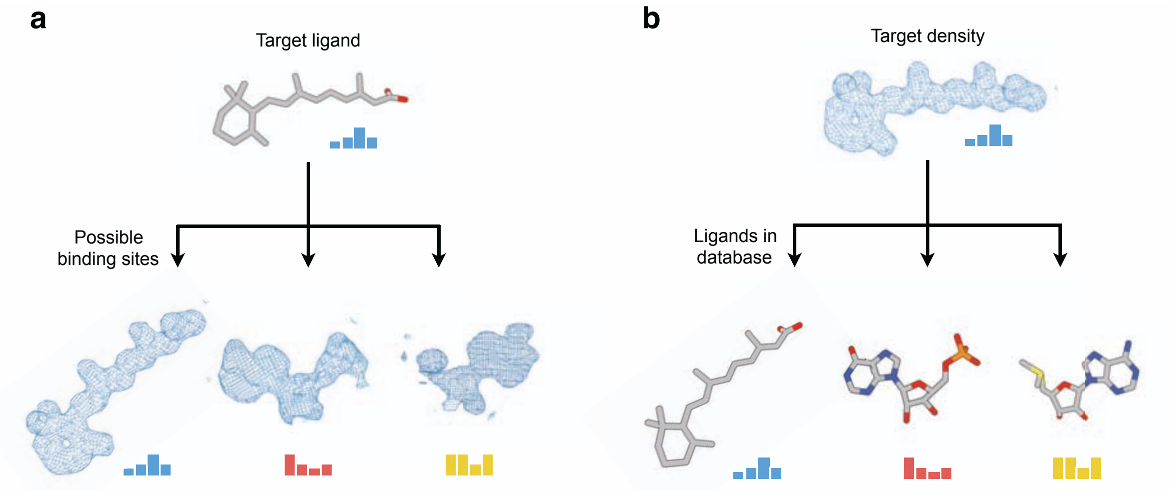


Figure 12 Shape descriptors in (a) retinoic acid (PDB ID: 1cbs) binding site identification and (b) ligand guessing. Coloured bars depict the shape descriptor fingerprint calculated for a set of density clusters and ligands. Ligands and density clusters with the same shape descriptor fingerprint correspond to each other.

Another common situation in MX is when a density cluster not explained by the protein model is observed but no ligand was expected. In this case, ARP/wARP can propose a possible ligand by comparing the density shape to those from a database with an approach similar to the shape matching method (Figure 12b) [166]. The shape of the mesh of the known density cluster is represented by 22 features, and these are compared to a database of 82 common crystallographic ligands in up to 200 conformations. Ligands in the database are then ranked according to their match to the density features and the top-ranking ligands in their best conformation superimposed on the sparse grid of the density map. Real-space refinement is then performed, and the ligand candidates are ranked by their fit to the density.

1.5. Limiting Factors in Crystallographic Protein Model Building

Given the fact that MX model building relies on the identification of known patterns in electron density maps, the resolution limit and the phase quality are the main limiting factors affecting the performance of automated approaches. While the phases can be improved, the resolution and the completeness of the experimental diffraction data stay.

As discussed in section 1.3, the resolution limit refers to the amount of information obtained in an MX experiment, which for a given crystal can be characterised by the total number of collected unique diffraction intensities. The higher the number of reflections, the more complete the data is, also in terms of the resolution (Figure 5a) [170]. Crystallographers struggle to obtain the experimental data to the highest possible resolution, but it is usually the crystal and the nature of the protein that defines the resolution limit. In particular, crystals of large proteins and their complexes, due to a lower surface-to-volume ratio (and thus less lattice contacts per molecule), higher flexibility and reduced number of molecules per crystal volume, tend to diffract to lower resolution than smaller proteins [171].

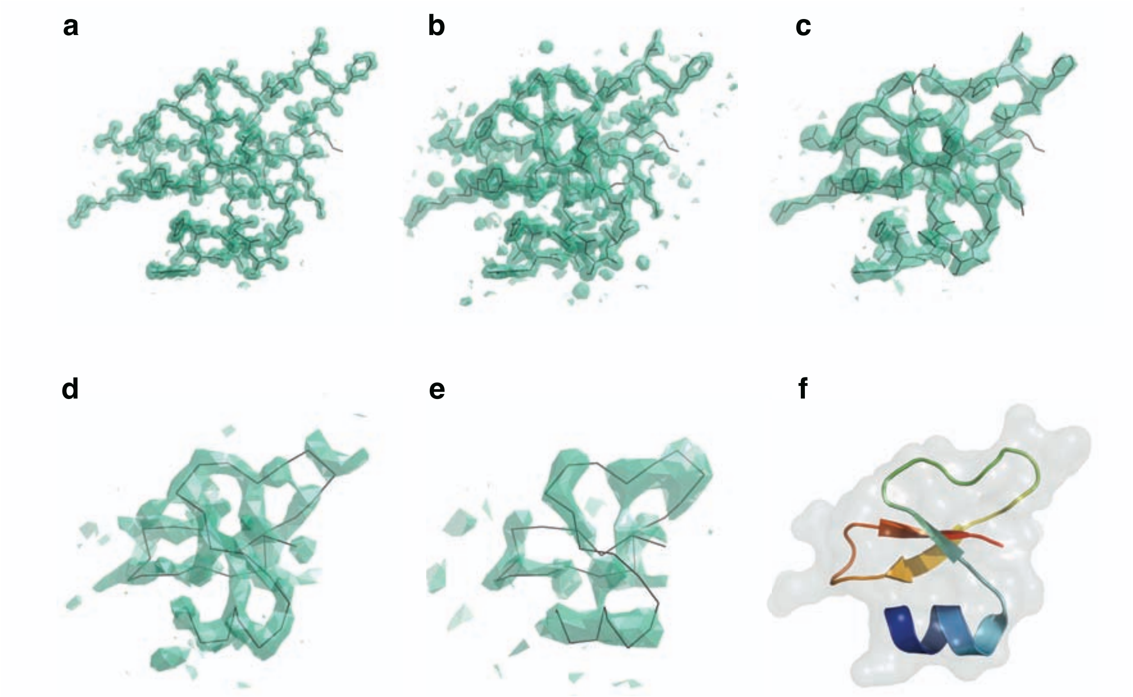


Figure 13 Electron density maps at different resolutions. (a) 1.0, (b) 2.0, (c) 3.0, (d) 4.0 and (e) 5.0. The maps were calculated from the final model of the human β-defensin-1 (f, PDB id 2nls). For the first 5 panels, the protein and Cα trace shown in grey thick lines and the map as transparent solid green-cyan.

At resolution lower than 3.5 Å, the number of observations available for structure refinement and calculation of an electron density is considerably smaller than the number of parameters to optimise [172], which requires then the use of additional data in a form of constraints or restraints. In addition, reduction of the resolution causes smoothing of density maps and a simultaneous loss of detectable features (Figure 13): at 4 Å peptide groups cannot be seen anymore, at 6-7 Å helices look as tubes and β-sheets as walls of density with no indication where the individual strands might be. It is thus difficult to trace the peptide main chain, taking into account that there are ambiguities in the direction of the chain and in the number of residues that make up various sections of the structure. The model building methods were historically developed with high-resolution diffraction data, and pattern-recognition-based map interpretation can be more intuitively applied at higher resolution. Automated protein model building in the medium-to-low resolution regime at best generates incomplete, inaccurate and fragmented macromolecular models [172] (Figure 14).

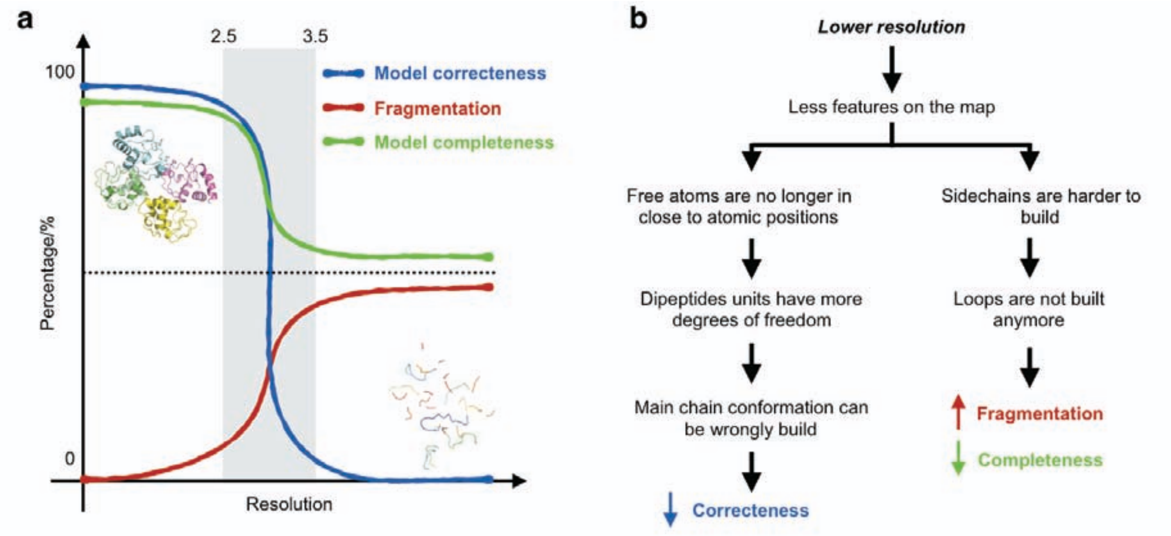

a
2.5
3.5
100
0
Percentage/%
Resolution
Model correctness
Fragmentation
Model completeness
b
Lower resolution
Less features on the map
Free atoms are no longer in close to atomic positions
Dipeptides units have more degrees of freedom
Main chain conformation can be wrongly build
Correcteness
Sidechains are harder to build
Loops are not built anymore
Fragmentation
Completeness

Figure 14 Conceptual effect of the resolution limit on the quality of protein models built by ARP/wARP. (a) Estimation of the model correctness, completeness and fragmentation at different resolution limits. (b) Schematic representation of how a low-resolution limit affects each of these quality indicators.

1.6. Challenges and Demands

There is an urgent need for efforts in the improvement of automated model building at medium-to-low resolution. The percentage of such entries in the PDB, although small, is rising, with the absolute numbers increasing rapidly (Figure 4b), demonstrating a growing interest and a need for structural information even at low levels of detail. A large number of important questions that structural biology attempts to address concern large macromolecules and their assemblies, and understanding of how their components interact. This does not necessarily require structures at near-atomic resolution, although it would clearly be more desirable to obtain them if possible [171], [172]. Availability of novel methodology and its robust software implementation for obtaining structural information from the low-resolution data will, in turn, increase the percentage of low-resolution models deposited in the PDB, and may also provide further uncovering of unknown parts of the protein folding space [30].

Often impressive results are reported for low-resolution structure determination, although seldom can a complete structure be built without user intervention (Figure 14). For example, with *phenix.autobuild* a completeness higher than 80% can be obtained for protein structures with data extending to resolution around 2.8 Å [87], dropping to about 60% at a resolution of 3.3 Å. BUCCANEER can build up to 87% of the model at resolution up to 3.2 Å and 76% of the model at resolution up to 3.6 Å [109], [173]. A similar behaviour has been observed for ARP/wARP [117], [133], [160] and the estimates from the ARP/wARP remote model-building web service suggest that protein model completeness of more than 90% is observed for resolution up to 2.5 Å, decreasing to 75% at 3.0 Å and 65% at 3.5 Å (Figure 14a).

Previous developments show that the combination of structural bioinformatics and modern X-ray data interpretation software is beneficial for the improvement of the completeness of automatically built protein models at a resolution worse than 2.5 Å. An example is the implementation of the structure extension module (PNSextender) in ARP/wARP, utilising NCS for dealing with fragmented protein models at a resolution range of 2.0-3.5 Å [169]. It is based on the observation that throughout structure determination of a protein with NCS, the NCS-related parts may be differently pronounced in the electron density. This results in the modelling of molecular fragments of variable length and accuracy, which can then be used to identify NCS relations and to extend other fragmented parts of the model. Another example is the FittOFF module, not yet implemented in ARP/wARP, that allows the extension of fragmented models and loop fitting in the absence of NCS and sequence docking [174]. Based on the prediction of the protein secondary structural content, the method identifies fragments that should be extended and connected by loops.

Given the increased uncertainty of the atomic positions at this range of resolution, the extension of fragmented models is of low value if the fragments are incorrectly built. It is necessary to first validate these models, mainly at the main-chain level. Automated protein model building softwares can provide that during main-chain tracing by selecting putative main-chain routes that show a conformation allowed in the Ramachandran plot. While this approach may be good at high-resolution, at lower resolution the higher coordinate error can lead the automated tracing algorithms selecting chains that do not follow the correct route (Figure 14b). It is important to develop tools that allow the automated protein model building to validate the built protein models on-the-fly. Such a method would most likely provide an increase of the model quality and completeness. Better models at early stages of model building provide useful information for the refinement, promoting an improvement of the electron density maps. This, in turn, can improve the identification of side-chains and the extension of fragments by the loop-building module.

1.7. Scope of This Thesis

The main aim of this thesis project is the development of computational methodologies within the ARP/wARP project for the improvement of the correctness and the accuracy of built models building at medium-to-low resolution. This was approached from two different sides: (1) research and development of a new general method for the validation of protein main-chain conformation that accounts for the coordinate error and the full range of degrees of freedom of trans-peptide Cα positions (*DipCheck*) and (2) technological implementation of novel developments into various ARP/wARP modules at medium-to-low resolution.

Although the development of DipCheck was performed within the ARP/wARP project, it can be used for general protein validation purposes. The project was triggered by the observation that the two Ramachandran-like angles represented by the Ramachandran-like plot are not informative enough, as they lack information about the angular geometry around each Cα in a dipeptide unit (Figure 3c). The improvement of ARP/wARP model building protocols at medium-to-low resolution was performed in several steps and in several points of the ARP/wARP protein model-building pipeline. Still in the scope of combined model building and validation, a small leap was also taken to the ligand Universe with a different method for the scoring of ligand molecules in electron density maps, which can be used for the validation of fitted ligands and as an additional scoring function during the fitting and guessing of ligands in density clusters.

Chapter 2

Introduction to Methodology

> "Molecules are the intellectual property of chemists and geometry is the province of mathematicians."
>
> — Tim Havel and Gordon Crippen

The development of computational methods to deal with molecular structural data relies on mathematical and computational approaches that facilitate the extraction of chemical and structural information as a set of unique parameters. From a simplistic point of view, molecules can be seen as graphs [175] (Figure 16): the atoms are the vertices of the graph and the bonds are the edges. The information about the molecule (e.g., topology, distances) can be stored in a matrix, which can be mathematically treated to obtain further information. This chapter presents an introduction to the mathematical concepts and tools used in this thesis.

2.1. Matrices, Eigenvalues and Eigenvectors

An $n \times m$ matrix M is a rectangular array of numbers arranged in n rows and m columns. If the number of rows and columns is the same, the matrix is said to be square. The individual items M_{ij}, in row i and column j in a matrix are called its elements. If $M_{ii} \neq 0$ and $M_{ij} = 0$, the matrix is called diagonal. If $M_{ij} = M_{ji}$, it is symmetric [176]. Given an $n \times n$ square matrix A, a set of scalars (eigenvalues) and vectors (eigenvectors) that highlight geometrical properties of the matrix can be calculated [177], [178]. A scalar λ is an eigenvalue of A if there is a non-zero column ($n \times 1$) vector $\boldsymbol{v}$ such that:

$$A\boldsymbol{v} = \lambda \boldsymbol{v} \tag{13}$$

$\boldsymbol{v}$ is then the eigenvector of matrix A associated with the eigenvalue λ. The eigenspectrum of the matrix A is the set of all its eigenvalues and has the same units as the elements of A. The absolute value of the largest eigenvalue is called the spectral radius $\rho(A)$ of the matrix (eq. 14) [178], [179]:

$$\rho(A) = max\{|\lambda_1|, |\lambda_2|, \dots, |\lambda_n|,\} \tag{14}$$

The decomposition of a matrix A into its eigenvalues and respective eigenvectors is called eigen-decomposition of A [178]. From equation 13, it follows that:

$$(A - \lambda I)\boldsymbol{v} = 0 \tag{15}$$

Where I is the $n \times n$ identity matrix. Equation 15 has a non-zero $\boldsymbol{v}$ solution if and only if the determinant of the matrix (A - λI) is zero and, therefore, the eigenvalues of A are the values of λ that satisfy equation 16:

$$|A - \lambda I| = 0 \tag{16}$$

This results in a polynomial function of degree n on the variable λ, called the characteristic polynomial of A (eq. 17):

$$\begin{gathered} c_0\lambda^n + c_1\lambda^{n-1} + \cdots + c_{n-1}\lambda + c_n = 0 \Leftrightarrow \\ \Leftrightarrow (\lambda - \lambda_1)(\lambda - \lambda_2) \dots (\lambda - \lambda_n) = 0 \end{gathered} \tag{17}$$

It has a maximum of n solutions and, therefore, a matrix of order n has a maximum of n different eigenvalues. The eigenvectors of A can then be computed with equation 15.

Geometrically, a matrix performs a linear transformation on vectors. An eigenvector corresponding to a real non-zero eigenvalue of a given matrix points in a direction that is re-scaled by the matrix (without rotation), with the eigenvalue as the factor of the re-scaling. If the eigenvalue is negative, the direction is reversed. For symmetric matrices (as the ones dealt with in this thesis), all eigenvalues are real and the respective eigenvectors are orthogonal to each other, spanning an n-dimensional real space encompassing all the information within the matrix [178].

The sum of all diagonal entries of a square matrix A is called the trace and it is proven to be equal to the sum of its eigenvalues (eq. 18, 19) [176], [180]:

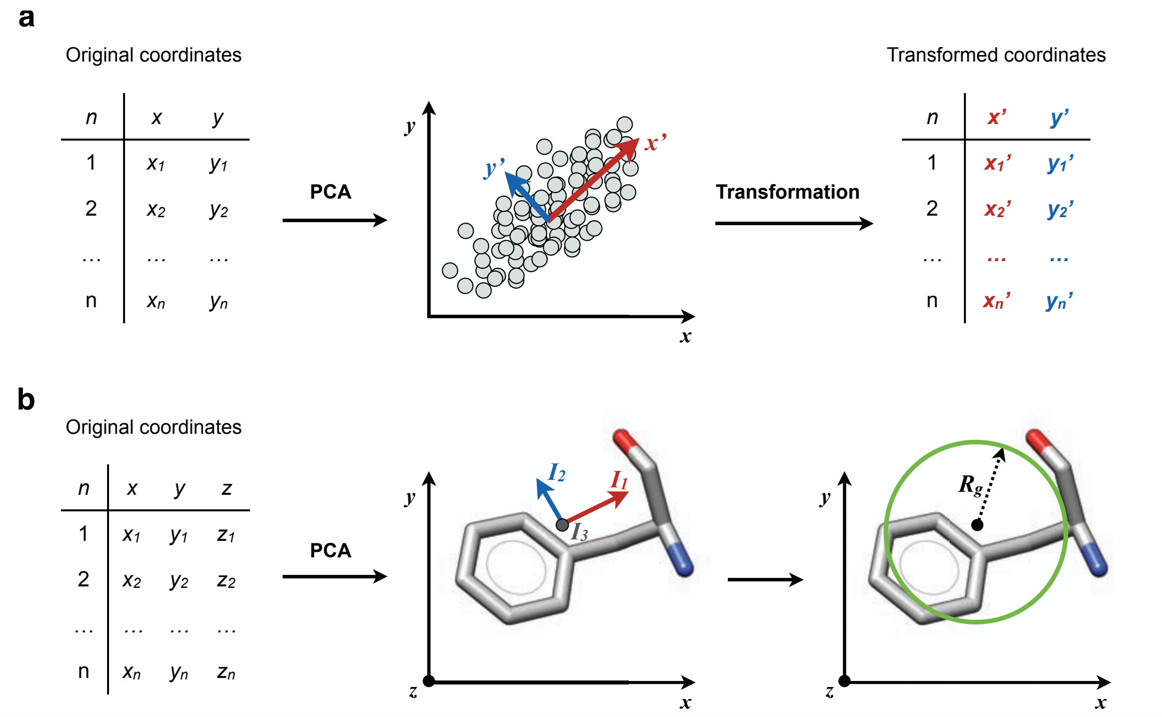
$$tr(A) = A_{11} + A_{22} + \cdots + A_{nn} = \sum_{i=1}^{n} A_{ii} \tag{18}$$

$$tr(A) = \lambda_1 + \lambda_2 + \cdots + \lambda_n = \sum_{i=1}^{n} \lambda_i \tag{19}$$

If a square matrix has all its diagonal entries equal to zero, it will have in general both positive and negative eigenvalues. If a matrix has at least one zero eigenvalue, it is referred to as singular and cannot be inverted [178]. The number of linearly independent rows (or columns) in a matrix is called rank and corresponds to the dimensionality of the Euclidean space that represents the matrix [181]. For square matrices the number of non-zero eigenvalues is equal to its rank [182].

2.2. Principal Component Analysis

Principal component analysis (PCA) is a statistical procedure that seeks the best summary of a dataset (followed by the second best and so on) by looking for new axes that can explain the maximum variance in the data [183]–[185]. The goal is to convert a set of correlated variables into a set of linearly uncorrelated variables, the principal components (Figure 15). The first principal component corresponds to the direction (vector) in the feature space along which the data vary most, with the second principal component giving the second direction, and so on. If the number of principal components is lower than the initial dimensionality of the data, the data can be transformed into a new coordinate system of uncorrelated variables with a lower number of dimensions. PCA can, thus, be used for data dimensionality reduction [185], [186].



n	x	y
1	x_1	y_1
2	x_2	y_2
…	…	…
n	x_n	y_n

n	x'	y'
1	x_1'	y_1'
2	x_2'	y_2'
…	…	…
n	x_n'	y_n'

n	x	y	z
1	x_1	y_1	z_1
2	x_2	y_2	z_2
…	…	…	…
n	x_n	y_n	z_n

Figure 15 Principal component analysis (PCA) over (a) a multi-dimensional (in this case, two-dimensional; *x,y*) data set for dimensionality reduction or (b) over atomic coordinates to identify the main axes of inertia I_i of a molecule and compute the radius of gyration R_g.

2.2.1. Getting a New Set of Axes

As reviewed by Morris [186], PCA is performed as the eigen-decomposition of the covariance matrix C of the data. Assuming a d-dimensional dataset X, with a total of d columns x_i, the covariance matrix is given as:

$$C = \begin{bmatrix} \mathrm{Cov}(x_1, x_1) & \cdots & \mathrm{Cov}(x_1, x_d) \\ \vdots & \ddots & \vdots \\ \mathrm{Cov}(x_d, x_1) & \cdots & \mathrm{Cov}(x_d, x_d) \end{bmatrix} \tag{20}$$

where the covariance Cov(·) can be estimated as the sample covariance cov(·):

$$\mathrm{Cov}(x_i, x_j) \approx \mathrm{cov}(x_i, x_j) = \frac{1}{n-1} \sum_{s=1}^{n} \left(x_{i_s} - \bar{x}_\imath\right) \left(x_{j_s} - \bar{x}_\jmath\right) \tag{21}$$

and $\bar{x}_i$ is the average value over all points, x_{i_s}, for the variable x_i. It is then a $d \times d$ square symmetric real matrix with a rank of d and eigenvalues λ_i and eigenvectors pc_i [186]:

$$\lambda_d \leq \lambda_{d-1} \leq \cdots \leq \lambda_2 \leq \lambda_1 \tag{22}$$

$$pc_d, pc_{d-1}, \ldots, pc_2, pc_1 \tag{23}$$

The eigenvector pc_1 with the largest eigenvalue λ_1 is the direction with the largest variance of the projected data and is, therefore, the first principal component. The eigenvector pc_2 with the second largest eigenvalue λ_2 is the second principal component, and so on [186]. When the eigenvalues of the covariance matrix are normalised by their sum, the explained variance of each principal component axis is obtained with equation 24:

$$\text{exp. variance}\left(pc_j\right) = \frac{\lambda_j}{\sum_{i=1}^{d} \lambda_i} \tag{24}$$

If it falls below a given threshold (e.g., 0.001; corresponding to 0.1% of the explained variance), pc_j can be neglected without a significant loss of information [186].

The data can be transformed into the new space described by the main principal components by rotation of the data [186]. If one places all the eigenvectors as rows in a matrix R, representing a $d\times D$ rotation matrix with d columns and D rows, the transformed data $X' = [x_1', \ldots, x_d']$ centred at the mean of each column of X can be obtained by applying the following operation (eq. 25):

$$X' = (X - \bar{X})R \tag{25}$$

where $\bar{X} = [\overline{x_1}, \ldots, \overline{x_d}]$ is the vector of all means of the elements of X.

2.2.2. Principal Component Analysis of Molecular Coordinates

PCA can be applied to the three-dimensional coordinates of a molecule to obtain information about conformational changes, protein-ligand interactions, etc. [187]. Let us consider a molecule M with n atoms $a = [x_1, y_1, z_1]$ in three-dimensional space (Figure 15b). We thus have a dataset made of three columns x, y, z with n lines. The covariance matrix C_M of this dataset is (eq. 26):

$$C_M = \begin{bmatrix} \text{cov}(x,x) & \text{cov}(y,x) & \text{cov}(z,x) \\ \text{cov}(x,y) & \text{cov}(y,y) & \text{cov}(z,y) \\ \text{cov}(x,z) & \text{cov}(y,z) & \text{cov}(z,z) \end{bmatrix} \tag{26}$$

a 3×3 square symmetric real matrix with a rank 3 for a three-dimensional object. The eigen-decomposition of C_M provides three eigenvectors, representing the three axis of variation of the coordinates. They are the three main principal components of the molecule, the three axes of inertia I_i, and the eigenvalues γ_i correspond to their squared lengths.

Given that the coordinates of the atoms in the molecule have Å units, the entries of C_M are in Å^2. The sum of its eigenvalues equals to the square of the molecule's radius of gyration R_g [188] (eq. 27):

$$R_g{}^2 = \sum_{i=1}^{3} \gamma_i \tag{27}$$

The R_g is a simple measure of the overall shape of an object (Figure 15b) and is defined as the root-mean-square distance from all its points to the centre of mass (eq. 28):

$$R_g = \frac{1}{n}\sqrt{\sum_{i=1}^{n} (\boldsymbol{a}_i - \overline{\boldsymbol{a}})^2} \tag{28}$$

Where $\overline{\boldsymbol{a}}$ are the coordinates of the centre of mass.

2.3. Distance Geometry and Molecular Conformation

The conformation of a molecule is defined by the relative position of all its atoms. If we know all distances between all atoms in a molecule, their relative positions and, therefore, the molecular structure can be derived (Figure 16b). This is the concept behind distance geometry [189] and is in widespread use in structural biology and chemistry, including NMR structure solution [189], protein structure prediction methods [190] and structure-based drug design [191]. However, different enantiomers of chiral molecules will present exactly the same inter-atomic distances. Thus, a distance-geometry based approach for the description of molecular conformation is always coupled with a chirality measure that is able to easily separate mirror-imaged conformations of the same molecule [189]. In the next sections the use of distances to describe molecular conformation will be discussed. Common chirality measures will be discussed later.

2.3.1. From Distances to Features

Euclidean distance matrices are $n \times n$ matrices representing the spacing of a set of n points in Euclidean space [189], [192], [193]. If D is a Euclidean distance matrix and the points $\boldsymbol{p_1}, \boldsymbol{p_2}, \boldsymbol{p_3}, \ldots, \boldsymbol{p_n}$ are defined in m-dimensional space, then the elements of D are given by:

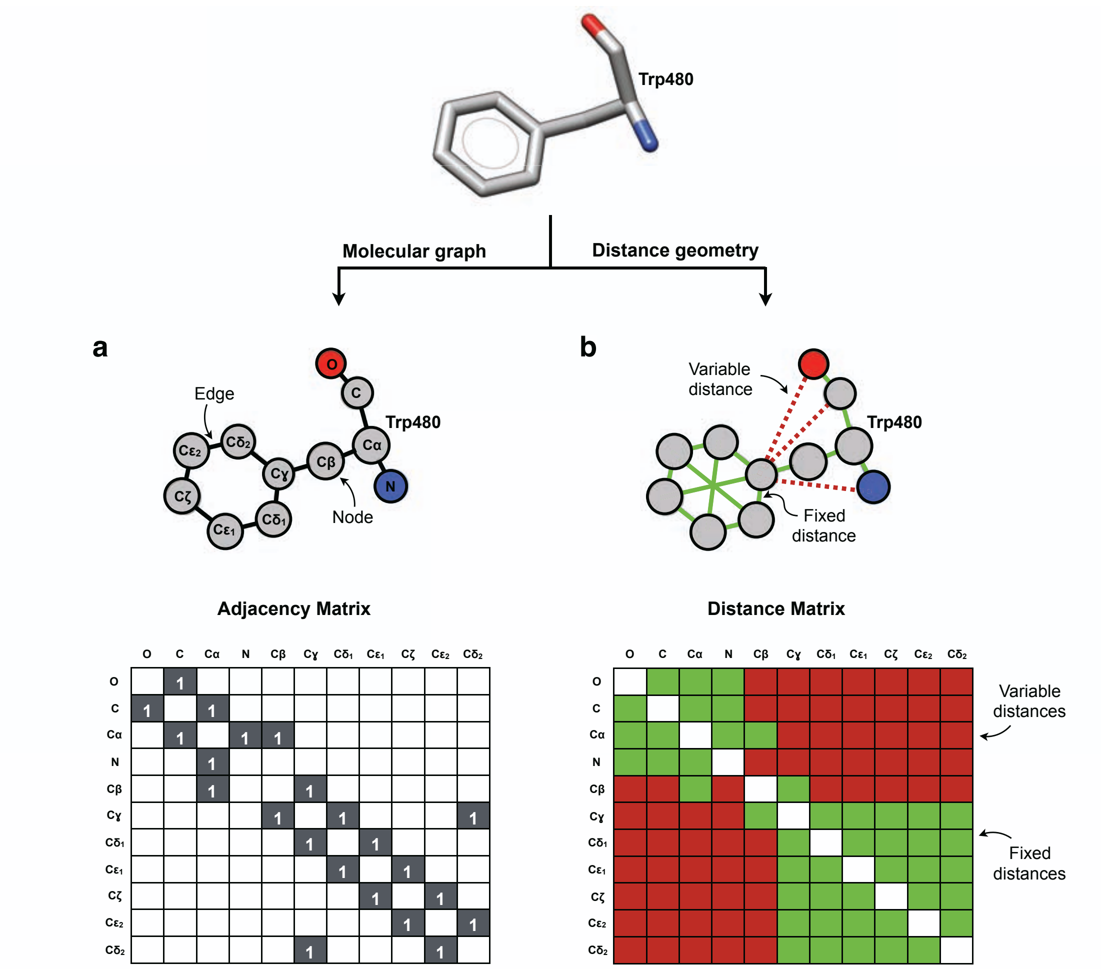
$$D = (a_{ij}) \tag{29}$$

$$a_{ij} = \left\|\boldsymbol{p_i} - \boldsymbol{p_j}\right\|^2 \tag{30}$$

where $\|.\|$ denotes the Euclidean norm on R^m. Therefore, the element a_{ij} describes the square of the distance between the i^{th} and j^{th} points, D is symmetric (i.e., $a_{ij} = a_{ji}$) and for any object in the 3-dimensional space, the elements of D are given by equation 31:

$$a_{ij} = \left(x_i - x_j\right)^2 - \left(y_i - y_j\right)^2 - \left(z_i - z_j\right)^2 \tag{31}$$

It is, therefore, a symmetric matrix with null diagonal and $\frac{n\times(n-1)}{2}$ unique entries.



	O	C	Cα	N	Cβ	Cγ	Cδ1	Cε1	Cζ	Cε2	Cδ2
O		1									
C	1		1								
Cα		1		1	1						
N			1								
Cβ			1			1					
Cγ					1		1				1
Cδ1						1		1			
Cε1							1		1		
Cζ								1		1	
Cε2									1		1
Cδ2						1				1	

	O	C	Cα	N	Cβ	Cγ	Cδ1	Cε1	Cζ	Cε2	Cδ2
O		🟩	🟩	🟩	🟥	🟥	🟥	🟥	🟥	🟥	🟥
C	🟩		🟩	🟩	🟥	🟥	🟥	🟥	🟥	🟥	🟥
Cα	🟩	🟩		🟩	🟩	🟥	🟥	🟥	🟥	🟥	🟥
N	🟩	🟩	🟩		🟥	🟥	🟥	🟥	🟥	🟥	🟥
Cβ	🟥	🟥	🟩	🟥		🟩	🟥	🟥	🟥	🟥	🟥
Cγ	🟥	🟥	🟥	🟥	🟩		🟩	🟩	🟩	🟩	🟩
Cδ1	🟥	🟥	🟥	🟥	🟥	🟩		🟩	🟩	🟩	🟩
Cε1	🟥	🟥	🟥	🟥	🟥	🟩	🟩		🟩	🟩	🟩
Cζ	🟥	🟥	🟥	🟥	🟥	🟩	🟩	🟩		🟩	🟩
Cε2	🟥	🟥	🟥	🟥	🟥	🟩	🟩	🟩	🟩		🟩
Cδ2	🟥	🟥	🟥	🟥	🟥	🟩	🟩	🟩	🟩	🟩	

Figure 16 Graph representation of Phe480 from leyshmanolysin (PDB ID 1lml) and the two types of matrices used in this thesis. (a) Molecular graph concept, showing the nodes and edges and the adjacency matrix computed. (b) Distance geometry concept, highlighting some variable (dashed red) and fixed (green) distances and the distance matrix computed.

For a given molecule, we can consider two types of distances: “fixed” and “variable” (Figure 16b). Fixed distances are formed between chemically bonded atoms that do not vary significantly from a reference value [123]. These may include atoms that are not forming a chemical bond with each other. For example, the atoms in the peptide plane are constrained by the planarity of the peptide bond and, therefore, their interatomic distances do not vary freely (Figure 2). The same applies to atoms in aromatic rings (Figure 16b). Variable distances between atoms are those that alter upon conformational changes.

Any $n \times n$ distance matrix (Euclidean or not) has one positive and $n-1$ negative eigenvalues and the rank of any Euclidean distance matrix D in m-dimensional space is given by eq. 32 [194]:

$$rank(D) \leq m + 2 \tag{32}$$

Hence, the maximum rank of any $n \times n$ Euclidean distance matrix in 3-dimensional space is 5, which means that there are 5 or less non-zero eigenvalues. For a non-planar molecule, there are 4 negative eigenvalues of matrix *D* ($\lambda_1, \lambda_2, \lambda_3, \lambda_4$) and 1 positive eigenvalue (λ_{max}). Given that all elements on the diagonal of *D* are zero, the trace of *D* is zero and the absolute sum of all 4 negative eigenvalues of *D* is equal to the positive eigenvalue (eq. 33):

$$\begin{gathered} trace(D) = \lambda_1 + \lambda_2 + \lambda_3 + \lambda_4 + \lambda_{max} \\ \Rightarrow |\lambda_1| + |\lambda_2| + |\lambda_3| + |\lambda_4| = \lambda_{max} \end{gathered} \tag{33}$$

with

$$|\lambda_1| > |\lambda_2| > |\lambda_3| > |\lambda_4| \tag{34}$$

Therefore, at least some of the information contained in the matrix *D* can be conveniently described by the four negative eigenvalues. Euclidean matrices can be calculated for any molecule and their eigenvalues and corresponding eigenvectors used as descriptors of molecular geometry [166], [190].

2.3.2. From Distances to Coordinates

The computation of the three-dimensional coordinates of each atom in a molecule can be preformed by a linear transformation of the Euclidean distance matrix, generating a *Gram matrix G* (eq. 35) [195]. Considering *P* as the matrix encompassing each $\boldsymbol{p_1}, \boldsymbol{p_2}, \boldsymbol{p_3}, \ldots, \boldsymbol{p_n}$ as defined in the previous section, with $P = [\boldsymbol{p_1}, \boldsymbol{p_2}, \boldsymbol{p_3}, \ldots, \boldsymbol{p_n}\ ]$, *G* is expressed as:

$$G = P^T P = \begin{bmatrix} \boldsymbol{p_1}^T \\ \vdots \\ \boldsymbol{p_n}^T \end{bmatrix} [\boldsymbol{p_1} \quad \cdots \quad \boldsymbol{p_n}] = \begin{bmatrix} \|\boldsymbol{p_1}\|^2 & \boldsymbol{p_1}^T\boldsymbol{p_2} & \boldsymbol{p_1}^T\boldsymbol{p_3} & \cdots & \boldsymbol{p_1}^T\boldsymbol{p_n} \\ \boldsymbol{p_2}^T\boldsymbol{p_1} & \|\boldsymbol{p_2}\|^2 & \boldsymbol{p_2}^T\boldsymbol{p_3} & \cdots & \boldsymbol{p_2}^T\boldsymbol{p_n} \\ \boldsymbol{p_3}^T\boldsymbol{p_1} & \boldsymbol{p_3}^T\boldsymbol{p_2} & \|\boldsymbol{p_3}\|^2 & \cdots & \boldsymbol{p_3}^T\boldsymbol{p_n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \boldsymbol{p_n}^T\boldsymbol{p_1} & \boldsymbol{p_n}^T\boldsymbol{p_2} & \boldsymbol{p_n}^T\boldsymbol{p_3} & \cdots & \|\boldsymbol{p_n}\|^2 \end{bmatrix} \tag{35}$$

From expanding the norm in equation 30, the entries a_{ij} of the Euclidean distances matrix *D* can also be expressed as [193]:

$$a_{ij} = (\boldsymbol{p_i} - \boldsymbol{p_j})^T(\boldsymbol{p_i} - \boldsymbol{p_j}) = \|\boldsymbol{p_i}\|^2 - \boldsymbol{2}\boldsymbol{p_i}^T\boldsymbol{p_j} + \|\boldsymbol{p_j}\|^2 \tag{36}$$

Therefore, the entries g_{ij} of the Gram matrix *G* and the entries a_{ij} of the Euclidean distances matrix *D* relate by equation 37:

$$a_{ij} = g_{ii}{}^2 - 2g_{ij} + g_{ij}{}^2 \tag{37}$$

The matrix *G* can be obtained from matrix *D* by equation 38, assuming that p_1 is at the origin [193]:

$$G = -\frac{1}{2}(D - 1a_1{}^T - a_1 1^T) \tag{38}$$

where a_1 is the first column of *D* and *1* the column vector of all ones.

Equation 38 allows the computation of the Gram matrix from the distances alone by considering that $\boldsymbol{p_1}$ is at the origin, and is the eigen-decomposition of G that allows the computation of the point set P. G is an $n \times n$ square symmetric matrix with all eigenvalues λ_i being non-negative and $\lambda_1, \lambda_2 > \ldots > \lambda_n$. It has as many non-zero eigenvalues as the dimensionality m of the point set P, and:

$$P = \begin{bmatrix} \lambda_1 & 0 & \ldots & 0 \\ 0 & \lambda_2 & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & \lambda_d \end{bmatrix} U^T \tag{39}$$

where d is the number of non-zero eigenvalues and U is the set of eigenvalues of G [193].

2.4. Estimation of Molecular Chirality

Chirality is a geometrical property of a molecule [1]. Four different substituents bonded to a tetrahedral (sp^3) carbon can have two different configurations (Figure 17), yielding two enantiomers that may have similar chemical properties but differ in physical and biological properties. These atoms are classified as asymmetric and, therefore, are referred to as chiral centres. They show exactly the same topology and inter-atomic distances (and, consequently, distance and Gram matrices) but are non-superimposable, as they are the mirror images of each other.

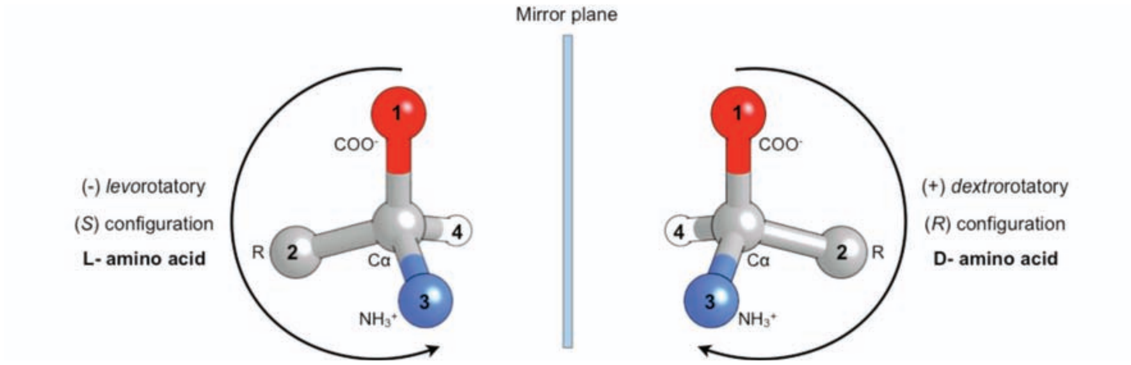


Figure 17 The two generic enantiomer forms of the asymmetric Cα atom in a general amino acid (not applicable to glycine). The priority given to each of the four substituents (1 is higher; 4 is lower), with an arrow depicting the direction of priority decrease [196], and the handedness of each enantiomer according to each nomenclature system is shown. (+) and (-) depict a clockwise or counter-clockwise rotation.

A molecule with only one chiral centre can have two enantiomers; when two or more (n) are present, there can be 2^n enantiomers. Enantiomers have nearly identical chemical reactivity but differ in a characteristic physical property; they rotate the plane of polarised light in opposite directions, while molecules without chiral centres do not [197]. Enantiomers that rotate polarised light to the left are referred as levorotatory (L-enantiomer) while those to the right as dextrorotatory (D-enantiomer), with this direction referred as their *handedness*. Given the chemical importance of stereochemistry, Cahn, Ingold and Prelog developed a set of priority rules that help on the nomenclature of different enantiomers [196]. Here, to each group attached to a chiral centre a priority is assigned, from 1 (highest) to 4 (lowest). If the priority of the groups reduces (from 1 to 4) in clockwise order, the configuration is (*R*), as from the Latin *rectus*; if in counter-clockwise order, the configuration is (*S*), as from the Latin *sinister*. A molecule with a single chiral centre, as an amino acid, can be named by either convention (Figure 17) [1]. Mathematically, enantiomers can be distinguished by several methods [189], [198].

2.4.1. Oriented and Chiral Volumes

The oriented and chiral volumes of asymmetric atoms are the simplest methods to distinguish between the two enantiomeric forms of a chiral centre. They are defined for a set of four points (e.g., atoms) and correspond to the volume of the tetrahedron defined by them (e.g., adjoining the central one) (oriented volume; Figure 18a) or to the volume of the parallelepiped formed when each of them mark its vertices (chiral volume; Figure 18b), with their sign representing the handedness of the asymmetric atom [189]. Although conceptually the same, they are computed differently and differ in their absolute values and signs.

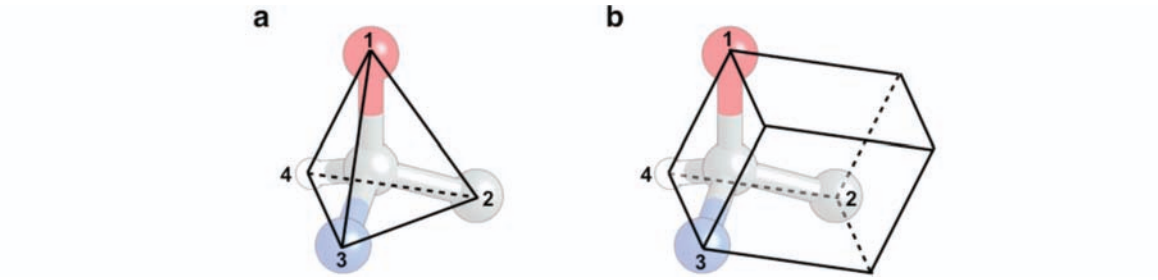


Figure 18 The solids for which the volume is computed with the (a) oriented volume and the (b) chiral volume, formed by a set of four substituent groups around an asymmetric chiral center.

The *oriented volume* (V_O) of a given central tetrahedral asymmetric centre is calculated as a determinant (eq. 40):

$$V_O(\boldsymbol{p_1}, \boldsymbol{p_2}, \boldsymbol{p_3}, \boldsymbol{p_4}) = \frac{1}{3!} det \begin{bmatrix} 1 & 1 & 1 & 1 \\ x_1 & x_2 & x_3 & x_4 \\ y_1 & y_2 & y_3 & y_4 \\ z_1 & z_2 & z_3 & z_4 \end{bmatrix} \tag{40}$$

With $\boldsymbol{p_i}$ = [x_i, y_i, z_i] being the vector that assigns to a substituent atom with a priority *i* its Cartesian coordinates. Its absolute value is independent on the particular embedding *p* used to define it but its sign provides a criterion to obtain the enantiomer handedness. The oriented volume is invariant under rotations and translations, the only way to change its sign is by reflecting the molecule by a plane without changing the orientation of the coordinate system. By definition, the coordinate system is seen as dextrorotatory and, with that, D-enantiomers have a positive (+) oriented volume and L-enantiomers a negative (-) one [189].

The *chiral volume* (V_C) is usually used as a stereochemical restraint in macromolecular refinement and computational chemistry [199], [200]. For the same system of points, it is given by equation 41:

$$V_C(\boldsymbol{p_1}, \boldsymbol{p_2}, \boldsymbol{p_3}, \boldsymbol{p_4}) = (\boldsymbol{p_1} - \boldsymbol{p_4}) \cdot [(\boldsymbol{p_2} - \boldsymbol{p_4}) \times (\boldsymbol{p_3} - \boldsymbol{p_4})] \tag{41}$$

Its absolute value is then six times the absolute of the oriented volume, and their signs are inversely related. If the labels of the substituents are in concordance with the Cahn-Ingold-Prelog system, the chiral volume will be negative (-) for the D-enantiomers and positive (+) for the L-enantiomers [200].

