3D modeling of the putative human surfactant proteins SP-G and SP-H and simulations in a pulmonary surfactant model system

Dissertation

zur Erlangung des Doktorgrades der Naturwissenschaften
(Dr. rer. nat.)

der

Naturwissenschaftlichen Fakultät II
Chemie, Physik und Mathematik

der Martin-Luther-Universität
Halle-Wittenberg

vorgelegt von

Herr Dipl.-Bioinf. Felix Rausch
geb. am 6.1.1985 in Bad Langensalza

Die vorliegende Arbeit wurde am Leibniz-Institut für Pflanzenbiochemie im Zeitraum von Oktober 2009 bis März 2013 angefertigt.

Gutachter:

1. Prof. Dr. Ludger A. Wessjohann
2. PD Dr. Harald Lanig

Tag der mündlichen Prüfung: 30.09.2015

Danksagung

Zunächst möchte ich mich vielmals bei *Prof. Dr. Ludger A. Wessjohann* bedanken, welcher mir die Anfertigung dieser Arbeit in seiner Abteilung am Leibniz-Institut für Pflanzenbiochemie ermöglicht hat und mir als betreuender Hochschullehrer jederzeit beratend zur Seite stand.

Ein besonderer Dank gilt *PD Dr. Wolfgang Brandt* für die Unterstützung und die Geduld in den letzten Jahren. Seine hervorragende Betreuung sowie seine ständige Hilfs- und Diskussionsbereitschaft sind maßgeblich für die Entstehung dieser Arbeit verantwortlich.

Prof. Dr. Lars Bräuer und *Prof. Dr. Friedrich Paulsen* danke ich vielmals dafür, dass ich an einem so interessanten und vielschichtigen Thema arbeiten durfte und auch weiterhin darf. Ihrer Hilfe und ihren wertvollen Ratschlägen ist ebenfalls das Gelingen dieser Arbeit zu verdanken.

Außerdem danke ich *Dr. Martin Schicht* für die Durchführung der experimentellen Arbeiten zu dieser Dissertation. Dank unserer zahlreichen angeregten Diskussionen habe ich auch einen Einblick in die praktische Seite des Themas im Labor erhalten können.

Ferner möchte ich mich herzlich bei *Juliane Fischer* und *Sebastian Brauch* für das Korrekturlesen des Manuskripts in Rekordzeit bedanken.

Bei allen aktuellen und ehemaligen Mitgliedern der Arbeitsgruppe Computerchemie möchte ich mich vielmals für die tolle Zeit bedanken. Ich hätte mir keine angenehmere und produktivere Atmosphäre wünschen können, als sie bei uns im „*Aquarium*“ vorhanden war.

Natürlich gilt mein Dank auch allen anderen Mitarbeitern der Abteilung Natur- und Wirkstoffchemie des Leibniz-Instituts für Pflanzenbiochemie.

Ein großer Dank gilt meinen Freunden, insbesondere *Juliane Fischer* und *Eva Schulze* für die tatkräftige Unterstützung während des Studiums und der Promotion, sowie für die schöne Zeit auch abseits der Arbeit.

Schließlich danke ich meinen lieben Eltern und meiner Familie, ohne deren fortwährende Unterstützung, das Verständnis und die viele Geduld diese Arbeit nicht möglich gewesen wäre.

Contents

Zusammenfassung ..... **I**

Abbreviations ..... **IV**

List of Figures ..... **V**

List of Tables ..... **VI**

1. Introduction **1**

1.1. The pulmonary surfactant system 1

1.2. Surfactant proteins 2

1.3. Computational modeling and simulation of surfactant proteins 5

1.4. Motivation and objectives 6

2. Methods **9**

2.1. Protein structure modeling 9

2.1.1. Homology modeling 10

2.1.2. Threading 10

2.1.3. *Ab initio* modeling 11

2.2. Protein model quality and validation tools 12

2.2.1. PROCHECK 12

2.2.2. ProSA II 13

2.2.3. ProQ 14

2.2.4. ERRAT 14

2.2.5. VERIFY-3D 15

2.2.6. Stability test with molecular dynamics simulations 15

2.3. Prediction of posttranslational modifications 16

2.4. Molecular dynamics simulations 21

2.4.1. DPPC simulation system setup 24

2.4.2. SP-G and SP-H simulation in a lipid environment 26

2.4.3. Molecular dynamics simulation analysis 27

3. Results 31
3.1. Protein structure modeling 31
3.2. Posttranslational modifications 36
3.3. Generation of specific antibodies 39
3.4. Preparation of the protein-lipid simulation system 42
3.5. Protein-lipid molecular dynamics simulation analysis 46
3.5.1. Detailed analysis for SP-G 47
3.5.2. Detailed analysis for SP-H 52
3.5.3. General findings and summary of the protein-lipid MD simulations 57
4. Discussion 59
4.1. Protein structure modeling and posttranslational modifications 59
4.2. Findings by the performed molecular dynamics simulations 60
4.3. Cooperation of computational and experimental studies 64
5. Summary 67
6. Prospective research suggestions 68
7. Literature 70
8. Appendix 83
9. Publications and lectures 101
9.1. Publications 101
9.2. Lectures 102

Curriculum vitae 103
Eidesstattliche Erklärung 104

Zusammenfassung

Im Rahmen dieser Arbeit wurden die Sequenzen der zwei putativen Surfactantproteine SP-G (SFTA2) und SP-H (SFTA3) erstmals mithilfe von computergestützten Modellierungs- und Simulationstechniken untersucht. Die Ergebnisse dieser theoretischen Proteinstruktur-modellierungen und Moleküldynamiksimulationen wurden anschließend genutzt, um biologische oder biochemische Experimente in der Praxis zu planen oder deren Resultate zu interpretieren. Durch die Kombination dieser beiden Disziplinen war es möglich, SP-G und SP-H in unterschiedlichen Organgeweben zu lokalisieren, welche typisch für das Vorkommen von Surfactantproteinen sind. Zudem legen die gewonnen Erkenntnisse nahe, dass die physikochemischen Eigenschaften von SP-G und SP-H vergleichbar mit denen der bereits bekannten Surfactantproteine sind und dass beide Proteine ebenfalls mit Lipidsystemen interagieren und dadurch Grenzflächeneigenschaften beeinflussen können.

Zu Beginn der Arbeiten wurden dreidimensionale Strukturmodelle für SP-G und SP-H erstellt. Dabei war der klassische Ansatz einer Homologiemodellierung nicht möglich, da zu dieser Zeit keine Proteinsequenzen mit einer hohen Sequenzähnlichkeit zu SP-G oder SP-H und einer bekannten 3D-Struktur in öffentlichen Datenbanken vorhanden waren, welche als Vorlage hätten dienen können. Stattdessen wurden die Modelle mithilfe des Servers „Robetta“ erzeugt, welcher eine online verfügbare Implementierung der *ab initio* Strukturvorhersage darstellt. Die erhaltenen Modelle benötigten nur geringfügige Optimierungen, um in den gängigen Programmen zur Bewertung der Modellqualität zufriedenstellende Ergebnisse zu liefern, welche u.a. eine native Faltung der Modelle nahelegen. Zusätzlich wurden Moleküldynamik-simulationen in Wasser durchgeführt, um die Stabilität der Proteinmodelle für SP-G und SP-H zu überprüfen.

Da in der Literatur das Vorhandensein von posttranslationalen Modifikationen als essentiell für die korrekte Funktion der Surfactantproteine beschrieben wird, wurden die Sequenzen der putativen Surfactantproteine SP-G und SP-H zusätzlich auf Proteinmodifikationen untersucht. Dazu wurden die Ergebnisse von verschiedenen sequenzbasierten Vorhersagealgorithmen ausgewertet, welche einige potentielle Modifizierungsstellen für Phosphorylierungen, Palmitoylierungen und verschiedene Arten von Glykosylierungen ergaben. Die Modelle für

SP-G und SP-H wurden anschließend entsprechend dieser Vorgaben manuell um diese Modifizierungen erweitert. Moleküldynamiksimulationen dieser Modelle wurden mit den zuvor durchgeführten Simulationen der unmodifizierten Modelle verglichen und ergaben, dass die posttranslationalen Modifikationen keinen signifikanten Einfluss auf die Faltung oder allgemeine Modellqualität zeigen.

Die Herstellung von spezifischen Antikörpern auf der Basis von Antigen-Peptiden, welche ohne Wissen über die 3D-Struktur des Proteins ausgewählt wurden, führen in vielen Fällen nicht zum gewünschten Ergebnis. Der als Antigen ausgewählte Proteinabschnitt könnte durch andere Teile des Proteins verdeckt werden oder posttranslationale Modifikationen tragen, welche die erwarteten Antigen-Antikörperinteraktionen blockieren. Aus diesen Gründen schlugen vorherige Versuche fehl, spezifische Antikörper gegen SP-G und SP-H herzustellen. Mit den in dieser Arbeit beschriebenen Proteinstrukturmodellen war es möglich, Sequenzabschnitte zu identifizieren, welche in der räumlichen Struktur an der Oberfläche des Proteins liegen, keine Modifikationen tragen und zahlreiche Möglichkeiten für Antigen-Antikörperinteraktionen (d.h. Aminosäuren mit polaren Seitenketten) bieten. Bei der anschließenden Antikörperherstellung führten diese potentiellen Antigensequenzen zu spezifischen Antikörpern gegen SP-G und SP-H. Die Antikörper stellen einen großen Fortschritt in der Erforschung dieser Proteine dar. Mit ihrer Hilfe war es möglich, beide Proteine in verschiedenen Geweben nachzuweisen, welche für die Expression von Surfactantproteinen typisch sind. Zudem erlaubten die Antikörper erste funktionelle Studien im Labor.

Um die Eigenschaften von SP-G und SP-H und ihr Verhalten in ihrer natürlichen Umgebung näher untersuchen zu können, wurde ein Modellsystem etabliert, welches die grundlegenden Eigenschaften des pulmonalen Surfactantsystems reproduzieren kann. Dieses besteht ausschließlich aus dem Lipid Dipalmitoylphosphatidylcholin (DPPC), welches den Hauptbestandteil des Lungensurfactants darstellt und in einer Einzelschicht angeordnet wurde. Die Parameter für DPPC wurden im G53a6-Kraftfeld entsprechend der aktuellen Literatur angepasst und die Simulationsparameter für GROMACS dahingehend optimiert, dass die Literaturwerte für ein DPPC-Lipidsystem reproduziert werden konnten. Zusätzlich wurde das Kraftfeld auch um Parameter für die modifizierten Aminosäuren der Proteinmodelle erweitert. Aus diesen Bemühungen resultierte ein Kraftfeld, welches für Lipide sowie die unmodifizierten als auch die modifizierten Proteinmodellen gleichermaßen verwendet werden kann. Ferner konnte ein Lipidsystem etabliert werden, welches grundlegende Eigenschaften des pulmonalen Surfactants widerspiegelt und über einen längeren Simulationszeitraum stabil bleibt.

Auf der Grundlage dieses Lipidsystems wurden anschließend Simulationen der SP-G- und SP-H-Modelle durchgeführt. Für beide Proteine wurden sowohl für das unmodifizierte als auch für das modifizierte Modell sechs Simulationen gestartet, welche zu Beginn der Rechnung unterschiedlichen Orientierungen des Proteins im Bezug zur Lipidschicht aufwiesen. Somit wurden insgesamt 24 Rechnungen zu je 50 ns durchgeführt. In allen 24 Simulationen konnte die Stabilität der Proteinmodelle festgestellt werden, so dass die Auswertung der Systeme nach Abschluss der Rechnungen keine allgemeine Entfaltung oder einen Qualitätsverlust der Proteinmodelle ergab. Weiterhin zeigte jede Simulation das Bestreben des Proteins, mit der Lipidschicht zu interagieren. Im Verlauf aller durchgeführten Simulationen bewegte sich das Protein durch die Wasserphase in Richtung Lipidschicht und wies am Ende der Simulation (nach 50 ns) direkten Kontakt zu den Kopfgruppen der Lipide auf. In einigen Rechnungen zeigte sich nur eine schwache Fixierung des Proteins auf der Lipidoberfläche, unterstützt durch wenige Interaktionen zwischen polaren Aminosäureseitenketten und Kopfgruppen der Lipide. Andere Simulationen zeigten hingegen eine starke Interaktion zwischen Protein und Lipidschicht, initiiert durch vereinzelte posttranslationale Modifikationen im Bereich der Interaktionsfläche, welche wie Anker tief in die Region der Lipidkopfgruppen eindrangen und das Protein so fest an der Lipidoberfläche fixierten. Diese Ergebnisse legen nahe, dass SP-G und SP-H tatsächlich in der Lage sind, mit einem Lipidsystem zu interagieren, wie es für bereits bekannte Surfactantproteine charakteristisch ist. Es bleibt aber festzuhalten, dass die Interaktionsflächen und die ausgebildeten Interaktionstypen (polar oder hydrophob) zwischen Protein und Lipiden sehr variabel waren und hochgradig von der Orientierung des Proteins zum Simulationsstart und den posttranslationalen Modifikationen abhingen. Zudem konnte aus den Simulationen kein direkter Einfluss der Proteine auf die Stabilität oder Ordnung der Lipidschicht festgestellt werden. Jedoch konnten die Rechnungen zeigen, dass die Oberflächeneigenschaften der Proteine (z.B. Ladungsverteilung) signifikant durch lokale Konformationsänderungen beeinflusst werden können. Dieser Effekt kann durch posttranslationale Modifikationen, insbesondere durch *N*-Glykosylierungen und Palmitoylierungen, noch verstärkt werden. Daraus könnten für SP-G und SP-H amphiphile Eigenschaften resultieren, wie sie für die bereits bekannten Surfactantproteine beschrieben werden. So könnten beide Proteine in einer wässrigen Umgebung einen hydrophilen Charakter aufweisen, in der Nähe einer Lipidschicht oder bei Einwanderung in ein hydrophobes Milieu aber durch geringfügige Änderungen der Struktur auch deutlich hydrophobe Bereiche präsentieren. Dieser amphiphile Charakter ist ein weiterer Hinweis auf die Zugehörigkeit von SP-G und SP-H zur Familie der Surfactantproteine, welcher mithilfe der computergestützten Simulation erlangt werden konnte.

Abbreviations

ARDS	Acute Respiratory Distress Syndrome
BLAST	Basic Local Alignment Search Tool
CASP	Critical Assessment of Techniques for Protein Structure Prediction
CRD	carbohydrate recognition domain
DPPC	dipalmitoylphosphatidylcholine
GalNAc	*N*-acetylgalactosamine
GlcNAc	*N*-acetylglucosamine
MD	molecular dynamics
MS	mass spectrometry
NMR	nuclear magnetic resonance
PDB	Protein Data Bank
PMDB	Protein Model Data Bank
PME	Particle-Mesh-Ewald method
PS	pulmonary surfactant
PTM(s)	posttranslational modification(s)
SFTA2, SFTA3	surfactant associated proteins 2 and 3, alternative denomination for SP-G and SP-H
SP(s)	surfactant protein(s)
SP-G, SP-H	surfactant protein G, surfactant protein H

Furthermore, the common three- and one-letter code for amino acids is used.

List of Figures

Figure 1: The pulmonary surfactant system. 2

Figure 2: Ramachandran plot produced by PROCHECK with the mapped distribution of the φ and ψ angles in native protein structures. 13

Figure 3: Schematic illustration of the simulation box layouts. 25

Figure 4: Overview of all 24 performed protein-lipid MD simulations. 26

Figure 5: Validation of the SP-G and SP-H protein model stability during a 20 ns MD simulation. 33

Figure 6: Structure presentation of the final protein model for SP-G. 34

Figure 7: Structure presentation of the final protein model for SP-H. 35

Figure 8: Protein model stability comparison for the SP-G and the SP-H model with PTMs and without PTMs. 38

Figure 9: Protein structure models of SP-G and SP-H with highlighted protein parts, which were suggested as antigens for antibody production. 40

Figure 10: Test of the anti-SP-G and anti-SP-H antibody by Western blot. 41

Figure 11: Plot of the area per lipid for each DPPC molecule in a bilayer patch with 128 lipids during a 25 MD simulation. 43

Figure 12: Influence of simulation temperature on the protein model stability. 44

Figure 13: Resulting structures of MD simulations of the SP-G and SP-H models in a lipid environment. 46

Figure 14: Detailed simulation results for the SP-G model without PTMs. 47

Figure 15: Protein-lipid interaction energy and backbone atoms RMSD plots for the SP-G model without and with PTMs. 48

Figure 16: Detailed simulation results for the SP-G model with PTMs. 49

Figure 17: Detailed simulation results for the pre-positioned SP-G model with PTMs. 50

Figure 18: Protein-lipid interaction energy and backbone atoms RMSD plots for the SP-G model with PTMs and most negative interaction energy and the pre-positioned SP-G model. 51

Figure 19: Detailed simulation results for the SP-H model without PTMs. 52

Figure 20: Protein-lipid interaction energy and backbone atoms RMSD plots for the SP-H model without and with PTMs. 53

Figure 21: Protein-lipid interaction energy and protein backbone RMSD for the SP-H model with PTMs and most negative interaction energy after 100 ns. 54

Figure 22: Detailed simulation results for the SP-H model with PTMs. 54

Figure 23: Detailed simulation results for the pre-positioned SP-H model with PTMs. 55

Figure 24: Protein-lipid interaction energy and backbone atoms RMSD plots for the SP-H model with PTMs and most negative interaction energy and the pre-positioned SP-H model..... 56

List of Tables

Table 1: Overview of posttranslational modification types that were considered in this work. 18

Table 2: Quality assessment results for the best scored homology model, threading model, and *ab initio* model for SP-G and SP-H. 32

Table 3: Predicted posttranslational modifications and their sequence positions in the SP-G and SP-H sequence..... 37

Table 4: Comparison of characteristics for a DPPC bilayer reported in the literature and values obtained from simulations in this work. 43

1. Introduction

As an essential for life, every breath supplies oxygen to the organism. The organ responsible for this process is the lung, which is therefore inevitably in direct contact with the air. Unfortunately, this exposes the lung surface to a large number of dangers, which could potentially damage the whole organ. Apart from the physical injury by inhaled particles, the evaporation of the surface and the underlying tissues as well as the infection by pathogens are the most threatening risks. To avert these threats, a thin liquid film lines the complete alveolar surface: the so-called pulmonary surfactant (PS).

1.1. The pulmonary surfactant system

Surfactant is an acronym for “surface active agent” and describes a complex mixture of approx. 90% lipids and approx. 10% proteins by weight [1]. It is part of the thin aqueous layer, which covers the air-liquid interface at the surface of the lung alveoli. The lipid component contains mostly phosphatidylcholines (between 70 and 80%) [2-4], from which the majority is dipalmitoylphosphatidylcholine (DPPC, 41-70%) [5,6]. The second most abundant lipids of the pulmonary surfactant (PS) are phosphatidylglycerols (7%), followed by phosphatidyl-ethanolamines that account for 5% of the total mass. Furthermore, phosphatidylinositol and sphingomyeline are present (2%) [5,6]. Neutral lipids, such as cholesterol, account for 5% of the lipid mixture. A summary of the average surfactant composition is depicted in Figure 1a. This lipid mixture forms a monolayer system with the polar lipid head groups facing the liquid interface and the hydrophobic carbonyl chains facing the air. Proteins of the PS, called surfactant proteins (SPs), are integrated into this monolayer or are lipid-associated in the aqueous phase. Figure 1b shows a simplified depiction of the current conception of the PS system setup, including the surfactant proteins, which are described in chapter 1.2.

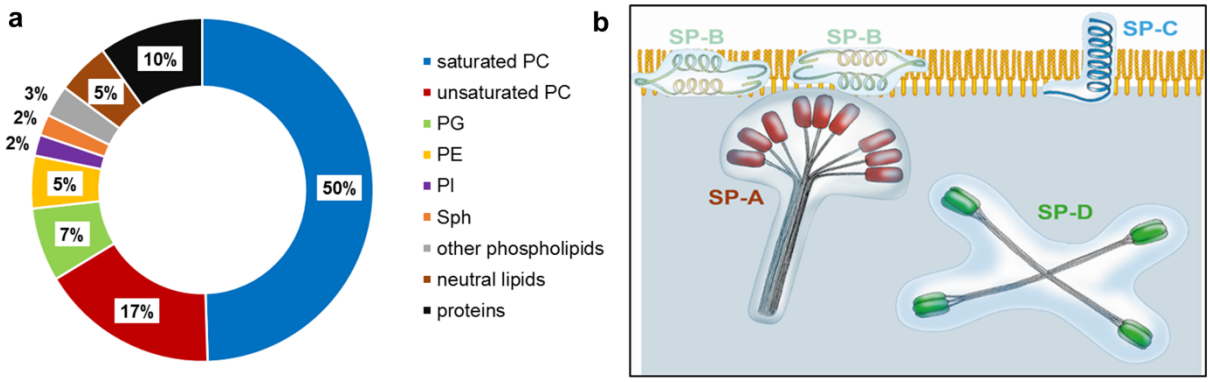


Figure 1: The pulmonary surfactant system. **(a)** Percentual composition of pulmonary surfactant from human bronchoalveolar lavage fluid (numbers from [1]). PC: phosphatidylcholine; PG: phosphatidylglycerol; PE: phosphatidylethanolamine; PI: phosphatidylinositol; Sph: sphingomyeline **(b)** Representation of the four currently known surfactant proteins SP-A, SP-B, SP-C and SP-D in vicinity of a pulmonary surfactant lipid monolayer (yellow) at the interphase between air (white area) and water (blue area).

The most important function of the PS is the lowering of the surface tension during the respiration process, which prevents the collapse of lung alveoli during expiration [7]. Hence, a fully functional PS system is essential for a proper lung function and surfactant dysfunction is associated with severe illnesses [8], for example the Adult Respiratory Distress Sydrome (ARDS) [9]. Furthermore, the PS provides an efficient protection against evaporation and shows mechanisms of host defense. Among the other components of the PS, the surfactant proteins are mainly responsible for the regulation of surface properties and immunological functions [10].

1.2. Surfactant proteins

Until now, four different surfactant proteins (SP-A, SP-B, SP-C and SP-D) have been identified, which can be divided into two classes. The surfactant proteins A and D are large hydrophilic proteins, which contain a carbohydrate recognition domain (CRD) and are part of the specific and non-specific immune defense mechanism of the pulmonary surfactant (PS) [11-14]. In contrast, surfactant proteins B and C are small and extremely hydrophobic proteins, whose functions are more related to general lipid organization and lipid layer stability [15,16]. To achieve their full functionality, SP-B and SP-C require a complex posttranslational modification pattern [17,18]. The interaction between the two surfactant protein (SP) classes seems to be necessary for a proper PS function. For example, the presence of SP-A showed a

supportive effect on SP-B activity [19] and the lack of SP-B results in a lower production of fully functional SP-C [20]. All four proteins were initially identified within lung tissue [11,12,21-23], but recently, SPs were also detected on the eye surface and in different tissues of the ocular system [24,25].

SP-A, encoded as a protein with 248 amino acids [26], is part of the C-type lectin family ("collectins"). Therefore, it shows a characteristic fold consisting of four regions [14]: The cysteine-containing *N*-terminus, which is important for oligomerization via intermolecular disulfide bridges, a collagen-like helical region, a short "neck region" with coiled-coil structure, and the *C*-terminus with CRD for Ca^{2+}-dependent binding of sugar moieties [27]. SP-A forms a characteristic bouquet-like 18-mer structure consisting of six homotrimer subunits (Figure 1b) [22]. *In vivo*, it is responsible for the formation of tubular myelin, an extracellular surfactant reservoir [28]. With that, SP-A is important for the spreading of lipids and the control of surface tension, especially in cooperation with SP-B [29]. Nevertheless, SP-A-deficient mice showed no alterations in PS stability [30], indicating that the other SPs can compensate an SP-A deficiency. Instead, the immunological functions of this protein are more important. As a part of the innate immune system, it stimulates the activity of macrophages [31], supports opsonization of microorganisms [5] and specifically binds to the surface of various pathogens by means of the CRD [32-34]. According to that, SP-A-deficient mice showed a reduced immune defense against pathogenic microorganisms [35].

SP-D, as a member of the C-type lectin family, contains the four regions of the general collectin fold as well. SP-D consists of 355 amino acids and its subunits show a high structural similarity to SP-A [12]. SP-D assembles as a dodecamer, consisting of four sets of triplet monomers, which are oriented in a cross-like complex (Figure 1b) [14]. With its CRD, it can specifically bind to carbohydrate moieties that are exposed for example on the surface of microorganisms. Therefore, SP-D is considered as part of the front line defense of the lung against inhaled pathogens. A direct interaction of SP-D with the *Influenza virus type A* [13], *Pseudomonas aeruginosa* and *Escherichia coli* [36,37] could be demonstrated. This emphasizes the importance of SP-D for the innate immune system, which is supported by experiments with SP-D-deficient mice that were more prone to infections with, for example, *Influenza virus type A* [38]. Furthermore, selective deletion of SP-D in mice [39] revealed its influence on the lipid homeostasis in the lung.

SP-B is a very small and extremely hydrophobic protein of the saposin superfamily. Posttranslational modification was shown to be mandatory to process the inactive precursor

protein with 381 amino acids into the fully functional protein [18]. The mature and active SP-B consists of only 79 amino acids, has a total charge of +7 and is organized in mainly α-helical structure [40]. Various cysteine residues stabilize the protein fold and allow the formation of oligomers of different sizes via intermolecular disulfide bridges [41]. SP-B is assumed to interact directly with a lipid monolayer, mediating lipid transfer and adsorption of single lipids into an existing layer. In this way, it influences actively the surface tension and stability of the PS during the respiration process [29,42]. Studies showed that a lack of SP-B is lethal for newborn mice [43] and causes fatal respiratory failure soon after birth in humans [44]. Recently, Yang *et. al.* demonstrated the role of SP-B in the activation of alveolar macrophages in the innate immune response in the lung [45]. All these facts emphasize the indispensability of SP-B for the regular breathing function. A short form of SP-B ("mini-B", residues 8-25 and 63-78) was shown to retain almost the complete activity of the full-length protein [46,47] and is therefore often used in experimental studies [48,49].

Despite the very short sequence length (35 amino acids), SP-C is one of the most hydrophobic proteins in nature known to date [50,51]. SP-C consists of an α-helix which may integrate into a lipid layer [52,53]. For this purpose, it possesses a high content of valine residues [53]. The hydrophobic character of SP-C is further increased by two palmitoyl moieties, which are attached to cysteine residues [54-58]. Additionally, other posttranslational modifications (PTMs), such as glycosylations, acylations or esterifications, were described for SP-C as well [17,59]. Similar to SP-B, SP-C is responsible for the stability of the PS, for the adsorption of lipids into an existing monolayer and for the reduction of the surface tension [11,21]. Different effects were demonstrated for SP-C-deficient mice, for example almost no change in PS stability compared to the wild type [60] in contrast to a higher susceptibility to inflammatory lung diseases [61]. This suggests a functional redundancy between SP-B and SP-C, where SP-B is the more effective protein [62], but SP-C showing additional immunological functions [63,64].

For the investigation of surfactant proteins (SPs), a comprehensive range of biochemical, biophysical and immunological methods were applied in many *in vitro* and *in vivo* experiments [30-39,46-49,60-64]. These studies led to new insights into characterization, localization, function and interaction of the different SPs with their environment. Despite these studies, there are still a lot of outstanding issues of interest in this field – not only because of the still unclear mode of action in detail. However, profound research on these proteins is very time consuming and requires a lot of experience, because the work with them is subjected to difficulties [65].

As for many proteins associated with a lipid system, the protein concentration *in vivo* is mostly very low, which prevents their direct purification from tissue. When overexpressed, some of these proteins tend to form aggregates, hence reducing the yield of stable and fully functional protein. Moreover, the recombinant expression in other host organisms often leads to posttranslational modification patterns that differ from the original organism or are missing completely. Depending on the protein, this could have a drastic effect on the protein activity in following experiments. Furthermore, especially the highly hydrophobic proteins SP-B and SP-C are difficult to handle in experiments due to their low solubility in aqueous media. All these aforementioned difficulties are also problematic for X-ray crystallography to obtain the overall protein structure. However, the knowledge of the 3D structure is a crucial step towards the understanding of the protein function. In fact, SP-C is the only surfactant protein with an X-ray structure of the full-length protein ("1spf" [66]) in the Protein Data Bank (PDB [67]). For SP-B, only very short fragments of the *N*-terminus ("1kmr" [68], "1dfw" [69], 15-25 of 79 amino acids), the *C*-terminus ("1rg3" [70], "1rg4" [70], 16 of 79 amino acids), and several versions of the truncated protein "mini-B" ("1ssz" [46], "2jou" [48], "2dwf" [48], 34 of 79 amino acids) are available in the PDB. The more hydrophilic character of SP-A and SP-D makes them less problematic to handle, but especially the *N*-terminus and the collagen-like region are still very difficult to resolve in X-ray experiments. Accordingly, only structures of the CRD-regions with "neck-domain" as single trimers are available in the PDB for SP-A ("1r13" [71], 148 of 248 amino acids) and SP-D ("1pw9" [72], 177 of 355 amino acids).

1.3. Computational modeling and simulation of surfactant proteins

The investigation of surfactant proteins (SPs) is an exemplary project, where the setup and realization of experiments is very complicated. In such cases, computational chemistry and protein modeling methods can effectively support experimental research. For example, modeling techniques can provide an atomistic three-dimensional model of a previously unknown protein structure. This model can give hints about the solubility of the protein or possible interactions with solutes in its environment, such as lipids, sugars or other proteins. Furthermore, a model could show which parts of the protein are exposed to the solvent. These solvent accessible residues will most likely possess posttranslational modifications, which may be essential for the protein function [73], as already described for the known surfactant proteins

[74-76]. Furthermore, a protein model can be used for molecular dynamics (MD) simulations. These calculations are able to show the time- and temperature-dependent behavior of a simulation system. This allows the observation of potential interactions of the protein with other compounds of its environment in a dynamic process.

Indeed, there are many examples in the literature for productive cooperation between theoretical and practical research: The assumption of SP-D being an immunological active protein could be supported by simulation studies, which showed the binding of different sugar moieties to the CRD region. Among these bound sugars were also glycans, which are presented on the surface of *Influenza virus type A* [77,78]. In more detail, simulations were able to show which amino acids are responsible for sugar binding and how the binding affinity is regulated [79,80]. For SP-B, various simulation studies were successfully performed, which showed the influence of the protein on systems consisting of different lipid species [81,82], determined the exact orientation of SP-B in proximity of a lipid layer [83,84] or observed which amino acids participate in the interaction with a lipid environment [85,86]. As a prerequisite for all these simulations, the possibility to reproduce a protein-free monolayer system consisting of PS lipids in a MD simulation was previously demonstrated by Javanainen *et. al.* [87]. By means of long time scale MD simulations, previously hypothesized SP-B functions, such as the support of lipid transfer and lipid reservoir building [88] or the mediation of lipid vesicle fusion [89], were confirmed as well. MD simulations of the SP-C structure in different media revealed the stability of the fold [90] and suggested SP-C to play an important role in the formation of bilayer reservoirs [91]. Finally, the cooperation of SP-B and SP-C observed in experimental studies was supported and visualized by MD simulations, which showed an increased fluidity of a membrane system and induced monolayer folding in presence of both proteins [92].

1.4. Motivation and objectives

With the decryption of whole genomes in the last years, a vast number of databases with information about putative gene sequences became available. This is also the case for the human genome. With the help of theoretical bioinformatics tools, these gene sequences were investigated and transformed into protein sequences with putative characteristics ascribed to them. Due to these studies, two new sequences for human proteins with putative surface regulatory activity were identified (UniProt [93] entries Q6UW10 and P0C7M3). According to

the order of their discovery, these two proteins were named surfactant-associated protein 2 (SFTA2) and surfactant-associated protein 3 (SFTA3) or alternatively, SP-G and SP-H [94,95]. The SP-G sequence comprises 78 amino acids with slightly hydrophobic character. It contains a predicted signal peptide of 19 amino acids at the *N*-terminus [96] that is essential for protein secretion [97]. In the UniProt entry, a potential *N*-linked glycosylation is suggested for position 37. Similar to SP-G, SP-H is a relatively short protein with 94 amino acids. However, the SP-H sequence shows an overall hydrophilic character. The amino acid sequences of SP-G and SP-H share only 23% identical residues. Their length of 78 and 94 amino acids is too short to show any similarity to the group of huge SPs (SP-A, SP-D). This suggests that they belong to the group of small surfactant proteins (SP-B, SP-C). However, they do not share any domains with SP-B or SP-C and the amino acid sequence identities are very low (about 10%). Unfortunately, no further information about characterization, localization, function or 3D structure was available for SP-G and SP-H at the beginning of this work. However, more information about these proteins might facilitate the understanding of the whole surfactant system. The localization of SP-G and SP-H on the lung surface or in associated tissue and the assignment of surface regulatory properties would verify their classification as surfactant proteins. Additional experimental studies and knowledge obtained about SP-G and SP-H could reveal new insights into the functionality of the pulmonary surfactant system. In this way, a detailed understanding of these proteins could point out completely new approaches for the treatment and therapy of surfactant dysfunction.

This work represents the theoretical part of an interdisciplinary project between the Leibniz Institute of Plant Biochemistry in Halle (PD Dr. W. Brandt) and the Institute of Anatomy II of the Friedrich-Alexander-University Erlangen-Nuremberg (Prof. L. Bräuer) to characterize the aforementioned proteins SP-G and SP-H, and to obtain first insights into the function of these novel and putative SPs. Therefore, the question if SP-G and SP-H are indeed surfactant proteins is the major issue of this work.

To address this question, the initial task is the generation of reliable 3D protein structure models for both proteins. The knowledge about the overall protein fold, the positions of potential posttranslational modifications (PTMs) and, consequently, hints about the surface reactivity (functional groups, hydrophobic spots) could be derived from these models. Based on these results, it should be possible to determine if SP-G and SP-H have any characteristics in common with the already known surfactant proteins. For example, two key features of surfactant proteins, the high grade of posttranslational modification (PTM) and the ability to interact with

lipid systems, should be deducible from the protein models. Furthermore, the models could be used to guide, support and interpret experimental studies, e.g. the generation of specific antibodies to enable the localization of both proteins in different tissues by immuno-histochemical methods. The localization of SP-G and SP-H in tissues that are typical for the presence of SP-A, SP-B, SP-C, and SP-D may further verify their classification as surfactant proteins.

However, the static representation of a protein model is not sufficient to investigate the interaction of SP-G and SP-H with a lipid system or to show if these proteins possess any surface regulatory activity, as it is typical for SPs. Therefore, the aim of this work is the application of computational simulation techniques on SP-G and SP-H in their natural environment. After establishing a lipid simulation system resembling the basic properties of the pulmonary surfactant, long-term MD simulations of the SP-G and SP-H models in this environment may indicate if these proteins are, in general, able to interact with lipids. Furthermore, these simulations could be able to show the protein-lipid interaction in detail (on an atomistic level) and might indicate the influence of the attached PTMs on the interaction interface and strength. These studies may also demonstrate if SP-G and SP-H are proteins that are associated on the surface of lipid systems, comparable to SP-A and SP-D, or if they are embedded into the lipid layers as known for SP-B and SP-C. Additionally, the influence of SP-G and SP-H on characteristics and stability of the lipid system could become apparent during these simulations. Altogether, the knowledge derived from MD simulations could help to classify these novel proteins with respect to the already known SPs and answer the initial question, if SP-G and SP-H show surface regulatory functionalities and thus are in fact members of the surfactant protein family.

2. Methods

2.1. Protein structure modeling

Knowing the exact three-dimensional structure of a protein is very important for the investigation of its characteristics and functionality. Therefore, nearly all 3D protein structures known today are stored in the Protein Data Bank (PDB) [67]. This repository is publicly available and provides coordinate files, literature references and various additional annotated information for each structure. Prior to release, every entry in the PDB is manually reviewed and assigned a four-letter code as unique identifier. The standard methods to obtain the 3D protein structure as deposited in the PDB are X-ray crystallography and nuclear magnetic resonance (NMR) spectroscopy. For crystallography, a beam of X-rays is directed at the protein in a crystalline state. The resulting diffraction pattern can be transformed into an electron density map, which is used to determine the atom positions within the crystal [98]. Due to a high flexibility (high degree of freedom) or poor solubility of the protein, it could be very difficult to find optimal crystallization conditions or even the formation of well-ordered crystals could be inhibited. For NMR spectroscopy, no protein crystals are needed and the protein is measured in a physiological (“natural”) solution, which may lead to more realistic protein structures [99]. However, the NMR technique is limited to small proteins and requires expensive equipment. Therefore, computer-assisted protein structure modeling tools were developed, which bypass the problems of experimental structure elucidation by constructing a model of the protein fold. In general, there are three methodologies currently available, which differ in their prerequisites, complexity and computational costs: homology modeling, protein threading and *ab initio* modeling [100]. However, the differentiation between these methods became blurred over the last decade and more and more protocols were presented that successfully combine elements of all approaches. The basic ideas of each method as well as the programs and tools used in this work will be presented in the following.

