Secondary Structure Motifs Made Searchable to Facilitate the Functional Peptide Design

Cheng-Yu Tsai[1,2,¶], Emmanuel O Salawu[1,3,¶], Hongchun Li[1,4,¶], Guan-Yu Lin[5,¶], Ting-Yu Kuo[5,¶], Liyin Voon[1,¶], Adarsh Sharma[1], Kai-Di Hu[1], Yi-Yun Cheng[6], Sobha Sahoo[1], Lutimba Stuart[1], Chih-Wei Chen[1], Yuan-Yu Chang[1], Yu-Lin Lu[1], Ximai Ke[1], Chen-Chi Wu[2], Chung-Yu Lan[5,7]*, Hua-Wen Fu[5,7]*, Lee-Wei Yang[1,3,7,8,]*

[1]Institute of Bioinformatics and Structural Biology, National Tsing Hua University, Hsinchu 30013, Taiwan; [2]Graduate Institute of Medical Genomics and Proteomics, National Taiwan University College of Medicine, Taipei 10055, Taiwan; [3]Bioinformatics Program, Institute of Information Sciences, Academia Sinica, Taipei 11529, Taiwan; [4]Department of Computational and Systems Biology, School of Medicine, University of Pittsburgh, Pittsburgh, PA, 15213, USA; [5]Institute of Molecular and Cellular Biology, National Tsing Hua University, Hsinchu 30013, Taiwan; [6]Praexisio Taiwan Inc. New Taipei 22180, Taiwan; [7]Department of Life Science, National Tsing Hua University, Hsinchu 30013, Taiwan; [8]Physics Division, National Center for Theoretical Sciences, Hsinchu 30013, Taiwan.

¶ Authors who have equal contributions

* To whom correspondence should be addressed: Lee-Wei Yang. Tel: +88635742467; Fax +88635715934; Email: lwyang@life.nthu.edu.tw; Hua-Wen Fu. Tel: +88635742485. Fax: +88635715934 Email: hwfu@life.nthu.edu.tw, and Chung-Yu Lan. Tel: +88635742473; Email: cylan@life.nthu.edu.tw

ABSTRACT

To ensure a physicochemically desired sequence motif to adapt a specific type of secondary structures, we compile an α-helix database allowing complicate search patterns to facilitate a data-driven design of therapeutic peptides. Nearly 1.7 million helical peptides in >130 thousand proteins are extracted along with their interacting partners from the protein data bank (PDB). The sequences of the peptides are indexed with patterns and gaps and deposited in our Therapeutic Peptide Design dataBase (TP-DB). We here demonstrate its utility in three medicinal design cases. By our pattern-based search engine but not PHI-BLAST, we can identify a pathogenic protein, *Helicobacter pylori* neutrophil-activating protein (HP-NAP), a virulence factor of *H. pylori*, which contains a motif DYKYLE that belongs to the affinity determinant motif DYKXX[DE] of the FLAG-tag and can be recognized by the anti-FLAG M2 antibody. By doing so, the known purification-tag-specific antibody is repurposed into a diagnostic kit for *H. pylori*. Also by leveraging TP-DB, we discovered a stretch of helical peptide matching the potent membrane-insertion pattern

WXXWXXW, elucidated by MD simulations. The newly synthesized peptide has a better minimal inhibitory concentration (MIC) and much lower cytotoxicity against *Candida albicans* (fungus) than that of previously characterized homologous antimicrobial peptides. In a similar vein, taking the discontinued anchoring residues in the helix-helix interaction interface as the search pattern, TP-DB returns several helical peptides as potential tumor suppressors of hepatocellular carcinoma (HCC) whose helicity and binding affinity were examined by MD simulations. Taken together, we believe that TP-DB and its pattern-based search engine provide a new opportunity for a (secondary-)structure-based design of peptide drugs and diagnostic kits for pathogens without inferring evolutionary homology between sequences sharing the same pattern. TP-DB is made available at http://dyn.life.nthu.edu.tw/design/.

1. Introduction

Secondary structures comprise a subset of amino acid sequences where each constituent residue can form hydrogen bonds (H-bonds) in specific orders with other constituent residues, at their peptidyl backbones, which grants a special mechanical property in our life-thriving nanomachines – proteins. Particularly, α-helices, where the *i*-th residue in the helix H-bonded with *i*+4-th residue, mediate various types of molecular functions including cell penetration[1, 2, 3, 4], subcellular localization[5], intra/inter-molecular allostery[6-9], special helix-DNA scaffolds[10] and protein-protein interaction[7, 11], provided that the primary sequence meets both the structural (main chain and side chain) and chemical (side chain) requirements for specific functions of interest. Especially at the interface, either water-lipid interface at the cell membrane, or protein-DNA/protein-protein interface, these helices have disparate physicochemical properties at their two sides to properly play their functional roles. Take amphiphilic antimicrobial peptides (AMPs) for example, if the *i*-th position is a membrane-insertion-promoting amino acid (say, W or Y), the *i*+3 or *i*+4-th position needs to be an amino acid also prompt for membrane insertion (say, another hydrophobic residues). In other words, these residues with similar physicochemical properties need to be aligned at the same side, while the other side faces bulk water and ions accumulated near the membrane. In this example, what we need is a stretch of amino acid sequence consisted of a pattern of (W/Y)X_1X_2(W/Y) or (W/Y)$X_1X_2X_3$(W/Y), where X_i can be any amino acid as long as the combination of X_i can still hold an helical structure, and (W/Y) means that position can be either

W or Y. From the point of view of protein design, either to design a better AMP or to propose stronger interface binders, and therefore blockers, for therapeutic purposes, it is not an easy task to allow constituent residues to address physicochemical needs, meeting a specific sequence pattern, while maintaining the required structural integrity. Suitable tools need to be developed to facilitate the process.

Currently, both (A) pattern-based (only) search engines locating desired functional motifs in primary sequences *regardless of their evolutionary conservation* and (B) dedicated secondary structural databases indexed for (A)'s purpose are lacking. Protein Secondary Structure database (PSS) proposed 30 years ago no longer serves the community[12], while DSSP annotates the secondary structures when provided with tertiary structures of proteins[13]. However, both DSSP and protein databank (PDB) do not provide a search engine for specific types of secondary structures (say, the entire query sequence should belong to an alpha-helix) for >170,000 structurally solved sequences, nor do they provide any search function that recognizes specific sequence patterns (see the beginning of Methods for our definition of "patterns"). To better integrate functional structural fragments into therapeutics[14, 15, 16], in this seminal work, we extracted ~1.7 million (1,676,117) structurally resolved helices and their contacting/interacting helical partners from the PDB, from which helical propensity and coordination (contact) number are derived (Table 1). We then establish a pattern-specific search engine (see below) to find indexed sequences bearing particular spatio-chemical properties instead of evolutionary homology (say, the purpose of Pattern Hit Initiated BLAST (PHI-BLAST)[17] that takes input of both residue pattern and the template sequence). We use this Therapeutic Peptide Design dataBase (TP-DB) to develop a new antifungal/antimicrobial peptide, helix-helix interface blockers that suppress hepatocellular carcinoma, and a new diagnostic reporter to detect *Helicobacter pylori* infection.

2. Methods

2.1. Extraction of Helical Peptide Sequences from the Protein Data Bank (PDB)

To obtain the amino acid sequences that fold into helices in nature, we processed the PDB[18] files of 130,000+ experimentally determined protein structures as of June, 2017 and extracted the locations of secondary structures from the header part of the PDB files. There are 96,711 unique

PDB IDs/files containing at least one alpha-helix documented at the header area to result in 1,676,087 structurally resolved helices having a length longer than 5 residues. In addition, for a given helix, its contacting neighbors in 3D space and the adjacent helical peptides that are within 4Å (per heavy atoms) were also collected and analyzed.

2.2. Indexing and Querying of TP-DB

Helical peptides that match specific patterns are carefully indexed in the database. A pattern could be of the form "Y***G**K", where "*" belongs to any of the 20 amino acids, equivalent to the regular expression "Y-x(3)-G-x(2)-K" in a PROSITE format that has also been used in the PHI-BLAST search (https://www.ncbi.nlm.nih.gov/blast/html/PHIsyntax.html). The location of a pattern is indexed by the PDB ID, the chain ID, and the position where the pattern begins in that chain (see below). Considering both flexibility and search comprehension, we let the simplest pattern be [A_1 m A_2 n A_3] where A_1, A_2, and A_3 are the single-letter codes of certain amino acids, in between which are the spacings represented by the whole numbers m and n. We currently allow m and n to be from zero to four. A key "Y3G2K" or its search pattern [Y 3 G 2 K] would indicate a "Y" is (3 + 1) residues upstream to a "G" that is (2 + 1) upstream to a "K". A key is paired with identifiers (values) referring to where this key can be found in all the stored peptides (**Fig. 1C, E**). Specifically, these values, stored in an array, are peptide identifiers and all the starting positions of a certain key in each of the peptides (**Fig. 1C**). These key-value pairs comprise our database (**Fig. 1**), TP-DB (http://dyn.life.nthu.edu.tw/design/), which could be searched by a simple pattern [A_1 m A_2 n A_3] or a pattern compounded by several of these simple patterns (**Fig. 1F**).

To query the database, the user specifies at least three anchor residues and the gaps that separate them, such that "ADE" or "K****VA" can be queried by [A 0 D 0 E] and [K 4 V 0 A], respectively (**Fig. 1D**). For more complicated queries, say "A/Y 3 G 2 K 3 H 4 K", it is first broken down into a combination of two subqueries "A 3 G 2 K 3 H 4 K" and "Y 3 G 2 K 3 H 4 K" (**Fig. 1F**). The five anchor residues in each subquery is treated as two concatenated three-anchor keys. The matched patterns co-localized in a peptide sequence are further parsed to take into account the overlapped regions before returning the results. TP-DB currently have two search engines to search the indexed peptides, one by pattern search, described above, and the other by BLAST (parameterized for short sequences). The use of the latter is beyond the scope of this paper.

A

```
Peptide p:    ADEKKFWGKYLYEVA        Peptide z:    TAFEGGILKKGHHCSYTKH
Positions:    012345678901234        Positions:    0123456789012345678
```

B

```
Keys in Peptide p                    |   Start Positions in Peptide p
A0D0E, D0E0K, ..., Y0E0V, E0V0A      |   0, 1, ..., 11, 12
A0D1K, D0E1K, ..., L0Y1V, Y0E1A      |   0, 1, ..., 10, 11
...                                  |   ...
A0D4W, D0E4G, ..., G0K4V, K0Y4A      |   0, 1, ..., 7, 8
...                                  |   ...
A4F0W, D4W0G, ..., G4E0V, K4V0A      |   0, 1, ..., 7, 8
...                                  |   ...
A4F4L, D4W4Y, ..., K4K4V, K4Y4A      |   0, 1, ..., 3, 4
```

C

```
DB_INDEX = { A0D0E: [ [Peptide p , 0], [Peptide p+1 , 17, 21, ...], ...],
             D0E0K: [ [Peptide p , 1], [Peptide p+9 , 13, 25, ...], ...],
             ...
             K4V0A: [ [Peptide p , 8], [Peptide p+78 , 1, 64, ...], ...],
             ...
             K4Y4A: [ [Peptide p , 4], [Peptide p+2 , 23, 41, ...], ...],
             ... }
```

D

Sample patterns of interest	Keys in the DB	Examples of peptides that the patterns can match
ADE	A0D0E	ADEKKFWGKYLYEVA
DEK	D0E0K	ADEKKFWGKYLYEVA
K----VA	K4V0A	ADEKKFWGKYLYEVA
K----Y----A	K4Y4A	ADEKKFWGKYLYEVA

E

```
Simple Query: K 4 Y 4 A, which is equivalent to K4Y4A or K----Y----A
Direct Fetch: K4Y4A from DB_INDEX in the RAM: →
                 [ [Peptide p , 4], [Peptide p+2 , 23, 41, ...], ...],
```

F

[1] Complex Query: Process into a Combination of Simple Queries

Example of Complex Queries	Matching Patterns	Constituent Simple Queries
A 3 G 2 K 3 H 4 K	A 3 G 2 K 3 H 4 K	A 3 G 2 K K 3 H 4 K
A/Y 3 G 2 K 3 H 4 K	A 3 G 2 K 3 H 4 K	A 3 G 2 K K 3 H 4 K
	Y 3 G 2 K 3 H 4 K	Y 3 G 2 K K 3 H 4 K
A 2,3 G 2 K 3,4 H 4 K	A 2 G 2 K 3 H 4 K	A 2 G 2 K K 3 H 4 K
	A 2 G 2 K 4 H 4 K	A 2 G 2 K K 4 H 4 K
	A 3 G 2 K 3 H 4 K	A 3 G 2 K K 3 H 4 K
	A 3 G 2 K 4 H 4 K	A 3 G 2 K K 4 H 4 K

[2] Simple Queries' results are aggregated into the required results

Fig. 1. Indexing the peptides and querying of TP-DB. (**A**) The positions of amino acids in the *p*-th helical peptide (peptide$_p$) are indexed from 0 to 14. (**B**) All the possible patterns that can be found in peptide$_p$,

stored as keys and values. (**C**) In TP-DB (or DB_{Index}), keys are paired values, and values are where these keys can be found, including in which peptides and the starting position of the keys. For instance, [A 0 D 0 E] can be found in the "0-th position" in the *p*-th peptide, while it can also be found in the 17th and 21st positions in the (*p*+1)-th peptide. (**D**) The patterns of interest are translated into "keys" comprising amino acids' single-letter codes separated by numbers, the sizes of the gaps. (**E**) Results for a simple query retrieved from TP-DB. (**F**) A compounded query can be broken down into subqueries and the result is a joint result from subqueries. For efficient search, two types of libraries are loaded in timely order. What is constantly loaded in the memory, with a small memory footprint, is a collection of key-value pairs that only index which proteins have certain keys but not their locations. Only when certain proteins are visited because they contained the queried keys, then libraries of a second type, inferring the locations of keys in certain proteins, are then loaded into the memory to report the found sequences, before the protein-relevant libraries are unloaded from the memory again soon after a search job is finished.

2.3. Physicochemical Properties and Ranking of the Pattern-Matched Helical Peptides in TP-DB

The helical peptides returned by a query can be ranked by (the logarithm of) their normalized helical propensity (default) or by their contact scores. Using the data collected in the TP-DB, each of the 20 amino acids has a normalized helical propensity HP_{NA} that is $HP_{a.a.}/NA_{a.a.}$ where $HP_{a.a.}$ is the helical propensity of a given amino acid type and $NA_{a.a}$ is natural abundance of that amino acid (**Table 1**). Both $HP_{a.a.}$ and $NA_{a.a.}$ are probabilities, where $HP_{a.a}$ is defined as the count of a certain amino acid type in all the extracted helices divided by the total number of amino acids in these helices (24,301,682 amino acids in 1,676,117 helical peptides), and $NA_{a.a.}$ is the count of that amino acid in the 130,000+ proteins in PDB, divided by the total number of residues in these proteins. A log-odds ratio, log (HP_{NA}) or log ($HP_{a.a.}/NA_{a.a.}$) can be found for each amino acid type, and the reported helical propensity for each peptide in TP-DB is the sum of log (HP_{NA}) for every constituent residue in this peptide (**Table 2**). The contact number is the amino acid contact within a 15Å range of a residue of interest, in the residue's original protein environment, where only the Cα atom of an amino acid is considered. The reported contact for a peptide in TP-DB is the per-residue average of its constituent residues'. An exemplified output table from the web server is shown in **Table 2**.

2.4. General MD Simulations Protocol

All MD simulations were performed using NAMD package 2.9[19] and OpenMM[20] in NPT ensemble with periodic boundary conditions. Latest CHARMM forcefields[21, 22] were employed for the peptide, water and lipids, respectively. All simulations were performed at a time step of 2 fs and trajectories were recorded every 5000 steps (i.e. at every 10 ps). RATTLE and SETTLE algorithms are applied to constrain hydrogen atoms in peptides and waters. Cutoff of 12 Å with switch distance 10 Å and pair list distance 14 Å are applied when calculating non-bonded interactions. With periodic boundary conditions, the Particle Mesh Ewald method[23] was employed for calculations of long-range electrostatic interactions. Temperature was maintained at 310 K using Langevin dynamics[24] and pressure was controlled at 1 atm using Nosé-Hoover Langevin piston[25].

2.5. Lipid Preparation and Equilibrium *in silico*

Zwitterionic membranes with 78 1-palmitoyl-2-oleoyl-sn-glycero-3-phosphocholine (POPC; also termed as "PC" hereafter) lipids to mimic eukaryotic membrane and with 30 POPC lipid + 10 2-oleoyl-1-pamlitoyl-sn-glyecro-3-glycerol (POPG; also termed as "PG") lipid molecules (see **Fig S1**) in each of the upper and lower lipid bilayers to mimic pathogenic prokaryotic membrane[26, 27] were prepared by CHARMM-GUI server[28]. The AMPs and lipid bilayers were solvated in explicit TIP3 water molecules[29]. Na^+ and Cl^- ions were added to neutralize the system. The initial area per lipid of PC was equilibrated to stay close to its experimental values measured at 303 K, which is 68.3 ± 1.5 $Å^2$ [30].

2.6. Natural insertion simulations with soft-boundary-confined diffusion for AMPs in the presence or absence of lipids

NMR-solved α-helical and fully extended (random-coil) AMPs were simulated together with PC and PC/PG (3:1) membranes in a water box containing 0.02 M NaCl. For the fully extended AMPs, they were first simulated in an NPT ensemble for 50 ns in the absence of lipids. The resulting snapshots of the AMPs were clustered into 5 groups based on their structural similarity using the "clustering" plug-in of the VMD software[31]. A representative conformation from the biggest cluster was selected as the initial conformation for an AMP before it was loaded with the

equilibrated membrane for the natural insertion simulations. The expression "natural" meant that no external pulling forces were exerted on the AMPs, contrasting to an accelerated approach we developed earlier[2] to first pull the AMPs into the lipids and release them in order to timely assess the insertion quality of AMPs.

