A novel computational method for head-to-tail peptide cyclization: application to urotensin II

Yasaman Karami[1], Samuel Murail[1], Julien Giribaldi[2], Benjamin Lefranc[3], Jérôme Leprince[3], Sjoerd J. de Vries[1,*], Pierre Tufféry[1,*]

[1] Université de Paris, CNRS UMR 8251, INSERM ERL U1133, Paris, France.
[2] Institut des Biomolécules Max Mousseron, UMR 5247, Université de Montpellier-CNRS, 34095 Montpellier, France.
[3] Normandie Université, UNIROUEN, INSERM U1239, Neuronal and Neuroendocrine Differentiation and Communication, PRIMACEN, Rouen, France.
[*] corresponding authors: sjoerd.de-vries@inserm.fr, pierre.tuffery@univ-paris-diderot.fr

Abstract

Peptides have recently re-gained interest as therapeutic candidates but their development remains confronted with several limitations including low bioavailability. Backbone head-to-tail cyclization is one effective strategy of peptide-based drug design to stabilize the conformation of bioactive peptides while preserving peptide properties in terms of low toxicity, binding affinity, target selectivity and preventing enzymatic degradation. However, very little is known about the sequence-structure relationship requirements of designing linkers for peptide cyclization in a rational manner. Recently, we have shown that large scale data-mining of available protein structures can lead to the precise identification of protein loop conformations, even from remote structural classes. Here, we transpose this approach to head-to-tail peptide cyclization. Firstly we show that given a linker sequence and the conformation of the linear peptide, it is possible to accurately predict the cyclized peptide conformation improving by over 1 Å over pre-existing protocols. Secondly, and more importantly, we show that is is possible to elaborate on the information inferred from protein structures to propose effective candidate linker sequences constrained by length and amino acid composition, providing the first framework for the rational peptide head-to-tail cyclization. As functional validation, we apply it to the design of a head-to-tail cyclized derivative of urotensin II, an 11-residue long peptide which exerts a broad array of biologic activities, making its cognate receptor a valuable and innovative therapeutic or diagnostic target. We propose a three amino acid candidate linker, leading to the first synthesized 14-residue long cyclic UII analogue with excellent retention of in vitro activity.

Introduction

Several naturally occuring cyclic peptides constitute alternatives to antibiotics and peptide backbone cyclization is frequently used in peptide-based drug design to convey druggable properties to linear bioactive sequences [1, 2]. Peptides in general combine high affinity with high target selectivity and low toxicity, and are a natural choice in the targeting of protein-protein interactions. While preserving these favorable properties, peptide cyclization additionally confers peptides with more rigid conformation and enhanced stability towards enzymatic proteolysis and improves the permeability through biological barriers [3–7]. Moreover, many natural-occurring cyclic peptides are known from different kingdoms of organisms, exhibiting diverse biological activities, including anti-tumor [8, 9], antimicrobial [10, 11] and antihelminthic activities [12–14]. Together, this has caused a growing interest toward cyclic peptides, thus the number of designed cyclic peptide drugs is growing [15].

When designing new cyclic peptides, there are broadly two strategies that can be followed: *i)* de novo design, or *ii)* cyclization of an existing peptide. For the first strategy, a number of experimental techniques are available, such as SICLOPPS [16], phage display, and mRNA display [17]. These are all based on libraries of random cyclic peptides that are subjected to *in vitro* selection. They can be complemented with library-based computational approaches such as from Slough et al. [18], CAESAR [19], Omega [20] and CycloPs [21] based on rdkit (`https://github.com/rdkit/rdkit`). Those approaches are conceptually similar to the molecular modeling of small ligands, with the corresponding strengths (arbitrary molecular topologies) and weaknesses (limited number of flexible bonds). For computational de novo design, an alternative approach is to perform peptide structure prediction, using one of the many fragment-based methods that are available, such as PLOP [22, 23], Peplook [24, 25], PEPstrMOD [26] or PEP-FOLD [27, 28], while imposing cyclization as a bond or distance restraint (see [29] for a review). Since these methods leverage the existing wealth of knowledge of protein and peptide structure, they can deal with larger peptides, but have difficulties where this knowledge falls short, *i.e.*, for unnatural amino acids.

For the second strategy, the starting point is an existing linear peptide of known structure. It is well established that small linear peptides generally exist in solution in an interchangeable conformational equilibrium. This flexibility provides to bioactive peptides the ability to interact with several types or subtypes of receptors for instance. Stabilizing a bioactive conformation is a challenge that can be tackled by a variety of cyclization strategies. On the one hand, this can be as straightforward as mutating two spatially close residues into cysteins with the aim of introducing a disulfide bond. On the other hand, sophisticated chemical scaffolds or cyclotides can be used for the grafting or stitching of peptides or cyclotides into rigid bioactive conformations [30, 31]. One particular successful strategy has been head-to-tail peptide backbone cyclization [32–41]. This involves the design of a sequence that links the N- and C-terminal extremities of the peptide. In principle, any amino acid can be part of the linker sequence, but Gly, Ala and Pro residues are often favored because they are small and their side chains cannot form hydrogen bonds, which could potentially disrupt the bioactive conformation.

Head-to-tail cyclization leads to cyclic peptides with improved pharmacological properties (affinity, potency, efficiency, selectivity) when compatible with target specificity (or bioactivity conservation). Whether the cyclic peptide is active or not, it is generally less sensitive to metabolic degradation. However, cyclization is often unsuccessful due to imposed conformational restriction that is too strict and too far from the bioactive structure. In order to avoid this, it is necessary to understand the general sequence-structure requirements; in particular: what is the allowed sequence space of the linker, and what will be the structure of the cyclized peptide? This is a challenging issue, and to the best of our knowledge, there is only one computational protocol that has been successfully applied to head-to-tail cyclization linker design, namely the Rosetta protocol used by Bhardwaj et al. [6]. However, in that study, the sequence and structure of the entire cyclic peptide were designed from scratch. Otherwise, we are not aware of any computational methods that can predict the sequence and structure of a head-to-tail cyclization linker, while preserving the sequence and structure of the linear peptide that is being cyclized.

Recently, we have developed DaReUS-Loop [42, 43], a fast data-based approach that identifies loop candidates mining the complete set of available experimental protein structures. This is done by treating the loop as a gap in the structure, and considering the flanking regions of the structure immediately before and after the gap. Loop candidates are then favored that *i)* superimpose well onto the flanks, and *ii)* have a compatible sequence. Recognizing the conceptual similarity, we have now developed PEP-Cyclizer, a method that extends the DaReUS-Loop approach and applies it to rational head-to-tail peptide cyclization. This method provides two complementary possibilities: *i)* given a sequence for the cyclization linker, PEP-Cyclizer can predict structural models for the cyclized peptide, *ii)* PEP-Cyclizer can propose candidate cyclization linker sequences, constrained by length and amino acid composition. PEP-Cyclizer is the first method that can propose the sequence or the structure of a head-to-tail cyclized peptide, starting from the linear peptide structure. For structure prediction, PEP-Cyclizer was validated on a benchmark of

five cyclic conotoxin structures for which a linear structure is available as well. With regard to the experimental structures, the predicted cyclized peptide models had a root-mean-square deviation (RMSD) of 2.0 Å (3.2 Å) for the top 20 (top 1) models, an improvement of more than 1 Å over the Rosetta Next-generation KIC (NGK) protocol [44], a high-resolution Rosetta protocol for the modelling of missing regions. For sequence prediction, PEP-Cyclizer was validated on the same benchmark and in result, experimental sequences were ranked significantly better than other sequences of the same length and composition.

As a functional validation, PEP-Cyclizer was used to design a cyclized peptide sequence of the human urotensin II (UII), that is an 11-residue long disulfide-bridged peptide [45]. UII exerts a broad array of biological activities, in particular in the central nervous system, the cardiovascular system, and the kidney. It has been suggested that the cognate receptor of UII (UT), may emerge as a valuable and innovative therapeutic or diagnostic target [46]. Indeed, high affinity, potent and selective UT peptide ligands have been designed, from structure-activity relationship studies [47] to further elucidate the pharmacology and biology of UII towards new therapeutic opportunities, such as the treatment of sepsis-induced lung damage [48]. In this context, introduction of a main conformational restraint through head-to-tail cyclisation has become a standard strategy to improve pharmacological profile of peptide ligands [49]. The NMR structure of the disulfide-bridged core of UII is well-defined, whereas the flanking linear extremities are very flexible [50–53]. Depending on the experimental environment (water or membrane mimetic micelles) and temperature, distinct conformations are stabilized within the disulfide-bridged core involving different sets of intramolecular hydrogen bonds. Here, using a linker predicted by PEP-Cyclizer, a head-to-tail cyclized UII peptide was synthesized and its activity validated, proposing the first bicyclic active UII analogue.

Results

PEP-Cyclizer considers all cyclization linker candidate structures that are compatible with the flanks of the uncyclized peptide structure. The sequences of these linker candidates, potentially filtered by *a priori* sequence constraints, are then used to build a linker sequence profile. This profile feeds a Hidden Markov Model from which it is possible to estimate the likelihood of candidate linkers using a forward-backtrack algorithm. Alternatively, if the linker candidates are restricted to one known linker sequence, they are clustered and superimposed onto the flanks, providing structural models of the cyclized peptide. **Figure 1** depicts the workflow of the method.

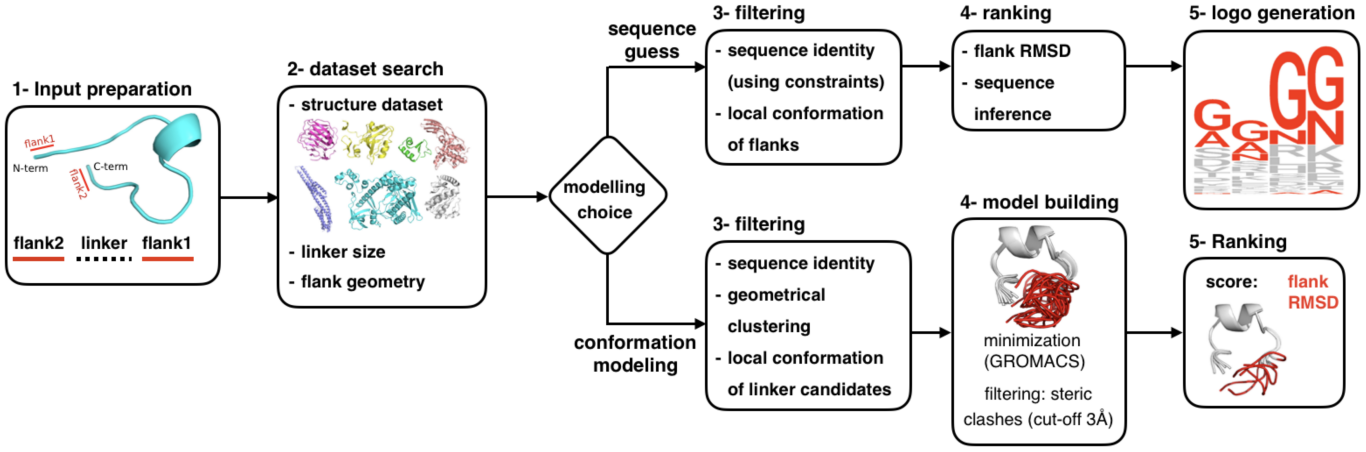

Figure 1: **The workflow of PEP-Cyclizer.** The workflow describes main steps for peptide head-to-tail cyclization. The method provides two possibilities: proposing candidate sequences for the linker, or modelling the 3D conformation. The steps of the workflow are: input preparation, linker candidate search, candidate filtering, model building, model selection and logo generation in case of sequence prediction. The inputs are a linear peptide and either the amino acid constraints for sequence prediction, or the linker sequence for conformation modelling. In the final step, for conformation modelling, the 20 best models are returned as the final predictions. For sequence prediction a logo is generated and a forward-backtrack algorithm is used to sample the sequence space and assess the likelihood of the candidate linkers. Note that the sequence logo serves strictly as a global visualization of the ensemble of generated sequence candidates, and has no predictive power by itself.

