Conformational Design of Cyclic Peptides

Bethany Atkinson

Degree in Chemistry

Submitted in fulfilment of the requirements for the Degree of Doctor of Philosophy

School of Chemistry

College of Science and Engineering

University of Glasgow

June 2022

Abstract

Due to their potential importance as drug molecules as well as other applications methods to design small cyclic peptides with a rigid well-defined conformation are useful. Computational methods to predict the conformation of cyclic peptides to identify those with a well-defined conformation allow for screening of potential sequences prior to expending the resources required to make the peptides. An alternative method to produce conformationally restricted cyclic peptides would be to include structural elements that prevent the peptide changing conformation. This thesis focuses on designing well-structured cyclic peptides, either through the use of computational techniques which were developed in order to help predict the structure of cyclic peptides or, through the introduction into the cyclic peptide structure of a β-turn mimic.

Chapter 1 covers current methods for the synthesis of cyclic peptides as well as methods for predicting the conformations a cyclic peptide is likely to form. β-turns, including β-turn mimics are also discussed.

Chapter 2 focuses on the modification of bias-exchange metadynamics (BE-META) simulations, which can be used to predict the conformation of cyclic peptides, to also predict the occurrence of cis proline within proline-containing cyclic peptides. An additional replica was added into the BE-META to allow for cis/trans isomerisation. A series of cyclic hexapeptides was synthesised and the cis to trans ratio of the proline within the peptides obtained by NMR. These results were then used to evaluate the computational predictions. It was found faster convergence in the simulations was reached using the additional replica, but the forcefield could not always accurately model the energy difference between the cis and trans proline states.

Chapter 3 presents the results of the analysis of β-turns found within a database. Cyclic hexapeptides are frequently observed to form a structure composed of two overlapping β-turns. It was hypothesised information on the β-turns extracted from the database could therefore be used to help design cyclic hexapeptides. Two peptides were designed based on the database analysis. The structure of the peptides, determined by NMR, show the amino acids in the peptides occupy the predicted positions in the major conformations.

Chapter 4 explores the introduction of restraints into BE-META simulations used to predict the conformations of cyclic peptides. The restrained simulations are used to infer the lowest energy structures a cyclic hexapeptide can adopt based on the backbone conformation of the peptide when a specific β-turn type is present within the peptide. The inclusion of chiral amino acids at specific positions within the cyclic peptide structure is seen to alter the most stable conformation.

In Chapter 5 a Random Forest machine learning algorithm was trained to predict the β-turn type a sequence will form based on the β-turns extracted from the database in Chapter 3. The Random Forest in combination with the lowest energy conformations determined by the restrained simulations in Chapter 4 is used to predict the conformations cyclic hexapeptides will adopt. A well-structured cyclic peptide containing the biologically active RGD motif was designed using the methods developed in this chapter. The use of the Random Forest allowed for fast filtering of potential sequences to identify those predicted to form only one major conformation.

Chapter 6 focuses on the incorporation of a β-turn mimic, which forms through a chemical ligation reaction, into cyclic peptides. Conditions were found to incorporate the β-turn mimic into the cyclic peptides which allow for the cyclisation of the peptide and formation of the β-turn mimic in a single step. The reaction has a broad sequence tolerance and was found to be suitable for peptide macrocycles of varying size.

Chapter 7 aims to introduce additional functionality to the β-turn mimic used in Chapter 6. The structure of the β-turn mimic was modified to include a fluorescent naphthalene group. A peptide containing the modified β-turn mimic was synthesised and circular dichroism (CD) analysis shows the modified β-turn mimic retains the β-turn structure. The fluorescent properties of the modified β-turn mimic were also analysed and found to be very similar to that of tryptophan.

Chapter 8 makes use of the β-turn mimic in order to design a cyclic WW Domain mimic which retains the ability to bind proline-rich ligands. Two β-strands from a WW Domain structure were cyclised using the methods developed in Chapter 6. Molecular dynamics simulations show the β-strand structure is retained in the cyclised peptide. Binding studies also demonstrate the cyclised WW domain retains the ability to bind to a ligand known to bind to the wildtype WW Domain.

Table of Contents

Acknowledgements

I would like to thank my supervisor Dr Drew Thomson. I have loved doing my PhD and (with the exception of covid) would do it all again if I could.

Additionally I would like to thank the research group for their help and support throughout this PhD. I would also like to thank Stuart for all his help and for setting up the computers so I was able to run all my simulations and Dr Brian Smith for all his help with peptide NMR.

I would also like to thank Dad and Mum for their support. I'd like to thank my sister for her mathematical advice and for being awesome in general. Becky also provided mathematical expertise and is an amazing friend. Thank you also to Caroline for always being willing to share her excessive amount of peptide-based knowledge and for fixing the HPLC all those times. Finally I'd like to thank all my family and friends that helped to keep me sane throughout the ups and downs of a PhD.

Author's Declaration

I declare that, except where explicit reference is made to the contribution of others, that the substance of this thesis is the result of my own work and has not been submitted for any other degree at the University of Glasgow or any other institution.

A portion of the work described herein has been published elsewhere as listed below:

B. C. Atkinson and A. R. Thomson, *Peptide Science*, 2022, e24266

____

Bethany Atkinson

Abbreviations

Ac	Acetyl
Ahx	Aminocaproic acid
Aib	α-Aminoisobutyric acid
Ala/A	Alanine
All	Allyl
Aloc	Allyloxycarbonyl
aMD	Accelerated molecular dynamics
Arg/R	Arginine
Asn/N	Asparagine
Asp/D	Aspartic acid
BAL	Backbone amide linker
BE-META	Bias-exchange metadynamics
BME	β-Mercaptoethanol
Boc	*tert*-Butyloxycarbonyl
BOP	(Benzotriazol-1-yloxy)tris(dimethylamino)phosphonium hexafluorophosphate
CD	Circular dichroism
COSY	Correlation spectroscopy
CV	Collective variable
Cys/C	Cysteine
DBU	1,8-Diazabicyclo(5.4.0)undec-7-ene
DCM	Dichloromethane
Dde	*N*-(1-(4,4-dimethyl-2,6-dioxocyclohexylidene)ethyl)
DFT	Density-functional theory
DIC	*N,N′*-Diisopropylcarbodiimide
DIPEA	*N,N*-Diisopropylethylamine
DMF	*N,N*-Dimethylformamide
DMSO	Dimethyl sulfoxide
EDC	1-ethyl-3-(-3-dimethylaminopropyl) carbodiimide hydrochloride
ESI	Electrospray ionisation
Fmoc	9-Fluorenylmethoxycarbonyl
GFP	Green fluorescent protein
Gln/Q	Glutamine
Glu/E	Glutamic acid
Gly/G	Glycine

GPCR	G-protein coupled receptor
HATU	1-[Bis(dimethylamino)methylene]-1H-1,2,3-triazolo[4,5-b]pyridinium 3-oxid hexafluorophosphate
HBTU	*N,N,N′,N′*-Tetramethyl-O-(1H-benzotriazol-1-yl)uronium hexafluorophosphate
HFIP	1,1,1,3,3,3-Hexafluoro-2-propanol
His/H	Histidine
HIV	Human immunodeficiency virus
HOBt	Hydroxybenzotriazole
HRMS	High resolution mass spectrometry
HSQC	Heteronuclear single quantum coherence
Ile/I	Isoleucine
ITC	Isothermal titration calorimetry
Kd	Equilibrium dissociation constant
KDE	Kernel density estimation
LCMS	Liquid chromatography mass spectrometry
Leu/L	Leucine
Lys/K	Lysine
m/z	Mass to charge ratio
MCC	Matthews correlation coefficient
MD	Molecular dynamics
Me	Methyl
Met/M	Methionine
MRE	Mean residue ellipticity
NBD	Nitrobenzoxadiazole
NCL	Native chemical ligation
NIP	Normalised integrated product
NMR	Nuclear magenetic resonance
NOE	Nuclear Overhauser effect
NOESY	Nuclear Overhauser effect spectroscopy
OPA	*Ortho*-phthalaldehyde
PBS	Phosphate buffered saline
PCA	Principal component analysis
PDB	Protein data bank
PEG	Polyethylene glycol
Ph	Phenyl
Phe/F	Phenylalanine

PPI	Protein-protein interaction
Pro/P	Proline
PyBOP	(Benzotriazol-1-yloxy)tripyrrolidinophosphonium hexafluorophosphate
REMD	Replica-exchange molecular dynamics
RF	Random forest
RMSD	Root-mean-square deviation
RNA	Ribonucleic acid
ROE	Rotating frame Overhauser effect
ROESY	Rotating frame Overhauser effect spectroscopy
RP-HPLC	Reverse phase high pressure liquid chromatography
rt	Room temperature
Ser/S	Serine
SPPS	Solid phase peptide synthesis
SVM	Support vector machine
TBAF	Tetrabutylammonium fluoride
*t*Bu	*tert*-Butyl
TES	triethylsilane
Tf	Trifluoromethanesulfonate
TFA	Trifluoroacetic acid
THF	Tetrahydrofuran
Thr/T	Threonine
TIPS	Triisopropylsilane
Tm	Melting temperature
TOCSY	Total correlation spectroscopy
Trp/W	Tryptophan
TrpZip	Tryptophan Zipper
Trt	Trityl
Tyr/Y	Tyrosine
UV	Ultraviolet
Val/V	Valine
WT	Wild type

1 Introduction

1.1 Cyclic Peptides

Cyclic peptides have found interest after being isolated from numerous animals and plants and being shown to have biological properties.[1, 2] Cyclic peptide hormones, including somatostatin, vasopressin and oxytocin, have been shown to be important for signalling pathways.[3-5] Several cyclic peptides are approved for use as drug molecules including the naturally occurring cyclic peptides gramicidin S, tyrocidine and vancomycin which are antibiotics and cyclosporin A which is an immunosuppressant.[3] Designed cyclic peptides have found many potential applications including in nanotechnology and the inhibition of protein-protein interactions (PPI). [3, 6-12]

Cyclic peptides possess several properties that make them suitable for the design of potential drug molecules. A less fixed structure compared to many available small molecule drugs allows for flexibility of the ligand to change shape and geometry in order to better bind a host meaning an increased specificity in binding can often be achieved by peptides compared to small molecules.[13-15] The body has mechanisms in place to degrade peptides, so bioavailability can be a problem. However studies have shown cyclic peptides often have much better bioavailability than the linear equivalent, being more resistant to hydrolysis and degradation by proteases.[16-18] The cyclisation of a peptide introduces conformational restraints resulting in stabilisation of specific structures. By cyclising the peptide, the number of conformations is reduced which can allow for the preorganisation of the peptide to the conformation that is necessary for binding to a target. This means there is reduced loss in entropy upon binding to a target thereby achieving increased potency.[19-21].

The ability to cover a larger area and have high binding affinity to relatively flat protein surfaces offers the potential for targeting protein-protein interactions (PPIs) which are difficult to bind using traditional small molecule drugs.[22, 23] This has allowed cyclic peptides to emerge as an area of interest in drug discovery where it is hoped peptides can bridge the gap between protein drugs and small molecule drugs.[3, 6-11, 24]

A problem remains in designing cyclic peptides, however. Cyclic peptides often adopt many conformations in solution, and for small cyclic peptides such as cyclic hexapeptides changing one amino acid in the sequence can greatly alter the conformation the cyclic peptide adopts. [14, 15, 22] This can make predicting the structure of cyclic peptides difficult. Computational methods currently provide the most efficient way of predicting the in-solution structure of cyclic peptides. [25]

1.2 Integrin-binding Cyclic Peptides

Cyclic peptide derivatives have been designed to bind amino acids and other biologically relevant molecules.[26, 27] Integrins are transmembrane receptors used by cells to interact with the extracellular matrix.[28, 29] They have been shown to participate in signal transduction upon binding, so they have been implicated in many diseases including cancer and autoimmune diseases.[30-33] Ligands which bind to integrins commonly contain the RGD tripeptide sequence which has led to the development of a number of RGD containing drugs candidates,[34-36] the most commonly known being the cyclic pentapeptide cilengitide which progressed to phase III of clinical trials as an anti-cancer agent.[37-39]

There are several different kinds of integrin. It has been shown that different conformations of the RGD motif bind to different integrins.[40] Spatial screening was carried out on a series of cyclic penta- and hexapeptides containing the RGD motif where a D-amino acid was incorporated into different positions.[41-44] Spatial screening is a technique where each amino acid in turn is swapped for a D-amino acid (or other modification).[41, 45] Inclusion of a D-amino acid within a cyclic penta- or hexapeptide is a method of reducing the conformations the peptide adopts as it often results in a well-structured conformation.[46] The substitution of an amino acid in a cyclic penta- or hexapeptide with the corresponding D-amino acid can therefore allow for some control of the conformation of the peptide. This means spatial screening can allow for the determination of the bioactive conformation of a peptide as different backbone conformations are screened. The structure of the peptides was determined by NMR and screening of the binding affinity of the peptides with known structures determined that the RGD motif in an elongated conformation binds αIIbβ3 integrins whereas a more bent motif binds to α5β1 and αvβ3 integrins (Figure 1). This was later confirmed with X-ray crystal structures of bound ligands.[47-49]

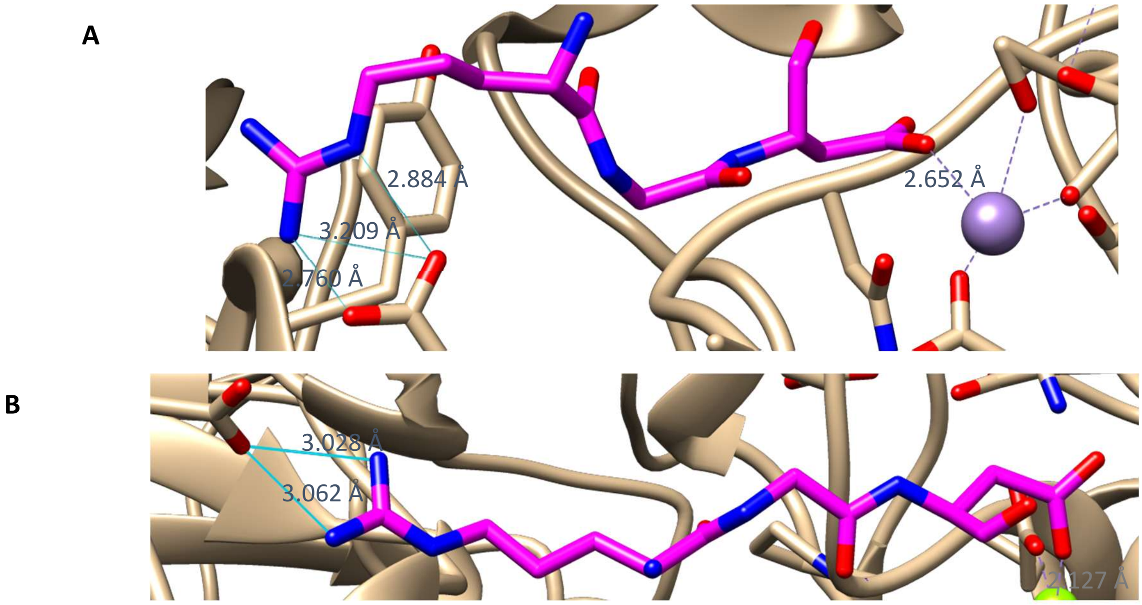


Figure 1: Bent RGD conformation bound to αvβ3 integrin (A, PDB:1L5G) and a linear RGD conformation bound to αIIβ3 integrin (B, PDB: 2VDR). The RGD peptide is shown in magenta and the immediate surrounding part of the integrin in light brown. Manganese ions found within the integrin proteins are shown in purple, calcium in green and magnesium in light green. Hydrogen bonding of the arginine sidechain with an aspartic acid sidechain of the integrin is shown in light blue.

1.2.1 Cilengitide

Due to the low molecular stability of linear peptides and to improve binding affinity, the RGD motif was included within cyclic peptide structures. As well as the RGD motif phenylalanine and valine were used to complete a cyclic pentapeptide as they are seen next to the RGD motif in vitronectin and fibrinogen, two proteins found in the extracellular matrix which bind to integrins.[29] Spatial screening using D-amino acids found the peptide *c*(RGDfV) to be the strongest antagonist of αvβ3 integrins, binding selectively over αIIbβ3 integrins.[50, 51] Further screening found *N*-methylation of the valine improved affinity and selectivity.[52] *N*-methylation has been shown to be able to reduce the conformational flexibility of peptides, and in some cases has been shown to improve bioavailability and stability of peptides.[53-55] The *N*-methylated peptide was named cilengitide (Figure 2).

Kessler *et al*. determined the NMR structure of cilengitide in water.[48] They found that the solution structure of cilengitide was similar to the bound conformation seen in the X-ray crystal structure of cilengitide bound to the extracellular segment of the integrin αvβ3, with a backbone atom RMSD of 0.407 Å.

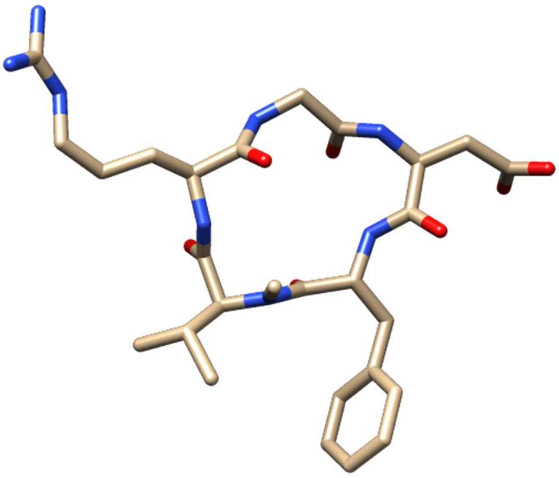


Figure 2: X-ray crystal structure of cilengitide. PDB: 1L5G

The incorporation of the RGD motif into cyclic peptides has been much studied with an emphasis on controlling the peptide conformation to bind specific integrins with high efficiency and specificity.[40] A large amount of information is therefore available about the structures of RGD containing cyclic peptides. They have therefore become benchmark structures when testing new computational methods of modelling cyclic peptides.[56]

1.3 Synthesis of Cyclic Peptides

There are many possibilities for the design of cyclic peptides including many combinations of both natural amino acids, unnatural amino acids and modifications. Furthermore, there are different types of cyclisation including head-to-tail, sidechain-to backbone and sidechain-to-sidechain (Figure 3). Therefore there are many different strategies for cyclisation that are compatible with solid phase peptide synthesis (SPPS).[57, 58]

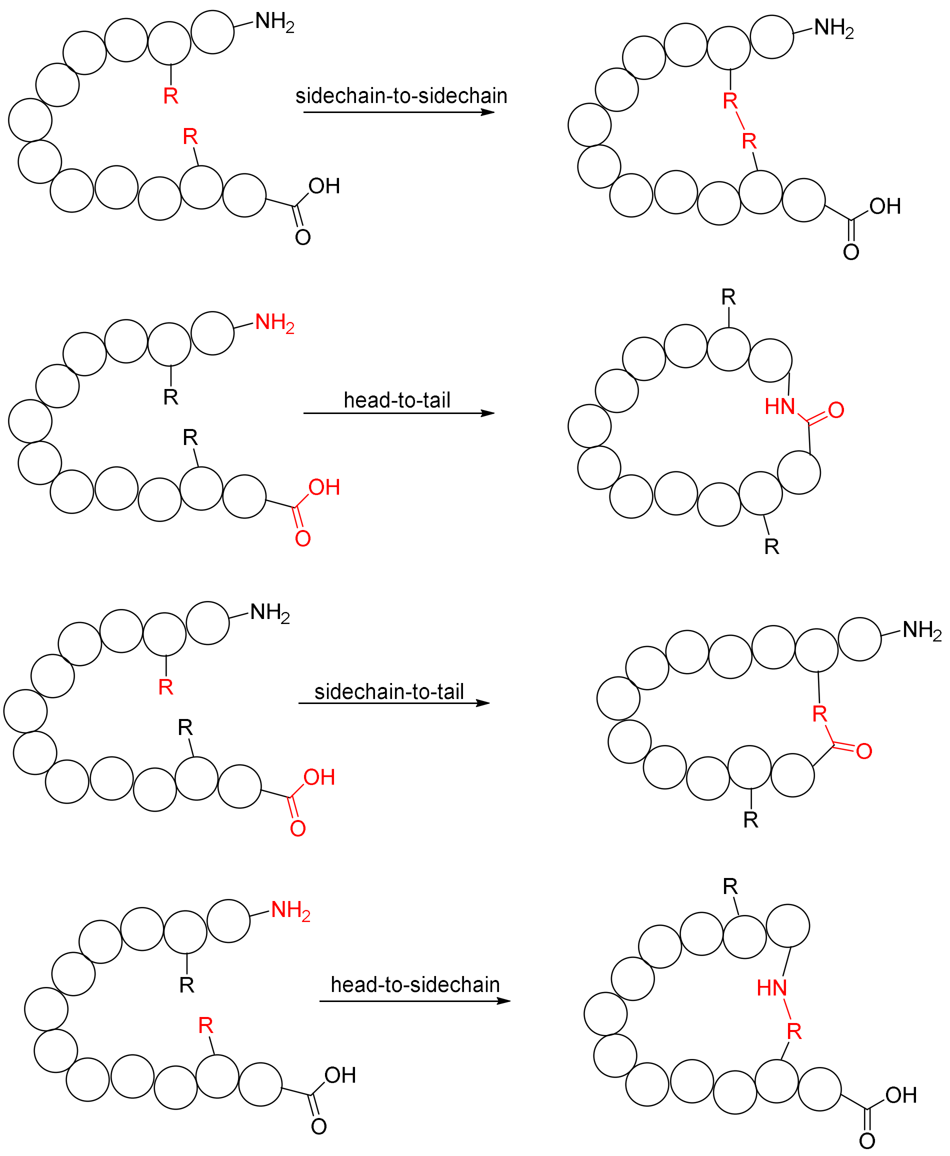
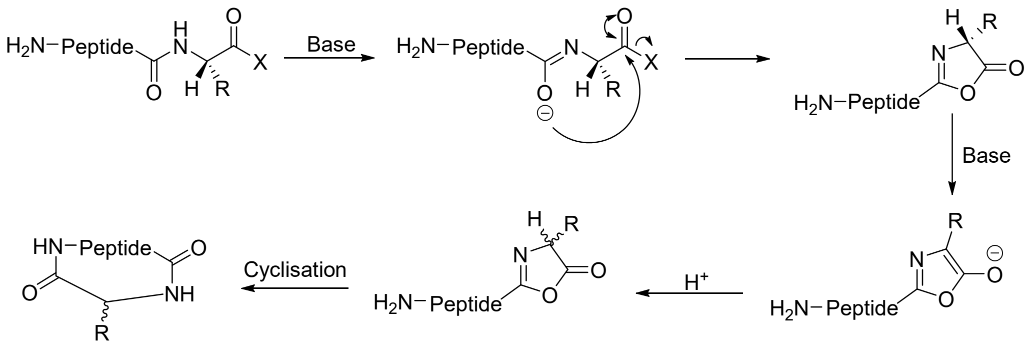


Figure 3: Different methods to cyclise cyclic peptides. Figure based on [59].

Peptide cyclisation can be performed in solution or on-resin. In solution phase cyclisation relatively large volumes of solvent are required to dilute the peptide to encourage intramolecular cyclisation rather than coupling with other peptides whereas on-resin cyclisation has the advantage of a pseudo-dilution effect. The pseudo-dilution effect can be particularly advantageous if the peptide to cyclise has a relatively high entropic barrier which must be overcome for the alignment of the N- and C-termini required for cyclisation. Two common methods for the synthesis of head-to-tail cyclic peptides are native chemical ligation (NCL) and the use of sidechain anchoring to resin to allow for on-resin cyclisation.

Common problems often occur during the cyclisation step during the synthesis of cyclic peptides. Oligomerisation is often seen and if amino acids besides glycine or proline are included at the C-terminus during head-to-tail cyclisation, epimerisation of the C-terminal amino acid is common. The most likely mechanism for this epimerisation is through the formation of an oxazolone from the activated C-terminal amino acid, leading to the loss of the α-carbons chirality (Scheme 1).[60] Various methods can be used to cyclise peptides that address these problems to different degrees and each has its own advantages and disadvantages such as restrictions on peptide sequence.



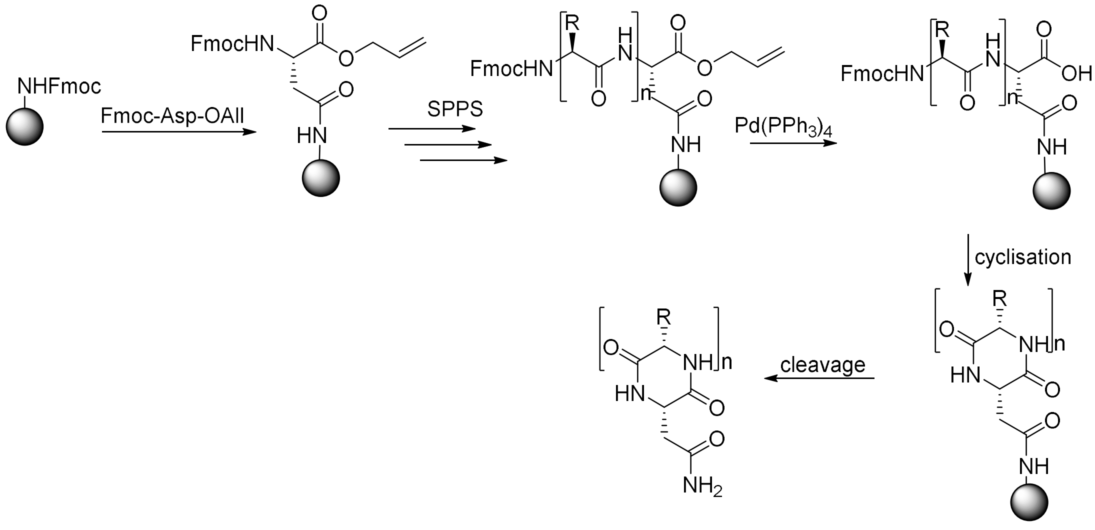
Scheme 1: Epimerisation of the C-terminal amino acid during cyclisation.

Many naturally occurring cyclic peptides are head-to-tail cyclic peptides.[3] Enhanced metabolic stability is often seen in these peptides as, due to the absence of N- and C-termini, the peptides are resistant to exopeptidases. Especially for head-to tail cyclisation, cyclisation is the most problematic step in the synthesis, particularly for cyclic peptides shorter than 7 residues in length as it requires dihedral angles not commonly accessed by most amino acids.[61-63] The sequence and point of cyclisation can affect the yield with sterically hindered amino acids at the point of cyclisation prone to decrease the yield,[64] and turn inducing elements such as proline able to increase yields.[57, 65] Glycine is another residue often utilised due to its flexibility. Another commonly used tactic for cyclisation is the use of D-amino acids, which when included in the peptide sequence can often lead to increased yields.[57, 58, 66-69] Inclusion of *N*-methylated amino acids can also aid cyclisation. [57, 58, 66-70]

1.3.1 On-resin Synthesis Using an Allyl Group to Protect the C-terminus

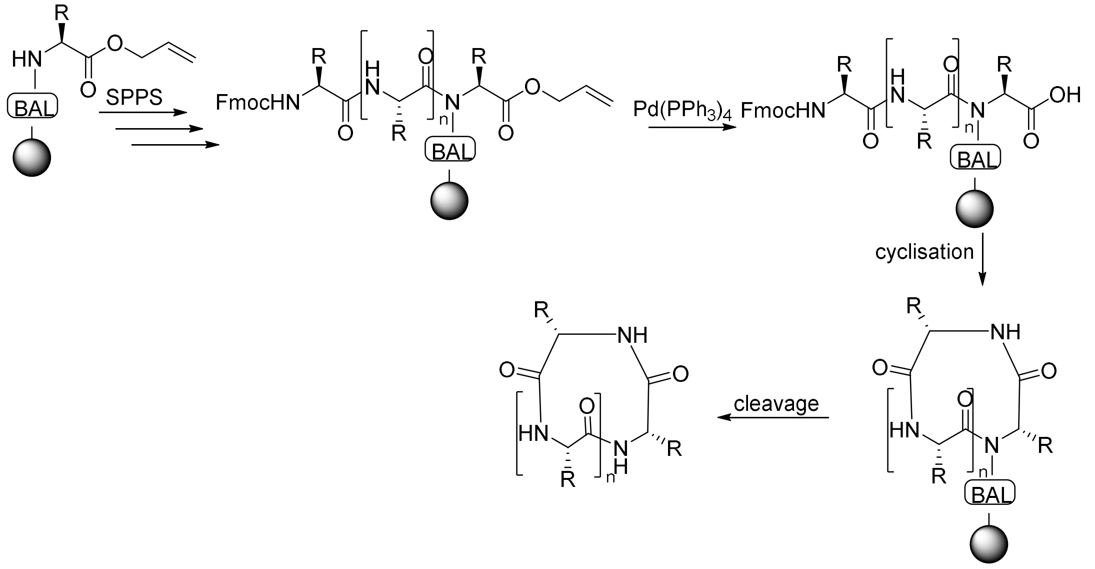
On-resin cyclisation for formation of head-to-tail cyclic peptides requires the sequence to contain an amino acid that can be attached to the resin via its sidechain leaving its C-terminus free for cyclisation. Three levels of chemical orthogonality are therefore required as the C-terminus must be protected by a protecting group that isn't affected by the synthesis procedure, which for Fmoc/*t*Bu SPPS contains protecting groups removed either by basic or acidic conditions.

One method used for on-resin cyclisation compatible with Fmoc/*t*Bu SPPS is carried out by using Fmoc amino acids with their C-terminus protected by an allyl group. Amino acids such as Fmoc-Asp-OAll and Fmoc-Glu-OAll can be coupled to a resin via the sidechain through formation of an amide bond. The allyl group on the C-terminus is stable in both basic and acidic conditions but can be removed using a palladium(0) catalyst, generally $Pd(PPh_3)_4$ with a phenylsilane scavenger (Scheme 2). The first example of using an allyl group as a third orthogonal protecting group for peptide cyclisation was reported in 1993 by Albericio *et al.*[71] Decapeptide c(AAA-*D*-F-PEDNYE) was synthesised in 71 % yield. The Asn residue was used to attach the peptide to the resin via a PAL linker and, the linear peptide was then synthesised by SPPS. The allyl group was removed using $Pd(PPh_3)_4$ in DMSO:THF:0.5 N aqueous HCl:morpholine (2:2:1 :0.1), for 2 h at 25 °C. The *N*-terminal Fmoc was removed and cyclisation carried out using BOP/HOBt/DIPEA followed by cleavage of the peptide from the resin. This method has since been adapted and used in the synthesis of many cyclic peptides including the total-synthesis of many naturally occurring cyclic peptides.[1, 59, 72] Other amino acids such as lysine and tyrosine have also been used for sidechain anchoring.[73-75]



Scheme 2: Asparagine sidechain anchoring for on-resin cyclisation.

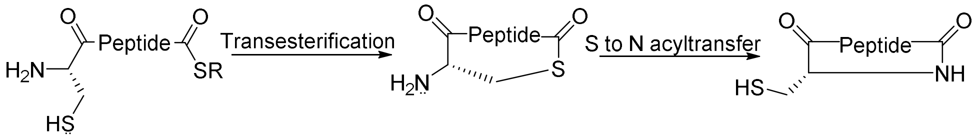
Rather than anchoring the peptide chain through an amino acid sidechain, the peptide can also be anchored to a resin through the backbone using a backbone amide linker (BAL). The first residue is attached via the amide group to the BAL which is anchored to the resin.[76] This leaves the C-terminus free for on-resin cyclisation (Scheme 3). As a secondary amine is produced when the first amino acid is attached to the BAL, coupling of an amino acid to this residue can often be difficult due to steric hinderance.



Scheme 3: Backbone amide linker method to synthesis cyclic peptides.

1.3.2 Native Chemical Ligation

A commonly used method for in-solution head-to-tail cyclisation is native chemical ligation (NCL).[77] NCL requires the peptide to contain a cysteine at the *N*-terminus and a thioester at the C-terminus. The sulfur of the cysteine attacks the carbonyl of the thioester forming a thiolactone intermediate. This intermediate then undergoes a S- to N-acyl shift, an irreversible step resulting in a normal peptide bond (Scheme 4). The sidechain of the C-terminal residue must be considered carefully as if it is a large or bulky group it will sterically hinder the reaction leading to slow reaction rates. There is a risk of epimerisation at the C-terminus when it is converted to a thioester. Providing there are no other residues in the peptide that are affected by the reaction, an alanine can be used instead as the cysteine can be reduced to alanine after the NCL reaction.[78]



Scheme 4: Native chemical ligation.

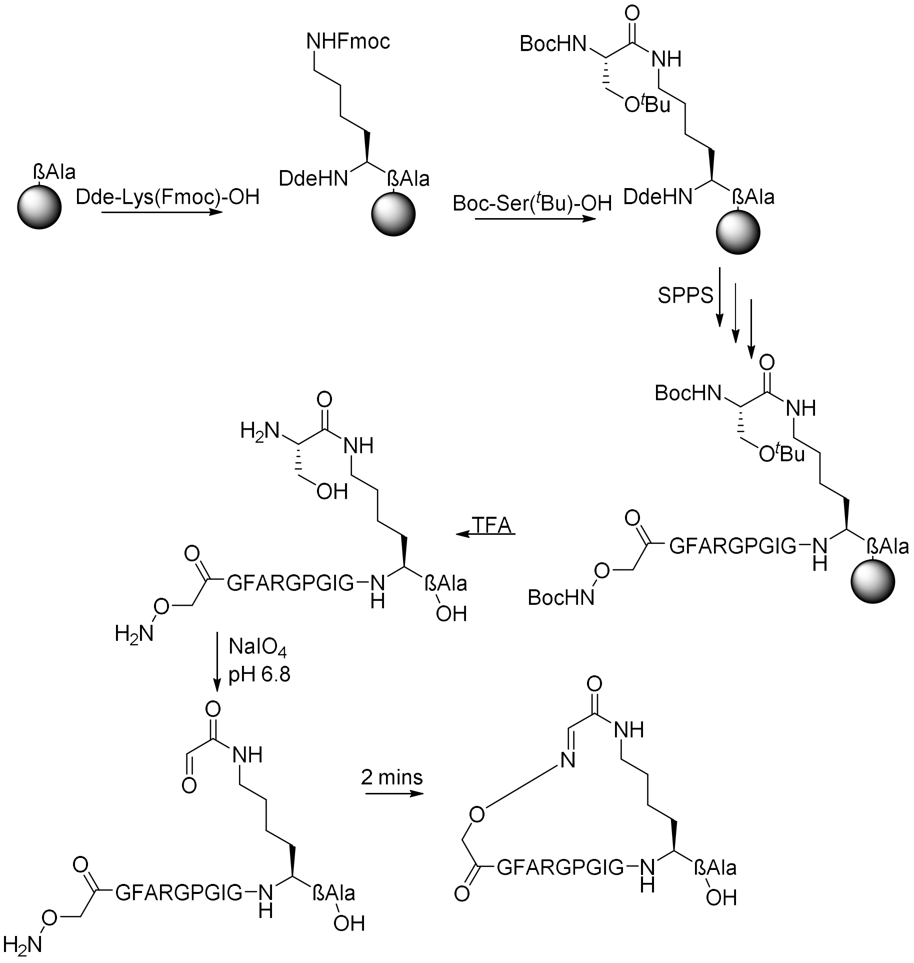
1.3.3 Peptidomimetics

Sidechain anchoring and NCL methods place restrictions on the sequence that can be used. Additionally for in-solution cyclisation often large volumes of solvent are required and often give low yields with many side-reactions possible. Therefore alternative methods which can be used to synthesis cyclic peptides are desirable. To address some of the inefficiencies of cyclisation through formation of a peptide bond the use of peptidomimetics offers an alternative. Besides NCL, ligation reactions developed for joining of peptide or protein fragments have often been adapted for the synthesis of cyclic peptides.[79] Such ligation reactions can be highly chemoselective and do not have the problem of epimerisation of the C-terminus. Many methods have been used to cyclise cyclic peptides including formation of disulfide bridges,[80-82] triazole formation[83-86] and cross-metathesis.[87-89]

1.3.3.1 Oxime/Hydrazone Cyclisation

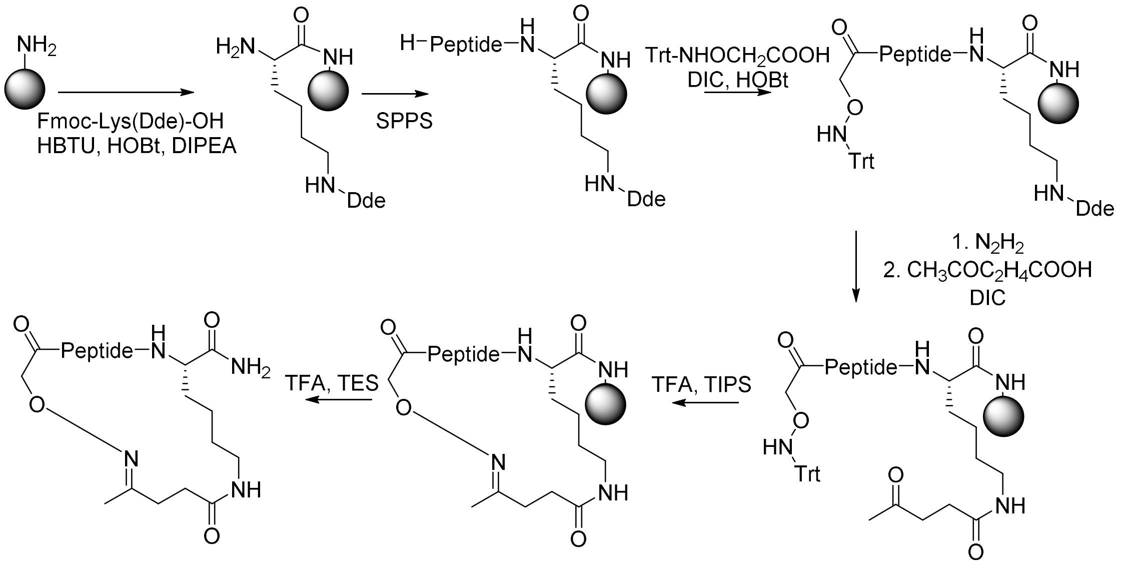
Many cyclisation strategies have been developed that involve the ligation reaction between a nucleophilic functional group and a C-terminal aldehyde.[59, 72] Oxime or hydrazone formation is one such ligation reaction relying on a C-terminal aldehyde that has been utilised to synthesise cyclic peptides. Oxime or hydrazone formation can occur through the reaction of an alkoxylamine or hydrazine with an aldehyde or ketone under relatively mild conditions.[90] Although more stable than imines due to the α-effect from the additional heteroatoms, oximes and hydrazones can be hydrolysed in aqueous conditions so reduction of the C=N double bond can be carried out. Oximes are generally more stable than hydrazones due to the greater electronegativity of oxygen.

Pallin *et al*. used oxime formation to synthesise a head-to-tail cyclic peptide (Scheme 5).[91] Dde-Lys(Fmoc)-OH was coupled to Fmoc-β-Ala-Wang resin. Following removal of the Fmoc group Boc-Ser(*t*Bu)-OH was coupled to the lysine sidechain. The Dde protecting group was removed using 1% hydrazine and the peptide synthesised by Fmoc/*t*Bu SPPS. The peptide was then cleaved from the resin using TFA. The 1,2-aminoalcohol of the serine residue was then oxidised using sodium periodate and the peptide was cyclised through the formation of an oxime. No side reactions were seen due to the unprotected sidechains during the cyclisation conditions.



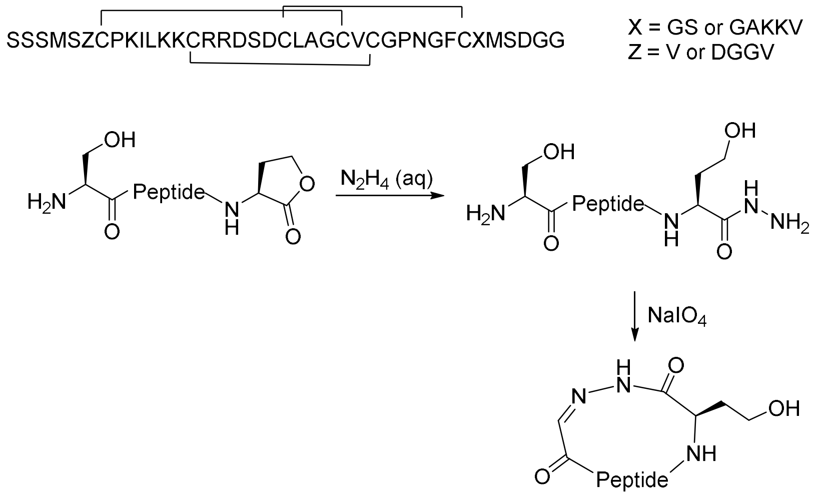
Scheme 5: Pallin et al. cyclisation of a cyclic peptide through formation of an oxime.[91]

Roberts *et al*. used oxime formation to create a library of head-to-sidechain cyclic peptides based on the DD/EXF motif found in the active site of type II restriction endonuclease *Eco*RV (Scheme 6).[92] A C-terminal lysine with a Dde protecting group on the sidechain was initially coupled to the resin. Fmoc/*t*Bu SPPS was used to couple the amino acids. (Trt-aminooxy)acetic acid was coupled to the *N*-terminus. Following removal of the Dde group the lysine sidechain reacted with levulinic anhydride. Mild acid removes the trityl group from the *N*-terminus allowing formation of an oxime on resin. The oxime was seen to be partially reduced under the cleavage conditions of TFA and triethylsilane (TES). TES in acidic conditions has previously been used for the reduction of non-peptide oximes.[93, 94] Less reduction was seen with less than 1% TES in the cleavage mixture or by using the less reactive TIPS as an alternative scavenger.



Scheme 6: Synthesis of a peptide library through hydrazone formation.[92]

The Kolmar group used hydrazone formation for the cyclisation of a 34-residue peptide which was found to inhibit β-II tryptase activity (Scheme 7).[95] The peptide was based on the cyclotide family of cyclic peptides. This family of peptides isolated from plants, are head-to-tail cyclised in addition to having a cysteine-knot motif formed through disulfide bonds of six conserved cysteine residues.[96] The miniprotein was synthesised using recombinant peptide synthesis then cleaved at the methionine sites using cyanogen bromide. The precursor contains a C-terminal γ-lactone which allowed for incorporation of the hydrazine group and the *N*-terminal residue serine was oxidised to generate a ketoaldehyde. Cyclisation occurred spontaneously in aqueous solution in minutes following oxidation of the serine. The peptide, prior to hydrazone formation, was likely already folded in an orientation with the N- and C-termini in close proximity to allow for the rapid cyclisation. The hydrazone remained stable at physiological pH. The hydrazone was found to be a good mimic for a peptide bond and tolerated in the biological system with increased inhibition of the tryptase seen over the peptide cyclised through the amide bond.

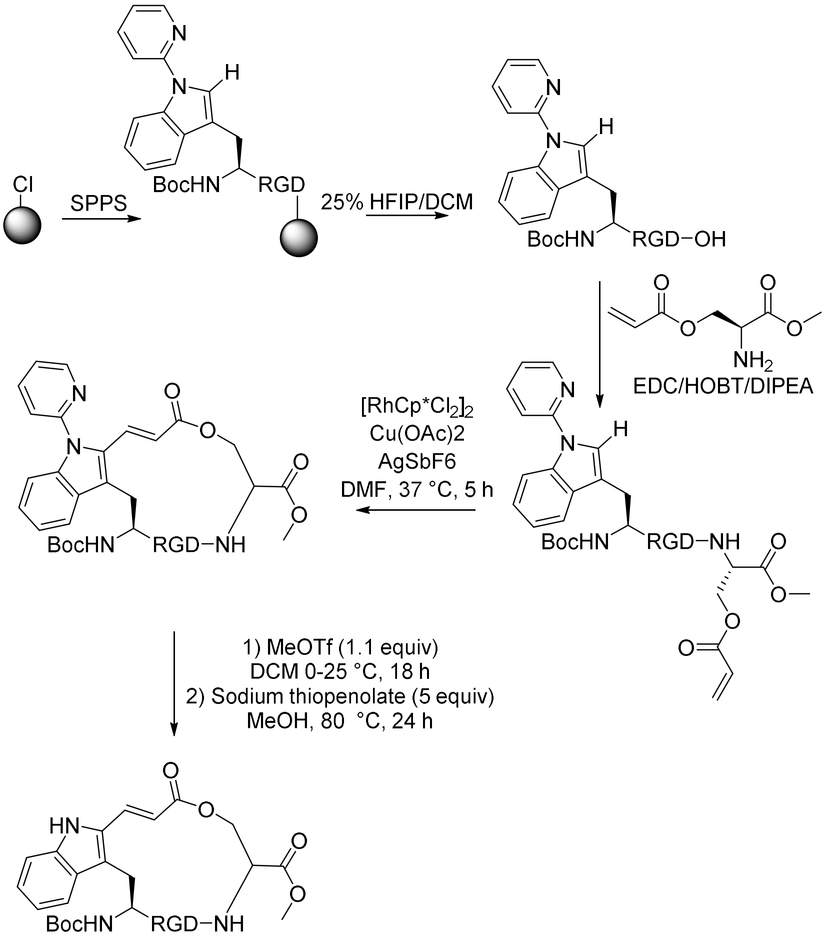


Scheme 7: Head-to-tail cyclisation through hydrazone formation of a cyclic peptide to inhibit β-*II tryptase.*[95]

1.3.3.2 Introducing Additional Functionality

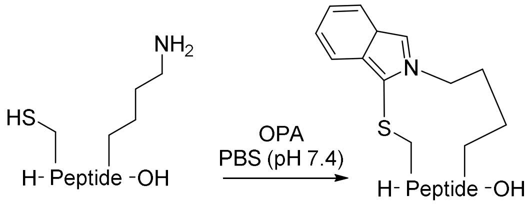
The incorporation of additional units for cyclisation allows for the potential to introduce additional functionality such as photoswitchable groups[97, 98] or fluorescence into the system.[99, 100] Fluorescent labelled proteins have become a powerful tool for studying the biological function of proteins. Various methods exist to introduce site-specific fluorescent labelling into peptides including installing N/C-terminal fluorescent dyes or unnatural amino acids with bio-orthogonal functional groups which can be linked to fluorophores.[101] These methods generally rely on large fluorophores which can alter the properties of the associated peptide. Therefore smaller fluorescent systems which can be incorporated directly into peptides are advantageous. A few methods which can be used for peptide cyclisation also directly incorporate fluorescence into the peptide during cyclisation without the need for ligation of a large fluorophore.

Tryptophan is a naturally occurring amino acid with intrinsic fluorescence. Its use in biological applications remains limited due to its suboptimal fluorescent properties. Liu *et al*. used Rh-catalysed C-H olefination in tryptophan residues for the cyclisation of cyclic peptides (Scheme 8).[102] A pyridyl group was introduced into the tryptophan to act as a directing group and increase reactivity and chemoselectivity. The E-alkene was produced. The pyridyl group could be removed following cyclisation. A red shift in the fluorescence of the tryptophan from 340 to 460 nm was seen upon cyclisation. A series of cyclic peptides were synthesised indicating the wide substrate scope of the reaction including the synthesis of a RGD containing peptide. The RGD cyclic peptide demonstrated strong binding and fluorescent labelling of αvβ3 integrins in cells.



Scheme 8: Synthesis of a RGD-containing cyclic peptide by Liu et al.[102]

Zhang *et al*. and Todorovic *et al*. simultaneously reported the use of *ortho*-phthalaldehyde(OPA) for cyclisation via reaction with the *N*-terminus or a lysine sidechain and a cysteine residue to form an isoindole (Scheme 9).[103, 104] The reaction proceeded in aqueous buffer on the deprotected peptide and had a broad sequence tolerance. However if the peptide contained both a free *N*-terminus and a sidechain amino group the reaction was not selective resulting in a mix of head-to-sidechain and sidechain-to-sidechain cyclic peptides so orthogonal protecting groups are required in such cases. The resulting isoindole moiety provides a scaffold for further modification of the peptide and provides intrinsic fluorescence. Although the isoindole resulting from the reaction of OPA gives very similar fluorescent properties to tryptophan, Todorovic *et al*. tested a series of substituted OPAs and the fluorescent properties of the resulting isoindoles.[103] The nitro-isoindoles were observed to have red-shifted maxima. However the regioselectivity of the isoindole formation from 3- and 4-substituted OPAs was limited and the regioisomers were difficult to separate.



Scheme 9: Ortho-phthaldehyde cyclisation of a peptide through a reaction with a cysteine and lysine sidechain.

1.4 β-turns

The secondary structure of a protein describes the local conformation of the chain which builds up to form the overall folded protein structure (tertiary structure). The ϕ, ψ and ω dihedral angles for each residue within a peptide chain can be used to define the conformation (Figure 4). Due to the delocalisation of the lone pair of electrons on the nitrogen atom the peptide bond has partial double bond character. This restricts the rotation around the amide bond putting limitations on the value of the ω dihedral angle, with most peptide bonds being in the trans conformation with an ω dihedral angle of 180°. The ϕ and ψ dihedral angles are therefore most commonly used in describing peptide conformation. The Ramachandran plot is used to visualise the ϕ and ψ dihedral angles with distinct secondary structures having unique allowed values. Of the 20 naturally occurring amino acids, with the exception of glycine and proline, most have a similar Ramachandran distribution in the absence of structural constraints.[105]

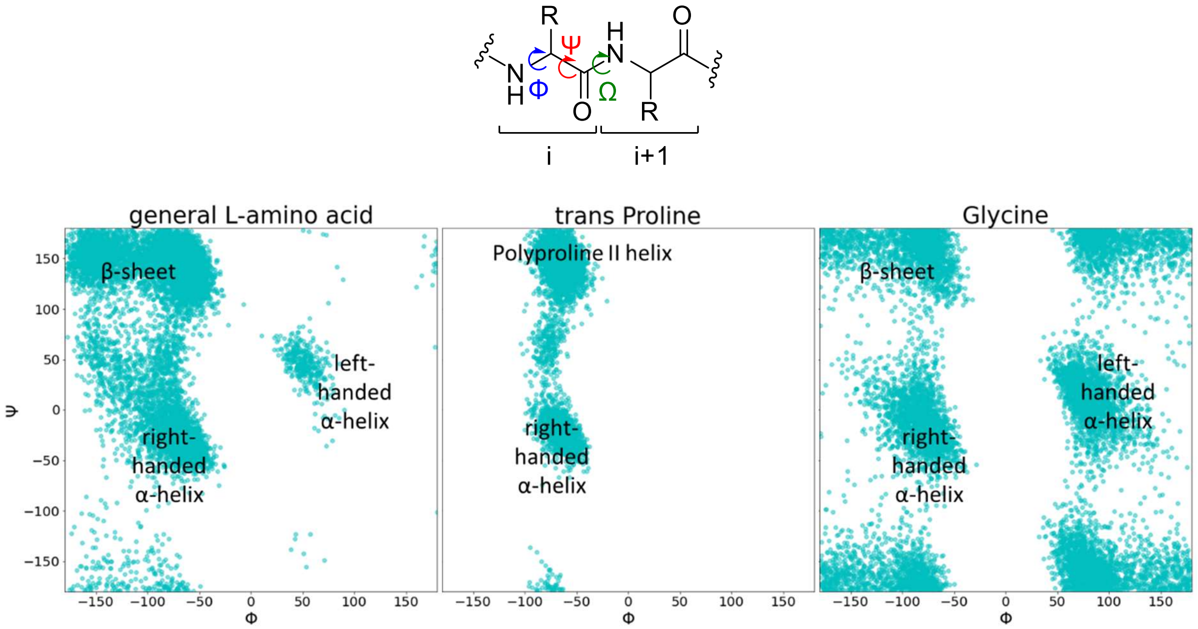


Figure 4: Peptide dihedral angles and the Ramachandran plots for a general L-amino acid, trans proline and glycine. Ramachandran plots obtained from the Loop Database (see section 10.1).

Reverse turns are an important part of protein secondary structure allowing the protein chain to change direction, as such they have been shown to be important for protein folding.[106-109] There are different types of turns based on the number of residues they contain. The smallest are δ-turns made up of two residues with a hydrogen bond between the NH(i) and CO(i+1). Slightly larger are γ-turns which contain three residues with the hydrogen bond between the CO(i) and NH(i+2). β-turns are composed of four residues - i to i+3 (Figure 5). They are defined by the presence of a hydrogen-bond between the CO(i) and NH(i+3), or an alternative definition uses a distance of 7 Å or less between the α-carbons of the i and i+3 residues so long as the i+1 and i+2 residues are not part of a helix or other form of secondary structure. α-turns contain 5 residues within the turn structure and, similar to β-turns, either the presence of a hydrogen-bond or less than 7 Å distance between the i and i+4 positions can be used to define them. Finally π-turns contain six residues and a hydrogen bond between residues i and i+5.[110]

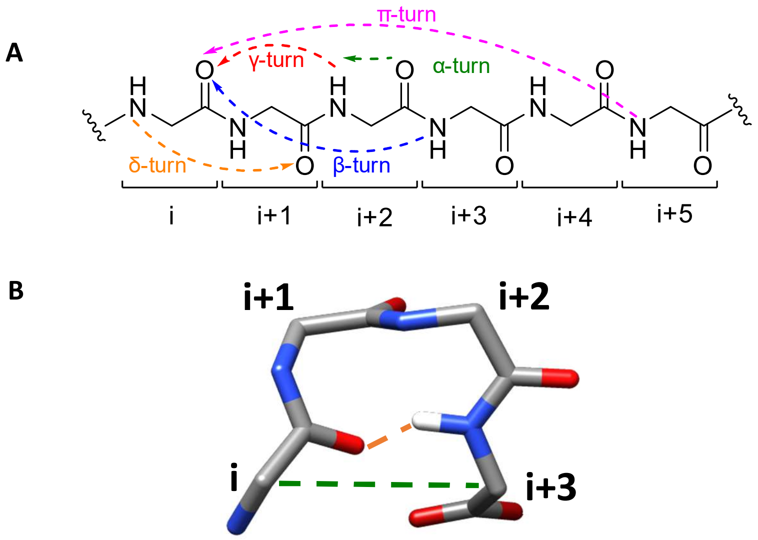


Figure 5: The different types of reverse turn (A) and the β-turn structure (B) which can have a hydrogen bond between the i and i+3 residues (orange dashed line), or a distance of 7 Å between the α-carbons of the i and i+3 residues (green dashed line).

β-turns are the most common reverse turn type.[111] They are often found on the surface of proteins where they have been implicated in protein-protein interactions (PPIs).[112, 113] Approximately 25% of β-turns are found to not contain the hydrogen bond between the i and i+3 positions,[114, 115] with type VIII turns in particular lacking the hydrogen bond.[116, 117]

1.4.1 Classification of β-turns

There are multiple β-turn types, defined by the dihedral angles of the i+1 and i+2 residues (Table 1).[118, 119] The commonly used turn type classifications were established in the 1960s and 1970s based on datasets of the β-turns observed in available protein structures.[118, 120, 121] Currently the β-turn types established by the Thornton group are used as standard.[114, 122] Eight different turn types were identified, with the additional type IV category for turns that do not fit any of the remaining categories.[114] Analysis of the datasets show type I turns are the most common followed by type II.

Turn type	ϕ_{i+1}	ψ_{i+1}	ϕ_{i+2}	ψ_{i+2}
I	-60	-30	-90	0
II	-60	120	80	0
I’	60	30	90	0
II’	60	-120	-80	0
VIII	-60	-30	-120	120
VI_{a1}	-60	120	-80	0
VI_{a2}	-120	120	-60	0
VI_{b}	-135	135	-75	160

Table 1: Ideal i+1 and i+2 dihedral angles for different β-turn types. Type VI turns require a proline to be at the i+2 position.

A more recent attempt to classify β-turn types was carried out by de Brevern in 2016.[123] A clustering approach was used to try and address the fact that, based on the dataset used, approximately 1/3 of β-turns fit into the miscellaneous type IV category. The type IV category was subdivided into 4 categories based on the clustering algorithm used (Table 2). However what was assigned as a type IV_1 turn resembles a type II turn, and the remaining three assigned type IV turn categories resemble a type I turn, so a different clustering algorithm may not always separate them as there is overlap in the Ramachandran plots of the turn types.

Type IV β-turn	ϕ_{i+1}	Ψ_{i+1}	ϕ_{i+2}	Ψ_{i+2}	% of type IV β-turns
IV_1	-120	130	55	41	16.08
IV_2	-85	-15	-125	55	12.44
IV_3	-71	-30	-72	-47	11.15
IV_4	-97	-2	-117	-11	8.50
IV_{misc}	-	-	-	-	51.83

Table 2: The categories of type IV turns identified by de Brevern.[123]

Shapovalov *et al*. carried out a density-peak based cluster analysis on a set of 1,074 high-resolution protein structures (1.2 Å resolution or higher).[124] This dataset contained 13,030 β-turns, and was clustered (using DBSCAN[125]) based on the dihedrals of the i to i+3 residues rather than just the i+1 and i+2. Most clusters were then further divided into smaller clusters using the *k*-medoids clustering algorithm.[126] This double clustering method was used to prevent some of the existing β-turn type classifications being omitted due to overlap in the densities of some clusters. 18 turn-types were identified, and the amino acid composition of each turn-type explored. 6 of these 18 were new turn types, often created by splitting other turn types. Unsurprisingly proline and glycine play important roles in the sequence breakdown of each turn-type. These researchers also undertook an analysis of the frequency of β-turns in loops and found that 63% of residues in loops are β-turns, with many turns overlapping with other turns.

1.4.2 Predicting β-Turn Type

The primary sequence of a protein influences the secondary structures that the protein adopts and this in turn affects the tertiary structure (and sometimes quaternary structure). A lot of effort has been put into determining the final structure of the protein from the sequence of amino acids. This includes the prediction of the location and type of β turns.

Most β-turn prediction methods are designed to scan through a full protein sequence and determine the location and then type of β-turns as a method towards predicting the conformation of a protein. Information such as the abundance of hydrophilic residues within β turns, multiple alignments and secondary structure predictions are commonly used rather than just sequence to improve predictive ability. Rather than one prediction being made for the tetrapeptide sequence that makes up the β-turn, generally a prediction is made for each amino acid individually. Window sizes for the number of amino acids looked at at a time vary between around 4 and 10 amino acids.

A dataset composed of 426 non-homologous proteins with structures determined by X-ray crystallography all solved with a resolution < 2 Å, known as BT426, is often used to test the predictive ability of new methods to allow direct comparison with previous algorithms. The Matthews correlation coefficient (MCC) is a metric that can be used to assess the performance of classifiers such as those used to assign β-turn type. The MCC has a value between -1 and 1 where 1 is 100% accurate classification and -1 is where there is 100% misclassification.

1.4.2.1 Statistical Methods

Statistical methods to determine the location of β-turns within proteins began in the 1970s and were expanded in the late 1980s to also be able to predict the type of β-turn.[127-130] Initial methods looked through a protein sequence and determined whether a sequence within the protein was likely to be within a β-turn or not. Methods were then developed to also assign likely β-turn types to the sequences which were predicted to be found within β-turns. Initially prediction of β-turn types was limited to types I, II and a non-specific category for the remaining turn types which are less common (or less well defined in the case of type IV). As more high resolution X-ray crystal structures were determined more data became available to improve the prediction methods and allow the inclusion of other β-turn types.

Statistical methods work by giving position specific potentials for each amino acid in the sequence based on known 3D protein structures. The likelihood of each amino acid appearing at each position within each β-turn type is determined and these propensities are used to predict β-turn type.

The first predictions of β-turn type were carried out in 1988 by Wilmot and Thornton.[127] Statistical data was obtained from 59 proteins containing 735 β-turns. This was sufficient data to determine propensities to predict whether a sequence was a type I or II β-turn or a non-specific turn (all remaining turn types). It was predicted whether a tetrapeptide within a protein was within a β-turn or not, and whether the β-turn was type I, II or non-specified in one procedure. Using propensities of individual turn types was found to improve the accuracy of predicting whether a tetrapeptide was within a β-turn or not over the previously used statistical methods. Of the β-turns that were correctly predicted by the algorithm 71% were correctly assigned to the right type. Variation in predicting turn type correctly was seen within the accuracy of the three turn type categories: 78 % of type I were predicted correctly, 80% type II and 63% non-specific. The lower accuracy for the non-specific category reflects the variation within the category.

Chou and Blinn extended the propensities method to include first-order residue coupling effects to predict whether a tetrapeptide sequence was part of a β-turn (and what type of β turn) or not.[131] They tested the method by predicting the structure of rubredoxin and comparing this to the known structure. The number of correctly assigned tetrapeptides was 82.4 % which was an improvement over previous methods.

An advance in statistical methods to predict β-turn type called COUDES was published in 2005 by Fuchs and Alix.[132] Propensities and multiple alignments were used to determine if a tetrapeptide was likely to be a β-turn or not, and then propensities used to predict β-turn type. The optimal length of a protein chain to look at to predict β-turns was 12 residues indicating that β-turns are influenced by their local environment. The commonly used dataset BT426 was used to test the predictive ability of the propensities. A MCC of 0.42 was achieved for predicting whether a sequence was likely to be a β-turn or not. β-turn types I, II, VIII, I’, II’ and IV had MCCs of 0.31, 0.30, 0.07, 0.23, 0.11 and 0.11 respectively.

Although initially more effective, statistical methods are now less common and give less accurate predictions than machine-learning methods which have improved based on the algorithms used and the type and quantity of input data used to train them.

1.4.2.2 Machine-Learning Methods

Starting from the late 1980s machine learning methods started to be developed and implemented towards structure prediction. Data used to train a machine learning algorithm is known as the training set. During machine learning, parameters will be tuned to best fit the model to the data in the training set. After training the performance of the model is tested on a set of data with known outcomes (the test set). Machine learning algorithms that assign β-turn types are classifiers as they assign a class label (the predicted β-turn type) to the input data.

Initially neural networks were used, and later support vector machines (SVMs) have also become common to predict the location and type of β-turns within proteins. Methods that are able to predict β-turn type (rather than just whether or not a sequence within a protein is likely to be a β-turn or not) include MOLEBRNN[133], BETATURNS[134], DEBT[135] and NetTurnP.[136]

1.4.2.2.1 Neural Network-Base Methods

Neural network algorithms were developed to loosely mimic the neural networks in a human brain.[137] Nodes/artificial neurons mimic biological neurons which can signal other neurons through synapses. The artificial neurons can similarly send signals to other artificial neurons they are connected to. This allows artificial neural networks to "learn" as they cluster and classify data. The neurons receive a signal in the form of a number and perform a function on it and can then signal other neurons they are connected to. The signals out of the nodes have different weights which affects the magnitude of the signal. Neural networks often have multiple layers where each layer may perform different transformations on their inputs. The input determines the state of the first layer, the final layer gives the output, and any layers in the network between these two layers are known as hidden layers. A network with no hidden layers is known as a perceptron.

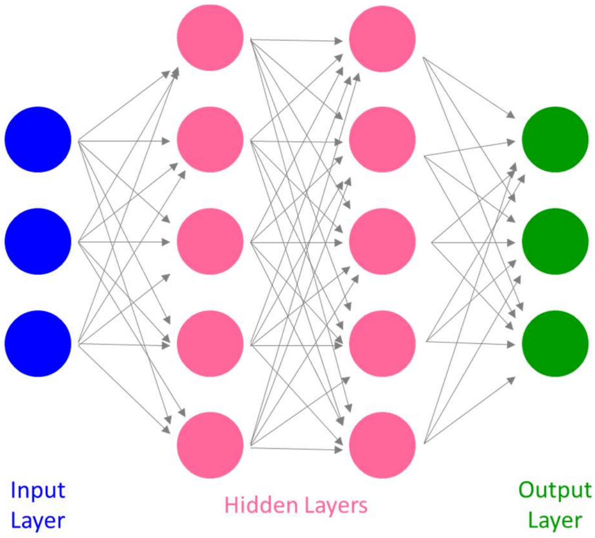


Figure 6: Neural network with two hidden layers.

The first machine learning algorithm applied specifically to the prediction of β-turns rather than more general prediction of secondary structure (α-helix, β-sheet or coil) was a neural network produced by McGregor *et al*. in 1989.[138] The BT426 dataset of proteins was used to test the algorithm. The amino acid at each position in a tetrapeptide was used as the input data for the neural network. The output is one of four categories: type I or II β turn, a non-specific β-turn (comprised of all the remaining β-turn categories) and non-β turn. The type I, II and non-specific β-turns were predicted with an accuracy of 69, 81 and 36 % respectively using a network with eight hidden units.

Kaur and Raghava used a neural network to predict the position and type of type I, II, VIII and IV β-turns.[134] They used the BT426 dataset to assess the predictive ability of the model.[111] The method called Betaturns uses data from multiple sequence alignments and secondary structure prediction for the neural network rather than directly using the protein sequence. The MCCs of the β-turn types I, II, IV and VIII are 0.25, 0.26, 0.18 and 0.13.

The use of multiple alignments and secondary structure prediction of the surrounding structure in Betaturns shows how additional data besides just the turn sequence can help improve prediction. NetTurnP is another artificial network method to predict β-turn types developed by Petersen *et al*.[136] Multiple alignments, predicted secondary structure and surface accessibility were used as inputs. Three types of artificial networks were trained based on this method – β-turn-G which predicts whether an amino acid is in a β-turn or not, β-turn-S which predicts what type of β-turn an amino acid is likely to be in and β-turn-P which predicts which position (i/i+1/i+2/i+3) within a β-turn an amino acid is in. MCCs for turns types I, I', II, II', IV, VIII, VI_{a1}, VI_{a2}, VI_{b} are 0.36, 0.23, 0.31, 0.16, 0.27, 0.16, 0.07, 0.03 and 0.11 respectively.

The neural network architecture and the number of hidden layers can also impact the predictive ability of a model. DeepDIN (deep dense inception network) is a deep neural network implemented in MUFold-BetaTurn, a method for predicting β-turn types.[139] It has many more layers than the shallow neural networks previously made for β-turn prediction. Data added to the neural network for prediction included physicochemical properties of the amino acids and secondary structure prediction. DeepDIN is designed to look at the whole protein at once and is able to read up to 700 residues at a time. As a deep neural network it can be used to explore features at multiple scales together, extracting long-range residue interactions, and discover high-level features including local and nonlocal interactions of residues. A DeepDIN network was trained for each of the nine different types of β turn. The method was not tested for β-turn type prediction on the BT426 dataset but the MCCs on a different dataset of 6376 protein chains were 0.30, 0.45, 0.35, 0.33, 0.20, 0.33, 0.25, 0.35 and 0.17 for turn types I, I', II, II', IV, VI_{a1}, VI_{a2}, VI_{b} and VIII respectively.

1.4.2.2.2 Support Vector Machine-Based Methods

Support vector machines (SVMs) are another machine learning method that have been used for the prediction of β turns, although most methods only predict whether a β-turn is likely to occur in a peptide or not rather than the type of β-turn a sequence is likely to form.[140-142] Figure 7 shows two clusters that can be separated by a straight line. The three dotted lines represent possible decision boundaries that could be assigned by a classifier. Although they work for the data shown, the orange and green lines may not work well for new data that is close to the boundary. To avoid misassignment it is beneficial to have a large gap between the two clusters. Support vector machines (SVMs) use a kernel function to map the input data into a new space where the greatest linear boundaries between the different classes can be found.[143]

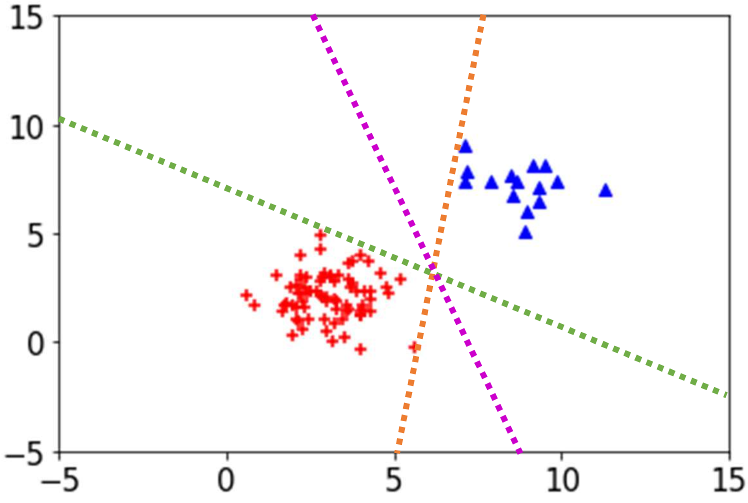


Figure 7: Potential decision boundaries to separate two clusters.

			MCC Values for BT426 Dataset where available								
	Method	Turn Types Predicted	I	II	I’	II’	VIII	VI$_{a1}$	VI$_{a2}$	VI$_{b}$	IV
Wilmot and Thornton	Propensities	I, II, IV									
Chou and Blinn	Propensities with first order coupling effects	I, II, I’, II’, VI, VIII									
COUDES	Propensities and multiple alignments	I, II, VIII, I’, II’, IV	0.31	0.30	0.23	0.11	0.07				0.11
McGregor *et al*.	Neural Network using tetrapeptide sequence	I, II, IV									
BetaTurns	Neural Network using multiple alignments and secondary structure prediction	I, II, VIII, IV	0.25	0.26			0.13				0.18
NetTurnP	Neural Network using multiple alignments, secondary structure prediction and surface accessibility	I, II, I’, II’, VIII, VI$_{a1}$, VI$_{a2}$, VI$_{b}$, IV	0.36	0.31	0.23	0.16	0.16	0.07	0.03	0.11	0.27
DEBT	SVM using multiple alignments, predicted secondary structure and predicted dihedral angles	I, II, VIII, IV	0.36	0.29			0.14				0.27
Nguyen *et al*.	SVM using multiple alignments and predicted structure	I, II, I’, II’, VI, IV	0.61	0.75	0.75	0.64	0.30	0.14			0.38
DeepDIN	Neural Network using physicochemical properties of the amino acids and secondary structure prediction	I, I’, II, II’, IV, VI$_{a1}$, VI$_{a2}$, VI$_{b}$, VIII	0.30	0.35	0.45	0.33	0.17	0.33	0.25	0.35	0.20
BetaTPred3	Random Forest using propensities, multiple alignments and secondary structure prediction	I, I’, II, II’, IV, VI$_{a1}$, VI$_{a2}$, VI$_{b,}$ VIII	0.39	0.42	0.47	0.31	0.14	0.27	0.13	0.38	0.26

Table 3: β-turn type prediction methods.

DEBT (dihedrally enhanced β-turn prediction) uses a SVM to predict β-turns and their types.[135] Information from multiple alignments and predicted secondary structures are used for the prediction as well as predicted dihedral angles from DISSPred[144]. As β-turn types are defined by the dihedral angles of the i+1 and i+2 residues the inclusion of dihedral information was thought to help increase the predictive ability of the method. DEBT predicts turn types I, II, IV, VIII. Information for 9 residues centred around the amino acid being assigned is used as the input to the SVM. For the BT426 dataset a MCC value of 0.49 was obtained for predicting β turn/non β turn. The MCCs for prediction of type I, II, IV, VIII and non-specific turns are 0.36, 0.29, 0.27 and 0.14. DISSPRED predicts helical conformations more accurately, therefore it predicts dihedral angles for type I and VIII turns more accurately as they are similar to helical conformations. This is reflected in the higher MCC values for the prediction of types I and VIII.

There are many more type I and II turns observed in proteins than there are other turn types. This means the prediction of the smaller classes of β-turns that occur less frequently is often less accurate. The algorithm can achieve high accuracy by assigning everything to the larger classes and there is less data available to "learn" what belongs to the smaller classes. Nguyen *et al*. developed a resampling technique named FOST (Flexible Over-Sampling Technique) to deal with the class imbalance between β-turn types.[145] FOST is a method to increase the density of a minority class (make more samples of the minority class) to improve prediction. A SVM was then trained and used to predict whether a sequence was a β-turn or not and what type of β turn. For the BT426 dataset the MCC values for turn types I, I', II, II', IV, VI and VIII are 0.61, 0.75, 0.75, 0.64, 0.38, 0.14 and 0.30 respectively.

1.4.2.2.3 Random Forest Based Methods

A Random Forest (RF) is another machine learning algorithm that is well-suited to classification problems such as β-turn assignment. A RF is made up of multiple decision trees which are trained on random subsets of data to make a prediction. The overall prediction given by the RF is based on the decision trees voting on which outcome is most likely.

Most methods discussed so far look at each amino acid individually and predict whether or not the amino acid is in a β-turn and if so what type of β-turn it is. By looking at each amino acid in a protein individually the algorithms sometimes assign only 3 or fewer amino acids in a row to a β turn, despite by definition β-turns requiring 4 amino acids (many β-turns overlap so greater than 4 amino acids may appear in a row as part of β-turn structures).

Singh *et al*. developed a method (BetaTPred3) that looks at all four positions in a β-turn at once.[146] Updated amino acid propensities were used as an input for a RF along with multiple alignments and predicted secondary structure data. A RF was used to predict whether a tetrapeptide was likely to form a β-turn or not. Further RFs for each turn type then predict whether the sequence forms the specific turn type or not in a binary classification. The RF that predicts the highest probability of the β-turn being that turn type is then used for assignment. Altering the window size the RF looked at from 4 residues to values between 6 and 20 to assess the effect of neighbouring residues on the prediction of β-turns did not alter the accuracy, so the window size was kept to the tetrapeptide that makes up the β turn. A MCC value for the BT426 dataset of 0.51 was obtained for predicting whether a tetrapeptide was a β-turn or not. The turn level prediction was converted into residue level prediction for comparison with previous β-turn type prediction methods. MCCs for turn types I, I', II, II', IV, VI_{a1}, VI_{a2}, VI_b and VIII were 0.39, 0.47, 0.42, 0.31, 0.26, 0.27, 0.13, 0.38 and 0.14 respectively.

1.4.3 β-Turn Mimics

As β-turns have been shown to be important for both protein folding and stability,[55, 147] many β-turn mimics have been developed which introduce non-natural fragments into the turn region.[148-150] The inclusion of a β-turn mimic can help control the conformation of a peptide potentially leading to the stabilisation of structures that would not otherwise occur. β-turn mimics are therefore useful for the study of protein conformation and have also been used to develop novel biologically active peptides such as antimicrobial peptides[151, 152] as well as inhibitors of PPIs.[153-157] Replacement of part of the peptide structure with small molecule mimics can potentially lead to improved bioavailability and pharmacokinetic properties.

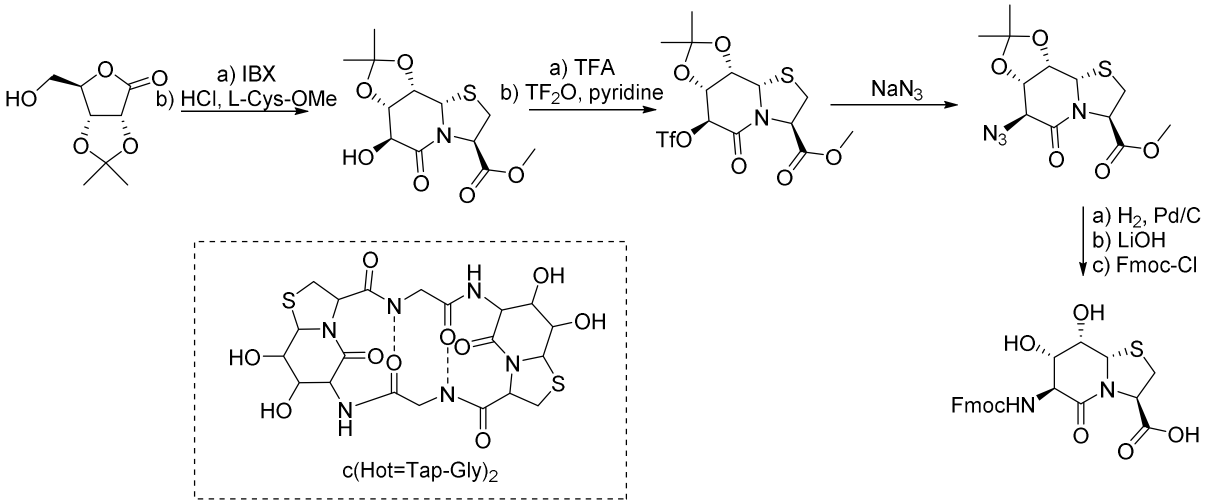
1.4.3.1 Turn inducing amino acids

The Pro-Xaa sequence, where Xaa is proline, glycine or asparagine is commonly observed to form β-turns within natural proteins.[114] Changing the L-Pro to a D-Pro at the i+1 position of the β-turn has been observed to help induce turn formation and lead to a more stable β-turn structure.[4, 43, 158-161] The dipeptide D-Pro-L-Pro is commonly used as a β-turn inducing element in peptides.[162-164] The D-Pro-L-Pro generally forms a type II' β-turn and has been shown to restrict the conformation of many, particularly cyclic, peptides by fixing the location of a β-turn.[163] This has allowed it on numerous occasions to be used to help determine the biologically active conformation of certain peptide sequences and incorporate biologically active motifs into β-hairpin structures.[164-166] Peptides designed to include the D-Pro-L-Pro motif include those with antimicrobial activity, antibody mimics as well as those targeting protein-RNA interactions and protein-protein interactions (PPIs).[165] The dipeptide sequence D-Pro-Gly is also frequently used to induce a β-turn structure in peptides.[165, 167]

In addition to D-proline, α-aminoisobutyric acid (Aib) is another frequently used amino acid for reducing conformational freedom and inducing turn structure.[168, 169] The Aib motif is often used in combination with other turn inducing amino acids such as D-Pro or glycine.[170, 171]

1.4.3.2 Small molecule-based mimics

Small molecule-based β-turn mimics aim to replicate the turn structure based on the spatial arrangements of their substituents. Bicyclic systems represent some of the earliest developed β-turn mimics where the i+1 and i+2 residues within a β-turn are replaced by the mimic. Originally designed by Nagai et. al.,[172] a bicyclic β-turn mimic was modified by Eckhardt et. al. by the introduction of hydroxyl groups (Scheme 10).[173] The new mimic, Hot=Tap, has improved turn-inducing properties, and was shown to model the β-turn structure within cyclic hexapeptides as well as a foldon protein based on the C-terminal domain of the T4 phage fibritin. Although primarily thought to mimic the type II' β-turn, the original β-turn mimic structure was shown to mimic multiple β-turn types depending on the protein environment.[174, 175] The additional substituents on Hot=Tap restrict the conformations caused by puckering of the δ-valerolactam ring which potentially means the structure better mimics the type II' β-turn. The Hot=tap mimic was introduced into the peptides using SPPS following synthesis of the unnatural amino acid.



Scheme 10: Synthesis of Hot=Tap.[173]

Other bicyclic scaffolds designed to mimic the i+1 and i+2 positions of a β-turn vary in the ring size and substituents used. Benzodiazepines are another commonly used class (Figure 8).[176] The 7-membered ring included in the structure allows some conformational freedom so benzodiazepine-based β-turn mimics can be used to mimic multiple different β-turn types.[177] Other small molecule-based β-turn mimics include those based on bicyclic lactams[178-181] and spirocyclic compounds.[182-185] A problem with the vast majority of these β-turn mimic systems is the multi-step synthesis which limits their application.

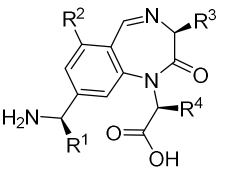
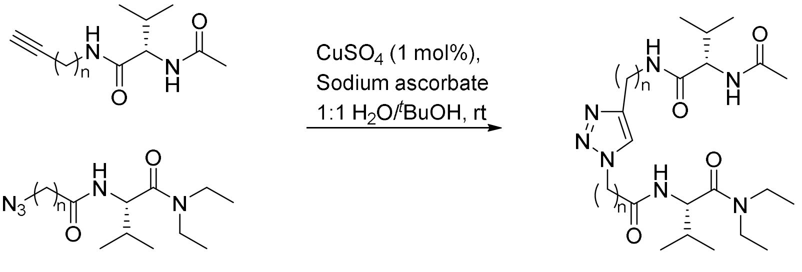


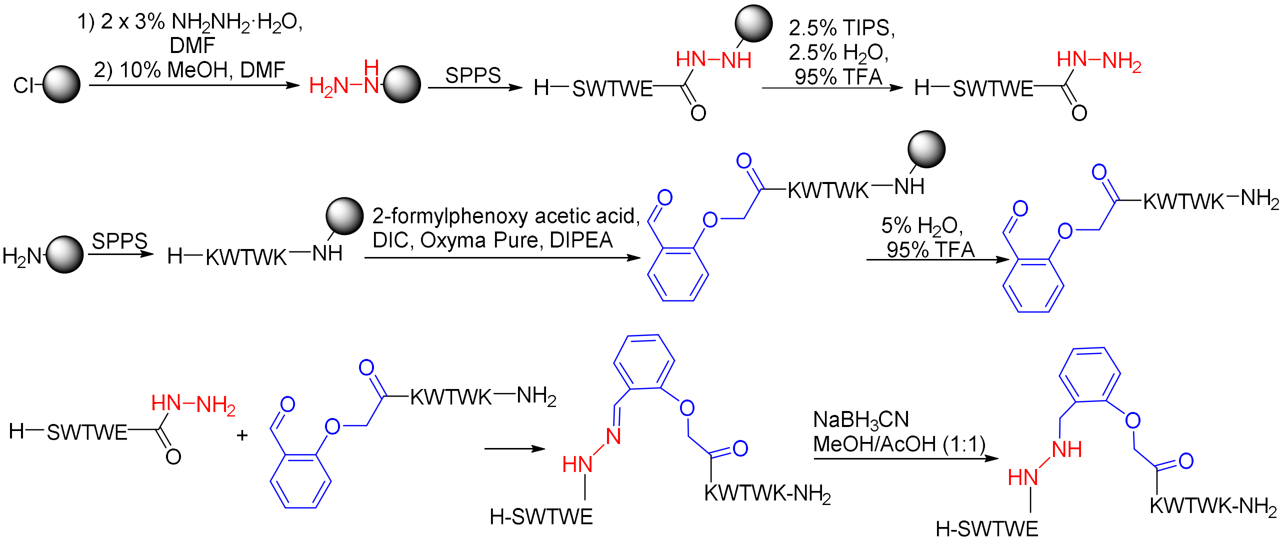
Figure 8: Benzodiazepine-based β*-turn mimics.*

Type VI turns are unique in that they are defined not just by the dihedral angles of the i+1 and i+2 positions, but also require the presence of a cis proline at the i+2 position. The presence of cis amide bonds has been shown to be important in protein folding and function and can mediate PPIs.[186-189] Mimics of type VI β-turns have been developed based on a 1,4-disubstituted[1,2,3]triazole.[190] The 1,4-disubstituted[1,2,3]triazole can be accessed through means of a copper(I)-catalysed reaction allowing formation of the β-turn mimic whilst ligating two peptide fragments together (Scheme 11).



Scheme 11: Synthesis of type VI β*-turn mimics.*[190]

Previously in the Thomson group a β-turn mimic (BTM) was developed based on hydrazine chemistry which can be used to ligate two peptide fragments together.[191] The BTM was incorporated into TrpZip1,[192] a 12 residue peptide that forms a stabilised structure containing a β-hairpin in solution with strong aromatic interactions between the tryptophan sidechains. The β-turn mimic replaced the GD hairpin turn in TrpZip1 whilst retaining the hydrogen bond and geometry characteristic of a β-turn. CD and NMR analysis showed that the BTM containing peptide showed similar conformational behaviour to TrpZip1.



Scheme 12: Synthesis of BTM-containing TrpZip based peptide.[191]

1.5 Cyclic Peptide Structure

Small cyclic peptides are generally lacking any large secondary structural elements such as α-helices or β-sheets. Reverse turns however feature heavily in their structure, meaning many small cyclic peptide structures can often be described by the type of turns present. Particularly for head-to-tail small cyclic peptides (<7 residues) similar backbone conformations are seen in the NMR and X-ray crystal structures of known cyclic peptides.

Cyclic dipeptides form a 2,5-diketopiperazine structure and are essentially rigid with only slight puckering of the heterocycle ring possible.[193] Cyclic tripeptides form strained 9-membered rings usually containing at least one cis amide bond.[194]

Cyclic tetrapeptides form a 12-membered ring which, as it contains four rigid amide groups, are typically very strained structures so can be difficult to synthesise.[58, 61, 195] One or more of the amide bonds may adopt a cis configuration which reduces strain but generally the peptide interconverts between many high energy conformations.[46]

Cyclic pentapeptides often form a structure made up of an overlapping β and γ turn,[50, 196] whereas cyclic hexapeptides frequently form a structure in solution that can be thought of as two β-turns overlayed at the i and i+3 positions (Figure 9).[197-200]

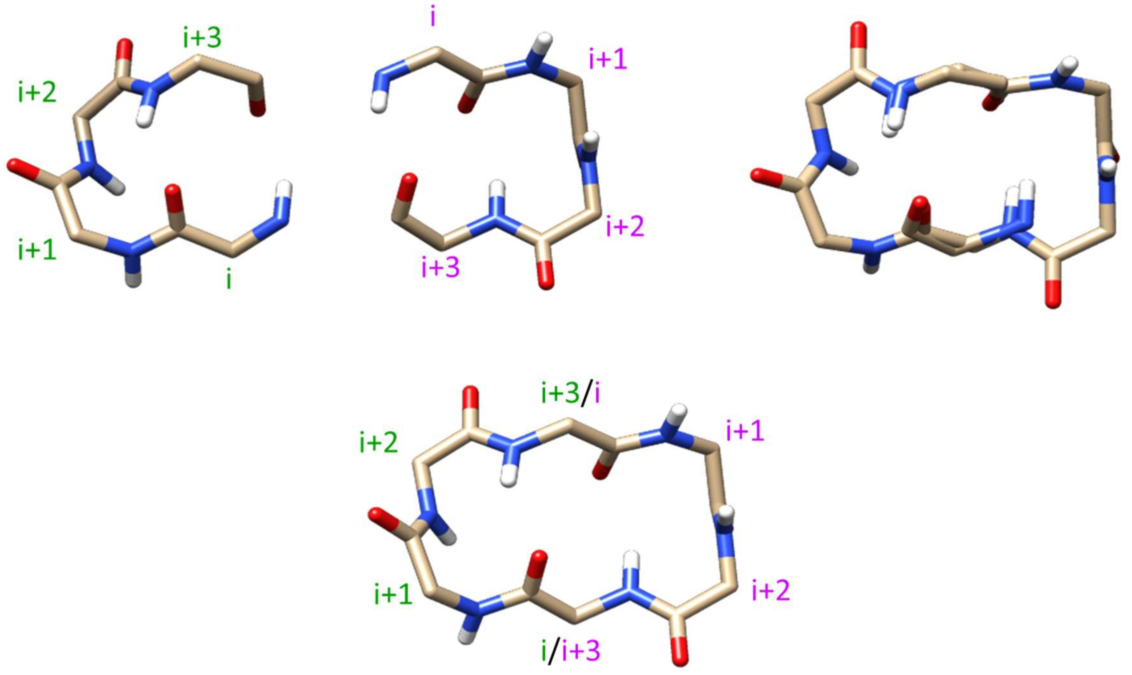


Figure 9: A cyclic hexapeptide often forms a structure composed of two overlaying β-turns.

For larger cyclic peptides there are generally more conformational possibilities with the larger structure presenting more opportunities for stabilising different conformations through various intramolecular hydrogen-bonding.[194] However, especially with the aid of turn-inducing elements such as β-turn mimics, β-hairpin conformations can often be observed, for example cyclic octapeptides often form a structure containing two β-turns when a β-turn inducing element such as D-Pro-L-Pro is included in the peptide sequence.[15, 201, 202]

1.5.1 Flexibility in Cyclic Peptide Structure

In the absence of constraining elements, cyclic peptides often convert between many conformations.[41, 53] The amino acids often change register within the turn structures e.g. in a cyclic hexapeptide an amino acid will move between the i+1 and i+2 positions of a β-turn. The type of turn can also change between conformations such as the interconversion of type I or II β-turns. Due to the flexibility of many cyclic peptides modifications can be made to reduce the conformational flexibility. Turn-inducing structures or stabilisers are particularly effective as they prevent the amino acids changing register within the structure. For example the inclusion of a single D-amino acid in a cyclic hexapeptide often leads to a structure containing a type II' β-turn with the D-amino acid at the i+1 position which fixes the register of the remaining amino acids.[44, 163, 203] *N*-methylation and the inclusion of β-turn mimics can also have an effect on the conformation.[46] For small cyclic peptides, due to the large solvent-exposed surface area the cyclic peptide structure is very dependent on the solvent it is dissolved in.[204, 205]

1.6 Fragment-based Algorithms for Cyclic Peptide Structure Prediction

Fragment-based methods for prediction of protein structure rely on the structural information from the protein data bank (PBD) to determine likely structures.[206, 207] Short fragments of known structures are used to assemble possible structures for a protein or peptide. Different algorithms then assess the generated conformations to determine the likely structures. Some fragment-based peptide prediction methods have been applied to the prediction of cyclic peptides. For smaller cyclic peptides this approach is much more difficult as, with the exception of small β-hairpins, they are too small to form secondary structural elements such as α-helices and β-sheets or form a hydrophobic core. The cyclisation can also greatly alter the conformation from unrestrained linear fragments. The accuracy of the fragment-based algorithms depends on their ability to predict all lowest energy conformations of the cyclic peptide.

I-Tasser[208, 209] selects potential structures for the peptide fragments by comparing the peptide sequence with the sequences of a non-redundant dataset of proteins from the PDB. A replica-exchange Monte-Carlo method is then used to assemble the identified fragments into potential structures. I-Tasser was primarily designed for proteins but can predict the sequence of peptides greater than 10 residues in length. The prediction method can be used on cyclic peptides by the introduction of distance restraints.

PepLook predicts the 3D structure of peptides of less than 30 residues based on a Boltzmann-Stochastic algorithm.[210] Random structures are generated using dihedral angles extracted from the structural alphabet for protein structures proposed by Etchebest *et al*.[211] Structural alphabets are libraries of structural fragments that represent local protein conformations.[212] For each iteration of randomly generated structures the probability of ϕ/ψ dihedral angles occurring is increased if they lead to lower energy structures and decreased if they lead to higher energy structures. The lowest energy structures are retained throughout the iterations and the iterations stopped when the mean probability of dihedral angles remains constant. This method was adapted for use with cyclic peptides by the introduction of distance restraints of 2.2 Å between sulfur atoms for disulfide bonds and 1.3 Å for amide bonds.[213] The structures of 38 cyclic peptides between 5 and 30 residues in length were predicted and compared to the available PDB structures. An average backbone RMSD of 3.8 Å was obtained for the rigid core region of the peptides (The N- and C-terminal ends of peptides cyclised through sidechains can be flexible so are omitted).

PEP-FOLD generates structures of peptides using a hidden Markov model to derive a structural alphabet for a peptide sequence.[214, 215] The fragments are assembled into the predicted peptide structure using a greedy algorithm[216] which uses the sOPEP coarse-grained forcefield.[217] A Monte Carlo procedure is then used to refine the structure. PEP-FOLD can be used to predict the conformation of peptides with 9 to 36 residues. It has been adapted for use with cyclic peptides, but only cyclized through cysteine sidechains via a disulfide bond. The 30 cyclic peptide structures used to test the algorithm contain between one and three disulfide bonds. An average RMSD of 3.7 Å was obtained for the rigid core regions of the peptides. The RMSD was worse for peptides with three disulfide bonds (5.4 Å). The sOPEP forcefield is not optimized for the prediction of cyclic peptides and the multiple disulfide bonds lead to a highly constrained structure. The highly constrained nature means the local conformation differs greatly from the free structure seen in NMR and the structural alphabet conformations. Presumably a similar effect would be seen with very small cyclic peptides.

The PEPstr algorithm was developed by Kaur *et al*. to predict the tertiary structure of peptides based on predicted secondary structure.[218] A further development of this system, called PEPstrMOD was designed to also be able to predict the structure of modified peptides including unnatural amino acids and cyclic peptides.[219] The algorithm uses PSIPRED[220] and BetaTurns[134] for secondary structure prediction which is then used to build possible peptide structures. The models then undergo an energy minimization and 100 ps MD simulation. The algorithm can be used to predict the structures of peptides between 7 and 25 residues in length. The PEPstrMOD algorithm was used to predict the conformation of a dataset of 34 cyclic peptides with known structures producing an average α-carbon RMSD of 3.69 Å for the rigid core regions. The structure of a peptide can be simulated in a vacuum, implicit solvent or explicit solvent. Implicit solvation represents the solvent as a continuous medium rather than simulating individual molecules as in explicit solvent. This allows for faster computational time. Repeating the MD step of the algorithm in implicit solvent rather than a vacuum improved the average RMSD of the test set to 3.53 Å. Explicit solvent was not tested.

APPTEST is a recently developed method for the prediction of the structures of peptides between 5 and 40 amino acids in length.[221] A neural network is used to derive predicted C_{α}-C_{α} and C_{β}-C_{β} distances which are then used as distance restraints in a simulated annealing protocol. Restraints can be included to allow for the prediction of cyclic peptides. Dihedral restraints were included only if those predicted by the neural network had a low standard deviation. For 34 cyclic peptide structures a backbone RMSD of 2.09 Å for the rigid core was obtained.

	Fragment	Refinement/evaluation	Size	Cyclisation Method
I-TASSER	Meta-threading using LOMETS[222]	Replica-exchange Monte Carlo	10-1500	Disulfide, head-to-tail
PepLook	Etchebest protein structural alphabet	Boltzmann-Stochastic algorithm	≤30	Disulfide, head-to-tail
PEP-FOLD	Hidden Markov model produced structural alphabet	Greedy algorithm using sOPEP coarse-grained forcefield then Monte-Carlo procedure.	9-36	Disulfide
PEPstrMOD	PSIPRED or BetaTurns predicted secondary structure	Energy minimization followed by 100 ps molecular dynamics simulation in implicit solvent	7-25	Disulfide, head-to-tail
APPTEST	Neural network derived restraints based on PDB	Simulated annealing	5-40	Disulfide, head-to-tail

Table 4: Fragment-based structure prediction methods for cyclic peptides.

Fragment-based methods are rarely able to predict the conformation of smaller peptides (< 7 residues). Even the algorithms which allow for the prediction of the structure of smaller peptides have primarily been tested on their ability to test larger cyclic peptides with test sets containing only one or two examples of the smaller peptides. The constrained nature of small peptides leads to conformations that the algorithms may be unlikely to predict as the fragments used to generate potential structures are based on linear unconstrained protein data. The cyclic peptides therefore may ultimately adopt conformations that are not explored by the algorithms. Additionally the structure of small cyclic peptides is often strongly influenced by solvent effects which are rarely modelled in detail in the fragment-based approaches.

1.7 Molecular Dynamics

Small cyclic peptides often form many conformations in solution so it can be beneficial to predict the full range of conformations and their relative stability to fully determine the likely behaviour of a peptide in a biological system. As they are small it is feasible to completely sample the conformational space of the peptides which is not possible with larger system. Molecular dynamics (MD) is therefore commonly used to predict the conformation of cyclic peptides.[56]

Cyclic peptides can take a large number of forms with different methods of cyclisation, different ring sizes, inclusion of unnatural amino acids, *N*-methylation and the introduction of non-peptidic groups. Small changes in the peptide sequence can hugely alter the conformation and therefore the properties of a cyclic peptide.[22] This means when designing cyclic peptides a trial and error approach is often used as it is difficult to rationally design and predict the structure of cyclic peptides given all these options. Computational methods that can accurately predict the conformation of a cyclic peptide in solution are therefore particularly advantageous.

Determining the conformation of small cyclic peptides by NMR can prove difficult as their low core-to-surface ratio typically means few relevant NOE cross-peaks appear. Additionally their tendency to adopt many conformations in solution means assignment can be challenging and information lost if the conformations interconvert faster than the NMR timescale. MD can therefore be a useful technique to use in combination with NMR as it can allow additional information to be obtained that would not be seen in the NMR as it can predict individual conformations. Synthesising cyclic peptides and obtaining and assigning NMR structures is a time-consuming process so screening potential sequences using MD can be a useful way to predict which sequences would likely form the desired conformation before significant time and resources are employed.

MD allows for the simulation of molecules using Newton's laws of motion.[223] Parameters needed for the calculations to describe the energy of the system based on the location of the atoms is stored in the forcefield.[224] Forcefield parameters typically are based on quantum mechanical calculations and experimental data.[225] The forcefield used during the MD simulations must accurately model the conformations of the peptides or the predicted structure will not match what is observed.

Cyclic peptides generally have a "rugged" conformational energy landscape with high energy barriers between stable conformations.[72] This means ordinary MD simulations are unable to completely explore the energy landscape of the system since the rare high energy events that are required to overcome the high energy barriers between conformations are rarely accessed in timescales used in MD simulations. Advanced sampling MD methods are designed to accelerate rare events to overcome the high energy transition states and allow complete exploration of the energy landscape of the system. This is important because otherwise depending on the starting conformation of the MD simulation different local minima/metastable states may be accessed rather than the global minima of the system.

Small cyclic peptides have a high proportion of solvent-exposed surface area with the potential to form hydrogen-bonds with water molecules. Hydrogen bonding with water can therefore be a significant determinant of energy of a conformation of a peptide, so an explicit solvent is generally preferred when modelling cyclic peptides, having been shown to improve the prediction of small cyclic peptides.[226]

Since cyclic peptides adopt multiple conformations in solution it is helpful to cluster the results of an advanced sampling technique into groups of similar conformations and look at the most common clusters and assess the stability/rigidity of the peptide in solution. Many different computational methods have been used to study cyclic peptides,[25, 56] including replica-exchange molecular dynamics (REMD)[26], accelerated molecular dynamics (aMD) [227], bias-exchange metadynamics (BE-META)[70, 226, 228] and others[56, 229].

1.7.1 BE-META

Bias-exchange metadynamcis (BE-META) is an advanced sampling technique for molecular dynamics that has been shown to be useful for studying cyclic peptides.[56]

In metadynamics simulations it is assumed that the system can be controlled by a few parameters known as collective variables (CVs) that describe the coordinates of the system in space. CVs include things such as bond angles, lengths and dihedrals.[230] CVs are chosen if they are seen to interchange relatively slowly but influence the different metastable states of the system. A history-dependent bias is introduced to the system by adding Gaussian hills to the CVs of the system at set time-intervals. The Gaussian hills add energy to the system to push the system out of previously explored conformational space (Figure 10). The height and width of the Gaussians can be set to get a balance between computational accuracy and efficiency. In this way the energy landscape is filled up allowing its exploration. Eventually the bias potential used converges to the to the negative of the free energy of the system as a function of the CVs used. The free energy landscape of the system can therefore be obtained by reversing the imposed bias potential.

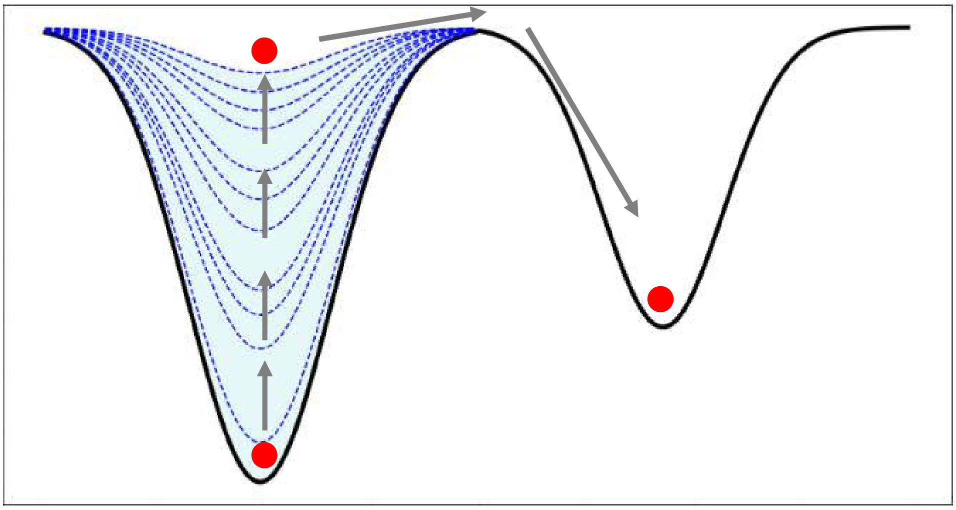


Figure 10: Metadynamics adds bias to a collective variable allowing the system to be pushed out of local minima and explore the full energy landscape.

A limitation of metadynamics is that computational time increases greatly with the number of CVs used so typically three or fewer are chosen. This can be a problem for complex systems when many CVs would be required to describe the conformation of the molecule accurately. BE-META circumvents this problem as several simulations are run in parallel.[231] Each replica has a different CV. At set time intervals replicas are randomly paired and an exchange attempt is made. Whether the exchange is successful or not is determined by a Metropolis style rule whereby the probability of finding the conformation in the other replica ultimately determines if the exchange was successful. If the exchange is successful the simulation then continues with a different bias acting on each replica (Figure 11). This helps ensure the full energy landscape of the system is explored. Additional replicas without bias can also be included in the simulation. These replicas can then be analysed directly without any additional steps to remove the effect of the bias along a CV.

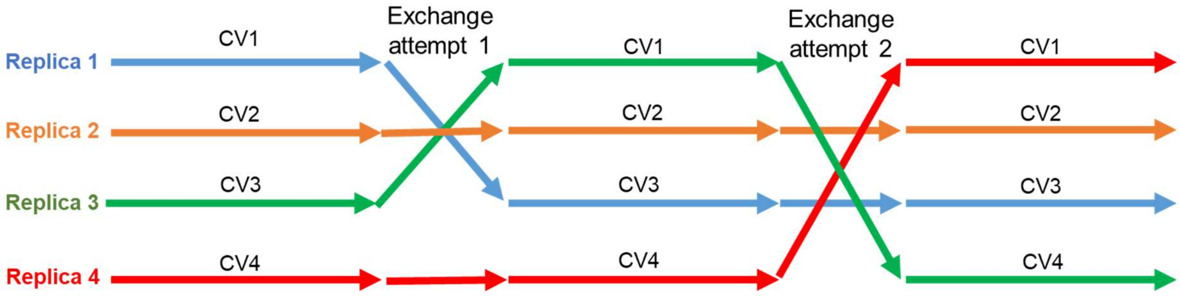


Figure 11: BE-META simulations exchange replicas with different CVs. Bias is therefore added to all CVs across the course of the simulation allowing full exploration of the energy landscape of the system.

1.7.1.1 Development of BE-META

The first use of BE-META, designed as a technique to study protein folding, was reported in 2007 by Piana *et al*.[231] They compared BE-META to the most commonly used method for studying protein folding in solution: replica-exchange molecular dynamics (REMD).[232-234] Similar to BE-META, in a REMD simulation several simulations of the same system are run in parallel and are randomly paired up and exchanged at intervals. However rather than bias on different CVs the replicas usually differ in temperature. The highest temperature is chosen to allow barriers to be crossed so the full conformational energy of the landscape can be explored. REMD on a peptide immersed in (particularly an explicit) solvent, often requires many replicas at high temperatures to properly explore a proteins energy landscape[231] as sufficient overlap in the potential energy is needed to effectively explore the energy landscape.[22] The computational time is therefore often large due to the number of replicas required, especially for large systems. BE-META generally requires fewer replicas than needed for REMD meaning the full energy landscape of the system can be explored in less computational time.

Piana *et al*. demonstrated the usefulness of their new MD method by using it to simulate the folding of a Trp cage (Figure 12) and compare it with REMD simulation.[231] The 20 residue Trp cage used was initially put into the BE-META simulation in an extended conformation and the folding in explicit solvent was looked at using 8 replicas and 5 CVs. This was compared with REMD using 48 replicas spaced between 298-576 K using the same extended starting structure used in the BE-META simulation. For both the BE-META and REMD the simulation time for each replica was 40 ns. Cluster analysis was performed on both the BE-META and REMD simulations giving similar results. The folded conformation made up the most populated cluster in both simulations having an occupancy of 45 % in BE-META and 42 % in REMD. The other clusters identified also showed similar occupancies between the simulations and represent possible metastable states the protein adopts whilst folding. Although similar results were obtained in both the BE-META and REMD simulations, the BE-META simulation had a significantly reduced computational time using approximately one sixth of the computational resources as the conventional REMD simulation.

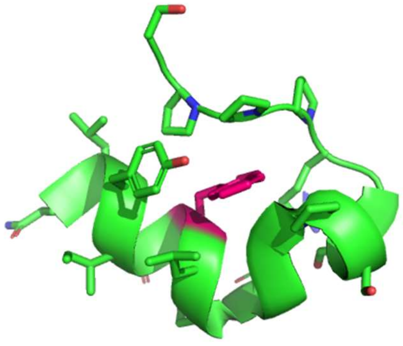


Figure 12: PDB entry 1L2Y.[235] Trp cage used in comparison of REMD and BE-META simulations. The Tryptophan of the Trp cage is shown in pink.

1.7.1.2 BE-META for Cyclic Peptide Structure Prediction

The use of BE-META to study cyclic peptides was first carried out in 2015 by Yu *et al*. who used it to model the cyclic peptide cNPF1 with the sequence c(YNPFEEGG).[22] Initially they demonstrated that other MD methods were insufficient to fully explore the energy landscape of the system in reasonable computational times. An ordinary MD simulation of 500 ns and a REMD simulation using

59 replicas were tested using an AMBER forcefield with an explicit solvent. To identify possible conformations after the simulation dihedral angle principal component analysis (PCA) and cluster analysis were performed on the last 50 ns of the trajectories of the 300 K replicas produced in the REMD simulation. Two different starting conformations were used in the REMD simulations and the two different conformations failed to converge to the same conformation even given the relatively long timescale of the simulation. BE-META was therefore tried as an alternative advanced-sampling MD technique. The φ, ψ and χ dihedral angles for each residue in the cyclic peptide were chosen as CVs and the peptide simulated for 300 ns. Unlike with REMD the two simulations from different starting conformations were seen to converge indicating the full energy landscape of the system had been explored.

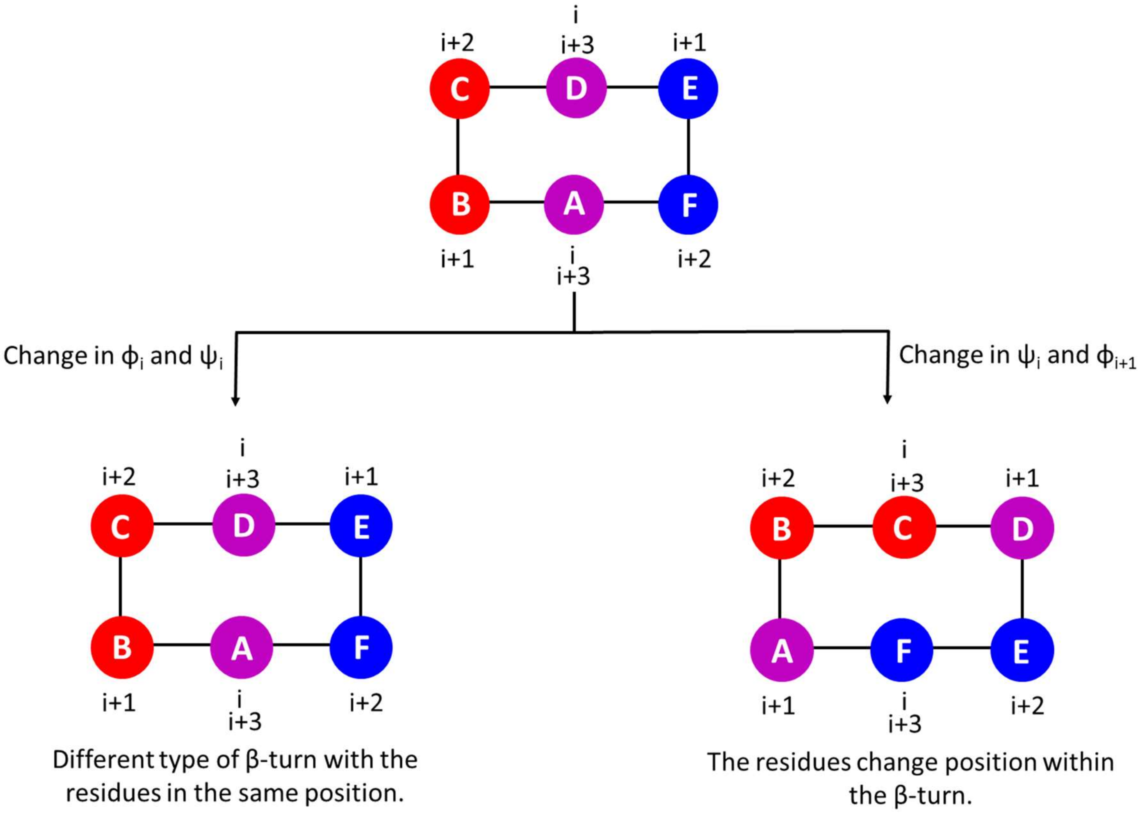
1.7.1.2.1 Choice of Forcefield

Further BE-META simulations were carried out on cNPF1 by Yu *et al*. using six different forcefields: Amber96[236], Amber99SB-ildn[237], Amber03[238], OPLS-AA/L[239] and GROMOS53a6[240] and RSFF1[241].[22] The conformations predicted by the BE-META were compared to the available NMR structure of the peptide to determine the ability of the different forcefields to accurately model the cyclic peptide structure. The forcefields tested tended to over stabilise dihedral angles in Ramachandran regions such as the α-helices and β-sheet regions which stabilise linear structures, but which do not necessarily stabilise cyclic peptide structure. This may help account for why the MD simulations indicated the cyclic peptide should adopt multiple highly populated clusters in solution rather than just the one major conformation as indicated by NMR.

Geng *et al*. tested the ability of different forcefields (OPLS-AA/L, AMBER-99SB-ildn, RSFF1 and RSFF2[242]) to correctly predict the structure of 20 cyclic peptides with known X-ray crystal structures using REMD.[243] RSFF2 performed the best producing clusters with a RMSD compared to the X-ray crystal structures of < 1.0 Å for 15 out of the 20 cyclic peptides tested. For the remaining forcefields tested only about half of the crystal structures could be accurately predicted with RMSD < 1.0 Å. The Wu group had developed the two residue-specific force fields, RSFF1 and RSFF2, based on modifying the OPLS-AA/L and Amber ff99SB forcefields respectively by editing some torsional and van der Waals parameters for specific residues so the model more accurately fits with the conformational free energy data obtained from a protein coil library.[241, 242] This may mean the better performance of RSFF2 may be down to improved backbone and sidechain torsional parameters. As RSFF2 was best able to replicate the structures of cyclic peptides it has since been used in BE-META simulations to predict the conformation of cyclic peptides.[56] RSFF2 has also been shown to allow for the most accurate structure prediction of *N*-methylated cyclic peptides.[70]

1.7.1.2.2 Collective Variables for BE-META on Cyclic Peptides

McHugh *et al*. further developed the BE-META method for studying cyclic peptides by carrying out simulations to determine the best CVs to fully explore the system whilst reducing the computational time required for convergence of two different starting structures.[244] The cyclic hexapeptide c(GGGGGG) was chosen as glycine is very flexible and is therefore ideal for studying how cyclic peptides change conformation. 100 µs ordinary MD simulations were run on the cyclic peptide. Cyclic hexapeptides often form a structure in solutions that resembles two overlapping β-turns so changes in the β-turns were monitored throughout the simulation. Two main pathways were found to be responsible for changes in its conformation (Figure 13). The cyclic peptide most commonly changes conformation by changing between different types of β-turn, with the residues at each position of the turn remaining the same. This process involved coupled changing of both the φ and ψ dihedral angles of the i+1 residue within β-turn of the peptide, with limited change to the remaining dihedral angles. Less commonly the amino acids change register within the β-turn leading to a different conformation. This process was associated with coupled changes of the ψ dihedral angle of residue at the i+1 position of a β-turn and the φ dihedral angle of the adjacent residue, i+2. It was therefore decided that when running BE-META simulations on cyclic peptides it would be best to use a 2D bias whereby bias was added to two CVs in each replica simultaneously. The CVs used would be the φ and ψ dihedral angles of a residue, as well as ψ dihedral angle of residue i, with the φ dihedral angle of the adjacent residue, i+1. So, for a hexapeptide, 12 biased replicas would be required: one for each residue with the bias on the φ and ψ dihedral angles of that residue, as well as one for each residue with a bias on the ψ dihedral angle of that residue in addition to the φ dihedral angle i+1 residue. Converged results were obtained for c(AAAAAA) and c(YNPFEEGG) in less time than was required if the 2D bias wasn't used.



i+2
i
i+3
i+1
C
D
E
B
A
F
i+1
i
i+3
i+2
Change in ϕ_i and ψ_i
Change in ψ_i and ϕ_{i+1}
i+2
i
i+3
i+1
C
D
E
B
A
F
i+1
i
i+3
i+2
Different type of β-turn with the residues in the same position.
i+2
i
i+3
i+1
B
C
D
A
F
E
i+1
i
i+3
i+2
The residues change position within the β-turn.

Figure 13: The two main pathways for changes in cyclic hexapeptide conformation.

1.7.1.3 Investigating Sequence-Structure Relationships of Cyclic Peptides Using BE-META

The use of the 2D bias for the ϕ and ψ CVs and the RSFF2 forcefield has been shown to be a useful method for accurately predicting the conformation of cyclic peptides. The BE-META simulations allow for quicker and sometimes more detailed information about the cyclic peptide structure to be determined than NMR. It is therefore an ideal method for helping to determine structural influences on the conformations a peptide will adopt.

Cummings *et al*. used BE-META to determine the ability of β-branched amino acids to help stabilise specific conformations of cyclic hexapeptides.[228] Specifically, while carrying out simulations on cyclic hexapeptides with varying quantities of glycine, alanine and valine (discussed in more detail in Chapter 4) they noticed c(VVGGVG) was predicted to have a much more stable conformation than the other cyclic peptides, with two type II turns with VG as the i+1 and i+2 residues in both turns.[245] NMR of c(VVGGVG) confirmed this to be the case. All 6 methyl groups among the 3 valines in c(VVGGVG) had unique chemical shifts, indicating a stable structure, as commonly the shifts of the methyl groups merge due to side-chain flexibility. The methyl residues of V1 had a difference of 0.19 ppm (peaks at 0.88 and 0.69), which is an unusually large difference. This along with the unusually large upshift of the methyl at 0.69 and the amide proton of this valine, suggests that this valine in particular is very well structured and therefore is important for the overall structure of the cyclic peptide. To test this, simulations were carried out where each of the valines was replaced by an alanine. All these simulations still had the two type II β-turns as the major conformation but to a lower extent. For the second and third valines this decrease was slight, but for V1, a more significant decrease was seen. They concluded that β-branching at this position is important for the stabilisation of this conformation. Further simulations swapping this valine for variously branched amino acids confirmed this.

McHugh *et al*. used BE-META to try to determine the sequence-structure relationships for cyclic pentapeptides.[196] Simulations were carried out on c(XAAAA) where X is one of the 20 naturally occurring amino acids. The most common structure contained a type II' β-turn and α_R tight turn. The location of amino acid X however varied with different amino acids occupying different registers within the turn structures. The cyclic peptide c(GFSEV) was designed based on the preference of each of the amino acids occupying each position in the βII'+α_R conformation. GF was predicted to be at the i+1 and i+2 positions of the β-turn with the α-turn centered on E. This structure was observed as the major conformation, however the cluster observed in the BE-META was not significantly greater than that of c(AAAAA) (55 rather than 53%). This indicates that neighboring amino acids influence the structural preferences of each other. The cyclic peptide series c(X_1X_2AAA) was next simulated where X=A, D, F, G, N, R or S chosen as representative of the 20 naturally occurring amino acids. A scoring function was developed based on the preferences of neighboring pairs of amino acids and used to select the sequence c(GNSRV) predicted to form a type II' β-turn at the GN and the α_R turn at R. A population of 67% was seen for this conformation in the BE-META. NMR experiments further supported these results.

