Doctoral Thesis

Structural information on secondary transporter proteins from distance measurements between spin labels

Author(s):
Guérin, Kamila

Publication Date:
2016

Permanent Link:
https://doi.org/10.3929/ethz-a-010788458

Rights / License:
In Copyright - Non-Commercial Use Permitted
Structural information on secondary transporter proteins from distance measurements between spin labels

A thesis submitted to attain the degree of
DOCTOR OF SCIENCES of ETH ZURICH
(Dr. Sc. ETH Zurich)

presented by

Kamila Guérin
M.Sc. Jagiellonian University, Krakow, Poland

born on 03.07.1986
citizen of Poland

accepted on the recommendation of
Prof. Dr. Gunnar Jeschke, examiner
Prof. Dr. Roland Riek, co-examiner
Dr. Yevhen Polyhach, co-examiner

2016
# Contents

Contents

Abstract

Zusammenfassung

Symbols and abbreviations

1 Introduction
   1.1 Secondary active transporters
       1.1.1 Insight into structure and transport mechanism
       1.1.2 Proline permease
       1.1.3 Importance of extracellular loop
   1.2 Protein architecture modelling with sparse distance constraints
   1.3 DEER measurements on membrane proteins
   1.4 Aim of the thesis

2 Theory
   2.1 Spin Hamiltonian
       2.1.1 Electron Zeeman interaction
       2.1.2 Nuclear Zeeman interaction
       2.1.3 Hyperfine interaction
       2.1.4 Electron dipole-dipole interaction
   2.2 DEER experiment
       2.2.1 Hahn echo
       2.2.2 Four-pulse DEER sequence
   2.3 Distance geometry
       2.3.1 Bounds on distances and triangle bound smoothing
       2.3.2 Embedding to Cartesian space and removal of bound violations

3 Materials and methods
   3.1 Modelling approach
       3.1.1 Improvement of computational efficiency
<table>
<thead>
<tr>
<th>Contents</th>
<th>iv</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.2 Sample preparation</td>
<td>33</td>
</tr>
<tr>
<td>3.2.1 Mutant choice</td>
<td>33</td>
</tr>
<tr>
<td>3.2.2 Preparation protocol</td>
<td>34</td>
</tr>
<tr>
<td>3.3 DEER measurements</td>
<td>35</td>
</tr>
<tr>
<td>3.3.1 X band</td>
<td>35</td>
</tr>
<tr>
<td>3.3.2 Q band</td>
<td>35</td>
</tr>
<tr>
<td>3.4 Data evaluation</td>
<td>36</td>
</tr>
<tr>
<td>3.5 Simulating distances with MMM</td>
<td>36</td>
</tr>
<tr>
<td>3.5.1 Helix ends definition</td>
<td>36</td>
</tr>
<tr>
<td>4 Results and discussion</td>
<td>37</td>
</tr>
<tr>
<td>4.1 Distance measurements of PutP for modelling purposes</td>
<td>37</td>
</tr>
<tr>
<td>4.1.1 List of measurements</td>
<td>37</td>
</tr>
<tr>
<td>4.2 Definition of helix ends</td>
<td>40</td>
</tr>
<tr>
<td>4.2.1 Outcome of different prediction programs</td>
<td>40</td>
</tr>
<tr>
<td>4.2.1.1 vSGLT</td>
<td>40</td>
</tr>
<tr>
<td>4.2.1.2 Rho</td>
<td>42</td>
</tr>
<tr>
<td>4.2.1.3 Modelling templates</td>
<td>42</td>
</tr>
<tr>
<td>4.2.2 Comparison of vSGLT structures</td>
<td>42</td>
</tr>
<tr>
<td>4.3 Modelling of protein architecture</td>
<td>43</td>
</tr>
<tr>
<td>4.3.1 Sparse distance constraints</td>
<td>43</td>
</tr>
<tr>
<td>4.3.2 Template data inclusion and its influence</td>
<td>44</td>
</tr>
<tr>
<td>4.3.2.1 Decoy models</td>
<td>46</td>
</tr>
<tr>
<td>4.3.2.2 Positioning of non-core helices</td>
<td>46</td>
</tr>
<tr>
<td>4.3.3 Tightness parameter and its meaning</td>
<td>46</td>
</tr>
<tr>
<td>4.3.3.1 Combining template constraints and reduced $c$</td>
<td>49</td>
</tr>
<tr>
<td>4.3.4 Including labels into modelling</td>
<td>50</td>
</tr>
<tr>
<td>4.3.5 Pre- and post-selection of the clustered structures</td>
<td>56</td>
</tr>
<tr>
<td>4.3.5.1 Reduction of the ensemble based on a geometrical criterion of selection</td>
<td>56</td>
</tr>
<tr>
<td>4.3.5.1.1 Swapped helices</td>
<td>59</td>
</tr>
<tr>
<td>4.3.5.2 EPR pre-selection</td>
<td>62</td>
</tr>
<tr>
<td>4.3.5.3 Pre-selection based on the simulated distance distribution</td>
<td>63</td>
</tr>
<tr>
<td>4.3.6 Penalty functions</td>
<td>63</td>
</tr>
<tr>
<td>4.3.7 Placing non-core helices</td>
<td>66</td>
</tr>
<tr>
<td>4.3.7.1 Modelling with different templates</td>
<td>74</td>
</tr>
<tr>
<td>4.3.7.2 Placement of non-core helices in PutP</td>
<td>76</td>
</tr>
<tr>
<td>4.4 Extracellular loop 4 of PutP</td>
<td>81</td>
</tr>
<tr>
<td>4.4.1 Following eL4 during Na(^+)/proline transport</td>
<td>81</td>
</tr>
<tr>
<td>4.4.2 Role of Glu311 in PutP</td>
<td>85</td>
</tr>
<tr>
<td>5 Summary and outlook</td>
<td>89</td>
</tr>
<tr>
<td>Contents</td>
<td></td>
</tr>
<tr>
<td>---------------------------------</td>
<td>-----</td>
</tr>
<tr>
<td><strong>A Geometric algebra</strong></td>
<td>93</td>
</tr>
<tr>
<td>A.1 Mathematical basis</td>
<td>93</td>
</tr>
<tr>
<td>A.2 Invariant theory</td>
<td>96</td>
</tr>
<tr>
<td><strong>B Results</strong></td>
<td>99</td>
</tr>
<tr>
<td>B.1 Helix/loop ends definition</td>
<td>99</td>
</tr>
<tr>
<td>B.1.1 LeuT</td>
<td>99</td>
</tr>
<tr>
<td>B.1.2 Mhp1</td>
<td>101</td>
</tr>
<tr>
<td>B.2 DEER data analysis for modelling purposes</td>
<td>102</td>
</tr>
<tr>
<td>B.3 Constraint list</td>
<td>109</td>
</tr>
<tr>
<td>B.4 Modelling results</td>
<td>110</td>
</tr>
<tr>
<td>B.4.1 vSGLT</td>
<td>110</td>
</tr>
<tr>
<td>B.4.1.1 Placing non-core helices</td>
<td>114</td>
</tr>
<tr>
<td>B.4.2 Rho</td>
<td>120</td>
</tr>
<tr>
<td>B.4.3 Ambiguities in placing non-core helix I of PutP</td>
<td>133</td>
</tr>
<tr>
<td>B.5 eL4 functional studies</td>
<td>134</td>
</tr>
<tr>
<td><strong>Bibliography</strong></td>
<td>137</td>
</tr>
<tr>
<td><strong>Acknowledgements</strong></td>
<td>149</td>
</tr>
<tr>
<td><strong>Curriculum Vitae</strong></td>
<td>151</td>
</tr>
</tbody>
</table>
Abstract

The classically used methods of structural biology like X-ray diffraction or NMR may have serious limitations when applied to membrane proteins. Among the factors responsible for that are big protein size, an often large conformational flexibility and the necessity to use detergent for crystallization. Site-directed spin labelling combined with pulse EPR measurements and computational methodology offer an independent, complementary approach to study structure and function of such systems.

For membrane proteins, distance measurements by DEER are favourable as the range of available distances is between 1.5 and 6 nm, which is of the order of their sizes. Electron spins coming from the labels introduced pairwise into the protein undergo manipulation via a four-pulse sequence which separates the dipolar interaction between electrons from other interactions. The strength of this interaction encodes the information about the distance distribution between the labels.

Secondary transporters are the proteins translocating a substrate across the biological membranes against its concentration gradient thanks to the energy released by the coupled transport of ions or another substrate along their respective concentration gradient. Although the variety of these transporters is very broad, biochemical and structural analysis indicated that they share an analogous transport mechanism as well as some structural features despite low primary sequence similarity. Functions of the members are as broad as their diversity, ranging from dietary glucose and galactose absorption to the drug assimilation mechanism as some of them are primary targets of medicinal and abused drugs. Our focus was the proline permease PutP, engaged among others in adaptation to the osmotic stress and in virulence of bacteria.

A modelling approach was developed to model the structure of PutP based on sparse distance constraints. The methodology was tested with simulated constraints on proteins with known structures, which were different either in the size or in the degree of deviation of the helical shape from ideal. On the example of bovine rhodopsin with seven relatively straight transmembrane helices, our method was shown to predict the
fold successfully when distances for all pairs of helix ends are known. The proper fold is still obtained when the distances are known only between spin labels attached to the helix ends. In the case of the sodium galactose transporter vSGLT, with a bigger number of transmembrane helices and bigger deviation from ideality, our approach worked as well, although with higher uncertainty, provided that all distances for helix end pairs are known. Uncertainty stemming from length and flexibility of the label linker is too large to find the correct fold with simulated label-to-label distances for vSGLT.

We applied our procedure to the sodium proline secondary symporter PutP, whose structure is unknown. As the number of available distance constraints (75) is not sufficient to uniquely position 13 transmembrane helices in space, similarity to other proteins that share presumably the same fold was included by using a template for a ten-helix core during calculations. Although the template information dominates the obtained fold of the core, our approach allowed to place two of the three non-core helices of PutP with respect to the protein core, based solely on the DEER distance data.

Additionally, we addressed the problem of the alternating access mechanism in PutP. A subset of measurements was performed in order to investigate the relevance of the extracellular loop 4 in the transport cycle, likewise the influence of substitution of two significant residues on the protein function. Though due to biochemical difficulties, the DEER experiment did not deliver direct information concerning the transport mechanism, it clearly proved the importance of two residues, Glu311 and Ala404, in stabilizing the resting state of the protein.
Zusammenfassung

Klassische Methoden der Strukturbiologie wie die Röntgenkristallographie und Kernspinresonanzspektroskopie (NMR) stossen für Membranproteine oft an frechen. Verantwortlich sind das hohe Molekulargewicht dieser Proteine, ihre oft grosse Flexibilität und die Notwendigkeit, zur Kristallisation Detergenzien zu benutzen. Die ortsspezifische Spinmarkierung in Kombination mit gepulster Elektronen-paramagnetischer Resonanz-Spektroskopie (EPR) und Modellierungsmethoden bietet ein unabhängiges, komplementäres Vorgehen für die Untersuchung der Struktur und Funktion von solchen Systemen.


Eine neue Methode ist entwickelt worden, um die PutP Struktur mittels einer kleinen Zahl von Abstandsrestriktionen zu modellieren. Diese Methodik ist an Proteinen mit
Zusammenfassung


Symbols and abbreviations

Symbols

$A$  set of atom pairs
$A$  hyperfine tensor
$A_{\text{eff}}$  effective hyperfine coupling
$a_{iso}$  isotropic hyperfine coupling value
$B_0$  static magnetic field vector
$B(t)$  background function
$D$  matrix of squared distance in all atom pairs in the system
$D_{DD}$  dipolar coupling tensor
$d_{ij}$  distance between atoms $i$ and $j$
$D_{ij}$  squared distance between atoms $i$ and $j$, $D_{ij} = d_{ij}^2$
$\Delta \tilde{\mu}_{H^+}$  proton gradient across the membrane
$\Delta \tilde{\mu}_{Na^+}$  sodium gradient across the membrane
$e$  elementary charge, $e = 1.602176 \cdot 10^{-19}$ C
$\hat{e}_i$  elementary vector
$E_{DD}$  dipolar energy
$E_{EZ}$  electron Zeeman energy
$F$  Faraday constant, $F = 9.6485 \times 10^4 \frac{C}{mol}$
$F(t)$  form factor (time domain data from intramolecular interactions)
$g$  $g$ tensor
$g_e$  g value of a free electron, $g_e = 2.00231930$
$g_{\text{eff}}$  effective g value
$g_n$  g value of a nucleus
$\gamma_n$  gyromagnetic ratio of a nucleus
$\hbar$  quantum of action for angular frequencies, $\hbar = 1.054572 \cdot 10^{-34}$ J-s
$\hat{H}_{DD}$  dipolar Hamiltonian
$\hat{H}_{EZ}$  electron Zeeman Hamiltonian
$\hat{H}_{HF}$  hyperfine Hamiltonian
$\hat{H}_{HF,\text{aniso}}$  anisotropic part of the hyperfine Hamiltonian
$\hat{H}_{HF,\text{iso}}$  isotropic part of the hyperfine Hamiltonian
Symbols and abbreviations

\( \hat{H}_{NZ} \) nuclear Zeeman Hamiltonian
\( \hat{I} \) nuclear spin vector operator
\( l_{ij} \) lower bound put on a distance between atom \( i \) and \( j \)
\( \vec{M} \) magnetization vector
\( m_e \) mass of electron, \( m_e = 9.109382 \cdot 10^{-31} \) kg
\( m_i \) mass of point \( i \)
\( m_I \) nuclear magnetic quantum number
\( m_S \) electron magnetic quantum number, \( m_S = \pm 1/2 \)
\( \mu_0 \) vacuum permeability
\( \mu_B \) Bohr magneton, \( \mu_B = 9.27400899(37) \cdot 10^{-24} \) JT\(^{-1} \)
\( \mu_e \) magnetic moment of a free electron
\( \mu_n \) magnetic moment of a nucleus, \( \mu_n = 5.05078317(20) \cdot 10^{-27} \) JT\(^{-1} \)
\( \omega_{DD} \) dipolar frequency
\( \omega_I \) nuclear Zeeman frequency
\( P_{\text{angle}} \) penalty function for poor packing angle of the helices
\( P_{\text{apv}} \) penalty function for improper planar arrangement of all helix endpoints
\( P_{\text{compact}} \) penalty function for poor compactness of the bundle
\( P_{\text{contact}} \) penalty function for poor contact of the helices
\( P_{\text{dist}} \) penalty function for improper distances between helix endpoints
\( P_{\text{mpv}} \) penalty function for improper planar arrangement of helix midpoints
\( P_{\text{pact}} \) penalty function for poor packing of the helices
\( P_{\text{plv}} \) penalty function for improper planar arrangement of helix end planes
\( P_{\text{vec}} \) total penalty function of the ensemble
\( P_{\text{vdW}} \) penalty function for violating van der Waals repulsion between the helices
\( \phi \) polar angle characterizing the direction of \( \vec{B}_0 \) in PAS
\( \vec{r} \) vector connecting electron and nuclear spin
\( R \) gas constant, \( R = 8.3145 \frac{J}{K\cdot mol} \)
\( \vec{r}_{12} \) interspin distance vector between electron spins 1 and 2
\( S \) electron spin, \( S = 1/2 \)
\( \hat{S} \) electron spin vector operator
\( \hat{S}^+ \) rising spin operator
\( \hat{S}^- \) lowering spin operator
\( T \) temperature
\( T_{DD} \) dipole-dipole coupling tensor between the electron and nuclear spin
\( \theta \) polar angle characterizing the direction of \( \vec{B}_0 \) in PAS
\( \theta_{12} \) angle between magnetic moments and the interspin distance vector \( \vec{r}_{12} \) for magnetic moments aligned with the magnetic field
\( \tau \) interpulse delay
\( l_{ij} \) upper bound put on a distance between atom \( i \) and \( j \)
\( V(t) \) primary time-domain data of a pulse EPR experiment
\( x_i \) position of atom \( i \) in the molecule
### Abbreviations

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>AdiC</td>
<td>arginine/agmatine antiporter</td>
</tr>
<tr>
<td>APC</td>
<td>amino acid-polyamine-organocation transporter</td>
</tr>
<tr>
<td>ApcT</td>
<td>H⁺/amino acid symporter</td>
</tr>
<tr>
<td>BAT</td>
<td>biogenic amine transporter</td>
</tr>
<tr>
<td>BetP</td>
<td>Na⁺/betaine-glycine symporter</td>
</tr>
<tr>
<td>BCCT</td>
<td>betaine-choline-carnitine transporter</td>
</tr>
<tr>
<td>CaiT</td>
<td>L-carnitine/γ-butyrobetaine antiporter</td>
</tr>
<tr>
<td>Cc</td>
<td>empty carrier inaccessible from any side</td>
</tr>
<tr>
<td>Ce</td>
<td>carrier open to external side of the membrane</td>
</tr>
<tr>
<td>CSc</td>
<td>occluded carrier in transitional state</td>
</tr>
<tr>
<td>CSe</td>
<td>carrier loaded with the substrate open to the external side of the membrane</td>
</tr>
<tr>
<td>CSec</td>
<td>occluded carrier loaded with the substrate and closed from the external side</td>
</tr>
<tr>
<td>Ci</td>
<td>carrier open to the internal side of the membrane</td>
</tr>
<tr>
<td>CSi</td>
<td>carrier loaded with the substrate open to the internal side of the membrane</td>
</tr>
<tr>
<td>CSec</td>
<td>occluded carrier loaded with the substrate and closed from the internal side</td>
</tr>
<tr>
<td>cTMD</td>
<td>transmembrane domain in the protein core</td>
</tr>
<tr>
<td>DAT</td>
<td>Na⁺/dopamine symporter</td>
</tr>
<tr>
<td>DDM</td>
<td>n-decyl-β-D-maltopyranoside</td>
</tr>
<tr>
<td>DEER</td>
<td>double electron electron resonance</td>
</tr>
<tr>
<td>d8-Gly</td>
<td>deuterated glycerol</td>
</tr>
<tr>
<td>eL4</td>
<td>extracellular loop 4</td>
</tr>
<tr>
<td>EPR</td>
<td>electron paramagnetic resonance</td>
</tr>
<tr>
<td>FID</td>
<td>free induction decay</td>
</tr>
<tr>
<td>FRET</td>
<td>Förster resonance energy transfer</td>
</tr>
<tr>
<td>GltP</td>
<td>glycerol-3-phosphate transporter</td>
</tr>
<tr>
<td>LacY</td>
<td>lactose permease</td>
</tr>
<tr>
<td>LeuBAT</td>
<td>biogenic Na⁺/leucine symporter</td>
</tr>
<tr>
<td>LeuT</td>
<td>Na⁺/leucine symporter</td>
</tr>
<tr>
<td>Mhp1</td>
<td>Na⁺/hydantoin symporter</td>
</tr>
<tr>
<td>MTSL</td>
<td>S-(1-oxyl-2,2,5,5-tetramethyl-2,5-dihydro-1H-pyrrol-3-yl)methyl methanesulfonothioate, spin label</td>
</tr>
<tr>
<td>mw</td>
<td>microwave</td>
</tr>
<tr>
<td>NCS-1</td>
<td>nucleobase/cation symport 1</td>
</tr>
<tr>
<td>ncTMD</td>
<td>transmembrane domain outside the protein core</td>
</tr>
<tr>
<td>NMR</td>
<td>nuclear magnetic resonance</td>
</tr>
<tr>
<td>Symbol</td>
<td>Abbreviation</td>
</tr>
<tr>
<td>--------</td>
<td>---------------------------------------</td>
</tr>
<tr>
<td>NSS</td>
<td>neurotransmitter/solute symporter</td>
</tr>
<tr>
<td>PAS</td>
<td>principal axes system</td>
</tr>
<tr>
<td>PCR</td>
<td>polymerase chain reaction</td>
</tr>
<tr>
<td>pmf</td>
<td>proton motive force</td>
</tr>
<tr>
<td>PutP</td>
<td>Na(^+)/proline symporter</td>
</tr>
<tr>
<td>Rho</td>
<td>rhodopsin</td>
</tr>
<tr>
<td>rmsd</td>
<td>root mean square deviation</td>
</tr>
<tr>
<td>R1</td>
<td>denotation of MTSL</td>
</tr>
<tr>
<td>SDSL</td>
<td>site directed spin labelling</td>
</tr>
<tr>
<td>smf</td>
<td>sodium motive force</td>
</tr>
<tr>
<td>SSS</td>
<td>sodium/solute symporter</td>
</tr>
<tr>
<td>TMD</td>
<td>transmembrane domain</td>
</tr>
<tr>
<td>vSGLT</td>
<td>Na(^+)/glucose symporter</td>
</tr>
</tbody>
</table>
Chapter 1

Introduction

The transport of substances across biological membranes is necessary for cell functioning, its metabolism, and signaling. Free diffusion through the lipid bilayer is possible only for small molecules whereas all solutes such as organic acids, peptides and sugars have to be transported either by a passive or by an active way.

Active transport of a solute against its concentration gradient requires input of energy. Depending on its source, three types of transport can be distinguished [1]. Energy-driven pumps and coupled-carriers are the most common. Whilst energy driven pumps use directly the energy provided by light, redox reactions, or ATP hydrolysis, they are called primary active transporters. Coupled carriers are secondary active transporters and use the electrochemical gradient of ions or a solute to drive energetically unfavourable substrate transport. Depending on transport directions of both solutes, uniport, symport, or antiport can be distinguished.

1.1 Secondary active transporters

Secondary transporters couple the translocation of a substrate across the membrane to the energetically favourable flow of ions or a second substrate down their electrochemical gradients. The range of transported substrates is very broad, e. g. sugars, amino acids, ions, nucleosides, organic acids, vitamins, neurotransmitters, peptides, tRNAs and others [2].

The electrochemical proton gradient $\Delta \tilde{u}$ across the membrane is the most important energy source for secondary transporters in prokaryotes [3]. Primary proton pumps are mostly responsible for maintaining this gradient. Translocation of protons is a source of proton motive force $pmf$:

$$ pmf = \frac{\Delta \tilde{u}_{H^+}}{F} = \frac{2.3RT}{F} \Delta pH + \Delta \psi $$

where $R$ is the gas constant ($R = 8.3145 \frac{J}{K\text{mol}}$), $F$ the Faraday constant ($F = 9.6485 \times 10^4 \frac{C}{\text{mol}}$) and $T$ the temperature in [K].
The energy stored in the \( pmf \) can be used to drive the energetically unfavourable transport of other substrates according to the chemiosmotic theory [4]. Nevertheless, more and more transporters are uncovered to rely on sodium as a coupling ion [5, 6]. In equation 1.1, \( \Delta \text{pH} \) is then substituted by \( \Delta \text{pNa}^+ \) giving a contribution to sodium motive force \( \text{smf} \). These secondary transporters are thus dependent on another primary system translocating sodium across the membrane, such as the firstly discovered oxaloacetate decarboxylase [7], ATPase [8] or many others. Additionally, \( \text{Na}^+/\text{H}^+ \) antiporters can generate \( \text{smf} \) as well by converting \( pmf \) to \( smf \) [9].

### 1.1.1 Insight into structure and transport mechanism

Although more than 200 distinct transporter families can be distinguished, biochemical, structural and kinetic analyses suggests that all secondary transporters work on the basis of the same alternating access mechanism [10]. The model assumes that the substrate binding site is accessible either from one or the other side of the membrane at any instant. It was suggested by Mitchell [11] that functioning of such transporters should be described rather as moving barrier than moving carrier - the movement of the substrate to the binding site is not decisive, rather the movement of extra- and intracellular gates and thus alternated accessibility of this site defines the transport mechanism.

Alternating access can be schematically depicted as in Figure 1.1. An carrier opened to the external site (Ce) of the membrane is loaded by ion(s) and then the substrate. Conformational change of some protein residues closes the access to the binding site from the outside (CSec) and so the occlusion occurs (called as well ”thin gate” closure). The protein then switches the conformation to inside-oriented (CSic) passing through the transitional, fully-occluded state CSc. When a ”thin gate” opens to the inside, ion(s) and the substrate can be released (CSI). The empty transporter is then in an inside-opened state (Ci) and can switch between inside- and outside-opened conformation through a transitional state of empty carrier not accessible from any site (Cc).

![Figure 1.1: Conformations during the substrate transport in secondary transporters.](image)

However, deep understanding of the transport mechanism was for a long time hindered by lack of high-resolution 3D structures in different transition states. Advances in structure determination methods allowed for successful resolution of transmembrane protein structures such as lactose permease LacY [12] or glycerol-3-phosphate transporter GlpT [13]. Subsequently appearing crystal structures of some secondary transporters belonging to different families resolved at atomic level revealed, despite very low similarity at the primary sequence level, a similar topological motif formed by ten
transmembrane domains (TMDs) out of which the first five are related to the remaining five TMDs by a pseudo-two-fold rotation axis in the membrane plane. Existence of these inverted 5-helix-motives in a protein structure is further called "LeuT fold". In LeuT the first two helices of each repeat (helix 1, 2, 6 and 7) meet in space to form a four-helix bundle. The third and fourth helix of every repeat (helix 3, 4, 8 and 9) form another 4-helix-domain called scaffold [14] or hash [15] which together with accompanying "arms" (helix 5 and 10) surrounds the bundle. Table 1.1 collects the secondary transporter structures with the LeuT-fold crystallized up to now.

Table 1.1: Crystal structures of secondary transporters with LeuT-fold. APC - amino acid-polyamine-organocation transporter, BAT - biogenic amine transporter, BCCT - betaine-choline-carnitine transporter, NCS-1 - nucleobase/cation symport 1, NSS - neurotransmitter/solute symporter, SSS - sodium/solute symporter.

<table>
<thead>
<tr>
<th>transporter/family</th>
<th>description</th>
<th>conformation</th>
<th>PDB index</th>
</tr>
</thead>
<tbody>
<tr>
<td>AdiC/APC</td>
<td>arginine/agmatine antiport</td>
<td>Ce</td>
<td>3NCY [16]</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>3LRB [17]</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>3LRC [17]</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>3OB6 [18]</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>3LIL [19]</td>
</tr>
<tr>
<td>ApcT/APC</td>
<td>H⁺/amino acid symport</td>
<td>Cie</td>
<td>3GIA [20]</td>
</tr>
<tr>
<td>BetP/BCCT</td>
<td>Na⁺/betaine-glycine symport</td>
<td>CScic</td>
<td>2WIT [21]</td>
</tr>
<tr>
<td>CaiT/BCCT</td>
<td>L-carnitine/γ-butyrobetaine antiport</td>
<td>Csi</td>
<td>3HFX [22]</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>2WSX [23]</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>2WSW [23]</td>
</tr>
<tr>
<td>DAT/NSS</td>
<td>Na⁺/dopamine symport</td>
<td>Ce</td>
<td>4M48 [24]</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>4XNU [25]</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>4XPA [26]</td>
</tr>
<tr>
<td>LeuBAT/BAT</td>
<td>Na⁺/leucine symport</td>
<td>Ce</td>
<td>3TT1 [28]</td>
</tr>
<tr>
<td>LeuT/NSS</td>
<td>Na⁺/leucine symport</td>
<td>Ce</td>
<td>3TT1 [28]</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>3F3A [29]</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>3QS4 [30]</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>2A65 [5]</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>2GEI [31]</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>2QJU [32]</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>3F48 [29]</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>3GWV [33]</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>3USG [34]</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>3TT3 [28]</td>
</tr>
<tr>
<td></td>
<td>S2 binding site</td>
<td>S2 binding site</td>
<td>3GJC [35]</td>
</tr>
<tr>
<td></td>
<td>no S2</td>
<td>no S2</td>
<td>4FXZ [36]</td>
</tr>
<tr>
<td>Mhp1/NCS1</td>
<td>Na⁺/hydantoin symport</td>
<td>Ce</td>
<td>2JLN [37]</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>2JLO [37]</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>2X79 [15]</td>
</tr>
<tr>
<td>vSGLT/SSS</td>
<td>Na⁺/glucose symport</td>
<td>Csic</td>
<td>3DH4 [38]</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>2XQ2 [39]</td>
</tr>
</tbody>
</table>
A number of comparative studies on alternating access mechanism in secondary transporters [40–43] as well as modelling attempts [14, 39] were undertaken to explain exact transport proceedings as both a gating-pore-type behaviour [44–46] and a rocking-bundle mechanism [14, 15] seemed possible. Whilst a gating-pore model emphasizes flexing of the first two helices in every repeat and thus controlling substrate binding site accessibility [44–46], a rocking-bundle model assumes that a 4-helix-bundle acts like a rigid body and rocks against the hash-motif [14, 15]. Later studies, conducted mainly on Mhp1 structures, which were solved in two “terminal” states of transport: Ce - empty carrier externally open (#2JLO) and Ci - empty carrier internally open (#2X79), tended to favour a rocking-bundle explanation [43, 47]. Nevertheless, based on simulations of an hSERT homology model [47] and covariance matrix analysis of elastic network models [43], hash and bundle appear to move not exactly like rigid bodies, as at least one TMD undergoes some reorientations as well as flexing and twisting of TMDs (fifth helix of every repeat) in order to occlude/open the access to the central binding site [43, 47].

Most of the studies of alternating access in secondary transporters are based on crystal structures captured in different conformational states. Beside the fact that many structures in relevant conformations are missing, one has to keep in mind that crystallization demands often physiologically unrealistic conditions (pH, harsh detergents, absence of lipid bilayer, high substrate concentration). It should not be then surprising that unnatural conformations are captured [35, 48] which may lead to inconsistencies in interpretation of the ion/substrate translocation described above. Thus, information gained on systems in more biologically relevant conditions is really appreciated. To the ”go-for” techniques belong, f. e. Förster resonance energy transfer (FRET) [46] and electron paramagnetic resonance (EPR) [49–54], where structural transitions can be observed upon addition of ions and substrate. Unfortunately, these methods offer only a small number of distance constraints due to high experimental demands (every individual constraint requires protein mutation, labelling and measurement). Because of this sparsity, modelling providing insight into transporter transitions can be then performed only under conditions of the reduction of the number of degrees of freedom. One of the ways is describing relevant protein conformational changes by a few normal modes [55], which are well predicted by coarse-grained elastic network models [56]. It was proven for some soluble proteins that the Zheng-Brooks algorithm [57] employing normal modes analysis performs reasonably well in predicting protein transitions. A small number of long-range distance constraints is demanded to specify forces in the analyzed system hence indicating EPR as an experimental method of choice. An algorithm adapted to EPR-derived distance constraints [58] was shown to predict the direction of Mhp1 conformational changes occurring during the transport correctly [43], though the exact amplitude and directions of movement might be missed for the motions possibly uncoupled from relative movements of hash and bundle (different loop conformations uncorrelated with analogous changes for connected helices). Also movement of the ”arms” (helix 5 and 10) that is generally speaking less collective, was reproduced worse [43]. However, the directions of the transport process are assigned properly. The same algorithm revealed conformational change of hyperpolarization-activated nucleotide-gated ion channels by binding of the cyclic nucleotide [59].
1.1.2 Proline permease

PutP is a Na\(^+\)/proline symporter, a protein belonging to the SSS (solute/sodium symporter) family. It consists of 502 residues (54.3 kDa) that are packed, according to gene fusion analyses, Cys accessibility studies, site-specific proteolysis and SDSL EPR measurements, in 13 transmembrane helices \[6, 60\]. Helices 2-11 are supposed to form a LeuT-motif. PutP utilizes \(smf\) given by

\[
smf = \frac{\Delta \tilde{\mu}_{Na^+}}{F} = -\frac{2.3RT}{F} \Delta pNa + \Delta \psi \tag{1.2}
\]

to drive proline transport uphill its gradient. Though no crystal stucture was obtained so far, a homology model was generated based on structure alignment with vSGLT - the closest crystallized relative of the same family \[61\]. Afterwards, additional constraints were included in order to improve the model \[62, 63\].

Found in bacteria and archaea \[64\], proline-specific SSS representants are engaged in amino acids uptake or proline transport during adaptation of the organism to the osmotic stress conditions \[65–67\]. Bacterial PutP plays an important role in human infectious diseases \[66, 68, 69\] and thus represents a target for the development of new drugs against pathogens \[70\].

1.1.3 Importance of extracellular loop 4

EPR studies on LeuT by D. Claxton et al. suggested that the external loop 4 (eL4) regulates the access of the ions and substrate to their binding site on the outer site of the membrane \[54\]. It was further confirmed by LeuT crystal structures in inward- and outward-facing conformations that the major rearrangement of transmembrane helices is accompanied by an inward movement of this loop \[28\]. Also the Mhp1 structure reports eL4 as a "thick gate" occluding the substrate binding site \[15\]. The term "thick gate" underlines a "massive" barrier to the central binding site \[41\], nevertheless it can be seamlessly omitted as it pertains to the main conformational change of the transporter (CSec \(\leftrightarrow\) Csic) \[42\].

