Structural Insights into Unnatural Proteins from Symmetry and Pseudosymmetry

By

Dale F. Kreitler

A dissertation in partial fulfillment of

the requirements for the degree of

Doctor of Philosophy

(Chemistry)

at the

UNIVERSITY OF WISCONSIN-MADISON

2017

Date of final oral examination: 02/28/2017

The dissertation is approved by the following members of the Final Oral Committee:

Sam H. Gellman, Professor, Chemistry

Katrina Forest, Professor, Chemistry and Bacteriology

Tom Record, Professor, Chemistry and Biochemistry

Sandro Mecozzi, Professor, Chemistry and Pharmaceutical Sciences

Judith Burstyn, Professor, Chemistry

To Lauren...

Acknowledgments

First and foremost I would like to thank Sam for his mentorship and for creating an environment where we can develop as independent scientists. I have been fortunate in that I was able to pursue a variety of interests during my graduate training and I am very grateful for that opportunity.

I am also thankful to the life sciences collaborative access team (LS-CAT) beamline staff at the advanced photon source (APS) at Argonne National Laboratory and also to the faculty at UW-Madison that make access to LS-CAT possible. A significant portion of the work in this thesis would not have been possible without the national laboratory system, so I am grateful for the work of those in the Department of Energy (DOE) that make facilities such as APS available to researchers such as myself.

I am lucky to have been able to collaborate with past and present Gellman group lab members, especially David Mortenson and Zhihui Yao, who also understand what the synthesis and purification of transmembrane peptides entail. I would be remiss if I did not mention Geoff Eddinger and our heroic efforts to keep the Gellman group preparative HPLCs functioning. Taking apart and rebuilding HPLC modules (and basically every instrument in general) has certainly enhanced my graduate education. Finally, I want to acknowledge Zhihui Yao, Kate Kurgan, and Nicole Thomas for keeping the fine tradition of UW-Chemistry macromolecular crystallography going, and I wish them the best in their endeavors.

Chapter 1 Introduction

For the better part of the last century, X-ray crystallography has been unparalleled in its ability to provide atomistic level insight into structures ranging from elemental packing arrangements to large macromolecular complexes. In molecular crystallography, molecular structure can be deduced from the X-ray diffraction pattern of crystals comprised of *N*-dimensional periodic arrangements of molecules. For molecular crystals, this periodicity is typically three-dimensional, but not necessarily. The discovery of quasicrystals in the latter half of the twentieth century made it clear that this periodicity can have a higher dimensionality. In 1992 the International Union of Crystallography (IUCr) altered the definition of a crystal from a solid 'having periodic atomic structure' to a solid 'having an essentially discrete diffraction diagram'.[1] X-ray crystal structures are highly sought after, especially in structural biology, because 'form ever follows function'. The information presented by these crystal structures is in the eye of the beholder, whether it is the chemist seeking the absolute configuration of a new synthetic target, a biologist searching for the presence of a new ligand in an enzyme active site, or a material scientist looking for a phase transition.

As chemists following the form-function paradigm, we often seek the molecular structure and not the crystal structure. However, there is much information beyond the three-dimensional coordinates that define the positions of atoms within a molecule, such as space group symmetry, packing configuration, displacement parameters, and crystal morphology, among others. Further insight into the molecular structure can often be gained by careful inspection of these sometimes overlooked parameters. Furthermore, these features are intimately related to and defined by the molecular structure. There are

some interesting historical examples where fundamental structural features were overlooked from both chemical and crystallographic perspectives. For instance, when Linus Pauling first reported the structure of the α-helix he assigned the wrong configuration to the C_α atoms, which resulted in a left-handed helix.[1–3] This oversight may be explained by Pauling's urgency to report the α-helix, but it is notable that the correct relative configuration of α-amino acids was known at the time and had been recently confirmed by Bijvoet's then newly reported anomalous scattering technique.[3] Another interesting oversight came several years later when Ramachandran reported abnormally short C-H···O bond lengths in collagen as 'normal' and not particularly noteworthy, effectively overlooking the first crystallographic evidence for C-H···O hydrogen bonds.[4,5] These oversights of course are easy to identify in hindsight, often with the added perspective on the original work, but they underscore the value in a holistic view of crystal structure when addressing questions pertaining to molecular structure.

In the mid-twentieth century, as the number of organic molecular crystal structures began to grow due to advances in instrumentation and an increase in the number of practitioners, it became clear that there was a non-random distribution of space groups adopted by these structures. There are 32 crystallographic point groups, which can be further expanded to 230 space groups when combined with translational symmetry operators that arise from lattice centering, glide planes, and screw axes.[6] So, although there are 230 mathematically possible space groups, only a select few are preferred among organic molecular crystals. Kitaigorodskii was one of the first researchers to seriously address this question of why some space groups are preferred over others.[7] His inductive argument is based on the principle of close-packing, where molecular crystals

are modelled as an array of tightly packed tri-axial ellipsoids, *i.e.*, packing arrangements that maximize density and decrease free volume are most favorable. The packing efficiency of uniform spheres, or 'the cannonball problem', had been a mathematical curiosity that long pre-dated the use of crystallography for molecular structure determination,[1,8] where the packing coefficient, defined as the ratio of the packing entity volume to the total volume, was known to be 0.74 for an array of hexagonally close packed (hcp) spheres.[9,10] Kitaigorodskii's key insight was that by including H atoms in the van der Waals radii of C atoms in organic molecules, the cannonball problem could be mapped to molecular solids.[7,11] The original treatise has been extensively reviewed,[12–14] debated,[15,16] and extended,[9,10,16] but here I will attempt to describe a few of the fundamental assumptions from which the preferred space groups are derived for molecules of arbitrary shape. This review is necessary to provide context for when the model begins to break down upon the transition from small organic molecules to large solvated macromolecules.

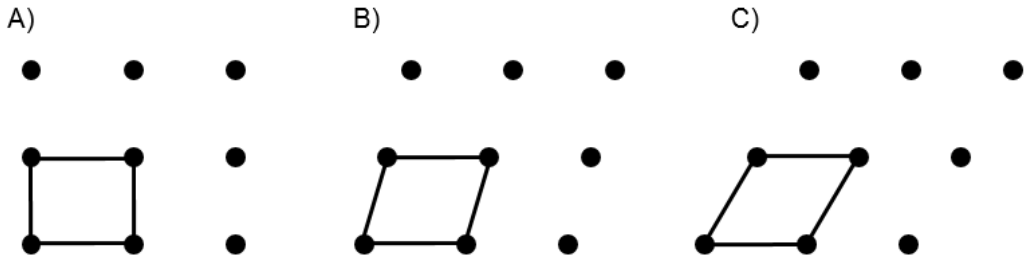


Figure 1.1. Schematic of A) rectangular, B) oblique, and C) hexagonal primitive lattices in two dimensions.

Kitaigorodskii's derivation starts with close packing layers in two dimensions and then builds into three dimensional networks by stacking the close-packed layers. Possible layers include rectangular, oblique, or hexagonal lattices (Figure 1.1). The first

simplification is to exclude layers with tetragonal or hexagonal symmetry, which would require a correlation between the two dimensional layer axes, thus making it impossible to pack objects of arbitrary shape with six-fold coordination, the optimal number of contacts between objects within layers. The next assumption is that horizontal mirror planes within layers are also highly disfavored; the duplication of layers due to a horizontal mirror plane results in knobs-to-knobs stacking whereas a 'knobs-to-hollows' packing is more favorable (Figure 1.2). Kitaigorodskii extends this 'knobs-to-knobs' argument to include horizontal glide planes (reflection then translation) to be undesirable in rectangular layers with the important distinction that they are acceptable in oblique layers. Specifically, in a rectangular lattice a horizontal glide will result in only four-fold coordination, which is less than the desired number of contacts, but the offset that occurs in an oblique lattice allows for six-fold coordination. These assumptions can be used to pare down the 48 possible rectangular and oblique layer planes into 10 distinct layer symmetry groups for molecules of arbitrary shape (Figure 1.3).

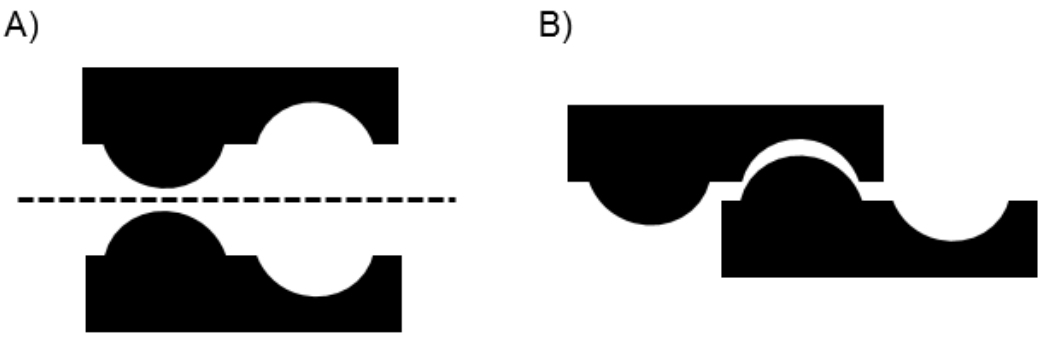


Figure 1.2. A) Horizontal mirror planes (dotted line) duplicate layers and result in unfavorable overlap of protrusions, B) whereas 'knobs-to-hollows' interactions, which require a translational offset, are more favorable.

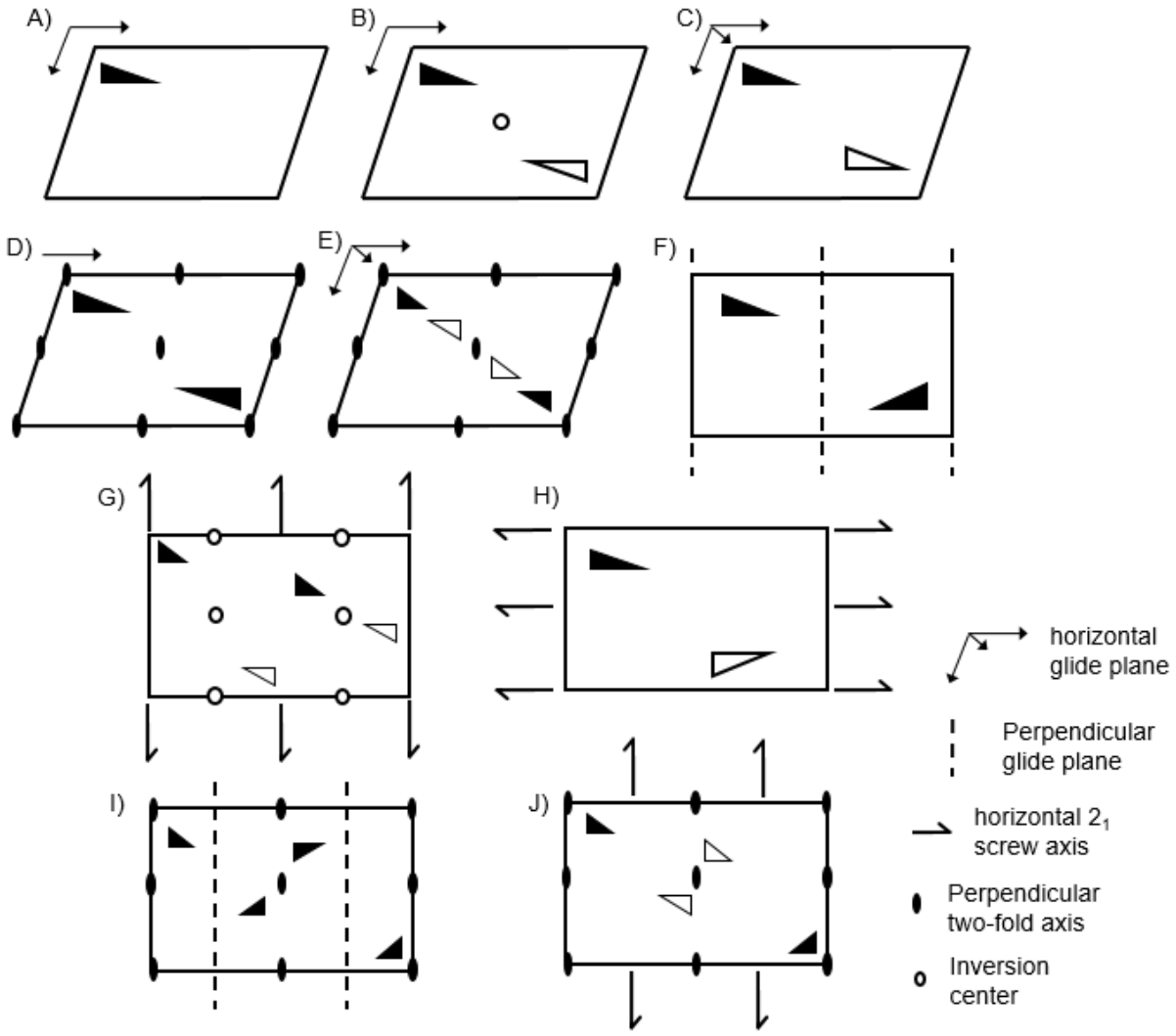


Figure 1.3. The ten distinct layer symmetry groups that result from applying the 'no horizontal mirror planes' and 'no horizontal glide planes in rectangular layers' assumptions to the 48 possible oblique and rectangular plane layers.[7] The triangle represents a polar molecule which is black on one side and white on the other. Symmetry operators are indicated in the figure legend.

Kitaigorodskii applies two final assumptions to further partition the 10 remaining layer symmetry groups into those that are closely packed (Figure 1.3: B,C,H) and those that are only 'coordination close-packed' (Figure 1.3: A,D,E,F,I,J). The first of these final two assumptions is that the polar layer symmetry groups, *i.e.*, groups with all black triangles in Figure 1.3, are undesirable for stacking. The final assumption is that two-fold

rotational axes perpendicular to the layer plane are also undesirable, because any void space in a molecule near a two-fold axis will be doubled due to rotational symmetry. Ultimately, there are three remaining layer groups compatible with close packing (Figure 1.3: B,C,H). Kitaigorodskii then performed the laborious task of cross checking each of the 74 lower symmetry space groups (triclinic, monoclinic, and orthorhombic) with the compatible layer groups. I have tabulated a more modern summary of this analysis in Table 1.1. The current top five most prevalent space groups in the Cambridge Structural Database (CSD; as of January 2017) show that $P2_1/c$ is by far the most prevalent space group. Another notable feature is that three of the top five prevalent space groups are centrosymmetric ($P2_1/c$, $P\bar{1}$, and $C2/c$), *i.e.*, they possess inversion symmetry.

Table 1.1. Space group ranking in the Cambridge Structural Database (CSD) and corresponding closely packed layer group.

Close-packed space group	Layer representation in Figure 1.3	% in CSD
$P2_1/c$	B,C,H	34.6
$P\bar{1}$	B	24.5
$C2/c$*	(G)	8.4
$P2_12_12_1$	H	7.2
$P2_1$	H	5.2

*$C2/c$ is a special case that requires the molecule to have inversion symmetry for the layer group G to apply, an interesting extension of this requirement has been described by Wilson.[16]
*** CSD = Cambridge Structural Database as of January 2017

Molecular crystals of small organic molecules clearly favor centrosymmetric space groups, as evidenced by the space group distribution in Table 1.1. These centrosymmetric space groups are only attainable from achiral molecules or racemic mixtures, and racemic structures are vastly more common than enantiopure structures. This phenomenon was noted in the literature by several others,[15] long before

Kitaigorodskii developed the close packing hypothesis. Before X-ray diffraction was even discovered, Wallach argued that racemic crystals are more stable than their chiral counterparts, based on the assumption stability was directly correlated with solubility.[15,17] If molecular crystals of small organic molecules favor space groups that minimize free volume and maximize packing density, then it seems reasonable to assume that centrosymmetric space groups are favored because they allow denser packing than do non-centrosymmetric space groups. Indeed, Brock has provided evidence via systematic CSD searches that centrosymmetric crystals are slightly denser (~1%) than their non-centrosymmetric counterparts for resolvable enantiomers.[15] However, Dunitz and Gavezzoti have more recently shown that for the 20 proteinogenic amino acids, racemic crystals are not significantly denser than their enantiopure chiral counterparts, but that the racemic crystals are however more thermodynamically stable.[17] These conflicting results are likely due to the prominent role that hydrogen bonding plays in amino acid crystal packing, where unfavorable decreases in packing density are offset by favorable positioning of functional groups for hydrogen bonding interactions.[9,10] From these myriad reports it becomes clear that the close packing model begins to break down when van der Waals interactions are no longer the dominant force in the crystal lattice, but centrosymmetric space groups are still favored because they allow for an increased number of energetically favorable contacts.

Kitaigorodskii's close-packing principle appears to stand the test of time with respect to the space group distribution of molecular crystals comprised of small organic molecules, where packing coefficients range from ~0.7-0.8 and coordination contacts range from 10 to 14 per molecule, and where van der Waals interactions dominate. For

a point of reference, an hcp lattice of identical spheres has a packing coefficient of 0.74 and coordination contact number of 12.[8] However, as previously discussed, the presence of functional groups which undergo enthalpically favorable electrostatic or hydrogen bonding interactions, which have distinct geometric preferences, will begin to invalidate many of the assumptions on which the close packing model is predicated. The violation of these close-packing principles becomes even more egregious for macromolecular crystals, where solvent content can range from 40-50% of the total volume.[18] Homochiral protein crystals are constrained to the 65 Sohncke space groups (chiral space groups), which do not contain symmetry operators of the second kind, *i.e.*, mirror/glide planes, roto-inversion axes, and inversion centers. A comparison of the space group distribution in the Protein Data Bank (PDB) with the Sohncke space group distribution in the Cambridge Structural Database (CSD) reveals a disparity in how space groups are distributed between the two databases, indicating that different principles govern the crystallization of macromolecules (Figure 1.4). The distribution of Sohncke space groups in the CSD is also much more highly skewed to favor $P2_12_12_1$ and $P2_1$, which combined account for almost 80% of CSD structures with Sohncke space groups.

Wukovitz and Yeates have put forth a compelling hypothesis to explain why protein crystals favor some space groups over others.[19] Mainly, these authors sought to address the question of why is $P2_12_12_1$ the most prevalent space group for protein crystals in the absence of close packing. Their hypothesis is based on an entropic view of crystallization where space groups are favored depending on inherent degrees of freedom that they offer to the first few molecules that begin to form a crystal. This entropic view is dissociated from any specific molecular packing interactions and is more reflective of a

mathematical property of space group symmetry. In fact, the only requirement pertaining to packing interactions is that packing entities must form an interconnected lattice spanning three dimensions.[19] This entropic model is defined by three parameters that are combined to yield a singular value, *D,* the rigid-body degrees of freedom for a particular space group. *D* is defined by the following expression:

$$D = S + L - C \qquad (1.1)$$

where *S* is the number of meaningful degrees of freedom for positioning a single molecule in the unit cell, *L* is the number of independent parameters for describing the unit cell, and *C* is the minimum number of unique contacts required to make an inter-connected three dimensional network of symmetry-related molecules (Figure 1.5). Each space group is assigned a particular *D* value based on these parameters.

Ultimately, proteins will preferentially crystallize in space groups with higher rigid-body degrees of freedom (Figure 1.4). In support of this hypothesis, the space group $P2_12_12_1$, which is the most prevalent among protein crystals, has the highest rigid-body degrees of freedom of any Sohncke space group (*D*=7).[19] The remainder of the top ten most prevalent protein crystal space groups have a slightly lower number of rigid-body degrees of freedom (*D*=6).[19] An interesting prediction follows from this analysis. If the possible space groups are expanded to include space groups that have improper symmetry elements, such as mirror planes and inversion centers, then the space group $P\bar{1}$ has the highest possible rigid-body degrees of freedom (*D*=8). This result implies that proteins will crystallize more readily from racemic solutions rather than homochiral solutions. With regard to racemic mixtures, the Kitaigorodskii close packing model predicts that $P2_1/c$ will be *one* of the most prevalent space groups among closely packed

molecular crystals, whereas the Wukovitz/Yeates entropic degrees of freedom model predicts that *P*$\bar{1}$ alone will be favored among racemic solvated macromolecular crystals.[7,19]

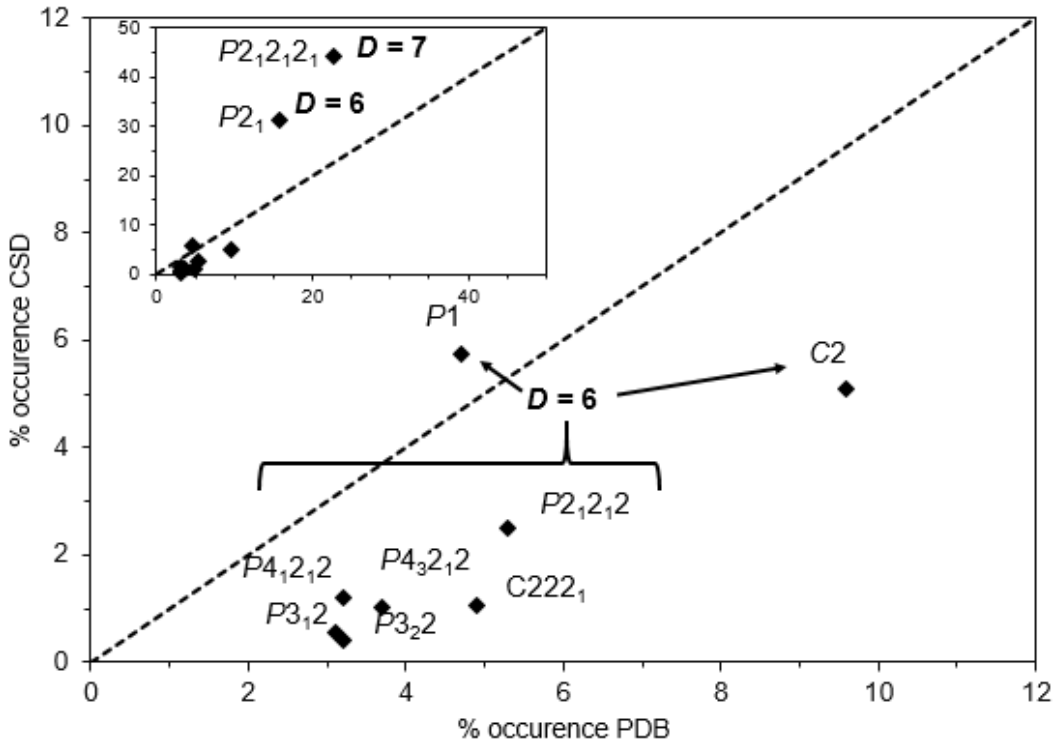


Figure 1.4. The occurrence of the top ten most prevalent space groups in the PDB does not show a one-to-one correspondence with the corresponding distribution in the CSD (normalized for Sohncke space groups). $P2_12_12_1$ has the highest rigid-body degrees of freedom (*D*=7). Wukovitz and Yeates performed a similar analysis in 1995. These data are updated to include all CSD structures as of January 2017.

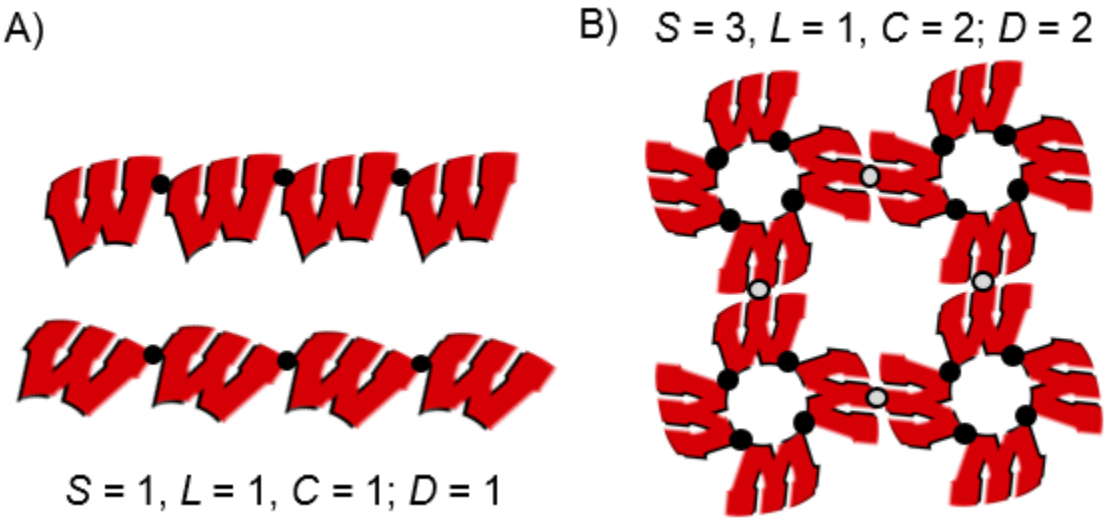


Figure 1.5. A) A one dimensional example of the Wukovitz/Yeates parameterization in a line lattice. There is one rotational degree of freedom (S=1), one unit cell parameter (L=1), and only one unique contact required for the 'W' to span a one dimensional lattice thus yielding D=1. B) In a two dimensional tetragonal plane lattice (p4) there are now two degrees of translational and one degree of rotational freedom (S=3), one lattice parameter due to the p4 symmetry (L=1), and now there are two unique contacts required (gray dot, black dot) to span the two dimensional plane (C=2) thus yielding D=2.

Zawadzke and Berg reported the first racemic protein crystal structure in 1993.[20,21] Their motivation seemed to primarily be amelioration of the phase problem that arises in macromolecular structure determination. They envisaged that a centrosymmetric protein crystal would, like small molecular crystals, be easier to phase given the centrosymmetric constraint on phase angles. Furthermore, the binary nature of centrosymmetric phases should result in higher quality electron density maps. Since these initial efforts a number of groups have pursued racemic protein crystal structures and in this endeavor, it appears the 'rate limiting step' has shifted from protein crystallogenesis to protein chemical synthesis, which is required to generate a protein comprising all D-amino acids.[22] Currently, there are 46 protein or nucleic acid structures in the PDB that occupy non-Sohncke space groups (Figure 1.6). Analysis of the space group distribution for these

racemic structures indicates that *P*ī may in fact prove to be the most favored space group among racemic macromolecular crystals; however, there are still too few structures to draw any meaningful statistical conclusions. Notably there have been four reported racemic macromolecular structures that crystallized in non-centrosymmetric space groups (*I*-4c2, *I*-4d2, *P*-3, and *R*-3). Protein homomers that contain inherent symmetry operators arising from quaternary contacts may follow a slightly different set of rules from those predicted by the Wukovitz/Yeates entropic degrees of freedom model, where protein monomers are implicitly assumed.[19] As more racemic homomeric protein structures become available, it will be very interesting to see the general trends that emerge.

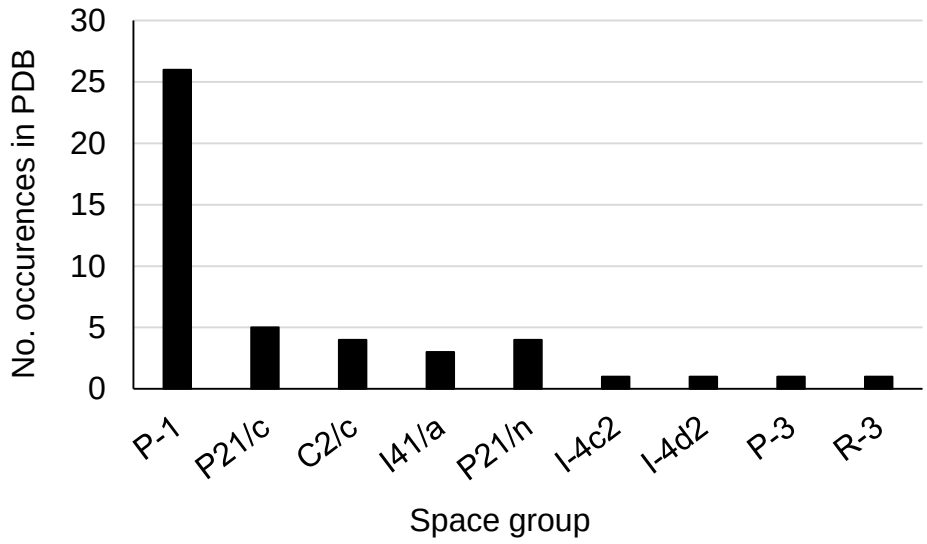


Figure 1.6. The distribution of space groups with improper symmetry elements in the PDB as of January 2017, there are a total of four non-centrosymmetric structures (I-4c2, I-4d2, P-3, and R-3)

The periodicity of a crystal lattice does not require that the constituent molecules within the lattice be identical or enantiomorphic (a molecule and its mirror image),

especially in light of the IUCr definition of a crystal as a solid 'having an essentially discrete diffraction diagram'. This flexibility suggests that nearly identical or enantiomorphic molecules could co-crystallize within the same lattice. Indeed, in the mid-nineteenth century Pasteur documented one of the first observations of this phenomenon during the course of his studies on the chirality of tartaric acid, when he co-crystallized a 1:1 *quasiracemic* mixture of ammonium (+)-bitartrate with ammonium (-)-bimalate (Figure 1.7).[23] Towards the end of the nineteenth century, before the advent of single crystal X-ray diffraction, Centnerszwer discovered that quasiracemic halogenated derivatives of succinic acid displayed altered melting point behavior relative to chiral mixtures of these derivatives.[14] In the mid-twentieth century, Fredga developed the 'steric correlation' method that allowed the relative configuration of compounds to be deduced from simple melting point diagrams.[24] Karle and Karle in collaboration with Fredga, provided the first single-crystal X-ray structure demonstrating that isosteric quasiracemates pack in a crystal lattice with pseudo-inversion symmetry (Figure 1.8).[25] Unfortunately, Fredga's steric correlation method appears to have been a victim of progress, the method was quickly displaced by Bijvoet's discovery that anomalous X-ray scattering could be used to unequivocally determine the absolute configuration of molecules with unknown stereochemistry.[26] More contemporary small molecule quasiracemic crystallization efforts by Davis and Wheeler involve the development of quasiracemic crystals that have non-linear optical properties.[27–29] Rational modifications in one of the quasi-enantiomers can tune the polarizability of a crystalline solid, and these 'tuned' quasiracemic crystals have potential applications in spectroscopy.

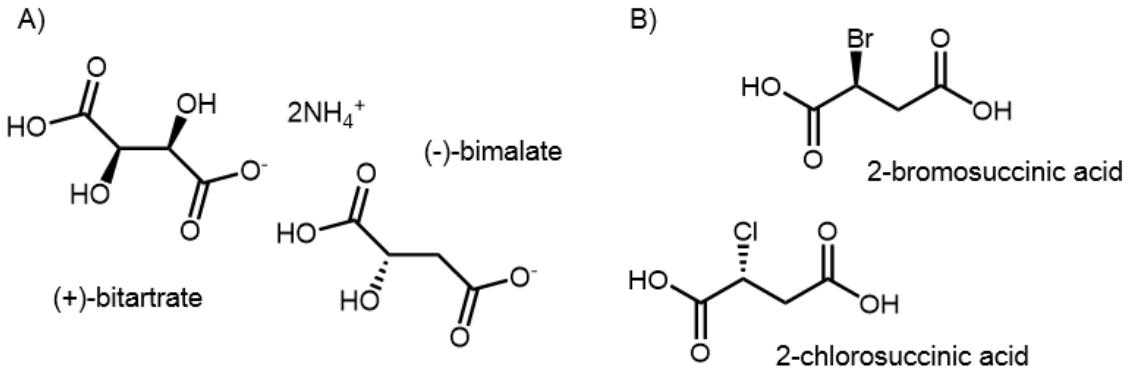


Figure 1.7. A) (+)-Bitartrate and (-)-bimalate, charge is neutralized with two ammonium ions. B) Quasiracemic halogenated derivatives of succinic acid.

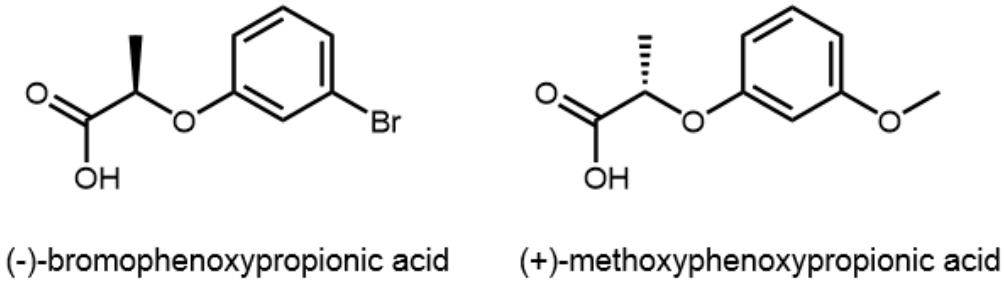


Figure 1.8. Karle and Karle published the quasiracemic structure of (-)-bromophenoxypropionic acid and (+)-methoxyphenoxypropionic acid, the first X-ray crystal structure of quasiracemates with pseudo-inversion symmetry. The structure was reported in space group C2 (presumably pseudo-C2/c).

A consistent theme has emerged: both small organic molecules and large solvated macromolecules, for varying reasons, tend to favor crystallization in centrosymmetric space groups. Given this fact, it seems reasonable to assume that quasiracemic protein mixtures would also tend to crystallize more readily in pseudo-centrosymmetric arrangements, similar to the packing arrangements observed for small molecule quasiracemates. Indeed, Pentelute *et al.* reported the first quasiracemic protein structure, of a snow flea anti-freeze protein, after initial attempts at homochiral crystallization had

failed (Figure 1.9A).[30,31] For the most part, protein crystallization is exceedingly more difficult than small molecule crystallization. Moreover, solution of the phase problem, now seen as trivial for small molecule crystal structure determination, still represents a significant challenge that must be overcome during macromolecular crystal structure determination. An explanation of the motivation behind the work of Pentelute *et al.* first requires a distinction between two philosophies concerning the application of the quasiracemic crystallization method to macromolecular crystallography. The first philosophy is concerned with alleviating difficulties associated with protein crystallogenesis and solving the phase problem, and this philosophy is evident upon reviewing the work of Pentelute *et al.* and others.[31–34] The second philosophy is more similar to the historical motivation behind small molecule quasiracemic crystallization, where quantifiable physical properties are ascribed to differences in molecular structure that arise from the point(s) of difference between quasienantiomers. Along these lines Mortenson *et al.* have suggested the application of quasiracemic crystallization as a tool for observing the structural differences between modified protein or peptide quasi-enantiomers (Figure 1.9B).[35]

Co-crystal structures of quasi-enantiomorphic pairs of proteins provide unique insights into the structural impact of modifications to the native protein structure. Specifically, these insights include the degree to which a quasi-enantiomorphic pair deviates from true inversion symmetry, assuming a centrosymmetric structure. Furthermore, the retention of quaternary contacts that preclude more ‘favorable’ symmetry operators, *i.e.*, inversion centers, is useful evidence for assessing the ‘intrinsic favorability’ of these quaternary contacts. Hayouka *et al.* have reported the first evidence

of such a scenario, where a quaternary contact between an antimicrobial peptide homodimer is retained and the modular dimer occupies a non-centrosymmetric space group, and the inherent dimeric molecular symmetry does not coincide with any crystallographic symmetry operators.[36]

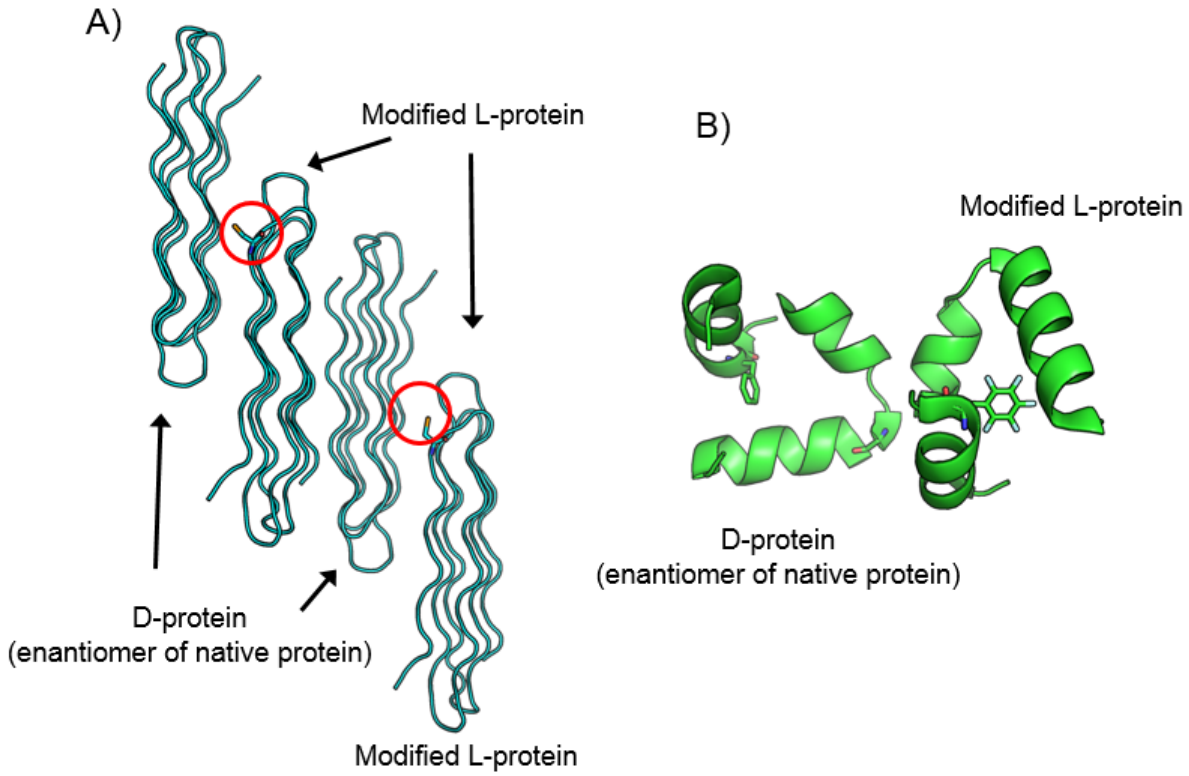


Figure 1.9. A) Quasiracemic protein structure of the snow flea anti-freeze protein reported by Pentelute et al. Incorporation of a selenium atom in the 'modified L-protein' results in a quasiracemic structure that yields an anomalous signal in the diffraction pattern, which can be exploited for experimental phase determination.[30] B) Mortenson et al. demonstrated that L-Villin headpiece with a fluorinated amino acid at a hydrophobic core position could co-crystallize with the enantiomer of the native Villin headpiece (D-VHP).[35]

Forward to Thesis

Each chapter in this thesis incorporates some aspect of the principles discussed in this introductory section, some more than others. In Chapter 2, I employ quasiracemic crystallization to explore the structural consequences of polyamide backbone

modifications within a mini-protein structure, and I also evaluate the impact these modifications have on the thermodynamic stability of the mini-protein. I use deviations from crystallographic inversion symmetry in each of the quasiracemic structures, in combination with the thermodynamic data, to identify the positions of backbone modification that are the most important for mini-protein stability. In Chapter 3, I demonstrate that the impact of a cyclic backbone constraint that varies in charge and polarity has a sequence-positional dependence on thermodynamic stability of the mini-protein system described in Chapter 2. The thermodynamic data presented in Chapters 2 and 3, combined with the quasiracemic structural data from Chapter 2, show that the modification position that is most tolerant in terms of stability also displays the smallest deviation from inversion symmetry. In Chapter 4, I describe a relationship between pseudo-inversion symmetry and the probability distributions of diffracted X-ray reflection intensities. I also discuss a method for quantifying pseudo-inversion symmetry in addition to the potential impact pseudo-inversion symmetry has on macromolecular crystal structure refinement and global quality indicators. Finally, in Chapter 5 I report three atomic-resolution crystal structures showing interactions between enantiomorphic peptide α-helices of opposite sense. I discuss the geometrical restraints of amino acid residues at the heterochiral interface in addition to a method for parameterizing such assemblies in which extracted model parameters are employed to corroborate these restraints.

1.2 References

(1) Dauter, Z.; Jaskolski, M. *FEBS J.* **2014**, *281* (18), 4010–4020.

(2) Pauling, L.; Corey, R. B.; Branson, H. R. *Proc. Natl. Acad. Sci. U. S. A.* **1951**, *37* (4), 205–211.

(3) Dunitz, J. D. *Nature* **2001**, *40*, 4167–4173.

(4) Desiraju, G. R. *J. Indian Inst. Sci.* **2014**, *94*, 1–8.

(5) Ramachandran, G. N.; Sasisekharan, V. *Biochim. Biophys. Acta - Biophys. Incl. Photosynth.* **1965**, *109* (1), 314–316.

(6) *International Tables for Crystallography*; Aroyo, M. I., Ed.; International Union of Crystallography: Chester, England, 2016; Vol. A.

(7) Kitaigorodskii, A. I. *Organic Chemical Crystallography*; Consultants Bureau: New York, 1961.

(8) Hales, T. C. *Discret. Comput. Geom.* **2006**, *36* (1), 5–20.

(9) Desiraju, G. R. *J. Am. Chem. Soc.* **2013**, *135* (27), 9952–9967.

(10) Desiraju, G. R. *Angew. Chemie - Int. Ed.* **2007**, *46* (44), 8342–8356.

(11) Kitaigorodskii, A. I. *Acta Crystallogr.* **1965**, *18* (4), 585–590.

(12) Brock, C. P.; Dunitz, J. D. *Chem. Mater.* **1994**, *6* (8), 1118–1127.

(13) Dunitz, J. D.; Gavezzotti, A. *Angew. Chemie - Int. Ed.* **2005**, *44* (12), 1766–1787.

(14) Spaniol, J. M.; Wheeler, K. A. *RSC Adv.* **2016**, *6* (69), 64921–64929.

(15) Brock, C. P.; Schweizer, W. B.; Dunitz, J. D. *J. Am. Chem. Soc.* **1991**, *113* (26), 9811–9820.

(16) Wilson, A. J. C. *Acta Crystallogr. Sect. A* **1993**, *49* (1), 210–212.

(17) Dunitz, J. D.; Gavezzotti, A. *J. Phys. Chem. B* **2012**, *116* (23), 6740–6750.

(18) Kantardjieff, K. A.; Rupp, B. *Protein Sci.* **2003**, *12* (9), 1865–1871.

(19) Wukovitz, S. W.; Yeates, T. O. *Nat. Struct. Biol.* **1995**, *2* (12), 1062–1067.

(20) Zawadzke, L. E.; Berg, J. M. *J. Am. Chem. Soc.* **1992**, *114* (10), 4002–4003.

(21) Zawadzke, L. E.; Berg, J. M. *Proteins Struct. Funct. Bioinforma.* **1993**, *16* (3), 301–305.

(22) Yeates, T. O.; Kent, S. B. H. *Annu. Rev. Biophys.* **2012**, *41*, 41–61.

(23) Wheeler, K. A.; Grove, R. C.; Davis, R. E.; Kassel, W. S. *Angew. Chemie Int. Ed.* **2008**, *47* (1), 78–81.

(24) Fredga, A. *Tetrahedron* **1960**, *8*, 126–144.

(25) Karle, I. L.; Karle, J. *J. Am. Chem. Soc.* **1966**, *88* (1), 24–27.

(26) Bijvoet, J. M.; Peerdeman, A. F.; van Bomme, A. J. *Nature* **1951**, *168* (4268), 271–272.

(27) Ashwell, G. *J. Mater. Chem.* **2000**, *9*, 1991–2003.

(28) Davis, F.; Hodge, P.; Liu, X.; Ali-Adib, Z. *Macromolecules* **1994**, *27* (7), 1957–1963.

(29) Ayyappan, P.; Evans, O. R.; Cui, Y.; Wheeler, K. A.; Lin, W. *Inorg. Chem.* **2002**, *41* (20), 4978–4980.

(30) Pentelute, B. L.; Gates, Z. P.; Tereshko, V.; Dashnau, J. L.; Vanderkooi, J. M.; Kossiakoff, A. A.; Kent, S. B. H. *J. Am. Chem. Soc.* **2008**, *130* (30), 9695–9701.

(31) Pentelute, B. L.; Gates, Z. P.; Dashnau, J. L.; Vanderkooi, J. M.; Kent, S. B. H. *J. Am. Chem. Soc.* **2008**, *130* (30), 9702–9707.

(32) Okamoto, R.; Mandal, K.; Sawaya, M. R.; Kajihara, Y.; Yeates, T. O.; Kent, S. B. *Angew Chem Int Ed Engl* **2014**, 5194–5198.

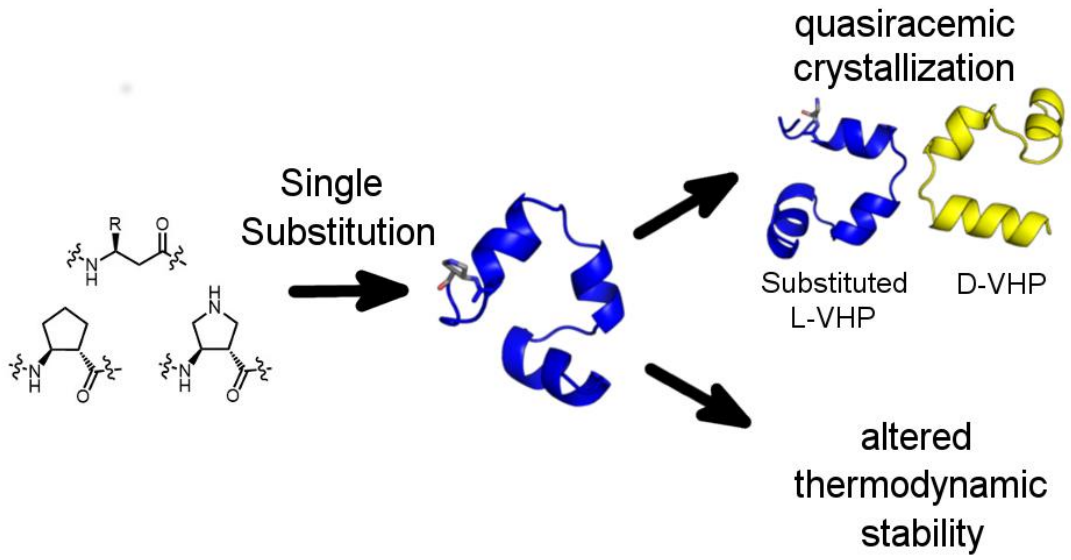
(33) Sawaya, M. R.; Pentelute, B. L.; Kent, S. B. H.; Yeates, T. O. *Acta Crystallogr. Sect. D Biol. Crystallogr.* **2012**, *68* (1), 62–68.

(34) Mandal, K.; Pentelute, B. L.; Bang, D.; Gates, Z. P.; Torbeev, V. Y.; Kent, S. B. H. *Angew. Chemie - Int. Ed.* **2012**, *51* (6), 1481–1486.

(35) Mortenson, D. E.; Satyshur, K. a; Guzei, I. a; Forest, K. T.; Gellman, S. H. *J. Am. Chem. Soc.* **2012**, *134* (5), 2473–2476.

(36) Hayouka, Z.; Thomas, N. C.; Mortenson, D. E.; Satyshur, K. A.; Weisblum, B.; Forest, K. T.; Gellman, S. H. *J. Am. Chem. Soc.* **2015**, *137* (37), 11884–11887.

Chapter 2 Effects of Single α to β Residue Replacements on Structure and Stability in a Small Protein: Insights from Quasiracemic Crystallization



Portions of this chapter have been published in:

Kreitler, D. F.; Mortenson, D. E.; Forest, K. T., Gellman, S. H. *J. Am. Chem. Soc.* **2016**, *138*, 6498-6505.

Abstract

Synthetic peptides that contain backbone modifications but nevertheless adopt folded structures similar to those of natural polypeptides are of fundamental interest and may provide a basis for biomedical applications. Such molecules can, for example, mimic the ability of natural prototypes to bind to specific target macromolecules but resist degradation by proteases. Previous results have shown that oligomers containing mixtures of α- and β-amino acid residues ("α/β-peptides") can mimic the α-helix secondary structure, and that properly designed α/β-peptides can bind to proteins that evolved to bind to α-helical partners. The work in this chapter describes fundamental studies that support the long-range goal of extending the α/β approach to tertiary structures. The impact of single α→β modifications on the structure and stability of the small and well-studied villin headpiece subdomain (VHP) was evaluated. The native state of this 35-residue polypeptide contains several α-helical segments packed around a small hydrophobic core; α→β substitution at five solvent-exposed positions, Gly11, Asn19, Trp23, Gln26 and Lys30 was examined. In each case except Gly11, both the β^3 homologue of the natural α residue and a cyclic β residue were evaluated; only a cyclic β residue was evaluated at Gly11. All α→β^3 substitutions caused significant destabilization of the tertiary structure as measured by variable-temperature circular dichroism, although at some of these positions, replacing the β^3 residue with a cyclic β residue led to improved stability. Atomic-resolution structures of four VHP analogues were obtained via quasiracemic crystallization. These findings contribute to a fundamental α/β-peptide knowledge-base by confirming that β^3-amino acid residues can serve as effective

structural mimics of homologous α-amino acid residues within a natural tertiary fold, which should support rational design of functional α/β analogues of natural poly-α-peptides.

2.1 Introduction

Proteins perform a broad range of functions; most activities depend on the ability of a polypeptide chain to adopt a specific shape and thereby present a particular set of side-chains in a defined three-dimensional arrangement. The wide variation in shape and function among natural proteins is achieved with a backbone that contains exclusively α-amino acid residues. Efforts to develop analogues of natural polypeptides that display altered activities have focused largely on side-chain modification, a choice that is based on both synthetic and structural considerations. Efficient ribosomal biosynthesis is limited to sequences that contain primarily L-α-amino acid residues,[1,2] and solid-phase synthesis, which offers access to greatly enhanced side-chain diversity at the cost of diminished chain length, has been heavily optimized for α-amino acid incorporation. The most common protein secondary structures, including α-helices, 3_{10}-helices, β-sheets and β-turns, depend on specific hydrogen bonding patterns involving backbone amides as donors and acceptors (Figure 2.1); therefore, it might be assumed that altering the covalent connectivity among backbone amide groups, as must occur if α-amino acid residues are replaced with other types of amino acid residues, would disrupt the folding patterns necessary for the functions of natural polypeptides.

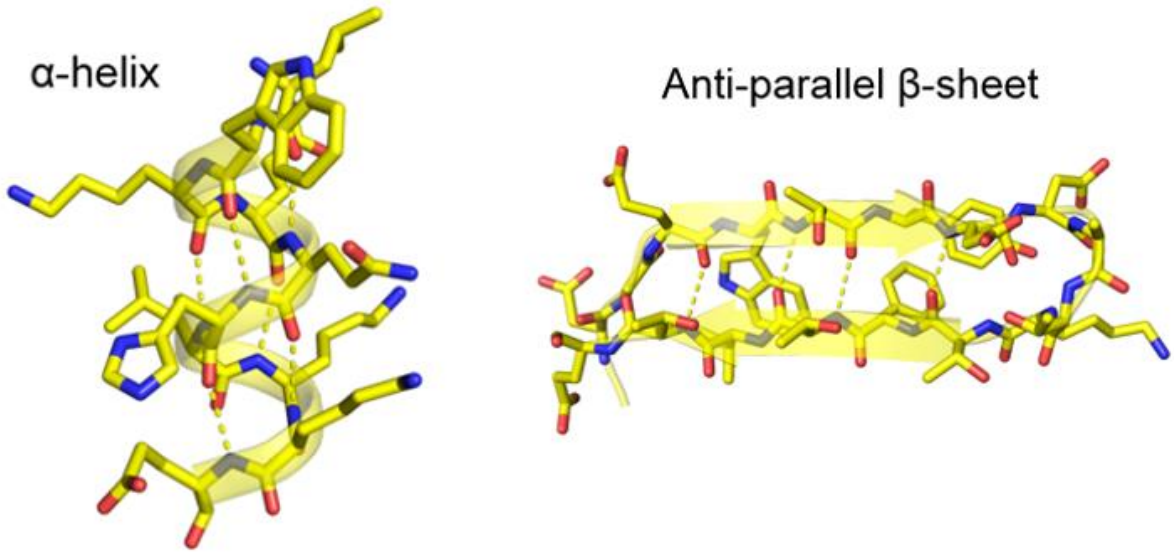


Figure 2.1. The i+4 H-bonding pattern in the α-helix can be modulated by modifications to the peptide backbone in the form of β-amino acids. Efforts to modulate H-bonding patterns in β-sheets with unnatural residues have been less successful. (left panel) An anti-parallel β-sheet extracted from PDB: 4KGR and (right panel) an α-helix extracted from PDB: 1YRF display the characteristic H-bonding patterns.

The work described in this chapter assesses the impact of replacing a single α-amino acid residue with a β-amino acid residue in the context of the small and well-studied tertiary structure formed by the villin headpiece subdomain (VHP).[3–5] Each variant containing an α→β replacement is compared with VHP itself in terms of stability, via thermal denaturation, and in several cases the structures are compared via X-ray crystallography. To place these findings in context, the precedents for α→β replacement within polypeptides are reviewed and knowledge gaps are identified.

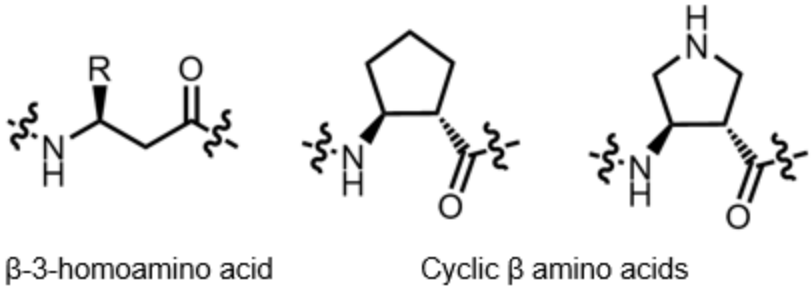


Figure 2.2. β-3-homoamino acids include the native α-amino acid side-chain with an additional methylene unit adjacent to the carbonyl. Hydrophobic and cationic cyclic β-amino acids can adopt dihedral angles that are compatible with an α-helix.

Efforts focused on α-helix mimicry by backbone-modified polypeptides have revealed that the structural and informational features of natural α-helices can be recapitulated with oligomers that contain α- and β-amino acid residues in simple patterns.[6–9] The secondary structure adopted by these α/β-peptides is very similar to the canonical α-helix despite the presence of 'extra' backbone carbon atoms (Figure 2.2), although the impact of α→β replacement on helix stability is not presently clear. Initial studies of α-helix-mimetic α/β-peptides were motivated by the hypothesis that such oligomers would maintain the recognition properties of a prototype α-helix but be less susceptible to enzymatic degradation.[10] This hypothesis has been borne out in several systems in which the recognition of multi-turn α-helices by partner proteins can be mimicked with α/β-peptides containing 25-33% β residues evenly distributed along the sequence.[6,11,12] An unanticipated outcome of these studies was the discovery that some patterns of α→β replacement engender unique recognition selectivities among sets of related binding proteins[13] or even among functionally distinct conformations of a single binding protein.[14,15]

Systematic evaluation of α→β replacements in the context of β-sheet secondary structure has been conducted by Horne et al.[16,17] This work showed that a two-stranded "β-hairpin" conformation can be maintained when cross-strand pairs of α residues are replaced with β residues; however, such insertions alter the global pattern of side-chain display relative to the prototype all-α hairpin. The side-chain display can be more effectively maintained if each β residue replaces a sequentially adjacent pair of α residues.[17] Martinek et al. have reported that the anti-angiogenic properties of a designed β-sheet-forming peptide can be retained after a few α→β replacements[18] despite the alterations in side-chain positioning that would be predicted based on the findings of Horne et al.[16,17]

A long term goal in this field is to understand impacts of α→β replacements not only on local secondary structure and binding properties but also on tertiary structure and biological function. To date, pioneering exploratory studies have led to variable outcomes; a unifying set of rules governing impacts of such replacements is still needed. Analogues of RNase A in which a two-residue reverse turn was replaced with a di-β-peptide displayed native-like catalytic function.[19] Similarly, David et al. described an IL-8 analogue in which the C-terminal α-helix was replaced with a geometrically divergent α-peptide helix or β-peptide segment; signaling activity, mediated by the N-terminal portion, was retained by these hybrid polypeptides.[20] However, Denton et al. found that replacing the α-helical C-terminal portion of the hormone GLP-1 with the same sort of α-peptide helix led to a million-fold loss in activity.[12] Similarly, experiments involving analogues of parathyroid hormone suggested that even limited and conservative backbone alterations could be highly deleterious to natural function.[11,21] Thus, strategic substitution of a natural

secondary structure element with a backbone-modified segment can be tolerated in some cases, but not all.

Recent reports have extended the periodic α→β replacement strategy to α-helices within model tertiary or quaternary structures. These studies are significant because the results pertain to larger and more irregular surfaces than those presented by a single helix. Horne et al. explored multiple α→β replacements within the small protein GB1.[22,23] Crystal structures of several backbone-modified GB1 variants showed that a native-like tertiary structure was retained in each case (Figure 2.3A). Hilvert et al. examined α→β replacements in an α-helical segment within an engineered two-component chorismate mutase.[24] This helix contributes a key arginine side-chain to the active site. Several α→β substitution patterns were found to support near-native levels of catalytic activity. Checco et al. developed α/β-peptides that adopt a helix-loop-helix tertiary structure and target specific binding surfaces on partner proteins (Figure 2.3B).[25,26] Crystallographic characterization of an α/β-peptide bound to a protein partner (vascular endothelial growth factor, VEGF) showed that α-helix-like secondary structure was maintained despite the presence of multiple β residues.

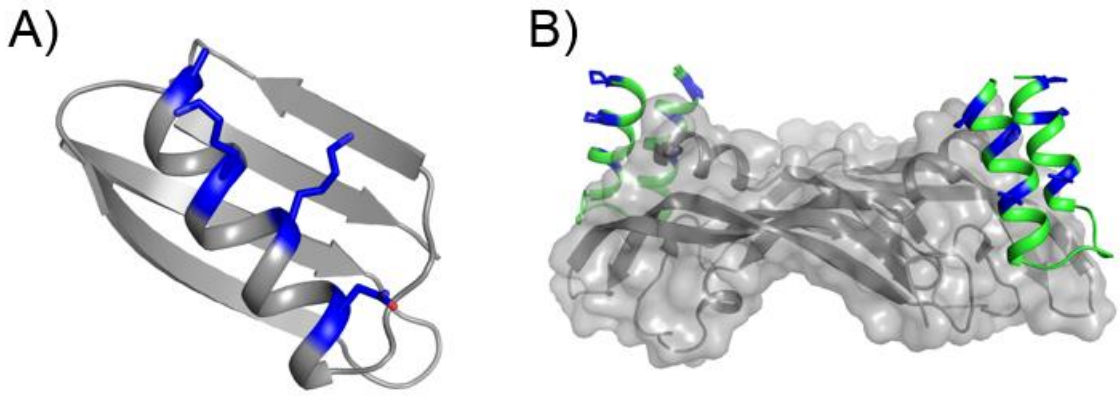


Figure 2.3. Higher order tertiary structures have recently been shown to accommodate backbone modifications in the form of β amino acids. A) X-ray crystal structure of a cooperatively folding tertiary structure with a heterogeneous backbone containing β-3-homoamino acid substitutions (shown in blue, PDB: 4KGR). B) An α/β peptide affibody (α-residues, green; β or Aib residues, blue; PDB: 4WPB) derived from an all α-residue scaffold in complex with vascular endothelial growth factor (VEGF, grey; α-residues, green).

Collectively, the studies of α→β replacement summarized above suggest that certain substitutions can result in partial or full retention of a prototype poly-α-peptide's structure and function; however, most of the precedents fail to provide insight on the impact of α→β replacement on conformational stability. The most incisive results in this regard involved α→β replacements within the lone α-helix of the GB1 tertiary structure. Each of the helix-modified GB1 variants contained four α→β replacements, which resulted in significant destabilization of the native fold.[22] My work complements the GB1 work in that a different tertiary structure, VHP, was employed and single α→β and acyclic → cyclic β residue replacements were evaluated.

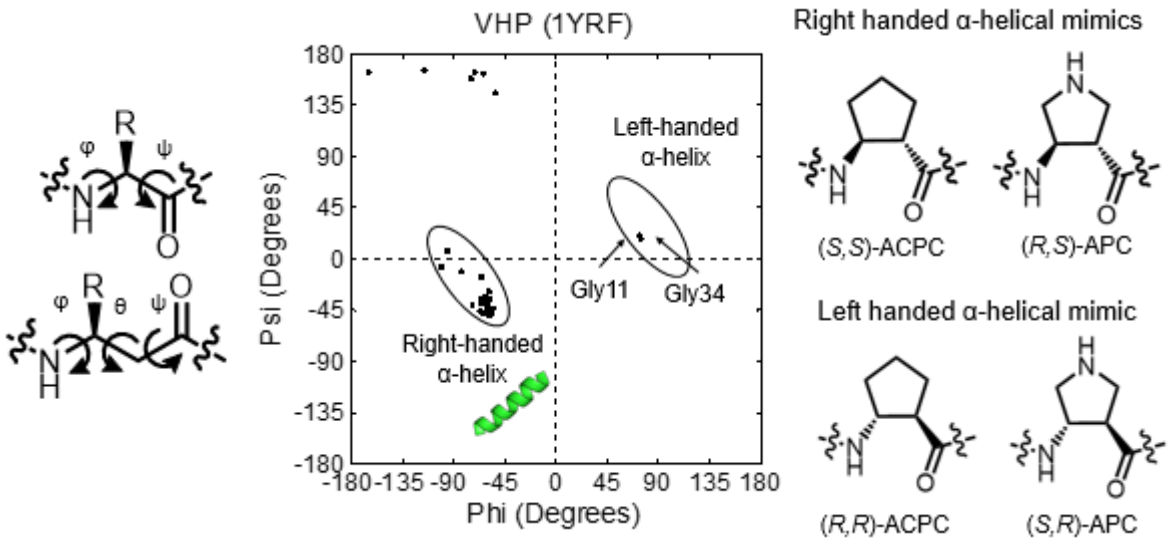


Figure 2.4. β-amino acids introduce an additional rotational degree of freedom in the peptide backbone. Cyclic β-amino acids can mimic dihedral angle space that correspond to an α-helix. In practice, (S,S)-ACPC and (R,S)-APC are predominantly employed for right-handed α-helical mimicry. In principle, left-handed α-helix mimicry is also possible via inversion of stereochemistry with the enantiomers (R,R)-ACPC and (S,R)-APC. The Ramachandran plot for VHP (PDB: 1YRF) indicates two positions that have positive φ values, both involving Gly residues.

The villin headpiece (VHP) is one of the smallest poly-α-peptides known to form a discrete tertiary structure without internal crosslinks resulting from disulfide formation or metal ion chelation, either of which could convolute the thermodynamic impact of backbone modifications on a tertiary structure.[3,4,27] The experimental design employed here focuses on single α→β replacements at five solvent-exposed positions within VHP; each position has helical dihedral angles (Figure 2.5). At four of the five positions, both a β^3 and a cyclic β residue were evaluated. Circular dichroism (CD) was used to determine the impact of each α→β replacement on the extent of α-helix formation, which provides an indirect indication of tertiary structure formation; variable-temperature CD data allow us to assess the effects of each α→β replacement on tertiary structural stability. For each of the four native α residues that were replaced with the β^3 homologue (identical side-

chain), the single-site backbone modification caused a decline in tertiary structure stability. At some but not all of the substitution sites, use of a conformationally constrained β residue led to a recovery of tertiary structural stability relative to the more flexible, acyclic β^3 substitution. Quasiracemic crystallography enabled high-resolution structural analysis of several VHP variants and facilitated comparisons between local conformations of β residues and the α-residues they replaced. Overall, these findings contribute to a small but fundamental knowledge-base that should inform rational design of functional α/β analogues of natural poly-α-peptides.

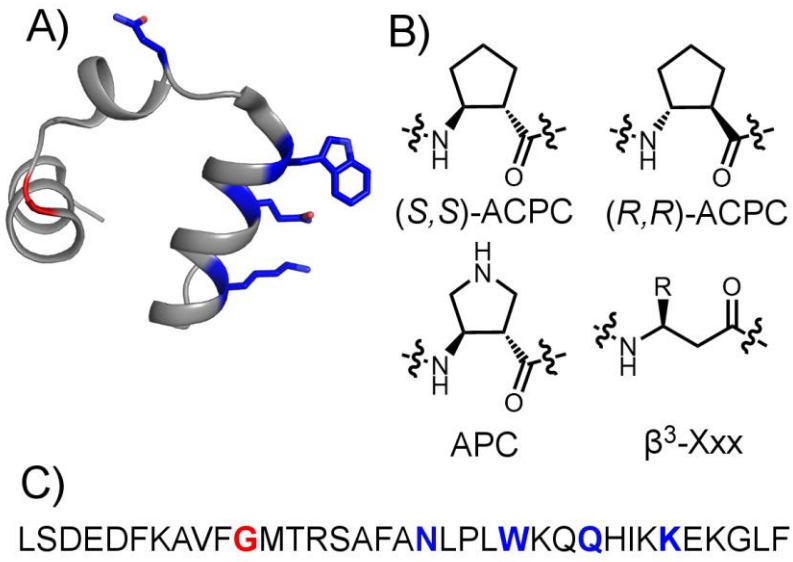


Figure 2.5. A) A previously reported crystal structure of the Villin headpiece subdomain (VHP; PDB entry 1YRF), with positions at which β residues were substituted in this study shown in red and blue. B) Structures of a generic β^3-homoamino acid residue, the *trans*-(1*S*,2*S*)-2-aminocyclopentyl-1-carboxylic acid (ACPC) residue and its enantiomer and the *trans*-(3*R*,4*S*)-4-aminopyrrolidine-3-carboxylic acid (APC) residue. C) β^3-homoaminoacid substitutions were made individually for native residue Asn19, Trp23, Gln26, or Lys30, the positions of which are shown in blue within the native sequence; in addition, ACPC substitutions were made individually for native residue Asn19, Trp23, or Gln26, and APC was substituted for Lys30. Finally, an (*R*,*R*)-ACPC substitution was made at Gly11.

2.2 Results and Discussion

2.2.1 Design of Substitution Positions

The first round of replacements in VHP involved β^3-homoamino acid residues homologous to the original α residues (Figure 1, blue positions); thus, Asn was replaced by β^3-hAsn, Trp by β^3-hTrp, etc. In these replacements, the side-chain found in the natural VHP domain is retained, but an extra methylene is inserted into the backbone. The second round of replacements involved cyclic β residues. In these cases, the native amino acid side-chains were not retained. For positions with uncharged side-chains (Asn19, Trp23 and Gln26), the cyclopentane-based β residue designated (*S*,*S*)-ACPC was used, while for the position with a basic side-chain (Lys30), the pyrrolidine-based β residue designated APC was used (Figure 1). Finally, Gly11 was replaced with (*R,R*)-ACPC, the enantiomer of (*S,S*)-ACPC, to accommodate the positive φ value of the Gly11 dihedral angle (Figure , which corresponds to a left-handed α-helix. All VHP analogues were prepared via standard Fmoc-based solid-phase synthesis. Previous work with peptides that adopt α-helical secondary structure has established that five-membered ring constraint with *trans* stereochemistry predisposes the β residue backbone to adopt an α-helix-like local conformation.[6] Crystallographic data from Horne et al. show that the α-helix-like local conformation of the ACPC residue is manifested also in backbone-modified variants of the GB1 tertiary structure.[22,23]

2.2.2 Circular Dichroism Spectra of β Substituted VHP Variants

The impact of each α→β replacement on tertiary structural stability was assessed via variable-temperature circular dichroism (CD) measurements made in 20 mM sodium

acetate (pH 5.0). VHP itself and each of the new variants displays a CD spectrum with two minima in the far-UV region, one near ~223 nm and the other near ~208 nm, as expected for an α-helix-rich folding pattern (Figure 6). The intensities of the CD signatures are generally similar among these polypeptides, but it is noteworthy that the ~208 nm minimum is a little more intense for every β-containing variant relative to VHP itself. Previous work has shown that α/β-peptides that contain multiple substitutions throughout the sequence and adopt α-helix-like conformations display only one far-UV CD minimum, near ~208 nm.[28] The G11(*R,R*)-ACPC substitution is not included in this analysis. The overall shape of the spectrum of this variant is consistent with the spectra of other variants (local minima at ~208 and 223 nm).

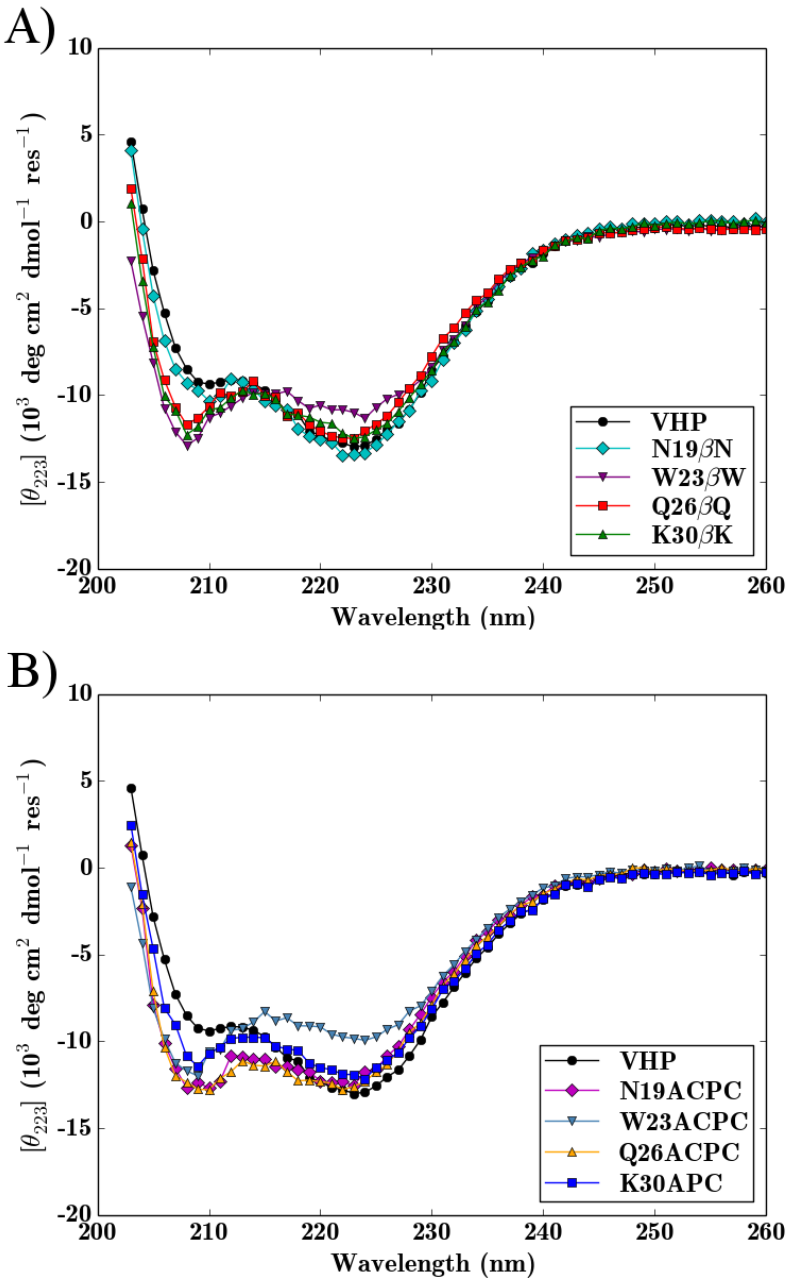


Figure 2.6. Far-UV circular dichroism (CD) data for β substituted VHP derivatives containing (A) acyclic (β^3) or (B) cyclic β-amino acid substitutions. CD data were obtained in 20 mM aqueous sodium acetate buffer, pH 5.0, at 50 μM peptide. Prior to each set of measurements the CD spectrometer was calibrated with a 1 mg/mL solution of (1S)-(+)-10-camphorsulfonic acid.

2.2.3 Thermodynamic Stability of β Substituted VHP Variants

Fitting the variation in CD intensity at 223 nm vs. temperature to a two-state model (folded vs. unfolded) allowed for the estimation of a melting temperature (T_m) and an unfolding enthalpy (ΔH_m) for each β substituted VHP variant (Figure 2.7, Table 2.1). For each of the four sites evaluated, replacing the original α residue with the β^3 homologue leads to a substantial decline in tertiary structure stability, with ΔT_m = [T_m(variant) – T_m(VHP)] ranging from -12° to -25°C. This trend is consistent with the substantial declines in tertiary structure stability observed by Horne *et al.* for GB1 variants containing four α→β^3 substitutions in the helical segment.[22] Replacing a β^3 residue with a cyclic β residue in the context of VHP led to divergent outcomes. For β^3hTrp23 and β^3hGln26, replacement with ACPC caused little change in T_m. In contrast, β^3→cyclic β was stabilizing at the other two positions: replacing β^3hAsn19 with ACPC increased T_m by ~10° C, and replacing β^3hLys30 with APC increased T_m by ~12°C. The variant bearing APC at position 30 is indistinguishable from native VHP in terms of thermal stability. A helix C-capping motif at Gly11, was replaced with (*R,R*)-ACPC, which resulted in a cooperatively folding variant that was destabilized by ~15°C relative to the native VHP fold.

Cyclically constrained β residues have been suggested to promote helical secondary structure through a preorganization of helical dihedral angles that lowers the entropic penalty of folding in isolated α-helices,[6] although this view has been questioned.[24] Our results show β^3→cyclic β replacement does not necessarily enhance conformational stability in the context of a tertiary structure, because considerable variation is observed among the four substitution sites in VHP. Chiu *et al.* demonstrated

that norleucine (Nle) substitution for Lys24 had a stabilizing effect on the folded conformation of VHP, presumably because of the removal of an unfavorable His27-Lys24 side-chain interaction (Coulomb repulsion) at pH 4.8.[4] Kubelka *et al.* showed that double Nle substitution at Lys24 and Lys29 provided additional stabilization relative to the single Lys24Nle substitution, presumably because the aliphatic Nle29 side-chain could pack into the hydrophobic core.[29] Therefore, perturbations of local packing or Coulombic interactions sites may explain, in part, the decrease in T_m that results from β^3 substitution for W23, Q26 or K30, and for cyclic β residue substitution for W23 or Q26.

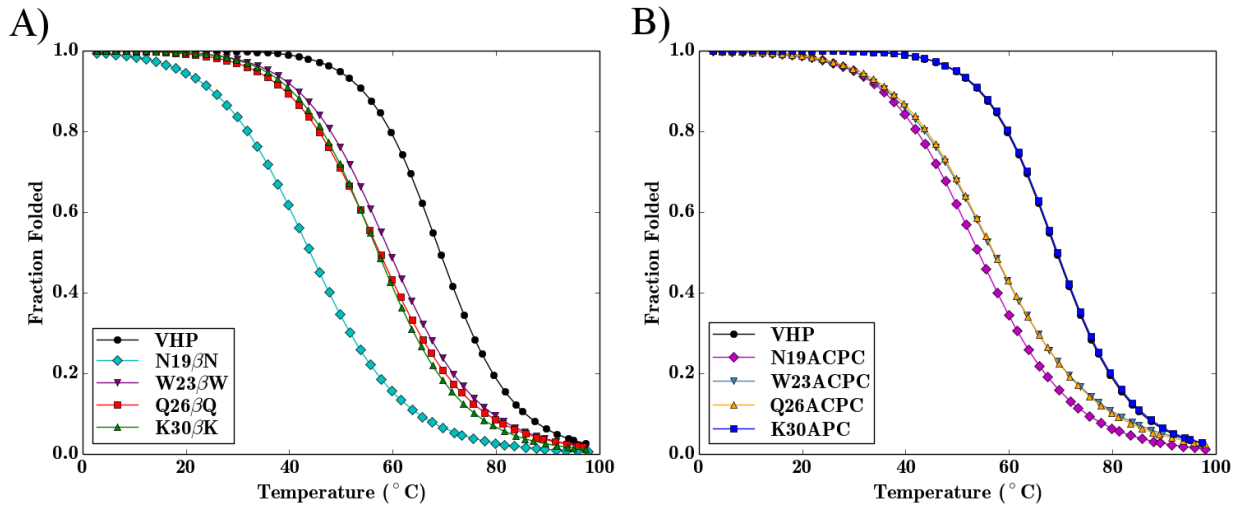


Figure 2.7. Variable-temperature circular dichroism (vt-CD) measurements were performed by monitoring the ellipticity at 223 nm for VHP and variants containing either A) a single acyclic (β^3) or B) a single cyclic β residue substitution. Data were acquired in 20 mM aqueous sodium acetate, pH 5.0, at 50 μM peptide. For each variant, model parameters (T_m, ΔH_m) and baselines were fit using non-linear regression analysis averaged over three independent measurements. Results are reported as fraction folded. VHP G11(R,R)-ACPC is not shown.

Table 2.1. CD-derived folding/unfolding parameters for native VHP and β substituted variants.[a]

Variant	T_m (°C)	ΔH_m (kcal/mol)	ΔT_m (°C)	$\Delta\Delta H_m$ (kcal/mol)
VHP	69.1 (±0.4)	32.8 (±1.2)	-	-
N19β^{3}N	44.2 (±1.7)	22.3 (±2.8)	-25.0 (±2.1)	-10.5 (±4.0)
W23β^{3}W	59.4 (±0.7)	25.9 (±0.6)	-9.7 (±1.1)	-6.9 (±1.8)
Q26β^{3}Q	57.4 (±0.9)	24.8 (±1.5)	-11.7 (±1.3)	-8.1 (±2.7)
K30β^{3}K	57.2 (±1.7)	27.2 (±1.3)	-11.9 (±2.1)	-5.6 (±1.6)
N19ACPC	54.0 (±1.1)	24.0 (±0.6)	-15.1 (±1.5)	-8.8 (±1.8)
W23ACPC	56.9 (±0.4)	22.0 (±3.9)	-12.2 (±0.8)	-10.8 (±1.6)
Q26ACPC	57.1 (±1.0)	22.1 (±0.5)	-12.1 (±1.4)	-10.7 (±1.7)
K30APC	69.4 (±1.2)	33.0 (±2.7)	0.3 (±1.6)	-0.1 (±3.9)
G11(*R,R*)ACPC	54.3 (±0.8)	23.0 (±1.5)	-14.8 (±1.2)	-9.8 (±2.7)

[a]The two-state folding parameters T_m and ΔH_m were derived by assuming that $\Delta C_p = 0$, ΔH_m is reported as the unfolding enthalpy (positive sign). ΔT_m and $\Delta\Delta H_m$ are defined as the value for the variant with the indicated β-substitution minus the value for native VHP. The standard deviation for each reported parameter was determined from three separate measurements and fitting sessions.

In practice, determining the change in heat capacity of folding, ΔC_p, from a single spectroscopic thermal denaturation experiment is difficult.[30] Typically, careful differential scanning calorimetric (DSC) measurements are required to accurately measure ΔC_p, and assigning folded and unfolded baselines can be problematic for small peptides that exhibit broad, marginally cooperative unfolding transitions.[31] In the present work, ΔC_p of the VHP β substituted variants is assumed to be zero. This assumption was made for two reasons. First, ΔC_p scales linearly with the change in both polar and non-polar solvent accessible surface area upon folding (ΔASA_p and ΔASA_{np}).[32,33] This phenomenon and its molecular basis have been extensively described and parameterized in the literature.[32–35] The change in solvent accessible surface area upon folding is linearly correlated with molecular weight, so that the magnitude of ΔC_p is small (~0.5 kcal/mol) for a small peptide

such as VHP (~4 kDa). Secondly, the uncertainty of ΔC_p from DSC measurements can range from 5% to 20% even in ideal cases.[33] Therefore, quantifying the differential impact of single β amino acid substitutions on ΔC_p would be extremely challenging. For instance, Kubelka *et al.* employed DSC measurements to determine the free-energy barrier height for the VHP two-state folding transition.[31] ΔC_p values of 0.457, 0.636, and 0.764 kcal/mol were reported, depending on the method used to determine the folded and unfolded baseline regions. Positive ΔC_p values result in a convex stability curve, $\Delta G(T)$ (Figure 2.8, red line), and when ΔC_p is assumed to be zero, $\Delta G(T)$ is a linear function of T with slope $-\Delta H_m/T_m$ (Figure 2.8, blue line). In the neighborhood of T_m a linear stability curve is a close approximation to a convex stability curve.[36]

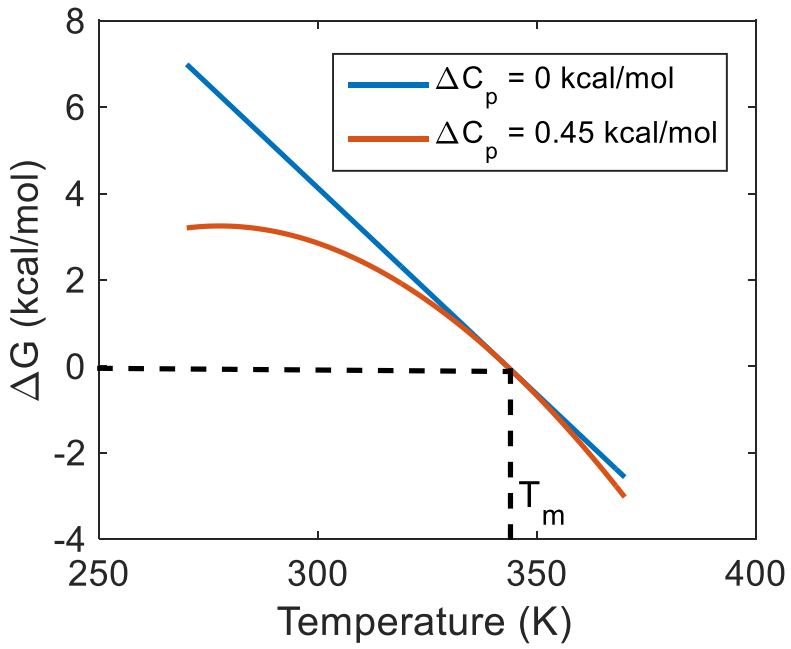


Figure 2.8. For L-VHP, a linear stability curve closely approximates a convex stability curve within the neighborhood of the transition temperature, Tm, which is indicated at 342.3 K.

2.2.4 Impact of β Substitutions on the Unfolding Enthalpy

The unfolding enthalpy, ΔH_m, primarily comprises two terms: disruption of internal non-covalent interactions, such as van der Waals or hydrogen bonding interactions, and hydration of buried polar functional groups. These terms have opposite signs and large magnitudes relative to the marginal difference between the two terms.[33] Additional non-covalent interactions that can also play a role include favorable salt-bridging or Coulombic interactions dependent upon the ionization state of key functional groups. In the present dataset, acyclic β^3 substitutions generally lowered the unfolding enthalpy, ΔH_m, by approximately 5-10 kcal/mol relative to the unfolding enthalpy of native VHP. The largest decrease in ΔH_m was observed at Asn19. In addition, cyclic β substitutions lowered ΔH_m by ~10 kcal/mol at positions Gly11, Asn19, Trp23, and Gln26. In contrast, at position Lys30 ΔH_m was unaffected. The difference in unfolding enthalpy, ΔH_m, between acyclic β^3 versus cyclic β substituted variants was dependent upon the substitution position. At Asn19 there was no measurable difference in ΔH_m between the acyclic β^3 substituted variant (N19β^3N) and the cyclic β substituted variant (N19ACPC); at Trp23 and Gln26, ΔH_m was slightly larger for the acyclic β^3 substituted variants (W23β^3W, Q26β^3Q) versus the cyclic β variants (W23ACPC, Q26ACPC); finally, at position Lys 30, ΔH_m was smaller for the acyclic β^3 substituted variant (K30β^3K) compared to the cyclic β substituted variant (K30APC). Each β substitution, with the exception of K30APC, decreased the magnitude of the unfolding enthalpy, relative to the unfolding enthalpy of native VHP (32.8 kcal/mol; Figure 2.9). The marginal difference of ΔH_m between native VHP and β substituted variants is represented as $\Delta\Delta H_m$.

These trends are consistent with similar observations made by Reinert and Horne on multiple α → acyclic β^3 and acyclic β^3 → cyclic β substitutions in the GB1 system; however, the $\Delta\Delta H_m$ values reported here are substantially larger than the values reported by Reinert and Horne, who postulated the following two mechanisms by which these substitutions decreased unfolding enthalpy. The first mechanism involved loss of enthalpically favorable electrostatic interactions upon acyclic β^3 to cyclic β substitution (β^3hLys → ACPC), and the second mechanism was the loss of a potential $n \rightarrow \pi^*$ interaction (due to the additional backbone methylene unit in the β residue) upon α → β substitution. In the present VHP-derived dataset, the former mechanism can be ruled out because cyclic β substitutions were made at residues without ionizable sidechains. The latter mechanism may account, in part, for the decreased unfolding enthalpy among each β substituted VHP variant, but it does not explain the sequence dependence of the marginal difference between the unfolding enthalpies of the β substituted VHP variants and the unfolding enthalpy of native VHP, $\Delta\Delta H_m$. In addition, acyclic β^3 → cyclic β substitutions inexorably result in sidechain removal, which could decrease the unfolding enthalpy if a particular sidechain is involved in an enthalpically favorable contact in the folded conformation.

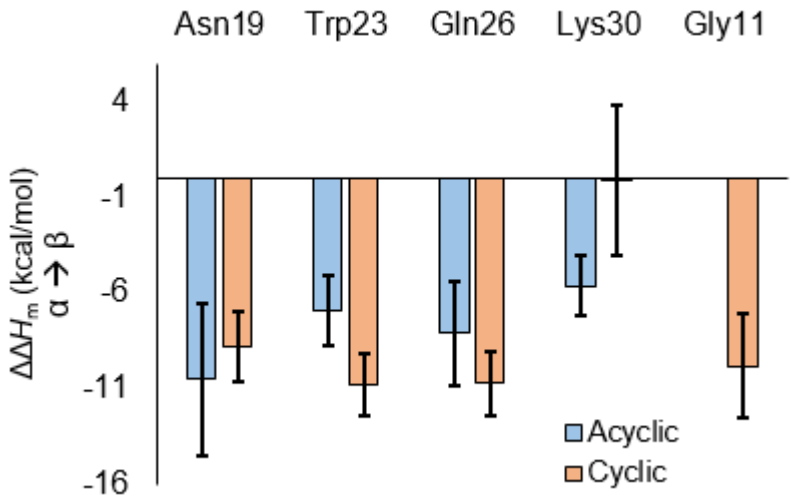


Figure 2.9. The difference in the unfolding enthalpy of β substituted variants and the unfolding enthalpy of native VHP, $\Delta\Delta H_m$, is shown for both acyclic and cyclic β substitutions at each position ($\Delta\Delta H_m = [\Delta H_{m,variant} - \Delta H_{m,VHP}]$). Error bars represent the sum of the uncertainty in individual ΔH_m values for VHP and the VHP variant.

2.2.5 Impact of β Substitutions on the Unfolding Entropy

Contrary to the notion that the additional rotational degree of freedom resulting from an acyclic β^3 substitution should increase the entropic penalty of folding, Reinert and Horne demonstrated that multiple $\alpha \rightarrow \beta^3$ substitutions in the GB1 system actually decreased the entropic penalty of folding.[23] This observation was ascribed to the contribution of solvation entropy. In the denatured ensemble, solvation of the 'extra' CH_2-unit in the β residue backbone may incur an entropic penalty that is relieved upon desolvation as the backbone is buried upon folding. In general, hydration entropy is considered to outweigh conformational entropy in the total unfolding entropy.[37] The present dataset is in agreement with the observations of Reinert and Horne. Singular $\alpha \rightarrow \beta^3$ substitutions in VHP decrease the unfolding entropy by approximately 13 to 25 cal/mol*K, depending upon the substitution position (Table 2.2). The uncertainty of ΔS_m among the β substituted VHP variants was too large to make significant comparisons

between unfolding entropies of acyclic β^3 and cyclic β substituted variants (uncertainty of ΔS_m was propagated from uncertainty in T_m and ΔH_m). The same general trends that were observed for the folding enthalpy are persistent in the folding entropy, which is reasonable considering that ΔS_m, which is independent of temperature due to the assumption that ΔC_p is zero, is simply the folding enthalpy scaled by the inverse of the fitted melting temperature. In addition to constraining backbone dihedral angles, acyclic β^3 → cyclic β substitutions remove the native side-chain. In the folded conformation the number of possible side-chain rotamers will be constrained relative to the number of accessible side-chain conformations in the denatured ensemble. This restriction incurs an additional penalty on the folding entropy. Therefore, replacement of flexible side-chains with side-chains that are pre-organized should have a favorable impact on the folding free energy. In the present data set however, the uncertainty of ΔS_m does not allow for any direct β^3 → cyclic β comparisons on the folding entropy among the variants (Table 2.2).

Table 2.2. The unfolding entropy, i.e. $\Delta H_m/T_m$, is listed for each β substituted VHP variant. $\Delta\Delta S_m$ is defined as $\Delta S_{m,acyclic} - \Delta S_{m,cyclic}$. The impact of the corresponding α side-chain has on folding entropy is also listed.

Variant	ΔS_m (cal/mol*K)	$\Delta\Delta S_m$ (cal/mol*K) α → β	Lee *et al*.[38] (α side-chain deletion, cal/mol*K)
VHP	95.8 (±3.5)	-	-
N19β^3N	70.3 (±8.8)	-25.6 (±12.3)	-
W23β^3W	77.9 (±1.8)	-18.0 (±5.3)	-
Q26β^3Q	75.0 (±4.5)	-20.8 (±8.1)	-
K30β^3K	82.3 (±4.0)	-13.5 (±7.5)	-
N19ACPC	73.4 (±1.9)	-22.5 (±5.4)	-3.3
W23ACPC	66.7 (±11.8)	-29.2 (±15.3)	-2.7
Q26ACPC	67.0 (±1.5)	0.5 (±5.0)	-5.0
K30APC	96.3 (±7.9)	-25.6 (±11.4)	-5.9
G11(*R,R*)ACPC	70.2 (±4.6)	-25.6 (±8.1)	-

2.2.6 Replacement of Gly11 with (R,R)-ACPC

Glycine at the C-terminal end of an α-helical segment is a recurring helix-termination motif.[39,40] The glycine amide proton is the last hydrogen bond donor as the peptide backbone 'peels away' from the helix to form an irregular conformation that connects the helix to another secondary structural element. Altering the direction of the peptide backbone in this manner requires a positive φ dihedral angle, which is commensurate with a left-handed α-helix. Raleigh *et al.* have demonstrated that in some cases, pre-organization of these positive φ values with D-Ala substitutions could be stabilizing for tertiary structures that contain Gly helix-capping motifs.[40,41] Given these considerations, I reasoned that a cyclic β with the ability to form positive φ values could possibly constrain the C-capping motif and thus stabilize the tertiary structure. To test this hypothesis, VHP Gly11 was replaced with (*R,R*)-ACPC, the enantiomer of the cyclic β residue used to pre-organize right handed α-helices. However, VHP Gly11(*R,R*)-ACPC folded 'less cooperatively', *i.e.* has a smaller unfolding enthalpy, ΔH_m, than does native VHP. In addition, the substitution lowered the melting temperature relative to native VHP by ~15°C, which is similar to the destabilization observed for the N19ACPC substituted VHP variant.

In a follow-up to the original demonstration that a Gly-to-D-Ala substitution could stabilize a mini-protein, Raleigh *et al.* developed a computational model that can predict qualitative trends in ΔT_m, *i.e.* $\Delta\Delta G_m$, of Gly-to-D-Ala C-capping pre-organization in a variety of mini-protein tertiary structures.[40] This predictive model is predicated on changes in the van der Waals (VDW) interaction energy, $\Delta\Delta E_{VDW}$, between molecular dynamics trajectories that represent folded and unfolded states. Therefore, a destabilizing D-Ala

substitution has a negative net impact on $\Delta\Delta E_{VDW}$, as manifested through an increasing number of VDW interactions that are either favorable within the unfolded ensemble or unfavorable within the folded state. Raleigh *et al.* showed that $\Delta\Delta E_{VDW}$ values derived from MD trajectories correlated strongly with experimentally determined $\Delta\Delta G_m$ values arising from Gly11-to-D-Ala among a series of four mini-proteins. Coincidentally, Raleigh *et al.* employed native VHP as a negative control to validate their predictive computational model. Gly11-to-D-Ala substitution in VHP lowered T_m by ~15°C, whereas C-capping Gly-to-D-Ala substitutions in two other mini-proteins (engrailed homeodomain, peripheral subunit-binding protein) were found to be stabilizing by approximately 5°C.[40] Gly11-to-D-Ala substitution in the fourth protein of the series (GA-albumin binding domain) was determined to be stabilizing via chemical denaturation.

The destabilizing effect of the VHP Gly11(*R,R*)-ACPC substitution, which is similar in magnitude to that of the Gly-to-D-Ala substitution,[40] likely arises because introduction of a five-membered ring within the peptide backbone causes a concomitant net decrease in $\Delta\Delta E_{VDW}$ through either less favorable direct VDW interactions with the β residue relative to VDW interactions with the original α residue in the folded state, or a perturbation of nonlocal VDW interactions in the folded state relative to the native folded state due to backbone elongation and rigidification.

2.2.7 Quasiracemic Crystallization

We pursued quasiracemic crystallization in order to obtain structural insight on the impact of α→β replacements in VHP. Crystallization trials for each variant, which is composed mostly of L-α-amino acid residues, were conducted with a 1:1 mixture

containing that variant and the enantiomer of VHP itself. Three considerations motivated this strategy. First, a racemic mixture is believed to be more susceptible to crystallization than an enantiomerically pure solution,[42,43] and this advantage may extend to quasiracemic mixtures.[44,45] Second, all but one of the VHP variants are less stable than VHP itself, and diminished conformational stability seems likely to diminish crystallization propensity. We hypothesized that co-crystallization of the less robustly folded variants with the more stable D-VHP would mitigate this problem. Third, co-occurrence of a VHP variant and D-VHP in the same crystal lattice would provide the optimum basis for comparing the local conformation of the β residue with the corresponding α residue in VHP itself. Previous efforts in our group have employed quasiracemic crystallography to compare α residues with a non-proteinogenic side-chain in the context of folded VHP,[45] and Kent et al. have applied quasiracemic crystallography to other types of side-chain-modified polypeptides,[46–48] but the present work represents the first examination of backbone modifications via quasiracemic crystallography.

Optimum crystal growth and cryoprotection conditions were found for the quasiracemates containing the VHP analogues harboring the β^{3}hGln26, ACPC26, β^{3}hLys30 or APC30 modification within a chain of L-α-amino acid residues paired with D-VHP (*Experimental Methods*). In each case crystals diffracted weakly or not at all on a Cu-Kα rotating anode X-ray source and required synchrotron radiation for unequivocal structure determination. Each quasiracemate crystallized in space group *P*1 with unique cell constants (Table 2.5-Table **2.8**). In each of the four successful conditions quasiracemate crystals were grown from 5.8 mg/mL total peptide, compared to the 50-100 mg/mL reported for L-VHP alone.[4]

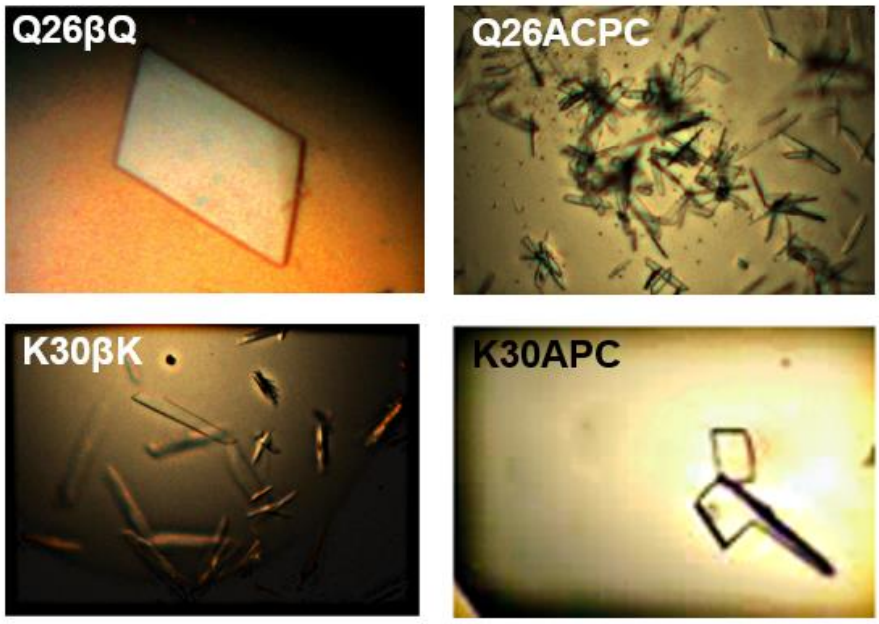


Figure 2.10. Quasiracemic crystals of β-substituted VHP variants were grown with hanging drop vapor diffusion from a variety of conditions (*Appendix*). After optimization of the crystallization conditions, quasiracemates Q26β^{3}hQ, Q26ACPC and K30β^{3}K yielded two dimensional plate-like crystals whereas quasiracemate K30APC yielded three dimensional crystals.

2.2.8 Crystal Structure Solution and Pseudo-Inversion Symmetry

Each structure was solved via molecular replacement with a polyalanine search model of D or L-VHP, refined in space group *P*1 and found to contain a distinct pseudo-inversion center that was detectable in the cumulative intensity distribution prior to structure solution (Chapter 4). The pseudo-inversion center combined with the primitive triclinic lattice indicates a pseudo-*P*$\bar{1}$ space group, which is consistent with the hypothesis of Wukovitz and Yeates that *P*$\bar{1}$ should be the most prevalent space group among truly racemic protein crystals.[42] Among the four quasiracemates, K30APC contains only four peptide chains in the unit cell and exhibits the closest approximation to true inversion symmetry as evidenced by the RMSD between pairs of chains related by pseudo-inversion symmetry (Table 2.3). Based on an approach originally developed for small

molecules,[49,50] I developed a framework to quantify the impact deviations of atomic coordinates from centrosymmetric positions would have on the distribution of normalized structure factor amplitudes, *i.e.*, to quantify the 'degree of centrosymmetry' prior to structure solution. The analysis indicated that among all the reported VHP quasiracemates, the K30APC quasiracemate distribution most closely mimicked that of a centrosymmetric structure (this pseudosymmetry analysis is discussed in detail in Chapter 4).

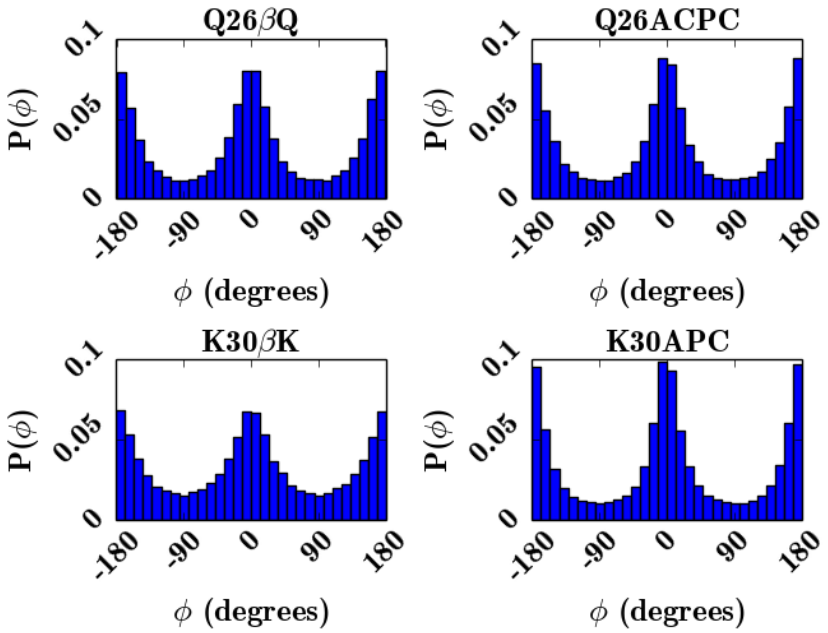


Figure 2.11. Phase angle distributions for each of the four quasiracemate models. Phase angles were binned (N=30) as normalized histograms at the resolutions reported in Tables 2.5-2.8.

Local and global pseudo-inversion centers can be present in structures with *P*$\bar{1}$ symmetry and a high *Z* number.[51] Therefore, to confirm that global pseudo-inversion centers were correctly identified, I translated putative inversion centers to the crystallographic origin and the resulting phase angles were plotted as normalized

histograms for each quasiracemate structure (Figure 2.11). Local pseudo-inversion centers were identified by determining the center-of-mass between pairs of L and D chains related by a possible inversion center. The observed bimodal distributions centered around 0° and 180° are consistent with the centrosymmetric phase angle constraint and pseudo-*P*$\bar{1}$ centering. In centrosymmetric crystals, the imaginary components of atomic scattering factors are cancelled out among pairs of atoms related by inversion symmetry, and the complex structure factor becomes real with phase 0 or 180°. The K30APC quasiracemate structure exhibits the sharpest phase angle distribution, which is consistent with the pseudo-inversion RMSD values and structure factor amplitude distribution pseudosymmetry metrics.

