Inverse design of cholesterol attracting transmembrane helices reveals a paradoxical role of hydrophobic length

Jeroen Methorst[a], **Niek van Hilten**[a], **and Herre Jelger Risselada**[a,b,*]

[a]Leiden Institute of Chemistry, Leiden University, The Netherlands; [b]Institute of Theoretical Physics, Georg-August University, Göttingen, Germany

This manuscript was compiled on July 5, 2021

The occurrence of linear cholesterol-recognition motifs in alpha-helical transmembrane domains has long been debated. Here, we demonstrate the ability of a genetic algorithm guided by coarse-grained molecular dynamics simulations—a method coined evolutionary molecular dynamics (evo-MD)—to directly resolve the sequence which maximally attracts/sorts cholesterol within a single-pass alpha-helical transmembrane domain (TMDs). We illustrate that the evolutionary landscape of cholesterol attraction in membrane proteins is characterized by a sharp, well-defined global optimum. Surprisingly, this optimal solution features an unusual short hydrophobic block, consisting of typically only eight short chain hydrophobic amino acids, surrounded by three successive lysines. Owing to the membrane thickening effect of cholesterol, cholesterol-enriched ordered phases favor TMDs characterized by a long rather than a short hydrophobic length. However, this short hydrophobic pattern evidently offers a pronounced net advantage for the binding of free cholesterol in both coarse-grained and atomistic simulations. Attraction is mediated by the unique ability of cholesterol to snorkel within the hydrophobic core of the membrane and thereby shield deeply located lysines from the unfavorable hydrophobic surrounding. Since this mechanism of attraction is of a thermodynamic nature and is not based on molecular shape specificity, a large diversity of sub-optimal cholesterol attracting sequences can exist. The puzzling sequence variability of proposed linear cholesterol-recognition motifs is thus consistent with sub-optimal, unspecific binding of cholesterol. Importantly, since evo-MD uniquely enables the targeted design of recognition motifs for distinct fluid lipid membranes, we foresee wide applications for evo-MD in the biological and biomedical fields.

genetic algorithm | molecular dynamics | artificial intelligence | evo-MD

Significance Statement

Our work demonstrates how a synergy between evolutionary algorithms and high-throughput coarse-grained molecular dynamics can yield fundamentally new insights into the evolutionary fingerprints of protein-mediated lipid sorting. We illustrate that the evolutionary landscape of cholesterol attraction in isolated transmembrane domains is characterized by a sharp, well-defined global optimum. In contrast, sub-optimal attraction of cholesterol is associated with a diverse solution space and features a high sequence variability despite binding the same unique molecule. The contrasting physicochemical nature of the resolved attraction optimum suggests that cholesterol attraction/binding via linear motifs does not pose a dominant pressure/constrain on the evolution of transmembrane proteins.

*To whom correspondence should be addressed. E-mail: hrissel@gwdg.de

Cholesterol serves as a major constituent of the mammalian plasma membrane. The overall fraction of cholesterol in the plasma membrane relative to total plasma membrane lipids is about 30 to 40 % in leukocytes, epithelial cells, neurons, and mesenchymal cells (1). The localization, trafficking, and functionality of membrane proteins involved in cholesterol-dependent pathways and cholesterol homeostasis may critically rely on their ability to attract and bind cholesterol molecules (2–10). Prediction of cholesterol binding affinity could therefore illuminate their role in diseases that are characterized by loss of cholesterol homeostasis (e.g. neurological diseases and cancer (11)), and pave the road for novel drug targets and strategies (6, 12–16). A compelling amount of data obtained by bioinformatic approaches, molecular modeling and simulations, and experiments have suggested the existence of cholesterol recognition amino acid consensus motifs (CRAC motifs) (3, 4, 17, 18), as well as its inverse CARC motif, in various membrane protein families such as, for example, viral membrane proteins (e.g. (12, 15)), ion channels (e.g. (19, 20)), and G protein-coupled receptors (GPCRs)—the most intensively studied drug target family (e.g. (6, 21–24)). However, the looseness of the CRAC definition, being a rather flexible algorithmic rule: (L/V)-X_{1-5}-(Y)-X_{1-5}-(K/R), is rather unexpected for a motif that mediates binding to a unique molecule, raising skepticism about its predictive value (3). Interestingly, a recent fluorescence microscopy study on model membranes suggested that 'slender' transmembrane domains (TMDs) comprised of hydrophobic amino-acids with short side chains such as, for example, leucine (L) or valine (V), preferentially bind to the cholesterol-enriched interface of liquid order (Lo) domains (5). These amino acids also play a vital role in the CRAC motif. The observation that attraction toward cholesterol-enriched phases strongly correlates to the accessible surface area of TMDs suggests that cholesterol recognition is not mediated by precise matching of molecular shape compatibility, as is common in protein-ligand docking, but is likely additionally mediated by alternative thermodynamic driving forces. This would explain why these amino acids occur on variable positions within the CRAC definition.

Evolutionary Molecular Dynamics (evo-MD). High-throughput screening of transmembrane sequences could provide an effective strategy to resolve the chemical and structural specificities which underpin cholesterol recognition and binding, as well as the relevant thermodynamic driving forces. However, the accessible chemical space of transmembrane domains is astronomical (about 20^{20} possibilities), warranting the employment of smart search strategies.

Directed evolution is a method used in protein engineering that mimics the process of natural selection to steer proteins or nucleic acids toward a user-defined goal (25). Evolutionary inverse design strategies see applications in a variety of fields due to their efficient exploration of search-space (26). These methods fall under the scope of reinforcement learning, adapting processes for optimal performance by reinforcing desired behavior (27). Of special interest are the genetic algorithms (GA), which model the mechanisms of darwinistic evolution in a computational algorithm, utilizing genetic elements such as recombination, cross-over, mutation, selection, and fitness (28). Since simulation of evolution is both time and labor intensive, it can quickly become intractable in a laboratory setting thereby limiting its application. In such a scenario, molecular dynamics (MD) simulations may provide an alternate *in silico* route for the high-throughput virtual screening of chemical space.

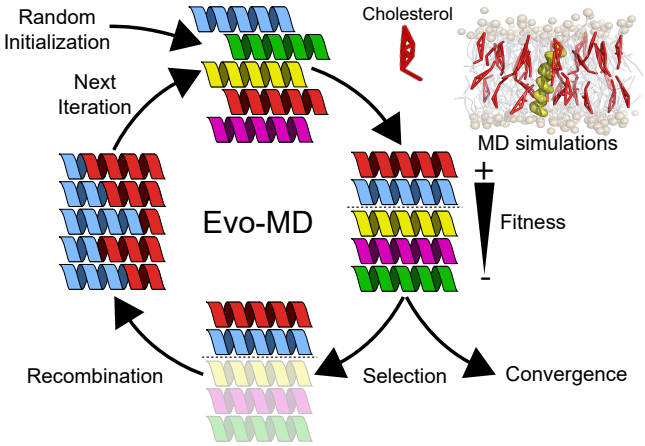

Fig. 1. Illustration of the basic concept of evolutionary molecular dynamics (evo-MD). Random peptide sequences self-evolve into optimal cholesterol attracting transmembrane domains in the coarse of the evolution. Generated peptides are iteratively ranked upon increasing fitness as being determined via ensemble averaging within molecular dynamics simulations.

