

## SUPPORTING INFORMATION

# Low-Density Lipoprotein: Structure, Dynamics and Lipid – ApoB-100 Interactions

Teemu Murtola, Timo A. Vuorela, Marja T. Hyvönen, Siewert J. Marrink, Mikko Karttunen, and Ilpo Vattulainen

## Part A: Simulation Details and Model Construction

In this part, the studied systems, and in particular, their construction, are presented in detail. This is followed by comparison of the constructed apolipoprotein B-100 structure (before any simulations) to available experimental data. Finally the details of the model (force fields) are described, together with the simulation protocol.

### 1 Model Construction

#### 1.1 Lipid Droplet

The simulated lipid droplet contained 630 1-palmitoyl-2-oleoyl-3-phosphatidylcholine (POPC), 80 1-palmitoyl-3-phosphatidylcholine (denoted as lysoPC in the main article), 600 cholesterol (CHOL), 1600 cholesteryl oleate (CO), and 180 trioleate (TO) molecules. This molecular composition is similar to the average physiological composition seen in experiments [17].

The lipid droplet was constructed by randomly placing 27th part of the lipids in a simulation box, which was then briefly simulated under NpT conditions. This box was then replicated three times in each direction to form a larger cube with all the lipids. The cube was then solvated in water to form the initial configuration for the simulation. We also experimented with completely random distributions of lipids and water, but these systems were slow to equilibrate, failing to form good droplets within a few microseconds. In contrast, the random cubical droplet reorganized into an amphiphilic surface monolayer and a hydrophobic core within 4  $\mu$ s, as described in more detail below.

The system was first equilibrated for 39 ns (effective time, see Section 3.2) with a Berendsen barostat, and then simulated for 1.3  $\mu$ s (effective time). At this point, the lipids had mostly self-organized such that the phospholipids and most of the free cholesterol were on the surface, and trioleate and cholesterol esters formed a core.

However, 216 POPC molecules were trapped inside the droplet, forming several inverted micelles. These were manually removed and placed on the surface of the droplet, and the simulation was continued for 3  $\mu\text{s}$  (effective time) before starting data collection for analysis, followed by 14  $\mu\text{s}$  (effective time) of data collection.

## 1.2 Apolipoprotein B-100

As described in the main article, two alternative models (models S1 and S2) were constructed for the protein part of LDL, i.e., the one instance of apolipoprotein B-100 bound to the particle surface. Model S1 was primarily modeled based on the work by Krisko and Etchebest [24]. For model S2, the first 2000 N-terminal residues were modeled using the homology model of Richardson *et al.* [37] for the first 1000 residues and sequence analysis methods for the latter 1000 residues (see below for details). The C-terminal residues after residue 2000 were modeled as in model S1.

As a first step in the construction, we placed the domains known from modeling on the surface of the lipid droplet based on experimental antibody data [7]. That is, the global layout of the protein on the surface was constructed based on electron microscope studies [7] where one mapped the relative placement of 11 epitopes of apoB-100 on the surface of LDL. Local optimization at domain level followed, including modeling of a disulfide bond between domains 4 and 5 and changes in the shape of domains 1 (model S2 only), 2c, and 8. Then, we constructed the unmodeled regions using input from secondary structure prediction (for model S2, an optimization scheme was also used for residues 1000–2000). Finally, the full model was briefly optimized. The following subsections discuss each of these steps in more detail. All the discussion applies to both models unless specifically indicated otherwise.

The general approach in constructing the models was to use Modeller 9v2 [38] to construct and optimize a rough atomic-scale model for the protein, and then transform the atomic-scale representation to the coarse-grained one in the end. This allowed us to use the well-developed tools available for atomic-scale modeling in the construction process. Visualization and manual tuning of the model was carried out with VMD [20].

During the whole optimization process, the lipid droplet was treated as a soft sphere of radius 10 nm with a weak attraction. In practice, this was achieved with additional constraints in Modeller: for each residue, its center-of-mass was constrained to be  $> 105 \text{ \AA}$  away from the center of the droplet with a standard deviation of 1  $\text{\AA}$ , and  $< 105 \text{ \AA}$  away from the center with a standard deviation of 5  $\text{\AA}$ . In the final optimization, an additional repulsive constraint was placed at 100  $\text{\AA}$  to remove any severe clashes with the lipids. Homology constraints from Modeller were only used if specifically indicated below, otherwise only standard bonded and non-bonded interactions were used together with rigid body restraints. At each optimization step, the default optimization schedule of Modeller was used.

### 1.2.1 Models for Protein Domains

The main ingredient for the models were the computational models for parts of apoB that have been previously published by Richardson *et al.* [37] and Krisko and Etchebest [24]. Richardson *et al.* have modeled the first 1000 residues of apoB, while Krisko and Etchebest divided the protein into eight domains, and constructed a total of ten models for different domains of the protein. Their first domain spans the residues 19–675, and it is nearly identical to the corresponding part of the model by Richardson *et al.*

ApoB-100 has also been analyzed extensively on the sequence level [39–41], and a five-domain structure with alternating  $\alpha$ -helix and  $\beta$ -sheet rich regions has been identified. The secondary structure of the domain models above is in good agreement with the predicted five-domain structure, although there is some mismatch at the boundaries of the domains. Also, most of the regions not modeled by Krisko and Etchebest [24] (453 of the 641 unmodeled residues) are located between residues 675 and 1900, where the sequence analysis predicts the protein to be rich in amphipathic  $\beta$ -sheets [39–41].

The models described in Ref. [24] were kindly provided by A. Krisko, while the model described in Ref. [37] was reproduced with Modeller, using the alignment given in the original reference.

The sequence of domain 3 that we obtained from A. Krisko had a minor deviation from the apoB-100 sequence in the SwissProt database (accession code P04114). To correct for this, we used two different approaches. For model S1, we constructed an alignment manually (there was a single insertion and a single deletion, both within the loop regions of the model), and used the automatic modeling in Modeller to construct the missing loop. For model S2, the domain was modeled with Modeller using the PDB structure 1MOZ [19] as a template, as with the original model [24]. The alignment was constructed with a similar approach as that used by GenThreader [22], i.e., a sequence profile was constructed for the template sequence by finding a set of similar sequences with BLASTP [1], constructing a multiple alignment for them using ClustalW [43], removing all insertions with respect to the template sequence, and finally aligning the domain with this profile, again with ClustalW.