2.4.2. Chiral Index and Chiral Invariant

Chiral molecules can have more than one chiral centre. If there are two such centres and all interatomic distances are known, only the sign of one chiral volume is necessary to define the chirality of the molecule [189], but if more than two are present, this can be more complicated. A chirality measure can be viewed as a tool for quantifying the difference in shape between the object and its mirror image [201]. The *universal chiral index* (G_O), derived by Osipov et al. [202], measures the chirality of a given object by an analogy with the optical activity of different enantiomers and has been used in the classification of protein secondary structure of short three-dimensional peptide fragments [203], [204]. Representing the molecule by a density distribution $\rho(\boldsymbol{r})$, a set of delta functions for a molecule consisting of point atoms, a universal chiral index is computed by the integration over all possible combinations of sets of four points in space, $\boldsymbol{r_1}$, $\boldsymbol{r_2}$, $\boldsymbol{r_3}$ and $\boldsymbol{r_4}$ (eq. 42):

$$\begin{gathered} G_O = \int \frac{(\boldsymbol{r}_{12} \times \boldsymbol{r}_{34} \cdot \boldsymbol{r}_{14})(\boldsymbol{r}_{12} \cdot \boldsymbol{r}_{23})(\boldsymbol{r}_{23} \cdot \boldsymbol{r}_{34})}{(r_{12} r_{23} r_{34})^a {r_{14}}^b} \\ \times \rho(\boldsymbol{r}_1)\rho(\boldsymbol{r}_2)\rho(\boldsymbol{r}_3)\rho(\boldsymbol{r}_4) d\boldsymbol{r}_1 d\boldsymbol{r}_2 d\boldsymbol{r}_3 d\boldsymbol{r}_4 \end{gathered} \tag{42}$$

where $\boldsymbol{r_i} = [x_i, y_i, z_i]^T$, $\boldsymbol{r_{ij}} = \boldsymbol{r_i} - \boldsymbol{r_j}$, $r_{ij} = \|\boldsymbol{r_{ij}}\|$, and a and b arbitrary integers. When $a = 2$ and $b = 1$, G_O is dimensionless. It changes sign under the space inversion and, therefore, is zero for achiral objects (which are invariant under this inversion). It is negative for D-enantiomers and positive for L-enantiomers.

The calculation of a universal chiral index is limited by the size of the object and the number of point atoms in the molecule. In order to overcome this, Hattne and Lamzin [198] derived the *chiral invariant* (*CI*), which is based on the use of moment invariants computed from the object, also represented as a density function. For any non-negative integers *l*, *m* and *n*, the raw moments

(M_{lmn}) of order $l+m+n$ of a three-dimensional density distribution function $\boldsymbol{\rho}(x,y,z)$ are computed as (eq. 43):

$$M_{lmn} = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} x^l y^m z^n \boldsymbol{\rho}(x,y,z)\, dxdydz \tag{43}$$

The central moments (μ_{lmn}), invariant under translation, are then taken with reference to the mean of the object (eq. 44):

$$\mu_{lmn} = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} (x-\bar{x})^l (y-\bar{y})^m (z-\bar{z})^n \boldsymbol{\rho}(x,y,z)\, dxdydz \tag{44}$$

The chiral invariant *CI* is expressed then as a sum of products of four moments (eq. 45):

$$CI = \sum \mu_{l_1m_1n_1}\, \mu_{l_2m_2n_2}\mu_{l_3m_3n_3}\mu_{l_4m_4n_4} \tag{45}$$

The chiral invariant is invariant under translation, rotation and scale of the object, and is zero for achiral objects. As the chiral index, it is negative for D-enantiomers and positive for L-enantiomers. It was shown to be useful for the understanding of protein backbone handedness and for an assessment of the quality of crystallographic electron density maps [198]. Therefore, the magnitude of both these chiral measures can serve as a description of the overall shape of the object, while their sign distinguishes two different mirror-imaged forms [198].

2.5. Graph Spectra and Molecular Topology

The topology of a molecule describes how atoms are forming chemical bonds between each other. It does not provide information about the three-dimensional structure of the molecule but about the connectivity and can be used to estimate chemical properties of the molecule itself [205]. The topology of a molecule can be represented as a graph (Figure 16a). Mathematically, a graph *G* can be defined as a pair (*V,E*) where *V* is a set of vertices representing the nodes (e.g, the atoms of the molecule) and *E* is a set of edges representing the connections between the nodes (e.g., the chemical bonds) [206]. Each element of *E* contains a pair *i,j* of elements of *V*. Two nodes are said to be adjacent if they are joined by an edge and two edges are adjacent if a node joins them. The degree of a node *v* is equal to the number of edges incident in it.

According to the relationships between edges and nodes, graphs can be classified into different categories (Figure 19) [207]. Graphs can be said to be directed or undirected if the direction of the edges is important or not, respectively. Graph nodes can be labelled or unlabelled and also be classified according to their degree. A complete graph is a graph such that every pair of vertices is joined by an edge, containing therefore all possible edges and a degree equal to the number of nodes minus 1. A regular graph is a graph in which each node has exactly the same number of edges and thus the same degree. A connected graph is a graph that has at least one edge connecting all nodes. A simple graph is an undirected graph in which multiple edges connecting the same pair of nodes and loops connecting one node to itself are disallowed. A bipartite graph is a graph whose vertices can be divided into two disjoint sets such that every edge connects a vertex in each set. A molecular graph is, therefore, a simple connected undirected labelled graph, where each node has its own identity (Figure 16a).

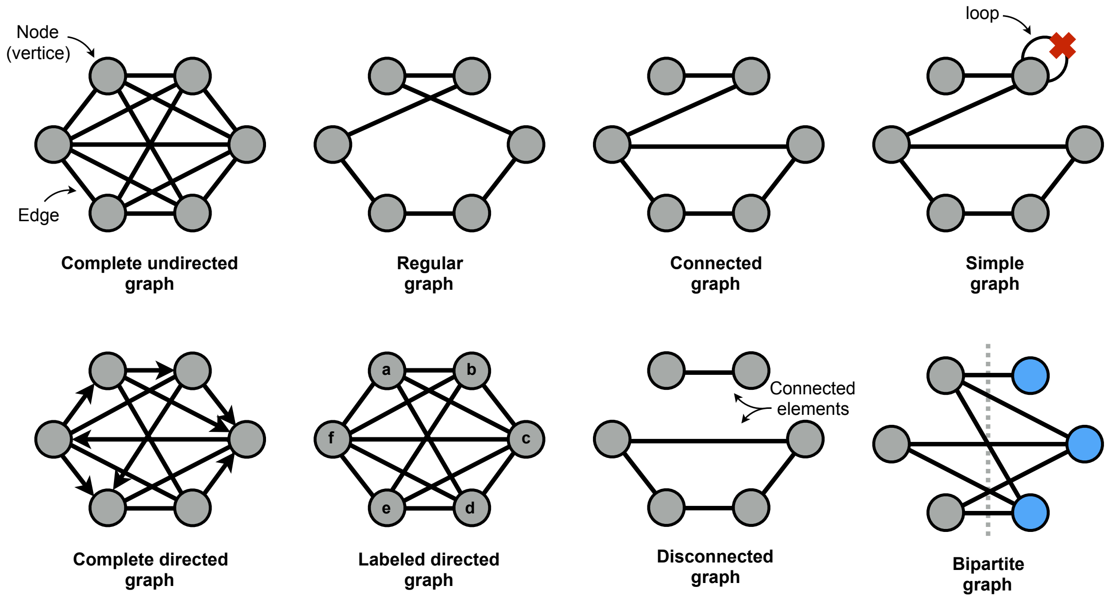


Figure 19 Example classes of graphs, as described in the text.

A graph can be represented using a set of different matrices, whose eigenspectra provide information about the graph structure [207]. The two main matrices used in graph theory for the representation and study of graphs are introduced below.

2.5.1. Adjacency Matrix

An adjacency matrix A is a matrix used to represent a finite graph, in which each entry indicates whether a pair of vertices is adjacent or not [208]. In the case of a finite simple undirected graph, such as a molecular graph, it is a symmetric (0,1)-matrix (eq. 46), where:

$$A_{ij} = \begin{cases} 1 & \text{if } v_i \rightarrow v_j \\ 0 & \text{otherwise} \end{cases} \tag{46}$$

For a graph with n nodes it is, therefore, an $n \times n$ symmetric real matrix with zeros on its diagonal (Figure 16). The set of individual eigenvalues, and their respective multiplicity, of an adjacency matrix A of a graph G is referred to as the spectrum of G, and is commonly used to evaluate the isomorphism of two graphs [209]. Two isomorphic graphs have the same spectra (they are isospectral), but two isospectral graphs do not necessarily need to be isomorphic. Relabeling the graph nodes does not alter its structure and, therefore, its set of eigenvalues and eigenvectors remain the same.

The adjacency matrix of a graph G with n number of vertices has a maximum of n non-zero real eigenvalues, which sum up to zero [210], with:

$$\lambda_n \leq \lambda_{n-1} \leq \cdots \leq \lambda_2 \leq \lambda_1 \tag{47}$$

Any eigenvalue of A lies in the interval [-d_{max}, d_{max}], where d_{max} is the maximum degree of any node in the graph and, therefore, equal to n-1 [210]. The largest eigenvalue (λ_1) is the spectral radius of the graph and relates to its average degree by equation 48 [211]:

$$\bar{d} \leq \lambda_1 \leq d_{max} \tag{48}$$

Where $\bar{d}$ is the average degree of all nodes. If the graph is complete, $\lambda_1 = d_{max}$. The multiplicity of λ_1 relates to the connectivity of the graph, with the graph being connected if λ_1 has a multiplicity of 1 (e.g., only appears once) [212].

The difference between λ_1 and λ_2 is called the spectral gap and is always smaller or equal to the number of real eigenvalues n (eq. 49) [210]:

$$\lambda_1 - \lambda_2 \leq n \tag{49}$$

The spectral gap characterises the robustness of the graph due to its relation to the algebraic connectivity. It is closely related to the number of graph cuts, the number of partitions of the graph vertices necessary to create two disjoint subsets of nodes. The smaller this value, the fewer the vertexes need to be removed in order to create a disconnected graph. Therefore, if the spectral gap is close to zero, the multiplicity of λ_1 is close to two and the graph is disconnected. The same way, the larger the spectral graph, the higher the connectivity of the graph [213].

The sum over all spacings between two consecutive eigenvalues equals to the difference between the largest and the smallest (λ_n) eigenvalues (eq. 50) [210]:

$$\sum_{i=1}^{n-1} (\lambda_i - \lambda_{i+1}) = \lambda_1 - \lambda_n \tag{50}$$

This is useful because a graph is bipartite if and only if its spectrum is symmetric about the origin, which means [211]:

$$\lambda_n = -\lambda_1 \tag{51}$$

With the sum over all spacings equalling $2\lambda_1$. The ratio between the largest and the smallest eigenvalues of connected graphs can be used to estimate the lower boundary of the chromatic number of G, $\chi(G)$, by (eq. 52) [214], [215]:

$$1 - \frac{\lambda_1}{\lambda_n} \leq \chi(G) \leq 1 + \lambda_1 \tag{52}$$

The chromatic number of a graph represents the smallest number of colours one would need to colour the nodes of *G* so that no two adjacent nodes share the same colour. It then represents the number of partitions, with a bipartite graph having a chromatic number of 2.

Finally, the number of distinct eigenvalues of *A* (*N*) can be used to estimate the diameter (*ρ*) of the graph *G* (eq. 53) [209], [210]:

$$\rho \leq N - 1 \tag{53}$$

The diameter of a graph *G* is the "longest shortest path" between any two nodes in the graph. In other words, it is the largest number of nodes that must be traversed in order to travel from one node to another when paths that backtrack, detour or loop are excluded from consideration.

2.5.2. Laplacian Matrix

The Laplacian matrix *L* of an undirected and unweighted simple graph *G* is an $n \times n$ symmetric matrix with one row and column for each node defined by (eq. 54) [209]:

$$L = D - A \tag{54}$$

where *A* is the adjacency matrix of *G* and *D* the degree matrix. *D* is a diagonal $n \times n$ matrix that contains information about the degree of each vertex (eq. 55), where:

$$D_{ij} = \begin{cases} \deg(v_i) & \text{if } i = \text{j} \\ 0 & \text{otherwise} \end{cases} \tag{55}$$

The diagonal elements L_{ii} of *L* are, therefore, equal to the degree of vertex v_i, off-diagonal elements L_{ij} are -1 if vertex v_i is adjacent to vertex v_j and zero otherwise (eq. 56):

$$L_{ij} = \begin{cases} \deg(v_i) & \text{if } i = j \\ -1 & \text{if } v_i \rightarrow v_j \\ 0 & \text{otherwise} \end{cases} \tag{56}$$

The Laplacian matrix measures the extent to which a graph differs at one vertex from its values at nearby vertices. For a graph with multiple connected components, *L* is a block-diagonal matrix, where each block is the respective Laplacian matrix for each component.

The Laplacian matrix of a graph *G* with *n* number of vertices has a maximum of *n*-1 non-zero real eigenvalues with:

$$\lambda_n \leq \lambda_{n-1} \leq \cdots \leq \lambda_2 \leq \lambda_1 \tag{57}$$

All eigenvalues of *L* are equal to or larger than zero and their sum is equal to twice the number of edges *m* of *G* (eq. 58):

$$tr(L) \neq 0 \Rightarrow tr(L) = \sum_{i=1}^{n} \lambda_i = 2m \tag{58}$$

Every row sum and column sum of L is zero. In consequence, $\lambda_n = 0$. The multiplicity of the null eigenvalue corresponds to the number of connected components in the graph [209]. A connected component of an undirected graph is a subgraph in which two vertices are connected to each other by paths and which is not connected to any additional vertices (Figure 19). A connected graph will, therefore, have only one zero eigenvalue. The second smallest eigenvalue is the algebraic connectivity, representing how well connected the overall graph is, and is non-zero only for connected graphs. The smallest non-zero eigenvalue is the spectral gap [216].

Chapter 3

Development of New Protein Main-Chain Conformational Descriptors

The use of only two parameters for the conformational description of each residue in the protein main-chain, as represented by the Ramachandran and Ramachandran-like plots, is insufficient [44], [50], [217]. When the polypeptide chain is seen as a combination of peptide planes connected at each Cα position and its conformation is described by the rotation of the two adjacent peptide planes, this dipeptide unit has three degrees of freedom (Figure 20a). Each of the nine atoms composing any dipeptide unit has a three-dimensional coordinate, totalling to $9 \times 3 = 27$ parameters. Excluding rigid-body rotations and translations, these reduce to $27 - 3 - 3 = 21$. The lengths of the four covalent bonds in each peptide plane (Cα-C', C'-O, C'-N and N-Cα), and four angle-bonded distances can, from a conformational point of view, be seen as fixed [123] thus reducing the number of degrees of freedom to $21 - 2 \times (4 + 4) = 5$. Assuming the planarity of the peptide unit, or that the ω does not vary much from a standard value, we arrive to the two possible configurations of the peptide plane: the trans ($\omega \approx 180°$) and the cis ($\omega \approx 0°$) (Figure 2).

Given that the trans-configuration is the more abundant, encompassing more than 99% of the peptide planes in the PDB [8], [9], [11], the number of degrees of freedom of a dipeptide unit reduces to three. If the first two are the Ramachandran dihedral angles, then the third degree of freedom is related to a variation of the τ(N-Cα-C') stretching angle (Figure 20a). This angle in refined protein structures varies in a narrow region, from 107.5° to 114.0° [50] and, therefore, could have been regarded of low importance. However, its value depends on the secondary structure to which the dipeptide unit belongs to [50] and on the chemical nature of the residue's side-chain [217].

Analogously, 5-atom dipeptide units used by ARP/wARP to assemble protein main-chain also have 3 degrees of freedom: the two dihedral angles and the $\tau_d(O_{i-1}\text{-}C\alpha_i\text{-}O_i)$ stretching angle (Figure 20b). Since the τ_d angle varies in a broader range compared to the τ angle due to the different vectors involved, it would be expected to be very informative for the description of the dipeptide units' conformation. However, the disadvantage of a direct use of the stretching angle is its dependence on the values of the two dihedral angles [50].

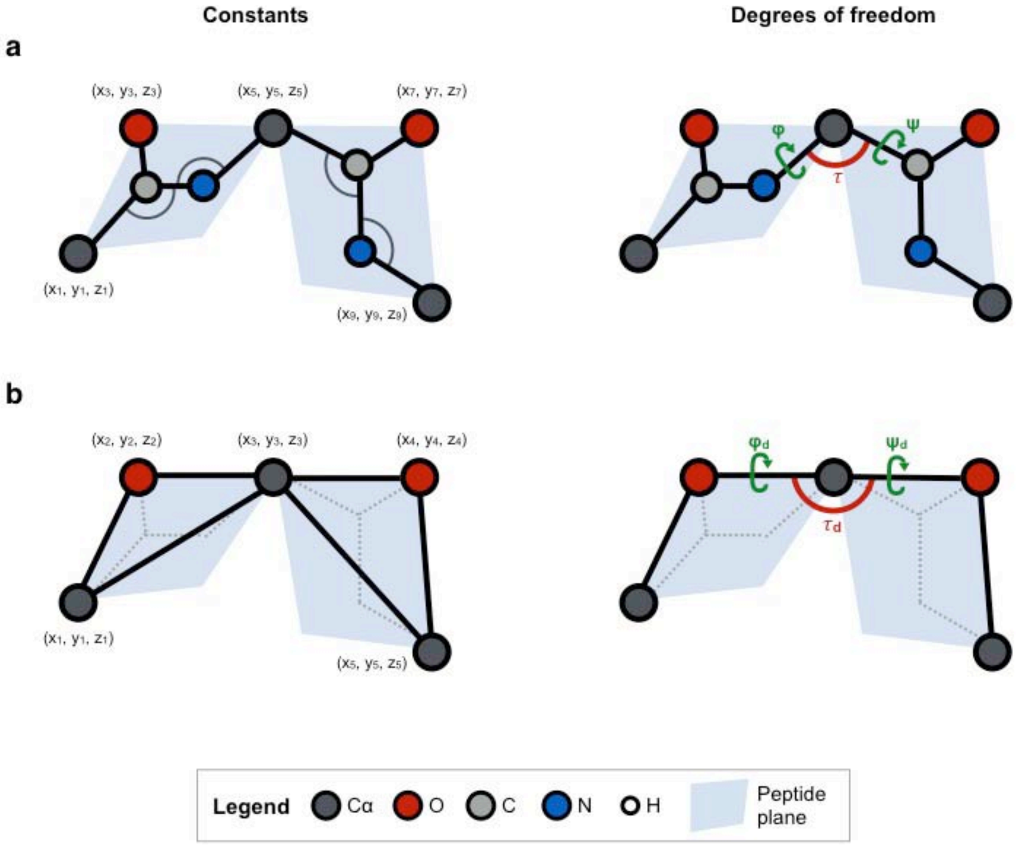


Figure 20 The degrees of freedom of (a) full atom and (b) 5-atom (in trans configuration) dipeptide units. Black lines mark the fixed distances and angles. Green arrows represent the degrees of freedom described by the (a) Ramachandran and (b) Ramachandran-like plots. Red arrows represent the additional degree of freedom of dipeptide units in the trans- configuration.

Consequently, novel descriptors of protein geometry that would be independent or at least weakly correlated with each other were investigated. The dipeptide unit was looked at from a different perspective, independent from angles. The applicability of a distance geometry-based approach, grounded on the eigenvalues of Euclidean distance matrices as described in 2.3, for the derivation of a set of uncorrelated dipeptide unit conformational descriptors was then studied, as described in the next sections. The main focus was given to the 5-atom model of a dipeptide unit for two reasons: (1) this is the simplest model of a dipeptide unit and (2) this is the model used by ARP/wARP during main-chain tracing, so that developed results could be more straightforwardly incorporated into a computational tool.

This investigation was performed in three steps: (1) the collection of a reliable set of dipeptide units from the PDB; (2) the identification of three uncorrelated geometrical descriptors able to discriminate dipeptide units with a different geometry; and (3) the identification of the best chirality measure to separate between dipeptide units that are mirror-images of each other.

3.1. Methods

3.1.1. Assembly of A Set of Dipeptide Units

For the collection of a set of dipeptide units, they were required to represent well the conformations present in the PDB and the structures used to extract them to be of sufficient geometrical quality. Protein chains with a pairwise sequence identity below 50% were obtained from the PDB [36] (as of September 30, 2014) using the PDB50 clusters [218], according to Table 1. For each chain, a DSSP file, containing its secondary structural content, was downloaded from the DSSP databank [219], [220]. Each selected protein chain was then broken into 5-atom dipeptide units, excluding those with atoms having high positional uncertainty or being in implausible conformations, based on the dipeptide units' interatomic distances (Table 1).

Table 1 Selection criteria for protein chain and dipeptide sampling from the PDB.

Selection parameter	Criterion
Protein chains	
Experimental method	X-ray crystallography
Resolution	Better than 2.5 Å
R-factor	Lower than 0.25
R-free – R-factor difference	Lower than 0.05
Clashscore and Ramachandran Outliers Percentiles	Higher than 40%
Dipeptides	
Occupancy for Cα and O atoms	Equal to 1.00
ADP for Cα and O atoms	Lower than 80 Å^2
Fixed distances	$\mu \pm 3\sigma$
Variable distances	Interval with 99.8% of the points

Filtering by bond-angle (fixed) distances

The "fixed distances" between atoms in the same peptide plane (Figure 21a), due to the chemical constraints affecting atomic bond geometry, should not vary much from their expected values, and these variations are expected to follow a Gaussian distribution. Bond lengths (as well as bond angles) and their respective standard deviation were already surveyed and are used as standard restraints in protein crystallography [123], [221]. However, in a 5-atom dipeptide unit, the atoms are 'angle-bonded'. Therefore, the expected angle-bonded distances were estimated using the tabulated bond distances and angles surveyed by Engh and Huber in 2006 [123] and then compared to those estimated by different statistical methods.

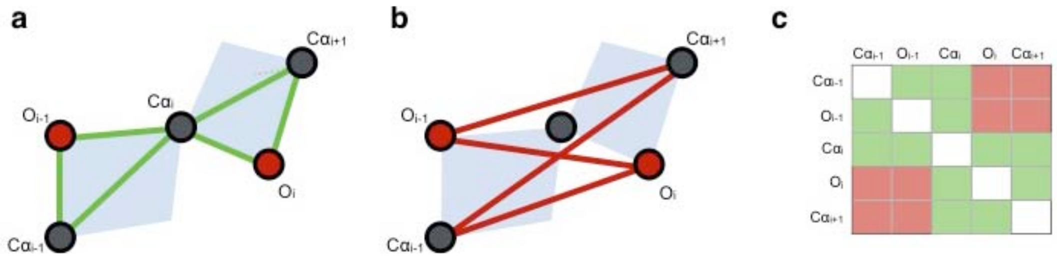


Figure 21 Distance geometry-based approach for the description of dipeptide unit geometry. 5-atom dipeptide units can be described by a combination of two different classes of distances: (a) fixed and (b) flexible distances. (c) When introduced into a distance matrix, these distances define two regions: the rigid region, including the information on coordinate error (green), and the flexible region, including the information on main-chain conformation (red).

For each of the peptide planes in the initially collected set of dipeptide units, all fixed distances were computed. Their mean, estimated standard deviation (SD), median and median absolute deviation (MAD) [222] were also computed. MAD is a very robust scale estimator, and is not much affected by the presence of extreme values in the data, compared to the standard deviation σ. Depending on the population distribution, the MAD value is multiplied by a consistency constant *b* (1.4826 for Gaussian distributions) being then referred to as MADe [222]. Given that in a Gaussian distribution approximately 99.8% of the data lie within 3σ from the mean, the 3SD and 3MADe intervals for each distance distribution were also calculated.

Additionally, the *normalmixEM* function from the *mixtools* R-package was used to investigate whether the fixed distance distributions could be described as a mixture of Gaussian distributions [223]. This function uses the EM (Expectation-Maximization) algorithm for Gaussian mixtures and starts by estimating a complete set of parameters for the given model and then proceeds by iteratively updating them until convergence. From this, the mean (μ), standard deviation (σ) and final mixing proportions (ν) for each Gaussian distribution composing the proposed model were obtained and compared to the expected values.

Filtering by variable distances

The 'variable distances' between atoms in different peptide planes (four per dipeptide unit; Figure 21b) are affected by the relative position of the peptide planes with respect to each other. These reflect the dipeptide unit conformation, and their distribution is not expected to follow a Gaussian distribution. As the SD and MAD approaches [222] are not suitable for non-symmetric distributions, they cannot be used for outlier removal from variable distances distributions. Here, the Highest Density Region (HDR) method, based on the estimation of the data density function, as implemented in the *hdrcde* R-package [224], was used. The R *hdr* function estimates the data density by a Kernel density function estimation [225], [226], with automatic bandwidth selection using the Samworth and Wand algorithm [227], and then finds the best region of highest density at the desired confidence level α (in this case $\alpha = 0.002$). This region should occupy the smallest possible volume in the sample space and every point inside the region should have the probability density at least as high as every point outside the region [224]. The 99.8% HDR was calculated for each variable-distance distribution and only those dipeptide units where all variable distances fall inside the correspondent HDR were accepted.

3.1.2. Obtaining The Sampled Space and Dipeptide Unit Structural and Chiral Information

To obtain the sampled space, describing the three-dimensional real space occupied by each atom in the dipeptide unit, all collected dipeptide units were aligned so that each $C\alpha_{i-1}$ was on the positive x axis, $C\alpha_i$ at the origin and O_{i-1} in the first quadrant of the xy plane. The three axes of inertia of each dipeptide unit were obtained by eigen-decomposition of its 3×3 variance-covariance coordinate matrix. The radius of gyration was computed as described in 2.2.2. The likely secondary structure class assigned by DSSP to the residue corresponding to the central Cα atom ($C\alpha_i$) and the precedent one ($C\alpha_{i-i}$) in each dipeptide unit was stored. A dipeptide unit was marked to be a part of a secondary structural element if and only if both residues were assigned to the same secondary structural class. Moreover, all dihedral (Ramachandran, Ramachandran-like and ω) as well as stretching (τ and τ_d) angles were computed. The chirality was evaluated by calculating the chiral volume of the first peptide (with atoms $C\alpha_{i-1}$, O_{i-1}, $C\alpha_i$ and O_i), and the second peptide (with atoms O_{i-1}, $C\alpha_i$, O_i and $C\alpha_{i+1}$), and the chiral invariant of the entire 5-atom dipeptide unit, as shown in Figure C.4a.

3.1.3. Eigen-decomposition of Distance-Squared Matrices and The Transformation to The DipSpace

For each dipeptide unit collected, a 5×5 Euclidian distance-squared matrix was computed and eigen-decomposed using the ARP/wARP software library (Figure 21c). The absolute values of the four negative eigenvalues were stored and referred as $\lambda_1 > \lambda_2 > \lambda_3 > \lambda_4$. Their square root was then taken to set their magnitudes on the Å scale. In order to evaluate whether their dimensionality could be reduced to three uncorrelated variables, they were subject to PCA over the entire set.

As described in section 3.2.5 and shown in Figure C.4c-d, three main principal components were obtained, which make the transformation matrix R (eq. 59):

$$R = \begin{bmatrix} -0.272 & 0.354 & -0.379 & 0.811 \\ 0.896 & -0.035 & -0.427 & 0.114 \\ 0.112 & -0.707 & 0.432 & 0.548 \end{bmatrix} \qquad (59)$$

These describe the three axes of the new protein backbone conformational space, the DipSpace, and for a given dipeptide unit its coordinates in this space (P) can be calculated using equation 60:

$$P= R(L-\overline{L}) \tag{60}$$

Where L is the vector of the four square-rooted eigenvalues (Λ_i) for the given dipeptide unit (eq. 61):

$$L = (\Lambda_1,\ \Lambda_2,\ \Lambda_3,\ \Lambda_4) \tag{61}$$

and $\overline{L}$ is the vector of their means in the collected set (eq. 62):

$$\overline{L} = (6.954, 3.502,2.425,0.884) \tag{62}$$

The square-rooted eigenvalues of distance matrices were divided into two groups following the sign of the chiral volume of the first or the second peptide or the sign of the chiral invariant. PCA was carried out over these new sets in order to check which chirality measure separated mirror-imaged dipeptide units without altering the DipSpace axes, but dividing the DipSpace into two 'parallel' subspaces.

3.1.4. Conformational Description by The DipSpace Axes

To relate the DipSpace axes to the Ramachandran plot, a set of dipeptide units and corresponding Ramachandran angles were inspected using two different approaches. Firstly, from the collected set of dipeptide units, five representatives from the negative subset that were approximately equally spaced along each DipSpace axis were selected so that they represent a route connecting highly populated regions in DipSpace and, at the same time, show a continuous path when projected on the Ramachandran plot. Target coordinates (t_1, t_2, t_3) in the DipSpace were estimated between the minimum and maximum of each principal component, equally spaced, and the best representative in the set (pc_1, pc_2, pc_3) identified by weighted least-squares minimisation (eq. 63):

$$r^2 = \frac{1}{2}\sum_{i=1}^{3} w_i(pc_i - t_i)^2 \tag{63}$$

With the weight w_i of each principal component calculated as the inverse of its variance over the entire dataset (σ_i^2) (eq. 64):

$$w_i = \frac{1}{\sigma_i{}^2} \tag{64}$$

For the path along the pc_1 axis, the pc_2 and pc_3 coordinates were set to 0.7 and 0.3, respectively. For the path along pc_2, both pc_1 and pc_3 were set to zero. For the path along pc_3, pc_1 and pc_2 were set to -0.7 and -0.2, respectively.

In the second approach, the first peptide plane in a dipeptide unit was aligned to lie in the *xy* plane with its Cα_i atom placed at the origin. The sampling was not restrained to any continuous path in the Ramachandran plot, the coordinates of the $C\alpha_{i+1}$ atom were placed on 196,418 nearly uniformly distributed points on a sphere with a radius of 3.8 Å using the Fibonacci sphere sampling method [228], [229], which allows an optimal and evenly distributed sampling of any number of points on the surface of a sphere [230]. For each of the coordinates of the $C\alpha_{i+1}$ atom, the second peptide unit was rotated around the $C\alpha_i$-$C\alpha_{i+1}$ axis with an angular increment of 0.46°. For each of the generated dipeptide units (in total ~150 million unique conformations), their coordinates in the DipSpace were evaluated. The dipeptide units were selected along each DipSpace axes if their one coordinate was closest to -1.5, -1.4, …, 1.5 while other two coordinates were closest to zero. Dipeptide units in a low populated area of the DipSpace (measured as the DipScore, described in chapter 4) were excluded, leading to the fact that the change of the conformations in the mapped dipeptides is discontinuous. The procedure was done separately for the positive and the negative subspaces and the conformations combined to generate three movies, available in the electronic version of this thesis and whose legends are listed in Appendix D.

3.1.5. Estimation of The Noise Level

In order to evaluate whether the DipSpace shape reflects the traces of the procedure used to obtain the eigenvalues and to carry out their PCA, dipeptides representing different noise models were also computed. A total of 30 000 dipeptide units were computed for each of the five different noise models. All interatomic distances and angles were stored, their distance-squared matrices computed, eigen-decomposed and transformed to the DipSpace, as described in 3.1.3 (Figure 28).

For Set A, 5 random coordinates (three Cα and two O atoms) inside a sphere of 4 Å radius centred at the origin were sampled. These represent the complete conformational space that can be occupied by 5 atoms inside the sphere delimited by the dipeptide sampling space when no restrains or constraints on the interatomic distances are present. For Set B, all fixed distances were constrained to their ideal values, by fixing the first peptide unit in the *xy* plane and fixing $C\alpha_i$ at the origin. Therefore, $C\alpha_{i+1}$ positions are represented by the surface of a sphere with 3.8 Å radius and O_{i-1} positions by the surface of a sphere of 2.4 Å radius. This set represents the conformational space that can be occupied by all combinations of geometrised dipeptide units when no additional constraints are present. Sampling of Set C was obtained as for Set B, but with the minimum distance between $C\alpha_{i-1}$ and $C\alpha_{i+1}$ constrained to be less than the value currently accepted during automated protein model building with ARP/wARP (4.72 Å). Set D was sampled by adding a further constraint on the $C\alpha_{i-1}$-O_i, O_{i-1}-$C\alpha_{i+1}$ and O_{i-1}-O_i distances by limiting them to the intervals determined for filtering of the fixed distances. Set E is similar to Set C but the $C\alpha_i$-$C\alpha_{i+1}$ distance was allowed to vary by 1.0 Å from its ideal value and the oxygen coordinates calculated accordingly by keeping the three fixed distances. This set represents incorrectly geometrised dipeptide units.

3.2. Results and Discussion

3.2.1. The Distances of The Sampled Dipeptide Units

A total of 4 639 chains were collected from the PDB, each representing a different PDB50 cluster. From these, a total of 1 392 719 dipeptide units were obtained before any distance-based filtering. Their average and median fixed distances are approximately the same and close to the trans-peptide theoretical values but show different associated deviations (Table 2). Their joint distribution is composed by two main clouds: one centred at the median value of each fixed distance distribution represents 99.6% of the data and comprises the space occupied by the trans-peptide units (Figure 22a-c); another at shorter $C\alpha_i$-$C\alpha_{i+1}$ distances and longer $C\alpha_{i+1}$-O_i distances represents about 0.3% of the data and comprises the space occupied by the cis-peptide units (mainly cis-prolines) (Figure 22a-c). Around these two clouds, some sparse points are found too, representing about 0.1% of the data and comprising peptide units with abnormal interatomic distances or non-standard residue numbering (Figure 22a-c). Therefore, more than 99% of the data is composed by trans-peptide units, explaining why the mean and the median fixed distances are closer to the trans-peptide theoretical values and the SD is always larger than the MADe.

Table 2 Theoretical and experimental first and second moments of $C\alpha_{i-1}$-O_{i-1}-$C\alpha_i$-O_i-$C\alpha_{i+1}$ bond-angle distances distributions. Theoretical values calculated based on [123].

	Theoretical μ (Å)		***Experimental moments (Å)***			
	Cis peptide	**Trans peptide**	**Mean**	**SD**	**Median**	**MADe**
$C\alpha_i$-O_i	2.391	2.391	2.399	0.018	2.399	0.013
$C\alpha_i$-$C\alpha_{i+1}$	2.802	3.804	3.807	0.052	3.808	0.021
O_i-$C\alpha_{i+1}$	3.539	2.748	2.776	0.062	2.774	0.037

The condensed core of the trans-peptide cloud suggests two types of peptide unit populations, probably arising from different weights applied to the geometrical restraints at different crystallographic resolutions or different model refinement software used (e.g., REFMAC5 or *phenix.refine*). All fixed distance distributions, and consequently the ω angle, can be well described by a sum of two Gaussian functions with different standard deviations (Figure 22d-f and Table B.1). The $C\alpha_i$-O_i distance distribution can be explained by a mix of two Gaussians, which are centred at the same value of 2.4 Å but have different standard deviations (Figure 22d). The distribution that contributes most, with a mixing proportion parameter of 0.64, has a σ value of 0.010 Å, which is very close to the MADe value. Thus, when the MADe method is applied for the estimation of the population σ, it detects only the contribution of the condensed central core of that distribution, ignoring the second population. The second distribution, with a mixing proportion parameter of 0.36, has a σ value of 0.026 and is partially caught by the standard deviation method, as the average standard deviation of these distributions is 0.018.

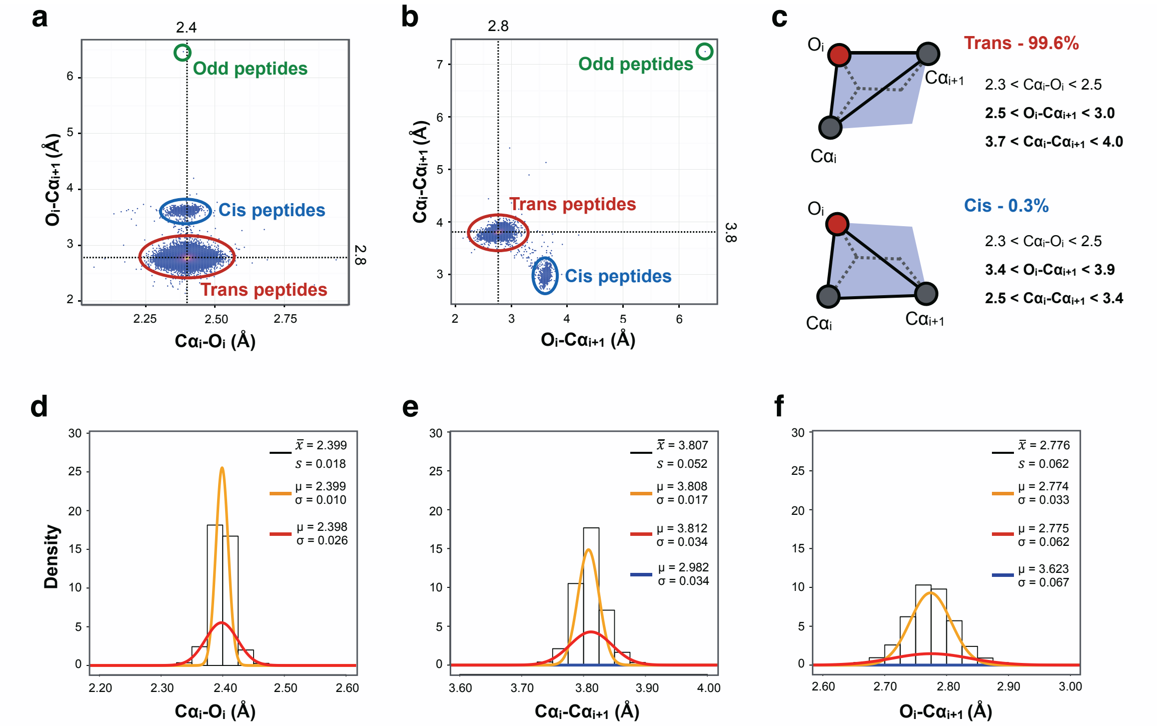


Figure 22 Distribution of the fixed distances in peptide units. (a-b) Joint distribution of $C\alpha_i$-O_i and $C\alpha_i$-$C\alpha_{i+1}$ distances, and O_i-$C\alpha_{i+1}$ and $C\alpha_i$-$C\alpha_{i+1}$ distances. Dashed lines mark the median value of each distribution. (c) The two main classes of peptide units with their corresponding frequency and characteristic fixed distances. (d-f) Normal mixed-model description of the distributions. The histogram shows the distribution of each fixed distance and continuous lines the Gaussian functions describing the data; orange represents the major component of the data (with mixing proportions of about 0.7), red are intermediate (with mixing proportions of about 0.3) and blue only a fraction (with mixing proportions less than 0.04).

For the other two fixed distances, three Gaussian functions are needed to describe the data. Two of them, with mixing proportions of 0.7 and 0.3, are centred at the mean value of the first cloud, and the third distribution (with a mixing proportion of 0.03) is centred at the mean value of the cis-peptide cloud (Figure 22e and f). Here the σ value of the distribution describing most of the density is also close to the MADe value, but the mean σ for both distributions is farther from the distribution SD due to the presence of cis-peptides. Therefore, the μ ± 3σ interval was applied to each fixed distances using the μ and σ values of the Gaussian distribution centred at the trans-peptide cloud with the largest σ (the red distributions in Figure 22d-f).

The four variable distances in the dipeptide units do not follow a Gaussian distribution and their distributions are multimodal (Figure 23b). One can identify two maxima for each variable distance, representing the two main secondary structural elements. Even after applying the fixed distance filtering, these four distributions still contain some outliers. By keeping 99.8% of the points for each distribution, a total of 1 360 370 dipeptides were selected, comprising 98% of the initial set.

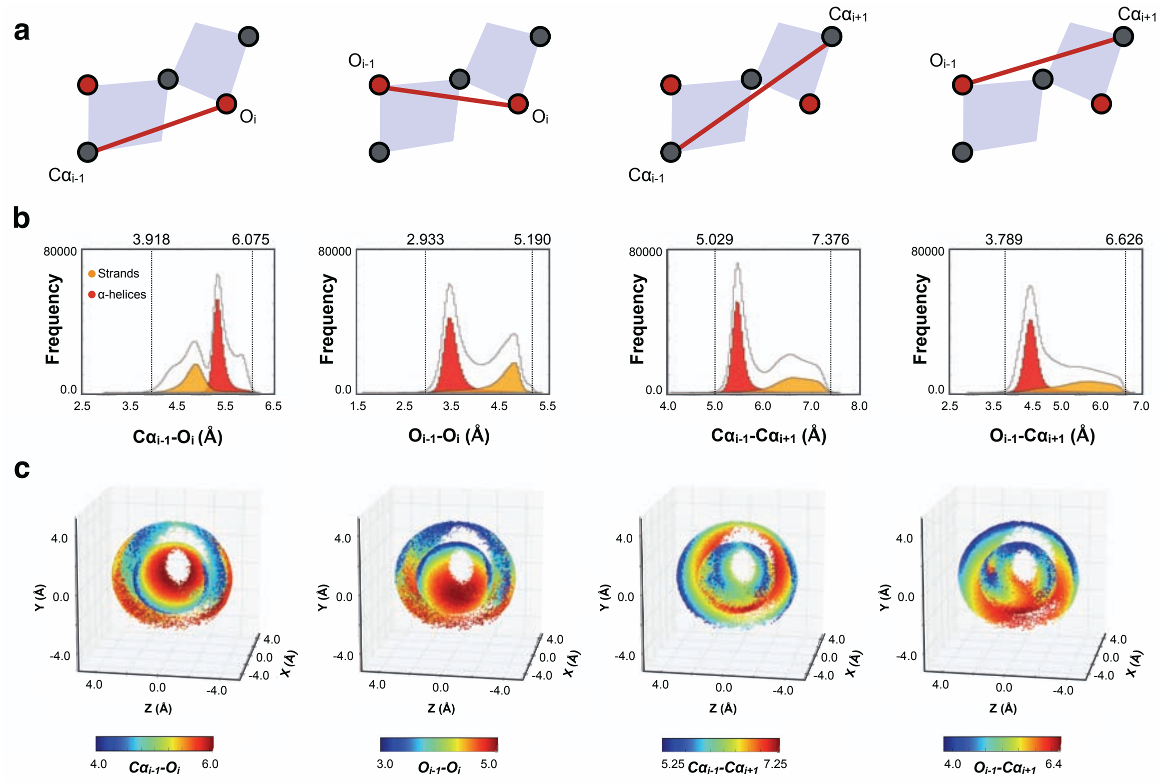


Figure 23 The variable distances. (a) The dipeptide unit atoms involved in them. (b) Histograms showing their one-dimensional distribution after filtering by the fixed distances. The distributions are multimodal, being influenced by the secondary structural preferences of the main-chain. The distributions can be divided into main two regions, the helical (red) and the stranded (orange) regions. Vertical dashed lines in each distribution mark the boundaries of the interval containing 99.8% of the points as determined using the HDR method [227]. (c) Their mapping on the sampled space.

3.2.2. The Angles of The Sampled Dipeptide Units

Comparing the two computed stretching angles, τ shows a unimodal distribution, while τ_d is bimodal (Figure C.1a and Figure C.2a). Among the collected dipeptides, the τ angle averages at 111.3°, ranging from about 85° to 135°. τ_d shows two peaks, corresponding to the helical and extended conformations (Figure C.2a) and spans 60° to more than 170°. It correlates with the O_{i-1}-O_{i+1} distance and shows a similar distribution shape. The distribution of the radius of gyration of the dipeptide units is also bimodal (Figure C.2a). The sampled Ramachandran and Ramachandran-like dihedral angles show a multimodal distribution (Figure C.1 and Figure C.2), with a sharp peak for the helical conformations and a good sampling of the allowed and favoured regions of the Ramachandran plot. Thus, the dipeptide unit sampling procedure although based only on the distances, indicates the conformational preferences of the protein backbone.

3.2.3. The Sampled Space

When the selected dipeptide units were moved to the same origin and their first peptide plane aligned at the first quadrant of the *xy* plane, the space occupied by the 5 atoms on the dipeptide unit describes the *sampled space* (Figure 24). It is a three-dimensional space showing the preferences of the atoms in the peptide planes composing the dipeptide unit. The space resembles half of a sphere, with two shells: the inner shell comprising all O_i atoms positions and the outer shell with all $C\alpha_{i+1}$ atoms (Figure 24a). In comparison with the Ramachandran plot, where mirror-imaged main-chain conformations are related by an inversion of the φ and ψ angles [35], here these are separated by the *xy* plane.

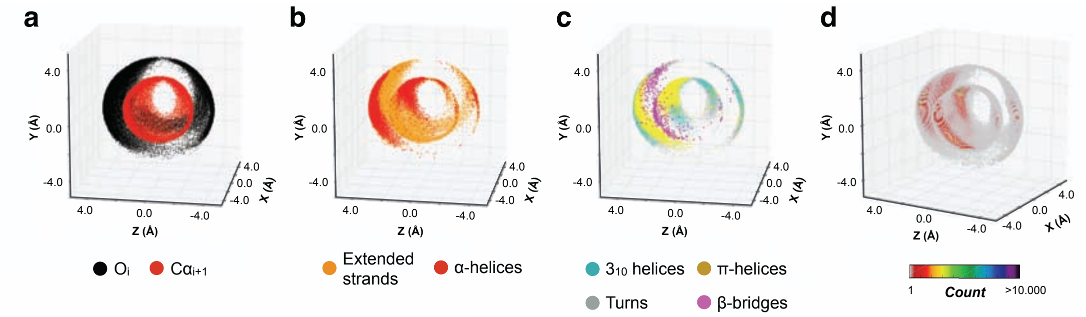


Figure 24 The sampled space coloured by (a) atom type, (b-c) secondary structure (according to the secondary structure annotated by DSSP) and (d) sampling frequency.

Specific areas of the sampled space are differently sampled and populated by distinct secondary structural elements (Figure 24b and c). There is a preference for the positive part of the *z* axis, with most of the dipeptide units being sampled in this area (Figure 24d). There is a strong cluster in the α-helical area, as expected from the distributions of distances and angles. By mapping the flexible distances onto the sampled space, a graphical representation of how the different flexible distances and their combinations describe the dipeptide unit conformation is obtained (Figure 23c). One important observation is that all distributions in the sampled space are mirrored by the *xy* plane, illustrating the notion that the interatomic distances alone are insufficient to describe dipeptide unit conformation [189], [200]. Although the description of the dipeptide units by the coordinates of their O_i and $C\alpha_{i+1}$ atoms in the sampled space provides already a good overview of dipeptide conformation, it lacks the information about the atoms in the first peptide plane as well as the coordinate error. The same can be said about the use of only variable distances because that would assume a too rigid view of the fixed distances, which also vary (Figure 22).

3.2.4. The Four Eigenvalues of Euclidean Distance Matrices

In order to derive a set of protein backbone conformational descriptors that are independent on each other, the eigenvalues of Euclidean distance-squared matrices were used. Their distributions have some resemblance to the distributions of the variable distances indicating that they contain information about the dipeptide unit shape but do not separate between mirror-imaged objects (Figure 25). It was suggested that the dominant eigenvalue of a Euclidean distance matrix is proportional to the mean squared distance of the points from the centre of mass and that the next three eigenvectors and their associated eigenvalues are then the principal components of the object [190]. This means that the largest eigenvalue, the positive λ_{max} eigenvalue, is proportional to the squared radius of gyration R_g^2 of the object. Since R_g^2 corresponds to the sum of the three principal lengths of the inertia ellipsoid of the object [188], λ_4 is then left without physical meaning.