2.2.9 Crystal Packing in VHP Quasiracemates Q26β³hQ, Q26ACPC and K30β³K

Intriguingly, three quasiracemate structures (those with the L-VHP variant containing β^{3}hGln26, ACPC26 or β^{3}hLys30) exhibit a type of crystal packing similar to that observed in a previously reported *P*$\bar{1}$ VHP racemate (PDB: 3TRW). However, for each of these three quasiracemates, the unit cell contains twice the number of VHP chains (eight vs. four) and is approximately double the volume of the VHP racemate unit cell (Figure 2.12-Figure **2.15**). Comparable unit cell expansion has been observed for small molecule quasiracemates relative to the corresponding racemate and is characterized by translational non-crystallographic symmetry.[51]

The enlarged unit cells in the three new quasiracemates relative to the true VHP racemate could represent pronounced cases of commensurate modulation, a phenomenon in which adjacent unit cells are regularly displaced relative to one another

in a manner that disrupts the translational symmetry of the crystal lattice.[52] Modulated structures can be viewed as being defined by a supercell, *i.e*. a unit cell that contains all of the smaller unit cells that comprise a full phase of the modulation function. For the three quasiracemates that include L-VHP variants containing β^{3}hGln26, ACPC26 or β^{3}hLys30, deviations from true inversion symmetry arising from each backbone modification result in the combination of two clusters of four polypeptide molecules to form the supercell (Figure 2.12-Figure **2.15**). The possibility that the three crystals featured incommensurate, *i.e.* irrational repeats, rather than commensurate modulation was considered. However, the stable refinement of each quasiracemate structure supports the hypothesis that each lattice is based on a supercell corresponding to a commensurately modulated sub-structure that mimics the four-chain subunit packing in the previously reported VHP racemate. Further evidence that supports the hypothesis of commensurate over incommensurate modulation, is the lack of satellite reflections in the diffraction images of each quasiracemate structure (Appendix). Modulated macromolecular crystals have been only rarely described in the past.[53]

The four-chain subunit comprises a planar L and D peptide pair associated along a closely packed interface between a helix-spanning loop region that features specific interchain hydrogen bonding interactions between side-chain and backbone amides (Figure 2.21-Figure **2.22**). Surrounding this planar pair are two L and D-peptide chains that effectively sandwich the planar chain pair, the 'sandwiching' chains are coaxial with a vector that is perpendicular to the plane defined by the planar L/D-chain pair (Figure **2.12**-Figure **2.16**). Although the four-chain subunit is common among the β^{3}hGln26, ACPC26 and β^{3}hLys30 quasiracemate structures, the relative position of each four-chain

subunit differs among the structures. Within the β^{3}hGln26 and ACPC26 quasiracemate structures, which have very similar packing arrangements, the four-chain subunits form an intimate contact between the C-termini of L and D-peptide chains that are in the 'sandwiching' pair of opposing subunits (Figure 2.12-Figure **2.13**). This closely packed interface is centered on a pseudo-inversion center. For each structure, these centers are placed at special positions that correspond to *P*ī symmetry. Specifically, pseudo-inversion centers are at positions: (0,0,0); (1/2,0,0); (0,1/2,0); (0,0,1/2); (1/2,1/2,0); (1/2,0,1/2); (0,1/2,1/2); (1/2,1/2,1/2).

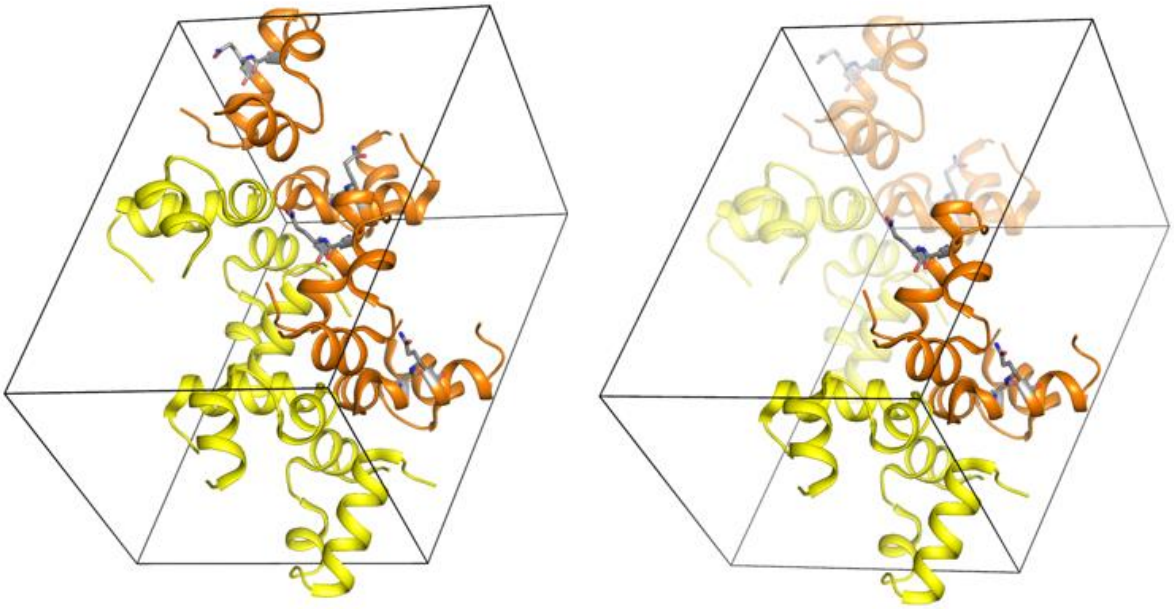


Figure 2.12. (left panel) The VHP Q26β^{3}Q unit cell with D-VHP (yellow) and VHP Q26β^{3}Q (orange). (right panel) The configuration of peptide chains that mimics the packing observed in the VHP racemate (PDB: 3TRW) is highlighted. The four-chain L/D subunit is positioned, relative to the other four-chain subunit, in a manner similar to the positioning in the Q26ACPC quasiracemate structure. The highlighted and transparent four-chain subunits are related by a pseudo-inversion center. There is not a true pseudo-inversion center between L/D chains within a four-chain subunit.

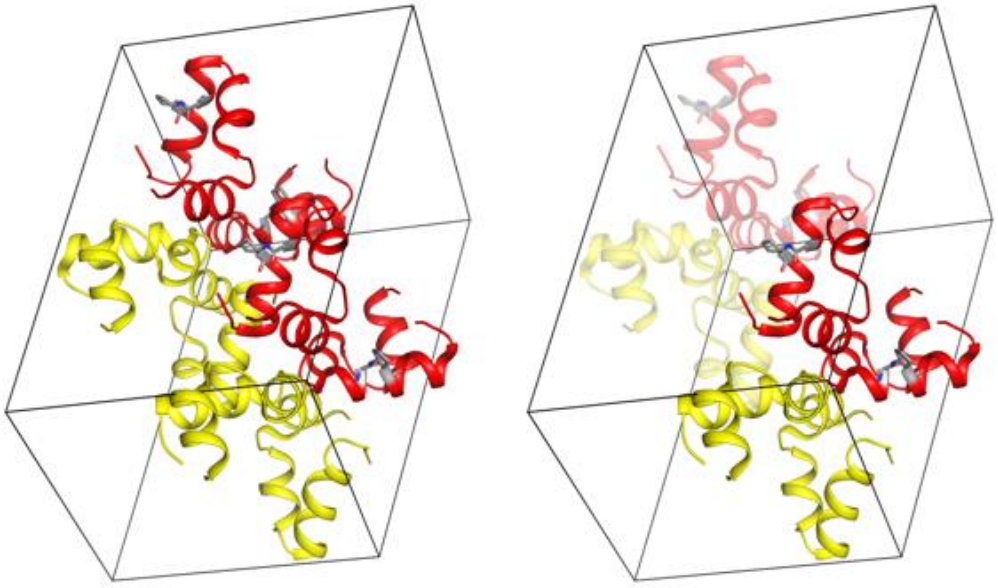


Figure 2.13. (left panel) The VHP Q26ACPC quasiracemate unit cell with D-VHP (yellow) and VHP Q26ACPC (red). (right panel) The configuration of peptide chains that mimics the packing observed in the VHP racemate (PDB: 3TRW) is highlighted. The four-chain L/D subunit is positioned, relative to the other four-chain subunit, in a manner similar to the positioning in the Q26β^{3}Q quasiracemate structure. The highlighted and transparent four-chain subunits are related by a pseudo-inversion center. There is not a true pseudo-inversion center between L/D chains within a four-chain subunit.

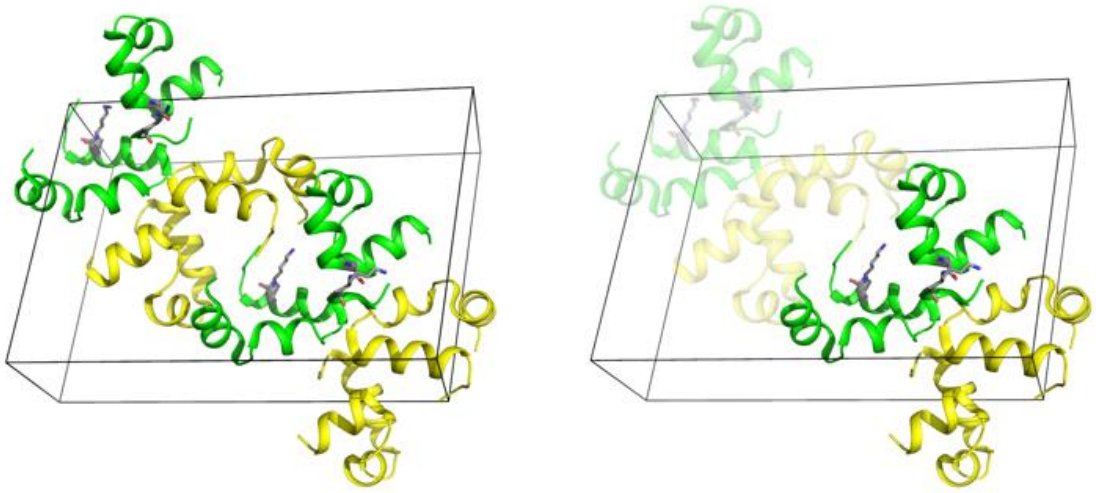


Figure 2.14. (left panel) The VHP K30β^{3}K quasiracemate unit cell with D-VHP (yellow) and VHP K30β^{3}K (orange). (right panel) The configuration of peptide chains that mimics the packing observed in the VHP racemate (PDB: 3TRW) is highlighted. The highlighted and transparent four-chain subunits are related by a pseudo-inversion center. While there is not a true pseudo-inversion center between L/D chains within a four-chain subunit. The relative positioning of the four-chain subunits is distinct from that of the positioning in the Q26β^{3}Q and Q26ACPC quasiracemates.

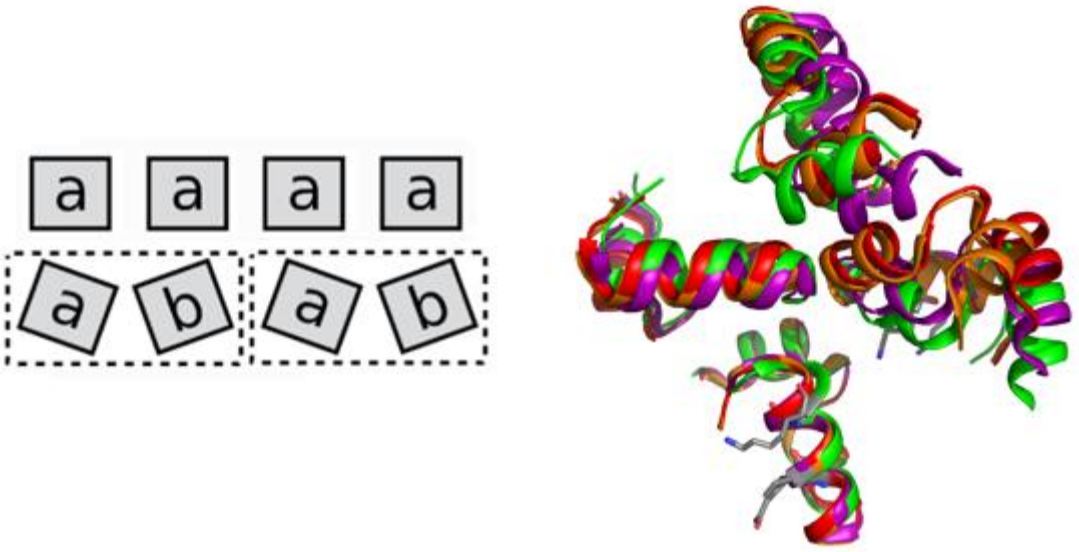


Figure 2.15. (left panel) A schematic representation of commensurate modulation where the modulation function extends over two subunit cells. Slight deviations from crystallographic inversion symmetry result in an expanded unit cell that encompasses both sub-unit cells (right panel) The four-chain subunit from the Pī racemate (PDB: 3TRW; purple) is overlaid with the analogous four-chain configuration from VHP quasiracemates Q26β^{3}hQ (PDB: 5i1N, orange), Q26ACPC (PDB: 5i1O, red) and K30β^{3}hK (PDB: 5i1N, green). β residues are shown in gray.

2.2.10 Crystal Packing in the VHP K30APC Quasiracemate

Crystal packing within the VHP K30APC quasiracemate is distinct from the packing configurations observed in the other three VHP β quasiracemate structures. The unit cell comprises four total peptide chains (two L and two D-chains), and planar pairs of L/D peptides are layered on top of each other with pseudo-translational symmetry (Figure 2.16). The packing arrangement is similar to that of a previously reported VHP quasiracemate that contains a pentafluorophenylalanine modification at a hydrophobic core position (PDB: 3TRV; Figure 2.16, *left panel*, cyan).[45] However, a major difference between the VHP K30APC and 3TRV structures is the number of chains in each unit cell. Akin to the unit cell expansion observed for the eight-chain VHP quasiracemates, the

K30APC quasiracemate contains four chains versus the two chains in the 3TRV structure. The fewer number of chains in the four-chain K30APC quasiracemate structure relative to the eight-chain quasiracemate structures results in a fewer number of unique interactions between chains and thus simplifies the identification of interactions that disrupt local inversion symmetry and expand the unit cell. In the two-chain 3TRV structure, the C-terminal phenylalanine residue achieves an extended conformation that orients the sidechain away from its respective chain and into the hydrophobic core of an adjacent chain.

Closely packed non-covalent interactions, such as the C-terminal Phe/hydrophobic core packing in 3TRV, are essential for pinning together a three dimensional lattice of macromolecules. In the K30APC quasiracemate, the planar, heterochiral dimer 'layers' feature one peptide chain that has the C-terminal Phe in an extended conformation while the other chain has the C-terminal Phe in a compact conformation (Figure 2.16, *right panel*). This bipartite combination among within planar pairs disrupts local inversion symmetry within the pair; however this feature is necessary to maintain three dimensional connectivity within the lattice (Figure 2.17). For instance, if members of pairs possessed the same conformation, *e.g.*, compact against compact, there would be a layer within the lattice that would be untethered to the adjacent layer and could thus slide indiscriminately with respect to the other layer. Presumably, this movement would result in a poorly ordered lattice that would either attenuate or preclude Bragg scattering.

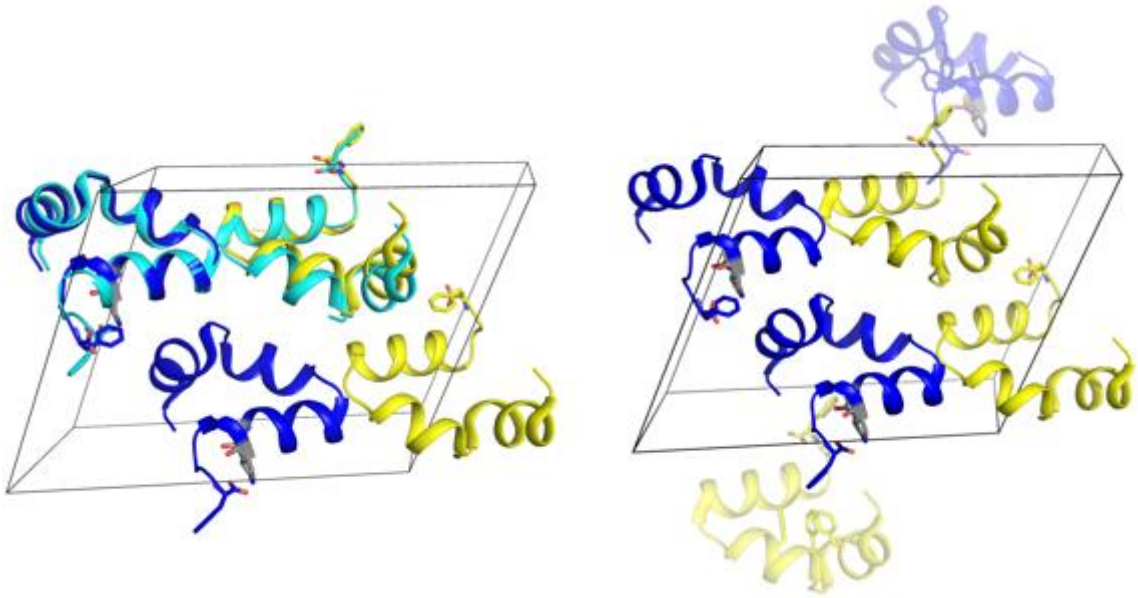


Figure 2.16. The VHP K30APC quasiracemate exhibits crystal packing that is distinct from the other three quasiracemate structures; however, the structure also exhibits modulation relative to a previously reported VHP structure (PDB: 3TRV, cyan; left panel). L/D-peptide pairs associate with near inversion symmetry, which is disrupted by the conformation of the C-terminus. (right panel) In one L and one D-chain the C-terminal phenylalanine side-chain is extended into the hydrophobic core of an adjacent peptide chain.

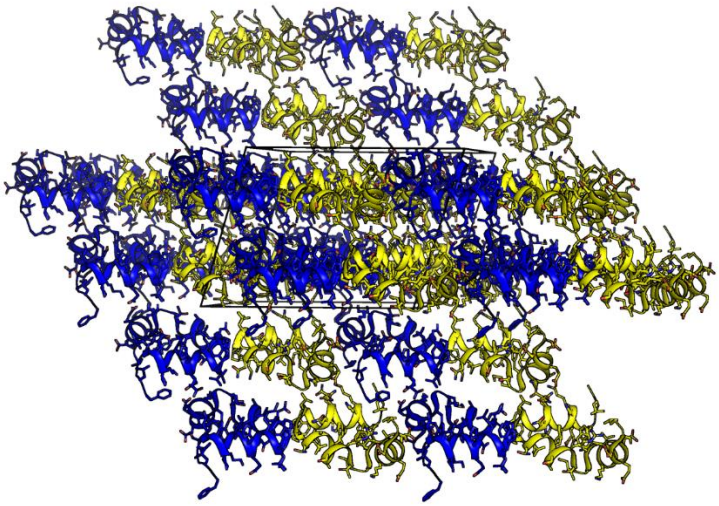


Figure 2.17. Layered packing in the triclinic K30APC quasiracemate lattice. Layers are pinned together by hydrophobic packing interactions between the side-chain of Phe35 and the phenylalanine core of chains in adjacent layers.

2.2.11 Packing Differences that Arise Due to the Nature and Position of β substitution

Slight perturbations from the local secondary structure of native VHP at the backbone modification site may have subtle but significant impacts on the global conformation of the polypeptide chain. Ultimately deviations from true crystallographic inversion symmetry within the VHP quasiracemates can be ascribed to each singular β substitution, relative to the D-α residue in D-VHP. Therefore, any deviation from inversion symmetry will represent a devation from the native conformation.

In the two-state folding model, the free energy of folding is dependent upon the energy difference between the folded and unfolded conformations. Therefore backbone modifications in the form of β-amino acid substitutions that decrease T_m with respect to the native system either destabilize the native folded conformation, stabilize the unfolded ensemble, or alter the relative energies of each state in a manner that decreases the free energy gap between states (folded vs. unfolded). Due to the nature of the denatured ensemble, quantitative metrics that describe the unfolded state are difficult to obtain much less employ for evaluating the impact of single atom modifications. However, significant conclusions can be drawn from high-resolution X-ray structures of the native state. For instance, there has been some success in combining large databases of thermodynamic measurements with high resolution structures to derive parametric equations that accurately predict thermodynamic properties of large proteins.[33,54] Given this success, there should be some correlation between the native structure and the thermodynamic properties of folding. Following this line of reasoning, within the VHP β-substituted quasiracemates, conformational disparity between the native and modified polypeptides can be ascribed to the particular β substitution and could explain, in part, the observed

ΔT_m for β substituted variants relative to native VHP where a structure of the β substituted variant is available.

The VHP K30APC quasiracemate, which has the smallest ΔT_m value, exhibits the highest level of pseudo-inversion symmetry based on the RMSD values between pseudo-inversion-related pairs of L/D-chains (Table 2.3) and the distribution of normalized structure factor amplitudes (Chapter 4). There are several factors that could explain why K30APC is stabilizing relative to the acyclic K30β^{3}hLys analogue, which include: preorganization of the backbone (entropic), removal of the flexible side-chain (entropic), and electrostatic interactions between the C-terminal carboxylate and the pyrrolidine side-chain that are possibly superior to those of the β^{3}hLys side-chain (enthalpic).

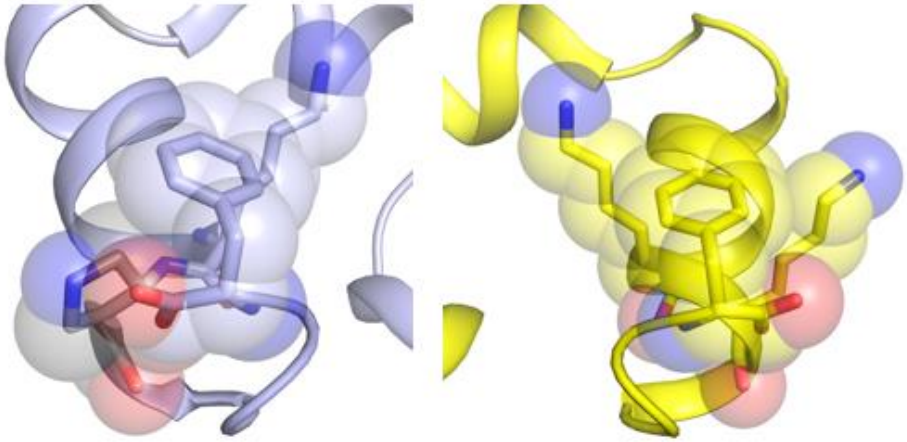


Figure 2.18. (*left panel*) The compact C-terminal Phe35 conformation features a Coulombic interaction between the C-terminal carboxylate and the APC pyrrolidine ring in addition to multiple van der Waals contacts between the Phe35 and Lys31 sidechains. (*right panel*) In contrast, the compact C-terminal Phe35 conformation in D-VHP features only VDW packing interactions between the Phe side-chain and the C-terminal helix.

The flexible sidechain hypothesis is not commensurate with the result that the acyclic β^{3}hGln → ACPC substitution had little impact on ΔT_m, because the entropic effect

of Gln and Lys side-chains are similar at -1.51 and -1.76 cal/mol*K, respectively.[38] Examination of the compact C-terminal Phe35 conformation for the K30APC quasiracemate reveals a favorable *i*→*i*+5 electrostatic interaction between the C-terminal carboxylate and the APC pyrrolidine ring within the L-peptide. No analogous interaction between Lys30 and Phe35 is present in the pseudo-inversion-symmetry related D-VHP C-terminus; the terminal carboxylate is oriented away from the helix in this case (Figure 2.18). The roles that backbone pre-organization and electrostatic interactions play in stabilizing the folded conformation are difficult to decouple; however, comparing backbone disorder within the C-termini for each quasiracemate structure explains, in part, some of the crystal packing differences observed within each structure, *i.e.* eight-chain vs. four-chain units (Q26β^{3}hQ, Q26ACPC, K30β^{3}hK) or four-chain vs. two-chain units (K30APC).

Among the VHP quasiracemates that exhibit the eight-chain packing configuration, Q26β^{3}hQ and Q26ACPC have a very similar arrangement of four-chain subunits. The K30β^{3}hK quasiracemate also features an eight-chain unit cell; however, the relative positioning of the four-chain subunits is distinct among the eight-chain configurations. Inspection of the K30β^{3}hK quasiracemate reveals an unraveling of the VHP C-terminus immediately following the β substitution position (residues 31-35). These residues display an irregular conformation that does not conform to any canonical secondary structural element. In the K30β^{3}hK quasiracemate this extension of C-terminal residues allows for a presumably favorable electrostatic contact between the C-terminal carboxylate and the Arg14 side-chain between two sets of the L/D planar pairs. This interaction, which is centered on a K30β^{3}hK quasiracemate pseudo-inversion center, is not present in the

Q26β^{3}hQ or Q26ACPC quasiracemate structures, which display more helical secondary structure around each respective pseudo-inversion center (Figure 2.19).

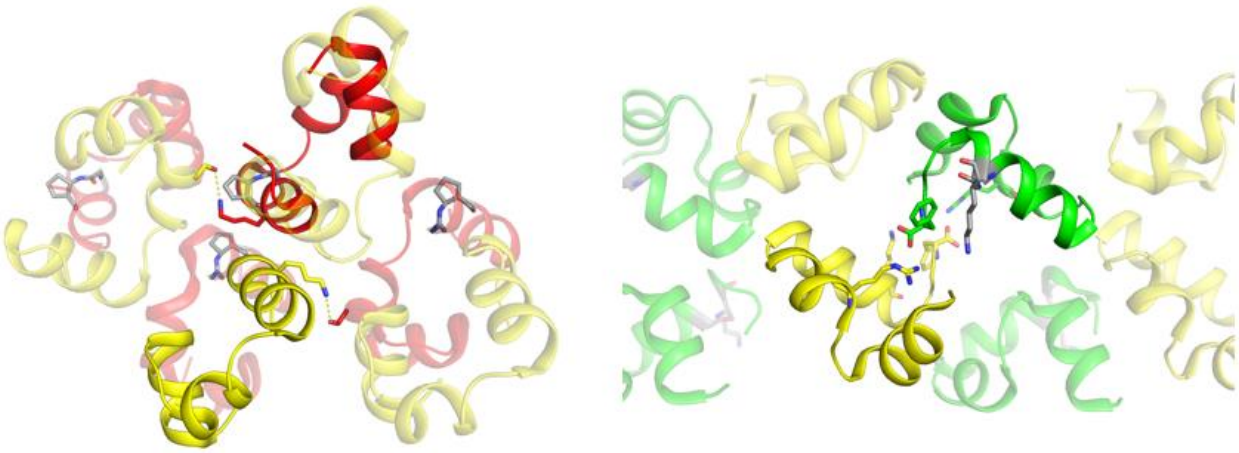


Figure 2.19. The crystallographic pseudo-inversion center is shown for VHP quasiracemates (left panel) Q26ACPC and (right panel) K30β^{3}hK. The Q26ACPC pseudo-inversion center involves L/D peptides from opposing 'sandwich' pairs whereas the K30β^{3}hK pseudo-inversion center involves L/D peptides from opposing planar pairs and features Coulombic interactions between the C-terminal Phe35 carboxylate and the Arg14 sidechain.

Table 2.3. Mean intra-crystal backbone atom RMSD (Å) values for pseudo-inversion related pairs of chains[a]

Variant	Pseudo-inversion	All possible L/D	D/3TRY
Q26β^{3}Q	0.10[b] (0.02)	0.28 (0.12)	0.34 (0.05)
Q26ACPC	0.16[b] (0.12)	0.40 (0.19)	0.41 (0.13)
K30β^{3}K	0.15[b] (0.04)	0.41 (0.22)	0.39 (0.11)
K30APC	0.070[b,c]	0.42 (0.41)	0.53[c]

[a]Mean RMSD values vary from 0.07 to 0.16 Å for pseudo-inversion related pairs of chains (first column) and from 0.28 to 0.42 Å for all possible L/D chain pairs within a crystal (second column). Values were calculated using backbone atoms (C, Cα, N, O); the 'extra' β residue C atom of each β residue was excluded. In addition to these intra-crystal comparisons (first two columns), we compared the D-VHP chains from each crystal with the D-chain in the previously reported VHP racemate structure (third column) (PDB: 3TRY). One standard deviation is reported in parentheses. [b]Values were less than the maximum-likelihood model error. [c]The two values were equivalent, thus no standard deviation is reported.

To quantify the degree of 'disorder' in the C-terminal residues (31-34), I plotted the isotropic atomic displacement parameters (ADPs) of backbone Cα atoms for both L and D-chains (related by pseudo-inversion symmetry) as a function of residue position (Figure 2.20). At Lys30, introduction of the cyclically constrained APC residue results in a lower set of ADP values compared to introduction of the acyclic β^{3}hLys. In contrast, at Gln26, introduction of the cyclically constrained ACPC residue has no discernible impact on the ADP values relative to introduction of the acyclic β^{3}hGln. These results are consistent with the hypothesis that pseudosymmetry is correlated with ΔT_m and may help explain the observation that the impact of backbone pre-organization is dependent on substitution position. The acyclic β^3 residue at the C-terminus results in an unraveling of helical secondary structure, whereas the cyclically constrained β residue results in a more ordered conformation that seems to influence crystal lattice packing.

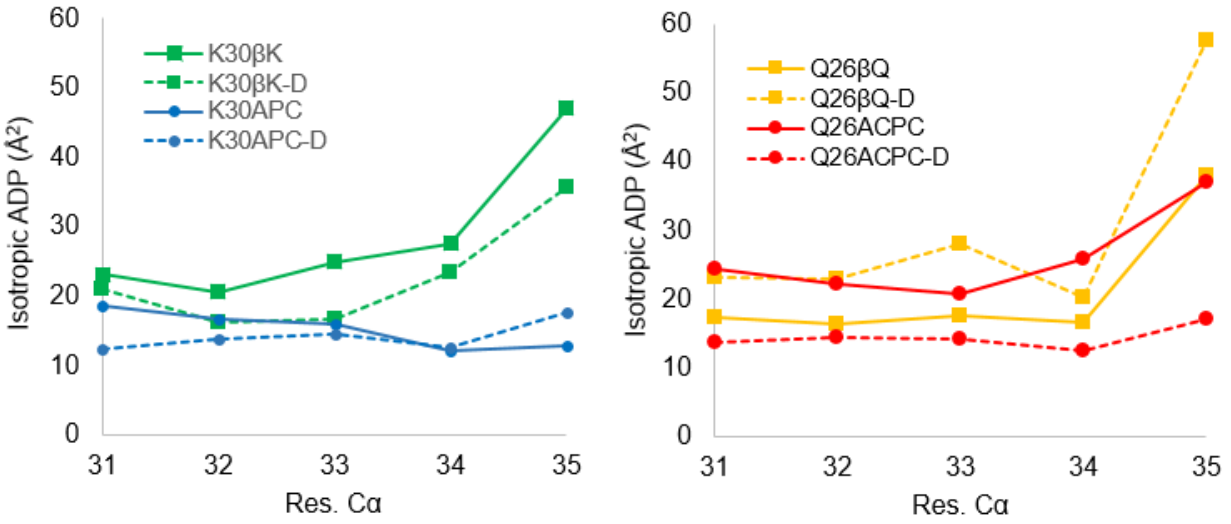
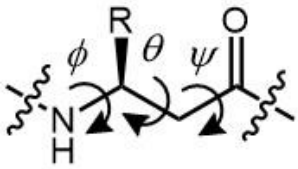


Figure 2.20. The isotropic atomic displacement parameter (ADP) is plotted for each Cα residue from residues 31-35. The cyclic versus acyclic β-substitution is compared at (left panel) position 30 and (right panel) position 26. At position 30 the cyclic β-substitution results in lower ADP values compared to the acyclic substitution, whereas at position 26 the identity of the β-residue (acyclic vs. cyclic) appears to have less bearing on the ADP values of the chain segment. In each case the corresponding D-VHP pseudo-inversion symmetry mate is also compared.

2.2.12 Helical H-bonding Patterns in VHP β Substituted Variants

The quasiracemates with L-VHP variants containing β^3hGln26, ACPC26, β^3hLys30 or APC30 exhibit mimicry of the canonical α-helix hydrogen bonding pattern at the site of substitution (Figure 2.21); however, the 'extra' backbone carbon atom present in each case results in multiple 14-atom hydrogen bond rings in place of the standard 13-atom hydrogen bond ring of the standard α-helix. The additional backbone bond in a β residue relative to an α residue results in a slight buckling of the backbone to accommodate the β residue without disruption of the helical H-bonds (Figure 2.22). The φ and ψ torsion angles of the β residues, which correspond with the conventional α residue angles, lie in the ranges -129.8° to -106.6° and -117.8° to -105.2° respectively (Table 2.4). An additional torsion angle θ, which arises from the C_α-C_β bond within each β residue, lies in the range 75.2° to 83.6°. These values are consistent with previously reported structures that contain multiple β residues in α-helix-like conformations.[22,23]

Table 2.4. Torsion angles for β residues in quasiracemic crystal structures.[a]



Variant	φ (°)	θ (°)	ψ (°)
Q26β3Q	-110.2, -106.7	75.2, 82.3	-117.8, -116.4
Q26ACPC	-111.7, -105.7	79.1, 83.6	-116.6, -112.8
K30β3K	-129.8, -110.6	75.9, 82.4	-107.2, -105.2
K30APC	-119.1, -106.6	77.4, 81.9	-112.0, -106.6

[a]Torsion angles were tabulated for each crystallographically independent β residue and are reported as [min, max] ranges. Conventional nomenclature was used for φ and ψ torsion angles; however these values differ significantly from that of a residues in a standard α-helix (φ = -57.8, ψ = -47.0) because of the extra C atom in each β residue. This extra backbone atom accounts also for the additional torsion angle, θ.

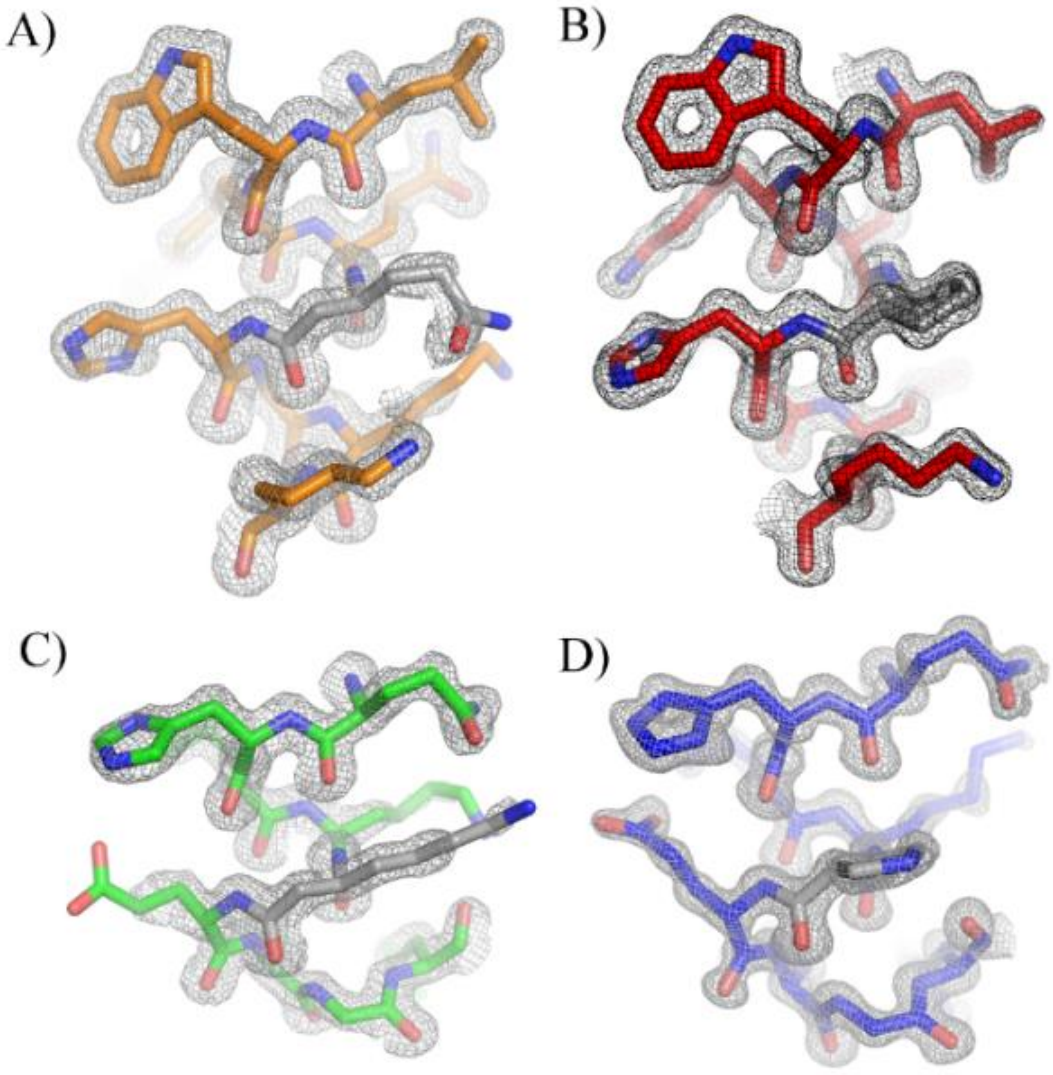


Figure 2.21. $2mF_o$-DF_c weighted electron density maps contoured at 1.5σ for A) Q26β^3Q (orange) and B) Q26ACPC (red) between residues Leu22 and Lys30; comparable images are shown for C) K30β^3K (green) and D) K30APC (blue) between Gln26 and Leu34. β residues are shown in gray.

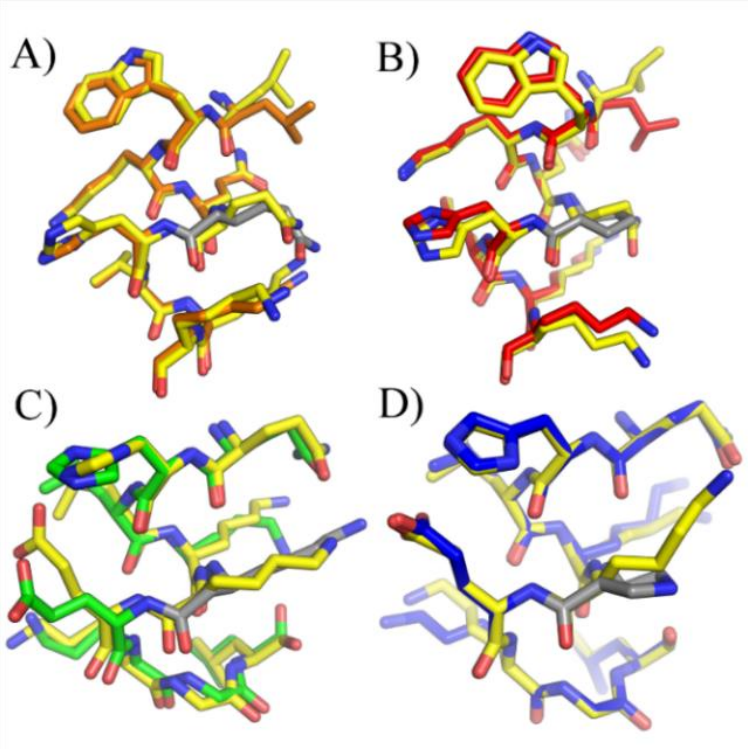


Figure 2.22. Overlay of inverted D-VHP (yellow) with the corresponding pseudo-inversion symmetry mate for A) Q26βQ (orange) and B) Q26ACPC (red) between residues Leu22 and Lys30; comparable overlays for C) K30βK (green) and D) K30APC (blue) between Gln26 and Leu34. β residues are shown in gray.

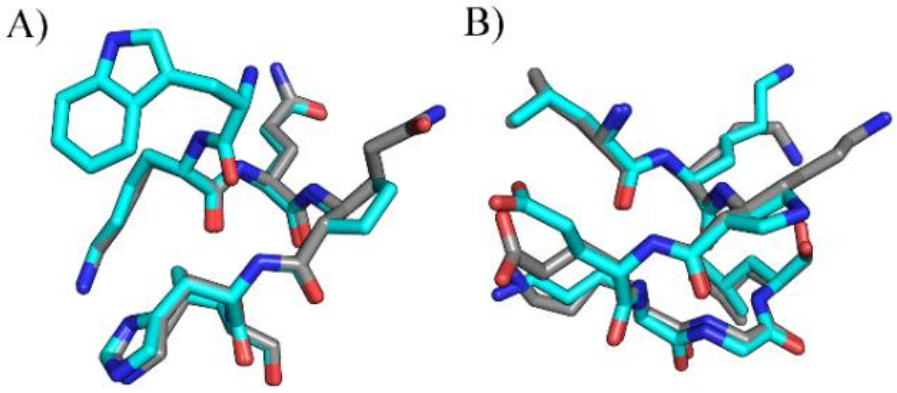


Figure 2.23. Cyclic β residues (cyan) compared with acyclic β3 residues (gray) by overlaying backbone atoms (C, Cα, N, O) for A) ACPC26 vs. β3h-Gln26 and B) APC30 vs. β3h-Lys30.

2.2.12 Heterochiral Packing Contacts

A heterochiral, hydrophobic packing interaction that involves the loop segment N19 to W23 is present in each of the four quasiracemate structures, regardless of the higher order lattice structure (Figure 2.24-Figure **2.25**). These heterochiral interactions, which are only possible from only racemic or quasiracemic mixtures, exemplify the packing arrangements that facilitate racemic and quasiracemic crystal growth compared to chiral crystals grown from enantiomerically pure solutions. Disruption of this interaction may explain my inability to optimize crystal growth conditions for quasiracemates with L-VHP variants containing α→β substitutions at Asn19 or Trp23. This hypothesis suggests that solution-phase native structure mimicry does not necessarily ensure quasiracemic crystal growth, because favorable packing contacts may be precluded by the point(s) of difference between the quasi-enantiomers. Lattice contacts that promote inversion or pseudo-inversion symmetry are crucial for racemic or quasiracemic crystal growth.

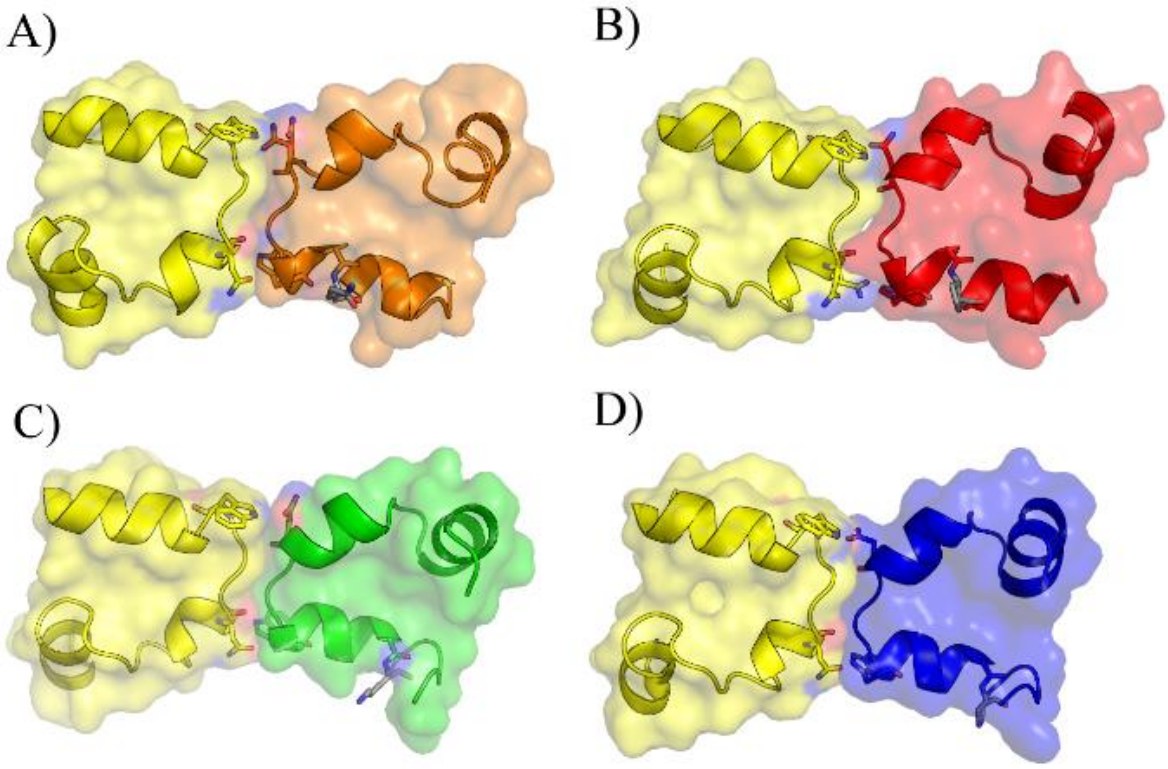


Figure 2.24. A heterochiral packing contact between residues Asn19 and Trp23 was observed in each quasiracemate structure reported here. A) Q26βQ (orange), B) Q26ACPC (red), C) K30βK (green), and D) K30APC (blue). D-VHP is shown in yellow. The side-chains of Asn and Trp residues are shown in each case.

Numerous heterochiral packing interactions, beyond the Asn19/Trp23 loop interaction, were observed that facilitated formation of the triclinic lattices. For instance, I observed a heterochiral interaction between L-His/Trp and D-Trp/His in the K30APC quasiracemate. Presumably the interaction is mediated by a cation-π interaction between L-His and D-Trp (or D-His and L-Trp) due to the acidic pH of the crystallization media (Figure 2.26). In addition Figure 2.27 illustrate the two distinct heterochiral packing modalities in the K30APC quasiracemate that disrupt inversion symmetry and result in modulation of the K30APC quasiracemate structure. Among the quasiracemic structures the heterochiral packing contacts range from specific hydrogen bonding interactions

between backbone and side-chain amides to glancing contacts of hydrophobic surface area akin to a ball and socket joint (Figure 2.25, Figure 2.28). As more racemic and quasiracemic protein structures become available it will be useful to catalog all of the observed L/D-interactions, to facilitate future efforts towards rational modifications that promote racemic or quasiracemic protein crystal growth.

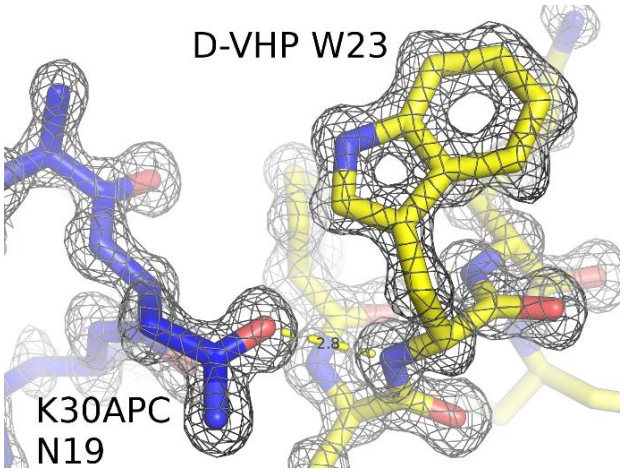


Figure 2.25. Hydrogen bonding at the heterochiral loop interface between asparagine 19 sidechain and tryptophan 23 backbone amide in the K30APC quasiracemate. Electron density is represented as a 2mFo-DFc weighted map contoured at 1.5σ.

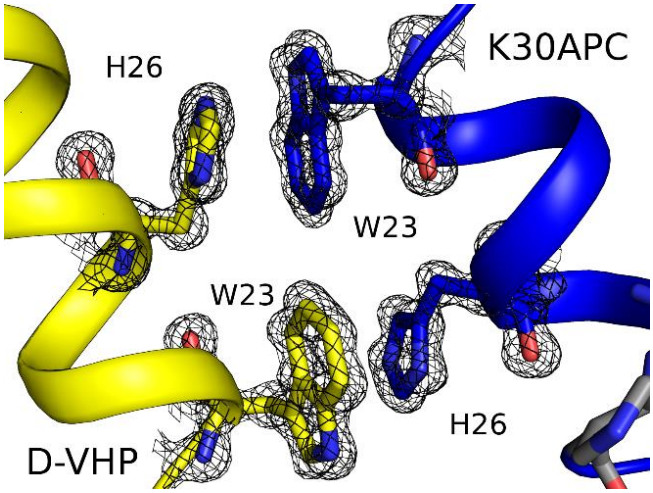


Figure 2.26. A heterochiral packing interaction between L/D-tryptophan 23 indoles and L/D-histidine 26 imidazoles. Electron density is represented as a 2mFo-DFc map contoured at 1.5σ.

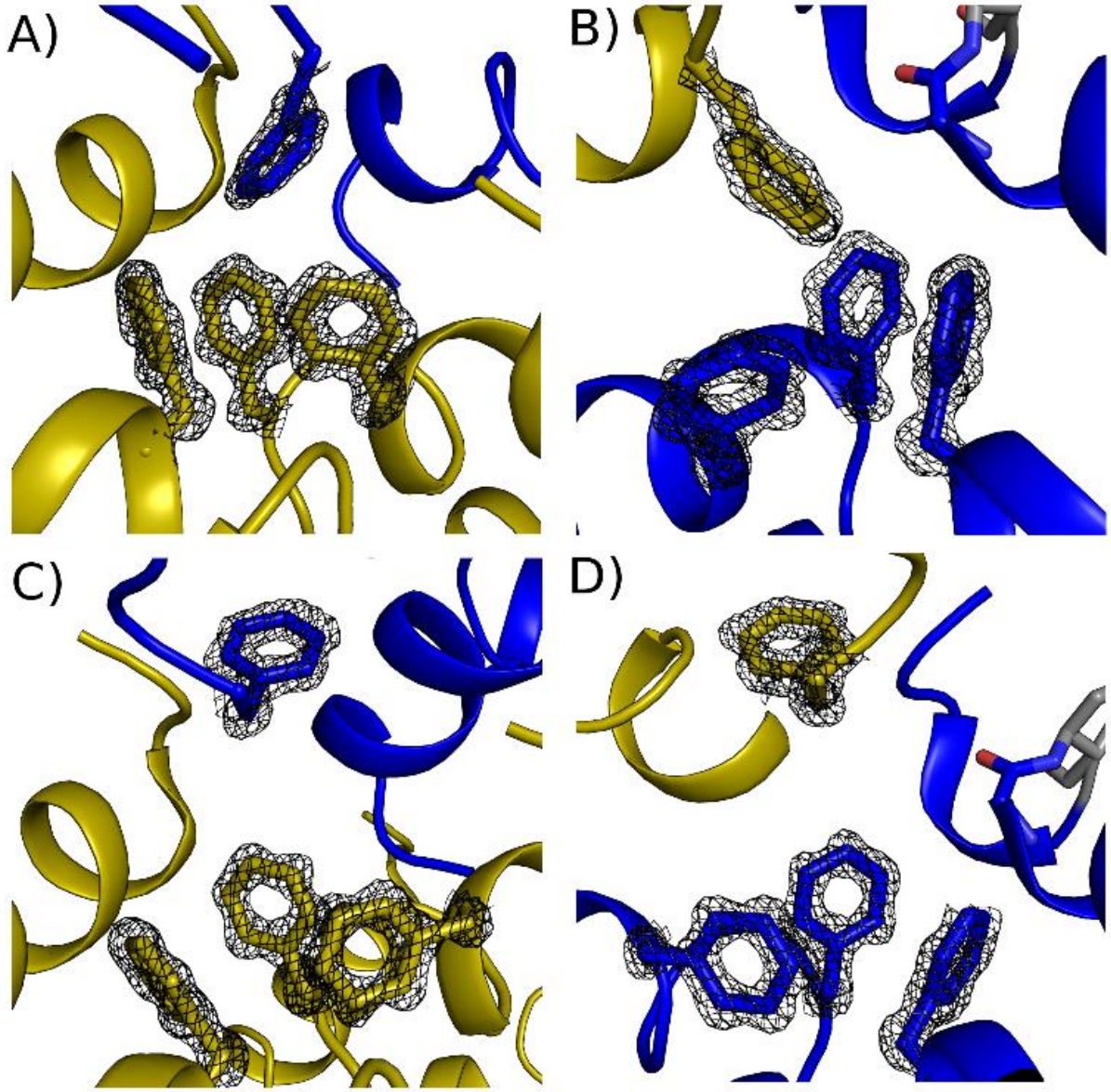


Figure 2.27. The two distinct VHP K30APC quasiracemate heterochiral packing interactions between A) the C-terminal Phe of chain A (L-VHP K30APC, blue) and the hydrophobic core of chain D (D-VHP, yellow), B) the C-terminal Phe of chain D (D-VHP) and the hydrophobic core of chain A (L-VHP K30APC, blue), C) the C-terminal Phe of chain B (L-VHP K30APC, blue) and the hydrophobic core of chain C (D-VHP, yellow), and D) the C-terminal Phe of chain C (D-VHP, yellow) and the hydrophobic core of chain B (L-VHP K30APC, blue). Electron density is represented as a 2mFo-DFc weighted map contoured at 1.5σ.

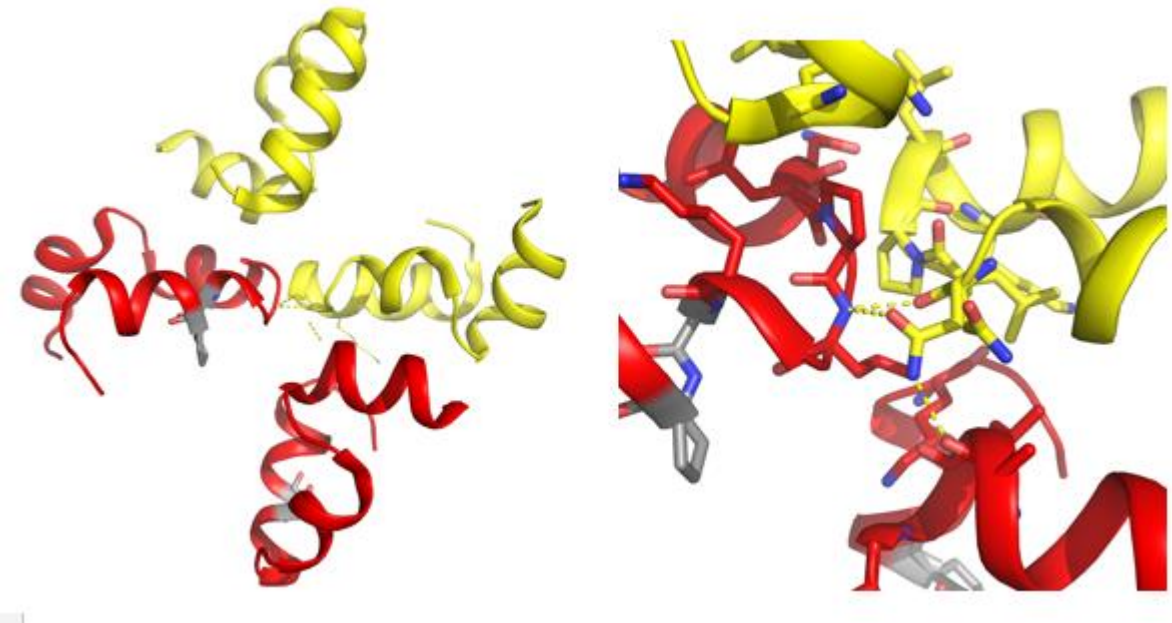


Figure 2.28. (left panel) The four-chain configuration from the VHP Q26ACPC quasiracemate features a planar L/D-peptide pair, associated by the Asn19/Trp23 loop motif, sandwiched between a separate L/D-peptide pair. (right panel) The sandwiching interfaces primarily comprise nonpolar contacts; however some sandwich interfaces feature H-bonding interactions between polar side-chains, such as that of Asn19, and peptide bonds that span two peptide chains together.

2.2.13 Influence of Quasiracemate Lattice on Tertiary Structures of Variants?

One might wonder whether the co-assembly of D-VHP and a β-containing variant of L-VHP into a pseudosymmetric crystal lattice could induce a deviation from the naturally preferred tertiary structure of one member of the quasiracemate pair. Since our four structures of D-VHP are very similar to one another and to the mirror image of the structure of L-VHP from a chiral crystal, we can dismiss the possibility of cocrystallization-induced distortion of D-VHP in any of the quasiracemates. Two lines of evidence argue against such distortion in the β-containing variants, although this possibility cannot be entirely excluded, particularly in terms of small-scale changes. First, each variant displays

a CD spectrum that is very similar to the CD spectrum of L-VHP (Figure 2.6); these data reflect the folding behavior of isolated polypeptides in solution. Second, the APC30 variant of L-VHP is comparable to L-VHP itself in terms of tertiary structure stability (Table 2.1), and it is not obvious why D-VHP would play a conformationally dominant role upon co-crystallization with the APC30 variant. Therefore, the similarity between the tertiary structures of L-VHP and the APC30 variant, and, by extension, the other β-containing variants, appears to reflect the intrinsic folding preferences of these variants.

2.3 Conclusions

These results suggest that single-site replacement of a native α-helical residue with its β^3 homologue is generally destabilizing to the tertiary folding pattern of VHP, although a native-like tertiary structure can be maintained despite the insertion of an 'extra' carbon atom into the polypeptide backbone. This type of structural mimicry has been well-established in previous studies involving multiple α→β^3 replacements in α-helical peptides.[55,56] The impact of α→β substitution on the stability of a tertiary structure however, has been evaluated directly in only one previous case, involving a discrete folding pattern different from the one we employed.[22] Prior studies with helical peptides have suggested that conformational stability can be recovered upon replacement of β^3 residues with cyclic β analogues such as ACPC or APC,[13,56–58] but in most such systems the link between cyclic β residues and conformational stabilization has been indirect. Our data reveal that the effect of β^3→cyclic β substitution varies as a function of position within the VHP sequence. At two sites, 19 and 30, this substitution increases stability, although only in the latter case is the stability of native VHP matched. At the other two sites, 23

and 26, β^3→cyclic β substitution has little impact on stability. In addition, cyclic β substitution at the site of a glycine with a positive φ value, using the enantiomer of (*R,R*)-ACPC destabilizes VHP. Divergent outcomes were observed for β^3→cyclic β substitution at positions 26 and 30. Inspection of the crystallographic data revealed an unraveling at position 30 upon acyclic β^3 substitution whereas the cyclic β substitution displayed a more ordered backbone as evidenced by lower isotropic backbone ADPs. Previous comparisons in the GB1 framework, all of which involved multiple sites of β^3→cyclic β substitution, are consistent with the trends we have observed for single substitutions in the VHP framework.[22,23] Collectively, these studies of α→β replacement in distinct tertiary contexts lay a foundation for the design of α/β tertiary structures intended to perform specific functions.[19,20,24,25,57] In the future it will be interesting to examine the impact of β residue substitutions in non-helical segments within discrete tertiary structures. Loops that connect α-helices and/or β-strands should be regions of particular interest because such loops are often involved in protein-protein interactions, and β substitutions at these positions may exert larger effects on non-local contacts within crystal lattices than is observed for substitutions within α-helices.

2.4 Appendix

2.4.1 VHP Peptide Synthesis and Purification

Peptides were prepared using microwave-assisted Fmoc solid-phase peptide synthesis. Fmoc-amino acids and coupling reagents were purchased from Chem-Impex International. Protected Fmoc-α-amino acids included: Asp(t-Bu ester), Glu(t-Bu ester), His(trityl), Lys(boc), Asn(trityl), Gln(trityl), Arg(Pbf), Ser(t-Bu ether), Thr(t-Bu ether), Trp(boc), and Arg(Pbf: 2,2,4,6,7-pentamethyldihydro-benzofuran-5-sulfonyl). Fmoc-β3-homoamino acids were purchased from PepTech (Bedford, MA). Fmoc-protected *trans*-(1*S*,2*S*)-2-aminocyclopentyl-1-carboxylic acid (ACPC) and *trans*-(1*R*,2*R*)-2-aminocyclopentyl-1-carboxylic acid (ACPC) was purchased from Chem-Impex (Wood Dale, IL). Fmoc-protected trans-N-Boc-(3*R*,4*S*)-4-aminopyrrolidine-3-carboxylic acid (APC) was synthesized as previously described.[59]

ChemMatrix Wang resin was preloaded with a coupling solution that comprised 5 equivalents of Fmoc-Phe-OH, 5 equivalents of 1-(mesitylene-2-sulfonyl)-3-nitro-1,2,4-triazole (MSNT), and 3.75 equivalents of methylimidazole in dichloromethane (DCM) at a final concentration of 100 mM Fmoc-Phe-OH. Resin was pre-swelled with DCM in a polypropylene fritted syringe, drained, and combined with coupling solution and allowed to nutate at room temperature for 5 h. The resin was then washed with 3 resin volumes of DCM. A fresh batch of coupling solution was added, and the resin was allowed to nutate at room temperature for 16 hours. Following C-terminal residue loading onto the resin and washing, subsequent coupling solutions comprised 4 equivalents of Fmoc-amino acid, 4 equivalents of 2-(6-chloro-1H-benzotriazole-1-yl)-1,1,3,3-tetramethylaminium hexafluorophosphate (HCTU), 4 equivalents of 6-chloro-1-hydroxy-

1H-benzotriazole (Cl-HOBt), and 8 equivalents of diisopropylethylamine (DIEA) in biotechnology-grade dimethylformamide (DMF) at a final concentration of 100 mM Fmoc-amino acid. Deprotection solutions were 20% (v/v) piperidine in biotechnolgy-grade DMF. Subsequent deprotection and coupling steps were carried out using microwave-assisted synthesis. Coupling reactions consisted of a 2 min ramp to 70°C followed by a 4 min hold at 70°C. Deprotection reactions consisted a 2 min ramp to 80°C followed by a 2 min hold at 80°C. The resin was washed with 3-5 resin volumes of biotechnology-grade DMF between each coupling and deprotection reaction.

Following the final deprotection step, resin was dried with DCM and mixed with a cleavage solution of 3.4 mL trifluoroacetic acid (TFA), 200 µL thioanisole, 200 µL water, 100 µL ethanedithiol, and 300 mg phenol per 50 µmol resin. The vessel was capped and allowed to nutate for 2-3 h at room temperature, after which the resulting crude solution was filtered through the fritted syringe, and the remaining resin was washed with 2-3 resin volumes of TFA that were combined with the crude solution. TFA was evaporated from the crude mixture under a stream of nitrogen to afford a cloudy brown oil. For a 50 umol-scale synthesis the crude oil was precipitated by addition of 50 mL cold diethyl ether and the precipitate was washed with 2x50 mL cold diethyl ether, centrifuged, and dried under nitrogen to afford a dry white powder. For a 50 µmol scale synthesis, the crude powder was redissolved in 5 mL of 1:1 water/acetonitrile, 0.1% (v/v) TFA and filtered through a 0.2 µm PTFE syringe filter for subsequent reverse-phase HPLC purification. Crude peptide mixtures were purified with a C18-functionalized semi-preparative HPLC column (250x21.2 mm) and a water/acetonitrile (0.1 % v/v TFA) gradient. The identity of each polypeptide was evaluated with MALDI-TOF mass spectrometry and purity (>90%)

was established via peak area integration in reverse phase HPLC or UPLC chromatograms (). MALDI-TOF spectra were collected on a Bruker Reflex II MALDI-TOF mass spectrometer in cases where an average [M+H] is reported in lieu of a monoisotopic [M+H]. MALDI-TOF spectra collected on a Bruker Ultraflex MALDI-TOF mass spectrometer are reported as the monoisotopic [M+H] due to the higher resolution afforded by the Bruker Ultraflex.

2.4.2 VHP Peptide Characterization

A solvent: H2O, 0.1% (v/v) TFA

B solvent: Acetonitrile, 0.1% (v/v) TFA

L-Villin Headpiece (L-VHP)
Sequence: H-LSDEDFKAVFGMTRSAFANLPLWKQQHLKKEKGLF-OH
Calculated monoisotopic [M+H]: 4080.0

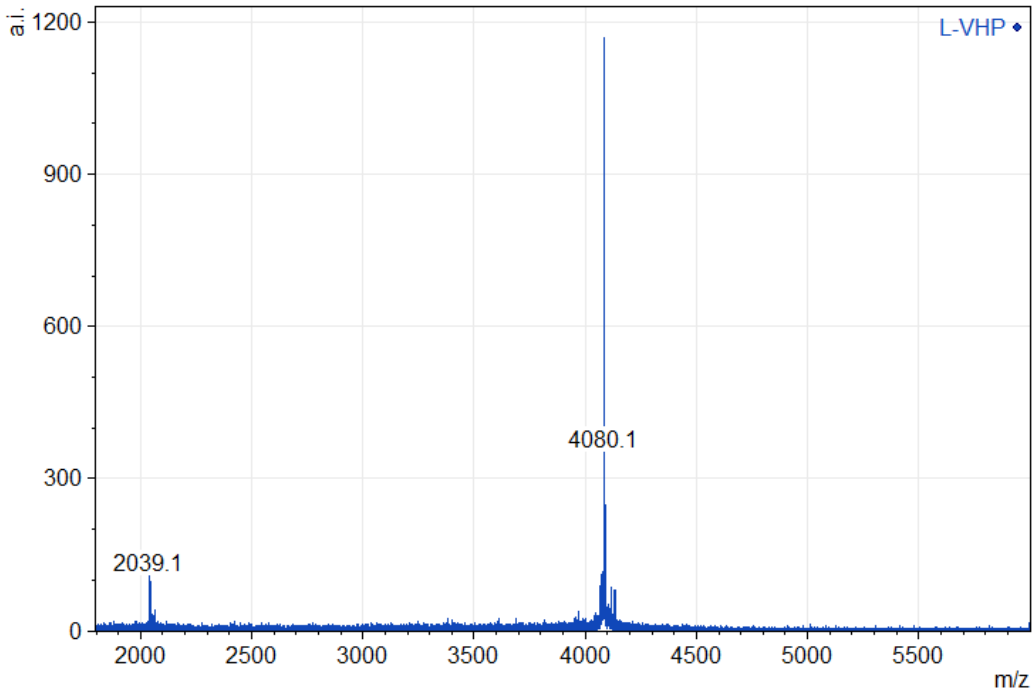


Figure 2.29. L-VHP MALDI-TOF monoisotopic [M+H] peak is labeled at 4080.1 and the largest [M+2H] peak is labeled at 2039.1.

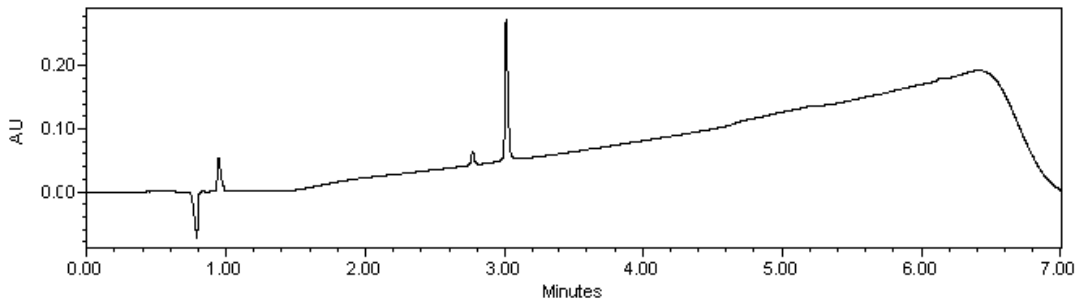


Figure 2.30. L-VHP purity analysis, UPLC gradient from 20-70% B solvent over 5 minutes (0.3 mL/min; column-Waters Acquity CSH C18 1.7 µm, 2.1 x 100 mm; purity, >90%).

D-Villin Headpiece (D-VHP)
Sequence: H-LSDEDFKAVFGMTRSAFANLPLWKQQHLKKEKGLF-OH
Calculated avg. [M+H]: 4082.7

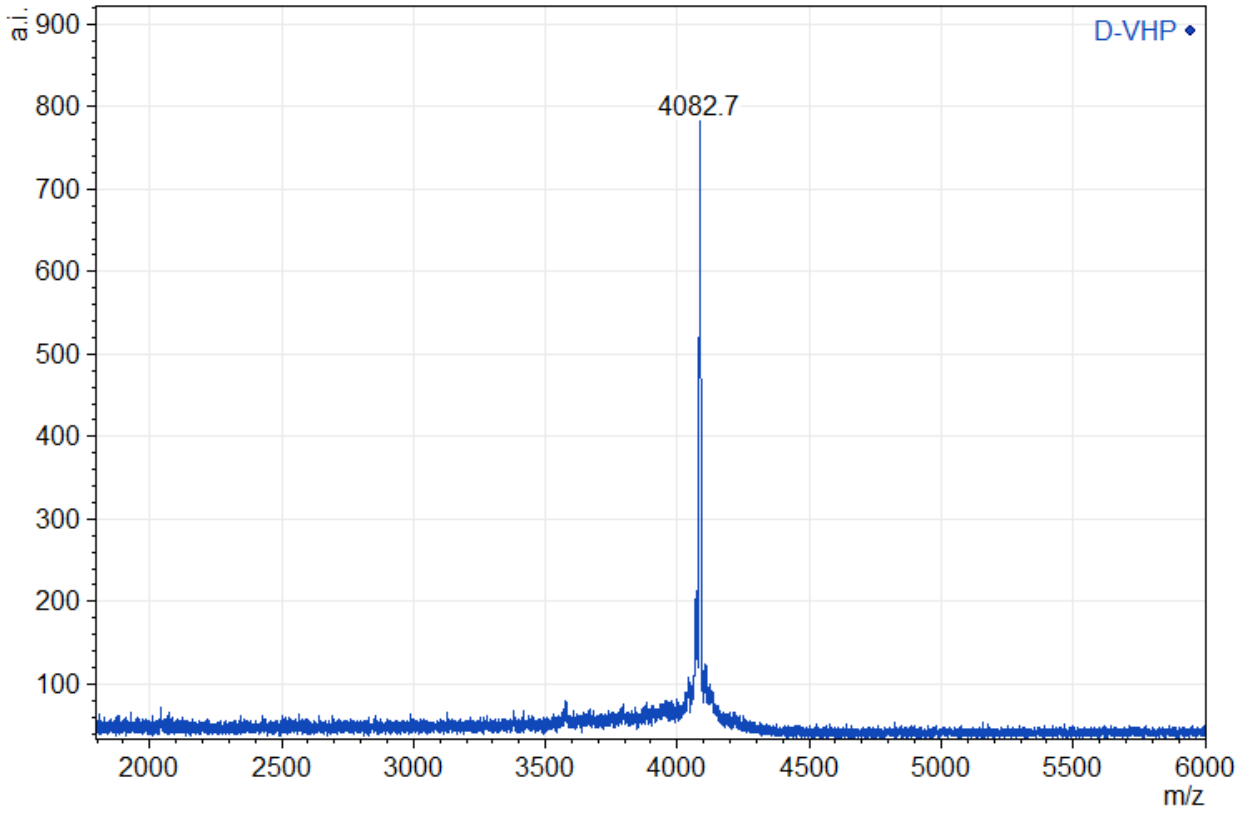


Figure 2.31. D-VHP MALDI-TOF average [M+H] peak is labeled at 4082.7.

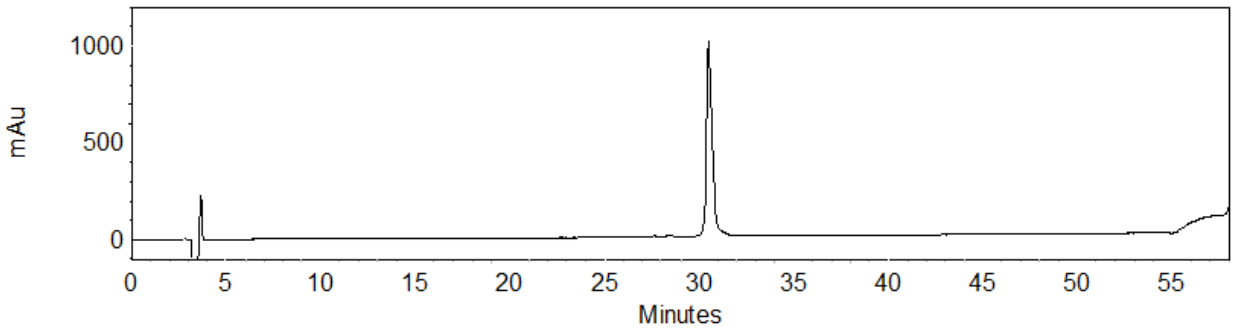


Figure 2.32. D-VHP HPLC purity analysis with a 10-60% B HPLC gradient over 50 minutes (1.0 mL/min; column-Waters XBridge Peptide BEHC18 130 Å, 5 µm, 4.6 x 250 mm; purity >95%).

N19β3N (β3N: β-3-homoasparigine)
Sequence: H-LSDEDFKAVFGMTRSAFA**β3N**LPLWKQQHLKKEKGLF-OH
Calculated monoisotopic [M+H]: 4094.2

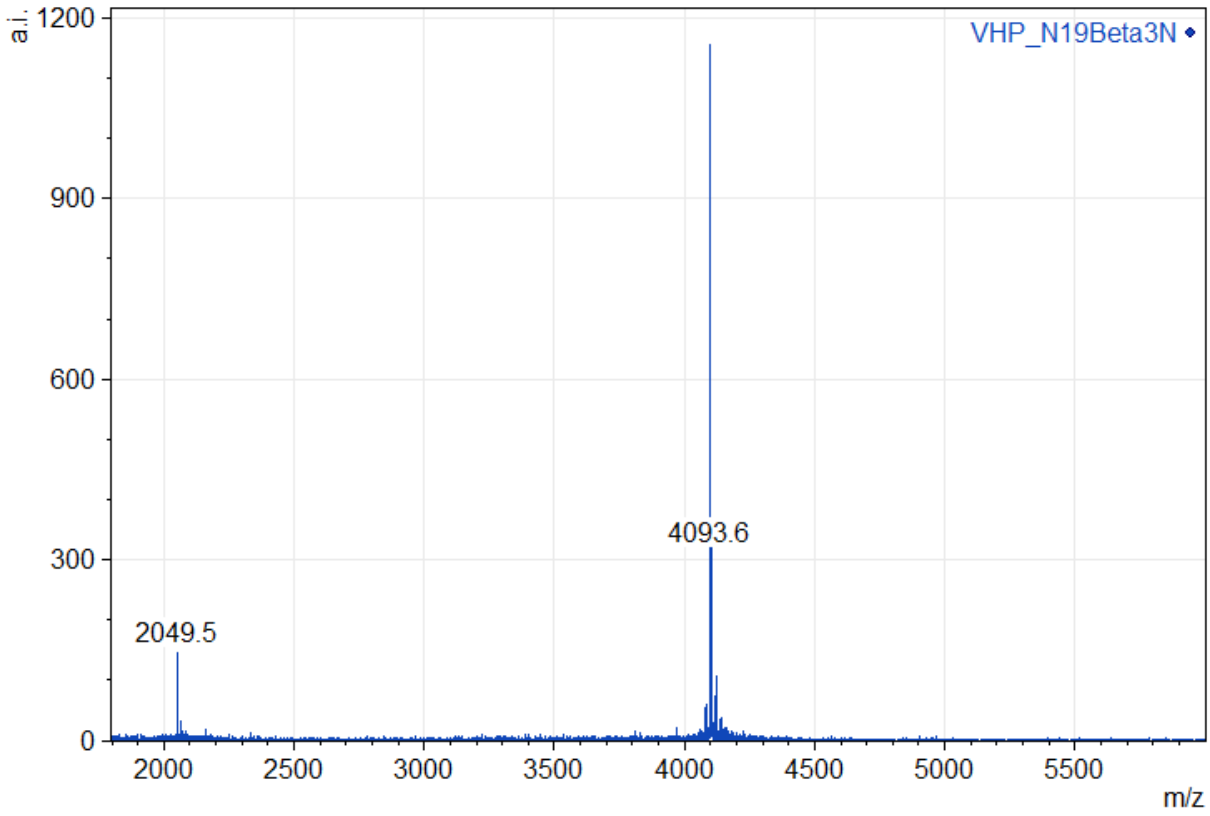


Figure 2.33. VHP N19β3N MALDI-TOF monoisotopic [M+H] peak is labeled at 4093.6 and the largest [M+2H] peak is labeled at 2049.5.

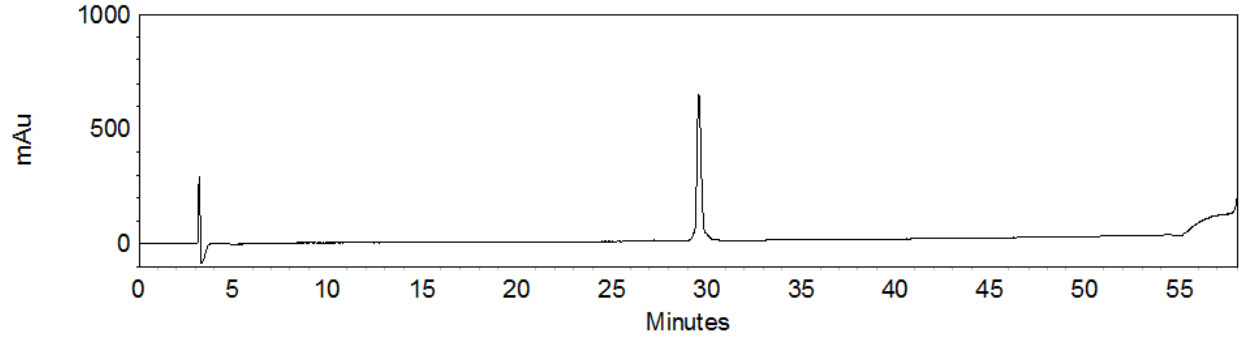


Figure 2.34. VHP N19β3N HPLC purity analysis with a 10-60% B HPLC gradient over 50 minutes (1.0 mL/min; column-Waters XBridge Peptide BEHC18 130 Å, 5 µm, 4.6 x 250 mm; purity >95%).

VHP W23β3W (β3W: β-3-homotryptophan)
Sequence: H-LSDEDFKAVFGMTRSAFANLPL**β3W**KQQHLKKEKGLF-OH
Calculated monoisotopic [M+H]: 4094.2

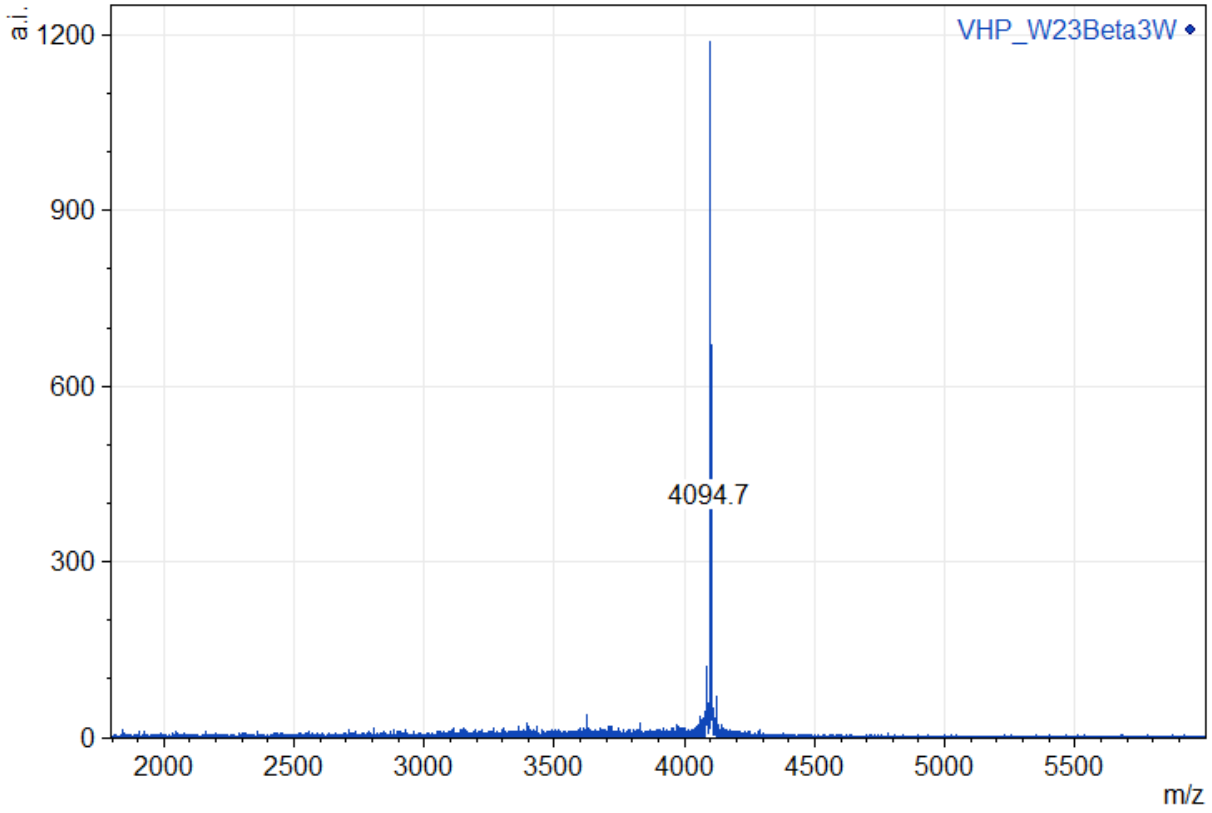


Figure 2.35. VHP W23β3W MALDI-TOF monoisotopic [M+H] peak is labeled at 4094.7.

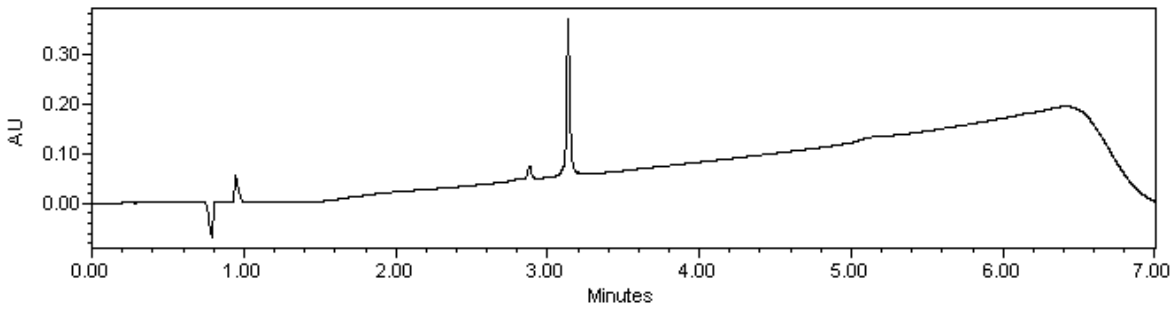


Figure 2.36. VHP W23β3W UPLC purity analysis, UPLC gradient from 20-70% B solvent over 5 minutes (0.3 mL/min; column-Waters Acquity CSH C18 1.7 µm, 2.1 x 100 mm; purity >90%).

VHP Q26β3Q (β3Q: β-3-homoglutamine)
Sequence: H-LSDEDFKAVFGMTRSAFANLPLWKQ**β3Q**HLKKEKGLF-OH
Calculated avg. [M+H]: 4096.8

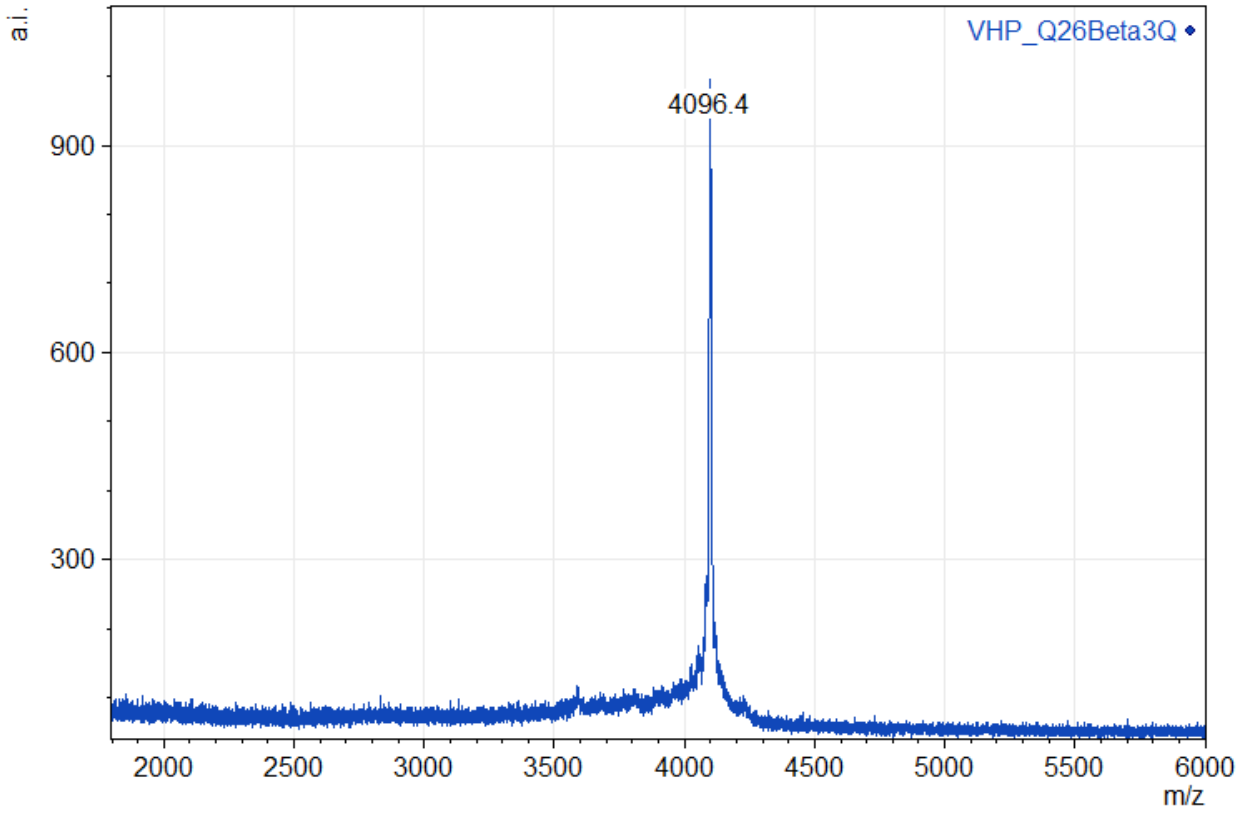


Figure 2.37. VHP Q26β3Q MALDI-TOF average [M+H] peak is labeled at 4096.4.

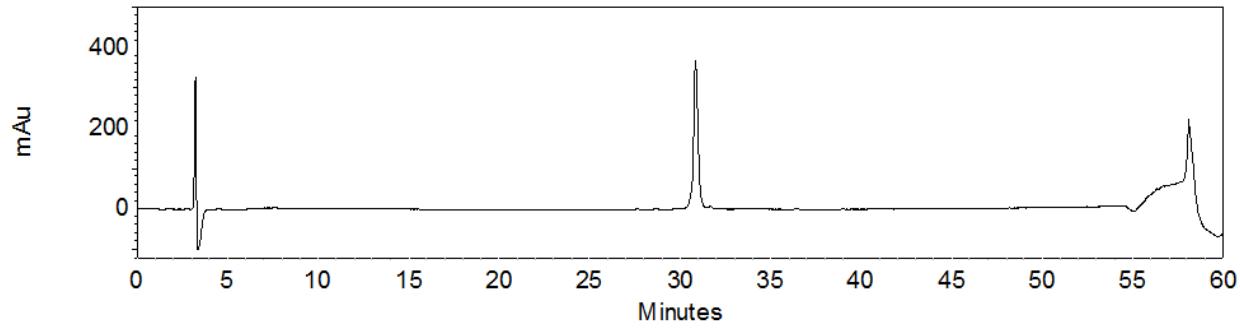


Figure 2.38. VHP Q26β3Q HPLC purity analysis with a 10-60% B HPLC gradient over 50 minutes (1.0 mL/min; column-Waters XBridge Peptide BEHC18 130 Å, 5 µm, 4.6 x 250 mm; purity >95%).

K30β3K (β3K: β-3-homolysine)
Sequence: H-LSDEDFKAVFGMTRSAFANLPLWKQQHLK**β3K**EKGLF-OH
Calculated monoisotopic [M+H]: 4094.2

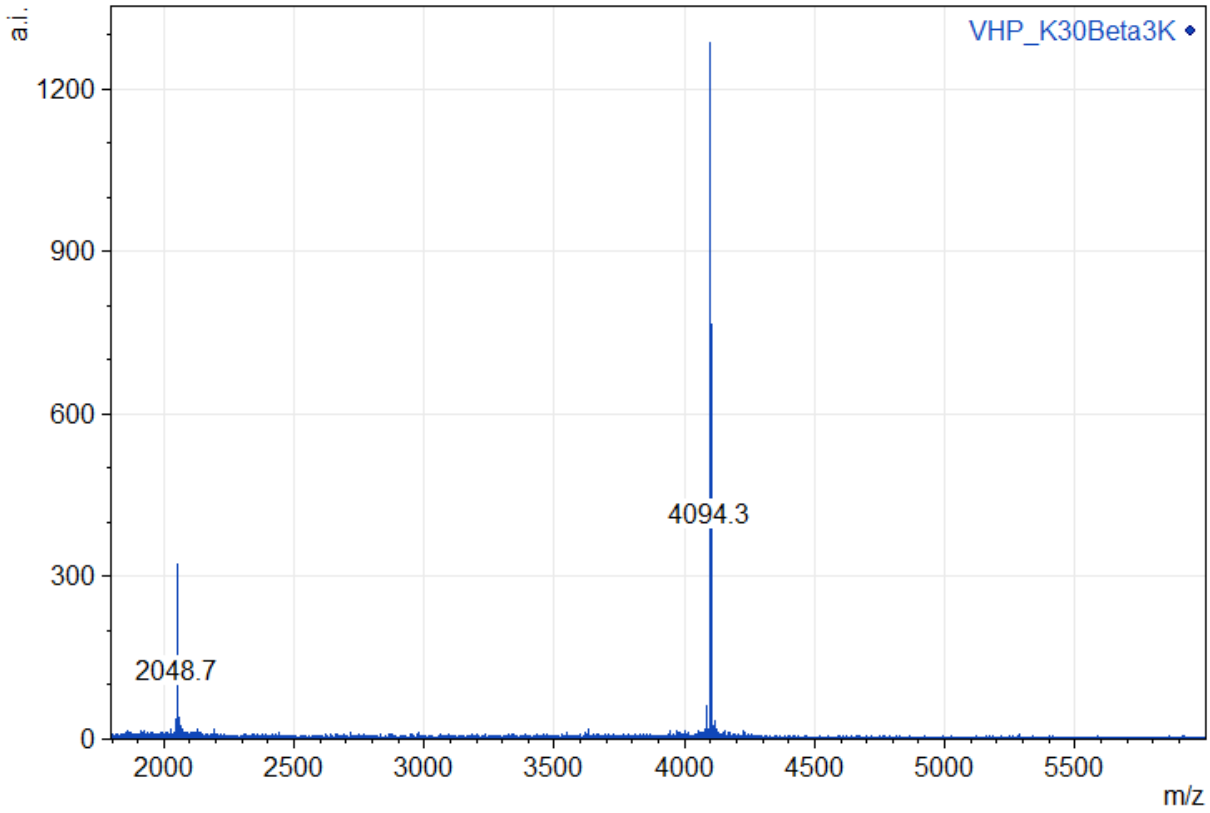


Figure 2.39. VHP K30β[3]K MALDI-TOF monoisotopic [M+H] peak is labeled at 4094.3 and the largest [M+2H] peak is labeled at 2048.7.

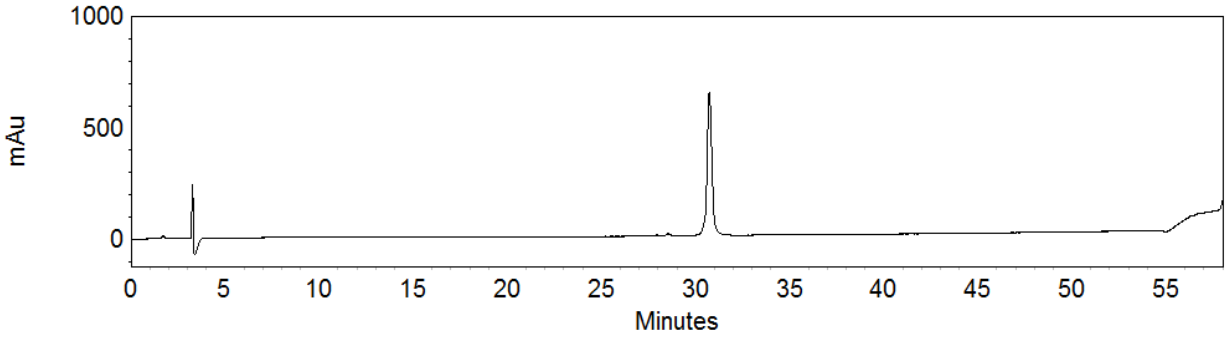


Figure 2.40. VHP K30β[3]K HPLC purity analysis with a 10-60% B HPLC gradient over 50 minutes (1.0 mL/min; column-Waters XBridge Peptide BEHC18 130 Å, 5 µm, 4.6 x 250 mm; purity >95%).

VHP N19ACPC (ACPC: (1S,2S)-2-aminocyclopentane carboxylic acid)
Sequence: H-LSDEDFKAVFGMTRSAFA(**ACPC**)LPLWKQQHLKKEKGLF-OH
Calculated monoisotopic [M+H]: 4077.2

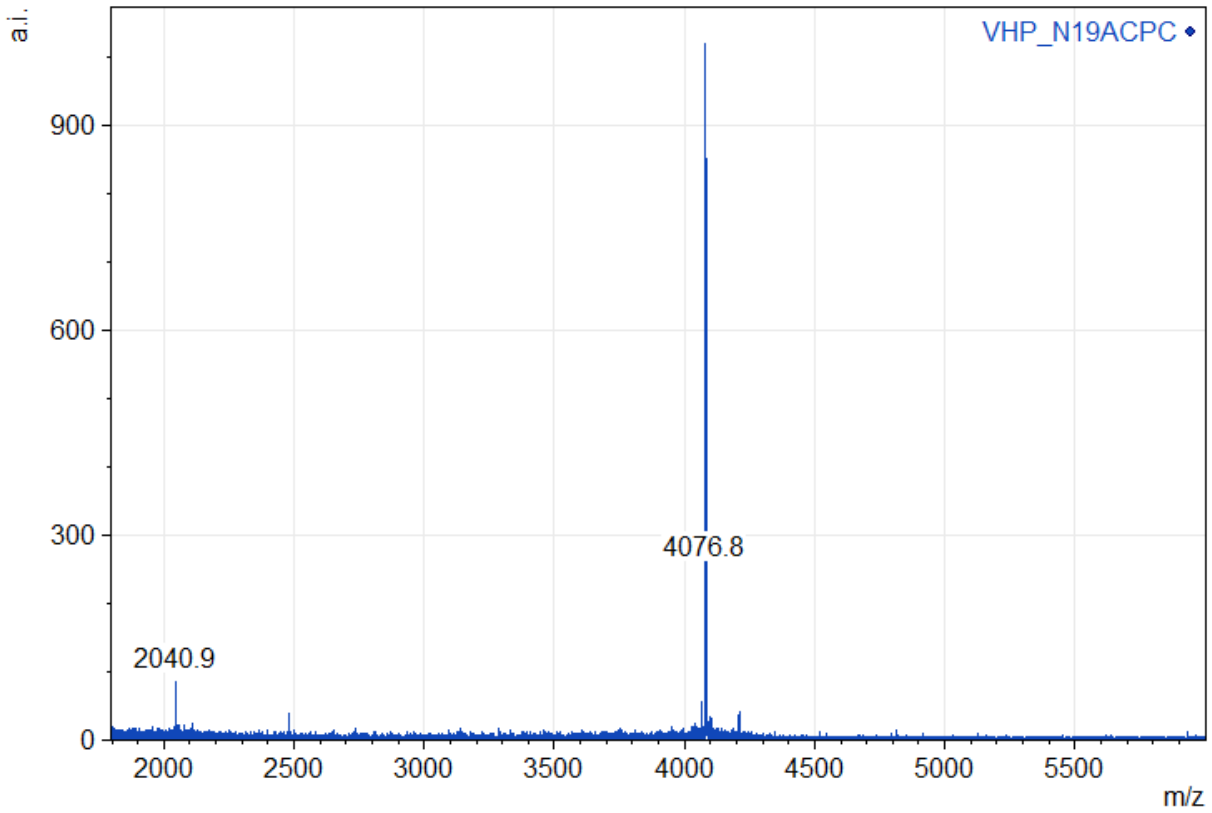


Figure 2.41. VHP N19ACPC MALDI-TOF monoisotopic [M+H] peak is labeled at 4076.8 and the largest [M+2H] peak is labeled at 2040.9.

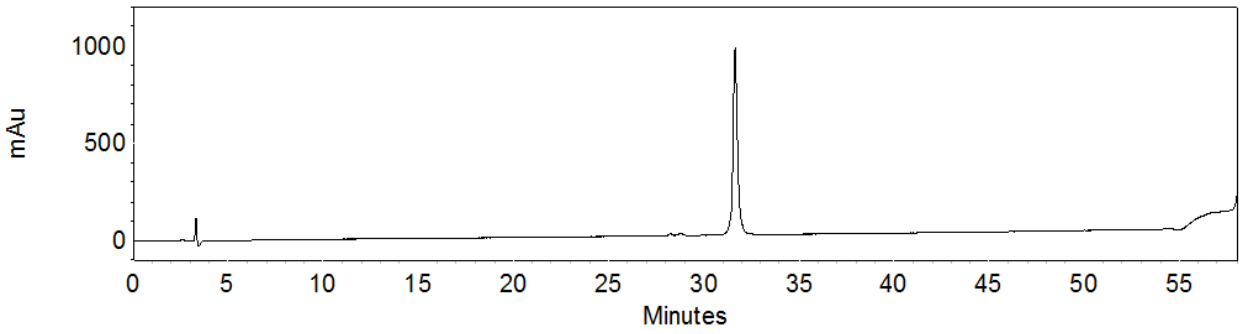


Figure 2.42. VHP N19ACPC HPLC purity analysis with a 10-60% B HPLC gradient over 50 minutes (1.0 mL/min; column-Waters XBridge Peptide BEHC18 130 Å, 5 µm, 4.6 x 250 mm; purity >95%).

VHP W23ACPC (ACPC: (1S,2S)-2-aminocyclopentane carboxylic acid)
Sequence: H-LSDEDFKAVFGMTRSAFANLPL(**ACPC**)KQQHLKKEKGLF-OH
Calculated monoisotopic [M+H]: 4005.1

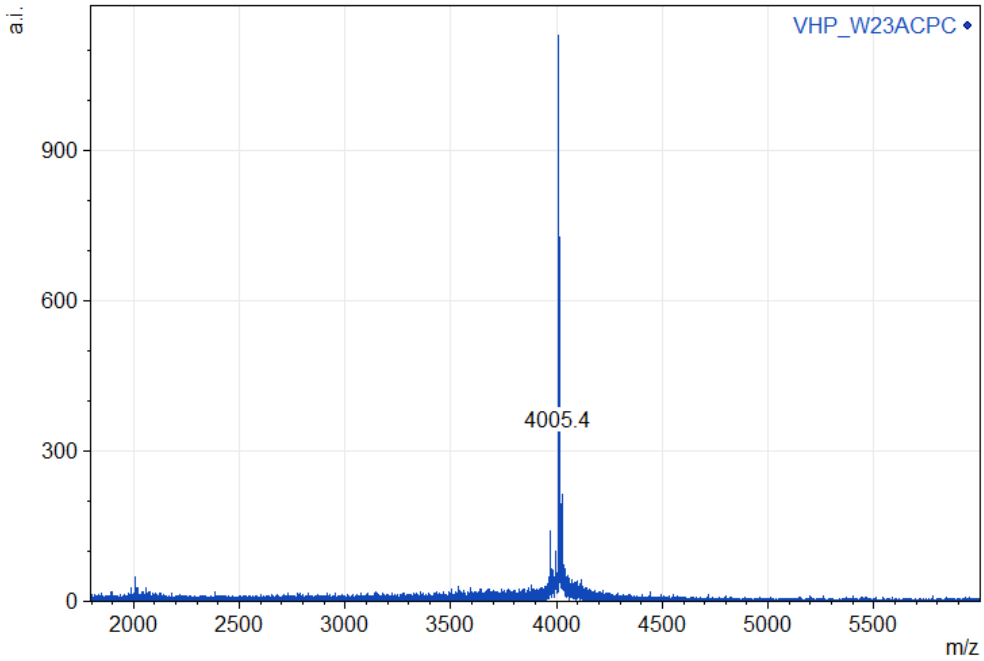


Figure 2.43. VHP W23ACPC MALDI-TOF monoisotopic [M+H] peak is labeled at 4005.4.