### 1.2.2 Surface Placement of Domains

The first step in constructing the model was to place the modeled domains on the surface of the lipid droplet in reasonable relative positions and orientations. The locations of the individual domains were determined based on electron microscopy experiments where Chatterton *et al.* mapped the average locations of 11 antibody epitopes on the LDL surface [7]. These locations allowed for rough placement of the individual domains on the surface, as well as the general orientation of the C- and N-termini of the domains. The locations of the antibodies were mirrored from those presented in Ref. [7] since this provided a better fit for the shape of several individual domains (most notably, the initial lipovitellin-like region that contains several epitopes, and the curvature of domain 2c of Ref. [24]).

The orientation of the domains, within the limits provided by the antibody locations, were determined based on the hydrophobic patch calculations of Krisko and Etchebest [24]. For domains 4 and 5, the presence of a disulfide bridge between the domains [49] was also taken into account when orienting the domains, placing the cysteine residues close to each other. At this stage, all the domains were treated as rigid bodies, and larger discrepancies with the desired geometry were left to be solved later (see below).

During all subsequent optimization, the information on the location of the antibodies was embedded as weak constraints in the optimization procedure: for each antibody, the location of the center-of-mass of all the residues forming the epitope was constrained to be within 1 nm of the measured position on the surface of the lipid droplet. No constraint was applied on the distance from the center of the droplet.

### 1.2.3 Optimization of Individual Domains

Once the domains were placed on the surface, Modeller was used to optimize the local features to better fit with the geometry of the lipid droplet as well as the locations of the antibody epitopes. These optimization steps are described in detail below. Interactions with other domains were only taken into account in for the lipovitellin-like domain and domain 8, since in the other cases the location being optimized was well isolated from other domains.

**Lipovitellin-like domain (model S2 only)** The homology model of Richardson *et al.* [37] for the first 1000 residues has a large hydrophobic cavity formed of two large  $\beta$ -sheets. For the lipovitellin template, it is well established that phospholipids bind in this cavity [44], and hence it is very likely that these  $\beta$ -sheets are the main lipid-interacting region in this domain [37]. However, the angle that these sheets form is not well suited for interaction with the low curvature of the LDL surface, and it has been proposed that as the lipid droplet grows, the angle between these sheets increases [37]. To better take this into account in the model, we optimized this domain in the presence of the lipids (through the spherical constraints) with homology restraints, which allowed the sheets to bend outwards while the center-of-mass of the domain was forced closer to the surface. To prevent severe clashes with the lipids, the spherical restraints were not applied to the  $\beta$ -barrel region (first 300 residues) because this region protrudes from the surface, and pressing it would force the other regions too deep within the lipid droplet. During this optimization, we also set secondary structure restraints on the unmodeled loop 670–745, which has been proposed to make a helix-turn-helix motif [37]. These restraints were set up based on the PSIPRED secondary structure prediction [23], which predicts the region to be predominantly helical.

**Domain 2c** After orientation by the hydrophobic patches discussed in Ref. [24], the curvature of domain 2c does not fit well with the surface of the sphere. To correct this, we treated the backbone of each helix in the domain as a rigid body, and let the domain relax under the constraints imposed by the lipid droplet (see above).

**Domains 4 and 5** ApoB-100 contains eight disulfide bridges [49]. Of these, seven are located within the first 1000 residues, and are contained in the models. The eighth connects domains 4 and 5 together by joining Cys3167 and Cys3297. To construct this disulfide bridge, the domains were placed such that the cysteines were close to each other and the C-terminus of domain 4 was close to N-terminus of domain 5. The latter is also important because there are no unmodeled residues between the domains. After manual orientation of the domains, the structure was optimized by keeping the backbones of the domains rigid except for 10 terminal residues and 10 residues flanking the cysteines.

**Domain 8** Domain 8 was far too compact to constitute the “bow” described by Chatterton *et al.* [7]. According to the EM measurements, the beginning of the domain is located close to the N-terminal globular region, and the rest of the domain constitutes a long ribbon that bends back and crosses the protein at around residue 3500 [7]. If the region were completely helical, it should consist of a single straight helix between the epitopes of MB43 and BSol16 to span the measured distance [7], and the latter half should also be significantly extended.

To obtain a structure that was consistent with the antibody measurements, we assumed that the secondary structure would be the same as in the original model for the domain, but we moved the helices relative to each other such that the constraints were satisfied. For both regions (between antibodies MB43 and BSol16 and between BSol16 and BSol17), the helices were placed such that the center-of-masses of the helices were on an isocircle connecting the measured antibody locations, and the helices formed a zigzag conformation. For the latter region, the distance from the center of the droplet was adjusted such that the helices were on the solvent side of the other parts of the protein, and that there were no close contacts. The loops between the helices were modeled from scratch with Modeller using linear interpolation and optimization. The structure was then optimized with a similar procedure to domain 2c, with two differences. First, it was not necessary to treat the spacers in this case as the beginning of the domain is placed correctly, and the domain ends at the C-terminus of the protein. Second, interactions with the rest of the protein were taken into account, with an additional weak constraint that Trp4369 would remain close (within 1 nm) to Arg3500. The last constraint was imposed because there is experimental evidence of interaction between these two residues [4].

#### 1.2.4 Construction of Unmodeled Regions

**Initial construction** The regions that were not modeled within any of the domains were first constructed using a simple combination of linear interpolation and optimization. In addition to the standard constraints, secondary structure constraints were used on the unmodeled regions. The secondary structure prediction was performed for the whole protein in overlapping 1500-residue segments using the PSIPRED server [6,23], and the predicted secondary structure was used to make constraints for the unmodeled regions. For model S1, the prediction was used as given by PSIPRED, but for model S2, the secondary structure within the unmodeled part of the  $\beta_1$  domain (residues 1000–2000) was optimized further as described below. The modeled domains were kept fixed during this phase.

**Secondary structure construction** The simple procedure described above results in mostly reasonable structures for helices, but  $\beta$ -strands are not stable in isolation, and the process does not even consider the possibility of forming  $\beta$ -sheets. To produce a more realistic secondary structure, we optimized the structure further. First, the coordinates of each predicted element (a helix or a strand) were modified so that they formed an “ideal” structure. Here, an ideal helix was taken to have  $(\phi, \psi) = (-57.8^\circ, -47.0^\circ)$ , while an ideal strand had  $(\phi, \psi) = (-139^\circ, 135^\circ)$ . The backbone of each element was then constrained as a rigid body, and the structure was reoptimized as above. After the optimization,  $\beta$ -sheets were manually constructed, again assuming an ideal geometry with a 5 Å separation between the strands. For residues 1000–2000 of model S2, the sheet geometry given by the optimization process (see below) was used with manually inserted breaks at points where the structure was visibly strained. For other residues, the sheets were constructed completely manually. In all cases but one (in the spacer between domains 3 and 4), antiparallel sheets seemed more reasonable, and were hence used. For most of the predicted strands, one side was significantly more hydrophobic than the other; care was also taken to orient the hydrophobic side towards the lipid when constructing the sheets. Finally, the structure was optimized with the helices and strands still kept rigid, but now additional constraints were used on the constructed sheets.