Here, we demonstrate the ability of GAs guided by coarse-grained MD simulations—a method which we coin evolutionary molecular dynamics (evo-MD)—to yield unique insights into the driving forces that underpin ligand recognition. Evo-MD effectively reduces the search for optimal ligand consensus motifs to solving a variational problem in high dimensional chemical space. While many implementations of GAs have had abstract purposes—often used as a mathematical rather than a biological approach—we effectively create an *in silico* molecular cell membrane environment subject to natural evolution. To this aim, we introduce EVOMD, a highly parallel software package for evolutionary molecular dynamics simulations that incorporates the GROMACS molecular dynamics engine into a custom, Python-based GA wrapper. EVOMD can adapt all elements of a simulation, be it structural (e.g. atoms, molecules), topological, or simulation parameters (e.g. force-field parameters), based on a reinforcement value measured during the simulation.

In this work, we apply evo-MD to explore the evolutionary landscape of cholesterol attraction for a transmembrane domain sequence with a *fixed* length of 20 amino acids. We illustrate that cholesterol attraction is mediated by enforcing snorkeling of lipid head groups in the direct vicinity of the TMD, which energetically favors cholesterol over POPC lipids. In native membrane proteins, the presence of additional protein-protein interactions likely poses a far larger evolutionary constraint on sequence than protein-lipid interactions. Our observation of a single, well-defined minimum within sequence space suggests that cholesterol attraction in native TMDs is only sub-optimal, and which enables cholesterol recognition motifs to be of a highly variable nature.

Results

Cholesterol sensing features a strong evolutionary conservation. Artificial evolution is simulated in a system consisting of a 30% cholesterol and 70% 1-palmitoyl-2-oleoyl-glycero-3-

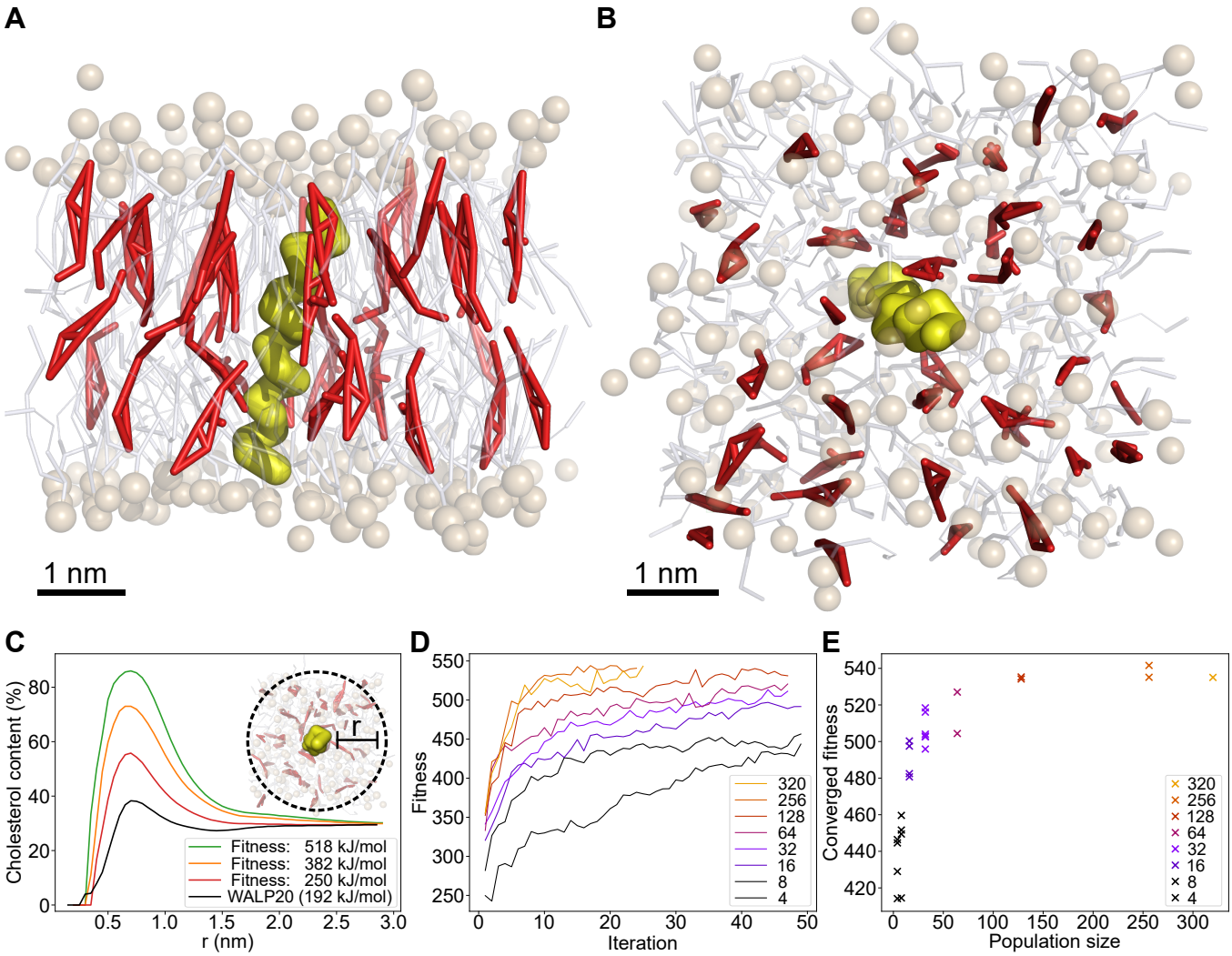

Fig. 2. Evolutionary molecular dynamics simulations of a cholesterol attracting transmembrane protein. (A, B): Snapshots of a transmembrane protein (yellow) embedded within a POPC (white/brown) membrane containing 30% cholesterol (red). (C): Ratio of the cholesterol content in a local radius around the protein (see methods). An increase in fitness correlates to an increase in local cholesterol. (D): Fitness development during protein evolution, shown for various population sizes. The fitness is expressed in terms of peptide-cholesterol interaction energy (Lennard-Jones energy). Fitness increases with GA iterations. Size of the population affects the height of the fitness plateau. (E): GA converges to different fitness values, depending on the size of the populations. Eventually, evolution converges to an optimal solution at population sizes greater than 128 individuals.

phosphocholine (POPC) membrane containing a single, 20 amino acid long peptide sequence positioned transversely through the membrane. Owing to the symmetry of the here-studied bilayer, generated sequences are mirror symmetric, i.e. only the first ten amino acids are independently chosen. Evolution is directed towards peptide sequences that increase the local density of cholesterol, visualized by the percentage cholesterol content of the membrane within a certain range from the peptide (see Fig. 2C). In practice, this is obtained by maximizing the ensemble-averaged non-bonded interaction energy between the peptide and cholesterol, i.e. this defines the *fitness*, in the course of sequence evolution.

Despite starting from *random* peptide sequences, the observed evolution eventually converges to an optimum, as being characterized by a plateau in fitness value. Convergence of genetic algorithms depends on a variety of factors, most notably the size of the population—which directly correlates to the area of the search space that is sampled each iteration, and the number of iterations that are performed. Either parameter requires some minimum value for convergence to occur. The population size should be large enough (in combination with mutation rate and other diversifying factors) to prevent premature convergence to sub-optimal solutions, and, with the evolutionary process taking place between iterations, a certain number of iterations is necessary. Ideally, both parameters are chosen as large as possible.