As a positive control, PEP-Cyclizer was applied to 64 cyclic peptides from the CyBase database (http://www.cybase.org.au/) [54, 55] (see **Table S1** for a complete list of studied peptides). 1147 linear peptides were artificially generated by removing segments of 2-7 residues from the 64 cyclic peptides, details are reported in **Supplementary Materials - CyBase benchmark**. Unlike a real-world situation, where a peptide may undergo conformational changes upon cyclization, these artificial linear peptides represent perfect conformations for modelling the removed linker conformation. For all linker sizes, PEP-Cyclizer was able to produce accurate models of the local linker conformation, with an average accuracy of 1 Å or better. This is comparable (although not fully equal in accuracy) to models obtained for the same peptides using Rosetta NGK (comparisons reported per peptide and linker size in **Table S2** and **S3, respectively**).

PEP-Cyclizer was then applied to a small benchmark of real-world cases, in the form of several conotoxin peptides where both cyclized and non-cyclized structures are available in the PDB. Seven distinct cyclized/non-cyclized pairs of peptide structures were identified (**Table 3** and **S4**). The range of backbone RMSD between the overlapping region of cyclized and uncyclized forms is 0.4-2.5Å. Using the known linker sequence, only the non-cyclized structure was used to model the linker. In this case, PEP-Cyclizer was able to return a model approximating the global structure of the linker at 2.01Å on average (1.01Å for the local linker conformation), as reported in **Table 1**. This is a considerable improvement over Rosetta NGK (3.48Å), which suffers from the structural imprecisions caused by conformational change upon cyclization. In contrast, our results show PEP-Cyclizer to be rather robust against such imprecisions. **Figure 2** illustrates the results for the best predictions - out of the top 20 - of PEP-Cyclizer (in green) and Rosetta NGK (in cyan), starting from the first NMR conformation of each uncyclized peptide.

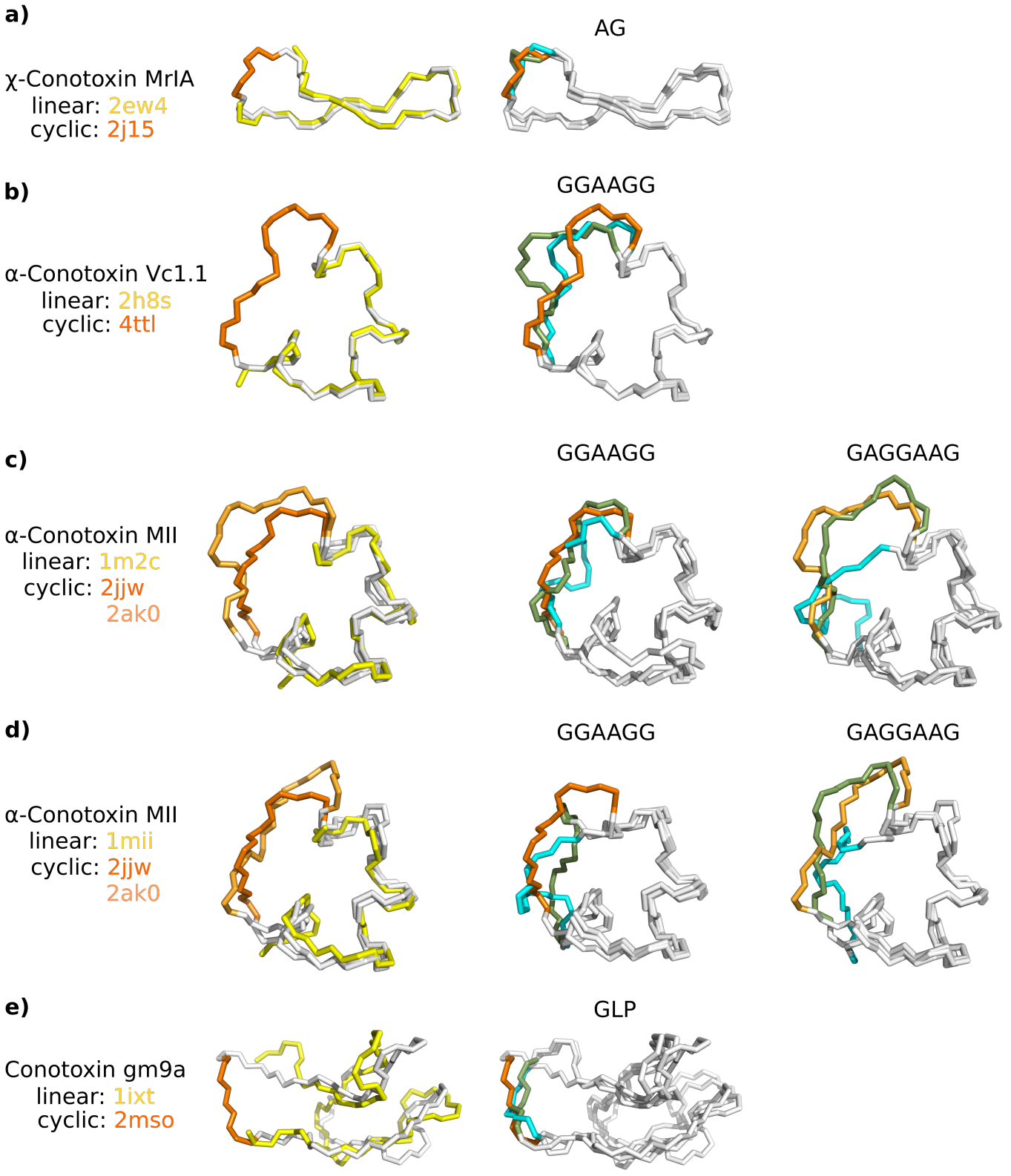

Figure 2: **Structure of the studied linear Conotoxins and their corresponding engineered cyclic peptides.** The native linear and cyclic peptides are shown at the left column, colored in yellow and orange, respectively. The structures on the middle and right columns, represent the comparison between the native linker (in orange), linkers modelled by PEP-Cyclizer (in green) and Rosetta NGK (in cyan). See **Table 1** for the corresponding $gRMSD_{20}$ (and $lRMSD_{20}$) values. The corresponding linker sequences are reported for every model, at the top.

Name	lsz	N_{model}	PEP-Cyclizer				Rosetta NGK			
			$lRMSD^*_{20}$	$gRMSD^*_{20}$	$lRMSD^*_1$	$gRMSD^*_1$	$lRMSD^*_{20}$	$gRMSD^*_{20}$	$lRMSD^*_1$	$gRMSD^*_1$
2ew4	2	20	0.47±0.13	2.03±0.84	0.67±0.18	3.57±1.35	0.31±0.22	2.53±1.25	0.39±0.26	3.10±1.25
1ixt	3	20	0.46±0.08	2.31±0.25	1.43±0.35	3.52±1.50	0.37±0.14	2.81±0.54	0.43±0.14	3.08±0.66
1m2c	6	14	1.31±0.15	1.99±0.16	1.66±0.12	2.58±0.57	1.33±0.17	4.58±0.90	1.53±0.16	6.08±1.53
1mii$^+$	6	20	1.35±0.01	1.72±0.01	1.75±0.50	3.11±1.74	1.56±0.11	5.76±0.55	1.72±0.12	7.22±0.45
2h8s	6	20	1.24±0.12	2.12±0.10	2.03±0.29	4.05±1.21	1.30±0.19	3.01±0.53	1.60±0.21	3.81±0.64
1m2c	7	14	1.59±0.16	1.97±0.25	2.09±0.56	3.66±1.56	1.84±0.37	5.34±1.30	2.12±0.26	6.53±1.29
1mii$^+$	7	20	1.51±0.08	1.89±0.06	1.69±0.01	2.30±0.02	1.56±0.03	5.51±0.72	1.64±0.06	7.80±0.67
average$^+$			1.01±0.46	2.01±0.43	1.49±0.52	3.24±1.26	0.95±0.63	3.48±1.39	1.13±0.71	4.28±1.77

Table 1: **Comparison of RMSD values for the predicted linkers using all the NMR models of the linear peptides.** For each peptide we report the average local and global RMSD values over the top 1 ($lRMSD^*_1$ and $gRMSD^*_1$) and best out of top 20 predictions ($lRMSD^*_{20}$ and $gRMSD^*_{20}$). The average values are measured over all $N_{uncyclized}$ NMR conformations of each linear peptide (see **Methods**). The RMSD values are calculated over the backbone atoms (N, C, Cα and O). $^+$The structure of 1mii and 1m2c correspond to the same protein (α-Conotoxin MII), and to avoid redundancies in reported values, we measured the average considering the best predictions between 1m2c and 1mii for each method.

Next, the ability of PEP-Cyclizer to propose peptide linker sequences was tested. The same conotoxin benchmark was used, adding ten cyclic sequences with available structures for the uncyclized but not the cyclized peptide, for a total of seventeen sequences. The details of the peptides are reported in **Tables 3** and **S4**. As potential linker sequences, all combinations of all amino acids that are present in the experimental linker sequence (typically only Gly and Ala) were considered, and ranked by the forward-backtrack algorithm. The results are shown in **Figure 3**. The experimental sequences were ranked significantly better (average percentile: 37.4, p=0.025) than other potential sequences.

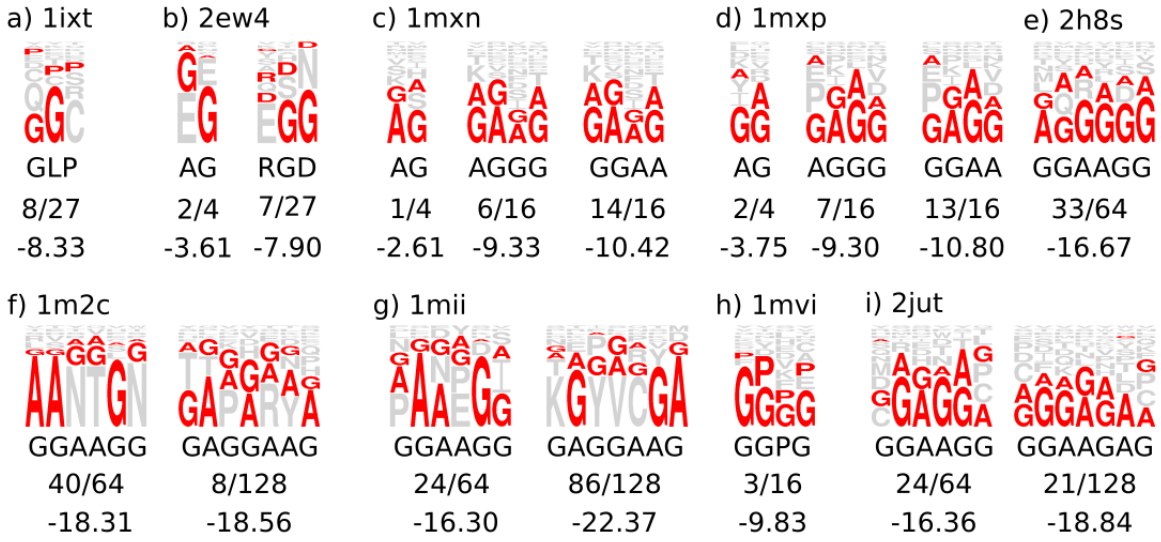

Figure 3: **Sequence logo generated by PEP-Cyclizer for the studied cases.** The pdb code of the linear peptides used as input are reported for each case. Below every logo, the desired linker sequence, its rank and score among the proposed sequences by the forward-backtrack algorithm are reported.