**Secondary structure optimization (model S2 only)** For model B, the secondary structure of residues 1000–2000 was constructed based on sequence analysis. Our approach is based on the proposition that this region is rich in amphipathic  $\beta$  strands and that these strands could form a nearly continuous sheet [40]. The predicted structure from PSIPRED was taken as the starting point, and other factors such as amphipathicity were taken into account through a scoring function (described in detail below). The scoring function was maximized using a simulated annealing technique. The annealing program also included a simple method for composing antiparallel  $\beta$  sheets out of individual strands and take them into account in the optimization. Several independent annealing runs were made, and the structure for each residue was taken as the most probable structure in these runs, and this structure was finally subjected to a short annealing run. The final structure has 78% identity with the PSIPRED prediction, and the similarity increased to 87% if the PSIPRED prediction confidence was taken into account, i.e., the sum of the prediction confidences of the correctly predicted locations was 87% of the total sum of the confidences.

The scoring function to determine the “best” secondary structure was constructed to take into account different factors. First, match with the PSIPRED prediction was favored, weighted with the square of the prediction confidence. Also, prolines in strands and helices were disfavored with a small penalty. For sheets and strands, the amphipathicity and average hydrophobicity of the nonpolar face (both calculated as in Ref. [41]) were both favored. Additionally, very long or very short strands and helices were disfavored. Finally, formation of sheets was favored with a small positive score for each hydrogen bond formed (for simplicity, each residue was assumed to have exactly one hydrogen bond to each neighboring strand), and rough edges were disfavored with a small penalty for each unsatisfied hydrogen bond. Only completely antiparallel sheets were considered for simplicity, and strands were placed in the sheets in the same order as they occur in the sequence, i.e., permutations of the strands within the sheet were not considered. The relative weights of these different terms are described in Table 1.

Table 1: Scoring function terms used in secondary structure optimization of residues 1000–2000 of model B.

Structure	Formula	Explanation
matching <sup>a</sup>	$\frac{1}{4}c_i^{2a}$	score for match with PSIPRED
H, E	-10 for each proline	penalty for prolines in structures
H, E	$4 \cdot (m + 2h)^b$	score for high amphiphilicity
H, $l < 6$ or $l > 20^c$	$5 \cdot (l - 6)$ or $20 - l$	penalty for too short/long helices
E, $l < 6$ or $l > 20^c$	$-(l - 6)^2$ or $3 \cdot (20 - l)$	penalty for too short/long strands
each pair of connected strands in sheets <sup>d</sup>	$M - 6 - 2(N_s^2 + N_e^2)$ $-4 l_1 - l_2 $	score for forming sheets

<sup>a</sup> score is 0 if the structure does not match the predicted one.  $c_i \in [0, 9]$  is the prediction confidence of PSIPRED. <sup>b</sup>  $m$  and  $h$  are the hydrophobic moment and the average hydrophobicity of the hydrophobic face of the helix/strand. The values are calculated as in Ref. [41], but are not normalized by the number of residues. <sup>c</sup>  $l$  is the length of the helix/strand. <sup>d</sup>  $M$  is the number of paired residues, and  $N_s$  and  $N_e$  are the number of unpaired residues at the ends.  $l_1$  and  $l_2$  are lengths of the strands.

### 1.2.5 Construction of the Coarse-Grained Model

Finally, the whole structure was optimized with homology constraints to relax the complete structure. After the final optimization, the coordinates for the coarse-grained beads were constructed as the center-of-mass coordinates of the relevant atoms, see Ref. [32] for details of the bead assignment. The VMD *CG Builder* plugin was used for the construction process. The topology file, i.e., bonds, angles, dihedrals, and non-bonded parameters, for the protein was constructed using the Perl script available from <http://md.chem.rug.nl/~marrink/coarsegrain.html>. For numerical stability reasons, the dihedrals generated by the script for  $\beta$ -strands were removed from the topology. Instead of dihedrals, an elastic network model (see below) was used to constrain the conformation of  $\beta$  structures. Post-translational modifications such as glycosylation and palmitoylation to the protein were not included in the model, but see discussion in Section 2.

In addition to the sequence of the protein, the construction of the topology requires the knowledge of the secondary structure of the protein [32]. We assigned the structure with the STRIDE program [10]. The secondary structure was determined for each domain before any optimization to minimize random effects as well as distortion from the constraints, while the secondary structure for the spacers was determined after the final optimization.

$\beta$  sheets were constrained with an elastic network model (ENM) as follows: continuous  $\beta$  sheets were identified using the hydrogen bonding network from STRIDE by putting two strands in the same sheet if there was at least one hydrogen bond between them. For each identified sheet, an ENM was constructed for the backbone beads with a cutoff of 8.0 Å, a force constant of 250 kJ/mol nm<sup>2</sup>, and equilibrium lengths determined from the locations of the  $C_\alpha$  atoms in the atomistic model. The cutoff was chosen such that for each residue there are (on average, based on an ideal sheet geometry) four intrastrand bonds, and three bonds to each neighbouring strand, but no direct bonds to the next-nearest neighbor strands.

### 1.3 Full LDL

After the coarse-grained model for apoB-100 was constructed, it was placed around the equilibrated lipid droplet. Two alternative approaches were taken. In the first one, the protein was placed as constructed, and water particles closer than 5 Å from the protein were removed. In the second, the protein was placed similarly, but all water particles were removed, and a 1 ns molecular dynamics run was performed with the lipids restrained to their initial positions. In such a run, the protein is rapidly attracted to the surface of the droplet, while in a simulation in water this process can be much slower. After this short run in “vacuum” (the actual simulation corresponds more closely to a simulation in an implicit solvent, since the dielectric constant was 15, as in all other CG simulations), the system was solvated as in the first approach, i.e., taking the water particles from the equilibrated lipid-only system and removing any that were closer than 5 Å to the solute. Finally, the system was neutralized with sodium ions, and 150 mM NaCl was added by replacing random water particles with coarse-grained ion particles. The models from the first approach are denoted as S1 and S2 in the main article, while those from the second approach are S1r and S2r.