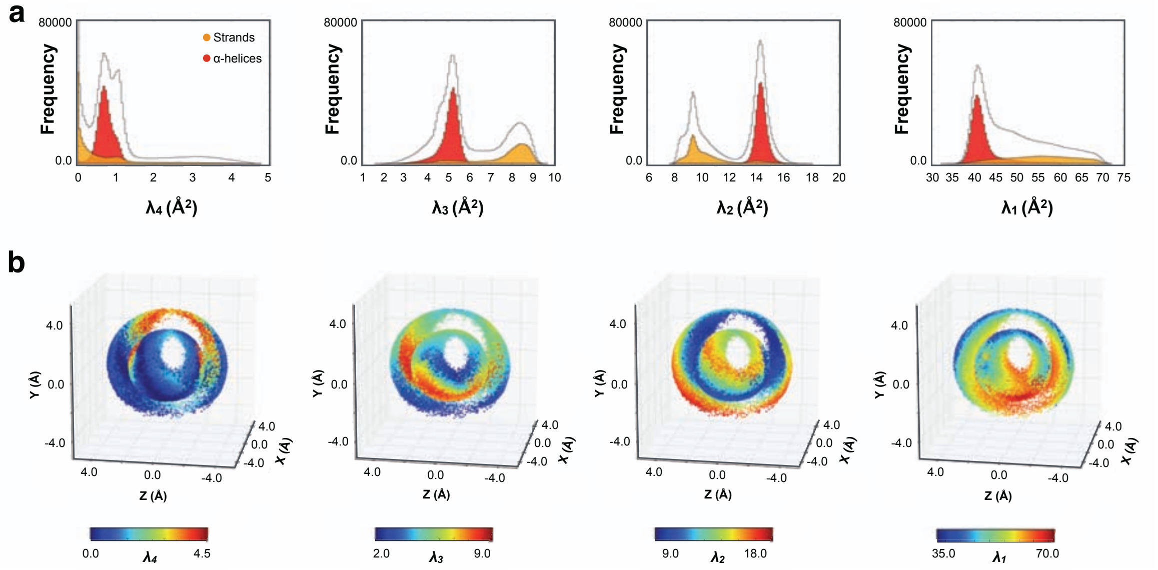


Figure 25 The four independent Euclidean distance matrix eigenvalues. (a) Histogram showing their one-dimensional distribution, depicting the contribution of the structural preferences of the main-chain (helices and strands). (b) Their mapping on the sampled space.

No significant linear correlation was found between the eigenvalues and the variable distances or the principal components dipeptide units' coordinates (Table B.2). Only λ_1 correlates strongly with the first axis of inertia of a dipeptide unit and R_g^2, and its square root with the O_{i-1}-$C\alpha_{i+1}$. λ_2 correlates with the second axis of inertia and its square root with the $C\alpha_{i-1}$-$C\alpha_{i+1}$ distance.

3.2.5. The DipSpace

The four eigenvalues calculated for each dipeptide unit vary in a correlated manner for the set of dipeptide units given that when PCA was carried out over the set of four eigenvalues calculated for units collected from the PDB, less than four Principal Components (pc_i) were obtained, with the main principal component containing approximately 90% of the information. Although the four eigenvalues have very different ranges and magnitudes (Figure 25b), this does not make any one of them less informative than the other. By carrying out PCA [186] over the square-rooted eigenvalues three principal components were identified, which in total account to 99.6% of the total variance (Figure C.4c-d). These define the basis of a three-dimensional space on the Å scale, denoted as DipSpace (*Dipeptide unit Space*) (Figure 26 and Figure 27) and with axes pc_1, pc_2 and pc_3. A variation of the data along the pc_1 axis of the DipSpace correlates with the length of the first principal component of the dipeptide unit ($r = 0.96$) and somewhat weaker with R_g ($r = 0.93$), suggesting that it describes the extension of the dipeptide unit. The pc_2 and the pc_3 axes of the DipSpace correlate weakly with the second ($r = -0.64$) and the third ($r = -0.50$) principal components of the dipeptide unit, respectively.

Overall, the three dimensions of the DipSpace embed the information about the dihedral and the stretching angles. Their mapping on the Ramachandran plot is shown in Figure 27b. Similarly, the mapping of various dihedral and torsion angles on the DipSpace shows their relation to each other (Figure C.3). Both show that a continuous walk through the DipSpace is not necessarily a continuous walk through the Ramachandran plot (Figure C.3c). The meaning of the DipSpace axes is further illustrated by fixing two DipSpace coordinates to a given value while varying the other from its minimum to the maximum (Figure 27). The pc_1 axis describes the extension of the dipeptide unit, e.g. the transition between a P_{II}-spiral and a β-strand [35] (Figure 27a and Video D.1) or a helical and extended conformation. The pc_2 direction describes the twist of the two peptide planes with respect to each other – as exemplified by the transition between a P_{II}-spiral and a γ'-turn [35] (Figure 27a and Video D.2) and the pc_3 axis describes the dipeptide bending – e.g., a transition between a helical conformation and a δ-turn [35] (Figure 27a and Video D.3).

The distribution of the conformations in the DipSpace resembles the shape of a hand with a flatter palm, a cylindrical thumb and a thin connecting layer (Figure 26a). The thumb lobe is highly condensed in one extremity and is mainly populated by helical conformations, with variable τ and φ angles but with ψ close to zero (Figure C.3). These dipeptide units have a moderate span of the twist but variable extension and bending (Figure 26b). The separation of 3_{10}- and π-helical conformation reflecting the change in the τ angle is shown in Figure 26c and Figure C.3c. The palm lobe is populated by turns and extended stranded conformations, with variable τ and φ angles but with ψ close to -180° and 180° (Figure C.3). The dipeptide units there have a moderate span of their bending but the twist and the extension vary considerably (Figure 26b). Since the most abundant conformation for a protein residue is α-helical, the DipSpace is centred close to the very condensed core of the thumb lobe and all other conformations can be seen relative to it.

Glycines and prolines (the identity of the residue linked to the middle $C\alpha_i$ atom of the dipeptide unit) are also distributed distinctively. Glycines are almost everywhere in the DipSpace cloud while prolines and residues preceding them fall into three very specific clusters, mainly in regions corresponding to lower τ angles (Figure 26d).

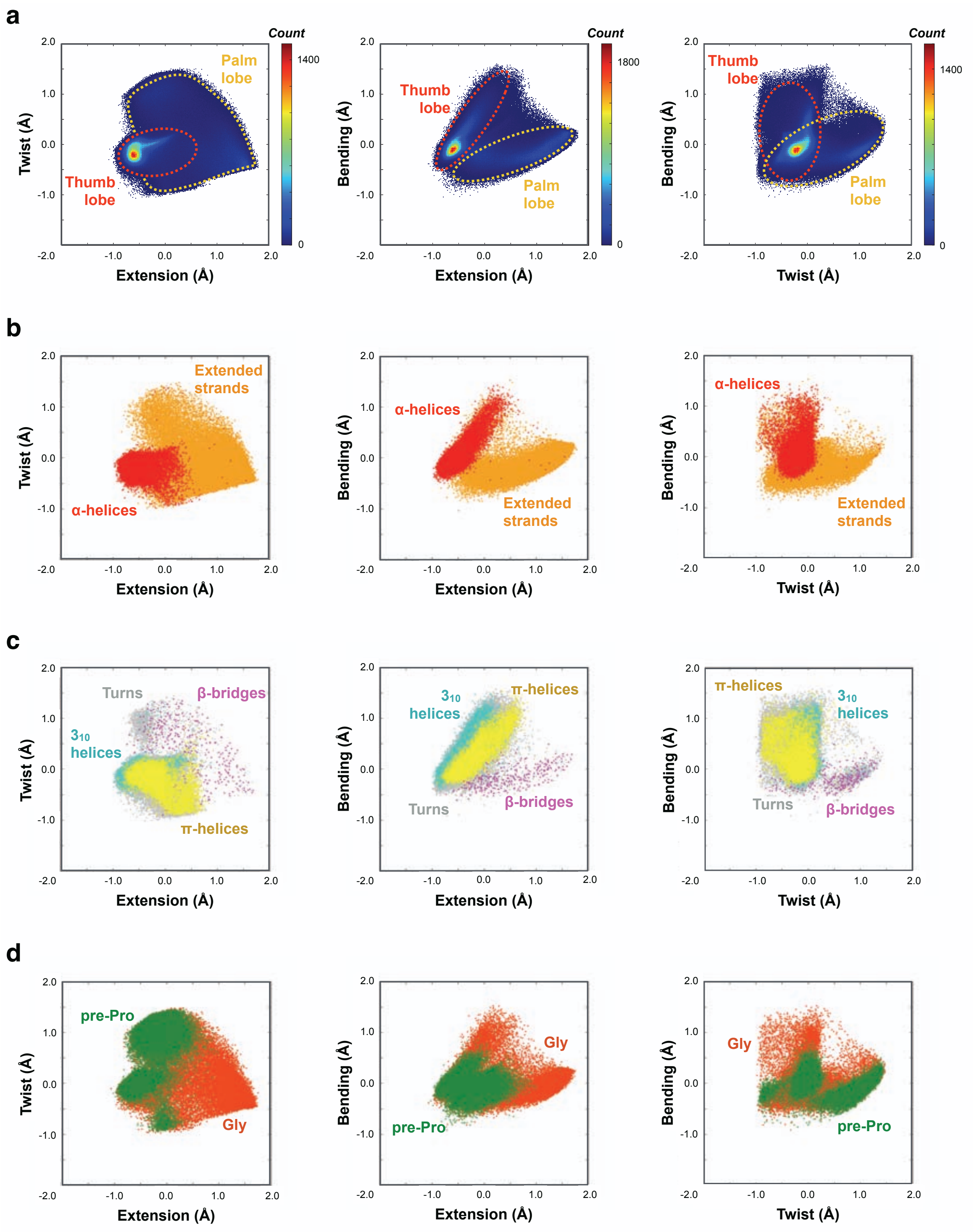


Figure 26 Projections of the three-dimensional DipSpace. (a) Joint distribution of pc_1 (extension) and pc_2 (twist), pc_1 (extension) and pc_3 (bending), and pc_2 (twist) and pc_3 (bending). The two main lobes are marked by dashed lines. Distribution of the (b-c) main secondary structural elements (α-helices and extended strands, as annotated by DSSP) and (d) glycine and pre-proline residues (the identity corresponding to the middle Cα atom of the dipeptide unit).

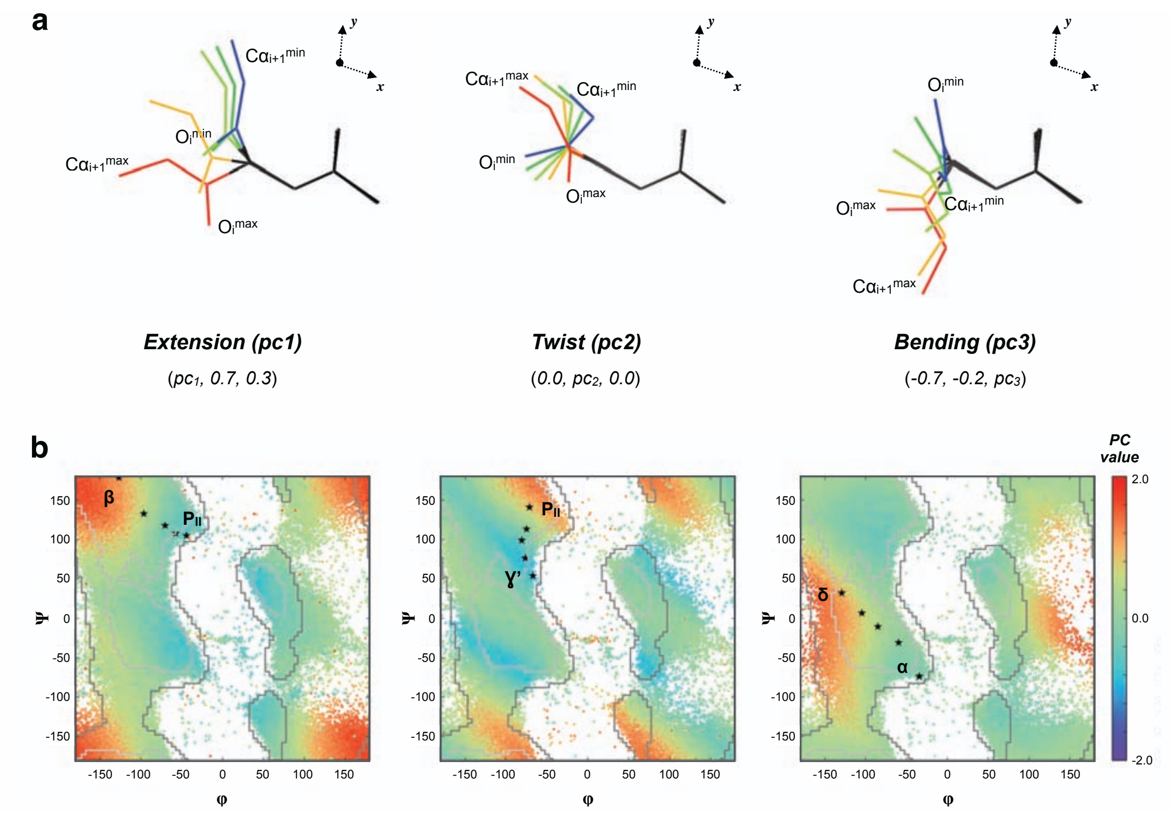


Figure 27 Description of the dipeptide unit geometry by the three DipSpace axes. (a) Representative dipeptide units selected by fixing two DipSpace axes in the negative subspace and allowing the other axis to vary between its minimum (blue) and maximum (red), so that they follow a continuous path in the Ramachandran plot. Target values for each axis are depicted in parentheses. (b) Projection of the DipSpace axes on the Ramachandran plot (the general limits, as described by Lovell [34], are shown). Stars mark the path highlighting the conformational transitions between different regions of the Ramachandran plot described by the dipeptide units in (a). The nomenclature follows Hollingsworth and Karplus [35], where β: β-strands; α: α-helices, γ': γ'-turns; δ: bridge region, several types of turns, P_{II}: P_{II}-spirals.

3.2.6. Separating Mirror-Imaged Dipeptide Units

Two different measures of object chirality were tested – the chiral volume of asymmetric atoms [200] and the dipeptide units chiral invariant [198] (Figure C.4). Since the magnitude of these values is dependent on the interatomic distances, only their sign is necessary to separate between mirror-imaged objects. A set of four points in three-dimensional Euclidean space has, generally, one chiral centre. 5-atom dipeptide units have two asymmetric points, one at the first peptide unit and another in the second peptide unit (Figure C.4a), but only the sign of one is needed, as the information about the other is embedded in the variable distances [189]. The distribution of the three chirality measures in the sampled space (Figure C.4b) suggests that the sign of the first peptide unit's chiral volume is the most intuitive to use, as it divides the sampled space into two mirrored subspaces.

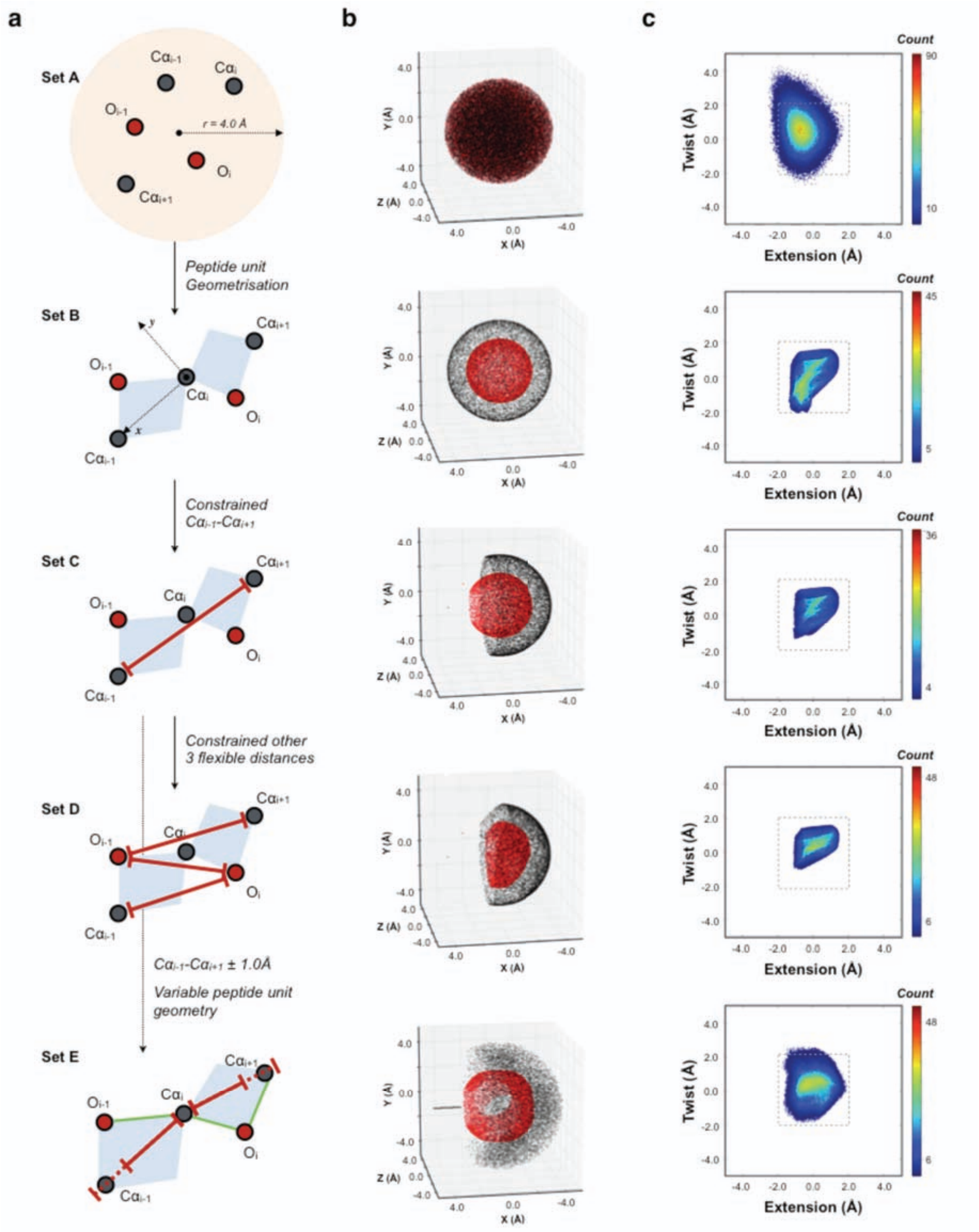


Figure 28 The different sets of randomly generated dipeptides. (a) Schematic representation of the different generated sets and applied constrains and restraints. (b) Corresponding sampled space and (c) distribution over the first two DipSpace axes.

However, if the dipeptide units had been aligned by their second peptide, it would be likely the sign of the second peptide unit chiral volume that would provide this separation. One would expect that the signs of the chiral volumes would correlate with the two Ramachandran-like angles. While this is observed for the magnitude of chiral volume of the second peptide unit and the ψ_d angle (r = -0.94), that is not observed for the first peptide unit chiral volume and φ_d (r = -0.57). Still, there is a perfect correlation between their signs (r = -1). Only when the dataset was separated according to the chiral volume sign of the first peptide the resulting principal components were the same as (or close to) the ones observed when the complete dataset is taken (Figure C.4c and d), meaning that it separates the space into two without changing it.

3.2.7. Separating 'Signal' From 'Noise'

The DipSpace highlights conformationally plausible dipeptide units and indicates the frequency of their occurrence. However, its shape may reflect the traces of the procedure used to obtain the eigenvalues and to carry out their PCA. In order to account for that, five different sets of dipeptide units were constructed (Figure 28). By starting from a random sampling of 5 atoms inside a 4.0 Å-radius sphere (Set A) and adding constraints to approximate these to atomic positions in fully geometrised dipeptide units (until Set D), the sampling space goes from a full sphere of oxygen and carbon atoms to two semi-shells resembling the sampled space observed for the PDB-derived data. The volume occupied by the Set A dipeptide units in the DipSpace (Figure 28c) marks the boundaries of the possible geometrical space occupied by 5 atoms inside the sphere marked by the sampled space limits. When the dipeptide units are geometrised and restrained to the allowed flexible distances intervals, their DipSpace volume decreases, and its shape becomes closer to the PDB-derived dipeptide units cloud. When the fixed and flexible distances of these dipeptide units are allowed to vary and differently restrained (set E), the sampled space looks "fuzzy" and the cloud in the DipSpace is extended.

Given this, dipeptide units in a plausible conformation are separated from those with a likely incorrect geometry in the DipSpace. Therefore, the DipSpace has a potential to be used for the validation of protein backbone.

3.3. Concluding Remarks

The obtained results suggest that distance geometry is also useful for the description of protein backbone conformational space. The presented method provides geometrical information about the backbone atoms around each Cα atom in a protein model within a unified orthogonal Euclidean three-dimensional space where the three axes are on the same scale and account for all degrees of freedom of 5-atom dipeptide units. The obtained DipSpace has the potential to be used as a validation tool of protein backbone geometry. Given that dipeptide units in a random conformation have a different distribution in the DipSpace, one can suggest that for a given Cα position in a protein backbone, its overall conformation can be evaluated by its coordinate in the DipSpace and the sign of the chiral volume of its first peptide unit. If this dipeptide unit falls in an area of the DipSpace populated by dipeptide units with good conformations and the same chirality, it will most likely have a correct geometry. In the opposite case, its conformation will most likely be incorrect.

The advantage of the developed descriptors compared to the presently used angular description of protein stereochemistry is that they are harder to manipulate and are affected by the distances between the atoms. One can imagine that the same set of dihedral angles can be calculated for a given set of points and another set with much larger distances. Their coordinates in the DipSpace, however, will be different. One would then expect it to be harder to 'twinkle' a protein model to force it to be in a highly populated area of the DipSpace by mere manipulation of the dihedral angles: the model will have to be properly geometrised and refined too. The application of the DipSpace as a validator of protein backbone conformation has the potential to allow and complement the detection of conformational problems otherwise not detected without the combination of different tools.

Chapter 4

Development of A New Protein Main-Chain Validation Method

The main goal of using DipSpace within the framework of the ARP/wARP software is to detect dipeptide units that are unlikely to be in correct conformation and exclude them from the model building process. At the same time, the DipSpace can also be used as a general validation tool. This is described in this chapter, alongside with the development of two scoring functions for local and overall protein backbone quality assessment.

One way to evaluate whether a given dipeptide unit is in a likely correct or incorrect conformation would be to check if it falls inside the cloud highly populated by dipeptide units from the PDB (Figure 26). However, there are areas of the DipSpace that are more populated than others both in the PDB-derived set and in the noise models. This leads to the idea of calculating the probability of a dipeptide unit to be in a likely correct conformation based on its location in the DipSpace, the denoted as *DipScore*. For a given position in the DipSpace, it is calculated from the relative densities of the PDB-derived points (d_{PDB}) and those from a noise model (d_{random}) (eq. 65):

$$\text{DipScore} = \frac{d_{PDB}}{d_{PDB} + d_{random}} \tag{65}$$

The values of d_{PDB} are determined from either the negative or the positive subspace, depending on the sign of the first chiral volume of the dipeptide unit, by counting the number of neighbours within a given distance and dividing it by the total number of points in the chiral subspace. The noise model – the values of d_{random} – are computed similarly from a chosen noise model, which is the same for both subspaces. This allows a dipeptide unit to be given a score that varies from 0 to 1 and reflects the validity of its conformation. A DipScore value close to 1.0 indicates a well-populated region of the PDB cloud with little contribution of the noise model. A dipeptide unit with such a DipScore is most likely in correct conformation. Conversely, a dipeptide unit with a DipScore close to zero is in a very unusual or incorrect conformation.

The overall geometry of the protein model can then be assessed by the shape of the distribution of the DipScores for the whole model. It would be expected that a good model has a negatively skewed distribution, with an average DipScore much higher than 0.5. The variance and the kurtosis of the distribution will depend on the variability of the DipScores. Therefore, for a given DipScore distribution, the average (m_1), variance (m_2), skewness (m_3) and kurtosis (m_4), can be computed. It would be expected that for a set of good models, these moments would follow a Gaussian distribution, allowing the calculation of four Z-scores (Z_i) using equation 66:

$$Z_i = \frac{m_i - \mu(m_i)}{\sigma(m_i)} \tag{66}$$

Where $\mu(m_i)$ is the mean and $\sigma(m_i)$ the standard deviation of each moment m_i, calculated from a set of good protein models.

Z-scores follow a standard normal distribution, with zero mean and unit variance. If they are independent, the square root of a sum of their squares (eq. 67):

$$\sqrt{\sum_{i=1}^{4} {Z_i}^2} \tag{67}$$

follows a χ-distribution with the number of degrees of freedom equal to the number of summed Z-scores. However, in our case the four Z_i values computed for a given DipScore distribution are correlated since their underlying four first order moments are not independent. They can be uncorrelated by PCA, reducing the number of degrees of freedom to n. The n uncorrelated Z-scores Zc_i share the same mean (of zero) but will now have different variance, with $\sigma^2(Zc_1) > \sigma^2(Zc_2) > \ldots > \sigma^2(Zc_n)$. By dividing each Zc_i by the variance of the largest component ($\sigma^2(Zc_1)$), their combination (eq. 68):

$$\chi = \sqrt{\frac{\sum_{i=1}^{n} {Zc_i}^2}{\sigma^2(Zc_1)}} \tag{68}$$

follows a χ-distribution with $\frac{\sum_{i=1}^{n} \sigma^2(Zc_i)}{\sigma^2(Zc_1)}$ degrees of freedom.

Z-scores calculated using equation 67 have a sign, indicating whether the respective moment is lower or higher than the average. Therefore, one would expect a perfect model to show a positive Z-score for the mean and the kurtosis, and a negative Z-score for the variance and the skewness. However, the squares in equation 68 remove that sign difference. By multiplying equation 68 by the sign of the highest uncorrelated component Zc_1 a signed chi-score χ_{score}, which differentiates between good and bad models (eq. 69), is obtained:

$$\chi_{score} = \frac{Zc_1}{|Zc_1|}\sqrt{\frac{\sum_{i=1}^{n} {Zc_i}^2}{\sigma^2(Zc_1)}} \tag{69}$$

It follows that models with a positive χ_{score} are better than the average while those with a negative χ_{score} are worse.

This approach for the validation of protein models was implemented as a standalone tool, DipCheck. The DipScore and χ_{score} thresholds had to be determined, as well as the parameters necessary to uncorrelate the Z-scores. This is presented in the next sections.

4.1. Methods

4.1.1. DipScore Calculation

The DipSpace was binned in two three-dimensional grids spanning -1.975 through 1.975 with a step of 0.05 Å, each representing a chiral subspace and containing in total 512 000 data points (Figure 29a). The value inside each bin was assigned to the number of points located within an empirically defined radius of 0.09 Å from the centre of the bin, allowing the overlap between the bins and thus smoothening the transition between them (Figure 29a). The number of neighbours was then normalised by the total number of the PDB-derived dipeptide units in the corresponding subspace. The same procedure was carried out for the Set A noise model (section 3.1.5) built of 1 200 000 randomly generated dipeptide units, resulting in the density of the noise model (Figure 29b). The DipScore for each bin was then calculated using equation 65. For a given dipeptide unit, its DipScore was calculated by computing its DipSpace coordinate in the corresponding subspace (as described in 3.1.3) and applying a parabolic 3×3×3 three-dimensional interpolation [231] on the three-dimensional grid.

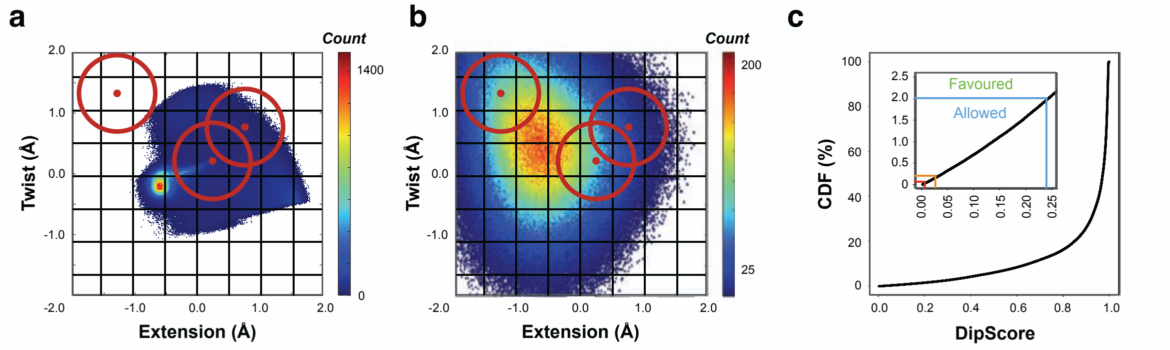


Figure 29 DipScore calculation and the DipScore thresholds. (a-b) Schematic two-dimensional grid representation for the calculation of the density of (a) PDB-derived dipeptide units and (b) Set A randomly generated dipeptide units for a given DipSpace chiral subspace. (c) Cumulative distribution, in percentage, of the DipScores calculated for each dipeptide unit used to compute the DipSpace. A zoom in the low DipScore region and the boundaries for each DipScore threshold are shown.

In order to define the DipScore thresholds, the DipScore was computed for each dipeptide unit selected in section 3.1.1 as described in the previous paragraph. The cumulative distribution of the DipScores in the complete set of 1 360 370 dipeptide units was obtained and the limits encompassing 98% (favoured), 99.8% (allowed) and 99.95% (generously allowed) of the data determined (Figure 29c). All dipeptide units with a DipScore lower than the generously allowed (the remaining 0.05% of the data) are then classified as outliers.

4.1.2. χ-Score Calculation

538 protein chains longer than 50 residues were randomly selected from the initial set of chains collected from the PDB in section 3.1.1. The chains comprised 17% all-alpha, 21% all-beta and 57% mixed alpha and beta (α/β and $\alpha+\beta$) models, corresponding well to the proportion of folds found in the PDB according to SCOPe [232] (Figure 30). The DipScores for each residue in each model, excluding cis-peptides and the residue preceding them, together with the first four moments of their distribution - mean (m_1), variance (m_2), skewness (m_3) and kurtosis (m_4) – were calculated. In order to access the differences between models with different folds, a two-sided t-test was carried out for the comparison between means (H_0: the means are equal; H_1: the means are different; H_0 rejected when $p < 0.005$), as implemented in the *t.test* R function [233].

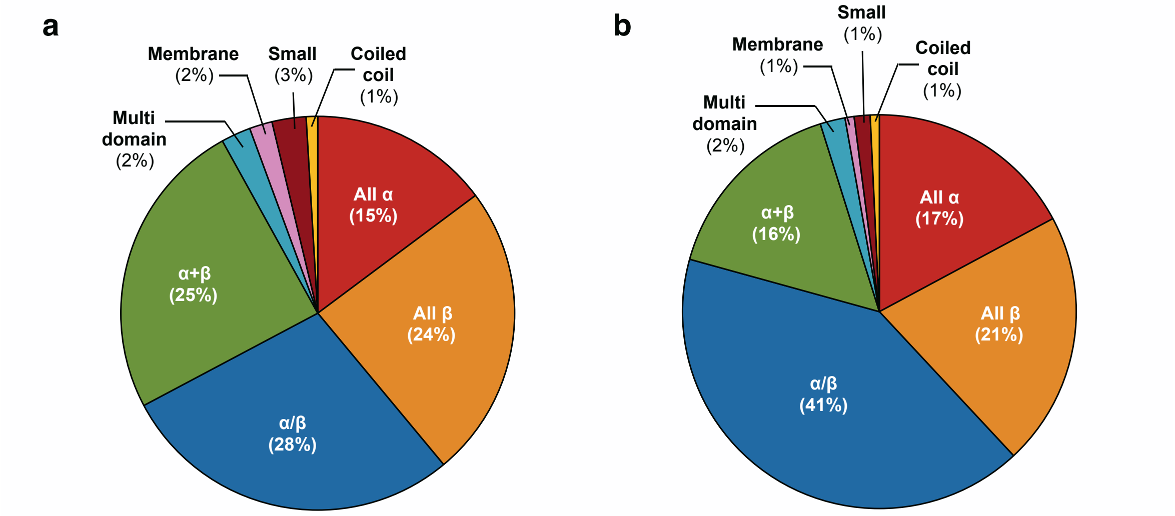


Figure 30 Frequency of protein chain folds according to SCOPe [232] in (a) the PDB (as of March 2016) and (b) the set of 538 chains randomly selected from the total of 4 639 chains as in section 3.1.1.

The median and the median absolute deviation (MADe) were used to estimate the population mean and standard deviation, respectively, for the distribution of each moment. 22 chains with at least one moment being further than 4.0 MADe away from the median were excluded. The mean (μ_i) and the standard deviation (σ_i) for the four first moments (m_i) of the remaining 516 chains were used to calculate a Z-score (Z_i) using equation 66. PCA was carried out to uncorrelate the Z-scores, as described in section 4.2.3. Two main principal components were obtained, using the transformation matrix *R'* (eq. 70):

$$R' = \begin{bmatrix} 0.5206 & -0.4596 & -0.5090 & 0.5085 \\ 0.3064 & -0.6860 & 0.4742 & -0.4589 \end{bmatrix} \qquad (70)$$

so that uncorrelated moments Z-scores (Zc_i) can be calculated using equation 71:

$$Zc = R'Z \qquad (71)$$

where $\boldsymbol{Z}$ is the vector of the four Z-scores (Z_i) for the given model (eq. 72):

$$Z = (Z_1, Z_2, Z_3, Z_4) \qquad (72)$$

Over the set of 516 chains, Zc_1 and Zc_2 have a mean value of zero but different variance ($\sigma^2(Zc_1)$ = 3.322 and $\sigma^2(Zc_2)$ = 0.585). Their combination (eq. 73):

$$\sqrt{\frac{{Zc_1}^2 + {Zc_2}^2}{\sigma^2(Zc_1)}} \qquad (73)$$

is expected to follow a χ-distribution with $\frac{\sigma^2(Zc_1)+\sigma^2(Zc_2)}{\sigma^2(Zc_1)} = 1.176$ degrees of freedom (Figure 31). The χ_{score} can then be computed using equation 69, where a sign is introduced based on the sign of Zc_1 and n = 2. Favoured (98%), allowed (99.8%) and generously allowed (99.95%) χ_{score} thresholds were computed in the same way as for the DipScores (Figure 31c).

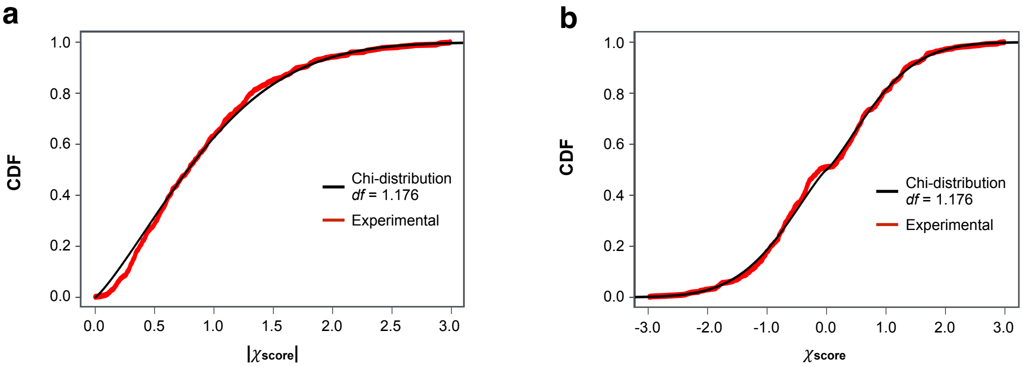


Figure 31 The χ-distribution with 1.176 degrees of freedom. (a) Cumulative density function (CDF) of a χ-distribution with 1.176 degrees of freedom, equivalent to the PDF of the absolute of the χ_{score}, in comparison to the CDF of the absolute χ_{score} computed for the set of 516 chains (experimental). (b) Cumulative density function (CDF) of the χ_{score}, modelled as the mixture of two χ-distributions mirrored at zero, in comparison to the CDF of the χ_{score} computed for the set of 516 chains (experimental).

4.1.3. Test case Selection

The coordinates of five test cases (PDB IDs 1bef, 1lml, 1n7s, 1qjp and 2fdq) were taken from the PDB. The experimental data for the 1lml entry were downloaded from the Uppsala Electron Density Server (EDS) [234] and refinement was carried out using REFMAC5 [122], applying default settings. The PDB_REDO report for the 2fdq model and the coordinates of the rebuilt structure were obtained from the PDB_REDO databank [219]. The WHAT_CHECK [145] and the PDB validation [40] reports for each model were obtained from the PDBe [235]. The number of non-glycine/non-proline Ramachandran plot outliers were computed using MolProbity [39]. The limits of the Ramachandran plot allowed and favoured areas were taken from http://kinemage.biochem.duke.edu/ [34], with a binning of 5˚.

4.2. Results and Discussion

4.2.1. The DipScore Thresholds

The overall distribution of DipScores calculated for the full set of 1,360,370 dipeptide units is highly negatively skewed (γ_1 = -2.6), with a mean value of 0.89 (Figure 29c). Following the classification suggested for the Ramachandran plot by Lovel *et al.* and Morris *et al.* in [34], [41], a residue is referred to be in a favoured region of the DipSpace if its DipScore is above 0.24, in an allowed region if its DipScore between 0.24 and 0.033 (Figure 29c) and in an generously allowed region if between 0.033 and 0.010. A residue with DipScore below 0.010 is regarded as an outlier.

4.2.2. Effect of Fold Class and Secondary Structural Content on DipScore Distribution

One would expect the DipScore distribution of a model to be not only affected by its stereochemical quality but also by the frequency and type of secondary structural elements that make the model. For example, a fully-helical model without any problematic residues may have most of its Cα atoms in the condensed core of the DipSpace thumb lobe (Figure 26a-b), which have a DipScore close to 1.0. On the contrary, Cα atoms in an all-beta model without geometrical problems have a broader area of allowed coordinates in the DipSpace, which are less populated than the helical region and, therefore, have lower DipScore values. Consequently, the DipScore distribution of an all-alpha model has different moments from that of an all-beta model but also mixed-alpha-beta models (Table B.3-7 and Figure 32). All-alpha models tend to show a more negatively skewed (γ_1 = -3.1) and peaked (γ_2 = 11) distribution, with a higher average ($\bar{x}$ = 0.92) DipScore than all-beta and other fold classes. The variance, however, does not seem to be significantly affected (Table B.6). All-beta models, on the other hand, show a less negatively skewed (γ_1 = -2.7) and less peaked (γ_2 = 8) distribution, with a consequently lower DipScore average ($\bar{x}$ = 0.90), but not a statistically different variance.

Coiled-coils stand out as extreme folds, pushing the DipScore distribution to its limits. Only four chains in a coiled-coil fold were used, representing only 1% of the set of chains, but all show a very high average DipScore, statistically different from any other fold class, even higher than all-alpha models (Figure 32). The variance, skewness and kurtosis are also extreme (very low variance, very negative skewness and high kurtosis), but given the small number of test cases, the differences to other fold classes are not statistically significant. This comes from the fact that coiled-coils are full helical structures, without loops or turns, while all-alpha models are composed by helical backbone stretches with loops and bends in between. The presence of these irregular structures allows the DipScore distribution to vary, while all coiled-coil residues are placed in highly populated areas of the DipSpace.

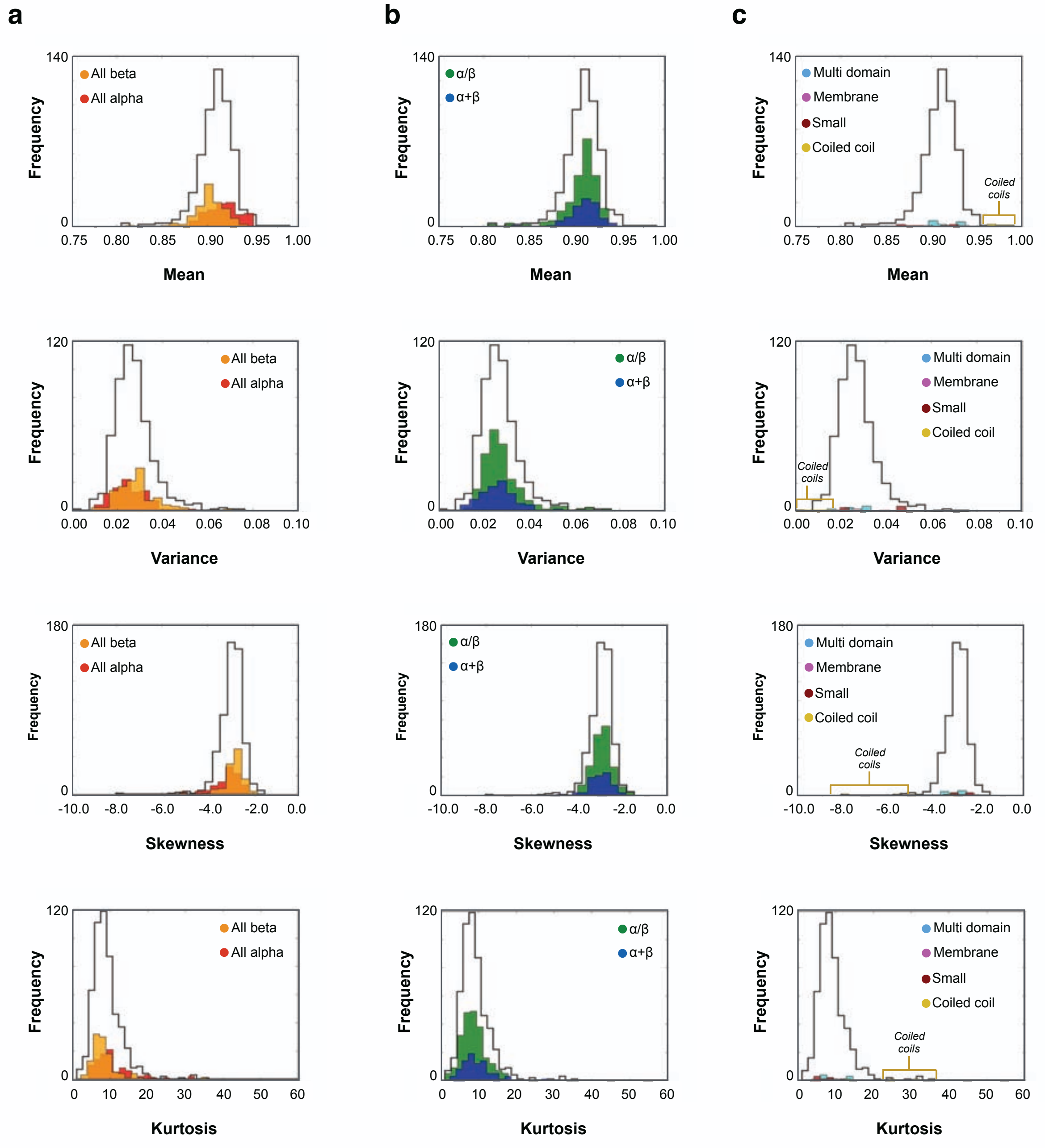


Figure 32 Distributions of the four central moments of DipScore distributions calculated for the set of 538 protein chains. The contribution of the different fold classes is depicted for (a) all-alpha and all-beta, (b) mixed-alpha-beta structures and (c) multidomain, transmembrane, small and coiledcoiled protein chains.

With all folds combined, the mean DipScore distribution for the 538 chains averages at 0.91, with a variance of 0.027, is negatively skewed (γ_1 = -2.9) and peaked (γ_2 = 9; leptokurtic) (Table B.3), following the same shape as the mixed models (Figure 32). These moments represent the shape of the DipScore distribution in the set of models and can be used for a comparison to the DipScore distribution of a given protein model. Given that the overall moment distributions show long tails, they cannot be used for direct Z-score calculation. 22 chains, which were more than 4.0 MADe away from the median for any of their four moments, were excluded. The mean (μ_i) and the standard deviation (σ_i) were re-calculated for the distributions (Table B.9) and used for the calculation of the Chi-score (χ_{score}), as described in the next section.

4.2.3. The Chi-Score Parameters and Thresholds

The four Z-scores computed with the moments of the average model DipScore distribution follow a standard normal distribution but are correlated, as shown in Table 3. By carrying out eigen-decomposition of the Z-scores variance-covariance matrix, two principal uncorrelated components - Zc_1 (83.2%) and Zc_2 (14.7%) - with the same mean (μ = 0) and different variance ($\sigma^2(Zc_1) > \sigma^2(Zc_2)$) were obtained. An increase of Zc_1 implies an increase in the mean and the kurtosis, with a decrease in the variance and the skewness (eq. 70). Therefore, Zc_1 'points' in the direction of the perfect models and a model with a positive Zc_1 is better than the average while a model with a negative Zc_1 represents a structure worse than the average. Multiplying equation 69 by the sign of Zc_1 allows the separation between 'better' and 'worse' models.

The combination of these two uncorrelated parameters (eq. 73) follows a χ-distribution with 1.176 degrees of freedom and when the sign is introduced (eq. 74) the probability function describing the density of χ_{score}, characterising the deviation of the DipScore distribution for the model in question from the 'average' DipScore distribution, is symmetric around zero (Figure 31). From this, a model is annotated as favoured if the χ_{score} is higher than -2.16 (this covers 98% of the distribution), as allowed if the score is between -2.16 and -2.97 and generously allowed if the score is between -2.97 and -3.38; otherwise it is an outlier.

Table 3 Correlation matrix of the four Z-scores (Z_i) calculated for the four first central moments of DipScore distributions for the 516 selected protein chains.

	Average (Z_1)	**Variance (Z_2)**	**Skewness (Z_3)**	**Kurtosis (Z_4)**
Average (Z_1)	1	-0.889	-0.803	0.776
Variance (Z_2)	-	1	0.584	-0.609
Skewness (Z_3)	-	-	1	-0.983
Kurtosis (Z_4)	-	-	-	1

4.2.4. Good, Bad and Ugly Structures

To demonstrate the applicability of the DipSpace, DipScore and χ_{score} to the validation of protein models, 5 test cases representing different scenarios in protein structural analysis were selected (Figure 33a and Table 4).

Table 4 Main-chain quality indicators for the five test cases (glycines and prolines are excluded). Dipeptide units that fall in allowed areas of the general Ramachandran plot are referred to as allowed (All.), and those that fall outside the allowed and favoured regions as outliers (Out.). Dipeptide units with a DipScore below 0.01 are outliers (Out.) and those with a score between 0.01 and 0.24 are allowed (All.; combining allowed and generously allowed). Their percentage is shown in parentheses. The model overall χ_{score} and WHAT_CHECK Ramachandran Z-score are also given.