The length of eL4 varies between 13 residues for Mhp1 \[15\] and 32 residues in vSGLT \[38\]. It consist of two domains - eL4a and eL4b. Helical eL4b is found in every structure with a LeuT-fold \[43\], whereas eL4a is less conserved and found only in some of them - LeuT \[5\], CaiT \[23\], ApcT \[20\] and DAT \[24\].

In PutP, substitution of chosen amino acids from eL4 influences the rate of transport substantially, thus suggesting importance of this loop \[62\]. At the same time it does not influence the substrate binding, meaning no direct ligand binding \[54, 62\]. However, eL4 stays in close contact with the substrate binding site, suggesting participation in mechanical closure of the gate from outside and transmission of this pulling force to open the transporter from the inside \[62\].
1.2 Protein architecture modelling with sparse distance constraints

Widely used methods of structure determination, such as X-ray crystallography and NMR meet difficulties in determining structures of specific proteins. For NMR, size of a protein is limiting, X-ray crystallography can only be conducted if crystals of sufficient size and quality can be obtained. Most problematic are proteins spanning the membrane as their crystallization is very often not possible. Interestingly, integral membrane proteins comprise around 30% of the proteosomes of living organisms and around 60% of pharmaceutically targeted proteins [71]. A lot of effort has been put to adjust techniques of structure biology to the need of determining membrane protein structures.

The search for membrane mimicking-systems not only suiting the investigated systems, but also dispersing cross-peaks well [72], allowed solution NMR for resolving structures with up to 9 transmembrane $\alpha$ helices [73] or for detecting conformational change in the 100 kD membrane transporter ClC-ec1 [74]. Crystallization procedures were verified and adjusted for transmembrane proteins [75], but success rate in obtaining highly-ordered structures is still disappointing. Fortunately, a tendency to accelerate is observed.

In the cases when none of the standard structure determination methods are helpful, development of other techniques based on distance data is very important. Distances could be determined by different means, f.e. chemical cross-linking, electron paramagnetic resonance (EPR), or Förster resonance energy transfer (FRET). Accompanying modelling of different types constrained by the experimental information is then capable to predict the structures. Anyway, one has to keep in mind that a set of distance constraints collected for the protein of interest will remain sparse, hence the structure will not be determined purely by the experimental information.

K. Sale et al. introduced Bundler [76, 77] - a two-step computational algorithm where, firstly, all possible bundles of the protein of interest are generated based on basic knowledge of membrane proteins. They are further verified and tested against the experimental distances [76]. Secondly, Monte Carlo simulated annealing is applied to every helix in order to minimize distance violations [77]. This method tested with simulated restraints on the Rhodopsin structure # 1F88 [78] gave promising results - $\text{C}_\alpha$ rmsd of 3.2 Å in comparison with the real crystal structure. The introduced algorithm is optimized for the template protein (Rho #1F88 [78]), so it is not surprising that it works well also for another proteins consisting of 7 helices. In addition, it was expected that Bundler could serve as a tool to model other classes of proteins as well [77].

RosettaEPR introduced by J. Meiler and co-workers is an Rosetta-based algorithm for de novo protein folding. A two-step process assumes low-resolution de novo folding followed by high-resolution refinement of the backbone and side chains [79]. An ensemble of structures is generated, the refinement allows for finding the lowest-energy models that have $\text{C}_\alpha$-rmsd from the native structure of 1.8 Å for T4-Lysozyme and 4.0 Å for $\alpha$A-Crystallin [79]. Unfortunatelly, Rosetta is limited to soluble proteins of maximally 150 amino acids [80].
Work on another approach to model membrane proteins was undertaken in the same group of J. Meiler. An algorithm based on Monte Carlo Metropolis was newly upgraded with EPR data inclusion (accessibility and distances) [81]. Models of a protein folded from smaller secondary structure elements were scored according to knowledge-based potentials [82] and additionally according to a potential describing amino acid exposure to an implicit membrane environment. EPR-restraining allowed for increasing the accuracy of de novo modelling by increasing the frequency of sampling the correct topology. The algorithm is capable to find a large fraction of correct conformations for proteins up to 215 amino acids [81].

Modeling Employing Limited Data (MELD) is an attempt to model a protein structure restrained by semireliable experimental data [83]. An algorithm based on Bayesian inference obeys strictly statistical mechanics rules and provides proper Boltzmann populations. The AMBER forcefield is used to generate protein structure which is then confronted with experimental information of unknown quality. The approach generates an minimum-free-energy ensemble of structures. Within such an ensemble, the αA-Crystallin structure with C_backbone-rmsd of 1.3 Å from the native is found. In comparison, the best T4-Lysozyme structure out of the generated ensemble has C_backbone-rmsd of 2.6 Å. One of the three most populous clusters has C_backbone-rmsd of 2.8 Å vs native for αA-Crystallin and 3.6 Å vs native for T4-Lysozyme, which is still sufficient to correctly predict the general protein fold. Quality of the structures is thus comparable with those of RosettaEPR with the advantage that the model is giving physical, Boltzmann populations.

Unfortunately, none of described approaches are suitable for PutP, a membrane protein consisting of 502 amino acids. Homology modelling, which became an important tool to predict protein structures, has been employed. At first PutP was modelled based on the vSGLT crystal structure #3DH4 [61]. Nevertheless, sequence identity between these two proteins is low, thus sequence alignment is not obvious. Moreover, for one transmembrane helix in #3DH4 residue numbers are not assigned in the crystal structure. Hence, the #2XQ2 structure, which is more complete, was used as a template to generate a new model by means of Modeler [84]. During this process experimental constraints were applied to improve the obtained model [62, 63].

1.3 DEER measurements on membrane proteins

Double Electron-Electron Resonance (DEER, known also as PELDOR - Pulsed EElectron-electron DOuble Resonance) is a useful tool to attempt the elucidation of structure and transport mechanism of secondary transporter proteins. These measurements have an advantage of high biological relevance, as samples are prepared in membrane-mimicking systems like liposomes or micelles, not demanding harsh conditions or crystallization. Nevertheless, most proteins, except some metaloproteins, are not directly suitable for EPR because of being "silent", as they contain no paramagnetic residues. In order to make this technique applicable, paramagnetic species have to be engineered into the protein. This is done by means of site-directed spin-labelling [85, 86], where a chosen residue (or some of them) is substituted by Cys and a specific spin label is further
attached to it. Amongst a broad variety of available spin labels, nitroxides with spin $S = 1/2$ are the most common. Often Cys occurring naturally in the protein of interest have to be mutated out unless the corresponding residues are favourable and sufficient for labelling.

In a sample with paramagnetic centers, magnetic dipole moments are interacting with each other. Exchange and dipole-dipole coupling may occur and the interactions can manifest themselves as relaxation enhancement, unresolved and resolved splittings in the spectra. It was shown by A. Milov et al. [87] that dipolar interactions can be experimentally separated from others in an EPR experiment, thus enabling the extraction of the distance $r$ between the paramagnetic residues. Nevertheless, fast averaging of orientation in solution also averages the dipole-dipole coupling. Hence, only frozen or at least immobilized protein samples are accessible to this technique.

The four-pulse DEER experiment introduced by M. Pannier et al. [88] is capable of delivering distances ranging from 1.5 to 6 nm or, in exceptional cases, even up to 8 nm [89, 90]. An additional $\pi$ pulse, introduced into the original three-pulse experiment [91–93], allows to avoid dead-time artifacts and thus the experiment offers characterization of even broad distributions. Hence, also flexible systems can be taken into consideration. Together with site-directed spin-labelling, this technique is suitable for distance measurements in biomacromolecules [94, 95].

Thanks to the magnetic dipole moment $\mu_e$ of the electron being much larger than the one of nuclei $\mu_i$, the DEER method is much more sensitive than NMR. Hence, it can approach longer distances between interacting spins as well as faster molecular motions. It offers also the advantage of small sample volumes and low concentrations needed [96].

1.4 Aim of the thesis

This thesis is committed to studies of proline permease PutP, a representative of secondary transporters, aiming at elucidation of the structure and transport mechanism. The ability to gain structural information from distance measurements between spin labels was tested. A model for the relative arrangement of all transmembrane helices of PutP is sought. To that end, spin labels are attached to the helix ends and distances are measured for spin pairs of helix ends. An experimental set of such constraints was available for PutP. Work on this set has started in 2004, shortly after DEER had first been demonstrated on PutP to work for membrane proteins reconstituted into liposomes [49]. Work on a helix bundle model was started at a time when no related transporter had been crystallized. Soon it was realized that the structure of LeuT obtained in 2005 [5] was relevant for understanding PutP and with publication of the first vSGLT structure in 2008 [38] it became apparent that LeuT and vSGLT, and thus, most likely PutP, shared the same ten-helix core. At that time, a large set of constraints had already been assembled and the question arose how well structure could be modelled from such a constraint set. This set of constraints is sparse, it does not include all helix end pair distances in the bundle. The aim of the thesis was to establish what information could be safely derived from this set of distance constraints and whether a full set of such constraints could have
provided a reliable model of the architecture of the bundle of transmembrane helices. Addressed in our studies was also the problem of the commonly accepted alternating access mechanism [10] in secondary transporters. States of PutP without ligand, with sodium and with sodium and proline were investigated in order to give an insight in different protein conformations that are adapted by the protein in the transport mechanism. Especially, the role of extracellular loop 4 (eL4) in this process and its interactions with other regions in the core of the protein were examined.
Chapter 2

Theory

Electron paramagnetic resonance is a well established experimental method for which the theoretical background got widely documented [97–100].

EPR is very similar in concepts to NMR with the major difference of excitation and detection of electron spins instead of nuclear spins. The magnetic moment of a free electron

\[
\hat{\mu}_e = -g_e \frac{e \hbar}{2m_e} \hat{S} = -g_e \mu_B \hat{S}
\]

(2.1)

where \( g_e = 2.00231930 \) is the g value of a free electron, \( e = 1.602176 \cdot 10^{-19} \) C the elementary charge, \( \hbar = \hbar/2\pi = 1.054572 \cdot 10^{-34} \) J·s the quantum of action for angular frequencies, \( m_e = 9.109382 \cdot 10^{-31} \) kg the mass of electron, \( \mu_B = 9.27400899(37) \cdot 10^{-24} \) JT\(^{-1}\) Bohr magneton and \( S = 1/2 \) the electron spin. Only unpaired electrons can give a contribution to an EPR spectrum, unlike for NMR where every magnetic nucleus gives a signal. This fact is at the same time a big advantage and disadvantage of the method. Naturally occurring free electron spins are quite rare and instable, they pertain mainly to radicals and transition metal complexes. Protein or RNA studies require additional attachment of paramagnetic groups. But this feature offers, at the same time, insensitivity to the EPR-silent solvents or neighbouring molecules, thus the signal comprises mainly characteristic features and assignment becomes simpler.

In the following chapter only the basics needed to describe the investigated system are introduced. These are the interactions between electrons with spin \( S = 1/2 \) and the electron quantum numbers \( m_S = \pm 1/2 \) with the nitrogen nucleus within the nitroxide with spin \( I = 1 \) and the nuclear quantum numbers \( m_I = -1, 0, +1 \). Only the \( ^{14}\text{N} \) isotope is considered as its natural abundance is 99.636%.

Distance geometry is a mathematical approach to solve 3-dimensional problems in 2-dimensional space of distances and ”translate” them back to 3-dimensional space of coordinates. The method pioneered by G. Crippen and T. Havel [101] got widely used in the beginning of NMR structure determination as it offers prediction of a 3-dimensional structure based on a set of experimental distances.
2.1 Spin Hamiltonian

Interactions within the nitroxide labels attached to the protein of interest are described by the Hamiltonian

\[ \hat{H} = \hat{H}_{EZ} + \hat{H}_{NZ} + \hat{H}_{HF} + \hat{H}_{DD} \]  

(2.2)

where \( \hat{H}_{EZ} \) is the electron Zeeman Hamiltonian, \( \hat{H}_{NZ} \) the nuclear Zeeman Hamiltonian, \( \hat{H}_{HF} \) the hyperfine Hamiltonian and \( \hat{H}_{DD} \) the dipolar Hamiltonian.

2.1.1 Electron Zeeman interaction

In the external magnetic field, the splitting of degenerate electron energy levels \( m_S \) takes place (Figure 2.1). This phenomenon is called the electron Zeeman interaction and can be described by the Hamiltonian:

\[ \hat{H}_{EZ} = \frac{\mu_B}{\hbar} \vec{B}_0^T \, g \hat{S} \]  

(2.3)

where \( \vec{B}_0^T \) is the transpose of the static magnetic field vector \( \vec{B}_0 = (B_{0x}, B_{0y}, B_{0z}) \), \( g \) the \( g \) tensor and \( \hat{S} \) the electron spin vector operator \( \hat{S} = (\hat{S}_x, \hat{S}_y, \hat{S}_z) \).

If the \( g \) tensor is anisotropic, the effective field felt by the molecule depends on orientation of the molecule in the magnetic field. Thus the effective \( g \) value is described as

\[ g_{\text{eff}} = \sqrt{g_x^2 \sin^2 \theta \cos^2 \phi + g_y^2 \sin^2 \theta \sin^2 \phi + g_z^2 \cos^2 \theta} \]  

(2.4)

where \( \theta \) and \( \phi \) are polar angles characterizing the direction of \( \vec{B}_0 \) in the principal axes system.

This Hamiltonian corresponds to the following eigenvalues

\[ E_{EZ} = m_S g_{\text{eff}} \mu_B B_0 = \pm \frac{1}{2} g_{\text{eff}} \mu_B B_0 \]  

(2.5)

The electron Zeeman interaction is the dominant one in the high-field approximation.

2.1.2 Nuclear Zeeman interaction

Nuclear Zeeman interaction is described by the following Hamiltonian

\[ \hat{H}_{NZ} = -\frac{g_n \mu_n}{\hbar} \vec{B}_0^T \hat{I} \]  

(2.6)

where \( g_n \) is the \( g \) value of a nucleus and \( \mu_n = 5.05078317(20) \cdot 10^{-27} \) JT\(^{-1}\) its magnetic moment. \( \hat{I} \) stands for the nuclear spin operator. In the laboratory frame this expression simplifies to

\[ \hat{H}_{NZ} = -\gamma_n B_0 \hat{I} = \omega_1 \hat{I}_z \]  

(2.7)
Figure 2.1: The electron Zeeman interaction. The external magnetic field $B_0$ causes splitting of the degenerate energy levels of an electron with spin $S = 1/2$. The energy difference $\Delta E$ needed for the transition scales linearly with the magnetic field. By convention, the magnetic field $B_0$ is set along $z$-axis thus $g_{\text{eff}} = g_e$. $h = 6.62607004 \cdot 10^{-34}$ Js is the Planck constant, $\nu$ the frequency of microwave (mw) irradiation.

The nuclear Zeeman interaction is mostly negligible in EPR as it does not contribute to the frequencies of allowed electron spin transitions to first order. In the high-field approximation (W-band - 95 GHz and higher) $^{14}$N atoms contribute the energy of $E_{NZ} = m_I \omega_I$ to the spin states.

### 2.1.3 Hyperfine interaction

Hyperfine interaction involves the through-space dipole-dipole couplings between the magnetic moments of the electron and nuclear spins as well as the Fermi contact interaction

$$\hat{H}_{HF} = \sum_k \hat{S}^T \hat{A}_k \hat{I}_k$$

$k$ sums over all nuclei with spin $I_k > 0$, $\hat{I}_k$ are the nuclear spin vector operators and $\hat{A}_k$ the hyperfine tensors. In the high-field approximation hyperfine coupling contributes with the energy of $m_S m_I A_{\text{eff}}$. Hyperfine interactions lead to a splitting of the electron energy levels into $2I + 1$ sublevels. For $^{14}$N with spin $I = 1$, each $m_S$ level will be splitted into $2I + 1 = 3$ sublevels. EPR selection rules ($\Delta m_S = 1$ and $\Delta m_I = 0$) allow then for 3 transitions (Figure 2.2).

The Fermi contact interaction arises from non-zero probability of finding the spin of an unpaired electron in an $s$ orbital at the same point in the space as the nuclear spin. It causes an isotropic coupling

$$\hat{H}_{HF,\text{iso}} = a_{\text{iso}} \hat{S}^T \hat{I}$$

The dipole-dipole interaction between electron and nuclear spins gives rise to an anisotropic part of the Hamiltonian

$$\hat{H}_{HF,\text{aniso}} = \hat{S}^T T_{DD} \hat{I}_k = \frac{\mu_0}{4\pi \hbar} g_e \mu_B g_n \mu_n \left[ \frac{(3 \hat{S}^T \hat{r})(\hat{r} \hat{I})}{r^5} - \frac{\hat{S}^T \hat{I}^2}{r^3} \right]$$

where $\mu_0$ is the vacuum permeability, the $T_{DD}$ dipole-dipole coupling tensor between the electron and nuclear magnetic moments and $\hat{r}$ the vector connecting electron and
nuclear spin.
The anisotropic part of the interaction is dominant for $^{14}\text{N}$ as the unpaired electron is located mostly in an $p$-orbital.
Unlike electron and nuclear Zeeman interaction, hyperfine coupling is field independent. Measurements at different frequencies can be then a requisite to separate hyperfine from other field dependent interactions.

### 2.1.4 Electron dipole-dipole interaction

The dipole-dipole interaction describes the coupling between the magnetic moments of two electrons $\mu_1$ and $\mu_2$. The corresponding spin Hamiltonian is assumed as

$$
\hat{H}_{DD} = S_1^T D_{DD} S_2 = \frac{1}{4\pi} \frac{\mu_1 \mu_2}{r_{12}^3} \frac{\mu_0}{\hbar} g_1 g_2 \mu_B^2 \left( \hat{S}_1 \hat{S}_2 - \frac{3}{r_{12}^2} (\hat{S}_1 \hat{r}_{12})(\hat{S}_2 \hat{r}_{12}) \right)
$$

where $D_{DD}$ is the dipolar tensor, $\hat{S}_1$ and $\hat{S}_2$ are the spin operators of electron 1 and 2 respectively and $\hat{r}_{12}$ is the interspin distance vector.

For a nitroxide radical, the dipole-dipole interaction can be approximated as a coupling between the point dipoles for distances longer than 1.5 nm. Thus, it can be described in analogy to the classical dipole-dipole interaction. For most transition metals and rare earth ions, the two magnetic moments are not parallel to the magnetic field due to the strong g-anisotropy. The dipolar interaction energy is then given by

$$
E_{DD} = -\frac{\mu_0}{4\pi} \frac{\mu_1 \mu_2}{r_{12}^3} \left( 2 \cos \theta_1 \cos \theta_2 - \sin \theta_1 \sin \theta_2 \cos \phi \right)
$$

where $\theta_1$ and $\theta_2$ are the angles between magnetic moments $\vec{\mu}_1$ and $\vec{\mu}_2$ with respect to the interspin distance vector $\vec{r}_{12}$ and $\phi$ is the dihedral angle set by both magnetic moment vectors and the interspin distance vector.

When the magnetic moments are aligned with the external magnetic field $\vec{B}_0$, $\phi$ becomes zero and $\theta_1 = \theta_2 = \theta_{12}$ defines an angle between $\vec{r}_{12}$ and $\vec{B}_0$. Eq. 2.12 can be simplified
as follows

$$E_{DD} \approx \frac{\mu_0}{4\pi} \frac{1}{r_{12}^3} (1 - 3 \cos^2 \theta_{12})$$  \hspace{1cm} (2.13)$$

The spin Hamiltonian of the dipolar interaction can be rewritten in terms of the so called "dipolar alphabet" based on the ladder operators $\hat{S}^+ = \hat{S}_x + i\hat{S}_y$ and $\hat{S}^- = \hat{S}_x - i\hat{S}_y$

$$\hat{H}_{dd} = \frac{\mu_0}{4\pi \hbar} \frac{g_1 g_2 \mu_B^2}{r_{12}^3} [\hat{A} + \hat{B} + \hat{C} + \hat{D} + \hat{E} + \hat{F}]$$  \hspace{1cm} (2.14)$$

where the terms are defined as

$$\hat{A} = \hat{S}_{1z} \hat{S}_{2z} (1 - 3 \cos^2 \theta_{12})$$
$$\hat{B} = -\frac{1}{4} [\hat{S}_1^+ \hat{S}_2^- + \hat{S}_1^- \hat{S}_2^+] (1 - 3 \cos^2 \theta_{12})$$
$$\hat{C} = -\frac{3}{2} = [\hat{S}_1^+ \hat{S}_2^+ + \hat{S}_1^- \hat{S}_2^-] \cos \theta_{12} \sin \theta_{12} e^{-i\phi}$$
$$\hat{D} = -\frac{3}{2} = [\hat{S}_1^+ \hat{S}_2^- + \hat{S}_1^- \hat{S}_2^+] \cos \theta_{12} \sin \theta_{12} e^{+i\phi}$$
$$\hat{E} = -\frac{3}{4} \hat{S}_1^+ \hat{S}_2^+ \sin^2 \theta_{12} e^{-2i\phi}$$
$$\hat{F} = -\frac{3}{4} \hat{S}_1^- \hat{S}_2^- \sin^2 \theta_{12} e^{+2i\phi}$$

Term $\hat{A}$ commutes with $\hat{H}_{EZ}$. Hence, the energy of the electron state is influenced, but no mixing of the levels takes place. Term $\hat{B}$ is responsible for the zero-quantum transition (flip-flop). Rising or lowering operators, likewise in terms $\hat{C}$ and $\hat{D}$, describe single quantum transitions (spin flip) with $\Delta m_S = \pm 1$. Or when they appear for both considered spins, like in terms $\hat{E}$ and $\hat{F}$, they represent double quantum transitions (flip-flip process) with $\Delta m_S = \pm 2$.

In the case of strong Zeeman interaction, terms that are time-dependent in the frame rotating with the excitation frequency can be neglected. These are the terms containing the ladder operators. The term $\hat{B}$ can be neglected if the difference between the Zeeman frequencies of the two electron spins is much larger than the dipole-dipole interaction. In the DEER experiment, this difference is approximately the difference between pump and observer frequency, which is much larger than coupling at distances larger than 1.5 nm. Thus $\hat{H}_{DD}$ simplifies significantly

$$\hat{H}_{DD} = \frac{\mu_0}{4\pi \hbar} \frac{g_1 g_2 \mu_B^2}{r_{12}^3} \hat{A} \approx \omega_{12} \hat{S}_{1z} \hat{S}_{2z}$$  \hspace{1cm} (2.15)$$

and the dipolar evolution frequency is given

$$\omega_{DD} = \frac{\mu_0}{4\pi \hbar} \frac{g_1 g_2 \mu_B^2}{r_{12}^3} (1 - 3 \cos^2 \theta_{12})$$  \hspace{1cm} (2.16)$$
When the spin-spin vector is perpendicular to the external magnetic field

\[ \omega_{\perp} = \frac{1}{r^3} \frac{\mu_0}{4\pi\hbar} g_1 g_2 \mu_B^2 \] (2.17)

If the g-value is approximated as the g-value of the free electron \( g_1 = g_2 = g_e \), which obeys for organic radicals, the dipolar coupling can be calculated as

\[ \omega_{\perp} = 2\pi \frac{52.04[\text{MHz}]}{r^3[\text{nm}^3]} \] (2.18)

### 2.2 DEER experiment

The double electron-electron resonance experiment is a tool to obtain distances between the spin labels attached to the protein of choice. It is performed by means of dipolar interaction measurements. Distance information can be then extracted for organic radicals according to Eq. 2.18.

DEER, as every pulse experiment, requires usage of short and intense irradiation. An mw pulse is applied to create non-equilibrium magnetization, which is then precessing about the magnetic field. This response to the radiation causes a free induction decay (FID). However, due to the dead time of the spectrometer the fast dephasing cannot be observed in EPR. Additional spin system manipulation is required and a sequence of pulses is applied.

#### 2.2.1 Hahn echo

The spin echo experiment, described firstly for nuclear spins by E. Hahn [102], is capable of reversing the dephasing of spin packets (Fig. 2.3). It is possible due to the fact that a spin packet - the ensemble of spins with the same Larmor frequency - experiences the same local field. The first electron spin echo signal was observed by R. Blume [103].

At the beginning of the experiment the magnetization vector \( \vec{M} \) is aligned with the external magnetic field \( \vec{B}_0 \), so typically with the z-axis. A \( \pi/2 \) pulse brings \( \vec{M} \) to the xy-plane (Fig. 2.3, lower panel). Due to different Larmor frequencies, spin packets start to dephase during the time \( \tau \). This process is reversed by applying a \( \pi \) pulse along the x-axis resulting in flipping all spin packets. As the result magnetization gets refocused, a spin echo is formed at the time \( \tau \) after the second pulse. It resembles two FID signals put “back-to-back” (Fig. 2.3).

#### 2.2.2 Four-pulse DEER sequence

The 4-pulse DEER sequence is based on the refocused echo scheme at the observer frequency (Fig. 2.4 A). Additionally, the \( \pi \) pulse applied at the pump frequency is being shifted in the time \( (t) \). Because the echo-detected field swept spectrum of a nitroxide radical is broader than the excitation bandwidth of a single pulse, it is possible to
selectively manipulate groups of spins that have different resonance frequencies. The π pulse applied at the pump frequency $\nu_{\text{pump}}$ inverts the pumped spins exclusively, thus the local magnetic field felt by the observer spins changes if the spins are within the detectable distance range (Fig. 2.4 B). This change leads to a frequency shift, the amplitude of the echo signal is decreased. For every time step $t$ the amplitude will vary and thus measurements taken with a series of different shifts $t$ will reveal the frequency with which the echo of the observer spins is modulated - the dipolar frequency.

As a result of the experiment, time domain data are obtained ($V(t)$). They resume intra- ($F(t)$) as well as inter-molecular ($B(t)$) interactions between the free electrons.
The intermolecular part of the signal arises from random distributions of spins and can be described by a stretched exponential decay

$$B(t) = \exp(-kt^D/\beta)$$

The background signal depends on the investigated system, concentration and spatial distribution of spins. Once known, the background contribution can be removed and by this way the intramolecular couplings can be separated as the form factor (Fig. 2.5)

$$F(t) = 1 - \lambda[1 - \cos(\omega_{DD}t)]$$

The form factor depends on inversion efficiency $\lambda$ and dipolar coupling frequency $\omega_{DD}$ and is thus a source of the distance information. Fourier transform of this signal leads to a sum of Pake patterns of every distance contributing to $F(t)$, so not only the distance, but a distribution of distances can be obtained. Usually, additional smoothing is applied during the transformation (Tikhonov regularization) to distance domain to counteract noise-induced numerical instabilities.

Nuclei in the vicinity of electron spins under investigation may give rise to nuclear modulation artefacts. Their contribution is suppressed by a systematic change of the interpulse delay $\tau_1$. For protonated solvents at X-band frequencies, adding the traces collected for eight different $\tau_1$ with an increment of $\Delta \tau_1 = 8$ ns is performed, similarly for deuterated ones with the only difference that $\Delta \tau_1$ is 56 ns. At Q-band frequencies, proton modulation is usually negligible (still $\Delta \tau_1 = 4$ ns might be used for averaging), the deuterium signal is averaged with $\Delta \tau_1 = 16$ ns.

## 2.3 Distance geometry

The molecular distance geometry problem (MDGP) is to find the coordinates of a set of points by using only the distances between them. Mathematically, it is formulated as follows:
Given a set \( A \) of atom pairs \((i, j)\) on a set of \( n \) atoms and distances \( d_{ij} \) defined over \( A \), find positions \( x_1, \ldots, x_n \in \mathbb{R}^3 \) of the atoms in the molecule such that

\[
\forall (i, j) \in A \quad \|x_i - x_j\| = d_{ij}
\]  

(2.22)

When the exact distances between all pairs of atoms in molecule are known, a unique 3-dimensional structure can be determined by a linear time algorithm [105]. Nevertheless, due to errors in given distances, a solution might not be unique or does not exist at all. As a consequence the molecular distance geometry problem becomes difficult to solve. For instance, the distance may be distributed if the molecule can adopt more than one conformation between the observed atoms \((i, j)\).

The approach to solve the MDGP was firstly introduced to consolidate all available distance data on the solution conformation of the biomacromolecules [106, 107]. Distance information could be gained from any means of experiments, e. g. magnetic resonance techniques, FRET, or chemical crosslinking studies. Most often these techniques offer a range for the distances, thus defining naturally upper and lower limits. Distances measured in a mobile molecule can as well be confined between well-defined borders - lower and upper bounds \( l(i, j), u(i, j) \). MDGP is then reformulated

\[
\forall (i, j) \in A \quad l_{ij} \leq \|x_i - x_j\| \leq u_{ij}
\]  

(2.23)

The solution of such problem is in principle possible for any Euclidian arrangement of points up to mirror reflection [101].

MDGP is a special case of geometric algebra of a 3-dimensional Euclidean vector space. It enriches linear algebra with a new geometric description. For a brief overview of the topic of geometric algebra the reader is directed to Appendix A, for in-depth description to the literature [108].

Due to the complexity of a problem exceeding a few atoms, direct calculations cannot be performed and numerical methods have to be employed in order to gain at least some insight into the structure. In general, a representative set of solutions satisfying the given bounds on distances is sought. It is further called a conformational ensemble. If the sampling is sufficient and random, it can be assumed that all persistent geometric relations were captured within the ensemble.

The procedure of obtaining the structure consists of three steps:

- obtaining a full set of \( l(i, j) \) and \( u(i, j) \) from sparse distance constraints during bounds smoothing
- choosing a random distance matrix from within given limits and computing the coordinates during embedding
- optimizing coordinates against an ”error” function
2.3.1 Bounds on distances and triangle bound smoothing

The obtained conformations do not need to satisfy the geometric conditions exactly, it is sufficient and most useful to constrain them only by lower and upper bounds on distances \( l(i,j), u(i,j) \).

The possibility of inconsistent distance information is a big difficulty when it comes to putting the theory into practice. Fulfillment of all triangle inequalities is tested every time after adding the constraints of different sources in order to check for possible inconsistencies. If they exist, they are further removed by forcing the relations

\[
\forall (i,j,k) \in A \quad \begin{align*}
  u(i,j) &\leq u(i,k) + u(j,k) \\
  l(i,j) &\leq l(i,k) + u(j,k)
\end{align*}
\] (2.24)

Eq. 2.24 is called triangle inequality as it pertains to the fact that the length of any side of a triangle cannot exceed the sum of the lengths of the remaining sides. For more details how it emerges, check Appendix A or [101].

2.3.2 Embedding to Cartesian space and removal of bound violations

When the consistent bounds on distances are obtained, the distance matrix, which is a premetric matrix, is set by random choice of values from within the limits

\[
d(A) = d(i,j,\ldots,k) = \begin{bmatrix}
  0 & d(i,j) & \ldots & d(i,k) \\
  d(i,j) & 0 & \ldots & d(j,k) \\
  \vdots & \vdots & \ddots & \vdots \\
  d(i,k) & d(j,k) & \ldots & 0
\end{bmatrix}
\] (2.25)

Metrization of \( d(A) \) takes place. Namely, selection of a random \( d(i,j) \) value is accompanied by setting the lower and upper limits equal to the chosen number. Then the triangle inequality limits are recomputed. This process is repeated iteratively for each distance. The algorithm results in \( l(A) = u(A) = d(A) \) satisfying all original limits and all triangle inequalities.

The 3-dimensional conformation is obtained by means of coordinatization algorithm called EMBED [109]. Though since then other algorithms appeared [110–112]), EMBED seems to perform comparably well.

Coordinates can be easily found by eigenvalues calculation. This approach gives also the advantage of avoiding the local minima. It is done by means of minimizing the function

\[
\sum_{1=i<j}^{n,n} (w_{ij}((\vec{x}_i - \vec{x}_j)^2 - D_{ij}))^2
\] (2.26)
where \( D_{ij} = D(i, j) = d_{ij}^2 \) are squared distances and \( w_{ij} \) are weights.

Eq. 2.26 can be further expanded to

\[
4 \sum_{1=i<j}^{n,n} w_{ij}^2 ((\vec{x}_i \cdot \vec{x}_j) - \frac{1}{2}(\vec{x}_i^2 + \vec{x}_j^2 - D_{ij}))^2 \tag{2.27}
\]

As we want to get a result independent on frame origin and orientation, the squared distances to the center of mass have to be considered

\[
D_{0i} = \frac{1}{M} \sum_{j=1}^{n} m_j D_{ij} - \frac{1}{M^2} \sum_{j=1}^{n,n} m_j m_k D_{jk} \tag{2.28}
\]

where \( m_j \) are the masses of points and \( m = \sum_j m_j \).