Each system was then minimized with 1000 steps of steepest descent, followed by a 100 ps simulation with a Berendsen barostat with a 20 ps time constant to relax the solvent. During both phases, everything but the solvent was restrained to their

original positions using harmonic springs with a spring constant of 1000 kJ/mol nm<sup>2</sup>. The system was further equilibrated with a 10 ns simulation without restraints and a Berendsen barostat. This was followed by the production runs.

## 2 Model Validation

In addition to the data used for constructing the apoB-100 model, a substantial amount of experimental data related to apoB-100 structure has been published. Here, we briefly compare our model to data that is most directly related to the atomic-scale structure of the protein. Note that the comparisons here are made before any simulation (with the exception of the short relaxation run in vacuum); results after the simulations are reported separately. Analysis of individual amino acids was carried out for the vacuum-relaxed (see Section 1.3) coarse-grained LDL particles, since the identification of solvent-accessible and lipid-interacting regions is most straightforward in this configuration. However, since the analysis is only carried out for a single configuration in which the domains are not allowed to reorient (during the vacuum simulation, the domains do not rotate significantly), possible changes in the local environment and in the orientation of the domains needs to be taken into account when interpreting the results.

Some additional discussion for the limitations of the model are also given in Part C. In the main paper, we show results for a number of quantities including dynamical properties that are found to be consistent with experimental data.

**Secondary structure** Table 2 shows the secondary structure content of the models, predicted secondary structure from PSIPRED [23] and SSPro [8, 36], and results from experiments that have measured the secondary structure content of apoB-100. The secondary structure of our model is in reasonable agreement with the experimental values, in particular when the variability in the experiments is taken into account.

Table 2: Secondary structure content of constructed apoB-100 models, compared to secondary structure prediction algorithms and experimental values.

System / Method	Helical	Extended	Turn	Coil
S1	34%	25%	27%	14%
S2	29%	31%	27%	14%
PSIPRED	30%	35%	35%	
SSPro	39%	22%	38%	
CD [21]	37%	17%	20%	25%
CD [11]	33%	12%	13%	43%
IR [11]	20%	42%	19%	19%

**Active lysines** NMR measurements indicate that 225 of the total 357 lysines in the protein are available for methylation on the LDL surface, and that of these, 53 are “active”, meaning that their pK values are significantly smaller than for typical lysines [27]. The active lysines were proposed to be located in clusters of basic amino acids, and they are less accessible to larger ions than other exposed lysines [27]. We briefly analyzed the environments of the lysines in our model to see whether they could have similar properties. As noted above, the results are only indicative, and flexibility of the

side chains and rotation of the domains may expose more lysines than revealed in this analysis.

Solvent-accessible surface area (SASA) calculations with a solvent radius of 2.64 Å (half the equilibrium distance between CG particles) show that for model S1 (S2) there are 180 (207) lysines whose sidechain is accessible to the solvent, which compares well with the experimental number of lysines available for methylation. 32 (37) of these are not accessible with a larger solvent radius of 4.5 Å. However, the difference in the average charge of the surrounding residues, calculated between all exposed residues and the 32 (37) that are not accessible to a larger solvent, is minor for both models: for model S1, there is no significant difference, while for model S2, the difference is some 0.3e. Calculating other statistics of the surrounding residues shows that ~ 60 (~ 90) lysines have at least one basic residue in their environment and ~ 25 (both models) have at least two. The average number of residues surrounding a lysine is ~ 7. Above, the surrounding residues were defined as those solvent-accessible residues (defined as for lysines) that were within 1.0 nm of the charged lysine bead.

**Glycosylation sites** Sequence analysis has identified 19 potential glycosylation sites in the apoB-100 sequence [48], and 16 of these were seen to be glycosylated when the protein was sequenced [48]. We have visually checked that in both models S1 and S2, all of these 16 confirmed glycosylation sites are on or near the surface of the protein, and local conformational changes should be able to expose them either to the water phase or to the lipids.

**Palmitoylation site** As a similar check, we also inspected the local conformation around Cys1085, which is palmitoylated in the final protein [50]. In model S1, this residue seems to be buried within domain 2a, but in model S2, it is facing the lipids in a  $\beta$  sheet.

**Receptor/proteoglycan binding sites** *In vitro* studies of delipidated apoB-100 have identified eight clusters of basic amino acids that could bind proteoglycans [47], but on the LDL surface, only one of them (site B-VII, residues 3359–3369) seems to be active [5]. However, this site is absent in apoB-48, which can still bind proteoglycans [9]. It appears that in the truncated protein, another binding site (site B-Ib, residues 84–94) is responsible for the binding, and in the full length protein, this site is masked by the C-terminal part of the protein [9]. These results provide another point of comparison for our model: we have visually inspected the locations and accessibilities of these eight clusters of amino acids. For both models, for sites B-III, B-IV, B-V, and B-VIII there is only limited access to the basic amino acids from the solvent. Sites B-Ia and B-II are partially exposed on the  $\beta$  barrel region of the lipovitellin homology region, but it is possible that the local surroundings do not favor binding, e.g., because of high flexibility. The location of site B-Ib agrees with the experiments: it is accessible to the solvent without the C-terminal region, but the C-terminal region mostly covers it. For model S2, this site is mostly exposed, but changes in the local conformation seem to make it possible for the C-terminal to block the site. Finally, sites B-VI and B-VII are located mostly on the surface of the particle. In our model, some of the basic residues in site B-VII are buried, while site B-VI has all the basic residues exposed.

**Tryptophans** As a final check, we visually inspected the locations of the 37 tryptophan residues in the protein. Tryptophans are unlikely to interact directly with the

solvent, and should most likely be either buried within the protein or be in a position to interact with the interfacial region of the surface lipids [26, 28]. In both models, most tryptophans are reasonably well buried or able to interact with the lipids. There are about five tryptophan residues in both models that are exposed to the solvent farther away from the lipids, but it is possible that once the system is simulated, the domains move to enable also these tryptophans to interact with the lipids.