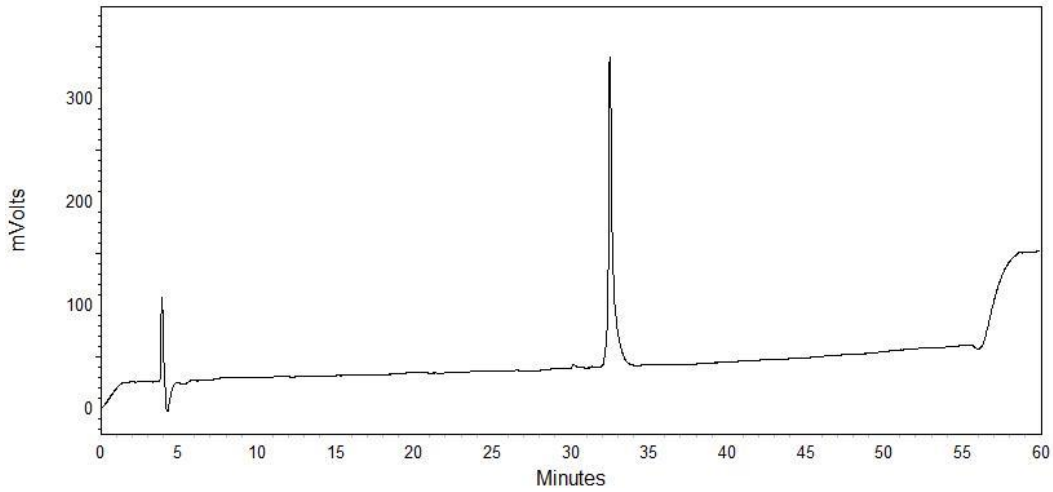


Figure 2.44. VHP W23ACPC HPLC purity analysis with a 10-60% B HPLC gradient over 50 minutes (1.0 mL/min; column-Waters XBridge Peptide BEHC18 130 Å, 5 µm, 4.6 x 250 mm; purity >95%).

VHP Q26ACPC (ACPC: (1*S*,2*S*)-2-aminocyclopentane carboxylic acid)
Sequence: H-LSDEDFKAVFGMTRSAFANLPLWKQ(**ACPC**)HLKKEKGLF-OH
Calculated monoisotopic [M+H]: 4063.2

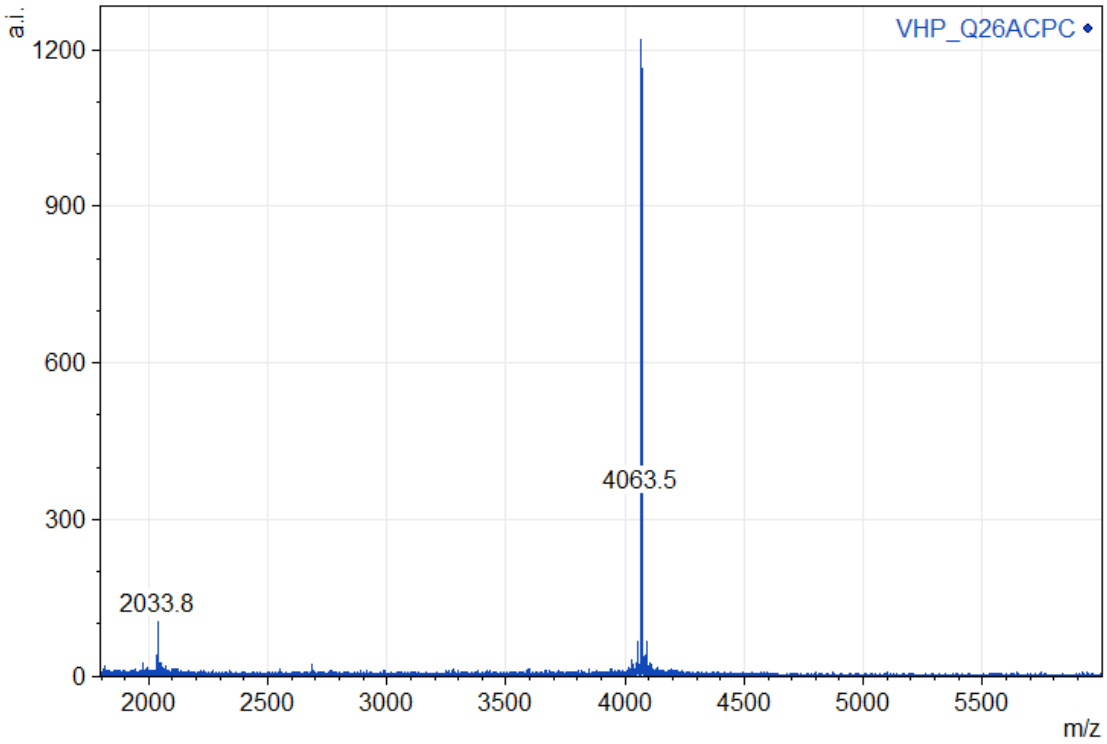


Figure 2.45. VHP Q26ACPC MALDI-TOF monoisotopic [M+H] peak is labeled at 4063.5 and the largest [M+2H] peak is labeled at 2033.8.

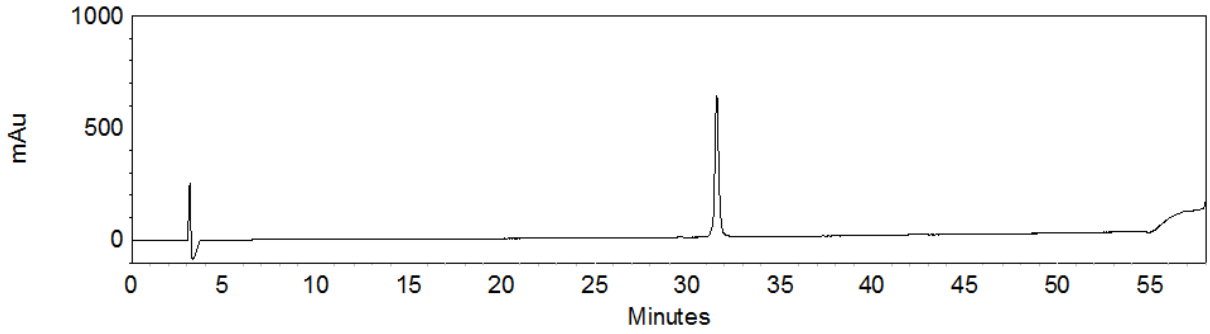


Figure 2.46. VHP Q26ACPC HPLC purity analysis with a 10-60% B HPLC gradient over 50 minutes (1.0 mL/min; column-Waters XBridge Peptide BEHC18 130 Å, 5 µm, 4.6 x 250 mm; purity >95%).

K30APC (circular dichroism) (APC: (1*S*,2*S*)-2-aminopyrrolidine carboxylic acid)
Sequence: H-LSDEDFKAVFGMTRSAFANLPLWKQQHLK(**APC**)EKGLF-OH
Calculated monoisotopic [M+H]: 4064.1

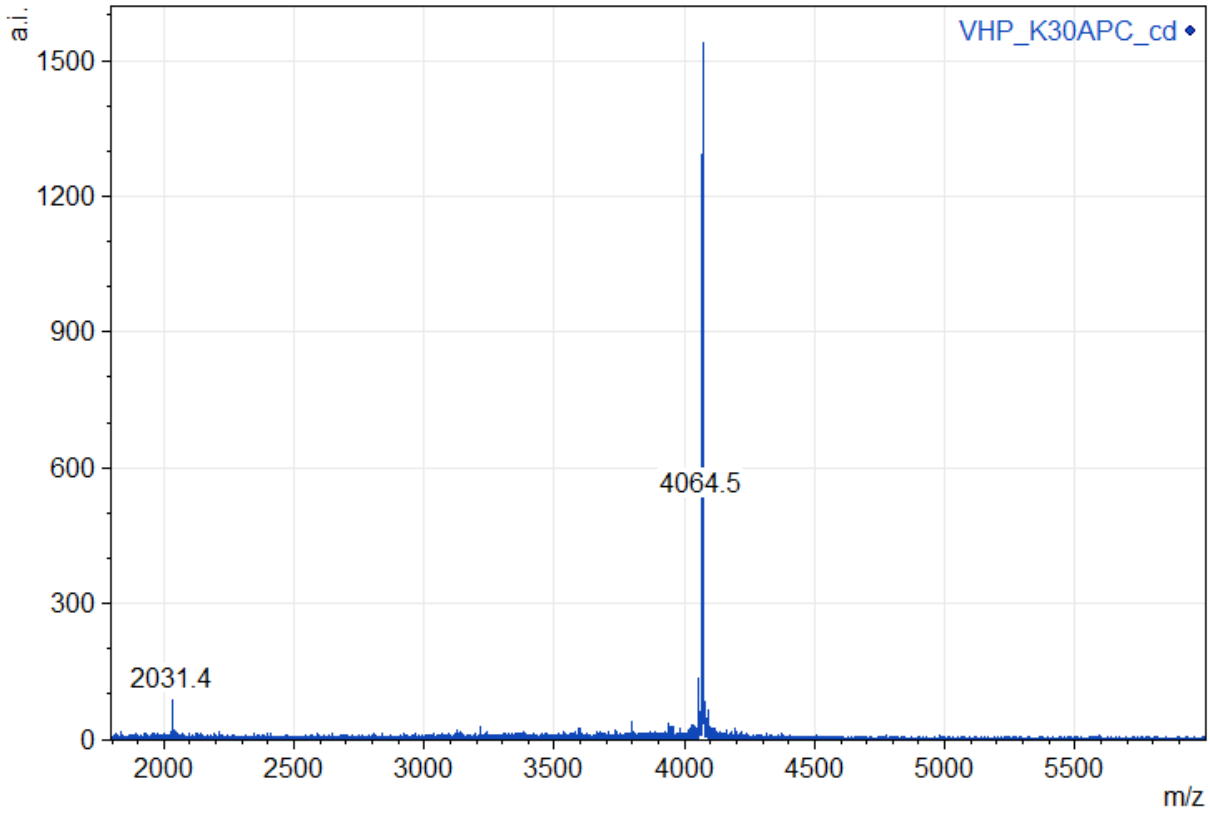


Figure 2.47. VHP K30APC MALDI-TOF monoisotopic [M+H] peak is labeled at 4064.5 and the largest [M+2H] peak is labeled at 2031.4.

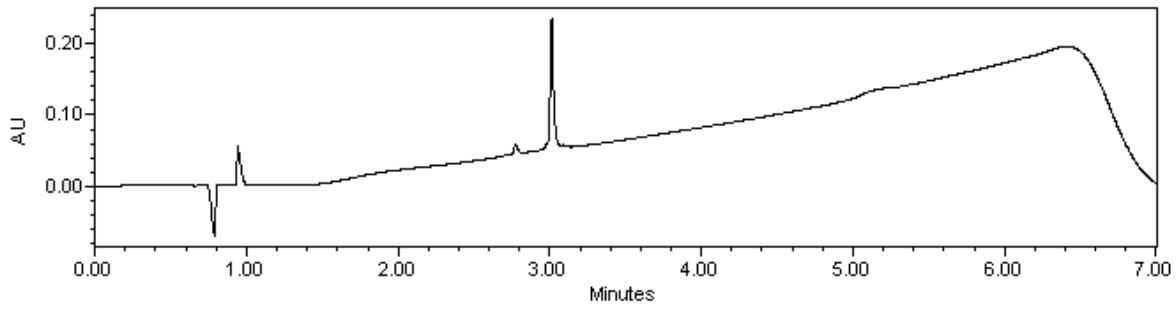


Figure 2.48. VHP K30APC purity analysis, UPLC gradient from 20-70% B solvent over 5 minutes (0.3 mL/min; column-Waters Acquity CSH C18 1.7 µm, 2.1 x 100 mm; purity >90%).

K30APC (crystallization) (APC: (1S,2S)-2-aminopyrrolidine carboxylic acid)
Sequence: H-LSDEDFKAVFGMTRSAFANLPLWKQQHLK(**APC**)EKGLF-OH
Calculated monoisotopic [M+H]: 4066.7

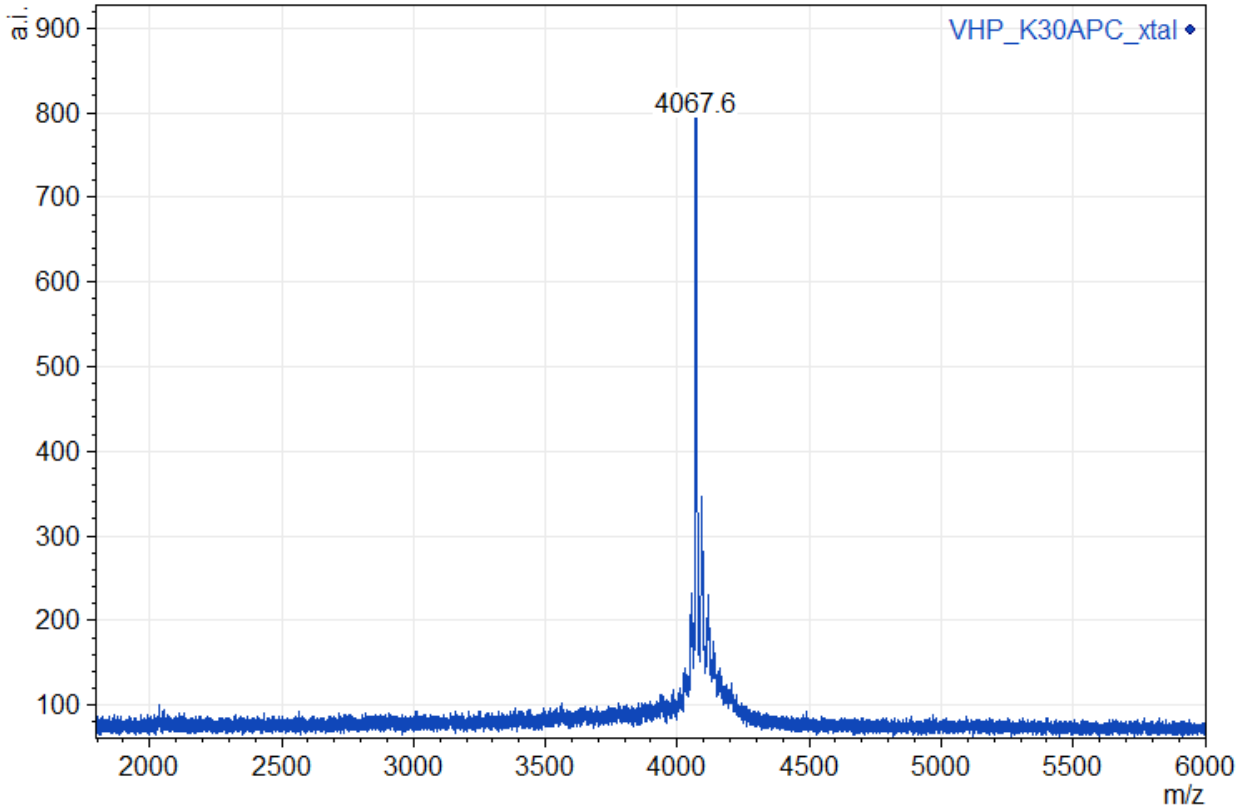


Figure 2.49. VHP K30APC MALDI-TOF average [M+H] peak is labeled at 4067.6.

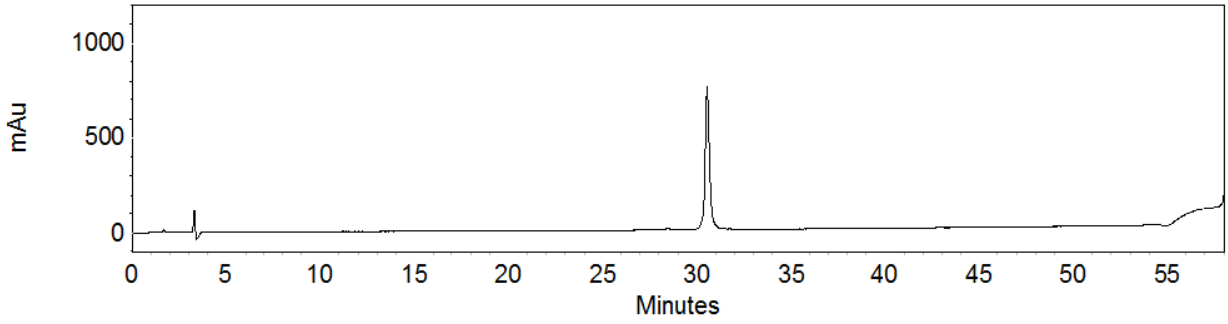


Figure 2.50. VHP K30APC HPLC purity analysis with a 10-60% B HPLC gradient over 50 minutes (1.0 mL/min; column-Waters XBridge Peptide BEHC18 130 Å, 5 µm, 4.6 x 250 mm; purity >95%).

VHP G11ACPC (ACPC: (1*R*,2*R*)-2-aminocyclopentane carboxylic acid)
Sequence: H-LSDEDFKAVFGMTRSAFANLPLWKQ(**ACPC**)HLKKEKGLF-OH
Calculated monoisotopic [M+H]: 4134.2

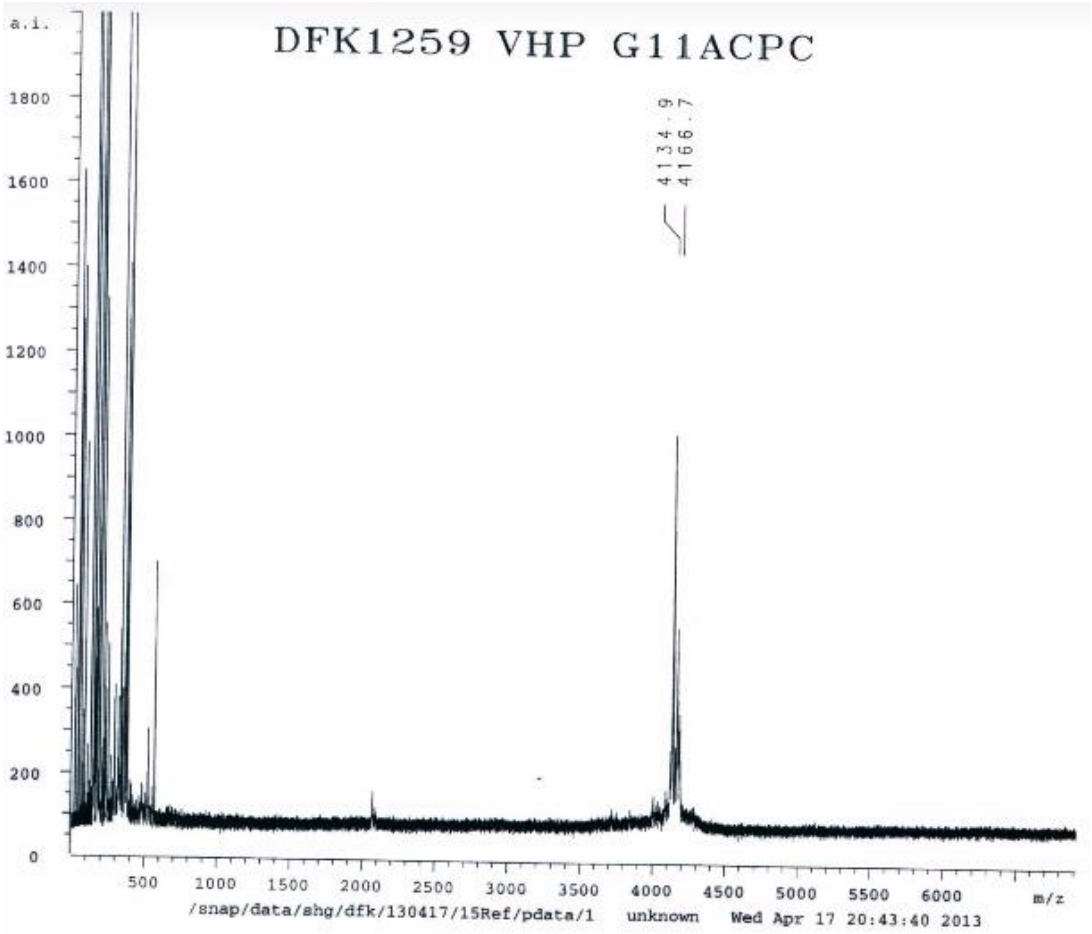


Figure 2.51. VHP G11(R,R)ACPC MALDI-TOF monoisotopic [M+H] peak is labeled at 4134.9, presumed oxidation (+32) peptide also appears at 4166.7.

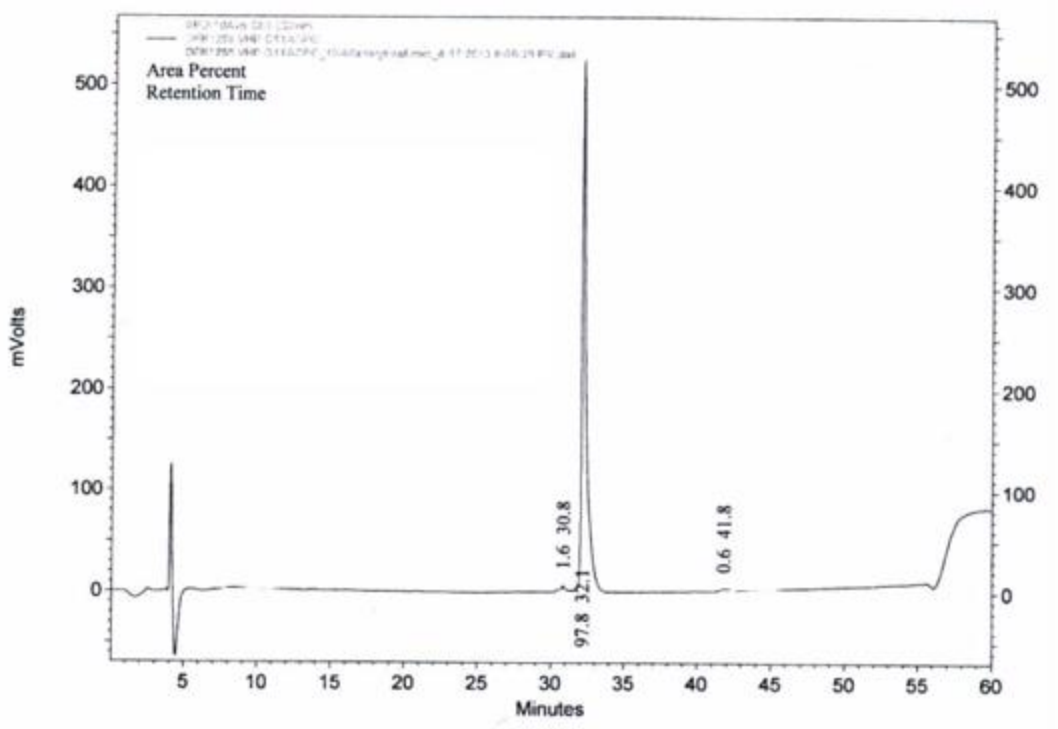


Figure 2.52. VHP G11(R,R)-ACPC HPLC purity analysis with a 10-60% B HPLC gradient over 50 minutes (1.0 mL/min; column-Waters XBridge Peptide BEHC18 130 Å, 5 µm, 4.6 x 250 mm; purity >95%).

2.4.3 Circular Dichroism Spectroscopy

All CD spectroscopy experiments were performed on an Aviv Biomedical model 420 CD spectrometer in 1 mm quartz strain-free cuvettes (Helma) with 50 µM peptide in 20 mM sodium acetate titrated to pH 5.0 with 1 M HCl and filtered through a 0.2 um PTFE syringe filter. Approximately ~1 mg of peptide/TFA salt was dissolved in 500 µL sodium acetate buffer, and peptide concentration was determined by measuring the absorbance at 280 nm in a 1:5 dilution of 6 M guanidine hydrochloride using an extinction coefficient of 5690 $M^{-1}cm^{-1}$ which corresponds to the single tryptophan side-chain.[60] The peptide stock solution was then diluted to 50 µM with the sodium acetate buffer. Due to the lack

of a chromophore in VHP W23ACPC, ε_{214} was estimated to be 73,790 $M^{-1}cm^{-1}$ by measuring the absorbance at 214 nm of an L-VHP solution of known concentration in the absence of the denaturant, and this value was used to determine the concentration of VHP W23ACPC Wavelength scans were collected from 260 nm to 190 nm with a 1 nm bandwidth and 10 second averaging time. The CD-spectrometer was calibrated with a 1 mg/mL solution of (+)-camphor sulfonic acid (CSA) in water prior to each wavelength scan.[61] For thermal denaturations, ellipticity was measured at 223 nm from 2-98°C in 2 degree increments with a 5 minute equilibration time and a 10 second averaging time.

2.4.4 Thermal Melt Modeling

For each thermal denaturation the data were fit to a two-state folding model. Previously reported calorimetric studies of VHP folding thermodynamics indicate that the folding of VHP can be accurately modeled as a two-state system.[31] The general procedure for modeling two-state folding thermodynamics described by Greenfield[30] was followed to fit the data collected in this study using Equations 2.1-2.18 and the following parameters:

θ_T measured ellipticity

θ_f ellipticity contribution from folded conformation

θ_u ellipticity contribution from unfolded conformation

θ_E ellipticity residual between two-state model and measured ellipticity

N_f The number of data points in the folded baseline region

N_u The number of data points in the unfolded baseline region

m_f slope of folded baseline; determined from linear regression on N_f points

b_f y-intercept of folded baseline; determined from linear regression on N_f points

m_u slope of unfolded baseline; determined from linear regression on N_u points

b_u y-intercept of unfolded baseline; determined from linear regression on N_u points

T Temperature

α fraction folded conformation

K equilibrium constant of folding

R ideal gas constant, 0.

ΔH_m enthalpy of folding

ΔS_m entropy of folding

ΔC_p heat capacity of folding

T_m melting temperature

The measured ellipticity is considered to be a linear combination of the signal from both the folded and unfolded conformation. The linear coefficients, as determined by the fraction of the peptide in the folded conformation, are nonlinear functions of the model parameters T_m and ΔH_m.

$$\theta_T = \alpha\theta_F + (1-\alpha)\theta_U \quad (1)$$

The folded and unfolded baselines are modeled as linear functions of temperature. The parameters m_f and b_f are determined by linear regression performed on a line-of-best-fit defined by the number of points in the folded region, N_f; similarly, m_u and b_u are determined by the number of points in the unfolded region, N_u.

$$\theta_F = m_f T + b_f \quad (2.2)$$

$$\theta_U = m_U T + b_U \quad (2.3)$$

The fraction of peptide in the folded conformation is defined by rearranging the expression for the equilibrium constant of folding (Equation 2.4), which is a function of the Gibbs free energy of unfolding (Equation 2.5), which can be obtained by combining Equations 2.6-2.8.

$$\alpha = \frac{1}{1+K} \quad (2.4)$$

$$K = e^{-\Delta G/RT} \quad (2.5)$$

$$\Delta H = \Delta H_m + \int_{T_m}^{T} C_p dT = \Delta H_m + C_p(T - T_m) \quad (2.6)$$

$$\Delta S = \Delta S_m + \int_{T_m}^{T} C_p dlnT = \frac{\Delta H_m}{T_m} + C_p ln\left(\frac{T}{T_m}\right) \quad (2.7)$$

The heat capacity of folding was assumed to be zero. We reason that this assumption is valid for two reasons: ΔC_p is relatively small for VHP compared to larger proteins,[31] and our data can be accurately modeled with only T_m and ΔH_m so there is no need to introduce additional parameters. Given these assumptions, Equation 2.8 describes the Gibbs free energy of unfolding as a function of T_m and ΔH_m.

$$\Delta G = \Delta H - T\Delta S = \Delta H_m \left(1 - \frac{T}{T_m}\right) \quad (2.8)$$

The measured ellipticity signal is represented as a function of the model parameters Tm, ΔH_m, m_u, b_u, m_f, and b_f (Equation 2.9).

$$\theta_T = f(T; \Delta H_m, T_m, m_f, b_f, m_u, b_u) \quad (2.9)$$

Estimation of the two-state model parameters, T_m and ΔH_m, requires nonlinear regression of a scalar function defined by the sum of the square residual error between the model

ellipticity and the measured ellipticity. The square residual error is weighted by the inverse of the standard deviation of each ellipticity measurement, σ_i.

$$\theta_E = \sum_{i=1}^{N} \frac{\left(\theta_{T,obs,i} - \theta_{T,model,i}\right)^2}{\sigma_i} \qquad (2.10)$$

Non-linear regression was performed using the Levenberg-Marquardt algorithm. The partial derivatives of the error function (Equation 2.10) with respect to the model parameters T_m and ΔH_m were determined analytically with the program *Mathematica* (Equations 2.11 and 2.12).

$$\left.\frac{\partial \theta_{E,i}}{\partial \Delta H_m}\right|_{T_m} = \frac{-e^{\frac{\Delta H_m}{RT}\left(\frac{T_i}{T_m}-1\right)}(T_i - T_m)\left(m_f T_i - m_u T_i + b_f - b_u\right)}{\left(1+e^{\frac{\Delta H_m}{RT}\left(\frac{T_i}{T_m}-1\right)}\right)^2 R T_i T_m} \qquad (2.11)$$

$$\left.\frac{\partial \theta_{E,i}}{\partial T_m}\right|_{\Delta H_m} = \frac{\Delta H_m e^{\frac{\Delta H_m}{RT_i}\left(\frac{T_i}{T_m}-1\right)}\left(m_f T_i - m_u T_i + b_f - b_u\right)}{\left(1+e^{\frac{\Delta H_m}{RT_i}\left(\frac{T_i}{T_m}-1\right)}\right)^2 R T_m^2} \qquad (2.12)$$

The partial derivatives represented as vector-valued functions are used to construct a 2x48 Jacobian (Equations 2.13 and 2.14) which corresponds to the two model parameters (T_m, ΔH_m) and the 48 data points collected during each thermal denaturation experiment.

$$\boldsymbol{J} = \begin{bmatrix} \frac{\partial \theta_{E,1}}{\partial \Delta H_m} & \cdots & \frac{\partial \theta_{E,n}}{\partial \Delta H_m} \\ \frac{\partial \theta_{E,1}}{\partial T_m} & \cdots & \frac{\partial \theta_{E,n}}{\partial T_m} \end{bmatrix} \qquad (2.13)$$

$$\boldsymbol{J}^T = \begin{bmatrix} \frac{\partial \theta_{E,1}}{\partial \Delta H_m} & \frac{\partial \theta_{E,1}}{\partial T_m} \\ \vdots & \vdots \\ \frac{\partial \theta_{E,1}}{\partial \Delta H_m} & \frac{\partial \theta_{E,n}}{\partial T_m} \end{bmatrix} \qquad (2.14)$$

An approximate 2x2 hessian is constructed from the product of the Jacobian and its transpose (Equation 2.15).

$$\boldsymbol{H} = \boldsymbol{J}\boldsymbol{J}^{\boldsymbol{T}} \qquad (2.15)$$

The Levenberg-Marquardt algorithm is a hybrid between Newton's method and gradient descent minimization. The parameter shift vector, δ, is solved for analytically to iteratively minimize the residual between the model ellipticity and the measured ellipticity (Equations 2.16-2.18). The parameter shift approximates a gradient descent when the residual between the model and the measurement is large; when the residual is smaller the parameter shift approximates Newton's method. This combination between the two procedures is achieved with the use of a dampening parameter, λ, which increases the magnitude of the parameter shift far from a solution and decreases the magnitude of the shift in the neighborhood of a solution.

$$(\boldsymbol{H} + \boldsymbol{\lambda}\, diag(\boldsymbol{H}))\boldsymbol{\delta} = \boldsymbol{J}(\boldsymbol{\theta}_T - \widehat{\boldsymbol{\theta}}_T) \qquad (2.16)$$

$$\boldsymbol{H}^* = \boldsymbol{H} + \boldsymbol{\lambda}\, diag(\boldsymbol{H}) \qquad (2.17)$$

$$\boldsymbol{\delta} = \boldsymbol{H}^{*-1}\boldsymbol{J}(\boldsymbol{\theta}_T - \widehat{\boldsymbol{\theta}}_T) \qquad (2.18)$$

Figure 2.53 outlines the Levenberg-Marquardt algorithm procedure used to fit the model parameters T_m and ΔH_m for each thermal denaturation experiment. Finally, for each denaturation experiment the optimal number of points for folded and unfolded baselines (N_f,N_u) was determined empirically with the use of a python script utilizing the Numpy module that fit T_m and ΔH_m for 484 different combinations of N_f and N_u which correspond to all possible combinations of folded and unfolded baselines given the constraints that unfolding transitions began before 48°C and that each baseline must contain at least two points (22 different N_f values multiplied by 22 different possible N_u values, 22^2=484).

Average two-state melting parameters were determined by performing three independent denaturation experiments and parameter fitting sessions. The average value of T_m and ΔH_m for each VHP variant was then reported as the average from the three fitting sessions. The result of each of these fitting sessions is shown below for each peptide.

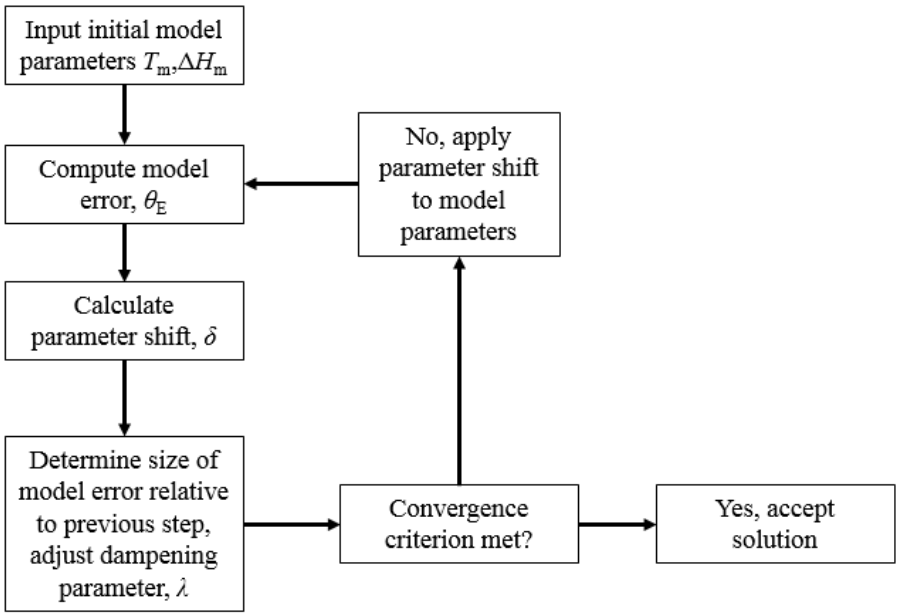


Figure 2.53. Levenberg-Marquardt algorithm used to perform nonlinear regression analysis of model parameters T_m and ΔH_m.

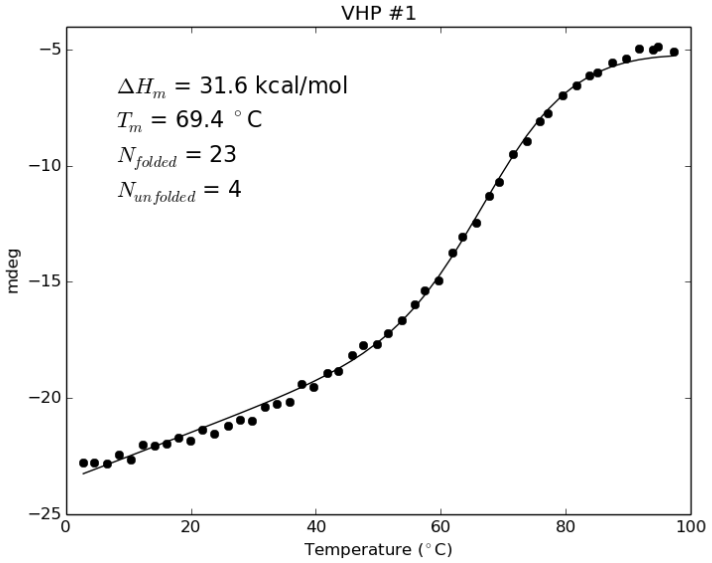


Figure 2.54. L-VHP variable temperature circular dichroism signal monitored at 223 nm fit to a two-state folding model.

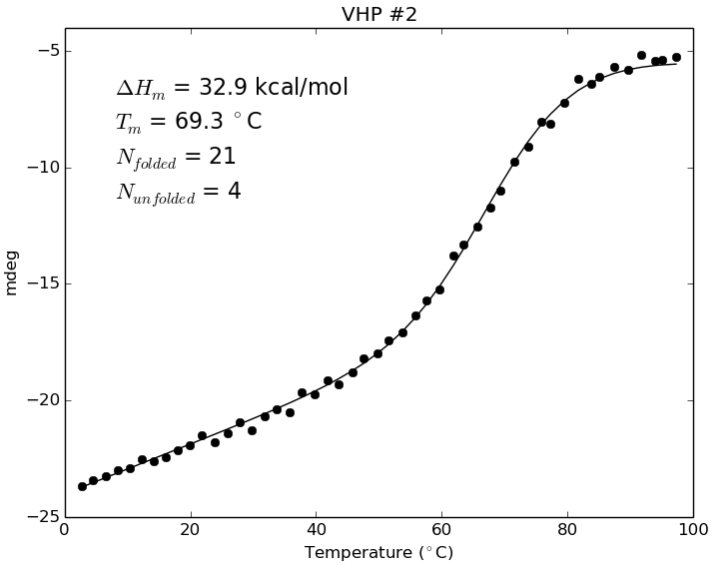


Figure 2.55. L-VHP variable temperature circular dichroism signal monitored at 223 nm fit to a two-state folding model.

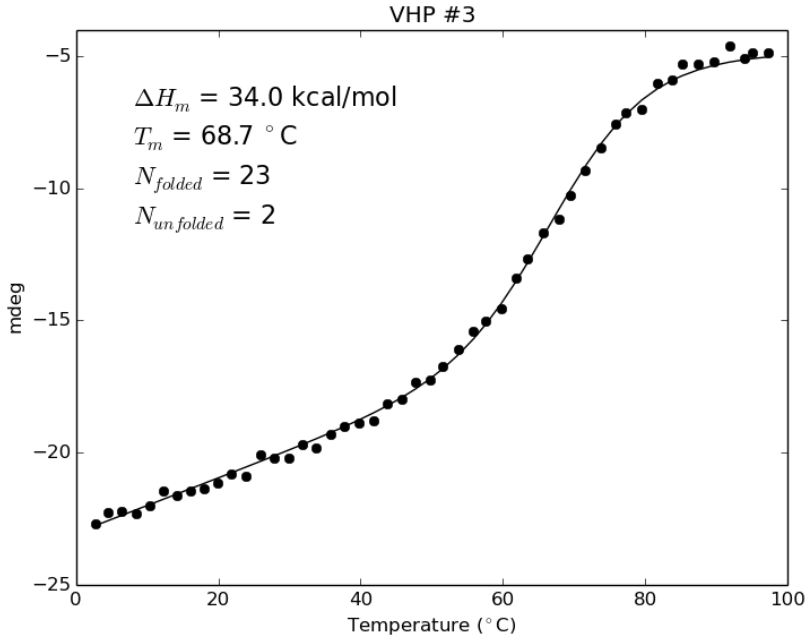


Figure 2.56. L-VHP variable temperature circular dichroism signal monitored at 223 nm fit to a two-state folding model.

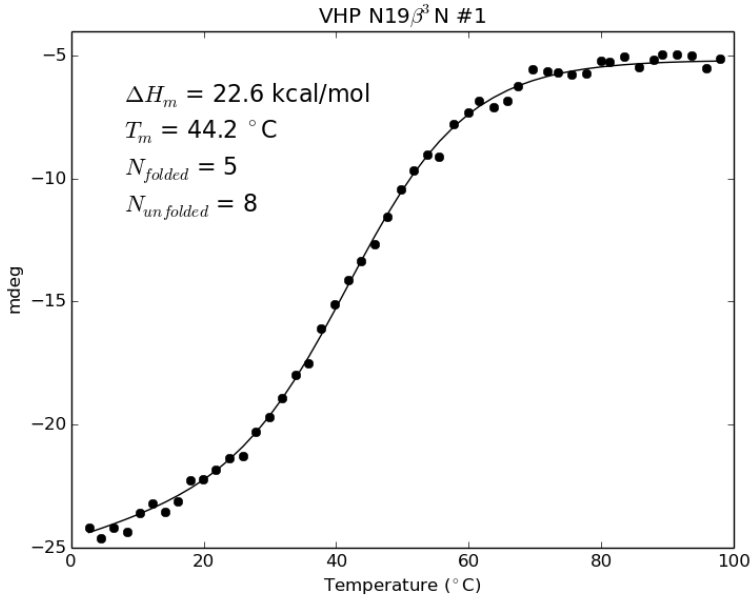


Figure 2.57. VHP β^{3}hAsn19 variable temperature circular dichroism signal monitored at 223 nm fit to a two-state folding model.

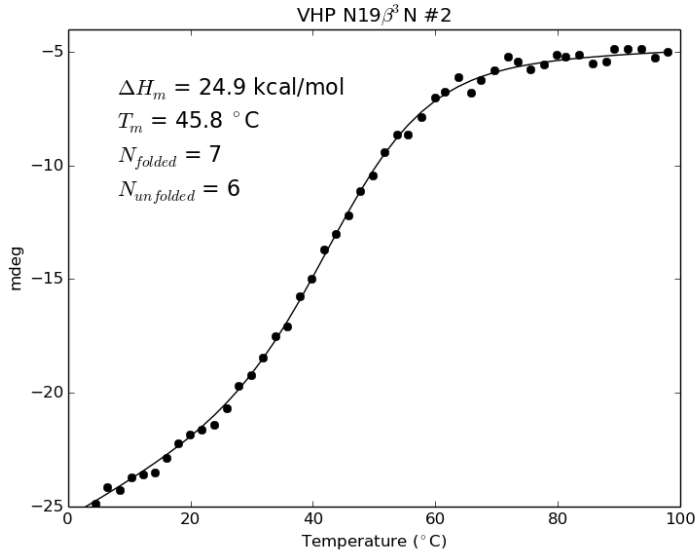


Figure 2.58. VHP β3hAsn19 variable temperature circular dichroism signal monitored at 223 nm fit to a two-state folding model.

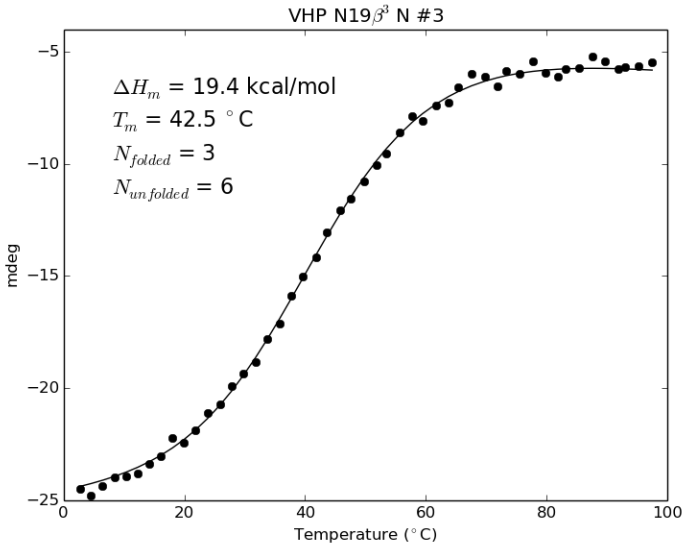


Figure 2.59. VHP β3hAsn19 variable temperature circular dichroism signal monitored at 223 nm fit to a two-state folding model.

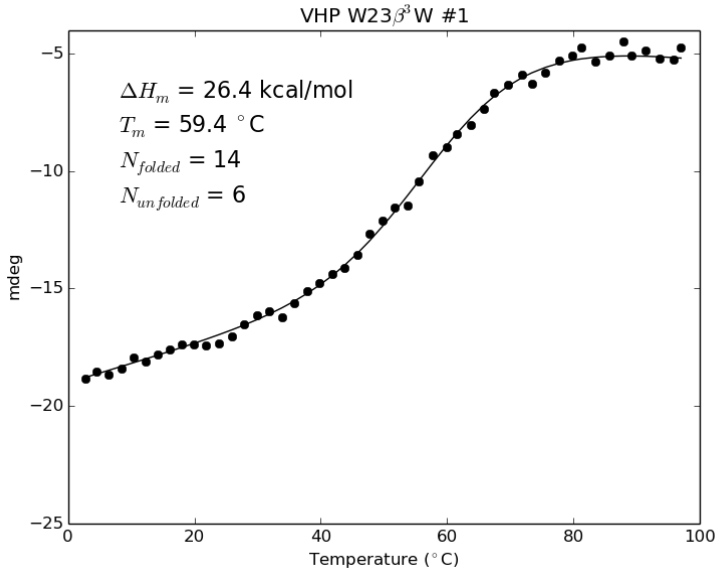


Figure 2.60. VHP β^{3}hTrp23 variable temperature circular dichroism signal monitored at 223 nm fit to a two-state folding model.

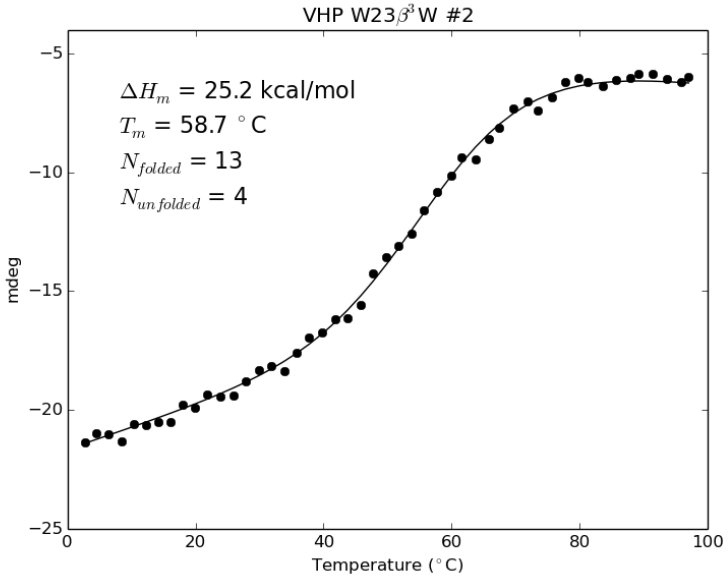


Figure 2.61. VHP β^{3}hTrp23 variable temperature circular dichroism signal monitored at 223 nm fit to a two-state folding model.

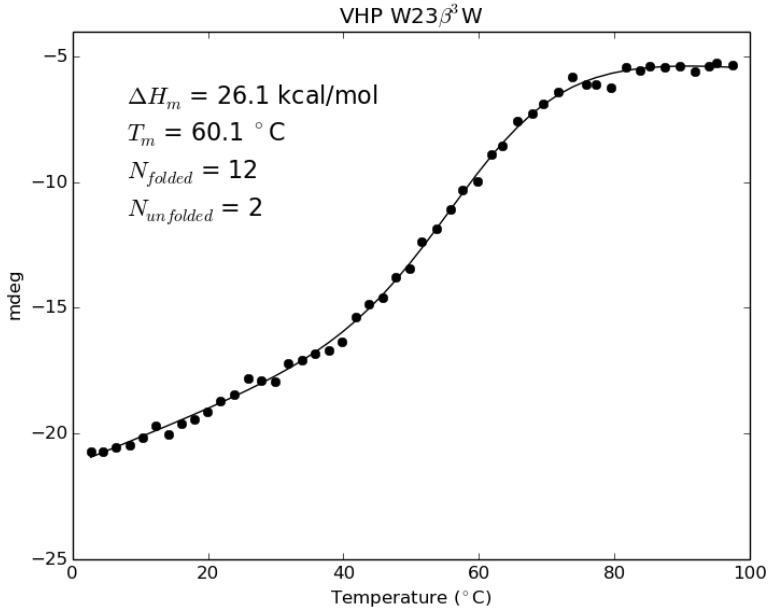


Figure 2.62. VHP β3hTrp23 variable temperature circular dichroism signal monitored at 223 nm fit to a two-state folding model.

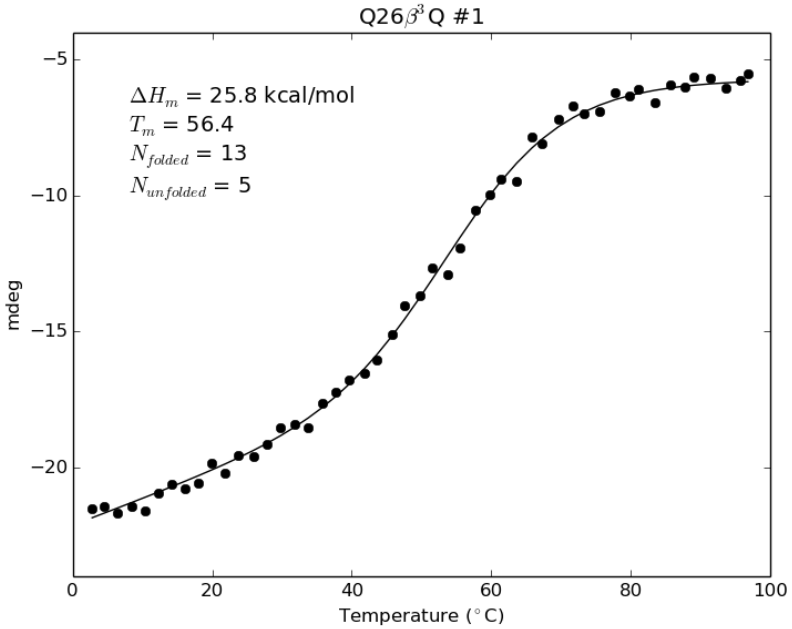


Figure 2.63. VHP β^{3}hGln26 variable temperature circular dichroism signal monitored at 223 nm fit to a two-state folding model.

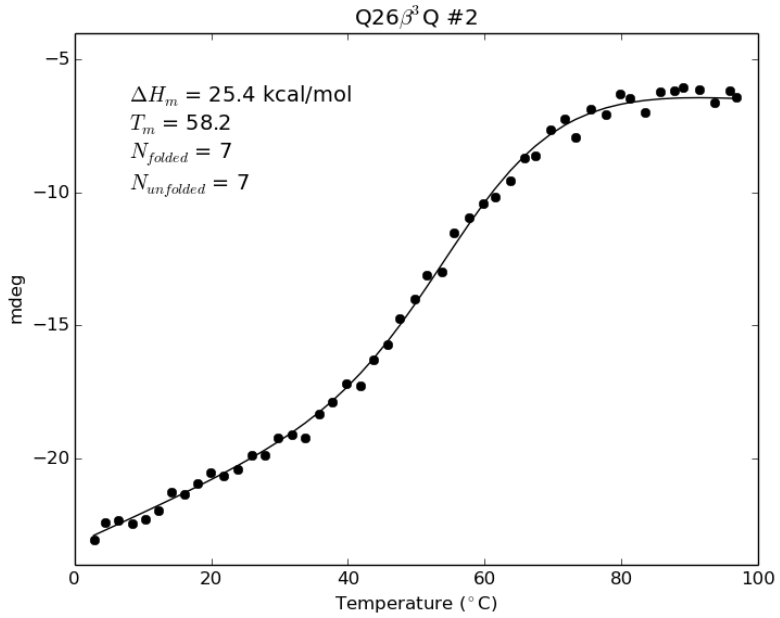


Figure 2.64. VHP β3hGln26 variable temperature circular dichroism signal monitored at 223 nm fit to a two-state folding model.

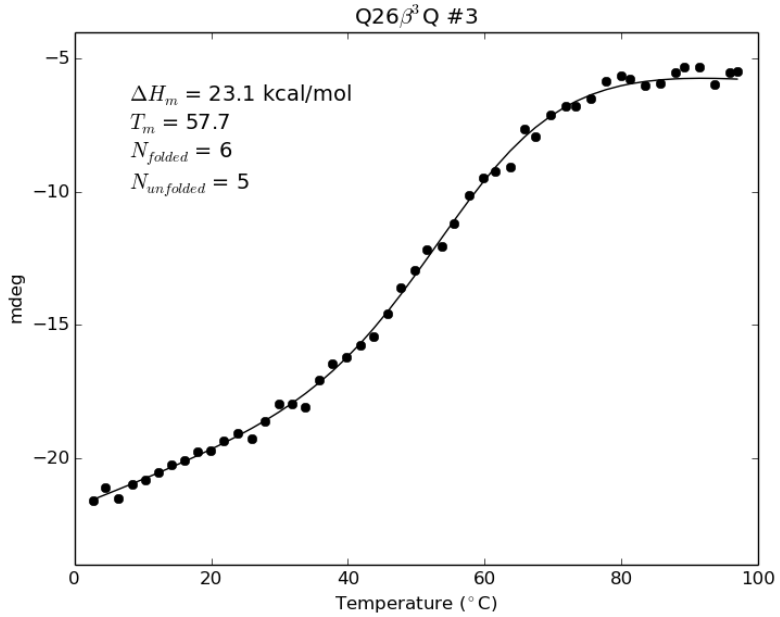


Figure 2.65. VHP β^{3}hGln26 variable temperature circular dichroism signal monitored at 223 nm fit to a two-state folding model.

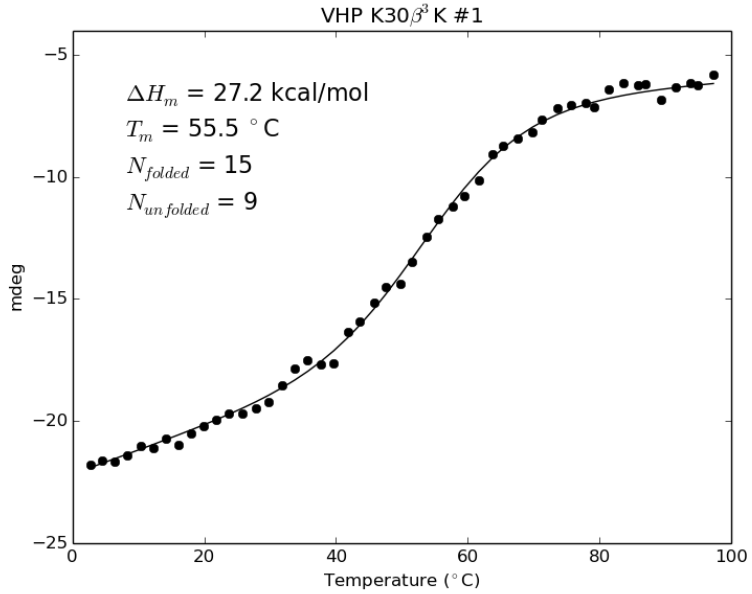


Figure 2.66. VHP β3hLys30 variable temperature circular dichroism signal monitored at 223 nm fit to a two-state folding model.

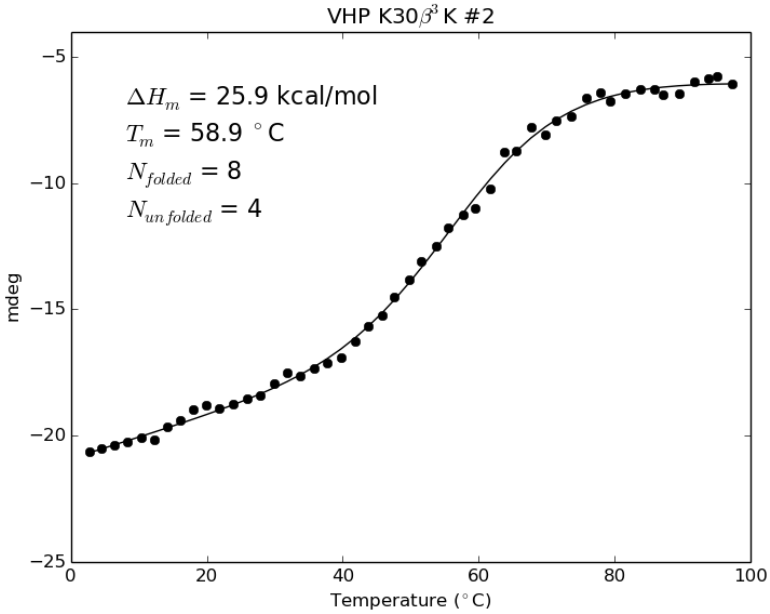


Figure 2.67. VHP β^{3}hLys30 variable temperature circular dichroism signal monitored at 223 nm fit to a two-state folding model.

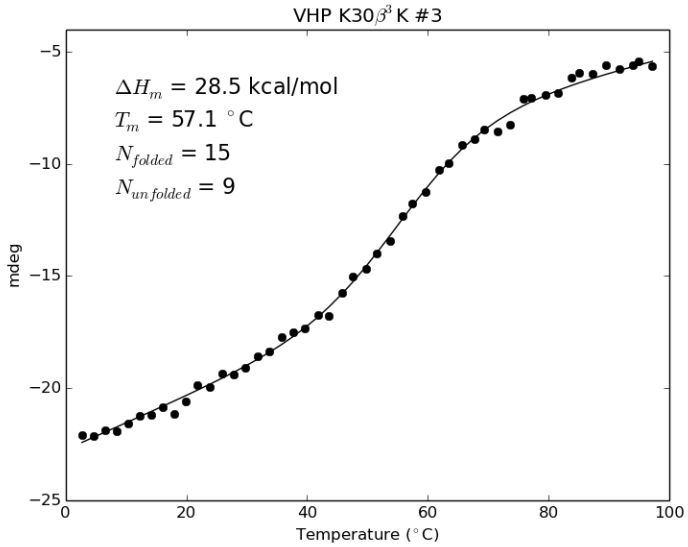


Figure 2.68. VHP β3hLys30 variable temperature circular dichroism signal monitored at 223 nm fit to a two-state folding model.

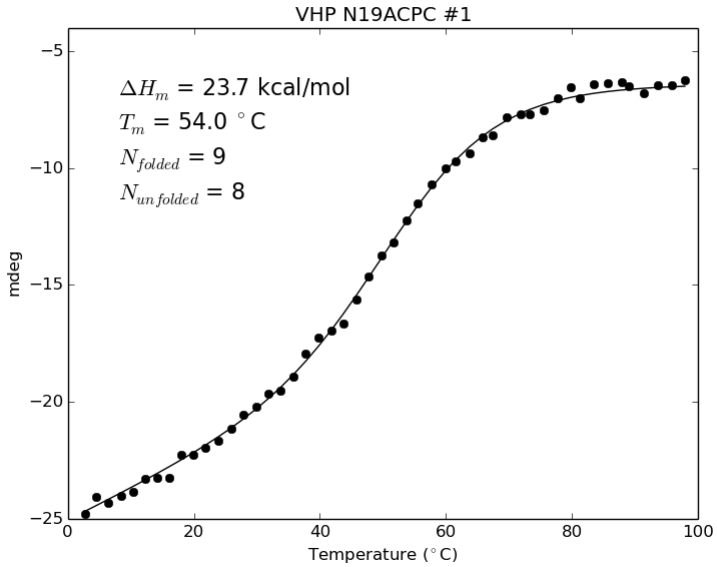


Figure 2.69. VHP ACPC19 variable temperature circular dichroism signal monitored at 223 nm fit to a two-state folding model.

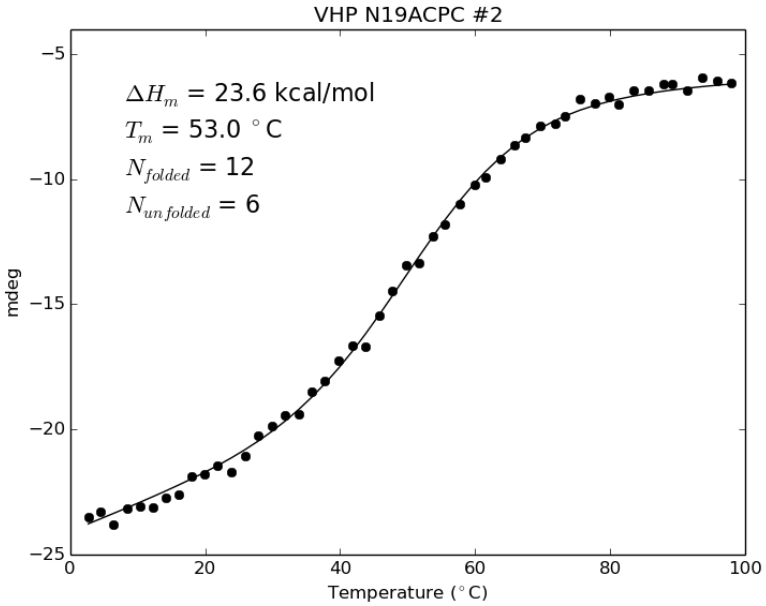


Figure 2.70. VHP ACPC19 variable temperature circular dichroism signal monitored at 223 nm fit to a two-state folding model.

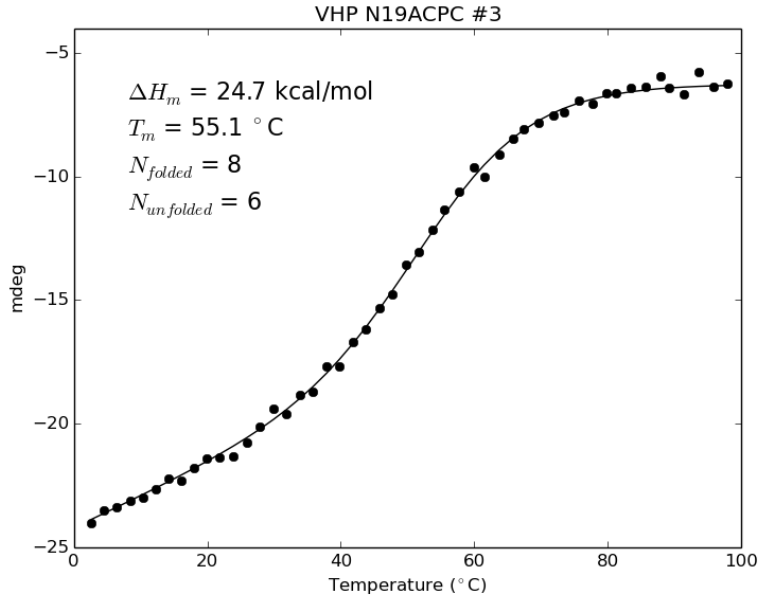


Figure 2.71. VHP ACPC19 variable temperature circular dichroism signal monitored at 223 nm fit to a two-state folding model.

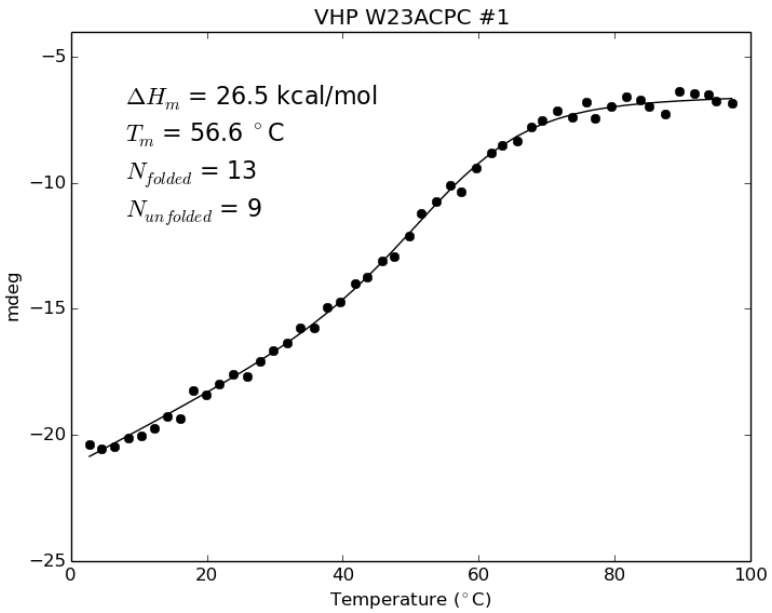


Figure 2.72. VHP ACPC23 variable temperature circular dichroism signal monitored at 223 nm fit to a two-state folding model.

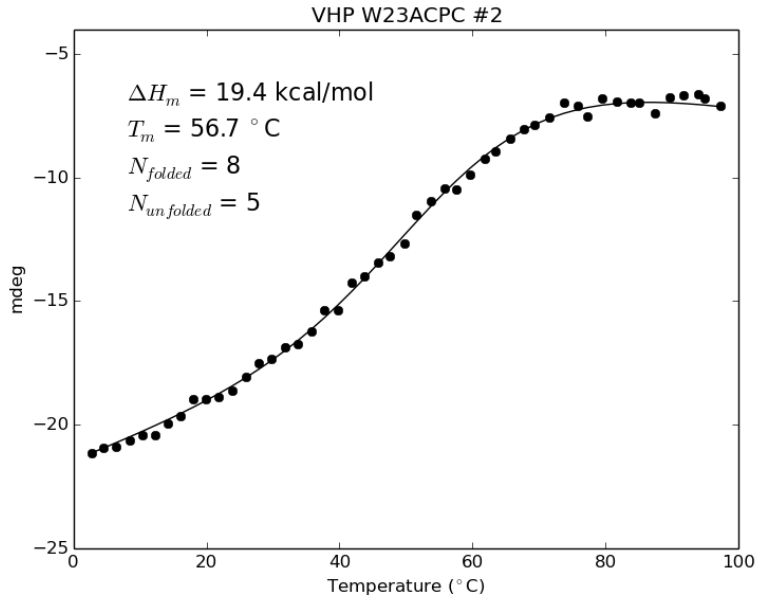


Figure 2.73. VHP ACPC23 variable temperature circular dichroism signal monitored at 223 nm fit to a two-state folding model.

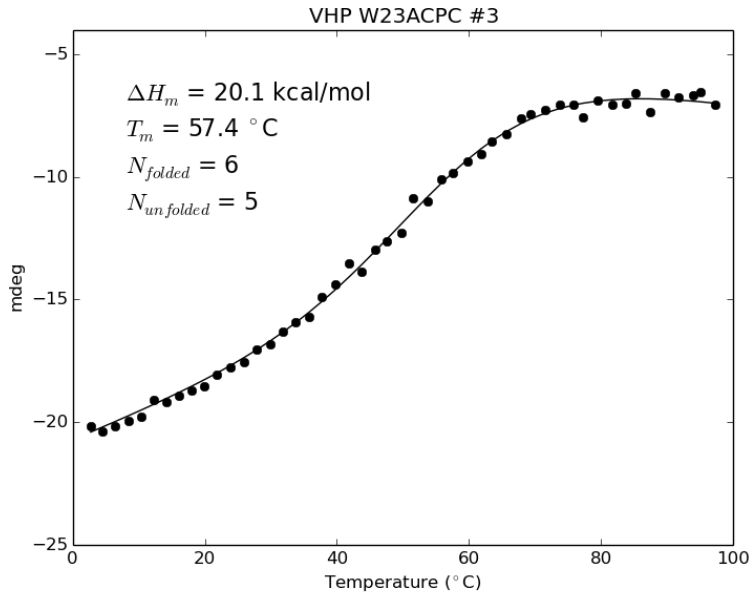


Figure 2.74. VHP ACPC23 variable temperature circular dichroism signal monitored at 223 nm fit to a two-state folding model.

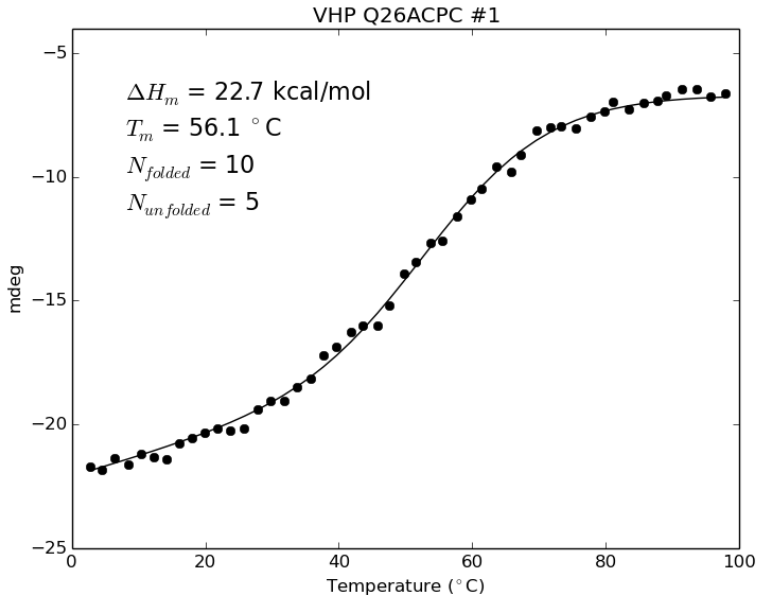


Figure 2.75. VHP ACPC26 variable temperature circular dichroism signal monitored at 223 nm fit to a two-state folding model.

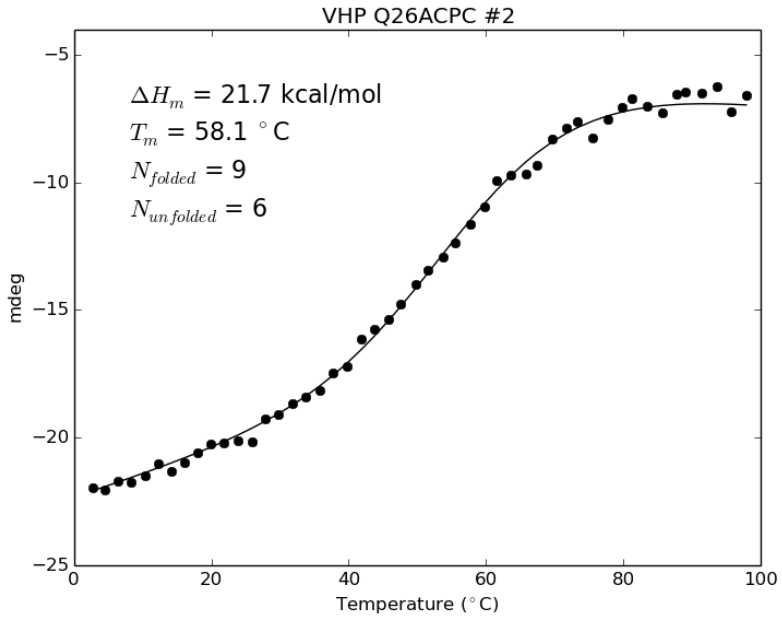


Figure 2.76. VHP ACPC26 variable temperature circular dichroism signal monitored at 223 nm fit to a two-state folding model.

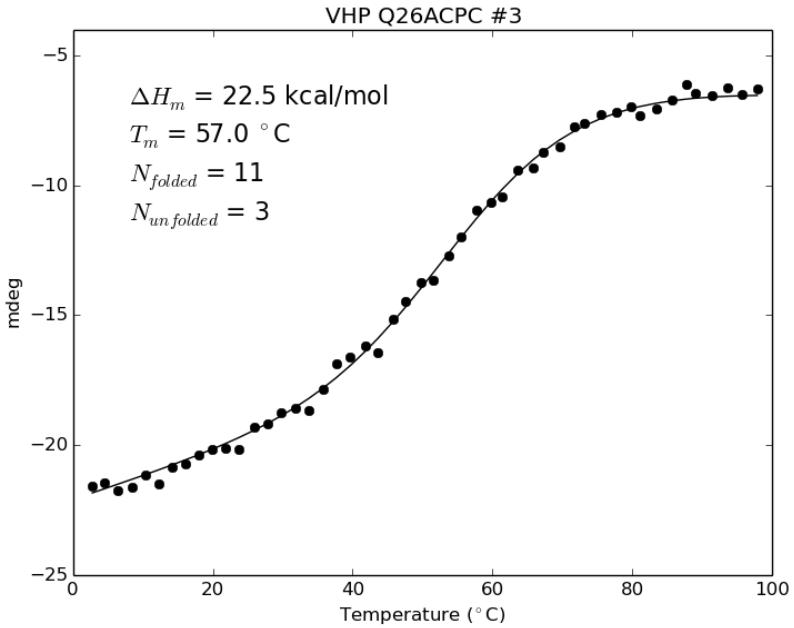


Figure 2.77. VHP ACPC26 variable temperature circular dichroism signal monitored at 223 nm fit to a two-state folding model.

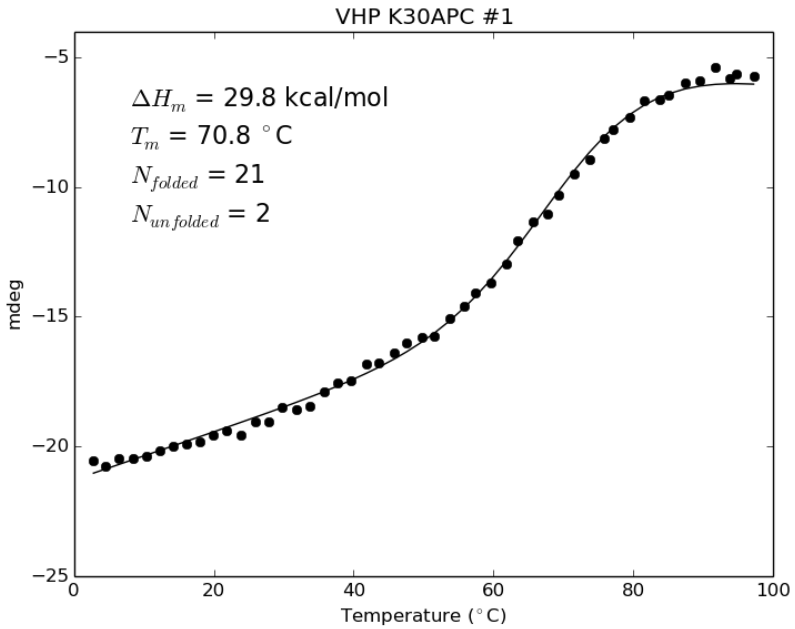


Figure 2.78. VHP APC30 variable temperature circular dichroism signal monitored at 223 nm fit to a two-state folding model.

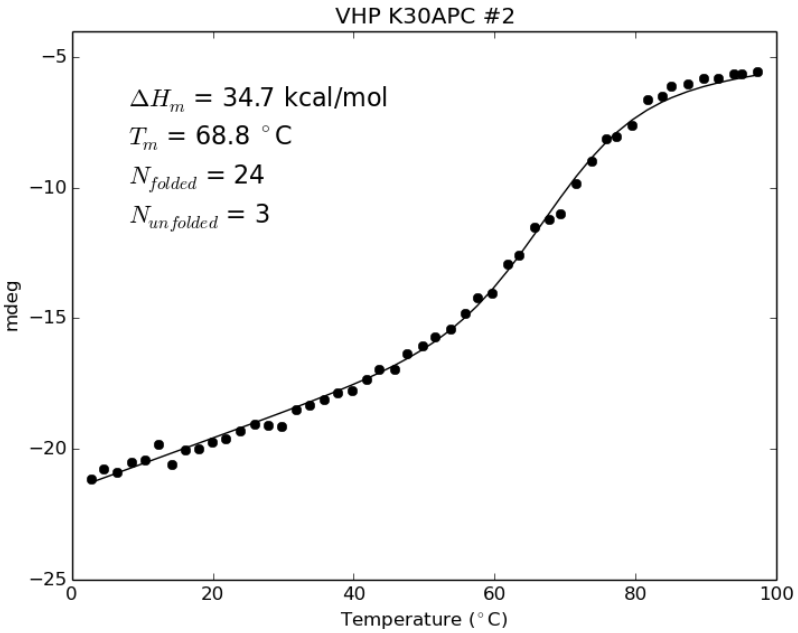


Figure 2.79. VHP APC30 variable temperature circular dichroism signal monitored at 223 nm fit to a two-state folding model.

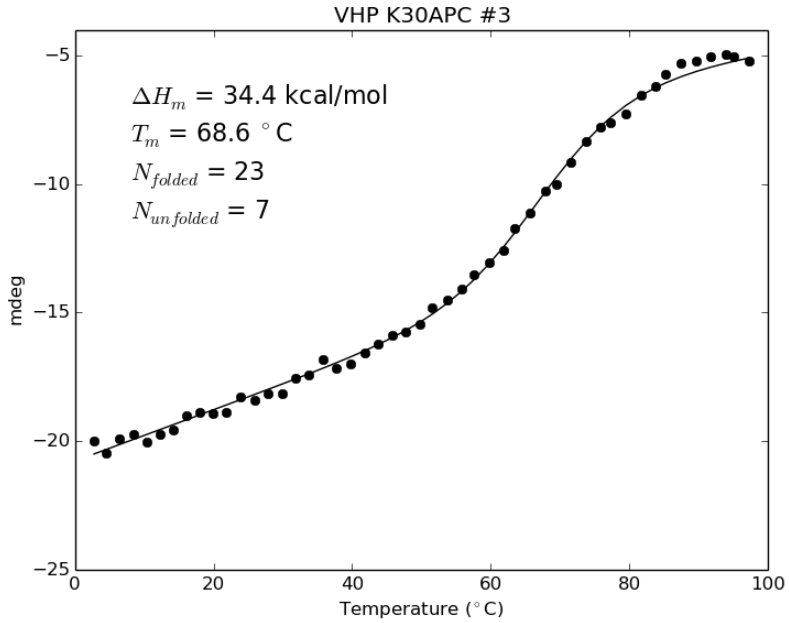


Figure 2.80. VHP APC30 variable temperature circular dichroism signal monitored at 223 nm fit to a two-state folding model.

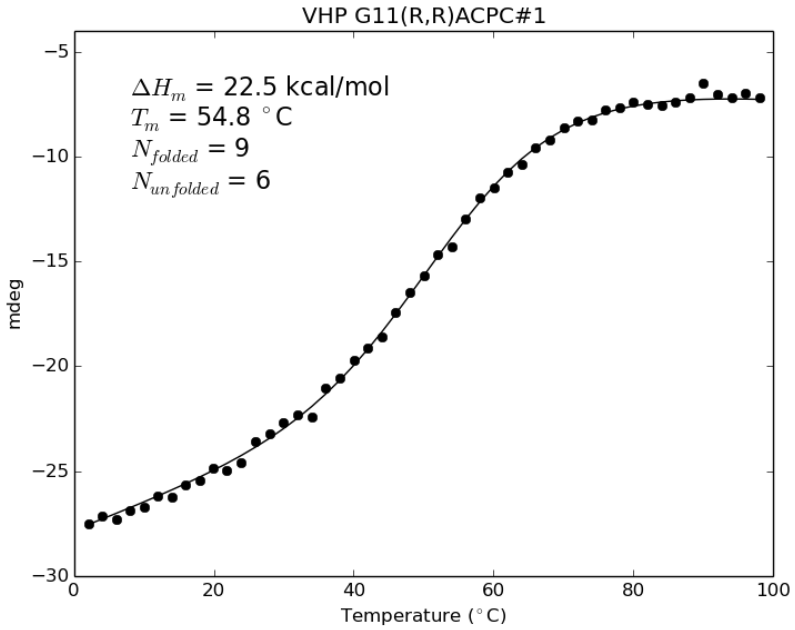


Figure 2.81. VHP (R,R)-ACPC11 variable temperature circular dichroism signal monitored at 223 nm fit to a two-state folding model.

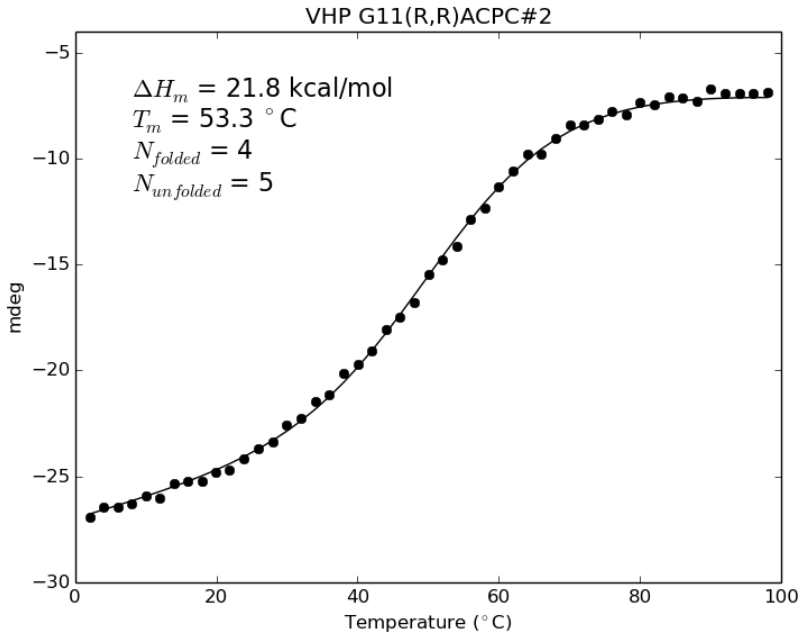


Figure 2.82. VHP (R,R)-ACPC11 variable temperature circular dichroism signal monitored at 223 nm fit to a two-state folding model.

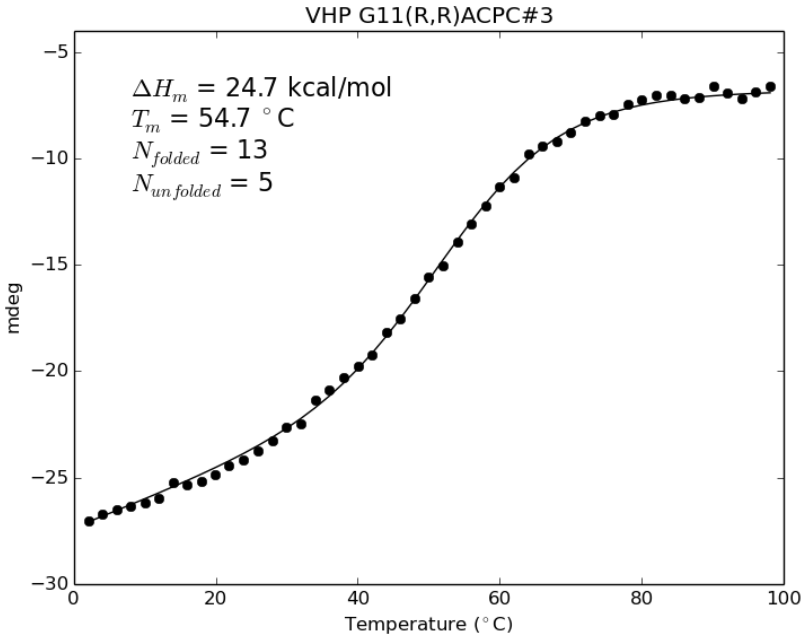


Figure 2.83. VHP (R,R)-ACPC11 variable temperature circular dichroism signal monitored at 223 nm fit to a two-state folding model.

2.4.5 Alternative Methods for Fitting Thermal Denaturation Data.

The previous section has described a procedure for determining T_m and ΔH_m of folding for a two-state system; ΔH_m is essentially the van't Hoff enthalpy of folding. Significant problems that arise when fitting a thermal denaturation sigmoid within a peptide system are the marginal folding cooperativity of such peptides and the linearity of a peptide signal in the pre-transition region as a function of temperature. This linearity becomes an issue when defining the end of the folded baseline and thus the beginning of the unfolding temperature range. Reliable selection of this region is critical for reproducible determination of the parameters T_m and ΔH_m. Authors will often circumvent this issue by listing the 'uncertainty of the fit' simply as the curvature of the parameter gradient at the parameter vector that minimizes the objective error function (the square of the sum of the residuals),[22,23] a common practice in nonlinear least squares regression. This practice however, does not address the reproducibility issue because the definition of the baseline region can alter the determined model parameters by several degrees.

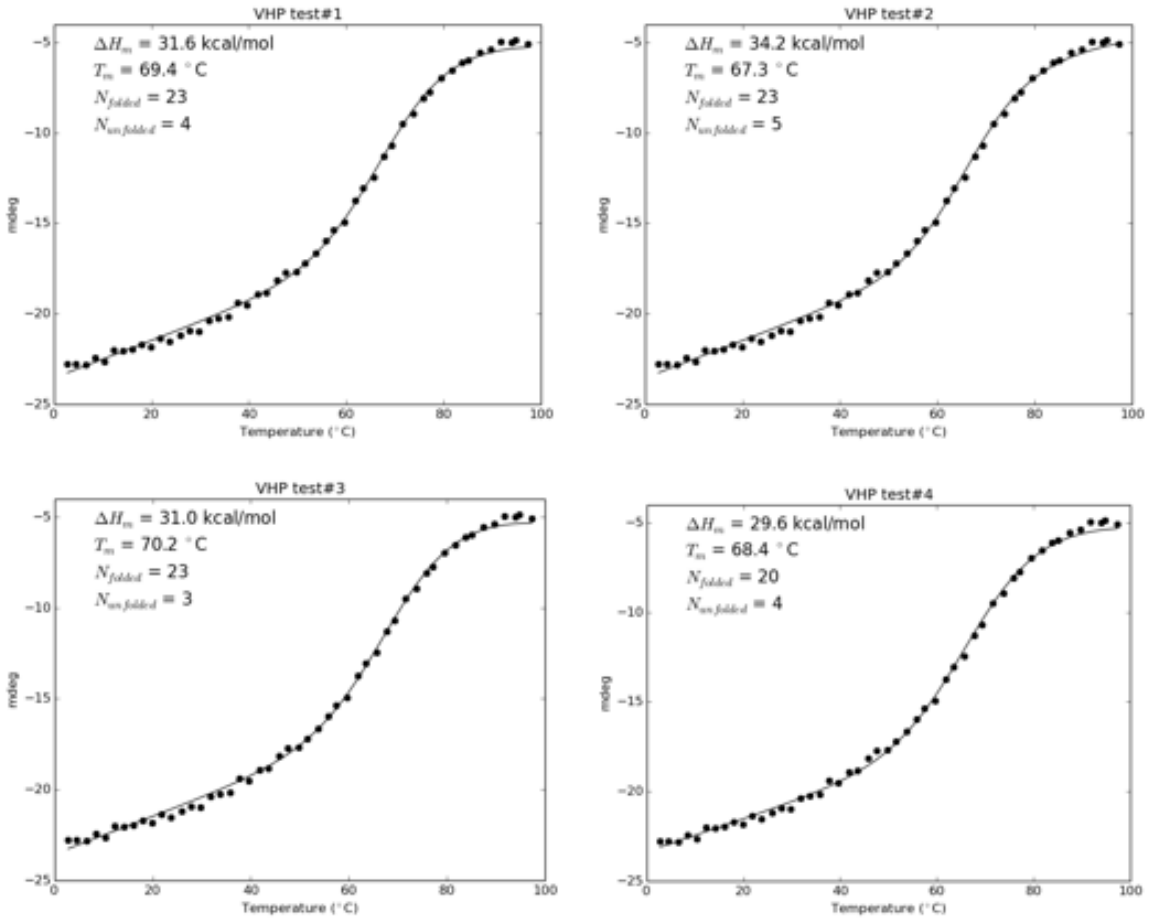


Figure 2.84. The same set of data points are fitted to compare the impact of slightly different folded and unfolded baseline definitions. The variation of T_m and ΔH_m is significant, even when the definition of the folded or unfolded baseline changes by only one or two points.

An alternative method of determining T_m depends on the fraction folded equation, which by definition has an inflection point when ΔG of folding equals zero. The problem of defining the starting and stopping point of the transition region is avoided, and T_m corresponds to a local maximum of the first derivative of the spectroscopic signal. This method is convenient when the baseline region is unclear; however in the presence of a linearly changing baseline, ΔH_m will not be easily determined. Another complication arises when the relative magnitude of the first derivative is small, such is the case in a peptide

system where the cooperativity, *i.e.* ΔH_m, is marginal, and the landscape of a first derivative spline function is therefore rough (Figure 2.85). This roughness results in multiple inflection point candidates, but the application of a Savitzky-Golay smoothing filter alleviates this problem (Figure 2.86). For comparison of calculated T_m values, the baseline Levenberg-Marquardt method yields a value of 69.4°C, and the numerical first derivative smoothing-filter method yields 68.5°C.

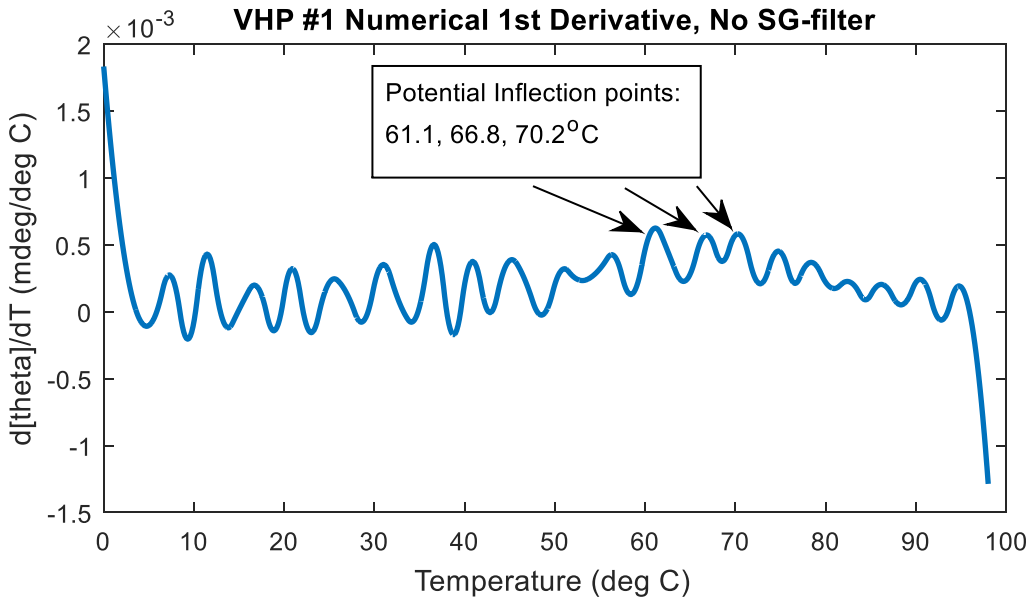


Figure 2.85. Local maxima of the first derivative of the dichroism signal as a function of temperature potential candidates for T_m values.

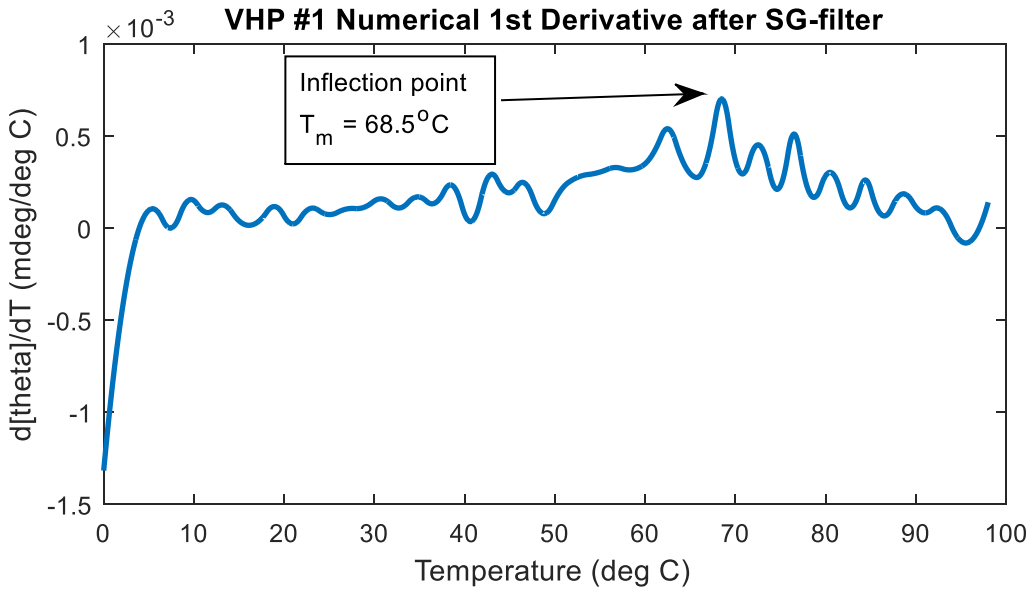


Figure 2.86. Application of a Savitzky-Golay smoothing filter to the numerical first derivative of the dichroism versus temperature yields a clear local maxima.

The work described here and in the subsequent chapter employs the baseline Levenberg-Marquardt method for two reasons. The first reason is that the enthalpy of folding, ΔH_m, which is a useful parameter for comparing the 'magnitude of cooperativity' among variants, is easily determined with the baseline procedure. The second reason is that the baseline definition is not arbitrary; the baseline is defined as the set of points that results in a minimum of the sum of the squares of the residual errors. This standardization minimizes the spread of the T_m values in a system where the signal changes linearly as a function of temperature in the pre- and post-transition region.

2.4.6 Propagating Uncertainty in ΔS_m

The uncertainty in ΔS_m is determined by first computing the fractional uncertainty, $\delta\Delta S_m/\Delta S_m$, by using the quotient rule for error propagation, and then scaling by ΔS_m:

$$\delta\Delta S_m = \Delta S_m \left(\frac{\delta\Delta H_m}{\Delta H_m} + \frac{\delta T_m}{T_m}\right) \qquad (2.19)$$

2.4.7 Crystallization Conditions

Each VHP quasiracemic crystal was grown using hanging drop vapor diffusion. Separate solutions of the β-substituted VHP and D-VHP were prepared by dissolving the lyophilized peptide powder in water to a final concentration of 5.8 mg/mL peptide, which was measured by UV absorbance using tryptophan as a chromophore.[60] The two separate solutions were then combined into a single quasiracemic solution. A 2 µL drop that comprised a 1:1 mixture of the quasiracemic solution and the optimized crystallization condition was placed on a glass cover slide that was then inverted to seal a well containing 250 µL of the crystallization condition. Buffers were made at 1 M concentration and then titrated to the appropriate pH with either 1 M HCl or 1 M NaOH. The final buffer concentration was achieved by performing a 1:10 dilution of the 1 M stock solution. PEG4000 solutions were made at 50% (w/v). Final crystallization solutions were filtered through a 0.2 µM syringe filter. Crystallization conditions are listed below:

VHP Q26 β^3Q quasiracemate: 0.1 M sodium acetate pH 4.6, 0.6 M ammonium acetate, 30% (w/v) PEG4000

VHP Q26ACPC quasiracemate: 0.1 M Tris-HCl pH 8.5, 0.1 M sodium acetate 30% (w/v) PEG4000

VHP K30β^3K quasiracemate: 0.1 M Tris-HCl pH 8.5, 20% (v/v) ethanol

VHP K30APC quasiracemate: 0.1 M sodium acetate pH 4.6, 1.5 M ammonium sulfate, 30% (v/v) PEG2000MME

2.4.8 X-ray Data Collection

Crystals were vitrified in liquid nitrogen after looping and pulling the crystal through the following cryoprotection conditions:

VHP Q26β^{3}Q quasiracemate: N/A, frozen directly

VHP Q26ACPC quasiracemate: 10% (v/v) glycerol in well solution

VHP K30β^{3}K quasiracemate: 30% (v/v) glycerol in well solution mixed with 3.5 M ammonium sulfate in a 1:1 ratio

VHP K30APC quasiracemate: 10% (v/v) glycerol in well solution

Diffraction data were collected at the Life Sciences Collaborative Access Team (LS-CAT) beamline 21-ID-F and 21-ID-G (where indicated) at the Advanced Photon Source (APS), Argonne National Laboratory. Datasets comprising a full 360º rotation were collected for each quasiracemate. Oscillation ranges are reported for each quasiracemate in Tables S1-S4. Data were subsequently indexed and integrated with the program XDS; data were then scaled and merged with the program XSCALE.[62]

2.4.9 Structure Solution and Refinement

Each quasiracemate crystallized in space group P1. Analysis of the cumulative intensity distribution for each dataset indicated the presence of a pseudo-inversion center. A molecular replacement solution was found with the program Phaser for each dataset using poly-alanine models of L-VHP (PDB: 3TRW) and D-VHP (PDB: 3TRY) as search models. Model refinement was carried out in the program phenix.refine in combination

with manual real-space model building and refinement in the program Coot (Crystallographic Object-Oriented Toolkit).[63,64] Simulated annealing and rigid body refinement were used for initial rounds of model refinement following manual model building in Coot. Anisotropic refinement of atomic displacement parameters (ADPs) was performed for VHP K30APC and Q26ACPC quasiracemates. TLS refinement was used for refinement of Q26β3Q, and K30β3K quasiracemate ADPs. Initially, a least-squares target function was employed for model refinement due to the nominally acentric reflections exhibiting centric intensity statistics; however, a maximum likelihood based target function with the conventional centric/acentric weighting scheme resulted in significantly lower R-values. Finally, for each quasiracemate structure the pseudo-inversion center, defined as the center of mass between pseudo-inversion pairs of chains, where each atom was weighted equally, was shifted to the origin for analysis of the phase angle distribution.

Restraints for β-amino acids were generated from the corresponding SMILES code with the program *phenix.elbow*.[65] Custom link restraints for β to α-amino acids were defined using standard peptide bond geometries (Figure 2.87). Restraints for D-amino acids were generated in *phenix.refine* by inverting the chiral center at C_α positions from the L-amino acid restraints to ensure bond length and angle restraints were applied consistently for both L and D-residues using the apply_cif_link setting in *phenix.refine*. Similarly, the chiral center was inverted at C_β for D-threonine restraints using the apply_cif_modification setting in *phenix.refine*. Riding hydrogen atoms were added at idealized position with the program *phenix.reduce*, and APC residues were modeled as

the pyrrolidine cation. Matthews coefficients were determined, using a peptide partial specific volume of 0.74 cm^3/g, as described by Kantardjieff and Rupp.[66]

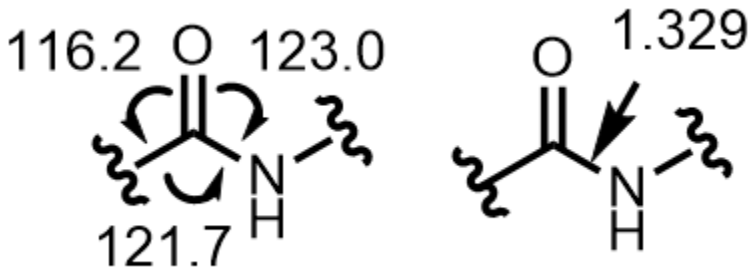


Figure 2.87. Peptide bond geometry restraints used for refinement of beta-amino acid linkages. Angles are listed in degrees (left) and distance is listed in Angstroms (right).

