Electronic Supplementary Material (ESI) for Physical Chemistry Chemical Physics.

Supplemental Information for

An Active Machine Learning Discovery Platform for Pore-forming Peptides

Alexander van Teijlingen[1], Daniel Edwards[2], Liao Hu[2], Annamaria Lilienkampf[2], Scott L. Cockroft[2] and Tell Tuttle[1*]

[1]Pure and Applied Chemistry, University of Strathclyde, 295 Cathedral Street, Glasgow, G1 1XL, UK.

[2]Department of Chemistry, University of Edinburgh, David Brewster Road, Edinburgh, EH9 3FJ, UK.

Corresponding author: tell.tuttle@strath.ac.uk

Table of Contents

1. Experimental Procedures

1.1 Peptide synthesis

The peptides were synthesized on a Wang-linker functionalized polystyrene resin using standard Fmoc chemistry. The crude peptides were purified by semi-preparative RP-HPLC. The lyophilized peptides were characterized by analytical RP-HPLC and MALDI-ToF or ESI MS (Table S1).

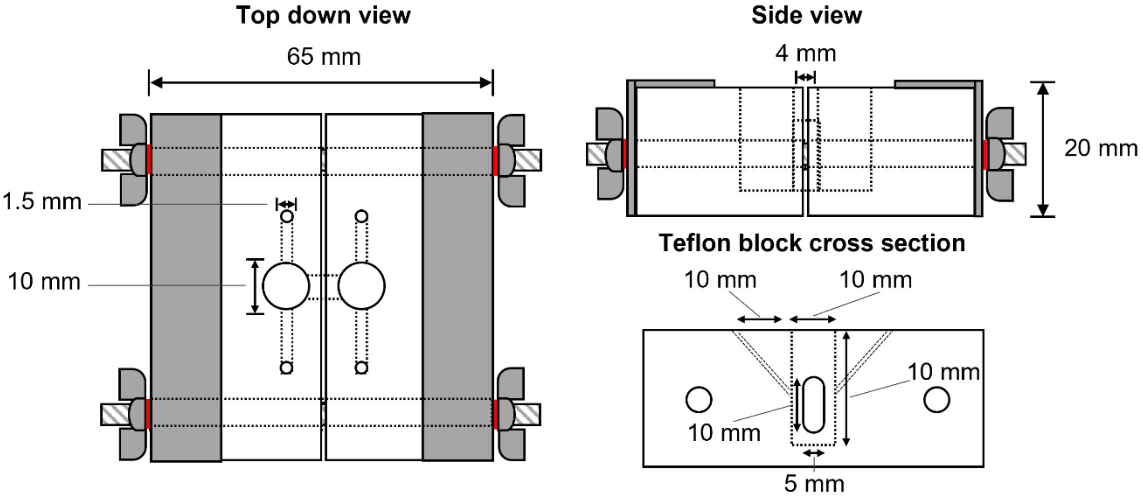
Table S1. Peptide characterization by RP-HPLC and MS.

Peptide	Sequence	HPLC (t_R min)[a]	Purity (%)	$[M+H]^+$ [b]
1	FFFLSRIF	3.4	94	1076.63
2	GSGTGSGT	0.8	99	623.30 [c]
3	CFTYFFRV	3.0	99	1082.71
4	VCVYWWRT	3.7	95	1112.76
5	FGYVLIRI	3.1	97	980.79
6	LSFMRFFF	3.4	99	1094.66
7	SAFWWFRF	3.5	99	1146.83
8	HGLFWWRF	3.1	99	1148.84
9	GIALKIVW	3.9	95	899.58
10	YCVLRLPF	3.2	97	1010.76
11	FFMSIRFF	3.4	98	1094.79
12	FFIMRFFS	3.4	99	1094.55
13	FFRISMFF	3.4	99	1094.63

[a] Detection at 220 nm; [b] MALDI-ToF MS; [c] ESI-MS

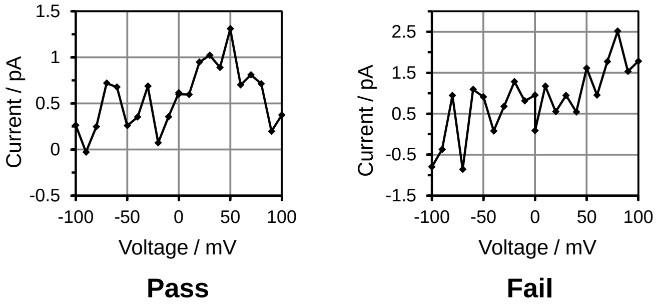
1.2 Planar lipid bilayer pore formation experiments

The ability of the peptides to disrupt a bilayer was then investigated by measuring the amount of current that could flow across the bilayer. Planar lipid bilayer recordings were performed in a custom Teflon cell equipped with two 1 mL compartments separated by a 20 μm thick Teflon film (Goodfellow) with an ~100 μm diameter aperture (Figure S1). A hanging drop of hexadecane in *n*-pentane (5 μL, 10%, v/v) was touched on each side of the Teflon sheet containing an aperture and allowed to dry for 1 min. The cell was placed into a Faraday cage, and Ag/AgCl electrodes (Warner) connected to a patch clamp amplifier (Axopatch 200B, Molecular Devices) were suspended either side of the Teflon sheet. KCl/MOPS buffer (600 μL) was added to the well on each side of the aperture. POPC lipid (1-palmitoyl-2-oleoyl-*sn*-glycero-3-phosphocholine) (approximately 8 μL, 5 μg μL^{-1} in *n*-pentane) was added to each side of the well and left for ~5 min to allow the pentane to evaporate. The buffer solution on both sides of the Teflon sheet was aspirated and dispensed using a Hamilton syringe to paint a phospholipid bilayer across the aperture. A ±1 mV pulse was applied at 1333 Hz to determine when a bilayer was obtained (capacitance >40 pF).

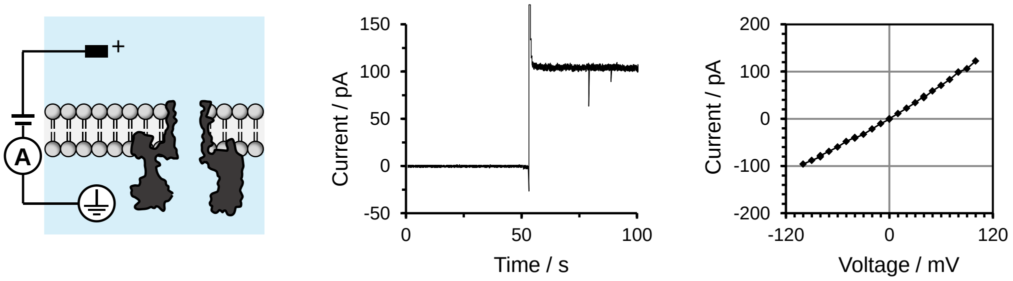


Supplemental Figure S1 Diagram of custom cell for planar lipid bilayer experiments. The cell consists of two mirror image Teflon blocks. Each block has a well bored into it, with a perpendicular channel connecting the two wells. Each well has two mitred channels at a 45° to allow for ease of access to the buffered well during experimentation. The blocks were held together using metal brackets. A tight seal was ensured by using silicon glue.

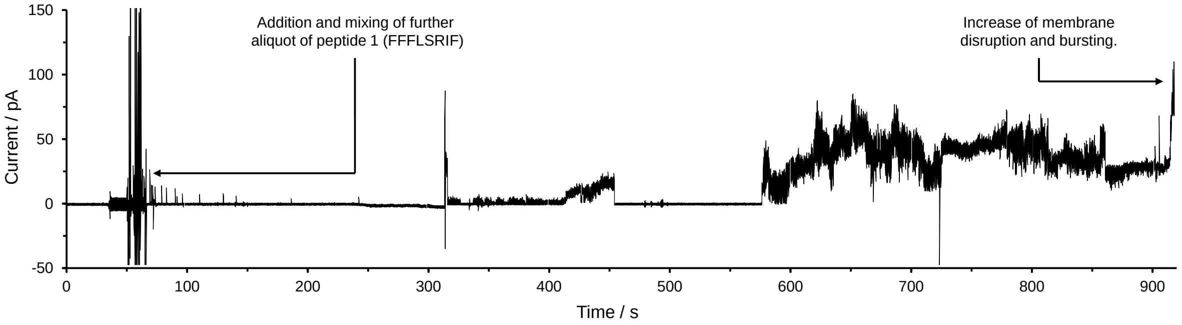
The membrane was characterized with successive 2 s sweeps under an applied potential ranging from +100 to −100 mV (Figure S2). The membrane seal was deemed acceptable if the range of current flow across the membrane measured <1 pA. Under an applied voltage (+10 mV), a solution of peptide (10 µL of a 50 µM solution, final concentration ~0.8 µM) was added to the *trans* well of a membrane-containing system.



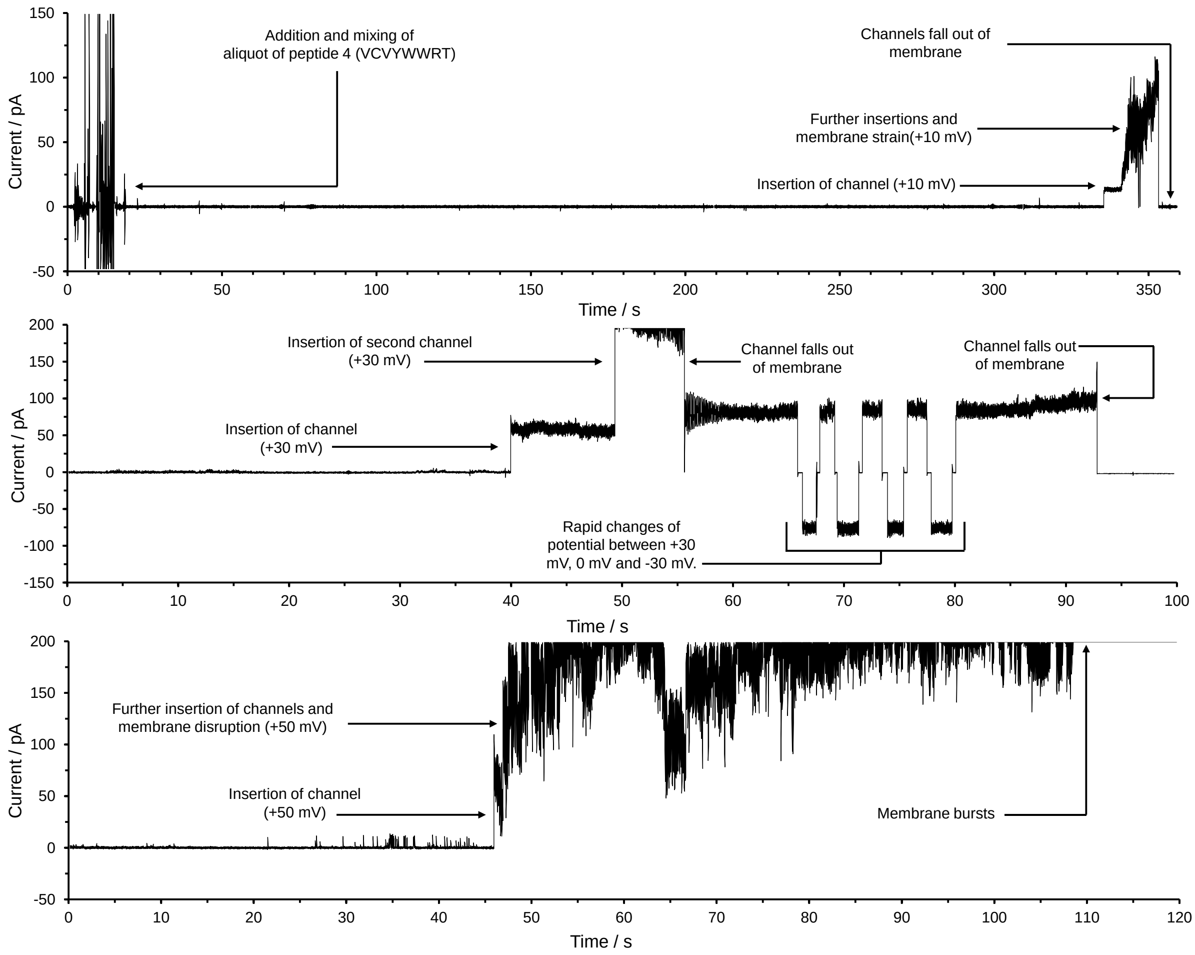
Supplemental Figure S2 Characterization of POPC Membranes. In order to interrogate the quality of the formed POPC bilayer a characterization is carried out by means of an IV sweep between the range -100 mV and +100 mV in 10 mV increments over 2 second sweeps. This protocol is repeated six times, and a mean average is calculated of the current at each value of voltage. A membrane with current flow deviation <1.5 pA is considered a pass and used for experiments, as demonstrated (left). In the case that there is current flow deviation >1.5 pA (as on right) this is considered a pass, and the membrane is zapped and reformed and the process repeated.



Supplemental Figure S3 Evidence of bilayer formation by insertion of a-HL transmembrane protein nanopore. Addition of α-HL (~0.2 µM) to a stable POPC phospholipid bilayer to isolate a single transmembrane channel (left). Current trace of channel insertion added to the ground well under a positive applied potential difference (+100 mV) in buffer (1 M KCl, 10 mM MOPS, pH 7.4) (middle). Current traces acquired at 2 kHz lowpass Bessel filter and x 50 output gain. Current voltage relationship of obtained α-HL single transmembrane channel stepwise from -100 mV to +100 mV (right).