### 3 Simulation Details

#### 3.1 Force Field

The simulations are based on the MARTINI force field [29, 30, 32], which has been parameterized for lipids and proteins using experimental information such as densities and partitioning coefficients. Briefly, roughly 3 to 4 heavy atoms (2 to 3 for ring-like compounds) are mapped to each coarse-grained bead. The coarse-grained particles are divided into four main groups based on whether the chemical group they describe is charged, polar, non-polar, or apolar, and further division into subgroups is done based on hydrogen bonding characteristics and polar affinity [30]. In total, there are 18 classes of particles. The interactions are described by electrostatics, Lennard-Jones, and bonded terms. Each coarse-grained bead inherits the total charge of its constituent atoms, and the electrostatic interactions are determined from a shifted Coulomb force with a dielectric constant 15. Lennard-Jones interactions are divided into ten levels of varying strength, and the levels have been assigned to the bead types based on, e.g., solvation free energies of model compounds. The bonded interactions are represented by simple harmonic bond and angle terms, and the strengths and equilibrium angles are specific to the system under study, and have been parameterized for several lipids [29, 30] and amino acids [32]. For a complete description of the model and its parameters, please see Refs. [30, 32].

The original MARTINI model includes parameters for all components of our system except for CO and TO. The force fields for these two molecules were constructed using the oleate chain and the cholesterol parameters from the MARTINI model as a starting point. For CO, the bead types were assigned as in cholesterol and an oleate tail, and the connecting ester bead was of type Na. Only one modification was made: the polar bead that contained the hydroxyl group was changed to the non-polar SC1 bead type. For TO, each oleate tail again had the same bead types, and each was connected to a central C1 bead (describing the glycerol backbone) through a Na ester bead. Intramolecular interactions involving the ester beads were tuned to match atomistic simulations of CO and TO [12, 13].

For the protein, secondary structure was modeled as described in Section 1.2.5. For bonds and angles between backbone atoms, as well as dihedral terms for helices, the parameters from Ref. [32] were used. For  $\beta$  structures, no dihedrals were used due to problems with numerical stability; instead, an elastic network model (ENM) was constructed for each  $\beta$ -sheet separately. The ENM was constructed based on the locations of the  $C_\alpha$  atoms in the atomistic model (see Section 1.2.5) with a cutoff of 8.0 Å and a force constant of 250 kJ/mol nm<sup>2</sup>.

### 3.2 Simulation Protocols

The simulations were performed using a development version of GROMACS 4 [3, 16, 25, 45] in the NpT ensemble. The reference temperature for all systems was 310 K, and different molecule types were separately coupled to the heat bath. The pressure was coupled isotropically at a reference pressure of 1 bar. Both Lennard-Jones and electrostatic interactions were cut off at 1.2 nm, with shifting from 0.9 nm for Lennard-Jones and 0 nm for electrostatic interactions [30, 32]. The time step for time integration was 20 fs. Constraints for rings in cholesterol and cholesteryl esters [30] and for ring-like amino acid sidechains [32] were applied with LINCS [14, 15].

During equilibration, i.e., the first 40 ns for the lipid droplet and the first 10 ns for the protein systems, the Berendsen thermostat and barostats were used [2]. If not stated otherwise above, the time constants were 1.0 ps for both the thermostat and the barostat.

After the equilibration runs, the thermostat was switched to the Nosé-Hoover [18, 33] thermostat, and the barostat to the Parrinello-Rahman [34, 35] barostat. The time constant was 1.0 ps for both algorithms.

The dynamics for the coarse-grained model is faster than that of an atomic model by a factor of 2 to 10 [30]. As a first approximation, the time axis can simply be scaled to obtain “real” dynamics. The standard factor for this scaling with the present model is 4, obtained from comparing the diffusion of water in the coarse-grained and in atomistic systems [30]. In this article, all simulation times are reported in as effective time unless otherwise stated, whereas the simulation parameters and lengths of short equilibration runs (< 100 ns) are in simulation times.

## Part B: Analysis Methods

This part gives details of the analysis methods used in the main manuscript.

### 4 Coverage of apoB-100

Several different quantities are reported separately for a section of the LDL under and not under apoB-100. To determine whether a point was under apoB-100, we constructed a vector from the center of mass of the lipids to the point of interest, and compared this vector to vectors from the center of mass of the lipids to each protein bead. If the vector of interest made an angle less than  $2^\circ$  from any of the protein vectors, it was deemed as being under the protein. The  $2^\circ$  translates into an approximate distance of 3 Å at the surface of the droplet. For typical regions in the protein, these cones corresponding to different protein beads overlap, with the result that the region essentially becomes a solid angle that is spanned by the protein.

The same method was also used to calculate the fraction of surface area covered by apoB-100. In this case, the spherical surface was divided into small squares with the longest edges corresponding to approximately  $1^\circ$ , resulting in approximately 18 000 squares. For each square, it was determined whether it overlaps with any of the  $2^\circ$  cones centered at the protein atoms. If the complete square was covered by a single cone, the full area was taken as covered by the protein. In the case that no single cone covered the whole square, but at least one cone overlapped with the square, half of its area was taken as covered by the protein. Although this calculation is not exact, the error should be small as partially covered squares only appear close to the edges

of the protein. Further, at least most of the error averages out because for a random distribution, the average coverage for the partially covered squares should be close to half of its total area because the squares are quite small.

## 5 Radial Density Distributions

Simple number densities of simulation beads were used for radial density distributions: the distance from the center of mass of the lipids to each simulation bead was calculated for each frame, and a histogram was created from the distances. For each bin, the number of beads was then normalized by the volume of the bin. For analysing the region under the protein, each bead was treated individually (i.e., a part of a molecule can be under apoB-100 while another part is not). Further, the area covered by the protein was taken into account when calculating the average volumes of bins for normalization.

## 6 Cholesterol Ring Order Parameters

Cholesterol ring order parameters were calculated as

$$S = \frac{1}{2} \langle 3 \cos^2 \theta - 1 \rangle \quad (1)$$

where  $\theta$  is the angle between the director of the cholesterol ring and the local normal. The director of the ring is defined as a vector pointing from the ring bead to which the short tail is connected to the ring bead that includes the hydroxyl oxygen. The local normal is just the vector from the center of mass of the lipids to the center of the director. The distance from the center of the droplet was also measured to the center of the director. The radial distance was divided into bins, and within each bin, the average order parameter was calculated to arrive at the order parameter versus radial distance plots. For analyzing the region under apoB-100, the molecule was treated as being under apoB-100 if the center of the director fulfilled the criterion.