The custom cif modification linkage that was used in the monomer library is the following:

```
data_link_LINK_AB
#
loop_
_chem_link_bond.link_id
_chem_link_bond.atom_1_comp_id
_chem_link_bond.atom_id_1
_chem_link_bond.atom_2_comp_id
_chem_link_bond.atom_id_2
_chem_link_bond.type
_chem_link_bond.value_dist
_chem_link_bond.value_dist_esd
 LINK_AB        1       C       2       N       single        1.329       0.014
loop_
_chem_link_angle.link_id
_chem_link_angle.atom_1_comp_id
_chem_link_angle.atom_id_1
_chem_link_angle.atom_2_comp_id
_chem_link_angle.atom_id_2
_chem_link_angle.atom_3_comp_id
_chem_link_angle.atom_id_3
_chem_link_angle.value_angle
_chem_link_angle.value_angle_esd
```

```
 LINK_AB         1       O          1       C          2       N          123.000         1.600
 LINK_AB         1       CA         1       C          2       N          116.200         2.000
 LINK_AB         1       C          2       N          2       H          124.300         3.000
 LINK_AB         1       C          2       N          2       CA         121.700         1.800
loop_
_chem_link_plane.link_id
_chem_link_plane.plane_id
_chem_link_plane.atom_comp_id
_chem_link_plane.atom_id
_chem_link_plane.dist_esd
 LINK_AB                    plane1                  1                  CA                  0.01
 LINK_AB                    plane1                  1                  C                   0.01
 LINK_AB                    plane1                  1                  O                   0.01
 LINK_AB                    plane1                  2                  N                   0.01
 LINK_AB                    plane1                  2                  CA                  0.01
 LINK_AB      plane1    2 H         0.01
```

Table 2.5. VHP β^3-hGln26 quasiracemate (PDB: 5I1N)

Data collection	
X-ray source	APS 21-ID-F
X-ray detector	MAR225
Detector distance (mm)	120
Oscillation range (º)	1
Wavelength (Å)	0.97872
Space group	P1
a / b / c (Å)	36.4 / 38.9 / 48.4
α / β / γ (º)	101.1 / 110.5 / 99.4
Volume (Å^3)	60900
Matthews coefficient (Å^3/Da)	1.86
Solvent content (%)	34.0
Molecules in asymmetric unit cell	8
Resolution range (Å)	36.92-1.30 (1.35-1.30)
Number of observations	210950 (16738)
Unique Reflections	53932 (4352)
Completeness	93.0% (75.1%)
Redundancy	3.9 (3.8)
Mean *I* / σ(*I*)	8.7 (0.8)
CC1/2*	0.99 (0.38)
R_{merge}	0.078 (1.71)
R_{meas}	0.09 (1.98)
R_{pim}	0.046 (1.01)
Wilson *B*-factor (Å^2)	13.9
Refinement statistics	
Refinement program	phenix.refine: 1.9_1692
Non-hydrogen protein atoms refined	2334
Resolution range (Å)	36.9-1.30 (1.33-1.30)
No. of reflections used in refinement	53887
Completeness (%)	95.2 (70.0)
Reflections in cross-validation set	1997
R-value (work)	20.6
R-value (free)	23.1
R-value (overall)	20.7
Coordinate error (ML, Å)	0.19
Mean ADP (Å^2)	24.1
RMSD	
Bond lengths (Å)	0.013
Bond angles (°)	1.40

Table 2.6. VHP ACPC26 quasiracemate (PDB: 5I1O)

Data collection	
X-ray source	APS 21-ID-G
X-ray detector	MAR300
Detector distance (mm)	180
Oscillation range (º)	1
Wavelength (Å)	0.97856
Space group	P1
a / b / c (Å)	36.9 / 40.0 / 48.4
α / β / γ (º)	111.4 / 110.2 / 94.7
Volume ($Å^3$)	60700
Matthews coefficient ($Å^3$/Da)	1.86
Solvent content (%)	34.0
Molecules in asymmetric unit cell	8
Resolution range (Å)	18.39-1.35 (1.40-1.35)
Number of observations	173564 (4378)
Unique Reflections	47784 (3183)
Completeness	92.2% (61.3%)
Redundancy	3.6 (1.4)
Mean I / σ(I)	11.8 (0.9)
CC1/2*	0.99 (0.61)
R_{merge}	0.055 (0.646)
R_{meas}	0.064 (0.912)
R_{pim}	0.033 (0.643)
Wilson B-factor ($Å^2$)	12.2
Refinement statistics	
Refinement program	phenix.refine: 1.9_1692
Non-hydrogen protein atoms refined	2295
Resolution range (Å)	18.39-1.35 (1.38-1.35)
No. of reflections used in refinement	47748
Completeness (%)	92.1 (56.0)
Reflections in cross-validation set	2018
R-value (work)	17.3
R-value (free)	20.4
R-value (overall)	17.5
Coordinate error (ML, Å)	0.16
Mean ADP ($Å^2$)	18.8
RMSD	
Bond lengths (Å)	0.01
Bond angles (°)	1.00

Table 2.7. VHP β^3-hLys30 quasiracemate (PDB: 5I1P)

Data collection	
X-ray source	APS 21-ID-F
X-ray detector	MAR225
Detector distance (mm)	120
Oscillation range (º)	1
Wavelength (Å)	0.97872
Space group	*P*1
a / b / c (Å)	24.6 / 40.6 / 59.0
α / β / γ (º)	96.5 / 90.7 / 95.4
Volume ($Å^3$)	58300
Matthews coefficient ($Å^3$/Da)	1.78
Solvent content (%)	31.0
Molecules in asymmetric unit cell	8
Resolution range (Å)	40.16-1.40 (1.45-1.40)
Number of observations	167264 (16735)
Unique Reflections	42610 (4221)
Completeness	95.8% (93.4%)
Redundancy	3.9 (4.0)
Mean I / σ(I)	9.3 (1.1)
CC1/2*	0.99 (0.59)
R_{merge}	0.067 (1.35)
R_{meas}	0.078 (1.56)
R_{pim}	0.039 (0.78)
Wilson B-factor ($Å^2$)	15.4
Refinement statistics	
Refinement program	phenix.refine: 1.9_1692
Non-hydrogen protein atoms refined	2584
Resolution range (Å)	40.16-1.40 (1.44-1.40)
No. of reflections used in refinement	42543
Completeness (%)	95.7 (93.0)
Reflections in cross-validation set	1994
R-value (work)	21.0
R-value (free)	23.5
R-value (overall)	21.1
Coordinate error (ML, Å)	0.18
Mean ADP ($Å^2$)	28.1
RMSD	
Bond lengths (Å)	0.014
Bond angles (°)	1.51

Table 2.8. VHP APC30 quasiracemate (PDB: 5I1S)

Data collection	
X-ray source	APS 21-ID-G
X-ray detector	MAR300
Detector distance (mm)	130
Oscillation range (º)	1.5
Wavelength (Å)	0.97856
Space group	P1
a / b / c (Å)	24.0 / 34.3 / 42.9
α / β / γ (º)	106.2 / 108.6 / 98.5
Volume ($Å^3$)	31000
Matthews coefficient ($Å^3$/Da)	1.90
Solvent content (%)	35.5
Molecules in asymmetric unit cell	4
Resolution range (Å)	38.09-1.12 (1.16-1.12)
Number of observations	155128 (4548)
Unique Reflections	42168 (2902)
Completeness	90.8% (62.5%)
Redundancy	3.7 (1.6)
Mean I / σ(I)	13.2 (1.5)
CC1/2*	0.99 (0.88)
R_{merge}	0.045 (0.390)
R_{meas}	0.052 (0.550)
R_{pim}	0.027 (0.380)
Wilson B-factor ($Å^2$)	9.7
Refinement statistics	
Refinement program	phenix.refine: 1.9_1692
Non-hydrogen protein atoms refined	1520
Resolution range (Å)	38.09-1.12 (1.15-1.12)
No. of reflections used in refinement	42095
Completeness (%)	90.6 (56.0)
Reflections in cross-validation set	2108
R-value (work)	15.5
R-value (free)	18.6
R-value (overall)	15.7
Coordinate error (ML, Å)	0.10
Mean ADP ($Å^2$)	15.6
RMSD	
Bond lengths (Å)	0.013
Bond angles (°)	1.54

2.4.10 Quasiracemate Diffraction Images

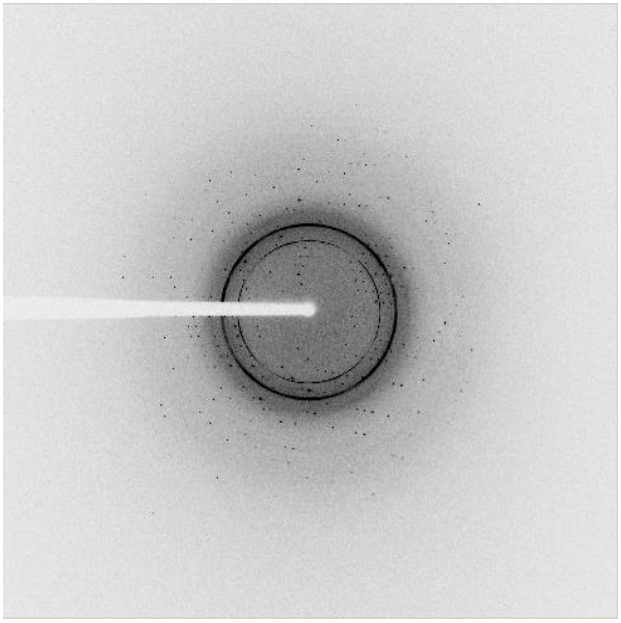


Figure 2.88. VHP β^3-hGln26 quasiracemate diffraction

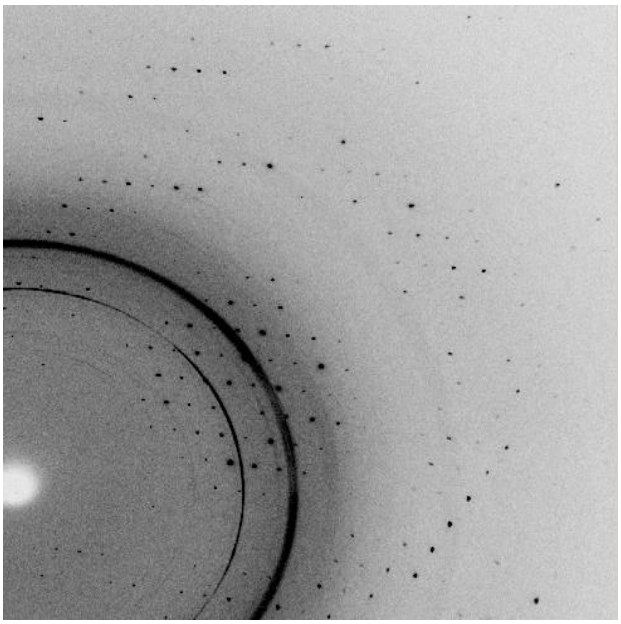


Figure 2.89. VHP β^3-hGln26 quasiracemate diffraction.

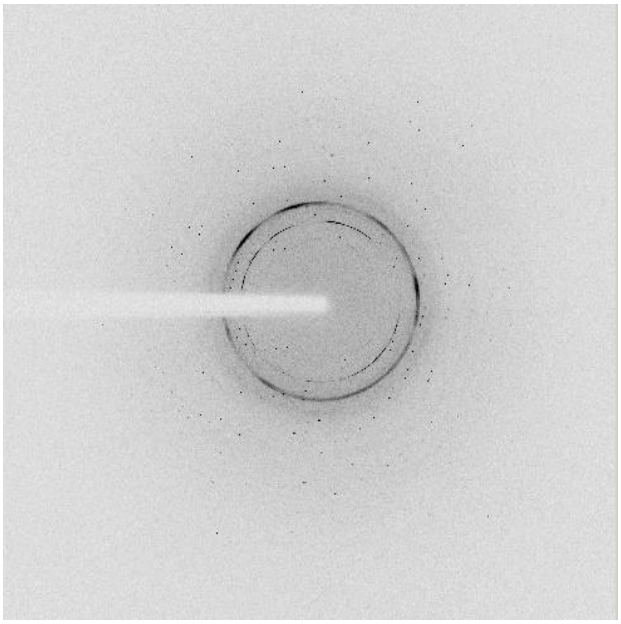


Figure 2.90. VHP 26ACPC quasiracemate diffraction.

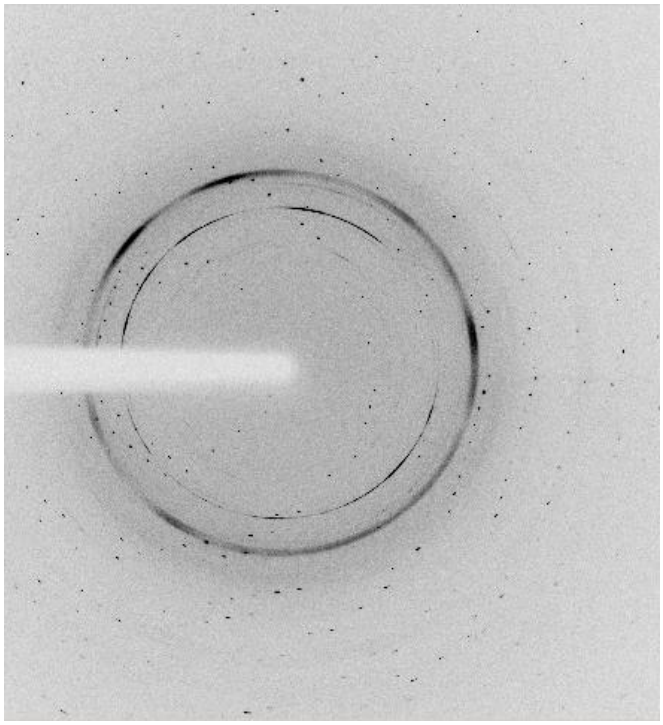


Figure 2.91. VHP 26ACPC quasiracemate diffraction.

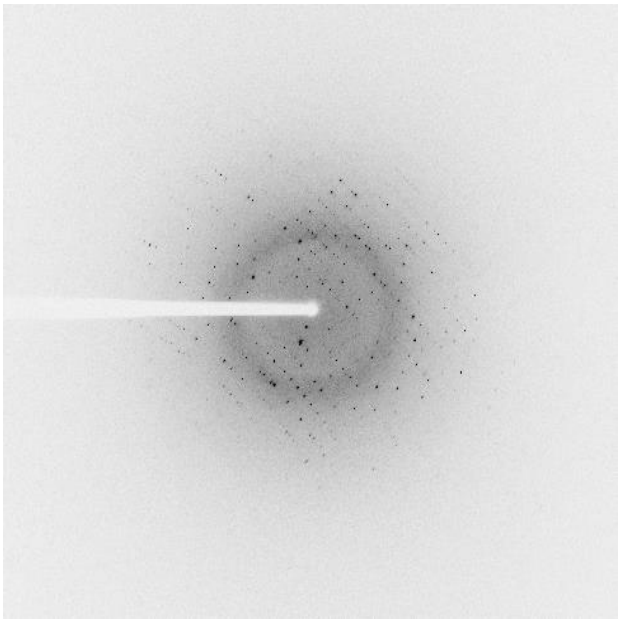


Figure 2.92. VHP β^3-hLys30 quasiracemate diffraction.

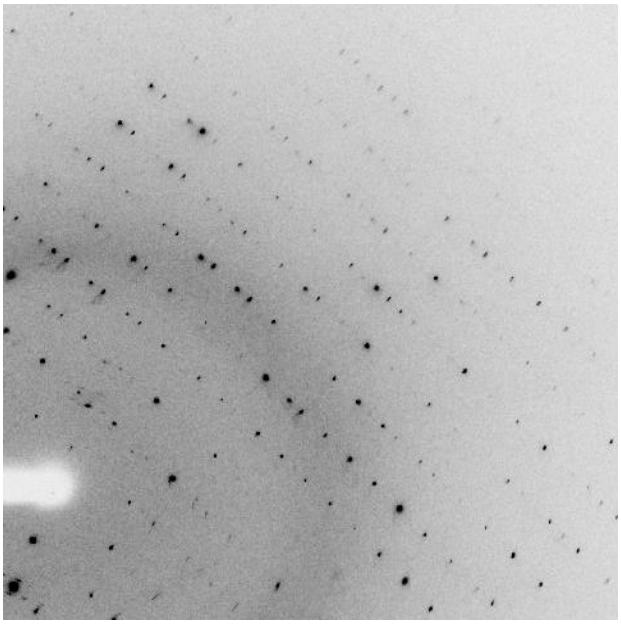


Figure 2.93. VHP β^3-hLys30 quasiracemate diffraction.

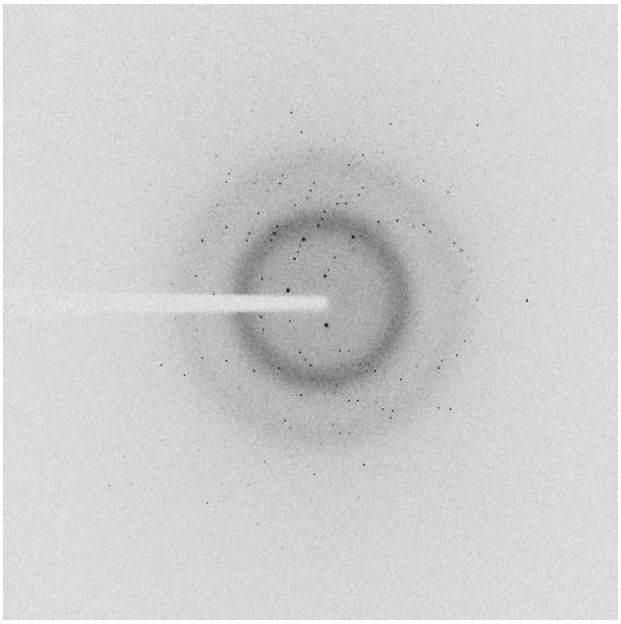


Figure 2.94. VHP 30APC quasiracemate diffraction

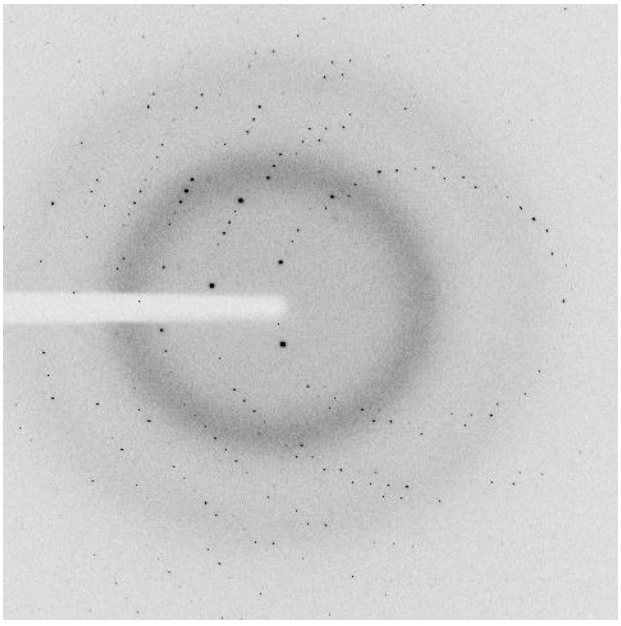


Figure 2.95. VHP 30APC quasiracemate diffraction.

2.5 References

(1) Maini, R.; Chowdhury, S. R.; Dedkova, L. M.; Roy, B.; Daskalova, S. M.; Paul, R.; Chen, S.; Hecht, S. M. *Biochemistry* **2015**, 150602071154009.

(2) Fujino, T.; Goto, Y.; Suga, H.; Murakami, H. *J. Am. Chem. Soc.* **2016**, *138* (6), 1962–1969.

(3) McKnight, C. J.; Matsudaira, P. T.; Kim, P. S. *Nat. Struct. Biol.* **1997**, *4* (3), 180–184.

(4) Chiu, T. K.; Kubelka, J.; Herbst-Irmer, R.; Eaton, W. A; Hofrichter, J.; Davies, D. R. *Proc. Natl. Acad. Sci. U. S. A.* **2005**, *102* (21), 7517–7522.

(5) Žoldák, G.; Stigler, J.; Pelz, B.; Li, H.; Rief, M. *Proc. Natl. Acad. Sci. U. S. A.* **2013**, *110* (45), 18156–18161.

(6) Johnson, L. M.; Gellman, S. H. *Methods Enzymol.* **2013**, *523*, 407–429.

(7) Horne, W. S.; Gellman, S. H. *Acc. Chem. Res.* **2008**, *41* (10), 1399–1408.

(8) Goodman, C. M.; Choi, S.; Shandler, S.; DeGrado, W. F. *Nat. Chem. Biol.* **2007**, *3* (5), 252–262.

(9) Guichard, G.; Huc, I. *Chem. Commun. (Camb).* **2011**, *47* (21), 5933–5941.

(10) Aguilar, M.-I.; Purcell, A. W.; Devi, R.; Lew, R.; Rossjohn, J.; Smith, a I.; Perlmutter, P. *Org. Biomol. Chem.* **2007**, *5* (18), 2884–2890.

(11) Schievano, E.; Mammi, S.; Carretta, E.; Fiori, N.; Corich, M.; Bisello, A.; Rosenblatt, M.; Chorev, M.; Peggion, E. *Biopolymers* **2003**, *70* (4), 534–547.

(12) Denton, E. V; Craig, C. J.; Pongratz, R. L.; Appelbaum, J. S.; Doerner, A. E.; Narayanan, A.; Shulman, G. I.; Cline, G. W.; Schepartz, A. *Org. Lett.* **2013**, *15* (20), 5318–5321.

(13) Boersma, M. D.; Haase, H. S.; Peterson-Kaufman, K. J.; Lee, E. F.; Clarke, O. B.; Colman, P. M.; Smith, B. J.; Horne, W. S.; Fairlie, W. D.; Gellman, S. H. *J. Am. Chem. Soc.* **2012**, *134* (1), 315–323.

(14) Cheloha, R. W.; Maeda, A.; Dean, T.; Gardella, T. J.; Gellman, S. H. *Nat. Biotechnol.* **2014**, *32* (7), 653–655.

(15) Hager, M. V.; Johnson, L. M.; Wootten, D.; Sexton, P. M.; Gellman, S. H. *J. Am. Chem. Soc.* **2016**, *138* (45), 14970–14979.

(16) Lengyel, G. a.; Frank, R. C.; Horne, W. S. *J. Am. Chem. Soc.* **2011**, *133* (12), 4246–4249.

(17) Lengyel, G. A.; Horne, W. S. *J. Am. Chem. Soc.* **2012**, *134*, 15906–15913.

(18) Hegedüs, Z.; Wéber, E.; Kriston-Pál, É.; Makra, I.; Czibula, Á.; Monostori, É.; Martinek, T. a. *J. Am. Chem. Soc.* **2013**, *135* (44), 16578–16584.

(19) Arnold, U.; Hinderaker, M. P.; Nilsson, B. L.; Huck, B. R.; Gellman, S. H.; Raines, R. T. *J. Am. Chem. Soc.* **2002**, *124* (29), 8522–8523.

(20) David, R.; Günther, R.; Baumann, L.; Lühmann, T.; Seebach, D.; Hofmann, H. J.; Beck-Sickinger, A. G. *J. Am. Chem. Soc.* **2008**, *130* (46), 15311–15317.

(21) Peggion, E.; Mammi, S.; Schievano, E.; Silvestri, L.; Schiebler, L.; Bisello, A.; Rosenblatt, M.; Chorev, M. *Biochemistry* **2002**, *41* (25), 8162–8175.

(22) Reinert, Z. E.; Lengyel, G. a.; Horne, W. S. *J. Am. Chem. Soc.* **2013**, *135* (34), 12528–12531.

(23) Reinert, Z. E.; Horne, W. S. *Chem. Sci.* **2014**, *5* (8), 3325–3330.

(24) Mayer, C.; Müller, M. M.; Gellman, S. H.; Hilvert, D. *Angew. Chemie - Int. Ed.* **2014**, *53* (27), 6978–6981.

(25) Checco, J. W.; Kreitler, D. F.; Thomas, N. C.; Belair, D. G.; Rettko, N. J.; Murphy, W. L.; Forest, K. T.; Gellman, S. H. *Proc. Natl. Acad. Sci.* **2015**, *112* (15), 4552–4557.

(26) Checco, J. W.; Gellman, S. H. *ChemBioChem* **2017**, *18* (3), 291–299.

(27) Kubelka, J.; Eaton, W. A.; Hofrichter, J. *J. Mol. Biol.* **2003**, *2836* (3), 625–630.

(28) Horne, W. S.; Price, J. L.; Gellman, S. H. *Proc. Natl. Acad. Sci. U. S. A.* **2008**, *105* (27), 9151–9156.

(29) Kubelka, J.; Chiu, T. K.; Davies, D. R.; Eaton, W. A.; Hofrichter, J. *J. Mol. Biol.* **2006**, *359*, 546–553.

(30) Greenfield, N. J. *Nat. Protoc.* **2007**, *1* (6), 2527–2535.

(31) Godoy-Ruiz, R.; Henry, E. R.; Kubelka, J.; Hofrichter, J.; Muñoz, V.; Sanchez-Ruiz, J. M.; Eaton, W. A. *J. Phys. Chem. B* **2008**, *112* (19), 5938–5949.

(32) Spolar, R. S.; Livingstone, J. R.; Record, M. T. *Biochemistry* **1992**, *31* (16), 3947–3955.

(33) Hilser, V. J.; Gómez, J.; Freire, E. *Proteins Struct. Funct. Genet.* **1996**, *26* (2), 123–133.

(34) Makhatadze, G. I.; Privalov, P. L. *Journal of molecular biology*. 1993, pp 639–659.

(35) Oobatake, M.; Ooi, T. *Prog Biophys Mol Biol* **1993**, *59* (3), 237–284.

(36) Becktel, W. J.; Schellman, J. A. *Biopolymers* **1987**, *26*, 1859–1877.

(37) Fitter, J. *Biophys. J.* **2003**, *84* (June), 3924–3930.

(38) Lee, K. H.; Xie, D.; Freire, E.; Amzel, L. M. *Proteins Struct. Funct. Bioinforma.* **1994**, *20* (1), 68–84.

(39) Aurora, R.; Rose, G. D. *Protein Sci.* **1998**, *7* (1), 21–38.

(40) Zou, J.; Song, B.; Simmerling, C.; Raleigh, D. *J. Am. Chem. Soc.* **2016**, *138*, 15682–15689.

(41) Anil, B.; Song, B.; Tang, Y.; Raleigh, D. P. *J. Am. Chem. Soc.* **2004**, *126* (41), 13194–13195.

(42) Wukovitz, S. W.; Yeates, T. O. *Nat. Struct. Biol.* **1995**, *2* (12), 1062–1067.

(43) Yeates, T. O.; Kent, S. B. H. *Annu. Rev. Biophys.* **2012**, *41*, 41–61.

(44) Pentelute, B. L.; Gates, Z. P.; Tereshko, V.; Dashnau, J. L.; Vanderkooi, J. M.; Kossiakoff, A. A.; Kent, S. B. H. *J. Am. Chem. Soc.* **2008**, *130* (30), 9695–9701.

(45) Mortenson, D. E.; Satyshur, K. A; Guzei, I. A; Forest, K. T.; Gellman, S. H. *J. Am. Chem. Soc.* **2012**, *134* (5), 2473–2476.

(46) Pentelute, B. L.; Gates, Z. P.; Dashnau, J. L.; Vanderkooi, J. M.; Kent, S. B. H. *J. Am. Chem. Soc.* **2008**, *130* (30), 9702–9707.

(47) Mandal, K.; Pentelute, B. L.; Bang, D.; Gates, Z. P.; Torbeev, V. Y.; Kent, S. B. H. *Angew. Chemie - Int. Ed.* **2012**, *51* (6), 1481–1486.

(48) Okamoto, R.; Mandal, K.; Sawaya, M. R.; Kajihara, Y.; Yeates, T. O.; Kent, S. B. *Angew Chem Int Ed Engl* **2014**, 5194–5198.

(49) Srinivasan, R.; Parthasarathy, S. *Some Statistical Applications in X-Ray Crystallography*; Pergamon Press: Elmsford, New York, USA, 1976.

(50) Srinivasan, R.; Swaminathan, P. *Curr. Sci.* **1975**, *44* (3), 71–73.

(51) Desiraju, G. R.; Calabrese, J. C.; Harlow, R. L. *Acta Crystallogr. Sect. B Struct. Sci.* **1991**, *341* (2), 77–86.

(52) Janssen, T.; Janner, A.; Looijenga-Vos, A.; de Wolff, P. M. In *International Tables for Crystallography*; Wilson, A. J. C., Prince, E., Eds.; Kluwer Academic Publishers: Dordrecht, Boston, London, 2006; Vol. C, pp 907–955.

(53) Lovelace, J. J.; Murphy, C. R.; Daniels, L.; Narayan, K.; Schutt, C. E.; Lindberg, U.; Svensson, C.; Borgstahl, G. E. O. *J. Appl. Crystallogr.* **2008**, *41* (3), 600–605.

(54) Luque, I.; Mayorga, O. L.; Freire, E. *Biochemistry* **1996**, *35* (42), 13681–13688.

(55) Horne, W. S.; Price, J. L.; Keck, J. L.; Gellman, S. H. *J. Am. Chem. Soc.* **2007**, *129* (14), 4178–4180.

(56) Horne, W. S.; Johnson, L. M.; Ketas, T. J.; Klasse, P. J.; Lu, M.; Moore, J. P.; Gellman, S. H. *Proc. Natl. Acad. Sci. U. S. A.* **2009**, *106* (35), 14751–14756.

(57) Haase, H. S.; Peterson-Kaufman, K. J.; Levengood, S. K. L.; Checco, J. W.; Murphy, W. L.; Gellman, S. H. *J. Am. Chem. Soc.* **2012**, *134* (18), 1–24.

(58) Peterson-Kaufman, K. J.; Haase, H. S.; Boersma, M. D.; Lee, E. F.; Fairlie, W. D.; Gellman, S. H. *ACS Chem. Biol.* **2015**, *10*, 1667–1675.

(59) Lee, H.; Leplae, P. R.; Porter, E. A.; Gellman, S. H. *J. Org. Chem.* **2001**, *66* (9), 3597–3599.

(60) Edelhoch, H. *Biochemistry* **1967**, *6* (7), 1948–1954.

(61) Greenfield, N. J. *Nat. Protoc.* **2006**, *1* (6), 2876–2890.

(62) Kabsch, W. *Acta Crystallogr. Sect. D Biol. Crystallogr.* **2010**, *66* (2), 125–132.

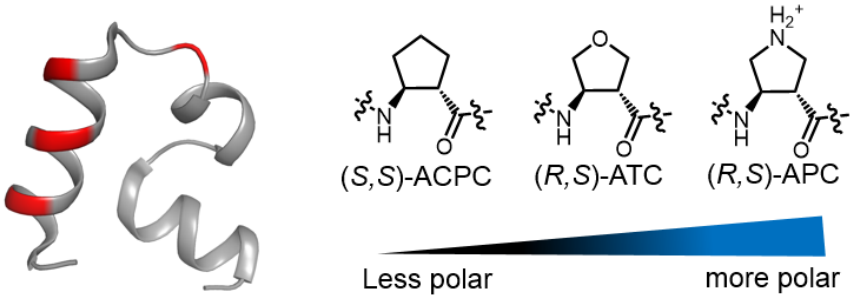
(63) Adams, P. D.; Afonine, P. V.; Bunkóczi, G.; Chen, V. B.; Davis, I. W.; Echols, N.; Headd, J. J.; Hung, L. W.; Kapral, G. J.; Grosse-Kunstleve, R. W.; McCoy, A. J.; Moriarty, N. W.; Oeffner, R.; Read, R. J.; Richardson, D. C.; Richardson, J. S.; Terwilliger, T. C.; Zwart, P. H. *Acta Crystallogr. Sect. D Biol. Crystallogr.* **2010**, *66* (2), 213–221.

(64) Emsley, P.; Cowtan, K. *Acta Crystallogr. Sect. D Biol. Crystallogr.* **2004**, *60* (12 I), 2126–2132.

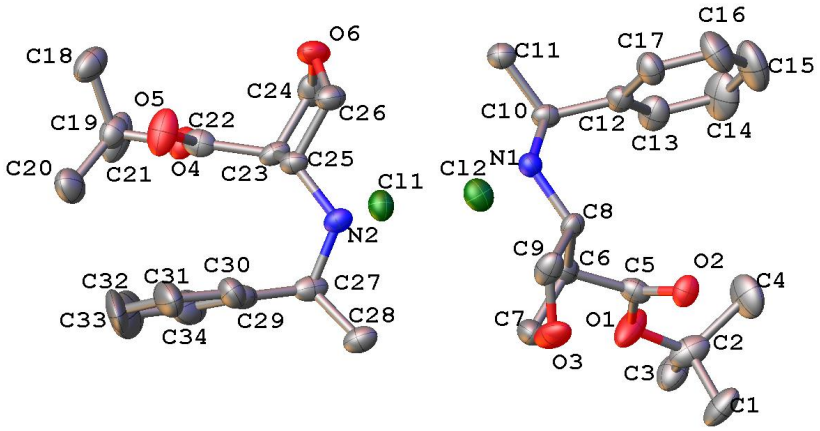
(65) Moriarty, N. W.; Grosse-Kunstleve, R. W.; Adams, P. D. *Acta Crystallogr. Sect. D Biol. Crystallogr.* **2009**, *65* (10), 1074–1080.

(66) Kantardjieff, K. A.; Rupp, B. *Protein Sci.* **2003**, *12* (9), 1865–1871.

Chapter 3 Exploring the Effects of Charge and Polarity on the Impact of Cyclic β Amino Acid Substitutions in a Mini-Protein



H-LSDEDFKAVFGMTRSAFA**N**LPL**W**KQ**Q**HLK**K**EKGLF-COOH



3.1 Introduction

In recent decades protein-based therapeutics have emerged as viable options for the treatment of diseases that have been outside the purview of traditional pharmaceutical intervention.[1] Specifically, monoclonal antibodies (mAbs) tailored to target specific pathogenic proteins have been developed to treat diseases caused by immune system dysfunction, *e.g.,* rheumatoid arthritis, Crohn's disease, multiple sclerosis, or cancers that are refractory to conventional chemotherapies.[1,2] For instance, Ipilimumab, developed by Bristol-Myers Squibb, targets a protein receptor, CTLA-4, that normally attenuates the immune system; blocking CTLA-4 function allows cytotoxic T lymphocytes (CTLs) to recognize and destroy cancer cells.[2] While promising, these antibody therapeutics possess some significant drawbacks, mainly immunogenicity and formulation stability.[3–5] Furthermore, the extraordinary capital investment required to deliver safe and efficacious antibody therapies has engendered ethical debates concerning such treatments.[1] Small peptide affibodies, small proteins engineered to mimic monoclonal antibodies, present an alternative class of compound that addresses some of these drawbacks while retaining the attractive features of mAb-based therapeutics, mainly the ability to target large, irregular protein interfaces.[6,7] Moreover, these peptide affibody scaffolds are accessible via chemical synthesis, which allows for unnatural chemical modifications and the generation of combinatorial libraries.

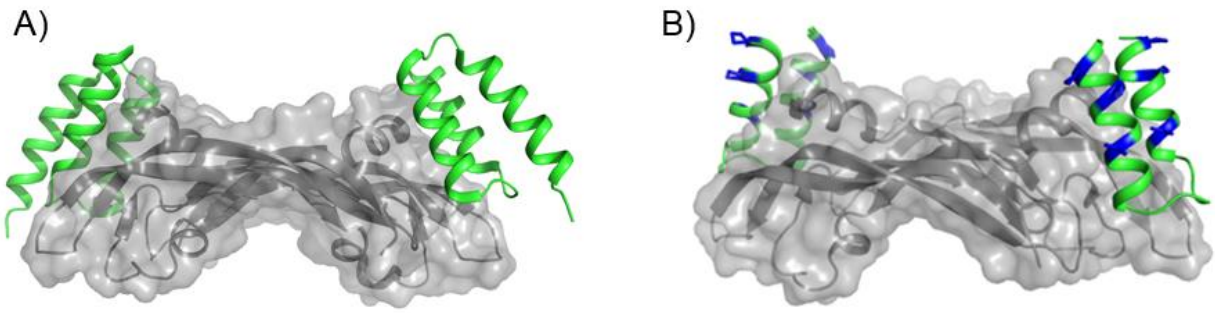


Figure 3.1. A) PDB: 3S1K, a three-helix peptide affibody, Z-domain, (green) developed by Fedorova et al. in complex with vascular endothelial growth factor (VEGF, gray).[6] B) The truncated Z-domain (green) reported by Checco et al. contains unnatural residues (blue) and is pre-organized by a disulfide bond between terminal residues (PDB: 4WPB).[7]

In 2010, Fedorova *et al.* reported a peptide affibody that targeted vascular endothelial growth factor (VEGF).[6] This three-helix 'Z-domain' was obtained via phage display, in which a peptide scaffold, derived from an *S. aureous* albumin-binding domain, underwent successive rounds of selection. Fedorova *et al.* also employed a truncation strategy, developed earlier by Starovasnik *et al.*, where the third helix along the 'back-face' of the peptide Z-domain was removed, yielding a more tractable helix-turn-helix, hairpin motif that could be stabilized by a disulfide bond between the N and C-termini.[6,8] These peptide affibodies, however, are susceptible to degradation by endogenous proteases. There are a multitude of chemical strategies to overcome this natural degradation process which include: PEGylation, N-terminal acetylation, backbone cyclization, or unnatural residue incorporation (β amino acids).[9–11] Checco *et al.* demonstrated that protease resistance of the Fedorova Z-domain could be enhanced, and the natural affinity for VEGF left largely intact, through the systematic incorporation of β and α,α-disubstituted amino acids.[7]

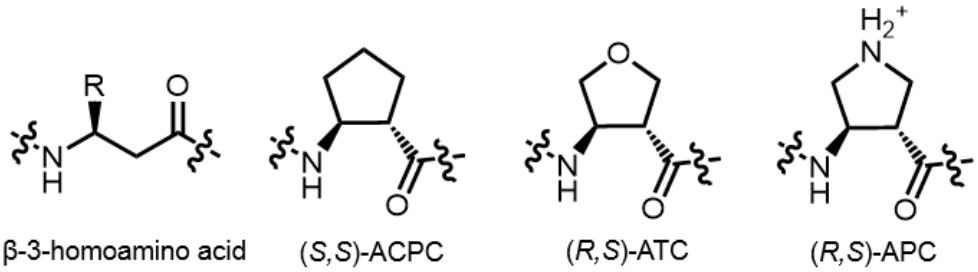


Figure 3.2. β-3-homoamino acids contain an additional flexible, methylene unit in the backbone. The charge and polarity of cyclically constrained β amino acids can be modulated with a ring heteroatom.

The 'sequence-based' β amino acid incorporation strategy has been developed primarily in α-helical systems, where flexible β-3-homoamino acid residues and conformationally pre-organized cyclic β residues have been employed.[12] These cyclic residues have predominantly been derived from either (3*S*,4*S*)-aminocyclopentane carboxylic acid (ACPC) or (3*R*,4*S*)-aminopyrrolidine carboxylic (APC), *i.e.*, hydrophobic or charged (pyrrolidinium pK_a ~11.3). Fundamental structural studies have shown that these five-membered rings promote α-helix-like secondary structure.[13,14] Recent evidence, however, has shown that within model tertiary structures this conformational constraint does not necessarily translate to increased thermodynamic stability and may actually be detrimental to overall stability.[15] In peptide affibody scaffolds, thermodynamic stability is likely to be correlated with binding efficacy.[16] Therefore, a deeper understanding of the impact of cyclic β residues on the stability of peptide tertiary structures will complement empirical efforts that employ the β residue incorporation strategy.

An approximation, implicit in the cyclic β incorporation strategy, is that the helix-promoting effects of ACPC versus APC are equivalent. However, this approximation may

not hold in small tertiary structures on the razor's edge of thermodynamic stability. For instance, slight variations in the enthalpic cost of burying polar cyclic β surface area in a folded helical backbone could disrupt the overall folding stability. Examining the VHP Q26ACPC structure from Chapter 2 reveals that approximately 70% of the solvent accessible surface area (SASA) of the ACPC substitution is buried in the folded VHP native state relative to an ACPC-containing tripeptide with an extended backbone, even though position 26 is 'solvent-exposed' (Figure 3.3). Structural and biophysical evidence presented in Chapter 2 suggests that hydrophobic (ACPC) or charged (APC) cyclic β substitutions may have a variable impact on the folded stability. An APC substitution at Lys30 in the Villin headpiece was the only tested β substitution that retained the native stability of the VHP, ACPC residues were generally destabilizing.

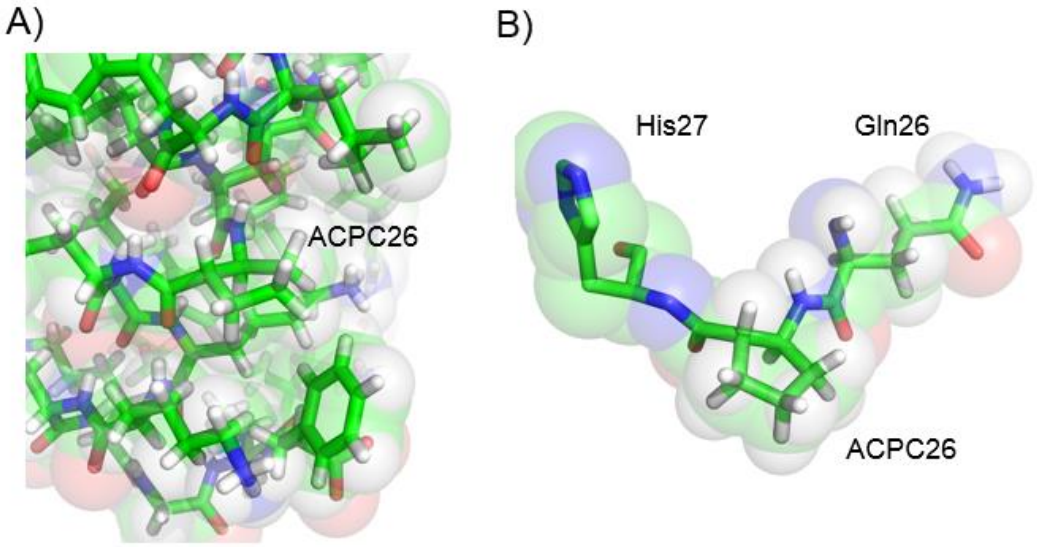


Figure 3.3. A) ACPC has a solvent accessible surface area of 47.9 Å^2 in the folded conformation (PDB: 5I1O) and B) 162.7 Å^2 in an extended tripeptide (His-ACPC-Gln). The solvent accessible surface area was calculated with the program VMD using a 1.4 Å probe radius. The model tripeptide was generated in phenix.elbow with an AM1 minimization engine.[17]

To test the hypothesis that polar/charged cyclic β residues are detrimental to stability relative to hydrophobic cyclic β residues, two-state folding parameters were determined for a series of β substituted VHP derivatives with cyclic β residues of varying charge and polarity (Figure 3.4). These new data complement the dataset reported in Chapter 2. In addition to the more commonly utilized cyclic β residues, ACPC and APC, I employed a heterocyclic β residue of intermediate polarity that has not been previously placed in a tertiary structure. This polar cyclic β residue contains an oxygen atom at the ring heteroatom position, and is termed ATC (*trans*-(3*R*,4*S*)-aminotetrahydrofuran carboxylic acid). Therefore the complete dataset comprises VHP with ACPC, APC, and ATC substitutions at positions Asn19, Trp23, Gln26, and Lys30 (Figure 3.4).

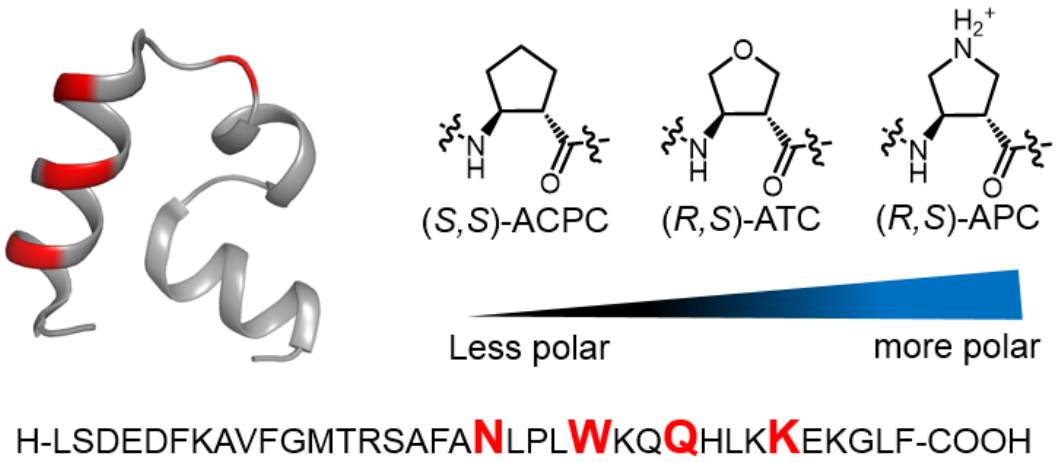
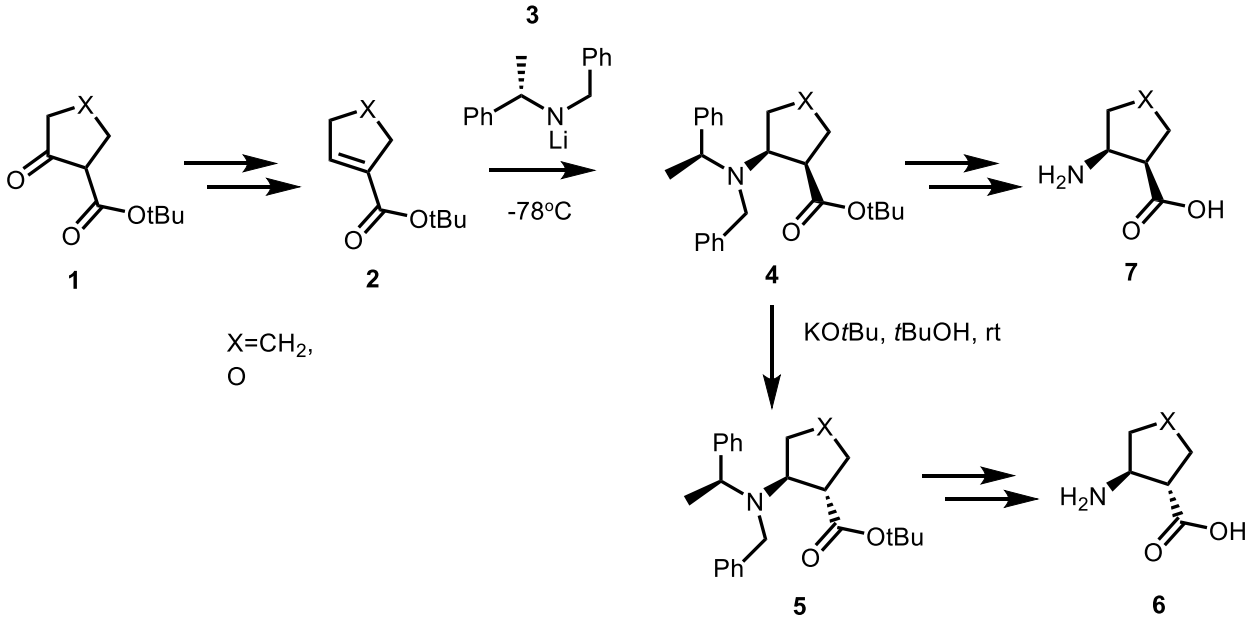


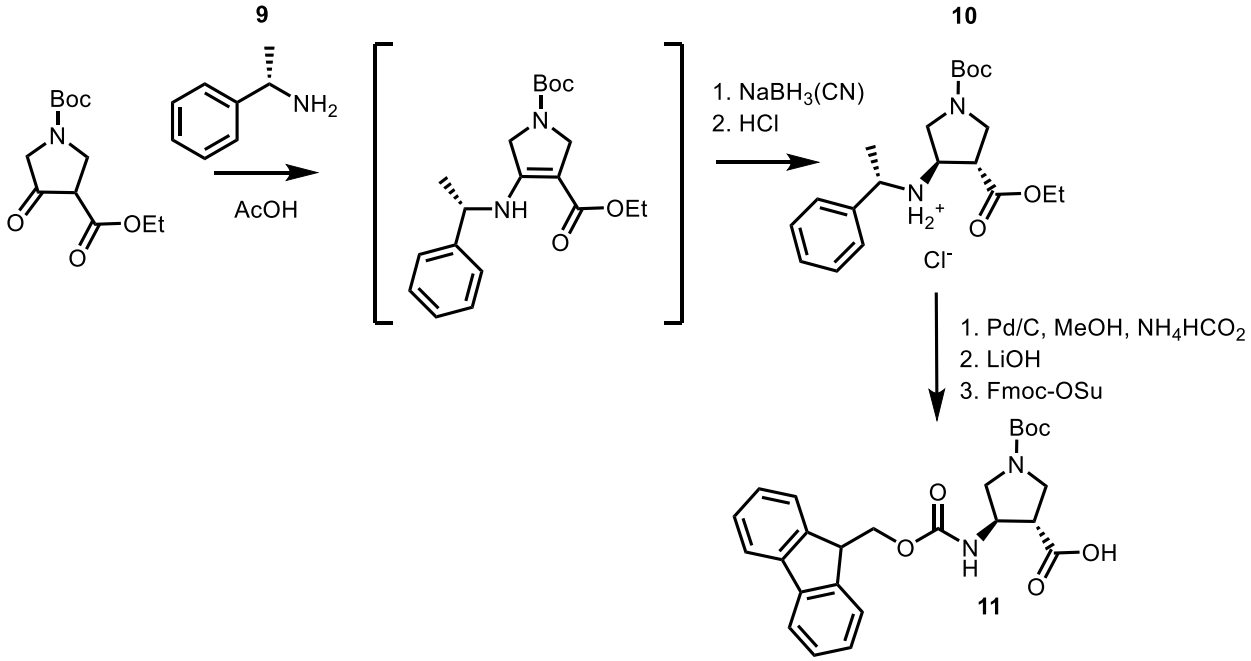
Figure 3.4. ACPC, ATC, and APC substitutions were made at VHP positions Asn19, Trp23, Gln26, and Lys30 (red).

These studies required a synthetic strategy for obtaining sufficient quantities of the Fmoc-protected ATC monomer for solid phase peptide synthesis. Towards this end, Geoff Eddinger and I followed a procedure previously described by Imamura *et al.* This procedure combines a route for preparing the racemic ATC β-ketoester, described by Bunnage *et al.*, and a strategy for the stereoselective reductive amination of heterocyclic

β-ketoesters, described by Lee *et al.*, to obtain Fmoc-*trans*-(3*S*,4*R*)-(ATC)-OH.[18,19] Here we established the absolute configuration of the β-aminoketoester intermediate. Bunnage *et al.* reported procedures for both *trans* and *cis*-ATC/ACPC via a stereoselective conjugate addition of lithium amide **3** to α,β-unsaturated ester **2** (Scheme 3.1). Presumably, this conjugate addition occurs through a 'butterfly' type transition state. Various routes for the preparation of carbocyclic and heterocyclic β amino acids have been reviewed extensively by Kiss and Fülöp.[20]



Scheme 3.1. The synthetic route reported by Bunnage et al. for cis and trans-(3R,4S)-aminotetrahydrofuran carboxylic acid (X=O), or cis and trans-(3S,4S)-aminocyclopentane carboxylic acid ($X=CH_2$).[18]



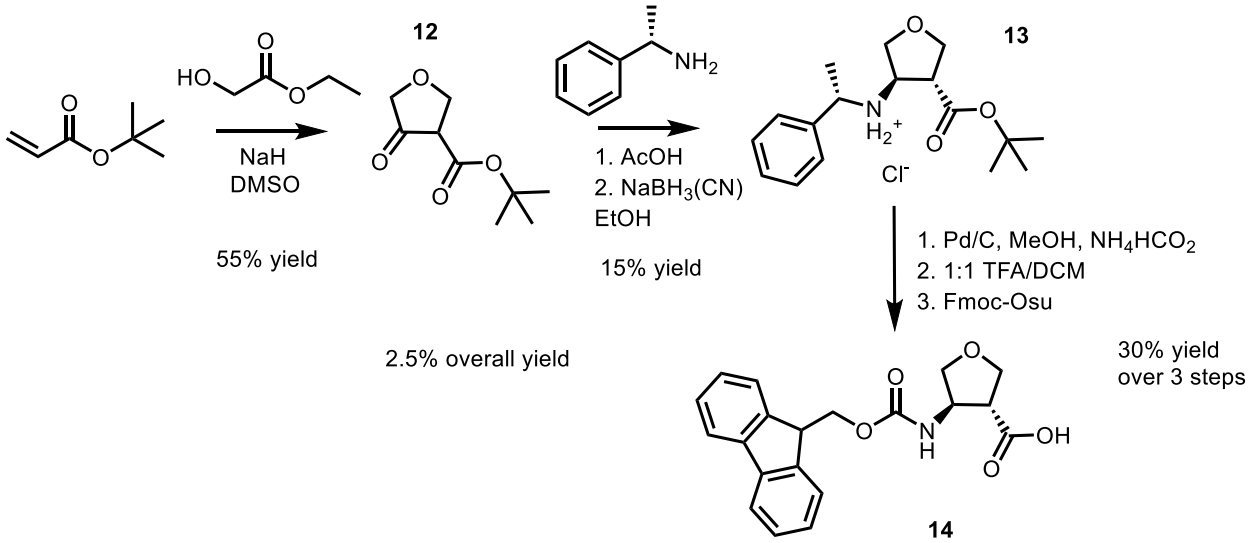
Scheme 3.2. The route reported by Lee et al. for the preparation of orthogonally protected Fmoc-(3R,4S)-trans-aminopyrrolidine carboxylic acid.[19]

3.2 Results and Discussion

3.2.1 The Asymmetric Synthesis of Fmoc-(3R,4S)-trans-aminotetrahydrofuran Carboxylic Acid

Although a route for the preparation of *trans*-(3*R*,4*S*)-aminotetrahydrofuran carboxylic acid (**6**) had been reported previously by Bunnage *et al.*, our aim was to employ a more robust procedure that avoids the use of hazardous organolithium reagents and is scalable for subsequent incorporation into synthetic peptide/protein libraries via solid phase synthesis. We combined procedures from Bunnage *et al.* (Scheme 3.1) and Lee *et al.* (Scheme 3.3) to generate Scheme 3.3. β ketoester **12**, prepared as described by Bunnage *et al.*, undergoes reductive amination with (*S*)-(-)-methylbenzylamine as a chiral auxiliary to afford the enantiomerically pure hydrochloride ammonium salt of β aminoester **13** after recrystallization. Imamura *et al.* have previously reported almost

the exact same procedure for Fmoc-*trans*-ATC-OH (unbeknownst to us during the development of our strategy), which differs from that of Scheme 3.3 only in the order of protecting group manipulations in the final three steps.[21] Here we established the absolute configuration of β aminoester **13** with single crystal X-ray crystallography (Appendix). Finally, through a series of protecting group manipulations we were able to obtain sufficient quantities (~1.0 g) of the Fmoc-protected ATC monomer **14** for the preparation of peptides via solid phase synthesis.



Scheme 3.3. Synthetic route employed in this work for Fmoc-trans-(3R,4S)-aminotetrahydrofuran carboxylic acid (**14**). The route is essentially the same as previously reported by Imamura et al. (steps 1 and 2 are reversed here in the final protecting group manipulations), which is similar in concept to the route described by Lee et al. after formation of the β-ketoester.

3.2.2 Circular Dichroism of Cyclic β Substituted VHP Variants

Each VHP variant displayed a CD wavelength spectrum consistent with α-helical secondary structure, *i.e.*, minima at ~208 and ~222 nm. The impact of each substitution on the normalized signal amplitude seemed to vary with position (Figure 3.5). At Asn19 and Gln26 the polar/charged (ATC/APC) substitutions resulted in a slight decrease in

absolute ellipticity relative to the hydrophobic ACPC variants (Figure 3.5A,C). In contrast, polar/charged substitutions at Trp23 and Lys30 caused an increase in ellipticity relative to the ACPC variants (Figure 3.5B,D). The ellipticity is a measure of helicity, so the 'degree of helicity' and the signal amplitude should be directly correlated. However, in the absence of some other form of structural information the link between the 'degree of helicity' and the 'degree of a specific tertiary structure' is not completely clear. Nevertheless, the overall ellipticity signal (local minimum at 223 nm) can be monitored as a function of temperature to determine two-state folding parameters.

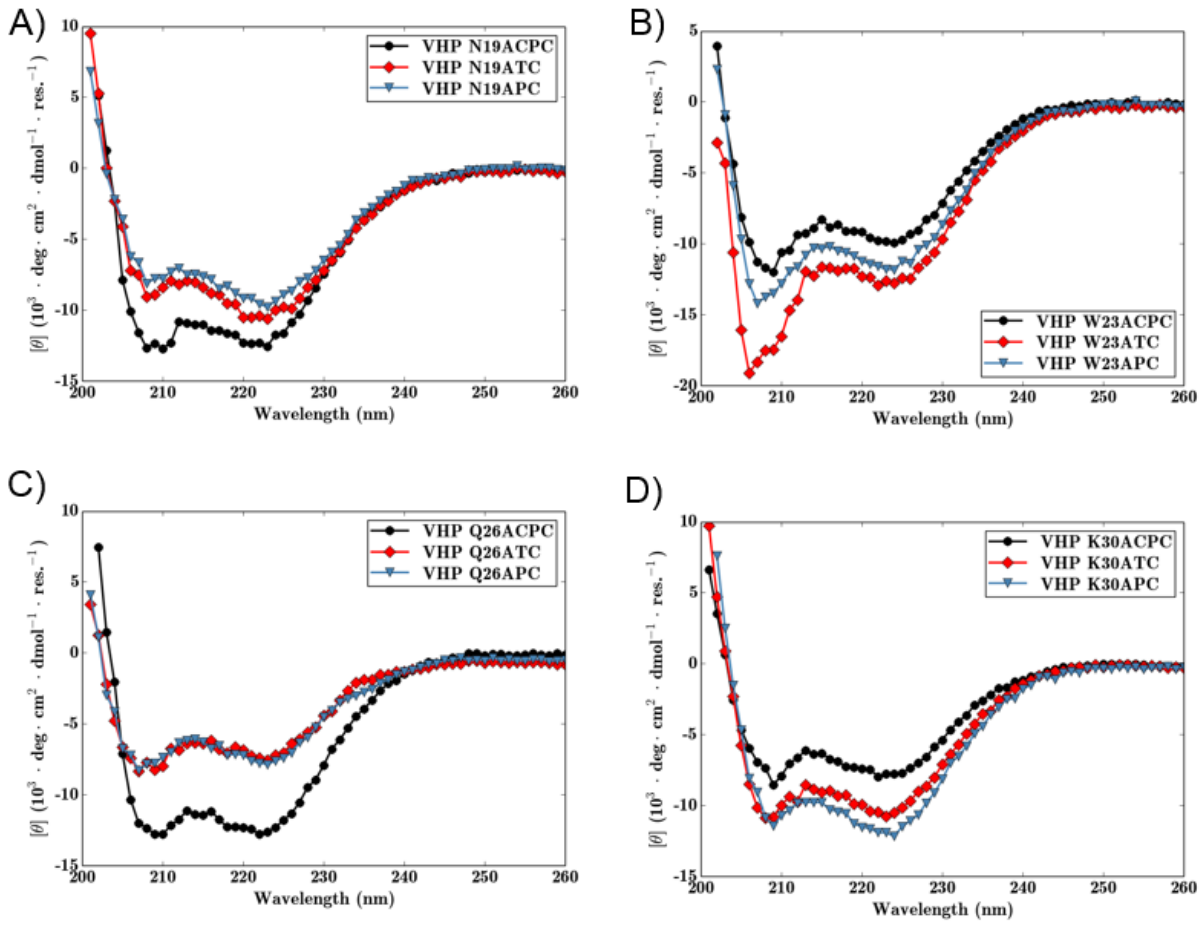


Figure 3.5. Circular dichroism wavelength scans were collected in 20 mM sodium acetate, pH 5.0 at 50 µM peptide concentration. Wavelength scans are shown for each ACPC, ATC, APC variant at positions A) Asn19, B) Trp23, C) Gln26, and D) Lys30.

3.2.3 The Variable Impact of Cyclic β Substitutions on the Two-State Folding Parameters of VHP Variants

Two-state folding parameters were determined for each cyclic β-substituted VHP variant via variable-temperature CD. There were small, but measurable, differences among variant melting temperatures. Substitutions at positions Trp23 and Gln26 exhibit the same general trend: where ATC/APC substitutions at those positions are destabilizing compared to the ACPC variant (Figure 3.6, Table 3.1). The magnitude of the destabilization at position Trp23 (~7°C) is slightly larger than the destabilization observed at Gln26 (~3-5°C). An interesting feature of the ATC/ACPC/APC26 variants is the increased magnitude of the unfolding enthalpy observed upon VHP ACPC26→ATC26 substitution. The increase in unfolding enthalpy is manifest as a steeper unfolding curve and suggests that a polar-polar contact may be formed upon burial of ATC surface area in the folded state.

A different trend was observed for variants containing substitutions at positions Asn19 and Lys30 (Figure 3.7, Table 3.1). In both cases the ACPC/ATC variants displayed similar folding parameters, whereas the APC variant was slightly more stable. For position Lys30, this result is consistent with the structural data reported in Chapter 2, where it was noted that the VHP C-terminal carboxylate formed a salt-bridging interaction with the APC pyrrolidine ring. Preservation of this salt-bridging interaction is evidently beneficial to the overall stability; however, the relative gain in stability is barely discernible given the experimental uncertainty of the melting temperature.

In Chapter 2, I observed that Lys30 was the only position where an acyclic β^3→cyclic β substitution recovered the native VHP thermal stability (VHP β^3hLys30 T_m, 57.2°C; VHP APC30 T_m, 69.4°C). With solely the Chapter 2 folding data, the origin of

this ‘stability recovery’ at position 30 upon hβ^{3}Lys →APC substitution was unclear, mainly due to the confounding effects of salt-bridging interactions involving the APC pyrrolidine ring and the C-terminal carboxylate. The VHP ATC/ACPC30 data presented here in this chapter support the notion that this ‘stability recovery’ originates mainly from some conformational effect ascribed to the cyclic β residue. The potential salt-bridging interaction between the C-terminal carboxylate and APC pyrrolidine ring is removed for VHP ATC30 and ACPC30 variants, yet these variants are still more stabilizing relative to ATC/ACPC substitutions at positions 23 and 26. Finally, no apparent hypothesis arises to explain why the VHP N19APC variant is more stable than the corresponding N19ATC/ACPC substitutions.

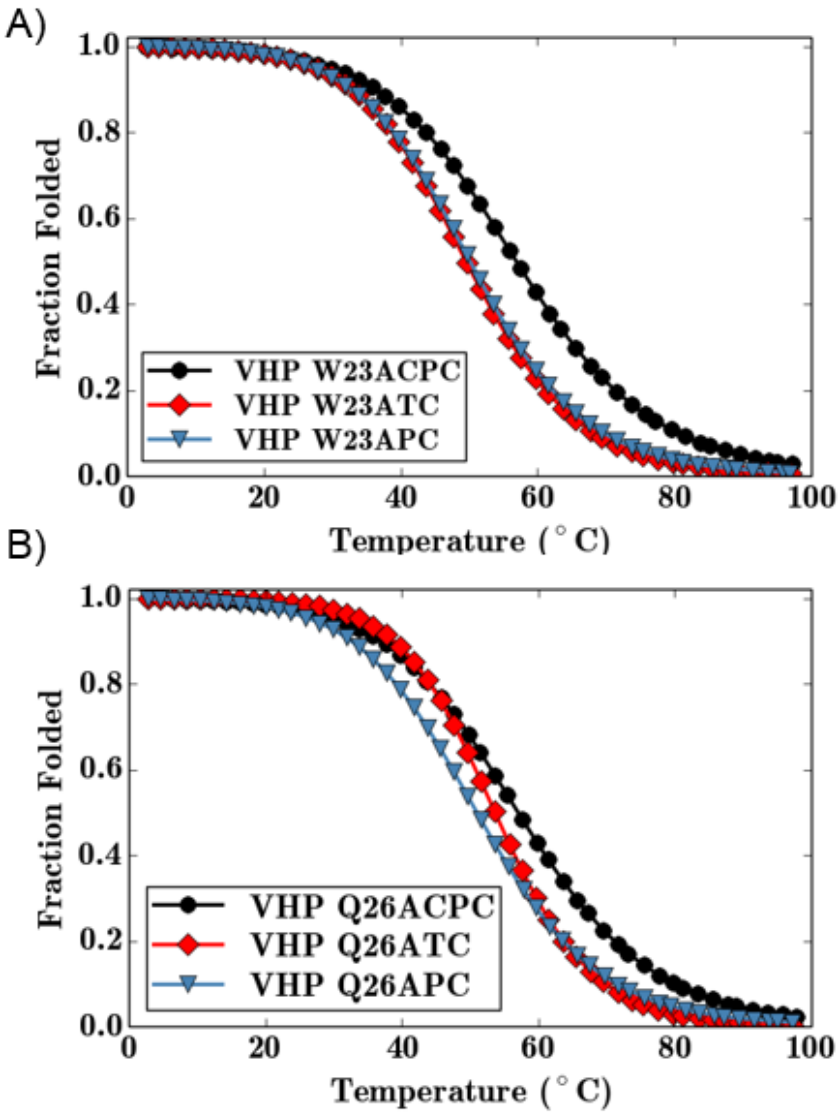


Figure 3.6. Variable-temperature CD denaturation curves show the fraction unfolded as a function of temperature for cyclic β substitutions at positions A) Trp23 and B) Gln26. In both cases the polar/charged cyclic β substitutions were slightly destabilizing relative to the ACPC variants.

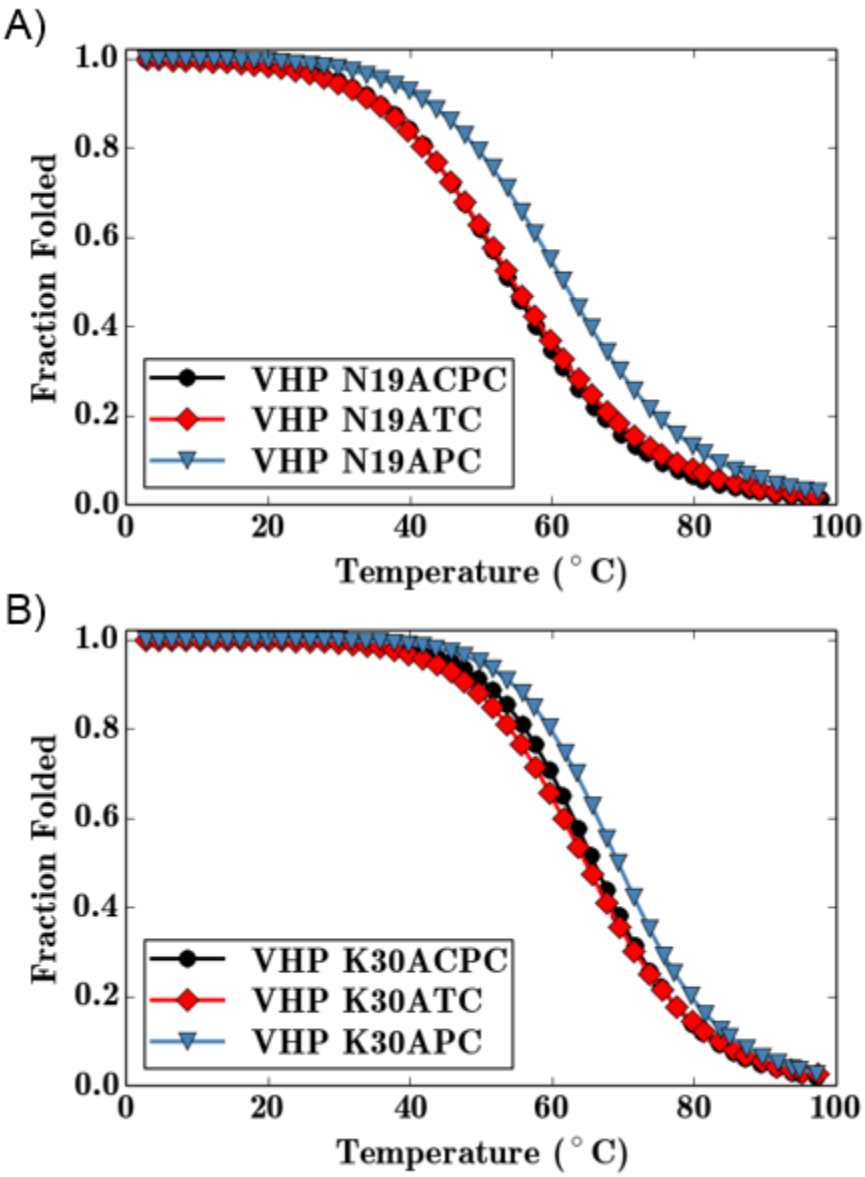


Figure 3.7. Variable-temperature CD denaturation curves show the fraction unfolded as a function of temperature for cyclic β substitutions at positions A) Asn19 and B) Lys30. In both cases, the APC variants are stabilized relative to the ACPC/ATC variants.

Table 3.1. Impact on VHP folding parameters of polar (ATC) or charged (APC) cyclic β substitutions relative to a hydrophobic (ACPC) cyclic β substitution.

Variant	T_m (°C)	ΔH_m (kcal/mol)	ΔT_m (°C)	$\Delta\Delta H_m$ (kcal/mol)
N19ACPC	54.0 (±1.1)	-24.0 (±0.6)	-	-
N19ATC	54.6 (±2.0)	-22.8 (±1.7)	0.6 (±3.1)	0.5 (±4.5)
N19APC	61.6 (±1.3)	-24.5 (±1.2)	7.6 (±2.3)	2.2 (±3.0)
W23ACPC	57.2 (±1.7)	-27.2 (±1.3)	-	-
W23ATC	49.6 (±0.7)	-25.8 (±1.6)	-7.6 (±2.4)	-1.4 (±2.9)
W23APC	50.2 (±2.5)	-25.1 (±2.5)	-7.0 (±4.2)	-2.1 (±3.8)
Q26ACPC	57.1 (±1.0)	-22.1 (±0.5)	-	-
Q26ATC	53.7 (±2.2)	-30.7 (±1.8)	-3.4 (±3.2)	8.6 (±2.3)
Q26APC	51.2 (±2.2)	-23.0 (±1.6)	-5.9 (±3.2)	0.9 (±2.1)
K30ACPC	66.0 (±0.6)	-31.8 (±4.2)	-	-
K30ATC	64.9 (±2.2)	-28.9 (±2.9)	-1.1 (±2.8)	-2.9 (±7.1)
K30APC	69.4 (±1.2)	-33.0 (±2.7)	3.4 (±2.2)	2.2 (±6.9)

3.3 Conclusions

A route towards Fmoc-(3*R*,4*S*)-*trans*-aminotetrahydrofuran carboxylic acid that is safer than a previously described procedure involving organolithium reagents[18] has been employed. The route employed here is essentially the same as that previously reported by Imamura *et al.*,[21] except here the absolute configuration of a key chiral intermediate was established. It was hypothesized that the partial burial of cyclic β residues within helical secondary structure would make the impact of polar (ATC) or charged (APC) cyclic β substitutions detrimental to folded stability, relative to the corresponding hydrophobic cyclic β (ACPC). The folding parameters observed for the VHP variants were consistent with this hypothesis for substitutions at positions Trp23 and Gln26. The trend observed for variants with substitutions at position Asn19 was reversed; however, this reversal is not inconsistent with the cyclic-β-backbone-burial hypothesis because Asn19 is primarily solvent exposed and not buried among the helical backbone. The small, but quantifiable stabilization of the APC substitution at Lys30 relative to ACPC/ATC was attributed to a

favorable salt-bridging interaction that was also observed in the quasiracemic crystal structure of VHP K30APC, but I found that the magnitude of this stabilization (~2°C) was barely detectable given the uncertainty of estimated T_m values. This small difference suggests that the enthalpic salt-bridging interaction, observed in the VHP K30APC quasiracemate structure, plays a minimal role in recovering the stability lost upon α→β^{3}hLys substitution (VHP T_m, 69.1°C; VHP K30β^{3}hLys T_m, 57.2°C; VHP K30APC T_m, 69.4°C). This observation implies that the recovered stabilization in VHP K30APC/ATC/ACPC relative to VHP K30β^{3}hLys has a primarily entropic origin. If the salt-bridging interaction observed in the VHP K30APC structure was the dominant factor in maintaining the native stability of VHP than removal of this interaction would have resulted in stabilities that were similar to cyclic β substitutions at other positions.

The data presented here suggest that polar or charged cyclic β residues are not simply interchangeable in a small tertiary structure if maintaining native stability is desired. Here single substitutions were evaluated, this disparity among cyclic β variant (hydrophobic/polar/charged) stabilities is likely to be amplified in the presence of multiple cyclic β substitutions.

3.4 Appendix

3.4.1 Peptide Synthesis and Purification

Peptides were synthesized using manual, microwave-assisted solid phase synthesis materials and procedures as described in Chapter 2. Similarly, peptides were purified using reverse phase HPLC as described in Chapter 2.

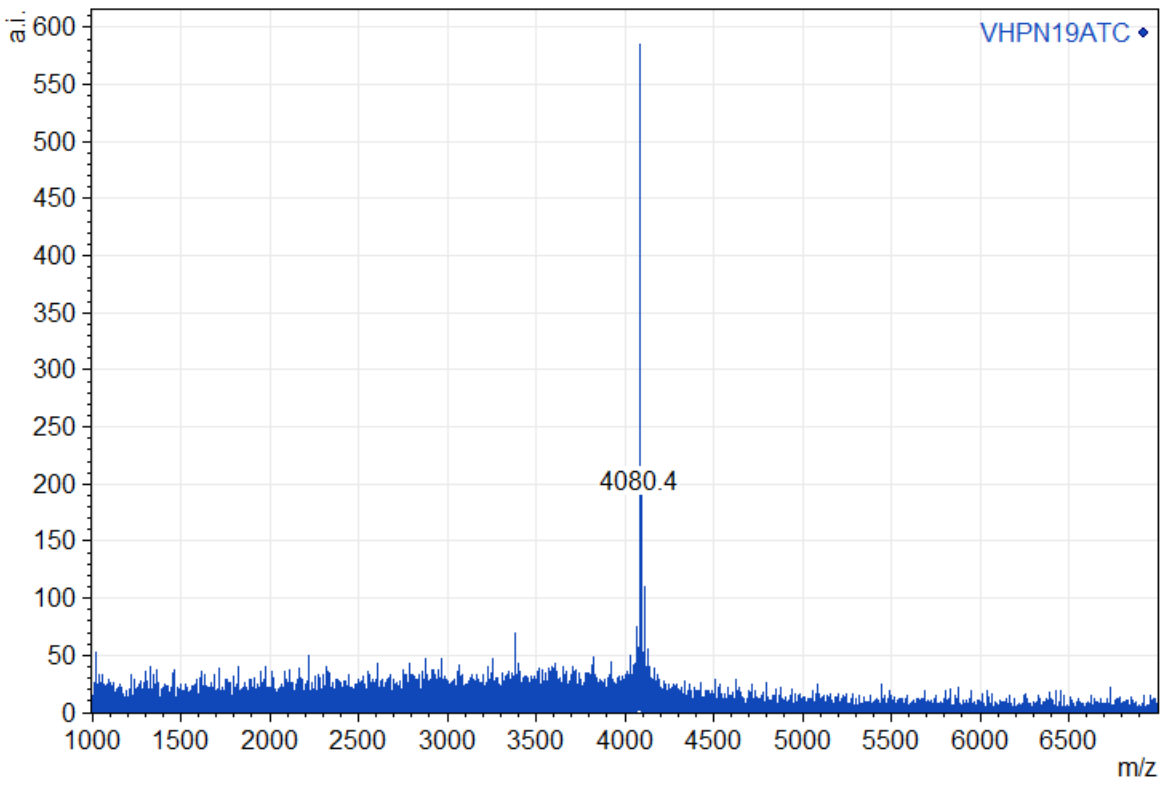
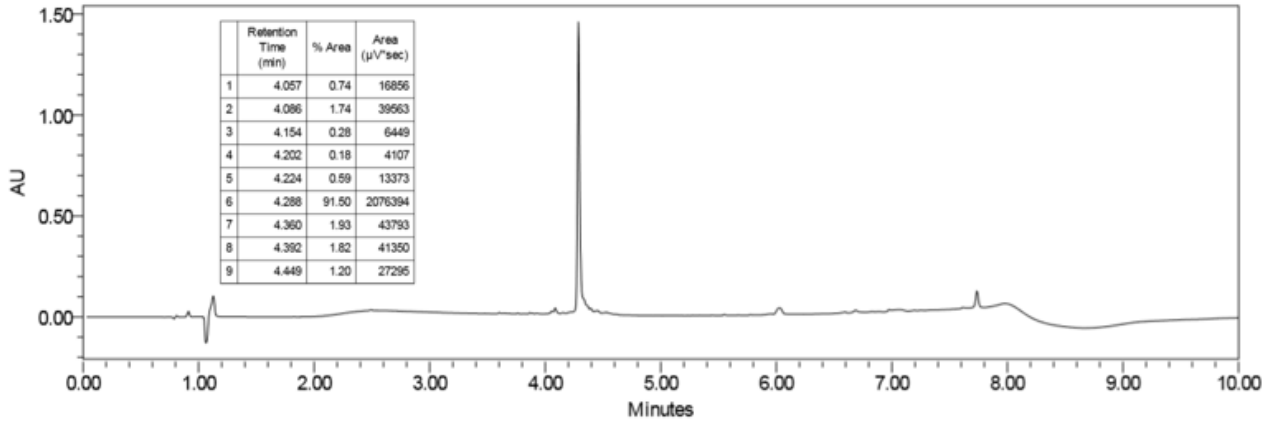


Figure 3.8. H-LSDEDFKAVFGMTRSAFA(**ATC**)LPLWKQQHLKKEKGLF-COOH. VHP N19ATC (ATC: trans-(3R,4S)-aminotetrahydrofuran carboxylic acid) calculated monoisotopic [M+H]=4079.2, observed monoisotopic [M+H]=4080.4.



	Retention Time (min)	% Area	Area (µV*sec)
1	4.057	0.74	16856
2	4.086	1.74	39563
3	4.154	0.28	6449
4	4.202	0.18	4107
5	4.224	0.59	13373
6	4.288	91.50	2076394
7	4.360	1.93	43793
8	4.392	1.82	41350
9	4.449	1.20	27295

Figure 3.9. VHP N19ATC UPLC purity analysis. UPLC gradient from 10-95% B solvent over 6 minutes (0.25 mL/min; column-Waters Acquity peptide CSH C18 1.7 µm, 2.1 x 100 mm, purity >90%).

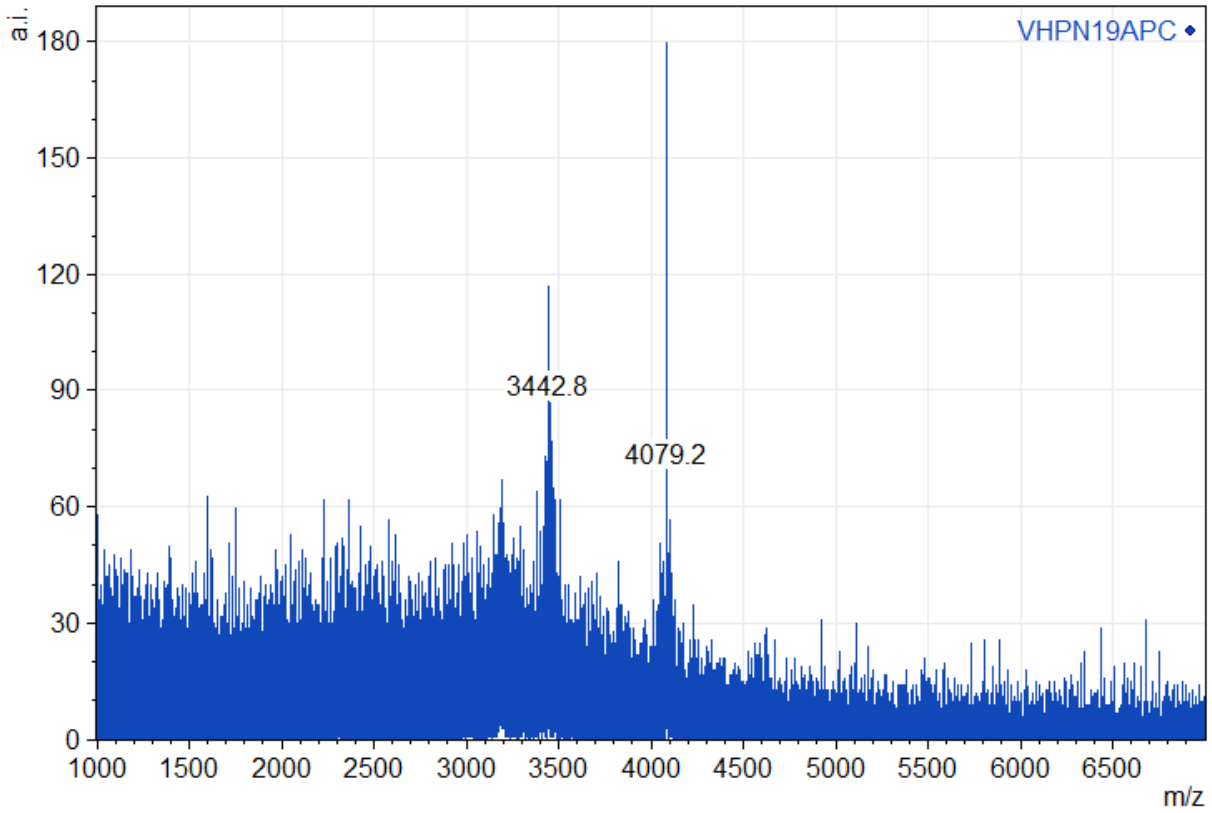


Figure 3.10. H-LSDEDFKAVFGMTRSAFA(**APC**)LPLWKQQHLKKEKGLF-COOH. VHP N19ATC (APC: trans-(3R,4S)-aminopyrrolidine carboxylic acid) calculated monoisotopic [M+H]=4078.2, observed monoisotopic [M+H]=4079.2.

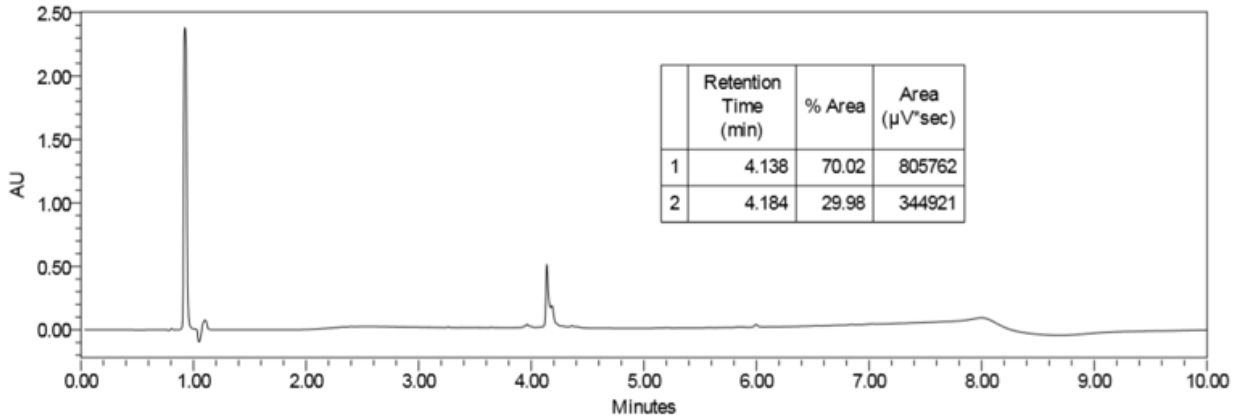


Figure 3.11. VHP N19APC UPLC purity analysis. UPLC gradient from 10-95% B solvent over 6 minutes (0.25 mL/min; column-Waters Acquity Peptide CSH C18 1.7 µm, 2.1 x 100 mm, purity >70%).

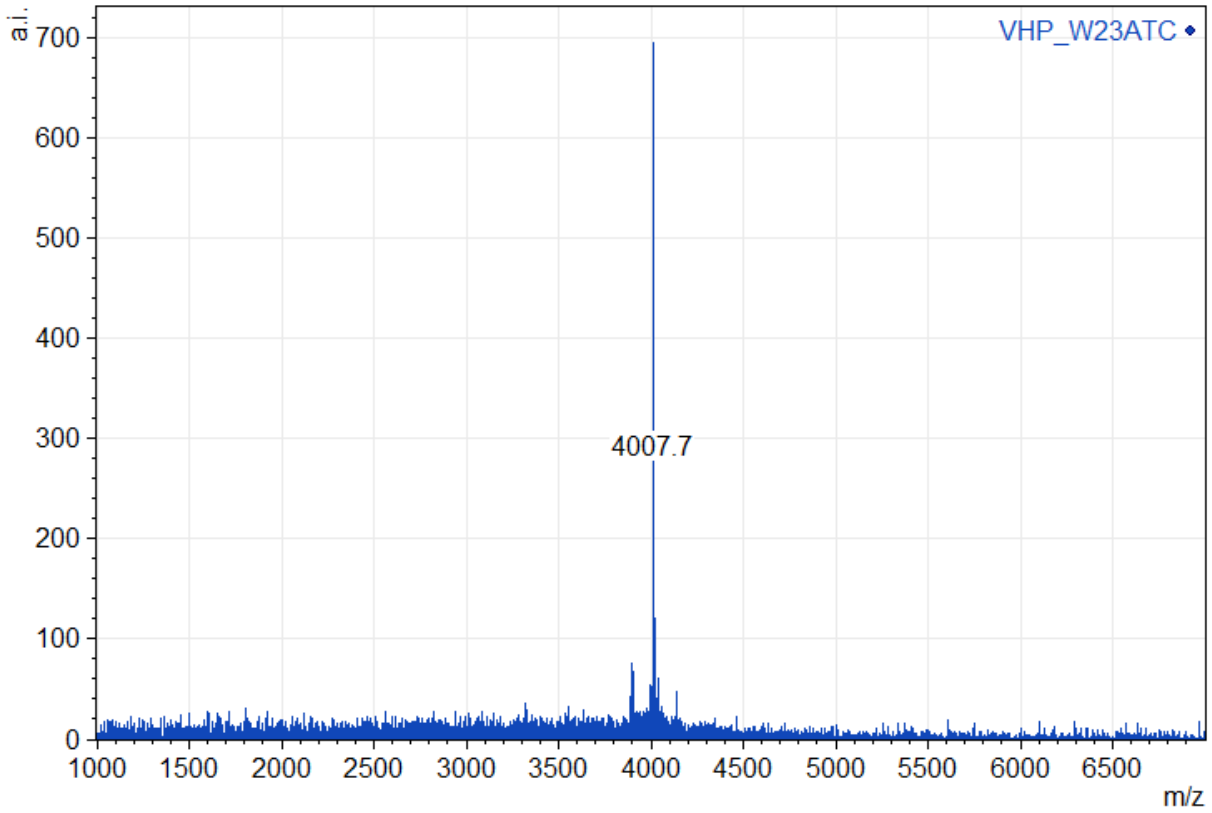
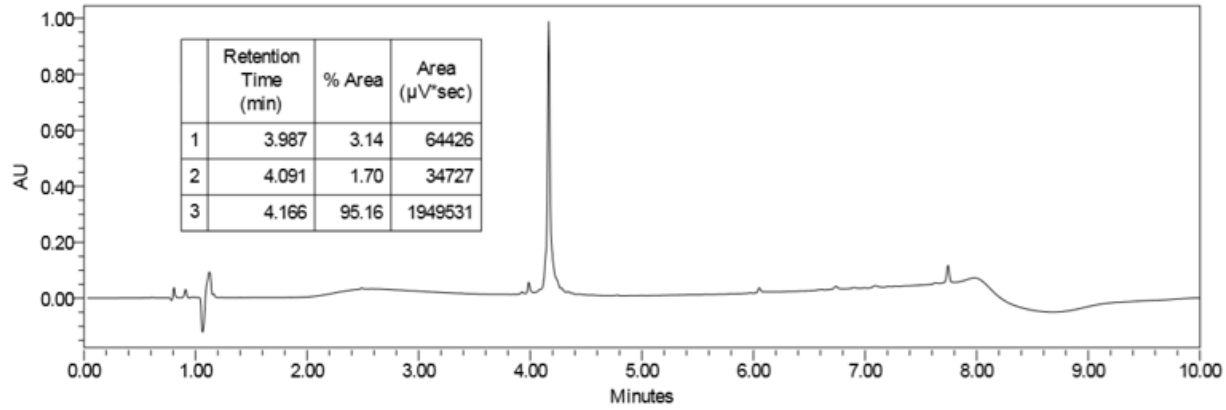


Figure 3.12. H-LSDEDFKAVFGMTRSAFANLPL(**ATC**)KQQHLKKEKGLF-COOH. VHP W23ATC (ATC: trans-(3R,4S)-aminotetrahydrofuran carboxylic acid) calculated monoisotopic [M+H]=4007.1, observed monoisotopic [M+H]=4007.7.



	Retention Time (min)	% Area	Area (µV*sec)
1	3.987	3.14	64426
2	4.091	1.70	34727
3	4.166	95.16	1949531

Figure 3.13. VHP W23ATC UPLC purity analysis. UPLC gradient from 10-95% B solvent over 6 minutes (0.25 mL/min; column-Waters Acquity Peptide CSH C18 1.7 µm, 2.1 x 100 mm, purity >95%).

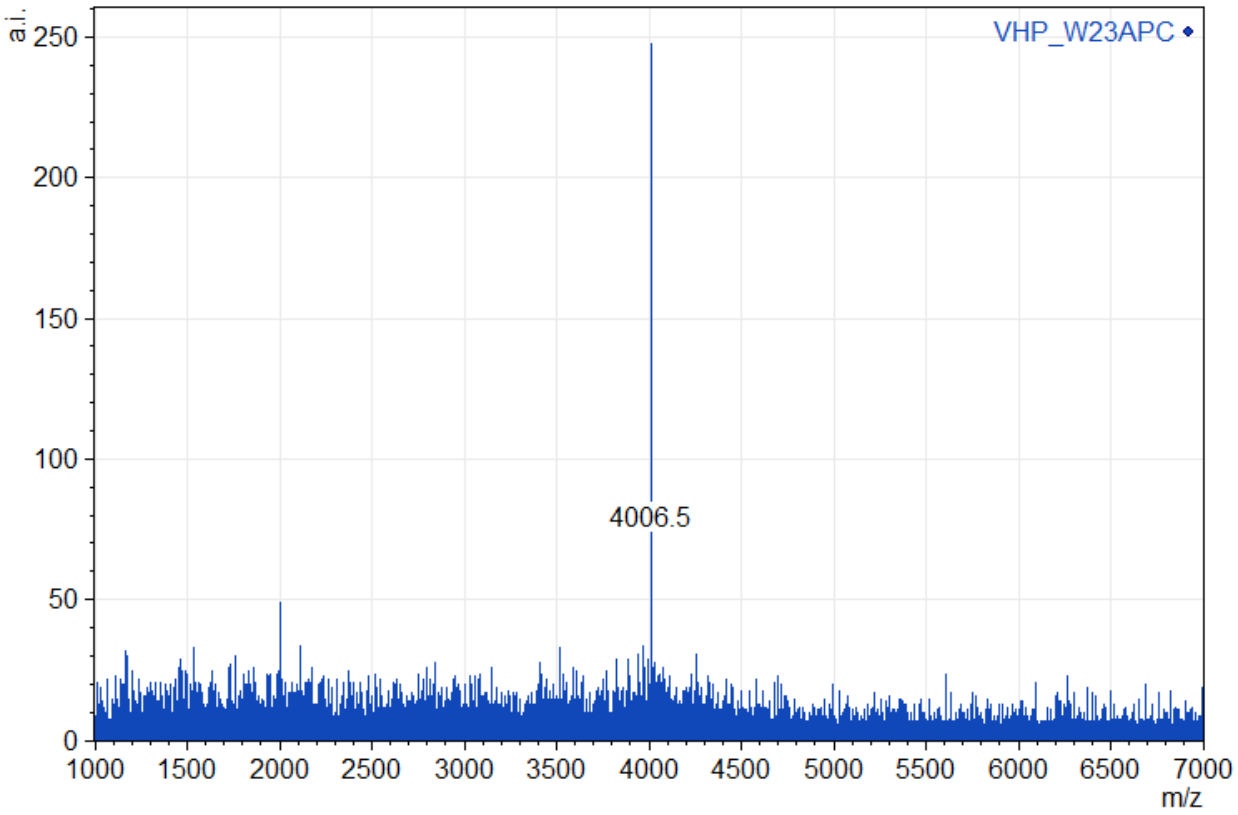
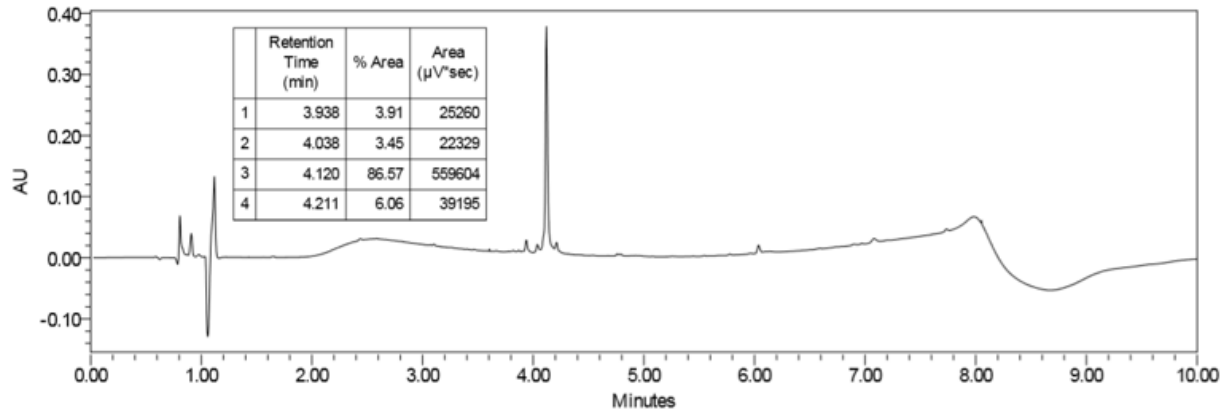


Figure 3.14. H-LSDEDFKAVFGMTRSAFANLPL(**APC**)KQQHLKKEKGLF-COOH. VHP W23APC (APC: trans-(3R,4S)-aminopyrrolidine carboxylic acid) calculated monoisotopic [M+H]=4006.1, observed monoisotopic [M+H]=4006.5.



	Retention Time (min)	% Area	Area (µV*sec)
1	3.938	3.91	25260
2	4.038	3.45	22329
3	4.120	86.57	559604
4	4.211	6.06	39195

Figure 3.15. VHP W23APC UPLC purity analysis. UPLC gradient from 10-95% B solvent over 6 minutes (0.25 mL/min; column-Waters Acquity Peptide CSH C18 1.7 µm, 2.1 x 100 mm, purity >85%).

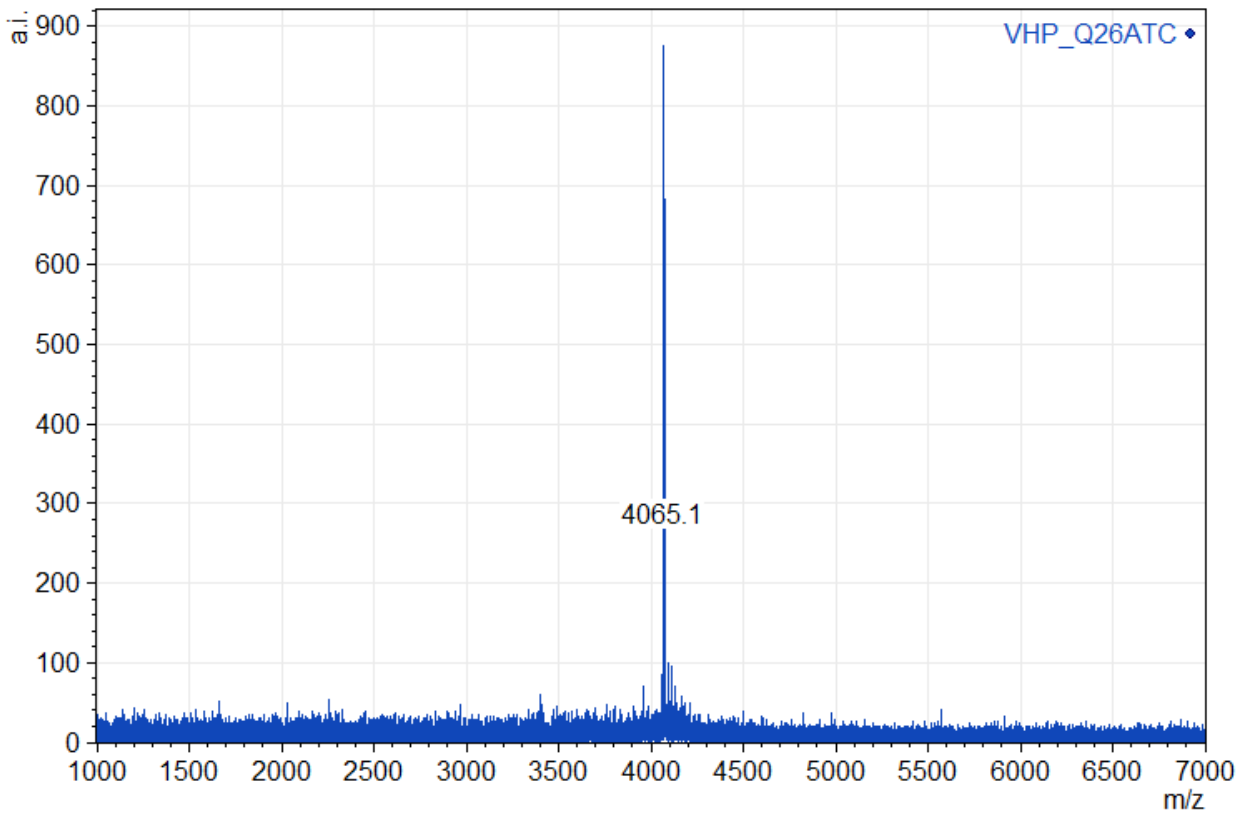


Figure 3.16. H-LSDEDFKAVFGMTRSAFANLPLWKQ(**ATC**)HLKKEKGLF-COOH. VHP Q26ATC (ATC: trans-(3R,4S)-aminotetrahydrofuran carboxylic acid) calculated monoisotopic [M+H]=4065.1, observed monoisotopic [M+H]=4065.1.

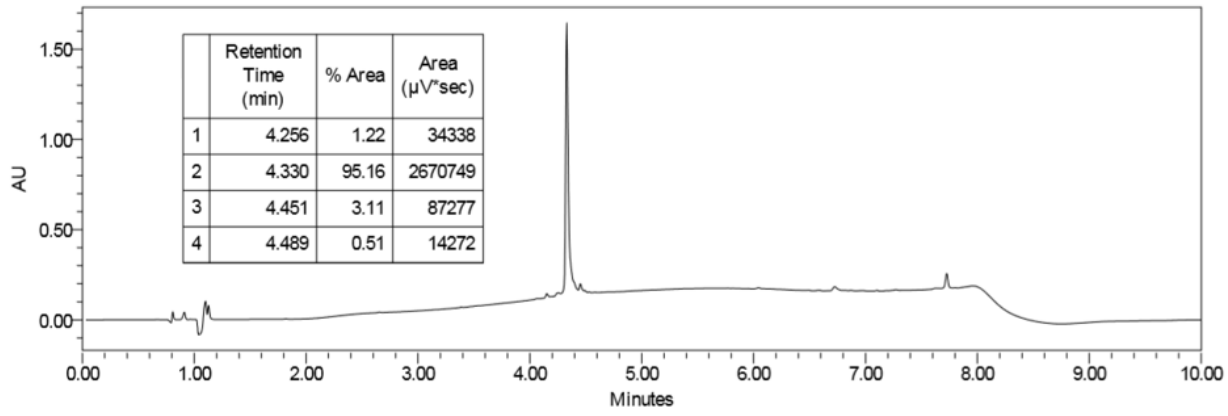


Figure 3.17. VHP Q26ATC UPLC purity analysis. UPLC gradient from 10-95% B solvent over 6 minutes (0.25 mL/min; column-Waters Acquity Peptide CSH C18 1.7 µm, 2.1 x 100 mm, purity >95%).

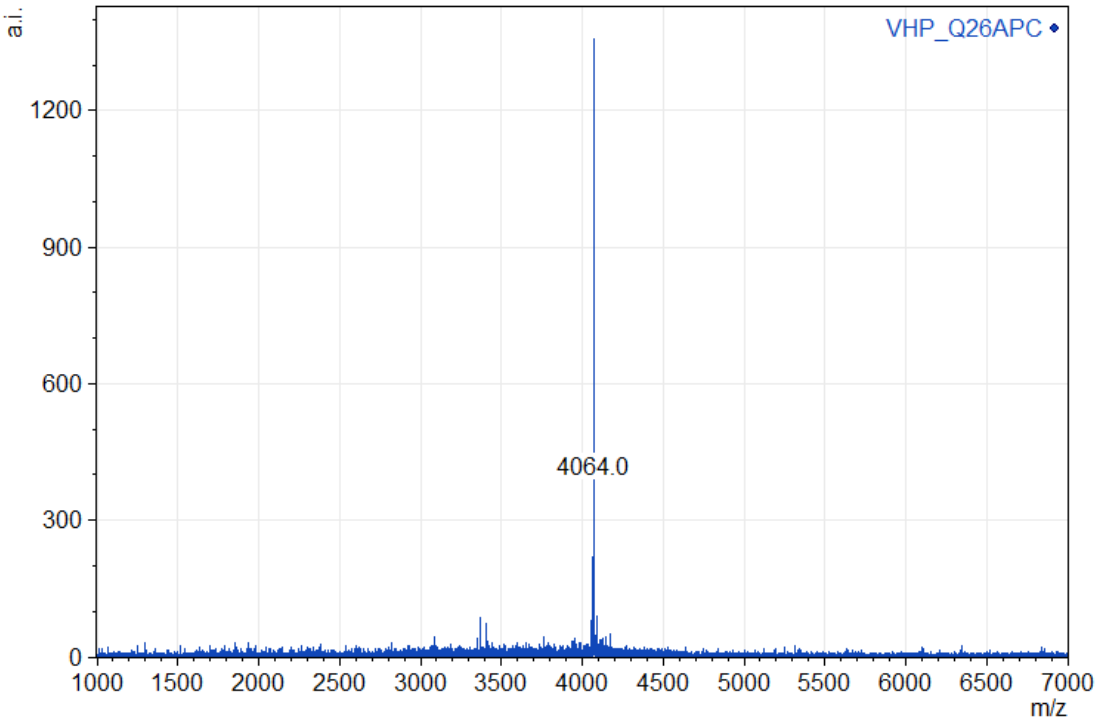


Figure 3.18. H-LSDEDFKAVFGMTRSAFANLPLWKQ(**APC**)HLKKEKGLF-COOH. VHP Q26ATC (ATC: trans-(3R,4S)-aminopyrrolidine carboxylic acid) calculated monoisotopic [M+H]=4064.1, observed monoisotopic [M+H]=4064.0.