In general \( D_{0i} \) is tolerant to errors in \( D_{ij} \) as it involves averaging

\[
D_{ij} = (\vec{x}_i - \vec{x}_j)^2 \Rightarrow D_{0i} = \vec{x}_i^2 \tag{2.29}
\]

Our problem is hence reformulated to minimizing the following function \( F(\vec{x}) \)

\[
F(\vec{x}) = \frac{1}{2} \| W(XX^T - A)W \|^2 \tag{2.30}
\]

Now, we define \( W = \text{diag}(w_1, \ldots, w_n) \), with \( A = [a_{ij}] \) and with \( X = [x_1, \ldots, x_n]^T \) being a matrix of coordinates. Eq. 2.30 can be then rewritten as

\[
F(X) = \frac{1}{2} \| W(XX^T - A)W \|^2 \tag{2.31}
\]

\( A \) is related to \( D = [D_{ij}] \)

\[
A = -\frac{1}{2}(I - 1m^T/M)D(I - m1^T/M) \tag{2.32}
\]

where \( 1 = [1, \ldots, 1]^T \), \( m = [m_1, \ldots, m_n]^T \) and \( I \) the identity matrix. \( A_X = XX^T \) and \( D_X = [(\vec{x}_i - \vec{x}_j)^2] \). Assuming also \( P = W(I - 1m^T/M)\sqrt{2} \) we obtain a function being a direct fit to the squared distances \[113]\]

\[
F(X) = \frac{1}{2} \| P(D_X - D)P \|^2 \tag{2.33}
\]

Variables can be changed as \( Y = WX \) and \( B \) can be defined as \( B = WAW \). Now

\[
F(Y) = \frac{1}{2} \| YY^T - B \|^2 \tag{2.34}
\]

Obviously at the global minimum gradient of the function \( F(\vec{Y}) \) has to become zero

\[
\frac{\partial F}{\partial y_{ij}} = (\vec{Y}Y^T - B)Y = 0 \Rightarrow BY = Y(Y^TY) \tag{2.35}
\]
Rotation of coordinates in space would lead to a diagonal matrix filled with eigenvalues

$$Y^T Y = \text{diag}(\lambda_1, \lambda_2, \lambda_3)$$  \hspace{1cm} (2.36)

For the column numbered $i$ we have

$$BY_i = \lambda_i Y_i$$  \hspace{1cm} (2.37)

Hence at any stationary point of the considered function $F(X)$

$$Y^T BY = (Y^T Y)^2 = (\text{diag}(\lambda_1, \lambda_2, \lambda_3))^2$$  \hspace{1cm} (2.38)

And so the solution of the problem is

$$F(Y) = \text{tr}(B^2) - \lambda_1^2 - \lambda_2^2 - \lambda_3^2$$  \hspace{1cm} (2.39)

As derived, the global minimum of the function $F(\vec{x})$ is given by the three biggest eigenvalues of the matrix $B$. The corresponding eigenvectors have to be scaled by their squareroots. Coming back to the original coordinates we obtain the coordinate set

$$X = W^{-1}Y$$  \hspace{1cm} (2.40)

The obtained solution, though fitting to the given distance constraints in the rmsd-sense, may violate some constraints. Thus bound violations (within some small tolerance) have to be removed. The way to do this is to construct a penalty function for which the global minimum corresponds to the structure fulfilling the bounds accurately, if existing. Then a nonlinear optimization has to be employed to find this minimum [107]. Such an optimization procedure is much easier than the energy minimization.

Bound violations can be removed by a conjugated gradient method [107]. Our implementation follows a simpler, but still computationally efficient approach. A correction matrix in distance space is built that has zero entries for all bounds that are fulfilled, negative entries corresponding to violations of an upper bound and positive entries corresponding to violation of lower bounds. Then, for each point coordinate a shift vector is constructed that is the sum of all vectors that would need to be added to this coordinate to remove bound violations for distances to all other points of the model. This shift vector is multiplied by a scaling factor $\lambda = k_{iter}/n_{iter}$, where $k_{iter}$ is an iteration index and $n_{iter}$ the maximum number of iterations, and added to the point coordinate. If bound violations are not reduced to less than $10^{-6}$ Å after $n_{iter}$ iterations, the model is discarded and the next Monte Carlo trial is started. We found empirically that good convergence is achieved with $n_{iter}$ being the sum of squares of all bound violations (in units of Å) of the originally embedded model, but at least 1000 and not more than 50'000.

It was empirically proven that the quantity of distances rather than their quality reduces effectively the range of available conformations. What is even more interesting, redundant information improves the convergence rate in the EMBED algorithm [107].
For the introduced methodology, inconsistent distance information shows up either as erroneous bounds \((u_{ij} \leq l_{ij})\) or a failure of EMBED to converge.
Chapter 3

Materials and methods

Careful measurements and cautious data analysis are very important when quantitative information has to be deduced. For the sake of modelling, mean distances and their standard deviations, which are extracted from the measured traces, are the source of constraints for the modelling. Though these distances are not the only given information for calculations, one has to remember that mistakes and inconsistencies may still lead to unexpected architectures.

Modelling of a protein architecture based on EPR distance information can be performed only after the problem has been simplified. Because of experimental and computational limitations, some approximations are required. Approximation of helices as straight cylinders with point labels attached allows for characterizing the protein fold, but does not provide a structure at atomic resolution.

3.1 Modelling approach

As experimental resources and computing power are limited, the developed modelling approach was originally aiming at finding the structure of proline permease PutP in its coarse-grained representation (Figure 3.1). An ideal two-point cylinder was attributed to every helix of the protein. Hence, PutP with its 13 helices would be relatively arranged in space by proper positioning of their 26 helix ends. These 26 helix ends correspond to $3 \cdot 26 = 78$ Cartesian coordinates. Removing local translational and rotational degrees of freedom, which are of no concern, 72 unknowns must be determined. In total $2n(2n - 1)/2 = 325$ distances between helix end pairs could, in principle, be measured, which would overdetermine the problem. Because we rely on labelling and the label-to-helix-end distance is too long to be neglected, we must define the model by $p = 52$ points (26 helix ends and 26 label sites) resulting in $3 \cdot 52 - 6 = 150$ unknown coordinates. Such a model could, in principle, still be overdetermined by accessible pair distances. But despite huge biochemical dedication such a number of constraints could not be gathered. The total number of 75 experimental restraints is one of the largest constraint sets obtained by EPR techniques for a
membrane protein, yet cannot fully define the structure, thus implying severe underdetermination of the problem, even on such a coarse-grained level.

Figure 3.1: Presented coarse-grained modelling approximates helices as regular cylinders and label position as the NO-bond midpoint. \( L, l_p, l_c, \theta_c \) and \( \theta_p \) are well defined based on LeuT PDB # 2A65 [5] and vSGLT PDB # 3DH4 [38] structures, \( \phi \) is unconstrained due to helix non-ideality.

The developed modelling approach is based on distance geometry. The general modelling scheme is depicted in Figure 3.2.

At the beginning of the modelling, the matrices of lower and upper bounds are set. They contain information on distances between all labels as well as labels and helix ends for both cytoplasmic and periplasmic sites, which leads to matrix dimensions \( 4k \times 4k \), where \( k \) is a number of helices. During the modelling process these matrices are filled with distance information. As mentioned, for PutP 75 distances out of 325 possible distances are available. The missing values in the matrices have to be complemented. Hence modelling has to be supported by additional information.

Intrinsic constraints were derived from two crystal structures of other secondary transporters existing at the time when the approach was developed - LeuT (PDB # 2A65 [5]) and vSGLT (PDB # 3DH4 [38]). By systematically analyzing these structures and possible spin label conformations, lower and upper bounds could be derived for the following parameters: the length of the helix - \( L \), the distance between helix end and NO-mean of the label on cytoplasmic and periplasmic side - \( l_p \) and \( l_c \), the angles between the central axis of the cylinder and the approximated label position - \( \theta_c \) and \( \theta_p \) (Figure 3.1). The angle between projections of the labels on xy-plane \( \phi \) remains unconstrained due to helix non-ideality.

As default bounds, the minimum and maximum distance between two labels was set from 3 to 80 Å, the distance between a label and the helix end point to which it was attached was constrained to the interval from 2 to 11 Å, transmembrane distances to the interval from 20 to 70 Å, distances between any helix end points between 8 and 70 Å, the helix length was calculated for every helix from the number of residues in the helix according to the equations:

\[
L_{\text{min}} = n \cdot l_{\text{min, res}} \\
L_{\text{max}} = n \cdot l_{\text{max, res}}
\]
where \( L \) is the length of the helix, \( l_{\min, \text{res}} = 1.35 \, \text{Å} \) the minimum length per residue of an \( \alpha \) helix, \( l_{\max, \text{res}} = 1.50 \, \text{Å} \) the maximum length per residue of an \( \alpha \) helix and \( n \) is the number of residues in the helix. Helix length was thus constrained to the interval from 18.9 to 34.5 \( \text{Å} \).

During the course of the project, crystal structures of a few more secondary transporters became available which allowed to further constrain the problem by structural analogy. Although these transporters are not from the same family and despite a very low similarity at the primary structure level, their structures seemed to share a topologically very similar motif of 10 transmembrane domains (TMDs), out of which the first five are related to the remaining five TMDs by a pseudo-two-fold rotation axis in the membrane plane [5, 21]. In this way, additional \( k_{\text{core}}(k_{\text{core}} - 1)/2 \) approximate distances can be

Figure 3.2: Schematic depiction of the developed modelling approach. For details see text.
derived from the template structure. Simulated distances are the distances between the mean NO-bond midpoint coordinates for the ensemble of all spin label rotamers. A template structure is translated into the cylindrical representation of it. The length of the helices is defined by the respective number of residues, its placement in Cartesian space is set by the "line" for which the sum of squares of the distances to all $C\alpha$ atoms is minimum. The midpoint is the mean $C\alpha$ coordinate of all residues belonging to the helix. Next, helix ends of a previously obtained bundle are adjusted such that they fit to the protein of interest, helices of a proper length are built up starting from the helix midpoints and elongating along the central lines of the cylinders.

Information on helix-to-helix distances from the templates should be weighted, since some regions of the protein core exhibit less variation amongst the 8 unique crystal structures (vSGLT # 2XQ2 [39], LeuT # 2A65 [5], Mhp1 # 2JLN [37], Mhp1 # 2X79 [15], ApcT # 3GIA [20], AdiC # 3NCY [16], BetP # 2WIT [21], CaiT # 3HFX [22]) than other regions. Therefore, a distance between two residues obtained from the template was allowed to vary within an interval whose width was related to the rmsd of this distance among the 8 structures by $2 \cdot \sqrt{\frac{1}{n-1} \sum (d_i - \bar{d})^2}$ for $n$ distances $d_i$ and their mean $\bar{d}$.

Lower and upper bound matrices are further filled with the experimental information. For that, distance distributions obtained from the DEER data were analyzed, parameters essential for modeling - mean distances $r_{\text{mean}}$ and standard deviations $\sigma_r$ - were extracted from the traces by their processing. Unfortunately, simple cases as shown in Figure 3.3 were rare, obtaining $r_{\text{mean}}$ and $\sigma_r$ was often ambiguous (Figure 3.4).

![Figure 3.3](image-url)

Figure 3.3: Exemplary DEER traces of liposomal PutP sample 326R1/164R1 acquired at Q band. From left to right: normalized time trace, background-corrected data and Tikhonov regularized distance distribution. $r_{\text{mean}}$ and $\sigma_r$ used for modelling are indicated.

Protein samples had been prepared with the protein:lipid ratio of 1:20 (w/w) which corresponds to the ratio of 1 protein for every 1250 lipids. This implied big protein density. Thus, in many of the traces a contribution from protein aggregates was observed, usually around $\sim 4.5$-$5$ nm, which had to be extracted and excluded in deriving the distance constraints (Figure 3.4 A).

Additionally, processing of often too short experimental traces introduced a peak at longer distances that are absent in the sample. Therefore, in the cases when a contribution from short distances to their distribution was dominant, a contribution from longer distances was neglected (Figure 3.4 B).

All original DEER data were reanalyzed in this thesis, using the same procedure, in order to reduce ambiguity in data interpretation. Data were of different quality, as they
had been acquired in the course of several years on different spectrometers. Generally, later acquired data had a better quality, as the spectrometers had improved.

Figure 3.4: DEER traces of liposomal PutP samples acquired at Q band. From left to right: normalized time trace, background-corrected data and Tikhonov regularized distance distribution. Range for obtaining $r_{\text{mean}}$ and $\sigma_r$ used for modelling is indicated.

The lower and upper bounds for experimental constraints with a mean distance $r_{\text{mean}}$ and standard deviation $\sigma_r$ are implemented with a tightness factor $c$

$$r = r_{\text{mean}} \pm c \cdot \sigma_r$$  \hspace{1cm} (3.1)

In early NMR-based protein modelling this factor was taken as 2.

The existing values in lower und upper bound matrices are updated only when this process causes tightening of the bound (increasing $l_{ij}$ or decreasing $u_{ij}$). After each change of the lower and upper bound matrices, triangle bound smoothing is performed.

To investigate the limitations of the methodology, distances between coordinates of the helix ends were considered. Such distances are, of course, experimentally not accessible, but such modelling can offer insight in the approach. Constraint lists for vSGLT and Rho modelling were generated by means of a homemade MATLAB script reading the Cα distances between helix ends residues from a PDB file. The number of coordinates was the same as in modelling with label coordinates. Both the complete set of coordinates and a set of constraints with a size available for PutP were generated. In this case no standard deviations of the distances are defined, thus they were set arbitrarily.

In the next step, a distance matrix is generated by randomly selecting all distances so that they fall within the bounds. This step initiates one trial in a Monte Carlo loop. The Monte Carlo loop repeats until a preset number of $n$ structures could be successfully metrized and corrected against the bound violations. Also helix-helix clashes are eliminated by shifting the helices along their normals. Some trials may fail at one of these steps.

In subsequent runs the random number seed can be derived from the time when the run is started, so that each time a different ensemble results.
An ensemble of generated structures is then clustered. The cluster algorithm starts by computing pairwise coordinate rmsd of all structures. The structures with smallest mean square coordinate deviation from all other structures ($\text{rmsd}_{\text{ensemble}}$) is chosen as a central structure and is taken as a representative of this ensemble. An ensemble can be reduced by removing the structures with biggest mean square coordinate deviation from all other structures (reduction based on a geometrical criterium of selection). Next all pairwise rmsds are recalculated and a new central structure is determined.

The original modelling algorithm was ment not to discriminate between different types of input information, i.e. the experimental constraints (for PutP) or simulated distance restraints (for the test cases Rho and vSGLT) are equally treated as f.e. template-derived restraints. Hence, the structures can be pre-selected before clustering by including only those with the best agreement (root mean square deviation between modelled distance and experimental $r_{\text{mean}}$) with the constraint list (EPR pre-selection). This pre-selection is based on the smallest rmsd of distances in each model vs mean distance in the constraint list. It assumes a normal distribution of the distances. However, experimental distributions often do not resemble a normal distribution, they can be bimodal or asymmetric. Thus, comparison to experimental distance distributions may be a better approach. For tests, these distributions can be simulated with a rotamer library approach [114] in the case of label constraints; for helix end constraints the problem does not arise.

Simulated distance distributions were normalized. For each model in an ensemble a loop was run. It was defined over all spin pairs $(k,l)$ for which simulated distributions existed. Within this loop, the difference between each distance $r_i$ present in the distribution and the distance obtained from the model $r_{\text{model}}$ was calculated

$$d_i = r_i - r_{\text{model}}$$ (3.2)

A distance from this distribution that was the closest to $r_{\text{model}}$ (corresponding to the smallest $d_i$) was searched and assigned as $r_{ic}$. The probability $P$ of finding distance $r_{\text{model}}$ in the distance distribution was approximated by a linear function between $P(r_{ic})$ and $P(r_{ic+1})$ if $r_{\text{model}} > r_{ic}$ or between $P(r_{ic})$ and $P(r_{ic-1})$ if $r_{\text{model}} < r_{ic}$ and hence

$$P(r_{\text{model}}) = \begin{cases} 
\frac{P(r_{ic+1}) - P(r_{ic})}{r_{ic+1} - r_{ic}} r_{\text{model}} + \frac{P(r_{ic})r_{ic+1} - P(r_{ic+1})r_{ic}}{r_{ic+1} - r_{ic}} & \text{if } r_{\text{model}} > r_{ic} \\
\frac{P(r_{ic}) - P(r_{ic-1})}{r_{ic} - r_{ic-1}} r_{\text{model}} + \frac{P(r_{ic-1})r_{ic} - P(r_{ic})r_{ic-1}}{r_{ic} - r_{ic-1}} & \text{if } r_{\text{model}} < r_{ic}
\end{cases}$$ (3.3)

Next, the sum of $P(r_{\text{model}})$ over spin pairs $(k,l)$ was calculated and assigned as a ”score” of this model. An ensemble could be further reduced based on scoring of the models.

The original ideal of inferring helix bundle topology from a small set of experimental constraints was inspired by [77]. The approach described there is a two-step algorithm. In the first step, an ensemble of models is generated based on experimental distances [76]. Then these initial conformations are refined by using a set of penalty functions.
Although we use a different approach for sampling the conformational space, it appeared feasible to use similar penalty functions. Functions penalizing improper distances ($P_{\text{dist}}$), poor packing ($P_{\text{pack}}$, $P_{\text{angle}}$), poor compactness ($P_{\text{compact}}$), wrong contact ($P_{\text{contact}}$), violations of van der Waals interactions ($P_{\text{vdW}}$) and improper planar arrangement ($P_{\text{apv}}$, $P_{\text{mpv}}$ and $P_{\text{plv}}$) were introduced. $P_{\text{apv}}$ is the square sum of the distance of all helix end points from their respective planes, either the cytoplasmic or the periplasmic mean helix end planes

$$P_{\text{apv}} = \sum_{2k} d_{ip}^2$$

$k$ is the number of helices and $d_{ip}$ the distance between the $i$-th helix end and its respective plane.

$P_{\text{mpv}}$ is the square sum of the distance of helix midpoints from the central bilayer plane

$$P_{\text{mpv}} = \sum_k d_{mp}^2$$

where $d_{mp}$ is the distance between the $m$-th helix midpoint and the central plane.

$P_{\text{plv}}$ is the penalty that is defined as an angle between the normal vectors of the cytoplasmic and periplasmic mean planes

$$P_{\text{plv}} = \frac{180}{\pi} \arccos \left| \sum_k \vec{c} \cdot \vec{p} \right|$$

where $\vec{c}$ and $\vec{p}$ are the normal vectors of the cytoplasmic and periplasmic planes, respectively.

$P_{\text{dist}}$ is the penalty for improper distances defined by

$$P_{\text{dist}} = k_{\text{dist}} \cdot \sum \left( \sum (d_{ij} - l_{ij})^2 + (u_{ij} - d_{ij})^2 \right)$$

where $k_{\text{dist}} = 500$ is the force constant for distance constraint violations, $l_{ij}$ and $u_{ij}$ are the lower and upper bounds for distances between helix end points $i$ and $j$, $d_{ij}$ is the distance between helix end points $i$ and $j$.

$P_{\text{pack}}$ is the penalty for poor packing given by

$$P_{ij,\text{pack}} = \begin{cases} (r_l - r_{ij})^2 & \text{if } r_{ij} < r_l \\ (r_{ij} - r_u)^2 & \text{if } r_{ij} > r_u \end{cases}$$

where $r_{ij}$ is the distance between the centers of mass of helix $i$ and $j$ in the current structure, $r_{\text{mean}} = 12.8 \, \text{Å}$ is the mean distance between the centers of mass of consecutive helices, $\sigma_r = 5.3 \, \text{Å}$ the standard deviation for this distance, $r_l = r_{\text{mean}} - 2 \cdot \sigma_r$ and $r_u = r_{\text{mean}} + 2 \cdot \sigma_r$. The packing distance penalties are then summed over all helix pairs as follows

$$P_{\text{pack}} = k_{\text{dist}} \sum P_{ij,\text{pack}}$$

where $k_{\text{pack}} = 50$ is the force constant for packing density.

If two helices are closer than it is allowed by the van der Waals repulsion, such a contact
is penalized
\[ P_{vdW} = k_{vdW} \sum (r_{ij} - r_{\text{min}})^4 \]  
(3.10)

where \( k_{vdW} = 100 \) is the force constant for van der Waals repulsion and \( r_{\text{min}} = 6 \, \text{Å} \) the minimum helix-helix distance.

\( P_{angle} \) is the penalty for poor helix packing angle
\[ P_{ij,angle} = \begin{cases} (\theta_l - \theta_{ij})^2 & \text{if } \theta_{ij} < \theta_l \\ (\theta_{ij} - \theta_u)^2 & \text{if } \theta_{ij} > \theta_u \end{cases} \]  
(3.11)

where an angle \( \theta_{ij} = \arccos(\vec{v}_1^T \vec{v}_2) \) and \( \vec{v}_i \) is the vector connecting the periplasmic and the cytoplasmic end of helix \( i \), \( \theta_{\text{mean}} = 30.9 \cdot \pi/180 \) is the mean packing angle, \( \sigma_\theta = 16.4 \cdot \pi/180 \) the standard deviation of the packing angle, \( \theta_l = \theta_{\text{mean}} - 2 \cdot \sigma_\theta \) and \( \theta_u = \theta_{\text{mean}} + 2 \cdot \sigma_\theta \). The packing angle penalty is then summed over all helix pairs as follows
\[ P_{angle} = k_{angle} \sum P_{ij,angle} \]  
(3.12)

where \( k_{angle} = 5 \cdot 180/\pi \) is a force constant for helix packing angle.

\( P_{contact} \) is a penalty for poor contact
\[ P_{contact} = k_{contact} \sum (2 - c) \]  
(3.13)

where \( k_{contact} = 500 \) is a force constant for contact penalty and \( c < 2 \) is the number of helices that are within the distance of 14.3 Å from the concerned helix.

\( P_{compact} \) is a penalty function for poor compactness
\[ P_{ij,compact} = \begin{cases} (b_l - b_{ij})^2 & \text{if } b_{ij} < b_l \\ (b_{ij} - b_u)^2 & \text{if } b_{ij} > b_u \end{cases} \]  
(3.14)

where \( b_{ij} \) is the fraction of distances between the point \( i \) on the helix and a helix end point \( j \) that are within the minimal distance of 6.65 Å, \( b_l = 0.17 \) and \( b_u = 0.86 \). The compactness penalty is then summed over all pairs of points as follows
\[ P_{compact} = k_{compact} \sum P_{ij,compact} \]  
(3.15)

where \( k_{compact} = 10000 \) is a force constant for compactness penalty.

The total penalty function \( P_{vec} \) is calculated as a sum of all of them
\[ P_{vec} = P_{apv} + P_{mpv} + P_{plv} + P_{dist} + P_{pack} + P_{vdW} + P_{angle} + P_{contact} + P_{compact} \]  
(3.16)

Modelling scripts were written in the MATLAB environment (The MathWorks Inc, Natick, MA, USA).
3.1.1 Improvement of computational efficiency

A change of floating point double number to integer numbers in triangle bound smoothing allowed to speed up the calculations without significant influence on the calculation result. Numbers are converted by multiplication by $10^4$ to shift the floating point, then switching to the uint32 number format of MATLAB. After bound smoothing, numbers are converted back by a division by $10^4$. This factor is sufficiently large to make the roundoff errors in the distances (originally given in units of nanometers) negligible and sufficiently small to avoid overflow of the unsigned 32 bit integer numbers. The procedure accelerates the calculations by factor of 6. The saving in computation time arises from comparing only integers instead of the mantissa exponent and sign of a floating point number. Since bound smoothing takes up only part of the total computation time, the speedup for the whole computation is smaller, 2-3 times.

3.2 Sample preparation

All projects described in this thesis demanded substantial biochemical and bioengineering effort. Modelling of PutP architecture as well as functional studies of PutP were possible thanks to the group of Prof. Dr. Heinrich Jung at LMU Munich, where all samples were prepared.

3.2.1 Mutant choice

PutP topology [6, 60] suggested a rational choice of the labelled positions for modelling purposes. With 13 helices each with 2 helix ends, 325 possible pairs could have been labelled and measured. In an earlier thesis by Daniel Hilger [115] at LMU Munich 50 of those pairs were produced and distance distributions were obtained for 45 of them. Labelling of the protein at the end of the helices was tried in order not to disrupt the structure [116]. When placing the label was not possible at the desired position according to the topology [6, 60], another trial was undertaken at some chosen residue placed more inside the helix. Positions accessed by labelling were as following (with highlighted reference sites):

- on the cytoplasmic site: 25, 41, 91, 128, 179, 190, 256, 276, 349, 371, 423, 432, 470,
- on the periplasmic site: 5, 62, 80, 149, 164, 210, 233, 294, 326, 391, 400, 446, 450.

Reference sites were chosen in order to better constrain the problem. Distances between reference sites and all other helix ends on the same side of the membrane were measured. A few pairs of residues on opposite sides of the membrane were chosen as additional constraints. The full list of constraints is given in Chapter 4, Table 4.2.
In order to investigate the process of substrate transport and the significance of extracellular loop 4 in this process, a set of spin-labelled double mutants was prepared. Residues undergoing labelling were chosen such that the contact between eL4 and structurally or functionally important protein regions was mapped (sites known for coordination of L-proline and \( \text{Na}^+ \) or being in their proximity):

- within eL4: 299/318
- between eL4 and inner or outer termini of TMs: 50/304, 294/326, 294/446, 298/446, else: 298/455, 294/403,
- between the termini: 326/446, else: 190/371, 91/371, 149/450, 41/118.

The internal bonding network of PutP was studied on the example of two mutations: Glu311 to Ala and Ala404 to Glu. Pairs chosen for labelling were the ones supposed to "see" the consequent conformational changes: 298/326, 298/391 and 190/371.

### 3.2.2 Preparation protocol

Samples were prepared by Dr. Daniel Hilger at LMU Munich. A bacterial Cys-less gene construct, where five naturally occurring Cys residues are substituted by different amino acids' residues, was chosen according to the criterium of highest activity of the expressed protein. The desired nucleotide substitution (site-directed mutagenesis) in the \( \text{putP} \) gene was generated by PCR. The obtained plasmid was then transferred into \( E. \, \text{coli} \), which overexpressed the mutated \( \text{putP} \) gene. Cell disruption and proper treatment allowed for extracting membranes containing, among other proteins, also PutP.

Membrane proteins solubilised by detergent were transferred to the column and washed with buffer containing MTSSL (MTSSL, Toronto Research Chemicals, Toronto, Canada). Spin labels got attached to Cys residues [85, 86], afterwards remaining non-bound labels and proteins other than PutP were washed out of the column (affinity chromatography). The purified protein was reconstituted into proteoliposomes prepared from \( E. \, \text{coli} \) polar lipids (67% phosphatidylethanolamine, 23.2% phosphatidyglycerol, and 9.8% cardiolipin; Avanti Polar Lipids, Alabaster, AL) at a protein:lipid ratio 1:20 (w/w) (1 protein molecule:1250 lipids), as described previously in [117]. Liposomes were first extruded through a 400 nm filter forming one-layered structures, what allowed for homogeneous distribution of the protein in the liposomes. For removal of the sparse detergent, bio-beads were used. Subsequently, they were removed by dialysis.

The detailed protocol of sample preparation can be found in [52, 115].

High density of the protein molecules (protein:lipid ratio of 1:20 (w/w)) caused presumable protein aggregation and unwanted densification of the spins. Therefore, liposomal samples for functional studies were prepared with a protein:lipid ratio of 1:40 (w/w) (1 protein molecule:2500 lipids). Final samples for measurements contained 10% glycerol-d8. A few micellar samples for comparison with liposomal ones were also prepared.
Chapter 3. Materials and methods

Proteins solubilised in DDM were spin-labelled and purified as stated above. Samples contained 10% glycerol + 10% glycerol-d8. Samples were prepared by Dr. Michael Raba or Susanne Bracher at LMU Munich, preparation details can be found in [62, 63].

3.3 DEER measurements

DEER traces were recorded at \( \sim 9.3 \) GHz (X band) and \( \sim 34.3 \) GHz (Q band). Because of sensitivity reasons, measurements for this thesis were mostly (always when available) conducted at Q-band frequencies, unless X-band measurement is indicated (the only X-band data in this thesis are highlighted in Table 4.2).

3.3.1 X band

Samples, which were shock-frozen in liquid nitrogen to avoid water crystal build-up were inserted into an Elexsys 580 spectrometer (Bruker Biospin GmbH, Karlsruhe, Germany) with a 3 mm split-ring resonator under conditions of strong overcoupling \((Q \approx 100)\). Four-pulse DEER measurements [88] were conducted at 50 K. The \( \pi/2 \) and \( \pi \) pulses at the observer frequency \( \nu_1 \) were set to either 16 ns and 32 ns or to an equal length of 32 ns in order to excite equal bandwidths. The \( \pi \) pulse at the pump frequency \( \nu_2 \) was set to 12 ns length to maximize modulation depth while still keeping the two excitation bandwidths separate. The long interpulse delay \( \tau_2 \) was typically 1500-2500 ns depending on the relaxation properties of the sample and on the distance. An initial value of \( t' = 80 \) ns and an increment of \( \Delta t' = 8 \) ns were used to acquire the time traces. Proton modulation was supressed by eight-step nuclear modulation averaging cycle with 8 ns increments [94]. A phase cycle \([+(+x),-(−x)]\) was applied to the first pulse. The pump frequency \( \nu_2 \) was set to the center of the resonator mode (typically 9.33 GHz) and coincided with the global maximum of the nitroxide spectrum [94], whereas the observer frequency \( \nu_1 \) was set to the local low-field maximum of the spectra, shifted by \(+65\) MHz from the pump frequency. Collected data were analyzed for dipolar evolution times \( t = t' - \tau_1 \geq 0 \). Accumulation times varied between 8 and 14 h.

3.3.2 Q band

A homebuilt or a commercial Q-band spectrometer (Bruker ELEXSYS-II E580) equipped with a pulsed TWT amplifier with nominal output power of 150 or 200 W, respectively, and a homebuilt TE102 rectangular resonator adjusted for oversized 3 mm sample tubes [96] were serving for Q-band DEER measurements of shock-frozen samples at 50 K. All pulse lengths were set to 12 ns, the pump pulse was applied at the maximum of the nitroxide spectrum, a frequency offset of (80..100) MHz was used for observer pulses. For the samples, where 10% d8-Gly was used, deuterium modulations were suppressed by an eight-step nuclear modulation averaging cycle with 16 ns increments,
whereas for non-deuterated samples this increment was 4 ns [96]. The long interpulse delay $\tau_2$ was typically between 1000 and 3800 ns depending on the relaxation properties of the sample, its deuteration and on the distance. Accumulation times varied between a few and $\sim 24$ h.

### 3.4 Data evaluation

All the distance measurements data were processed with DeerAnalysis2010, 2011 or 2013 - an open source software for analysis of pulsed dipolar spectroscopy data [118]. As part of the data was collected before the timeframe of this PhD thesis, all previously existing data were re-processed and a peak at $\sim 4.5$ nm stemming from protein aggregation was consistently excluded from the traces. Where such a peak could not be distinguished from the intramolecular label-to-label distance, data were rejected and not used as constraints. Additionally, for the distance distributions that were obtained from insufficiently long traces and were dominated by a contribution from short distances, a contribution from long distances was neglected. Dimensionality of the homogeneous background function for singly labelled mutants in liposomes varied from 1.69 to 2.5. In difficult cases, the validation tool of DeerAnalysis was applied to better define the distance distribution. For micellar samples background dimensionality was mostly equal to 3.

### 3.5 Simulating distances with MMM

MMM (Multiscale Modeling of Macromolecules) is an open source program for modelling spin label conformations [114] which allows, among else for simulating distance distributions and expected DEER traces. MMM was used to simulate the distances between chosen residues of Rho and vSGLT, crystallized proteins which served as best structures for exploring the methodology. Simulation were done with the rotamer library R1A and at 175 K.

#### 3.5.1 Helix ends definition

To simulate the distances between labels attached to the protein with MMM, residues for labelling have to be selected. As in the case of PutP, where experimental data was available, for Rho and vSGLT virtual labels were attached to the helix ends. Helix ends were defined by use of different prediction programs like DSSP [119] (implemented in MMM [114]) and Swiss-PdbViewer [120] or a visualization program Accelrys Discovery Studio [121]. The obtained results were compared and one set of helix ends was chosen.
Chapter 4

Results and discussion

4.1 Distance measurements of PutP for modelling purposes

4.1.1 List of measurements

75 distance constraints had been collected by means of DEER measurements by Daniel Hilger (45 constraints) and as a part of this thesis (remaining ones) in order to model protein architecture (Table 4.2). Data were collected at both X- and Q-band frequencies (check Table 4.2, Figure 4.1). Parameters essential for modelling - mean distances $r_{\text{mean}}$ and standard deviations $\sigma_r$ - were extracted from the traces as described in Chapter 3.1.