Application to urotensin II

Finally, PEP-Cyclizer was applied to predict a head-to-tail cyclization linker sequence for UII. So far, only the structures of a fragment corresponding to the eight last amino acids of UII and its N-methylated tryptophan counterpart, [(N-Me)Trp^7]U-II$_{4-11}$ in polar conditions (PDB entries 6HVB and 6HVC) have been solved by NMR. Since our goal was to obtain a head-to-tail cyclized version of UII, we decided to start from 3D models of the complete linear UII (11 amino acids). Therefore two ensembles of 8 and 5 conformations were generated using two distinct strategies: *i)* molecular dynamics simulations (MD) and *ii)* PEP-FOLD [28]. The models are highly structurally divergent, with typical RMSD values in excess of 2 Å both within and between the ensembles (**Supplementary Table S6**). Consequently all those models were used as the starting points for the cyclization (see **Methods**). Based on the average distance between the N- and C-terminus of the models (7.27+/-2.14 Å), a linker of size 3 was considered, accepting only alanine, proline and glycines. **Table 2** presents the results cumulated for each of the two ensembles of models. As can be observed, it is striking that despite the diversity of the conformations and the way they were obtained, those two independent ensembles of models resulted in a rather stable ranking of the predicted sequences. This is reflected by the fact that in both cases, the top 4 consists of the same four sequences, as well as by the high overall correlation of the ranks (Spearman r=0.98).

Table 2: **The likelihood of each of the possible 27 linker sequences.** Two independent series of models (8 generated using MD and 5 using PEP-FOLD) were used as starting points.

PEP-FOLD		**MD**	
linker	L	linker	L
AGG	-7.42	AGG	-7.21
APG	-7.65	GAG	-7.52
GAG	-7.67	PGG	-7.52
PGG	-7.73	APG	-7.61
AGA	-7.80	GGG	-7.64
PAG	-7.81	PAG	-7.69
GGG	-7.89	AAG	-7.78
AAG	-7.90	AGA	-7.80
PGA	-8.11	PGA	-8.10
GAP	-8.12	PPG	-8.11
PPG	-8.15	GAP	-8.21
AGP	-8.19	AGP	-8.22
PAP	-8.26	GGA	-8.23
GGA	-8.27	GPG	-8.36
APP	-8.29	PAP	-8.38
APA	-8.32	AAP	-8.47
AAP	-8.34	APA	-8.49
GPG	-8.44	APP	-8.50
PGP	-8.50	PGP	-8.53
GGP	-8.66	GGP	-8.65
GAA	-8.75	GAA	-8.79
PPP	-8.80	PAA	-8.96
PPA	-8.82	PPA	-8.99
PAA	-8.88	PPP	-9.00
AAA	-8.97	AAA	-9.06
GPP	-9.08	GPA	-9.24
GPA	-9.11	GPP	-9.25

To test the significance of our approach, we analyzed the impact of one linker in a functional assay. Since the repetition of similar consecutive amino acids can lead to some difficulties [56], the top-ranked sequence AGG was discarded; instead, UII was cyclized with the sequence GAG, leading to the LV-4130 cyclic peptide. The head-to-tail cyclized peptide underwent synthesis and functional tests (see **Methods**). Briefly, the linear precursor (CFWKYCV-

GAGETPD) was first assembled on a Fmoc-Asp(Wang resin)-OAl. After selective deprotection of both extremities and cysteine residues, intramolecular backbone and side-chain to side-chain (S-S) cyclizations, respectively, were successively carried out. Finally, after resin cleavage and side-chain deprotection and purification, highly pure bicyclic UII (LV-4130) was obtained with <1% yield. The pharmacological profile of this synthesis-challenging compound was assessed by testing its ability to increase intracellular calcium concentration $[Ca^{2+}]i$ in human UT-transfected CHO cells (Eurofins-Cerep), as previously described [57]. As shown in **Figure 4**, UII and LV-4130 induced a dose-dependent increase in $[Ca^{2+}]i$ with EC_{50} of 0.7 and 46 nM.

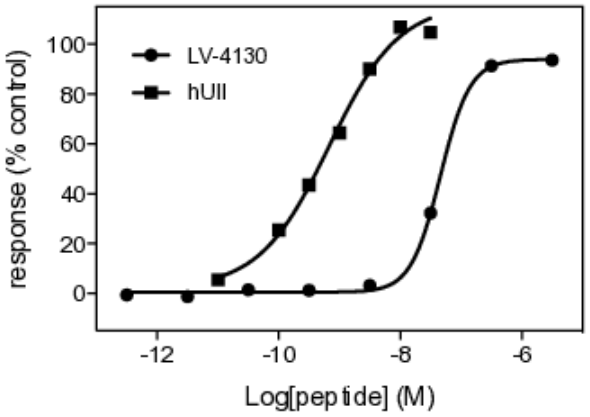

Figure 4: **Concentration-dependent agonist-evoked Ca^{2+} responses on UT-transfected CHO cells.** Agonist responses were expressed as a percent of the response observed with a maximally effective concentration of UII (100 nM). Data points represent mean of duplicate.

Our analysis shows that LV-4130, a first bicyclic UII analogue, retained a substantial ability to increase $[Ca^{2+}]i$ in UT-transfected CHO cells. While there is a shift in potency, the EC_{50} is less than 2 orders of magnitude lower, and LV-4130 is a nanomolar active UT agonist of peculiar interest. Indeed, its backbone cyclic structure may confer a less susceptibility to metabolic degradation and a better selectivity for UT or a subset of UT's signaling cascade that deserves to be investigated.

Discussion

Recently, we have demonstrated that the current structural information available in the Protein Data Bank (PDB) [58] is sufficient to propose accurate protein loop candidates, in a manner that is robust for conformational inaccuracies. In the present study, this is shown to be true for peptide cyclization linkers as well. We propose the first computational method to assist the design of head-to-tail cyclization of an existing peptide structure, a well-known strategy to enhance peptide resistance to enzymatic degradation and thus peptide bioavailability. The method addresses two complementary questions, namely : *(i)* proposing candidate sequences for the linker, a facility to assist medicinal chemists, and *(ii)* predicting the 3D conformation of the linker, for further peptide conformational stability analysis or peptide-receptor docking. Up to now, there has been an evident lack of computational methods to answer those questions. Existing methods [18–28] are oriented towards de novo design and do not perform head-to-tail cyclization of existing structures. We are aware of a single existing computational method, the Rosetta protocol by Bhardwaj et al. [6], that is able to design head-to-tail cyclization linkers for pre-existing peptide structures. However, in that method, what is pre-defined is the complete structure of the entire cyclic peptide, including the linker; also, the sequence of the entire peptide is designed from scratch, and not just that of the linker. In contrast, PEP-Cyclizer takes an existing structure of a linear peptide, and predicts the sequence or structure of a cyclization linker, while leaving the rest of the peptide undisturbed. To the best of our knowledge, PEP-Cyclizer is the first computational method designed to

do this.

The performance of PEP-Cyclizer was validated on a set of conotoxins for which both linear and cyclic peptide structures are known. For comparison, we also evaluated a Rosetta protocol, not the one from Bhardwaj et al. [6], but the Rosetta NGK protocol [44], a state-of-the-art protocol for building missing loops in crystal structures. It must be mentioned that Rosetta NGK is not designed for peptide cyclization and we had to modify the input data and convert the head-to-tail cyclization to loop modelling (*i.e.*, dividing the peptides in two segments and switching them to generate a gapped structure). In all cases, the peptide linker models generated by PEP-Cyclizer had a significantly better global accuracy. This is especially evident for the two longest (7 amino acids) linkers, where the RMSD was $<$ 2 Å, while $>$ 5 Å for Rosetta NGK.

PEP-Cyclizer is the extension of the DaReUS-Loop algorithm for the problem of head-to-tail peptide cyclization; details about the algorithm are reported in our previous study [42]. Essentially, the linker/loop is treated as a gap in the structure, and a structural database search is performed using the flank regions on either side. Like DaReUS-Loop, PEP-Cyclizer is a consensus method that considers both structural compatibility (*i.e.*, good superposition of the linker candidate onto the flanks) and sequence compatibility. Therefore, when using PEP-Cyclizer to predict linker conformations, it is essential to consider all 20 candidate structures. When PEP-Cyclizer is forced to make a single prediction, the quality deteriorates considerably (from 2.0 to 3.2 Å). While a top-1 accuracy is naturally less favorable than a best-of-20 for any prediction method, it is specifically true for PEP-Cyclizer, as the effect is much weaker for Rosetta NGK (from 3.5 to 4.3 Å).

In contrast, PEP-Cyclizer is shown to be very robust against conformational changes. For the conotoxin benchmark, the range of backbone RMSD between the overlapping region of cyclized and uncyclized forms is 0.4-2.5Å. This is to be compared with the positive control, where this RMSD is zero. However, the global accuracy of the PEP-Cyclizer models is essentially the same between the two (2.0 Å vs 1.87 Å). This is a stark contrast to Rosetta NGK, which performs very well on the positive control (1.33 Å), but poorly on the real-world conotoxin benchmark (3.5 Å). This is an expected result, as Rosetta NGK is primarily designed to complete missing regions in otherwise high-quality crystal structures. Note that as a high-resolution protocol, Rosetta NGK does a good job in generating accurate local linker conformations; it is the global positioning of the linker onto the rest of the conotoxin structure where Rosetta NGK is outperformed by PEP-Cyclizer.

The robustness of PEP-Cyclizer for conformational change is also apparent for the prediction of linker sequences. For UII, sequence prediction was performed on two different structure ensembles that were of different origin and highly divergent, with very similar results. Note that it is inherently complicated to evaluate linker sequence predictions, as we only have a few positive cases and no negatives, *i.e.*, we normally do not know that a sequence does *not* cyclize. In addition, we must stress that PEP-Cyclizer proposes linker candidates based on likely sequence and structure only; in contrast, it cannot predict if a proposed cyclized peptide is likely to fold or not (or otherwise preserve its stability) if synthesized *in vitro*. Future research will focus on the prediction of the most likely length of the linker sequence, for which the current protocol does not show significant predictive power. Still, the result that experimental sequences were on average better ranked shows that PEP-Cyclizer has at least some predictive power. More importantly, the activity of the head-to-tail cyclic UII peptide LV-4130 demonstrates that PEP-Cyclizer has direct practical ability in cyclic peptide-based drug design.