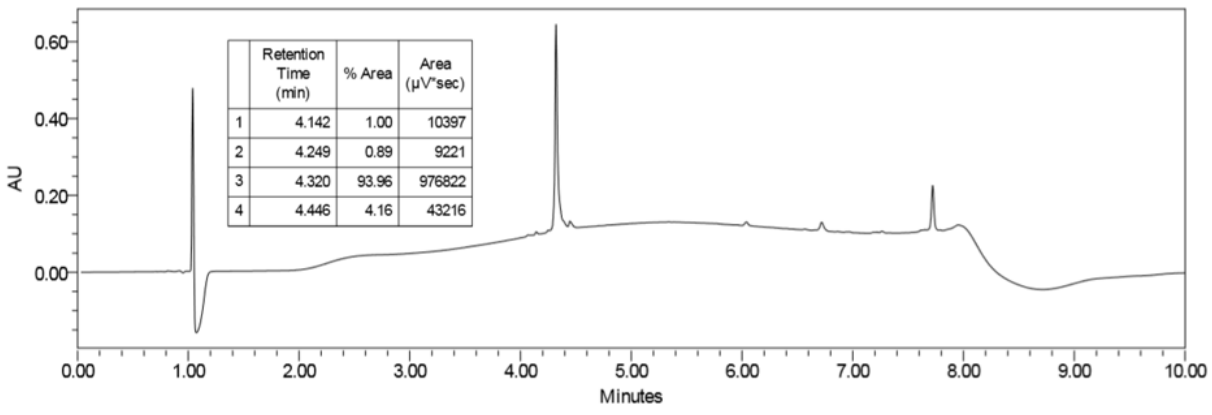


Figure 3.19. VHP Q26APC UPLC purity analysis. UPLC gradient from 10-95% B solvent over 6 minutes (0.25 mL/min; column-Waters Acquity Peptide CSH C18 1.7 µm, 2.1 x 100 mm, purity >90%)

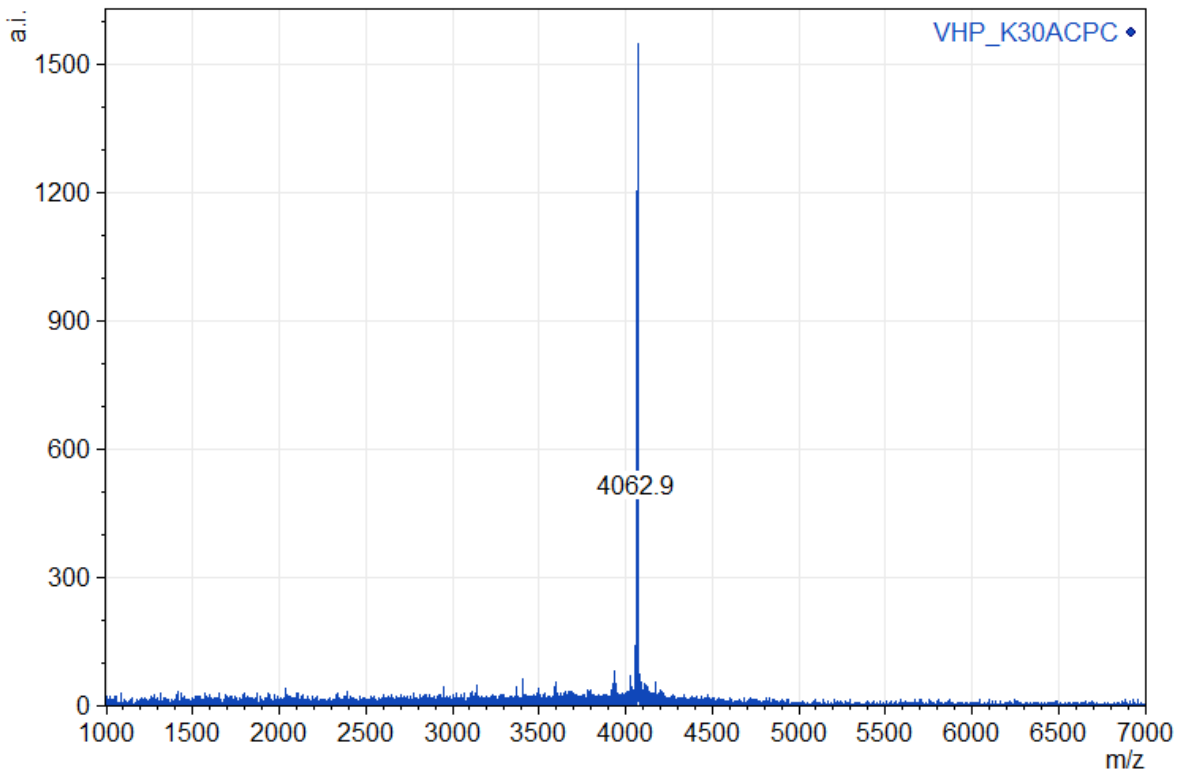


Figure 3.20. H-LSDEDFKAVFGMTRSAFANLPLWKQQHLK(**ACPC**)EKGLF-COOH. VHP K30ACPC (ACPC: trans-(3S,4S)-aminocyclopentane carboxylic acid) calculated monoisotopic [M+H]=4063.1, observed monoisotopic [M+H]=4062.9.

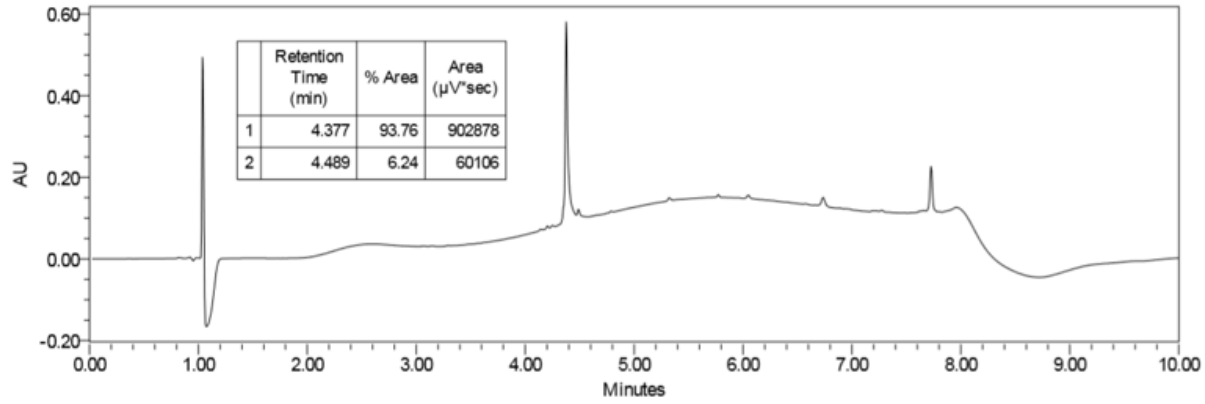


Figure 3.21. VHP K30ACPC UPLC purity analysis. UPLC gradient from 10-95% B solvent over 6 minutes (0.25 mL/min; column-Waters Acquity Peptide CSH C18 1.7 µm, 2.1 x 100 mm, purity >90%).

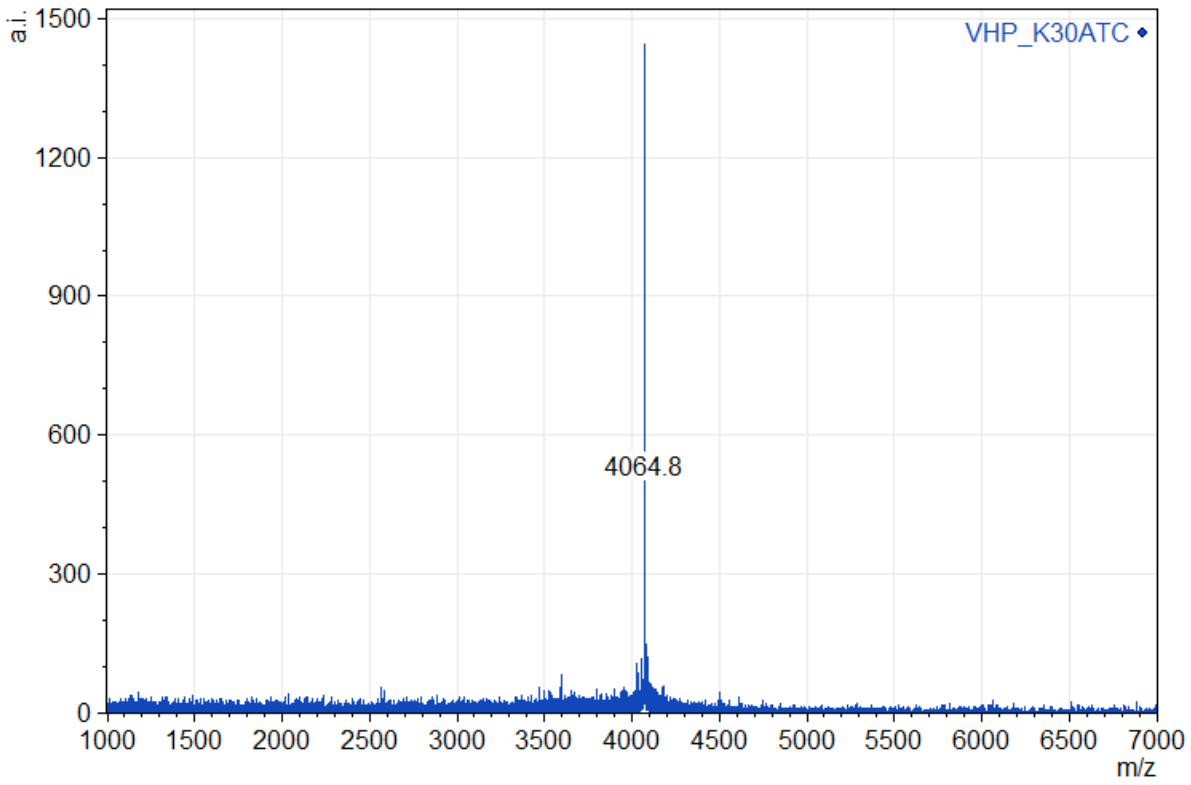


Figure 5.3.22. H-LSDEDFKAVFGMTRSAFANLPLWKQQHLK(**ATC**)EKGLF-COOH. VHP K30ATC (ATC: trans-(3R,4S)-aminotetrahydrofuran carboxylic acid) calculated monoisotopic [M+H]=4063.1, observed monoisotopic [M+H]=4062.9.

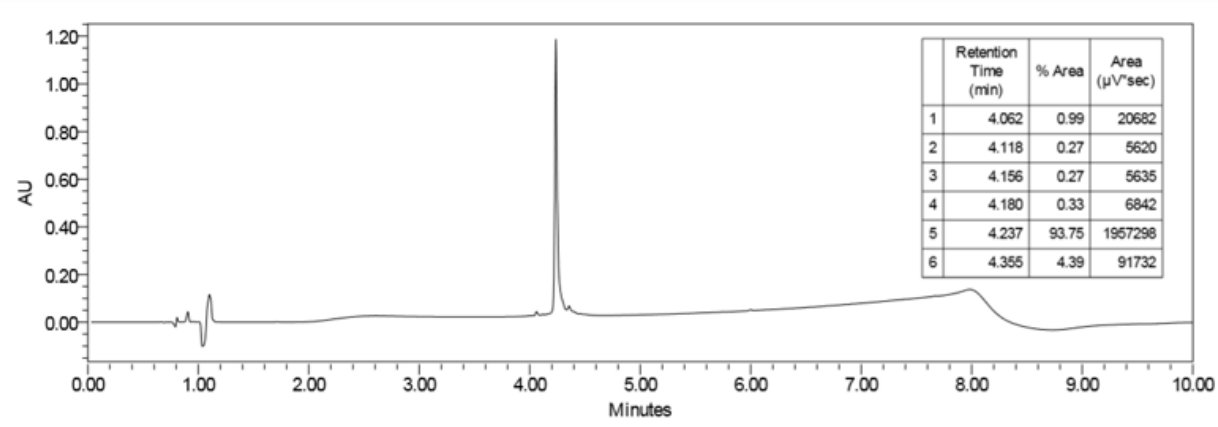


Figure 3.23. VHP K30ACPC UPLC purity analysis. UPLC gradient from 10-95% B solvent over 6 minutes (0.25 mL/min; column-Waters Acquity Peptide CSH C18 1.7 µm, 2.1 x 100 mm, purity >90%).

3.4.2 CD-Thermal Denaturation of Cyclic β Substituted VHP Variants

Variable temperature CD-thermal denaturation data were obtained as described in Chapter 2. Experiments were performed in 20 mM sodium acetate pH 5.0 at 50 μM peptide concentration.

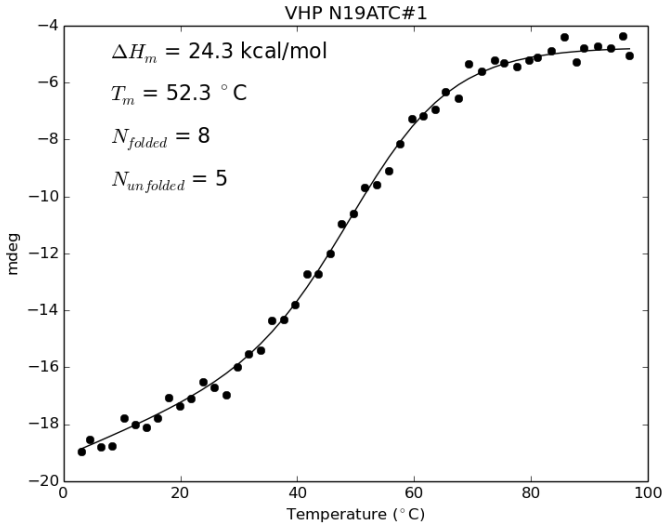


Figure 3.24. VHP N19ATC variable temperature circular dichroism signal monitored at 223 nm fit to a two-state folding model.

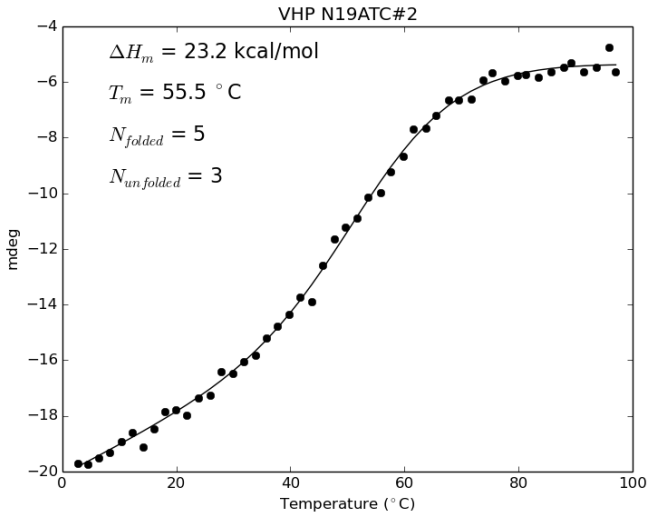


Figure 3.25. VHP N19ATC variable temperature circular dichroism signal monitored at 223 nm fit to a two-state folding model.

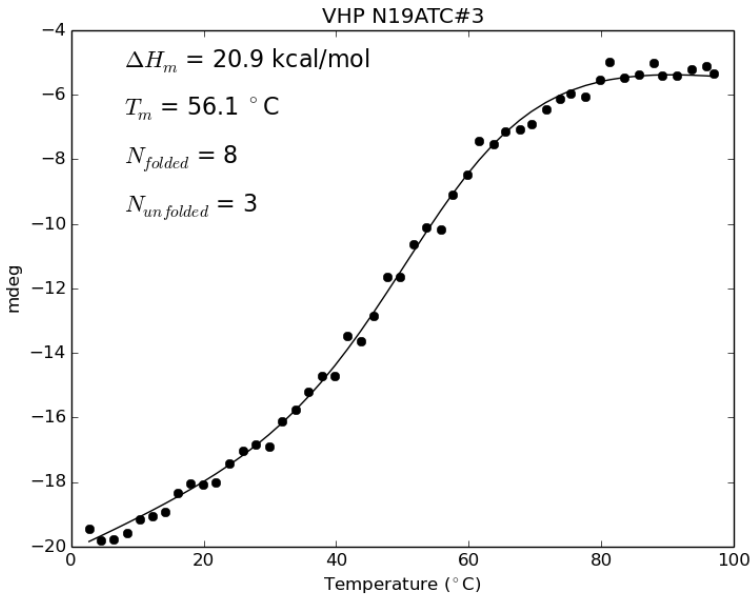


Figure 3.26. VHP N19ATC variable temperature circular dichroism signal monitored at 223 nm fit to a two-state folding model.

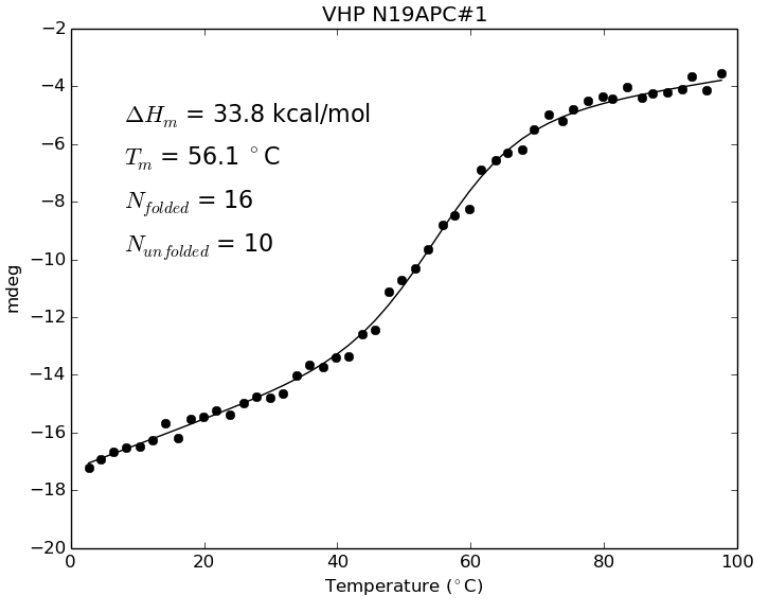


Figure 3.27. VHP N19APC variable temperature circular dichroism signal monitored at 223 nm fit to a two-state folding model.

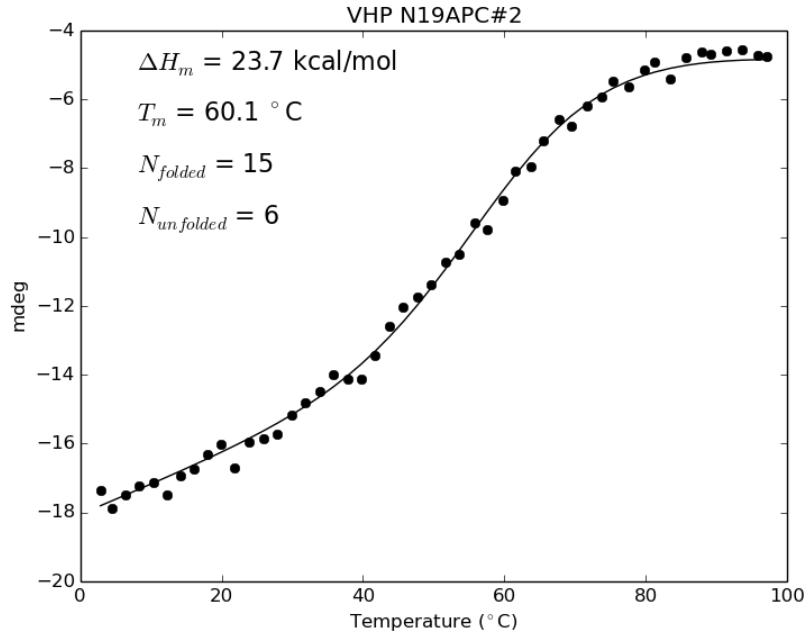


Figure 3.28. VHP N19APC variable temperature circular dichroism signal monitored at 223 nm fit to a two-state folding model.

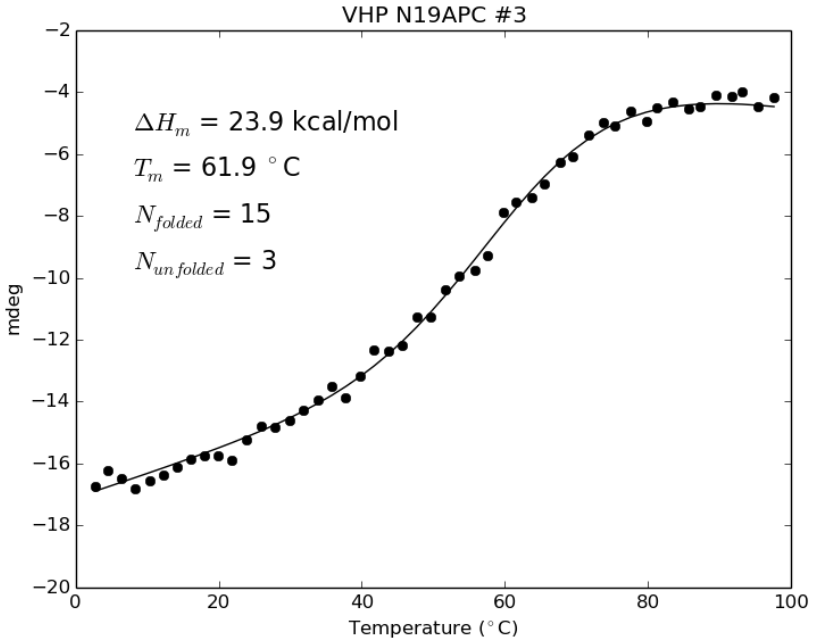


Figure 3.29. VHP N19APC variable temperature circular dichroism signal monitored at 223 nm fit to a two-state folding model.

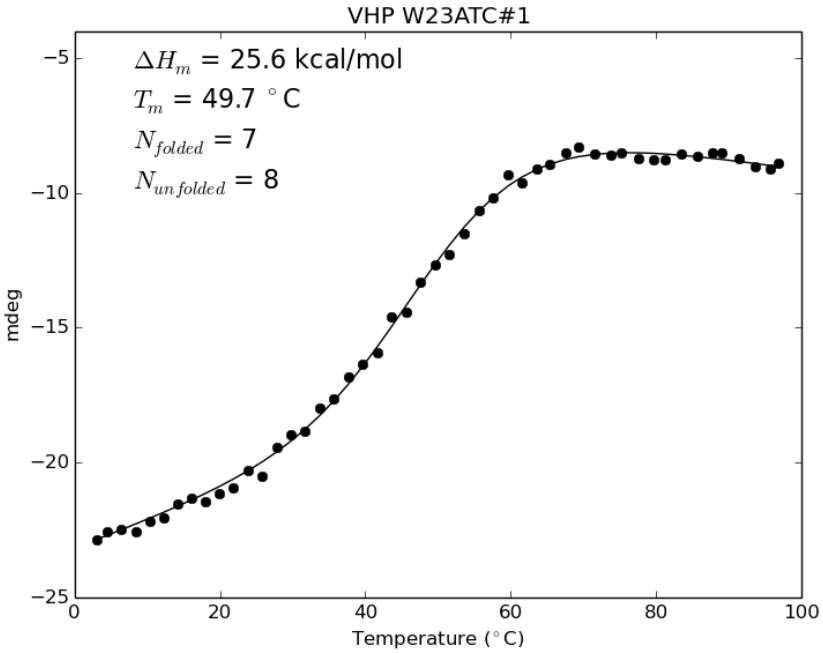


Figure 3.30. VHP W23ATC variable temperature circular dichroism signal monitored at 223 nm fit to a two-state folding model.

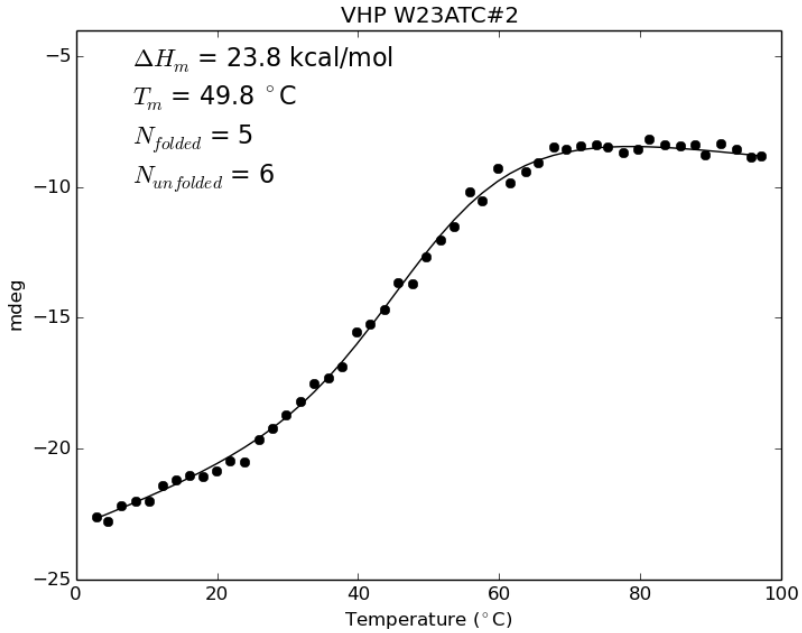


Figure 3.31. VHP W23ATC variable temperature circular dichroism signal monitored at 223 nm fit to a two-state folding model.

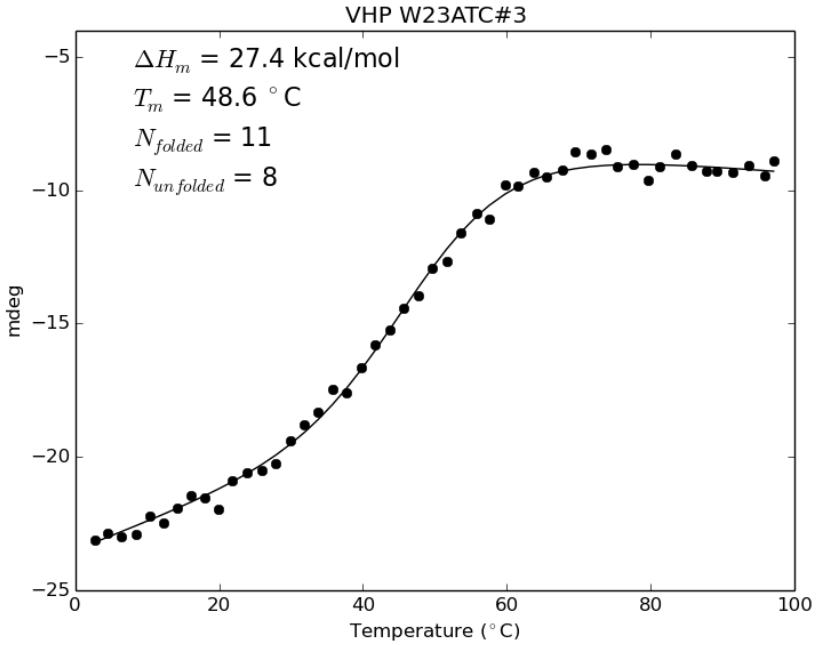


Figure 3.32. VHP W23ATC variable temperature circular dichroism signal monitored at 223 nm fit to a two-state folding model.

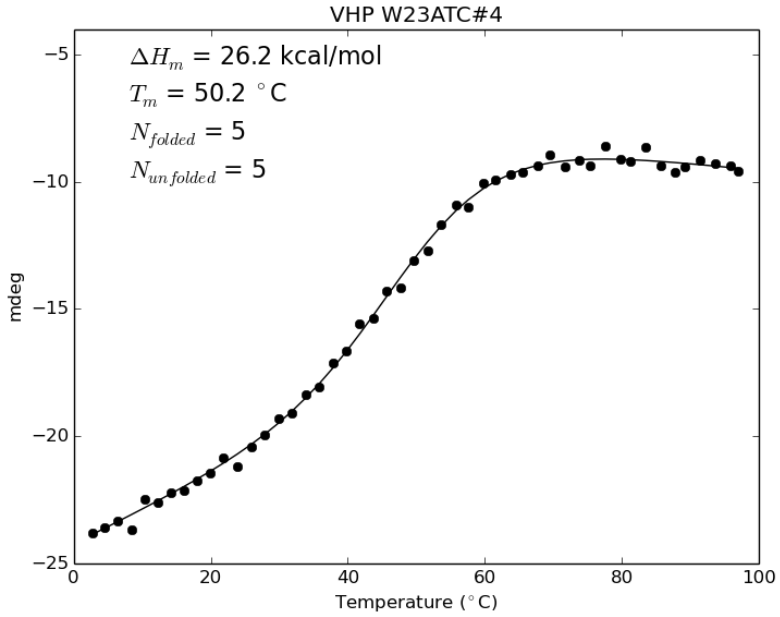


Figure 3.33. VHP W23ATC variable temperature circular dichroism signal monitored at 223 nm fit to a two-state folding model.

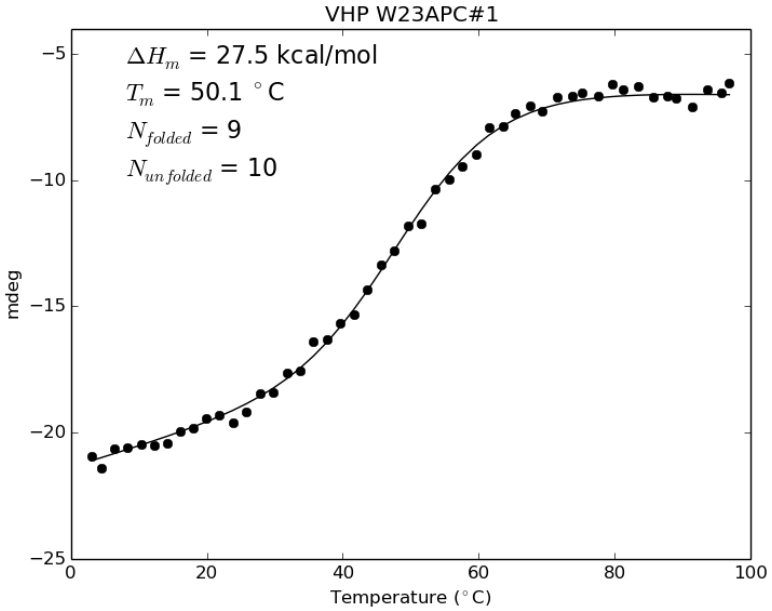


Figure 3.34. VHP W23APC variable temperature circular dichroism signal monitored at 223 nm fit to a two-state folding model.

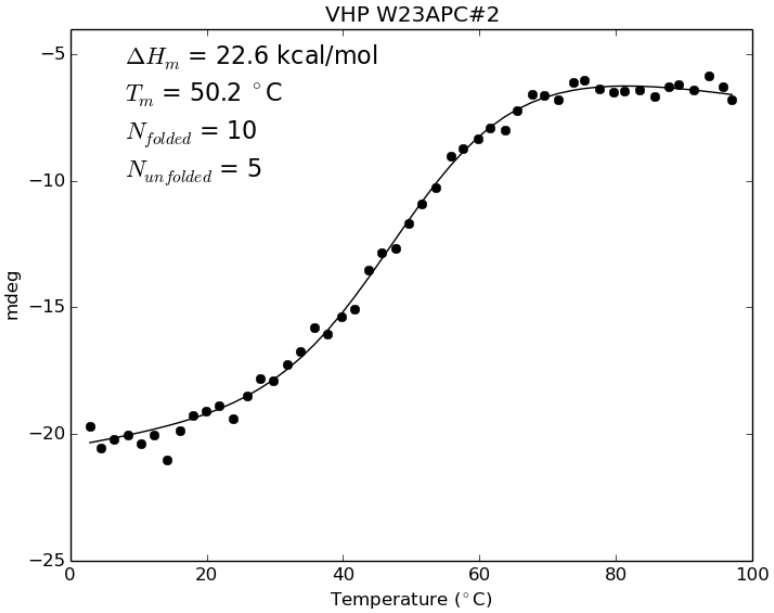


Figure 3.35. VHP W23APC variable temperature circular dichroism signal monitored at 223 nm fit to a two-state folding model.

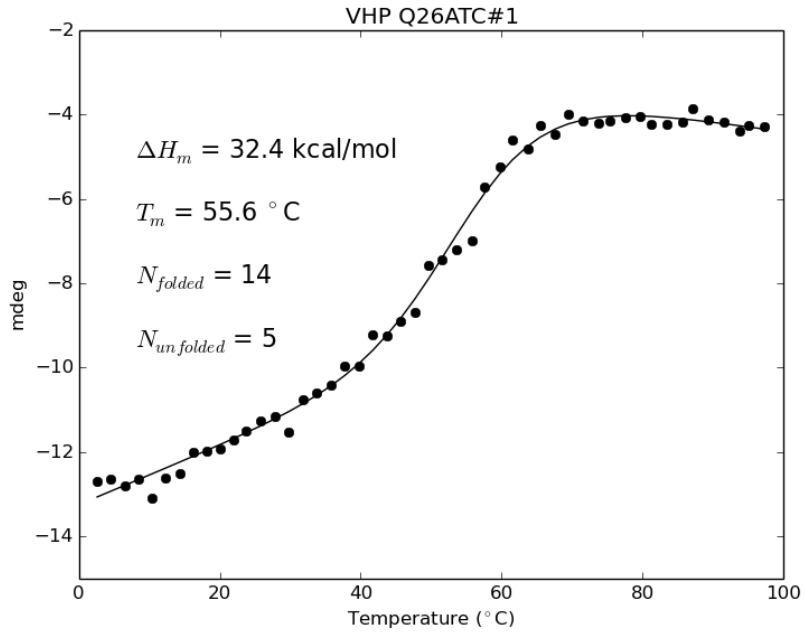


Figure 3.36. VHP Q26APC variable temperature circular dichroism signal monitored at 223 nm fit to a two-state folding model.

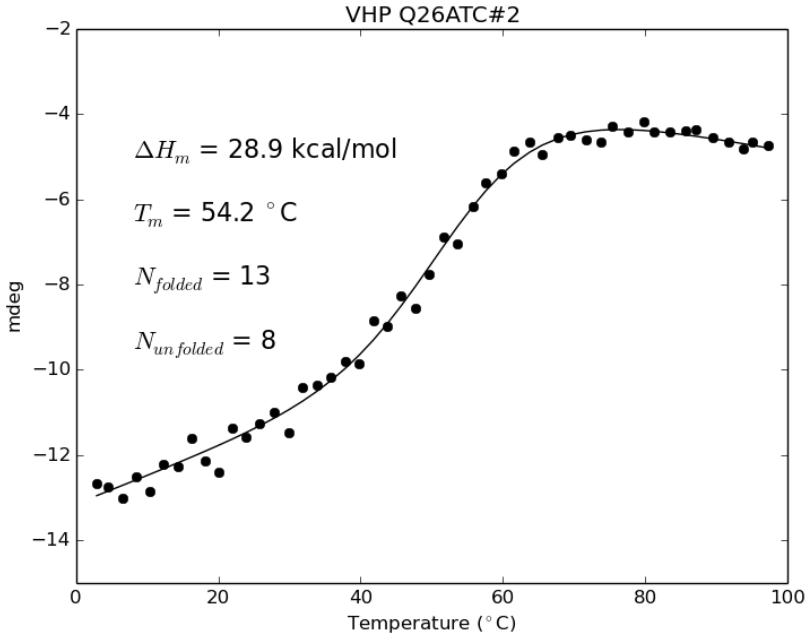


Figure 3.37. VHP W23APC variable temperature circular dichroism signal monitored at 223 nm fit to a two-state folding model.

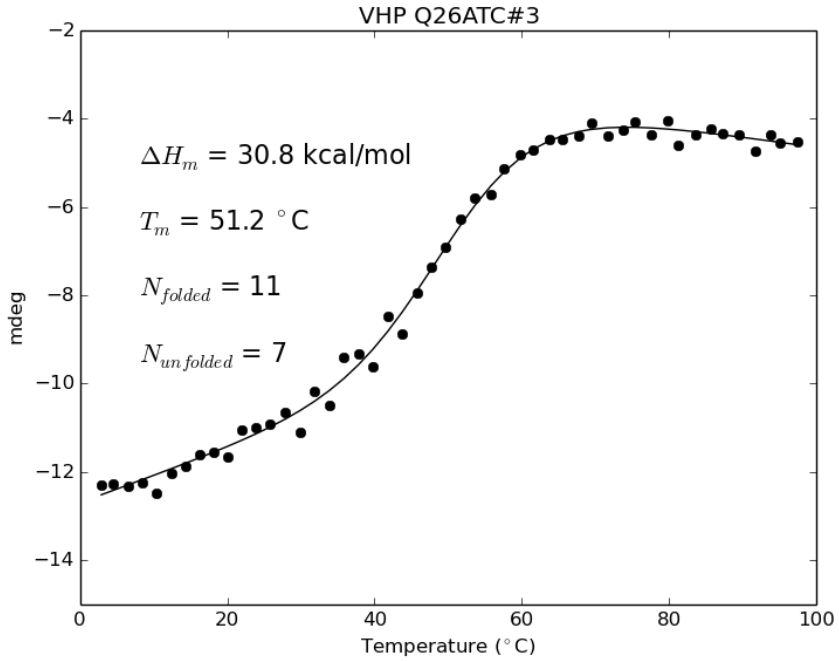


Figure 3.38. VHP Q26ATC variable temperature circular dichroism signal monitored at 223 nm fit to a two-state folding model.

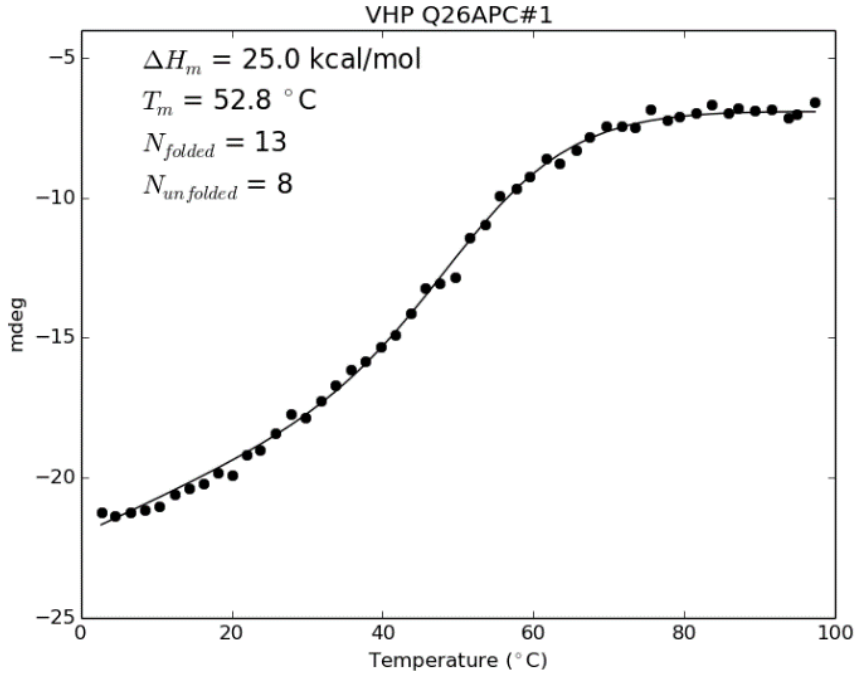


Figure 3.39. VHP Q26APC variable temperature circular dichroism signal monitored at 223 nm fit to a two-state folding model.

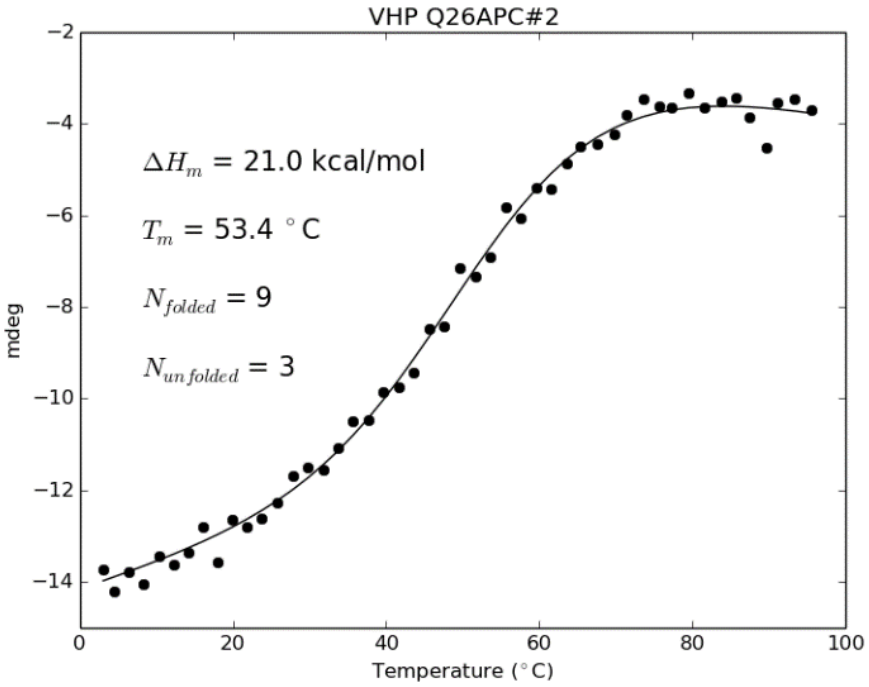


Figure 3.40. VHP Q26APC variable temperature circular dichroism signal monitored at 223 nm fit to a two-state folding model.

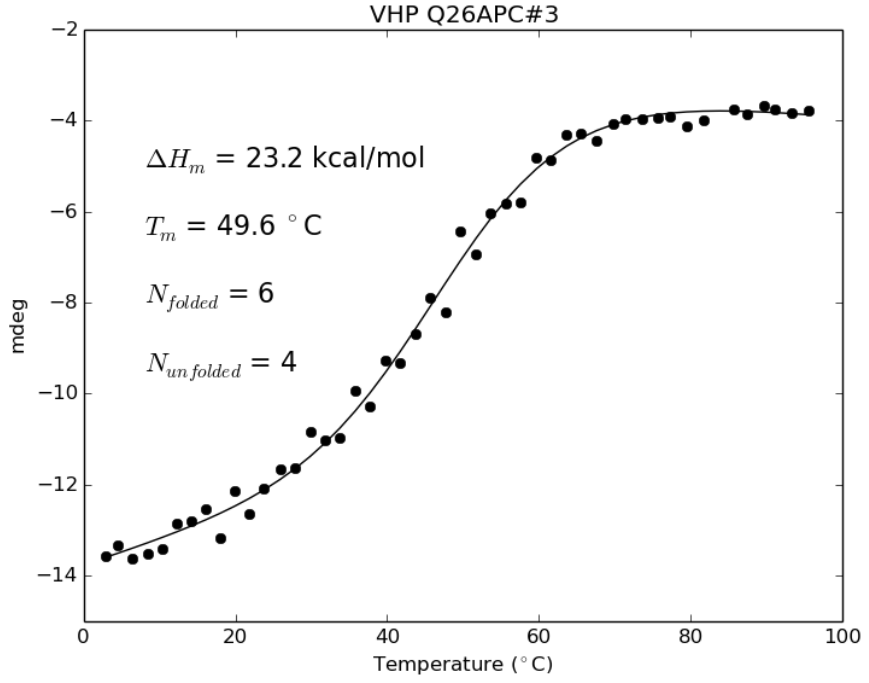


Figure 3.41. VHP Q26APC variable temperature circular dichroism signal monitored at 223 nm fit to a two-state folding model.

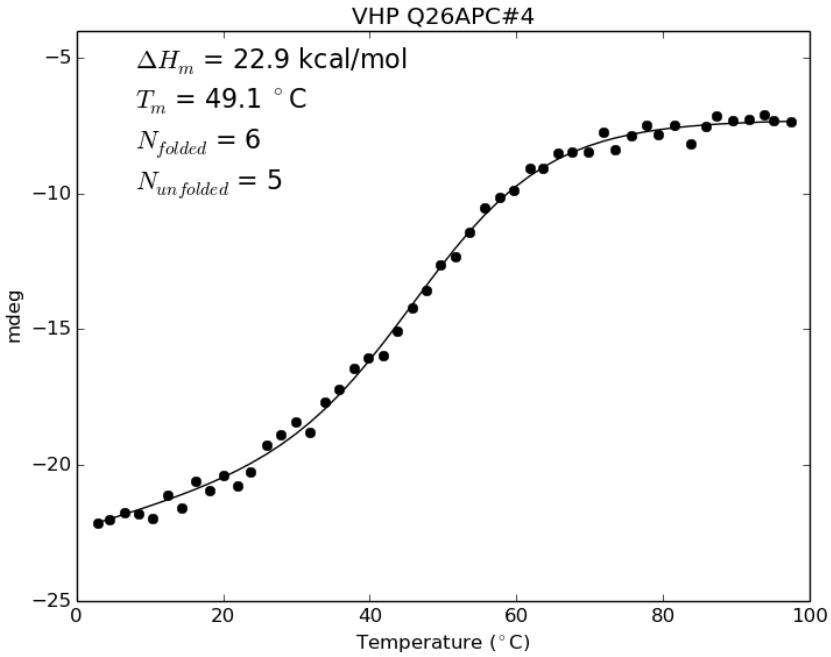


Figure 3.42. VHP Q26APC variable temperature circular dichroism signal monitored at 223 nm fit to a two-state folding model.

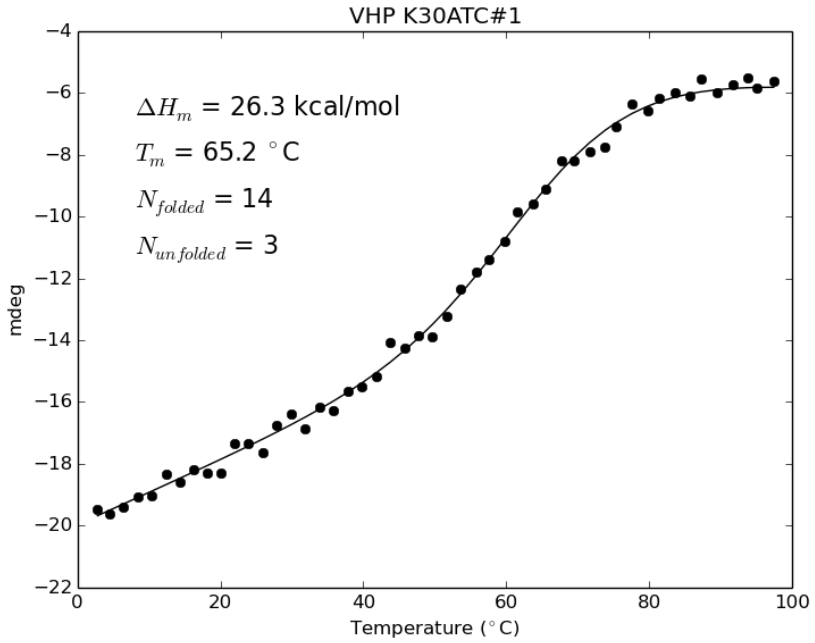


Figure 3.43. VHP K30ATC variable temperature circular dichroism signal monitored at 223 nm fit to a two-state folding model.

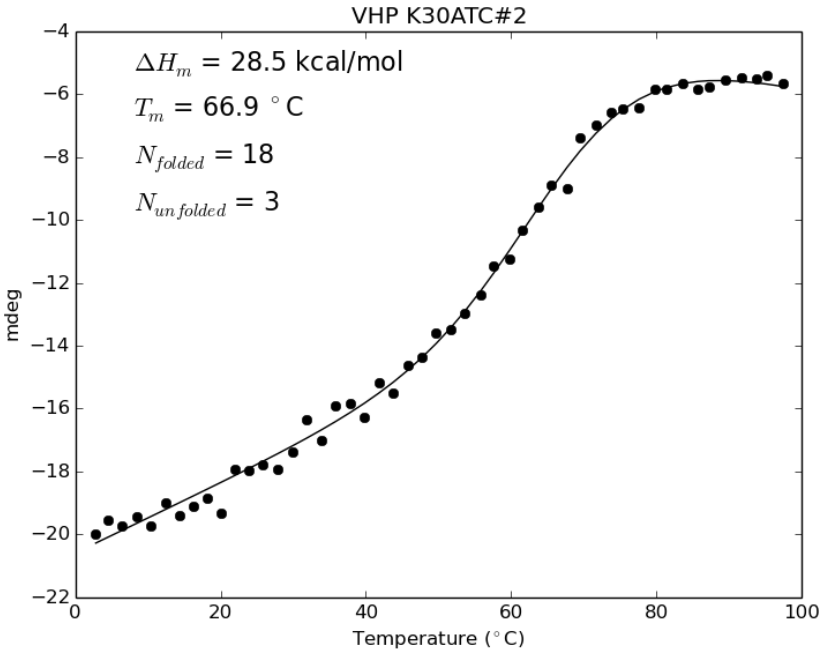


Figure 3.44. VHP K30ATC variable temperature circular dichroism signal monitored at 223 nm fit to a two-state folding model.

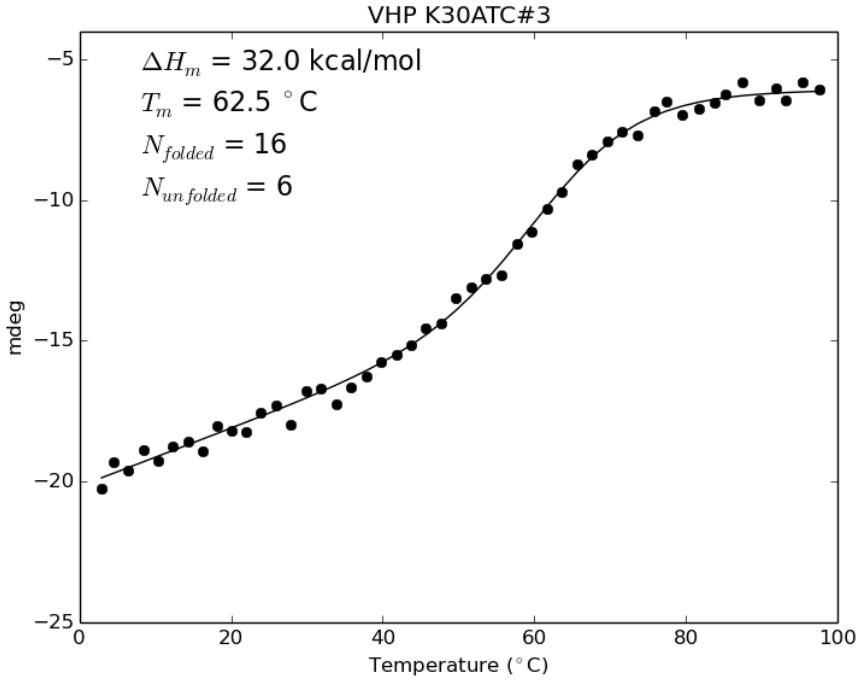


Figure 3.45. VHP K30ATC variable temperature circular dichroism signal monitored at 223 nm fit to a two-state folding model.

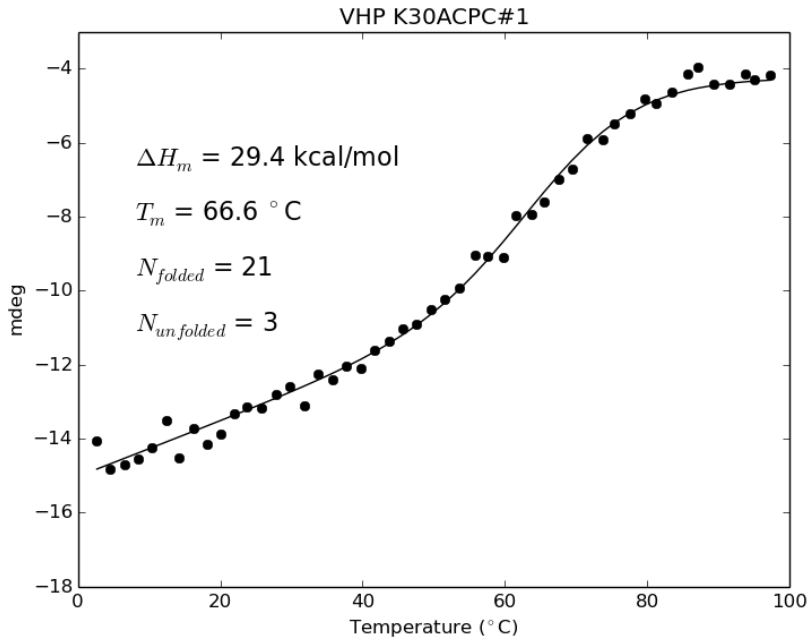


Figure 3.46. VHP K30ACPC variable temperature circular dichroism signal monitored at 223 nm fit to a two-state folding model.

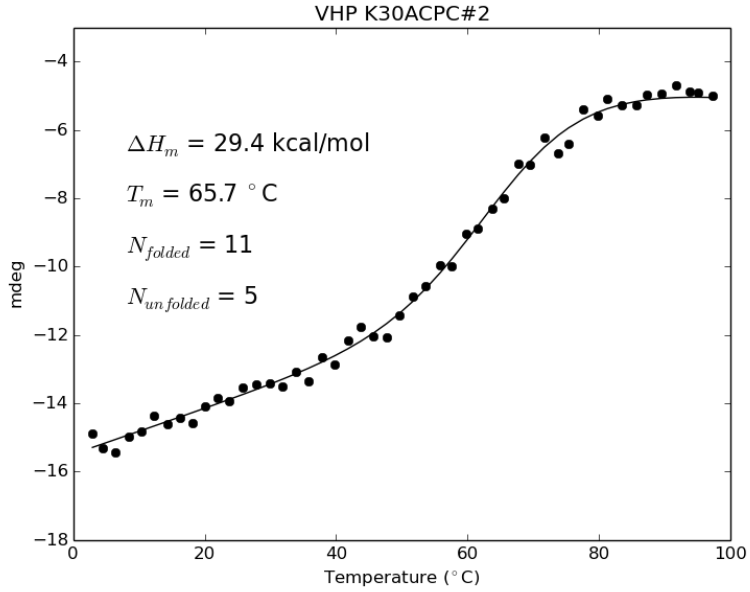


Figure 3.47. VHP K30ACPC variable temperature circular dichroism signal monitored at 223 nm fit to a two-state folding model.

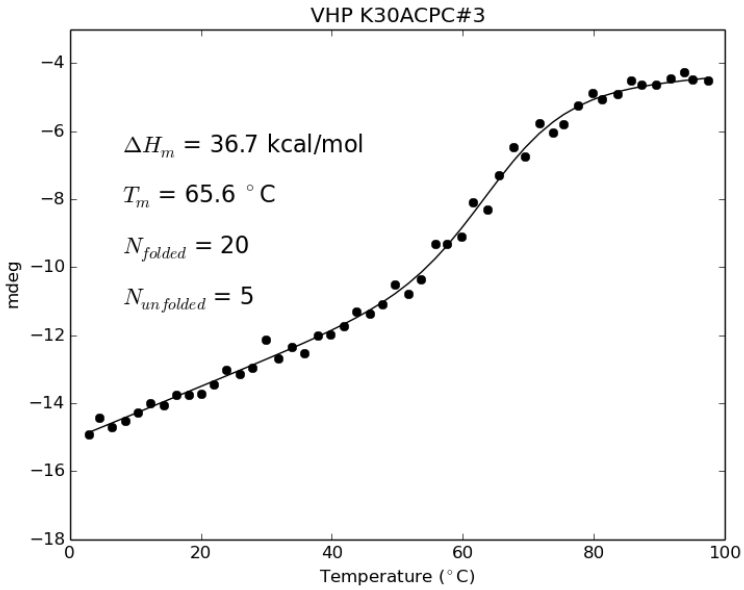
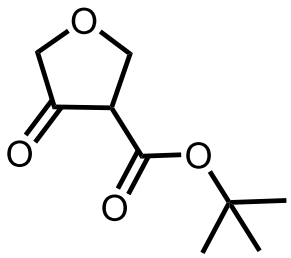


Figure 3.48. VHP K30ACPC variable temperature circular dichroism signal monitored at 223 nm fit to a two-state folding model.

Table 3.2. CD-derived folding parameters for native VHP and cyclic β-substituted variants.

Variant	T_m (°C)	ΔH_m (kcal/mol)	ΔT_m (°C)	$\Delta\Delta H_m$ (kcal/mol)
VHP	69.1 (±0.4)	32.8 (±1.2)	-	-
N19ACPC	54.0 (±1.1)	24.0 (±0.6)	-15.1 (±1.5)	-8.8 (±1.8)
N19ATC	54.6 (±2.0)	22.8 (±1.7)	-14.5 (±2.4)	-10.0 (±2.9)
N19APC	61.6 (±1.3)	24.5 (±1.2)	-7.5 (±0.7)	-8.3 (±2.4)
W23ACPC	57.2 (±1.7)	27.2 (±1.3)	-11.9 (±2.1)	-5.6 (±1.6)
W23ATC	49.6 (±0.7)	25.8 (±1.6)	-19.5 (±1.1)	-7.0 (±2.9)
W23APC	50.2 (±2.5)	25.1 (±2.5)	-18.9 (±2.9)	-7.7 (±2.7)
Q26ACPC	57.1 (±1.0)	22.1 (±0.5)	-12.1 (±1.4)	-10.7 (±1.7)
Q26ATC	53.7 (±2.2)	30.7 (±1.8)	-15.4 (±2.6)	-2.1 (±3.0)
Q26APC	51.2 (±2.2)	23.0 (±1.6)	-17.9 (±2.6)	-9.8 (±2.8)
K30ACPC	66.0 (±0.6)	31.8 (±4.2)	-3.1 (±1.0)	-1.0 (±5.6)
K30ATC	64.9 (±2.2)	28.9 (±2.9)	-4.2 (±2.6)	-3.9 (±4.1)
K30APC	69.4 (±1.2)	33.0 (±2.7)	0.3 (±1.6)	-0.1 (±3.9)

3.4.3 ATC Synthesis



ATC β-ketoester

Formula: $C_9H_{14}O_4$

IUPAC: tert-butyl 4-oxooxolane-3-carboxylate

SMILES: CC(C)(C)OC(=O)C1COCC1=O

Procedure: The reaction procedure comprises two main steps: (1) generation of the ethyl glycolate alkoxide nucleophile and (2) conjugate addition/cyclization.

(1) A 250 mL flame-dried round bottom flask was charged with a mineral oil suspension of 60% w/w NaH (3.84 g, 96 mmol, 1.0 eq.). The NaH suspension was rinsed with pentanes (2x50 mL). The powdery metal hydride was re-suspended in 75 mL anhydrous diethyl ether (dried with an alumina plug). Ethyl glycolate (10.0 g, 9.1 mL, 96 mmol, 1.0 eq.) was added dropwise to the NaH/ether suspension. (Note: hydrogen gas evolves rapidly so adequate ventilation of the flask must be ensured). The heterogeneous slurry was stirred at room temperature for 30 minutes, after which ether was removed with rotary evaporation to afford a white powder.

(2) A solution of *tert*-butyl acrylate (14.8 g, 16.9 mL, 115 mmol, 1.2 eq.) in DMSO (75 mL, 1.3 M, anhydrous, alumina plug) was added to the flask in one lot, and the mixture was stirred at 0°C for 15 min. The flask was allowed to warm to room temperature, and the mixture was stirred for an additional 30 min. H_2SO_4 (50 mL

5% v/v aq.) was added, and the heterogeneous mixture was extracted with diethyl ether (3x100 mL). The organic layers were combined and washed with saturated brine (100 mL). The organic layer was dried with $MgSO_4$ and evaporated to afford a light brown oil.

Purification: Silica flash chromatography with 10% v/v EtOAc in pentanes (visualized on TLC plate with PMA stain, R_f ~ 0.1 with column conditions). The product was concentrated to afford a thick pink oil that forms a white crystalline solid when stored at -20°C (9.6 g, 55% yield).

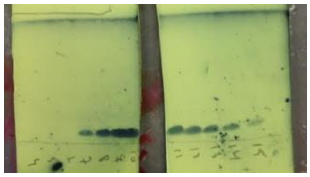


Figure 3.49. PMA stain, TLC (10% v/v EtOAc in pentanes)

^{1}H NMR (400 MHz, $CDCl_3$): δ = 4.46-4.31 (m, 2H), 3.99-3.85 (ABq, J_{AB}=16.8 Hz, 2H) 3.36 (t, *J*=8.2 Hz, 1H) 1.43 (s, 9H)

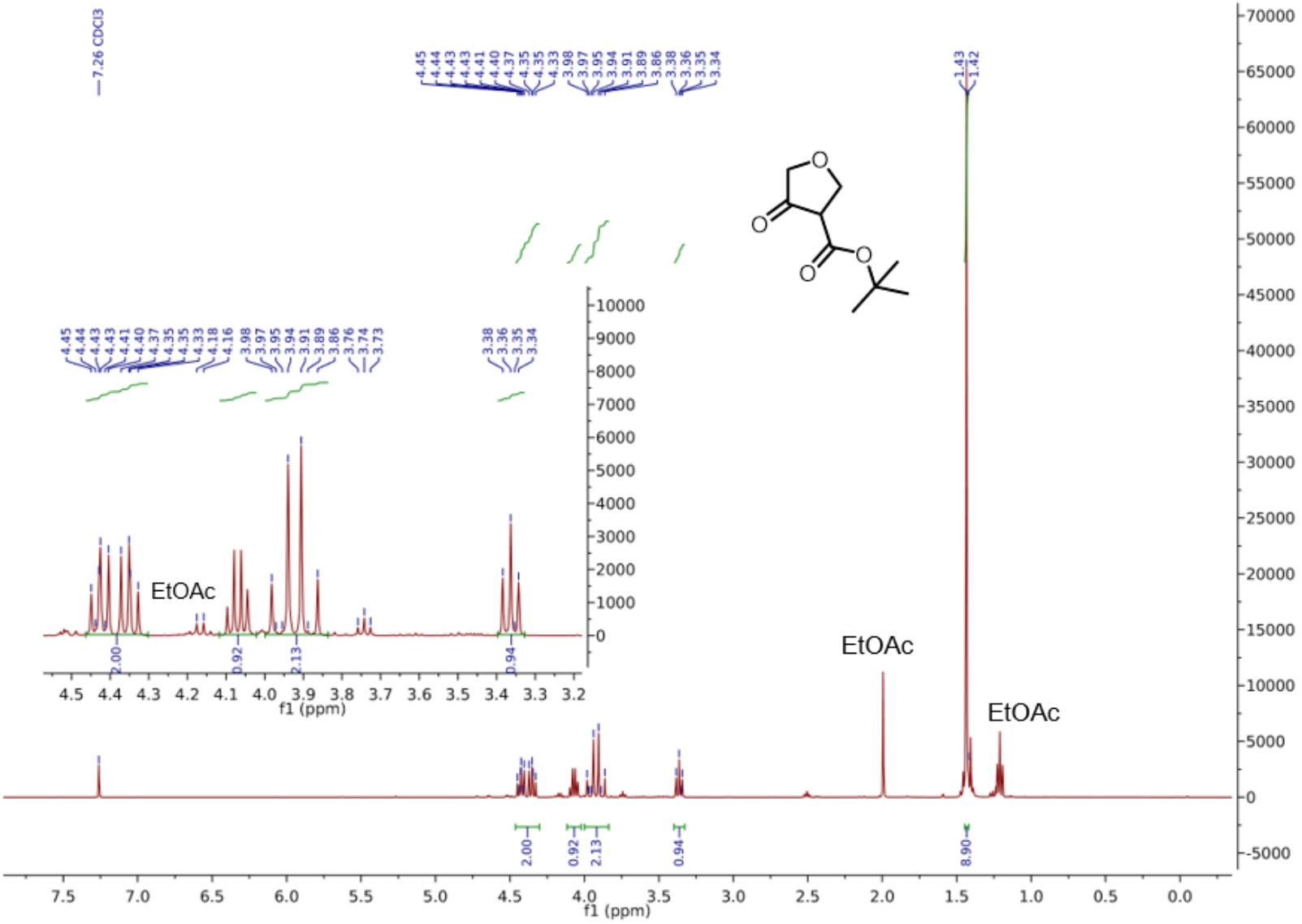
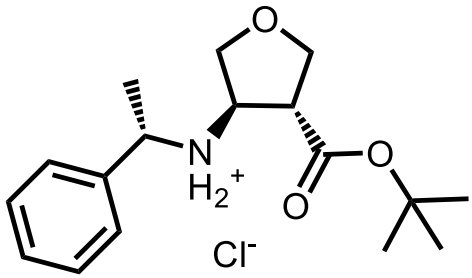


Figure 3.50. ^{1}H NMR spectrum of ATC β-ketoester.



ATC β aminoester

Formula: $C_{17}H_{26}ClNO_3$

IUPAC: *tert*-butyl (3*R*,4*R*)-4-{[(1S)-1-phenylethyl]amino}oxolane-3-carboxylate

SMILES: C[C@H](N[C@H]1COC[C@@H]1C(=O)OC(C)(C)C)c2ccccc2

Procedure: A 500 mL round bottom was charged with *tert*-butyl 4-oxooxolane-3-carboxylate (8.00 g, 4.30 mmol, 1.0 eq.) and the oily solid was dissolved in absolute ethanol (200 mL, 2.15 M). While the mixture was stirred, (*S*)-(-)-methylbenzylamine was added by pipet (11.40 mL, 8.60 mmol, 2 eq.) followed by acetic acid (4.91 mL, 8.60 mmol, 2.0 eq.). The reaction was stirred at room temperature while imine formation was monitored via TLC (10% v/v EtOAc in hexanes; imine R_f ~ 0.6, β-ketoester R_f ~ 0.4, iodine or PMA stain; imine formation was complete in approximately 4 h under these conditions). Imine formation was confirmed, sodium cyanoborohydride (10.80 g, 170 mmol, 4.0 eq.) was added portion wise over 2 min, and then the mixture was stirred overnight (~14 h). Ethanol was removed slowly via rotary evaporation, and then the slurry was diluted with water (200 mL, millipore). The aqueous solution was extracted with diethyl ether (3x100 mL). The combined organic layers were washed with brine (1x25 mL) and then dried over sodium sulfate. The dried organic phase was then concentrated slowly to approximately one-half volume.

Precipitation: In a 500 mL round-bottom flask, the organic phase was cooled on ice. HCl gas was generated via dropwise addition of H_2SO_4 to NaCl in the apparatus shown in

Figure 3.51 and passed over/through the ethereal amine solution. (Another commonly employed procedure is the addition of a dioxane/HCl solution; however, the product may be marginally soluble as increasing amounts of dioxane are added, so results may vary. Also conc. aqueous HCl can be used in lieu of NaCl). Passing gaseous HCl over/through the solution generated a white precipitate. HCl formation was ceased once the white precipitate stopped forming. (If the ammonium chloride 'oils out', cease HCl flow once oil layer stops forming). One quarter-volume of ether was added to the precipitate-laden flask and cooled to -20°C for 2-3 h. The crude white precipitate was filtered and dried overnight under hi-vacuum.

Recrystallization: The crude precipitate was dissolved in a minimal amount of acetonitrile (HPLC-grade; approximately 50 mL per 10 g of crude solid) with gentle heating. The solution was stored at -20°C overnight after which white crystals had formed. The white crystalline solid was filtered and dried under vacuum (2.35 g, 17% yield).

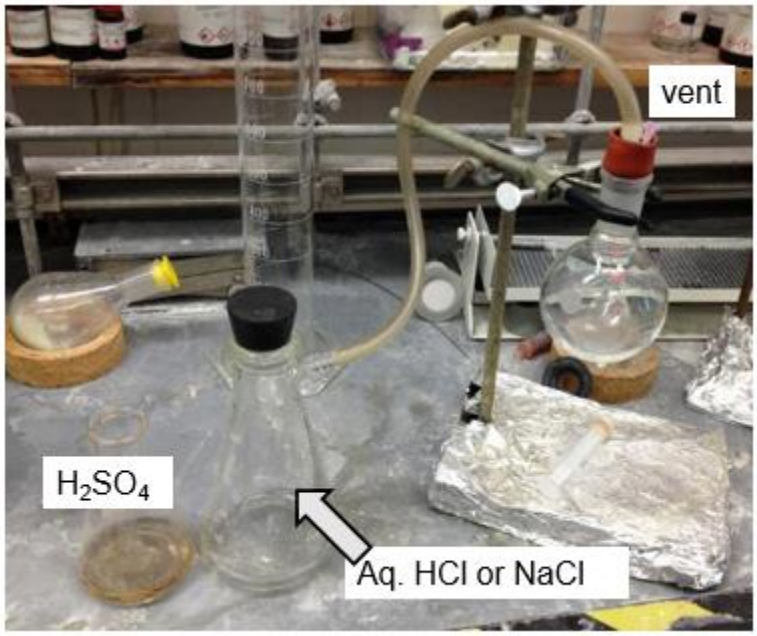


Figure 3.51. Gaseous HCl can be generated in a vacuum flask and passed through tygon tubing into the vented round-bottom to precipitate the ammonium chloride salt of the orthogonally protected ATC intermediate.

^{1}H NMR (400 MHz, $CDCl_3$): δ = 7.66-7.61 (m, 2H), 7.46-7.36 (m, 3H), 4.44-4.27 (m, 3H), 3.79 (dd, *J*=8.9, 5.1 Hz, 1H), 3.73-3.61 (m, 3H), 1.96 (d, *J*=6.8 Hz, 3H), 1.42 (s, 9H)

^{13}C NMR (400 MHz, $CDCl_3$): δ = 170.33, 135.87, 129.67, 129.61, 128.27, 82.46, 71.12, 70.69, 59.08, 58.86, 48.74, 28.07, 28.04, 20.99

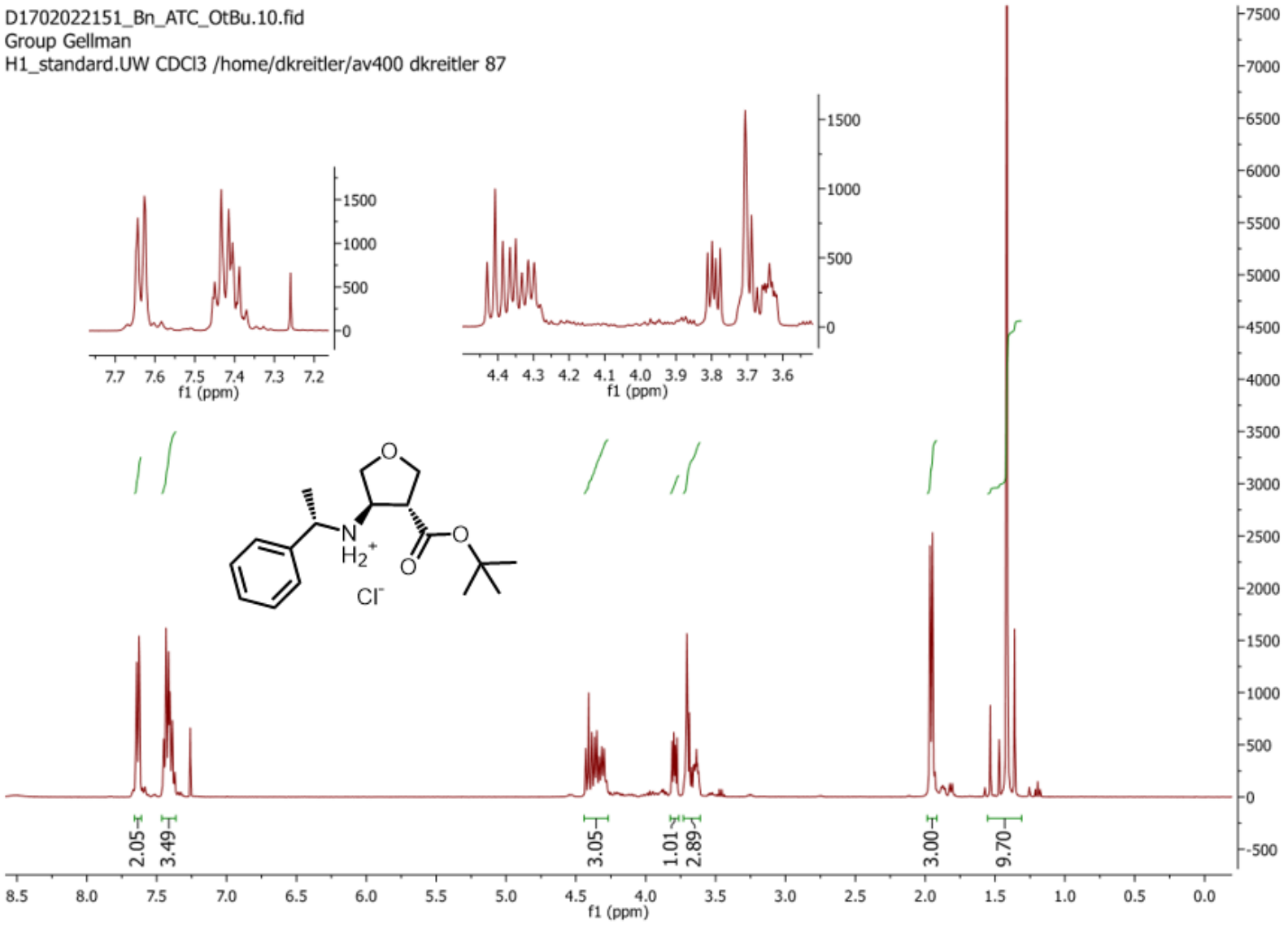


Figure 3.52. ^{1}H-NMR spectrum of ATC β-aminoketoester.

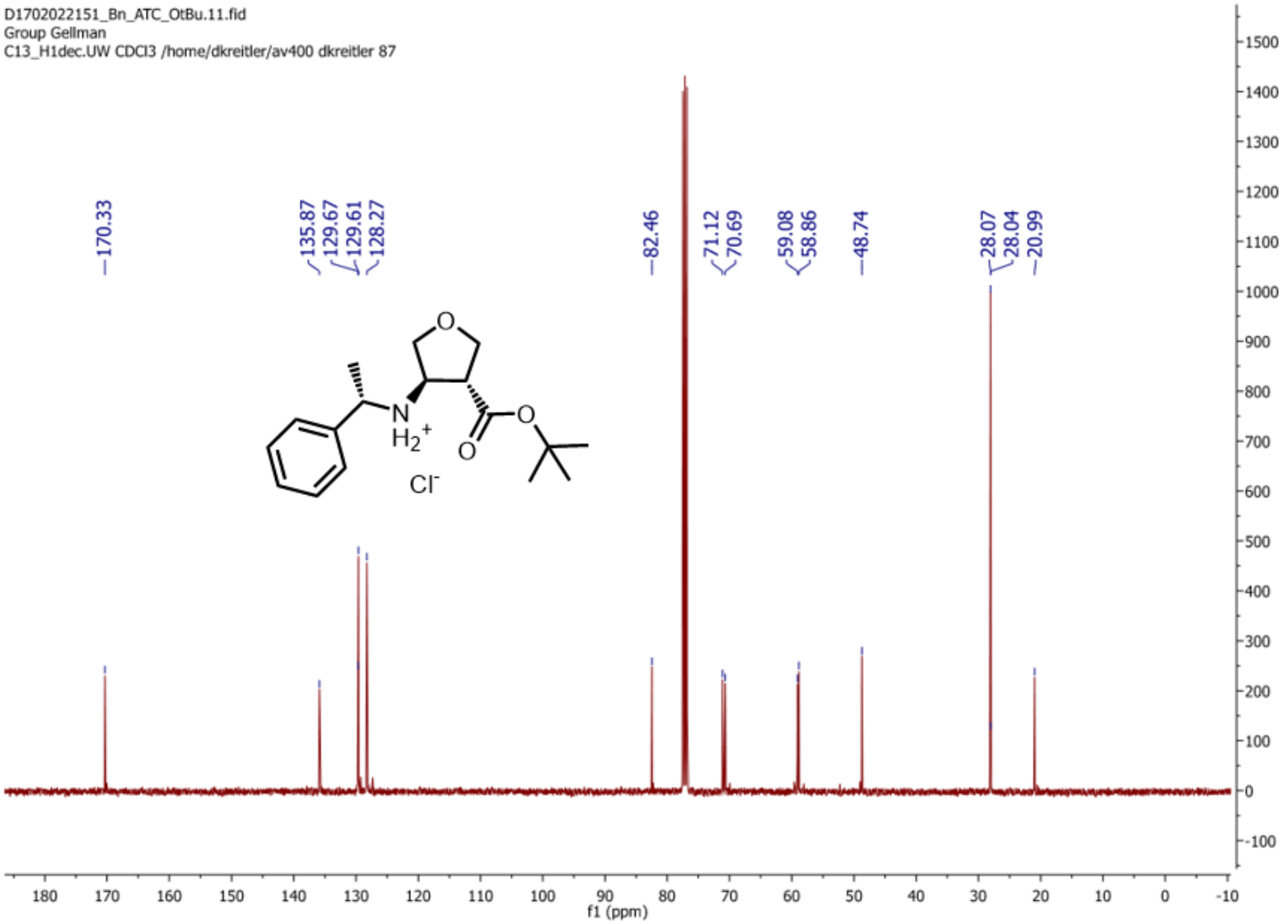
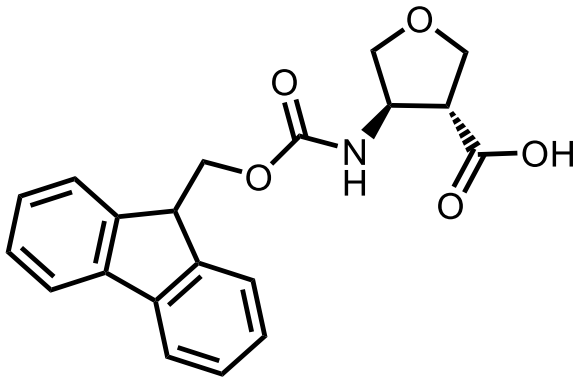


Figure 3.53. ^{13}C-NMR spectrum of ATC β-aminoketoester.



Fmoc-*trans*-(3*R*, 4*S*)-aminotetrahydrofuran carboxylic acid

Formula: $C_{20}H_{19}NO_5$

IUPAC: (3*R*,4*R*)-4-({[(9H-fluoren-9-yl)methoxy]carbonyl}amino)oxolane-3-carboxylic acid

SMILES: O=C(O)[C@H]1COC[C@@H]1NC(=O)OCC4c2ccccc2c3ccccc34

Procedure: This procedure is a series of serial protecting group manipulations with no intermediate purification steps:

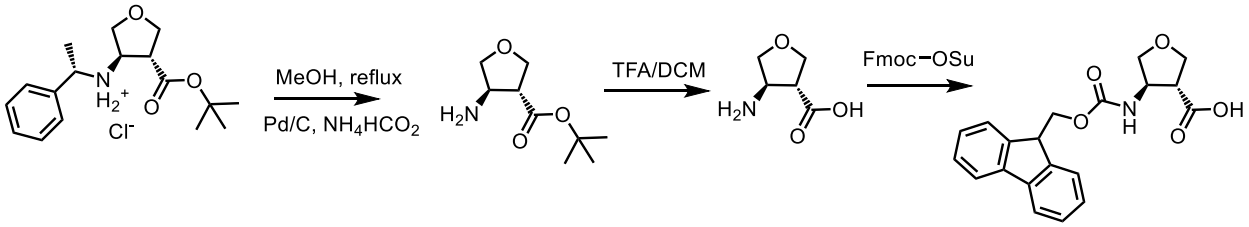


Figure 3.54. The first step is removal of the chiral auxiliary via reductive hydrogenation, followed by ester hydrolysis, and finally Fmoc protection.

<u>*Debenzylation*</u>*:* A 250 mL round-bottom flask was charged with *tert*-butyl (3*R*,4*R*)-4-{[(1S)-1-phenylethyl]amino}oxolane-3-carboxylate (1.7 g, 5.19 mmol, 1.0 eq.) and subsequently dissolved in methanol (50 mL, 0.1 M). 10% w/w Pd/C (0.42 g) was added slowly as a slurry in MeOH, and then ammonium formate (2.58 g, 4.08 mmol) was added portion wise. The flask was outfitted with a reflux condenser and the solution refluxed overnight. Finally, the solution was filtered through celite and concentrated to afford an oily solid residue.

Ester Hydrolysis: The solid residue (1.10 g) from the previous step was dissolved in 20 mL of 1:1 TFA/DCM and stirred for 5 h at room temperature. TFA/DCM was removed by slow evaporation under an N_2 stream to afford an oily residue (0.77 g).

Fmoc protection: The oily residue (0.77 g, 5.84 mmol, 1.0 eq.) from the previous step was placed under a hi-vacuum line, outfitted with a KOH trap, for ~30 min. The solid residue was then dissolved in chilled water:acetone (1:1 v/v, 150 mL) in a 250 mL round-bottom flask to which sodium bicarbonate (2.45 g, 29.2 mmol, 5 eq.) was added portion-wise. Once effervescence had ceased, Fmoc-OSu (2.56 g, 7.60 mmol, 1.3 eq.) was added in one lot, and the resulting heterogeneous solution was allowed to warm to room temperature and stir overnight.

Purification: The heterogeneous solution was diluted with water (50 mL, de-ionized), and acetone was removed via rotary evaporation. The aqueous solution was then washed with diethyl ether (3x75 mL). The aqueous solution was chilled on ice and the Fmoc-protected amino acid was then precipitated slowly via dropwise addition of 1M HCl (aq.). A white precipitate formed which was then extracted with EtOAc (2x100 mL; the precipitate is soluble in EtOAc). The organic layers were combined and washed with brine (1x50 mL) and dried over sodium sulfate. The solution was concentrated via rotary evaporation to afford a crude foamy solid. Finally, the crude solid was dissolved in a minimal amount of DCM and purified via silica chromatography (mobile phase, 7% v/v MeOH/DCM) to yield a white crystalline foam (0.4 g, 30% yield over three steps).

^{1}H NMR (500 MHz, $CDCl_3$): δ 7.75 (d, *J*=7.5 Hz, 2H), 7.58 (d, *J*=7.5 Hz, 2H), 7.39 (t, *J* 7.5 Hz, 2H), 7.30 (t, *J*=7.4 Hz, 2H), 4.47 (m, 1H), 4.41 (m, 2H), 4.18 (m, 1H), 4.11 (m, 1H), 3.97 (m, 2H), 3.67 (m, *J* = 6.5 Hz, 1H), 2.98 (m, 1H).

^{13}C NMR (500 MHz, $CDCl_3$): δ = 174.07, 156.44, 143.71, 141.24, 127.68, 127.03, 124.88, 119.91, 72.91, 69.69, 66.63, 55.13, 51.00, 47.11

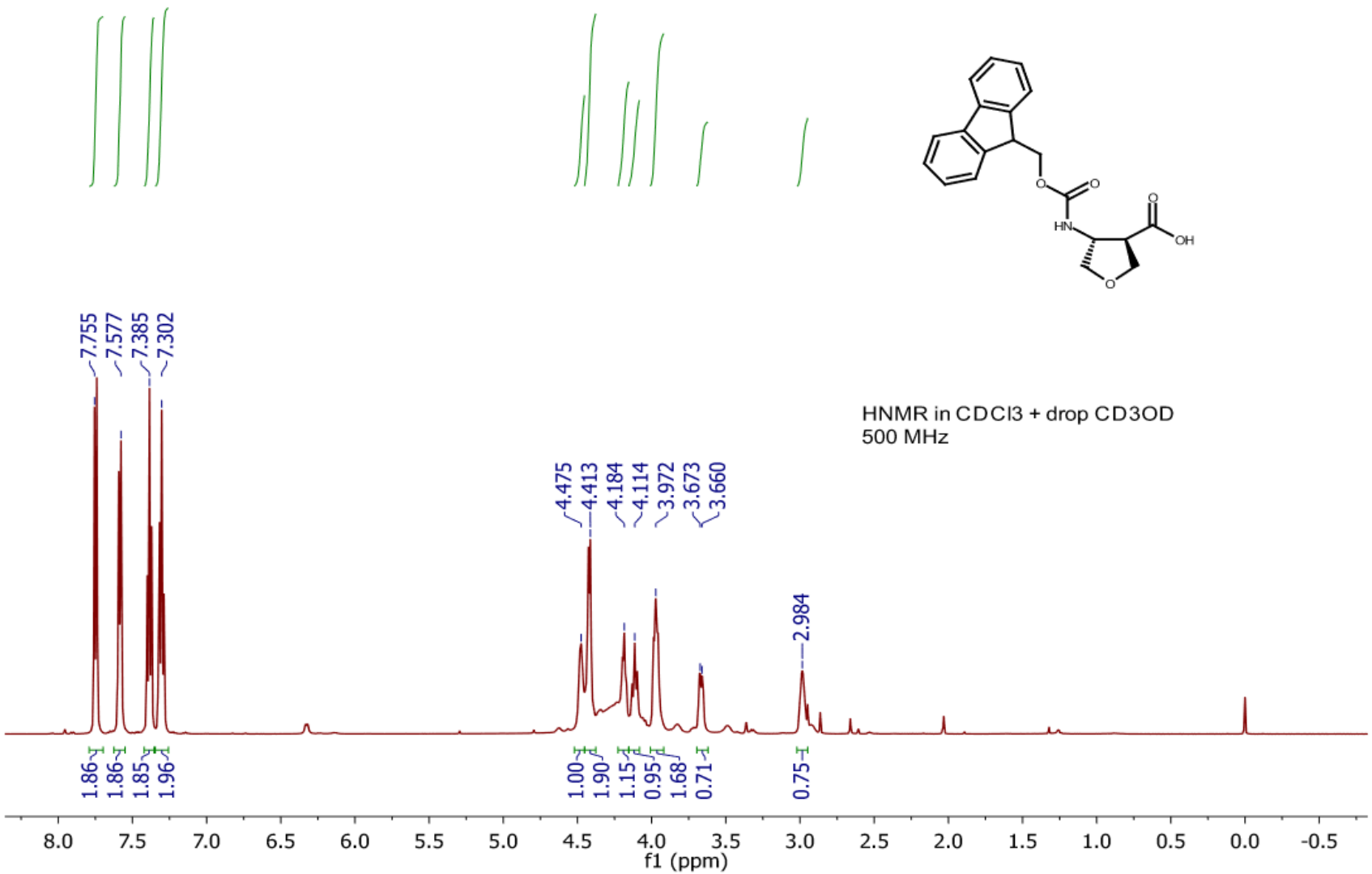


Figure 3.55. Fmoc-ATC-OH ^{1}H-NMR spectrum (spectrum recorded by Brian Fisher).

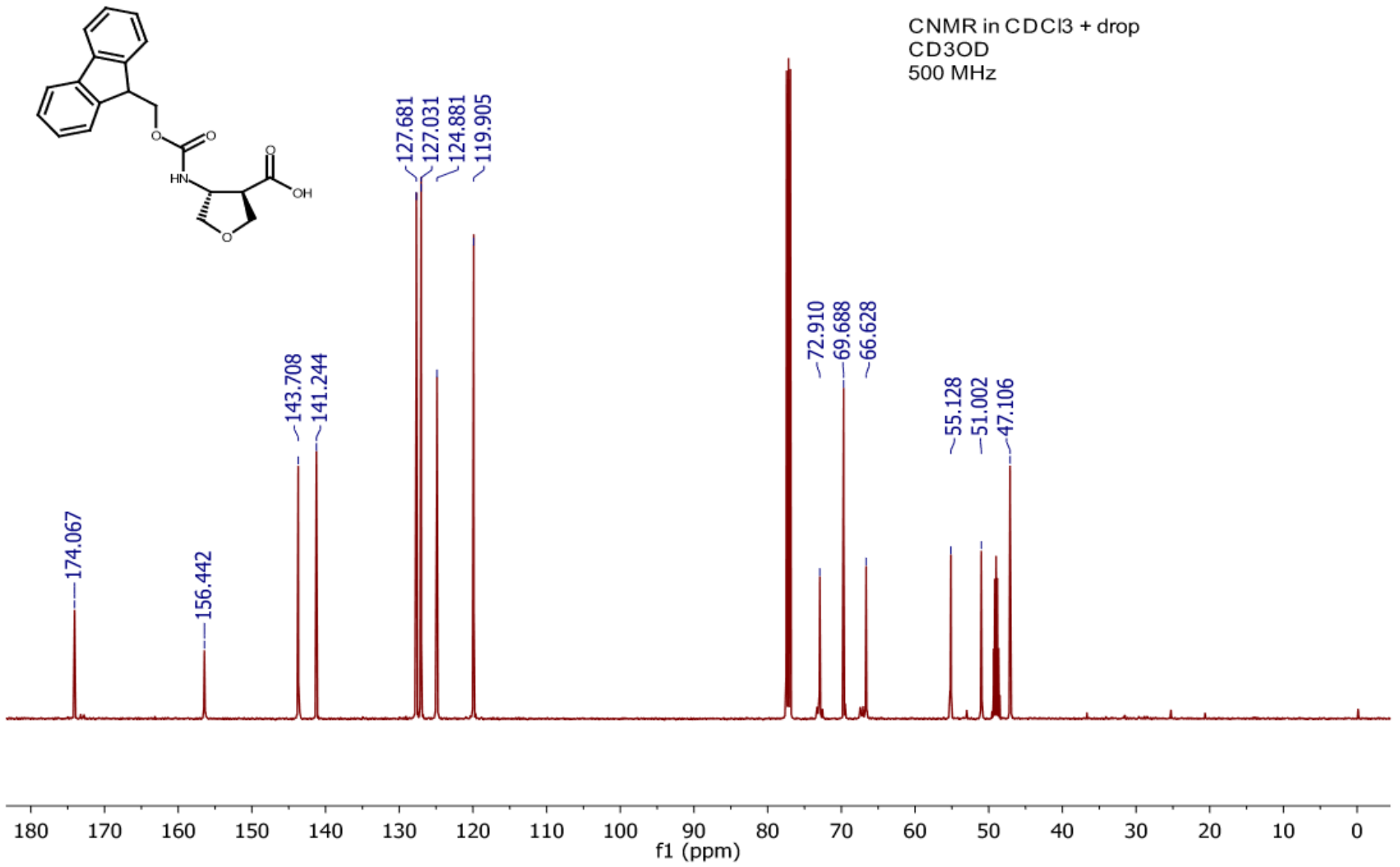


Figure 3.56. Fmoc-ATC-OH ^{13}C-NMR spectrum (spectrum recorded by Brian Fisher).

3.4.4 X-ray Data Collection

Data were collected on a Bruker Photon 100 diffractometer with Cu-K_α radiation (λ=1.54178 Å). Data were indexed and integrated with SAINT and scaling and absorption correction were performed in XPREP. Diffraction data were collected by Ilia Guzei and Kelsey Miles at Bruker AXS.

3.4.5 Structure Solution and Refinement

Systematic absences in the diffraction data were consistent with space group $P2_1$ and $P2_1/c$. *E*-statistics indicated that the non-centrosymmetric space group $P2_1$ was the appropriate space group which is also consistent with the hypothesis that the crystals were comprised of the enantiomerically pure protected amino acid. The structure was solved using the dual space method in SHELXD. Non-hydrogen atoms were manually built in using difference Fourier maps and iterative rounds of refinement in SHELXL. The solution yielded a chemically reasonable structure that resulted in stable refinement.

All non-hydrogen atoms were refined with anisotropic atomic displacement parameters. There is positional disorder in the structure mainly corresponding to rotational disorder from the *tert*-butyl group around the O1-C2 bond and there are also slight deviations from sp2 planarity around the C5-O2 bond, presumably a result of the rotational disorder of the *tert*-butyl group (Figure 2).

Analysis of the cumulative intensity distribution suggested the presence of twinning. Refinement with a twin law consistent with pseudomerohedral twinning (-1,0,0;0,-1,0;0,0,1) resulted in a significant reduction of residuals *R* and *wR*.

The proposed (*S,S*) relative configuration of the enantiomerically pure protected amino acid was confirmed by the correct configuration of the (*S*)-methylbenzylamine moiety. The Flack parameter of 0.069 also corroborates the proposed absolute configuration. The final least-squares refinement of 450 parameters against 6913 data resulted in residuals *R* (based on F^2 for $I \geq 2\sigma$) and *wR* (based on F^2 for all data) of 0.0603 and 0.1407, respectively.

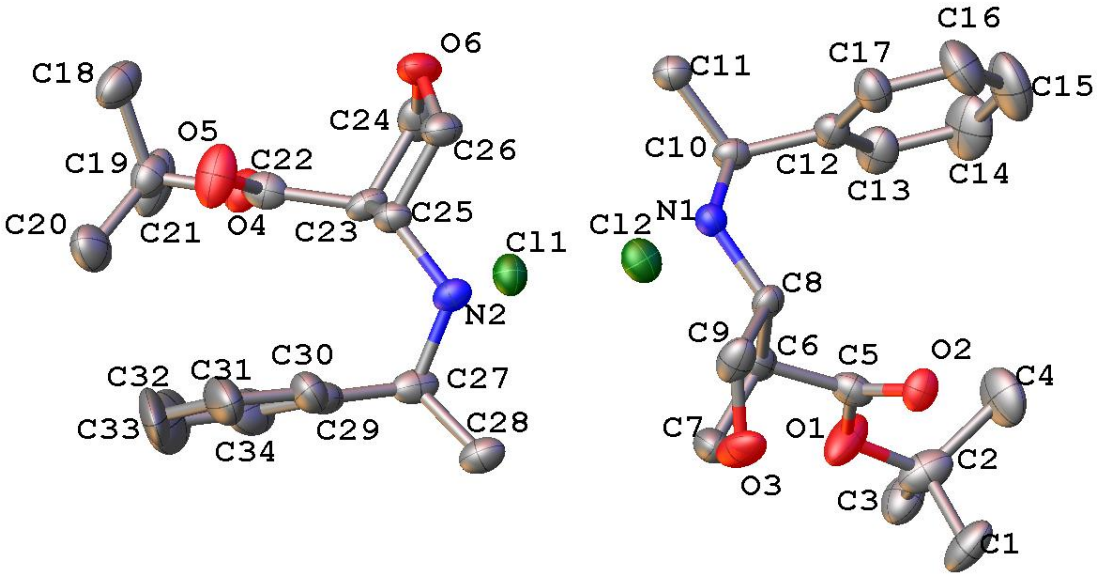


Figure 3.57. β aminoester **13** with thermal ellipsoids drawn at 50% probability. Positional disorder is omitted for clarity.

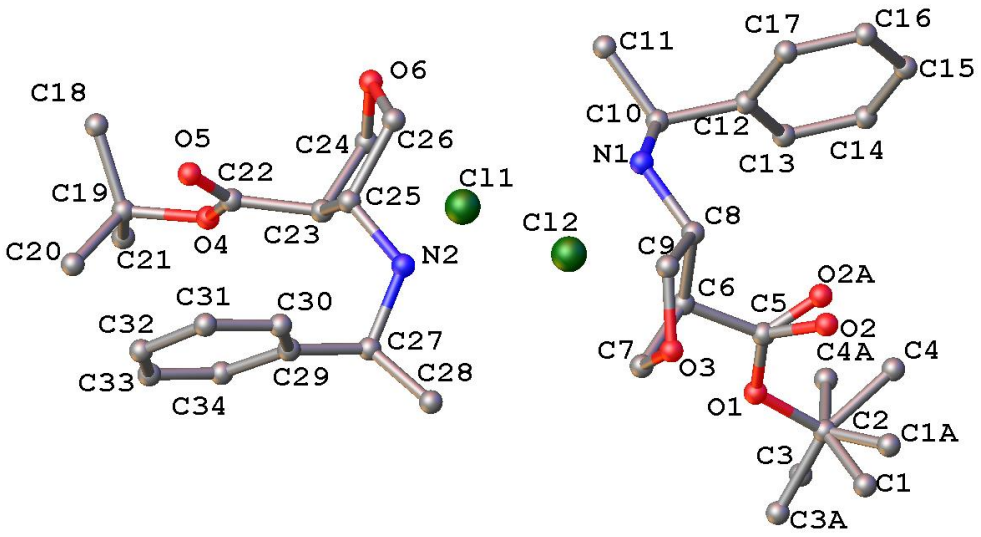


Figure 3.58. Positional disorder around one of the tert-butyl protecting groups in the asymmetric unit.

Table 3.3. Crystal data and structure refinement for β aminoester 13

Identification code	Gellman163
Empirical formula	$C_{17}H_{26}NO_3Cl$
Formula weight	326.60
Temperature/K	100.01
Crystal system	Monoclinic
Space group	$P2_1$
a/Å	7.1944(14)
b/Å	28.739(5)
c/Å	8.8893(18)
α/°	90
β/°	90.210(7)
γ/°	90
Volume/Å^3	1838.0(6)
Z	4
ρ_{calc}g/cm^3	1.1802
μ/mm^{-1}	1.929
F(000)	704.7
Radiation	Cu Kα (λ = 1.54178)
2Θ range for data collection/°	6.16 to 140.42
Index ranges	-8 ≤ h ≤ 8, -35 ≤ k ≤ 35, -10 ≤ l ≤ 10
Reflections collected	18150
Independent reflections	6913 [R_{int} = 0.0673, R_{sigma} = 0.0744]

Data/restraints/parameters	6913/0/450
Goodness-of-fit on F^2	1.048
Final R indexes [I>=2σ (I)]	R_1 = 0.0603, wR_2 = 0.1407
Final R indexes [all data]	R_1 = 0.0760, wR_2 = 0.1540
Largest diff. peak/hole / e $Å^{-3}$	0.41/-0.27
Flack parameter	0.069(11)

Table 3.4. Fractional Atomic Coordinates (×10^4) and Equivalent Isotropic Displacement Parameters ($Å^2$×10^3) for β aminoester 13. U_{eq} is defined as 1/3 of of the trace of the orthogonalised U_{IJ} tensor.

Atom	*x*	*y*	*z*	U(eq)
O3	-7600(6)	-2751.3(14)	-7547(4)	51.3(10)
N1	-7628(5)	-3012.6(13)	-3652(4)	26.4(8)
C12	-6974(7)	-3819.9(17)	-2706(5)	32.7(10)
C8	-7501(6)	-3158.4(16)	-5262(5)	28.1(9)
C5	-5014(6)	-3573.6(17)	-6715(5)	33.5(10)
C6	-5538(6)	-3129.2(16)	-5913(5)	30.4(10)
C9	-8661(7)	-2821.3(18)	-6235(6)	41.9(13)
C10	-6580(6)	-3302.0(16)	-2515(5)	29.2(10)
C13	-5502(8)	-4106(2)	-3013(6)	46.4(14)
C11	-7090(7)	-3117.6(18)	-965(5)	34.3(11)
C16	-9063(10)	-4467(2)	-2672(8)	65.3(18)
C17	-8744(7)	-3996.9(18)	-2534(6)	42.4(12)
C7	-5751(8)	-2722.2(18)	-7043(6)	44.2(13)
C14	-5837(11)	-4588(2)	-3159(9)	69(2)
C15	-7562(12)	-4767(3)	-2982(10)	80(2)
O5	-1176(5)	-991.9(14)	626(5)	56.6(11)
O4	1710(5)	-1276.7(12)	713(4)	37.4(8)
O6	-2491(5)	-2129.0(13)	1128(4)	44.0(9)
N2	-2642(5)	-1875.7(14)	-2763(4)	28.9(8)
C19	2517(7)	-881(2)	1548(6)	39.7(18)
C27	-2188(7)	-1763(2)	-5461(5)	39.9(12)
C18	1600(9)	-835(2)	3056(7)	55.1(19)
C21	2296(9)	-443(2)	638(7)	54.7(19)
C23	-646(7)	-2141.3(19)	614(6)	42.3(12)
C44	-519(6)	-1738.0(17)	-506(5)	30.7(10)
C24	-2490(6)	-1726.0(17)	-1161(5)	29.6(10)
C25	-3613(7)	-2063.9(18)	-153(5)	37.6(12)
C22	-49(6)	-1284.3(17)	316(5)	33.7(10)
C20	4506(8)	-1012(3)	1744(9)	64(2)

Cl2	-1819.7(15)	-2941.7(5)	-3024.4(14)	38.1(3)
Cl1	-6845.8(14)	-1948.5(5)	-3398.7(13)	33.4(2)
O1	-3331(5)	-3554.8(15)	-7239(5)	58.9(11)
C2	-2453(7)	-3952(3)	-8015(6)	50.6(15)
C1	-3250(20)	-3983(5)	-9584(16)	55(4)
C3	-383(15)	-3912(5)	-8024(14)	52(4)
C4	-2897(17)	-4450(4)	-7268(16)	62(4)
C52A	-806(19)	-3649(5)	-8833(17)	47(4)
C54A	-1643(19)	-4157(4)	-6694(16)	54(4)
C53A	-3540(30)	-4185(7)	-9000(20)	65(6)
C28	-2082(8)	-1073(2)	-3697(6)	46.2(14)
C29	-3889(9)	-903.0(19)	-3827(6)	47.5(13)
C30	-4241(14)	-436(2)	-3586(9)	81(3)
C31	-2750(30)	-148(3)	-3261(13)	132(6)
C32	-940(20)	-317(4)	-3112(14)	139(6)
C33	-648(12)	-767(3)	-3322(8)	81(3)
C26	-1670(6)	-1581(2)	-3910(5)	37.1(11)
O2	-6232(14)	-3843(5)	-7270(20)	33(4)
O9A	-5760(30)	-3922(3)	-6490(30)	60(7)

Table 3.5. Anisotropic Displacement Parameters ($Å^2×10^3$) for β aminoester 13. The Anisotropic displacement factor exponent takes the form: $-2π^2[h^2a^{*2}U_{11}+2hka^*b^*U_{12}+…]$

Atom	U_{11}	U_{22}	U_{33}	U_{12}	U_{13}	U_{23}
O3	75(3)	52(2)	27.1(19)	25(2)	-2.5(17)	7.8(17)
N1	28.2(17)	25(2)	26.3(18)	-0.7(15)	-1.1(13)	0.6(15)
C12	39(2)	30(3)	29(2)	5(2)	-4.8(19)	3.2(19)
C8	31(2)	29(2)	24(2)	-1.1(18)	-5.1(17)	-4.1(18)
C5	29(2)	35(3)	36(3)	3(2)	-0.6(19)	12(2)
C6	34(2)	32(2)	25(2)	-8.3(19)	-6.3(17)	-2.9(18)
C9	49(3)	40(3)	36(3)	14(2)	-15(2)	-6(2)
C10	27(2)	36(3)	24(2)	-0.5(18)	-3.6(17)	4.0(19)
C13	49(3)	41(3)	49(3)	13(3)	-12(2)	1(2)
C11	36(2)	37(3)	30(2)	1(2)	0.3(18)	1.8(19)
C16	66(4)	42(4)	88(5)	-14(3)	1(4)	11(3)
C17	47(3)	31(3)	49(3)	1(2)	-6(2)	4(2)
C7	64(3)	32(3)	36(3)	-7(2)	10(2)	3(2)
C14	68(4)	47(4)	93(5)	30(3)	-20(4)	-5(3)
C15	90(5)	34(4)	116(7)	4(4)	-22(5)	9(4)
O5	32.1(19)	54(2)	84(3)	11.4(17)	-11.9(18)	-27(2)

O4	31.3(16)	42(2)	39.0(19)	5.2(15)	-2.9(14)	-10.5(16)
O6	55(2)	51(2)	26.0(18)	-12.5(17)	-2.0(15)	6.5(16)
N2	26.1(17)	37(2)	23.3(18)	5.2(16)	-3.0(13)	-0.4(16)
C19	35(3)	43(3)	41(3)	-4(2)	-2(2)	-9(2)
C27	39(3)	52(3)	28(2)	7(2)	-0.6(19)	0(2)
C18	50(3)	70(4)	45(3)	-14(3)	-1(3)	-19(3)
C21	62(4)	47(4)	55(4)	-20(3)	-8(3)	1(3)
C23	48(3)	39(3)	40(3)	1(2)	-18(2)	1(2)
C44	26(2)	45(3)	21(2)	4.7(19)	-1.6(16)	-2.4(19)
C24	31(2)	34(3)	24(2)	3.4(19)	-0.4(17)	3.3(18)
C25	39(3)	45(3)	29(2)	-5(2)	2.2(19)	1(2)
C22	21(2)	43(3)	38(3)	5.0(19)	-0.6(18)	7(2)
C20	33(3)	68(5)	90(5)	2(3)	-15(3)	-40(4)
Cl2	25.2(5)	39.4(7)	49.7(7)	-3.0(5)	-4.5(4)	2.1(5)
Cl1	26.5(5)	30.3(6)	43.3(6)	-1.1(5)	0.1(4)	-3.8(5)
O1	43(2)	77(3)	57(2)	-22(2)	18.3(18)	-36(2)
C2	35(3)	77(4)	39(3)	13(3)	-1(2)	-7(3)
C1	59(8)	66(10)	41(8)	27(8)	-9(6)	-24(6)
C3	44(6)	55(8)	56(7)	3(5)	8(5)	-17(6)
C4	57(7)	36(6)	94(10)	19(5)	4(6)	4(6)
C52A	52(7)	38(8)	50(8)	0(6)	24(6)	-15(6)
C54A	48(7)	45(8)	69(9)	17(6)	3(6)	-12(6)
C53A	75(11)	56(12)	63(13)	-2(9)	13(10)	-41(10)
C28	61(3)	52(4)	26(3)	-24(3)	2(2)	2(2)
C29	66(4)	31(3)	45(3)	-6(3)	1(3)	4(2)
C30	138(8)	32(4)	72(5)	-7(4)	27(5)	1(3)
C31	253(16)	35(5)	107(8)	-64(8)	61(10)	-20(4)
C32	201(14)	102(10)	115(9)	-112(10)	51(10)	-41(7)
C33	97(6)	86(6)	60(4)	-63(5)	15(4)	-8(4)
C26	24(2)	60(3)	28(2)	-6(2)	-1.0(18)	2(2)
O2	24(5)	36(6)	38(8)	4(3)	-8(4)	-8(5)
O9A	58(7)	26(4)	98(15)	-2(4)	32(9)	-9(5)

Table 3.6. Table 4 Bond Lengths for β aminoester 13.