2.7. Determination of the Minimum Inhibitory Concentration (MIC)

To measure antimicrobial activity of peptides, minimum inhibitory concentrations (MIC) were determined by broth-dilution assay as described in Clinical and Laboratory Standards Institute (CLSI) documents M27-A3 for yeasts[32] and M07-A9 for bacteria[33] with minor modifications. Briefly, *E. coli* ATCC 25922 and *C. albicans* SC5314 were recovered from frozen stock by growing overnight at 37°C on LB (Luria-Bertani) and YPD (Yeast Extract-Peptone-Dextrose) agar plates, separately. A single colony of *E. coli* was inoculated into 5 ml LB medium and grown overnight at 37°C with shaking at 220 rpm. Cells were then diluted into MHB (Mueller–Hinton broth) at 1:100, and grew until the optical density at 600 nm (OD_{600}) reached 0.4 to 0.8. For *C. albicans* SC5314, a single colony was inoculated in YPD broth and grown overnight at 30°C with shaking at 180 rpm. Cells from the overnight culture were subcultured into fresh YPD and grew to mid-log phase. Finally, *E. coli* and *C. albicans* cells obtained from above-described were diluted into MHB and LYM, a modified RPMI 1640 medium[34] to a concentration of ~5×10^5 cells/ml and ~4×10^4 cells/ml, respectively. To determine the MIC values, one hundred microliters of cells was placed in each well of a clear bottom 96-well microplate, followed by independently adding 100 µl peptides with different concentrations (in MHB or LYM). The final concentrations in the wells were 120, 60, 30, 15, 7.5, 3.75 and 1.88 µg/ml for peptides, ~ 2.5×10^5 cells/ml for *E. coli* and ~2×10^4 cells/ml for *C. albicans*. The plate was incubated at 37°C with shaking at 220 rpm for 18 h (*E. coli*) and at 180 rpm for 24 h (*C. albicans*). The OD_{600} values were measured using a iMark™ Microplate Absorbance Reader (Bio-Rad). The MIC_{90} value was determined by 90% reduction in growth compared with that of the peptide-free growth control. Experiments were performed in pentaplicate, and MIC_{90} was determined as the majority value out of 5 repeats.

2.8. Assays for Minimum Hemolytic Concentration (MHC)

The hemolytic toxicity was determined by the hemolysis against human red blood cells (RBCs) as previous publication[35]. First, RBCs from healthy donors were mixed with EDTA in the collecting tubes to inhibit the coagulation. Then the RBCs were washed for three repeats as follows - the centrifugation at RCF 800g for 10 minutes, the removement of the supernatant, and the slight resuspension with PBS buffer. Such washed RBCs stock was finally diluted with a PBS buffer to be the 10% (v/v) RBCs solution for assays. Then total eight samples of various final concentrations (240, 120, 60, 30, 15, 7.5, 3.75 and 1.88 μg/ml) for each AMP in 96-well microplate were prepared by mixing 10% RBCs and diluted AMP solutions, respectively, at ratio 1:1 (v/v) into final volume 200 μL. All samples were placed under the condition of 37°C for 1 h and then centrifuged at RCF 800g for 10 minutes. Finally, all these supernatants from samples and controls were picked out for the measurement of optical density at the 450nm wavelength. The data processing was performed to calculate the hemolytic percentage defined as $[A_{sample} - A_{control-}]/[A_{control+} - A_{control-}]$, where $A_{control+}$ and $A_{control-}$ are the measurement at wavelength 450nm for the positive and negative controls, that are 100 μL 10% RBCs suspended in 100 μL PBS buffer with and without 2% (v/v) Triton X-100 (a known detergent to lyse the RBCs), respectively. The final hemolytic percentage for AMPs at various concentrations were determined by the mean values of three repeated experiments, and MHC_5 values that the minimal concentration of AMPs leading to 5% hemolysis of RBCs were determined. The protocol was pre-approved by Research Ethics Committee of the National Taiwan University Hospital (approval number: 201810004RINA), and blood collections were performed by trained phlebotomists to confirm the safety of the donors.

3. Results and Discussions

3.1. Database Features and Statistics

TP currently comprises 1,676,117 helical peptides (24,301,682 amino acids in total) extracted from 130,000+ experimentally solved protein structures stored in PDB. The apparent goal of establishing search engines for protein secondary structures (starting from all types of helical configurations, e.g., α-helices and 3_{10} helices.) is to find stretches of protein sequences that match the queried pattern and adapt a helical conformation. Since protein sequences assume structures in the context of their 3-dimensional environment full of residue contacts, it is quite likely the helices are not as helical as when they are part of their parent proteins. We herein design two metrics to evaluate and prioritize the peptides, in the query results, which are highly likely to still stay helical when expressed or synthesized alone - helical propensity (HP) and contact number. A higher HP and/or lower contact number would suggest a good helicity for peptides even in their isolated forms. For the former, we use a normalized helical propensity (see Methods), HP_{NA}, defined as $HP_{a.a.}/NA_{a.a.}$ where $HP_{a.a.}$ is the helical propensity of a given amino acid type and $NA_{a.a}$ is natural abundance of that amino acid (**Table 1**). Both $HP_{a.a.}$ and $NA_{a.a.}$ are probabilities, where $HP_{a.a}$ is defined as the fraction of a certain amino acid type in TP-DB and $NA_{a.a.}$ is the fraction of the same type of amino acid (https://web.expasy.org/protscale/pscale/A.A.Swiss-Prot.html) from the based on the ExPASy tool "ProtScale"[36]. A log-odds ratio, log (HP_{NA}) or log ($HP_{a.a.}/NA_{a.a.}$) is found for each amino acid type, and the reported helical propensity for each peptide in TP-DB is the sum of log (HP_{NA}) for every constituent residue in this peptide (**Table 2**). The second metric is the contact number, defined as the average number of amino acids, represented by corresponding C_α atoms, contacting a helix of interest within 15Å from it. The lower contact number of a helix in its parent protein suggests a higher probability for it to stay in the helical form in isolation.

Table 1. Amino acids sorted by their indexes of helical propensity

AA*	$HP_{a.a.}$	$NA_{a.a.}$	HP_{NA} ($HP_{a.a.}/NA_{a.a.}$)	Helical Propensity, log (HP_{NA})
A	0.109	0.083	1.320	0.277
W	0.014	0.011	1.284	0.250
E	0.084	0.067	1.246	0.220
L	0.115	0.097	1.192	0.175
Q	0.045	0.039	1.145	0.135
Y	0.033	0.029	1.130	0.122
K	0.064	0.058	1.100	0.095
M	0.026	0.024	1.079	0.076
R	0.059	0.055	1.067	0.065
F	0.038	0.039	0.984	-0.016
I	0.058	0.059	0.978	-0.022
H	0.022	0.023	0.969	-0.031
D	0.052	0.055	0.952	-0.049
N	0.038	0.041	0.936	-0.066
T	0.049	0.053	0.918	-0.086
V	0.062	0.069	0.902	-0.103
C	0.012	0.014	0.876	-0.132
S	0.053	0.066	0.804	-0.218
G	0.046	0.071	0.650	-0.431
P	0.021	0.047	0.445	-0.810

* AA = Amino Acid; $HP_{a.a.}$ = Helical Propensity for each amino acid;
$NA_{a.a.}$ = Natural Abundance of each amino acid; HP_{NA} = $HP_{a.a.}$ normalized by $NA_{a.a}$

3.2. Anti-FLAG antibody is repurposed for pathogen diagnosis

To find specific patterns in a biological sequence while inferring homology, Pattern Hit Initiated (PHI)-BLAST has been the main (if not only) bioinformatics tool to perform such[18]. PHI-BLAST was designed to address evolutionary relevance of two proteins carrying the same or similar motifs, so it renders results considering both the pattern match and local sequence similarity with statistically significance. To showcase the prowess and disparate utility of the TP-DB, we demonstrate below how a purification tag (herein FLAG-tag) and its antibody (herein M2 antibody) can be repurposed into a diagnostic kit for detecting human pathogens. The relevant experimental methods are listed in Supplementary Materials. FLAG-tag is known to have a sequence DYKDDDDK[37] with the pattern D-Y-K-x-x-[DE] being experimentally confirmed as the main affinity determinant motif[38]. "DYK" here is the strongest determinant while the last D (or E) is of the secondary importance. We would like to search TP-DB for proteins that (i) are not included in the PHI-BLAST search results and (ii) belong to proteins in human pathogens but not in non-pathogenic bacteria.

3.2.1. Confirmed by ELISA and Western Blot, HP-1 is Recognized by anti-FLAG M2 Antibody Specifically at a stretch containing the D-Y-K-x-x-[DE] motif

When querying the pattern "D 0 Y 0 K 2 D/E" against the TP-DB, we can find totally 93 sequences with unique 19 sequences. While using PHI-BLAST to search NCBI's non-redundant protein sequences (nr) with the pattern D-Y-K-x(2)-[DE] (in PROSITE format) and FLAG sequence "DYKDDDDK", we could not find any sequence with an E-value less than 100. On the other hand, among the 19 TP-DB-identified sequences, HP-NAP, containing the sequence "**DYK**YL**E**", can be found in *Helicobacter pylori* strain 26695 (*accession no.* AE000543, ATCC) but not in regular non-pathogenic bacteria. If the anti-FLAG M2 antibody can recognize this randomly selected pathogenic protein for its containing a FLAG-determinant motif (herein D-Y-K-x-x-[DE]), we can potentially use the same concept to repurpose the original use (say, purification) of any known pair of antibody and antigen recognition motif (say, M2 antibody and D-Y-K-x-x-[DE]) into a new use, herein the potential diagnostic kit for human pathogens. To our heartfelt delight, the anti-FLAG M2 antibody was indeed found to be capable of recognizing HP-NAP as analyzed by ELISA (**Fig. 2A**) and Western blot (**Fig. 2B**). Further analysis showed that M2 antibody detected maltose-

binding protein (MBP)-tagged HP-NAP$_{R77\text{-}E116}$ but not MBP (**Fig. 2B**), indicating that M2 antibody binds to the polypeptide fragment of HP-NAP containing residues Arg77 to Glu116, which contains the antigen recognition motif, **DYK**YL**E,** as shown in **Fig. 2C**.

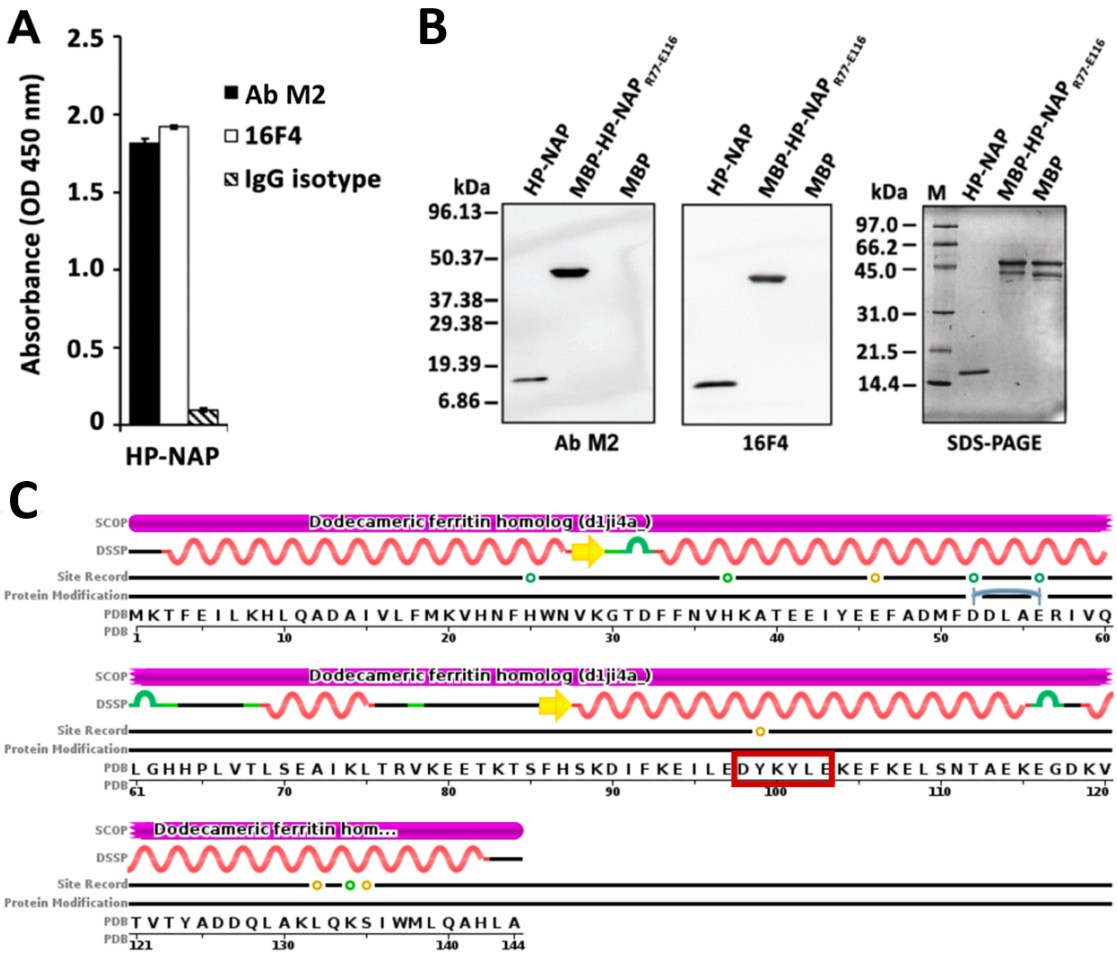


Fig. 2. Detection of HP-NAP by anti-FLAG M2 antibody. (**A**) Detection of HP-NAP by anti-FLAG M2 antibody using recombinant HP-NAP-based ELISA. Recombinant HP-NAP was purified by a small-scale DEAE Sephadex negative mode batch chromatography. Purified HP-NAP as an antigen was coated on an ELISA plate as described in Methods. The anti-FLAG M2 antibody, its corresponding mouse IgG antibody, and the antibody 16F4 against HP-NAP, as positive control, were subjected to recombinant HP-NAP-based ELISA. The result is expressed as absorbance at $OD_{450\ nm}$. Data were presented as the mean ± S.D. of one experiment in duplicate. Similar results were obtained in three independent experiments. (**B**) Detection of recombinant HP-NAP by anti-FLAG M2 antibody using Western blot analysis. Recombinant HP-NAP was purified by two consecutive gel-filtration chromatography. MBP-tagged HP-NAP$_{R77\text{-}E116}$ and MBP were partially purified by the MBPTrap HP column using ÄKTA Purifier. These purified proteins, 1 µg each, were subjected to SDS-PAGE on a 15 % gel and Western blot analysis with anti-FLAG M2 antibody and the antibody 16F4. Molecular masses (M) in kDa are indicated in the left of the blots and the gel. Similar results were obtained in two independent experiments. **(C)** The sequence of HP-NAP. Note that

"DYKYLE", highlighted in the red box, matches the pattern D-Y-K-x-x-[DE] and is included in the R77-E116 stretch. The image is download from PDB (code:1JI4) with minor modifications.

As a result, the antibody (herein M2 antibody) of a purification kit can be *repurposed into a diagnostic kit* for detecting human pathogens, herein *H. pylori*, through recognizing native HP-NAP by ELISA or HP-NAP monomer as a 17 kDa band by Western blot.

3.3. α-Helical Antimicrobial Peptide (AMP) Design

Pan-drug-resistant bacteria is no less a threat to mankind than viral infections that could cause a pandemic[39]. A discouraging forecast has suggested that by 2050, 10 million people could die from these super bugs each year, with half of them in Asia[40]. Antimicrobial peptides (AMPs) that target bacterial lipid membranes, evolving at a much slower pace than bacterial protein receptors, appear to be a promising alternative for antibiotics. There have been substantial efforts made to study AMPs' insertion mechanism[26, 27, 41, 42] and analyze the determining characteristics for their efficacy[43, 44]. To join these collective efforts and showcase the good use of TP-DB, we first examined in details the insertion processes of two previously reported helical antimicrobial peptides (AMPs), W3_p1 (sequence: KK **W**RK **W**LK **W**LA KK[45]) and W3_p2 (sequence: KK **W**LK **W**LK **W**LK KK[2]) by MD simulations. The two amphiphilic AMPs have 3 equally spaced tryptophans at their hydrophobic face, while positively charged lysines/arginines are on the other side, which is a typical feature and requirement for helical AMPs. The goal of the simulations is to understand each residue's physicochemical role in the bacteria-membrane insertion process while earlier reports[2, 45, 46, 47, 48] have suggested the deeper a peptide penetrates the lipid membrane, the higher bactericidal potency it can have.

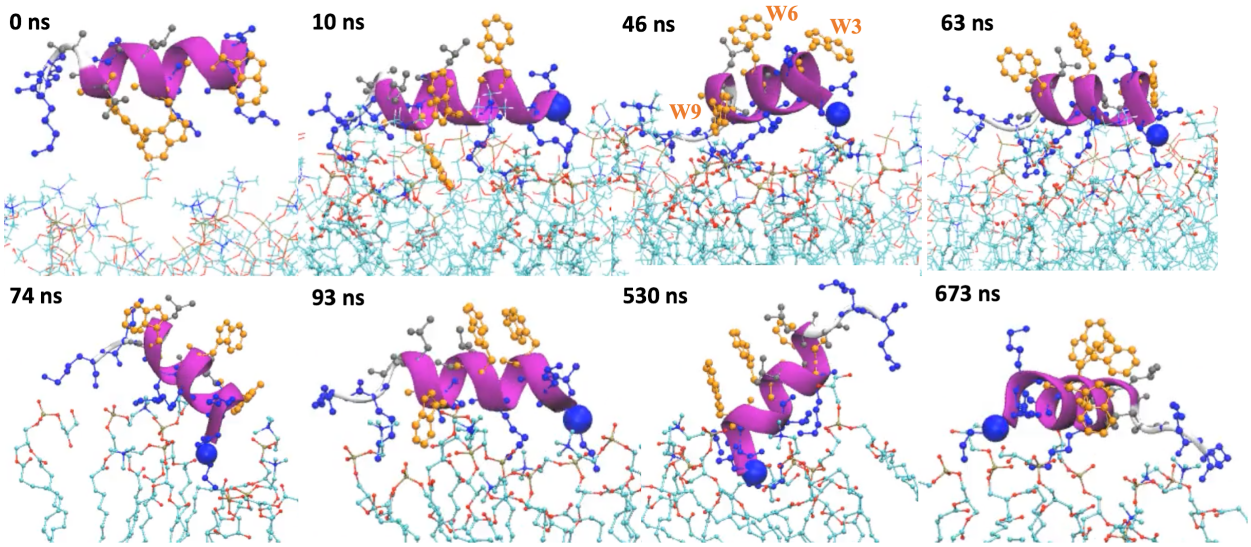
With microsecond simulations, we revealed types of contact modes and time-resolved atomic interaction between AMPs and lipid bilayers before AMPs' final insertion into the bacterial membrane (see **Movie S1 and Fig S2**). Taking W3_p1 for example, as shown in Figure 3, lysine and arginine with long positively charged side chains preferentially interact with the phosphate groups in PC and PG (see **Fig S1**). These interactions not only "raise" some interacting lipid molecules out of the upper leaflet but also push away these phosphate groups due to the long arms of these positively charged side chains. This creates what we call "hydrophobic cradle" which is a

temporary vacuum created due to the aforementioned Lys/Arg-lipid interaction where at the bedding of the cradle lie the exposed aliphatic lipid tails whose head groups have been pushed either side way or upward. One can see such cradles form at 46, 673, 678 and 715 ns in **Figure 3**. This provides an important opportunity for adjacent tryptophan residue to flip in to enjoy such a hydrophobic interaction only when the orientation is right. Either W3 or W9 (see 46 ns in **Fig 3**) close to N- or C-terminal double lysine residues can be the first tryptophan to take advantage of these hydrophobic cradles for insertion. However, W9 insertion near the C-terminus is transient (starting right in the beginning, at 10 ns, drifting away at 46 ns, resuming the insertion at 93 ns but then drifting away again soon after) (**Movie S1 and Fig 3**). It cannot be maintained because the horizontal insertion pops the other two tryptophan up or side way toward the bulk solvent. On the other hand, W3-initiated insertion benefits not only from the transiently formed hydrophobic cradles but also from the downward facing N-terminus (see 530 ns, 715 ns and thereafter in Fig 3) that further brings W6 to contact the deeper part of a lipid (e.g. 715 ns and thereafter). The N-terminus- rather than C-terminus-initiated penetration is consistent with aforementioned experimental results[49] and also common to most of the membrane anchoring signals[50]. N-terminus seems to better interact with the phosphate groups of lipids (e.g. at 678ns) and transiently interact with the lipid tails better than the C-terminus.