Materials and Methods

In this section we explain the details of PEP-Cyclizer, that is an extension of DaReUS-Loop, a data-based approach using remote or unrelated structures for loop modelling [42, 43]. Starting from the geometry of flank residues, *i*.e., four residues before and four residues after the loop of interest, PEP-Cyclizer mines a structure database and identifies all possible candidates. It then integrates a filtering step, and in the end, ranks the candidates and proposes a final set of top models (structures or sequences). PEP-Cyclizer implements two complementary and new functionalities: (*i*) guessing the linker sequence and (*ii*) modelling the conformation of the linker. The details of those functionalities are

depicted in **Figure 1**, and explained in the followings.

Structure Database

We employed two different structure databases. The first one is the database to search for linker candidates, which contains the entire set of protein structures available in the PDB. In March 2017, it consisted of 123,417 PDB entries, corresponding to 338,613 chains in total. The second database, is the one to search for linker sequences and contains the entire set of protein structures available in pdb70. For every database, each chain was split into segments that correspond to consecutive regions separated by gaps or non-standard residues, but accepting seleno-methionines. This led to two databases with 758,143 and 172,693 protein segments, respectively.

Test sets

To validate our approach, we have searched for cases for which both structures of the un-cyclized and cyclized peptides are available. Backbone cyclization has been applied to few conotoxins, as reported in [56], and to the best of our knowledge, the structures (NMR/Xray) of only five engineered cyclic conotoxins for which the structure of the un-cyclized form exists have been deposited in the PDB database [58]. For one of the cases, two structures of the open form have been deposited in PDB (1m2c and 1mii), and their structures deviate by 1Å, and we have included both structures in our test set. For 3 additional peptides, the structure of the un-cyclized conformation and information about successful linkers are available. **Table 3** reports the details of those studied cases. Of note, the structure of all the linear and cyclic peptides in this test set have been determined using NMR, at the exception of one case (4ttl) for which it has been solved by X-ray crystallography.

Since all the structures of the un-cyclized forms of the peptides have been determined using NMR and have $N_{uncyclized}$ conformations, we have performed the head-to-tail cyclization starting from all $N_{uncyclized}$ NMR conformations. The final predictions for the cyclized forms of the peptides have been in turn compared with all the $N_{cyclized}$ conformations of the cyclized structures. **Table 1** summarises the average local and global $RMSD^*_{20}$ (best out of top 20) and $RMSD^*_1$ (top 1) values obtained for each linker (averaged over $N_{uncyclized}$ conformations).

Table 3: **The list of real cases for head-to-tail cyclization.** The PDB code of the un-cyclized and cyclized peptides (if available) are reported. With the exception of 4ttl, all the other structures are obtained using NMR and has several models. The average RMSD values are measured between all the models of the un-cyclized and cyclized conformations. In some cases more than one linker sequence exist, as reported in the last column of the table.

un-cyclized	#NMR models	cyclized	#NMR models	RMSD (Å)	un-cyclized size	cyclized size	linker sequence
1m2c	14	2ajw	20	1.22 +/- 0.10	16	22	GGAAGG
1m2c	14	2ak0	20	1.09 +/- 0.12	16	23	GAGGAAG
1mii	20	2ajw	20	1.26 +/- 0.09	16	22	GGAAGG
1mii	20	2ak0	20	1.03 +/- 0.12	16	23	GAGGAAG
2h8s	20	4ttl	1	0.40 +/- 0.00	16	22	GGAAGG
1ixt	20	2mso	20	2.45 +/- 0.07	27	30	GLP
2ew4	20	2j15	21	1.03 +/- 0.40	13	15	AG
2ew4	20	-	-	-	13	16	RGD
1mxn (1mxp)	20	-	-	-	15	17, 19, 19	AG, AGGG, GGAA
2jut	20	-	-	-	13	19, 20	GGAAGG, GGAAGAG
1mvi	15	-	-	-	25	28	GLP

Input preparation and candidate search

We consider head-to-tail cyclization as a loop modelling problem, where the loop flanks are the first and the last four residues in the N-terminus and C-terminus, respectively. Accordingly, the minimum acceptable size for the input linear

peptide is 8 residues. We then, switch the flanks and search for linker candidates that match those flanks. We employ the method that was previously introduced to mine the database using a Binet-Cauchy (BC) kernel and a Rigidity score [59] (detail in **Supporting Materials**).

Candidate filtering

In most cases the number of candidates returned by BCLoopSearch is too large to be tractable, which implies to limit their number. Different filters were sequentially applied in our protocol for each mode of prediction:

Modelling the conformation of the linker

- **Sequence similarity:** The sequence similarity of a linker candidate with the query linker sequence using BLOSUM62 score. Candidates with negative scores were discarded.

- **Geometrical clustering:** We used the python Numpy library to measure the pairwise distances (RMSD) between all the candidates [60]. In addition, we used the python Scipy package to perform hierarchical clustering [61]. A RMSD cut-off of 1Å was used to group similar linker candidates. To consider memory constraints, we applied an iterative clustering over subsets of 25,000 candidates, until at most 25,000 clusters were obtained. Finally, one representative linker candidate with the highest sequence similarity to the query linker was selected for each cluster. The computational time of our clustering protocol is in the range of 1-5 minutes, however it depends directly on the number of candidates detected by BCLoopSearch. In extreme cases, the needed time may increase up to 10-15 minutes.

- **Local conformation:** Previously, Shen et al. have shown that local conformation profiles predicted from sequence and profile-profile comparison can be employed to accurately distinguish similar structural fragments [62]. Consequently, we pre-computed a collection of profiles for all the protein chains in the structure dataset, and for all proteins of the test sets. For each linker candidate, it is thus possible to extract the sub-profiles P and Q, corresponding to the query and candidate linker, and to measure the Jensen Shannon divergence ($JS(P,Q)$) between these profiles:

$$JS(P,Q) = \frac{1}{2}D_{KL}(P,M) + \frac{1}{2}D_{KL}(Q,M) \tag{1}$$

  where M corresponds to $1/2(P+Q)$ and D_{KL} is the Kullback-Leibler divergence:

$$D_{KL}(P,Q) = \sum_{1 \leq i \leq 27} P(i) ln(P(i)/Q(i)) \tag{2}$$

  $P(i)$ is the probability of SA letter i. Then we measured the average Jensen Shannon divergence (JSD) over the paired series of query and candidate profiles:

$$JSD(P,Q) = \sum_{1 \leq i \leq n} JS(P_i, Q_i)/n \tag{3}$$

  where P_i and Q_j are the two profiles corresponding to positions 1 to L on the query and candidate linker sequences. Note that a JSD of 0 indicates a perfect identity of the profiles. This procedure was applied on each linker candidate and those with a $JSD > 0.40$ were discarded from the remaining set.

- **steric clash detection:** After modelling the complete structure, models with steric clashes were discarded considering the C_α distance between linker residues and other residues of the protein, using a cut-off value of 3 Å.

Predicting the linker sequence

- **Sequence similarity:** If sequence constraints are given, a subset of sequences that represent at least 50% sequence identity to any of the constraint amino acid types, regardless of their position, are kept.
- **Local conformation:** Measuring the local conformation of flanks (query and candidate flanks) and discarding candidates with flank $JSD > 0.40$.

Sequence constraints

Throughout the study, linker sequences were predicted using the following sequence constraints. At each position of the linker, the set of amino acids of the entire experimental linker was considered - for instance, for the RGD linker of 2ew4, the amino acids Arg, Gly and Asp were considered at all three positions, *i.e.*, 3^3 different linker sequences are possible.

Model building

Final energy minimisation was conducted using Gromacs 2018 [63], the CHARMM36m force field [64] and the steepest descent algorithm for 1000 steps. All bonds were constrained using the LINCS algorithm. The particle mesh Ewald algorithm was used to handle electrostatics with a 10 Å cutoff for the short-range part and a grid spacing of 1.2 Å for the long-range contribution in reciprocal space. The Verlet buffer scheme was used for non-bonded interactions, the neighbour list was updated every 20 steps.

Model selection

To rank the models, we considered the RMSD of the flanks. In case of conformation modelling, our procedure returns a maximum of 20 models with the lowest *flank RMSD* score. And for sequence guessing, it returns a set of 30 sequences with the lowest *flank RMSD* score. From this set and considering the sequence constraints, we apply the sequence inference procedure (as explained below) to propose final set of likely sequences for the linker.

Candidate sequence inference

To draw candidate sequences given the sequences of the candidate linkers identified, we have used a forward-backtrack procedure. One advantage of such a procedure is to provide both sequences and their likelihood. The probabilities $p^l_{aa,linker}$ of observing each amino acid type aa at position l of the *linker* can be estimated from the amino acid sequences of the candidate linkers satisfying the condition of peptide cyclization. However, when a reduced number of amino acids is considered at a given position, these estimates can be performed on a rather low number of sequences. Consequently, we have estimated pseudo-frequencies, with $p^l_{aa} = \alpha.p^l_{aa,linker} + (1-\alpha).p^l_{aa,db}$ where α is a value between 0 and 1, and $p^l_{aa,db}$ is the frequency of amino acid type aa as observed in a large collection of sequences named db. For db, we have considered the sequences of the loops of 123,417 PDB entries (758,143 protein segments), identified using the procedure described in [42]. Alternatively, we have also considered db_s, which corresponds to the subset of db corresponding to a loop size of s. Transition probabilities have been estimated similarly. Pseudo transition probabilities $p(aa^l/aa^{l-1})$ were estimated as $p(aa^l/aa^{l-1}) = \beta.p(aa^l_{linker}/aa^{l-1}_{linker}) + (1-\beta).p(aa^l_{db}/aa^{l-1}_{db})$, where β is a value between 0 and 1. Given estimates of p^l_{aa} and $p(aa^l/aa^{l-1})$ we have used the forward-backtrack algorithm to infer series of amino acids that fit best the estimates. We prefer such procedure to for instance the $viterbi_{kbest}$ procedure that, in our experience [65], usually returns less diverse sequences.

Linker quality assessment

To assess the quality of the final linker structures, we use the global RMSD of the linker candidates main chain heavy atoms (N, C, Cα and O), *i.e.*, the modeled cyclic peptides are superimposed on the native structure excluding the linker region, then the RMSD is calculated over the linker.

Statistical testing

To test the prediction of linker sequences of the conotoxin benchmark, the rank of the experimental linker sequences were determined. To avoid pseudo-replication, five duplicate cyclic sequences were eliminated; using the remainder of the benchmark, the overall ranking of the experimental linker sequences was tested for statistical significance. With the total number of linker sequences varying from case to case, and many instances of tied ranks, it was not feasible to compute an analytical p-value based on hypergeometric distributions. Instead, random ranks were simulated by sampling from flat rank distributions, converted to percentiles, and it was evaluated how often the overall mean percentile was better than the observed mean percentile (37.4) for the experimental linker sequences. This was the case in 2518/100000 random simulations, *i.e.*, a p-value of 0.025.

Comparison with other approaches

In this work we compare the performance of our linker modelling protocol with the Rosetta NGK [44]. The Rosetta NGK runs were performed using the protocol provided by [44], and Rosetta energy values were employed for ranking the models. Considering the fact that Rosetta NGK is not designed for peptide cyclization, we converted the head-to-tail cyclization to loop modelling, by breaking every peptide into two segments and switching the two.