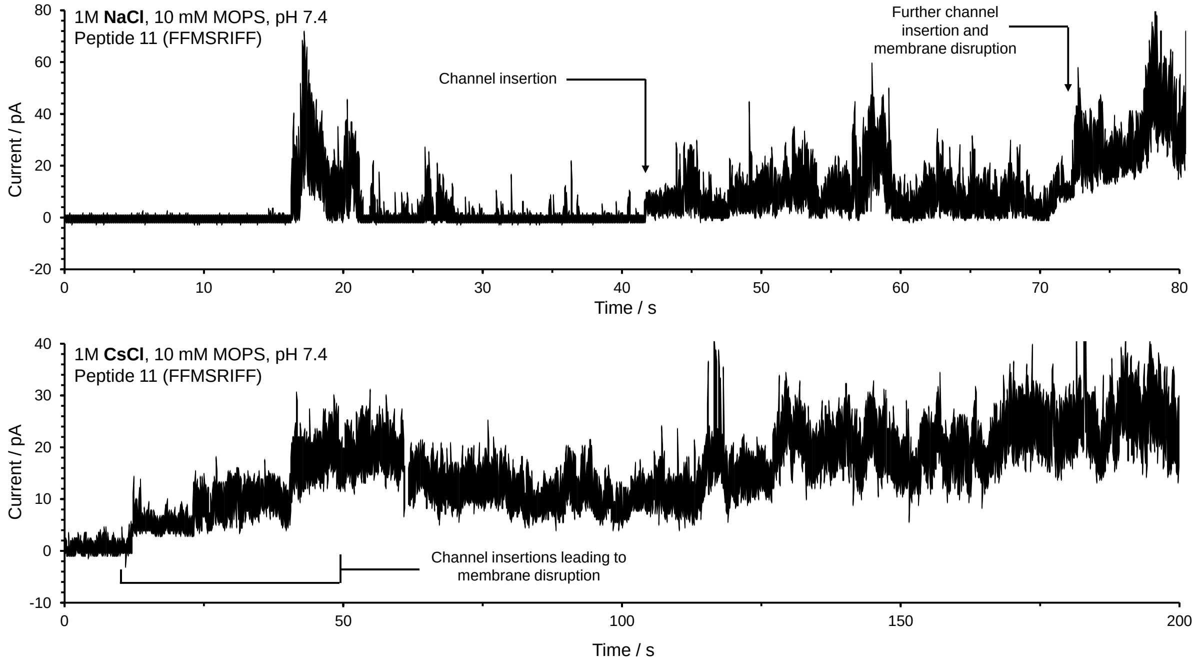
Supplemental Figure S4 Positive control peptide 1 FFFLSRIF. Current traces of membrane disruption with peptide 1 (FFFLSRIF) using 1 KCl, 10 mM MOPS, pH 7.4. Current traces acquired at 2 kHz lowpass Bessel filter and x 50 output gain. Membrane formed using DPhPC.



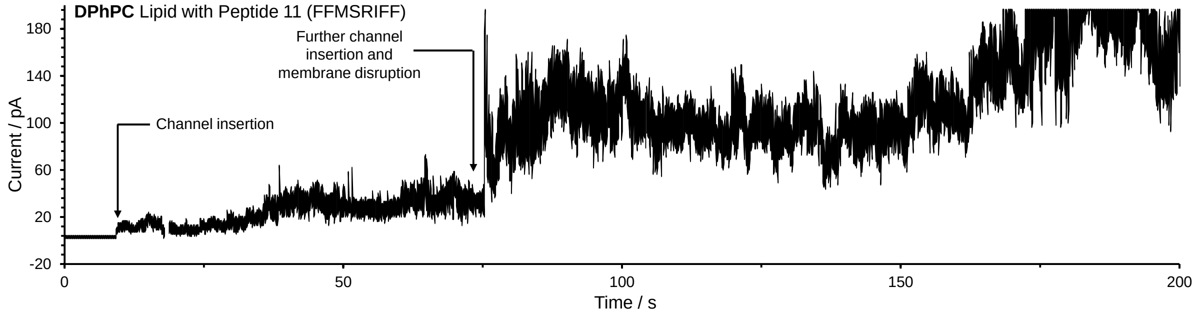
Supplemental Figure S5 Peptide 4 channel forming behavior. Current traces of membrane disruption with peptide 4 (VCVYWWRT) using 1 M KCl,10 mM MOPS, pH 7.4. Current traces acquired at 2 kHz lowpass Bessel filter and x 50 output gain. Traces visualized with lowpass filter type (Gaussian) with 750 Hz cut-off. Varying insertions observed at +10 mV (top), +30 mV (middle) and +50 mV (bottom) eventually leading to the membrane bursting.

1.3 FFMSRIFF, experimental pore-formation with alternative ions and phospholipids

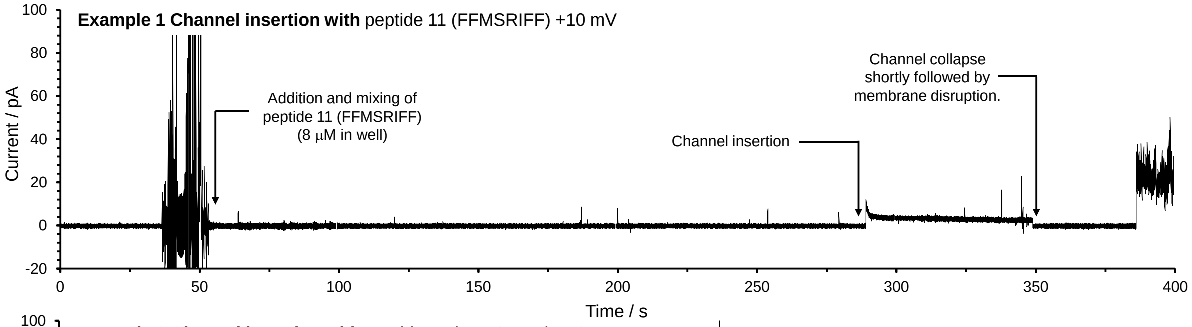
Membrane disruption sustained with variation of buffer components from KCl to NaCl and CsCl but stability of the channels was greatly decreased (Figure S6). Channel-like behavior was preserved with alternative phospholipid 1,2-diphytanoyl-*sn*-glycero-3-phosphocholine (DPhPC), a more prevalent candidate for single-channel studies as a result of increased stability over POPC[1] (Figure S7).

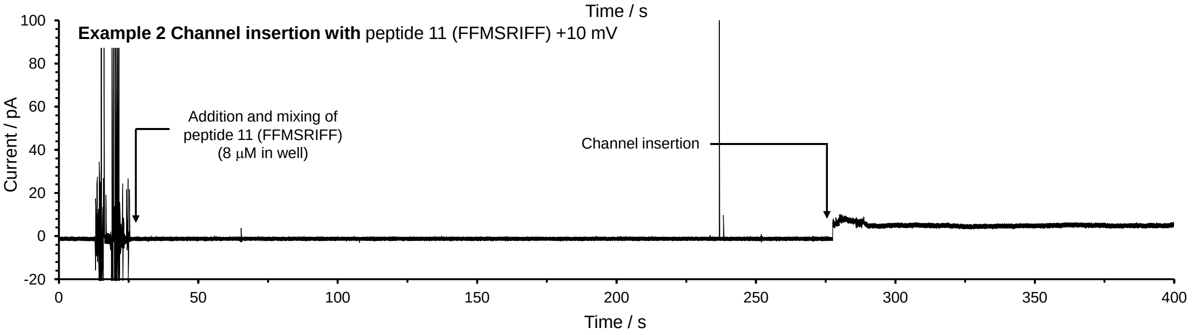


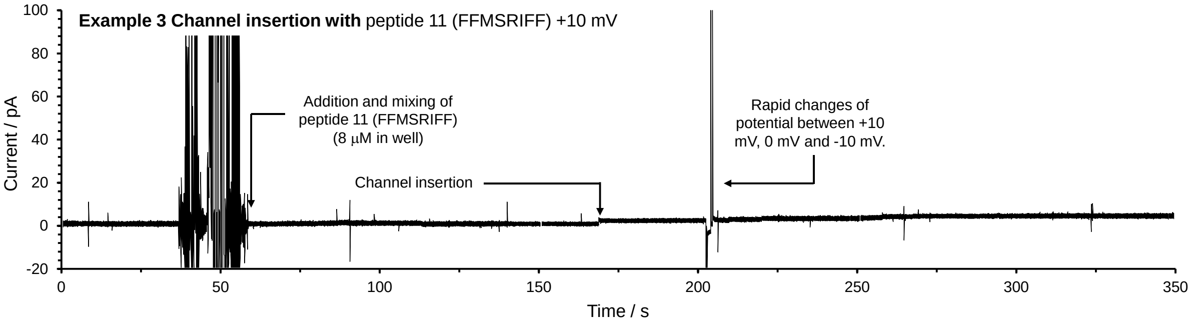
Supplemental Figure S6 Membrane disruption observed with octapeptide 11 FFMSIRFF under varying buffer concentration. Current traces of membrane disruption with peptide 11 (FFMSRIFF) using 1 M NaCl (top) and 1M CsCl (bottom) with 10 mM MOPS, pH 7.4. Current traces acquired at 2 kHz lowpass Bessel filter and x 50 output gain.

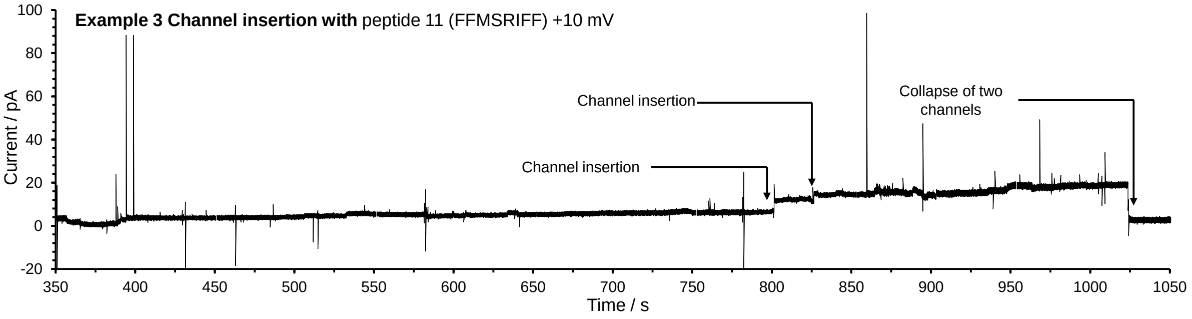


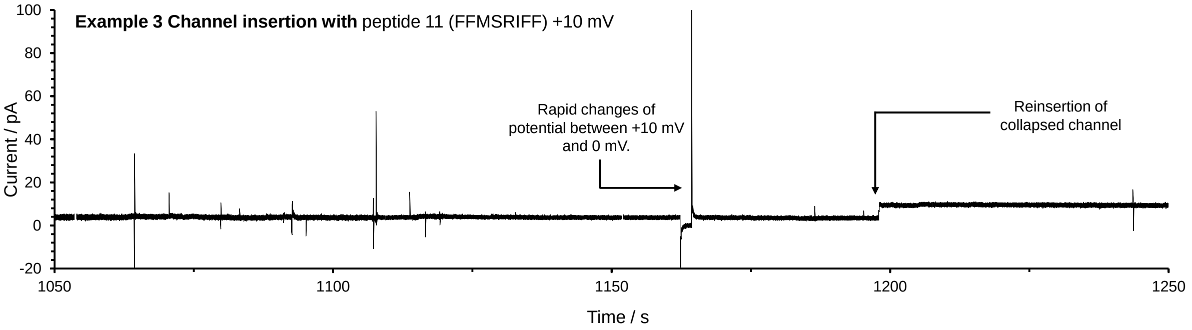
Supplemental Figure S7 Membrane disruption observed with octapeptide 11 FFMSIRFF with DPhPC lipid. Current traces of membrane disruption with peptide 11 (FFMSRIFF) using 1 KCl, 10 mM MOPS, pH 7.4. Current traces acquired at 2 kHz lowpass Bessel filter and x 50 output gain. Membrane formed using DPhPC.

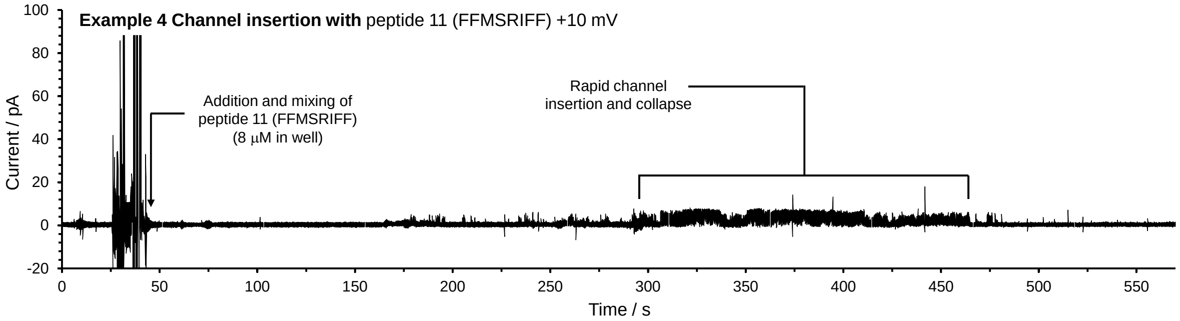

Example 1 Channel insertion with peptide 11 (FFMSRIFF) +10 mV
Addition and mixing of peptide 11 (FFMSRIFF) (8 μM in well)
Channel insertion
Channel collapse shortly followed by membrane disruption.
Current / pA
100
80
60
40
20
0
-20
Time / s
0
50
100
150
200
250
300
350
400

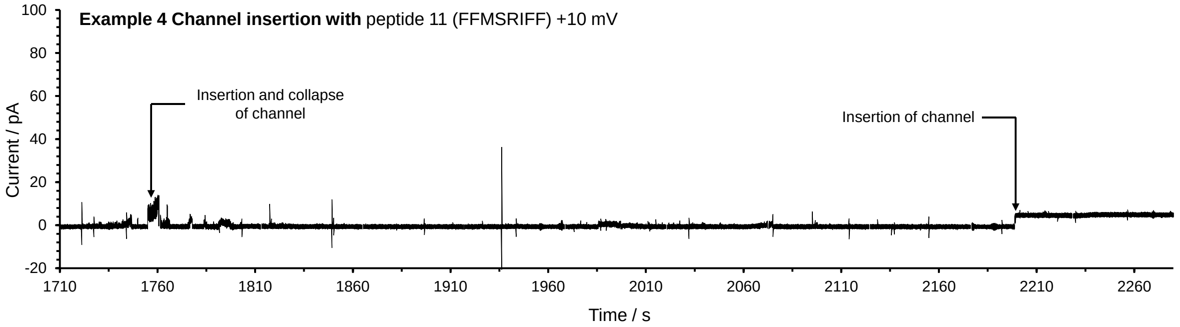

Example 2 Channel insertion with peptide 11 (FFMSRIFF) +10 mV
Addition and mixing of peptide 11 (FFMSRIFF) (8 μM in well)
Channel insertion
Current / pA
100
80
60
40
20
0
-20
Time / s
0
50
100
150
200
250
300
350
400