2.1.1. Homology modeling

With homology or comparative modeling, an atomistic structure model for a given amino acid sequence ("target") is generated based on at least one protein structure with high sequence similarity and already known 3D structure ("template"). The idea of this method is based on the observation that evolutionary related proteins with similar sequences often share a similar fold [101] and that local changes in the protein sequence (e.g. single mutations) do not necessarily influence the overall structure of a protein [102]. For a successful structure prediction, target and template should have a sequence identity (i.e. amount of identical amino acids) of at least 20% [103]. To identify possible template structures, a search with the BLAST algorithm ("Basic Local Alignment Search Tool") [104,105] is the common procedure. Thereby, the target sequence is compared to all sequences of proteins with known 3D structure in the PDB. The similarity between two sequences is calculated as a score based on identity and coverage after an alignment. Filter options allow to show only hits above a defined threshold as results.

In this work, the homology modeling protocol as implemented in YASARA [105-108] was used. It contains an automated template search, secondary structure prediction for the target sequence [109], and an alignment protocol to align target and template sequences. The final models are refined by energy minimizations and short MD simulations with the YASARA2 force field [110,111], which was developed especially to optimize protein structure geometries. Finally, an internal overall quality score ranks all resulting models. A special feature of the YASARA modeling protocol is the generation of a "hybrid model", which combines the best-scored parts of all obtained models.

2.1.2. Threading

When no template structures with a sequence identity above 20% are available in the PDB, the homology modeling will probably fail. The threading method expands the idea of homology modeling by classification (protein family), secondary structure prediction and fold recognition (domain identification) of the target protein. In general, homology modeling and threading are both template-based processes. However, whereas the homology modeling considers only

sequence similarity, the threading approach focusses on structural similarity for template identification and structural alignment. Threading routines are often provided by online servers for academic use. In this work, the “iterative threading assembly refinement” server, in short I-TASSER [112,113], was used for model generation. I-TASSER was ranked as best server for protein structure prediction in four consecutive “Critical Assessment of Techniques for Protein Structure Prediction” experiments (CASP7 [114], CASP8 [115], CASP9 [116], CASP10 [117]). The CASP experiments are organized as annual competitions, where all participants try to model the same predefined target as accurate as possible. Although the good performance of I-TASSER in these experiments, the server is still continuously improved. The target sequence can be submitted via web interface and the whole structure prediction and model building process is multi-phased and fully automated.

2.1.3. *Ab initio* modeling

If the requirements for homology modeling or threading cannot be fulfilled or the generation of a reliable model failed for other reasons, the *ab initio* or *de novo* protein modeling can be used to build a structure model. In the ideal vision of *ab initio* modeling, the protein structure is predicted “from scratch”, i.e. the prediction is solely based on physical and chemical principles of the amino acid sequence rather than already known structures or fragment libraries (knowledge-based information). The success of this method depends on the availability of an efficient method to explore all possible conformations of a peptide and a realistic energy function to obtain the energy landscape and to rank the individual conformations [118]. Since the conformational search is increasingly extensive for longer peptides, *ab initio* modeling is computationally very expensive. Therefore, the combination with knowledge-based information and high-performance computers is necessary to produce models in reasonable time, even for medium-sized proteins. ROBETTA [119] is the only folding server available for academic use that offers state-of-the-art *ab initio* modeling protocols (evaluated by CASP [120]) and the required computational power. After submission of the amino acid sequence using the online user interface [119], sequences are processed completely automated.

2.2. Protein model quality and validation tools

One of the most crucial questions in the process of protein modeling arises after the models were built with any of the previously described methods: How reliable is the obtained protein structure model? The numerous methods that were developed to answer this question use for example statistics about natively folded proteins (i.e. crystal structures), geometric properties of the amino acids (e.g. stereochemistry) or empirical energy functions. These different methods are able to show problematic protein regions, which deviate from a native-like state. Very often, several factors are calculated and combined to an overall quality score, which facilitates the comparison of multiple models. This information can be used to improve or correct the corresponding model (or parts of it), for example by refinement with energy minimization or molecular dynamics methods, using an alternative template for the modeling process or even switching to a more sophisticated modeling approach. In the following paragraphs, the protein structure validation and quality assessment tools used to evaluate the obtained protein models in this work will be introduced.

2.2.1. PROCHECK

The stereochemical quality of a protein model can be validated with PROCHECK [121]. The statistical analysis of known protein structures showed that native-like folds feature specific geometry patterns. The program calculates, for example, bond length and bond angles for the backbone atoms and checks the planarity of all peptide bonds or ring systems of amino acid side chains. The results are then compared with the statistics of native protein structures and are presented in various plots with (if present) highlighted problematic residues. The most important graph produced by PROCHECK is the “Ramachandran plot” (Figure 2). In this diagram, the two backbone dihedral angles φ and ψ of each amino acid residue are plotted against each other. Statistical analyses showed that not all possible combinations of φ and ψ occur evenly distributed, but that specific regions of the Ramachandran plot are preferred in native protein structures (Figure 2) [122]. Red or yellow regions represent typical and allowed torsion angle combinations. Light yellow regions show generously allowed angles, which are not very often found in known 3D structures, but which are still present in native proteins.

Ideally, over 90% of the protein amino acids should be located in the red (favored) regions. An amino acid with a φ-ψ-combination in one of the white areas is called "outlier" and the geometry of this amino acid itself but also of its environment should be checked carefully. Since a high stereochemical quality is an essential, but not sufficient prerequisite for a native-like protein model, other quality assessment tools were applied as well.

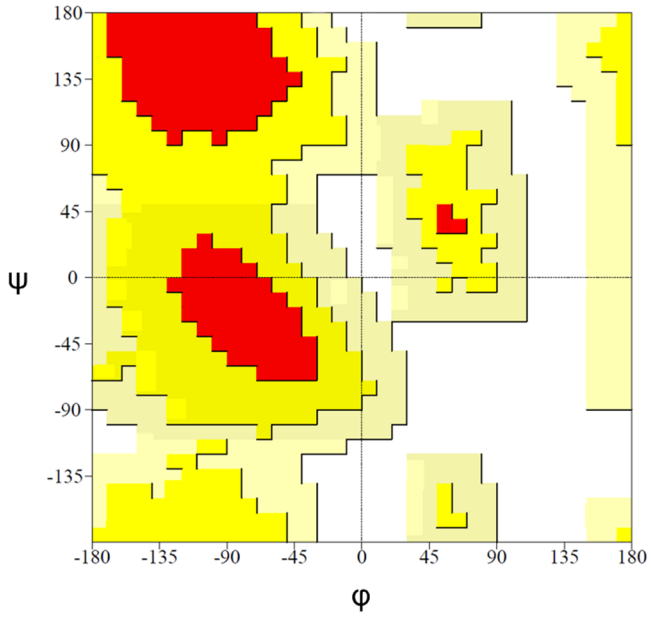


Figure 2: Ramachandran plot produced by PROCHECK with the mapped distribution of the φ and ψ angles in native protein structures. Red regions are "favored", yellow regions are "additionally allowed", and light yellow are "generously allowed" areas. All white regions represent "disallowed" angle combinations. For a protein with native fold, at least 90% of the amino acids should reside in the "favored" regions.

2.2.2. ProSA II

ProSA II [123] was the main criterion in this work to assess the protein model quality. With the help of knowledge-based energy potentials obtained from statistical analysis of known protein structures (X-ray and NMR from the PDB), the ProSA II program is able to estimate the fold quality of a protein structure model. The overall quality of the model is represented by a calculated "Z-score". This score is dependent on the protein length and the pair, surface, or a combined pair and surface potential. For the potential calculation, only the C_α atoms, only the C_β atoms or a combination of both can be used. For proteins with similar length and a native fold, the Z-scores are in a characteristic range, so that this score can give a hint if a protein model shows a native-like fold. If the calculated Z-score for a protein model is outside of this range, it very likely contains misfolded parts or erroneous regions. Additionally to the overall quality measure via Z-score, a local model quality is calculated with the energy potential as a

function of the sequence position. The result is presented in a plot with variable amino acid residue sliding window. In general, the plot should have a negative value for all positions in the amino acid sequence. Regions with positive energy values indicate problematic or non-native elements that should be checked and refined carefully.

2.2.3. ProQ

The Protein Quality Predictor [124] is a neuronal network based method to identify a correct model from a large subset of models with incorrect fold. To determine the model quality, two different scores are combined: *LGscore* [125] and *MaxSub* [126]. Both are sequence length dependent measures for the distance between a model and a correct target structure. Both can result in values between 0 and 1, but whereas the *LGscore* for two identical structures would be 0, the *MaxSub* would have a value of 1 and vice versa for two unrelated structures. In ProQ, the negative logarithm of the *LGscore* is used for computational efficiency. The reason for combining two different scores is the fact that all quality measures developed so far have different advantages and disadvantages (review in Cristobal *et. al.*, 2001 [125]). The most prominent problem is the influence of the protein sequence length on the accuracy of the method. In the case of *LGscore* and *MaxSub*, this dependence is contrary. While long proteins are more likely to achieve a good *LGscore*, short protein sequences are more likely to achieve a good *MaxSub* score. The idea of combining both measures in ProQ is to balance out the length dependency to obtain a more reliable protein quality measure. In practice, a correct model is defined by a combination of *LGscore* above 1.5 and *MaxSub* greater than 0.1, whereas an incorrect model should have an *LGscore* below 1.5 and a *MaxSub* lower than 0.1.

2.2.4. ERRAT

ERRAT [127] is an algorithm for protein structure verification, which concentrates on the statistical analysis of non-bonded pairwise atom interactions within a protein structure. The distribution of three different atom types (carbon, oxygen, and nitrogen) among the protein model structure is evaluated with a quadratic error function and is subsequently compared with results of 96 reliable protein structures. A bar plot of the error value is produced for the pairwise atom interactions of a nine-residue sliding window. Bars with a value above 95% indicate

residue windows with problematic atom type distribution. All regions with an error value above 99% or no successful error value calculation should be reviewed carefully. Additionally, an overall quality factor for the whole protein structure is calculated (between 0 and 100), which represents the percentage of protein residues with an error value below 95%. For natively folded proteins, this overall quality factor should be around a value of 95 or higher.

2.2.5. VERIFY-3D

VERIFY-3D [128,129] is able to generate a 3D profile for a given protein structure. Each residue of the structure is categorized into an "environment class" according to three criteria: the area of the side chain buried by other protein atoms, the percentage of this area that is buried by polar atoms or water, and the local secondary structure [128]. In this way, three-dimensional structure information is mapped to an one-dimensional information string that can be compared to an amino acid sequence. Therefore, VERIFY-3D can check if a given protein model (3D) is compatible with the corresponding amino acid sequence (1D). The compatibility is calculated as "3D-1D score", which is plotted versus the sequence number in a 21-residue sliding window. This allows an easy identification of regions with a problematic fold (i.e. incompatibility between structure and sequence). The score calculation for the first and last nine residues is not possible. For a good protein model, the 3D-1D score should be above 0.2 for at least 75% of all scored protein residues.

2.2.6. Stability test with molecular dynamics simulations

MD simulations are able to calculate the time-dependent behavior of a system and are therefore suitable to show dynamic process (see chapter 2.4.). Thus, MD simulations can give hints about the stability of a protein model. Extensive and permanent movements in protein regions, a loss of secondary structure elements or complete unfolding of the model during the simulation can indicate a poor model quality. Furthermore, the results of the previously described quality assessment tools for the model before and after the simulation can be compared. A significant degradation in those measures may suggest an unreliable protein model. To check the stability of the SP-G and SP-H protein models, MD simulations were performed with YASARA [105-108] and the YASARA2 force field [110,111]. Each protein model was placed separately

in a water box with a physiological NaCl concentration of 0.9% for a simulation time of 20 ns. The models of the final simulation snapshots were compared to the initial models. For more information about MD simulation analysis, please see chapter 2.4.3.

2.3. Prediction of posttranslational modifications

Many proteins of the proteome are chemically modified after or during their biosynthesis. About 400 different posttranslational modification (PTM) types are known today, so that the chemical space of the proteome is considerably expanded beyond the possibilities of the proteinogenic amino acids [130]. These covalently attached functional groups can significantly influence the stability and functionality of proteins. For many proteins, the full functionality is only reached after the addition of all PTMs. Even the control of complete protein activation and inactivation is possible due to the reversibility of most PTMs. In the following, a selection of PTMs considered in this work is briefly described.

One of the most important PTMs is the attachment of different carbohydrates ("glycans") to amino acid side chains. These "*glycosylations*" play an important role for protein targeting and transit. Furthermore, they are necessary for different signaling processes and can influence protein folding and activity [131]. Different types of glycans exist, whose complexity ranges from single monosaccharide moieties up to very huge structures with multiple sugar types, branches, and intermolecular bonds. The sugar moieties can be bound in different ways to the protein. The most prominent types are the linkage to an amine group of asparagine ("*N*-linked glycosylation") or to a hydroxyl group of serine, threonine or tyrosine ("*O*-linked glycosylation") [131]. "*Phosphorylation*" is the addition of a phosphate group, most often to the side chains of serine, threonine or tyrosine residues. Since it is a very flexible and reversible process mediated by protein kinases and phosphatases, phosphorylation is an essential mechanism to activate or deactivate enzymes or receptors, for example in signaling pathways [132]. Estimations indicate that 30% of all cellular proteins contain at least one phosphorylated residue, which emphasizes the importance of this PTM type [133,134]. "*Acetylation*" is the addition of an acetyl group to the *N*-terminus of the protein or, less frequently, to the ε-amino group of lysine [135]. The effects of *N*-terminal acetylation are not completely understood so far, but it may influence the protein stability, metabolism, and degradation [136,137]. The acetylation of lysine side chains, however, is a reversible process, which is associated with gene

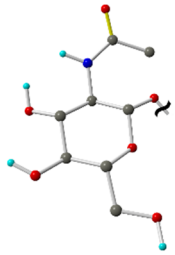
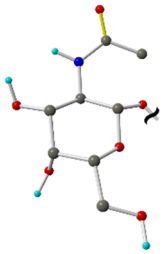
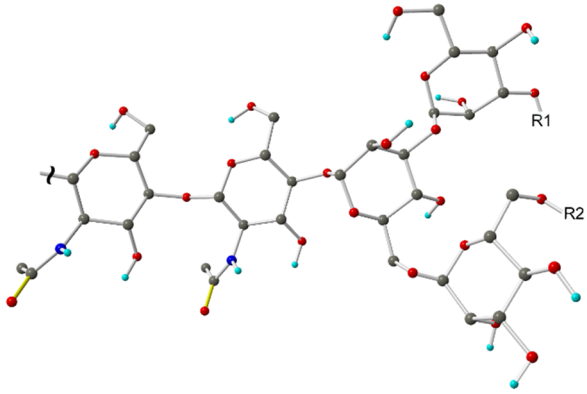
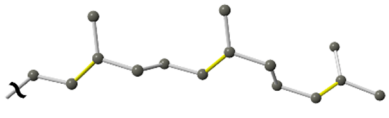
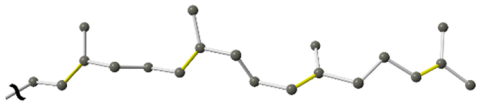
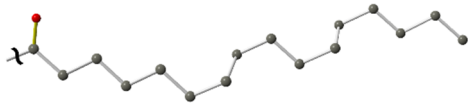
regulation and is often accompanied with other PTMs to modify the functions of the protein [138]. The addition of a sulfate group to the hydroxyl group of tyrosine is called "*sulfation*". Whereas it is known that sulfation is responsible for the strengthening of protein-protein interactions, its influence on protein function is still uncertain [139]. In the case of a "*prenylation*", a farnesyl or geranylgeranyl moiety is added to a cysteine residue by a thioester bond. Since the addition of these compounds may create a huge hydrophobic area on the protein surface, prenylation is typical for membrane-bound or membrane-integrated proteins [140]. Furthermore, prenylations may be important for specific protein-protein interactions [141]. The same holds for "*palmitoylations*", where palmitic acid as hydrophobic component is bound to a cysteine residue by a thioester bond. Analog to prenylations, this PTM is often part of membrane-associated proteins, because the fatty acid carbon chain enhances the hydrophobicity of the protein surface and is able to act as an anchor in a membrane compartment [142]. A comprehensive overview of effects and functions of palmitoylations (and prenylations) is presented by Dunphy *et al.* [143].

The method of choice to detect PTMs for a given protein is the (tandem) mass spectrometry (MS) analysis [144]. After the tryptic cleavage of the protein, mass shifts in the resulting peptide fragments indicate type and position of the PTMs. To enhance the chances for a successful PTM identification, pure and high-enriched protein samples are necessary. Therefore, the combination of these MS experiments with advanced chromatography and immuno-histochemistry methods is often essential, which requires to a very sophisticated preparation process.

As an alternative to complicated and expensive experimental studies, PTMs can be predicted based on existing knowledge. There are many tools available, which use data from sequence motifs or positions of known PTMs to recognize potential modification sites in proteins with unknown modification pattern. The majority of these tools performs a sequence-based prediction of a certain modification by means of a neuronal network, which was trained with a data set of experimentally investigated modification patterns. These sequence-based prediction tools for PTMs are available as online servers and are listed on the ExPASy bioinformatics resource portal [145]. The input is typically the raw amino acid sequence and the results are shown on html-webpages.

Table 1 gives an overview of all types of posttranslational modifications, which were considered in this study by prediction tools. Furthermore, amino acids that are a target for modification and example structures for attached functional groups are shown.

Table 1: Overview of posttranslational modification (PTM) types that were considered in this work.

PTM type	linkages	example structure
***O*-linked glycosylation**	Ser, Thr, Tyr	 *N*-acetylglucosamine  *N*-acetylgalactosamine
***N*-linked glycosylation**	Asn	 GlcNAc GlcNAc 3x mannose
phosphorylation	Ser, Thr, Tyr	~ PO_3^-
acetylation	*N*-terminus, Lys (ε-amino group)	~ $COCH_3$
sulfation	Tyr	~ SO_3^-
prenylation	Cys	 farnesyl group  geranylgeranyl group
palmitoylation	Cys	 palmitoyl group

Chemical structures are shown without aliphatic hydrogens. Single bonds are light grey and double bonds are yellow. The "~" symbol marks the bond that connects protein and PTM. Atom color code: carbon: grey; oxygen: red; nitrogen: blue; hydrogen: cyan.

In the following, all sequence-based prediction tools used in this work for the sequences of SP-G and SP-H will be described briefly:

NetPhos 2.0 [146]: With the help of a neuronal network, the probability for a phosphorylation of serine, threonine or tyrosine in a given eukaryotic sequence is predicted. This prediction is based on a large set of experimentally verified phosphorylation sites. The sensitivity of the method ranges between 69 and 96%, depending on the residue type.

NetOGlyc 3.1 [147]: For mammalian proteins, possible glycosylations of hydroxyl groups for serine or threonine residues with *N*-acetylgalactosamine (GalNAc) are predicted. The prediction is based on a neuronal network which is trained with the sequence itself and sequence derived features (surface accessibility, secondary structure, and distance constraints prediction). According to the developer's results, the method is able to predict 76% of the glycosylated and 93% of the not glycosylated residues within an unknown sequence.

YinOYang 1.2 [148]: The glycosylation of protein hydroxyl groups with an *N*-acetyl-glucosamine (GlcNAc) moiety is predicted based on an algorithm that is very similar to NetOGlyc (neuronal network). Since the modification sites for glycosylation and phosphorylation are overlapping (serine or threonine side chains), YinOYang can make use of the NetPhos server to identify and consider residues with positive prediction for both modifications.

NetAcet 1.0 [149]: This server predicts the *N*-terminal acetylation as performed by the *N*-acetyltransfase A (NatA) with a sensitivity up to 74% for mammalian data. The used neuronal network is trained with a data set derived from the yeast NatA, whose modification patterns were shown to be transferrable to mammalian NatA orthologs. The acetylation of internal lysine ε-amino groups or other acetyltransferases is not considered.

NetCGlyc 1.0 [150]: The NetCGlyc 1.0 server predicts the modification of the indole C2 atom of a tryptophane residue with a α-mannopyranosyl moiety via C-C coupling. Again, the prediction is performed by a neuronal network, which was trained with experimentally verified modification sites. About 93% of both positive and negative C-mannosylation sites are predicted correctly.

NetNGlyc 1.0 [148]: With this tool, the *N*-glycosylation of asparagine in human proteins is predicted. The prerequisite to identify a modification site is an Asn-Xaa-Ser/Thr motif. Based

on neuronal networks provided with known *N*-glycosylation data, the server reaches a cross-validated overall accuracy of 76%.

Sulfinator [151]: The sulfation of tyrosine residues in proteins is very hard to predict, because there are no clearly defined sequence motifs for this modification. Sulfinator combines four different Hidden Markov Models, which were trained with data of experimentally observed sulfations to predict possible modification sites in a protein sequence with an accuracy of 98%.

PrePS [152]: The "Prenylation Prediction Suite" is a web-application which combines the prediction for farnesylation or geranylgeranylation by proteins with CAAX-box motif [153]. Based on the already known substrates for these proteins, PrePS can predict if a given sequence might be a substrate as well. The results can be cross-checked with PRENbase [154], an annotated database with predicted and known prenylated proteins.

CSS-Palm 2.0 [155]: Based on a "Clustering and Scoring Strategy" (CSS) algorithm, the modification of a free cysteine sulfur atom with a palmitoyl group (saturated C16 fatty acid) is predicted. Since the prediction of such a modification site is very difficult due to the lack of unique sequence motifs, the performance of the predecessor of this program [156] was considerably improved by training the algorithm with a data set of 263 verified palmitoylation sites. In a cross-validation to a comprehensive experimental study [157], about 75% of the palmitoylations were predicted correctly by CSS-Palm 2.0.

The predicted posttranslational modifications (PTMs) were manually added to the final protein structure models of SP-G and SP-H, followed by an energy minimization in YASARA [105-108] with the YASARA2 force field. [110,111]. The stability of the added PTMs and their influence on the protein model structure was checked by MD simulations in YASARA (20 ns, water box with 0.9% NaCl, YASARA2 force field [110,111]). The results were compared to the protein model simulations without PTMs.

2.4. Molecular dynamics simulations

A protein structure model, even if it has an outstanding quality, represents only a static picture of a natural scenario. However, dynamic processes are very important for the protein conformation and the progress of chemical reactions in nature. Molecular dynamics (MD) simulations as a computational method can be used to consider these natural dynamics and show the behavior of a protein model over the course of a defined time period. This typically comprises several hundreds of picoseconds (ps) up to the microseconds (µs) scale, depending on the system size and available computational power. Since proteins usually reside in an aqueous environment, protein models are typically simulated in a box filled with water instead of vacuum. If there is already information about the protein environment available, the simulation system can be adapted to this knowledge. For example, a physiological salt concentration can be added to the solvent fraction or in the case of a transmembrane protein, the model can be integrated into a lipid system.

For a MD simulation, the movement of each atom of the system is calculated by solving Newton's equations of motion temperature-dependent and in defined time intervals ("time step"). Therefore, the force for every atom is calculated as the negative derivation of potential energy functions, which are provided for all elements of the system by force fields. The parameter sets of force fields can be derived empirically (based on experimental data) or by accurate *ab initio* calculations [158]. In general, there is no "optimal" force field for all purposes. However, many force fields were parameterized for a special scope of application. For example, the MMFF94 force field is only suitable for small organic molecules [159]. In contrast to that, the GROMOS [160] or AMBER [161] force fields were especially parameterized to accurately simulate protein structures and nucleotides (subsequently extended for other organic molecules). The choice of a suitable force field for the own research project is up to the user and may have a significant influence on the simulation results [162].

For the GROMOS force field [160] which is used in this work, the potential energy functions are represented as the sum of three different terms: *bonded* interactions, *non-bonded* interactions, and *restraints*. As the name suggests, the *bonded* term comprises the interaction energy of covalent atom bonds. Thus, the parameters of bond length as well as bond angles, dihedral angles ("proper"), and in-plane torsion angles ("improper") are available in this force

field as harmonic potentials for all possible combinations of atom types. High frequency oscillations of bonded interactions are a common problem in MD simulations. In combination with an unfavorable or too high time step, these oscillations may induce the breakdown of the whole simulation system. In GROMACS [163,164], the implemented LINCS algorithm [165,166] can be used to constrain the bond length between atoms of defined types or all atom types. This stabilizes the simulation and allows a higher time step. The *non-bonded* energy contains a repulsive and a dispersion term for van-der-Waals interactions, which are present in the form of Lennard-Jones potentials with parameters from the force field [163,164]. Furthermore, a Coulomb term is responsible to take (electrostatic) interactions between atoms with partial charges into account. To determine which atoms are interacting, GROMACS uses so-called "neighbor lists". These lists contain all non-bonded atoms within a certain radius and are updated in regular intervals (pre-defined, not necessarily in every simulation step). Since non-bonded interactions show effects over long distances, i.e. the neighbor list radius has to be very large, their calculation is a very time-consuming task in a MD simulation. Therefore, van-der-Waals interactions are typically only considered up to a defined distance ("cutoff") or are progressively switched off in a defined distance interval [167]. For the calculation of Coulomb interactions, however, even high cutoff distances might result in simulation artefacts [168]. Thus, sophisticated calculation schemes, such as the Particle-Mesh-Ewald (PME) method [169,170] used in this study, are necessary to calculate electrostatic interactions with the desired accuracy in reasonable time. The *restraints* term allows the user-defined manipulation of the potential energy for different reasons. With this value, it is possible to lock distances or angles between atoms during a simulation. Furthermore, the position of atoms in the coordinate system can be fixed. This is often used in the first phase of a simulation, where extreme fluctuations of properties, e.g. temperature or pressure, may threaten to damage the system.

For a MD simulation, the term "ensemble" is defined as a set of environmental assumptions that produces statistical representative conformations for the simulation system under the given conditions. In the so-called NVT (canonical) ensemble, the number of atoms (N), the volume of the simulation box (V) and the temperature (T) are conserved. This requires a very careful setup of the simulation box, since the box boundaries and consequently the density of the simulation system is not allowed to change during the MD. To keep the system temperature constant, a "thermostat" is introduced, which couples the system to an external heat bath by introducing a scaling factor to the calculated energies. Depending on the thermostat choice, the coupling algorithm is either very effective in adjusting a system to a target temperature, which

is important for system equilibration phases, or very efficient in terms of resembling a canonical ensemble [171]. In GROMACS, it is possible to couple different groups of atoms to different heat baths, e.g. protein and solvent atoms separately. Consequently, artefacts introduced due to the imperfect energy exchange between different system components may be avoided [172]. In the case of a so-called NPT (isothermal-isobaric) ensemble, number of atoms (N), pressure (P) and temperature (T) are constant. In addition to a thermostat, the scaling of the box dimensions is now allowed in order to reach a target value for the system pressure. The methodology is very similar to the idea of a heat bath and the algorithm responsible for this "pressure bath" is called "barostat". The box scaling can be achieved in different combinations. With isotropic scaling, size changes are applied equally in x, y and z dimension of the simulation box. Whereas this setup may be suitable for simple systems, it causes problems for simulations with a membrane, for example. In this case, the changes in x and y dimension of the membrane plane (membrane surface area) are not necessarily identical to the scaling needed in z dimension. This could lead to a deformed membrane, misleading results and artefacts. Alternatively, semi-isotropic scaling can be used, where only the x and y directions are associated and the z dimension is able to change independently. Furthermore, anisotropic coupling is possible, where six dimensions (including diagonal compressibility) can change independently. However, this complicated variant may lead to extremely deformed simulation boxes.

Independently from the choice of the ensemble applied, the box layout can have a significant influence on the simulation results. Whereas a rectangular box is the most suitable shape for simulations with lipid systems in the x/y plane, a rhombic dodecahedron or truncated octahedron may be more appropriate for proteins [173]. These layouts are closer to the shape of most proteins than a rectangular box and thus require less solvent molecules to fill it up. Nevertheless, the box dimensions should be large enough so that the solute can exhibit a reasonable far distance to the box boundaries, since a too small distance may introduce artefacts. This is very problematic for the membrane simulations planned in this work, because the lipid layers will be in direct contact with the boundaries of a rectangular box (cf. Figure 3). However, the usage of periodic boundary conditions [174] can completely avoid disturbing boundary effects. With this method, multiple translated copies of itself surround the simulation box, so that the atoms of the "original" box can "feel" the atoms of the adjacent copy, i.e. the atoms of the opposite box side, during the simulation [174]. Thus, there are no boundaries in the resulting "infinite" simulation system. Periodic boundary conditions in all three dimensions are necessary to use the Particle-Mesh-Ewald (PME) method for the calculation of electrostatic interactions, since it was developed for periodic systems [169,170].

All simulations with lipid or protein-lipid systems in this work were carried out with the GROMACS package version 4.5.4 [163,164]. The main reasons to use GROMACS were its free availability, the high performance in multi-threaded simulations and the united-atom G53a6 force field [160]. In contrast to an all-atom force field, the united-atom approach integrates the parameters for aliphatic hydrogens into the values of the carbon atom to which they are bound to. Therefore, aliphatic hydrogens can be omitted during the simulation, which reduces the number of atoms in the system and thus speeds up the calculation. Especially for the lipid systems used in this study, the speed up is significant due to omitting all hydrogen atoms of the lipid hydrocarbon chains. The simulation parameter files (.mdp) for all performed simulations are presented in Appendix 1-Appendix 5.

2.4.1. DPPC simulation system setup

The system, which is required for investigating possible interactions between the protein model and a lipid environment, should be as close as possible to the native state. For this reason, a basic dipalmitoylphosphatidylcholine (DPPC) lipid layer was established to simulate the SP-G and SP-H models in a natural environment. DPPC is the most abundant lipid in the pulmonary surfactant [2,3]. Accordingly, the literature describes that a lipid layer consisting solely of DPPC lipids is suitable to reproduce the basic properties of the lung surfactant in MD simulations [92,175-177]. The standard DPPC parameter set of the G53a6 force field was slightly modified in consideration of the results of Kukol (2009) [178] to produce a reliable lipid system. The initial bilayer consisted of 128 DPPC molecules per layer (256 lipids in total) and was generated with the CELLmicrocosmos MembraneEditor 2.2 [179]. The bilayer was placed in the center of a simulation box and solvated with water (Figure 3a). During the simulation, the water molecules are represented by the SPC water model [180]. A simulation of 75 ns indicated that the chosen lipid parameters and simulation settings are able to reproduce a stable lipid bilayer system. The MD simulation was performed with the Nosé-Hoover thermostat [181,182] at 323 K and the Parrinello-Rahman barostat [183,184] with semi-isotropic coupling and a reference pressure of 1 bar. The LINCS constraint algorithm [165,166] was used to fix the stretching of all bonds, allowing a time step of 4 fs. Electrostatic interactions were calculated with the PME algorithm [169,170] as implemented in GROMACS with a cutoff of 1.2 nm. The van-der-Waals potential was switched off between 1.2 and 1.3 nm. The neighbor

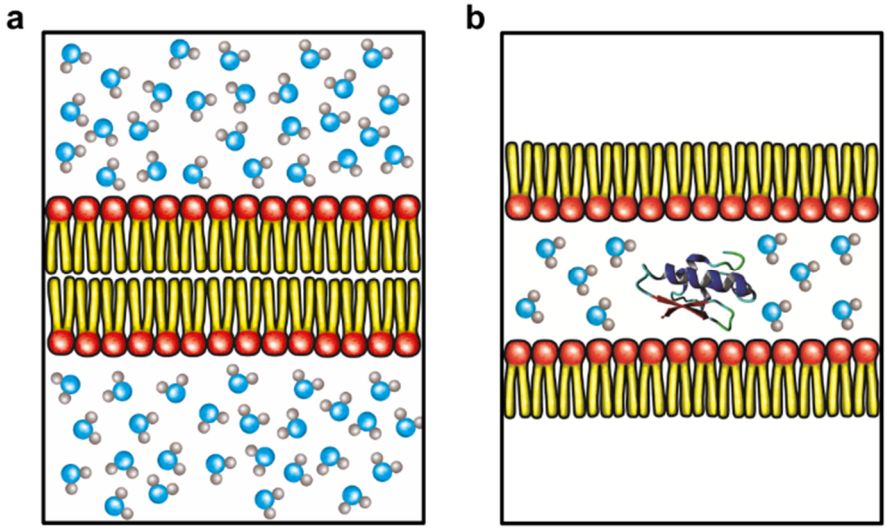


Figure 3: Schematic illustration of the simulation box layouts. The polar lipid head groups are red and the lipid tails are yellow. Water is depicted as blue oxygen and grey hydrogen atoms. The secondary structure of the protein is shown in ribbon representation. **(a)** Simulation of a lipid bilayer in water. **(b)** Simulation of a lipid monolayer system with protein. The polar head groups of two distinct monolayers face each other, creating a polar phase for the protein and water molecules. On the lipid tail side, the monolayers are separated by a broad vacuum phase.

list was updated every five steps, energy and pressure dispersion correction was applied. The last 25 ns of the simulation were used to calculate the area and volume per lipid, the lateral diffusion coefficient, and the area compressibility. In order to estimate the simulation quality, these values were compared to literature data (area and volume per lipid [185], lateral diffusion coefficient [186], and area compressibility [185,187]). The last snapshot of this 75 ns MD simulation was used to build the DPPC monolayer system. The membrane layer with the lipids 1-128 was rotated by 180 degrees so that the polar lipid head groups were facing each other. Afterwards, the layers were separated from each other generating space between the lipid head groups. Two systems were built, one with the lipid layers approx. 6.5 nm apart, hereafter referred to as “small system”, and one with approx. 9.5 nm space between the DPPC layers, hereafter referred to as “big system”. Both systems were placed in a simulation box with the lipid layers parallel to the x/y-plane. The z dimension of the box was set big enough to generate a 4-5 nm vacuum phase between the hydrophobic lipid tails due to the applied periodic boundary conditions. The space between the lipid head groups was filled with SPC water molecules [180]. A 25 ns MD simulation was performed to equilibrate the monolayer systems and check their stability. The compressibility of the systems in z direction was set to zero to preserve the vacuum layer between the lipid tails. Apart from that, the other simulation settings were identical to the bilayer calculations. The resulting monolayer systems were used to build the initial protein-lipid simulation layouts by placing the prepared and equilibrated (cf. 2.4.2) protein models in the water phase between the lipid head groups (Figure 3b).

2.4.2. SP-G and SP-H simulation in a lipid environment

All four protein models (SP-G and SP-H, each without and with PTMs) were equilibrated by a 20 ns MD simulation in a water box with the G53a6 force field [160] at 323 K. For this purpose, the force field was further modified with parameters for the attached PTM residues, namely phosphorylated serine, threonine and tyrosine, palmitoylated cysteine, and serine or threonine residues that are *O*-glycosylated with GlcNAc or GalNAc as well as *N*-glycosylated asparagine. The *N*-glycosylation residue consists of a pentasaccaride core with two GlcNAc and three mannose moieties (-GlcNAc-GlcNAc-mannose-$(mannose)_2$, cf. Table 1). The parameters for these residues were taken from original building blocks of the G53a6 force field, for example the glucose or mannose building block, and combined with standard amino acid building blocks to describe the modified residue. Missing values for the connection between those parts were complemented manually with parameter sets from the original force field. A derivation of novel force field parameters was not necessary. In the case of the phosphorylated amino acids, parameters were taken from the G43a1p force field [188]. The equilibrated protein models were placed in arbitrary orientations in the SPC water phase [180] between the DPPC monolayers.

Overall, six different starting orientations per model were generated and each system contained only one copy of the respective protein model (Figure 4). For each protein model, four different systems were built based on the “small system” and two based on the “big system”. As a special case, in one starting structure based on the “small system” for each modified protein, the model was manually positioned in a way that the palmitoylated cysteine residues are interacting with the lipid layer (“positioned”). For the SP-G model with PTMs, the palmitoyl moiety of Cys76 is in contact with the DPPC layer 1-128 at the simulation start. For the modified SP-H model, the palmitoylations of Cys45 and Cys56 are interacting with the DPPC layer 129-256 at simulation start.

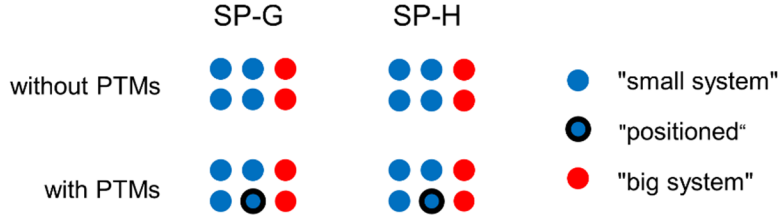


Figure 4: Overview of all 24 performed protein-lipid MD simulations. Every point represents a simulation. The different types of simulation systems are color-coded.