Urotensin II cyclization

Model generation

Two sets of 3D models were used. The first one was generated using PEP-FOLD server [28], a *de novo* approach to peptide structure prediction. Five independent runs of 3D generation (100 models) were run, and five models showing closed disufide bonds in the PEP-FOLD coarse grained representation were then submitted to refinement using MD, with the aim to stabilize the disulfide bond in the all atom representation. The model topology was created using the Gromacs pdb2gmx command, which did not include the disulphide bond. The topology was further modified to include the disulfide bond parameter using the gromacs_py library [66]. Simulations were performed using the CHARMM-36 force field [67] and the TIP3P model for water. The Gromacs 2018 sofware [63] was used to run the simulations. The five models were minimized two times for 10,000 steps with the steepest descent algorithm. During the first minimisation the bonds were not constraints, as in the second and following steps, all bonds were constrained using the LINCS algorithm. The five models were solvated in a water box and roughly 150 mM of NaCl. Systems were again minimized in two similar steps. And then equilibrated in three successive steps, *(i)* 100 ps with position restraints of 1000 $kJmol^{-1}nm^{-2}$ applied on the peptide heavy atoms and an integration time step of 1 fs *(ii)* 500 ps with position restraints of 1000 $kJmol^{-1}nm^{-2}$ applied on the peptide Cα atoms, the integration time step was fixed to 2 fs *(iii)* 1 ns with position restraints of 100 $kJmol^{-1}nm^{-2}$ applied on the Cα atoms. Production runs were finally computed for 100 ns. The five 100 ns trajectories were then analysed using MDAnalysis library [68]. PCA of backbone atoms coordinates were computed and the fifteen first components were used to cluster the coordinates. The clustering DBSCAN algorithm [69] was used using a min_sample of 20, and sigma value of 5. A total of 13 clusters was identified, the cluster centroids were chosen by taking the closest element in terms of RMSD to the average cluster structure. The conformations generated using this protocol are available as supplementary information. All models underwent sequence guessing to cyclize the peptide.

Another set of models was kindly provided by D. Chatenet and co-workers, at INRS Quebec, Canada. It consists of a set of 8 representative structures of UII displaying the heterogeneous conformational ensemble of this peptide. The three-dimensional structure of UII was generated from the sequence using the pdbutilities server https://spin.niddk.nih.gov/bax/nmrserver/pdbutil/. System preparation and MD simulations were performed using AMBER v16 [70] and the ff14SB force field [71]. Simulations were performed at 300 K under constant energy (NVE) conditions using a 2 fs timestep. The peptide was solvated using the SPC(E) water model in a rectangular box with periodic boundary conditions. The system was neutralized through the addition of counter ions (Na+). The pre-processing

steps were followed by equilibration steps, as described previously [72]. All simulations were performed using the GPU-enabled version of the AMBER simulation engine pmemd. A Particle Mesh Ewald cut-off of 8 Å was used for the GPU-enabled simulations [73]. The peptide was simulated for a total of 100 ns. Representative structures were selected by clustering simulation ensembles obtained from the MD simulation trajectory. Clustering was performed using the hierarchical agglomerative approach with an epsilon cutoff of 3 Å, which represents the minimum distance between the clusters.

Peptide synthesis and functional test

Linear peptide precursor of LV-4130 was synthesized by Fmoc solid phase methodology on a Liberty microwave assisted automated peptide synthesizer (CEM, Saclay, France) using the standard manufacturer's procedure at 0.1 mmol scale on a preloaded Fmoc-Asp(Wang resin)-OAl as previously described [74]. Reactive side chains were protected as follow: Thr, Tyr, tert-butyl (tBu) ether; Glu, tert-butyl (OtBu) ester; Lys, Trp, tert-butyloxycarbonyl (Boc) carbamate; Cys, p-methoxytrityl (Mmt) thioether. After completion of the chain assembly, deprotection of the allyl ester was performed manually. A solution of $PheSiH_3$ (24 equiv) in DCM (1.3 mL) was added to the H-peptidyl(resin)-OAl using an Ar flushed gas-tight syringe and gently agitated at room temperature. The $Pd(PPh_3)_4$ catalyst (0.3 equiv) in DCM (3.9 mL) was added and the mixture was stirred for 1 hour. The resin was then washed sequentially with sodium diethyldithio-carbamate (0.02 M in DMF), DMF and DCM, and dried in vacuo. Head-to-tail cyclisation was performed on-resin by *in situ* activation of the free carboxyl group with HATU (5 eq), HOAt (5 eq) and DiEA (10 eq) in 10 mL of DMF, overnight at room temperature. The disulfide bridge was then formed on-resin by selective deprotection of the Mmt group and subsequent treatment with N-chlorosuccinimide (NCS) as previously described [75]. Briefly, the resin-bound cyclopeptide was treated five times for 2 min with a solution of 2% trifluoroacetic acid (TFA) in DCM (5 mL) and washed with DCM. A solution of NCS (2 eq) in DMF (10 mL) was added and left at room temperature for 15 min, then the resin was washed with DMF and DCM. Finally, the bicyclic peptide was deprotected and cleaved from the resin by adding 10 ml of the mixture TFA/TIS/H_2O (9.5:0.25:0.25) for 180 min at room temperature. After filtration, crude peptide was washed thrice by precipitation in TBME followed by centrifugation (4500 rpm, 15 min). The synthetic peptide was purified by reversed-phase HPLC on a 21.2 x 250 mm Jupiter C18 (5 μm, 300 Å) column (Phenomenex, Le Pecq, France) using a linear gradient (30-80% over 45 min) of acetonitrile/TFA (99.9:0.1) at a flow rate of 10 mL/min. The purified peptides were then characterized by MALDI-TOF mass spectrometry on a ultrafleXtreme (Bruker, Strasbourg, France) in the reflector mode using α-cyano-4-hydroxycinnamic acid as a matrix. Analytical RP-HPLC, performed on a 4.6 x 250 mm Jupiter C18 (5 μm, 300 Å) column, indicated that the purity of the peptide was >99%.

Intracellular calcium assay

Ligand-stimulated intracellular calcium responses were measured at the human UT receptor expressed in transfected CHO cells using a fluorimetric detection method according to Eurofins-Cerep standard assay protocols (catalog ref. G099-1376). The assays were performed in duplicate. The results were expressed as a percent of the control response to 100 nM human UII and plotted using Prism software (GraphPad, San Diego, CA).

Acknowledgements

ANR-10-BINF-0003 (BipBip); ANR-14-2011-IFB; INSERM [U 1133]; Ressource Parisienne en Bioinformatique Structurale (RPBS). The authors thank David Chatenet, Nicolas Doucet and Chitra Narayanan, INRS, Quebec, Canada for communicating sample conformations of UII.

Competing interests

The authors declare no competing interests.

References

1. A. Zorzi, K. Deyle, and C. Heinis. Cyclic peptide therapeutics: past, present and future. *Curr Opin Chem Biol*, 38:24–29, Jun 2017.

2. A. Falanga, E. Nigro, M. G. De Biasi, A. Daniele, G. Morelli, S. Galdiero, and O. Scudiero. Cyclic Peptides as Novel Therapeutic Microbicides: Engineering of Human Defensin Mimetics. *Molecules*, 22(7), 07 2017.

3. N L Daly and D J Craik. Acyclic permutants of naturally occurring cyclic proteins characterization of cystine knot and β-sheet formation in the macrocyclic polypeptide kalata b1. *Journal of Biological Chemistry*, 275(25): 19068–19075, 2000.

4. D. G. Barry, N. L. Daly, R. J. Clark, L. Sando, and D. J. Craik. Linearization of a naturally occurring circular protein maintains structure but eliminates hemolytic activity. *Biochemistry*, 42(22):6688–6695, 2003.

5. Y. Zhang, Q. Zhang, C. T. T. Wong, and X. Li. Chemoselective Peptide Cyclization and Bicyclization Directly on Unprotected Peptides. *J. Am. Chem. Soc.*, Jul 2019.

6. G. Bhardwaj, V. K. Mulligan, C. D. Bahl, J. M. Gilmore, P. J. Harvey, O. Cheneval, G. W. Buchko, S. V. Pulavarti, Q. Kaas, A. Eletsky, P. S. Huang, W. A. Johnsen, P. J. Greisen, G. J. Rocklin, Y. Song, T. W. Linsky, A. Watkins, S. A. Rettie, X. Xu, L. P. Carter, R. Bonneau, J. M. Olson, E. Coutsias, C. E. Correnti, T. Szyperski, D. J. Craik, and D. Baker. Accurate de novo design of hyperstable constrained peptides. *Nature*, 538(7625):329–335, Oct 2016.

7. C. K. Wang and D. J. Craik. Designing macrocyclic disulfide-rich peptides for biotechnological applications. *Nat Chem Biol*, 14(5):417–427, 05 2018.

8. P. Lindholm, U. Göransson, S. Johansson, P. Claeson, J. Gullbo, R. Larsson, L. Bohlin, and A. Backlund. Cyclotides: a novel type of cytotoxic agents. *Mol Cancer Ther*, 1(6):365–369, Apr 2002.

9. J. Tang, C. K. Wang, X. Pan, H. Yan, G. Zeng, W. Xu, W. He, N. L. Daly, D. J. Craik, and N. Tan. Isolation and characterization of cytotoxic cyclotides from Viola tricolor. *Peptides*, 31(8):1434–1440, Aug 2010.

10. James P Tam, Yi-An Lu, Jin-Long Yang, and Kou-Wei Chiu. An unusual structural motif of antimicrobial peptides containing end-to-end macrocycle and cystine-knot disulfides. *Proceedings of the National Academy of Sciences*, 96(16):8913–8918, 1999.

11. Lorents Gran, Knut Sletten, and Lars Skjeldal. Cyclic peptides from oldenlandia affinis dc. molecular and biological properties. *Chemistry & biodiversity*, 5(10):2014–2022, 2008.

12. Michelle L Colgrave, Andrew C Kotze, David C Ireland, Conan K Wang, and David J Craik. The anthelmintic activity of the cyclotides: natural variants with enhanced activity. *Chembiochem*, 9(12):1939–1945, 2008.

13. Michelle L Colgrave, Andrew C Kotze, Yen-Hua Huang, John O?Grady, Shane M Simonsen, and David J Craik. Cyclotides: natural, circular plant peptides that possess significant activity against gastrointestinal nematode parasites of sheep. *Biochemistry*, 47(20):5581–5589, 2008.

14. R. Dahiya, S. Singh, A. Sharma, S. V. Chennupati, and S. Maharaj. First Total Synthesis and Biological Screening of a Proline-Rich Cyclopeptide from a Caribbean Marine Sponge. *Mar Drugs*, 14(12), Dec 2016.

15. X. Jing and K. Jin. A gold mine for drug discovery: Strategies to develop cyclic peptides into therapies. *Med Res Rev*, 40(2):753–810, Mar 2020.

16. Ali Tavassoli. Siclopps cyclic peptide libraries in drug discovery. *Current Opinion in Chemical Biology*, 38: 30–35, 2017.

17. Toby Passioura, Takayuki Katoh, Yuki Goto, and Hiroaki Suga. Selection-based discovery of druglike macrocyclic peptides. *Annual review of biochemistry*, 83:727–752, 2014.

18. Diana P Slough, Sean M McHugh, Ashleigh E Cummings, Peng Dai, Bradley L Pentelute, Joshua A Kritzer, and Yu-Shan Lin. Designing well-structured cyclic pentapeptides based on sequence–structure relationships. *The Journal of Physical Chemistry B*, 122(14):3908–3919, 2018.