Atom	Atom	Length/Å	Atom	Atom	Length/Å
O3	C9	1.411(6)	N2	C24	1.492(5)
O3	C7	1.404(7)	N2	C26	1.500(6)
N1	C8	1.494(5)	C19	C18	1.502(8)
N1	C10	1.509(5)	C19	C21	1.504(8)
C12	C10	1.524(7)	C19	C20	1.490(8)

C12	C13	1.370(7)	C27	C26	1.520(7)
C12	C17	1.381(7)	C23	C44	1.531(7)
C8	C6	1.531(6)	C44	C24	1.531(6)
C8	C9	1.542(6)	C44	C22	1.532(7)
C5	C6	1.511(7)	C24	C25	1.550(6)
C5	O1	1.300(6)	O1	C2	1.477(7)
C5	O2	1.267(11)	C2	C1	1.508(14)
C5	O9A	1.154(9)	C2	C3	1.494(12)
C6	C7	1.549(7)	C2	C4	1.609(13)
C10	C11	1.522(6)	C2	C52A	1.642(13)
C13	C14	1.410(9)	C2	C54A	1.435(15)
C16	C17	1.375(8)	C2	C53A	1.35(2)
C16	C15	1.411(11)	C28	C29	1.394(8)
C14	C15	1.354(12)	C28	C33	1.396(8)
O5	C22	1.201(6)	C28	C26	1.502(8)
O4	C19	1.477(6)	C29	C30	1.382(9)
O4	C22	1.313(5)	C30	C31	1.384(16)
O6	C23	1.406(7)	C31	C32	1.40(2)
O6	C25	1.407(6)	C32	C33	1.323(16)

Table 3.7. Table 5 Bond Angles for β aminoester 13.

Atom	Atom	Atom	Angle/˚	Atom	Atom	Atom	Angle/˚
C7	O3	C9	105.0(4)	C24	C44	C23	101.9(4)
C10	N1	C8	117.1(3)	C22	C44	C23	110.3(4)
C13	C12	C10	117.8(5)	C22	C44	C24	111.4(4)
C17	C12	C10	121.2(4)	C44	C24	N2	114.9(4)
C17	C12	C13	121.0(5)	C25	C24	N2	109.5(4)
C6	C8	N1	114.0(3)	C25	C24	C44	104.5(4)
C9	C8	N1	109.1(4)	C24	C25	O6	104.6(4)
C9	C8	C6	104.6(4)	O4	C22	O5	125.3(5)
O1	C5	C6	111.6(4)	C44	C22	O5	123.8(4)
C5	C6	C8	111.4(4)	C44	C22	O4	110.7(4)
C7	C6	C8	101.4(4)	C2	O1	C5	122.4(4)
C7	C6	C5	110.9(4)	C3	C2	C1	112.0(8)
C8	C9	O3	105.1(4)	C4	C2	C1	104.8(9)
C12	C10	N1	111.8(3)	C4	C2	C3	105.7(8)
C11	C10	N1	107.0(4)	C54A	C2	C52A	106.7(9)
C11	C10	C12	113.3(4)	C53A	C2	C52A	113.1(10)
C14	C13	C12	118.4(6)	C53A	C2	C54A	124.1(12)
C15	C16	C17	119.4(6)	C33	C28	C29	119.1(7)

C16	C17	C12	120.4(5)	C26	C28	C29	121.0(5)
C6	C7	O3	104.7(4)	C26	C28	C33	119.8(6)
C15	C14	C13	121.4(7)	C30	C29	C28	119.9(7)
C14	C15	C16	119.4(7)	C31	C30	C29	118.2(10)
C25	O6	C23	106.2(4)	C32	C31	C30	122.2(9)
C26	N2	C24	117.0(4)	C33	C32	C31	118.3(10)
C21	C19	C18	111.2(5)	C32	C33	C28	122.3(11)
C20	C19	C18	110.0(5)	C27	C26	N2	108.0(4)
C20	C19	C21	112.0(6)	C28	C26	N2	111.7(4)
C44	C23	O6	104.6(4)	C28	C26	C27	113.6(4)

Table 3.8. Hydrogen Atom Coordinates (Å×104) and Isotropic Displacement Parameters (Å2×103) for β aminoester 13

Atom	***x***	***y***	***z***	**U(eq)**
H1a	-7208(5)	-2711.1(13)	-3580(4)	31.6(9)
H1b	-8862(5)	-3011.8(13)	-3385(4)	31.6(9)
H8	-7991(6)	-3482.1(16)	-5378(5)	33.7(11)
H6	-4609(6)	-3051.5(16)	-5112(5)	36.5(12)
H9a	-8863(7)	-2523.4(18)	-5700(6)	50.3(15)
H9b	-9885(7)	-2959.2(18)	-6486(6)	50.3(15)
H10	-5220(6)	-3250.0(16)	-2670(5)	35.0(11)
H13	-4284(8)	-3984(2)	-3125(6)	55.7(16)
H11a	-6770(50)	-2787(4)	-899(15)	51.4(16)
H11b	-8428(11)	-3156(12)	-804(17)	51.4(16)
H11c	-6400(40)	-3291(9)	-193(6)	51.4(16)
H16	-10283(10)	-4588(2)	-2561(8)	78(2)
H17	-9748(7)	-3792.9(18)	-2319(6)	50.9(14)
H7a	-4875(8)	-2757.3(18)	-7892(6)	53.1(15)
H7b	-5516(8)	-2419.9(18)	-6543(6)	53.1(15)
H14	-4832(11)	-4790(2)	-3386(9)	83(2)
H15	-7760(12)	-5093(3)	-3065(10)	96(3)
H4	2150(70)	-1547(10)	700(200)	56.0(12)
H2a	-2184(5)	-2173.7(14)	-2835(4)	34.7(10)
H2b	-3883(5)	-1887.2(14)	-3014(4)	34.7(10)
H27a	-1920(50)	-2097(3)	-5517(15)	59.8(18)
H27b	-3517(13)	-1712(12)	-5642(18)	59.8(18)
H27c	-1470(40)	-1598(10)	-6225(6)	59.8(18)
H18a	1730(60)	-1128(6)	3610(20)	83(3)

H18b	2190(50)	-584(12)	3630(20)	83(3)
H18c	279(18)	-764(18)	2915(7)	83(3)
H21a	2830(60)	-489(6)	-363(19)	82(3)
H21b	973(10)	-368(9)	540(40)	82(3)
H21c	2940(60)	-186(4)	1150(30)	82(3)
H23a	239(7)	-2097.3(19)	1457(6)	50.7(15)
H23b	-375(7)	-2441.5(19)	114(6)	50.7(15)
H44	415(6)	-1805.8(17)	-1307(5)	36.9(12)
H24	-3007(6)	-1404.7(17)	-1055(5)	35.6(12)
H25a	-3834(7)	-2363.1(18)	-677(5)	45.1(14)
H25b	-4826(7)	-1926.4(18)	124(5)	45.1(14)
H20a	4585(8)	-1322(8)	2200(60)	95(4)
H20b	5120(20)	-1017(18)	761(11)	95(4)
H20c	5120(20)	-785(11)	2400(50)	95(4)
H1c	-4580(30)	-4050(30)	-9529(16)	83(6)
H1d	-3050(110)	-3687(12)	-10110(40)	83(6)
H1e	-2630(90)	-4230(20)	-10140(40)	83(6)
H3a	-29(15)	-3592(8)	-8270(90)	78(5)
H3b	109(17)	-3990(30)	-7030(30)	78(5)
H3c	130(16)	-4120(20)	-8780(70)	78(5)
H4a	-2640(120)	-4437(10)	-6190(20)	94(6)
H4b	-4210(40)	-4527(15)	-7430(90)	94(6)
H4c	-2110(90)	-4689(6)	-7730(80)	94(6)
H52a	-1367(19)	-3420(20)	-9500(80)	70(6)
H52b	-50(80)	-3490(30)	-8064(17)	70(6)
H52c	-20(80)	-3858(6)	-9420(80)	70(6)
H54a	-1380(140)	-3913(6)	-5950(50)	81(6)
H54b	-2510(60)	-4380(30)	-6260(70)	81(6)
H54c	-490(80)	-4320(30)	-6960(30)	81(6)
H53a	-4060(140)	-3966(10)	-9740(90)	97(8)
H53b	-2800(50)	-4420(30)	-9520(100)	97(8)
H53c	-4560(110)	-4340(40)	-8460(30)	97(8)
H29	-4879(9)	-1107.6(19)	-4081(6)	56.9(16)
H30	-5470(14)	-317(2)	-3641(9)	97(3)
H31	-2970(30)	176(3)	-3135(13)	158(7)
H32	60(20)	-114(4)	-2865(14)	167(7)
H33	578(12)	-885(3)	-3215(8)	97(3)
H26	-301(6)	-1627(2)	-3772(5)	44.6(14)

Table 3.9. Atomic Occupancy for β aminoester 13.

Atom	*Occupancy*	Atom	*Occupancy*	Atom	*Occupancy*
H4	0.32(11)	C19	1.01(2)	C18	0.936(11)
H18a	0.936(11)	H18b	0.936(11)	H18c	0.936(11)
C21	0.936(11)	H21a	0.936(11)	H21b	0.936(11)
H21c	0.936(11)	C20	0.936(11)	H20a	0.936(11)
H20b	0.936(11)	H20c	0.936(11)	C1	0.557(14)
H1c	0.557(14)	H1d	0.557(14)	H1e	0.557(14)
C3	0.557(14)	H3a	0.557(14)	H3b	0.557(14)
H3c	0.557(14)	C4	0.557(14)	H4a	0.557(14)
H4b	0.557(14)	H4c	0.557(14)	C52A	0.443(14)
H52a	0.443(14)	H52b	0.443(14)	H52c	0.443(14)
C54A	0.443(14)	H54a	0.443(14)	H54b	0.443(14)
H54c	0.443(14)	C53A	0.443(14)	H53a	0.443(14)
H53b	0.443(14)	H53c	0.443(14)	O2	0.42(4)
O9A	0.58(4)				

3.5 References

(1) Reichert, J. M. *Curr. Pharm. Biotechnol.* **2008**, *9* (January 2009), 423–430.

(2) Nelson, A. L.; Dhimolea, E.; Reichert, J. M. *Nat. Rev. Drug Discov.* **2010**, *9* (10), 767–774.

(3) Chi, E. Y.; Krishnan, S.; Randolph, T. W.; Carpenter, J. F. *Pharm. Res.* **2003**, *20* (9), 1325–1336.

(4) Roberts, C. J. *Trends Biotechnol.* **2014**, *32* (7), 372–380.

(5) Frokjaer, S.; Otzen, D. E. *Nat. Rev. Drug Discov.* **2005**, *4* (4), 298–306.

(6) Fedorova, A.; Zobel, K.; Gill, H. S.; Ogasawara, A.; Flores, J. E.; Tinianow, J. N.; Vanderbilt, A. N.; Wu, P.; Meng, Y. G.; Williams, S. P.; Wiesmann, C.; Murray, J.; Marik, J.; Deshayes, K. *Chem. Biol.* **2011**, *18* (7), 839–845.

(7) Checco, J. W.; Kreitler, D. F.; Thomas, N. C.; Belair, D. G.; Rettko, N. J.; Murphy, W. L.; Forest, K. T.; Gellman, S. H. *Proc. Natl. Acad. Sci.* **2015**, *112* (15), 4552–4557.

(8) Starovasnik, M. A.; Braisted, A. C.; Wells, J. A. *Proc. Natl. Acad. Sci. U. S. A.* **1997**, *94* (19), 10080–10085.

(9) Hamley, I. W. *Biomacromolecules* **2014**, *15* (5), 1543–1559.

(10) Cheloha, R. W.; Maeda, A.; Dean, T.; Gardella, T. J.; Gellman, S. H. *Nat. Biotechnol.* **2014**, *32* (7), 653–655.

(11) Doig, A. J.; Chakrabartty, A.; Klingler, T. M.; Baldwin, R. L. *Biochemistry* **1994**, *33* (11), 3396–3403.

(12) Johnson, L. M.; Gellman, S. H. *Methods Enzymol.* **2013**, *523*, 407–429.

(13) Horne, W. S.; Price, J. L.; Gellman, S. H. *Proc. Natl. Acad. Sci. U. S. A.* **2008**, *105* (27), 9151–9156.

(14) Price, J. L.; Horne, W. S.; Gellman, S. H. *J. Am. Chem. Soc.* **2010**, *132* (35), 12378–12387.

(15) Kreitler, D. F.; Mortenson, D. E.; Forest, K. T.; Gellman, S. H. *J. Am. Chem. Soc.* **2016**, *138* (20), 6498–6505.

(16) Araya, C. L.; Fowler, D. M.; Chen, W.; Muniez, I.; Kelly, J. W.; Fields, S. *Proc. Natl. Acad. Sci.* **2012**, *109* (42), 16858–16863.

(17) Moriarty, N. W.; Grosse-Kunstleve, R. W.; Adams, P. D. *Acta Crystallogr. Sect. D Biol. Crystallogr.* **2009**, *65* (10), 1074–1080.

(18) Bunnage, M. E.; Davies, S. G.; Roberts, P. M.; Smith, D.; Withey, J. M. *Org. Biomol. Chem.* **2004**, *2*, 2763–2776.

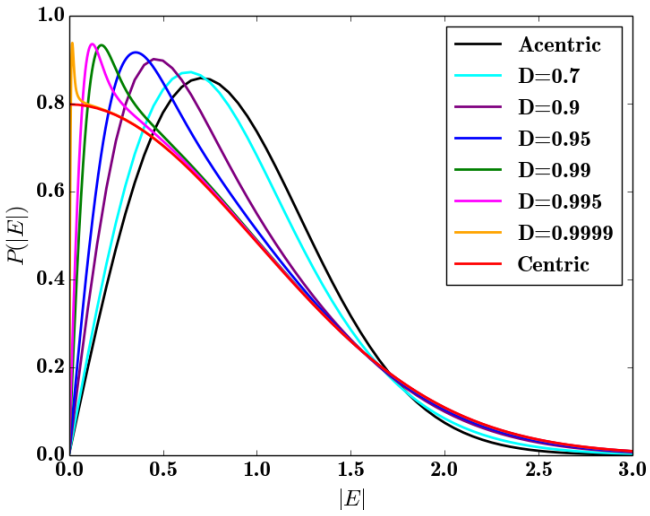
(19) Lee, H.; Leplae, P. R.; Porter, E. A.; Gellman, S. H. *J. Org. Chem.* **2001**, *66* (9),

3597–3599.

(20) Kiss, L.; Fülöp, F. *Chem. Rev.* **2014**, *114* (2), 1116–1169.

(21) Imamura, Y.; Umezawa, N.; Osawa, S.; Shimada, N.; Higo, T.; Yokoshima, S.; Fukuyama, T.; Iwatsubo, T.; Kato, N.; Tomita, T.; Higuchi, T. *J. Med. Chem.* **2013**, *56* (4), 1443–1454.

Chapter 4 The Impact of Pseudo-Inversion Symmetry on Structure Factor Amplitude Statistics



4.1 Introduction

Since the middle of the twentieth century, probabilistic methods have played an essential role in the development and success of X-ray crystallography, particularly single crystal diffraction.[1–3] These methods can be arbitrarily divided into two classes; those concerned purely with the statistical properties of diffracted X-ray intensities, and those bent towards solving the phase problem via probabilistic relations between certain classes of diffracted X-ray reflections. Investigations defined by the former class, aligned with the primary focus of this work, were first described by Wilson,[2] and investigations concerning the latter class were soon after brought to fruition by Sayre, Cochran, Hauptman, and Karle.[4] In the latter half of the twentieth century, statistical concepts such as maximum likelihood parameter estimation (MLE) and Bayesian inference have emerged as integral tools for macromolecular X-ray crystal structure solution and refinement.[3,5] This later development was primarily born out of necessity due to the less favorable data-to-parameter ratio that plagues macromolecular structure determination (relative to the data-to-parameter ratio in small molecule crystallography). Ultimately, the successful application of these probabilistic methods requires that a key statistical assumption is satisfied: the positions of random atoms in a model unit cell are uncorrelated so that the real and imaginary components of a complex structure factor are independently distributed.

The presence of pseudo-inversion symmetry, which is inexorable under the purview of quasiracemic crystallization, invalidates this central statistical assumption. Pseudo-inversion symmetry dictates that the positions of the atoms are correlated to some degree. The work described in this chapter assesses the impact that pseudo-

inversion symmetry has on the normalized structure factor amplitude probability density functions. Essentially, the degree to which the atomic position independence assumption is invalidated can be quantified and potentially exploited. To place these findings into context, the fundamental principles that become altered in the presence of pseudo-inversion symmetry are reviewed.

Although, there is an arbitrary distinction between philosophies concerning the application of probabilistic methods in X-ray crystallography, there is a commonality in objective, which is to extract as much information as possible about the molecular details of the crystal lattice in the absence of phase information. Diffracted X-rays are represented with complex structure factors of the form:

$$\boldsymbol{F}(hkl) = \sum_{i=1}^{N} f_i T_i e^{2\pi i(hx_i+ky_i+lz_i)} \qquad (4.1)$$

which is a summation over the *N* atoms in the unit cell where f_i is the atomic scattering factor, T_i is the temperature factor, and the complex exponential term represents the inner product of each set of atomic real space coordinates (*x,y,z*) with the corresponding reciprocal lattice vector (*h,k,l*) so that every atom in the unit cell contributes to a particular structure factor. In practice, Equation 4.1 is the Fourier-transform of the discrete electron density function, $\rho(x,y,z)$:

$$\boldsymbol{F}(hkl) = \sum_{x=0}^{1}\sum_{y=0}^{1}\sum_{z=0}^{1} \rho(x,y,z) e^{2\pi i(hx_i+ky_i+lz_i)} \qquad (4.2)$$

Therefore it follows that the real-space electron density function is the inverse Fourier transform of Equation 4.2:

$$\rho(x,y,z) = \sum_h \sum_k \sum_l \boldsymbol{F}(hkl) e^{-2\pi i(hx+ky+lz)} \qquad (4.3)$$

$$\rho(x,y,z) = \sum_h \sum_k \sum_l |\boldsymbol{F}(hkl)| e^{-2\pi i(hx+ky+lz+\varphi_{hkl})} \qquad (4.4)$$

where the complex structure factor with phase, φ_{hkl}, has been separated into real and imaginary components (Equation 4.4). The phase problem stems from the fact that only the modulus of the complex structure factor can be determined via elastic scattering of X-ray photons, and the phases are required to reconstruct the electron density function. Knowledge of the underlying symmetry within the crystal lattice is essential for successful phase reconstruction and structure refinement, and much can be obtained solely from the measured intensities, which are the squared norms of the complex structure factors (Equation 4.1).

In organic molecular crystals comprised of small molecules, the presence of a center of symmetry can be confirmed or refuted solely from the measured diffraction intensities.[1] In macromolecular crystals comprised of homochiral species, improper symmetry operators such as inversion centers and mirror planes are not possible; however, when racemic protein crystallization is employed, improper symmetry operations are now relevant and must be taken into account. Wilson was the first to recognize that the solution to the two-dimensional random walk problem could be applicable to understanding probability density functions of structure factor amplitudes. The key assumption is that a large number of randomly distributed atoms within a unit cell will result in randomly oriented atomic scattering factors (Figure 4.1). The contribution from each random atom is a complex random variable that is comprised of real and imaginary components. By the Central Limit Theorem both real and imaginary components (*A* and *B* respectively) of the resultant structure factor, ***F***, comprise sums of random variables and thus approximate normal distributions with variance, Σ, as *N* becomes large:

$$\boldsymbol{F} = \sum_{i=1}^{N} f_i e^{2\pi i \boldsymbol{h}\cdot\boldsymbol{x}_i} = A + iB \qquad (4.5)$$

$$P(A) = \frac{1}{\sqrt{\pi\Sigma}} e^{-\frac{\boldsymbol{A}^2}{\Sigma}} \qquad (4.6)$$

$$P(B) = \frac{1}{\sqrt{\pi\Sigma}} e^{-\frac{\boldsymbol{B}^2}{\Sigma}} \qquad (4.7)$$

$$P(\boldsymbol{F}) = \frac{1}{\pi\Sigma} e^{-\frac{\boldsymbol{A}^2+\boldsymbol{B}^2}{\Sigma}} = \frac{1}{\pi\Sigma} e^{-\frac{\boldsymbol{F}^2}{\Sigma}} \qquad (4.8)$$

$$P(F) = \frac{2F}{\Sigma} e^{-\frac{F^2}{\Sigma}} \qquad (4.9)$$

Equation 4.8 represents the probability density function for the complex structure factor, ***F***, for the acentric case, *i.e.*, no center of symmetry. Conversion to polar coordinates amounts to integration over an annulus with infinitesimal area, *2πFdF*, to yield the density function for the structure factor amplitude, *F* (Equation 4.9). For the centric case, the center of symmetry imposes the constraint, (*x,y,z*) → (*-x,-y,-z*), which results in the cancellation of the imaginary structure factor components for pairs of atoms related by inversion symmetry. This constraint also enhances the variance by twofold:

$$P(\boldsymbol{F}) = \frac{1}{\sqrt{\pi 2\Sigma}} e^{-\frac{\boldsymbol{F}^2}{2\Sigma}} \qquad (4.10)$$

$$P(F) = \sqrt{\frac{2}{\pi}} e^{-\frac{F^2}{2\Sigma}} \qquad (4.11)$$

Complete derivations of both acentric and centric distributions have been described by Wilson.[2]

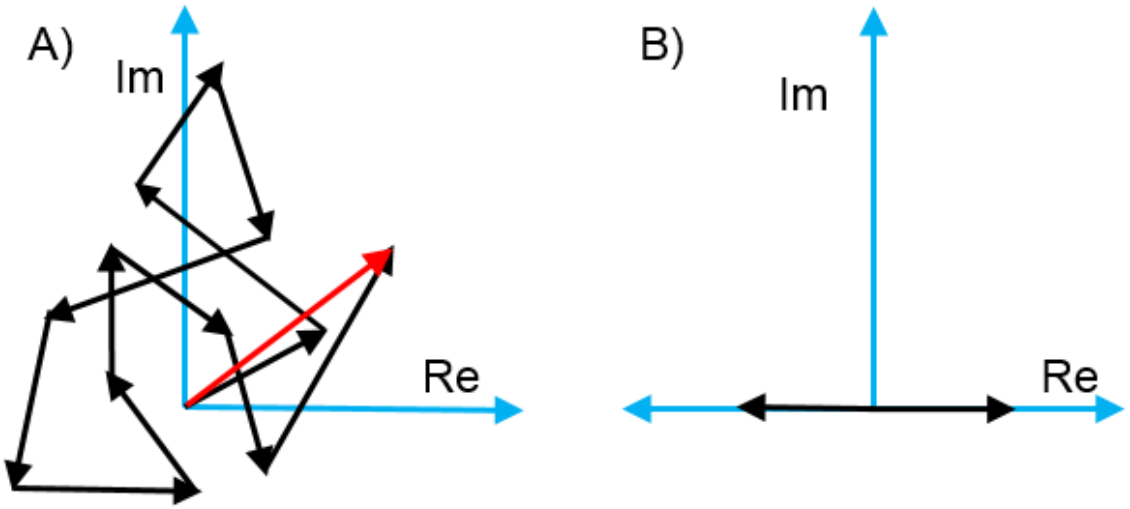


Figure 4.1. A) The two dimensional random walk applies to centric structure factors (atomic contributions, black; resultant structure factor, red), B) while inversion symmetry simplifies centric structure factors to a binary, one dimensional random walk.

The atomic scattering factor, f_i, and temperature factor, T_i, both depend on the angle of diffraction; therefore, the contribution from each term to the structure factor amplitude in Equation 4.1 is attenuated at higher scattering angles (Figure 4.2). Normalized structure factors that account for this resolution dependence are obtained with the following expression:

$$\boldsymbol{E} = \frac{\boldsymbol{F}}{\sqrt{\sum_{i=1}^{N} f_i^2}} \qquad (4.12)$$

Given this definition, the mean value of the squared normalized structure factor amplitude, $\langle E^2 \rangle$, is equal to one. Reflections are binned together within resolution shells so that the structure factor amplitudes are invariant with respect to resolution. Finally, Equation 4.10 can be combined with the acentric (Equation 4.8) and centric (Equation 4.9) probability distributions to yield the following expressions (Figure 4.3):

$$P(E) = 2Ee^{-E^2} \qquad (4.12)$$

$$P(E) = \sqrt{\frac{2}{\pi}} e^{-\frac{E^2}{2\Sigma}} \qquad (4.13)$$

Based on these distributions the presence of an inversion center can be detected by analyzing the value of parameters such as $<|E^2-1|>$, the shape of the distribution around the mean value $<E^2> = 1$, or higher order moments of $P(E)$, $<E^n>$.[1]

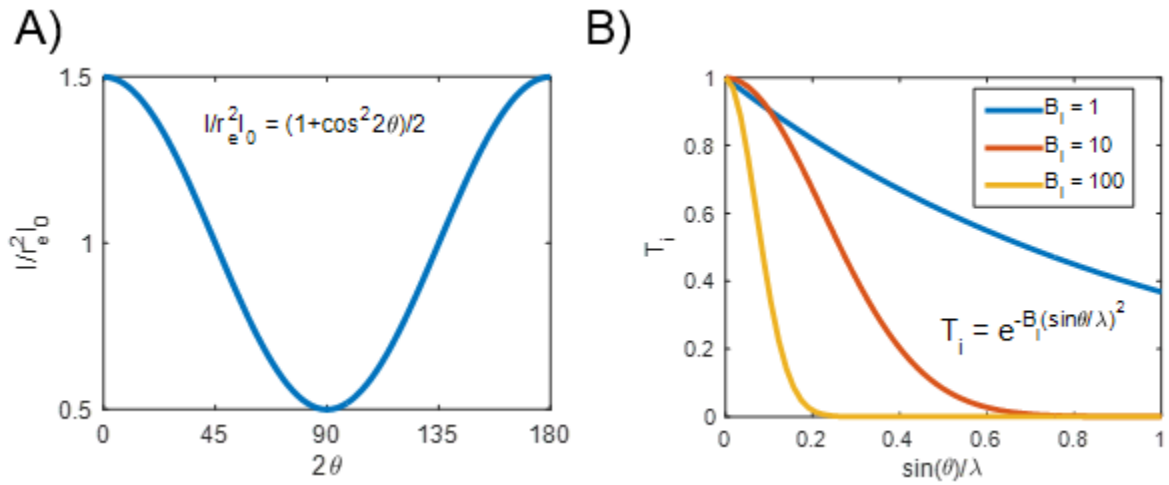


Figure 4.2. A) The angular dependence of the intensity of X-rays diffracted by a single electron is symmetrical, but the function decreases within the relevant resolution range [0°,90°]. B) The resolution-dependent attenuation in structure factor amplitude, attributed to temperature-dependent vibration of atoms within a crystalline lattice, increases in magnitude with increasing isotropic displacement parameter, B_i, values.

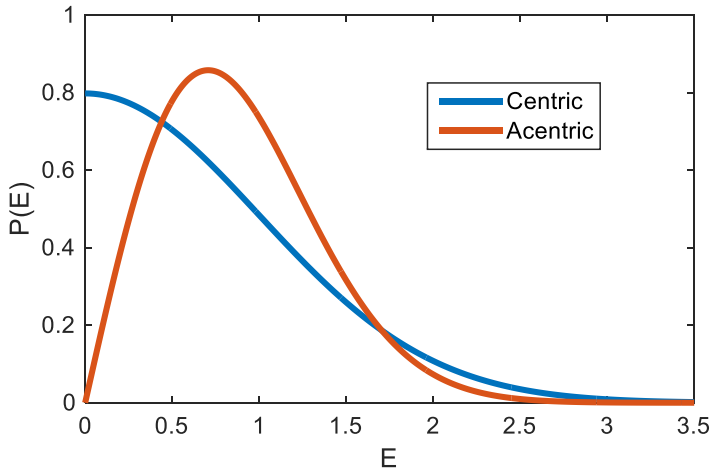


Figure 4.3. The probability density function of normalized structure factor amplitudes is illustrated for centric (blue) and acentric (red) structure factors.

Knowledge of the presence or absence of an inversion center is extremely useful, especially in small molecule structure determination; however, a more powerful result lies in the extension of Equations 4.12 and 4.13 to reformulate the distribution of a complex structure factor, ***E*** (normalized) or ***F*** (un-normalized), in terms of a conditional probability distribution, *i.e.*, to define some probabilistic relationship between a hypothetical model and the data. These conditional distributions are especially helpful in each step of structure solution and refinement, where the goal is to determine model parameters that result in a chemically sensible model that agrees closely with the data. Historically, least squares methods have worked well when there is a favorable number of measured reflections compared to the number of model parameters, such as in a small molecule structure.[6] However, when the number of reflections decreases relative to the number of model parameters, parameter estimation becomes more difficult or even intractable with least squares methods.[3,5,7] For this reason, maximum likelihood parameter estimation (MLE) has become the predominate choice in macromolecular crystallography.

Implementation of this error minimization goal at each stage requires some sort of error function that can be minimized with respect to the model parameters for each atom, namely x,y,z and B_i (isotropic). In practice there are three main sources of error that result in a discrepancy between the measured structure factor amplitudes, F^o, and the amplitudes derived from the hypothetical atomic coordinate model, F^c. These sources include: error in the coordinates of atoms within the model, missing atoms from the model, and error in the measurement of reflection intensities.[3,7] The aforementioned model/measurement amplitude discrepancy is typically quantified with a discrepancy index, or R-factor:

$$R = \frac{\sum_h \left| |F^o| - |F^c| \right|}{\sum_h |F^o|} \quad (4.14)$$

Srinivasan and Ramachandran demonstrated that the impacts of both coordinate errors and missing atoms on the normalized structure factor variance are mathematically equivalent, a term they defined as σ_A.[1] Therefore these effects can be combined with a normalization and experimental measurement error term to define an empirical parameter, β, which modulates the variance of $P(\boldsymbol{F})$ within a particular resolution shell.[7] In addition, the model structure factor, $\boldsymbol{F}^c$, is proportional to the mean value of the observed structure factor, $\langle \boldsymbol{F}^o \rangle$. Conceptually, the probability distribution of an observed structure factor amplitude, $\boldsymbol{F}^o$, given a model-derived structure factor, $\boldsymbol{F}^c$, is equivalent to the distribution of the difference between these vectors in the complex plane, $\Delta\boldsymbol{F}$ (Figure 4.4). An expression for the probability that a particular amplitude is measured, *i. e.* the data, given a particular model can be generated by taking the product of the distributions for real and imaginary components of $\Delta\boldsymbol{F}$, which are assumed to be independent; converting to polar coordinates (Equation 4.15); then integrating out the phase term (Equation 4.16), which is unknown. Finally, Equation 4.17 is the conditional distribution, P(data given model) that can be combined with Bayesian inference to estimate model parameters via MLE, and I_0 is a modified Bessel function of the first kind.

$$P(F^o, \varphi | F^c) = \frac{F}{\pi\beta} \exp\left[-\frac{(F^o)^2 + (\alpha F^c)^2}{\beta} \right] \exp\left[2\frac{\alpha F^o F^c}{\beta} \cos(\varphi - \varphi^c) \right]$$

(4.15)

$$P(F^o | F^c) = \frac{F}{\pi\beta} \exp\left[-\frac{(F^o)^2 + (\alpha F^c)^2}{\beta} \right] \int_0^{2\pi} \exp\left[2\frac{\alpha F^o F^c}{\beta} \cos(\varphi - \varphi^c) \right] d\varphi$$

(4.16)

$$P(F^o|F^c) = \frac{2F}{\pi\beta}\exp\left[-\frac{(F^o)^2 + (\alpha F^c)^2}{\beta}\right] \mathrm{I}_0\left[2\frac{\alpha F^o F^c}{\beta}\right]$$

(4.17)

Different investigators have employed different forms of Equation 4.17 for MLE of model parameters for macromolecular structures. Notably, Murshudov and Reed have employed the σ_A/*E* formalism for the programs *REFMAC* (macromolecular refinement) and *Phaser* (molecular replacement) respectively,[5,8] whereas others have employed the α,β/*F* formalism in the CCTBX/Phenix program suite.[7,9,10] Procedures for estimating σ_A or α,β are dependent upon the implicit assumption that atomic positions are uncorrelated, so invalidation of this assumption could possibly affect the proper estimation of these variance and scaling parameters.

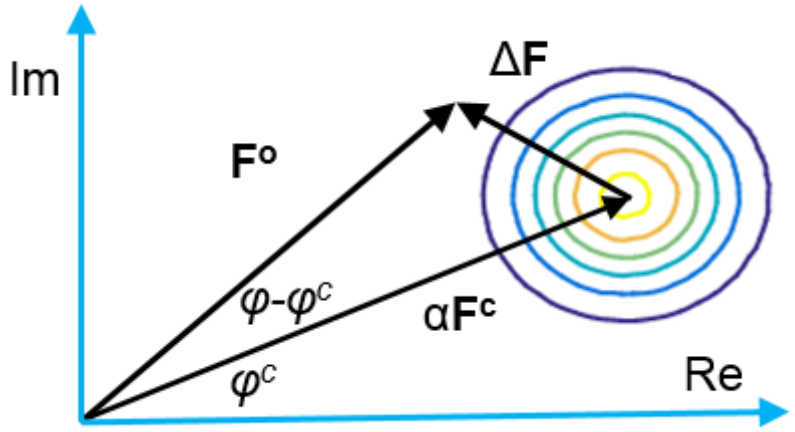


Figure 4.4. A schematic representation of the relationship between $\mathbf{F}^o$ and $\mathbf{F}^c$ illustrates that the conditional distribution $P(\mathbf{F}^o,\varphi|\mathbf{F}^c)$ is represented by the distribution of Δ**F**, which is subject to the cosine rule. The contoured lines represent a two-dimensional probability density function centered at $\alpha\mathbf{F}^c$.

Bayesian inference yields the probability of observing some event given some data as the product of a *prior probability* and a *likelihood* function derived from some statistical model (Equation 4.18). This probability is known as the *posterior probability,* the

probability that some hypothesis is true given the observed data. With regard to X-ray diffraction data, Equation 4.17 is the likelihood function that represents the probabilistic model connecting model parameters, *i. e.* atomic coordinates and displacement parameters, to the structure factor amplitude of acentric reflections (Equation 4.19). The *prior probability* term, or *P*(model), incorporates chemical constraints such as bond angles and lengths. The *P*(data) term represents the marginal likelihood and is a constant value for all events considered, so the term is removed when comparing probabilities (see McCoy for a review of all these concepts[11]).

$$P(model|data) = \frac{P(data|model)P(model)}{P(data)} \qquad (4.18)$$

$$P(F^c|F^o) = P(F^o|F^c)P(F^c) \qquad (4.19)$$

$$P(model|data) = P(Xray)P(Chemistry) \qquad (4.20)$$

Ultimately Bayesian inference allows for the construction of a likelihood function with the assumption that the likelihood distributions (Equation 4.17) for measured reflections are independent of each other. Therefore a global likelihood function can be established as a product of the likelihood function for each reflection (Equation 4.21); the negative logarithm of the likelihood function is computed to make the likelihood function more tractable computationally (Equation 4.22). The search for the optimal model parameters during structure solution or refinement becomes a minimization problem based on a probabilistic relationship between structure factor amplitudes and model parameters. Conventional construction of the log-likelihood function for acentric reflections (Equation 4.22) is predicated on the acentric probability distribution, an assumption that may be invalidated in the presence of non-crystallographic symmetry.[12,13]

$$L = \prod_h P(F_h^c | F_h^o) \qquad (4.21)$$

$$\ln(L) = -\sum_h \ln[P(F_h^c | F_h^o)] \qquad (4.22)$$

Quasiracemic crystallization, by definition, involves the co-crystallization of a molecule with its quasienantiomer. These pairs of molecules can range from small molecule isosteres to peptides or proteins that differ via a side chain or backbone modification. Statistical evidence from both small molecule and protein structural data banks suggests that racemic mixtures of molecules more readily crystallize in space groups that contain inversion centers versus chiral space groups. The triclinic space group, *P* $\bar{1}$, and the monoclinic space group, $P2_1/c$, are predicted to be the most commonly observed centrosymmetric space groups.[14,15] In either case pairs of enantiomers typically are configured around an inversion center, giving rise to reflections that obey the centric distribution (Equation 4.13). Quasi-enantiomers achieve similar centrosymmetric packing configurations; however, true inversion symmetry is precluded by the point of difference between the molecules. These differences disrupt the crystallographic inversion symmetry and therefore alter the structure factor amplitude distribution to some degree. Srinivasan and coworkers described a probability density function for an acentric structure with some 'degree of centrosymmetry' (Figure 4.5).[16–18] The probability density function has the following form:

$$P(E) = \frac{2E}{\sqrt{1-D^4}} \exp\left[-\frac{E^2}{1-D^4}\right] I_0\left[\frac{D^2E^2}{1-D^2}\right] \qquad (4.23)$$

$$D = \langle \cos 2\pi \Delta \boldsymbol{r} \vec{\boldsymbol{s}} \rangle \qquad (4.24)$$

where $\Delta\boldsymbol{r}$ is the deviation in atomic coordinates from centrosymmetric positions, and $\vec{\boldsymbol{s}}$ is the scattering vector with magnitude, sin(θ)/λ. The parameter *D* quantifies the 'degree of

centrosymmetry'.[16] For $D \rightarrow 0$, $P(E)$ is the acentric distribution (Equation 4.12), and for $D \rightarrow 1$, $P(E)$ approaches the centric distribution (Equation 4.13). Interest in these theoretical pseudo-centric distributions appears to have peaked in a period that spans the early 1970s to the mid-1980s,[16–19] but despite this interest, experimental validation or application of these distributions appears to be lacking in the literature.

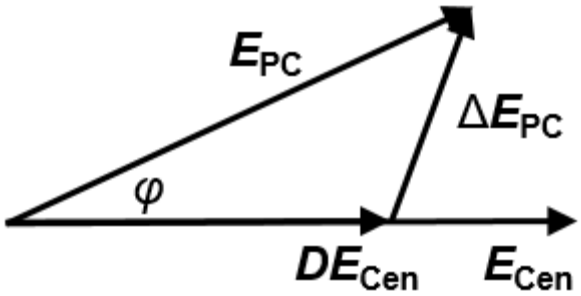


Figure 4.5. The centric structure factor represented by DE_{Cen} has a 0° phase angle and is attenuated in magnitude by the parameter D. E_{PC} represents the pseudo-centric structure factor where the inversion symmetry constraint has been removed. The conditional probability distribution can be determined for E_{PC} by applying the cosine rule to ΔE_{pc}, where $\Delta E_{pc}^2 = E_{PC}^2 + 2D^2E_{Cen}^2 - 2DE_{PC}E_{Cen}\cos(\varphi)$.

Fundamentally, the objective of my crystallization efforts is to obtain some structure-based hypothesis pertaining to the native function or stability of proteins that are difficult to crystallize. Quasiracemic crystallization introduces a unique perspective, unavailable from purely chiral structures, which is the degree of deviation of a modified structure, *e.g.*, a polypeptide containing a backbone modification, from some reference structure. Ancillary efforts are aimed at exploiting this deviation to address the phase problem that arises in macromolecular crystallography, which requires some framework for handling pseudo-inversion symmetry. In the present work, the 'degree of centrosymmetry' or 'degree of pseudosymmetry' has been quantified for the series of β substituted villin headpiece (VHP) quasiracemic structures described in Chapter 2.

Similarly, I determined the degree of pseudosymmetry for a series of quasiracemic L-ubiquitin dimers/D-ubiquitin monomer structures reported by Liu *et al.*[20] In that case, the propensity of quasiracemic mixtures to afford centrosymmetric crystals was exploited to obtain atomic resolution structures of synthetic ubiquitin dimers, covalently linked at various branching points. This impressive array of structures from Liu *et al.* underscores the utility of quasiracemic crystallization.[20,21]

To further structure solution and refinement via quasiracemic crystallization, I describe a method for deriving a conditional phase angle probability distribution strictly from measured pseudo-centric structure factor amplitudes. This distribution was found to be in good agreement with phase angles derived from the β-substituted VHP quasiracemic structures described in Chapter 2. Furthermore, the impact that pseudo-inversion symmetry has on the largest likely R-factor (Equation 4.14), an important term in structure refinement, is also described for different degrees of pseudo-inversion symmetry. Finally, efforts to assess the impact pseudo-inversion symmetry has on maximum likelihood-based structure refinement conclude with a comparison of the likelihood function (Equation 4.17) for simulated pseudo-centric data versus uncorrelated acentric data.

This chapter concludes with a case study of pseudo-inversion symmetry manifest in a homochiral structure, where the pseudo-centric symmetry arises from a pseudo-two-fold rotational axis.

4.2 Results and Discussion

4.2.1 Quantifying Pseudosymmetry Among β Substituted VHP Quasiracemates

Numerical determination of the pseudosymmetry parameter, *D*, from Equation 4.23 is possible for each quasiracemate dataset with a brute-force grid search. *D* was determined by finding the value that minimized the sum of the squares of residuals (Equation 4.25) between the marginal distribution (Equation 4.23) and the data, $P(E_{data})$. $P(E_{data})$ was determined by first binning and normalizing the structure factors from each dataset with the *CCP4* program *ecalc* and then scaling the fraction of reflections with a given *E* by a normalization factor, so that the sum of the scaled values was equal to one, corresponding to a normalized probability distribution function (Equation 4.26). The width of each bin was 0.05, and the set of reflections for each dataset was truncated so that the R-factor was no greater than 10%. Otherwise, inclusion of noisier reflections at higher resolutions resulted in distributions that were skewed disproportionately towards lower *E* values. Shmueli and Wilson have described a moment-generating function for pseudo-centric distributions such as Equation 4.23,[17] which would likely be more tractable than the least squares procedure employed here for implementation in a refinement or scaling program.

$$\chi = \sum_{i=1}^{k}[P(E_D) - P(E_{data})]^2 \quad (4.25)$$

$$P\left(E_{data,i}\right) = \frac{E_{data,i}}{(0.05)N_{bin}} \quad (4.26)$$

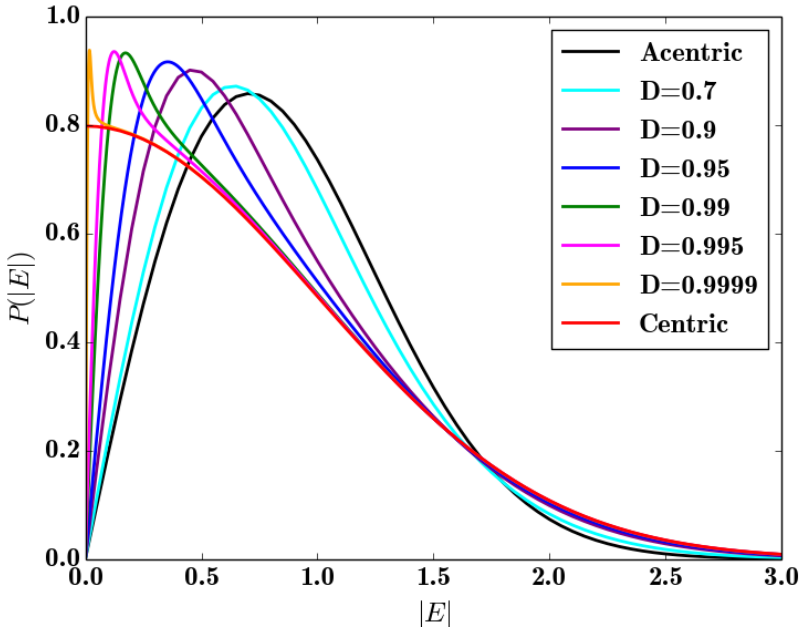


Figure 4.6. Theoretical normalized structure factor amplitude distributions for the marginal distribution described in (Equation 4.23) plotted for varying values of D. Centric and acentric Wilson distributions are also shown as limiting cases.

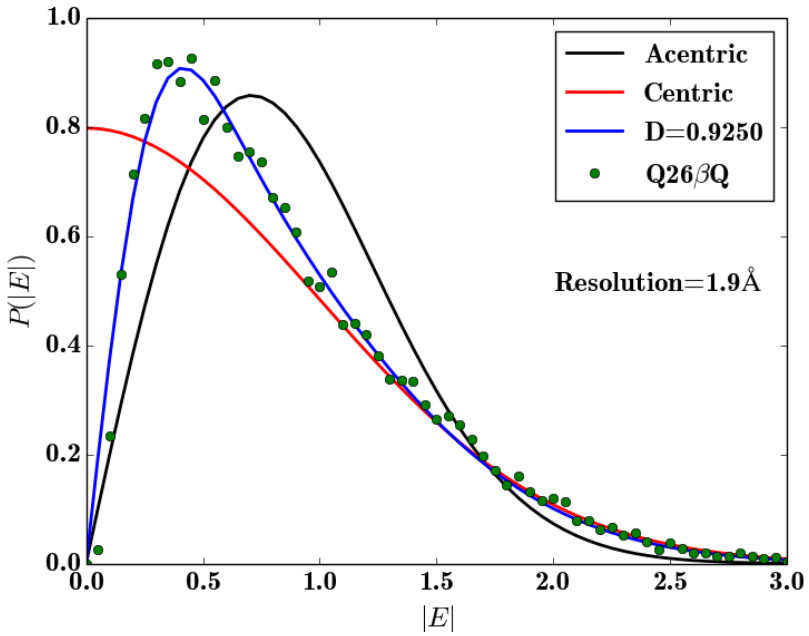


Figure 4.7. P(E) curve for VHP Q26β[3]hQ.

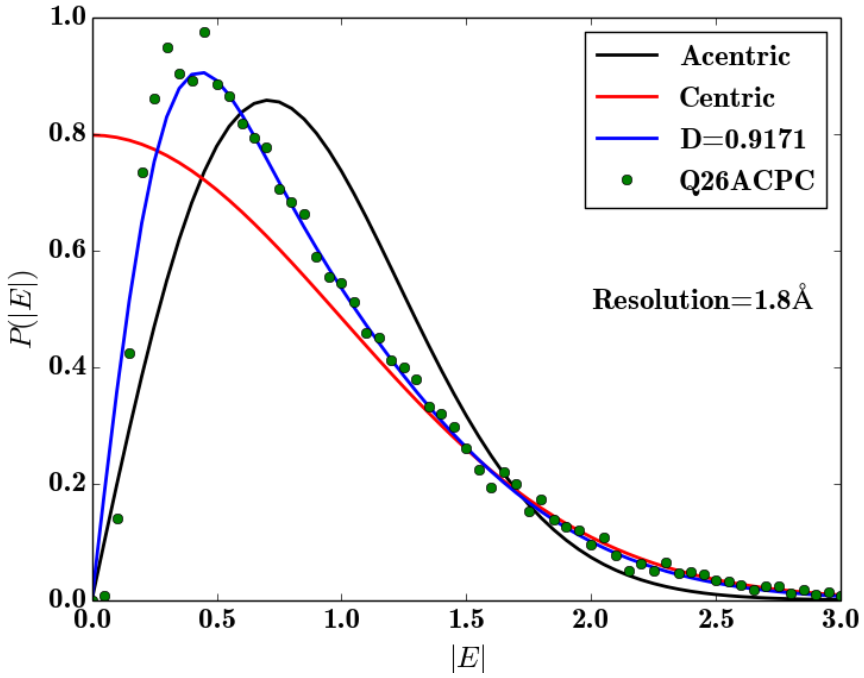


Figure 4.8. P(E) curve for VHP Q26ACPC.

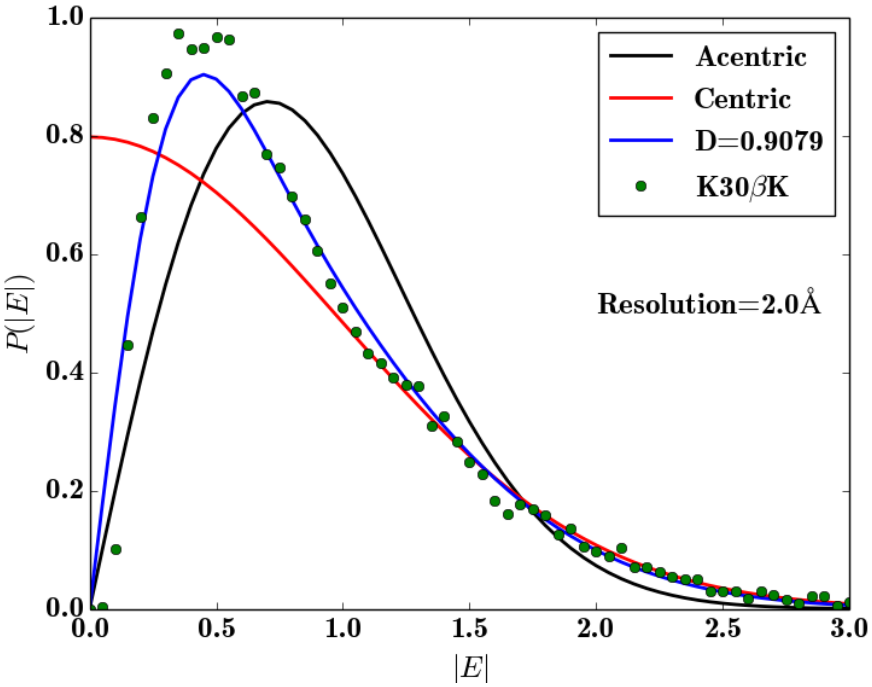


Figure 4.9. P(E) curve for VHP K30β^{3}hK.

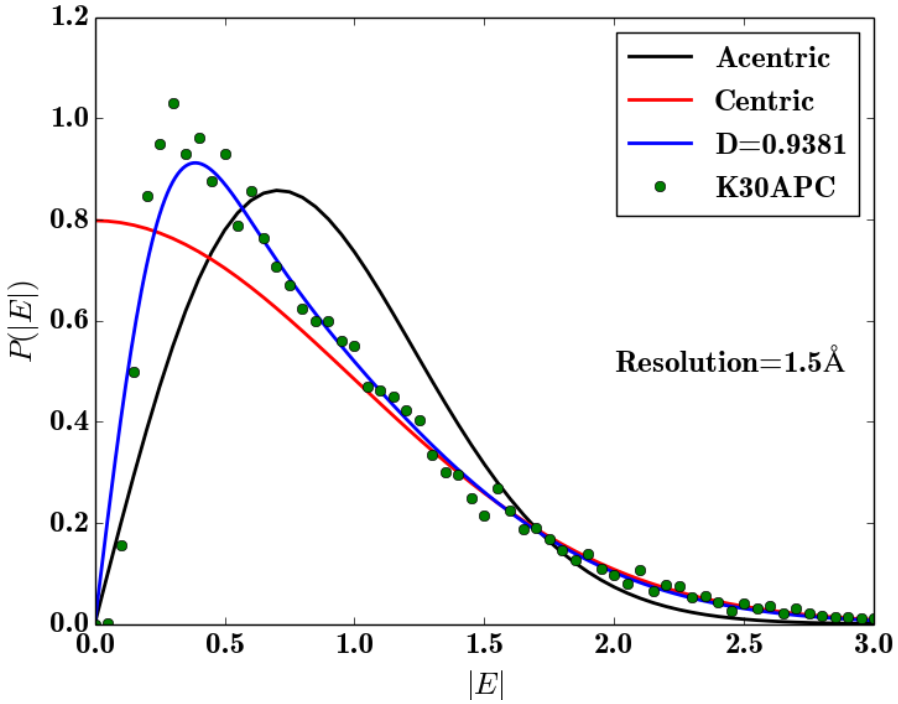


Figure 4.10. P(E) curve for VHP K30APC.

The VHP K30APC quasiracemate has the highest degree of pseudosymmetry with the *D* value closest to 1 (0.9381), where *D*=1 corresponds to a truly centrosymmetric structure. This result is consistent with the other metrics of pseudosymmetry, namely the RMSDs between pseudo-inversion-related pairs of peptide chains and the phase angle distribution, where the K30APC quasiracemate also exhibited the highest level of pseudosymmetry (Chapter 2). However, a marked difference between the model-based measures of pseudosymmetry and the structure factor amplitude distribution analysis is that the latter is made *a priori*, *i.e.*, prior to structure solution and refinement.

A shortcoming of the amplitude distribution analysis is the lack of a straightforward method to determine a standard error for the parameter *D* within a single dataset. Given the favorable data-to-parameter ratio using the least squares method, non-linear regression, rather than a brute force search, could possibly yield a numerical second

derivative and thus allow for some estimation of uncertainty for a given *D*-value. An alternative approach may include maximum likelihood estimation of the parameter *D*.

The fitted pseudo-centric distributions overlap well with the normalized amplitude probability values. To further assess the validity of the distribution fits for the quasiracemic datasets, RMSDs between backbone atoms (C_α, C, N, O) of pseudo-inversion related pairs of chains were plotted as a function of the corresponding fitted *D*-values. As $D \rightarrow 1$, the corresponding distribution becomes centric, *i. e.* RMSD $\rightarrow 0$ (Figure 4.6). Linear regression of the four points from each β substituted VHP quasiracemate atomic coordinate model and structure factor amplitude distributions show a reasonable R^2 value (0.83) when extrapolated to include the limiting centric case. The line has a slope of approximately -0.17 Å per 0.1 *D*.

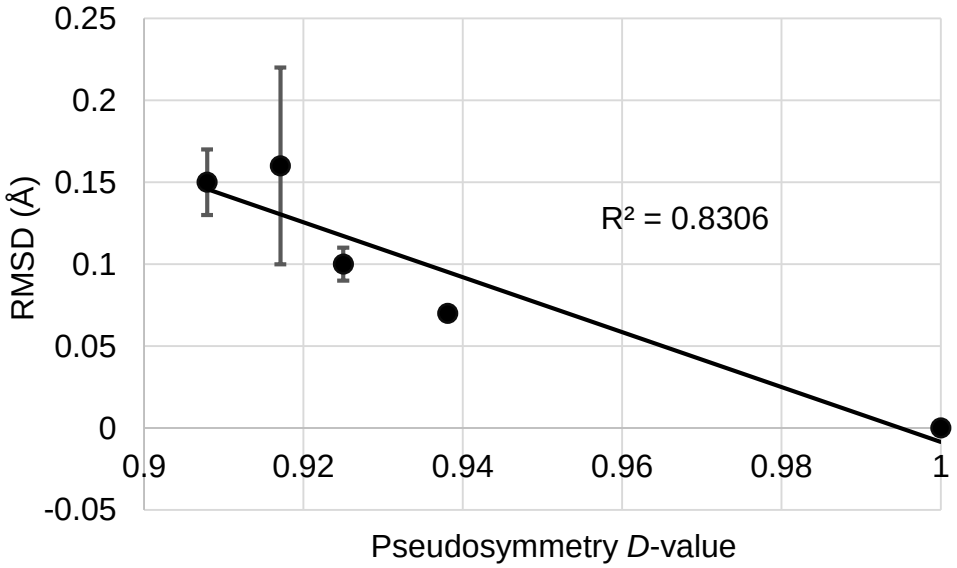


Figure 4.11. The RMSD between backbone atoms (C_α, N, C, O) related by a pseudo-inversion center is plotted as a function of the pseudosymmetry D-value among the series of β-substituted quasiracemates. The point that corresponds to a purely centrosymmetric structure (D=1, RMSD=0 Å) was included in the linear regression fit. Error bars were determined when there were multiple pairs of pseudo-inversion related chains.

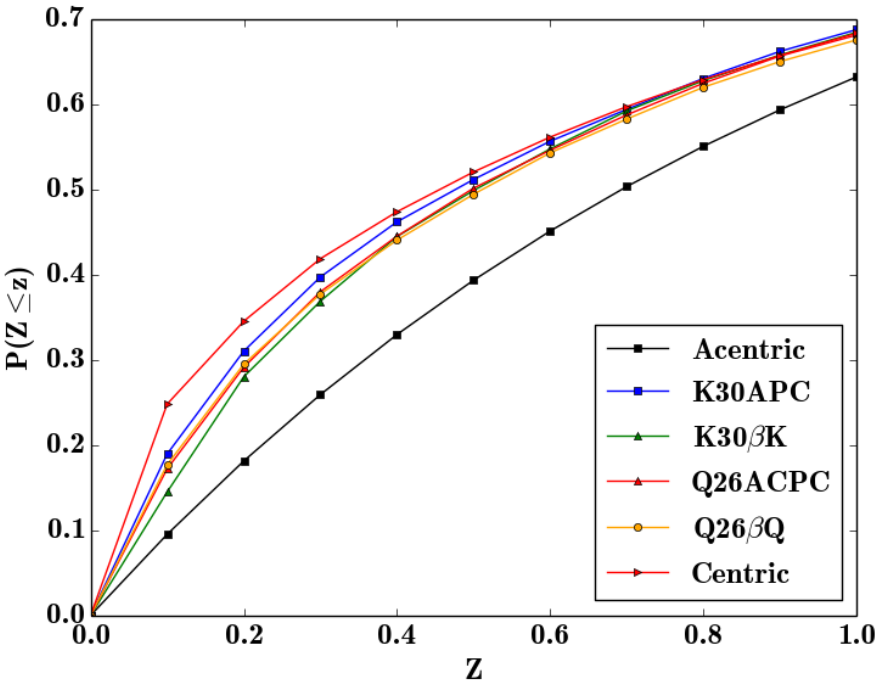


Figure 4.12. The cumulative normalized intensity distributions for the β substituted VHP quasiracemates also display a pseudo-centric distribution. At larger normalized intensity values the pseudo-centric distributions more closely approximate a centric distribution.

4.2.2 Quantifying Pseudo-inversion Symmetry in Proteins: Branched Ubiquitin Quasiracemates

An attractive feature of quasiracemic crystallization is that a single enantiomer of a native sequence can promote crystallization among an array of modified versions of that particular sequence. This versatility is especially useful given the economic constraints surrounding the synthesis of large proteins, comprised of D-amino acids, through multiple segment condensations. Recently, Liu *et al.* have pushed this aspect of quasiracemic crystallization to the limit by crystallizing solutions comprised of 2:1 molar ratios of D-ubiquitin monomers with L-ubiquitin dimers (Figure 4.13).[20] The motivation was to access atomic-resolution structures of ubiquitin connected at different branching points after conventional chiral crystallization had failed. This work produced several high

resolution structures of quasiracemic ubiquitin dimers, which in turn provided an opportunity to analyze the distribution of normalized structure factor amplitudes.

Several open questions still remain pertaining to quasiracemic crystallization. How 'different' can two quasi-enantiomers be and still crystallize with a pseudo-centrosymmetric configuration? What effects will the pseudo-inversion symmetry have on conventional structure solution/refinement programs, algorithms, and weighting schemes? Analysis of the structure factor amplitude distribution begins to address both of these questions, because the degree of deviation from centrosymmetry can be quantified, and the degree of deviation will in turn affect the refinement scheme. At first glance, the structure factor amplitude distributions for the ubiquitin quasiracemates appear similar to those of the β substituted VHP variants. However, the distributions are qualitatively more centric in character. Close inspection of the curves shows that deviation of the empirical $P(E)$ from the purely centric distribution begins much earlier ($E < \sim 0.9$), whereas for the β-substituted VHP quasiracemates the deviation begins significantly later ($E > 1.0$). Intuitively, this trend is reasonable since the point(s) of difference within the ubiquitin quasiracemate is confined to a flexible region and is relatively small compared to the ~150 residue dimer. This observation is significant because the distributions more closely resemble a truly centric distribution when scaling and measurement errors are taken into account. There is a possibility that weighting the nominally acentric reflections *P*1/pseudo-*P*$\bar{1}$ as centric reflections could improve the overall refinement and resultant models.

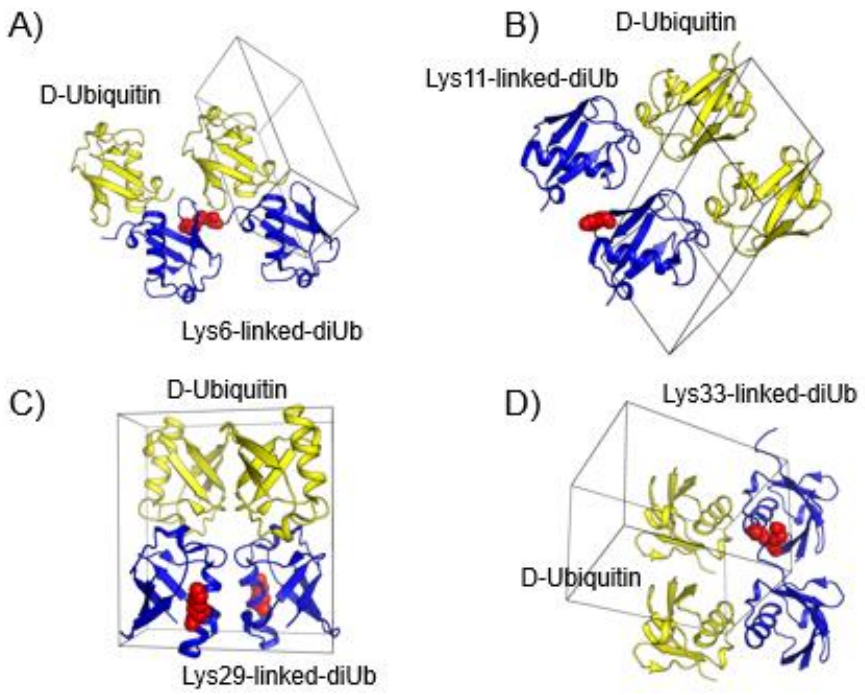


Figure 4.13. The reported D-mono-ubiquitin (yellow), L-di-ubiquitin (blue) quasiracemates feature L-di-ubiquitin connected via an amide linkage between a C-terminus and the side chain at A) Lys6, B) Lys11, C) Lys29, and D) Lys33. Each crystal has P1 symmetry (pseudo-P$\bar{1}$) except for the Lys29 linked quasiracemate which has P2 symmetry (pseudo-$P2_1/c$). The lysine side chains involved in the linkage are shown as red spheres. Individual structures are deposited in the PDB.[20]

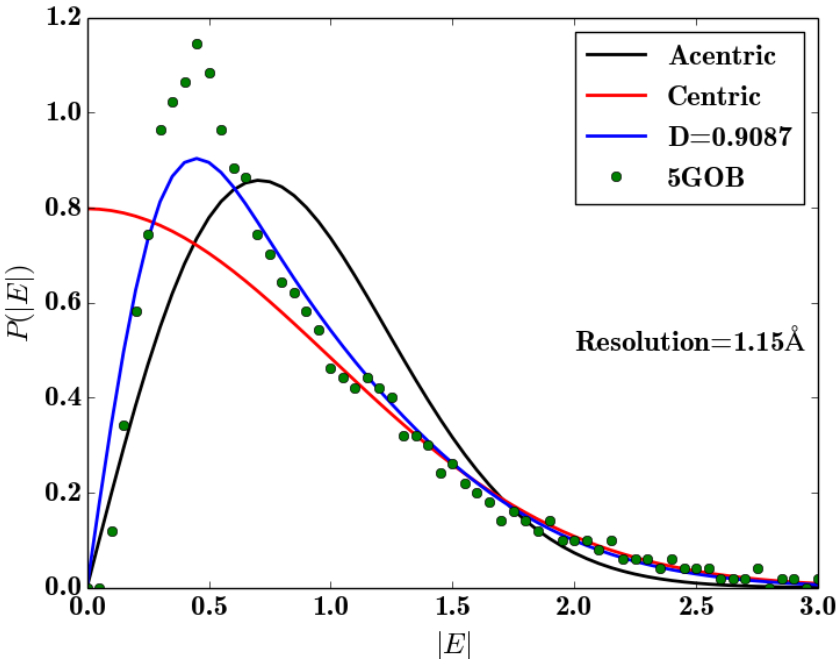


Figure 4.14. Quasiracemic Lys6-linked di-L-Ubiqutin with D-Ubiquitin monomer normalized structure factor amplitude distribution (PDB: 5GOB).

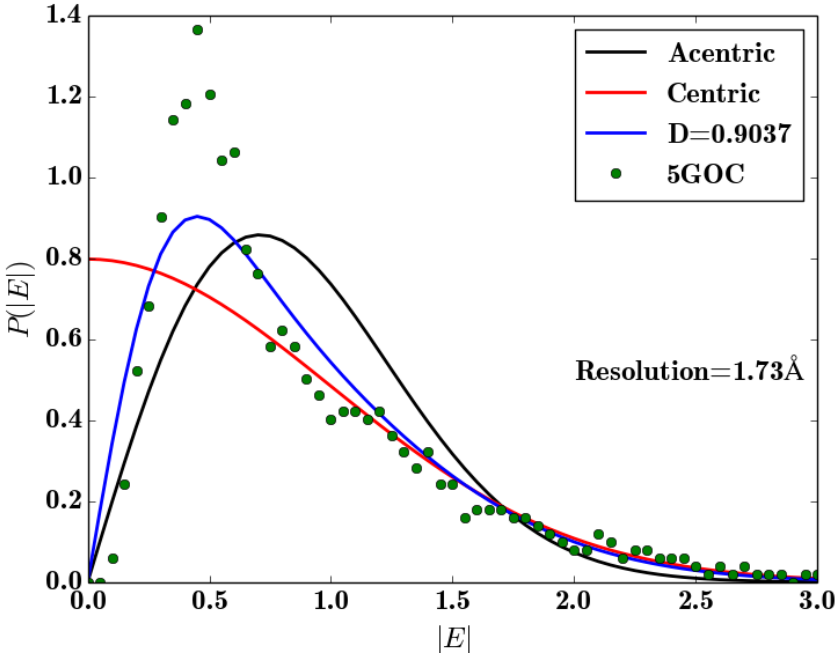


Figure 4.15. Quasiracemic Lys11-linked di-L-Ubiqutin with D-Ubiquitin monomer normalized structure factor amplitude distribution (PDB: 5GOC).

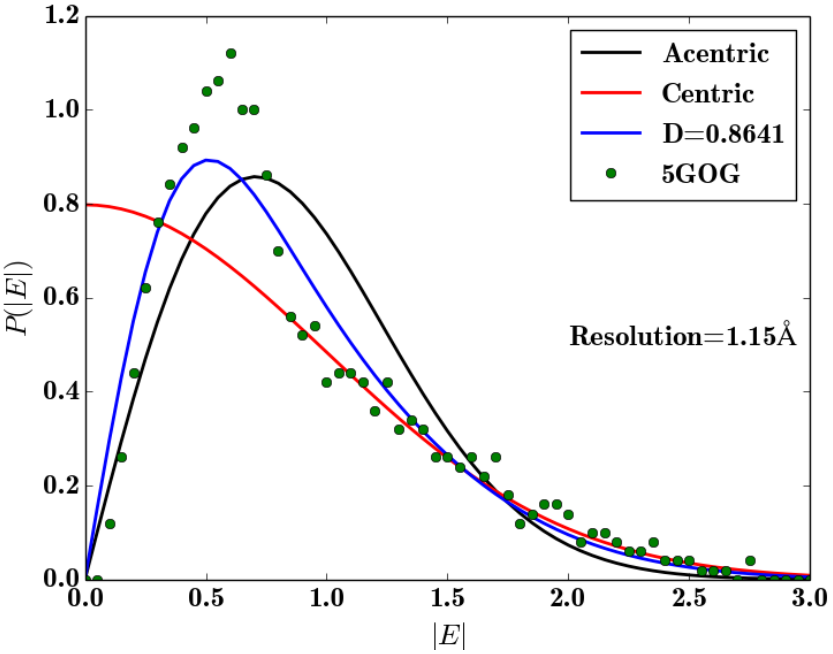


Figure 4.16. Quasiracemic Lys29-linked di-L-Ubiqutin with D-Ubiquitin monomer normalized structure factor amplitude distribution (5GOG).

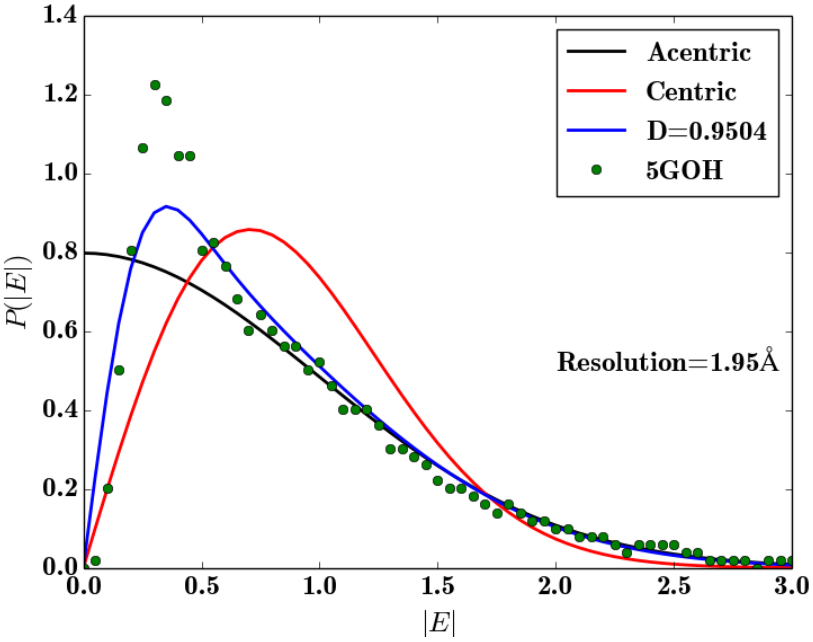


Figure 4.17. Quasiracemic Lys33-linked di-L-Ubiqutin with D-Ubiquitin monomer normalized structure factor amplitude distribution (5GOH).

4.2.3 An Experimental Phase Probability Density Function Derived from Pseudosymmetry

Quasiracemic protein crystallization has been demonstrated as a potentially useful tool for experimental phasing of protein X-ray crystal structures. In terms of experimental phasing strategies, quasiracemic crystallization provides a distinct advantage over racemic crystallization, where the centrosymmetric constraint leads to a loss of any anomalous signal. Pentelute *et al.* showed that the benefits of racemic protein crystallization could be combined with multi-wavelength anomalous dispersion (MAD) phasing.[12] Similarly, Sawaya *et al*. insightfully demonstrated that the pseudo-inversion symmetry could be exploited to break the phase angle ambiguity that arises during single-wavelength anomalous dispersion (SAD) experiments (Figure 4.18).[22]

During a SAD phasing experiment, atomic scattering factors, f_i, of atoms that undergo anomalous dispersion are altered by a wavelength-dependent scattering factor with real and imaginary components, $f_{\lambda i}$ and $f_{\lambda i}''$. This shift in scattering factors results in a concomitant shift in structure factors that can be measured in the observed intensities (and thus amplitudes). Moreover, this shift in amplitude for each reflection is consistent with two different phase angles, where one of the two possible phase angles corresponds to the true model phase. The measured amplitudes are then used to construct a bimodal phase probability density function.[23] Conventional methods for discriminating true model phases from alternative ‘noise’ solutions typically include some form of density modification, a procedure where prior knowledge about the electron density disparity between protein and solvent regions is exploited.[24]

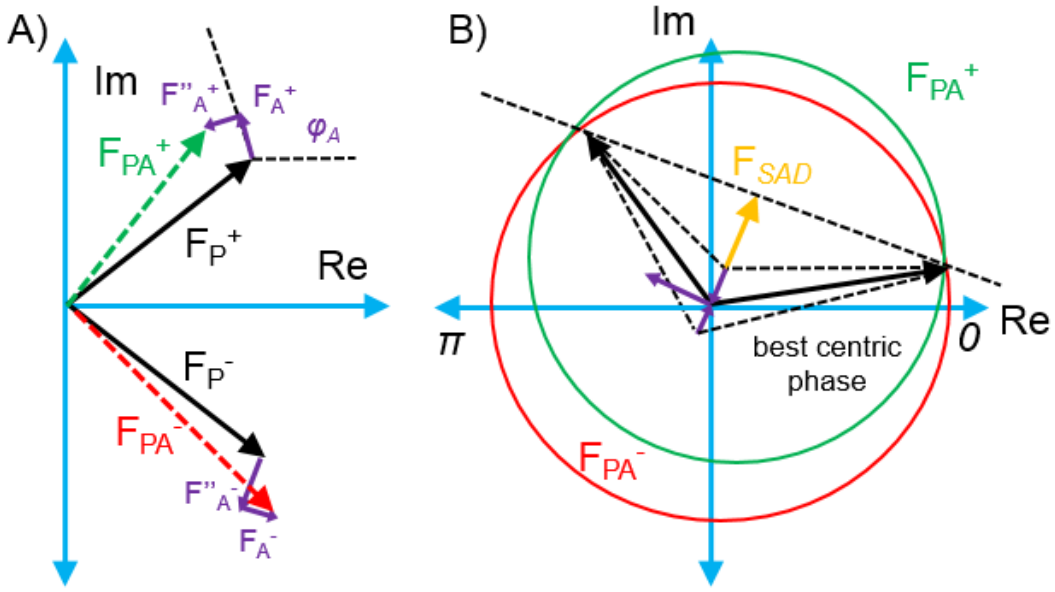


Figure 4.18. A) The structure factor shifts caused by anomalous dispersion. B) Sawaya et al. demonstrated that choosing the SAD best phase closest to either 0 or π radians resulted in higher quality maps than traditional SAD phases.

As mentioned previously, Sawaya *et al.* used a centric symmetry approximation, rather than density modification, to distinguish the true model phase from the incorrect noise phase.[22] This method may be advantageous in cases where density modification fails, such as in densely packed crystals where solvent content is lower than normal. In the centric approximation method, SAD phase determination is relegated to a binary choice between the ambiguous solutions where the phase closest to either 0° or 180° is accepted as the 'SAD best centric phase'. Sawaya *et al.* speculated that in principle, a 'more robust' procedure would require some statistical method to predict the phase deviation from centrosymmetry in advance.[22]

That statistical method is described here. The pseudosymmetry *D*-value, determined from experimental structure factor amplitude distributions, can be used to construct a phase probability distribution that agrees with the model (initially phased with

molecular replacement) phases for the series of VHP β substituted quasiracemates. Furthermore, the validity of this new phase probability distribution is underscored by the fact that the VHP quasiracemates were initially phased with molecular replacement and refined with a conventional maximum-likelihood target function, both of which make no implicit assumptions concerning the presence of pseudosymmetry.

Empirical phase probability distributions were determined by binning phase angles for each refined VHP β substituted quasiracemate structure, where the pseudo-inversion center in each model was shifted to the origin to qualitatively assess the shape of the distribution. Histograms of binned phase angles were normalized to generate a probability density function to accommodate the varying number of reflections among datasets. Each model exhibits a bimodal distribution of phase angles centered on 0 and π radians (0° or 180°); the VHP K30APC quasiracemate has the sharpest distribution, which is in line with this structure having the lowest RMSD value between pseudo-inversion related pairs of chains among each structure. These phase angle distributions combined with the low RMSD values (<1 Å) between pseudo-inversion related pairs of chains are indicative of pseudo-*P*$\bar{1}$ space group symmetry for each of the four structures that were examined.

The procedure for generating the phase probability distribution for pseudo-centric structure factor amplitudes is conceptually similar to the procedure first described by Sim for purely acentric or centric distributions.[9,25] Sim was interested in understanding and accounting for the impact that missing atoms have on the mean phase error. For the pseudo-centric case, the conditional phase probability distribution, $P(\varphi|E,E_c)$, where E is an acentric structure factor and E_c is a centric structure factor, is constructed by using the relationships for joint and conditional distributions (Equations 4.27-4.29). At any step

where the phase is marginalized, *i.e.*, removed as a variable by summing over all possible values, a modified Bessel function of the first kind is introduced due to the identity, $\int_0^{2\pi} e^{-a\cos\alpha} d\alpha = 2\pi I_0[a]$.[26]

$$P(\varphi, E|E_c) = \frac{E^o}{\pi(1-D^2)} \exp\left[-\frac{E^2+D^2E_c^2-2DEE_c\cos\varphi}{1-D^2}\right] \quad (4.27)$$

$$P(E|E_c) = \frac{E}{\pi(1-D^2)} \exp\left[-\frac{E^2+D^2E_c^2}{1-D^2}\right] \int_0^{2\pi} \exp\left[\frac{2DEE_c}{1-D^2}\right] d\varphi \quad (4.28)$$

$$P(E|E_c) = \frac{2E}{(1-D^2)} \exp\left[-\frac{E^2+D^2E_c^2}{1-D^2}\right] I_0\left[\frac{2DEE_c}{1-D^2}\right] \quad (4.29)$$

Equation 4.27 and 4.29 can be combined to obtain a conditional phase distribution for a given *E* and E_c value, which reflects the fact that a conditional probability is simply a normalized joint probability. Centric reflections are constrained to phases of 0 or π radians; a pseudo-centric distribution will necessarily reflect this constraint in the form of a bimodal distribution centered on 0 and π. This bipartite distribution is represented as a sum of two independent distributions centered on 0 and π; however, the resulting distribution must be renormalized (Equation 4.33-4.34). Ultimately, the conditional phase distribution is useful, because now the conditional random variables can be marginalized (Equations 4.33 and 4.35) with the appropriate distributions (Equations 4.12-4.13) to obtain *P*(φ|E) and *P*(φ).

$$P_0(\varphi|E, E_c) = \frac{P(\varphi, E|E_c)}{P(E|E_c)} = \frac{1}{2\pi} I_0^{-1}\left[\frac{2DEE_c}{1-D^2}\right] \exp\left[\frac{2DEE_c\cos\varphi}{1-D^2}\right] \quad (4.30)$$

$$P_{\pi}(\varphi|E,E_c) = \frac{1}{2\pi} I_0^{-1}\left[\frac{2DEE_c}{1-D^2}\right] \exp\left[\frac{2DEE_c\cos(\varphi+\pi)}{1-D^2}\right] \quad (4.31)$$

$$P(\varphi = 0\ or\ \varphi = \pi) = P(\varphi = 0) + P(\varphi = \pi) \quad (4.32)$$

$$P(\varphi|E) = \frac{1}{\tau}\int_0^{\infty}[P_0(\varphi|E,E_c) + P_{\pi}(\varphi|E,E_c)]\,P(E_c)dE_c \quad (4.33)$$

$$\tau = \int_0^{2\pi}\int_0^{\infty}[P_0(\varphi|E,E_c) + P_{\pi}(\varphi|E,E_c)]\,P(E_c)dE_c d\varphi \quad (4.34)$$

$$P(\varphi) = \int_0^{\infty} P(\varphi|E)\,P(E)dE \quad (4.35)$$

The existence of a closed form, analytic solution to *P*(φ), which results from marginalization of *E* and E^c, is unclear. Therefore numerical integration of the double integral (Equation 4.36) was implemented in *MATLAB* (details regarding numerical integration are described in *Appendix*).

$$P(\varphi) = \frac{\sqrt{2}}{\pi}\iint_0^{\infty} I_0^{-1}\left[\frac{2DEE_c}{1-D^2}\right]\exp\left[\frac{2DEE_c\cos\varphi}{1-D^2}\right]\exp\left[-\frac{2E^2+E_c^2}{2}\right]dE_c dE$$

(4.36)

The conditional distribution *P*(φ|*E*) is potentially useful for the phase recombination procedure described by Sawaya *et al.*[22] To assess the impact that the magnitude of *E*, the normalized pseudo-centric structure factor amplitude, has on the shape of the conditional distribution, *P*(φ|*E*) was plotted for varying values of *E* at a constant *D*-value (Figure 4.19). As *E* increases, the phase distribution becomes more 'centric', and the distribution decreases in width around 0 and *π* radians. In the procedure described by Sawaya *et al.*, the 'best centric phase' was assumed to be equally valid for all reflections;

however, given the current observation, the approximation is more likely to be valid for stronger reflections.

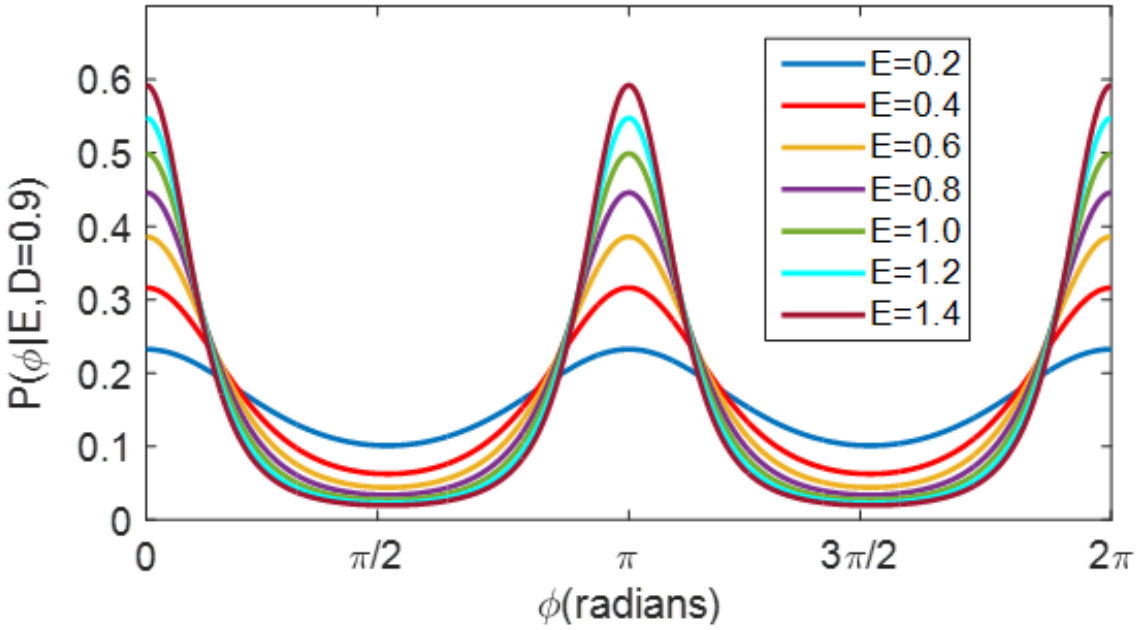


Figure 4.19. The conditional distribution, P(φ|E), shows an increase in 'centric character' for higher E values, i. e. for stronger reflections.

To further test the validity of *P*(φ|*E*), and by extension *P*(φ), distributions were plotted at varying pseudosymmetry *D*-values given a constant *E* (Figure 4.20). As expected, higher *D*-values yielded sharper, 'more centric' distributions compared to distributions with lower *D*-values. Another interesting feature of the distribution set is the variation in relative increase in the height, and thus decrease in width, for each incremental increase in *D*-value. For instance, the increase in the distribution height as *D* changes from 0.8 to 0.9 is substantially larger than the increase in height as *D* changes from 0.5 to 0.6. So, although there appears to be a linear relationship between the pseudosymmetry *D*-value and the deviation of atomic coordinates from centrosymmetric

positions, Δ***r***, it appears the change in 'distribution character', *i.e.*, centric vs. acentric, resembles a more exponential transition as a function of the pseudosymmetry *D*-value.

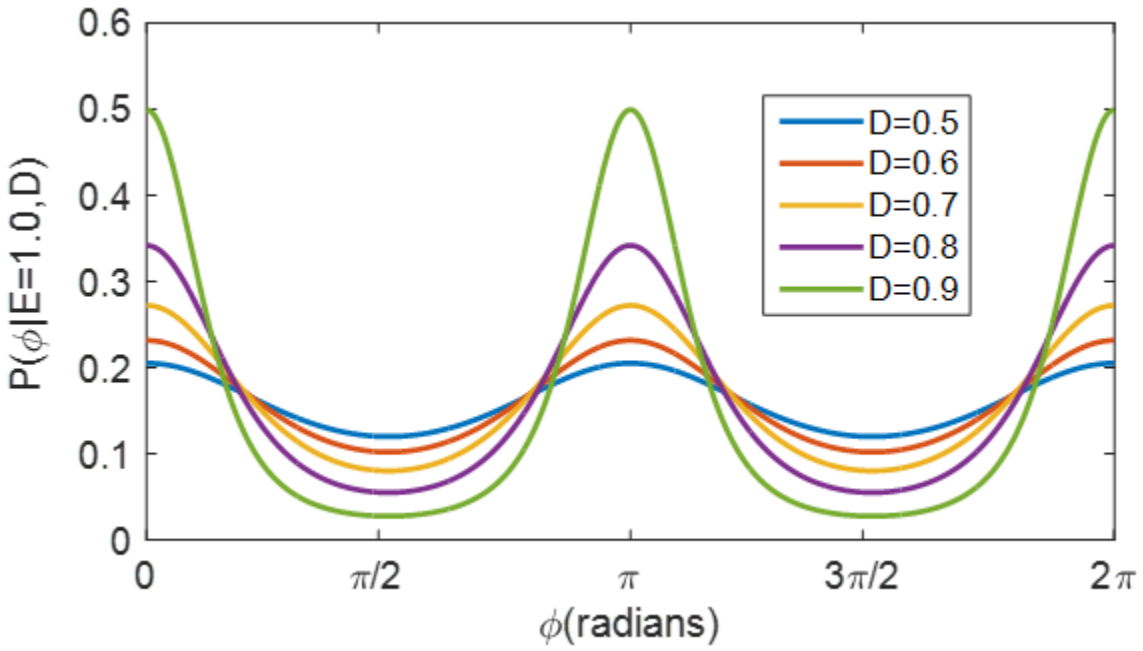


Figure 4.20. The conditional distribution, P(φ|E), plotted at varying pseudosymmetry D-values shows an exponential increase in 'centric character' as D → 1.

Finally, the theoretical probability distributions, *P*(φ), were compared to experimental distributions derived from each β substituted VHP quasiracemate structure. *D*-values were initially determined from the measured normalized structure factor amplitudes and in turn used to generate *P*(φ) via numerical integration. For clarification, the experimental phase distribution histograms were phase-shifted by +*π* radians for convenient comparison with the theoretical *P*(φ). This shift has no bearing on the results due to the fact that the distribution is symmetrical about 0 and *π*. In two cases (VHP quasiracemates Q26ACPC and K30APC), the theoretical distributions overlay well with the model-derived histograms. In the other two cases (VHP quasiracemates Q26β^{3}hQ

and K30β^{3}K; Figure 4.21 and Figure 4.9), the amplitude-derived *D*-value seems to overestimate the degree of centrosymmetry within the model. Intuitively, this overestimation seems reasonable given that no pseudosymmetry restraints were employed during structure refinement; therefore, the model was free to diverge from centrosymmetry during refinement. In the two instances where the amplitude-derived *D*-value overestimated the pseudosymmetry, a second *D*-value was determined that fit more closely fit the model phases (VHP Q26β^{3}hQ: amplitude *D*=0.925; model phase *D*,=0.87; VHP K30β^{3}hK: amplitude *D*=0.907, model phase *D*=0.820). Furthermore, these models also displayed the largest deviation from centrosymmetry among pseudo-inversion related pairs of chains (Figure 4.11). The ability to predict the phase deviation from centrosymmetry based on the measured structure factor amplitudes will likely be useful for generating higher quality maps from SAD phasing, which could open the door to combining quasiracemic crystallization with weak SAD signals such as those that arise from sulfur or phosphorus. Another exciting possibility is that the pseudo-centric phase distribution could be combined with density modification or radiation-induced damage phasing (RIP) to improve electron density map quality or possibly even to solve the phase problem *de novo*.[27]

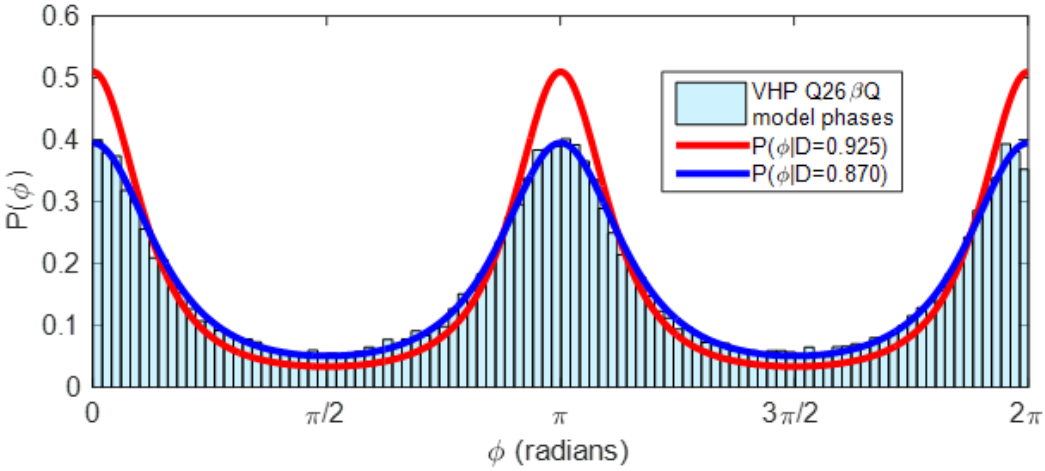


Figure 4.21. The VHP Q26β^{3}hQ quasiracemate model derived phase angles were plotted as a normalized histogram and compared with the probability density function generated from numerical integration of Equation 4.36. The amplitude derived pseudosymmetry *D*-value (red) appeared to overestimate the degree of pseudosymmetry. A second *D*-value (blue) was determined that more closely matched the model phases. The 0.05 difference in *D*-values corresponds to a ~0.1 Å RMSD from centrosymmetric positions based on Figure 4.11.

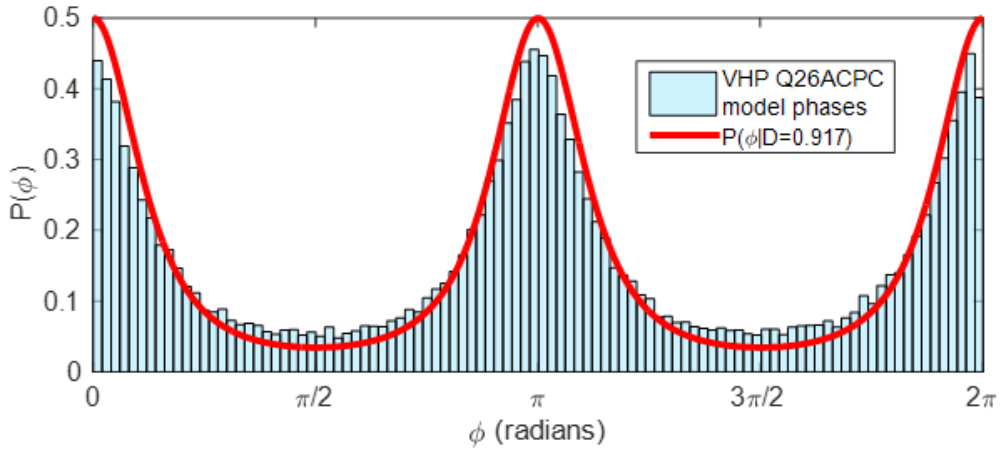


Figure 4.22. The VHP Q26ACPC quasiracemate model derived phase angles were plotted as a normalized histogram and compared with the probability density function generated from numerical integration of Equation 4.36. The amplitude derived pseudosymmetry *D*-value (red) appeared to overlay well with the model derived phase angles.

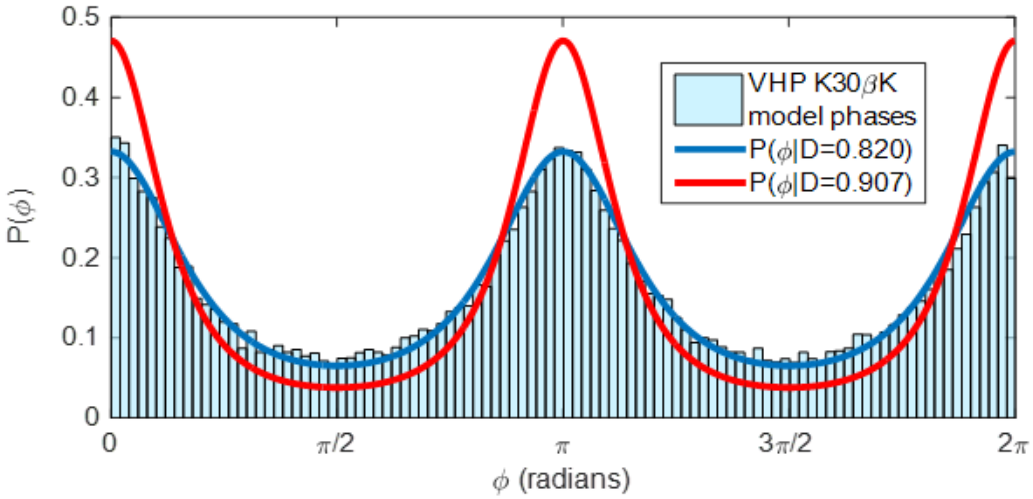


Figure 4.23. The VHP K30β[3]hK quasiracemate model derived phase angles were plotted as a normalized histogram and compared with the probability density function generated from numerical integration of Equation 4.36. The amplitude derived pseudosymmetry *D*-value (red) appeared to overestimate the degree of pseudosymmetry. A second *D*-value (blue) was determined that more closely matched the model phases. The 0.08 difference in *D*-values corresponds to a ~0.15 Å RMSD from centrosymmetric positions based on Figure 4.11.

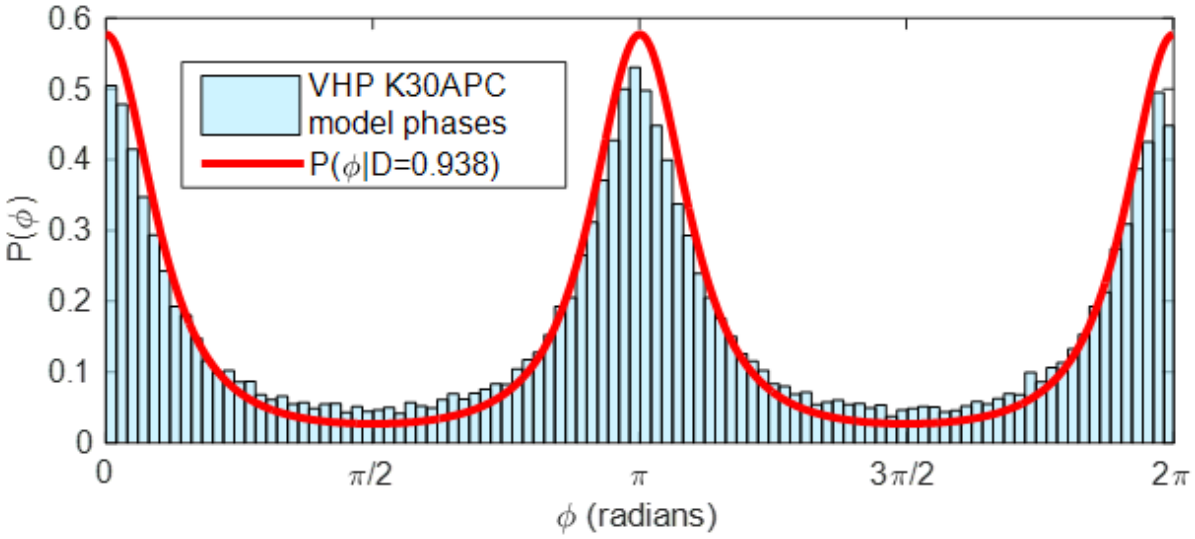


Figure 4.24. The VHP Q26ACPC quasiracemate model derived phase angles were plotted as a normalized histogram and compared with the probability density function generated from numerical integration of Equation 4.36. The amplitude derived pseudosymmetry D-value (red) appeared to overlay well with the model derived phase angles.

4.2.4 The Impact of Pseudosymmetry on the Largest Likely R-factor

The R-factor is the most ubiquitous global quality indicator for X-ray crystal structures. Many different variations of the R-factor have been described, but in general the aim is to quantify the discrepancy between the structure factor amplitudes predicted by the model and the amplitudes deduced from the measured intensities. Wilson was the first to consider the largest likely values of the R-factor for structures of completely random atoms based on the probability distributions for centrosymmetric and non-centrosymmetric reflections. It was discovered that the largest likely R-factors were 0.828 (centrosymmetric) and 0.586 (non-centrosymmetric). Presumably, this disparity can be ascribed to the greater dispersion in centrosymmetric amplitude distributions.

This initial result lends itself to the hypothesis that structures with some degree of centrosymmetry will have higher R-factors than purely non-centrosymmetric structures. Moreover, the width of a pseudo-centric distribution seems to approach that of a centric distribution in the limiting case ($D \rightarrow 1$); therefore, the largest possible R-factor for a pseudo-centric distribution likely assumes some intermediate value between 0.586 and 0.828. To test this hypothesis, the procedure followed by Wilson to determine the largest likely R-factor was combined with the pseudo-centric probability distribution at varying pseudosymmetry *D*-values.[28,29] The R-factor for a structure comprised of random atoms can be determined with the expression:

$$R = \frac{\langle X \rangle}{\langle |F| \rangle} \qquad (4.37)$$

where X is a random variable that can be achieved by either $F_2=F_1+X$ or $F_2=F_1-X$, and $\langle|F|\rangle$ is the mean value for the structure factor amplitude. From this fact a new probability distribution $Q(X)$ can be described, where $Q(X)$ is the probability that X lies between X and $X+dX$:

$$Q(X) = \int_0^\infty P(F)P(F+X)dF + \int_X^\infty P(F)P(F-X)dF \quad (4.38)$$

The complete derivation has been described by Wilson.[28] The general procedure for explicitly calculating the R-factor from Equation 4.37 involves the following two expressions:

$$R = 2 - 4\frac{\langle G(F)\rangle}{\langle|F|\rangle} \quad (4.39)$$

$$\langle G(F)\rangle = \int_0^\infty \left(\int_0^F FP(F)dF\right)dF \quad (4.40)$$

For the typical centric and acentric distributions, straightforward analytic expressions can be derived for $\langle G(F)\rangle$ and $\langle|F|\rangle$ via integration by parts; however, for the pseudo-centric distribution (Equation 4.23) a simple closed form expression is unknown; therefore, numerical integration was required to solve for R (details regarding numerical integration of the numerator and denominator of Equation 4.39 are provided in the appendix).

Finally, R was determined numerically for pseudosymmetry D-values ranging from 0 to 1 (Figure 4.25). Satisfyingly, the largest likely R-factor approaches 0.586 as $D \rightarrow 0$ (acentric limiting case) and 0.828 as $D \rightarrow 1$ (centric limiting case). Numerical integration fails in the neighborhood of the centric limiting case because as $D \rightarrow 1$, $I_0[X] \rightarrow \infty$. Knowledge of the largest likely R-factor is useful in the early stages of structure solution and refinement. For instance, a completely wrong molecular replacement solution will have an R-factor of ~0.6 for a chiral space group and ~0.8 for an achiral space group.