## 7 Diffusion Coefficients

Diffusion coefficients were calculated using jump length distributions: a histogram of the displacements of the centers of mass of the selected molecules were calculated, calculating the displacements over a time scale  $t$ . The obtained distribution was then fitted to the theoretical curves for two- and three-dimensional random walkers, i.e., to the Gaussian distributions

$$P_{2d}(d; t) = \frac{r}{2Dt} \exp(-r^2/4Dt) \quad (2)$$

and

$$P_{3d}(d; t) = \frac{4\pi r^2}{(4\pi Dt)^{3/2}} \exp(-r^2/4Dt). \quad (3)$$

The total translation of the center of mass of the whole droplet was removed before calculating the diffusion coefficients. For the lipids in the surface monolayer (POPC, lysoPC, CHOL), the 2D curve fits better, while for the core lipids (CO and TO), the 3D curve is more suitable, in agreement with the intuitive idea of how the molecules move.

The time scale of  $t = 200$  ns was used to obtain the values in the main article, but other values were also tested, and they gave qualitatively similar results. However, at much shorter time scales the behavior starts to deviate from a random walk due to interactions with other molecules, while at longer time scales the limited size of the droplet causes deviations. The time scale of 200 ns was chosen as a reasonable compromise between these limits.

To analyze diffusion under apoB-100 and/or within a certain distance from the droplet center, a single jump for a molecule was treated as being (not being) under the protein if either end position fulfilled (did not fulfill) the criterion.

Errors were estimated by partitioning the trajectory into four parts, calculating the diffusion coefficients independently for each part, and calculating the standard error of the mean from these four data points.

## 8 Contacts Between Protein and Lipids

The number of contacts between two groups of beads was calculated by finding the pairs of beads (one from each group) whose distance from one another was less than 8 Å. For contacts between individual amino acids and a certain lipid class, this number was then divided by the number of how many of that amino acid are contained in apoB-100.

## Part C: Extra Data on Validation, Equilibration and Diffusion

First, let us briefly discuss the validity and limitations of our model. The MARTINI lipids have been widely applied and shown to reproduce experimental properties in many contexts [31] and references therein). Also, in the LD simulation the lipids spontaneously self-organize into a surface monolayer and a hydrophobic core. The protein remains stable in all the simulations. Fixed bond potentials are only used to hold individual helices and  $\beta$ -sheets together, while the tertiary structure is held together only by non-bonded interactions that have been parameterized on partitioning data. Hence, the stability of the protein structure that we observed in our simulations implies that the local environment favors the present protein structure. However, this does not imply that the structure is correct, only that it is feasible. Hence, any detailed studies of the protein itself would be highly speculative. This is in contrast to HDL simulations (see, e.g., refs. [42, 46]) where the protein (apolipoprotein A-1) consists of a single  $\alpha$ -helix and is much smaller. Despite the limitations from the size of the protein and uncertainty in its structure, partitioning and large-scale properties are expected to be correct because of the way the interactions are parameterized. This is the central reason why in this work we mostly focus on the generic features of protein-lipid interaction and the distribution of lipids in LDL. Concerning the latter, all the simulations show consistent long-time behavior, e.g., in how the lipids distribute.

Figure 1 shows the fraction of molecules located on the LDL surface as a function of time, calculated as the number of molecules on the surface divided by the total number of molecules. A molecule is defined to be on the surface if its distance from the center of mass of the lipids is larger than 7.5 nm. Each panel shows the numbers for

a single molecule species. POPC and lysoPC are not shown because all the molecules are on the surface.

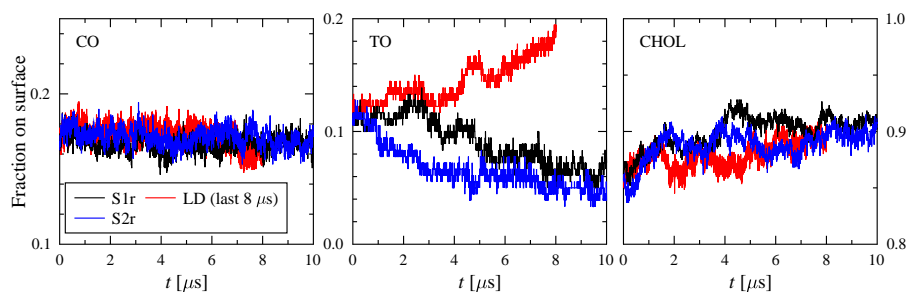


Figure 1: Fraction of molecules on LDL surface as a function of time. Different molecule species are shown in different panels, and line color distinguishes the different simulations. A molecule was defined to be on the surface if its center of mass was further than 7.5 nm away from the center of mass of all the lipids.

Figure 2 shows the ellipticity of the lipid droplet as a function of time for different simulations. The ellipticity is defined as  $\sqrt{1 - b^2/a^2}$ , where  $a$  and  $b$  are the longest and shortest radii of gyration along the principal axes of inertia.

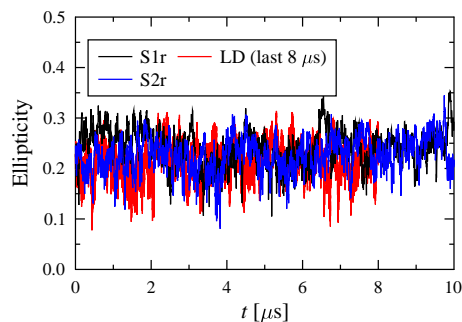


Figure 2: Ellipticity of the lipid droplet as a function of time. Different lines correspond to different simulations.

Figure 3 shows the fraction of LDL surface covered by apoB-100 as a function of time for the simulations S1r and S2r.

Figure 4 shows representative jump length distributions for different groups of molecules, together with the best fits to the theoretical curves for random walks. Each panel shows distributions for one type of molecules, with different colors corresponding to molecules in different parts of the droplet. Dots show the measured distributions, and for each data set the solid line of the same color shows the theoretical fit. The diffusion coefficients associated with the theoretical curves can be found in the table in the main article.

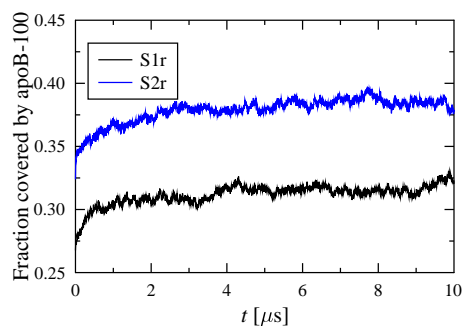


Figure 3: Fraction of LDL surface covered by apoB-100 as a function of time.