All 24 starting orientations (simulation systems, Figure 4) were neutralized with counter ions (Na^+/Cl^-) and submitted to a 250 ps equilibration run with NVT ensemble and the Berendsen thermostat [189] at 323 K, followed by a 250 ps equilibration run with NPT ensemble and the Berendsen thermostat at 323 K and barostat at 1 bar. Afterwards, a 50 ns production run was performed for all 24 orientations. The LINCS constraint algorithm [165,166] was applied on all bonds involving hydrogens and the simulation time step was set to 2 fs. The Nosé-Hoover thermostat [181,182] at 323 K and the Parrinello-Rahman barostat [183,184] with semi-isotropic coupling and a reference pressure of 1 bar were used for temperature and pressure coupling. Similar to the monolayer equilibration MD, the compressibility in z dimension was set to zero to maintain the simulation box layout. Electrostatic interactions were calculated with a cutoff at 1.2 nm with the Particle-Mesh-Ewald (PME) algorithm [169,170]. The van-der-Waals potential was switched off between 1.2 and 1.3 nm. The neighbor list was updated every 10 steps and no dispersion correction was applied. Trajectories of the system were saved every 10 ps.

2.4.3. Molecular dynamics simulation analysis

The evaluation of the MD simulation results and trajectories was done with tools included in GROMACS [163,164]. For an efficient analysis workflow, the tools were performed sequentially with a bash script (Appendix 6). The following values were obtained from all simulations:

Simulation box parameters: During the calculation, parameters regarding the simulation box are written to a log file ("energy file") in defined intervals. This comprises temperature, pressure and density of the system, as well as box dimensions and box volume. Furthermore, the energy of the system according to the force field parameters (in kJ/mol) is recorded as a single value (total energy) or divided in separate energy terms (bond, angle, torsion, Lennard-Jones, Coulomb energy etc.). Following the energy of individual system parts is also possible by defining them as "energy groups" prior to the MD simulation. All or only a selection of these values can be extracted from the energy file with the GROMACS tool "*g_energy*". The output is a tabulated file with the selected data series, which can be used to generate data plots and diagrams.

Root mean square deviation (RMSD): The RMSD of atomic positions is used to compare two (protein) structures and represents the most important value to observe the behavior of a protein during the MD simulation. Each simulation snapshot (frame) is superimposed with the starting structure ("fitted") and the spatial deviation is calculated for each (selected) protein atom. These values are averaged over all (selected) atoms in the snapshot, so that a single RMSD value (in nm) results for each simulation frame. Thus, plotting these values versus the simulation time gives an impression of the protein movement and stability during the MD simulation. If the RMSD plot is essentially stable over a longer simulation period until the end of the MD, the protein is referred to as stable or "equilibrated". This means that there are only minor movements in the protein structure without any expected significant changes. An instable graph with continuous fluctuations, however, may indicate problems in the protein model quality (regions with non-natural fold) or that the simulation conditions were not suitable for the examined protein (e.g. hydrophobic protein in polar solvent, too high system temperature or pressure). In this work, all protein atoms were used for the fitting process and the RMSD was calculated solely for the protein backbone atoms (*N*-C_α-*C*) using the GROMACS tool "*g_rms*".

Root mean square fluctuation (RMSF): The RMSF calculates the spatial distances (in nm) between the atoms of a simulation snapshot (frame) and a reference, which is in general the simulation starting structure. In contrast to the RMSD, the RMSF averages the distances for every atom over the whole simulation time, i.e. over all frames. This results in a single value for each atom, which indicates its range of motion during the MD simulation. Accordingly, the average value of all atoms in an amino acid residue represents the movement of each amino acid in the system. In this way, stable or instable regions of the protein can be identified. The RMSF values were calculated per residue with the GROMACS tool "*g_rmsf*" (with option -res).

Interaction energy: Prior to the simulations, two energy groups "PROTEIN" including all protein atoms and "DPPC" comprising all lipid atoms were defined in the simulation settings. As a result, the energy terms for these two groups are listed separately in the energy file (cf. "Simulation box parameters"). Since values for the energy terms between these two groups are recorded to the energy file as well, the interaction between protein and lipid layer can be monitored. It has to be noted that these calculated interaction energies are no definitive values. The force field-based calculation of non-bonded interactions still shows deficiencies in accuracy compared to experimental data [190-193]. However, plotting this protein-lipid

interaction energy versus simulation time can give an overview at which point of the simulation the interactions started and how stable these interactions are. Furthermore, it allows the comparison of the interaction strength between different simulations.

Area per lipid: The area per lipid is one of the most important structural parameters to describe a lipid layer system. The amount of space a single lipid is allowed to take up in an ordered layer structure is well defined and depends on the lipid layer composition. Thus, a steady area per lipid may indicate a stable lipid layer and that the simulation parameters are selected appropriately. Otherwise, sub-optimal simulation settings or changes in the layer structure due to the interaction of a protein with the lipid surface may result in an area per lipid change. The area per lipid is calculated as the product of the x and y box dimensions, where the membrane in parallel to the x/y-plane, divided by the number of lipids per layer. The box dimensions were extracted from the energy file (cf. "Simulation box parameters") and the area per lipid (in nm^2) was calculated for every simulation frame to track the stability of the lipid layers used in the performed MD simulations.

Secondary structure assignment: The DSSP algorithm [194,195] can assign a secondary structure element (α-helix, β-sheet, turn, coil etc.) to each amino acid of a given protein. The assignment process is based on the hydrogen-bonding pattern of the protein backbone. Thus, by performing DSSP for each MD snapshot, the stability of secondary structure elements over time is obtained. For a stable protein structure, the distribution of secondary structure elements should not change. For a protein with instable regions or a protein that starts to unfold during the simulation, more and more residues would be assigned as "coil". In this work, the GROMACS tool "*do_dssp*" was used to determine major changes in the protein secondary structure. This tool calculates the number of residues with assigned structure (sum of residues with assignment as α-helix, β-sheet, β-bridge or turn) and plots the results versus simulation time to assess the protein stability. Furthermore, the tool calculates the overall secondary structure element distribution (in %).

Visualization: Manual illustration of simulation snapshots is essential for the simulation analysis. Watching the system at simulation speed can reveal how different system components influence each other in their motion and show dynamic processes for example of protein loops or termini. Due to the atomic representation of the simulation system, observation of interactions between single atoms is possible, for example the formation or regression of hydrogen bonds. Eventually, the visualization can help to understand results and phenomena observed in the statistic-based MD analysis. For the visualization of simulation results as well

as the generation of figures, VMD [196] and YASARA [106,111] were used. Both programs are able to show the snapshots as interactive sequence (“movie clip”), can draw systems with different visualization styles, and can color, hide or label specific system components.

3. Results

3.1. Protein structure modeling

At the beginning, online BLAST [104] searches were performed for human SP-G and SP-H amino acid sequences to find possible homolog proteins. For SP-G, the full length protein sequence including the *N*-terminal 19 amino acid signal peptide was used, because it is unknown in which form the protein is present at its site of action. A BLAST search in the UniProt database [93] for sequences with a high identity compared to SP-G resulted solely in hits that are most likely uncharacterized or putative SP-G-homologs of other mammals. This was also the case for SP-H, however, a number of hits with a low score for short sections of putative regulatory proteins from different *Pseudomonas* species were detected as well. As expected, BLAST searches for both proteins to find similar sequences with already known 3D structure in the PDB were not successful. The identified hits had either a too low sequence identity (< 20%) or a poor coverage (only 18-25 of 78 residues for SP-G and 28-33 of 94 residues for SP-H). Since there were no reliable templates found and the sequence identity of SP-G and SP-H to the already known SPs is very low (ca. 10%), first attempts to obtain the 3D structure by homology modeling failed. These attempts were performed with the YASARA homology modeling routine, which automatically searched the PDB and identified “1em7” for SP-G and “1vj0” for SP-H as possible templates for the modeling process. However, the sequences identities to the target sequences were below 10% in both cases. Accordingly, the resulting models for SP-G and SP-H showed problematic ProSA II Z-scores and energy plots (see Table 2 and Appendix 7). While PROCHECK and ERRAT resulted in acceptable values, the scores of VERIFY-3D and ProQ showed serious deficiencies and could not be improved by MD refinements. Therefore, the homology models were discarded and the SP-G and SP-H sequences were sent to the online threading server I-TASSER. Despite the threading models for both proteins showed significantly better results for the combined Z-score and ProQ as well as in ProSA II plots, they revealed issues in the PROCHECK and ERRAT values. Especially the SP-G model showed a problematic Ramachandran plot with only 68.2% of all amino acids in the favored regions and five outliers (Table 2a). Even MD refinements, manual editing and

MDs to relax the system could not improve the quality of the SP-G and SP-H models. As a result, the threading models were considered insufficient for further studies.

Table 2: Quality assessment results for the best scored homology model, threading model, and *ab initio* model for **(a)** SP-G and **(b)** SP-H.

a

SP-G models	combined Z-score	PROCHECK		ERRAT-score	VERIFY-3D	ProQ	
		fav. regions	outlier			*LGscore*	*MaxSub*
homology modeling	-0.65	92.4%	0	90.7	21.5%	-0.358	-0.008
threading	-5.50	68.2%	5	80.9	54.4%	1.881	0.126
ab initio modeling	-6.16	95.5%	0	100.0	97.5%	3.579	0.141

b

SP-H models	combined Z-score	PROCHECK		ERRAT-score	VERIFY-3D	ProQ	
		fav. regions	outlier			*LGscore*	*MaxSub*
homology modeling	-3.90	91.7%	0	97.7	21.1%	0.448	0.062
threading	-5.10	85.7%	1	89.5	72.6%	1.736	0.030
ab initio modeling	-5.72	94.0%	0	93.0	48.4%	1.804	0.131

Overall, these validation values provide hints about the quality of generated models. The length-dependent average value of the combined Z-score is -7.77 for SP-G (**a**) and -8.0 for SP-H (**b**). More than 90% of all residues should reside in the most favored regions of the Ramachandran plot without any outliers. The ERRAT score should be 95 or higher and the VERIFY-3D percentage above 75% for a natively folded structure. For a good model, the *LGscore* should be above 1.5 and the *MaxSub* greater than 0.1.

Finally, the amino acid sequences of SP-G and SP-H were submitted to the *ab initio* modeling server ROBETTA. The obtained protein models showed a significant quality improvement in comparison to the threading models. Only minor local energy minimizations and a MD refinement with YASARA were needed to achieve acceptable results with structure validation tools. ProSA II produced a negative plot for the whole SP-G structure model (Appendix 7) and a combined Z-score of -6.16, which is close to the average value for proteins of this length (-7.77) (Table 2a). PROCHECK determined 95.5% of the 78 amino acids with a dihedral angle in the favored regions of the Ramachandran plot and no outliers. The ERRAT overall quality

factor reached the best possible value (100) and VERIFY-3D showed a very good result with 97.5% of the residues having a 3D-1D score above the threshold. ProQ calculated an *LGscore* of 3.579 and a *MaxSub* score of 0.141, which indicated a "very good" and "fairly good" model, respectively. Altogether, this suggested a reliable SP-G model structure.

For the model of SP-H, the ProSA II plot was also completely negative (Appendix 7) and the Z-score (-5.72) was in acceptable distance to the length-dependent average value (-8.0), indicating a native-like folding of the model (Table 2b). In addition, the Ramachandran plot showed 94% of the 94 amino acids in the favored regions without any outlier, which implied a very high stereochemical quality. The overall quality factor of ERRAT was 93. Furthermore, the ProQ *LGscore* of 1.804 and the *MaxSub* score of 0.131 indicated a "fairly good" model. The only drawback of this model for SP-H was the VERIFY-3D result, which was clearly below the optimal value (75%) with only 48.4%. This may be due to the methodology of this tool. Especially for small proteins, already a single polar amino acid (partially) buried by hydrophobic side chains can reduce the score drastically, although this would not necessarily indicate a problem with the overall fold of the protein model. However, since the other four tools did not suggest major quality problems, the obtained protein structure model for SP-H was considered reliable.

Both final protein models based on *ab initio* modeling were subjected to a 20 ns MD simulation in a water box with YASARA to determine the model stability. The analysis of the RMSD of the protein backbone atoms revealed that both protein models reached a stable conformation within a reasonable simulation time. For SP-G, the RMSD first showed a stable phase between 4 and 9 ns, before it rose to a plateau after about 10 ns (Figure 5, black plot). The RMSD was

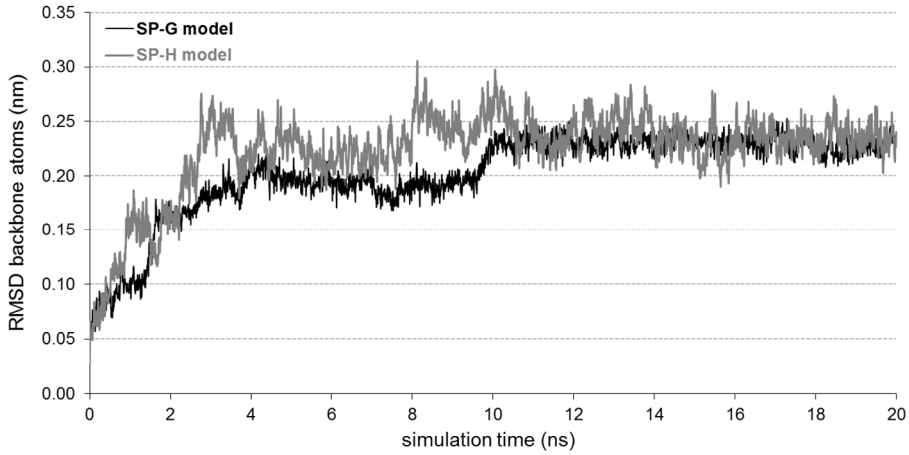


Figure 5: Validation of the SP-G (black) and SP-H (grey) protein model stability during a 20 ns MD simulation. The RMSD of the protein backbone atoms (in nm) was used as a measure for the model stability. Minor fluctuations of the RMSD indicate a stable protein model.

very stable on this level with only minor fluctuations until the end of the simulation. For SP-H, the RMSD plot reached a plateau already after 4 ns, where it remained stable until the end of the simulation (Figure 5, grey plot). For both protein models, no significant change in the secondary structure or unfolding of the protein was observed. Additionally, this stability was reflected by the secondary structure element percentages, which remained unchanged during the simulation (47% helix, 19% sheet and 34% coil for SP-G and 50% helix, 8% sheet and 42% coil for SP-H). The results of the validation programs thereby were comparable to the pre-MD results, some ratings are even improved, for example the ProQ results for SP-G or the Z-score and ERRAT score for SP-H (Appendix 8). Therefore, the stable models were deposited at the Protein Model Data Base PMDB [197] for public download and received the PMDB id PM0078341 for SP-G and PM0079092 for SP-H. With this, the three dimensional models could give first insights into the structures of SP-G and SP-H.

The 3D model for SP-G (Figure 6a) is dominated by an α-helix (amino acids 41-56) and an antiparallel β-sheet structure spanning the residues 63-68 and 72-78, respectively. The hydrophobic part of the *N*-terminal signal peptide is modeled as a short α-helix (8-13). This helix as well as the other residues of the signal peptide (1-19) are loosely attached to the surface

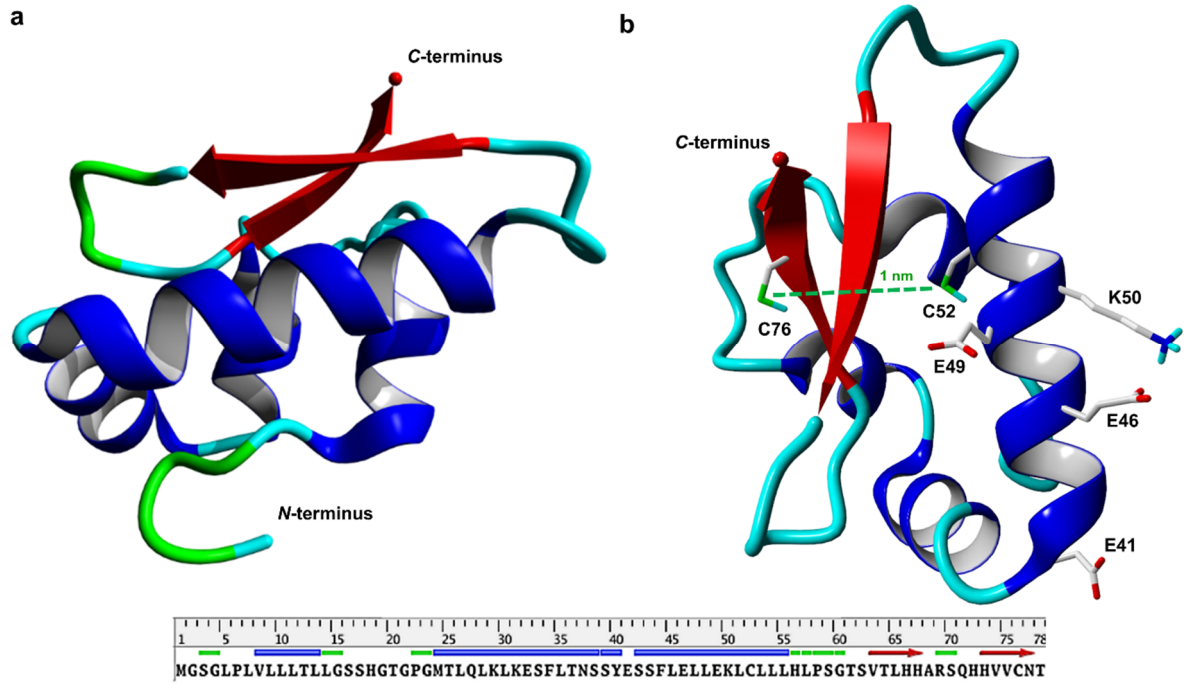


Figure 6: Structure presentation of the final protein model for SP-G. **(a)** Only the protein backbone is shown in ribbon presentation. **(b)** SP-G model with highlighted cysteine residues and selected charged amino acids. The view was rotated in comparison to **(a)** for a better understanding. Amino acid side chains are shown in stick representation without aliphatic hydrogens (carbon: grey; oxygen: red; nitrogen: blue; sulfur: green; hydrogen: cyan). In both pictures, α-helices are blue, β-sheets are red, turns and random elements are green and cyan. The same color code is used on the sequence bar on the bottom of the figure, which shows the secondary structure elements.

and cover the hydrophobic core of the protein. The fixation of this *N*-terminus on the protein is not very strong, so that a high flexibility is possible, which is needed for a signal peptide to interact with or be embedded into a lipid system due to its hydrophobic character. The α-helix 41-56 also contains many hydrophobic residues (seven leucine residues and one phenylalanine). In addition, it contains three glutamate residues and one lysine, which could possibly interact with the polar head groups of lipid molecules (Figure 6b). Furthermore, the structure model shows that the only two cysteine residues of the sequence are about 1 nm apart from each other, which is too distant for the formation of a stabilizing intramolecular disulfide bridge. However, Cys76 is located on the surface of the protein and could be able to form an intermolecular disulfide bond to another SP-G monomer. This would result in a covalently connected protein dimer. Although there is no surface region predestined for interactions with another monomer, a non-covalent oligomerization of SP-G cannot be excluded based on the protein structure model. There are no structure similarities to the already known SPs observable.

The most prominent structural features of the SP-H model (Figure 7a) are a long and stable α-helix of the amino acids 7-31 and an antiparallel β-sheet spanning the amino acids 55-62 and 67-73, respectively. The α-helix shows a high content of polar or even charged amino acids

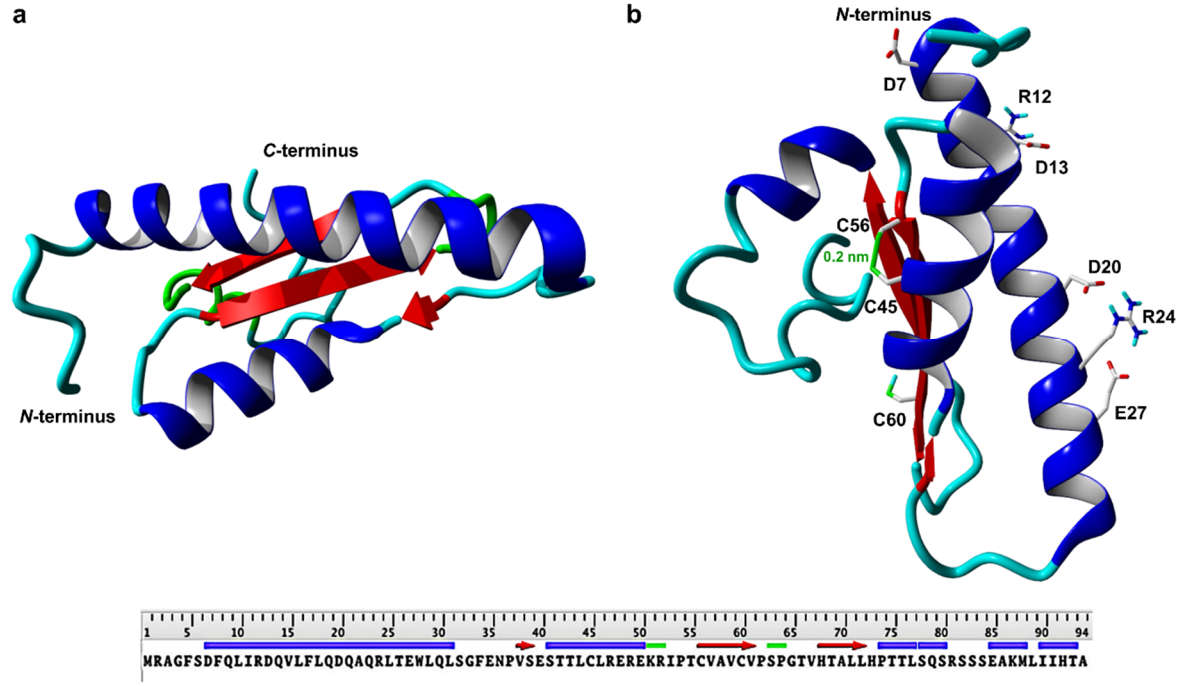


Figure 7: Structure presentation of the final protein model for SP-H. **(a)** Only the protein backbone is shown in ribbon presentation. **(b)** SP-H model with highlighted cysteine residues and selected charged amino acids. The view was rotated in comparison to **(a)** for a better understanding. Amino acid side chains are shown in stick representation without aliphatic hydrogens (carbon: grey; oxygen: red; nitrogen: blue; sulfur: green; hydrogen: cyan). In both pictures, α-helices are blue, β-sheets are red, turns and random elements are green and cyan. The same color code is used on the sequence bar on the bottom of the figure, which shows the secondary structure elements.

(four glutamine residues, three asparagine residues, two arginine residues, and one glutamate), which are present on the protein surface and could interact with polar lipid head groups (Figure 7b). On the remaining protein surface, no extensive hydrophobic domains or regions are observable, which could interact with a hydrophobic membrane fraction. However, there are single hydrophobic spots on the protein surface formed only by single amino acids or short sequence parts (e.g. Phe5, Trp28, Leu31, Phe34 or positions 88-91). Furthermore, indications for transmembrane regions in the protein were not found. The first five *N*-terminal residues are not able to form strong interactions with the nearby residues and are thus very flexible. The cysteine residues on position 45 and 56 could form a structure stabilizing intramolecular disulfide bridge (Figure 7b). Since all three available cysteine residues are accessible on the protein surface, intermolecular disulfide bonds, for example with other SP-H monomers, could be possible. Hence, an oligomerization of SP-H cannot be excluded based on the protein structure model. Similar to the SP-G model, the SP-H model shows no structural similarities to the already known SPs.

3.2. Posttranslational modifications

Since it is known that posttranslational modifications are very important for the function of the already known surfactant proteins, the SP-G and SP-H sequences were also analyzed for PTMs with various statistic-based online prediction tools.

For SP-G, NetNGlyc predicted an *N*-glycosylation on Asn37, which is already noted in the UniProt entry of SP-G. NetOGlyc predicted an *O*-glycosylation with *N*-acteylgalactosamine (GalNAc) on the *C*-terminal residue Thr78. Overall, YinOYang predicted five *O*-glycosylations with *N*-acetylglucosamine (GlcNAc) as sugar moiety. Thereby, the probability for a GlcNAc modification was moderate for Ser38, Ser39, Ser62, and Ser70 and high for a modification at Thr78. Given the results of NetPhos, the amino acids Ser17, Ser38, Ser39, and Tyr40 are most likely phosphorylated. Finally, the CSS-Palm server showed a possibly palmitoylated Cys76. The servers NetAcet, NetCGlyc, Sulfinator, and PrePS did not predict any modification site for the SP-G sequence. At this point, it is noticeable that only one PTM was predicted for amino acids of the *N*-terminal signal peptide. However, the phosphorylation of Ser17 is already very close to the signal peptide cleavage site.

Scanning the SP-H sequence for possible PTM sites gave the following results: The NetOGlyc server suggested six threonine residues at the positions 55, 66, 69, 75, 76, and 93 to be modified with a GalNAc moiety. The YinOYang prediction indicated a GlcNAc modification on Ser39, Thr76, and Ser78 with a high and on Ser82, Ser83, and Ser93 with a low probability. NetPhos predicted seven phosphorylation sites, namely Ser32, Ser39, Thr55, Ser80, Ser82, Ser83, and Ser84, with the last four having a high probability. The CSS-Palm server showed that two of the three available cysteine residues (45 and 56) might be palmitoylated. NetNGlyc showed no potential *N*-glycosylation and NetAcet did not predict any acetylation. Finally, NetCGlyc as well as Sulfinator and PrePS showed no potential modification sites.

Subsequently, the predicted PTMs were added manually to the final protein models of SP-G and SP-H. For this process, the following two conventions were applied: First, if there was more than one modification predicted for the same position, only the modification with the highest probability was considered. Second, only solvent accessible amino acids, i.e. side chains on the protein surface, were modified, since the addition of e.g. a bulky glycosyl moiety would have caused steric problems and noticeably changes of the protein structure. All PTMs that fulfilled these requirements are summarized in Table 3 and were actually added to the protein models. Two phosphorylations, three *O*-glycosylations with GlcNAc, one palmitoylation and one *N*-glycosylation were added to the SP-G model (Table 3). For the SP-H sequence, six phosphorylation sites, six *O*-glycosylations (two with GlcNAc and four with GalNAc) as well as two palmitoylated residues were predicted and attached to the protein model (Table 3).

Table 3: Predicted posttranslational modifications and their sequence positions in the SP-G and SP-H sequence.

sequence position SP-G	Ser17	Asn37	Tyr40	Ser62	Ser70	Cys76	Thr78
modification	PHOS	*N*-GLC	PHOS	*O*-GLC	*O*-GLC	PALM	*O*-GLC

sequence position SP-H	Ser32	Ser39	Cys45	Thr55	Cys56	Thr66	Thr69
modification	PHOS	*O*-GLC	PALM	PHOS	PALM	*O*-GAL	*O*-GAL
sequence position SP-H	Thr75	Ser78	Ser80	Ser82	Ser83	Ser84	Thr93
modification	*O*-GAL	*O*-GLC	PHOS	PHOS	PHOS	PHOS	*O*-GAL

Present modification types: phosphorylation (PHOS), palmitoylation (PALM), *O*-glycosylation with GlcNAc (*O*-GLC) or GalNAc (*O*-GAL), *N*-glycosylation with a pentasaccaride core consisting of two GlcNAc and three mannose moieties (*N*-GLC).

After the manual addition of the PTMs, the protein models were submitted to a 20 ns MD simulation in YASARA to relax the attached modifications and check their influence on the protein model stability in comparison to the unmodified models. The RMSD plot for the modified SP-G model (Figure 8a, grey plot) showed that the structure is very robust in this simulation system, reaching an equilibrium phase after 8 ns with only small RMSD fluctuations thereafter. As for the unmodified protein model (Figure 8a, black plot), no significant secondary structure changes or hints for an unfolding of the protein structure were observed.

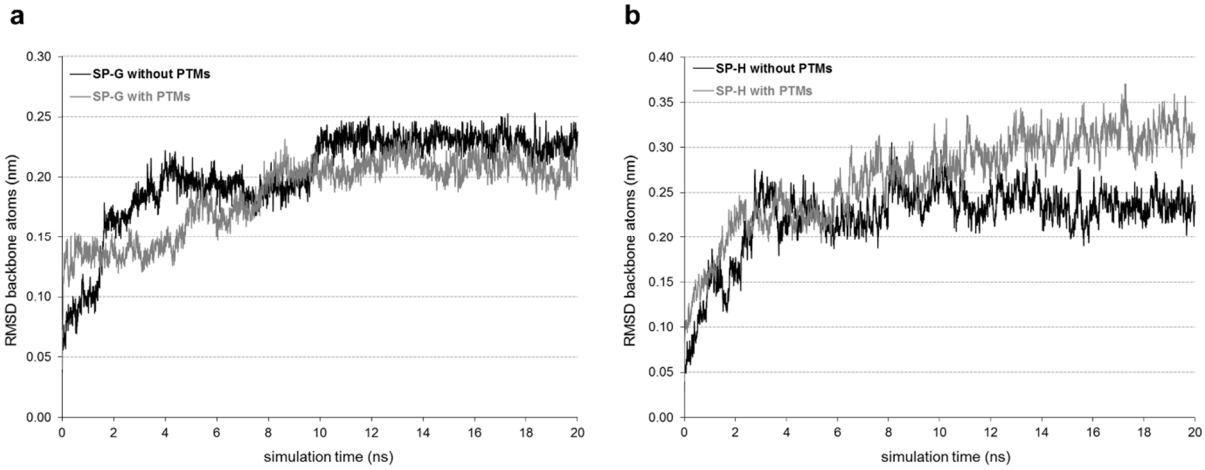


Figure 8: Protein model stability comparison for **(a)** the SP-G and **(b)** the SP-H model with PTMs (grey) and without PTMs (black). Plots of the protein backbone atoms RMSD (in nm) are used as a measure for the protein model stability during a 20 ns MD simulation.

The RMSD plot for SP-H also showed a stable protein model with additional PTMs (Figure 8b, grey plot). Until a simulation time of 11 ns, the RMSD values were almost identical to the unmodified model (Figure 8b, black plot). Thereafter, the plot for the modified model showed a higher fluctuation due to the influence of the large and numerous PTMs attached to the protein. Nevertheless, no significant secondary structure changes or an unfolding of the protein was visible.

Overall, two model variants for each protein (with and without PTMs) were obtained, which maintain their good model quality during MD simulations (Appendix 8). Therefore, all four models are suitable for sophisticated computational chemistry studies in a lipid environment. The protein models with attached PTMs resulting from the MD simulations were deposited at the PMDB [197] and received the PMDB id PM0078342 for SP-G and PM0079093 for SP-H.

In summary, the number of predictions determined in this work suggest that SP-G and SP-H show a high grade of posttranslational modification, comparable to the already known surfactant proteins. The possibility of protein-lipid interactions is significantly increased by the

numerous added sugar and phosphate moieties, which could interact with polar lipid head groups. Furthermore, the attached palmitoylations could play an important role for the interactions of the proteins with the hydrophobic lipid compartment. Thus, the obtained findings justify the classification of SP-G and SP-H as surfactant proteins.

3.3. Generation of specific antibodies

To perform immunohistochemical staining methods, e.g. protein localization studies, a specific antibody for the protein of interest is needed. For the production of such an antibody, the selection of a peptide sequence (15-20 amino acids) of the target protein is necessary, which can serve as antigen for the immunization process. The choice of this antigen is crucial for the function and specificity of the resulting antibody. In a first attempt, the antigen for the production of a specific anti-SP-G antibody was chosen solely based on the primary sequence. Therefore, regions with many amino acids that are able to form hydrogen bonds or electrostatic interactions between antigen and antibody residues were selected. However, the resulting antibody for SP-G was not specific in first experiments. This might be due to criteria other than the types of amino acids in the antigen sequence, which are important for a suitable antigen. The protein part used as antigen has to be located on the surface of the protein structure (*solvent-accessible*). If the antigen is part of the protein core region, the antibody may be unable to bind to the antigen sequence. Furthermore, the peptide should be free of PTMs, since bulky modifications, such as glycosylations or palmitoylations, could inhibit proper binding of the antibody to the antigen sequence due to steric clashes (*PTM-free*). Finally yet importantly, the antigen sequence has to be unique within all proteins present in the experimental sample (*unique*). Otherwise, binding of the antibody to the target sequence being part of another protein could be possible as well. This would affect the antibody specificity and most certainly would lead to false positive results. With the help of the previously obtained protein structure models of SP-G and SP-H, putative antigen sequences could be identified, which met all the aforementioned criteria and thus may lead to successful and specific antibodies.

For SP-G, two areas were identified as potential antigens (Figure 9a). The first suggestion covers an α-helix ranging from sequence position 40 to 57 (YESSFLELLEKLCLLLHL) and the second suggestion comprises a β-strand of the amino acids 60 to 70 (GTSVTLHHARS). The results of BLAST searches showed no identical hits, which suggested that both peptide

sequences are *unique* within the human proteome. The first suggestion (40-57) contains not only a lysine and a histidine, but also three negatively charged glutamate residues. These residues are very likely to form electrostatic interactions or hydrogen bonds with residues of the antibody. The only predicted PTM for this region is a phosphorylation at position 40, which should not affect a potential antibody-antigen binding. The second suggestion (60-70) is rather short and contains only one arginine and two histidine residues, which could interact considerably with an antibody. The rest of this sequence part contains mainly hydrophobic amino acids. Unfortunately, glycosylations are predicted for the positions 62 and 70, and the peptide may be partially buried by the adjacent palmitoylation on position 76 and a potential glycosylation on position 78. This situation may prevent the formation of a correct antibody-antigen complex. Thus, only the first suggestion (YESSFLELLEKLCLLLHL) is *solvent-accessible* and *PTM-free* and was suggested for antibody production.

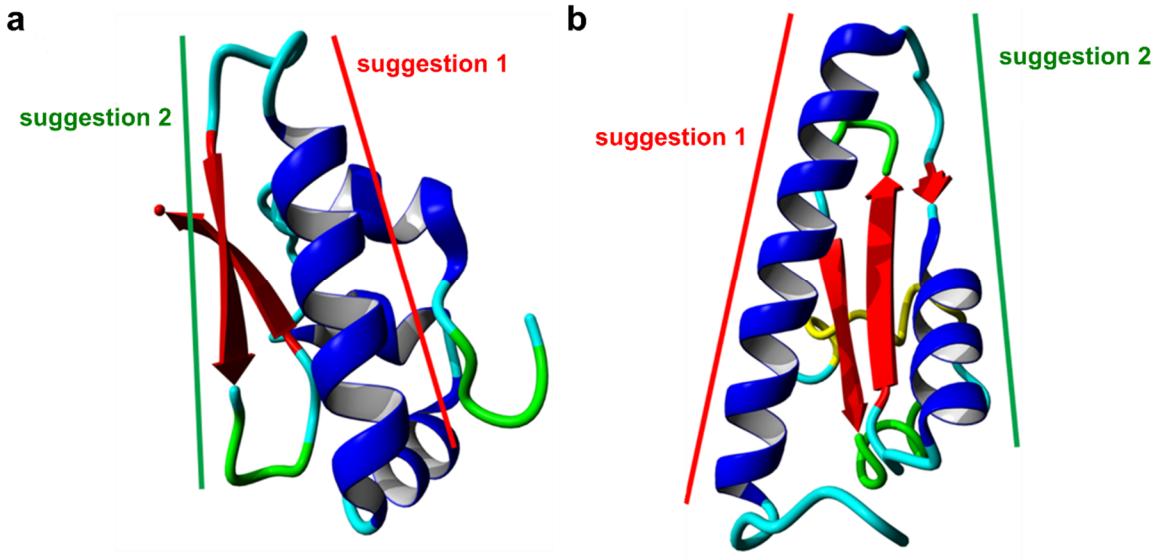


Figure 9: Protein structure models of **(a)** SP-G and **(b)** SP-H with highlighted protein parts, which were suggested as antigens for antibody production. For each protein, two suggestions were selected. The protein models are shown in ribbon representation with α-helices in blue, β-sheets in red, turns in green and random coil elements in cyan.

Two potential antigen sequences could be identified for SP-H as well (Figure 9b). The first suggestion covers the very stable *N*-terminal α-helix from position 7 to 31 (DFQLIRDQVLFLQDQAQRLTEWLQL) and the second suggestion comprises the amino acid positions 35 to 51 (ENPVSESTTLCLREREK). BLAST searches indicated that both peptide sequences are *unique* within the human proteome. Furthermore, both sequences contain various amino acids with functional groups that would allow a specific binding of an antibody. However, whereas there are no predicted PTMs for the first suggestion (*PTM-free*), the predicted *O*-glycosylation on position 39 and palmitoylation on position 45 could interfere with

a proper binding of the antibody to the area of the second suggestion. Furthermore, the spatial proximity of the palmitoylation on position 56 may also cause steric hindrances (*solvent-accessible*). Accordingly, only the first peptide (DFQLIRDQVLFLQDQAQRLTEWLQL) fulfills the criteria for a promising antibody-antigen interaction and was suggested for antibody production.