Recently the group updated the scoring function used to select the c(GNSRV) sequence.[246] Rather than basing it on the dataset of c(X_1X_2AAA) peptides they used c(X_1X_2GGG). In addition to A, V, F, N, S, D and R they included D-amino acids so X could also be a, v, f, n, s, d or r. They divided the Ramachandran space into ten regions and used a scoring function based on the population of an amino acid occupying each region in c(X_1X_2GGG) dataset to predict the conformation of 50 random cyclic pentapeptides. Only 11 of these test sequences had accurately predicted most-populated structures. The scoring function was found to be best able to predict well-structured cyclic peptides but for peptides more prone to forming multiple conformations in solution a poor correlation was seen between the predicted and observed structures and it was unable to predict the lower populated conformations accurately. They therefore developed Structural Ensembles Achieved by Molecular Dynamics and Machine Learning model (StrEAMM) to improve the predictive ability.

StrEAMM used weighted least squares fitting (a type of linear regression) to assign weights to represent the free energy contribution of each of the neighboring interactions when adopting a particular structure. The contributions from the different neighboring pairs was assumed to be additive. A partition function and exponential operation were used to convert the predicted free energy for a conformation into a predicted population. The model was fit to 106 c(X_1X_2GGG) peptides as the training set and then tested on a set of 50 c($X_1X_2X_3X_4X_5$) peptides with structures determined using BE-META. The model did not accurately predict the conformation of non-well-structured cyclic peptides, only successfully predicting the most populated structures of 12 peptides. Even when the most populated cluster was accurately predicted the predictions for the remaining conformations is often inaccurate. They hypothesized improved predictions could be obtained by incorporating higher order longer-range interactions rather than just including the nearest neighbor effect on conformation. Interactions between i and i+3 residues (1,3) as well as the 1,2 interactions were therefore accounted for.

The results of the BE-META simulations on c(X_1GX_2GG) peptides were incorporated into the training set for the machine learning algorithm (StrEAMM (1,2)+(1,3)/sys). The dihedral angles of the amino acid inbetween the two residues but not the amino acid itself was taken into account when calculating 1,3 interactions as the dihedral angles of the middle residue are likely to affect the relative distance and orientation of the two amino acids. A second training set was also devised of 705 semi-random c($X_1X_2X_3X_4X_5$) peptides containing all $X_1X_2X_3$ patterns and with all X_1X_2 and X_1X_3 combinations appearing at least 15 times (StrEAMM (1,2)+(1,3)/random). Both models were then used to predict the conformations of the 50 peptides previously used as the test set. StrEAMM (1,2)+(1,3)/sys correctly predicted the largest cluster for 30 of the 50 sequences while StrEAMM (1,2)+(1,3)/random successfully predicted 43. StrEAMM (1,2)+(1,3)/random successfully predicted the conformations formed for both well-structured and non-well-structured peptides. The model therefore allows for rapid prediction of the populations of the conformations a cyclic pentapeptide is likely to adopt, with a prediction generated in under a second compared to the days it takes to run and analyse a MD simulation in explicit solvent.

The same system could be applied to larger cyclic peptide systems with the adaption of the inclusion of other long-range interactions. For example cyclic hexapeptides often form a structure composed of two overlapping β-turns so 1,4 interactions may be important as the i/i+3 positions commonly form intramolecular hydrogen-bonds. Large quantities of data are required to accurately train machine-learning algorithms however and there are limited numbers of available solution structures of small cyclic peptides. The StrEAMM algorithm uses data produced by BE-META simulations to predict the results of further BE-META simulations. Its accuracy for predicting the conformation of cyclic peptides is therefore highly dependent on the results of the BE-META simulations being accurate models of the actual cyclic peptide structure. The RSFF2 forcefield used has previously been shown to allow for structure prediction of cyclic peptides but StrEAMM could be updated if a new forcefield is proven to allow for more accurate modelling of cyclic peptide structure.

1.7.2 Examples of Using Molecular Dynamics to Design Biologically Active Cyclic Peptides

Computational methods currently offer the best solutions for predicting the conformation of cyclic peptides designed with a specific biological function. Without the inclusion of specific groups known to occupy particular secondary structures such as β-turn mimics it can be difficult to predict the specific register and conformation the peptides will adopt. Limited understanding of sequence-structure relationships currently limits prediction meaning a trial-and-error approach remains common for designing cyclic peptides for a specific purpose. This is a costly and time-consuming process, especially due to the tendency of cyclic peptides to adopt multiple conformations in solution which can make structural characterisation difficult and complicate understanding of how the sequence alters the conformation.

Macaluso and Glen used REMD to predict the conformation of a series of cyclic peptides based on Apelin-13,[247] a small peptide known to activate the G-protein coupled receptor (GPCR) APJ which is a coreceptor for HIV cellular entry.[248] The Apelin-13 peptide (QRPRLSHKGPMPF) has been implicated in cardiovascular disease, HIV infection and cancer.[249-251] Four head-to-tail cyclic peptides were designed based on the Apelin-13 peptide and their structure determined using REMD. By cyclizing the structure the backbone conformation of the peptides was constrained inducing the formation of specific turns within the sequence. The smaller two peptides (c(QRPRLS) and c(QRPRLSH)) formed conformations with well-defined type II turns at the RPRL sequence throughout most of the simulation. The larger two peptides (c(QRPRLSHK) and c(QRPRLSHKG)) however showed a preference for formation of a turn structure at the RLSH sequence stabilized by the hydrogen-bonding of the hydroxy sidechain of the serine at the i+2 position with the peptide backbone. Such hydrogen bonding has previously been shown to help stabilize type I β-turns.[114] Binding assays were carried out with the APJ receptor. Only the peptides with the β-turn centered at the RPRL sequence showed significant binding to the receptor supporting the importance of a β-turn within the Apelin-13 structure for binding. In such a way computational modelling can be used to rationalise the binding affinity of a series of peptides to provide information for future designs. The results also demonstrate that cyclisation can lead to reduced affinity of binding if the wrong structure is stabilized. Accurate prediction methods are therefore important for designing cyclic peptides with a specific structure.

Razavi *et al*. used computational screening to design cyclic peptides designed to mimic the β-hairpin conformation seen in the LapD protein, found to be important for bacterial biofilm formation.[252] The cyclic peptides were covalently crosslinked through amino acid sidechains so could theoretically be synthesized through olefin metathesis. Two peptide designs were investigated based either on the sequence VSRGWEQAA with two crosslinks: one between the Ser2 and Gln7 sidechains and the Val1 and Ala9 backbones (Figure 14). The second design was based on the shorter SRGWEQ sequence with only the one crosslink between the Ser and Ala backbones. Variations were introduced based on the inclusion of L- or D-valine or the E/Z isomers of the double bond in the crosslinks. REMD was used to model the cyclic peptides using an AMBER-96 forcefield for the standard amino acids and the GAFF[253] forcefield for the crosslinks. Free energy profiles of the resulting structures were generated as a function of the backbone and β-carbon RMSD compared to the native LapD hairpin structure for the SRGWEQ sequence. The cyclic peptides containing 9 rather than 7 residues were found to have lower RMSDs. The two crosslinks likely help enforce the hairpin conformation. A different preferred geometry was seen for the two crosslinkers, the central one gave lower RMSD values with an E- rather than Z-olefin geometry whereas the reverse is seen for the peripheral linker. Although an entirely computational study the process used shows how computational methods can be used to help screen and potentially design drug molecules.

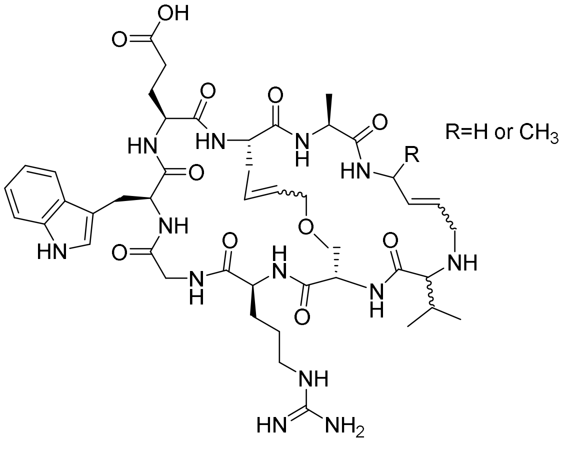


Figure 14: The 9-mer cyclic peptides investigated by Razavi et al.[252]

Huang *et al*. carried out a series of BE-META simulations on poly-glycine peptides between 5 and 15 residues in length.[254] Cyclisation of peptides can improve binding affinity as the cyclic structure is more rigid so less entropy is lost upon binding. By comparing linear and cyclic peptides Huang *et al*. estimated the entropic effects between linear and cyclic peptides based on the backbone dihedral angles. They found for cyclic peptides greater than 9 residues in length the effect on configurational entropy was minimal. The reduction in entropy for smaller cyclic peptides primarily resulted from correlated changes in dihedral angles rather than a smaller dihedral angle distribution. In particular changes in ψ_i and ϕ_{i+1} were highly linked for cyclic penta- and hexapeptides. Despite the similar broad dihedral angle distributions some differences in preferred dihedral angles were seen for the smaller cyclic peptides indicating certain dihedral angle patterns may be necessary to obtain ring closure. It was assumed that glycine being very flexible would explore many possible conformations cyclic peptides are able to adopt. Hot loop regions were therefore compared with the conformations seen in the poly-glycine simulations to determine if a cyclic peptide could be used to mimic the structure. 120 out of the 193 loops tested had a structure similar to a conformation that appeared in one of the BE-META simulations. A similar conformation may indicate a cyclic peptide size to choose when incorporating loop regions into cyclic peptides. This potentially offers a starting point for the design of biologically active cyclic peptides designed to target PPIs based on hot loop regions. If a conformation does not occur in the poly-glycine simulations it may still be possible to design cyclic peptides which mimic the structure especially with the use of unique sequences such as proline or *N*-methylation which induce restraints on the structure.

1.7.3 Modelling cis/trans Proline Isomerisation

The inclusion of a proline residue has been shown to help stabilise the conformation of cyclic peptides due to the limited dihedral angles that can be achieved in the ring system.[46] Additionally the inclusion of a proline can help in the synthesis of cyclic peptides as it introduces a kink in the peptide chain that can help bring the ends together for the cyclisation steps. [57, 58, 255] Proline is unique among the naturally occurring amino acids in that it is more frequently found in the cis form, with around 6% of xP proline residues found in the cis conformation, with some evidence cis proline is more common in cyclic peptides.[256, 257] Whether Proline is cis or trans is determined by the ω dihedral angle which is at 0° for cis and 180° for trans. A relatively high energy barrier separates the cis and trans conformations so when using MD advanced sampling techniques are necessary to model the cis/trans isomerisation.

Despite the usefulness of proline in synthesising and controlling the conformation of cyclic peptides there are limited examples where cis/trans isomerisation of proline has been modelled in a cyclic peptide. Primarily only the trans form is considered.

Kamenik *et al*. used accelerated molecular dynamins (aMD) to predict the conformation of c(PSlDV),[227] a small cyclic peptide known to bind integrin.[258] The NMR structure of c(PSlDV) contains both a cis and trans conformation. The aMD predicted three conformations for c(PSlDV): one containing trans proline and two containing cis proline. The cis to trans ratio seen in the simulation was 1:0.3, similar to the 1:0.25 ratio seen in the NMR structure. The cis conformation is the major structure so MD techniques that do not allow for cis/trans isomerisation would not be able to predict the major conformations of the peptide.

Wu *et al*. used parallel tampering to predict the conformation of four cis proline-containing cyclic peptides.[259] They used an Amber forcefield[260] with an implicit solvent. Although structures seen in the NMR were also seen in the simulations, it is difficult to determine if the cis to trans ratio of conformations was correctly predicted as water was used as a solvent for the simulations, but DMSO, chloroform and acetonitrile were used to obtain the NMR structures. The solvent effects were seen to influence cis/trans isomerisation so using explicit solvent could potentially lead to better modelling of the system.

Despite rarely being used in cyclic peptide prediction cis/trans isomerisation of proline residues has been shown to be important in the folding and function of a variety of proteins.[186, 261, 262] Various advanced sampling techniques have therefore been used to model cis/trans isomerisation of proline in proteins, usually to determine the mechanism for isomerisation which is often catalysed by another protein.[263-266] Advanced sampling techniques used to model proline cis/trans isomerisation include aMD,[267, 268] umbrella sampling[269] and metadynamics.[264, 270, 271]

1.7.3.1 Metadynamics for cis/trans Proline Isomerisation

Metadynamics has previously been used to predict the ratio of cis to trans proline in peptides. As the ω dihedral is used to describe if proline is cis or trans it would be an intuitive choice for the CV to use. However the ω dihedral angle of the proline is coupled to the out-of-plane deformation of the proline nitrogen. This can be described by the improper dihedral angle η between the α-carbon of the residue prior to the proline, the α-carbon of the proline, the proline nitrogen atom and the δ-carbon of the proline (Figure 15). This angle should therefore be included with the ω angle when performing metadynamics on a proline-containing peptide to obtain the full energy landscape. The ψ angle of the Proline has also been shown to be an important factor in cis-trans isomerisation of the proline residue so should be included as a CV.[272, 273] The ψ angle describes the orientation of the C-terminal amide which can interact with the lone pair of the proline nitrogen or the carbonyl oxygen of the residue preceding proline. This may decrease the double bond character of the bond between the carbonyl carbon and the nitrogen of the proline, lowering the barrier to isomerisation.

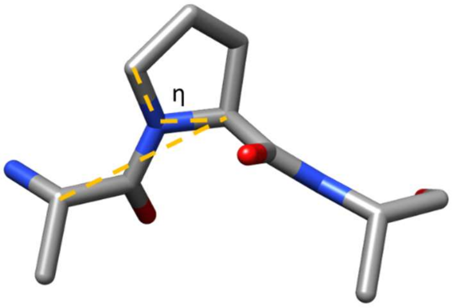


Figure 15: The η improper dihedral angle in proline.

Rather than using the three dihedral angles ω, η and ψ as CVs in a metadynamics simulation, Fischer *et al*. proposed an alternative to the ω and η CVs.[272] The improper dihedral angle ζ (Figure 16) takes into account cis-trans isomerisation described by ω and the nitrogen pyramidalization described by η.[272] This allows for the metadynamics to be carried out with two rather than three CVs. As computational time required for running metadynamics simulations increases exponentially with number of CVs this is an advantage. As such a combination of ζ and ψ are generally chosen as CVs when running metadynamics to look at the cis-trans isomerisation of proline.[270, 274]

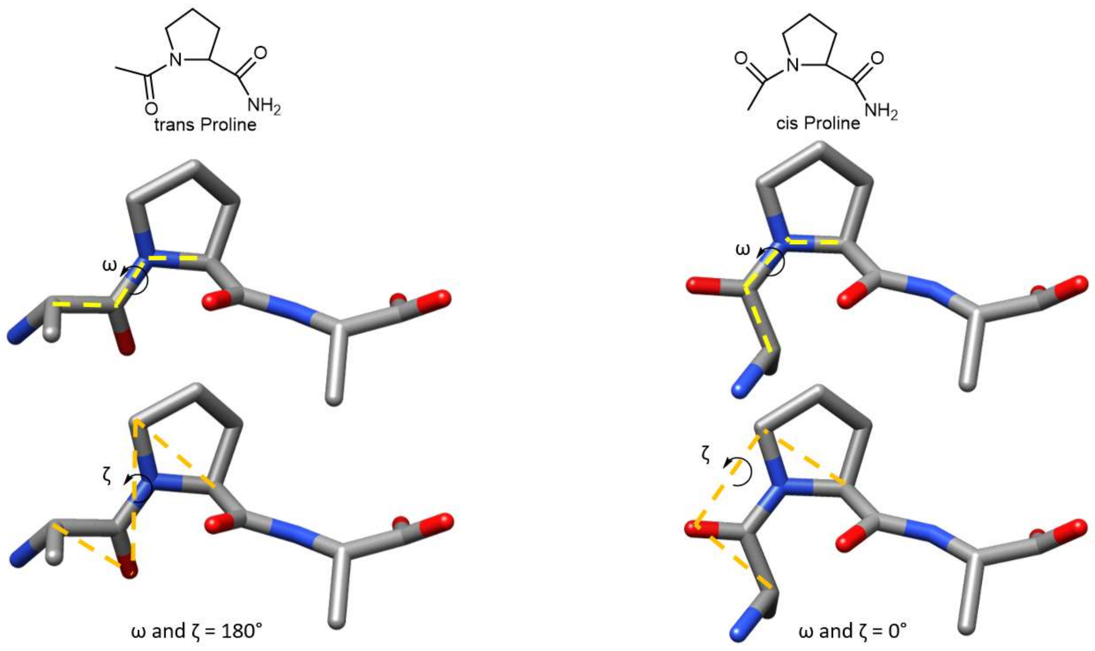


Figure 16: ω and ζ dihedral angles in trans and cis proline.

Explicit water is required as it has been shown that water influences the cis-trans isomerisation of proline by preventing an intramolecular hydrogen-bond within the peptide being as favoured in the trans conformation.[270] It is expected that there will be a slight error in predicting the energy barrier for cis-trans isomerisation as electronic structure effects that occur in transition states between the isomers require the partial double bond character between the carbonyl carbon and proline nitrogen to be broken. This cannot be simulated in an ordinary metadynamics simulation based on Newtonian mechanics. Density-functional theory (DFT) based calculations would offer more accurate results but would be significantly more time consuming, especially on larger peptides, and previous comparisons between Newtonian-based metadynamics and DFT calculations have shown that metadynamics still offers a reasonable estimate of cis-trans isomerisation.[270]

1.8 Thesis Scope

Despite their usefulness as potential drug molecules it still remains difficult to predict the structure of small cyclic peptides, especially those composed of only the 20 naturally occurring amino acids. BE-META has proven to be a useful technique for determining the effect of small changes on the structure of cyclic peptides. Current implementations however do not take into account cis/trans proline isomerization. Additionally although a lot faster than synthesising and determining a peptide structure by NMR, BE-META still requires a few days of computational time per cyclic peptide. Screening methods to help narrow down cyclic peptide structure prior to MD are therefore desirable.

Synthesis of small cyclic peptides can often be difficult with side reactions including oligomerization and epimerization common. As the cyclisation step is the most problematic, alternative ligation methods to cyclise a peptide can often lead to increased yields than can be obtained from standard amide bond formation. Available cyclisation methods may put restrictions on peptide sequence, may not be compatible with head-to-tail cyclic peptides or not allow for on-resin cyclisation making the use of large-volumes of solvent necessary. Multistep synthesis of unnatural amino acids may also be necessary to introduce functional groups required for the cyclisation reaction.

Cyclic peptides often form many conformations in solution. Having a single rigid conformation can lead to increased affinity in binding to a target as less entropy is lost upon binding and the conformation can be preorganised into the correct orientation for binding. Determining sequences likely to form cyclic peptides with a single major conformation compatible with a desired structure is therefore an aim of structure prediction. Introduction of β-turn mimics within the cyclic peptide structure can offer an alternative means of controlling cyclic peptide conformation.

This thesis looks at the use of BE-META in combination with analysis of β-turns from a database for the prediction of cyclic peptide structure. As cyclic hexapeptides often form a structure in solution composed of two overlapping β-turns, information about the structure may be determined using available information about β-turns extracted from a database. The introduction of a β-turn mimic into a cyclic peptide whereby the formation of the β-turn mimic occurs through a ligation reaction that also causes cyclisation of the peptide is also explored. This allows for both an efficient cyclisation reaction and induces structure allowing for cyclic peptides with a more rigid structure. The design of well-structured cyclic peptides either through the use of predicting sequences likely to form well-defined structures or the addition of the β-turn mimic to add additional functional elements are therefore addressed.

2 Inclusion of a Proline Replica in BE-META of Cyclic Peptides

Previous studies have shown that RSFF2 is the best forcefield at replicating the structure of cyclic peptides.[243] However these studies were on peptides containing all trans peptide bonds. Proline is unique amongst the other naturally occurring amino acids in that the cis amide bond is seen to occur in protein structures approximately 6% of the time.[256] In the relatively strained structure of cyclic peptides this percentage is likely to increase, with cyclic tripeptides and tetrapeptides generally containing at least one cis peptide bond even if they do not contain proline.[226] Proline is also useful for the synthesis of cyclic peptides, generally leading to increased yields, as it introduces a kink in the peptide structure which can help bring the ends together for cyclisation.[275] It would therefore be beneficial for computational techniques to be able to predict the conformation of cyclic peptides containing cis proline. Despite this there are very limited examples of computational methods used to predict the cis/trans ratio of proline in cyclic peptides.[227, 259]

Advanced sampling techniques are necessary for modelling proline cis/trans isomerisation as there is a relatively high energy barrier between the two energy minima (approximately 20 kcal/mol)[276] so isomerisation occurs across longer timescales than can be reasonably modelled by MD. Advanced sampling techniques are already necessary to fully explore the energy landscape of cyclic peptides so any advanced sampling techniques for modelling proline cis/trans isomerisation would have to be compatible with existing methods for modelling the rest of the peptide.

Slough *et al*. used bias-exchange metadynamics (BE-META) to predict the conformation of the *N*-methylated cyclic peptides c(aAAAAA) and c(aAAAAA) where the *N*-methylated residues are underlined, and D-amino acids are represented by lowercase letters.[70] *N*-methylated amino acids, similar to proline, have a higher proportion of cis amide bonds. Both the peptides which were modelled have structures which have been determined by NMR with c(aAAAAA) having a cis amide bond for the second *N*-methylated alanine and the remaining amide bonds in the two peptides being trans. During BE-META of cyclic peptides bias is added to the ϕ and ψ dihedral angles of each residue to allow full exploration of the energy landscape of the system. Some cis/trans isomerisation of the *N*-methylated residues was seen to occur in the BE-META due to the bias added to the ϕ and ψ dihedral angles throughout the simulation despite there being no bias added to the ω dihedral angle. When all the conformations with the same cis or trans amide bonds as seen in the NMR structures were extracted from the simulation (i.e. c(aAAAAA) conformations containing a trans/cis amide bond for the *N*-methylated residues respectively and c(aAAAAA) conformations with all trans amide bonds) then the RSFF2 forcefield correctly predicted the solution conformations of the peptides. Simulations were also carried out where bias was added to the ω dihedral angle of the *N*-methylated residues to determine if BE-META could be used to predict the cis/trans isomers seen in the final structures. Although the RSFF2 forcefield correctly predicted the conformation of the peptides when in either the correct cis or trans conformations, it did not accurately predict the presence of a cis amide bond in c(aAAAAA). Similar results were seen with different water models so it is likely the energetic terms for the peptide in the forcefield overpredicts the stability of the trans conformation for *N*-methylated amino acids.

2.1 Inclusion of a Proline Replica in BE-META

One of the most commonly used computational methods to predict the conformation of small cyclic peptides in solution is BE-META; [22, 244] however it currently has not been used to predict the presence of cis proline in cyclic peptides. BE-META using RSFF2 has been shown to lead to accurate prediction of the structure of small cyclic peptides.[243] Currently however this only applies to peptides with all trans peptide bonds. RSFF2 may not be able to accurately predict the structure of peptides containing cis proline. Fischer *et al*. previously suggested the use of the improper dihedral angle, ζ (Figure 17), to use as a CV during metadynamics to predict the cis/trans ratio of proline in proteins.[272] It should therefore be possible to include an extra replica in the BE-META of proline-containing cyclic peptides which biases the ζ angle of proline. Inclusion of such a replica may allow for the prediction of peptide structures containing cis proline such as those with a type VI β-turn. For peptides which form many conformations it may be possible to predict the cis/trans ratio of prolines within the peptide.

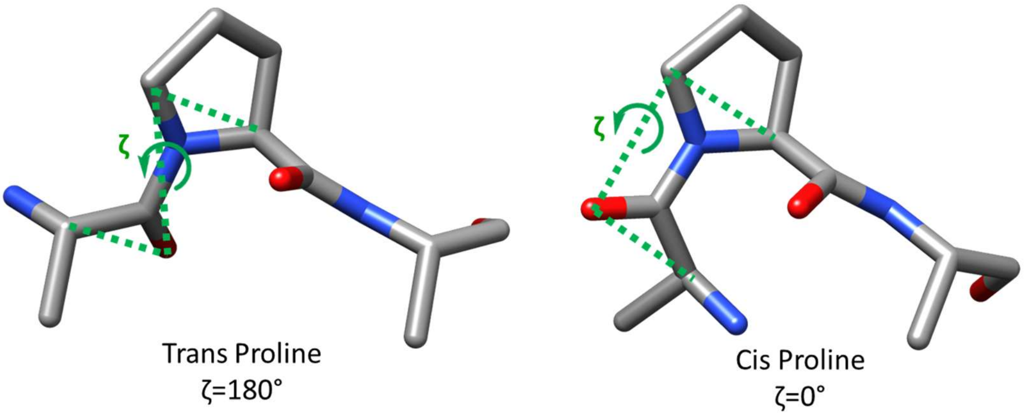


Figure 17: The improper dihedral angle ζ in trans and cis proline.

To choose sequences to test the BE-META of cyclic peptides with an additional proline replica, overlays were used. As cyclic hexapeptides often form a structure that is the equivalent of two β-turns overlapping at the i/i+3 positions, extracting β-turns from a database and overlaying them at these positions may help design cyclic hexapeptides (Figure 18). The amino acids in the overlay have already been seen to adopt the dihedral angles that they potentially would in a cyclic peptide environment. Overlays were created based on the RMSD of the overlapping i/i+3 residues. BE-META was used to predict the conformations of the peptide which were determined by NMR.

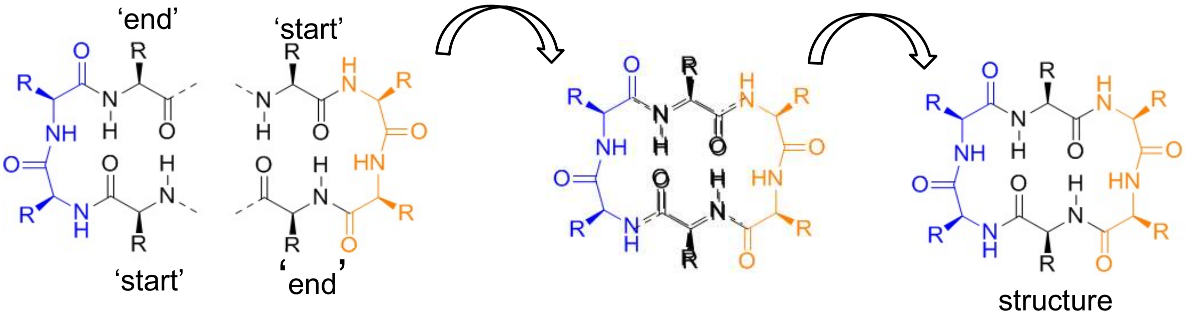


Figure 18: Formation of an overlay from two β-turns to predict the structure of a cyclic hexapeptide.

2.1.1 YPWG-RNKE Series of Overlays

Two β-turns with the sequences YPWG and RNKE were found with a backbone RMSD of the overlapping i/i+3 residues of 0.94 Å. The YPWG β-turn (from PDB: 4KV9) contains a proline residue at the i+1 position of one of the turns to test the inclusion of the proline switching replica in the BE-META. The inclusion of tyrosine and tryptophan may allow for a higher percentage of cis proline as CH-π stacking with proline has previously been shown to increase the occurrence of cis proline.[277, 278] RNKE (found in PDB: 5ILB) contains asparagine and lysine at the i+1 and i+2 positions of the β-turn to aid with synthesis of the peptides - lysine helps with solubility, and asparagine allows for sidechain anchoring to allow cyclisation on-resin.[71, 275]

There are four possible sequences from overlaying the two β-turns at the i/i+3 positions depending on which of the amino acids at the overlaying positions are included (Figure 19). If the overlay can be used to predict the structure of the peptide then the amino acids will remain in the same register with the PW and NK subsequences forming a type I and I’ turn respectively. Alternatively if the sequences are more flexible and other conformations are seen, the inclusion of the proline replica in BE-META may allow for the prediction of cis proline-containing sequences that otherwise may not be predicted.

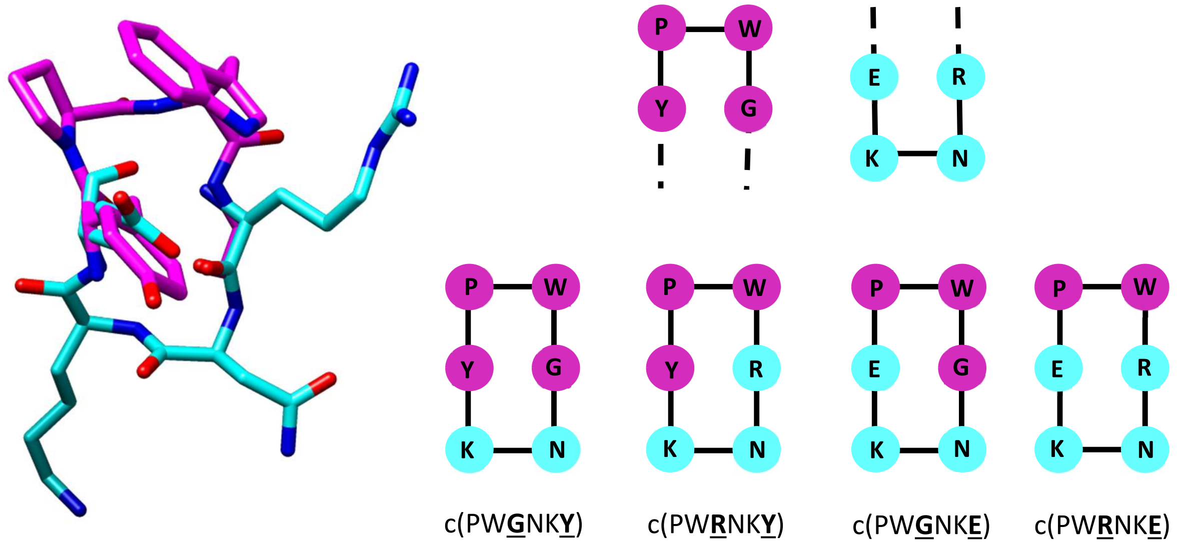


Figure 19: The overlay between YPWG and RNKE β-turns to produce the four possible cyclic hexapeptides.

2.2 BE-META

2.2.1 BE-META with the Inclusion of a Proline Replica

BE-META was carried out on the four sequences: c(PWGNKY), c(PWRNKY), c(PWGNKE) and c(PWRNKE). The CVs to allow for full exploration of the energy landscape of small cyclic peptides include a replica for each amino acid in the sequence biasing the ϕ and ψ dihedral angles as well as a replica for each amino acid biasing the ψ dihedral angle as well as the ϕ dihedral angle of the next residue in the sequence. As well as these replicas an additional replica was used with bias added to the ζ improper dihedral angle and the ψ dihedral angle of the proline to allow for cis/trans isomerisation (see section 10.5.3.1 for details of the bias added throughout the simulation). Five unbiased replicas were also included for analysis. This means 18 replicas were used in total for the cyclic hexapeptides (Figure 20).

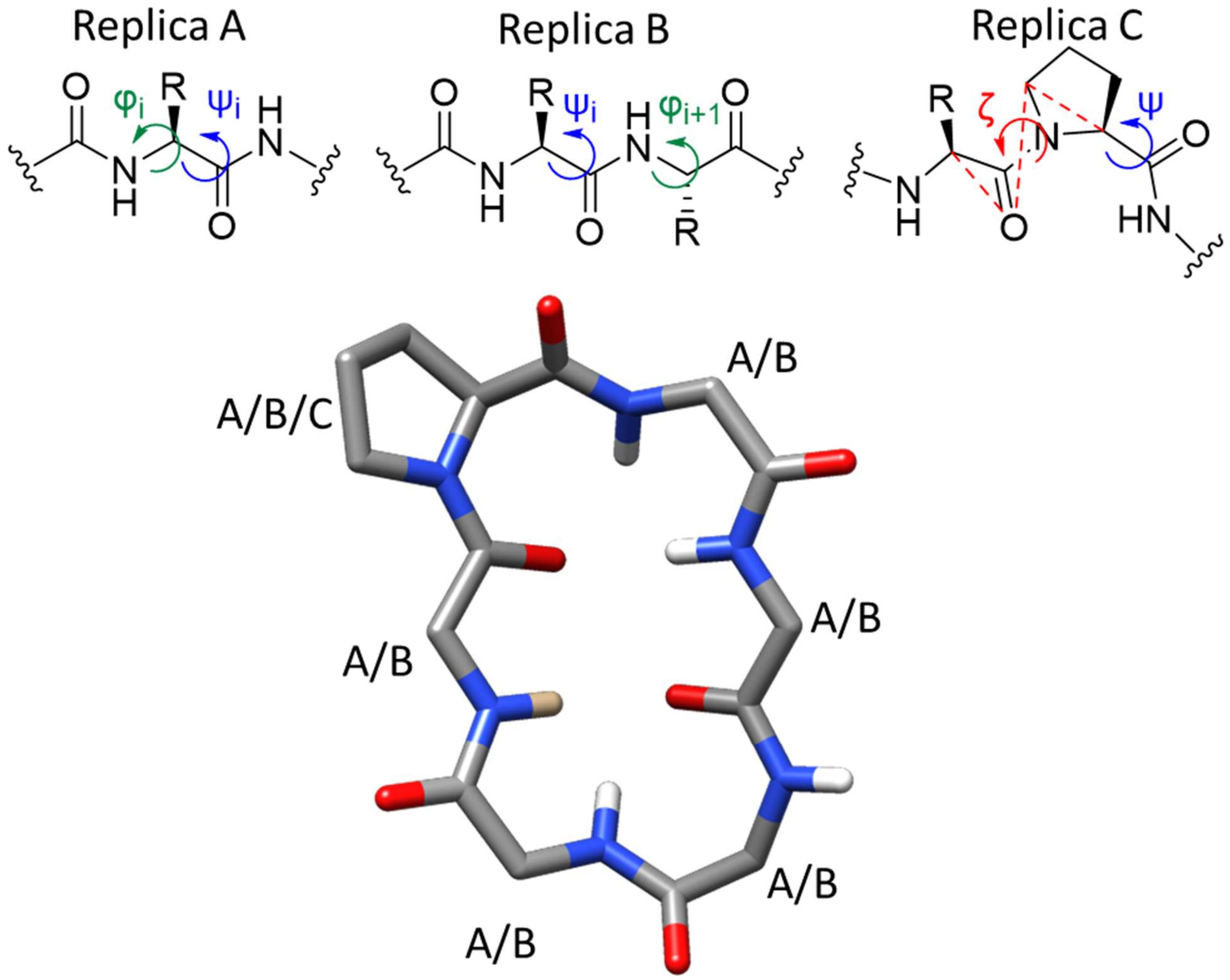


Figure 20: Replicas used in the BE-META of proline-containing cyclic peptides. There is a replica for each residue with the φ and ψ dihedral angles as CVs (replica A) as well as a replica for each residue with bias on the ψ dihedral angle in addition to the φ of the i+1 residue (replica B). Proline has an additional replica (replica C) with bias on the ζ and φ dihedrals. Five unbiased replicas are used for analysis.

If the proline replica prevents full exploration of the system then convergence will not be reached. To test convergence of the BE-META simulations two simulations are run per sequence from different starting conformations. The trajectories are clustered to identify the different conformations the peptide forms (see section 10.6). If the same final clusters in similar proportions are obtained from each simulation convergence has been reached. For the four proline-containing sequences tested convergence was reached within 200 ns so the additional proline replica does not prevent convergence. The starting conformation of the peptide can originally contain either cis or trans proline and the same conformations in similar proportions will be seen.

The predicted conformations for c(PWGNKY), c(PWRNKY), c(PWGNKE) and c(PWRNKE) are shown in Table 5. The register is written as the i to i+3 positions of the two overlapping β-turns so for example, for c(PWGNKY) cluster 1 the conformation contains a type VI and a type II' turn. The i to i+3 positions of the type VI turn are KYPW and the i to i+3 positions of the II' turn are WGNK. W and K occupy the two overlapping i/i+3 positions in this conformation. The BE-META predicts multiple conformations for each sequence, with a mixture of conformations containing cis or trans proline.

Varying the sequence by a small amount (one amino acid) can significantly alter the predicted conformation of the cyclic peptide and the cis/trans ratio of proline. For example c(PWRNKY) is predicted to have 100% cis proline-containing clusters whereas swapping the tyrosine for glutamic acid changes the prediction to a completely different conformation with predominately trans proline.

	Cluster	Population %	Register	Conformation	% cis proline-containing clusters
c(PWGNKY)	1	41	KYPW/WGNK	VI+II’	54
	2	33	YPWG/GNKY	I+I’	
	3	13	KYPW/WGNK	VI+I’	
	4	13	YPWG/GNKY	I+IV$_3$	
c(PWGNKE)	1	51	EPWG/GNKE	I+IV$_3$	49
	2	40	KEPW/WGNK	VI+II’	
	3	9	KEPW/WGNK	VI+I’	
c(PWRNKE)	1	86	EPWR/RNKE	I+IV$_3$	14
	2	14	KEPW/WRNK	VI+II	
c(PWRNKY)	1	66	KYPW/WRNK	VI+I’	100
	3	34	KYPW/WRNK	VI+I	

Table 5: The clusters seen in the BE-META simulations of c(PWGNKY), c(PWGNKE), c(PWRNKE) and c(PWRNKY). Clusters containing a cis proline are shown in blue.

Only one cluster across the four sequences (c(PWGNKY) cluster 2) contains the I+I’ conformation seen in the overlayed β-turns used to generate the sequences. The I+IV$_3$ conformation is much more common for that register where the NK sequence forms a type IV$_3$ turn rather than a type I’ turn. It is therefore unlikely that the overlaying of two β-turns based on lowest RMSD of the i/i+3 positions allows for prediction of what structure will occur in a cyclic hexapeptide. The protein environment in which the RNKE β-turn was found is very different to the cyclic peptide environment potentially leading to the different preferred turn type.

2.2.2 BE-META without the Proline Replica

Simulations were performed on the same sequences without the inclusion of the proline replica. It was thought the proline replica was necessary for cis/trans isomerisation of the proline residue, however cis proline is still seen in the BE-META without its inclusion. The same conformations in very similar proportions are seen in the simulations. For example the same major clusters are seen in the BE-META of c(PWRNKY) with/without the proline replica (Table 6). There is an additional small cluster seen in the simulation without the proline replica but the variation is within the noise of the simulations. However, without the inclusion of the proline replica, convergence takes longer to be reached. The c(PWRNKY) simulation reaches convergence when the proline replica is included in the BE-META after 200 ns, but without the proline replica convergence is not reached until 300 ns. The remaining sequences also take 100 ns longer to reach convergence without the inclusion of the proline replica, with the exception of c(PWGNKE) which still hasn’t reached convergence after 300 ns despite reaching convergence after 100 ns when the proline replica is included.

Conformation	Population with Proline Replica (%)	Population without Proline Replica (%)
VI+I’	66	70
VI+I	34	21
I+I’	0	9

Table 6: Clusters seen in the BE-META of c(PWRNKY) with and without the inclusion of the proline replica.

The appearance of cis proline occurs much more rapidly in the simulations with the inclusion of the proline replica (Figure 21). This is likely why the inclusion of a proline replica allows for convergence to be reached 100 ns earlier. Running metadynamics simulations for extended lengths of time, in addition to taking up significant amounts of computational time, also risks pushing the system into configurational space which is no longer physically meaningful.[279]

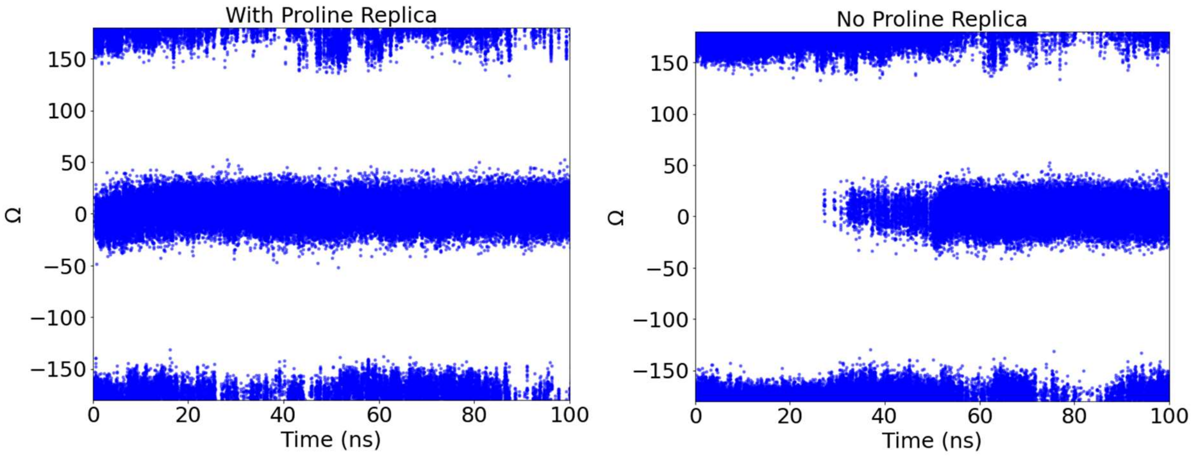


Figure 21: Changes in the ω dihedral angle of proline in an unbiased replica of the BE-META for c(PWGNKY) with and without the inclusion of the proline replica.

The appearance of cis proline despite the lack of the proline replica could potentially be due to the size of the macrocycle. The cyclic hexapeptide is relatively small so there may be enough strain within the system for the proline to switch to the cis conformation without the bias added to the ζ CV. Alternatively the existing CVs necessary for exploring the energy landscape of the cyclic peptide contribute to proline cis/trans isomerisation. The ψ dihedral angle of the proline has previously been associated with cis/trans isomerisation.[272] Two replicas, other than the proline replica, within the BE-META simulation add bias to the proline ψ dihedral. They could therefore potentially be contributing to the appearance of cis proline in the simulations without a proline replica.

In order to determine if the bias on the ψ dihedral angle of the proline was contributing to cis/trans isomerisation, the BE-META simulation on c(PWRNKY) was repeated after the removal of the bias. The replica with bias on the ζ improper dihedral angle of proline was also not included. Cis/trans isomerisation of the proline was still seen to occur, although much later than seen in the simulations with bias on the ψ of the proline (Figure 22). Convergence was not reached even after 300 ns. Due to the relatively strained nature of small cyclic peptides coupled changes in the dihedral angles are often seen. This is the basis behind the use of 2D bias (two CVs are biased per replica) for fully exploring the energy landscape of cyclic peptides.[22] This means that despite the lack of bias on the ψ dihedral angle, the bias on the ϕ dihedral angle of the proline replica is associated with changes in the ψ dihedral angle. The ϕ dihedral of the i+1 residue is also coupled to the changes of the proline ψ dihedral angle. The ψ dihedral angle is in turn associated with changes in the ω dihedral angle of proline so could potentially be leading to cis/trans isomerisation. This is consistent with no cis proline being seen in extended simulations of small cyclic peptides in the absence of biased CVs. For larger cyclic peptides with ten or more amino acids less of a coupling effect is seen,[254] therefore the proline replica is likely to be needed for cis proline to occur.

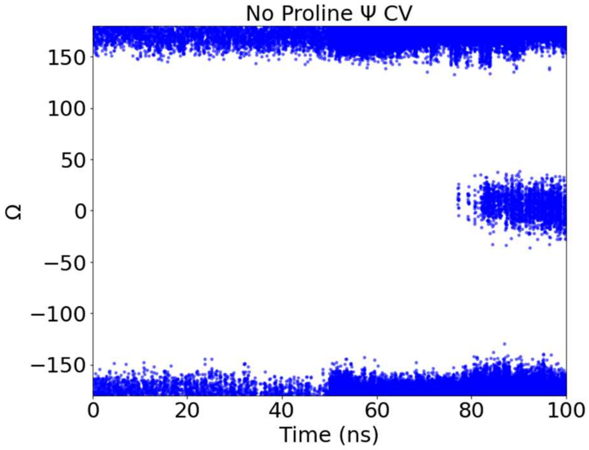


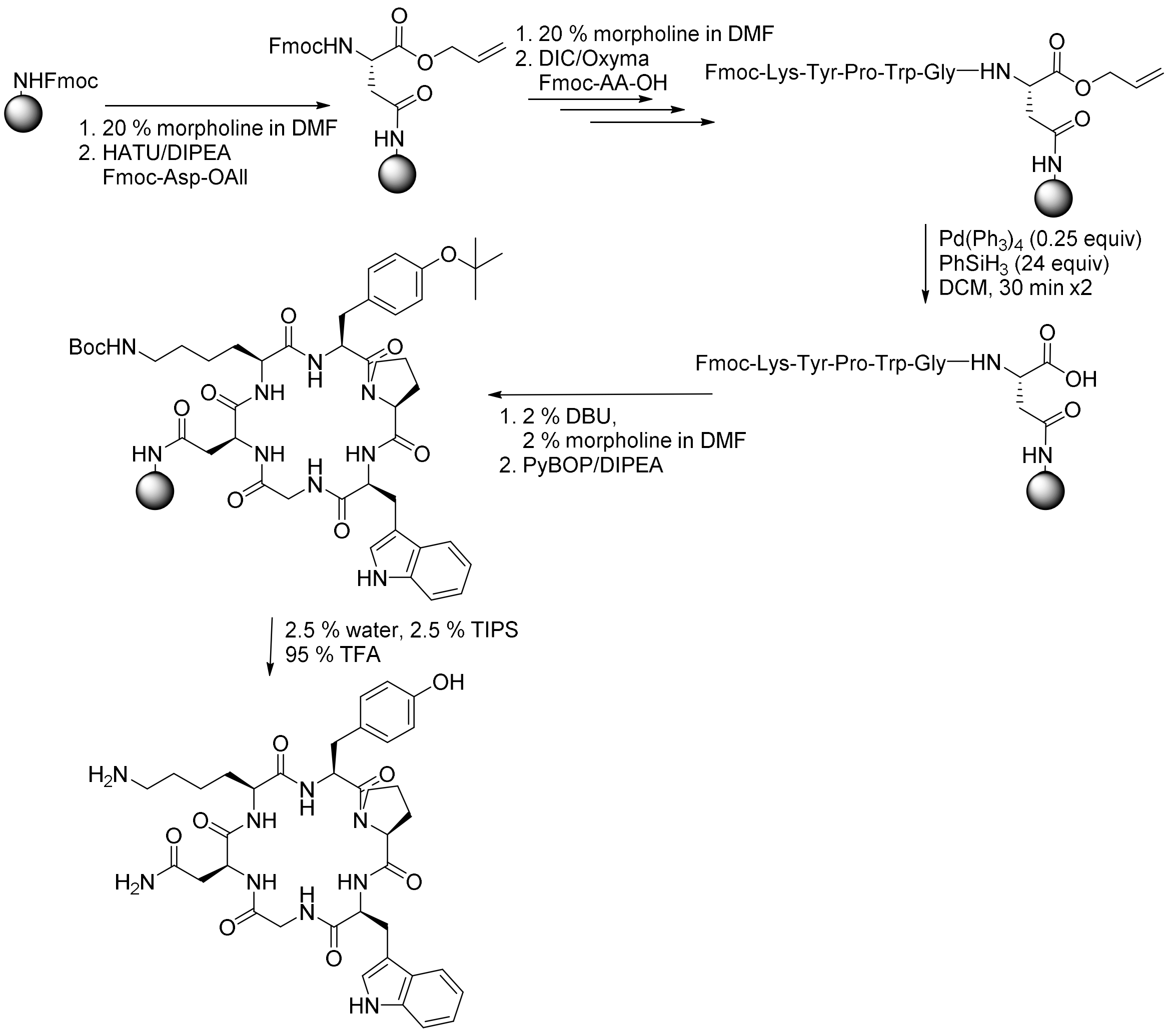
Figure 22: Proline ω changes with no bias on the ψ of proline in the BE-META of c(PWRNKY).

To summarise four cyclic peptides were designed based on the overlay of two β-turns. The overlays were not predictive of the conformations formed by the cyclic hexapeptides. The inclusion of a proline replica allows for faster convergence of simulations to determine the conformation of proline-containing cyclic peptides. If the proline replica is not included cis/trans isomerisation is still seen however due to the small, strained nature of the cyclic peptides allowing for coupled changes in dihedral angles. The four cyclic peptides were next synthesised to determine if the BE-META prediction accurately reflects the conformations the cyclic peptides will adopt in solution.

2.2.3 NMR

The inclusion of a proline replica in the BE-META of proline-containing cyclic peptides can help convergence be reached. However the RSFF2 forcefield which has previously been shown to give the most accurate predictions of cyclic peptide structure has not been tested on its ability to predict the presence of cis proline. The c(PWGNKY), c(PWRNKY), c(PWGNKE) and c(PWRNKE) peptides were therefore synthesised and their structure determined by NMR to compare with the BE-META predictions.

The peptides were synthesised using a triply orthogonal protecting group strategy to allow for on-resin cyclisation (Scheme 13). Fmoc-Asp-OAll was coupled to Rink Amide AM resin. The allyl protecting group is stable under basic and acidic conditions so remained in place during SPPS used to couple the remaining residues. The C-terminal allyl protecting group was then removed using a palladium catalyst and phenyl silane. To remove the final Fmoc group rather than 20% morpholine in DMF it was found 2% DBU and 2% morpholine in DMF gave a better yield. DBU is more reactive than morpholine and has previously been found to be useful where formation of structures such as β-turns and sterics result in more difficult Fmoc removal.[280] The cyclisation could then be carried out on-resin using PyBOP. After cleavage from the resin and purification by HPLC the NMR of each peptide was obtained to assess the accuracy of the BE-META to predict their conformations in solution.



Scheme 13: Synthesis of c(PWGNKY).

NMR was carried out in 5% D_2O in H_2O at 278 K with a peptide concentration of 1 mM in potassium phosphate buffer at pH 7.4. COSY, NOESY, TOCSY as well as HSQC NMR spectra were obtained. A problem with NMR of small peptides is that due to the size of the molecules the tumbling speed is close to the region where the NOEs approach zero. This means limited information could be obtained from the NMR due to the limited NOE cross-peaks. One way of changing the tumbling speed of the peptide is to decrease or increase the temperature. The NMR was carried out at 278 K to reduce the exchange of amide protons with the solvent. However cross-peaks for many amide protons were still missing or not possible to assign for certain sequences. As small peptides have a low core-to-surface ratio faster proton exchange generally occurs as the protons are more solvent exposed. Increasing the temperature to alter the tumbling speed is therefore not a viable option and lowering the temperature further may cause the solvent to freeze. Using a lower pH may help reduce amide proton exchange but may alter the conformation seen so may not be an accurate test of the BE-META simulations. ROESY is another NMR technique that can be used to see through-space couplings, however unlike NOESY, ROESY does not switch sign with rotational correlation time. Few additional cross-peaks were seen in the ROESY experiments on the peptides. The lack of NOE or ROE cross-peaks and peaks being masked meant more detailed assignment of what turn types were likely occurring in the peptides remains difficult (Figure 23). NMR could be repeated on higher concentrations of the peptide to see if more information could be obtained.

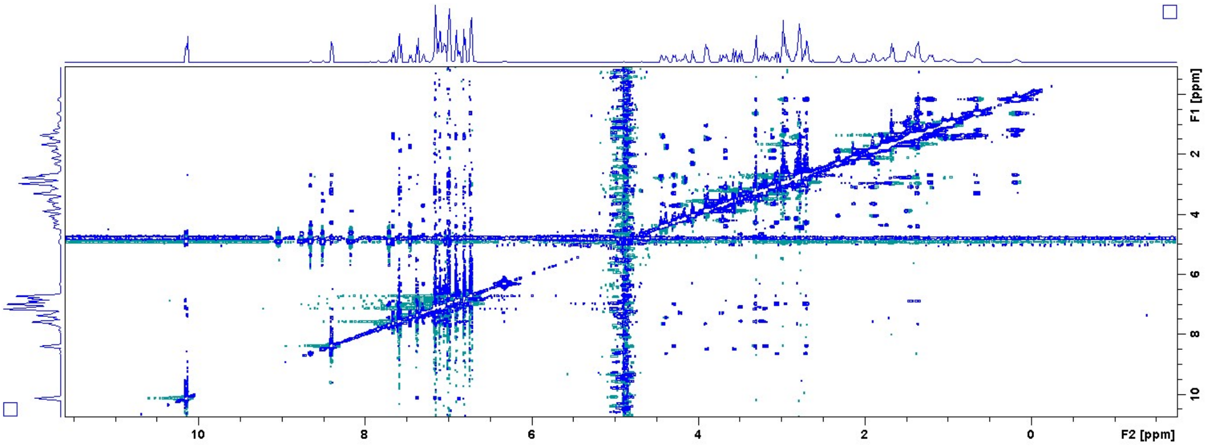


Figure 23: NOESY spectrum of c(PWGNKY). Water signals and missing NOE cross-peaks make assignment difficult.

Although it is difficult to determine the exact conformations formed by the cyclic peptides using the NMR, it was possible to distinguish between conformations containing either cis or trans proline based on NOESY cross-peaks. For trans proline a NOESY cross-peak is seen between the α-carbon of the preceding residue and the δ-carbons of the proline. For cis proline a NOESY cross-peak is seen between the α-carbon of the preceding residue and the α-carbon of the proline (Figure 24). The cis to trans ratio of each peptide was determined and compared to the predicted ratio from the BE-META simulations. Therefore despite the more limited information about conformation obtained from the NMR it is possible to determine if the BE-META can be used to accurately predict the cis to trans ratio.

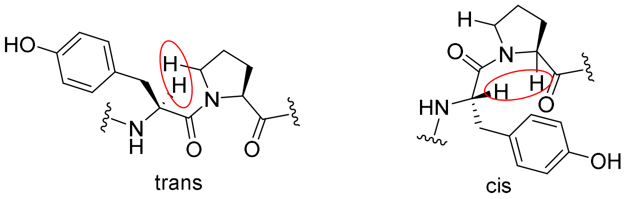


Figure 24: For trans proline a NOESY cross-peak is seen between the α-carbon of the preceding residue and the δ-carbon of the proline. For cis proline a NOESY cross-peak is seen between the α-carbon of the preceding residue and the α-carbon of the proline.

Table 7 shows the proline cis to trans ratio predicted by the BE-META and the observed ratio seen in the NMR. The NMR assignment of the peptides and a comparison of the BE-META and NMR results is discussed below.

Sequence	BE-META Predicted cis to trans ratio	BE-META ΔG (kJ/mol)	NMR cis to trans ratio	NMR ΔG (kJ/mol)
c(PWGNKY)	1:0.9	0.3	1:0.7	0.8
c(PWGNKE)	1:1	0	1:3	-3
c(PWRNKE)	1:6	-4	1:0.4	2
c(PWRNKY)	1:0	-	1:0	-

Table 7: Predicted cis to trans proline ratios compared to the ratios seen in the NMR.

2.2.3.1 c(PWGNKY)

Four clusters are seen in the BE-META of c(PWGNKY). Two contain trans proline and the remaining two contain cis proline with the two cis proline-containing clusters making up 54% of the clustered data. If the BE-META can be used to accurately predict the proportion of cis proline then two conformations will be seen in the NMR in roughly equal proportions. One will contain a cis proline and the other trans proline as cis/trans proline isomerisation is relatively slow on the NMR timescale. Clusters with low populations potentially would not show up in the NMR but for c(PWGNKY) the two smaller clusters (each 13%) contain cis and trans proline respectively, so not including them would not significantly alter the predicted cis to trans ratio.

Two conformations are clearly seen in the NMR of c(PWGNKY). One NMR conformation contains a NOESY cross-peak between the α-carbon of the tyrosine and the α-carbon of the proline, indicating a cis proline, whereas the other conformation has a cross-peak between the α-carbon of the tyrosine and the δ-carbon of the proline, indicating a trans proline. There are a few places in the 1D spectra without significant peak overlap that can be clearly assigned. By integrating these peaks the relative populations of the cis and trans conformations were obtained, showing a cis to trans ratio of 1 : 0.7, similar to the predicted cis to trans ratio based on the BE-META of 1 : 0.9.

The cis proline has unusually low chemical shifts. For example the γ protons of proline typically have a chemical shift between 1.8 and 2.2 ppm but a shift of 1.21 ppm is seen in the cis proline-containing conformation of c(PWGNKY). It is possible this is due to aromatic shielding with tryptophan and tyrosine. Such stacking is seen in the BE-META (Figure 25). A similar low chemical shift is also seen in the cis proline conformations of the remaining three peptide sequences. Aromatic and cis proline residues in the PDB are often seen stacked on top of each other due to a CH-π interaction.[278] Such CH-π interactions can result in large upfield or downfield shifts from typical values.

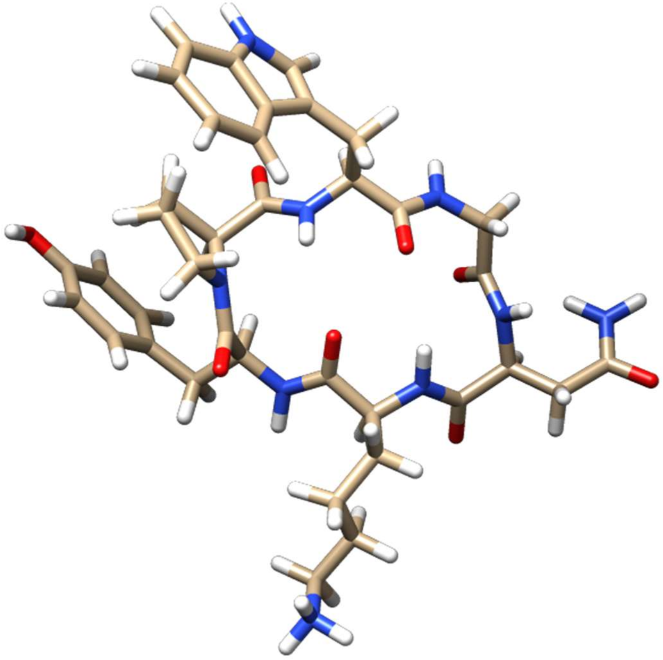


Figure 25: Stacking between tyrosine, cis proline and tryptophan is seen in the BE-META of c(PWGNKY).

The typical amide proton NMR shift in a random coil is around 8.3 ppm. A wide chemical shift dispersion around this value indicates a more structured conformation, whereas a narrower range of amide shifts is more indicative of a more flexible structure as the protons interchange leading to average ppm values. A similar effect may be seen with the sidechain protons. For sidechains with two β-hydrogens, if the sidechain rotates freely, the two β-hydrogens will have the same or similar shifts. Alternatively if the sidechain is in a particular conformation, possibly due to sidechain interactions, then the two hydrogens may show very different shifts. Additionally different strength NOE cross-peaks could be seen between the two β-hydrogens if one is in closer proximity to another hydrogen than the other in a particular conformation.

c(PWGNKY) has a relatively wide range of amide chemical shifts for both the cis and trans proline conformations. Both conformations are therefore likely to be well structured. There is also a large difference in chemical shift for the tyrosine β-hydrogens of 0.36 ppm in the cis proline conformation. One of the β-hydrogens has a very low chemical shift of 2.70 ppm. This would be consistent with aromatic shielding with proline in this conformation with a very well-defined tyrosine sidechain conformation.

2.2.3.2 c(PWGNKE)

In the NMR of c(PWGNKE) two conformations are seen. In the largest conformation there is a NOE cross-peak between the α-carbon of the glutamic acid and the δ-carbon of the proline, indicating a trans proline. The α-carbon of the glutamic acid of the minor conformation is masked by the water signals. It is likely it is a conformation containing a cis proline however it is possible it is another conformation of the peptide which exchanges slowly relative to the NMR time scale. Assuming it is a cis conformation the cis to trans ratio between the two conformations in the NMR is 1:3. The predicted cis to trans ratio from BE-META being 1:0.9.

The trans proline-containing conformation for c(PWGNKE) has a wide range of assigned amide proton shifts. This indicates a relatively stable conformation. The cis proline-containing conformation however has a narrower chemical shift dispersion so is potentially a less rigid conformation and is changing between multiple conformations containing a cis proline. Similar to c(PWGNKY) very low chemical shifts are seen for the proline protons in the cis conformation. The tryptophan and proline residues are seen to interact throughout the BE-META simulation which could potentially be altering the chemical shift. In the cis conformation the two tryptophan β-protons have chemical shifts of of 3.25 and 3.48 ppm. A difference between the two shifts in the cis proline conformation of 0.23 is relatively large which would be consistent with the tryptophan being in a relatively rigid conformation due to stacking with the proline. It is possible the cis proline is part of a type VI turn but the GN turn is more flexible forming multiple turn types as predicted by the BE-META.

2.2.3.3 c(PWRNKE)

Two conformations are seen in the NMR of c(PWRNKE) with one much smaller than the other. The major conformation has a NOE cross-peak between the α-carbon of the glutamic acid and the α-carbon of the proline, indicating a cis proline. The smaller conformation has a NOE cross-peak between the α-carbon of the glutamic acid and the δ-carbon of the proline, indicating a trans proline. The cis to trans ratio seen in the NMR is 1:0.4 which is significantly different from the predicted ratio of 1:6.

For c(PWRNKE) the cis conformation appears to be very well structured based on the wide range of amide chemical shift dispersions from 7.18 to 9.12 ppm. The tryptophan β-hydrogens have shifts of 3.60 and 3.88 ppm whereas only one peak is seen for both hydrogens in the trans conformation at 3.33 ppm. This is consistent with the tryptophan having a relatively fixed position in the cis proline conformation due to stacking with proline but adopts more conformations in the trans proline conformations. The δ-protons of the cis proline in c(PWRNKE) have unusually low shifts (2.8 and 3.1 compared to an average of 3.6 ppm), likely due to its interaction with the adjacent tryptophan, which has a high shifts. This interaction is further supported by a ROE cross-peak between the proline and tryptophan δ-protons. In the BE-META simulations stacking of the proline and tryptophan is seen, so it is likely this stacking is occurring and is possibly contributing to why the major conformation contains a cis proline.

Strong NOE cross-peaks are typically seen between the amide protons of i+2 and i+3 positions of β-turns. An amide-to amide cross-peak is seen between the asparagine and lysine in the conformation containing cis proline indicting a β-turn at the other side of the peptide. An additional strong NOE is also seen between the α-proton of arginine and the amide proton of asparagine. This is consistent with a type II β-turn with the arginine and asparagine at the i+1 and i+2 positions. Other β-turn types either do not have this NOE cross-peak or it is weak due to a greater distance between the two protons (Figure 26). This would be consistent with the BE-META prediction of a VI+II conformation. For the trans conformation the lysine-glutamic acid amide-to-amide NOE is seen showing a β-turn is present but the tryptophan to arginine cross-peak is masked by the water. It may be that BE-META can accurately predict the conformations likely to be seen despite not accurately predicting the cis to trans ratio.

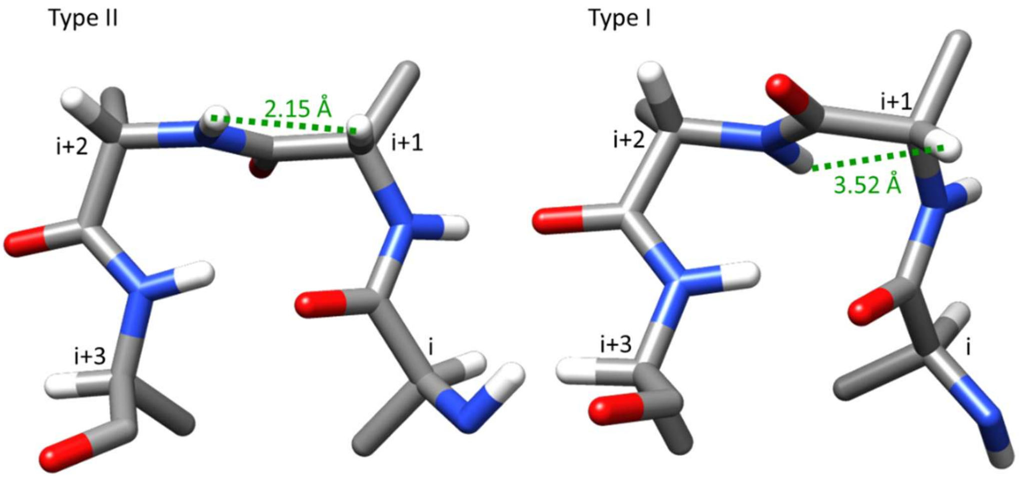


Figure 26: Type II turns have a strong NOE cross-peak between the α-proton of the i+1 position and the amide proton of the i+2 position that is not seen in other β-turn types such as type I where there is a larger distance between the protons.

2.2.3.4 c(PWRNKY)

Only one conformation is seen in the NMR of c(PWRNKY). It has a NOE cross-peak between the α-carbon of the tyrosine and the α-carbon of the proline, indicating a cis proline. The BE-META simulations predicted all conformations of c(PWRNKY) to contain cis proline.

Similar to the other sequences proline has low chemical shifts. Both the tryptophan and tyrosine have distinct β-hydrogen chemical shifts with differences of 0.33 and 0.35 ppm between the two β-hydrogens respectively. This is a large difference in chemical shift for two β-hydrogens so indicates a well-structured cyclic peptide with the sidechains in a relatively rigid conformation. This would be consistent with the tyrosine-proline-tryptophan stacking seen in the BE-META simulations. One major conformation is predicted (66%) with type VI and I’ turns.

2.2.4 Comparison of the BE-META and NMR Conformations

Table 7 shows the predicted cis to trans ratios of proline within the peptides based on the clusters seen in the BE-META compared to the cis to trans ratio seen for the peptides in the NMR. The sequences c(PWGNKY) and c(PWRNKY) predict similar cis to trans ratios as seen in the NMR. However c(PWGNKE) is predicted to have a larger proportion of cis proline than seen in the NMR whereas c(PWRNKE) underpredicts the amount of cis proline that occurs.

The BE-META both with and without the proline replica included predicted the same conformations for each of the sequences. Convergence was also reached in each case. This means it is unlikely the differences in the predicted and observed cis to trans ratio are due to the proline replica or it is very unlikely the same proportions of each conformation would occur across the multiple simulations. The discrepancy between the predicted and observed cis to trans ratio for c(PWGNKE) and c(PWRNKE) is therefore likely caused by the RSFF2 forcefield parameters not accurately modelling the energy differences between the cis and trans proline states. RSFF2 has previously been shown to be the most accurate forcefield for modelling small cyclic peptides so it may be that it is able to accurately predict the conformation with either cis or trans proline, however it may not be able to accurately predict the energy difference between the two states leading to incorrect predictions for the cis to trans ratio. So for example the NMR of c(PWRNKE) showed the cis proline-containing conformation likely had a conformation equivalent to overlapping type VI and type II β-turns. This is consistent with the minor conformation seen in the BE-META. So although predicted to occur in smaller quantities than seen in the NMR, the correct conformation was predicted.

The distinct conformations predicted by the BE-META have very small energy differences between them. They therefore require an accurate method of determining free energy differences. Due to the very small size of the peptides the accuracy of the forcefield becomes very important for accurately predicting the conformation and therefore the free energy between states, with even small inaccuracies in the potential energy function of the forcefield producing noticeable differences in conformation.[281] The differences between the predicted and observed proline cis to trans ratio correspond to only small differences in free energy, with the largest energy difference between the cis and trans states from the BE-META and NMR of 6 kJ/mol. Even small errors in the forcefields ability to accurately model the peptide could lead to the differences seen.

There are many potential features of the forcefield which could lead to the energy differences between the cis and trans proline states not being accurately modelled for the cyclic peptide system with terms for bond stretching, bond angle flexibility, torsions, van der Waals forces and electrostatic interaction. Forcefield parameters are generally based on quantum calculations on model systems and optimized for predicting protein structure. They also may have parameters fit to those seen in large protein databases. Torsional parameters for the amino acid sidechains may not accurately reflect what is seen in the small peptides with an absence of secondary structure. For example it has been shown that in general forcefields oversample the α-helical region of Ramachandran space.[282]

RSFF2 is based on the Amber99SB[283] forcefield with improved torsional parameters from a coil database.[242] Cis proline appears infrequently in proteins, with trans proline appearing approximately 30-45 times more.[284, 285] The torsional parameters for cis proline may therefore not be accurate as there is limited experimental data.[286] Compared to the Amber forcefield the torsional parameters of the ω dihedral angle of the proline are not updated in RSFF2. Doshi *et al*. modified the torsional parameters for the ω dihedral angle of proline in the Amber99SB forcefield. [267] They used the new parameters in accelerated MD on the model systems Ala-Pro and Phe-Pro. A better match between experimental values for the equilibrium cis to trans ratio and the energy barrier for cis to trans isomerisation was obtained. Adding the new torsional parameters into RSFF2 may therefore lead to improved predictions for the cis to trans ratio of cyclic peptides. However other factors can also alter the cis to trans ratio at equilibrium so new forcefield parameters are possibly also needed. Wu *et al*. found the flexibility of the peptide bond allowed throughout the simulation could greatly alter the predicted cis to trans ratio of the peptide c(Phe-Phe-Aib-Leu-Pro).[259] Rigid bond angles were used throughout the simulation and the predicted amount of trans proline changed from 42 to 14% when the ω dihedral angle of the proline was changed from 6 to 0°.

Polarisation parameters are another common weakness seen in current forcefields.[287] As aromatic stacking is seen between the proline and aromatic sidechains in the four peptides tested an inaccurate description of these interactions could be leading to the differences between the computational and NMR values. The RSFF2 forcefield used (like most commonly used MD forcefields) uses one point-charges for each atom which fails to capture the anisotropic nature of electronic features required for an accurate description of π-systems.[288]

Other forcefields could be tested to see if they are better at modelling the cis/trans systems in cyclic peptides. However as RSFF2 has previously been shown to lead to the most accurate predictions of cyclic peptide structure,[70, 289] other forcefields would likely not accurately predict the conformation even if they led to improved modelling of cis/trans isomerisation. New forcefields with improved parameters are therefore required to accurately predict the relative energies of cis and trans proline-containing cyclic peptides.

Rather than the discrepancy in NMR and BE-META cis to trans ratio being due to the properties of the forcefield not accurately modelling the peptide, it could potentially be caused by the peptides interacting with one another. If the peptides interact in solution this could potentially alter the cis to trans ratio if one conformation allows for stronger interactions with the surrounding peptides. This effect is obviously not encapsulated by the BE-META where only a single peptide is simulated. The two peptides with significantly different predicted cis to trans ratios than seen in the NMR are the two that contain glutamic acid. As the peptides also contain lysine its theoretically possible salt bridges between the peptides could be forming. The peptides could also be interacting due to the hydrophobic effect, where stacking of the tryptophan and proline residues would be likely to occur. Wu et. al. used REMD to model tentoxin, a small cyclic tetrapeptide which although it doesn't contain proline contains two *N*-methylated residues which adopt cis amide bonds.[259] Tentoxin begins to aggregate above 35 µM in water. Therefore the prediction was found to not match the major NMR structures as the NMR was carried out at with a protein concentration of 250 µM above when aggregates form. If aggregates form shielding effects often cause changes in chemical shifts. Different chemical shifts would also be seen and if the NMR was rerun at lower concentrations where aggregates didn't form. Adding organic solvents has also been shown to disrupt aggregation but may cause other alterations in the peptide conformation.[290-292]

2.3 Conclusions

Four peptides (c(PWGNKY), c(PWRNKY), c(PWGNKE) and c(PWRNKE)) were simulated by BE-META and had their structures determined by NMR to compare the predicted and observed proline cis to trans ratios. Inclusion of a replica with bias on the improper dihedral angle ζ allowed for proline cis/trans isomerisation in the BE-META of cyclic hexapeptides. Generally the same structures were seen as when the additional proline replica was not included, but convergence was reached faster meaning less computational time was required. In some instances, without the inclusion of the additional replica convergence was not reached at all. As proline is often a beneficial amino acid to include in cyclic peptides, seen to aid both synthesis and reduce the conformational flexibility of cyclic peptides,[46] being able to model proline-containing cyclic peptides is important. As the additional proline replica allows for faster convergence it is beneficial to include within the BE-META.

Although two of the tested sequences gave similar cis to trans ratios in the BE-META and NMR, two sequences did not. As similar results are obtained in the BE-META with and without the proline replica it is unlikely the proline replica is leading to the peptide becoming stuck in an unrealistic energy minimum leading to the differences seen in the BE-META and NMR. The difference may therefore be due to other reasons such as the forcefields ability to accurately model the energy difference between the cis and trans conformations. The greatest difference in predicted and observed cis to trans ratio only represents a free energy difference of 6 kJ/mol between the cis and trans proline states. This is a relatively small difference so highly accurate forcefields are required to accurately model such small energy differences.

3 Database Analysis

Known structures provide a good starting point for the design of cyclic peptides with biological function. Loop regions involved in protein-protein interactions (PPIs) are increasingly being incorporated into peptide macrocycles for development of potential drug molecules.[80-82, 293-297] Many peptide sequences shown to be important for PPIs have been shown to contain a β-turn.[112] Cyclic hexapeptides often form a structure that can be thought of as two overlapping β-turns. Design of cyclic hexapeptides with a specific structure often remains difficult however, with many different conformations often forming in solution and often only small changes in peptide sequence leading to very different conformations. By analysing β-turns from known structures stored in a database, insights into which β-turns may be incorporated into cyclic peptides to give more stable structures may be obtained. For example inclusion of proline and other residues has led to reduced numbers of conformations.[46, 158, 199] If certain sequences are found in the database to be particularly favourable at forming β-turns then they may also form β-turns in cyclic peptides leading to more stable structures. It may also be possible to determine sequence-structure relationships of different β-turn types to help determine the conformations a cyclic peptide is likely to adopt.

3.1 Loop Database

A database was generated by Dr Drew Thomson which contains high-resolution, non-redundant protein crystal structures (see section 10.1). Loop regions of the protein structure were identified that contain no persistent secondary structure but are flanked on each side by secondary structures for inclusion in the database. It was hypothesised that the conformation of a loop should be dominated by local interactions rather than long-range non-local interactions that can affect structure in the rest of the protein. The Loop Database was then used to obtain data on β-turns which might be helpful in the design of cyclic hexapeptides.

3.2 β-Hairpins

Cyclic hexapeptides are a relatively constrained system and frequently contain two intrapeptide hydrogen bonds.[245] β-turns found in β-hairpins (β-turn joining two antiparallel β-strands) are also in a relatively constrained system with multiple hydrogen bonds seen. Therefore β-hairpins were initially searched for within the Loop Database (see section 10.1.1). Specifically β-hairpins containing a β-turn with a hydrogen bond between the i and i+3 positions and at least three additional hydrogen bonds following the one that makes up part of the turn (Figure 27) were searched for. It is possible that the tight turn in a β-hairpin is a good model for the β-turns found in cyclic peptides due to the additional geometrical constraints from the hydrogen-bonded β-sheet structure.

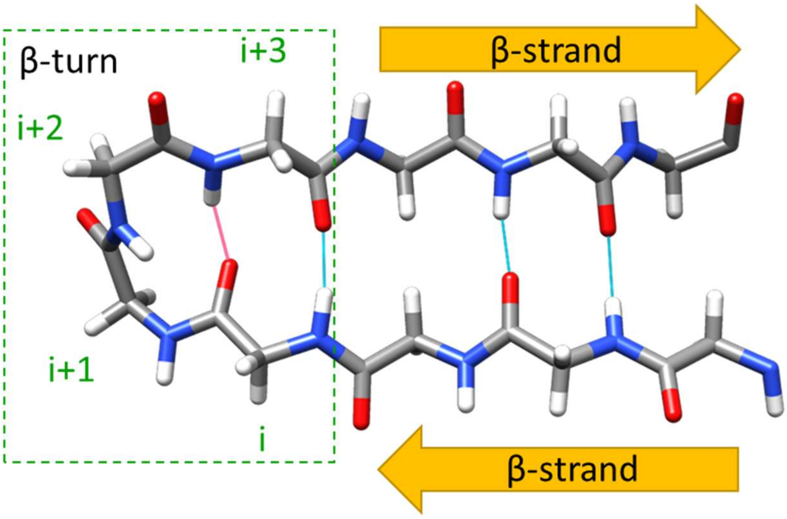


Figure 27: β-hairpin with the β-turn joining two antiparallel β-strands. The i to i+3 positions of the β-turn are labelled. β-turns were extracted from the database which have three additional hydrogen bonds (shown in light blue) besides the one between the i and i+3 positions of the β-turn (pink).

Previously there has been analysis of β-turns in proteins to determine sequence propensities to help with structure prediction and validation.[127, 131, 132, 146] Although it has been noticed the different prevalent β-turn types within β-hairpins compared to other β-turns within the protein environment, there has been more limited analysis on the amino acid composition found within the hairpin environment and how this compares to other β-turns.

10,923 β-hairpins were found in the Loop Database using the specified criteria. The amino acid frequency at each position of the β-turn shows an abundance of glycine in β-hairpins especially at the i+1 and i+2 positions (Figure 28). Asparagine and aspartic acid also appear frequently at the i+1 and i+2 positions. Proline has a relatively low abundance in the turn region despite being frequently seen to stabilise turn regions.[298-300]

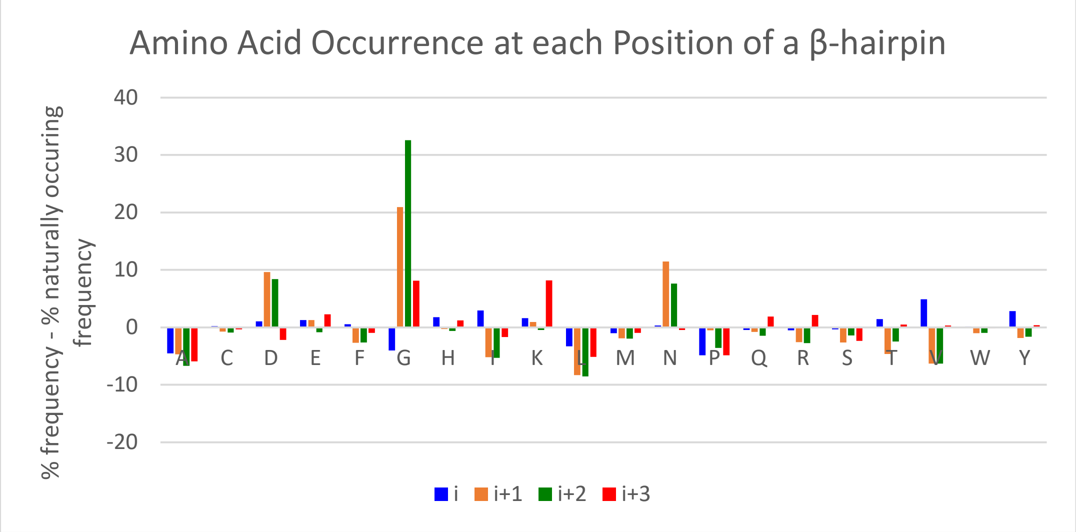
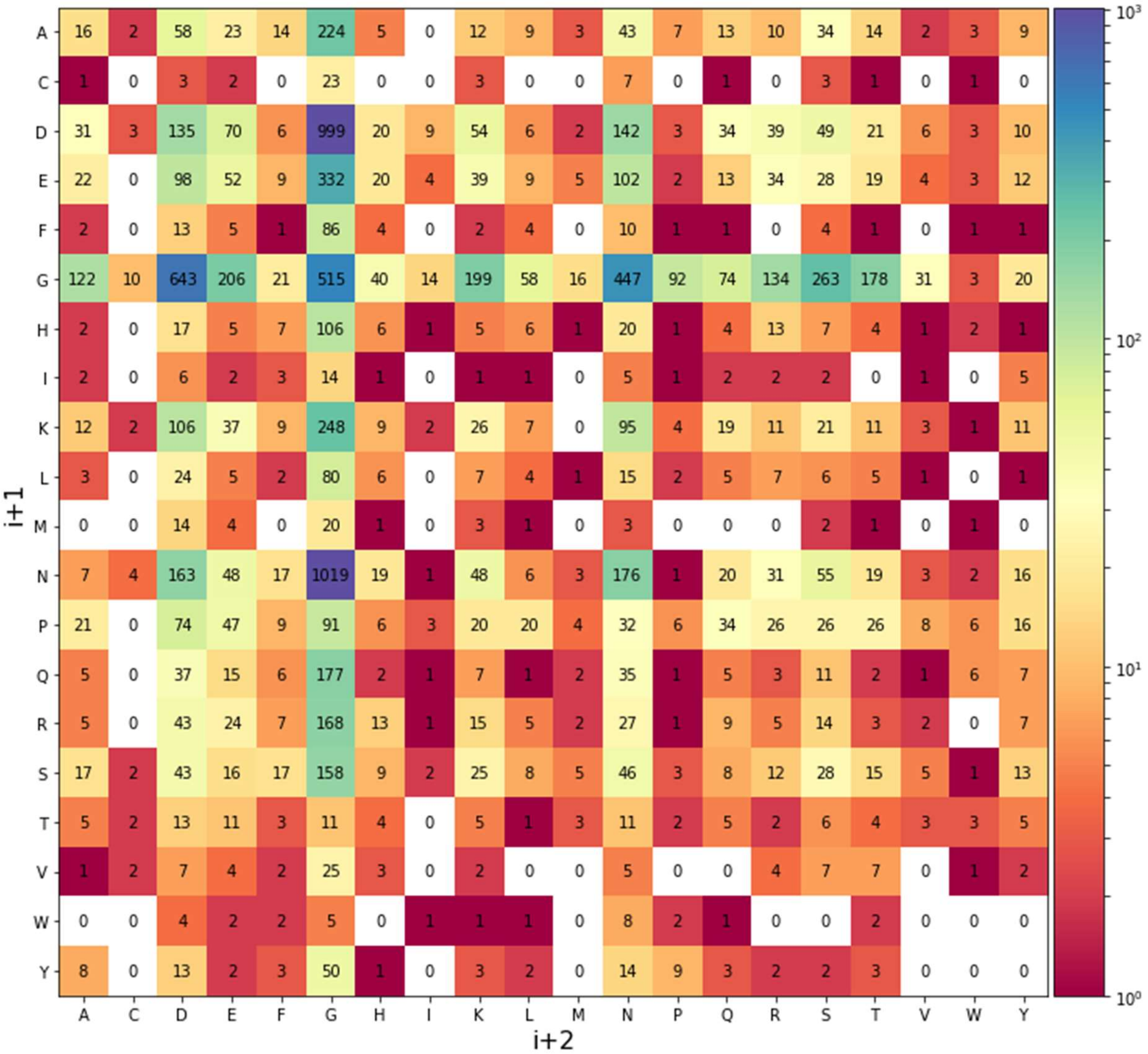


Figure 28: The occurrence of amino acids at each position of the β-turn in the hairpin dataset, corrected for the naturally occurring frequency of each amino acid. Naturally occurring frequency of amino acids based on all available structures in the PDB taken from [301].

Taking the i+1 and i+2 positions which are used to define the β-turn, there are 400 possible sequence combinations using the 20 naturally occurring amino acids. A search for all the sequences was carried out. NG and DG are the most commonly occurring sequences at the i+1 and i+2 positions (Figure 29). GG, GD and GN are also common, appearing over 400 times. When designing cyclic peptides inclusion of these subsequences may help enforce a specific amino acid register if they are more likely to occupy the i+1 and i+2 positions of a β-turn.



	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
A	16	2	58	23	14	224	5	0	12	9	3	43	7	13	10	34	14	2	3	9
C	1	0	3	2	0	23	0	0	3	0	0	7	0	1	0	3	1	0	1	0
D	31	3	135	70	6	999	20	9	54	6	2	142	3	34	39	49	21	6	3	10
E	22	0	98	52	9	332	20	4	39	9	5	102	2	13	34	28	19	4	3	12
F	2	0	13	5	1	86	4	0	2	4	0	10	1	1	0	4	1	0	1	1
G	122	10	643	206	21	515	40	14	199	58	16	447	92	74	134	263	178	31	3	20
H	2	0	17	5	7	106	6	1	5	6	1	20	1	4	13	7	4	1	2	1
I	2	0	6	2	3	14	1	0	1	1	0	5	1	2	2	2	0	1	0	5
K	12	2	106	37	9	248	9	2	26	7	0	95	4	19	11	21	11	3	1	11
L	3	0	24	5	2	80	6	0	7	4	1	15	2	5	7	6	5	1	0	1
M	0	0	14	4	0	20	1	0	3	1	0	3	0	0	0	2	1	0	1	0
N	7	4	163	48	17	1019	19	1	48	6	3	176	1	20	31	55	19	3	2	16
P	21	0	74	47	9	91	6	3	20	20	4	32	6	34	26	26	26	8	6	16
Q	5	0	37	15	6	177	2	1	7	1	2	35	1	5	3	11	2	1	6	7
R	5	0	43	24	7	168	13	1	15	5	2	27	1	9	5	14	3	2	0	7
S	17	2	43	16	17	158	9	2	25	8	5	46	3	8	12	28	15	5	1	13
T	5	2	13	11	3	11	4	0	5	1	3	11	2	5	2	6	4	3	3	5
V	1	2	7	4	2	25	3	0	2	0	0	5	0	0	4	7	7	0	1	2
W	0	0	4	2	2	5	0	1	1	1	0	8	2	1	0	0	2	0	0	0
Y	8	0	13	2	3	50	1	0	3	2	0	14	9	3	2	2	3	0	0	0

Figure 29: The sequence occurrence at the i+1 and i+2 positions of β*-hairpins in the Loop Database.*

The most commonly occurring β-hairpin sequences appeared in the database between 23 and 53 times (Figure 30). All of the most common sequences have glycine at the i+2 position and either N or D at the i+1 position. Lysine also appears frequently at the i+3 position and valine at the i position. The hydrophobic amino acids valine, isoleucine and leucine have previously been observed to have a high relative abundance in β-strands.[302]

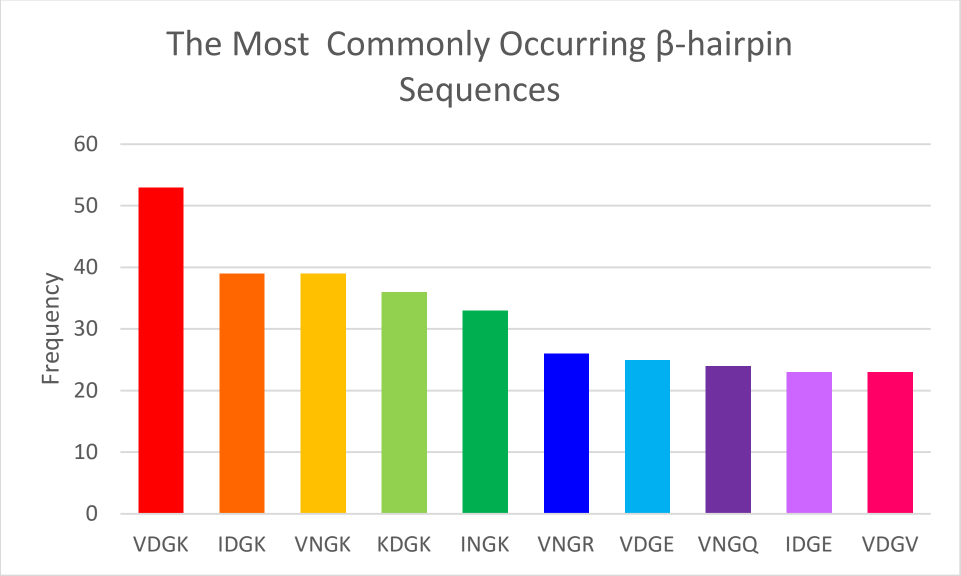
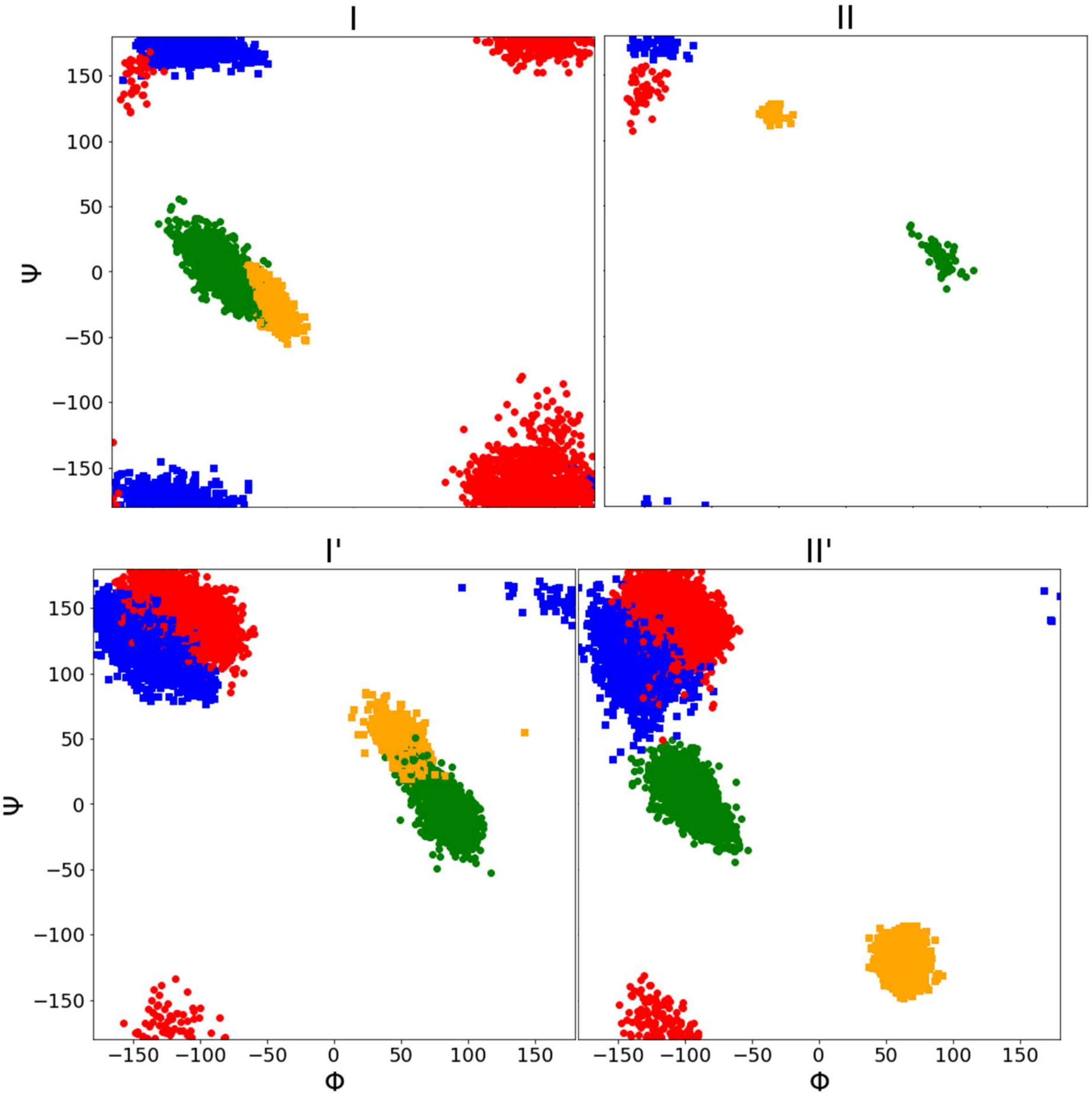


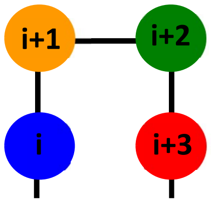
Figure 30: The most commonly occurring sequences that make up β-hairpins in the database.

3.2.1 Clustering of the β-hairpin Data

The β-turns in the hairpin dataset form a variety of different β-turn types which are classified based on the i+1 and i+2 ϕ and ψ dihedral angles. A clustering algorithm can be used to find the different β-turn types within the dataset. Different amino acid sequences may be seen for the different clusters and therefore potentially be used to select sequences that form certain turn types.

A density based clustering algorithm[303] was used to cluster the β-turns in the hairpin dataset based on the ϕ and ψ dihedral angles of the i to i+3 positions. Prior to clustering the dihedral angles were processed by a sin and cos transformation followed by principal component analysis (PCA).[304] The sin and cos transformation converts the dihedral angles, which are circular variables, into linear metric coordinate space which is necessary for the PCA algorithm. The PCA was used to reduce the number of dimensions of the data to three whilst retaining as much variation in the data as possible. The results of the PCA were used in clustering as the time taken for clustering increases greatly with number of dimensions. Four clusters were identified (Figure 31). All clusters could be assigned to β-turn types based on the Ramachandran angles of the i+1 and i+2 residues. The major cluster representing 50% of the turns is a type I’ β turn. Type II’ turns are the second most common (36%), followed by type I with very few type II turns seen. These are very different frequencies compared to those generally expected for β-turns, where the types II’ and I’ turns are much more uncommon than types I and II.[123] This difference in the prevalence of different β-turn types has previously been observed in β-hairpins.[127, 305] Type I and II turns have a left-handed twist, whereas the I’ and II’ turns have a righthanded twist which is more compatible with the right-handed twist seen in β-hairpins.[167, 306] The high occurrence of the relatively flexible glycine, asparagine and aspartic acid at the i+1 and i+2 positions contributes to this reversal in most common β-turn types as they are more able to access the necessary Ramachandran regions for the type I’ and II’ turns.





Type	Population	% of Clustered	Non-Hairpin Frequency for Turn Type (%)
I	1,378	14.07	38.21
II	52	0.53	11.81
I’	4,880	49.82	4.10
II’	3,486	35.56	2.54

Figure 31: The four clusters obtained from the β-hairpins. The colours represent the I to i+3 positions of a β-turn as shown in the diagram. Frequency for β-turn types not in a β-hairpin taken from [123]*.*

Of the eight generally accepted β-turn types and the ninth miscellaneous type IV category, only the I, II, I’ and II’ types are observed. The remaining turn types may be present but in too low numbers to be picked up by the clustering algorithm using this dataset. Searching within the unclustered data for β-hairpins with a cis proline at the i+2 position finds type VI turns, although all are type VI_{a1} with no VI_b or VI_{a2} present.

The amino acids that occur at each position of each β-turn type corrected by size of cluster and naturally occurring frequency of the amino acids are shown in Figure 32. The sequence propensities begin to show the differences between the clusters that leads to the different turn types. Type I β-turns have an abundance of threonine at the i position, proline at the i+1 position and glycine at the i+3 position. Only 40 of the 1,378 type I β-turns (2.90%) have an amino acid other than glycine at the i+3 position. The i+3 position of the type I cluster occupies the Ramachandran region in the top righthand corner of the plot in Figure 31, an area not frequently occupied by chiral amino acids. When designing cyclic hexapeptides with a type I β-turn it may be that including glycine at the i+3 position would lead to a more stable structure. Type II β-hairpins predominately have the PG sequence at the i+1 and i+2 positions but as there are only 52 type II turns there is limited data. Type I' β-turns frequently have asparagine or aspartic acid at the i+1 position and glycine at the i+2 position (NG or DG), and to a lesser extent lysine at the i+3 position. Type I' turns make up approximately 50% of the dataset, so this is consistent with the NG and DG turns being most common. Type II' turns predominantly have glycine at the i+1 position and either asparagine or aspartic acid at the i+2 position (GN or GD).

Analysis of the β-hairpins finds an abundance of type I' and II' turns compared to type I and II turns. β-hairpins form a tight β-turn so potentially would make a good model for cyclic peptides. If this is the case, then a higher proportion of type I' and II' turns might be expected to be seen in cyclic peptides than otherwise expected. The relatively flexible glycine, asparagine and aspartic acid are commonly found at the i+1 and i+2 positions of the β-turn so could potentially be incorporated into those positions in a cyclic hexapeptide. The different turn types have different sequence profiles which could allow for the prediction of which turn types are likely to form in a cyclic peptide conformation.

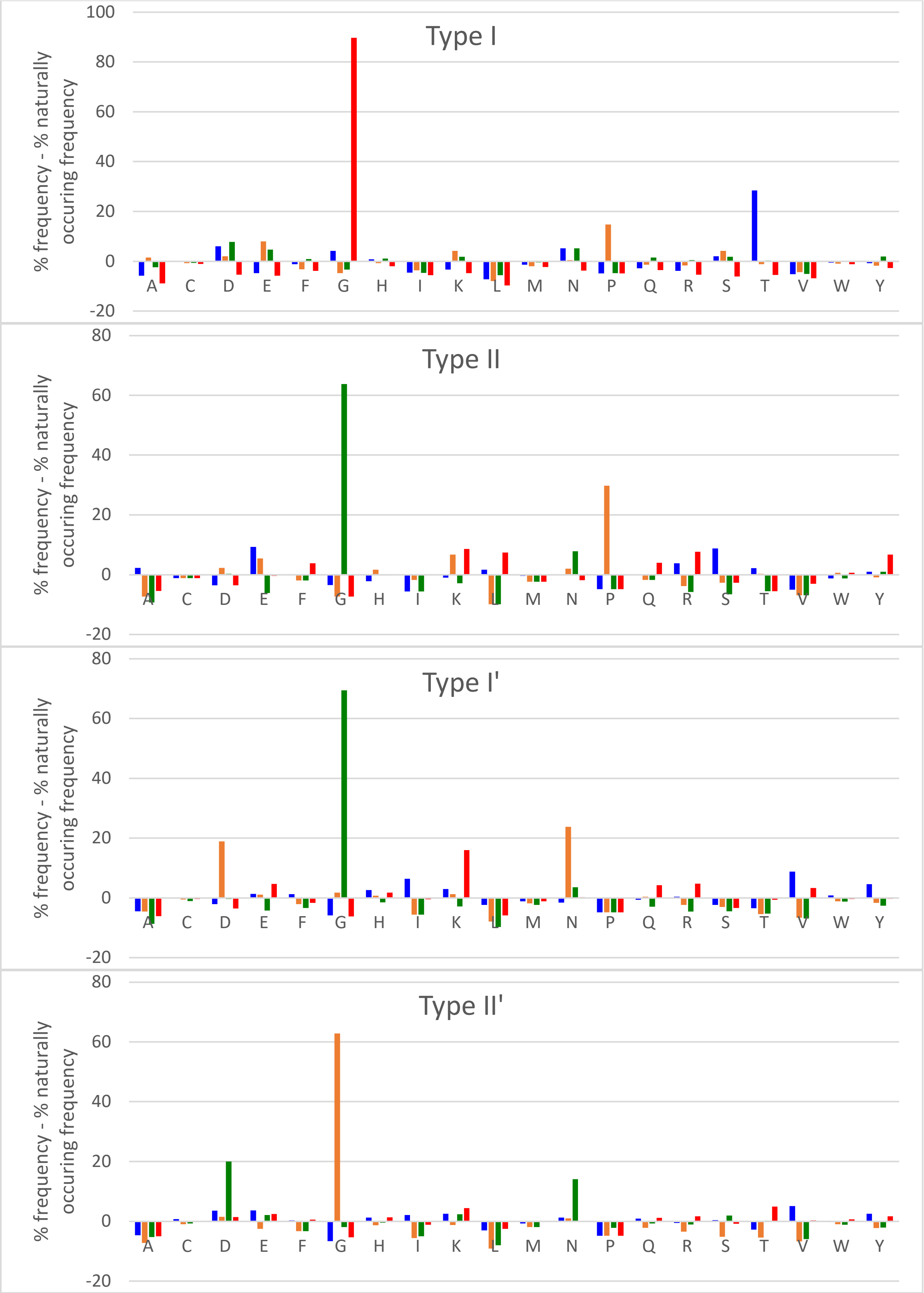


Figure 32: The amino acid composition of the clusters of the β-hairpins. Blue is the i position of the β-turn, orange the i+1, green i+2 and red i+3. Naturally occurring frequency of amino acids based on all available structures in the PDB taken from [301].

3.2.2 7 Å β-turn Definition β-Hairpins

Initially β-hairpins were found which had at least three additional hydrogen-bonds besides the one forming part of the β-turn. Few examples of type II turns were seen. The hairpin dataset was therefore expanded. The β-hairpins looked at so far all have a hydrogen bond between the i and i+3 positions of a β-turn. An alternative definition of a β-turn requires a distance of less than 7 Å between the α-carbons of the i and i+3 positions rather than the hydrogen bond. When this β-turn definition was used in combination with a distance less than 7 Å between the α-carbons of 3 subsequent amino acid pairs (Figure 33), 15,121 β-turns are found. Of these turns 4,226 are not in the original β-hairpin dataset.

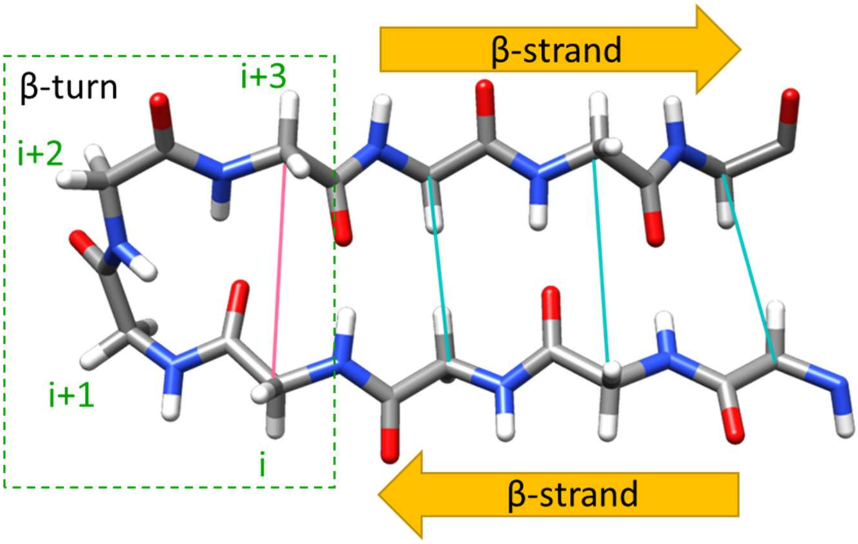


Figure 33: β-hairpins were identified with less than 7 Å distance between the α-carbons of the i and i+3 positions of the β-turn (pink) in addition to a distance of less than 7 Å between the three subsequent amino acid α-carbons (light blue). The β-turn i to i+3 positions are labelled in green.

When the additional turns are clustered, four clusters representing type I, II, I' and II' turns are still found but the percentage represented by each cluster differs slightly. There are relatively fewer type II' and more type I (Table 8). The definition used to define a β-turn not only changes the number of β-turns found but the proportion of β-turn types. The assignment criteria has previously been shown to significantly alter the occurrence of β-turns identified within a database.[132, 307]

	% β-turn type		
	Hydrogen-bonded β-hairpins	7 Å distance β-hairpins which do not have a hydrogen bond (additional β-hairpins not found in the hydrogen-bonded dataset)	All β-hairpins
I	14.07	28.54	18.20
II	0.53	10.61	3.41
I’	49.82	45.09	48.47
II’	35.56	15.75	29.93

Table 8: Percentage of clustered hairpin datasets belonging to each turn type category.

The sequence profiles seen for the 4,226 β-hairpins without a hydrogen bond between the i and i+3 positions are very similar to those with the hydrogen-bond (Figure 34). There are slight differences seen for the type I turns which no longer have such a high prevalence of threonine at the i position and glycine at the i+3 position. This is reflected in the Ramachandran plot of the i+3 position which now has a higher density in the top left rather than the top right region previously seen.

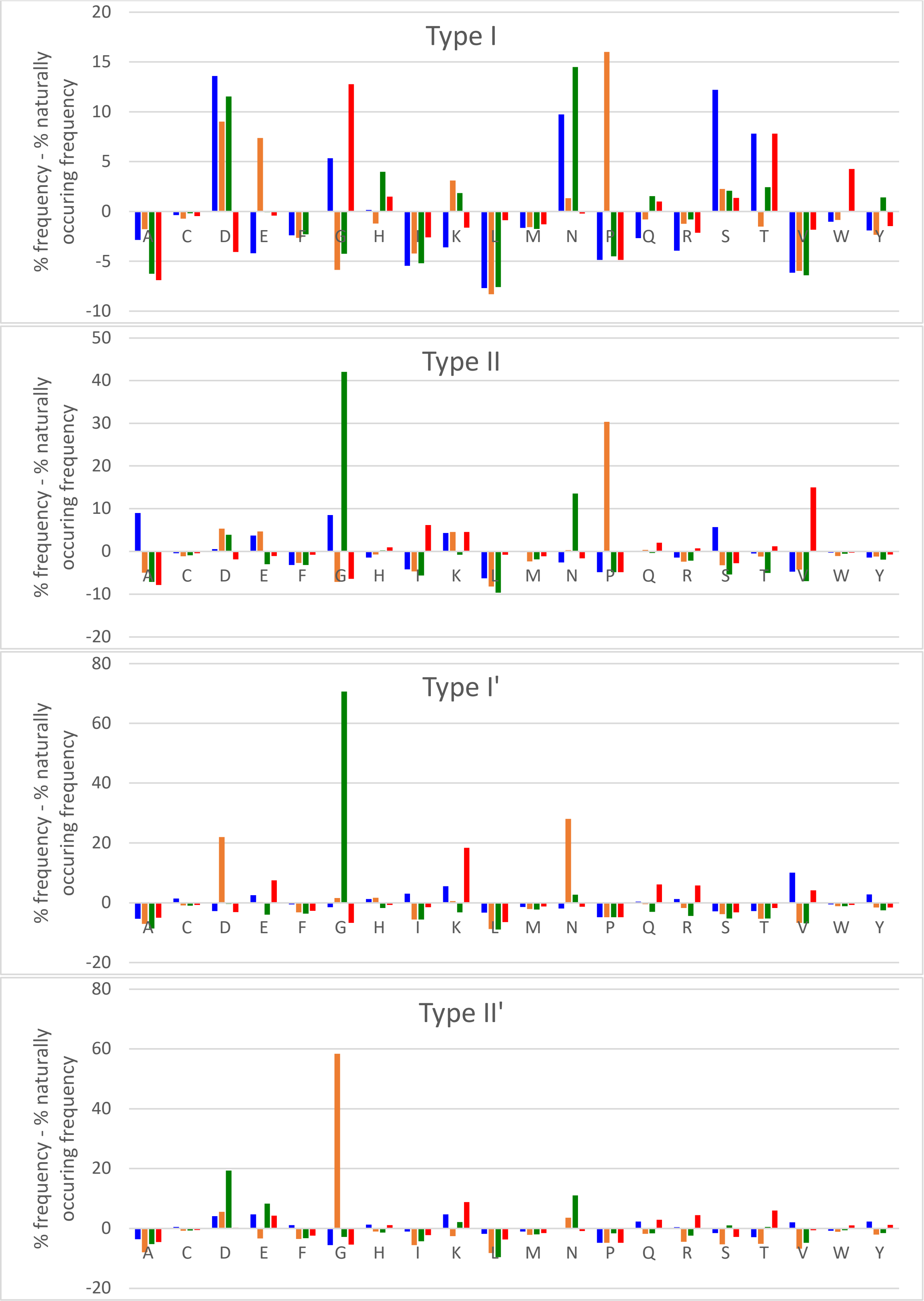


Figure 34: Amino acid composition of the β-hairpins without a hydrogen-bond between the i and i+3 positions. Blue is the i position of the β-turn, orange the i+1, green i+2 and red i+3. Naturally occurring frequency of amino acids based on all available structures in the PDB taken from [301].

3.3 Other β-turns in the Database

The hairpin dataset contains relatively few type II β-turns. β-turns which are not in a β-hairpin were therefore searched for within the Loop Database to expand the available number of β-turns (see section 10.1.2).

β-turns within larger loops in the database were searched for using the hydrogen-bonded definition of a β-turn. A total of 152,226 β-turns were found. This is too large a data sample for the clustering algorithm to complete in a reasonable timescale so a sample of 5,023 was used (a thirtieth of the data). Nine clusters are found when the dihedral angles from i to i+3 residues were included in the clustering (Table 9). In the hairpin dataset most of the i and i+3 dihedrals are in the β region of the Ramachandran plot so contribute little to the separation into different clusters. However for the β-turns the i and i+3 dihedrals contribute to the separation into different clusters. All the clusters that are identified by the algorithm are type I or II β-turns but are also separated based on whether the i and i+3 dihedrals are in the α, β or left-handed α region of the Ramachandran plot. Type I’ and II’ turns are present in too few numbers to be picked out by the clustering algorithm when the i and i+3 dihedrals are included.

Cluster	Turn Type	i/i+3 dihedrals	Population	% of β-turns
1	I	ββ	1,143	22.52
2	I	βα / βLα	894	17.62
3	I	αβ / Lαβ	700	13.79
4	I	αα / αLα	646	12.73
5	II	ββ	499	9.83
6	I	βLα	378	7.45
7	II	βα	157	3.09
8	II	αβ	139	2.74
9	II	αα	24	0.47

Table 9: The clusters found when β-turns from larger loops within the database are clustered based on the i to i+3 dihedral angles.

As β-turn type is determined by the i+1 and i+2 residue dihedrals only, and to try and pick out the type I’ and II’ turns to get clusters compatible with the hairpin turns, the turns were clustered based on the dihedral angles of the i+1 and i+2 residues only. Five clusters are seen: types I, II, I’, II’ and VI (Table 10). The β-turns dataset has a very different distribution of β-turn types compared to the β-hairpin dataset. Type I turns make up the vast majority of the turns found, whereas type I’ and II’ occur infrequently despite being the largest categories in the hairpin dataset.

Turn Type	Population	% of β-turns
I	3,775	74.38
II	920	18.13
I’	195	3.84
II’	107	2.11
VI	26	0.51

Table 10: The clusters found when β-turns from larger loops within the database are clustered based on the i+1 and i+2 dihedral angles.

The same clusters in similar ratios are seen when other samples of the β-turns are clustered. Therefore to assign all the turns to a given turn type, all points in the principal component subspace following PCA were initially added to a cluster based on which cluster centre from the clustered sample they were closest to. The area occupied by each cluster was then divided into grids and grids with low populations were discarded to create the final clustered data (Figure 35). Type VI turns differ from the other β-turn types in that the generally accepted definition for type VI turns, rather than just being based on the i+1 and i+2 dihedral angles, also requires cis proline to occupy the i+2 position. The type VI cluster was therefore further modified to remove any that do not have a proline at this position. Although uncommon non-proline amino acids are sometimes observed to have cis amide bonds, so this process removed 7.65% of the type VI cluster. Higher resolution X-ray crystal structures have been observed to have a higher frequency of cis amide bonds than seen in the lower resolution structures available when the first β-turn type definitions were developed,[308-310] so the extra stipulation that the i+2 position be proline is perhaps redundant. As the β-turns from the database will be used for cyclic peptide structure prediction, it was decided to keep the requirement for proline at the i+2 position with the assumption made that the majority of non-proline cis amide bonds are due to longer-range interactions in the wider protein structural environment and would not occur in the cyclic peptide environment. It has previously been observed many non-proline cis amide bonds occur close to the protein active site with sidechain-to-sidechain interactions of the residue common, especially with aromatic amino acids.[257] The final dataset is made up of 145,479 turns: 108,872 (74.8%) type I, 27,346 (18.8%) II, 2,941 (2.0%) II’, 5,233 (3.6%) I’ and 1,087 (0.7%) VI. Overall this is 96% of all β-turns found within the database.

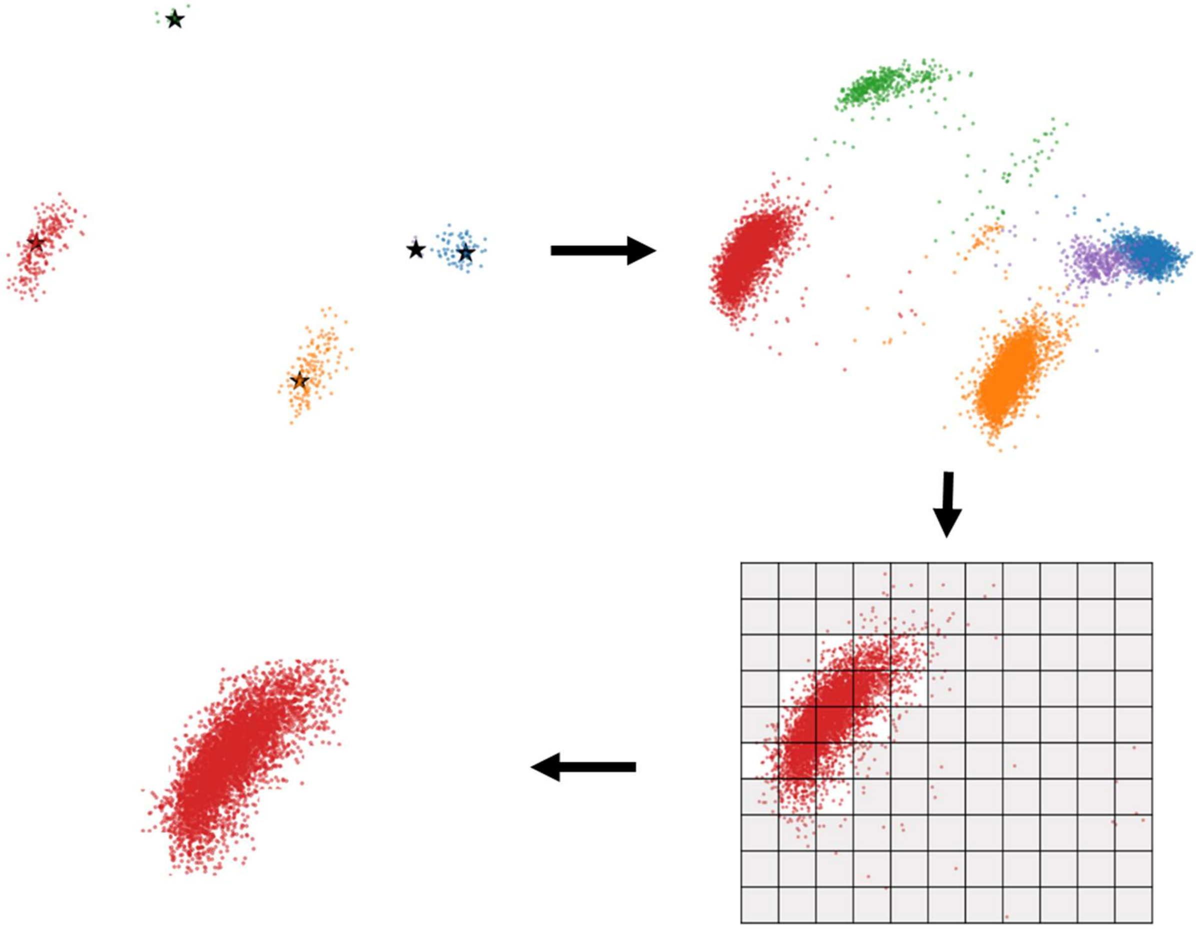


Figure 35: The process used to expand clusters to include all available data.

3.3.1 Sequence Propensities

The amino acid distribution for each position (i to i+3) in the β-turns for the turn types identified in the β-turns data set are shown in Figure 36. Generally similar amino acid compositions are seen to those in the β-hairpin dataset.