19. Jiabo Li, Tedman Ehlers, Jon Sutter, Shikha Varma-O'Brien, and Johannes Kirchmair. Caesar: a new conformer generation algorithm based on recursive buildup and local rotational symmetry consideration. *Journal of chemical information and modeling*, 47(5):1923–1932, 2007.

20. Paul CD Hawkins, A Geoffrey Skillman, Gregory L Warren, Benjamin A Ellingson, and Matthew T Stahl. Conformer generation with omega: algorithm and validation using high quality structures from the protein databank and cambridge structural database. *Journal of chemical information and modeling*, 50(4):572–584, 2010.

21. Fergal J Duffy, Mélanie Verniere, Marc Devocelle, Elise Bernard, Denis C Shields, and Anthony J Chubb. Cyclops: generating virtual libraries of cyclized and constrained peptides including nonnatural amino acids. *Journal of chemical information and modeling*, 51(4):829–836, 2011.

22. M. P. Jacobson, D. L. Pincus, C. S. Rapp, T. JF Day, B. Honig, D. E. Shaw, and R. A. Friesner. A hierarchical approach to all-atom protein loop prediction. *Proteins: Structure, Function, and Bioinformatics*, 55(2):351–367, 2004.

23. Daniel J Mandell, Evangelos A Coutsias, and Tanja Kortemme. Sub-angstrom accuracy in protein loop reconstruction by robotics-inspired conformational sampling. *Nature methods*, 6(8):551–552, 2009.

24. Annick Thomas, Sébastien Deshayes, Marc Decaffmeyer, Marie Hélène Van Eyck, Benoit Charloteaux, and Robert Brasseur. Prediction of peptide structure: how far are we? *Proteins: Structure, Function, and Bioinformatics*, 65(4):889–897, 2006.

25. Jérôme Beaufays, Laurence Lins, Annick Thomas, and Robert Brasseur. In silico predictions of 3d structures of linear and cyclic peptides with natural and non-proteinogenic residues. *Journal of Peptide Science*, 18(1): 17–24, 2012.

26. Sandeep Singh, Harinder Singh, Abhishek Tuknait, Kumardeep Chaudhary, Balvinder Singh, S Kumaran, and Gajendra PS Raghava. Pepstrmod: structure prediction of peptides containing natural, non-natural and modified residues. *Biology direct*, 10(1):73, 2015.

27. Pierre Thevenet, Yimin Shen, Julien Maupetit, Frédéric Guyon, Philippe Derreumaux, and Pierre Tuffery. Pepfold: an updated de novo structure prediction server for both linear and disulfide bonded cyclic peptides. *Nucleic acids research*, 40(W1):W288–W293, 2012.

28. Yimin Shen, Julien Maupetit, Philippe Derreumaux, and Pierre Tufféry. Improved pep-fold approach for peptide and miniprotein structure prediction. *Journal of chemical theory and computation*, 10(10):4745–4758, 2014.

29. Sean M McHugh, Julia R Rogers, Sarah A Solomon, Hongtao Yu, and Yu-Shan Lin. Computational methods to design cyclic peptides. *Current opinion in chemical biology*, 34:95–102, 2016.

30. A. G. Poth, L. Y. Chan, and D. J. Craik. Cyclotides as grafting frameworks for protein engineering and drug design applications. *Biopolymers*, 100(5):480–491, Sep 2013.

31. Eric Valeur, Stéphanie M Guéret, Hélène Adihou, Ranganath Gopalakrishnan, Malin Lemurell, Herbert Waldmann, Tom N Grossmann, and Alleyn T Plowright. New modalities for challenging targets in drug discovery. *Angewandte Chemie International Edition*, 56(35):10294–10323, 2017.

32. Richard J Clark, Harald Fischer, Louise Dempster, Norelle L Daly, K Johan Rosengren, Simon T Nevin, Frederic A Meunier, David J Adams, and David J Craik. Engineering stable peptide toxins by means of backbone cyclization: stabilization of the α-conotoxin mii. *Proceedings of the National Academy of Sciences*, 102(39): 13767–13772, 2005.

33. Christopher J Armishaw, Julie L Dutton, David J Craik, and Paul F Alewood. Establishing regiocontrol of disulfide bond isomers of α-conotoxin imi via the synthesis of n-to-c cyclic analogs. *Peptide Science: Original Research on Biomolecules*, 94(3):307–313, 2010.

34. Richard J Clark, Jonas Jensen, Simon T Nevin, Brid P Callaghan, David J Adams, and David J Craik. The engineering of an orally active conotoxin for the treatment of neuropathic pain. *Angewandte Chemie International Edition*, 49(37):6545–6548, 2010.

35. Reena Halai, Brid Callaghan, N. L. Daly, Richard J Clark, David J Adams, and David J Craik. Effects of cyclization on stability, structure, and activity of α-conotoxin rgia at the $\alpha9\alpha10$ nicotinic acetylcholine receptor and gabab receptor. *Journal of medicinal chemistry*, 54(19):6984–6992, 2011.

36. C. J. Armishaw, A. A. Jensen, L. D. Balle, K. C. Scott, L. Sérensen, and K. Str?mgaard. Improving the stability of α-conotoxin AuIB through N-to-C cyclization: the effect of linker length on stability and activity at nicotinic acetylcholine receptors. *Antioxid. Redox Signal.*, 14(1):65–76, Jan 2011.

37. E. S. Lovelace, S. Gunasekera, C. Alvarmo, R. J. Clark, S. T. Nevin, A. A. Grishin, D. J. Adams, D. J. Craik, and N. L. Daly. Stabilization of α-conotoxin AuIB: influences of disulfide connectivity and backbone cyclization. *Antioxid. Redox Signal.*, 14(1):87–95, Jan 2011.

38. E. S. Lovelace, C. J. Armishaw, M. L. Colgrave, M. E. Wahlstrom, P. F. Alewood, N. L. Daly, and D. J. Craik. Cyclic MrIA: a stable and potent cyclic conotoxin with a novel topological fold that targets the norepinephrine transporter. *J. Med. Chem.*, 49(22):6561–6568, Nov 2006.

39. Z. Dekan, C. I. Wang, R. K. Andrews, R. J. Lewis, and P. F. Alewood. Conotoxin engineering: dual pharmacophoric noradrenaline transport inhibitor/integrin binding peptide with improved stability. *Org. Biomol. Chem.*, 10(30):5791–5794, Aug 2012.

40. X. Hemu, M. Taichi, Y. Qiu, D. X. Liu, and J. P. Tam. Biomimetic synthesis of cyclic peptides using novel thioester surrogates. *Biopolymers*, 100(5):492–501, Sep 2013.

41. Muharrem Akcan, Richard J Clark, Norelle L Daly, Anne C Conibear, Andrew de Faoite, Mari D Heghinian, Talwar Sahil, David J Adams, Frank Marí, and David J Craik. Transforming conotoxins into cyclotides: Backbone cyclization of p-superfamily conotoxins. *Peptide Science*, 104(6):682–692, 2015.

42. Y. Karami, F. Guyon, S. De Vries, and P. Tufféry. DaReUS-Loop: accurate loop modeling using fragments from remote or unrelated proteins. *Sci Rep*, 8(1):13673, Sep 2018.

43. Y. Karami, J. Rey, G. Postic, S. Murail, P. Tufféry, and S. J. de Vries. DaReUS-Loop: a web server to model multiple loops in homology models. *Nucleic Acids Res.*, 47(W1):W423–W428, Jul 2019.

44. Amelie Stein and Tanja Kortemme. Improvements to robotics-inspired conformational sampling in rosetta. *PLoS One*, 8(5):e63090, 2013.

45. Y. Coulouarn, I. Lihrmann, S. Jegou, Y. Anouar, H. Tostivint, J. C. Beauvillain, J. M. Conlon, H. A. Bern, and H. Vaudry. Cloning of the cDNA encoding the urotensin II precursor in frog and human reveals intense expression of the urotensin II gene in motoneurons of the spinal cord. *Proc Natl Acad Sci U S A*, 95(26): 15803–15808, Dec 1998.

46. H. Vaudry, J. Leprince, D. Chatenet, A. Fournier, D. G. Lambert, J. C. Le Mevel, E. H. Ohlstein, A. Schwertani, H. Tostivint, and D. Vaudry. International Union of Basic and Clinical Pharmacology. XCII. Urotensin II, urotensin II-related peptide, and their receptor: from structure to function. *Pharmacol Rev*, 67(1):214–258, 2015.

47. J. Leprince, D. Chatenet, C. Dubessy, A. Fournier, B. Pfeiffer, E. Scalbert, P. Renard, P. Pacaud, H. Oulyadi, I. Segalas-Milazzo, L. Guilhaudis, D. Davoust, M. C. Tonon, and H. Vaudry. Structure-activity relationships of urotensin II and URP. *Peptides*, 29(5):658–673, May 2008.

48. E. Cadirci, R. A. Ugan, B. Dincer, B. Gundogdu, I. Cinar, E. Akpinar, and Z. Halici. Urotensin receptors as a new target for CLP induced septic lung injury in mice. *Naunyn Schmiedebergs Arch Pharmacol*, 392(2):135–145, 02 2019.

49. A. Tapeinou, M. T. Matsoukas, C. Simal, and T. Tselios. Review cyclic peptides on a merry-go-round; towards drug design. *Biopolymers*, 104(5):453–461, Sep 2015.

50. R. Bhaskaran, A. I. Arunkumar, and C. Yu. NMR and dynamical simulated annealing studies on the solution conformation of urotensin II. *Biochim Biophys Acta*, 1199(2):115–122, Mar 1994.

51. S. Flohr, M. Kurz, E. Kostenis, A. Brkovich, A. Fournier, and T. Klabunde. Identification of nonpeptidic urotensin II receptor antagonists by virtual screening based on a pharmacophore model derived from structure-activity relationships and nuclear magnetic resonance studies on urotensin II. *J Med Chem*, 45(9):1799–1805, Apr 2002.

52. E. Lescot, J. Sopkova-de Oliveira Santos, C. Dubessy, H. Oulyadi, A. Lesnard, H. Vaudry, R. Bureau, and S. Rault. Definition of new pharmacophores for nonpeptide antagonists of human urotensin-II. Comparison with the 3D-structure of human urotensin-II and URP. *J Chem Inf Model*, 47(2):602–612, 2007.

53. A. Carotenuto, P. Grieco, P. Campiglia, E. Novellino, and P. Rovero. Unraveling the active conformation of urotensin II. *J Med Chem*, 47(7):1652–1661, Mar 2004.

54. Jason P Mulvenna, Conan Wang, and David J Craik. Cybase: a database of cyclic protein sequence and structure. *Nucleic acids research*, 34(suppl_1):D192–D194, 2006.

55. Conan KL Wang, Quentin Kaas, Laurent Chiche, and David J Craik. Cybase: a database of cyclic protein sequences and structures, with applications in protein discovery and engineering. *Nucleic acids research*, 36 (suppl_1):D206–D210, 2007.

56. Xiaosa Wu, Yen-Hua Huang, Quentin Kaas, and David J Craik. Cyclisation of disulfide-rich conotoxins in drug design applications. *European Journal of Organic Chemistry*, 2016(21):3462–3472, 2016.