	1LML Alpha-beta	**1QJP Purely Beta**	**1BEF Fabricated**	**2FDQ Before PDB_REDO**	**2FDQ After PDB_REDO**	**1N7S Purely helical**
Ramachandran (% is shown in parentheses)						
Out.	0 (0.0)	0 (0.0)	1 (0.8)	12 (5.2)	0 (0.0)	1 (0.4)
All.	8 (2.1)	0 (0.0)	6 (4.6)	62 (26.8)	12 (5.3)	2 (0.8)
DipSpace (DipScore) (% is shown in parentheses)						
Out.	0 (0.0)	0 (0.0)	1 (0.8)	13 (5.6)	0 (0.0)	0 (0.0)
All.	5 (1.3)	1 (1.0)	3 (2.3)	41 (17.7)	7 (3.1)	1 (0.4)
Model χ_{score} (overall DipScore distribution)						
χ_{score}	-1.05	2.22	-4.13	-12.89	-0.46	9.97
Ramachandran Z-score						
Z-score	-1.43	-0.301	-5.41	-6.69	-0.54	3.99

The armadillo acyl-CoA-binding protein (ACBP) [236] (PDB ID 2fdq) is (as of November 2015) the protein model with the highest geometrical improvement obtained by running PDB_REDO [219]. This is an all-alpha protein complex refined at 3.5 Å resolution. It has a WHAT_CHECK Ramachandran Z-score of -6.69, with 12 Ramachandran outliers out of 225 non-glycine/non-proline residues. The DipScore indicates 13 DipSpace outliers, mainly in helical conformations, which are not the same as the Ramachandran outliers (Table B.10). For example, Tyr31C is located in a favoured helical region of the Ramachandran plot, but has a τ angle of 106.8^o and is an outlier in the DipSpace due to too short variable distances ($C\alpha_{i-1}$-$C\alpha_{i+1}$ of 4.9 and O_i-$C\alpha_{i+1}$ of 3.6 Å; Figure 33c and Figure 23). On the other hand, Thr64A is a Ramachandran outlier but DipSpace favoured (Figure 33c). This is because the dipeptide interatomic distances fall in the peaks of their distributions, except O_{i-1}-O_i (2.6 Å), pulling the residue to the DipScore favoured region. In the Ramachandran plot this short O_{i-1}-O_i distance pulls it to the border of the allowed region (Figure 33b).

After rebuilding with PDB_REDO, the ACBP model showed no outliers, neither in the Ramachandran plot nor in the DipSpace (Figure 33a and Table 4). The short O_{i-1}-O_i distance around Thr64A increased by about 1.0 Å without a distortion of other intra-dipeptide distances. The τ angle for Tyr31C increased to 110.5°, with a concurrent increase of the $C\alpha_{i-1}$-$C\alpha_{i+1}$ and O_i-$C\alpha_{i+1}$ distances. The improvement of the ACBP backbone is also shown by the distribution of its DipScore (Figure 33d) and the overall χ_{score} (Figure 33a and Table 4).

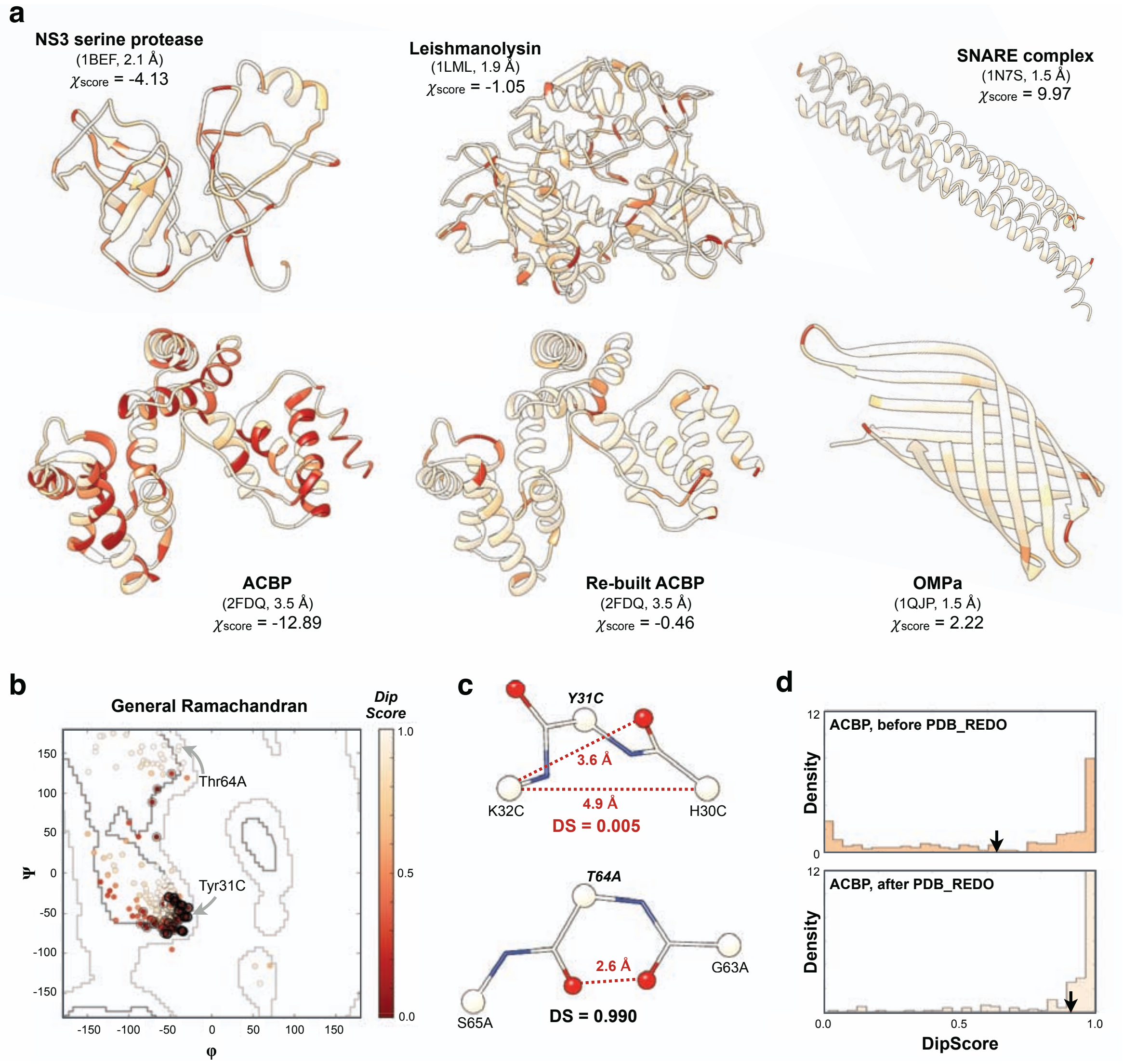


Figure 33 Local and overall protein model validation using the DipSpace. (a) Cartoon representation of the test cases, coloured by the local DipScore. PDB ID and the resolution of the models are indicated. (b) General (non-glycine/non-proline) Ramachandran plot for the ACBP model before rebuilding and refinement using PDB_REDO. The allowed (grey) and favoured (dark grey) boundaries according to Lovell [34] are marked. The points are coloured according to the corresponding DipScore. Outliers (DipScore < 0.010) are circled in black and those in allowed and generously allowed regions (DipScore between 0.010 and 0.240) in light grey. (c) Ball-and-stick representation of ACBP Tyr31C and Thr64A dipeptide units, highlighting their DipScore and problematic distances. (d) DipScore histograms for the ACBP models before and after PDB_REDO. Arrows mark the model's average DipScore.

The crystallographic model of the dengue virus NS3 serine protease (PDB ID 1bef) was retracted from the PDB as a fabricated case and presented several geometrical problems [62]. It is an all-beta model according to SCOPe, with a Ramachandran Z-score of -5.41 but with only one Ramachandran outlier (Table 4). In spite of the low number of Ramachandran and DipSpace outliers, this model shows an unusual DipScore distribution, with an average DipScore of 0.83, about five standard deviations lower than the expected average of 0.91 (Table B.3) and a broad DipScore variance (0.056; four standard deviations higher than the expected value), being then flatter and less skewed than the average protein model, thus supporting its classification as a problematic model.

The crystallographic models of Leishamolysin (PDB ID 1lml), the outer membrane protein A (OMPa; PDB ID 1qjp) and the truncated neuronal SNARE complex (PDB ID 1n7s) have no major geometrical problems but different fold classifications (Table 4 and Figure 33a). Leyshmanolysin is an alpha-beta protein with both favoured χ_{score} and Ramachandran Z-score. Given its mixed nature, it shows a score lower than the all-beta OMPa and the all-alpha SNARE models. The SNARE complex, being a perfect helical bundle, shows a remarkably high χ_{score}, as it would be expected.

4.2.5. DipCheck: A New Tool for The Validation of Protein Models

Given the applicability of the DipSpace as a validation tool for protein backbone geometry and stereochemistry, the developed scoring methods and functions were implemented in a standalone tool: DipCheck (Figure 34). The tool is available as a web service (http://cluster.embl-hamburg.de/dipcheck; Figure 35).

DipCheck only requires on input a PDB or a CIF file of the protein model to be evaluated. It then extracts all dipeptide units, but excludes those with at least one cis-peptide plane (Figure 2 and Table 2). For each dipeptide unit a DipScore is computed using a uniform noise model (as described in 4.1.1). The DipScore distribution is evaluated and a χ_{score} computed for each chain in the model but also for the whole model. The output of DipCheck contains a list of DipScores for each residue, the number of Cα atoms in the model and the number of those evaluated, the number and the percentage of residues in favoured, allowed, generously allowed and disallowed DipSpace areas, and the moments of the DipScore distribution, with their corresponding expected values and Z-scores. Optionally, DipCheck generates a PDB file with the coordinates of the full-atom model, excluding the residues not evaluated, where the B-factor column is replaced by the DipScores computed for each residue. A figure can then be generated with any visualisation software (e.g., UCSF Chimera [237], Pymol, etc.), where the model is residue-coloured according to the DipScores, providing a quick visualisation of the local geometrical quality.

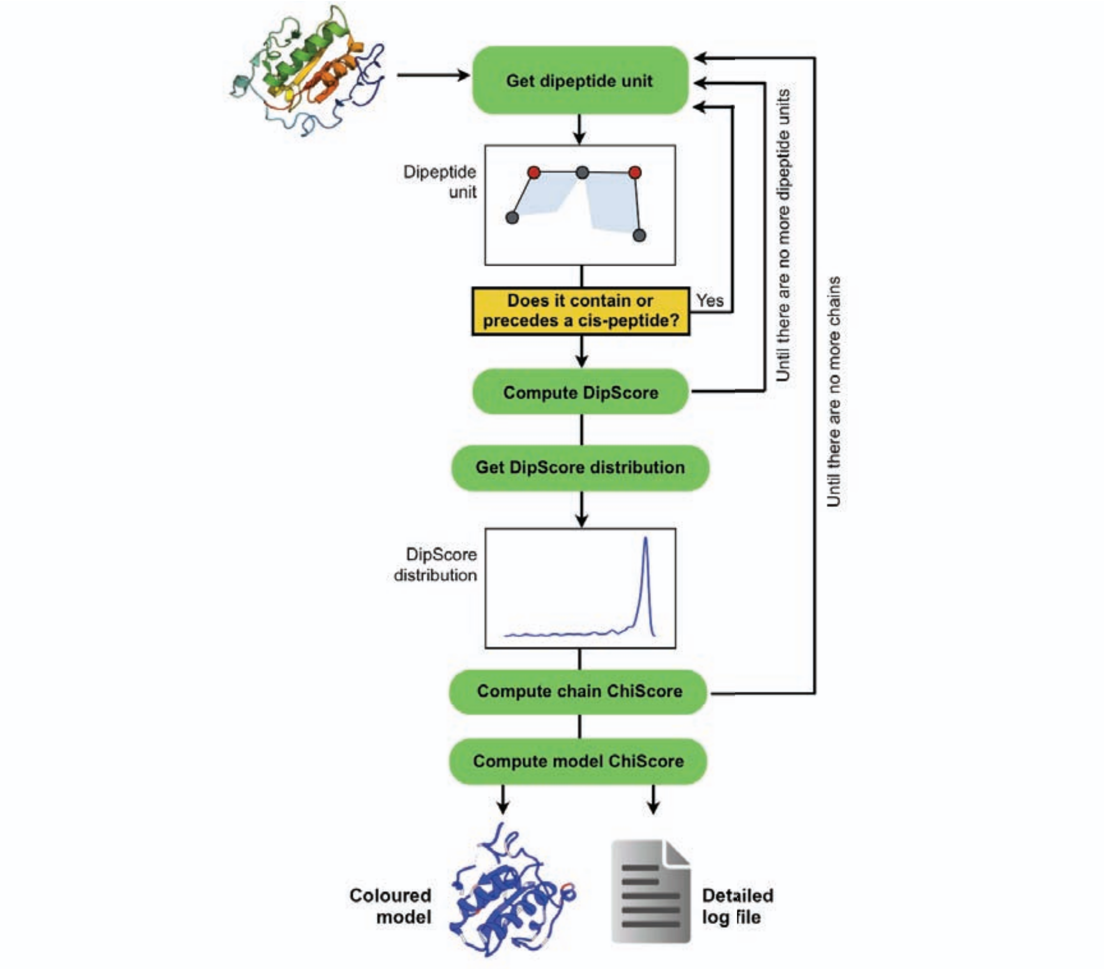
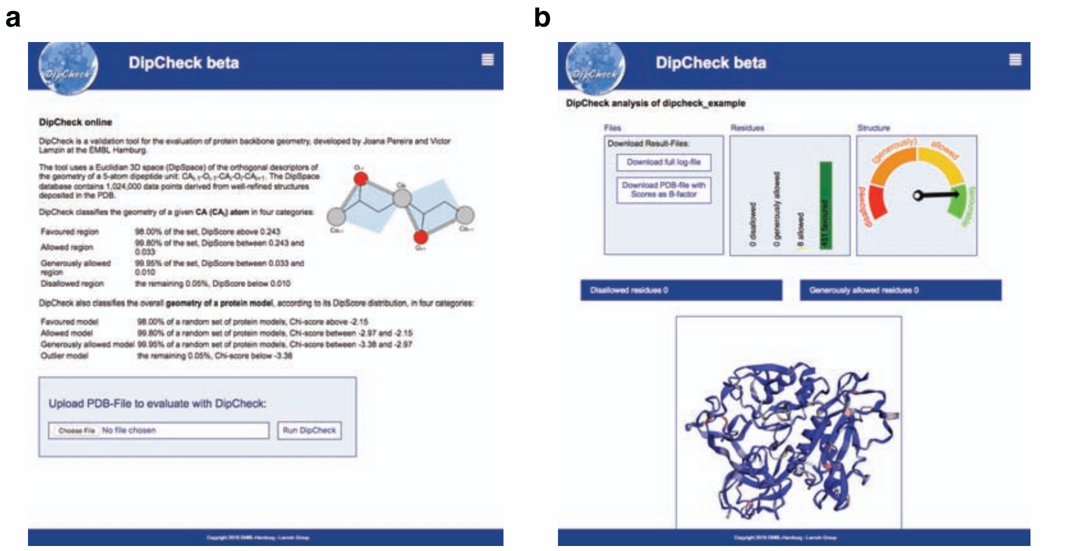


Figure 34 General DipCheck workflow. The yellow box marks the elimination of cis-peptides and those preceding them.



DipCheck online

DipCheck is a validation tool for the evaluation of protein backbone geometry, developed by Joana Pereira and Victor Lamzin at the EMBL Hamburg.

The tool uses a Euclidian 3D space (DipSpace) of the orthogonal descriptors of the geometry of a 5-atom dipeptide unit: CA_{i-1}-O_{i-1}-CA_i-O_i-CA_{i+1}. The DipSpace database contains 1,024,000 data points derived from well-refined structures deposited in the PDB.

DipCheck classifies the geometry of a given **CA (CA_i) atom** in four categories:

Favoured region	98.00% of the set, DipScore above 0.243
Allowed region	99.80% of the set, DipScore between 0.243 and 0.033
Generously allowed region	99.95% of the set, DipScore between 0.033 and 0.010
Disallowed region	the remaining 0.05%, DipScore below 0.010

DipCheck also classifies the overall **geometry of a protein model**, according to its DipScore distribution, in four categories:

Favoured model	98.00% of a random set of protein models, Chi-score above -2.15
Allowed model	99.80% of a random set of protein models, Chi-score between -2.97 and -2.15
Generously allowed model	99.95% of a random set of protein models, Chi-score between -3.38 and -2.97
Outlier model	the remaining 0.05%, Chi-score below -3.38

Upload PDB-File to evaluate with DipCheck:

Choose File No file chosen | Run DipCheck

Copyright 2016 EMBL-Hamburg - Lamzin Group

DipCheck analysis of dipcheck_example

Files

Download Result-Files:

Download full log-file

Download PDB-file with Scores as B-factor

Residues

0 disallowed
0 generously allowed
8 allowed
451 favoured

Structure

disallowed
(generously)
allowed
favoured

Disallowed residues 0

Generously allowed residues 0

Copyright 2016 EMBL-Hamburg - Lamzin Group

Figure 35 The DipCheck web service (a) main page and (b) results page.

The main page of the DipCheck web service (Figure 35a) includes a brief description of the DipScore and χ_{score} thresholds, and a link to the upload of the input pdb file. The results page (Figure 35b) shows summary statistics as well as a half-wheel scheme indicating the classification of the model according to the overall χ_{score} and an interactive window showing the model coloured by local DipScore (red: 0.0; white: 0.5; blue: 1.0). From this page, the detailed results file as well as the modified pdb file can be downloaded.

4.3. Concluding Remarks

The DipScore for a residue can be straightforwardly obtained from the DipSpace and while it is sufficiently informative on its own, it complements other tools, such as the Ramachandran plot. The overall χ_{score} provides a measure of the overall model quality, both at the conformation and geometrisation levels. Although it may be seen related to, for example, the WHAT_CHECK Ramachandran Z-score [149], it uses different information and has other statistical properties: while Z-scores are distributed normally, the χ_{score} follows a Chi-distribution where a sign is introduced to separate models which are better or worse than the average. The method can therefore be used for the detection of protein models with regions of unusual conformations or geometry of trans-peptide units. One would generally expect the models with a poor Ramachandran Z-score to also display a poor DipSpace χ_{score} but variations can be observed, as shown by the five test cases presented (Table 4).

The presented way to compute the DipScore does not differentiate the identity of the residue. It will certainly be of interest to investigate the DipScore distributions for glycines, prolines and cis-prolines. Another direction to pursue could be the addition of weights or deliberate narrowing of the distributions of the intra-dipeptide distances, so that the DipSpace becomes tuned to a particular geometrical feature, for example the O_{i-1}-O_i distance. The use of other noise models could also adjust the method towards different approaches for model building and validation.

Chapter 5

Improvement of Automated Model Building Protocols at Medium-to-Low resolution

The aim of this project was the development of tools and computational methodologies that would allow the improvement of automated model building and the interpretation of MX electron density maps at medium-to-low resolution. Therefore, a considerable part of the work was devoted to the implementation of the developments in ARP/wARP. This work was performed in several steps and is described below.

5.1. The Effect of Resolution on ARP/wARP Automated Protein Model Building

In order to understand what is needed to improve automated protein model building at medium-to-low resolution, three main questions where addressed:

- How does the quality of the final model change with a decrease in the resolution of the X-ray data?
- How does the model evolve over the cycles of model building?
- Do the fragments built in each cycle differ considerably from those built in previous cycles?

5.1.1. Selection of Test Cases and Benchmark Analysis

Two jobs submitted to the EMBL cluster with a non-confidential dissemination level were used as test cases with diffraction data truncated to different resolution between 1.6/1.8 Å and 5.0 Å. All calculations were carried out with the beta-version of the 'classic protein model building' protocol of ARP/wARP 7.4 (as of March 2013) [117] implemented in ArpNavigator [238]. The data contained experimental phases, which were used to build the initial free-atoms model. The first dataset selected (referred to as 06234) extended to a resolution of 1.8 Å and corresponded to a homo-tetramer, with 104 residues per chain, totalling to 416 residues. The second dataset (referred to as 14417) extended to a resolution of 1.6 Å and corresponded to a protein with 407 residues.

Automated protein model building was carried out at 8 different resolutions of the data (Figure 36 and Figure 37), using default parameters. The total number of residues and fragments built, and the model correctness as estimated by ARP/wARP were stored and analysed. The estimated correctness of the model is an empirical score based on the number and lengths of built fragments and the resolution of the data. Model completeness was calculated as the ratio of the number of residues built to the total number of expected residues. In order to monitor which parts of the model were built at each model building cycle, the ARP/wARP *watercomp* routine was used, which calculates the nearest-neighbour root-mean-square deviation (N.N.r.m.s.d.) between two models taking into account symmetry transformations. Here, the comparison was done between a partially built protein model and the final model at high resolution. Only distances below 2.0 Å were considered for calculation as they were deemed to represent the correct Cα(model$_1$)-Cα(model$_2$) correspondences.

5.1.2. Understanding The ‘Effect of Resolution’

Figure 36 and Figure 37 show that at reduced resolution the models are built to a lower completeness, become more fragmented and are estimated to be less accurate. However, different behaviour is observed for each dataset. For the test case 06234 (Figure 36), either with or without the use of the PNSextender module, the completeness of the built model and the estimated correctness are higher than 90% at better than 3.0 Å resolution, but drops rapidly to zero correctness and 40% completeness at worse than 3.5 Å resolution. The number of chain fragments stays below 10 for the entire model at <3.0 Å resolution, and increases to 30-40 fragments at >3.5 Å resolution. This clearly demonstrates that there is a need to improve model building at resolution worse than 3.0 Å.

However, for the test case 14417 (Figure 37) the estimated model correctness falls from 100% at 1.6 Å to around 40% at 3.0 Å and further to 0% at 3.5 Å. Model completeness drops dramatically from 98% at 1.6 Å to around 50% at 2.0 Å, although generally at this resolution ARP/wARP builds more than 90% of the structure. The investigation on how the three parameters (model fragmentation, completeness and estimated correctness), as well as the crystallographic R-factor, evolve along the model building cycles (Figure 38) shows that at high resolution (Figure 38a) (1) the model built at the first cycle contains about 30 fragments, which represent 45% of the protein, with an estimated correctness of over 90%; (2) then, at each cycle there is a steady increase in model completeness and estimated correctness, with a concurrent decrease in the fragmentation and a decrease in R-factor, with the jumps corresponding to the steps where the model is re-built; (3) at the end, the model is fully complete and correct, and consists of one fragment.

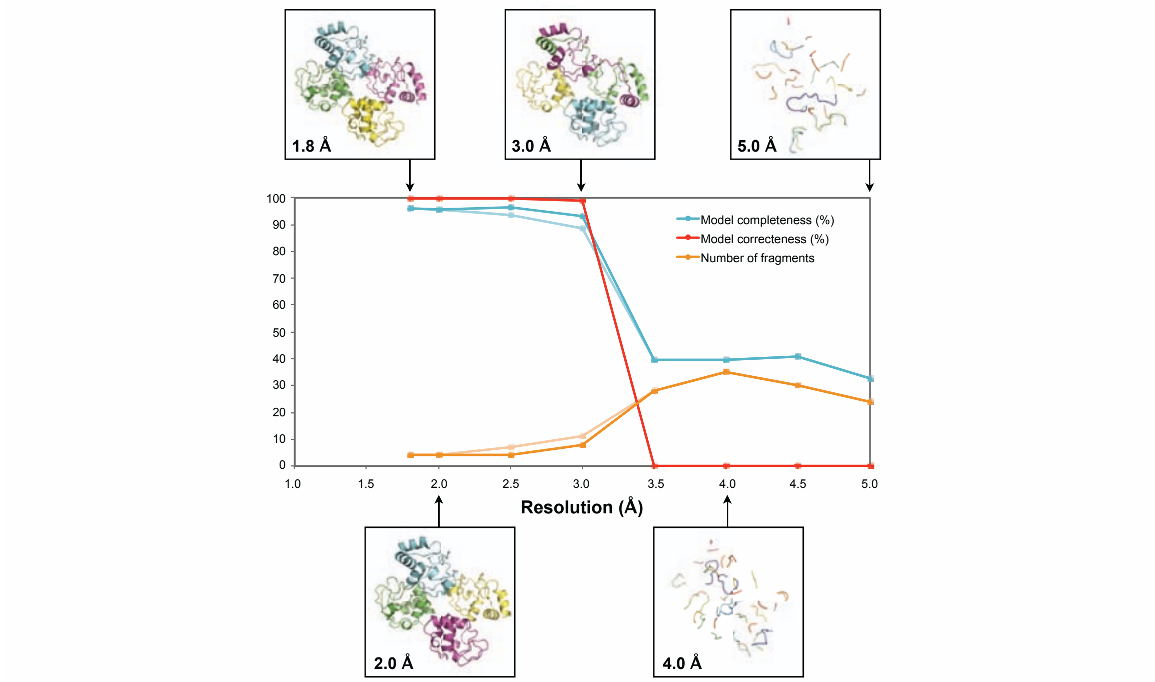


Figure 36 The quality of the final model for the test case 06234, evaluated at different resolution. Solid lines indicate protocols involving PNSextender; faded lines without it.

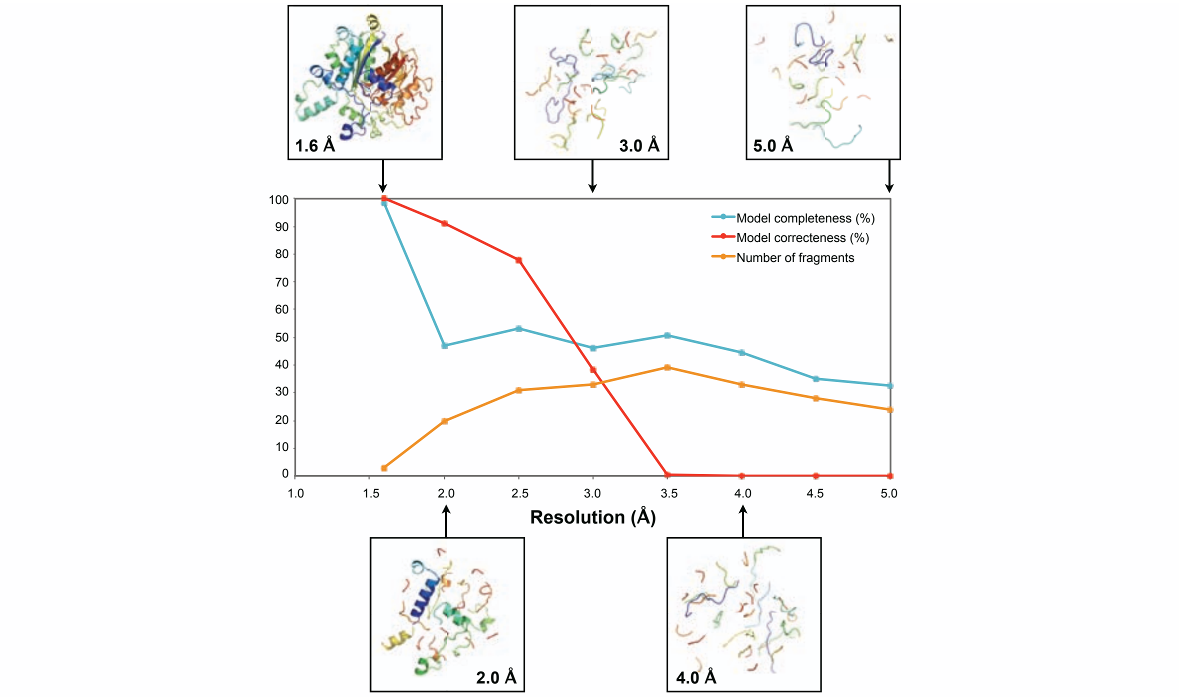


Figure 37 The quality of the final model for the test case 14117, evaluated at different resolution. Solid lines indicate protocols involving PNSextender; faded lines without it.

Such evolution is not observed at other resolutions. At 2.0 Å (Figure 38b), the model completeness slightly reduces. The fragmentation improves and the estimated model correctness improves marginally. At 3.0 Å (Figure 38c), the first 4-5 building cycles generate more complete and accurate models, which are less fragmented. However, during the next building cycles the structure deteriorates, resulting in a model similar or even worse compared to the one built in the first cycle. Finally, at 4.0 (Figure 38d) and 5.0 Å resolution, there is an overall decrease in the model completeness and estimated accuracy, although the fragmentation has slightly improved. Still, having about 30 fragments (with 5.5 residues per fragment on average) is not much help for subsequent manual model completion. At this resolution range, the model correctness is zero already after the first building cycle, and this does not improve at further cycles. It is also noteworthy that the zipper-type pattern of R-factor reduction is not observed. This indicates that it is getting difficult for ARP/wARP to build any model at this resolution. Once again, one would expect this behaviour since ARP/wARP was not designed to build protein models at 5.0 Å resolution.

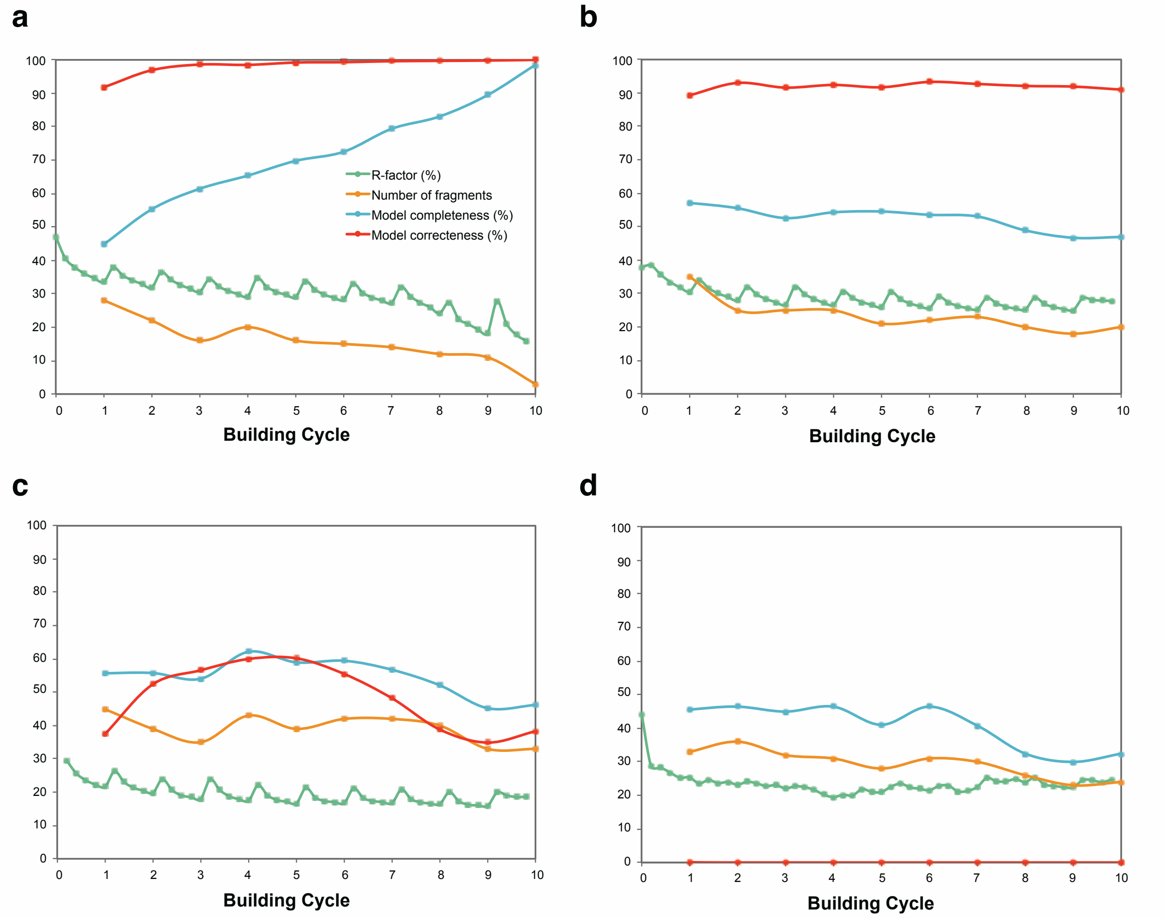


Figure 38 Evolution of the model building for the test case 14417 at (a) 1.6 Å, (b) 2.0 Å, (c) 3.0 Å and (d) 4.0Å resolution.

5.1.3. A Closer Look into Each Building Cycle

The evolution of the built fragments during each cycle at different resolutions, specifically at 2.0 and 3.0 Å (Figure 39), was investigated comparing them to the final model at 1.6 Å. Here, one of the two outcomes were expected: (1) either at each cycle the modelled fragments stay the same or increase in size and connect to other built fragments, or (2) their location is not retained and they are re-built at other places in the sequence space.

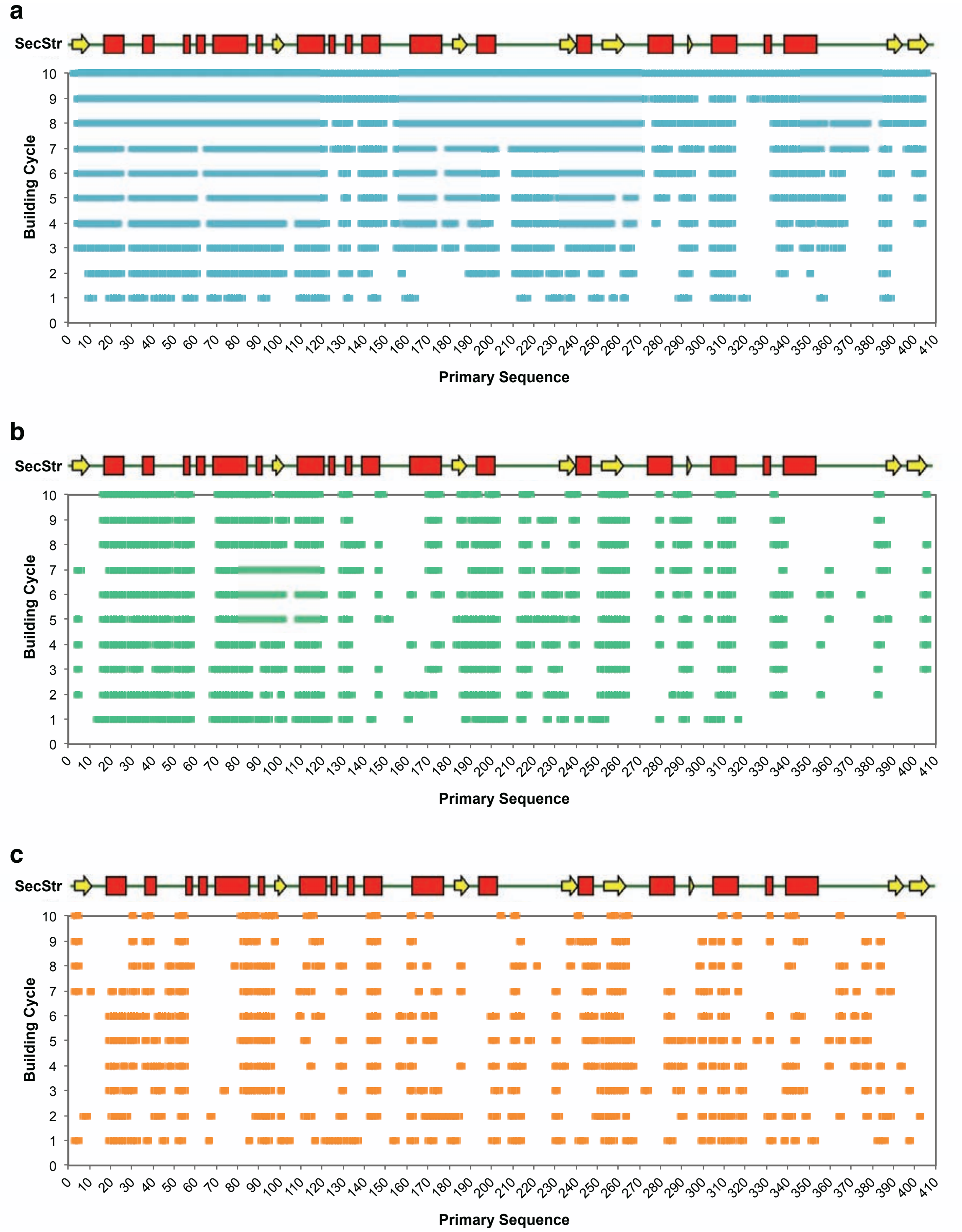


Figure 39 Evolution of the model building for the test case 14417, modelled at (a) 1.6, (b) 2.0 and (c) 3.0 Å resolution. SecStr: secondary structure (green line: loop; red cylinder: helix; yellow arrow: strand), as in the final model at 1.6 Å, annotated by DSSP [239].

At 1.6 Å, the fragments become, indeed, elongated at each round; at the same time new fragments are modelled (Figure 39a). This means that at each cycle the modelled fragments correspond to previously built fragments in the same sequence space and are becoming connected by newly-modelled loops. Therefore, at each cycle the number of fragments decreases while their length increases, converging to a complete model in a single chain. Conversely, at 2.0 Å (Figure 39b), there is an increase in the number of modelled fragments during the first two cycles, but no further positive evolution: all the fragments during subsequent cycles are roughly the same as those modelled in the second cycle. In fact, the model built in the last round has fewer fragments due to disappearance of some previously modelled fragments. At 3.0 Å (Figure 39c), the quality of the model is low so that it is difficult to compare the intermediate models to the final model at 1.6 Å resolution. Correspondences of 30 residues were only found in the N-terminal part of the protein, one region always well modelled at 1.6 and 2.0 Å. However, in other places almost every correspondence is randomly associated to a protein fragment, yielding a fragmented pattern that would suggest a random (in the sequence space) fragment building process at each cycle.

Figure 39 also shows that there are regions which are always built in each round at any resolution, others that are never built at 2.0 and 3.0 Å, and even regions that are never built at 2.0 Å but are at 3.0 Å. In order to obtain a deeper understanding, the region between residues 50 and 90, which folds into a small motif composed by two small and one long helices connected by two loops, was looked in more detail (Figure 39 and Figure 40a). At 1.6 Å, this region is initially modelled in four fragments, which are then extended during subsequent cycles; at the 5th cycle the longest helix is built and the motif is modelled in two fragments separated by a gap of 7 non-modelled residues. Although the connection between both fragments is not yet built at this cycle, one can see that the electron density in the connecting region is improved and a higher number of free atoms is placed at or near the correct Cα positions (Figure 40a).

At 2.0 Å (Figure 40b) there is a steady improvement in map quality at each round, although the density remains poor in the region corresponding to the missing fragment. In fact, this region is formed by blobs of density that are filled with free atoms placed too far from the correct Cα positions and do not form a traceable part of the structure. At 3.0 Å resolution (Figure 40c), three small fragments composed of 2 to 3 residues are first built. However, they are not located close to each other. In the subsequent cycles, three fragments are modelled close to each other and comprise the beginning of the longest helix and the small helix, which are always built at higher resolution (Figure 40). Contrasting with what happens at higher resolution, there is no improvement of the electron density: it remains fairly the same but more ‘blobby’ and discontinuous. Additionally, there is also a reduction in the number of free atoms placed in the region corresponding to the non-built linker.

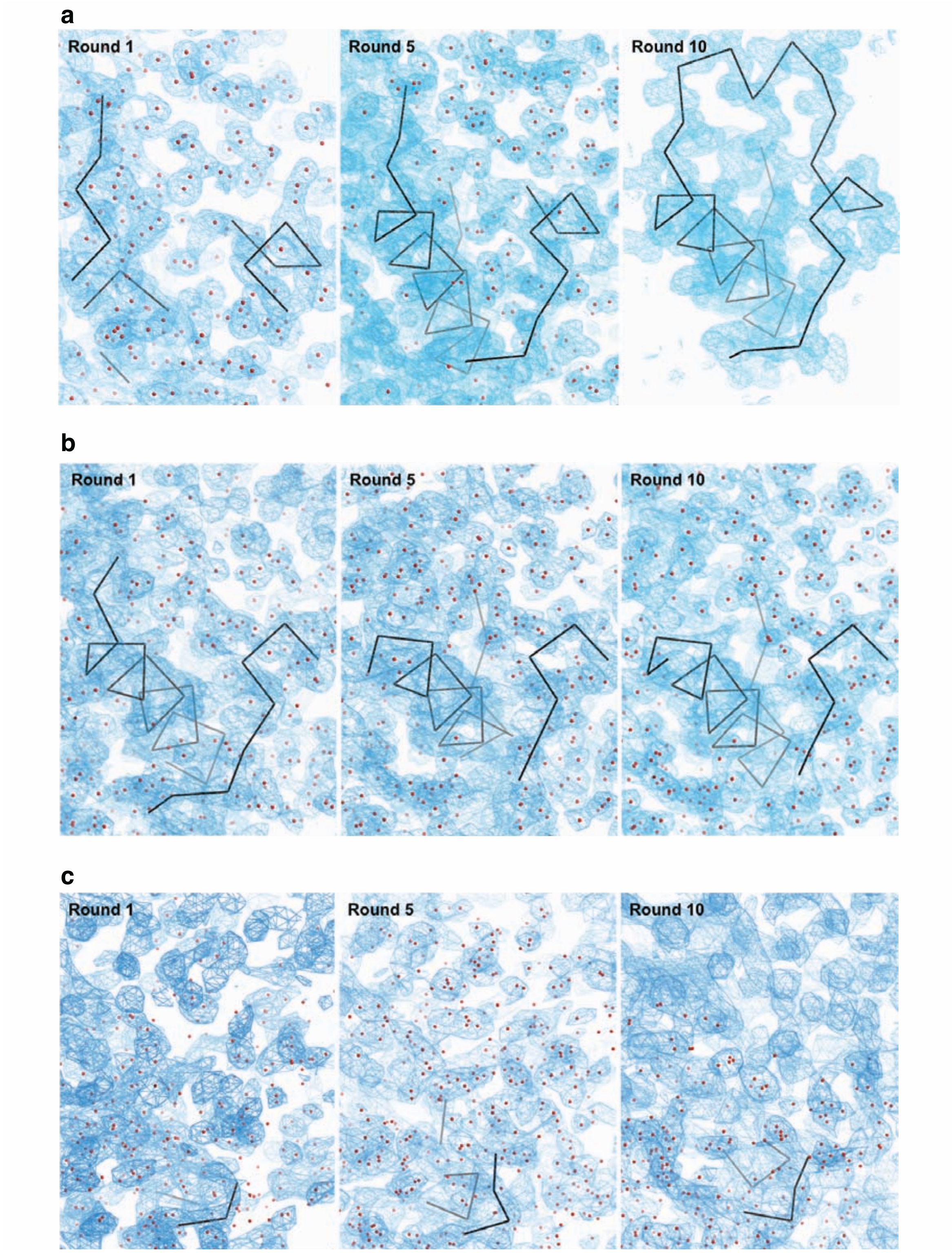


Figure 40 Close-up view of fragment evolution for the test case 14417 between residues 50 and 90, modelled at (a) 1.6, (b) 2.0 and (c) 3.0 Å resolution. Cα trace of main-chain fragments is represented in black thick lines with free atoms as red points. Maps are in blue, drawn at 1.5σ above the mean (0.789 e/Å^3).

5.1.4. Possible Directions for The Improvement of ARP/wARP Protein Model Building Protocol

These results indicate that one problem relates to the inaccuracy (with respect to the true atomic positions) with which the free atoms are placed into the map at lower resolution. This has two consequences: (1) the likelihood of identifying correct main-chain paths is reduced; (2) during the sequence docking the identification of the correct connectivity vectors becomes more difficult. From the ARP/wARP workflow (section 1.4), there are three main points of action (Figure 41):

1. Improvement of the accuracy of the peptide units, which are identified from the free atoms during main-chain tracing; this may be attempted during the geometrisation of the identified peptide units. One would expect that the deeper the level of geometrisation, the closer to the target geometry the peptide units converge, and the more accurate the final protein model could be.

2. Improvement of the main-chain conformation by supplementing the use of the Ramachandran-like plot with the DipSpace and the DipCheck model validation. Given the fact that at medium-to-lower resolution the main-chain third degree of freedom, represented by the τ_d angle, becomes significant for the description of protein main-chain conformation, the application of the DipSpace as a validation tool should improve the quality of the built models.

3. Improvement of the sequence docking by replacing a one-dimensional connectivity vector with a more elaborate method. One way to do this could be a use of a density shape-based method similar to those applied for the identification of ligands and ligand binding sites.

In the next sections the main focus is given to the first two action points, although preliminary results obtained for the third one are also presented.

5.2. Geometrical Analysis of Candidate Dipeptide Units

Geometrisation of a dipeptide unit is an important step in automated protein model building with ARP/wARP. The pipeline as of version 7.4 uses the Ramachandran-like plot, which was obtained from properly geometrised polypeptide chains. As ARP/wARP builds a putative peptide units on two free atoms that are 3.8±1.0 Å apart and then finds a putative dipeptide unit if two peptide units share one Cα and follow each other, the candidate dipeptide units may be based on an object with geometrical properties far from their ideal values. Therefore, geometrisation of the dipeptide is needed before its conformation can be evaluated. As a default, only one round of geometrisation has been implemented in the ARP/wARP version 7.4.

Although the geometry of the dipeptide unit is already good after one round of geometrisation, the fact that the position of the central Cα ($C\alpha_i$) atom is obtained by averaging its position from connecting peptides, may disturb the overall geometry of the dipeptide unit (Figure 9c). By performing additional geometrisation rounds, new middle Cα positions are generated and the new average is taken. The dependence between the level of geometrisation and the number of correctly traced residues was then studied. One would expect that, the higher the level of geometrisation, the better the geometry of the fragments and the resulting ARP/wARP models could be (Figure 42). At the same time, too many rounds of geometrisation may result in a dipeptide being driven too far from the initial free atoms positions.