Type I turns in the hydrogen-bonded hairpin dataset often had threonine at the i position and glycine at the i+3 position of the β-turn. This is no longer seen in the type I turns of the β-turn dataset. It may be that without the restraints of the hairpin system other amino acids are more able to occupy the i+3 position. A much larger number of type II turns is available compared to the hairpin dataset but PG remains the dominant sequence for the i+1 and i+2 positions. Similar to the hairpin dataset, type I' turns predominantly have NG or DG at the i+1 and i+2 positions, and type II' turns are mainly GN or GD. For type VI turns, by definition, they require a proline at the i+2 position. Aromatic amino acids including tyrosine and phenylalanine are frequently seen on either side of the cis proline. Sidechain interactions with aromatic rings has previously been observed to help stabilise the cis proline conformation.[278]

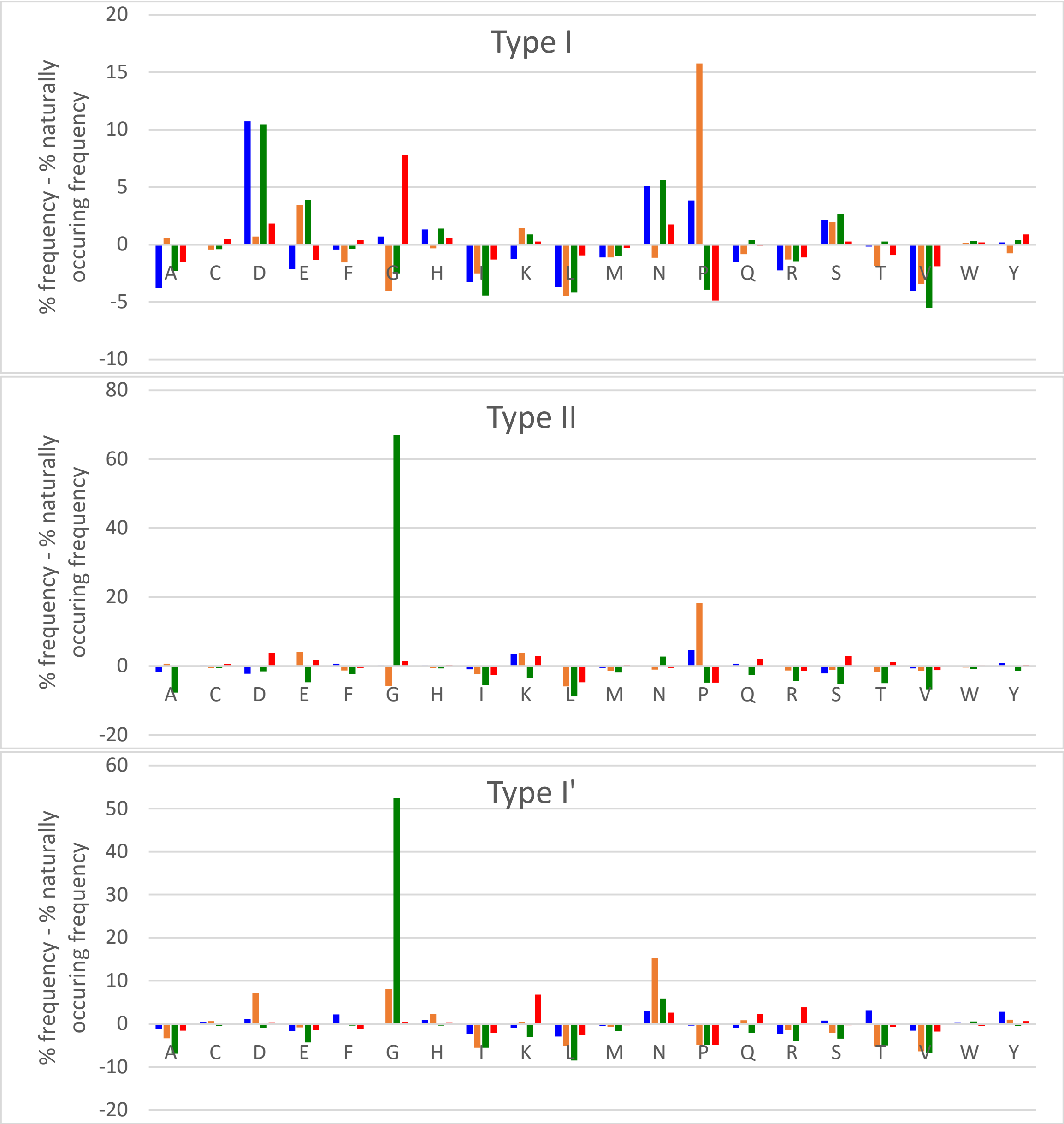


Figure 36: Amino acid propensity for the β-turn types found in the β-turn dataset. Frequency is corrected for size of cluster and naturally occurring frequency of the amino acid. The i position of the β-turn is blue, i+1 orange, i+2 green and i+3 red. Naturally occurring frequency of amino acids based on all available structures in the PDB taken from [301].

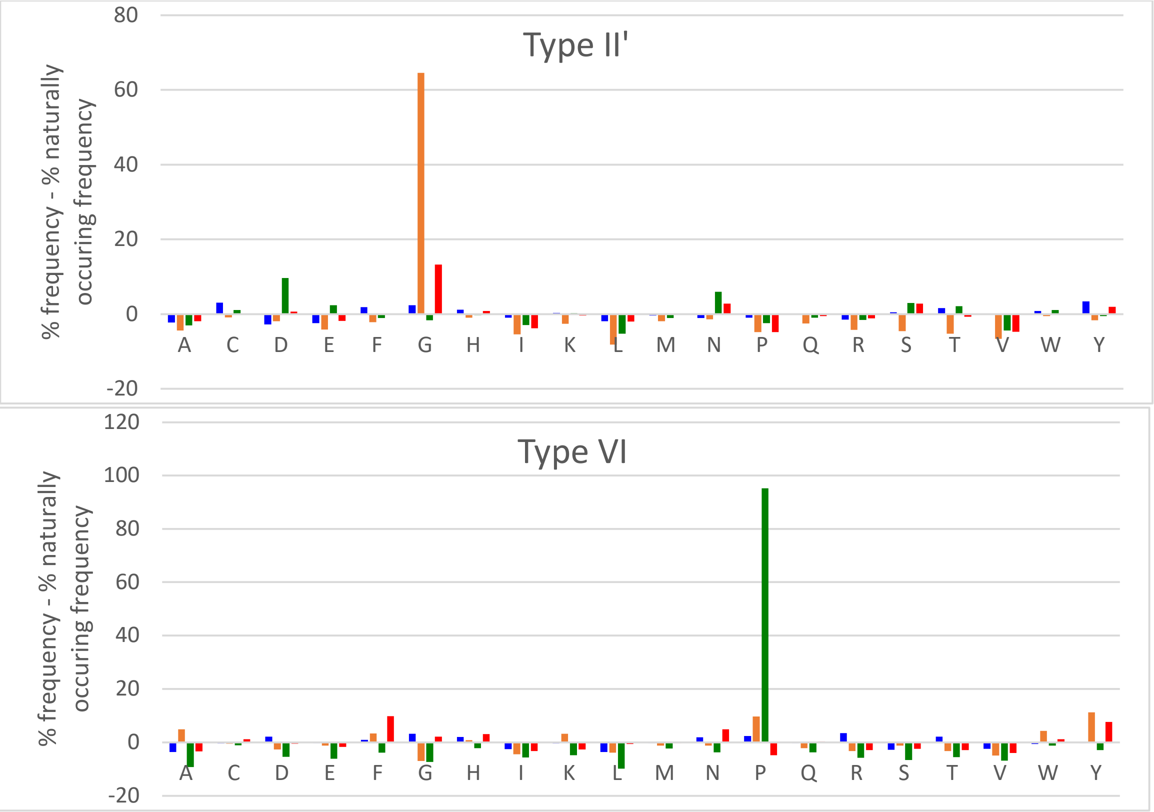


Figure 36 continued: Amino acid propensity for the β-turn types found in the β-turn dataset. Frequency is corrected for size of cluster and naturally occurring frequency of the amino acid. The i position of the β-turn is blue, i+1 orange, i+2 green and i+3 red. Naturally occurring frequency of amino acids based on all available structures in the PDB taken from [301].

Using the hydrogen-bonded definition of a β-turn seems to exclude many turn types. The clustering algorithm used does not assign all points to a cluster so it is possible other turn types appear in insufficient quantities to form a cluster. If no hydrogen bond was required between the i and i+3 positions a much different distribution of turn types may be seen.

3.3.2 7 Å Definition of β-turns

When β-turns are searched for within loops with a distance of less than 7 Å between the α-carbons at the i and i+3 positions, an additional 139,409 β-turns are found than when a hydrogen-bond is required. A tenth of the additional β-turns which do not have a hydrogen bond were clustered and the clusters expanded to include all the data using the method described previously. Clustering the additional β-turn types which do not have a hydrogen-bond produced a very different distribution of turn types than seen in the other datasets (Table 11).

Turn Type	Population (%)
I+IV	49.26
VIII	31.13
II	14.02
IV	2.9
VI_b	2.01
VI_{a1}	0.68

Table 11: The different clusters seen when the additional β-turns identified in loops with a distance of less than 7 Å between α-carbons but no hydrogen-bond are clustered based on the i+1 and i+2 dihedral angles.

The type I' and II' turns appear in too few numbers to form a cluster. Additionally the clusters seen tend to be broader. For example, the cluster containing the type I β-turns is shown in Figure 37. Using the traditional definitions of β-turn types some datapoints assigned to this cluster would not meet the criteria to be assigned as a type I β-turn. Based on de Brevens classification of β-turns this cluster contains type I turns as well as type IV_3 and IV_4 β-turns.[123] It is possible the cluster could be subdivided into different turn types with further clustering. However it would be necessary to use a different clustering algorithm to do so as when further clustering of the cluster was attempted with the current algorithm it could not be separated. The type IV turn which makes up 2.90% of the clustered data is also shown in Figure 37. The average ϕ/ψ dihedral angles for the type IV turn are 85/-20 for the i+1 position and -108/-26 for the i+2 position.

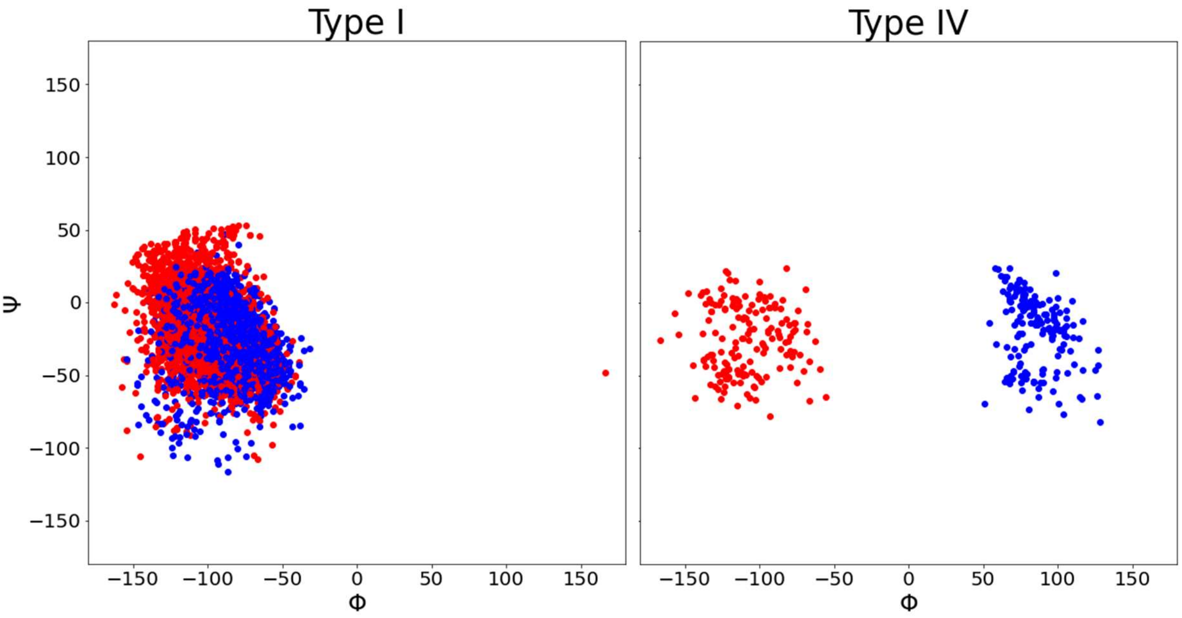


Figure 37: The cluster containing type I turns and the type IV cluster identified when the β-turns found within loops in the Loop Database with a distance less than 7 Å between i/i+3 α-carbons but which had no hydrogen bond were clustered. Blue is the i+1 position of the β-turn and red is the i+2 position.

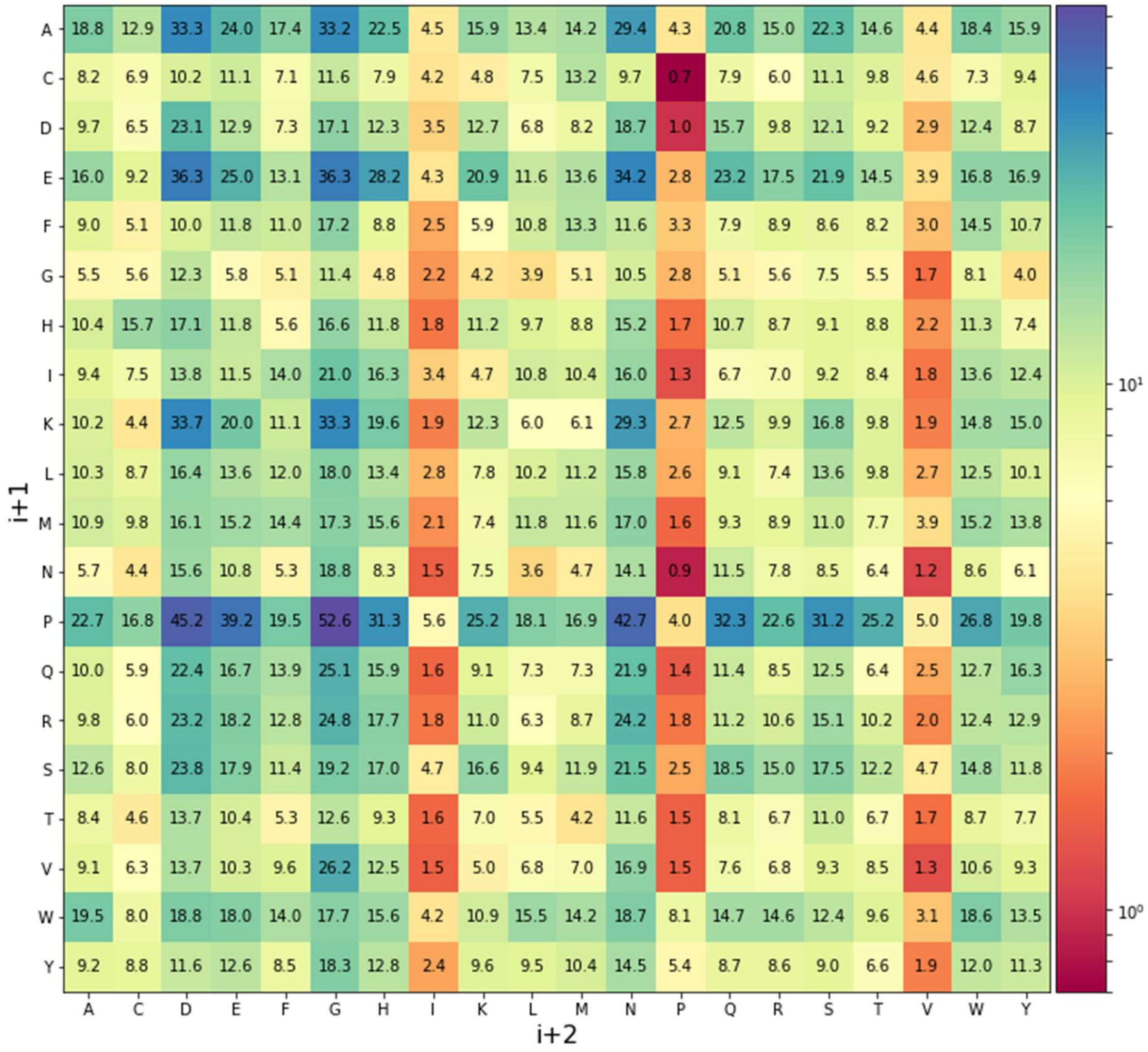
As broad clusters and such a different distribution of β-turn types was seen this dataset of β-turns was not explored further. Without any additional restraints on the system, such as the presence of a hydrogen bond or the β-hairpin structure, the assumption was made the β-turn would no longer be a good model for the β-turns found within the relatively restrained cyclic peptide environment.

Depending on whether the hydrogen bonded or 7 Å distance definition of a β-turn is used different distributions of β-turn types are found. Whether the β-turns are found in β-hairpins or in wider loop regions also leads to vastly different β-turn type distributions. Despite the different prevalence of different turn types depending on where/how they are searched for, the same turn types, with few exceptions, generally have similar sequence distributions. For example type II turns have a high frequency of proline and glycine at the i+1 and i+2 positions respectively regardless of whether the type II turns were from β-hairpins or β-turns found within the wider loop regions and whether the turn does or does not have the hydrogen bond present between the i and i+3 positions.

3.4 Turn Propensity

Searching for all occurrences of a given dipeptide sequence within loop regions in the Loop Database determined the percentage occurrence of the dipeptide appearing at the i+1 and i+2 position of a β-turn. Using sequences that have a high rate of occurrence in a β-turn may help design cyclic hexapeptides with a more stable structure. Such sequences may be more likely to occupy the i+1 and i+2 positions of a β-turn and therefore could potentially be used to define the register of a cyclic peptide.

Proline at the i+1 position or glycine at the i+2 position is present in many of the dipeptide sequences that more frequently occupy the i+1 and i+2 positions of β-turns (Figure 38). Proline introduces a kink in the sequence and glycine is very flexible and able to access many areas of the Ramachandran plot not typically accessed by other amino acids, therefore it is unsurprising that they increase the proportion of β-turns. Valine, proline and isoleucine at the i+2 position on the other hand have a very low rate of occurrence in β-turns.



	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
A	18.8	12.9	33.3	24.0	17.4	33.2	22.5	4.5	15.9	13.4	14.2	29.4	4.3	20.8	15.0	22.3	14.6	4.4	18.4	15.9
C	8.2	6.9	10.2	11.1	7.1	11.6	7.9	4.2	4.8	7.5	13.2	9.7	0.7	7.9	6.0	11.1	9.8	4.6	7.3	9.4
D	9.7	6.5	23.1	12.9	7.3	17.1	12.3	3.5	12.7	6.8	8.2	18.7	1.0	15.7	9.8	12.1	9.2	2.9	12.4	8.7
E	16.0	9.2	36.3	25.0	13.1	36.3	28.2	4.3	20.9	11.6	13.6	34.2	2.8	23.2	17.5	21.9	14.5	3.9	16.8	16.9
F	9.0	5.1	10.0	11.8	11.0	17.2	8.8	2.5	5.9	10.8	13.3	11.6	3.3	7.9	8.9	8.6	8.2	3.0	14.5	10.7
G	5.5	5.6	12.3	5.8	5.1	11.4	4.8	2.2	4.2	3.9	5.1	10.5	2.8	5.1	5.6	7.5	5.5	1.7	8.1	4.0
H	10.4	15.7	17.1	11.8	5.6	16.6	11.8	1.8	11.2	9.7	8.8	15.2	1.7	10.7	8.7	9.1	8.8	2.2	11.3	7.4
I	9.4	7.5	13.8	11.5	14.0	21.0	16.3	3.4	4.7	10.8	10.4	16.0	1.3	6.7	7.0	9.2	8.4	1.8	13.6	12.4
K	10.2	4.4	33.7	20.0	11.1	33.3	19.6	1.9	12.3	6.0	6.1	29.3	2.7	12.5	9.9	16.8	9.8	1.9	14.8	15.0
L	10.3	8.7	16.4	13.6	12.0	18.0	13.4	2.8	7.8	10.2	11.2	15.8	2.6	9.1	7.4	13.6	9.8	2.7	12.5	10.1
M	10.9	9.8	16.1	15.2	14.4	17.3	15.6	2.1	7.4	11.8	11.6	17.0	1.6	9.3	8.9	11.0	7.7	3.9	15.2	13.8
N	5.7	4.4	15.6	10.8	5.3	18.8	8.3	1.5	7.5	3.6	4.7	14.1	0.9	11.5	7.8	8.5	6.4	1.2	8.6	6.1
P	22.7	16.8	45.2	39.2	19.5	52.6	31.3	5.6	25.2	18.1	16.9	42.7	4.0	32.3	22.6	31.2	25.2	5.0	26.8	19.8
Q	10.0	5.9	22.4	16.7	13.9	25.1	15.9	1.6	9.1	7.3	7.3	21.9	1.4	11.4	8.5	12.5	6.4	2.5	12.7	16.3
R	9.8	6.0	23.2	18.2	12.8	24.8	17.7	1.8	11.0	6.3	8.7	24.2	1.8	11.2	10.6	15.1	10.2	2.0	12.4	12.9
S	12.6	8.0	23.8	17.9	11.4	19.2	17.0	4.7	16.6	9.4	11.9	21.5	2.5	18.5	15.0	17.5	12.2	4.7	14.8	11.8
T	8.4	4.6	13.7	10.4	5.3	12.6	9.3	1.6	7.0	5.5	4.2	11.6	1.5	8.1	6.7	11.0	6.7	1.7	8.7	7.7
V	9.1	6.3	13.7	10.3	9.6	26.2	12.5	1.5	5.0	6.8	7.0	16.9	1.5	7.6	6.8	9.3	8.5	1.3	10.6	9.3
W	19.5	8.0	18.8	18.0	14.0	17.7	15.6	4.2	10.9	15.5	14.2	18.7	8.1	14.7	14.6	12.4	9.6	3.1	18.6	13.5
Y	9.2	8.8	11.6	12.6	8.5	18.3	12.8	2.4	9.6	9.5	10.4	14.5	5.4	8.7	8.6	9.0	6.6	1.9	12.0	11.3

Figure 38: Percentage occurrence of dipeptide sequences at the i+1 and i+2 positions β-turns in the database compared to all occurrences of that sequence in the database.

3.5 Using the Loop Database to Design Cyclic Peptides

As cyclic hexapeptides often form a structure in solution that contains two overlapping β-turns, the analysis of β-turns from the database could potentially help design the cyclic peptide structure. The RGD motif is commonly found in integrin binding proteins.[40] The RGD motif must be in the correct orientation for binding to the relevant integrin to occur.[40] Data from the Loop Database was used to design two cyclic peptides each with the RGD motif held in a different orientation.

3.5.1 Design

When designing a stable cyclic hexapeptide with a conformation equivalent to two overlapping β-turns, the amino acids should remain in the same register i.e. the amino acid at the i+1 position of a β-turn should remain at the i+1 position rather than exchanging between different conformations where it occupies the i+2 or i/i+3 positions. The percentage occurrence of dipeptide sequences at the i+1 and i+2 positions of β-turns in the database was used as an estimate of turn propensity for a given sequence. Sequences with a higher turn propensity should be more likely to occupy the i+1 and i+2 positions of a cyclic peptide and so could potentially be used to design cyclic peptides with fewer conformations.

The RGD motif was to be incorporated within a cyclic peptide with the arginine at a i+2 position of a β-turn, glycine at one of the bridging i/i+3 position and aspartic acid at the i+1 position of the other β-turn. The RGD sequence is therefore in an elongated orientation down one side of the cyclic peptide structure (Figure 39). When arginine is required at the i+2 position of a β-turn, choosing a proline at the i+1 position gives the highest percentage occurrence of the dipeptide sequence being in a β-turn in the Loop Database. 22.6% of all PR sequences within the loops in the database are found at the i+1 and i+2 position of a β-turn. When aspartic acid is at the i+1 position DD, DG and DN have the highest percentage β-turn formation. DD has a higher percentage than DN but DN was selected for ease of synthesis. For the remaining i/i+3 position valine was chosen. The VP and NV sequences have a low percentage occurrence at the i+1 and i+2 positions so this may make the alternative registers more unfavourable. Valine is also seen frequently at the i position of the β-hairpin dataset and has previously been shown to lead to cyclic hexapeptides with a more stable structure.[228] This gives a sequence of c(VPRGDN).

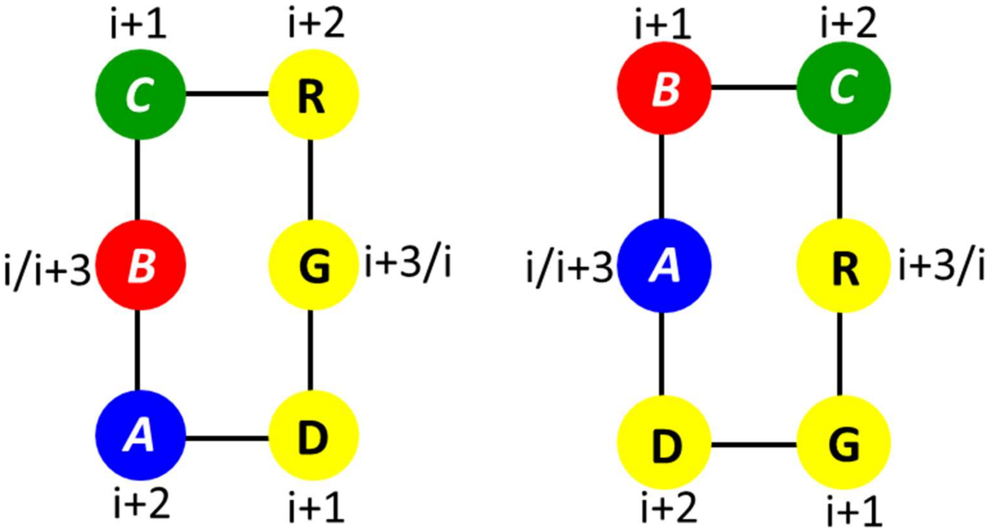


Figure 39: Two RGD registers within a cyclic hexapeptide.

A second sequence was designed to have the RGD sequence at the i to i+2 positions of a β-turn. The RGD motif would therefore be in a bent rather than elongated orientation which could alter its binding properties. For this orientation one of the β-turns within the cyclic hexapeptide requires the sequence GD to be at the i+1 and i+2 positions. The GD sequence occupies the i+1 and i+2 positions of a β-turn 12.3% of the time within the Loop Database. In order to prevent the register of the amino acids changing within the cyclic peptide and keep the RGD motif occupying the i to i+2 positions of a β-turn, the other β-turn should ideally have a high turn propensity to induce this conformation. The dipeptide sequences most frequently found at the i+1 and i+2 positions of β-turns are PD, PG and PN. PN was chosen as it would allow for sidechain anchoring when the peptide was synthesised. Valine was once again chosen to occupy the remaining i/i+3 position. This gives the sequence c(VPNRGD). The two peptides therefore contain the same amino acids but in a different order. Small changes in amino acid sequence have previously been shown to be able to significantly change the conformation of small cyclic peptides.[41, 245]

For the two sequences the estimated turn propensities based on the Loop Database analysis for the three possible registers that could form in the cyclic hexapeptide structure made up of two overlapping β-turns are shown in Table 12. In each case register 3 is unlikely to occur as it would require proline at one of the i/i+3 positions of the cyclic hexapeptide. As seen in the database analysis proline very rarely occupies the i/i+3 positions of the β-turns, especially in the hairpin dataset. Having proline prevents formation of a hydrogen-bond between the i and i+3 positions of a β-turn as the proline does not have an amide hydrogen. Additionally the proline structure introduces a kink into the sequence so much more favourably occupies the i+1 turn position. The estimated turn propensity does not include the i/i+3 positions so could potentially be expanded to include such values. In each case register 1 has the highest estimated turn propensity. For c(VPNRGD) the RG turn in register 2 has a relatively high turn propensity so could potentially form a minor cluster. However it occurs in conjunction with the VP turn which has a low estimated turn propensity.

c(VPRGDN)

Register	i+1/i+2	% occurrence in β-turn	i+1/i+2	% occurrence in β-turn
1	PR	22.6	DN	18.7
2	VP	1.5	GD	12.3
3	NV	1.2	RG	24.8

c(VPNRGD)

Register	i+1/i+2	% occurrence in β-turn	i+1/i+2	% occurrence in β-turn
1	PN	42.7	GD	12.3
2	VP	1.5	RG	24.8
3	DV	2.9	NR	7.8

Table 12: Turn propensity for the possible registers for c(VPNRGD) and c(VPRGDN) based on the percentage occurrence of dipeptide sequences at the i+1 and i+2 positions of a β-turn in the Loop Database.

3.5.2 BE-META

Both the c(VPNRGD) and c(VPRGDN) sequences were designed based on the ability of the sequences to form turns. If the percentage occurrence a dipeptide sequence is found at the i+1 and i+2 positions of a β-turn within the Loop Database can be used as an estimate of turn propensity then both peptides should form conformations with only one amino acid register induced by the turn sequences chosen. Despite the very similar sequences they should have very different conformations with the RGD motif held in different orientations. BE-META was used to predict the conformations in solution.

For c(VPRGDN) all clusters seen in the BE-META had the amino acids at the predicted registers with the RGD motif in the elongated orientation down one side of the peptide structure (Table 13). The major conformation (70%) contained two type II β-turns.

Cluster	Population (%)	Conformation
1	70	II + II
2	21	II + I
3	9	I + II

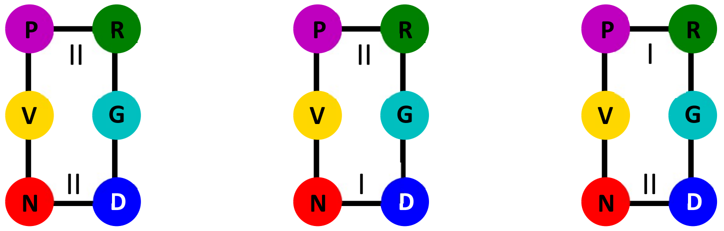


Table 13: Conformations seen in the BE-META of c(VPRGDN).

Three clusters are seen in the BE-META for c(VPNRGD) all with the amino acids in the predicted register based on the turn propensity of the amino acids. The largest cluster contained a type II' GD turn and a type II PN turn. PN also formed a type II turn in the remaining clusters, with GD forming multiple turn types.

Cluster	Population (%)	Conformation
1	51	II' + II
2	34	IV + II
3	15	I + II

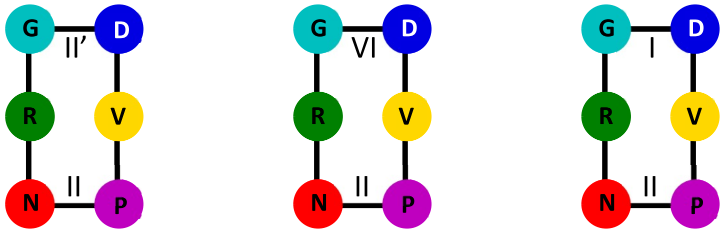
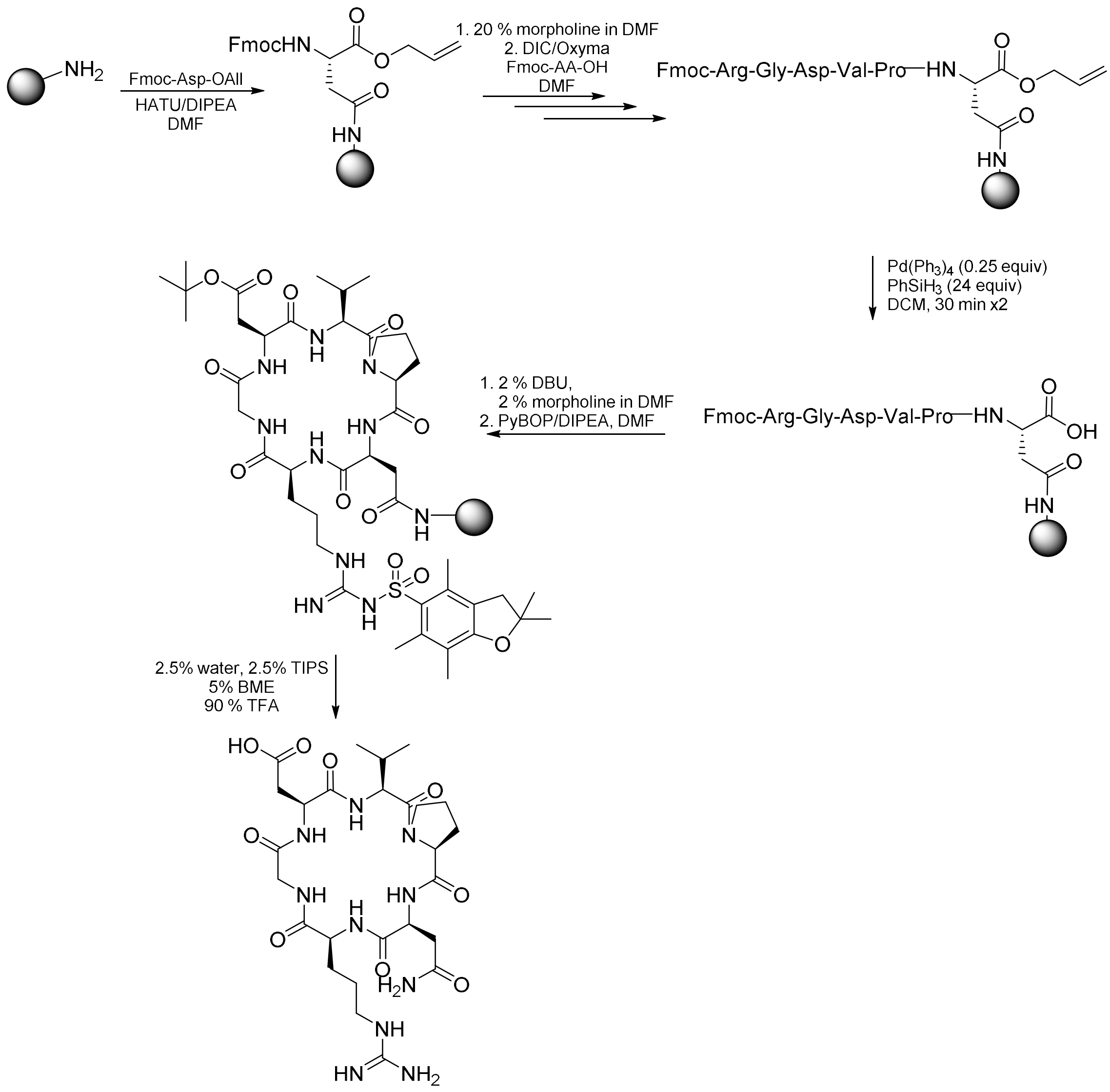


Table 14: Conformations seen in the BE-META of c(VPNRGD).

The BE-META predicts that both c(VPNRGD) and c(VPRGDN) will adopt conformations in solution with registers that are predicted based on the turn propensities obtained from analysis of the Loop Database. To confirm the BE-META predictions the peptides were synthesised and their conformations determined by NMR.

3.5.3 Synthesis of c(VPNRGD) and c(VPRGDN)

Both c(VPNRGD) and c(VPRGDN) were synthesised using a three orthogonal protecting group strategy with sidechain anchoring through the asparagine sidechain used to allow for on-resin cyclisation (Scheme 14). Fmoc-Asp-OAll was coupled to Rink Amide AM resin through its sidechain. The remaining residues were coupled using SPPS. The allyl protecting group was then removed from the C-terminus using a palladium catalyst. After removal of the final *N*-terminal Fmoc the peptide can then be cyclised prior to cleavage from the resin. Upon cleavage from the resin, rather than aspartic acid, asparagine is produced.



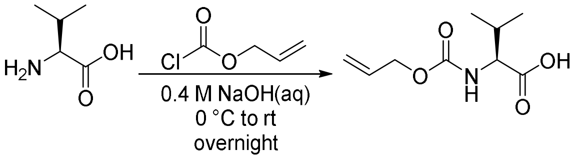
Scheme 14: Synthesis of c(VPNRGD).

The final Fmoc deprotection in the synthesis is carried out with 2% DBU and 2% morpholine in DMF rather than 20% morpholine in DMF. This is used in cases where Fmoc deprotection can be difficult due to sterics.[311] However for c(VPRGDN) even 2% DBU and 2% morpholine was insufficient to remove the Fmoc group from the valine to allow for cyclisation. Sequences with VP and PP have previously been shown to have difficult to remove Fmoc groups due to the propensity of these sequences to form β-turns.[280] The formation of intra molecular hydrogen bonds and hydrophobic interactions can make it difficult to remove the Fmoc group. Other deprotection solutions also failed to remove the Fmoc group (Table 15).

Base	Temperature (°C)	Time (h)
2% DBU, 2% morpholine	rt	0.5
20% morpholine, 5% DBU	rt	1
20% morpholine	80	0.5
20% piperidine, 5% DBU	60	0.5
50% morpholine	60	0.5
0.1M TBAF	rt	0.5
44% piperidine, 1% DBU	90	0.5
20% morpholine	rt	12

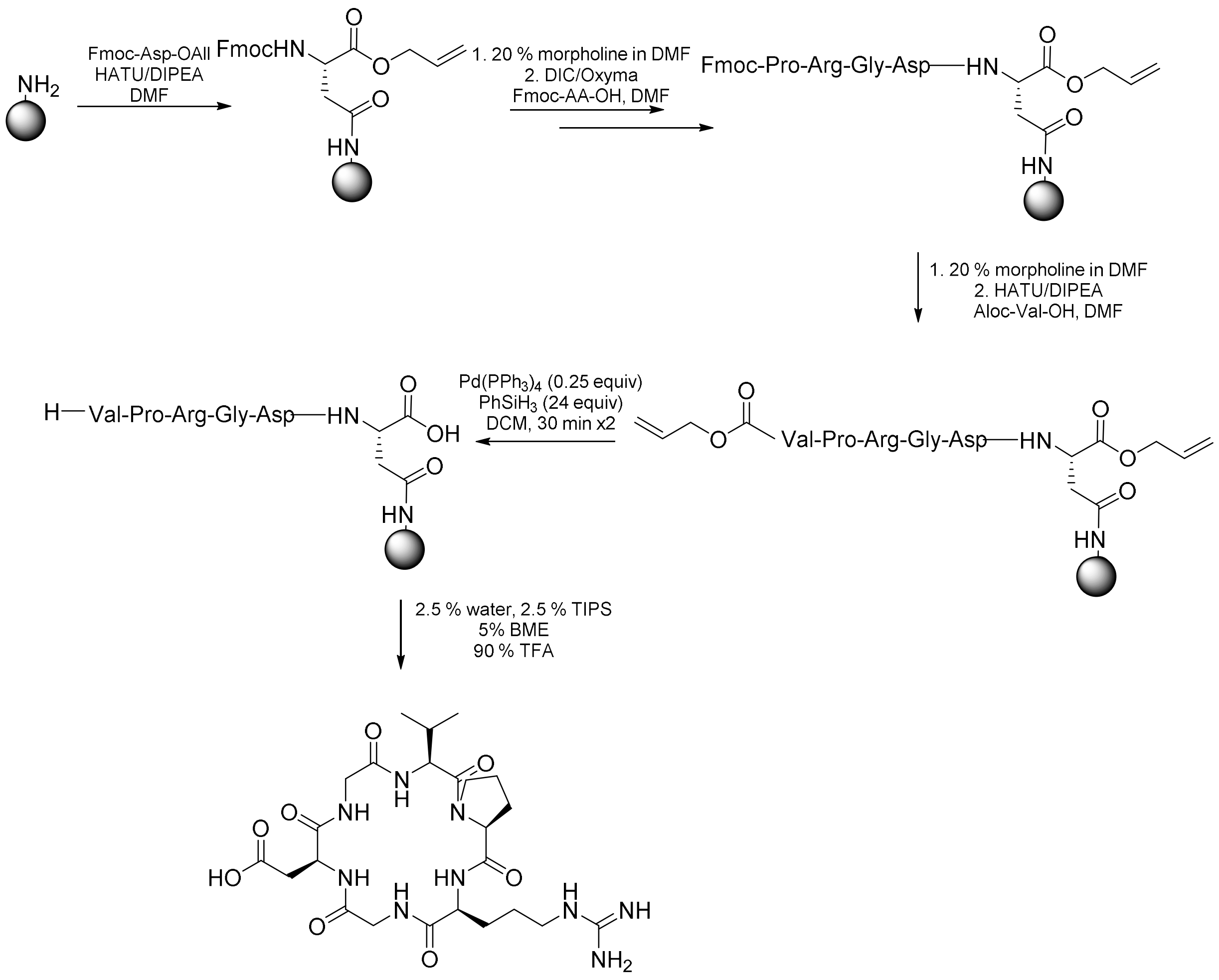
Table 15: Conditions tested to remove the final Fmoc during the synthesis of c(VPRGDN). DMF was used as the solvent for all reactions.

Since the final Fmoc group could not be removed from the valine an alternative synthesis was designed for c(VPRGDN). Aloc-valine was synthesised from valine using one equivalent of allyl chloroformate in 0.4 M sodium hydroxide solution (Scheme 15) and obtained in a 69% yield. SPPS could then be used to couple the aloc-valine into the peptide and the smaller aloc protecting group could be removed in the same step used to remove the allyl protecting group from the C-terminus (Scheme 16).



Scheme 15: Synthesis of Aloc-Val-OH.

To couple the Aloc-valine to the peptide, two equivalents were used with HATU as the coupling reagent. To remove the allyl protecting group from the C-terminus 0.25 equivalents of $Pd(PPh_3)_4$ catalyst was used in the presence of 24 equivalents of phenylsilane scavenger in dry DCM. This reaction was repeated. It was found that the aloc group was removed from the valine *N*-terminus under the same conditions as removal of the allyl group with no additional catalyst required. The peptide could then be cyclised on resin using PyBOP as for the previous peptide. Both c(VPNRGD) and c(VPRGDN) were purified by HPLC and their NMR spectra obtained.



Scheme 16: Synthesis of c(VPRGDN).

3.5.4 NMR of c(VPNRGD) and c(VPRGDN)

NMR of c(VPNRGD) and c(VPRGDN) were obtained at a concentration of 1 mM in 5% D_2O in water with pH 7.4 potassium phosphate buffer at 278 K. The structure of the peptides was determined and compared to the predicted structure from the BE-META simulations.

3.5.4.1 c(VPNRGD)

The NMR of c(VPNRGD) shows two conformations. Cis/trans proline isomerisation is relatively slow on the NMR timescale so, as the peptide contains a proline residue, could account for the two conformations. The major conformation has a ROE cross-peak between the valine α-carbon and proline δ-carbons showing the conformation contains a trans proline. If the preceding residue is large then cross-peaks may also be seen between the sidechain protons and the trans proline sidechain (Figure 40). A ROE cross-peak is seen between the valine β-hydrogens and the proline δ-hydrogens further supporting the conformation containing a trans proline. The minor conformation has a ROE cross-peak between the valine and proline α-carbons showing it contains a cis proline. The ratio between the cis and trans proline-containing conformations is 1:2.2.

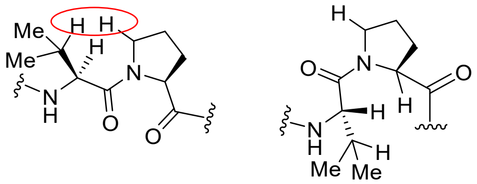


Figure 40: The valine sidechain protons are only close in space to the proline δ-protons in the trans conformation.

Strong ROE cross-peaks are generally seen between the amide hydrogens of the i+2 and i+3 positions in β-turns as they are relatively close in space (Figure 41). In more well-structured cyclic peptides, where the structure is typically made up of two overlapping well-defined β-turns, only two amide-to-amide ROE cross-peaks are therefore commonly seen. This would decrease to only one amide-to-amide ROE if the peptide contains a type VI turn as the proline at the i+2 position does not have the amide proton. If however the peptide forms multiple conformations with the position of the amino acids in the β-turns changing (e.g an amino acid at position i+1 in a β-turn moves to the i+2 position), then typically more amide-to-amide ROE cross-peaks are seen.

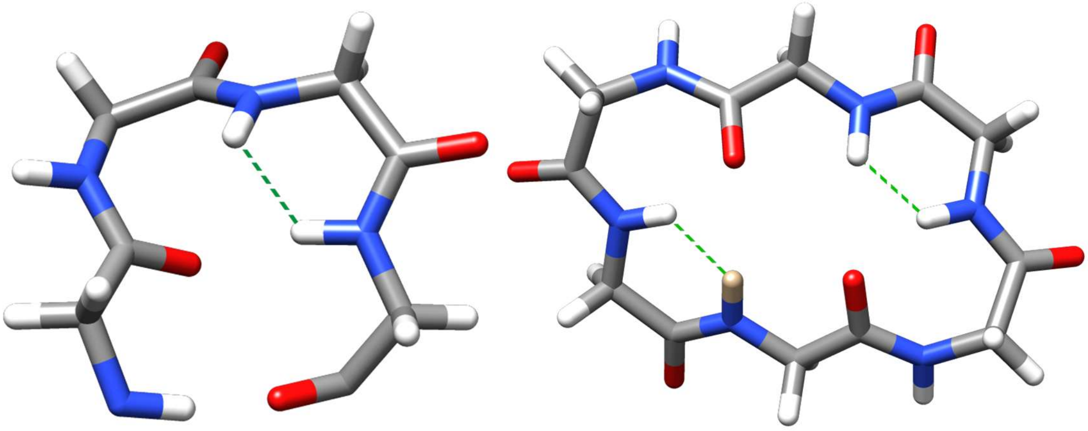


Figure 41: ROE cross-peaks are generally seen between the i+2 and i+3 positions of a β-turn. In well-structured cyclic hexapeptides only two amide-to-amide ROE cross-peaks are seen.

Assignment of the NMR of c(VPNRGD) was difficult due to many peaks overlapping with each other or being masked by the water signal. Therefore it was not possible to assign the amide-to-amide cross-peaks that belong to the major conformation. For the minor conformation two such cross-peaks are seen: one between aspartic acid and valine and a stronger one between aspartic acid and glycine. The minor conformation may therefore be interchanging between multiple conformations all with a cis proline. This is supported by the valine methyl groups having only one carbon peak and only a small chemical shift between the protons. In less structured peptides it is common for the same/very similar shifts to be seen for the two methyl groups due to sidechain flexibility, whereas in very structured peptides the two methyl groups often have distinct shifts. As the strongest amide-to-amide ROE cross-peak is between aspartic acid and glycine it is likely the conformation mainly exists with the cis proline at the i+2 position of a type VI β-turn and another β-turn at the other side of the cyclic peptide with the glycine and aspartic acid at the i+2 and i+3 positions respectively (Figure 42).

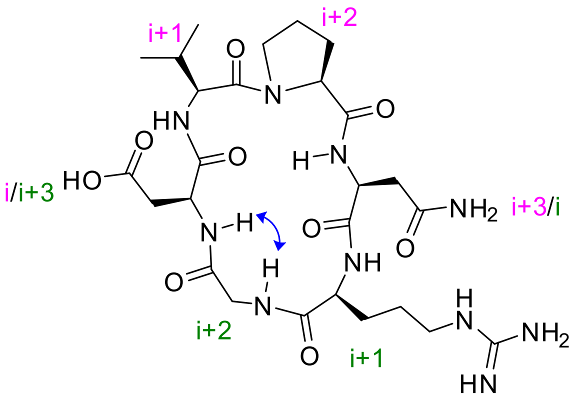


Figure 42: Potential cis proline-containing conformation of c(VPNRGD). Proline occupies the i+2 position of a type VI β-turn. The blue arrow shows the close proximity of the glycine and aspartic acid amide protons which would result in a ROE cross-peak in the NMR.

The BE-META of c(VPNRGD) did not predict a cis proline-containing conformation but as discussed in Chapter 2 the RSFF2 forcefield does not always accurately model the energy difference between the cis and trans proline states in cyclic hexapeptides. The major conformation contains a trans proline and likely has the same register as predicted by the BE-META and turn propensity from the database analysis as trans proline frequently occupies the i+1 position of β-turns. Unfortunately due to many peaks overlapping and signals being masked by the water limited additional information could be obtained from the NMR.

3.5.4.2 c(VPRGDN)

The NMR of c(VPRGDN) has only one conformation. A ROE cross-peak between the α-hydrogen of the valine and the δ-hydrogen of the proline is seen showing the conformation contains a trans proline. This is further supported by a ROE cross-peak between the valine β-hydrogen and proline δ-hydrogens.

The single conformation seen is likely to be well-structured as there is a broad range of amide hydrogen shifts and the two methyl groups on the valine have distinct shifts with a relatively large difference of 0.3 ppm. This indicates the conformation is relatively well structured as otherwise the peaks would merge and average values would be seen.

Figure 43 shows the amide-to-amide (d_{NN}) and α-proton-to-amide ($d_{\alpha N}$) NOE cross-peaks seen in different types of β-turns. The thicker the line the stronger the NOE between the positions in the β-turn. The ^{3}J values between the amide and α-hydrogens is also shown.

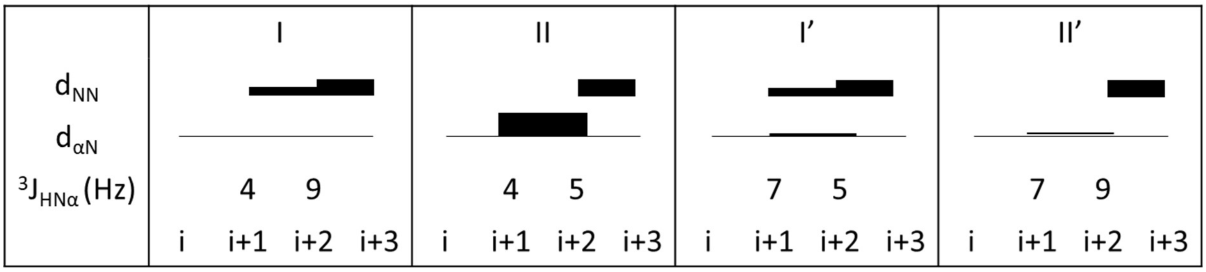


Figure 43: d_{NN} and $d_{\alpha N}$ NOE cross-peaks seen between the i to i+3 positions in a β-turn. Figure based on [312].

Amide-to-amide (d_{NN}) ROE cross-peaks are seen between asparagine and valine as well as arginine and glycine. No other d_{NN} ROEs are seen which is consistent with the cyclic peptide being well-structured with the amino acids staying in defined registers. For type I turns weak d_{NN} ROE may also be seen between the i+1 and i+2 positions of the β-turn, however no such cross-peaks are seen. It is therefore possible the β-turns are type II turns as predicted by the BE-META. Strong cross-peaks can generally be seen between the α-hydrogen of the i+1 position and the amide hydrogen ($d_{\alpha N}$) of the i+2 position for type II turns, whereas only weak ROEs are seen for other β-turn types. Such strong ROE cross-peaks were seen between proline and arginine as well as aspartic acid and asparagine. It is therefore likely the cyclic peptide is forming a conformation in solution made up of two type II β-turns. Due to overlapping and missing peaks the $^3J_{HN\alpha}$ values could not be determined but would have helped confirm the type II turn assignment.

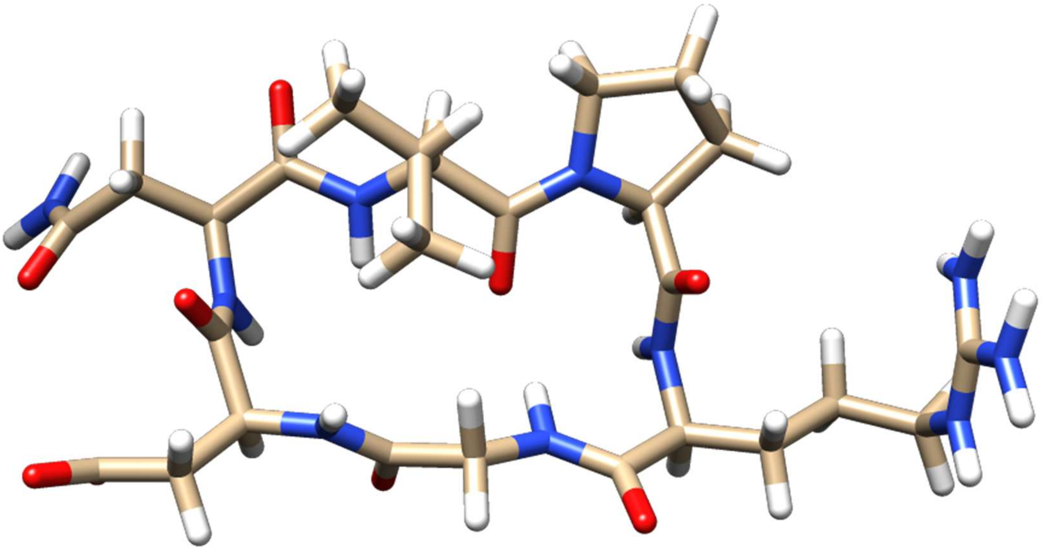


Figure 44: c(VPRGDN) with a conformation of two overlapping type II β-turns seen in the BE-META.

Only one major cluster is seen in the BE-META of c(VPRGDN) which is composed of two overlapping type II turns. This is consistent with what is seen in the NMR. The peptide adopts the register predicted by the turn propensities determined by the analysis of the Loop Database.

Based on the NMR it is likely the major conformation for both c(VPNRGD) and c(VPRGDN) has the predicted amino acid register based on the sequence turn propensity estimated from analysis of the Loop Database. A minor conformation is seen for c(VPNRGD) containing cis proline. The RG subsequence is predicted to have a relatively high turn propensity which may make this conformation more favorable despite the low predicted turn propensity for the VP sequence, but the NMR indicates multiple cis proline-containing conformations are interchanging. The cyclic hexapeptide environment may make the VP turn more favourable than estimated based on the Loop Database analysis as the type VI turn has previously been shown to be more favourable in small cyclic peptide environments.[194]

3.6 Conclusions

As cyclic hexapeptides often form a structure composed of two overlapping β-turns, β-turns within a Loop Database were analysed to see if common features could be extracted to help design well-structured cyclic peptides. Type I, II, I’ and II’ turns were very common within the database. The β-hairpin and β-turn datasets have very different distributions of turn types but generally similar amino acid sequences occur for the same turn types across the two datasets.

The percentage occurrence of a dipeptide sequence occupying the i+1 and i+2 positions of β-turns within the Loop Database was used as an estimate of turn propensity to design two cyclic peptides containing the integrin-binding RGD motif. The two peptides were designed to have the RGD motif in different orientations. For cases where Fmoc removal becomes very difficult due to sterics, it was found an alternative aloc protecting group can be used. For c(VPRGDN) the conformation observed in the NMR matched that seen in the BE-META and predicted based on the turn propensities. For c(VPRGDN) the major conformation likely matches the predicted register but a smaller conformation containing cis proline was also seen. The relatively restrained cyclic peptide environment may change which β-turn types will occur from those which would be predicted by the database analysis.

Both c(VPNRGD) and c(VPRGDN) were originally designed based on the RGD motif occupying a specific register within the cyclic peptide. The β-turn type however was not considered. Analysis of the database shows each β-turn type has unique amino acid propensities. A machine learning algorithm can therefore be trained to predict the β-turn type that will form from a given sequence. The prediction could then possibly be used to predict the β-turn types likely to form within a cyclic hexapeptide sequence. This is explored in Chapter 5.

4 Restrained Simulations

Cyclic peptides are increasingly being investigated as potential drug molecules as they have the potential to target protein-protein interactions (PPIs).[99] However the design of cyclic peptides for such purposes is currently difficult. Structure prediction remains a problem as does the tendency of cyclic peptides to form many conformations in solution rather than a single major conformation. Cyclic hexapeptides have frequently been observed to form a structure in solution made up of two overlapping β-turns, often with two intramolecular hydrogen-bonds.[163, 194, 197, 199, 200, 203, 313-335] The structure can therefore be summarised as a combination of two β-turns. For example if a type II β-turn is seen at one side of the cyclic peptide and a type II' turn at the other the structure can be described as having a II+II' conformation (Figure 45).

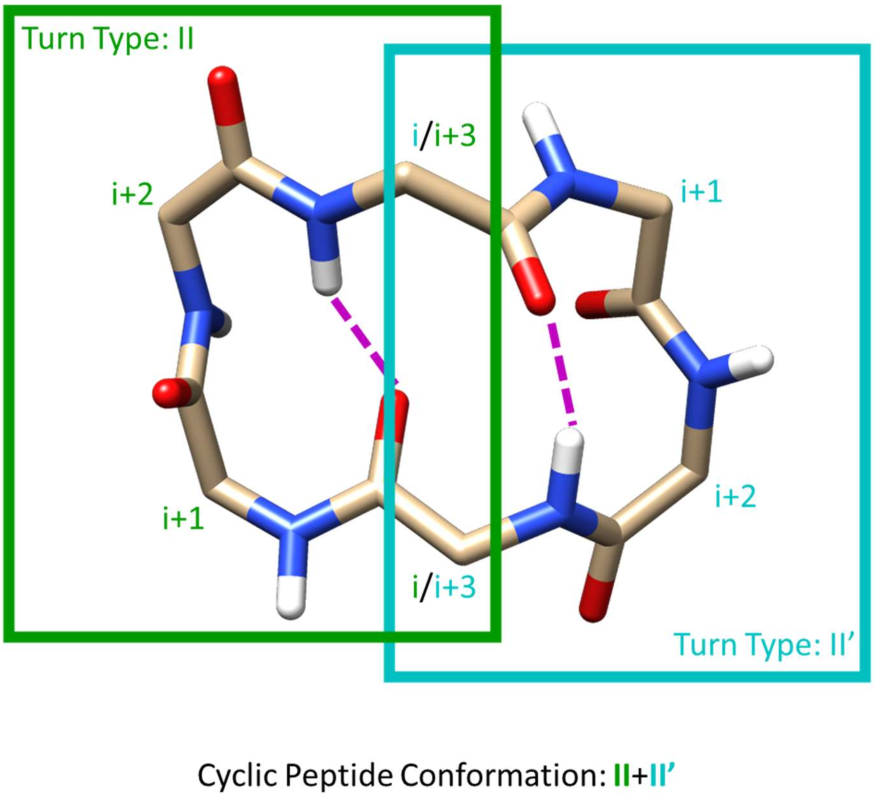


Figure 45: A cyclic peptide with a structure made up of overlapping type II and II' β-turns has a II+II' conformation. Hydrogen bonds are shown as magenta dashed lines.

It has been observed in known NMR or X-ray structures of cyclic peptides that certain conformations, such as those with a I+I or I'+I' turn-combination, occur less frequently than others.[245] It may be that such conformations are seen infrequently because the backbone conformation necessary for these structures is unfavourable. When designing cyclic peptides it may be beneficial to choose a conformation made up of a more favourable turn type combination as this may lead to one prevalent conformation rather than the multiple conformations usually seen. This could lead to stronger binding of a target molecule if the peptide conformation is preorganised into the right orientation for interaction with a ligand. A method to determine the most stable turn type combinations could therefore benefit the design of cyclic peptides with a stable structure. This chapter presents the results of restrained simulations, which were designed to determine the lowest energy backbone conformations cyclic hexapeptides can adopt when a type I, II, I' or II' turn is included within the structure. These turn types have previously been shown to occur frequently within cyclic hexapeptide structure.[245]

4.1 Conformation of c(GGGGGG) and Related Sequences

In order to try to determine sequence-structure relationships of cyclic hexapeptides McHugh et al. carried out a series of BE-META simulations on c(GGGGGG) and related sequences with alanine or valine substitutions.[245] The results of their simulations are discussed below. Common features were seen throughout the simulations but, due to the tendency of cyclic hexapeptides in solution to adopt many conformations, the effects of substituting an amino acid can be complex and hard to interpret. The restrained simulations were therefore designed to look at small changes in the cyclic peptide structure.

The peptide c(GGGGGG) should show the lowest energy conformations of cyclic hexapeptides in the absence of sidechain-specific effects. The I+I' and I+II'/I'+II combinations occurred most throughout the simulation and as such are lower energy than the other conformations for c(GGGGGG).[245] McHugh *et al*. suggested this was due to a reduction in the coulombic interactions of the C=O which point above and below the plane of the ring in these conformations (Figure 46). The distance between the α-carbons of the i and i+3 positions in β-turns has been measured in a variety of models and datasets and type II/II' turns are generally found to be 0.1-0.4 Å wider than type I/I' turns.[124, 336, 337] As the type I and I' turns are narrower than type II and II' they suggested that cyclic peptides with a I+I or I'+I' conformation are seen less frequently than those with other turn combinations due to the increased coulombic repulsion in these conformations.

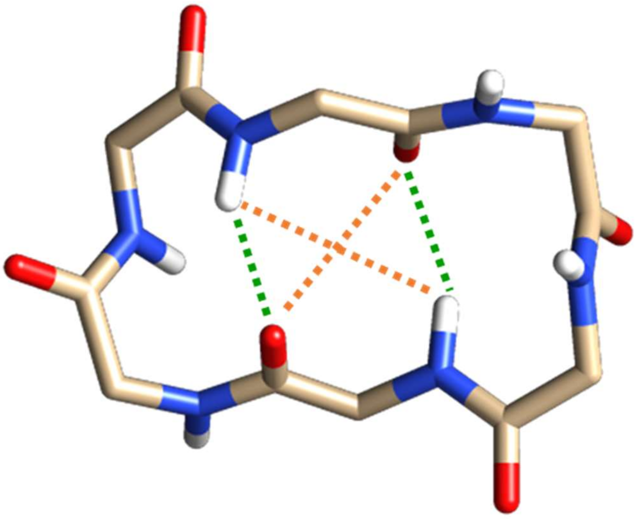


Figure 46: C=O of the i/i+3 residues point above and below the plane of the ring. CO/CO and NH/NH repulsive interactions shown as orange dashed lines, the CO/NH attractive interactions are shown as green dashed lines.

As the glycines were substituted with increasing numbers of alanine type I' β-turns became less favourable and type I and II became more favoured.[245] This is due to the required dihedral angles necessary for type I and II turns being more favourably occupied by L-amino acids than the type I' and II' turns which require dihedral angles in sparsely occupied regions of the Ramachandran plots (Figure 47). Table 16 shows the many different conformations seen in the major clusters of the BE-META of the alanine-containing sequences, demonstrating the complexities of cyclic peptide conformation prediction. Multiple conformations are seen for each sequence with the same sequences able to form multiple types of β-turn. Some sequences are more likely to form certain turn types than others, but Ramachandran preferences are not always the determining factor for turn type. For example although dipeptide subsequence AA commonly forms a type I turn, the two most populated clusters seen in the simulation of c(AAAAAA) are a II+ II conformation (22.5%) and I+II' (20.5%). The Ramachandran distributions of individual amino acids therefore do not directly dictate the overall structure of the cyclic peptide or the conformation seen for this sequence would be I+I.

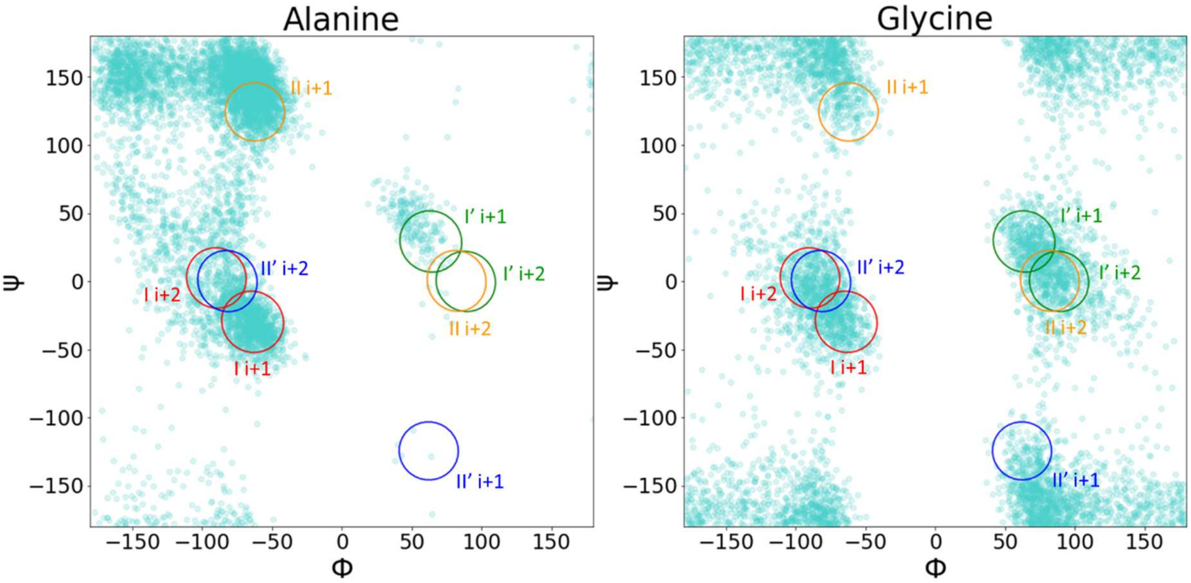


Figure 47: Differences in the Ramachandran plots of L-amino acids and achiral glycine and the i+1 and i+2 position for type I, II, I' and II' β-turns. Ramachandran plots obtained from the Loop Database.

A similar wide range of conformations was seen in the equivalent sequences where the glycines were substituted by valine, with even small changes to the sequence able to vastly change the conformations seen (Table 17). The tendency of cyclic peptides to adopt multiple conformations in solution is seen in the simulations on c(GGGGGG) with alanine or valine substitutions where the major conformations population is rarely more than 30%. With the valine substitutions, the cyclic peptide structure was now not always made up of a combination of two type I, II, I' or II' β-turns. Type IV β-turns or even γ turns were now seen. Type IV β-turns are a miscellaneous category for β-turns which do not fit into any of the other traditional categories. The type IV turns seen for the valine-containing sequences do not always show a hydrogen-bond between the i and i+3 residues but still retain a distance of less than 7 Å between the α-carbons. The type IV and γ turns were most commonly seen in sequences with three valines in a row, likely due to steric hinderance.

Sequence	Population (%)	Sequence	Population (%)	Sequence	Population (%)
GGGGGG	32.5	GGGGGG	20.8	GGGGGG	19.2
AGGGGG	20.9	AGGGGG	10.7	AGGGGG	8.7
AAGGGG	16.8	AAGGGG	15.5	AAGGGG	12.3
AGAGGG	15.1	AGAGGG	9.3	AGAGGG	8.2
AGGAGG	31.6	AGGAGG	20.8	AGGAGG	9.8
AAAGGG	16.5	AAAGGG	12.6	AAAGGG	8.0
AAGAGG	25.0	AAGAGG	11.6	AAGAGG	5.8
AAGGAG	19.6	AAGGAG	10.7	AAGGAG	8.0
AGAGAG	41.4	AGAGAG	20.8	AGAGAG	5.4
AAAAGG	21.3	AAAAGG	16.6	AAAAGG	11.9
AAAGAG	26.6	AAAGAG	17.5	AAAGAG	8.4
AAGAAG	28.4	AAGAAG	7.2	AAGAAG	3.5
AAAAAG	19.0	AAAAAG	13.5	AAAAAG	13.3
AAAAAA	22.5	AAAAAA	20.5	AAAAAA	15.4

Table 16: The three largest clusters seen in the BE-META simulations carried out by McHugh et al. on alanine-containing cyclic peptides. Each sequence shows cyclic peptides can form many conformations. The i+1 and i+2 positions of each β-turn is coloured. Red is a type I turn, green is type II, orange type I' and blue is type II'. Table based on [245].

The c(GGGGGG) simulation shows how the backbone structure of the peptide influences the lowest energy structures. The further simulations with alanine or valine being included in the cyclic peptide show how the presence of chiral amino acids can change the lowest energy conformations, with small changes in the cyclic peptide sequence potentially leading to very different structures.

Sequence	Population (%)	Sequence	Population (%)	Sequence	Population (%)
GGGGGG	32.5	GGGGGG	20.8	GGGGGG	19.2
VGGGGG	38.5	VGGGGG	9.8	VGGGGG	7.8
VVGGGG	18.7	VVGGGG	10.0	VVGGGG	8.8
VGVGGG	14.7	VGVGGG	12.2	VGVGGG	8.3
VGGVGG	45.6	VGGVGG	9.2	VGGVGG	7.3
VVVGGG	29.7	VVVGGG	15.5	VVVGGG	10.2
VVGVGG	37.2	VVGVGG	21.9	VVGVGG	5.0
VVGGVG	81.8	VVGGVG	1.3	VVGGVG	0.3
VGVGVG	36.0	VGVGVG	10.5	VGVGVG	5.2
VVVVGG	47.3	VVVVGG	33.2	VVVVGG	0.7
VVVGVG	47.6	VVVGVG	7.9	VVVGVG	6.2
VVGVVG	34.5	VVGVVG	8.7	VVGVVG	7.9
VVVVVG	18.7	VVVVVG	14.5	VVVVVG	14.1
VVVVVV	22.3	VVVVVV	17.6	VVVVVV	11.7

Table 17: The three highest clusters obtained from the BE-META performed by McHugh et al. for valine substitutes sequences. The i+1 and i+2 positions of each β-turn are coloured. Red is a type I turn, green is type II, orange type I' and blue is type II'. Table based on [245].

The BE-META of these cyclic hexapeptide sequences demonstrate the prevalence of conformations made up of two β-turns, especially those made up of type I, II, I' and II' turns. Only the valine-containing sequences contain other turn types, with 23 out of the 84 possible turns across the three major clusters for each sequence containing at least one type IV or γ-turn. All but one of the sequences contains at least one type I, II, I' or II' turn. These other turn types usually occur at the VVV subsequence with the valine-containing sequences otherwise showing similarity to those containing alanine with the same conformations generally seen in different proportions. It is therefore likely type I, II, I' and II' turns will still be predominant in sequences without multiple bulky sidechains in a row.

Not all turn type combinations were seen in the simulations carried out by McHugh et. al. Some β-turn type combinations lead to an overall more stable structure and so are more likely to occur whereas the least favourable combinations are not seen in the simulations. Certain turn combinations being lower energy structures contributes to the structure formed by the peptide along with the propensity for a given turn type shown by the amino acid sequence. This would explain why, for example, the c(AAAAAA) sequences' major conformation is II+II despite the AA subsequence more favourably forming a type I turn. The II+II conformation is likely a lower energy turn combination than a I+I conformation. This is certainly the case for c(GGGGGG) as seen in the BE-META simulation but due to the flexibility of glycine the I+I conformation with glycine occupying the relatively constrained i/i+3 positions is not necessarily the same as the I+I conformations with chiral amino acids in those positions. When only a few alanine or valine residues are present in the cyclic peptide sequence, they do not occupy the i/i+3 positions of the β-turns making up the peptide structure and often occupy the i+1 positions. The effect on the most stable conformation when chiral amino acids occupy the i/i+3 positions are therefore unknown without chiral amino acids also occupying i+1 and/or i+2 positions. In such cases it is difficult to know whether the turn combination is the most favourable or if the chiral amino acids preference for forming type I (or II) turns is leading to the conformation seen.

4.2 Restrained Simulations Theory

In order to determine the most stable turn-type combinations restrained simulations were designed. BE-META used for determination of cyclic peptide structure was modified by the inclusion of restraints at one end of the cyclic peptide to incorporate a given β-turn type. The most compatible turns with the restrained turn type occur at the unrestrained side of the cyclic peptide with lower energy conformations occurring more frequently.

Two sets of collective variables (CVs) are typically used in the BE-META of cyclic peptides.[244] A bias on the ϕ and ψ dihedral angles for each amino acid in the sequence is associated with changes in β-turn type seen within the structure. Replicas with a bias on the ψ dihedral of residue i and ϕ dihedral of residue i+1 are also included and are linked to changes of the amino acid register within the sequence. By removing the biased replicas associated with changes in amino acid register and by constraining one end of a cyclic hexapeptide to a certain turn type, each amino acid will be kept within the same positions within the β-turns making up the structure of the peptide. A biased replica for the remaining unrestrained four residues of the peptide (Figure 48) can then be used to explore the conformations the peptide forms when a given β-turn is present.

The i+1 and i+2 positions at one end of the cyclic peptide were restrained to the ideal dihedral angles for a particular β-turn type using a harmonic force constant of 50 kJ/mol/rad^2 (see section 10.5.3.2). To be assigned to a β-turn type typically dihedral angles for the i+1 and i+2 positions should be within 30° of the ideal angles for three of the dihedrals with the fourth able to vary by 45°. Around 99% of the dihedrals in the restrained turn fit within this definition with some flexibility for exploration of the system.

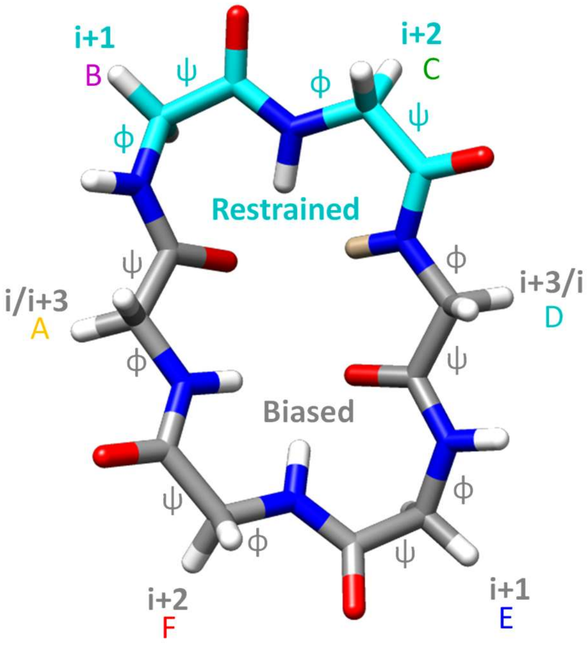


Figure 48: For the restrained BE-META simulations the i+1 and i+2 residues at one end of the β-turn (shown in light blue) have restrained φ and ψ dihedral angles, the rest of the cyclic peptide has bias acting on the remaining dihedrals to allow full exploration of conformations compatible with the restrained turn.

The hydrogen bond between the i and i+3 residues of the restrained turn is present throughout the simulations showing the restraints are sufficient to maintain a β-turn within the cyclic hexapeptide structure. To assign the unrestrained turn-types the results of the BE-META were clustered and a cut-off value was used whereby at least 70% of the cluster had to fit within the definition of a turn type for it to be classified as such. This means a proportion of some clusters would not be classified as a given turn type based on the traditional β-turn definitions despite forming part of the same cluster. In each case however the cluster centre remains within the traditional definition of a β-turn with the remaining cluster points distributed in a continuum. This is a limitation of the traditionally used definition of β-turn types as these points should not be treated as a separate turn type as there is no large energy difference or sequence variation between them.

For the restrained simulations the following naming system was used: the six amino acids are listed starting from the i position of the restrained turn (this is the i+3 position of the unrestrained turn) followed by the i+1 and i+2 positions of the restrained turn continuing around the peptide until the i+2 position of the unrestrained turn. So, as seen in Figure 48, for a peptide c(ABCDEF) BC would be at the i+1 and i+2 positions of the restrained turn, EF would be at the i+1 and i+2 positions of the unrestrained turn and A and D would be at each of the i/i+3 positions. Turn type combinations for the restrained simulations are written first with the β-turn that is restrained followed by the turn-type that forms at the unrestrained side of the cyclic peptide. For example in a restrained simulation with the restrained turn set to a type I turn, if a type II' turn is observed at the unrestrained side the turn combination is written as I+II'. If the restrained turn is type II' however, and a type I turn is observed the turn combination is written as II'+I. These two conformations will be equivalent in the restrained simulations if the same amino acids occupy both of the i/i+3 positions and the same Ramachandran distribution is seen.

4.3 Poly-Glycine Simulations

The restrained simulations were first carried out on c(GGGGGG) to determine the lowest energy structures a cyclic hexapeptide can adopt in the absence of sidechain specific effects. Four restrained simulations were carried out with the dihedral angles in the restrained turn restrained to the ideal angles in a type I, II, I’ and II’ β-turn respectively in each of the simulations.

As glycine is achiral types I and I’ will give similar populations as will type II and II’. This is seen in the simulations (Table 18). No type I+I or type I’+I’ is seen, which is consistent with what is observed in cyclic hexapeptides with known structures where these conformations are rare.[38] Type I+I’ makes up the majority of the population for the type I/I’ restrained systems followed by I+II’/I’+II. For the II/II’ restrained systems the major conformation is the I+II’/I’+II followed by the type II + II’ conformation which makes up around 20% of the population. A small cluster is seen in the type II simulation of a I+II conformation. This would be equivalent to a I’+II’ conformation. This gives a ranking of the most stable turn-type combinations of I+I’ > I’+II = I+II’ > II’+II > II’+I’ = II+I. The I+I, I’+I’, II+II, II’+II’ combinations are not seen in any of the restrained simulations. They are therefore higher energy conformations than the other conformations.

Restrained Turn Type	Unrestrained Turn	% Observed in Restrained Simulations
I	I	-
	II	-
	I’	57
	II’	35
II	I	0.33
	II	-
	I’	66
	II’	13
I’	I	50
	II	43
	I’	-
	II’	-
II’	I	65
	II	20
	I’	-
	II’	-

Table 18: The clusters seen in BE-META of c(GGGGGG) when the i+1 and i+2 positions at one end of the cyclic hexapeptide are restrained to a particular turn type.

The II+I conformation is equivalent to the II’+I’ simulation as glycine is achiral. Only a very small II+I cluster is seen in the type II simulation but a II’+I’ cluster is not seen in the type II’ simulation. Noise in the simulation means that although the II+I conformation appears enough to form a cluster in the type II simulation, the equivalent II’+I’ conformation does not appear in the type II’ simulation.

The restrained turn was restrained to the ideal dihedral angles used to define β-turn types. These are idealised values rounded to the nearest 10 °. It is therefore unlikely the actual highest density of ϕ/ψ dihedral angles would occur at these values. The harmonic potential used allows for some variation around the ideal value, but slightly different results may be seen with the ϕ and ψ dihedral angles based on cyclic peptide structures. Analysing the β-turn types in the Loop Database (see Chapter 3) shows average ϕ and ψ values slightly different from the ideal values. Table 19 shows the average dihedral angles seen for each turn type in the β-hairpin dataset and the β-turn dataset obtained from the database.

β-turn type	Dataset	Φ_{i+1}	Ψ_{i+1}	Φ_{i+2}	Ψ_{i+2}
I	Ideal values	-60	-30	-90	0
	β-hairpin	-61	-25	-96	0
	β-turns	-62	-24	-85	-5
II	Ideal values	-60	120	80	0
	β-hairpin	-54	120	71	11
	β-turns	-56	132	83	-1
I’	Ideal values	60	30	90	0
	β-hairpin	52	43	78	4
	β-turns	56	34	72	14
II’	Ideal values	60	-120	-80	0
	β-hairpin	60	-123	-94	6
	β-turns	56	-130	-87	3

Table 19: Ideal β-turn dihedral angles compared to the average values seen in the β-hairpins and other β-turns extracted from the Loop Database.

The average dihedral angles seen in the database can vary by up to 20° from the ideal angles for β-turn types, with the largest differences seen for the type I’ i+2 dihedral angles. The context in which the β-turn forms is likely to alter the preferred dihedral angles with the dihedral angles seen between the hairpin dataset and the β-turns which are not found in a hairpin varying by around 10° in some instances. The hairpin dataset contains predominately type I’ and II’ turns with relatively few type I and II turns whereas the reverse is seen for the other dataset so a larger data sample may lead to more similar results. In the restrained cyclic peptide structure different average dihedral angle values may be found potentially altering the results of the restrained simulations.

The angular RMSD around the ideal dihedral angles for a given β-turn type differs between the restrained and unrestrained turn. Whereas the restrained turn has an RSMD of approximately 11° around the ideal dihedral angles the turn is restrained to, the unrestrained turn gives RMSD values of around 20° around the ideal dihedral angles of the β-turn type the turn is assigned to. This may be because the ideal dihedral angles for a given turn type do not reflect the actual lowest energy values as seen by the variation in the average dihedral angle values compared to the ideal values seen in the hairpin and β-turn datasets. The unrestrained β-turn is centred around slightly different dihedral angles and more variation is seen in the dihedrals.

Regardless of small differences in the average ϕ and ψ dihedral angle values and their distributions between the restrained and unrestrained turn, the restrained simulations still show which turn types are most compatible. The β-turns seen at the restrained and unrestrained turns are similar enough that the same conformation with the same Ramachandran distribution at all 6 positions is seen when a turn type combination occurs in different restrained simulations. For example, if a type II’ turn is observed at the unrestrained end of the peptide in the type II simulation (II+II’ conformation) the same structure is seen in the simulation where the restrained turn is type II’ but a type II turn is seen at the unrestrained side (II’+II conformation, Figure 49). The two can therefore be treated as equivalent with the small differences discarded as the overall structure is very similar.

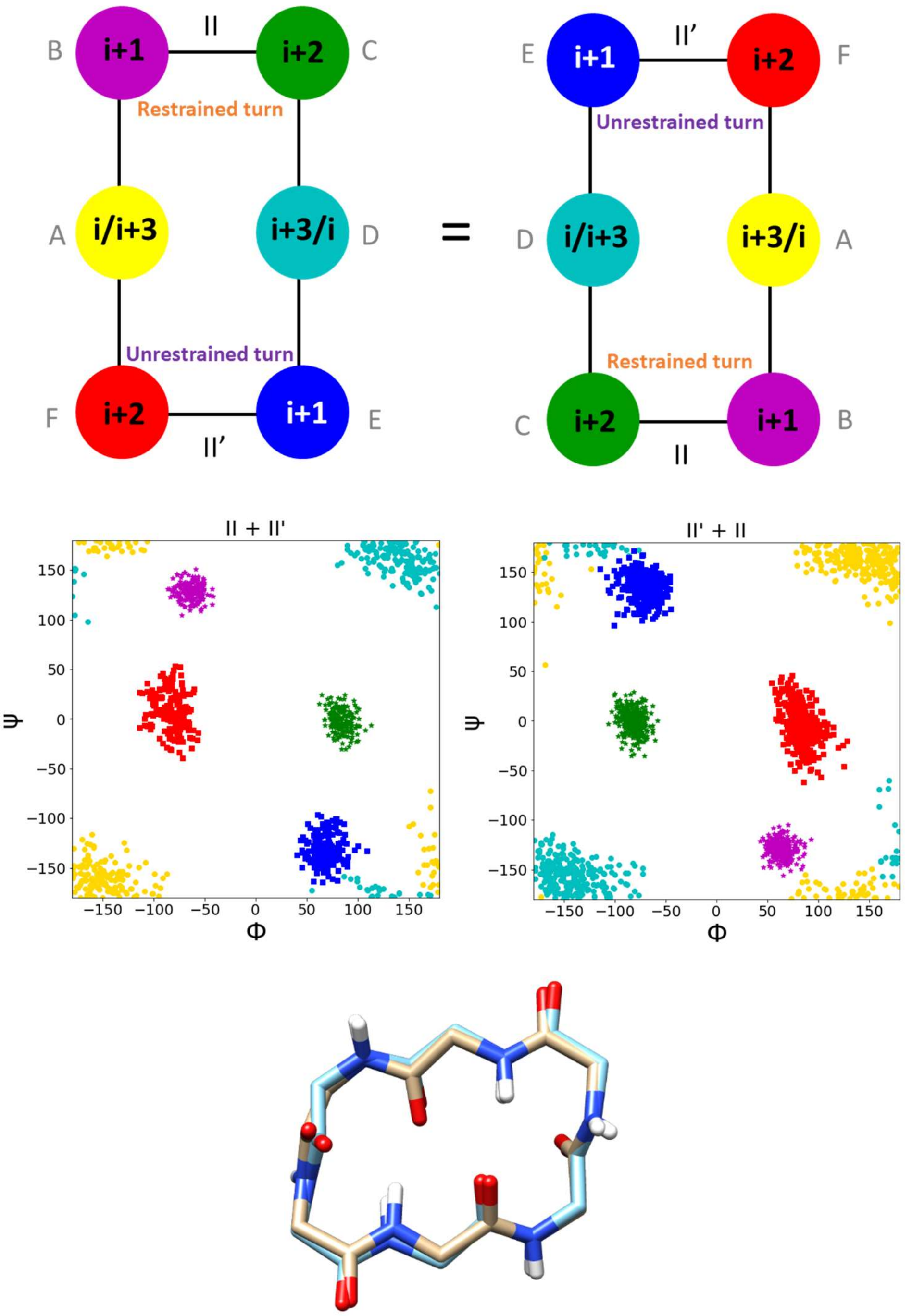


Figure 49: The II+II’ conformation from the type II restrained simulation is equivalent to the II’+II conformation in the type II’ restrained simulation as shown by the similar Ramachandran plots produced form the restrained simulations trajectories.

For all turn type combinations, the Ramachandran positions occupied by the i/i+3 positions are similar across the different c(GGGGGG) restrained simulations. If the restrained turn is a type I or II then the i position of the restrained turn (i+3 of the unrestrained turn) occupies the yellow part of the Ramachandran plot in Figure 50. The i+3 position of the restrained turn has ϕ and ψ dihedral angles which occupy the cyan region in Figure 50. The reverse is seen when the restrained turn type is I’ or II’ with the i position of the restrained turn occupying the blue region and the i+3 position occupying the yellow region. Generally the type I or II turns are always seen with either a I’ or II’ turn and vice versa. The exception is the small II+I cluster. This cluster has a similar Ramachandran distribution to the other conformations, with the i position of the type II restrained turn occupying the cyan region and the i+3 position occupying the yellow region of the Ramachandran plot.

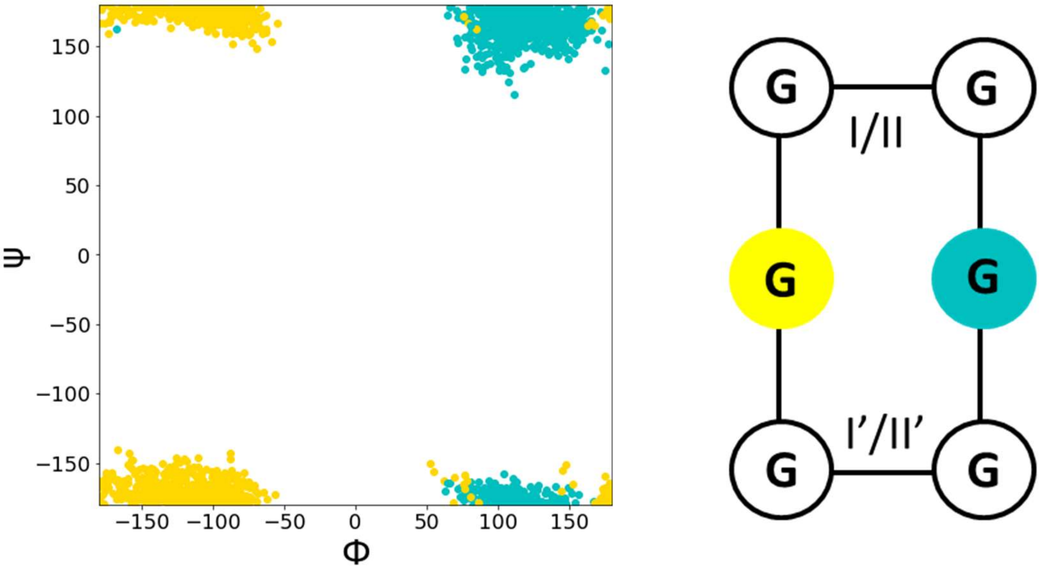


Figure 50: The Ramachandran distributions of the i/i+3 positions in the restrained simulations of c(GGGGGG).

As the size of each cluster seen in the restrained simulations is related to the free energy of the conformation, with lower energy conformations appearing more in the simulation, the Boltzmann equation can be used to estimate the ΔG value between the conformations seen in a simulation. The ΔG value between the conformations seen in each of the restrained simulations was therefore calculated using the ratio of the occurrence of each conformation relative to the largest cluster. Using equivalent clusters with the same Ramachandran distributions which appear in different restrained simulations it is possible to estimate the ΔG values between conformations seen in different restrained simulations. For example, the type I restrained simulation shows two clusters: I+I’ and I+II’. The I+II’ conformation is equivalent to the II’+I conformation seen in the type II’ restrained simulation. A type II’+II conformation is also seen in the type II’ restrained simulation. The energy difference between the I+I’ and II’+II conformation can therefore be determined despite occurring in different restrained simulations as the ΔG value between both of them and the I+II’/II’+I conformation is known. The energy differences between the clusters seen across multiple restrained simulations can therefore be estimated and plotted. More variation in the unrestrained dihedral angles then the restrained dihedral angles means two clusters which appear in multiple restrained simulations are not exactly the same but were treated as equivalent due to the overall similarity of the structures. The relative energies for all the conformations seen in the restrained simulations of c(GGGGGG) are shown in Figure 51.

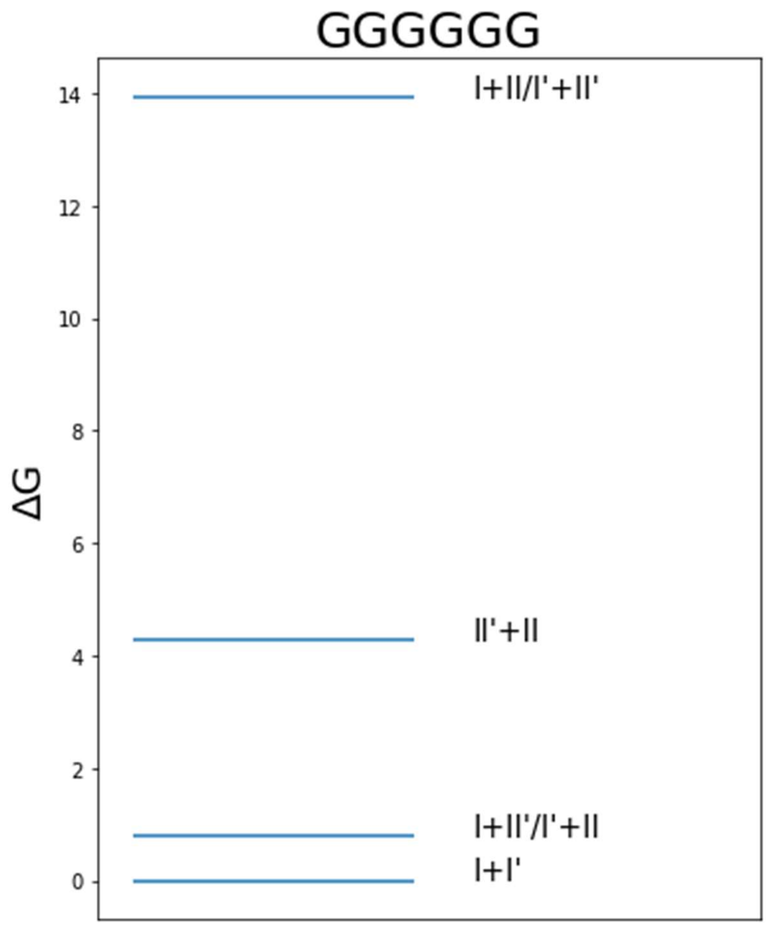


Figure 51: The relative energies of the conformations seen in the restrained simulations of c(GGGGGG).

The three most populated clusters in the unrestrained BE-META of c(GGGGGG) carried out McHugh *et al*. are I+I’ followed by I’+II and I+II’ (Table 20).[245] The I’+II and I+II’ conformations are equivalent appearing in similar quantities, likely only differing due to the noise in the simulation. The Boltzmann equation can be used to estimate difference in energy between these clusters. This gives a value of approximately 1 kJ/mol which is similar to the value calculated by the restrained simulations. The restrained simulations therefore can be used to predict the structure of c(GGGGGG). Additionally they allow the detection of conformations that do not appear if unrestrained BE-META is used – more of the energy landscape of the system is determined. This could potentially help design stable cyclic peptides. For example if designing a cyclic peptide containing a type II turn the II+I’ conformation is the lowest energy turn type combination but the restrained simulations show that a II+II’ conformation would be the next best option if a compatible I’ turn cannot be found.

Conformation	Population (%)
I + I’	32.5
I’ + II	20.8
I + II’	19.2

Table 20: The lowest energy conformations seen in the unrestrained BE-META of c(GGGGGG) carried out by McHugh et al.[245]

The restrained simulations on c(GGGGGG) show the lowest energy cyclic peptide conformations when a given β-turn type is at one side of the cyclic peptide and glycine is at the i/i+3 positions. If chiral amino acids are included at the i/i+3 positions different favourable combinations of β turns may be seen. Therefore the restrained simulations were initially repeated with a chiral amino acid at either the i or i+3 position of the restrained turn.

4.4 Restrained BE-META on c(XGGGGG) and c(GGGXGG)

Simulations on c(XGGGGG) and c(GGGXGG) were carried out where X occupies the i or i+3 position of the restrained turn respectively and X = A, Q, V or F. These restrained simulations identify compatible turn-types when there is only one chiral amino acid at either the i or i+3 position with the different amino acids chosen to determine if the same conformations would be seen. A, Q, V and F represent a range of amino acid sidechain characteristics: alanine is relatively small, valine has a large β-branched sidechain, glutamine is polar and has the potential to hydrogen bond to other sidechains or the peptide backbone, and phenylalanine is aromatic. There are 16 possible combinations of I, II, I’ and II’ turns depending on whether the chiral amino acid is at the i or i+3 position of the restrained β-turn (Table 21). The results of the simulations are shown in Table 22.

		Glycine			
		I	II	I’	II’
Chiral Amino Acid	I	I + I	I + II	I + I’	I + II’
	II	II + I	II + II	II + I’	II + II’
	I’	I’ + I	I’ + II	I’ + I’	I’ + II’
	II’	II’ + I	II’ + II	II’ + I’	II’ + II’

Table 21: 16 possible combinations of I, II, I’ and II’ β*-turns for c(XGGGGG) or c(GGGXGG).*

The inclusion of a chiral amino acid at a i/i+3 position, in some instances alters the most favourable conformations, with different turn-type combinations seen as the lowest energy structure than seen in the c(GGGGGG) simulations. There is no longer a degeneracy in multiple turn-type combinations. For example the I+II’ and I’+II combinations (which are very stable in the c(GGGGGG) simulations) no longer appear in similar amounts, with the I’+II combination no longer seen in any of the c(XGGGGG) simulations.

When the chiral amino acid is at the i position of the restrained turn (c(XGGGGG) simulations) and the restrained turn type is a type I or II turn then the Ramachandran plot of the resulting conformation is similar to that seen for the c(GGGGGG) simulations. Similarly when the chiral amino acid is at the i+3 position of the restrained turn (c(GGGXGG simulations) and the restrained turn type is a type I’ or II’ turn the results of these simulations are also similar to the c(GGGGGG) simulations. The clusters in these simulations are equivalent and show the Glycine appearing in the righthand side of the Ramachandran plot (Figure 52). When the chiral amino acid is at the i position of a type I’ or II’ restrained turn/the i+3 position of a type I or II turn, then the chiral amino acid would have to occupy this position on the top righthand side of the Ramachandran plot to produce a cluster with the same conformation as seen in the c(GGGGGG) simulations. This is more unfavourable for L-amino acids than glycine so different clusters are now seen. In most instances both the i/i+3 positions dihedral angles are now found in the top lefthand corner of the Ramachandran plot. Although the same turn type combination may be seen, due to the different i/i+3 dihedral angles, a different backbone conformation is therefore seen compared to c(GGGGGG).

		% Observed in Restrained Simulations							
Restrained turn	Unrestrained turn	AGGGGG	GGGAGG	FGGGGG	GGGFGG	QGGGGG	GGGQGG	VGGGGG	GGGVGG
I	I	-	-	-	1	-	-	-	0.3
	II	-	-	6	17	-	10	-	33
	I’	49	-	50	-	49	9	47	-
	II’	47	74	35	65	46	44	43	51
	IV	-	-	-	-	-	1	-	-
II	I	-	15	7	22	-	8	-	-
	II	1	12	44	40	-	13	25	40
	I’	51	2	31	-	61	2	52	-
	II’	16	57	5	18	10	42	1	26
	IV	10	7	-	-	0.3	21	6	14
I’	I	-	32	-	29	-	33	-	10
	II	-	36	-	32	-	28	-	48
	I’	21	2	19	5	30	1	38	1
	II’	4	8	-	18	-	3	-	12
	IV	62	-	58	1	50	4	43	2
II’	I	-	63	-	66	-	67	-	67
	II	11	18	17	10	14	16	15	17
	I’	5	-	10	-	16	-	15	-
	II’	-	1	-	0.2	-	-	-	-
	IV	59	-	53	-	48	-	50	-

Table 22: c(XGGGGG) and c(GGGXGG) constrained simulations.

The clusters that resemble those seen in the c(GGGGGG) simulations occur in similar quantities to those in the c(GGGGGG) simulations. For example the type I restrained simulations of c(AGGGGG) and c(GGGGGG) each show the same I+I’ and I+II’ clusters in similar proportions to each other. It is however more common for smaller clusters to also occur in addition to those seen in the c(GGGGGG) simulations. Across the c(XGGGGG) and c(GGGXGG) simulations all sixteen possible conformations made up of combinations of type I, II, I’ and II’ β-turns were seen in at least one of the restrained simulations. It is possible there are multiple possible i/i+3 dihedral angle distributions which could be used to combine a given turn type combination. However the same Ramachandran distributions are seen for the same turn type combinations when different chiral amino acids are used. The i/i+3 dihedral angles in the restrained simulations represent the lowest energy conformation to join the two turn types. Unlike in the c(GGGGGG) simulations, turn types other than the I, II, I’ and II’ turns are seen in some of the c(XGGGGG)/c(GGGXGG) simulations. They are type IV turns which is a broad class which encompasses all β-turns which don’t fit any of the other classifications. The c(XGGGGG)/c(GGGXGG) simulations show more conformations than seen in the c(GGGGGG) simulations possibly due to competing factors of minimising the backbone energy and the sidechain interactions of the chiral amino acid leading to small differences in energy between similar structures.

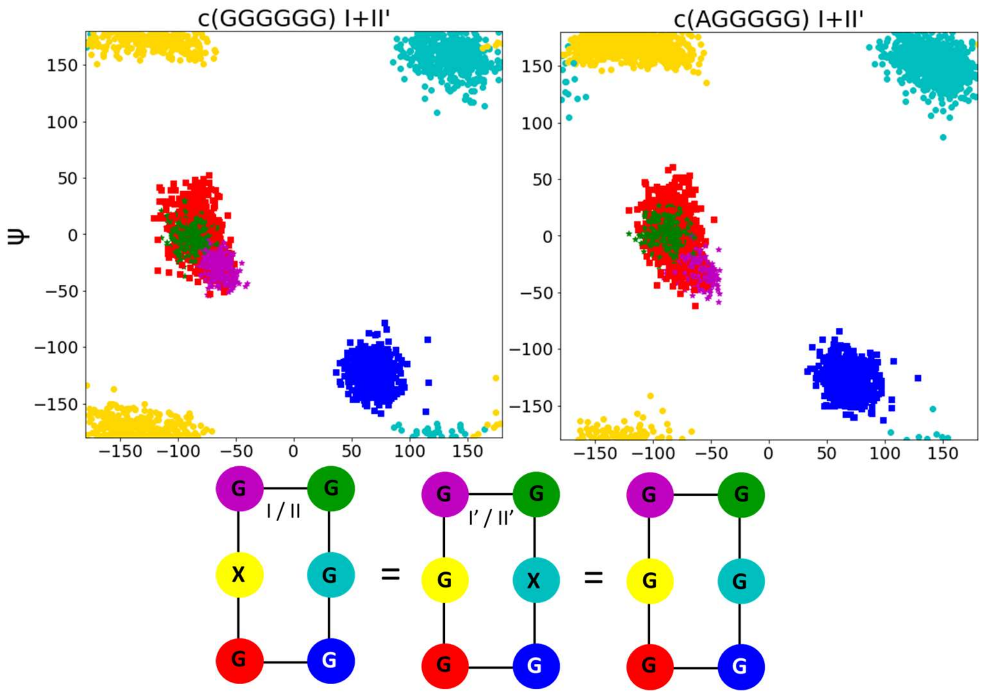


Figure 52: When a type I or II restrained turn is used in the c(XGGGGG) simulations or a type I' or II' turn is used in the c(GGGXGG) simulations, the same conformations as seen in the c(GGGGGG) simulations occur. This is seen in the similar Ramachandran plots produced for each conformation.

Where a turn combination is seen in both the c(XGGGGG) and c(GGGXGG) simulations for a particular amino acid, the conformation from each simulation is treated as equivalent as they have very similar Ramachandran distributions. For example the I+I' combination where the chiral amino acid is at the i position of the type I turn is seen in both the c(XGGGGG) type I simulation as well as the c(GGGXGG) type I' simulation. As such all the energy difference between conformations seen across the different restrained simulations can be estimated. The relative energies of the conformations seen in the restrained simulations for the c(XGGGGG) simulations are shown in Figure 53.

When designing a cyclic peptide with glycine occupying only one of the i/i+3 positions choosing a sequence likely to form a I+I', I+II' or II+I' turn combination may lead to a more stable conformation as these are consistently found in the lowest energy conformations across all the sequences. The choice of amino acid can lead to small differences in energy between conformations. A different choice of conformation may therefore be needed when designing a cyclic peptide if a β-branched amino acid such as valine is needed at one of the i/i+3 positions. The lowest energy conformations are generally those with structures which match the conformations seen in the c(GGGGGG) simulations. In these conformations the L-amino acid can occupy one of the i/i+3 positions without having to alter the backbone conformation from that seen in the simulations of c(GGGGGG). As these conformations generally have the lowest energy backbone conformation it can be inferred that having at least one glycine at the i/i+3 position could potentially lead to more stable conformations.

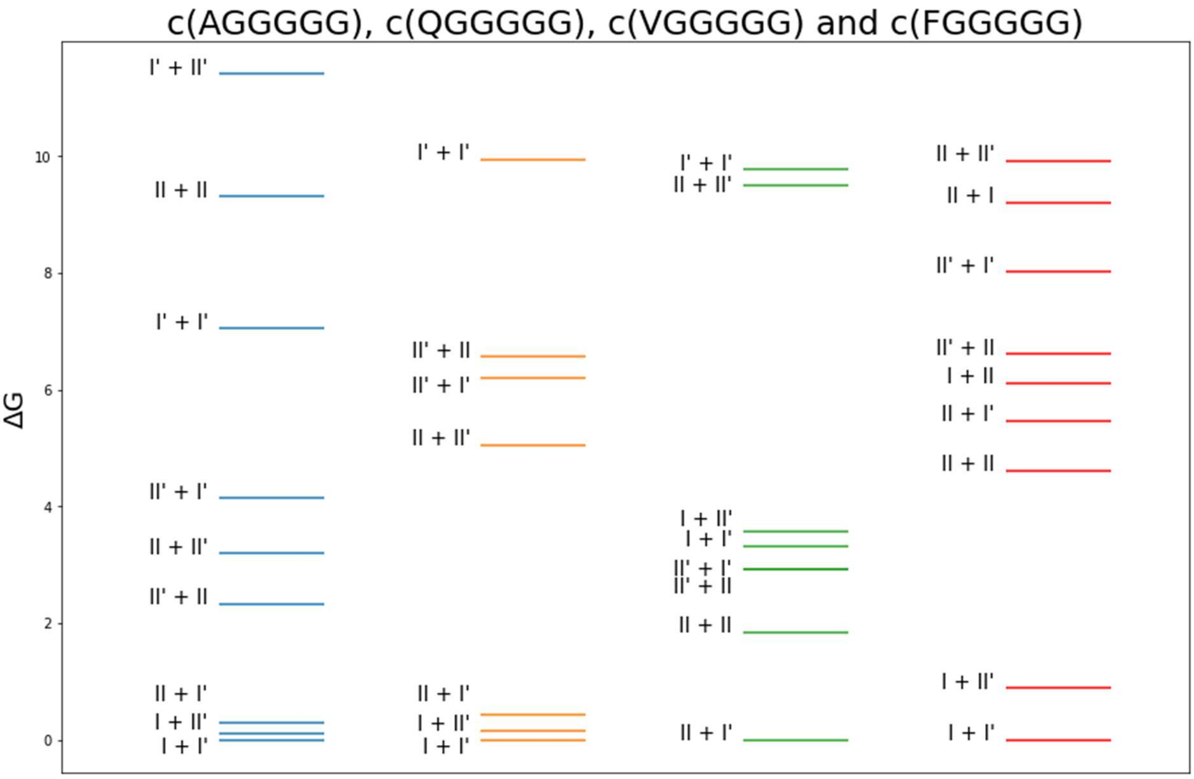


Figure 53: The relative energy differences between the clusters seen in the restrained simulations of c(AGGGGG) (blue), c(QGGGGG) (orange), c(VGGGGG) (green) and c(FGGGGG) (red).

The restrained simulations on c(XGGGGG) and c(GGGXGG) show how the presence of a chiral amino acid at one of the i/i+3 positions can alter the lowest energy conformations of a cyclic peptide. To determine the effect on the most favourable conformations when both i/i+3 positions are occupied by chiral amino acids further restrained simulations were carried out.

4.5 Restrained BE-META on c(XGGXGG)

The restrained simulations were repeated on c(XGGXGG), where X occupies the i and i+3 positions and X = A, F, Q or V. This gives the most stable conformations for a cyclic hexapeptide to adopt when chiral amino acids are at both the i/i+3 positions of the structure made up of two overlapping β-turns. Glycine residues are required at the i+1 and i+2 positions as, due to their flexible nature they can form all β-turn types. When, for example, the simulations were carried out on c(AAAAAA) predominately type I turns were seen at the unrestrained side of the cyclic peptide, not because type I is most compatible with the restrained turn type, but because a β-turn composed of alanines forms a type I turn. The AA turn sequence was altering the conformation of the peptide away from the lowest energy structure based on the backbone conformation. Such sidechain interactions often alter the turn type that forms e.g. steric interactions between the β-carbon of L-amino acids at the i+2 position with the carbonyl group of the peptide bond of the i+1 residue mean glycine is often found at the i+2 position of type II turns.[299] The results of the c(XGGXGG) restrained simulations are shown in Table 23.

		% Observed in Restrained Simulations			
Restrained turn	Unrestrained turn	AGGAGG	FGGFGG	QGGQGG	VGGVGG
I	I	-	-	-	-
	II	5	-	-	8
	I’	-	-	-	-
	II’	74	84	79	65
	IV	9	-	-	10
II	I	3	-	-	-
	II	9	-	-	-
	I’	-	-	-	-
	II’	73	63	45	61
	IV	2	14	44	23
I’	I	-	-	-	-
	II	-	-	-	-
	I’	49	63	68	81
	II’	26	14	3	-
	IV	-		-	-
II’	I	-	-	-	-
	II	31	13	21	17
	I’	12	17	13	20
	II’	23	47	44	52
	IV	27	14	11	-

Table 23: The clusters seen in the restrained simulations of c(XGGXGG) where X=A, Q, V or F.

Figure 54 shows the energy diagram of the relative energies of the clusters for the c(AGGAGG), c(QGGQGG) and c(FGGFGG) simulations. For c(VGGVGG) there were not enough conformations which are the same as each other in the different restrained simulations to determine the ΔG value between conformations seen in different simulations. The c(VGGVGG) conformations are therefore not included in the diagram. Although the relative energies of the conformations in different restrained simulations can be estimated when the sequence is the same, the relative energies of clusters from different sequences is unknown – e.g. the I’+I’ conformation for c(AGGAGG) could be higher or lower energy than c(QGGQGG). The energy due to the backbone conformation will be the same if the same conformation is seen but there may be differences caused by the sidechains.

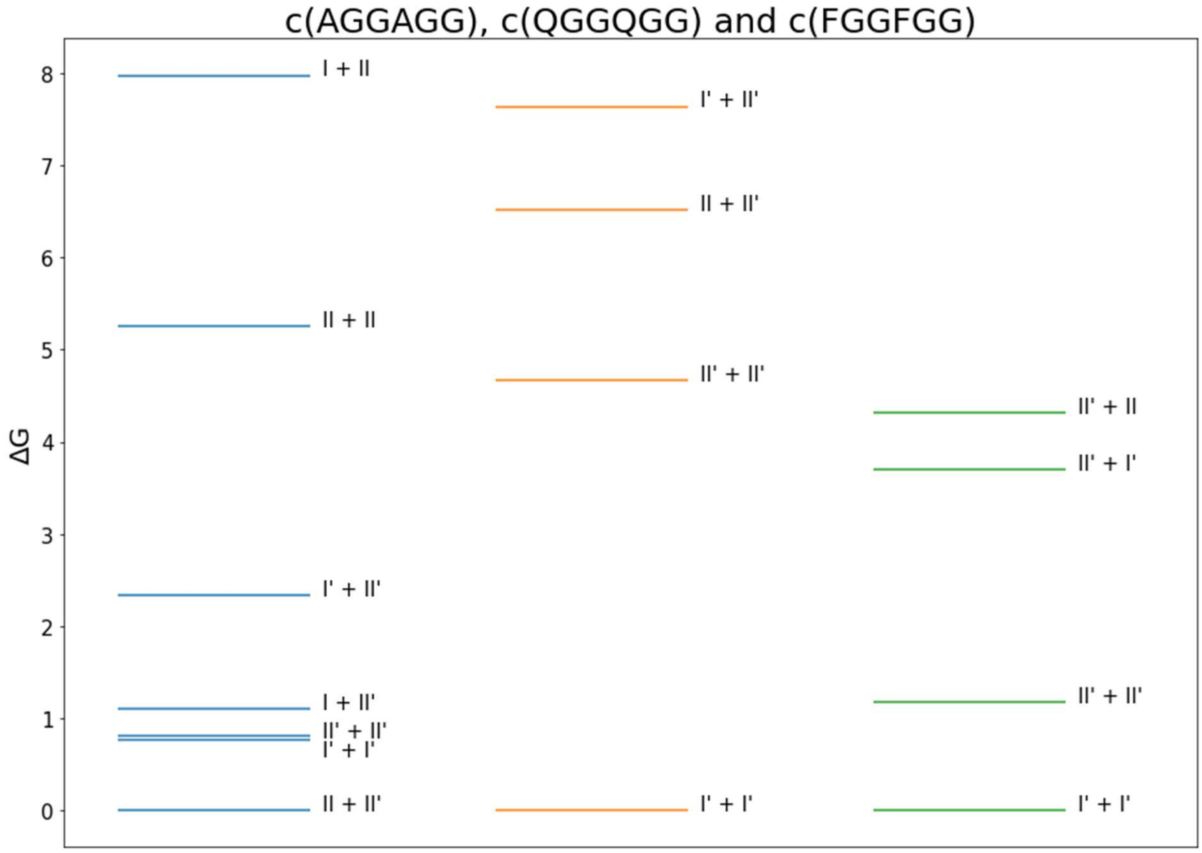


Figure 54: The relative energies between the conformations seen in the restrained simulations of c(AGGAGG) (blue), c(QGGQGG) (orange) and c(FGGFGG) (green). The I+II' conformations of c(QGGQGG) and c(FGGFGG) are not shown as the energy difference could not be estimated.

4.5.1 Differences Compared to the c(GGGGGG) Simulations

Similar low energy clusters are seen in the simulations with the different amino acids with the I'+I', II'+II', I'+II' and II+II' all generally forming the lowest energy clusters. These are different low energy conformations compared to the c(GGGGGG) and c(XGGGGG) simulations where the I+I', I'+II and I+II' conformations were generally the lowest energy. For a given turn type combination by definition the two i+1 and i+2 positions must remain the same as these define the β-turn types. Differences in the Ramachandran distributions of the i/i+3 positions can however be seen for the same turn type combinations. These differences in the Ramachandran distributions of the i/i+3 positions may be one of the reasons different conformations are seen compared to the c(GGGGGG) simulations. For c(GGGGGG) one of the i/i+3 positions often occupies Ramachandran positions that are particularly unfavourable for L-amino acids. For c(XGGXGG), for the majority of conformations, both the i/i+3 ϕ and ψ dihedral angles occur in the β-region of the Ramachandran plot. Inclusion of chiral amino acids at the i/i+3 positions therefore alters the structure of the peptide to a slightly different conformation than when glycine is present at the i/i+3 positions, even if the same turn-type combination occurs.

More variation in the i/i+3 dihedral angles of the c(XGGXGG) sequences is seen than in the c(GGGGGG) simulations. Depending on which turn combination occurs in the restrained simulations, distinct i/i+3 dihedral angles are seen. Although each conformation has unique favoured i/i+3 positions they generally fit into one of three categories shown in Figure 55. The I+II and II+II conformations require one of the i/i+3 positions to have dihedral angles in the bottom lefthand corner of the Ramachandran plot (cyan in Figure 55A). This is a sparsely populated region of the

Ramachandran plot for chiral amino acids which may be one of the reasons such clusters are typically higher energy. The I’+I’, II’+II’ and I’+II’ conformations (Figure 55C) may now be appearing as favourable turn-type combinations, despite the c(GGGGGG) showing them as high energy conformations, as both the i/i+3 dihedral angles are in a more populated region of the Ramachandran plot for L-amino acids compared to the other combinations.

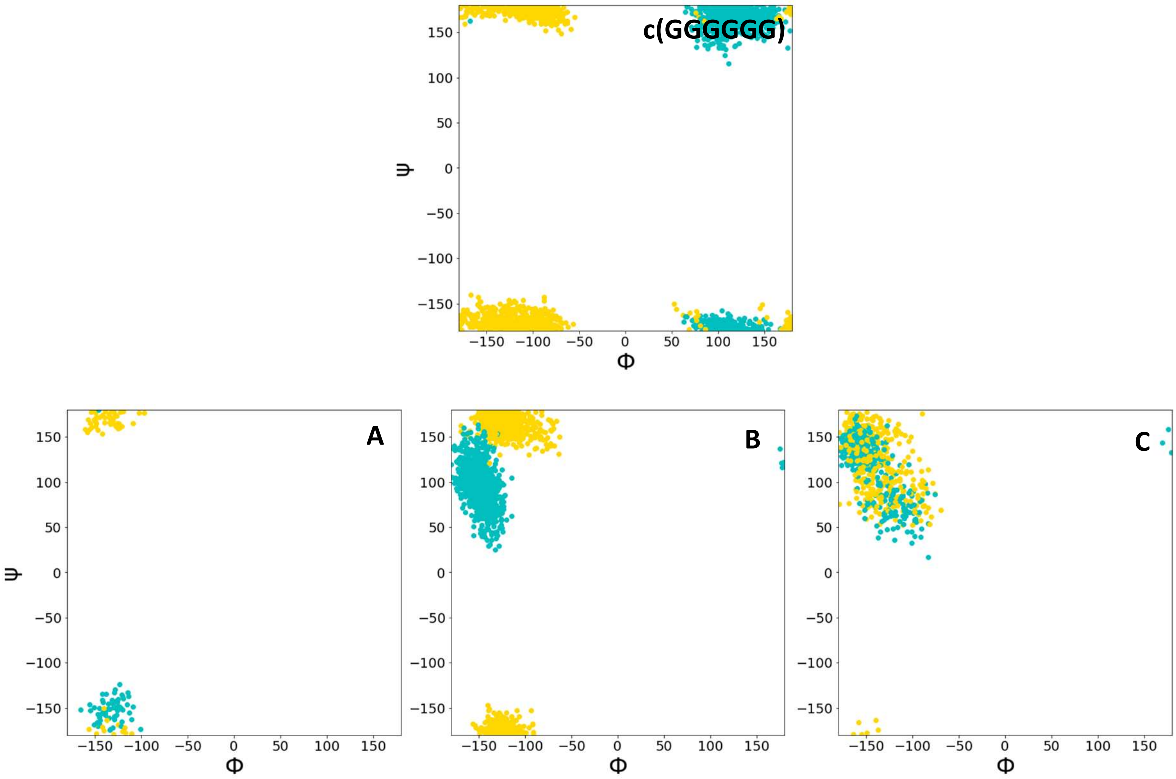


Figure 55: The Ramachandran plots of the i/i+3 positions in c(GGGGGG) compared to the three i/i+3 dihedral angles distributions seen in the simulations of c(AGGAGG). A shows the i/i+3 positions found in the I+II and II+II conformations, B the II+II’ and I+II’ conformations and C the I’+I’, I’+II’ and II’+II’ conformations. Yellow is the i position of the restrained turn and cyan is the i+3 position.

The conformations seen in the c(XGGXGG) simulations resemble some of the conformations seen in the c(XGGGGG)/c(GGGXGG) simulations (Figure 56). Whereas the clusters in the c(XGGGGG)/c(GGGXGG) simulations that resembled those seen in the c(GGGGGG) simulations occurred in similar quantities to those in the c(GGGGGG) simulations, the clusters that are more similar to the c(XGGXGG) simulations appeared in different proportions with completely different conformations sometimes seen. For example the c(AGGAGG) and c(AGGGGG) type II’ restrained simulations have different major clusters with the II’+IV_3 conformation appearing approximately twice as much in the c(AGGGGG) trajectory. One possible reason for this is that as glycine occupies one of the i/i+3 positions in the c(XGGGGG)/c(GGGXGG) simulations different lowest energy conformations are made possible due to the flexibility of glycine. Additionally sidechain-sidechain interactions could be altering the energy of the conformations seen in the c(XGGXGG) simulations.

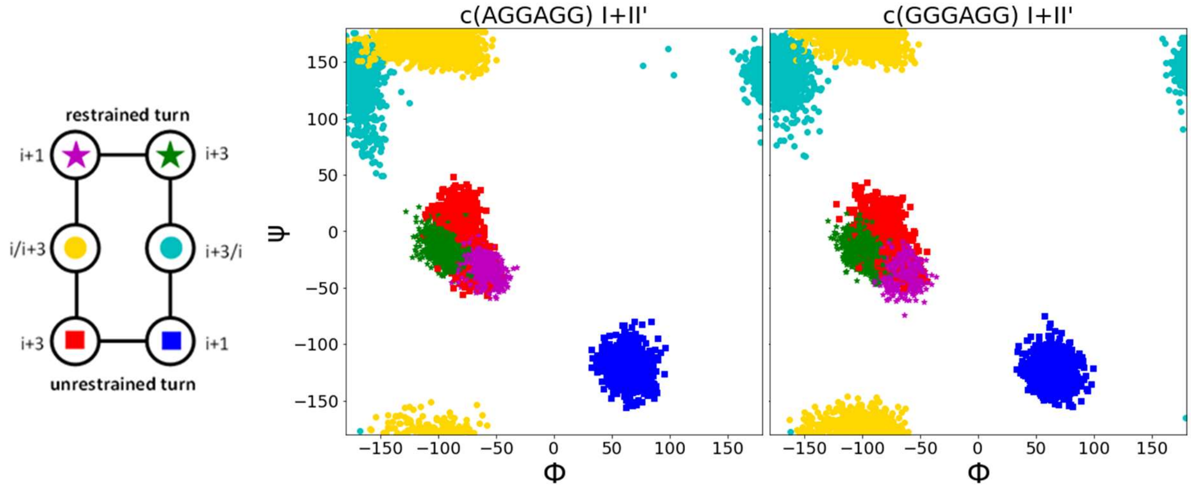


Figure 56: As seen by the similarity of the Ramachandran plots, similar conformations are seen in some of the c(XGGGGG)/c(GGGXGG) and c(XGGXGG) restrained simulations.