As W3 starts to insert, its irreplaceable chemical importance in membrane insertion becomes immediately apparent in these time-resolved snapshots. As shown at 678 ns in Figure 3, it can first interact with the choline groups of PC lipids through the cation–π interaction. Then, it can interact with the phosphate groups (715 ns) or deeper side of the glycerol moiety (735 ns) in the same or different lipids via hydrogen bonds. Eventually W3 reaches and interacts stably with lipid tails due to the hydrophobic interaction (741 ns onward). As W3 reaches the lipid phosphate groups or glycerol moieties, W6 can start to interact with the choline group of the same lipid (e.g. at 715 ns). When W6 starts to reach the phosphate group (735 ns) and lipid tails (741 ns), the originally vertical W3_p1 starts to lean down with the N-terminus still diving a bit deeper than the C-terminus, which brings W9 closer to the lipid on the side way. At 756 ns, we see that W9 starts to interact with two choline groups of PC lipids. As W9 starts to sink deeper to interact with the phosphate group of a lipid (e.g. at 866 ns), to allow W3 and W6 being still stably anchored in the lipid tails, the N-terminus starts to open up and becomes less helical (866 ns onward), which helps split W3 and W6 further apart and bring W9 to rotate further downward and better align W6 and W9 on the

same side. This also synchronizes with positively charged residues being upward facing and sticking their side chains toward the bulk solvent (968 ns onward).

In summary, W3_p first interacts with lipids via electrostatic interaction with the phosphate groups of the lipids popping up tryptophans in the bulk solvent. As N-terminus dives downward in the hydrophobic cradle, created by the long positively charged side chains to push the lipid heads away and expose their lipid tails, the W3 is brought together to interact with the deeper part of the lipids, which makes the AMP stand vertically on the lipid surface. As W6 further crawls in and interacts with phosphate groups of the lipids, or deeper, the AMP starts to lean down with N-terminus in turn to become "loopy" and less helical, which helps W9 be better aligned with W6 and insert deeper. As also evident in the Movie S1, within the 1 microsecond, all the 3 tryptophan residues in W3_p1 are inserted into the bacterial membrane while W3_p2, which is strictly helical, has only two tryptophan residues (W3 and W6) inserted into the membrane (especially after 800 ns). In short, these AMPs insert the bacteria membrane heavily relying on the *three tryptophan anchors* equally spaced with two amino acids.



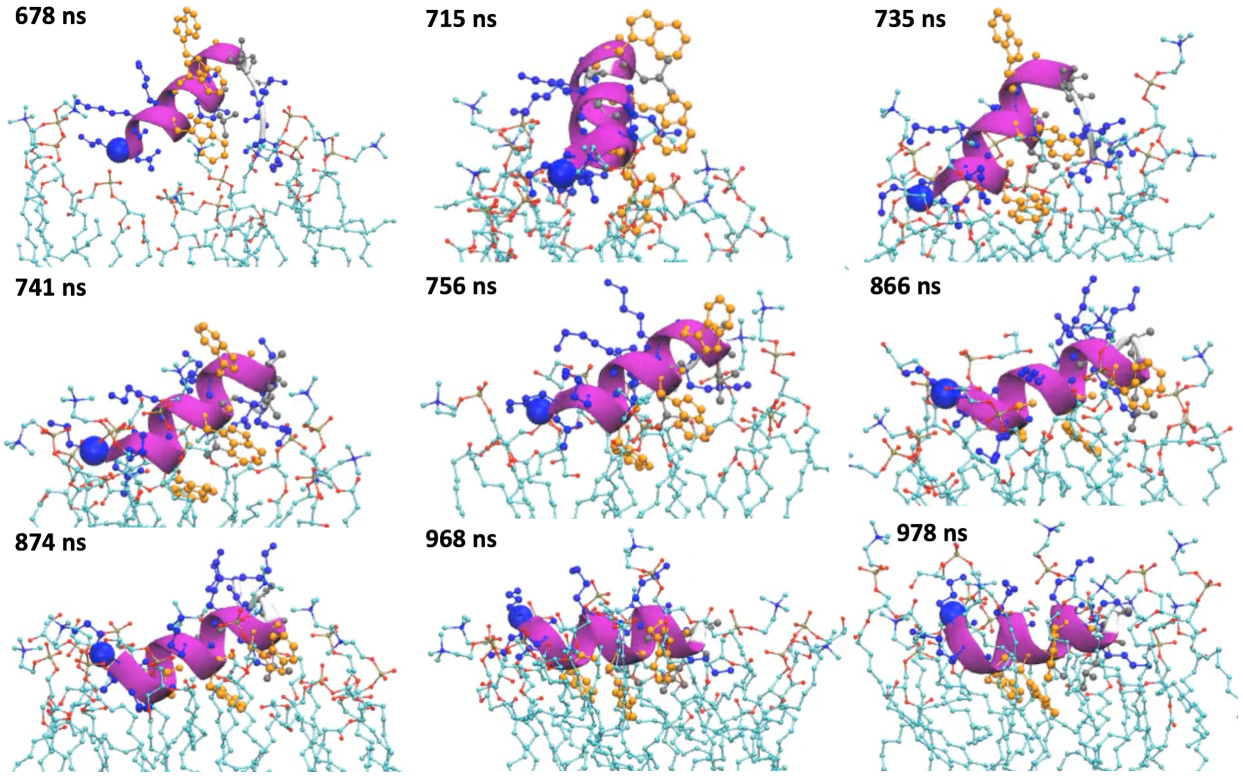


Fig. 3. Mechanistic details of W3_p1's insertion into the bacterial membrane revealed by 1 μs MD simulations. The Cα atom of the first residues (N-terminus) is highlighted with a blue VDW ball, from which the three tryptophan residues, shown in golden ball-and-stick, are W3, W6 and W9 (see also the panel of 46 ns). The positively charged residues in the AMP are represented in blue CPK. Starting from the 74th nanosecond, only the lipids staying within 3 Å from the AMPs are shown while part of their aliphatic tails can be removed for clarity. The heavy atoms of the lipids are represented in thin CPK; O, P and C atoms are colored in red, brown and cyan, respectively. The N atoms in the choline groups of POPC lipids that engage cation-π interactions with the tryptophan are colored in blue.

Knowing the aforementioned insertion process and the importance of the pattern "W**W**W" (or [W 2 W 2 W] as the input query in TP-DB) in a helical structure, we are interested in finding other helical peptides in TP-DB with the following three characteristics - (1) containing the W2W2W motif, (2) maintaining helicity in isolation (with relative high HP) and (3) enriched with positively charged residues, so that the corresponding part in W3_p1 or W3_p2 can be replaced by the TP-found helical stretches (**Table 2**) to preserve the terminal lysine anchors.

105 helical peptides from different PDB IDs (and chain IDs) containing 11 unique **"**W**W**W**"** motifs can be found through the query [W 2 W 2 W] (**Table 2).** The peptide **"**WKCWARWRL" (the 5th in the #U column) having the most positively charged residues and a relatively high HP is

chosen as our top AMP candidate. The peptide is derived from the mycobacterium tuberculosis Zinc-dependent metalloprotease-1 (Zmp1) (PDB: 3ZUK), which is a soluble protein and does not associate with the membrane. The peptide sits at the relatively outerior side of the protein but is not wide open to the bulk solvent. The flanking double Lys residues are further added to extend this motif into a full-fledged AMP candidate. This newly designed AMP, named as W3_db5, is then investigated by MD simulations and synthesized for the subsequent bactericidal experiments.

Table 2. Results obtained from the TP-DB for the query [W 2 W 2 W]

#	U#	Matched Sequence (MS)	Matched Pattern	Full Helix (FH)	PDB ID: Chain	Positions in PDB (MS), (FH)	Helical Propensity	Contact	Interacting Partners
1	1	WLEWIRW	W2W2W	DVQQLLIWLEWIRWESD	4eba:F	(287, 293), (280, 296)	1.189	8.857	4eba:F (324, 338)
10	2	WQQWLNW	W2W2W	DLATFWQQWLNW	3tsu:A	(651, 657), (646, 657)	1.131	7.857	
11	3	WQRWENW	W2W2W	VNSYLWQRWENWFWNVTLR	2lzq:A	(10, 16), (5, 23)	1.105	8.000	
12	4	WDAWLNW	W2W2W	WDAWLNWFR	1z8h:D	(198, 204), (198, 206)	1.089	7.286	
13	5	WKCWARW	W2W2W	NVEDWKCWARWRLIRARA	3zuk:B	(283, 289), (279, 296)	1.055	9.143	3zuk:B (415, 434)
15	6	WAWWISW	W2W2W	AGEWLGSWTIFYWAWWISWSPFVGMFLAR	4llh:A	(371, 377), (359, 387)	1.039	10.571	4llh:A (449, 479) 4llh:A (511, 525) 4llh:A (527, 545)
25	7	WVDWWQW	W2W2W	QWVDWWQWWVK	4n7k:L	(259, 265), (258, 268)	0.985	7.286	4n7k:M (81, 88) 4n7k:M (178, 192)
86	8	WKDWESW	W2W2W	SVLRKALHDSLHDCSHWFYTRWKDWESWYS QSF	3zrh:A	(481, 487), (460, 492)	0.799	8.857	3zrh:A (522, 533)
87	9	WTAWSTW	W2W2W	WTAWSTWKYC	3cb7:A	(106, 112), (106, 115)	0.638	7.714	
88	10	WPEWWNW	W2W2W	GWPEWWNWWLE	3wmm:L	(268, 274), (267, 277)	0.345	8.857	3wmm:M (82, 90) 3wmm:M (180, 193)
89	11	WPEWWGW	W2W2W	WPEWWGWWL	7prc:L	(259, 265), (259, 267)	-0.020	8.429	

As shown in https://dyn.life.nthu.edu.tw/design/result?JobID=602a761eg, totally 105 helical peptides can be found with the query [W 2 W 2 W]. Only the unique 11 sequences (indexed in the U# column) are shown here and the red block highlights the new helical stretch we used for the AMP design.

Contrasting to the simulation results of W3_p1 and W3_p2, W3_db5 has only one Trp (W3) inserted into the bacterial membrane by the end of 1 microsecond simulation (Movie S1 and Fig S2). To further examine the bactericidal and antifungal activities of these AMPs, we measure the minimal inhibitory concentration (MIC_{90}) at which the growth of 90% of the pathogenic fungus *C. albicans* SC5314 and *E. coli* ATCC 25922 can be inhibited. Experiments are performed in pentaplicate and MIC_{90} is determined as the majority results of the 5 replicates (see Methods). It is worth mentioning that we could reproduce the reported MIC_{90} values for three published AMPs when using the same culture protocol to culture *E. coli* ATCC 25922. They are CM15, Anoplin-R5KT8W and FK13 with reported MIC values of 4[51], 15.13[52] and 27.5[53] μg/ml, respectively, while we obtained 6.25, 12.5 and 25 μg/ml for the same. Our results (**Table 3**) show that the newly designed AMP W3_db5 (KK **W**KC **W**AR **W**RL KK) has an anti-fungal MIC_{90} of 3.75 μg/ml, which is twice better than that of W3_p1 (7.5 μg/ml) and W3_p2 (7.5 μg/ml), while it is a level

lower in its bactericidality with a MIC_{90} of 30 μg/ml than that of W3_p1 (15 μg/ml) and W3_p2 (15 μg/ml). The latter result is consistent with the MD-based natural insertion results (Fig S2).

Although our newly designed W3_db5 has similar anti-bacterial and anti-fungal activities as its mother templates', it has a surprisingly low cytotoxicity as compared with the template AMPs. As shown in Figure 4, at the concentration of 240 μg/ml, W3_p2, W3_p1 and W3_db5 lyse 20.3%, 13.0% and 0.6%, respectively, the red blood cells (RBCs). W3_db5 is at least 20-fold less hemolytic than W3_p1 and W3_p2 (at least 12-fold less hemolytic at 120 μg/ml); in other words, we can use 20-fold higher concentration than W3_db5's MIC to still have a smaller cytotoxicity than its W3 templates. In Table 3, we also list AMPs' minimal hemolytic concentration (MHC_5) to lyse 5% of the RBCs, where W3_db5's MHC_5 can be much larger than 240 μg/ml, which implies even a high concentration of W3_db5 can still be potentially used in medication or personal hygiene products.

To further validate the use of TP-DB, as a control study, we borrow the conventional PHI-BLAST to find a sequence homologous to W3_p1 and containing the motif "W2W2W". Here we have to note that PHI-BLAST has to take two input to initiate a search - a protein sequence S and a regular expression pattern P occurring in S. Given the two inputs, PHI-BLAST helps finding protein sequences both containing an occurrence of P and are homologous to S in the vicinity of where the pattern P occurs. Here, P as "W-x(2)-W-x(2)-W" and S as the sequence of W3_p1 are used to search the PDB via the NCBI PHI-BLAST website. A single chain protein (PDB: 4F54) from the DUF4136 family with a relatively high resolution (1.60 Å) contains a stretch of "**W**GN **W**GG **W**", located at the end of a helix connecting toward a beta-sheet through a short loop. We synthesize a peptide "KK WGN WGG WRL KK", coined as "W3_n1", which replaces the "**W**KC **W**AR **W**" part in W3_db5 with the "**W**GN **W**GG **W**". When examining its MIC, we fail to find its MIC_{90} given the concentration range we have tested (the highest at 120 μg/ml) (see **Table 3**), suggesting not any non-helical "W2W2W" sequence can work as a functional AMP motif. As a second control study, we take another helical "W2W2W" stretch in **Table 2**, herein the 8th unique peptide "**W**KD **W**ES **W**", to a new "AMP". That means we replace the "**W**KC **W**AR **W**" part in W3_db5 with "**W**KD **W**ES **W**" to create the peptide "KK **W**KD **W**ES **W**RL KK", which is termed as "W3_n2". The "**W**KD **W**ES **W**RL" stretch has one less positively charged residue and two more negatively charged residues (D and E) as compared with the "**W**KC **W**AR **W**RL" motif in W3_db5.

Again, we do not find its MIC_{90} within the concentration range we have tested (the highest at 120 μg/ml) (see **Table 3**), which suggest a helical stretch containing the W2W2W motif does not necessarily guarantee its bactericidal activity if there are not enough positive charges in the designed AMPs. Note that the TP-DB-found W3_db5 has the same number of positive charges as its W3 templates (i.e. W3_p and W3_p2), while W3_n2 could be 3 positive charges short to become an AMP. In the MIC_{90} experiments, W3_n1 and W3_n2 respectively kill 2.5% and 9.5% of the *E. coli* ATCC 25922 at a concentration of 30 μg/ml, which is the MIC_{90} value of W3_db5 for the same bacteria.

Table 3. The anti-fungal, antibacterial activity and hemolytic toxicity of TP-DB-designed AMP (W3_db5) and other AMPs containing the "W2W2W" motif.

Name	Sequence	MIC_{90} vs. *C. albicans* (μg/mL)	MIC_{90} vs. *E. coli* (μg/mL)	MHC_5 (μg/mL)
W3_p1	KK **W**RK **W**LK **W**LA KK	7.5	15	120
W3_p2	KK **W**LK **W**LK **W**LK KK	7.5	15	90
W3_db5	KK **W**KC **W**AR **W**RL KK	3.75	30	>> 240
W3_n1	KK **W**GN **W**GG **W**RL KK	N/A	>> 120	N/A
W3_n2	KK **W**KD **W**ES **W**RL KK	N/A	>> 120	N/A

MIC_{90} means the minimal concentration of AMP to inhibit 90% growth of pathogens, while MHC_5 means the minimal concentration of AMP to induce 5% hemolysis against human erythrocytes.