To assess whether the convergence of evolution is either sub-optimal (i.e, a local solution) or optimal (i.e., a global solution), we conducted a set of evolutionary runs with population sizes ranging from 4 to 320 individuals until no further convergence of fitness was observed. Figure 2D shows how the fitness of the best performing sequences changes with each generation. As expected, increasing population size increases the optimum fitness, as evident from a higher plateau value reached after convergence of fitness (see Fig. 2E). This increase in optimal fitness leveled off once the population size began exceeding

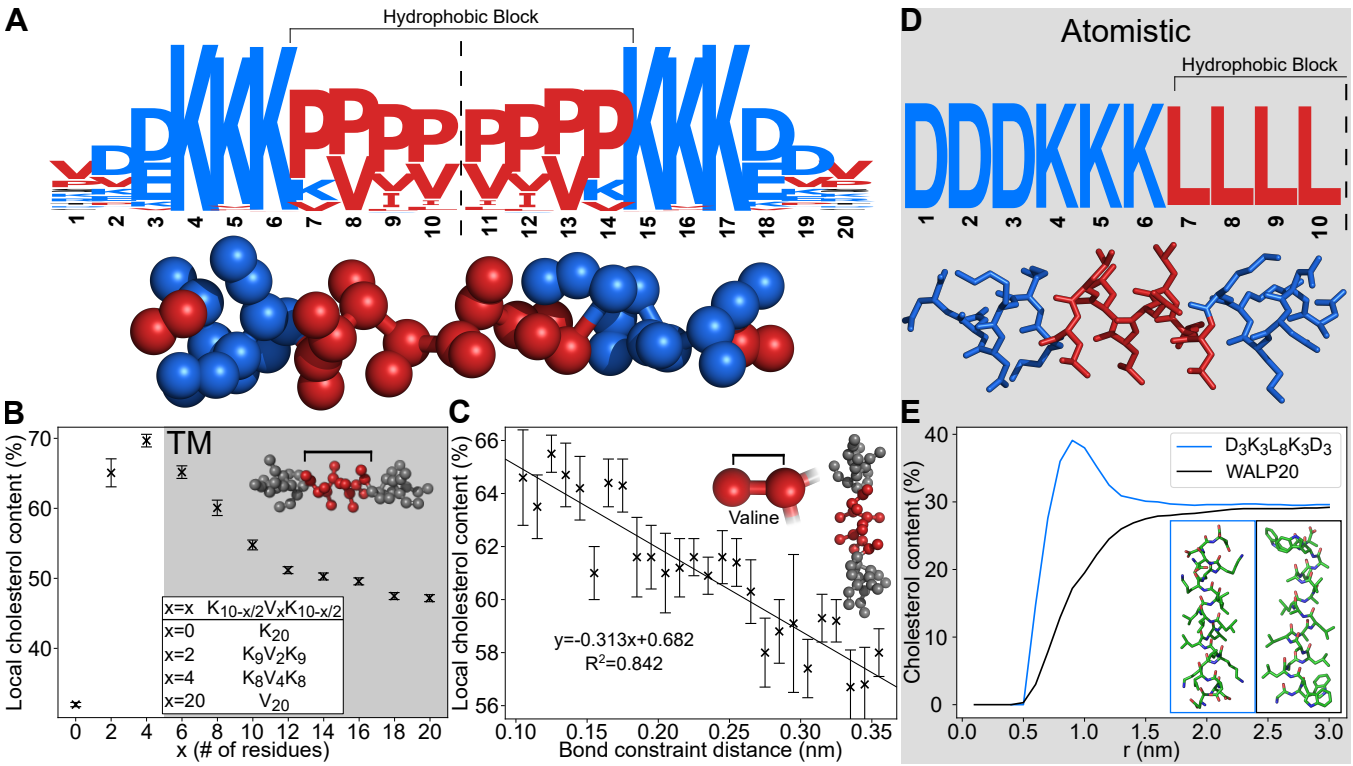

Fig. 3. Sequence and chemical features of the optimal cholesterol attractor. (A): Sequence logo computed from all high-fitness (≥ 500 kJ/mol) peptide sequences reveals a highly conserved hydrophobic pattern (red = hydrophobic; blue = hydrophilic). Owing to the symmetry of the here-used bilayer, sequences are mirrored around the center as indicated by the dashed line. (B): Short, hydrophobic blocks result in high local (1.0 nm radius) cholesterol composition of the membrane. Sequences adhere to the following motif: $K_{(10-x/2)}$–V_x–$K_{(10-x/2)}$ (x=0, 2, 4 etc.). The gray region indicates the regime where a transmembrane (TM) topology is thermodynamically stable (see Fig. S9). Bars represent the standard error of the mean. (C): The cross-sectional thickness of proteins markedly affects the local cholesterol content (defined at a 1.0 nm radius) of the membrane surrounding the protein. Force field was adapted to vary the distance between valine backbone and sidechain beads in a dummy peptide (sequence: $K_6V_8K_6$). Bars represent the standard error of the mean. (D): Atomistic sequence of a cholesterol sensor. (E): A rationally designed motif $D_3K_3L_8K_3D_3$, containing the conserved poly-lysine patch, is compared to the stereotypical membrane peptide $GW_2[LA]_7W_2A$ (WALP20). WALP20 is characterized by a slender hydrophobic motif rich in leucines $[LA]_7$. Nevertheless, a pronounced cholesterol attraction is only observed for the designed motif. The resolved pattern attracts cholesterol in all-atom simulations.

128 individuals, which we took as the baseline population size for GA convergence. Data from GA runs containing 128+ individuals and at least 40 generations was used for sequence analysis.

Accompanying the convergence in fitness with respect to population size, we observed a similarity in sequences being produced by distinct GA runs. While GA runs with lower population sizes (< 64) did eventually converge to some fitness value, their respective peptides showed large diversity in sequence, indicating that the algorithms converged to local optima in the solution space. This diversity in sequence decreases as population size increases, with very similar sequences being obtained as population sizes increase to 128 individuals and above. Furthermore, at such population sizes, starting the evolution from different initial populations consisting of randomly generated sequences yields a consistent result. On these grounds, we can conclude that the GA successfully converges to a global optimum.

To gain detailed insights into the resolved evolutionary landscape, high-fitness sequences from all GA runs with populations of 128+ individuals were combined to generate a sequence logo out of the sampled sequence space (see Fig. 3A). Such a logo expresses the degree of amino acid conservation at each position within the sequence in terms of the concomitant Shannon entropy (bits) by scaling the character height of the corresponding amino acid. Randomly occurring amino acids at a certain position contain no information, corresponding to a small letter, whereas a more frequently occurring amino acid encodes information, corresponding to a larger letter.

Interestingly, the global solution converges to a distinct pattern featuring a *short* conserved hydrophobic block consisting of proline (P), valine (V), and occasionally (iso)leucine (I/L), located in the center of the peptide, embedded within two hydrophilic blocks consisting of three highly conserved positively charged lysines (K). The first and last three terminal residues are less conserved and predominantly feature negatively charged aspartic acids (D). Finally, it is important to emphasize that the here-resolved solution space is subject to a constraint in secondary structure, i.e. all sequences are assumed to be alpha-helical (5). We will extensively address the transferability of solution space within a later section of the manuscript.

A short hydrophobic block is crucial for maximal cholesterol attraction. We observe the emergence of a specific hydrophobic pattern in the high-fitness peptide sequences. Most notable is the high conservation of the lysine (K) residues at positions 4-6 (and 15-17), which is conserved early in the evolution of the sequence and is present in all high fitness sequences. Surprisingly, a strong competition between lysine and arginine (R) is not observed, in contrast to their hypothesised equivalent

role within the CRAC definition (18).