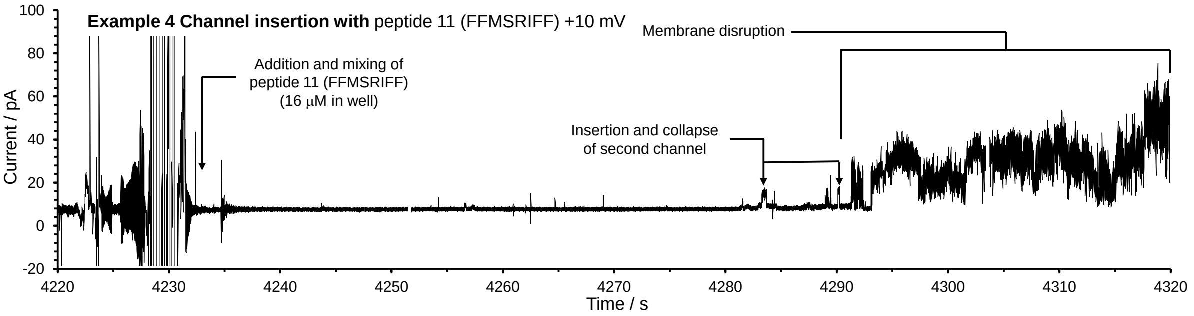

Example 3 Channel insertion with peptide 11 (FFMSRIFF) +10 mV
Addition and mixing of peptide 11 (FFMSRIFF) (8 μM in well)
Channel insertion
Rapid changes of potential between +10 mV, 0 mV and -10 mV.
Current / pA
100
80
60
40
20
0
-20
Time / s
0
50
100
150
200
250
300
350

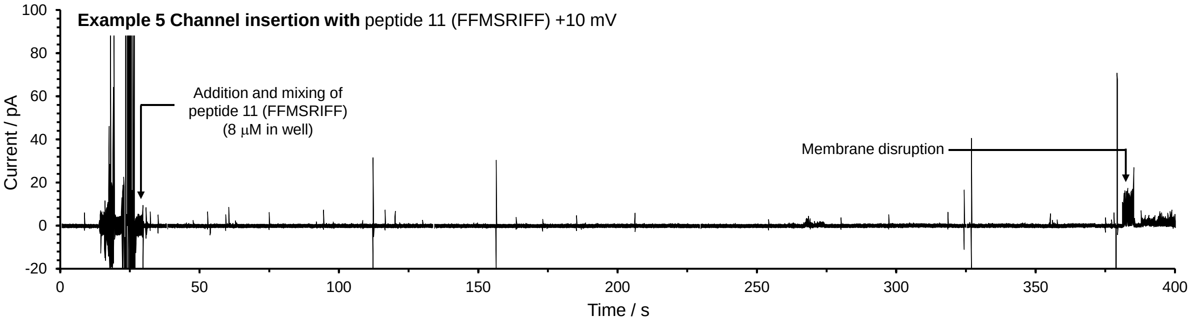

Example 3 Channel insertion with peptide 11 (FFMSRIFF) +10 mV
Channel insertion
Channel insertion
Collapse of two channels
Current / pA
100
80
60
40
20
0
-20
Time / s
350
400
450
500
550
600
650
700
750
800
850
900
950
1000
1050

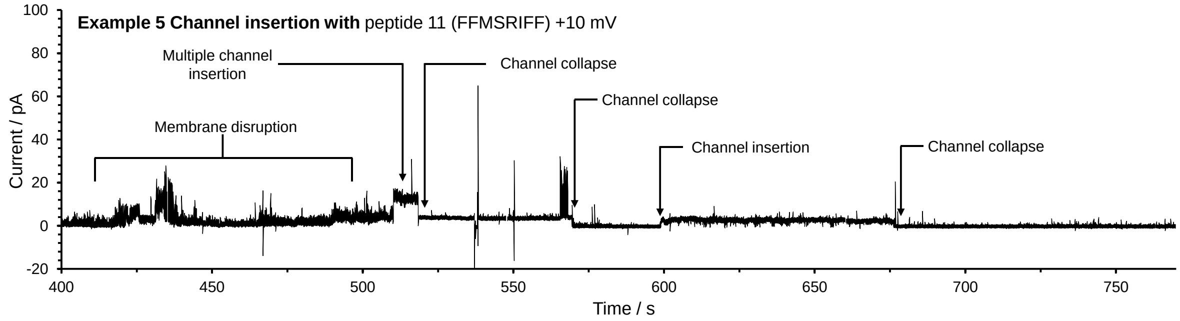

Example 3 Channel insertion with peptide 11 (FFMSRIFF) +10 mV
Rapid changes of potential between +10 mV and 0 mV.
Reinsertion of collapsed channel
Current / pA
100
80
60
40
20
0
-20
Time / s
1050
1100
1150
1200
1250

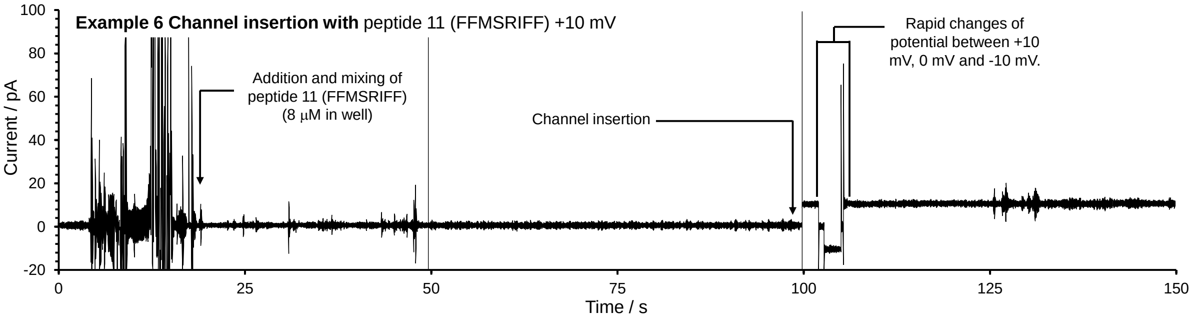

Example 4 Channel insertion with peptide 11 (FFMSRIFF) +10 mV
Addition and mixing of peptide 11 (FFMSRIFF) (8 µM in well)
Rapid channel insertion and collapse
Current / pA
100
80
60
40
20
0
-20
0
50
100
150
200
250
300
350
400
450
500
550
Time / s

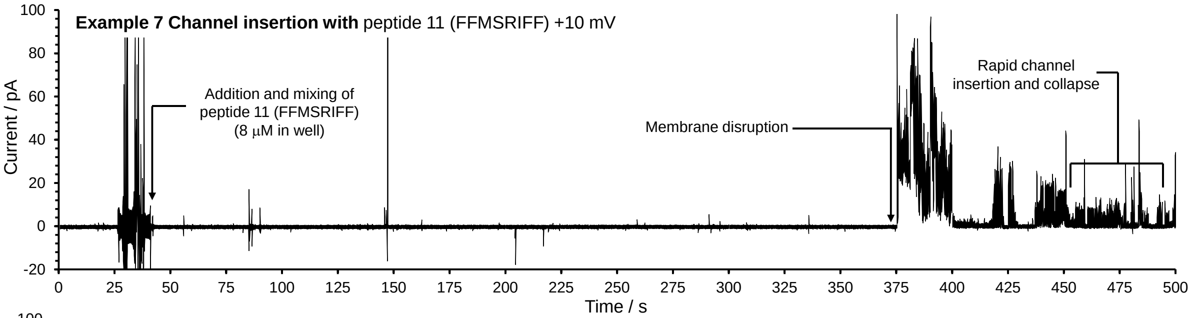

Example 4 Channel insertion with peptide 11 (FFMSRIFF) +10 mV
Insertion and collapse of channel
Insertion of channel
Current / pA
100
80
60
40
20
0
-20
1710
1760
1810
1860
1910
1960
2010
2060
2110
2160
2210
2260
Time / s

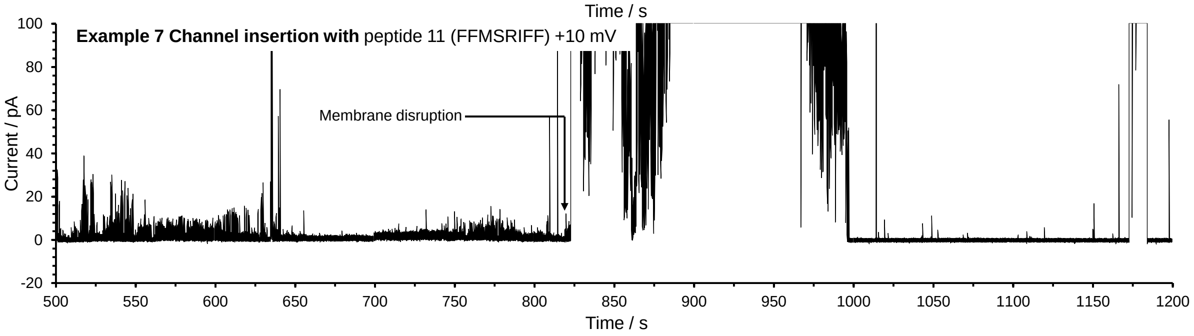

Example 4 Channel insertion with peptide 11 (FFMSRIFF) +10 mV
Membrane disruption
Addition and mixing of peptide 11 (FFMSRIFF) (16 µM in well)
Insertion and collapse of second channel
Current / pA
100
80
60
40
20
0
-20
4220
4230
4240
4250
4260
4270
4280
4290
4300
4310
4320
Time / s

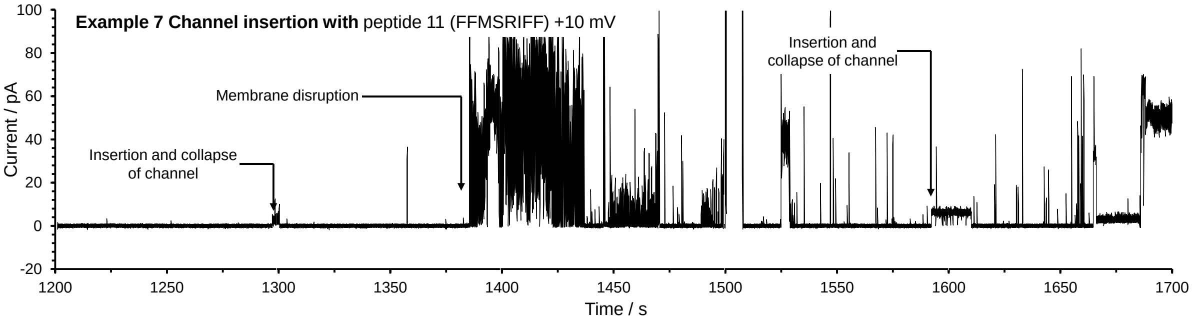

Example 5 Channel insertion with peptide 11 (FFMSRIFF) +10 mV
Addition and mixing of peptide 11 (FFMSRIFF) (8 µM in well)
Membrane disruption
Current / pA
100
80
60
40
20
0
-20
0
50
100
150
200
250
300
350
400
Time / s

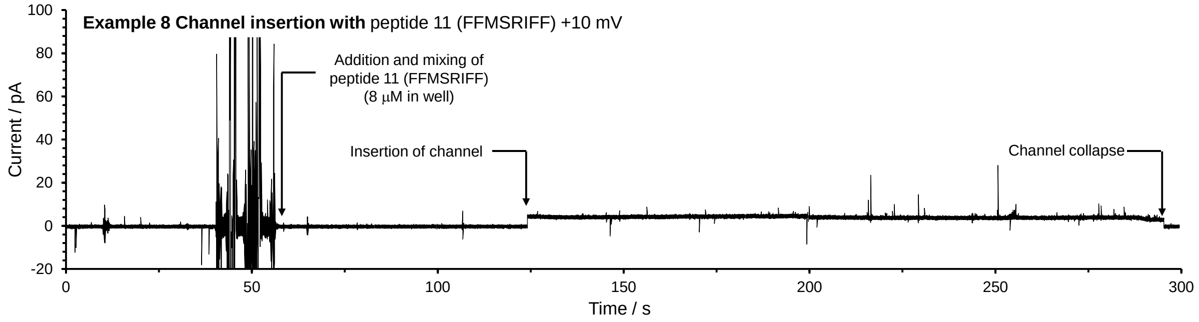

Example 5 Channel insertion with peptide 11 (FFMSRIFF) +10 mV
Multiple channel insertion
Channel collapse
Channel collapse
Membrane disruption
Channel insertion
Channel collapse
Current / pA
100
80
60
40
20
0
-20
400
450
500
550
600
650
700
750
Time / s


Example 6 Channel insertion with peptide 11 (FFMSRIFF) +10 mV
Addition and mixing of peptide 11 (FFMSRIFF) (8 μM in well)
Channel insertion
Rapid changes of potential between +10 mV, 0 mV and -10 mV.
Current / pA
100
80
60
40
20
0
-20
0
25
50
75
100
125
150
Time / s