Figure 4.1 depicts a few traces collected at X band (B) and Q band (C). Q-band measurements were in favour due to the higher sensitivity and thus longer $\tau_2$ times available. All collected traces and their fitting can be found in Appendix, Figure B.2, B.3, B.4, B.5, B.6, B.7 and B.8.

Non-core helices are modelled based on only experimental distance constraints except for a few rather loose internal constraints, no constraints extracted from a template structure are available for them. The existing set of constraints for non-core helices of PutP had to be complemented during the course of this thesis in order to avoid ambiguities in spatial arrangement of these helices. Earlier modelling attempts indicated that this problem concerned especially helix I for which five experimental distances (three constraints from the cytoplasmic end of helix I - residue 25 and two from the periplasmic end of helix I - residue 5) were not enough to distinguish on which site of the bundle helix I is placed (Figure B.42). And so, additional mutants between the periplasmic ends of helix I and VII (PutP 5R1/233R1), between the periplasmic ends of helix I and helix VIII (PutP 5R1/294R1), between the cytoplasmic ends of helix I and helix VIII (PutP 25R1/276R1) and between the cytoplasmic ends of helix I and helix XI (PutP 25R1/423R1) were prepared and the DEER traces had been acquired (Figure B.8). Mutant 5R1/294R1 indicated very low modulation depth and a very broad distance distribution (data not shown) and hence, it was left out.
Chapter 4. Results and discussion

Figure 4.1: PutP DEER traces of liposomal samples. From left to right: normalized time trace, background-corrected data and Tikhonov regularized distance distribution of A, X-band spectra, B, Q-band spectra. * denotes an aggregation peak whose contribution was neglected in defining \( r_{\text{mean}} \) and \( \sigma_r \).

The total number of distance constraints collected for non-core helices of PutP was: 8 for helix I (5 on the cytoplasmic and 3 on the periplasmic side) - Table 4.2 in blue, 18 for helix XII (3 on the cytoplasmic and 15 on the periplasmic site) - Table 4.2 in red, 6 for helix XIII (3 on the cytoplasmic and 3 on the periplasmic site) - Table 4.2 in green.

Table 4.1: Helix ends of PutP according to the labelling scheme, topology prediction and homology model. Non-core helices highlighted in blue.

<table>
<thead>
<tr>
<th>labelling scheme</th>
<th>topology prediction [6, 60]</th>
<th>homology model [63]</th>
</tr>
</thead>
<tbody>
<tr>
<td>5-25</td>
<td>4-26</td>
<td>5-24</td>
</tr>
<tr>
<td>41-62</td>
<td>41-65</td>
<td>44-69</td>
</tr>
<tr>
<td>80-91</td>
<td>77-95</td>
<td>73-94</td>
</tr>
<tr>
<td>128-149</td>
<td>125-150</td>
<td>123-156</td>
</tr>
<tr>
<td>164-179</td>
<td>163-184</td>
<td>160-176</td>
</tr>
<tr>
<td>190-210</td>
<td>190-213</td>
<td>184-210</td>
</tr>
<tr>
<td>233-256</td>
<td>232-256</td>
<td>234-248</td>
</tr>
<tr>
<td>276-294</td>
<td>275-295</td>
<td>264-293</td>
</tr>
<tr>
<td>326-349</td>
<td>324-351</td>
<td>326-355</td>
</tr>
<tr>
<td>371-391</td>
<td>371-391</td>
<td>368-390</td>
</tr>
<tr>
<td>400-423</td>
<td>398-423</td>
<td>404-430</td>
</tr>
<tr>
<td>432-446</td>
<td>429-447</td>
<td>436-448</td>
</tr>
<tr>
<td>450-470</td>
<td>450-472</td>
<td></td>
</tr>
</tbody>
</table>
Table 4.2: Constraint list for PutP modelling. Helix ends of helix I highlighted in blue, of helix XII in red and of helix XIII in green.

<table>
<thead>
<tr>
<th>res1</th>
<th>res2</th>
<th>$r_{mean}$ [nm]</th>
<th>$\sigma_r$ [nm]</th>
<th>res1</th>
<th>res2</th>
<th>$r_{mean}$ [nm]</th>
<th>$\sigma_r$ [nm]</th>
</tr>
</thead>
<tbody>
<tr>
<td>41</td>
<td>25</td>
<td>2.46</td>
<td>0.45</td>
<td>446</td>
<td>294</td>
<td>4.56</td>
<td>0.88</td>
</tr>
<tr>
<td>41</td>
<td>62</td>
<td>3.23</td>
<td>0.49</td>
<td>446</td>
<td>326</td>
<td>4.36</td>
<td>0.71</td>
</tr>
<tr>
<td>41</td>
<td>91</td>
<td>3.19</td>
<td>0.49</td>
<td>446</td>
<td>391</td>
<td>1.73</td>
<td>0.48</td>
</tr>
<tr>
<td>41</td>
<td>179</td>
<td>4.37</td>
<td>0.60</td>
<td>446</td>
<td>400</td>
<td>1.40</td>
<td>0.27</td>
</tr>
<tr>
<td>41</td>
<td>190</td>
<td>3.08</td>
<td>0.45</td>
<td>446</td>
<td>450</td>
<td>2.30</td>
<td>0.25</td>
</tr>
<tr>
<td>41</td>
<td>276</td>
<td>2.75</td>
<td>0.52</td>
<td>349</td>
<td>25</td>
<td>3.05</td>
<td>0.50</td>
</tr>
<tr>
<td>41</td>
<td>349</td>
<td>3.90</td>
<td>0.79</td>
<td>349</td>
<td>62</td>
<td>2.81</td>
<td>0.77</td>
</tr>
<tr>
<td>41</td>
<td>371</td>
<td>5.21</td>
<td>0.31</td>
<td>349</td>
<td>91</td>
<td>2.83</td>
<td>0.28</td>
</tr>
<tr>
<td>41</td>
<td>423</td>
<td>4.68</td>
<td>0.80</td>
<td>349</td>
<td>128</td>
<td>1.89</td>
<td>0.17</td>
</tr>
<tr>
<td>41</td>
<td>446</td>
<td>4.74</td>
<td>0.64</td>
<td>349</td>
<td>179</td>
<td>1.99</td>
<td>0.47</td>
</tr>
<tr>
<td>41</td>
<td>470</td>
<td>3.46</td>
<td>0.88</td>
<td>349</td>
<td>190</td>
<td>2.16</td>
<td>0.36</td>
</tr>
<tr>
<td>371</td>
<td>25</td>
<td>3.26</td>
<td>0.46</td>
<td>349</td>
<td>256</td>
<td>1.80</td>
<td>0.27</td>
</tr>
<tr>
<td>371</td>
<td>91</td>
<td>3.85</td>
<td>0.61</td>
<td>349</td>
<td>276</td>
<td>2.86</td>
<td>0.35</td>
</tr>
<tr>
<td>371</td>
<td>128</td>
<td>1.99</td>
<td>0.34</td>
<td>349</td>
<td>423</td>
<td>3.39</td>
<td>0.27</td>
</tr>
<tr>
<td>371</td>
<td>179</td>
<td>2.10</td>
<td>0.51</td>
<td>349</td>
<td>432</td>
<td>2.73</td>
<td>0.26</td>
</tr>
<tr>
<td>371</td>
<td>190</td>
<td>2.89</td>
<td>0.33</td>
<td>349</td>
<td>446</td>
<td>2.78</td>
<td>0.29</td>
</tr>
<tr>
<td>371</td>
<td>276</td>
<td>4.01</td>
<td>0.37</td>
<td>349</td>
<td>470</td>
<td>2.59</td>
<td>0.66</td>
</tr>
<tr>
<td>371</td>
<td>349</td>
<td>2.09</td>
<td>0.60</td>
<td>326</td>
<td>5</td>
<td>2.27</td>
<td>0.19</td>
</tr>
<tr>
<td>371</td>
<td>423</td>
<td>3.84</td>
<td>0.21</td>
<td>326</td>
<td>41</td>
<td>2.81</td>
<td>0.32</td>
</tr>
<tr>
<td>371</td>
<td>446</td>
<td>3.29</td>
<td>0.36</td>
<td>326</td>
<td>80</td>
<td>2.75</td>
<td>0.39</td>
</tr>
<tr>
<td>371</td>
<td>470</td>
<td>2.85</td>
<td>0.63</td>
<td>326</td>
<td>149</td>
<td>2.65</td>
<td>0.30</td>
</tr>
<tr>
<td>62</td>
<td>80</td>
<td>1.92</td>
<td>0.40</td>
<td>326</td>
<td>164</td>
<td>3.29</td>
<td>0.16</td>
</tr>
<tr>
<td>62</td>
<td>149</td>
<td>3.28</td>
<td>0.61</td>
<td>326</td>
<td>210</td>
<td>2.59</td>
<td>0.45</td>
</tr>
<tr>
<td>62</td>
<td>164</td>
<td>3.50</td>
<td>0.99</td>
<td>326</td>
<td>233</td>
<td>4.11</td>
<td>0.49</td>
</tr>
<tr>
<td>62</td>
<td>210</td>
<td>3.01</td>
<td>0.43</td>
<td>326</td>
<td>294</td>
<td>3.14</td>
<td>0.63</td>
</tr>
<tr>
<td>62</td>
<td>233</td>
<td>3.45</td>
<td>0.35</td>
<td>326</td>
<td>349</td>
<td>3.18</td>
<td>0.41</td>
</tr>
<tr>
<td>62</td>
<td>294</td>
<td>3.21</td>
<td>0.55</td>
<td>326</td>
<td>371</td>
<td>4.06</td>
<td>0.35</td>
</tr>
<tr>
<td>62</td>
<td>326</td>
<td>2.71</td>
<td>0.66</td>
<td>326</td>
<td>391</td>
<td>3.53</td>
<td>0.21</td>
</tr>
<tr>
<td>62</td>
<td>371</td>
<td>4.61</td>
<td>0.73</td>
<td>326</td>
<td>400</td>
<td>3.45</td>
<td>0.24</td>
</tr>
<tr>
<td>62</td>
<td>391</td>
<td>3.04</td>
<td>0.67</td>
<td>326</td>
<td>450</td>
<td>4.55</td>
<td>1.00</td>
</tr>
<tr>
<td>62</td>
<td>446</td>
<td>3.47</td>
<td>0.33</td>
<td>41</td>
<td>256</td>
<td>2.72</td>
<td>0.50</td>
</tr>
<tr>
<td>62</td>
<td>450</td>
<td>3.66</td>
<td>0.35</td>
<td>41</td>
<td>432</td>
<td>4.83</td>
<td>0.56</td>
</tr>
<tr>
<td>446</td>
<td>5</td>
<td>5.02</td>
<td>0.71</td>
<td>371</td>
<td>256</td>
<td>1.66</td>
<td>0.12</td>
</tr>
<tr>
<td>446</td>
<td>80</td>
<td>3.30</td>
<td>0.46</td>
<td>371</td>
<td>432</td>
<td>3.03</td>
<td>0.19</td>
</tr>
<tr>
<td>446</td>
<td>149</td>
<td>2.59</td>
<td>0.56</td>
<td>25</td>
<td>276</td>
<td>3.14</td>
<td>0.40</td>
</tr>
<tr>
<td>446</td>
<td>164</td>
<td>2.25</td>
<td>0.47</td>
<td>25</td>
<td>423</td>
<td>5.28</td>
<td>0.42</td>
</tr>
<tr>
<td>446</td>
<td>210</td>
<td>4.40</td>
<td>0.81</td>
<td>5</td>
<td>233</td>
<td>5.67</td>
<td>0.35</td>
</tr>
<tr>
<td>446</td>
<td>233</td>
<td>3.51</td>
<td>0.32</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Chapter 4. Results and discussion

4.2 Definition of helix ends

The modelling approach assumes that all helices are labelled at their respective ends. For PutP, the corresponding sites had been selected in the experimental study by Daniel Hilger, based on an existing model for the topology of the protein. To test the modelling approach, helix ends also needed to be defined for the protein Rho and vSGLT, for which crystal structures are available.

4.2.1 Outcome of different prediction programs

Definition of helix ends of a protein may be often ambiguous despite the existence of a crystal structure. This is because different sets of criteria exist for assigning helical secondary structure, e.g., hydrogen bond estimation - DSSP program [119] or Ramachandran plots - Swiss-PdbViewer [120]. The assignment of helices in PDB structure files often does not conform to DSSP or standard Ramachandran plot definitions, but may have a subjective component introduced by the authors of the structure file, who tend to include residues with "nearly" helical parameters into the helices. Programs for PDB structure visualization often use the assignment provided in the PDB file.

4.2.1.1 vSGLT

Table 4.3: Helix ends definition of vSGLT. "swiss" stands for SwissPDB-Viewer, "dssp" for DSSP. The overall resolution of the crystal structures is ∼3 Å for #3DH4 and 2.7 Å for #2XQ2.

<table>
<thead>
<tr>
<th>3DH4 paper [38]</th>
<th>3DH4 swiss</th>
<th>3DH4 dssp</th>
<th>2XQ2 swiss</th>
<th>2XQ2 dssp</th>
<th>taken for modelling</th>
</tr>
</thead>
<tbody>
<tr>
<td>4-18</td>
<td>4-18</td>
<td>4-18</td>
<td>11-29</td>
<td>11-28</td>
<td>10-28</td>
</tr>
<tr>
<td>54-78</td>
<td>55-79</td>
<td>53-79</td>
<td>53-79</td>
<td>53-79</td>
<td>53-79</td>
</tr>
<tr>
<td>83-106</td>
<td>83-108</td>
<td>82-108</td>
<td>86-108</td>
<td>82-108</td>
<td>82-108</td>
</tr>
<tr>
<td>163-177</td>
<td>162-175</td>
<td>162-175</td>
<td>162-178</td>
<td>162-178</td>
<td>162-178</td>
</tr>
<tr>
<td>251-276</td>
<td>250-275</td>
<td>250-275</td>
<td>250-275</td>
<td>250-275</td>
<td>250-275</td>
</tr>
<tr>
<td>281-312</td>
<td>280-311</td>
<td>280-313</td>
<td>280-311</td>
<td>280-312</td>
<td>280-312</td>
</tr>
<tr>
<td>394-416</td>
<td>394-417</td>
<td>394-419</td>
<td>392-417</td>
<td>392-419</td>
<td>392-419</td>
</tr>
<tr>
<td>524-544</td>
<td>524-543</td>
<td>522-543</td>
<td>522-544</td>
<td>522-544</td>
<td>522-544</td>
</tr>
</tbody>
</table>

Helix ends of vSGLT were assigned by considering result of both a DSSP analysis and a Ramachandran plot analysis on both available crystal structures (Table 4.3). Both
SwissPDB-Viewer and DSSP analysis is mostly consistent. Inconsistencies appear only in a few cases:

- helix I is defined differently for PDB# 3DH4 and 2XQ2 structure (highlighted in red in Table 4.3, discussed in Section 4.2.2), because in the former structure, register of this helix with the amino acid sequence could not be established,

- the cytoplasmic end of helix VI is not properly defined in the paper describing the crystal structure PDB# 3DH4 [38] (check Figure 4.2), the same end is completely missed by DSSP for the PDB# 2XQ2 structure (both highlighted in red in Table 4.3),

- residue 383 being the cytoplasmic end of IX and residue 416 being the periplasmic end of helix X according to [38] (highlighted in green in Table 4.3 and on the Figure 4.2) are slightly displaced (by 3 residues each) in comparison with a DSSP and a Ramachandran plots analysis.

Figure 4.2: Depiction of the helix ends of vSGLT in the crystal structure [38] is not in complete agreement with results of mathematical analysis of atom coordinates. The helix ends assigned in the PDB files are indicated in green on A. PDB#2XQ2 structure, B. PDB# 3DH4 structure. Visualization by MMM.
4.2.1.2 Rho

For Rho, DSSP analysis is in rather good agreement with secondary structure assignment by the authors of crystal structure PDB# 1F88 (Table 4.4). This is probably due to the better resolution of the crystal structure of Rho and due to the fact that helices in Rho are closer to ideal \( \alpha \) helices than the helices in vSGLT.

<table>
<thead>
<tr>
<th>1F88 paper [78]</th>
<th>1F88 dssp</th>
<th>taken for modelling</th>
</tr>
</thead>
<tbody>
<tr>
<td>33-65</td>
<td>34-64</td>
<td></td>
</tr>
<tr>
<td>70-101</td>
<td>71-100</td>
<td></td>
</tr>
<tr>
<td>105-140</td>
<td>106-139</td>
<td></td>
</tr>
<tr>
<td>149-173</td>
<td>150-173</td>
<td></td>
</tr>
<tr>
<td>199-226</td>
<td>200-225</td>
<td></td>
</tr>
<tr>
<td>245-278</td>
<td>246-277</td>
<td></td>
</tr>
<tr>
<td>284-309</td>
<td>285-309</td>
<td></td>
</tr>
</tbody>
</table>

Table 4.4: Helix ends definition of Rho PDB# 1F88. "dssp" for DSSP. The overall resolution of the crystal structures #1F88 is 2.8 Å.

4.2.1.3 Modelling templates

Definition of helix ends of the proteins used as a modelling template of vSGLT can be found in Appendix B.1.

4.2.2 Comparison of vSGLT structures

The first vSGLT structure was crystallized in an inward-facing conformation with a resolution of \( \sim 3 \) Å(PDB # 3DH4, [38]). Nevertheless, some regions of the protein have poor resolution (\( \sim 3.7 \) Å), side chains could not be assigned and therefore, register with the amino acid sequence could not be established.

The second vSGLT structure also corresponds to an inward-facing conformation and is, in general, better resolved (\( \sim 2.7 \) Å, PDB # 2XQ2, [39]). However, the first eight N-terminal residues are still unresolved.

Comparisons of the two structures led to the conclusion that residue 4 (helix end) in PDB # 3DH4 structure corresponds to residue 11 in PDB # 2XQ2 structure as well as residue 18 (helix end) in PDB # 3DH4 corresponds to residue 25 in PDB # 2XQ2 (check Figure 4.3 and Table 4.3).
Chapter 4. Results and discussion

4.3 Modelling of protein architecture

The present chapter will introduce the reader into the problems with modelling under conditions of the underdetermination of the problem and will try to answer apparent questions with the help of two model proteins - vSGLT [39] and Rho [78]. Why exactly these ones? vSGLT is the closest relative of PutP for which the structure is known. Additionally, it contains a similar number of transmembrane helices and supposedly has a similar general architecture. In comparison, Rho with a smaller number of helices, which fit the cylindrical approximation very well, lets us explore the very basics of the algorithm.

4.3.1 Sparse distance constraints

To check the amount of information that is sufficient for modelling, vSGLT modelling was performed with 75 simulated constraints and tightness factor $c = 2$. The list of distances simulated with MMM 2013 comprised pairs analogous to those of PutP (distances between ends of the same helices) - Figure B.4. The result of this modelling is shown in Figure 4.4.

The result of modelling under the described conditions is disappointing - after including the simulated experimental-like constraints and internal constraints the algorithm is not able to predict a protein fold (overall rmsd for 13 helices in comparison to the real
Figure 4.4: Result of vSGLT modelling with 75 simulated constraints, $r = r_{\text{mean}} \pm 2\sigma_r$.
Superposition of the vSGLT central structure (transparent) and vSGLT target structure (opaque) highlighted for 13 constrained helices. Unconstrained helix XIV is not shown. The algorithm is not able to recognize the protein fold - rmsd vs real $r_{\text{real}} = 23.29$ Å (13 helices), ensemble rmsd $r_{\text{ensemble}} = 24.95$ Å. Red dots (structure)/pink dots (model) - cytoplasmic helix ends, blue points (structure)/violet points (model) - periplasmic helix ends.

structure $r_{\text{real}} = 23.29$ Å). The high ensemble rmsd indicates that the coarse-grained structure varies to an extent that would make any interpretation of the model futile.

But does this mean that the methodology is completely useless for the considered problem? In order to answer this question a run with 378 simulated distances ($n(2n - 1)$ for $n=14$ helices) was performed (the constraint list generated by MMM 2013 comprised all possible combinations of paired helix ends, not shown). Surprisingly, Figure 4.5 proves that the algorithm is still incapable of recognizing the protein fold - the rmsd vs real structure $r_{\text{real}} = 24.21$ Å and the ensemble rmsd is $r_{\text{ensemble}} = 28.58$ Å. Thus, failure of modelling is not only due to a lack of experimental constraints, but must have an additional reason.

4.3.2 Template data inclusion and its influence

The central structure of an ensemble modelled with the inclusion of template data is superimposed with the target protein structure of vSGLT in Figure 4.6. This central structure has an rmsd $r_{\text{real}} = 14.21$ Å from the true structure. Despite the rmsd is still poor, relative placement of the helices is now correct. Hence, the template constraints improved the result significantly.

Note that the helices that were constrained additionally by the template (for vSGLT, core helices are 2-11) have a much better agreement with the native structure than the non-core helices. To explain this phenomenon the modelling result is compared to the
Chapter 4. Results and discussion

Figure 4.5: Result of vSGLT modelling with 378 simulated constraints, $r = r_{mean} \pm 2\sigma_r$. Superposition of the vSGLT central structure (transparent) and vSGLT target structure (opaque) highlighted for all helices. The algorithm can not recognize the protein fold - rmsd vs real rmsd$_{real} = 24.21$ Å (14 helices), ensemble rmsd rmsd$_{ensemble} = 28.58$ Å. Red dots (structure)/pink dots (model) - cytoplasmic helix ends, blue points (structure)/violet points (model) - periplasmic helix ends.

Figure 4.6: Result of vSGLT modelling with 75 simulated constraints and template constraints, $r = r_{mean} \pm 2\sigma_r$. Superposition of the vSGLT central structure (transparent) and vSGLT target structure (opaque) highlighted for 13 constrained helices. The algorithm now recognizes the protein fold - rmsd vs real rmsd$_{real} = 14.21$ Å (13 helices), ensemble rmsd rmsd$_{ensemble} = 15.07$ Å. Red dots (structure)/pink dots (model) - cytoplasmic helix ends, blue points (structure)/violet points (model) - periplasmic helix ends.

LeuT template. Figure 4.7 shows how the core of a vSGLT model follows the fold of LeuT. The simulated constraints do not influence the model that much, it is largely determined by the template structure.
4.3.2.1 Decoy models

In order to check whether the approach would produce wrong architectures from wrong templates, such improper templates were generated by permutation of helices (decoy models). 1'000'000 decoys were tested for vSGLT modelling with the LeuT template for both constant tolerance of 6 Å as well as for a tolerance computed from the variation of experimental structures. Such modelling resulted in no decoys consistent with a given constraint list. This test shows that the algorithm can recognize whether the template constraints are consistent with experimental constraints. In other words, successful modelling with template constraints provides evidence that relative arrangement of helices is indeed the same as in the template. For PutP, it can thus be tested whether the transporter also has the ten-helix core structure.

4.3.2.2 Positioning of non-core helices

The topology in the core region of a protein is not exclusively determined by the measured distances, it is mostly a "copy" of the template core (Figure 4.7). On the other hand, modelling without any template gives a bunch of structures which are weakly correlated (high ensemble rmsd) and the protein fold cannot be reconstructed (Figure 4.4). Besides testing whether PutP has the ten-helix core fold, we can thus only hope to place the non-core helices relative to the ten-helix core. For this, optimal modelling conditions must be found.

4.3.3 Tightness parameter and its meaning

Runs for vSGLT were performed with the full set of simulated restraints (378) at various tightness factors $c$ ($c = 1$ - Figure B.9, $c = 0.6$ - Figure B.10 and $c = 0.58$ - Figure 4.8). That way a limiting value for the tightness parameter $c$ was found - for $c < 0.58$ embedding into Cartesian space invariably failed. A structure with best rmsd from
native that can be found within the ensemble generated with $c = 0.58$ has $\text{rmsd}_{\text{real}} = 7.17 \, \text{Å}$ (Figure B.11).

Table 4.5: Comparison of modelling runs of vSGLT with full set of simulated constraints (378) and no template data included.

<table>
<thead>
<tr>
<th>$c$</th>
<th>2</th>
<th>1</th>
<th>0.6</th>
<th>0.58</th>
</tr>
</thead>
<tbody>
<tr>
<td>success rate [%]</td>
<td>99</td>
<td>100</td>
<td>78</td>
<td>8</td>
</tr>
<tr>
<td>$\text{rmsd}_{\text{real}}$ (14 helices) [Å]</td>
<td>24.21</td>
<td>7.61</td>
<td>8.92</td>
<td>8.33</td>
</tr>
<tr>
<td>$\text{rmsd}_{\text{ensemble}}$ [Å]</td>
<td>28.58</td>
<td>12.35</td>
<td>10.38</td>
<td>9.87</td>
</tr>
</tbody>
</table>

Figure 4.8: Result of vSGLT modelling with 378 simulated constraints, $r = r_{\text{mean}} \pm 0.58 \cdot \sigma_r$. Superposition of the vSGLT central structure (transparent) and vSGLT target structure (opaque) highlighted for all helices. Rmsd vs real $\text{rmsd}_{\text{real}} = 8.33 \, \text{Å}$ (14 helices), ensemble rmsd $\text{rmsd}_{\text{ensemble}} = 9.87 \, \text{Å}$. Red dots (structure)/pink dots (model) - cytoplasmic helix ends, blue points (structure)/violet points (model) - periplasmic helix ends.

Table 4.5 shows that, as long as embedding has a reasonable success rate, both ensemble rmsd and rmsd from the real target structure tend to decrease with decreasing $c$. This suggests to work at the lowest tightness level $c$ where an ensemble of sufficient size can still be obtained. In the particular case of vSGLT, reducing $c$ to below 1 does no longer improve rmsd from the target.

At this point it was attempted to model the vSGLT structure once more without the template data, but with an adjusted tightness parameter. A run with 75 simulated constraints and tightened $c$ was performed. The results are shown in Figure 4.9. Despite the lower $c$-value the proper topology was not found. Therefore, it cannot be expected that the experimental constraint set for PutP allows for modelling without including additional template constraints.

The influence of the tightness parameter on the result was also explored for Rhodopsin. Similarly, a full set of constraints was used (91 for 7 transmembrane helices, the constraint list generated by MMM 2013 comprised all possible combinations of paired helix
Figure 4.9: Result of vSGLT modelling with 75 simulated constraints, $r = r_{\text{mean}} \pm 0.58 \cdot \sigma_r$. Superposition of the vSGLT central structure (transparent) and vSGLT target structure (opaque) highlighted for 13 modelled helices. Rmsd vs real $r_{\text{real}} = 23.83 \AA$ (13 helices), ensemble rmsd $r_{\text{ensemble}} = 28.30 \AA$. Red dots (structure)/pink dots (model) - cytoplasmic helix ends, blue points (structure)/violet points (model) - periplasmic helix ends.

ends, not shown). In this case, the tightest $c$ possible for vSGLT - 0.58 (Figure 4.10) allowed embedding of almost all generated structures, which is probably due to the smaller size of the protein and the more straight helices. A structure with best rmsd vs native that can be found within this ensemble has $r_{\text{real}} = 6.88 \AA$ (Figure B.26). Additional runs were performed with $c = 0.328$ (Figure B.21) and $c = 0.31$ (Figure 4.11). Similar to the case of vSGLT, the success rate of embedding decreases when $c$ is further decreased (Table 4.6). In contrast to the case of vSGLT, decreasing tightness improves neither rmsd vs the target structure nor the ensemble rmsd.

Structures with best agreement vs native that can be found within generated ensembles have $r_{\text{real}}$ of 6.38 Å for $c = 0.328$ (Figure B.27) and 7.42 Å for $c = 0.31$ (Figure B.28).

Table 4.6: Comparison of modelling runs of Rho with full set of simulated constraints (91) and no template data included.

<table>
<thead>
<tr>
<th>$c$</th>
<th>0.58</th>
<th>0.328</th>
<th>0.31</th>
</tr>
</thead>
<tbody>
<tr>
<td>success rate [%]</td>
<td>99</td>
<td>91</td>
<td>3</td>
</tr>
<tr>
<td>$r_{\text{real}}$ [Å]</td>
<td>8.32</td>
<td>7.25</td>
<td>8.81</td>
</tr>
<tr>
<td>$r_{\text{ensemble}}$ [Å]</td>
<td>13.63</td>
<td>12.85</td>
<td>13.67</td>
</tr>
</tbody>
</table>
4.3.3.1 Combining template constraints and reduced $c$

For vSGLT, a limiting tightness $c = 0.58$ giving 8% of successfully embedded structures was found. With this tightness, the variance of the structures in the ensemble was reduced. It was then attempted to position the non-core helices with respect to the core by using 75 simulated constraints as well as template data (based on PDB #2A65) in the calculations.

The result presented in Figure 4.12 demonstrates how template inclusion improves the agreement of modelling results with the target structure (compare with 4.9) - template addition reduces $\text{rmsd}_{\text{real}}$ by nearly 6 Å and $\text{rmsd}_{\text{ensemble}}$ by 10 Å. A structure with best $\text{rmsd}$ from the target that is found in this ensemble has $\text{rmsd}_{\text{real}} = 9.56$ Å for 13 constrained helices (Figure B.12).

On the other hand, once template data are included, decreasing the tightness parameter value from 2 (compare with Figure 4.6) to 0.58 does not lead to closer approach to the
Chapter 4. **Results and discussion**

Figure 4.12: Result of vSGLT modelling with 75 simulated constraints, \( r = r_{\text{mean}} \pm 0.58 \cdot \sigma_r \) and template constraints included. Superposition of the vSGLT central structure (transparent) and vSGLT target structure (opaque) highlighted for 13 modelled helices.

Rmsd vs real rmsd \( \text{rmsd}_{\text{real}} = 18.44 \, \text{Å} \) (13 helices), ensemble rmsd \( \text{rmsd}_{\text{ensemble}} = 18.00 \, \text{Å} \). Red dots (structure)/pink dots (model) - cytoplasmic helix ends, blue points (structure)/violet points (model) - periplasmic helix ends.

For vSGLT with a full constraints set the approach can reach an accuracy of only about 8-9 Å in comparison with the target structure, for vSGLT with a constraints set of similar size as for PutP (75 distances) and template constraints this number is about 18 å.

**Table 4.7: Comparison of modelling runs of vSGLT with 75 simulated constraints under different conditions.**

<table>
<thead>
<tr>
<th>( c )</th>
<th>( 2 )</th>
<th>0.58</th>
</tr>
</thead>
<tbody>
<tr>
<td>template</td>
<td>yes</td>
<td>yes</td>
</tr>
<tr>
<td>success rate [%]</td>
<td>64</td>
<td>89</td>
</tr>
<tr>
<td>( \text{rmsd}_{\text{real}} ) (13 helices) [Å]</td>
<td>14.21</td>
<td>18.44</td>
</tr>
<tr>
<td>( \text{rmsd}_{\text{ensemble}} ) [Å]</td>
<td>15.07</td>
<td>18.00</td>
</tr>
</tbody>
</table>

As stated in the previous section, the final aim of the modelling with a template is positioning of non-core helices. Already a quick glance on Figure 4.12 shows that this is not possible with the used constraint set either. This result will be discussed more in detail in Section 4.3.7.

### 4.3.4 Including labels into modelling

For vSGLT with a full constraints set the approach can reach an accuracy of only about 8-9 Å in comparison with the target structure, for vSGLT with a constraints set of similar size as for PutP (75 distances) and template constraints this number is about 18 å.
18-19 Å. In order to test, whether this is clue to flexibility of the labels, the simulated distances between helix ends were used in the modelling.

A tightness parameter $c = 0.58$ was found for modelling with labels attached as the minimum possible $c$ and hence, vSGLT simulations were run with $\sigma_r = 5.8$ Å corresponding to $c = 0.58$ for distances between the helix ends. The result is shown in Figure 4.13.

![Figure 4.13: Result of vSGLT modelling with 378 simulated constraints between helix ends, $r = r_{\text{mean}} \pm 5.8$ Å. Superposition of the vSGLT central structure (transparent) and vSGLT target structure (opaque) highlighted for 14 modelled helices. Rmsd vs real $\text{rmsd}_{\text{real}} = 4.98$ Å (14 helices), ensemble rmsd $\text{rmsd}_{\text{ensemble}} = 4.00$ Å. Red dots (structure)/pink dots (model) - cytoplasmic helix ends, blue points (structure)/violet points (model) - periplasmic helix ends.](image)

With helix end constraints rather than label constraints the success rate increased from 8 to 92% (Table 4.8). Obviously, the reduced uncertainty of a model without flexible labels reduces the probability that embedding fails. Hence, conditions under which only ~10% of the structures are successfully embedded were searched. At $\sigma_r = 3.8$ Å a success rate of 4% was found.