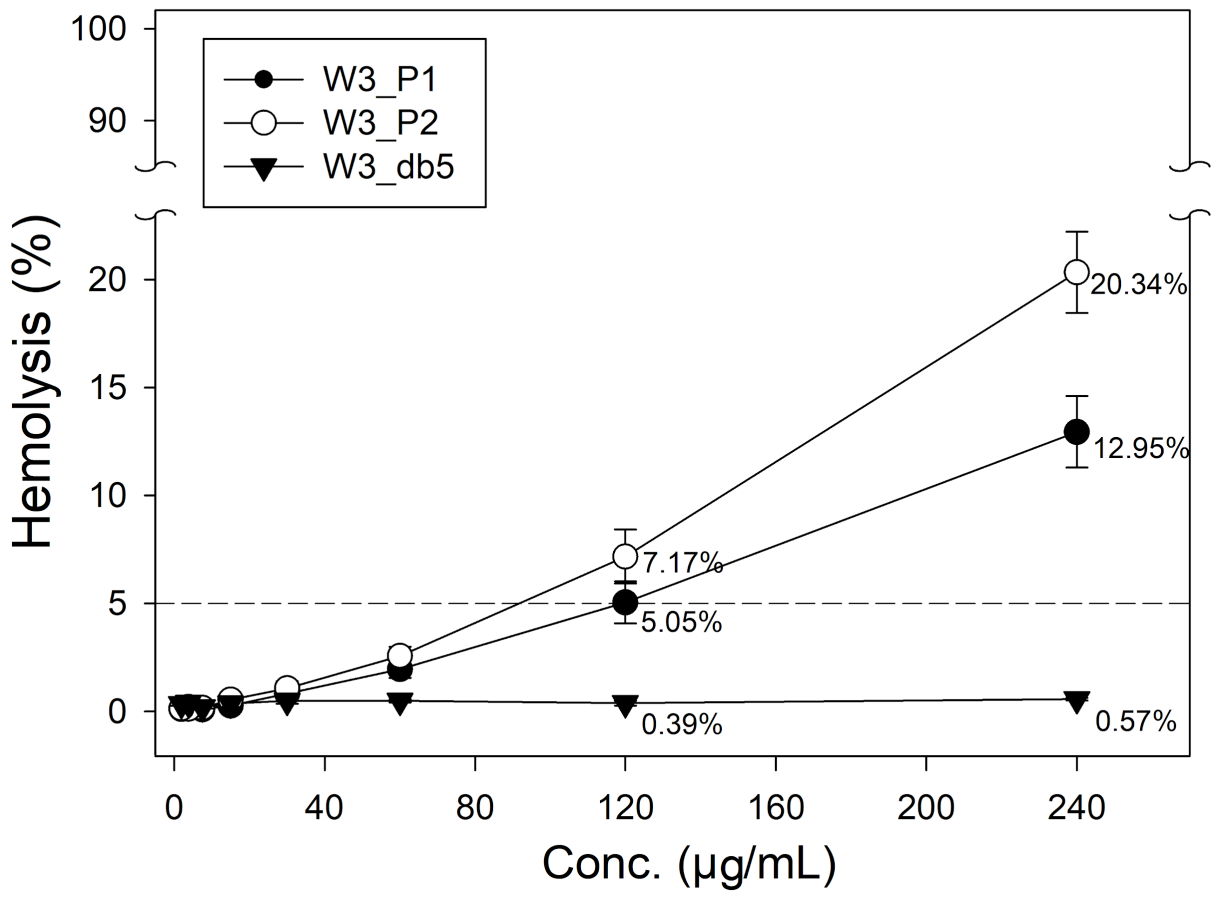


Figure 4. The hemolytic percentage as a function of AMP concentrations for W3_p1, W3_p2 and W3_db5. W3_db5 has a much lower hemolysis of red blood cells (RBCs) than that of W3_p1 and W3_p2. The dash line indicates the 5% hemolytic threshold that defines MHC_5. Data are presented by the mean ± S.D. from three repeats.

It should be noted that the original protein (mycobacterium tuberculosis Zinc-dependent metalloprotease-1; PDB ID: 3ZUK) that contains this potent stretch "**W**KC **W**AR **W**RL" in W3_db5 is not a transmembrane or membrane-bound protein; the helical stretch is situated close but not fully exposed to the water-protein interface. Therefore, TP-DB brings an interesting opportunity for researchers to put together structurally resolved elements, matching a desired pattern, for the rational design of new therapeutics, herein showcased by a new class of AMP that is 6 point mutations away from its original templates - W3_p1 and W3_p2.

From the design point of view, these 6 point mutations do not seem to change much the required insertion thermodynamics[54] in the new AMP, W3_db5, which can still penetrate the peptidoglycans, reach the lipid-water interface, stay amphiphilic and helical, and eventually insert into the cell membrane. This is difficult to achieve by mutagenesis experiments or MD simulations without the guidance of TP-DB.

3.4. Design Peptide Inhibitors for Tumorigenic PPI

Our statistics in **Table 1** also reveal the interesting fact that alanine is indeed a safe choice for point mutation while still maintaining the integrity of a helix (see HP_{NA}). However, with the pattern-based search engine (see Methods section 2.2) made available in TP-DB, one can do point mutation(s), by choosing the residues suggested by TP-DB, in a helix to abolish or enhance (see Supplementary Material) its interaction with certain drugs, peptides or proteins while still maintaining the helicity of the helix, as well as to abolish the catalytic function of a residue on an α-helix while maintaining its binding affinity with certain substrate or ligand in the scenario where a structural biologist co-crystalizes an enzyme with its cognate substrate. In the Supplementary Material, we demonstrate an *in-silico* study on how peptide blockers can be designed by TP-DB to block a helix-helix interface in order to suppress the late-stage hepatocellular carcinoma.

4. Conclusions

The pattern-based search engine without relying on sequence homology, realized in the newly introduced TP-DB, has been shown to provide a new avenue to the re-use or re-assembly of structured peptide fragments into new therapeutics or diagnostic kits. New therapeutics can therefore be designed not only with known functional sequence motifs but also with desired secondary structures. We also foresee its future extension to incorporate search engines for beta-sheet elements containing specific sequences and/or patterns to meet desired physicochemical traits with a design purpose.

Acknowledgements

E.O.S. acknowledges financial support from MOST and Taiwan International Graduate Program, Academia Sinica, Taipei, Taiwan. We thank vast computational resources for this project provided by High Performance Computing Infrastructure (HPCI), Japan and National Center for High-performance Computing (NCHC) of National Applied Research Laboratories (NARLabs) of Taiwan. This work was supported by the Ministry of Science and Technology, Taiwan (103-2627-M-007-001 and 104-2113-M-007-019 to L.W.Y; 104-2311-B-007-003 and 108-2311-B-007-004 to H.W.F) and the research program of National Tsing Hua University (104N2052E1 to H.W.F).

References

1. Respondek M, Madl T, Göbl C, Golser R, Zangger K. Mapping the orientation of helices in micelle-bound peptides by paramagnetic relaxation waves. *Journal of the American Chemical Society* **129**, 5228-5234 (2007).

2. Yang L-W, Cheng J-W, Li H-C, Tsai C-Y, Hui-Yuan Y. Evaluation system for the efficacy of antimicrobial peptides and the use thereof. *United States Patent* *https://patentsgooglecom/patent/US20170061288A1/en*, (2020).

3. Tossi A, Sandri L, Giangaspero A. Amphipathic, α-helical antimicrobial peptides. *Peptide Science* **55**, 4-30 (2000).

4. Porto WF*, et al.* In silico optimization of a guava antimicrobial peptide enables combinatorial exploration for peptide design. *Nature communications* **9**, 1-12 (2018).

5. Vandromme M, Cavadore J-C, Bonnieu A, Froeschlé A, Lamb N, Fernandez A. Two nuclear localization signals present in the basic-helix 1 domains of MyoD promote its active nuclear translocation and can function independently. *Proceedings of the National Academy of Sciences* **92**, 4646-4650 (1995).

6. Yang L-W, Kitao A, Huang B-C, Gō N. Ligand-induced protein responses and mechanical signal propagation described by linear response theories. *Biophysical journal* **107**, 1415-1425 (2014).

7. Johnson LN, Barford D. Glycogen phosphorylase: the structural basis of the allosteric response and comparison with other allosteric proteins. *The Journal of biological chemistry (Print)* **265**, 2409-2412 (1990).

8. Chang K-C, Salawu EO, Chang Y-Y, Wen J-D, Yang L-W. Resolution-exchanged structural modeling and simulations jointly unravel that subunit rolling underlies the mechanism of programmed ribosomal frameshifting. *Bioinformatics* **35**, 945-952 (2019).

9. Huang B-C, Yang L-W. Molecular dynamics simulations and linear response theories jointly describe biphasic responses of myoglobin relaxation and reveal evolutionarily conserved frequent communicators. *Biophysics and physicobiology* **16**, 473-484 (2019).

10. Lee EY, *et al.* Helical antimicrobial peptides assemble into protofibril scaffolds that present ordered dsDNA to TLR9. *Nature communications* **10**, 1-10 (2019).

11. Xu Z, *et al.* Structure and function of the PP2A-shugoshin interaction. *Molecular cell* **35**, 426-441 (2009).

12. Suzuki H, Kolaskar A, Samuel S, Otsuka J, Tsugita A. A protein secondary structure database (PSS). *Protein sequences & data analysis* **4**, 97-104 (1991).

13. Joosten RP, *et al.* A series of PDB related databases for everyday needs. *Nucleic acids research* **39**, D411-D419 (2010).

14. Jencks WP. On the attribution and additivity of binding energies. *Proceedings of the National Academy of Sciences* **78**, 4046-4050 (1981).

15. Erlanson DA. Introduction to fragment-based drug discovery. *Fragment-based drug discovery and X-ray crystallography*, 1-32 (2011).

16. Kutchukian PS, Lou D, Shakhnovich EI. In Silico Fragment-Based Generation of Drug-Like Compounds. In: *Library Design, Search Methods, and Applications of Fragment-Based Drug Design*). ACS Publications (2011).

17. Zhang Z, *et al.* Protein sequence similarity searches using patterns as seeds. *Nucleic acids research* **26**, 3986-3990 (1998).

18. Berman HM, *et al.* The protein data bank. *Nucleic acids research* **28**, 235-242 (2000).

19. Phillips JC, *et al.* Scalable molecular dynamics with NAMD. *Journal of computational chemistry* **26**, 1781-1802 (2005).

20. Eastman P, *et al.* OpenMM 7: Rapid development of high performance algorithms for molecular dynamics. *PLoS computational biology* **13**, e1005659 (2017).

21. Klauda JB, *et al.* Update of the CHARMM all-atom additive force field for lipids: validation on six lipid types. *The journal of physical chemistry B* **114**, 7830-7843 (2010).

22. Best RB, *et al.* Optimization of the additive CHARMM all-atom protein force field targeting improved sampling of the backbone ϕ, ψ and side-chain $\chi 1$ and $\chi 2$ dihedral angles. *Journal of chemical theory and computation* **8**, 3257-3273 (2012).

23. Darden T, York D, Pedersen L. Particle mesh Ewald: An N· log (N) method for Ewald sums in large systems. *The Journal of chemical physics* **98**, 10089-10092 (1993).

24. McQuarrie D, Allan D. Statistical mechanics, II ed. *University Science Books, Sausalito, Calif*, (2000).

25. Feller SE, Zhang Y, Pastor RW, Brooks BR. Constant pressure molecular dynamics simulation: the Langevin piston method. *The Journal of chemical physics* **103**, 4613-4621 (1995).

26. Tsai C-W, *et al.* Coupling molecular dynamics simulations with experiments for the rational design of indolicidin-analogous antimicrobial peptides. *Journal of molecular biology* **392**, 837-854 (2009).

27. Wang Y, Schlamadinger DE, Kim JE, McCammon JA. Comparative molecular dynamics simulations of the antimicrobial peptide CM15 in model lipid bilayers. *Biochimica et Biophysica Acta (BBA)-Biomembranes* **1818**, 1402-1409 (2012).

28. Jo S, Lim JB, Klauda JB, Im W. CHARMM-GUI Membrane Builder for mixed bilayers and its application to yeast membranes. *Biophysical journal* **97**, 50-58 (2009).

29. Mahoney MW, Jorgensen WL. A five-site model for liquid water and the reproduction of the density anomaly by rigid, nonpolarizable potential functions. *The Journal of chemical physics* **112**, 8910-8922 (2000).

30. Kučerka N, Tristram-Nagle S, Nagle JF. Structure of fully hydrated fluid phase lipid bilayers with monounsaturated chains. *The Journal of membrane biology* **208**, 193-202 (2006).

31. Humphrey W, Dalke A, Schulten K. VMD: visual molecular dynamics. *Journal of molecular graphics* **14**, 33-38 (1996).

32. Reference Method for Broth Dilution Antifungal Susceptibility Testing of Yeasts, Approved Standard—Third Edition. *Clinical Laboratory Standards Institute* **M27-A3**, (2008).

33. Methods for Dilution Antimicrobial Susceptibility Tests for Bacteria That Grow Aerobically, Approved Standard—Ninth Edition. *Clinical Laboratory Standards Institute* **M07-A9**, (2012).

34. Rothstein DM, *et al.* Anticandida activity is retained in P-113, a 12-amino-acid fragment of histatin 5. *Antimicrobial agents and chemotherapy* **45**, 1367-1373 (2001).

35. Tramer F, Da Ros T, Passamonti S. Screening of fullerene toxicity by hemolysis assay. In: *Nanotoxicity*). Springer (2012).

36. Gasteiger E, Hoogland C, Gattiker A, Wilkins MR, Appel RD, Bairoch A. Protein identification and analysis tools on the ExPASy server. *The proteomics protocols handbook*, 571-607 (2005).

37. Hopp TP, *et al.* A short polypeptide marker sequence useful for recombinant protein identification and purification. *Bio/technology* **6**, 1204-1210 (1988).

38. Slootstra J, Kuperus D, Plückthun A, Meloen R. Identification of new tag sequences with differential and selective recognition properties for the anti-FLAG monoclonal antibodies M1, M2 and M5. *Molecular diversity* **2**, 156-164 (1997).

39. McGann P, *et al.* Escherichia coli harboring mcr-1 and blaCTX-M on a novel IncF plasmid: first report of mcr-1 in the United States. *Antimicrobial agents and chemotherapy* **60**, 4420-4421 (2016).

40. Antimicrobial resistance: tackling a crisis for the health and wealth of nations. *Review on Antimicrobial Resistance*https://wellcomecollectionorg/works/rdpck35v, (2014).

41. Ulmschneider JP, Ulmschneider MB. Molecular dynamics simulations are redefining our view of peptides interacting with biological membranes. *Accounts of chemical research* **51**, 1106-1116 (2018).

42. Chen CH*, et al.* Simulation-guided rational de novo design of a small pore-forming antimicrobial peptide. *Journal of the American Chemical Society* **141**, 4839-4848 (2019).

43. Chung C-R*, et al.* Characterization and identification of natural antimicrobial peptides on different organisms. *International journal of molecular sciences* **21**, 986 (2020).

44. Chung C-R, Kuo T-R, Wu L-C, Lee T-Y, Horng J-T. Characterization and identification of antimicrobial peptides with different functional activities. *Briefings in bioinformatics* **21**, 1098-1114 (2020).

45. Yu HY*, et al.* Rational design of tryptophan-rich antimicrobial peptides with enhanced antimicrobial activities and specificities. *Chembiochem* **11**, 2273-2282 (2010).

46. De Santis E*, et al.* Antimicrobial peptide capsids of de novo design. *Nature communications* **8**, 1-11 (2017).

47. Pyne A*, et al.* Engineering monolayer poration for rapid exfoliation of microbial membranes. *Chemical science* **8**, 1105-1115 (2017).

48. Chan DI, Prenner EJ, Vogel HJ. Tryptophan-and arginine-rich antimicrobial peptides: structures and mechanisms of action. *Biochimica et Biophysica Acta (BBA)-Biomembranes* **1758**, 1184-1202 (2006).

49. Li Y*, et al.* Single-molecule visualization of dynamic transitions of pore-forming peptides among multiple transmembrane positions. *Nature communications* **7**, 1-9 (2016).

50. Pfanner N, Neupert W. The mitochondrial protein import apparatus. *Annual review of biochemistry* **59**, 331-353 (1990).

51. Moghaddam MM, Abolhassani F, Babavalian H, Mirnejad R, Barjini KA, Amani J. Comparison of in vitro antibacterial activities of two cationic peptides CM15 and CM11 against five pathogenic bacteria: Pseudomonas aeruginosa, Staphylococcus aureus, Vibrio cholerae, Acinetobacter baumannii, and Escherichia coli. *Probiotics and antimicrobial proteins* **4**, 133-139 (2012).

52. Uggerhøj LE*, et al.* Rational Design of Alpha-Helical Antimicrobial Peptides: Do's and Don'ts. *ChemBioChem* **16**, 242-253 (2015).

53. Rajasekaran G, Kim EY, Shin SY. LL-37-derived membrane-active FK-13 analogs possessing cell selectivity, anti-biofilm activity and synergy with chloramphenicol and anti-inflammatory activity. *Biochimica et Biophysica Acta (BBA)-Biomembranes* **1859**, 722-733 (2017).

54. Almeida PF, Pokorny A. Mechanisms of antimicrobial, cytolytic, and cell-penetrating peptides: from kinetics to thermodynamics. *Biochemistry* **48**, 8083-8093 (2009).

SUPPLEMENTARY DATA

for

Secondary Structure Motifs Made Searchable to Facilitate the Functional Peptide Design

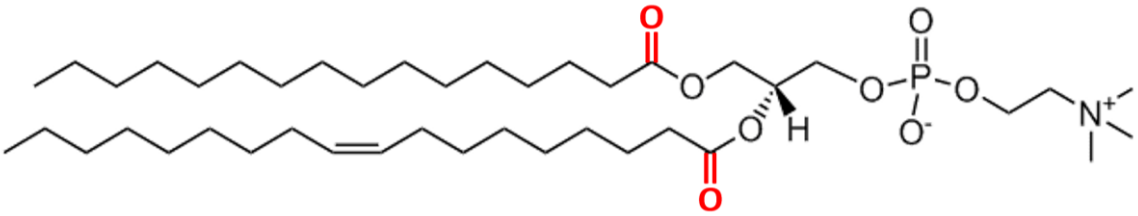
Cheng-Yu Tsai[1,2,¶], Emmanuel O Salawu[1,3,¶], Hongchun Li[1,4,¶], Guan-Yu Lin[5,¶], Ting-Yu Kuo[5,¶], Liyin Voon[1,¶], Adarsh Sharma[1], Kai-Di Hu[1], Yi-Yun Cheng[6], Sobha Sahoo[1], Lutimba Stuart[1], Chih-Wei Chen[1], Yuan-Yu Chang[1], Yu-Lin Lu[1], Ximai Ke[1], Chen-Chi Wu[2], Chung-Yu Lan[5,7]*, Hua-Wen Fu[5,7]*, Lee-Wei Yang[1,3,7,8,]*

[1]Institute of Bioinformatics and Structural Biology, National Tsing-Hua University, Hsinchu 30013, Taiwan; [2]Graduate Institute of Medical Genomics and Proteomics, National Taiwan University College of Medicine, Taipei 10055, Taiwan; [3]Bioinformatics Program, Institute of Information Sciences, Academia Sinica, Taipei 11529, Taiwan; [4]Department of Computational and Systems Biology, School of Medicine, University of Pittsburgh, Pittsburgh, PA, 15213, USA; [5]Institute of Molecular and Cellular Biology, National Tsing Hua University, Hsinchu 30013, Taiwan; [6]Praexisio Taiwan Inc. New Taipei 22180, Taiwan; [7]Department of Life Science, National Tsing Hua University, Hsinchu 30013, Taiwan; *[8]Physics Division, National Center for Theoretical Sciences, Hsinchu 30013, Taiwan.*

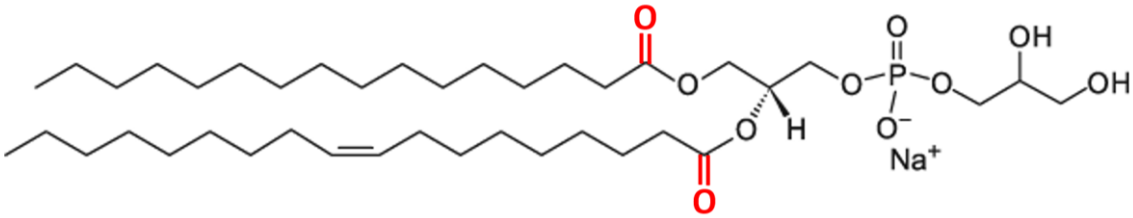
¶ Authors share equal contributions

Supplementary Results

Chemical structures of POPC and POPG lipids



1-palmitoyl-2-oleoyl-glycero-3-phosphocholine, POPC (PC)



1-palmitoyl-2-oleoyl-sn-glycero-3-phosphoglycerol, POPG (PG)

Figure. S1 Chemical structure of POPC (PC) and POPG (PG). All images modified from Avanti® product page (Product ID: 840457 and 850457). The carbonyl oxygen in the lipids are highlighted in red, and these oxygen atoms are represented as red points in Fig. 3A. The atoms of the inserted AMPs below these oxygen atoms are considered to interact with the aliphatic tails of the lipids.