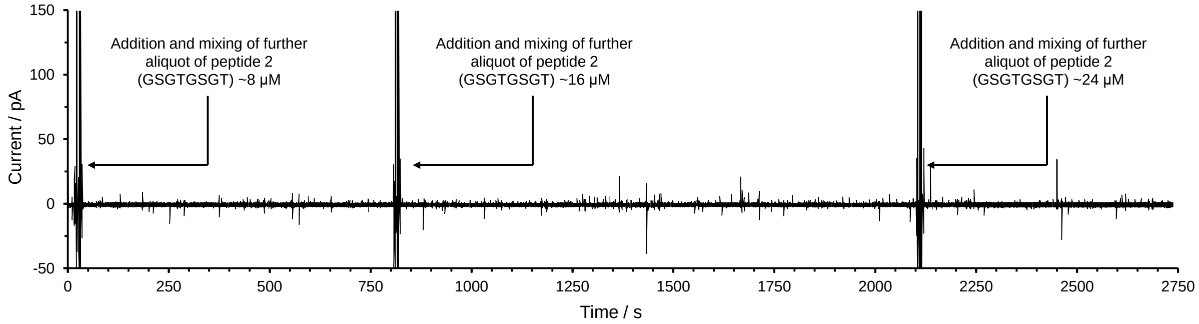

Example 7 Channel insertion with peptide 11 (FFMSRIFF) +10 mV
Addition and mixing of peptide 11 (FFMSRIFF) (8 μM in well)
Membrane disruption
Rapid channel insertion and collapse
Current / pA
100
80
60
40
20
0
-20
0
25
50
75
100
125
150
175
200
225
250
275
300
325
350
375
400
425
450
475
500
Time / s


Example 7 Channel insertion with peptide 11 (FFMSRIFF) +10 mV
Membrane disruption
Current / pA
100
80
60
40
20
0
-20
500
550
600
650
700
750
800
850
900
950
1000
1050
1100
1150
1200
Time / s

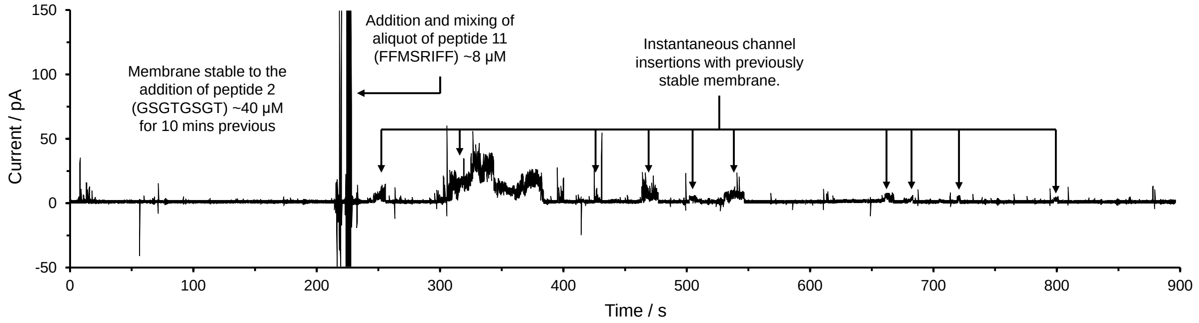

Example 7 Channel insertion with peptide 11 (FFMSRIFF) +10 mV
Membrane disruption
Insertion and collapse of channel
Insertion and collapse of channel
Current / pA
100
80
60
40
20
0
-20
1200
1250
1300
1350
1400
1450
1500
1550
1600
1650
1700
Time / s



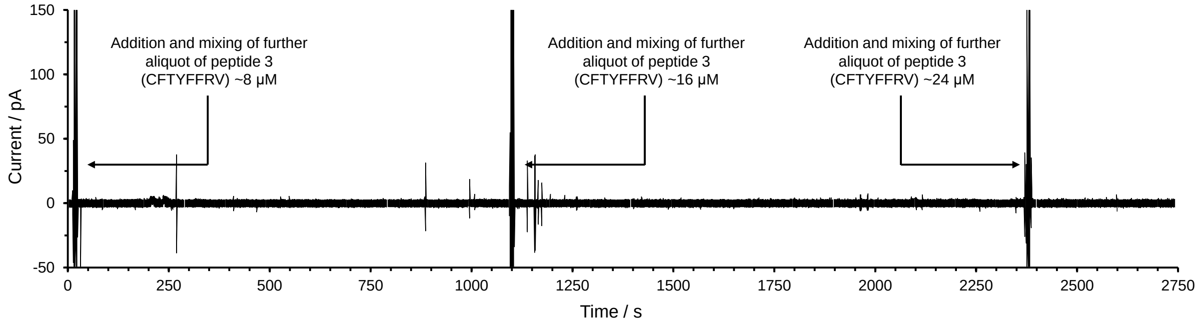
Supplemental Figure S8 Further examples of channel insertion with peptide 11 FFMSRIFF. Current traces of membrane disruption with peptide 11 (FFMSRIFF) using 1 M KCl,10 mM MOPS, pH 7.4. Current traces acquired at 2 kHz lowpass Bessel filter and x 50 output gain. Traces visualized with lowpass filter type (Gaussian) with 750 Hz cut-off. Examples of eight experiments shown each demonstrating channel forming behavior and membrane disruption.



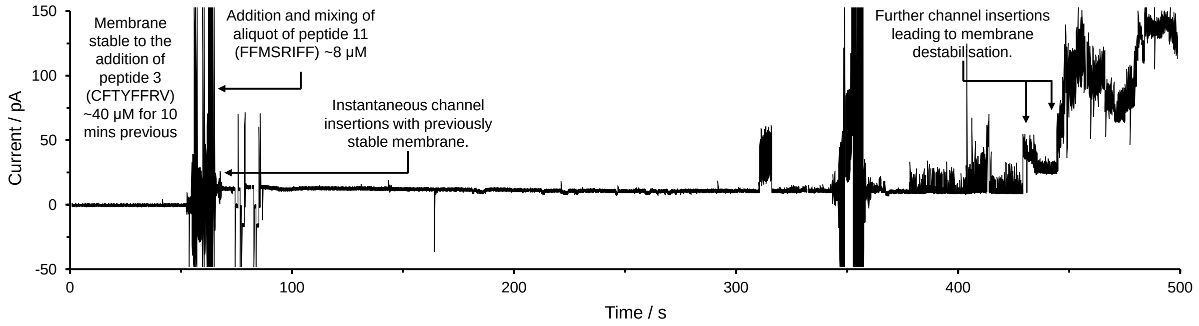
Supplementary figure S9 Additional data with negative control peptide 2. Current traces of stable membrane over multiple additions with negative control peptide 2 (GSGTGSGT) using 1 M KCl,10 mM MOPS, pH 7.4. Current traces acquired at 2 kHz lowpass Bessel filter and x 50 output gain. Traces visualised with lowpass filter type (Gaussian) with 750 Hz cuttoff.



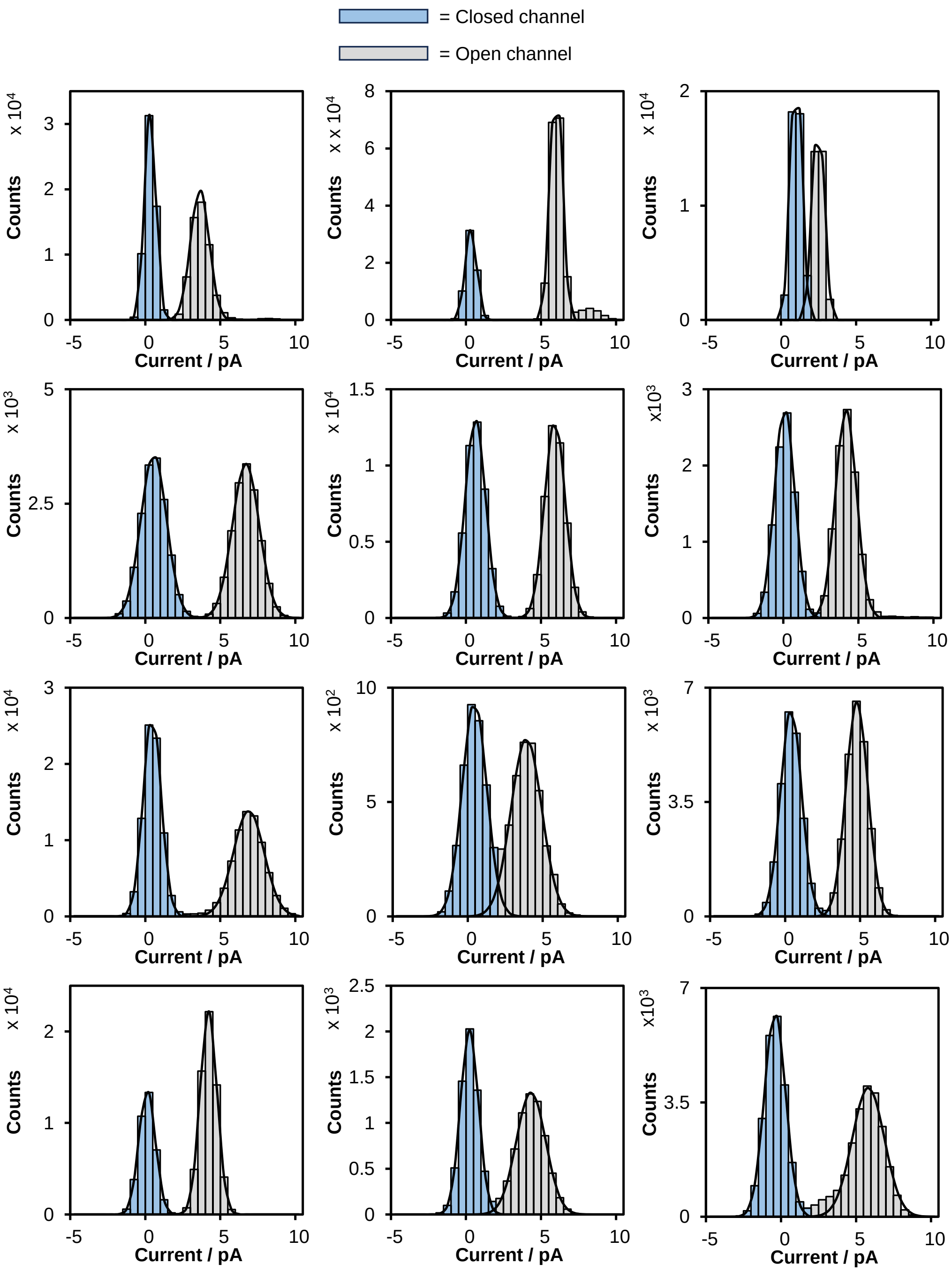
Supplementary figure S10 Additional data with negative control peptide 2 and then peptide 11. Current traces of stable membrane over multiple additions with negative control peptide 2 (GSGTGSGT) which then demonstrate sporadic channel formation following the addition of peptide 11 (FFMSRIFF) using 1 M KCl,10 mM MOPS, pH 7.4. Current traces acquired at 2 kHz lowpass Bessel filter and x 50 output gain. Traces visualised with lowpass filter type (Gaussian) with 750 Hz cuttoff.



Supplementary figure S11 Additional data with negative control peptide 3. Current traces of stable membrane over multiple additions with negative control peptide 3 (CFTYFFRV) using 1 M KCl,10 mM MOPS, pH 7.4. Current traces acquired at 2 kHz lowpass Bessel filter and x 50 output gain. Traces visualised with lowpass filter type (Gaussian) with 750 Hz cuttoff.



Supplementary figure S12 Additional data with negative control peptide 3 then peptide 11. Current traces of stable membrane over multiple additions with negative control peptide 3 (CFTYFFRV) which then demonstrate sporadic channel formation following the addition of peptide 11 (FFMSRIFF) using 1 M KCl,10 mM MOPS, pH 7.4. Current traces acquired at 2 kHz lowpass Bessel filter and x 50 output gain. Traces visualised with lowpass filter type (Gaussian) with 750 Hz cuttoff.



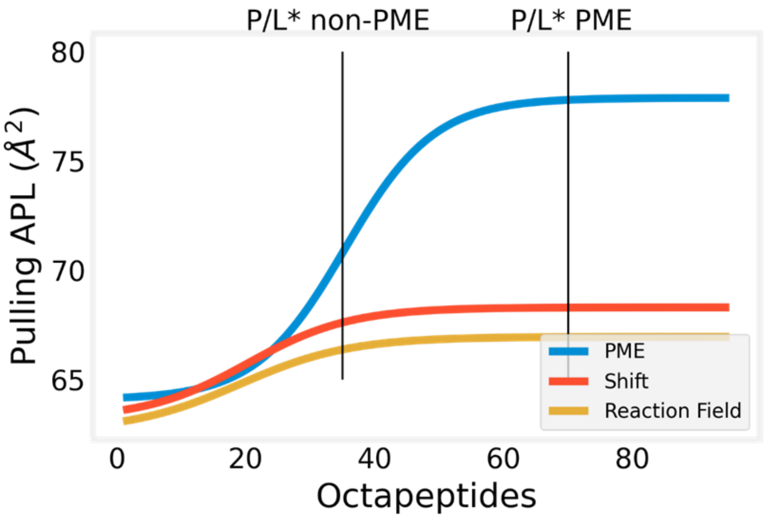
Supplementary figure S13 Inspection of channel insertion conductance with peptide 11. Residual ion current histogram distributions of twelve examples of channels observed from the addition of peptide **11** at +10 mV in 1 M KCl, 10 mM MOPS, pH 7.4. Closed channel (blue) and open channel (grey) current distributions interrogated with normal distributions determined by analysis with Solver plugin.

2. Computational Simulations

2.1 Coarse-grained molecular dynamics

The MARTINI forcefield (version 2.1) parameters for coarse-grained (CG) peptides, phospholipids, water & ions are used with a predefined α-helical secondary peptide structure[2, 3]. The peptide atoms are mapped one-to-four in corresponding heavy atoms-to-beads, water beads represent four water molecules for the purpose of computational efficiency and the ion beads represent one ion. This causes an inevitable loss of detail but leaves a much more computationally efficient method of studying biological systems as the atom properties (polarity, molecular shape, bond lengths etc.) are implied *via* the coarse-grained representation. However, water does not enter pores in the simulations performed herein due to the coarse-grained model being too bulky to allow water to enter. Another consequence of this is that many of our results are sinking rafts as barrel-stave is only accessible if a single peptide can span the bilayer.