The result of this run is shown in Figure 4.14. A structure with best $\text{rmsd}$ vs native that is found in the ensemble has $\text{rmsd}_{\text{real}} = 2.87$ Å (Figure B.13). Runs with both $\sigma_r = 5.8$ Å and $\sigma_r = 3.8$ Å gave satisfactory results.

Keeping the same tightness parameter ($c = 0.58$ corresponding to $\sigma_r = 5.8$ Å) with helix end constraints improves agreement with the target structure already significantly (Table 4.8) - from 8.33 to 4.98 Å. The ensemble is also much tighter - 4.00 instead of 9.87 Å. At a similar fraction of successful embeddings, an even bigger improvement is observed - $\text{rmsd}_{\text{real}}$ drops from 8.33 to 4.09 Å, $\text{rmsd}_{\text{ensemble}}$ from 9.87 to 2.55 Å. A structure with $\text{rmsd}_{\text{real}} = 2.87$ from the native is the structure with best agreement vs target that could be found in this ensemble. An $\text{rmsd}$ of about 4-5 Å is acceptable for such a coarse-grained model and would allow for biological interpretation. Apparently,
Chapter 4. Results and discussion

Figure 4.14: Result of vSGLT modelling with 378 simulated constraints between helix ends, $r = r_{\text{mean}} \pm 3.8 \text{ Å}$. Superposition of the vSGLT central structure (transparent) and vSGLT target structure (opaque) highlighted for 14 modelled helices. \text{Rmsd vs real } r_{\text{real}} = 4.09 \text{ Å} (14 helices), ensemble r_{\text{ensemble}} = 2.55 \text{ Å}. Red dots (structure)/pink dots (model) - cytoplasmic helix ends, blue points (structure)/violet points (model) - periplasmic helix ends.

it is the flexibility of the label and its distance from the helix ends that lead to failure of the approach with simulated label-to-label constraints.

Table 4.8: Comparison of modelling runs of vSGLT with 378 simulated constraints under different conditions.

<table>
<thead>
<tr>
<th>constraint type</th>
<th>labels</th>
<th>helix ends</th>
</tr>
</thead>
<tbody>
<tr>
<td>$c$ or $\sigma_{\tau,\text{nm}}$ (no label)</td>
<td>2</td>
<td>0.58</td>
</tr>
<tr>
<td>success rate [%]</td>
<td>99</td>
<td>8</td>
</tr>
<tr>
<td>$r_{\text{real}}$ (14 helices) [Å]</td>
<td>24.21</td>
<td>8.33</td>
</tr>
<tr>
<td>$r_{\text{ensemble}}$ [Å]</td>
<td>28.58</td>
<td>9.87</td>
</tr>
</tbody>
</table>

Modelling was repeated for the case when template data and 75 constraints are included in order to check if sparse helix end constraints are also sufficient to find the proper topology (Figure 4.15). The overall fit to the real structure is much better than for calculations with the labels and the ensemble is much tighter (Table 4.9). Although the r_{\text{rmsd}} to the target structure increases to 9.5 Å, general topology can still be recognized. The structure ensemble contains a structure that has the best agreement with the target and $r_{\text{rmsd,real}} = 6.14$ Å for 13 modelled helices (Figure B.14). Placement of non-core helices is reasonable. A more in depth discussion can be found in Section 4.3.7.
Figure 4.15: Result of vSGLT modelling with 75 simulated constraints between helix ends, \( r = r_{mean} \pm 5.8 \text{ Å} \), template data included. Superposition of the vSGLT central structure (transparent) and vSGLT target structure (opaque) highlighted for 13 constrained helices. Rmsd vs real \( \text{rmsd}_{\text{real}} = 9.53 \text{ Å} \) (13 helices), ensemble \( \text{rmsd}_{\text{ensemble}} = 5.12 \text{ Å} \). Red dots (structure)/pink dots (model) - cytoplasmic helix ends, blue points (structure)/violet points (model) - periplasmic helix ends.

Table 4.9: Comparison of modelling runs of vSGLT with 75 simulated constraints and template information with or without the labels.

<table>
<thead>
<tr>
<th>Constraint type</th>
<th>0.58</th>
</tr>
</thead>
<tbody>
<tr>
<td>( c ) or ( \sigma_{r_{nm}} ) (no label)</td>
<td>constraint type</td>
</tr>
<tr>
<td>( c ) or ( \sigma_{r_{nm}} ) (no label)</td>
<td>labels</td>
</tr>
<tr>
<td>success rate [%]</td>
<td>99</td>
</tr>
<tr>
<td>( \text{rmsd}_{\text{real}} ) (13 helices) [Å]</td>
<td>18.44</td>
</tr>
<tr>
<td>( \text{rmsd}_{\text{ensemble}} ) [Å]</td>
<td>18.00</td>
</tr>
</tbody>
</table>

Agreement with the native vSGLT structure improves significantly if the label constraints are replaced by helix end constraints, both with a full set of constraints (Table 4.8) and with a set of similar size as the set of experimental constraints for PutP (Table 4.9). For the full set of distances and a \( c \) parameter giving a few percent of successfully embedded structures, the real structure is approached with an \( \text{rmsd}_{\text{real}} \) of 4.09 Å and a structure with smallest \( \text{rmsd} \) vs native existing in this ensemble has \( \text{rmsd}_{\text{real}} = 2.87 \text{ Å} \). The reason for the remaining deviation even with a full set of accurate helix end restraints and minimal possible tightness intervals may be that in vSGLT not all helices are well approximated by a cylindrical model. A bend in some of the vSGLT helices is outside the scope of the cylinder model (Figure 4.16).
Chapter 4. Results and discussion

To test this assumption modelling for Rho with a full set of simulated constraints between helix ends was done.

σ_r = 5.8 Å is such a loose parameter for a smaller protein with straight helices as Rho that practically all built structures get embedded to the Cartesian space. The result of this run is shown in Figure 4.17. Agreement of the model with the cylindrical structure representation is good and rmsd of 5.38 Å vs the target structure is comparable with that of vSGLT modelling with the same tightness (check Figure 4.13). A structure with
rmsd_{real} = 3.23 Å is the structure with best agreement with the target structure that can be found in this ensemble (Figure B.29). A set of simulation runs with different $\sigma_r$ was performed in order to find conditions where the success rate was $\sim$10 %. Results can be found in Figures: $\sigma_r = 3$ Å - B.22, $\sigma_r = 2$ Å - B.23, $\sigma_r = 1$ Å - B.24, $\sigma_r = 0.143$ Å - B.25 and $\sigma_r = 0.142$ Å - 4.18. The same trend is observed - the smaller $c$ the tighter the ensemble is and the better agreement with the target structure (Table 4.10).

A stepwise search led to the value of $\sigma_r = 0.142$ Å or $\sigma_r = 0.143$ Å resulting in 3 or 15% of successful embeddings, respectively. One has to realize that this tightness parameter leaves practically no space for variations allowing each distance to be varied by only 0.1 Å. Accordingly, the ensemble rmsds of 0.01 - 0.02 Å confirm that all structures in the ensemble are practically the same. Indeed, a structure with the smallest rmsd vs the target structure that can be found within this ensemble has rmsd_{real} = 2.78 Å (Figure 4.18).

For the modelling with $\sigma_r = 1$ Å the ensemble contains a structure with minimum rmsd_{real} = 2.47 Å vs the target structure (Figure B.30). At the end, agreement with the target structure reaches 2.70 - 2.78 Å, whereas it could reach optimally 2.47 Å, which indicates that the ultimate accuracy of the approach is not better than that.

Figure 4.18: Result of Rho modelling with 91 constraints between helix ends, $r = r_{mean} \pm 0.142$ Å. Superposition of the Rho central structure (transparent) and Rho real structure (opaque) highlighted for all helices. Rmsd vs target rmsd_{real} = 2.78 Å, ensemble rmsd rmsd_{ensemble} = 0.01 Å. Red dots (structure)/pink dots (model) - cytoplasmic helix ends, blue points (structure)/violet points (model) - periplasmic helix ends.

Table 4.10: Comparison of modelling runs of Rho with 91 simulated constraints between helix ends.

<table>
<thead>
<tr>
<th>$\sigma_{r,nm}$</th>
<th>0.58</th>
<th>0.3</th>
<th>0.2</th>
<th>0.1</th>
<th>0.0143</th>
<th>0.0142</th>
</tr>
</thead>
<tbody>
<tr>
<td>success rate [%]</td>
<td>99</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>15</td>
<td>3</td>
</tr>
<tr>
<td>rmsd_{real} [Å]</td>
<td>5.38</td>
<td>3.49</td>
<td>2.90</td>
<td>2.70</td>
<td>2.78</td>
<td>2.78</td>
</tr>
<tr>
<td>rmsd_{ensemble} [Å]</td>
<td>11.02</td>
<td>9.24</td>
<td>5.97</td>
<td>0.71</td>
<td>0.02</td>
<td>0.01</td>
</tr>
</tbody>
</table>

Even for a small protein with relatively straight helices and true distances between helix ends known (from PDB file) the approach results in an accuracy that would not allow for interpretation of the structure in atomic detail. Yet, architecture of the helix bundle is nicely reproduced under such conditions. Taken the labels into consideration leads
to a quality drop. Table 4.11 summarizes the results: for 100 % successful embeddings label attachment causes 3-5 Å quality drop (compare with Figure 4.10), for tight $c$ allowing only a few percent structures to pass, 5 Å (compare with Figure 4.11). This finding again indicates that more rigid labels or a better label description are required.

Table 4.11: Comparison of modelling runs of Rho with 91 simulated constraints.

<table>
<thead>
<tr>
<th>constraint type</th>
<th>helix ends</th>
<th>spin labels</th>
</tr>
</thead>
<tbody>
<tr>
<td>$c$ or $\sigma_{r,nm}$ (no label)</td>
<td>0.58 0.1 0.0142</td>
<td>0.58 0.31</td>
</tr>
<tr>
<td>success rate [%]</td>
<td>99 100 3</td>
<td>99 3</td>
</tr>
<tr>
<td>rmsd$_{real}$ [Å]</td>
<td>5.38 2.70 2.78</td>
<td>8.32 8.81</td>
</tr>
<tr>
<td>rmsd$_{ensemble}$ [Å]</td>
<td>11.02 0.71 0.01</td>
<td>13.63 13.67</td>
</tr>
</tbody>
</table>

4.3.5 Pre- and post-selection of the clustered structures

Since the modelling results obtained under the limiting tight conditions are not satisfactory, other ways of improving the approach were sought. The task was to find optimal selection conditions that allowed for picking structures that better fitted the target structure ensemble.

4.3.5.1 Reduction of the ensemble based on a geometrical criterium of selection

An algorithm was tested for the best reduction rate as described in 3.1, firstly for Rho modelling with full sets of constraints between helix ends or between labels. The results summary is collected in Table 4.12.

Table 4.12: Impact of re-clustering of structures chosen according to best rmsd vs the central structure of the original ensemble on modelling result quality shown for Rho with 91 simulated constraints.

<table>
<thead>
<tr>
<th>constraint type</th>
<th>helix ends</th>
<th>spin labels</th>
</tr>
</thead>
<tbody>
<tr>
<td>$c$ or $\sigma_{r,nm}$ (no label)</td>
<td>0.58 0.3 0.1</td>
<td>0.58 0.31</td>
</tr>
<tr>
<td>success rate</td>
<td>99 100 100</td>
<td>99 3</td>
</tr>
<tr>
<td>rmsd$_{real}$ [Å]</td>
<td>all 5.38 3.49 2.70</td>
<td>8.32 8.81</td>
</tr>
<tr>
<td></td>
<td>70% 4.19 3.15 2.70</td>
<td>9.70 8.76</td>
</tr>
<tr>
<td></td>
<td>50% 5.39 3.43 2.70</td>
<td>8.56 9.77</td>
</tr>
<tr>
<td></td>
<td>30% 4.69 3.22 2.70</td>
<td>9.51 8.52</td>
</tr>
<tr>
<td>rmsd$_{ensemble}$ [Å]</td>
<td>all 11.02 9.24 0.71</td>
<td>13.63 13.67</td>
</tr>
<tr>
<td></td>
<td>70% 8.01 3.00 0.67</td>
<td>10.08 8.74</td>
</tr>
<tr>
<td></td>
<td>50% 4.67 2.21 0.64</td>
<td>9.10 8.45</td>
</tr>
<tr>
<td></td>
<td>30% 4.74 2.22 0.62</td>
<td>8.98 7.90</td>
</tr>
</tbody>
</table>
Generally, no consistent and significant improvement was found. In the case where the constraints between helix ends were already tight \((\sigma_r = 1 \text{ Å})\), the central structure of the ensemble did not change. A graphical overlap of models and the cylindrical representation of the Rho target structure can be found in the Appendix \((\sigma_r = 5.8 \text{ Å} - B.31, \sigma_r = 3 \text{ Å} - B.32 \text{ and } \sigma_r = 1 \text{ Å} - B.33)\).

Cluster rmsd plots for Rho modelling with constraints between helix ends and different \(c\) parameters are shown in Figure 4.19. The ensemble becomes more "uniform" after cutting all "tail" structures that are far from the central one. When the primary ensemble is tight and does not reveal a "stepwise" cluster rmsd plot, cutting the ensemble leads to only a very slight reduction of ensemble rmsd. The ensemble rmsd for reduced ensembles for all reduction levels are given in Table 4.12.

![Figure 4.19: Cluster rmsd plots for modelling of Rho with constraints between helix ends and \(A \sigma_r=5.8 \text{ Å}, B \sigma_r=3 \text{ Å}, C \sigma_r=1 \text{ Å}. Most left plot: clustering of all embedded structures. Next plots: geometrical post-selection of the embedded structures done by choosing ones being the closest to the central structure and re-clustering them - 70% (green), 50% (red) and 30% (cyan) taken.](image)

A similar conclusion applies for modelling constraints between spin labels (graphical modelling results to be found in Appendix: \(c = 0.58 - B.34, c = 0.31 - B.35\)). Removing structures furthest from the central one does not clearly improve agreement of the central structure of the ensemble with the target structure, but it tightens the ensemble (Table 4.12). Removal of the "tail" of clustered structures helps to find a more uniform ensemble, where the central structure is not far from any other in the group (Figure 4.20).

In conclusion, reduction of the clustered ensemble based on the geometrical criterion (structures closest to the central one of the original ensemble are being kept) does not allow for finding a new central structure that significantly better fits the cylindrical model of Rho. The optimal reduction rate depends on the specific ensemble and thus on the \(c\) parameter used. "Cutting the tail", which means removing structures being
Figure 4.20: Cluster rmsd plots for modelling of Rho with label constraints and tightness factor of A $c = 0.58$, B $c = 0.31$. Most left plot: clustering of all embedded structures. Next plots: geometrical post-selection of the embedded structures done by choosing ones being the closest to the central structure and re-clustering them - 70% (green), 50% (red) and 30% (cyan) taken.

part of the second step on the cluster rmsd plot, can be recommended as it leads to an ensemble rmsd that better represents the rmsd of the central structure from the target structure.

However, for vSGLT modelling, reducing the ensemble does not lead to significant improvement - the central structure remains the same or changes only slightly and tightening of the ensemble is very minor (Table 4.13).

Table 4.13: Impact of re-clustering of structures chosen according to best rmsd vs the central structure of the original ensemble on modelling result quality shown for vSGLT with 378 simulated constraints.

<table>
<thead>
<tr>
<th>constraint type</th>
<th>helix ends</th>
<th>spin labels</th>
</tr>
</thead>
<tbody>
<tr>
<td>$c$ or $\sigma_{r,nm}$ (no label)</td>
<td>0.58</td>
<td>0.38</td>
</tr>
<tr>
<td>success rate</td>
<td>92</td>
<td>4</td>
</tr>
<tr>
<td>$\text{rmsd}_{\text{real}}$ [Å]</td>
<td>4.98</td>
<td>4.09</td>
</tr>
<tr>
<td>all</td>
<td>24.21</td>
<td>8.33</td>
</tr>
<tr>
<td>70%</td>
<td>4.98</td>
<td>4.03</td>
</tr>
<tr>
<td>50%</td>
<td>4.98</td>
<td>4.03</td>
</tr>
<tr>
<td>30%</td>
<td>4.89</td>
<td>4.09</td>
</tr>
<tr>
<td>$\text{rmsd}_{\text{ensemble}}$ [Å]</td>
<td>4.00</td>
<td>2.55</td>
</tr>
<tr>
<td>all</td>
<td>28.58</td>
<td>9.87</td>
</tr>
<tr>
<td>70%</td>
<td>3.61</td>
<td>2.14</td>
</tr>
<tr>
<td>50%</td>
<td>3.39</td>
<td>1.94</td>
</tr>
<tr>
<td>30%</td>
<td>3.16</td>
<td>1.78</td>
</tr>
</tbody>
</table>
4.3.5.1.1 Swapped helices

Further analysis of the simulated ensembles revealed some specific properties of the approach. Namely, if the modelling was run with a "loose" $c$ parameter, with constraints between either helix ends or labels, the structures "in the tail" of the cluster rmsd plot contain pairs of swapped helices. Such solutions are allowed in distance space as swapping can be balanced by tilting other helices, nevertheless embedding to the Cartesian space gives completely wrong solutions.

An analysis of the presented issue is shown on the example of Rho modelling with $\sigma_r = 5.8$ Å and constraints between helix ends (Figure 4.21).

Figure 4.21: Modelling of Rho with $\sigma_r = 5.8$ Å and constraints between helix ends, A Cluster rmsd plot of all embedded structures, structures with smallest rmsd vs the central ensemble (cyan) and structures "in the tail" having worst rmsd vs ensemble (red), B Rmsd vs target structure of part of the ensemble and its central structure for the cyan selection and structure with worst rmsd$_{real}$ for red part, C closer look at the last structure "in the tail", cytoplasmic and periplasmic view, respectively.

Figure 4.21 A shows the cluster rmsd plots. Division of the structures into two groups and their further clustering reveals that structures from the first part of the plot are sharing a similar fold (small rmsd$_{ensemble}$ values), the same happens with the structures from second part of the plot. The "step" in the cluster plot of all structures is then a border between two significantly distinct groups of structures, where the first group corresponds to the true protein fold (low rmsd$_{real}$) and the second one to an completely unrealistic fold with swapped helices (high rmsd$_{real}$ values) - Figure 4.21 B.

Figure 4.21 C zooms into the structure with worst rmsd vs target structure. A combination of tilting the helices and swapping helix pairs allows to fulfill the distance boundaries. Removal of such structures significantly improves the result (check Table 4.14 and Figure B.36 vs 4.17 for the visualization).
Chapter 4. **Results and discussion**

60

Figure 4.22: Modelling of Rho with \( c = 0.58 \) and constraints between labels. **A** Cluster rmsd plot of all embedded structures, structures with smallest rmsd vs the central ensemble (cyan) and structures "in the tail" having worst rmsd vs ensemble (red). **B** Rmsd vs target structure of part of the ensemble and its central structure for the cyan selection and structure with worst rmsd_{real} for red part, **C** closer look at the last structure "in the tail", cytoplasmic and periplasmic view, respectively.

Figure 4.22 presents a similar analysis with constraints between spin labels. Though the overall conclusions remain similar, here the situation is not so unambiguous anymore - cluster plots of grouped structures show high rmsd_{ensemble} proving that the group members are quite diverse (Figure 4.22 **A**). Also rmsd_{real} plots in Figure 4.22 **B** suggest that the structures from the first group do not give much better agreement with the cylindrical Rho representation. In this case, removal of the "tail" does not improve agreement of the central structure (compare Figure B.37 and 4.10). However, it does not do any damage either and it does lead to the ensemble rmsd better representing the rmsd of the central structure vs the target structure (Table 4.14). Hence, such "tail" removal can be recommended.

**Table 4.14:** Impact of removing wrong architectures from the ensemble of Rho modelled with 91 simulated constraints.

<table>
<thead>
<tr>
<th>constraint type</th>
<th>helix ends</th>
<th>labels</th>
</tr>
</thead>
<tbody>
<tr>
<td>( c ) or ( \sigma_{r,nm} ) (no label)</td>
<td>0.58</td>
<td>0.58</td>
</tr>
<tr>
<td>success rate</td>
<td>99</td>
<td>99</td>
</tr>
<tr>
<td>rmsd_{real}</td>
<td>5.38</td>
<td>8.32</td>
</tr>
<tr>
<td>rmsd_{real} after removing &quot;tail&quot;</td>
<td>3.52</td>
<td>8.84</td>
</tr>
<tr>
<td>rmsd_{ensemble}</td>
<td>11.02</td>
<td>13.67</td>
</tr>
<tr>
<td>rmsd_{ensemble} after removing &quot;tail&quot;</td>
<td>4.64</td>
<td>9.19</td>
</tr>
</tbody>
</table>
For the tight $c$ parameter used for calculations, the cluster plot has no more "step" and all structures in the ensemble converge to a proper fold. To illustrate that simulation runs for Rho with constraints between helix ends and $\sigma_r = 1 \, \text{Å}"$ ("smooth" cluster rmsd plot) and $\sigma_r = 3 \, \text{Å}"$ (still some "step" on the cluster rmsd plot) are analysed. Structures with the worst cluster rmsd are shown in Figure 4.23 A and B, respectively.

Figure 4.23: A "step" in the cluster rmsd plot implies existence of structures with swapped helices for both Rho and vSGLT, for the cases of constraints between helix ends and between labels. Cluster rmsd plot and superposition of structure with worst rmsd$_{ensemble}$ and the cylindrical representation of a target structure for modelling of A Rho, constraints between helix ends, $\sigma_r = 1 \, \text{Å}$, B Rho, constraints between helix ends, $\sigma_r = 3 \, \text{Å}$, C Rho, constraints between labels, $c = 0.31$, D vSGLT, constraints between helix ends, $\sigma_r = 5.8 \, \text{Å}$, E vSGLT, constraints between labels, $c = 0.58$, are shown. Closer look at the last structure "in the tail" from cytoplasmic and periplasmic view for improper architectures presented.

For constraints between helix ends, a sufficiently tight $\sigma_r$ to avoid the "step" on the cluster plot was not searched, as $cc = 0.31$ was giving only 3% of successfully embedded structures. In this case, the cluster plot shows still double-cluster behaviour and part of the structures present improper fold - Figure 4.23 C.
These conclusions are confirmed by modelling of vSGLT. Smooth cluster rmsd plots obtained for tight $\sigma_r$ do not show a grouping into two significantly different cluster fractions for the case of constraints between helix ends (Figure 4.23 D). As looking for the tightest $\sigma_r$ was the primary aim of these simulation runs, no data with very loose $\sigma_r$ resulting in a "step" in the cluster rmsd plot are available here. For simulations with constraints between spin labels (Figure 4.23 E) $c = 0.58$ gave a few structures with wrong architecture. Ensemble reduction removes these structures.

### 4.3.5.2 EPR pre-selection

Table 4.15: Impact of clustering of structures chosen according to the best rmsd vs the constraint list on quality of modelling results shown for Rho with 91 simulated constraints.

<table>
<thead>
<tr>
<th>constraint type</th>
<th>helix ends</th>
<th>labels</th>
</tr>
</thead>
<tbody>
<tr>
<td>$c$ or $\sigma_{\text{rms}}$ (no label)</td>
<td>0.58</td>
<td>0.3</td>
</tr>
<tr>
<td>success rate</td>
<td>99</td>
<td>100</td>
</tr>
<tr>
<td>$\text{rmsd}_{\text{real}}$ [Å]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>all</td>
<td>5.38</td>
<td>3.49</td>
</tr>
<tr>
<td>70%</td>
<td>5.38</td>
<td>2.93</td>
</tr>
<tr>
<td>50%</td>
<td>4.19</td>
<td>2.93</td>
</tr>
<tr>
<td>30%</td>
<td>4.19</td>
<td>2.88</td>
</tr>
<tr>
<td>$\text{rmsd}_{\text{ensemble}}$ [Å]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>all</td>
<td>11.02</td>
<td>9.24</td>
</tr>
<tr>
<td>70%</td>
<td>10.90</td>
<td>9.35</td>
</tr>
<tr>
<td>50%</td>
<td>11.07</td>
<td>9.34</td>
</tr>
<tr>
<td>30%</td>
<td>10.94</td>
<td>9.36</td>
</tr>
</tbody>
</table>

Table 4.15 summarizes the results of EPR pre-selection on Rho modelling. Generally speaking, such selection does not influence results much and brings either no improvement or only a very slight non-systematic improvement of $\text{rmsd}_{\text{real}}$ (graphical comparison vs target structure is presented on Figure B.38 throughout B.41). The impact on Rho modelling with label constraints and $c = 0.58$ is not shown, as the central structure does not change in comparison to clustering the full ensemble - 4.10). The impact on ensemble tightening is also rather negligible.

Results for reducing the vSGLT ensemble on the basis of best agreement with the simulated distance constraints are collated in Table 4.16. Also in this case, EPR pre-selection does not cause improvement of the rmsd of the central structure vs the target structure or of the rmsd of the ensemble.

As visualized in Figure 4.24, best agreement with the constraint list does not correlate with the best correspondence to the target structure (compare the first and the second column). Sorting $\text{rmsd}_{\text{real}}$ according to the indices of the best match with the constraint list for both distances between labels and the helix ends, does not separate the structures into two groups with better or worse agreement with the target one (Figure 4.24, second throughout last column). Hence, the approach of reducing the structural ensemble according to the best agreement with the constraint list does not improve for the fraction of structures that are close to the target structure.
Table 4.16: Impact of clustering of structures chosen according to best rmsd vs the constraint list on quality of modelling results for vSGLT with 378 simulated constraints.

<table>
<thead>
<tr>
<th>constrain type</th>
<th>helix ends</th>
<th>labels</th>
</tr>
</thead>
<tbody>
<tr>
<td>c or $\sigma_{r,nm}$ (no label)</td>
<td>0.58</td>
<td>0.38</td>
</tr>
<tr>
<td>success rate</td>
<td>92</td>
<td>4</td>
</tr>
<tr>
<td>$\text{rmsd}_{\text{real}}$ [Å]</td>
<td>all</td>
<td>4.98</td>
</tr>
<tr>
<td></td>
<td>70%</td>
<td>4.64</td>
</tr>
<tr>
<td></td>
<td>50%</td>
<td>4.64</td>
</tr>
<tr>
<td></td>
<td>30%</td>
<td>4.64</td>
</tr>
<tr>
<td>$\text{rmsd}_{\text{ensemble}}$ [Å]</td>
<td>all</td>
<td>4.00</td>
</tr>
<tr>
<td></td>
<td>70%</td>
<td>3.84</td>
</tr>
<tr>
<td></td>
<td>50%</td>
<td>3.80</td>
</tr>
<tr>
<td></td>
<td>30%</td>
<td>3.77</td>
</tr>
</tbody>
</table>

4.3.5.3 Pre-selection based on the simulated distance distribution

Pre-selection based on agreement of the distances within a structure with the simulated distance distributions was supposed to better select the structures closest to the target structure. Unfortunately, such an improved ensemble reduction scheme of an ensemble does not lead to any improvement, neither in agreement of the central structure with the target structure nor for tightening the ensemble, see Table 4.17.

Table 4.17: Impact of clustering of structures chosen according to the best agreement with simulated distance distributions on modelling result quality shown for Rho with 91 simulated label constraints.

<table>
<thead>
<tr>
<th>c</th>
<th>0.58</th>
<th>0.31</th>
</tr>
</thead>
<tbody>
<tr>
<td>success rate</td>
<td>99</td>
<td>3</td>
</tr>
<tr>
<td>$\text{rmsd}_{\text{real}}$ [Å]</td>
<td>all</td>
<td>8.32</td>
</tr>
<tr>
<td></td>
<td>70%</td>
<td>8.89</td>
</tr>
<tr>
<td></td>
<td>50%</td>
<td>8.42</td>
</tr>
<tr>
<td></td>
<td>30%</td>
<td>9.65</td>
</tr>
<tr>
<td>$\text{rmsd}_{\text{ensemble}}$ [Å]</td>
<td>all</td>
<td>13.63</td>
</tr>
<tr>
<td></td>
<td>70%</td>
<td>13.71</td>
</tr>
<tr>
<td></td>
<td>50%</td>
<td>13.97</td>
</tr>
<tr>
<td></td>
<td>30%</td>
<td>14.01</td>
</tr>
</tbody>
</table>

4.3.6 Penalty functions

The penalty functions may be specific to a certain class of membrane proteins. Since Sale \textit{et al.} [77] based their approach on an analysis of rhodopsin-like protein structures, usefulness of the penalty functions was first tested for Rho, assuming constraints between helix ends and with a "loose" standard deviation put on distances, $\sigma_r = 5.8$
Figure 4.24: Reduction based on best agreement with the constraint list of the Rho ensemble modelled with different tightness and 91 simulated constraints does not improve agreement between the central structure and the target structure. Left: rmsd between distances within a model and the constraint list, sorted ascending; second left: rmsd vs the target structure rmsd_{real} sorted according to the structure indices that were extracted from ascendent sorting of the structures based on rmsd vs constraints list (first column); subsequent plots: rmsd vs the target structure rmsd_{real} shown for all embedded structures, 70%, 50% and 30% of the embedded structures with the best agreement with the constraint list (colors agree with the arrows): A constraints between helix ends, $\sigma_r = 5.8 \text{ Å}$, B constraints between helix ends, $\sigma_r = 3 \text{ Å}$, C constraints between helix ends, $\sigma_r = 1 \text{ Å}$, D constraints between the labels, $c = 0.58$, E constraints between the labels, $c = 0.31$.

Å, which provides a sufficiently broad ensemble for detecting a potential improvement (Figure 4.25).

Cluster rmsd in this case indeed seems to correlate slightly with $P_{apv}$, $P_{mpv}$, $P_{plv}$, $P_{vec}$ and possibly also $P_{dist}$. For the other functions, no significant correlation is observed. Despite these correlations, a reduction of the ensemble based on these penalty functions does not improve agreement of the central structure with the target structure and does not significantly reduce ensemble rmsd (Table 4.18).

For Rho modelling based on label constraints, not even a correlation between any penalty function and rmsd was found (Figure 4.26).

For completeness, correlation between the penalty functions and rmsd was also tested for vSLGT modelling with helix ends constraints. For ”loose” $\sigma_r$, modelling runs with a full
Chapter 4. Results and discussion

65

Figure 4.25: Individual penalty functions sorted according to the structure indices that were extracted from ascendent sorting of the structures based on cluster rmsd for Rho modelling with 91 simulated distances between helix ends and $\sigma_r = 5.8 \text{ Å}$.

Table 4.18: Impact of clustering of structures chosen according to the smallest mentioned penalty value on modelling result quality shown for Rho with 91 simulated constraints between helix ends, $\sigma_r = 5.8 \text{ Å}$.

<table>
<thead>
<tr>
<th>50% best of...</th>
<th>all</th>
<th>EPR</th>
<th>cluster</th>
<th>$P_{vec}$</th>
<th>$P_{compact}$</th>
<th>$P_{contact}$</th>
<th>$P_{ave}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>rmsd$_{real}$ [Å]</td>
<td>5.38</td>
<td>4.19</td>
<td>5.39</td>
<td>4.27</td>
<td>4.30</td>
<td>5.38</td>
<td>5.38</td>
</tr>
<tr>
<td>rmsd$_{ensemble}$ [Å]</td>
<td>11.02</td>
<td>11.07</td>
<td>4.97</td>
<td>11.30</td>
<td>10.97</td>
<td>11.31</td>
<td>11.03</td>
</tr>
</tbody>
</table>

Figure 4.26: Individual penalty functions sorted according to the structure indices that were extracted from ascendent sorting of the structures based on cluster rmsd for Rho modelling with 91 simulated distances between lables and $c = 0.58$.

constraint set as well as a sparse constraint set complemented by template constraints were tested. No correlation between most penalty functions and rmsd was found, except for very slight correlations for $P_{dist}$ and $P_{mpw}$ (Figure 4.27 and 4.28, respectively). A fully constrained result was tested against chosen penalty functions giving no improvement of the result (Table 4.19).
Chapter 4. Results and discussion

Figure 4.27: Individual penalty functions sorted according to the structure indices that were extracted from ascendent sorting of the structures based on cluster rmsd for vSGLT modelling with 378 simulated distances between helix ends, $\sigma_r = 5.8$ Å.

Table 4.19: Impact of clustering of structures chosen according to the smallest mentioned penalty value on modelling result shown for vSGLT with 378 simulated constraints between helix ends, $\sigma_r = 5.8$ Å.