Natural insertion of three AMPs into bacteria membrane

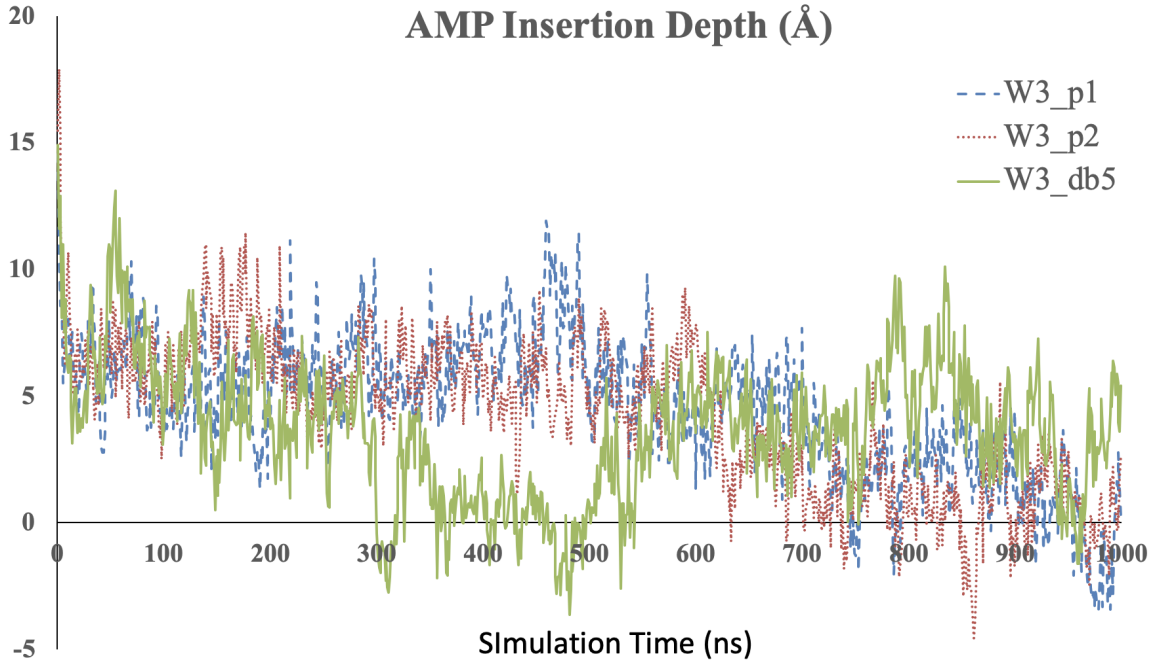


Figure S2. The relative position of the center of mass (COM) of AMPs to the COM of phosphor atoms in the upper leaflet consisting of 40 POPC/POPG (3:1) lipid molecules, mimicking bacteria membrane, is plotted against the simulation time. All the MD simulations were conducted for 1 microsecond (us) at body temperature and normal pressure (1 atm) using OpenMM[20] with CHARMM36 forcefield[21,22].

Supplementary Movie S1

All the three AMPs were first created as fully extended AMPs and they were then simulated in an NPT ensemble for 50 ns in the absence of lipids. The resulting snapshots of the AMPs were clustered into 5 groups based on their structural similarity using the "clustering" plug-in of the VMD software[31]. The representative conformation from the biggest cluster in each of the three AMPs was selected as the initial conformation to be simulated from time zero in an NPT ensemble in the presence of the already equilibrated PC/PG (3:1) membrane. At time zero, as seen in the video, all the peptides have been folded into a helical shape in the CHARMM36 forcefield[21,22]. Within 1 microsecond, the tryptophan residues in W3_p1 had all three Trp and W3_p2 had two Trp (W3 and W6) inserted into the membrane (especially after 800 ns), while only one Trp (W3) in W3_db5 had been inserted into the membrane.

Design of peptide inhibitor against Sgo1-PP2A interaction

Below we demonstrate how TP-DB can be used to design helical blockers to prevent disease-related protein-protein interaction.

In late-stage hepatocellular carcinoma (HCC), patients rely only on targeted therapy and chemotherapy, which were proved generally ineffective. Novel therapeutic targets for treating HCC are urgently needed. Researchers previously found that protein Shugoshin-1 (Sgo1), protecting the stability of centromeric cohesin at centromere and ensuring proper chromosome separation[55], was up-regulated in HCC. Notably, hepatoma cells were found more sensitive to Sgo1 deficiency than normal hepatocytes[56]. These results could suggest that Sgo1 can be a potential therapeutic target for HCC. It has been known that phosphorylated-Sgo1 recruits protein phosphatase 2A (PP2A) to centromere during early mitosis[57] and the Sgo1-PP2A complex sequestered SA2 subunit of cohesin from Plk1-mediated phosphorylation to maintain its stability[58]. Additionally, Sgo1-PP2A complex also maintains Sororin in a hypo- phosphorylated state and keeps the interaction of Sororin-Pds5 to counteract WAPL[57]. These regulations ensure centromeric cohesin tethering sister chromatids together, until the onset of anaphase (Fig S3). As a result, an inhibitor (possibly a peptide) that blocks the Sgo1-PP2A interaction could potentially suppress the rapidly growing hepatoma cells, susceptible to improper Sgo1-PP2A association, in mitosis.

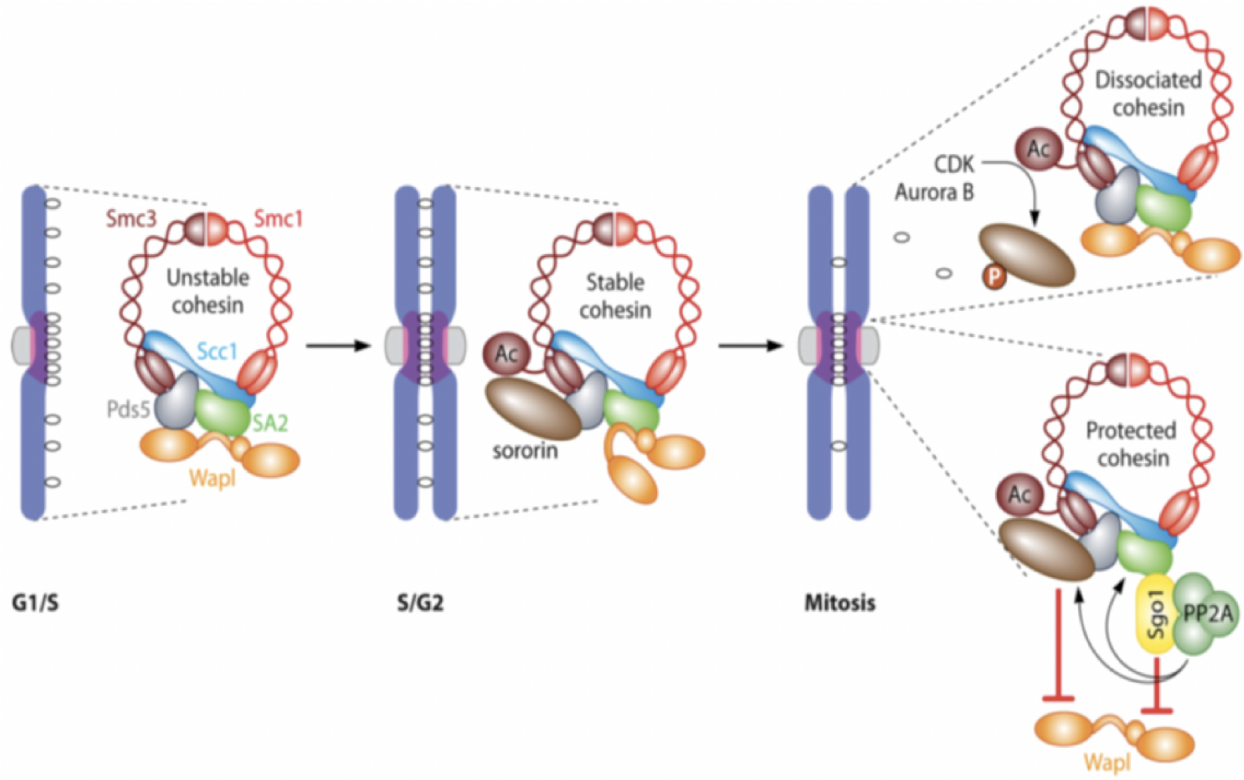


Figure S3 The overview of Sgo1-PP2A protecting centromeric cohesin. During mitosis, phosphorylation of SA2 and Sororin cause cohesin to dissociate from chromosome arm. Cdk1-mediated phospho-Sgo1 recruits PP2A at centromere to counteract phosphorylation of SA2 and Sororin and maintains centromeric cohesin to tether sister chromatids together. Image is adopted from Figure 3 in Marston's review article[59].

We noted that the complex of Sgo1-PP2A has been solved by x-ray crystallography (PDB ID: 3FGA) and the two molecules bind with each other through an helix-helix interface (one helix from each molecule)[11]. To design anticancer peptides as blocker against Sgo1-PP2A interaction, the helical stretch in PP2A (residues from 352-392 in chain B of PDB 3FGA), known to form the helix-helix interaction with another helix in Sgo1 (residues from 75 to 92 in chain D of PDB 3FGA) (Fig S4), is chosen as the candidate target for further investigation. Evidenced from the x-ray structure, the helix-helix interaction is visibly apparent to be the most important essential interacting elements at the protein-protein interface. As shown in Figure S3a, K374, G378 and Y381 in the helical stretch "KTIHGLIY" could be the anchoring residues on PP2A to mediate the Sgo1-PP2A interaction, which is supported by our calculation of energetics using *in silico* alanine scanning (Fig S4b).

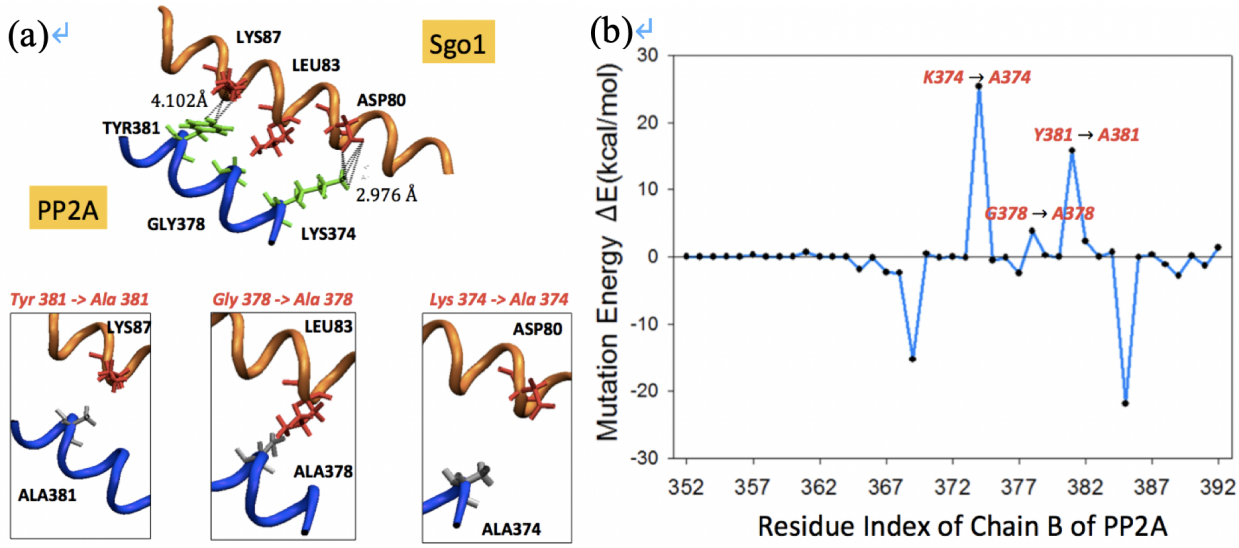


Figure S4 Protein-protein interaction (PPI) between Sgo1 and PP2A through helix-helix interactions. Panel **(a, upper)** shows only the PPI interface between Sgo1 and PP2A, while the rest of proteins (3FGA; Xu et al., 2009) are hidden for clarity. Sgo1-PP2A interaction is mediated by two helices, one from Sgo1 (PDB: 3FGA; V_{75}KEAQDIILQLRKECYYL_{92} in chain D) and the other from PP2A (PDB: 3FGA; K_{369}THMNKTIHGLIYNALK$_{385}$ in chain B) . Computational alanine scanning for every residue in the PP2A helical fragment (from residue 352 to 392 of chain B) is carried out, and in panel (**b**) energetic contribution for each residue mutated into alanine is assessed (from residue 369 to 385 of chain B) where a large positive value (e.g. for K374A, G378A and Y381A) indicates the importance of the residues in binding, and a large negative value suggests mutation into alanine favors the binding. The non-bonded energy is evaluated by NAMD package (Phillips et al., 2005) using CHARMM 36 forcefield (Huang and MacKerell, 2013). Here in panel **(b),** the mutation energy for the n-th residue, ΔEn, is defined as ΔEn = Yn - X0 where X0 is the non-bonded energy of original complex between Sgo1 and PP2A (for the 352-392 fragment), and Yn is the energy of the complex with the *n*-th residue in PP2A being mutated into alanine, after a short energy minimization.

To block the protein-protein interaction (PPI) between Sgo1 and PP2A, one could potentially modify part of the most essential interacting element in PP2A into a peptide that interacts with Sgo1 better than its template. As shown in Figure S4b, computational alanine scanning helps detecting a stretch from 370 to 384 in PP2A (chain B) dominantly contributing the affinity in the PPI of Sgo1-PP2A complex, especially from K374, G378 and Y381 (i.e. aforementioned essential interacting element in PP2A). We then used this helical stretch, or the extended one with its flanking region from residue 364 to 388, including two energetically favored point mutations, K369A and K385A, as the starting template.

To find the helical peptides that bind the interface helix of Sgo1 stronger than PP2A does, we search the TP-DB for PP2A-like helical peptides using the queries summarized in Table S1.

As detailed in Supporting Methods, we first used the pattern "**K***G**Y**" to search TP-DB and discover 69 unique PP2A-like helical peptides (Table S1 and S2); we then carried out molecular dynamics simulations (MD) to assess how these peptides interact with Sgo1 at the helix-helix interface. Helices H60, H54, H61, H33 and H29 (**Fig S5**) are energetically promising for further experimental confirmation. Additionally, we examined other sets of peptides that met the patterns consisting of three anchoring residues ("A***K***G" and "G**Y***A") or consisting of at least four residues ("A***K***G**Y", "K***G**Y***A", "A***K***G**Y***A" and "A**/***K**/***G**/***Y**/***A") in TP-DB. All the corresponding results obtained from the queries are illustrated in **Table S3** to **S8**. Among them only the "A***K***G**Y***A" pattern (according to Figure S3b) returns no result from TP-DB.

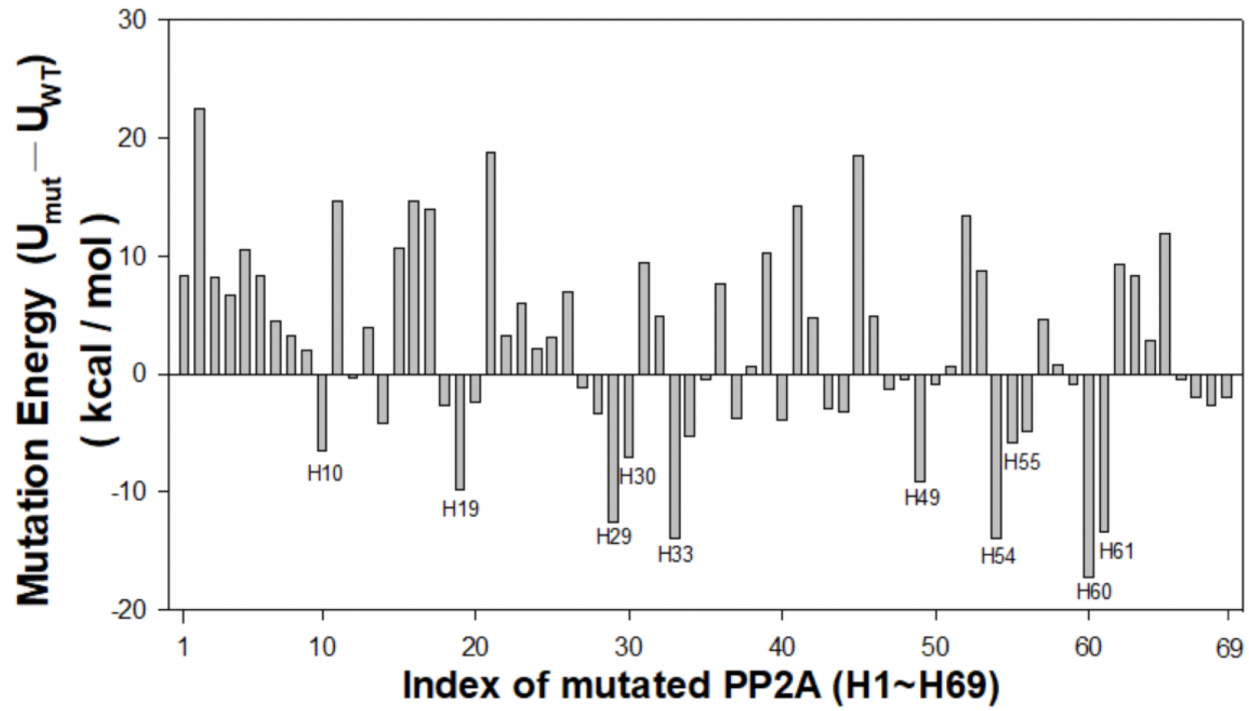


Figure S5. Interactions between Sgo1 helix and PP2A-like peptides. U_{mut} is the interaction energy between targeted Sgo1 helix and each of the 69 PP2A-like peptides (from H1 to H69) that match the pattern "K3G2Y". In detail, the energy is calculated by replacing the residue T375, I376, H377, L379 and I380 in the "$\mathbf{K}_{374}$TIH**G**LI$\mathbf{Y}_{381}$" stretch of PP2A with the corresponding residues in each of the 69 pattern-matched (K3G2Y) helical peptides returned from TP-DB. U_{WT} is the interaction energy between Sgo1 helix and PP2A. Larger difference of interaction energies (U_{mut} - U_{WT}) suggest stronger binding of PP2A-like peptides

with the Sgo1 helix than that between PP2A and the Sgo1 helix. Top ten values are further labeled in the graph.