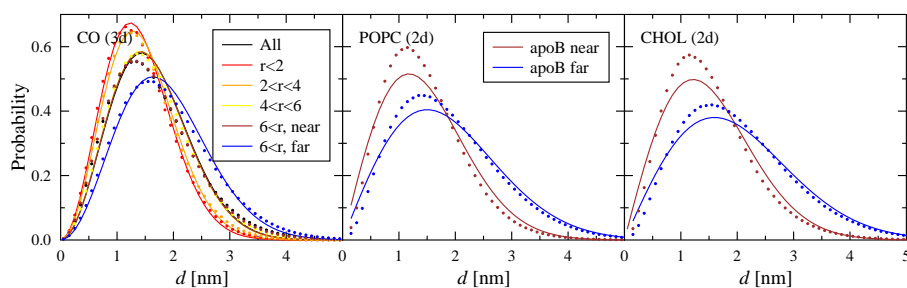


Figure 4: Jump length distributions for different groups of molecules. Each panel corresponds to one molecule type, and each color to molecules within a certain region of LDL. Dots show the measured distributions, while solid lines show the best theoretical fits to the corresponding data.

## References

- [1] S. E. Altschul, W. Gush, W. Miller, E. W. Myers, and D. J. Lipman. Basic local alignment search tool. *J. Mol. Biol.*, 215:403–310, 1990.
- [2] H. J. C. Berendsen, J. P. M. Postma, A. DiNola, and J. R. Haak. Molecular dynamics with coupling to an external bath. *J. Chem. Phys.*, 81:3684–3690, 1984.
- [3] H. J. C. Berendsen, D. van der Spoel, and R. van Drunen. GROMACS: A message-passing parallel molecular dynamics implementation. *Comp. Phys. Comm.*, 91:43–56, 1995.
- [4] J. Borén, U. Ekström, B. Ågren, P. Nilsson-Ehle, and T. L. Innerarity. The molecular mechanism for the genetic disorder familial defective apolipoprotein B100. *J. Biol. Chem.*, 276:9214–9218, 2001.
- [5] J. Borén, K. Olin, I. Lee, A. Chait, T. N. Wight, and T. L. Innerarity. Identification of the principal proteoglycan-binding site in LDL. *J. Clin. Invest.*, 101:2658–2664, 1998.

- [6] K. Bryson, L. J. McGuffin, R. L. Marsden, J. J. Ward, J. S. Sodhi, and D. T. Jones. Protein structure prediction servers at University College London. *Nucleic Acids Res.*, 33:W36–38, 2005.
- [7] J. E. Chatterton, M. L. Phillips, L. K. Curtiss, R. Milne, J.-C. Fruchart, and V. N. Schumaker. Immunoelectron microscopy of low density lipoproteins yields a ribbon and a bow model for the conformation of apolipoprotein B on the lipoprotein surface. *J. Lipid Res.*, 36:2027–2037, 1995.
- [8] J. Cheng, A. Randall, M. Swerodoski, and P. Baldi. SCRATCH: A protein structure and structural feature prediction server. *Nucleic Acids Res.*, 33:W72–76, 2005.
- [9] C. Flood, M. Gustafsson, P. E. Richardson, S. C. Harvey, J. P. Segrest, and J. Borén. Identification of the proteoglycan binding site in apolipoprotein B48. *J. Biol. Chem.*, 277:32228–32233, 2002.
- [10] D. Frishman and P. Argos. Knowledge-based secondary structure assignment. *Proteins: Struct. Funct. Genet.*, 23:566–579, 1995.
- [11] E. Goormaghtigh, J. D. Meutter, B. Vanloo, R. Brassuer, M. Rosseneu, and J.-M. Ruyschaert. Evaluation of the secondary structure of apo B-100 in low-density lipoprotein (LDL) by infrared spectroscopy. *Biochim. Biophys. Acta*, 1006:147–150, 1989.
- [12] A. Hall, J. Repakova, and I. Vattulainen. Modeling of the triglyceride-rich core in lipoprotein particles. *J. Phys. Chem. B*, 112:13772–13782, 2008.
- [13] M. Heikelä, I. Vattulainen, and M. T. Hyvönen. Atomistic simulation studies of cholesteryl oleates: Model for the core of lipoprotein particles. *Biophys. J.*, 90:2247–2257, 2006.
- [14] B. Hess. P-LINCS: A parallel linear constraint solver for molecular simulation. *J. Chem. Theory Comput.*, 4:116–122, 2008.
- [15] B. Hess, H. Bekker, H. J. C. Berendsen, and J. G. E. M. Fraaije. LINCS: A linear constraint solver for molecular simulations. *J. Comp. Chem.*, 18:1463–1472, 1997.
- [16] B. Hess, C. Kutzner, D. van der Spoel, and E. Lindahl. GROMACS 4: Algorithms for highly efficient, load-balanced, and scalable molecular simulation. *J. Chem. Theory Comput.*, 4:435–447, 2008.
- [17] T. Hevonoja, M. O. Pentikäinen, M. T. Hyvönen, P. T. Kovanen, and M. Ala-Korpela. Structure of low density lipoprotein (LDL) particles: Basis for understanding molecular changes in modified LDL. *Biochim. Biophys. Acta*, 1488:189–210, 2000.
- [18] W. G. Hoover. Canonical dynamics: Equilibrium phase-space distribution. *Phys. Rev. A*, 31:1695–1697, 1985.
- [19] E. G. Huizinga, S. Tsuji, R. A. Romijn, M. E. Schiphorst, P. G. de Groot, J. J. Sixma, and P. Gros. Structures of glycoprotein Ib $\alpha$  and its complex with von Willebrand factor A1 domain. *Science*, 297:1176–1179, 2002.