<table>
<thead>
<tr>
<th>50% best of...</th>
<th>all</th>
<th>EPR</th>
<th>cluster</th>
<th>$P_{vec}$</th>
<th>$P_{compact}$</th>
<th>$P_{contact}$</th>
<th>$P_{apv}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>rmsd$_{real}$ [Å]</td>
<td>4.98</td>
<td>4.64</td>
<td>4.98</td>
<td>4.98</td>
<td>4.49</td>
<td>4.64</td>
<td>4.27</td>
</tr>
<tr>
<td>rmsd$_{ensemble}$ [Å]</td>
<td>4.00</td>
<td>3.80</td>
<td>3.39</td>
<td>3.71</td>
<td>3.95</td>
<td>4.13</td>
<td>4.09</td>
</tr>
</tbody>
</table>

Figure 4.28: Individual penalty functions sorted according to the structure indices that were extracted from ascendent sorting of the structures based on cluster rmsd for vSGLT modelling with 75 simulated helix end distances and template constraints, $c = 0.58$.

4.3.7 Placing non-core helices

Modelling with a sparse set of distance constraints is unable to provide a correct helix bundle model. Including of template constraints into modelling enhances the quality of the result, but it drives the solution very strongly towards the template fold. Such modelling can prove or disprove consistency with the template, but does not generate a fold model ab initio.
The best we could hope for under these circumstances was to position the non-core helices of PutP with respect to the core based on the existing set of experimental constraints. In order to test whether at least this was feasible and would lead to a reliable result, we tested different approaches to reducing the primary ensemble for vSGLT that we had generated with 75 simulated constraints between spin labels. For this, the different reduction approaches discussed above were tested individually and in diverse combinations. Conditions which offer the most realistic placement of all non-core helices at once are accepted as the optimal ones.

All results are presented in Figures 4.29-4.34.

For $c = 0.58$, placement of helix I is the least satisfactory out of all non-core helices. An r.m.s.d between 27.66 and 47.22 Å vs the target structure was found (Figure 4.29). Better results are achieved when the ensemble is reduced to $\leq 50\%$ of the structures that have the best agreement with mean distances or distance distributions. An additional removal of structures based on cluster r.m.s.d should remain at the level of $\leq 50\%$ (EPR pre-selection) or even $\geq 70\%$ (distribution pre-selection).

Placement of helix XII is quite satisfactory in general, regardless of reduction levels (with few exceptions - see Figure 4.30).

Helix XIII with its r.m.s.d$_{real}$ spread between 16.63 and 34.42 Å is rather "unstable" (Figure 4.31). EPR pre-selection does not bring improvement for the result, cutting the structures at cluster r.m.s.d of $\leq 50\%$ does. EPR pre-selection $\leq 30\%$ pushes the central structure toward a non-realistic placement of helix XIII.

In comparison to that even a slight distribution pre-selection helps to find a better position of this helix (especially when the reduction is based on cluster r.m.s.d $\geq 30\%$).

The situation differs a bit for modelling with tighter $c$ factor. Generally, placement of helices is more accurate in this case.

Helix I is placed with r.m.s.d$_{real}$ of 9.65-15.27 Å (Figure 4.32). Best results are obtained when there is no EPR pre-selection or its level is equal to the additional reduction according to best cluster r.m.s.d level. Cutting the ensemble at EPR pre-selection $\leq 30\%$ leads to improper helix position.

Distribution pre-selection at 50-70% and an optional cluster r.m.s.d-based reduction ($>30\%$) improves helix placement.

In contrast to helix I, helix XII can be placed better when the ensemble is reduced by EPR pre-selection $\leq 50\%$ regardless of the level of cluster r.m.s.d-based selection. When distribution pre-selection is used, results at a level of 70% or by combining level of 50% with an additional 50-70% cluster post-selection is most accurate (Figure 4.33).

Similarly, as for modelling with $c = 0.58$, helix XIII placement varies the most depending on the applied reduction conditions (Figure 4.34). Best results are found for EPR pre-selection of 50% (any cluster r.m.s.d-based reduction) or 70% distribution pre-selection (any cluster r.m.s.d-based reduction).

Summarizing and generalizing, the optimal approach to ensemble reduction is either 50% EPR pre-selection and 50% clustering or $\leq 50\%$ distribution pre-selection and 70%
or no clustering based reduction for the "loose" $c$ parameter used during calculations. For "tight" $c$, findings are different. Here also 50% EPR pre-selection and 50% clustering provide reasonable results, but taking distribution pre-selection at a level of 70% gives the best result, which does not depend on the choice of cluster rmsd based reduction.
A "tight" $c$ parameter used for modelling results in the central structure that has smaller rmsd vs the target structure for non-core helices than for the "loose" $c$ and thus, it is recommended over using "loose"$c$ that results in the bundle with high $\text{rmsd}_{\text{ensemble}}$. For "tight" $c$, placing of non-core helices can be assumed as sufficient even without a
Table: Clustering Results

<table>
<thead>
<tr>
<th>Clustering Method</th>
<th>100%</th>
<th>70%</th>
<th>50%</th>
<th>30%</th>
</tr>
</thead>
<tbody>
<tr>
<td>100%</td>
<td>34.42 Å</td>
<td>34.42 Å</td>
<td>24.81 Å</td>
<td>20.97 Å</td>
</tr>
<tr>
<td>70%</td>
<td>34.42 Å</td>
<td>28.19 Å</td>
<td>24.81 Å</td>
<td>22.18 Å</td>
</tr>
<tr>
<td>50%</td>
<td>17.53 Å</td>
<td>28.19 Å</td>
<td>17.53 Å</td>
<td>22.08 Å</td>
</tr>
</tbody>
</table>

Figure 4.31: Superposition of vSGLT models obtained with 75 constraints, template information and $c = 0.58$ (transparent, pink helix ends) and the target structure (opaque, red helix ends) shown for helix XIII under various reduction conditions; cytoplasmic view (periplasmic view to be found in Appendix: Figure B.17).

Reduction of the ensemble, which is not the case for "loose" $c$ parameter. However, by means of cluster reducing based on diverse approaches, it is possible to place non-core helices correctly in both of the cases. The most general conditions that seem to lead to the correct placement of non-core helices for both "tight" and "loose" $c$ factor is 50%.
Figure 4.32: Superposition of vSGLT models obtained with 75 constraints, template information and \( c = 0.15 \) (transparent, pink helix ends) and the target structure (opaque, red helix ends) shown for helix I under various reduction conditions; cytoplasmic view (periplasmic view to be found in Appendix: Figure B.18).

EPR pre-selection and 50% geometrical reduction of a cluster. For vSGLT, placement of helix I, XII and XIII under these conditions is shown in Figure 4.36 (3\textsuperscript{rd} panel for \( c = 0.58 \) and 4\textsuperscript{th} panel for \( c = 0.15 \)).
vSGLT comprises a long loop of 24 residues between helices I and II. This is supposedly the reason for the difficulties in placement of helix I, as a short loop would constrain at least the C-terminal helix end coordinates. It is therefore reasonable to assume that
Chapter 4. Results and discussion

Figure 4.34: Superposition of vSGLT models obtained with 75 constraints, template information and $c = 0.15$ (transparent, pink helix ends) and the target structure (opaque, red helix ends) shown for helix XIII under various reduction conditions; cytoplasmic view (periplasmic view to be found in Appendix: Figure B.20).

for PutP with a loop length of only 15 residues, helix I placement is more accurate. In general, placement of helix XIII is least satisfactory.
4.3.7.1 Modelling with different templates

One of the concerns during modelling of vSGLT with template information was the fact that the crystal structure PDB #2XQ2 occurs in an occluded/inward open conformation, whereas the template for its modelling, LeuT PDB #2A65, was captured in an outward open state. Another template captured in the same conformation was searched. As the closest, the Mhp1 structure PDB #2JLO was found, offering not the identical, but the most similar state, an occluded state. According to the findings from the previous section, results for 50% EPR pre-selection and 50% ensemble reduction according to the best clustering rmsd are taken as representative (Figure 4.36).

Once the LeuT crystal structure in an inward open state became available [28], it was used as an alternative modelling template as well. Superposition of the cylindrical representation of LeuT structures in both conformations is presented in Figure 4.35. Despite the different states, the helix bundle topologies are very similar at our level of coarse graining. No rmsd between them is given as the helix ends definition slightly differs (check Tables B.1 and B.2), but it is in the order of only 4 Å. Reminding the agreement between modelled vSGLT helices in the core with the 2A65 template (Figure 4.7), which is worse than the difference between the two states, it can be suspected that template change will not improve the result significantly in this case. Nevertheless, modelling with the new template was performed, the result is shown in Figure 4.36.

As already seen, reducing the tightness factor $c$ for calculations improves the $\text{rmsd}_{\text{real}}$ of the helix bundle model a lot. The exact gain differs depending on the helix under consideration and on the used template.

It cannot be judged unambiguously that one of the tested templates is better than the others. For $c = 0.58$, best helix I positioning is obtained with the 3TT3-based template, while helices XII and XIII are better placed with the 2A65 template. At smaller $c = 0.15$ best helix I placement is still observed for the 3TT3-based template, best helix XII placement still for the 2A65 template, but helix XIII placement for the Mhp1 2JLO-based template (very close with the result obtained for 3TT3 template).
Chapter 4. Results and discussion

Figure 4.36: Matching non-core helices of the modelled vSGLT structures under conditions of 50% EPR pre-selection and 50% cluster rmsd cut (transparent, pink/violet helix ends) with the target structure (opaque, red/blue helix ends). Modelling done with 75 simulated constraints and template constraints (mentioned on the Figure). c used for calculations and rate of successful embeddings are depicted as well as rmsd for helix I, XII and XIII.

Hence, the least favourable template would be Mhp1 #2JLO in this case, but 2A65 and 3TT3 perform equally well for vSGLT modelling.

The results for vSGLT indicate that positioning of non-core helices with respect to the core can be successfully approached at a coarse level, by means of modelling with a
suitable template and with a tightness parameter $c$ chosen such that only a few percent of structures gets successfully embedded to the Cartesian space. The choice of template is not critical. Furthermore, an ensemble reduction by two criteria is most appropriate. First, the 50% of all models should be reduce that satisfy experimental or simulated constraints best and second, 50% of the structures that are geometrically the most uniform.

### 4.3.7.2 Placement of non-core helices in PutP

Based on the findings in the previous section, the PutP bundle was modelled and analyzed. As a template for core helices (helices 2-11 in vSGLT and PutP), the vSGLT structure with PDB ID 2XQ2 was used. 75 experimentally derived restraints were utilized. 3000 trials with $c = 0.55$ were run in order to achieve 127 successful embeddings ($\sim 4\%$). The central structure remained unchanged when the ensemble was reduced to 50% of all models based on rmsd to the distance constraint list and to 50% based on geometric similarity, although this reduction allowed for tightening the ensemble from $\text{rmsd}_{\text{ensemble}} = 9.12 \text{ Å}$ to $\text{rmsd}_{\text{ensemble}} = 7.43 \text{ Å}$. The resulting central structure superimposed with 2XQ2 template is shown in Figure 4.37.

As seen on the example of vSGLT (Figure 4.36), a secondary transporter belonging to the SSS family, our methodology is able to place the non-core helices I and XII with respect to the core helices with reasonable agreement to the native structure. The rmsd values for the chosen ensemble reduction (50% EPR-preselection + 50% reduction according to $\text{rmsd}_{\text{ensemble}}$) is of the order of $8 \text{ Å}$, $7 \text{ Å}$ and $14 \text{ Å}$ for helices I, XII and XIII, respectively, which is a satisfactory result given the coarse-grained nature of our approach.

Our model provides evidence that helix I and XII in PutP are similarly positioned with respect to the ten-helix core as in vSGLT (Figure 4.37 B and C). This indicates that they play similar mechanistic roles in these two transporters in the SSS family. The situation is less clear cut for helix XIII, whose periplasmic ends appear to nearly coincide (helix end points separated by $7.31 \text{ Å}$) while the cytoplasmics do not (helix ends separated by $47.89 \text{ Å}$) (Figure 4.37 D). The “misplaced” cytoplasmic end of helix XIII in PutP is constrained by 3 distances (to the cytoplasmic end of helix II, to the cytoplasmic end of helix IX).

Although in most modelling runs, helix I was roughly placed in the same position as displayed in Figure 4.37, in an earlier run with only 72 restraints (5 constraints for helix I - 3 constraints for the cytoplasmic end and 2 constraints for the periplasmic end) we encountered ambiguous placement (see Figure B.42 in Appendix B). This shows that positioning of non-core helices depends critically on the number of constraints.

The obtained result can be compared with a homology model [63] (data obtained from G. Jeschke, internal communication), but only for helices I and XII as helix XIII is missing in the homology model because of very low sequence homology in this region. Figure 4.38 depicts the superimposition of them.

Whilst positioning of helix I is the same in both models, positioning of helix XII is clearly different when the homology model was translated into the cylindrical representation...
Figure 4.37: A Result of PutP modelling with $c = 0.55$, 75 experimental constraints and a ten-helix core template taken from vSGLT structure #2XQ2. Ensemble $\text{rmsd}_{\text{ensemble}} = 9.12 \, \text{Å}$ for non-reduced ensemble and $\text{rmsd}_{\text{ensemble}} = 7.43 \, \text{Å}$ for the reduced ensemble (50% EPR pre-selection and 50% geometrical post-selection). Non-core helices of the model highlighted on the superimposition of PutP and the #2XQ2-based target bundle (vSGLT-based template in opaque colours, PutP in transparent) for B helix I (blue), C helix XII (orange) and C helix XIII (red), shown also from a cytoplasmic (second column) and periplasmic (third column) view. Rmsd between a helix of the central structure of the modelled ensemble and a helix of the template is indicated.

according to the topology prediction (Table 4.1), see Figure 4.38 B. There are two reasons for that. First of all, helix XI has a very significant bent in the homology model (Figure 4.39), which has caused a shift of helix XII in the cylindrical approximation as an effect of removing helix-helix clashes. Secondly, the homology model was translated onto the cylindrical representation by means of helix ends definition according to the original topology model (Table 4.1). As depicted on Figure 4.39, ends of helix XI and XII in the topology model do not coincide with the ones of the homology model. Redefinition of the ends of helix XI and XII according to the homology model (Table 4.1) for translation of the homology model onto the cylindrical representation improves the agreement between the two models (Figure 4.38 C).

Positioning of non-core helices can be also seen in a direct comparison between our model and the homology model (Figure 4.40, helix I in blue, helix XII in red). Helix
Figure 4.38: Matching non-core helices I and XII of the modelled PutP structure under conditions of 50% EPR pre-selection and 50% cluster rmsd cut (transparent, pink/violet helix ends) with the homology model [63] (opaque, red/blue helix ends). Modelling done with 75 experimental constraints and the 2XQ2-based template of non-core helices, \( c = 0.55 \). Superimposition of the structures highlighted for A helix I (blue) and B, C helix XII (red). Whilst B shows the result obtained when the homology model was translated onto the cylindrical representation by means of helix ends definition according to the topology prediction (see 4.1), C illustrates the superimposition when the homology model was translated onto the cylindrical representation by means of helix ends as predicted by the homology model (see 4.1). In C, \( \text{rmsd}_{\text{real}} \) indicated on the figure is just an approximation as the length of helix XII differs in our model and the homology model. Helix positioning is shown also from a cytoplasmic (second column) and periplasmic (third column) site.

XIII of the bundle model is evidently misplaced as it is positioned in the middle of the core helices in the homology model (Figure 4.40).

Position of a label attached to a protein can be calculated by means of multilateration if a set of distance constraints is available for the residue where the label should be attached. It is shown on the example of vSGLT that a set of simulated constraints that has a size of the set available for PutP allows for calculating the localization of a label that corresponds to chosen residues (Figure 4.41). Position of the labels attached to the periplasmic ends of helix XII and XIII could not have been calculated due to an ambiguous localization that is caused by the sparsity of distance information (3 distances constraints for each residue). On the other hand, the periplasmic end of helix XII, for which a larger number of constraints exist, can be very nicely localized.
Chapter 4. **Results and discussion**

Figure 4.39: Helix XI and XII of the homology model of PutP [63] coloured according to the original topology model (Table 4.1). Helix XI (yellow) has a significant bent, helix XII according to the topology (red) does not match the homology model.

Figure 4.40: Positioning of the spin label (golden) attached to the residue 446 that is the periplasmic end of helix XII in PutP calculated by multilateration. Helix I highlighted in blue and helix XII in red in both the homology model (helices) and the modelled bundle (cylinders). Helix XIII of the modelled bundle shown as a pale-green cylinder.

Based on these results, localization of the labels attached to the helix ends of PutP was tried. Unfortunatelly, it was only successful in one case, for the periplasmic end of helix XII (residue 446), see the golden cloud in Figure 4.40. In the other cases the distance constraints were not sufficient to place the labels unambiguously. The label localization for the residue 446 suggests that the bundle model positioned helix XII more accurately than the homology model.
Figure 4.41: Localization of the spin labels attached to helix end residues of non-core helices I, XII and XIII in the vSGlT structure with PDB ID #2XQ2 calculated by multilateration. Results of multilateration for non-core helices I (blue), XII (orange) and XIII (red), shown for successful label localization at helix ends at the periplasmic (cyan) and the cytoplasmic (red) sites of A helix I, B helix XII and C helix XIII. Lime-green clouds indicate the possible localization of the label attached to the coloured residue.

We can thus conclude that our modelling approach was able to position non-core helices I and XII well at the coarse-grained level, whereas the cytoplasmic end of helix XIII was clearly misplaced.
4.4 Extracellular loop 4 of PutP

The role of the extracellular loop 4 (eL4) in the widely accepted alternating access mechanism was examined for secondary transporters with a LeuT-fold, suggesting that it is an extracellular gate participating in occlusion of the substrate binding site and consequent substrate binding [28, 37, 54, 122, 123]. The role of eL4 in PutP is addressed here.

A broad spectrum of experiments including cysteine-scanning mutagenesis between residues 294-326, site-directed spin labelling and EPR spectroscopy, which are known for being sensitive to the conformational changes [86, 95, 124, 125], was conducted in order to provide information on structure and dynamics of eL4. Some of the information obtained by our collaboration partners is not discussed here, but in reference [62].

4.4.1 Following eL4 during Na⁺/proline transport

The goal of the DEER studies was to follow dynamic rearrangement of eL4 during the transport process. To that end, DEER traces were recorded for a few selected double mutants in the substrate-free state (apo) and upon Na⁺ or Na⁺/proline addition.

The selected pair in the eL4 loop - mutant 299R1/318R1 revealed a distance of 2.7 nm (Figure 4.42), suggesting that the helix-loop-helix motif has a hairpin-like architecture. The broad distance distribution did not exhibit any significant change upon Na⁺ or Na⁺/proline addition.

![Figure 4.42: DEER distance measurements of doubly-labelled PutP mutant 299R1/318R1 within the eL4 loop in liposomes.](image)

A few distances between eL4 residues and residues near helix ends were acquired: 50R1 in cTM1/304R1 in eL4, 294R1 in eL4/326R1 in cTM8 (only apo), 294R1 in eL4/446R1 in the first non-core TM - ncTM11 (only apo), 298R1 in eL4/446R1 in ncTM11, 298R1 in eL4/455R1 in ncTM12 and 294R1 in eL4/403R1 in cTM10 (Figure 4.43).

For the 298R1/446R1 mutant, addition of NaCl to the liposomal sample shifted the distance distribution slightly towards longer distances (Figure 4.43). Additional inclusion of proline caused a further shift towards longer distances, which appears clearly also in the background corrected time trace. However, the micellar preparation does
not contain a contribution from the 5 nm distance at all. This might suggest protein oligomerization or aggregation happening in liposomes. The observed change may thus be due to a change in aggregation rather than due to a change in eL4 conformation. Neither of the pairs 50R1/304R1 or 294R1/403R1 exhibited any sensitivity to the presence of Na\(^+\) or Na\(^+\)/proline (Figure 4.43). Though the labelled pair 294R1/403R1 seems to be sensitive to Na\(^+\) or/and substrate addition, it exhibits such a broad distance distribution that no firm conclusions can be drawn.

The PutP variant 298R1/455R1 was not labelled successfully (Figure B.44 A), the fitted background dimension of 2.09 (apo)/2.14 (Na\(^+\)) agreed with the one used for data correction of the liposomal sample. This indicates that only one of the two sites could be spin labelled.

The mutants 294R1/326R1 and 294R1/446R1 measured only in the substrate-free state were acquired in order to constrain homology modelling of the eL4 loop (Figure B.44 B). Both of them demonstrate broad distributions of distances.

Distances measured between the PutP helix ends were determined in absence or presence of NaCl/proline in order to test for conformational change of the helix bundle. Pairs 326R1 in cTM8/446R1 ncTM11 (only apo), 190R1 in cTM5/371R1 in cTM9, 62R1 in cTM1/400R1 in cTM10, 91R1 in cTM2/371R1 in cTM9, 149R1 in cTM3/450R1 in ncTM12, 41R1 in cTM1/118R1 in cTM3 were studied.

For variant 190R1/371R1 no changes in the distance distributions were found upon Na\(^+\) or Na\(^+\)/proline addition (Figure 4.44).

The double mutant 91R1/371R1 in liposomes exhibited a slight distance shortening after Na\(^+\)/proline addition. Because of the very small magnitude, the effect should be interpreted cautiously as it might be just an effect of experimental uncertainty. Micellar samples of the same mutant gave very broad spectra and no distance change upon ion/substrate addition was observed (Figure 4.44).

The supposedly interesting pair 149R1/450R1 of PutP, being analogous to the pair 136R1/480R1 in LeuT previously studied by Claxton et al. [54], exhibited a broad distance distribution indifferent to Na\(^+\) or Na\(^+\)/proline addition (Figure 4.44). This is in contrast to the LeuT transporter, where leucine binding shifts the distance distribution towards manifestly shorter distances and reduces it to one component [54].

The double mutant 41R1/118R1 delivered traces with very small modulation depth and broad distance distributions. No influence of ion/substrate addition was observed (Figure 4.44).

For 326R1/446R1, a trace was acquired only in the apo state as an additional constraint for homology modelling. It showed a broad distance range (Figure B.44 B).

The mutant 62R1/400R1 was unsuccessfully labelled as the fitted background dimensions of 1.66-1.72 correspond quite accurately to those used for background substraction (Figure B.44 A), indicating presence of a label at only one of the two sites.

For LeuT, alterations between the outward-occluded substrate-bound (#2A65 [5]) and inward-open state (#3TT3 [28]) are pronounced for distances between cTM1 and eL4, cTM 2 and 9, cTM 5 and 9, cTM3 and ncTM 11. The corresponding distances in PutP remain unaltered in our studies, thus excluding that the same structural change takes place.
Figure 4.43: DEER distance measurements of doubly-labelled PutP mutants between eL4 residues and residues near helix ends. Variant indicated on the figure measured in ligand-free state (apo, violet), upon addition of 50 mM NaCl (Na$^+$, pink) and with 50 mM NaCl and 10 mM proline (Na$^+$/P, green). Left column: normalized primary DEER data with background fits; middle: form factors after background correction scaled to the same modulation depth; right column: Tikhonov regularized distance distributions.

Distance changes between the outward-open apo structure (#3TT1 [28]) and outward-occluded substrate-bound structure (#2A65 [5]) of LeuT or between the vSGLT inward-occluded substrate-bound state (#3DH4 [38]) and the inward-open apo state (#2XQ2 [39]) are too small to be detected by DEER, hence such a transition could neither be confirmed nor be excluded in our studies.

Any observed changes of experimental parameters of PutP (for all experiments see [62]) are either very minor or can be explained by changes in aggregation. We may thus exclude global rearrangement of the protein conformation under our experimental conditions. Although small local conformational changes could be interpreted as possible subtle transitions towards an outward-open or inward-open state, they cannot be definitely assigned to any such change. Since the distance distributions are broad and remain broad under all conditions, a superposition of inward-open and outward-open states cannot be excluded. Such a case was observed in EPR studies of the structurally
Figure 4.44: DEER distance measurements of doubly-labelled PutP mutants between chosen protein helix ends. Variant indicated on the figure measured in ligand-free state (apo, violet), upon addition of 50 mM NaCl (Na$^+$, pink) and with 50 mM NaCl and 10 mM proline (Na$^+$/P, green). Left column: normalized primary DEER data with background fits; middle: form factors after background correction scaled to the same modulation depth; right column: Tikhonov regularized distance distributions.

unrelated Na$^+$/aspartate transporter Glt$P_h$ [126, 127].

For secondary transporters, the energetically uphill substrate transport is coupled with movement of ions along their gradient. Unfortunately, for PutP inserted into liposomes, protein orientation in the membrane is completely random and no control over the electrochemical gradient across the membrane can be exercised (Na$^+$ ions spread in- and outside of the vesicles). Possibly applying an electrical potential would help to
capture different transport stages. Co-reconstitution of a potassium or proton pump together with PutP in order to generate a membrane potential in liposomes would be desired. Channelrhodopsin, bacteriorhodopsin or the lactate-dehydrogenase would be possible pumps that could work.

Our findings are in contrast to other studies of secondary transporters with the LeuT-fold, where addition of ions/substrate results in global protein rearrangement in vitro without an electrochemical gradient being present. As found in LeuT [54, 122], addition of Na^+ causes opening of the protein to the extracellular milieu and simultaneous closing of the inward gate. Upon subsequent substrate binding, the protein rearranges closing the extracellular gate. Thus occlusion from both sides occurs. Protein fluctuations are enabled and inward opening follows on, ions are supposed to dissociate according to their gradient if such gradient is present. The substrate that has low affinity to the transporter when Na^+ unplugs dissociates to the intracellular side as well. The protein can fluctuate between inward- and outward-open states and thus the transport cycle can start again.

In comparison, in the Mhp1 transport cycle both Na^+ and substrate are bound by the protein simultaneously [123]. A rocking-bundle mechanism is supposed, where protein fluctuations allow for inward opening and the stepwise dissociation of ions and the substrate follows.

Restraints on secondary structure and water accessibility of eL4 in PutP obtained by our collaborators together with a few of our DEER constraints allowed for refinement of an existing homology model of PutP, in particular for defining the structure of eL4 and for deriving a hypothesis on its interaction with the protein core and its role in the transport mechanism [62].

### 4.4.2 Role of Glu311 in PutP

Once the role of eL4 in the gating process of PutP was proven [62], more detailed questions could be raised. To answer these questions, an extensive analysis of possible interactions between amino acids of eL4 and other protein domains was performed.

Scanning mutagenesis revealed the importance of Glu311 for the transport process, but its exact role remained uncharted [62]. The restraint-based structural model suggested an interaction with the peptide backbone close to the outer end of core transmembrane helix cTM10 [63]. The distance between the side chain carboxyl of Glu311 and the protein backbone at the periplasmic end of cTM10 in our model ( [62], improved by new data [63]), being as small as 2.6 Å, suggested hydrogen bonding. This suspicion is supported by crystal structures of vSGLT #3DH4 [38] and LeuT # 3TT3 [28], where analogous distances are in the H-bond range as well. The shortest distance found for PutP connects Glu311 and the backbone atom NH of Ala404. The latter residue was found to be crucial for membrane insertion, stability and function of the protein [63].

A few site pairs were scanned in order to test for an influence of Glu311 and Ala404 mutations on the transport cycle. Residue 298 (eL4a) was chosen in conjunction with residues 326 (outer end of cTM8) or 391 (outer end of cTM9) as well as the 190R1/371R1
mutant presenting the distance between inner ends of cTM5 and cTM9. This choice of the positions was based on analysis of structural transitions of proteins with a LeuT fold [43] that suggested pairs giving insight on movement of eL4, as well as the accessibility of the pair distances to the DEER experiment (well-defined interspin distances according to the existing homology model).

Measurements of 190R1/371R1 made in micelles (Figure 4.45 A) led to very broad distance distributions. The original traces, despite low nominal protein concentration, reveal a strong background. Such decay could suggest a contribution from distances of 6-7 nm, which are not expected in this case. Alternatively, this behaviour could be explained by formation of aggregates. Removal of free labels was undertaken to improve matters. Desalting of the sample as well as dialysis yielded a slightly increased modulation depth, but the decay in the time trace remained (Figure B.43 A). As no big improvement in data quality was found, the 190R1/371R1 mutant was discarded.

Similarly as for the 190/371 labelled pair, time traces of 298/391 (Figure 4.45 B) show a decay which should be absent for such low nominal concentrations (here assigned by Bradford protocol as being between 3 and 7 µM). The mutation A404E leads to shortening of the distance. Nevertheless, a firm conclusion cannot be drawn, as the signal quality is too poor as a consequence of severely reduced protein expression of this mutant and very low modulation depth (labelling efficiency around 20%).

The problem could be addressed successfully with double mutant 298R1/326R1. In this case, the additional mutations E311A and A404E cause elongation of the distance between the labels (Figure 4.45 C and B.43 B). The observed effect is unambiguous, oscillations present in the time trace of non-mutated 298R1/326R1 vanish completely upon mutations. Low protein concentration in connection with low labelling efficiency, also for A404E, where protein expression is drastically inhibited, could not explain the observed behaviour.

Samples obtained in liposomes exhibit the same type of distance shortening for E311A and A404E mutants (Figure 4.45 D). Although the signal-to-noise ratio of the 298R1/326R1/A404E trace is very low, the decay is significantly different from the one observed in the pseudo-wild type.

In our experiments on PutP, residue 298 from eL4 and 326 at the outer end of cTM8 move further apart when residue 311 or 404 is mutated out. Thus opening of the extracellular gate and simultaneous closing of the inner cavity [61, 128, 129] as a consequence of disrupting, most probably, an H-bond between Glu311 and Ala404 can be deduced. Importantly, PutP is most stable in an inward-open conformation [61, 128–130], which underlines the importance of this H-bond. In contrary, LeuT being most stable in the outward-open conformation is driven towards an inward-open conformation upon Y268A or R5A mutation [122]. Bonding networks in the LeuT and in the dopamine transporter have been previously studied [46, 131].

Nevertheless, PutP data interpretation requires caution. Our restraint-supported homology model [61] is based on a locally uncertain sequence alignment with vSGLT in cTM10 - our alignment differs from the one given by Faham et al. [38]. In the latter alignment, Leu398 is the extracellular end of cTM10, so that the Ala404 residue resides
more inside the membrane. Moreover, Leu398 is also of high importance (proposed participation in proline binding [132]) and an interaction between this amino acid and eL4 would connect substrate binding with the dynamics of gating. This or that way, residue 404 was experimentally proven to be significant [63] and its substitution capable to influence global conformational changes. Such observations are not surprising. In vSGLT1 it was proposed that the side chains placed at the gate are not only a "passive" barrier but take part "actively" in coordinating sodium and the substrate [133].

Figure 4.45: Influence of E311A (blue) and A404E (red) mutations on the interspin distances measured by DEER between doubly-labelled PutP variants indicated on the figure. Data for the pseudo-wild type WT* (spin-labelled, but with E311 and A404) are shown in black. Left column: normalized primary DEER data with background fits; middle: form factors after background correction scaled to the same modulation depth; right column: Tikhonov regularized distance distributions.
Chapter 5

Summary and outlook

In this thesis it was shown that distance measurements by the double electron electron resonance technique supported by a modelling approach are capable of delivering structural information on PutP, a secondary transporter protein that symports proline and sodium ions in *E. coli*.

The PutP fold was investigated by means of a modelling approach based on distance geometry, which is an algebraic methodology that allows to translate the set of distances into the coordinates of the points for which the distances were given. In our case lower and upper bounds derived from the experimental data, mean distances $r_{\text{mean}}$ and standard deviations $\sigma_r$, were available. Random setting of a distance matrix and its embedding to Cartesian space offered an ensemble of the structures that fits given bounds. The central structure that is the structure with the smallest pairwise rmsd within the ensemble was taken as its representative.

The available set of 75 interspin distances was too sparse to build a model at atomic resolution and thus, the problem had to be simplified as a coarse-grained helix bundle model. In this way, all 13 transmembrane helices of PutP were redefined as the straight cylinders that could have been positioned in Cartesian space only by setting two end points of each helix. Even then, the number of constraints was proved to be too small for deriving a stable model of the helix bundle. In order to find the reasons for this failure of the modelling approach and ways to overcome the problem, the methodology was analyzed in depth. For this, it was tested with simulated constraints against two proteins with their crystal structures known, the galactose transporter vSGLT and the bovine rhodopsin Rho. Although the architecture of their helix bundle is very different and the number of helices and deviation of their shape from the shape of an ideal straight $\alpha$ helix also differs, conclusions are consistent.

A set of distance constraints equivalent to the one experimentally available in the PutP case was simulated for vSGLT. For that, spin labels were virtually attached to the protein structure by means of the rotamer library approach and the distances between the mean NO-bond midpoint coordinates were obtained. Also in that case, modelling did not yield a stable helix bundle and the approach turned out solutions that differed strongly from the known bundle architecture. In order to fix this problem, additional
constraints were derived. Namely, a set of distances was generated from the helix arrangement within a conserved ten-helix core, that is common to several secondary transporters despite their low sequence similarity. We found that the experimental constraints are insufficient to detect the subtle changes in the ten-helix core structure between different secondary transporters and hence, no information about the relative arrangement of helices within a protein core can be obtained. The core architecture is dominated by the template constraints.