Table S1. PP2A-like patterns searched against the developed helical peptide database

Pattern	Query	# of Peptides Found	# of Unique Peptides Found
K *** G ** Y	K 3 G 2 Y	398	69
A *** K *** G	A 3 K 3 G	2981	391
G ** Y *** A	G 2 Y 3 A	2079	225
A *** K *** G ** Y	A 3 K 3 G 2 Y	15	3
K *** G ** Y *** A	K 3 G 2 Y 3 A	3	2
A *** K *** G ** Y *** A	A 3 K 3 G 2 Y 3 A	0	0
A**/***K**/***G**/***Y**/***A	A 2,3 K 2,3 G 2,3 Y 2,3 A	34	3

Table S2. 69 unique helical sequences obtained from the TP-DB matching the pattern K*G**Y given the query of [K 3 G 2 Y]**

#	U#	Matched Sequence (MS)	Matched Pattern	Full Helix (FH)	PDB ID: Chain	Positions in PDB (MS), (FH)	Helical Propensity	Contact	Interacting Partners
1	1	KAAEGLKY	K3G2Y	GESLKKAAEGLKYEDFAK	1exe:A	(87, 94), (82, 99)	0.831	10.250	
2	2	KAAQGAMY	K3G2Y	SRDLRLSLAKAITQRVILKAAQGAMY	1t3q:C	(278, 285), (260, 285)	0.829	8.750	
4	3	KNAWGEAY	K3G2Y	SPEMKNAWGEAYDQLVAAIKS	3qqq:B	(139, 146), (135, 155)	0.745	9.750	
6	4	KERLGEAY	K3G2Y	QESAVANTKERLGEAYQA	4poo:A	(62, 69), (54, 71)	0.744	7.375	
7	5	KVAWGQAY	K3G2Y	SPEMKVAWGQAYDHLVAAIKAEMNLS	3zhw:B	(138, 145), (134, 159)	0.624	9.625	
9	6	KLQLGQLY	K3G2Y	TDTEVERAKSLLKLQLGQLYE	3cxh:A	(371, 378), (359, 379)	0.583	8.875	
20	7	KQKLGKAY	K3G2Y	KQKLGKAYFQVQKIEAELYQLIKVSH	1skv:D	(39, 46), (39, 64)	0.564	7.500	
24	8	KEQLGEFY	K3G2Y	TKEQLGEFYEALDCL	3apw:B	(135, 142), (134, 148)	0.522	8.125	
31	9	KRKMGEAY	K3G2Y	LHDSASKECKRKMGEAYALNKKM	3fy4:C	(490, 497), (481, 503)	0.519	8.875	
34	10	KMYLGYEY	K3G2Y	DAKMYLGYEYVTAIRNLRE	3fgs:A	(312, 319), (310, 328)	0.502	7.375	
84	11	KTLAGQLY	K3G2Y	NKTLAGQLYSELKEFF	1d9z:A	(67, 74), (66, 81)	0.464	9.000	
94	12	KQQMGKEY	K3G2Y	KQQMGKEYREKIEAELQDICNDVLELLDKYLIPNA	2bq0:B	(77, 84), (77, 111)	0.448	8.375	
97	13	KLMKGQLY	K3G2Y	CRFVKLMKGQLYIDTVAA	1q4q:H	(294, 301), (290, 307)	0.443	7.375	
109	14	KERRGEMY	K3G2Y	GNLKERRGEMYHYFYDQLLTRYYFERLTN	3gwj:D	(243, 250), (240, 268)	0.432	9.625	3gwj:D (347, 358) 3gwj:D (359, 363) 3gwj:D (374, 385) 3gwj:D (409, 428) 3gwj:D (533, 535) 3gwj:D (595, 597)
112	15	KNAWGVAY	K3G2Y	SPEMKNAWGVAYDQLVAAIKFE	3qqr:B	(139, 146), (135, 156)	0.422	9.500	
114	16	KMLLGATY	K3G2Y	KMLLGATYFLPIYATVFIVGGFWEVLFCMVR	4p6v:B	(121, 128), (121, 151)	0.404	11.375	
115	17	KASAGKLY	K3G2Y	SKASAGKLYEAG	1p0n:B	(192, 199), (191, 202)	0.393	8.625	1p0n:B (237, 249) 1p0n:B (329, 339)
119	18	KSEWGRAY	K3G2Y	GVLEGLKVKSEWGRAYG	1x9f:J	(13, 20), (5, 21)	0.381	8.625	1x9f:J (58, 78) 1x9f:J (102, 122) 1x9f:J (126, 143) 1x9f:K (27, 47)
125	19	KHEKGREY	K3G2Y	LEEILEKHEKGREYG	2cxi:A	(182, 189), (176, 190)	0.355	8.250	
128	20	KEFLGDEY	K3G2Y	YKEFLGDEYYASVASLNTC	1aov:A	(654, 661), (653, 671)	0.337	8.125	
129	21	KALAGVTY	K3G2Y	RSEEMHRLTENVYKTIMEQFNPSLRNFIAMGKNYEKALAGVTYAAKGYFDALVKMGELASE	2ykt:A	(40, 47), (5, 65)	0.328	8.000	2ykt:A (159, 229)
132	22	KHILGELY	K3G2Y	NDKQKKHILGELYLFLN	3nzp:B	(590, 597), (585, 601)	0.304	8.375	
134	23	KVMQGYAY	K3G2Y	KVMQGYAYILTH	1rp9:A	(305, 312), (305, 316)	0.295	8.875	1rp9:A (332, 346)
149	24	KAFVGEYY	K3G2Y	PERTLDLAMQRLKAFVGEYY	3piu:A	(426, 433), (414, 433)	0.288	7.250	
153	25	KQACGFEY	K3G2Y	SDDAEASMISKLKQACGFEYTSKLQRMFQDIGVSKDLNEQFKKHLT	4f52:A	(493, 500), (481, 526)	0.271	7.250	
158	26	KDKLGAIY	K3G2Y	KDKLGAIYYRANT	1npc:A	(261, 268), (261, 273)	0.263	8.625	1npc:A (302, 313)
160	27	KALHGMHY	K3G2Y	NKEWGLRQRDKALHGMHYFDTVLRER	3erg:B	(158, 165), (148, 173)	0.252	9.375	3erg:B (206, 217)
164	28	KLIEGYTY	K3G2Y	DPOAAEKLNKLIEGYTY	2mln:B	(57, 64), (48, 64)	0.196	7.375	

:

347	56	KTIHGLIY	K3G2Y	KTIHGLIYNALKLFME	2iae:B	(384, 391), (384, 399)	-0.200	8.750	
349	57	KWGMGQHY	K3G2Y	KWGMGQHYELGSYIRRRYGRFL	1rpt:A	(40, 47), (40, 61)	-0.215	10.125	1rpt:A (78, 91)
351	58	KKIVGESY	K3G2Y	LDSERDKARKEVEEYVKKIVGESYAKS	2m6u:A	(17, 24), (1, 27)	-0.241	8.250	2m6u:A (58, 81)
352	59	KDILGHVY	K3G2Y	SKDILGHVYEYFLGQFALAEG	3khk:B	(201, 208), (200, 220)	-0.243	8.500	3khk:B (255, 267) 3khk:B (291, 305)
353	60	KSCYGKKY	K3G2Y	KSCYGKKYG	1b8t:A	(59, 66), (59, 67)	-0.252	7.625	
354	61	KFDKGYSY	K3G2Y	DPDKFDKGYSYNIRHSF	3q36:B	(340, 347), (337, 353)	-0.279	13.250	
356	62	KGTAGYIY	K3G2Y	SWDDWDYAAKVTLANSQKGTAGYIYRFLHDVSEG	2wxt:A	(230, 237), (213, 246)	-0.354	9.500	
368	63	KDIIGICY	K3G2Y	LRWLAKSFFELTVLKDIIGICY	3phf:U	(329, 336), (315, 336)	-0.462	61.750	3phf:U (368, 383) 3phf:U (415, 434)
384	64	KKPKGRAY	K3G2Y	TEKPKKPKGRAYKRL	3izb:Z	(28, 35), (24, 38)	-0.492		
385	65	KASLGTGY	K3G2Y	DTAEALATAKASLGTGYKKVSKSYTGEWKKYLNSL	1ulv:A	(246, 253), (237, 271)	-0.496	8.750	1ulv:A (330, 345) 1ulv:A (346, 361)
387	66	KPEKGVQY	K3G2Y	DVIRKRHYRIGLNLFNKKPEKGVQYLIER	4c0a:F	(416, 423), (399, 427)	-0.676	9.250	4c0a:F (448, 457) 4c0a:F (461, 474) 4c0a:F (595, 607) 4c0a:F (724, 757)
391	67	KTYLGPQY	K3G2Y	TYKTYLGPQYLTLMDN	3mc2:D	(650, 657), (648, 663)	-0.677	7.625	
395	68	KPKRGIQY	K3G2Y	QQKEIIEQGIDLFNKKPKRGIQYLQEQ	3ltl:A	(712, 719), (697, 723)	-0.751	9.375	3ltl:A (744, 753) 3ltl:B (700, 723)
397	69	KGTSGYIY	K3G2Y	SWDEWDYAAKVALANSQKGTSGYIYRFLHDVSD	1kho:B	(230, 237), (213, 245)	-0.849	9.625	

The result page is available online at https://dyn.life.nthu.edu.tw/design/result?JobID=602bae01v

Table S3. 390 unique helical sequences obtained from the TP-DB for the pattern A*K***G given the query of [A 3 K 3 G]**

#	U#	Matched Sequence (MS)	Matched Pattern	Full Helix (FH)	PDB ID: Chain	Positions in PDB (MS), (FH)	Helical Propensity	Contact	Interacting Partners
1	1	AELAKAEAG	A3K3G	DHAAELAKAEAGVAEMAAKR	3pij:B	(19, 27), (16, 35)	1.389	8.889	3pij:B (326, 331)
5	2	AYLAKEAWG	A3K3G	DSDAIGSSVAFAYLAKEAWG	2enx:A	(24, 32), (13, 32)	1.264	8.778	2enx:A (44, 54) 2enx:A (121, 132) 2enx:A (137, 153)
6	3	ALAKKAAEG	A3K3G	GPSGGAAVKALAKKAAEG	3vsd:A	(348, 356), (339, 356)	1.264	9.444	3vsd:A (370, 373) 3vsd:A (374, 382)
10	4	AAAAKAMKG	A3K3G	GEHVRDALLFAALAAAAKAMKG	1ua4:A	(368, 376), (355, 376)	1.221	10.556	1ua4:A (388, 390) 1ua4:A (437, 453)
14	5	AAALKWLKG	A3K3G	SGAFLEQIVAIVKLKGEAAALKWLKGLKEY G	3tyh:F	(156, 164), (139, 169)	1.192	9.444	3tyh:F (220, 223) 3tyh:F (292, 306)
51	6	ARAEKKAAG	A3K3G	KQEKKQRLLARAEKKAAG	3j3b:G	(116, 124), (107, 124)	1.153		
52	7	ALAEKMALG	A3K3G	APQLSKIMWALAEKMALG	3ql2:D	(210, 218), (201, 218)	1.143	7.778	
54	8	ALARKALEG	A3K3G	ALARKALEG	2dsj:B	(287, 295), (287, 295)	1.132	8.889	
56	9	AAFAKQAEG	A3K3G	AAFAKQAEGT	2lsh:A	(16, 24), (16, 25)	1.113	9.111	2lsh:A (70, 74)
57	10	AEAAKNALG	A3K3G	SYKVIVAEAAKNALG	3gyx:G	(175, 183), (169, 183)	1.103	9.667	3gyx:G (470, 493)
63	11	AAREKIAAG	A3K3G	SVIAAREKIAAG	1f76:D	(303, 311), (300, 311)	1.036	54.889	
65	12	AYEMKEALG	A3K3G	GGMPAAYEMKEALG	3t31:A	(17, 25), (12, 25)	1.032	9.556	3t31:A (62, 68) 3t31:A (321, 340) 3t31:A (385, 405)
81	13	AKLAKEQLG	A3K3G	GNAGYYTAKLAKEQLG	1euz:F	(228, 236), (221, 236)	1.020	9.444	1euz:F (256, 268)
83	14	AALDKELAG	A3K3G	SDEAKAAAAESVKNHLAALDKELAG	1f2e:D	(134, 142), (118, 142)	1.018	8.444	1f2e:D (175, 187)
85	15	AAAIKAFAG	A3K3G	EAAAIKAFAGEFD	3exz:E	(36, 44), (35, 47)	1.013	9.000	
90	16	ADAEKAYEG	A3K3G	GADAEKAYEGFLAFKDVVKKS	1ppr:O	(130, 138), (129, 149)	1.010	9.111	
98	17	AQLEKELQG	A3K3G	MEELQAQLEKELQGG	3rde:B	(263, 271), (258, 272)	1.003	8.222	
102	18	AEELKALIG	A3K3G	EAEELKALIGS	4wsr:F	(98, 106), (97, 107)	0.987	8.667	
114	19	AAFRKAEEG	A3K3G	DLVAAFRKAEEG	3p0b:A	(126, 134), (123, 134)	0.985	8.667	3p0b:A (154, 174)
116	20	ADELKEEWG	A3K3G	EIPDIMFKLKTQPNKIADELKEEWG	3t7z:A	(52, 60), (36, 60)	0.979	7.889	
117	21	ALRLKEELG	A3K3G	RRMLAAREEALRLKEELG	2efy:B	(127, 135), (118, 135)	0.972	8.444	
122	22	AEAAKMAVG	A3K3G	SYKPIIAEAAKMAVG	2fje:C	(2158, 2166), (2152, 2166)	0.967	10.222	2fje:C (2447, 2471)

:

2765	363	ANVCKKITG	A3K3G	LAANVCKKITGRLTSAIA	1u6g:C	(128, 136), (126, 143)	-0.373	8.333	
2768	364	APNLKALVG	A3K3G	DQTDIEVVRQLAEEIDRLYQLRKKLEDYID RAMDDVAPNLKALVGAKLAARLISLAG	3nvi:A	(257, 265), (221, 277)	-0.409	9.889	3nvi:A (313, 317) 3nvi:A (318, 323) 3nvi:A (328, 350) 3nvi:A (354, 372) 3nvi:C (130, 161) 3nvi:C (221, 277)
2780	365	AEHGKTVMG	A3K3G	NAKVAEHGKTVMGGLDRAVQ	1spg:B	(61, 69), (57, 76)	-0.414	9.667	1spg:B (99, 117) 1spg:B (124, 144)
2781	366	AFSVKEISG	A3K3G	LRLAFSVKEISG	4fiq:E	(203, 211), (200, 211)	-0.415	9.000	4fiq:E (228, 246)
2783	367	ASGKKFIEG	A3K3G	SRVIASGKKFIEGFQSTKL	1ihp:A	(145, 153), (141, 159)	-0.431	11.000	
2788	368	ANSQKGTAG	A3K3G	SWDDWDYAAKVTLANSQKGTAGYIYRFLHD VSE	1qmd:B	(226, 234), (213, 245)	-0.447	10.556	
2800	369	AFLPKLASG	A3K3G	SEAQKEAFLPKLASG	2cx9:D	(111, 119), (105, 119)	-0.474	10.444	
2801	370	AHGIKVLHG	A3K3G	NANVAAHGIKVLHG	2h8f:D	(62, 70), (57, 70)	-0.502	8.778	2h8f:D (100, 119)
2829	371	APAYKKDTG	A3K3G	TSTDNTGLLDVLAPAYKKDTG	3kn3:C	(22, 30), (10, 30)	-0.509	8.111	3kn3:C (121, 136) 3kn3:C (214, 226)
2832	372	APYLKVFLG	A3K3G	LIAPYLKVFLGFI	1t5b:B	(154, 162), (152, 164)	-0.514	8.889	
2834	373	APTVKAFAG	A3K3G	APTVKAFAGLLREG	1wlo:A	(72, 80), (72, 85)	-0.518	9.444	1wlo:A (107, 135)
2835	374	AGVVKFKKG	A3K3G	DAGVVKFKKG	1lrz:A	(377, 385), (376, 385)	-0.521	8.667	
2836	375	APFDKYLRG	A3K3G	APFDKYLRG	1rz5:A	(168, 176), (168, 176)	-0.571	9.778	1rz5:A (285, 298)
2837	376	AAHGKVVCG	A3K3G	PKVAAHGKVVCGALDKAVK	1ouu:B	(61, 69), (58, 76)	-0.582	9.444	1ouu:B (124, 144)
2840	377	AHGVKVLHG	A3K3G	TPDAIKGNAKIAAHGVKVLHG	1v4w:D	(62, 70), (50, 70)	-0.583	9.000	1v4w:D (100, 119)
2844	378	APILKLIHG	A3K3G	SAPILKLIHGDLENQFNE	3qhq:A	(74, 82), (73, 90)	-0.594	8.111	3qhq:A (126, 135) 3qhq:A (144, 159)
2846	379	AVFEKMMPG	A3K3G	PKLITDGLAVFEKMMPGYVSVLESNLT	2a0b:A	(685, 693), (677, 703)	-0.615	7.667	2a0b:A (744, 773)
2850	380	AFNAKPKNG	A3K3G	HPEIEKAQREIIEAFNAKPKNGINKIKEIC EQY	1xt0:B	(15, 23), (2, 34)	-0.644	9.333	1xt0:B (50, 53) 1xt0:B (54, 63) 1xt0:B (66, 79)
2855	381	AHHGKVVMG	A3K3G	NPKVAHHGKVVMGGLERAIKN	3bcq:D	(61, 69), (57, 77)	-0.682	9.778	3bcq:D (100, 119) 3bcq:D (124, 143)
2857	382	AHGKKVAGG	A3K3G	SAQVKAHGKKVAGGLSEAANH	2zfb:A	(57, 65), (52, 72)	-0.683	9.556	2zfb:A (95, 113) 2zfb:A (118, 136)
2858	383	AILGKVLGG	A3K3G	PVVLVAGTVVTVIAILGKVLGGFLGA	4bwz:A	(301, 309), (288, 313)	-0.695	11.778	4bwz:A (331, 345) 4bwz:A (349, 384)
2859	384	APYLKTRVG	A3K3G	AAPYLKTRVG	4efz:A	(77, 85), (76, 85)	-0.695	18.222	
2861	385	ADPVKAWSG	A3K3G	DEFLKKNPEKVRKFLKAIKKATDYVLADPVKAWSGYIDFK	4h67:H	(233, 241), (207, 246)	-0.710	7.667	
2871	386	ASVGKMIDG	A3K3G	ATEIRASVGKMIDGIGRFYIQMCTEL	4dyt:C	(27, 35), (22, 47)	-0.806	10.444	4dyt:C (112, 122) 4dyt:C (277, 285)
2933	387	AHGKKVMGG	A3K3G	SPNIKAHGKKVMGGIALAVSK	3ng6:C	(58, 66), (53, 73)	-0.884	9.222	3ng6:C (119, 138)
2966	388	ANTGKSCKG	A3K3G	DQNNIKQAFANTGKSCKGCHDVYK	4ulv:B	(112, 120), (103, 126)	-0.898	8.667	
2970	389	AHSIKGGAG	A3K3G	AEQLNAIFRAAHSIKGGAGTF	1i5n:B	(47, 55), (37, 57)	-0.915	9.667	
2975	390	ANSQKGTSG	A3K3G	SWDEWDYAAKVALANSQKGTSGYIYRFLHD VSD	1kho:A	(226, 234), (213, 245)	-0.943	10.444	