The curve in Figure 4.25 provides an alternative perspective on the phenomenon observed in the previous section, the exponential model error as a function of the pseudosymmetry *D*-value. As *D* goes from 0.4 to 0.5, the largest likely *R*-factor increases by ~0.02, whereas when *D* goes from 0.8 to 0.9, the largest likely *R*-factor increases by ~0.18, almost an order of magnitude change in the relative increase.

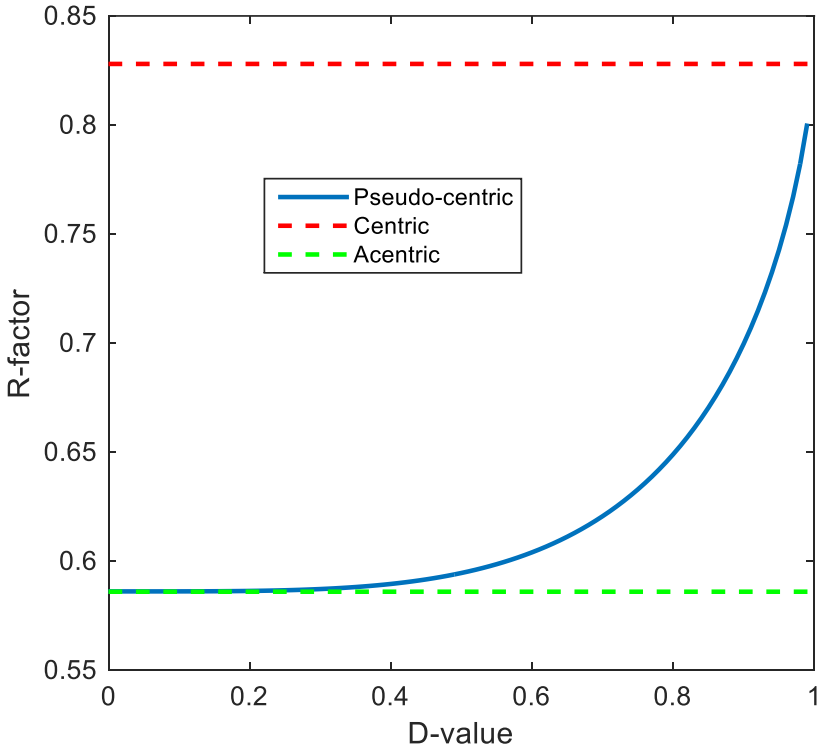


Figure 4.25. The largest likely R-factor for a structure comprised of random atoms as a function of pseudosymmetry. The dotted lines represent the limiting centric (red, R=0.828) and acentric (green, R=0.586) cases. Numerical integration fails in the neighborhood of the centric limiting case due to the argument in the modified Bessel function term, $I_0[X]$, because as D → 1, $I_0[X]$ → ∞.

4.2.5 Simulating the Impact of Pseudo-Inversion Symmetry on Maximum Likelihood Refinement

Maximum-Likelihood based refinement of macromolecular crystal structures involves minimization of the negative log-likelihood function (Equation 4.17 and 4.22).[5] The conditional distribution that comprises the basis for the likelihood function is based on the assumption that the real and imaginary components of the complex structure factor are independent and thus uncorrelated. These two uncorrelated random variables can be used to construct a bivariate normal distribution for which the correlation coefficient, ρ, is equal to zero. In a non-centrosymmetric structure this independence implies a uniform phase distribution, *i.e.*, the structure factor can rotate freely in the complex plane. However, in the presence of pseudo-inversion symmetry, where atoms deviate slightly from centrosymmetric positions, the phase angle distribution is altered so that now a narrow band of phases centered on 0 or π radians is favored, which leads to the question, what effect does this altered distribution have on the likelihood function (Equation 4.17)?

. The conventional likelihood function for purely acentric structure factors (Equation 4.17) has the form of a Rician distribution, originally developed to model signal decay in wireless communication networks.[30] The distribution is essentially an uncorrelated bivariate normal distribution centered on an arbitrary vector in the complex plane (Figure 4.4). As the length of the arbitrary vector increases, the Rician distribution begins to approximate a normal distribution. The tools to address the question of what impact does a *correlated* bivariate normal distribution have on the likelihood function can be drawn from the signal processing field. Beckmann first described a generalized version of the Rician distribution that includes a non-zero correlation coefficient in the off-origin bivariate normal distribution.[30] However, a shortcoming of Beckmann's distribution is the

lack of a convenient, closed-form expression. Nevertheless, the Beckmann distribution can be simulated with a multivariate random number generator and a covariance matrix of the form:

$$\Sigma = \begin{bmatrix} \sigma^2 & \rho\sigma^2 \\ \rho\sigma^2 & \sigma^2 \end{bmatrix} \qquad (4.41)$$

$$P(\boldsymbol{E}) = \frac{1}{2\pi\sqrt{|\Sigma|}} \exp\left[-\frac{1}{2}(\boldsymbol{E}-\boldsymbol{\mu})^T \Sigma^{-1}(\boldsymbol{E}-\boldsymbol{\mu})\right] \qquad (4.42)$$

For the general acentric distribution, the correlation coefficient, ρ, is equal to zero, and the distribution is centered at the origin (Figure 4.26). The effects of pseudo-centric symmetry on a bivariate distribution, such as a complex structure factor, can be replicated with a non-zero correlation coefficient, a fact that was recognized by Wilson in 1980.[18] To confirm the validity of simulating normalized structure factor amplitude probability distributions with a random number generator, the acentric Wilson distribution was reproduced (Figure 4.27).

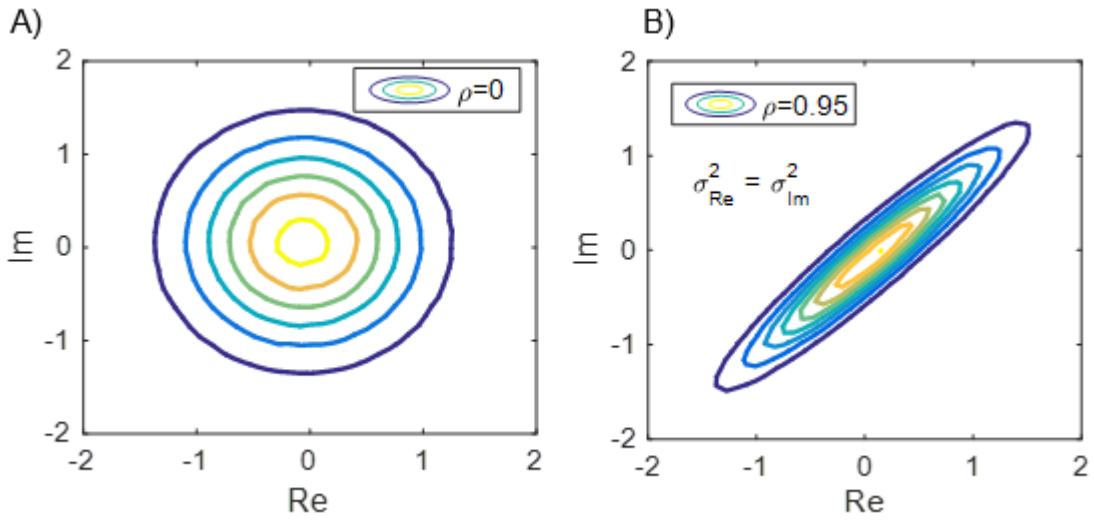


Figure 4.26. A) Uncorrelated (ρ=0) and B) correlated (ρ=0.95) bivariate normal distributions are shown with means centered at the origin.

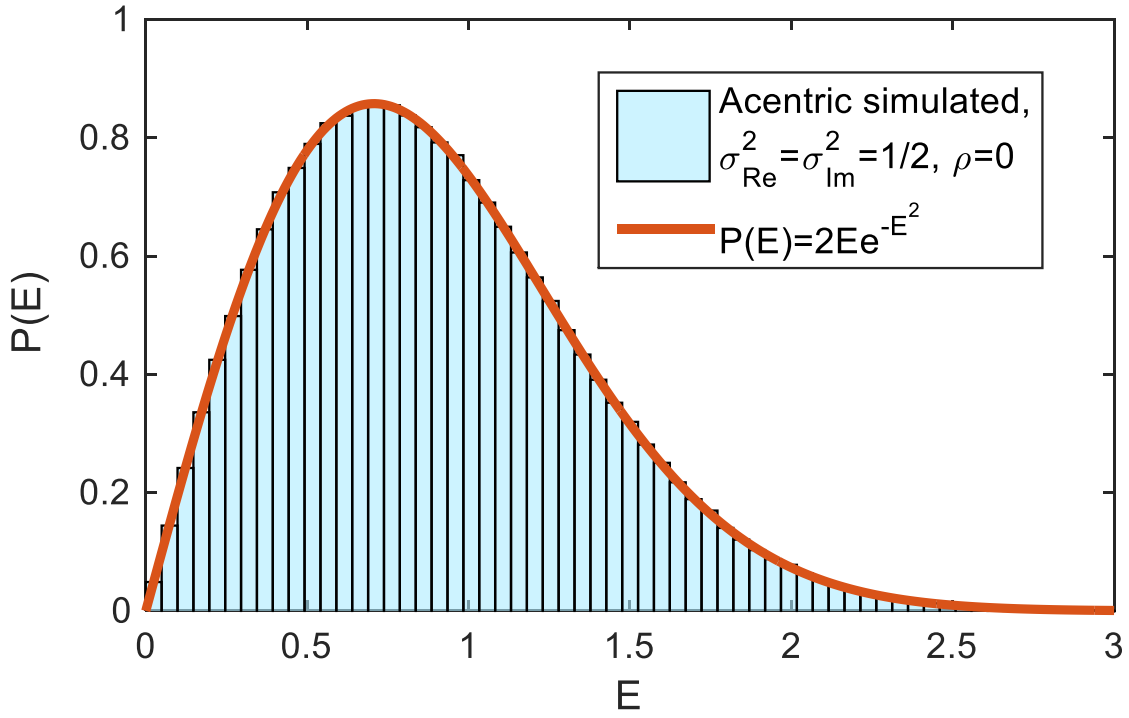


Figure 4.27. The acentric Wilson distribution can be reproduced with a random number generator. The real and imaginary components are simulated as random numbers with $\sigma_{Re}^2 = \sigma_{Im}^2 = 1/2$ and $\mu_{Re} = \mu_{Im}=0$. The random number generator creates a 2xN matrix and the amplitude is the root mean square of each row vector within the matrix.

The simulated random vectors can be rotated to align with the real and imaginary axes to reflect the centrosymmetric phase constraint of 0 and π radians. Structures with pseudo-inversion symmetry display a small amount of dispersion that results in an imaginary contribution to the structure factor, and this fact is reflected in the elliptical shape of the distribution (Figure 4.28). Initial generation of the vectors assumes the variances of both real and imaginary components are equal, but the rotation causes a change in the relative magnitude of each variance, where the contribution from the real component is now much larger than that of the imaginary component. Finally, a phase angle distribution of the rotated system reveals a symmetrical, bimodal distribution centered on 0 and π radians (Figure 4.29). The width of the phase distribution can be

modulated with the correlation coefficient, ρ; as $\rho \rightarrow 0$, a particular band of phases is no longer favored, and the distribution simply becomes a uniform distribution.

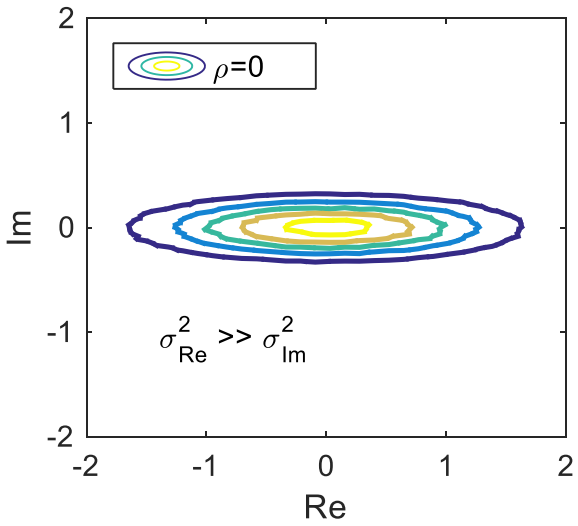


Figure 4.28. A 45° rotation of each random vector aligns the major and minor axes of the elliptical probability distribution with the real and imaginary axes. The two random variables are now uncorrelated but they are not independent. The lengths of the elliptical axes are attenuated by the variance for each random variable, so the rotation has the effect of altering the real and imaginary component variances.

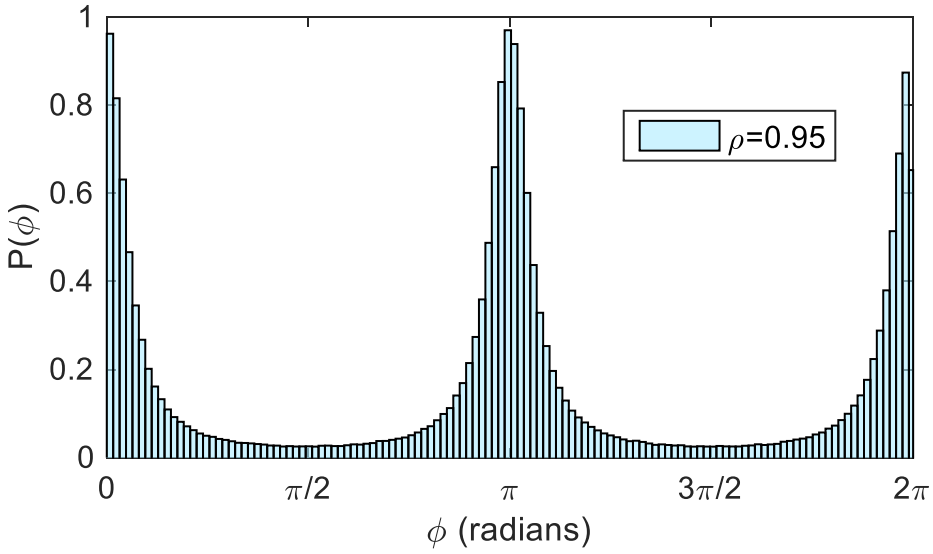


Figure 4.29. The phase angle distribution after a 45° rotation so that the elliptical axes of the distribution are aligned with the real and imaginary axes.

Differences in the shape of acentric and pseudo-centric distributions will ultimately be manifest in the conditional probability distribution used as the basis for the likelihood function (Equation 4.17). This conditional distribution, simply stated, takes an observed and a model structure factor amplitude and computes the probability that there is a vector with the appropriate amplitude to span the difference (Figure 4.4). Therefore, the shape of the distribution governing the 'difference spanning' vector will likely affect the conditional distribution. Furthermore, the amplitude and phase angle of the given model structure factor will also play a role in shaping the conditional distribution. To gain insight into the effect that pseudo-inversion symmetry has on the likelihood function, and by extension the overall refinement, conditional distributions were simulated by centering a rotated, *correlated* bivariate normal distribution onto hypothetical model vectors of varying magnitude and phase angle. The resulting distributions were then qualitatively compared to standard Rician distributions, which represent the truly acentric case.

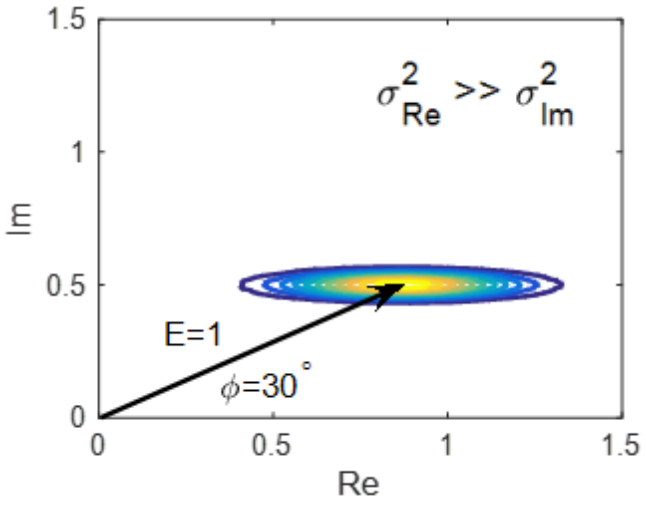


Figure 4.30. Graphical example of the simulated conditional distribution. A bivariate distribution comprised of dependent random variables is centered at the end of the structure factor depicted with magnitude E=1 and φ=30°. A histogram is then generated from the magnitude of the vectors defined by the distribution.

A pseudo-centric distribution was first generated by scaling a correlated (ρ=0.95) bivariate normal distribution ($\sigma_{Re}^2 = \sigma_{Im}^2 = 1/2$, $\mu_{Re} = \mu_{Im}$=) by 0.2 and then centering the resulting distribution on the end of two sets of vectors (E=0.2, E=1.0). Each of the two sets of starting vectors were rotated through phase angles of 0°, 30°, 60°, and 90°. As an example, Figure 4.30 illustrates the case where E=1.0 and φ=30°. Finally, each of the eight distributions was fitted to a Rician distribution in *MATLAB*.

For the E=1.0 case, the starting structure factor is significantly larger than the simulated error, *i.e.*, $E >> \Delta E$. The resulting distributions overlayed well with the fitted Rician distributions at lower phase angles (0° and 30°; Figure 4.31). As the phase angles increased (60° and 90°), the distribution narrowed and deviated slightly from the fitted Rician distribution (Figure 4.32). A comparison of the 0° and 90° cases displays the severe narrowing of the distribution as the phase angle aligns with the imaginary axis (Figure 4.33). Intuitively this narrowing makes sense, because the major component of the total variance is aligned with the real axis, so the error shrinks as the real component makes a diminishing contribution to the starting structure factor, but overall it seems that a Rician distribution is a good approximation in the case of severe pseudo-centric symmetry when $E >> \Delta E$.

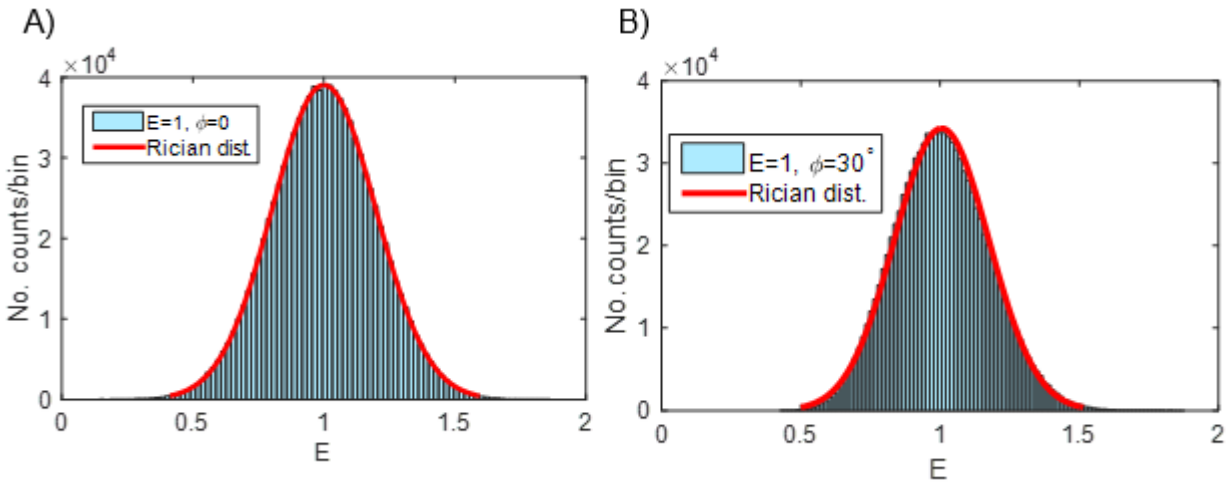


Figure 4.31. The simulated histograms for A) E=1.0, φ=0° and B) E=1.0, φ=30° show an overall agreement with the fitted Rician distributions.

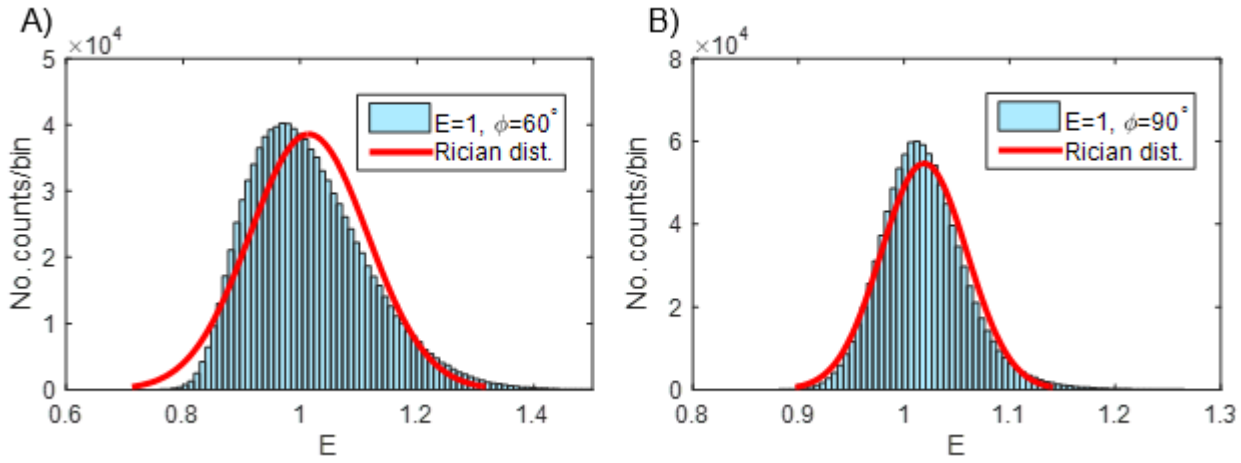


Figure 4.32. The simulated histograms for A) E=1.0, φ=60° and B) E=1.0, φ=90° show a slight deviation from the fitted Rician distributions.

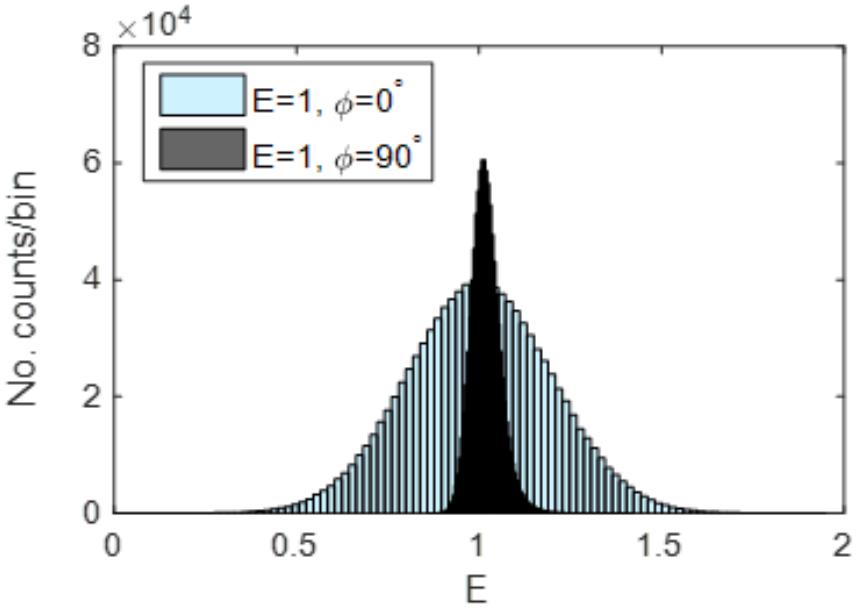


Figure 4.33. The width of the conditional distribution is inversely proportional to the phase angle between 0° and 90°.

For the *E*=0.2 case, the starting structure factor is similar in magnitude to the simulated error, *i.e.*, $E \approx \Delta E$. However, even at lower phase angles (0° and 30°) the simulated distributions differ markedly from the fitted Rician distributions (Figure 4.34). As the phase angle gradually increases (60° and 90°) the peak of the distribution begins to coalesce around *E*=0.2 (Figure 4.35). The sharpening trend seen in the *E*=1.0 case is apparent here; as φ increases, the distribution width decreases. These results indicate that if the starting model error is large, such as a partial molecular replacement structure solution, then weaker reflections may have a disproportionately deleterious effect on the refinement.

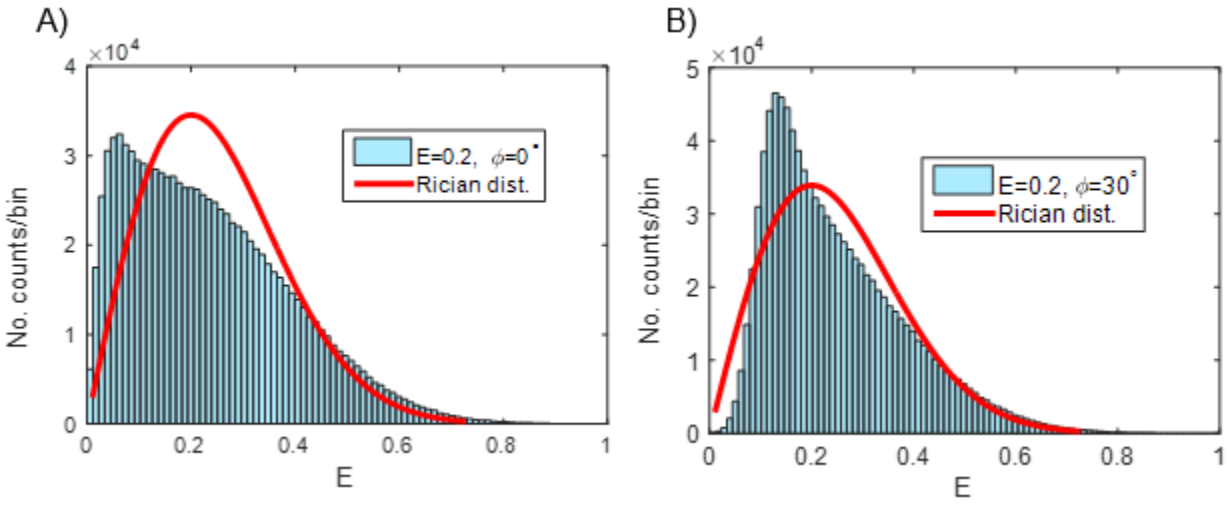


Figure 4.34. The simulated histograms for A) E=0.2, φ=0° and B) E=0.2, φ=30° show a substantial deviation from the fitted Rician distributions.

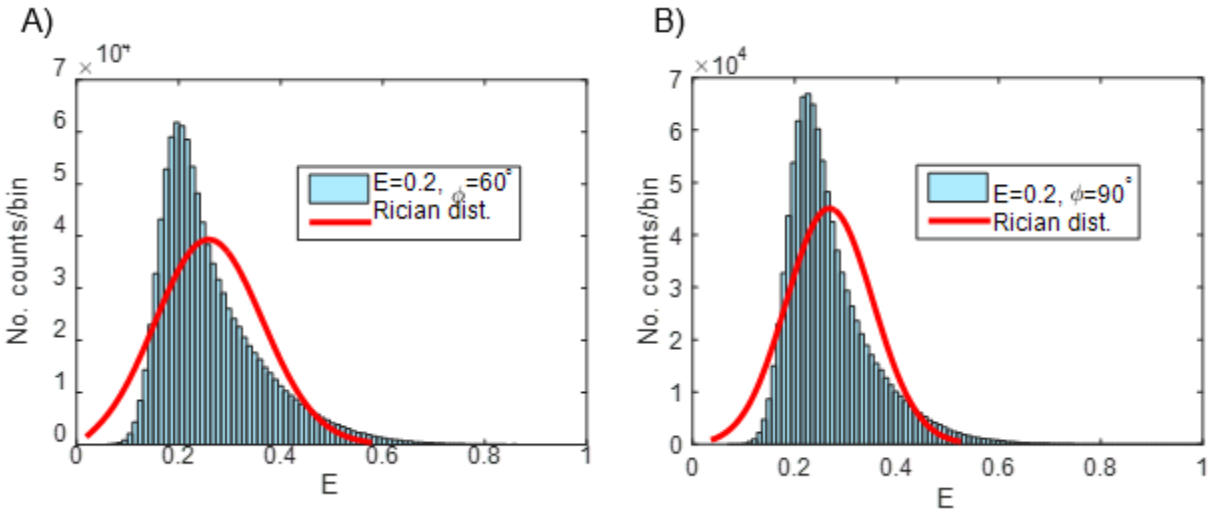


Figure 4.35. The simulated histograms for A) E=0.2, φ=60° and B) E=0.2, φ=90° show a substantial deviation from the fitted Rician distributions.

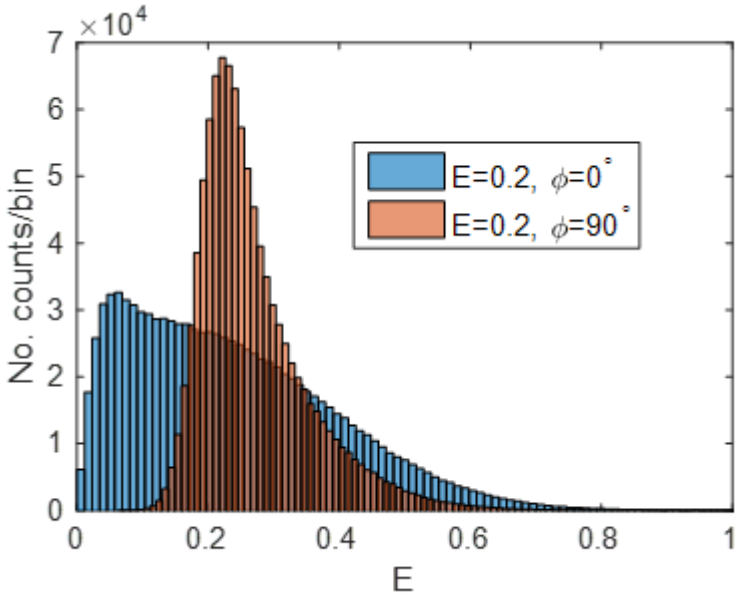


Figure 4.36. Similar to the case where E=1.0, the width of the distribution is inversely related to the phase angle; however, here the distributions are asymmetric and skewed toward lower E values.

Weaker reflections may have a negative impact on the early stages of structure refinement based on this simplistic analysis (the structure factors employed in the simulations were assumed to be normalized, so weaker normalized reflections will be present within each resolution shell and not localized to a specific region of the reciprocal lattice). Surprisingly, stronger reflections appear to still maintain the shape of a Rician distribution; however, at this time it is still unclear what impact pseudo-centric symmetry has on the width of the distribution, which plays a significant role in maximum-likelihood based refinement. These results suggest that a possible strategy for the early stages of quasiracemic structure solution and refinement may include flagging weak reflections for cross-validation in the initial rounds of refinement and then removing those flags as the model becomes more complete and the relative model error decreases.

4.2.6 Pseudo-Inversion Symmetry in Homochiral Structures

Centric intensity distributions are not strictly limited to centrosymmetric structures. Sohncke space groups (the 65 chiral space groups) can contain subsets of reflections that follow centric intensity distributions. For instance, a two-fold rotational symmetry axes mapped onto a plane perpendicular to the axis produces reflections that obey centric distributions. So in the case of a monoclinic space group such as $P2_1$, the second most prevalent space group among protein crystals, *hkl* → *lkh* by the two-fold symmetry, and a rotation of the two-fold axis projection onto the perpendicular plane (*hkl* → *h0l*) corresponds to *h0l* → *l0h*, which is equivalent to the inversion symmetry operation (*x,y,z*) → (*-x,-y,-z*). Given this information, the same type of pseudo-inversion symmetry statistics should apply to *0kl* or *hk0* reflections if a non-crystallographic two-fold symmetry axis is perpendicular to a crystallographic two-fold symmetry axis.

To test this hypothesis, a Protein Databank (PDB) search was performed among homodimeric structures that crystallized in space group $P2_1$. The search was narrowed to only include structures where β, the angle between *a* and *c* lattice vectors, was within ~2-3° of 90°. The reported structure PDB: 5BP2, a homodimeric polyketide synthase, fits the narrowed criteria.[31] 5BP2 is a primitive, monoclinic crystal with two of the functional homodimers in the asymmetric unit (Figure 4.37). Interestingly, the structure did not crystallize with the crystallographic 2_1-screw axis aligned with the natural C2 axis of the functional homodimer, thus making the structure perfectly suitable for pseudo-inversion symmetry analysis.

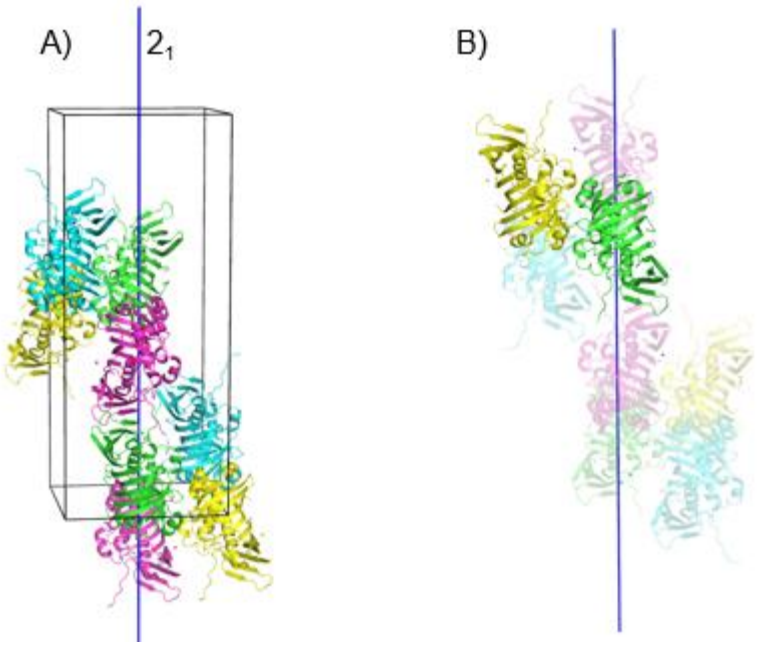


Figure 4.37. A) All chains present in the 5BP2 unit cell are shown. Cyan/magenta and yellow/green color combinations show the functional homodimer not related by any crystallographic symmetry. B) The homodimer related by a pseudo-two-fold axis, parallel with the a lattice vector, is shown for a yellow/green homodimer. In each case the 2_1 screw axis is shown as a dark blue line.

First, measured amplitudes were normalized with the CCP4 program *ecalc* and then *0kl* reflections were extracted from the reflection file with the program *mtzutils*. An empirical probability distribution was determined with *MATLAB* for the 5BP2 *0kl* reflections, fitted with a pseudosymmetry *D*-value, and then compared to the centric and acentric distributions. A *D*-value of 0.875 was found for the pseudo-two-fold symmetry axis, which is consistent with the pseudo-inversion symmetry hypothesis (Figure 4.38). This result suggests that knowledge of *D*-values could potentially enhance current methods for detecting non-crystallographic symmetry, where applicable.

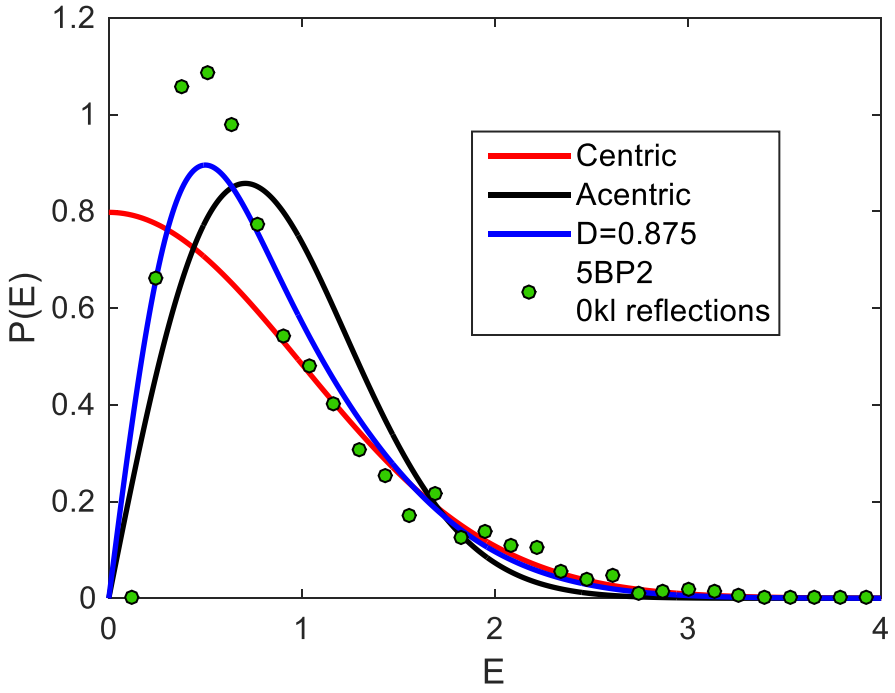


Figure 4.38. The normalized structure factor amplitude distribution for 0kl reflections from the deposited 5BP2 dataset.

4.3 Conclusions

Pseudo-inversion symmetry was identified and quantified with a single parameter for the series of β-substituted VHP quasiracemic structures described in Chapter 2 and for a series of mono-ubiquitin/di-ubiquitin quasiracemate structures reported by Liu *et al.*,[20]. For the β-substituted VHP quasiracemic structures, the empirical normalized structure factor amplitude probability distributions agreed well with the 'pseudo-centric' or 'sub-centric' distribution, fitted with the appropriate pseudosymmetry *D*-value, described by Srinivasan.[16] There was also a direct correlation between the RMSD value between pseudosymmetry related backbone atoms and the pseudosymmetry *D*-value, which suggests that the pseudosymmetry *D*-value is aptly suited for quantifying pseudosymmetry based solely on normalized structure factor amplitudes. The mono-

ubiquitin/di-ubiquitin quasiracemate structure normalized amplitude distribution displayed more centric character (rather than pseudo-centric). I ascribed this observation to the relative size and position of the symmetry-disrupting modification. This result is consistent with the notion that deviations from pseudo-inversion symmetry are manifest in the structure factor amplitude probability distribution. The symmetry-disrupting modification in the ubiquitin quasiracemates is small (a single residue in a flexible region) compared to the 150 residue dimer. Next, theoretical phase distributions based on the pseudosymmetry *D*-value were found to agree with quasiracemate model phases for the β-substituted VHP series. These efforts should complement the quasiracemic-SAD phasing strategy described by Sawaya *et al.*[22] Furthermore, it was shown that the largest likely *R*-factor for a quasiracemic structure increases exponentially as a function of the pseudosymmetry *D*-value after $D > 0.5$, until the *R*-factor matches that of a purely centrosymmetric structure. Finally, pseudo-inversion symmetry may alter the shape of the likelihood function employed in maximum likelihood refinement when a reflection amplitude derived from the model is similar in magnitude to the model error amplitude. This result suggests that flagging weaker, normalized reflections ($E < 1.0$) for cross validation during early rounds of structure solution and refinement may lead to higher quality electron density maps. Finally, a pseudo-centric distribution was shown to arise from a pseudo-two-fold rotational axis in a homochiral structure. This result illustrates the utility of the pseudosymmetry *D*-value for quantifying non-crystallographic symmetry in macromolecular crystals comprised of homochiral molecules.

4.4 Appendix

4.4.1 Derivation of the Pseudo-Centric Normalized Amplitude Distribution

The complex structure factor can be separated into real and imaginary components *A* and *B* respectively:

$$\boldsymbol{F} = \sum_{i=1}^{N} f_i e^{2\pi i \boldsymbol{h}\cdot\boldsymbol{x}_i} = A + iB \qquad (4.43)$$

The probability distribution of $\Delta\boldsymbol{F}$ can in turn be described based on the probability distributions of the real and imaginary components of the structure factor (Equation 4.44).

$$P(\Delta\boldsymbol{F})d(\Delta\boldsymbol{F}) = P(\Delta\boldsymbol{A})P(\Delta\boldsymbol{B})d(\Delta\boldsymbol{A})d(\Delta\boldsymbol{B}) \qquad (4.44)$$

Where the probability distribution of *A* and *B* have been described previously by Luzzati.

$$P(\Delta\boldsymbol{A}) = \frac{1}{\sqrt{\pi\Sigma(1-D^2)}} e^{-\frac{\Delta\boldsymbol{A}^2}{\Sigma(1-D^2)}} \quad (4.45)$$

$$P(\Delta\boldsymbol{B}) = \frac{1}{\sqrt{\pi\Sigma(1-D^2)}} e^{-\frac{\Delta\boldsymbol{B}^2}{\Sigma(1-D^2)}} \quad (4.46)$$

$$\Sigma = \sum_{i=1}^{N} f_i^2 \qquad (4.47)$$

$$D = \langle \cos 2\pi\Delta\vec{\boldsymbol{r}}\cdot\vec{\boldsymbol{s}}\rangle, \;\; 0 \le D \le 1 \qquad (4.48)$$

Σ is the sum of the square of the atomic scattering factors of the *N* atoms in the unit cell and *D* is analogous to the Luzzati factor with the important distinction that Δ***r*** is the deviation from centrosymmetry; Δ***r*** is assumed to follow a normal distribution and to be equivalent for each atom in the unit cell. It must be noted that a value of *D*=1 corresponds to a purely centrosymmetric structure and *D*=0 corresponds to a purely non-centrosymmetric structure. The reciprocal space vector ***s*** is the conventional term where |***s***| = sin *θ*/*λ*; *θ* is the scattering angle and *λ* is the wavelength of the incident X-rays. Equations 4.45 and 4.46 can be combined to describe the probability distribution of the resultant structure factor arising from the deviation from centrosymmetry.

$$P(\Delta \boldsymbol{F}) = \frac{1}{\pi\Sigma(1-D^2)} e^{-\frac{\Delta \boldsymbol{A}^2 + \Delta \boldsymbol{B}^2}{\Sigma(1-D^2)}} = \frac{1}{\pi\Sigma(1-D^2)} e^{-\frac{\Delta \boldsymbol{F}^2}{\Sigma(1-D^2)}} \qquad (4.49)$$

Equation 4.49 can be modified to describe a conditional probability of the resulting acentric structure factor $\boldsymbol{F}_N$ by applying the cosine rule to the diagram depicted in Figure 4.5.

$$P(\boldsymbol{F}_N; \boldsymbol{F}_N^C) = \frac{1}{\pi\Sigma(1-D^2)} e^{-\frac{|\boldsymbol{F}_N|^2 + |\boldsymbol{F}_N^C|^2 - 2D|\boldsymbol{F}_N||\boldsymbol{F}_N^C|\cos\alpha}{\Sigma(1-D^2)}} \qquad (4.50)$$

To compare the structure factor statistics between the different quasiracemates it is necessary to bring everything into the same scale by using normalized structure factors.

$$\boldsymbol{E} = \frac{\boldsymbol{F}}{\sqrt{\sum_{i=1}^{N} f_i^2}} \qquad (4.51)$$

Structure factor normalization and inclusion of the phase angle, α, between the centric and acentric structure factors results in the following conditional, joint probability distribution where the term $|\boldsymbol{E}_N|$ has been introduced to accommodate the switch from Cartesian to polar coordinates.

$$P(|\boldsymbol{E}_N|, \alpha; \boldsymbol{E}_N^C) = \frac{|\boldsymbol{E}_N|}{\pi(1-D^2)} e^{-\frac{|\boldsymbol{E}_N|^2 + |\boldsymbol{E}_N^C|^2 - 2D|\boldsymbol{E}_N||\boldsymbol{E}_N^C|\cos\alpha}{(1-D^2)}} \qquad (4.52)$$

The phase angle variable, α, is removed by integration over the range [0,2π] which results in the modified Bessel function of the first kind, I_0.

$$P(|\boldsymbol{E}_N|; \boldsymbol{E}_N^C) = \frac{2|\boldsymbol{E}_N|}{(1-D^2)} e^{-\frac{|\boldsymbol{E}_N|^2 + |\boldsymbol{E}_N^C|^2}{(1-D^2)}} I_0\left[\frac{2D|\boldsymbol{E}_N||\boldsymbol{E}_N^C|}{1-D^2}\right] \qquad (4.53)$$

Equation 4.53 can be combined with the centric Wilson distribution (Equation 4.54) to create a joint probability distribution (Equation 4.55).

$$P(|\boldsymbol{E}_N^C|) = \sqrt{\frac{2}{\pi}} e^{-\frac{|\boldsymbol{E}_N^C|^2}{2}} \qquad (4.54)$$

$$P(|\boldsymbol{E}_N|, |\boldsymbol{E}_N^C|) = \sqrt{2/\pi}\, \frac{2|\boldsymbol{E}_N|}{(1-D^2)} e^{-\frac{2|\boldsymbol{E}_N|^2+(1+D^2)|\boldsymbol{E}_N^C|^2}{2(1-D^2)}} I_0\left[\frac{2D|\boldsymbol{E}_N||\boldsymbol{E}_N^C|}{1-D^2}\right] \qquad (4.55)$$

Finally, a marginal probability distribution that describes the distribution of $|\boldsymbol{E}_N|$ independent of $|\boldsymbol{E}_N^C|$ can be generated by integrating out the $|\boldsymbol{E}_N^C|$ term with the use of the definite integral described by Gradshteyn and coworkers (Equation 4.56).

$$\int_0^\infty e^{-\gamma x^2} I_0(\beta x) dx = \frac{\pi}{2\gamma} e^{\frac{\beta^2}{8\gamma}} I_0\left(\frac{\beta^2}{8\gamma}\right) \; [Re\, \gamma > 0] \qquad (4.56)$$

$$P(|\boldsymbol{E}_N|) = \frac{2|\boldsymbol{E}_N|}{\sqrt{1-D^4}} e^{-\frac{|\boldsymbol{E}_N|^2}{(1-D^4)}} I_0\left[\frac{D^2|\boldsymbol{E}_N|^2}{1-D^4}\right] \qquad (4.57)$$

4.4.2 Numerical Integration of Equation 4.33 in MATLAB

```
function int_bin = phase_pdf_pseudo(E,D)
phi = 0:0.01:2*pi;
Ec = 0:0.01:5;
P_Ec = sqrt(2/pi)*exp(-Ec.^2/2);
int_bin = zeros(size(phi));
%integrate P(phi|E,Ec)P(Ec)dEc, then normalize
for k = 1:length(phi)
    term1 =((1/2*pi)*exp(2*D*E*Ec*cos(phi(k))/(1D^2))./...
            besseli(0,2*D*E*Ec/(1D^2))).*P_E;
    term2 = ((1/2*pi)*exp(2*D*E*Ec*cos(phi(k)+pi)/(1-D^2))./...
            besseli(0,2*D*E*Ec/(1-D^2))).*P_Ec;
    term = term1 + term2;
    int_bin(k) = trapz(term)./(0.2*length(Ec));
end
norm_term = trapz(int_bin)*2*pi./(length(phi));
int_bin = int_bin/norm_term;
end
```

4.4.3 Numerical Integration of Equation 4.36 in MATLAB

```
function [P_phi] = phase_pdf_pseudo_unconditional(D)
Ec = 0:0.01:5;
P_ED = (2*Ec).*exp(-Ec.^2);
phi=0:0.01:2*pi;
int_bin= zeros(length(Ec),length(phi));
for k = 1:length(Ec)
    P_phi_given_E = phase_pdf_pseudo(Ec(k),D);
    term = P_phi_given_E.*P_ED(k); %vector of length phi
    int_bin(k,:) = term; % int_bin will be a Ec x phi sized matrix, need to
                         % integrate down columns
end
P_phi = trapz(int_bin)./(0.2*length(Ec));
end
```

4.4.4 Determination of the Largest Likely R-factors: Numerical Integration of Q(X)

```
function [ G, meanG, meanF, R_factor] = G_gen( F,D )

%function workflow:
%F is a normalized structure factor from 0:0.01:5;
%D a scalar value that correspond to pseudosymmetry;
%1) generate <G(F)>
%2) generate <|F|>
%3) compute R=2-4*<G(F)>/<|F|>
I0 = besseli(0,D^2*F.^2/(1-D^4));
expo = exp(-F.^2/(1-D^4));
P = (2*F.^2/sqrt(1-D^4)).*expo.*I0;
G=zeros(size(F));
for k=1:length(F)
    G(k) = trapz(P(1:k));
end
G=G/(0.25*length(F));
P_meanG = (2*F./sqrt(1-D^4)).*expo.*I0;
meanG = trapz(G.*P_meanG)/(0.25*length(F));
meanF = trapz(F.*P_meanG)/(0.25*length(F));
R_factor = 2-4*(meanG/meanF);
```

4.5 References

(1) Srinivasan, R.; Parthasarathy, S. *Some Statistical Applications in X-Ray Crystallography*; Pergamon Press: Elmsford, New York, USA, 1976.

(2) Wilson, A. J. C. *Acta Crystallogr.* **1949**, *2* (5), 318–321.

(3) Pannu, N. S.; Read, R. J. *Acta Crystallogr. Sect. A Found. Crystallogr.* **1996**, *A52* (5), 659–668.

(4) Giacovazzo, C. *Phasing in Crystallography: A Modern Perspective*; Oxford University Press: New York, NY, USA, 2013.

(5) Murshudov, G. N.; Vagin, A. A.; Dodson, E. J. *Acta Crystallogr. Sect. D Biol. Crystallogr.* **1997**, *53* (3), 240–255.

(6) Sheldrick, G. M. *Acta Crystallogr. Sect. C Struct. Chem.* **2015**, *71* (Md), 3–8.

(7) Lunin, V. Y.; Afonine, P. V.; Urzhumtsev, A. G. *Acta Crystallogr. Sect. A Found. Crystallogr.* **2002**, *58* (3), 270–282.

(8) Read, R. J. *Acta Crystallogr.* **1986**, *42*, 140–149.

(9) Lunin, V. Y.; Skovoroda, T. P. *Acta Crystallogr. Sect. A* **1995**, *51* (6), 880–887.

(10) Adams, P. D.; Afonine, P. V.; Bunkóczi, G.; Chen, V. B.; Davis, I. W.; Echols, N.; Headd, J. J.; Hung, L. W.; Kapral, G. J.; Grosse-Kunstleve, R. W.; McCoy, A. J.; Moriarty, N. W.; Oeffner, R.; Read, R. J.; Richardson, D. C.; Richardson, J. S.; Terwilliger, T. C.; Zwart, P. H. *Acta Crystallogr. Sect. D Biol. Crystallogr.* **2010**, *66* (2), 213–221.

(11) McCoy, A. J. *Acta Crystallogr. Sect. D Biol. Crystallogr.* **2004**, *60* (12 I), 2169–2183.

(12) Zwart, P. H.; Grosse-Kunstleve, R. W.; Lebedev, A. A.; Murshudov, G. N.; Adams, P. D. *Acta Crystallogr. Sect. D Biol. Crystallogr.* **2007**, *64* (1), 99–107.

(13) Read, R. J.; Adams, P. D.; McCoy, A. J. *Acta Crystallogr. Sect. D Biol. Crystallogr.* **2013**, *69* (2), 176–183.

(14) Brock, C. P. *Acta Crystallogr. Sect. B Struct. Sci.* **2016**, *72*, 807–821.

(15) Wukovitz, S. W.; Yeates, T. O. *Nat. Struct. Biol.* **1995**, *2* (12), 1062–1067.

(16) Srinivasan, R.; Swaminathan, P. *Curr. Sci.* **1975**, *44* (3), 71–73.

(17) Shmueli, U.; Wilson, A. J. C. *Acta Crystallogr. Sect. A* **1983**, *39*, 225–233.

(18) Wilson, A. J. C. *Acta Crystallogr. Sect. A* **1980**, *36* (6), 945–946.

(19) Giacovazzo, C.; Cascarano, G.; Zheng Chao-de. *Acta Crystallogr. Sect. A Found. Crystallogr.* **1988**, *44* (3), 291–294.

(20) Wang, J.; Gao, S.; Pan, M.; Zheng, Y.; Huang, Y.; Zheng, Q.; Sun, D.; Lu, L.; Tan, X.; Tan, X.; Lan, H.; Wang, J.; Wang, T.; Liu, L. *J. Am. Chem. Soc.* **2016**, *138* (43), 14497–14502.

(21) Pan, M.; Gao, S.; Zheng, Y.; Tan, X.; Lan, H.; Tan, X.; Sun, D.; Lu, L.; Wang, T.; Zheng, Q.; Huang, Y.; Wang, J.; Liu, L. *J. Am. Chem. Soc.* **2016**, *138* (23), 7429–7435.

(22) Sawaya, M. R.; Pentelute, B. L.; Kent, S. B. H.; Yeates, T. O. *Acta Crystallogr. Sect. D Biol. Crystallogr.* **2012**, *68* (1), 62–68.

(23) Rose, J. P.; Wang, B. C. *Arch. Biochem. Biophys.* **2016**, *602*, 80–94.

(24) Terwilliger, T. C. *Acta Crystallogr. Sect. D Biol. Crystallogr.* **2001**, *57* (12), 1763–

1775.

(25) Sim, G. A. *Acta Crystallogr.* **1960**, *13* (6), 511–512.

(26) Abramowitz, M.; Stegun, I. a.; Miller, D. *Journal of Applied Mechanics*. 1965, p 239.

(27) Nanao, M. H.; Sheldrick, G. M.; Ravelli, R. B. G. *Acta Crystallogr. Sect. D Biol. Crystallogr.* **2005**, *61* (9), 1227–1237.

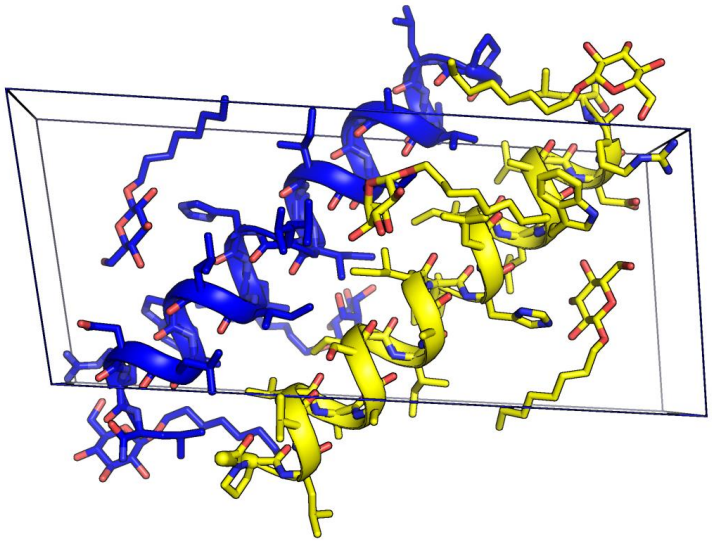
(28) Wilson, A. *Acta Crystallogr.* **1950**, *3* (5), 397–398.

(29) Phillips, D. C.; Rogers, D.; Wilson, A. J. C. *Acta Crystallogr.* **1950**, *3* (5), 398–399.

(30) Beckmann, P. *J. Res. Natl. Bereau Stand. Radio Propag.* **1962**, *66D* (3), 231–240.

(31) Herbst, D. A.; Jakob, R. P.; Zähringer, F.; Maier, T. *Nature* **2016**, *531* (7595), 533–537.

Chapter 5 A Hendecad Motif is Commensurate with Heterochiral Coiled Coil Formation: Evidence from Racemic Protein Crystallography



Portions of this chapter appear in:

Kreitler, D. F., Yao, J., Steinkruger, J. D., Mortenson, D. E., Forest, K. T., Gellman, S. H. A Hendecad Motif is Commensurate with Heterochiral Coiled Coil Formation: Evidence from Racemic Protein Crystallography. *Manuscript in preparation.*

5.1 Introduction

Polypeptides that contain D-amino acids have generated interest in part due to their ability to resist degradation by natural processes *in vivo*, a major shortcoming of their L-polypeptide counterparts.[1–3] Rational incorporation of D-amino acid residues at select positions in tertiary or quaternary L-peptide folds can also enhance folded stability or provide access to unique topologies that are unobtainable to purely L-polypeptides.[4–10] Concomitant with this strategy is the development of peptides comprised entirely of D-amino acids. These D-peptides often exhibit altered activity or immunogenicity relative to the corresponding L-peptide enantiomers.[8,11] Inversion of stereochemistry often diminishes or abolishes affinity to the native target binding partner or substrate due to the stereospecific nature of many biological interactions. In general, a selection or design process is required to generate useful versions of these D-peptides. Mirror image phage display has been demonstrated as a viable method for identifying D-polypeptides with a high affinity for a natural L-protein target.[2,12] Bacteriophages are iteratively evolved to interact with a synthetic D-protein target, and when the optimal sequence is identified the enantiomer of the phage derived L-peptide invariably binds to the natural L-protein target with the same affinity. Retro-inverso peptides, where the chirality is inverted from L to D-amino acids and the peptide sequence is reversed with respect to the amide backbone, have also emerged as a promising strategy for mimicking the activity of a natural polypeptide (Figure 5.1B,C).[13,14] For instance, retro-inverso peptides have been shown to bind antibodies raised against the parent L-peptides, a remarkable cross reactivity.[13] Despite these initial limited successes there still remains much to be understood about how D-polypeptides interact with their L-polypeptide counterparts.

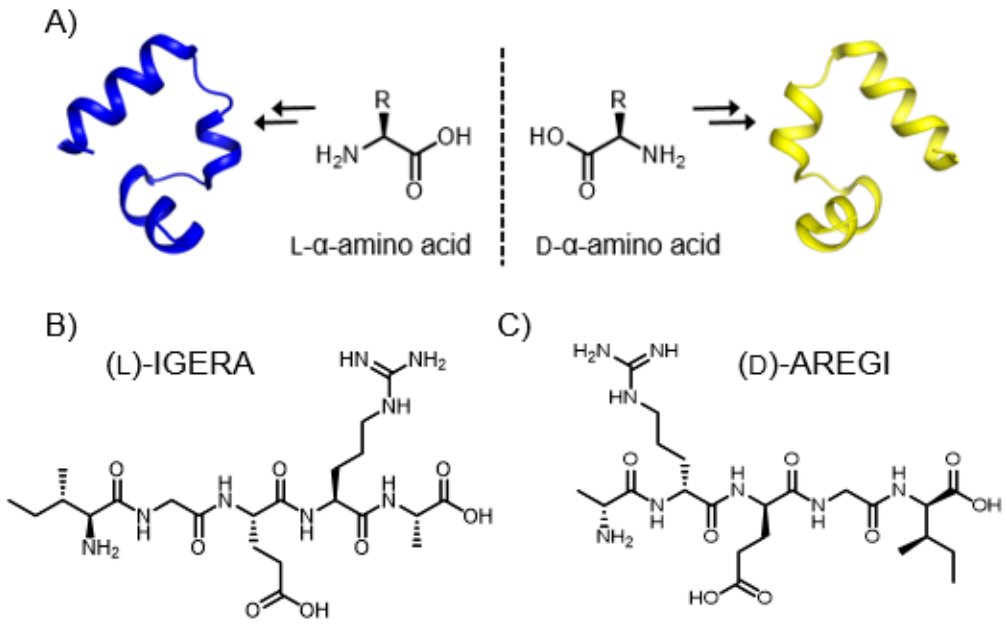


Figure 5.1. A) Proteins of opposite handedness can be synthesized from L or D-α-amino acids D-VHP (PDB: 3TRW; blue), L-VHP (PDB: 3TRY; yellow). The binding properties of a B) model peptide antigen with little secondary structure can be retained by C) an inversion of stereochemistry at the C_α position and a reversal of the direction of the amide backbone with respect to the native sequence (retro-inverso peptide).

Coiled coils, oligomeric assemblies of *n* peptide helices, are arguably the most well understood protein folding motif, and they play essential roles in membrane fusion, cytoskeletal movement, and DNA transcription.[15–17] In canonical coiled coils, apolar residues spaced three and four residues apart are displayed on the same side of an α-helix, and the helices twist around a central superhelical axis to ensure maximum burial of hydrophobic surface area. Regular spacing of hydrophobic core residues every three to four residues results in a seven-residue or 'heptad' motif that gives rise to a predictable superhelical structure (Figure 5.2).[16,18] The relative orientation of strands, *i.e.* parallel vs. antiparallel, can also be controlled by strategic placement of polar or charged residues at positions flanking hydrophobic core positions (or in some cases, at hydrophobic core

residues) so that favorable Coulombic or hydrogen bonding interactions can only occur in certain orientations, which adds an extra layer of control beyond simple burial of hydrophobic surface area.[19,20] These principles are well-known and have been reviewed extensively in the literature since Crick described a set of parametric equations that accurately describes coiled coil geometry in 1953.[15–18,21] Recent computational efforts based on various iterations of the Crick equations demonstrate that *de novo* design of thermodynamically stable, homochiral coiled coil assemblies is possible.[9,22–24] Given these considerations, a coiled coil system should be appropriate for initial efforts in elucidating the principles that govern heterochiral interactions between L and D-peptides. In 1953, Crick noted that helices of opposite handedness should 'mesh together' in a manner akin to two gears of opposite sense, *i.e.* the helices should remain parallel to each other (parallel refers only to the alignments of helical axes, not the orientation of the amide backbone).[21] Despite this early prediction and more recent efforts towards obtaining biologically active D-peptides, fundamental design principles for creating heterochiral complexes are lacking relative to what is known for homochiral coiled coils.

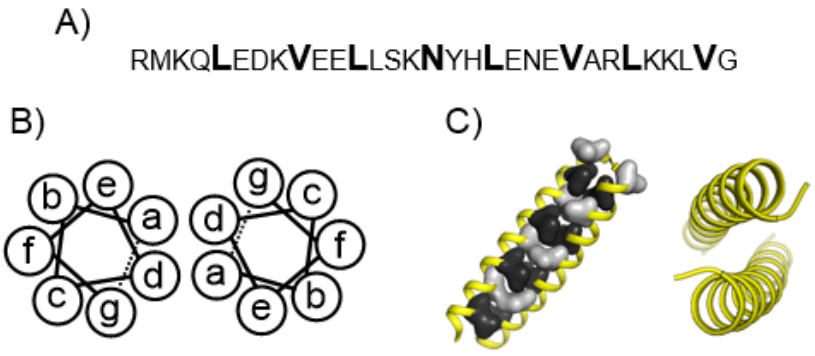


Figure 5.2. A) The primary sequence of GCN4, an antiparallel homodimeric yeast transcription factor, features hydrophobic residues at a and d positions. B) Helical wheel diagrams illustrate interfacial residues in canonical coiled coils. C) The atomic coordinate model of GCN4 displays left handed supercoiling around the interhelical axis, a common feature of canonical coiled coils.

An idealized alpha-helix has 3.63 residues/turn.[25] Therefore residue spacing of three or four residues will position the sidechains of *i*+3 and *i*+4 residues on approximately the same side of an α-helix as the *i*-th residue. In canonical coiled coils, hydrophobic residues are spaced three and four residues apart to form a seven-residue, heptad registry, that spans two turns of the α-helix. The heptad positions are denoted *abcdefg*, where *a* and *d* are conventionally the hydrophobic core residues (Figure 5.2). This 3,4-spacing results in a periodicity of 3.5 residues/turn, which is lower than the 3.63 residues/turn pitch of a linear, idealized α-helix. This discrepancy (3.5 vs. 3.63 residues/turn) imparts a chirality to the coiled-coil interface. In a right-handed L-helix there is a left-handed drift of the hydrophobic interface. The corollary is then that a left-handed D-helix will result in a right-handed drift of the hydrophobic interface (Figure 5.3).

The helices within natural, homochiral coiled-coils compensate for this drift by supercoiling around the interhelical axis. A 3,4-spacing pattern is less likely to be compatible with a heterochiral coiled coil beyond approximately four turns of the alpha-helix due to the opposite ‘handedness’ of each hydrophobic interface. Despite this fact there has been some limited success in designing heterochiral coiled coils based on a 3,4-spacing principle.[26,27] Introduction of a four residue ‘stutter’ continues the 3,4-spacing motif across another turn and encompasses 11 residues that results in a periodicity of 3.66 residues/turn. The discrepancy between the 3.66 periodicity of a 3,4,4-spacing and the 3.63 residue/turn α-helix is marginal, so that the hydrophobic interface displayed by a right-handed α-helix with 3,4,4-spacing is effectively a straight stripe along the side of the helix*, i.e.* the marginal difference between the periodicities results in a marginal handedness of the hydrophobic surface area display (Figure 5.3). The conventional

abcdefg heptad can then be expanded to *abcdefg**x**efg* so that the residue that would be in position *a* is shifted towards the core of the interface, the side-chains from residues at this position constitute what is known as an '*x* layer'. Residues at *a* and *d* positions are projected towards either side of the interface and form a '*da* layer' (Figure 5.3).[17,28,29]

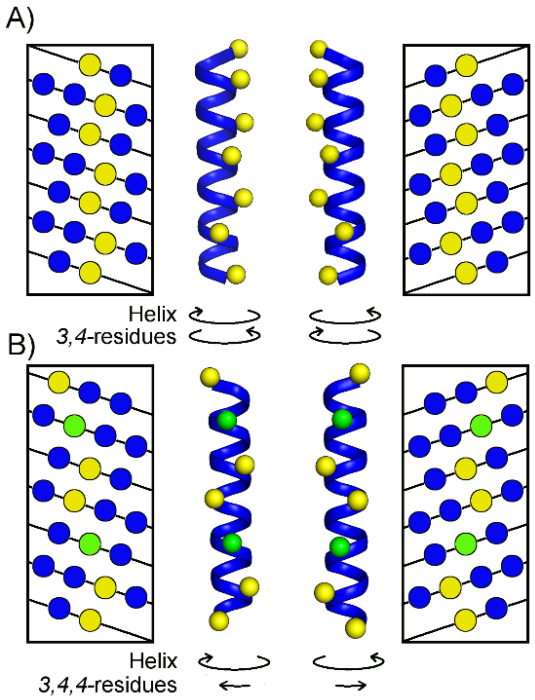


Figure 5.3. Helical net diagrams map the 3D surface of a helix to a 2D display with periodic boundary conditions based on three- and four-residue repeats (approximately one turn of the helix). A) Knob positions are indicated in yellow for a 3,4-spacing pattern for a left-handed (left panel) and right-handed (right panel) α-helix. B) Insertion of a four residue stutter results in the formation of an x layer where the x position (green) is flanked by two four residue spacings for a left-handed (left panel) and right-handed (right panel) helix. In both A) and B) the helical chirality is compared to the chirality of the hydrophobic interface.

Recently, Mortenson *et al.* reported structures of a heterochiral coiled coil obtained via racemic crystallization of the Matrix-2 proton channel transmembrane domain (M2-TM).[30] Two distinct high-resolution structures were obtained, one from a lipid cubic phase (LCP; PDB: 4RWB), and one from a racemic solution of L/D-β-octyl-glucoside (OG;

PDB: 4RWC). The LCP-derived structure displayed a 3,4,4-spacing pattern, and the OG-derived structure featured a geometry similar to that of the LCP structure; however, the helices were offset and exhibited a 3,4-spacing over four turns. From this pair of structures it was unclear whether the observed hendecad motif in the LCP-derived structure was an 'artifact of crystallization' or an inherent feature of a heterochiral coiled coil interface.

The identity of the hydrophobic core residues can play a role in determining the oligomerization state of homochiral coiled coils; different oligomeric states of homochiral coiled coils have different steric requirements for hydrophobic core positions.[15] Residues with bulky side-chains, such as isoleucine, can bias assemblies towards specific oligomerization states.[15,31] We reasoned that attenuation of side-chain steric bulk at core positions in the M2-TM sequence, via Ile→Ala substitution, would alter the observed heterochiral packing arrangement (Figure 5.4). Therefore, we sought to obtain high resolution X-ray crystal structures of these altered assemblies to elucidate structural principles that would account for any altered packing arrangement. These structural principles should also support the *de novo* design of heterochiral coiled coils systems.

Racemic crystallization of the Ile→Ala substituted M2-TM variants enabled high-resolution structure determination of three different substituted variants. Each structure contained a heterochiral dimer with a hendecad packing motif, similar to the previously reported M2-TM LCP structure (PDB: 4RWB). Substitutions that decreased steric bulk, *i.e.* Ile → Ala, at *x* layers (knobs) or at *g* positions (holes), favored formation of the hendecad interface. Finally, we derived parametric equations that describe a heterochiral coiled coil assembly were derived and used to determine the helical frequency

(residues/turn) for each substitution. In each case, the frequency was depressed relative to the native homochiral M2-TM structure and consistent with a hendecad (3+4+4) repeat motif.

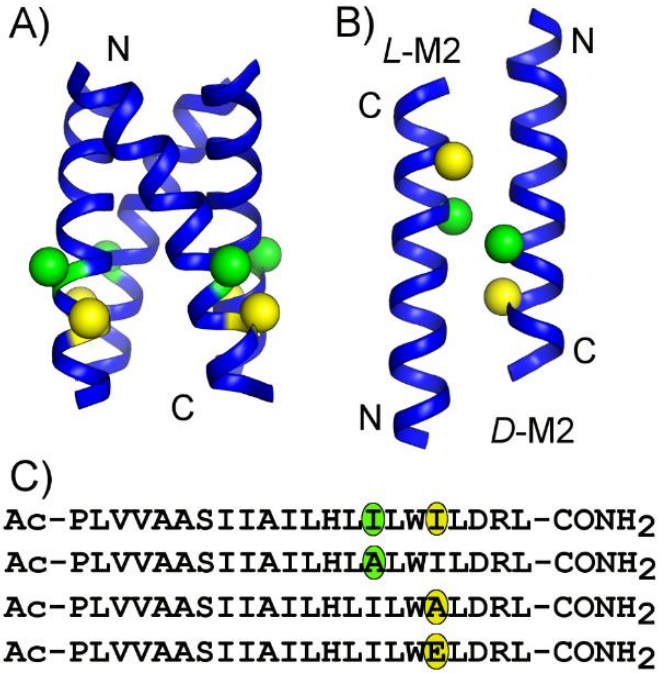


Figure 5.4. A) Substitutions were made at residues 39 (green spheres) and 42 (yellow spheres) which are highlighted in the heterochiral, dimeric M2-TM OG structure. B) Four different pairs of racemic M2-TM derivatives were prepared in this study; substitutions were made at either x-layer (green ovals) or da-layer positions (yellow ovals).

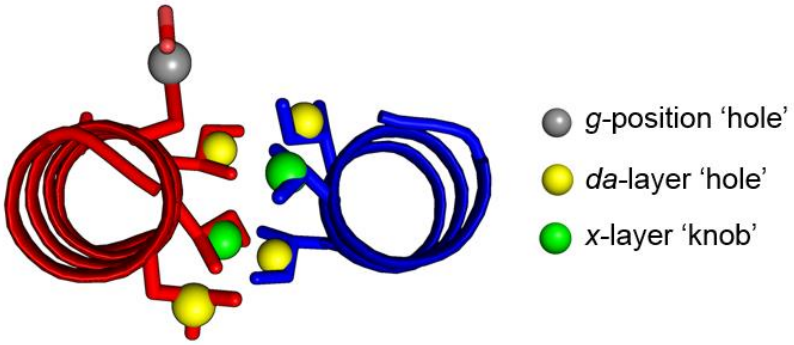


Figure 5.5. A hendecad motif between an antiparallel, heterochiral coiled coil dimer (L-helix, blue; D-helix, red); x-layer residues project their side-chains (green spheres) into the interhelical axis where they are flanked by da-layer residue side-chains (yellow spheres) and g position knobs (gray spheres).

In addition to providing useful insights on the design principles governing interactions between L and D-helices, the structures described in this chapter augment a small, but growing knowledge bank of transmembrane domain X-ray crystal structures. Integral membrane proteins (IMPs) account for ~30% of protein-encoding genes in humans and are an important class of targets for drug development, yet comparatively there are fewer structures of IMPs relative to soluble proteins.[32] This disparity in structural information is primarily due to the refractory nature of IMPs, particularly their low solubility, which complicates each step of the structure-determination process. A tenable workaround to this problem involves a reductionist approach to obtaining IMP structures in which discrete, soluble extracellular and intracellular domains are structurally characterized instead of the full length IMP. However, this approach necessarily excludes the transmembrane (TM) domain, which often plays an integral role in signal transduction across lipid bilayers in single-pass IMPs.[32] Several research groups have pursued an extension of this compartmentalization approach, focused on isolated TM domains, obtained via chemical synthesis or heterologous expression. These TM domain peptides can be crystallized from either a lipidic cubic phase or non-ionic detergents. These efforts have yielded ten high-resolution structures, excluding the three structures reported in this chapter (Table 5.1). These ten structures were derived from three different TM sequences: glycophorin A (parallel homodimer; antiparallel homodimer), M2-TM (parallel tetramer, homochiral; antiparallel dimer, heterochiral), and DAP12 (parallel homotrimer; parallel homotetramer). Although these three sequences represent a minute fraction of the available transmembrane proteome, some general properties that govern TM domain crystal lattice assemblies begin to emerge through inspection of the various crystal forms,

which are unique for each structure (Table 5.1; Appendix). The salient features of these assemblies are outlined to provide context for the packing assemblies observed for the three new structures reported in this chapter.

Table 5.1. Summary of reported TM domain X-ray crystal structures

PDB entry	Sequence	Space group	Type packing	Method	No. res.	Antiparallel interface	(Year) Ref.
5EH4	GpA (p)	$P3_112$	I	LCP	30	✓	(2015)[33]
5EH6	GpA (ap)	$R32$	I	LCP	30	✓	(2015)[33]
4WOL	DAP12	$P2_12_12_1$	I	LCP	32	✓	(2015)[34]
4WO1	DAP12	$P2_12_12$	I	LCP	32	✓	(2015)[34]
4QK7	M2	$I4$	I	LCP	25	X	(2015)[35]
4QKC	M2	$I4$	I	LCP	25	X	(2015)[35]
3LBW	M2	$C222_1$	I	OG	22	✓	(2010)[36]
5C02	M2 S31N	$I4$	I	LCP	25	X	(2016)[37]
4RWB	rac-M2	$P2_1/c$	I	LCP	22	✓	(2015)[30]
4RWC	rac-M2	$P\bar{1}$	I	rac-OG	22	✓	(2015)[30]
M2-I39A	rac-M2	$P\bar{1}$	I	rac-OG	22	✓	This work
M2-I42A	rac-M2	$P\bar{1}$	I	rac-OG	22	✓	This work
M2-I42E	rac-M2	$P\bar{1}$	II	rac-OG	22	✓	This work

GpA = Glycophorin A, LCP = lipidic cubic phase, OG = D-β-octyl-glucoside, rac = racemic

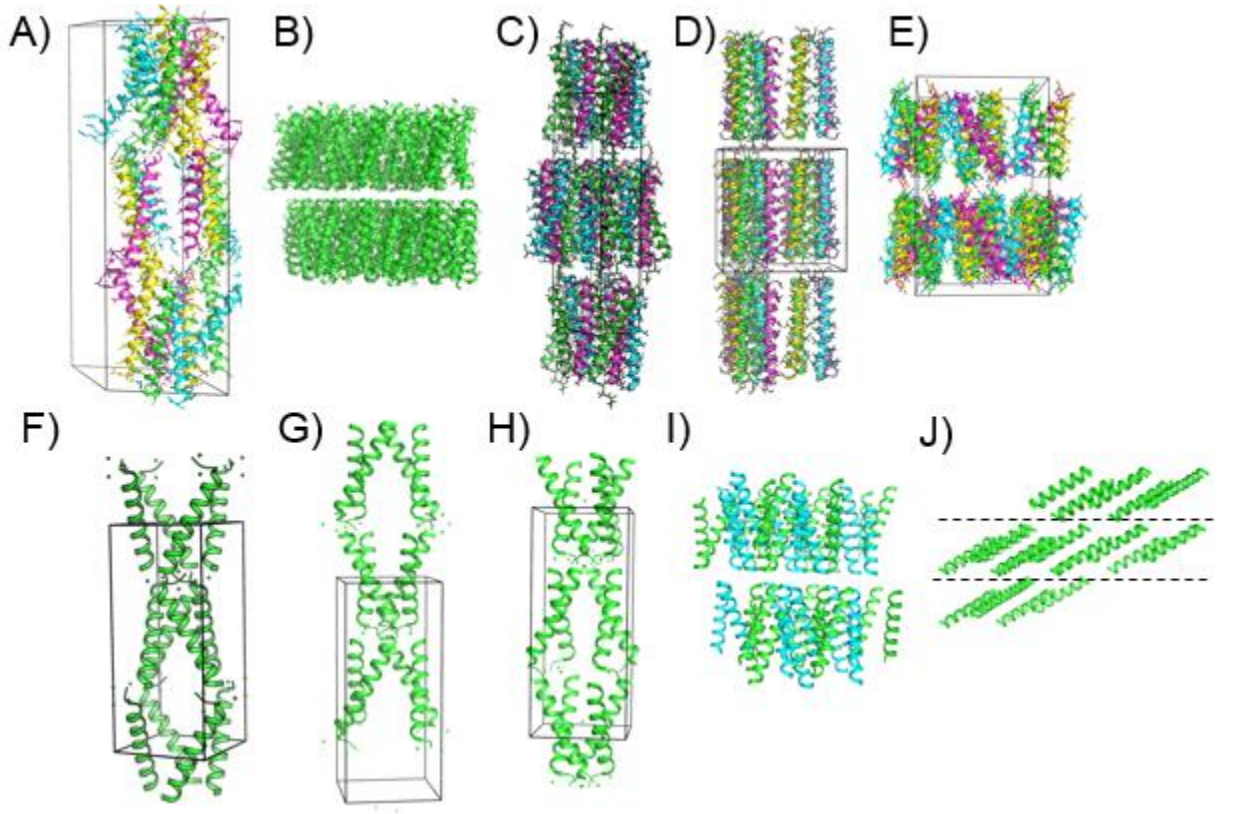


Figure 5.6. Each previously reported TM domain structure, both homochiral and racemic, feature type I crystal packing: PDB A) 5EH4, B) 5EH6, C) 4WOL, D) 4WO1, E) 3LBW, F) 4QK7, G) 4KQC, H) 5CO2, I) 4RWB (racemic), and J) 4RWC (racemic).

Each previously reported TM structure displays type I crystal packing, in which lamellar sheets of peptide helices interact through tightly packed lateral van der Waals (VDW) interactions. These sheets are stacked on top of one another and pinned together by some combination of Coulombic or hydrogen bonding interactions between terminal residue side chains and/or backbones between adjacent layers (Figure 5.6, Appendix). Only one structure, derived from glycophorin A (5EH6), lacks any discernible contact between layers. Another interesting feature is the preponderance of two-fold rotational symmetry among the structures (six out of ten homochiral structures). For dimeric structures, two-fold symmetry may provide a means to maximize the number of favorable

contacts between layers (Figure 5.7). In either case (parallel or antiparallel dimer) the presence of horizontal two-fold rotational symmetry in a type I lamellar packing assembly necessitates the presence of an antiparallel interface (Figure 5.7). In the homochiral M2-TM structures derived from the lipid cubic phase the lipid tethers layers together. The aliphatic lipid tail forms close contacts with the hydrophobic side chains within one layer, and the polar lipid head group extends to form hydrogen bonding contacts with polar or charged side chains in the adjacent layer (Figure 5.8). Given this observation, an intriguing hypothesis emerges: perhaps covalent tethering of alkyl groups to N or C-terminal side chains would enhance the formation of these types of layered assemblies.

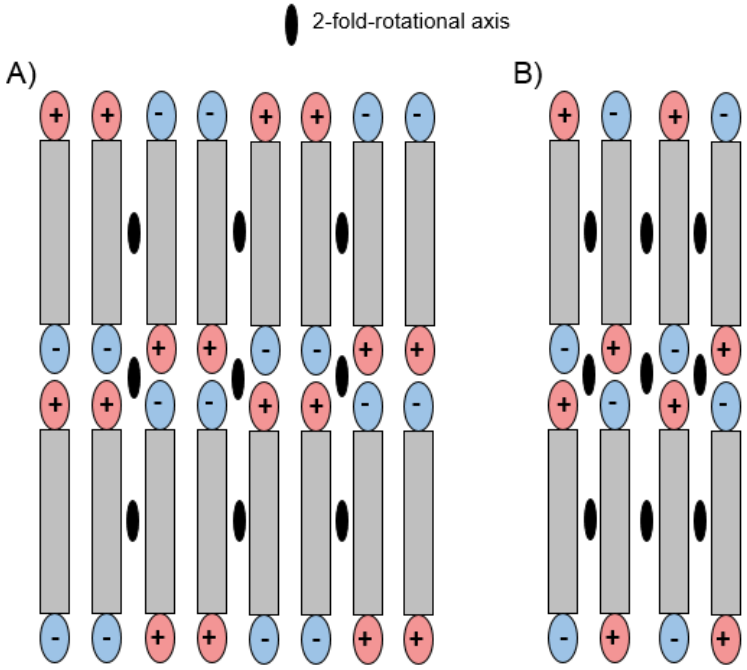


Figure 5.7. A) Simplified schematic of the packing scenario observed in the glycophorin A 5EH4 structure, where the presumed functionally relevant parallel homodimer was observed. B) An antiparallel homodimer was observed in the glycophorin A 5EH6 structure. In either case, the presence of two-fold rotational symmetry necessitates the presence of an antiparallel interface.

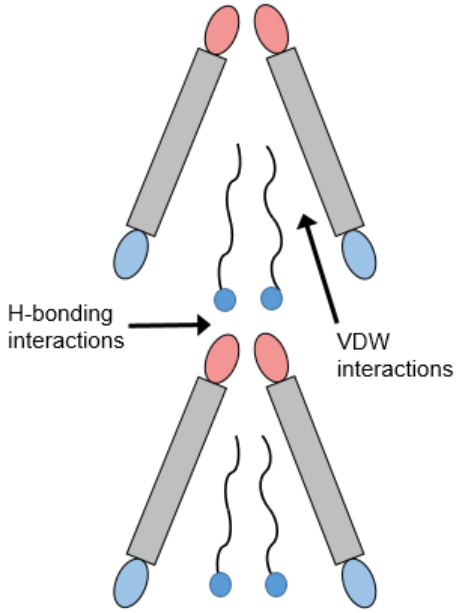


Figure 5.8. Interlayer packing was facilitated by monoolein in the lipid cubic phase derived, homochiral M2-TM structures (4KQ7, 4KQC, and 5C02). Polar lipid head groups extend beyond the peptide layer while the alkyl chain anchors the lipid into the layer through contacts with hydrophobic side chains.

Presently the lipidic cubic phase dominates as the employed crystallization medium among reported TM structures, and so the results may be skewed to over-represent type I packing, which is concomitant with the putative mechanism of protein crystallization from lipid mesophases.[38] As more TM domain structures are reported from detergents, mesophases, or other media, *e.g.*, bicelles or nanodiscs, it will be crucial to catalogue layer-spanning interactions. Ultimately, rational incorporation of small di or tri-peptide motifs at TM domain termini may be a useful strategy for crystallizing these recalcitrant sequences. Chemical synthesis affords the unique advantage of incorporating unnatural layer-spanning motifs that are unobtainable via heterologous expression. To further these cataloguing efforts, interactions are described here that were observed to cause a switch from anisotropic (disordered layering) to isotropic (more

ordered layering) diffraction in the heterochiral M2-TM variants reported here that displayed type I packing. Finally, the first example of a TM domain structure, M2-TM I42E, (crystallized by Jennifer Yao) that displays type II packing (3-dimensional packing), is also described.

5.2 Results and Discussion

5.2.1 Design of Substitution Positions

Canonical coiled coils are predicated on knob-into-hole (KIH) interactions where intercalated side-chain 'knobs' pack into a four-residue 'hole' comprised of offset side-chains on adjacent helices. The identity of residues at these core positions can dictate the oligomerization state of the coiled coil. For instance, in the parallel coiled coil GCN4 system, derivatives with Leu and Ile at *a* and *d* positions respectively form dimers, and derivatives with Ile at both *a* and *d* positions form trimers.[31] This selectivity among oligomerization states is rationalized by steric constraints imposed upon KIH interactions, which can be classified as perpendicular, parallel, or acute, depending on the angle formed between the C_α-C_β vector of the knob with the horizontal C_α-C_α vector of the hole (Figure 5.10). We reasoned that substitutions at residues 39 (Figure 5.4, green) and 42 (Figure 5.4, yellow) would be disruptive to the close packing interactions in the OG heterochiral coiled coil (PDB: 4WRC) yet have a minimal impact on the native tetramer (PDB: 3LBW) where residues 39 and 42 are positioned along the periphery of the proton channel (Figure 5.4). The native tetrameric M2-TM is not a canonical coiled coil; individual helices curve around an articulation point centered at Ile35, thus standard coiled coil packing rules do not apply.[36,39] Ile→Ala and Ile→Glu substitutions were made separately

at these positions 39 and 42 to try to perturb packing interactions that seem to be crucial for heterochiral coiled coil formation.

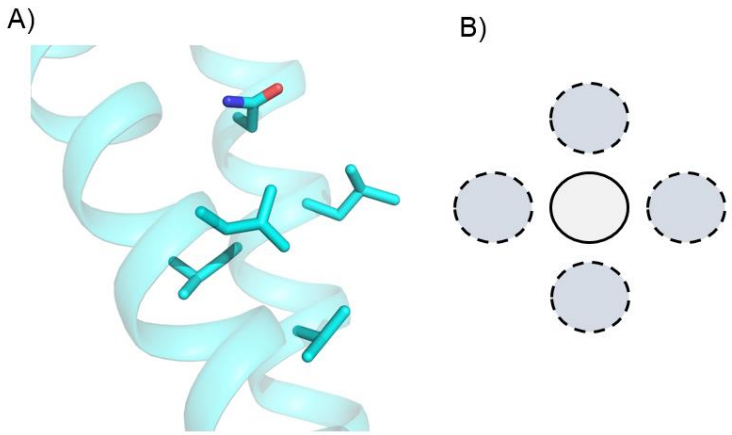


Figure 5.9. A) Canonical knob-into-hole (KIH) interaction in the homochiral GCN4 system. B) KIH interactions (knob, solid circle; hole, dotted circle) are often represented as 2D projections in helical net diagrams.

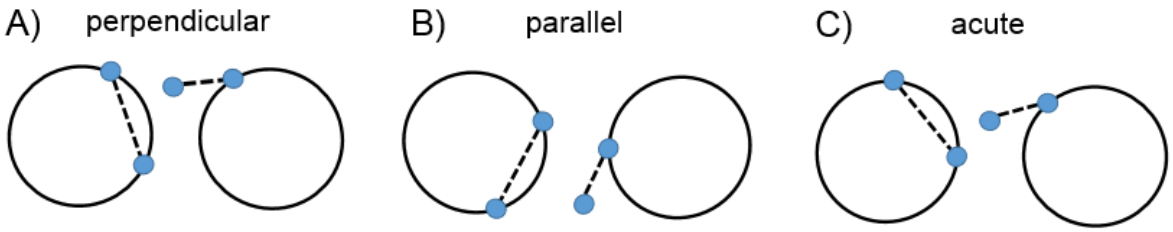


Figure 5.10. Diagram depicting the three possible knob-into-hole interaction orientations. C_α-C_β vectors can be either A) perpendicular B) parallel or C) acute to the horizontal C_α-C_α vector on an adjacent helix.

5.2.2 Racemic Crystallization

Racemic crystallization is a technique that is thought to promote growth of diffraction quality crystals of peptides and proteins that are recalcitrant to conventional crystallization techniques.[40] The presence of both enantiomers makes racemic crystallization well-suited for structural studies of heterochiral peptide assemblies.

Crystallization of racemic solutions of three variants of M2-TM I39A, I42A, and I42E, afforded atomic-resolution X-ray crystal structures of 1.55, 1.40, and 1.40 Å respectively. A 1:1 solution of each peptide enantiomer was crystallized as a racemic mixture from an aqueous solution of racemic-β-octyl-glucoside (2% w/v) using hanging drop vapor diffusion (for details regarding peptide synthesis, purification, and crystallization procedures see *Appendix*).

Each structure crystallized in space group $P\bar{1}$ and contains antiparallel heterochiral dimeric coiled coils, in which individual helices within a dimer are related by a crystallographic inversion center. M2-TM I39A and I42A both contain one L and one D-peptide in the unit cell, and both crystallized with type I crystal packing, which features a lamellar assembly of hydrophobic peptide layers and is commonly observed among membrane protein crystals that are grown from lipid mesophases.[41] In the M2-TM I39A and I42A structures, lamellae are comprised of heterochiral dimers that are aligned with the normal vector of the lamellae at a slightly acute angle (Figure 5.11); water molecules are occluded from the lamellae by intimate, lateral van der Waals interactions between peptides and occur mainly in the interstitial space that spans each layer. The M2-TM I42E structure contains two L-peptides in the asymmetric unit and thus two symmetry-independent heterochiral interfaces. The structure also displays type II crystal packing, which more closely resembles the packing observed in soluble protein crystals. Type II crystal packing is manifest as a 3D lattice pinned together by polar contacts between molecules and is more commonly observed among membrane protein crystals derived from a detergent solution relative to membrane protein crystals derived from LCP.[41] The type II packing observed in the M2-TM I42E racemate is the first such example

observed in a transmembrane domain crystal structure. In all three structures, detergent molecule head-groups interact with polar residue side chains of M2-TM in a stereospecific manner that is different for each structure (Figure 5.15).

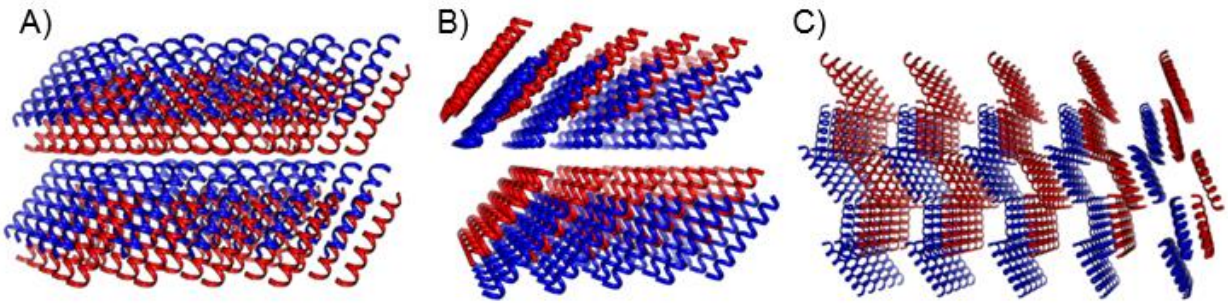


Figure 5.11. Type I (lamellar) crystal packing was observed in M2-TM A) I39A and B) I42A. C) Type II (3-dimensional) packing was observed in M2-TM I42E. Each panel displays two unit cells in the c-direction and five unit cells in the a and b-directions. All unit cells are oriented so that the ab-plane is horizontal and perpendicular with the plane of the page.

Despite their varied packing arrangements, all three structures share a heterochiral interface that is similar to the hendecad motif present in M2-TM LCP and distinct from the M2-TM OG interface. A common concern is whether or not an interaction observed in a crystal structure is reflective of what happens in solution or merely an 'artifact of crystallization'.[42] Although the latter possibility cannot be completely ruled out solely with X-ray crystal structures, the size (~1200 $Å^2$) and persistence of the heterochiral coiled coil interface among four different crystal forms (M2-TM LCP, I39A, I42A, and I42E) are both evidence against this interface being an 'artifact of crystallization'.

5.2.3 Interlayer Contacts that Impact Lamellar Ordering

Single crystal X-ray diffraction is dependent upon the regular positioning of atoms in three-dimensions. Any deviation from this translational symmetry, such as from thermal

motion, domain shifts, or crystal mosaicity will result in an attenuation of the Bragg reflection intensity. This attenuation can be orientation dependent, where disorder in a specific direction causes diffuse scattering in the region of reciprocal space that corresponds to the disordered real-space direction. In light of this phenomenon, disorder between lamella in TM domain crystals that display type I packing is manifest as an attenuation of reflection intensity for reflections corresponding to the reciprocal lattice vector that is perpendicular to the lamella. In the triclinic M2-TM I39A and I42A racemate structures, which display this layered packing, the plane spanned by the real-space ***a***, ***b*** lattice vectors corresponds to the layered electron density so that the ***c**** reciprocal lattice vector is perpendicular to the plane (***c**** is defined as the cross product of ***a*** and ***b***; Figure 5.12A). Therefore, as the magnitude of the dot product between ***c**** and a Bragg reflection increases, the impact of disorder will also increase, and there will be a decrease in signal intensity. So for the case of the M2-TM I39A and I42A racemates, a loss of signal in the ***c**** direction would be evidence for disorder between layers.

To test for this layered disordering, anisotropy plots were generated that display the signal-to-noise versus resolution (Å) for each reciprocal lattice basis vector. The signal-to-noise is defined as the structure factor amplitude, *F*, divided by the measurement error, sigma*F* (Figure 5.12). Anisotropic diffraction was observed for the M2-TM I39A racemate, whereas the M2-TM I42A racemate displayed relatively isotropic diffraction. These results indicate that the M2-TM I42A lattice displays ‘more ordered’ layers relative to the M2-TM I39A lattice.

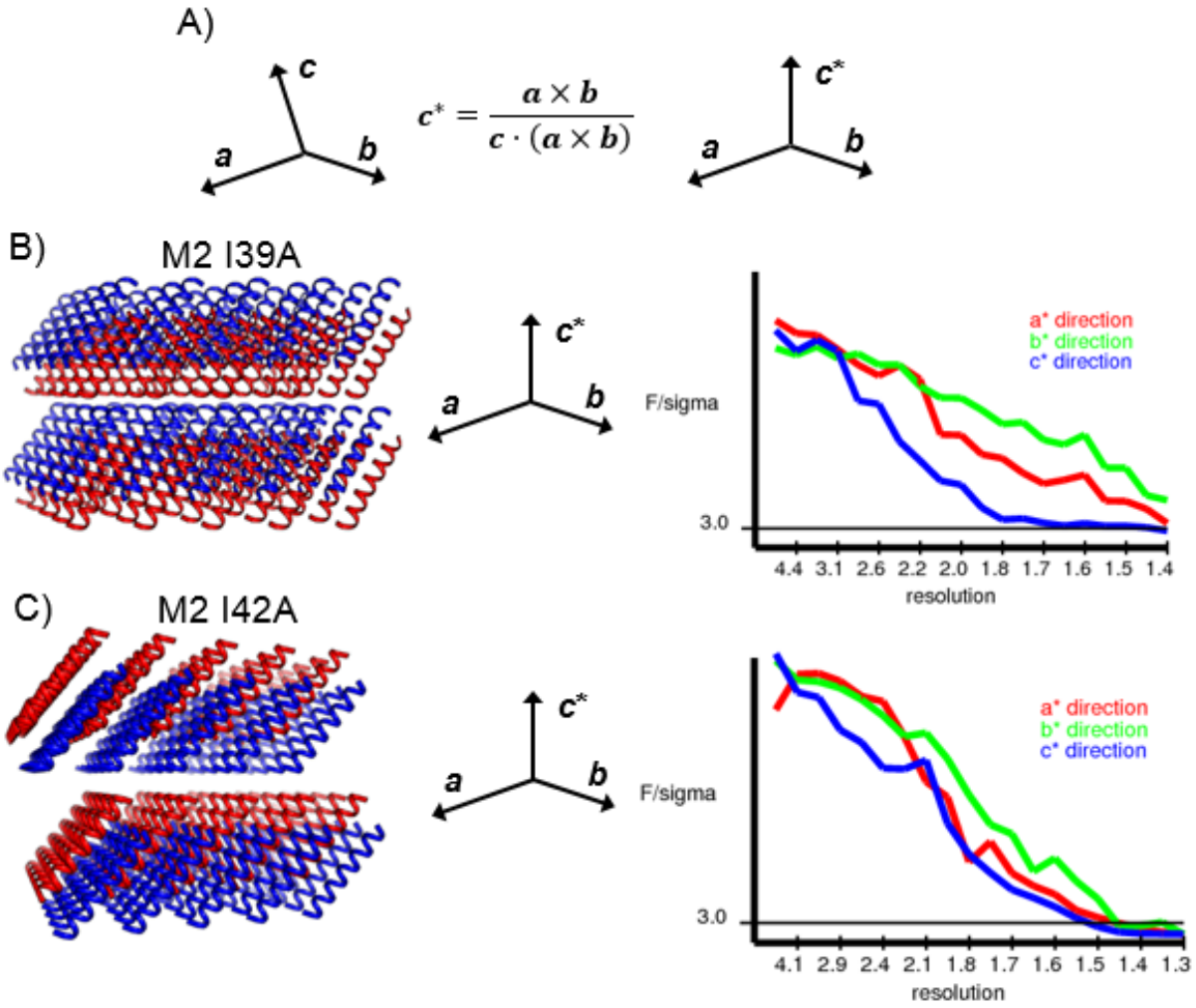


Figure 5.12. A) The reciprocal lattice vector **c*** is defined as the cross product of the **a** and **b** real space lattice vectors, therefore **c*** is perpendicular to the plan spanned by **a**, **b**. In the racemic B) M2 I39A and C) M2 I42A structures type I packing (lamellar) was observed where the layers correspond to **ab** planes. In the M2 I39A case, disorder between layers is manifest as anisotropic diffraction where structure factor amplitudes are attenuated the more they align with **c***. Anisotropy plots, which display the measured amplitude signal-to-noise (F/sigmaF) versus resolution (Å), were generated with the UCLA MBI diffraction anisotropy server (http://services.mbi.ucla.edu/anisoscale/).

The primary contact between layers in the M2-TM I39A and I42A racemates occurs at a crystallographic inversion center between the C-termini of L and D-peptide helices (Figure 5.13). These helices, anchored by this C-terminal-to-C-terminal contact, then interact laterally in their respective layers via hydrophobic interactions. The inter-layer

interaction is mediated primarily by the three C-terminal residues, Arg-Asp-Leu-$CONH_2$. In each crystal, the position of the Arg45 side-chain, especially, appears to play an integral role in the lattice-pinning interaction and thus the overall lattice order. Arginines frequently appear at positions flanking TM domains; presumably the energetic cost associated with guanidinium insertion into the lipid bilayer provides an anchor for the TM domain.[43] In the M2-I39A racemate structure, an *i*+4 cation-π interaction sequesters the arginine side chain away from the layer interface, which results in a large offset between the closest interlayer pair of helices (Figure 5.13). This offset arises in part to accommodate an interhelical salt-bridging interaction between the D-Arg45 guanidinium and the L-Asp44 carboxylate and backbone carbonyl (Figure 5.14A). An identical interaction is present between the L-Arg45 guanidinium and the D-Asp44 carboxylate and backbone carbonyl due to the presence of inversion symmetry. In the M2-TM I42A structure, which lacks the cation-π interaction, the Arg45 side chain is 'free' to extend towards the layer interface and form hydrogen bonding interactions with the Asp44 and Arg45 backbone carbonyls of the nearest peptide in the adjacent layer (Figure 5.14B). Based on the diffraction anisotropy data, this L-Arg45 side chain, D-Asp44/Arg45 backbone (or D-Arg45 side chain, L-Asp44/Arg45 backbone) interaction is evidently more favorable for layer ordering than the L-Arg45/D-Asp44 (or D-Arg45/L-Asp44) side chain interaction that occurs in the M2-TM I39A racemate structure. The identification and exploitation of these types of motifs will likely be essential for engineering modifications to TM domain sequences that enhance the likelihood of type I crystal growth.

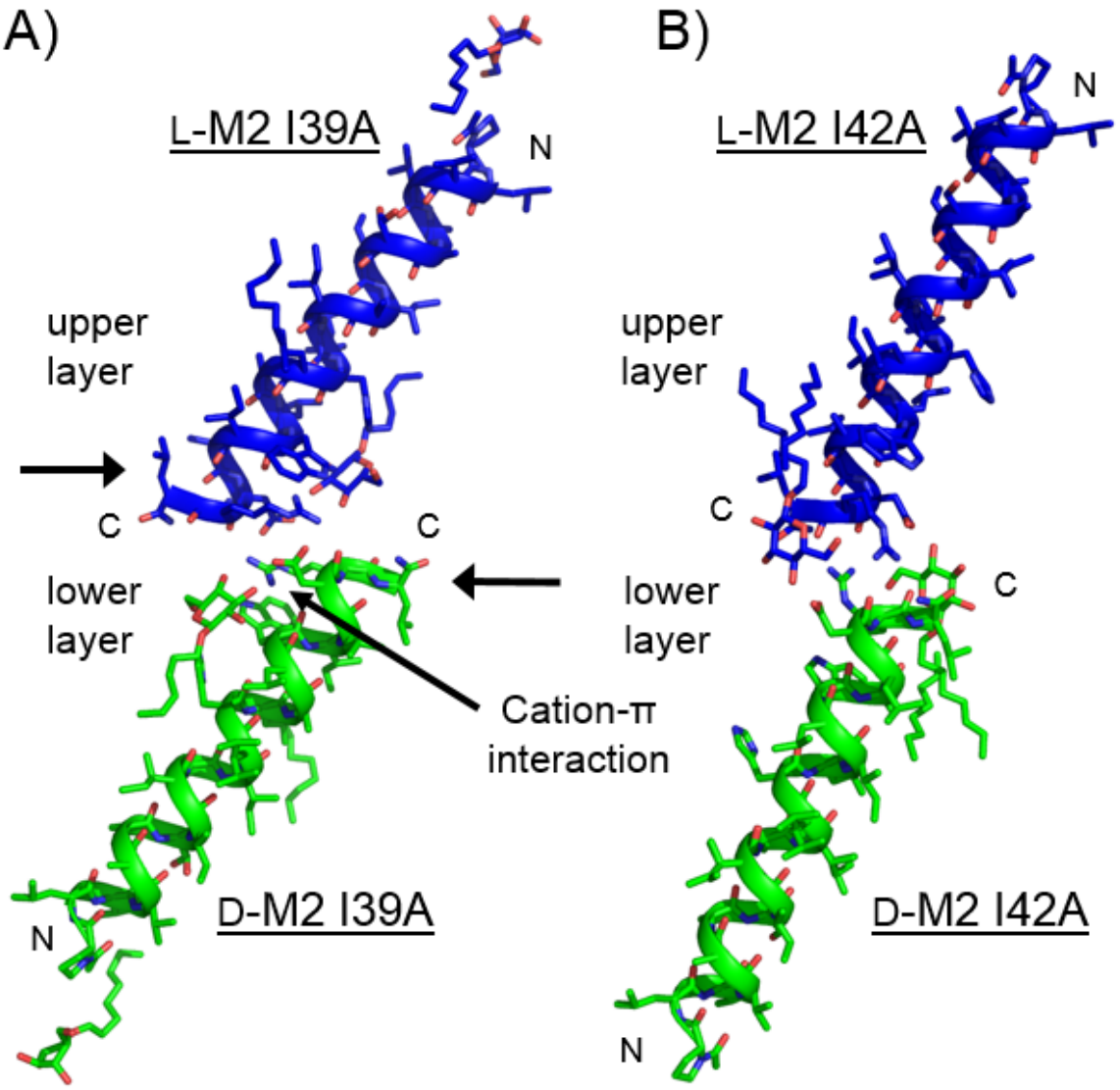


Figure 5.13. C-terminal end-to-end stacking mediates contacts between layers in the A) M2-TM I39A and B) M2-TM I42A racemates. In M2-TM I39A interlayer contacts are mediated primarily through side chain/side chain interactions salt-bridging interactions whereas in M2-TM I42A contacts are mediated mainly through side chain/backbone hydrogen bonds. The horizontal arrows in A) represent the direction in which the helices shift between M2-TM I39A and I42A structures.