It was also tested whether a full set of distance constraints between labels attached to the helix ends of PutP would allow for deriving a reliable model of the helix bundle architecture. Tests were performed with 91 and 378 simulated distance constraints for Rho and vSGLT, respectively. The quality of models was, in general, better for Rho with 7 nearly ideal helices than for vSGLT with 14 less ideal helices. The limiting accuracy was found at the order of 8 Å for both model proteins. In the case of vSGLT, with some helices that are bent in reality, the value appeared reasonable within the approximation of straight cylinders. However, for Rho with almost perfectly straight helices, the deviation pointed to other uncertainties than only helix shape. We conjectured that length and flexibility of the label linker were to blame and investigated the approach upon virtual removal of the labels. To check that, we considered distance constraints between Cα atoms at the helix ends. Indeed, the result obtained for the biggest possible sets of constraints was then improved, the rmsd between the central structure of the ensemble and the native structure was decreased to 4 Å for vSGLT and 3 Å for Rho. The obtained accuracy is satisfactory at the given coarse-grain level. The result obtained for Rho is comparable to the number of 3.2 Å offered by another approach [77]. This finding indicates that either more rigid labels or labels with shorter linkers are required for the introduced modelling approach to deliver better results or that independent information on label position from other experimental techniques may be needed.

We looked whether reduction of the originally computed ensemble of models by different criteria would lead to a central structure of the ensemble that better represented the target structure. We considered removal of structures with high values of penalty functions that are based on common structural features of membrane proteins. Unfortunately, we found no impact on the ensemble quality.

We tested also removal of those structures that have the largest deviation from EPR distance constraints (EPR pre-selection). This deviation was estimated either by rmsd for each distance from the constraint list vs the distance in the model or by the probability of achieving a model distance in the simulated distance distribution. EPR pre-selection brought only minor improvement in agreement of the central structure with the target structure.

We checked also impact of removal of structures that are geometrically most dissimilar from other models in the ensemble (geometrical post-selection). The similarity of structures was defined by pairwise rmsd between all structures in the ensemble. In this case, removal of the structures with highest $\text{rmsd}_{\text{ensemble}}$ was essential to unify the ensemble. Only upon such reduction definition of the central structure as the representative of the ensemble is meaningful. Anyway, central structures of the reduced ensembles did not represent the target structure much better.
In this situation, placement of the non-core helices with respect to the ten-helix core was the new information on PutP that could still be sought. Additional tests with simulated constraints for the vSGLT case revealed that the reliability of such placement depends strongly on the uncertainty $c$ assumed for the mean distances between the spin labels, $r = r_{\text{mean}} \pm c \cdot \sigma_r$. Accuracy of the models improves on tightening of the bounds as long as the constraints are still sufficiently consistent to allow for embedding of the distance matrix to Cartesian space. Additionally, tightening the $c$ factor was crucial to avoid wrong protein architectures that contained some swapped helices. The reliability of placement of non-core helices with respect to the ten-helix core could be improved by combining 50% EPR pre-selection vs the constraint list with an additional 50% geometrical post-selection.

Under these conditions we placed non-core helices of PutP in relation to the core helices. Placement of non-core helices was compared to those of the homology model of PutP and of the vSGLT structure. Our approach allowed to position helix I well at our coarse level. Multilateration that was used to localize the spin label attached to the periplasmic end of helix XII indicated that the placement of this helix is better in our bundle model than in the homology model. The cytoplasmic end of helix XIII is misplaced in our bundle model. This indicates that additional constraints between core helix ends and ends of helix XIII would be required in order to position this helix.

We also addressed the problem of the commonly accepted alternating access mechanism [10] in secondary transporters and, in particular, the role of extracellular loop 4 (eL4) in this mechanism. PutP samples without ligand, with sodium and with both sodium and proline were investigated. Unfortunately, the experiments did not reveal states with different conformations, most likely due to the lack of a membrane potential. Still, our data combined with data from a spin labelling site scan performed by our collaborators allowed for modelling eL4 and for deriving a hypothesis on its role in the mechanism and could further improve the model of eL4 as well as the model of its interactions with the ten-helix core.

The mutagenesis of Glu311 to Ala or Ala404 to Glu resulted in drastically elongated distance within the 298R1/326R1 labelled pair. This observation underlined the significance of residues Glu311 at eL4 and Ala404 at cTM10 in stabilizing the inward-open conformation of PutP via, presumably, a hydrogen bond [63]. It can be instigated that the interaction between the tip of eL4 and cTM10 is important for the transduction of the conformational changes while the substrate is transported.
Appendix A

Geometric algebra

A.1 Mathematical basis

The geometric product of three vectors $a$, $b$ and $c$ is given by

\[
(\vec{a}\vec{b})\vec{c} = \vec{a}(\vec{b}\vec{c})
\]

\[
\vec{a}(\vec{b} + \vec{c}) = \vec{a}\vec{b} + \vec{a}\vec{c}
\]

\[
(\vec{a} + \vec{b})\vec{c} = \vec{a}\vec{c} + \vec{b}\vec{c}
\]

\[
\vec{a}^2 = \|\vec{a}\|^2 \Rightarrow \vec{a}^{-1} = \vec{a}/\vec{a}^2
\]

The inner product of vectors, by the law of cosines, is the symmetric part of their geometrical product $\vec{a}\vec{b}$

\[
\vec{a} \cdot \vec{b} = \frac{1}{2}(\vec{a}\vec{b} + \vec{b}\vec{a}) \tag{A.1}
\]

The definition of the outer product is

\[
\vec{a} \wedge \vec{b} = \frac{1}{2}(\vec{a}\vec{b} - \vec{b}\vec{a}) \tag{A.2}
\]

This outer product is clearly anticommutative

\[
\vec{a} \wedge \vec{b} = -\vec{b} \wedge \vec{a} \tag{A.3}
\]

and it describes a new quantity called a bivector.

The orthonormal basis $\hat{e}_1$, $\hat{e}_2$ of $\hat{e}_3$ defines the space of the vector. Within this, the outer product can be expanded

\[
\vec{a} \wedge \vec{b} = (a_1\hat{e}_1 + a_2\hat{e}_2 + a_3\hat{e}_3) \wedge (b_1\hat{e}_1 + b_2\hat{e}_2 + b_3\hat{e}_3)
\]

\[
= (a_1b_2 - b_1a_2)\hat{e}_1\hat{e}_2 + (a_3b_1 - b_3a_1)\hat{e}_2\hat{e}_3 + (a_2b_3 - b_2a_3)\hat{e}_2\hat{e}_3
\]

So every bivector can be expanded by use of the elementary bivectors $\hat{e}_1 \wedge \hat{e}_2 = \hat{e}_1\hat{e}_2$, $\hat{e}_3 \wedge \hat{e}_1 = \hat{e}_3\hat{e}_1$ and $\hat{e}_2 \wedge \hat{e}_3 = \hat{e}_2\hat{e}_3$. 
It can be also shown that
\[(\alpha \hat{e}_2 + \beta \hat{e}_3 \hat{e}_1 + \gamma \hat{e}_2 \hat{e}_3)^2 = -\alpha^2 - \beta^2 - \gamma^2\] (A.4)
which proves that these bivectors are linearly independent and their space is most likely 3-dimensional.

The outer product of three vectors (trivector) can be defined as
\[\vec{a} \wedge \vec{b} \wedge \vec{c} = \frac{1}{6}(\vec{a}\vec{b}\vec{c} - \vec{a}\vec{c}\vec{b} + \vec{b}\vec{a}\vec{c} - \vec{b}\vec{c}\vec{a} + \vec{c}\vec{a}\vec{b} - \vec{c}\vec{b}\vec{a}) = \det(\vec{a}, \vec{b}, \vec{c}) \hat{e}_1 \hat{e}_2 \hat{e}_3 \] (A.5)

All trivectors can be presented by means of the trivector \(\hat{\mathbf{i}} = \hat{e}_1 \hat{e}_2 \hat{e}_3\). This trivector has similarities with imaginary unit, namely \(\hat{\mathbf{i}}^2 = -1\).

Products in geometric algebra display parity, hence
\[\vec{a} \wedge (\vec{b} \wedge \vec{c}) = \frac{1}{2}(\vec{a}(\vec{b} \wedge \vec{c}) + (\vec{b} \wedge \vec{c})\vec{a}) = \vec{a} \wedge \vec{b} \wedge \vec{c} = (\vec{a} \wedge \vec{b}) \wedge \vec{c} \] (A.6)

Calculation on inner and outer products can be joined, f.e.
\[\vec{a} \bullet (\vec{b} \wedge \vec{c}) = \frac{1}{2}(\vec{a}(\vec{b} \wedge \vec{c}) - (\vec{b} \wedge \vec{c})\vec{a}) = (\vec{a} \bullet \vec{b})\vec{c} - (\vec{a} \bullet \vec{c})\vec{b} \] (A.7)

Also, the inner product of two bivectors can be calculated
\[(\vec{a} \wedge \vec{b}) \bullet (\vec{a} \wedge \vec{c}) = \vec{a} \bullet (\vec{b} \bullet (\vec{a} \wedge \vec{c})) = (\vec{a} \bullet \vec{c})(\vec{b} \bullet \vec{d}) - (\vec{a} \bullet \vec{d})(\vec{b} \bullet \vec{c}) = \det \begin{bmatrix} \vec{a} \bullet \vec{c} & \vec{a} \bullet \vec{d} \\ \vec{b} \bullet \vec{c} & \vec{b} \bullet \vec{d} \end{bmatrix} \] (A.8)

And similarly, the inner product of two trivectors can be presented as
\[(\vec{a} \wedge \vec{b} \wedge \vec{c}) \bullet (\vec{f} \wedge \vec{e} \wedge \vec{d}) = \det \begin{bmatrix} \vec{a} \bullet \vec{d} & \vec{a} \bullet \vec{e} & \vec{a} \bullet \vec{f} \\ \vec{b} \bullet \vec{d} & \vec{b} \bullet \vec{e} & \vec{b} \bullet \vec{f} \\ \vec{c} \bullet \vec{d} & \vec{c} \bullet \vec{e} & \vec{c} \bullet \vec{f} \end{bmatrix} \] (A.8)

These determinants of inner product matrices are called Gramians.

The useful feature of geometric algebra is that purely algebraic quantities have physical, geometric meaning. New relations can be then easily derived and interpreted, f.e. \(\vec{a} \bullet \vec{b} = 0 \implies \text{perpendicular [108]}\).

It can be shown that other types of algebra, f.i. the Gibbs’ vector algebra, fit to the geometric algebra. The inner product existing in Gibbs’ algebra can be omitted, the outer product of two vectors can be rewritten by usage of the identity relation and Eq. A.2
\[- \hat{\mathbf{i}} \hat{e}_1 \hat{e}_2 = (\hat{e}_3 \hat{e}_2 \hat{e}_1)\hat{e}_1 \hat{e}_2 = (\hat{e}_3 \hat{e}_2)\hat{e}_1 \hat{e}_2 = \hat{e}_3 \hat{e}_2\hat{e}_1 = \hat{e}_3 \] (A.9)

Furthermore, it can be shown that
\[- \hat{\mathbf{i}}(\vec{a} \wedge \vec{b}) = (a_2 b_3 - b_2 a_3)\hat{e}_1 + (a_3 b_1 - b_3 a_1)\hat{e}_2 + (a_1 b_2 - b_1 a_2)\hat{e}_3 = \vec{a} \times \vec{b} \] (A.10)
which is a Gibbs’ cross product represented in coordinates. Vectors are assigned to
directed line segments. Thus, bivectors can be translated onto directed plane segments.
Contrary to the cross product, the direction of a bivector is not changed by the inversion
according to the identity relation

\[ i(\vec{a} \times \vec{b}) = \vec{a} \wedge \vec{b} \]  \hspace{1cm} (A.11)

The Gibbs’ triple product is

\[ \vec{a} \cdot (\vec{b} \times \vec{c}) = -\frac{1}{2}(\vec{a}((\vec{b} \wedge \vec{c}) + i(\vec{b} \wedge \vec{c})\vec{a}) = -\frac{i}{4}(\vec{a}\vec{b}\vec{c} - \vec{a}\vec{c}\vec{b} + \vec{b}\vec{c}\vec{a} - \vec{c}\vec{b}\vec{a}) = -i(\vec{a} \wedge \vec{b} \wedge \vec{c}) \]  \hspace{1cm} (A.12)

The trivector can be hence interpreted as directed space segment.
If we take an arbitrary plane in space, f.e. \( \hat{\vec{e}}_1\hat{\vec{e}}_2 = \hat{\vec{e}}_1 \wedge \hat{\vec{e}}_2 \) and define \( \vec{a} = a_1\hat{\vec{e}}_1 + a_2\hat{\vec{e}}_2 \) as
any vector in this space, then

\[ \hat{\vec{e}}_1\vec{a} = \hat{\vec{e}}_1 \cdot \vec{a} + \hat{\vec{e}}_1 \wedge \vec{a} = a_1 + a_2\hat{\vec{e}}_1\hat{\vec{e}}_2 \]
\[ (\hat{\vec{e}}_1\hat{\vec{e}}_2)^2 = -(\hat{\vec{e}}_1\hat{\vec{e}}_2)(\hat{\vec{e}}_2\hat{\vec{e}}_1) = -1 \]

which means that any vector relative to some vector fixed in the plane can be described
as a ”complex number”. Namely, rotation by 1/4 of a full turn is made by multiplication
by the imaginary unit of the plane \( \hat{\vec{e}}_1\hat{\vec{e}}_2 \).
The inner product of a vector and a bivector is then interpreted as a projection into
the bivector plane followed by 1/4 turn and scaling by the bivector magnitude. The
projection is given by

\[ (\vec{a} \cdot (\hat{\vec{e}}_1\hat{\vec{e}}_2))(\hat{\vec{e}}_2\hat{\vec{e}}_1) = (a_1\hat{\vec{e}}_2 - a_2\hat{\vec{e}}_1)(\hat{\vec{e}}_2\hat{\vec{e}}_1) = a_1\hat{\vec{e}}_1 + a_2\hat{\vec{e}}_2 = (\hat{\vec{a}}\hat{\vec{e}}_1)\hat{\vec{e}}_1 + (\hat{\vec{a}}\hat{\vec{e}}_2)\hat{\vec{e}}_2 \]  \hspace{1cm} (A.13)

The inner product of two bivectors

\[ (\vec{a} \wedge \vec{b}) \cdot (\vec{b} \wedge \vec{c}) = (\vec{a} \cdot \vec{c})\Vert \vec{b} \Vert^2 - (\vec{a} \cdot \vec{b})(\vec{b} \cdot \vec{c}) = (\cos \theta_{ac} - \cos \theta_{ab} \cos \theta_{bc})\|\vec{a}\|\|\vec{b}\|\|\vec{c}\| \]
\[ = -(\sin \theta_{ab} \sin \theta_{bc} \cos \phi)\|\vec{a}\|\|\vec{b}\|\|\vec{c}\| = -\|\vec{a} \wedge \vec{b} \wedge \vec{c}\| \cos \phi \]

where \( \phi \) is the dihedral angle between the planes \( \vec{a} \wedge \vec{b} \) and \( \vec{b} \wedge \vec{c} \).
The reflexion of \( \vec{a} \) in the plane perpendicular to an arbitrary unit vector \( \hat{\vec{e}} \)

\[ -\hat{\vec{e}}\vec{a}\hat{\vec{e}} = -\hat{\vec{e}}(\vec{a} \cdot \hat{\vec{e}} + \vec{a} \wedge \hat{\vec{e}}) = \vec{a} - 2(\vec{a} \cdot \hat{\vec{e}})\hat{\vec{e}} \]  \hspace{1cm} (A.14)

According to the rule that the product of two reflections is a rotation \( \vec{R} \) about the axis
in which their planes intersect and by twice the smaller angle between the planes, for
any two unit vectors \( \hat{\vec{e}} \) and \( \hat{\vec{f}} \)

\[ (\hat{\vec{e}}\hat{\vec{f}})\vec{a}(\hat{\vec{e}}\hat{\vec{f}}) = (\hat{\vec{e}} \cdot \hat{\vec{f}} - \hat{\vec{e}} \wedge \hat{\vec{f}})\vec{a}(\hat{\vec{e}} \cdot \hat{\vec{f}} + \hat{\vec{e}} \wedge \hat{\vec{f}}) = \vec{R}\vec{a}\vec{R} \]  \hspace{1cm} (A.15)

where the rotation of \( \vec{a} \) is by the angle \( 2 \arccos(\hat{\vec{e}} \hat{\vec{f}}) \).
As a consequence, any rotation can be presented as a sum of a scalar and a bivector

\[ \vec{R} = \sigma + \rho \cdot \vec{v} \]  \hspace{1cm} (A.16)
where \( \hat{r} \) is a unit vector along the axis of rotation and \( \sigma^2 + \rho^2 = 1 \) such that \( \overline{RR} = 1 \). Up to the sign this description is unique.

## A.2 Invariant theory

Geometric algebra offers a set of quantities staying invariant upon rotations and translations and these are of particular interest. Namely, we seek for scalar-valued expressions generated by the interpoint vectors. Such invariant quantities can be always reduced to multivariate polynomials in the squared interpoint distances and signed volumes. Relations between these invariants are given by Cayley-Menger determinants \([134, 135]\).

In Cartesian coordinates, the required features hold for squared distances between points and signed volumes and they provide a complete system. Hence any invariant can be described as a multivariate polynomial in them \([136]\).

Any set of multivectors in 3-dimensional space can be described by a system of scalar-valued entities in these fundamental invariants. It can be always split into scalar, vector, bivector and trivector parts. Further, any set of vectors is determined by their Gram matrix of inner product. Coordinates of a freely chosen vector \( \vec{x} \) are determined by a maximal linearly independent subset of vectors \( \vec{a}, \vec{b}, \ldots \) versus the basis

\[
\begin{bmatrix}
\vec{a}^2 & \vec{a} \cdot \vec{b} & \ldots & \vec{a} \cdot \vec{c} \\
\vec{a} \cdot \vec{b} & \vec{b}^2 & \ldots & \vec{b} \cdot \vec{c} \\
\vdots & \vdots & \ddots & \vdots \\
\vec{a} \cdot \vec{c} & \vec{b} \cdot \vec{c} & \ldots & \vec{c}^2 \\
\end{bmatrix}
\begin{bmatrix}
x_a \\
x_b \\
\vdots \\
x_c \\
\end{bmatrix}
= 
\begin{bmatrix}
\vec{a} \cdot \vec{x} \\
\vec{b} \cdot \vec{x} \\
\vdots \\
\vec{c} \cdot \vec{x} \\
\end{bmatrix}
\tag{A.17}
\]

Any expression can be expanded to a polynomial in the inner products, though it can be done in a few ways corresponding to different choices of interpoint vectors. The simplest one assumes squared distances \( \|\vec{c} - \vec{d}\|^2 = D_{ad} \) etc.

\[
(\vec{a} - \vec{b}) \cdot (\vec{c} - \vec{d}) = \frac{1}{2}(\|\vec{a} - \vec{d}\|^2 + \|\vec{b} - \vec{c}\|^2 - \|\vec{a} - \vec{c}\|^2 - \|\vec{b} - \vec{d}\|^2) = \frac{1}{2}(D_{ad} + D_{bc} - D_{ac} - D_{bd})
\tag{A.18}
\]

The number of squared distances and signed volumes amongst a set of \( n \) 3-dimensional points is larger than the number of degrees of freedom \( 3n - 6 \). A complete set of algebraic relations among the squared distances and oriented volumes can be presented be means of Cayley-Menger determinants. The most important of these relations is

\[
(\vec{b} - \vec{a}) \wedge (\vec{c} - \vec{a}) \wedge (\vec{d} - \vec{a}) \wedge (\vec{c} - \vec{a}) = 0
\tag{A.19}
\]

showing that the hypervolume spanned by any other interpoint vector is always zero. As an example we consider a single pair of interpoint vectors \( \vec{x} = \vec{b} - \vec{a} \) and \( \vec{y} = \vec{c} - \vec{a} \). It can be expanded as a Gramian and according to Cauchy-Schwarz relation (\( \|\vec{u} \cdot \vec{v}\| \leq \|\vec{u}\|\|\vec{v}\| \))

\[
0 \leq (\vec{x} \wedge \vec{y}) \cdot (\vec{y} \wedge \vec{x}) = \det
\begin{bmatrix}
\vec{x}^2 & \vec{x} \cdot \vec{y} \\
\vec{x} \cdot \vec{y} & \vec{y}^2 \\
\end{bmatrix}
\tag{A.20}
\]
By proper extension and some basic operations this determinant can be rewritten

\[
\begin{vmatrix}
1 & 0 & 0 & 0 \\
0 & 1 & 0 & 0 \\
0 & 0 & \vec{x}^2 & \vec{x} \cdot \vec{y} \\
0 & 0 & \vec{x} \cdot \vec{y} & \vec{y}^2
\end{vmatrix} = - \det
\begin{vmatrix}
0 & 1 & 1 & 1 \\
1 & 0 & 0 & 0 \\
1 & 0 & \vec{x}^2 & \vec{x} \cdot \vec{y} \\
1 & 0 & \vec{x} \cdot \vec{y} & \vec{y}^2
\end{vmatrix}
\]

\[
= - \det
\begin{vmatrix}
0 & 1 & 1 & 1 \\
1 & 0 & -\frac{1}{2} \vec{x}^2 & -\frac{1}{2} \vec{x} \cdot \vec{y} \\
1 & -\frac{1}{2} \vec{x}^2 & 0 & \vec{y}^2 \\
1 & -\frac{1}{2} \vec{y}^2 & \vec{x} \cdot \vec{y} - \frac{1}{2} (\vec{x}^2 + \vec{y}^2) & 0
\end{vmatrix}
\]

which leads us to the the three-point Cayley-Menger determinant \( D(a, b, c) \)

\[
D(a, b, c) = - \det
\begin{vmatrix}
0 & 1 & 1 & 1 & 1 \\
1 & 0 & -D_{ab}/2 & -D_{ac}/2 & 1 \\
1 & -D_{ab}/2 & 0 & -D_{bc}/2 & 1 \\
1 & -D_{ac}/2 & -D_{bc}/2 & 0 & 1
\end{vmatrix}
\]

\[
= \frac{1}{4} (d_{ab} + d_{ac} + d_{bc})(d_{ab} + d_{ac} - d_{bc})(d_{ac} + d_{bc} - d_{ab}) \times (d_{bc} + d_{ab} + d_{ac})
\]

As the determinant has to be nonnegative, three triangle inequalities among the points arise

\[
d_{ab} \leq d_{ac} + d_{bc} \\
d_{ac} \leq d_{bc} + d_{ab} \\
d_{bc} \leq d_{ab} + d_{ac}
\]

In general, the \( n \)-dimensional Cayley-Menger determinant is given by

\[
D(a, b, \ldots, c; p, q, \ldots, r) = 2 \left( \frac{-1}{2} \right)^n \det
\begin{vmatrix}
0 & 1 & 1 & \ldots & 1 \\
1 & D_{ap} & D_{aq} & \ldots & D_{ar} \\
1 & D_{bp} & D_{bq} & \ldots & D_{br} \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
1 & D_{cp} & D_{cq} & \ldots & D_{cr}
\end{vmatrix}
\]

\[
= ((\vec{b} - \vec{a}) \wedge \ldots \wedge (\vec{c} - \vec{a})) \cdot ((\vec{q} - \vec{p}) \wedge \ldots \wedge (\vec{r} - \vec{p}))
\]

The five-point Cayley-Menger determinants are the Gramians of four interpoint vectors and thus vanish for the 3-dimensional case as well as every higher-order determinants do.

Though not easy to prove, it can be stated that nonnegativity of the symmetric determinant up to 4-point and vanishing of the determinants with higher order is sufficient that any symmetric matrix of real numbers with zeros down the diagonal is a matrix of squared distances in 3-dimensional Euclidean space.
Appendix B

Results

B.1 Helix/loop ends definition

Helix ends were defined for modelling templates, unless mentioned differently.

B.1.1 LeuT

Table B.1: Helix ends definition of LeuT PDB #2A65. "swiss" stands for SwissPDB-Viewer, "dssp" for DSSP. The resolution of #2A65 is 1.65 Å.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>11-37</td>
<td>11-37</td>
<td>11-37</td>
<td>11-37</td>
<td></td>
</tr>
<tr>
<td>41-70</td>
<td>41-70</td>
<td>41-70</td>
<td>41-71</td>
<td></td>
</tr>
<tr>
<td>88-124</td>
<td>89-124</td>
<td>88-124</td>
<td>88-124</td>
<td></td>
</tr>
<tr>
<td>166-183</td>
<td>166-184</td>
<td>166-184</td>
<td>166-184</td>
<td></td>
</tr>
<tr>
<td>191-213</td>
<td>191-213</td>
<td>191-213</td>
<td>191-213</td>
<td></td>
</tr>
<tr>
<td>241-266</td>
<td>242-267</td>
<td>241-266</td>
<td>241-266</td>
<td></td>
</tr>
<tr>
<td>276-306</td>
<td>276-305</td>
<td>276-305</td>
<td>276-305</td>
<td></td>
</tr>
<tr>
<td>337-369</td>
<td>337-370</td>
<td>336-369</td>
<td>336-369</td>
<td></td>
</tr>
<tr>
<td>375-395</td>
<td>375-394</td>
<td>375-395</td>
<td>375-395</td>
<td></td>
</tr>
<tr>
<td>399-424</td>
<td>399-424</td>
<td>399-424</td>
<td>399-424</td>
<td></td>
</tr>
<tr>
<td>447-477</td>
<td>447-476</td>
<td>447-477</td>
<td>447-477</td>
<td></td>
</tr>
<tr>
<td>483-513</td>
<td>482-514</td>
<td>483-513</td>
<td>483-513</td>
<td></td>
</tr>
</tbody>
</table>
Table B.2: Helix ends definition of LeuT PDB #3TT3. "accelrys" stands for Accelrys Discovery Studio, "dssp" for DSSP. The resolution of #3TT3 is 3.22 Å

<table>
<thead>
<tr>
<th>2A65 paper [5]</th>
<th>3TT3 accelrys</th>
<th>3TT3 dssp</th>
<th>taken for modelling</th>
</tr>
</thead>
<tbody>
<tr>
<td>11-37</td>
<td>11-37</td>
<td>12-37</td>
<td>12-37</td>
</tr>
<tr>
<td>41-70</td>
<td>46-71/72</td>
<td>44-72</td>
<td>44-72</td>
</tr>
<tr>
<td>88-124</td>
<td>88-123/124</td>
<td>88-123</td>
<td>88-123</td>
</tr>
<tr>
<td>166-183</td>
<td>166-184</td>
<td>166-184</td>
<td>166-184</td>
</tr>
<tr>
<td>191-213</td>
<td>191/192-213</td>
<td>192-213</td>
<td>192-213</td>
</tr>
<tr>
<td>241-266</td>
<td>241-266</td>
<td>241-266</td>
<td>241-266</td>
</tr>
<tr>
<td>276-306</td>
<td>276-305/306</td>
<td>276-306</td>
<td>276-306</td>
</tr>
<tr>
<td>337-369</td>
<td>337-369</td>
<td>337-369</td>
<td>337-369</td>
</tr>
<tr>
<td>375-395</td>
<td>375-395</td>
<td>375-395</td>
<td>375-395</td>
</tr>
<tr>
<td>399-424</td>
<td>399-423/424</td>
<td>399-423</td>
<td>399-423</td>
</tr>
<tr>
<td>447-477</td>
<td>447-476/477</td>
<td>447-476</td>
<td>447-476</td>
</tr>
<tr>
<td>483-513</td>
<td>483-511</td>
<td>483-510</td>
<td>483-510</td>
</tr>
</tbody>
</table>

Figure B.1: Superposition of LeuT PDB #3TT3 (pink/violet helix ends, transparent) with 2A65 (red/blue helix ends, more opaque colors).
### B.1.2 Mhp1

Table B.3: Helix ends definition of Mhp1 PDB #2JLO. "accelrys" stands for Accelrys Discovery Studio, "dssp" for DSSP. The resolution of #2JLO is 2.85 Å

<table>
<thead>
<tr>
<th>2JLO paper [37]</th>
<th>2JLO accelrys</th>
<th>2JLO dssp</th>
<th>taken for modelling</th>
</tr>
</thead>
<tbody>
<tr>
<td>28-55</td>
<td>29-54</td>
<td>29-55</td>
<td>29-55</td>
</tr>
<tr>
<td>57-86</td>
<td>58-84</td>
<td>57-84</td>
<td>57-84</td>
</tr>
<tr>
<td>101-137</td>
<td>104-134</td>
<td>104-134</td>
<td>104-134</td>
</tr>
<tr>
<td>142-159</td>
<td>142-158</td>
<td>142-158</td>
<td>142-158</td>
</tr>
<tr>
<td>161-191</td>
<td>160-189</td>
<td>160-189</td>
<td>160-189</td>
</tr>
<tr>
<td>209-231</td>
<td>209-231</td>
<td>209-231</td>
<td>209-231</td>
</tr>
<tr>
<td>243-268</td>
<td>242-278</td>
<td>242-278</td>
<td>242-278</td>
</tr>
<tr>
<td>297-326</td>
<td>296-327</td>
<td>296-327</td>
<td>296-327</td>
</tr>
<tr>
<td>336-350</td>
<td>336-349/349</td>
<td>336-348</td>
<td>336-348</td>
</tr>
<tr>
<td>359-385</td>
<td>371-382</td>
<td>371-382</td>
<td>371-382</td>
</tr>
<tr>
<td>409-424</td>
<td>409-420</td>
<td>409-420</td>
<td>409-420</td>
</tr>
<tr>
<td>429-451</td>
<td>429-446/450</td>
<td>426-445</td>
<td>429-445</td>
</tr>
</tbody>
</table>
B.2 DEER data analysis for modelling purposes

Figure B.2: PutP DEER traces of liposomal samples of reference residue 41. From left to right: normalized time trace, background-corrected data and Tikhonov regularized distance distribution. Traces below the dashed line were measured at Q-band frequencies. All samples prepared and all X-band measurements performed by Daniel Hilger. Q-band traces acquired in the course of this thesis.
Figure B.3: PutP DEER traces of liposomal samples of reference residue 371. From left to right: normalized time trace, background-corrected data and Tikhonov regularized distance distribution. Traces below the dashed line were measured at Q-band frequencies. All samples prepared and all X-band measurements performed by Daniel Hilger. Q-band traces acquired in the course of this thesis.
Figure B.4: PutP DEER X-band traces of liposomal samples of reference residue 62. From left to right: normalized time trace, background-corrected data and Tikhonov regularized distance distribution. All samples prepared and all measurements performed by Daniel Hilger.
Figure B.5: PutP DEER X-band traces of liposomal samples of reference residue 446. From left to right: normalized time trace, background-corrected data and Tikhonov regularized distance distribution. All samples prepared and all measurements performed by Daniel Hilger.
Appendix B. Results

Figure B.6: PutP DEER Q-band traces of liposomal samples of reference residue 349. From left to right: normalized time trace, background-corrected data and Tikhonov regularized distance distribution. All samples prepared by Daniel Hilger. Experimental traces acquired in the course of this thesis.
Figure B.7: PutP DEER Q-band traces of liposomal samples of reference residue 326. From left to right: normalized time trace, background-corrected data and Tikhonov regularized distance distribution. All samples prepared by Daniel Hilger. Experimental traces acquired in the course of this thesis.
Figure B.8: PutP DEER Q-band traces of liposomal samples of helix I (additional constraints). From left to right: normalized time trace, background-corrected data and Tikhonov regularized distance distribution. All samples prepared by Daniel Hilger. Experimental traces acquired in the course of this thesis.
## B.3 Constraint list

Table B.4: Constraint list for vSGLT modelling, 75 distances.