The company SeqLab (Göttingen, Germany) produced anti-peptide antibodies for SP-G and SP-H with the help of the suggested antigen peptides (YESSFLELLEKLCLLLHL for SP-G, CDFQLIRDQVLFLQDQAQE for SP-H). Martin Schicht [198] verified the specificity of the resulting SP-G antibody by means of Western blot analysis, using protein isolated from lung tissue (30 µg) and the recombinantly synthesized SP-G protein (not purified, 30 µg) (Figure 10a). The purified antibody showed distinct protein bands in lung for SP-G at 11 kDa, 20 kDa and 30 kDa. A distinct protein band for recombinantly synthesized SP-G was visible at about 12 kDa. Lung tissue was used as specific positive control for surfactant proteins. All obtained bands deviated from the molecular weight of the pure SP-G sequence (9 kDa). However, a molecular weight of 11 kDa is calculated when taking the previously predicted PTMs into account (cf. 3.2). Therefore, the band at 11 kDa seemed to represent the mature protein monomer. Since the formation of oligomers could not be excluded based on the modeling results, the bands at 20 and 30 kDa may represent dimer and trimer complexes of SP-G, respectively.

The specificity of the obtained SP-H antibody was also tested by Martin Schicht [199] with protein isolated from lung tissue and bronchoalveolar lavage (30 µg) (Figure 10b). The antibody showed distinct protein bands in lung for SP-H at 13 kDa, 26 kDa and 43 kDa. Lung

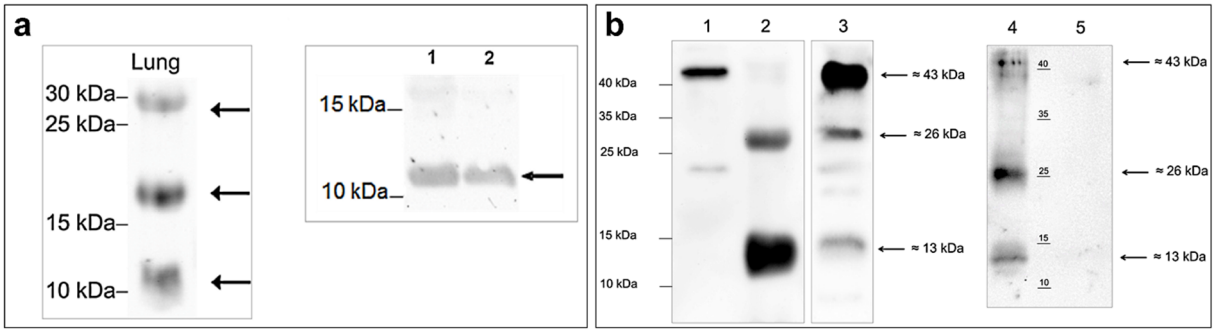


Figure 10: Test of the **(a)** anti-SP-G and **(b)** anti-SP-H antibody by Western blot. **(a)** Figure from [198]. Proteins extracted from lung tissue (positive control) show distinct bands for SP-G at the theoretically expected molecular weights of 11, 20 and 30 kDA (left). The antibody detects a distinct band at 11 kDa for the recombinantly expressed SP-G protein (right) at 28°C (1) and 37°C (2). Arrows indicate positive evidence of SP-G. **(b)** Figure from [199]. Arrows indicate positive evidence of SP-H at molecular weights of 13, 26 and 43 kDa. Results are shown for A549 cells [218] (1), lung tissue (2), bronchoalveolar lavage (3); lung tissue without (4) and after (5) pre-incubation with antigen-peptide.

tissue was used as specific positive control for surfactant proteins. The analysis of bronchoalveolar lavage showed distinct bands at 13 kDa, 26 kDa and 43 kDa. Analog to SP-G, the results deviated from the molecular weight of the pure SP-H sequence (10 kDa). Nevertheless, considering the predicted PTMs resulted in a molecular weigth of 13 kDa. Thus, the band at 13 kDa may represent the monomer and the band at 26 kDa the homodimer of mature SP-H. It is speculative if the band at 43 kDa indicated a SP-H trimer complex with an altered PTM pattern or a complex of SP-H with another protein.

Eventually, the obtained specific antibodies were crucial for the realization of immunohistochemical staining experiments, which demonstrated the presence of SP-G and SP-H in different tissues that typically contain surfactant proteins [198,199]. The occurrence of both proteins in these tissues (among them tissue of the respiratory tract) are a further strong indication that SP-G and SP-H are indeed members of the surfactant protein family.

3.4. Preparation of the protein-lipid simulation system

The DPPC lipid layers used in this work were built from scratch instead of using external sources with pre-equilibrated membrane systems. Therefore, these membrane systems had to be equilibrated prior to the actual protein-lipid simulations. To produce realistic starting structures for DPPC monolayer systems and to verify the cooperation between simulation settings and modified G53a6 force field, a 75 ns MD simulation with a DPPC bilayer was performed with GROMACS. To ensure a reference for future calculations, this simulation should be able to reproduce experimentally determined literature values for DPPC bilayers. Indeed, the trajectories of the last 25 ns of this simulation resemble typical bilayer characteristics (Table 4).

The average value for the volume that each lipid occupies in the layer plane settled at 1.221 nm^3, which was very similar to the experimental literature value of 1.232 nm [185]. The lateral diffusion coefficient of $9.2e^{-8}$ cm^2/s, which describes the movement of single lipids within a layer, nearly matched the experimental value of $9.7e^{-8}$ cm^2/s [186]. The area compressibility of 533 mN/m was far off the experimental value of 231 mN/m, but was in the typical range of reported values for MD simulations (200-600 mN/m) [185,187].

Table 4: Comparison of characteristics for a DPPC bilayer reported in the literature and values obtained from simulations in this work.

value	literature	simulation
volume per lipid (nm^3)	1.232 [185]	1.221
lateral diffusion coefficient (cm^2/s)	$9.7 \cdot e^{-8}$ [186]	$9.2 \cdot e^{-8}$
area compressibility (mN/m)	200-600 [185,187]	533

Moreover, as the primary criteria for the stability of a bilayer system, the averaged area per lipid was calculated for the last 25 ns of the MD simulation. In this simulation, the area per lipid showed only minor fluctuations and remained stable at a level of about 0.625 nm^2 (Figure 11). This was very close to the experimentally determined value of 0.64 nm^2 reported in the literature (blue line in Figure 11) [185]. Altogether, these analyses showed that the chosen force field parameters and simulation settings are able to reproduce a native DPPC bilayer correctly. Furthermore, they suggested that the equilibrated lipid bilayer system could be used as a starting point for the construction of monolayer systems for the protein-lipid MD simulations.

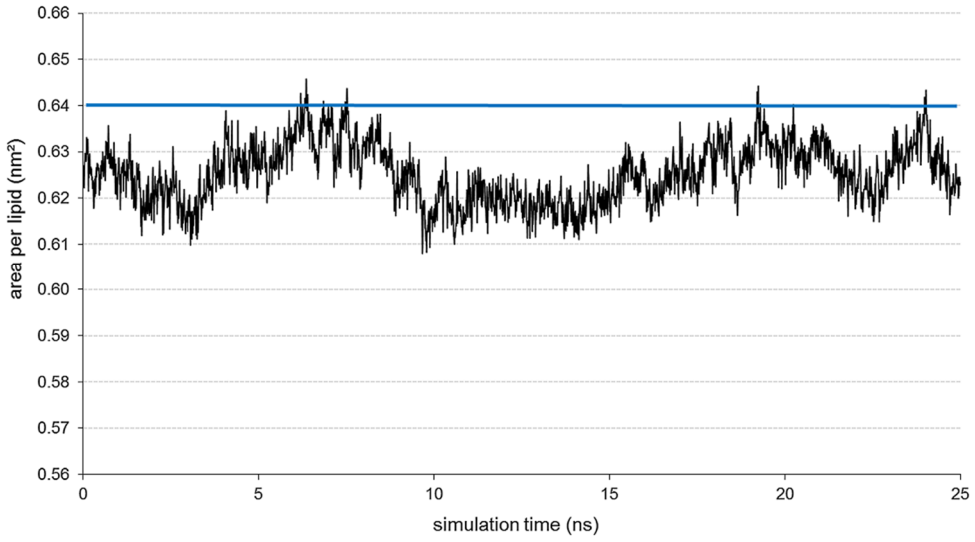


Figure 11: Plot of the area per lipid (in nm^2) for each DPPC molecule in a bilayer patch with 128 lipids during a 25 MD simulation to verify the chosen parameters and simulation settings. The blue line denotes the experimental literature value for a DPPC molecule in a bilayer of 0.64 nm^2 at 323 K [185].

To achieve a good agreement between the characteristics of the lipid simulation system and a natural bilayer, the simulation temperature was set to 323 K. This temperature is above the phase transition temperature for DPPC, so that the system achieves the biologically relevant fluid L_α state instead of the more ordered gel or subgel state [200-202]. However, the question

arose if this high temperature would have any negative effect on the stability of the protein models. To address this question, MD simulations for all four models (SP-G and SP-H, each with and without PTMs) were performed at 298 K and 323 K for 20 ns. The results of the simulations with both temperatures were compared for each model to identify differences in the model stability.

For the SP-G model without PTMs, the level of RMSD values was slightly higher at 323 K in comparison to the simulation at 298 K (Figure 12a). The reason for this may be the overall higher energy in the system due to the increased temperature. However, the model seems to be equally stable with only minor fluctuations in the plot after 10 ns. The RMSD plot showed that the SP-G model with PTMs was more stable at 298 K compared to 323 K (Figure 12b). The higher system temperature enhanced the movements of the attached PTMs. Especially the bulky *N*-glycosylation on position 37 induced structure fluctuations in the nearby protein regions.

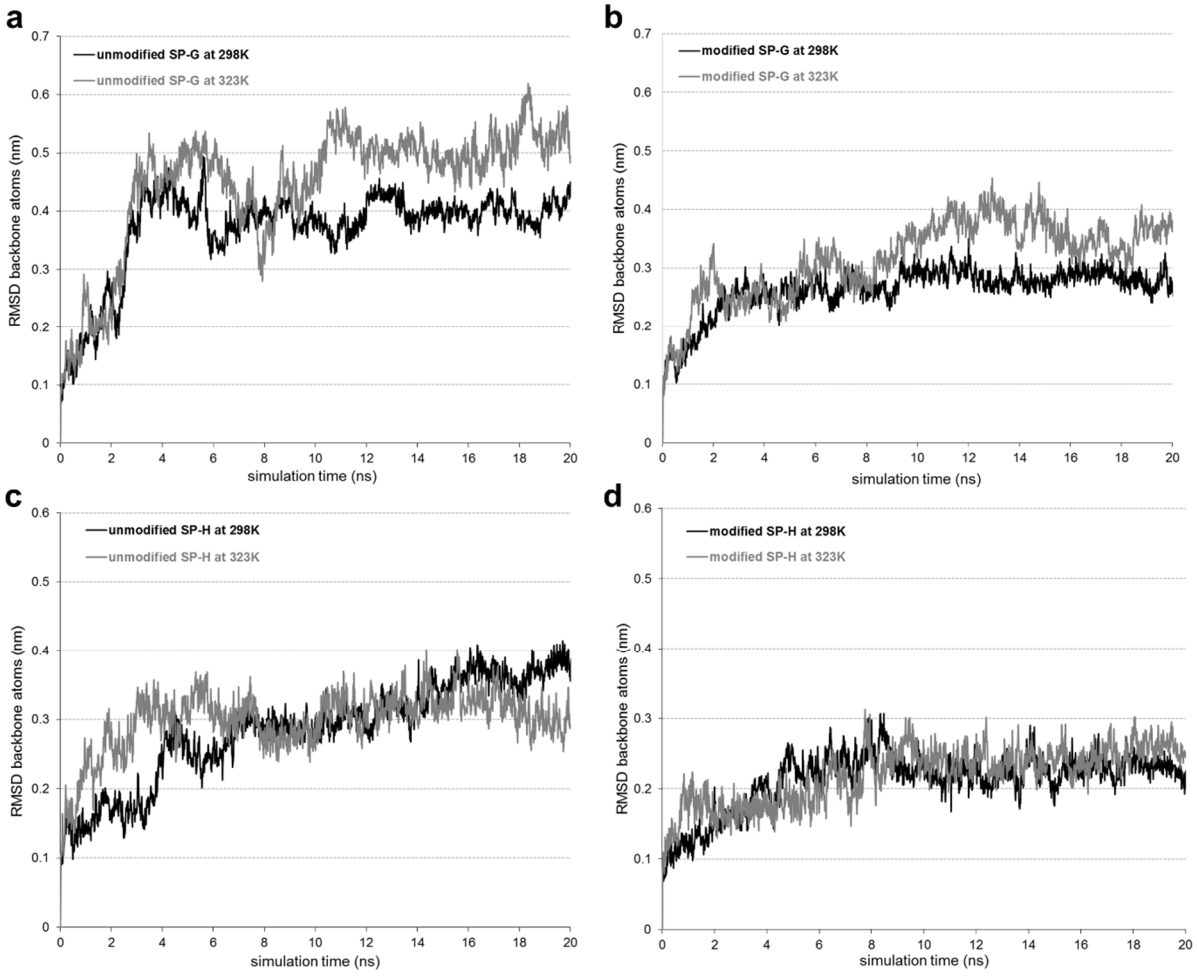


Figure 12: Influence of simulation temperature on the protein model stability. The RMSD of the backbone atoms (in nm) of **(a)** the SP-G model without PTMs, **(b)** the SP-G model with PTMs, **(c)** the SP-H model without PTMs, and **(d)** the SP-H model with PTMs is compared. 20 ns MD simulation with 298 K (black plots) and 323 K (grey plots) were performed using GROMACS to investigate the influence of higher temperatures on the protein models.

Nevertheless, the “core regions” of the protein were quite stable and no secondary structure changes or a general unfolding of the protein was observed.

Remarkably, the RMSD plot for the SP-H model without PTMs at 323 K showed a more stable progression compared to the simulation at 298 K (Figure 12c). Whereas the plot for 298 K was slightly increasing after 14 ns until the end of the simulation, the RMSD at 323 K remained on the same level after 5 ns. For the SP-H model with PTMs, no differences between the RMSD plots at 298 K and 323 K were observable (Figure 12d). For this model, the temperature seemed not to have any influence on the protein stability. In both cases, the RMSD plot showed a very stable progression and was equilibrated after 8 ns.

In conclusion, the higher temperature of 323 K showed no significant influence on the protein stability of any of the four models. Therefore, the protein models were combined with the previously established lipid layer to generate the starting structures for the protein-lipid simulations (cf. 2.4.2). For each of the four protein models (SP-G with and without PTMs, SP-H with and without PTMs), six different starting orientations were created, which resulted in overall 24 different simulation systems for the following 50 ns MD simulations (cf. Figure 4).

3.5. Protein-lipid molecular dynamics simulation analysis

After 50 ns of protein-lipid MD simulation, the final trajectory snapshots and statistics files were analyzed for all 24 performed simulations (cf. Figure 4). The most important finding thereby is that in all 24 orientations the protein models started to interact with the lipid layer. However, the protein parts that were responsible for the protein-lipid interactions were highly diverse. In the final trajectory overlay of all six simulations per model (Figure 13), no specific interaction site or "consensus orientation" could be identified for any of the four models. To select a representative result for each of the four cases, the protein-lipid interaction energy calculated by the force field was used as major criterion (Appendix 9, Appendix 10). Additionally, the protein stability measured by the RMSD (backbone atoms) was checked as well (Appendix 11, Appendix 12). In the following sections, the obtained results are described in more detail, separately for SP-G and SP-H. Additionally, the results of the orientations are presented, where the SP-G or SP-H model with PTMs was manually positioned to interact with the lipid layer already at simulation start ("positioned").

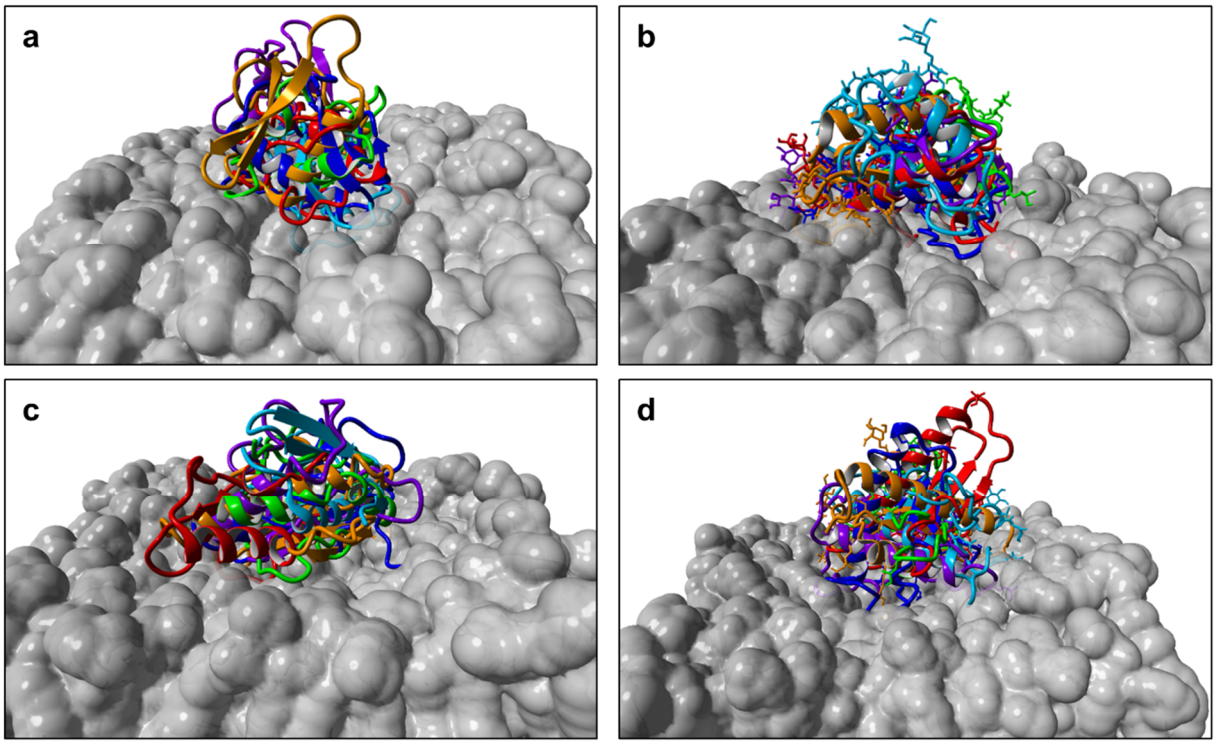


Figure 13: Resulting structures of MD simulations of the SP-G and SP-H models in a lipid environment. The final trajectories of all six performed simulations (orientations) for each model are superimposed in one picture, separately for **(a)** the SP-G model without PTMs, **(b)** the SP-G model with PTMs, **(c)** the SP-H model without PTMs, and **(d)** the SP-H model with PTMs. The DPPC lipids are shown as a grey surface. Protein backbone and atoms of the PTMs (in **(b)** and **(d)**) are colored differently for each orientation.

3.5.1. Detailed analysis for SP-G

In the SP-G model without PTMs with the most negative protein-lipid interaction energy (-1100 kJ/mol, orientation 2), parts of the *N*-terminal signal peptide (1-14) and residues of the α-helix 41-58 were mostly responsible for the protein-lipid interaction (Figure 14a). Thereby, the signal peptide was aligned parallel to the lipid surface after it reached the polar lipid head groups. The α-helical conformation of the *N*-terminus was lost during this process (cf. Figure 6). Furthermore, the resulting positioning of the protein allowed the interaction of the α-helix 41-58 with the monolayer in an almost parallel orientation. The first interactions established after 6 ns, as visible in the interaction energy plot (Figure 15a, black plot). After 30 ns, the interaction energy remained stable at about -1100 kJ/mol. The protein backbone RMSD plot for this simulation was not completely equilibrated, but almost constant with only minor fluctuations after 25 ns, which indicated a stable protein structure (Figure 15b, black plot). During this

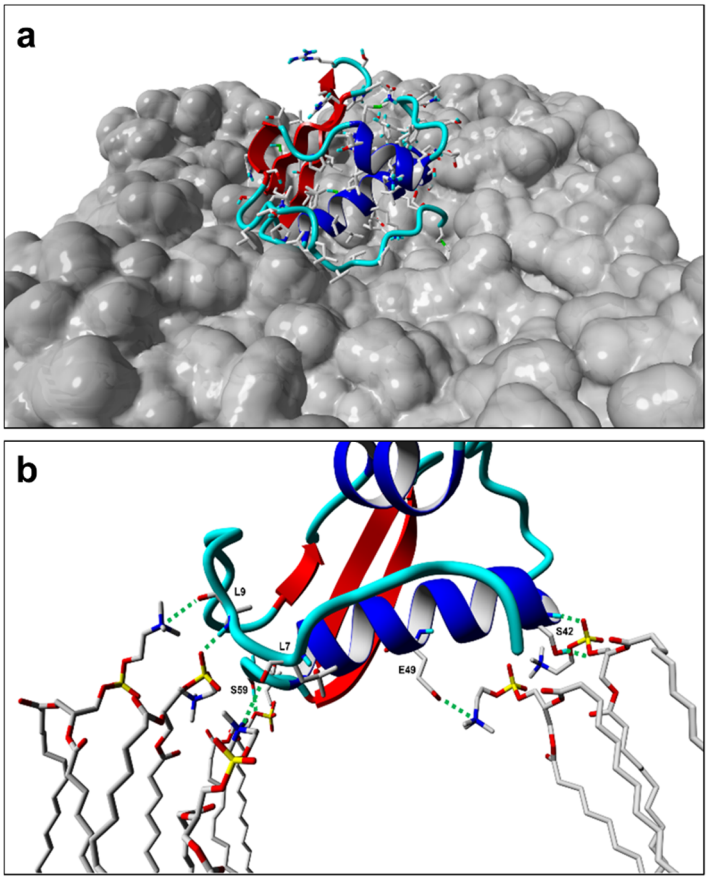


Figure 14: Detailed simulation results for the SP-G model without PTMs. **(a)** Representation of the SP-G model without PTMs with the most negative protein-lipid interaction energy after 50 ns of MD simulation. The DPPC lipids are shown as grey surface. **(b)** Detailed representation of the protein-lipid interaction site of the system in **(a)**. For clarity reasons, the view was slightly rotated. Green dashed lines indicate interactions between amino acid side chains and lipids. In both pictures, the amino acids and lipids are shown in stick representation without aliphatic hydrogens.

steady phase, the hydrophobic residues of the signal peptide penetrated deeper into the polar lipid head groups and reached the top of the aliphatic lipid tails. However, a closer investigation of the protein-lipid interaction site revealed that only five amino acids with polar side chains interact with the lipid head groups (Figure 14b). In the final simulation snapshot, three hydrogen bonds and four ionic interactions between protein side chains and lipid phosphate or choline moieties were responsible for a moderate fixation of the protein on the lipid surface.

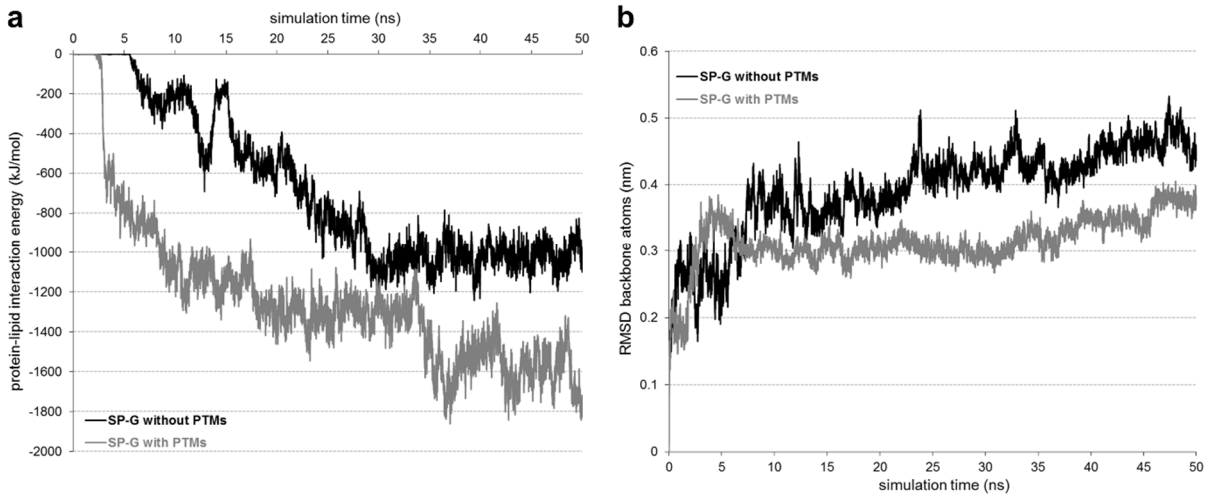


Figure 15: (a) Protein-lipid interaction energy (in kJ/mol) and **(b)** backbone atoms RMSD plots (in nm) for the SP-G model without (black plots) and with PTMs (grey plots). In both cases, only the results for the orientations with the most negative protein-lipid interaction energy after 50 ns MD simulation are shown.

For the SP-G model with PTMs and the most negative interaction energy (orientation 3), mainly the 18 *N*-terminal signal peptide residues as well as the amino acids 29-43 were in contact with the lipid layer (Figure 16a). In contrast to the simulation without PTMs, the signal peptide maintained its α-helical fraction. During the simulation, the protein approached the monolayer very quickly. First protein-lipid interactions were visible after 3 ns (Figure 15a, grey plot) and increased quickly thereafter. Unfortunately, the interaction energy was not stable at the end of the simulation. If the simulation would be continued, the interaction energy probably might trend towards a more negative value. The fact that the RMSD plot did not equilibrate after 50 ns (Figure 15b, grey plot) reflects this trend as well. Conformational changes of the protein while approaching the lipid layer surface to optimize atomic interactions certainly caused these fluctuations in both graphs. However, the interaction energy of ca. -1800 kJ/mol at the end of the simulation with PTMs attached to the SP-G model was already significantly more negative than the energy observed for the SP-G model simulation without PTMs (-1100 kJ/mol). This high interaction energy was also apparent from the protein-lipid interaction site (Figure 16b).

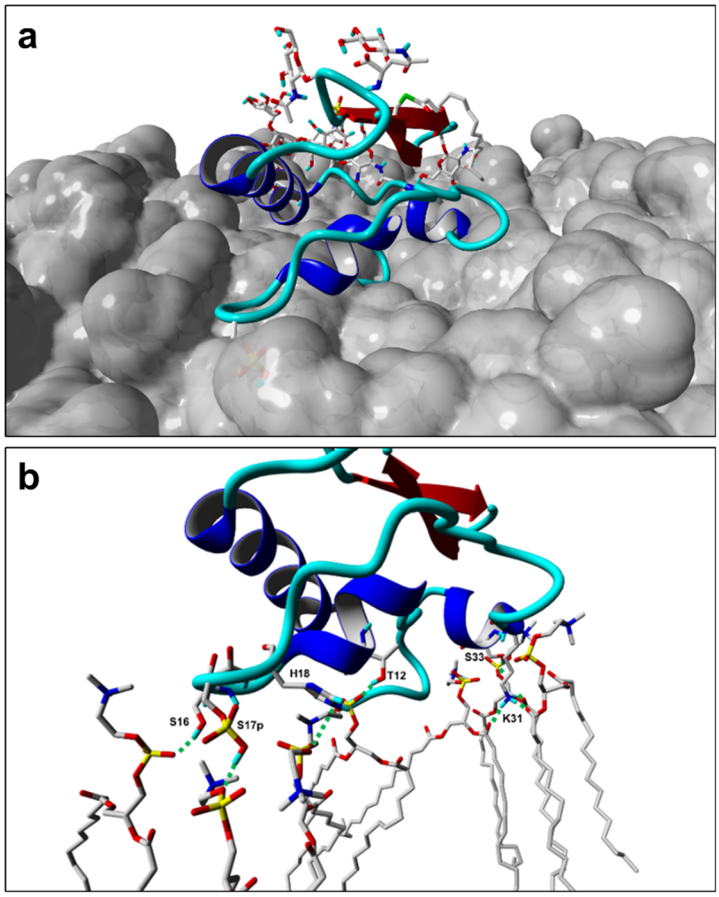


Figure 16: Detailed simulation results for the SP-G model with PTMs. **(a)** Representation of the SP-G model with PTMs with the most negative protein-lipid interaction energy after 50 ns of MD simulation. The DPPC lipids are shown as grey surface. **(b)** Detailed representation of the interaction site of the system in **(a)**. For clarity reasons, the view was slightly rotated. Green dashed lines indicate interactions between amino acid side chains and lipids. "p" labels a phosphorylated residue. In both pictures, the amino acids and lipids are shown in stick representation without aliphatic hydrogens.

Compared to the results of the unmodified SP-G model, the number of interacting amino acids was increased (nine instead of five). Due to clarity reasons, the interactions of Gly2, Ser3, and Glu46 are not shown in Figure 16b. Hydrogen bonds were the dominant interaction type and Lys31 alone interacted with fatty acid carbonyl groups of three different lipids. However, only one modified residue (phosphorylated Ser17) interacted with a lipid molecule, all other PTMs (cf. Table 3) resided in the water phase. This is also true for the palmitoylation on Cys76, which was located opposite to the lipids (Figure 16a, green sulfur atom marks Cys76) and resided in hydrophobic cavities on the protein surface to avoid unfavorable contact with the water.

In the special case, where the SP-G model with PTMs already interacted with the lipids at simulation start ("positioned"), the protein was positioned in a way that the palmitoylation on Cys76 crossed the polar region of the lipid head groups and reached into the area of the aliphatic lipid chains. Accordingly, a completely different interaction site between protein and lipids

resulted from 50 ns MD simulation (Figure 17a) compared to the previously discussed MDs. The α-helix 41-58 was located on the surface of the complex and interacted with the solvent and not with the lipid surface as described for the orientations before (Figure 14a, Figure 16a). Instead, the *C*-terminal region and the protein part ranging from position 15 to 30 interacted considerably with the lipids, where the latter immersed deeply into the monolayer (Figure 17a, red surfaces). The α-helical character of the signal peptide was maintained over the whole simulation. The signal peptide covered the hydrophobic protein core and was nicely stabilized in this conformation. The position of the whole protein enabled the interaction of glycosylations with the lipids instead of the water phase. Accordingly, the protein was fixed on three points by the attached glycosylations of Asn37, Ser62, Ser70, and Thr78 (Figure 17a, red surfaces). These

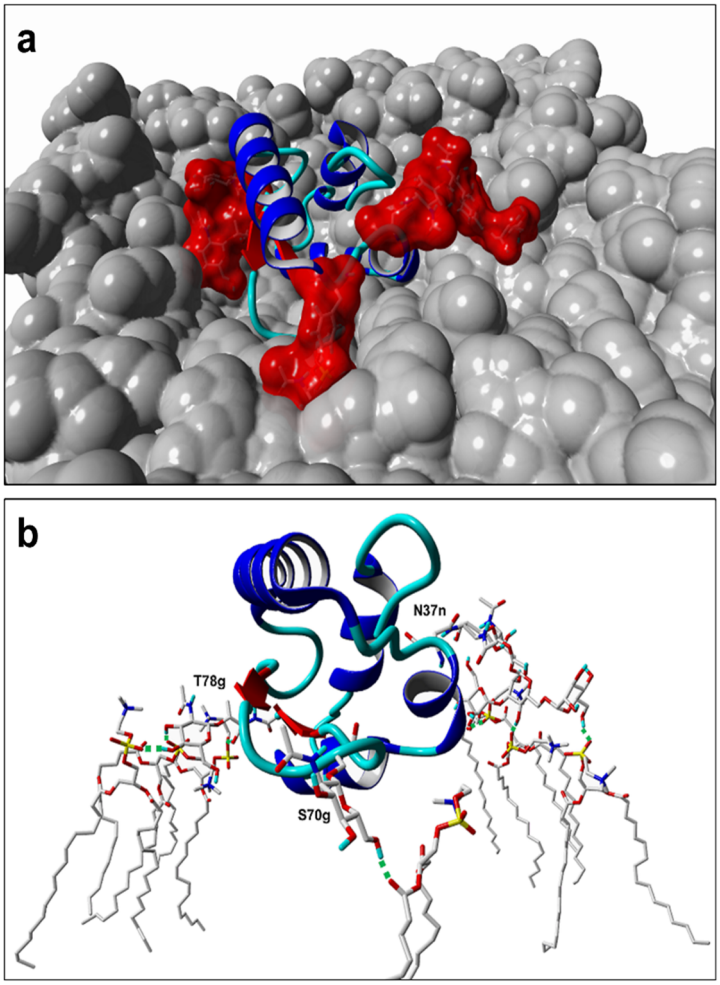


Figure 17: Detailed simulation results for the pre-positioned SP-G model with PTMs. **(a)** Representation of the SP-G model with PTMs, which was positioned to interact with the lipid layer prior to MD simulation. The picture was taken after 50 ns. The DPPC lipids are shown as grey and the major interacting protein residues as red surface. **(b)** Detailed representation of the protein-lipid interaction site of the system in **(a)**. Due to clarity reasons, the view was slightly rotated. Green dashed lines indicate interactions between amino acid side chains and lipids. The label "g" indicates an *O*-glycosylation, the label "n" an *N*-glycosylation. In both pictures, the amino acids and lipids are shown in stick representation without aliphatic hydrogens.

glycosylations formed numerous hydrogen bonds with the phosphate groups or ester bond regions of DPPC and acted like anchors for the protein on the lipid layer surface (Figure 17b). Due to clarity reasons, the interactions of phosphorylated Tyr40 and glycosylated Ser62 are not shown in Figure 17b. Surprisingly, the protein-lipid interaction energy of the resulting complex was not significantly higher than for the previously described orientation with modified SP-G (Figure 18a). The reason may be the low number of observed interactions between amino acid side chains and lipids – apart from the mentioned glycosylated residues. Another reason could be repulsive energy terms due to the palmitoylation on Cys76. Although it was placed to be inside the lipid tail region at the simulation start, the interactions were not sufficient to stabilize this position. Therefore, it left the hydrophobic area due to conformational changes and is located unfavorably in the polar lipid head group region at the end of the simulation. This may be a hint that SP-G is a membrane-associated instead of a membrane-integrated protein. However, the interaction energy was very stable after 30 ns. This was also true for the protein RMSD plot (Figure 18b), which was strikingly stable due to the "glycosylic anchors".

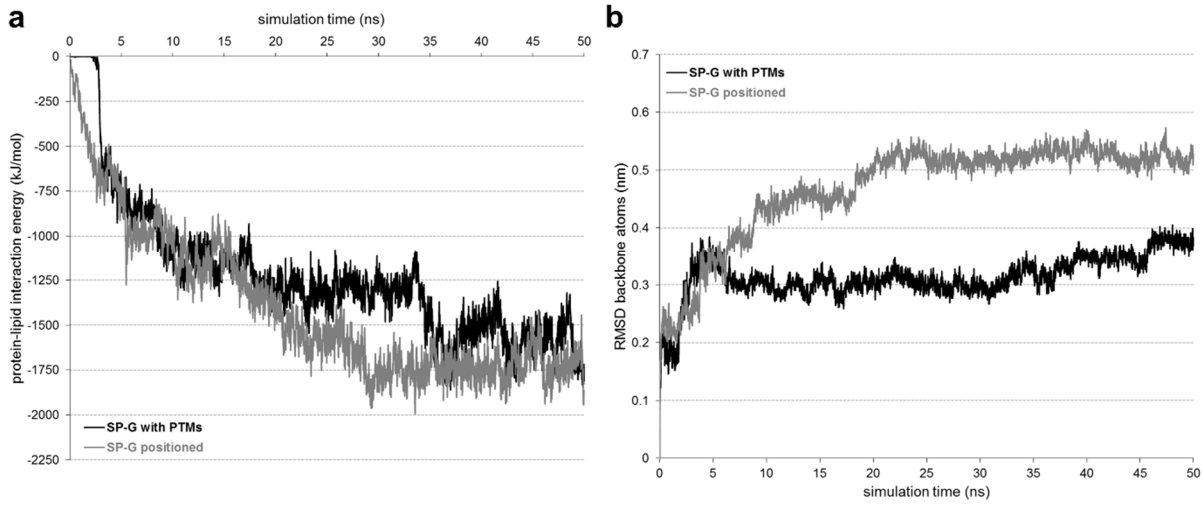


Figure 18: (a) Protein-lipid interaction energy (in kJ/mol) and **(b)** backbone atoms RMSD plots (in nm) for the SP-G model with PTMs and most negative interaction energy (black plots) and the pre-positioned SP-G model (grey plots).