4.5.2 Differences in the c(XGGXGG) Simulations due to the Different Amino Acids

Although generally the same lowest energy conformations are seen in the c(XGGXGG) simulations when X = A, F, Q or V, there is some variation in the proportions of each cluster as the amino acid changes e.g. the I'+II' cluster represents 26% of the data in the c(AGGAGG) type I' simulation but does not appear at all in the corresponding simulation on c(VGGVGG). The differences in the simulations show how the choice of amino acid at the i/i+3 positions alters the most stable conformation.

Differences are seen in the relative energies of the clusters between the different sequences, with the alanine clusters being closer in energy than those in the other simulations. This reflects the different amino acids influence on the most stable conformation of the peptide with alanine potentially leading to more flexible cyclic peptides when at the i/i+3 position as there is less energy difference between the conformations.

The Ramachandran positions necessary for each turn type combination remain the same regardless of which chiral amino acids occupy the i/i+3 positions. For example the same i/i+3 positions are occupied for the I'+I' conformation formed by c(AGGAGG) as the I'+I' conformation formed by c(VGGVGG). These represent the lowest energy structures for each turn type combination when chiral amino acids are at both the i/i+3 positions in a cyclic hexapeptide. The Ramachandran positions are distinct for each turn type combination: for example the I'+I' conformation has slightly different i/i+3 dihedral angles than the I'+II' conformation (Figure 57).

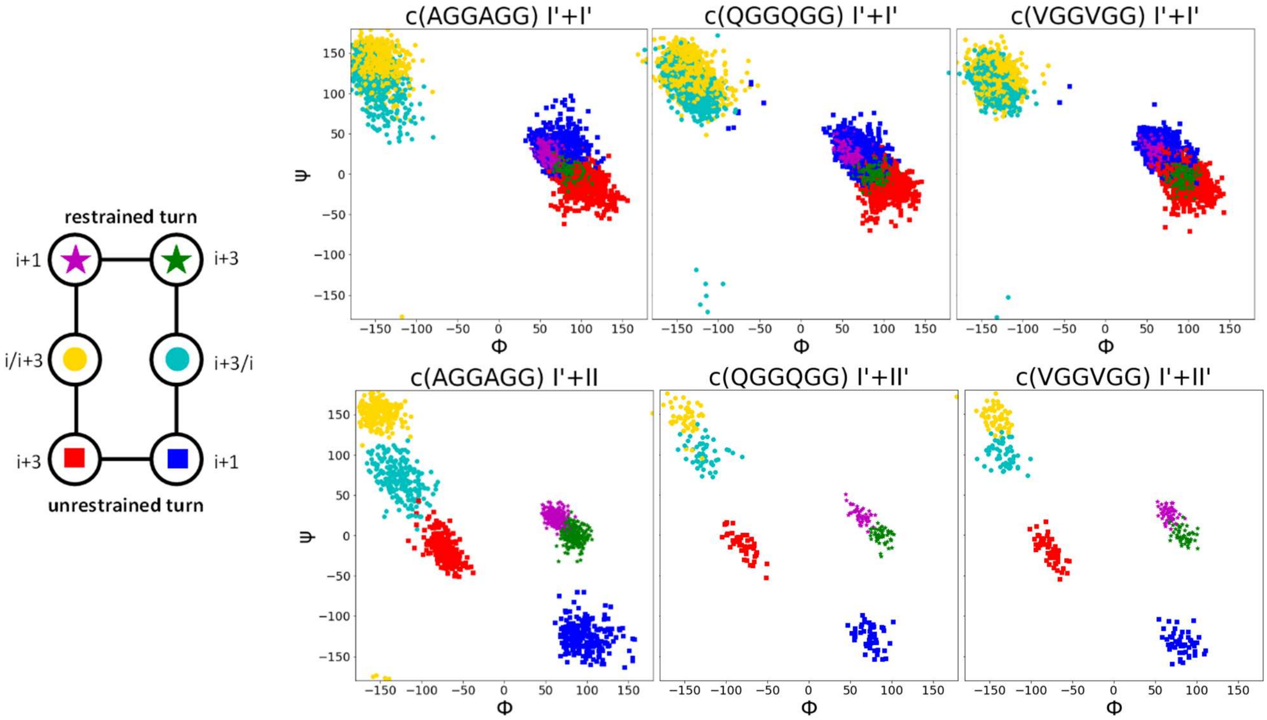


Figure 57: The same Ramachandran distribution is seen for the I'+I' conformation whether A, Q or V is at the i/i+3 positions. This conformation is different to the i/i+3 Ramachandran positions seen in the other conformations such as the I'+II'.

The different proportions of the conformations seen when different amino acids are used could potentially be due to sidechain interactions. In different conformations the sidechains could be pointing away from or towards each other resulting in changes in energy of the conformation due to the hydrophobic effect or sterics. These small energy differences could potentially alter the distribution between conformations a lot. Each turn type combination has unique i/i+3 Ramachandran distributions. It is possible the differences in the amount of each conformation seen depending on which chiral amino acids are used are due to the differences in Ramachandran distributions of each amino acid and the similarity with the necessary areas of the Ramachandran plot for each turn type combination. All turn combinations will have an inherent stability but if an amino acid has a Ramachandran distribution better suited to one turn combination over another, it may mean a higher proportion of that conformation is seen compared to an amino acid with a less compatible Ramachandran distribution. Valine as a β-branched amino acid shows a different Ramachandran distribution to Alanine and Glutamate, which although different show more similarity to each other than valine. This may be why much more similarity is seen in the restrained simulations of c(AGGAGG) and c(QGGQGG) than that of c(VGGVGG).

4.5.2.1 c(QGGQGG)

To test whether the difference between the alanine and glutamine simulations was due to the ability of the amide bond functionality on glutamine hydrogen-bonding to the peptide backbone or the other glutamate sidechain, the hydrogen-bond was looked for throughout the simulation. If a hydrogen bond is present the H-O distance should be less than 2.5 Å.[338] Plotting histograms of the distances extracted from the simulations shows hydrogen bonding is infrequent so likely not occurring enough to explain the differences in the proportions of the conformations. Figure 58 shows a representative example of these distances taken from the type II restrained simulation as this shows the greatest difference in the percentage of each cluster seen compared to the type II restrained simulation of c(AGGAGG). Measuring the distance between the other potential hydrogen-bonding configurations similarly shows little evidence of hydrogen-bonding with few distances below 2.5 Å measured.

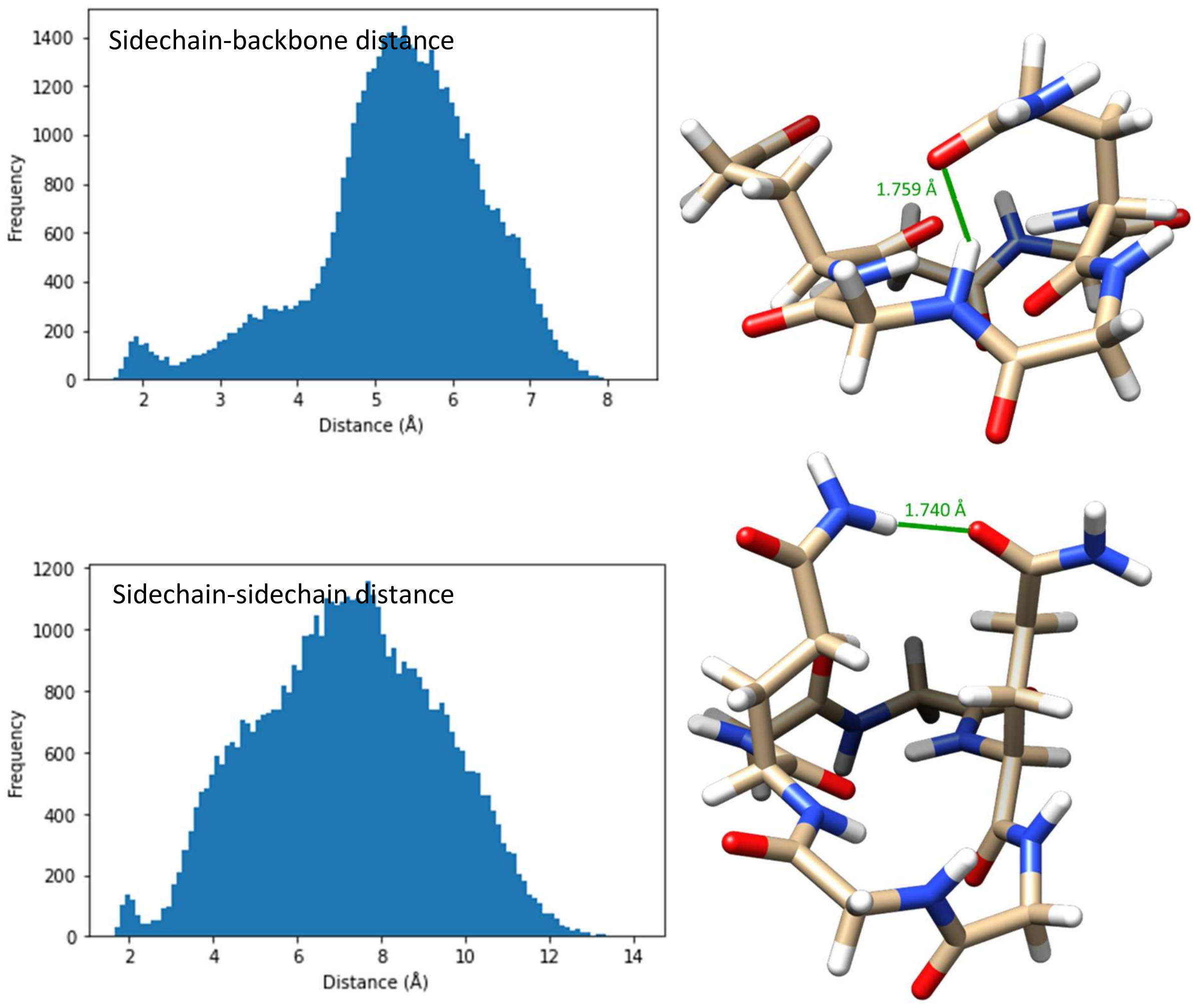


Figure 58: Representative examples of hydrogen bonding by the glutamine sidechain. The distance between the hydrogen and carbonyl groups in the restrained simulations rarely reaches less than 2.5 Å between the glutamine sidechain and the backbone or between the two glutamine sidechains.

4.5.2.2 Type IV Turns

Type IV turns are seen to appear across the c(XGGXGG) restrained simulations. The same type IV turns appear across the different simulations, with three different turn types being the most common (Figure 59). The most common is a distorted type I turn where the ψ dihedral angles of the i+2 residue are lower than seen in a type I turn (Figure 59A). This turn was classified as a type IV_3 turn by de Brevern.[123] The type IV_3 turns were assigned to the same category as type I turns by the clustering algorithm used for analysis of the Loop Database, but they only appeared when the 7 Å between the i and i+3 α-carbons definition of a β-turn is used rather than the hydrogen-bonded definition and they are not seen in the hairpin dataset. The second type IV cluster (Figure 59B) was seen as a cluster in the Loop Database but similar to the type IV_3 turns was only seen when the 7 Å definition of a β-turn was used and did not appear in the hairpin dataset. The remaining type IV turn type appeared very infrequently in the Loop Database and was not seen to form a cluster. It may be that it is a more favourable turn type in the relatively restrained system of a cyclic peptide than in a protein. Small cyclic peptides are also more solvent exposed than many amino acids found in the hydrophobic core of a protein which could alter the distribution of turn types seen.

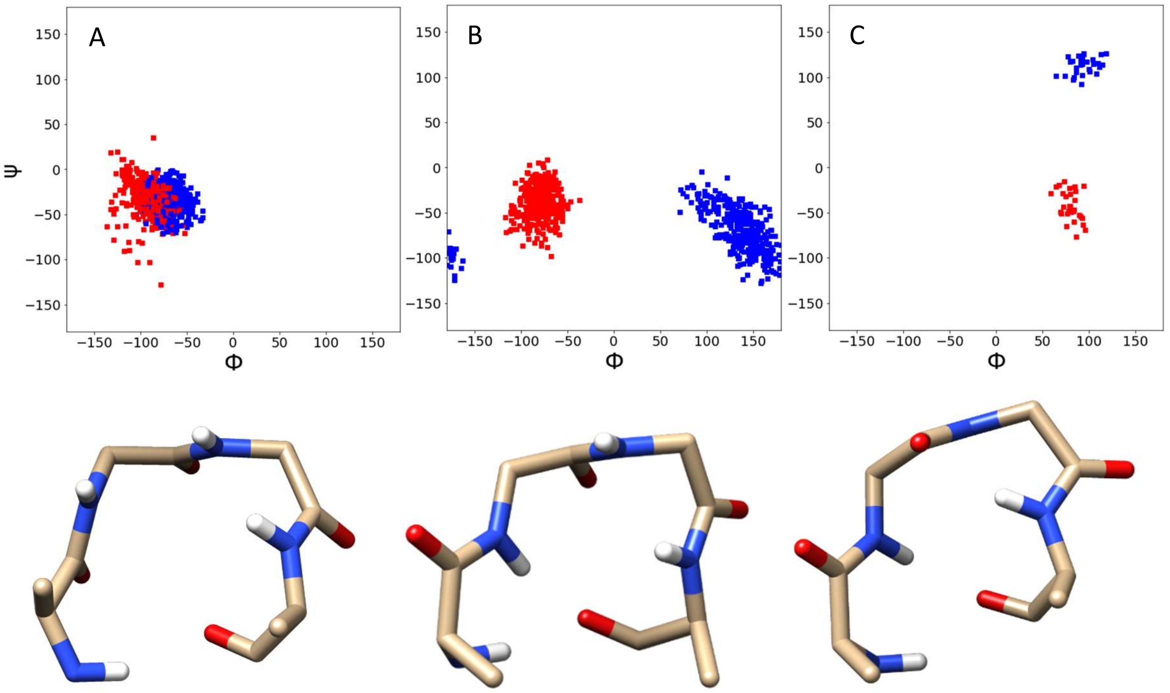


Figure 59: The three most common type IV turns in the restrained simulations. A is the most common, followed by B then C. In the Ramachandran plots blue is the i+1 position and red is i+2.

It may be that some of these type IV turns only appear in the restrained simulations due to the i+1 and i+2 positions of the unrestrained turn being made up of glycine residues. Glycine, being unsubstituted, is flexible so has a broader Ramachandran distribution than the other amino acids. The second most common type IV turn (B in Figure 59) has the i+1 position in the righthand region of the Ramachandran plot, a region not usually occupied by L-amino acids. The final type IV turn that appears across multiple simulations (C in Figure 59) has both the i+1 and i+2 positions in the right-hand region of the Ramachandran plot. Type I' turns also have dihedral angles where both the i+1 and i+2 positions occur in the right-hand region of the Ramachandran plot. A glycine usually occupies at least one of the i+1 or i+2 positions of most observed type I' turns. Therefore these type IV turn types may not occur outside of the restrained simulations despite forming a favourable conformation. This demonstrates the ability of the restrained simulations to find the lowest energy conformations in the absence of sequence preferences.

4.5.3 Unrestrained BE-META of c(AGGAGG)

The structure of c(AGGAGG) without the effect of the restrained turn was obtained by running BE-META. The alanines generally appear at the i+1 positions of the β-turns, with one cluster where they appear at the i+2 positions of the β-turns (Table 24). As the peptide is mainly composed of glycine it is very flexible and many clusters are seen. The alanines alter the conformation from c(GGGGGG) showing how small changes can alter the conformation of cyclic peptides. The smaller clusters seen in this simulation differ slightly to those seen in the equivalent simulation by McHugh *et al.*,[245] possibly due to differences in processing the data prior to clustering. McHugh *et al.* used an additional noise-filtering step prior to clustering to try to reduce the computational time required for clustering. Additionally noise in the simulations can lead to slight difference in the size of clusters seen.

β-turn position	Population (%)	Type
c(AGGAGG)	46	II+I’
c(AGGAGG)	21	I+II’
c(AGGAGG)	17	I+I’
c(AGGAGG)	16	I+II

Table 24: The clusters obtained in the BE-META of c(AGGAGG). The i+1 and i+2 positions of each β-turn is coloured. Red is a type I turn, green is type II, orange type I’ and blue is type II’.

None of the conformations have the alanines at the i/i+3 positions. The restrained simulations on c(AGGAGG) cannot therefore be used to predict the conformation of c(AGGAGG) as they could be for c(GGGGGG) as they keep the alanines at the i/i+3 positions. It is likely if the restrained simulations were repeated on c(AGGAGG) but the restraints were moved so the alanines occupied the i+1 position of the β-turns then they would show the same lowest energy structures as seen in the unrestrained BE-META on c(AGGAGG) (with the exception of the conformation where the alanines occupy the i+2 positions of the β-turns). By keeping the alanines at the i/i+3 positions the restrained simulations can be used to look at the relative energies of conformations that would not occur without an energetic bias. By looking at these conformations in the simpler systems used for the restrained simulations it is easier to infer patterns than in more complex systems. They could therefore be applied to other systems with different changes made to the peptide and the effect on the conformations seen observed. For example the i+1 and i+2 positions could be substituted in turn to determine the effects on the most stable conformations.

4.6 c(X^1GGX^2GG) simulations

Further simulations were run with different chiral amino acids at the i/i+3 positions (Table 25). Alanine, glutamate, valine and phenylalanine were used to complete all possible restrained simulations with these amino acids occupying the i/i+3 positions i.e. simulations were carried out on c(X^1GGX^2GG) where X^1X^2 = AQ, QA, AV, VA, QV, VQ, AF, FA, QF, FQ, VF, FV. This shows the effect of different types of amino acids at each of the i/i+3 positions.

		% Observed in Restrained Simulations											
Restrained turn	Unrestrained turn	AGGFGG	AGGQGG	AGGVGG	FGGAGG	FGGQGG	FGGVGG	QGGAGG	QGGFGG	QGGVGG	VGGAGG	VGGQGG	VGGFGG
I	I	-	-	-	-	-	-	-	-	-	-	-	-
	II	11	-	8	2	7	14	-	3	18	10	5	33
	I’	-	-	-	-	18	4	-	-	-	-	24	1
	II’	74	71	60	70	57	70	80	81	69	59	50	51
	IV	1	1	-	15	-	-	-	-	-	23	4	
II	I	18	-	-	14	-	3	-	5	-	-	-	14
	II	5	7	14	6	11	11	5	-	3	23	8	12
	I’	-	-	-	-	-	-	-	-	-	-	-	-
	II’	66	44	30	67	53	52	72	73	51	59	45	63
	IV	-	39	49	-	29	18	1	2	32	9	39	-
I’	I	-	-	-	-	-	-	-	-	-	-	-	-
	II	-	-	-	-	-	-	-	-	-	-	-	-
	I’	44	36	43	69	61	68	58	68	71	66	72	75
	II’	34	39	36	-	10	8	21	11	5	10	3	4
	IV	-	10	3	-	-	-	-	-	-	-	-	-
II’	I	-	27	11	-	-	-	-	-	-	-	-	-
	II	26	31	27	21	25	20	19	16	14	15	19	10
	I’	3	2	-	29	12	10	29	22	9	38	16	20
	II’	41	24	43	32	42	57	31	46	62	25	45	52
	IV	23	-	12	13	12	-	16	5	-	7	5	2

Table 25: Restrained simulations on c(X^1GGX^2GG). If multiple type IV turns occur they have been grouped into one category.

For the previous c(XGGXGG) simulations the ΔG between conformations which occur in different restrained simulations could be determined as long as one conformation which is the same in the two simulations is seen. Similarly for the c(X^1GGX^2GG) sequences, pairs of simulations can be used to determine the energy differences between conformations if the c(X^1GGX^2GG) and c(X^2GGX^1GG) simulations contain an equivalent cluster. The two different amino acids must occupy the same i/i+3 position in the conformation and the same Ramachandran distribution must be seen. For example the c(AGGQGG) II’+I conformation is not equivalent to the c(AGGQGG) I+II’ conformation as alanine occupies different i/i+3 positions in each. However the I+II’ conformation of c(AGGQGG) will be the same conformation as c(QGGAGG) II’+I. In both conformations it is the alanine which occupies the i position of the type I turn and glutamine which occupies the i+3 position of the type II’ turn (Figure 60).

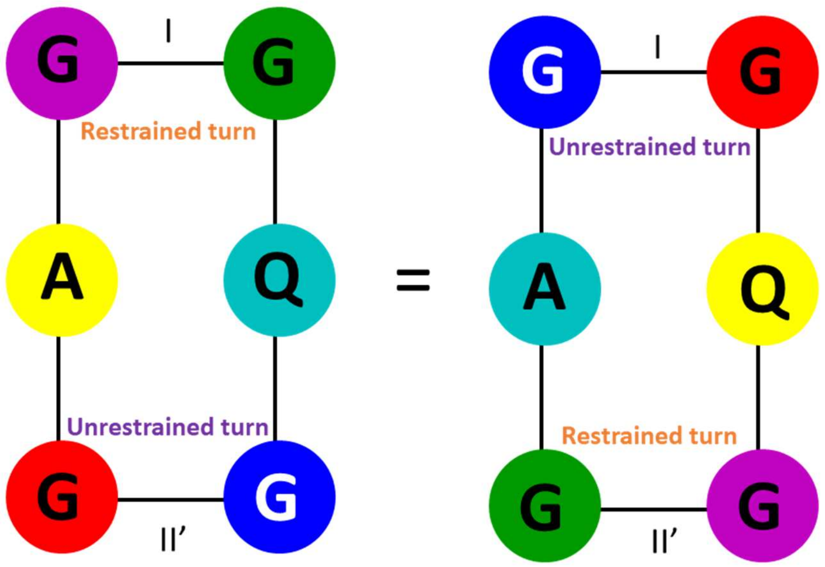


Figure 60: The c(AGGQGG) I+II’ conformation shows the same Ramachandran distribution and is equivalent to the c(QGGAGG) II’+I conformation.

The energy differences between the conformations were calculated and are shown in Figure 61. The relative energies of some of the conformations could not be determined as they had no equivalent conformation seen in another simulation. Usually these conformations were from the type I restrained simulation as there is no x+I conformation in another simulation. As the type I conformation isn’t generally seen at the unrestrained side of the cyclic peptide this means more favourable turn types are possible, with the type I turn being a higher energy conformation. The majority of β-turns found in the Loop Database were type I turns. As the majority of sequences form type I turns but type I turns rarely form part of the lowest energy conformations of the backbone structure of cyclic hexapeptides, this is potentially one of the reasons cyclic hexapeptides typically form multiple conformations in solution. If there is a small energy difference between the sequence forming a type I turn compared to forming another turn type which allows for the backbone conformation of the peptide to form a more stable conformation, then multiple conformations could be seen.

In general the order of conformations is very similar between sequences. The I’+I’ conformation is usually the lowest energy conformation despite not being seen frequently in known NMR and X-ray crystal structures of cyclic peptides. Relatively few type I’ β-turns were found in the Loop Database however so this may reflect the fact that for most sequences other turn types are more favourably formed. Other low energy conformations include II’+II’, II+II’ and I’+II’, the same low energy conformations seen in the c(AGGAGG), c(QGGQGG), c(FGGFGG) and c(VGGVGG) simulations.

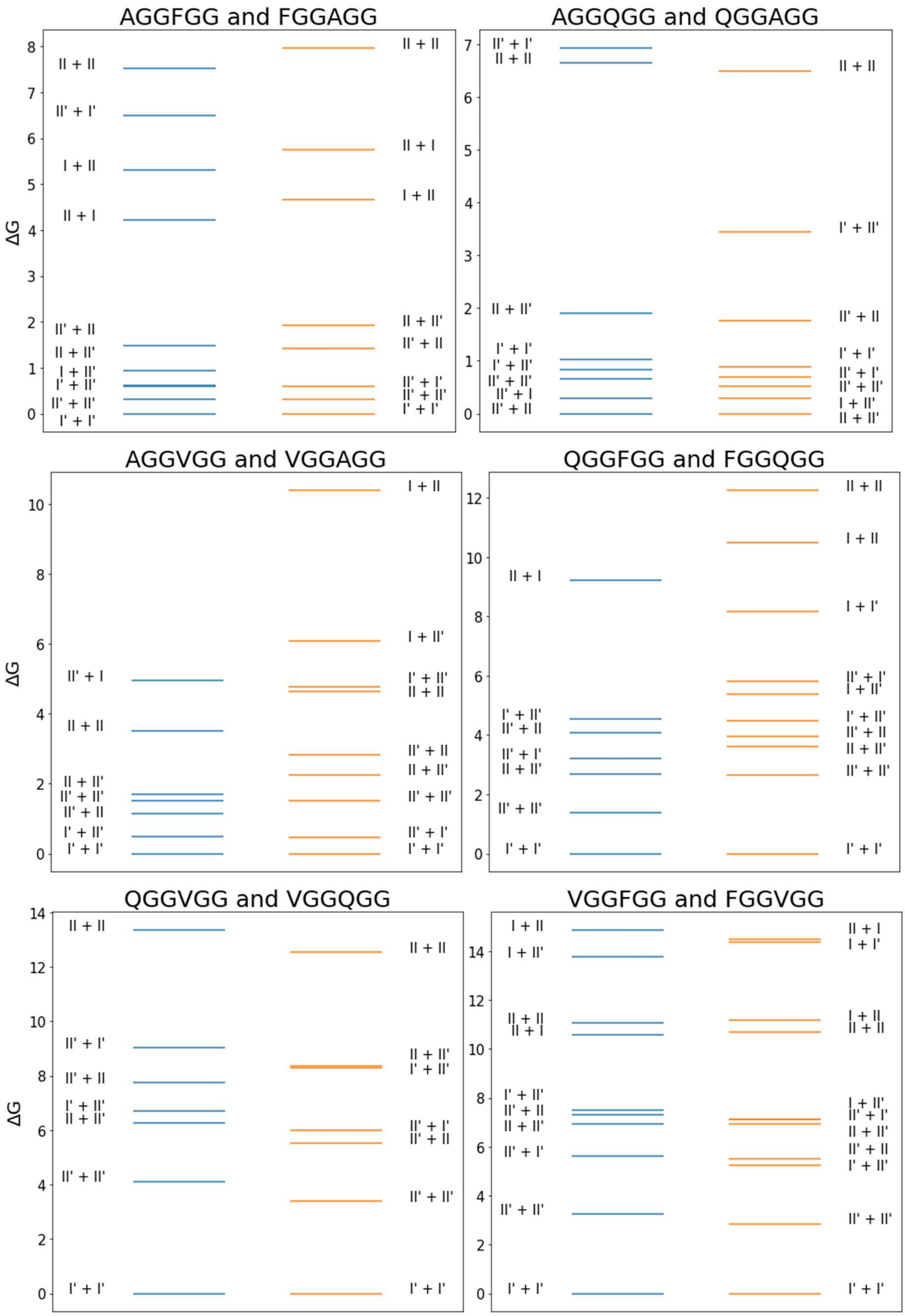


Figure 61: Energy differences between conformations in the c(X^{1}GGX^{2}GG) restrained simulations.

For some sequences there is a relatively large energy difference between what should be equivalent conformations such as the c(AGGQGG) II’+I’ conformation which is equivalent to the c(QGGAGG) I’+II’ conformations. The largest energy differences tend to be in the highest energy conformations which only appear in small amounts in the restrained simulations. For example the II’+I’ conformation is only observed 2-3% of the time in the II’ restrained simulations of c(AGGQGG) and c(QGGAGG). The size of the clusters in the constrained simulations can vary by approximately 5, so the II’+I’ clusters which are observed 2-3% could potentially not appear at all in the restrained simulations or may be seen up to 8% of the time. This means noise in the simulation proportionally affects smaller clusters more so the calculated ΔG value has a larger error. The energy difference between equivalent conformations can also vary slightly due to the small differences in the conformations based on which of the turns is restrained. The unrestrained turn has a wider distribution around the ideal dihedral angles for a given turn type and the dihedral angles may be centred around slightly different values. Reducing the force constant on the restrained turn could potentially make the unrestrained and restrained turns more similar to reduce this effect.

4.7 Double Restrained Simulations

There are ten possible conformations made up of the I, II, I’ and II’ turn types when chiral amino acids are at the i/i+3 positions (Table 26), as for example, the I+II combination shows the same Ramachandran distribution as the II+I conformation. Although most of the ten possible turn combinations made up of I, II, I’ and II’ turns were seen in the restrained simulations, some turn type combinations (I+I and I’+II) did not appear in any of the restrained simulations no matter which amino acids were present at the i/i+3 positions. These conformations are therefore higher energy than the other conformations that appear in the restrained simulations.

	I	II	I’	II’
I	I+I	■	■	■
II	I+II	II+II	■	■
I’	I+I’	II+I’	I’+I’	■
II’	I+II’	II+II’	I’+II’	II’+II’

Table 26: The 10 possible β-turn combinations.

Although the I+I conformation was also found to be very high energy in the c(GGGGGG) simulations, the I’+II combination was one of the most favourable structures appearing in similar quantities to the I+II’ conformation. With the inclusion of chiral amino acids the I+II’ conformation remains the lowest energy in the type I simulation despite the I’+II conformation not appearing in either the I’ or II restrained simulations where lower energy conformations are now able to form. It may be that although in the type II and I’ simulations a new more favourable alternative is available, when the peptide contains a type I turn this is not the case so the I+II’ conformation is still seen. Alternatively these conformations are unfavourable due to particular sidechain interactions or the Ramachandran positions necessary for these conformations to form when chiral amino acids are present at the i/i+3 positions. Double restrained simulations were designed to find which i/i+3 dihedral angles are lowest energy for these conformations.

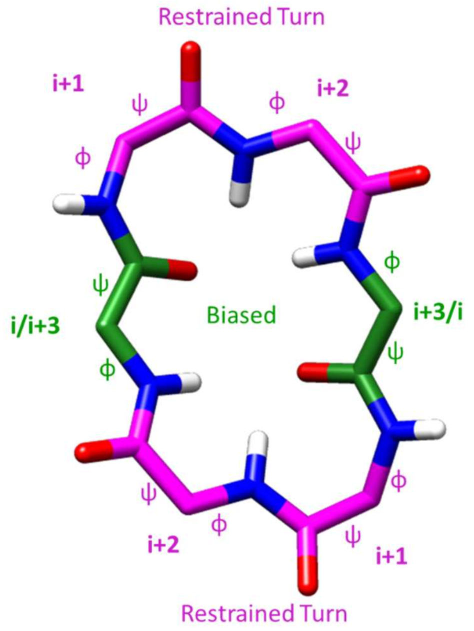


Figure 62: Double restrained simulations. Both β*-turns are restrained to ideal values for a given* β*-turn type. The two i/i+3 positions have bias acting on their φ and ψ dihedral angles to find the lowest energy conformation of the peptide when the specified* β*-turn types are present.*

For the double constrained simulations both i+1 and i+2 positions within the cyclic peptide were restrained to the necessary turn types and biased replicas were only kept for the i/i+3 positions (Figure 62). These simulations were carried out for c(XGGXGG) where X is A, F, Q or V. These simulations show the necessary i/i+3 dihedrals required for these turn combinations to occur.

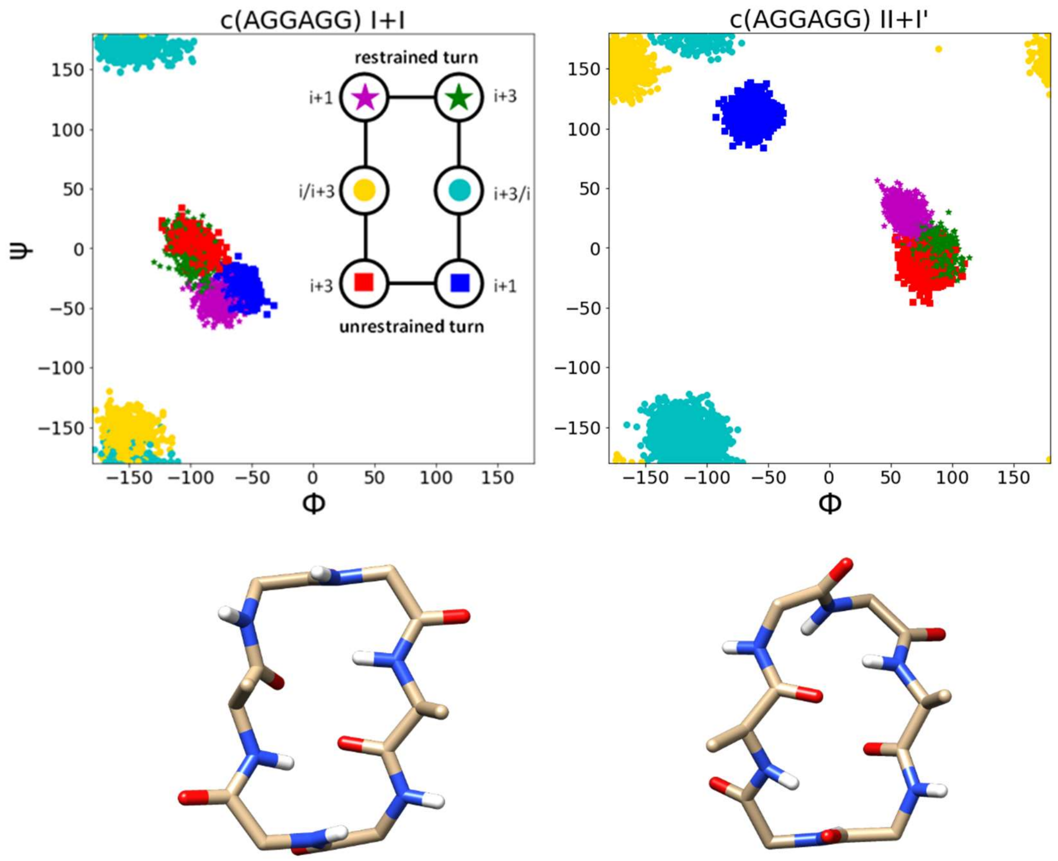


Figure 63: c(AGGAGG) I+I and II+I’ conformations.

Only one conformation was seen in each simulation showing the most favourable conformation with the specified β-turn type combination (Figure 63). The exception was the I’+II simulation which produced two clusters for c(QGGQGG). A small cluster, not seen in the alanine or valine simulations, has one glutamate with dihedral angles occupying a position on the righthand side of the Ramachandran plot (Figure 64). As this cluster only appears in small numbers in just the glutamate simulation the larger cluster was chosen as representative of the I’ + II conformation. The smaller cluster does not contain significant amounts of hydrogen-bonding of the glutamate sidechain. Although the restrained simulations show the lowest energy conformation for each turn type combination remains the same when different amino acids occupy the i/i+3 positions, alternative conformations made up of the same turn type may be possible. Like the minor I’+II conformation they will be higher energy based on the backbone structure of the peptide but may be seen in the structure of cyclic peptides due to sequence specific effects.

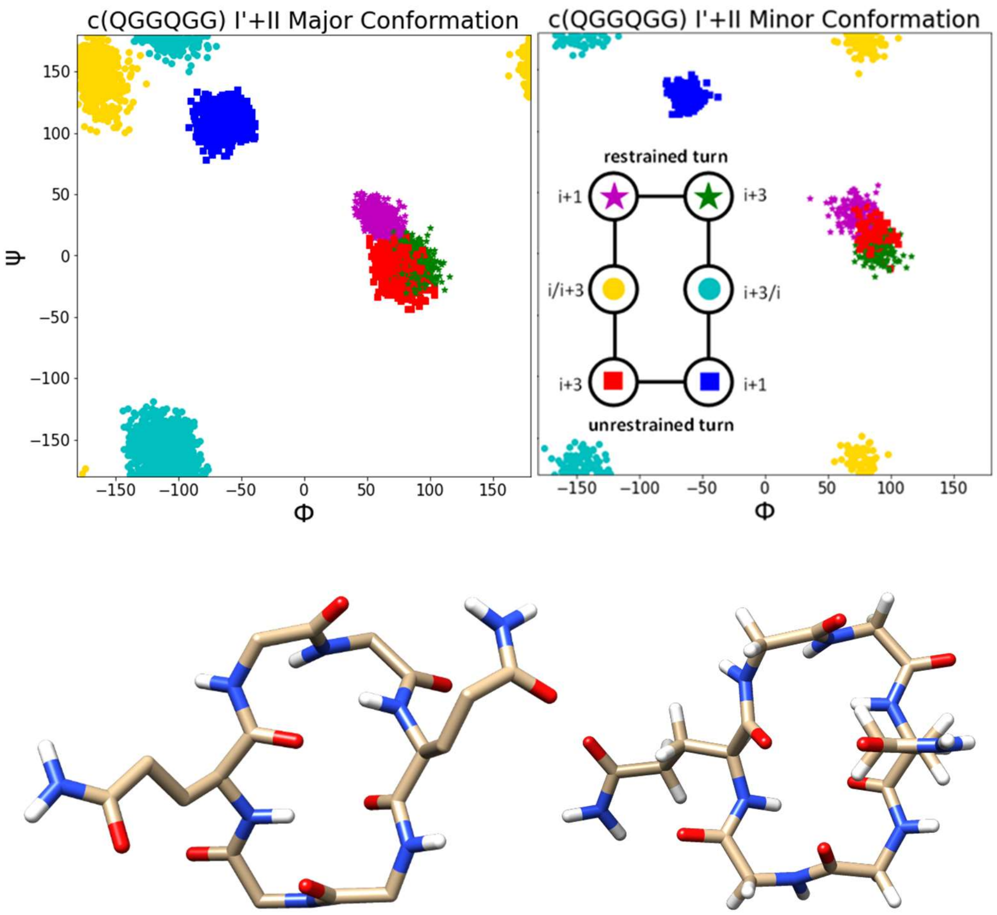


Figure 64: The major cluster seen in the I’+II double restrained simulation of c(QGGQGG) is the same as seen in the simulations of c(AGGAGG) and c(VGGVGG). A smaller cluster is also seen with different i/i+3 Ramachandran angles.

The energy differences between clusters seen in the double restrained simulations compared to the clusters seen in the restrained simulations cannot be determined. It can be inferred that since they do not appear in the restrained simulations there is a lower energy conformation available. This means these conformations may not be observed frequently in cyclic peptides and when designing a peptide with a single dominant conformation it may be best to avoid sequences that would form these conformations.

Both the I+I and I’+II conformations have one i/i+3 position which requires dihedral angles in the bottom lefthand corner of the Ramachandran plot. This is an unfavourable part of the Ramachandran plot for L-amino acids. This is similar to the I+II and II+II conformations which were seen in the restrained simulations as the highest energy conformations. Two intrapeptide hydrogen bonds are usually seen in the cyclic hexapeptide conformations. For the I’+II conformation the slightly twisted structure means the hydrogen bond at the type II turn is not present throughout much of the simulation. For the I+II’ conformation however two hydrogen bonds are present throughout most of the simulation which could potentially be one factor in why the I+II’ conformation is seen in the restrained simulations whereas the I’+II conformation is not despite both being low energy conformations in the c(GGGGGG) simulations.

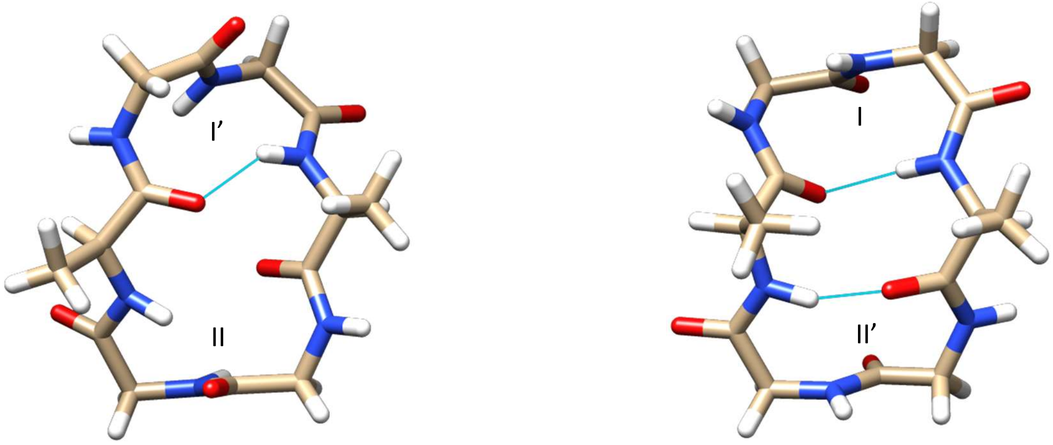


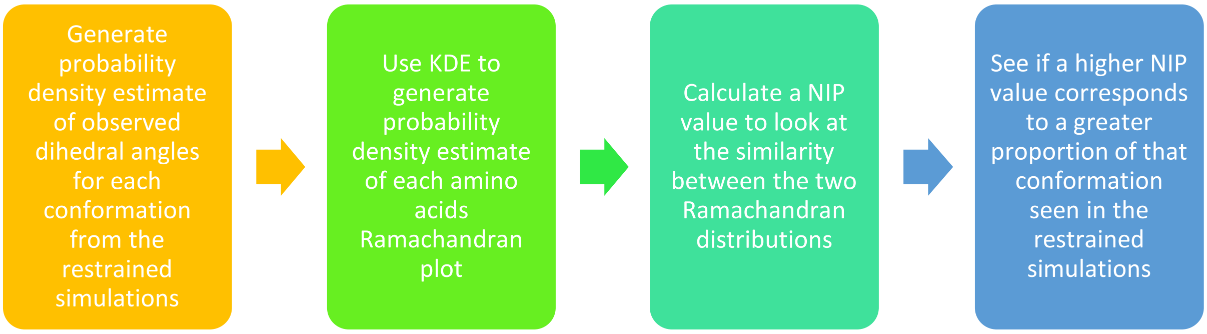
Figure 65: One intrapeptide hydrogen bond (shown in light blue) is typically seen in the II+I’ conformation whereas two hydrogen-bonds are seen within the I+II’ conformation.

To summarise restrained BE-META simulations were devised to determine the lowest energy conformations of cyclic peptides when a particular turn type is included in the structure. Different structures are seen when glycine occupies one or both of the i/i+3 positions which should be considered when designing cyclic peptides. The restrained simulations on c(XGGXGG) show how each cyclic hexapeptide conformation made up of type I, II, I’ and II’ β-turns has unique i/i+3 dihedral angles. These dihedral angles vary between different turn type combinations but for a particular turn type combination the same dihedral angle distribution is seen independent of the choice of chiral amino acids at the i/i+3 positions. The choice of L-amino acid at the i/i+3 position generally doesn’t change the lowest energy structures, but slight differences are seen in the energy differences between conformations. To explore if these differences are due to the different amino acids preferences for different dihedral angles making some conformations more favourable the similarity between the Ramachandran plot of the amino acids with the required dihedral angles for a particular conformation was next investigated. One method to look at the similarity between two distributions is the normalised integrated product (NIP).

4.8 Using the Normalised Integrated Product to Compare Distributions

4.8.1 Overview

To look at the similarity between Ramachandran plots to determine if the Ramachandran plot of individual amino acids alters their predisposition to certain cyclic peptide conformations the normalised integrated product (NIP) can be used. First a probability distribution is generated from a Ramachandran plot using kernel density estimation (KDE) to give a smoothed estimation of the likelihood of given dihedral angles occurring. The overlap of two probability distributions is then generated by obtaining the NIP. The overlap of each amino acids intrinsic Ramachandran distribution is compared with each conformation observed i/i+3 dihedral angles as obtained from the restrained simulations. The proportion of each conformation seen in the restrained simulations can then be compared to the NIP value to determine if this overlap can be used to help predict the conformation of cyclic peptides.



Generate probability density estimate of observed dihedral angles for each conformation from the restrained simulations

Use KDE to generate probability density estimate of each amino acids Ramachandran plot

Calculate a NIP value to look at the similarity between the two Ramachandran distributions

See if a higher NIP value corresponds to a greater proportion of that conformation seen in the restrained simulations

Figure 66: Flowchart of the process to generate NIPs and determine if they can be used to predict the proportion of a particular conformation.

4.8.2 Kernel Density Estimation

Kernel density estimation (KDE) is a statistical method which can be used to generate a smoothed probability distribution to estimate the likelihood of dihedral angles occurring based on individual data points (Figure 67). KDE adds a kernel function (generally a Gaussian kernel) at each data point. By summing all the Gaussians within a grid point of the data an estimate of the density at that point can be made leading to a smoothed probability distribution.

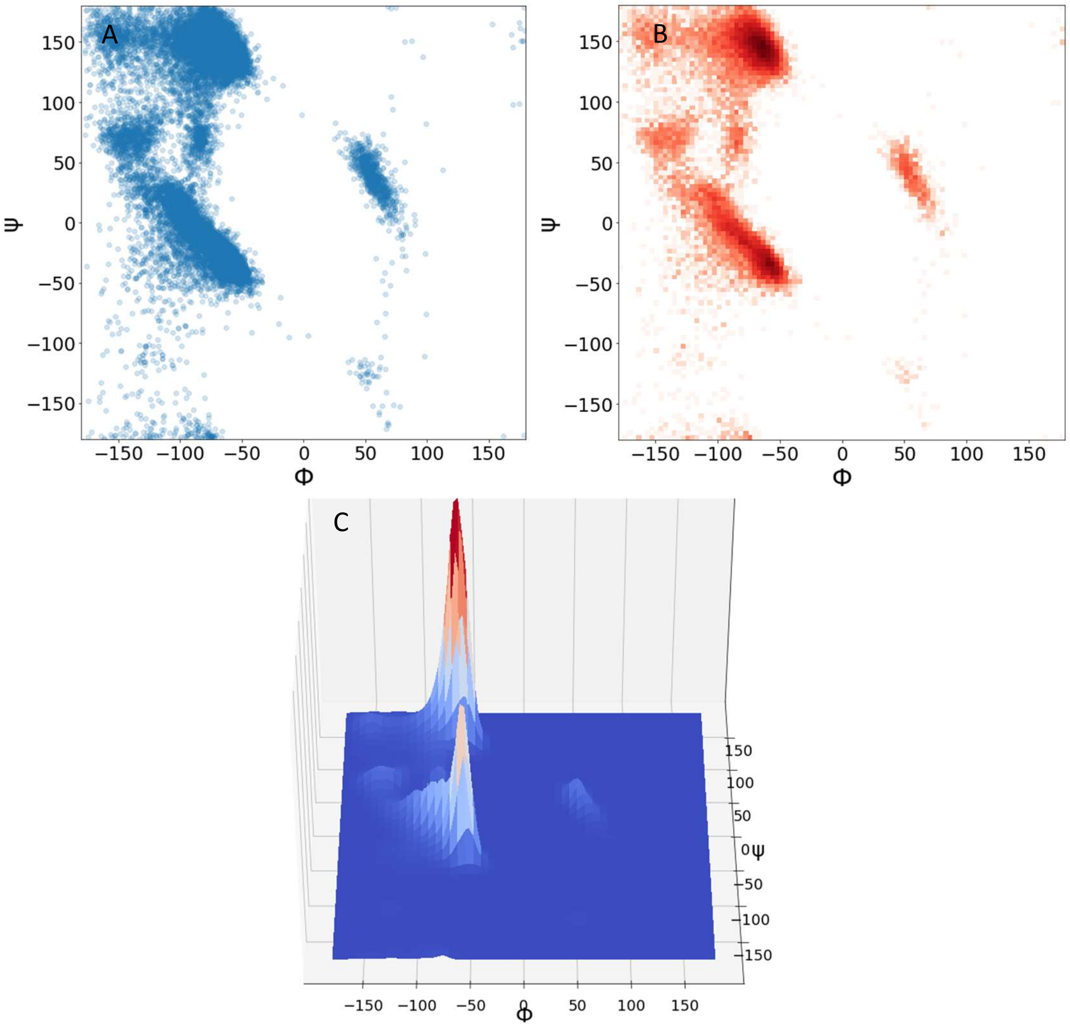


Figure 67: A: The Ramachandran plot of alanine. B: the same Ramachandran plot shown as a histogram and C: the probability density estimate of the Ramachandran plot produced by KDE.

The calculation using KDE to generate a probability at a point r (P(r)) with a Gaussian kernel estimator from a set of n points is shown in equation 1. The kernel bandwidth is represented by the letter h, d is the dimensional space of the data and r_k is the position of point k in space.

$$\mathrm{P(r)} = \frac{1}{n}\sum_{k=1}^{n}\prod_{\alpha=1}^{d}\frac{1}{h\sqrt{2\pi}}\ e^{\left[-\frac{(r_\alpha - r_{\alpha k})^2}{2h^2}\right]} \qquad (1)$$

As dihedral angles are circular variables this probability density function needs to be adapted for use on a Ramachandran plot. Fisher proposed a model based on linear data methods whereby the point r is replaced by angle θ.[339] Rather than measuring distance in Euclidean space (r-r_k) the distance between the two angles (θ-θ_k) is measured using the angles between the two vectors (equation 2).

$$\|\theta - \theta_k\| = \min\left(|\theta - \theta_k|, 2\pi - \ |\theta - \theta_k|\right) \qquad (2)$$

4.8.2.1 Kernel Bandwidth

The choice of the kernel bandwidth (h) can drastically alter the probability distribution obtained. If a too large bandwidth is used oversmoothing of the data will occur and features of the dataset will be lost. Conversely if the bandwidth is too small undersmoothing of the data will result in artifacts that do not reflect the true density of the data (Figure 68). For this reason many methods to generate the best bandwidth value to use have been developed. The two most commonly used statistical methods are Scott's rule of thumb (equation 3),[340] and Silverman's rule of thumb (equation 4),[341]. The standard deviation of the data is represented by σ, n is the size of the sample and IQR is the interquartile range.

$$h \ \approx 1.06\,\hat{\sigma}\,n^{-\frac{1}{5}} \qquad (3)$$

$$h \ \approx 0.9\min\left(\hat{\sigma}, \frac{IQR}{1.34}\right)n^{-\frac{1}{5}} \qquad (4)$$

Scott's rule gives relatively accurate results on normal distributions but can give less accurate results on data that is skewed or bimodal. Silverman's rule has been shown to be more robust generally producing probability distributions that better represent the data than Scott's rule so is generally preferred.[342] Despite this, Silverman's rule is still based on the assumption that the underlying shape of the data is a normal distribution and in more complicated datasets can give poor bandwidth selection. Therefore various methods have been developed to find the best bandwidth value without making assumptions about the shape of the data. Cross-validation is one such method which is frequently used. The cross-validation method for bandwidth selection splits the data into samples and then fits different bandwidth values to a sample of the data and assesses how well the bandwidth then fits another sample of the data. The bandwidth that best fits the data samples is selected.

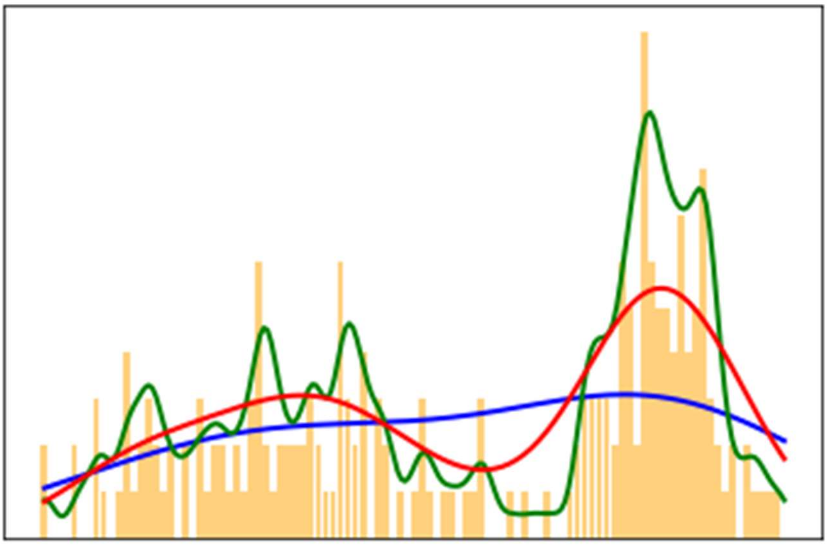


Figure 68: The effect of the kernel bandwidth. A set of data shown as a histogram overlayed with probability density distributions produced by KDE with different values of kernel bandwidth. Undersmoothing is seen with too small a kernel bandwidth (green) and oversmoothing seen with too large a kernel bandwidth (blue).

The underlying distribution of the Ramachandran plots is unknown and they are multi-modal meaning that Silverman's rule of thumb may lead to an inaccurate bandwidth. Cross-validation was therefore used to determine the best bandwidth for KDE. The difference between using Silverman's rule and cross validation for bandwidth selection to produce a probability distribution from a Ramachandran plot is shown in Figure 69. Silverman's rule gave a bandwidth value of 0.6, much lower than the cross-validation bandwidth of 4.1, leading to undersmoothing of the data.

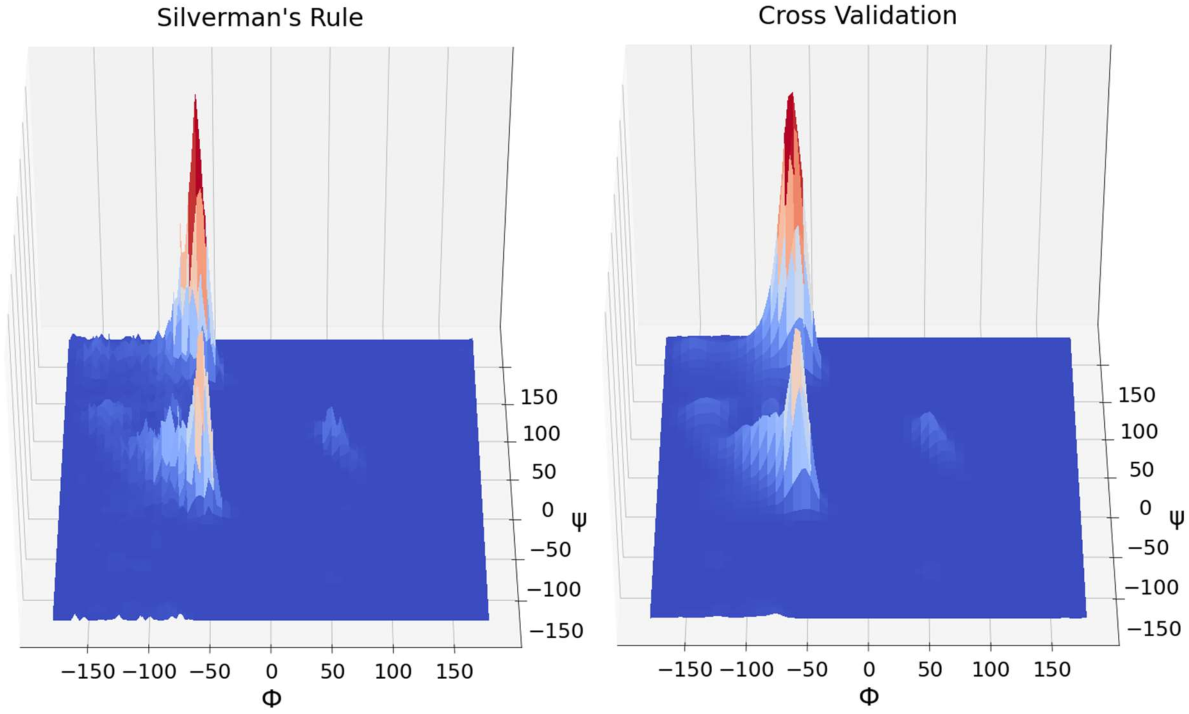


Figure 69: Silverman's vs Cross Validation bandwidth selection influence on the probability distribution.

4.8.2.2 Normalised Integrated Product of KDEs

Generating a normalised integrated product (NIP) is a way of comparing the similarity of two probability estimates. The NIP takes a value between 0 and 1, where 1 is perfect overlap between the two distributions being compared and 0 is no overlap. For example Figure 70 shows the plot of three probability distributions generated by KDE. A NIP value was generated to compare the orange and green probability densities with the blue distribution. The blue distribution is much more similar to the orange distribution so gives a high NIP value of 0.9 whereas the green distribution shows very little overlap with the blue distribution so gives a lower NIP of 0.2.

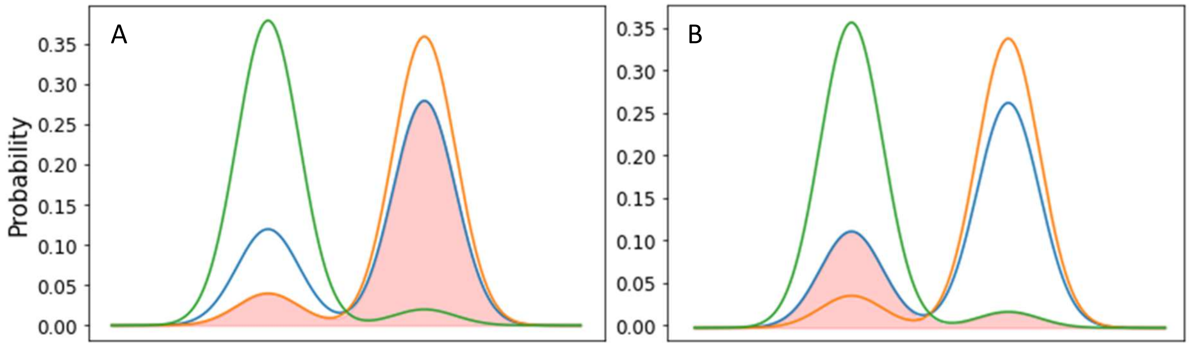


Figure 70: The overlap of the orange and blue distributions (A overlap shaded in pink) is relatively large giving a NIP value close to 1, whereas the NIP between the blue and green distributions (B) is much lower as they have little overlap (overlap shaded in pink).

The normalised integrated product (NIP) of two kernel density probability functions of two Ramachandran plots can be used to compare the similarity of the plots. It is calculated as follows where P1 and P2 are probability functions generated by KDE.

$$\frac{\int P_1(r)P_2(r)dr}{\int P_1^2(r)dr \quad \int P_2^2(r)dr} \qquad (5)$$

The integrated product of two densities used to calculate the NIP is shown below where the sample size is represented by n, kernel bandwidth is h, d is the number of dimensions and r represents the position of each of the points in each of the probability functions.

$$\int P_1\,(r)P_2(r)dr = \frac{1}{n_1 n_2}\sum_{k=1}^{n_1}\sum_{l=1}^{n_2}\prod_{\alpha=1}^{d}\left(\int \frac{1}{h_1 h_2 2\pi}\, e^{\left[-\frac{(r_\alpha - r_{k\alpha})^2}{2h_1^2} - \frac{(r_\alpha - r_{l\alpha})^2}{2h_2^2}\right]} dr_\alpha\right)$$

$$= \frac{1}{n_1 n_2}\sum_{k=1}^{n_1}\sum_{l=1}^{n_2}\prod_{\alpha=1}^{d}\frac{1}{\sqrt{h_1^2+h_2^2\sqrt{2\pi}}}\, e^{\left[-\frac{(r_{k\alpha} - r_{l\alpha})^2}{2(h_1^2+h_2^2)}\right]}$$

$$= \frac{1}{n_1 n_2}\left[(h_1^2+h_2^2)2\pi\right]^{-\frac{d}{2}}\sum_{k=1}^{n_1}\sum_{l=1}^{n_2} e^{[-\frac{\|r_k - r_l\|^2}{2(h_1^2+h_2^2)}]} \qquad (6)$$

For Ramachandran plots point r represents each dihedral angle. Rather than the distance between two points being measured by Euclidean distance, Fishers approximation for adapting KDE to circular variables was used. If two Ramachandran plots are similar the NIP comparing their probability distributions will be close to one.

4.8.3 NIPs for the restrained simulations

The results from the 16 c(X^{1}GGX^{2}GG) restrained simulations (where X^{1}X^2 = AA, QQ, VV, AQ, QA, AV, VA, QV, VQ, FF, FA, AF, FQ, QF, VQ and QV) show the dihedral angles that will occur in a particular cyclic hexapeptide conformation when chiral amino acids are at the i/i+3 positions. The size of each conformation alters slightly depending on which amino acids are present at the i and i+3 positions. However no matter which amino acids are used at the i/i+3 positions the same Ramachandran positions are occupied for a particular turn type combination despite the range of amino acids tested (Figure 71). The preferred dihedral angle values for the i/i+3 positions can therefore be extracted from the restrained simulations for a given turn type conformation.

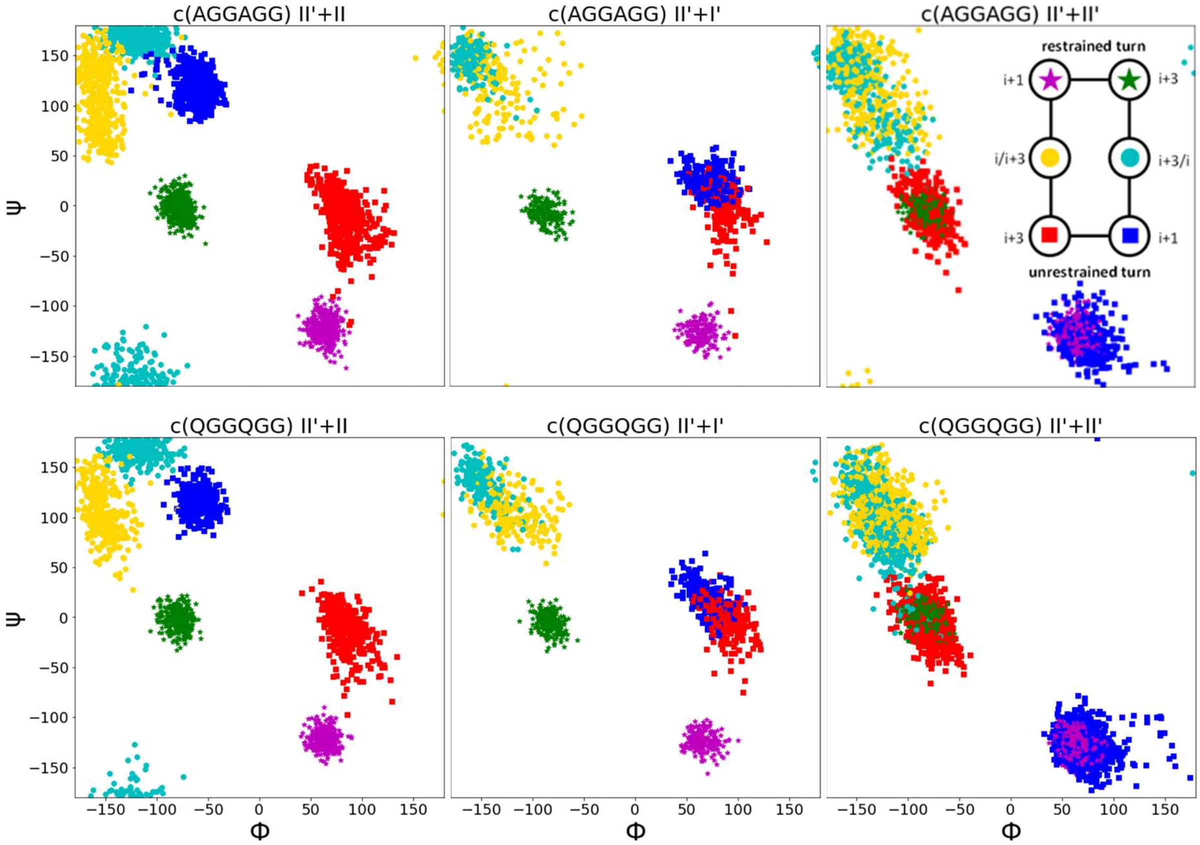


Figure 71: Ramachandran plots of the c(AGGAGG) and c(QGGQGG) type II' restrained simulation clusters. Alanine and glutamine occupy the same Ramachandran space in a particular conformation.

The energy difference between conformations due to the structure of the backbone of the peptide should remain the same when different amino acids are seen at the i/i+3 positions. However different amino acids have different ϕ and ψ dihedral angle preferences. To test whether these preferences, rather than specific side chain interactions, are the cause of the remaining energy differences between the conformations seen in the restrained simulations when different amino acid combinations are used at the i/i+3 positions, NIPs were used. The Ramachandran distribution for an amino acid in the absence of the constraints of a cyclic peptide is used as a model for its preferred dihedral angle preferences which are compared with the observed dihedral angles necessary for the amino acid to occupy a given position in a cyclic peptide. The assumption was made that a higher NIP for a particular i/i+3 dihedral angle distribution compared to the NIP values for the other positions that the amino acid could occupy in the different turn combinations would mean it is more favourable for the amino acid to be in that position. Therefore a higher percentage of that conformation may be seen in the restrained simulation. If this is the case it could be used to help predict the conformation of cyclic peptides. The restrained simulations offer a simplified system to test the NIPs on as the amino acids are in fixed registers so cannot change positions within a β-turn and only two amino acids are changed each time. Therefore just the NIP of the overlap of the two i/i+3 positions dihedral angles with the intrinsic preference of the amino acids which occupy those positions in the restrained simulations may predict the size of the cluster of that conformation that will occur. The remaining amino acids which make up the peptide in the restrained simulations are all glycine so as they are kept constant do not need to be factored in to get a value associated with the peptide adopting a particular conformation.

There are ten possible conformations for the c(XGGXGG) sequences made up of the 4 turn types I, II, I' and II'. This gives 20 Ramachandran plots in total: the two i/i+3 positions for each of the 10 possible conformations. The overlap of each of these 20 plots with the Ramachandran plot for each of the 19 naturally occurring chiral amino acids can be used to generate a NIP. Amino acids with a greater overlap with a position in a conformation will give a higher NIP value. The i/i+3 dihedral angles for each turn type combination seen in the c(AGGAGG) simulations were used to generate the NIP. The i/i+3 dihedral angles for combinations which did not occur in the restrained simulations were taken from the double restrained simulations.

The NIP values for a given amino acid (aa) being in a given position (p) in a given conformation (C), $NIP_{(aa, p, C)}$ were calculated. The $NIP_{(aa, p, C)}$ considers the similarity between the amino acids preferred Ramachandran distribution with the observed Ramachandran distribution of one of the i/i+3 positions in a conformation. The $NIP_{(aa, p, C)}$ values were then tested to see if they may help predict the conformation of cyclic hexapeptides. A higher $NIP_{(aa, p, C)}$ (closer to 1) means the amino acid has a greater overlap in the Ramachandran region necessary to achieve a position in a conformation. It may therefore be inferred, due to the closer similarity between the probability distributions, that the amino acid is lower energy in this particular turn combination thus contributing to the overall structure of the peptide.

4.8.4 Choice of database for generating NIPs

It is important for the right dataset to be used to obtain the Ramachandran plot showing the dihedral angle preferences for each amino acid that will be used to generate the $NIP_{(aa, p, C).}$ If an amino acids Ramachandran plot is influenced by external factors such as secondary structure the NIP it produces will not reflect the value intrinsic to that amino acid but that amino acid in a specific context. For example the Ramachandran plot of Alanine found in β-sheets would be very different to that of alanine found in α-helices. Therefore the source of the Ramachandran plot used to generate the $NIP_{(aa, p, C)}$ has the potential to give very different results so must be chosen carefully or bias in the plot could lead to incorrect predictions.

4.8.4.1 Hairpin dataset

The i/i+3 positions in the hairpin dataset extracted from the Loop Database (see Chapter 3) occupy very similar Ramachandran regions to those seen in the i/i+3 positions of the cyclic hexapeptide. They may therefore be suitable to be used as a model for the i/i+3 positions in a cyclic hexapeptide. For each amino acid the dihedral angles where it was found in either the i or i+3 position of the hairpin were extracted and combined to generate a Ramachandran plot. The number of each amino acid found at each position is shown in Table 27. The Ramachandran plot of the combined positions was used to obtain $NIP_{(aa, p, C)}$ values. The values for glycine were not included as glycine occupies a different region of the Ramachandran plot leading to different preferred conformations. When searching for β-hairpins within the Loop Database, the presence of a hydogen bond between the i and i+3 positions of the β-turn was searched for. This means proline is also not included at any of the i/i+3 positions in the dataset as it's missing the amide hydrogen necessary for the i/i+3 hydrogen bond to form. Additionally the proline structure naturally introduces a kink into the peptide chain meaning it is generally found at the i+1 or i+2 turn position of β-turns.[114]

Amino acid	i population	i+3 population	Total population
A	517	365	882
C	153	90	243
D	715	361	1,076
E	808	914	1,722
F	483	319	802
H	432	367	799
I	933	433	1,366
K	701	1,422	2,123
L	722	520	1,242
M	145	151	296
N	446	361	807
Q	358	615	973
R	575	869	1,444
S	684	462	1,146
T	759	654	1,413
V	1,288	792	2,080
W	159	130	289
Y	624	353	977
total	10,870	10,864	21,734

Table 27: The amino acids found in the hairpin dataset at the i/i+3 positions.

Figure 72 shows the $NIP_{(aa, p, C)}$ values generated from the hairpin dataset. Each of the 10 turn combinations has two positions labelled A and B. Position A (pA) is at the i position of the first listed turn and position B (pB) is the i+3 position of the first listed turn. For example pA of the I+II conformation is the i position of the type I turn/the i+3 position of the type II turn and pB is at the i position of the type II turn/the i+3 position of the type I turn.

The $NIP_{(aa, p, C)}$ values vary from ~0 to ~1. Values close to zero show that there is very little overlap between the Ramachandran values for that particular amino acid in the hairpin dataset with the required position for a given conformation in a cyclic peptide. By choosing the amino acid with the highest $NIP_{(aa, p, C)}$ value for each position in a conformation the most favourable amino acids for that conformation (based on overlap in Ramachandran distributions) can be determined. For example for the I+II’ conformation position A, the greatest $NIP_{(aa, p, C)}$ value is seen for threonine with a value just below 0.6. This means for a I+II’ conformation, that based on the NIPs, threonine would be the best amino acid to have at the i/i+3 position going into the type I turn. The best amino acid to have at position B (the i/i+3 position coming out of the type I turn) is serine with a $NIP_{(aa, p, C)}$ value of approximately 0.4. Greater $NIP_{(aa, p, C)}$ are seen for threonine and serine in other conformations, such as the II+II’ conformation, which has a $NIP_{(aa, p, C)}$ just below 0.75 for threonine at the A position and serine at position B has a value of around 0.55. This means based on these $NIP_{(aa, p, C)}$ values a II+II’ conformation with threonine at position A (at the i position of the type II turn) and serine at position B (at the i+3 position of the type II turn) would be more favourable than a I+II’ conformation.

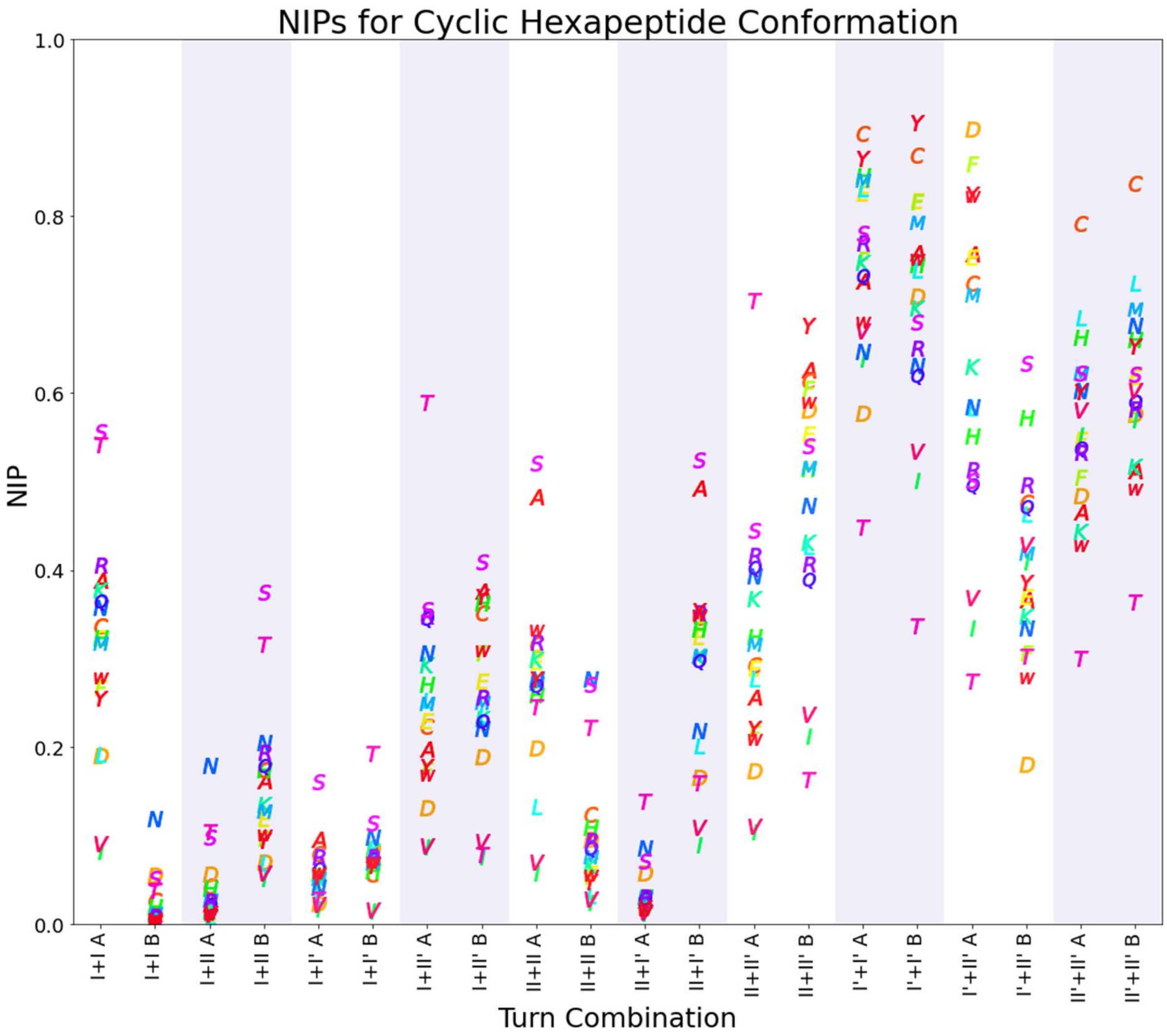


Figure 72: NIPs for each turn combination based on the overlap with the hairpin dataset.

The values close to one show that these amino acids occupy very similar distributions in the hairpin datset as they would need to for a particular cyclic hexapeptide conformation. The highest values are generally seen for conformations containing type I' and II' turns. These are the most common turn types seen in the hairpin dataset. This may mean bias in the dataset towards these turn types is being seen. However the I'+I', I'+II' and II'+II' conformations which give the greatest $\text{NIP}_{(aa,\ p,\ C)}$ values often appear as the lowest energy clusters in the restrained simulation energy diagrams. The high $\text{NIP}_{(aa,\ p,\ C)}$ values for these conformations may therefore be one of the reasons why these turn types are relatively low energy.

Similar $\text{NIP}_{(aa,\ p,\ C)}$ values for a particular amino acid between turn combinations (e.g. valine at pA of the I+I conformation and pA of the I+II' conformation each have a NIP value of approximately 0.1) show how the Ramachandran preference for some amino acids could lead to only very small differences in energy between conformations.

β-branched amino acids (valine, isoleucine and threonine) tend to give lower $\text{NIP}_{(aa,\ p,\ C)}$ values. Such β-branched amino acids are frequently found at the i/i+3 positions, as shown by the most common turn sequences in the Loop Database where 6 of the 10 most commonly occurring β-hairpin sequences have either valine or isoleucine at the i position of a β-turn. This could mean that although these amino acids are frequently seen at the i position in β-hairpins this is no longer a favourable position for them in cyclic peptides despite the similarities between the two, meaning β-hairpins are not a good model for the β-turn found within cyclic peptides. Alternatively, as it is the relative $\mathrm{NIP}_{(aa,\ p,\ C)}$ of all possible options that may determine conformation, despite a low $\mathrm{NIP}_{(aa,\ p,\ C)}$ for these positions, an even lower overlap could be seen for these amino acids with the i+1 and i+2 positions of different β-turns making occupying the i/i+3 positions preferable in comparison. The hairpin dataset for the i/i+3 positions only covers a small portion of the β-region of the Ramachandran plot, using a different dataset which shows the dihedral angle preferences of an amino acid covering the full Ramachandran space may help show the balance for an amino acids preference between different regions which may then change the order of the amino acids $\mathrm{NIP}_{(aa,\ p,\ C)}$.

The occurrence of some amino acids within this dataset is relatively low, for example there are only 243 occurrences of cysteine at either a i or i+3 position in a β-hairpin within the dataset. A low data sample may mean the KDE of the data does not generate a probability estimate that reflects the actual values and would subsequently generate an inaccurate $\mathrm{NIP}_{(aa,\ p,\ C)}$. To test if the sample sizes were sufficient, NIP values comparing the Ramachandran distribution of randomly selected resampled data were generated. If the resampled data retains a similar distribution a NIP value close to 1 will be seen and the dataset is large enough to reflect the actual Ramachandran distribution for that amino acid. This means similar $\mathrm{NIP}_{(aa,\ p,\ C)}$ values would be generated if the resampled data was used to look at the overlap with each turn conformations position. For cysteine, methionine and tryptophan (the three smallest datasets) the NIP decreases with sample size a lot more rapidly than the other amino acids when increasingly smaller proportions of the data is resampled (Table 28). A NIP below 0.9 is seen for these datasets with 50% resampling. It would therefore be preferable to use larger datasets to generate the $\mathrm{NIP}_{(aa,\ p,\ C)}$ values.

Amino acid	Total population before resampling	NIP value between random samples		
		90% Resampling	50% Resampling	10% Resampling
A	882	0.992	0.943	0.677
C	243	0.976	0.784	0.341
D	1,076	0.992	0.941	0.586
E	1,722	0.997	0.971	0.803
F	802	0.994	0.942	0.676
H	799	0.995	0.963	0.675
I	1,366	0.999	0.991	0.917
K	2,123	0.997	0.981	0.855
L	1,242	0.996	0.970	0.781
M	296	0.979	0.890	0.377
N	807	0.992	0.915	0.629
Q	973	0.993	0.950	0.700
R	1,444	0.995	0.970	0.726
S	1,146	0.993	0.953	0.677
T	1,413	0.997	0.972	0.799
V	2,080	0.999	0.990	0.895
W	289	0.982	0.895	0.484
Y	977	0.995	0.967	0.735

Table 28: The NIP value comparing different random samples of the hairpin dataset.

Although showing similar Ramachandran distributions the hairpin dataset will differ from the preferred regions in a cyclic hexapeptide. A large proportion of β-turns are involved in binding interfaces.[112] The hairpins in the database may therefore be interacting with ligands or the remaining protein structure, modifying their Ramachandran distribution. For example the cysteines could be forming a disulfide bond with other cysteines in the protein which could alter the conformation of the peptide and the dihedral angles seen for cysteine. Other amino acids could be involved in similar interactions within the dataset, such as histidine binding to metal centres, which would bias the data. The hairpin datas*et al*so contains predominately type I' and II' turns which may bias the distribution seen. A Ramachandran distribution that reflects the amino acids distribution in the absence of structural influences would therefore be useful.

4.8.4.2 MD and Database Ramachandran Distributions

Two new datasets were generated to calculate NIP$_{(aa,\ p,\ C)}$ values without the small sample sizes and contextual influence seen in the hairpin dataset. A sample of 3,000 dihedrals for each type of amino acid was taken randomly from the Loop Database rather than just from β-hairpins. The first two amino acids at the beginning of a loop were ignored in case they retained any of the influence of the secondary structure preceding the loop. The second dataset was made by running MD simulations on Ac-GGXGG-NH$_2$ where X is one of the 20 naturally occurring amino acids and then extracting the dihedrals for each amino acid. It is hoped that this sequence, a short peptide with the flexible achiral glycine, will give the Ramachandran distribution of each amino acid without the influence of secondary structure. Previously this and similar sequences have been used to obtain the Ramachandran distribution of amino acids,[343-349] but not using the RSFF2 forcefield which has been shown to be the best forcefield for modelling cyclic peptides.[243]

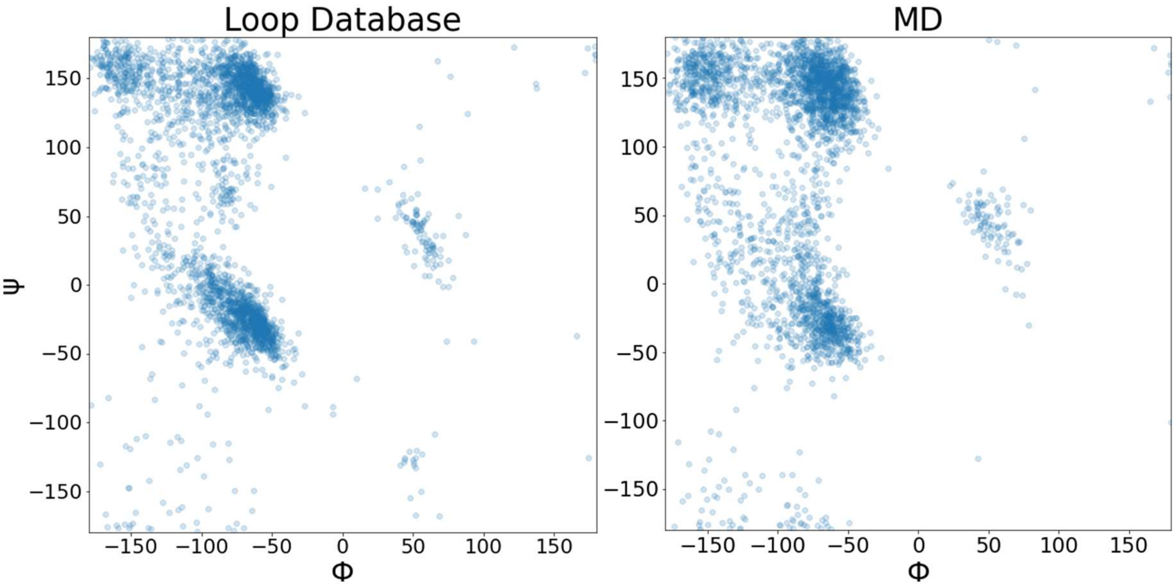


Figure 73: The database and MD Ramachandran plot of Alanine.

Both these datasets gave very similar results (Figure 73). Generating a NIP looking at the overlap between alanine's Ramachandran plot from the database and MD gives a value of 0.95 which is close to 1 showing the similarity of the two datasets. Similar NIP values are seen for the other amino acids. Figure 74 shows the NIP$_{(aa,\ p,\ C)}$ values generated by the overlap between the Ramachandran distribution of each amino acid in the database dataset compared with the i/i+3 angles for each turn combination and Figure 75 shows the equivalent NIP$_{(aa,\ p,\ C)}$ values produced from the MD dataset.

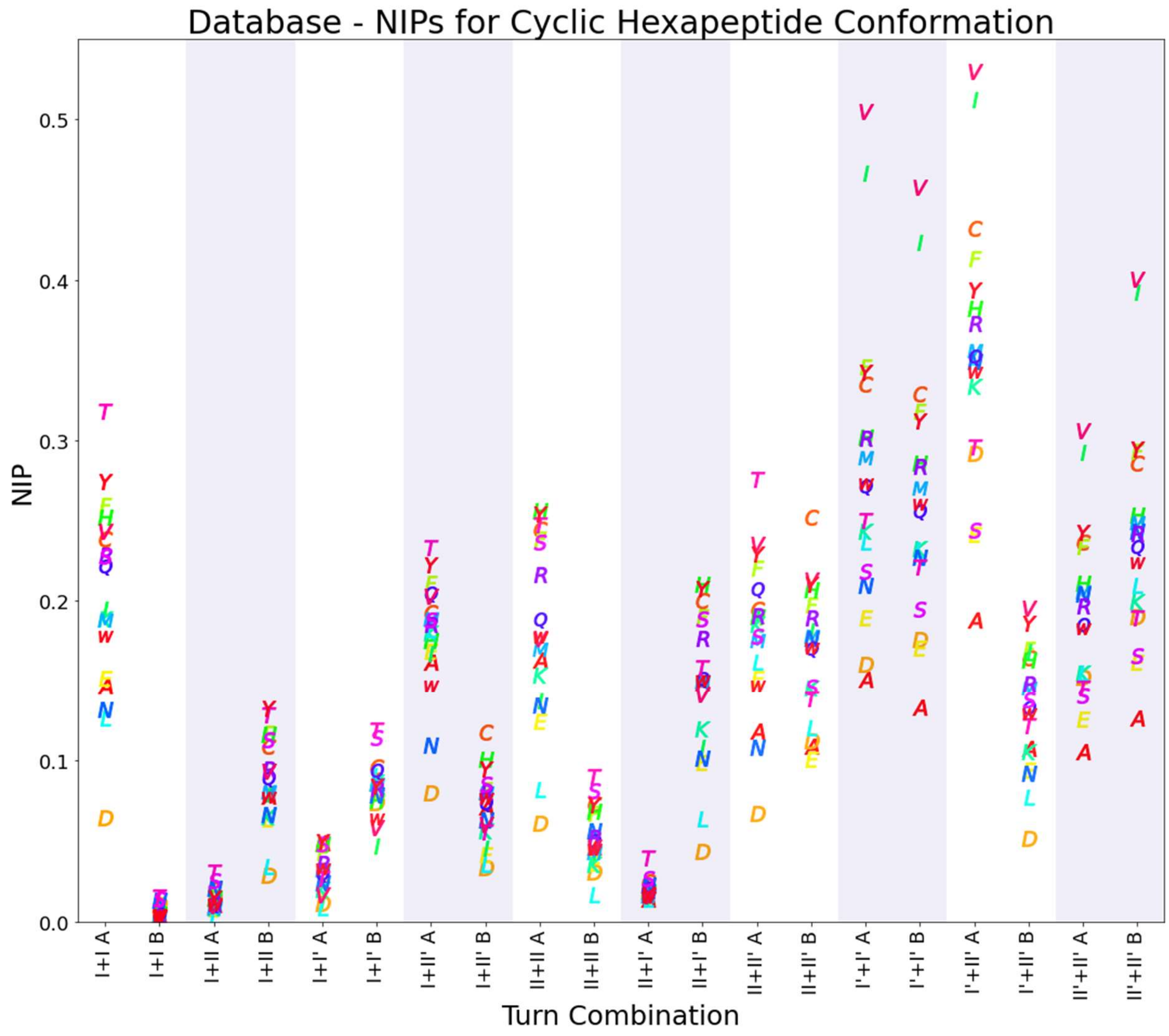


Figure 74: NIPs for each amino acid in each turn combination generated from the database dataset.

For both datasets the NIP$_{(aa, p, C)}$ values cover a range of ~0 to ~0.5, lower than seen in the hairpin dataset as the amino acids Ramachandran distributions cover a larger area so there is less overlap. The relative NIPs of the different positions in the different conformations are similar to those of the hairpin dataset: I+I position B has the lowest NIP$_{(aa, p, C)}$ values in the datasets and the I'+I', I'+II' and II'+II' combinations have on average greater NIP$_{(aa, p, C)}$ values than the other combinations. For the MD and database datasets within a conformation, aspartic acid and asparagine generally give the lowest NIP values. Both these amino acids, based on the β-turns found in the Loop Database, are generally found at the i+1 or i+2 position of β-turns. For the I'+I', I'+II' and II'+II' conformations, which were generally the lowest energy in the c(XGGXGG) restrained simulations, valine and isoleucine have the greatest NIP values. This may mean having one of these amino acids at such a position in a cyclic peptide will lead to a particularly stable conformation.

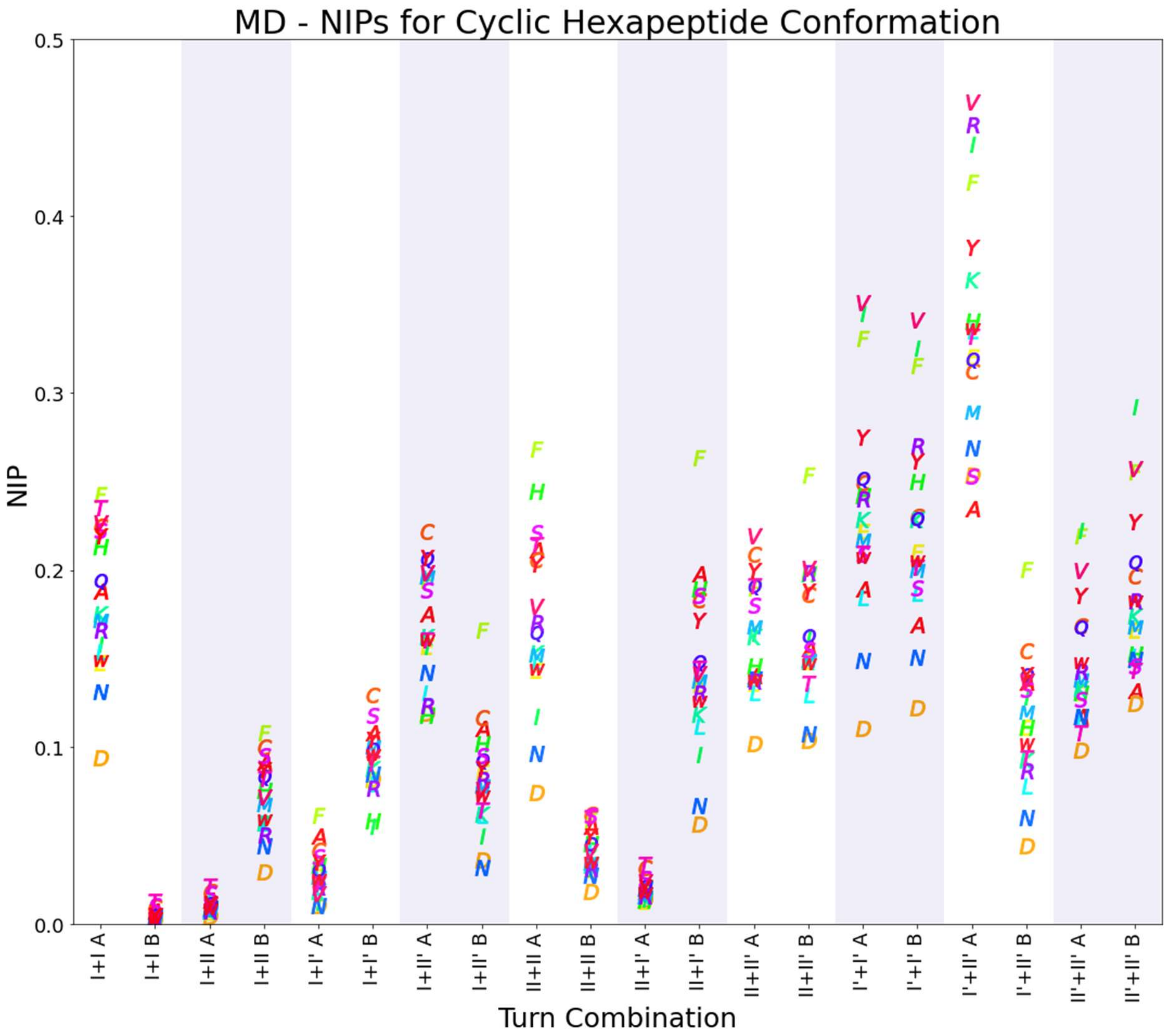


Figure 75: The NIPs generated from the MD of Ac-GGXGG-NH₂.

The differences in the NIP$_{(aa, p, C)}$ generated from either the database dataset or the MD dataset are shown in Figure 76. All values differ by less than 0.15 showing the similarity of the datasets Ramachandran distributions for each amino acid. The MD dataset is based on RSFF2, a forcefield modified from the Amber99SB forcefield with improved torsional and non-bonded parameters for some amino acids to better align with the dihedrals seen in a database of the coil residues of proteins.[242] These modifications allow for better modelling of various structures including those that form predominately α-helices and β-sheets as well as better reproducing the ^{3}J values for small dipeptides and the structures of cyclic peptides.[70, 242, 243, 289] The Loop Database and RSFF2 therefore might be expected to give similar Ramachandran distributions as RSFF2 is based on a the protein coil library (both are based on databases which omit α-helices and β-sheets from the Ramachandran distributions of the amino acids).[350]

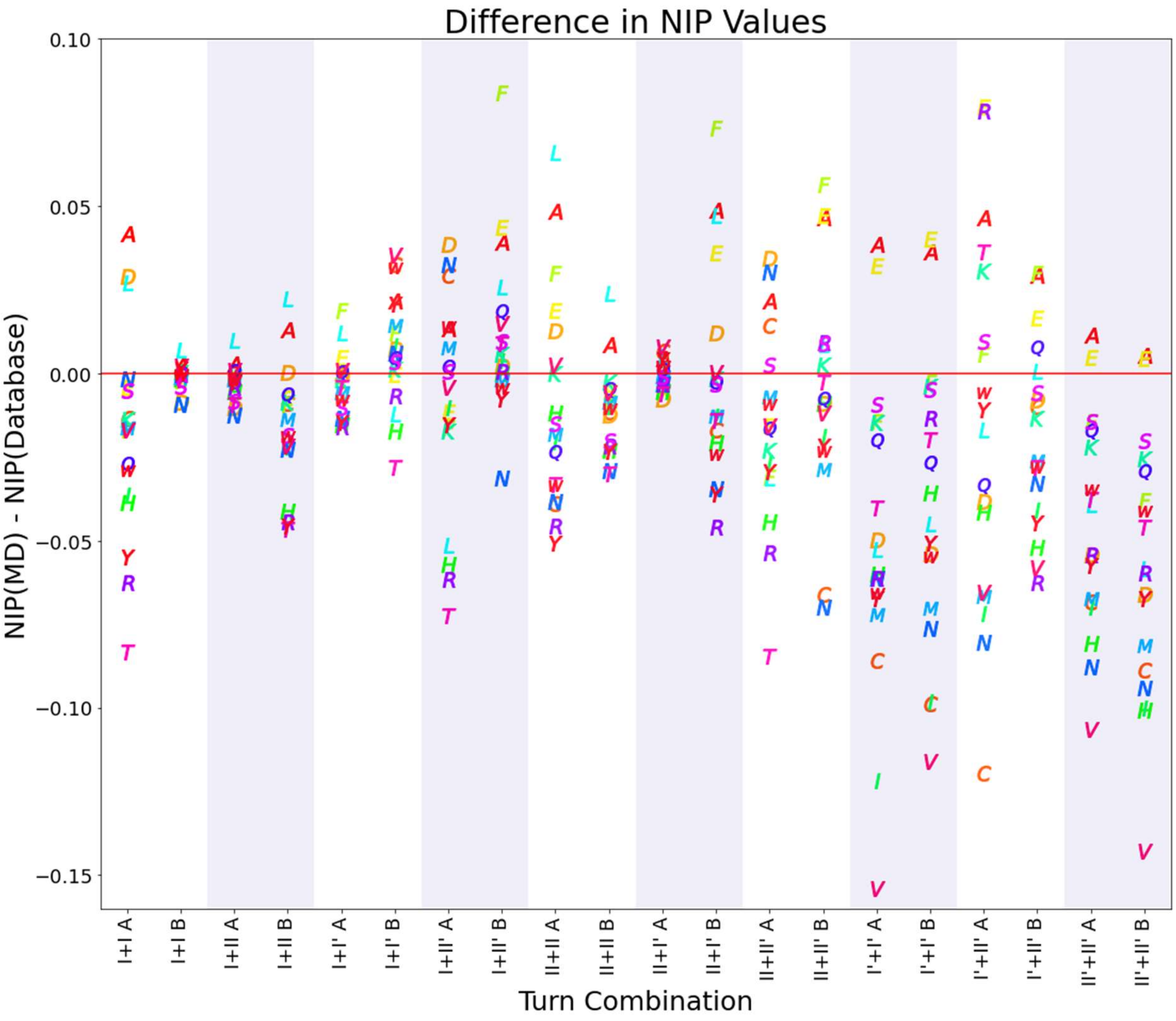
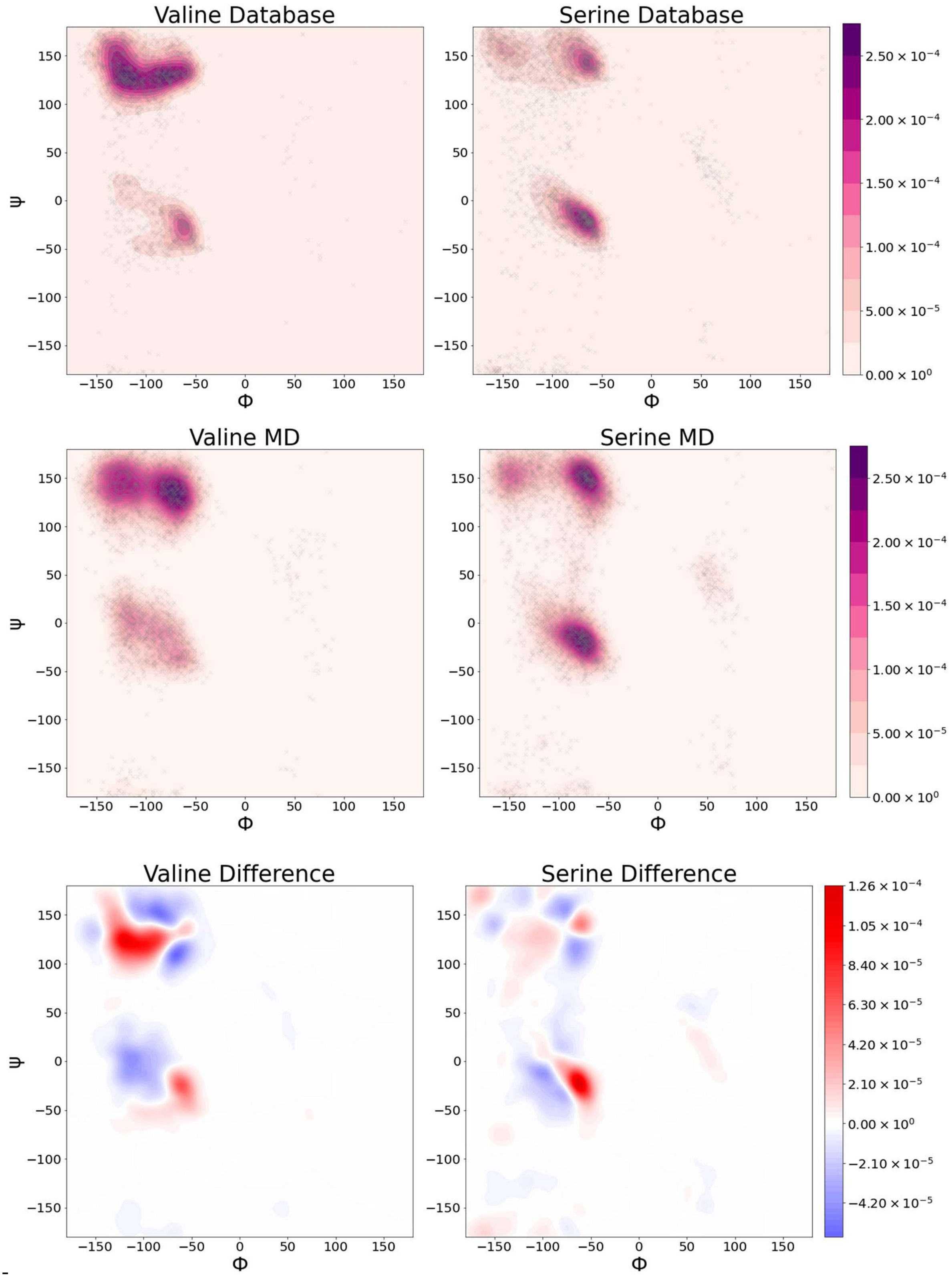


Figure 76: The difference in the NIPs generated from the database or MD datasets.

Some positions show more similarity than others such as the I+I position B, I+II position A and II+I' position A where all the amino acids are closely centred around zero with very similar NIP$_{(aa, p, C)}$ values seen between the MD and database datasets. These three positions all have their dihedral angles predominately at the bottom lefthand corner of the Ramachandran plot, with negative ψ values close to -180°. These positions in general show the least difference between the NIP generated from the MD and database datasets. This is a relatively sparsely populated region of the Ramachandran plot, generally being unfavourable for L-amino acids.

Other positions show a more spread-out distribution of the amino acids. Similar differences are seen for the I'+I', I'+II' and II'+II' positions as these conformations all show similar Ramachandran distributions leading to similar NIP$_{(aa, p, C)}$ differences between the datasets. Valine generally shows the largest difference between the MD and database datasets for these positions. Serine generally shows very little difference across all positions. The difference between two probability distributions can be looked at by subtracting one probability distribution away from the other. Looking at the difference in the probability distributions generated by KDE on the MD and database datasets for serine and valine (Figure 77), the main difference for valine is seen in the β-region of the Ramachandran plot whereas for serine it is in the α region. As the i/i+3 positions in the cyclic peptide conformations require dihedral angles predominately in the β-region of the Ramachandran plot valine shows a greater difference in the NIPs generated using the MD or database datasets.



-

Figure 77: The probability distributions of valine and serine for the MD and database datasets and the difference between the distributions.

The small differences between the MD and Loop Database probability distributions leads to the slight differences in NIP$_{(aa,\ p,\ C)}$ values. This could lead to different predictions for cyclic peptide conformation. The NIP$_{(aa,\ p,\ C)}$ values are therefore very sensitive to the dataset used. However the trend in the NIP$_{(aa,\ p,\ C)}$ values for each conformation remains similar between the two datasets.

The MD dataset was chosen to test if the NIP$_{(aa,\ p,\ C)}$ values could be used to predict the amount of each conformation in the restrained simulations. The database dataset may still retain influence of the protein environment altering the amino acids Ramachandran distribution away from intrinsic distribution. Additionally the i/i+3 dihedral angle distributions in the cyclic hexapeptide conformations were generated by MD using the RSFF2 forcefield, the same forcefield used to generate the MD datasets Ramachandran distributions.

4.8.5 The I+I Restrained Simulations

In order to obtain the lowest energy i/i+3 dihedral angles for the I+I conformation a double restrained simulation was carried out as the conformation was too high energy to occur in the type I singly restrained simulation. These simulations were performed on c(XGGXGG) where X=A, F, Q and V. For each of these four sequences two clusters of approximately equal size were seen in the I+I double restrained simulations which were equivalent to each other. This is because the same amino acid is present at each of the i/i+3 positions. If different amino acids are used the two clusters would no longer be equivalent and it would likely be preferable for one amino acid to be in the yellow i/i+3 position and the other at the cyan i/i+3 position as seen in cluster 1 in Figure 78. Therefore only one cluster would likely be seen in the I+I restrained simulation if two different amino acids were used at the i/i+3 positions or the two clusters may appear but one would be lower energy than the other so would be larger.

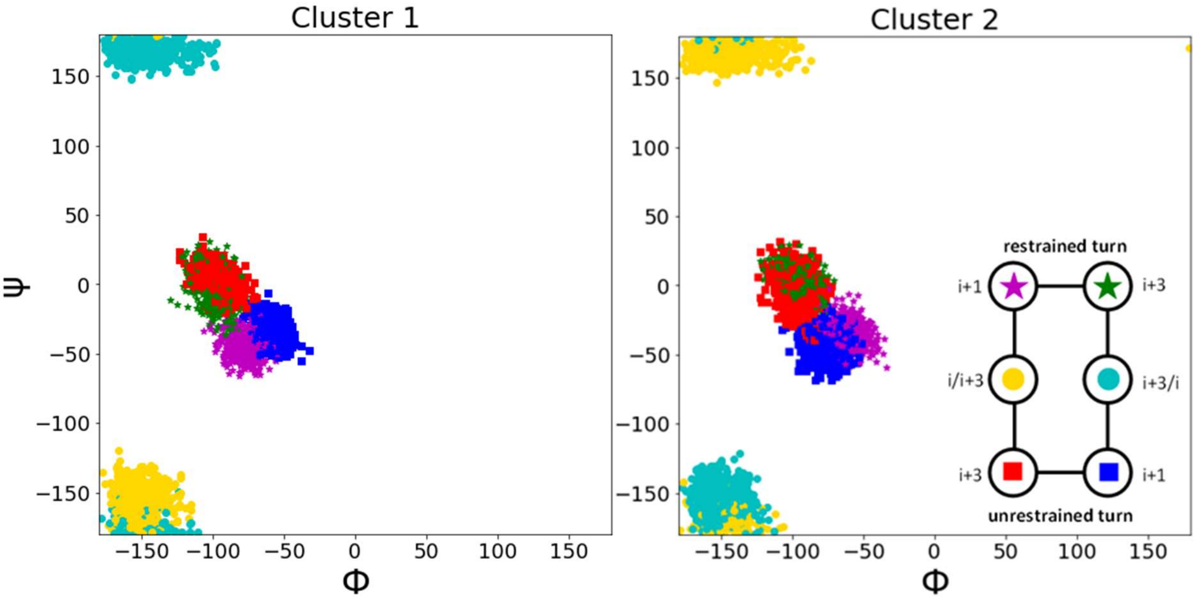


Figure 78: The Ramachandran plots of the two equivalent clusters seen in the I+I double restrained simulation of c(AGGAGG).

The I+I double restrained simulations were therefore repeated on c(X^1GGX^2GG) where X^1X^2 = AQ, AV, AF, QV, QF and VF. In each of the c(X^1GGX^2GG) simulations the clusters have the same dihedral angles as previously seen for the I+I conformation but now generally only one major cluster is seen in the restrained simulations rather than two clusters of equal size. It is now lower energy to have one of the two amino acids at one of the i/i+3 positions and the other amino acid at the other position.

In order to determine if it is possible to predict the stability of an amino acid when occupying a particular i/i+3 position, the overlap in the amino acids Ramachandran plot from the Ac-GGXGG-NH_2 simulations with each of the Ramachandran plots of the i/i+3 positions was looked at. Table 29 shows the NIP values for each of the i/i+3 positions in the I+I conformation for A, Q, V and F. Position B (yellow in cluster 1 in Figure 78) shows very low $NIP_{(aa, p, C)}$ values which may be one of the reasons why this cluster was too high energy to appear in the type I singly restrained simulation. Position A (cyan in cluster 1 in Figure 78) has greater $NIP_{(aa, p, C)}$ values due to a greater overlap between the amino acids Ramachandran plot with the Ramachandran plot of this position. Phenylalanine has the greatest $NIP_{(aa, p, C)}$ value followed by valine with glutamine and alanine having slightly lower $NIP_{(aa, p, C)}$ values.

	Position A	Position B
A	0.188	0.007
Q	0.194	0.004
V	0.227	0.005
F	0.243	0.004

Table 29: The NIP values for each of the two i/i+3 positions in the I+I conformations generated using the MD dataset.

For each conformation seen in the c(X^1GGX^2GG) I+I doubly restrained simulations the $NIP_{(aa, p, C)}$ values were compared to which amino acids occupied each of the i/i+3 positions. Table 30 shows which amino acid occurs at each of the two i/i+3 positions in the major conformation seen in the simulations. If a minor conformation occurs in the simulation, it has the same Ramachandran positions but the amino acids occupying each of the two i/i+3 positions (pA and pB) are reversed.

Sequence	Amino Acid at Position A	Amino Acid at Position B	% Major conformation	Position A difference in $NIP_{(aa, pA, C)}$	Position B difference in $NIP_{(aa, pB, C)}$
c(AGGQGG)	Q	A	91	0.006	0.003
c(AGGVGG)	V	A	99	0.039	0.002
c(AGGFGG)	F	A	100	0.055	0.003
c(QGGVGG)	V	Q	73	0.033	-0.001
c(QGGFGG)	F	Q	55	0.049	0.000
c(VGGFGG)	V	F	69	-0.016	-0.001

Table 30: The c(X^1GGX^2GG) I+I double restrained simulations major conformation. Each amino acid more favourably occupies one of the i/i+3 positions. If a minor conformation occurs the positions of the amino acids are reversed. The difference in NIP value calculated for the two positions is also shown.

For the sequences containing an alanine, alanine is always at pB. Alanine has the lowest $NIP_{(aa, p, C)}$ value for pA and the highest $NIP_{(aa, p, C)}$ value for pB. The NIPs can therefore be used to predict that alanine would be at pB for these sequences with the other amino acid at pA. Only small differences are seen in the $NIP_{(aa, p, C)}$ for pB as there is very little overlap in the Ramachandran plots in that region. For pA then the larger the difference in the $NIP_{(aa, p, C)}$ between the two amino acids the larger the size of the major conformations cluster. For example the difference between the alanine and glutamine $NIP_{(aa, p, C)}$ at pA is 0.006. The major conformation seen in the simulation of c(AGGQGG) makes up 91% of the clustered data with a small cluster (9%) where alanine is at pA and glutamine at pB. Valine has a greater $NIP_{(aa, p, C)}$ difference from alanine at pA of 0.039. The major conformation (with valine at pA and alanine at pB) now makes up a larger proportion of the clustered data. It may be that the larger the difference between the two $NIP_{(aa, p, C)}$ values the more favourable it is for an amino acid to be at one position meaning it makes up a larger proportion of the clustered data.

For the remaining sequences not containing alanine (c(QGGVGG), c(QGGFGG) and c(VGGFGG)), the major conformation makes up a smaller proportion of the clustered data compared to the sequences containing alanine. The amino acid with the greater NIP$_{(aa,\ p,\ C)}$ value at position A is still seen at pA in the major cluster, with the exception of c(VGGFGG) where valine is at pA in the major cluster and phenylalanine at pB despite phenylalanine having a greater NIP$_{(aa,\ p,\ C)}$ value than valine for pA. This indicates there are other factors contributing to the most stable structure besides the preferred Ramachandran positions of the amino acids.

As the results of the alanine containing sequences indicate that in some cases the NIP$_{(aa,\ p,\ C)}$ values may be able to be used to predict the conformation the use of NIPs to predict the results of different restrained simulations was further investigated.

4.8.6 Using NIPs to Predict the Results of the Restrained Simulations

The NIP$_{(aa,\ p,\ C)}$ values generated by the Ramachandran distributions of the MD dataset and the i/i+3 dihedral angles seen for each turn combination in the restrained simulations show which amino acids have the most similarity between the positions they must occupy in a cyclic hexapeptide conformation and the intrinsic preference of the amino acids. A correlation was looked for between the NIP$_{(aa,\ p,\ C)}$ values and the amount of each conformation which appeared in the restrained simulations to see if the Ramachandran preferences of the amino acids could be used to help predict the conformation of cyclic peptides.

Each turn type combination has two i/i+3 positions: position A (pA) and position B (pB). For a given turn type combination one position does not occur without the other with both positions needed to describe the conformation seen in the c(XGGXGG) restrained simulations. The two NIP$_{(aa,\ p,\ C)}$ values associated with a turn type combination therefore need to be combined to give a single value associated with that conformation. That is the NIP for an amino acid X at position A in a given turn type combination, C, (NIP$_{(aaX,\ pA,\ C)}$) needs to be combined with the NIP for an amino acid Y at position B in the same conformation (NIP$_{(aaY,\ pB,\ C)}$) to give an overall NIP value associated with the turn type combination. This overall NIP value (NIP$_{(aaX,\ aaY,\ C)}$) will give a single value associated with having amino acid X at position A and amino acid Y at position B in conformation C.

To obtain a NIP$_{(aaX,\ aaY,\ C)}$ value an average of the NIPs for each position in a given conformation (NIP$_{(aaX,\ pA,\ C)}$ and NIP$_{(aaY,\ pB,\ C)}$) was taken (equation 7). By combining the two NIP values for pA and pB for a particular conformation the overall NIP$_{(aaX,\ aaY,\ C)}$ value may be correlated with amount of this conformation that will occur in the restrained simulation. This NIP$_{(aaX,\ aaY,\ C)}$ value can then be compared with other NIP$_{(aaX,\ aaY,\ C)}$ values such as the same conformation occurring with two different amino acids each with different NIP$_{(aa,\ pA,\ C1)}$ and NIP$_{(aa,\ pB,\ C)}$ values.

$$NIP_{(aaX,aaY,C1)} = \frac{NIP_{(aaX,pA,C)}+NIP_{(aaY,pB,C)}}{2} \qquad (7)$$

As it is the relative energies of all conformations which can occur which determines the amount of each conformation seen, then if the NIP$_{(aaX,aaY,C)}$ is related to the most low energy conformation, it will also be the relative NIP$_{(aaX,aaY,C)}$ values of all possible conformations that will determine the amount of a conformation. The NIP$_{(aaX,\ aaY,\ C)}$ value therefore needs to be compared to the other NIP$_{(aaX,\ aaY,\ C)}$ values of the possible conformations which can occur. The NIP ratio was therefore calculated as the ratio of the NIP$_{(aaX,\ aaY,\ C)}$ compared to the sum of all the conformations in the restrained simulation (equation 8). This gives a value which takes into account all possible conformations that can occur with given amino acids. Outliers in the data extracted from the restrained simulations are not assigned to a cluster (less than 100% of the data is clustered) so a

NIP$_{(aaX, aaY, C)}$ value was also calculated for the unclustered positions. Although not occurring enough to form a cluster these unclustered points represent possible conformations of the peptide. A higher NIP ratio means that conformation may be more favourable than the other ones that are possible.

$$NIP\ ratio = \frac{NIP_{(aaX,aaY,C1)}}{NIP_{(aaX,aaY,C1)} + NIP_{(aaX,aaY,C2)} + \cdots NIP_{(aaX,aaY,Cx)}}$$

$$= \frac{NIP_{(aaX,aaY,C1)}}{\sum NIP_{(aaX,aaY,Cx)}} \qquad (8)$$

The plots showing the NIP ratio compared to the percentage occurrence of different clusters in the restrained simulations are shown in Figure 79. All the conformations that appear regularly across the different sequences are included. Some conformations appear in more than one restrained simulation. For example the I’+II’ simulation occurs in the type I’ restrained simulation with a type II’ turn appearing at the unrestrained end of the cyclic peptide. This is the same conformation as the II’+I’ cluster that appears in the type II’ restrained simulation where a type I’ turn appears at the unrestrained side. The combined i/i+3 positions for the I’+II’ and II’+I’ conformations give a NIP value of 0.99 showing they are equivalent despite appearing in separate restrained simulations.