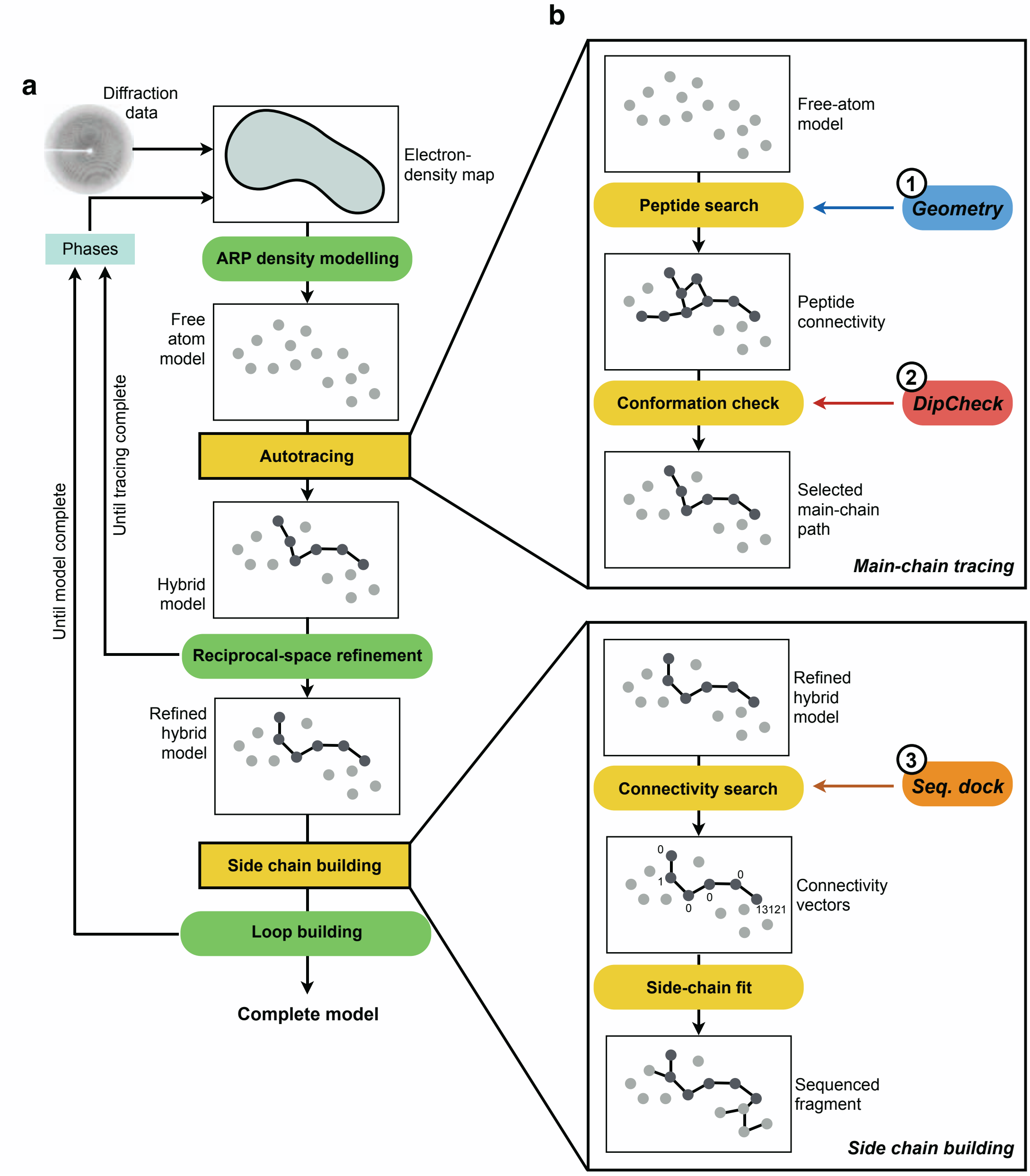


Figure 41 Main points of action for the improvement of protein automated model building at medium-to-low resolution with ARP/wARP. (a) General ARP/wARP workflow (described in detail in section 1.4). Yellow rectangular boxes mark the main two steps that should be focused on. (b) Zoom into the two main steps, highlighting the three main points of action, numbered according to their sequential order in the ARP/wARP pipeline.

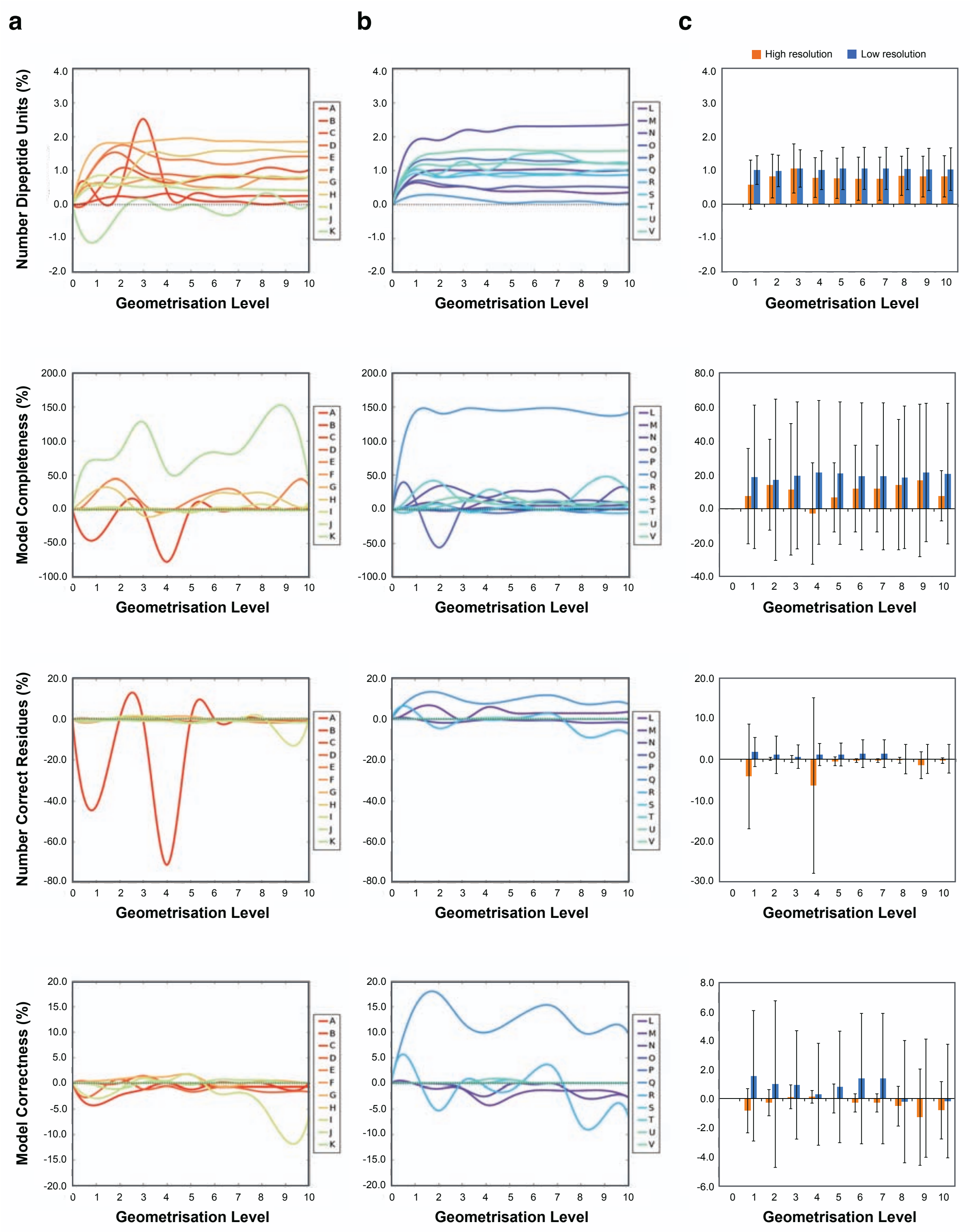


Figure 42 Effect of the level of geometrisation on the number of correctly built peptides in the first round of model building, total residues built and the correctly built residues in the final model, as well as the overall correctness of the finally built model. The absolute improvement for each model for (a) the high-resolution data, (b) the low-resolution data, (c) the average improvement over no rounds of geometrisation (rounds = 0) represented by bars and the minimum and the maximum improvement denoted by lines.

5.2.1. Test Case Selection and Benchmark Analysis

A number of jobs submitted to the EMBL cluster (as of September 2014) with a non-confidential dissemination level, at a resolution higher than 4.0 Å and a reference structure found in the PDB, were selected as test cases. The reference structure was identified using *blastp* [240] against the PDB, provided that the sequence identity was higher than 99%, the test case and the reference structure had the same number of chains and the total number of residues did not differ by more than 5%. A total of 22 test cases were selected, with the resolution of the data ranging from 1.2 to 3.7 Å, containing 91 to 930 residues in 1 to 4 chains in the asymmetric unit (Table A.1). These were further divided in the high-resolution group (comprising 10 cases with the data better that 2.5 Å) and the low-resolution group (the remaining 12).

For each case, 11 different model-building jobs were executed with different number of geometrisation rounds: ranging from 0 (no geometrisation at all) through 1 (default in ARP/wARP 7.4) up to 10; and the number of identified dipeptides and the final model quality were compared. In order to evaluate the model correctness, the built model was compared to the reference structure by using ARP/wARP *peptcomp*, which denotes a residue as correctly built if the displacement of its Cα atom with respect to the correct one is not higher than 1.0 Å, the chain direction is correct and the O atom is located in the correct 'hemisphere' of the peptide unit. It outputs the number of correctly and incorrectly built residues allowing an estimation of the model completeness and correctness as follows:

$$\text{Model completness (\%)} = \frac{T_{residues}}{P_{residues}} \times 100 \qquad (74)$$

$$\text{Model correctness (\%)} = \frac{C_{residues}}{T_{residues}} \times 100 \qquad (75)$$

with $T_{residues}$ as the total number of built residues, $P_{residues}$ the expected number of residues and $C_{residues}$ the number of correctly built residues.

5.2.2. The Optimum Level of Geometrisation

Looking at the averages of the benchmarking tests, it is not straightforward to point to the optimum level of geometrisation, as the effect is highly dependent on the individual test cases (Figure 42). It is evident that an increase in the number of geometrisation rounds promotes an increase in the number of dipeptides identified in correct conformation as defined by the Ramachandran-like plot (around 1.0% improvement). This has a higher effect on the low-resolution cases, but given the variations observed the observation is not statistically significant. From 0 to 3-4 geometrisation rounds there is an overall increase in the number of dipeptides found, reaching a plateau with additional rounds (Figure 42). With this, there is an increase in the number of residues built by 20% on average for the low-resolution cases and 15% for the high-resolution group, but not the number of correctly built residues or the overall model correctness. The average change in the total number of residues built is always positive in overall, except for the high-resolution group at the fourth round of geometrisation (Figure 42c).

Closer inspection showed that this negative influence is only due to test case A (Figure 42a). When it is ignored, the average effect becomes positive and higher than the previous levels. The same is observed for the average change in the number of correctly built residues and model correctness. This may be due to the data itself or an indication that this level of geometrisation is already too high for the atomic detail of the map at 1.2 Å resolution. At this resolution regime, the free atoms are already placed close to the real atomic positions and this level of geometrisation may be too high already and drives the free atoms to positions too far from the real atomic positions. Another factor affecting the model correctness may be the 1.0 Å displacement threshold used with *peptcomp* for the classification of a dipeptide to be correct, which may be too stringent at lower resolution. The level of 4 rounds of geometrisation was then implemented in ARP/wARP (as of version 7.5). This level provides the maximum improvement for the test cases studied in the number of correctly built residues, both at low and high resolution. It is evident, that further investigation on this topic is required in the future.

5.3. Validation of The Conformation of Dipeptide Units During Protein Automated Model Building

Within the ARP/wARP pipeline, the Ramachandran–like plot is used in two steps of autotracing (Figure 41 and Figure 43): *hmain* and *pept*, but as discussed in chapter 3 it lacks important conformational information. ARP/wARP *hmain* searches for putative peptide planes located on the free-atoms mesh representation of the electron density and evaluates their conformation (Figure 43a). The best fragments are selected by *pept* according to their length and conformation, breaking potentially incorrect links (circular fragments or Cα candidates that have more than one incoming or outgoing connections, etc; Figure 43b). With this, the possibility of improving the quality of the automatically traced protein models by complementing the use of the Ramachandran-like plot with the use of a DipScore threshold was tested.

5.3.1. Implementation of The DipSpace-Based Validation Method

The DipSpace 3-dimensional space described in section 4.1.1 was implemented to be used by both ARP/wARP *hmain* and *pept* in order to compute the DipScore for each possible dipeptide unit. While *pept* as well as DipCheck use a random uniform noise model, *hmain* uses the noise model represented by set E (Figure 28): a random sampling of not yet geometrised dipeptide units. With this, for a given dipeptide unit, its coordinate in the DipSpace and the sign of the chiral volume of the first peptide unit are computed and the DipScore of the corresponding point in the DipSpace subspace is then obtained using 3D parabolic interpolation. If the DipScore value is below a given threshold, the dipeptide unit is excluded from further consideration.

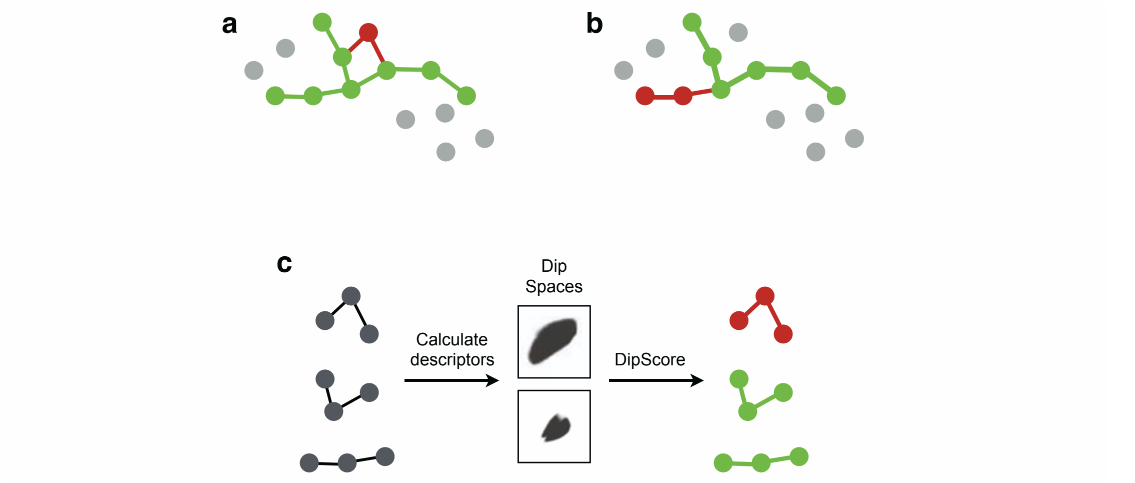


Figure 43 ARP/wARP auto-tracing. (a) The concept behind *hmain*. The connectivity of the individual peptide units is checked so that if two peptides form a dipeptide unit with an unlikely conformation, their connectivity (in red) is excluded from further consideration. (b) The concept behind *pept*. After all possible routes for the main chain are found and the fragments built, overlapping links and alternative routes are eliminated if a given dipeptide unit formed by two linking peptides has improbable conformation. (c) The decision-making based on the calculation of the distance geometry-derived conformational descriptors and the derived DipScore from the noise model and the PDB-derived DipSpace clouds. If the DipScore is below a given threshold, the dipeptide unit is excluded (in red).

5.3.2. Test Case Selection and Benchmark Analysis

In order to test the ARP/wARP performance with the different DipScore thresholds for both *pept* and *hmain*, a set of real test data was used. The set included the jobs submitted to the EMBL cluster (as of November 2015) with a non-confidential dissemination level, a resolution higher than 3.0 Å and a reference structure found in the PDB. A total of 16 test cases were selected, ranging from 1.2 to 3.0 Å resolution, having 1 to 4 chains in the asymmetric unit and the total number of residues ranging from 91 to 888. This set was further divided into two the high-resolution group (better or equal to 2.5 Å) and the low-resolution group with a resolution worse than 2.5 Å. A summary of the test cases used is given in Table A.2.

In chapter 4, three DipScore thresholds were defined (the favoured, allowed and generously allowed limits). These values cannot be used by ARP/wARP given the different noise models implemented. Consequently, the effect of different *hmain* and *pept* DipScore thresholds on the quality of the protein models built with ARP/wARP was evaluated. Starting from a zero threshold (no DipScore-based exclusion), DipScore thresholds up to 0.05 were tested with a step of 0.005.

For each of the 16 test cases, a total of 121 ARP/wARP jobs were run, using default parameters and the NCS-based extension [169] of fragmented models. At each cycle of model building, the obtained model was compared to the reference structure with the ARP/wARP *peptcomp* routine. Given the likelihood of an improvement of the model geometry at the first stages of model building to also affect the placement of the side-chains, model completeness and correctness were estimated with equations 74 and 75 together with the average fragment size and sequence coverage with equations 76 and 77:

$$\text{Average fragment size} = \frac{T_{residues}}{T_{chains}} \tag{76}$$

$$\text{Sequence coverage (\%)} = \frac{D_{residues}}{T_{residues}} \times 100 \tag{77}$$

where $T_{residues}$ is the total number of built residues, $D_{residues}$ the number of correctly docked amino acids and T_{chains} the number of built fragments.

For each model, the χ_{score} (eq. 69) was computed with DipCheck, the percentage of the Ramachandran outliers with MolProbity [39] and the N.N.r.m.s.d. between the built models and the reference structure with the ARP/wARP *watercomp* routine using the maximum distance of 2.0 Å around each atom. Comparison was performed for both the main-chain only and all atom models, excluding solvent, ligand and free atoms. For each measure, the median deviation was used to evaluate how the different DipScore combinations affect the model building.

5.3.3. Optimisation of DipScore Thresholds

One would expect that an increase in the DipScore threshold in *hmain* would promote an improvement of the model correctness and a reduction of the main-chain r.m.s.d. to the reference structure. An increase in the *pept* DipScore threshold would, on the other hand, affect the fragmentation and the overall completeness, but also improve model correctness. The results from the carried out benchmarking are summarised in Figure C.5-7. The first observation is that a gradual change of the *pept* and *hmain* thresholds does not translate to a smooth change in the model parameters. However, an improvement in the median model correctness (evaluated by *peptcomp*), the r.m.s.d. of Cα positions, the χ_{score} and the percentage of Ramachandran outliers is always observed for all threshold combinations (Figure C.6-6). This is more prominent for the lower resolution cases than for the higher resolution set, as the models built with high resolution data even without the DipSpace validation methods are already more than 90% complete and correct, have a low r.m.s.d. to the reference model and consist of long fragments.

The choice for the best combination of the thresholds was based on a compromise between model completeness, average fragment size, correctness and proper geometry, as the variation of each parameter within each combination tested is always very high (e.g., Figure 44). Applying a higher DipScore threshold in *pept* than in *hmain* would not make much sense, as *hmain* performs its task before *pept* and, therefore, a lower threshold in *hmain* could result in building less plausible fragments. On the contrary, by using a higher DipScore threshold in *hmain*, unlikely peptide unit connections are filtered at the first stage and *pept* then only needs to find which overlapping fragments have to be disconnected. With this, a DipScore threshold of 0.035 was implemented in *hmain* and 0.010 in *pept* (Figure C.5-7 and Figure 44).

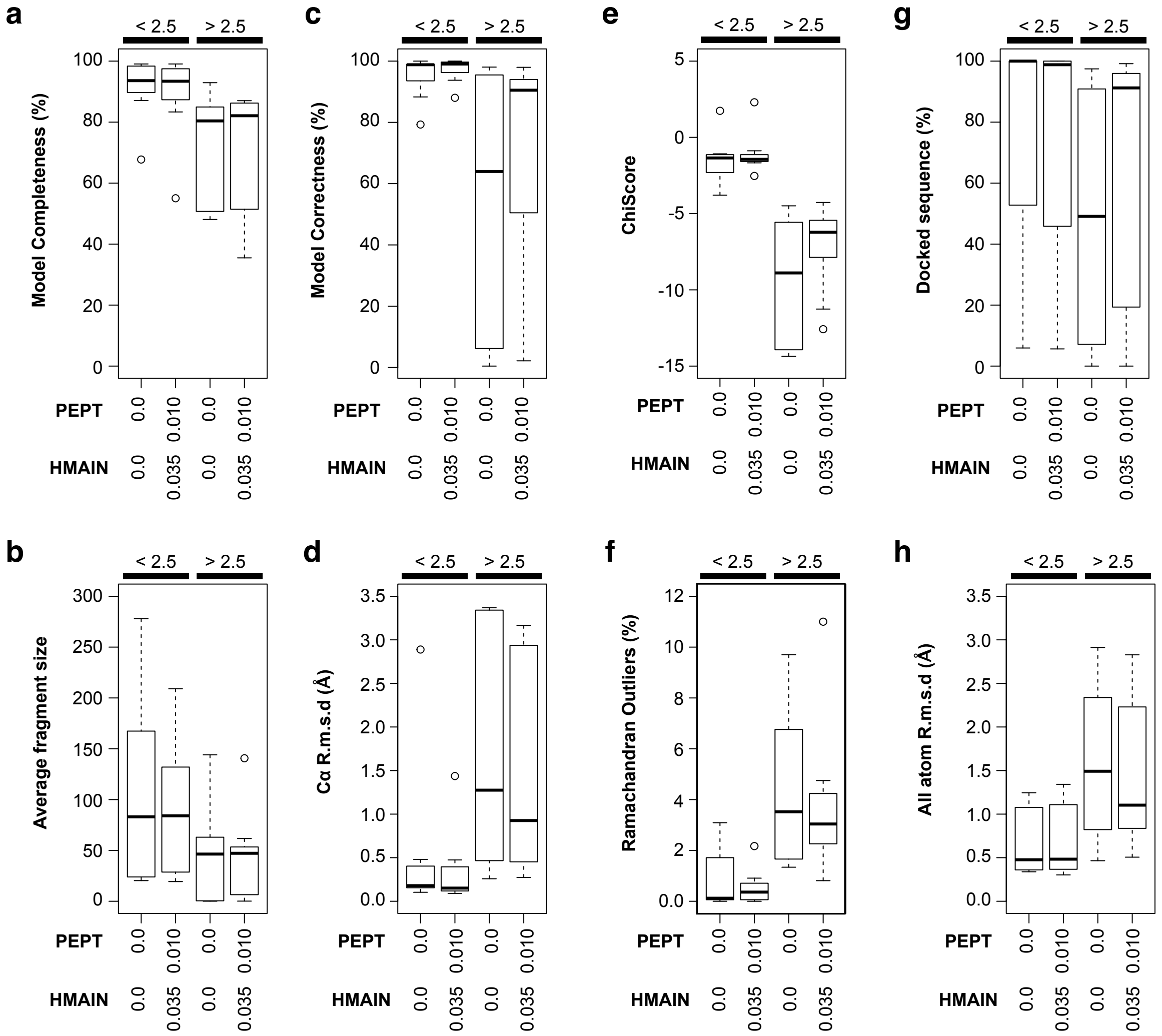


Figure 44 Boxplots depicting the effect of use of the different DipSpace thresholds on the quality of the protein models built with ARP/wARP for the high (< 2.5 Å) and lower (> 2.5 Å) resolution.

5.3.4. Effects at High and Medium-to-Low Resolution

The selected combination of the DipScore thresholds does not affect the quality of the models built from high-resolution data (Figure 44). The median model completeness stays, although it varies more when DipSpace is activated. The median model correctness increases slightly and most of the test cases have a correctness higher than 95%, with a lower local r.m.s.d., a χ_{score} close to zero and a very low percentage of the Ramachandran outliers. More striking results are obtained for the lower resolution set. The use of the DipSpace promoted a median increase of the model correctness of lower resolution models from approximately 60% to 90%, with a corresponding reduction of the Cα r.m.s.d. to the reference model, but also with a slight increase of the model completeness. The median χ_{score} also improved drastically from about -10 to -5, with lower variation and a concurrent reduction in the percentage of the Ramachandran outliers. Another noticeable result is a considerable improvement in the docked sequence from about 50% to more than 90% in the lower resolution set, and the reduction of the all-atom r.m.s.d. (Figure 44). While DipSpace does not affect the sequence docking directly, the more correct tracing of the main-chain is essential for the sequence docking, which can be explained by the fact that the current ARP/wARP sequence docking is based on the location of the free atoms and on the density quality around Cα positions [162].

5.3.5. An Extreme Test Case

From the benchmark analysis, for one of the medium-to-low resolution test cases an extreme improvement was obtained when DipSpace was used with the selected thresholds: test case O (Table A.2). It has 100% sequence identity to the chicken C-Src kinase domain (PDB ID 2oiq). The reference model is an alpha-beta protein complex composed by two chains, A and B (Figure 45a), each with about 300 residues, which was refined at a resolution of 2.1 Å. Here, the data extend to 2.9 Å resolution. Without the use of DipSpace the autobuilt model (Figure 45b) was highly fragmented, with an average fragment length of 6 residues, only 54% complete and with only 5.4% of the built residues identified as correct. DipCheck identified 5.3% of the residues in disallowed regions and 2.4% and 20.8% in generously allowed and allowed areas, respectively, classifying the overall model as disallowed given its χ_{score} of -14.4. The use of DipSpace allowed the building of 82.2% of the model, with an average fragment size of 40 residues, a correctness of 94% and an average r.m.s.d. of 0.5 Å to the Cα positions in the reference model (Figure 45c). Only two residues were now in the disallowed region of the DipSpace and 92.4% have a favoured DipScore. The χ_{score} also improved considerably to -5.4. Although it is still disallowed, the improvement indicates that the geometry has become more favourable.

The alignment of the autobuilt and the reference models showed three main regions with high r.m.s.d. (Figure 45d-h). The first one is the stretch between Glu275 and Cys277 (Figure 45d-f), in both A and B chains. In chain B, this region is not built, but in chain A it is modelled and was assigned to sequence, with Gly276 deviating by 1.5 Å from its reference position. When both the reference and the modelled molecules are checked against the density (Figure 45f) none of them fits perfectly but the automatically traced model seems to fit poorly and shows a poorer geometry. In the second region (Figure 45g), the automatically traced main-chain (and also side-chains) fit the electron density well.

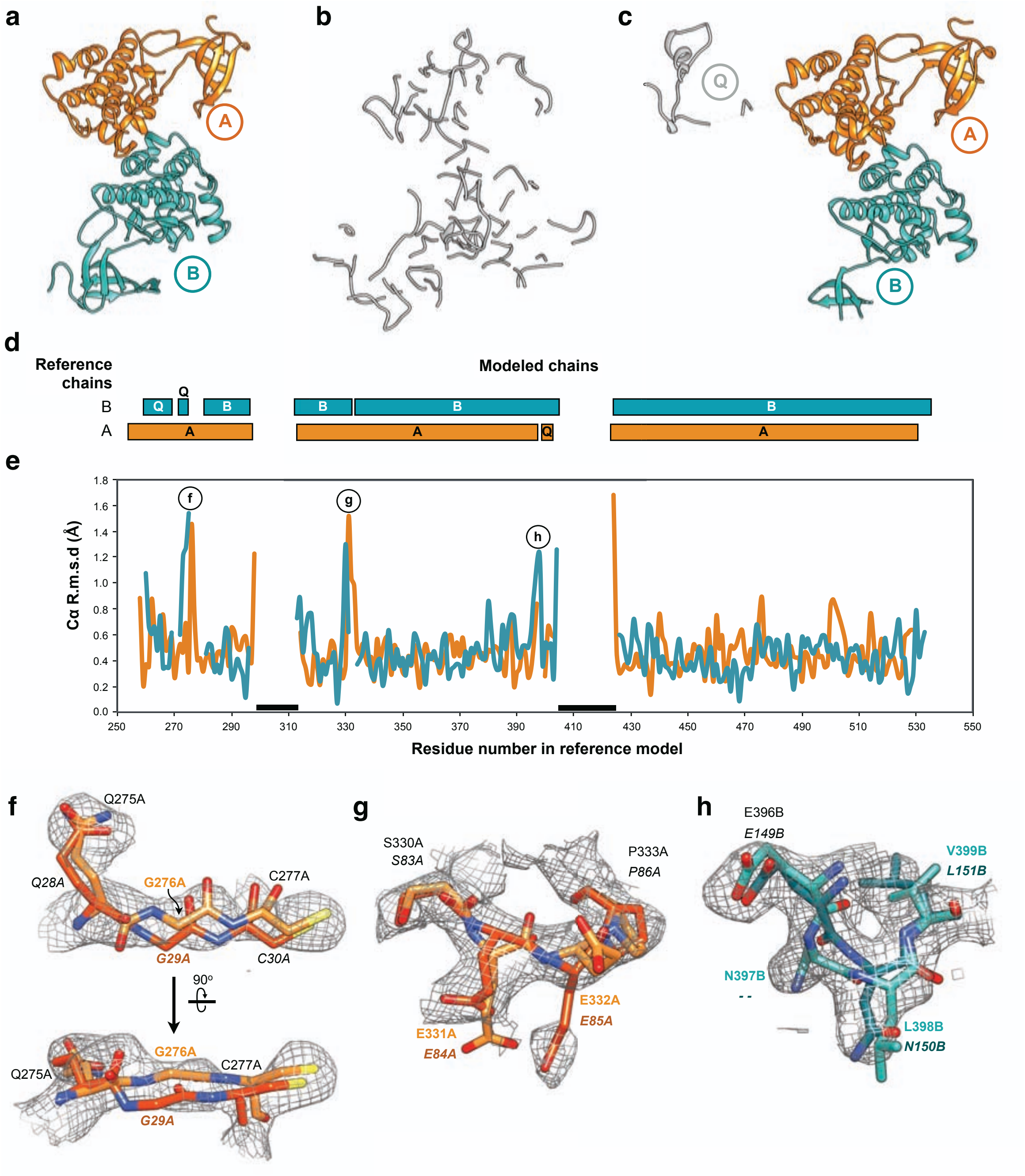


Figure 45 Test case O. (a) The reference model with two chains in the structure (PDB ID: 2oiq). (b) Automatically traced model without DipSpace. (c) Automatically traced model with DipSpace. (d) Alignment of the reference and the automatically traced models (with DipSpace activated). (e) Local Cα N.N.r.m.s.d between the reference and the automatically traced models. Orange lines correspond to the reference chain A and blue lines to chain B. Two black bars mark non-modelled regions where the electron density is absent. (f-g) Close up of the three regions with higher r.m.s.d., whose positions in the reference model are depicted in (e). The residues with the higher r.m.s.d. are marked in coloured bold. Orange corresponds to chain A and blue to chain B, darker colours to the automatically traced model and lighter colours to the reference model. Labels in italic correspond to the automatically traced model.

The third region (Figure 45h) is broken for chain A but modelled for chain B with a residue-shift error. This region has four residues in the reference structure but was automatically traced with three residues only, resulting in a sequence mismatch. This region is characterised by somewhat broader density, which presumably confuses the main-chain tracing. One important observation about these three regions is that they are not equally represented in both chains: there is always a break in one of the chains that cannot be improved by the NCS-related parts. Also, there are two weak-density regions that are never built for both chains (Figure 45d-e).

This is an extreme case in the benchmark demonstrating the potential of the DipSpace as a complement to the Ramachandran-like plot in ARP/wARP for building of lower resolution models. A deeper understanding of the effect of the DipSpace on the different stages of the model building process may provide means for further exploitation of its potential. It would be also interesting to further evaluate its effect on sequence docking.

5.4. Eigenvalue-Based Identification of Side-Chains

At medium-to-low resolution, the free atoms not accounted for main-chain building and representing the side-chains density are not necessarily placed close enough to the true atomic positions. More importantly, the number of such free atoms can differ from the true number of atoms in this part of the model (Figure 10b). Therefore, one-dimensional connectivity vectors used for side-chain docking are not powerful at this range of resolution. At this stage, ARP/wARP is faced with a mesh of k free atoms, which represents a density grid and the shape of the side-chains density cluster. A possible alternative would be a description of topology of each amino acid directly by a set of density points. They can be seen as a three-dimensional graph with n points (with $n > k$) that can be mathematically treated to obtain information about the density shape. Distances between these points provide information about the conformation of the side-chain while their connectivity reveals the topology, independently on the conformation. The underlying idea is that a use of higher number of graph points (vertices) may compensate for the coordinate error in the position of each vertex.

As overviewed in section 2.5, several matrices can be used to describe the shape of a molecule when we look at it as a graph. In the case of side-chain topology and electron density shape description, the 3-dimensional mesh is a simple, non-directional and finite graph and its connectivity can be described by a symmetric squared adjacency (0,1)-matrix with zeros on its diagonal. Given that the eigenvalue spectra of such matrices are widely used to study the properties of graphs (e.g., isomorphism) [241], the eigenspectra of density meshes could potentially be applied for the derivation of a fingerprint for each side-chain topology at a given resolution range. The concept behind the proposed approach is shown in Figure 46. A given side-chain density cluster would be first represented as a mesh of n density points, whose connectivity is stored in an adjacency matrix. The eigen-decomposition of the adjacency matrix provides a set of *n* real eigenvalues; with *i* zero eigenvalues (corresponding to the number of connected clusters) and *n-i* non-zero eigenvalues. By comparing the number and the magnitude of the non-zero eigenvalues with other density clusters, a sequence identity could be given to it. The application of this concept for the differentiation between different density clusters corresponding to different side-chains at medium resolution was investigated, by calculating the eigenspectra of adjacency matrices for a few different side-chain topologies in a given protein model. The preliminary results obtained are described in the next sections.

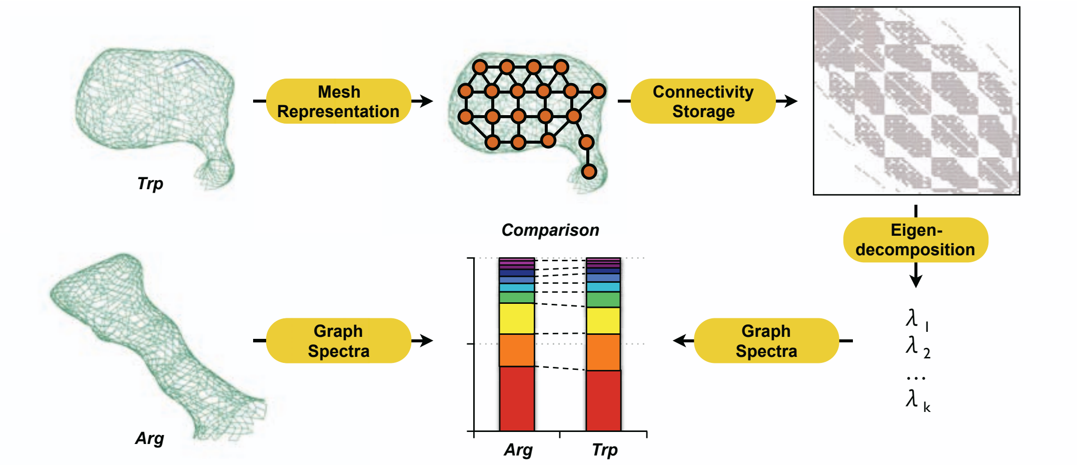


Figure 46 The concept of the proposed new method for side-chain identification and sequence docking with ARP/wARP.

5.4.1. Test Case Selection and Eigenspectra Computation

The crystallographic model of *Leishmania major* leishmanolysin (PDB ID 1lml) was re-refined with REFMAC5 [122] using default settings and cutting the highest resolution limit to 2.0 Å. The density clusters of four side-chains with different topologies and distinct shapes were investigated: (1) tryptophan, characterised by a large and bulky density cluster; (2) valine, having a small density cluster; (3) asparagine, similar to valine but longer; and (4) arginine, a long and flexible side-chain that can show density clusters with a wide range of shapes.

To obtain the density cluster for a given side-chain, a sphere of a given radius (*r*; different for different side-chains) was centred in a middle position, defined as the central atom in the side-chain or the average position between the two most central atoms of the side-chain in real space. The density inside this sphere was selected using ARP/wARP *mapplus*. While for tryptophans, valines and asparagines only one sphere (r = 3.0 Å, 2.0 Å and 2.5 Å, respectively) was enough to obtain its density cluster without the interference of the surroundings, in the case of arginine its long chain had to be broken into two fragments at its Cδ-Cε bond and the process repeated for both fragments (as shown in Figure 47).

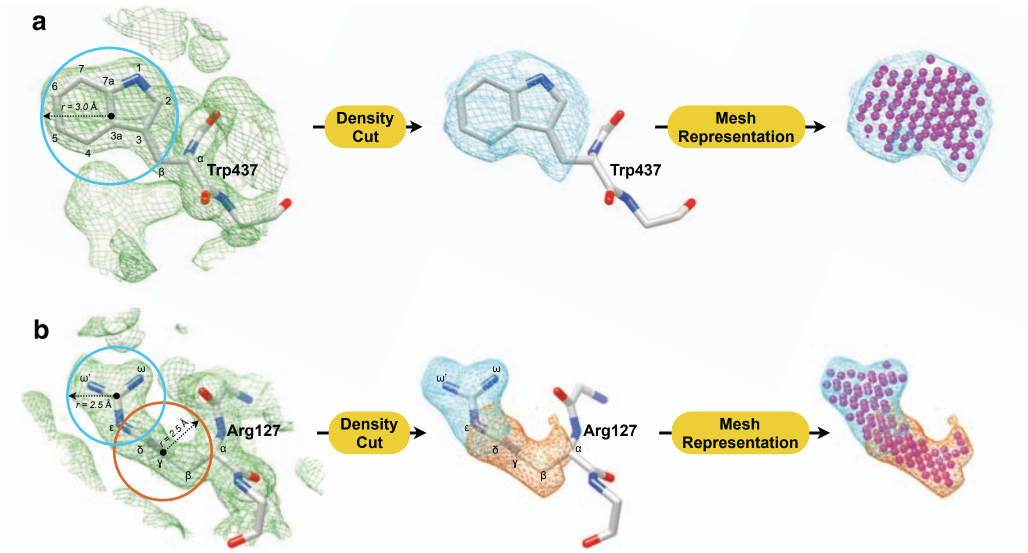


Figure 47 Side-chain density extraction and mesh representation for (a) Trp437 and (b) Arg127 from the 2.0 Å resolution model of leishmanolysin.

The obtained side-chain density clusters were then represented as a mesh of free atoms with the ARP/wARP *mapread* routine. Those with the density below 0.6 e/Å^3 above the mean were excluded. The adjacency matrix of the collected mesh was computed and its eigenvalue spectra calculated. Two mesh points were marked as adjacent if the distance between them was less than 2.0 Å (Figure 48a). A total of 5 tryptophans, 10 valines, 5 asparagines and 7 arginines were analysed. For each mesh, the number of non-zero (non-null) and total eigenvalues, the magnitude of the largest and the smallest eigenvalues as well as the spectral gap and chromatic number (section 2.5.1) were compared. As one would expect, some of the variations observed for the largest and the smallest eigenvalues, the spectral gap and the chromatic number are affected by the total number of nodes in the graph and vary due to the different quality of the density clusters; these were normalised by dividing their values by the total number of points in the mesh. In order to assess if there are significant differences between different side-chain meshes, two-sided t-tests were carried out for the difference in the means (H_0: the means are equal; H_1: the means are different), as implemented in the *t.test* R function [233].

5.4.2. Side-Chain Mesh Representation and Eigenvalue Spectra

The results obtained are summarised in Figure 48 and Table B.11-16. Meshes describing the density clusters of larger residues are on average composed by a larger number of points (Figure 48b). These meshes have as many eigenvalues as the number of points (Figure 48c). Therefore, their adjacency matrix has its maximum rank. The largest eigenvalue (λ_1), related to the meshes average degree [241], [242], is also always higher for larger side-chains (Figure 48d). It is, for any side-chain, always smaller than the number of nodes, meaning that the mesh is not a complete graph but a regular graph where not all nodes are connected to all others.

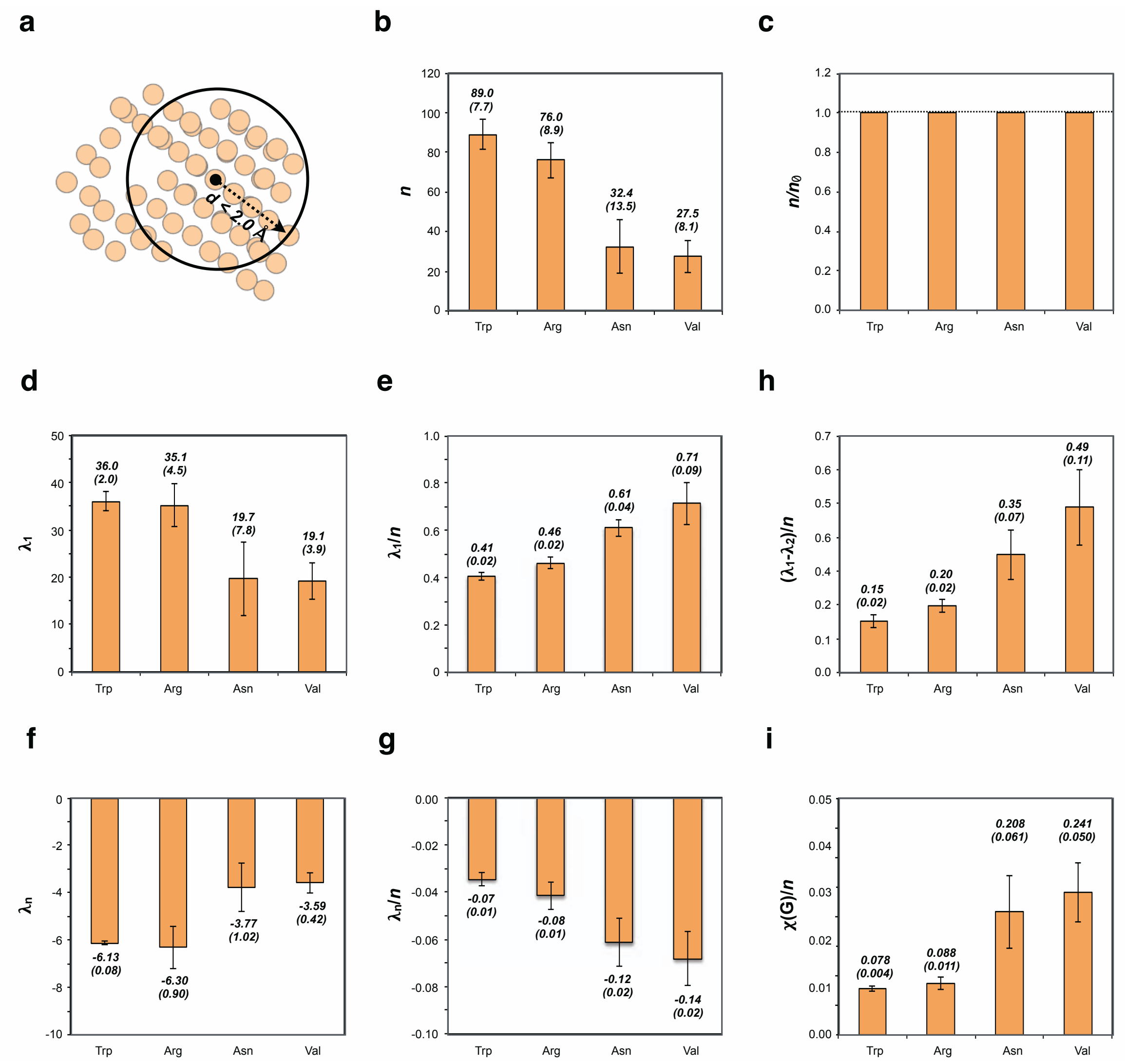


Figure 48 Eigenspectra of side-chain electron density meshes. (a) Two-dimensional projection of the side-chain mesh computed for leishmanolysin Trp7, depicting the maximum distance two mesh-points should have in between to be considered adjacent. (b-i) Average, and respective standard deviation in parentheses, (b) number of nodes (n), (c) number of nodes over the number of non-null eigenvalues ($n/n_{ø}$), (d) largest (λ_1) and (e) normalised largest (λ_1/n) eigenvalue, (f) smallest (λ_n) and (g) normalised smallest (λ_n/n) eigenvalue, (h) normalised spectral gap ($(\lambda_1$-$\lambda_2)/n$) and normalised chromatic number ($\chi(G)/n$).

While large side-chains (tryptophans and arginines) have a λ_1 significantly larger than small side-chains (asparagines and valines), the same is not observed within each group (Table B.12). This is resolved by dividing the largest eigenvalue by the total number of mesh points (Figure 48e and Table B.14), providing also an estimate of the average percentage of mesh points that are adjacent to each other (Table B.11). On average, each mesh point in a tryptophan density cluster is adjacent to about 40% of mesh points, while in arginines this increases to 45%. In asparagines each mesh point is adjacent to about 60% of the points and in valines to about 70%. The spectral gap, related to the connectivity of the mesh [241], [242] is approximately the same for each side-chain mesh (12 to 14). This means that all meshes are highly connected, without any disconnected set of mesh points. The normalisation by the number of points shows the proportion of mesh points that would be disconnected by the removal of the edges and increases the difference between the different side-chains (Figure 48h). As one would expect, a higher proportion of mesh points would be disconnected for smaller side-chains, with valines and asparagines separated with a confidence level of 99% (Table B.16).

The smallest eigenvalue (λ_n) also has a higher value for larger side-chains, almost without any variation for tryptophans (Figure 48f and Table B.11). This eigenvalue relates to the bipartiteness of the mesh [241], [242] and, as $|\lambda_n| < \lambda_1$ for all side-chains, none of them represents a bipartite graph: none of them can be divided into two different sets of edges without the removal of edges. The variation observed due to the variation in the number of mesh points does not allow the separation of different side-chain meshes with approximately the same size (Table B.13) but the normalisation by the number of points does (Figure 48g and Table B.14). The ratio between the largest and the smallest eigenvalue relates to the graph chromatic number by equation 52. As the spectral gap, it is of approximately the same value (about 6-7) for all side-chain meshes. The normalisation by the number of points allowed a better separation of the different groups (large and small side-chains) but not of side-chains with approximately the same size (Table B.17).

These preliminary results suggest that the eigenspectra of side-chain meshes, representing the shape of the electron density clusters, have a potential to be used as a fingerprint for the assignment of the protein sequence, with the normalised largest eigenvalue and spectral gap showing a stronger signal for the separation between the four different side-chains evaluated (Table B.14 and Table B.16). However, the population of side-chains tested was small and, therefore, the significance of the obtained results is yet to be proved. It would be of high interest to test this approach over a larger set of side-chains, and also to compare these results with those of the molecular graph representing the full topology of the side-chain and the same residue at different resolution ranges.

5.5. Use of Energy Term For The Enhancement of Ligand Fitting

In addition to the development and improvement of ARP/wARP protocols for the automated building of protein models, a short visit was made to the ‘ligands world’. This side project was preformed collaboratively with Dr. Daria Beshnova from the Lamzin group at the EMBL in Hamburg and was driven by the joint usage of validation tools and model building for the improvement of the quality of the small molecule models automatically built by ARP/wARP. Daria Beshnova implemented the method and preformed all calculations, while I contributed to the design of the method, the analysis and the interpretation of the results.

As described in section 1.3.4, a number of tools exist that provide means to study protein-ligand interactions and MX is used as the main experimental technique for structural analysis. However, the methods for the automatic interpretation of the electron density representing the ligands are not as well developed as those for protein model building. Additionally, there is a consensus within the structural biology community that more attention should be devoted to the analysis of protein-ligand contacts and the interpretation of corresponding electron density before the structural model is deposited in the PDB. In particular, additional means for the validation of the built models are seen to be necessary [243].

The methods for ligand fitting and identification maximise a scoring function that takes into account the shape of the electron density in the ligand binding site. While this is an obvious approach to follow, chemically different ligands can have similar shape and, therefore, an incorrect ligand can be built by mistake. This is of higher importance during ligand guessing. One hypothesis is that by taking into account the environment in the binding site, ligand guess could be improved. Therefore, a novel approach, *LigEnergy*, for the evaluation of protein-ligand binding in MX was developed. It can be used either for the validation of the already built ligands or as a supplementary scoring function during the process of guessing the ligands and the evaluation of their fit to the electron density. The method is based on the estimation of the inter-molecular energy of the protein-ligand binding calculated as the sum of Van der Waals (VDW), H-bond and electrostatic interaction energy terms [244]. It offers a one-parameter estimation of the quality of protein-ligand models that allows the fast scanning of large databases and the identification of 'questionable' structures. It also allows improving the identification and fitting of ligands into specified electron density with ARP/wARP, as described below.