This sharp positional convergence of lysine prompted us to investigate what role the length of the hydrophobic block plays in the cholesterol sensing ability of the sequence. We created dummy peptides according to the $K_{(10-x/2)}$–V_x–$K_{(10-x/2)}$ motif with each peptide consisting of 20 amino acids in total. Here, valines form the hydrophobic block of the peptides, with lysines functioning as the hydrophilic edges. By varying the number of valines and lysines, we effectively vary the length of the hydrophobic block.

Interestingly, cholesterol affinity increases with decreasing hydrophobic block length, with an optimal effect at a specific length ($K_8V_4K_8$), see figure 3B. This pattern seems to arise from a trade-off between short block length and transmembrane (meta)stability, with a further decrease in block length resulting in a sharp decline in functionality. Artificially restraining a transmembrane orientation/topology for these motifs ($K_9V_2K_9$, and even K_{20}) restores the functionality to levels similar to the best performer ($K_8V_4K_8$) (see Fig. S8). This suggests that cholesterol attraction is predominantly mediated by the lysine residues, conform with their high conservation, and that their positioning, i.e. the conserved hydrophobic block length, ensures a transmembrane topology in the course of evolution. Owing to the membrane thickening effect of cholesterol (**?**), cholesterol-enriched phases such as the liquid ordered (Lo) phase generally favors TMDs characterized by a long rather than short hydrophobic length (5, 29–32). This suggest that optimal binding of free membrane cholesterol or cholesterol enriched ordered phases are, in fact, mediated by distinct physicochemical driving forces.

Cholesterol clustering favors a slender hydrophobic block. The presence of the specific amino acids in the sequence relates to the coarse-grained forcefield used in the simulations. This is particularly relevant for the hydrophobic block, which, in contrast to the highly conserved lysines, displays a split between several amino acids.

The common ground between these peptides is that they are constructed of small hydrophobic amino acids, containing only a single sidechain bead, suggesting that the GA seeks to optimize for cholesterol attraction by creating a thin, hydrophobic part within the peptide. This preference for small hydrophobic amino acid is further emphasised by the increased frequency of proline and valine compared to leucine and isoleucine, with the only difference between these amino acids within the coarse-grained forcefield being the bond length between the sidechain and backbone bead (33).

We further explore the presence for a thin hydrophobic block by creating a series of dummy peptides with sequence $K_6V_8K_6$, where we vary the bond length of the valine residues, allowing us to artificially change the thickness of the hydrophobic block. Indeed, we observe that cholesterol attraction increases even further when we artificially decrease the bond distance of the valine residues, see figure 3C. Likewise, when we increase the bond distance, cholesterol attraction reduces. This is in line with our previous findings, and it is consistent with the notion that reduction of TMD surface area increases cholesterol affinity (5).

The resolved hydrophobic pattern attracts cholesterol in unbiased atomistic simulations. In this work, we resolved the essential physicochemical driving forces that underpin cholesterol recognition/binding in transmembrane domains within homogeneous model membranes. The here-resolved chemical features of the optimal cholesterol attractor are subsequently translated into realistic peptide sequences by anticipating for the following two model approximations:

(I) Since transmembrane domains are predominantly alpha-helical, an alpha-helical secondary structure restraint was imposed on generated sequences. While this assumption simplifies the search space by avoiding the problem of secondary structure prediction, it introduces the possibility for amino acids with large penalties to alpha-helical propensity (e.g. proline, valine) to occur in the generated sequences, producing non-helical peptides in unrestrained atomistic simulations. In line with the here-resolved properties of cholesterol attractors, i.e. the presence of a short and thin hydrophobic block, we devised a more realistic sequence by replacing these amino acids with helix-favoring alternatives such as leucine.

(II) Electrostatic interactions are underestimated in the coarse-grained simulations, enabling the formation of sequences with a high net charge. To obtain a sequence with net zero charge, we balance the conserved lysines patches by adding three aspartic acids to both terminal ends.

Altogether, this leads to the more realistic sequence, $D_3K_3L_8K_3D_3$, which retains all the design features put forth by the GA. Indeed, atomistic simulations confirm that such a sequence exhibits a pronounced cholesterol sensing functionality, as is shown in figure 3E. A peak in cholesterol content near the peptide (within 0.9 nm) is observed, showing a 33% increase relative to the overall proportion of cholesterol molecules in the system. Though slightly diminished, we thus observe that the functionality persists between the resolutions. Since the coarse-grained model is parametrized on reproducing thermodynamic properties while only coarsely captures structural features, the observed resolution independence strongly suggest that the mechanism of cholesterol attraction is mainly of a thermodynamic nature.

Discussion

We have demonstrated the ability of evo-MD to yield unique insights into the driving forces that underpin ligand recognition in membrane proteins. Inverse design of molecules is based on the principle that most if not all of the physical driving forces that govern functionality are inherently encoded within the complexity of independently parameterized classical molecular force fields. Our approach therefore substantially differs from existing data-driven quantitative structure-activity relationship (QSAR) based inverse design approaches, where optima in an abstract high-dimensional latent space are translated into corresponding chemical structures using machine learning-based variational encoders (26). In fact, evo-MD can be exploited to further the development of more cost efficient latent space-based models by resolving relevant physicochemical descriptors in retrospect after optimization.

So far distinct motifs associated with cholesterol binding have been identified in native membrane proteins such as, for example, the CRAC motif (3) or the glycine zipper motif (34). Surprisingly, we observe that optimal clustering of free membrane cholesterol to isolated transmembrane domains prefers a short slender hydrophobic region flanked by lysine residues. This pattern evidently offers a pronounced net advantage for cholesterol binding in both coarse-grained and atomistic sim-