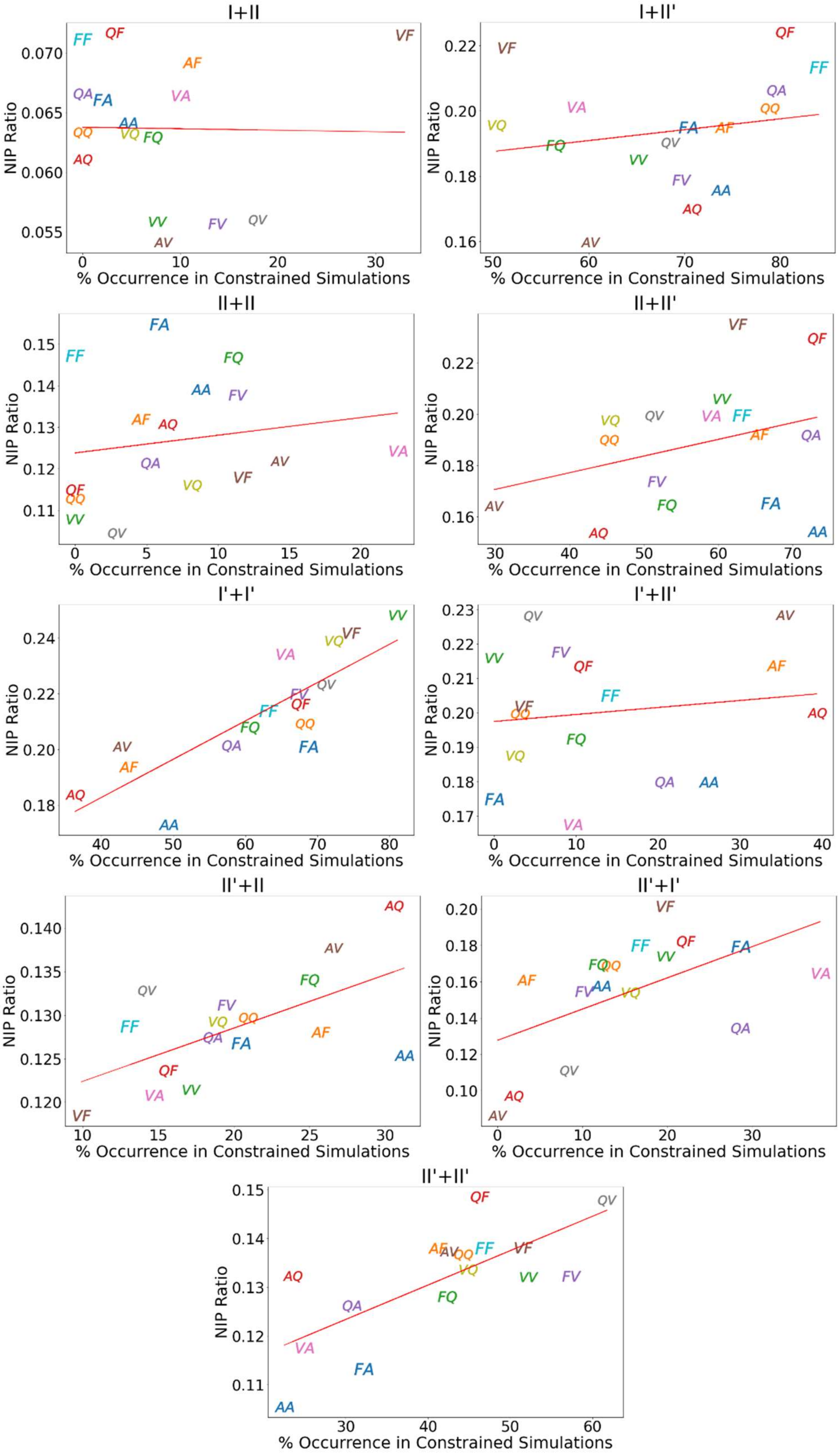


Figure 79: The NIP ratio compared to the % occurrence of conformations in the restrained simulations.

Although there is a general increase in the NIP ratio with the percentage of a conformation seen in the restrained simulations for most conformations, the graphs show that the NIP ratio would not be useful for predicting the conformation of the peptide as there is no consistent correlation between the NIP ratio and the amount of a conformation seen in the restrained simulation.

4.8.6.1 Difficulties of Adapting NIPs for use on Circular Variables

There are limitations in the implementation of KDE and therefore the NIP values used to look at the similarity between Ramachandran distributions. These limitations mainly arise from the difficulty of adapting these methods for use on circular variables and the kernel bandwidth selection. It may be that the $\mathrm{NIP}_{(\mathrm{aa,\ p,\ C})}$ values do not correlate well with the amount of each conformation seen in the restrained simulations due to KDE leading to inaccurate probability estimates of the Ramachandran plots.

The dihedrals angles are circular variables so Fishers modification of KDE was used to generate the probability distributions. There are however alternative methods to generate probability distributions of circular variables which may lead to more accurate results. The von Mices distribution (circular normal distribution) can be used for KDE which may lead to improved probability distributions but its current implementation in Python packages is limited and it is a more complex model requiring further approximations leading to increased computational time.[351] As the pA and pB dihedral angles in cyclic peptides occur at the boundaries of the Ramachandran plot using a von Mices distribution would have the greatest effect here and likely improve the results due to being able to better deal with the "wrap-around" effect of using circular variables than the Gaussian kernel used.[352] Transformation of the data to predominately lie in a region away from the boundary to minimise effects where the "wrapping around" of the data makes accurate probability distribution estimates difficult may be an alternative approach where Gaussian kernels could still be used. The data was therefore transformed so the pA and pB dihedral angles occur in the centre of the Ramachandran plot rather than the edges to determine if the current KDE implementations inability to accurately generate probability estimates of circular data is leading to errors in the NIP values. The Ramachandran plots of the amino acids from the MD dataset were likewise transformed prior to KDE (Figure 80). Although the MD dataset after transformation now has the largest density (what was the α-region of the Ramachandran plot) close to the edge, as this region shows little overlap with the pA and pB positions, the ability of the Gaussian kernels used to deal with the "wrap-around" effect should hopefully be minimised.

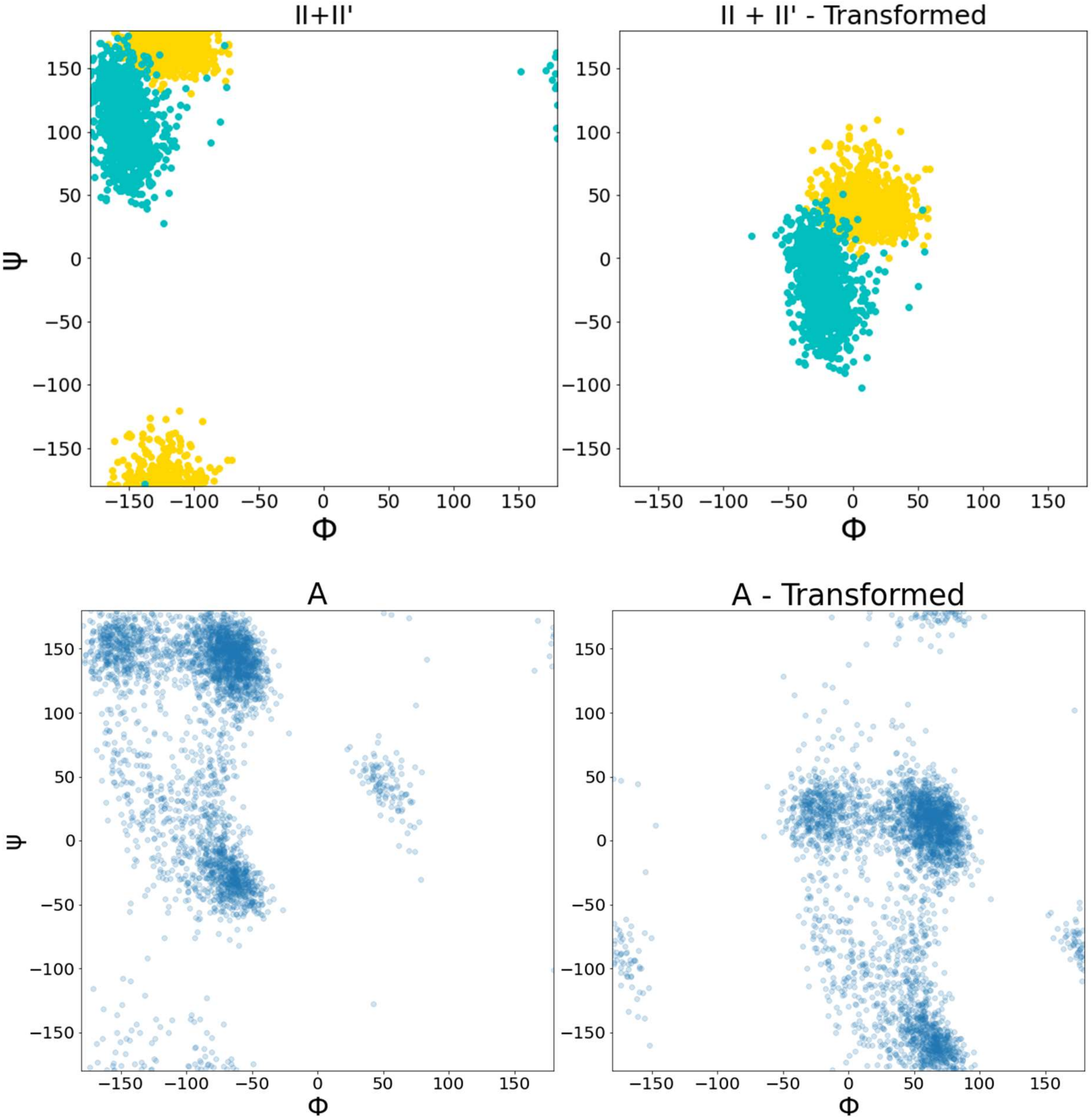


Figure 80: Transformed Ramachandran plot to centre the pA and pB positions. 130° was added to the φ values then 360° taken away from values greater than 180°. 130° was subtracted from the ψ values and 360° added to values below -180°.

The NIP ratios were recalculated using the $\mathrm{NIP}_{(aa,\ p,\ C)}$ values generated from the transformed data. Although transforming the data altered the $\mathrm{NIP}_{(aa,\ p,\ C)}$ values slightly, showing how the current implementation could be better adapted to circular variables, the same trends in the $\mathrm{NIP}_{(aa,\ p,\ C)}$ values and therefore similar NIP ratios were seen (Figure 81). Therefore it is unlikely the NIP ratios do not correlate well with the amount of the conformation seen due to the difficulty of KDE for use on circular variables.

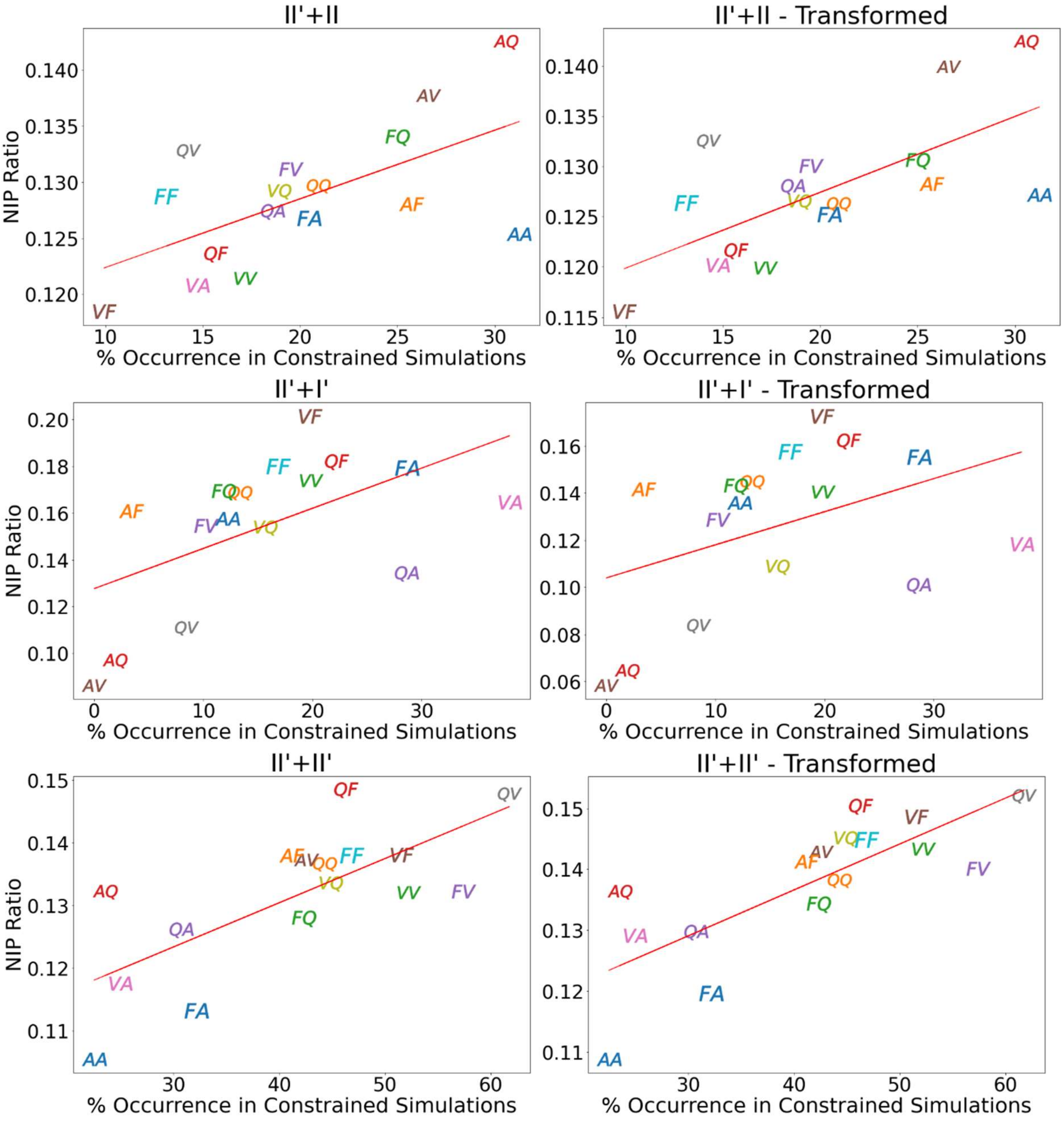


Figure 81: The NIP ratio generated from the Ramachandran plots before and after transformation are very similar as shown by the comparison of the NIP ration for the three conformations commonly seen in the type II' restrained simulations.

The calculated size of the kernel bandwidth is seen to vary between the transformed and untransformed datasets. A cross-validation method was chosen to select kernel bandwidth due to the complex underlying distribution of the Ramachandran plot making reference rules unsuitable. However the kernel bandwidth selection did not take into account the fact that the dihedral angles used are circular variables. There are many methods that may be better designed to give bandwidth values for circular variables but their implementation currently remains difficult.[353-357] These methods are generally made for spherical circular variables (circular variables on the surface of a sphere) but Marzio *et al*. developed a method for KDE for use on a torus (the 3D shape of the Ramachandran plot).[358] They implemented a cross-validation method for bandwidth selection adapted for use on circular variables. The difficulties of choosing bandwidth may therefore be leading to inaccurate NIP values.

Rather than using KDE to generate the probability density estimates, Dirichlet processes (a type of Bayesian nonparametric method) could be used instead.[359] Whereas KDE builds up a density estimate from component densities placed at each data point, the Dirichlet process mixture model also uses multiple density functions which are combined to give the overall density estimate (a distribution made up of distributions) but the number of component densities is unknown and inferred from the data. Dirichlet processes do not require a choice of bandwidth however they require assumptions about the underlying shape of the data, often take longer to compute and similar to KDE, has not yet been fully adapted for use on circular variables. Ting *et al*. used a hierarchical Dirichlet process to generate the probability density estimates of the Ramachandran plots of amino acids and then used the Hellinger distance (H), a method very similar/related to the NIP, to look at the similarity between distributions.[360] The equation for calculation of the Hellinger distance is shown in equation 9 where P_1 and P_2 represent two probability distributions. The Dirichlet process could therefore offer an alternative to using the NIP.

$$H^2 = \frac{1}{2}\int(\sqrt{P_1} - \sqrt{P_2})^2\, dx \qquad (9)$$

4.8.6.2 Coupling Effects

If the $\text{NIP}_{(\text{aa, p, C})}$ values reflect the accurate similarity between distributions then the NIP ratios may not be predictive of the amount of each conformation seen in the restrained simulations due to the amino acid sidechains interacting with each other leading to different preferred conformations than based on just the Ramachandran distributions alone.

Amino acids often show coupling effects, where neighbouring amino acids alter the Ramachandran preference of an amino acid. This is most noticeably seen with pre-proline residues which frequently show reduced density of dihedral angles in the α and left-handed regions of the Ramachandran plot than they otherwise would.[298, 361] Coupling effects have also been seen with other residues.[360] Testing the use of NIPs on the restrained simulations should have eliminated this effect as much as possible. Only glycine is adjacent to the i/i+3 positions and this is consistent between all the restrained simulations being tested. Additionally the NIPs are calculated using the Ramachandran distribution generated using the Ac-GGXGG-NH_2 MD simulations which also have glycine adjacent to the amino acid being investigated. Therefore neighbouring effects should be minimal. However in the cyclic hexapeptide, as a small restrained system, neighbouring effects may not be limited to adjacent amino acids. The i/i+3 positions are held in relatively close proximity due to the structure of the peptide. If the sidechains are large they could interact: the hydrophobic effect could contribute to a lower energy structure and in the phenylalanine containing sequences CH-π interactions are possible.

The use of NIPs treated the A and B positions in a conformation as independent variables. If the probabilities are not independent this means they cannot accurately be combined into information about the overall most probable state. If such interactions between sidechains are occurring, then the assumption made that the two amino acids NIPs can be treated as independent variables will not work as the Ramachandran distribution of one will influence the other. Depending on the magnitude of the energy associated with the sidechain interaction it may or may not be possible to discount this influence on the Ramachandran distribution. If the sidechain interaction in a conformation leads to only a minor alteration in the stability of that conformation, then this effect may be negligible with the NIPs still being predictive of the amount of the conformation seen. If the same sidechain coupling effects occur across all the conformations or if the Ramachandran distributions have a greater impact on the most stable conformation, then the NIPs would still be a good approximation. However different conformations are likely to have different sidechain coupling effects. Although each conformation has unique i/i+3 dihedral angles there is some overlap between different conformations with, for example, the I'+II', I'+I' and II'+II' conformations all having very similar pA and pB dihedral angle distributions. There is likely to be little energy difference between conformations with similar pA and pB dihedral angles due to Ramachandran preferences of the amino acids as only very small differences are seen. Therefore small differences due to sterics or sidechain interactions could have a comparatively large effect since the difference in i/i+3 dihedral angles remains very small.

One way to test if the NIP Ratio is not predictive of the amount of a particular conformation due to interactions between the two amino acids at the i/i+3 positions is to try the method on either the c(XGGGGG) or c(GGGXGG) restrained simulations. In these simulations glycine replaces one of the chiral amino acids preventing any interactions between large sidechains which may be altering the results. However with only one amino acid changing between the different simulations there are generally much smaller differences seen between the simulations. There are also only four sequences to compare: A, Q, V and F making it more difficult to look for a correlation between the NIPs and the amount of a conformation so further restrained simulations using different amnio acids would have to be carried out to test this.

4.8.6.3 Variation in the Restrained Simulations

The restrained simulations show that the i/i+3 dihedral angles for each conformation remain the same regardless of which chiral amino acids occupy those positions. There is however some noise and variation in the data. The amount of some of the conformations that appear in the restrained simulations is relatively small for many sequences, usually making up less than 10% of the conformations seen throughout the simulation. Especially for the smaller clusters, noise in the data may impact the results, as the noise in the data is proportionally large compared to their size. It is therefore more difficult to use the NIP ratio to predict the amount of these conformations. The size of the clusters in the restrained simulations can vary by approximately ±5 when the restrained simulations are repeated. Running multiple simulations on each sequence and then taking an average may help reduce the effect of noise in the restrained simulations but this requires much more computational time. Adaptive kernels or noise filtering could also be used. Adaptive kernel density estimation alters the size of the kernels used to generate the probability density of the data based on where the point is in space. This can prevent localised undersmoothing or oversmoothing of particularly long-tailed or multimodal distributions.

The dihedral angles for the i/i+3 positions used to generate the $\mathrm{NIP}_{(\mathrm{aa,\ p,\ C})}$ values were obtained from the c(AGGAGG) restrained simulations. Although the same dihedrals for a particular conformation was seen with other amino acids, there may be small differences in the distributions which mean, although very similar, they cannot be treated as the same. Small differences are seen when the distributions are compared using the NIP (Table 31). Comparing the i/i+3 dihedral angles for a particular conformation between the different sequences gives NIP values that vary from around 0.95 to 1. These values are close to one showing they are very similar between all the simulations. However there is only a small energy difference between the different conformations that can occur. It may be that these small differences in the observed i/i+3 dihedral angles seen when different sequences are used are sufficiently large when looking at the very small energy differences between the conformations that using alanine as representative values does not allow for accurate prediction.

Position	NIP			
	II’+II	II’+I’	II’+II’	II’+IV
i/i+3	0.99654	0.96519	0.95371	0.99412
Restrained turn i+1	0.99994	0.99926	0.99910	0.99998
Restrained turn i+2	0.99998	0.99910	0.99952	0.99990
i+3/i	0.99813	0.98248	0.95846	0.99790
Restrained turn i+1	0.99901	0.99771	0.99748	0.99863
Restrained turn i+2	0.99915	0.99973	0.99994	0.99298

Table 31: NIP comparison of the Ramachandran plots of c(AGGAGG) and c(QGGQGG) for the type II’ restrained simulation.

The double restrained simulations should have the most consistent conformations across the different sequences as both β-turns are restrained to ideal values. Two intrapeptide hydrogen bonds are often seen in cyclic peptides as a hydrogen bond forms between each of the i and i+3 positions of the two β-turns. The distances between the carbonyl group and the amide hydrogen of the i and i+3 positions for the c(AGGAGG) and c(QGGQGG) II+I’ double restrained simulations trajectories are shown in Figure 82. Hydrogen bonding is dependent on distance so the differences in the histograms distributions could potentially be altering the stability of the conformation with the different amino acids with different amounts of hydrogen-bonding occurring. Similar slight differences are seen across the conformations seen in the restrained simulations. In such a way despite the similarities between them, small differences in Ramachandran plots could lead to different energy structures.

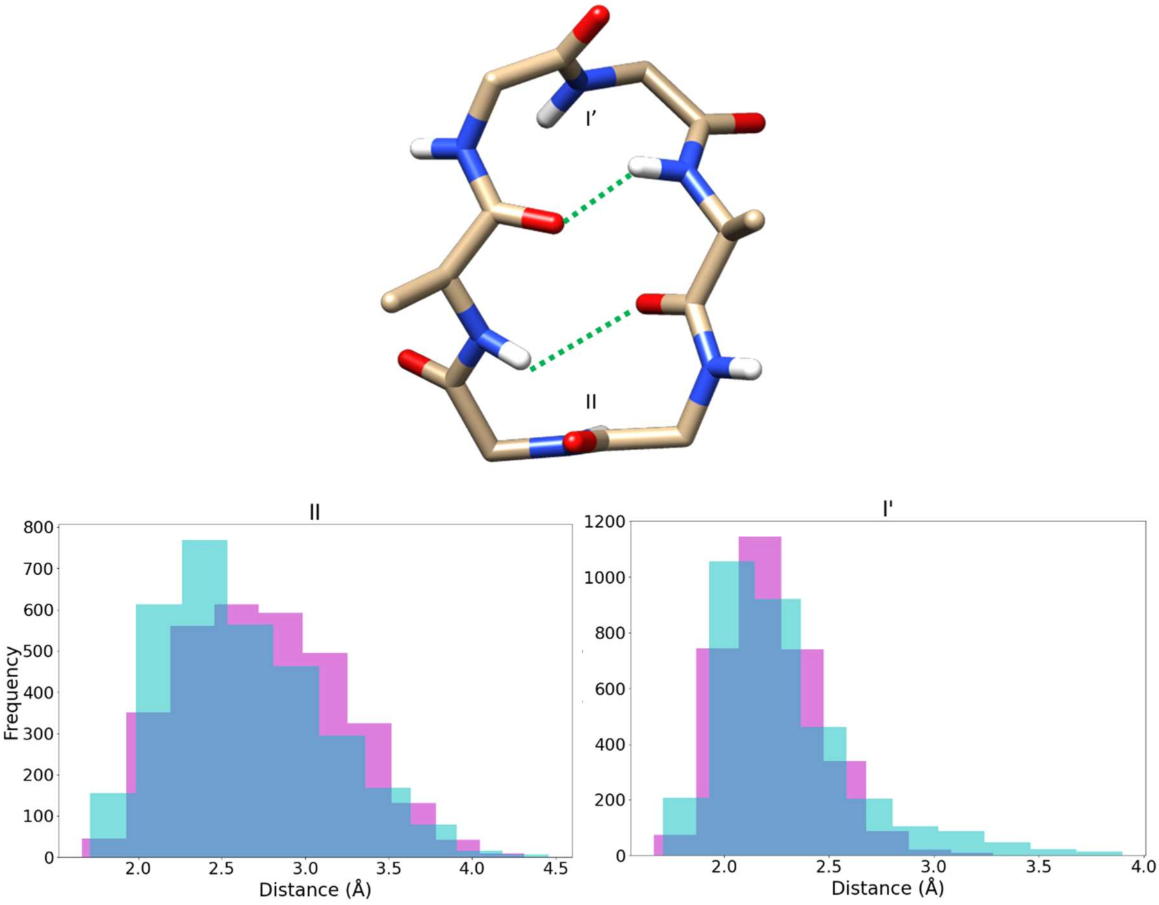


Figure 82: Differences in the distance between the carbonyl group of the i position and amide hydrogen of the i+3 position of the two turns in the I’+II conformation for c(AGGAGG) (pink) and c(QGGQGG) (blue).

Throughout the trajectories from the restrained simulations limited rotation of the glutamine, phenylalanine and valine sidechains is often seen. The sidechain orientation may alter the relative energies of the β-turn types seen. A particular sidechain orientation may result from sidechain interactions such as the hydrophobic effect which could make a particular conformation more favourable than would be predicted by the NIPs.

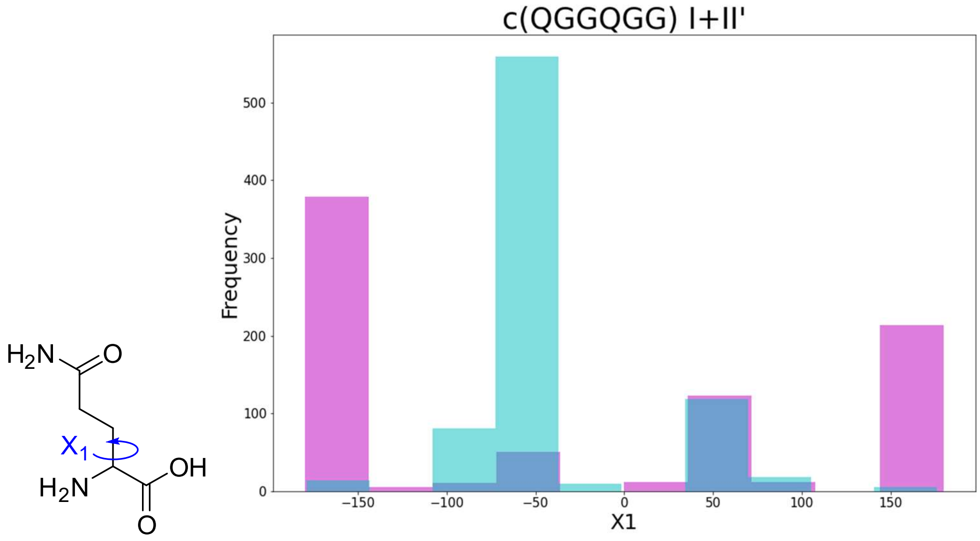


Figure 83: The χ_1 dihedral angles of the two glutamine residues in the c(QGGQGG) I+II' conformation seen in the type I restrained simulation. Limited rotation of the sidechain is seen.

Although each conformation has very similar i/i+3 Ramachandran distributions when different amino acids are used, there are very small differences between them. These small differences can lead to differences in the amount of intrapeptide hydrogen bonding between the i and i+3 positions as well as potentially slightly altering any backbone sterics and interactions of the peptide with the solvent. In addition larger amino acids have restricted χ values which vary between conformations and sidechains. As the energy difference between conformations is so small, even factors that only slightly alter the energy of the conformation can have large effects on the relative energies. These slight differences between simulations therefore may be a potential reason why the overlap in Ramachandran region for each conformation cannot be used to predict the proportion of the conformation that occurs in the constrained simulations.

4.9 Conclusions

Due to the tendency of cyclic peptides to adopt multiple conformations in solution, with the amino acids changing register within β-turns, it can be difficult to determine individual effects of certain amino acids occupying given positions within the structure. The restrained simulations show which β-turn types are most compatible in the cyclic hexapeptide structure due to the backbone conformation of the peptide. The presence of chiral amino acids at the i/i+3 positions alters which turn types are most compatible which should be taken into account when choosing sequences for cyclic hexapeptides. The effect of chiral amino acids at the i/i+3 positions on the most stable backbone structure could not be determined in an unrestrained system due to a prevalence of the chiral amino acids to occupy the i+1 position and alter β-turn type. The restrained simulations therefore allow for exploration of small changes in the cyclic peptide system which would otherwise be difficult to observe.

- For the c(GGGGGG) simulations the I+I' and I+II'/I'+II conformations were lowest energy

- For the c(XGGXGG) simulations different backbone conformations were seen compared to c(GGGGGG) as the i/i+3 dihedrals were different for the same turn type combinations
- The I'+I', II'+II', II'+II and I+II conformations tended to be most favourable for the c(XGGXGG) simulations
- For the c(XGGGGG)/c(GGGXGG) simulations similar lowest energy conformations were seen to the c(GGGGGG) simulations. Higher energy conformations resembled those seen in the c(XGGXGG) simulations

Although similar lowest energy conformations are generally seen when different L-amino acids are included at the i/i+3 positions in the cyclic peptide sequence, different proportions of each conformation are seen. Inclusion of valine generally leads to the greatest differences which should be taken into account when β-branched amino acids are included in a cyclic peptide. However the Ramachandran positions necessary for each turn-type combination remain the same regardless of which chiral amino acids occupy the i/i+3 positions. As the only difference between the restrained simulations was the choice of amino acids at the i/i+3 positions, a correlation between the overlap of the Ramachandran plot of each amino acid generated from MD simulations on Ac-GGXGG-NH_2 with the required positions for a given conformation and the amount of that conformation which appeared in the restrained simulations was looked for using the normalised integrated product (NIP) of the two Ramachandran distributions.

It was found that the NIP cannot be used to predict the amount of a particular conformation. Possible reasons for this include coupling effects of the two amino acids, sidechain interactions and the sterics in the relatively restrained system. All of which could potentially alter the most stable conformation and have a greater effect on which is the lowest energy conformation than the intrinsic Ramachandran preferences of the amino acids. Additionally although each conformation has unique i/i+3 dihedral angles there are very small differences between the same conformations with different amino acids. Since there are only small energy differences between the different conformations these small differences could impact the results preventing the use of alanine as representative of a particular conformation.

As the restrained simulations show which turn type combinations form the lowest energy conformations, choosing sequences more likely to form compatible turn types should help make a more conformationally restricted cyclic peptide. For example when incorporating a type II' turn into a cyclic peptide it might be best to choose a sequence which forms a type II or II' turn to complete the cyclic peptide if chiral amino acids are at the i/i+3 positions. This may lead to a cyclic peptide with a more stable structure. Machine learning based methods can be used to predict the β-turn type a sequence is likely to form. In the next chapter a random forest (RF) machine learning algorithm is trained to predict the β-turn type a tetrapeptide sequence is likely to form. The compatible turn types determined from the constrained simulations are then used in conjunction with the RF to attempt to predict the conformation of cyclic hexapeptides.

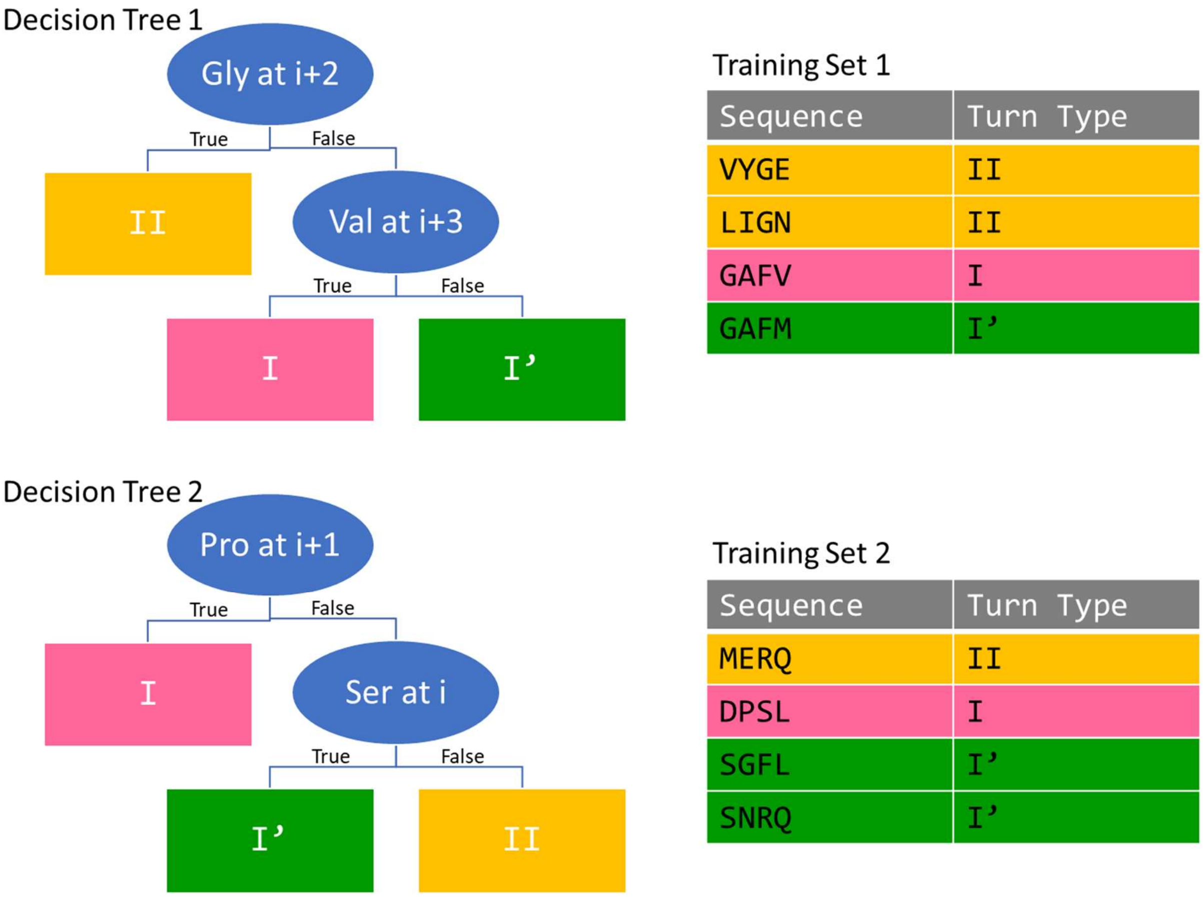
5 Using a Random Forest to Predict Cyclic Peptide Structure

As cyclic hexapeptides often form a structure in solution which can be thought of as two overlapping β-turns, it may be possible to predict the structure of a cyclic peptide by prediction of which β-turn types the sequence is likely to form. There are currently numerous methods based on machine learning algorithms which predict the β-turn type likely to form but only in the wider context of a protein.[134, 139, 140, 362, 363] As they are designed for protein structure prediction and validation these methods require the input of relatively large peptide sequences and other factors such as secondary structure prediction rather than just the tetrapeptide sequence that makes up the β-turn. This means existing β-turn type prediction methods are not suitable for predicting the β-turn type likely to form in a small cyclic peptide. A Random Forest machine learning algorithm was therefore trained to predict β-turn type specifically for use in cyclic peptide prediction (see section 10.2).

5.1 Decision Trees and Random Forests

Analysis of the β-turns found in the Loop Database show that different β-turn types often have distinct sequence profiles, with proline and glycine often playing an important role in determining turn type. As such it is possible to train a machine learning algorithm to predict β-turn type from the sequence. By extracting features which increase the probability of sequences being a certain turn type, the algorithm can then use this data to predict the turn type of any given sequence. As a cyclic hexapeptide is frequently observed to form conformations that can be thought of as two overlapping β-turns, the machine learning algorithm could then be used to predict the β-turn types likely to form within the cyclic peptide based upon the sequence. This would help choose which sequences to use when designing cyclic peptides.

A decision tree is a machine learning algorithm which can be trained on a dataset to predict the outcome of a classification problem. In a decision tree the leaves represent the outcome. The tree is trained to reach the leaves by splitting the data at nodes based on features of the dataset. The nodes of the decision trees are chosen based on attribute selection measures which find the features of the data that best split the data into the relevant categories. The Gini method is one such attribute selection measure which is commonly used.[364] Figure 84 shows two very simple examples of decision trees trained on 4 sequences which form either a type I, II or I’ β-turn. In decision tree 1 the sequences are first split depending on if a glycine is present at the i+2 position of the β-turn sequence. If glycine is present the sequence is assigned as type II. If there is not a glycine at the i+2 position the second node splits the remaining sequences into type I and I’ turns. The second decision tree, trained on a different dataset, uses alternative criteria to separate the sequences into the different categories.



Training Set 1

Sequence	Turn Type
VYGE	II
LIGN	II
GAFV	I
GAFM	I’

Training Set 2

Sequence	Turn Type
MERQ	II
DPSL	I
SGFL	I’
SNRQ	I’

Figure 84: Two decision trees trained on separate data sets. Nodes are represented as blue ovals and the leaves are the assigned β*-turn type.*

The sequence GGFH is present in the Loop Database as a type II turn. Decision tree 1 would classify this sequence as a type I’ turn. Decision tree 2 however would correctly assign this sequence as a type II turn. A Random Forest (RF) randomly splits a large dataset into separate training sets (the same sequence can occur in multiple training sets). Multiple decision trees are then trained on the separate training sets. When the RF is given a new sequence to classify, each decision tree in the RF assigns the sequence to a particular category. The final result is then based on the input of all the decision trees with the decision trees “voting” on the outcome they consider most likely based on their past experiences with their separate training sets (Figure 85). The proportion of the decision trees which vote to assign a datapoint to a particular class is often used to represent the probability of the datapoint occurring in that category. In this case the predicted probability would represent the probability of the sequence forming that β-turn type. For a RF made up of the two decision trees in Figure 84 the sequence GGFH would have a predicted probability of forming a type II turn of 0.5. By using a RF rather than a single decision tree a more accurate assignment can usually be obtained as overfitting is avoided. So although decision tree 1 would assign the GGFH sequence incorrectly, if the majority of the decision trees in a RF assign it as a type II turn the correct overall assignment would still be seen.

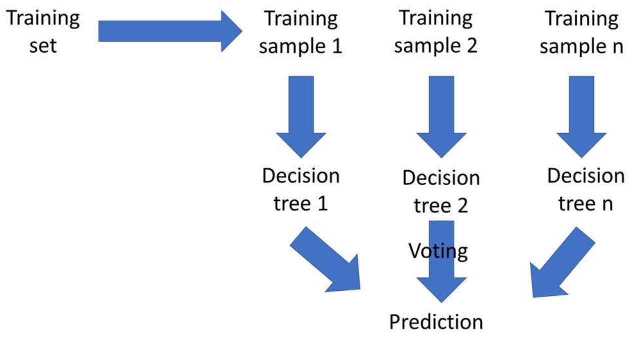


Figure 85: A RF is made up of multiple decision trees.

Unlike some other machine learning algorithms such as neural networks, the decision-making process in a decision tree (and therefore a RF) is interpretable. They are also fast to train, do not rely on any prior assumptions about the input data and are able to handle high-dimensional data. A RF was therefore chosen to be trained to predict the β-turn types an amino acid sequence is likely to form in order to determine if this prediction correlates with the β-turn types seen within a cyclic hexapeptide structure (Figure 86).

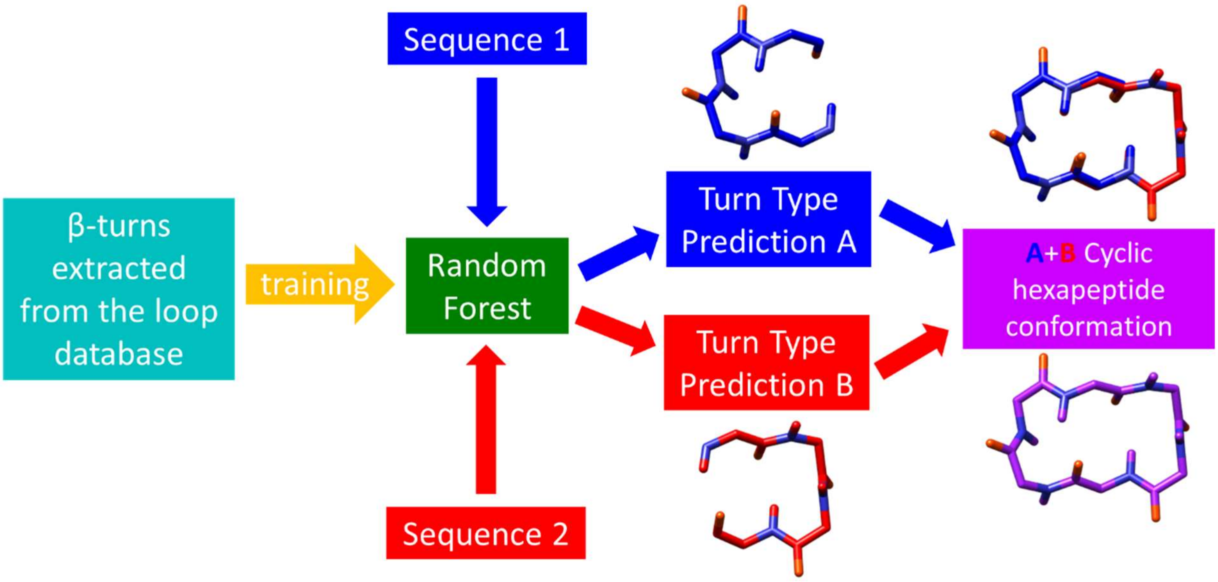


Figure 86: Random Forest for prediction of β-turn types in a cyclic hexapeptide.

5.2 Training the Random Forest

The data put into a RF algorithm is initially split into a training set and a test set. The training set makes up the vast majority of the data and is randomly split into samples for training of the individual decision trees which make up the RF. After the RF has been trained it then predicts the category the remaining data in the test set will fit into. How well these predictions match up with the actual categories the test set fit into can then be used to evaluate the ability of the RF for accurate prediction. To evaluate a RF classifier the precision, recall, F1-score and overall accuracy can be assessed. Precision is the number of correct positive results divided by all positive results (correct positive results + false positives) identified by the random forest. A high precision indicates the RF produces a low proportion of false positives when making predictions. A high recall on the other hand means a low number of false negatives are predicted – recall is the number of correct positive results divided by the number of results if everything had been assigned correctly (correct positive results + false negatives). A F1-score takes into account both precision and recall and has a value between 0 and 1, 1 being perfect precision and recall.

5.2.1 Random Forests using β-hairpin Data

β-turn sequences and their corresponding turn types were extracted from the Loop Database. The RF algorithm was initially trained on sequences from the β-hairpin dataset. As the β-turns require an additional hydrogen bond adjacent to the one found in the β-turn they show some similarity to the cyclic hexapeptide structure which often forms two overlapping β-turns with two intramolecular hydrogen-bonds.

The expanded β-hairpin dataset where β-hairpins using either the hydrogen bonded definition of a β-turn or a distance of less than 7 Å between the i/i+3 α-carbons was used to train the RF (Table 32). This gives as large a dataset as possible which is important for accurate training of a machine learning algorithm, especially because in this dataset there are few examples of type II turns which would make prediction of this turn type difficult. Although the clustering algorithm used to determine the β-turn types in the hairpin dataset did not pick out type VI turns, searching for turns with an ω dihedral angle between 20 and -20 ° and proline at the i+2 position in the expanded β-hairpin dataset found all the type VI β-hairpins. The type VI turns were therefore included as a separate category to train the RF on besides the type I, II, I’ and II’ turns found when the hairpin dataset was clustered. This could be useful as inclusion of a proline can often aid cyclic peptide synthesis.

Turn Type	Number of Sequences
I	2,628
II	488
I’	6,679
II’	4,018
VI	103
Total	13,916

Table 32: The number of each β-turn type found in the expanded hairpin dataset.

To train the RF, 90% of the data was used as a training set with the remaining 10% used as the test set. 100 decision trees were used in the RF. Altering the size of the training set or number of trees did not alter the prediction ability of the algorithm. As a β-turn is made up of four amino acids, i to i+3, the tetrapeptide sequences which make up the β-turns in the expanded hairpin dataset were used to train the RF. An 81% successful prediction of the β-turn types in the test set was seen. The RF produces a value known as feature importance which is a measure of which of the input features are most important for determining the overall classification. In this case the feature importance determines which of the i to i+3 positions in the tetrapeptide sequences which make up the β-turn are most important for determining which β-turn type the sequence will form. The feature importance for each of the i to i+3 positions was: 10:23:35:31. The i+2 position is therefore most important for predicting turn type. Despite β-turn type being defined based on only the i+1 and i+2 dihedral angles, the i+3 position has a greater feature importance than the i+1 position.

	Precision	Recall	F1-score	Support
I	0.8	0.69	0.74	254
II	0.53	0.16	0.24	51
I’	0.83	0.91	0.87	692
II’	0.78	0.79	0.79	384
VI	1	0.82	0.9	11

Table 33: Random forest classifier on β-turns from the β-hairpin dataset with a 7 Å distance between i and i+3 α-carbons.

Only a very small proportion (3.5%) of the hairpins present in the training set are type II, therefore prediction of type II turns remains poor as seen in the evaluation of the different categories where an F1 score of 0.24 is seen (Table 33). If type II turns have a similar sequence profile to other turn types but there are few examples of type II turns then there is little opportunity for the algorithm to learn the differences between similar sequences and instead a high overall accuracy can be obtained by assigning sequences to the larger categories. This is why overall accuracy is a poor indicator of prediction ability when there is large discrepancy between the sizes of the input data categories, as is the case here. The high overall accuracy does not reflect the prediction ability for the smaller categories which remains relatively poor. By using a different larger dataset from the database which includes more type II β-turns more data will be available to train the RF to obtain more accurate predictions across the smaller categories.

5.2.2 Random Forests on an Expanded Dataset

The dataset of β-turns found in larger loops within the Loop Database rather than in β-hairpins was used to train the RF algorithm. This is a larger dataset containing approximately ten times as many turns. A larger dataset means more examples to train the RF with so should lead to improved predictions. The β-turns however may not resemble those seen in the cyclic peptide structure as much as those found in the hairpin dataset as there are less restraints on their structure. As previously the data was split into 90% training set and 10% test set with 100 trees. Overall an 86 % accuracy is seen, a slight increase from the 81% accuracy seen when the hairpin dataset is used to train the RF. However the II’ and I’ categories, with F1-scores of 0.29 and 0.28 respectively are now poorly predicted rather than the type II category (Table 34). This is likely due to the limited number of type II’ and I’ turns found in the β-turns in the database. Whereas in the hairpin dataset the high level of accuracy was due to the dominance of the I’ category, here the same is seen with the type I category. More type II’ and I’ turns are seen in the 15,695 hairpins in the database than in all the 152,226 β-turns in bigger loops. Joining the datasets together therefore may help the prediction of the smaller categories.

	Precision	Recall	F1-score	Support
I	0.89	0.94	0.92	10,900
II	0.74	0.72	0.73	2,719
I’	0.49	0.20	0.28	526
II’	0.50	0.20	0.29	287
VI	0.69	0.63	0.66	116

Table 34: Prediction ability of the random forest algorithm when trained on the β-turns found within larger loops in the Loop Database.

When the datasets are combined the prediction ability for each category is somewhere between the two different datasets separately (Table 35). An accuracy of 82% is seen overall. A drop in overall accuracy means very little as there is still a large discrepancy in the size of the input categories. The prediction of the type I’ and II’ sequences is slightly improved but remains much lower than the other categories. Although very small the type VI category consistently has relatively high F1 scores. By definition type VI turns require a proline at the i+2 position which may be one of the reasons a relatively good F1-score is seen despite the small sample size. There may be more sequence similarity seen between the other sequences making their prediction more difficult. The feature importances are very different than those in the hairpin dataset: 16.3%, 18.4%, 46.4% and 18.8 % for the i to i+3 positions respectively. The i+2 position is still the most important for determining β-turn type. The differences in feature importance from the hairpin dataset may reflect the different breakdown in β-turn types seen between the categories.

	Precision	Recall	F1-score	Support
I	0.89	0.92	0.90	11,092
II	0.69	0.67	0.68	2,781
I’	0.56	0.38	0.45	1,022
II’	0.57	0.52	0.54	625
VI	0.60	0.64	0.62	110

Table 35: Prediction ability of the random forest algorithm when trained on the β*-turns found within larger loops in the Loop Database and the* β*-hairpin dataset.*

5.2.3 Resampling

Due to the large difference in category sizes used to train the RF a high overall accuracy can be obtained by predicting everything to belong the largest categories. This leads to very poor prediction of the smaller categories. Resampling can sometimes be used to overcome this issue by altering the data to give a training set with more even class sizes. The two most simple resampling methods are undersampling and oversampling. Undersampling requires taking fewer samples from the larger categories when training the RF, whereas oversampling requires randomly multiplying examples from the smaller categories to make them as large as the bigger categories, again making the category sizes more equal.

Both undersampling and oversampling were used on the β-turns used to train the RF to try to improve the predictive ability of the RF towards the smaller type I’ and II’ categories. The smallest category is the type II’ turns so the remaining categories were randomly undersampled to have the same number of sequences as in the type II’ category. The undersampled data sets were then used to train the RF. Conversely the largest category is the type I β-turns. For oversampling, sequences from the smaller categories were randomly resampled to make those categories the same size as the type I category prior to training the RF.

Undersampling gives an overall accuracy of 70% and oversampling gives an overall accuracy of 76%. Both lower than the 82% accuracy seen without resampling. Although the recall is improved for the smaller I’ and II’ categories (Table 36) there is no significant improvement so it was decided to not use resampling when training the RF.

Undersampling:				
	Precision	Recall	F1-score	Support
I	0.93	0.72	0.81	11,092
II	0.54	0.60	0.57	2,781
I’	0.30	0.68	0.41	1,022
II’	0.33	0.68	0.44	625
VI	0.50	0.97	0.66	110

Oversampling:				
	Precision	Recall	F1-score	Support
I	0.91	0.82	0.86	11,092
II	0.58	0.63	0.61	2,781
I’	0.37	0.56	0.44	1,022
II’	0.39	0.61	0.48	625
VI	0.62	0.76	0.68	110

Table 36: Predictive ability of the RF trained on resampled datasets.

5.2.4 Duplicates within the dataset

An overall 100% accuracy could not be achieved by the RF as it is possible for the same sequence to form multiple β-turn types (Table 37). 44,934 of the sequences from the combined β-hairpin and β-turn datasets form just one turn type (these sequences represent 97,290 β-turns). 10,113 sequences (59,003 turns) are seen to form at least two turn types in the datasets. Of the 10,113 sequences that form multiple turn types: 8,506 form 2 turn types (84.1 %), 1486 form 3 turn types (14.7 %) and 121 sequences form 4 turn types (1.2 %). None of the sequences that form 4 turn types are type VI turns (all are either I, II, I' or II').

Turn Type	% of turns in the combined β-hairpin and β-turn datasets	% of the 59,003 turns that form multiple turn types	% of the 1,398 turns that form all 4 turn types
I	74.8	44.4	18.2
II	18.8	35.5	36.6
I'	3.6	13.3	33.5
II'	2.0	6.4	11.7
VI	0.7	0.3	0

Table 37: The breakdown of turn types formed by sequences which form multiple turn types.

The amino acids found at the i+1 and i+2 positions in the β-turns which form multiple turn types corrected for category size and the amino acids naturally occurring frequency are shown in Figure 87. Having a D, G or N at the i+2 position seems to give the most flexibility, with these amino acids also frequently occurring at the i+1 positions as well. The i+2 position has the highest feature importance in the RF (almost 50 %) so it is likely that having more flexible amino acids, that have smaller preference for particular regions of the Ramachandran plot, allows for greater flexibility allowing the sequences to form multiple turn types. Type I are usually sequences made up of all chiral amino acids which is potentially one of the reasons type I turns form a much lower percentage of the sequences that form multiple turn types than in the dataset as a whole.

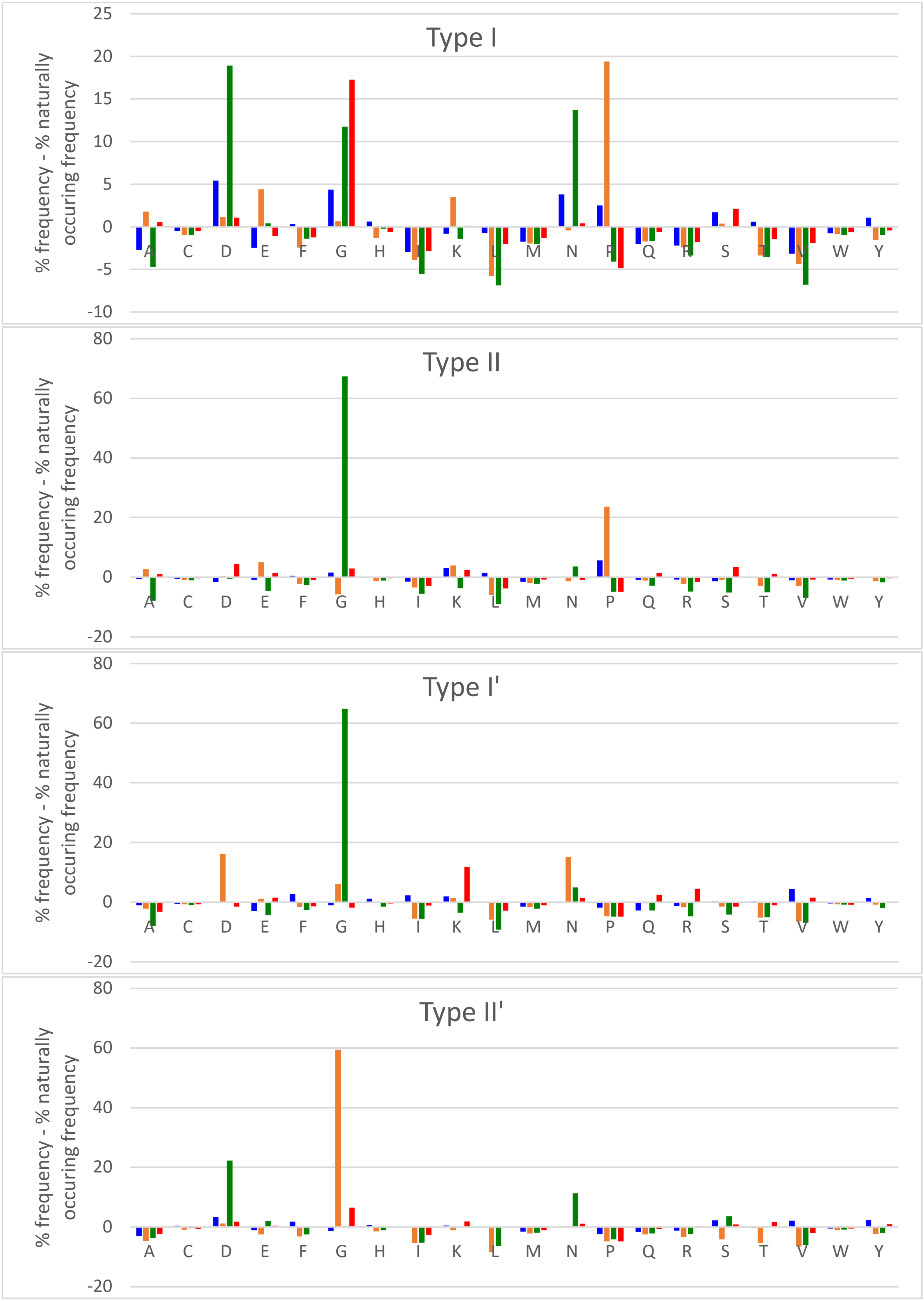


Figure 87: Sequence profiles of the sequences which form multiple turn types. The i position is blue, i+1 is orange, i+2 is green and i+3 is red. Naturally occurring frequency of amino acids based on all available structures in the PDB taken from [301].

The ability of certain sequences to form multiple β-turn types could reduce the accuracy of the RF as it is difficult to predict what turn type these sequences will form. Of the 44,934 sequences that form just one turn type 84,581 type I turns are seen, 6,713 type II, 2,335 type I', 2,679 type II' and 982 type VI turns. Using just these sequences to train the RF a 99.5% accuracy is seen when both the training set and test set are made up of just the sequences that form only one turn type, with very high F1 scores seen across all categories (Table 38).

	Precision	Recall	F1-score	Support
I	1.00	1.00	1.00	8,452
II	0.99	0.97	0.98	656
II'	0.98	0.97	0.98	267
I'	0.98	0.95	0.97	255
VI	1.00	0.98	0.99	99

Table 38: Very high accuracy, precision, recall and F1 scores are seen when the RF is trained and tested on sequences which do not form multiple turn types.

To allow a direct comparison between the RF trained on all the data, the same test set containing the sequences which can form multiple turn types should be used. In this case an improved 87% overall accuracy (compared to 82%) was seen and slightly higher F1-scores across all the categories (Table 39). Removing ambiguous sequences that can form multiple turn types therefore improves the predictive ability of the RF.

	Precision	Recall	F1-score	Support
I	0.92	0.95	0.93	11,092
II	0.73	0.72	0.73	2,781
II'	0.77	0.64	0.70	625
I'	0.64	0.47	0.55	1,022
VI	0.89	0.91	0.90	110

Table 39: The results of the RF trained on sequences which do not form multiple turn types but tested on a dataset containing sequences which do form multiple turn types.

For the sequences that form multiple turn types, there is often a dominant turn type that is formed and only one or two examples of the sequence forming another turn type. For example, the sequence VPGE appears as a β-turn forming sequence 19 times within the database. 18 of those times it appears as a type II turn and only once appearing as a type I turn. Adding such sequences after removing the anomaly where it forms the lesser occurring turn type may further increase the accuracy of the RF.

Of the 8,506 sequences that formed 2 turns types, 1,136 of these sequences formed their dominant turn type 5 or more times, and the other turn type only once. The one occurrence of the other turn type was removed as well as one occurrence of the dominant turn type and then the sequences added to the dataset of sequences that only form one turn type. This increased the dataset by 9,334 from 97,290 to 106,624 turns. When the RF algorithm was trained on this dataset there was very little difference from that trained on the original dataset of turns that only form one turn type (Table 40). An overall accuracy of 95 % on the test set without sequences that form multiple turn types, and an overall accuracy of 87% on the original training set used for the dataset with all turns regardless of if they form multiple turn types was seen. Therefore since there is little extra benefit to be gained from adding sequences that from multiple turn types once selecting for their most dominant turn type they were not included when training the final version of the RF.

Test set from dominant single turn dataset				
	Precision	Recall	F1-score	Support
I	0.96	0.99	0.97	9,167
II	0.91	0.79	0.84	867
II’	0.81	0.71	0.75	247
I’	0.80	0.52	0.64	282
VI	0.77	0.79	0.78	100

Test set with repeat sequences				
	Precision	Recall	F1-score	Support
I	0.92	0.95	0.93	11,092
II	0.74	0.73	0.73	2,781
II’	0.76	0.65	0.71	625
I’	0.64	0.49	0.56	1,022
VI	0.86	0.92	0.89	110

Table 40: Evaluation of the RF algorithm trained on the dataset including sequences with a dominant turn type despite forming multiple different types of β*-turns.*

The vast majority of 20^4 possible tetrapeptide sequences made up of the 20 naturally occurring amino acids are predicted to form type I turns (Table 41), which is consistent with the dataset used to train the RF where the majority of turns were also type I. The RF could potentially overpredict the occurrence of type I turns. The dataset used to train the RF contained around 75% type I turns but around 85% of sequences are predicted to form type I turns. However this could also be due to the different breakdown of sequences in the training set, which obviously does not contain all of the tetrapeptide sequences but turns found in the database with multiples of the same sequence appearing. The RF predicts if a sequence were to occur in a β-turn what turn it is likely to form, not if the sequence is likely to form a turn or not. Other turn types typically include glycine in them and so may be better able to form turns than the sequences predicted to form a type I turn by the RF. Hence most sequences are predicted to form type I turns despite not appearing quite so frequently in the database. The main differences in predictions of the RF versions where sequences which formed multiple turn types were or were not included in the training set, is seen as an increase in type I turns. Fewer sequences were also predicted to form type I’ turns when the sequences are not included. This possibly reflects the sequence similarity between the type I’ turns and the more flexible sequences as both have an abundance of glycine, aspartic acid and asparagine.

	Sequences which form multiple turn types in training set		Sequences which form multiple turn types not in training set	
	Number of sequences predicted to form turn type	Number of sequences with a RF score of 1	Number of sequences predicted to form turn type	Number of sequences with a RF score of 1
I	137,793 (86%)	19,898	138,560 (87%)	34,241
II	9,313 (6%)	310	9,360 (6%)	774
I’	4,227 (3%)	43	3,668 (2%)	51
II’	4,929 (3%)	61	4,848 (3%)	108
VI	3,738 (2%)	45	3,564 (2%)	60
total	160,000	20,357	160,000	35,234

Table 41: The turn types predicted for all possible tetrapeptides made up of the 20 naturally occurring amino acids.

The RF trained on the combined β-hairpin and β-turn datasets with the sequences that formed multiple turn types removed provides the largest number of sequences and the best ability of predicting all turn-types rather than just favouring the largest category. It was therefore selected as the final version of the RF and used to predict all future β-turn types from tetrapeptide sequences.

5.3 BT426 Dataset

BT426 is a commonly used dataset for testing the ability of prediction methods to correctly assign β-turn type. Made up of 426 proteins it has been used to test algorithms which predict the location and then the type of β-turn within the proteins. The RF is different to these algorithms in that it doesn’t predict the location of β-turns/whether a subsequence in a protein is likely to be in a β-turn but predicts the β-turn type that a sequence would form if it were in a β-turn.

The RF was used to predict the turn type of the β-turns found within the BT426 dataset (Table 42). For each category it was able to correctly predict most of the turn types. Similar to the test set when training the RF, prediction of the type I category remains the most accurate, whereas the I’ category has the lowest accuracy with only 42% of the type I’ turns seen in BT426 being correctly assigned. The I’ category is the smallest and there is some similarity between the sequence profiles of the sequences which form multiple turn types and the type I’ turns which would further contribute to making prediction of this category difficult. If the sequences are flexible enough to form multiple turn types then which turn is seen may be context dependent. As the only input to the RF is sequence it has no information about the surrounding environment which could help determine which turn type will be seen.

		Observed turn type				
		I	II	I’	II’	VI
Predicted turn type (%)	I	94	26	21	30	32
	II	4	68	35	4	1
	I’	1	5	42	2	0
	II’	1	1	2	65	0
	VI	0	0	0	0	67

Table 42: The RFs prediction of β*-turn types for the sequences seen in the BT426 dataset compared to the observed turn type.*

Other than precision, recall, F1 score and accuracy the Matthews correlation coefficient (MCC) has been used to assess the overall performance of classifiers in β-turn prediction. It treats the true class and the predicted class as two separate binary variables and calculates the correlation coefficient between them (equation 10). The closer to one the MCC value the better the classifier. Overall a Matthews correlation coefficient (MCC) of 0.66 was obtained based on the assignment of β-turns in BT426. This is a relatively high value showing in most cases the RF is able to accurately predict the β-turn type a sequence will form.

$$MCC = \frac{TP \times TN - F \quad \times FN}{\sqrt{(TP+F \quad)(TP+FN)(TN+FP)(TN+FN)}} \qquad (10)$$

The MCC value cannot be compared to the MCC value of other methods used to predict β-turn type. These methods are designed for structure prediction and validation so frequently look at larger frames of a protein sequence as well as other factors (rather than just the tetrapeptide sequence that forms a β-turn) to determine if an amino acid is predicted to be within a β-turn and if so the β-turn type. The MCC value for these other prediction algorithms is then given for individual residues i.e. whether each residue examined individually is part of the predicted turn rather than looking at the turn as a whole. This differs from the RF where the tetrapeptide sequence that will form the i to i+3 positions of a β-turn is predicted together rather than each amino acid within the turn treated individually.

Method	Predicted Turn Types	MCC
COUDES	I, II, VIII, I’, II’ and IV	0.41
MOLEBRNN	I, II, VIII, IV, I’ and II’	0.45
DEBT	I, II, IV, VIII	0.48
NetTurnP	I, I’, II, II’, IV, VIII, VI_{a1}, VI_{a2}, VI_b	0.50
BetaTPred3	I, I’, II, II’, IV, VI_{a1}, VI_{a2}, VI_b and VIII	0.51

Table 43: MCC values for predicting whether an amino acid is in a β-turn or not for methods used to predict β-turn location and type within a protein sequence.

The RF can only predict a sequence to form a type I, II, I’, II’ or VI turn. These turn types were chosen as those were the turn types found by clustering the β-turns extracted from the Loop Database. Other β-turn prediction methods may be trained to assign other β-turn categories. Most prediction methods have a type IV category which is missing from the current version of the RF. Type IV turns are often difficult to predict as they are a broad class which includes all turns which do not fit any other category and therefore they likely have a wider range of features than the other categories.

5.4 Using the Random Forest to Design Cyclic Peptides

For a given tetrapeptide sequence the RF predicts which β-turn type it is likely to form. As the RF is made up of multiple decision trees the RF can return a value corresponding to how many of the different decision trees would classify the sequence to a particular category (how many decision trees voted for the sequence to belong to a category). This is returned as a value between 0 and 1 and such values are often used as an estimate of the probability that a datapoint belongs to a certain class. A scoring system of the highest probability minus the probabilities for all the remaining categories was used as an estimate of how likely a given sequence is to form the turn type specified by the highest probability for that sequence. For example the tetrapeptide sequence GTGC is predicted by the RF to have a 30% chance of forming a type I turn and a 70% chance of forming a type II turn. The highest probability is the 70% chance of forming the type II turn. Therefore the RF score is 0.7 minus the probabilities the sequence will form the other turn types: 0.7 – 0.3. This gives an overall RF score of 0.4 for a type II turn (Figure 88).

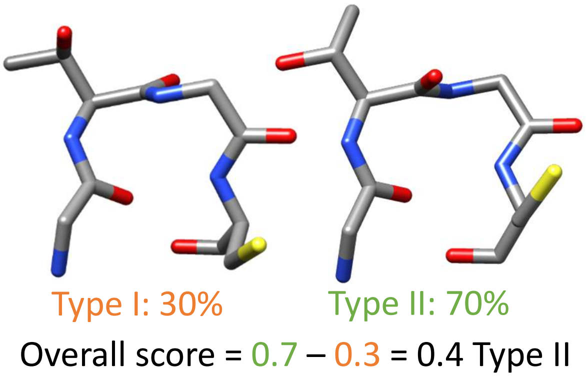
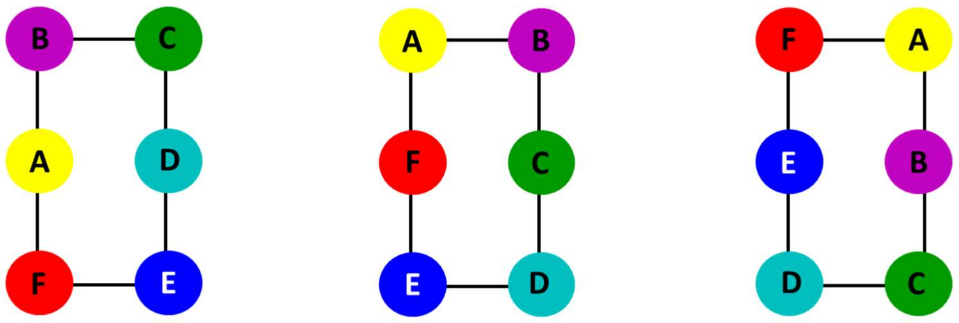


Figure 88: RF score for GTGC.

There are three possible registers for the amino acids in the overlapping β-turns that make up a cyclic hexapeptide depending on whether the amino acid is at a i/i+3, i+1 or i+2 position in the β-turn (Figure 89). By looking at the predictions for all possible overlapping tetrapeptide sequences that would make up the i to i+3 positions of a β-turn in the hexapeptide it may be possible to predict the conformation of the cyclic hexapeptide.



Sequence 1	**ABCD**	**FABC**	**EFAB**
Sequence 2	**DEFA**	**CDEF**	**BCDE**

Figure 89: The three registers amino acids in a cyclic hexapeptide made up of two β-turns can occupy.

It was hypothesised that by using higher scoring sequences a conformation made up of the predicted turn types was more likely to occur. Whereas sequences with lower scores, which are predicted to be able to form multiple β-turn types, could potentially form a cyclic peptide able to form many more conformations in solution. Therefore the RF could be used to find sequences with a high likelihood of forming one turn type combination in order to design structured cyclic peptides with only one major conformation in solution.

Using the 20 naturally occurring amino acids there are 20^4 possible tetrapeptide sequences. The RF was used to obtain a RF score for each sequence. The vast majority of sequences have very high RF scores (Figure 90). To make a cyclic hexapeptide the two β-turns must be overlayed. When looking for potential cyclic hexapeptide sequences, sequences were therefore searched for where the amino acid at the i position of the first sequence was the same as the amino acid at the i+3 positions of the second sequence and vice versa.

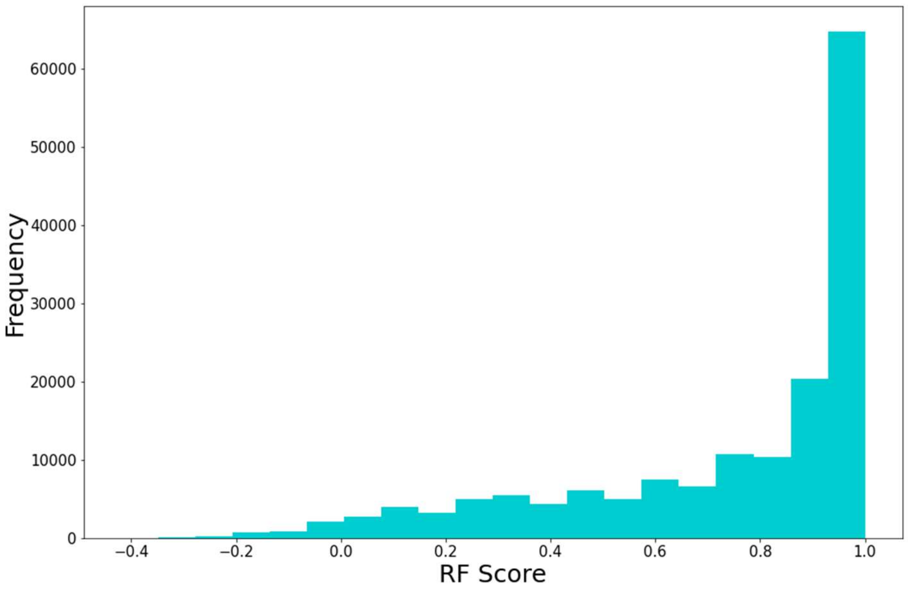


Figure 90: Most sequences have very high RF Scores.

There are currently limited examples of cyclic hexapeptides with known structures in solution. Computational methods such as BE-META are currently most used for predicting the conformation of cyclic peptides.[56] They are faster and use less resources than making the peptides and determining their conformation by NMR so offer the potential to screen sequences prior to making them. They also often allow for a more detailed description of the conformation as due to the tendency of cyclic peptides to adopt many conformations in solution it can be difficult to resolve the precise structures using NMR. BE-META carried out using RSFF2 as the forcefield has been shown to allow for the most accurate determination of cyclic peptide structure.[70] It was therefore decided to use BE-META to determine the structure of cyclic peptides and compare the results with the RF predictions. It should be noted however BE-META is also a prediction technique and can give wrong predictions. Compared to BE-META the RF is a much faster prediction technique, able to predict potential structures of a cyclic hexapeptide in under a second compared to the days it takes to run BE-META, so the RF could allow for filtering of sequences prior to simulating them.

5.4.1 c(WGTCGS)

A higher RF score may mean that the sequence is more likely to form that turn type, whereas a lower score shows a sequence has less of a preference for a particular turn type. Using the RF to find sequences that have high scores for the desired conformation may therefore help design structured cyclic peptides (Figure 91).

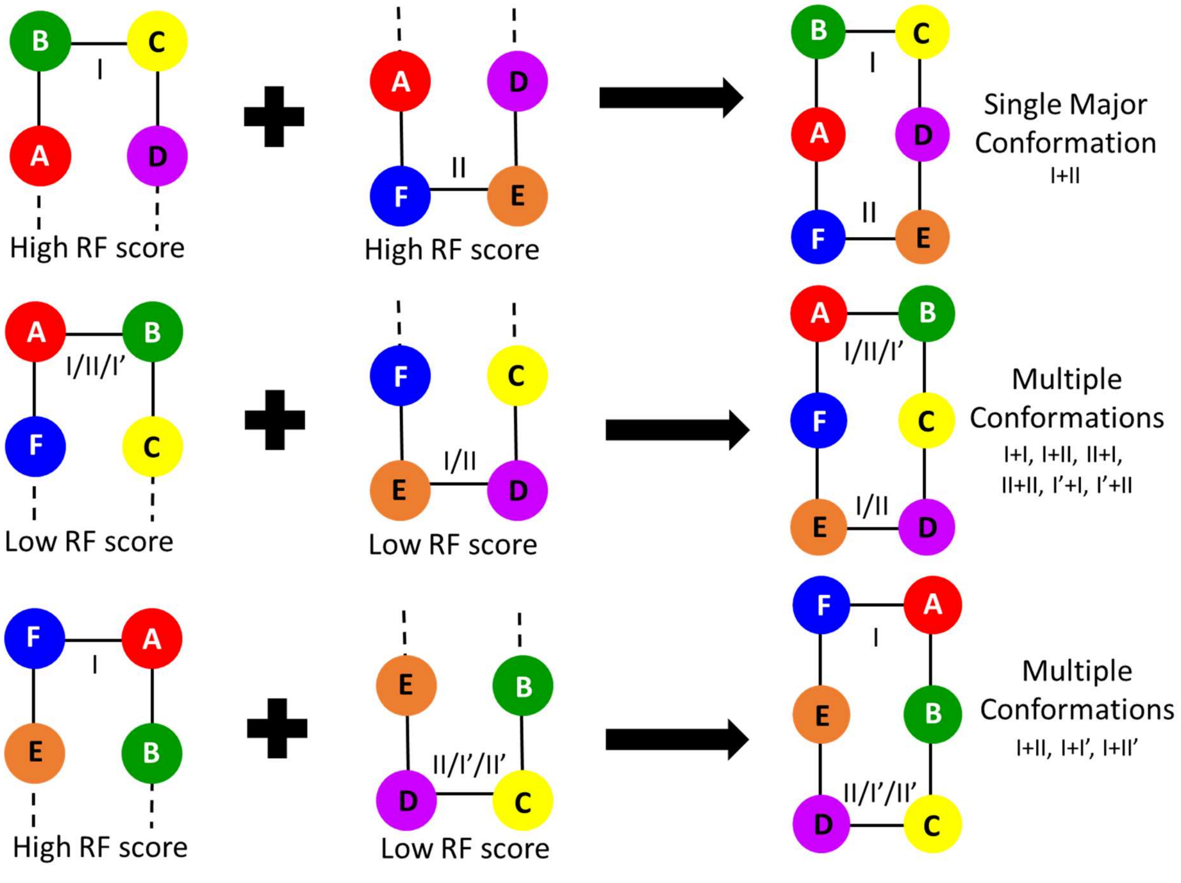


Figure 91: A register with two β-turns with high RF scores may lead to a cyclic peptide with fewer conformations in solution than a register containing a turn with a low RF score.

The sequence c(WGTGCS) showed a high RF score for WGTG to form a type II' turn and GCSW a high score of forming a type I turn (Table 44). The remaining possible amino acid registers have lower RF scores, with at least one of the tetrapeptide sequences having a relatively low score. If a high RF score means a β-turn is more likely to form that particular turn type, and the RF can be used to predict the turn types likely to form within a cyclic hexapeptide, then the WGTG/GCSW register will form only one major conformation containing I and II’ turns. The remaining registers if they occur may be more likely to adopt multiple turn types.

Sequence	Type	Score	Sequence	Type	Score
WGTG	II’	0.86	GCSW	I	1.00
GTGC	II	0.40	CSWG	I	0.92
TGCS	II’	0.12	SWGT	II	0.60

Table 44: RF Scores for the three possible registers of c(WGTGCS).

Three clusters are seen in the BE-META of c(WGTGCS) (Table 46). As predicted the I+II’ turn combination is seen as the only conformation for the WGTG/GCSW register. The GTGC/CSWG II+I conformation forms the major cluster (60%). A smaller cluster with the amino acids in the same GTGC/CSWG register is also seen with a I+I conformation which is consistent with the lower RF scores for this register. The TGCS/SWGT register does not appear.

	Prediction				
Sequence	I	II	I’	II’	VI
WGTG	0.07	0	0	0.93	0
GCSW	1	0	0	0	0
GTGC	0.30	0.70	0	0	0
CSWG	0.96	0.02	0	0.02	0
TGCS	0.40	0.01	0.03	0.56	0
SWGT	0	0.80	0.20	0	0

Table 45: The proportion of decision trees in the RF which assigned the c(WGTGCS) sequences to each turn type category.

30% of the decision trees in the RF predicted GTGC would form a type I turn (Table 45) which is seen in the minor conformation for the GTGC/CSWG register. All turn types seen therefore correspond to predictions made by the RF for the given sequences. Although the RF scores predicted the WGTG/GCSW register would form only one conformation with the other registers more likely to form multiple conformations, the RF scores do not predict which registers are likely to form/why is the GTGC/CSWG register the major one seen rather than the WGTG/GCSW register and why does the TGCS/SWGT register not occur.

Cluster	Population (%)	Turn-type Combination
1	60	II + I
2	27	II’ + I
3	13	I + I

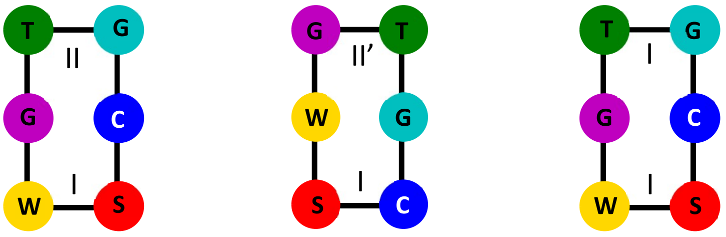


Table 46: Conformations seen in the BE-META of c(WGTGCS).

The restrained simulations (see Chapter 4) on c(GGGGGG) allow conformations that aren’t seen in the restrained simulations of c(XGGXGG). The simulations on c(XGGGGG) indicate that these conformations are lower in energy than those that appear in the c(XGGXGG) simulations but are prevented from occurring by the presence of chiral amino acids. This may mean that the TGCS/SWGT register isn't seen as it would require chiral amino acids to be at both the i/i+3 positions whereas the other sequences have only one chiral amino acid at a i/i+3 position and the other one occupied by glycine.

The WGTC/GCSW II'+I conformation seen in the BE-META has a chiral amino acid at the i position of the type II' turn/i+3 position of a type I turn and glycine at the i+3 position of the type II' turn/I position of the type I turn. The c(XGGGGG) II'+I conformation from the restrained simulations (or the c(GGGXGG) I+II' conformation as they are equivalent) can therefore be used to estimate the relative energy of the conformation compared to other possible conformations seen in the restrained simulations. This gives an estimate of the relative energies of the conformations based on the backbone structure. The GTGC/CSWG II+I conformation has glycine at the i position of the type II/i+3 position of the type I turn and the chiral cysteine at the i+3 position of the type II turn/i position of a type I turn. It therefore matches the II+I conformation from the c(GGGXGG) or I+II conformation from the c(XGGGGG) restrained simulations. The I+II' conformation isn't seen in the c(GGGXGG) restrained simulation. However the II+I conformation is, this likely means it is a lower energy conformation. This might be why there is a larger proportion of the GTGC/CSWG conformation than the WGTG/GCSW conformation in the BE-META.

The GTGC/CSWG register shows interaction of the sidechains of cysteine and tryptophan which are in relatively close proximity to each other (Figure 92). This interaction is not seen in the WGTG/GCSW register so could also be a contributing factor as to why the GTGC/CSWG II+I conformation is the major conformation seen in the BE-META. If a sidechain interaction contributes to the turn type a sequence is likely to form this may be reflected in the RF predictions depending on the available examples in the dataset used to train the RF. For example serine is frequently observed at the i+2 position of type I β-turns due to the ability of the hydroxy sidechain to hydrogen-bond to the peptide backbone in this conformation.[114] If the training set contains many examples of type I turns with serine at the i+2 position it will learn that having serine at this position increases the chance of the turn being type I. Pairwise sidechain interactions will be less well represented within the training set due to the lower probability of the two amino acids occurring together in the relevant positions in the β-turn to form interactions that alter the β-turn type formed. There will therefore be few examples containing such interactions when training the RF so such interactions may not be reflected when predicting the turn types of new sequences. Tryptophan and cysteine have low naturally occurring frequencies so interaction between the two sidechains and any effect on the predicted β-turn type is unlikely to be reflected in the RF as few if any examples will be seen in the training set. Despite this the RF still accurately predicted the β-turn type the sequences would form.

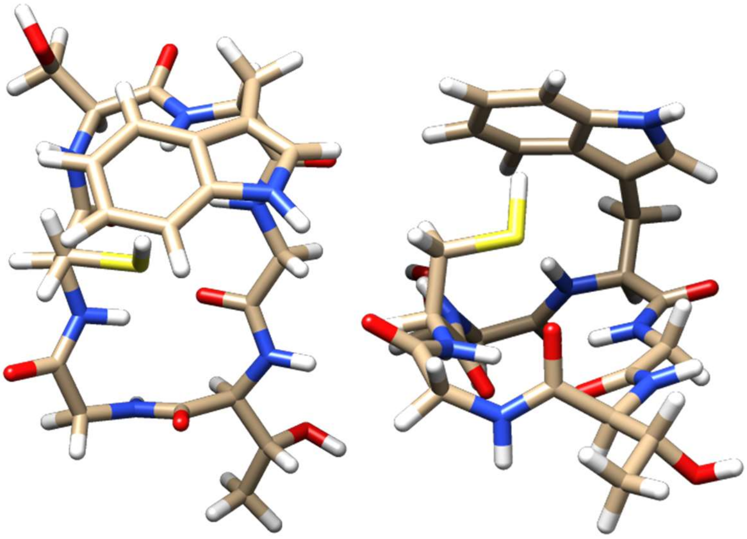


Figure 92: Interaction seen between the cysteine and tryptophan sidechains in the GTGC/CSWG register of c(WGTGCS).

The BE-META of c(WGTGCS) shows that the β-turn types predicted by the RF for a given sequence are seen in the cyclic hexapeptide structure. The RF score alone though is not enough to predict the major conformation seen as it does not predict the amino acid register most likely to form. The restrained simulations were used to determine which turn type combinations produced the lowest energy cyclic hexapeptide structures based on the backbone conformation. The turn combinations that are lower energy structures appeared frequently throughout the different restrained simulations, whereas higher energy turn combinations either made up minor clusters in the simulations or did not appear at all, such as the I+I conformation. A conformation with a lower energy turn type combination based on the restrained simulations may make a particular register more favourable. Choosing a lower energy turn combination based on the restrained simulations in combination with using the RF scores may therefore lead to better predictions. A series of conditions were selected to test the effects of combining turn combination with RF score.

Sequences were searched for whereby one of the three possible registers had different RF scores and/or turn-type combinations from the other two registers. BE-META was then used to see if that register was the major conformation seen. Various case studies were used to choose the sequences based on whether the selected register had high/low RF scores and a predicted high/low energy turn type combination compared to other possible registers (Table 47). Four different combinations were tested:

1. A low energy turn combination with high RF scores for one register with the remaining registers having both low RF scores and high energy turn combination
2. A high energy turn combination with high RF scores for one register with the remaining registers having lower energy turn combinations but low RF scores
3. A low energy turn combination with high RF scores for one register with the remaining registers having high energy turn combinations and high RF scores (High RF scores for all registers but each register has different predicted turn combinations)
4. A low energy turn combination with low RF scores for one register with the remaining registers having high energy turn combinations but high RF scores

Case study	Test Sequences	Register 1 Turn Combination Energy	Register 1 RF Score	Register 2 & 3 Turn Combination Energy	Register 2 & 3 RF Score	Predicted Outcome
1	NFEWSG RGNQPG	Low	High	High	Low	Register 1 has both "better" RF Scores and turn combination so should be the major conformation
2	RGSQGW NWQNVA	High	High	Low	Low	Register 1 won't be main register as high energy turn combination so instead register 2 or 3 will form with multiple conformations
3	EGDSAR	Low	High	High	High	All registers have high RF Scores so the register with lower energy turn combination is predicted to form the major conformations
4	NSKSED	Low	Low	High	High	Register 1 will be the main register but multiple conformations will be seen

Table 47: Sequences used to test different combinations of RF Scores and high/low energy turn type combinations.

5.4.2 Case study 1

The restrained simulations show which conformations are more energetically stable based on the backbone structure of the peptide. The RF predicts what turn types a sequence is likely to form which is related to the energetic preference of that sequence for a given turn type. By combining a stable turn type combination from the restrained simulations with sequences predicted to form those turn types, then both the backbone and sequences should be in a relatively low energy conformation allowing for the design of a more stable cyclic peptide. Other registers which do not form the stable turn combination or have less strong predictions were therefore hypothesised to be less likely to occur.

Although the RF was able to predict the β-turn types seen in c(WGTGCS) the major conformation could not be predicted based on RF scores alone. The restrained simulations determined the lowest energy turn type combination for a cyclic peptide to adopt. The c(WGTGCS) structure was seen to form the lowest energy turn combination out of the β-turn types the RF predicted it was able to form. Further sequences were therefore searched for which were predicted to form low energy turn type combinations based on the restrained simulations. To try favour one register only, sequences where the other registers were predicted to form relatively high energy turn combinations were selected. The sequences were further narrowed down by making the register which formed the low energy turn combination have high RF Scores compared to the remaining registers. This should allow for the cyclic peptide to form only one major conformation in solution and so is potentially a way of designing structurally stable cyclic peptides.

5.4.2.1 c(NFEWSG)

The vast majority of sequences are predicted to form a type I β-turn by the RF. A low energy turn combination containing a type I turn was therefore chosen to increase the number of sequences found. The I+II' combination was shown to be relatively low in energy compared to other turn combinations in the c(XGGXGG) restrained simulations, being the major conformation seen in the type I restrained simulations. The RF was therefore used to search for sequences with a very high prediction of forming a type I turn and sequences that have a high probability of forming a type II' turn with overlapping chiral amino acids at the i/i+3 positions. The sequences were then narrowed down based on the remaining registers being predicted to form a high energy turn combination and the I+II' register having a very high RF score compared to the remaining registers. The sequence c(NFEWSG) has two overlapping β-turns predicted to form a type I and a type II' turn respectively each with a RF score of 1 (Table 48). The other sequences that would form a β-turn if the register changed have a RF score below 0.4 and are predicted to form the relatively high energy I+I turn combination.

Sequence	Type	Score	Sequence	Type	Score
FEWS	I	1.0	SGNF	II'	1.0
GNFE	I	0.22	EWSG	I	0.32
NFEW	I	0.38	WSGN	I	0.22

Table 48: Predicted turn types and RF scores of the three possible registers of c(NFEWSG).

The FEWS/SGNF register forming a type I+II' conformation should be seen in the BE-META of c(NFEWSG) if the RF turn type predictions in combination with selecting low energy turn combinations based on the restrained simulations can be used to predict the conformation of the cyclic peptide. The other registers with low scores shouldn't occur, or if they do, they should only be minor conformations.

	Prediction				
Sequence	I	II	I'	II'	VI
SGNF	0	0	0	1	0
FEWS	1	0	0	0	0
GNFE	0.61	0.06	0.33	0	0
EWSG	0.66	0	0	0.34	0
NFEW	0.69	0.12	0	0.19	0
WSGN	0.61	0.13	0.26	0	0

Table 49: The proportion of the decision trees in the RF which assigned each of the tetrapeptide sequences to each class.

The BE-META of c(NFEWSG) shows two clusters (Table 50). Neither of the possible amino acid registers which have sequences with low RF scores and are predicted to form high energy turn combinations appear. The predicted conformation with a I+II' conformation is seen, but it is not the major cluster. The major cluster has the amino acids in the predicted FEWS/SGNF register but has a I+IV conformation. The EW subsequence still forms the predicted type I turn but the GN subsequence forms a type IV turn which is similar to a type I turn but distorted away from ideal values. The type IV turn is equivalent to a type IV_3 turn based on de Breverns classification of β-turns.[123] The type IV$_3$ turn often appeared at the unrestrained side of the cyclic peptide during the restrained simulations, appearing more frequently than a type I turn. It may therefore be a relatively low energy turn-type in the context of cyclic peptides. The type IV$_3$ turn only appears in the Loop Database when the < 7 Å distance between α-carbons rather than hydrogen-bonded definition of a β-turn is used and only in the β-turns rather than the hairpin dataset. The prevalence of the type IV$_3$ β-turn in cyclic peptides but not in the β-hairpin dataset may indicate β-hairpins do not make good models for the β-turn structure seen in cyclic peptides. The density-based clustering algorithm used to cluster the β-turns in the Loop Database did not separate the type IV$_3$ turns from the type I turns in the datasets they appeared in. The RF does not predict the likelihood of a tetrapeptide sequence forming a type IV$_3$ turn. The RF would have to be trained on a dataset containing type IV$_3$ turns as a category in order to be able to predict the occurrence of this conformation.

Cluster	Population (%)	Turn-type Combination
1	68	IV + I
2	32	II’ + I

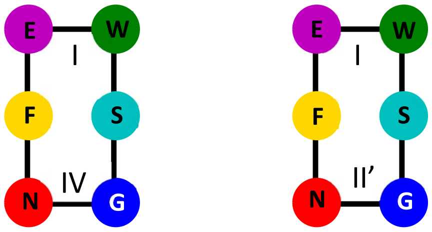


Table 50: Clusters seen in the BE-META of c(NFEWSG). The predicted conformation is highlighted in blue.

Based on the restrained simulations the I+II’ conformation should be a lower energy turn combination than the I+IV$_3$ conformation for the FEWS/SGNF register - the I+II’ combination consistently forms a larger cluster in the type I c(XGGXGG) restrained simulations. Despite being a higher energy conformation, if the SGNF sequence more favourably forms the type IV$_3$ turn type, then the I+IV$_3$ conformation could be predicted to form the major cluster. The type IV$_3$ turn forms at the GN turn. Although predicted to form a type II’ turn GN is a particularly flexible sequence with many of the sequences seen in the database which form multiple turn types (especially those which were seen to form four turn-types) containing the GN subsequence at the i+1 and i+2 positions of the β-turn.

As GN is a particularly flexible sequence it may be expected to form a type II’ turn in order to form the lowest energy turn combination rather than showing a sequence preference for a type IV$_3$ turn. The feature importance of the RF shows that the i/i+3 positions can alter the turn type a sequence will form so, despite containing the flexible GN subsequence, the β-turn could potentially have a strong preference for a given β-turn type. The smallest difference in energy between the I+II’ and I+IV$_3$ conformations seen in the c(XGGXGG) restrained simulations is approximately 5 kJ/mol. This is a sufficiently small energy difference that other factors such as sidechain interactions could overcome selecting the lowest energy turn combination.

The BE-META shows that the phenylalanine and tryptophan sidechains interact throughout the trajectory (Figure 93). This interaction is only seen in the I+IV$_3$ conformation and not the I+II’ conformation where the glutamic acid sidechain prevents close proximity of the phenylalanine and tryptophan sidechains. The I+II’ conformation however shows the serine residue forming more hydrogen-bonds with the peptide backbone. Although the restrained simulations predict the I+II’ conformation to be lower energy this is based on backbone structure. The sidechain interactions seen here mean the I+IV$_3$ conformation is lower energy for this sequence. Although π-stacking between the phenylalanine and tryptophan sidechains is seen only throughout the I+IV$_3$ conformation, both the I+IV$_3$ and I+II' conformations contain the FEWS subsequence as part of a type I turn. The tryptophan therefore should have very similar dihedral angles between the two conformations. The phenylalanine however has slightly different dihedral angles (Figure 94) which could be why the π-stacking is not seen in the I+II' conformation. The two turns are able to affect the conformation of each other allowing for different sidechain interactions to occur in different conformations.

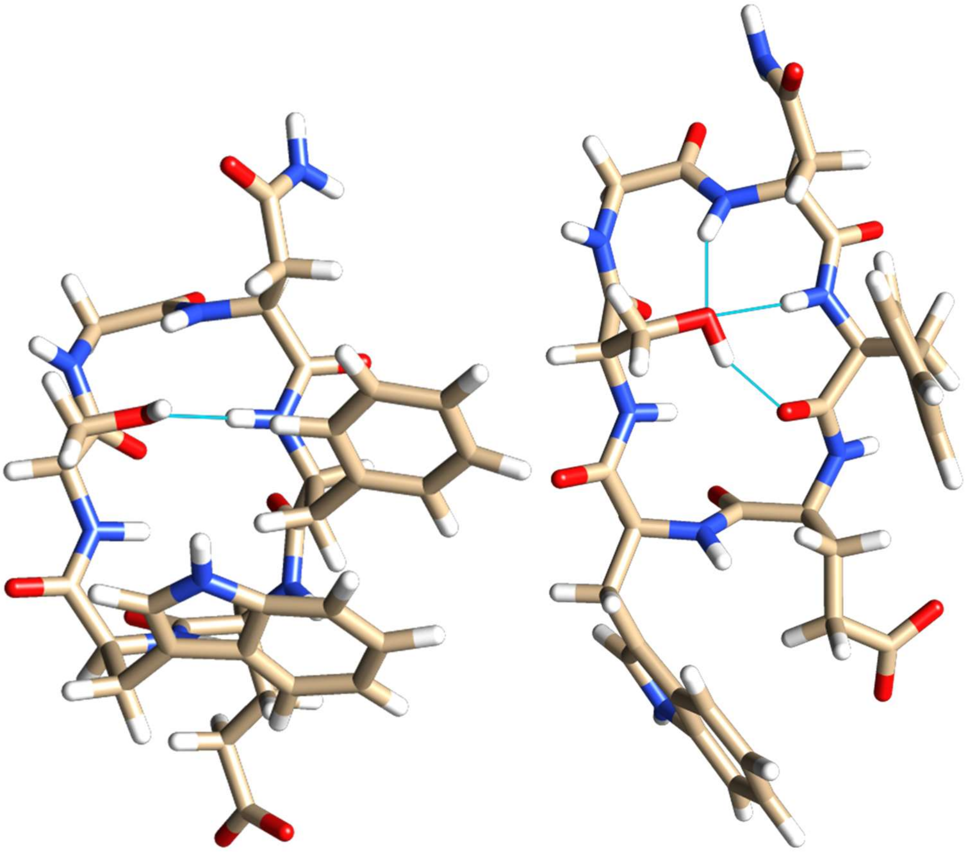


Figure 93: I+IV (left) and I+II' (right) conformations.

The predicted conformation makes up approximately one third of the clustered conformations. As the only other conformation seen contains a turn type the RF cannot predict, using a high score for the desired turn combination and low scores for the other registers is still a useful way of finding potential sequences, but the occurrence of other turn-types is possible.

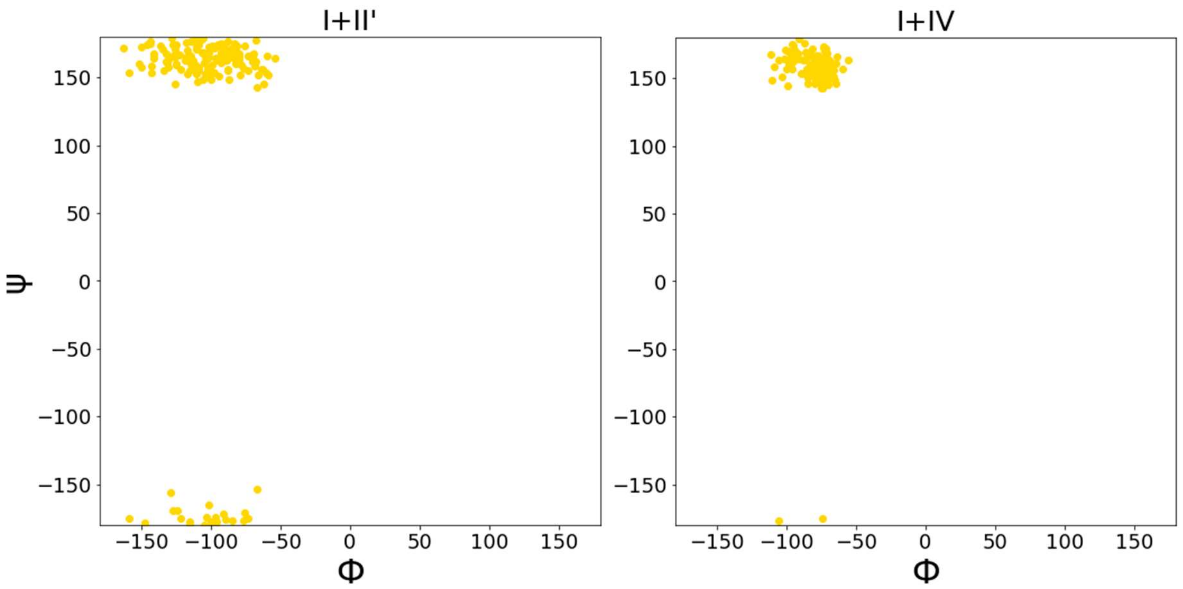


Figure 94: Phenylalanine dihedral angles in the I+II' and I+IV conformations.

5.4.2.2 c(RGNQPG)

To further test the use of high RF scores to select for a desired conformation a search was carried out to identify a sequence likely to form a II+II' conformation. The II+II' conformation is also a relatively low energy conformation based on the restrained simulations. There are no sequences where the other registers RF scores are as low as they are for c(NFEWSG). The sequence c(RGNQPG) however has low RF scores for all other tetrapeptide sequences in the remaining registers, with the exception of GNQP which has a score of 0.62 (Table 51). Proline shows unique Ramachandran preferences compared to the other L-amino acids and often induces a turn in peptide structure, as shown by the analysis of the turns in the Loop Database, so is very unlikely to occur at the i/i+3 positions therefore this combination of turns/amino acid register would probably not occur. If the RF prediction is correct the BE-META should show the II+II' conformation as the major cluster.

Sequence	Type	Score	Sequence	Type	Score
RGNQ	II'	1	QPGR	II	1
GNQP	I	0.62	PGRG	I	0.16
NQPG	I	0.30	GRGN	II	-0.28

Table 51: Predicted turn types and RF scores for c(RGNQPG).

BE-META on c(RGNQPG) shows three clusters (Table 53). The major cluster, making up 52% of the clustered conformations, is the II'+II predicted by the RF in combination with choosing low energy turn combinations from the restrained simulations. A smaller cluster is also seen (17%) with the amino acids in the same register. The QPGR subsequence still forms a type II turn but the RGNQ sequence, similar to c(NFEWSG) where the GN turn is seen to form multiple turn types, forms a type I turn rather than the predicted type II' turn. The RF predicts most sequences to form a type I turn but none of the decision trees in the RF predicted the GN turn would form a type I rather than a type II turn (Table 52). Many examples of sequences containing GN turns were removed from the training set as they formed multiple different types of β-turns. Although this led to improved accuracy of the RF, it may actually be beneficial to include such sequences to better predict what turn types more flexible sequences will form. The inclusion of glycine, aspartic acid and asparagine could contribute to cyclic peptide structures which interconvert more readily between β-turn types based on the fact they make up the majority of sequences which formed multiple turn-types. In the c(XGGXGG) restrained simulations the II+II’ conformation is lower energy than the II+I conformation.

	% Prediction				
Sequence	I	II	I’	II’	VI
RGNQ	0	0	0	1	0
QPGR	0	1	0	0	0
GNQP	0.81	0.01	0.18	0	0
PGRG	0.58	0.03	0.02	0.37	0
NQPG	0.65	0	0	0	0.35
GRGN	0.28	0.36	0.36	0	0

Table 52: The proportion of the decision trees in the RF which assigned each of the tetrapeptide sequences to each class for c(RGNQPG).

The second largest cluster (31%) contains a cis proline and has a VI+II conformation. Type VI turns have been observed to appear more in small cyclic peptides than typically seen in linear peptides.[194, 198] The RF is trained on β-turns from loops not cyclic peptides so this may lead to differences in what is predicted and what is observed. 35% of the trees in the RF predicted the NQPG sequence would form a type VI turn with the remaining trees predicting it would form a type I turn. The largest class in the dataset used to train the RF was type I turns so there may be an overprediction for these turn types as a high accuracy can be achieved when training the RF by assigning most sequences to the major class. There was also limited examples of type VI turns in the training dataset which could make prediction of type VI turns more difficult.

Cluster	Population (%)	Turn-type Combination
1	52	II’ + II
2	31	VI + II
3	17	I + II

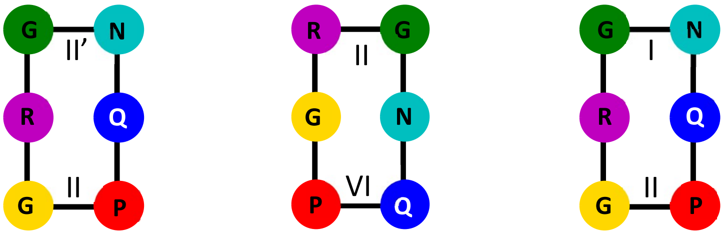


Table 53: Clusters seen in the BE-META of c(RGNQPG). Predicted cluster is highlighted in blue.

Type VI turns were not included in the restrained simulations so it is not possible to predict the relative energy of conformations containing type VI turns compared to the major II+II’ conformation. The differences between the c(XGGXGG) and c(XGGGGG) restrained simulations show that having glycine at the i/i+3 position can lead to a more favourable conformation. The NQPG/GRGN register may therefore be forming a stable turn type conformation. Having glycine in the cyclic peptide sequence can lead to increased flexibility which can make structure prediction more difficult.

Based on these examples, although other conformations are also seen, using the RF scores to look for sequences which have a high score for the desired combination and low scores for the other registers, when used in combination with picking low energy turn type combinations for the high scoring register selects sequences that form the predicted conformation. β-turn types the RF cannot predict may appear however and some β-turn types may be more favourable than predicted by the RF due to the strained cyclic peptide environment. Further sequences were selected to test whether a less structured cyclic peptide would result when, rather than the high scoring register being predicted to form a low energy turn combination, the high scoring register has a high energy turn combination based on the restrained simulations. The other registers with lower RF scores may then become more favourable. Even if they are predicted to also form a high energy turn combination, the lower RF scores may mean the sequences are more likely to form other turn types. Other turn types which form a lower energy turn type combination may therefore be seen for the lower scoring registers.

5.4.3 Case study 2

The previous sequences look for a high RF score for a particular conformation and a low score for the remaining registers but the conformation with high scores was always selected as a relatively low energy conformation based on the restrained simulations. A sequence was selected which had a high energy conformation as the register with high RF scores. The cyclic peptides may therefore form multiple conformations in solution as the other registers are predicted to form lower energy turn combinations.

5.4.3.1 c(RGSQGW)

A sequence was selected which had a register with RF scores of 1, predicted to form a I+I conformation. This has been shown to be a high energy conformation in the restrained simulations, and as the majority of sequences are predicted to form type I turns by the RF there are more sequences to choose from. The available sequences were narrowed down so the remaining registers were made up of sequences with the lowest RF scores possible. The sequence c(RGSQGW) was found. The GWRG/GSQG register is predicted to form a I+I conformation with RF scores of 1 for both tetrapeptides (Table 54). The remaining registers have RF scores below 0.05. The WRGS/SQSW register is predicted to form a I'+II conformation which (similar to the I+I conformation) is very high energy and did not appear in the c(XGGXGG) restrained simulations. The final RGSQ/QGWR register is predicted to form a I+II' conformation which is a relatively low energy turn type combination based on the restrained simulations.

Sequence	Type	Score	Sequence	Type	Score
GWRG	I	1.0	GSQG	I	1.0
WRGS	I'	0.04	SQGW	II	0.04
RGSQ	I	0.02	QGWR	II'	0.02

Table 54: RF scores for c(RGSQGW).

The RGSQ/QGWR register is predicted to form the lowest energy conformation based on the restrained simulations and may therefore be predicted to form the major conformation. The low RF scores however mean the sequence may be flexible so it may be seen to form other turn types. Additionally although predicted to form the unfavourable I'+II conformation the WRGS/SQFW register may also occur as a different turn-type combination due to the predicted flexibility of the register. The RGSQ/QGWR register is not predicted to occur based on the RF scores predicting it to have a high probability of forming the high energy I+I conformation.

	Prediction				
Sequence	I	II	I’	II’	VI
GWRG	1	0	0	0	0
GSQG	1	0	0	0	0
WRGS	0.01	0.47	0.52	0	0
SQGW	0.02	0.52	0.46	0	0
RGSQ	0.51	0.09	0.03	0.37	0
QGWR	0.48	0	0.01	0.51	0

Table 55: Proportion of the decision trees in the RF which assigned the tetrapeptide sequences to different β-turn types for c(RGSQGW).

The BE-META of c(RGSQGW) shows 4 conformations (Table 56). The I+I conformation with the high RF scores is seen but is a small cluster making up around 19% of the clustered data. Despite the very high RF score other turn types are also seen for the GSQG and GWRG sequences with the remaining two smaller clusters (cluster 2 and 4) having the same amino acid register as the I+I conformation but form a I+II and IV+I conformation respectively. The type IV turn is not a type IV_3 turn this time but is a type IV turn that very infrequently appeared in some of the c(XGGXGG) simulations. The GSQG/GWRG register however has glycine rather than chiral amino acids at the i/i+3 positions so should be more analogous to the c(GGGGGG) restrained simulations than the c(XGGXGG) simulations. The sequences may be forming other turn types despite the high RF score due to the RF potentially overpredicting type I turns. Alternatively the preference for more favourable turn combinations may lead to sequences adopting other turn types than they otherwise would.

Cluster	Population (%)	Turn-type Combination
1	48	II’+II’
2	21	I+II
3	19	I+I
4	13	VI+I

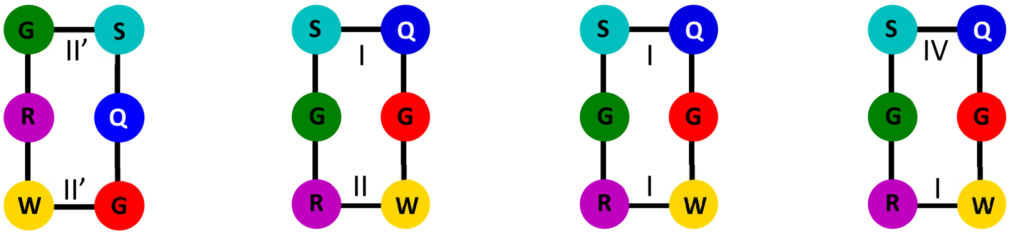


Table 56: Clusters seen in the BE-META of c(RGSQGW).

The major conformation making up almost 50% of the clusters is a II’+II’ conformation with the RGSQ/QGWR amino acid register. Although the RGSQ sequence was predicted to form a type I turn the distribution of the number of trees in the RF which classifies the sequence shows both the RGSQ and QGWR sequences have a relatively high chance of being a type II’ turn (as do most sequences with Glycine at the i+1 position) (Table 55). The II’+II’ conformation is consistently one of the lowest energy turn combinations in the c(XGGXGG) restrained simulations. It is difficult to tell if the conformation is predicted to be higher or lower energy than the c(GGGGGG) I+I conformation however as including a glycine at the i/i+3 positions can lead to lower energy structures but the I+I conformation is a high energy turn combination based on the restrained simulations. Assuming the II’+II’ conformation is lower energy then the major conformation of c(RGSQGW) is not predicted by the highest RF scores but instead is a lower energy turn-type combination. This is preferred even to glycine being at both the i/i+3 positions in this sequence.

The adjacent arginine and tryptophan sidechains are seen to interact through π-stacking in the BE-META for all turn type combinations (Figure 95). Hydrogen bonds are also seen between the serine and glutamine sidechains. This could also potentially alter which turn types are seen. If the sidechain interactions alter the most stable turn type combinations, the RF prediction of turn type will only be accurate if the influence of the sidechain interactions on which turn type forms is seen in the dataset used to train the RF.

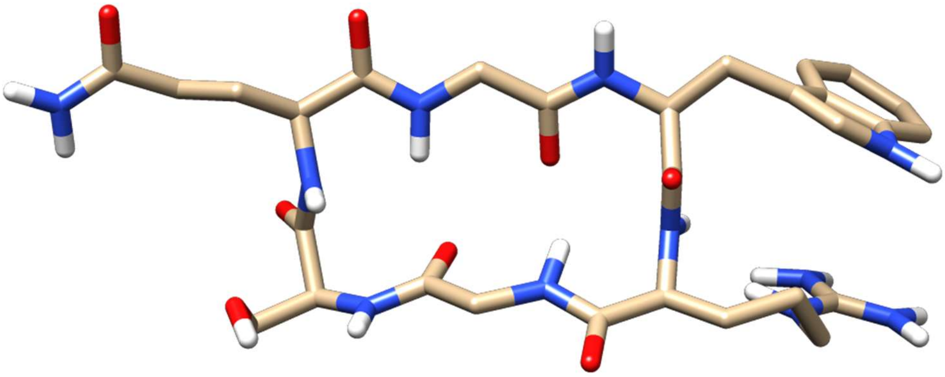


Figure 95: Arg-Trp π-stacking in the I+II conformation.

Although the high RF scoring register is seen in the BE-META it formed only a minor cluster. Instead a lower scoring register with what is likely a lower energy turn combination formed the major conformation of the peptide. This may indicate that when using the RF to design cyclic peptides it is better to choose a low energy turn combination with high RF scores like the c(NFEWSG) and c(RGNQPG) sequences for more reliable prediction. In the c(XGGXGG) restrained simulations each turn combination, with few exceptions, produced only one conformation regardless of the amino acids present at the i/i+3 positions. It may be that when glycine occupies both positions as in the GSQG/GWRG register of c(RGSQGW), due to the flexible nature of glycine, more conformations are able to form with smaller energy differences between them as there are more ways to join the two turn types. As the c(RGSQGW) sequence contains two glycine residues a further sequence with high RF scores for only one register with a high energy turn combination was searched for but which contained no glycine residues.

5.4.3.2 c(NWQNVA)

The NWQN/NVAN register of c(NWQNVA) is predicted to form a I+I combination with each β-turn having a score of 1 (Table 57). The remaining registers are also predicted to form I+I conformations but have low scores. The RF predicts the QNVA sequence will form a type I turn despite a 50:50 split of the decision tree votes between the type I and I’ categories (Table 58). Where there is a tie between categories such as this the RF assigns the sequence to the largest of the categories based on the data in the training set. As there were more type I than type I’ β-turns in the training set the QNVA sequence is therefore assigned as a type I turn.

Sequence	Type	Score	Sequence	Type	Score
NWQN	I	1	NVAN	I	1
WQNV	I	-0.06	VANW	I	0.04
QNVA	I	0	ANWQ	I	0.02

Table 57: Predicted turn types and RF scores for c(NWQNVA).

Although all registers are predicted to form the high energy I+I turn type combination, it may be the peptide will form a conformation with one of the lower scoring RF registers. As they have lower RF scores they are predicted to be able to form other turn types. They therefore may be able to form β-turns which allow for the cyclic peptide to adopt a lower energy turn type combination.

	Prediction				
Sequence	I	II	I’	II’	VI
NWQN	1	0	0	0	0
NVAN	1	0	0	0	0
WQNV	0.47	0.20	0.20	0.13	0
VANW	0.52	0	0	0.48	0
QNVA	0.50	0	0.50	0	0
ANWQ	0.51	0	0.49	0	0

Table 58: The proportion of decision trees in the RF which assigned each sequence to the different β-turn classes.

The BE-META shows three clusters none of which is a I+I conformation (Table 59). All have the same register: WQNV/VANW. Although predicted to form a I+I conformation the WQNV/VANW register has very low RF scores meaning there is a large difference in the number of decision trees in the RF that assigned the sequences to each class. The VANW sequence, based on the decision tree votes, also has a high chance of being a type II’ turn whereas the WQNV sequence is potentially very flexible with decision trees assigning it to four different classes. This flexibility is seen in the observed clusters where the WQNV subsequence form a different turn type in each cluster. Having low scoring sequences may not mean they are poor at forming turns so it may be preferable for low scoring registers to form turns over higher scoring registers. Especially when the high scoring registers are predicted to form unfavourable turn types - a lower energy conformation may be formed by alternative registers.

Cluster	Population (%)	Turn-type Combination
1	62	II+II
2	21	II’+II
3	18	I+II

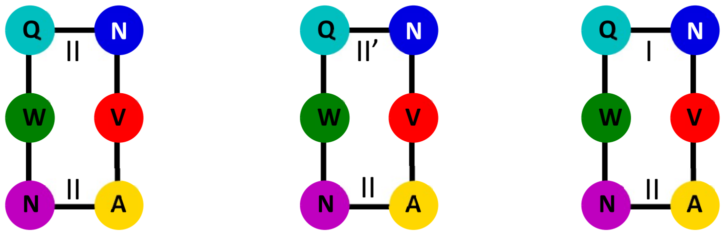


Table 59: Clusters seen in the BE-META of c(NWQNVA).

The three clusters seen all form a lower energy turn type combination than the predicted I+I conformation based on the c(XGGXGG) restrained simulations. Despite being predicted to form a type I or I’ turn VANW is only seen to form a type II turn. It has previously been noticed having a valine at the i/i+3 position often favors cyclic peptides with a relatively stable II+II conformation.[228] As the β-turns used to train the RF are not found in cyclic peptides this may be why the type II conformation is not predicted as it is a structure unique to the cyclic peptide environment. The restrained simulations, including those on c(VGGVGG), indicate the II+II’ conformation is lower energy than the II+II conformation. It may be that when chiral amino acids are further included at the i+1/i+2 positions this changes and this is why the II+II rather than the II+II’ forms the major cluster in the BE-META. Further constrained simulations could be run where the i+1 and i+2 positions are alternatively substituted for chiral amino acids and the effects on the lowest energy conformations determined.

When the RF was used to select cyclic peptide sequences with only one register having high RF scores, if the high scoring register was predicted to form a relatively low energy turn combination it was observed in the BE-META of the sequence. However the c(RGSQGW) and c(NWQNVA) sequences indicate that when the high scoring register is predicted to form a high energy turn combination, it is much less likely to be the major conformation seen in the cyclic peptide. This could potentially be due to the lower scoring registers being more flexible and therefore allowing for the formation of a lower energy structure. Having a high RF score does not mean that it is more favourable for that sequence to form a β-turn. The RF predicts the β-turn type a sequence will form if it were to form a β-turn not how likely a sequence is to form a β-turn. As the high energy turn combination was shown to occur infrequently despite high RF scores it may be possible to select for a given register by selecting sequences whereby the undesired registers have high scores for an unfavourable turn type combination.

5.4.4 Case study 3

If having a high RF score means a sequence is very inflexible and very unlikely to form another turn type, then the registers which would require the peptide to form a high energy turn combination will not occur. Instead the register where the peptide is able to form a lower energy conformation based on the restrained simulations will be the major conformation. Selecting sequences where certain registers do not occur because they can only form a high energy conformation would therefore be a potential way of designing cyclic peptides with a more stable conformation as some registers are inaccessible.

5.4.4.1 c(EGDSAR)

A sequence was searched for which has a register with a high RF score for a favourable turn-type combination based on the restrained simulations, but the remaining registers have high scores for a less stable turn-type combination. The I+II' combination is low energy in the c(XGGXGG) restrained simulation (unless valine is present at a i/i+3 position). Therefore a sequence was looked for which has chiral amino acids (which aren't β-branched) at the i/i+3 positions which is predicted to form a I+II' turn combination with high RF scores. The sequences were then filtered based on the remaining registers having a high score for a I+I or I+II combination which are high energy conformations in the restrained simulations. The sequence c(EGDSAR) was found (Table 60). If the RF predictions are correct, then the EGDS/SARE register should form the major conformation of the peptide as it allows for the lowest energy conformation based on the restrained simulations.

Sequence	Type	Score	Sequence	Type	Score
EGDS	II'	1.0	SARE	I	1.0
GDSA	I	1.0	AREG	I	1.0
DSAR	I	1.0	REGD	II	0.98

Table 60: RF turn type predictions and scores for c(EGDSAR).

Only one cluster was seen in the BE-META of c(EGDSAR) with the cyclic peptide having a rigid structure with only one major conformation in solution (Table 62). Neither of the registers predicted to form high energy conformations are seen. Designing stable cyclic peptides with only one major conformation in solution may therefore be possible by choosing sequences where two of the three registers are very unfavourable due to having strong preferences for forming less compatible turn types. However although the EGDS/SARE register is seen in the final structure the sequence does not form the predicted I+II' conformation. Instead, although the EGDS sequence forms the predicted type II' turn, the SARE subsequence form a type IV_3 turn. A IV_3+II' conformation is therefore seen.

	Prediction				
Sequence	I	II	I'	II'	VI
EGDS	0	0	0	1	0
SARE	1	0	0	0	0
GDSA	1	0	0	0	0
AREG	1	0	0	0	0
DSAR	1	0	0	0	0
REGD	0.01	0.99	0	0	0

Table 61: The proportion of decision trees in the RF which assigned the tetrapeptide sequences to each class.