5.5.1. Test Case Selection

In order to test weather the intermolecular energy of the observed binding mode can be used for the validation and detection of plausible or 'questionable' protein-ligand complexes, 100 models were randomly selected from the PDB excluding covalently bound ligands. The ligand size varied from 10 to 50 atoms and the structures were solved at a maximum resolution between 1.5 and 3.5 Å. Solvent molecules and ions were excluded. The model coordinates, crystallographic data and the RSCC values were obtained from the EDS [245]. The RSCC varied from 0.49 to 0.99, as reported by the server. The full list of PDB ID codes, 3-letter ligand identifiers, ligand size, maximum MX resolution and RSCC are provided in Table A.3. To test the approach for the identification of crystallographic ligands in a specified electron density cluster, three test cases (PDB IDs 1cx4, 2x4o and 2of1) were used where the ligand guess method implemented in ARP/wARP v.7.6 identified ligands incorrectly.

5.5.2. Inter-Molecular Energy Calculation and Protein-Ligand Complex Validation

For all protein-ligand complexes considered, the energy of pair-wise atomic interactions (*V*) was computed using AutoDockTools, as described in [244] and given by equation 78:

$$V = W_{VDW} \sum_{i,j} \left(\frac{A_{ij}}{{r_{ij}}^{12}} - \frac{B_{ij}}{{r_{ij}}^{6}} \right) + W_{hbond} \sum_{i,j} \left(\frac{C_{ij}}{{r_{ij}}^{12}} - \frac{D_{ij}}{{r_{ij}}^{6}} \right) + W_{ele} \sum_{i,j} \left(\frac{q_i q_j}{\varepsilon(r_{ij}) r_{ij}} \right) \tag{78}$$

The first term corresponds to the 6/12 Lennard-Jones potential for the dispersion/repulsion interactions, with parameters A and B taken from the Amber force field [246]. The second term presents an H-bond energy estimated by a 10/12 potential [247]. The last term is the energy of electrostatic interactions, based on the Coulomb potential, with the distance-dependent dielectric constant $\varepsilon(r_{ij})$ [248]. The parameters C, D, $\varepsilon(t)$ and optimised weights W_{vdw}, W_{hbond}, W_{ele} were calculated according to [244].

The addition of hydrogen atoms and the assignment of partial charges to the ligand and the protein atoms are essential for the estimation of the hydrogen bonds and the electrostatic terms. Marsili-Gasteiger partial charges were calculated on the basis of electronegativity equilibration using the ‘partial equalization of orbital electronegativities’ method [249]–[251]. Since the positions of hydrogen atoms in structures obtained by X-ray techniques are not well defined, here we considered only polar hydrogen atoms, which are important for the H-bonding. After the polar hydrogen atoms were placed, the partial charges computed using AutoDockTools [252], [253].

Given that larger ligands tend to have higher energy of interactions, with a linear correlation of 0.73 to the number of non-hydrogen atoms in the collected 100 ligands (Figure 49a), the calculated binding energy was normalised by the number of atoms in the ligand. This normalised energy was used further for the evaluation of the correctness of the ligand-binding mode, combined with the RSCC as an estimator of the overall fit of the model to the electron density. This was both used to validate the 100 protein-ligand complexes selected from the PDB but also to provide a final ranking of the 40 identified candidate compounds with ARP/wARP 7.6 ligand-guess method for the three test cases. In the next release of ARP/wARP we intend to provide the assignment of partial charges and the addition of hydrogen atoms using local geometry [254] and the estimation of the protein-ligand interaction energy with equation 78.

5.5.3. Energetic Validation of Deposited Models

The normalised intermolecular energy computed for the test set of 100 protein-ligand complexes has a bell-shaped distribution (Figure 49b). The majority of the models show negative interaction energy, below -0.15 kcal/mol per atom, and a good fit to the density (RSCC ≥ 0.85). Five of them (PDB ID: 4dma, 4dt2, 4gcq, 4dkp, 3sgy) show an RSCC between 0.71 and 0.80 and their detailed visual inspection didn't reveal any unusual contacts between the ligand and protein atoms. Two cases (with RSCC of 0.95 and 0.49, respectively), however, showed positive or close to zero interatomic energy values (Figure 49b and Figure 50). These two models were, therefore, classified as 'questionable' and were considered in more detail.

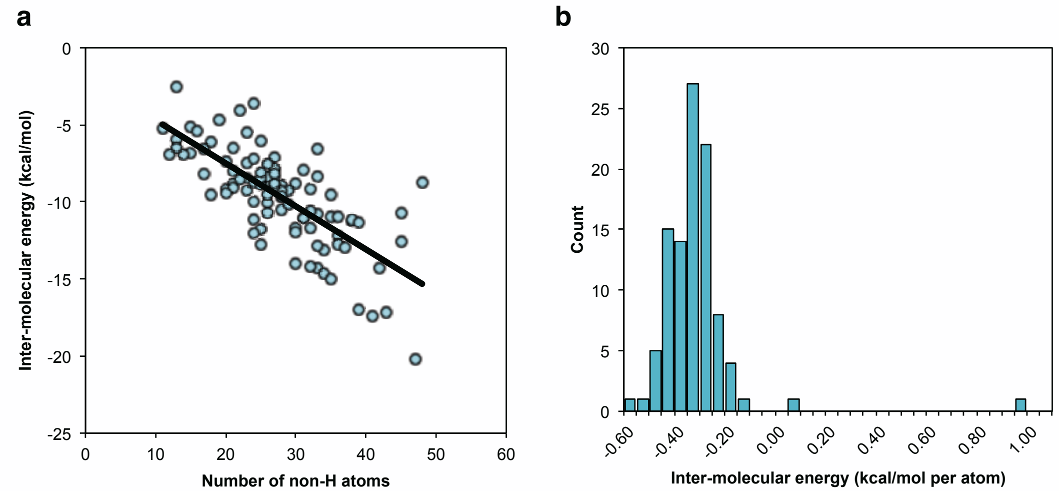


Figure 49 Estimated intermolecular energy of 100 randomly selected protein-ligand complexes from the PDB. (a) Correlation between the number of non-hydrogen atoms in the ligand and the overall protein-ligand interaction energy. Two outliers with positive intermolecular energy were excluded. (b) Distribution of the estimated intermolecular energy per atom.

The first complex (PDB ID 3ue8) is that of the human kynurenine aminotransferase II and the 09M[1] ligand, regarded as a strong inhibitor of this enzyme and as a putative drug for the treatment of schizophrenia [255]. The ligand associated with chain A fits well into the electron density, with an RSCC of 0.95. However, its intermolecular energy is positive, 0.93 kcal/mol per atom, as a result of an unfavourable VDW contact (2.1 Å) between the O7 atom in the ligand phosphate group and the hydroxyl group of Tyr74B (Figure 50a). Some rotation of this phosphate group around its O3-P bond could eliminate this clash. The second ligand molecule (in chain B) looks unproblematic with good interaction energy of -0.42 kcal/mol per atom and RSCC of 0.94.

The second complex is between the *Pseudomonas aeroginosa* ceramidase and C2-ceramide (PDB ID: 2zxc) [256]; it has an unfavourable interaction energy close to zero, 0.07 kcal/mol per atom. The modelled ligand, associated with chain A, has poor electron density support, with an RSCC of 0.49, and several unfavourable contacts (Figure 50b): 2.6 Å distance from its C4 atom to the Ser334A CB atom and to the N atom of Thr335A. The second ligand, associated with chain B, also has positive interaction energy and low RSCC of 0.41.

While the overall normalised intermolecular energy per atom allows the detection of 'questionable' structures - those with non-negative values of the overall protein-ligand interaction

[1] (5-hydroxy-4-{[(1-hydroxy-2-oxo-1,2-dihydroquinolin-3-yl)amino]methyl}-6-methylpyridin-3-yl)methyl dihydrogen phosphate

energy - the inspection of energies for each individual atom in the ligand can also be informative. For example, the complex between the Drosophila class III PI3-kinase VPS34 and the 093[2] ligand [257] (PDB id 2x6j) has an overall energy of -0.15 kcal/mol per atom. However, its chlorine atom has a highly positive intermolecular energy of 4.9 kcal/mol, while all the other ligand atoms have their energies within the 0.0 to -0.7 kcal/mol range. The reason is a close contact (2.4 Å) between the chlorine and the amino group of the Lys698A side-chain (Table B.18). The majority of the described problems can be corrected with modern refinement tools. For example, a re-refinement of these three models using the PDB-REDO server [258] improved their geometry and eliminated the interatomic clashes, reducing the computed energies to -0.4 kcal/mol per atom too.

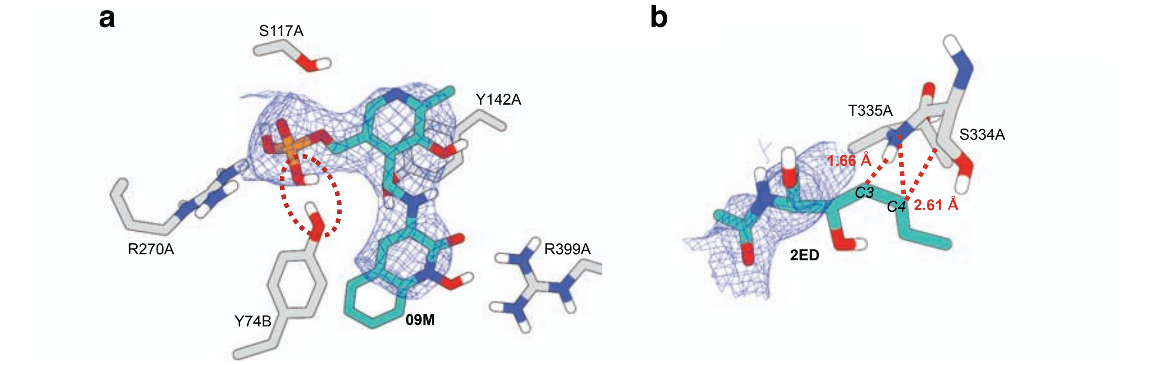


Figure 50 The binding site of the deposited models of (a) human kynurenine aminotransferase II in complex with 09M ($2mF_o$-mF_c map contoured at 1.4σ above the mean (0.28 e/Å^3) in blue mesh) and (b) *Pseudomonas aeruginosa* ceramidase in complex with C2-ceramide ($2mF_o$-mF_c map contoured at 0.9σ above the mean (0.28 e/Å^3) in blue mesh). The red dotted lines and the elliptical shape indicate problematic contacts. Grey dashed lines point to favourable interatomic interactions.

5.5.4. Enhanced Ligand Rebuilding and Fitting

Further investigations were carried out to test weather the intermolecular energy can be used to rank and identify the correct ligand in a given binding site with ARP/wARP, using therefore equation 78 as a second scoring function in addition to the RSCC. Three test cases where the ARP/wARP ligand guessing method was unable to detect the correct ligand were used (Figure 51). The first case is the complex between the mutant of the type II beta regulatory subunit of the murine cAMP-dependent protein kinase and adenosine-3',5'-cyclic-monophosphate (CMP) (PDBID 1cx4) [259]. The ligand-guessing protocol identified adenosine-5’-diphosphate (ADP) as the most likely binder for this enzyme (Table B.19), a structurally similar molecule to the correct ligand. However, this is not the most energetically favoured ligand (Table B.19). By using the LigEnergy approach, the ligand with the best interaction energy (-0.36 kcal/mol per atom) was indeed CMP, the deposited ligand (Figure 51a and Table B.19).

[2] N-(5-(4-chloro-3-(2-hydroxy-ethylsulfamoyl)-phenylthiazole-2-yl)-acetamide

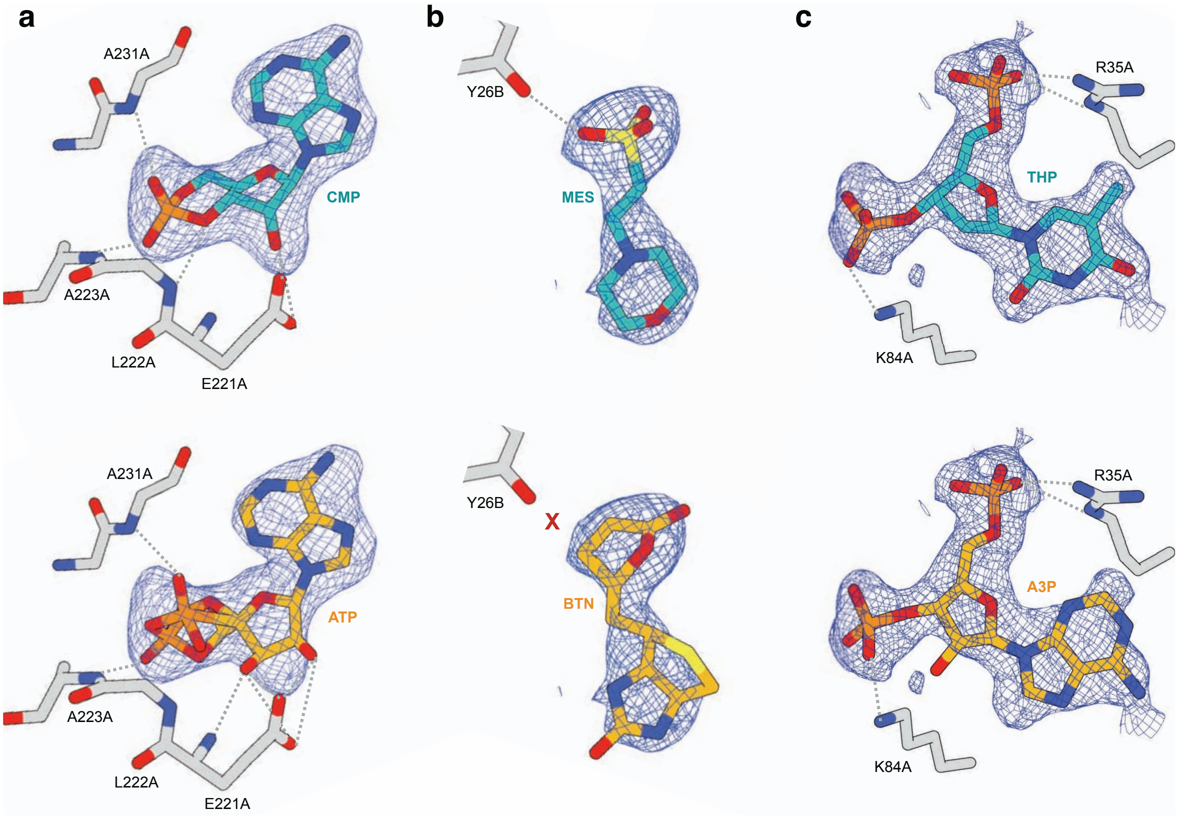


Figure 51 Ligand-guessing without (orange) and with (blue) the use of the estimated binding energy as an additional scoring function. Superposition between the two identified ligands for (a) the cAMP-dependent protein kinase (PDB ID 1cx4; 2mF$_o$-mF$_c$ map contoured at 2.5 σ above the mean (0.14 e/Å^3) in blue mesh), (b) the complex of MHC class I HLA-A2.1 and HIV-1 envelope peptide ENV120-128 (PDB ID 2x4o; 2mF$_o$-mF$_c$ map contoured at 3.6 σ above the mean (0.25 e/Å^3) in blue mesh) and (c) Staphylococcal nuclease variant truncated Delta+PHS I92W (PDB ID 2of1; 2mF$_o$-mF$_c$ map contoured at 2.5 σ above the mean (0.21 e/Å^3) in blue mesh). Dashed grey lines indicate favourable contacts.

The second case is the complex between 2-(N-morpholino)-ethanesulfonic acid (MES) and the complex of MHC class I HLA-A2.1 with the HIV-1 envelope peptide ENV120-128 (PDB ID 2x4o) [260]. The ligand-guess method identified biotin (BTN) as the most likely binder (Figure 51b) but, according to *LigEnergy* (Table B.20), MES was identified instead as the best interacting ligand with an energy of -0.21 kcal/mol per atom. This is due to the presence of hydrogen bonds and electrostatic interactions formed between MES and the protein residues, which are absent in biotin (Figure 51b). Another indication that biotin is perhaps not the right ligand is the formation of several close contacts between the ligand atoms as a result of its fit into the insufficiently large density cluster.

The third test case is the complex between thymidine-3',5'-diphosphate (THP) and the Staphylococcal nuclease variant Delta+PHS I92W (PDBID 2of1). The ligand-guessing method suggested adenosine-3’,5’-diphosphate (A3P) as the compound with the highest shape similarity to the selected protein binding site and the highest RSCC (Figure 51c and Table B.22). Using the

LigEnergy approach, A3P became second in the ranking and THP, the correct ligand, was identified instead, with the interaction energy of -0.28 kcal/mol per atom (Figure 51c and Table B.22). This result is particularly encouraging since THP and A3P are structurally very similar.

5.6. Concluding Remarks

In the beginning of this chapter, it was indicated that there are a number of areas where the improvement of the automated building of MX protein models by ARP/wARP could be attempted. The appropriate geometrisation of the identified putative dipeptide units in ARP/wARP proved beneficial and as from version 7.5, four rounds of geometrisation have been implemented. This improved the number of correctly built residues (more 15-20% on average), both for high- and for low-resolution cases. Further advances were obtained by complementing the Ramachandran-like plot with the DipSpace validation method during the protein auto-tracing step. Remarkable results were obtained for the test cases at a resolution lower than 2.5 Å. In this study only the final quality of the final model was addressed, as it is indeed the result the user aims to achieve, but the analysis of how the DipSpace mode and the different thresholds affect the model building run at each cycle and round, will most likely provide good insights for further improvement of the method.

Novel approaches for the interpretation of side-chain density clusters may supersede the connectivity vectors currently used for sequence docking. The preliminary results indicate that the normalised largest eigenvalue and the spectral gap of adjacency matrices allow for the discrimination between the different side-chains. Further tests are required at different resolutions for all residue types.

The use of a simple function to estimate the intermolecular energy of ligand binding pointed to problematic areas in already deposited ligand structures, and helped improve the automated guessing of ligand identity in a binding site. The results demonstrate that indeed the estimated energy of intermolecular interaction can serve as an additional scoring parameter for the identification of the most likely ligand for a selected site in a given protein model and that LigEnergy method has the potential for the improvement of a ligand-guessing protocol during the automated identification of ligands. Its use is appropriate at the final stage of ligand ranking, although in combination with other measures like the RSCC it may also be beneficial during the earlier stages. The joint use of the shape similarity between the ligand and the density cluster with the estimated intermolecular energy increases the chances of automatically identifying the most likely ligand representing the unknown binder. While the RSCC may not be sufficient, the energy should not be used by itself either, and these two important parameters should better be used together.

Chapter 6

Conclusions and Outlook

The aim of this project was the development of tools and computational methodologies for the improvement and validation of automated model building and the interpretation of MX electron density map at medium-to-low resolution. The main achievement was the development of the novel descriptor DipSpace and the associated DipScore. Complementing the Ramachandran-like plot with the DipSpace and eliminating dipeptide units and fragment links with low DipScore, proved to be a promising route to follow for the improvement of the geometry of the automatically traced models with ARP/wARP. The fact that at medium-to-lower resolution there are cases that could only be built if the DipSpace is employed is an indicator that the DipSpace is indeed a useful addition for the automated interpretation of electron density maps. In combination with improved geometrisation of candidate dipeptide units, ARP/wARP is now able to build better models than four years ago.

The DipSpace presents a new way of looking at the protein main-chain conformational space and has a potential to be used in structural biology hands-in-hands with the Ramachandran plot. Angles by themselves are independent on the scale and, therefore, one can design a dipeptide unit with perfect angular geometry but with interatomic distances that do not make sense from a chemical point of view. Therefore, tools that take into account the known distributions of bond distances are needed to better validate protein models. The DipSpace presents distance information in a unified three-dimensional Euclidean conformational space, which has a great potential for validation of protein models. The DipSpace is based on a representation of protein main-chain to a set of trans-peptide planes connected at the Cα positions that can be defined by Cα and oxygen atoms. Accounting for the coordinate error and the variation of the interatomic distances, it accounts for the natural variation observed in trans-peptide planes (Figure 22). This system allows then the faster rendering of the protein main-chain (due to the reduced number of points) but still includes all important geometrical information.

Due to their rare occurrence, cis-peptides were not considered, but their location in the DipSpace would be interesting to evaluate in the future. They would most likely not change the main DipSpace axes but given their different fixed distances they would most likely populate a different volume of the DipSpace. Additionally, as discussed in 3.3, the DipSpace does not discriminate between different amino acids. While this can certainly be considered for specific cases, such generality allows the application of DipSpace during early stages of model building, before any sequence space is considered. The DipSpace is thus a representation of all possible main-chain conformations and their frequency, independently on the residue type.

These features of the DipSpace and the uniform noise model used makes all residue types to be scored with the same weight, without taking into account the effect of the side-chains in the main-chain geometry and discarding all residues in the cis- conformation. The computation and proper weighting of the DipSpace coordinates of different residues and noise could provide a way to improve the validation of protein models already deposited, but should not be used during the automated building of protein models where the residue identity of a given Cα position is still not known and would most likely affect the identification of the correct main-chain paths.

Overall, the DipScore is a powerful measure of the dipeptide unit conformation, providing a likelihood of belonging to the same population as present in the PDB. The method can be used for the validation of a protein model obtained with any Structural Biology technique as long as main-chain Cα and carbonyl oxygen atoms are present in the model. This includes MX models, but also NMR model ensemblies, high-resolution EM models and structures obtained using *ab initio* or homology modelling. The DipCheck tool will be freely available to the scientific community via its dedicated web service. It is also offered to the PDBe as an additional validation tool. A recent collaboration with the group of Dr. Igor Barsukov from the University of Liverpool is formed to develop DipCheck to deal properly with NMR protein models and to provide information that may help the modelling and the analysis of NMR ensembles, alongside with other validation tools, e.g. Procheck-NMR [261] and MolProbity [39].

The application of the distance-geometry approach to the derivation of a protein main-chain conformational space opens doors for the development of other protein main-chain conformational spaces. One space to study could be that of the full-atom protein main chain, where instead of using the Cα and main-chain oxygen atom positions to represent dipeptide units, all main-chain atoms are taken into account, excluding then the simplified view of the peptide plane. The Euclidian space of the dipeptide units composed of 9 atoms may have properties similar to that presented in this thesis. This space could also be useful for the validation of protein models, but less likely for the automated model building of protein models. A very intriguing direction would be the construction of a conformational space for Cα-only protein models. The elaboration of the distance-geometry approach used in this thesis may allow the identification of a three-dimensional conformational space that may be of use for the validation of low-resolution EM models and the automated building of protein models in the low-resolution MX electron density maps.

Who says ‘main-chain conformational space’ can also say ‘side-chain conformational and topological space’. An approach similar to the one described in this thesis was already used for the description of small molecule conformation and implemented for the fitting and guessing of ligands in known electron density clusters [166]; in the same way it can be applied to the description of side-chain conformational space and ligand fingerprinting. A representation of side-chain density clusters as a mesh of points, placed either regularly or in the density peaks, together with the eigenspectra of their adjacency matrices have a good potential for side-chain fingerprinting. It could also make the identification of side-chain density more reliable, providing the possibility to increase the chances of the loop extension modules to extend fragmented protein models, and therefore enhance the model building procedure, particularly at lower resolution.

Bibliography

[1] D. L. Nelson and M. M. Cox, *Lehninger Principles of Biochemistry*, 6th ed. W. H. Freeman and Company, 2013.

[2] J. A. G. Ranea, A. Sillero, J. M. Thornton, and C. A. Orengo, "Protein Superfamily Evolution and the Last Universal Common Ancestor (LUCA)," *J. Mol. Evol.*, vol. 63, no. 4, pp. 513–525, Oct. 2006.

[3] B. Alberts, "The Cell as a Collection of Protein Machines: Preparing the Next Generation of Molecular Biologists," *Cell*, vol. 92, no. 3, pp. 291–294, Feb. 1998.

[4] T. E. Creighton, "Chemical Properties of Polypeptides," in *Proteins: Structures and Molecular Properties*, 2nd ed., W. H. Freeman and Company, 1993, pp. 1–47.

[5] D. L. Nelson and M. M. Cox, "Amino Acids , Peptides and Proteins," in *Lehninger Principles of Biochemistry*, 6th ed., W. H. Freeman and Company, 2013, pp. 75–150.

[6] M. T. Englander, J. L. Avins, R. C. Fleisher, B. Liu, P. R. Effraim, J. Wang, K. Schulten, T. S. Leyh, R. L. Gonzalez, and V. W. Cornish, "The ribosome can discriminate the chirality of amino acids within its peptidyl-transferase center.," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 112, no. 19, pp. 6038–43, May 2015.

[7] L. Pauling, R. B. Corey, and H. R. Branson, "The structure of proteins: two hydrogen-bonded helical configurations of the polypeptide chain.," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 37, no. 4, pp. 205–11, Apr. 1951.

[8] G. N. Ramachandran and A. K. Mitra, "An explanation for the rare occurrence of cis peptide units in proteins and polypeptides.," *J. Mol. Biol.*, vol. 107, no. 1, pp. 85–92, Oct. 1976.

[9] M. W. MacArthur and J. M. Thornton, "Deviations from planarity of the peptide bond in peptides and proteins.," *J. Mol. Biol.*, vol. 264, no. 5, pp. 1180–95, Dec. 1996.

[10] D. Pal and P. Chakrabarti, "Cis peptide bonds in proteins: residues involved, their conformations, interactions and locations.," *J. Mol. Biol.*, vol. 294, no. 1, pp. 271–88, Nov. 1999.

[11] R. B. Corey and L. Pauling, "Fundamental Dimensions of Polypeptide Chains," *Proc. R. Soc. B Biol. Sci.*, vol. 141, no. 902, pp. 10–20, Mar. 1953.

[12] L. Pauling and R. B. Corey, "The pleated sheet, a new layer configuration of polypeptide chains.," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 37, no. 5, pp. 251–6, May 1951.

[13] D. L. Nelson and M. M. Cox, "The Three-Dimensional Structure of Proteins," in *Lehninger Principles of Biochemistry*, 6th ed., W. H. Freeman and Company, 2013, pp. 115–150.

[14] B. W. Low and R. B. Baybutt, "The π helix—A hydrogen bonded configuration of the polypeptide chain," *J. Am. Chem. Soc.*, vol. 74, no. 22, pp. 5806–5807, Nov. 1952.

[15] R. B. Cooley, D. J. Arp, and P. A. Karplus, "Evolutionary origin of a secondary structure: π-helices as cryptic but widespread insertional variations of α-helices that enhance protein functionality.," *J. Mol. Biol.*, vol. 404, no. 2, pp. 232–46, Nov. 2010.

[16] J. Donohue, "Hydrogen Bonded Helical Configurations of the Polypeptide Chain.," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 39, no. 6, pp. 470–8, Jun. 1953.

[17] D. J. Barlow and J. M. Thornton, "Helix geometry in proteins," *J. Mol. Biol.*, vol. 201, no. 3, pp. 601–619, Jun. 1988.

[18] M. N. Fodje and S. Al-Karadaghi, "Occurrence, conformational features and amino acid propensities for the pi-helix.," *Protein Eng.*, vol. 15, no. 5, pp. 353–8, May 2002.

[19] L. Pauling and R. B. Corey, "Configurations of Polypeptide Chains With Favored Orientations Around Single Bonds: Two New Pleated Sheets.," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 37, no. 11, pp. 729–40, Nov. 1951.

[20] B. L. Sibanda and J. M. Thornton, "β-Hairpin families in globular proteins," *Nature*, vol. 316, no. 6024, pp. 170–174, Jul. 1985.

[21] A. C. Martin, K. Toda, H. J. Stirk, and J. M. Thornton, "Long loops in proteins.," *Protein Eng.*, vol. 8, no. 11, pp. 1093–101, Nov. 1995.

[22] a G. Murzin, S. E. Brenner, T. Hubbard, and C. Chothia, "SCOP: a structural classification of proteins database for the investigation of sequences and structures.," *J. Mol. Biol.*, vol. 247, no. 4, pp. 536–540, Apr. 1995.

[23] C. Orengo, A. Michie, S. Jones, D. Jones, M. Swindells, and J. Thornton, "CATH – a hierarchic classification of protein domain structures," *Structure*, vol. 5, no. 8, pp. 1093–1109, Aug. 1997.

[24] J. Hou, G. E. Sims, C. Zhang, and S.-H. Kim, "A global representation of the protein fold space.," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 100, pp. 2386–2390, 2003.

[25] R. Kolodny, L. Pereyaslavets, A. O. Samson, and M. Levitt, "On the universe of protein folds.,"

Annu. Rev. Biophys., vol. 42, pp. 559–82, Jan. 2013.

[26] V. Alva, M. Remmert, A. Biegert, A. N. Lupas, and J. Söding, "A galaxy of folds.," *Protein Sci.*, vol. 19, no. 1, pp. 124–30, Jan. 2010.

[27] M. Beck, M. Topf, Z. Frazier, H. Tjong, M. Xu, S. Zhang, and F. Alber, "Exploring the spatial and temporal organization of a cell's proteome.," *J. Struct. Biol.*, vol. 173, no. 3, pp. 483–96, Mar. 2011.

[28] P. E. Wright and H. J. Dyson, "Intrinsically disordered proteins in cellular signalling and regulation," *Nat. Rev. Mol. Cell Biol.*, vol. 16, no. 1, pp. 18–29, Dec. 2014.

[29] J. Habchi, P. Tompa, S. Longhi, and V. N. Uversky, "Introducing Protein Intrinsic Disorder," *Chem. Rev.*, vol. 114, no. 13, pp. 6561–6588, Jul. 2014.

[30] N. Perdigao, J. Heinrich, C. Stolte, K. S. Sabir, M. J. Buckley, B. Tabor, B. Signal, B. S. Gloss, C. J. Hammang, B. Rost, A. Schafferhans, and S. I. O'Donoghue, "Unexpected features of the dark proteome," *Proc. Natl. Acad. Sci.*, vol. 112, no. 52, pp. 15898–15903, Nov. 2015.

[31] F. H. C. Crick, "The packing of α-helices: simple coiled-coils," *Acta Crystallogr.*, vol. 6, no. 8, pp. 689–697, 1953.

[32] G. N. Ramachandran, C. Ramakrishnan, and V. Sasisekharan, "Stereochemistry of polypeptide chain configurations.," *J. Mol. Biol.*, vol. 7, pp. 95–99, 1963.

[33] G. J. Kleywegt, "Validation of protein models from Cα coordinates alone," *J. Mol. Biol.*, vol. 273, no. 2, pp. 371–376, Oct. 1997.

[34] S. C. Lovell, I. W. Davis, W. B. Arendall, P. I. W. De Bakker, J. M. Word, M. G. Prisant, J. S. Richardson, and D. C. Richardson, "Structure validation by Cα geometry: φ,ψ and Cβ deviation," *Proteins Struct. Funct. Genet.*, vol. 50, no. 3, pp. 437–450, 2003.

[35] S. A. Hollingsworth and P. A. Karplus, "A fresh look at the Ramachandran plot and the occurrence of standard structures in proteins.," *Biomol. Concepts*, vol. 1, no. 3–4, pp. 271–283, Oct. 2010.

[36] H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne, "The Protein Data Bank.," *Nucleic Acids Res.*, vol. 28, no. 1, pp. 235–242, Mar. 2000.

[37] O. Carugo and K. Djinovic-Carugo, "Half a century of Ramachandran plots.," *Acta Crystallogr. D. Biol. Crystallogr.*, vol. 69, no. Pt 8, pp. 1333–41, Aug. 2013.

[38] R. A. Laskowski, M. W. Macarthur, D. S. Moss, and J. M. Thornton, "PROCHECK: a program to check the steroechemical quality of protein structures," *J. Appl. Crystallogr.*, vol. 26, no. 2, pp. 283–291, 1993.

[39] V. B. Chen, W. B. Arendall, J. J. Headd, D. A. Keedy, R. M. Immormino, G. J. Kapral, L. W. Murray, J. S. Richardson, and D. C. Richardson, "MolProbity: all-atom structure validation for macromolecular crystallography.," *Acta Crystallogr. D. Biol. Crystallogr.*, vol. 66, no. Pt 1, pp. 12–21, Jan. 2010.

[40] R. J. Read, P. D. Adams, W. B. Arendall, A. T. Brunger, P. Emsley, R. P. Joosten, G. J. Kleywegt, E. B. Krissinel, T. Lütteke, Z. Otwinowski, A. Perrakis, J. S. Richardson, W. H. Sheffler, J. L. Smith, I. J. Tickle, G. Vriend, and P. H. Zwart, "A new generation of crystallographic validation tools for the protein data bank.," *Structure*, vol. 19, no. 10, pp. 1395–412, Oct. 2011.

[41] a L. Morris, M. W. MacArthur, E. G. Hutchinson, and J. M. Thornton, "Stereochemical quality of protein structure coordinates.," *Proteins*, vol. 12, no. 4, pp. 345–64, Apr. 1992.

[42] R. W. Woody, "Circular dichroism spectrum of peptides in the poly(Pro)II conformation.," *J. Am. Chem. Soc.*, vol. 131, no. 23, pp. 8234–45, Jun. 2009.

[43] G. Némethy and M. P. Printz, "The γ Turn, a Possible Folded Conformation of the Polypeptide Chain. Comparison with the β Turn," *Macromolecules*, vol. 5, no. 6, pp. 755–758, Nov. 1972.

[44] P. A. Karplus, "Experimentally observed conformation-dependent geometry and hidden strain in proteins.," *Protein Sci.*, vol. 5, no. 7, pp. 1406–1420, 1996.

[45] R. L. Dunbrack and M. Karplus, "Backbone-dependent rotamer library for proteins. Application to side-chain prediction.," *J. Mol. Biol.*, vol. 230, no. 2, pp. 543–74, Mar. 1993.

[46] H. Schrauber, F. Eisenhaber, and P. Argos, "Rotamers: to be or not to be? An analysis of amino acid side-chain conformations in globular proteins.," *J. Mol. Biol.*, vol. 230, no. 2, pp. 592–612, Mar. 1993.

[47] M. W. MacArthur and J. M. Thornton, "Influence of proline residues on protein conformation," *J. Mol. Biol.*, vol. 218, no. 2, pp. 397–412, Mar. 1991.

[48] A. K. Jha, A. Colubri, M. H. Zaman, S. Koide, T. R. Sosnick, and K. F. Freed, "Helix, sheet, and polyproline II frequencies and strong nearest neighbor effects in a restricted coil library.," *Biochemistry*, vol. 44, no. 28, pp. 9691–702, Jul. 2005.

[49] D. Ting, G. Wang, M. Shapovalov, R. Mitra, M. I. Jordan, and R. L. Dunbrack, "Neighbor-dependent Ramachandran probability distributions of amino acids developed from a hierarchical Dirichlet process model.," *PLoS Comput. Biol.*, vol. 6, no. 4, p. e1000763, Apr. 2010.

[50] D. S. Berkholz, M V Shapovalov, R. L. Dunbrack, and P. A. Karplus, "Conformation dependence

of backbone geometry in proteins.," *Structure*, vol. 17, no. 10, pp. 1316–25, Oct. 2009.

[51] K. S. Wilson, S. Butterworth, Z. Dauter, V. S. Lamzin, M. Walsh, S. Wodak, J. Pontius, J. Richelle, A. Vaguine, C. Sander, R. W. W. Hooft, G. Vriend, J. M. Thornton, R. A. Laskowski, M. W. MacArthur, E. J. Dodson, G. Murshudov, T. J. Oldfield, R. Kaptein, and J. A. C. Rullmann, "Who checks the checkers? four validation tools applied to eight atomic resolution structures," *J. Mol. Biol.*, vol. 276, no. 2, pp. 417–436, Feb. 1998.

[52] S. Rackovsky and H. A. Scheraga, "Differential Geometry and Polymer Conformation. 1. Comparison of Protein Conformations," *Macromolecules*, vol. 11, no. 6, pp. 1168–1174, 1978.

[53] S. Rackovsky and H. A. Scheraga, "Differential Geometry And Protein Folding," *Acc. Chem. Res*, vol. 17, no. 061/7, p. 209, 1984.

[54] M. Levitt, "A simplified representation of protein conformations for rapid simulation of protein folding," *J. Mol. Biol.*, vol. 104, no. 1, pp. 59–107, Jun. 1976.

[55] T. J. Oldfield and R. E. Hubbard, "Analysis of Cα geometry in protein structures," *Proteins Struct. Funct. Genet.*, vol. 18, no. 4, pp. 324–337, 1994.

[56] X. Peng, A. Chenani, S. Hu, Y. Zhou, and A. J. Niemi, "A three dimensional visualisation approach to protein heavy-atom structure reconstruction.," *BMC Struct. Biol.*, vol. 14, no. 1, p. 27, Jan. 2014.

[57] R. C. Penner, E. S. Andersen, J. L. Jensen, A. K. Kantcheva, M. Bublitz, P. Nissen, A. M. H. Rasmussen, K. L. Svane, B. Hammer, R. Rezazadegan, N. C. Nielsen, J. T. Nielsen, and J. E. Andersen, "Hydrogen bond rotations as a uniform structural tool for analyzing protein architecture," *Nat. Commun.*, vol. 5, p. 5803, Dec. 2014.

[58] I. D. Campbell, "The march of structural biology," *Mol. Cell Biol.*, vol. 3, no. May, 2002.

[59] E. F. Garman, "Developments in x-ray crystallographic structure determination of biological macromolecules.," *Science*, vol. 343, no. 6175, pp. 1102–8, Mar. 2014.

[60] M. G. Strauss, E. M. Westbrook, I. Naday, T. A. Coleman, M. L. Westbrook, D. J. Travis, R. M. Sweet, J. W. Pflugrath, and M. Stanton, "CCD-based detector for protein crystallography with synchrotron X-rays," *Nucl. Instruments Methods Phys. Res. Sect. A Accel. Spectrometers, Detect. Assoc. Equip.*, vol. 297, no. 1–2, pp. 275–295, Nov. 1990.

[61] C. Broennimann, E. F. Eikenberry, B. Henrich, R. Horisberger, G. Huelsen, E. Pohl, B. Schmitt, C. Schulze-Briese, M. Suzuki, T. Tomizaki, H. Toyokawa, and A. Wagner, "The PILATUS 1M detector.," *J. Synchrotron Radiat.*, vol. 13, no. Pt 2, pp. 120–30, Mar. 2006.

[62] A. Wlodawer, W. Minor, Z. Dauter, and M. Jaskolski, "Protein crystallography for non-crystallographers, or how to get the best (but not more) from published macromolecular structures," *FEBS J.*, vol. 275, no. 1, pp. 1–21, 2008.

[63] W. H. Bragg and W. L. Bragg, "The Reflection of X-rays by Crystals," *Proc. R. Soc. A Math. Phys. Eng. Sci.*, vol. 88, no. 605, pp. 428–438, Jul. 1913.

[64] L. Ooi, *Principles of X-ray Crystallography*, 1st ed. Oxford: Oxford University Press, 2010.

[65] G. Taylor, "The phase problem," *Acta Crystallogr. Sect. D Biol. Crystallogr.*, vol. 59, no. 11, pp. 1881–1890, Oct. 2003.

[66] R. E. Stenkamp and L. H. Jensen, "Resolution revisited: limit of detail in electron density maps," *Acta Crystallogr. Sect. A Found. Crystallogr.*, vol. 40, no. 3, pp. 251–254, May 1984.

[67] S. M. Swanson, "Effective resolution of macromolecular X-ray diffraction data," *Acta Crystallogr. Sect. A Found. Crystallogr.*, vol. 44, no. 4, pp. 437–442, Jul. 1988.

[68] J. C. Kendrew, G. Bodo, H. M. Dintzis, R. G. Parrish, and H. Wyckoff, "A three-dimensional model of the myoglobin molecule obtained by x-ray analysis," *Nature*, vol. 181, no. 4610, pp. 662–666, 1958.

[69] G. Bodo, H. M. Dintzis, J. C. Kendrew, and H. W. Wyckoff, "The Crystal Structure of Myoglobin. V. A Low-Resolution Three-Dimensional Fourier Synthesis of Sperm-Whale Myoglobin Crystals," *Proc. R. Soc. A Math. Phys. Eng. Sci.*, vol. 253, no. 1272, pp. 70–102, 1959.

[70] A. Wlodawer, W. Minor, Z. Dauter, and M. Jaskolski, "Protein crystallography for aspiring crystallographers or how to avoid pitfalls and traps in macromolecular structure determination," *FEBS J.*, vol. 280, no. 22, pp. 5705–5736, Nov. 2013.

[71] G. L. Taylor, "Introduction to phasing.," *Acta Crystallogr. D. Biol. Crystallogr.*, vol. 66, no. Pt 4, pp. 325–38, Apr. 2010.

[72] D. M. Blow and M. G. Rossmann, "The single isomorphous replacement method," *Acta Crystallogr.*, vol. 14, no. 11, pp. 1195–1202, Nov. 1961.

[73] M. G. Rossmann, "The position of anomalous scatterers in protein crystals," *Acta Crystallogr.*, vol. 14, no. 4, pp. 383–388, Apr. 1961.

[74] W. Hendrickson, "Determination of macromolecular structures from anomalous diffraction of synchrotron radiation," *Science (80-.)*., vol. 254, no. 5028, pp. 51–58, Oct. 1991.

[75] M. G. Rossmann and D. M. Blow, "The detection of sub-units within the crystallographic

asymmetric unit," *Acta Crystallogr.*, vol. 15, no. 1, pp. 24–31, Jan. 1962.

[76] G. Scapin, "Molecular replacement then and now.," *Acta Crystallogr. D. Biol. Crystallogr.*, vol. 69, no. Pt 11, pp. 2266–75, Nov. 2013.

[77] J. Navaza, "AMoRe: an automated package for molecular replacement," *Acta Crystallogr. Sect. A Found. Crystallogr.*, vol. 50, no. 2, pp. 157–163, Mar. 1994.

[78] A. Vagin and A. Teplyakov, "MOLREP : an Automated Program for Molecular Replacement," *J. Appl. Crystallogr.*, vol. 30, no. 6, pp. 1022–1025, Dec. 1997.

[79] A. J. McCoy, R. W. Grosse-Kunstleve, P. D. Adams, M. D. Winn, L. C. Storoni, and R. J. Read, "Phaser crystallographic software.," *J. Appl. Crystallogr.*, vol. 40, no. Pt 4, pp. 658–674, Aug. 2007.

[80] R. M. Keegan and M. D. Winn, "MrBUMP: an automated pipeline for molecular replacement.," *Acta Crystallogr. D. Biol. Crystallogr.*, vol. 64, no. Pt 1, pp. 119–24, Jan. 2008.

[81] F. Long, A. A. Vagin, P. Young, and G. N. Murshudov, "BALBES: a molecular-replacement pipeline.," *Acta Crystallogr. D. Biol. Crystallogr.*, vol. 64, no. Pt 1, pp. 125–32, Jan. 2008.

[82] J. Bibby, R. M. Keegan, O. Mayans, M. D. Winn, and D. J. Rigden, "AMPLE: a cluster-and-truncate approach to solve the crystal structures of small proteins using rapidly computed ab initio models.," *Acta Crystallogr. D. Biol. Crystallogr.*, vol. 68, no. Pt 12, pp. 1622–31, Dec. 2012.

[83] D. D. Rodríguez, C. Grosse, S. Himmel, C. González, I. M. De Ilarduya, S. Becker, G. M. Sheldrick, and I. Usón, "Crystallographic ab initio protein structure solution below atomic resolution.," *Nat. Methods*, vol. 6, no. 9, pp. 651–653, Sep. 2009.

[84] D. Rodríguez, M. Sammito, K. Meindl, I. M. de Ilarduya, M. Potratz, G. M. Sheldrick, and I. Usón, "Practical structure solution with ARCIMBOLDO.," *Acta Crystallogr. D. Biol. Crystallogr.*, vol. 68, no. 4, pp. 336–43, Apr. 2012.

[85] K. Cowtan and P. Main, "Miscellaneous Algorithms for Density Modification," *Acta Crystallogr. Sect. D Biol. Crystallogr.*, vol. 54, no. 4, pp. 487–493, Jul. 1998.

[86] T. C. Terwilliger, "Statistical density modification with non-crystallographic symmetry," *Acta Crystallogr. Sect. D Biol. Crystallogr.*, vol. 58, no. Pt 12, pp. 2082–2086, Nov. 2002.

[87] T. C. Terwilliger, R. W. Grosse-Kunstleve, P. V Afonine, N. W. Moriarty, P. H. Zwart, L.-W. Hung, R. J. Read, and P. D. Adams, "Iterative model building, structure refinement and density modification with the PHENIX AutoBuild wizard," *Acta Crystallogr. Sect. D Biol. Crystallogr.*, vol. 64, no. 1, pp. 61–69, Jan. 2008.

[88] R. M. Esnouf, "Polyalanine reconstruction from Calpha positions using the program CALPHA can aid initial phasing of data by molecular replacement procedures.," *Acta Crystallogr. D. Biol. Crystallogr.*, vol. 53, no. Pt 6, pp. 665–72, Nov. 1997.

[89] M. F. Perutz, M. G. Rossmann, A. F. Cullis, H. Muirhead, W. Georg, and A. C. T. North, "Structure of Hæmoglobin: A Three-Dimensional Fourier Synthesis at 5.5-Å Resolution, Obtained by X-Ray Analysis," *Nature*, vol. 185, no. 4711, pp. 416–422, Feb. 1960.

[90] J. C. Kendrew, R. E. Dickerson, B. E. Strandberg, R. G. Hart, D. R. Davies, D. C. Phillips, and V. C. Shore, "Structure of Myoglobin: A Three-Dimensional Fourier Synthesis at 2 Å Resolution," *Nature*, vol. 185, no. 4711, pp. 422–427, Feb. 1960.

[91] M. F. Perutz, H. Muirhead, J. M. Cox, and L. C. Goaman, "Three-dimensional Fourier synthesis of horse oxyhaemoglobin at 2.8 A resolution: the atomic model.," *Nature*, vol. 219, no. 5150, pp. 131–9, Jul. 1968.

[92] F. M. Richards, "The matching of physical models to three-dimensional electron-density maps: A simple optical device," *J. Mol. Biol.*, vol. 37, no. 1, pp. 225–230, Oct. 1968.

[93] T. A. Jones, "A graphics model building and refinement system for macromolecules," *J. Appl. Crystallogr.*, vol. 11, no. 4, pp. 268–272, Aug. 1978.

[94] T. A. Jones, J. Y. Zou, S. W. Cowan, and M. Kjeldgaard, "Improved methods for building protein models in electron density maps and the location of errors in these models," *Acta Crystallogr. Sect. A Found. Crystallogr.*, vol. 47, no. 2, pp. 110–119, Mar. 1991.

[95] D. E. McRee, "XtalView/Xfit--A versatile program for manipulating atomic coordinates and electron density.," *J. Struct. Biol.*, vol. 125, no. 2–3, pp. 156–65, Jan. 1999.