The result page is available online at https://dyn.life.nthu.edu.tw/design/result?JobID=602bae5cu

Table S4. 224 unique helical sequences obtained from the TP-DB for the pattern GY***A given the query of [G 2 Y 3 A]**

#	U#	Matched Sequence (MS)	Matched Pattern	Full Helix (FH)	PDB ID: Chain	Positions in PDB (MS), (FH)	Helical Propensity	Contact	Interacting Partners
1	1	GAAYAEIA	G2Y3A	PGAAYAEIALAAATDTF	4p7p:A	(969, 976), (968, 984)	0.998	9.250	4p7p:A (1117, 1121) 4p7p:A (1124, 1134)
2	2	GEQYAELA	G2Y3A	TFIAAQFVLYGEQYAELAARR	3s4c:A	(529, 536), (519, 539)	0.997	10.000	3s4c:A (592, 601) 3s4c:A (666, 678) 3s4c:A (732, 744)
12	3	GANYAAWA	G2Y3A	GANYAAWAVNVAQV	1mdt:B	(275, 282), (275, 288)	0.985	10.000	1mdt:B (310, 314) 1mdt:B (326, 347) 1mdt:B (358, 376)
24	4	GARYEALA	G2Y3A	AGARYEALATEIDRGLRFMSACG	2ypq:A	(211, 218), (210, 232)	0.983	7.625	
39	5	GAEYQRAA	G2Y3A	SAGAEYQRAALISALQTLY	3c0k:A	(132, 139), (130, 148)	0.944	9.000	3c0k:A (160, 165)
41	6	GLLYRAAA	G2Y3A	SGLLYRAAAFLALRAG	3w90:A	(34, 41), (33, 48)	0.939	9.625	3w90:A (85, 88) 3w90:A (90, 102) 3w90:A (103, 115)
46	7	GMAYLEEA	G2Y3A	AAETLLGMCLDVCEGMAYLEEA	4rfm:A	(469, 476), (455, 476)	0.937	9.125	4rfm:A (536, 553) 4rfm:A (604, 618)
102	8	GQAYKLWA	G2Y3A	TKEDLKRRGQAYKLWADQN	2yym:A	(84, 91), (76, 94)	0.902	10.250	2yym:A (267, 310) 2yym:A (315, 342)
108	9	GAVYLAAA	G2Y3A	GAVYLAAALAAHSQ	1fui:C	(117, 124), (117, 130)	0.873	11.625	1fui:C (565, 570)
114	10	GEAYQAIA	G2Y3A	TPEIINAWGEAYQAIADIFITVEKKMYEEA	4glv:A	(123, 130), (115, 144)	0.857	8.375	
119	11	GLAYMAMA	G2Y3A	PQYEYLVAMFIPIWSGLAYMAMAI	4tl3:A	(48, 55), (33, 56)	0.850	10.000	4tl3:A (69, 92) 4tl3:A (194, 226)
125	12	GEEYFALA	G2Y3A	APDGEEYFALALFFASHRWG	3a3v:A	(129, 136), (126, 145)	0.846	9.875	3a3v:A (195, 199) 3a3v:A (200, 211) 3a3v:A (265, 279) 3a3v:A (357, 373)
133	13	GAEYATLA	G2Y3A	VFIAAQFVLYGAEYATLAERR	2cqt:B	(529, 536), (519, 539)	0.833	9.750	2cqt:B (592, 601) 2cqt:B (667, 679) 2cqt:B (732, 744)
149	14	GLEYLREA	G2Y3A	GEKGLEYLREAADKY	3pg8:B	(146, 153), (143, 157)	0.824	9.625	3pg8:B (170, 178)
165	15	GAVYAELA	G2Y3A	DNYERDAELQQRTGAVYAELAAQG	1mrs:A	(176, 183), (163, 186)	0.816	9.500	
188	16	GAHYYWEA	G2Y3A	GAHYYWEA	4tkx:L	(577, 584), (577, 584)	0.807	8.625	
190	17	GAQYIQAA	G2Y3A	GAQYIQAAGVALGLKMR	3duf:G	(145, 152), (145, 161)	0.772	12.125	3duf:G (178, 191) 3duf:H (58, 72) 3duf:H (89, 95)
210	18	GELYLWIA	G2Y3A	YGELYLWIAKALDEADQN	3fdu:F	(37, 44), (36, 53)	0.768	9.125	3fdu:F (92, 104) 3fdu:F (118, 123)
216	19	GWCYYAAA	G2Y3A	GGWCYYAAA	3eve:A	(86, 93), (85, 93)	0.763	10.750	3eve:A (120, 122)
222	20	GYQYQLEA	G2Y3A	PGYQYQLEAVFNAKRN	3vsv:D	(285, 292), (284, 299)	0.757	9.875	
						:			
1879	194	GPYYQRYA	G2Y3A	GPYYQRYALMPFIWFAKAIETNE	4ok4:A	(255, 262), (255, 277)	-0.397	10.000	
1883	195	GPRYRQLA	G2Y3A	GPRYRQLARNWVPTAGLWGAVGAVGLVWATD	1sqp:K	(7, 14), (7, 37)	-0.401	7.875	
1895	196	GKGYSRKA	G2Y3A	WGKGYSRKAAALEFL	1e1w:A	(72, 79), (71, 85)	-0.426	28.125	1e1w:A (105, 118)
1897	197	GHLYGHVA	G2Y3A	ARHMRELLTQMGHLYGHVADELA	1brm:C	(180, 187), (169, 191)	-0.453	8.500	
1903	198	GFGYALGA	G2Y3A	AGFGYALGAAADMIRG	2aj9:B	(124, 131), (123, 138)	-0.457	10.250	
1927	199	GQPYNQLA	G2Y3A	GQPYNQLAILASSK	1ya0:A	(186, 193), (186, 199)	-0.461	11.375	1ya0:A (219, 234)
1929	200	GGIYTLNA	G2Y3A	GGIYTLNAARSFQTFVREV	4b46:A	(143, 150), (143, 161)	-0.462	9.625	4b46:A (352, 367)
1930	201	GGLYHTNA	G2Y3A	HGGLYHTNAEVKL	3t6g:A	(633, 640), (632, 644)	-0.471	9.875	3t6g:C (543, 574)
1932	202	GSTYTMCA	G2Y3A	GSTYTMCAQ	3gyx:G	(258, 265), (258, 266)	-0.478	11.000	3gyx:G (587, 614)
1938	203	GSAYSSVA	G2Y3A	GSAYSSVAPQVLCFLQ	2vn4:A	(210, 217), (210, 225)	-0.510	9.375	2vn4:A (443, 448)
1940	204	GSEYGSQA	G2Y3A	YGSEYGSQALAFVKSK	2axe:A	(191, 198), (190, 205)	-0.543	8.000	
1943	205	GGPYALAA	G2Y3A	GGPYALAAAKALLRNT	1m4y:C	(125, 132), (125, 140)	-0.543	10.250	
1946	206	GGVYGLAA	G2Y3A	GGVYGLAA	1kb0:A	(542, 549), (542, 549)	-0.544	11.375	1kb0:A (592, 609)
1947	207	GPYYFYRA	G2Y3A	GPYYFYRAAVRLGVAAFVFSEAARRDRRASA	2xqy:E	(262, 269), (262, 292)	-0.548	9.125	
1949	208	GMGYAIGA	G2Y3A	IGMGYAIGASVTSG	2q28:B	(426, 433), (425, 438)	-0.563	11.625	2q28:B (453, 455) 2q28:B (456, 463) 2q28:B (518, 533)
1955	209	GSFYGRRA	G2Y3A	GTSLGDGSFYGRRAGISAAK	4at2:A	(475, 482), (469, 488)	-0.567	12.500	
1957	210	GPFYATKA	G2Y3A	YGPFYATKAAIEKAG	4n45:B	(291, 298), (290, 304)	-0.571	11.000	4n45:B (319, 331)
1961	211	GGPYAQAA	G2Y3A	GGPYAQAAARALLENT	1g4b:N	(125, 132), (125, 140)	-0.583	10.250	
2005	212	GDVYTSVA	G2Y3A	TVGDVYTSVAV	2cgq:A	(63, 70), (61, 71)	-0.589	8.250	
2006	213	GQIYGGLA	G2Y3A	NKLVVDGQIYGGLAQGVGLALS	1dgj:A	(806, 813), (800, 821)	-0.605	10.750	1dgj:A (840, 844) 1dgj:A (869, 873) 1dgj:A (874, 888) 1dgj:A (897, 904)
2018	214	GSGYRFIA	G2Y3A	GSGYRFIADKVIEIDR	4pw4:A	(799, 806), (799, 814)	-0.654	9.375	4pw4:A (826, 828) 4pw4:A (834, 850) 4pw4:A (855, 867)
2027	215	GVAYSSGA	G2Y3A	YAYSGVAYSSGALN	2yof:A	(104, 111), (100, 113)	-0.724	10.750	2yof:A (155, 167) 2yof:A (168, 170)
2045	216	GPYYIDRA	G2Y3A	WGPYYIDRAKAVLD	3s99:A	(250, 257), (249, 262)	-0.725	8.500	
2046	217	GGSYGWQA	G2Y3A	GGSYGWQAAE	1s20:F	(120, 127), (120, 129)	-0.726	10.000	1s20:F (304, 317)
2056	218	GKGYGQTA	G2Y3A	GKGYGQTASGRSRGAVAAAEK	3iib:A	(167, 174), (167, 187)	-0.750	8.750	3iib:A (419, 423)
2057	219	GQRYPVIA	G2Y3A	GWGQRYPVIAQMVQYH	4clr:C	(49, 56), (47, 62)	-0.766	17.125	4clr:C (73, 84)
2061	220	GQPYQEGA	G2Y3A	LGQPYQEGAG	4q6w:B	(264, 271), (263, 272)	-0.782	9.000	
2065	221	GKGYSLGA	G2Y3A	TSDVGKGYSLGAELAFKYFNEKG	3eaf:A	(62, 69), (58, 80)	-0.841	11.250	3eaf:A (273, 276) 3eaf:A (279, 285) 3eaf:A (323, 327) 3eaf:A (328, 346)
2066	222	GPCYSTKA	G2Y3A	GIGPCYSTKAS	1p9b:A	(146, 153), (144, 154)	-1.183	9.375	
2067	223	GHPYYSSA	G2Y3A	GVEEEAKIKAPILEQYEHQGHPYYSSAR	3u9r:B	(517, 524), (498, 525)	-1.186	11.000	
2068	224	GPTYSSKA	G2Y3A	GIGPTYSSKAA	1j4b:A	(168, 175), (166, 176)	-1.268	8.875	

The result page is available online at https://dyn.life.nthu.edu.tw/design/result?JobID=602bb16dr

Table S5. Results obtained from the TP-DB for pattern A*K***G**Y with query [A 3 K 3 G 2 Y]**

#	U#	Matched Sequence (MS)	Matched Pattern	Full Helix (FH)	PDB ID: Chain	Positions in PDB (MS), (FH)	Helical Propensity	Contact	Interacting Partners
1	1	ANKLKQAVGDIY	A3K3G2Y	DANKLKQAVGDIYN	4nzt:M	(366, 377), (365, 378)	0.507	8.667	4nzt:M (406, 420)
2	2	ANSQKGTAGYIY	A3K3G2Y	SWDDWDYAAKVTLANSQKGTAGYIYRFLHD VSE	1qmd:B	(226, 237), (213, 245)	-0.225	10.000	
14	3	ANSQKGTSGYIY	A3K3G2Y	SWDEWDYAAKVALANSQKGTSGYIYRFLHD VSD	1kho:A	(226, 237), (213, 245)	-0.720	10.000	

The result page is available online at https://dyn.life.nthu.edu.tw/design/result?JobID=602bb1e0r

Table S6. Results obtained from the TP-DB for pattern K*G**Y***A with query [K 3 G 2 Y 3 A]**

#	U#	Matched Sequence (MS)	Matched Pattern	Full Helix (FH)	PDB ID: Chain	Positions in PDB (MS), (FH)	Helical Propensity	Contact	Interacting Partners
1	1	KAAEGLKYEDFA	K3G2Y3A	GESLKKAAEGLKYEDFAK	1exe:A	(87, 98), (82, 99)	1.264	10.000	
2	2	KKEIGRSYRFIA	K3G2Y3A	KKEIGRSYRFIARN	1ais:B	(1183, 1194), (1183, 1196)	0.230	9.167	

The result page is available online at https://dyn.life.nthu.edu.tw/design/result?JobID=602bb2e9z

Table S7. Results obtained from the TP-DB for pattern A*K***G**Y***A with query [A 3 K 3 G 2 Y 3 A]**

#	U#	Matched Sequence (MS)	Matched Pattern	Full Helix (FH)	PDB ID: Chain	Positions in PDB (MS), (FH)	Helical Propensity	Contact	Interacting Partners

No peptide found
The result page is available online at https://dyn.life.nthu.edu.tw/design/result?JobID=602bb312l

Table S8. Results obtained from the TP-DB for pattern A/***K**/***G**/***Y**/***A with query [A 2,3 K 2,3 G 2,3 Y 2,3 A]**

#	U#	Matched Sequence (MS)	Matched Pattern	Full Helix (FH)	PDB ID: Chain	Positions in PDB (MS), (FH)	Helical Propensity↕	Contact↕	Interacting Partners
1	1	ADEKKFWGKYL	A3K2G2A	EPDTLCAEFKADEKKFWGKYLYEVARRH	4luf:A	(128, 138), (118, 145)	0.835	8.182	4luf:A (172, 222)
21	2	AYAKKVEGDMY	A3K2G2A	DRRMENLVAYAKKVEGDMYESAN	2lxt:A	(630, 640), (622, 644)	0.703	8.727	2lxt:B (846, 856)
31	3	ARFRKDRGALY	A3K2G2A	DLAAEARFRKDRGALYREAE	1wiy:B	(181, 191), (176, 195)	0.646	9.364	1wiy:B (207, 238)

The result page is available online at https://dyn.life.nthu.edu.tw/design/result?JobID=602bb335c

We selected the helical PP2A-like peptides found from queries [A 3 K 3 G 2 Y] (**Table S5**), [K 3 G 2 Y 3 A] (**Table S6**), [A 2,3 K 2,3 G 2,3 Y 2,3 A] (**Table S8**), as well as the wide-type PP2A helix as control, for further investigations through all-atom MD-based binding free energy change (ΔG) evaluation by MM/PBSA (see Supporting Methods). The results (**Table S9**) showed that seven of the newly found helixes, with a binding ΔΔG < 0.0 kcal/mol, bound Sgo1 stronger than the wide-type PP2A did, suggesting a potential PPI blocking function for these peptides. Among them, the number one peptide "**A**YA**K**KVE**G**DM**Y**", with a binding affinity ~10 kcal/mol stronger than the wild-type peptide, can be a promising peptide subject to further experimental validation by isothermal calorimetry (ITC) assays and/ or NMR HSQC spectra. Peptide design is an elaborated miniature of protein design, and we expect that herein proposed methodologies can be applied for the design of therapeutic α-helical peptides for other medicinal purposes.

Table S9. Seven of the newly found helixes, with binding ΔΔG less than 0.0 kcal/mol, have the potentials of binding stronger to SGO1's helix than the control

Candidate Helix (Matched Pattern)	Helix Source, PDB ID: Chain	Helical Propensity Score	Contact Number	Binding ΔΔG (kcal/mol)	In which Table above* (U#)
AYA**K**KVE**G**DM**Y** **(A 2,3 K 2,3 G 2,3 Y)**	2lxt:A	0.703	8.545	-10.11 ± 0.43	**Table S8** (2')
KKEI**G**RS**Y**RFI**A** **(K 3 G 2 Y 3 A)**	1ais:B	0.230	9.167	-9.20 ± 0.32	**Table S6** (2)
KAAE**G**LK**Y**EDF**A** **(K 3 G 2 Y 3 A)**	1exe:A	1.264	10.000	-8.04 ± 0.29	**Table S6** (1)
AYA**K**KVE**G**DM**Y**ES**A** **(A 2,3 K 2,3 G 2,3 Y 2,3 A)**	2lxt:A	0.703	8.727	-6.64 ± 0.45	**Table S8** (2)
A**N**SQ**K**GTA**G**YI**Y** **(A 3 K 3 G 2 Y)**	1qmd:B	-0.225	10.000	-5.63 ± 0.33	**Table S5** (2)
ARFR**K**DR**G**AL**Y**RE**A** **(A 2,3 K 2,3 G 2,3 Y 2,3 A)**	1wiy:B	0.646	9.364	-1.96 ± 0.46	**Table S8** (3)
ARFR**K**DR**G**AL**Y** **(A 2,3 K 2,3 G 2,3 Y)**	1wiy:B	0.957	8.500	-1.14 ± 0.38	**Table S8** (3')
KTIH**G**LI**Y** **Control, from PP2A**	3fga:B	-0.200	9.250	0.00	
A**N**SQ**K**GTS**G**YI**Y** **(A 3 K 3 G 2 Y)**	1kho:A	-0.720	10.000	1.38 ± 0.39	**Table S5** (3)
ADEK**K**FW**G**KYL**Y** **(A 2,3 K 2,3 G 2,3 Y)**	4luf:A	0.957	8.500	4.31 ± 0.29	**Table S8** (1')
ADEK**K**FW**G**KYL**Y**EV**A** **(A 2,3 K 2,3 G 2,3 Y 2,3 A)**	4luf:A	0.835	8.182	4.66 ± 0.86	**Table S8** (1)
ANKL**K**QAV**G**DI**Y** **(A 3 K 3 G 2 Y)**	4nzt:M	0.507	8.667	5.11 ± 0.32	**Table S5** (1)

*The earlier Tables that listed the matched sequences are denoted at the rightmost column. The number in the parenthesis is the serial number for the unique (U#) sequence in the corresponding Table of interest. A prime following a U#, such as 2', indicates that the sequence in that row of this table is a subsequence of the original sequence presented in the earlier Tables.