Based on the restrained simulations the IV_3+II' conformation is lower energy than the I+I or I+II conformations predicted for the other registers. It is very similar to the I+II' predicted conformation but is lower energy. Although technically classified as separate turn types there is some overlap in the dihedral angles of the i+1 and i+2 positions used to define the type I and IV_3 turns. The clusters seen in the restrained simulations also show the two conformations are very similar, with only the ψ dihedral angles of the i+2 residue of the type I or IV_3 turn differing between the I+II' and IV_3+II' conformations (Figure 96). It is therefore likely that for this conformation there is a continuum between the type I and IV_3 turns and the type IV_3 turn is lower energy. The two turn types are not separated by the clustering algorithm used on the Loop Database showing the similarity between them. So although the RF cannot predict the occurrence of type IV_3 turns they are very similar to type I turns so could be treated as equivalent as long as the conformation does not differ at any other positions.

Cluster	Population (%)	Turn-type Combination
1	100	IV_3+II'

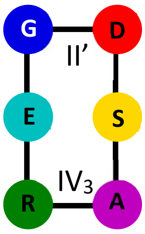


Table 62: A II'+IV cluster is seen in the BE-META of c(EGDSAR).

Looking at the BE-META trajectory, sidechain interactions are seen which could alter the conformation observed from that predicted from the RF (Figure 97). Although the arginine and aspartic acid are only involved in hydrogen-bonding for a small proportion of the simulation, serine is seen to hydrogen bond with the peptide backbone for 65% of the simulation.

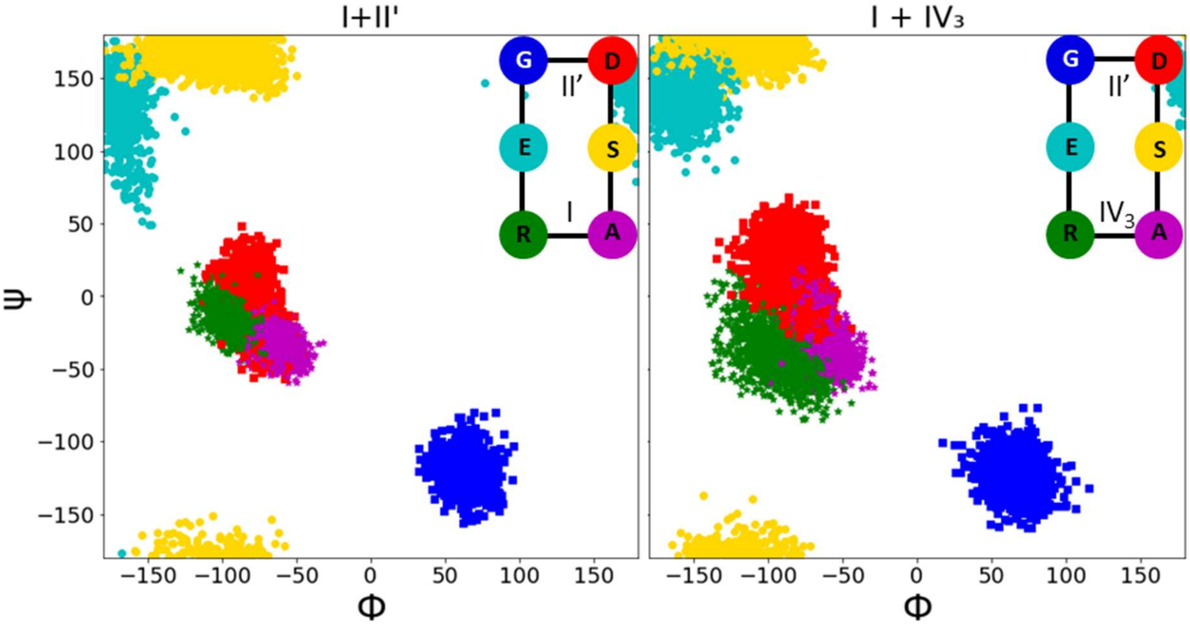


Figure 96: The Ramachandran plot for the IV₃+II’ conformation adopted by c(EGDSAR) and the I+II’ conformation seen in the c(AGGAGG) type I restrained simulation.

Although a type IV₃ turn rather than a type I turn was seen, by choosing a sequence with one register with high RF scores for a favourable turn combination and the remaining registers having high scores for a high energy turn conformation, a cyclic peptide with a single major conformation was identified. If sequences are able to only form one β-turn type and if combining the two turn types leads to a very high energy conformation this can make a register unfavourable. The peptides c(NFEWSG) and c(RGNQPG) were designed by selecting sequences where only the desired register has a high RF score and low energy turn combination with the remaining registers having low RF scores for more unfavourable turn combinations. The two methods differ depending on whether high or low RF scores are used for the undesired registers. Based on these very limited examples it is difficult to conclude which method is better. The peptide c(EGDSAR) only had one major conformation whereas c(NFEWSG) and c(RGNQPG) formed multiple conformations so it is possible that making a register unfavourable due to a high probability of forming a high energy turn combination is a better way of designing cyclic peptides with only one major conformation in solution. Registers with lower scores are more flexible so may be able to form a low energy turn combination despite the overall RF prediction.

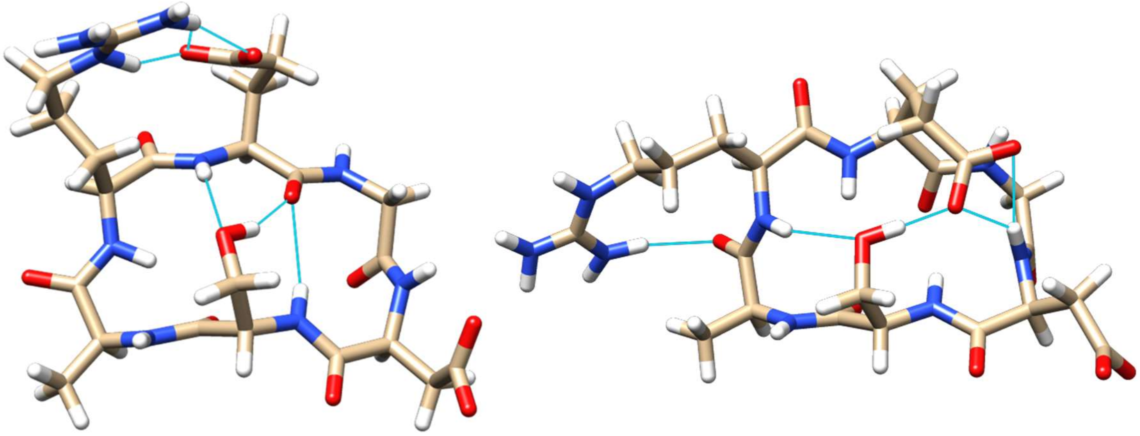


Figure 97: Sidechain hydrogen bonding in c(EGDSAR).

5.4.5 Case study 4

5.4.5.1 c(NSKSED)

By selecting a sequence with two registers with very high RF scores for an unfavourable turn combination the conformationally rigid cyclic peptide c(EGDSAR) was identified. The register predicted to form a low energy turn type combination for c(EGDSAR) also had very high RF scores. If the low energy conformation register has lower RF scores the peptide may be more likely to form more conformations in solution. To test whether the register with the low energy turn combination would still be seen to form the predicted conformation if it had much lower RF scores, a sequence was next chosen which also has high RF scores for the high energy I+I conformations for two registers, but low RF scores for the remaining register which was predicted to form a lower energy turn combination. If the lower energy turn-type combination is still favoured despite the low RF scores it will be the major cluster in the BE-META. The peptide may be less conformationally rigid however as the lower RF scores may mean the lower energy register may form multiple turn type combinations.

Sequence	Type	Score	Sequence	Type	Score
EDNS	I	0.44	SKSE	II’	0.34
KSED	I	1.00	DNSK	I	1.00
SEDN	I	1.00	NSKS	I	1.00

Table 63: RF predictions and scores for c(NSKSED).

The peptide c(NSKSED) was predicted to form the relatively low energy I+II’ conformation for one register, with RF scores less than 0.5 (Table 63). The remaining registers are predicted to form I+I conformations with RF scores of 1. Three clusters are seen in the BE-META of c(NSKSED) (Table 65). The EDNS/SKSE register is seen in cluster 2 but the predicted I+II’ conformation isn’t seen. Instead a I+IV$_3$ conformation is seen. The type IV$_3$ turn is similar to a type I turn and the SKSE turn, although predicted to have the highest chance of forming a type II’ turn, 32% of the decision trees in the RF predicted it would form a type I turn (Table 64). The major conformation (56%) is the SEDN/NSKS register which forms the predicted I+I conformation despite being an unfavourable turn combination. Cluster 3 is a similar cluster where the NSKS sequence forms a type IV$_3$ rather than a type I turn.

	% Prediction				
Sequence	I	II	I'	II'	VI
EDNS	0.72	0.13	0.13	0.02	0
SKSE	0.32	0	0.01	0.67	0
KSED	1	0	0	0	0
DNSK	1	0	0	0	0
SEDN	1	0	0	0	0
NSKS	1	0	0	0	0

Table 64: The proportion of the decision trees in the RF which predicted the sequences would form each turn type.

Only one of the I+I high scoring registers was seen. The assumption was made that a high RF score means a sequence is more likely to form a given turn type. It might be useful however to include a different parameter to estimate the likelihood of a sequence forming a β-turn. If the SEDN/NSKS sequences are more likely to form β-turns than the KSED/DNSK sequences it could explain why only the SEDN/NSKS register was seen despite the two having the same RF scores and predictions.

Cluster	Population (%)	Turn-type Combination
1	56	I+I
2	30	I+IV$_3$
3	14	I+IV$_3$

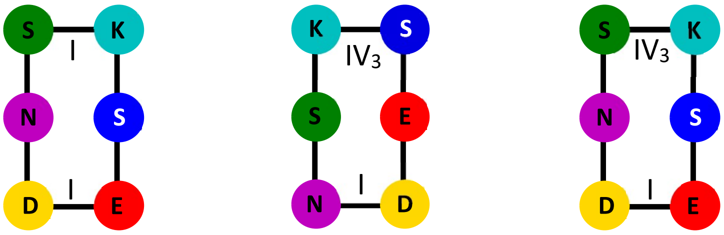


Table 65: Clusters seen in the BE-META of c(NSKSED).

Analysis of the Loop Database found the percentage occurrence of a dipeptide occupying the i+1 and i+2 positions of a β-turn compared to all occurrences of that dipeptide. This was used as an estimation for the propensity for each of the sequences to form a β-turn with sequences with a higher percentage occurrence in a β-turn assumed to have a higher propensity for forming a β-turn. The ED and SK sequences (part of the SEDN/NSKS register) give values of 36.3 and 16.6 % respectively (Table 66). The SE and NS sequences (part of the KSED/DNSK register) give values of 17.9 and 8.5 %. It may therefore be that the SEDN/NSKS I+I conformation is seen in the BE-META of c(NSKSED) and the KSED/DNSK is not due to the SEDN/NSKS register being better able to form β-turns. The EDNS/SKSE register has estimated turn propensity values of 18.7% for DN and 16.8% for KS. This is somewhere between the values for the two other registers as although not as high as seen for the ED subsequence in the SEDN/NSKS register the values are higher than the 8.5% seen for the

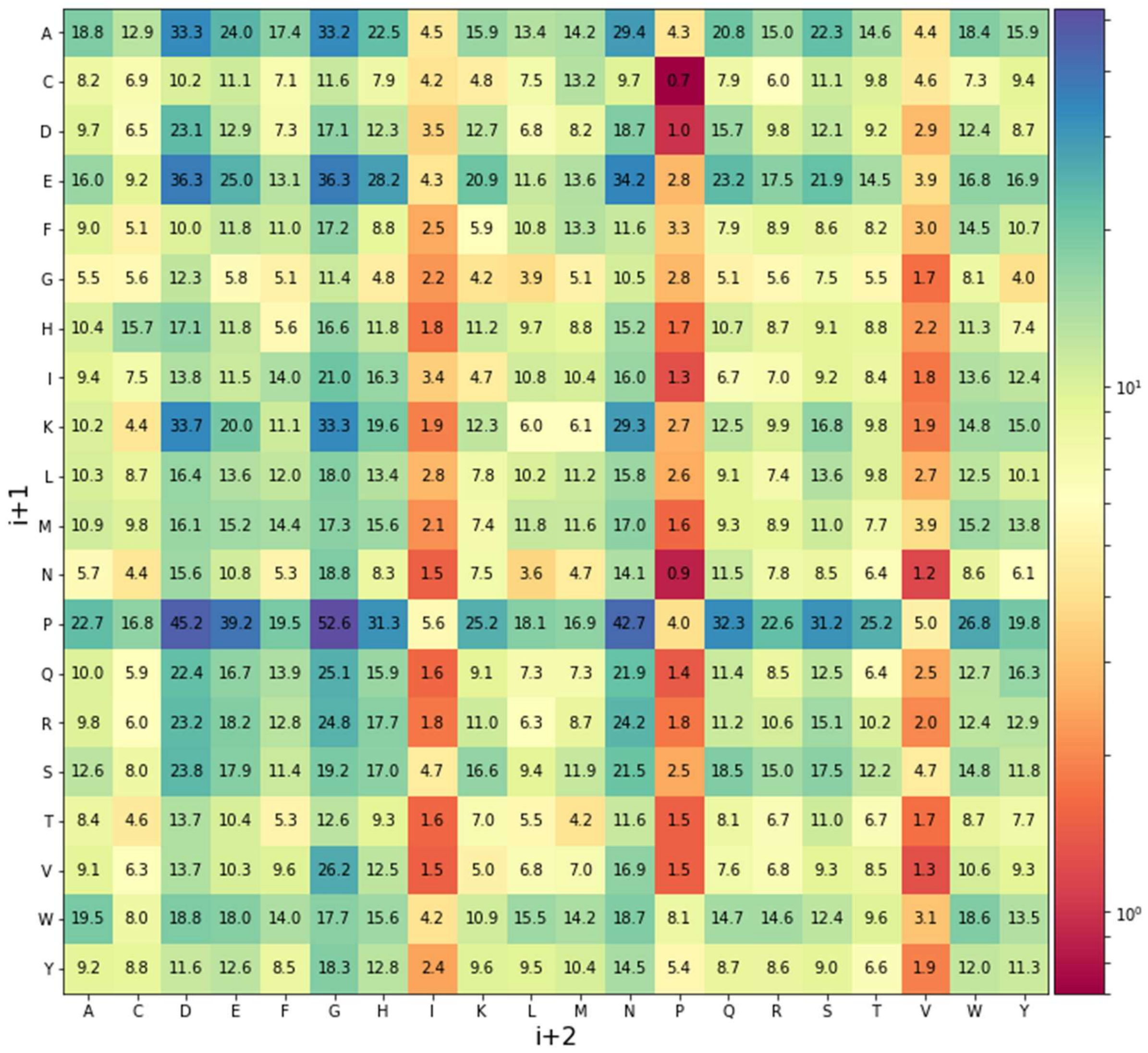
NS sequence which forms part of the KSED/DNSK register. This may be why the EDNS/SKSE register forms a minor cluster.

Sequence	Turn Propensity (%)	Sequence	Turn Propensity (%)
EDNS	18.7	SKSE	16.8
KSED	17.9	DNSK	8.5
SEDN	36.3	NSKS	16.6

Table 66: Estimated turn propensity for the registers of c(NSKSED) based on the percentage occurrence a dipeptide sequence is found at the i+1 and i+2 positions of a β-turn in the Loop Database.

5.4.6 Turn Propensity

Using the RF score in combination with the restrained simulations does not always lead to the predicted conformations. The amino acids usually remain in the same register if multiple conformations are seen. This suggests some amino acids more favourably occupy certain β-turn positions than others so some sequences will be more likely to form β-turns than others. A way of narrowing down the register likely to occur by looking at the likelihood a sequence will form a turn could potentially lead to improved predictions. The percentage occurrence of each dipeptide made up of the 20 naturally occurring amino acids in the i+1/i+2 positions of a β-turn compared to how often the dipeptide was observed elsewhere in the Loop Database was used as an estimate of turn propensity of each sequence (Figure 98).



	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
A	18.8	12.9	33.3	24.0	17.4	33.2	22.5	4.5	15.9	13.4	14.2	29.4	4.3	20.8	15.0	22.3	14.6	4.4	18.4	15.9
C	8.2	6.9	10.2	11.1	7.1	11.6	7.9	4.2	4.8	7.5	13.2	9.7	0.7	7.9	6.0	11.1	9.8	4.6	7.3	9.4
D	9.7	6.5	23.1	12.9	7.3	17.1	12.3	3.5	12.7	6.8	8.2	18.7	1.0	15.7	9.8	12.1	9.2	2.9	12.4	8.7
E	16.0	9.2	36.3	25.0	13.1	36.3	28.2	4.3	20.9	11.6	13.6	34.2	2.8	23.2	17.5	21.9	14.5	3.9	16.8	16.9
F	9.0	5.1	10.0	11.8	11.0	17.2	8.8	2.5	5.9	10.8	13.3	11.6	3.3	7.9	8.9	8.6	8.2	3.0	14.5	10.7
G	5.5	5.6	12.3	5.8	5.1	11.4	4.8	2.2	4.2	3.9	5.1	10.5	2.8	5.1	5.6	7.5	5.5	1.7	8.1	4.0
H	10.4	15.7	17.1	11.8	5.6	16.6	11.8	1.8	11.2	9.7	8.8	15.2	1.7	10.7	8.7	9.1	8.8	2.2	11.3	7.4
I	9.4	7.5	13.8	11.5	14.0	21.0	16.3	3.4	4.7	10.8	10.4	16.0	1.3	6.7	7.0	9.2	8.4	1.8	13.6	12.4
K	10.2	4.4	33.7	20.0	11.1	33.3	19.6	1.9	12.3	6.0	6.1	29.3	2.7	12.5	9.9	16.8	9.8	1.9	14.8	15.0
L	10.3	8.7	16.4	13.6	12.0	18.0	13.4	2.8	7.8	10.2	11.2	15.8	2.6	9.1	7.4	13.6	9.8	2.7	12.5	10.1
M	10.9	9.8	16.1	15.2	14.4	17.3	15.6	2.1	7.4	11.8	11.6	17.0	1.6	9.3	8.9	11.0	7.7	3.9	15.2	13.8
N	5.7	4.4	15.6	10.8	5.3	18.8	8.3	1.5	7.5	3.6	4.7	14.1	0.9	11.5	7.8	8.5	6.4	1.2	8.6	6.1
P	22.7	16.8	45.2	39.2	19.5	52.6	31.3	5.6	25.2	18.1	16.9	42.7	4.0	32.3	22.6	31.2	25.2	5.0	26.8	19.8
Q	10.0	5.9	22.4	16.7	13.9	25.1	15.9	1.6	9.1	7.3	7.3	21.9	1.4	11.4	8.5	12.5	6.4	2.5	12.7	16.3
R	9.8	6.0	23.2	18.2	12.8	24.8	17.7	1.8	11.0	6.3	8.7	24.2	1.8	11.2	10.6	15.1	10.2	2.0	12.4	12.9
S	12.6	8.0	23.8	17.9	11.4	19.2	17.0	4.7	16.6	9.4	11.9	21.5	2.5	18.5	15.0	17.5	12.2	4.7	14.8	11.8
T	8.4	4.6	13.7	10.4	5.3	12.6	9.3	1.6	7.0	5.5	4.2	11.6	1.5	8.1	6.7	11.0	6.7	1.7	8.7	7.7
V	9.1	6.3	13.7	10.3	9.6	26.2	12.5	1.5	5.0	6.8	7.0	16.9	1.5	7.6	6.8	9.3	8.5	1.3	10.6	9.3
W	19.5	8.0	18.8	18.0	14.0	17.7	15.6	4.2	10.9	15.5	14.2	18.7	8.1	14.7	14.6	12.4	9.6	3.1	18.6	13.5
Y	9.2	8.8	11.6	12.6	8.5	18.3	12.8	2.4	9.6	9.5	10.4	14.5	5.4	8.7	8.6	9.0	6.6	1.9	12.0	11.3

Figure 98: Percentage occurrence of dipeptide sequences at the i+1 and i+2 positions β-turns in the database.

It should be noted however that using the amount a sequence forms a turn in the Loop Database may not be best measure for turn propensity in a cyclic peptide environment. For example the sequences with glycine at the i+2 position generally have higher estimated turn propensities. The restrained simulations show that in a cyclic hexapeptide environment more stable conformations can be obtained with having glycine at one of the i/i+3 positions. This contextual difference means both the estimated turn propensity and information from the restrained simulations needs to be taken into account. As glycine is particularly flexible it is able to occupy any position within the cyclic peptide environment not just the turn region. Although commonly found in turns in the Loop Database, in a cyclic peptide where the i/i+3 positions are relatively restrained positions then it sometimes may be more favourable for glycine to occupy a i/i+3 position rather than a i+1 or i+2 position. For example based on turn propensity the sequence c(RGSQGW) simulated above would be predicted to adopt the WRGS/SQGW register (Table 67). This register is not seen however but the GWRG/GSQG register is. This is why glycine (and aspartic acid and asparagine) can make cyclic peptide structure prediction particularly difficult.

Sequence	Turn Propensity (%)	Sequence	Turn Propensity (%)
GWRG	14.6	GSQG	18.5
WRGS	24.8	SQGW	25.1
RGSQ	7.5	QGWR	8.1

Table 67: Estimated turn propensity for c(RGSQGW).

5.4.6.1 c(DSWDKA)

By looking for a sequence with the same turn type and same score for all the registers, but with one register with a much greater turn propensity than the others, then only the register with the greatest turn propensity should be seen. As type I turns are the most common and most frequently give a RF score of 1 a sequence was chosen whereby all registers were made up of β-turns predicted to form a type I turn with a score of 1. Approximately 25% of the dipeptide sequences occur at the i+1 and i+2 positions of β-turns more than 15% of time. The potential sequences were therefore narrowed down by looking for a sequence which has a > 15 % occurrence at the i+1 and i+2 positions in a β-turn in the database for one register, with the remaining registers having sequences with lower turn percentages and thus a lower turn propensity. The sequence c(DSWDKA) was selected (Table 68).

Sequence	Type	Score	Turn Propensity	Sequence	Type	Score	Turn Propensity
KADS	I	1	33.3	SWDK	I	1	18.8
ADSW	I	1	12.1	WDKA	I	1	12.7
DSWD	I	1	14.8	DKAD	I	1	10.2

Table 68: RF prediction and scores for c(DSWDKA) and turn propensity for each register.

BE-META on c(DSWDKA) shows two clusters (Table 69). The major conformation is the one predicted based on the RF scores and turn propensities with a I+I conformation and the KADS/SWDK register. The remaining smaller cluster (8%) has the same amino acid register but a II'+IV$_3$ conformation. The other two registers with the lower turn propensities are not seen despite the same RF scores and predictions so including turn propensity may help predict the conformation seen.

Cluster	Population (%)	Turn-type Combination
1	92	I+I
2	8	II'+IV$_3$

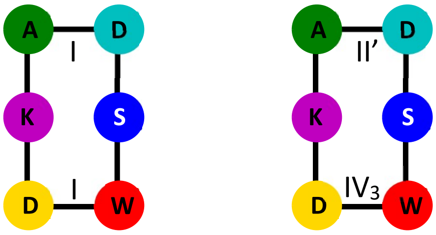


Table 69: Clusters seen in the BE-META of c(DSWDKS).

Although based on a limited number of examples these initial results suggest using the RF scoring system in combination with the restrained simulations may help predict the conformation of cyclic peptides. Turn types that were not used to train the RF can appear however that cannot be predicted. As cyclic peptides often form multiple conformations in solution, using the RF scores in combination with the restrained simulations can potentially help find sequences which form one dominant conformation in solution. However the RF does not predict how likely it is for a sequence to form a β-turn meaning prediction of which register will form can still be difficult. A cut-off of 15%

of a dipeptide sequence occurring at the i+1 and i+2 positions in a β-turn was used to select a sequence with one register having a higher turn propensity than the others. The register predicted to have the highest turn propensity was the only one seen in the BE-META of c(DSWDKS) so incorporating this extra parameter into structure prediction may help improve the results.

5.5 Predicting the Conformation of a RGD-containing Cyclic Peptide

Previous sequences were selected based on their predicted ability to form a single stable conformation. To see if the same selection principles could be applied to design a cyclic peptide incorporating a biologically active motif the RF was used to design a peptide containing the integrin binding RGD motif. A linear RGD motif is needed to bind to αIIbβ3 integrins. The RF was therefore used for designing a cyclic peptide with a conformation where the RGD motif is in the correct linear orientation for binding. If the peptide doesn’t form a conformation with the correct register then the RGD will be in the wrong orientation for binding (Figure 99).

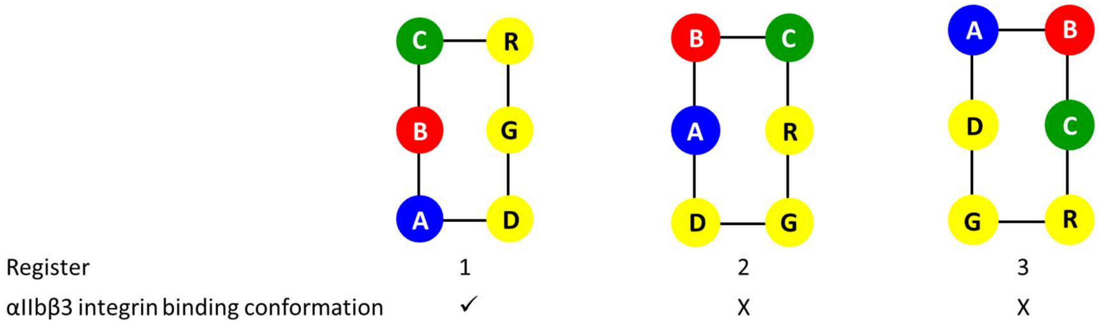


Figure 99: Only one cyclic peptide register would allow for binding of the RGD motif to αIIbβ3 integrins.

The X-ray crystal structure of a small peptide containing the RGD motif bound to an αIIbβ3 integrin (PDB: 2VDR) was used to obtain the dihedral angles necessary for the peptide to bind (Figure 100).[365] The RGD would have to occupy the position down one side of the cyclic peptide to have the elongated conformation, with the Arg at the i+2 position of one turn and aspartic acid at the i+1 position of another. As the arginine and aspartic acid would be at the i+2 and i+1 positions of β-turns respectively the dihedral angles they occupy were used to determine the desired turn-type combination for the cyclic peptide. The ϕ and ψ dihedral angles in the bound conformation are -85.7 and -3.9° for arginine and -85.4 and 76.3 for aspartic acid. To fit in with the dihedral angles seen in the crystal structure this would require the arginine to be part of a type I or a type II’ turn and the aspartic acid to be part of a type II turn (Figure 101). The dihedral angles for the glycine residue are also compatible with occupying the i/i+3 position within a cyclic hexapeptide.

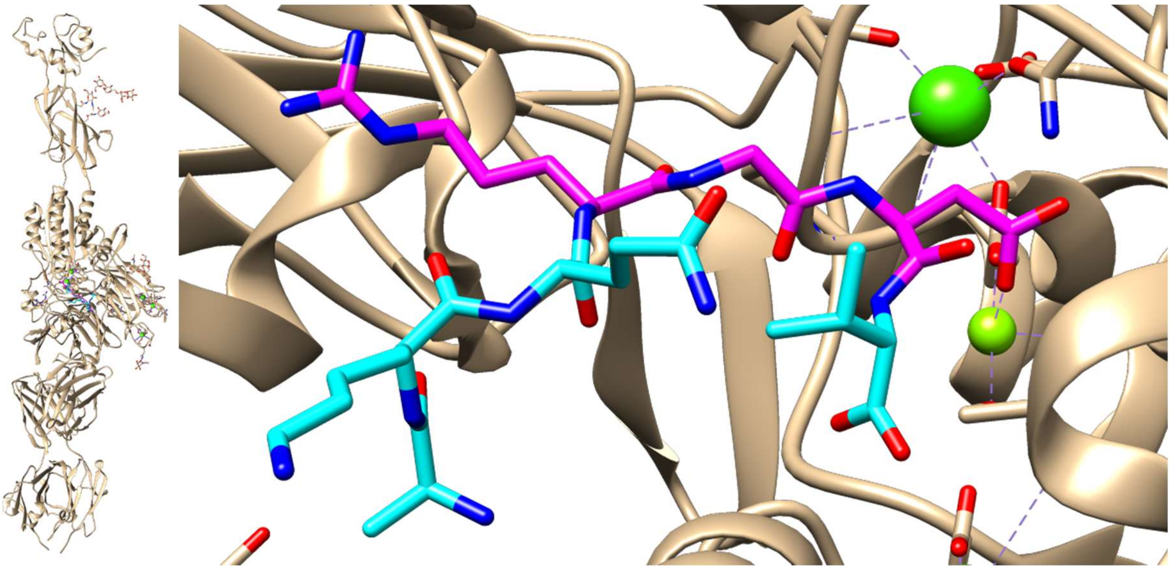
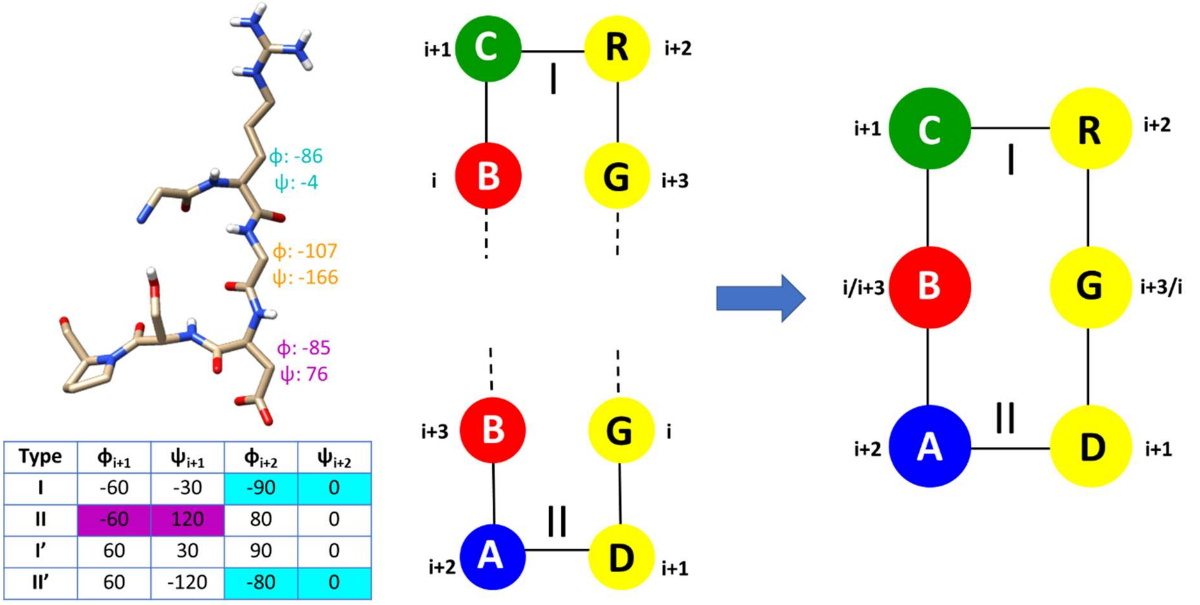


Figure 100: X-ray crystal structure of a RGD-containing ligand bound to αIIbβ3 integrin (PDB 2VDR). The ligand is shown in light blue with the RGD portion in magenta.

The restrained simulations where a glycine occupies one of the i/i+3 positions suggest that the II+II’ conformation would be lower energy than the I+II conformation. Very few sequences are predicted to form a type II’ turn however. The RF predictions therefore did not find any sequences with arginine at the i+2 position and glycine at the i+3 position predicted to form a type II’ turn with a high RF score. There are many more sequences predicted to form a type I turn than a type II’ turn so a cyclic hexapeptide with a type I+II conformation rather than a II’+II combination was designed.



Type	ϕ_{i+1}	ψ_{i+1}	ϕ_{i+2}	ψ_{i+2}
I	-60	-30	-90	0
II	-60	120	80	0
I’	60	30	90	0
II’	60	-120	-80	0

Figure 101: The RGD motif could be incorporated into a cyclic peptide with a I+II conformation.

The RF was used to search for tetrapeptide sequences which are predicted to form a type I turn with RG at the i+2 and i+3 positions. 358 sequences were found. Searching for sequences with GD at the i and i+1 positions of a type II turn resulted in 28 sequences. The i+3 of the type II turn must be the same amino acid as the amino acids at the i position of the type I turn in order for the two turns to overlap. Compatible sequences were filtered based on the RF score for the desired I+II conformation being high, and the score for the other possible amino acid registers in the peptide being low.

c(TPRGDG) gave a score of 1 for TPRG forming a type I turn and a score of 0.97 for GDGT forming a type II turn (Table 70). The scores for the other registers were all much lower. The TPRG/GDGT register also has the highest turn propensities with the exception of PRGD which would be very unlikely to occur as proline would be required to occupy the i/i+3 positions of the cyclic peptide. A method of determining turn propensity which includes the contribution of the i/i+3 positions of the β-turn could therefore lead to better predictions.

Sequence	Type	Score	Turn Propensity	Sequence	Type	Score	Turn Propensity
TPRG	I	1	23	GDGT	II	0.97	17
PRGD	II	0.09	25	DGTP	II’	0.04	5
RGDG	I	0.11	12	GTPR	VI	0.22	2

Table 70: RF scores for c(TPRGDG).

BE-META was carried out on the sequence to determine its conformation to see if it matched the RF prediction. Two clusters were seen in the BE-META (Table 72). One was the predicted I+II combination, the other retained the same register and the type I turn for the TPRG sequence, but the GDGT sequence formed a type I’ rather than a type II turn. The I+I’ combination occurred more frequently (87%).

	Prediction				
Sequence	I	II	I’	II’	VI
TPRG	1	0	0	0	0
GDGT	0.01	0.98	0.01	0	0
PRGD	0.44	0.54	0.02	0	0
DGTP	0.48	0	0	0.52	0
RGDG	0.55	0	0	0.45	0
GTPR	0.39	0	0	0	0.61

Table 71: The proportion of the decision trees which make up the RF which assigned each sequence to a given class for c(TPRGDG).

The GDGT sequence based on the RF was predicted to only have a 1% chance of being a type I’ turn (Table 71). The II+I combination is less stable than the I’+I conformation based on the c(GGGXGG) restrained simulations (Figure 102) which may be why the GDGT sequence forms a type I’ rather than a type II turn. The β-turn environment may be important for accurate prediction. Type I and II turns are in general much more common than I’ and II’ turns, with the exception of in β-hairpins. In β-hairpins this trend is reversed and type I’ and II’ turns are more common. The RF is trained on all turn types so type I and II turns are more commonly predicted. It may be however that cyclic hexapeptides are an environment more similar to β-hairpins than β-turn in general so type I’ and II’ β-turns may occur more frequently than predicted by the forest.

Cluster	Population (%)	Turn-type Combination
1	87	I + I’
2	13	I + II

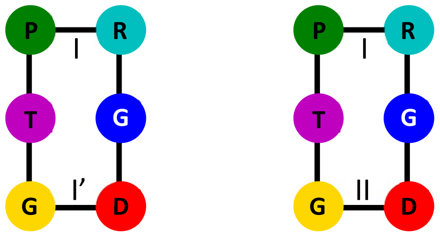


Table 72: Two clusters are seen in the BE-META of c(TPRGDG).

The c(TPRGDG) sequence has two glycines in it which potentially makes it more difficult to predict the conformation as glycine is very flexible. The majority of sequences which were able to form multiple turn types contain glycine, asparagine or aspartic acid at the i+1 and/or i+2 positions. As sequences which form multiple β-turn types were removed from the training set in order to improve the predictive ability of the RF it is likely this improved accuracy actually came at the expense of predicting the flexibility of certain turn types.

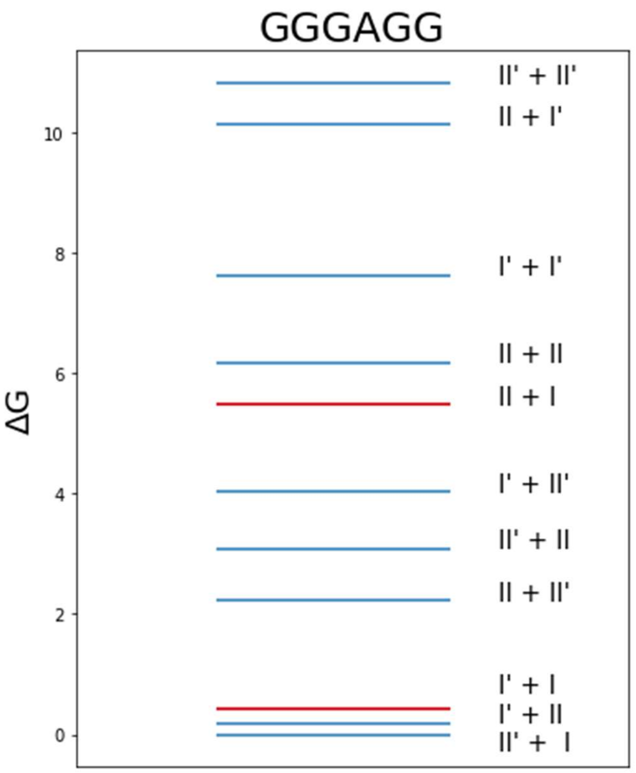


Figure 102: The I’+I conformation is a lower energy conformation than the II+I conformation in the c(GGGXGG) restrained simulations.

If the peptide is being designed to bind something, then having the amino acids in the right register but having some flexibility in the conformations seen could potentially still lead to binding. The backbone atom RMSD between the RGD motif of the I+II and I+I’ conformations of c(TPRGDG) with the conformation of the LGGAKQRGDV ligand bound to αIIbβ3 integrin is 0.22 and 0.68 Å respectively (Figure 103).

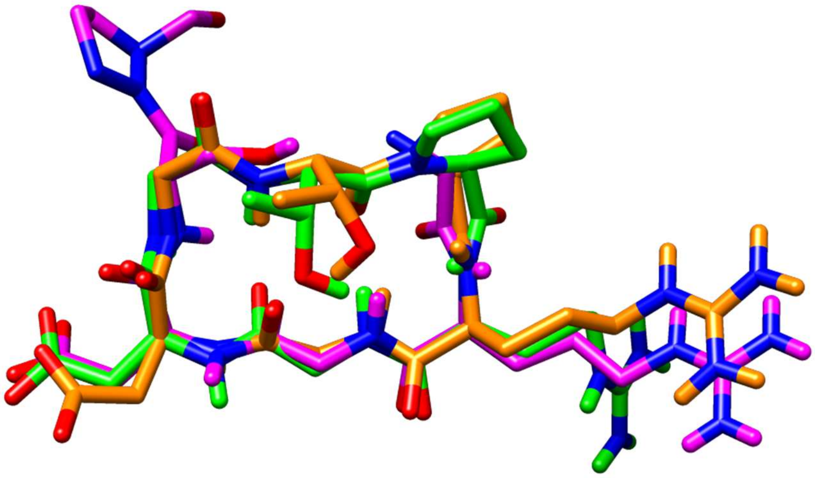


Figure 103: The LGGAKQRGDV ligand (magenta) from PDB: 2VDR overlayed with the I+II (green) and I+I' (orange) conformations of c(TPRGDG).

5.6 Limitations of the Random Forest

The RF was trained on β-turns extracted from the Loop Database. Differences in the β-turns used to train the RF and the strained cyclic peptide environment can lead to differences in the turn types seen. β-turns in the centre of a protein will be in a very different environment from a β-turn in a cyclic peptide. Some sequences have a very small energy difference between the different β-turn types and were observed to form multiple β-turn types in the Loop Database, which can make accurate prediction difficult. This may be why sequences with glycine, aspartic acid and asparagine are particularly poorly predicted by the RF. They allow for flexible turns rather than a preference for a single turn type. Although removing sequences which formed multiple turn types from the training set of the RF appeared to allow for improved prediction, it likely made prediction of these sequences worse as information about their flexibility is lost by removing them. Additionally as they are flexible they may form multiple turn types in solution, but the sequences the RF is trained on are from X-ray crystal structures which may not accurately reflect the solution behaviour of the turns.

Figure 104 shows the rate of errors in the RF predictions based on the test set. The squares represent sequences belonging to each turn type category and what they were incorrectly assigned as, with darker squares representing a higher proportion of incorrectly assigned sequences. The first column is relatively dark as many sequences which actually form other turn types are predicted to form a type I turn. Sequences which form II' turns are most commonly incorrectly assigned as type I turns. Sequences that form type I β-turns are usually correctly assigned however, with only a few examples of type I turns being misassigned as type II turns. The dataset used to train the RF contained predominantly type I turns, with very few type II' turns. The RF can therefore obtain a high accuracy by assigning most turns as type I. This is likely why there is a higher error for other turn types, especially the small type II' category, being assigned as type I. There may also be some sequence overlap between the two categories making prediction more difficult.

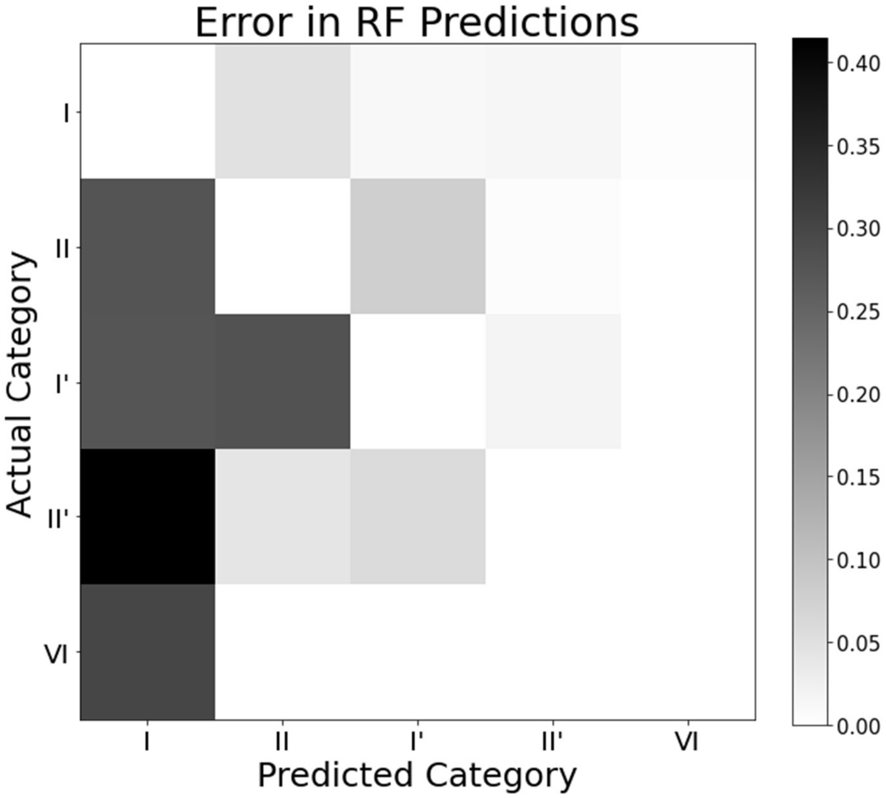


Figure 104: Sequences which form other turn types are often incorrectly assigned as type I turns. The scale shows the fraction of misassigned β*-turns.*

The main error in the RF predictions ultimately comes from the class imbalance in the training set where most β-turns were type I turns. One way to reduce the error in the RF predictions would be to add additional examples of data that belong to the minority categories to the training set allowing the RF to learn how to better distinguish the different turn types rather than over assigning the type I category. There are limits on the available data based on the number of structures in the PDB making obtaining such data difficult. Resampling to make the category sizes more even did not allow for better prediction but there are many other types of resampling that could potentially have an effect. Some resampling techniques, such as the synthetic minority oversampling technique (SMOTE),[366] aim to deal with class imbalance by creating new data points which belong to the smaller categories. The new data is generated by sampling attributes from the minority classes to produce further examples which are likely to be part of that class. This can result in ambiguous examples if there is overlap between classes, which is likely in this case given that many sequences are seen to form multiple β-turn types.

As it is possible for sequences to form multiple turn types, new features to better distinguish which turn type a sequence forms rather than just the tetrapeptide sequence would also likely lead to improved predictions. As the turns need to be applicable to the cyclic peptide environment this limits any additional data that can be added. For example the secondary structural environment the β-turn is found within in the Loop Database will not be the same as in the cyclic peptide environment. However some information may be relevant such as how solvent exposed the β-turn is. If a sequence is more likely to form a given turn type when found at the surface of the protein than in a hydrophobic core it may be more likely to form that turn type in the cyclic peptide which as a small peptide interacts with the solvent environment.

5.7 Conclusions

A Random Forest was trained to predict the type of β-turn a given tetrapeptide would form if it were to form a β-turn. Although not comparable to other β-turn type prediction methods which are designed to predict if an individual amino acid would be part of a β-turn and if so what type in the larger context of a protein structure, the RF performs well with a MCC value of 0.66 despite the only input being the tetrapeptide sequence with no additional information such as secondary structure prediction included.

As cyclic hexapeptides often form a structure in solution which can be thought of as two overlapping β-turns, the RF was tested to see if it could be used to help predict the conformation of cyclic peptides by determining the β-turn types likely to form from a given sequence. The RF scoring system was devised whereby a sequence with a higher prediction for a given turn type was assumed to more favourably be able to form that turn type. It was found that the RF scores should be used in combination with low energy turn combinations based on the restrained simulations to allow for improved prediction of which cyclic peptide conformation would occur. BE-META takes a significant amount of computational time to run so using the RF to find sequences most likely to form the desired conformations could potentially help with the design of cyclic peptides. These initial results show the RF is capable of predicting the β-turn types sequences are likely to form but further test sequences are needed. There were some common features that appeared in many of the sequences tested so far which are important when using the RF to design cyclic peptides:

- High scoring sequences are more likely to form the predicted turn type than lower scoring sequences which can potentially be more flexible.
- The RF should be used in combination with the restrained simulations. High scoring sequences predicted to form lower energy turn type combinations give more stable structures.
- The RF does not discern between which sequences are better able to form a turn so a measure of turn propensity can aid prediction. Inclusion of a proline can help as it often occupies the i+1 position of a β-turn and rules out the register where proline would be at the i/i+3 positions.
- The RF is only trained to predict turn types I, II, I', II' and VI. It therefore cannot accurately predict the conformation of sequences that ultimately form other turn types. The type IV$_3$ turn type isn't predicted by the RF but is overrepresented by the cyclic peptide restrained simulations so can lead to conformations that aren't predicted by RF scores.
- The RF is trained on many β-turns but all in a different context from a cyclic peptide. The local environment in a cyclic peptide may lead to different turn-types than predicted.
- Most sequences are predicted to form type I turns so the RF may overpredict the occurrence of type I turns. Since most sequences form type I turns it can be useful to include a turn compatible with type I turns when designing cyclic peptides
- Glycine plays an important role in β-turn type prediction. There is often a glycine in the sequence if turn types other than type I are searched for. Glycine can make the prediction of the conformation more difficult. Especially turns composed of glycine, aspartic acid and asparagine are particularly flexible and prone to forming multiple turn types. This flexibility may not be well captured by the RF particularly because all sequences which formed multiple turn types were removed from the final training dataset. Sequences containing adjacent G/N/D residues should probably be avoided when designing conformationally stable cyclic peptides.

RF score, turn type combination and turn propensity were all shown to contribute to cyclic peptide structure prediction. For the test sequences for simplicity they were treated as categorical variables with either a high/low RF score etc, but as there are small energy differences between conformations a way of processing them as quantitative data would likely lead to improved predictions. If sufficient data on the structure of cyclic peptides were available it may be possible to use a machine learning algorithm such as linear regression or a neural network to process these three pieces of data that seem to contribute to cyclic peptide structure. There may also be other factors which could influence the structure however.

The RF was used to find potential sequences which would form a I+II conformation with the biologically active RGD motif in a linear conformation. This conformation, based on an available X-ray crystal structure of a ligand bound to αIIbβ3 integrin, demonstrates how the RF could be used to design cyclic peptides with a specific function. Using the 20 naturally occurring amino acids to occupy the three remaining positions in the cyclic hexapeptide there are 8,000 possible combinations. The RF was able to identify a sequence in seconds likely to form the correct β-turn types. The BE-META of the c(TPRGDG) sequence shows the peptide is likely to adopt a conformation with the RGD motif in the desired register and the I+II conformation was seen as a cluster. The RF therefore offers a rapid way of screening potential sequences prior to simulating or making them.

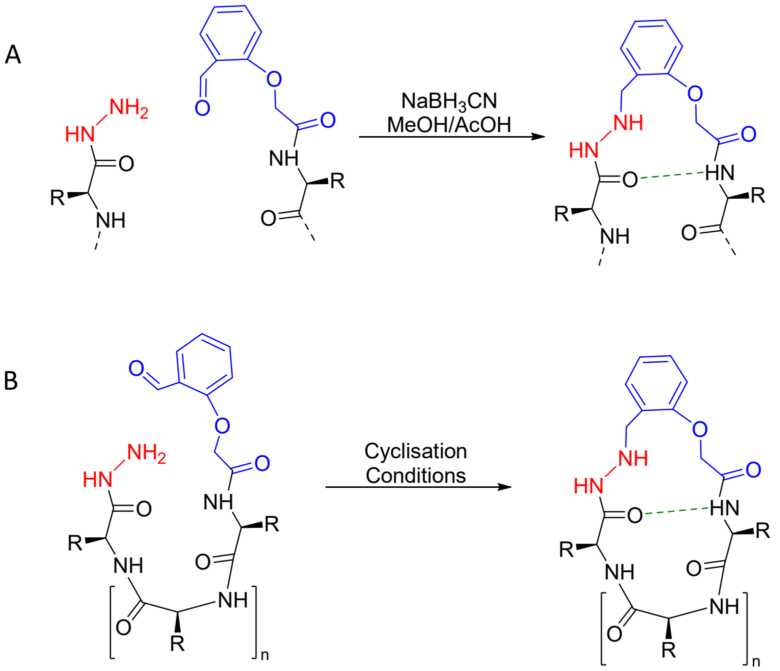
Although applied to cyclic hexapeptides, the RF could potentially be adapted for use on other macrocycle sizes. Other reverse turn structures could be extracted from the Loop Database such as γ-turns. Whereas a cyclic hexapeptide frequently makes up a structure in solution that can be thought of as two overlapping β-turns, cyclic pentapeptides frequently form structures composed of overlapping β- and γ-turns. By training a RF to predict the γ-turn a sequence is likely to form it could therefore be possible to predict the conformation of cyclic pentapeptides.

6 Incorporation of a β-turn Mimic within a Cyclic Peptide

Small cyclic peptides often form multiple conformations in solution.[202, 313, 316, 367] Currently when designing cyclic peptides with reduced conformational freedom many variations of the same cyclic peptide are synthesised and screened as it can be difficult to predict the effect of a particular modification, such as *N*-methylation, on the conformation. Small cyclic peptides frequently form structures in solution containing β-turns. For example cyclic pentapeptides often form a structure composed of overlapping β and γ-turns and cyclic hexapeptides two overlapping β-turns.[46] One of the most commonly used turn structures is D-Pro-L-Pro which forms a type II' turn and has previously been effective in forming cyclic peptides with defined β-hairpin structures designed to mimic biologically active structures.[11, 164-166] Chapter 4 (Restrained simulations) shows how the incorporation of a type II' turn could potentially alter the β-turn type formed at the other side of a cyclic hexapeptide.

Cyclisation is often used as a strategy to reduce the conformational freedom of peptides and induce conformation. However cyclisation alone is sometimes insufficient to induce the desired structure. Therefore cyclisation methods that also induce secondary structure can provide another means of designing peptides with a desired conformation. If a ligation method is used to cyclise the peptide whilst also introducing a structural element that can help control the conformation of the cyclic peptide this would offer an advantage. As cyclic peptides often form multiple conformations in solution structural elements such as β-turn mimics can potentially reduce the conformational flexibility of a peptide and induce structure.

Previously in the group a chemical ligation method was developed and used to join two peptide fragments together whilst incorporating a β-turn mimic (BTM) in the process.[191] The chemical ligation used commercially available materials and mild conditions to incorporate the BTM within the peptide. Inclusion of a BTM could induce structure within a cyclic peptide thereby helping restrict the number of conformations. It could also potentially aid synthesis if the chemical ligation that forms the BTM is used to cyclise the peptide. Methods were therefore developed to incorporate the BTM in cyclic peptides.

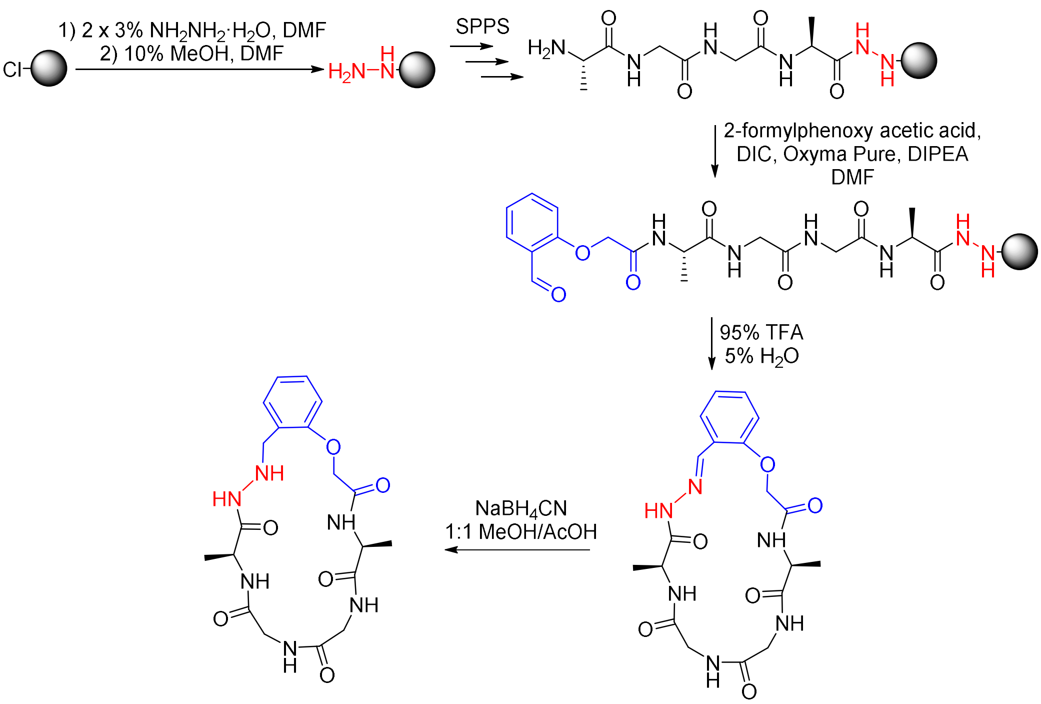


Scheme 17: Incorporation of a BTM into a peptide through chemical ligation original (A) and cyclic (B).

6.1 Incorporation of a β-turn Mimic into a Cyclic Peptide

The BTM unit is formed through the chemical ligation of two peptide fragments by the reaction of an acylhydrazine group with an aldehyde. To synthesise the BTM containing cyclic peptides hydrazine hydrate was reacted with 2-Cl-trityl resin to form a hydrazine resin. SPPS was then used to couple the amino acids. The aldehyde group was introduced by coupling 2-formylphenoxyacetic acid to the peptide as the final residue. Following cleavage of the peptide from the resin the aldehyde group at the *N*-terminus and the hydrazine group at the C-terminus can react to cyclise the peptide through formation of an acylhydrazone. The acylhydrazone can then be reduced to form the BTM using sodium cyanoborohydride (Scheme 18).

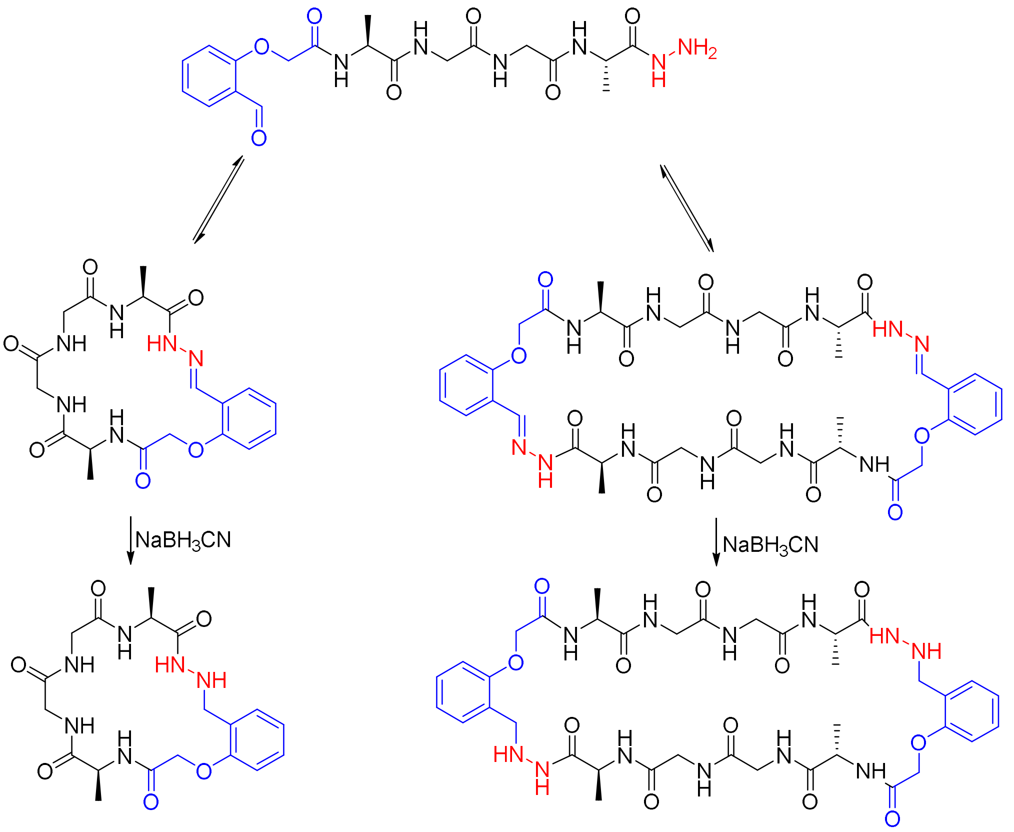
The short peptide sequence AGGA was initially used as a model for the incorporation of the BTM into a cyclic peptide. It was found that following cleavage of the peptide from the resin the linear peptide was not observed, instead the cyclic peptide formed directly. Spontaneous cyclisation of the AGGA peptide is therefore favourable.



Scheme 18: Synthesis of c(BTM-AGGA).

6.1.1 The Formation of Dimers

In addition to c(BTM-AGGA), the joining of two peptides to form a cyclic "dimer" of c(BTM-AGGA)$_2$ was observed following reduction of the acylhydrazone with sodium cyanoborohydride (Scheme 19). A ratio of 1:0.77 of the monomer to dimer was detected by HPLC integration. Hydrazone formation is reversible in acidic conditions such as the 1:1 methanol/acetic acid used for the sodium cyanoborohydride reduction step.



Scheme 19: Formation of c(BTM-AGGA) and c(BTM-AGGA)$_2$.

Following cleavage of the peptide from the resin (prior to the sodium cyanoborohydride step) LCMS analysis showed there is no dimer present before the peptide was freeze-dried. After freeze-drying dimer was observed. In order to freeze-dry the peptide it was dissolved in 1:1 water/acetonitrile and frozen in liquid nitrogen. The Gibbs free energy between states can be calculated by $\Delta G = \Delta H - T\Delta S$. The $T\Delta S$ term means that temperature is closely linked with the difference in entropy between the monomer and dimer states so, depending on the magnitude of the ΔS term between the two states, the temperature can greatly affect the free energy difference between states. The dimer may therefore be more prone to form at lower temperatures while the solution the peptide is in is being frozen.

Several solvent systems were tested to determine if the solvent affects the ratio of monomer to dimer by altering the equilibrium between the two states. Following cleavage of the peptide from the resin, in addition to the 1:1 methanol/acetic acid previously used, a sample was dissolved in 1:1 water/methanol, 1:1 water/acetonitrile or phosphate buffered saline (PBS, pH 7.4) at a concentration of 0.15 M. The solutions were left to equilibrate for a week at room temperature prior to addition of sodium cyanoborohydride. Due to residual TFA left over from the cleave the water/methanol and water/acetonitrile solutions had a pH of approximately 5. Sodium cyanoborohydride was able to reduce the hydrazone to the hydrazine under all solvent systems tested. A longer reaction time was required for the less acidic solutions such as PBS where complete reduction of the acylhydrazone was seen after two days rather than the 15 minutes required for the methanol/acetic acid solution. The same ratio of monomer to dimer was seen for each solvent system. Solvent therefore does not significantly alter the monomer to dimer ratio at room temperature for this system. The equilibrium between monomer and dimer states relies on the reversibility of hydrazone formation. It may be that minimal hydrolysis of the hydrazone is seen in the solvent systems tested.

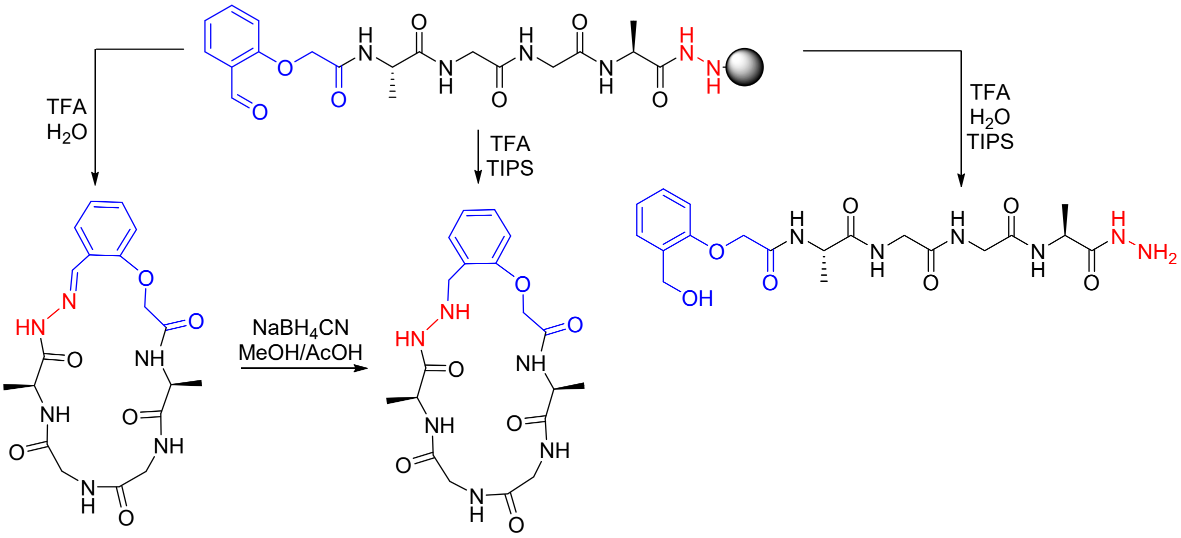
If the lower temperatures used during freeze-drying the peptide allow for dimer formation then no dimer should be seen if the peptide is dissolved and reduced with sodium cyanoborohydride without freeze drying in between the two steps. The synthesis of c(BTM-AGGA) was therefore repeated without freeze-drying the peptide prior to reduction of the acylhydrazone. All other conditions were kept the same. The dimer was no longer observed. To further test the effects of temperature on the formation of the dimer, following cleavage from the resin and without freeze drying, 0.15 M samples of the peptide were dissolved in 2:1 methanol/water. One sample was held at -20 °C and the other was kept at room temperature for one week. Sodium cyanoborohydride was then added to reduce the acylhydrazone. Dimer formation was observed in the sample kept at -20 °C (1:0.05 monomer to dimer) but not in the one kept at room temperature.

It is likely an increased proportion of dimer may be seen with a higher concentration of peptide and lower temperatures. Solvent effects could also potentially impact the monomer to dimer ratio as peptide folding is different in different solvents. Dimer formation was avoided by omitting freeze-drying the peptide prior to reduction using sodium cyanoborohydride.

6.1.2 Cyclisation Conditions

During the initial development of the BTM triisopropylsilane (TIPS) had been removed from the cleavage conditions as it was causing partial reduction of the aldehyde group. As the AGGA peptide was seen only in the cyclic form prior to reduction of the acylhydrazone, TIPS was introduced back into the cleavage mixture. The sodium cyanoborohydride step to reduce the acylhydrazone may be unnecessary if TIPS in the cleavage mixture could be used to reduce the acylhydrazone directly. The peptide would have to cyclise and the hydrazone be reduced by the TIPS faster than the rate the aldehyde would be reduced by the TIPS.

Standard cleavage conditions of 95% TFA, 2.5% water and 2.5% TIPS were initially tested. The cyclic peptide with the hydrazone reduced to the acylhydrazine was the major product, however the linear peptide with the aldehyde reduced to an alcohol, and the non-reduced cyclic peptide are also seen (ratio 1:0.47:0.62). No c(BTM-AGGA)$_2$ dimer was seen. It was found that removing water from the cleavage cocktail prevented the formation of the linear peptide with the aldehyde reduced. As including TIPS was found to yield the product directly in one step whilst cleaving the peptide from the resin further conditions with varying amounts of TIPS were tested.



Scheme 20: Cyclisation conditions for c(BTM-AGGA).

The cleavage/cyclisation conditions tested are shown in Table 73. Increasing the percentage of TIPS leads to an improvement in the percentage of product seen only up to a certain point. As TIPS and TFA are not fully miscible 10% DCM was added into a reaction mixture of 80% TFA and 10% TIPS but no increase in the reduction of the hydrazone was seen. Increasing the temperature had the greatest effect with complete reduction of the acylhydrazone to acylhydrazine after 30 minutes at 50 °C with 90% TFA and 10% TIPS.

Conditions	% TIPS in TFA	% DCM	Temperature (°C)	Time (h)	% hydrazone reduced to hydrazine in the cyclic peptide	Crude Yield %
A	5	0	rt	0.5	64	57
B	10	0	rt	0.5	74	69
C	20	0	rt	0.5	64	60
D	10	10	rt	0.5	64	57
E	10	0	rt	2	86	64
F	10	0	15	2	60	55
G	10	0	50	0.5	100	83

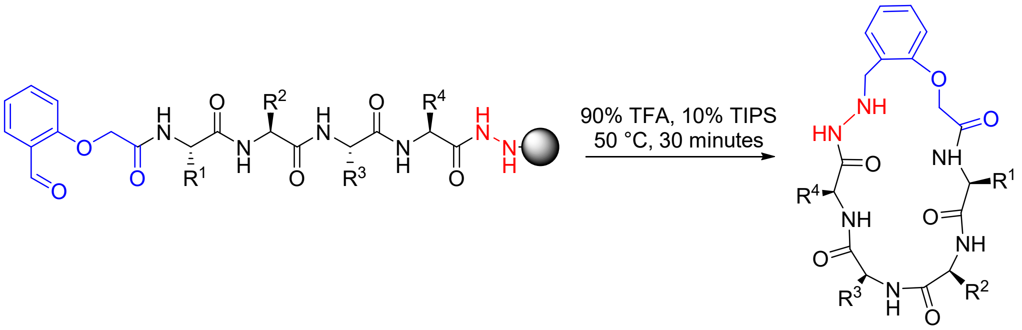
Table 73: Cleavage/cyclisation conditions.

Triethylsilane (TES) is more reactive than TIPS and has previously been used to reduce non-peptidic oximes or hydrazones in similarly acidic conditions. [93, 94] Triethylsilane was therefore tried as an alternative reducing agent to TIPS. With 95% TFA and 5% TES, 85% of the hydrazone was reduced to the hydrazine at room temperature. However, unlike with TIPS, the linear peptide with the aldehyde reduced was seen as a side product. This increase in side products meant a 56% crude yield was obtained.

90% TFA with 10% TIPS at 50 °C for 30 minutes allowed for complete reduction of the acylhydrazone and the highest crude yield so these conditions were therefore selected to synthesise BTM-containing cyclic peptides.

6.1.3 Sequence Tolerance

The sequence tolerance of the reductive cleave/cyclisation conditions was tested on a series of peptides (Scheme 21). Specifically the inclusion of different amino acids adjacent to the acylhydrazine group were examined. Amino acids with large or potentially nucleophilic sidechains were incorporated as they have potential to interfere with or compete with the hydrazine group by reacting with the aldehyde prior to formation of the BTM. Most amino acids with potentially reactive sidechains were tolerated at this position including serine, arginine and lysine, as were bulky β-branched amino acids such as valine (Table 74).



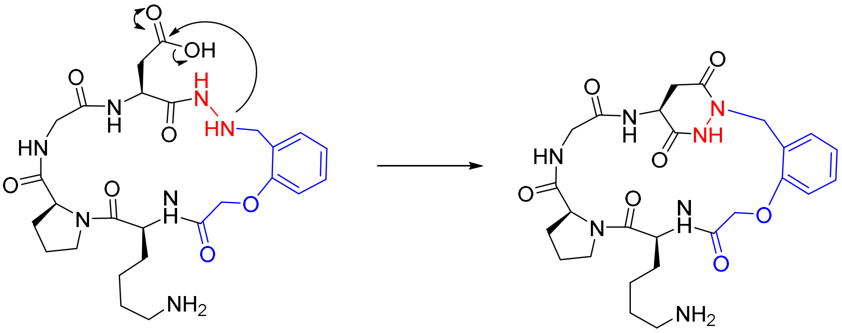
Scheme 21: Testing the sequence tolerance of the reductive cleave/cyclisation conditions. The influence of the amino acid next to the acylhydrazine (R^4) was examined.

It was found that asparagine in the position next to the hydrazine lead to decomposition of the product to an alternative peptide with a mass 17 Da lower. It is likely the nucleophilic nitrogen of the hydrazine functionality attacks the amide carbonyl of the asparagine sidechain leading to formation of a 6-membered ring and loss of ammonia (Scheme 22). This is a reaction analogous to aspartimide formation occasionally observed in peptide synthesis.[368-370] The decomposition occurred over time so is likely an inherent property of the peptide rather than due to the cleavage conditions used. The equivalent reaction is not seen when aspartic acid, glutamine or glutamic acid occupy this position.

Peptide Sequence	Crude Yield (%)
c(*BTM*-AGGA)	83
c(*BTM*-KPGN)	-
c(*BTM*-KPGD)	64
c(*BTM*-KPGQ)	81
c(*BTM*-KPGE)	77
c(*BTM*-QGPK)	77
c(*BTM*-KPGC)	7/63*
c(*BTM*-KPGV)	83
c(*BTM*-KPGR)	84

*Table 74: Crude yields observed for the different sequences tested. *cleavage at room temperature.*

A low yield of 7% was seen for c(BTM-KPGC). The thiol sidechain presumably participates in side reactions leading to the reduced yield. In order to try increase the yield the synthesis was repeated at room temperature. A longer time is required for full reduction of the acylhydrazone (2 hours) at the lower temperature, but 63 % crude yield was obtained.



Scheme 22: Proposed side reaction for c(BTM-KPGD).

Various potentially reactive or bulky amino acids were tested under the optimised cyclisation conditions. In most cases high yields are seen. The peptides tested so far are all equivalent to cyclic hexapeptides. The incorporation of the BTM into cyclic peptides with varying macrocycle ring size was next tested.

6.1.4 Varying Macrocycle Ring Size

The ring-size tolerance of the cyclisation conditions were tested by the inclusion of two to eight amino acids which, with the inclusion of the BTM which mimics the i+1 and i+2 positions of a β-turn, represent mimics of cyclic tetra- to decapeptides. The results are shown in Table 75.

Peptide sequence	Product (%)
c(*BTM*-PG)	0
c(*BTM*-KPG)	38
c(*BTM*-KAG)	55
c(*BTM*-APG)	55
c(*BTM*-AKPGA)	65
c(*BTM*-AKPGSA)	54
c(*BTM*-SAKPGASA)	17/77*

*Table 75: Crude yields for the BTM-containing cyclic peptides with varying macrocycle ring sizes. *2-step cyclisation using sodium cyanoborohydride as the reducing agent.*

6.1.4.1 c(BTM-PG)$_2$

For the PG sequence, rather than c(BTM-PG) forming, two peptides joined together to form c(BTM-PG)$_2$ as the major product. The equivalent of a cyclic octapeptide rather than a cyclic tetrapeptide was therefore formed. Tetrapeptides generally interconvert between various high energy conformations,[46] usually containing a cis peptide bond, with no β-turn structure present. Cyclic octapeptides, when they contain turn-inducing elements, generally form structures containing two β-turns.[194]

The PG subsequence has previously been observed as a turn-inducing sequence within peptides.[371] As the (PG-BTM)$_2$ peptide is composed of PG and the BTM which is designed to induce a turn within a peptide sequence, bias-exchange molecular dynamics (BE-META) simulations were carried out on this sequence to see if the PG turns or the BTM would be the dominant turn unit in the structure of the peptide.

Two major conformations are seen in the BE-META of c(BTM-PG)$_2$ (Figure 105). The major conformation (61%) resembles a conformation of two conjoined β-turns with the BTM occupying what would be the i+1 and i+2 positions of the β-turns at either side of the peptide with PG along the sides of the cyclic peptide occupying the i and i+3 positions. The major conformation shows the BTM is a preferable turn structure to PG in this peptide. The hydrogen-bond between the i and i+3 positions of the β-turn is not possible as proline does not have the amide hydrogen necessary to form the hydrogen-bond. Instead a hydrogen-bond is seen between the glycine amide hydrogen and the carbonyl group of each of the i positions of the two β-turns. This orientation is likely only due to the flexible nature of glycine. The minor conformation does not contain any β-turn structures. The wider conformation allows the proline residue to have more favourable dihedral angles which occur in the polyproline II (PPII) rather than the α-region of the Ramachandran plot.

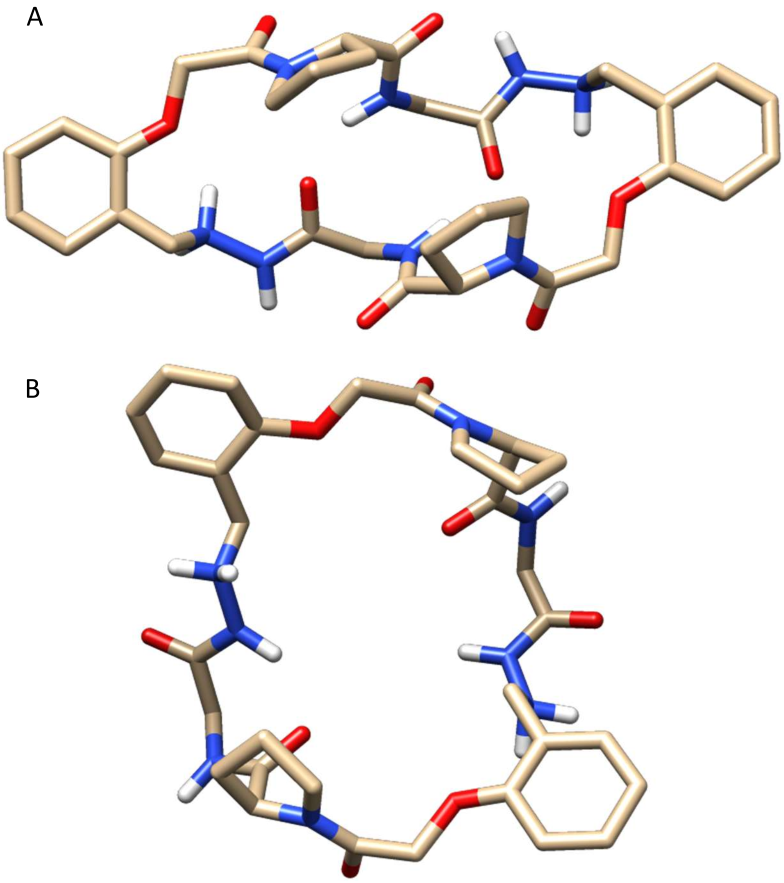
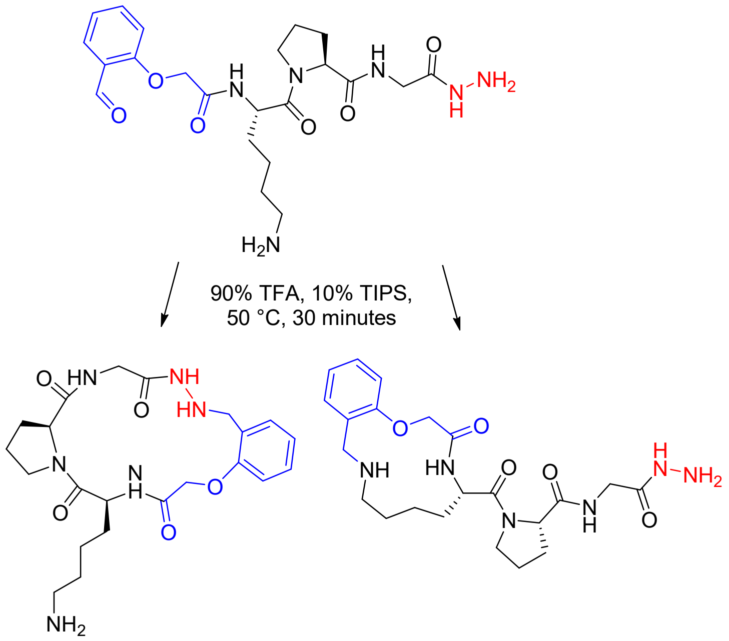


Figure 105: Conformations seen in the BE-META of c(BTM-PG)$_2$. Major conformation A (61%) and minor conformation B (39%).

6.1.4.2 Cyclic Pentapeptide Equivalents

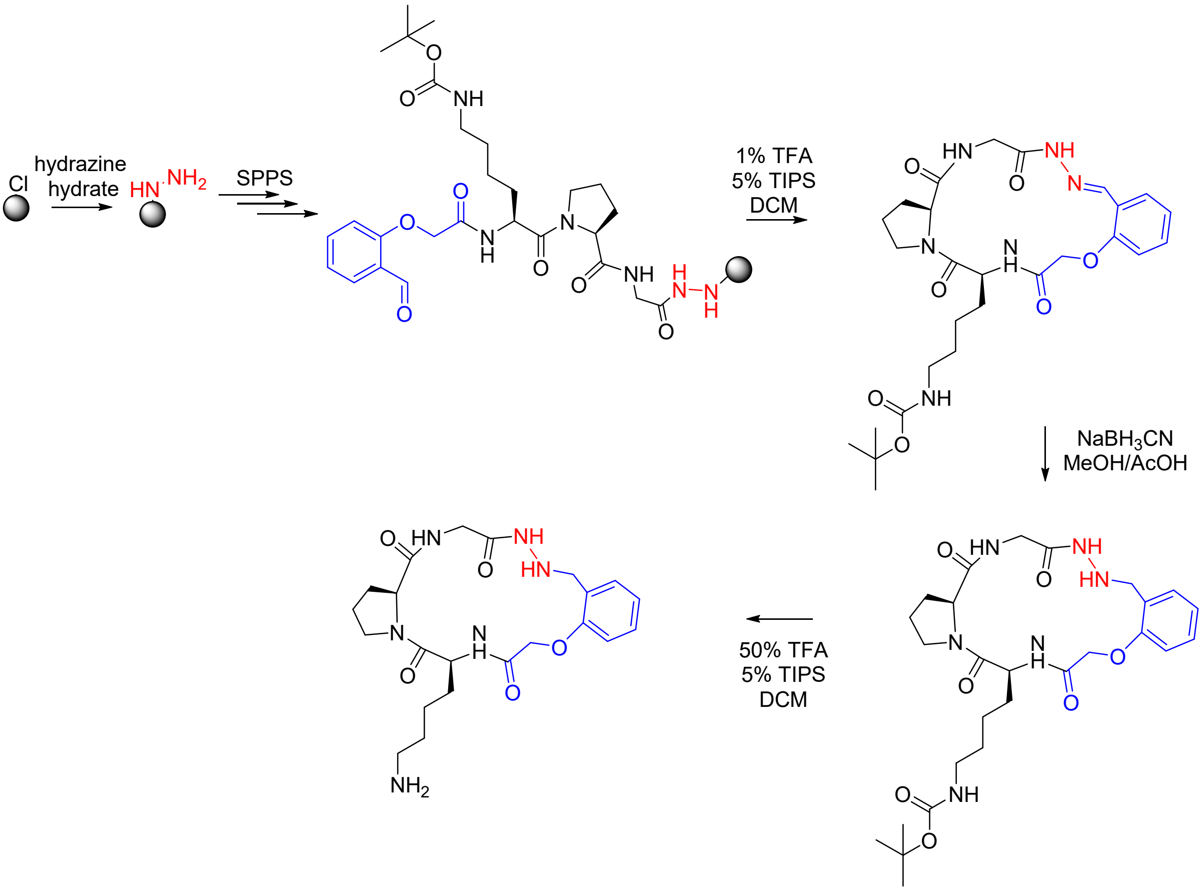
The peptide c(BTM-KPG) was synthesised which is the equivalent of a cyclic pentapeptide. Cyclic pentapeptides generally form a structure in solution that is made up of overlapping β- and γ-turns.[46] It is therefore expected the BTM will occupy what would be the i+1 and i+2 positions of the β-turn and the 3 amino acids form a γ-turn. The BTM would restrict the conformational freedom of the peptide preventing the amino acids changing register within the peptide.

For c(BTM-KPG) two peaks were seen on the LCMS with the product mass. The peaks were separable by HPLC and only one peak is seen for the similar sequence c(BTM-APG). It is therefore unlikely the two peaks are due to cis and trans proline isomerisation and instead the lysine sidechain reacted with the aldehyde rather than the hydrazine to produce the second product with the same mass (Scheme 23).



Scheme 23: Reaction of the aldehyde group with either the hydrazine or lysine sidechain.

In order to confirm cyclisation through the lysine sidechain was occurring during the synthesis of c(BTM-KPG) the peptide was cleaved from the resin using 1% TFA and 5% TIPS in DCM for 10 minutes at room temperature to yield the cyclised hydrazone with the Boc protecting group still present on the lysine. The hydrazone was reduced using sodium cyanoborohydride. The Boc protecting group was then removed using 50% TFA, 5% TIPS in DCM (Scheme 24). As cyclisation through the lysine sidechain is not possible by this method it allowed identification of the peaks cyclised through the hydrazone and the lysine sidechain seen in the one-step cleave and reduction method. The ratio of product cyclised through the hydrazine to that cyclised through the lysine sidechain was 1:1.3. The two-step method gave an improved crude yield of 61% rather than 38% seen in the one-step method.



Scheme 24: Alternative synthesis of c(BTM-KPG).

The lysine sidechain is less reactive than the hydrazine group and is not observed to react with the aldehyde in any of the other sequences tested. The peptide c(BTM-KAG) was synthesised and only one product peak was seen. The lysine sidechain reacting with the aldehyde group therefore seems to be a specific property of the c(BTM-KPG) sequence. The difference in crude yields seen for c(BTM-KPG) (38%) compared to c(BTM-APG) (55%) and c(BTM-KAG) (55%) demonstrates that the amino acid sequence chosen for the peptide can alter the cyclisation efficiency of the peptide through the BTM. Proline is often found in turn structures due to the relatively rigid ring structure of the amino acid making it ideal for introducing a kink into the peptide sequence. It may be that for this peptide the proline structure does not help bring the hydrazine and aldehyde into close proximity to allow for formation of the acylhydrazone but instead orientates the lysine sidechain closer to the aldehyde allowing the competing reaction to occur despite the reduced nucleophilicity.

6.1.4.3 Larger Macrocycles

The cyclic peptides c(BTM-AKPGA), c(BTM-AKPGSA) and c(BTM-SAKPGASA) were synthesised using the one-step reductive cleave conditions. With increasing macrocycle size the yield decreases with multiple side products being observed.

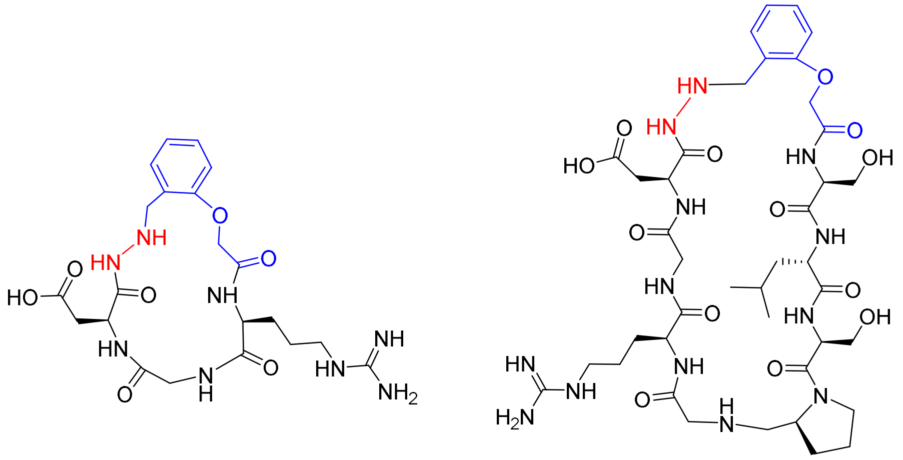
The c(BTM-SASKPGASA) peptide, which is equivalent to a cyclic decapeptide, was synthesised by the two-step method to see if an increased yield could be obtained. The peptide was cleaved from the resin using non-reductive conditions of 95% TFA, 5% water and then the resultant acylhydrazone reduced in a second step using sodium cyanoborohydride. An increased yield of 77% was seen compared to the 17% seen using the on-resin cyclisation with TIPS. The two-step method therefore presents an alternative synthetic route for larger peptides.

The BTM was next incorporated into cyclic peptides containing the biologically active RGD integrin-binding motif.

6.2 Incorporation of the β-Turn Mimic into Integrin-Binding Cyclic Peptides

Integrin binding proteins often contain the RGD sequence.[40] The RGD motif is observed to adopt different conformations depending on which type of integrin it binds.[40] Many cyclic peptides have been designed containing the RGD motif.[26, 35] As small cyclic peptides often form many conformations in solution, the biological activity of the RGD-containing cyclic peptides can be unpredictable so many variants are often tested during the design process.[41] Modifications such as the inclusion of D-amino acids or *N*-methylation which reduce the conformational flexibility of the peptide and help stabilise the conformation necessary for binding to a target are often beneficial.[45, 51] The BTM should help induce structure and reduce the flexibility of cyclic peptides. A series of RGD containing peptides were therefore designed making use of the BTM to help enforce the RGD motif to adopt the different conformations necessary to bind to different types of integrin.

Cilengitide is a small cyclic pentapeptide containing the RGD motif in a turn structure and is an agonist to αvβ3 and αvβ5 integrins.[37] The RGD motif in ligands which bind αIIbβ3 integrins however adopt an extended linear conformation.[40] The peptide c(BTM-RGD) was designed to mimic the structure of cilengitide and the peptide c(BTM-SLSPGRGD) was designed to mimic the elongated RGD conformation found in the ligands of αIIbβ3 integrins. A cyclic β-hairpin structure was expected for c(BTM-SLSPGRGD) with the PG subsequence forming a β-turn opposite the BTM. Serine and leucine are frequently observed in β-strand structures,[305] so were chosen to occupy the positions opposite to the RGD motif.



c(BTM-RGD)

c(BTM-SLSPGRGD)

Figure 106: RGD-containing cyclic peptides incorporating the BTM.

The peptides were synthesised using the one-step cleave/cyclisation method. Crude yields of 77% and 34% were obtained for c(BTM-RGD) and c(BTM-SLSPGRGD) respectively.

Peptide Sequence	Product (%)
c(BTM-RGD)	77
c(BTM-SLSPGRGD)	34

Table 76: RGD-containing cyclic peptides

6.2.1 BE-META

Bias exchange molecular dynamics (BE-META) simulations were used to explore the conformations of the RGD-containing cyclic peptides. The BTM should help reduce the conformational freedom of the cyclic peptides and help induce structure to retain the RGD motif in either the bent or linear orientations. The conformations observed in the BE-META were compared to the available crystal structures of cilengitide bound to the extracellular segment of αVβ3 integrin (PDB: 1L5G)[47] or the short peptide ligand LGGAKQRGDV bound to αIIbβ3 integrin (PDB: 2VDR).[365]

Inclusion of the BTM in a cyclic peptide may mean that additional CVs are required in the BE-META simulations for full exploration of the energy landscape of the system. A 500 ns simulation was carried out on c(BTM-AGGA) to assess whether it was necessary to include further CVs when determining the conformation of cyclic peptides containing the BTM. The dihedral angles within the BTM were monitored throughout the simulation as potential CVs. It was observed that the BTM dihedral angles, although helping to describe the metastable states of the system, interconvert readily even within the relatively restrained cyclic peptide. Therefore no additional CVs are necessary in the simulations than those already used in the BE-META of cyclic peptides. In order to confirm this a BE-META simulation was run on c(BTM-RGD) both with and without the inclusion of the BTM dihedral angles as CVs. The same results were seen in both simulations further supporting that additional CVs aren't necessary to fully explore the energy landscape of BTM containing cyclic peptides.

6.2.1.1 c(BTM-RGD)

Two major conformations are seen in the simulation of c(BTM-RGD) in an approximate 2:1 ratio. Both conformations are similar but differ slightly at the aspartic acid residue which switches from the β-sheet region in the major conformation to the α-region of the Ramachandran plot in the minor conformation. Both conformations have a similar RGD conformation to that of cilengitide with backbone RMSD values of 0.201 Å and 0.281 Å respectively (Figure 107). It is therefore likely c(BTM-RGD) would be able to bind to αvβ3 and αvβ5 integrins.

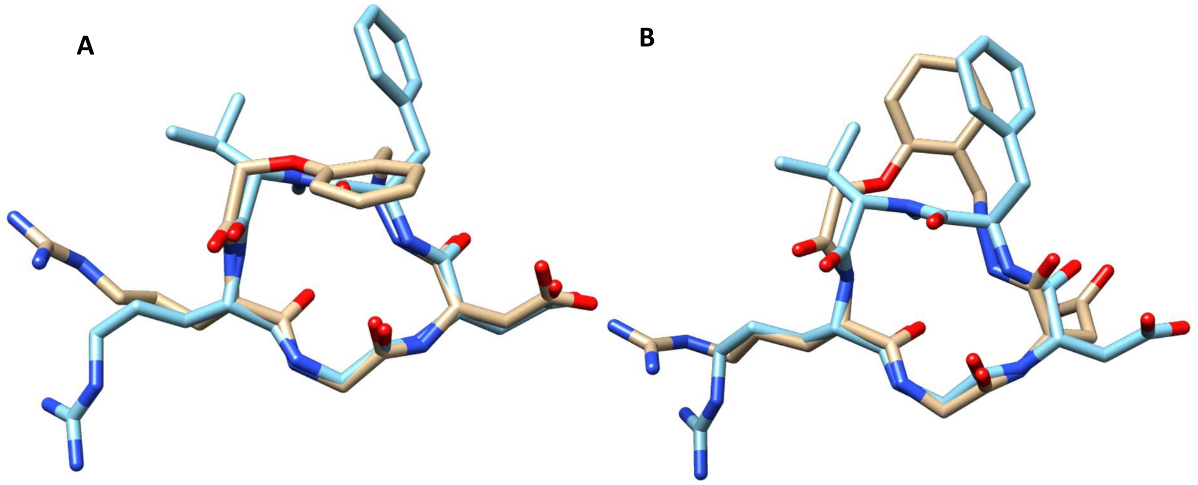


Figure 107: Cilengitide (blue) overlayed with the major conformation (A) and minor conformations (B) of c(BTM-RGD).

6.2.1.2 c(BTM-SLSPGRGD)

One major conformation is seen in the BE-META of c(SLSPGRGD-BTM). The PG subsequence forms a type II turn at the opposite side of the cyclic peptide from the BTM with the regions between the two turn structures adopting a β-sheet structure. A backbone RMSD of the RGD sequence with the crystal structure of LGGAKQRGDV bound to αIIbβ3 integrin of 0.903 Å is seen (Figure 108).

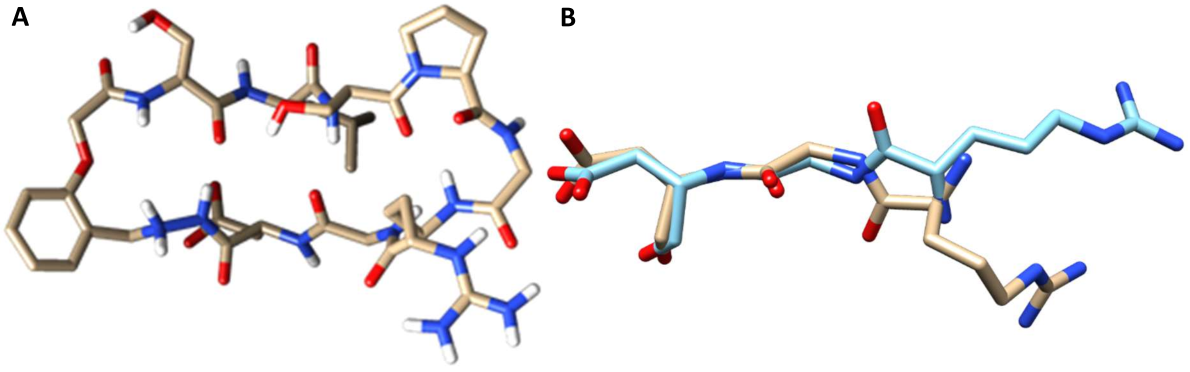


Figure 108: c(BTM-SLSPGRGD) (A) and the overlay of the RGD subsequence of c(BTM-SLSPGRGD) and the LGGAKQRGDV ligand (B).

The X-ray crystal structure of the LGGAKQRGDV peptide bound to αIIbβ3 integrin contains the RGD motif in an elongated linear conformation but does not have a β-hairpin structure. The dihedral angles of the RGD sequence therefore do not all occupy the β-region of the Ramachandran plot (Figure 109). This is likely why the cilengitide/c(BTM-RGD) overlay produced lower RMDS values.

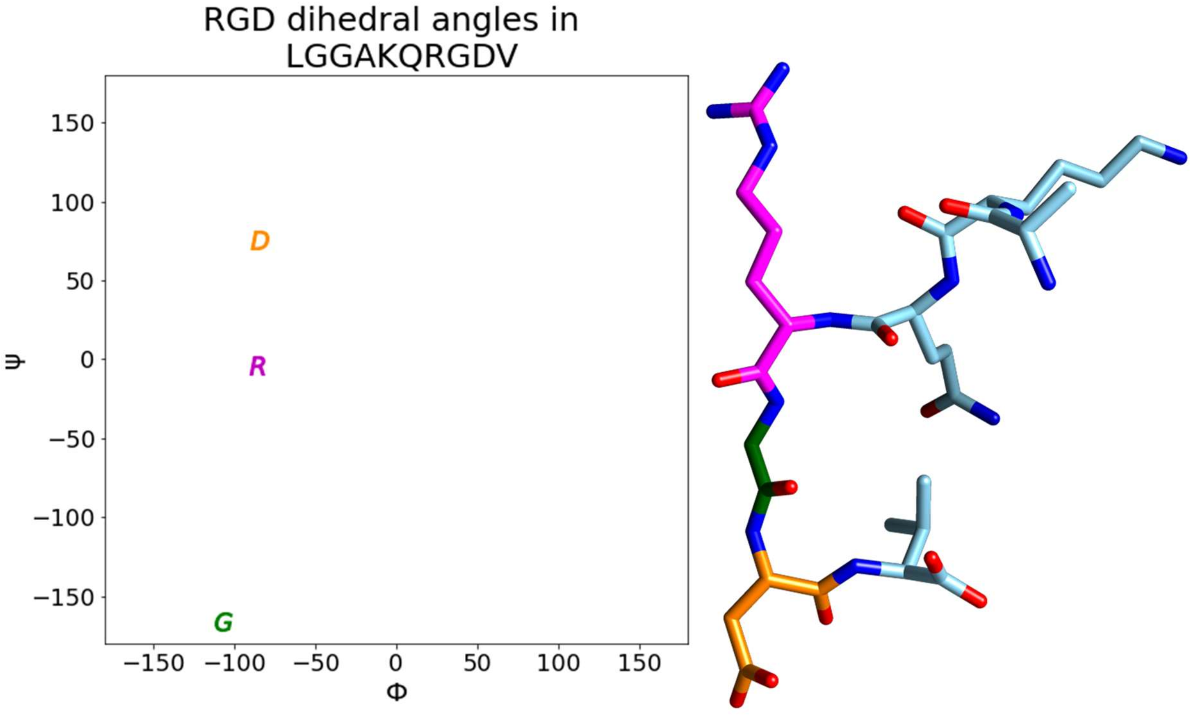


Figure 109: Dihedral angles of the RGD motif in the LGGAKQRGDV ligand when bound to αIIbβ3 integrin.

These results show that the BTM can be used as a means of cyclisation and help control conformations within cyclic peptides. The amino acids selected to complete the cyclic peptide should be chosen to help produce the desired conformation.

6.3 Conclusions

Conditions were found for the incorporation of the BTM into cyclic peptides directly during cleavage of the peptide from the resin following SPPS. The equivalent of 5-10 residue containing cyclic peptides can be synthesised in high yields. For larger sequences (greater than the equivalent of a heptapeptide) an alternative two-step method where the acylhydrazone is reduced by sodium cyanoborohydride following cleavage from the resin produces higher yields. Asparagine at the position adjacent to the ligation junction led to decomposition of the product but the remaining potentially nucleophilic amino acids tested were tolerated.

The cyclic peptides c(BTM-RGD) and c(BTM-SLSPGRGD) were designed to contain the integrin-binding RGD motif in different conformations. BE-META simulations of the peptides show the BTM, in addition to providing a means of cyclisation, helps induce structure leading to well-defined conformations.

As the BTM can be used to both induce structure and cyclise a peptide it could potentially be incorporated into many other biologically active cyclic peptides. Chapter 8 makes use of the BTM in the design of a cyclic WW domain mimic. Modification of the structure of the BTM unit could potentially be used to introduce additional functionality into a peptide. In chapter 7 the BTM is modified by the incorporation of a fluorescent naphthalene unit.

7 Naphthalene Containing β-turn Mimic

Fluorescence has been used extensively to help study proteins including studies of structure, ligand-binding effects and dynamics.[372-378] Of the 20 naturally occurring amino acids phenylalanine, tyrosine and tryptophan are intrinsically fluorescent. However they have a low natural abundance and have suboptimal fluorescent properties. Other methods to introduce fluorescence into proteins have therefore been established. Fluorescent amino acids have been developed and incorporated into proteins by replacing natural amino acids within the protein sequence.[372, 375] This has the advantage over covalently linking large fluorescent units such as green fluorescent protein (GFP) to a protein, as they are less likely to alter the properties being monitored such as function, localization, and native interactions of the target protein. Additionally they can be incorporated using standard SPPS. The BTM unit allows for ligation of two peptides whilst allowing secondary structure within the peptide by inducing a β-turn-like structure. Further versions of the BTM could be developed which include a small fluorescent unit adding additional functionality to the peptide.

7.1 Modification of the BTM to Include Naphthalene

Polycyclic aromatic hydrocarbons have intrinsic fluorescence,[379] derivatives of which can therefore be incorporated into fluorophores.[380-385] One of the simplest fluorophores of this class is naphthalene. The original BTM was synthesised using 2-formylphenoxyacetic which could be adapted by addition of another aromatic ring to incorporate a naphthalene group into the BTM (Figure 110). This would hopefully incorporate fluorescence into the BTM.

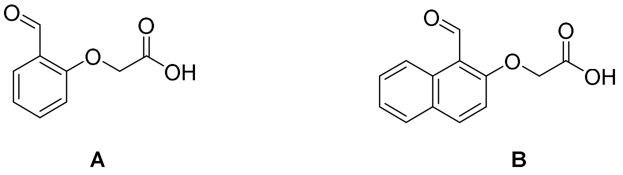
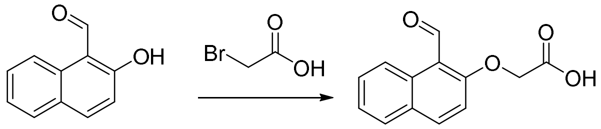


Figure 110: 2-formylphenoxyacetic acid (A) and 2-[(1-formylnaphthalen-2-yl)oxy]acetic acid (B).

7.1.1 Synthesis of 2-[(1-formylnaphthalen-2-yl)oxy]acetic acid

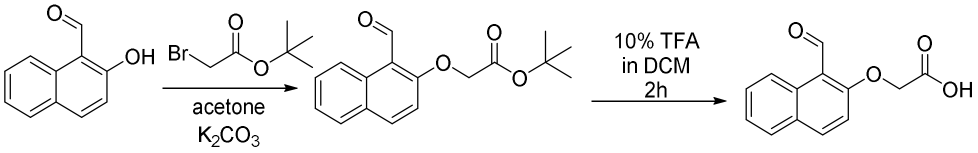
2-hydroxy-1-naphthaldehyde is commercially available and it was hoped a structure analogous to the 2-formylphenoxy acetic acid used to make the original BTM could be synthesised through a reaction with bromoacetic acid (Scheme 25). The resulting 2-[(1-formylnaphthalen-2-yl)oxy]acetic acid could then be coupled into a peptide by SPPS in order to make the BTM.



Scheme 25: Synthesis of 2-[(1-formylnaphthalen-2-yl)oxy]acetic acid.

Initially ten equivalents of bromoacetic acid was refluxed overnight with 2-hydroxy-1-naphaldehyde in acetone with 10 equivalents of potassium carbonate. Only starting material was recovered. Using 0.5 M aqueous sodium hydroxide as an alternative solvent/base to potassium carbonate also failed to produce the desired product. An alternative synthesis was therefore planned. Rather than coupling bromoacetic acid directly, *tert*-butyl bromoacetate was used. The *tert*-butyl group can then be removed in a separate step to give the carboxylic acid (Scheme 26).

When 2-hydroxy-1-naphthaldehyde was reacted with 1 equivalent of *tert*-Butyl bromoacetate with 1 equivalent of anhydrous potassium carbonate the reaction proceeded in acetone when refluxed overnight to give a 92% yield. The *tert*-Butyl group was then removed using 10% TFA in DCM for 2 hours to give 2-[(1-formylnaphthalen-2-yl)oxy]acetic acid in an 80% yield.



Scheme 26: Alternative synthesis of 2-[(1-formylnaphthalen-2-yl)oxy]acetic acid.

7.2 TrpZip

The naphthalene based BTM was initially incorporated into a TrpZip based peptide. TrpZip peptides form a stable β-hairpin motif stabilised by tryptophan cross-strand pairs.[192] They are therefore an ideal system to test whether the incorporation of a naphthalene moiety disrupts the ability of the BTM to form a stable β-turn like structure. TrpZip1 has the sequence SWTWSGNKWTWK with the GN occupying the i+1 and i+2 positions of a type II' β-turn. TrpZip1 was used as a model peptide where the GN turn was replaced with the naphthalene BTM (Figure 111).

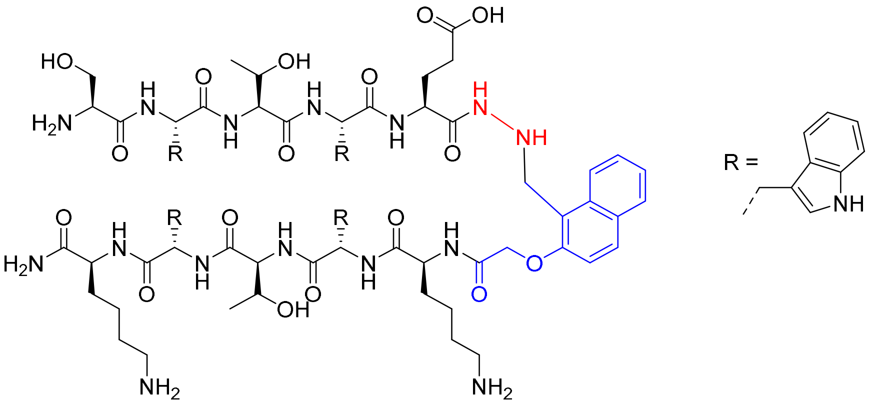
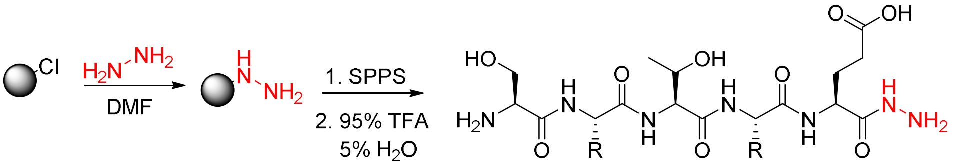


Figure 111: TrpZip incorporating the naphthalene-based BTM at the i+1 and i+2 positions of the β-turn.

7.2.1 Synthesis of TrpZip Peptide

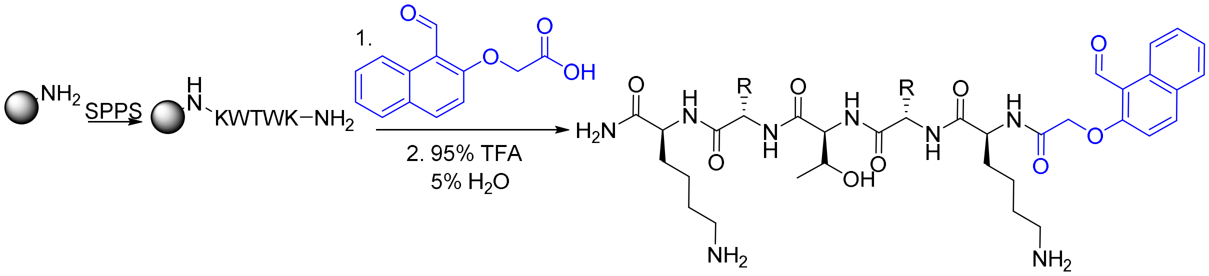
The TrpZip peptide was synthesised in two fragments which are then ligated together to form the BTM. The first TrpZip fragment was synthesised by the same method as previously developed for formation of the BTM. Hydrazine hydrate was reacted with Cl-trityl resin to create a hydrazine resin. SPPS was then used to couple the amino acids. The peptide can then be cleaved from the resin ready for ligation (Scheme 27).



Scheme 27: Synthesis of TrpZip fragment 1.

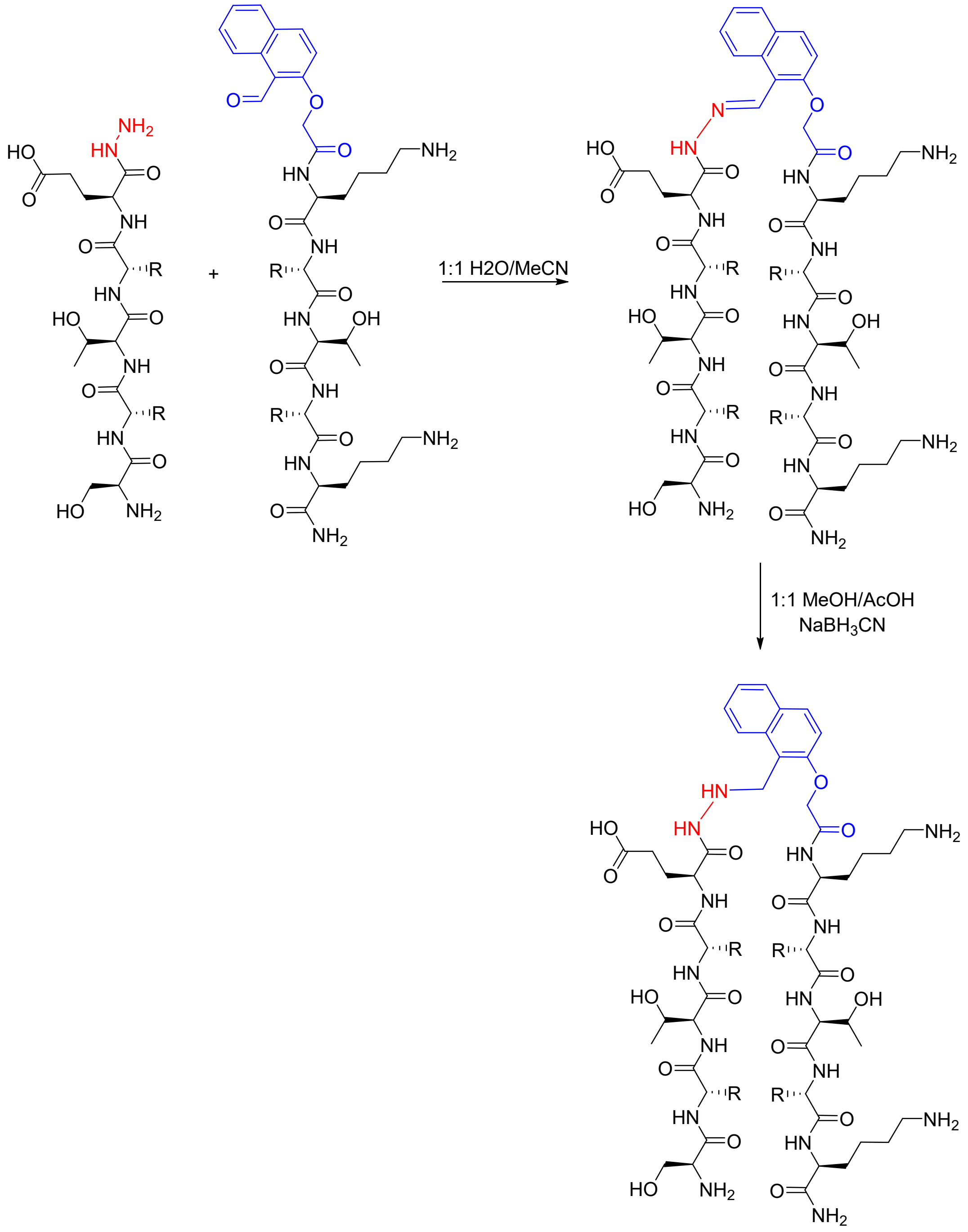
The second TrpZip fragment was synthesised by SPPS using Rink Amide AM resin. The 2-[(1-formylnaphthalen-2-yl)oxy]acetic acid can be coupled into the peptide following the other residues through the carboxylic acid functionality by standard SPPS procedures (Scheme 28). For the original BTM the 2-formylphenoxyacetic acid was coupled into the peptide using 4.5 equivalents DIC, 4.5

equivalents oxyma and 6 equivalents DIPEA. These conditions were therefore initially tested to couple the 2-[(1-formylnaphthalen-2-yl)oxy]acetic acid, however as naphthalene is a bulkier substituent the reaction was allowed to react overnight rather than for 2.5 hours. Despite being allowed to react for an increased time, incomplete coupling was seen with approximately 47% coupled peptide produced. Double coupling the naphthalene improved this to 63%. HATU and PyBOP were tried as alternative coupling reagents. However when HATU/DIPEA and PyBOP/DIPEA were used no product was seen but rather a mass -18 Da from the product. The peptide was cleaved from the resin using 95% TFA, 5% water.



Scheme 28: Synthesis of TrpZip fragment 2.

The two TrpZip peptide fragments were purified and lyophilised together after mixing in 1:1 water/acetonitrile. The hydrazine group on the first TrpZip fragment reacts with the aldehyde group on the second fragment to form a hydrazone. The acylhydrazone can then be reduced by sodium cyanoborohydride to an acylhydrazine (Scheme 29). Although coupling the naphthalene to the peptide was difficult, possibly due to the added steric bulk of the additional benzene ring, the ligation reaction worked well with all the peptide reacting with the other fragment.



Scheme 29: Ligation of the two TrpZip fragments.

7.3 CD

Circular dichroism (CD) spectroscopy was used to characterize the structure and thermal stability of the naphthalene TrpZip peptide. The CD spectra were obtained for 190-260 nm UV with a peptide concentration of 10 µM in MOPS buffer at pH 7.4. The spectra of TrpZip1 and TrpZip with the original BTM previously obtained in the group were compared to the naphthalene BTM TrpZip.[191]

β-hairpins produce CD spectra with a characteristic shape, with a minimum at around 215 nm.[386, 387] The TrpZip structure is partially stabilised by interactions between the tryptophan sidechains. Such interactions can affect the CD spectra, often leading to a positive band in the 225-235 region.[387, 388] The CD spectra for the naphthalene TrpZip peptide (Figure 112) is indicative of a β-strand structure with tryptophan sidechain interactions occurring, but the CD spectra differs slightly from the CD spectra of the TrpZip peptide containing the original BTM and TrpZip1. A change in conformation of the tryptophan sidechains is therefore possibly being seen due to the large naphthalene group. The naphthalene group presents an alternative area the indole sidechains could potentially interact with. Additionally MOPS rather than potassium phosphate buffer was used which could potentially affect the conformation of the peptide and therefore the CD spectra.

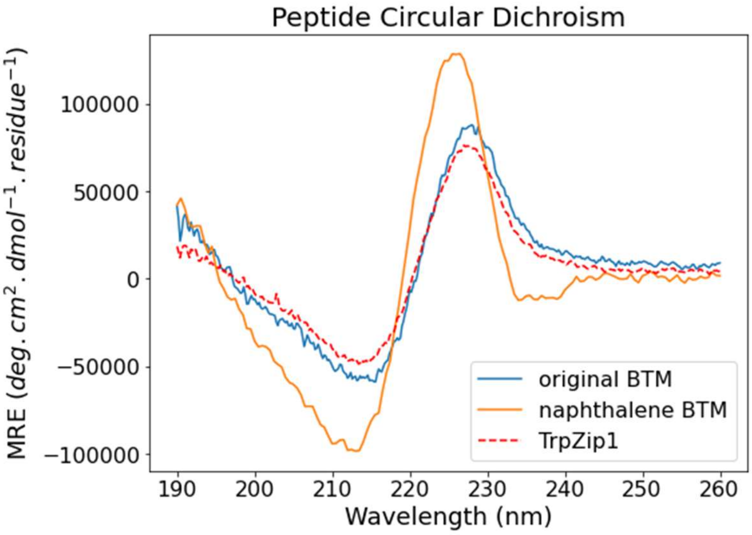


Figure 112: CD spectra of TrpZip1 and the BTM containing derivatives.

The thermal stability of the peptide was tested by obtaining the melt curves. The mean residue ellipticity (MRE) was measured at 225 nm across a temperature range of 5 to 80 °C. The CD spectra of the peptide before and after the melt were superimposable, demonstrating the reversibility of the peptide folding. The melting curve obtained was similar to that of TrpZip1 and the TrpZip peptide containing the original BTM.

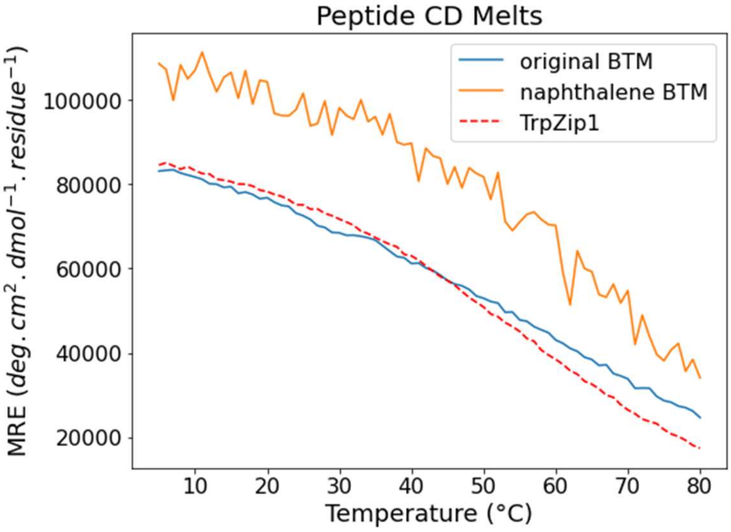


Figure 113: Melt curves for the TrpZip peptides.

The melting temperature (T_m) for the naphthalene-containing peptide is equal to when half the peptide is in an unfolded state and was obtained by measuring the fraction of folding by the following equation based on a two-state unfolding equilibrium:

$$\alpha = \frac{[F]}{[F]+[U]} = \frac{(\theta_t - \theta_U)}{(\theta_F - \theta_U)} \qquad (11)$$

Where the fraction of folded peptide, α, is equal to the concentration of the folded peptide [F] over the total concentration of peptide (folded and unfolded ([U]) peptide). The fraction can be determined by the ellipticity at a specific temperature, θ_t, if the ellipticity of the folded (θ_F) and unfolded states are known (θ_U). The ellipticity of the unfolded state cannot be determined as the shape of the melting curve shows the peptide had still not reached a completely unfolded state by 80 °C. A MRE of zero was therefore used to represent the unfolded state.