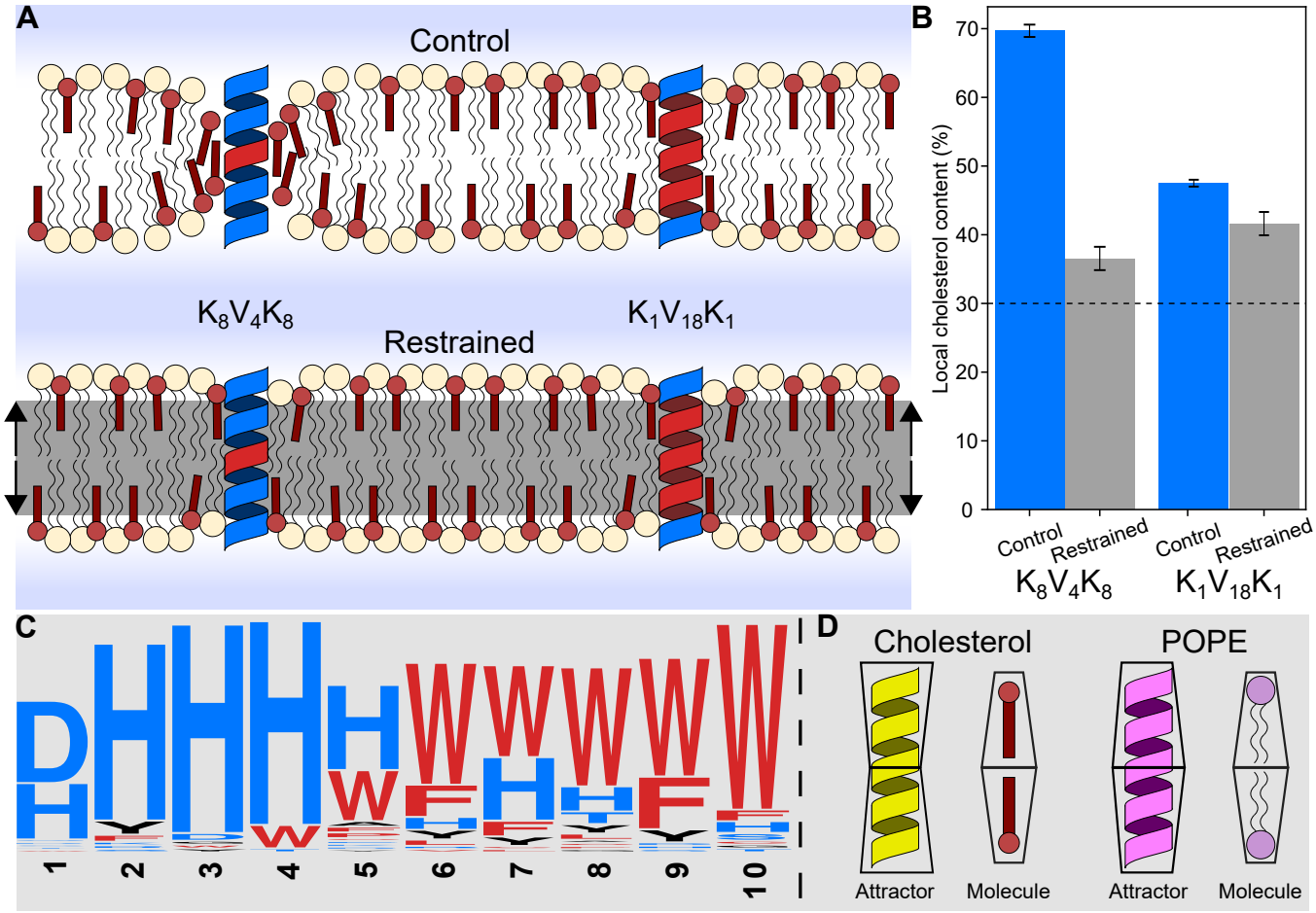

Fig. 4. Mechanistic model for maximal cholesterol attraction. (A, Control): Cholesterol (red) has a lower energy penalty for membrane snorkeling compared to POPC (beige), allowing for more favorable shielding of deeply located lysine residues. High-affinity cholesterol attractors utilize this effect by increasing deep lysine interactions, leading to local accumulation of cholesterol molecules. (A, Restrained): Application of a force to lipid headgroups within a specific distance from the membrane center prevents cholesterol snorkeling/flip-flopping. (B): Removal of cholesterol snorkeling (Restrained) leads to a large drop in functionality for cholesterol attractors with short hydrophobic blocks ($K_8V_4K_8$), while attractors with longer hydrophobic blocks ($K_1V_{18}K_1$) are less affected. The dashed line indicates the average cholesterol content of the system (30%). Bars represent the standard error of the mean. (C): Sequence logo of a (mirrored) POPE attractor, computed from all high-fitness (≥ 630 kJ/mol) peptide sequences, reveals a sequence distinct from the cholesterol attractor. (D): While cholesterol and POPE exhibit similarity in effective shape, corresponding attractor peptides vary notably in shape, suggesting a shape unspecific mechanism of attraction.

ulations. The traditional lock and key hypothesis postulates that protein-ligand binding is mediated by matching of complementary molecular shapes. The optimal cholesterol attractor resolved in our evo-MD simulations features a concave overall shape because of its voluminous hydrophilic ends and slender hydrophobic midsection. Interestingly, such a molecular shape is in principle compatible with the intrinsic cone-shape of cholesterol, i.e. small head group versus a voluminous hydrophobic moiety (35). Since both cholesterol and POPE lipids are cone-shaped molecules (35), and if binding is indeed effective molecular shape determined, then also POPE lipids are expected to bind the resolved motif. However, we do not observe an increased attraction for POPE lipids (see Fig. S11). As a matter of fact, an TMD optimized for the attraction of POPE lipids in fact features a highly contrasting structure (see Fig. 4C,D), i.e. a voluminous hydrophobic region comprised of tryptophans (W) with a slender intermediate hydrophilic region comprised of histidines (H). This suggests that the observed binding and recognition of cholesterol molecules as well as POPE molecules is not based on effective molecular shape and is thus facilitated by a very different mechanism.

In contrast to POPC lipids, cholesterol faces a low free energy barrier to flip-flop between the two membrane leaflets. Consequently, the head group of cholesterol is better able to bind to the deeply located lysine patch by snorkeling within the hydrophobic region of the membrane. Such positioning of cholesterol molecules shields the lysine patch from unfavorable interactions with the hydrophobic lipid tails (see Fig. 4A). Indeed, artificial restriction of lipid snorkeling confirms this hypothesis: High fitness sequences which contain a short hydrophobic block, and which would thus highly depend on this snorkeling effect, indeed show a large decrease in functionality, while 'longer' less optimal attractors, whose ability to attract cholesterol mostly relies on the slenderness of the hydrophobic section, remain relatively unaffected (see Fig. 4B). We therefore attribute the enhanced attraction of cholesterol to the enforced snorkeling of lipid head groups in the direct vicinity of the TMD. Within such a mechanism, a slender central hydrophobic TMD region comprised of, for example, leucines likely eases a closer binding of the cholesterol head groups to

the lysine patch and thereby increases the associated enthalpic gain. Notably, also the optimal POPE attractor (see Fig. 4C) can be attributed to a differential difference in snorkeling between POPE and POPC lipids with the distinction that snorkeling now exploits a favorable enthalpic interaction between POPE head groups and the centrally located tryptophan region.

In essence, our cholesterol attracting sequence can be regarded as a superposition of multiple CRAC motifs (3), however without the aromatic residues tyrosine. Here, it is important to emphasize that our sequence has evolutionarily evolved to *maximally* cluster cholesterol within the membrane, rather than closely binding a single cholesterol molecule. Tyrosine may favorably bind the hydrophobic ring structure of a cholesterol molecule and thereby offer an advantage under the latter evolutionary pressure. However, with tyrosine being a voluminous amino acid that is far less hydrophilic than lysine, stacking multiple tyrosines together with lysine seems counterproductive for maximizing cholesterol clustering. Notably, also the presence of multiple valines/leucines and a single lysine facilitates a pronounced cholesterol attraction even in the absence of tyrosine or a short hydrophobic block (Fig. 4B). This suggests that linear cholesterol-recognition motifs can be extremely variable and may be less-defined than hitherto assumed, in correspondence with findings in other studies (10, 15).

Surprising is the absence of arginine in cholesterol attractor sequences, with the CRAC motif suggesting a similar role for the lysine and arginine residues. Although substitution of lysine with arginine retains most of the functionality, a notable decrease can be observed (see figure S10). This preference for lysine seems to stem from a larger difference in attraction/repulsion between cholesterol and POPC in the force field. While lysine and arginine have a charged residue end in common, the core of the residue differs, with arginine being more polar than lysine. The apolar lysine sidechain provides a larger difference in attraction relative to arginine, with charged POPC headgroups being less attracted than the simply polar cholesterol headgroup. With the main feature of attraction satisfied by either residue (i.e., high hydrophilicity), we propose that this difference in sidechain further discriminates between the two lipids. We hypothesize that such an effect is more pronounced in strong cholesterol clustering sequences, in contrast to the CRAC motif, due to the larger number of cholesterol molecules that are simultaneously attracted.