Supplementary Methods

Production of recombinant Helicobactor pylori neutrophil-activating protein (HP-NAP) and maltose-binding protein (MBP)

Recombinant *H. pylori* neutrophil-activating protein (HP-NAP) was expressed in *E. coli* BL21(DE3) cells harboring the expression plasmid pET42a-NAP and purified by either two consecutive gel-filtration chromatography as previously described[60] or a small-scale DEAE Sephadex negative mode batch chromatography as previously described[61]. Maltose-binding protein (MBP) was prepared the same as the procedure for production of MBP fused with the polypeptide containing residues Arg77 to Glu116 of HP-NAP as described below except that *E. coli* BL21(DE3) cells harboring the pMALc2g expression vector was used for expression.

Cloning of HP-NAP$_{R77-E116}$ into a MBP fusion protein expression vector

The plasmid DNA pET42a-NAP encoding a *napA* gene from *H. pylori* strain 26695 [GenBank:AE000543.1, Gene: HP0243] was prepared as previously described[60]. The DNA fragment coding for polypeptide containing residues Arg77 to Glu116 of HP-NAP (HP-NAP$_{R77-E116}$), which contains the D-Y-K-x-x-[DE] motif, was amplified by PCR from the plasmid pET42a-NAP using the forward and reverse primers containing BamHI and HindIII site, respectively. The forward primer is 5'-ATAAGGATCCCGTGTTAAAGAAGAAACTAAAAC-3' and the reversed primer is 5'-TTAATAAGCTTTAATTCTTTTTCAGCGGTGTTAGAG-3'. The PCR reaction was carried out with 10 ng plasmid DNA pET42a-NAP as a template and KAPA HiFi PCR Kit (Kapa Biosystems, Inc.) in a Mastercycler Gradient 5331 (Eppendorf, Germany). An initial denaturing phase of 95 °C for 5 min was followed by 39 cycles of 98 °C for 20 sec, 67 °C for 15 sec, and 72 °C for 15 sec. A final elongation phase of 72 °C for 2 min was also included. The amplified DNA fragments encoding HP-NAP$_{R77-E116}$ were then cloned into pJET1.2/blunt vectors using the CloneJET PCR Cloning Kit (Thermo Fisher Scientific Inc.). The resulting plasmid was designated as pJET1.2/blunt-HP-NAP$_{R77-E116}$. The insert was sequenced to confirm the correct DNA sequence. The correct insert was digested from pJET1.2/blunt-HP-NAP$_{R77-E116}$ with BamHI and HindIII and then cloned into the pMALc2g expression vector[62]. The resulting plasmid was designated as pMALc2g-HP-NAP$_{R77-E116}$.

Production of MBP-tagged HP-NAP$_{R77-E116}$

E. coli BL21(DE3) cells harboring pMALc2g-HP-NAP$_{R77-E116}$ were streaked on a lysogeny broth (LB) agar plate containing 100 μg/ml ampicillin and incubated at 37 °C for 16 hr. A single colony was picked and inoculated into 4 ml of LB containing 100 μg/ml ampicillin and the culture was incubated at 37 °C with shaking at 170 rpm for 16 hr. A volume of 2 ml of the overnight culture was inoculated into 200 ml LB containing 100 μg/ml ampicillin and the inoculated culture was incubated at 37 °C with shaking at 170 rpm for 2 hr until the OD_{600} reached 0.5. The expression of MBP-tagged HP-NA$_{PR77-E116}$ was induced by the addition of isopropyl β-D-1-thiogalactopyranoside (IPTG) to a final concentration of 0.3 mM and the culture was incubated at 37 °C with shaking at 180 rpm for 3 h until the OD_{600} reached 1.7. Then, the cells were centrifuged at 6,000 x g at 4 °C for 15 minutes to remove the supernatant and the cell pellets were stored at -70 °C.

The cell pellet from a 200 ml culture of *E. coli* expressing recombinant MBP-tagged HP-NAP$_{R77-E116}$ were re-suspended in 20 ml of ice-cold buffer containing 20 mM Tris-HCl, pH 7.4, 200 mM NaCl, 1 mM ethylenediaminetetraacetic acid (EDTA), and 1 mM dithiothreitol (DTT), plus 0.1% (v/v) protease inhibitor mixture (PI mix). The PI mix contained 0.13 M phenylmethylsulfonyl fluoride (PMSF), 0.03 M N-alpha-tosyl-L-lysyl-chloromethyl ketone (TLCK), and 0.03 M N-tosyl-L-phenylalanyl-chloromethyl ketone (TPCK). The bacterial suspensions were disrupted by Emulsiflex C3 high-pressure homogenizer (Avestin) operated at a range of 15,000-20,000 psi for 7 times at 4 °C. The lysates were centrifuged at 30,000 x g at 4 °C for 1 hr to separate insoluble and soluble proteins by using a Hitachi himac CP80WX ultracentrifuge (Hitachi Koki Co. Ltd., Tokyo, Japan). Then, 5 mL supernatant containing the soluble proteins were loaded onto a 1-ml MBPTrap HP column (GE Healthcare Bio-Sciences), which was pre-equilibrated with 20 mM Tris-HCl, pH 7.4, 200 mM NaCl, 1 mM EDTA, and 1 mM DTT, at a flow rate of 0.5 ml/min at 4 °C by ÄKTA Purifier. The column was eluted with 20 mM Tris-HCl, pH 7.4, 200 mM NaCl, 1 mM EDTA, 1 mM DTT, and 10 mM maltose at a flow rate of 1 ml/min at 4 °C by ÄKTA Purifier. The flow-through and elution fractions were analyzed by SDS-PAGE on a 12% gel. The elution fractions containing recombinant MBP-tagged HP-NAP$_{R77-E116}$ were collected and concentrated to a concentration higher than 1 mg/ml.

Recombinant HP-NAP-based enzyme linked immunosorbent assay (ELISA)

Nunc MaxiSorp ninety-six-well enzyme linked immunosorbent assay (ELISA) plates (Nunc, Rochester, NY, USA) were coated with 0.3 μg of recombinant HP-NAP in 100 μl of bicarbonate buffer (pH 9.0) for each well at room temperature for 16 hr. Each well was washed with 300 μl of phosphate buffered saline (PBS), pH 7.4, containing 20 mM Na_2HPO_4, 1.47 mM KH_2PO_4, 137 mM NaCl, and 2.7 mM KCl, with the addition of 0.1% tween-20 (PBS-T) three times for 10 min each time. The wells were blocked with 250 μl

PBS with 1% bovine serum albumin (BSA) for 2 hr and then washed with 300 µl of PBS-T buffer three times for 10 min each time. The anti-FLAG M2 antibody (Sigma-Aldrich, Cat# F-3165) and its corresponding mouse IgG antibody (Sigma-Aldrich, Cat# I5381) at a concentration of 600 ng/ml and the hybridoma culture supernatant containing mouse monoclonal antibody MAb 16F4[63] against HP-NAP at a dilution of 1:5000 in 100 µl of PBS-T buffer containing 1% BSA were added into each well. The plate was incubated at room temperature for 1 hr and then the wells were washed with 300 µl of PBS-T buffer three times for 10 min each time. The horseradish peroxidase-conjugated goat anti-mouse secondary antibody (Jackson ImmunoResearch) at a dilution of 1:10000 in 100 µl of PBS-T buffer containing 1% BSA was loaded into each well. The plate was incubated at room temperature for 1 hr and then the wells were washed with 300 µl of PBS-T buffer three times for 10 min each time. The color was developed using 3,3',5,5'-tetramethylbenzidine (TMB) peroxidase substrate (Thermo Scientific). The reaction was terminated by the addition of 2 N H_2SO_4, and the absorbance at 450 nm was measured by Bio-rad iMark microplate absorbance reader (Hercules, CA).

Western blot analysis

Western blotting was performed essentially the same as previously described[64]. The membrane was probed with either anti-FLAG M2 antibody (Sigma-Aldrich) at a concentration of 1 µg/ml or the hybridoma culture supernatant containing mouse monoclonal antibody MAb 16F4[63] against HP-NAP at a dilution of 1:2000.

Finding PP2A-like Peptides

Knowing that the K374, G378, and Y381 of PP2A plays important roles in PP2A's favourable interactions with Sgo1 (see **Fig S4**), we set out to search for a peptide with a similar amino acid pattern and which may interact better than with Sgo1 and thereby out compete/displace PP2A and bind to Sgo1. Therefore, we searched the developed TP-DB for the following queries: **[K 3 G 2 Y]**, **[A 3 K 3 G], [G 2 Y 3 A], [A 3 K 3 G 2 Y], [K 3 G 2 Y 3 A], [A 3 K 3 G 2 Y 3 A], and [A 2,3 K 2,3 G 2,3 Y 2,3 A].** The resulting peptides from each query are ranked based on their helical propensity score and contact number.

Minimization and MD simulations of Martini Coarse-grained (CG) models in vacuum and assessing the stability of complexes

To search for PP2A-like helical peptides to block the interaction between Sgo1 and PP2A, 69 sequences (https://dyn.life.nthu.edu.tw/design/result?JobID=602bae01v) that matched the pattern "K***G**Y" in PP2A helix (see **Fig S4**; the three anchoring residues of PP2A helix $\mathbf{K}_{374}\mathbf{T}_{375}\mathbf{I}_{376}\mathbf{H}_{377}\mathbf{G}_{378}\mathbf{L}_{379}\mathbf{I}_{380}\mathbf{Y}_{381}$ that interacts with Sgo1 are K374, G378 and Y381) were chosen for affinity assessment. Methodologically, it was equivalent to computationally modify (mutate) the five spacing residues T375, I376, H377, L379 and I380 in PP2A helix "KTIHGLIY" into corresponding residues in every of these 69 sequences. In other words, we only replaced the alpha-helical segment "K_{374}TIHGLIY_{381}" in PP2A [PDB:3FGA; chain B] by every of the 69 helical stretches that matched the pattern "K***G**Y" and the segment was shown to interact with the helix "V_{75}KEAQDIILQLRKECYYL_{92}" in Sgo1 [PDB:3FGA; chain D]. Subsequently, all 69 complexes comprising the Sgo1 helix and mutated PP2A helix were converted into coarse-grained (CG) models by the web service CHARMM-GUI[65]. Both terminals of the helical peptides in all complexes were set to be electrically neutral, and 69 CG models with PP2A mutants as well as the counterpart of wide-type were energy-minimized and then briefly equilibrated for 1 ps at 20 fs time step by MD simulations running on GROMACS package[66] with MARTINI force field[67, 68]. The pressure and temperature were maintained at 1 bar and 310K, respectively. The cutoff distances for both Coulomb and Van der Waal were set to be 12Å.

MD simulations were conducted in vacuum with modest positional restraints (spring constant = 2.5 kcal/mol/$Å^2$) exerted on the backbone beads for 1 ps at 20 fs time step, and the potential energies at the final step were recorded for the analysis to obtain the binding stability between the targeted helix in Sgo1 and mutated (or wild-type) PP2A segments.

Besides the top-ranked peptides by Martini forcefield from the aforementioned 69 complexes, we also searched potential helical binders from TP-DB using at least 4 anchoring residues in the search pattern (**Tables S5** to **S8**). We carried out all-atom molecular dynamics simulations for these peptides to assess their interactions with the targeted helix of Sgo1 using MM/PBSA[69, 70] (see below). The top 11 identified helices (**Table S9**), searched from the queries [A 3 K 3 G 2 Y], [K 3 G 2 Y 3 A] and [A 2,3 K 2,3 G 2,3 Y 2,3 A], were first superimposed onto the template helix of the PP2A [PDB: 3FGA; chain B; residue 372 to 383] with their common residues K374, G378, and Y381 (as shown in Fig S3), before the simulations.

We carried out the all-atom molecular dynamics simulations by AMBER16[71, 72] package, using ff14SB forcefield[73] for protein and ionsjc_tip3p for ions[74] in explicit solvent. Each molecular system to be simulated

was placed in a periodic box and solvated with TIP3P water. All the input files for MD simulations were prepared using tLeap[71] from AmberTools16.

Energy minimization was performed in three stages - first, with weak harmonic positional restraints (spring constant = 0.5 kcal/mol/Å^2) on all atoms except for the water/solvent atoms; second, with weak harmonic restraints on the CA atoms of amino acids; lastly, without restraints. Each of the first two stages of energy minimization is composed of 5,000 steps, with the first 2500 steps (for each energy minimization stage) carried out using steepest descent algorithm and the remaining 2,500 steps carried out using conjugate gradient algorithm.

Following the energy minimization, each simulation system was slowly heated to 310K while applying weak harmonic positional restraints on the CA atoms. Each system was then equilibrated at the temperature reached without any positional restraints prior to production MD simulations. Each production MD simulation ran for 500 ns at 2 fs time step. Non-bonded interactions were evaluated up to a cut-off distance of 10 Å where it was switched off with a cubic spline switch function. Particle Mesh Ewald method[23] was used to calculate full electrostatic interactions energies. All temperature regulations were done using Langevin thermostat (with a collision frequency, γ, of 2 ps^{-1}) and all pressure control was done with Berendsen barostat[75].
Binding free energy change (binding ΔG) was calculated from the MD simulations trajectories using AMBER16's implementation of Molecular Mechanics/Poisson-Boltzmann Surface Area (MM/PBSA)[69, 70].

Supplementary References

55. Salic A, Waters JC, Mitchison TJ. Vertebrate shugoshin links sister centromere cohesion and kinetochore microtubule stability in mitosis. *Cell* **118**, 567-578 (2004).

56. Wang L-H, Yen C-J, Li T-N, Elowe S, Wang W-C, Wang LH-C. Sgo1 is a potential therapeutic target for hepatocellular carcinoma. *Oncotarget* **6**, 2023 (2015).

57. Liu H, Rankin S, Yu H. Phosphorylation-enabled binding of SGO1–PP2A to cohesin protects sororin and centromeric cohesion during mitosis. *Nature cell biology* **15**, 40-49 (2013).

58. Tang Z, Shu H, Qi W, Mahmood NA, Mumby MC, Yu H. PP2A is required for centromeric localization of Sgo1 and proper chromosome segregation. *Developmental cell* **10**, 575-585 (2006).

59. Marston AL. Shugoshins: tension-sensitive pericentromeric adaptors safeguarding chromosome segregation. *Molecular and cellular biology* **35**, 634-648 (2015).

60. Wang C-A, Liu Y-C, Du S-Y, Lin C-W, Fu H-W. Helicobacter pylori neutrophil-activating protein promotes myeloperoxidase release from human neutrophils. *Biochemical and biophysical research communications* **377**, 52-56 (2008).

61. Yang Y-C*, et al.* High yield purification of Helicobacter pylori neutrophil-activating protein overexpressed in Escherichia coli. *BMC biotechnology* **15**, 1-11 (2015).

62. Walker IH, Hsieh P-c, Riggs PD. Mutations in maltose-binding protein that alter affinity and solubility properties. *Applied microbiology and biotechnology* **88**, 187-197 (2010).

63. Iankov ID, Haralambieva IH, Galanis E. Immunogenicity of attenuated measles virus engineered to express Helicobacter pylori neutrophil-activating protein. *Vaccine* **29**, 1710-1720 (2011).

64. Hong Z-W, Yang Y-C, Pan T, Tzeng H-F, Fu H-W. Differential effects of DEAE negative mode chromatography and gel-filtration chromatography on the charge status of Helicobacter pylori neutrophil-activating protein. *PloS one* **12**, e0173632 (2017).

65. Lee J*, et al.* CHARMM-GUI input generator for NAMD, GROMACS, AMBER, OpenMM, and CHARMM/OpenMM simulations using the CHARMM36 additive force field. *Journal of chemical theory and computation* **12**, 405-413 (2016).

66. Pronk S*, et al.* GROMACS 4.5: a high-throughput and highly parallel open source molecular simulation toolkit. *Bioinformatics* **29**, 845-854 (2013).

67. de Jong DH*, et al.* Improved parameters for the martini coarse-grained protein force field. *Journal of chemical theory and computation* **9**, 687-697 (2013).

68. Wassenaar TA, Ingólfsson HI, Böckmann RA, Tieleman DP, Marrink SJ. Computational lipidomics with insane: a versatile tool for generating custom membranes for molecular simulations. *Journal of chemical theory and computation* **11**, 2144-2155 (2015).

69. Kollman PA*, et al.* Calculating structures and free energies of complex molecules: combining molecular mechanics and continuum models. *Accounts of chemical research* **33**, 889-897 (2000).

70. Wang J, Hou T, Xu X. Recent advances in free energy calculations with a combination of molecular mechanics and continuum models. *Current Computer-Aided Drug Design* **2**, 287-306 (2006).

71. Salomon-Ferrer R, Case DA, Walker RC. An overview of the Amber biomolecular simulation package. *Wiley Interdisciplinary Reviews: Computational Molecular Science* **3**, 198-210 (2013).

72. Salomon-Ferrer R, Götz AW, Poole D, Le Grand S, Walker RC. Routine microsecond molecular dynamics simulations with AMBER on GPUs. 2. Explicit solvent particle mesh Ewald. *Journal of chemical theory and computation* **9**, 3878-3888 (2013).

73. Maier JA, Martinez C, Kasavajhala K, Wickstrom L, Hauser KE, Simmerling C. ff14SB: improving the accuracy of protein side chain and backbone parameters from ff99SB. *Journal of chemical theory and computation* **11**, 3696-3713 (2015).

74. Joung IS, Cheatham III TE. Determination of alkali and halide monovalent ion parameters for use in explicitly solvated biomolecular simulations. *The journal of physical chemistry B* **112**, 9020-9041 (2008).

75. Berendsen HJ, Postma Jv, van Gunsteren WF, DiNola A, Haak JR. Molecular dynamics with coupling to an external bath. *The Journal of chemical physics* **81**, 3684-3690 (1984).