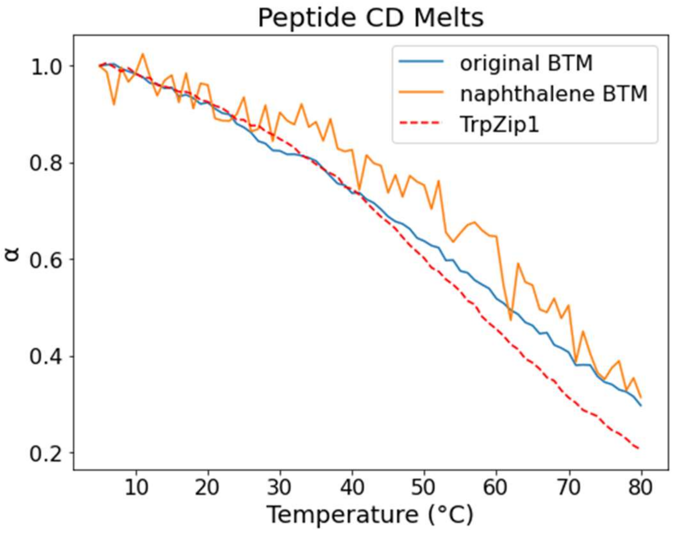


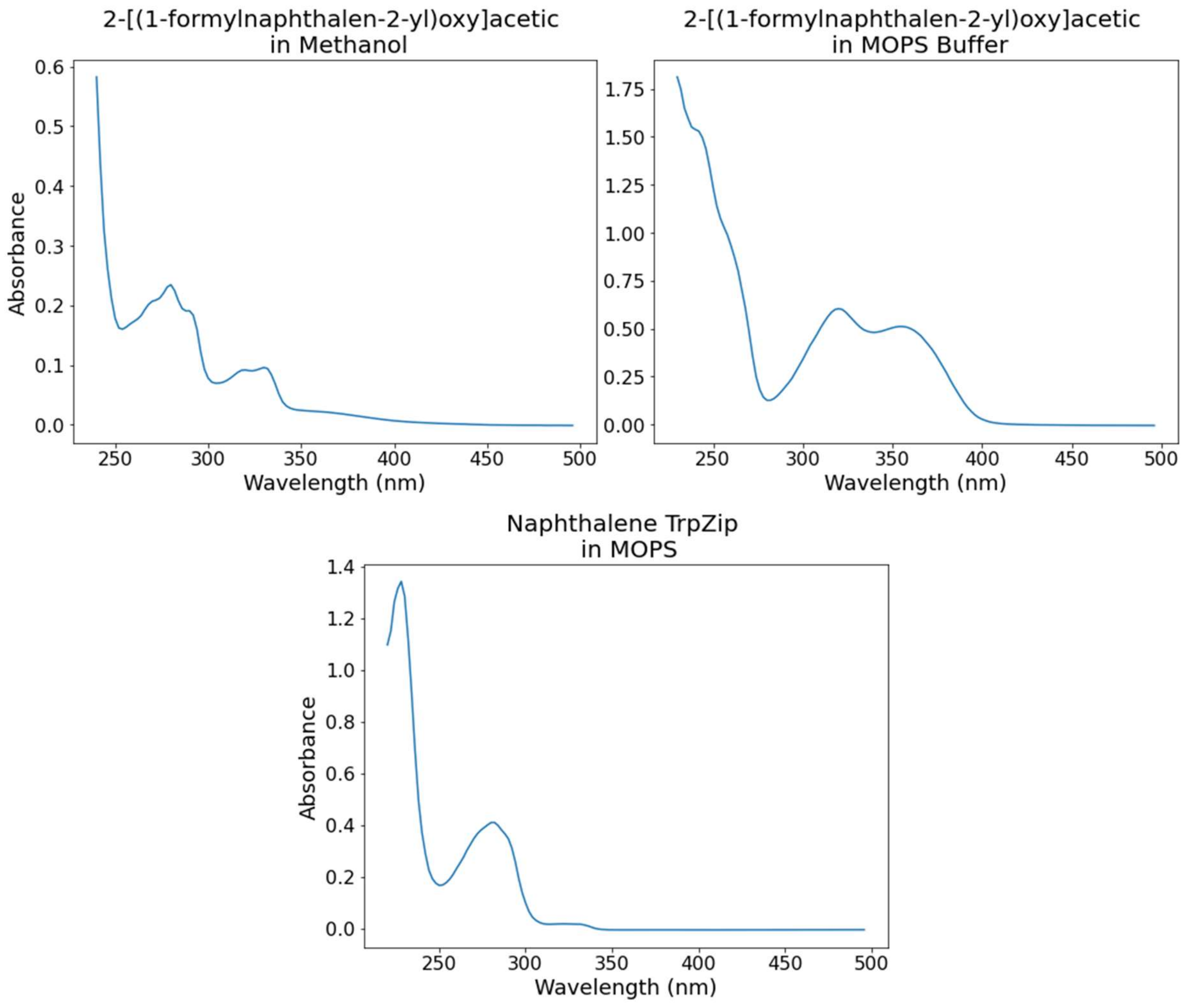
Figure 114: Fraction of folded peptide with temperature.

The T_m values for TrpZip1, the original BTM TrpZip peptide and the naphthalene BTM TrpZip are 57, 61 and 67 °C respectively. The naphthalene containing BTM TrpZip peptide therefore has increased thermal stability compared to the other two peptides. If the tryptophan sidechains are interacting with the naphthalene group this could potentially lead to increased stability of the peptide.

7.4 Fluorescence

The naphthalene BTM allows for retention of the β-turn structure based on the CD spectra of the TrpZip based peptide. The fluorescence spectra were next obtained to determine if the naphthalene BTM was introducing fluorescence into the peptide.

The absorption spectra for the 2-[(1-formylnaphthalen-2-yl)oxy]acetic acid and the naphthalene TrpZip are shown in Figure 115. The spectrum for 2-[(1-formylnaphthalen-2-yl)oxy]acetic acid was obtained in MOPS buffer at pH 7.4 to match the conditions used for the peptide as well as in methanol due to limited solubility in the aqueous solution. The spectrum in MOPS differs from that in methanol. It is possible the naphthalene units are interacting in MOPS causing changes in the spectrum. The spectrum for the naphthalene TrpZip shows a maxima around 280 nm which is likely resulting from the absorption of the naphthalene unit as well as the tryptophan residues within the peptide. A small peak is also seen around 330 nm which could potentially be due to the naphthalene as a similar peak was seen for the 2-[(1-formylnaphthalen-2-yl)oxy]acetic acid spectrum in methanol. The absorption spectrum of TrpZip1 would have to be obtained to allow for direct comparison to help determine if this peak is due to the naphthalene.


2-[(1-formylnaphthalen-2-yl)oxy]acetic in Methanol

2-[(1-formylnaphthalen-2-yl)oxy]acetic in MOPS Buffer

Naphthalene TrpZip in MOPS

Figure 115: The absorption spectra for 2-[(1-formylnaphthalen-2-yl)oxy]acetic acid in both methanol and MOPS buffer and the spectrum of the naphthalene TrpZip.

The fluorescence spectrum of the 2-[(1-formylnaphthalen-2-yl)oxy]acetic acid prior to coupling into a peptide was initially obtained to determine the fluorescent properties of the BTM building block and if they would be retained upon changes to the structure when incorporated into a peptide (Figure 116). The emission spectrum of the 2-[(1-formylnaphthalen-2-yl)oxy]acetic has two peaks in methanol at approximately 340 and 350 nm. In buffer the spectrum is less well defined but a similar maxima around 350 nm is also seen. Differences in the fluorescence of the naphthalene due to solvent effects may also mean that different fluorescence properties are seen in different environments when incorporated into a peptide. Environment sensitive fluorescence can be a useful feature of fluorescent probes.

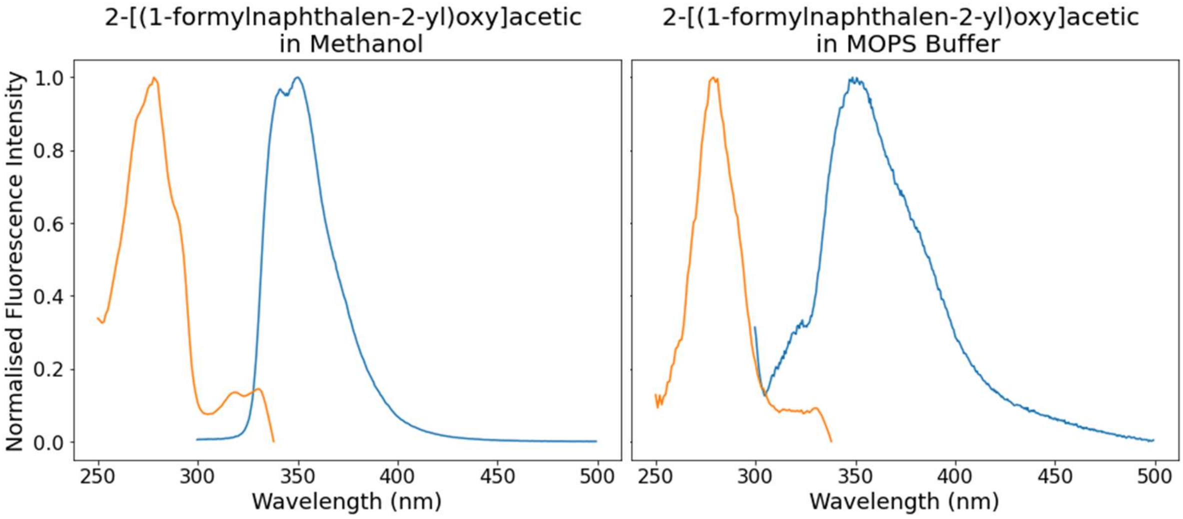


Figure 116: Excitation and Emission spectra (orange and blue respectively) for 2-[(1-formylnaphthalen-2-yl)oxy]acetic acid dissolved in either methanol or MOPS buffer. The emission spectra were obtained using an excitation wavelength of 270 nm and the excitation spectra measured at 350 nm.

Incorporating 2-[(1-formylnaphthalen-2-yl)oxy]acetic acid into a peptide will alter its fluorescent properties as the naphthalene substituents are changed with both the coupling and ligation steps. The fluorescent properties of the TrpZip peptide containing the naphthalene BTM were therefore also measured (Figure 117).

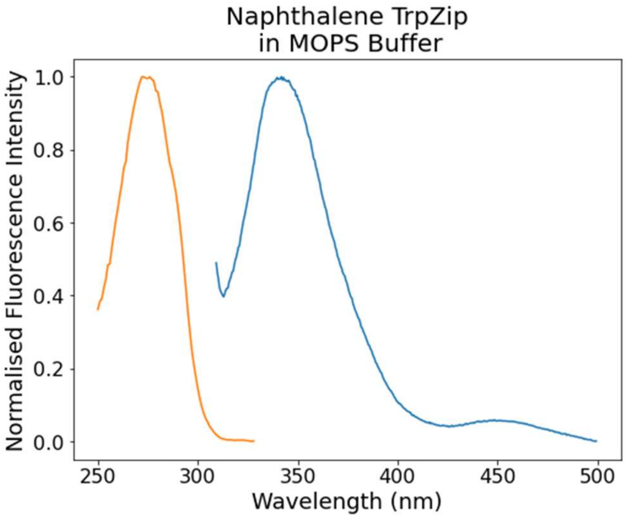


Figure 117: Excitation and emission spectra (orange and blue respectively) for the naphthalene-BTM containing TrpZip peptide. The emission spectra was obtained using an excitation wavelength of 290 nm and the excitation spectra measured at 325 nm.

The TrpZip peptide contains four tryptophan residues. Tryptophan has intrinsic fluorescence. One of the challenges of making fluorescent amino acids or incorporating fluorescence into peptides therefore is making sure they do not overlap with the fluorescent properties of tryptophan and other natural amino acids with intrinsic fluorescence. An emission maxima at approximately 350 nm is still seen following the incorporation of the 2-[(1-formylnaphthalen-2-yl)oxy]acetic acid into the TrpZip peptide. However tryptophan is also know to have a peak in its emission spectra at 350 nm so it is difficult to distinguish the fluorescent properties of the tryptophan residues and the naphthalene group.

A range of excitation wavelengths were tested to obtain the emission spectrum for the peptide due to the naphthalene whilst minimising the effects of tryptophan. If an excitation wavelength could be used for the naphthalene which did not overlap with that of tryptophan, then the fluorescence of the naphthalene could be monitored without the influence of the tryptophan residues in the protein. The naphthalene BTM and tryptophan have overlapping excitation and emission spectra however so the fluorescence properties of the naphthalene and tryptophan are too similar to monitor the effects of just the naphthalene fluorescence. It is likely both the tryptophan and naphthalene are contributing to the fluorescence of the peptide but an alternative peptide without any of the naturally occurring amino acids with intrinsic fluorescence would need to be made to confirm the fluorescent properties of the naphthalene BTM once incorporated into a peptide.

7.5 Conclusions

Substitution of the BTM could allow for incorporation of additional functionality into peptides. The BTM was therefore modified to incorporate a fluorescent naphthalene group. 2-[(1-formylnaphthalen-2-yl)oxy]acetic acid was synthesised and incorporated into a small TrpZip based peptide. The larger substituent made coupling of the aldehyde into the peptide more difficult, but once coupled did not prevent ligation of the two peptide fragments to form the BTM.

The CD spectra indicate the BTM retained its structural ability to mimic the i+1 and i+2 positions of a β-turn within the peptide as the β-hairpin structure of the TrpZip peptide was retained. A slight increased thermal stability of the peptide was seen compared to the original BTM and TrpZip1. Differences in the CD spectra show the naphthalene BTM has the potential to slightly change the conformation of peptides it is incorporated into possibly by interacting with aromatic sidechains of the tryptophan. Molecular dynamics (MD) simulations or NMR could help determine the effects of the naphthalene group on the conformations of the peptide.

The naphthalene BTM has fluorescent properties similar to that of tryptophan so its usefulness in incorporating fluorescence into peptides is limited. The naphthalene BTM could potentially still function in a peptide without the naturally occurring fluorescent amino acids but would have limited applications in a cellular environment. An alternative fluorescent group could potentially be incorporated into the BTM with fluorescent spectra which did not overlap with that of tryptophan (or tyrosine/phenylalanine). Potential fluorescent groups that have previously been incorporated into fluorescent amino acids include those based on the dansyl group, nitrobenzoxadiazoles (NBDs) and BODIPY.[375] For the polycyclic aromatic hydrocarbons other fluorophores are generally much larger than naphthalene and therefore could potentially begin to alter the conformation of the peptide it is incorporated into. Additionally coupling of the naphthalene group into the peptide already did not go to completion, likely due to sterics, so the efficiency of the coupling step may need to be improved if a larger fluorescent group is to be incorporated. An alternative method could potentially be to develop a BTM with a further ligation point which could be used to link a fluorophore in a separate step. Despite the suboptimal fluorescent properties, the naphthalene BTM demonstrates how the BTM structure can be adapted to add additional functionality.

8 Design of a Cyclic WW Domain

Chapter 6 shows how the BTM can be incorporated into cyclic peptides allowing for cyclisation and induction of structure. The β-turn structure of the BTM could be particularly useful for maintaining a β-hairpin structure by joining two antiparallel β-strands. This could potentially allow for the design of peptides mimicking β-hairpin regions of a protein. By enforcing the presence of a β-turn a β-hairpin structure may be more likely to be retained when an amino acid sequence is removed from the protein structure, especially when used in addition to cyclising the peptide, which could further reduce the degrees of freedom of the peptide. If the β-hairpin structure of a protein contains amino acids involved in ligand binding, the BTM could potentially be used to design a peptide mimicking the β-hairpin structure necessary to retain the ability to bind to a ligand.

WW domains are small proteins, involved in a variety of protein-protein interactions (PPIs), with a structure containing β-strands. They therefore offer an ideal system to test if incorporating the BTM into a cyclic peptide would help retain a β-hairpin structure and allow for the design of a peptide which retains the ability of the protein to bind to a ligand. This chapter looks at the design of a cyclic WW domain mimic using the BTM to retain the β-sheet structure necessary for binding.

8.1 WW Domains

WW domains are small proteins, approximately 40 amino acids in length, which form a three stranded antiparallel β-turn structure (Figure 118). They contain two conserved tryptophan residues, around 20-23 residues apart (Figure 119), the first of which is important for the protein structure and the second is involved in ligand binding.[389, 390]

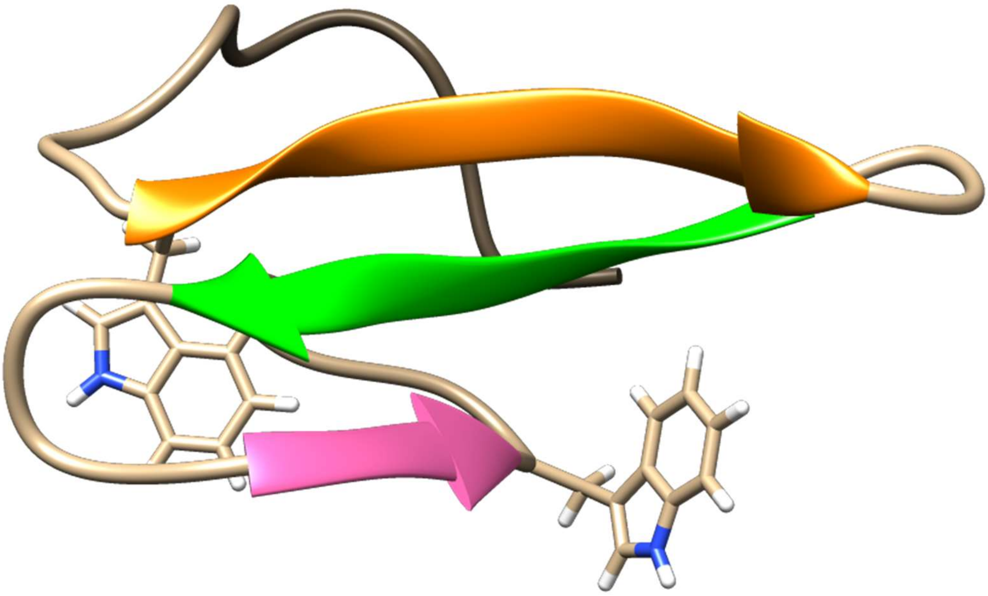
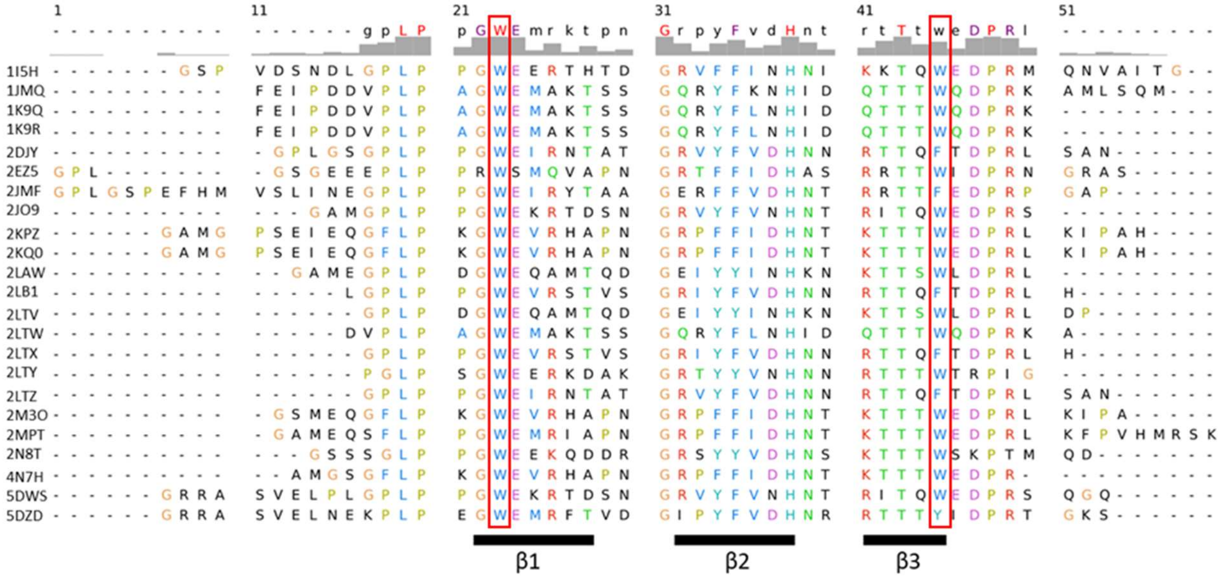


Figure 118: PDB: 1JMQ. The three strands that make up the β-sheet structure are shown in orange, green and pink respectively. The conserved tryptophan residues are also shown.

There are several classes of WW domains which mediate PPIs and have therefore been implicated in various diseases including Alzheimer's and cancer.[391-394] The WW domain classes are based on the consensus sequence of the ligands they bind, all of which are proline-rich sequences.[395, 396] The largest class of WW domains (Class I) bind ligands containing the sequence (L/P)PxY where x is usually proline and the tyrosine may be phosphorylated. The phosphorylation of this tyrosine is thought to be involved in regulating the binding of the ligand to WW domains.[397-399] The PPxY motif necessary for binding of Class I WW domains is often found on a relatively unstructured loop region of proteins and is therefore relatively flexible.[395]



```
      1          11         21         31         41         51
      ---------- ------gpLP pGWEmrktpn GrpyFvdHnt rtTtweDPRI ---------
1I5H  -------GSP VDSNDLGPLP PGWEERTHTD GRVFFINHNI KKTQWEDPRM QNVAITG--
1JMQ  ---------- FEIPDDVPLP AGWEMAKTSS GQRYFKNHID QTTTWQDPRK AMLSQM---
1K9Q  ---------- FEIPDDVPLP AGWEMAKTSS GQRYFLNHID QTTTWQDPRK ---------
1K9R  ---------- FEIPDDVPLP AGWEMAKTSS GQRYFLNHID QTTTWQDPRK ---------
2DJY  ---------- -GPLGSGPLP PGWEIRNTAT GRVYFVDHNN RTTQFTDPRL SAN------
2EZ5  GPL------- -GSGEEEPLP PRWSMQVAPN GRTFFIDHAS RRTTWIDPRN GRAS-----
2JMF  GPLGSPEFHM VSLINEGPLP PGWEIRYTAA GERFFVDHNT RRTTFEDPRP GAP------
2JO9  ---------- ---GAMGPLP PGWEKRTDSN GRVYFVNHNT RITQWEDPRS ---------
2KPZ  ------GAMG PSEIEQGFLP KGWEVRHAPN GRPFFIDHNT KTTTWEDPRL KIPAH----
2KQ0  ------GAMG PSEIEQGFLP KGWEVRHAPN GRPFFIDHNT KTTTWEDPRL KIPAH----
2LAW  ---------- --GAMEGPLP DGWEQAMTQD GEIYYINHKN KTTSWLDPRL ---------
2LB1  ---------- -----LGPLP PGWEVRSTVS GRIYFVDHNN RTTQFTDPRL H--------
2LTV  ---------- ------GPLP DGWEQAMTQD GEIYYINHKN KTTSWLDPRL DP-------
2LTW  ---------- -----DVPLP AGWEMAKTSS GQRYFLNHID QTTTWQDPRK A--------
2LTX  ---------- ------GPLP PGWEVRSTVS GRIYFVDHNN RTTQFTDPRL H--------
2LTY  ---------- ------PGLP SGWEERKDAK GRTYYVNHNN RTTTWTRPIG ---------
2LTZ  ---------- ------GPLP PGWEIRNTAT GRVYFVDHNN RTTQFTDPRL SAN------
2M3O  ---------- -GSMEQGFLP KGWEVRHAPN GRPFFIDHNT KTTTWEDPRL KIPA-----
2MPT  ---------- -GAMEQSFLP PGWEMRIAPN GRPFFIDHNT KTTTWEDPRL KFPVHMRSK
2N8T  ---------- ---GSSSGLP PGWEEKQDDR GRSYYVDHNS KTTTWSKPTM QD-------
4N7H  ---------- --AMGSGFLP KGWEVRHAPN GRPFFIDHNT KTTTWEDPR- ---------
5DWS  ------GRRA SVELPLGPLP PGWEKRTDSN GRVYFVNHNT RITQWEDPRS QGQ------
5DZD  ------GRRA SVELNEKPLP EGWEMRFTVD GIPYFVDHNR RTTTYIDPRT GKS------
```

Figure 119: Sequence alignment of a selection of Class I WW Domains in complex with a PPxY ligand. The conserved tryptophan residues are shown in red boxes.

8.1.1 hYAP WW Domain Complex

The Yes-associated protein YAP is a transcription cofactor which has two isoforms, YAP1 and YAP2 which differ due to the number of associated WW domains.[400-402] YAP1 has one associated WW domain whereas YAP2 has two.[403, 404] The WW domains are Class I WW domains and so bind PPxY motifs.[396, 405, 406] Such proline-rich motifs occur in many transcription factors.[399, 404, 407-411] Mutations in YAP and its WW domains have therefore been associated with many types of cancer.[401, 403, 412-414]

The proline-rich ligand GTPPPPYTVG was used to determine the first structure of a human YAP1 WW domain complex by NMR.[415] It was determined from this structure that the tyrosine in the ligand formed an important interaction with a hydrophobic pocket in the WW domain but further information was limited by a lack of intermolecular NOEs. Pires *et al*. found the L30K mutant of the WW domain showed improved solubility and stability and so were able to further determine the structure of the complex between the WW domain and GTPPPPYTVG (PDB: 1JMQ).[416] Using NOE values from the NOESY of the complex to determine distance restraints, the NMR structure of the L30K WW Domain was determined using a simulated annealing protocol (Figure 120). The NMR structures of the complex show the tyrosine in the GTPPPPYTVG ligand is bound in a hydrophobic pocket formed by K30, H32 and Q35 of the WW domain (numbering based on the originally isolated wildtype (WT) YAP1 WW domain). The conserved W39 and Y28 found in the WW domain stack with the two proline residues that make up the core PPxY motif in the ligand sequence necessary for binding. T37 is seen to hydrogen-bond with the carbonyl group of P5 in the ligand (the second proline in the core PPxY motif). Studies have shown this threonine residue is very highly conserved in WW domains and can only be replaced by serine without significant reduction of ligand binding.[417] All the residues which contribute most to binding of the ligand are on the β2 and β3 strands of the WW domain and these interactions have subsequently been observed in many other Class I WW domain complexes with PPxY ligands.

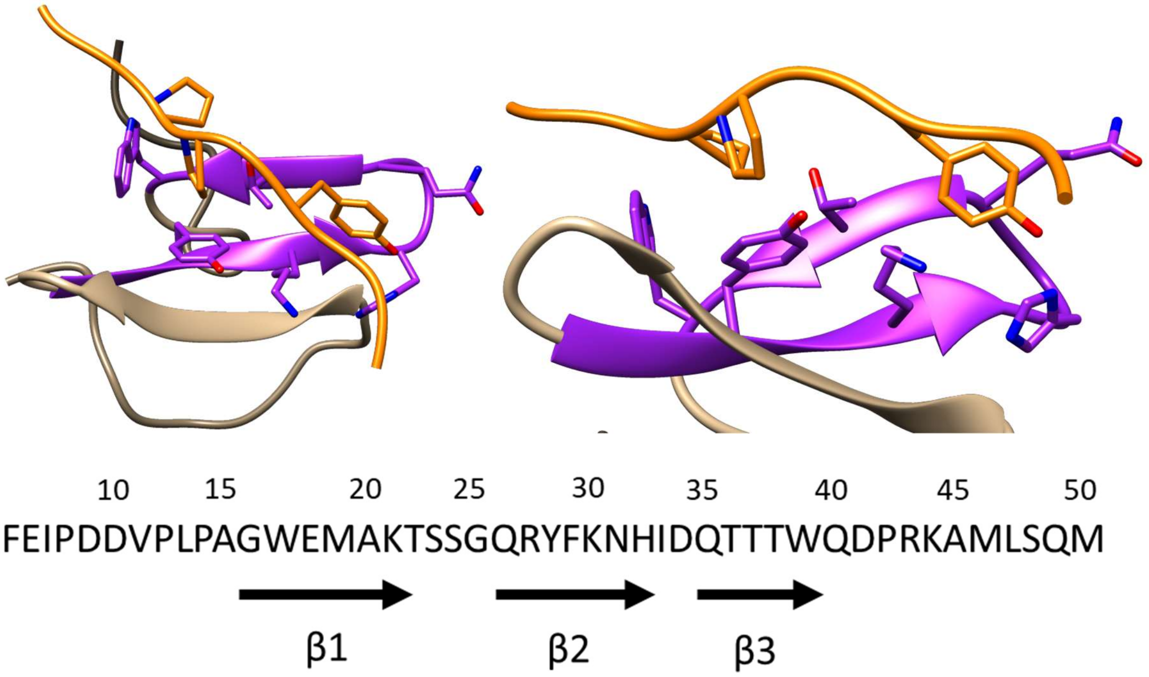


Figure 120: PDB: 1JMQ. GTPPPPYTV (orange) bound to the L30K human YAP1 WW domain (β2 and β3 strands in purple).

Fluorescent titration using the intrinsic fluorescence of the tryptophan in the WT WW domain was used to measure the binding of GTPPPPYTVG. A dissociation constant (K_d) of 52 µM was measured. For the L30K mutant a K_d of 40 µM was obtained. The improved binding of the L30K mutant could potentially be due to the lysine hydrogen-bonding to the C-terminus of the GTPPPPYTVG ligand which is close in proximity, however this is not seen in the NMR structures.

Noticing all the key amino acids for ligand binding are on the β2 and β3 strands Strijowski *et al*. cyclised the β2 and β3 strands of the human YAP1 WW domain through the formation of an amide bond between a *N*-terminal succinyl group and the sidechain of a lysine residue added to the end of the β3 strand sequence (Figure 121).[418] An Ahx linker and cysteine residue were used to allow for the peptide to be immobilised in a plate for binding assays. The cyclised β2 and β3 strands retained approximately 20% of the binding affinity of the WT YAP1 WW domain for the proline-rich ligand Biotin-Ahx-EYPPYPPPPYPSG-NH_2. NMR studies on the cyclic peptide (without the Ahx/Cysteine) show the cyclic peptide has a structure which does not retain the β-sheets seen in the wildtype peptide, with the Y28 and W39 residues important for binding having a large distance between them which is probably the main contributing factor to the reduced binding of the cyclic peptide to the ligand. A linear peptide composed of the same residues used to make the cyclic WW domain mimic showed no binding to the ligand.

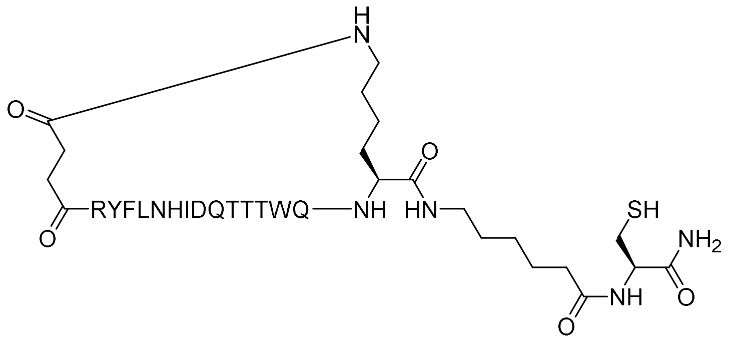


Figure 121: c(succinyl-RYFLNHIDQTTTWQ-Lys)-Ahx-Cys-NH$_2$. Cyclic WW domain mimic by Strijowski et al.[418]

The succinyl group and lysine used to cyclise the β2 and β3 strands of the WW domain forms a relatively flexible linker so may be one of the reasons the cyclic WW domain mimic designed by Strijowski *et al*. showed limited binding. The BTM has been shown to induce a β-turn structure so could potentially be a better way of cyclising the β2 and β3 strands of the YAP1 WW domain whilst retaining the β-sheet structure, allowing for higher affinity binding.

The L30K WW domain showed improved solubility and stability than the WT WW domain. The GTPPPPYTVG is a short ligand containing the key PPxY motif that mimics the poly-proline helix type structure seen in many WW domain ligands thus acting as a good model for more complex ligands and has available NMR structures of its interaction with the L30K YAP1 WW domain. The key interactions for ligand binding occur between amino acids which are on the β2 and β3 strands of the WW domain. The NMR structure of the complex was therefore used to design a cyclic WW Domain mimic to determine if cyclising the β2 and β3 strands with the BTM could retain the necessary β-sheet structure to bind the GTPPPPYTVG ligand.

8.2 Design

The PDB structure 1JMQ was used to design a cyclic WW domain mimic, c(β2β3-BTM), which could bind ligands containing the PPxY motif. As all the key residues for ligand binding are found on β-strands two and three of the WW domain it was envisioned that the two strands could be cyclised in order to retain the β-strand structure for binding without the rest of the protein (Figure 122). The BTM has been shown to allow for the cyclisation of peptides whilst restricting the conformational freedom of the peptide and inducing structure. c(β2β3-BTM) could therefore have the potential for protein-protein interactions allowing the principle to be applied to other WW domain systems.

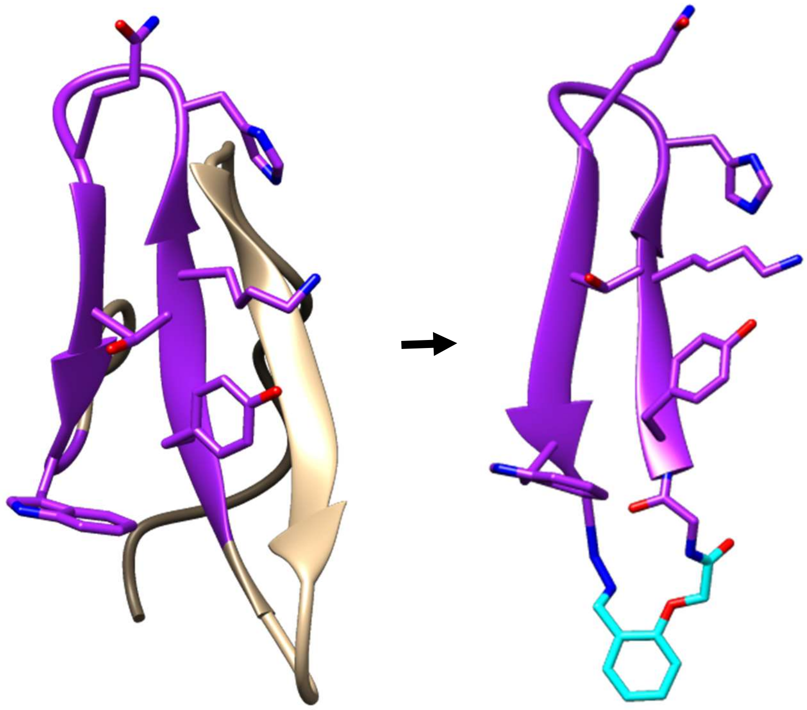


Figure 122: Cyclised WW domain design. The β2 and β3 strands of the WW domain to be included in the cyclic peptide design are shown in purple. The BTM used to cyclise the two strands is shown in light blue.

8.2.1 Alanine Scan

An alanine scan can be used to measure which residues are most important for the binding between a protein host and a ligand.[419, 420] Each residue in the protein is alternatively swapped for alanine and the difference in free energy of the interaction of the original residue compared to the interaction of alanine and the ligand measured. Alanine retains very similar Ramachandran preferences to the other chiral amino acids so should hopefully not alter the structure significantly and has only a small sidechain without chemically reactive functional groups. Residues with larger ΔΔG values contribute more to binding of the ligand and should therefore be conserved in the design of the cyclic WW domain mimic. Making a protein with each residue of interest replaced with an alanine and measuring the effects is a very time consuming process, especially if every single amino acid were to be mutated. Therefore computational methods for computational alanine-scanning mutagenesis have been developed.[421-426]

The BAlaS webserver[427] was used to estimate the free energy associated with each residue in the WW domain for binding to GTPPPPYTVG (see section 10.3). An alanine scan for the 20 NMR structures available for 1JMQ was performed (Figure 123). The only residue on the β1 strand which could potentially be contributing to ligand binding is T22, however compared to the residues on the β2 and β3 strands involved in binding it has a small ΔΔG value, so a cyclic WW domain based on just the β2 and β3 strands could still bind the GTPPPPYTVG ligand. Y28 is close to the beginning of β2 and W39 is at the end of β3 (Figure 124) so must be included in the cyclic WW domain design. Q26 could also be contributing to ligand binding but is omitted from the cyclic WW domain mimic as the β3 strand is shorter than the β2 strand and the WW is kept as small as possible whilst still retaining all the key residues for binding.

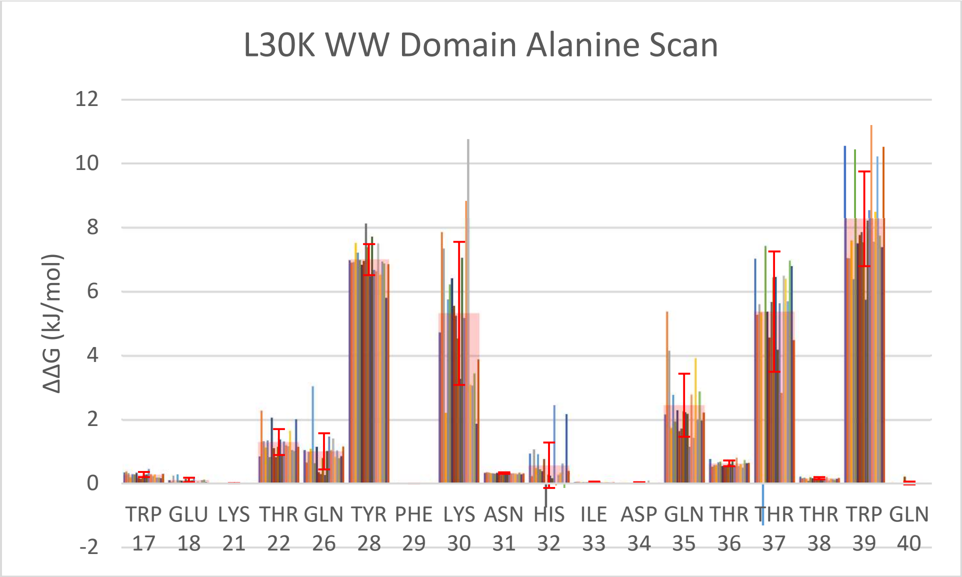


Figure 123: Alanine scan of 1JMQ. Calculated ΔΔG value for each of the 20 NMR structures are overlayed with the average values with standard deviation shown in red.

The WW domain subsequence RYFKNHIDQTTTWQ was selected to be incorporated into a cyclic WW domain mimic. It contains all the key residues for binding of the proline-rich ligand whilst incorporating as few residues from the protein as possible.

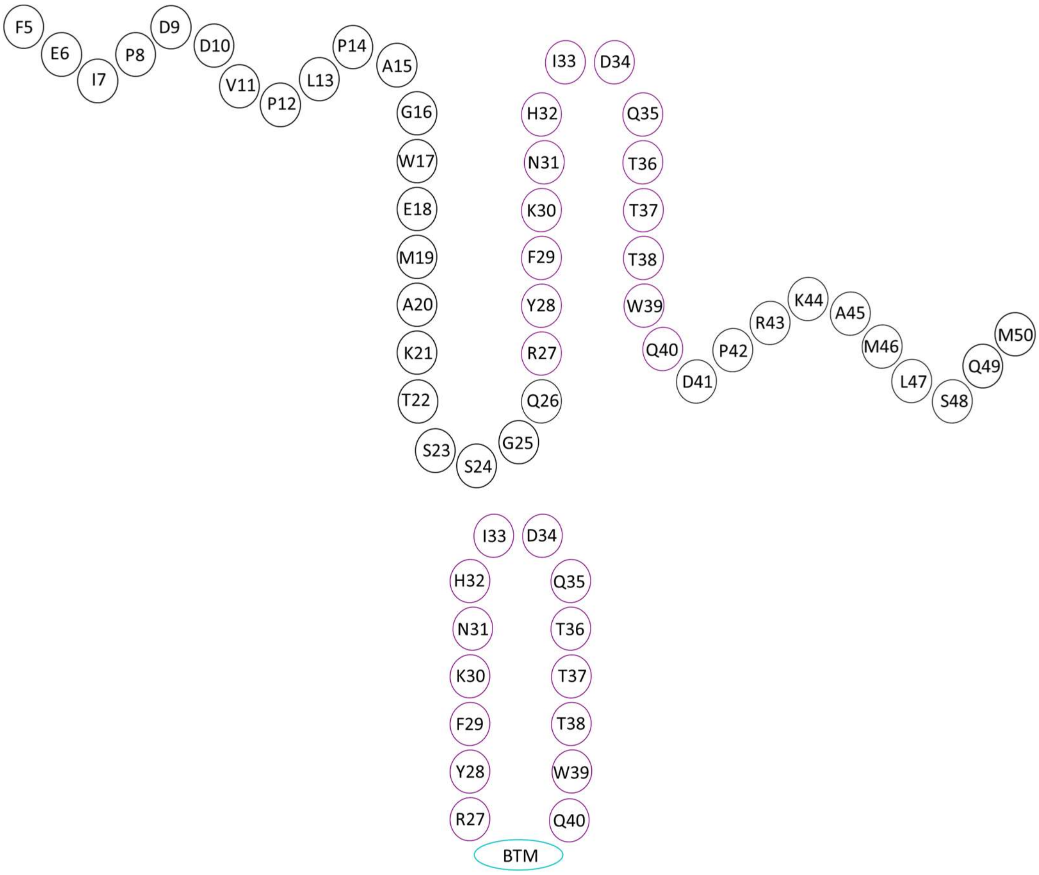


Figure 124: L30 K YAP1 WW Domain (above) and c(β2β3-BTM) (below). Residue numbering is based on the originally isolated YAP1 WW domain. The β2 and β3 strands of the WW domain included in the cyclic mimic are shown in purple with the BTM used to cyclise the peptide in blue.

An alanine scan on the NMR structures of the L30K YAP1 WW domain where each residue in the ligand is substituted by alanine confirms the importance of the PPxY motif for binding (Figure 125). The Tyrosine has the largest ΔΔG value.

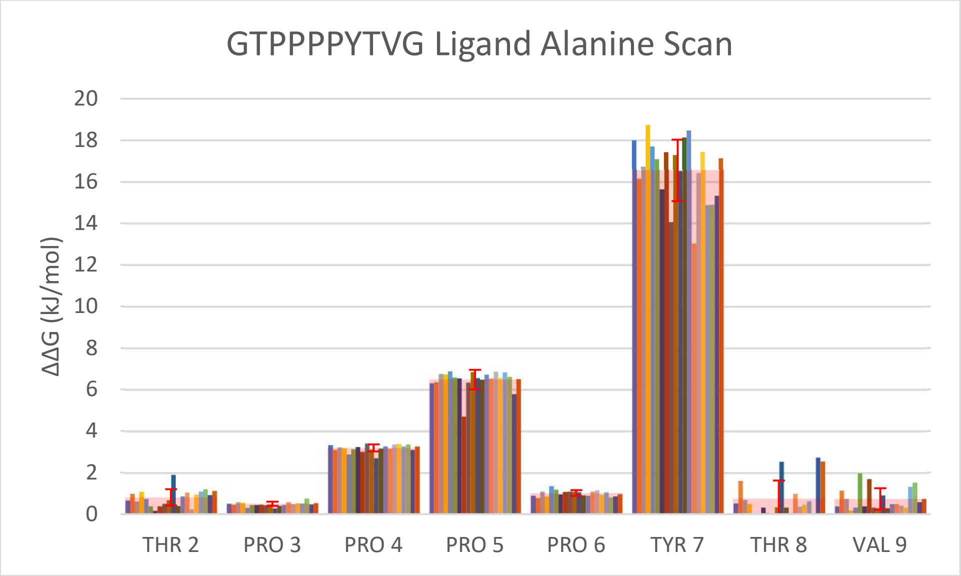


Figure 125: Alanine scan of GTPPPPYTVG in the 20 NMR structures for 1JMQ. Individual values are shown as well as the average values in red with the standard deviation.

8.3 c(β2β3-BTM)

8.3.1 MD

A 500 ns MD simulation was used to determine the conformation of the cyclic WW domain mimic (c(β2β3-BTM)) in solution. One major conformation was found. The structure was overlayed with the L30K YAP1 WW domain NMR structures to determine if the mimic retained the key features of the WW domain necessary for binding. The WW domain has an average backbone RMSD of 2.8 Å across the 14 amino acids in c(β2β3-BTM). The RMSD just across the six key residues for interaction with the ligand – Y28, K30, H32, Q35, T37 and W39 – is around 3.1 Å. The β-sheet structure of the β2 and β3 strands is retained so it is likely the c(β2β3-BTM) would be able to bind the GTPPPPYTVG ligand as the conformation it is predicted to adopt by the MD simulation is similar to the bound structure (Figure 126).

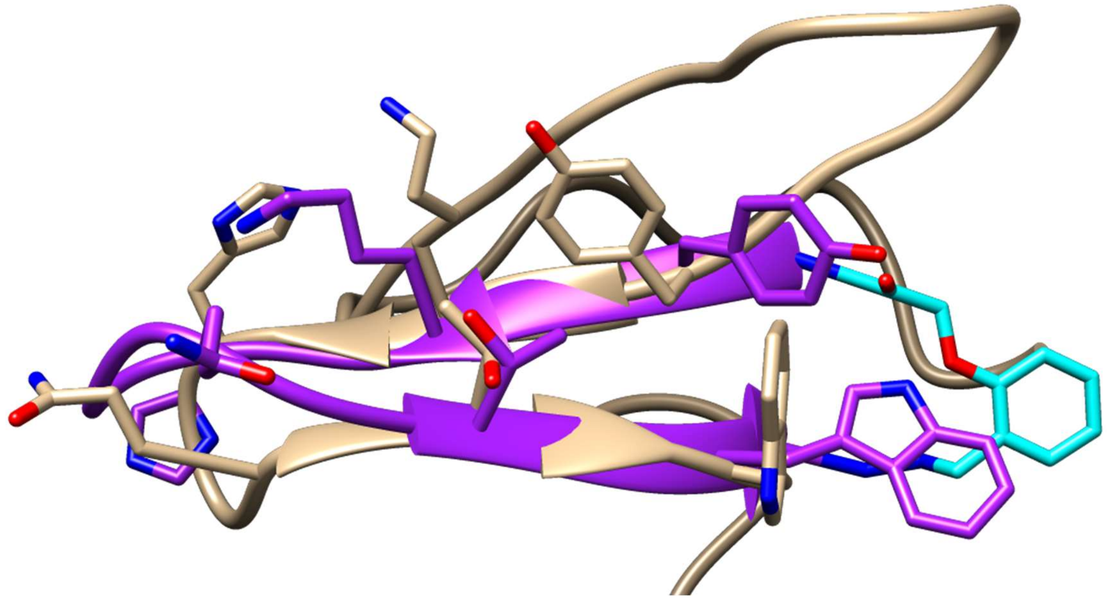


Figure 126: The overlay of 1JMQ (beige) and c(β2β3-BTM) (purple with the BTM shown in light blue).

The largest difference in the structures is at the β-turn region at the other end of the peptide from the β-turn mimic. In the L30K YAP1 WW domain the two β-strands are linked by a type I β-turn made up of the dipeptide sequence Ile-Asp. In the c(β2β3-BTM) conformation the turn region twists and the type I turn centred on ID is not seen (Figure 127). This means H32 (and Q35) which forms part of the hydrophobic pocket which binds the tyrosine of the PPxY motif have slightly different positions.

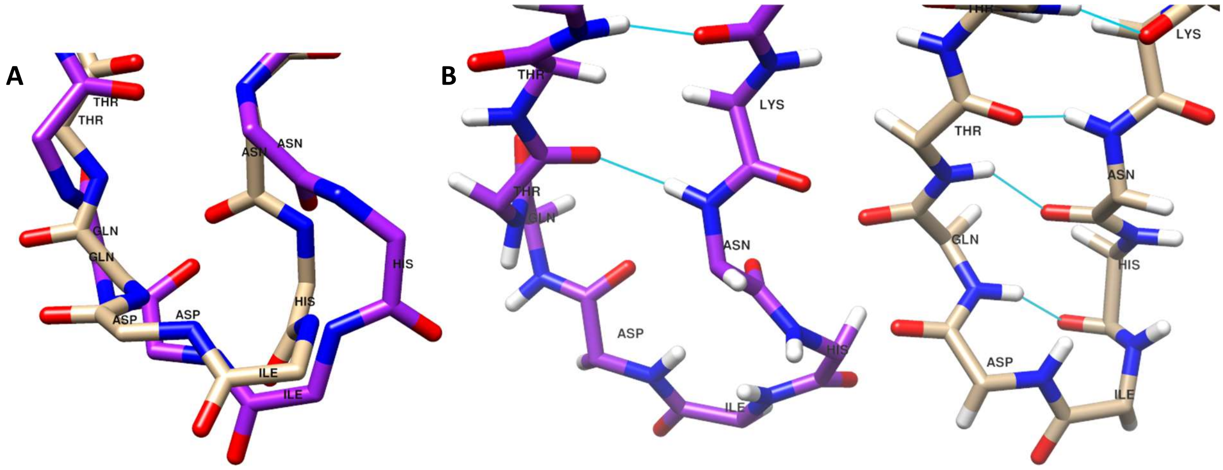


Figure 127: Peptide backbone of the turn region of L30K YAP1 WW domain (brown) and c(β2β3-BTM) (purple). The overlay (A) and the two turn regions with intramolecular hydrogen-bonds shown in blue (B).

8.3.2 MD on the Bound Conformation of c(β2β3-BTM) and Alanine Scan

The NMR structure of the L30K WW domain was used as a model to build the structure of c(β2β3-BTM) in complex with the GTPPPPYTVG ligand. A 10 ns MD simulation was used to equilibrate the structure. During the 10 ns simulation the ligand remained in a complex with c(β2β3-BTM) despite no distance restraints being used. The backbone atom RMSD of the c(β2β3-BTM) complex and the NMR structures of 1JMQ is around 2.9 Å across all atoms in both the ligand and the 14 residues the c(β2β3-BTM) and L30K WW domain have in common. In both complexes the GTPPPPYTVG ligand binds diagonally across the two β2 and β3 sheets. The proline residues of the ligands core PPxY motif interact with Y28 and W39 and the key tyrosine residue interacts with K30, H32 and Q35 (Figure 128).

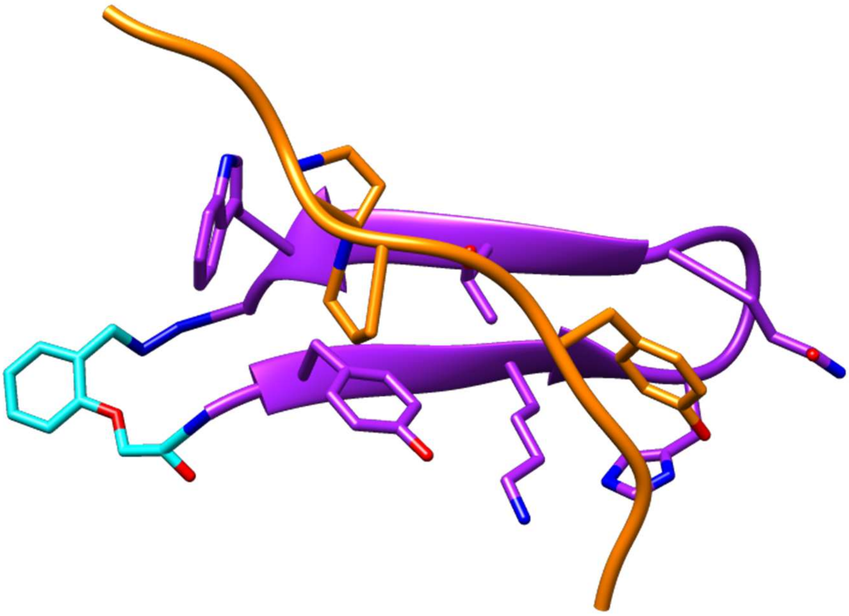


Figure 128: c(β2β3-BTM) bound to GTPPPPYTVG (orange).

Randomly selected conformations from the last 5 ns of the simulation were taken and used for an alanine scan (Figure 129). The results of the alanine scan show similar results to that of the alanine scan of the L30K NMR structures. Y28, K30, Q35, T37 and W39 are most important for ligand binding. H32 possibly contributes more to binding in c(β2β3-BTM) with a slightly higher ΔΔG seen in the alanine scan. The c(β2β3-BTM) therefore has potential to bind PPxY ligands. Similarly an alanine scan of the ligand in the structure shows it is still the two proline residues and the tyrosine in the core PPxY motif that contribute the most to binding of the ligand. The cyclic WW domain could therefore potentially bind to other ligands containing this motif. The tyrosine in the PPxY core motif in the ligand contributed most to binding in the WW domain with a ΔΔG of approximately 16 kJ/mol. A slightly lower value is seen in the alanine scan of c(β2β3-BTM) so it may not bind as strongly.

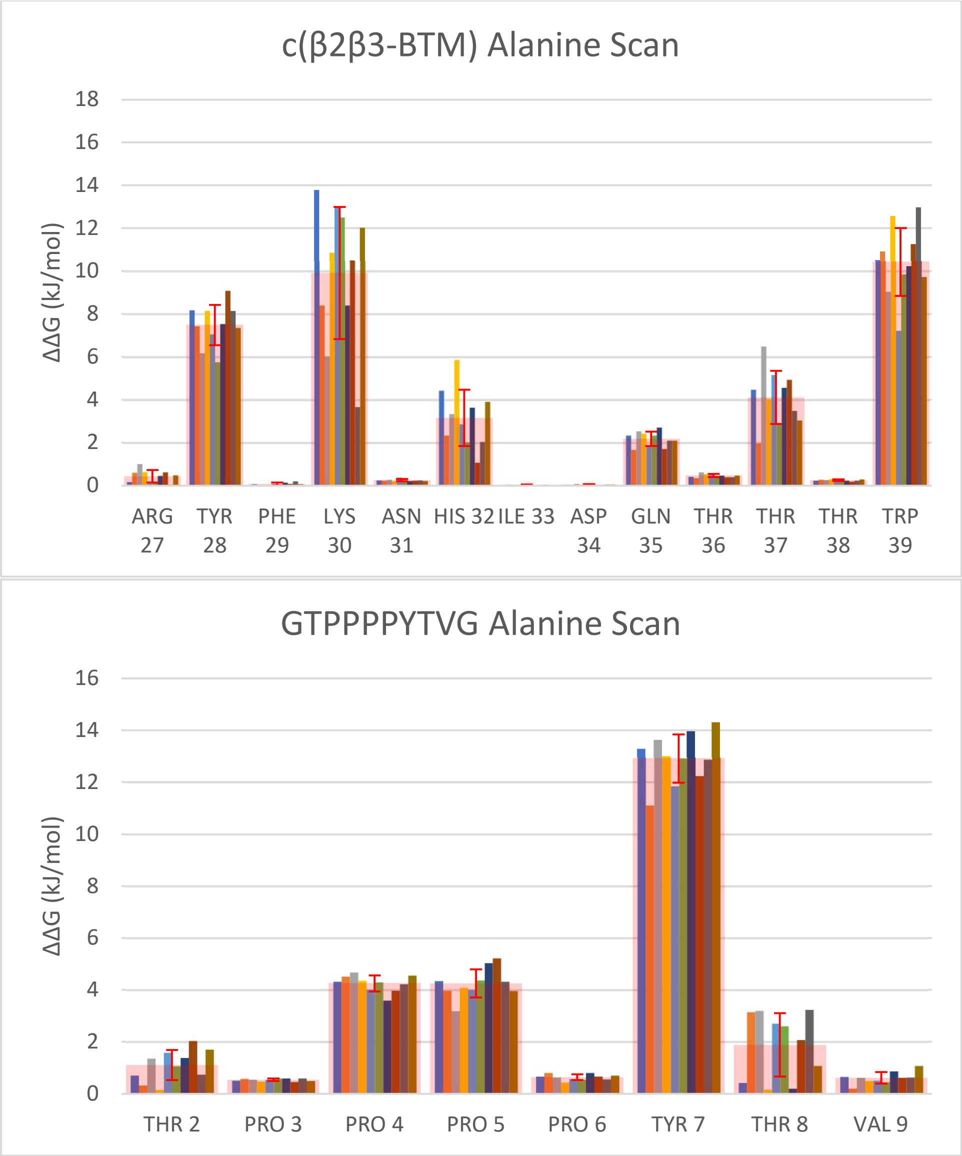


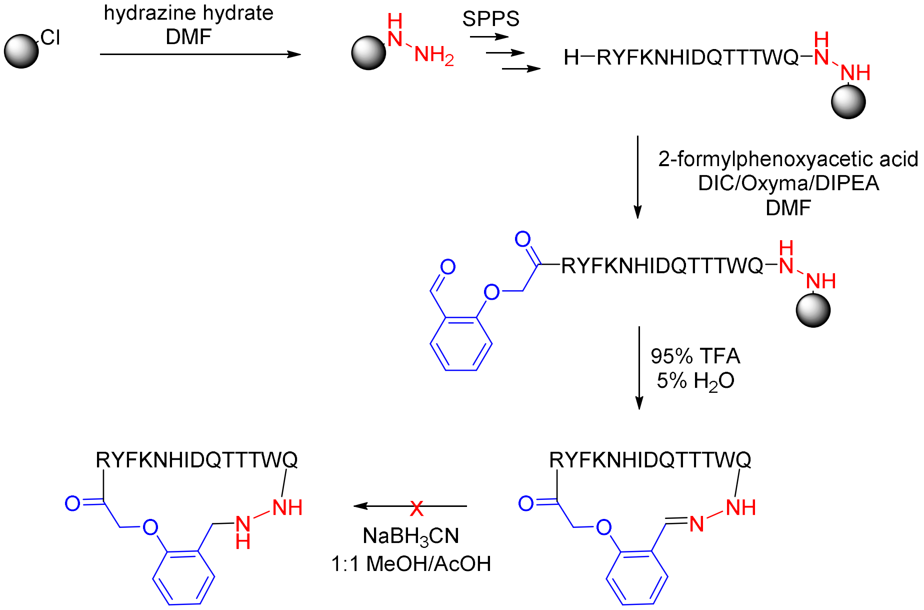
Figure 129: Alanine Scans for the c(β2β3-BTM) in complex with the ligand.

A backbone atom RMSD between the c(β2β3-BTM) conformation obtained in the 500 ns simulation and the c(β2β3-BTM) /GTPPPPYTVG complex of approximately 1.4 Å is seen. The predicted conformation of c(β2β3-BTM) in solution is therefore very similar to the predicted conformation of the bound structure.

To summarise the β2 and β3 strands of the YAP1 WW domain contain the majority of residues necessary for binding of PPxY ligands. A cyclic peptide was therefore designed using the BTM to cyclise the two strands. MD simulations and alanine scans show the c(β2β3-BTM) is likely to retain the β-strand structure and the ability to bind the GTPPPPYTVG ligand in solution.

8.3.3 Synthesis

As the peptide contains 14 residues the two-step method for synthesising cyclic peptides containing the BTM was used as this has been shown to give higher yields for larger cyclic peptides. The hydrazine resin was synthesised and the residues coupled by SPPS, followed by coupling of 2-formylphenoxyacetic acid. The peptide was then cleaved from the resin using 95% TFA, 5% water for 3 hours. The peptide would then be dissolved in 1:1 methanol/acetic acid and reduced using ten equivalents of sodium cyanoborohydride (Scheme 30). However after coupling of the 2-formylphenoxyacetic acid the peptide was very insoluble and could not be redissolved in order to reduce the hydrazone to the hydrazine.



Scheme 30: Synthesis of c(β2β3-BTM).

8.4 Version 2 – Improved Solubility

In order to improve the solubility of the WW domain some amino acids were substituted. Amino acids not involved in the binding of the ligand can be substituted as long as they do not drastically alter the structure of the peptide. Four amino acids were substituted:

Version 1:	R	Y	F	K	N	H	I	D	Q	T	T	T	W	Q
Version 2:	R	Y	K	K	N	H	P	N	Q	T	T	E	W	Q

The Ile-Asp subsequence that occupies the i+1 and i+2 positions of the type I β-turn between the β2 and β3 strands was replaced by PN. The greatest difference in RMSD between the c(β2β3-BTM) and L30K YAP1 WW domain was due to the Ile-Asp β-turn changing conformation so by replacing it with proline it is hoped this turn region will now better match the L30K YAP1 WW domain structure. Proline was chosen as an alternative as it commonly occupies the i+1 position of β-turns, which are also often type I turns when a chiral amino acid is at the i+2 position. Isoleucine is a bulky hydrophobic amino acid so it is hoped improved solubility would be seen without its inclusion in the sequence. Replacing aspartic acid with asparagine removes a negative charge so the overall charge of the peptide is now more positive hopefully also improving solubility.

Additionally the phenylalanine was replaced with a lysine and the final threonine (T38) was replaced with a glutamic acid. These are opposite each other on the non-binding side of the β-sheet structure of the peptide. The lysine and glutamic acid sidechains are therefore in close proximity and could potentially form a salt-bridge. Such salt-bridges have been observed to help stabilise β-hairpin structures.[428-431]

The phenylalanine is highly conserved in WW domains but is on the opposite side of the β-sheet to ligand binding. It has been shown to be important for forming the three-stranded β-sheet structure of the WW domain, interacting in a hydrophobic interaction with the first tryptophan.[390] In the cyclic version without this first tryptophan it is therefore not necessary to keep this phenylalanine despite it being highly conserved in WW domains. The final threonine which was replaced by a glutamic acid is often conserved in WW domains but sometimes replaced by a glutamine. Glutamic acid and glutamine are very similar so replacing this threonine with glutamic acid is unlikely to drastically alter the structure of the peptide.

8.4.1 MD

A 500 ns MD simulation was used to determine the conformation of version 2 of the cyclic WW domain. The results of the simulation were clustered and one major conformation was seen. The cluster centre was used as representative of the conformation. Figure 130 shows the overlay of the conformation of the cyclic WW domain mimic with the β2 and β3 strands of the NMR structure of the L30K WW domain. Backbone RMSD values of approximately 1.6 Å are seen for each available NMR structure. The sidechains are therefore in approximately the right position for binding to the PPxY ligand. For just the Y28, K30, H32, Q35, T37 and W39 sidechains (those most important for binding) a RMSD of around 1.2 Å is seen. These are lower RMDS values than seen for version 1 of the c(β2β3-BTM). The PN turn forms a type I β-turn similar to the β-turn seen in the L30K WW domain structure. Substitution of the residues in the β-turn which are not involved in ligand binding therefore allowed for improved design of the c(β2β3-BTM).

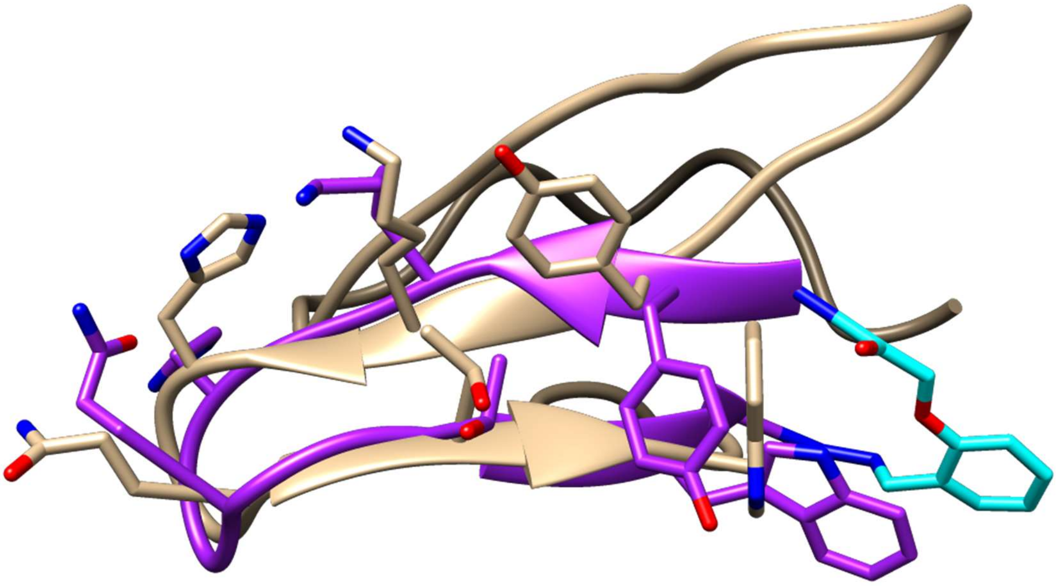


Figure 130: Overlay of the cyclic WW domain mimic (purple with the BTM shown in blue) with the β2 and β3 strands of the L30K WW Domain.

W39 is in close proximity to the BTM so could potentially interact with it through π-stacking. This presents an alternative interaction which could potentially lower the binding affinity of the cyclic WW domain mimic.

8.4.2 MD of the c(β2β3-BTM) – GTPPPPYTVG Complex and Alanine Scan

A model of the c(β2β3-BTM) version 2 in complex with the GTPPPPYTVG ligand was built based on the PDB structure of the L30K YAP1 WW domain complex. A short MD simulation of 10 ns was used to minimise the bound structure. Similar to the L30K YAP1 WW domain, the ligand remains bound to the structure diagonally (Figure 131). A backbone atom RMSD between the bound structures of the L30K YAP1 WW domain and c(β2β3-BTM) of around 1.2 Å is seen. The simulation was extended to

100 ns and the same conformation was seen, so although short the 10 ns simulation seems sufficient to find the predicted conformation of the complex. Although based on the NMR structure of the L30K YAP1 WW domain it is however still possible that the ligand could potentially bind in other orientations.

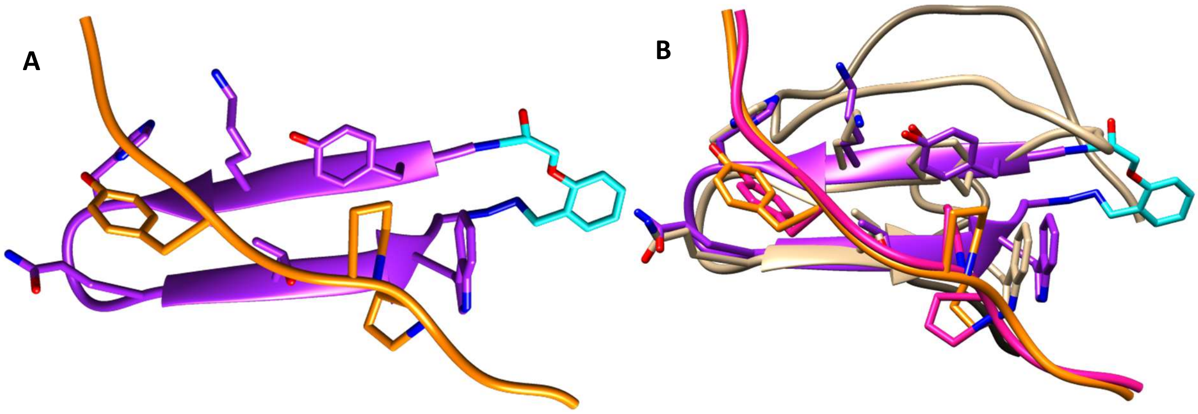


Figure 131: A: GTPPPPYTVG ligand (orange) bound to c(β2β3-BTM) version2. B: L30K YAP1 WW Domain (brown) bound to GTPPPPYTVG (pink) overlayed with c(β2β3-BTM) version 2 (purple with BTM in cyan) bound to GTPPPPYTVG (orange).

Randomly selected conformations from the last 5 ns of the 10 ns simulation were used in an alanine scan (Figure 132). Similar results are seen to the alanine scans of the L30K YAP1 WW Domain complex for both c(β2β3-BTM) and the GTPPPPYTVG ligand.

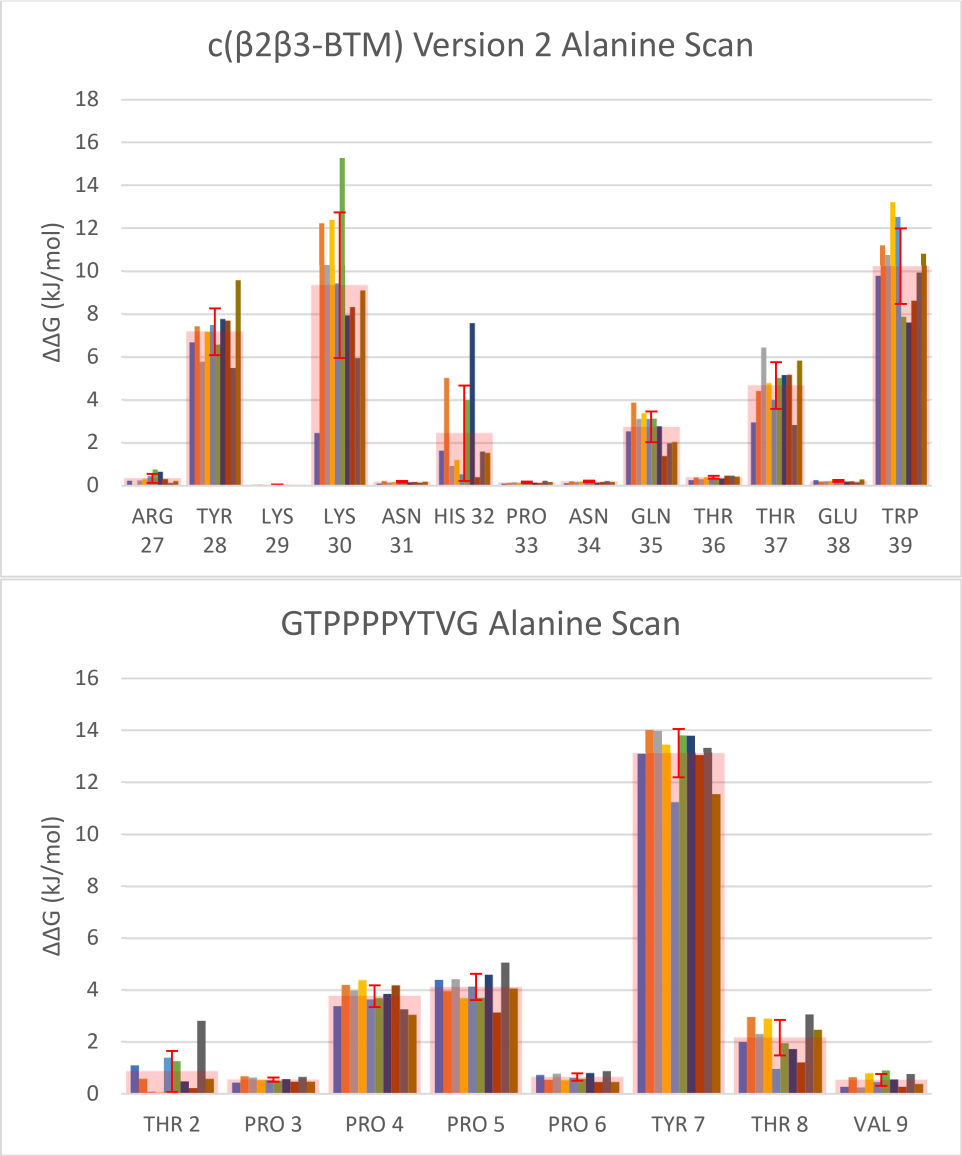


Figure 132: Alanine Scans of c(β2β3-BTM) version 2 and the ligand in a complex. Average values with the standard deviations are shown in red overlayed over the individual conformations.

The 500 ns MD simulation on c(β2β3-BTM) predicted the conformation the peptide is likely to form in solution when not bound to a ligand. The structure obtained is relatively similar to the β2 and β3 strands in the YAP1 L30K WW domain NMR structures. The predicted c(β2β3-BTM) solution conformation obtained during the 500 ns simulation was therefore overlayed with the YAP1 L30K WW domain in the NMR structure. The WW domain was then deleted to produce a model of the c(β2β3-BTM) in complex with the GTPPPPYTVG ligand. Scwrl4[432] was used to repack the sidechains in the structure to remove any steric clashes that may be present and help orientate the sidechains into the correct position for binding. A 100 ns MD simulation of the complex was then carried out to determine if the same peptide structure in complex with the ligand is seen when the model is built from a different starting conformation. As the starting conformation represents the predicted structure of the peptide in solution before binding to the ligand, if the same bound conformation is seen this supports the binding model previously seen based on the NMR structure.

The same bound conformation of c(β2β3-BTM) is seen in both 100 ns simulations starting from the different conformations based on either the bound NMR structure of the YAP1 L30K WW domain or based on the predicted solution structure of c(β2β3-BTM) from the 500 ns simulation. The c(β2β3-BTM) could therefore potentially bind the GTPPPPYTVG ligand in this conformation. Both structures were based on having the ligand in the same initial orientation, diagonally across the two β-sheets. It is therefore possible the ligand can also bind in other orientations. However since there was a strong enough interaction between c(β2β3-BTM) and the PPxY ligand for the ligand to remain in the complex for the full 100 ns simulations despite no distance restraints being used and both initial structures minimise to the same conformation it is likely the ligand is able to bind to the c(β2β3-BTM) in this orientation.

The predicted unbound conformation of c(β2β3-BTM) is also able to easily rearrange to the preferred orientation for interaction with the ligand, with convergence between the two simulations from different starting conformations seen within 10 ns. The backbone atom RMSD between the conformation of c(β2β3-BTM) obtained from the 500 ns MD simulation and the conformation seen in the bound complex is 1.4 Å. The conformation of the c(β2β3-BTM) is therefore relatively similar to the possible binding conformation.

8.4.3 Synthesis

Version 2 of the c(β2β3-BTM) was synthesised by the same method as the first version up to the reduction of the hydrazone. The new WW domain sequence showed improved solubility and was therefore able to be redissolved in 5 mL 1:1 methanol/acetic acid after cleavage from the resin and reduced to the final product using 10 equivalents of sodium cyanoborohydride. Prior to reduction the peptide was only observed in the cyclised form. The product was isolated and purified by RP-HPLC.

In order to determine if the c(β2β3-BTM) was able to bind a PPxY ligand the peptide GTPPPPYTVG was also synthesised and purified. The L30K YAP1 WW domain was also synthesised to compare any results obtained for the cyclised version. There are several possible constructs of the WT YAP1 WW domain which often vary at the C-terminus. The C-terminus residues, after the β3 strand, form an unstructured region of the protein. A shorter construct of the L30K YAP1 WW domain was therefore synthesised than originally used by Pires *et al*. to determine the binding of the GTPPPPYTVG ligand. This shorter construct means the sequence contains two fewer methionine residues which were making purification of the peptide difficult as they were being oxidised in air but had a very similar retention time.

8.5 Fluorescence Titration

Changes in the intrinsic fluorescence of tryptophan residues in a protein can be measured with titration of a ligand. A binding curve can therefore be measured and the dissociation constant (K_d) calculated. Fluorescent titration was carried out to determine if c(β2β3-BTM) retains the ability of the L30K YAP1 WW domain to bind the GTPPPPYTVG ligand.

Fluorescent titration was carried out using 10 µM of the L30K YAP1 WW domain or c(β2β3-BTM). A 5 mM solution containing the GTPPPPYTVG ligand was titrated into the WW domain solution. Both solutions were made up using potassium phosphate buffer (pH 7.4) and 100 mM potassium chloride, the same conditions used by Pires *et al*.[416] An excitation wavelength of 298 nm was used and the change in fluorescence upon addition of the ligand measured at 340 nm. A repeat of the fluorescent titration gave the same results.

8.5.1 L30K YAP1 WW Domain

Fluorescence titration of the L30K YAP1 WW domain gave an initial increase in fluorescence as the GTPPPPYTVG ligand was added (Figure 133). A similar fluorescence increase is seen to what was reported by Pires *et al*.[416] The K_d value for this initial part of the curve is around 52 ± 15 µM similar to the reported value of 40 µM.

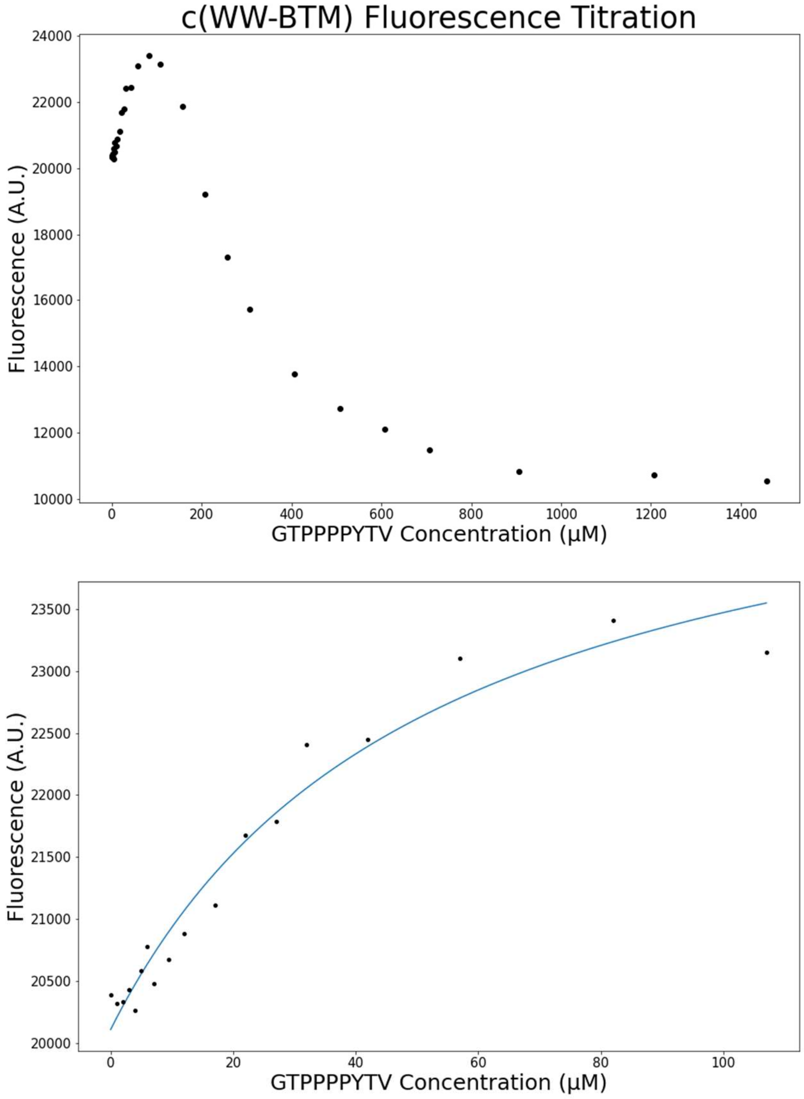


Figure 133: L30K YAP1 WW Domain fluorescent titration. Full titration curve (above) and the initial increase in fluorescence used to measure the K_d value (below).

After this initial increase in fluorescence the fluorescence decreases rapidly as further ligand is titrated in. This is not mentioned by Pires *et al*. regarding their fluorescence titrations on the L30K YAP1 WW domain. The WW domains differ slightly at the C-terminus, but as the C-terminus is a very flexible region and is not involved in binding, the different WW domain constructs should behave the same. If multiple binding events are occurring with opposite effects on fluorescence this could be why an initial increase in fluorescence is seen followed by a decrease. Alternatively HPLC traces of the WW Domain show one peak at lower concentrations but more peaks forming at higher concentrations, so it is potentially forming other interactions or changing conformation at higher concentrations of ligand.

Changes in the emission spectra are seen upon addition of the GTPPPPYTVG ligand (Figure 134). Similar to what is reported by Pires *et al*. the emission maxima shifts from around 345 to 340 nm as the ligand is added during the initial stage where fluorescence increases.

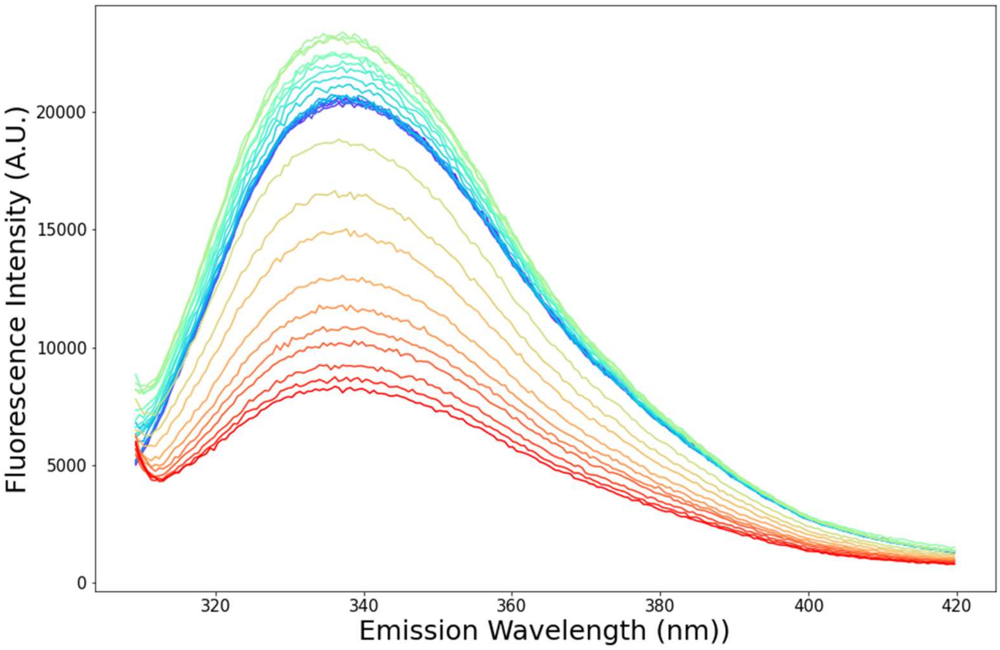


Figure 134: Change in the emission spectrum of the L30K YAP1 WW Domain upon addition of the GTPPPPYTV ligand (plot changes from blue to red upon addition of the ligand).

8.5.2 c(β2β3-BTM) Fluorescence Titration

The cyclic WW domain shows a lower total fluorescence than the full WW domain. The sequence is much shorter and does not have the first conserved tryptophan residue (or a phenylalanine), so a lower fluorescence is not unexpected. Whereas for the L30K YAP1 WW domain an initial increase in fluorescence is seen followed by a larger decrease in fluorescence, for the c(β2β3-BTM) there is no initial increase in fluorescence. The fluorescence decreases as the GTPPPPYTVG is titrated into the c(β2β3-BTM) containing solution (Figure 135). As a ligand-dependant change in fluorescence is seen it is therefore likely the cyclic WW domain retains the ability to bind to the ligand.

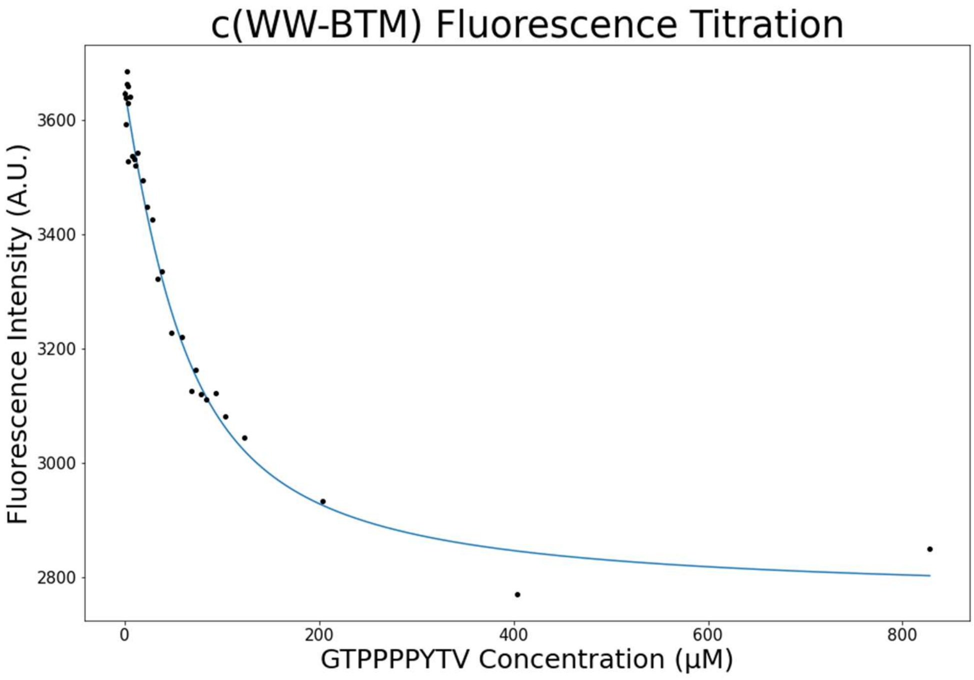


Figure 135: Fluorescence titration of c(β2β3-BTM).

Non-linear least square fitting was used to fit the binding curve based on the equation below where I is the fluorescence intensity, I^0 is the initial fluorescence, I^∞ is the fluorescence at binding saturation, [L] is the concentration of the GTPPPPYTVG ligand and K_d is the dissociation constant. A K_d value of 66 µM ± 7 was calculated. This is similar to the value of 52 µM calculated for the full WW domain. It is therefore likely, as the MD simulations indicate, the c(β2β3-BTM) retains the β-hairpin structure with the amino acids necessary for binding held in similar positions to the L30K WW domain.

$$I = I^0 + \frac{I^\infty \times [L]}{[L] + K_d} \qquad (12)$$

The Alanine scan on the L30K YAP1 WW Domain showed that although the residues which contributed most to ligand binding were on the β2 and β3 strands of the WW domain, T22 on β1 and Q26 at the start of β2 also potentially contributed to binding of the ligand. These two residues were not included in the c(β2β3-BTM) so are potentially one reason why a slightly higher K_d value is seen.

Changes in the emission spectrum are seen upon addition of the ligand (Figure 136). Similar to the L30K YAP1 WW domain, the maxima of the peak shifts from around 345 to 340 nm as the ligand is added. A blue shift in the emission maxima of tryptophan during fluorescent titrations is generally seen due to the tryptophan being in a more hydrophobic environment after binding.[377, 378, 433] The change in emission maxima seen during the titration is therefore consistent with ligand binding.

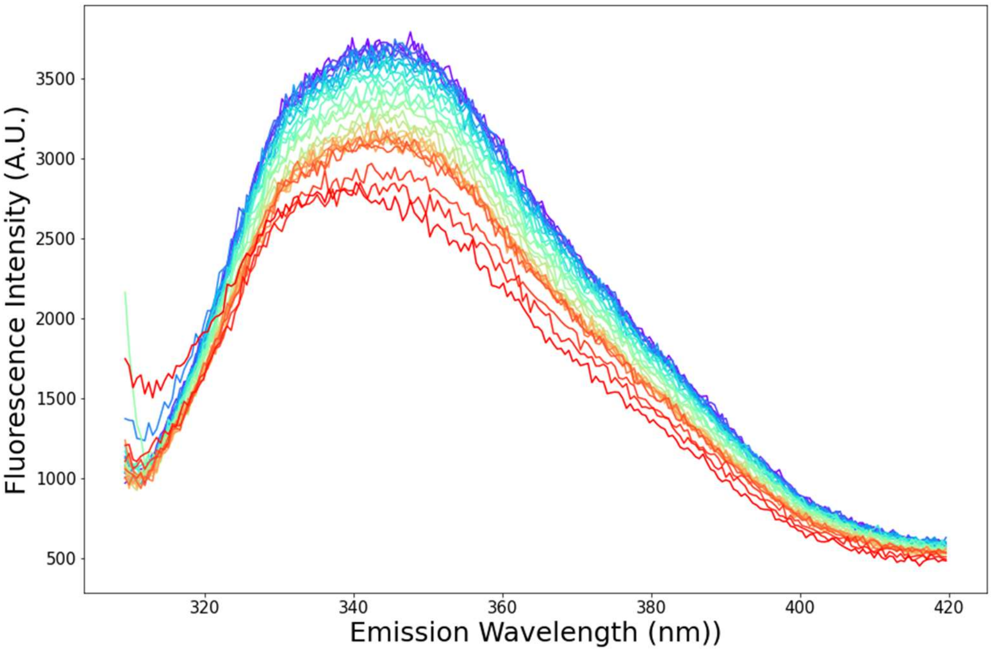


Figure 136: Emission spectrum of c(β2β3-BTM) upon addition of GTPPPPYTV ligand. Spectrum changes from blue to red upon addition of the ligand.

8.6 Conclusions

A cyclic peptide was designed to incorporate the key residues from the YAP1 L30K WW domain for binding of a small PPxY ligand. Previous work from Chapter 6 on incorporation of the BTM into cyclic peptides allowed for efficient synthesis of the c(β2β3-BTM). Version 1 was very insoluble and the β-turn joining the two β-strands was predicted to form a different turn structure than seen in the native WW domain. The Random Forest (Chapter 5) and Loop Database analysis (Chapter 3) predicted the PN subsequence would form a type I turn. As a type I turn joins the β-hairpin structure in the native YAP1 WW domain, replacing the ID turn subsequence allowed for improved structure design as well as solubility.

Initial studies on version 2 of c(β2β3-BTM) show it interacting with the proline-rich ligand GTPPPPYTVG in both MD simulations and during fluorescence titration. Using the BTM to cyclise the β2 and β3 strands was able to retain the β-hairpin structure of the WW domain and therefore is a potential method for the design of cyclic peptides to target PPIs.

CD could be used to confirm the predicted β-sheet structure of the c(β2β3-BTM) as could NMR assignment. The NMR structure could be determined for the c(β2β3-BTM) in complex with the ligand to determine if it binds the proline-rich ligand in the same orientation as seen in native WW domains. Changes in the CD and in NMR shift upon addition of the ligand could also be used to confirm binding. Pires et. al. observed large changes in the chemical shift of the indole NH of W39 upon addition of proline-rich ligands to the YAP1 WW domain.[416] It is therefore likely a similar large change in shift would be observed for the cyclic WW Domain mimic. Further binding experiments such as isothermal titration calorimetry (ITC) could also confirm binding of the ligand to the cyclic WW domain mimic.

As WW domains are involved in many PPIs they are associated with numerous diseases such as cancer and Alzheimer's.[393] The c(β2β3-BTM) was based on the YAP1 WW domain but similar binding mechanisms are seen for the other Class I WW domains. The same design principles could therefore be applied to other WW domains. Future versions of the BTM could be used to incorporate other functionalities to the WW domain. Fluorescence could be incorporated into the WW domain, similar to the naphthalene based BTM discussed in chapter 7, in order to monitor the pathways of disease-associated WW domains. Further ligation points could also be incorporated into the BTM to be used to join other functional groups or proteins to the WW domain. Alternatively non-binding residues in the c(β2β3-BTM) could potentially be replaced with a cysteine residue which could also be used as a ligation point. The PPIs associated with the WW domain could then be targeted through binding of the c(β2β3-BTM).

9 Conclusions

In this thesis the prediction of cyclic hexapeptide structure has been discussed with the aim of identifying sequences which lead to well-structured cyclic peptides. The use of a β-turn mimic (BTM) to design peptides with a single major conformation was also explored.

Bias-exchange metadynamics (BE-META) simulations are a useful method for predicting the conformation of cyclic peptides. Current implementations of BE-META however have not explored the cis/trans isomerisation of proline. As the inclusion of proline within a cyclic peptide can lead to reduced conformational freedom of the peptide and can improve synthesis, Chapter 2 explored the addition of an extra replica in the BE-META to allow for cis/trans proline isomerisation. A series of four proline-containing peptides were synthesised and had their structures determined by NMR. The forcefield used throughout the simulations could not always accurately model the energy difference between the cis and trans proline states. Small energy differences between the two states for some of the peptides means a highly accurate forcefield is needed to model the peptides or large differences in predicted cis to trans proline ratios will be seen. Addition of the proline replica in the BE-META however allowed for faster convergence of the simulations so can still be a beneficial addition when BE-META is used to predict the structures of cyclic peptides.

Chapter 3 introduced the Loop Database. As cyclic hexapeptides commonly form a structure in solution made up of two overlapping β-turns, β-turns were extracted from the Loop Database to determine features that may help design cyclic hexapeptides. Two datasets of β-turns were created: the β-hairpin dataset comprised of β-turns from β-hairpin structures and the β-turns dataset containing β-turns extracted from wider loop regions. The different datasets were clustered to identify different β-turn types. All dipeptide sequences made up of the 20 naturally occurring amino acids were searched for throughout the Database. The percentage that a dipeptide sequence occupied the i+1 and i+2 positions of a β-turn was then used as an estimate of turn propensity. Two peptides were designed based on the estimated turn propensities: c(VPNRGD) and c(VPRGDN). BE-META simulations indicated the peptides would both adopt conformations with the amino acids in the predicted registers based on the estimated turn propensity. Both peptides had their conformations determined by NMR. It was found during the synthesis of c(VPRGDN) that Aloc-valine provides a convenient route to the synthesis of the cyclic peptide when the Fmoc group could not be removed from the valine, likely due to sterics. The NMR for both peptides indicated the major conformation had the amino acids in the predicted register. A small cis proline-containing conformation was also seen in the NMR of c(VPNRGD).

In Chapter 4 restrained BE-META simulations were designed to explore the lowest energy conformations of cyclic hexapeptides. Restraints were introduced to enforce a particular β-turn type at one end of the peptide. The lowest energy structures based on the backbone energy of the peptide was then determined by observing the most common β-turn type combinations. Introducing chiral amino acids at the i/i+3 positions altered the most favourable turn combinations. When different chiral amino acids occupied the i/i+3 positions the same Ramachandran distributions were seen for a particular turn type combination. The similarity of the Ramachandran distribution necessary for a particular combination was compared with the Ramachandran distribution of a particular amino acid in the absence of structural restraints using the normalised integrated product (NIP). It was hypothesised that having a more similar distribution (higher NIP) for a particular conformation would mean a higher occurrence of that conformation in the restrained simulation of that particular amino acid. NIP values were therefore generated based on the MD simulations of Ac-GGXGG-NH_2 where X is one of the 19 naturally occurring chiral amino acids and the Ramachandran distributions of c(AGGAGG) for each turn type combination made up of type I, II, I’ and II’ turns. It was found the NIP value could not be used to predict the proportion of the different conformations in the restrained simulations. When different chiral amino acids are included in the peptide structure, although the Ramachandran distributions for a particular turn type combination are very similar, very small differences were present. This means, due to the very small energy differences examined, the NIPs cannot be used to predict the amount of a particular conformation. The conformations observed in the restrained simulations provide information that could be used in the design of cyclic hexapeptides. For example, the D-Pro-L-Pro motif, which is commonly used as a β-turn inducing element, frequently forms a type II' β-turn. The restrained simulations indicate the presence of a type II' turn within a cyclic hexapeptide could alter the overall most stable conformation. If incorporating a β-turn forming sequence into a cyclic peptide such information should be taken into account.

Using information from the database analysis in Chapter 3, in Chapter 5 a Random Forest (RF) was trained to predict the β-turn type a tetrapeptide sequence was likely to form. For the final version of the RF both the β-turns from the β-hairpin and β-turns datasets were used to train the RF, with ambiguous sequences which form multiple turn types removed from the training set as this produced the highest accuracy across the different categories. The RF scoring system was devised with the hypothesis that higher scoring sequences would be more likely to form the predicted turn type, which could potentially lead to a cyclic hexapeptide with fewer conformations, whereas lower scoring sequences would be more likely to form multiple turn types leading to more conformations. The RF was used in combination with the restrained simulations to predict the conformations likely to be adopted by various sequences. By choosing sequences likely to form low energy turn combinations based on the restrained simulations with high RF scores, peptides were identified which based on BE-META simulations formed the predicted conformation. Several factors were determined to be important for structure prediction of cyclic hexapeptides including the backbone conformation, the β-turn type a sequence was likely to form and the turn propensity of a sequence. Such information could be beneficial to future structure prediction methods and applied to other macrocycle sizes. As BE-META can take several days to run, methods such as the RF can offer much faster filtering of potential sequences, identifying candidates likely to form a desired structure in seconds.

In Chapter 6 conditions were identified which allowed for the incorporation of a BTM into cyclic peptides. As the BTM forms through a chemical ligation reaction the formation of the BTM allowed for cyclisation of the peptide and introduction of a structural constraining element at the same time. For shorter peptides the BTM could be formed during the cleavage of the peptide from the resin using TIPS as the reducing agent. For larger peptides greater yields were obtained when the peptide was initially cleaved from the resin using non-reductive conditions and then the acylhydrazone reduced to form the final BTM in a separate step using sodium cyanoborohydride. The reaction was found to have a broad sequence tolerance, but, inclusion of asparagine adjacent to the hydrazine led to decomposition of the product. Two RGD-containing cyclic peptides were designed using the BTM. BE-META simulations on the peptides demonstrated how, as well as allowing for efficient cyclisation, the BTM also acts as a structure-inducing element leading to the two peptides being relatively well-structured. As cyclisation of peptides can often prove difficult the use of the BTM, which can be made using commercially available materials compatible with SPPS, provides an alternative route to synthesise peptides. It could therefore potentially be used for the synthesis of peptide libraries and in the design of potentially biologically active peptides.

Chapter 7 aimed to incorporate fluorescence into the BTM through the use of a naphthalene unit. A TrpZip-based peptide was synthesised to incorporate the naphthalene-containing BTM. 2-[(1-

formylnaphthalen-2-yl)oxy]acetic acid was synthesised and coupled into a TrpZip peptide fragment. Incomplete coupling was seen, likely due to the added steric bulk of the naphthalene unit compared to the original BTM. Two peptide fragments were ligated together to produce the final peptide. The CD spectra of the naphthalene BTM TrpZip peptide was compared to that of TrpZip1 and the peptide containing the original BTM. A β-sheet structure is seen in all three peptides demonstrating the naphthalene BTM allows for retention of the BTM structure. Differences in the CD spectra indicated the interaction of the tryptophan sidechains with the naphthalene group. Fluorescence spectra of the naphthalene TrpZip were obtained. The naphthalene unit has very similar fluorescence properties to tryptophan. Modification of the BTM, as demonstrated by the addition of the naphthalene unit, offers the potential to introduce additional functionality into the ligation junction. Future versions of the BTM could therefore be used to introduce additional properties into peptides. As fluorescence is often used to monitor biological pathways in cells, future versions of the BTM could have wide ranging applications.

Finally, the β2 and β3 strands of a YAP1 WW domain were observed to contain all the residues necessary for binding of a proline-rich ligand. A cyclic WW Domain mimic was therefore designed in Chapter 8 by cyclising the two β2 and β3 strands using the BTM. Due to limited solubility four amino acids on the non-binding side of the β-strands were substituted. BE-META simulations indicate the β-sheet structure of the β-strands was retained. A fluorescence titration with a proline-rich ligand shows the cyclic WW domain mimic retained the ability to bind to a ligand. As WW domains are involved in many protein-protein interactions (PPIs), a synthetically designed WW domain could potentially be used to monitor the PPI pathways. The designed WW domain therefore offers a proof of principle which could be further developed and applied to different WW domain signalling pathways.

Due to their potential applications, including in the inhibition of PPIs, methods to design cyclic peptides with a specific well-structured conformation are advantageous. Computational methods which predict the structure of cyclic peptides can be used to identify potential sequences which will form a desired structure before expending resources on synthesising peptides. In this work information from a Loop Database and BE-META simulations was used to predict the structure of cyclic hexapeptides. The use of the BTM provides an alternative strategy to design cyclic peptides. In addition to providing a structural element that can reduce the number of conformations of the peptide, formation of the BTM through a chemical ligation reaction was shown to allow for efficient cyclisation of peptides of a variety of sizes. The methods developed in this work therefore have potential to be used in the design of cyclic peptides with a variety of applications.

10 Experimental

10.1 Loop Database

The database was constructed by Dr Drew Thomson. Using custom Python scripts the PISCES[434] server was used to identify a subset of high-resolution, non-sequence redundant protein crystal structures (resolution 2.5 Å or better, R-factor 1.0 or better, PDB version 28/10/2019). Regions of secondary structure were defined as having four or more contiguous residues of the same secondary structural assignment with DSSP[435, 436]. A 'loop' was defined as unstructured by DSSP, including regions containing up to three contiguous residues with the same DSSP secondary structural assignment, in addition to blocks of mixed secondary structures. *N*- and *C*-terminal unstructured regions were excluded by the requirement that a loop be flanked on each side by a region of secondary structure. Loops were filtered to exclude any in which the loop was discontinuous due to missing residues or atoms in the crystal structure. A database entry for each identified loop was generated containing the 3D coordinates for the loop, as well as the four residues of secondary structure on either side. The database entry also contained information such as the sequence, DSSP assignment, and vectors representing the end-to-end separation and orientation of flanking secondary structure for each loop. In total the database contains 310,085 loops.

To extract β-turns from the database either the presence of a hydrogen-bond between the i and i+3 positions of four residues within a loop or a distance of less than 7 Å between the i and i+3 α-carbons was searched for. To search for the presence of a hydrogen-bond the vector of the C=O bond was mapped to the N-H position. If the angle between the two vectors was greater than 120 ° and the distance between the oxygen and hydrogen atoms less than 2.5 Å a hydrogen bond was considered to be present.

10.1.1 Hairpin Dataset

Loops were extracted from the database with a loop length of 2 residues and β-sheet assigned secondary structure on either side of the loop region. To create the hairpin dataset these loops were filtered to identify all β-turns with a hydrogen-bond as well as at least three additional hydrogen bonded pairs of residues adjacent to those in the β-turn. The expanded hairpin dataset includes all turns identified using the hydrogen-bonded definition of a β-turn as well as all turns identified using the 7 Å distance definition of the β-turn.

10.1.2 β-turns Dataset

The β-turns dataset includes all β-turns within the loop regions in the database using the hydrogen-bonded definition of a β-turn. All four residues which make up the β-turn must be classified as part of a loop region to be identified, so no overlap with the hairpin dataset is seen. The β-turns dataset is approximately 10 times larger than the expanded β-hairpin dataset containing over 150,000 β-turns.

10.1.3 Cluster Analysis

For the hairpin dataset a sin and cos transformation was performed on the ϕ and ψ dihedral angles of the i to i+3 positions of the β-turns. Principal component analysis (PCA) was then used to reduce the number of dimensions to three. A density-based clustering algorithm[303] was then used to separate the different types of β-turns.

For the β-turns dataset only the i+1 and i+2 dihedral angles were included. Due to the large size of the dataset one thirtieth of the data was clustered. To identify the remaining turns belonging to each cluster, all datapoints in the principal component subspace were initially added to a cluster based on the closest cluster center of the data that was clustered. The principal component subspace of each cluster was then divided into grids and the population of each grid counted. Outliers which did not belong to any cluster were then removed by removing the least populated grids from the cluster.

10.2 Random Forest

The Random Forest to predict the β-turn type a tetrapeptide sequence was likely to form was built using scikit-learns built in Random Forest Classifier function.[437] 100 trees were used with the Gini importance measure to determine split points of the data. All other parameters were kept as default. The β-turns extracted from the Loop Database were randomly split into 10 % test set and 90% training set.

10.3 Alanine Scan

PDB files of structures were used for computational alanine scans using the Balas Webserver.[427] For peptides containing the BTM, the BTM unit was deleted from the PDB file prior to submission to the alanine scan as the Balas Webserver currently cannot perform alanine scans on non-canonical amino acids. Each residue of the selected peptide is swapped for alanine in turn and the ΔΔG value measuring the interaction of the mutant with a ligand is calculated using the BUDE[438] all-atom forcefield.

10.4 Sequence Alignment

Sequence alignment of protein sequences from the PDB was performed using the Align Chain Sequences function in Chimera.[439] The function aligns sequences using a Clustal Omega web service.[440]

10.5 Molecular Dynamics

10.5.1 Set up and Equilibration

All models were initially built using Chimera.[439] Molecular Dynamics (MD) simulations were carried out using Gromacs 4[40] with the Plumed 2.5[41] plugin. After immersion of the peptide in a cubic box of explicit water, with a minimum distance of 1.0 nm between the peptide and the edge of the box, a steepest descent algorithm was used. A four-stage equilibration process was then carried out. First a 50 ps NVT ensemble was carried out followed by a 50 ps NPT ensemble. Both ensembles were run with 1000 kJ/mol/nm^2 restraints on the heavy atoms. Next a 100 ps NVT ensemble followed by a 100 ps NPT ensemble was carried out, this time without restraints. To prevent the hot solvent, cold solute problem[441] a separate V-rescale thermostat[442] was used for both the peptide and the solvent. A Berendsen barostat[443] with an isothermal compressibility of 4.5 x 10^{-5} bar^{-1} was used throughout the NPT ensembles and the MD simulations. The final trajectory file from the NPT 100 ps ensemble was used as the initial input for the MD simulations.

10.5.2 Conventional MD simulations

A 2 fs timestep was used throughout the simulations. Simulations were carried out at 300 K at 1 bar with the explicit TIP3P[444] water model. An Amber99SB[283] or RSFF2[242] forcefield was used. For the duration of the simulation all bonds to hydrogen atoms were restrained to equilibrium values using a LINCS[445] algorithm. Nonbonded interactions (both Lennard-Jones and electrostatic) were truncated at 1 nm. For long range electrostatics a particle mesh Ewald (PME)[446] with an order of 4 and Fourier spacing of 0.12 nm was used.

10.5.3 BE-META

10.5.3.1 Cyclic Hexapeptides

The BE-META simulations on cyclic hexapeptides were carried out with 12 biased replicas: a replica for each residue biased along its ϕ and ψ angles (ψ_i and ϕ_i), and then a replica for each residue biased along its ψ dihedral and the ϕ dihedral of the next residue along (ψ_i and ϕ_{i+1}). Gaussian hills of height 0.1 kJ/mol and width 0.3 radians were added every 4 ps. An additional biased replica was added when the cyclic peptide contained a proline residue. This additional replica biased the improper dihedral angle ζ of the Proline as well as the proline ψ angle to allow for cis/trans isomerisation. The height of the Gaussians of this biased replica was 0.2 kJ/mol and the sigma values 0.2 radians, and the Gaussians were added every 4 ps. Five additional unbiased replicas were also included for analysis to give a total of seventeen (or eighteen for proline-containing peptides) replicas per simulation. A leapfrog algorithm[447] with a timestep of 2 fs was used. Replica exchanges were attempted every 5 ps. All simulations were carried out using the RSFF2 force field at 300 K and 1 bar. For the duration of the simulation all bonds to hydrogen atoms were restrained to equilibrium values using a LINCS algorithm. Nonbonded interactions (both Lennard-Jones and electrostatic) were truncated at 1 nm. For long range electrostatics a particle mesh Ewald (PME) with an order of 4 and Fourier spacing of 0.12 nm was used.

10.5.3.2 Restrained BE-META

For the restrained BE-META simulations the six replicas with bias for each residue along its ψ dihedral and the ϕ dihedral of the next residue along (ψ_i and ϕ_{i+1}) were no longer used. Additionally the replicas biasing the ϕ and ψ dihedral angles of the restrained residues were removed. The restraints were added to the ϕ and ψ dihedral angles of the same two amino acids across all replicas at 50 kJ/mol to the ideal dihedral angles of either a type I, II, I' or II' β-turn. No restraints were used during initialisation and equilibration. Three unbiased replicas were used for analysis, giving a total of 7 replicas per restrained BE-META simulation on a cyclic hexapeptide. All simulations converged within 100 ns. The starting structure must resemble one with two β-turns already present or the simulation may fail in the first stages as the restraints force it too quickly into a different conformation. The same conditions as stated above for the BE-META simulations on cyclic hexapeptides were used in each replica. 70 kJ/mol restraints were included for all ω dihedral angles.

For the doubly restrained simulations four of the six amino acids within a cyclic hexapeptide have restrained ϕ and ψ dihedral angles. A replica for each of the two remaining amino acids biases the ϕ and ψ dihedral angles of those residues. Two unbiased replicas were used for analysis (four replicas were used in total).

10.5.3.3 BE-META for BTM-Containing Cyclic Peptides

Partial charges for the BTM unit were generated by Dr Drew Thomson using antechamber with the AM1-BCC charge scheme.[448]

The same conditions as used for the BE-META on cyclic hexapeptides were used but with the Amber99SB rather than the RSFF2 forcefield. It was found the BTM was sufficiently flexible that no additional CVs were necessary.

10.6 Analysis of MD Trajectories

The ϕ and ψ dihedral angles for each residue were extracted from the trajectory of the conventional MD simulations or for BE-META simulations, the unbiased trajectories. A sin and cos transformation was performed on each dihedral angle to convert the dihedrals from circular variables to a linear metric coordinate space and principal component analysis used to reduce the number of dimensions to three.[42] The first three principal components accounted for around 56 – 98 % of the variance of the data. Every 100th value of the first three principal components was clustered using a density-based clustering algorithm.[43] Each cluster represents a conformation of the peptide in solution. The centre of each cluster (the highest density point of the cluster) was also extracted and used to represent each cluster. The dihedral angles of all the frames from each cluster were extracted and Ramachandran diagrams plotted. Backbone atom RMSD values were calculated using the match function in Chimera.[439]

10.6.1 Checking for Convergence

For each peptide two different initial starting conformations were built. If convergence has been reached then the two simulations on each of the starting conformations will produce the same peptide conformations in similar proportions. To check for convergence the normalised integrated product (NIP) of the dihedral angle principal component subspace of the clusters identified in the two simulations was measured.[449] If the simulation had not reached convergence after 100 ns the simulation was extended by a further 100 ns until convergence was reached. The time taken to reach convergence and the NIP value measuring the similarity of the principal component subspace of the clusters identified in the two simulations on the peptides with different starting conformations is shown in Table 77. Convergence was assumed to have been reached when a NIP value above 0.9 was obtained.

Peptide	Time (ns)	NIP
c(PWGNKY)	100	0.946
c(PWGNKY)*	200	0.964
c(PWRNKY)	200	0.922
c(PWRNKY)*	300	0.958
c(PWGNKE)	100	0.953
c(PWRNKE)	100	0.977
c(PWRNKE)*	200	0.984
c(VPRGDN)	200	0.919
c(VPNRGD)	100	0.957
c(GGGGGG)	200	0.935
c(WGTGCS)	100	0.969
c(NFEWSG)	100	0.993
c(RGNQPG)	100	0.986
c(RGSQGW)	200	0.901
c(NWQNVA)	100	0.974
c(EGDSAR)	100	0.973
c(NSKSED)	100	0.902
c(TPRGDG)	100	0.999
c(BTM-PG)$_2$	200	0.949
c(BTM-RGD)	100	0.964
c(BTM-SLSPGRGD)	200	0.986

*Table 77: Time taken for each BE-META simulations to reach convergence and the NIP value showing the similarity of the clusters identified in the two simulations. *BE-META simulation without the inclusion of a proline replica.*

10.7 General Information

Unless otherwise stated, all chemicals used were bought from Sigma Aldrich. $Pd(PPh_3)_4$ was purchased from Fluorochem. 2-mercaptoethanol was purchased from Alfa Aesar. Fmoc-Gly-OH and Fmoc-Ala-OH were bought from Iris Biotech GmbH and Fmoc-Lys(Boc)-OH from Activotec. Fmoc-Gln(Trt)-OH and Fmoc-Pro-OH were bought from Pepceuticals and Fmoc-Trp(Boc)-OH, Fmoc-

Tyr(*t*Bu)-OH and Fmoc-Cys(Trt)-OH were purchased from CEM. Peptide grade DMF was purchased from Rathburn. All other solvents were purchased from VWR. Commercially available starting materials were used without purification unless otherwise stated. All amino acids are of L-configuration unless otherwise stated. Dry solvents were purified using a PureSolv 500 MD solvent purification system.

Peptides were purified on a reverse-phase Dionex HPLC system equipped with Dionex P680 pumps and a Dionex UVD170U UV-vis detector (monitoring at 214 nm and 280 nm). A Phenomenex, Gemini, C18, 5 µm, 250 x 21.2 mm column at a flow rate of 3 mL/min or a Phenomenex, Luna, C8(2), 5 µm, 250 x 10 mm column at a flow rate of 8 mL/min was used. Gradients were run using a solvent system consisting of A (H_2O + 0.1 % TFA) and B (MeCN + 0.1 % TFA). Collected fractions were lyophilised on a Christ A 2-4 LO plus freeze dryer.

Pure peptides were analysed on a Shimadzu reverse-phase HPLC system equipped with Shimadzu LC-20AT pumps, a Shimadzu SIL-20A autosampler and a Shimadzu SPD-20A UV-vis detector (monitoring at 214 nm and 280 nm). A Phenomenex, Aeris, 5 µm, peptide XB-C18, 150 x 4.6 mm column at a flow rate of 1 mL/min was used for analysis of the peptides. Gradients were run using a solvent system consisting of solution A (H_2O + 0.1% TFA) and B (MeCN + 0.1% TFA). A 5-50% B gradient over 30 minutes was used for analysis.

LC-MS analysis was performed on a Thermo Scientific Dionex Ultimate 3000 LC system coupled to a Thermo Scientific LCQ Fleet quadrupole mass spectrometer using positive mode electrospray ionisation (ESI+). A Dr Maisch ReproSil Gold 120 C18, 110 Å, 3 µM, 150 x 4 mm column was used. Gradients were run using a solvent system consisting of solution A (95/5 H_2O/MeCN and 0.1% TFA) and B (95/5 MeCN/H_2O and 0.1% TFA).

High resolution mass spectrometry (HRMS) was performed by the analytical service of the University of Glasgow. Unless otherwise stated a Bruker microTOF-Q II High Resolution Mass Spectrometer using ESI+ ionisation in positive mode was used. HRMS data are reported as mass to charge ratio.

Peptide content was analysed on a Thermo Scientific NanoDrop One UV-Vis spectrophotometer.

Fluorescence spectra were recorded on a HORIBA duetta bio fluorescence and absorbance spectrometer using a cuvette with a 1 cm pathlength.

Circular dichroism (CD) spectra were obtained for 190-260 nm UV using a JASCO J-810 spectropolarimeter fitted with a Peltier temperature controller. A cuvette with a 0.1 cm pathlength was used and spectra obtained at 5 °C unless otherwise stated. A peptide concentration of 10 µM in MOPS buffer at pH 7.4 was used. Thermal denaturation experiments measured the mean residue ellipticity (MRE^{450}) at 225 nm across a temperature range of 5 to 80 °C monitoring every 1 °C.

Small molecule nuclear magnetic resonance (NMR) spectra were recorded on a Bruker AVI 400MHz spectrometer (400 MHz for ^{1}H-NMR and 100 MHZ for ^{13}C-NMR). Peptide NMR were recorded on a Bruker AVANCE 600 MHz spectrometer equipped with a TCI cryoprobe. Peptide NMR samples contained a peptide concentration of 100 µM in 600 µL of water with 5% D_2O and potassium phosphate buffer (pH 7.4). Chemical shifts (δ) are reported in parts per million (ppm) relative to an internal standard or the solvent residual peaks. For δH these correspond to Me_4Si at 0 ppm for the internal standard or the solvent residual peaks of $CDCl_3$: 7.26 ppm, $(CD_3)_2SO$: 2.50 ppm, D_2O δH: 4.79 ppm. For δC the internal standard Me_4Si at 0 ppm or the solvent residual peaks of $CDCl_3$: 77.16 ppm or $(CD_3)_2SO$: 39.52 ppm. Chemical shifts for the small molecules were assigned using proton, carbon, Correlation Spectroscopy (COSY) and Heteronuclear Single Quantum Coherence (HSQC) experiments. For peptide NMR rotating-frame nuclear Overhauser effect correlation spectroscopy (ROESY) was also used for assignment. Splitting patterns are abbreviated as follows: singlet (s), doublet (d), triplet (t), quartet (q), multiplet (m), broad (b), or a combination of these.

10.8 General Procedure for Microwave-assisted SPPS

Peptides were synthesised on a CEM Liberty Blue microwave-assisted peptide synthesiser using the Fmoc/*t*Bu protecting group strategy. Unless otherwise stated all peptides were synthesised on a 0.1 mmol synthetic scale. Fmoc deprotection was performed with a solution of 20% morpholine in DMF (4 mL), at rt and 0 W for 5 sec followed by 78 °C and 100 W for 30 sec, 88 °C and 70 W for 20 sec, and 90°C and 25 W for 60 sec. The peptide was washed after deprotection. DIC (5 equiv., 0.5 M in DMF) and Oxyma Pure (5 equiv., 1 M in DMF) were used to couple N-α-Fmoc protected amino acids (5 equiv., 0.2 M in DMF). To couple the amino acid, the reaction vessel was held at rt at 0 W for 5 sec, followed by being heated to 80 °C at 100 W for 30 sec, 86 °C at 70 W for 20 sec and 90 °C at 25 W for 120 sec. Coupling of Fmoc-Cys(Trt)-OH and Fmoc-His(Trt)-OH was carried out without heating at 0 W for 120 sec and then at 50 °C at 50 W for 480 sec. Following coupling the resin was washed with DMF.