Each 12.5×12.5×17.0 nm NPT box was setup with a bilayer of 360 POPC and 90 POPS phospholipids in a bilayer using INSANE[4]. 80 zwitterionic peptides were inserted within 2 nm of the top side of the equilibrated bilayer with at least 0.3 nm between their centers of geometry (COG) with the box neutralized with Na/Cl ions and solvated with MARTINI water. The temperature and pressure were kept constant at 323 K (higher temperature used help to perturb the bilayer) [5] & 1 bar respectively *via* a v-rescale thermostat & Berendsen semiisotropic barostat[6, 7]. Bond lengths between backbone and side-chain for peptides I, V & Y as well as within aromatic side chains were constrained *via* the LINCS algorithm.[8] The boxes were minimized using the steepest descent integrator and steered molecular dynamics (SMD) was used to pull peptides. PME electrostatics with an order of 4 and Fourier spacing of 0.12 nm was used for computing long-range electrostatic interactions (Figure S14). Due to the relationship between the diffusion constants of the MARTINI coarse-grained and atomistic simulations, the effective simulation time is 4 times greater than the formal simulation time. Herein we refer to the effective simulation time and not the formal time.



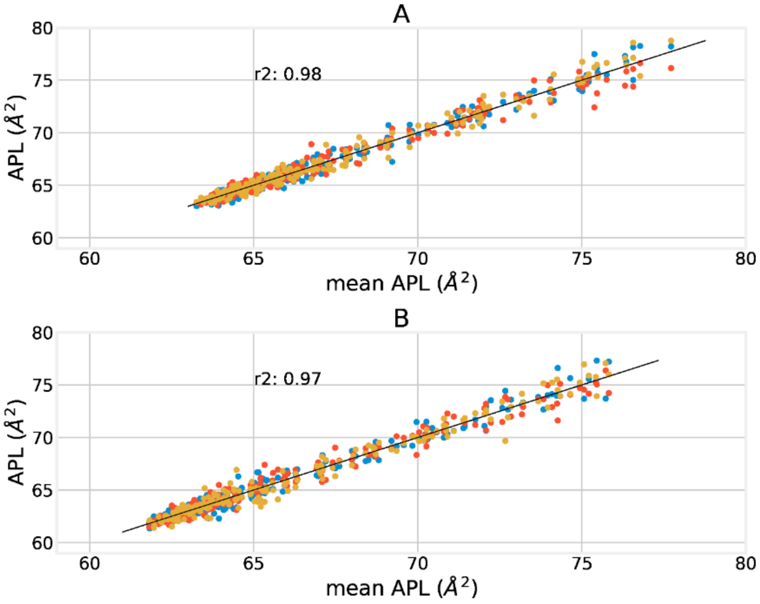
Supplemental Figure S14 Molecular dynamics electrostatics methods and their effect on peptide-bilayer binding. A segment of melittin (ALISWIKR) was added to POPC/POPS bilayer simulations using different electrostatic evaluation methods (PME, shift & reaction field). The P/L* ratios for the short-ranged methods (shift & reaction field) are around 35 while PME allows far more peptides (~ 60) to adsorb onto the surface while increasing APL.

2.2 Steered coarse-grained molecular dynamics

Herein we will use a combination of SMD and CGMD for fast and accurate evaluation of peptides that cause increased perpendicular pressure on phospholipid bilayers and increased APL[9]. Peptides are placed within 2 nm of the bilayer and pulled along the Z-axis to allow them to quickly aggregate on the surface. SMD was used to pull peptides (*via* umbrella sampling) in the Z-dimension towards the bilayer at a constant rate of 0.1 kJmol^{-1}nm^{-2} for 400 ns, followed by relaxation for 400 ns and finally CpHMD for up to 2,000 ns. This technique has previously been used to decrease the equilibration portion of simulations where peptides find their way to the surface of a bilayer[5, 10-13]. Following this we relax the simulation by removing the pulling forces, enabling peptides to dissociate.

2.2.1 SMD repeatability

With steered molecular dynamics (SMD) there is known to be a lower degree of reproducibility compared with equilibration, therefore when fitting hyperparameters or for any kind of measurement of parameter usefulness we take the averages of 3 runs and use the average of 2 runs for the active learning algorithm. We do find however that since we apply only a very weak force constant our results are quite repeatable (Figure S15).



Supplemental Figure S15 SMD repeatability. SMD results are highly reproducible between the three runs (blue, orange, red) of the random set of 200 systems in both the pulling (A) and relaxation stages (B).

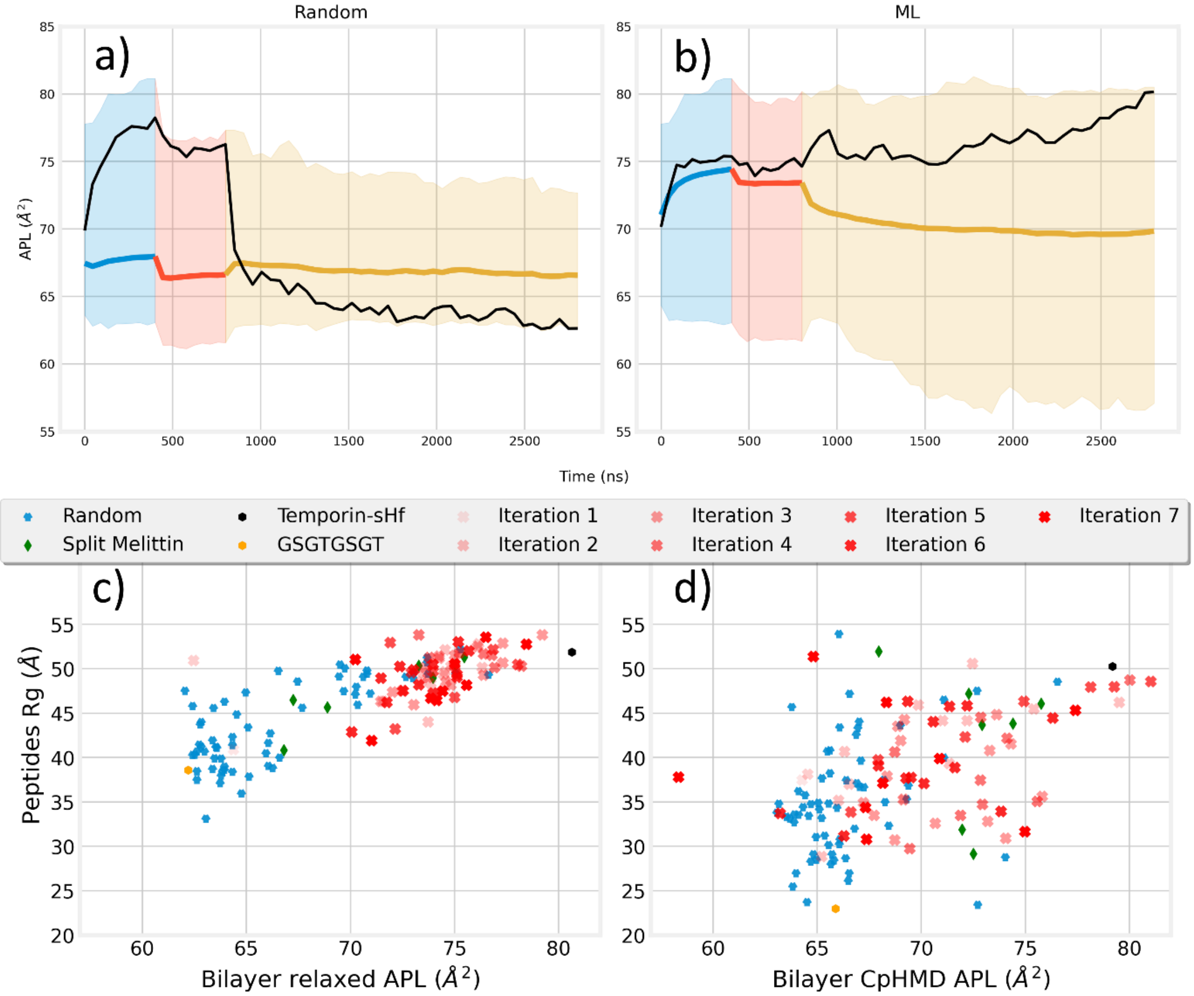
2.3 Constant pH molecular dynamics

To capture a wider range of transitions than are constricted by probability in the original CpHMD algorithm proposed by Radak *et al.* we modified this algorithm, such that the weighted choice of which titratable residue to titrate ($\lambda \rightarrow \lambda'$) is determined by the distance from the bilayer center rather than the theoretical pK_a. Temperature (311 K) and semiisotropic pressure (1.01325 bar) were coupled using Langevin dynamics [14, 15] and a modified Nose-Hoover Langevin piston.[16] The electrostatic interactions where evaluated using particle-mesh Ewald summation and non-bonded LJ terms by a cutoff at 1.2 nm with a shifted modified beginning at 1.0 nm. Each simulation was run at pH 7.0 for up to 2,000 ns or until APL had converged which was made up of 10,000 iterations of 2,000 time steps of 25 fs, at each CGMD iteration a non-equilibrium switch between two titratable states is proposed and run for 200 steps.

$$Weights = \frac{1}{\left||res_z - bilayer_z|\right|} : res \ \in titratable\ residues$$

2.4 Pore-formation *via* CpHMD

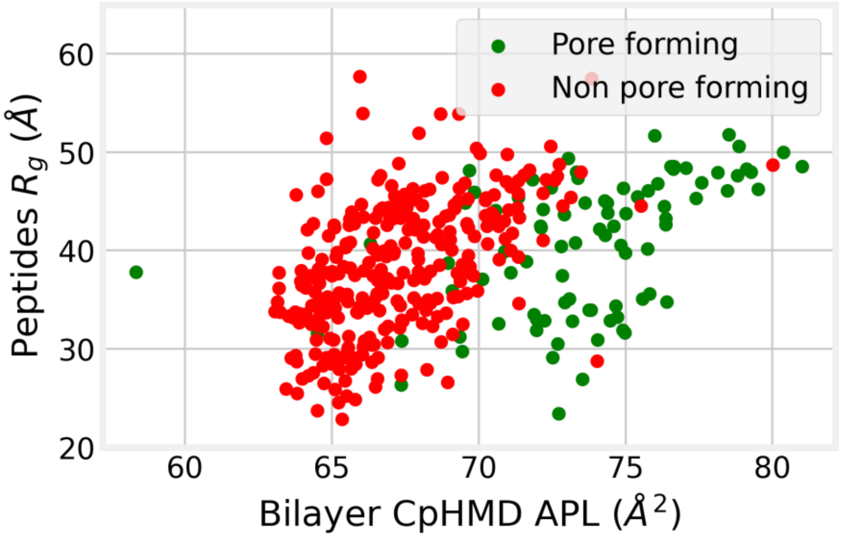
Very few of the SMD+CGMD simulations predicted pore-forming behavior. Hence, CpHMD was used to further simulate the systems starting from the final frame of each CGMD simulation. The CpHMD simulation were found to act as a discriminator of pore-forming behavior; it progresses the formation of pores for some systems while for others not only do pores not form, but peptides also tend to dissociate slightly from the bilayer if they are non-pore-forming (Figure S16 C, D).



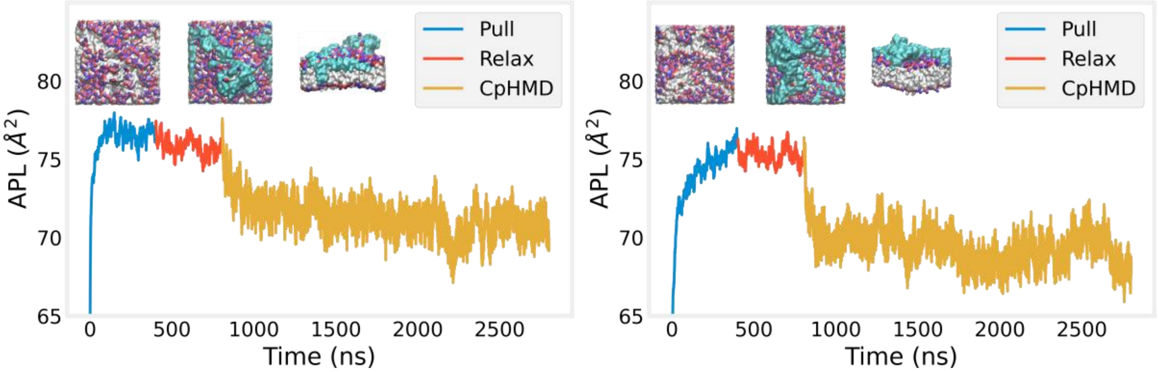
Supplemental Figure S16 Active learning vs random screening with CpHMD discriminator. a,b) Bold lines show the mean APL for the three stages, pulling (blue), relaxing (red), CpHMD (yellow) of peptide-bilayer simulation while the filled area shows the range of the random set of octapeptides (a) and machine learning selected (b). The black line is a selected peptide (a: CIIWKWFT, b: FGYVLIRI) from each set demonstrating how CpHMD can significantly alter the result. c,d) Combining APL and R_g can help to identify PFPs, every octapeptide in the top right corner (d) forms pores or carpets membranes when CpHMD is performed, comparison of (c) & (d) shows that CpHMD is much better at discriminating between pore-forming and inactive octapeptides. The anomalous result of where an octapeptide (FFMSIRFF) has a very low (< 60) APL yet is pore-forming demonstrates a short coming of the APL descriptor.

2.5 APL as a metric

We scored the ability of each peptide to form transmembrane pores in the CpHMD simulations by assessing the bilayer area-per-lipid (APL). The use APL is based on work by Huang[17] who proposed that transmembrane peptide pores assemble when the peptide-to-lipid ratio (P/L) reaches a critical point (P/L*). According to this model, thinning of the membrane increases in proportion to the P/L ratio, but remains constant above the critical ratio when pores form. Since membrane thinning increases APL, we anticipated that the APL should provide a measure of the pore-forming ability of a peptide. Hence, we used the APL to score the ability of a peptide to form transmembrane pores. Additional parameters such as radius of gyration (R_g, Figure S17), solvent accessible surface area (SASA, Figure S19) and aggregation propensity (AP, Figure S20).

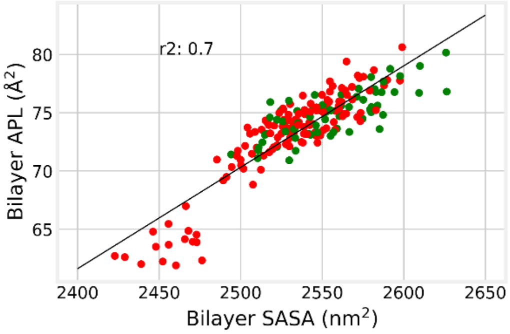


Supplemental Figure S17 APL and R_g predictiveness of pore-forming. Pore-forming ability is strongly correlated with APL but not R_g, however the R_g measurement can be useful in determining the nature of the pore(s) formed.