<table>
<thead>
<tr>
<th>res1</th>
<th>res2</th>
<th>$r_{\text{mean}}$ [nm]</th>
<th>$\sigma_r$ [nm]</th>
<th>res1</th>
<th>res2</th>
<th>$r_{\text{mean}}$ [nm]</th>
<th>$\sigma_r$ [nm]</th>
</tr>
</thead>
<tbody>
<tr>
<td>28</td>
<td>53</td>
<td>3.85</td>
<td>0.22</td>
<td>348</td>
<td>479</td>
<td>4.53</td>
<td>0.44</td>
</tr>
<tr>
<td>28</td>
<td>280</td>
<td>4.30</td>
<td>0.36</td>
<td>380</td>
<td>28</td>
<td>2.91</td>
<td>0.42</td>
</tr>
<tr>
<td>28</td>
<td>392</td>
<td>2.38</td>
<td>0.55</td>
<td>380</td>
<td>79</td>
<td>3.86</td>
<td>0.24</td>
</tr>
<tr>
<td>28</td>
<td>447</td>
<td>4.93</td>
<td>0.48</td>
<td>380</td>
<td>108</td>
<td>2.57</td>
<td>0.14</td>
</tr>
<tr>
<td>53</td>
<td>79</td>
<td>4.30</td>
<td>0.20</td>
<td>380</td>
<td>124</td>
<td>1.63</td>
<td>0.20</td>
</tr>
<tr>
<td>53</td>
<td>108</td>
<td>2.64</td>
<td>0.32</td>
<td>380</td>
<td>178</td>
<td>2.44</td>
<td>0.26</td>
</tr>
<tr>
<td>53</td>
<td>178</td>
<td>4.76</td>
<td>0.17</td>
<td>380</td>
<td>187</td>
<td>2.97</td>
<td>0.28</td>
</tr>
<tr>
<td>53</td>
<td>187</td>
<td>2.28</td>
<td>0.32</td>
<td>380</td>
<td>275</td>
<td>1.59</td>
<td>0.33</td>
</tr>
<tr>
<td>53</td>
<td>275</td>
<td>2.16</td>
<td>0.22</td>
<td>380</td>
<td>280</td>
<td>2.65</td>
<td>0.31</td>
</tr>
<tr>
<td>53</td>
<td>280</td>
<td>1.85</td>
<td>0.18</td>
<td>380</td>
<td>447</td>
<td>2.43</td>
<td>0.30</td>
</tr>
<tr>
<td>53</td>
<td>380</td>
<td>3.33</td>
<td>0.10</td>
<td>380</td>
<td>453</td>
<td>2.17</td>
<td>0.16</td>
</tr>
<tr>
<td>53</td>
<td>392</td>
<td>3.89</td>
<td>0.31</td>
<td>380</td>
<td>472</td>
<td>3.14</td>
<td>0.36</td>
</tr>
<tr>
<td>53</td>
<td>447</td>
<td>3.12</td>
<td>0.39</td>
<td>380</td>
<td>501</td>
<td>2.95</td>
<td>0.35</td>
</tr>
<tr>
<td>53</td>
<td>453</td>
<td>4.69</td>
<td>0.16</td>
<td>392</td>
<td>108</td>
<td>3.14</td>
<td>0.35</td>
</tr>
<tr>
<td>53</td>
<td>472</td>
<td>5.22</td>
<td>0.37</td>
<td>392</td>
<td>124</td>
<td>2.29</td>
<td>0.45</td>
</tr>
<tr>
<td>53</td>
<td>501</td>
<td>5.27</td>
<td>0.31</td>
<td>392</td>
<td>178</td>
<td>2.06</td>
<td>0.45</td>
</tr>
<tr>
<td>79</td>
<td>82</td>
<td>2.20</td>
<td>0.11</td>
<td>392</td>
<td>187</td>
<td>2.82</td>
<td>0.47</td>
</tr>
<tr>
<td>79</td>
<td>158</td>
<td>3.19</td>
<td>0.19</td>
<td>392</td>
<td>275</td>
<td>2.41</td>
<td>0.49</td>
</tr>
<tr>
<td>79</td>
<td>162</td>
<td>2.16</td>
<td>0.23</td>
<td>392</td>
<td>280</td>
<td>3.30</td>
<td>0.40</td>
</tr>
<tr>
<td>79</td>
<td>212</td>
<td>2.91</td>
<td>0.25</td>
<td>392</td>
<td>380</td>
<td>1.68</td>
<td>0.35</td>
</tr>
<tr>
<td>79</td>
<td>250</td>
<td>2.29</td>
<td>0.31</td>
<td>392</td>
<td>447</td>
<td>3.79</td>
<td>0.45</td>
</tr>
<tr>
<td>79</td>
<td>312</td>
<td>1.13</td>
<td>0.14</td>
<td>392</td>
<td>472</td>
<td>4.53</td>
<td>0.47</td>
</tr>
<tr>
<td>79</td>
<td>348</td>
<td>3.59</td>
<td>0.39</td>
<td>392</td>
<td>453</td>
<td>3.17</td>
<td>0.40</td>
</tr>
<tr>
<td>79</td>
<td>392</td>
<td>5.15</td>
<td>0.41</td>
<td>392</td>
<td>501</td>
<td>4.12</td>
<td>0.52</td>
</tr>
<tr>
<td>79</td>
<td>419</td>
<td>1.64</td>
<td>0.25</td>
<td>472</td>
<td>82</td>
<td>3.85</td>
<td>0.32</td>
</tr>
<tr>
<td>79</td>
<td>472</td>
<td>2.46</td>
<td>0.47</td>
<td>472</td>
<td>158</td>
<td>4.25</td>
<td>0.50</td>
</tr>
<tr>
<td>79</td>
<td>479</td>
<td>1.53</td>
<td>0.31</td>
<td>472</td>
<td>162</td>
<td>2.62</td>
<td>0.45</td>
</tr>
<tr>
<td>348</td>
<td>53</td>
<td>4.16</td>
<td>0.33</td>
<td>472</td>
<td>212</td>
<td>5.08</td>
<td>0.41</td>
</tr>
<tr>
<td>348</td>
<td>82</td>
<td>3.74</td>
<td>0.21</td>
<td>472</td>
<td>250</td>
<td>2.55</td>
<td>0.23</td>
</tr>
<tr>
<td>348</td>
<td>158</td>
<td>1.78</td>
<td>0.43</td>
<td>472</td>
<td>312</td>
<td>3.06</td>
<td>0.41</td>
</tr>
<tr>
<td>348</td>
<td>162</td>
<td>3.22</td>
<td>0.43</td>
<td>472</td>
<td>348</td>
<td>5.32</td>
<td>0.54</td>
</tr>
<tr>
<td>348</td>
<td>212</td>
<td>1.60</td>
<td>0.23</td>
<td>472</td>
<td>419</td>
<td>2.15</td>
<td>0.58</td>
</tr>
<tr>
<td>348</td>
<td>250</td>
<td>5.29</td>
<td>0.27</td>
<td>472</td>
<td>423</td>
<td>2.72</td>
<td>0.42</td>
</tr>
<tr>
<td>348</td>
<td>312</td>
<td>3.16</td>
<td>0.37</td>
<td>472</td>
<td>479</td>
<td>1.58</td>
<td>0.43</td>
</tr>
<tr>
<td>348</td>
<td>380</td>
<td>4.83</td>
<td>0.24</td>
<td>10</td>
<td>250</td>
<td>5.85</td>
<td>0.26</td>
</tr>
<tr>
<td>348</td>
<td>392</td>
<td>5.16</td>
<td>0.46</td>
<td>10</td>
<td>348</td>
<td>1.45</td>
<td>0.27</td>
</tr>
<tr>
<td>348</td>
<td>419</td>
<td>3.58</td>
<td>0.58</td>
<td>472</td>
<td>5.50</td>
<td>0.50</td>
<td></td>
</tr>
<tr>
<td>348</td>
<td>423</td>
<td>3.05</td>
<td>0.24</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
B.4 Modelling results

B.4.1 vSGLT

Figure B.9: Result of vSGLT modelling with 378 simulated constraints between labels, $c = 1$. Superposition of vSGLT the central structure of the modelled ensemble (transparent) and vSGLT target structure (opaque) highlighted for all helices. Rmsd vs real $\text{rmsd}_{\text{real}} = 7.61 \, \text{Å} \; (14 \, \text{helices})$, ensemble $\text{rmsd}_{\text{ensemble}} = 12.35 \, \text{Å}$. Red dots (structure)/pink dots (model) - cytoplasmic helix ends, blue points (structure)/violet points (model) - periplasmic helix ends.
Appendix B. Results

Figure B.10: Result of vSGLT modelling with 378 simulated constraints between labels, $c = 0.6$. Superposition of the vSGLT central structure in the modelled ensemble (transparent) and vSGLT target structure (opaque) highlighted for all helices. Rmsd vs real $\text{rmsd}_{\text{real}} = 8.92 \, \text{Å}$ (14 helices), ensemble $\text{rmsd}_{\text{ensemble}} = 10.38 \, \text{Å}$. Red dots (structure)/pink dots (model) - cytoplasmic helix ends, blue points (structure)/violet points (model) - periplasmic helix ends.

Figure B.11: Result of vSGLT modelling with 378 simulated constraints between labels, $c = 0.58$. Superposition of the vSGLT structure with best $\text{rmsd}$ from native that can be found within the modelled ensemble (transparent) and vSGLT target structure (opaque) highlighted for all helices. $\text{rmsd}_{\text{real}} = 7.17 \, \text{Å}$ (14 helices). Red dots (structure)/pink dots (model) - cytoplasmic helix ends, blue points (structure)/violet points (model) - periplasmic helix ends.
Figure B.12: Result of vSGLT modelling with 75 simulated constraints between labels, $c = 0.58$ and template data included. Superposition of the vSGLT structure with best rmsd from native that can be found within the modelled ensemble (transparent) and vSGLT target structure (opaque) highlighted for all helices. Rmsd vs real rmsd$_{\text{real}} = 9.56$ Å (13 helices). Red dots (structure)/pink dots (model) - cytoplasmic helix ends, blue points (structure)/violet points (model) - periplasmic helix ends.

Figure B.13: Result of vSGLT modelling with 378 simulated constraints between helix ends, $r = r_{\text{mean}} \pm 3.8$ Å. Superposition of the vSGLT structure with best rmsd from native that can be found within the modelled ensemble (transparent) and vSGLT target structure (opaque) highlighted for all helices. Rmsd vs real rmsd$_{\text{real}} = 2.87$ Å (14 helices). Red dots (structure)/pink dots (model) - cytoplasmic helix ends, blue points (structure)/violet points (model) - periplasmic helix ends.
Figure B.14: Result of vSGLT modelling with 75 simulated constraints between helix ends, \( r = r_{\text{mean}} \pm 5.8 \, \text{Å} \) and template data included. Superposition of the vSGLT structure with best rmsd from native that can be found within the modelled ensemble (transparent) and vSGLT target structure (opaque) highlighted for all helices. Rmsd vs real rmsd\(_{\text{real}} = 6.14 \, \text{Å} \) (13 helices). Red dots (structure)/pink dots (model) - cytoplasmic helix ends, blue points (structure)/violet points (model) - periplasmic helix ends.
B.4.1.1 Placing non-core helices

<table>
<thead>
<tr>
<th>Clustering</th>
<th>100%</th>
<th>70%</th>
<th>50%</th>
<th>30%</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>100%</strong></td>
<td><img src="image1" alt="Image" /></td>
<td><img src="image2" alt="Image" /></td>
<td><img src="image3" alt="Image" /></td>
<td><img src="image4" alt="Image" /></td>
</tr>
<tr>
<td>36.20 Å</td>
<td>36.20 Å</td>
<td>42.85 Å</td>
<td>47.22 Å</td>
<td></td>
</tr>
<tr>
<td><strong>70%</strong></td>
<td><img src="image5" alt="Image" /></td>
<td><img src="image6" alt="Image" /></td>
<td><img src="image7" alt="Image" /></td>
<td><img src="image8" alt="Image" /></td>
</tr>
<tr>
<td>36.20 Å</td>
<td>29.79 Å</td>
<td>42.85 Å</td>
<td>30.99 Å</td>
<td></td>
</tr>
<tr>
<td><strong>50%</strong></td>
<td><img src="image9" alt="Image" /></td>
<td><img src="image10" alt="Image" /></td>
<td><img src="image11" alt="Image" /></td>
<td><img src="image12" alt="Image" /></td>
</tr>
<tr>
<td>27.66 Å</td>
<td>29.79 Å</td>
<td>27.66 Å</td>
<td>44.21 Å</td>
<td></td>
</tr>
<tr>
<td><strong>30%</strong></td>
<td><img src="image13" alt="Image" /></td>
<td><img src="image14" alt="Image" /></td>
<td><img src="image15" alt="Image" /></td>
<td><img src="image16" alt="Image" /></td>
</tr>
<tr>
<td>22.75 Å</td>
<td>29.79 Å</td>
<td>29.79 Å</td>
<td>29.79 Å</td>
<td></td>
</tr>
<tr>
<td><strong>EPR pre-selection</strong></td>
<td><img src="image17" alt="Image" /></td>
<td><img src="image18" alt="Image" /></td>
<td><img src="image19" alt="Image" /></td>
<td><img src="image20" alt="Image" /></td>
</tr>
<tr>
<td>36.20 Å</td>
<td>30.99 Å</td>
<td>41.46 Å</td>
<td>44.83 Å</td>
<td></td>
</tr>
<tr>
<td><strong>50%</strong></td>
<td><img src="image21" alt="Image" /></td>
<td><img src="image22" alt="Image" /></td>
<td><img src="image23" alt="Image" /></td>
<td><img src="image24" alt="Image" /></td>
</tr>
<tr>
<td>27.66 Å</td>
<td>27.66 Å</td>
<td>42.20 Å</td>
<td>44.83 Å</td>
<td></td>
</tr>
<tr>
<td><strong>30%</strong></td>
<td><img src="image25" alt="Image" /></td>
<td><img src="image26" alt="Image" /></td>
<td><img src="image27" alt="Image" /></td>
<td><img src="image28" alt="Image" /></td>
</tr>
<tr>
<td>27.66 Å</td>
<td>27.66 Å</td>
<td>27.66 Å</td>
<td>27.66 Å</td>
<td></td>
</tr>
<tr>
<td><strong>exp distribution pre-selection</strong></td>
<td><img src="image29" alt="Image" /></td>
<td><img src="image30" alt="Image" /></td>
<td><img src="image31" alt="Image" /></td>
<td><img src="image32" alt="Image" /></td>
</tr>
<tr>
<td>36.20 Å</td>
<td>30.99 Å</td>
<td>41.46 Å</td>
<td>44.83 Å</td>
<td></td>
</tr>
<tr>
<td><strong>50%</strong></td>
<td><img src="image33" alt="Image" /></td>
<td><img src="image34" alt="Image" /></td>
<td><img src="image35" alt="Image" /></td>
<td><img src="image36" alt="Image" /></td>
</tr>
<tr>
<td>27.66 Å</td>
<td>27.66 Å</td>
<td>42.20 Å</td>
<td>44.83 Å</td>
<td></td>
</tr>
<tr>
<td><strong>30%</strong></td>
<td><img src="image37" alt="Image" /></td>
<td><img src="image38" alt="Image" /></td>
<td><img src="image39" alt="Image" /></td>
<td><img src="image40" alt="Image" /></td>
</tr>
<tr>
<td>27.66 Å</td>
<td>27.66 Å</td>
<td>27.66 Å</td>
<td>27.66 Å</td>
<td></td>
</tr>
</tbody>
</table>

Figure B.15: Superposition of vSGLT models obtained with 75 simulated constraints between labels, template information and $c = 0.58$ (transparent, pink helix ends) and target structure (opaque, red helix ends) shown for helix I under various reduction conditions; periplasmic view.
Figure B.16: Superposition of vSGLT models obtained with 75 simulated constraints between labels, template information and $c = 0.58$ (transparent, pink helix ends) and target structure (opaque, red helix ends) shown for helix XII under various reduction conditions; periplasmic view.
Figure B.17: Superposition of vSGLT models obtained with 75 simulated constraints between labels, template information and $c = 0.58$ (transparent, pink helix ends) and target structure (opaque, red helix ends) shown for helix XIII under various reduction conditions; periplasmic view.
Figure B.18: Superposition of vSGLT models obtained with 75 simulated constraints between labels, template information and $c = 0.15$ (transparent, pink helix ends) and target structure (opaque, red helix ends) shown for helix I under various reduction conditions; periplasmic view.
Figure B.19: Superposition of vSGLT models obtained with 75 simulated constraints between labels, template information and $c = 0.15$ (transparent, pink helix ends) and target structure (opaque, red helix ends) shown for helix XII under various reduction conditions; periplasmic view.
Figure B.20: Superposition of vSGLT models obtained with 75 simulated constraints between labels, template information and $c = 0.15$ (transparent, pink helix ends) and target structure (opaque, red helix ends) shown for helix XIII under various reduction conditions; periplasmic view.
B.4.2 Rho

Figure B.21: Result of Rho modelling with 91 simulated constraints between labels, \( c = 0.328 \). Superposition of the Rho central structure of the modelled ensemble (transparent) and Rho target structure (opaque) highlighted for all helices. Rmsd vs real \( \text{rmsd}_{\text{real}} = 7.25 \) Å, ensemble rmsd \( \text{rmsd}_{\text{ensemble}} = 12.82 \) Å. Red dots (structure)/pink dots (model) - cytoplasmic helix ends, blue points (structure)/violet points (model) - periplasmic helix ends.

Figure B.22: Result of Rho modelling with 91 simulated constraints between helix ends, \( r = r_{\text{mean}} \pm 3 \) Å. Superposition of the Rho central structure of the modelled ensemble (transparent) and Rho target structure (opaque) highlighted for all helices. Rmsd vs real \( \text{rmsd}_{\text{real}} = 3.49 \) Å, ensemble rmsd \( \text{rmsd}_{\text{ensemble}} = 9.24 \) Å. Red dots (structure)/pink dots (model) - cytoplasmic helix ends, blue points (structure)/violet points (model) - periplasmic helix ends.
Figure B.23: Result of Rho modelling with 91 simulated constraints between helix ends, \( r = r_{\text{mean}} \pm 2 \text{ Å} \). Superposition of the Rho central structure of the modelled ensemble (transparent) and Rho target structure (opaque) highlighted for all helices. Rmsd vs real \( \text{rmsd}_{\text{real}} = 2.90 \text{ Å} \), ensemble rmsd \( \text{rmsd}_{\text{ensemble}} = 5.97 \text{ Å} \). Red dots (structure)/pink dots (model) - cytoplasmic helix ends, blue points (structure)/violet points (model) - periplasmic helix ends.

Figure B.24: Result of Rho modelling with 91 simulated constraints between helix ends, \( r = r_{\text{mean}} \pm 1 \text{ Å} \). Superposition of the Rho central structure of the modelled ensemble (transparent) and Rho target structure (opaque) highlighted for all helices. Rmsd vs real \( \text{rmsd}_{\text{real}} = 2.70 \text{ Å} \), ensemble rmsd \( \text{rmsd}_{\text{ensemble}} = 0.71 \text{ Å} \). Red dots (structure)/pink dots (model) - cytoplasmic helix ends, blue points (structure)/violet points (model) - periplasmic helix ends.
Figure B.25: Result of Rho modelling with 91 simulated constraints between helix ends, \( r = r_{\text{mean}} \pm 0.143 \, \text{Å} \). Superposition of the Rho central structure of the modelled ensemble (transparent) and Rho target structure (opaque) highlighted for all helices. Rmsd vs real \( \text{rmsd}_{\text{real}} = 2.78 \, \text{Å} \), ensemble \( \text{rmsd}_{\text{ensemble}} = 0.02 \, \text{Å} \). Red dots (structure)/pink dots (model) - cytoplasmic helix ends, blue points (structure)/violet points (model) - periplasmic helix ends.

Figure B.26: Result of Rho modelling with 91 simulated constraints between labels, \( c = 0.58 \). Superposition of the Rho structure with best rmsd from native that can be found within the modelled ensemble (transparent) and Rho target structure (opaque). Rmsd vs real \( \text{rmsd}_{\text{real}} = 6.88 \, \text{Å} \). Red dots (structure)/pink dots (model) - cytoplasmic helix ends, blue points (structure)/violet points (model) - periplasmic helix ends.
Appendix B. Results

Figure B.27: Result of Rho modelling with 91 simulated constraints between labels, $c = 0.328$. Superposition of the Rho structure with best rmsd from native that can be found within the modelled ensemble (transparent) and Rho target structure (opaque). Rmsd vs real rmsd, $r_{real} = 6.38$ Å. Red dots (structure)/pink dots (model) - cytoplasmic helix ends, blue points (structure)/violet points (model) - periplasmic helix ends.

Figure B.28: Result of Rho modelling with 91 simulated constraints between labels, $c = 0.31$. Superposition of the Rho structure with best rmsd from native that can be found within the modelled ensemble (transparent) and Rho target structure (opaque). Rmsd vs real rmsd, $r_{real} = 7.42$ Å. Red dots (structure)/pink dots (model) - cytoplasmic helix ends, blue points (structure)/violet points (model) - periplasmic helix ends.
Appendix B. Results

Figure B.29: Result of Rho modelling with 91 simulated constraints between helix ends, $r = r_{\text{mean}} \pm 5.8$ Å. Superposition of the Rho structure with best rmsd from native that can be found within the modelled ensemble (transparent) and Rho target structure (opaque). Rmsd vs real rmsd$_{\text{real}} = 3.22$ Å. Red dots (structure)/pink dots (model) - cytoplasmic helix ends, blue points (structure)/violet points (model) - periplasmic helix ends.

Figure B.30: Result of Rho modelling with 91 simulated constraints between helix ends, $r = r_{\text{mean}} \pm 1$ Å. Superposition of the Rho structure with best rmsd from native that can be found within the modelled ensemble (transparent) and Rho target structure (opaque). Rmsd vs real rmsd$_{\text{real}} = 2.47$ Å. Red dots (structure)/pink dots (model) - cytoplasmic helix ends, blue points (structure)/violet points (model) - periplasmic helix ends.
Figure B.31: Result of Rho modelling with 91 simulated constraints between helix ends, \( r = r_{mean} \pm 5.8 \, \text{Å} \). Reduction based on geometrical criterion - the number of structures taken for re-clustering indicated. Superposition of the Rho central structure in the modelled ensemble (transparent) and Rho target structure (opaque) highlighted for full structures and all helices separately. Red dots (structure)/pink dots (model) - cytoplasmic helix ends, blue points (structure)/violet points (model) - periplasmic helix ends.
Appendix B. Results

Figure B.32: Result of Rho modelling with 91 simulated constraints between helix ends, $r = r_{\text{mean}} \pm 3 \text{ Å}$. Reduction based on geometrical criterion - the number of structures taken for re-clustering indicated. Superposition of the Rho central structure in the modelled ensemble (transparent) and Rho target structure (opaque) highlighted for full structures and all helices separately. Red dots (structure)/pink dots (model) - cytoplasmic helix ends, blue points (structure)/violet points (model) - periplasmic helix ends.
Figure B.33: Result of Rho modelling with 91 simulated constraints between helix ends, \( r = r_{\text{mean}} \pm 1 \text{ Å} \). Reduction based on geometrical criterion - the number of structures taken for re-clustering indicated. Superposition of the Rho central structure in the modelled ensemble (transparent) and Rho target structure (opaque) highlighted for full structures and all helices separately. Red dots (structure)/pink dots (model) - cytoplasmic helix ends, blue points (structure)/violet points (model) - periplasmic helix ends.
Figure B.34: Result of Rho modelling with 91 simulated constraints between labels, \( c = 0.58 \). Reduction based on geometrical criterion - the number of structures taken for re-clustering indicated. Superposition of the Rho central structure in the modelled ensemble (transparent) and Rho target structure (opaque) highlighted for full structures and all helices separately. Red dots (structure)/pink dots (model) - cytoplasmic helix ends, blue points (structure)/violet points (model) - periplasmic helix ends.
Figure B.35: Result of Rho modelling with 91 simulated constraints between labels, $c = 0.31$. Reduction based on geometrical criterion - the number of structures taken for re-clustering indicated. Superposition of the Rho central structure in the modelled ensemble (transparent) and Rho target structure (opaque) highlighted for full structures and all helices separately. Red dots (structure)/pink dots (model) - cytoplasmic helix ends, blue points (structure)/violet points (model) - periplasmic helix ends.
Figure B.36: Effect of removing improper structures from the ensemble of Rho modelled with 91 simulated constraints between helix ends, \( r = r_{\text{mean}} \pm 5.8 \text{ Å} \). Superposition of the Rho central structure of the modelled ensemble (transparent) and Rho target structure (opaque) highlighted for full structures and all helices separately. Red dots (structure)/pink dots (model) - cytoplasmic helix ends, blue points (structure)/violet points (model) - periplasmic helix ends.

Figure B.37: Effect of removing improper structures from the ensemble of Rho modelled with 91 simulated constraints between labels, \( c = 0.58 \). Superposition of the Rho central structure of the modelled ensemble (transparent) and Rho target structure (opaque) highlighted for full structures and all helices separately. Red dots (structure)/pink dots (model) - cytoplasmic helix ends, blue points (structure)/violet points (model) - periplasmic helix ends.
Appendix B. Results

Figure B.38: Result of Rho modelling with 91 simulated constraints between helix ends, \( r = r_{\text{mean}} \pm 5.8 \, \text{Å} \). Reduction based on EPR pre-selection - the number of structures fulfilling experimental constraints best that were taken for clustering indicated. Selecting 70% of best EPR structures results in the central structure of the modelled ensemble as for all structures clustered (4.17). Superposition of the Rho central structure of the ensemble (transparent) and Rho target structure (opaque) highlighted for full structures and all helices separately. Red dots (structure)/pink dots (model) - cytoplasmic helix ends, blue points (structure)/violet points (model) - periplasmic helix ends.

Figure B.39: Result of Rho modelling with 91 simulated constraints between helix ends, \( r = r_{\text{mean}} \pm 3 \, \text{Å} \). Reduction based on EPR pre-selection - the number of structures fulfilling experimental constraints best that were taken for clustering indicated. Superposition of the Rho central structure of the modelled ensemble (transparent) and Rho target structure (opaque) highlighted for full structures and all helices separately. Red dots (structure)/pink dots (model) - cytoplasmic helix ends, blue points (structure)/violet points (model) - periplasmic helix ends.
Appendix B. Results

Figure B.40: Result of Rho modelling with 91 simulated constraints between helix ends, $r = r_{\text{mean}} \pm 1$ Å. Reduction based on EPR pre-selection - the number of structures fulfilling experimental constraints best that were taken for clustering indicated. Superposition of the Rho central structure of the modelled ensemble (transparent) and Rho target structure (opaque) highlighted for full structures and all helices separately. Red dots (structure)/pink dots (model) - cytoplasmic helix ends, blue points (structure)/violet points (model) - periplasmic helix ends.

Figure B.41: Result of Rho modelling with 91 simulated constraints between spin labels, $c = 0.31$. Reduction based on EPR pre-selection - the number of structures fulfilling experimental constraints best that were taken for clustering indicated. Selecting 70% of best EPR structures results in the central structure of the modelled ensemble as for all structures clustered (4.11). Superposition of the Rho central structure of the ensemble (transparent) and Rho target structure (opaque) highlighted for full structures and all helices separately. Red dots (structure)/pink dots (model) - cytoplasmic helix ends, blue points (structure)/violet points (model) - periplasmic helix ends.
B.4.3 Ambiguities in placing non-core helix I of PutP

Figure B.42: Result of PutP modelling with 72 experimental constraints and template constraints generated from the vSGLT #2XQ2 template, $r = r_{\text{mean}} \pm 2\sigma_r$. Measurements between indicated helix ends are supposed to distinguish a correct arrangement of helix I.
B.5 eL4 functional studies

Figure B.43: A Influence of free label removal in the 190R1/371R1 variant on DEER traces. B Second preparation of 298R1/326R1 and its mutants. Left column: normalized primary DEER data with background fits; middle: form factors after background correction scaled to the same modulation depth; right column: Tikhonov regularized distance distributions.
Figure B.44: A DEER distance measurements of doubly-labelled PutP variants. Mutants indicated on the figure were measured in the ligand-free state (apo, violet), upon addition of 50 mM NaCl (Na\(^+\), pink) and with 50 mM NaCl and 10 mM proline (Na\(^+\)/P, green). Left column: normalized primary DEER data with background fits; middle: form factors after background correction scaled to the same modulation depth; right column: Tikhonov regularized distance distributions. B Complementary DEER traces in ligand-free state used to restrain the modelling. Columns assign as in A. Data collected either at Q band (294R1/326R1) or X band (remaining).
Bibliography


mechanism of the glycerol-3-phosphate transporter from *Escherichia coli*. *Science*,

[14] L. R. Forrest, Y.-W. Zhang, M. T. Jacobs, J. Gesmonde, L. Xie, B. H. Honig, and
G. Rudnick. A mechanism for alternating access in neurotransmitter transporters.

[15] T. Shimamura, S. Weyand, O. Beckstein, N. G. Rutherford, J. M. Hadden,
D. Sharples, M. S. P. Sansom, S. Iwata, P. J. F. Henderson, and A. D.
Cameron. Molecular basis of alternating access membrane transport by the

[16] Y. Fang, H. Jayaram, T. Shane, L. Kolmakova-Partensky, F. Wu, C. Williams,
Y. Xiong, and C. Miller. Structure of a prokaryotic virtual proton pump at 3.2 A


[18] L. Kowalczyk, M. Ratera, A. Paladino, P. Bartoccion, E. Errasti-Murugarren,
Vázquez-Ibar, and M. Palacín. Molecular basis of substrate-induced permeation

[19] L. Devel, S. Garcia, B. Czarny, F. Beau, E. LaJeunesse, L. Vera, D. Georgiadis,
E. Stura, and V. Dive. Insights from selective non-phosphinic inhibitors of MMP-


Molecular basis of transport and regulation in the Na⁺/betaine symporter BetP.

transporter and insights into the antiport mechanism. *Nat Struct Mol Biol*, 17

[23] S. Schulze, S. Köster, U. Geldmacher, A. C. Terwisscha van Scheltinga, and
W. Kühlbrandt. Structural basis of Na⁺-independent and cooperative sub-


dopamine transporter in complex with nisoxetine and reboxetine. *Nat Struct


Acknowledgements

First of all, I would like to thank Prof. Gunnar Jeschke for giving me the opportunity to work in his group. For being there always to provide an invaluable advice and at the same time not restricting the directions of work. His engagement in supervised projects ensured constant help and proper guidance of the research without any redundant pressure. His open-mindedness enabled to build a friendly and amazing group, where people enjoy not only working together, but also spending time together. And lastly, I need to thank for his kindness, understanding and flexibility, without which it would not have been possible for me to reconcile the PhD studies and the family. It was a great pleasure to be a part of his group.

I would also like to thank Prof. Roland Riek for finding the time to review my thesis despite many other scientific obligations.

My huge gratitude goes to Dr. Yevhen Polyhach as my direct supervisor. I am very thankful for his support, introduction to new type of experiments and modelling. Long hours that he devoted to discussions regarding results and further proceedings of my project were always fruitful. His assistance during all this time was inestimable. Personal conversations were entertaining and bringing an instant “relief” at the stressful stages of my work. I also truly appreciate all his understanding to my personal ups and downs.

I would like to thank Prof. Enrica Bordignon for all the knowledge she shared with me during a short-standing common project and unconditioned help I could have always asked for. Her support and encouraging words were buoying me up and making believing that everything is possible. Enjoyable conversations complemented the amazing time I had chance to spend with her.

I also thank all the people from the group of Prof. Heinrich Hung (Munich, Germany) that were engaged in the common project: Dr. Daniel Hilger, Dr. Michael Raba and Susanne Bracher. I could have conducted all my measurements, data analysis and modelling thanks to their immense biochemical commitment.

A big thank goes to Simon Böhm for the support, help and long discussion, not only scientific. And to Gabriela Drabik for amazing cheering and believe in me even after leaving ETH.
I also thank all present and former members of the EPR group at ETH for the enjoyable conversations and support. Some of them aided me at the first stages of my PhD, another cheered when this thesis was being written. My acknowledgements go to Kristina Comiotto, Dr. Maxim Yulikov, Dr. Inés Garcia Rubio, Dr. Daniel Klose, Dr. Takuya Segawa, Dr. Vidmantas Kalendra, Dr. Benesh Joseph, Dr. Tona von Hagens, Dr. Petra Lüders, Dr. Sahand Razzaghi, Dr. Andrin Doll, Dr. Luca Garbuio, Dr. Udo Kielmann, Oliver Oberhänsli, Rene Tschaggelar, Thomas Kohn, Stephan Pribitzer, Irina Ritsch, Katharina Keller, Frauke Breitgoff, Christoph Gmeiner, Janne Soetbeer and all the master and semester students I had opportunity to meet. It was amazing and joyful time to work amongst them.

I would like to thank Veronika Sieger for help and support concerning administrative issues as well as for all warm words I got from her.

My deepest gratitude goes to my parents. For the everlasting love and support throughout my life. They have always believed in me and were directing invisibly in the proper directions since the childhood. Without the motivation and encouragement I got from them I would not be who I am.

Finally, I would like to thank my best friends, Agnieszka Adamska-Venkatesh, Dominika Żurek-Biesiada and Magdalena Grażul. They have always open heart and mind for me despite long distances between us. They serve a good advice and discussions on any topic. I can count on them in any life situation.