10.9 General Procedure for Peptide Cleavage

The peptide bound to the resin was washed with DCM. To cleave the peptide from the resin and remove any sidechain protecting groups a cleavage cocktail of 95% TFA, 2.5% water and 2.5% triisopropylsilane (TIPS) was used (10 mL total volume) and the mixture was left to react for 2 h at rt with mixing. After filtration, the cleavage cocktail was reduced in volume to approximately 5 mL using a stream of nitrogen and the peptide was precipitated from solution with ice cold diethyl ether and centrifuged at 3700 rpm for 5 min. The precipitated peptide was dissolved in 1:1 water/acetonitrile and lyophilised.

For arginine containing cyclic peptides a cleavage mixture of 2.5% water, 2.5% TIPS and 3% β-mercaptoethanol (BME) in TFA was used and the resin mixed for 3 h.

10.10 General Procedure for Synthesis of BTM-Containing Cyclic Peptides

10.10.1 Reactivation of the Cl-trityl Resin

Following swelling of the Cl-trityl resin (0.25 g, 0.8 mmol/g) in 1:1 DMF/DCM for 15 minutes, thionyl chloride (0.1 mL) in DCM (4 mL) was added to the resin and left to react for 30 minutes. The resin was washed four times with DCM then the reaction repeated.

10.10.2 Synthesis of the Hydrazine Resin

Hydrazine hydrate (20 equiv, ~80%) in DMF (6 mL) was added to Cl-trityl resin (0.25 g, 0.8 mmol/g) and left to react for 30 minutes. After washing the resin four times with DMF the reaction was repeated. 10% methanol in DMF (6 mL) was then added to the resin and left to react for 30 minutes to cap any remaining unreacted sites.

10.10.3 Coupling of the Aldehyde

2-formylphenoxyacetic acid (5 equiv) was dissolved with, DIC (4.5 equiv) and Oxyma Pure (4.5 equiv) in DMF. DIPEA (6 equiv) was then added. The solution was stirred for 5 min for preactivation and then it was added to the resin for reaction (2 h).

10.10.4 Reductive Cleave/Cyclisation Conditions

90% TFA, 10% TIPS (5 mL) was added to the resin and heated to 50 °C with gentle stirring for 30 minutes or at rt for 2 h.

10.10.5 Non-Reductive Cleave Conditions

The peptide was cleaved from the resin using 95% TFA, 5% water (5 mL) with mixing for 2 h.

10.10.6 Non-Reductive Cleave Conditions without Removal of Sidechain Protecting Groups

To cleave the peptide from the resin without removal of the amino acid sidechain protecting groups, 1% TFA, 5% TIPS in DCM (5 mL total volume) was added to the resin and mixed for 10 minutes. The sidechain protecting groups can then be removed in a second cleave with 50% TFA and 5% TIPS in DCM with stirring for 30 minutes.

10.10.7 Sodium Cyanoborohydride Reduction

The peptide was dissolved in 1:1 MeOH/AcOH (5 mL) and sodium cyanoborohydride (10 equiv) added. The solution was stirred for 15 minutes.

10.11 General Procedure for Cyclic Peptide Synthesis

Fmoc-Asp-OAll (2 equiv) was dissolved with DIPEA (4 equiv) and HATU (1.9 equiv) in DMF (5 mL) and left to pre-activate for 5 minutes before adding to Rink Amide AM resin (0.1 mmol scale, 0.42 mmol/g). The remaining residues were coupled by Microwave-assisted SPPS. Following washing of the resin with DCM, the allyl group was removed from the C-terminus using $Pd(PPh_3)_4$ (0.25 equiv) and phenylsilane (24 equiv) for 30 minutes in dry DCM (5 mL). This reaction was then repeated to ensure complete deprotection of the allyl group. The Fmoc group on the *N*-terminus was removed using 2 % morpholine and 2 % DBU in DMF (5 mL total volume) and cyclisation achieved using PyBOP(1.5 equiv) and DIPEA (1.5 equiv) in DMF (5 mL) for 2 h. The peptides were purified by RP-HPLC and obtained as a white solid.

10.12 Chapter 2 Inclusion of a Proline Replica in BE-META of Cyclic Peptides

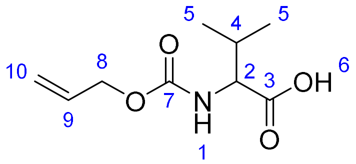
All peptides were synthesised using the general procedure for cyclic peptide synthesis.

Peptide	Empirical Formula	Calculated M+H$^+$ mass	Observed M+H$^+$ mass	HPLC Retention Time (min)
c(PWGNKY)	$C_{37}H_{48}N_9O_8$	746.3620	746.3583	19.9
c(PWRNKY)	$C_{41}H_{57}N_{12}O_8$	845.4417	845.4367	19.4
c(PWGNKE)	$C_{33}H_{46}N_9O_9$	712.3413	712.3486	20.6
c(PWRNKE)	$C_{37}H_{55}N_{12}O_9$	811.4209	811.4199	18.8

Table 78: Characterisation data for c(PWGNKY), c(PWRNKY), c(PWGNKE) and c(PWRNKE).

10.13 Chapter 3 Database Analysis Peptides and Small Molecules

Aloc-Valine



H-Val-OH (50 mg) was dissolved in 30 mL of 0.4 M NaOH(aq). Allyl chloroformate (56 µL, 1.2 equiv) was added dropwise at 0 °C. The solution was allowed to warm to rt and stirred overnight. After washing with diethyl ether, the aqueous layer was acidified to around pH 4 with 1 M HCl(aq). The product was extracted using ethyl acetate and the organic layer dried with magnesium sulfate. The mixture was then filtered and the solvent evaporated *in vacuo*. The product was purified using RP-HPLC and obtained as a colourless oil in a 69% yield.

HRMS (ESI+): $C_9H_{15}NO_4$ M+Na$^+$ calculated 224.0899, observed 224.0935

1H NMR (400 MHz, $CDCl_3$) δ_H 8.05 (s, 1H, H-6), 5.85 (m, 1H, H-9), 5.20 (dd, 2H, H-10), 4.51 (d, 2H, H-8), 4.20 (dd, 1H, H-2), 2.10 (m, 1H, H-4), 0.88 (d, 3H, H-5a), 0.80 (d, 3H, H-5b)

13C NMR (100 MHz, $CDCl_3$) δ_C 176.93 (C-3), 156.26 (C-7), 132.51 (C-9), 118.05 (C-10), 66.05 (C-8), 58.78(C-2), 31.00 (C-4), 19.04 (C-5a), 17.35 (C-5b)

Peptides:

c(VPNRGD) was synthesised using the general procedure for cyclic peptide synthesis. To synthesise c(VPRGDN) Fmoc-Asp-OAll (2 equiv) was coupled to Rink Amide AM resin (0.1 mmol scale, 0.42 mmol/g) using microwave-assisted SPPS. P, R, G and D were also coupled using microwave-assisted SPPS. Aloc-Val (2 equiv) was dissolved in DMF (5 mL) with HATU (1.9 equiv) and DIPEA (4 equiv) and left to preactivate for 5 minutes before addition to the resin. The resin was mixed for 2 h and then washed 4 times with DMF and once with DCM. The Aloc and allyl group were removed from the N- and C-termini respectively in one step using $Pd(PPh_3)_4$ (0.25 equiv) and phenylsilane (24 equiv) for 30 minutes in dry DCM (5 mL). This reaction was then repeated to ensure complete deprotection of the allyl group. PyBOP(1.5 equiv) and DIPEA (1.5 equiv) dissolved in DMF (5 mL) was added to the resin which was mixed for 2 h to allow for cyclisation. The resin was washed with DMF and DCM. The peptides following cleavage from the resin were purified by RP-HPLC and obtained as white solids.

Peptide	Empirical Formula	Calculated M+H$^+$ mass	Observed M+H$^+$ mass	HPLC Retention Time (min)
c(VPNRGD)	$C_{26}H_{43}N_{10}O_9$	639.3209	639.3211	14.9
c(VPRGDN)	$C_{26}H_{43}N_{10}O_9$	639.3209	639.3218	19.0

Table 79: Characterisation data for c(VPNRGD) and c(VPRGDN).

10.14 Chapter 6 Incorporation of a β-turn Mimic within a Cyclic Peptide

All peptides were synthesised using the general procedure for the synthesis of cyclic peptides containing the BTM on a 0.03 mmol scale. The reductive cleave/cyclisation conditions at 50 °C for 30 minutes were used with the exception of c(BTM-KPGC), c(BTM-KPG) and c(BTM-SAKPGASA).

c(BTM-KPGC) was cleaved from the resin using the reductive cleave conditions at room temperature for 2 h.

c(BTM-KPG was cleaved from the resin using 1% TFA, 5% TIPS in DCM (5 mL total volume) for 10 minutes. Following evaporation of the cleavage mixture and precipitation in ice cold diethyl ether the peptide was redissolved in 1:1 MeOH/AcOH (5 mL) and sodium cyanoborohydride (10 equiv) added. Following evaporation of the 1:1 MeOH/AcOH the peptide was redissolved in 1:1 water/acetonitrile and freeze dried. The Boc protecting group was then removed from the lysine sidechain using 50% TFA and 5% TIPS in DCM (5 mL) with stirring for 30 minutes. The second cleavage mixture was evaporated using a stream of nitrogen and the peptide freeze-dried again prior to purification using RP-HPLC.

The two-step non-reductive cleave/sodium cyanoborohydride reduction method was used to synthesise c(SAKPGASA).

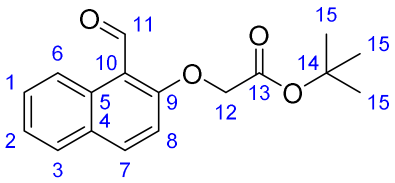
Peptide	Empirical formula	Calculated M+H$^+$ mass	Observed M+H$^+$ mass	HPLC Retention Time (min)
c(*BTM*-AGGA)	$C_{19}H_{26}N_6O_6$	457.1806	457.1962	16.7
c(*BTM*-KPGD)	$C_{27}H_{40}N_8O_7$	576.2776	576.7700	23.8
c(*BTM*-KPGQ)	$C_{27}H_{39}N_7O_8$	589.3093	589.3087	16.3
c(*BTM*-KPGE)	$C_{27}H_{40}N_8O_7$	590.2933	590.2930	16.5
c(*BTM*-QGPK)	$C_{27}H_{40}N_8O_7$	589.3093	589.3086	18.5
c(*BTM*-KPGC)	$C_{25}H_{37}N_7O_6S$	564.2599	564.2597	18.3
c(*BTM*-KPGV)	$C_{27}H_{41}N_7O_6$	560.3191	560.3191	18.5
c(*BTM*-KPGR)	$C_{28}H_{44}N_{10}O_6$	617.3518	617.3516	16.7
c(*BTM*-PG)$_2$	$C_{32}H_{40}N_8O_8$	687.2861	687.2852	20.3
c(*BTM*-KPG)	$C_{22}H_{32}N_6O_5$	461.2507	461.2507	16.0
c(*BTM*-KAG)	$C_{20}H_{30}N_6O_5$	435.2350	435.2351	14.5
c(*BTM*-APG)	$C_{19}H_{25}N_5O_5$	426.1748	426.1748	16.6
c(*BTM*-AKPGA)	$C_{28}H_{42}N_8O_7$	603.3249	603.3238	17.3
c(*BTM*-AKPGSA)	$C_{31}H_{47}N_9O_9$	690.3570	690.3565	16.7
c(*BTM*-SAKPGASA)	$C_{37}H_{57}N_{11}O_{13}$	862.4065	862.4074	17.0
c(*BTM*-RGD)	$C_{21}H_{30}N_8O_7$	507.2310	507.2305	16.0
c(*BTM*-SLSPGRGD)	$C_{40}H_{61}N_{13}O_{14}$	948.4534	948.4516	21.0

Table 80: Characterisation data for the BTM-containing cyclic peptides in Chapter 6.

HRMS is not shown for c(BTM-KPGD) due to decomposition of the peptide.

10.15 Chapter 7 Naphthalene Containing β-turn Mimic Peptides and Small Molecules

tert-butyl 2-[(1-formylnaphthalen-2-yl)oxy]acetate



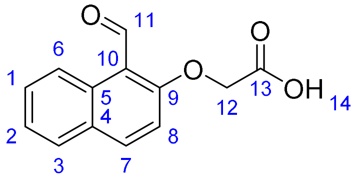
2-hydroxynaphthalene-1-carbaldehyde (500 mg) was dissolved in acetone (50 mL). Potassium carbonate (400 mg, 1 equiv) and tert-butyl bromoacetate (0.4 mL, 1 equiv) were added and the mixture refluxed overnight. After cooling to rt the potassium carbonate was removed by filtration and the acetone removed under vacuum. The crude product was redissolved in ethyl acetate and washed with water. The organic layer was dried with magnesium sulfate and the ethyl acetate removed by rotary evaporation. The product was obtained as a pale yellow solid in a 92 % yield. The crude product was used in the next reaction directly. Alternatively the product can be purified by column chromatography using 94% petroleum ether/6% ethyl acetate.

HRMS (ESI+): $C_{17}H_{18}O_4$, M+Na$^+$ calculated 309.1103 observed 309.1096

1H NMR (400 MHz, $CDCl_3$) δ_H 10.89 (s, 1H, H-11), 9.20 (d, 1H, H-6), 7.92 (d, 1H, H-8), 7.67 (d, 1H, H-3), 7.53 (m, 1H, H-1 or H-2), 7.34 (m, 1H, H-1 or H-2), 7.01 (d, 1H, H-7), 4.66 (s, 2H, H-12), 1.39 (s, 9H, H-15)

13C NMR (100 MHz, $CDCl_3$) δ_C 192.18 (C-11), 167.26 (C-13), 162.27 (C-10), 137.34 (C-8), 131.49 (C-4 or C-5 or C-9), 129.12 (C1 or C-2), 128.99 (C-4 or C-5 or C-9), 128.24 (C-3), 125.13 (C-6), 125.10 (C1 or C-2), 117.44 (C-4 or C-5 or C-9), 113.18 (C-7), 83.05 (C-14), 66.62 (C-12), 28.04 (C-15)

2-[(1-formylnaphthalen-2-yl)oxy]acetic acid



tert-butyl 2-[(1-formylnaphthalen-2-yl)oxy]acetate (600 mg) was dissolved in 10% TFA in DCM (50 mL) and stirred for 2 hours. The solvent was then removed by rotary evaporation. The product was purified by column chromatography using 5% AcOH in DCM in a 98% yield as a green solid. To ensure complete removal of the AcOH the product was redissolved in water and freeze dried.

HRMS (ESI-): $C_{13}H_9O_4$, M-H$^-$ calculated 229.0506 observed 229.0508

1H NMR (400 MHz, $CD_3)_2SO$) δ_H 10.88 (s, 1H, H-11), 9.12 (m, 1H, H-6), 8.26 (d, 1H, H-8), 7.95 (m, 1H, H-3), 7.66 (m, 1H, H-1 or H-2), 7.53 (d, 1H, H-7), 7.49 (m, 1H, H-1 or H-2), 5.07 (s, 2H, H-12)

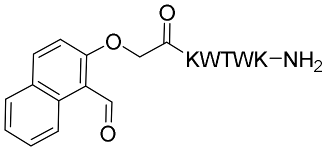
13C NMR (100 MHz, $(CD_3)_2SO$) δ_C 192.09 (C-11), 170.39 (C-13), 163.29 (C-10), 138.10 (C-8), 131.06 (C-4 or C-5 or C-9), 130.25 (C-1 or C-2), 129.05 (C-3), 128.90 (C-4 or C-5 or C-9),,), 125.35 (C-1 or C-2), 124.41 (C-6), 116.59 (C-4 or C-5 or C-9), 115.17 (C-7), 66.26 (C-12)

Naphthalene TrpZip Fragment 1

H-SWTWE-NH-NH_2

The naphthalene TrpZip was synthesised in two fragments which were ligated together to form the naphthalene BTM. To synthesise the first TrpZip fragment hydrazine hydrate (20 equiv, ~80%) in DMF (6 mL) was added to Cl-trityl resin (0.25 g, 0.8 mmol/g) and left to react for 30 minutes. After washing the resin four times with DMF the reaction was repeated. 10% methanol in DMF (6 mL) was then added to the resin and left to react for 30 minutes to cap any remaining unreacted sites. Microwave assisted Fmoc/*t*Bu SPPS was used to couple the SWTWE amino acids. The fragment was cleaved from the resin using 5% water in TFA (10 mL) for 2 hours. The peptide was purified by RP-HPLC and obtained as a white solid.

Naphthalene TrpZip Fragment 2



To make the second TrpZip fragment microwave-assisted SPPS was used to couple KWTWK to Rink Amide AM resin (0.1 mmol scale, 0.54 mmol/g). The 2-[(1-formylnaphthalen-2-yl)oxy]acetic acid (2.5 equiv) was dissolved with DIC (3.5 equiv), Oxyma Pure (2.5 equiv) and DIPEA (4 equiv) and left to preactivate for 5 minutes prior to addition to the resin. The aldehyde was left to couple to the peptide overnight. After washing with DMF the reaction was repeated overnight. The peptide was cleaved from the resin using 5% water in TFA for 2 h. The peptide was purified by RP-HPLC and obtained as a pale green solid.

Naphthalene TrpZip

H-SWTWE-nBTM-KWTWK-NH_2 where nBTM represents the naphthalene version of the BTM.

The naphthalene TrpZip peptide fragments 1 and 2 were dissolved in 1:1 water/acetonitrile and lyophilised together. Following freeze drying the peptide was dissolved in 1:1 MeOH/AcOH and sodium cyanoborohydride added (10 equiv). After 15 minutes the solvent was evaporated, the peptide redissolved in 1:1 water/acetonitrile and freeze dried. Following purification with RP-HPLC the peptide was obtained as a white solid.

Peptide	Empirical Formula	Calculated $M+2H^{2+}$ mass	Observed $M+2H^{2+}$ mass	Calculated $M+H^{+}$ mass	Observed $M+H^{+}$ mass	HPLC Retention Time (min)
Naphthalene TrpZip Fragment 1	$C_{34}H_{44}N_{9}O_{9}$	-	-	722.3257	722.3260	22.1
Naphthalene TrpZip Fragment 2	$C_{51}H_{64}N_{10}O_{9}$	960.4847	960.4805	-	-	26.3
Naphthalene TrpZip	$C_{85}H_{107}N_{19}O_{17}$	832.9041	832.9023	-	-	23.2

Table 81: Characterisation data of the naphthalene TrpZip peptide.

Circular dichroism spectra for the naphthalene TrpZip peptide:

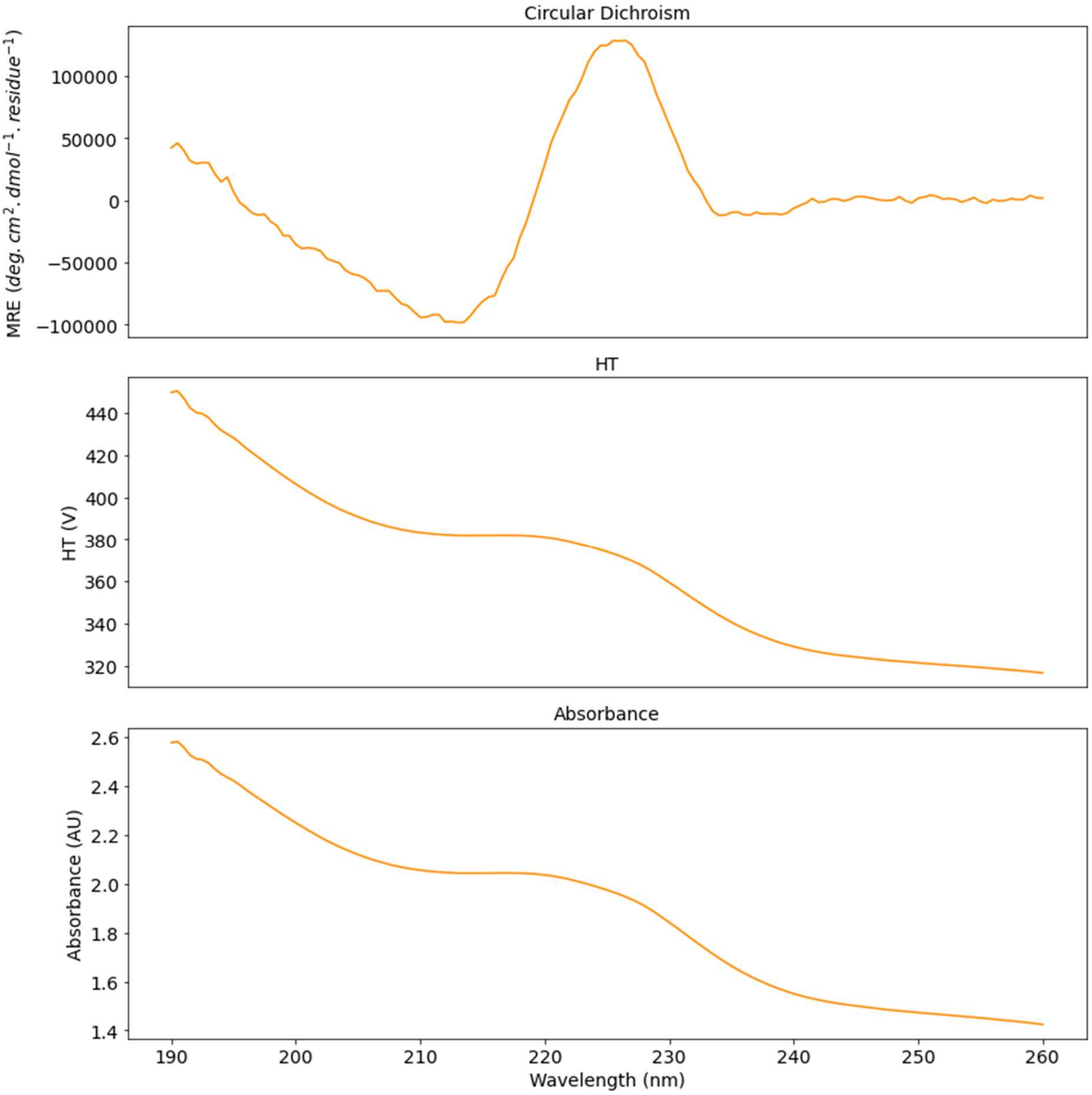


Figure 137: Circular dichroism of the naphthalene TrpZip peptide.

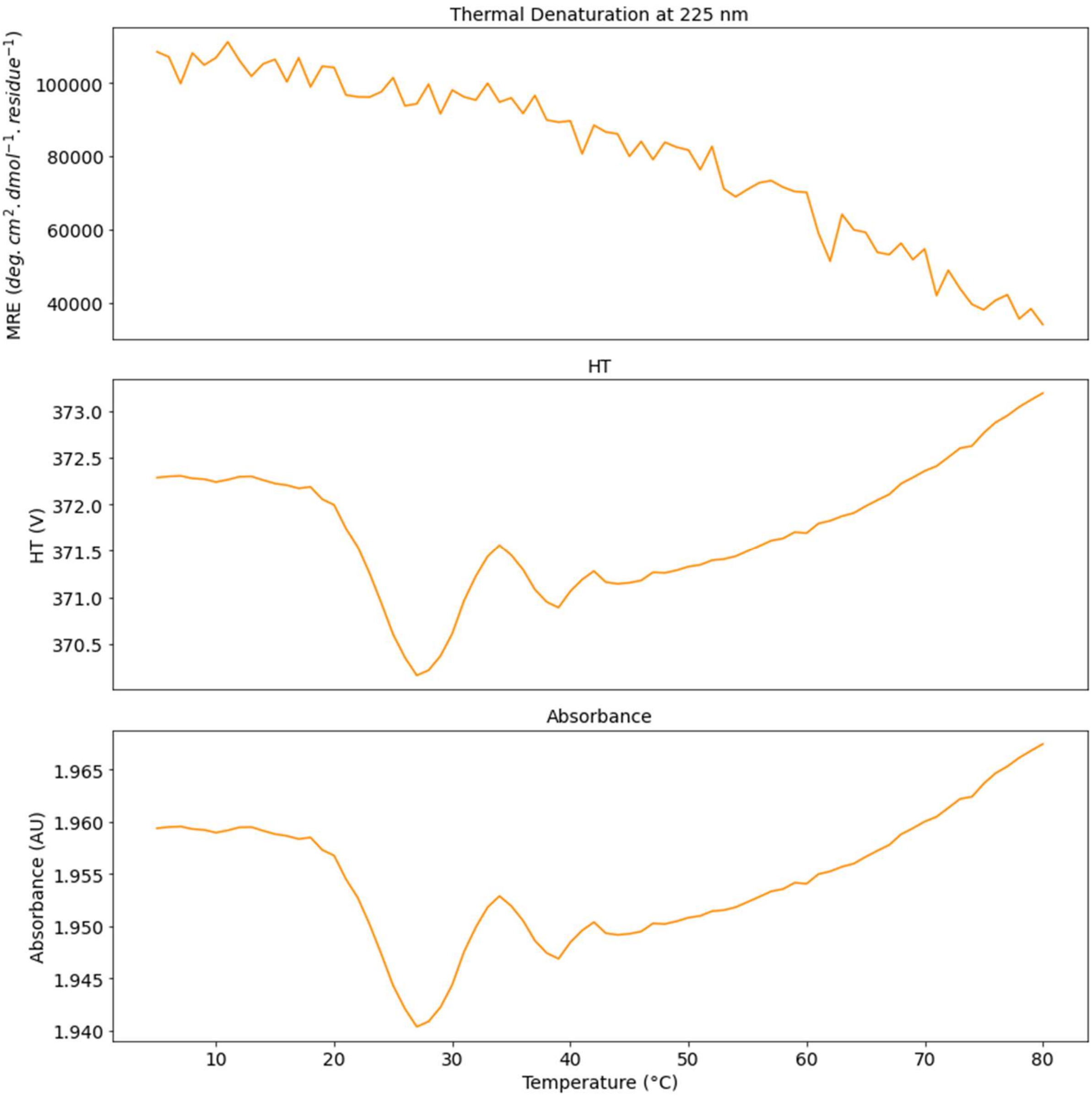


Figure 138: Thermal denaturation of the naphthalene TrpZip peptide measured at 225 nm.

10.16 Chapter 8 Design of a Cyclic WW Domain Peptides

c(β2β3-BTM)

c(BTM-RYKKNHPNQTTEWQ)

The peptide was synthesised using the general procedure for BTM-containing cyclic peptides on a 0.1 mmol scale using the two-step non-reductive cleave followed by reduction with sodium cyanoborohydride method. Following purification with RP-HPLC the peptide was obtained as a white solid.

L30K WW Domain Construct

H-FEIPDDVPLPAGWEMAKTSSGQRYFKNHIDQTTTWQDPRK-OH

The peptide was synthesised on a 0.1 mmol scale using microwave-assisted SPPS using H-Lys(Boc)-HMPB-ChemMatrix resin (0.4-0.65 mmol/g). The final 18 residues were double coupled. The peptide was cleaved from the resin using 2.5% TIPS, 2.5% water, 5% BME and 90% TFA for 4 h. Following purification with RP-HPLC the peptide was obtained as a white solid.

GTPPPPYTVG

GTPPPPYTVG was synthesised using microwave-assisted SPPS with Fmoc-Gly-Wang resin (0.1 mmol scale, 0.79 mmol/g). Standard cleave conditions were used. Following purification with RP-HPLC the peptide was obtained as a white solid.

Peptide	Empirical Formula	Calculated M+H$^+$ mass	Observed M+H$^+$ mass	Calculated M-3H^{3-} mass	Observed M-3H^{3-} mass*	HPLC Retention Time (min)
c(β2β3-BTM)	$C_{89}H_{129}N_{28}O_{25}$	1989.9678	1989.9669	-	-	24.1
L30K WW Domain Construct	$C_{210}H_{310}N_{57}O_{64}S$	-	-	1562.0828	1562.0779	16.0
GTPPPPYTVG	$C_{46}H_{69}N_{10}O_{14}$	985.4989	985.5002	-	-	23.5

*Table 82: Characterisation data of the WW Domain peptides. *ESI(-)*

10.17 Peptide NMR

The ROE cross-peaks for the assigned peptide NMR are shown below.

c(VPNRGD)

Conformation A: Asn and arginine couldn't be assigned

Coordinate 1	Assignment 1	Coordinate 2	Assignment 2
Val-Hga	0.57	Val-Hb	2.04
Val-Hgb	0.63	Val-Hb	2.04
Val-Hb	2.04	Val-Ha	4.38
Val-Ha	4.38	Pro-Hda	3.31
Val-Ha	4.38	Pro-Hdb	3.49
Val-Hb	2.04	Pro-Hda	3.31
Val-Hb	2.04	Pro-Hdb	3.49
Pro-Hda	3.31	Pro-Hdb	3.49
Pro-Hda	3.31	Pro-Hga	1.64
Pro-Hda	3.31	Pro-Hgb	1.81
Pro-Hdb	3.49	Pro-Hga	1.64
Pro-Hdb	3.49	Pro-Hgb	1.81
Gly-Haa	3.34	Gly-NH	8.59
Gly-Hab	3.76	Gly-NH	8.59
Gly-Hab	3.76	Asp-NH	8.38
Asp-NH	8.38	Asp-Ha	4.57
Asp-Ha	4.57	Asp-Hba	2.42
Asp-Ha	4.57	Asp-Hbb	2.60

Table 83: ROE cross-peaks for the major conformation of c(VPNRGD). The proline is in the trans conformation.

Conformation B:

Coordinate 1	Assignment 1	Coordinate 2	Assignment 2
Val-Hga	0.69	Val-Hb	1.76
Val-Hgb	0.76	Val-Hb	1.76
Val-Hb	1.76	Val-Ha	3.77
Val-Ha	3.77	Pro-Ha	4.35
Pro-Hda	3.30	Pro-Hdb	3.49
Pro-Hda	3.30	Pro-Hga	1.64
Pro-Hda	3.30	Pro-Hgb	1.81
Gly-Haa	3.43	Gly-NH	8.66
Gly-Hab	3.74	Gly-NH	8.66
Gly-Hab	3.74	Asp-NH	7.25
Gly-NH	8.66	Asp-NH	7.25
Asp-NH	7.25	Asp-Ha	4.42
Asp-Ha	4.42	Asp-Hba	2.39
Asp-Ha	4.42	Asp-Hbb	2.48
Asp-Ha	4.42	Val-H	8.24

Table 84: ROE cross-peaks for the minor conformation of c(VPNRGD). The proline is in the cis conformation.

c(VPRGDN)

Coordinate 1	Assignment 1	Coordinate 2	Assignment 2
Val-Hga	0.31	Val-Hb	1.95
Val-Hgb	0.61	Val-Hb	1.95
Val-Hb	1.95	Val-Ha	4.39
Val-Ha	4.39	Pro-Hdb	3.50
Val-Hb	1.95	Pro-Hda	3.28
Pro-Hda	1.82	Pro-Hdb	3.50
Pro-Hda	1.82	Pro-Hga	0.31
Pro-Hdb	1.64	Pro-Hgb	0.61
Pro-Hba	1.50	Pro-Ha	3.96
Pro-Hbb	2.01	Pro-Ha	3.96
Pro-Ha	3.96	Arg-NH	8.74
Arg-NH	8.74	Arg-Ha	3.67
Arg-Ha	3.67	Arg-Hb	1.73
Arg-Hb	1.73	Arg-Hg	1.25
Arg-Hg	1.25	Arg-Hda	2.87
Arg-Hg	1.25	Arg-Hdb	2.94
Arg-NH	8.74	Gly-NH	7.85
Arg-Ha	3.67	Gly-NH	7.85
Gly-Haa	3.24	Gly-NH	7.85
Gly-Hab	4.09	Gly-NH	7.85
Gly-Haa	3.24	Asp-NH	8.63
Asp-NH	8.63	Asp-Ha	4.39
Asp-Ha	4.39	Asp-Hba	2.44
Asp-Ha	4.39	Asp-Hbb	2.57
Asp-Ha	4.39	Asn-NH	8.74
Asn-NH	8.74	Asn-Ha	4.39
Asn-Ha	4.39	Asn-Hba	2.62
Asn-Ha	4.39	Asn-Hbb	2.71
Asn-NH	8.75	Val-NH	7.04
Asn-Ha	4.15	Val-NH	7.04

Table 85: Assigned ROE cross-peaks in the NMR of c(VPRGDN).

c(PWRNKY)

Coordinate 1	Assignment 1	Coordinate 2	Assignment 2
Pro-Hda	2.69	Pro-Hdb	2.93
Pro-Hda	2.69	Pro-Hga	0.04
Pro-Hdb	2.93	Pro-Hgb	1.18
Pro-Hda	2.69	Pro-Hgb	1.18
Pro-Hdb	2.93	Pro-Hga	0.04
Pro-Hga	0.04	Pro-Hba	0.79
Pro-Hba	0.79	Pro-Hbb	1.39
Pro-Hgb	1.18	Pro-Hbb	1.39
Pro-Hga	0.04	Pro-Hbb	1.39
Pro-Hba	0.79	Pro-Ha	3.06
Trp-NH	8.78	Trp-Ha	4.82
Trp-Ha	4.82	Trp-Hba	3.26
Trp-Ha	4.82	Trp-Hbb	3.59
Trp-Hba	3.26	Trp-Hbb	3.59
Trp-Hbb	3.59	Trp-Hd1	7.23
Trp-Hba	3.26	Trp-Hd1	7.23
Trp-Hd1	7.23	Trp-He1	10.17
Trp-Hz2	7.40	Trp-He3	7.15
Trp-Hz2	7.40	Trp-Hh2	7.07
Trp-Hh2	7.07	Trp-Hz3	7.69
Arg-Ha	3.98	Arg-Hba	1.76
Arg-Ha	3.98	Arg-Hbb	1.87
Arg-Hba	1.76	Arg-Hbb	1.87
Arg-Hba	1.76	Arg-Hga	1.57
Arg-Hba	1.76	Arg-Hgb	1.63
Arg-Hga	1.57	Arg-Hd	3.17
Arg-Hgb	1.63	Arg-Hd	3.17
Asn-Ha	4.56	Asn-Hb	2.83
Lys-NH	7.15	Lys-Ha	4.25
Lys-Ha	4.25	Lys-Hba	1.77
Lys-Ha	4.25	Lys-Hbb	1.83
Lys-Hba	1.77	Lys-Hg	1.40
Lys-Hbb	1.83	Lys-Hg	1.40
Lys-Hg	1.40	Lys-Hd	1.70
Lys-Hd	1.70	Lys-He	3.01
Tyr-Ha	4.30	Tyr-Hba	2.75
Tyr-Ha	4.30	Tyr-Hbb	3.10
Tyr-Hd	7.01	Tyr-He	6.74
Tyr-Ha	4.31	Pro-Ha	3.06

Table 86: ROE cross-peaks in the NMR of c(PWRNKY).

c(PWGNKE)

Conformation A:

Coordinate 1	Assignment 1	Coordinate 2	Assignment 2
Pro-Hda	3.04	Pro-Hdb	3.54
Trp-Ha	4.68	Trp-Hba	3.25
Trp-Ha	4.68	Trp-Hbb	3.48
Trp-Hba	3.25	Trp-Hbb	3.48
Gly-Haa	3.80	Gly-NH	7.33
Gly-Hab	4.13	Gly-NH	7.33
Gly-Haa	3.80	Asn-NH	9.00
Gly-Hab	4.13	Asn-NH	9.00
Asn-Ha	4.27	Asn-NH	9.00
Asn-Ha	4.27	Asn-Hba	2.76
Asn-Ha	4.27	Asn-Hbb	2.88
Asn-Ha	4.27	Lys-NH	8.52
Lys-NH	8.52	Lys-Ha	4.01
Lys-Ha	4.01	Lys-Hba	1.82
Lys-Ha	4.01	Lys-Hbb	1.89
Lys-Hba	1.82	Lys-Hg	0.57
Lys-Hbb	1.89	Lys-Hg	0.57
Lys-Hg	0.57	Lys-Hd	1.67
Lys-Hd	1.67	Lys-He	2.98
Lys-Ha	4.02	Glu-NH	7.55
Lys-NH	8.52	Glu-NH	7.55
Glu-NH	7.55	Glu-Ha	4.51
Glu-Ha	4.50	Pro-Hda	3.04
Glu-Ha	4.50	Pro-Hdb	3.54

Table 87: ROE cross-peaks of the major conformation of c(PWGNKE). The conformation contains trans proline.

Conformation B:

Coordinate 1	Assignment 1	Coordinate 2	Assignment 2
Pro-Hda	2.91	Pro-Hdb	3.18
Pro-Hda	2.91	Pro-Hga	1.28
Pro-Hda	2.91	Pro-Hgb	1.34
Pro-Hga	1.28	Pro-Hba	1.82
Pro-Hba	1.82	Pro-Hbb	1.89
Pro-Hba	1.82	Pro-Ha	4.56
Pro-Hbb	1.89	Pro-Ha	4.56
Trp-Ha	4.86	Trp-Hba	3.26
Trp-Ha	4.86	Trp-Hbb	3.36
Trp-Hba	3.26	Trp-Hbb	3.36
Gly-Haa	3.60	Gly-NH	8.29
Gly-Hab	3.93	Gly-NH	8.29
Gly-Haa	3.60	Asn-NH	8.75
Gly-Hab	3.93	Asn-NH	8.75
Asn-Ha	4.39	Asn-NH	8.75
Asn-Ha	4.39	Asn-Hba	2.70
Asn-Ha	4.39	Asn-Hbb	2.78
Asn-Ha	4.39	Lys-NH	8.55
Lys-NH	8.56	Lys-Ha	4.11
Lys-Ha	4.11	Lys-Hba	1.85
Lys-Ha	4.11	Lys-Hbb	1.91
Lys-Hba	1.85	Lys-Hga	1.28
Lys-Hbb	1.91	Lys-Hgb	1.34
Lys-Hba	1.85	Lys-Hgb	1.34
Lys-Hbb	1.91	Lys-Hga	1.28
Lys-Hga	1.28	Lys-Hda	1.56
Lys-Hga	1.28	Lys-Hdb	1.60
Lys-Hda	1.56	Lys-Hea	2.90
Lys-Hdb	1.60	Lys-Heb	2.97
Lys-Ha	4.10	Glu-NH	8.21
Glu-NH	8.21	Glu-Ha	4.87

Table 88: ROE cross-peaks seen in the minor conformation of c(PWGNKE). It is likely the proline is present as cis proline but the necessary cross-peak is masked by water.

c(PWGNKY)

Conformation A:

Coordinate 1	Assignment 1	Coordinate 2	Assignment 2
Pro-Hda	2.70	Pro-Hdb	2.95
Pro-Hda	2.70	Pro-Hga	0.18
Pro-Hdb	2.95	Pro-Hgb	1.21
Pro-Hdb	2.95	Pro-Hga	0.18
Pro-Hga	0.18	Pro-Hba	0.65
Pro-Hba	0.65	Pro-Ha	3.31
Pro-Hbb	1.37	Pro-Ha	3.31
Trp-Ha	4.79	Trp-Hba	3.20
Trp-Ha	4.79	Trp-Hbb	3.31
Trp-Hba	3.20	Trp-Hbb	3.31
Trp-Hbb	3.32	Trp-Hd1	7.16
Trp-Hba	3.20	Trp-Hd1	7.16
Trp-Hd1	7.16	Trp-He1	10.13
Gly-Haa	3.57	Gly-NH	8.51
Gly-Hab	3.90	Gly-NH	8.51
Gly-Haa	3.57	Asn-NH	9.04
Gly-Hab	3.90	Asn-NH	9.04
Asn-NH	9.04	Asn-Ha	4.16
Asn-Ha	4.16	Asn-Hba	2.83
Asn-Ha	4.16	Asn-Hbb	2.92
Asn-Ha	4.16	Lys-NH	7.46
Lys-NH	7.46	Lys-Ha	4.38
Lys-Ha	4.38	Lys-Hba	1.73
Lys-Ha	4.38	Lys-Hbb	1.77
Lys-Hba	1.73	Lys-Hg	1.36
Lys-Hbb	1.77	Lys-Hg	1.36
Lys-Hg	1.36	Lys-Hd	1.68
Lys-Hd	1.68	Lys-He	2.98
Lys-Ha	4.38	Tyr-NH	8.40
Tyr-NH	8.40	Tyr-Ha	4.29
Tyr-Ha	4.29	Tyr-Hba	2.70
Tyr-Ha	4.29	Tyr-Hbb	3.06
Tyr-Hd	7.00	Tyr-He	6.73
Tyr-Ha	4.29	Pro-Ha	3.31

Table 89: ROE cross-peaks for the major conformation of c(PWGNKY). The conformation contains cis proline.

Conformation B:

Coordinate 1	Assignment 1	Coordinate 2	Assignment 2
Pro-Hda	3.11	Pro-Hdb	3.67
Pro-Hda	3.11	Pro-Hga	1.89
Pro-Hdb	3.67	Pro-Hgb	2.13
Pro-Hda	3.11	Pro-Hgb	2.13
Pro-Hdb	3.67	Pro-Hga	1.89
Pro-Hga	1.89	Pro-Hba	1.90
Pro-Hba	1.36	Pro-Hbb	2.32
Pro-Hgb	2.13	Pro-Hbb	2.32
Pro-Hga	1.89	Pro-Hbb	2.32
Pro-Hba	1.91	Pro-Ha	4.06
Pro-Hbb	2.32	Pro-Ha	4.06
Pro-Ha	4.06	Trp-NH	6.33
Trp-NH	6.33	Trp-Ha	4.06
Trp-Ha	4.06	Trp-Hba	3.24
Trp-Ha	4.06	Trp-Hbb	3.50
Trp-Hba	3.24	Trp-Hbb	3.50
Trp-Hbb	3.50	Trp-Hd1	7.10
Trp-Hba	3.25	Trp-Hd1	7.10
Trp-Hd1	7.10	Trp-He1	10.15
Trp-NH	6.33	Gly-NH	7.30
Gly-Haa	3.72	Gly-NH	7.30
Gly-Hab	4.21	Gly-NH	7.30
Asn-Ha	4.45	Asn-Hba	2.77
Asn-Ha	4.45	Asn-Hbb	2.80
Asn-Ha	4.45	Lys-NH	8.17
Lys-NH	8.17	Lys-Ha	3.90
Lys-Ha	3.90	Lys-Hba	2.77
Lys-Ha	3.90	Lys-Hbb	1.48
Lys-Hba	2.77	Lys-Hga	0.96
Lys-Hbb	1.48	Lys-Hgb	1.04
Lys-Hba	2.77	Lys-Hgb	1.04
Lys-Hbb	1.48	Lys-Hga	0.96
Lys-Hga	0.96	Lys-Hd	1.43
Lys-Hgb	1.04	Lys-Hd	1.43
Lys-Hd	1.43	Lys-He	2.79
Lys-Ha	3.90	Tyr-NH	7.66
Tyr-NH	7.66	Tyr-Ha	4.42
Tyr-Ha	4.42	Tyr-Hba	1.36
Tyr-Ha	4.42	Tyr-Hbb	1.45
Tyr-Hd	6.91	Tyr-He	6.80
Tyr-Ha	4.42	Pro-Hdb	3.67

Table 90: ROE cross-peaks assigned for the minor conformation of c(PWGNKY). Proline in this conformation has a trans amide bond.

c(PWRNKE)

Conformation A:

Coordinate 1	Assignment 1	Coordinate 2	Assignment 2
Pro-Hda	2.80	Pro-Hdb	3.08
Pro-Hda	2.80	Pro-Hga	0.17
Pro-Hdb	3.08	Pro-Hgb	1.40
Pro-Hda	2.80	Pro-Hgb	1.40
Pro-Hdb	3.08	Pro-Hga	0.17
Pro-Hga	0.17	Pro-Hb	1.81
Pro-Hgb	1.40	Pro-Hb	1.81
Pro-Hb	1.81	Pro-Ha	4.45
Trp-Ha	4.88	Trp-Hba	3.28
Trp-Ha	4.88	Trp-Hbb	3.60
Trp-Hba	3.28	Trp-Hbb	3.60
Trp-Hbb	3.60	Trp-Hd1	7.26
Trp-Hba	3.28	Trp-Hd1	7.26
Trp-Hd1	7.26	Trp-He1	10.21
Arg-NH	8.60	Arg-Ha	3.95
Arg-Ha	3.95	Arg-Hba	1.74
Arg-Ha	3.95	Arg-Hbb	1.84
Arg-Hba	1.74	Arg-Hbb	1.84
Asn-NH	8.74	Asn-Ha	4.50
Asn-Ha	4.50	Asn-Hb	2.80
Asn-Ha	4.50	Lys-NH	7.18
Asn-NH	8.74	Lys-NH	7.18
Lys-NH	7.18	Lys-Ha	4.20
Lys-Ha	4.20	Lys-Hba	1.74
Lys-Ha	4.20	Lys-Hbb	1.80
Lys-Hba	1.74	Lys-Hg	1.36
Lys-Hbb	1.80	Lys-Hg	1.36
Lys-Hg	1.36	Lys-Hd	1.66
Lys-Hd	1.66	Lys-He	2.96
Lys-Ha	4.20	Glu-NH	9.12
Glu-NH	9.12	Glu-Ha	4.22
Glu-Ha	4.22	Glu-Hb	2.87
Glu-Ha	4.22	Pro-Ha	4.45
Pro-Hda	2.80	Trp-Hh2	7.18
Pro-Hda	2.80	Trp-H1	7.26

Table 91: c(PWRNKE) ROE cross-peaks for the major conformation. The conformation contains a cis proline.

Conformation B:

Asn not assigned along with Arg, Lys sidechains

Coordinate 1	Assignment 1	Coordinate 2	Assignment 2
Pro-Hda	3.24	Pro-Hdb	3.60
Pro-Hda	3.24	Pro-Hga	1.83
Pro-Hdb	3.60	Pro-Hgb	2.04
Pro-Hda	3.24	Pro-Hgb	2.04
Pro-Hdb	3.60	Pro-Hga	1.83
Pro-Hgb	2.04	Pro-Hb	2.09
Pro-Hb	2.09	Pro-Ha	3.93
Trp-NH	7.04	Trp-Ha	4.56
Trp-Ha	4.56	Trp-Hb	3.33
Trp-Hb	3.33	Trp-Hd1	7.09
Trp-Hd1	7.09	Trp-He1	10.27
Trp-Ha	4.56	Arg-NH	7.61
Lys-NH	8.40	Lys-Ha	4.13
Lys-Ha	4.13	Lys-Hba	1.74
Lys-Ha	4.13	Lys-Hbb	1.90
Lys-NH	8.40	Glu-NH	7.83
Glu-NH	7.83	Glu-Ha	4.50
Glu-Ha	4.50	Glu-Hba	1.11
Glu-Ha	4.50	Glu-Hbb	1.20
Glu-Hba	1.11	Glu-Hg	2.22
Glu-Hbb	1.20	Glu-Hg	2.22
Glu-Ha	4.50	Pro-Hdb	3.60

Table 92: ROE cross-peaks for the minor conformation of c(PWRNKE). Contains trans proline.

11 References

1. Y. Hamada and T. Shioiri, *Chemical Reviews*, 2005, **105**, 4441-4482.
2. N.-H. Tan and J. Zhou, *Chemical Reviews*, 2006, **106**, 840-895.
3. A. Zorzi, K. Deyle and C. Heinis, *Current Opinion in Chemical Biology*, 2017, **38**, 24-29.
4. G. Weckbecker, I. Lewis, R. Albert, H. A. Schmid, D. Hoyer and C. Bruns, *Nature Reviews Drug Discovery*, 2003, **2**, 999-1017.
5. A. Paananen, K. Järvinen, T. Sareneva, M. S. Salkinoja-Salonen, T. Timonen and E. Hölttä, *Toxicology*, 2005, **212**, 37-45.
6. V. J. Thombare and C. A. Hutton, *Angewandte Chemie International Edition*, 2019, **58**, 4998-5002.
7. F. Giordanetto and J. Kihlberg, *Journal of Medicinal Chemistry*, 2014, **57**, 278-295.
8. T. A. F. Cardote and A. Ciulli, *ChemMedChem*, 2016, **11**, 787-794.
9. D. Leenheer, P. ten Dijke and C. J. Hipolito, *Peptide Science*, 2016, **106**, 889-900.
10. A. K. Yudin, *Chemical Science*, 2015, **6**, 30-49.
11. T. A. Hill, N. E. Shepherd, F. Diness and D. P. Fairlie, *Angewandte Chemie International Edition*, 2014, **53**, 13020-13041.
12. W.-H. Hsieh and J. Liaw, *Journal of Food and Drug Analysis*, 2019, **27**, 32-47.
13. G. Hummel, U. Reineke and U. Reimer, *Molecular BioSystems*, 2006, **2**, 499-508.
14. P. S. Cremer, A. H. Flood, B. C. Gibb and D. L. Mobley, *Nature Chemistry*, 2017, **10**, 8.
15. R. M. Cusack, L. Grøndahl, D. P. Fairlie, L. R. Gahan and G. R. Hanson, *Journal of the Chemical Society, Perkin Transactions **2***, 2002, 556-563.
16. R. Tugyi, G. Mezö, E. Fellinger, D. Andreu and F. Hudecz, *Journal of Peptide Science*, 2005, **11**, 642-649.
17. K. Shibata, T. Suzawa, S. Soga, T. Mizukami, K. Yamada, N. Hanai and M. Yamasaki, *Bioorganic & Medicinal Chemistry Letters*, 2003, **13**, 2583-2586.
18. D. Wang, W. Liao and P. S. Arora, *Angewandte Chemie International Edition*, 2005, **44**, 6525-6529.
19. M. Dathe, H. Nikolenko, J. Klose and M. Bienert, *Biochemistry*, 2004, **43**, 9140-9150.
20. D. P. Fairlie, J. D. A. Tyndall, R. C. Reid, A. K. Wong, G. Abbenante, M. J. Scanlon, D. R. March, D. A. Bergman, C. L. L. Chai and B. A. Burkett, *Journal of Medicinal Chemistry*, 2000, **43**, 1271-1281.
21. J. Mallinson and I. Collins, *Future Medicinal Chemistry*, 2012, **4**, 1409-1438.
22. H. Yu and Y.-S. Lin, *Physical Chemistry Chemical Physics*, 2015, **17**, 4210-4219.
23. R. González-Muñiz, M. Á. Bonache and M. J. Pérez de Vega, *Molecules*, 2021, **26**, 445.
24. A. A. Vinogradov, Y. Yin and H. Suga, *Journal of the American Chemical Society*, 2019, **141**, 4167-4181.
25. S. M. McHugh, J. R. Rogers, S. A. Solomon, H. Yu and Y.-S. Lin, *Current Opinion in Chemical Biology*, 2016, **34**, 95-102.
26. A. E. Wakefield, W. M. Wuest and V. A. Voelz, *Journal of Chemical Information and Modeling*, 2015, **55**, 806-813.
27. Y. Zhu, M. Tong, C. Liu, C. Song, D. Wei, Q. Zhao and M. Tang, *Computational and Theoretical Chemistry*, 2014, **1027**, 46-52.
28. R. O. Hynes, *Cell*, 2002, **110**, 673-687.
29. J. D. Humphries, A. Byron and M. J. Humphries, *Journal of Cell Science*, 2006, **119**, 3901-3903.
30. D. Cox, M. Brennan and N. Moran, *Nature Reviews Drug Discovery*, 2010, **9**, 804-820.
31. J. S. Desgrosellier and D. A. Cheresh, *Nature Reviews Cancer*, 2010, **10**, 9-22.
32. H. Jin and J. Varner, *Br J Cancer*, 2004, **90**, 561-565.
33. C. J. Avraamides, B. Garmy-Susini and J. A. Varner, *Nature Reviews Cancer*, 2008, **8**, 604-617.
34. S. Raab-Westphal, J. F. Marshall and S. L. Goodman, *Cancers*, 2017, **9**, 110-136.
35. C.-C. Sun, X.-J. Qu and Z.-H. Gao, *American Journal of Therapeutics*, 2016, **23**, 198-207.

36. K. Ley, J. Rivera-Nieves, W. J. Sandborn and S. Shattil, *Nature Reviews Drug Discovery*, 2016, **15**, 173-183.
37. C. Mas-Moruno, F. Rechenmacher and H. Kessler, *Anticancer Agents Med Chem*, 2010, **10**, 753-768.
38. D. A. Reardon and D. Cheresh, *Genes & Cancer*, 2011, **2**, 1159-1165.
39. R. Stupp, M. E. Hegi, T. Gorlia, S. C. Erridge, J. Perry, Y.-K. Hong, K. D. Aldape, B. Lhermitte, T. Pietsch, D. Grujicic, J. P. Steinbach, W. Wick, R. Tarnawski, D.-H. Nam, P. Hau, A. Weyerbrock, M. J. B. Taphoorn, C.-C. Shen, N. Rao, L. Thurzo, U. Herrlinger, T. Gupta, R.-D. Kortmann, K. Adamska, C. McBain, A. A. Brandes, J. C. Tonn, O. Schnell, T. Wiegel, C.-Y. Kim, L. B. Nabors, D. A. Reardon, M. J. van den Bent, C. Hicking, A. Markivskyy, M. Picard and M. Weller, *The Lancet Oncology*, 2014, **15**, 1100-1108.
40. T. G. Kapp, F. Rechenmacher, S. Neubauer, O. V. Maltsev, E. A. Cavalcanti-Adam, R. Zarka, U. Reuning, J. Notni, H.-J. Wester, C. Mas-Moruno, J. Spatz, B. Geiger and H. Kessler, *Scientific reports*, 2017, **7**, 39805-39815.
41. T. Weide, A. Modlinger and H. Kessler, in *Bioactive Conformation I*, ed. T. Peters, Springer Berlin Heidelberg, Berlin, Heidelberg, 2007, pp. 1-50.
42. G. Müller, M. Gurrath and H. Kessler, *Journal of Computer-Aided Molecular Design*, 1994, **8**, 709-730.
43. M. Aumailley, M. Gurrath, G. Müller, J. Calvete, R. Timpl and H. Kessler, *FEBS Letters*, 1991, **291**, 50-54.
44. M. Pfaff, K. Tangemann, B. Müller, M. Gurrath, G. Müller, H. Kessler, R. Timpl and J. Engel, *Journal of Biological Chemistry*, 1994, **269**, 20233-20238.
45. H. Kessler, R. Gratias, G. Hessler, M. Gurrath and G. Müller, *Pure and Applied Chemistry*, 1996, **68**, 1201-1205.
46. R. Jwad, D. Weissberger and L. Hunter, *Chemical Reviews*, 2020, **120**, 9743-9789.
47. J.-P. Xiong, T. Stehle, R. Zhang, A. Joachimiak, M. Frech, S. L. Goodman and M. A. Arnaout, *Science*, 2002, **296**, 151.
48. U. K. Marelli, A. O. Frank, B. Wahl, V. La Pietra, E. Novellino, L. Marinelli, E. Herdtweck, M. Groll and H. Kessler, *Chemistry – A European Journal*, 2014, **20**, 14201-14206.
49. J. Zhu, B.-H. Luo, T. Xiao, C. Zhang, N. Nishida and T. A. Springer, *Molecular cell*, 2008, **32**, 849-861.
50. R. Haubner, R. Gratias, B. Diefenbach, S. L. Goodman, A. Jonczyk and H. Kessler, *Journal of the American Chemical Society*, 1996, **118**, 7461-7472.
51. R. Haubner, D. Finsinger and H. Kessler, *Angewandte Chemie International Edition in English*, 1997, **36**, 1374-1389.
52. M. A. Dechantsreiter, E. Planker, B. Mathä, E. Lohof, G. Hölzemann, A. Jonczyk, S. L. Goodman and H. Kessler, *Journal of Medicinal Chemistry*, 1999, **42**, 3033-3040.
53. J. Chatterjee, D. F. Mierke and H. Kessler, *Chemistry – A European Journal*, 2008, **14**, 1508-1517.
54. J. Chatterjee, C. Gilon, A. Hoffman and H. Kessler, *Accounts of Chemical Research*, 2008, **41**, 1331-1342.
55. P. Lahiri, H. Verma, A. Ravikumar and J. Chatterjee, *Chemical Science*, 2018, **9**, 4600-4609.
56. J. Damjanovic, J. Miao, H. Huang and Y.-S. Lin, *Chemical Reviews*, 2021, **121**, 2292-2324.
57. C. J. White and A. K. Yudin, *Nature Chemistry*, 2011, **3**, 509.
58. L. M. De Leon Rodriguez, A. J. Weidkamp and M. A. Brimble, *Organic & Biomolecular Chemistry*, 2015, **13**, 6906-6921.
59. H. C. Hayes, L. Y. P. Luk and Y.-H. Tsai, *Organic & biomolecular chemistry*, 2021, **19**, 3983-4001.
60. D. S. Kemp and J. Rebek, *Journal of the American Chemical Society*, 1970, **92**, 5792-5793.
61. D. Besser, S. Reissmann, R. Olender, R. Rosenfeld and O. Arad, *The Journal of Peptide Research*, 2000, **56**, 337-345.

62. M. El Haddadi, F. Cavelier, E. Vives, A. Azmani, J. Verducci and J. Martinez, *Journal of Peptide Science*, 2000, **6**, 560-570.
63. A. Thakkar, T. B. Trinh and D. Pei, *ACS Combinatorial Science*, 2013, **15**, 120-129.
64. J. M. Humphrey and A. R. Chamberlin, *Chemical Reviews*, 1997, **97**, 2243-2266.
65. J. Blankenstein and J. Zhu, *European Journal of Organic Chemistry*, 2005, 1949-1964.
66. R. Chapman, K. A. Jolliffe and S. Perrier, *Advanced Materials*, 2013, **25**, 1170-1172.
67. R. J. Brea, C. Reiriz and J. R. Granja, *Chemical Society Reviews*, 2010, **39**, 1448-1456.
68. R. Chapman, M. Danial, M. L. Koh, K. A. Jolliffe and S. Perrier, *Chemical Society Reviews*, 2012, **41**, 6023-6041.
69. F. Cavelier-Frontin, S. Achmad, J. Verducci, R. Jacquier and G. Pèpe, *Journal of Molecular Structure: THEOCHEM*, 1993, **286**, 125-130.
70. D. P. Slough, H. Yu, S. M. McHugh and Y.-S. Lin, *Physical Chemistry Chemical Physics*, 2017, **19**, 5377-5388.
71. S. A. Kates, N. A. Solé, C. R. Johnson, D. Hudson, G. Barany and F. Albericio, *Tetrahedron Letters*, 1993, **34**, 1549-1552.
72. C. Bechtler and C. Lamers, *RSC Medicinal Chemistry*, 2021, **12**, 1325-1351.
73. W.-L. Xu, A. L. Cui, X.-X. Hu, X.-F. You, Z.-R. Li and J.-S. Zheng, *Tetrahedron Letters*, 2015, **56**, 4796-4799.
74. E. Peterse, N. Meeuwenoord, H. van den Elst, G. A. van der Marel, H. S. Overkleeft and D. V. Filippov, *European Journal of Organic Chemistry*, 2022, e202101341.
75. C. Cabrele, M. Langer and A. G. Beck-Sickinger, *The Journal of Organic Chemistry*, 1999, **64**, 4353-4361.
76. U. Boas, J. Brask and K. J. Jensen, *Chemical Reviews*, 2009, **109**, 2092-2118.
77. V. Agouridas, O. El Mahdi, V. Diemer, M. Cargoët, J.-C. M. Monbaliu and O. Melnyk, *Chemical Reviews*, 2019, **119**, 7328-7443.
78. Q. Wan and S. J. Danishefsky, *Angewandte Chemie International Edition*, 2007, **46**, 9248-9252.
79. H. Y. Chow, Y. Zhang, E. Matheson and X. Li, *Chemical Reviews*, 2019, **119**, 9971-10001.
80. S. Di Maro, A. M. Trotta, D. Brancaccio, F. S. Di Leva, V. La Pietra, C. Ieranò, M. Napolitano, L. Portella, C. D'Alterio, R. A. Siciliano, D. Sementa, S. Tomassi, A. Carotenuto, E. Novellino, S. Scala and L. Marinelli, *Journal of Medicinal Chemistry*, 2016, **59**, 8369-8380.
81. C. G. Joseph, X. S. Wang, J. W. Scott, R. M. Bauzo, Z. Xiang, N. G. Richards and C. Haskell-Luevano, *Journal of Medicinal Chemistry*, 2004, **47**, 6702-6710.
82. B. K. Yap, E. W. W. Leung, H. Yagi, C. A. Galea, S. Chhabra, D. K. Chalmers, S. E. Nicholson, P. E. Thompson and R. S. Norton, *Journal of Medicinal Chemistry*, 2014, **57**, 7006-7015.
83. R. A. Turner, A. G. Oliver and R. S. Lokey, *Organic Letters*, 2007, **9**, 5011-5014.
84. V. D. Bock, R. Perciaccante, T. P. Jansen, H. Hiemstra and J. H. van Maarseveen, *Organic Letters*, 2006, **8**, 919-922.
85. P. J. LeValley, E. M. Ovadia, C. A. Bresette, L. A. Sawicki, E. Maverakis, S. Bai and A. M. Kloxin, *Chemical Communications*, 2018, **54**, 6923-6926.
86. A. Stefanucci, W. Lei, S. Pieretti, E. Novellino, M. P. Dimmito, F. Marzoli, J. M. Streicher and A. Mollica, *Scientific Reports*, 2019, **9**, 5771.
87. C. Cai, F. Wang, X. Xiao, W. Sheng, S. Liu, J. Chen, J. Zheng, R. Xie, Z. Bai and H. Wang, *Chemical Communications*, 2022, **58**, 4861-4864.
88. Z. Bai, C. Cai, Z. Yu and H. Wang, *Angewandte Chemie International Edition*, 2018, **57**, 13912-13916.
89. J. Liu, P. Wang, Z. Yan, J. Yan, Kenry and Q. Zhu, *ChemBioChem*, 2021, **22**, 2762-2771.
90. D. K. Kölmel and E. T. Kool, *Chemical Reviews*, 2017, **117**, 10358-10376.
91. T. D. Pallin and J. P. Tam, *Journal of the Chemical Society, Chemical Communications*, 1995, 2021-2022.

92. K. D. Roberts, J. N. Lambert, N. J. Ede and A. M. Bray, *Journal of Peptide Science*, 2004, **10**, 659-665.
93. D. D. Sternbach and W. C. L. Jamison, *Tetrahedron Letters*, 1981, **22**, 3331-3334.
94. M. Fujita, H. Oishi and T. Hiyama, *Chemistry Letters*, 1986, **15**, 837-838.
95. O. Avrutina, H.-U. Schmoldt, D. Gabrijelcic-Geiger, A. Wentzel, H. Frauendorf, C. P. Sommerhoff, U. Diederichsen and H. Kolmar, *ChemBioChem*, 2008, **9**, 33-37.
96. S. J. de Veer, M.-W. Kan and D. J. Craik, *Chemical Reviews*, 2019, **119**, 12375-12421.
97. L. Albert and O. Vázquez, *Chemical Communications*, 2019, **55**, 10192-10213.
98. O. Nwajiobi, A. K. Verma and M. Raj, *Journal of the American Chemical Society*, 2022, **144**, 4633-4641.
99. K. Bozovičar and T. Bratkovič, *International Journal of Molecular Sciences*, 2021, **22**, 1611.
100. Y. Wu, H.-F. Chau, W. Thor, K. H. Y. Chan, X. Ma, W.-L. Chan, N. J. Long and K.-L. Wong, *Angewandte Chemie International Edition*, 2021, **60**, 20301-20307.
101. L. Mendive-Tapia, J. Wang and M. Vendrell, *Peptide Science*, 2021, **113**, e24181.
102. J. Liu, X. Liu, F. Zhang, J. Qu, H. Sun and Q. Zhu, *Chemistry – A European Journal*, 2020, **26**, 16122-16128.
103. M. Todorovic, K. D. Schwab, J. Zeisler, C. Zhang, F. Bénard and D. M. Perrin, *Angewandte Chemie International Edition*, 2019, **58**, 14120-14124.
104. Y. Zhang, Q. Zhang, C. T. T. Wong and X. Li, *Journal of the American Chemical Society*, 2019, **141**, 12274-12279.
105. S. A. Hollingsworth and P. A. Karplus, *Biomol Concepts*, 2010, **1**, 271-283.
106. S. Gupta and Y. U. Sasidhar, *The Journal of Physical Chemistry B*, 2017, **121**, 1268-1283.
107. S. R. Griffiths-Jones, A. J. Maynard and G. J. Sharman, *Chemical Communications*, 1998, 789-790.
108. T. Sharpe, A. L. Jonsson, T. J. Rutherford, V. Daggett and A. R. Fersht, *Protein Science*, 2007, **16**, 2233-2239.
109. N. N. W. Kuo, J. J. T. Huang, J. Miksovska, R. P. Y. Chen, R. W. Larsen and S. I. Chan, *Journal of the American Chemical Society*, 2005, **127**, 16945-16954.
110. K. R. Rajashankar and S. Ramakumar, *Protein Science*, 1996, **5**, 932-946.
111. K. Guruprasad and S. Rajkumar, *Journal of Biosciences*, 2000, **25**, 143-156.
112. G. Ruiz-Gómez, J. D. A. Tyndall, B. Pfeiffer, G. Abbenante and D. P. Fairlie, *Chemical Reviews*, 2010, **110**, PR1-PR41.
113. M. W. Franklin and J. S. G. Slusky, *Journal of molecular biology*, 2018, **430**, 3251-3265.
114. E. G. Hutchinson and J. M. Thornton, *Protein Science*, 1994, **3**, 2207-2216.
115. K.-C. Chou, *Analytical Biochemistry*, 2000, **286**, 1-16.
116. H. Santa, M. Ylisirniö, T. Hassinen, R. Laatikainen and M. Peräkylä, *Protein Engineering, Design and Selection*, 2002, **15**, 651-657.
117. P. F. J. Fuchs, A. M. J. J. Bonvin, B. Bochicchio, A. Pepe, A. J. P. Alix and A. M. Tamburro, *Biophysical journal*, 2006, **90**, 2745-2759.
118. C. M. Venkatachalam, *Biopolymers*, 1968, **6**, 1425-1436.
119. J. S. Richardson, *Advances in Protein Chemistry*, 1981, **34**, 167-339.
120. J. L. Crawford, W. N. Lipscomb and C. G. Schellman, *Proceedings of the National Academy of Sciences*, 1973, **70**, 538-542.
121. P. N. Lewis, F. A. Momany and H. A. Scheraga, *Biochimica et Biophysica Acta (BBA) - Protein Structure*, 1973, **303**, 211-229.
122. A. W. Chan, E. G. Hutchinson, D. Harris and J. M. Thornton, *Protein Science*, 1993, **2**, 1574-1590.
123. A. G. de Brevern, *Scientific Reports*, 2016, **6**, 33191.
124. M. Shapovalov, S. Vucetic and R. L. Dunbrack, Jr., *PLOS Computational Biology*, 2019, **15**, e1006844.

125. A. Ram, J. Sunita, A. Jalal and K. Manoj, *International Journal of Computer Applications*, 2010, **3**, 1-4.
126. X. Jin and J. Han, in *Encyclopedia of Machine Learning*, eds. C. Sammut and G. I. Webb, Springer US, Boston, MA, 2010, pp. 564-565.
127. C. M. Wilmot and J. M. Thornton, *Journal of Molecular Biology*, 1988, **203**, 221-232.
128. P. Y. Chou and G. D. Fasman, *Biophysical journal*, 1979, **26**, 367-383.
129. P. Y. Chou and G. D. Fasman, *Biochemistry*, 1974, **13**, 211-222.
130. G. D. Rose, L. M. Glerasch and J. A. Smith, in *Advances in Protein Chemistry*, eds. C. B. Anfinsen, J. T. Edsall and F. M. Richards, Academic Press, 1985, vol. 37, pp. 1-109.
131. K.-C. Chou and J. R. Blinn, *Journal of Protein Chemistry*, 1997, **16**, 575-595.
132. P. F. J. Fuchs and A. J. P. Alix, *Proteins: Structure, Function, and Bioinformatics*, 2005, **59**, 828-839.
133. A. Kirschner and D. Frishman, *Gene*, 2008, **422**, 22-29.
134. H. Kaur and G. P. S. Raghava, *Bioinformatics*, 2004, **20**, 2751-2758.
135. P. Kountouris and J. D. Hirst, *BMC Bioinformatics*, 2010, **11**, 407-407.
136. B. Petersen, C. Lundegaard and T. N. Petersen, *PLoS One*, 2010, **5**, e15079.
137. F. Emmert-Streib, Z. Yang, H. Feng, S. Tripathi and M. Dehmer, *Frontiers in Artificial Intelligence*, 2020, **3**.
138. M. J. McGregor, T. P. Flores and M. J. E. Sternberg, *Protein Engineering, Design and Selection*, 1989, **2**, 521-526.
139. C. Fang, Y. Shang and D. Xu, *Proteins*, 2020, **88**, 143-151.
140. Y.-D. Cai, X.-J. Liu, Y.-X. Li, X.-b. Xu and K.-C. Chou, *Peptides*, 2003, **24**, 665-669.
141. T. Pham, K. Satou and T. Ho, *Genome informatics. International Conference on Genome Informatics*, 2003, **14**, 196-205.
142. C. Zheng and L. Kurgan, *BMC Bioinformatics*, 2008, **9**, 430.
143. J. Cervantes, F. Garcia-Lamont, L. Rodríguez-Mazahua and A. Lopez, *Neurocomputing*, 2020, **408**, 189-215.
144. P. Kountouris and J. D. Hirst, *BMC Bioinformatics*, 2009, **10**, 437.
145. L. Nguyen, X. Dang, T. Le, T. Saethang, V. Tran, D. Ngo, S. Gavrilov, N. Ngoc Giang, M. Kubo, Y. Yamada and K. Satou, *Journal of Biomedical Science and Engineering*, 2014, **07**, 927-940.
146. H. Singh, S. Singh and G. P. S. Raghava, *Proteins: Structure, Function, and Bioinformatics*, 2015, **83**, 910-921.
147. A. M. C. Marcelino and L. M. Gierasch, *Biopolymers*, 2008, **89**, 380-391.
148. J. Venkatraman, S. C. Shankaramma and P. Balaram, *Chemical Reviews*, 2001, **101**, 3131-3152.
149. O. Khakshoor and J. S. Nowick, *Current opinion in chemical biology*, 2008, **12**, 722-729.
150. R. V. Nair, S. B. Baravkar, T. S. Ingole and G. J. Sanjayan, *Chemical Communications*, 2014, **50**, 13874-13884.
151. N. Srinivas, K. Moehle, K. Abou-Hadeed, D. Obrecht and J. A. Robinson, *Organic & Biomolecular Chemistry*, 2007, **5**, 3100-3105.
152. S. C. Shankaramma, Z. Athanassiou, O. Zerbe, K. Moehle, C. Mouton, F. Bernardini, J. W. Vrijbloed, D. Obrecht and J. A. Robinson, *ChemBioChem*, 2002, **3**, 1126-1133.
153. M. Pelay-Gimeno, A. Glas, O. Koch and T. N. Grossmann, *Angewandte Chemie International Edition*, 2015, **54**, 8896-8927.
154. L. R. Whitby and D. L. Boger, *Accounts of chemical research*, 2012, **45**, 1698-1709.
155. R. Fasan, R. L. A. Dias, K. Moehle, O. Zerbe, J. W. Vrijbloed, D. Obrecht and J. A. Robinson, *Angewandte Chemie International Edition*, 2004, **43**, 2109-2112.
156. S. Deike, S. Rothemund, B. Voigt, S. Samantray, B. Strodel and W. H. Binder, *Bioorganic Chemistry*, 2020, **101**, 104012.
157. C. Kim, J. Jung, T. T. Tung and S. B. Park, *Chemical Science*, 2016, **7**, 2753-2761.

158. B. Laufer, J. Chatterjee, A. O. Frank and H. Kessler, *Journal of Peptide Science*, 2009, **15**, 141-146.
159. T. S. Haque, J. C. Little and S. H. Gellman, *Journal of the American Chemical Society*, 1996, **118**, 6975-6985.
160. D. F. Veber, R. M. Freidinger, D. S. Perlow, W. J. Paleveda, F. W. Holly, R. G. Strachan, R. F. Nutt, B. H. Arison, C. Homnick, W. C. Randall, M. S. Glitzer, R. Saperstein and R. Hirschmann, *Nature*, 1981, **292**, 55-58.
161. I. L. Karle, S. K. Awasthi and P. Balaram, *Proceedings of the National Academy of Sciences*, 1996, **93**, 8189-8193.
162. C. M. Nair, M. Vijayan, Y. V. Venkatachalapathi and P. Balaram, *Journal of the Chemical Society, Chemical Communications*, 1979, 1183-1184.
163. J. W. Bean, K. D. Kopple and C. E. Peishoff, *Journal of the American Chemical Society*, 1992, **114**, 5328-5334.
164. J. Wen, H. Liao, K. Stachowski, J. P. Hempfling, Z. Qian, C. Yuan, M. P. Foster and D. Pei, *Bioorganic & Medicinal Chemistry*, 2020, **28**, 115711.
165. J. A. Robinson, *Accounts of Chemical Research*, 2008, **41**, 1278-1288.
166. I. L. Batalha, I. Lychko, R. J. F. Branco, O. Iranzo and A. C. A. Roque, *Organic & Biomolecular Chemistry*, 2019, **17**, 3996-4004.
167. T. S. Haque and S. H. Gellman, *Journal of the American Chemical Society*, 1997, **119**, 2303-2304.
168. S. Aravinda, N. Shamala, R. Rajkishore, H. N. Gopi and P. Balaram, *Angewandte Chemie International Edition*, 2002, **41**, 3863-3865.
169. U. Raghavender, S. Aravinda, R. Rai, N. Shamala and P. Balaram, *Organic & biomolecular chemistry*, 2010, **8**, 3133-3135.
170. L. R. Masterson, M. A. Etienne, F. Porcelli, G. Barany, R. P. Hammer and G. Veglia, *Peptide Science*, 2007, **88**, 746-753.
171. E. Benedetti, A. Bavoso, B. Di Blasio, V. Pavone, C. Pedone, C. Toniolo and G. M. Bonora, *Proceedings of the National Academy of Sciences*, 1982, **79**, 7951-7954.
172. U. Nagai and K. Sato, *Tetrahedron Letters*, 1985, **26**, 647-650.
173. B. Eckhardt, W. Grosse, L.-O. Essen and A. Geyer, *Proceedings of the National Academy of Sciences*, 2010, **107**, 18336-18341.
174. F. A. Etzkorn, T. Guo, M. A. Lipton, S. D. Goldberg and P. A. Bartlett, *Journal of the American Chemical Society*, 1994, **116**, 10412-10425.
175. G. Müller, G. Hessler and H. Y. Decornez, *Angewandte Chemie International Edition*, 2000, **39**, 894-896.
176. W. C. Ripka, G. V. De Lucca, A. C. Bach, R. S. Pottorf and J. M. Blaney, *Tetrahedron*, 1993, **49**, 3593-3608.
177. M. Hata and G. R. Marshall, *Journal of Computer-Aided Molecular Design*, 2006, **20**, 321-331.
178. L. Belvisi, C. Gennari, A. Mielgo, D. Potenza and C. Scolastico, *European Journal of Organic Chemistry*, 1999, 389-400.
179. L. Belvisi, A. Bernardi, L. Manzoni, D. Potenza and C. Scolastico, *European Journal of Organic Chemistry*, 2000, 2563-2569.
180. L. Belvisi, C. Gennari, A. Madder, A. Mielgo, D. Potenza and C. Scolastico, *European Journal of Organic Chemistry*, 2000, 695-699.
181. K. Kim, J.-p. Dumas and J. P. Germanas, *The Journal of Organic Chemistry*, 1996, **61**, 3138-3144.
182. M. G. Hinds, N. G. J. Richards and J. A. Robinson, *Journal of the Chemical Society, Chemical Communications*, 1988, 1447-1449.
183. H. Bittermann and P. Gmeiner, *The Journal of Organic Chemistry*, 2006, **71**, 97-102.
184. L. R. Whitby, Y. Ando, V. Setola, P. K. Vogt, B. L. Roth and D. L. Boger, *Journal of the American Chemical Society*, 2011, **133**, 10184-10194.

185. L. Lomlim, J. Einsiedel, F. W. Heinemann, K. Meyer and P. Gmeiner, *The Journal of Organic Chemistry*, 2008, **73**, 3608-3611.
186. W. J. Wedemeyer, E. Welker and H. A. Scheraga, *Biochemistry*, 2002, **41**, 14637-14644.
187. P. A. M. Schmidpeter and F. X. Schmid, *Journal of Molecular Biology*, 2015, **427**, 1609-1631.
188. A. H. Andreotti, *Biochemistry*, 2003, **42**, 9515-9524.
189. M. Keller, C. Boissard, L. Patiny, N. N. Chung, C. Lemieux, M. Mutter and P. W. Schiller, *Journal of Medicinal Chemistry*, 2001, **44**, 3896-3903.
190. K. Oh and Z. Guan, *Chemical Communications*, 2006, 3069-3071.
191. S. Crecente-Garcia, A. Neckebroeck, J. S. Clark, B. O. Smith and A. R. Thomson, *Organic Letters*, 2020, **22**, 4424-4428.
192. A. G. Cochran, N. J. Skelton and M. A. Starovasnik, *Proceedings of the National Academy of Sciences*, 2001, **98**, 5578.
193. J. Ciarkowski, *Biopolymers*, 1984, **23**, 397-407.
194. C. Ramakrishnan, P. K. C. Paul and K. Ramnarayan, *Journal of Biosciences*, 1985, **8**, 239-251.
195. V. Sarojini, A. J. Cameron, K. G. Varnava, W. A. Denny and G. Sanjayan, *Chemical Reviews*, 2019, **119**, 10318-10359.
196. D. P. Slough, S. M. McHugh, A. E. Cummings, P. Dai, B. L. Pentelute, J. A. Kritzer and Y.-S. Lin, *The Journal of Physical Chemistry B*, 2018, **122**, 3908-3919.
197. J. N. Brown and C. H. Yang, *Journal of the American Chemical Society*, 1979, **101**, 445-449.
198. S. Ciudad, N. Bayó-Puxán, M. Varese, J. Seco, M. Teixidó, J. García and E. Giralt, *ChemistrySelect*, 2018, **3**, 2343-2351.
199. L. M. Gierasch, C. M. Deber, V. Madison, C.-H. Niu and E. R. Blout, *Biochemistry*, 1981, **20**, 4730-4738.
200. L. Wu, Y. Lu, Q.-T. Zheng, N.-H. Tan, C.-M. Li and J. Zhou, *Journal of Molecular Structure*, 2007, **827**, 145-148.
201. K. D. Kopple, G. Kartha, K. K. Bhandary and K. Romanowska, *Journal of the American Chemical Society*, 1985, **107**, 4893-4897.
202. K. D. Kopple, Y. S. Wang, A. G. Cheng and K. K. Bhandary, *Journal of the American Chemical Society*, 1988, **110**, 4168-4176.
203. C. L. Barnes and D. van der Helm, *Acta Crystallographica Section B*, 1982, **38**, 2589-2595.
204. J. Witek, B. G. Keller, M. Blatter, A. Meissner, T. Wagner and S. Riniker, *Journal of Chemical Information and Modeling*, 2016, **56**, 1547-1562.
205. C. Merten, F. Li, K. Bravo-Rodriguez, E. Sanchez-Garcia, Y. Xu and W. Sander, *Physical Chemistry Chemical Physics*, 2014, **16**, 5627-5633.
206. J. M. Bujnicki, *ChemBioChem*, 2006, **7**, 19-27.
207. B. Kuhlman and P. Bradley, *Nature Reviews Molecular Cell Biology*, 2019, **20**, 681-697.
208. Y. Zhang and J. Skolnick, *Proceedings of the National Academy of Sciences*, 2004, **101**, 7594-7599.
209. Y. Zhang, *BMC Bioinformatics*, 2008, **9**, 40.
210. A. Thomas, S. Deshayes, M. Decaffmeyer, M. H. Van Eyck, B. Charloteaux and R. Brasseur, *Proteins: Structure, Function, and Bioinformatics*, 2006, **65**, 889-897.
211. C. Etchebest, C. Benros, S. Hazout and A. G. de Brevern, *Proteins: Structure, Function, and Bioinformatics*, 2005, **59**, 810-827.
212. B. Offmann, M. Tyagi and G. A. de Brevern, *Current Bioinformatics*, 2007, **2**, 165-202.
213. J. Beaufays, L. Lins, A. Thomas and R. Brasseur, *Journal of Peptide Science*, 2012, **18**, 17-24.
214. P. Thévenet, Y. Shen, J. Maupetit, F. Guyon, P. Derreumaux and P. Tufféry, *Nucleic Acids Research*, 2012, **40**, W288-W293.
215. J. Maupetit, P. Derreumaux and P. Tuffery, *Nucleic Acids Research*, 2009, **37**, W498-W503.
216. P. Tuffery, F. Guyon and P. Derreumaux, *Journal of Computational Chemistry*, 2005, **26**, 506-513.

217. J. Maupetit, P. Tuffery and P. Derreumaux, *Proteins: Structure, Function, and Bioinformatics*, 2007, **69**, 394-408.
218. K. Harpreet, G. Aarti and G. P. S. Raghava, *Protein & Peptide Letters*, 2007, **14**, 626-631.
219. S. Singh, H. Singh, A. Tuknait, K. Chaudhary, B. Singh, S. Kumaran and G. P. S. Raghava, *Biology Direct*, 2015, **10**, 73.
220. L. J. McGuffin, K. Bryson and D. T. Jones, *Bioinformatics*, 2000, **16**, 404-405.
221. P. B. Timmons and C. M. Hewage, *Briefings in Bioinformatics*, 2021, **22**, bbab308.
222. S. Wu and Y. Zhang, *Nucleic acids research*, 2007, **35**, 3375-3382.
223. A. Hospital, J. R. Goñi, M. Orozco and J. L. Gelpí, *Advances and Applications in Bioinformatics and Chemistry*, 2015, **8**, 37-47.
224. M. A. González, *JDN*, 2011, **12**, 169-200.
225. J. W. Ponder and D. A. Case, in *Advances in Protein Chemistry*, Academic Press, 2003, vol. 66, pp. 27-85.
226. D. P. Slough, S. M. McHugh and Y.-S. Lin, *Biopolymers*, 2018, **109**, e23113.
227. A. S. Kamenik, U. Lessel, J. E. Fuchs, T. Fox and K. R. Liedl, *Journal of Chemical Information and Modeling*, 2018, **58**, 982-992.
228. A. E. Cummings, J. Miao, D. P. Slough, S. M. McHugh, J. A. Kritzer and Y.-S. Lin, *Biophysical Journal*, 2019, **116**, 433-444.
229. M. Jusot, D. Stratmann, M. Vaisset, J. Chomilier and J. Cortés, *Journal of Chemical Information and Modeling*, 2018, **58**, 2355-2368.
230. G. Bussi and A. Laio, *Nature Reviews Physics*, 2020, **2**, 200-212.
231. S. Piana and A. Laio, *The Journal of Physical Chemistry B*, 2007, **111**, 4553-4559.
232. Y. Sugita and Y. Okamoto, *Chemical Physics Letters*, 1999, **314**, 141-151.
233. A. E. Garcia, H. Herce and D. Paschek, in *Annual Reports in Computational Chemistry*, ed. D. C. Spellmeyer, Elsevier, 2006, vol. 2, pp. 83-95.
234. K. Ostermeir and M. Zacharias, *Biochimica et Biophysica Acta (BBA) - Proteins and Proteomics*, 2013, **1834**, 847-853.
235. J. W. Neidigh, R. M. Fesinmeyer and N. H. Andersen, *Nature Structural Biology*, 2002, **9**, 425.
236. P. A. Kollman, *Accounts of Chemical Research*, 1996, **29**, 461-469.
237. K. Lindorff-Larsen, S. Piana, K. Palmo, P. Maragakis, J. L. Klepeis, R. O. Dror and D. E. Shaw, *Proteins*, 2010, **78**, 1950-1958.
238. Y. Duan, C. Wu, S. Chowdhury, M. C. Lee, G. Xiong, W. Zhang, R. Yang, P. Cieplak, R. Luo, T. Lee, J. Caldwell, J. Wang and P. Kollman, *Journal of Computational Chemistry*, 2003, **24**, 1999-2012.
239. G. A. Kaminski, R. A. Friesner, J. Tirado-Rives and W. L. Jorgensen, *The Journal of Physical Chemistry B*, 2001, **105**, 6474-6487.
240. C. Oostenbrink, A. Villa, A. E. Mark and W. F. Van Gunsteren, *Journal of Computational Chemistry*, 2004, **25**, 1656-1676.
241. F. Jiang, C.-Y. Zhou and Y.-D. Wu, *The Journal of Physical Chemistry B*, 2014, **118**, 6983-6998.
242. C.-Y. Zhou, F. Jiang and Y.-D. Wu, *The Journal of Physical Chemistry B*, 2015, **119**, 1035-1047.
243. H. Geng, F. Jiang and Y.-D. Wu, *The Journal of Physical Chemistry Letters*, 2016, **7**, 1805-1810.
244. S. M. McHugh, J. R. Rogers, H. Yu and Y.-S. Lin, *Journal of Chemical Theory and Computation*, 2016, **12**, 2480-2488.
245. S. M. McHugh, H. Yu, D. P. Slough and Y.-S. Lin, *Physical Chemistry Chemical Physics*, 2017, **19**, 3315-3324.
246. J. Miao, M. L. Descoteaux and Y.-S. Lin, *Chemical Science*, 2021, **12**, 14927-14936.
247. N. J. M. Macaluso and R. C. Glen, *ChemMedChem*, 2010, **5**, 1247-1253.
248. N. A. Chapman, D. J. Dupré and J. K. Rainey, *Biochemistry and Cell Biology*, 2014, **92**, 431-440.
249. H. Antushevich and M. Wójcik, *Clinica Chimica Acta*, 2018, **483**, 241-248.
250. M. B. Wysocka, K. Pietraszek-Gremplewicz and D. Nowak, *Frontiers in Physiology*, 2018, **9**.

251. G. Hu, Z. Wang, R. Zhang, W. Sun and X. Chen, *Frontiers in Physiology*, 2021, **12**.
252. A. M. Razavi, W. M. Wuest and V. A. Voelz, *Journal of Chemical Information and Modeling*, 2014, **54**, 1425-1432.
253. J. Wang, R. M. Wolf, J. W. Caldwell, P. A. Kollman and D. A. Case, *Journal of Computational Chemistry*, 2004, **25**, 1157-1174.
254. H. Huang, J. Damjanovic, J. Miao and Y.-S. Lin, *Physical Chemistry Chemical Physics*, 2021, **23**, 607-616.
255. K. Jolliffe, *Australian Journal of Chemistry, 2018,* **71**, 723-730.
256. D. Pal and P. Chakrabarti, *Journal of Molecular Biology*, 1999, **294**, 271-288.
257. A. Jabs, M. S. Weiss and R. Hilgenfeld, *Journal of Molecular Biology*, 1999, **286**, 291-304.
258. J. H. Viles, J. B. O. Mitchell, S. L. Gough, P. M. Doyle, C. J. Harris, P. J. Sadler and J. M. Thornton, *European Journal of Biochemistry*, 1996, **242**, 352-362.
259. M. G. Wu and M. W. Deem, *The Journal of Chemical Physics*, 1999, **111**, 6625-6632.
260. S. J. Weiner, P. A. Kollman, D. T. Nguyen and D. A. Case, *Journal of Computational Chemistry*, 1986, **7**, 230-252.
261. P. Craveur, A. P. Joseph, P. Poulain, A. G. de Brevern and J. Rebehmed, *Amino Acids*, 2013, **45**, 279-289.
262. F. Edlich and G. Fischer, in *Molecular Chaperones in Health and Disease*, eds. K. Starke and M. Gaestel, Springer Berlin Heidelberg, Berlin, Heidelberg, 2006, pp. 359-404.
263. C. Melis, G. Bussi, S. C. R. Lummis and C. Molteni, *The Journal of Physical Chemistry B*, 2009, **113**, 12148-12153.
264. J. Xia and R. M. Levy, *The Journal of Physical Chemistry B*, 2014, **118**, 4535-4545.
265. K. Shinoda and H. Fujitani, *Scientific Reports*, 2017, **7**, 16964.
266. S. Kawagoe, H. Nakagawa, H. Kumeta, K. Ishimori and T. Saio, *Journal of Biological Chemistry*, 2018, **293**, 15095-15106.
267. U. Doshi and D. Hamelberg, *The Journal of Physical Chemistry B*, 2009, **113**, 16590-16595.
268. D. Hamelberg, T. Shen and J. A. McCammon, *Journal of the American Chemical Society*, 2005, **127**, 1969-1974.
269. G. P. Di Martino, M. Masetti, A. Cavalli and M. Recanatini, *Proteins: Structure, Function, and Bioinformatics*, 2014, **82**, 2943-2956.
270. C. Melis, G. Bussi, S. C. R. Lummis and C. Molteni, *The Journal of Physical Chemistry B* 2009, **113**, 12148-12153.
271. Y. Yonezawa, H. Shimoyama and H. Nakamura, *Chemical Physics Letters*, 2011, **501**, 598-602.
272. S. Fischer, R. L. Dunbrack and M. Karplus, *Journal of the American Chemical Society*, 1994, **116**, 11931-11937.
273. C. Cox, V. G. Young and T. Lectka, *Journal of the American Chemical Society*, 1997, **119**, 2307-2308.
274. V. Leone, G. Lattanzi, C. Molteni and P. Carloni, *PLOS Computational Biology*, 2009, **5**, e1000309.
275. J. N. Lambert, J. P. Mitchell and K. D. Roberts, *Journal of the Chemical Society, Perkin Transactions 1*, 2001, 471-484.
276. H. N. Cheng and F. A. Bovey, *Biopolymers*, 1977, **16**, 1465-1472.
277. W.-J. Wu and D. P. Raleigh, *Biopolymers*, 1998, **45**, 381-394.
278. N. J. Zondlo, *Accounts of chemical research*, 2013, **46**, 1039-1049.
279. A. Barducci, G. Bussi and M. Parrinello, *Physical Review Letters*, 2008, **100**, 020603.
280. E. F. Kolesanova, M. A. Sanzhakov and O. N. Kharybin, *International Journal of Peptides*, 2013, 197317.
281. J. L. Klepeis, K. Lindorff-Larsen, R. O. Dror and D. E. Shaw, *Current Opinion in Structural Biology*, 2009, **19**, 120-127.
282. R. B. Best, N.-V. Buchete and G. Hummer, *Biophysical Journal*, 2008, **95**, L07-L09.

283. V. Hornak, R. Abel, A. Okur, B. Strockbine, A. Roitberg and C. Simmerling, *Proteins*, 2006, **65**, 712-725.

284. T. Drakenberg and S. Forsén, *Journal of the Chemical Society D: Chemical Communications*, 1971, 1404-1405.

285. A. Radzicka, L. Pedersen and R. Wolfenden, *Biochemistry*, 1988, **27**, 4538-4541.

286. L. J. Weiser and E. E. Santiso, *Journal of Computational Chemistry*, 2019, **40**, 1946-1956.

287. P. Cieplak, F.-Y. Dupradeau, Y. Duan and J. Wang, *Journal of Physics: Condensed Matter*, 2009, **21**, 333102-333102.

288. S. Cardamone, T. J. Hughes and P. L. A. Popelier, *Physical Chemistry Chemical Physics*, 2014, **16**, 10367-10387.

289. S. Li and A. H. Elcock, *The Journal of Physical Chemistry Letters*, 2015, **6**, 2127-2133.

290. C. M. Slupsky, C. M. Kay, F. C. Reinach, L. B. Smillie and B. D. Sykes, *Biochemistry*, 1995, **34**, 7365-7375.

291. C. M. Slupsky, F. C. Reinach, L. B. Smillie and B. D. Sykes, *Protein Science*, 1995, **4**, 1279-1290.

292. T. Tengel, I. Sethson and M. S. Francis, *European Journal of Biochemistry*, 2002, **269**, 3659-3668.

293. S. M. Guéret, S. Thavam, R. J. Carbajo, M. Potowski, N. Larsson, G. Dahl, A. Dellsén, T. N. Grossmann, A. T. Plowright, E. Valeur, M. Lemurell and H. Waldmann, *Journal of the American Chemical Society*, 2020, **142**, 4904-4915.

294. J. R. Frost, C. C. G. Scully and A. K. Yudin, *Nature Chemistry*, 2016, **8**, 1105-1111.

295. E. A. Villar, D. Beglov, S. Chennamadhavuni, J. A. Porco, D. Kozakov, S. Vajda and A. Whitty, *Nature Chemical Biology*, 2014, **10**, 723-731.

296. J. R. Harjani, B. K. Yap, E. W. W. Leung, A. Lucke, S. E. Nicholson, M. J. Scanlon, D. K. Chalmers, P. E. Thompson, R. S. Norton and J. B. Baell, *Journal of Medicinal Chemistry*, 2016, **59**, 5799-5809.

297. A. E. Owens, I. de Paola, W. A. Hansen, Y.-W. Liu, S. D. Khare and R. Fasan, *Journal of the American Chemical Society*, 2017, **139**, 12559-12568.

298. M. W. Macarthur and J. M. Thornton, *Journal of molecular biology*, 1991, **218 2**, 397-412.

299. S. R. Trevino, S. Schaefer, J. M. Scholtz and C. N. Pace, *Journal of Molecular Biology*, 2007, **373**, 211-218.

300. S. Melnikov, J. Mailliot, L. Rigger, S. Neuner, B.-S. Shin, G. Yusupova, T. E. Dever, R. Micura and M. Yusupov, *EMBO reports*, 2016, **17**, 1776-1784.

301. C. The UniProt, *Nucleic Acids Research*, 2018, **47**, D506-D515.

302. M. Tsutsumi and J. M. Otaki, *Journal of Chemical Information and Modeling*, 2011, **51**, 1457-1464.

303. A. Rodriguez and A. Laio, *Science*, 2014, **344**, 1492.

304. F. Sittel, A. Jain and G. Stock, *The Journal of Chemical Physics*, 2014, **141**, 014111.

305. C. D. DuPai, B. W. Davies and C. O. Wilke, *bioRxiv*, 2020, 2020.2010.2028.359612.

306. A.-S. Yang and B. Honig, *Journal of Molecular Biology*, 1995, **252**, 366-376.

307. A. Bornot and A. G. de Brevern, *Bioinformation*, 2006, **1**, 153-155.

308. D. E. Stewart, A. Sarkar and J. E. Wampler, *Journal of Molecular Biology*, 1990, **214**, 253-260.

309. M. S. Weiss, A. Jabs and R. Hilgenfeld, *Nature Structural Biology*, 1998, **5**, 676-676.

310. J. Zhang and M. W. Germann, *Biopolymers*, 2011, **95**, 755-762.

311. J. D. Wade, J. Bedford, R. C. Sheppard and G. W. Tregear, *Peptide Research*, 1991, **4**, 194-199.

312. K. Wuthrich, *NMR of proteins and nucleic acids*, 1986.

313. A. E. Tonelli and A. I. Brewster, *Journal of the American Chemical Society*, 1972, **94**, 2851-2854.

314. K. D. Kopple, A. Go and T. J. Schamper, *Journal of the American Chemical Society*, 1978, **100**, 4289-4295.

315. K. I. Varughese, G. Kartha and K. D. Kopple, *Journal of the American Chemical Society*, 1981, **103**, 3310-3313.
316. C.-H. Yang, J. N. Brown and K. D. Kopple, *Journal of the American Chemical Society*, 1981, **103**, 1715-1719.
317. I. L. Karle and J. Karle, *Acta Crystallographica*, 1963, **16**, 969-975.
318. I. L. Karle, J. W. Gibson and J. Karle, *Journal of the American Chemical Society*, 1970, **92**, 3755-3760.
319. F. A. Bovey, A. I. Brewster, D. J. Patel, A. E. Tonelli and D. A. Torchia, *Accounts of Chemical Research*, 1972, **5**, 193-200.
320. D. Torchia, A. Di Corato, S. Wong, C. Deber and E. Blout, *Journal of the American Chemical Society*, 1972, **94**, 609-615.
321. J. N. Brown and R. G. Teller, *Journal of the American Chemical Society*, 1976, **98**, 7565-7569.
322. M. B. Hossain and D. Van der Helm, *Journal of the American Chemical Society*, 1978, **100**, 5191-5198.
323. E. C. Kostansek, W. N. Lipscomb and W. E. Thiessen, *Journal of the American Chemical Society*, 1979, **101**, 834-837.
324. E. C. Kostansek, W. E. Thiessen, D. Schomburg and W. N. Lipscomb, *Journal of the American Chemical Society*, 1979, **101**, 5811-5815.
325. H. Kessler, M. Klein, A. Müller, K. Wagner, J. W. Bats, K. Ziegler and M. Frimmer, *Angewandte Chemie International Edition in English*, 1986, **25**, 997-999.
326. H. Kessler, J. W. Bats, C. Griesinger, S. Koll, M. Will and K. Wagner, *Journal of the American Chemical Society*, 1988, **110**, 1033-1049.
327. V. Pavone, E. Benedetti, B. Di Blasio, A. Lombardi, C. Pedone, L. Tomasich and G. P. Lorenzi, *Biopolymers*, 1989, **28**, 215-223.
328. C. L. Barnes, M. B. Hossain, K. Fidelis and D. van der Helm, *Acta Crystallographica Section B*, 1990, **46**, 238-246.
329. J.-P. Declercq, B. Tinant, S. Bashwira and C. Hootele, *Acta Crystallographica Section C*, 1990, **46**, 1259-1262.
330. H. Kessler, H. Matter, G. Gemmecker, M. Kottenhahn and J. W. Bats, *Journal of the American Chemical Society*, 1992, **114**, 4805-4818.
331. H. Morita, T. Kayashita, A. Shishido, K. Takeya, H. Itokawa and M. Shiro, *Tetrahedron*, 1996, **52**, 1165-1176.
332. B. Dittrich, T. Koritsanszky, M. Grosche, W. Scherer, R. Flaig, A. Wagner, H. G. Krane, H. Kessler, C. Riemer, A. M. M. Schreurs and P. Luger, *Acta Crystallographica Section B*, 2002, **58**, 721-727.
333. C. Appelt, A. Wessolowski, J. A. Söderhäll, M. Dathe and P. Schmieder, *ChemBioChem*, 2005, **6**, 1654-1662.
334. Y. Tong, J.-G. Luo, R. Wang, X.-B. Wang and L.-Y. Kong, *Bioorganic & Medicinal Chemistry Letters*, 2012, **22**, 1908-1911.
335. D. S. Nielsen, R.-J. Lohman, H. N. Hoang, T. A. Hill, A. Jones, A. J. Lucke and D. P. Fairlie, *ChemBioChem*, 2015, **16**, 2289-2293.
336. S. L. Garland and P. M. Dean, *Journal of Computer-Aided Molecular Design*, 1999, **13**, 485-498.
337. K. Möhle, M. Gußmann and H.-J. Hofmann, *Journal of Computational Chemistry*, 1997, **18**, 1415-1430.
338. G. A. Jeffrey, *An Introduction to Hydrogen Bonding*, Oxford University Press, 1997.
339. N. I. Fisher, *Journal of Structural Geology*, 1989, **11**, 775-778.
340. D. Scott, *Multivariate density estimation: Theory, practice, and visualization: Second edition*, John Wiley and Sons, 2015.
341. H. Läuter, *Biometrical Journal*, 1988, **30**, 876-877.

342. N.-B. Heidenreich, A. Schindler and S. Sperlich, *AStA Advances in Statistical Analysis*, 2013, **97**, 403-433.
343. D. A. C. Beck, D. O. V. Alonso, D. Inoyama and V. Daggett, *Proceedings of the National Academy of Sciences*, 2008, **105**, 12259.
344. A. M. Firestine, V. M. Chellgren, S. J. Rucker, T. E. Lester and T. P. Creamer, *Biochemistry*, 2008, **47**, 3216-3224.
345. Z. Shi, K. Chen, Z. Liu and N. R. Kallenbach, *Chemical Reviews*, 2006, **106**, 1877-1897.
346. K. Plaxco, C. Morton, S. Grimshaw, J. Jones, M. Pitkeathly, I. Campbell and C. Dobson, *Journal of biomolecular NMR*, 1997, **10**, 221-230.
347. G. Merutka, J. Dyson and P. Wright, *Journal of biomolecular NMR*, 1995, **5**, 14-24.
348. S. Schwarzinger, G. Kroon, T. Foss, P. Wright and J. Dyson, *Journal of Biomolecular NMR*, 2000, **18**, 43-48.
349. M. C. Childers, C.-L. Towse and V. Daggett, *Protein Engineering, Design and Selection*, 2016, **29**, 271-280.
350. N. C. Fitzkee, P. J. Fleming and G. D. Rose, *Proteins: Structure, Function, and Bioinformatics*, 2005, **58**, 852-854.
351. S. Sra, *Computational Statistics*, 2012, **27**, 177-190.
352. K. V. Mardia and P. E. Jupp, *Directional Statistics*, John Wiley and Sons, 2000.
353. P. Hall, G. S. Watson and J. Cabrera, *Biometrika*, 1987, **74**, 751-762.
354. M. Oliveira, R. M. Crujeiras and A. Rodríguez-Casal, *Computational Statistics & Data Analysis*, 2012, **56**, 3898-3908.
355. T. M. Pham Ngoc, *Journal of Multivariate Analysis*, 2019, **173**, 248-267.
356. C. C. Taylor, *Computational Statistics & Data Analysis*, 2008, **52**, 3493-3500.
357. K. Bedouhene and N. Zougab, *Monte Carlo Methods and Applications*, 2020, **26**, 69-82.
358. M. Di Marzio, A. Panzera and C. C. Taylor, *Journal of Statistical Planning and Inference*, 2011, **141**, 2156-2173.
359. D. Görür and C. Edward Rasmussen, *Journal of Computer Science and Technology*, 2010, **25**, 653-664.
360. D. Ting, G. Wang, M. Shapovalov, R. Mitra, M. I. Jordan and R. L. Dunbrack, Jr., *PLoS computational biology*, 2010, **6**, e1000763.
361. S. C. Lovell, I. W. Davis, W. B. Arendall Iii, P. I. W. de Bakker, J. M. Word, M. G. Prisant, J. S. Richardson and D. C. Richardson, *Proteins: Structure, Function, and Bioinformatics*, 2003, **50**, 437-450.
362. Y.-D. Cai, X.-J. Liu, X.-b. Xu and K.-C. Chou, *Journal of Peptide Science*, 2002, **8**, 297-301.
363. A. J. Shepherd, D. Gorse and J. M. Thornton, *Protein Science*, 1999, **8**, 1045-1055.
364. L. Breiman, J. H. Friedman, R. A. Olshen and C. J. Stone, *Classification and Regression Trees*, Wadsworth and Brooks, 1984.
365. T. A. Springer, J. Zhu and T. Xiao, *Journal of Cell Biology*, 2008, **182**, 791-800.
366. N. Chawla, K. Bowyer, L. Hall and W. Kegelmeyer, *Journal of Artificial Intelligence Research*, 2002, **16**, 321-357.
367. E. R. Blout, *Biopolymers*, 1981, **20**, 1901-1912.
368. T. Michels, R. Dölling, U. Haberkorn and W. Mier, *Organic Letters*, 2012, **14**, 5218-5221.
369. D. Samson, D. Rentsch, M. Minuth, T. Meier and G. Loidl, *Journal of Peptide Science*, 2019, **25**, e3193.
370. K. Neumann, J. Farnung, S. Baldauf and J. W. Bode, *Nature Communications*, 2020, **11**, 982.
371. H. E. Stanger and S. H. Gellman, *Journal of the American Chemical Society*, 1998, **120**, 4236-4237.
372. S. M. Twine and A. G. Szabo, in *Methods in Enzymology*, Academic Press, 2003, vol. 360, pp. 104-127.
373. L. D. Lavis and R. T. Raines, *ACS Chemical Biology*, 2008, **3**, 142-155.
374. A. R. Katritzky and T. Narindoshvili, *Organic & Biomolecular Chemistry*, 2009, **7**, 627-634.

375. A. T. Krueger and B. Imperiali, *ChemBioChem*, 2013, **14**, 788-799.
376. R. W. Sinkeldam, N. J. Greco and Y. Tor, *Chemical reviews*, 2010, **110**, 2579-2619.
377. A. B. T. Ghisaidoobe and S. J. Chung, *International Journal of Molecular Sciences*, 2014, **15**, 22518-22538.
378. C. A. Royer, *Chemical Reviews*, 2006, **106**, 1769-1784.
379. R. Rieger and K. Müllen, *Journal of Physical Organic Chemistry*, 2010, **23**, 315-325.
380. M. Yoshizawa and J. K. Klosterman, *Chemical Society Reviews*, 2014, **43**, 1885-1898.
381. Y. Zou, D. D. Young, A. Cruz-Montanez and A. Deiters, *Organic Letters*, 2008, **10**, 4661-4664.
382. K. Ayyavoo and P. Velusamy, *New Journal of Chemistry*, 2021, **45**, 10997-11017.
383. H. Maeda, T. Maeda, K. Mizuno, K. Fujimoto, H. Shimizu and M. Inouye, *Chemistry – A European Journal*, 2006, **12**, 824-831.
384. E. Benedetti, L. S. Kocsis and K. M. Brummond, *Journal of the American Chemical Society*, 2012, **134**, 12418-12421.
385. J. Feng, X. Chen, Q. Han, H. Wang, P. Lu and Y. Wang, *Journal of Luminescence*, 2011, **131**, 2775-2783.
386. N. Sreerama and R. W. Woody, in *Methods in Enzymology*, Academic Press, 2004, vol. 383, pp. 318-351.
387. P. Morales and M. A. Jiménez, *Archives of Biochemistry and Biophysics*, 2019, **661**, 149-167.
388. I. B. Grishina and R. W. Woody, *Faraday Discussions*, 1994, **99**, 245-262.
389. S. Martinez-Rodriguez, J. Bacarizo, I. Luque and A. Camara-Artigas, *Journal of Structural Biology*, 2015, **191**, 381-387.
390. M. J. Macias, V. Gervais, C. Civera and H. Oschkinat, *Nature Structural Biology*, 2000, **7**, 375-379.
391. A. Jamous and Z. Salah, *Frontiers in Oncology*, 2018, **8**.
392. S.-S. Huang, L.-J. Hsu and N.-S. Chang, *FASEB BioAdvances*, 2020, **2**, 234-253.
393. N.-S. Chang, R. Lin, C.-I. Sze and R. I. Aqeilan, *Frontiers in oncology*, 2019, **9**, 719-719.
394. C.-Y. Hsu, K.-T. Lee, T.-Y. Sun, C.-I. Sze, S.-S. Huang, L.-J. Hsu and N.-S. Chang, *Cells*, 2021, **10**, 1781.
395. Z. Salah, A. Alian and R. Aqeilan, *Frontiers in bioscience : a journal and virtual library*, 2012, **17**, 331-348.
396. L. Otte, U. Wiedemann, B. Schlegel, J. R. Pires, M. Beyermann, P. Schmieder, G. Krause, R. Volkmer-Engert, J. Schneider-Mergener and H. Oschkinat, *Protein Science*, 2003, **12**, 491-500.
397. J. L. Ilsley, M. Sudol and S. J. Winder, *Cellular Signalling*, 2001, **13**, 625-632.
398. N. Reuven, M. Shanzer and Y. Shaul, *Experimental Biology and Medicine*, 2015, **240**, 375-382.
399. B. Kay, M. Williamson and M. Sudol, *FASEB journal,* 2000, **14**, 231-241.
400. M. Iglesias-Bexiga, F. Castillo, E. S. Cobos, T. Oka, M. Sudol and I. Luque, *PLoS One*, 2015, **10**, e0113828.
401. A. Komuro, M. Nagai, N. E. Navin and M. Sudol, *Journal of Biological Chemistry*, 2003, **278**, 33334-33341.
402. M. Sudol, H. I. Chen, C. Bougeret, A. Einbond and P. Bork, *FEBS Letters*, 1995, **369**, 67-71.
403. R. Yagi, L. F. Chen, K. Shigesada, Y. Murakami and Y. Ito, *The EMBO Journal*, 1999, **18**, 2551-2562.
404. H. I. Chen, A. Einbond, S.-J. Kwak, H. Linn, E. Koepf, S. Peterson, J. W. Kelly and M. Sudol, *Journal of Biological Chemistry*, 1997, **272**, 17070-17077.
405. M. Sudol, D. C. Shields and A. Farooq, *Seminars in Cell & Developmental Biology*, 2012, **23**, 827-833.
406. A. Verma, F. Jing-Song, M. L. Finch-Edmondson, A. Velazquez-Campoy, S. Balasegaran, M. Sudol and J. Sivaraman, *Oncotarget*, 2018, **9**, 8068-8080.
407. C. B. McDonald, S. K. N. McIntosh, D. C. Mikles, V. Bhat, B. J. Deegan, K. L. Seldeen, A. M. Saeed, L. Buffa, M. Sudol, Z. Nawaz and A. Farooq, *Biochemistry*, 2011, **50**, 9616-9627.

408. S. Sangadala, R. Metpally and B. Reddy, *Journal of biomolecular structure & dynamics*, 2007, **25**, 11-23.
409. T. Kim, D. Hwang, D. Lee, J.-H. Kim, S.-Y. Kim and D.-S. Lim, *The EMBO Journal*, 2017, **36**, 520-535.
410. H. Oh, M. Slattery, L. Ma, K. P. White, R. S. Mann and K. D. Irvine, *Cell Reports*, 2014, **8**, 449-459.
411. M. J. Macias, S. Wiesner and M. Sudol, *FEBS Letters*, 2002, **513**, 30-37.
412. J. Avruch, D. Zhou and N. Bardeesy, *Cell Cycle*, 2012, **11**, 1090-1096.
413. S. Strano, E. Munarriz, M. Rossi, L. Castagnoli, Y. Shaul, A. Sacchi, M. Oren, M. Sudol, G. Cesareni and G. Blandino, *Journal of Biological Chemistry*, 2001, **276**, 15164-15173.
414. E. Rozengurt and G. Eibl, *World J Gastroenterol*, 2019, **25**, 1797-1816.
415. M. J. Macias, M. Hyvönen, E. Baraldi, J. Schultz, M. Sudol, M. Saraste and H. Oschkinat, *Nature*, 1996, **382**, 646-649.
416. J. R. Pires, F. Taha-Nejad, F. Toepert, T. Ast, U. Hoffmüller, J. Schneider-Mergener, R. Kühne, M. J. Macias and H. Oschkinat, *Journal of Molecular Biology*, 2001, **314**, 1147-1156.
417. F. Toepert, J. R. Pires, C. Landgraf, H. Oschkinat and J. Schneider-Mergener, *Angewandte Chemie International Edition*, 2001, **40**, 897-900.
418. U. Strijowski, T. Hirsch, A. Quintilla, W. Wenzel and J. Eichler, *International Journal of Peptide Research and Therapeutics*, 2007, **13**, 245-250.
419. C. Cunningham Brian and A. Wells James, *Science*, 1989, **244**, 1081-1085.
420. T. Clackson and A. Wells James, *Science*, 1995, **267**, 383-386.
421. I. Massova and P. A. Kollman, *Journal of the American Chemical Society*, 1999, **121**, 8133-8143.
422. Y. Dehouck, J. M. Kwasigroch, M. Rooman and D. Gilis, *Nucleic Acids Research*, 2013, **41**, W333-W339.
423. D. E. V. Pires, T. L. Blundell and D. B. Ascher, *Scientific Reports*, 2016, **6**, 29575.
424. J. Schymkowitz, J. Borg, F. Stricher, R. Nys, F. Rousseau and L. Serrano, *Nucleic Acids Research*, 2005, **33**, W382-W388.
425. T. Kortemme, E. Kim David and D. Baker, *Science's STKE*, 2004, pl2.
426. I. S. Moreira, P. A. Fernandes and M. J. Ramos, *Journal of Computational Chemistry*, 2007, **28**, 644-654.
427. C. W. Wood, A. A. Ibarra, G. J. Bartlett, A. J. Wilson, D. N. Woolfson and R. B. Sessions, *Bioinformatics*, 2020, **36**, 2917-2919.
428. B. Ciani, M. Jourdan and M. S. Searle, *Journal of the American Chemical Society*, 2003, **125**, 9038-9047.
429. H. R. Bosshard, D. N. Marti and I. Jelesarov, *Journal of Molecular Recognition*, 2004, **17**, 1-16.
430. J. E. Donald, D. W. Kulp and W. F. DeGrado, *Proteins: Structure, Function, and Bioinformatics*, 2011, **79**, 898-915.
431. S. Pylaeva, M. Brehm and D. Sebastiani, *Scientific Reports*, 2018, **8**, 13626.
432. Q. Wang, A. A. Canutescu and R. L. Dunbrack, Jr., *Nature Protocols*, 2008, **3**, 1832-1847.
433. P. A. Chong, H. Lin, L. Wrana Jeffrey and D. Forman-Kay Julie, *Proceedings of the National Academy of Sciences*, 2010, **107**, 18404-18409.
434. G. Wang and R. L. Dunbrack, Jr., *Bioinformatics*, 2003, **19**, 1589-1591.
435. W. Kabsch and C. Sander, *Biopolymers*, 1983, **22**, 2577-2637.
436. R. P. Joosten, T. A. H. te Beek, E. Krieger, M. L. Hekkelman, R. W. W. Hooft, R. Schneider, C. Sander and G. Vriend, *Nucleic acids research*, 2011, **39**, D411-D419.
437. F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss and V. Dubourg, *the Journal of machine Learning research*, 2011, **12**, 2825-2830.
438. S. McIntosh-Smith, J. Price, R. B. Sessions and A. A. Ibarra, *International Journal of High Performance Computing Applications*, 2015, **29**, 119-134.

439. E. F. Pettersen, T. D. Goddard, C. C. Huang, G. S. Couch, D. M. Greenblatt, E. C. Meng and T. E. Ferrin, *Journal of Computational Chemistry*, 2004, **25**, 1605-1612.

440. F. Sievers, A. Wilm, D. Dineen, T. J. Gibson, K. Karplus, W. Li, R. Lopez, H. McWilliam, M. Remmert, J. Söding, J. D. Thompson and D. G. Higgins, *Molecular Systems Biology*, 2011, **7**, 539.

441. A. Cheng and K. M. Merz, *The Journal of Physical Chemistry*, 1996, **100**, 1927-1937.

442. G. Bussi, D. Donadio and M. Parrinello, *The Journal of Chemical Physics*, 2007, **126**, 014101.

443. H. J. C. Berendsen, J. P. M. Postma, W. F. van Gunsteren, A. DiNola and J. R. Haak, *The Journal of Chemical Physics*, 1984, **81**, 3684-3690.

444. W. L. Jorgensen, J. Chandrasekhar, J. D. Madura, R. W. Impey and M. L. Klein, *The Journal of Chemical Physics*, 1983, **79**, 926-935.

445. B. Hess, H. Bekker, H. J. C. Berendsen and J. G. E. M. Fraaije, *Journal of Computational Chemistry*, 1997, **18**, 1463-1472.

446. U. Essmann, L. Perera, M. L. Berkowitz, T. Darden, H. Lee and L. G. Pedersen, *The Journal of Chemical Physics*, 1995, **103**, 8577-8593.

447. R. W. Hockney, S. P. Goel and J. W. Eastwood, *Journal of Computational Physics*, 1974, **14**, 148-158.

448. J. Wang, W. Wang, P. A. Kollman and D. A. Case, *Journal of Molecular Graphics and Modelling*, 2006, **25**, 247-260.

449. J. M. Damas, L. C. S. Filipe, S. R. R. Campos, D. Lousa, B. L. Victor, A. M. Baptista and C. M. Soares, *Journal of Chemical Theory and Computation*, 2013, **9**, 5148-5157.

450. S. M. Kelly, T. J. Jess and N. C. Price, *Biochimica et Biophysica Acta (BBA) - Proteins and Proteomics*, 2005, **1751**, 119-139.