Of course, the question remains to which extend the here-resolved hydrophobic pattern can prevail in nature, where evolution is subject to many additional constraints. Purely composition-wise, verification with a coarse-grained native epithelial membrane model (36) suggests that the here-resolved functionality persists in more realistic membrane environments (see Fig. S7). However, proteins targeting cholesterol-enriched liquid ordered phases (lipid rafts) are expected to prefer longer transmembrane domains (5, 29–32), likely to dampen hydrophobic mismatches between itself and the lipid rafts. The here-resolved motif is therefore not expected to optimally bind towards the interface of cholesterol-enriched liquid ordered domains (5, 37) (see Fig. S12). Nevertheless, the clustering of cholesterol is itself membrane phase independent and equally occurs when the resolved TMD is embedded within a liquid ordered DPPC:cholesterol mixture (see fig. S8). Although the hydrophobic pattern sequence adopts some of the basic features of the CRAC motif, the sequences of many native cholesterol-associated membrane proteins have evidently evolved on a respectable distance from the here-resolved attraction optimum. An obvious reason for this is that the true fitness function in nature will be subject to multiple distinct evolutionary pressures related to both protein trafficking and functionality, such as the potential interaction with raft domains and *especially* the interaction with other membrane proteins. Particularly, protein-protein interactions are highly specific and thus pose a far larger constraint on sequence variability than protein-lipid interactions. Although the isolated evolutionary landscape of cholesterol attraction is well-defined and sharply peaked, the observed mechanism of attraction is evidently not based on the precise matching of complementary molecular shapes but rather exploits the differential ability of lipids to snorkel within the membrane. Consequently, a large diversity of sub-optimal cholesterol attracting sequences can evolve evolutionarily. However, it is important to emphasize that our work solely considers the binding of cholesterol to isolated transmembrane domains. We speculate that binding of cholesterol to pockets present between clustered transmembrane domains (e.g., dimers and higher oligomeric species) will allow for a higher binding specificity and lower sequence variability since such a binding renders the matching of complementary molecular shapes more relevant over the here-observed thermodynamic mechanism (7–10).

To summarize, we have demonstrated the ability of evo-MD approach to isolate the evolutionary fingerprints of protein-lipid interactions within membrane proteins. This unique property enables evo-MD to yield valuable insights into *how* proteins may recognize and bind distinct membrane lipids or lipid soluble ligands such as hormones and vitamins within the highly crowded environment of lipid membranes. We therefore expect evo-MD to yield novel insights into the mechanisms that underpin the molecular organization of biological membranes and protein trafficking. Particularly, we foresee important near-future applications of evo-MD in designing peptide motifs that are able to selective target characteristic membrane features, such as membrane curvature and lipid composition. Importantly, this will pave the road for the inverse design of peptide drugs and peptide based drug vehicles which are able to selectively target the fluid membranes of viruses, microbes and cancer cells since their membrane leaflets are characterized by pronounced differences in curvature and/or lipid composition. Notably, structure-based molecular design approaches are rendered ineffective beyond the level of targeting individual lipid species because of the diffusive, fluid nature of lipid membranes. Finally, as evo-MD effectively extracts or mines implicit information from the underlying force field, integration of improved (coarse-grained) force fields—for example, the recent Martini 3 (38, 39) and the Spica force-field (40, 41) —in conjunction with the integration of groundbreaking protein structure prediction methodology— for example, the Alphafold 2 project (42)—could further facilitate these applications.

Materials and Methods

Software. Coarse-Grained Simulations were performed with the Martini 2.2 CG force field using the GROMACS 2019.1 molecular dy-

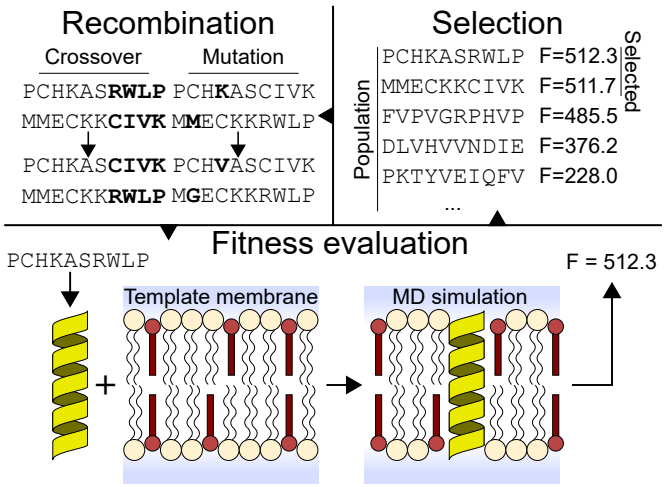

Fig. 5. Graphical overview of *EVOMD*. (Bottom): Peptide sequences are evaluated by means of MD simulation. A peptide structure (yellow) is generated from sequence and inserted into a POPC (beige) and cholesterol (red) bilayer membrane. The fitness is then computed from the resulting trajectory. (Top-right): Highest fitness sequences are selected from the evaluated population. (Top-left): Through recombination (involving crossover and mutation operations) of the selected sequences, a new population is generated.

namics package. All-atom simulations were performed using the CHARMM36 force field. *EVOMD* is written in Python 3.6.8 and depends on the *NumPy* and *MPI for Python* packages for functionality. Peptide topologies are generated using *seq2itp* (33). Input parameters for the coarse-grained simulations are based on the Martini 2 'New-RF' parameters (43), with exceptions detailed in the relevant sections below.

Implementation of simulated evolution: EVOMD. *EVOMD* was developed as a framework for the simulated evolution of MD simulation systems. Simulated evolution is a type of optimization problem involving the optimization of some property of the simulated system, by means of iteratively tuning a set of parameters. The performance (*fitness*) of such a parameter set is then measured by means of a fitness function, which generally consists of one or more MD simulations followed by an analysis step.

Using GAs, we can manage large, hyper-dimensional optimization problems through efficient exploration of the search space. Analogous to the method's origin in genetics, we label each possible solution as a *chromosome*, which consists of a unique set of parameters encoded into a (bit)string sequence. The algorithm iteratively samples parts of the search space by forming a *population* of chromosomes and measuring their fitnesses. In line with evolution, individuals with high fitnesses are selected to *recombine* and form a new population. As the new population is based on a highest fitness subset of the previous population, it is assumed that the average fitness of the population increases each iteration. This process is visualized in figure 5.

Implementation of the cholesterol sensing project is summarized in figure S6. Each candidate peptide is encoded as a sequence of one-letter amino acid codes. For faster convergence, the sequence is mirrored to produce a palindromic sequence, effectively reducing the search space for a peptide 20 amino acids in length from 20^{20} to 20^{10} (assuming 20 amino acid types). The GA is initialized by generating a random population of N_{pop} sequences, after which each sequence is evaluated in parallel according to the fitness function.

The fitness function takes a sequence as argument and returns a single float value representing the sequence's fitness. This function involves several simulation steps (*modules*): *generate_peptide*, *insert_peptide*, *production*, and *compute_fitness*. *Generate_peptide* generates a peptide structure and topology using the *seq2itp* tool (33), followed by energy minimization and peptide-membrane alignment. *Insert_peptide* combines the peptide structure with an existing equilibrated membrane structure containing 128 lipid molecules (90 POPC, 38 cholesterol) and 1598 Martini water beads, and places the peptide transversely through the membrane. Collisions between peptide and membrane structures are resolved by partially decoupling the non-bonded interactions—combined with soft-core potentials—and running a steepest descent algorithm. The *production* module adds ions to neutralize any net charge on the system, after which equilibration and production simulations are performed. The *compute_fitness* module then measures the ensemble-averaged short-ranged Lennard-Jones interactions between peptide and cholesterol (Coulomb interactions with cholesterol are absent within the CG model.) molecules from the simulation trajectory, which is returned as the fitness. Notable, such a fitness is the direct outcome of the competition between cholesterol and POPC lipids to interact with the peptide. Therefore, its value is directly proportional to the adopted cholesterol concentration and thus the *relative* binding free energy.