3.5.2. Detailed analysis for SP-H

The simulation of the SP-H model without PTMs that possessed the most negative protein-lipid interaction energy (orientation 3) showed a huge contact area between protein and lipids (Figure 19a). In detail, especially the 27 *N*-terminal and nine *C*-terminal amino acids were in close contact with the lipid layer. Accordingly, the interaction energy plot showed a steady increase after the first contact at 2 ns until it reached a plateau after 40 ns at ca. -2300 kJ/mol (Figure 20a, black plot). The protein model, meanwhile, is extremely stable in this simulation. There are no major fluctuations of the RMSD plot after a simulation time of 10 ns and the model can be denoted as equilibrated after 20 ns (Figure 20b, black plot). The hydrophobic amino acids of the α-helix caused a hollow on the monolayer surface, which enabled the immersion of the *C*-terminal protein parts below the head group region. The reason for the model stability could be the different interactions between numerous amino acids and lipid head groups, which

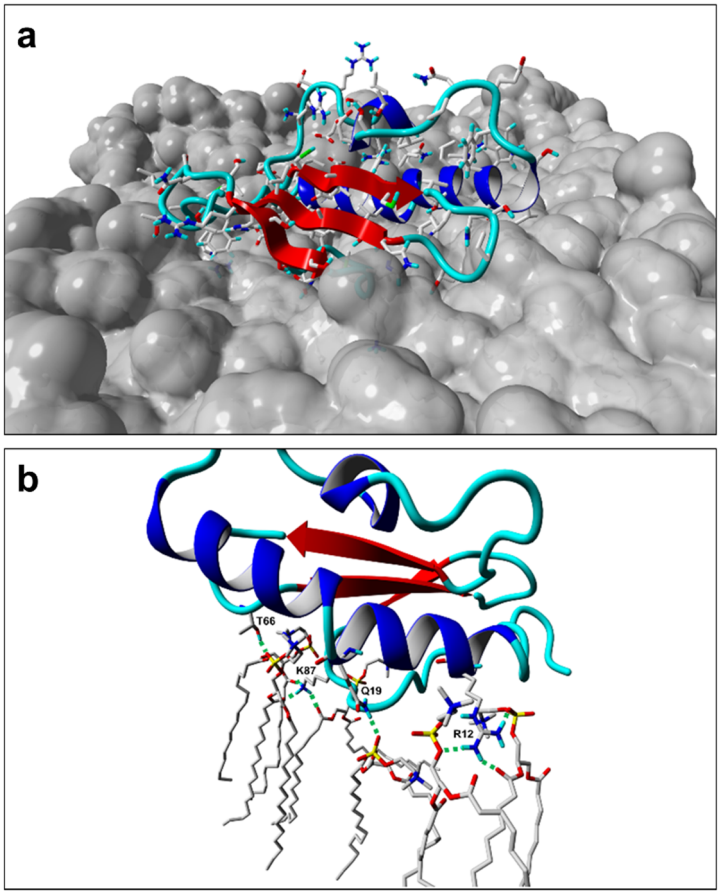


Figure 19: Detailed simulation results for the SP-H model without PTMs. **(a)** Representation of the SP-H model without PTMs with the most negative protein-lipid interaction energy after 50 ns of MD simulation. The DPPC lipids are shown as grey surface. **(b)** Detailed representation of the protein-lipid interaction site of the system in **(a)** Due to clarity reasons, the view was slightly rotated. Green dashed lines indicate interactions between amino acid side chains and lipids. In both pictures, the amino acids and lipids are shown in stick representation without aliphatic hydrogens.

fixed the protein on the lipid surface (Figure 19b). Among others, positively charged amino acid side chains (Arg2, Arg12, Arg87) formed three of nine observed interactions and served as fixation points in the ester bond region of the lipid layer. Due to clarity reasons, the residues Arg2, Gln23, Glu27, Met88, and Leu89 are not shown in Figure 19b.

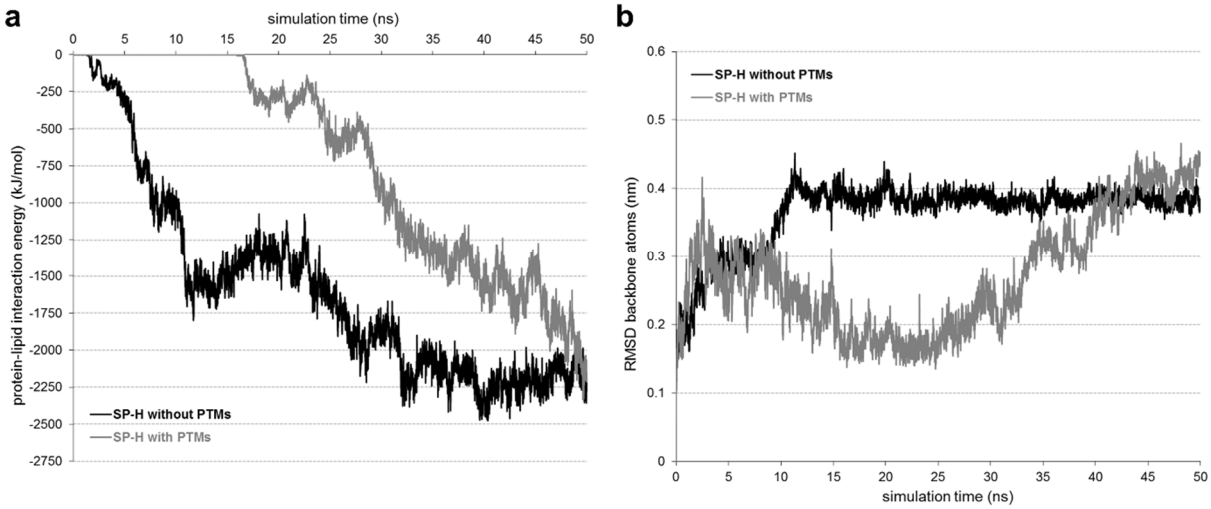


Figure 20: (a) Protein-lipid interaction energy (in kJ/mol) and **(b)** backbone atoms RMSD plots (in nm) for the SP-H model without (black plots) and with PTMs (grey plots). In both cases, only the results for the orientations with the most negative protein-lipid interaction energy after 50 ns MD simulation are plotted.

The SP-H model with PTMs and the most negative interaction energy (orientation 5) also showed a large contact area mainly with the residues 32-51 and the *N*-terminus, but phosphorylated *C*-terminal residues were very important as well (Figure 22a, phosphorylated Ser80, Ser82, and Ser84 not shown). The first protein-lipid contact was observable in the interaction energy plot after 16 ns (Figure 20a, grey plot). The value was quickly increasing to a level comparable to the simulation without PTMs (-2300 kJ/mol). Unfortunately, the interaction energy was not stable at the end of the calculation. This instability was also reflected in the RMSD plot (Figure 20b, grey plot), which showed significant fluctuations until the end of the simulation. However, this orientation showed the most negative interaction energy in comparison to the other five simulations with modified SP-H (Appendix 10b). Therefore, this simulation was extended until 100 ns to estimate the reliability of the results after 50 ns. The results showed a stable interaction energy of about -2300 kJ/mol (Figure 21a) and an equilibrated protein model with respect to the RMSD after 60 ns (Figure 21b). Since there were no major changes to the results, the values obtained after 50 ns, although not equilibrated, could be compared to the simulation without PTMs. This comparison indicated almost no difference in the most negative interaction energy between the SP-H model without and with PTMs. The detailed investigation of the interaction site (Figure 22b) showed that the number of interacting

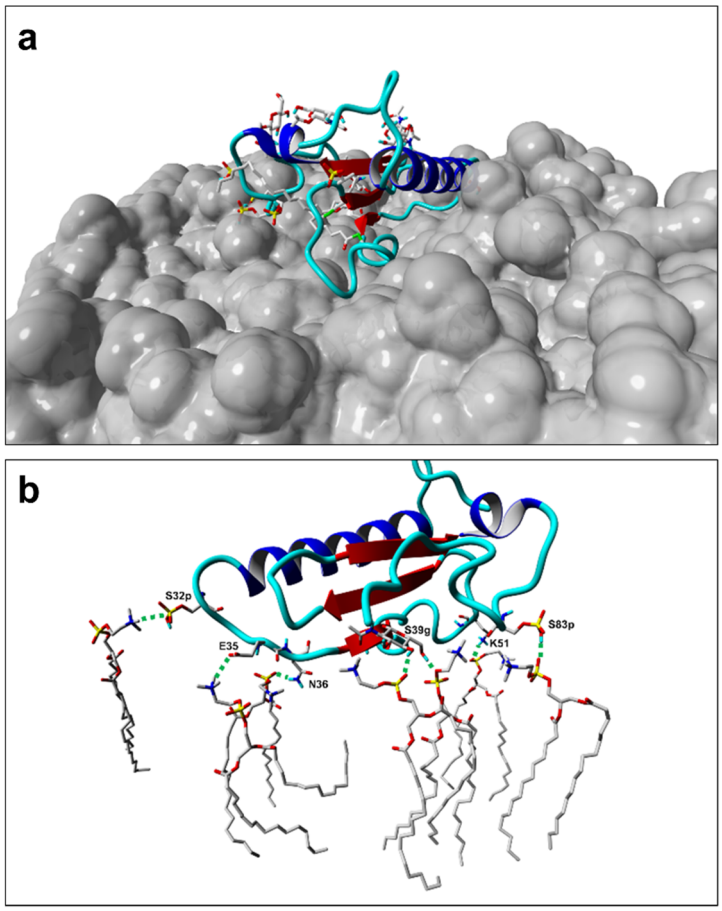


Figure 22: Detailed simulation results for the SP-H model with PTMs. **(a)** Representation of the SP-H model with PTMs with the most negative protein-lipid interaction energy after 50 ns of MD simulation. The DPPC lipids are shown as grey surface. **(b)** Detailed representation of the protein-lipid interaction site of the system in **(a)**. For clarity reasons, the view was slightly rotated. Green dashed lines indicate interactions between amino acid side chains and lipids. The label "g" indicates an *O*-glycosylation and the label "p" a phosphorylation. In both pictures, the amino acids and lipids are shown in stick representation without aliphatic hydrogens.

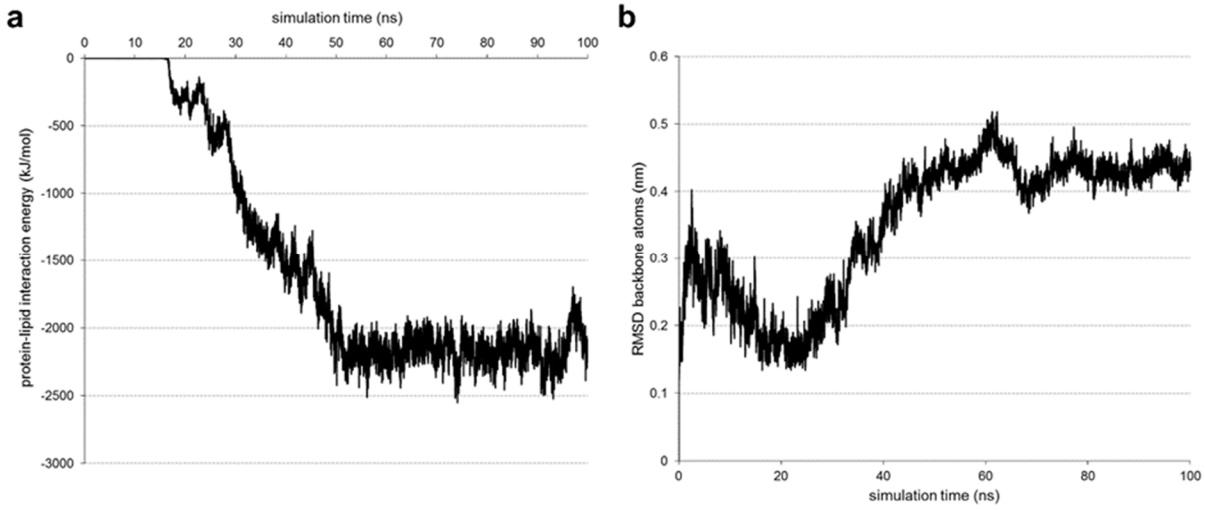


Figure 21: (a) Protein-lipid interaction energy (in kJ/mol) and **(b)** protein backbone RMSD (in nm) for the SP-H model with PTMs and most negative protein-lipid interaction energy. Since interaction energy and RMSD were unstable after 50 ns, this simulation was extended until 100 ns.

amino acids was significantly increased from nine (unmodified SP-H) to 14 (modified SP-H). Among them, two glycosylated residues (Ser39 and Thr93) and two phosphorylated amino acids (Ser32 and Ser83) were responsible for a huge part of the protein-lipid interaction energy. Due to clarity reasons, the interactions of Arg24, Trp28, Leu31, Thr42, Arg49, Glu50, Ala94, and glycosylated Ser39 are not shown in Figure 22b. The palmitoylations on Cys45 and Cys56 did not seem to play any role for the protein-lipid interaction. The hydrophobic moieties were located near the lipid surface, but they were integrated into the hydrophobic core of the protein, consequently avoiding the polar lipid head groups in this way.

The resulting interaction complex of the pre-positioned orientation (Figure 23a), where the palmitoylations at Cys45 and Cys56 were determined to interact directly with the lipids at simulation start, looked very similar to the described SP-H model with PTMs and most negative

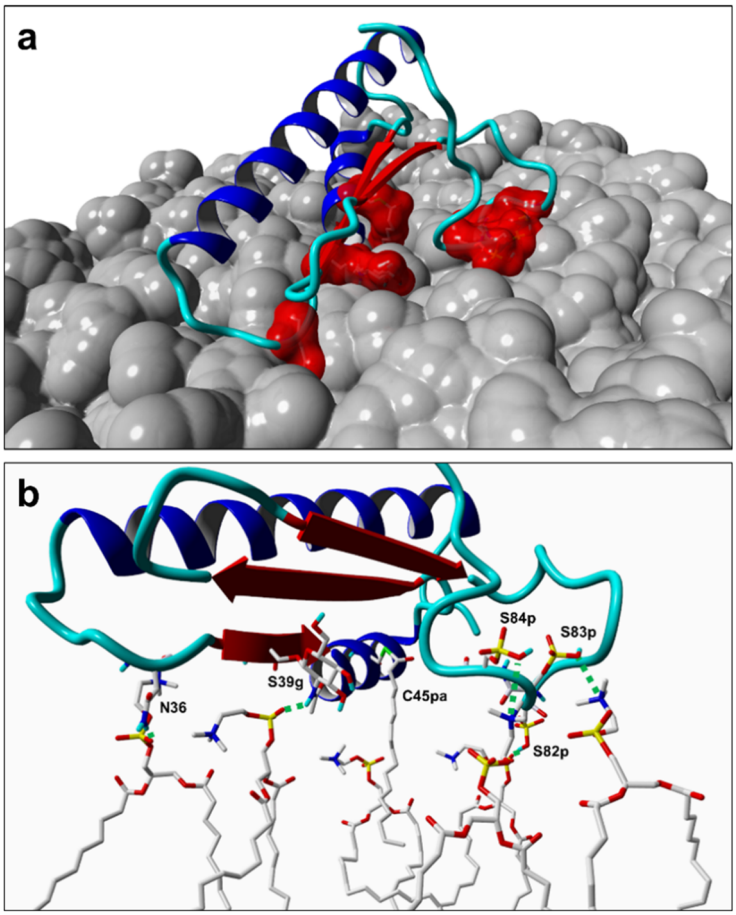


Figure 23: Detailed simulation results for the pre-positioned SP-H model with PTMs. **(a)** Representation of the SP-H model with PTMs, which was positioned to interact with the lipid layer at simulation start. The picture was taken after 50 ns. The DPPC lipids are shown as grey and the major interacting protein residues as red surface. **(b)** Detailed representation of the protein-lipid interaction site of the system in **(a)**. Due to clarity reasons, the view was slightly rotated. Green dashed lines indicate interactions between amino acid side chains and lipids. The label "g" indicates an *O*- glycosylation, the label "p" a phosphorylation, and the label "pa" a palmitoylation. Amino acid, PTM, and lipid atoms are shown in stick representation without aliphatic hydrogens.

interaction energy (cf. Figure 22b and Figure 23b). However, a small deviation in the contact angle between protein and lipid layer led to a quite different interaction pattern. Positioning of the palmitoylations at Cys45 and Cys56 into the hydrophobic lipid phase resulted in a main interaction spot of amino acids around these residues (Figure 23a, red surfaces). Furthermore, single amino acids between the positions 35 and 40, and the residues 80-85 were very important for the protein-lipid interaction. Most of the interactions could be found in the cluster of phosphorylated serine residues at positions 80, 82, 83, and 84 (Figure 23b, due to clarity reasons, Ser80 is not shown). Furthermore, Asn36 and the glycosylated Ser39 stabilized the protein on the lipid surface. The palmitoylation on Cys45 still penetrated the lipid head group region and was in contact with the hydrophobic lipid tails at the end of the simulation. In contrast to that, the palmitoylation on Cys56 left the lipid layer during the simulation. Due to the pre-positioning, the interaction energy started at -500 kJ/mol and continued to decrease until a level of about -1250 kJ/mol was reached, where it remained stable until the simulation end (Figure 24a). Although the interaction energy of the pre-positioned modified SP-H showed only minor fluctuations, it was significantly less negative compared to the previously described simulation for the SP-H model with PTMs in Figure 22. This was mainly the result of a strongly reduced number of directly interacting residues (7 instead of 14). Nevertheless, the result of the pre-positioned model was robustly fixed on the lipid layer, which suggested that SP-H is membrane-associated. As expected from the stable interaction energy, the RMSD showed only minor fluctuations and indicated a stable protein structure (Figure 24b).

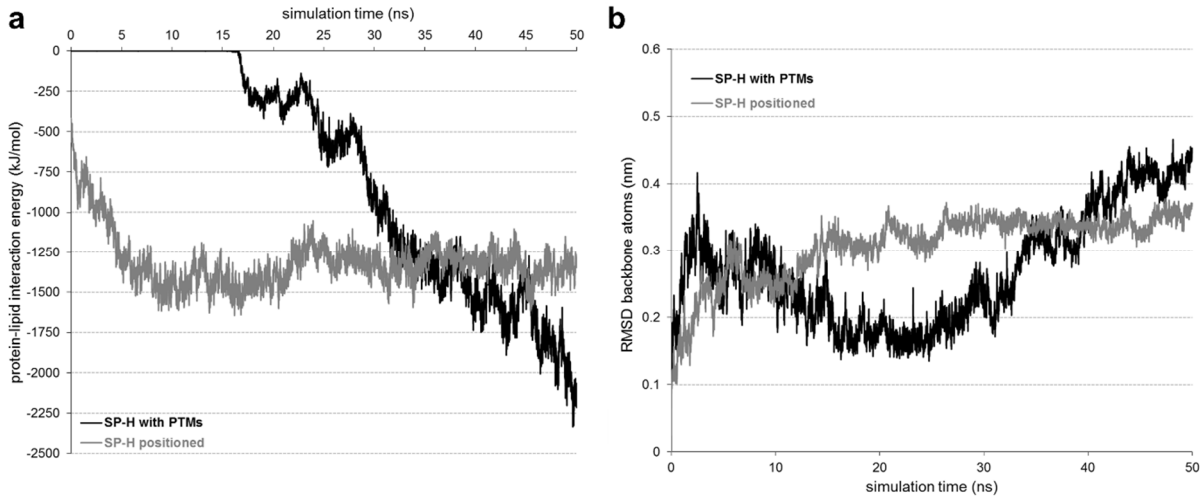


Figure 24: (a) Protein-lipid interaction energy (in kJ/mol) and **(b)** backbone atoms RMSD plots (in nm) for the SP-H model with PTMs and most negative interaction energy (black plots) and the pre-positioned SP-H model (grey plots).

3.5.3. General findings and summary of the protein-lipid MD simulations

The fluctuation analysis of each protein residue during the simulation (RMSF) for all 24 orientations (Appendix 13, Appendix 14) indicated a generally reduced fluctuation of protein parts that were directly interacting with polar lipid head groups. This was due to hydrogen bonds or electrostatic interactions of the amino acid side chain atoms. In some cases, even protein backbone atoms interacted with lipid head groups, which stabilized the protein backbone and further reduced the fluctuations of these amino acids. Polar PTMs, such as phosphorylations or glycosylations, might enhanced this effect. In contrast to that, these PTMs increased the fluctuation of their attached protein parts if they were oriented towards the water phase.

The area per lipid was also monitored in all simulations (Appendix 15, Appendix 16), but a general influence of the protein-lipid interactions on the area per lipid value could not be observed. For all simulations, the area per lipid reached approximately 0.54 nm^2 with a fluctuation of about ± 0.02 nm^2. There are cases, where the binding of the protein caused an effect. Especially when the protein penetrated deep into the lipid head group region, the area per lipid plot may be influenced. However, the changes were not significant and got lost in the general “fluctuation noise” caused by the MD methodology.

DSSP was used to investigate the secondary structure elements of the protein models during the simulation. The number of residues with assigned secondary structure (α-helix, β-sheet, β-bridge or turn) remained almost constant for all performed simulations (Appendix 17, Appendix 18). Neither attached PTMs nor the interaction with lipids seemed to have an influence on the stability of the core structure elements of the proteins. There were some orientations with small plot fluctuations, but they did not suggest a major change in the protein fold or an unfolding of the structure. Thus, the observed stability of the protein models was a very important result of the MD simulations. The knowledge about the consistency of the protein fold will be very important for computational studies in future.

In summary, it can be noted that by means of the performed MD simulations, three potential poses for each protein could be identified, which demonstrated the possibility of SP-G and SP-H to interact with a lipid system. The simulations showed that the PTMs added to the protein models have a significant influence on the possible protein-lipid interactions. According to the simulation results of this work, the attached palmitoylations could determine the interaction site, and the glycosylations may be very important to stabilize a resulting protein-lipid complex. Additionally, the simulations indicated that SP-G and SP-H are associated on the surface of a

lipid layer, similar to SP-A and SP-D, in contrast to layer-integrated surfactant proteins such as SP-B or SP-C. This is evident from the fact that the proteins did not completely integrate into the layer during the MD simulations and remained in the lipid head group region. Similar results were obtained if the protein models were placed very close to the lipid layer at the beginning of the simulation ("positioned" orientations). However, longer MD simulations may be needed to observe more negative interaction energies and, probably, a deeper immersion of the protein into the lipid layer in some cases. Unfortunately, an effect of the proteins on the lipid layer stability was not observed, which might be due to the methodology. Nevertheless, all analyzed situations are valuable suggestions, which will have to be further investigated by experimental studies. The results of the performed MD simulations indicated that SP-G and SP-H have an increased probability to interact with lipid systems and that this interaction potential is dependent on the attached PTMs. Moreover, they seem to be associated to a lipid layer. All these points further justify the classification of SP-G and SP-H as members of the surfactant protein family, since the aforementioned observations represent typical characteristics for surfactant proteins.

4. Discussion

4.1. Protein structure modeling and posttranslational modifications

Although there were no proteins with an already known 3D structure and a sufficient sequence identity available, comparative modeling was performed for the SP-G and SP-H sequence. As expected, this approach failed; however, these homology models were presented in this study. Especially for SP-G, the model quality was problematic in three out of five quality evaluation tools (cf. Table 2). Furthermore, the models obtained by the more sophisticated threading approach showed also deficiencies in their quality. In particular, the stereochemical quality (Ramachandran plot results) of these threading models was inacceptable, especially in direct comparison to the stereochemical quality of the homology models (cf. Table 2). An improvement of these values by YASARA refinement MDs was not possible, which might be due to the tight packing of the protein structures in the I-TASSER modeling process [112,113]. Finally, *ab initio* protein structure modeling with ROBETTA was able to generate reliable structure models for SP-G and SP-H. Common evaluation tools and a 20 ns MD simulation showed the good quality and stability of both models (Table 2, Figure 5). These results demonstrated that ROBETTA is able to produce high-quality models for practically oriented studies as well, besides the excellent performance in structure modeling contests (CASP) [203].

In the literature, the high impact of posttranslational modifications (PTMs) regarding the stability and function of surfactant proteins is well known [17,18]. Depending on the modification type, polar (phosphorylation, glycosylation) or hydrophobic (palmitoylation) regions could be formed on the protein surface, which could change the solubility of the protein in an aqueous environment. Moreover, these modified regions could significantly influence the interaction potential of a protein to a lipid system. Furthermore, PTMs might be responsible for amphiphilic properties of known surfactant proteins [17,18]. To consider this possibility for SP-G and SP-H, the PTMs obtained by sequence-based prediction tools were attached to the final SP-G and SP-H models (Table 3). The predictions indicated a high grade of modification, similar to the already known surfactant proteins. Apart from polar modifications, such as

phosphorylations and glycosylations, the prediction servers suggested hydrophobic modifications as well, which might influence the properties of the protein surface significantly. For example, whereas the amino acid sequence of SP-G suggested a slightly hydrophobic protein [94], Mittal *et al.* recently postulated hydrophilic properties [204]. This amphiphilic character might evolve from the attached PTMs. Although a conclusive experimental proof of the determined and attached modifications is still pending, the reliability of the applied prediction algorithms is between 75 and 93% [146-152,154-157]. Regarding to the central question of this work, if SP-G and SP-H are part of the surfactant protein family, the here obtained modification patterns represent a key feature of the analyzed proteins and the already known surfactant proteins. Moreover, MD simulations showed that the attached PTMs did not deteriorate the protein model quality or the model stability compared to the models without added PTMs. These modifications were even able to stabilize protein regions that would have shown high fluctuations without PTMs. In case of SP-H, the MD simulation of the model with PTMs even improved the model quality measured by ProSA II and ProQ (Appendix 8).

4.2. Findings by the performed molecular dynamics simulations

A typical feature of surfactant proteins is the ability to interact with lipids, as reported in previous studies especially for SP-B and SP-C [86,89,91,92]. Accordingly, the SP-G and SP-H models were simulated in the presence of a DPPC monolayer, which corresponds to the current understanding of the pulmonary surfactant layout. DPPC as major lipid component of the pulmonary surfactant [2,3] was already shown to adequately reproduce the surfactant system of the lung in various MD simulations [92,175-177]. In these studies, parameters and settings for MD simulations were extensively studied and could be adapted to the simulation system used in this work. Therefore, only minor changes in the G53a6 force field were necessary for the PTMs attached to the protein models. All calculations were performed at a temperature of 323 K, which is above the phase transition temperature of DPPC [200,201]. This ensured that the lipid system was present in the biologically relevant fluid L_α state found *in vivo* instead of the more ordered gel or subgel state of a lipid layer [202]. To estimate the influence of the higher temperature on the protein stability, MD simulations of all four final models were performed in a water box at 298 K in comparison to MD simulations for the protein structures at 323 K. The results showed only a minor increase of fluctuations in the RMSD plot for SP-G

and no significant changes in stability for SP-H (cf. Figure 12). The lipid layer system for this work was built from scratch to obtain a lipid layer patch with the appropriate dimensions for the protein sizes, and in addition, to ensure the usage of correct simulation setup and parameters. During the lipid layer preparation, literature values for comparable systems were reproduced successfully (cf. Table 4).

All final models for SP-G and SP-H, with and without PTMs, were used to perform overall 24 MD simulations in the generated lipid environment (Figure 4). These simulations mostly showed the stability of the protein model fold in the RMSD plots (Appendix 11, Appendix 12) and demonstrated the influence of the PTMs on the physicochemical properties of the proteins via conformational changes during the MD simulations. SP-G and SP-H might be amphiphilic proteins, which are able to show a hydrophobic as well as a hydrophilic character. A dynamic process could manage this switch between both properties, which was suggested by the MD simulations of the models containing PTMs. Depending on type and conformation of the posttranslational modifications, polar or hydrophobic areas on the protein surface with a significant impact on water solubility or the protein-lipid interaction potential could be formed. When the protein is not residing in proximity to a lipid system, the palmitoylations for example could be embedded into the hydrophobic protein core, which would significantly change the protein surface properties. In case of SP-G, the hydrophobic *N*-terminal signal peptide is an important factor, since it could protrude from the protein surface or could be tightly bound to it. Not only is the shape of the protein probably altered in this way, but also the positions of hydrophobic spots on the protein surface. Furthermore, the PTMs showed an influence on the secondary structure stability. On the one hand, bulky modifications, such as *N*- or *O*-glycosylations, could introduce flexibility to the connected protein regions due to the rapid change of hydrogen bonding partners (i.e. water molecules) in free solution. On the other hand, these bulky modifications could significantly stabilize a protein region by forming mostly hydrogen bonds with the polar head groups of DPPC molecules. These described options demonstrate the influence of the PTMs on the stability as well as the interaction potential of both proteins.

The special cases of MD simulations with the modified protein models positioned into the lipid layer at simulation start (“positioned”) support these findings. In both simulations, the models were manually positioned in a way that the straightened palmitoyl moieties were sticking into the hydrophobic region of the lipid tails. This simulation setup was inspired by simulation studies from the literature, which showed different roles of posttranslational palmitoylation in

protein-lipid interaction [205] or the ability of lipid modifications to regulate the protein activity [206]. Since it is not assumed that SP-G and SP-H possess any transmembrane elements, the positioning of the palmitoylations in the hydrocarbon tail regions led to a very close proximity between the remaining protein structure and the polar lipid head groups (cf. Figure 17, Figure 23). Although this positioning was arbitrary, the protein models remained stable concerning their structure and position on the lipid surface during the simulations. Unfortunately, the contact between palmitoylations and lipid tails diminished during the simulation. For SP-G, the palmitoyl group left the hydrocarbon chain region, avoiding the polar environment of the lipid head groups by interacting with hydrophobic amino acids of the protein core (cf. Figure 17). After MD simulation of the pre-positioned SP-H model with PTMs, the palmitoylation was still spanning the head group region; however, interactions were only visible for the uppermost methyl groups of the lipid tails. The reason for this observation could be simplifications of the lipid layer model, which were necessary in this study. Due to the simplifications, the model system may miss a component that is essential to maintain the interactions. For example, lipids with short or unsaturated chains could enhance the layer fluidity and support the integration of the palmitoyl chain into the layer. A further possible reason could be a general underestimation of hydrophobic interaction energy in calculations using empirical force fields [191-193], which could prevent the palmitoyl chain from finding an energetically favored position. Another cause for the loss of palmitoyl-lipid interactions could be the polar PTMs. The formation of strong interactions between these polar PTMs and the lipid head groups could prevent the protein from penetrating deeper into the lipid layer. For the observed orientations, the palmitoylations are very flexible and the introduced “kinks” might lead to a too short hydrocarbon chain that cannot reach the lipid tails. Although a conclusive function of the hydrophobic palmitoylations was not observable, the importance of the polar glycosylations and phosphorylations for the stability and interaction potential of SP-G and SP-H was demonstrated in this study. These PTMs fixed the protein models like anchors on the lipid surface during the simulations. In general, the results suggested that both proteins are not completely integrated into the lipid layer, as reported for SP-B and SP-C. Instead, SP-G and SP-H seem to be layer-associated and remain on the surface of the lipids, where the numerous PTMs interact with the polar lipid head groups.

In the MDs without manual protein model positioning, most of the interactions were also established between polar amino acid side chains or PTMs and the polar head groups of the lipid molecules. Almost no contact of protein parts with the hydrophobic lipid tail region was observed. The results of the simulations showed no direct impact of the protein-lipid interaction

on the layer stability or lipid ordering. The literature suggests that longer simulations in the microsecond range might be required to observe protein-mediated events, such as lipid layer folding or lipid vesicle fusion [89,92]. Such long simulations would be computationally too expensive for the united-atom approach used in this study. So called coarse-grained simulations [207,208] with reduced complexity developed especially for such long-term simulations would be the technique of choice for future experiments. For this, the knowledge about the 3D protein structure is very important and a required input, because currently, the most commonly used MARTINI coarse-grained force field is unable to consider changes in the secondary structure of a protein during the simulation [208,209]. Thus, the secondary structure assigned to each amino acid at the simulation start remains unchanged throughout the simulation. However, the simulation results of this work demonstrated the stability of the SP-G and SP-H protein fold in most cases, even during the formation of interactions between protein and lipid layer. Therefore, the here performed calculations provide the requirements for coarse-grained simulations.

Although at the simulation start, the protein models were located in a distance between 1.5 and 3.5 nm to the lipid layer, all models started to interact with the lipids mostly within the first 25 ns of the 50 ns MD simulations. In some cases, the first interactions were already observed after less than 5 ns. This process was traceable by monitoring the protein-lipid interaction energy (Appendix 9, Appendix 10). In this way, it was possible to discriminate between different interaction scenarios and to visualize the influence of polar amino acids and PTMs on the interaction strength. However, the energies calculated based on force field parameters can only give a rough estimation of the *in vivo* energies, since the accuracy of force fields reproducing intermolecular (i.e. non-bonded) interaction energies is limited [190-193]. For more detailed insights, more specialized computational chemistry techniques, such as free energy calculations [210,211] or the experimental isothermal titration calorimetry (ITC [212]) would be advantageous. Nevertheless, the fact that all 24 performed simulations showed an interaction between the protein models and the lipid layer is a major result of this work. Thereby, the hypothesis that SP-G and SP-H are indeed able to interact with lipids and may exhibit surface-regulatory properties is strongly supported.

4.3. Cooperation of computational and experimental studies

At the beginning of this work, the classification of SP-G and SP-H as surfactant proteins was solely based on hypothetical predictions. There was no experimental evidence or verification of this classification available. With the help of the computational modeling and simulation techniques applied in this work, however, first indications could be observed that confirm the membership of SP-G and SP-H to the surfactant protein family. As results of the studies performed in this thesis, the existing lipid interaction potential, the high degree of posttranslational modification and with that, the possible amphiphilic character of SP-G and SP-H were consistent with the classification as surfactant proteins. At this point, the localization of both proteins in tissues that are typical for the presence of surfactant proteins would have been a further hint for the function of these putative surfactant proteins. Unfortunately, no commercial anti-SP-G or anti-SP-H antibody was available, and previous attempts to produce specific antibodies for localization studies failed. However, with the here obtained knowledge of the 3D structure of both proteins and the potential modification pattern, it was possible to identify PTM-free sequence regions on the protein surface. Their use as antigen peptides led to specific antibodies for SP-G and SP-H. The successful production of these antibodies on the one hand indicated a high reliability of the protein models and on the other hand allowed first localization and functional studies.

Human lung tissue was used to test the antibody specificity, because the already known surfactant proteins were initially described in the lung [11,12,21-23]. The corresponding Western blot analysis (M. Schicht [198]) for SP-G showed specific bands at 11 kDa, 20 kDa, and 30 kDa. All three values deviated from the molecular weight of 9 kDa, which was calculated based on the amino acid sequence. However, considering that the protein might be posttranslationally modified with glycosyl, phosphate, and palmitoyl moieties, the distinct protein band at 11 kDa seems to represent the mature protein. Based on the modeling results, the formation of oligomers cannot be excluded. Therefore, the two higher molecular weights (20 kDa and 30 kDa) might represent homodimers or homotrimers of SP-G.

For SP-H, the Western blot analysis showed three bands at 13 kDa, 26 kDa, and 43 kDa (M. Schicht [199]). Again, none of the three values matched the calculated molecular weight of 10 kDa for SP-H. Analog to SP-G, the correction of the molecular weight with the predicted

posttranslational modifications (PTMs) suggests the band with 13 kDa to be the mature monomer of SP-H and the band at 26 kDa the mature homodimer. Whether the band at 43 kDa represents a homotrimer of SP-H with an altered PTM pattern or a complex of SP-H with another protein remains speculative. Nevertheless, due to the high PTM rate of these proteins, a broad range of observed molecular weights is very common for small surfactant proteins. The protein sizes for SP-B vary from 8 kDa and 25 kDa in the lung up to 35 kDa in tissues of the ocular system [24]. This observation accounts for SP-C as well, which shows molecular weight differences in the range from 7 kDa [213], 21 kDa [25] up to 26 kDa [214].

For a better understanding of the wide range of molecular weights obtained from Western blot experiments, knowledge about the assembly of monomers to multimers would be advantageous. Therefore, several protein-protein docking algorithms exist to obtain a homodimer structure [215]. Unfortunately, the currently available methods were not able to build reliable complexes for SP-G or SP-H. In most cases, the obtained docking conformations were not reproducible. Furthermore, the majority of algorithms ignored amino acids with PTMs, resulting in broken protein structures.

For our cooperation partners (Institute of Anatomy II, FAU Erlangen), the SP-G and SP-H antibodies obtained by the help of protein modeling were crucial for immunohistochemical experiments. The first protein localization studies for these proteins via immunohistochemical staining showed that SP-G [198] and SP-H [199] are present in tissues of the human lung and eye, among many others. In these tissues, the already known surfactant proteins are also present and play a crucial role [10,11,13,14,21,24,25,37]. Within lung tissue, the distribution of SP-G as a superficial layer of the epithelium of the bronchioles was demonstrated, which additionally indicated surface activity of the protein [198]. It could be shown that SP-H is present in alveolar cell macrophages and in the cytosol of epithelial cells [199]. These results are in line with the occurrence of already known surfactant proteins [216,217]. Additionally, immunohistochemistry and immunogold electron microscopy experiments with A549 cells [218] suggested that SP-H is also secreted and presented on the cell membrane. This fact indicates that SP-H remains cytoplasmic because of its physicochemical properties and that it is secreted after modification. In the here presented theoretical studies, SP-H showed a palmitoylation potential, which would allow SP-H to interact with a lipid membrane similar to SP-B or SP-C [219].