- [20] W. Humphrey, A. Dalke, and K. Schulten. VMD – Visual Molecular Dynamics. *J. Molec. Graphics*, 14:33–38, 1996.
- [21] A. Johs, M. Hammel, I. Waldner, R. P. May, P. Laggner, and R. Prassl. Modular structure of solubilized human apolipoprotein B-100. *J. Biol. Chem.*, 281:19732–19739, 2006.
- [22] D. T. Jones. GenTHREADER: An efficient and reliable protein fold recognition method for genomic sequences. *J. Mol. Biol.*, 287:797–815, 1999.
- [23] D. T. Jones. Protein secondary structure prediction based on position-specific scoring matrices. *J. Mol. Biol.*, 292:195–202, 1999.
- [24] A. Krisko and C. Etchebest. Theoretical model of human apolipoprotein B100 tertiary structure. *Proteins: Struct. Funct. Bio.*, 66:342–358, 2007.
- [25] E. Lindahl, B. Hess, and D. van der Spoel. GROMACS 3.0: A package for molecular simulation and trajectory analysis. *J. Mol. Mod.*, 7:306–317, 2001.
- [26] W. Liu and M. Caffrey. Interactions of tryptophan, tryptophan peptides, and tryptophan alkyl esters at curved membrane interfaces. *Biochemistry*, 45:11713–11726, 2006.
- [27] S. Lund-Katz, J. A. Ibdah, J. Y. Letizia, M. F. Thomas, and M. C. Phillips. A  $^{13}\text{C}$  NMR characterization of lysine residues in apolipoprotein B and their role in binding to the low density lipoprotein receptor. *J. Biol. Chem.*, 263:13831–13838, 1988.
- [28] J. L. MacCallum, W. F. D. Bennett, and D. P. Tieleman. Distribution of amino acids in a lipid bilayer from computer simulations. *Biophys. J.*, 94:3393–3404, 2008.
- [29] S. J. Marrink, A. H. de Vries, and A. E. Mark. Coarse grained model for semi-quantitative lipid simulations. *J. Phys. Chem. B*, 108:750–760, 2004.
- [30] S. J. Marrink, H. J. Risselada, S. Yefimov, D. P. Tieleman, and A. H. de Vries. The MARTINI force field: Coarse grained model for biomolecular simulations. *J. Phys. Chem. B*, 111:7812–7824, 2007.
- [31] J. R. McNamara, D. M. Small, Z. Li, and E. J. Schaefer. Differences in LDL subspecies involve alterations in lipid composition and conformational changes in apolipoprotein b. *J. Lipid Res.*, 37:1924, 1996.
- [32] L. Monticelli, S. K. Kandasamy, X. Periole, R. G. Larson, D. P. Tieleman, and S.-J. Marrink. The MARTINI coarse-grained force field: Extension to proteins. *J. Chem. Theory Comput.*, 4:819–834, 2008.
- [33] S. Nosé. A molecular dynamics method for simulations in the canonical ensemble. *Mol. Phys.*, 52:255–268, 1984.
- [34] S. Nosé and M. L. Klein. Constant pressure molecular dynamics for molecular systems. *Mol. Phys.*, 50:1055–1076, 1983.
- [35] M. Parrinello and A. Rahman. Polymorphic transitions in single crystals: A new molecular dynamics method. *J. Appl. Phys.*, 52:7182–7190, 1981.

- [36] G. Pollastri, D. Przybulski, B. Rost, and P. Baldi. Improving the prediction of protein secondary structure in three and eight classes using recurrent neural networks and profiles. *Proteins: Struct. Funct. Genet.*, 47:228–235, 2002.
- [37] P. E. Richardson, M. Manchekar, N. Dashti, M. K. Jones, A. Beigneux, S. G. Yong, S. C. Harvey, and J. P. Segrest. Assembly of lipoprotein particles containing apolipoprotein-B: Structural model for the nascent lipoprotein particle. *Biophys. J.*, 88:2789–2800, 2005.
- [38] A. Sali and T. L. Blundell. Comparative protein modelling by satisfaction of spatial restraints. *J. Mol. Biol.*, 234:779–815, 1993.
- [39] J. P. Segrest, M. K. Jones, and N. Dashti. N-terminal domain of apolipoprotein B has structural homology to lipovitellin and microsomal triglyceride transfer protein: A “lipid pocket” model for self-assembly of apoB-containing lipoprotein particles. *J. Lipid Res.*, 40:1401–1416, 1999.
- [40] J. P. Segrest, M. K. Jones, H. D. Loof, and N. Dashti. Structure of apolipoprotein B-100 in low density lipoproteins. *J. Lipid Res.*, 42:1346–1367, 2001.
- [41] J. P. Segrest, M. K. Jones, V. K. Mishra, G. M. Anantharamaiah, and D. W. Garber. ApoB-100 has a pentapartite structure composed of three amphipathic  $\alpha$ -helical domains alternating with two amphipathic  $\beta$ -strand domains. Detection by the computer program LOCATE. *Arterioscler. Thromb.*, 14:1674–1685, 1994.
- [42] A. Y. Shih, I. G. Denisov, and J. C. Phillips. Molecular dynamics simulations of discoidal bilayers assembled from truncated human lipoproteins. *Biophys. J.*, 88:548, 2005.
- [43] J. D. Thompson, D. G. Higgins, and T. J. Gibson. CLUSTAL W: Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, positions-specific gap penalties and weight matrix choice. *Nucleic Acids Res.*, 22:4673–4680, 1994.
- [44] J. R. Thompson and L. J. Banaszak. Lipid-protein interactions in lipovitellin. *Biochemistry*, 41:9398–9409, 2002.
- [45] D. van der Spoel, E. Lindahl, B. Hess, G. Groenhof, A. E. Mark, and H. J. C. Berendsen. GROMACS: Fast, flexible and free. *J. Comp. Chem.*, 26:1701–1719, 2005.
- [46] T. Vuorela, A. Catte, P. S. Niemela, A. Hall, M. T. Hyvonen, S. J. Marrink, M. Karttunen, and I. Vattulainen. Role of lipids in spheroidal high density lipoproteins. *PLoS Comput. Biol.*, 6:e1000964, 2010.
- [47] K. H. Weisgraber and S. C. Rall, Jr. Human apolipoprotein B-100 heparin-binding sites. *J. Biol. Chem.*, 262:11097–11103, 1987.
- [48] C. Y. Yang, Z. W. Gu, S. A. Weng, T. K. Kim, S. H. Chen, H. J. Pownall, P. M. Sharp, S. W. Liu, W. H. Li, and A. M. Gorro, Jr. Structure of apolipoprotein B-100 of human low density lipoproteins. *Arterioscler. Thromb.*, 9:96–108, 1989.
- [49] C.-Y. Yang, T. W. Kim, S.-A. Weng, B. Lee, M. Yang, and A. M. Gorro, Jr. Isolation and characterization of sulfhydryl and disulfide peptides of human apolipoprotein B-100. *Proc. Natl. Acad. Sci. USA*, 87:5523–5527, 1990.

- [50] Y. Zhao, J. B. McCabe, J. Vance, and L. G. Berthiaume. Palmitoylation of apolipoprotein B is required for proper intracellular sorting and transport of cholesteryl esters and triglycerides. *Mol. Biol. Cell*, 11:721–734, 2000.