Once all sequences in the population have been evaluated, the algorithm proceeds by selecting the best N performers to serve as parents for the next population. A new sequence is generated by recombining two randomly selected sequences from the parent pool, which involves a *cross-over* operation and a *mutation* operation. During the cross-over operation, a random position is selected in the new sequence. The part to the left of that position is inherited from the first parent, while the rest of the sequence is inherited from the second parent. Afterwards, the mutation operation ensures that each position in the sequence has a $1/len(sequence)$ chance of being replaced with a random amino acid. New sequences are created in this manner until a new population of size N_{pop} is produced. This process of population fitness evaluation and recombination of the highest fitness candidates into a new population is then repeated until a desired number of iterations is achieved.

A *rerun* mechanism was implemented to account for possible undersampling during fitness evaluation. If a sequence reoccurs in a future generation, its fitness value will be computed from the weighted average of the current and all prior fitness evaluations. With the chance of re-occurrence increasing as the algorithm converges, this mechanism serves to increase confidence in the final fitness value.

Membrane setup. The membrane template structure consists of a 5.6x5.6x10 nm simulation box, containing a bilayer membrane in water solvent. The membrane consists of 90 POPC molecules and 38 cholesterol molecules. The solvent consists of 1598 Martini water beads.

Evo-MD modules.

Module: generate_peptide. As the seq2itp tool only produces topology files, a structure file for the peptide is generated by stacking hardcoded amino acid structures along the Z-axis and performing a 1.5 ps simulation at low time step (0.05 fs) using the GROMACS 2019.1 'sd' stochastic dynamics integrator. This allows the hardcoded structure to slowly relax to a more reasonable conformation according to the generated topology.

Module: insert_peptide. *Insert_peptide* centers the peptide in the membrane box and merges the two structures together. A steepest descent, combined with a partial decoupling of the non-bonded interactions ($\lambda = 0.75$) and soft-core potentials, is then performed on the merged structure to remove collisions between the peptide and the membrane structures.

Module: production. A final steepest descent is performed without soft-core potentials. A short, 1.5 ps simulation is performed at low time step (0.05 fs) using the stochastic dynamics integrator to prevent blowing up of the system before the actual simulation is performed. The production simulation is a 500 ns NPT MD simulation with 30 fs timestep, of which the first 50 ns are used for equilibration. Temperature is coupled to 300 K using velocity rescaling ($\tau = 1$ ps with separate coupling groups for the membrane, peptide, and solvent), Pressure is coupled semi-isotropically to 1 bar using the Berendsen algorithm ($\tau = 8$ ps), with compressibility set to $4.5 \cdot 10^{-5}$ bar^{-1}.

Module: compute_fitness. Evaluation of the sequence's fitness is finalized by computation of a fitness value from the produced simulation trajectory. GROMACS' *gmx energy* tool is used to extract

Table 1. Overview of GA run parameters.

Population	Repeats	Parents	Iteration elites	Rerun elites	Mutation frequency*
4	4	2	1	1	1/20
8	4	2	1	1	1/20
16	4	4	1	1	1/20
32	4	8	2	2	1/20
64	2	16	2	2	1/20
128	2	16	2	2	1/20
256	2	16	2	2	1/20
320	1	16	2	2	1/20

*

the ensemble average of the non-bonded interaction energies from the production trajectory. The absolute value is then returned to the GA.

Quantification of sequence cholesterol clustering capability was performed by measuring the ratio of cholesterol molecules to membrane molecules within a cylinder of radius r centered on the peptide center-of-mass (COM). GROMACS' *gmx rdf* tool was used to compute a cumulative number radial distribution function ($g_{CN}(r)$) for cholesterol COMs and POPC COMs, both with respect to the peptide COM. The final ratio figures are created by computing $f_{ratio}(r) = \frac{g_{CN,CHOL}(r)}{g_{CN,CHOL}(r)+g_{CN,POPC}(r)}$.

Comparisons between multiple ratio figures were taken at a cylinder radius of 1.0 nm, chosen as a middle-ground between local-sampling (low r) and sufficient sampling (high r).

GA parameters. Production runs of the GA were performed according to the parameters as described in Table 1. *Parents* indicates the size of the selection pool, from which parents were selected at random for the recombination step. *Iteration elites* describe the number of highest fitness sequences which pass unaltered into the next generation. *Rerun elites* keeps track of a list of sequences which have been evaluated more than once, and allows several highest fitness sequences to proceed to the next generation unaltered. The total number of elites is equal to the sum of iteration and rerun elites.

Backmapping. Backmapping from coarse-grained Martini forcefield to all-atom CHARMM36 forcefield was performed using the *backward.py* tool, according to the workflow described by Wassenaar et al. (44).

Atomistic simulations. All-atom simulations were performed using the CHARMM36 forcefield. 500 ns NPT MD simulations were performed using a 0.002 fs timestep on the backmapped structures. The temperature was coupled to 303.15 K using velocity rescaling ($\tau = 1$ ps with separate coupling groups for the membrane, peptide, and solvent), Pressure was coupled semi-isotropically (parallel and orthogonal to the membrane) to 1 bar using the Berendsen algorithm ($\tau = 8$ ps), with compressibility set to $4.5 \cdot 10^{-5}$ bar^{-1}.

Restraining lipid snorkeling/flip-flopping. To investigate the cholesterol snorkeling mechanism, removal of lipid snorkeling and flip-flopping was facilitated by application of an inverse flat-bottomed position restraint to the first beads of both POPC (NC3 bead) and cholesterol (ROH bead). The position restraint consists of a layer, parallel to the membrane and centered on the bilayer center. A harmonic force with force constant 1000 kJ $\cdot$ mol^{-1} $\cdot$ nm^{-2}, directed away from the bilayer center, is applied to affected beads that come within 2.0 nm (NC3) or 1.5 nm (ROH) of the center of the bilayer.