Finally, the antibodies enabled first functional studies which showed that inflammatory cytokines influence the SP-H expression level [199]. This could indicate an immunoregulatory function of SP-H comparable to SP-A and SP-D [10,14]. However, a role of SP-G in

inflammation and immunological defense is speculative, because immune regulatory domains have not been identified yet, neither with the applied computational methods nor with the performed molecular-biological methods.

5. Summary

With the help of *ab initio* protein structure prediction it was possible to obtain 3D models for the two putative surfactant proteins SP-G (SFTA2) and SP-H (SFTA3). Common quality assessment tools indicated a native-like folding of the protein models, and subsequent molecular dynamics simulations demonstrated the stability of the fold of these SP-G and SP-H models. The models were extended by posttranslational modifications (PTMs), because the high importance of PTMs for the function of the already known surfactant proteins was described in the literature. Sequence-based prediction tools indicated numerous phosphorylations, glycosylations, and palmitoylations for SP-G and SP-H, which were manually added to the protein models and did not influence the overall model stability in molecular dynamics simulations.

Previous attempts to obtain specific antibodies for SP-G and SP-H failed due to the lack of knowledge about the 3D protein structure. The models obtained in this work revealed sequence parts on the surface of the proteins without any PTM, which were suitable antigens for the production of specific antibodies. Therefore, computational modeling significantly supported the experimental work, because the obtained antibodies allowed the first localization of SP-G and SP-H in different cell tissues where the already known surfactant proteins are present as well. Furthermore, they could be used in first functional studies.

To mimic the basic properties of the pulmonary surfactant, a simulation system containing a DPPC lipid monolayer was established. This system was used to study the characteristics of the SP-G and SP-H models in their natural environment, each with and without PTMs. Overall, 24 MD simulations of 50 ns each were performed. Although the strength of the interactions and contact areas on the protein surface were dependent on the starting structure and attached PTMs, all performed simulations indicated a high probability of protein-lipid interactions. Furthermore, the calculation results suggested that the positions and conformations of PTMs could be responsible for an amphiphilic character of SP-G and SP-H, which was described for the already known surfactant proteins as well. The high theoretical lipid interaction potential determined by the presented simulations could be used to support and discuss the outcome of experimental characterization and localization studies, which suggest that SP-G and SP-H are indeed part of the surfactant protein family.

6. Prospective research suggestions

Although the results of this work represent completely new and valuable insights into the structure and characterization of SP-G and SP-H, they provide numerous interesting starting points for future research projects. For example, the simulation system established in this work to resemble the pulmonary surfactant system is very basic. It contains only one lipid species and no other compounds, although the natural pulmonary surfactant composition is much more complex. For further research, the here presented DPPC layer system could represent the basis for the development of a more sophisticated pulmonary surfactant model, which might consist of a huge variety of compounds shown to be present *in vivo*: phosphatidylcholines with different carbon chain length or saturation levels, phosphatidylinositol, phosphatidylglycerol, cholesterol or even other proteins [2,3]. Such a detailed simulation system could increase the reliability of performed MD studies and may give a more comprehensive impression of the characteristics and functions of the investigated proteins. Apart from SP-G and SP-H, the introduction of other proteins into the simulation system opens the wide field of protein-protein interaction analysis. These simulation types could show the interactions between proteins on the atomic scale or even demonstrate what effect the cooperation of two different proteins could have, as previously described in the literature for SP-B and SP-C [92].

The simulation studies performed in this thesis covered the time scale of tens of nanoseconds. This is sufficient to demonstrate the model stability and establish the protein-lipid contact by atomic interactions. However, there may be other events of the protein or protein-lipid interaction, which need a much longer time scale to take place. For this purpose, the coarse-grained simulation technique could be used, which allows the simulation of a system up to the microsecond scale [207,208]. Therefore, the degrees of freedom in the system and necessary calculation efforts are reduced by grouping nearby atoms with similar attributes or functional groups into one “bead”. Drawbacks of this method are the loss of precision due to the aggregation of atoms and the fact that the secondary structure of a protein is not allowed to change during such a simulation. However, the results presented in this work can already present detailed interaction information and show the stability of the secondary structure of the SP-G and SP-H models. With these assumptions, coarse-grained simulations could be used in

future studies to perform simulations on a long time scale (several microseconds) to observe even very complex processes, such as the reordering of single lipids or a whole lipid layer. Moreover, the cooperation of two or more proteins could be investigated. A recent study impressively demonstrated the benefits of combining all-atom and coarse-grained methods to get insights into protein-lipid interactions [220].

Since SP-G and SP-H were found in the tear film and tissues of the ocular surface [24,25] as well, the knowledge transfer from the presented simulations in a pulmonary surfactant system to a tear film model system would be very interesting. Although the functions of surfactant proteins (lipid organization, lipid layer stability, immunological functions) are equally essential in both systems, the layout and composition of the air-water interface differ very much in the pulmonary surfactant and in the tear film. Accordingly, a completely new simulation system would be necessary, because the tear film structure is much more complicated than the pulmonary surfactant system. Establishing a tear film model system would include the parameterization of predominantly nonpolar lipid species for a force field. Moreover, the development of a stable multilayered system consisting of glycocalyx, aqueous, amphiphilic, and nonpolar layer [221] would be necessary. Finally, this complex tear film system would allow a detailed investigation of the whole transition process of SP-G and SP-H from an aqueous, through an amphiphilic, to a nonpolar environment, which includes the conformational and surface potential changes suggested in this work

7. Literature

1. Akella A, Deshpande SB (2013) Pulmonary surfactants and their role in pathophysiology of lung disorders. Indian J Exp Biol 51: 5-22.

2. Goerke J (1998) Pulmonary surfactant: functions and molecular composition. Biochim Biophys Acta 1408: 79-89.

3. Veldhuizen R, Nag K, Orgeig S, Possmayer F (1998) The role of lipids in pulmonary surfactant. Biochim Biophys Acta 1408: 90-108.

4. Notter RH (2000) Lung Surfactants - Basic Science and Clinical Applications: Marcel Dekker, Inc.

5. Creuwels LA, van Golde LM, Haagsman HP (1997) The pulmonary surfactant system: biochemical and clinical aspects. Lung 175: 1-39.

6. Kahn MC, Anderson GJ, Anyan WR, Hall SB (1995) Phosphatidylcholine molecular species of calf lung surfactant. Am J Physiol 269: L567-573.

7. Wright JR, Clements JA (1987) Metabolism and turnover of lung surfactant. Am Rev Respir Dis 136: 426-444.

8. Griese M (1999) Pulmonary surfactant in health and human lung diseases: state of the art. Eur Respir J 13: 1455-1476.

9. Spragg RG, Gilliard N, Richman P, Smith RM, Hite RD, *et al.* (1994) Acute effects of a single dose of porcine surfactant on patients with the adult respiratory distress syndrome. Chest 105: 195-202.

10. Wright JR (2005) Immunoregulatory functions of surfactant proteins. Nat Rev Immunol 5: 58-68.

11. Yu SH, Possmayer F (1990) Role of bovine pulmonary surfactant-associated proteins in the surface-active property of phospholipid mixtures. Biochim Biophys Acta 1046: 233-241.

12. Crouch E, Wright JR (2001) Surfactant proteins a and d and pulmonary host defense. Annu Rev Physiol 63: 521-554.

13. Hartshorn KL, Crouch E, White MR, Colamussi ML, Kakkanatt A, *et al.* (1998) Pulmonary surfactant proteins A and D enhance neutrophil uptake of bacteria. Am J Physiol 274: L958-969.

14. Kishore U, Greenhough TJ, Waters P, Shrive AK, Ghai R, *et al.* (2006) Surfactant proteins SP-A and SP-D: structure, function and receptors. Mol Immunol 43: 1293-1315.

15. Ding J, Takamoto DY, von Nahmen A, Lipp MM, Lee KY, *et al.* (2001) Effects of lung surfactant proteins, SP-B and SP-C, and palmitic acid on monolayer stability. Biophys J 80: 2262-2272.

16. Schurch D, Ospina OL, Cruz A, Perez-Gil J (2010) Combined and independent action of proteins SP-B and SP-C in the surface behavior and mechanical stability of pulmonary surfactant films. Biophys J 99: 3290-3299.

17. Glasser SW, Korfhagen TR, Perme CM, Pilot-Matias TJ, Kister SE, *et al.* (1988) Two SP-C genes encoding human pulmonary surfactant proteolipid. J Biol Chem 263: 10326-10331.

18. Voorhout WF, Veenendaal T, Haagsman HP, Weaver TE, Whitsett JA, *et al.* (1992) Intracellular processing of pulmonary surfactant protein B in an endosomal/lysosomal compartment. Am J Physiol 263: 479-486.

19. Veldhuizen EJ, Batenburg JJ, van Golde LM, Haagsman HP (2000) The role of surfactant proteins in DPPC enrichment of surface films. Biophys J 79: 3164-3171.

20. Li J, Ikegami M, Na CL, Hamvas A, Espinassous Q, *et al.* (2004) N-terminally extended surfactant protein (SP) C isolated from SP-B-deficient children has reduced surface activity and inhibited lipopolysaccharide binding. Biochemistry 43: 3891-3898.

21. Notter RH, Shapiro DL, Ohning B, Whitsett JA (1987) Biophysical activity of synthetic phospholipids combined with purified lung surfactant 6000 dalton apoprotein. Chem Phys Lipids 44: 1-17.

22. King RJ, Simon D, Horowitz PM (1989) Aspects of secondary and quaternary structure of surfactant protein A from canine lung. Biochim Biophys Acta 1001: 294-301.

23. van Iwaarden F, Welmers B, Verhoef J, Haagsman HP, van Golde LM (1990) Pulmonary surfactant protein A enhances the host-defense mechanism of rat alveolar macrophages. Am J Respir Cell Mol Biol 2: 91-98.

24. Bräuer L, Johl M, Borgermann J, Pleyer U, Tsokos M, *et al.* (2007) Detection and localization of the hydrophobic surfactant proteins B and C in human tear fluid and the human lacrimal system. Curr Eye Res 32: 931-938.

25. Bräuer L, Kindler C, Jager K, Sel S, Nolle B, *et al.* (2007) Detection of surfactant proteins A and D in human tear fluid and the human lacrimal system. Invest Ophthalmol Vis Sci 48: 3945-3953.

26. White RT, Damm D, Miller J, Spratt K, Schilling J, *et al.* (1985) Isolation and characterization of the human pulmonary surfactant apoprotein gene. Nature 317: 361-363.

27. Day AJ (1994) The C-type carbohydrate recognition domain (CRD) superfamily. Biochem Soc Trans 22: 83-88.

28. Suzuki Y, Fujita Y, Kogishi K (1989) Reconstitution of tubular myelin from synthetic lipids and proteins associated with pig pulmonary surfactant. Am Rev Respir Dis 140: 75-81.

29. Kobayashi T, Nitta K, Takahashi R, Kurashima K, Robertson B, *et al.* (1991) Activity of pulmonary surfactant after blocking the associated proteins SP-A and SP-B. J Appl Physiol 71: 530-536.

30. Korfhagen TR, Bruno MD, Ross GF, Huelsman KM, Ikegami M, *et al.* (1996) Altered surfactant function and structure in SP-A gene targeted mice. Proc Natl Acad Sci USA 93: 9594-9599.

31. Wright JR, Youmans DC (1993) Pulmonary surfactant protein A stimulates chemotaxis of alveolar macrophage. Am J Physiol 264: L338-344.

32. Williams MD, Wright JR, March KL, Martin WJ, 2nd (1996) Human surfactant protein A enhances attachment of Pneumocystis carinii to rat alveolar macrophages. Am J Respir Cell Mol Biol 14: 232-238.

33. Kabha K, Schmegner J, Keisari Y, Parolis H, Schlepper-Schaeffer J, *et al.* (1997) SP-A enhances phagocytosis of Klebsiella by interaction with capsular polysaccharides and alveolar macrophages. Am J Physiol 272: L344-352.

34. Madan T, Kishore U, Shah A, Eggleton P, Strong P, *et al.* (1997) Lung surfactant proteins A and D can inhibit specific IgE binding to the allergens of Aspergillus fumigatus and block allergen-induced histamine release from human basophils. Am J Clin Exp Immunol 110: 241-249.

35. Benson HL, Mobashery S, Chang M, Kheradmand F, Hong JS, *et al.* (2011) Endogenous matrix metalloproteinases 2 and 9 regulate activation of CD4+ and CD8+ T cells. Am J Respir Cell Mol Biol 44: 700-708.

36. Sastry K, Ezekowitz RA (1993) Collectins: pattern recognition molecules involved in first line host defense. Curr Opin Immunol 5: 59-66.

37. Ferguson JS, Voelker DR, McCormack FX, Schlesinger LS (1999) Surfactant protein D binds to Mycobacterium tuberculosis bacilli and lipoarabinomannan via carbohydrate-lectin interactions resulting in reduced phagocytosis of the bacteria by macrophages. J Immunol 163: 312-321.

38. LeVine AM, Whitsett JA, Hartshorn KL, Crouch EC, Korfhagen TR (2001) Surfactant protein D enhances clearance of influenza A virus from the lung in vivo. J Immunol 167: 5868-5873.

39. Botas C, Poulain F, Akiyama J, Brown C, Allen L, *et al.* (1998) Altered surfactant homeostasis and alveolar type II cell morphology in mice lacking surfactant protein D. Proc Natl Acad Sci USA 95: 11869-11874.

40. Hawgood S, Derrick M, Poulain F (1998) Structure and properties of surfactant protein B. Biochim Biophys Acta 1408: 150-160.

41. Whitsett JA, Ohning BL, Ross G, Meuth J, Weaver T, *et al.* (1986) Hydrophobic surfactant-associated protein in whole lung surfactant and its importance for biophysical activity in lung surfactant extracts used for replacement therapy. J Pediatric Res 20: 460-467.

42. Oosterlaken-Dijksterhuis MA, Haagsman HP, van Golde LM, Demel RA (1991) Characterization of lipid insertion into monomolecular layers mediated by lung surfactant proteins SP-B and SP-C. Biochemistry 30: 10965-10971.

43. Clark JC, Wert SE, Bachurski CJ, Stahlman MT, Stripp BR, *et al.* (1995) Targeted disruption of the surfactant protein B gene disrupts surfactant homeostasis, causing respiratory failure in newborn mice. Proc Natl Acad Sci USA 92: 7794-7798.

44. Nogee LM, Garnier G, Dietz HC, Singer L, Murphy AM, *et al.* (1994) A mutation in the surfactant protein B gene responsible for fatal neonatal respiratory disease in multiple kindreds. J Clin Invest 93: 1860-1863.

45. Yang L, Johansson J, Ridsdale R, Willander H, Fitzen M, *et al.* (2010) Surfactant protein B propeptide contains a saposin-like protein domain with antimicrobial activity at low pH. J Immunol 184: 975-983.

46. Waring AJ, Walther FJ, Gordon LM, Hernandez-Juviel JM, Hong T, *et al.* (2005) The role of charged amphipathic helices in the structure and function of surfactant protein B. J Pept Res 66: 364-374.

47. Walther FJ, Waring AJ, Hernandez-Juviel JM, Gordon LM, Wang Z, *et al.* (2010) Critical structural and functional roles for the N-terminal insertion sequence in surfactant protein B analogs. PLoS One 5: e8672.

48. Sarker M, Waring AJ, Walther FJ, Keough KM, Booth V (2007) Structure of mini-B, a functional fragment of surfactant protein B, in detergent micelles. Biochemistry 46: 11047-11056.

49. Palleboina D, Waring AJ, Notter RH, Booth V, Morrow M (2012) Effects of the lung surfactant protein B construct Mini-B on lipid bilayer order and topography. Eur Biophys J 41: 755-767.

50. Weaver TE, Conkright JJ (2001) Function of surfactant proteins B and C. Annu Rev Physiol 63: 555-578.

51. Beers MF, Fisher AB (1992) Surfactant protein C: a review of its unique properties and metabolism. Am J Physiol 263: L151-160.

52. Vandenbussche G, Clercx A, Curstedt T, Johansson J, Jornvall H, *et al.* (1992) Structure and orientation of the surfactant-associated protein C in a lipid bilayer. Eur J Biochem 203: 201-209.

53. Johansson J (1998) Structure and properties of surfactant protein C. Biochim Biophys Acta 1408: 161-172.

54. Curstedt T, Johansson J, Persson P, Eklund A, Robertson B, *et al.* (1990) Hydrophobic surfactant-associated polypeptides: SP-C is a lipopeptide with two palmitoylated cysteine residues, whereas SP-B lacks covalently linked fatty acyl groups. Proc Natl Acad Sci USA 87: 2985-2989.

55. Stults JT, Griffin PR, Lesikar DD, Naidu A, Moffat B, *et al.* (1991) Lung surfactant protein SP-C from human, bovine, and canine sources contains palmityl cysteine thioester linkages. Am J Physiol 261: L118-125.

56. Vorbroker DK, Dey C, Weaver TE, Whitsett JA (1992) Surfactant protein C precursor is palmitoylated and associates with subcellular membranes. Biochim Biophys Acta 1105: 161-169.

57. Creuwels LA, Demel RA, van Golde LM, Benson BJ, Haagsman HP (1993) Effect of acylation on structure and function of surfactant protein C at the air-liquid interface. J Biol Chem 268: 26752-26758.

58. Qanbar R, Possmayer F (1995) On the surface activity of surfactant-associated protein C (SP-C): effects of palmitoylation and pH. Biochim Biophys Acta 1255: 251-259.

59. Jacobs KA, Phelps DS, Steinbrink R, Fisch J, Kriz R, *et al.* (1987) Isolation of a cDNA clone encoding a high molecular weight precursor to a 6-kDa pulmonary surfactant-associated protein. J Biol Chem 262: 9808-9811.

60. Glasser SW, Burhans MS, Korfhagen TR, Na CL, Sly PD, *et al.* (2001) Altered stability of pulmonary surfactant in SP-C-deficient mice. Proc Natl Acad Sci USA 98: 6366-6371.

61. Glasser SW, Detmer EA, Ikegami M, Na CL, Stahlman MT, *et al.* (2003) Pneumonitis and emphysema in sp-C gene targeted mice. J Biol Chem 278: 14291-14298.

62. Oosterlaken-Dijksterhuis MA, Haagsman HP, van Golde LM, Demel RA (1991) Interaction of lipid vesicles with monomolecular layers containing lung surfactant proteins SP-B or SP-C. Biochemistry 30: 8276-8281.

63. Augusto L, Le Blay K, Auger G, Blanot D, Chaby R (2001) Interaction of bacterial lipopolysaccharide with mouse surfactant protein C inserted into lipid vesicles. Am J Physiol Lung Cell Mol Physiol 281: L776-785.

64. Whitsett JA, Weaver TE (2002) Hydrophobic surfactant proteins in lung function and disease. N Engl J Med 347: 2141-2148.

65. Seddon AM, Curnow P, Booth PJ (2004) Membrane proteins, lipids and detergents: not just a soap opera. Biochim Biophys Acta 1666: 105-117.

66. Johansson J, Szyperski T, Curstedt T, Wuthrich K (1994) The NMR structure of the pulmonary surfactant-associated polypeptide SP-C in an apolar solvent contains a valyl-rich alpha-helix. Biochemistry 33: 6015-6023.

67. Bernstein FC, Koetzle TF, Williams GJ, Meyer EF, Jr., Brice MD, *et al.* (1977) The Protein Data Bank: a computer-based archival file for macromolecular structures. J Mol Biol 112: 535-542.

68. Kurutz JW, Lee KY (2002) NMR structure of lung surfactant peptide SP-B(11-25). Biochemistry 41: 9627-9636.

69. Gordon LM, Lee KY, Lipp MM, Zasadzinski JA, Walther FJ, *et al.* (2000) Conformational mapping of the N-terminal segment of surfactant protein B in lipid using 13C-enhanced Fourier transform infrared spectroscopy. J Pept Res 55: 330-347.

70. Booth V, Waring AJ, Walther FJ, Keough KM (2004) NMR structures of the C-terminal segment of surfactant protein B in detergent micelles and hexafluoro-2-propanol. Biochemistry 43: 15187-15194.

71. Head JF, Mealy TR, McCormack FX, Seaton BA (2003) Crystal structure of trimeric carbohydrate recognition and neck domains of surfactant protein A. J Biol Chem 278: 43254-43260.

72. Shrive AK, Tharia HA, Strong P, Kishore U, Burns I, *et al.* (2003) High-resolution structural insights into ligand binding and immune cell recognition by human lung surfactant protein D. J Mol Biol 331: 509-523.

73. Kaleem A, Hoessli DC, Haq IU, Walker-Nasir E, Butt A, *et al.* (2011) CREB in long-term potentiation in hippocampus: role of post-translational modifications-studies In silico. J Cell Biochem 112: 138-146.

74. Weaver TE (1998) Synthesis, processing and secretion of surfactant proteins B and C. Biochim Biophys Acta 1408: 173-179.

75. Guttentag S, Robinson L, Zhang P, Brasch F, Buhling F, *et al*. (2003) Cysteine protease activity is required for surfactant protein B processing and lamellar body genesis. Am J Respir Cell Mol Biol 28: 69-79.

76. Beers MF, Mulugeta S (2005) Surfactant protein C biosynthesis and its emerging role in conformational lung disease. Annu Rev Physiol 67: 663-696.

77. Zhang JL, Zheng QC, Zhang HX (2010) Unbinding of glucose from human pulmonary surfactant protein D studied by steered molecular dynamics simulations. Chem Phy Lett 484: 338-343.

78. van Eijk M, Rynkiewicz MJ, White MR, Hartshorn KL, Zou X, *et al.* (2012) A unique sugar-binding site mediates the distinct anti-influenza activity of pig surfactant protein D. J Biol Chem 287: 26666-26677.

79. Allen MJ, Laederach A, Reilly PJ, Mason RJ, Voelker DR (2004) Arg343 in human surfactant protein D governs discrimination between glucose and N-acetylglucosamine ligands. Glycobiology 14: 693-700.

80. Zhang J, Zheng Q, Zhang H (2010) Insight into the dynamic interaction of different carbohydrates with human surfactant protein D: molecular dynamics simulations.J Phys Chem B 114: 7383-7390.

81. Lee H, Kandasamy SK, Larson RG (2005) Molecular dynamics simulations of the anchoring and tilting of the lung-surfactant peptide SP-B1-25 in palmitic acid monolayers. Biophys J 89: 3807-3821.

82. Bertani P, Vidovic V, Yang TC, Rendell J, Gordon LM, *et al.* (2012) Orientation and depth of surfactant protein B C-terminal helix in lung surfactant bilayers. Biochim Biophys Acta 1818: 1165-1172.

83. Kandasamy SK, Larson RG (2005) Molecular dynamics study of the lung surfactant peptide SP-B1-25 with DPPC monolayers: insights into interactions and peptide position and orientation. Biophys J 88: 1577-1592.

84. Kim HI, Kim H, Shin YS, Beegle LW, Jang SS, *et al.* (2010) Interfacial reactions of ozone with surfactant protein B in a model lung surfactant system. J Am Chem Soc 132: 2254-2263.

85. Kaznessis YN, Kim S, Larson RG (2002) Specific mode of interaction between components of model pulmonary surfactants using computer simulations. J Mol Biol 322: 569-582.

86. Freites JA, Choi Y, Tobias DJ (2003) Molecular dynamics simulations of a pulmonary surfactant protein B peptide in a lipid monolayer. Biophys J 84: 2169-2180.

87. Javanainen M, Monticelli L, Bernardino de la Serna J, Vattulainen I (2010) Free volume theory applied to lateral diffusion in Langmuir monolayers: atomistic simulations for a protein-free model of lung surfactant. Langmuir 26: 15436-15444.

88. Baoukina S, Tieleman DP (2011) Lung surfactant protein SP-B promotes formation of bilayer reservoirs from monolayer and lipid transfer between the interface and subphase. Biophys J 100: 1678-1687.

89. Baoukina S, Tieleman DP (2010) Direct simulation of protein-mediated vesicle fusion: lung surfactant protein B. Biophys J 99: 2134-2142.

90. Kovacs H, Mark AE, Johansson J, van Gunsteren WF (1995) The effect of environment on the stability of an integral membrane helix: molecular dynamics simulations of surfactant protein C in chloroform, methanol and water. J Mol Biol 247: 808-822.

91. Baoukina S, Monticelli L, Amrein M, Tieleman DP (2007) The molecular mechanism of monolayer-bilayer transformations of lung surfactant from molecular dynamics simulations. Biophys J 93: 3775-3782.

92. Duncan SL, Larson RG (2010) Folding of lipid monolayers containing lung surfactant proteins SP-B(1-25) and SP-C studied via coarse-grained molecular dynamics simulations. Biochim Biophys Acta 1798: 1632-1650.

93. Consortium U (2012) Reorganizing the protein space at the Universal Protein Resource (UniProt). Nucleic Acids Res 40: D71-75.

94. Heilig R, Eckenberg R, Petit JL, Fonknechten N, Da Silva C, *et al.* (2003) The DNA sequence and analysis of human chromosome 14. Nature 421: 601-607.

95. Mungall AJ, Palmer SA, Sims SK, Edwards CA, Ashurst JL, *et al.* (2003) The DNA sequence and analysis of human chromosome 6. Nature 425: 805-811.

96. Zhang Z, Henzel WJ (2004) Signal peptide prediction based on analysis of experimentally verified cleavage sites. Protein Sci 13: 2819-2824.

97. Nakai K (2000) Protein sorting signals and prediction of subcellular localization. Adv Protein Chem 54: 277-344.

98. Matthews BW (1976) X-Ray Crystallographic Studies of Proteins. Annu Rev Phys Chem 27: 493-523.

99. Morris GA (1986) Modern Nmr Techniques for Structure Elucidation. Magn Reson Chem 24: 371-403.

100. Baker D, Sali A (2001) Protein structure prediction and structural genomics. Science 294: 93-96.

101. Kaczanowski S, Zielenkiewicz P (2010) Why similar protein sequences encode similar three-dimensional structures? Theor Chem Acc 125: 643-650.

102. Chothia C, Lesk AM (1986) The relation between the divergence of sequence and structure in proteins. EMBO J 5: 823-826.

103. Rost B (1999) Twilight zone of protein sequence alignments. Protein Eng 12: 85-94.

104. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. J Mol Biol 215: 403-410.

105. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, *et al.* (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25: 3389-3402.

106. Krieger E, Koraimann G, Vriend G (2002) Increasing the precision of comparative models with YASARA NOVA--a self-parameterizing force field. Proteins 47: 393-402.

107. Muckstein U, Hofacker IL, Stadler PF (2002) Stochastic pairwise alignments. Bioinformatics 18: S153-S160.

108. Qiu J, Elber R (2006) SSALN: an alignment algorithm using structure-dependent substitution matrices and gap penalties learned from structurally aligned protein pairs. Proteins 62: 881-891.

109. Jones DT (1999) Protein secondary structure prediction based on position-specific scoring matrices. J Mol Biol 292: 195-202.

110. Krieger E, Darden T, Nabuurs SB, Finkelstein A, Vriend G (2004) Making optimal use of empirical energy functions: force-field parameterization in crystal space. Proteins 57: 678-683.

111. Krieger E, Joo K, Lee J, Lee J, Raman S, *et al.* (2009) Improving physical realism, stereochemistry, and side-chain accuracy in homology modeling: Four approaches that performed well in CASP8. Proteins 77 Suppl 9: 114-122.

112. Zhang Y (2008) I-TASSER server for protein 3D structure prediction. BMC Bioinformatics 9: 40.

113. Roy A, Kucukural A, Zhang Y (2010) I-TASSER: a unified platform for automated protein structure and function prediction. Nat Protoc 5: 725-738.

114. Moult J, Fidelis K, Kryshtafovych A, Rost B, Hubbard T, *et al.* (2007) Critical assessment of methods of protein structure prediction-Round VII. Proteins 69 Suppl 8: 3-9.

115. Moult J, Fidelis K, Kryshtafovych A, Rost B, Tramontano A (2009) Critical assessment of methods of protein structure prediction - Round VIII. Proteins 77 Suppl 9: 1-4.

116. Moult J, Fidelis K, Kryshtafovych A, Tramontano A (2011) Critical assessment of methods of protein structure prediction (CASP)--round IX. Proteins 79 Suppl 10: 1-5.

117. Moult J, Fidelis K, Kryshtafovych A, Schwede T, Tramontano A (2014) Critical assessment of methods of protein structure prediction (CASP)--round x. Proteins 82 Suppl 2: 1-6.

118. Lee J, Wu S, Zhang Y (2009) Ab Initio Protein Structure Prediction. In: Rigden D, editor. From Protein Structure to Function with Bioinformatics: Springer Netherlands. pp. 3-25.

119. Kim DE, Chivian D, Baker D (2004) Protein structure prediction and analysis using the Robetta server. Nucleic Acids Res 32: W526-531.

120. Chivian D, Kim DE, Malmstrom L, Bradley P, Robertson T, *et al.* (2003) Automated prediction of CASP-5 structures using the Robetta server. Proteins 53 Suppl 6: 524-533.

121. Laskowski RA, MacArthur DS, Moss DS, Thornton JM (1993) PROCHECK: a program to check the stereochemical quality of protein structures. J Appl Cryst 26: 283-291.

122. Morris AL, MacArthur MW, Hutchinson EG, Thornton JM (1992) Stereochemical quality of protein structure coordinates. Proteins 12: 345-364.

123. Sippl MJ (1993) Recognition of errors in three-dimensional structures of proteins. Proteins 17: 355-362.

124. Wallner B, Elofsson A (2003) Can correct protein models be identified? Protein Sci 12: 1073-1086.

125. Cristobal S, Zemla A, Fischer D, Rychlewski L, Elofsson A (2001) A study of quality measures for protein threading models. BMC Bioinformatics 2: 5.

126. Siew N, Elofsson A, Rychlewski L, Fischer D (2000) MaxSub: an automated measure for the assessment of protein structure prediction quality. Bioinformatics 16: 776-785.

127. Colovos C, Yeates TO (1993) Verification of protein structures: patterns of nonbonded atomic interactions. Protein Sci 2: 1511-1519.

128. Bowie JU, Luthy R, Eisenberg D (1991) A method to identify protein sequences that fold into a known three-dimensional structure. Science 253: 164-170.

129. Luthy R, Bowie JU, Eisenberg D (1992) Assessment of protein models with three-dimensional profiles. Nature 356: 83-85.

130. Minguez P, Parca L, Diella F, Mende DR, Kumar R, *et al.* (2012) Deciphering a global network of functionally associated post-translational modifications. Mol Syst Biol 8: 599.

131. Moremen KW, Tiemeyer M, Nairn AV (2012) Vertebrate protein glycosylation: diversity, synthesis and function. Nat Rev Mol Cell Biol 13: 448-462.

132. Ubersax JA, Ferrell JE, Jr. (2007) Mechanisms of specificity in protein phosphorylation. Nat Rev Mol Cell Biol 8: 530-541.

133. Pinna LA, Ruzzene M (1996) How do protein kinases recognize their substrates? Biochim Biophys Acta 1314: 191-225.

134. Cohen P (2000) The regulation of protein function by multisite phosphorylation--a 25 year update. Trends Biochem Sci 25: 596-601.

135. Glozak MA, Sengupta N, Zhang X, Seto E (2005) Acetylation and deacetylation of non-histone proteins. Gene 363: 15-23.

136. Hershko A, Heller H, Eytan E, Kaklij G, Rose IA (1984) Role of the alpha-amino group of protein in ubiquitin-mediated protein breakdown. Proc Natl Acad Sci USA 81: 7021-7025.

137. Hollebeke J, Van Damme P, Gevaert K (2012) N-terminal acetylation and other functions of Nalpha-acetyltransferases. Biol Chem 393: 291-298.

138. Yang XJ, Seto E (2008) Lysine acetylation: codified crosstalk with other posttranslational modifications. Mol Cell 31: 449-461.

139. Moore KL (2003) The biology and enzymology of protein tyrosine O-sulfation. J Biol Chem 278: 24243-24246.

140. Maltese WA (1990) Posttranslational modification of proteins by isoprenoids in mammalian cells. FASEB J 4: 3319-3328.

141. Novelli G, D'Apice MR (2012) Protein farnesylation and disease. J Inherit Metab Dis 35: 917-926.

142. Basu J (2004) Protein palmitoylation and dynamic modulation of protein function. Curr Sci India 87: 212-217.

143. Dunphy JT, Linder ME (1998) Signalling functions of protein palmitoylation. Biochim Biophys Acta 1436: 245-261.

144. Silva AM, Vitorino R, Domingues MR, Spickett CM, Domingues P (2013) Post-translational modifications and mass spectrometry detection. Free Radical Biol Med 65: 925-941.

145. Artimo P, Jonnalagedda M, Arnold K, Baratin D, Csardi G, *et al.* (2012) ExPASy: SIB bioinformatics resource portal. Nucleic Acids Res 40: W597-603.

146. Blom N, Gammeltoft S, Brunak S (1999) Sequence and structure-based prediction of eukaryotic protein phosphorylation sites. J Mol Biol 294: 1351-1362.

147. Julenius K, Molgaard A, Gupta R, Brunak S (2005) Prediction, conservation analysis, and structural characterization of mammalian mucin-type O-glycosylation sites. Glycobiology 15: 153-164.

148. Gupta R, Brunak S (2002) Prediction of glycosylation across the human proteome and the correlation to protein function. Pac Symp Biocomput: 310-322.

149. Kiemer L, Bendtsen JD, Blom N (2005) NetAcet: prediction of N-terminal acetylation sites. Bioinformatics 21: 1269-1270.

150. Julenius K (2007) NetCGlyc 1.0: prediction of mammalian C-mannosylation sites. Glycobiology 17: 868-876.

151. Monigatti F, Gasteiger E, Bairoch A, Jung E (2002) The Sulfinator: predicting tyrosine sulfation sites in protein sequences. Bioinformatics 18: 769-770.

152. Maurer-Stroh S, Eisenhaber F (2005) Refinement and prediction of protein prenylation motifs. Genome Biol 6: R55.

153. Gao J, Liao J, Yang GY (2009) CAAX-box protein, prenylation process and carcinogenesis. Am J Transl Res 1: 312-325.

154. Maurer-Stroh S, Koranda M, Benetka W, Schneider G, Sirota FL, *et al.* (2007) Towards complete sets of farnesylated and geranylgeranylated proteins. PLoS Comput Biol 3: e66.

155. Ren J, Wen L, Gao X, Jin C, Xue Y, *et al.* (2008) CSS-Palm 2.0: an updated software for palmitoylation sites prediction. Protein Eng Des Sel 21: 639-644.

156. Zhou F, Xue Y, Yao X, Xu Y (2006) CSS-Palm: palmitoylation site prediction with a clustering and scoring strategy (CSS). Bioinformatics 22: 894-896.

157. Roth AF, Wan J, Bailey AO, Sun B, Kuchar JA, *et al.* (2006) Global analysis of protein palmitoylation in yeast. Cell 125: 1003-1013.

158. Pinilla C, Irani AH, Seriani N, Scandolo S (2012) Ab initio parameterization of an all-atom polarizable and dissociable force field for water. J Chem Phys 136: 114511.

159. Halgren TA (1999) MMFF VII. Characterization of MMFF94, MMFF94s, and other widely available force fields for conformational energies and for intermolecular-interaction energies and geometries. J Comput Chem 20: 730-748.

160. Oostenbrink C, Villa A, Mark AE, van Gunsteren WF (2004) A biomolecular force field based on the free enthalpy of hydration and solvation: the GROMOS force-field parameter sets 53A5 and 53A6. J Comput Chem 25: 1656-1676.

161. Ponder JW, Case DA (2003) Force fields for protein simulations. Adv Protein Chem 66: 27-85.

162. Mackerell AD, Jr. (2004) Empirical force fields for biological macromolecules: overview and issues. J Comput Chem 25: 1584-1604.

163. Van Der Spoel D, Lindahl E, Hess B, Groenhof G, Mark AE, *et al.* (2005) GROMACS: fast, flexible, and free. J Comput Chem 26: 1701-1718.

164. Hess B, Kutzner C, van der Spoel D, Lindahl E (2008) GROMACS 4: Algorithms for highly efficient, load-balanced, and scalable molecular simulation. J Chem Theory Comput 4: 435-447.

165. Hess B, Bekker H, Berendsen HJC, Fraaije JGEM (1997) LINCS: A linear constraint solver for molecular simulations. J Comput Chem 18: 1463-1472.

166. Hess B (2008) P-LINCS: A parallel linear constraint solver for molecular simulation. J Chem Theory Comput 4: 116-122.

167. van der Spoel D, van Maaren PJ (2006) The origin of layer structure artifacts in simulations of liquid water. J Chem Theory Comput 2: 1-11.