[96] P. Emsley and K. Cowtan, "Coot: model-building tools for molecular graphics.," *Acta Crystallogr. D. Biol. Crystallogr.*, vol. 60, no. Pt 12 Pt 1, pp. 2126–32, Dec. 2004.

[97] P. Emsley, B. Lohkamp, W. G. Scott, and K. Cowtan, "Features and development of Coot.," *Acta Crystallogr. D. Biol. Crystallogr.*, vol. 66, no. Pt 4, pp. 486–501, Apr. 2010.

[98] J. Greer, "Three-dimensional pattern recognition: An approach to automated interpretation of electron density maps of proteins," *J. Mol. Biol.*, vol. 82, no. 3, pp. 279–301, Jan. 1974.

[99] T. Oldfield, "Pattern-recognition methods to identify secondary structure within X-ray crystallographic electron-density maps," *Acta Crystallogr. Sect. D Biol. Crystallogr.*, vol. 58, no. 3,

pp. 487–493, Feb. 2002.

[100] T. J. Oldfield, "Automated tracing of electron-density maps of proteins," *Acta Crystallogr. Sect. D Biol. Crystallogr.*, vol. 59, no. 3, pp. 483–491, Feb. 2003.

[101] T. Holton, T. R. Ioerger, J. A. Christopher, and J. C. Sacchettini, "Determining protein structure from electron-density maps using pattern matching," *Acta Crystallogr. Sect. D Biol. Crystallogr.*, vol. 56, no. 6, pp. 722–734, Jun. 2000.

[102] T. R. Ioerger and J. C. Sacchettini, "Automatic modeling of protein backbones in electron-density maps via prediction of Calpha coordinates.," *Acta Crystallogr. Sect. D Biol. Crystallogr.*, vol. 58, no. Pt 12, pp. 2043–2054, Nov. 2002.

[103] T. R. Ioerger and J. C. Sacchettini, "TEXTAL system: artificial intelligence techniques for automated protein model building.," *Methods Enzymol.*, vol. 374, no. 1998, pp. 244–70, Jan. 2003.

[104] T. Romo, K. Gopal, E. McKee, L. Kanbi, J. Smith, J. Sacchettini, and T. Ioerger, "TEXTAL: AI-Based Structural Determination for X-ray Protein Crystallography," *IEEE Intell. Syst.*, vol. 20, no. 6, pp. 59–63, Nov. 2005.

[105] T. D. Romo, J. C. Sacchettini, and T. R. Ioerger, "Improving amino-acid identification, fit and C(alpha) prediction using the Simplex method in automated model building.," *Acta Crystallogr. D. Biol. Crystallogr.*, vol. 62, no. Pt 11, pp. 1401–6, Nov. 2006.

[106] G. J. Kleywegt and T. A. Jones, "Template convolution to enhance or detect structural features in macromolecular electron-density maps," *Acta Crystallogr. Sect. D Biol. Crystallogr.*, vol. 53, no. 2, pp. 179–185, 1997.

[107] N. W. Isaacs and R. C. Agarwal, "Free atom insertion and refinement as a means of extending and refining phases," *Methods Enzymol.*, vol. 115, pp. 112–117, 1985.

[108] K. D. Cowtan, "Fast Fourier feature recognition," *Acta Crystallogr. Sect. D Biol. Crystallogr.*, vol. 57, no. Pt 10, pp. 1435–1444, Sep. 2001.

[109] K. Cowtan, "The Buccaneer software for automated model building. 1. Tracing protein chains," *Acta Crystallogr. Sect. D Biol. Crystallogr.*, vol. 62, no. Pt 9, pp. 1002–1011, Sep. 2006.

[110] T. C. Terwilliger, "Automated main-chain model building by template matching and iterative fragment extension," *Acta Crystallogr. Sect. D Biol. Crystallogr.*, vol. 59, no. 1, pp. 38–44, Dec. 2002.

[111] T. C. Terwilliger, "Automated side-chain model building and sequence assignment by template matching," *Acta Crystallogr. Sect. D Biol. Crystallogr.*, vol. 59, no. 1, pp. 45–49, Dec. 2002.

[112] T. C. Terwilliger, "SOLVE and RESOLVE: automated structure solution and density modification.," *Methods Enzymol.*, vol. 374, pp. 22–37, Jan. 2003.

[113] P. H. Zwart, P V Afonine, R. W. Grosse-Kunstleve, L.-W. Hung, T. R. Ioerger, A. J. McCoy, E. McKee, N. W. Moriarty, R. J. Read, J. C. Sacchettini, N. K. Sauter, L. C. Storoni, T. C. Terwilliger, and P. D. Adams, "Automated structure solution with the PHENIX suite.," *Methods Mol. Biol.*, vol. 426, pp. 419–35, Jan. 2008.

[114] P. D. Adams, V. Pavel, V. B. Chen, W. Ian, N. Echols, N. W. Moriarty, R. J. Read, D. C. Richardson, S. Jane, and C. Thomas, "PHENIX : a comprehensive Python-based system for macromolecular structure solution," *Acta Crystallogr. Sect. D Biol. Crystallogr.*, pp. 213–221, 2010.

[115] V. S. Lamzin and K. S. Wilson, "Automated refinement of protein models.," *Acta Crystallogr. Sect. D Biol. Crystallogr.*, vol. 49, no. Pt 1, pp. 129–147, 1993.

[116] A. Perrakis, T. K. Sixma, K. S. Wilson, and V. S. Lamzin, "wARP: improvement and extension of crystallographic phases by weighted averaging of multiple-refined dummy atomic models.," *Acta Crystallogr. D. Biol. Crystallogr.*, vol. 53, no. Pt 4, pp. 448–55, Jul. 1997.

[117] A. Perrakis, R. Morris, and V. S. Lamzin, "Automated protein model building combined with iterative structure refinement.," *Nat. Struct. Biol.*, vol. 6, no. 5, pp. 458–63, May 1999.

[118] V. S. Lamzin and K. S. Wilson, "Automated refinement for protein crystallography.," *Methods Enzymol.*, vol. 277, no. 1990, pp. 269–305, Jan. 1997.

[119] F. DiMaio, J. Shavlik, and G. N. Phillips, "A probabilistic approach to protein backbone tracing in electron density maps.," *Bioinformatics*, vol. 22, no. 14, pp. e81-9, Jul. 2006.

[120] F. DiMaio, D. a Kondrashov, E. Bitto, A. Soni, C. a Bingman, G. N. Phillips, and J. W. Shavlik, "Creating protein models from electron-density maps using particle-filtering methods.," *Bioinformatics*, vol. 23, no. 21, pp. 2851–2858, Nov. 2007.

[121] I. J. Tickle, "Statistical quality indicators for electron-density maps.," *Acta Crystallogr. D. Biol. Crystallogr.*, vol. 68, no. Pt 4, pp. 454–67, Apr. 2012.

[122] G. N. Murshudov, P. Skubák, A. a Lebedev, N. S. Pannu, R. a Steiner, R. a Nicholls, M. D. Winn, F. Long, and A. a Vagin, "REFMAC5 for the refinement of macromolecular crystal structures.," *Acta Crystallogr. D. Biol. Crystallogr.*, vol. 67, no. Pt 4, pp. 355–67, Apr. 2011.

[123] R. A. Engh and R. Huber, “Structure quality and target parameters,” *Int. Tables Crystallogr. Vol F*, vol. F, pp. 382–392, 2006.

[124] P. V Afonine, R. W. Grosse-Kunstleve, N. Echols, J. J. Headd, N. W. Moriarty, M. Mustyakimov, T. C. Terwilliger, A. Urzhumtsev, P. H. Zwart, and P. D. Adams, “Towards automated crystallographic structure refinement with phenix.refine.,” *Acta Crystallogr. D. Biol. Crystallogr.*, vol. 68, no. Pt 4, pp. 352–67, Apr. 2012.

[125] A. Brünger, “Free R value: a novel statistical quantity for assessing the accuracy of crystal structures,” *Nature*, vol. 355, pp. 472–475, 1992.

[126] A. Brünger, “Free R value: Cross-validation in crystallography,” *Methods Enzymol.*, vol. 277, no. 1990, pp. 366–396, 1997.

[127] K. D. Watenpaugh, L. C. Sieker, J. R. Herriott, and L. H. Jensen, “Refinement of the model of a protein: rubredoxin at 1.5 Å resolution,” *Acta Crystallogr. Sect. B Struct. Crystallogr. Cryst. Chem.*, vol. 29, no. 5, pp. 943–956, May 1973.

[128] J. Deisenhofer and W. Steigemann, “Crystallographic refinement of the structure of bovine pancreatic trypsin inhibitor at l.5 Å resolution,” *Acta Crystallogr. Sect. B Struct. Crystallogr. Cryst. Chem.*, vol. 31, no. 1, pp. 238–250, Jan. 1975.

[129] R. Diamond, “A real-space refinement procedure for proteins,” *Acta Crystallogr. Sect. A*, vol. 27, no. 5, pp. 436–452, Sep. 1971.

[130] D. Sayre, “On least-squares refinement of the phases of crystallographic structure factors,” *Acta Crystallogr. Sect. A*, vol. 28, no. 2, pp. 210–212, Mar. 1972.

[131] G. N. Murshudov, a a Vagin, and E. J. Dodson, “Refinement of macromolecular structures by the maximum-likelihood method.,” *Acta Crystallogr. Sect. D Biol. Crystallogr.*, vol. 53, no. Pt 3, pp. 240–255, May 1997.

[132] T. C. Terwilliger, R. W. Grosse-Kunstleve, P. V Afonine, N. W. Moriarty, P. H. Zwart, L. W. Hung, R. J. Read, and P. D. Adams, “Iterative model building, structure refinement and density modification with the PHENIX AutoBuild wizard.,” *Acta Crystallogr. D. Biol. Crystallogr.*, vol. 64, no. Pt 1, pp. 61–9, Jan. 2008.

[133] G. Langer, S. X. Cohen, V. S. Lamzin, and A. Perrakis, “Automated macromolecular model building for X-ray crystallography using ARP/wARP version 7.,” *Nat. Protoc.*, vol. 3, no. 7, pp. 1171–9, Jan. 2008.

[134] J. É. Debreczeni and P. Emsley, “Handling ligands with Coot.,” *Acta Crystallogr. D. Biol. Crystallogr.*, vol. 68, no. Pt 4, pp. 425–30, Apr. 2012.

[135] J. Agirre, J. Iglesias-Fernández, C. Rovira, G. J. Davies, K. S. Wilson, and K. D. Cowtan, “Privateer: software for the conformational validation of carbohydrate structures.,” *Nat. Struct. Mol. Biol.*, vol. 22, no. 11, pp. 833–4, Nov. 2015.

[136] P. H. Zwart, G. G. Langer, and V. S. Lamzin, “Modelling bound ligands in protein crystal structures.,” *Acta Crystallogr. D. Biol. Crystallogr.*, vol. 60, no. Pt 12 Pt 1, pp. 2230–9, Dec. 2004.

[137] G. G. Langer, G. X. Evrard, C. G. Carolan, and V. S. Lamzin, “Fragmentation-tree density representation for crystallographic modelling of bound ligands,” *J. Mol. Biol.*, vol. 419, no. 3–4, pp. 211–22, Jun. 2012.

[138] T. C. Terwilliger, H. Klei, P. D. Adams, N. W. Moriarty, and J. D. Cohn, “Automated ligand fitting by core-fragment fitting and extension into density.,” *Acta Crystallogr. D. Biol. Crystallogr.*, vol. 62, no. Pt 8, pp. 915–22, Aug. 2006.

[139] T. C. Terwilliger, P. D. Adams, N. W. Moriarty, and J. D. Cohn, “Ligand identification using electron-density map correlations.,” *Acta Crystallogr. D. Biol. Crystallogr.*, vol. 63, no. Pt 1, pp. 101–7, Jan. 2007.

[140] G. J. Kleywegt and T. A. Jones, “Efficient rebuilding of protein structures.,” *Acta Crystallogr. D. Biol. Crystallogr.*, vol. 52, no. Pt 4, pp. 829–32, Jul. 1996.

[141] A. M. Davis, S. J. Teague, and G. J. Kleywegt, “Application and limitations of X-ray crystallographic data in structure-based ligand and drug design.,” *Angew. Chem. Int. Ed. Engl.*, vol. 42, no. 24, pp. 2718–36, Jun. 2003.

[142] C.-I. Bränden and T. Alwyn Jones, “Between objectivity and subjectivity,” *Nature*, vol. 343, no. 6260, pp. 687–689, 1990.

[143] W. G. Touw, R. P. Joosten, and G. Vriend, “New biological insights from better structure models.,” *J. Mol. Biol.*, vol. 428, no. 6, pp. 1375–1393, Feb. 2016.

[144] F. H. Allen, “The Cambridge Structural Database: A quarter of a million crystal structures and rising,” *Acta Crystallogr. Sect. B Struct. Sci.*, vol. 58, no. 3 PART 1, pp. 380–388, 2002.

[145] R. W. Hooft, G. Vriend, C. Sander, and E. E. Abola, “Errors in protein structures.,” *Nature*, vol. 381, no. 6580, p. 272, May 1996.

[146] I. W. Davis, A. Leaver-Fay, V. B. Chen, J. N. Block, G. J. Kapral, X. Wang, L. W. Murray, W. B.

Arendall, J. Snoeyink, J. S. Richardson, and D. C. Richardson, “MolProbity: All-atom contacts and structure validation for proteins and nucleic acids,” *Nucleic Acids Res.*, vol. 35, no. SUPPL.2, pp. 375–383, 2007.

[147] L. Urzhumtseva, P. V Afonine, P. D. Adams, and A. Urzhumtsev, “Crystallographic model quality at a glance.,” *Acta Crystallogr. D. Biol. Crystallogr.*, vol. 65, no. Pt 3, pp. 297–300, Mar. 2009.

[148] G. Vriend, “WHAT IF: A molecular modeling and drug design program,” *J. Mol. Graph.*, vol. 8, no. 1, pp. 52–56, Mar. 1990.

[149] R. W. Hooft, C. Sander, and G. Vriend, “Objectively judging the quality of a protein structure from a Ramachandran plot.,” *Comput. Appl. Biosci.*, vol. 13, no. 4, pp. 425–430, 1997.

[150] R. P. Joosten and G. Vriend, “PDB improvement starts with data deposition.,” *Science*, vol. 317, no. 5835, pp. 195–6, Jul. 2007.

[151] R. P. Joosten, T. Womack, G. Vriend, and G. Bricogne, “Re-refinement from deposited X-ray data can deliver improved models for most PDB entries.,” *Acta Crystallogr. D. Biol. Crystallogr.*, vol. 65, no. Pt 2, pp. 176–85, Feb. 2009.

[152] J. W. M. Nissink, C. Murray, M. Hartshorn, M. L. Verdonk, J. C. Cole, and R. Taylor, “A new test set for validating predictions of protein-ligand interaction.,” *Proteins*, vol. 49, no. 4, pp. 457–71, Dec. 2002.

[153] G. J. Kleywegt, “Crystallographic refinement of ligand complexes.,” *Acta Crystallogr. D. Biol. Crystallogr.*, vol. 63, no. Pt 1, pp. 94–100, Jan. 2007.

[154] J. Liebeschuetz, J. Hennemann, T. Olsson, and C. R. Groom, “The good, the bad and the twisted: a survey of ligand geometry in protein crystal structures.,” *J. Comput. Aided. Mol. Des.*, vol. 26, no. 2, pp. 169–83, Feb. 2012.

[155] I. J. Bruno, J. C. Cole, M. Kessler, J. Luo, W. D. S. Motherwell, L. H. Purkis, B. R. Smith, R. Taylor, R. I. Cooper, S. E. Harris, and A. G. Orpen, “Retrieval of crystallographically-derived molecular geometry information.,” *J. Chem. Inf. Comput. Sci.*, vol. 44, no. 6, pp. 2133–44, Jan. 2004.

[156] A. W. Schüttelkopf and D. M. F. van Aalten, “PRODRG: a tool for high-throughput crystallography of protein-ligand complexes.,” *Acta Crystallogr. D. Biol. Crystallogr.*, vol. 60, no. Pt 8, pp. 1355–63, Aug. 2004.

[157] G. J. Kleywegt and M. R. Harris, “ValLigURL: a server for ligand-structure comparison and validation.,” *Acta Crystallogr. D. Biol. Crystallogr.*, vol. 63, no. Pt 8, pp. 935–8, Aug. 2007.

[158] A. Perrakis, M. Harkiolaki, K. S. Wilson, and V. S. Lamzin, “ARP/wARP and molecular replacement,” *Acta Crystallogr. Sect. D*, vol. 57, no. 10, pp. 1445–1450, Sep. 2001.

[159] R. J. Morris, A. Perrakis, and V. S. Lamzin, “ARP / wARP ’s model-building algorithms. I. The main chain,” *Acta Crystallogr. Sect. D Biol. Crystallogr.*, vol. 58, no. 6, pp. 968–975, May 2002.

[160] R. J. Morris, A. Perrakis, and V. S. Lamzin, “ARP / wARP and Automatic Interpretation of Protein Electron Density Maps,” *Methods Enzymol.*, vol. 374, pp. 229–244, 2003.

[161] R. J. Morris, P. H. Zwart, S. Cohen, F. J. Fernandez, M. Kakaris, O. Kirillova, C. Vonrhein, A. Perrakis, and V. S. Lamzin, “Breaking good resolutions with ARP/wARP,” *J. Synchrotron Radiat.*, vol. 11, no. 1, pp. 56–59, Nov. 2003.

[162] S. X. Cohen, R. J. Morris, F. J. Fernandez, M. Ben Jelloul, M. Kakaris, V. Parthasarathy, V. S. Lamzin, G. J. Kleywegt, and A. Perrakis, “Towards complete validated models in the next generation of ARP/wARP.,” *Acta Crystallogr. D. Biol. Crystallogr.*, vol. 60, no. Pt 12 Pt 1, pp. 2222–9, Dec. 2004.

[163] S. X. Cohen, M. Ben Jelloul, F. Long, A. Vagin, P. Knipscheer, J. Lebbink, T. K. Sixma, V. S. Lamzin, G. N. Murshudov, and A. Perrakis, “ARP/wARP and molecular replacement: the next generation,” *Acta Crystallogr. Sect. D Biol. Crystallogr.*, vol. 64, no. Pt 1, pp. 49–60, Jan. 2008.

[164] J. Hattne and V. S. Lamzin, “Pattern-recognition-based detection of planar objects in three-dimensional electron-density maps.,” *Acta Crystallogr. Sect. D Biol. Crystallogr.*, vol. D64, no. Pt 8, pp. 834–842, Aug. 2008.

[165] G. X. Evrard, G. G. Langer, A. Perrakis, and V. S. Lamzin, “Assessment of automatic ligand building in ARP/wARP.,” *Acta Crystallogr. D. Biol. Crystallogr.*, vol. 63, no. Pt 1, pp. 108–17, Jan. 2007.

[166] C. G. Carolan and V. S. Lamzin, “Automated identification of crystallographic ligands using sparse-density representations.,” *Acta Crystallogr. D. Biol. Crystallogr.*, vol. 70, no. Pt 7, pp. 1844–53, Jul. 2014.

[167] J. Y. Zou and T. A. Jones, “Towards the automatic interpretation of macromolecular electron-density maps: qualitative and quantitative matching of protein sequence to map.,” *Acta Crystallogr. D. Biol. Crystallogr.*, vol. 52, no. Pt 4, pp. 833–41, Jul. 1996.

[168] K. Joosten, S. X. Cohen, P. Emsley, W. Mooij, V. S. Lamzin, and A. Perrakis, “A knowledge-

driven approach for crystallographic protein model completion," *Acta Crystallogr. Sect. D Biol. Crystallogr.*, vol. 64, no. Pt 4, pp. 416–424, Apr. 2008.

[169] T. Wiegels and V. S. Lamzin, "Use of noncrystallographic symmetry for automated model building at medium to low resolution.," *Acta Crystallogr. D. Biol. Crystallogr.*, vol. 68, no. Pt 4, pp. 446–53, Apr. 2012.

[170] M. Jaskolski, "From atomic resolution to molecular giants: An overview of crystallographic studies of biological macromolecules with synchrotron radiation," *Acta Phys. Pol. A*, vol. 117, no. 2, pp. 257–263, 2010.

[171] F. Dyda, "Developments in low-resolution biological X-ray crystallography," *F1000 Biol. Rep.*, vol. 2, no. November, p. 80, Jan. 2010.

[172] A. M. Karmali, T. L. Blundell, and N. Furnham, "Model-building strategies for low-resolution X-ray crystallographic data," *Acta Crystallogr. Sect. D Biol. Crystallogr.*, vol. 65, no. Pt 2, pp. 121–127, Feb. 2009.

[173] K. Cowtan, "Completion of autobuilt protein models using a database of protein fragments," *Acta Crystallogr. Sect. D Biol. Crystallogr.*, vol. 68, no. 4, pp. 328–335, Apr. 2012.

[174] T. Wiegels, "Better Models in Macromolecular Crystal Structure Determination," Universitat Hamburg (PhD Thesis), 2012.

[175] J. J. Sylvester, "Chemistry and Algebra," *Nature*, vol. 17, no. 432. pp. 277–296, 1878.

[176] H. Anton and C. Rorres, "Systems of Linear Equations and Matrices," in *Elementary Linear Algebra: Applications Version*, 10th ed., John Wiley & Sons, 2010, pp. 1–92.

[177] E. D. Nering, "Determinants, Eigenvalues, and Similarity Transformations," in *Linear Algebra and Matrix Theory*, 2nd ed., John Wiley & Sons, 1969, pp. 85–127.

[178] H. Anton and C. Rorres, "Eigenvalues and Eigenvectors," in *Elementary Linear Algebra: Applications Version*, 10th ed., John Wiley & Sons, 2010, pp. 295–334.

[179] I. S. Gradshteyn and I. M. Ryzhik, "Norms," in *Table of Integrals, Series, and Products*, 6th ed., A. Jeffrey and D. Zwillinger, Eds. Academic Press Ltd, 2000, pp. 1071–1081.

[180] R. Bronson and G. B. Costa, "Eigenvalues, Eigenvalues and Differential Equations," in *Linear Algebra: An Introduction*, 2nd ed., Academic Press Ltd, 2007, p. 219.

[181] H. Anton and C. Rorres, "General Vector Space," in *Elementary Linear Algebra: Applications Version*, 10th ed., John Wiley & Sons, 2010, pp. 171–294.

[182] F. Chatelin, "Supplements from Linear Algebra," in *Eigenvalues of Matrices: Revised Edition*, Revised., SIAM, 2012, pp. 1–43.

[183] K. Pearson, "On lines and planes of closest fit to systems of points in space," *Philos. Mag.*, vol. 2, pp. 559–572, 1901.

[184] H. Hotelling, "Analysis of a complex of statistical variables into principal components.," *J. Educ. Psychol.*, vol. 24, no. 6, pp. 417–441, 1933.

[185] S. Wold, K. Esbensen, and P. Geladi, "Principal component analysis," *Chemom. Intell. Lab. Syst.*, vol. 2, no. 1–3, pp. 37–52, 1987.

[186] R. J. Morris, "Statistical pattern recognition for macromolecular crystallographers.," *Acta Crystallogr. D. Biol. Crystallogr.*, vol. 60, no. Pt 12 Pt 1, pp. 2133–43, Dec. 2004.

[187] S. A. M. Stein, A. E. Loccisano, S. M. Firestine, and J. D. Evanseck, "Principal Components Analysis: A Review of its Application on Molecular Dynamics Data," *Annu. Rep. Comput. Chem.*, vol. 2, pp. 233–248, 2006.

[188] H.-G. Elias, "Conformation," in *Macromolecules: Structure and Properties, Volume 1*, Springer, 1977, pp. 93–152.

[189] G. M. Crippen and T. F. Havel, *Distance geometry and molecular conformation*, vol. 74. Research Studies Press Somerset, England, 1988.

[190] A. Kloczkowski, R. L. Jernigan, Z. Wu, G. Song, L. Yang, A. Kolinski, and P. Pokarowski, "Distance matrix-based approach to protein structure prediction," *J. Struct. Funct. Genomics*, vol. 10, no. 1, pp. 67–81, 2009.

[191] S. L. Dixon, "Pharmacophore methods," in *Drug Design*, K. M. Merz, D. Ringe, and C. H. Reynolds, Eds. Cambridge: Cambridge University Press, 2010, pp. 137–150.

[192] T. F. Havel, "Distance Geometry : Theory, Algorithms, and Chemical Applications," in *Encyclopedia of Computational Chemistry*, 2002, pp. 723–743.

[193] I. Dokmanic, R. Parhizkar, J. Ranieri, and M. Vetterli, "Euclidean Distance Matrices: Essential Theory, Algorithms and Applications," *IEEE Signal Process. Mag.*, vol. 32, no. 6, p. 17, Feb. 2015.

[194] M. Marcus and T. R. Smith, "A note on the determinants and eigenvalues of distance matrices," *Linear Multilinear Algebr.*, vol. 25, no. 3, pp. 219–230, Oct. 1989.

[195] J. Dattorro, "Euclidean Distance Matrix," in *Convex Optimization & Euclidean Distance Geometry*, Meboo Publishing, 2006, pp. 357–438.

[196] R. S. Cahn, C. Ingold, and V. Prelog, “Spezifikation der molekularen Chiralität,” *Angew. Chemie*, vol. 78, no. 8, pp. 413–447, Apr. 1966.

[197] H. D. Flack, “Louis Pasteur’s discovery of molecular chirality and spontaneous resolution in 1848, together with a complete review of his crystallographic and chemical work.,” *Acta Crystallogr. A.*, vol. 65, no. Pt 5, pp. 371–89, Sep. 2009.

[198] J. Hattne and V. S. Lamzin, “A moment invariant for evaluating the chirality of three-dimensional objects.,” *J. R. Soc. Interface*, vol. 8, no. July 2010, pp. 144–151, Jan. 2011.

[199] E. Prince, L. W. Finger, and J. H. Konnert, “Chapter 8.3. Constraints and restraints in refinement,” in *International Tables for Crystallography*, vol. C, E. Prince, Ed. Chester, England: International Union of Crystallography, 2006, pp. 694–701.

[200] A. R. Leach, “A Survey of Methods for Searching the Conformational Space of Small and Medium-Sized Molecules,” in *Reviews in Computational Chemistry, Volume 2*, B. L. Kenny and D. B. Boyd, Eds. John Wiley & Sons, 1991, pp. 1–47.

[201] A. B. Buda and K. Mislow, “A Hausdorff chirality measure,” *J. Am. Chem. Soc.*, vol. 114, no. 15, pp. 6006–6012, Jul. 1992.

[202] M. A. Osipov, B. T. Pickup, and D. A. Dunmur, “A new twist to molecular chirality: intrinsic chirality indices,” *Mol. Phys.*, vol. 84, no. May 2016, pp. 1193–1206, 1995.

[203] A. Pietropaolo, L. Muccioli, R. Berardi, and C. Zannoni, “A chirality index for investigating protein secondary structures and their time evolution.,” *Proteins*, vol. 70, no. 3, pp. 667–77, Feb. 2008.

[204] M. Solymosi, R. J. Low, M. Grayson, and M. P. Neal, “A generalized scaling of a chiral index for molecules,” *J. Chem. Phys.*, vol. 116, no. 22, p. 9875, May 2002.

[205] A. T. Balaban, “Applications of graph theory in chemistry,” *J. Chem. Inf. Model.*, vol. 25, no. 3, pp. 334–343, Aug. 1985.

[206] W. Huber, V. J. Carey, L. Long, S. Falcon, and R. Gentleman, “Graphs in molecular biology.,” *BMC Bioinformatics*, vol. 8 Suppl 6, p. S8, Jan. 2007.

[207] R. Diestel, *Graph Theory*. Springer Science & Business Media, 2006.

[208] P. Van Mieghem, “Algebraic graph theory,” in *Graph spectra for complex networks*, 1st ed., Cambridge: Cambridge University Press, 2011, pp. 13–25.

[209] A. E. Brouwer and W. H. Haemers, “Graph spectrum,” in *Spectra of graphs*, Springer Science & Business Media, 2011, pp. 1–16.

[210] P. Van Mieghem, “Eigenvalues of the adjacency matrix,” in *Graph spectra for complex networks*, 1st ed., Cambridge: Cambridge University Press, 2011, pp. 29–63.

[211] A. E. Brouwer and W. H. Haemers, “Eigenvalues and eigenvectors,” in *Spectra of graphs*, Springer Science & Business Media, 2011, pp. 33–63.

[212] D. M. Cvetković, P. Rowlinson, and S. Simić, “A background in graph spectra,” in *Eigenspaces of graphs*, 1st ed., Cambridge: Cambridge University Press, 1997, pp. 1–20.

[213] D. M. Cvetković, M. Doob, and H. Sachs, *Spectra of graphs: theory and application*. 1980.

[214] H. S. Wilf, “The eigenvalues of a graph and its chromatic number,” *J. London Math. Soc.*, vol. 42, pp. 330–332, 1967.

[215] A. J. Hoffman, “On eigenvalues and colouring of graphs,” in *Graph Theory and its Applications*, B. Harris, Ed. New York: Academic Press Ltd, 1970, pp. 79–91.

[216] L. Barrière, C. Huemer, D. Mitsche, and D. Orden, “On the Fiedler value of large planar graphs,” *Linear Algebra Appl.*, vol. 439, no. 7, pp. 2070–2084, Oct. 2013.

[217] S. M. Malathy Sony, K. Saraboji, N. Sukumar, and M. N. Ponnuswamy, “Role of amino acid properties to determine backbone tau(N-Calpha-C’) stretching angle in peptides and proteins.,” *Biophys. Chem.*, vol. 120, no. 1, pp. 24–31, 2006.

[218] W. Li and A. Godzik, “Cd-hit: A fast program for clustering and comparing large sets of protein or nucleotide sequences,” *Bioinformatics*, vol. 22, no. 13, pp. 1658–1659, 2006.

[219] W. G. Touw, C. Baakman, J. Black, T. A. H. te Beek, E. Krieger, R. P. Joosten, and G. Vriend, “A series of PDB-related databanks for everyday needs.,” *Nucleic Acids Res.*, vol. 43, no. Database issue, pp. D364-8, Jan. 2015.

[220] W. Kabsch and C. Sander, “Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features.,” *Biopolymers*, vol. 22, no. 12, pp. 2577–2637, 1983.

[221] R. A. Engh and R. Huber, “Accurate bond and angle parameters for X-ray protein structure refinement,” *Acta Crystallogr. Sect. A Found. Crystallogr.*, vol. 47, no. 4, pp. 392–400, Jul. 1991.

[222] F. R. Hampel, “The Influence Curve and its Role in Robust Estimation,” *J. Am. Stat. Assoc.*, vol. 69, no. 346, pp. 383–393, 1974.

[223] T. Benaglia, D. Chauveau, D. R. Hunter, and D. S. Young, “mixtools: An R Package for Analyzing Mixture Models,” *J. Stat. Softw.*, vol. 32, no. 6, 2009.

[224] R. J. Hyndman, “Computing and Graphing Highest Density Regions,” *Am. Stat.*, vol. 50, no. 2, pp.

120–126, 1996.

[225] M. Rosenblatt, “Remarks on Some Nonparametric Estimates of a Density Function,” *The Annals of Mathematical Statistics*, vol. 27, no. 3. pp. 832–837, 1956.

[226] E. Parzen, “On Estimation of a Probability Density Function and Mode,” *The Annals of Mathematical Statistics*, vol. 33, no. 3. pp. 1065–1076, 1962.

[227] R. J. Samworth and M. P. Wand, “Asymptotics and optimal bandwidth selection for highest density region estimation,” *Ann. Stat.*, vol. 38, no. 3, pp. 1767–1792, 2010.

[228] D. I. Svergun, “Solution scattering from biopolymers: advanced contrast-variation data analysis,” *Acta Crystallogr. Sect. A Found. Crystallogr.*, vol. 50, no. 3, pp. 391–402, May 1994.

[229] Á. González, “Measurement of Areas on a Sphere Using Fibonacci and Latitude–Longitude Lattices,” *Math. Geosci.*, vol. 42, no. 1, pp. 49–64, Nov. 2009.

[230] R. Swinbank and R. James Purser, “Fibonacci grids: A novel approach to global modelling,” *Q. J. R. Meteorol. Soc.*, vol. 132, no. 619, pp. 1769–1793, Jul. 2006.

[231] W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery, *Numerical Recipes in Fortran 77: The Art of Scientific Computing*, 2nd ed. Cambridge: University of Cambridge, 1999.

[232] N. K. Fox, S. E. Brenner, and J.-M. Chandonia, “SCOPe: Structural Classification of Proteins--extended, integrating SCOP and ASTRAL data and classification of new structures.,” *Nucleic Acids Res.*, vol. 42, no. Database issue, pp. D304-9, Jan. 2014.

[233] R Developement Core Team, “R: A Language and Environment for Statistical Computing,” *R Foundation for Statistical Computing*, vol. 1. p. 409, 2015.

[234] G. J. Kleywegt, M. R. Harris, J. Y. Zou, T. C. Taylor, A. Wählby, and T. A. Jones, “The Uppsala Electron-Density Server.,” *Acta Crystallogr. D. Biol. Crystallogr.*, vol. 60, no. Pt 12 Pt 1, pp. 2240–9, Dec. 2004.

[235] S. Velankar, C. Best, B. Beuth, C. H. Boutselakis, N. Cobley, A. W. Sousa Da Silva, D. Dimitropoulos, A. Golovin, M. Hirshberg, M. John, E. B. Krissinel, R. Newman, T. Oldfield, A. Pajon, C. J. Penkett, J. Pineda-Castillo, G. Sahni, S. Sen, R. Slowley, A. Suarez-Uruena, J. Swaminathan, G. van Ginkel, W. F. Vranken, K. Henrick, and G. J. Kleywegt, “PDBe: Protein Data Bank in Europe.,” *Nucleic Acids Res.*, vol. 38, no. Database issue, pp. D308-17, Jan. 2010.

[236] M. D. Costabel, M. R. Ermácora, J. A. Santomé, P. M. Alzari, and D. M. A. Guérin, “Structure of armadillo ACBP: a new member of the acyl-CoA-binding protein family.,” *Acta Crystallogr. Sect. F. Struct. Biol. Cryst. Commun.*, vol. 62, no. Pt 10, pp. 958–61, Oct. 2006.

[237] E. F. Pettersen, T. D. Goddard, C. C. Huang, G. S. Couch, D. M. Greenblatt, E. C. Meng, and T. E. Ferrin, “UCSF Chimera—A visualization system for exploratory research and analysis,” *J. Comput. Chem.*, vol. 25, no. 13, pp. 1605–1612, 2004.

[238] G. G. Langer, S. Hazledine, T. Wiegels, C. Carolan, and V. S. Lamzin, “Visual automated macromolecular model building,” *Acta Crystallogr. Sect. D Biol. Crystallogr.*, vol. 69, no. 4, pp. 635–641, Mar. 2013.

[239] D. Frishman and P. Argos, “Knowledge-based protein secondary structure assignment.,” *Proteins*, vol. 23, no. 4, pp. 566–79, Dec. 1995.

[240] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman, “Basic local alignment search tool.,” *J. Mol. Biol.*, vol. 215, no. 3, pp. 403–410, 1990.

[241] Piet Van Mieghem, *Graph Spectra for Complex Networks*, 1st ed. Cambridge: Cambridge University Press, 2012.

[242] A. E. Brouwer and W. H. Haemers, *Spectra of Graphs*. Springer Science & Business Media, 2011.

[243] P. D. Adams, K. Aertgeerts, C. Bauer, J. A. Bell, H. M. Berman, T. N. Bhat, J. M. Blaney, E. Bolton, G. Bricogne, D. Brown, S. K. Burley, D. A. Case, K. L. Clark, T. Darden, P. Emsley, V. A. Feher, Z. Feng, C. R. Groom, S. F. Harris, J. Hendle, T. Holder, A. Joachimiak, G. J. Kleywegt, T. Krojer, J. Marcotrigiano, A. E. Mark, J. L. Markley, M. Miller, W. Minor, G. T. Montelione, G. Murshudov, A. Nakagawa, H. Nakamura, A. Nicholls, M. Nicklaus, R. T. Nolte, A. K. Padyana, C. E. Peishoff, S. Pieniazek, R. J. Read, C. Shao, S. Sheriff, O. Smart, S. Soisson, J. Spurlino, T. Stouch, R. Svobodova, W. Tempel, T. C. Terwilliger, D. Tronrud, S. Velankar, S. C. Ward, G. L. Warren, J. D. Westbrook, P. Williams, H. Yang, and J. Young, “Outcome of the First wwPDB/CCDC/D3R Ligand Validation Workshop.,” *Structure*, vol. 24, no. 4, pp. 502–8, Apr. 2016.

[244] R. Huey, G. M. Morris, A. J. Olson, and D. S. Goodsell, “A semiempirical free energy force field with charge-based desolvation.,” *J. Comput. Chem.*, vol. 28, no. 6, pp. 1145–52, Apr. 2007.

[245] G. J. Kleywegt, M. R. Harris, J. Y. Zou, T. C. Taylor, A. Wählby, and T. A. Jones, “The Uppsala Electron-Density Server.,” *Acta Crystallogr. D. Biol. Crystallogr.*, vol. 60, no. Pt 12 Pt 1, pp. 2240–9, Dec. 2004.

[246] S. J. Weiner, P. A. Kollman, D. A. Case, U. C. Singh, C. Ghio, G. Alagona, S. Profeta, and P.

Weiner, “A new force field for molecular mechanical simulation of nucleic acids and proteins,” *J. Am. Chem. Soc.*, vol. 106, no. 3, pp. 765–784, Feb. 1984.

[247] P. J. Goodford, “A computational procedure for determining energetically favorable binding sites on biologically important macromolecules,” *J. Med. Chem.*, vol. 28, no. 7, pp. 849–857, Jul. 1985.

[248] E. L. Mehler and T. Solmajer, “Electrostatic effects in proteins: comparison of dielectric and charge models.,” *Protein Eng.*, vol. 4, no. 8, pp. 903–10, Dec. 1991.

[249] J. Hinze and H. H. Jaffe, “Electronegativity. I. Orbital Electronegativity of Neutral Atoms,” *J. Am. Chem. Soc.*, vol. 84, no. 4, pp. 540–546, Feb. 1962.

[250] J. Hinze, M. A. Whitehead, and H. H. Jaffe, “Electronegativity. II. Bond and Orbital Electronegativities,” *J. Am. Chem. Soc.*, vol. 85, no. 2, pp. 148–154, Jan. 1963.

[251] J. Gasteiger and M. Marsili, “Iterative partial equalization of orbital electronegativity—a rapid access to atomic charges,” *Tetrahedron*, vol. 36, no. 22, pp. 3219–3228, Jan. 1980.

[252] M. F. Sanner, “Python: a programming language for software integration and development.,” *J. Mol. Graph. Model.*, vol. 17, no. 1, pp. 57–61, Feb. 1999.

[253] G. M. Morris, R. Huey, W. Lindstrom, M. F. Sanner, R. K. Belew, D. S. Goodsell, and A. J. Olson, “AutoDock4 and AutoDockTools4: Automated docking with selective receptor flexibility.,” *J. Comput. Chem.*, vol. 30, no. 16, pp. 2785–91, Dec. 2009.

[254] J. M. Word, S. C. Lovell, J. S. Richardson, and D. C. Richardson, “Asparagine and glutamine: using hydrogen atom contacts in the choice of side-chain amide orientation.,” *J. Mol. Biol.*, vol. 285, no. 4, pp. 1735–1747, 1999.

[255] A. B. Dounay, M. Anderson, B. M. Bechle, B. M. Campbell, M. M. Claffey, A. Evdokimov, E. Evrard, K. R. Fonseca, X. Gan, S. Ghosh, M. M. Hayward, W. Horner, J.-Y. Kim, L. A. McAllister, J. Pandit, V. Paradis, V. D. Parikh, M. R. Reese, S. Rong, M. A. Salafia, K. Schuyten, C. A. Strick, J. B. Tuttle, J. Valentine, H. Wang, L. E. Zawadzke, and P. R. Verhoest, “Discovery of Brain-Penetrant, Irreversible Kynurenine Aminotransferase II Inhibitors for Schizophrenia.,” *ACS Med. Chem. Lett.*, vol. 3, no. 3, pp. 187–92, Mar. 2012.

[256] T. Inoue, N. Okino, Y. Kakuta, A. Hijikata, H. Okano, H. M. Goda, M. Tani, N. Sueyoshi, K. Kambayashi, H. Matsumura, Y. Kai, and M. Ito, “Mechanistic insights into the hydrolysis and synthesis of ceramide by neutral ceramidase.,” *J. Biol. Chem.*, vol. 284, no. 14, pp. 9566–77, Apr. 2009.

[257] S. Miller, B. Tavshanjian, A. Oleksy, O. Perisic, B. T. Houseman, K. M. Shokat, and R. L. Williams, “Shaping development of autophagy inhibitors with the structure of the lipid kinase Vps34.,” *Science*, vol. 327, no. 5973, pp. 1638–1642, 2010.

[258] R. P. Joosten, J. Salzemann, V. Bloch, H. Stockinger, A.-C. Berglund, C. Blanchet, E. Bongcam-Rudloff, C. Combet, A. L. Da Costa, G. Deleage, M. Diarena, R. Fabbretti, G. Fettahi, V. Flegel, A. Gisel, V. Kasam, T. Kervinen, E. Korpelainen, K. Mattila, M. Pagni, M. Reichstadt, V. Breton, I. J. Tickle, and G. Vriend, “PDB_REDO: automated re-refinement of X-ray structure models in the PDB.,” *J. Appl. Crystallogr.*, vol. 42, no. Pt 3, pp. 376–384, Jun. 2009.

[259] T. C. Diller, N.-H. Xuong, and S. S. Taylor, “Molecular Basis for Regulatory Subunit Diversity in cAMP-Dependent Protein Kinase,” *Structure*, vol. 9, no. 1, pp. 73–82, Jan. 2001.

[260] P. H. N. Celie, M. Toebes, B. Rodenko, H. Ovaa, A. Perrakis, and T. N. M. Schumacher, “UV-induced ligand exchange in MHC class I protein crystals.,” *J. Am. Chem. Soc.*, vol. 131, no. 34, pp. 12298–304, Sep. 2009.

[261] R. A. Laskowski, J. A. Rullmannn, M. W. MacArthur, R. Kaptein, and J. M. Thornton, “AQUA and PROCHECK-NMR: programs for checking the quality of protein structures solved by NMR.,” *J. Biomol. NMR*, vol. 8, no. 4, pp. 477–86, Dec. 1996.

Appendix A

Test Cases

Table A.1 The ARP/wARP web service user test cases (as of September 2014) and reference structures used for the determination of the optimum number of geometrisation rounds during automated protein model building. A dashed line divides the high resolution (<2.5 Å) from the lower (medium) resolution (>2.5 Å) structures.

Test case	Resolution (Å)	No. residues	NCS operators	Reference structure
A	1.2	91	1	1GXT
B	1.5	744	2	3TJ4
C	1.6	152	1	3TOW
D	1.6	298	1	1H01
E	1.7	318	1	2REI
F	1.9	122	1	1H6M
G	1.9	241	1	1H2H
H	2.0	154	2	3ZXC
I	2.1	260	2	3OSF
J	2.2	852	2	3QPF
K	2.5	129	1	1AT6
L	2.7	424	1	3SZY
M	2.8	561	1	2XCV
N	2.8	332	2	4AKM
O	2.8	930	4	3MEL
P	2.8	339	1	4AV8
Q	2.9	652	2	2OIQ
R	2.9	240	1	3KWY
S	3.0	339	1	4EXV
T	3.0	154	2	3ZXC
U	3.3	339	1	4AV8
V	3.7	347	1	3ZH9

Table A.3 The full list of PDB ID codes, 3-letter ligand identifiers, ligand size, resolution of the X-ray data and real-space correlation coefficients (RSCC) for 100 protein-ligand structures used in a this work (*continues*).

PDBID	Resolution	Ligand identifier	Number of atoms	Intermolecular energy (kcal/mol per atom)	RSCC
1j4r	1.80	001	45	-0.24	0.93
2fjn	2.20	073	43	-0.40	0.95
2jfz	1.86	003	33	-0.33	0.91
2pjl	2.30	047	25	-0.51	0.92
2x6i	3.40	090	26	-0.32	0.92
2x6j	3.50	093	24	-0.15	0.90
2xyw	3.14	08S	34	-0.38	0.94
2zxc	2.20	2ED	13	0.07	0.49
3ckr	2.70	009	45	-0.28	0.87
3ik3	1.90	0LI	39	-0.43	0.91
3r2b	2.90	05B	30	-0.29	0.92
3rcd	3.21	03P	38	-0.30	0.95
3roa	2.30	06V	28	-0.37	0.90
3sgy	2.60	06W	28	-0.35	0.80
3ske	1.97	054	35	-0.27	0.98
3skh	2.50	058	27	-0.31	0.94
3tgs	2.70	03G	23	-0.37	0.96
3tjc	2.40	0TP	26	-0.33	0.92
3twj	2.90	07R	23	-0.24	0.94
3ty0	2.00	082	36	-0.34	0.86
3tym	2.00	08R	27	-0.26	0.94
3u4o	1.77	08E	28	-0.32	0.96
3ue8	3.22	09M	28	0.93	0.95
3unz	2.80	0BZ	24	-0.36	0.97
3uoh	2.80	0C4	24	-0.41	0.97
3uok	2.95	0C6	25	-0.35	0.97
3uol	2.40	0C7	26	-0.32	0.96
3utd	1.70	0CJ	15	-0.46	0.99
3uuo	2.11	0CV	23	-0.32	0.91
3uuz	2.10	0CB	38	-0.29	0.92
3uv3	1.60	0CM	13	-0.46	0.99

Table A.3 The full list of PDB ID codes, 3-letter ligand identifiers, ligand size, resolution of the X-ray data and real-space correlation coefficients (RSCC) for 100 protein-ligand structures used in a this work (*cont.*).