57. C. L. Herold, D. J. Behm, P. T. Buckley, J. J. Foley, W. E. Wixted, H. M. Sarau, and S. A. Douglas. The neuromedin B receptor antagonist, BIM-23127, is a potent antagonist at human and rat urotensin-II receptors. *Br J Pharmacol*, 139(2):203–207, May 2003.

58. Helen M. Berman, John Westbrook, Zukang Feng, Gary Gilliland, T. N. Bhat, Helge Weissig, Ilya N. Shindyalov, and Philip E. Bourne. The Protein Data Bank. *Nucleic Acids Research*, 28(1):235–242, 01 2000.

59. F. Guyon, F. Martz, M. Vavrusa, J. Becot, J. Rey, and P. Tufféry. BCSearch: fast structural fragment mining over large collections of protein structures. *Nucleic Acids Res.*, 43(W1):W378–382, Jul 2015.

60. Travis E Oliphant. *A guide to NumPy*, volume 1. Trelgol Publishing USA, 2006.

61. Eric Jones, Travis Oliphant, Pearu Peterson, et al. SciPy: Open source scientific tools for Python, 2001–. URL http://www.scipy.org/.

62. Yimin Shen, Géraldine Picord, Frédéric Guyon, and Pierre Tufféry. Detecting protein candidate fragments using a structural alphabet profile comparison approach. *PloS one*, 8(11):e80493, 2013.

63. Mark James Abraham, Teemu Murtola, Roland Schulz, Szilárd Páll, Jeremy C Smith, Berk Hess, and Erik Lindahl. Gromacs: High performance molecular simulations through multi-level parallelism from laptops to supercomputers. *SoftwareX*, 1:19–25, 2015.

64. J. Huang, S. Rauscher, G. Nawrocki, T. Ran, M. Feig, B. L. de Groot, H. Grubmuller, and A. D. MacKerell. CHARMM36m: an improved force field for folded and intrinsically disordered proteins. *Nat. Methods*, 14(1): 71–73, 01 2017.

65. Alexis Lamiable, Pierre Thévenet, and Pierre Tufféry. A critical assessment of hidden markov model sub-optimal sampling strategies applied to the generation of peptide 3d models. *Journal of computational chemistry*, 37(21): 2006–2016, 2016.

66. Samuel Murail. gromacs_py: A gromacs python wrapper, October 2018. URL https://doi.org/10.5281/zenodo.1455734.

67. Jing Huang and Alexander D MacKerell Jr. Charmm36 all-atom additive protein force field: Validation based on comparison to nmr data. *Journal of computational chemistry*, 34(25):2135–2145, 2013.

68. Richard J Gowers, Max Linke, Jonathan Barnoud, Tyler John Edward Reddy, Manuel N Melo, Sean L Seyler, Jan Domanski, David L Dotson, Sébastien Buchoux, Ian M Kenney, et al. Mdanalysis: a python package for the rapid analysis of molecular dynamics simulations. Technical report, Los Alamos National Lab.(LANL), Los Alamos, NM (United States), 2019.

69. K Mahesh Kumar and A Rama Mohan Reddy. A fast dbscan clustering algorithm by accelerating neighbor searching using groups method. *Pattern Recognition*, 58:39–48, 2016.

70. DA Case, Josh Berryman, RM Betz, DS Cerutti, TE Cheatham Iii, TA Darden, RE Duke, TJ Giese, H Gohlke, AW Goetz, et al. Amber 2015. 2015.

71. James A Maier, Carmenza Martinez, Koushik Kasavajhala, Lauren Wickstrom, Kevin E Hauser, and Carlos Simmerling. ff14sb: improving the accuracy of protein side chain and backbone parameters from ff99sb. *Journal of chemical theory and computation*, 11(8):3696–3713, 2015.

72. Pratul K Agarwal. Cis/trans isomerization in hiv-1 capsid protein catalyzed by cyclophilin a: Insights from computational and theoretical studies. *Proteins: Structure, Function, and Bioinformatics*, 56(3):449–463, 2004.

73. Romelia Salomon-Ferrer, Andreas W Gotz, Duncan Poole, Scott Le Grand, and Ross C Walker. Routine microsecond molecular dynamics simulations with amber on gpus. 2. explicit solvent particle mesh ewald. *Journal of chemical theory and computation*, 9(9):3878–3888, 2013.

598 74. Axel Touchard, Samira R Aili, Nathan Téné, Valentine Barassé, Christophe Klopp, Alain Dejean, R Manjunatha
599 Kini, Mrinalini, Laurent Coquet, Thierry Jouenne, et al. Venom peptide repertoire of the european myrmicine
600 ant manica rubida: identification of insecticidal toxins. *Journal of Proteome Research*, 19(4):1800–1811, 2020.

601 75. Tobias M Postma and Fernando Albericio. N-chlorosuccinimide, an efficient reagent for on-resin disulfide for-
602 mation in solid-phase peptide synthesis. *Organic letters*, 15(3):616–619, 2013.

Supplementary Materials

Database search

We previously introduced the BCLoopSearch protocol, to mine large protein structure datasets and retrieve loop candidates, given two disjoint fragments (loop flanks) [59]. It is based on a Binet-Cauchy (BC) kernel and a Rigidity score:

$$BC(X,Y) = \frac{det(X^T Y)}{\sqrt{det(X^T X) det(Y^T Y)}} \tag{4}$$

where X and Y are C_α coordinates of the flanks and dataset fragments, respectively and they are centered at the origin. Note that a BC score of 1 indicates a perfect match. *Rigidity* score $R(X,Y)$ is defined as:

$$R'(X,Y) = max_{1 \leq i \leq N} |\|X_i - Y_i\|| \tag{5}$$

$$R(X,Y) = max\big\{R'(X,Y), |\|X_N - X_1\| - \|Y_N - Y_1\|| \big\} \tag{6}$$

where X_i and Y_i are C_α coordinates of the ith residues of the flanks and dataset fragments and $||.||$ is the euclidean norm. Rigidity score is the maximum variation of intra-distances between: *(i)* residues and geometric center and *(ii)* intra-distances between terminal C_α. In addition, we also measured the RMSD between query and candidate flanks for the fragments returned. In total, four cut-offs values related to *(i)* flank size, *(ii)* flank BC score, *(iii)* flank Rigidity and *(iv)* flank RMSD, have been considered to limit the number of loop candidates. In this study we used: a flank size of 4 residues, Rigidity ≤ 2.5, flank RMSD ≤ 4 Å and the minimal flank BC score cut-off of 0.8.

CyBase benchmark

CyBase (http://www.cybase.org.au/) [54, 55] provides a set of existing naturally occurring cyclic peptides. Presently, 64 3D structures of cyclic peptides from 25 different species are reported. We applied a filtering step on the list to keep only those that are *i)* head-to-tail cyclized, *ii)* without modified amino acids and *iii)* not identical (filtering out entries with identical sequences), resulting in a final set of 35 structures. Residues from the N- and/or C-terminal extremities of each cyclic peptide were removed to generate linear peptides (here by N- and C-terminal extremity, we refer to the head and tail residues from the sequence). We considered all possible combinations of truncating two to seven residues from the N- and/or C-termini (*i.e.*, removing two residues from N-terminus or two residues from C-terminus or one residue from each side), generating 33 different linear peptides from every cyclic target. We also excluded the cases where the size of generated linear peptide was less than 8 residues, that is size limit of our protocol. Finally we obtained a total of 1147 linear peptides, where the corresponding linkers are in the range of 2-7 residue long. The details of those structures are reported in **Supplementary Table S1**.

Table S1: **The list of cyclic structures from CyBase.** Structures with identical sequences were discarded and only one representative was considered. For each cyclic peptide, we generated a total of 33 linear peptides by truncating two to seven residues from N- and/or C-term. The total number of linear peptides for each target, as well as those modelled with PEP-Cyclizer and Rosetta NGK are reported.

Protein name	Class	Type	PDBcode	size	#linkers	#linkers (PEP-Cyclizer)	#linkers (NGK)
kalata-B1	Cyclotide	NMR	1NB1	29	33	33	33
kalata-B1	Cyclotide	NMR	1K48	29	33	33	32
kalata-B1	Cyclotide	NMR	1KAL	29	33	33	32
[P20D,V21K]-kalata-B1	Cyclotide	NMR	2F2I	29	33	33	33
[W19K,-P20N,-V21K]-kalata-B1	Cyclotide	NMR	2F2J	29	33	33	33
kalata-B2	Cyclotide	NMR	1PT4	29	33	33	33
kalata-B5	Cyclotide	NMR	2KUX	30	33	32	33
kalata-B7	Cyclotide	NMR	2JWM	29	33	33	33
kalata-B7	Cyclotide	NMR	2M9O	29	33	33	33
kalata-B8	Cyclotide	NMR	2B38	31	33	33	33
kalata-B12	Cyclotide	NMR	2KVX	28	33	32	33
cycloviolacin-O1	Cyclotide	NMR	1NBJ	30	33	32	33
cycloviolacin-O1	Cyclotide	NMR	1DF6	30	33	33	33
cycloviolacin-O2	Cyclotide	NMR	2KNM	30	33	32	33
cycloviolacin-O14	Cyclotide	NMR	2GJ0	31	33	33	33
MCoTI-II	Squash-trypsin-inhibitor	XRAY	4GUX	34	33	33	33
circulin-A	Cyclotide	NMR	1BH4	30	33	32	33
circulin-B	Cyclotide	NMR	2ERI	31	33	33	33
kB1[GHFRWG;23-28]	Cyclotide	NMR	2LUR	29	33	32	32
[Ala1,15]kB1	Cyclotide	NMR	1N1U	29	33	33	33
des(24-28)kB1	Cyclotide	NMR	1ORX	24	33	33	32
SFTI-1	BBI-like-trypsin-inhibitor	XRAY	3P8F	14	25	25	25
Ent-AS-48	Bacterial	XRAY	1O82	70	33	33	33
vhl-1	Cyclotide	NMR	1ZA8	31	33	33	33
vhl-2	Cyclotide	NMR	2KUK	30	33	33	33
varv-peptide-F	Cyclotide	NMR	2K7G	29	33	33	33
varv-peptide-F	Cyclotide	XRAY	3E4H	29	33	33	33
BiKK	BBI-like-trypsin-inhibitor	NMR	2BEY	16	33	33	33
RTD-1	Primate	NMR	1HVZ	18	33	33	33
palicourein	Cyclotide	NMR	1R1F	37	33	33	33
vhr1	Cyclotide	NMR	1VB8	30	33	31	33
tricyclon-A	Cyclotide	NMR	1YP8	33	33	33	33
Cter-M	Cyclotide	NMR	2LAM	29	33	33	33
MCo-PMI	Squash-trypsin-inhibitor	NMR	2M86	51	33	33	31
Carnocyclin-A	Bacterial	NMR	2KJF	60	33	33	33
total					1147	1141	1141

We applied both our protocol and Rosetta NGK to the CyBase test set to model all the linkers. Over the 1147 cases,

both our data-mining and Rosetta NGK failed to model the linker for 6 different cases (0.5%). In fact, our protocol identified candidates in all cases, but discarded all the candidates with a correct geometry but a non satisfactory sequence similarity in 6 cases. Thus overall, in terms of ability to identify linkers, the data-mining strategy seems to perform as well as a pure *ab initio* procedure. Then, we compared both protocols using the 1135 over 1147 (99%) cases for which the linker could be modelled by both methods. All heavy backbone atoms (N, C, Cα, O) were considered. The local RMSD corresponds to RMSD obtained by superposing the model linker on the native conformation using a best fit procedure, whereas the global RMSD corresponds to RMSD observed after superposing the linear part of the peptide (*i.e.*, without the linker). The best RMSD over the top 20 predictions by each method were retained.