168. Patra M, Karttunen M, Hyvonen MT, Falck E, Lindqvist P, *et al.* (2003) Molecular dynamics simulations of lipid bilayers: major artifacts due to truncating electrostatic interactions. Biophys J 84: 3636-3645.

169. Darden T, York D, Pedersen L (1993) Particle Mesh Ewald - an N.Log(N) Method for Ewald Sums in Large Systems. J Chem Phys 98: 10089-10092.

170. Essmann U, Perera L, Berkowitz ML, Darden T, Lee H, *et al.* (1995) A Smooth Particle Mesh Ewald Method. J Chem Phys 103: 8577-8593.

171. Berendsen HJC (1991) Transport Properties Computed by Linear Response through Weak Coupling to a Bath. In: Meyer M, Pontikis V, editors. Computer Simulation in Materials Science: Springer Netherlands. pp. 139-155.

172. Eastwood MP, Stafford KA, Lippert RA, Jensen MO, Maragakis P, *et al.* (2010) Equipartition and the Calculation of Temperature in Biomolecular Simulations. J Chem Theory Comput 6: 2045-2058.

173. Bekker H, Dijkstra EJ, Renardus MKR, Berendsen HJC (1995) An Efficient, Box Shape Independent Nonbonded Force and Virial Algorithm for Molecular-Dynamics. Mol Simulat 14: 137-151.

174. van Gunsteren WF, Berendsen HJC (1990) Moleküldynamik-Computersimulationen; Methodik, Anwendungen und Perspektiven in der Chemie. Angew Chem 102: 1020-1055.

175. Knecht V, Muller M, Bonn M, Marrink SJ, Mark AE (2005) Simulation studies of pore and domain formation in a phospholipid monolayer. J Chem Phys 122: 024704.

176. Mohammad-Aghaie D, Mace E, Sennoga CA, Seddon JM, Bresme F (2010) Molecular dynamics simulations of liquid condensed to liquid expanded transitions in DPPC monolayers. J Phys Chem B 114: 1325-1335.

177. Rose D, Rendell J, Lee D, Nag K, Booth V (2008) Molecular dynamics simulations of lung surfactant lipid monolayers. Biophys Chem 138: 67-77.

178. Kukol A (2009) Lipid Models for United-Atom Molecular Dynamics Simulations of Proteins. J Chem Theory Comput 5: 615-626.

179. Sommer B, Dingersen T, Gamroth C, Schneider SE, Rubert S, *et al.* (2011) CELLmicrocosmos 2.2 MembraneEditor: a modular interactive shape-based software approach to solve heterogeneous membrane packing problems. J Chem Inf Model 51: 1165-1182.

180. Berendsen HJC, Postma, J. P. M., van Gunsteren, W. F., Hermans, J. (1981) Interaction models for water in relation to protein hydration. In: Pullman B, editor. Intermolecular Forces: Reidel Publishing Company Dordrecht. pp. 331–342.

181. Nose S (1984) A molecular dynamics method for simulations in the canonical ensemble. Mol Phys 52: 255-268.

182. Hoover WG (1985) Canonical dynamics: Equilibrium phase-space distributions. Phys Rev A 31: 1695-1697.

183. Parrinello M, Rahman A (1981) Polymorphic transitions in single crystals: A new molecular dynamics method. J Appl Phys 52: 7182-7190.

184. Nose S, Klein ML (1983) Constant pressure molecular dynamics for molecular systems. Mol Phys 50: 1055-1076.

185. Nagle JF, Tristram-Nagle S (2000) Structure of lipid bilayers. Biochim Biophys Acta 1469: 159-195.

186. Konig S, Pfeiffer W, Bayerl T, Richter D, Sackmann E (1992) Molecular-Dynamics of Lipid Bilayers Studied by Incoherent Quasi-Elastic Neutron-Scattering. J Phys Ii 2: 1589-1615.

187. Anezo C, de Vries AH, Holtje HD, Tieleman DP, Marrink SJ (2003) Methodological issues in lipid bilayer simulations. J Phys Chem B 107: 9424-9433.

188. Smith GR (2002) G43a1 force field modified to contain phosphorylated Ser, Thr and Tyr. GROMACS User Contributions, Available: http://www.gromacs.org/Downloads/User_contributions/Force_fields, *Accessed 26.11.2014*

189. Bussi G, Donadio D, Parrinello M (2007) Canonical sampling through velocity rescaling. J Chem Phys 126: 014101

190. Almlof M, Brandsdal BO, Aqvist J (2004) Binding affinity prediction with different force fields: examination of the linear interaction energy method. J Comput Chem 25: 1242-1254.

191. Grossman JC, Schwegler E, Galli G (2004) Quantum and classical molecular dynamics simulations of hydrophobic hydration structure around small solutes. J Phys Chem B 108: 15865-15872.

192. Allesch M, Schwegler E, Galli G (2007) Structure of hydrophobic hydration of benzene and hexafluorobenzene from first principles. J Phys Chem B 111: 1081-1089.

193. Li JL, Car R, Tang C, Wingreen NS (2007) Hydrophobic interaction and hydrogen-bond network for a methane pair in liquid water. Proc Natl Acad Sci USA 104: 2626-2630.

194. Kabsch W, Sander C (1983) Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. Biopolymers 22: 2577-2637.

195. Joosten RP, te Beek TA, Krieger E, Hekkelman ML, Hooft RW, *et al.* (2011) A series of PDB related databases for everyday needs. Nucleic Acids Res 39: D411-419.

196. Humphrey W, Dalke A, Schulten K (1996) VMD: visual molecular dynamics. J Mol Graph 14: 27-38.

197. Castrignano T, De Meo PD, Cozzetto D, Talamo IG, Tramontano A (2006) The PMDB Protein Model Database. Nucleic Acids Res 34: D306-309.

198. Rausch F, Schicht M, Paulsen F, Ngueya I, Bräuer L, *et al.* (2012) "SP-G", a putative new surfactant protein--tissue localization and 3D structure. PLoS One 7: e47789.

199. Schicht M, Rausch F, Finotto S, Mathews M, Mattil A, *et al.* (2014) SFTA3, a novel protein of the lung: 3D-Structure, Characterization and Immune activation. Eur Respir J 44: 447-456.

200. Nagle JF (1993) Area/lipid of bilayers from NMR. Biophys J 64: 1476-1481.

201. Biltonen RL, Lichtenberg D (1993) The Use of Differential Scanning Calorimetry as a Tool to Characterize Liposome Preparations. Chem Phys Lipids 64: 129-142.

202. Kranenburg M, Smit B (2005) Phase behavior of model lipid bilayers. J Phys Chem B 109: 6553-6563.

203. Kryshtafovych A, Fidelis K, Moult J (2014) CASP10 results compared to those of previous CASP experiments. Proteins: Struct, Funct, Bioinf 82: 164-174.

204. Mittal RA, Hammel M, Schwarz J, Heschl KM, Bretschneider N, *et al.* (2012) SFTA2-A Novel Secretory Peptide Highly Expressed in the Lung-Is Modulated by Lipopolysaccharide but Not Hyperoxia. PLoS One 7: e40011.

205. Olausson BE, Grossfield A, Pitman MC, Brown MF, Feller SE, *et al.* (2012) Molecular dynamics simulations reveal specific interactions of post-translational palmitoyl modifications with rhodopsin in membranes. J Am Chem Soc 134: 4324-4331.

206. Meyer T, York JD (1999) Calcium-myristoyl switches turn on new lights. Nature Cell Biol 1: E93-95.

207. Tozzini V (2005) Coarse-grained models for proteins. Curr Opin Struct Biol 15: 144-150.

208. Marrink SJ, Risselada HJ, Yefimov S, Tieleman DP, de Vries AH (2007) The MARTINI force field: coarse grained model for biomolecular simulations. J Phys Chem B 111: 7812-7824.

209. Bradley R, Radhakrishnan R (2013) Coarse-Grained Models for Protein-Cell Membrane Interactions. Polymers 5: 890-936.

210. Christ CD, Mark AE, van Gunsteren WF (2010) Basic ingredients of free energy calculations: a review. J Comput Chem 31: 1569-1582.

211. Hansen N, van Gunsteren WF (2014) Practical Aspects of Free-Energy Calculations: A Review. J Chem Theory Comput 10: 2632-2647.

212. Leavitt S, Freire E (2001) Direct measurement of protein binding energetics by isothermal titration calorimetry. Curr Opin Struct Biol 11: 560-566.

213. ten Brinke A, Vaandrager AB, Haagsman HP, Ridder AN, van Golde LM, *et al.* (2002) Structural requirements for palmitoylation of surfactant protein C precursor. Biochem J 361: 663-671.

214. Vorbroker DK, Voorhout WF, Weaver TE, Whitsett JA (1995) Posttranslational processing of surfactant protein C in rat type II cells. Am J Physiol 269: L727-733.

215. Smith GR, Sternberg MJ (2002) Prediction of protein-protein interactions by docking methods. Curr Opin Struct Biol 12: 28-35.

216. Madsen J, Kliem A, Tornoe I, Skjodt K, Koch C, *et al.* (2000) Localization of lung surfactant protein D on mucosal surfaces in human tissues. J Immunol 164: 5866-5870.

217. Schicht M, Knipping S, Hirt R, Beileke S, Sel S, *et al.* (2013) Detection of surfactant proteins A, B, C, and D in human nasal mucosa and their regulation in chronic rhinosinusitis with polyps. Am J Rhinol Allergy 27: 24-29.

218. Foster KA, Oster CG, Mayer MM, Avery ML, Audus KL (1998) Characterization of the A549 cell line as a type II pulmonary epithelial cell model for drug metabolism. Exp Cell Res 243: 359-366.

219. Parra E, Alcaraz A, Cruz A, Aguilella VM, Perez-Gil J (2013) Hydrophobic pulmonary surfactant proteins SP-B and SP-C induce pore formation in planar lipid membranes: evidence for proteolipid pores. Biophys J 104: 146-155.

220. Horn JN, Kao TC, Grossfield A (2014) Coarse-grained molecular dynamics provides insight into the interactions of lipids and cholesterol with rhodopsin. Adv Exp Med Biol 796: 75-94.

221. Butovich IA (2013) Tear film lipids. Exp Eye Res 117: 4-27.

8. Appendix

Appendix 1: Template of a GROMACS simulation parameter file (.mdp) for an energy minimization run.

```
title               = minimization run
; Run Control
integrator          = steep             ; Steepest descent
emtol               = 100.0             ; Stop if force < emtol (10 kJ/(mol*nm))
emstep              = 0.01              ; Initial step size (0.01 nm)
nsteps              = 2500              ; Maximum number of minimization steps
; Output Control
nstlog              = 10                ; Output to .log
nstenergy           = 10                ; Output to .edr
; Neighbor Searching
ns_type             = grid              ; Search neighboring grid cells
nstlist             = 1                 ; Neighbor list updated every step
rlist               = 1.4               ; Short range neighbor list cutoff (nm)
pbc                 = xyz               ; PBC in all directions
; Electrostatics
coulombtype         = PME               ; Particle Mesh Ewald summation
rcoulomb            = 1.4               ; Short range electrostatics cutoff
; Van-der-Waals
vdw-type            = cut-off           ; LJ potential with plain cutoff
rvdw                = 1.4               ; Short range van-der-Waals cutoff (nm)
```

Appendix 2: Template of a GROMACS simulation parameter file (.mdp) for a 250 ps equilibration run with NVT ensemble (constant particle number, volume and temperature).

```
title                    = NVT equilibration run
; Run Control
continuation             = no               ; No restart
constraint_algorithm     = LINCS            ; LINCS on
constraints              = hbonds           ; Bonds with hydrogens constrained
lincs_iter               = 1                ; Number of iterations for LINCS
lincs_order              = 4                ; LINCS accuracy
integrator               = md
dt                       = 0.002            ; Time step
nsteps                   = 125000           ; Number of MD steps
; Position restraints ON
define                   = -DPOSRES_PROT -DPOSRES_DPPC   ; Restraint on all
                                                         ; lipids and protein
; Generate velocities
gen_vel                  = yes              ; Generate velocities
gen_temp                 = 323              ; Temp. for Maxwell distribution
gen_seed                 = -1               ; Generate random seed
; Output Control
nstxout                  = 1000             ; Coordinates output to .trr
nstvout                  = 1000             ; Velocities output to .trr
nstfout                  = 1000             ; Forces output to .trr
nstlog                   = 500              ; Output to .log
nstxtcout                = 500              ; Output to .xtc
nstenergy                = 500              ; Output to .edr
; Neighbor Searching
ns_type                  = grid             ; Search neighboring grid cells
nstlist                  = 5                ; Time step dependent! 10fs
pbc                      = xyz              ; PBC in all directions
rlist                    = 1.2              ; Short range neighbor list cutoff
rlistlong                = 1.4
; Electrostatics
coulombtype              = PME              ; Particle Mesh Ewald
rcoulomb                 = 1.2              ; Short range cutoff
pme_order                = 4                ; Cubic interpolation
fourierspacing           = 0.16             ; Grid spacing for FFT
; van-der-Waals
vdw-type                 = switch           ; LJ is switched off smoothly
rvdw_switch              = 1.2              ; Begin of potential switch off
rvdw                     = 1.3              ; Short range cutoff (LJ = 0)
dispcorr                 = no               ; Dispersion correction off
; Temperature Coupling is ON = constant temperature
tcoupl                   = V-rescale        ; Modified Berendsen thermostat
tc-grps                  = protein other water     ; Coupling groups
tau-t                    =   0.1    0.1   0.1      ; Coupling time constant
ref-t                    =   323    323   323      ; Reference temperature
; Pressure Coupling is OFF = constant volume
pcoupl                   = no               ; No pressure coupling for NVT
```

Appendix 3: Template of a GROMACS simulation parameter file (.mdp) for a 250 ps equilibration run with NPT ensemble (constant number of particles, pressure and temperature).

```
title                    = NPT equilibration run
; Run Control
continuation             = yes              ; Restart from NVT equilibration
constraint_algorithm     = LINCS            ; LINCS on
constraints              = hbonds           ; Only bonds involving hydrogens
lincs_iter               = 1                ; Number of iterations in LINCS
lincs_order              = 4                ; LINCS accuracy
integrator               = md
dt                       = 0.002            ; Time step
nsteps                   = 125000           ; Number of MD steps
; Generate velocities
gen_vel                  = no               ; Since continuation = yes
; Output Control
nstlog                   = 500              ; Output to .log
nstxtcout                = 500              ; Output to .xtc
nstenergy                = 500              ; Output to .edr
; Neighbor Searching
ns_type                  = grid             ; Search neighboring grid cells
nstlist                  = 5                ; Time step dependent! 10fs
pbc                      = xyz              ; PBC in all directions
rlist                    = 1.2              ; Short range neighbor list cutoff
rlistlong                = 1.4
; Electrostatics
coulombtype              = PME              ; Particle Mesh Ewald
rcoulomb                 = 1.2              ; Short range cutoff
pme_order                = 4                ; Cubic interpolation
fourierspacing           = 0.16             ; Grid spacing for FFT
; van-der-Waals
vdw-type                 = switch           ; LJ is switched off smoothly
rvdw_switch              = 1.2              ; Begin of potential switch
rvdw                     = 1.3              ; Short range cutoff (LJ = 0)
dispcorr                 = no               ; Dispersion correction off
; Temperature Coupling is ON = constant temperature
tcoupl                   = V-rescale        ; Modified Berendsen thermostat
tc-grps                  = protein other water     ; Coupling groups
tau-t                    =   0.1    0.1   0.1      ; Coupling time constant
ref-t                    =   323    323   323      ; Reference temperature
; Pressure Coupling is now ON = constant pressure
pcoupl                   = Berendsen        ; Equilibration with Berendsen
pcoupltype               = semiisotropic    ; Isotropic only in x and y
tau-p                    = 1.0              ; Constant coupling in x/y and z
compressibility          = 4.5e-5   0       ; Water standard in x/y and 0 in z
ref-p                    = 1.0     1.0      ; Reference pressure
```

Appendix 4: Template of a GROMACS simulation parameter file (.mdp) for a 25 ns production run of a DPPC bilayer system.

```
title                    = DPPC bilayer production run 25 ns
; Run Control
continuation             = yes              ; Restart from NPT equilibration
constraint_algorithm     = LINCS            ; LINCS on
constraints              = all-bonds        ; All bonds constrained
lincs_iter               = 1                ; Number of iterations in LINCS
lincs_order              = 4                ; LINCS accuracy
integrator               = md
dt                       = 0.004            ; Time step
nsteps                   = 6250000          ; Number of MD steps
; Position restraints OFF
; Generate velocities
gen_vel                  = no               ; Since continuation = yes
; Output Control
nstlog                   = 2500             ; Output to .log
nstxtcout                = 2500             ; Output to .xtc
nstenergy                = 2500             ; Output to .edr
; Neighbor Searching
ns_type                  = grid             ; Search neighboring grid cells
nstlist                  = 5                ; Time step dependent! 20fs
pbc                      = xyz              ; PBC in all directions
rlist                    = 1.2              ; Short range neighbor list cutoff
rlistlong                = 1.4
; Electrostatics
coulombtype              = PME              ; Particle Mesh Ewald
rcoulomb                 = 1.2              ; Short range cutoff
pme_order                = 4                ; Cubic interpolation
fourierspacing           = 0.16             ; Grid spacing for FFT
; van-der-Waals
vdw-type                 = switch           ; LJ is switched off smoothly
rvdw_switch              = 1.2              ; Begin of potential switch off
rvdw                     = 1.3              ; Short range cutoff (LJ = 0)
dispcorr                 = EnerPres         ; Dispersion correction ON
; Temperature Coupling is ON = constant temperature
tcoupl                   = Nose-Hoover      ; Resembling canonical ensemble
nh-chain-length          = 1                ; >1 not supported for leap frog
tc-grps                  = DPPC water       ; Coupling groups
tau-t                    = 0.4   0.4        ; Time constant for coupling
ref-t                    = 323   323        ; Reference temperature
; Pressure Coupling is ON = constant pressure
pcoupl                   = Parrinello-Rahman; Pressure Coupling on
pcoupltype               = semiisotropic    ; Isotropic only in x and y
tau-p                    = 2.0              ; Coupling constant in x/y and z
compressibility          = 4.5e-5   4.5e-5  ; Water standard in x/y and z
ref-p                    = 1.0       1.0    ; Reference pressure
```

Appendix 5: Template of a GROMACS simulation parameter file (.mdp) for a 50 ns production run of a protein-lipid system.

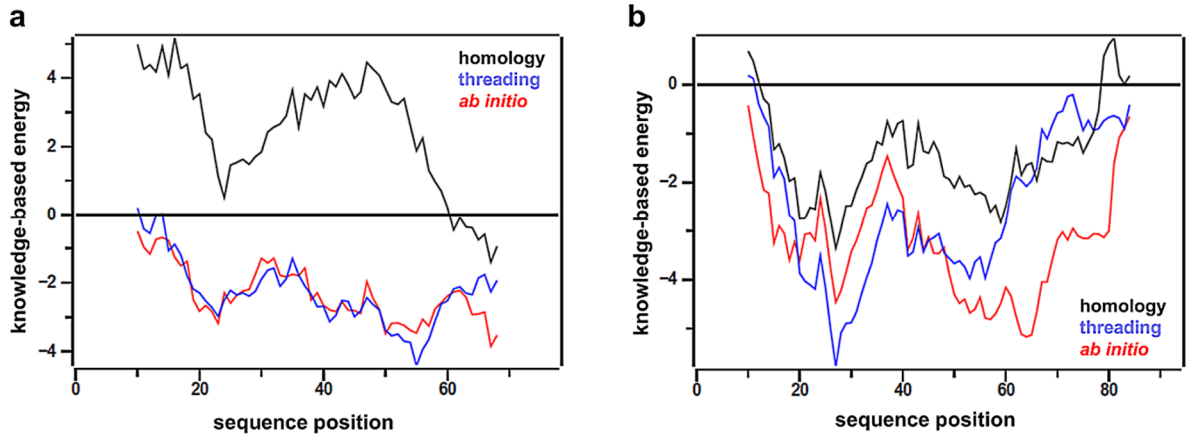
```
title                    = protein-monolayer production run 50 ns
; Run Control
continuation             = yes              ; Restart from NPT equilibration
constraint_algorithm     = LINCS            ; LINCS on
constraints              = h-bonds          ; All bonds constrained
lincs_iter               = 1                ; Number of iterations
lincs_order              = 4                ; LINCS accuracy
integrator               = md
dt                       = 0.002            ; Time step
nsteps                   = 25000000         ; Number of MD steps
; Position restraints OFF
; Generate velocities
gen_vel                  = no               ; Since continuation = yes
; Output Control
nstlog                   = 5000             ; Output to .log
nstxtcout                = 5000             ; Output to .xtc
nstenergy                = 5000             ; Output to .edr
energygrps               = DPPC protein     ; Separate energy groups
; Neighbor Searching
ns_type                  = grid             ; Search neighboring grid cells
nstlist                  = 10               ; Time step dependent! 20fs
pbc                      = xyz              ; PBC in all directions
rlist                    = 1.2              ; Short range neighbor list cutoff
rlistlong                = 1.4
; Electrostatics
coulombtype              = PME              ; Particle Mesh Ewald
rcoulomb                 = 1.2              ; Short range cutoff
pme_order                = 4                ; Cubic interpolation
fourierspacing           = 0.12             ; Grid spacing for FFT
; van-der-Waals
vdw-type                 = switch           ; LJ is switched off smoothly
rvdw_switch              = 1.2              ; Begin of potential switch off
rvdw                     = 1.3              ; Short range cutoff (LJ = 0)
dispcorr                 = no               ; Dispersion correction off
; Temperature Coupling is ON = constant temperature
tcoupl                   = Nose-Hoover      ; Resembling canonical ensemble
nh-chain-length          = 1                ; >1 not supported for leap frog
tc-grps                  = protein other water     ; Coupling Groups
tau-t                    =   0.4    0.4   0.4      ; Coupling time constant
ref-t                    =   323    323   323      ; Reference temperature
; Pressure Coupling is ON = constant pressure
pcoupl                   = Parrinello-Rahman       ; Pressure Coupling on
pcoupltype               = semiisotropic    ; Isotropic only in x and y
tau-p                    = 2.0              ; Coupling constant in x/y and z
compressibility          = 4.5e-5     0     ; Water standard in x/y and 0 in z
ref-p                    = 1.0        1.0   ; Reference pressure
```

Appendix 6: *bash*-script for the evaluation of MD simulations. GROMACS analysis tools are called sequentially with customizable options.

```
#!/bin/bash -l
#
# Analyze the results of a GROMACS MD simulation
#
# Script Arguments: 1: Common name of all files;
#
# This script contains:
#            - g_energy for equiNVT, equiNPT and production run
#            - RMSD and RMSF calculation
#            - DSSP calculation for whole protein
#            - minimization of the last snapshot (.gro), conversion to .pdb
#
#
# Root file name
export NAME=$1

# Number of processors for energy minimization
export PROCS=8

#
# Read out energy files
#
g_energy -f $NAME'_equiNVT.edr' -s $NAME'_equiNVT.tpr' -o $NAME'_equiNVT.xvg' >
simulation_analysis.txt 2>&1 <<EOF
Potential
Total-Energy
Temperature
EOF

g_energy -f $NAME'_equiNPT.edr' -s $NAME'_equiNPT.tpr' -o $NAME'_equiNPT.xvg'
>> simulation_analysis.txt 2>&1 <<EOF
Potential
Total-Energy
Temperature
Pressure
Volume
EOF

g_energy -f $NAME'_production.edr' -s $NAME'_production.tpr' -o
$NAME'_production.xvg' >> simulation_analysis.txt 2>&1 <<EOF
Potential
Total-Energy
Temperature
Pressure
Volume
EOF

echo "Finshed g_energy."
```

```
#
# RMSD calculation for complete protein
#
echo "Running g_rms..."
g_rms -f $NAME'_production.xtc' -s $NAME'_production.tpr' -o
$NAME'_production_RMSD_full.xvg' >> simulation_analysis.txt 2>&1 <<EOF
Protein
Backbone
EOF
echo "Finished RMSD calculation."

#
# RMSF calculation for complete simulation
#
echo "Running g_rmsf..."
g_rmsf -f $NAME'_production.xtc' -s $NAME'_production.tpr' -o
$NAME'_production_RMSF_full.xvg' -res >> simulation_analysis.txt 2>&1 <<EOF
Protein
EOF
echo "Finished RMSF calculation."

#
# Calculate secondary structure (DSSP) for whole protein
#
echo "Running do_dssp.."
do_dssp -f $NAME'_production.xtc' -s $NAME'_production.tpr' -o
$NAME'_production_DSSP.xpm' -sc $NAME'_production_DSSPcount.xvg' >>
simulation_analysis.txt 2>&1 <<EOF
Protein
EOF
echo "Finished DSSP calculation."

#
# Minimize final snapshot
#
echo "Running minimization on $PROCS cores..."
grompp -f minimization.mdp -c $NAME'_production.gro' -p $NAME.top -o
$NAME'_production_EM.tpr' >> simulation_analysis.txt 2>&1

mdrun -nt $PROCS -deffnm $NAME'_production_EM' >> simulation_analysis.txt 2>&1

trjconv -f $NAME'_production_EM.gro' -s $NAME'_production_EM.tpr' -o
$NAME'_production_EM.pdb' -pbc res -ur compact -center >>
simulation_analysis.txt 2>&1 <<EOF
Protein
System
EOF
echo "Minimization finished."

echo "MD ANALYSIS FINISHED"
```

Appendix 7: Comparison of plots of the ProSA II knowledge-based energy for the best models for **(a)** SP-G and **(b)** SP-H obtained by homology modeling (black), threading (blue), and *ab initio* modeling (red).



Appendix 8: Results of protein quality assessment tools for the models of **(a)** SP-G and **(b)** SP-H with and without attached PTMs after the 20 ns stability test MD simulation in comparison to the structures without PTMs directly after the modeling process ("final models").

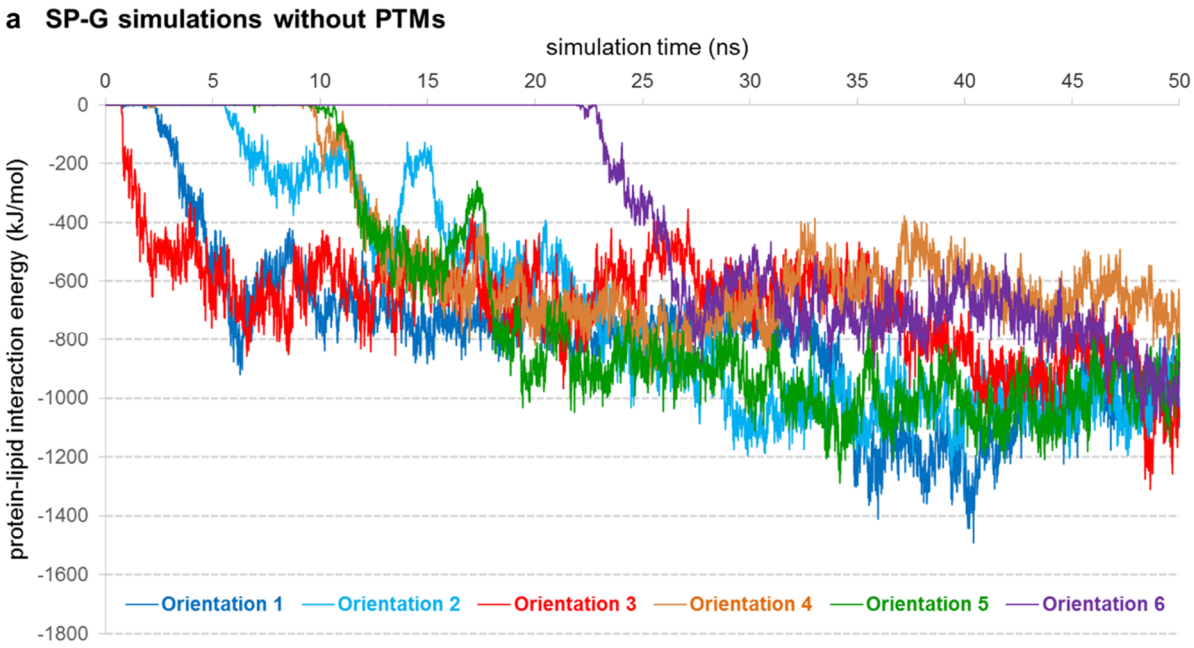
a

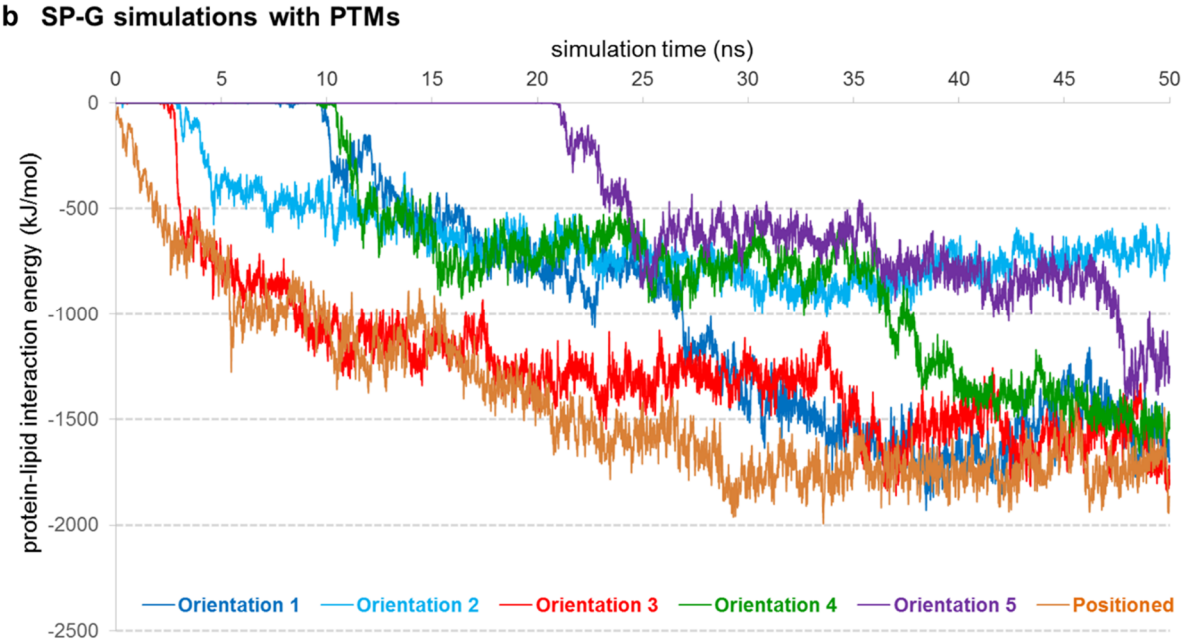
SP-G models	combined Z-score	PROCHECK		ERRAT-score	VERIFY-3D	ProQ	
		fav. regions	outlier			*LGscore*	*MaxSub*
final model	-6.16	95.5%	0	100.0	97.5%	3.579	0.141
after MD without PTMs	-5,84	92,4%	0	100.0	87,3%	4.023	0.185
after MD with PTMs	-6,00	92,4%	0	100.0	92,3%	4.611	0.158

b

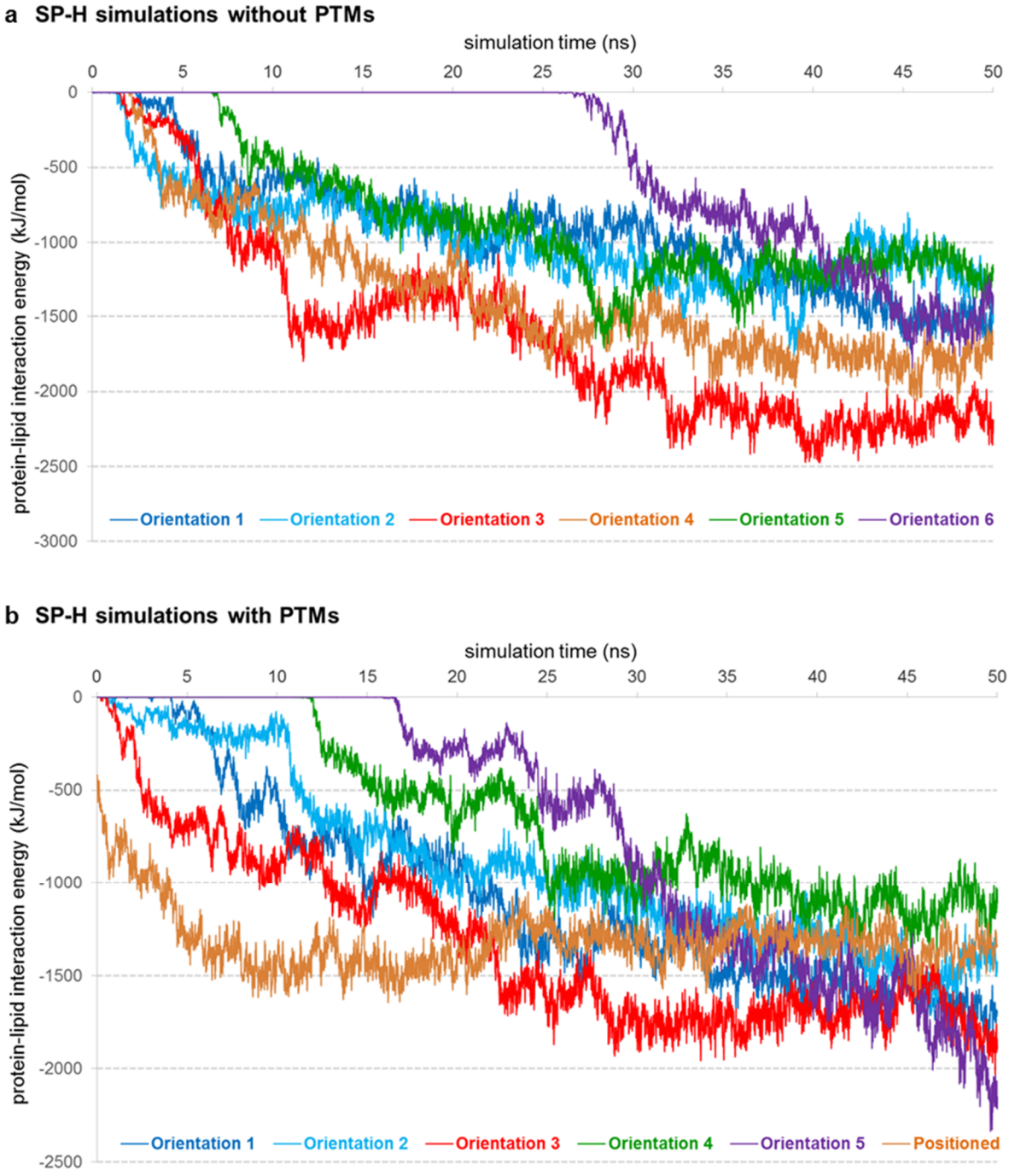
SP-H models	combined Z-score	PROCHECK		ERRAT-score	VERIFY-3D	ProQ	
		fav. regions	outlier			*LGscore*	*MaxSub*
final model	-5.72	94.0%	0	93.0	48.4%	1.804	0.131
after MD without PTMs	-6,10	94.0%	0	94.2	48.4%	2.334	0.075
after MD with PTMs	-6,00	94.0%	0	84.9	36.8%	3.527	0.118

Appendix 9: Protein-lipid interaction energy (in kJ/mol) versus simulation time (in ns) for all six orientations of **(a)** the SP-G model without PTMs and **(b)** the SP-G model with PTMs.