PDBID	Resolution	Ligand identifier	Number of atoms	Intermolecular energy (kcal/mol per atom)	RSCC
3uv6	1.70	0CH	14	-0.49	0.97
3uv7	1.60	0CN	13	-0.50	0.98
3uwk	1.91	0DF	17	-0.48	0.95
3uwo	1.70	0DJ	20	-0.46	0.96
3uz5	1.90	0CU	11	-0.47	0.95
3v1s	2.33	0LH	12	-0.58	0.98
3v5j	2.59	0F2	32	-0.33	0.96
3v5l	1.86	0G1	31	-0.36	0.95
3v7t	2.09	0GX	30	-0.39	0.97
3v8s	2.29	0HD	21	-0.42	0.97
3v8w	2.27	0G2	29	-0.32	0.96
4d7o	1.78	0GD	25	-0.24	0.94
4d8z	2.20	0J2	19	-0.24	0.89
4d9m	2.50	0JO	21	-0.38	0.95
4daf	2.50	0J4	18	-0.34	0.91
4daj	3.40	0HK	26	-0.39	0.98
4dce	2.03	0JF	36	-0.35	0.89
4ddl	2.07	0JQ	31	-0.26	0.89
4dhf	2.80	0K6	33	-0.25	0.91
4di2	2.00	0K9	42	-0.34	0.94
4djw	1.90	0KP	26	-0.37	0.90
4djy	1.86	0KR	26	-0.41	0.97
4dk7	2.45	0KS	27	-0.29	0.85
4dk8	2.75	0KT	32	-0.37	0.89
4dkp	1.80	0LL	24	-0.30	0.80
4dkr	1.80	0LJ	27	-0.29	0.91
4dma	2.30	0L8	21	-0.43	0.71
4dt2	2.70	0LV	13	-0.19	0.74
4dvw	2.20	0M4	23	-0.4	0.95
4dya	2.75	0MF	39	-0.29	0.95
4dyp	2.82	0MS	29	-0.35	0.95

Table A.3 The full list of PDB ID codes, 3-letter ligand identifiers, ligand size, resolution of the X-ray data and real-space correlation coefficients (RSCC) for 100 protein-ligand structures used in a this work (*cont.*).

PDBID	Resolution	Ligand identifier	Number of atoms	Intermolecular energy (kcal/mol per atom)	RSCC
4e4n	1.90	0NL	25	-0.36	0.96
4e6q	1.95	0NV	25	-0.34	0.95
4eb3	1.90	0O3	15	-0.34	0.96
4edz	2.00	0O5	28	-0.33	0.81
4ee0	1.75	0O4	27	-0.30	0.88
4ei4	2.22	0Q2	20	-0.37	0.96
4eo6	1.79	0S2	27	-0.33	0.96
4eo8	1.80	0S3	27	-0.33	0.95
4ere	1.80	0R2	35	-0.31	0.82
4erf	2.00	0R3	32	-0.29	0.85
4f1q	2.80	0RZ	17	-0.38	0.81
4fam	2.00	0SZ	22	-0.39	0.98
4fcd	2.02	0T6	25	-0.32	0.94
4fes	2.00	0T9	31	-0.36	0.94
4ffg	2.30	0U8	22	-0.19	0.92
4fic	2.50	0UL	16	-0.34	0.96
4frs	1.70	0V6	25	-0.47	0.96
4g0k	2.56	0VS	33	-0.20	0.88
4g2l	3.00	0WL	27	-0.32	0.94
4gcq	2.20	0JM	26	-0.29	0.76
4gdy	2.89	0X1	37	-0.35	0.96
4ge4	2.41	0KE	30	-0.47	0.97
4geb	2.15	0LD	33	-0.39	0.96
4gid	2.00	0GH	47	-0.43	0.94
4gj5	2.40	0LR	24	-0.50	0.96
4gj7	2.80	0LT	33	-0.43	0.94
4gja	2.60	0M3	34	-0.43	0.95
4gjb	2.75	0ME	24	-0.46	0.97
4gjc	2.40	0MJ	35	-0.43	0.94
4gjd	2.65	0N0	41	-0.42	0.98

Table A.3 The full list of PDB ID codes, 3-letter ligand identifiers, ligand size, resolution of the X-ray data and real-space correlation coefficients (RSCC) for 100 protein-ligand structures used in a this work.

PDBID	Resolution	Ligand identifier	Number of atoms	Intermolecular energy (kcal/mol per atom)	RSCC
4gqs	2.87	0XV	21	-0.31	0.95
4gsy	1.71	0Y5	36	-0.31	0.97
4hbm	1.90	0Y7	30	-0.40	0.94
4j4b	1.90	0TF	18	-0.53	0.89
4jqc	2.80	0WE	28	-0.34	0.92
4r09	2.62	06S	48	-0.18	0.94
4tkg	1.95	09L	32	-0.44	0.88
5c5p	1.75	0E0	20	-0.47	0.95

Appendix B

Supplementary Result Tables

Table B.1 Parameters of the Gaussian functions describing the distribution of the three fixed distances and the ω dihedral angle in trans-peptide units. μ, the mean (in Å); σ, the standard deviation; ν, the mixing proportion.

	Cα_i-O$_i$	**Cα_i- Cα_{i+1}**	**O$_i$-Cα_{i+1}**	**ω**
Major Gaussian Function				
μ	2.399	3.808	2.774	179.4
σ	0.010	0.017	0.033	2.6
ν	0.64	0.63	0.77	0.56
Minor Gaussian Function				
μ	2.398	3.812	2.775	179.2
σ	0.026	0.034	0.062	7.4
ν	0.36	0.36	0.23	0.44

Table B.2 Linear correlation coefficients (*r*) between the square-root of the absolute values of the four negative eigenvalues of Euclidean matrices and the variable distances (in Å) as well as between their absolute values and the three principal components of the *xyz* variance-covariance matrices (*Dc$_i$*) of 5-atom dipeptide units (in Å^2).

	$\sqrt{\lambda_1}$	$\sqrt{\lambda_2}$	$\sqrt{\lambda_3}$	$\sqrt{\lambda_4}$
Variable distances				
Cα_{i-1}-O$_i$	-0.297	0.864	-0.666	-0.259
O$_{i-1}$-O$_i$	0.808	-0.297	0.420	-0.058
Cα_{i-1}-Cα_{i+1}	0.889	-0.905	0.705	-0.077
O$_{i-1}$-Cα_{i+1}	0.958	-0.580	0.528	-0.423
	λ_1	**λ_2**	**λ_3**	**λ_4**
Dipeptide unit principal components (Dc$_i$)				
Dc$_1$	1.000	-0.666	0.573	-0.161
Dc$_2$	-0.586	0.941	-0.657	-0.434
Dc$_3$	-0.705	0.216	-0.085	0.188
Rg^2	0.980	-0.533	0.525	-0.304

Table B.3 Average moments and associated standard deviation (in parenthesis) of the DipScore distributions computed for the set of 538 chains (all chains) and for the different subsets representing different SCOPe fold classes.

	N	Average (m_1)	Variance (m_2)	Skewness (m_3)	Kurtosis (m_4)
All chains	538	0.91 (0.02)	0.027 (0.009)	-2.9 (0.5)	9 (5)
All alpha	92	0.92 (0.02)	0.026 (0.008)	-3.1 (0.6)	11 (5)
All beta	112	0.90 (0.02)	0.028 (0.009)	-2.8 (0.4)	8 (4)
α/β	221	0.91 (0.02)	0.028 (0.009)	-2.9 (0.4)	9 (3)
α+β	85	0.91 (0.02)	0.026 (0.008)	-3.0 (0.5)	10 (4)
Coiled coil	4	0.98 (0.01)	0.008 (0.007)	-5.3 (2.1)	33 (23)

Table B.4 Median moments and associated median absolute deviation (MADe; in parenthesis) of the DipScore distributions computed for the set of 538 chains (all chains) and for the different subsets representing different SCOPe fold classes.

	N	Average (m_1)	Variance (m_2)	Skewness (m_3)	Kurtosis (m_4)
All chains	538	0.91 (0.02)	0.026 (0.007)	-2.9 (0.4)	9 (3)
All alpha	92	0.92 (0.02)	0.025 (0.008)	-3.1 (0.5)	10 (4)
All beta	112	0.90 (0.01)	0.028 (0.007)	-2.7 (0.3)	8 (2)
α/β	221	0.91 (0.01)	0.026 (0.006)	-2.9 (0.4)	8 (3)
α+β	85	0.91 (0.01)	0.026 (0.007)	-2.9 (0.4)	9 (3)
Coiled coil	4	0.97 (0.01)	0.009 (0.006)	-5.2 (1.8)	28 (15)

Table B.5 p-values resulting from the two-sided student t-test for the comparison between the mean average (m_1) DipScore for each set (H_0: the means are equal; H_1: the means are different).

	All alpha	All beta	Mixed	Coiled coil
All chains	9.5×10^{-5}	2.4×10^{-4}	0.60	0.002
All alpha	-	4.8×10^{-9}	5.2×10^{-5}	0.002
All beta	-	-	0.002	9.1×10^{-4}
Mixed	-	-	-	0.002

Table B.6 p-values resulting from the two-sided student t-test for the comparison between the mean DipScore variance (m_2) for each set (H_0: the means are equal; H_1: the means are different).

	All alpha	All beta	Mixed	Coiled coil
All chains	0.05	0.26	0.63	0.01
All alpha	-	0.02	0.03	0.01
All beta	-	-	0.45	0.008
Mixed	-	-	-	0.01

Table B.7 p-values resulting from the two-sided student t-test for the comparison between the mean DipScore skewness (m_3) for each set (H_0: the means are equal; H_1: the means are different).

	All alpha	All beta	Mixed	Coiled coil
All chains	0.002	0.002	0.30	0.10
All alpha	-	3.0×10^{-6}	3.6×10^{-4}	0.12
All beta	-	-	0.020	0.09
Mixed	-	-	-	0.10

Table B.8 p-values resulting from the two-sided student t-test for the comparison between the mean DipScore kurtosis (m_4) for each set (H_0: the means are equal; H_1: the means are different).

	All alpha	All beta	Mixed	Coiled coil
All chains	0.01	0.02	0.19	0.12
All alpha	-	2.8×10^{-4}	0.002	0.14
All beta	-	-	0.15	0.11
Mixed	-	-	-	0.12

Table B.9 Target values for the calculation of the Z-scores for the first four central moments of the DipScore distribution.

	Average (m_1)	Variance (m_2)	Skewness (m_3)	Kurtosis (m_4)
μ	0.9002	0.02686	-2.916	8.998
σ	0.0156	0.00717	0.410	3.128

Table B.10 Non proline/glycine Ramachandran-plot or DipSpace outliers for the MX model of the armadillo acyl-CoA-binding protein (ACBP)[236] (PDB ID 2fdq), out of a total of 231 residues evaluated.

Residue	φ, ψ, τ (°)	Ramachandran status	DipScore	DipCheck status
Ala9A	-56.3, -75.7, 105.1	Outlier	0.036	Allowed
Glu10A	-39.3, -47.2, 110.0	Allowed	0.004	Outlier
Val12A	-37.2, -39.6, 113.7	Allowed	0.000	Outlier
Lys16A	-37.9, -29.4, 118.7	Outlier	0.773	Favored
Asp22A	-35.1, -40.6, 111.3	Allowed	0.000	Outlier
Ile39A	-33.8, 137.6, 115.3	Outlier	0.990	Favored
Thr64A	-40.3, 153.5, 113.1	Outlier	0.990	Favored
Ala20B	-39.2, 157.9, 111.1	Outlier	0.984	Favored
Asp22B	-53.3, -75.8, 104.8	Outlier	0.001	Outlier
Glu23B	-32.1, -55.0, 111.3	Allowed	0.001	Outlier
Phe26B	-32.1, -54.1, 109.7	Allowed	0.001	Outlier
Asp48B	-66.3, 44.7, 108.6	Outlier	0.029	Gen. allowed
Lys66B	-29.0, -54.9, 106.7	Allowed	0.000	Outlier
Tyr73B	-47.7, -95.5, 108.7	Outlier	0.407	Favored
Ile74B	-28.0, -43.4, 110.5	Outlier	0.000	Outlier
Ile27C	-57.2, -74.8, 109.1	Outlier	0.121	Allowed
Tyr28C	-34.3, -72.8, 111.7	Outlier	0.005	Outlier
Tyr31C	-51.3, -29.8, 106.8	Favored	0.005	Outlier
Gln33C	-47.9, 2.1, 113.6	Outlier	0.857	Favored
Lys52C	-44.4, -34.8, 109.7	Favored	0.009	Outlier
Gln60C	-45.2, -31.1, 110.6	Allowed	0.004	Outlier
Asp75C	-36.0, -71.8, 107.3	Allowed	0.000	Outlier

Table B.11 Average, and associated standard deviation (in parenthesis), of the total number mesh points (n), normalised (λ_1/n) and non-normalised (λ_1) largest, normalised (λ_n/n) and non-normalised (λ_n) smallest eigenvalues, normalised spectral gap ((λ_1-λ_2)/n) and chromatic number (χ(G)/n) computed for the mesh representation of side-chains at 2.0 Å resolution.

	N	n	λ_1	λ_1/n	λ_n	λ_n/n	(λ_1- λ_2)/n	χ(G)/n
Tryptophan	5	89.0 (7.7)	36.0 (2.0)	0.41 (0.02)	-6.13 (0.08)	-0.07 (0.01)	0.15 (0.02)	0.009 (0.001)
Arginine	7	76.0 (8.9)	35.1 (4.5)	0.46 (0.02)	-6.30 (0.90)	-0.08 (0.01)	0.20 (0.02)	0.011 (0.001)
Asparagine	5	32.4 (13.5)	19.7 (7.8)	0.61 (0.04)	-3.77 (1.02)	-0.12 (0.02)	0.35 (0.07)	0.028 (0.011)
Valine	10	27.5 (8.1)	19.1 (3.9)	0.71 (0.09)	-3.59 (0.42)	-0.14 (0.02)	0.49 (0.11)	0.031 (0.007)

Table B.12 p-values resulting from the two-sided student t-test for the comparison between the mean side-chain mesh largest eigenvalue.

	Arginine	Asparagine	Valine
Tryptophan	0.65	0.008	5.2×10^{-8}
Arginine	-	0.008	7.0×10^{-6}
Asparagine	-	-	0.88

Table B.13 p-values resulting from the two-sided student t-test for the comparison between the mean side-chain mesh smallest eigenvalue.

	Arginine	Asparagine	Valine
Tryptophan	0.63	0.007	3.6×10^{-9}
Arginine	-	0.002	8.4×10^{-5}
Asparagine	-	-	0.72

Table B.14 p-values resulting from the two-sided student t-test for the comparison between the mean side-chain mesh normalised largest eigenvalue.

	Arginine	Asparagine	Valine
Tryptophan	8.2×10^{-4}	3.5 x 10-5	7.5×10^{-7}
Arginine	-	1.1 x 10-4	3.7×10^{-6}
Asparagine	-	-	0.007

Table B.15 p-values resulting from the two-sided student t-test for the comparison between the mean side-chain mesh normalised smallest eigenvalue.

	Arginine	Asparagine	Valine
Tryptophan	0.02	0.003	2.6×10^{-6}
Arginine	-	0.008	1.9×10^{-5}
Asparagine	-	-	0.27

Table B.16 p-values resulting from the two-sided student t-test for the comparison between the mean side-chain mesh normalised spectral gap.

	Arginine	Asparagine	Valine
Tryptophan	0.004	0.003	3.0×10^{-6}
Arginine	-	0.008	1.2×10^{-5}
Asparagine	-	-	0.01

Table B.17 p-values resulting from the two-sided student t-test for the comparison between the mean side-chain mesh normalised chromatic number.

	Arginine	Asparagine	Valine
Tryptophan	0.05	0.009	2.5×10^{-6}
Arginine	-	0.01	2.7×10^{-6}
Asparagine	-	-	0.33

Table B.18 Individual energetic contribution computed for each atom in ligand 093 binding to the Drosophila class III PI3-kinase VPS34 (PDB ID 2x6j).

Ligand atom	VDW+H-bond+electrostatic energies (kcal/mol)
CAF	-0.21
CAB	-0.33
CL	4.87
CAC	-0.42
CAD	-0.47
CAG	-0.32
CAH	-0.36
CAI	-0.34
CAJ	-0.49
CAE	-0.68
NAK	-0.61
CAQ	-0.35
SAP	-0.54
NAR	-0.37
CAS	-0.26
OAL	-0.44
CAT	-0.43
SAN	-0.03
OAM	-0.52
OAO	-0.40
NAU	-0.44
CAV	-0.16
CAW	-0.21
OAX	-0.11

Table B.19 The intermolecular energies calculated for the ligands identified by the ligand-guess method with the mutant of the type II beta regulatory subunit of cAMP-dependent protein kinase (PDB ID 1cx4).

Ligand identifier	Number of atoms	Ranking by shape similarity	RSCC	Inter-molecular energy (kcal/mol per atom)
CMP	22	9	0.88	-0.36
LDA	16	16	0.83	-0.31
5GP	24	15	0.92	-0.25
P6G	19	7	0.85	-0.24
AMP	23	1	0.82	-0.03
SAM	27	4	0.88	0.10
MYR	16	19	0.84	0.35
RET	20	5	0.89	2.52
OLA	20	12	0.75	5.41
TYD	25	8	0.82	43.05

Table B.20 The intermolecular energies calculated for ligands identified by the ligand-guess method with the complex of MHC class I HLA-A2.1 and HIV-1 envelope peptide ENV120-128 (PDB ID 2x4o).

Ligand identifier	Number of atoms	Ranking by shape similarity	RSCC	Inter-molecular energy (kcal/mol per atom)
MES	12	24	0.85	-0.21
NHE	13	20	0.84	-0.19
CXS	14	12	0.76	-0.17
PG4	13	25	0.70	-0.16
1PE	16	19	0.74	-0.15
GSH	20	6	0.83	-0.13
ADN	19	28	0.82	-0.13
HC4	12	31	0.86	-0.07
AKG	10	37	0.88	-0.04
PLP	16	30	0.78	-0.01

Table B.22 The intermolecular energies calculated for ligands identified by ligand-guess method with the Staphylococcal nuclease variant truncated Delta+PHS I92W (PDB ID 2of1).

Ligand identifier	Number of atoms	Ranking by shape similarity	RSCC	Inter-molecular energy (kcal/mol per atom)
THP	25	27	0.78	-0.28
A3P	27	3	0.88	-0.22
MYR	16	37	0.76	-0.21
1PE	16	24	0.72	-0.17
PEG	7	4	0.83	-0.15
FPP	24	23	0.72	-0.15
TAM	11	28	0.60	-0.15
P6G	19	18	0.80	-0.07
BCL	66	15	0.74	-0.01
HEM	43	21	0.72	0.00

Appendix C

Supplementary Result Figures

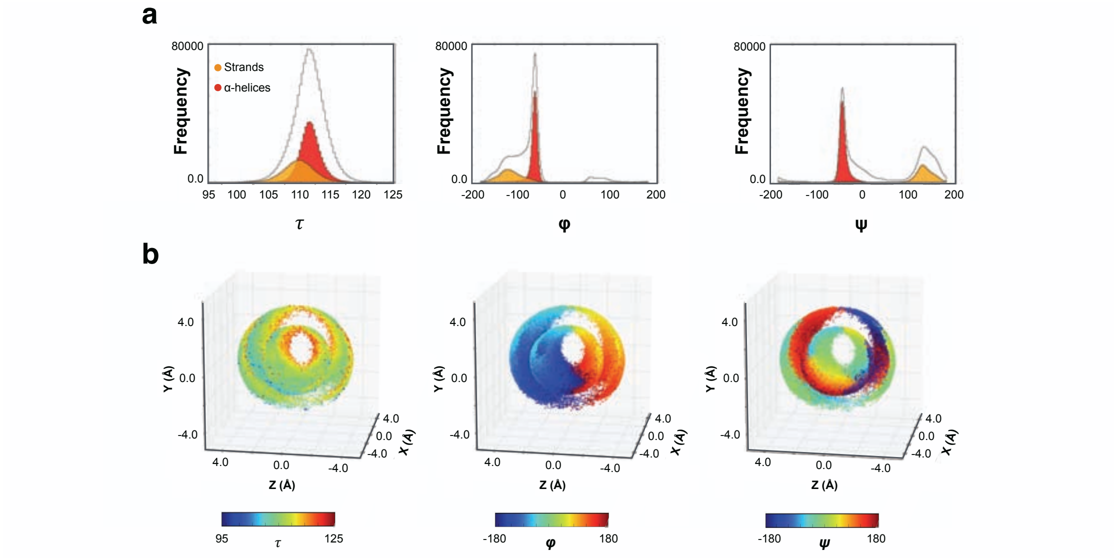


Figure C.1 The three degrees of freedom of full atom trans dipeptide units. (a) Histogram showing their one-dimensional distribution on the collected set of dipeptide units, depicting the contribution of the structural preferences of the main-chain (helices and strands). (b) Their mapping on the sampled space.

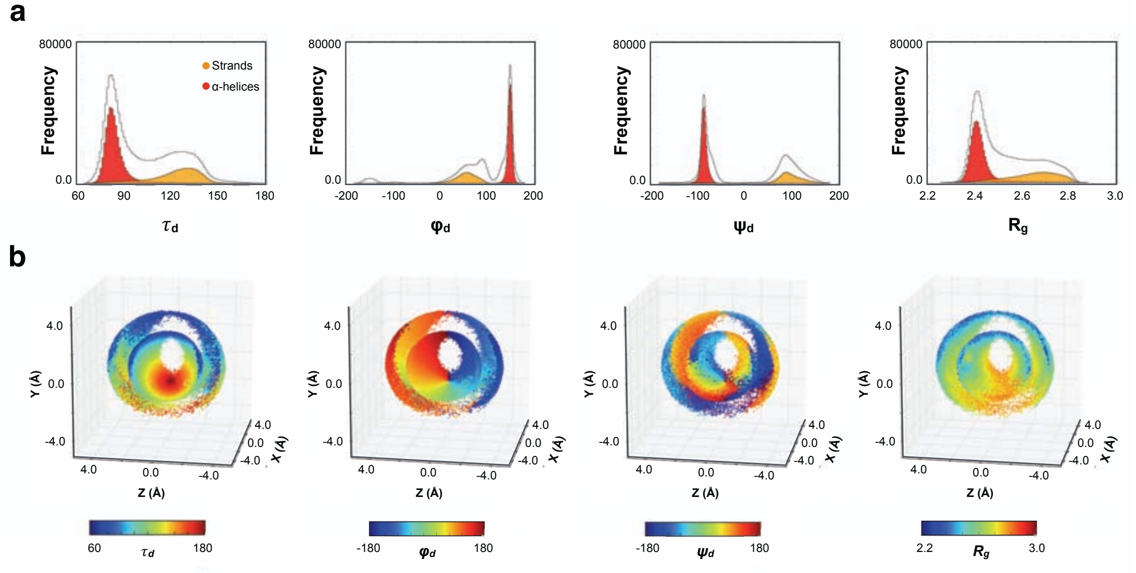


Figure C.2 The three degrees of freedom and the radius of gyration (R_g) of 5-atom trans dipeptide units. (a) Histogram showing their one-dimensional distribution on the collected set of dipeptide units, depicting the contribution of the structural preferences of the main-chain (helices and strands). (b) Their mapping on the sampled space.

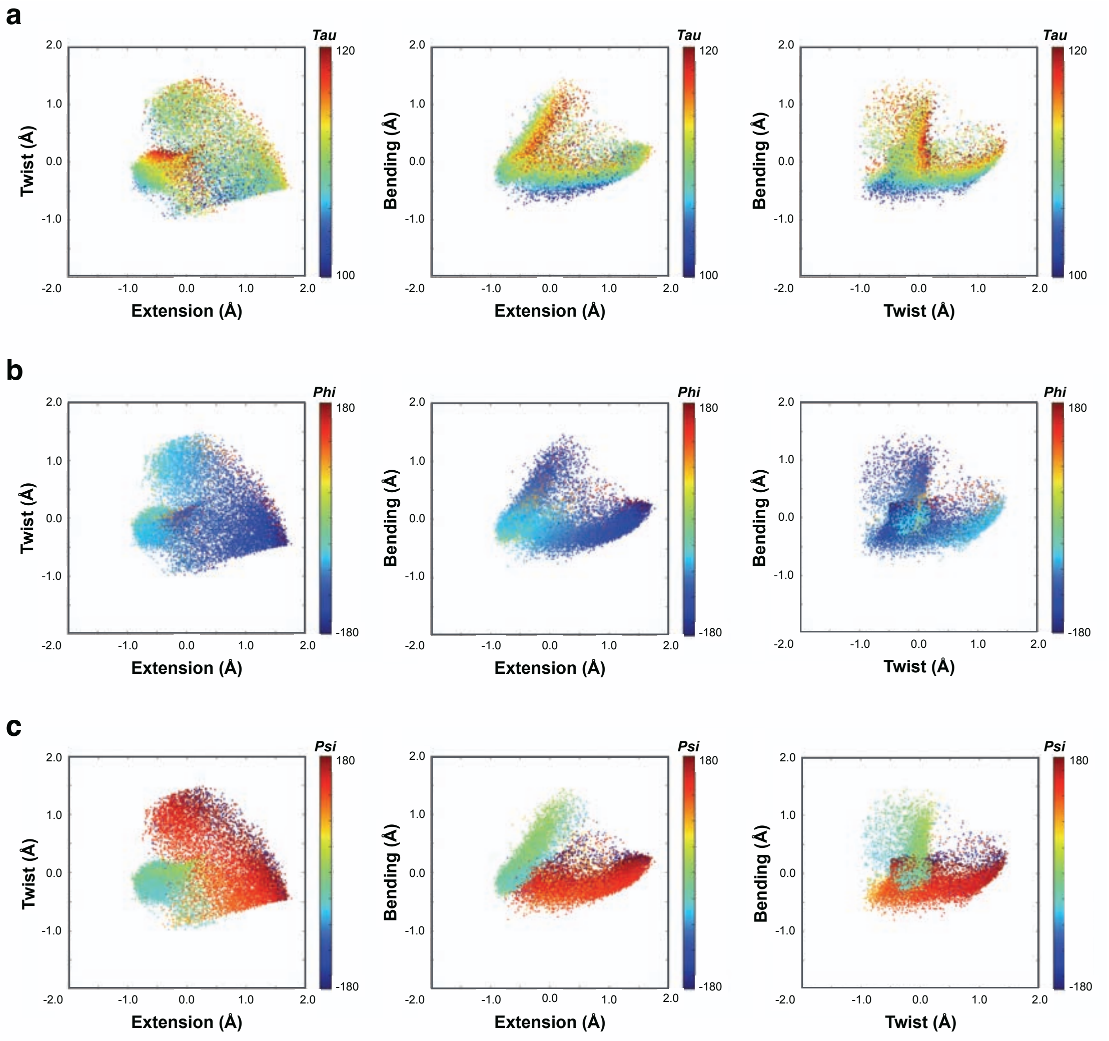


Figure C.3 The DipSpace coloured according to (a) the τ stretching angle, (b) the Ramachandran φ dihedral angle, and (c) the Ramachandran ψ dihedral angle.

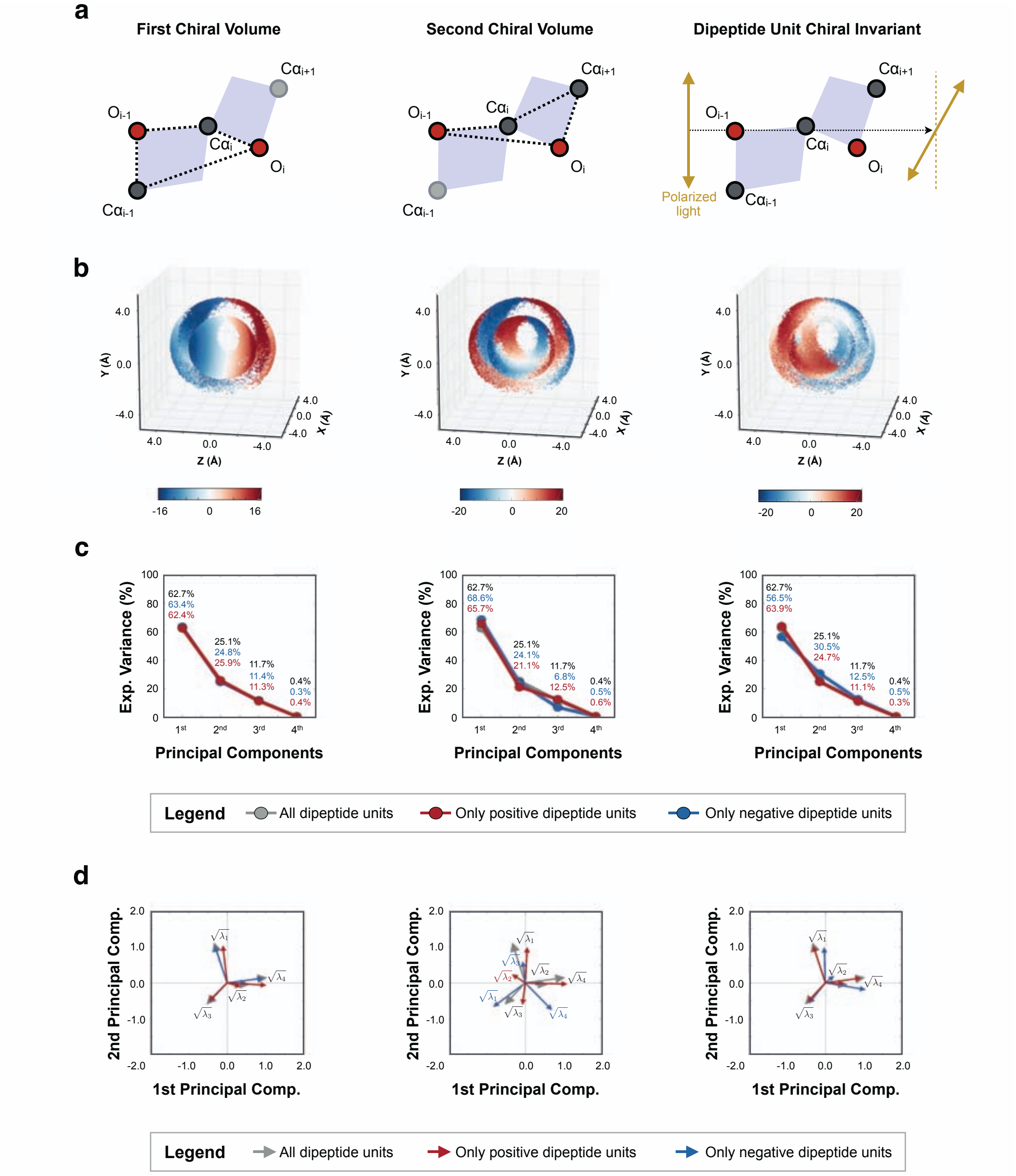


Figure C.4 Separating mirror imaged dipeptide units by a chirality measure. (a) The three different chirality measures tested and (b) the sampled space coloured according to them (Red: positive chirality; blue: negative chirality; light grey: close to achiral). (c) Principal components explained variance and (d) projections obtained when the distance-squared matrix eigenvalue dataset is separated into two (negative and positive chirality) according to each chiral measure and used for dimensionality reduction by PCA.

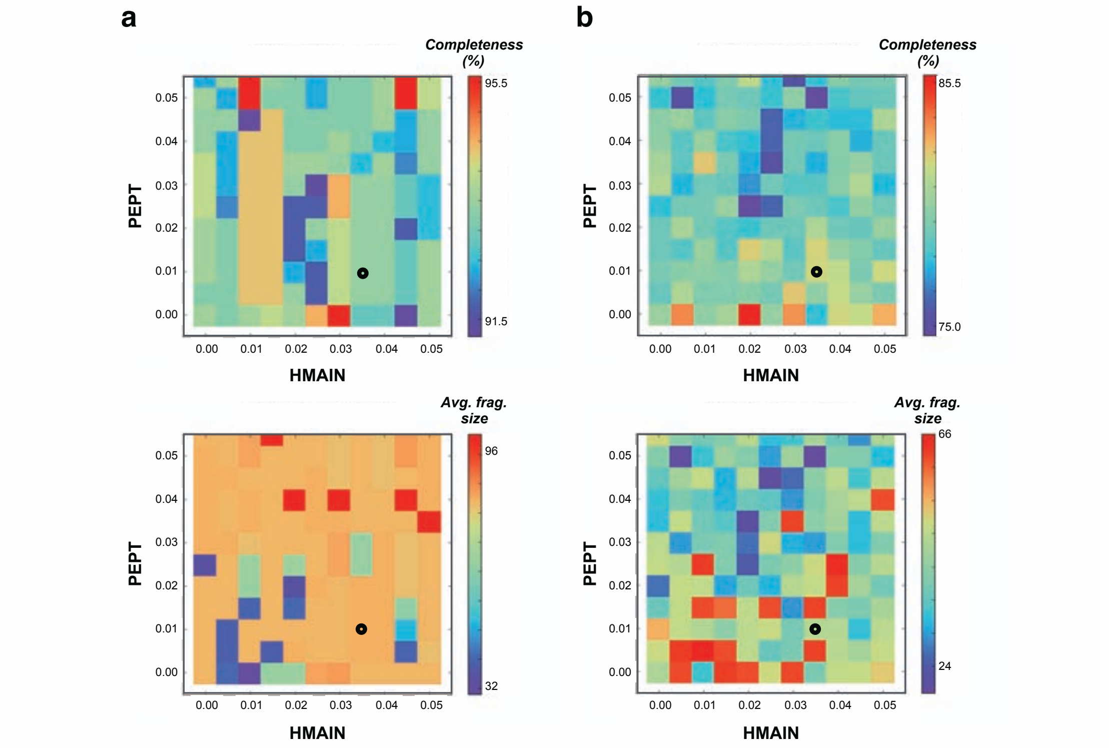


Figure C.5 Heat maps describing the final model completeness and average fragment size for the (a) high resolution (< 2.5 Å) and (b) lower resolution sets (2.5-3.0 Å). A black circle marks the *hmain/pept* DipScore thresholds implemented in ARP/wARP 7.6.

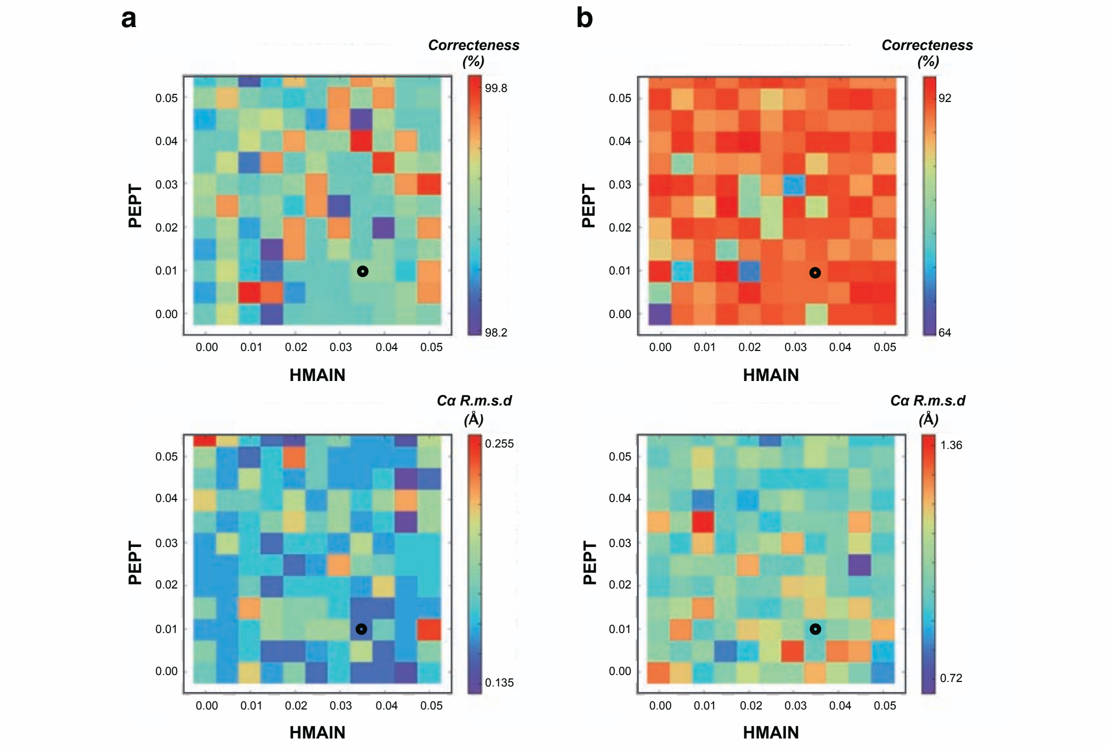


Figure C.6 Heat maps describing the final model correctness and Cα r.m.s.d. to the reference for the (a) high resolution (< 2.5 Å) and (b) lower resolution sets (2.5-3.0 Å). A black circle marks the *hmain/pept* DipScore thresholds implemented in ARP/wARP 7.6.

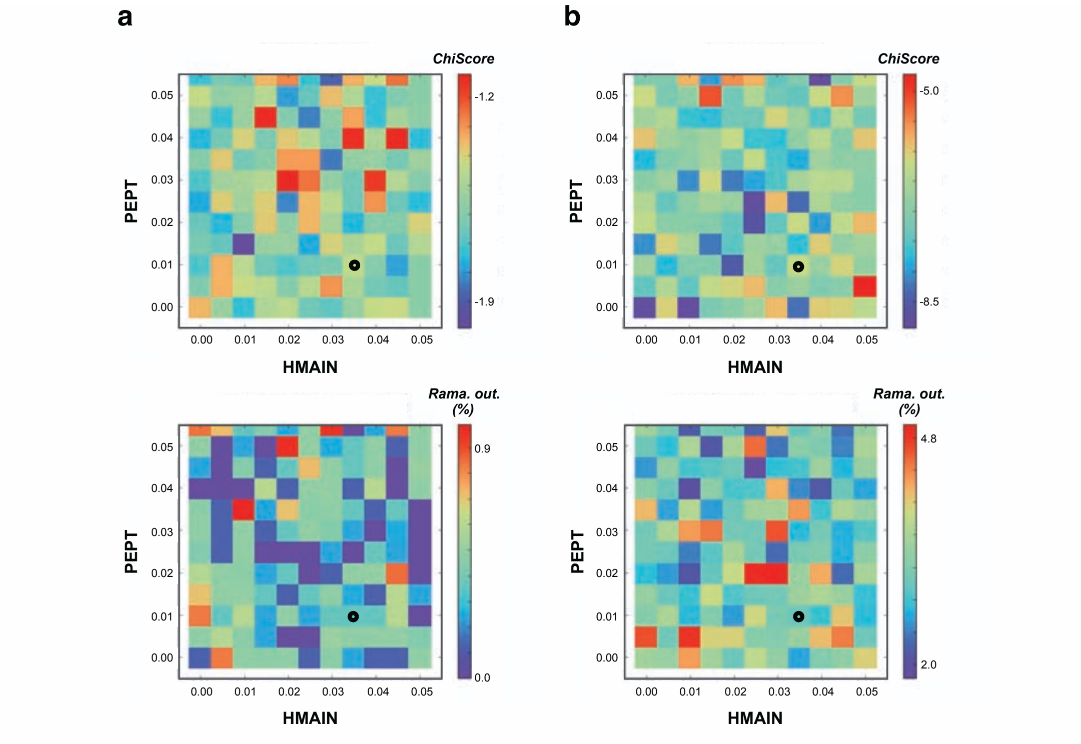


Figure C.7 Heat maps describing the final model ChiScore and percentage of Ramachandran outliers for the (a) high resolution (< 2.5 Å) and (b) lower resolution sets (2.5-3.0 Å). A black circle marks the DipScore thresholds implemented in ARP/wARP 7.6.

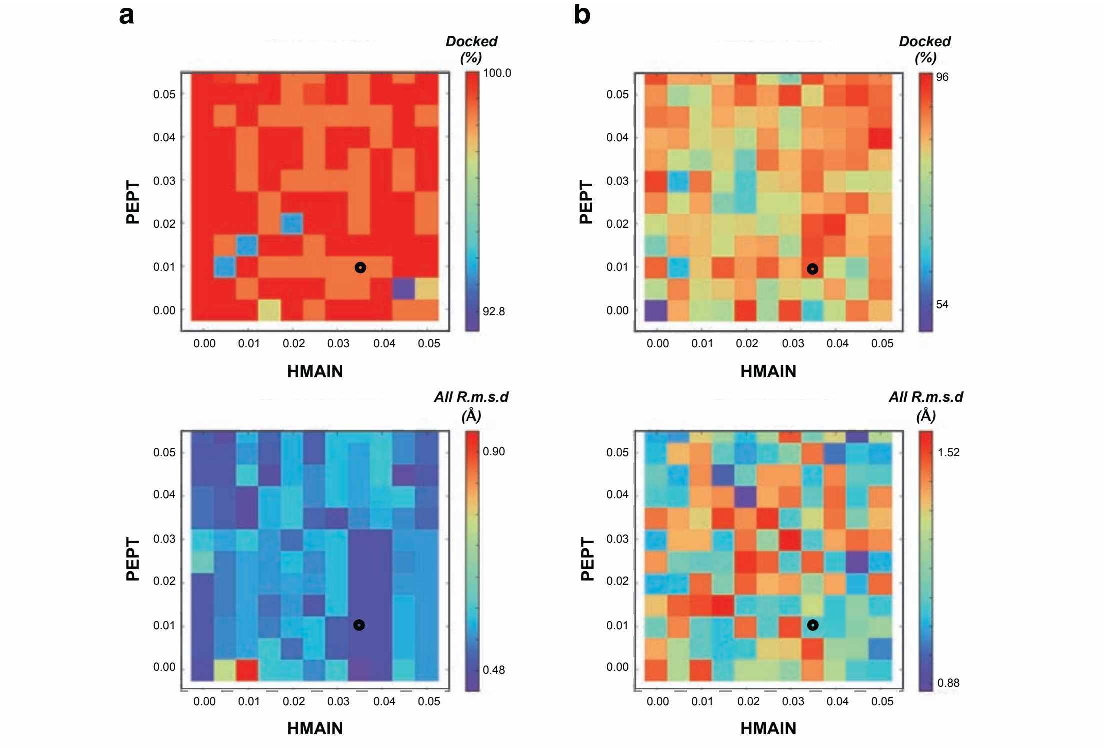


Figure C.8 Heat maps describing the final model percentage of correctly docked residues and all atom r.m.s.d. to the reference for the (a) high resolution (< 2.5 Å) and (b) lower resolution sets (2.5-3.0 Å). A black circle marks the DipScore thresholds implemented.

Appendix D

Legends of Supplementary Result Videos

Video D.1 Conformational transition of the dipeptide units along the DipSpace *pc1* axis, demonstrating their ‘extension’.

Video D.2 Conformational transition of the dipeptide units along the DipSpace *pc2* axis, demonstrating their ‘twist’.

Video D.3 Conformational transition of the dipeptide units along the DipSpace *pc3* axis, demonstrating their ‘bending’.

Appendix E

List of Hazardous Substances

The presented work is purely theoretical. Therefore, no laboratory experiments were carried out and no hazardous, carcinogenic, mutagenic or toxic substances according to GHS were used.

Appendix F

Acknowledgements

After four years working at the EMBL, I have a great list of people who contributed in some way to this thesis, and to which I would like to express my gratitude.

First of all, I would like to thank Dr. Victor Lamzin for giving me the opportunity to work in his group at EMBL and believing that a biochemist without any programming skill would be able to learn, develop methods for and work in bioinformatics. I am very grateful that him and his group made such an effort to teach me about programming, crystallography and math, and supported my attendance to a number of international meetings and conferences. I also would like to thank Victor, the CCP4 guys and David Aragão from the Australian Synchrotron for encouraging, inviting and funding me to travel to the four corners of the World to teach about ARP/wARP. With this I could visit several synchrotrons and research institutes, meet several of the big crystallography names, get incredible feedback about my work and ARP/wARP, get insights into the crystallographic field otherwise very hard and do a round trip to planet Earth!

According to the EMBL regulations, every PhD student needs to have a University PhD supervisor and Prof. Dr. Andrew E. Torda happily accepted that role. I am very thankful to Andrew for all the helpful discussions and all the help with University matters. I can not forget the other three members of my Thesis Advisory Committee (TAC) - Thomas Schneider, Gerard Kleywegt and Inari Kursula - for the amazing feedback during our yearly meetings that helped shaping the project too. Thank you also for all the effort to be always present in these meetings and for in some cases even waking up at 4 a.m. just to travel to Hamburg.

I also would like to thank all members of the Lamzin group, especially my office mate Claudia Hackenberg, without whom these four years in the office would be boring. Thank you for all the laughs and friendship. I am very happy I made you love Portugal as much as I do. Office 116 will always be the ‘Champions Office’! A special thank you goes also to Philipp Heuser, Grzegorz Chojnowski, Tim Wiegels and Daria Beshnova for helpful discussions and feedback on my thesis work and great company during the ARP/wARP ‘world tours’. Philipp and Grezgorz, thank you for proofreading all my written work too (papers and thesis), Daria thank you for including me, and my ideas, on your ligands’ project and taking care of most of the experiments for LigEnergy. Philipp Heuser and Umut Oezugurel a big thanks for helping with setting up the DipCheck webserver.

Life in Hamburg wouldn't be the same without the EMBL family. Here I made great friends that I will take with me everywhere I go. I want to thank Margret Fischer for taking care of all of the EMBL members in Hamburg as a mother does, for baking for us and making Fridays mornings so special with freshly baked scones, but also for caring about our personal issues and making everything smooth. For that I also want to thank every member of the EMBL graduate office. To Rosemary Wilson, I would like to thank for always coming to me to share about my work and experiences around the world, it was a great opportunity to share my stories with you and the EMBL community. To my friends Diana Mendes Freire, Natasha Giannopoulou, Kate Beckhman, Anne-Sophie Huart and Anna Polyakova, thank you for the great lunch-hour everyday and our after work events. Diana Mendes Freire, my 'sister from another mister', I will never forget these four years with you, as those sharing an awesome Portuguese apartment, helping and caring for each other like only family does. To your mother, Maria Armanda Mendes, I am very thankful for taking care of me as a 'daughter' even when you needed all the attention. You are my second family and I will always be with you as I was during these four years.

Still, I couldn't reach where I am now and accomplish so much without my family. I want to thank my parents, Maria de Fátima Soares and José Paulo Pereira, and my brother, João Paulo Pereira, for their unconditional love and patience during these four years, specially for all the support and opening their doors so that I would adventure myself in Germany by myself. Thank you for all the calls every night, the visits and words of love and confidence that keep me going all these years doing what I love. To my uncles, Alice Soares and Domingos Ferreira, cousins, Carla Ferreira and Pedro Ferreira, and grandmothers, Joana Garcia and Maria do Rosário Moreira, for all their words of love, hope and strength. I missed you all so much! Finally but not least, I owe my deepest gratitude to my boyfriend, João Barros. Even more than 1.200 km apart, you were with me every single day, either via message or skype, showing that a long distance relationship can work if we want to. These four years would have been much harder if I did not have you everyday and did not have your love and friendship. Thank you for all the brainstorming, help and support. For being such a big part of what I managed to reach today, I dedicate my thesis to all of you!

Lastly, I would like to thank all of the others I love and supported me but that by space limitation are not mentioned here and the EMBL for funding my PhD fellowship, allowing me to work on such a great environment and giving me so many opportunities during four years and three months.

Appendix G

Declaration Upon Oath

I hereby declare on oath, that I have written the present dissertation by my own and have not used other than the acknowledged resources and aids. The submitted written version corresponds to the version on the electronic storage medium. I hereby declare that I have not previously applied or pursued for a doctorate (Ph.D. studies).

Date: ______________________________

Signature: ___________________________

(Joana Pereira)