1. Levental I, Veatch SL (2016) The continuing mystery of lipid rafts. *J. Mol. Biol.* 428(24):4749–4764.
2. Midzak A, Papadopoulos V (2014) Binding domain-driven intracellular trafficking of sterols for synthesis of steroid hormones, bile acids and oxysterols. *Traffic* 15(9):895–914.
3. Fantini J, Barrantes FJ (2013) How cholesterol interacts with membrane proteins: an exploration of cholesterol-binding sites including CRAC, CARC, and tilted domains. *Front. Physiol.* 4.
4. Scala CD, et al. (2017) Relevance of CARC and CRAC cholesterol-recognition motifs in the nicotinic acetylcholine receptor and other membrane-bound receptors. *Curr. Top. Membr.* 80:3–23.
5. Lorent JH, et al. (2017) Structural determinants and functional consequences of protein affinity for membrane rafts. *Nat. Commun.* 8(1):1219.
6. Fatakia SN, Sarkar P, Chattopadhyay A (2019) A collage of cholesterol interaction motifs in the serotonin1a receptor: An evolutionary implication for differential cholesterol interaction. *Chem. Phys. Lipids* 221:184–192.
7. Bukiya AN, Dopico AM (2017) Common structural features of cholesterol binding sites in crystallized soluble proteins. *J. Lipid Res.* 58(6):1044–1054.
8. Wang C, Ralko A, Ren Z, Rosenhouse-Dantsker A, Yang X (2019) Modes of cholesterol binding in membrane proteins: a joint analysis of 73 crystal structures. *Adv. Exp. Med. Biol.* pp. 67–86.
9. Dubey V, Bozorg B, Wüstner D, Khandelia H (2020) Cholesterol binding to the sterol-sensing region of niemann pick c1 protein confines dynamics of its n-terminal domain. *PLoS Comp. Biol.* 16(10):e1007554.
10. Marlow B, Kuenze G, Li B, Sanders CR, Meiler J (2021) Structural determinants of cholesterol recognition in helical integral membrane proteins. *Biophysical Journal* 120(9):1592–1604.
11. Luo J, Yang H, Song BL (2019) Mechanisms and regulation of cholesterol homeostasis. *Nat. Rev. Mol. Cell Biol.* 21(4):225–245.
12. Li H, x. Yao Z, Degenhardt B, Teper G, Papadopoulos V (2001) Cholesterol binding at the cholesterol recognition/ interaction amino acid consensus (CRAC) of the peripheral-type benzodiazepine receptor and inhibition of steroidogenesis by an HIV TAT-CRAC peptide. *Proc. Natl. Acad. Sci. USA* 98(3):1267–1272.
13. Nierzwicki Ł, Czub J (2015) Specific binding of cholesterol to the amyloid precursor protein: structure of the complex and driving forces characterized in molecular detail. *J. Phys. Chem. Lett.* 6(5):784–790.
14. Koufos E, Chang EH, Rasti ES, Krueger E, Brown AC (2016) Use of a cholesterol recognition amino acid consensus peptide to inhibit binding of a bacterial toxin to cholesterol. *Biochemistry* 55(34):4787–4797.
15. Elkins MR, et al. (2017) Cholesterol-binding site of the influenza m2 protein in lipid bilayers from solid-state NMR. *Proc. Natl. Acad. Sci. USA* 114(49):12946–12951.
16. Castellano BM, et al. (2017) Lysosomal cholesterol activates mTORC1 via an SLC38a9–niemann-pick c1 signaling complex. *Science* 355(6331):1306–1311.
17. Epand RM (2006) Cholesterol and the interaction of proteins with membrane domains. *Prog. Lipid Res.* 45(4):279–294.
18. Fantini J, Epand RM, Barrantes FJ (2019) Cholesterol-recognition motifs in membrane proteins. *Adv. Exp. Med. Biol.* 1135:3–25.
19. Rosenhouse-Dantsker A, Noskov S, Durdagi S, Logothetis DE, Levitan I (2013) Identification of novel cholesterol-binding regions in kir2 channels. *J. Biol. Chem.* 288(43):31154–31164.
20. Singh AK, et al. (2012) Multiple cholesterol recognition/interaction amino acid consensus (CRAC) motifs in cytosolic c tail of slo1 subunit determine cholesterol sensitivity of ca2- and voltage-gated k(BK) channels. *J. Biol. Chem.* 287(24):20509–20521.
21. Jafurulla M, Tiwari S, Chattopadhyay A (2011) Identification of cholesterol recognition amino acid consensus (CRAC) motif in g-protein coupled receptors. *Biochem. Biophys. Res. Commun.* 404(1):569–573.
22. Sengupta D, Chattopadhyay A (2012) Identification of cholesterol binding sites in the serotonin1a receptor. *J. Phys. Chem. B* 116(43):12991–12996.
23. Hedger G, et al. (2019) Cholesterol interaction sites on the transmembrane domain of the hedgehog signal transducer and class f g protein-coupled receptor smoothened. *Structure* 27(3):549–559.e2.
24. Sejdiu BI, Tieleman DP (2020) Lipid-protein interactions are a unique property and defining feature of g protein-coupled receptors. *Biophys. J.* 118(8):1887–1900.
25. Lutz S (2010) Beyond directed evolution—semi-rational protein engineering and design. *Curr. Opin. Biotechnol.* 21(6):734–743.
26. Sanchez-Lengeling B, Aspuru-Guzik A (2018) Inverse molecular design using machine learning: Generative models for matter engineering. *Science* 361(6400):360–365.
27. Kaelbling LP, Littman ML, Moore AW (1996) Reinforcement learning: A survey. *Journal of Artificial Intelligence Research* 4:237–285.
28. Sloss AN, Gustafson S (2019) 2019 evolutionary algorithms review.
29. Schafer LV, et al. (2011) Lipid packing drives the segregation of transmembrane helices into disordered lipid domains in model membranes. *Proc. Natl. Acad. Sci. USA* 108(4):1343–1348.
30. Kaiser HJ, et al. (2011) Lateral sorting in model membranes by cholesterol-mediated hydrophobic matching. *Proc. Natl. Acad. Sci. USA* 108(40):16628–16633.
31. Milovanovic D, et al. (2015) Hydrophobic mismatch sorts SNARE proteins into distinct membrane domains. *Nat. Commun.* 6(1).
32. Chakraborty S, et al. (2020) How cholesterol stiffens unsaturated lipid membranes. *Proc. Natl. Acad. Sci. USA* 117(36):21896–21905.
33. Monticelli L, et al. (2008) The martini coarse-grained force field: Extension to proteins. *Journal of Chemical Theory and Computation* 4(5):819–834.
34. Barrett PJ, et al. (2012) The amyloid precursor protein has a flexible transmembrane domain and binds cholesterol. *Science* 336(6085):1168–1171.
35. Kollmitzer B, Heftberger P, Rappolt M, Pabst G (2013) Monolayer spontaneous curvature of raft-forming membrane lipids. *Soft matter* 9(45):10877–10884.
36. Wilson KA, et al. (2021) The role of plasmalogens, forssman lipids, and sphingolipid hydroxylation in modulating the biophysical properties of the epithelial plasma membrane. *The Journal of Chemical Physics* 154(9):095101.
37. Risselada HJ, Marrink SJ (2008) The molecular face of lipid rafts in model membranes. *Proceedings of the National Academy of Sciences* 105(45):17367–17372.
38. Souza PC, et al. (2021) Martini 3: a general purpose force field for coarse-grained molecular dynamics. *Nat. Method.* pp. 1–7.
39. Risselada HJ (2021) Martini 3: a coarse-grained force field with an eye for atomic detail. *Nat. Methods.* 18(4):342–343.

40. Shinoda W, DeVane R, Klein ML (2007) Multi-property fitting and parameterization of a coarse grained model for aqueous surfactants. *Molecular Simulation* 33(1-2):27–36.
41. Seo S, Shinoda W (2019) Spica force field for lipid membranes: Domain formation induced by cholesterol. *J. Chem. Theory Comput.* 15(1):762–774. PMID: 30514078.
42. Senior AW, et al. (2020) Improved protein structure prediction using potentials from deep learning. *Nature* 577(7792):706–710.
43. de Jong DH, Baoukina S, Ingólfsson HI, Marrink SJ (2016) Martini straight: Boosting performance using a shorter cutoff and gpus. *Computer Physics Communications* 199:1–7.
44. Wassenaar TA, Pluhackova K, Böckmann RA, Marrink SJ, Tieleman DP (2014) Going backward: A flexible geometric approach to reverse transformation from coarse grained to atomistic models. *Journal of Chemical Theory and Computation* 10(2):676–690. PMID: 26580045.