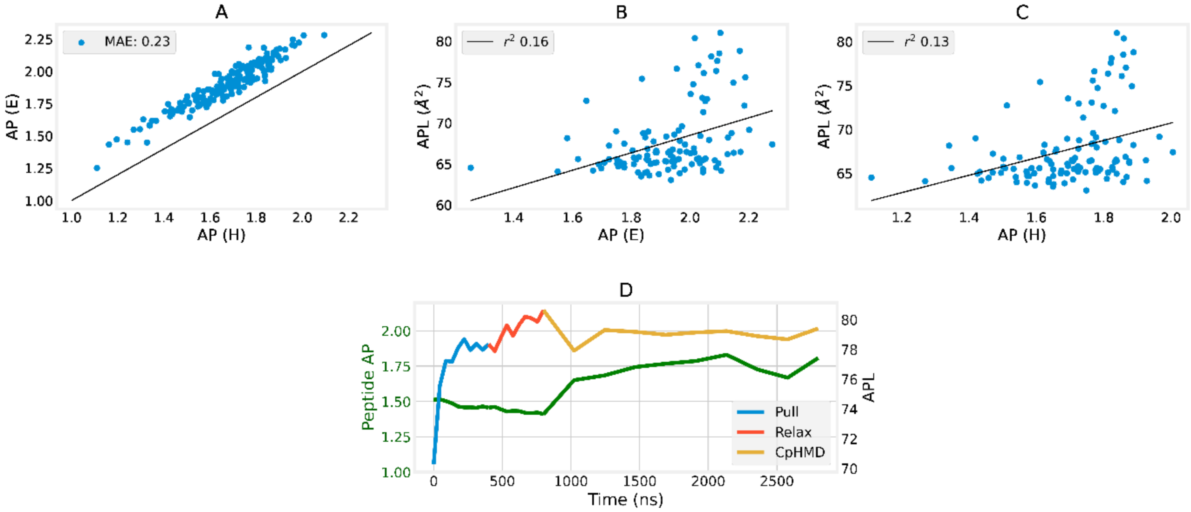
Supplemental Figure S18 Validation against PFPs found in nature. Left) Melittin binds with the bilayer surface and forms a pore. Right) LL-37 binds with the bilayer surface but does not enter the bilayer. Both simulations have very similar APL profiles, but manual inspection shows the LL-37 does not form pores while Melittin does.

We also compared with APL as a metric with solvent accessible surface area (SASA) of the bilayer. Both of these metrics are highly correlated and seem to be able to provide some indication of pore-formation though neither are definitive (Figure S19).



Supplemental Figure S19 Comparison of using APL and SASA as measurements of bilayer disruption. While APL is only indicative of pore-forming, the SASA measurement is equally only indicative of pore-formation. This is shown by the red (non-pore-forming) and green (pore-forming) peptide-bilayer system measurements having no clear cut-off point using either method.

We found that the aggregation propensity (AP) of peptides was weakly correlated with the APL score (Figure S20). This result is biased by the starting positions of the peptides being close to the bilayer and close to each other.



Supplemental Figure S20 Area per lipid (APL) vs aggregation propensity (AP). 115 octapeptides randomly selected from the machine learning, random and validation sets. The potential for these peptides to aggregate in water was simulated using the method described by Frederix *et al.*[18] using both helical (H) and extended (E) secondary structure. The aggregation score (AP) was used to quantify the degree of aggregation experienced by each system. A) We find that AP scores for H/E secondary structure is highly correlated, B, C) however neither AP scores correlate strongly with APL with low r² scores for a linear trendline. D) This correlation was also compared over the duration of peptide **1** simulations showing only a minor increase in AP score compared to a much larger increase in APL.

2.6 Phospholipid bilayers vs stands-ins.

There has been discussion of using model substitutes for phospholipids in peptide-bilayer simulations such as cylamine/octanoic acid, octanol/water, octadecanoic acid/octadecylamine[19]. This is done often to decrease equilibration time as the transition from out of the bilayer to inside of it crosses over a small transition so that the proportions of what's on each side can be realized in less time. However, we find this is inappropriate for searching for PFPs as these, despite being simple peptides, are complex molecules which may be unable to pass the transition barrier formed by the head groups.

3. Machine learning / Artificial intelligence

3.1 Active learning cycle

We applied our active-learning cycle with three different model combinations (Table S1). Each iteration consists of scanning the entire octapeptide (20^8) dataset using the Judred parameters via the cheap model, this is done by loading in each "chunk" of the dataset (2,560,000 rows) and selecting the peptide with highest predicted APL (total of 10,000 peptides). For these 10,000 peptides the higher resolution parameters are generated and a second APL prediction is made via the expensive model. From here either the top 10 are taken or a Monte Carlo (MC) function is applied to the APL predictions by multiplying all values by a random number between 0 and 1 from a uniform distribution.

$$\mathrm{MC} = \mathrm{APL} \times \mathrm{U}(0,1)$$

This has the effect of randomly dropping out top performers to increase diversity while acting on a large enough set so as not to promote any low-to-mid performers to the top of the list. This helps to escape local maxima. These top 10 are simulated and the resulting mean APL measurement for each system (run in duplicate) are used to retrain both models and make the next iteration of prediction. The combinations of machine learning algorithms tested for the Judred dataset screening and higher resolution dataset are given in Table S2.

Table S2. Machine learning model combinations. Three combinations of machine learning models used in this study and whether a MC step was used.

Model	Cheap model	Expensive model	Monte Carlo
A	ETR	ETR	X
B	ETR	ETR	✓
C	XGB	ETR	✓

Many high APL octapeptides were discovered. However, without the Monte Carlo function, model A over-selected for octapeptides containing Tyr, with 66/70 containing at least one Tyr residue and 29/70 containing a YYYY motif. This is due to highly localized optimization and not desirable as we aim to discover PFPs with a diverse range of sequences to better understand sequence-dependent activity.

3.2 Machine learning peptide descriptors

Given the previously reported issues with generating large datasets of descriptors for the scales of searching employed herein (octapeptides, 20^8) using methods that require generating complex molecular representations (Mordred[20], PyBioMed[21], PyMolSAR[22] & PaDELPy[23, 24]), we utilize a two-stage strategy that allows the entire octapeptide (20^8) space to be examined at a lower resolution (Judred parameters) and to then re-examine with on-the-fly generated higher resolution parameters (Table S3). This method has previously been successfully deployed in iteratively learning which peptides will self-assemble in peptide space up to hexapeptides[25]. To generate the high-resolution dataset, the five molecular descriptor (Judred, Mordred, PyBioMed, PyMolSAR, PaDELPy) programs were used to generate exhaustive (1330) parameters for 200 randomly selected octapeptides and reduced *via* 5-CV reductive feature elimination to 20 remaining features.

Enormous amounts of disk and RAM space are saved by implicit indexing based on position rather than storing peptide sequences, or even explicit indexes. This is achieved by Judred[25] preparing the dataset in an ordered fashion such that any peptide sequence can be calculated by knowing its position in the dataset and the length of the peptide. Since the entire dataset is screened on each iteration this is a feasible scheme that is employed throughout.

The Judred parameters were generated for each peptide with the addition of the isoelectric point (pI) using CuPY[26] and Apache Parquet[27], with a chunk size of 2,560,000, due to the trade-off in calculation speed and RAM size requirements. The source code which has been made available at https://github.com/avanteijlingen/Judred.

Table S3. Machine learning model metrics comparison. RMSE and r2 scores for various machine learning algorithms evaluated for use within machine learning. Each model was trained on 200 randomly selected octapeptides and their associated APL scores and validated against a previously unseen validation set of 72 systems with above average APL. Both the ability to predict APL from Judred and Mordred parameters are measured and considered.

	RMSE		r2	
	Train	High APL (Validation)	Train	High APL (Validation)
Judred				
Extra trees regressor	1.25	1.93	0.89	0.51
Gradient boosted decision trees	0.57	1.82	0.98	0.57
Random Forest	1.63	2.45	0.82	0.22
Decision tree regressor	1.3	2.13	0.89	0.41
Linear SVM	3.36	4.74	0.24	-1.92
Linear regressor	2.19	2.74	0.68	0.03
Gaussian process regressor	1.23	2.48	0.9	0.21
Elastic Net	3.29	4.39	0.27	-1.5
$\text{NuSVM}_{\text{rbf}}$	4.06	3.37	-0.11	-0.48
SVM_{rbf}	2.36	3.13	0.62	-0.27
SVM_{poly}	2.23	2.9	0.66	-0.09
Ridge regressor	2.41	2.94	0.61	-0.12
Multi-layer perceptron	1.52	2.1	0.84	0.43
Mordred				
	Train	High APL (Validation)	Train	High APL (Validation)
Extra trees regressor	0.97	1.74	0.94	0.61
Gradient boosted decision trees	0.81	1.88	0.96	0.54

Random Forest	1.13	1.95	0.91	0.51
Decision tree regressor	1.11	2.06	0.92	0.45
LinearSVM	2.21	2.64	0.67	0.09
Linear regressor	1.85	2.78	0.77	0
Gaussian process regressor	0.46	3.38	0.99	-0.48
Elastic Net	2.62	3.5	0.54	-0.59
NuSVM$_{rbf}$	4.12	3.56	-0.15	-0.64
SVM$_{rbf}$	3.2	3.68	0.31	-0.76
SVM$_{poly}$	2.97	3.77	0.4	-0.85
Ridge regressor	3.42	4.06	0.21	-1.14
Multi-layer perceptron	2.52	4.4	0.57	-1.51
Mean of predicted values				
	Train	High APL (Validation)	Train	High APL (Validation)
Extra trees regressor	1.07	1.75	0.92	0.6
Gradient boosted decision trees	0.68	1.83	0.97	0.56
Random Forest	1.32	2.09	0.88	0.43
Decision tree regressor	1.05	1.87	0.93	0.54
LinearSVM	2.52	3.49	0.57	-0.58
Linear regressor	1.94	2.68	0.75	0.07
Gaussian process regressor	0.76	2.49	0.96	0.2
Elastic Net	2.76	3.8	0.48	-0.87
NuSVM$_{rbf}$	4.08	3.46	-0.12	-0.55
SVM$_{rbf}$	2.65	3.27	0.53	-0.39
SVM$_{poly}$	2.43	3.14	0.6	-0.28
Ridge regressor	2.71	3.33	0.51	-0.44
Multi-layer perceptron	1.74	2.94	0.8	-0.12

3.3 Machine learning results

Table S4 and S5 contain the results from the machine learning algorithms (Models A-C), Figures S16 &S17 contain top-down snapshots of ML selected bilayers at the end of their CpHMD simulations.

Table S4. Pore-forming peptides found via active learning. The 71 PFPs found with the active learning and the model combination that found them, their disruption mechanism and APL and R_g values. Short peptides are not able to form pores via toroidal or barrel-stave mechanisms as this requires that the peptides enter one-by-one and that they are able to individually span the bilayer. From the 160 position specific residues, many from the top 20 identified occur frequently.

	APL (MD)	APL (CpHMD)	R_g	Model	Mechanism	Top residues	Z/AA

FFMSIRFF	71.3	58.3	37.8	B	Carpet	F1	0.125
QWCIKSKG	71.5	64.5	31.6	C	Sinking Raft	K5	0.25
FSVLFFRW	73.4	66.3	40.7	B	Sinking Raft	F1, F6, R7	0.125
YPKMFAFF	72.2	67.4	30.8	B	Sinking Raft	-	0.125
IKKFYYYY	77.9	69.1	35.9	A	Sinking Raft	-	0.25
GMFFWGKY	71.6	69.3	31.2	B	Sinking Raft	-	0.125
CLWPKQIL	76.3	69.4	29.7	C	Sinking Raft	K5, L8	0.125
KWPFYYRY	72.8	69.5	44.9	A	Sinking Raft	R7	0.25
LIKFYYMR	73.1	69.7	48.1	B	Sinking Raft	-	0.25
WCWFSLRW	74.6	69.8	45.9	B	Sinking Raft	R7	0.125
YYMVKLYI	74.2	70.1	37.1	C	Sinking Raft	V4, K5	0.125
YYMVFMRM	72.0	70.6	44.1	B	Sinking Raft	V4, R7	0.125
FWYGFGKF	76.4	70.7	32.6	B	Sinking Raft	F1	0.125
VFHKIVTL	75.1	70.9	39.9	C	Sinking Raft	L8	0.125
MYLMMFRY	75.1	71.1	37.8	B	Sinking Raft	L3, F6, R7	0.125
QWVVKILV	73.8	71.3	45.4	C	Sinking Raft	V4, K5	0.125
VMMRYYMF	73.8	71.6	38.9	B	Sinking Raft	-	0.125
LRKCWFFC	74.5	71.8	47.2	B	Sinking Raft	F6	0.25
MKKFYYYY	74.5	71.9	33.5	A	Sinking Raft	-	0.25
LHFVKTVL	76.7	72	33	C	Sinking Raft	V4, K5, L8	0.125
YGIVRFIL	77.0	72.1	42.3	C	Sinking Raft	G2, V4, F6, L8	0.125
MMMFYYRV	74.5	72.1	42.5	B	Sinking Raft	R7	0.125
PYKCFPFG	73.3	72.2	32.9	B	Sinking Raft	-	0.125
YCAFFLRY	72.8	72.2	44.2	B	Sinking Raft	R7	0.125
HGLFWWRF	71.8	72.5	46.4	B	Sinking Raft	G2, L3, R7	0.125
QFLMKCLL	77.4	72.7	30.5	C	Sinking Raft	L3, K5, L8	0.125
PKFYWRYY	75.1	72.8	40.4	A	Sinking Raft	-	0.25
CPWFYMKY	72.9	72.8	37.5	B	Sinking Raft	-	0.125
GIALKIVW	73.5	72.9	34.7	C	Sinking Raft	K5	0.125
QIWMVIKV	74.2	73.1	49.3	C	Sinking Raft	-	0.125
FGFGYPKW	75.0	73.2	32.8	B	Sinking Raft	F1, G2	0.125
VGGFYFKF	74.5	73.3	40.8	B	Sinking Raft	G2, F6	0.125
KPWFYYYR	74.3	73.3	48	A	Sinking Raft	-	0.25
SAFWWFRF	74.0	73.4	47.4	B	Sinking Raft	F6, R7	0.125
FWAFYAKG	73.8	73.6	44.8	B	Sinking Raft	F1	0.125
LHVVKFTL	74.5	73.8	33.9	C	Sinking Raft	V4, K5, F6, L8	0.125
MKKALFGT	73.8	73.8	34	B	Sinking Raft	F6	0.25
TVLYAKFW	72.2	74	30.9	C	Sinking Raft	L3	0.125
FIFMYPRY	75.8	74.1	42.2	B	Sinking Raft	F1, R7	0.125
FFPYGWKG	77.8	74.3	41.5	B	Sinking Raft	F1	0.125
LFVHKVLT	79.7	74.6	42.5	C	Sinking Raft	K5	0.125