Table S2: **The RMSD values for all the linkers of each structure from CyBase.** The average local and global RMSD values are measured over the backbone atoms (N, C, Cα, O) for the linkers modelled by both PEP-Cyclizer and Rosetta NGK. For each cyclic peptide, we generated a total of 33 linear peptides by truncating two to seven residues from N- and/or C-terminal extremities. For each target, the number of linear peptides that were cyclized by both PEP-Cyclizer and Rosetta NGK are reported (out of the total 33 linkers).

Protein name	**number of linkers**	local RMSD (Å)		global RMSD (Å)	
		PEP-Cyclizer	**NGK**	**PEP-Cyclizer**	**NGK**
kalata-B1	33	0.53±0.22	0.56±0.37	1.17±0.45	1.05±0.46
kalata-B1	32	0.69±0.34	0.70±0.57	2.05±1.40	1.77±2.43
kalata-B1	32	0.71±0.26	1.04±0.55	1.45±0.47	1.71±0.96
[P20D,V21K]-kalata-B1	33	0.65±0.37	0.52±0.38	1.40±0.70	0.99±0.77
[W19K,-P20N,-V21K]-kalata-B1	33	0.54±0.28	0.71±0.40	1.31±0.70	1.24±0.65
kalata-B2	33	0.49±0.17	0.40±0.38	1.12±0.47	0.73±0.66
kalata-B5	32	0.74±0.45	0.45±0.57	1.83±1.53	0.88±1.54
kalata-B7	33	0.58±0.19	0.82±0.65	1.15±0.32	1.13±0.73
kalata-B7	33	0.79±0.36	0.56±0.58	2.11±1.52	1.45±2.06
kalata-B8	33	1.12±0.57	1.14±0.61	2.57±1.13	2.11±1.22
kalata-B12	32	0.74±0.38	0.69±0.37	1.57±1.07	1.40±1.32
cycloviolacin-O1	32	1.08±0.68	0.43±0.17	2.83±1.98	0.91±0.33
cycloviolacin-O1	33	0.98±0.43	1.15±0.43	2.14±1.24	1.99±1.17
cycloviolacin-O2	32	0.75±0.43	0.38±0.39	2.07±1.76	0.75±1.30
cycloviolacin-O14	33	0.94±0.57	0.79±0.89	2.61±1.91	2.01±2.54
MCoTI-II	33	0.73±0.43	0.31±0.44	1.70±0.78	0.64±1.13
circulin-A	33	0.97±0.34	0.88±0.34	2.37±0.96	1.59±0.87
circulin-B	33	1.02±0.44	0.37±0.17	2.22±0.91	0.61±0.30
kB1[GHFRWG;23-28]	32	1.08±0.45	1.16±0.40	2.61±1.23	2.18±0.87
[Ala1,15]kB1	33	0.93±0.39	0.88±0.33	1.94±0.73	1.22±0.51
des(24-28)kB1	32	1.34±0.55	1.60±0.64	2.44±0.81	2.88±1.37
SFTI-1	25	0.50±0.41	0.22±0.14	1.48±1.19	0.57±0.48
Ent-AS-48	33	0.54±0.29	0.14±0.08	1.16±0.53	0.23±0.10
vhl-1	33	0.87±0.46	0.36±0.27	1.88±1.01	0.66±0.32
vhl-2	33	0.53±0.25	0.71±0.74	1.24±0.52	1.11±0.95
varv-peptide-F	33	0.61±0.42	0.51±0.64	1.86±1.62	1.24±2.12
varv-peptide-F	33	0.50±0.21	0.41±0.28	1.17±0.52	0.79±0.35
BiKK	33	0.64±0.47	1.05±0.75	1.64±1.27	1.87±1.36
RTD-1	33	0.72±0.37	0.74±0.43	1.98±0.84	1.77±0.94
palicourein	33	1.12±0.36	1.11±0.32	2.24±1.02	1.93±0.59
vhr1	31	0.97±0.58	0.61±0.34	2.36±1.59	1.23±0.48
tricyclon-A	33	0.79±0.30	0.68±0.48	1.64±0.57	1.13±0.86
Cter-M	33	0.60±0.42	0.54±0.69	1.62±1.43	1.45±2.08
MCo-PMI	31	1.11±0.48	1.09±0.51	2.33±1.22	1.93±1.19
Carnocyclin-A	33	0.52±0.26	0.23±0.13	1.21±0.47	0.48±0.22

linker size (# gaps)		2 (101)	3 (139)	4 (175)	5(208)	6 (241)	7 (271)
$lRMSD_{20}$	PEP-Cyclizer	0.32±0.19	0.51±0.22	0.66±0.28	0.77±0.35	0.92±0.44	1.11±0.49
	Rosetta NGK	0.26±0.33	0.38±0.36	0.52±0.43	0.66±0.48	0.83±0.54	1.04±0.66
$gRMSD_{20}$	PEP-Cyclizer	1.38±0.57	1.45±0.50	1.52±0.59	1.64±0.78	2.02±1.25	2.55±1.70
	Rosetta NGK	0.73±0.78	0.84±0.70	0.97±0.68	1.15±0.81	1.60±1.51	1.93±1.74
$lRMSD_1$	PEP-Cyclizer	0.64±0.29	0.94±0.39	1.14±0.50	1.21±0.57	1.43±0.67	1.79±0.80
	Rosetta NGK	0.34±0.35	0.55±0.53	0.73±0.61	0.89±0.63	1.12±0.72	1.35±0.80
$gRMSD_1$	PEP-Cyclizer	2.67±1.40	2.80±1.47	3.06±1.85	3.06±2.18	3.99±2.67	4.90±3.14
	Rosetta NGK	0.89±0.89	1.14±0.99	1.44±1.47	1.68±1.74	2.18±1.87	2.65±2.25

Table S3: **RMSD and ranks over the CyBase test set.** For every case the best RMSD out of top 20 and the top 1 were considered. The average and standard deviations of best local ($lRMSD_{20}$, $lRMSD_1$) and global ($gRMSD_{20}$, $gRMSD_1$) RMSD values are reported for every gap size.

Table S4: **Summary of cyclic linkers for conotoxins.** Data is collected from [56] and additional details are added from the mentioned references. The last column reports the pdb code of the available engineered cyclic peptides. linkers sequences in bold correspond to the functional variants that were considered in this study.

name	linear peptide	linker	activity	structure	stability	pdb code
α-Conotoxin MII	1m2c (1mii)	GGAAG (cMII-5) [32]	not active	not similar	-	-
		GGAAGG (cMII-6) [32]	similar	similar	improved	**2ajw**
		GAGGAAG (cMII-7) [32]	smiliar	similar	improved	**2ak0**
α-Conotoxin ImI	1cnl	A [33]	-	-	slightly improved	-
		βA [33]	-	-	improved	-
		AG [33]	-	-	slightly improved	-
		AGG [33]	-	-	slightly improved	-
α-Conotoxin Vc1.1	2h8s	GGAAG [34]	substantial loss	similar	-	-
		GGAAGG [34]	similar/higher	similar	improved	**4ttl**
α-Conotoxin RgIA	2JUT	GAA [35]	reduced	not similar	-	-
		GAAG [35]	reduced	not similar	-	-
		GAAGG [35]	reduced	similar	-	-
		GGAAGG [35]	similar	similar	improved	-
		GGAAGAG [35]	similar	similar	improved	-
α-Conotoxin AuIB	1mxn 1mxp	A [36]	reduced	-	-	-
		AG [36]	reduced	-	improved	-
		AGG [36]	reduced	-	improved	-
		AGGG [36]	reduced	-	improved	-
		GGAAG [36]	reduced	-	improved	-
		GAGAAG [36]	reduced	-	improved	-
		GGAGGAG [36]	reduced	-	improved	-
		GGAA [37]	reduced	similar	improved	-
		AGAGA [37]	reduced	similar	improved	-
		GGAAGG [37]	reduced	similar	improved	-
		GGAAAGG [37]	reduced	-	improved	-
χ-Conotoxin MrIA	2ew4	**AG** [38]	similar	similar	improved	2j15
		RGD [39]	similar	similar	improved	-
ω-Conotoxin MVIIA	1mvi	**GGPG** [40]	-	-	-	-
Conotoxin gm9a	1ixt	**GLP** [41]	-	similar	similar	2mso
Conotoxin bru9a	-	**GLP** [41]	-	-	similar	2msq

Table S5: **Average ranks of the cyclic linkers for conotoxins, using forward-backtrack algorithm.**

name	linear peptide	linker	ranks
α-Conotoxin MII	1m2c (1mii)	GGAAG (cMII-5)	19/32
		GGAAGG (cMII-6)	40/64
		GAGGAAG (cMII-7)	8/128
α-Conotoxin ImI	1cnl	A	-
		βA	-
		AG	2/4
		AGG	3/8
α-Conotoxin Vc1.1	2h8s	GGAAG	23/32
		GGAAGG	33/64
α-Conotoxin RgIA	2JUT	GAA	8/8
		GAAG	14/16
		GAAGG	28/32
		GGAAGG	24/64
		GGAAGAG	21/128
α-Conotoxin AuIB	1mxn 1mxp	A	-
		AG	1/4
		AGG	2/8
		AGGG	6/16
		GGAAG	25/32
		GAGAAG	24/64
		GGAGGAG	8/128
		GGAA	14/16
		AGAGA	3/32
		GGAAGG	37/64
		GGAAAGG	35/128
χ-Conotoxin MrIA	2ew4	**AG**	2/4
		RGD	7/27
ω-Conotoxin MVIIA	1mvi	**GGPG**	3/16
Conotoxin gm9a	1ixt	**GLP**	8/27
Conotoxin bru9a	-	**GLP**	-

Table S6: **The RMSD between the 7 UII models generated by MD (M1-M7) and 5 UII models generated by PEP-FOLD (M1-M5), used as input to PEP-Cyclizer.**

		M1											
	M2	1.07	**M2**										
	M3	1.10	0.93	**M3**									
	M4	3.20	3.34	3.15	**M4**								
MD	**M5**	2.64	2.91	2.59	1.04	**M5**							
	M6	2.23	2.52	2.25	1.45	0.77	**M6**						
	M7	2.49	2.70	2.43	1.38	1.00	1.15	**M7**					
	M8	2.53	2.81	2.50	1.70	1.22	1.26	0.83	**M8**				
	M1	1.98	2.34	2.03	1.96	1.36	1.16	1.45	1.69	**M1**			
	M2	1.92	2.18	2.16	3.42	2.85	2.53	2.93	2.86	2.47	**M2**		
PEP-FOLD	**M3**	1.79	2.16	1.97	2.98	2.27	2.00	2.49	2.38	1.87	1.70	**M3**	
	M4	2.32	2.52	2.64	3.44	3.11	2.78	2.65	2.73	2.79	2.67	2.99	**M4**
	M5	2.07	2.51	2.49	4.12	3.44	3.09	3.35	3.36	2.97	2.12	2.30	2.50