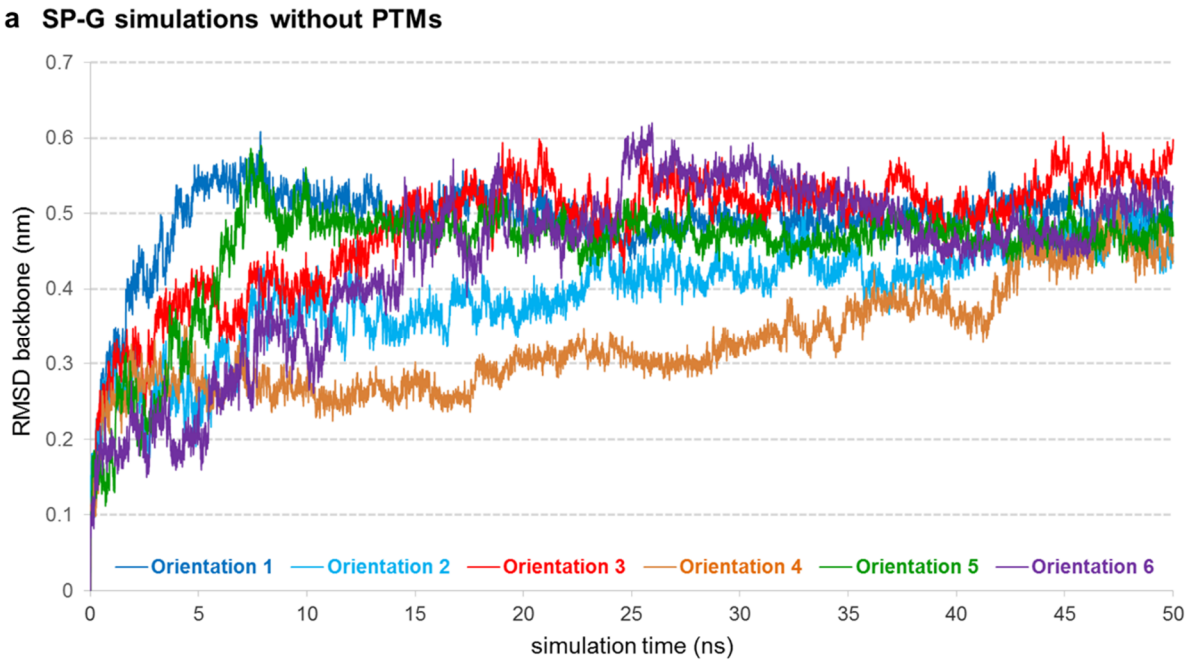


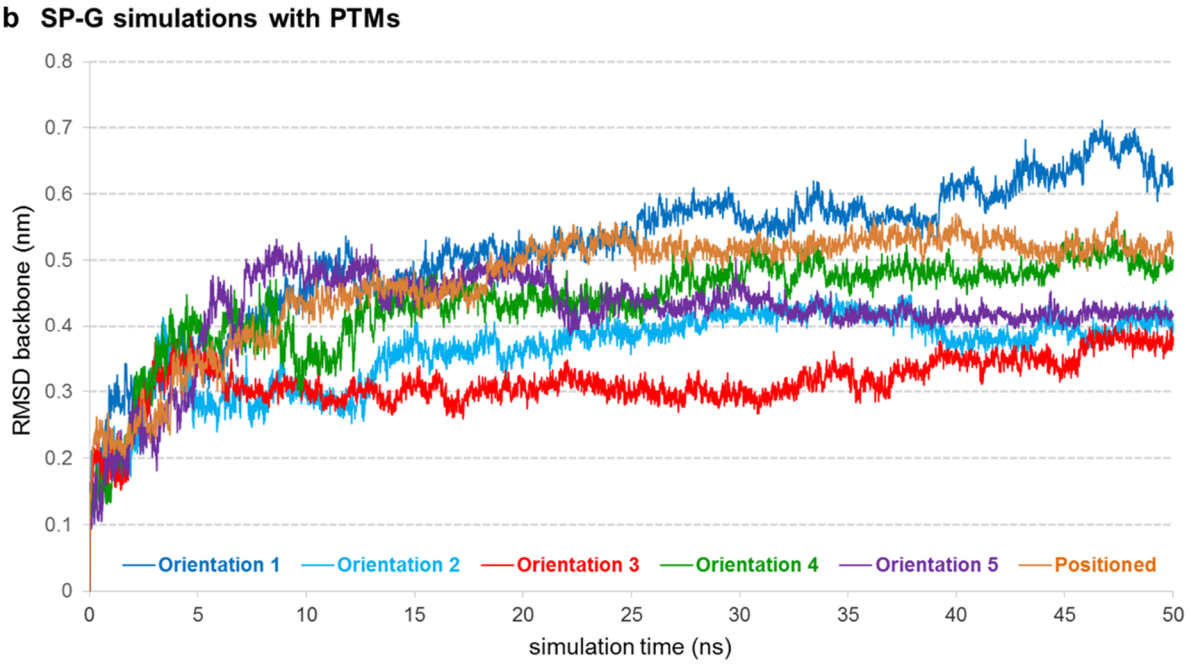


Appendix 10: Protein-lipid interaction energy (in kJ/mol) versus simulation time (in ns) for all six orientations of **(a)** the SP-H model without PTMs and **(b)** the SP-H model with PTMs.

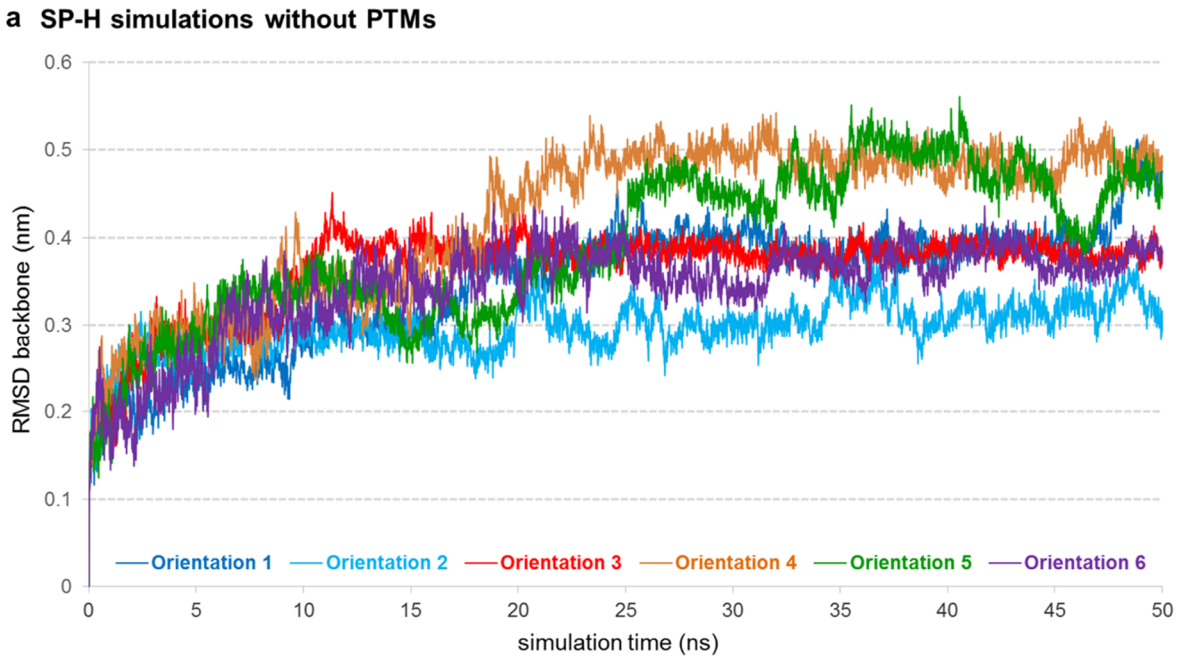


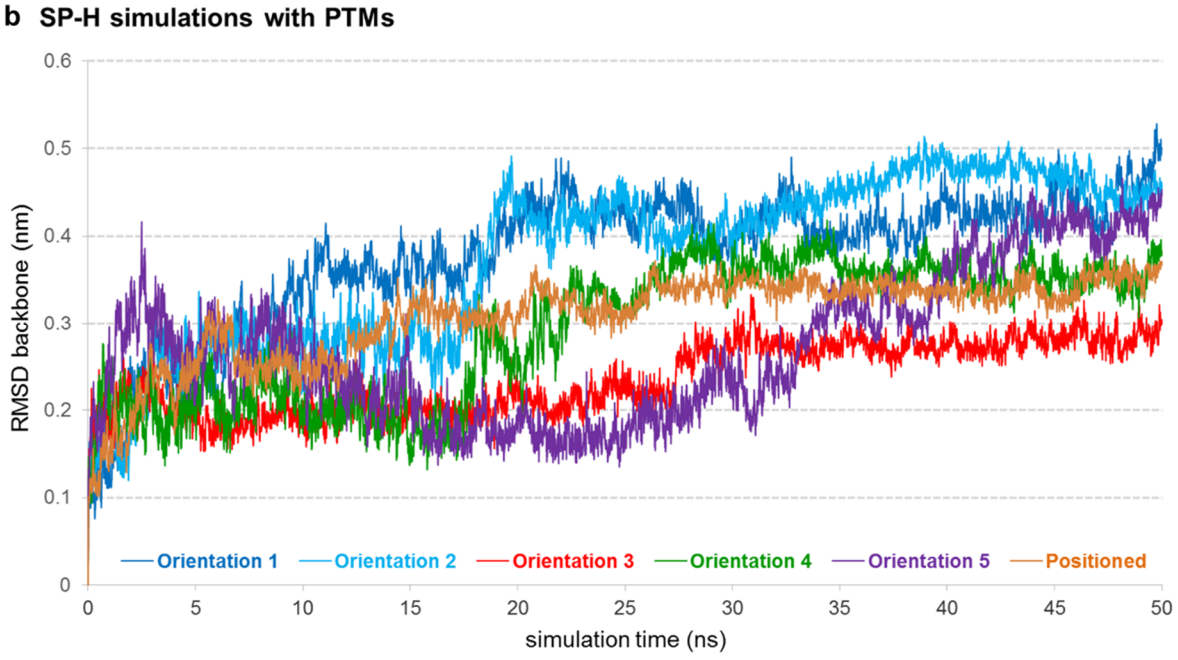
Appendix 11: RMSD of the protein backbone atoms (in nm) versus simulation time (in ns) for all six orientations of **(a)** the SP-G model without PTMs and **(b)** the SP-G model with PTMs.



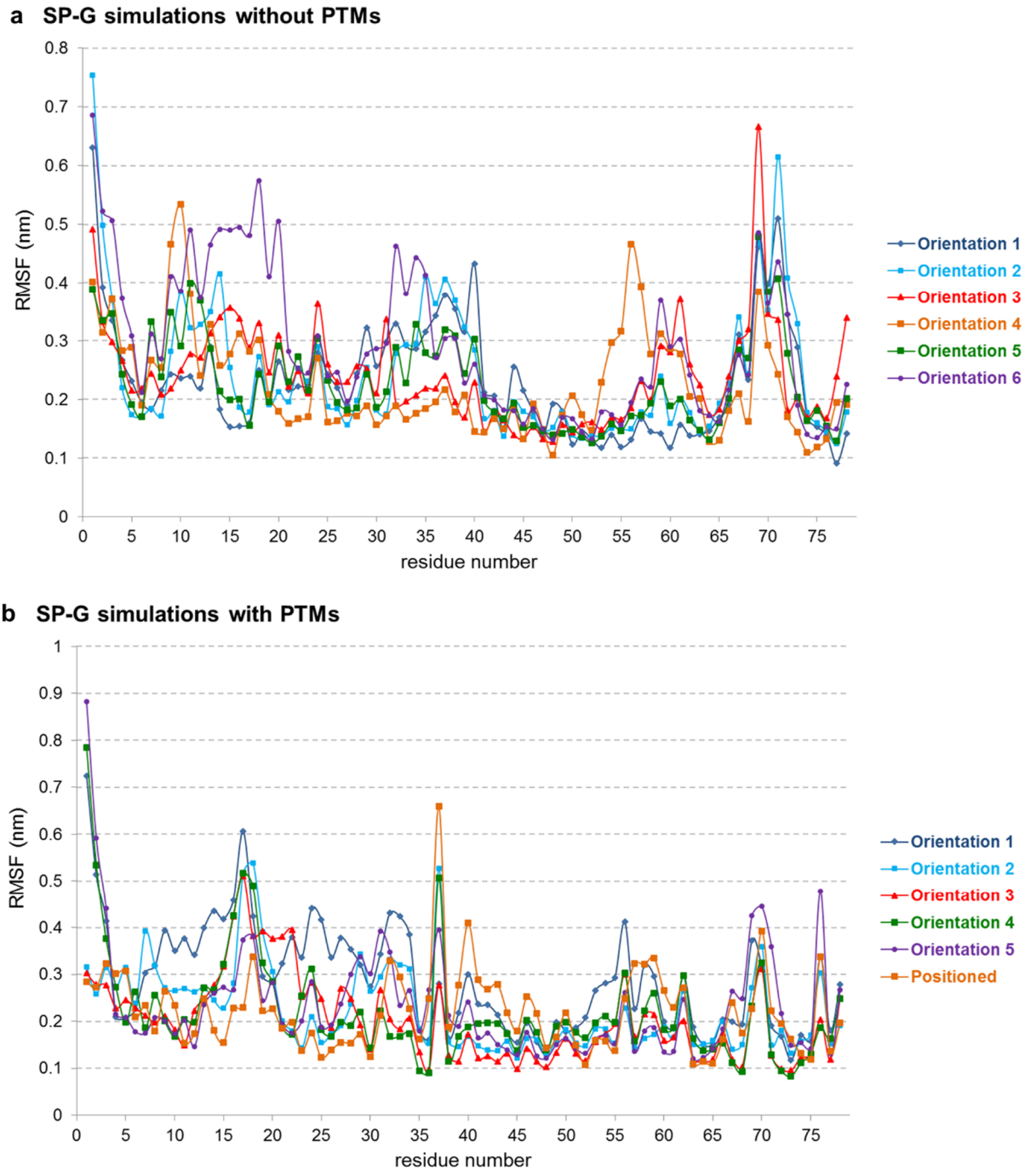


Appendix 12: RMSD of the protein backbone atoms (in nm) versus simulation time (in ns) for all six orientations of **(a)** the SP-H model without PTMs and **(b)** the SP-H model with PTMs.

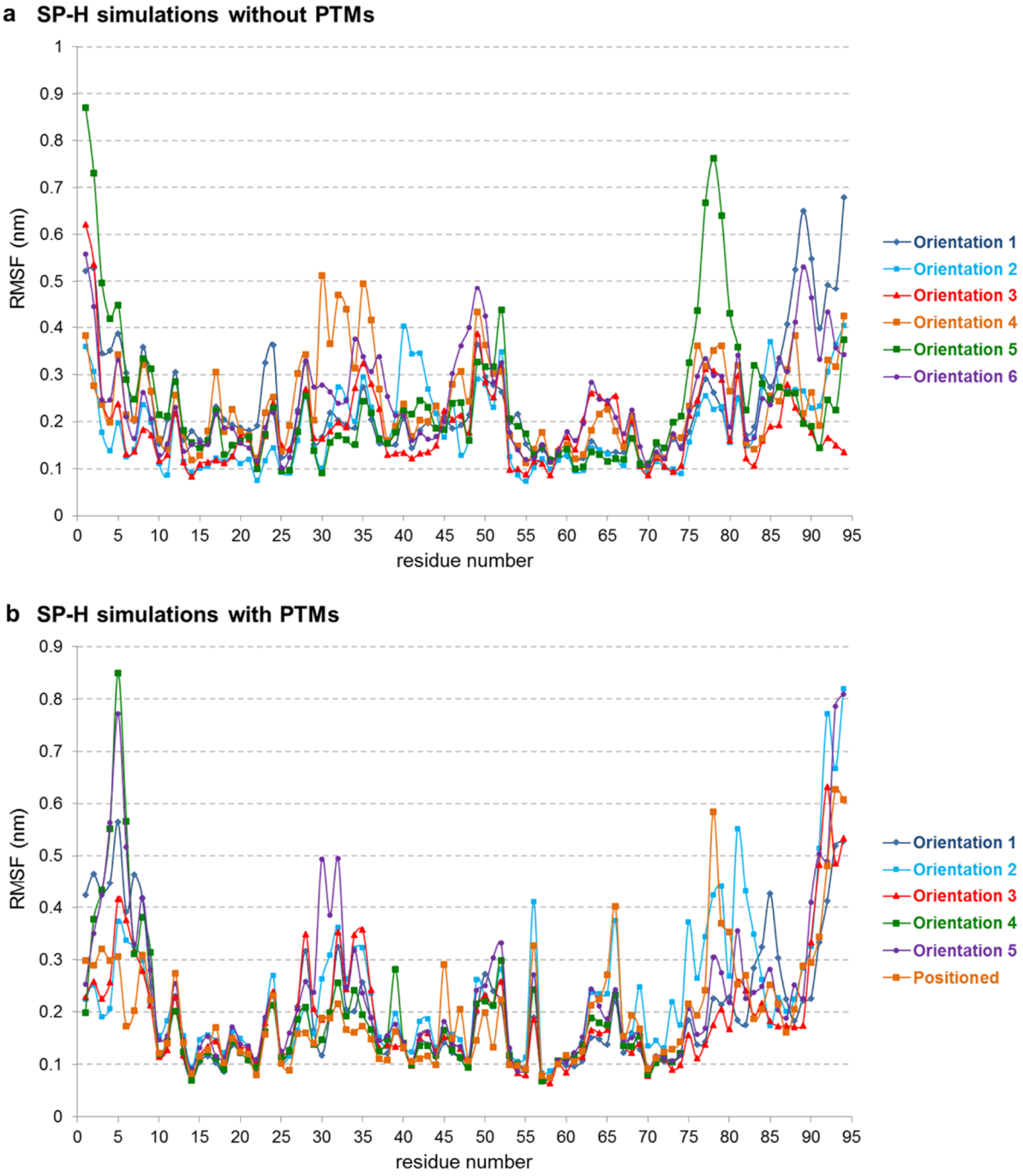




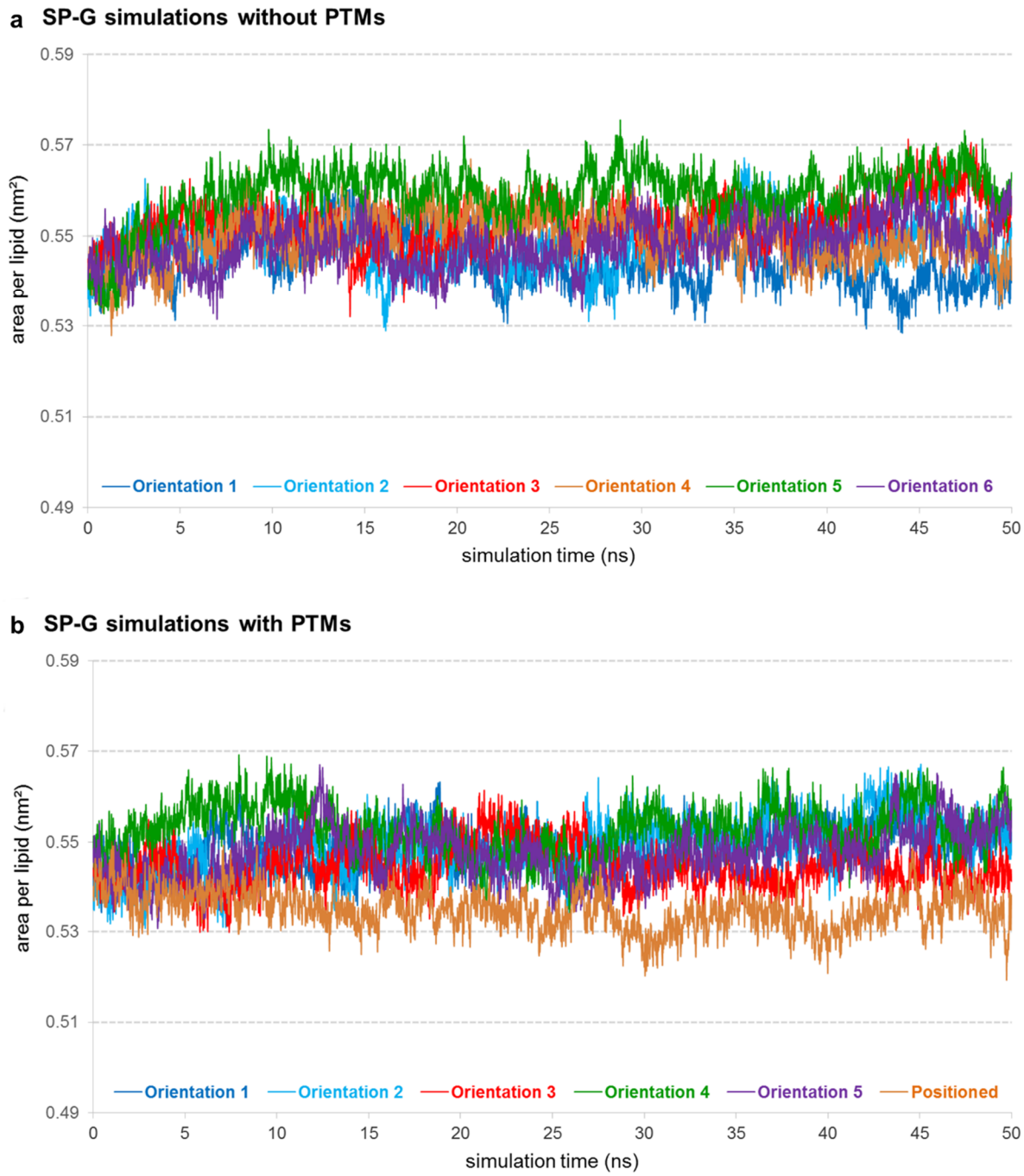
Appendix 13: RMSF for each amino acid residue (in nm) versus simulation time (in ns) for all six orientations of **(a)** the SP-G model without PTMs and **(b)** the SP-G model with PTMs.



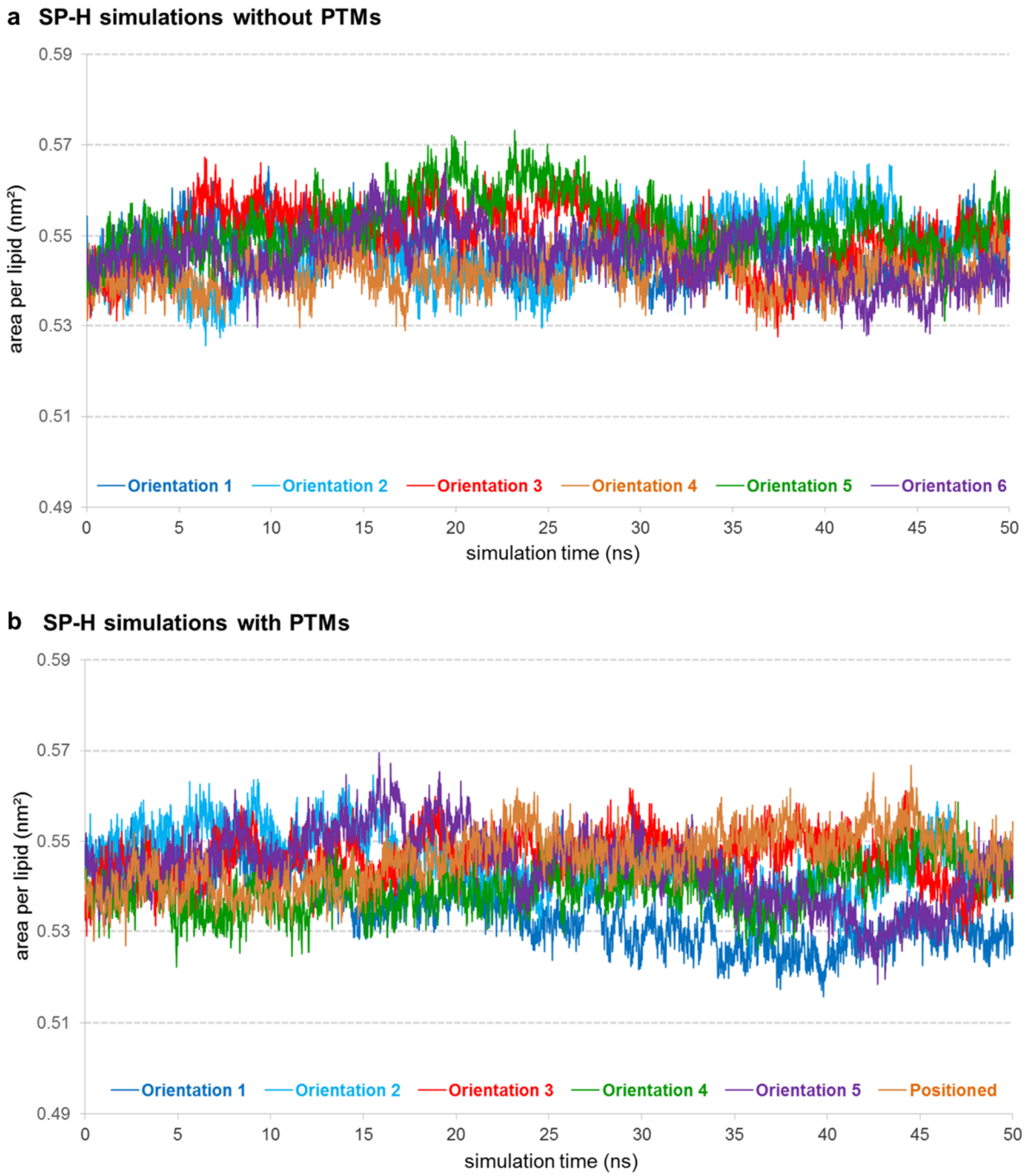
Appendix 14: RMSF for each amino acid residue (in nm) versus simulation time (in ns) for all six orientations of **(a)** the SP-H model without PTMs and **(b)** the SP-H model with PTMs.



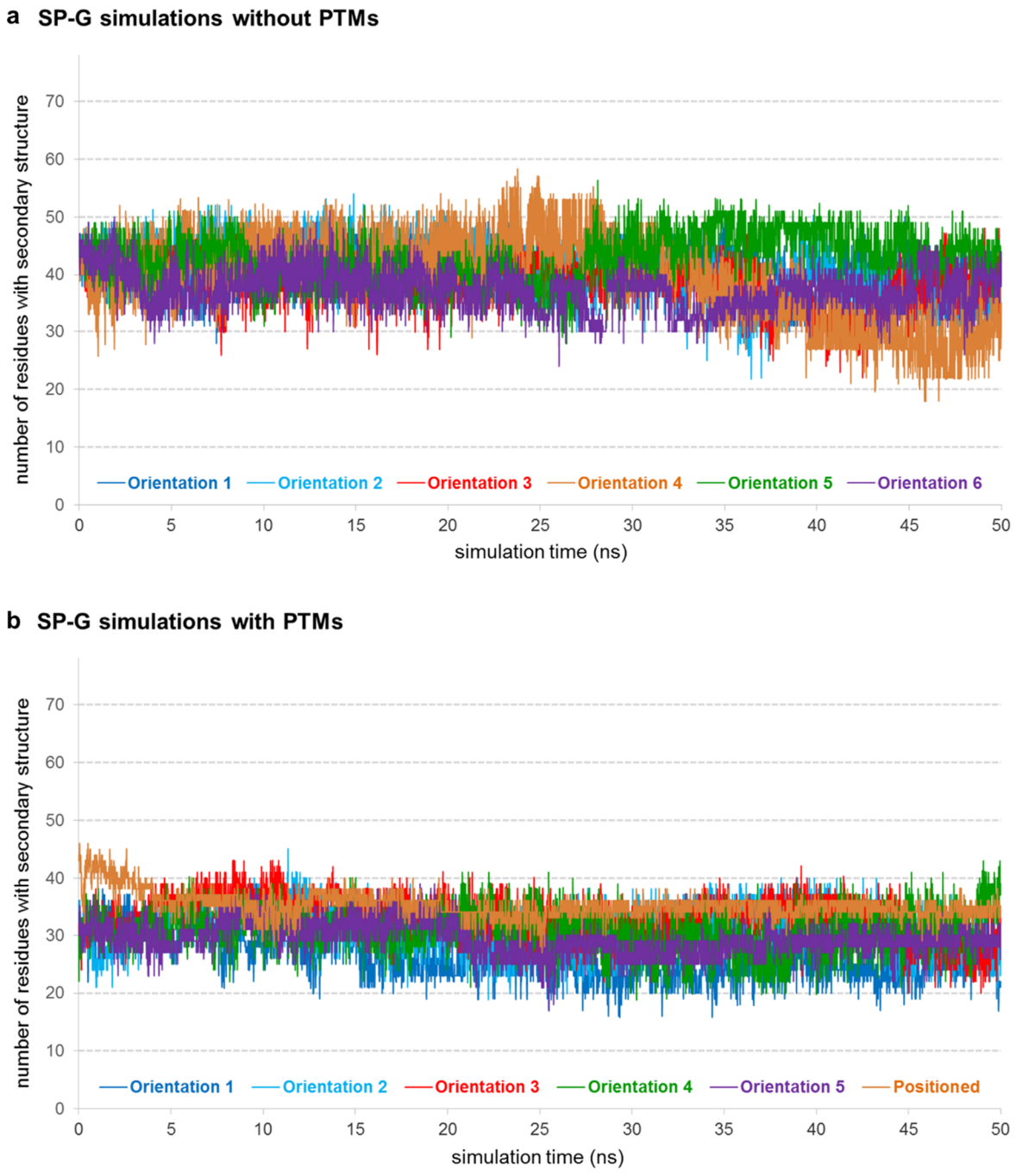
Appendix 15: Area per lipid in the monolayer (in nm²) versus simulation time (in ns) for all six orientations of **(a)** the SP-G model without PTMs and **(b)** the SP-G model with PTMs.



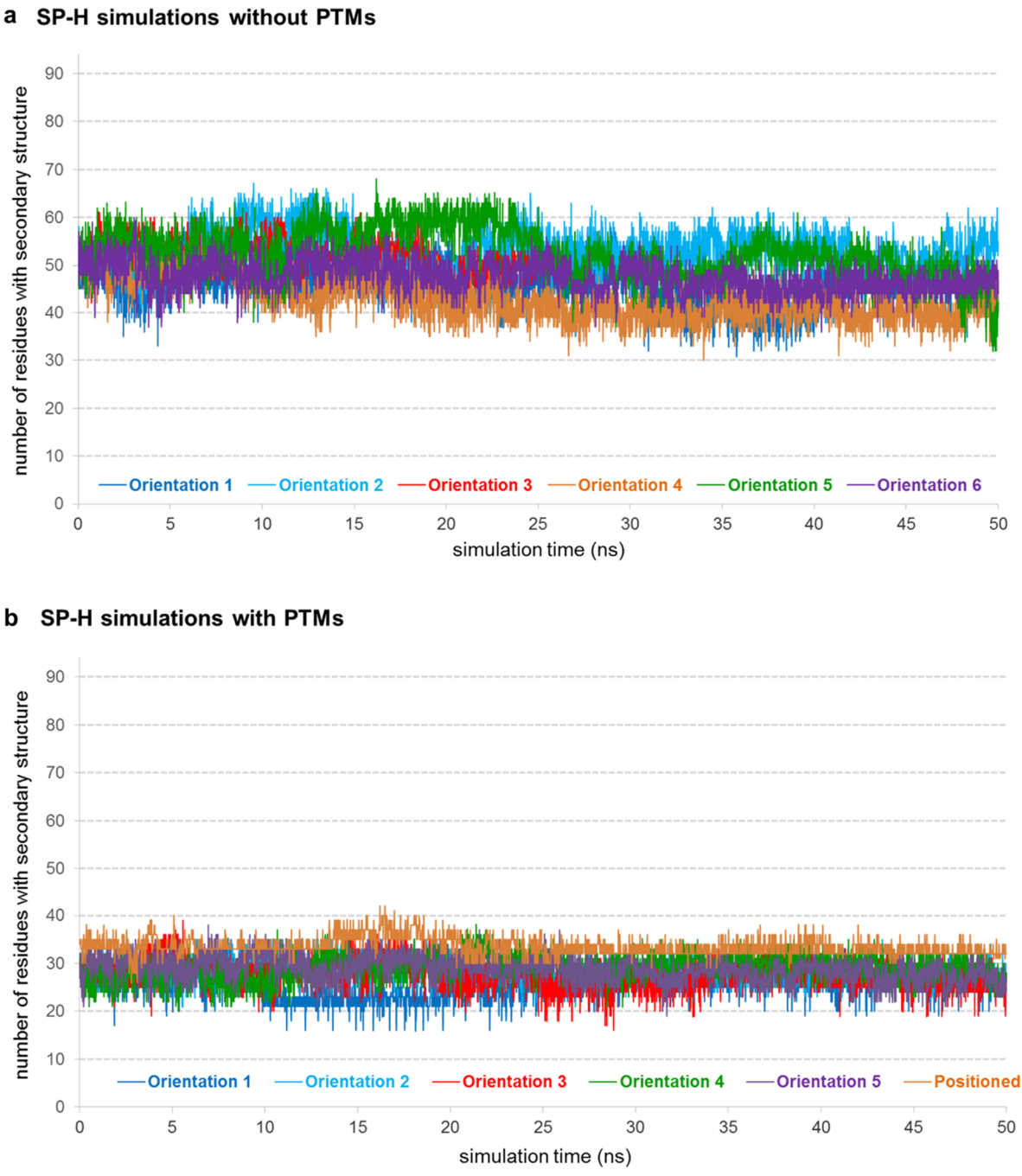
Appendix 16: Area per lipid in the monolayer (in nm²) versus simulation time (in ns) for all six orientations of **(a)** the SP-H model without PTMs and **(b)** the SP-H model with PTMs.



Appendix 17: Number of residues that were assigned as secondary structure element by DSSP (α-helix, β-sheet, β-bridge or turn) versus simulation time (in ns) for all six orientations of **(a)** the SP-G model without PTMs and **(b)** the SP-G model with PTMs.



Appendix 18: Number of residues that were assigned as secondary structure element by DSSP (α-helix, β-sheet, β-bridge or turn) versus simulation time (in ns) for all six orientations of **(a)** the SP-H model without PTMs and **(b)** the SP-H model with PTMs.



9. Publications and lectures

9.1. Publications

Protein Modeling and Molecular Dynamics Simulation of the Two Novel Surfactant Proteins SP-G and SP-H
Rausch F, Schicht M, Paulsen F, Bräuer L, Brandt W (2014) J Mol Model 20: 2513.

The distribution of human surfactant proteins within the oral cavity and their role during infectious diseases of the gingiva.
Schicht M, Stengl C, Sel S, Heinemann F, Götz W, Petschelt A, Pelka M, Scholz M, Rausch F, Paulsen F, Bräuer L (2014) Ann Anat *in press*.

SFTA3, a novel protein of the lung: 3D-Structure, Characterization and Immune activation
Schicht M and Rausch F, Finotto S, Mathews M, Mattil A, Schubert M, Koch B, Traxdorf M, Bohr C, Worlitzsch D, Brandt W, Garreis F, Sel S, Paulsen F, Bräuer L (2014) Eur Respir J 44: 447-56

Protein modeling and molecular dynamic studies of two new surfactant proteins
Rausch F, Brandt W, Schicht M, Bräuer L, Paulsen F (2013) J Cheminform 5: 2

Molekülsimulation von Surfactant-Proteinen im Special “Trockenes Auge”
Rausch F (2012) Ophthalmologische Nachrichten 12.2012: 13

"SP-G", a putative new surfactant protein - tissue localization and 3D structure
Rausch F and Schicht M, Paulsen F, Ngueya I, Bräuer L, Brandt W (2012) PLoS One 7: e47789.

9.2. Lectures

8th German Conference on Chemoinformatics, Goslar, 2012
“Protein Modeling and Molecular Dynamics Studies of Two New Surfactant Proteins”, 12.11.2012

Workshop on Computer Simulation and Theory of Macromolecules, Hünfeld, 2012
“Protein Modeling and Molecular Dynamics Studies of Two New Surfactant Proteins”, 21.04.2012

26th Molecular Modelling Workshop, Erlangen, 2012
“Protein Modeling and Molecular Dynamics Studies of Two New Surfactant Proteins”, 13.03.2012

Institute of Anatomy II, Friedrich-Alexander University Erlangen-Nuremberg, 2011
Colloquium "Computersimulation von oberflächenaktiven Proteinen", 18.11.2011

Curriculum vitae

Name:	Felix Rausch
Geburtsdatum:	06.01.1985
Geburtsort:	Bad Langensalza

seit 04/2013	**Wissenschaftlicher Mitarbeiter** an der Friedrich-Alexander-Universität Erlangen-Nürnberg, Institut für Anatomie II unter der Leitung von Prof. F. Paulsen, AG Oberflächen-aktive Proteine von Prof. L. Bräuer
10/2009 – 03/2013	**Promotion** am Leibniz-Institut für Pflanzenbiochemie in Halle (Saale), Abteilung Natur- und Wirkstoffchemie von Prof. L. A. Wessjohann, AG Computerchemie von PD Dr. W. Brandt Thema: *„3D modeling of the putative human surfactant proteins SP-G and SP-H and simulations in a pulmonary surfactant model system"*
09/2012	Sicca-Förderpreisträger 2012 des Ressorts Trockenes Auge im Berufsverband der Augenärzte Deutschlands
10/2004 – 09/2009	**Studium der Bioinformatik** an der Martin-Luther-Universität Halle-Wittenberg
09/2009	Abschluss: Diplom-Bioinformatiker Diplomarbeit: *„Detaillierte Untersuchungen zum Katalysemechanismus verschiedener Monoterpensynthasen mittels kombinierter quanten- und molekülmechanischer Methoden"* angefertigt am Leibniz-Institut für Pflanzenbiochemie in Halle (Saale)
08/1994 – 07/2003	**Abitur** Oskar-Gründler-Gymnasium Gebesee

Eidesstattliche Erklärung

Hiermit erkläre ich an Eides statt, dass ich die vorliegende Arbeit selbständig und nur unter Verwendung der angegebenen Hilfsmittel und Quellen angefertigt habe. Die aus den genutzten Werken wörtlich oder inhaltlich entnommenen Stellen wurden als solche gekennzeichnet. Diese Arbeit wurde von mir an keiner anderen wissenschaftlichen Institution vorgelegt.

Erlangen,

F. Rausch