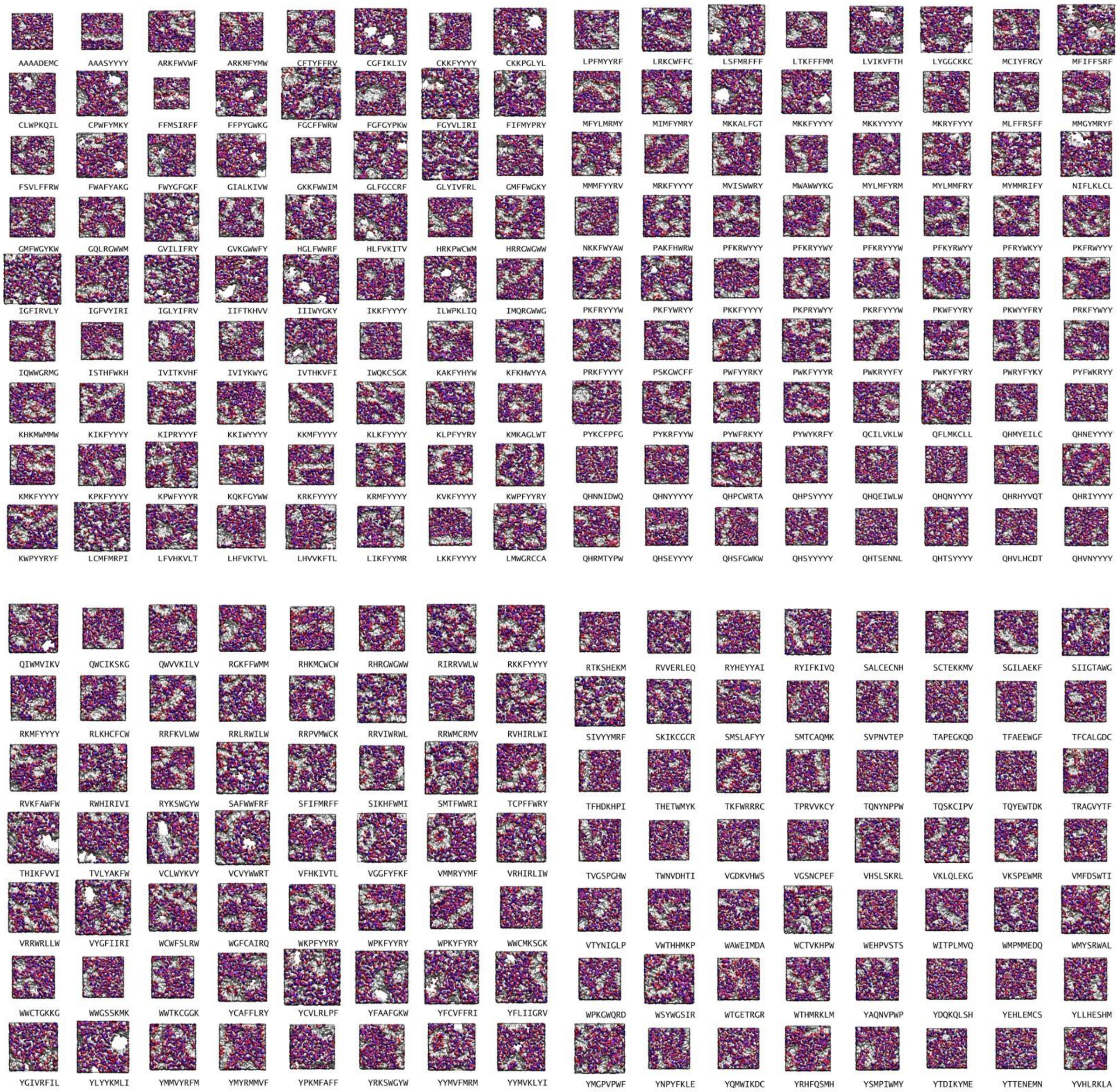
NIFLKLCL	75.9	74.7	33.2	C	Sinking Raft	K5, L8	0.125
VCLWYKVY	74.5	74.7	34.4	C	Sinking Raft	L3	0.125
YLYYKMLI	76.1	74.8	40.5	C	Sinking Raft	K5	0.125
IVTHKVFI	75.2	74.9	31.9	C	Sinking Raft	K5	0.125
YFLIIGRV	75.2	74.9	46.3	C	Sinking Raft	L3, R7	0.125
THIKFVVI	73.1	75	31.7	C	Sinking Raft	-	0.125
ILWPKLIQ	75.9	75	39.8	C	Sinking Raft	K5	0.125
LYGGCKKC	71.7	75.4	45.5	C	Sinking Raft	-	0.25
CGFIKLIV	75.1	75.6	35.1	C	Sinking Raft	G2, K5	0.125
IIFTKHVV	75.4	75.7	40.2	C	Sinking Raft	K5	0.125
YFAAFGKW	74.5	75.8	35.6	B	Sinking Raft	-	0.125
CKKPGLYL	75.0	76.3	44.5	B	Sinking Raft	L8	0.25
HLFVKITV	75.7	76.4	34.8	C	Sinking Raft	V4, K5	0.125
LVIKVFTH	78.4	76.4	42.6	C	Sinking Raft	F6	0.125
IGFVYIRI	75.2	76.4	43.2	C	Sinking Raft	G2, V4, R7	0.125
LMWGRCCA	75.0	76.6	48.3	C	Sinking Raft	-	0.125
VCVYWWRT	73.6	76.6	48.5	B	Sinking Raft	R7	0.125
GLFGCCRF	76.5	77.1	48.4	C	Sinking Raft	R7	0.125
YFCVFFRI	74.5	77.4	45.3	B	Sinking Raft	V4, F6, R7	0.125
IGLYIFRV	72.1	77.6	46.9	C	Sinking Raft	G2, L3, F6, R7	0.125
GVILIFRY	72.1	78.1	47.9	C	Sinking Raft	F6, R7	0.125
IIIWYGKY	77.3	78.5	46.1	C	Sinking Raft	-	0.125
GLYIVFRL	76.1	78.5	51.8	C	Sinking Raft	F6, R7, L8	0.125
VYGFIIRI	75.1	78.8	47.6	C	Sinking Raft	R7	0.125
LCMFMRPI	76.0	78.9	50.6	C	Sinking Raft	-	0.125
LSFMRFFF	75.0	79.1	48.3	B	Sinking Raft	F6	0.125
YCVLRLPF	77.3	79.3	48	C	Sinking Raft	-	0.125
MFIFFSRF	75.6	79.5	46.2	B	Sinking Raft	R7	0.125
IGFIRVLY	76.5	80.4	50	C	Sinking Raft	G2	0.125
FGYVLIRI	76.0	81	48.6	C	Sinking Raft	F1, G2, V4, R7	0.125

Table S5 Peptides found to be non-pore forming. The 129 non-PFPs found with the active learning, the model combination that found them, and APL and R_g values. Of the 160 positional residues, very few from the top 20 identified occur.

	APL (CpHMD)	Rg	Model	Top residues	Z/AA		APL (CpHMD)	Rg	Model	Top residues	Z/AA
CKKFYYYY	63.2	33.7	A		0.25	PFRYWKYY	67.8	35.2	A		0.25
GKKFWWIM	63.2	36.2	B		0.25	IVIYKWYG	67.9	34.6	C	K5	0.125
QHPSYYYY	63.2	37.7	A		0	PKWYYFRY	67.9	39.7	A	F6, R7	0.25

WWGSSKMK	63.9	38	C		0.25	MKRYFYYY	67.9	39.1	A		0.25
LTKFFFMM	64	36.5	B	F6	0.125	WKPFYYRY	68.1	37.1	A	R7	0.25
AAASYYYY	64.3	37.5	A		0	RRFKVLWW	68.1	41.4	C		0.375
QHNEYYYY	64.3	37.5	A		-0.125	RKKFYYYY	68.2	34	A		0.375
QHQNYYYY	64.4	37.1	A		0	WPKFYYRY	68.3	46.2	A	R7	0.25
AAAADEMC	64.5	38.1	B		-0.25	PRKFYWYY	68.4	37.9	A		0.25
QHNNIDWQ	64.6	36.7	B		-0.125	PFKRWYYY	68.5	41.4	A		0.25
MWAWWYKG	64.7	39.1	B		0.125	RLKHCFCW	68.6	41.6	B	F6	0.25
QHVLHCDT	64.7	26.5	B		-0.125	PKFRWYYY	68.7	38.5	A		0.25
QHSYYYYY	64.8	33.7	A		0	RRWMCRMV	68.7	53.9	B		0.375
QHSFGWKW	64.8	37.7	B		0.125	GMFWGYKW	68.7	30.7	B		0.125
RYKSWGYW	64.8	47.2	A		0.25	PRKFYYYY	68.7	40.6	A		0.25
WPKYFYRY	64.8	51.4	A	R7	0.25	QHRMTYPW	68.9	33.6	B		0.125
QHSEYYYY	64.9	35.3	A		-0.125	PKPRYWYY	68.9	43.5	A		0.25
LKKFYYYY	65	37.5	A		0.25	GQLRGWWM	69	41.9	C	L3	0.125
ISTHFWKH	65	37.9	C		0.125	PFKRYYWY	69	41.1	A		0.25
IWQKCSGK	65	30.9	C		0.25	QCILVKLW	69.1	31.5	C		0.125
QHNYYYYY	65.1	42.6	A		0	PKFRYYYW	69.1	40.1	A		0.25
WWTKCGGK	65.2	35.7	C		0.25	MCIYFRGY	69.1	35.3	C		0.125
QHTSENNL	65.2	24.5	B	L8	-0.125	ARKFWVWF	69.1	35.3	A		0.25
QHQEIWLW	65.2	28	B		-0.125	PKWFYYRY	69.2	44.3	A	R7	0.25
WWCMKSGK	65.2	28.9	C	K5	0.25	IQWWGRMG	69.2	44.6	C		0.125
QHTSYYYY	65.5	29.7	A		0	PWFYYRKY	69.3	53.9	A		0.25
NKKFWYAW	65.5	35.3	A		0.25	YMYRMMVF	69.3	46.3	B		0.125
QHVNYYYY	65.5	29.1	A		0	YMMVYRFM	69.3	37.7	B	V4	0.125
RHKMCWCW	65.6	31.3	B		0.25	PYKRFYYW	69.4	38.6	A		0.25
MRKFYYYY	65.7	43.9	A		0.25	RVKFAWFW	69.5	38.4	A		0.25
QHMYEILC	65.8	24.9	B		-0.125	RGKFFWMM	69.5	46.8	B	G2	0.25
WWCTGKKG	66	35.2	C		0.25	KKIWYYYY	69.5	32.5	A		0.25
PYFWKRYY	66.1	37.9	A	K5	0.25	MFYLMRMY	69.5	37.8	B		0.125
PWKYFYRY	66.3	38.8	A	R7	0.25	IVITKVHF	69.6	45.2	C	K5	0.125
MKKYYYYY	66.3	31.2	A		0.25	QHRIYYYY	69.6	39.5	A		0.125
PKKFYYYY	66.5	35.4	A		0.25	PAKFHWRW	69.6	35.7	B	R7	0.25
YRKSWGYW	66.5	37	A		0.25	KPKFYYYY	69.6	42.3	A		0.25

KKMFYYYY	66.5	43.3	A		0.25	RRLRWILW	69.9	50.4	C	L3	0.375
QHRHYVQT	66.6	44.1	B		0.125	KLKFYYYY	69.9	41.5	A		0.25
PYWYKRFY	66.6	47.6	A	K5	0.25	IMQRGWWG	70	49.9	C		0.125
ARKMFYMW	66.6	33.9	B		0.25	MYLMFYRM	70	35.9	B	L3, R7	0.125
KIKFYYYY	66.7	44.3	A		0.25	HRRGWGWW	70.2	43.6	A		0.25
HRKPWCWM	66.7	35.7	B		0.25	MVISWWRY	70.3	42.5	B	R7	0.125
LPFMYYRF	66.7	36.3	B	R7	0.125	SFIFMRFF	70.3	40.7	B		0.125
KMKFYYYY	66.9	32.3	A		0.25	KRMFYYYY	70.4	43.8	A		0.25
KVKFYYYY	66.9	44.4	A		0.25	KFKHWYYA	70.5	43.5	A		0.25
PFKRYYYW	67	46.6	A		0.25	PWKFYYYR	70.6	47.7	A		0.25
PFKYRWYY	67	40.5	A		0.25	KLPFYYRY	70.7	42.5	A	R7	0.25
KRKFYYYY	67.1	44.3	A		0.375	RWHIRIVI	70.9	41.2	C		0.25
KQKFGYWW	67.1	45.5	A		0.25	CFTYFFRV	71	44.1	B	F6, R7	0.125
PWRYFYKY	67.1	39.9	A		0.25	KIPRYYYF	71	49.8	A		0.25
GVKGWWFY	67.2	34.9	B		0.125	MMGYMRYF	71.1	41.8	C		0.125
PWKRYYFY	67.2	44.8	A		0.25	VRHIRLIW	71.3	45.8	C		0.25
PSKGWCFF	67.3	34.4	B		0.125	RRVIWRWL	71.3	44.3	C	L8	0.375
RKMFYYYY	67.3	32.2	A		0.25	RIRRVWLW	71.3	47.1	C		0.375
PYWFRKYY	67.3	43.3	A		0.25	TCPFFWRY	71.3	39.3	B	R7	0.125
KMKAGLWT	67.5	44.4	B		0.25	KWPYYRYF	71.4	43.4	A		0.25
RRPVMWCK	67.5	44.9	B	V4	0.375	VRRWRLLW	72.2	45.8	C		0.375
MLFFRSFF	67.6	43.5	B		0.125	MYMMRIFY	72.2	41	B		0.125
SIKHFWMI	67.6	42.1	B		0.125	WGFCAIRQ	72.5	50.6	C	G2, R7	0.125
PKRFYYYW	67.7	33.5	A		0.25	RVHIRLWI	72.6	47.2	C		0.25
RHRGWGWW	67.8	45.2	A		0.25	MIMFYMRY	72.8	44.5	B	R7	0.125
KHKMWMMW	67.8	37	B		0.25	QHPCWRTA	73.5	48	B		0.125
KAKFYHYW	67.8	43.3	A		0.25	SMTFWWRI	75.5	44.5	B	R7	0.125
FGCFFWRW	80	48.7	B	F1, G2, R7	0.125						



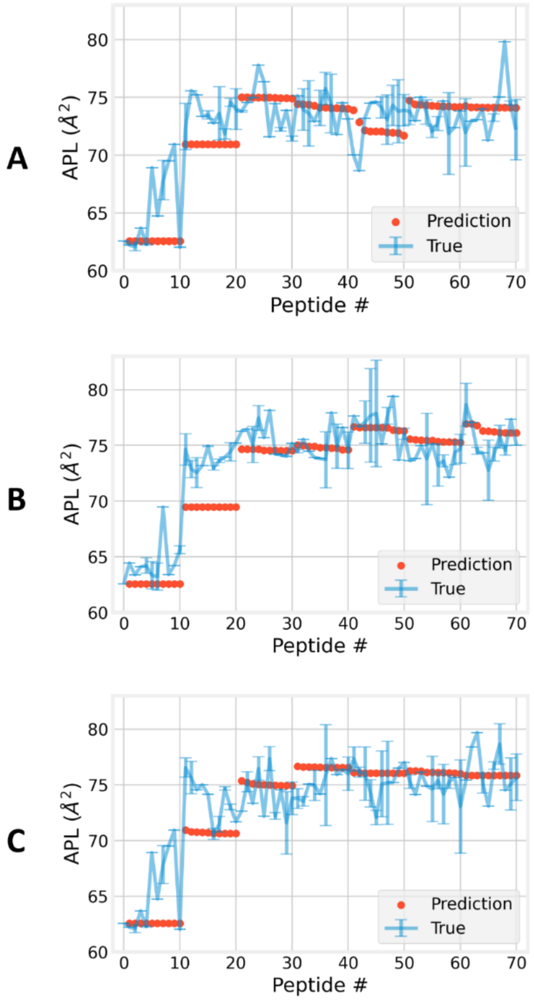
Supplemental Figure S21 Additional peptide-bilayer snapshots. Snapshots of randomly selected peptide-bilayer systems at the end of each simulation, peptides have been removed to better show their effect on the bilayers.

IMHEPDVN
ISFHLCHT
ISNCAKSD
IYLQLDIF
KDVFSYQL
KGHSSYNF
KGIVGFCE
KGSWHSAK
KKMRDEWV
KNEITWMT
KPPPFSSM
KQLNLRSD
KWQFNMYW
KYPYPRMH
LDDPDQVE
LIWSMIMD
LIYDTHQK
LIYYMNPT
LLHADDFH
LNGGWQDI
LQKHHKLT
LTFWAGME
LWGWRTNR
LWVMQMSP
LWWRDECS
LYLKSWAP
MFEPGLVN
MHHVFYTE
MHTMECIW
MKIDGSGR
MLFMRRIR
MLINATKI
MNPDCEYR
MNWRIQMN
MPMTMNCW
MSYDSGGT
NDVCLMII
NDYEQPSV
NEMQRNSW
NHYLCVHM
NLPGWHAA
NLSCGGLF
NNHMLDCH
NNVQHTYV
NSGTHFVC
PADYPDWD
PAFTNTYQ
PARHLVIF
PFMLGACT
PGSFKLLK
PLSWHLCH
QAPAHRCC
QCKLDKTI
QEARISTQ
QEFMLRYM
QGIQMWSC
QLHAHCCF
QLKCRTGY
QLNANIHK
QNHGYDYN
QPLLMTYG
QWAQGLML
QYWQMIST
RFADSHDV
RGSYAPQC
RIANMQCA
RLKPVTPY
RNVGTNVD
RPDDYYPA
RPGPWAYH
RQKNFFLV
RQSCRPEY
AFSGMLER
AGGIMNDV
AKQHSRLD
AKWGQTNH
AMHYHWHQ
AMKWVYWG
ARYISAFK
ASQIRTDT
AWPPAKEG
AYQLAADK
AYYIPCCG
CEDWSNVP
CETTREKH
CHPIFGTK
CIIWKWFT
CLCQIMFT
CRLEFGHC
CSFKFLLP
CWSGAADD
DCHPWAPY
DCMLDYFY
DDYAACKH
DFCRAWAL
DMHIPQID
DMRWRCLG
DNSATDLT
DPGETFPD
DPWDKGVL
DSKYDPRY
DSTDPRDG
EEAQMYLA
EINFLGHT
ELYFQYVS
ENCCGQWF
EPGTERIQ
EQTIIAGE
ESGQYLWK
ETFKMPAF
EVWNVDNM
EYFYMSYD
EYWLQRKK
FESDRNFA
FEYSYCHL
FIQRWLDV
FNYLDHAP
FSGEGGCM
FWGTCDGT
FWQTLCDQ
GARHWFVD
GEHNTVCM
GGLFVSCQ
GKLENIYH
GLESCLYI
GLFLAAGH
GMEHKLSY
GMKFHTAC
GMNCDGLA
GPGREIEL
GPYWRWCP
HDFNFAAT
HHCPWADP
HKPQAVAI
HLMMAVKV
HMAINQTL
HPYRGEQV
HTFQHEAM
HVHAWSIC
ICRYECEP
IEPNGFTP
IFEWRQIC
IHWPFTLH
IIFNTADQ

Supplemental Figure S22 Additional peptide-bilayer snapshots. Snapshots of randomly selected peptide-bilayer systems at the end of each simulation, peptides have been removed to better show their effect on the bilayers.

3.3.1 Machine learning model improvement over time

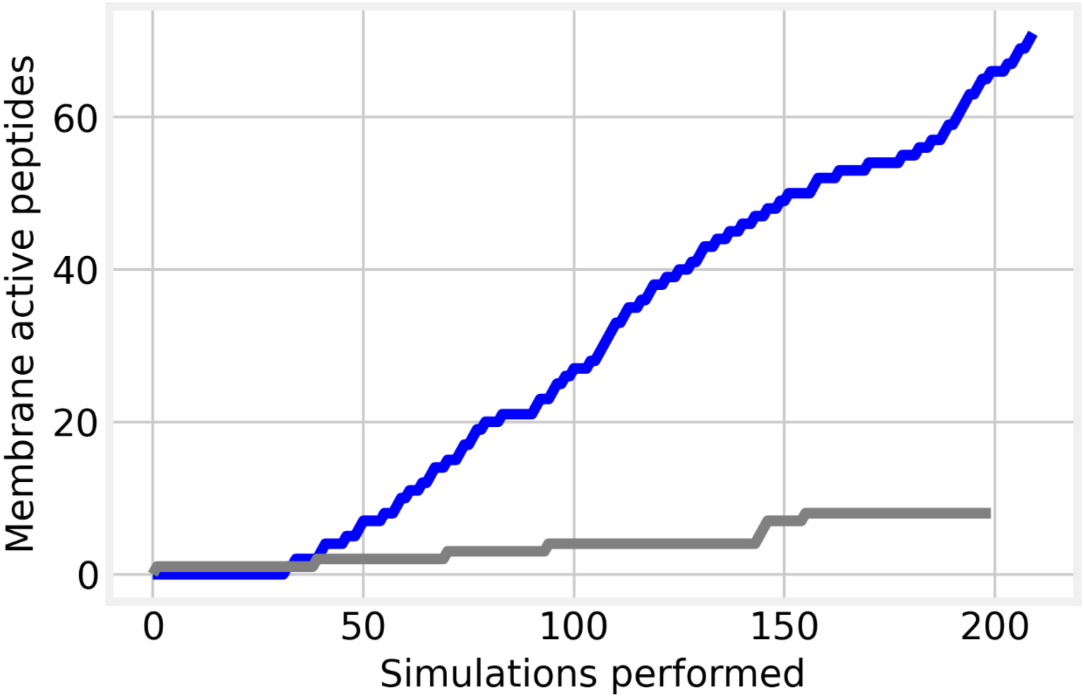
All three active learning paradigms (Table S2) showed rapid improvement in ability to select for high APL scoring peptides, including the ability to correct for a relatively bad set of predictions (Figure S23).



Supplemental Figure S23 Active learning iterations. Iterative cycle of predicting and testing the top performing peptides with the results of the predicted (red) and measured APL values from duplicate runs (blue) for our three model combinations of models A, B and C.

3.4 Machine learning acceleration

To determine the relative acceleration of using the active learning algorithm described herein we compare the random set of 200 peptides with the active learning algorithms' peptides. We find that the active learning algorithms initially do not outperform random screening during the early phases where the models are learning the chemical space. However, as the active learning models improve the rate of discovery (active peptides found vs simulations performed) increases (Figure S24).



Supplemental Figure S24 Active learning vs random selection. The rate of discovery of membrane active peptides is low for random screening (grey) as well the initial stages of active learning where the peptide selections are no better than random. However, as the active learning models (blue) improve the rate of discovery of membrane active peptides increases dramatically.

4. Data underpinning this publication

All data underpinning this publication are openly available from the University of Strathclyde KnowledgeBase at doi.org/10.15129/e9b5fb03-b07a-46d2-9336-14cfaa1fea31

5. Supplemental References

1. K. R. Mahendran, Springer US, 2021, vol. 2186, pp. 19-32.
2. S. J. Marrink, H. J. Risselada, S. Yefimov and D. P. a. Tieleman, *J. Phys. Chem. B*, 2007, **111**, 7812--7824.
3. L. Monticelli, S. K. Kandasamy, X. Periole, R. G. Larson, D. P. Tieleman and S. J. Marrink, *J. Chem. Theory Comput.*, 2008, **4**, 819--834.
4. T. A. Wassenaar, H. I. Inglfsson, R. A. Bckmann, D. P. Tieleman and S. J. Marrink, *J. Chem. Theory Comput.*, 2015, **11**, 2144--2155.
5. A. J. Rzepiela, D. Sengupta, N. Goga and S. J. Marrink, *Faraday Discuss.*, 2010, **144**, 431--443.
6. G. Bussi, D. Donadio and M. Parrinello, *J. Chem. Phys.*, 2007, **126**, 014101.
7. H. J. C. Berendsen, J. P. M. Postma, W. F. van Gunsteren, A. DiNola and J. R. Haak, *J. Chem. Phys.*, 1984, **81**, 3684-3690.
8. B. Hess, *J. Chem. Theory Comput.*, 2008, **4**, 116-122.
9. M. Sonora, E. E. Barrera and S. Pantano, *Biochim. Biophys. Acta, Biomembr.*, 2022, **1864**, 183804.
10. K. P. Santo, S. J. Irudayam and M. L. Berkowitz, *J. Phys. Chem. B*, 2013, **117**, 5031-5042.
11. S. J. Irudayam and M. L. Berkowitz, *Biochim. Biophys. Acta*, 2012, **1818**, 2975-2981.
12. L. Thogersen, B. Schiott, T. Vosegaard, N. C. Nielsen and E. Tajkhorshid, *Biophys. J.*, 2008, **95**, 4337-4347.
13. E. Han and H. Lee, *RSC Advances*, 2015, **5**, 2047-2055.
14. N. Goga and A. J. a. Rzepiela, *J. Chem. Theory Comput.*, 2012, **8**, 3637--3649.
15. J. A. Lzaguirre, D. P. Catarello, J. M. Wozniak and R. D. Skeel, *J. Chem. Phys.*, 2001, **114**, 2090--2098.
16. S. E. Feller, Y. Zhang, R. W. Pastor and B. R. Brooks, *J. Chem. Phys.*, 1995, **103**, 4613--4621.
17. H. W. Huang, *Biochim. Biophys. Acta*, 2006, **1758**, 1292-1302.
18. P. W. Frederix, G. G. Scott, Y. M. Abul-Haija, D. Kalafatovic, C. G. Pappas, N. Javid, N. T. Hunt, R. V. Ulijn and T. Tuttle, *Nat. Chem.*, 2015, **7**, 30-37.
19. M. J. Davila and C. Mayer, *Life (Basel)*, 2022, **12**.
20. H. Moriwaki, Y. S. Tian, N. Kawashita and T. Takagi, *J. Cheminformatics*, 2018, **10**, 4.
21. J. Dong, Z. J. Yao, L. Zhang, F. Luo, Q. Lin, A. P. Lu, A. F. Chen and D. S. Cao, *J. Cheminformatics*, 2018, **10**, 16.
22. R. Avadhoot, PyMolSAR, https://github.com/BeckResearchLab/PyMolSAR ,urldate = 2021-06-01).
23. T. Kessler, PaDELPy: A Python wrapper for PaDEL-Descriptor software, https://github.com/ECRL/PaDELPy ,urldate = 2021-06-01).
24. C. W. Yap, *J. Comput. Chem.*, 2011, **32**, 1466-1474.
25. A. Van Teijlingen and T. Tuttle, *J. Chem. Theory Comput.*, 2021, **17**, 3221-3232.
26. Okuta, Ryosuke, Unno, Yuya, Nishino, Daisuke, Hido, Shohei, Loomis and Crissman, presented in part at the Proc. Work. Mach. Learn. Syst. Thirty-first Annu. Conf. Neural Inf. Process. Syst., 2017.
27. D. Vohra, in *Practical Hadoop Ecosystem*, Apress, Berkeley, CA, 2016, ch. 8, pp. 325-335.