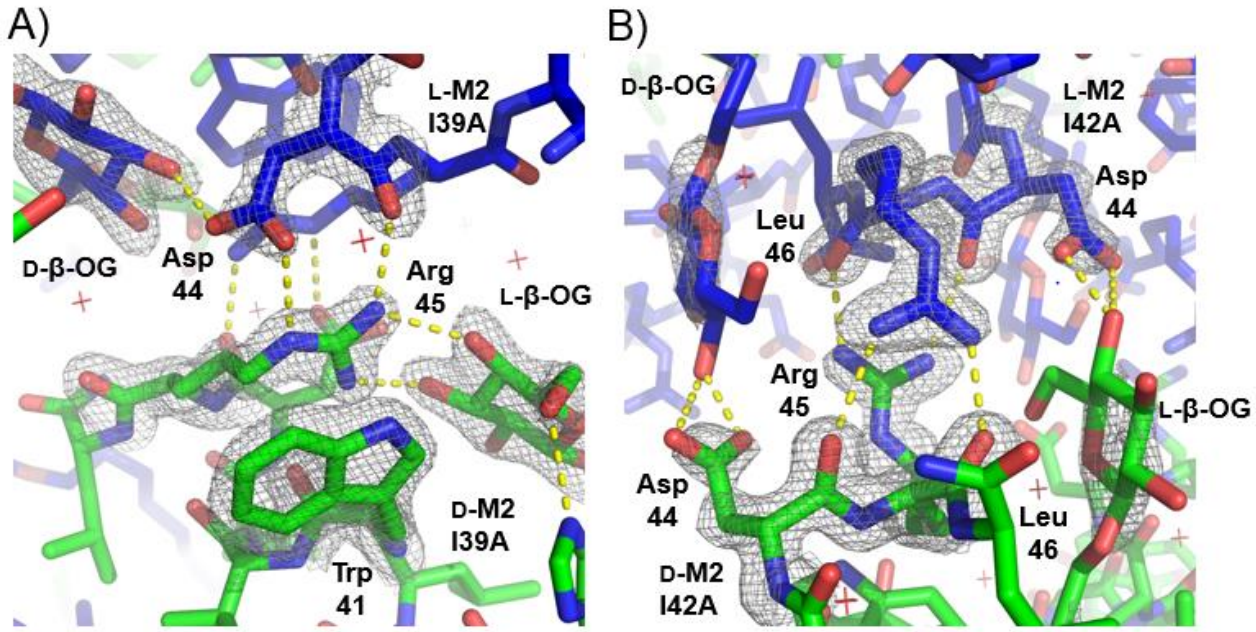


Figure 5.14. Interlayer salt-bridging/hydrogen bonding networks for the M2-TM racemate structures that displayed type I crystal packing (L-peptides, blue; D-peptides, green). A) The M2-TM I39A racemate features an interlayer salt-bridging interaction between Arg45 and Asp44 in addition to a hydrogen bonding interaction between the Asp44 backbone carbonyl and Arg45 side chain. A cation-π interaction sequesters the Arg45 side chain, which also forms hydrogen bonds with an L-β-octyl glucoside molecule. B) The M2-TM I42A racemate features interlayer contacts between the Arg45 side chain and Asp44/Leu46 backbone carbonyls from the adjacent layer. In both cases these interlayer contacts are centered on a crystallographic inversion center. Electron density maps are contoured at 1.5σ and were generated with $2mF_o$-DF_c weighted map coefficients. Dashed yellow lines indicate distances between 2.9-3.1 Å.

5.2.4 Knob-into-Hole (KIH) Interactions at the Heterochiral Interface

KIH interactions between antiparallel heterochiral dimers were measured and catalogued for each structure by quantifying the distance between the side-chain centers-of-mass (COM) of putative pairs of knob and hole residues (Figure 5.15). Knobs were identified at residues Val28, Ile32, Ile35, Xxx39, and Leu43, where Xxx is either Ile or Ala. This set of residues corresponds to a spacing of 4,3,4,4, which is consistent with a four-residue stutter insertion or a single instance of a hendecad motif (3+4+4).

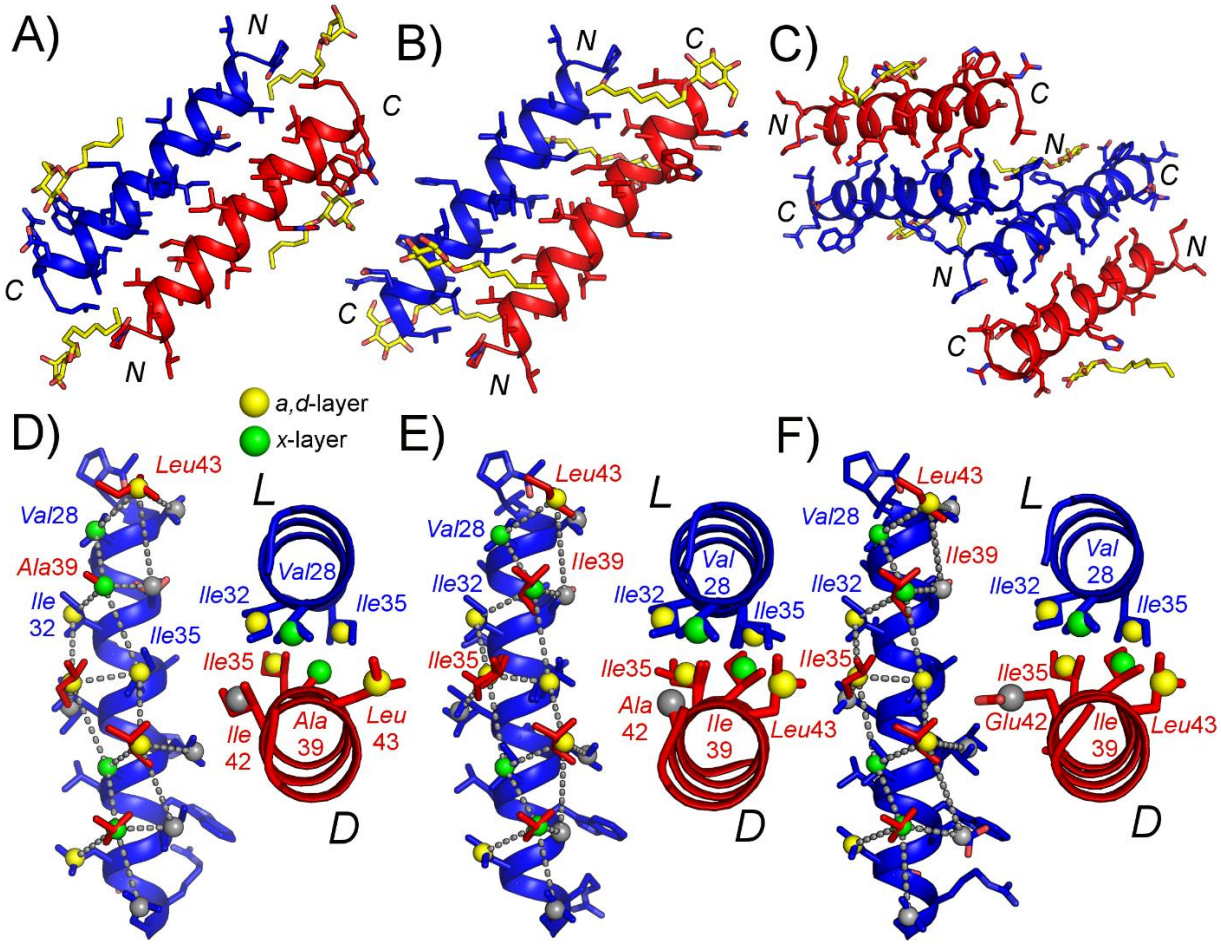


Figure 5.15. Heterochiral interfaces, including β-D/L-octyl-glucoside detergent molecules (yellow sticks) and feature one L-helix (blue) and one D-helix (red) are shown for M2-TM variants A) I39A, B) I42A, and C) I42E. Knob-into-hole (KIH) interactions are indicated by the gray dotted line for M2-TM variants: D) I39A, E) I42A, and F) one I42E interface. x layer knobs are shown as green spheres, da layer knobs as yellow spheres, and the modified g position hole is shown as a gray sphere.

Insertion of four residues into a canonical two-turn heptad, with a periodicity of 102.9°/residue (720°/7 residues), results in a three-turn hendecad with a periodicity of 98.2°/residue (1080°/11 residues). This decrease in the angular frequency of residue display upon switching from a heptad (102.9°/residue, 3.50 residues/turn) to a hendecad (98.2°/residue, 3.67 residues/turn) results in a local unwinding of the motif so that the side-chain of the residue in the middle of the 4,4 spacing, *i.e.* the *x* position, is projected

into the interface (Figure 5.15, green spheres). Based on the 4,3,4,4 spacing assigned to knob residues, Val28 and Xxx42 (Xxx: Ala or Ile) comprise the *x* layer, and the remaining knobs, Ile32, Ile35, and Leu43, comprise the *da* layer. This categorization of knob residues into either *x* or *da* layers lends insight to why a specific heterochiral interface may be preferred. Residues with smaller sidechains such as Ala, Val, or Gly are likely more easily accommodated at *x* positions compared to residues with bulkier side-chains such as Ile. In line with this steric complementarity hypothesis, we observe that substitution of *x* position Ile42 with Ala results in a shift from the offset 3,4-spacing in M2-TM OG to the 3,4,4-spacing observed in the hendecad motif, where Ala is aligned as an *x* position. As an extension of this complementarity hypothesis, we propose that a modification to the interface that increases the size of a particular hole should have a similar effect as decreasing the size of the corresponding knob. Indeed, the M2-TM I42A and I42E structures feature the hendecad motif which has shifted from the 3,4-M2-TM OG interface so that the substituted residue is at the *g* position adjacent to the Ile *x* layer knob (Figure 5.16).

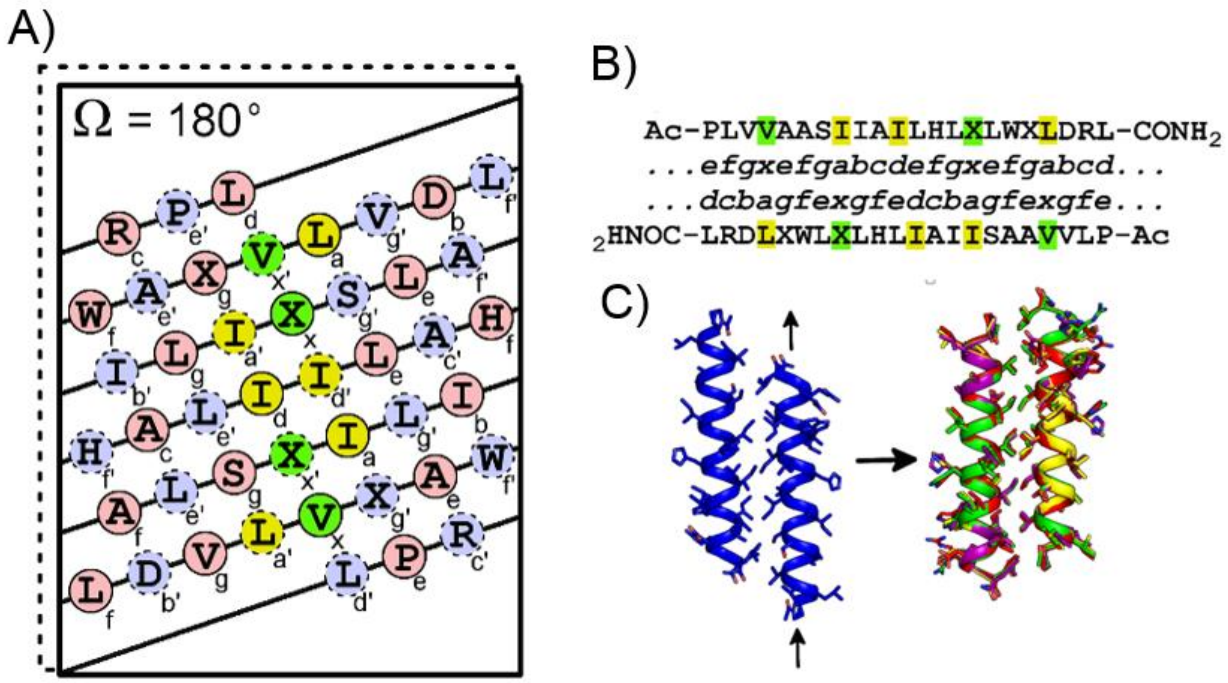


Figure 5.16. A) A helical net diagram for the antiparallel hendecad interface (180° crossing angle) shows that substituted residues become aligned at x (knob) and g positions (hole). Substitution positions are demarcated by 'X'. B) Finally, a sequence alignment of M2-TM overlaid with a hendecad registry corresponds to the net diagram. C) Finally, each substitution resulted in a shift from the 3,4-spacing observed in M2-TM OG (blue) to the hendecad motif (I39A, red; I42A, green; I42E1, yellow; I42E2, purple).

5.2.5 Knob-to-Knob (KTK) Interactions

The stated principles that govern dimeric heterochiral coiled coil assembly thus far include: a straight display of hydrophobic surface area along the side of an α-helix, namely in the form of a hendecad repeat motif, and a sterically compatible KIH interaction between *x* layer knobs and *ga* holes. However, neither of these principles necessarily discriminates between a heterochiral over a homochiral association or *vice versa*. Therefore a more detailed description of the geometry underlying each interface is required.

In homochiral coiled coils, non-canonical hendecad motifs result in the projection of an *x* layer side-chain towards the central axis of the coiled coil interface. This projection

results in a knob-to-knob (KTK) interaction between *x* position side-chains that disrupts canonical KIH packing. Consideration of the angles formed between the C_α-C_β vectors on opposing *x* position residues may in part explain some of the observed preferences for homochiral versus heterochiral and parallel versus antiparallel pairings of helices.

In the antiparallel case, the *z*-components of the C_α-C_β vectors on opposing *x* residues have opposite signs due to the orientation of the amide backbone (Figure 5.17A-C, left panel). However, the *y*-components of *x* residue C_α-C_β vectors have the same sign for the homochiral case (Figure 5.17A-B, right panel), but opposite sign for the heterochiral case (Figure 5.17C, right panel). Therefore, the angles between opposing C_α-C_β vectors are ~120° for homochiral dimers and ~180° for heterochiral dimers; the latter values were observed in each of the M2-TM derived heterochiral dimers. From a purely steric perspective, the heterochiral dimer seems more favorable than the homochiral dimer; however the fact that side-chains are oriented in completely opposite directions precludes any other type of interaction between side-chains, such as a salt-bridging, a cation-pi pairing, or a hydrogen bonding type interaction.

In the parallel case, the *z*-components of the *x* residue C_α-C_β vectors have the same sign for heterochiral versus homochiral (Figure 5.17D, left panel), and although no structure currently exists for the parallel heterochiral case, I predict that the orientation of the C_α-C_β vector *z*-component would be unaffected by inversion of stereochemistry. However, for a parallel dimer, the *y*-components of the C_α-C_β vectors are of opposite sign for the homochiral case and of the same sign for the hypothetical heterochiral case (Figure 5.16D). Thus, from a purely steric perspective, there is maximum repulsion for the heterochiral case but less steric overlap for the homochiral case. The parallel, homochiral

x residues have less steric repulsion than those of the antiparallel homochiral case, but there is also less room for favorable non-covalent interactions. Based on this simplistic C_α-C_β vector analysis, an antiparallel heterochiral coiled coil pair is the only pairing combination that completely mitigates sterically unfavorable KTK interactions in the presence of 3,4,4-spacing (Table 5.2).

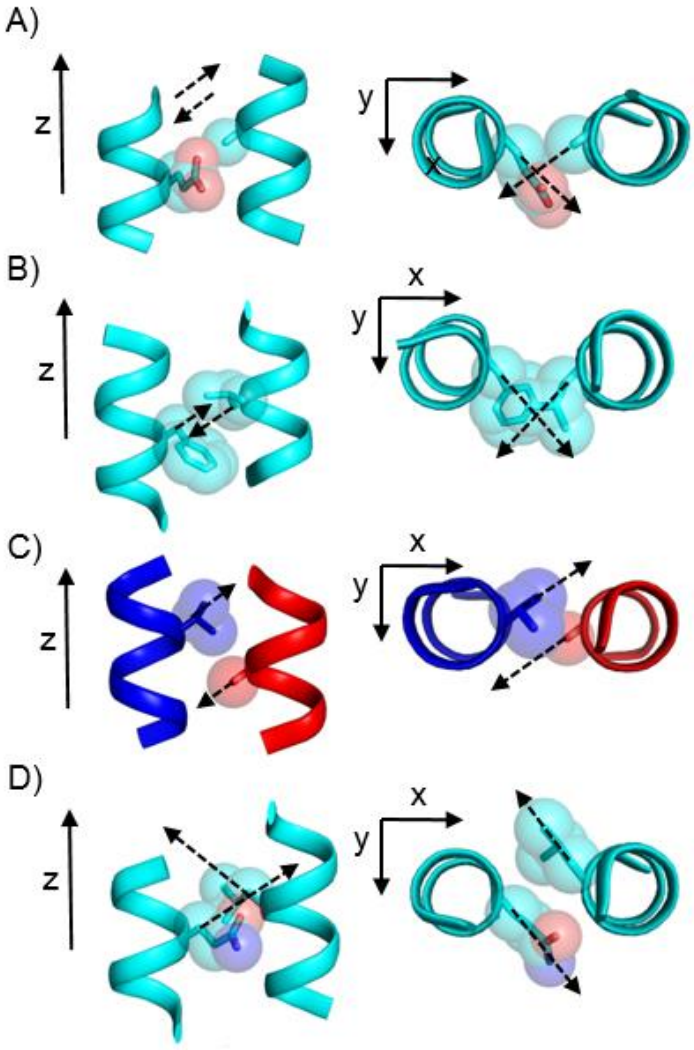


Figure 5.17. Longitudinal and axial views of dimeric knob-to-knob (KTK) interactions at hendecad x-positions are shown for the antiparallel, homochiral case for A) PDB: 4LTB Glu267/Ala234 and B) PDB: 3VEM Phe1776/Leu1765. C) The antiparallel heterochiral case is shown for M2-TM I39A (L-peptide, blue; D-peptide, red) between Val28 and Ala39 knobs. D) Finally, the parallel homochiral KTK interaction is shown for PDB: 4HH3 Gln134/Leu130. In each case, the Cartesian coordinate directions referenced in the text are shown in addition to the C_α-C_β vectors (dashed arrows).

Table 5.2. Predicted pairing preferences based on KTK interaction steric clash

Orientation	Pairing	*z*-component	*y*-component	Figure 5.17
antiparallel	homochiral	✓	X	A, B
	heterochiral	✓	✓	C
parallel	homochiral	X	✓	D
	heterochiral	X	X	-

✓ no predicted clash
X predicted clash

5.2.6 Parameterization of a Heterochiral Helical Assembly

Quantitative comparisons of structural features among the series of heterochiral M2-TM variant coordinate models require an abstraction of heterochiral coiled coil assemblies based on objectifiable model parameters. Pertinent model parameters include some value to quantify the 'tightness of packing' among interfaces to test our steric complementarity hypothesis, in addition to a helical frequency parameter to test the hendecad hypothesis. In general, the parameterization of folded protein structures and assemblies is useful for *de novo* design purposes. Mapping these complex structures to an *N*-dimensional model, where *N* is relatively small compared to the number of atoms in the structure, allows for grid searching of the available structural space. Several successful coiled coil design strategies have employed the Crick coiled coil parametric equations to generate backbone conformations that can sample available structural space at a reduced computational cost.[22–24,44] The Crick equations for a canonical coiled coil have the general form:

$$x = R_0 \cos\omega_0 - R_1 \cos\omega_0 \sin\varphi + R_1 \cos\alpha \sin\omega_0 \sin\varphi \quad (5.1)$$

$$y = R_0 \sin\omega_0 - R_1 \sin\omega_0 \cos\varphi - R_1 \cos\alpha \cos\omega_0 \sin\varphi \quad (5.2)$$

$$z = P(\omega/2\pi) + R_1 \sin\alpha \sin\varphi \quad (5.3)$$

$$x_i = R_0 \cos(2\pi z/P + \tau_i) + x\cos(2\pi z/P + \tau_i) - y\cos\alpha \sin(2\pi z/P + \tau_i) \quad (5.4)$$

$$y_i = R_0 \sin(2\pi z/P + \tau_i) + x\sin(2\pi z/P + \tau_i) - y\cos\alpha \cos(2\pi z/P + \tau_i) \quad (5.5)$$

$$z_i = z - y\sin\alpha \quad (5.6)$$

$$\tau_i = 2\pi(i-1)/n \quad (5.7)$$

where *P* is the pitch of the superhelix (Figure 5.18), *x,y,z* define points along the superhelix and x_i,y_i,z_i define points along the *n*-th helix. The remaining parameters are described in Figure 5.19. Ultimately these equations can be adapted with several key assumptions to describe an antiparallel, heterochiral dimer.

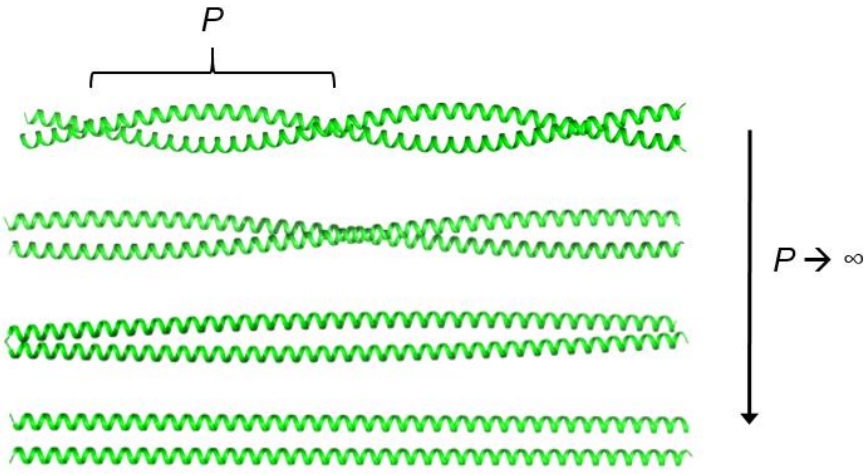


Figure 5.18. As the pitch, P, of a coiled coil super helix becomes increasingly large the angular frequency of the superhelix goes to zero (as P → ∞, ω_0 → 0). This phenomenon occurs in a heterochiral coiled coil where the crossing angle is 0 or π radians due to stereochemical constraints.

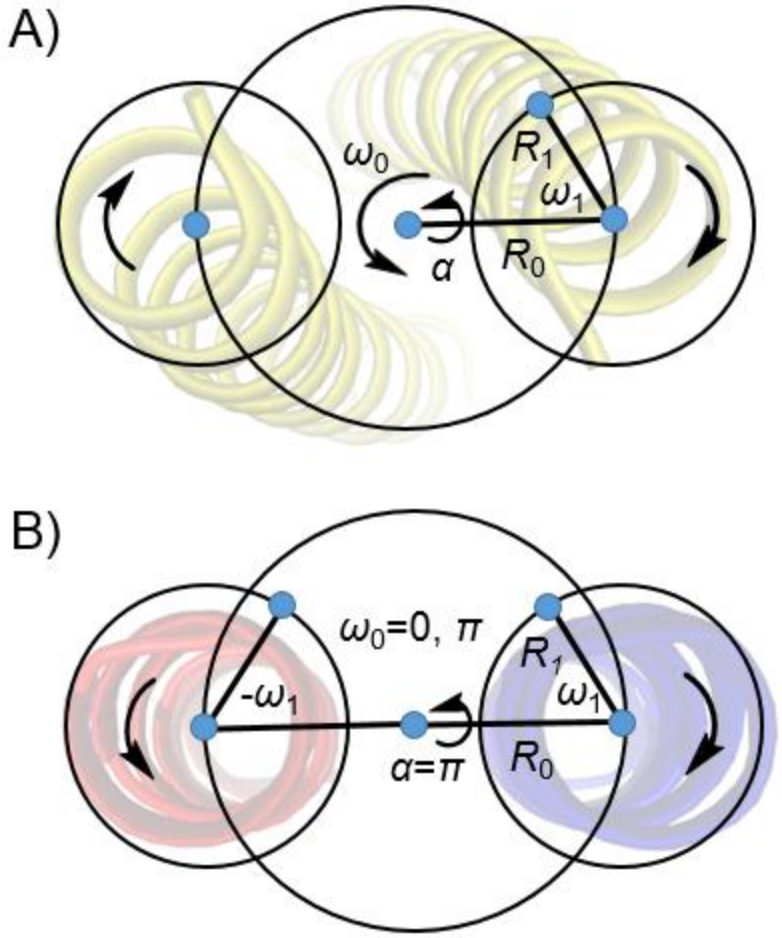


Figure 5.19. A) Schematic definition of the parameters employed in the Crick equations for a canonical coiled coil; ω_0 is the superhelical frequency, R_0 is the superhelical radius, R_1 is the helical radius, ω_1 is the helical frequency and α is the pitch angle between a helical axis and the superhelical axis, which displays left-handed supercoiling. B) In the heterochiral case the helical frequencies have opposite signs and are decoupled from the superhelical frequency which is zero. There are only two possible phases (0, π), which correspond to the right or left-handed helix. In the antiparallel case, applicable to this work, α is equal to π radians (180°).

The Crick parametric equations for a coiled coil are based on the principle of supercoiling, which causes multiple α-helices to twist around a superhelical axis (Figure 5.18). This supercoiling is defined by an angular frequency, ω_0, which depends on the pitch of the supercoil, *P* (the unit distance periodicity). The position, such as that of a

backbone atom, on one of the *n* helices (*n*=2 in Figure 5.19) can be determined if the initial phase of a helix, φ_i, and the helical frequency, ω_1, are known. To determine pertinent model parameters and to compare parameters between M2-TM substitutions, I modified the Crick coiled coil parametric equations for the case of a heterochiral coiled coil system so that the parameters include: R_0, the radius of the coiled coil; R_1, the radius of an individual α-helix; ω, the frequency of each α-helix; φ, the initial phase of each α-helix; *d*, the rise/residue of each α-helix; and Δz, the offset between α-helices. The parametric equations for a heterochiral coiled coil are therefore:

$$x_l = R_0 + R_1 \cos(\omega t + \varphi_l) \quad (5.8)$$

$$y_l = R_1 \sin(\omega t + \varphi_l) \quad (5.9)$$

$$z_l = dt + \Delta z \quad (5.10)$$

$$x_r = -R_0 - R_1 \cos(-\omega t + \varphi_r) \quad (5.11)$$

$$y_r = R_1 \sin(-\omega t + \varphi_r) \quad (5.12)$$

$$z_r = dt \quad (5.13)$$

where *t* is the residue index (*t*=1,2,…,22 for the M2-TM variants). A major difference from the original Crick equations is that in the heterochiral coiled coil system, there is no supercoiling due to the 180° crossing angle, so the *z*-coordinate is set based on the rise per residue of an individual α-helix. The 'superhelix' is simply two parallel lines that have phases of 0 or π radians. Another departure from the Crick system is the inversion of the sign of the helical phase due to the presence of a left-handed α-helix. We adopted the Δz and ΔZ_{off} terms described by Grigoryan, where the sign of Δz depends on the direction the left-handed helix shifts with respect to the right-handed helix; a shift in the C→N direction results in a positive value.[44] ΔZ_{off} is defined as the minimum distance between *z*-values on the parametric helical curves that have a phase of 180°, *i.e.* the minimal

distance (*z*-direction) between points along the curve that point inward towards the central 'superhelical' axis (Figure 5.20). Due to the crystallographic inversion center at the center of the interface, the helical radius, R_1, and helical frequency, ω, are constrained to one value for the two-helix system. The RMSD between the parametric helical curves and the atomic coordinates of the C_α-atoms was used as the objective function during nonlinear regression to determine the parametric model parameters for each heterochiral M2-TM structure (fitting procedure described in Appendix).

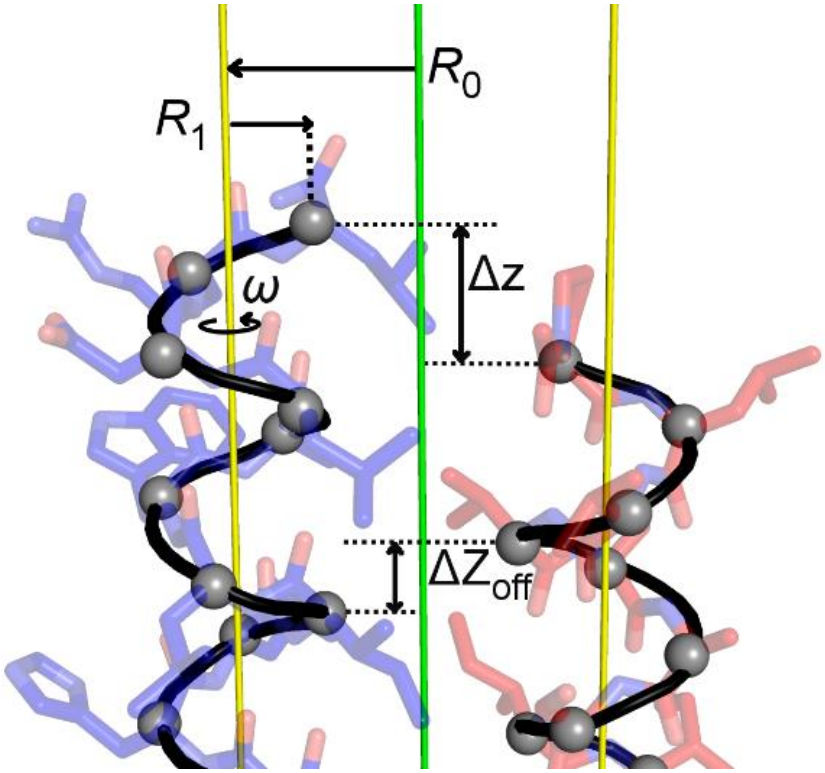


Figure 5.20. The parameterized model for a heterochiral coiled coil is shown for M2-TM I42A. $\mathbf{R}_0$ is the radius of the superhelix, centered on the green axis. R_1 is the radius of an individual helix, centered on the yellow axis. The model assumes no supercoiling, so that the yellow axes are at phases of 0 and 180° with respect to the green axis and correspond to each individual helical axis. The parametric curves (black) trace through the Cα atoms (gray spheres), and the helicity depends only on the helical frequency, ω, and the handedness of the helix. Δz is an offset term that allows for translational motion parallel to the superhelical axis (green). ΔZ_{off} describes the closest offset distance between portions of the parametric curves that are in phase. ΔZ_{off} by definition is constrained between [-2.7 Å, 2.7 Å] for a hendecad.

Variant	R_0 (Å)	R_1 (Å)	ω (°/res.)	φ (°)	d (Å/res.)	Δz (Å)	ΔZ_{off} (Å)	RMSD (Å)
I39A	4.73	2.31	98.0	143.3	1.50	3.37	2.14	0.19
I42A	4.90	2.30	98.5	142.1	1.50	3.98	1.48	0.28
I42E1	5.00	2.28	99.1	138.1	1.50	3.96	1.49	0.32
I42E2	5.04	2.28	100.3	129.0	1.50	3.93	1.47	0.33
M2LCP1	5.01	2.29	98.5	143.2	1.50	3.96	1.50	0.17
M2LCP2	5.06	2.31	98.0	146.7	1.49	4.06	1.41	0.22
M2OG	5.14	2.30	98.0	162.1	1.50	-6.26	0.771	0.23

Table 5.3. Parameterized heterochiral coiled coil model values for all heterochiral interfaces derived from M2-TM. Best fits are shown in Figure 5.21-Figure 5.27.

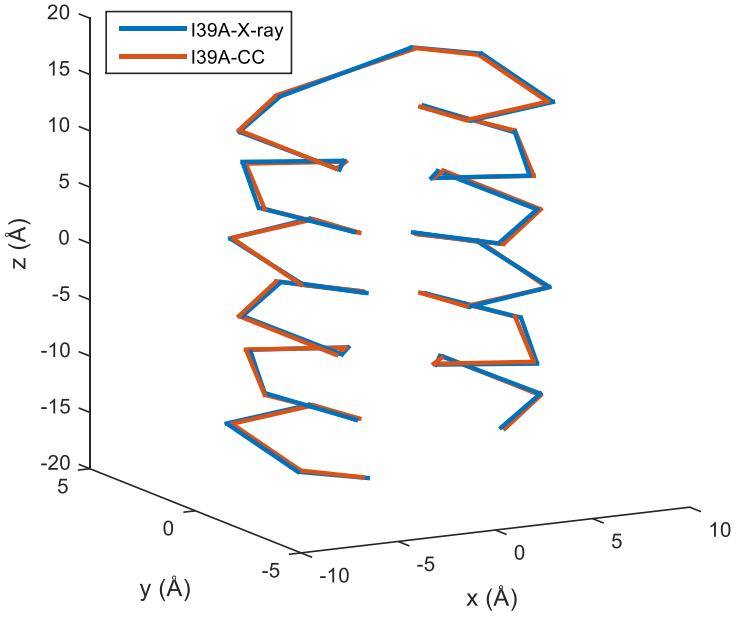


Figure 5.21. Heterochiral, coiled coil parametric equation fit (I39A-CC, red) compared to C_α coordinates extracted from the M2-TM I39A racemate structure (I39A-X-ray, blue).

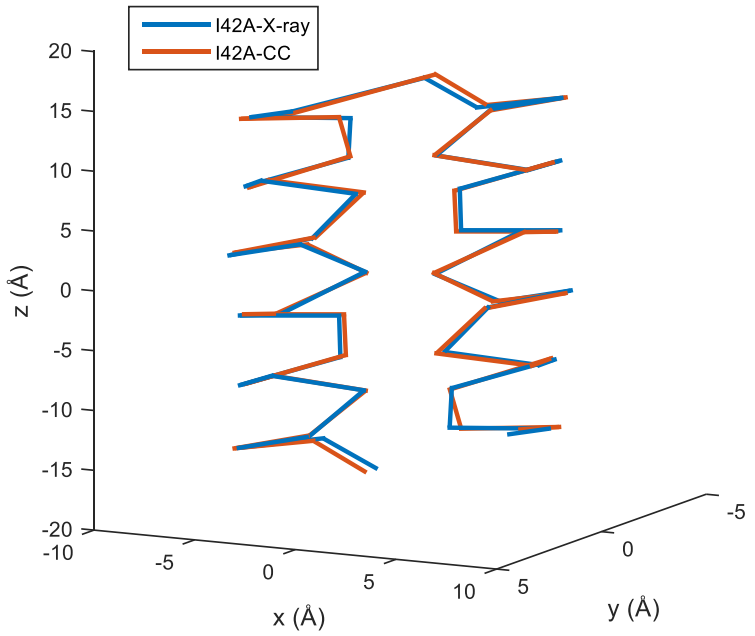


Figure 5.22. Heterochiral, coiled coil parametric equation fit (I42A-CC, red) compared to C_α coordinates extracted from the M2-TM I39A racemate structure (I42A-X-ray, blue).

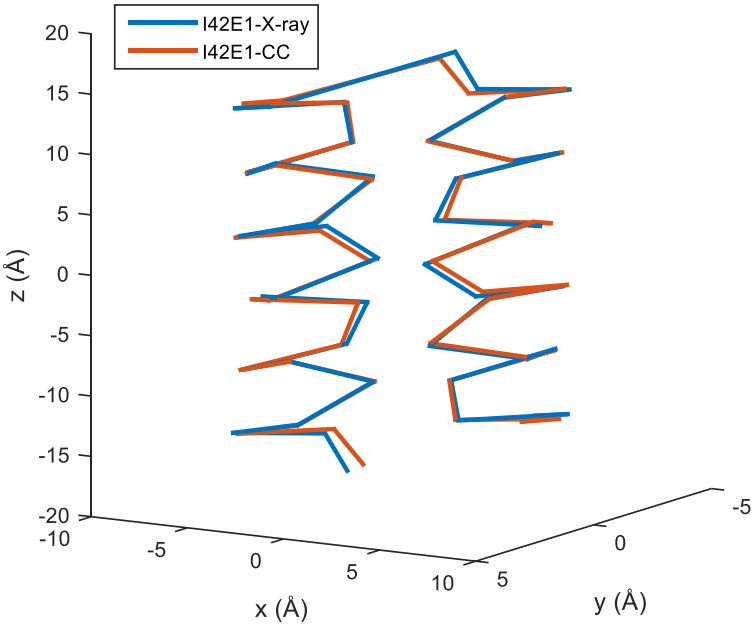


Figure 5.23. Heterochiral, coiled coil parametric equation fit (I42E1-CC, red) compared to C_α coordinates extracted from the M2-TM I39E racemate structure, interface #1 (I42E1-X-ray, blue).

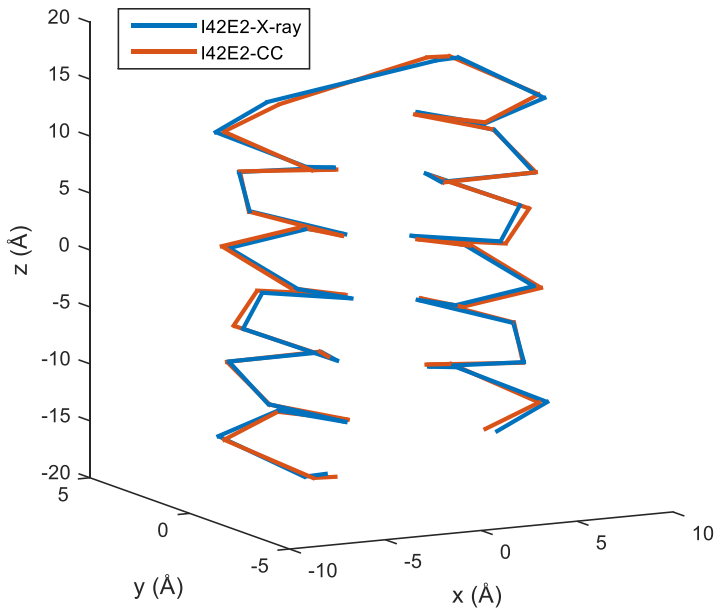


Figure 5.24. Heterochiral, coiled coil parametric equation fit (I42E2-CC, red) compared to C_α coordinates extracted from the M2-TM I39E racemate structure, interface #2 (I42E2-X-ray, blue).

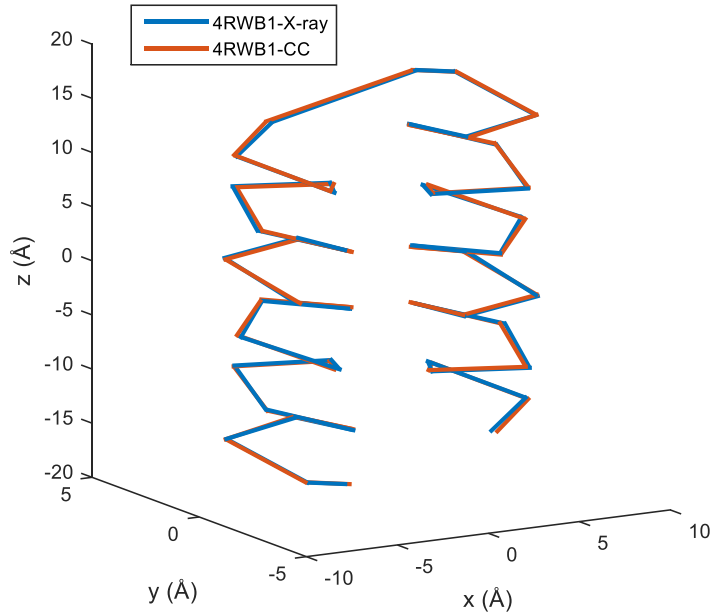


Figure 5.25. Heterochiral, coiled coil parametric equation fit (4RWB-CC, red) compared to C_α coordinates extracted from the M2-TM LCP racemate structure, interface #1 (4RWB1-X-ray, blue).

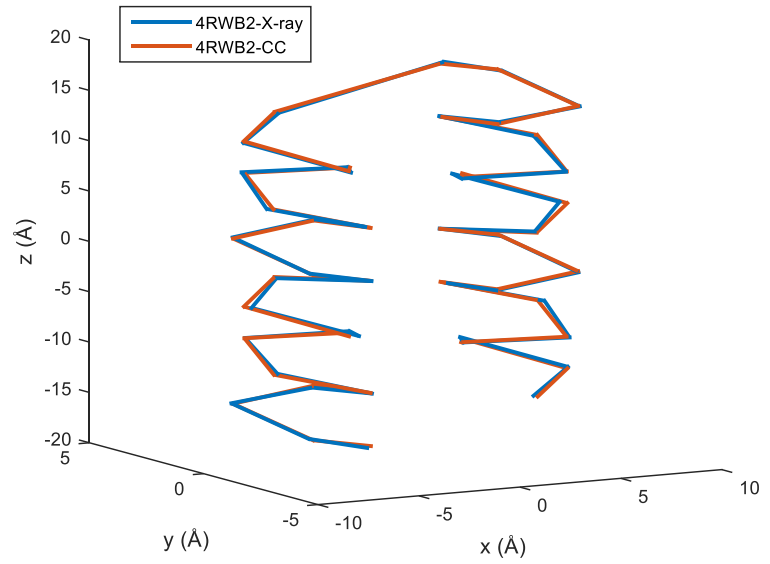


Figure 5.26. Heterochiral, coiled coil parametric equation fit (4RWB-CC, red) compared to C_α coordinates extracted from the M2-TM LCP racemate structure, interface #2 (4RWB2-X-ray, blue).

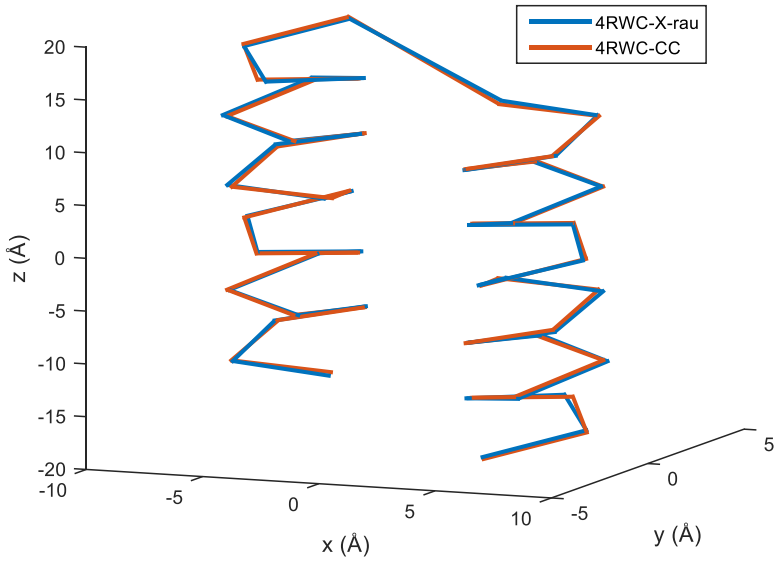


Figure 5.27. Heterochiral, coiled coil parametric equation fit (4RWC-CC, red) compared to C_α coordinates extracted from the M2-TM OG racemate structure (4RWC-X-ray, blue).

The value for R_0 essentially quantifies the 'tightness' of packing between two helices. More closely packed helices will have a smaller R_0 value, which ranges from 4.72 Å (M2-TM I39A) to 5.14 Å (M2-TM OG). This result is consistent with our rationale for the observed shift from a 3,4-spacing in the M2-TM OG to a 3,4,4 hendecad in M2-TM I39A where the Ile→Ala substitution at the *x* position results in a more closely packed interface. Similarly the M2-TM I42A R_0 value of 4.90 Å falls in between the range bounded by M2-TM I39A (4.72 Å) and M2-TM OG (5.14 Å), indicating that smaller knobs are likely more beneficial in terms of close packing than smaller holes. Close packing at interfaces involving Ala residues has been documented in the Alacoil motif found in globular proteins, where alanine side-chains are projected into the upper or lower triad of residues that form the corresponding hole which results in extremely small R_0 values (<4.5 Å).[45] Intriguingly the two M2-TM I42E R_0 values, 5.00 and 5.04 Å, are smaller than those of M2-TM OG, but are similar in size to the values in the M2-TM LCP structure (PDB: 4RWB), 5.01 and 5.06 Å, which contains Ile at both the *x* position knob and the *g* position hole. For a point of reference, the median distribution of R_0 was reported to be 4.82 Å for antiparallel, *homochiral* coiled coil dimers from the CC+ database.[44,46]

The helical frequencies observed for M2-TM I39A (98.0°/res.), I42A (98.5°/res.), and one of the I42E interfaces (99.1°/res.) are consistent with the hendecad hypothesis, which predicts a frequency of 98.2°/res. The second M2-I42E interface frequency value (100.2°/res.) has the largest deviation from 98.2°/res.; however, the margin of difference is greater than the margin between the I42E frequency and the canonical heptad frequency (102.9°/res. two 360° turns over seven residues). The previously reported M2-TM OG and LCP structures also have frequencies consistent with a hendecad repeat.

5.2.7 Motifs Beyond the Hendecad Compatible with Heterochiral Coiled Coils

The main feature of a hendecad motif that makes it compatible with an antiparallel, heterochiral coiled coil is the marginal difference between the residue frequency of the motif (3.67 residues/turn) and that of an ideal α-helix (3.63 residues/turn). The distortion of the α-helix required to accommodate the hendecad motif is less than the distortion required to accommodate the heptad motif (3.50 residues/turn); therefore, helices can readily adopt a conformation that results in a straight display of hydrophobic surface area side-chains. These concepts suggest other possible combinations of three and four residue spacings that can achieve this result. For example, an 18-residue spacing over five turns requires a periodicity of approximately 100°/residue or 3.60 residues/turn, a frequency that was observed for one of the M2-TM I42E interfaces. Recent *de novo* computational efforts have already exploited this 18 residue motif for the design of homochiral coiled coils.[9] Motifs that result in minimal supercoiling, or constrained models in general, may be more tractable from a design standpoint in part due to a decreased search space compared to coiled coil systems that display supercoiling.

In 1953, Crick recognized that dimeric heterochiral coiled coils would necessarily have 0° or 180° crossing angles akin to helical gears of opposite sense in an automobile transmission.[21] Each of the aforementioned motifs (hendecad and 18-residue) is predicated on a homodimeric interaction; however, there is a third possibility for a heterochiral interaction if different sequences are considered, the dimeric pair might not be enantiomers. In a system comprised entirely of L-amino acids, if the periodicity of an ideal α-helix is considered as a ‘baseline’, the ‘drift’ of a heptad motif (3.50 residues/turn) with respect to the baseline is left-handed and approximately -0.13 residues/turn.

Similarly, a 15 residue, pentadecad, motif over four turns has a periodicity of 3.75 residues/turn and thus a right-handed drift of approximately 0.12 residues/turn with respect to the baseline. Therefore, inversion of the chirality of a pentadecad helix, via D-amino acids, will in turn invert the handedness of the hydrophobic residue display so that it is now left-handed and commensurate with the periodicity of a heptad helix comprised of L-amino acids. Other possible heterodimeric, heterochiral pairings are also possible, such as an L/D hendecad motif paired with a D/L 18 residue motif.

5.3 Conclusions

Substitutions at a heterochiral interface that displayed 3,4-spacing of KIH interactions caused the interface to shift to a 3,4,4-hendecad motif in X-ray crystal structures obtained via racemic crystallography. Smaller *x* position knobs (Ile→Ala) or *g* position holes (Ile→Ala, Ile→Glu) appeared to bias the heterochiral dimer towards a hendecad motif where the presence of the hendecad motif was confirmed by cataloguing knob-into-hole interactions and parameterizing the X-ray crystal structures to determine the helical frequency of each structure. The helical frequencies from three of the interfaces, M2-TM I39A, I42A and I42E1 were found to be in agreement with the frequency of a hendecad repeat (98.2°/residue), and the interface that had an alanine residue at an *x* position knob (M2-TM I39A) showed the tightest coiled coil radius of 4.72 Å. We reasoned that in a dimeric coiled coil that features hendecad hydrophobic side-chain display, antiparallel heterochiral coiled coils are the preferred pairing arrangement because they minimize steric overlap in knob-to-knob interactions between *x*-layer residues. The small difference between the frequency of a hendecad repeat

(98.2°/residue) and the frequency of an α-helix (99.2°/residue) suggests the possibility for heterochiral coiled coil interfaces that extend over multiple hendecad repeats. Based solely on the geometry of an α-helix and the geometry of hendecad hydrophobic side-chain display there is nothing limiting the length of such an interface; solubility (of a long fiber) is likely to be the main design constraint. Extended, specific interfaces could potentially be designed by careful placement of *x*-layer residues.

Two of the newly reported M2-TM variant structures (M2-TM I39A and I42A) displayed type I (layered) crystal packing where the I39A variant featured more disordered layering compared to the I42A variant as determined by diffraction anisotropy. The more ordered layering in the M2-TM I42A racemate structure was ascribed to interactions between L-Arg45 side chains and D-Arg45/Asp44 carbonyls (or L-Arg45 side chains and L-Arg45/Asp44 carbonyls) centered on an inversion center. The predisposition of macromolecular and small organic molecular crystals for centrosymmetric space groups has been established (Chapter 1). With regard to the physical principles that govern space group preferences, TM domain peptide crystals (solvent volume ~20-30%) occupy the middle of a spectrum flanked by closely packed organic molecular crystals and solvated macromolecular crystals (solvent volume ~50%). Inversion symmetry is likely to enhance the likelihood of TM domain crystallogenesis and facilitate TM domain crystal growth. In the context of lamellar type I crystal packing, this enhancement likely will be mediated through both extensive antiparallel heterochiral interfaces (intralayer) and side-chain/backbone interactions between ionizable functional groups (interlayer), such as the salt-bridging interaction between L-Arg45 side chains and D-Arg45/Asp44 carbonyls observed in the M2-TM I42A racemate structure. Whether or not natural,

homochiral interactions persist in these heterochiral packing assemblies, remains to be seen.

5.4 Appendix

Zhihui Yao synthesized, purified, crystallized and refined the structure for racemic M2 I42E.

5.4.1 Peptide Synthesis and Purification

L and D-Fmoc amino acids were purchased from Chem-Impex Intl. (Woodale, IL, USA). Peptides were prepared using either manual, microwave-assisted (L-M2 I42E, D-M2 I42E) or automated (L-M2 I39A, L-M2 I42A, D-M2 I39A, D-M2 I42A) solid phase peptide synthesis. L-peptides were synthesized on a Symphony Automated Peptide Synthesizer (Protein Technologies Inc. Tuscon, AZ, USA). D-peptides were synthesized on a Prelude™ Automated Peptide Synthesizer (Protein Technologies Inc.) charged with Fmoc-D-amino acids. In both cases ChemMatrix® H-PAL resin (~0.5 mmol/g; PCAS Biosciences Montreal, Quebec, CA) was used as the solid support. Peptides were synthesized on a 100 umol scale, and the following coupling/deprotection/acetylation procedures were employed:

Automated (N_2 mixing, room temperature)

A: 0.1 M HCTU (ChemImpex Intl. Woodale, IL, USA), 0.4 M *N*-methylmorpholine (Sigma-Aldrich, Milwaukee, WI, USA) in DMF (biotechnology grade, Sigma-Aldrich).

B: 0.1 M Fmoc-amino acid (ChemImpex Intl.) in DMF (biotechnology grade)

C: 20% v/v piperidine (Sigma-Aldrich) in DMF (ACS reagent grade, Sigma-Aldrich)

Coupling: 2x1 h, 8 mL 1:1 A:B

D: ACS reagent grade DMF (Sigma-Aldrich)

Deprotection: 1) 5 min 2) 15 min 4 mL C

Coupling wash: 2x30 sec D

Deprotection wash: 5x30 sec D

Manual Microwave-Assisted Synthesis

Manual synthesis was performed in fritted 12 mL polypropylene syringes (Torviq, Niles, MI, USA). 400 µmol Fmoc-amino acid and HATU (ChemImpex Intl.) were weighed out into a 14 mL polycarbonate tube and then dissolved in 4 mL of DMF (biotechnology grade). Subsequently, 800 µmol of neat *N,N*-diisopropylethylamine (DIEA; Sigma-Aldrich) was pipetted into the Fmoc-amino acid/HATU solution. This 'activated solution' was vortexed for 5-10 seconds and then transferred to the fritted syringe. A CEM Mars400 microwave reactor, outfitted with a fiber optic temperature probe, was used to accelerate coupling and deprotection steps. Stirring was achieved with a magnetic flea stir bar in the resin bed. Coupling procedures comprised 2x12 min reactions held at 70°C with a 2 min ramp time. Between reactions the resin was drained and fresh coupling solution was added. For deprotection steps approximately 3-5 mL of 20% v/v piperidine in DMF (ACS reagent grade) was added to the resin. Deprotection procedures comprised 1x4 min reactions (2 min ramp time to 80°C followed by a 2 min hold). Finally, between coupling and deprotection steps the resin was washed ~3 times with DMF (ACS reagent grade).

Capping/Acetylation

Following the final deprotection step the resin was rinsed 3 times with DMF (ACS reagent grade) and then rinsed 3 times with DCM (ACS reagent grade, Sigma-Aldrich) and allowed to dry for approximately 10 min. Finally, 5 mL of an 8:2:1 solution comprised of

DMF (ACS reagent grade): DIEA: acetic anhydride (ACS reagent grade) was added to the resin bed, which was stirred for approximately 15 min, then drained and rinsed 3 times with DMF (ACS reagent grade) followed by rinsing 3 times with DCM (ACS reagent grade).

Following synthesis, peptides were cleaved from the resin with 8 mL of a 95:2.5:2.5 solution comprised of 2,2,2-trifluoroacetic acid (TFA; ACS reagent grade, Sigma-Aldrich): triisopropylsilane (TIPS; ACS reagent grade, Sigma-Aldrich): water (millipore filtered). The cleavage cocktail was transferred to the polypropylene syringe containing the dry resin, which was sealed with the syringe plunger and allowed to nutate for approximately 3 h. The cleavage solution was then filtered into a 50 mL conical polycarbonate tube through the syringe frit and concentrated with an N_2 stream to afford a thick brown oil. The crude peptide oil was precipitated in 2x40 mL diethyl ether and sedimented via centrifugation in the 50 mL polycarbonate tube (4,000 rpm); ether was decanted after each step. Residual solvent was removed from the off-white crude precipitate with a gentle N_2 stream to afford an off-white powder. The crude solid was then dissolved in ~1 mL 1:1 TFA:2,2,2-trifluoroethanol (TFE; ACS reagent, Sigma-Aldrich) and diluted to 10 mL with acetonitrile (HPLC grade, Sigma-Aldrich) for immediate reverse phase HPLC purification.

Reverse phase HPLC purification was used to purify each M2 variant following solid phase synthesis. A binary gradient was employed with the following solvent system:

A solvent: 9:1 H_2O (millipore filtered):acetonitrile (HPLC grade, Sigma-Aldrich), 0.05% v/v TFA (ACS reagent grade, Sigma-Aldrich)

B solvent: 6:3:1, isopropanol (HPLC grade, Sigma-Aldrich): acetonitrile (HPLC grade, Sigma-Aldrich): H_2O (millipore filtered), 0.05% v/v TFA (ACS reagent grade, Sigma-Aldrich)

Typical HPLC methods comprised an 85→95% B solvent over 20 min gradient followed by a 5 min isocratic flush (95% B solvent) and 5 min re-equilibration (85% B solvent). Flow rates ranged from 9-10 mL/min with a Supelco C5 column (21.2x250 mm), and peaks were collected based on UV absorbance at 220 nm. Peak identities were established with MALDI-TOF mass spectrometry. Confirmed peaks were pooled into a 250 mL round-bottom flask and slowly concentrated via rotary evaporation (room temperature, water aspirator) to remove organic volatiles prior to lyophilization. Pooled, concentrated fractions were diluted with water (so that the total volume was approximately 150% of the volume prior to dilution), frozen with dry ice, and lyophilized to afford a white powder that was subjected to mass (MALDI-TOF) and purity (UPLC) analysis.

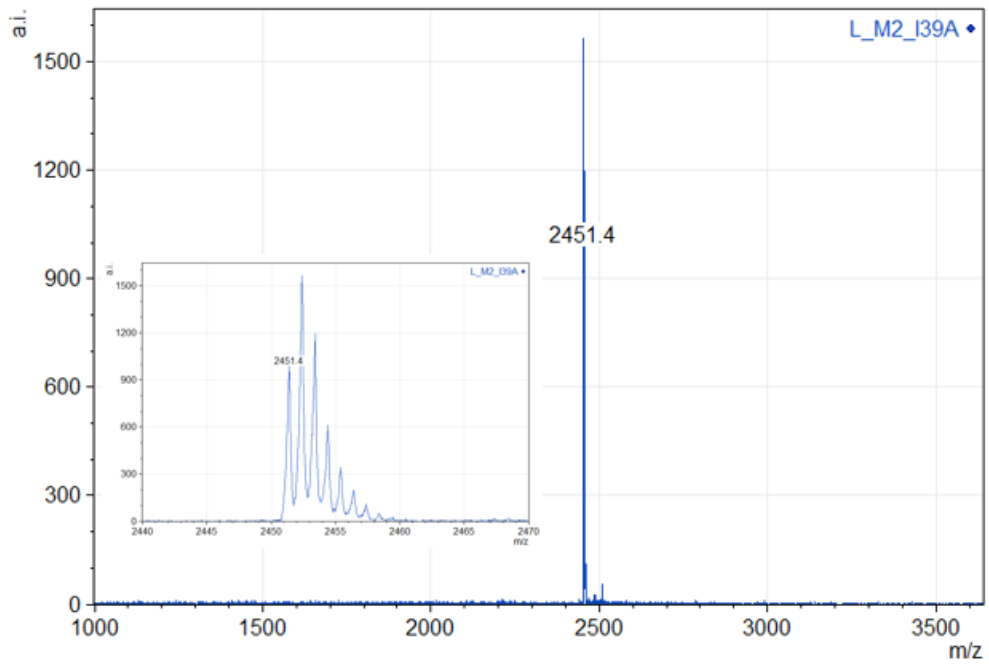
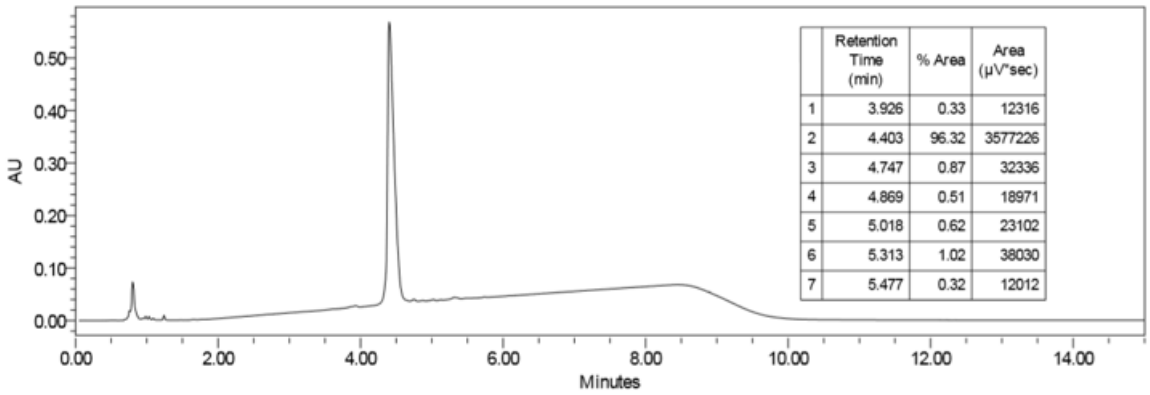


Figure 5.28. Ac-PLVVAASIIAILHL**A**LWILDRL-CONH$_2$ L-M2 I39A calculated monoisotopic [M+H]=2451.5, observed monoisotopic [M+H]=2451.4.



	Retention Time (min)	% Area	Area (µV*sec)
1	3.926	0.33	12316
2	4.403	96.32	3577226
3	4.747	0.87	32336
4	4.869	0.51	18971
5	5.018	0.62	23102
6	5.313	1.02	38030
7	5.477	0.32	12012

Figure 5.29. L-M2 I39A UPLC purity analysis, UPLC gradient from 10-98% B solvent over 7 minutes (0.35 mL/min, 50°C; column-Waters Acquity BEH C4 1.7 µm, 2.1 x 100 mm, purity >95%).

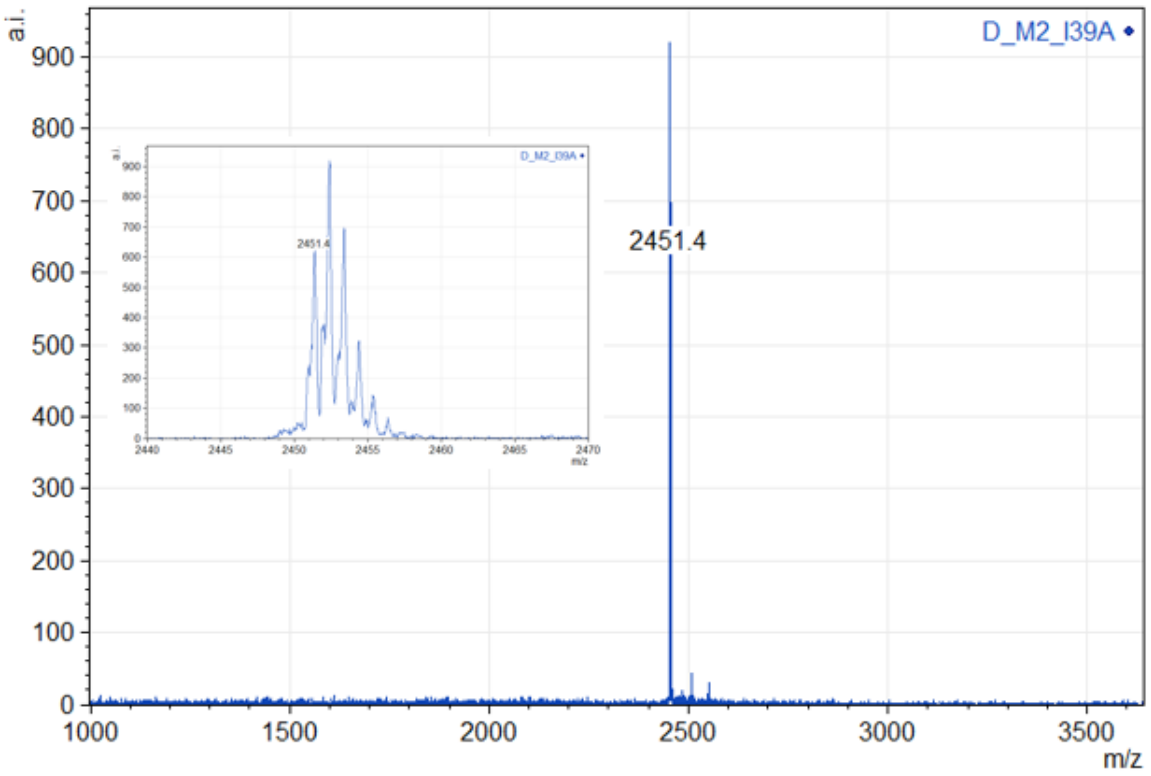
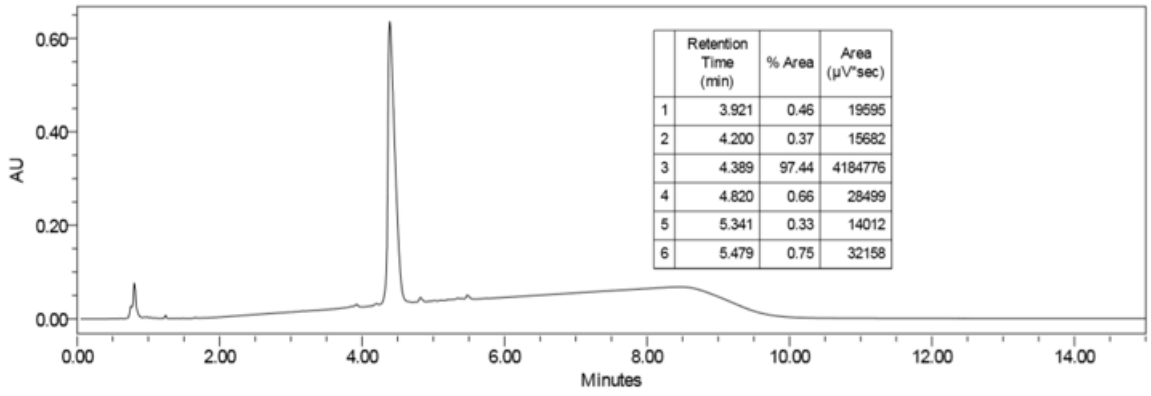


Figure 5.30. Ac-PLVVAASIIAILHL**A**LWILDRL-CONH$_2$ D-M2 I39A calculated monoisotopic [M+H]=2451.5, observed monoisotopic [M+H]=2451.4.



	Retention Time (min)	% Area	Area (µV*sec)
1	3.921	0.46	19595
2	4.200	0.37	15682
3	4.389	97.44	4184776
4	4.820	0.66	28499
5	5.341	0.33	14012
6	5.479	0.75	32158

Figure 5.31. D-M2 I39A UPLC purity analysis, UPLC gradient from 10-98% B solvent over 7 minutes (0.35 mL/min, 50°C; column-Waters Acquity BEH C4 1.7 µm, 2.1 x 100 mm, purity >95%).

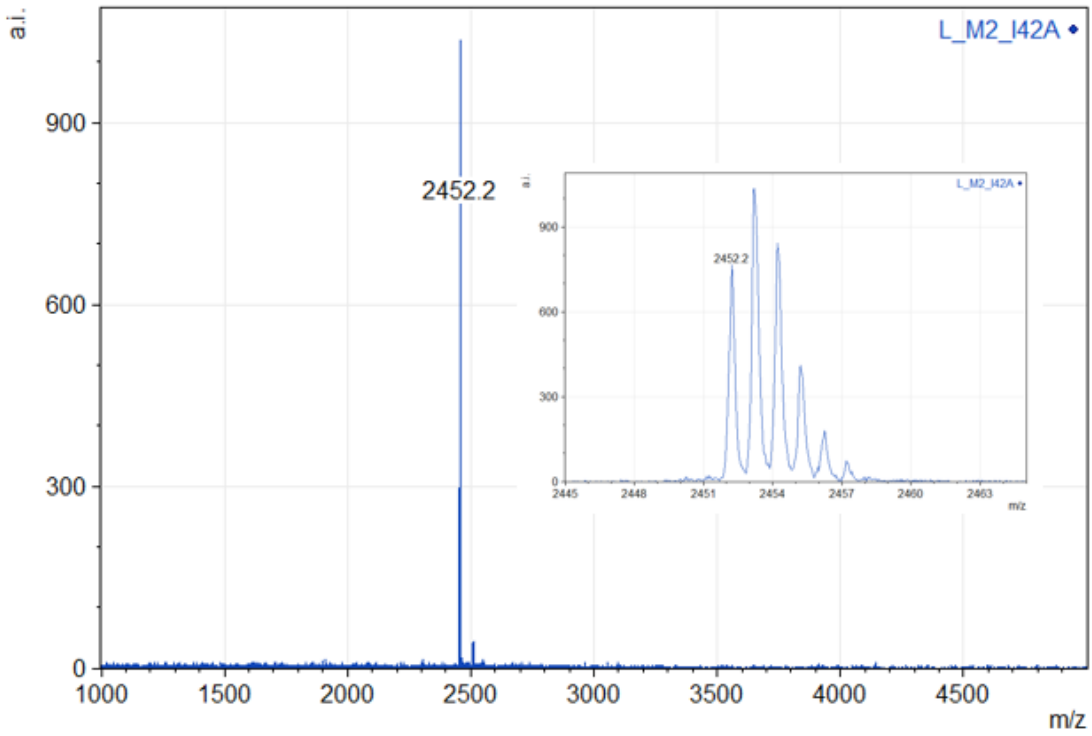
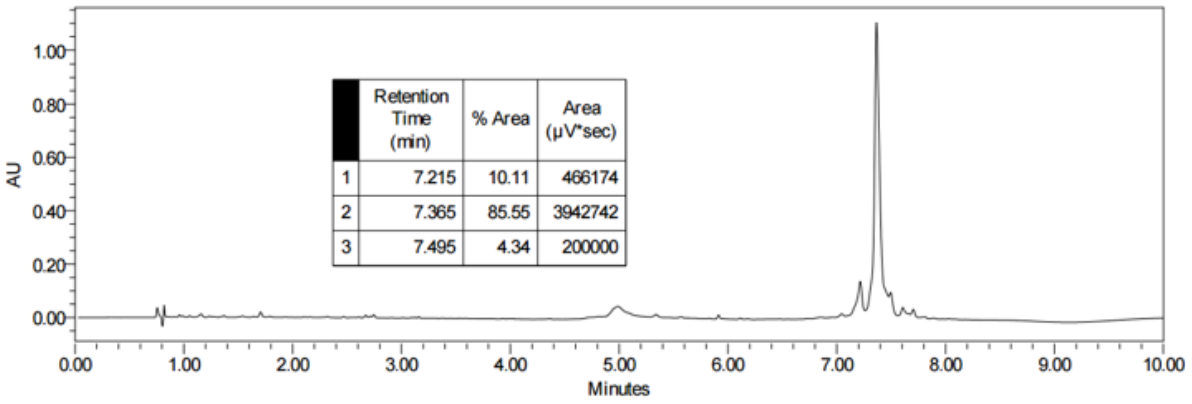


Figure 5.32 Ac-PLVVAASIIAILHLALW**A**LDRL-CONH$_2$ L-M2 I42A calculated monoisotopic [M+H]=2451.5, observed monoisotopic [M+H]=2452.2.



	Retention Time (min)	% Area	Area (µV*sec)
1	7.215	10.11	466174
2	7.365	85.55	3942742
3	7.495	4.34	200000

Figure 5.33. L-M2 I42A UPLC purity analysis, UPLC gradient from 10-98% B solvent over 7 minutes (0.35 mL/min, 50°C; column-Waters Acquity BEH C4 1.7 µm, 2.1 x 100 mm, purity >85%).

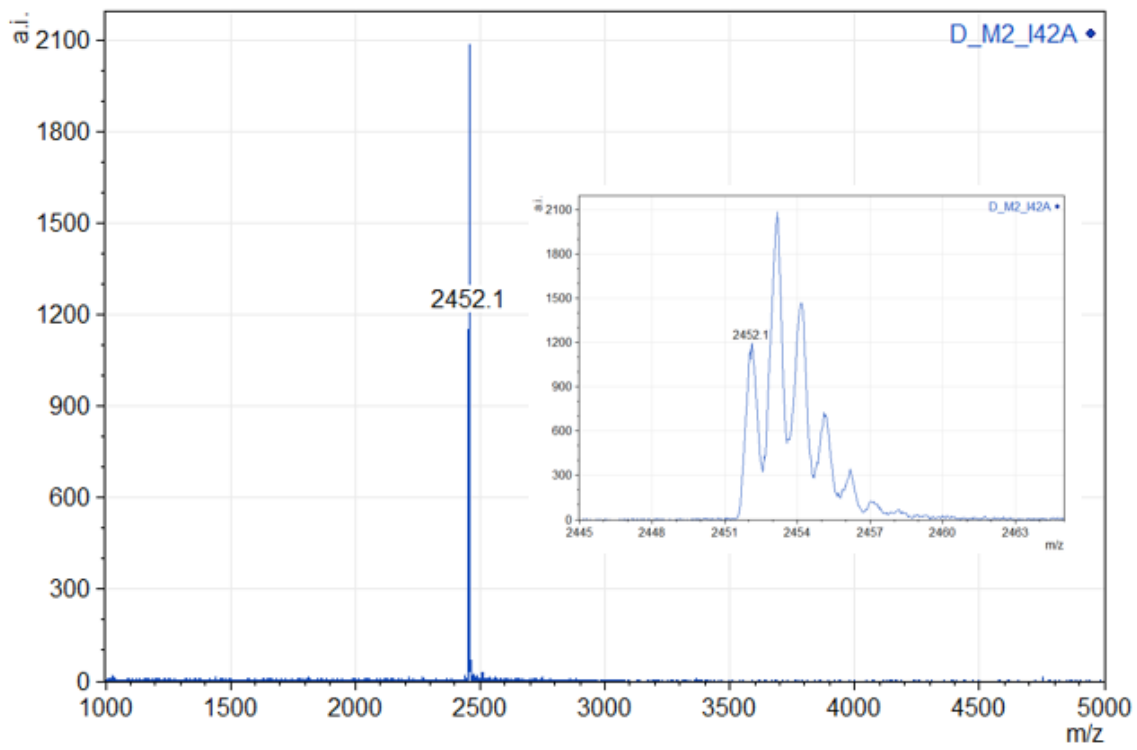


Figure 5.34. Ac-PLVVAASIIAILHLALWILDRL-CONH$_2$ D-M2 I42A calculated monoisotopic [M+H]=2451.5, observed monoisotopic [M+H]=2452.1.

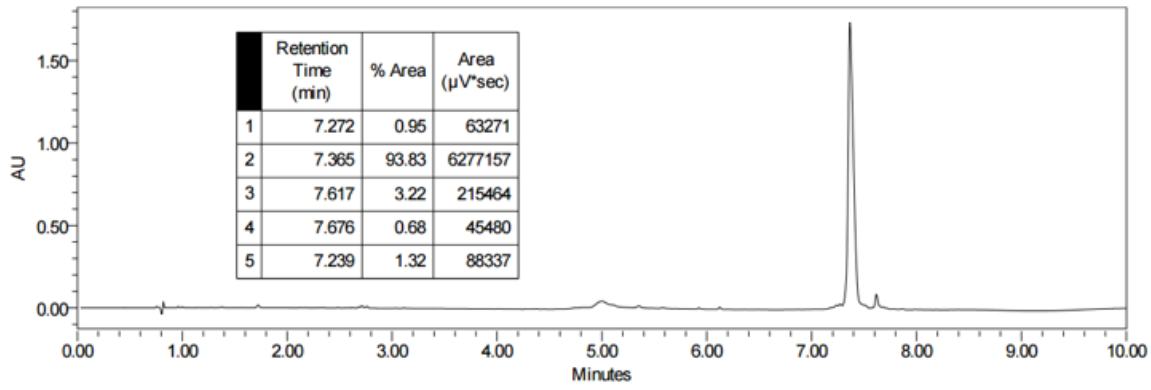


Figure 5.35. D-M2 I42A UPLC purity analysis, UPLC gradient from 10-98% B solvent over 7 minutes (0.35 mL/min, 50°C; column-Waters Acquity BEH C4 1.7 µm, 2.1 x 100 mm, purity >90%).

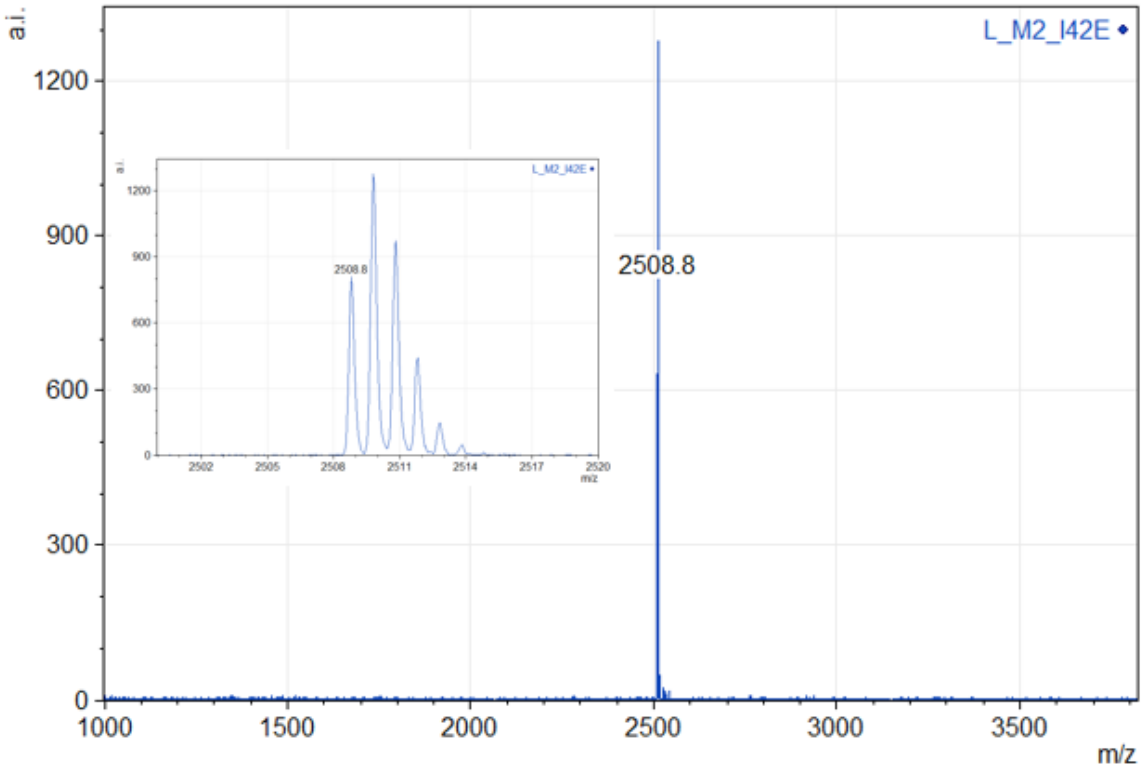


Figure 5.36. Ac-PLVVAASIIAILHLALW**E**LDRL-CONH$_2$ L-M2 I42E calculated monoisotopic [M+H]=2509.5, observed monoisotopic [M+H]=2508.8.

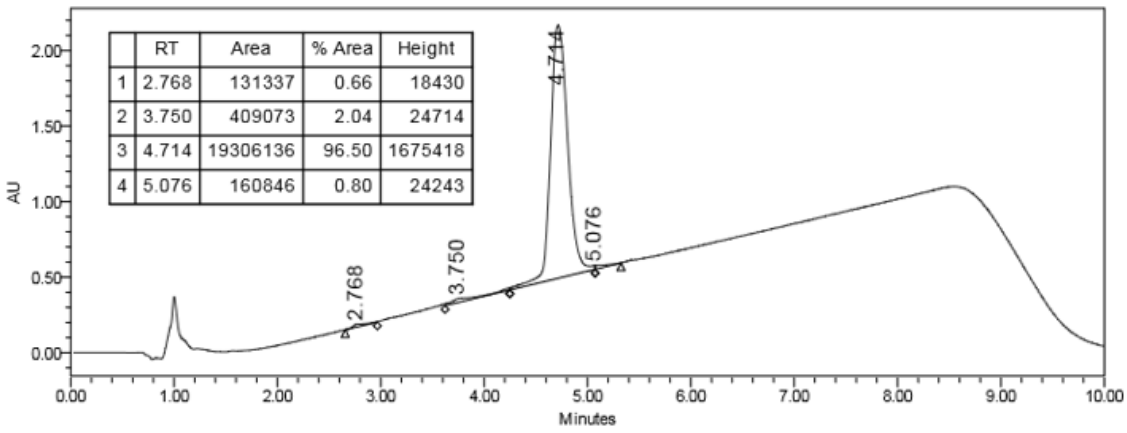


Figure 5.37. L-M2 I42E UPLC purity analysis, UPLC gradient from 10-60% B solvent over 7 minutes (0.35 mL/min, 55°C; column-Waters Acquity BEH C4 1.7 µm, 2.1 x 100 mm, purity >95%).

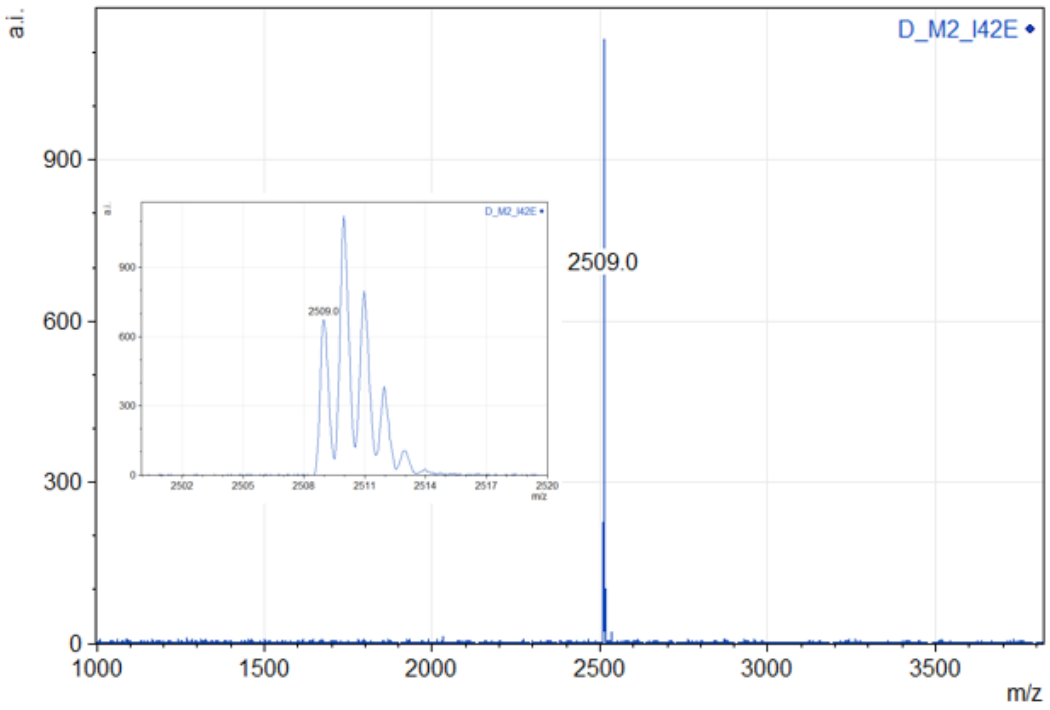


Figure 5.38. Ac-PLVVAASIIAILHLALW**E**LDRL-CONH$_2$ L-M2 I42E calculated monoisotopic [M+H]=2509.5, observed monoisotopic [M+H]=2509.0.

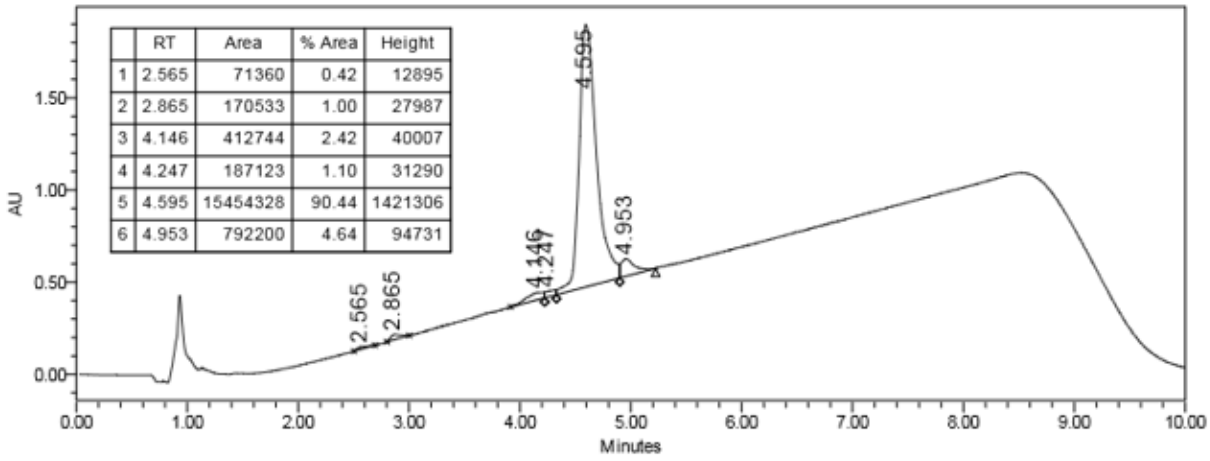


Figure 5.39. D-M2 I42E UPLC purity analysis, UPLC gradient from 10-60% B solvent over 7 minutes (0.35 mL/min, 55°C; column-Waters Acquity BEH C4 1.7 µm, 2.1 x 100 mm, purity >90%).

5.4.2 Peptide Crystallization

Racemic peptide solutions in racemic 2% w/v β-octyl-glucoside solution (OG; D-β-octyl glucoside, purchased from Sigma-Aldrich; L-β-octyl glucoside, provided by Anatrace) were prepared with the following procedure. First, racemic OG solutions were made up by separately dissolving each crystalline OG enantiomer in millipore water and then combining each enantiomerically pure solution in a 1:1 volumetric ratio. The racemic OG solution was then used to separately dissolve each peptide enantiomer (lyophilized powder). Concentrations were ~2.5 mg/mL for I39A/I42A variants and ~3.0 mg/mL for I42E variants. Finally, racemic-OG/L-peptide and racemic-OG/D-peptide solutions were combined in a 1:1 molar ratio, as determined by UPLC (*vide infra*), and the resulting racemic-OG-racemic-peptide solution was used for subsequent hanging drop vapor diffusion crystallization.

Before combining racemic-OG/L-peptide and racemic-OG/D-peptide solutions, Peak areas were quantified for separate 1 µL injections of each peptide solution (same day). The volume required for a 1:1 molar ratio was determined based on these peak areas and finally, this volume amount was used to prepare the racemic solution used for crystallization.

Table 5.4. Optimized crystallization conditions for M2 racemates I39A, I42A, and I42E

Variant	Tray	Well Volume	Well Solution	Drop Composition
I39A	48 well VDX48 (Hampton Research)	350 µL +100 µL Al's oil	20% v/v MPD*, 0.1 M ADA* pH 6.5	1.5 µL well solution: 1.5 µL racemic peptide/OG solution
I42A	48 well VDX48 (Hampton Research)	200 µL	0.1 M sodium citrate tribasic, 0.1 M TRIS-HCl pH 8.5, 5% v/v PEG400	1.0 µL well solution: 1.0 µL racemic peptide/OG solution
I42E	48 well VDX48 (Hampton Research)	500 µL	0.8 M imidazole, pH 7.0	1.0 µL well solution: 1.0 µL racemic peptide/OG solution

*ADA = *N*-(2-acetamido)iminodiacetic acid (BioUltra, Sigma-Aldrich); note, 1.0 M stock ADA solutions had to be diluted initially with 1 M NaOH instead of millipore water.
*MPD = racemic 2-methyl-2,4-pentanediol

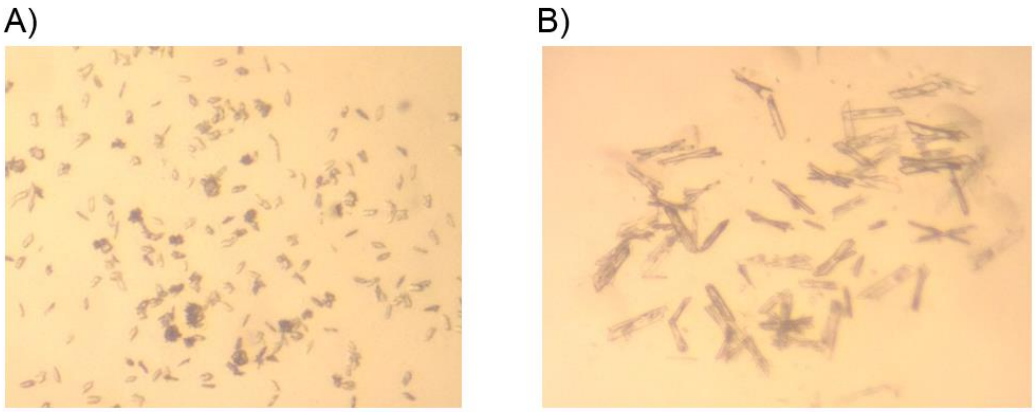


Figure 5.40. A) Lead condition for optimization from initial crystallization screens of racemic M2 I39A (hanging drop vaport diffusion; 0.1 M ADA pH 6.5, 12% MPD). B) The addition of Al's oil (1:1 silicone:paraffin oil) appeared to accelerate crystal growth in two dimensions (optimized well solution; Table 5.4). Photographs were taken with an iPhone 4S through a microscope lens at ~60x magnification.

5.4.3 X-ray Data Collection

X-ray data were collected at the Advanced Photon Source (Argonne National Laboratory) Sector 21 ID-F (M2 I42E) and ID-G (M2 I39A, M2 I42A). The following optimized cryo-protection strategies were employed for each M2 racemate:

M2 I39A: Crystals were harvested from the drop with a MicroLoops™E mount (MiTeGen, Ithaca, NY, USA); soaked for ~30 seconds in a ~1 µL droplet of 1:1 glycerol: well solution; and plunged into liquid N_2 for vitrification and storage prior to data collection. The MiTeGen Microloops™ E mount was essential for supporting the thin plate-like morphology of racemic M2 I39A crystals during data collection (Figure 5.40).

M2 I42A: Crystals were harvested from the drop with a MicroMeshes™ mount (MiTeGen); dragged through a ~1 µL droplet containing 30% v/v glycerol in the well solution; and plunged into liquid N_2 for vitrification and storage prior to data collection.

M2 I42E: Crystals were harvested from the droplet with a nylon loop; dipped in a droplet of 30% v/v PEG400 in the well solution; and plunged into liquid N_2 for vitrification and storage prior to data collection.

5.4.4 Structure Solution and Refinement

360° datasets comprised of 1° oscillations with varying detector distance and exposure time were collected for each racemic crystal (Table 5.5,Table ***5.6***,Table ***5.7***). Data were indexed, integrated, and scaled with the program *XDS*.[47] Friedel mates and identical reflections were merged with the program *XSCALE*.[47] The oblique lattice constants for each dataset indicated a triclinic crystal system, which simplified space group assignment to a binary choice of either *P*1 (no symmetry) or *P*$\bar{1}$ (inversion

symmetry). Analysis of the cumulative intensity distributions with the program *xprep* provided evidence for the presence of a center of symmetry (<|E^2-1|> values: M2 I39A, 1.240; M2 I42A, 1.040; M2 I42E, 0.942; theoretical centric, 0.968). The <|E^2-1|> value is effectively a measure for the spread of a normalized intensity distribution, so the <|E^2-1|> value for the M2 I39A dataset is likely slightly elevated due to the observed anisotropic diffraction.

Initial molecular replacement attempts in space group *P*$\bar{1}$ were unsuccessful and resulted in poor quality electron density maps that were indecipherable. Therefore, M2 I39A and M2 I42A structures were solved with molecular replacement in space group *P*1 using the program *Phaser*[48] with both L and D-polyalanine M2-TM, derived from PDB: 4RWC, as search models. A problem then arose during the conversion from *P*1 to *P*$\bar{1}$ due to the arbitrary origin in a *P*1 lattice. In a primitive lattice, *i.e.* in the absence of translational symmetry from lattice centering, inversion centers must be located at the origin for the operator *x*,*y*,*z* → -*x*,-*y*,-*z* to remain valid. Likewise, in a *P*2_1 space group the origin can be located anywhere along the *b* axis (aligned with the 2_1 screw axis), whereas in the centrosymmetric space group P2_1/c, the inversion center must be located at the origin (along the *b* axis). To overcome this origin placement problem in the *P*1 *Phaser* solution, the inversion center was assumed to be at the interface between an L and a D-helix. The fractional coordinates of a numerical center-of-mass (COM) were computed between L and a D-helices in the *P*1 solution, where each atom was weighted equally regardless of identity. These fractional coordinates were then added to each atom in the model, effectively moving the 'inversion center' to the crystallographic origin. Finally, D-helix atoms were removed from the model and the model was subjected to rigid-body

refinement in space group $P\bar{1}$ to yield an initial *R*-factor of ~60% (a completely random centrosymmetric structure has an *R*-factor of 0.828; Figure 5.41, Figure 5.42).

Following structure solution, missing side chain atoms were built in manually with iterative real-space refinement in *Coot*.[49] After placement of side chain atoms iterative rounds of *x,y,z* coordinate and atomic displacement parameter (ADP) refinement were carried out in *phenix.refine*.[50] The favorable data-to-parameter ratio allowed for anisotropic refinement of displacement parameters. L and D-β-octyl-glucoside restraints were generated with *phenix.elbow*[51] and placed in positions indicated by the $2mF_o$-DF_c weighted electron density maps. Finally, relative target function weights of X-ray and chemical restraint contributions were optimized with the protocol provided in *phenix.refine*.

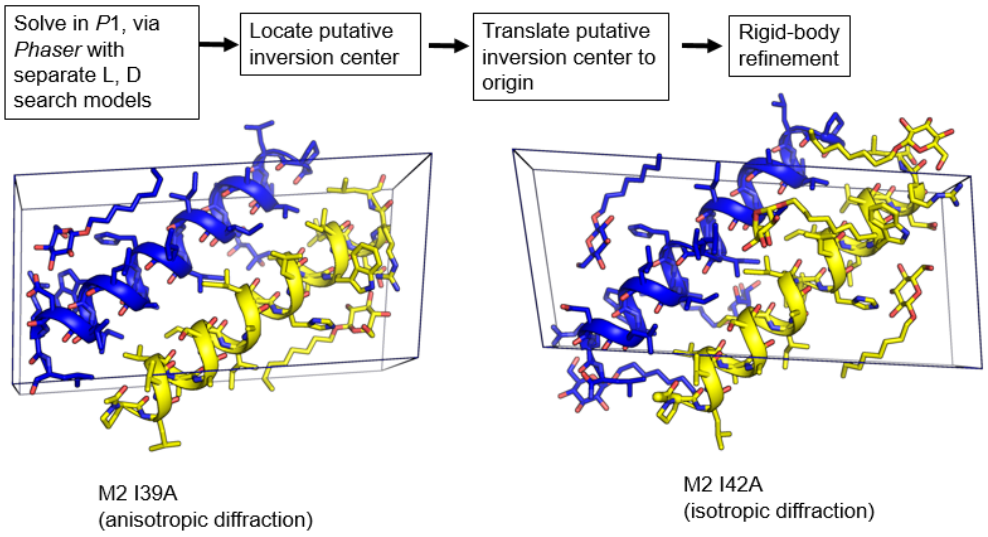


Figure 5.41. Flow diagram of the procedure used to phase M2 I39A and M2 I42A racemate structures with molecular replacement.

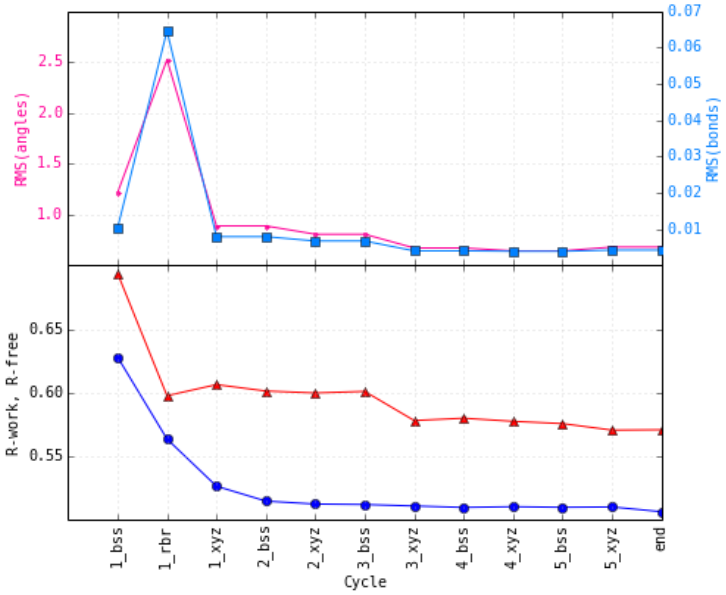


Figure 5.42. The initial rigid body refinement cycle (1_rbr) of the origin-shifted P$\bar{1}$ M2 I39A structure carried out in phenix.refine locates the true inversion center (R-free, red; R-work, blue). The numeral indicates the refinement cycle and the three letter acronym indicates the refinement step (bss, bulk solvent scaling; rbr, rigid-body refinement; xyz, atomic coordinate refinement).

Table 5.5. M2-TM I39A

Data collection	
X-ray source	APS 21-ID-G
X-ray detector	MAR300
Detector distance (mm)	180
Exposure time (s)	1
Oscillation range (º)	1
Wavelength (Å)	0.97856
Space group	$P\bar{1}$
a / b / c (Å)	17.1 / 18.8 /37.0
α / β / γ (º)	94.3 / 90.0 / 109.5
Volume (Å^3)	11220
Matthews coefficient (Å^3/Da)	1.85
Solvent content (%)	33.5
Molecules in asymmetric unit	1
Resolution range (Å)	18.45-1.55 (1.605-1.55)
Number of Reflections	19780 (1986)
Unique Reflections	5992 (581)
Completeness	95.2% (94.3%)
Redundancy	3.3 (3.4)
Mean *I* / σ(*I*)	11.7 (5.0)
CC1/2*	0.998 (0.993)
R_{merge}	0.0440 (0.150)
R_{meas}	0.0529
Wilson *B*-factor (Å^2)	49.0
Refinement statistics	
Refinement program	phenix.refine: 1.9_1692
Non-hydrogen protein atoms refined	240
Resolution range (Å)	18.45-1.55 (1.71-1.55)
No. of reflections used in refinement	5979
Completeness (%)	95.7 (93.0)
Reflections in cross-validation set	607
R-value (work)	0.239
R-value (free)	0.276
R-value (overall)	0.243
Coordinate error (ML, Å)	0.05
Mean ADP (Å^2)	33.6
Mean ADP macromolecule (Å^2)	27.7
RMSD Bond lengths (Å)	0.010
RMSD Bond angles (°)	2.50

Table 5.6. M2-TM I42A

Table 5.2. M2-TM I42A	
Data collection	
X-ray source	APS 21-ID-G
X-ray detector	MAR300
Detector distance (mm)	170
Exposure time (s)	2
Oscillation range (º)	1
Wavelength (Å)	0.97856
Space group	$P\bar{1}$
a / b / c (Å)	17.3 / 18.5 / 40.2
α / β / γ (º)	81.2 / 88.0 / 69.1
Volume (Å^3)	11900
Matthews coefficient (Å^3/Da)	1.96
Solvent content (%)	37.3
Molecules in asymmetric unit	1
Resolution range (Å)	17.11-1.40 (1.45-1.40)
Number of Reflections	32816 (3353)
Unique Reflections	8609 (891)
Completeness	94.6% (94.2%)
Redundancy	3.8 (3.8)
Mean *I* / σ(*I*)	10.9 (2.3)
CC1/2*	0.999 (0.882)
R_{merge}	0.0501 (0.624)
R_{meas}	0.0585
Wilson *B*-factor (Å^2)	48.4
Refinement statistics	
Refinement program	phenix.refine: 1.9_1692
Non-hydrogen protein atoms refined	249
Resolution range (Å)	17.11-1.40 (1.49-1.40)
No. of reflections used in refinement	8604
Completeness (%)	94.6 (94.0)
Reflections in cross-validation set	854
R-value (work)	0.223
R-value (free)	0.231
R-value (overall)	0.224
Coordinate error (ML, Å)	0.05
Mean ADP (Å^2)	26.5
Mean ADP macromolecule (Å^2)	19.0
RMSD Bond lengths (Å)	0.003
RMSD Bond angles (°)	0.74

Table 5.7. M2-TM I42E

Table 5.3. M2-TM I42E	
Data collection	
X-ray source	APS 21-ID-F
X-ray detector	MAR300
Detector distance (mm)	200
Exposure time (s)	2
Oscillation range (º)	1
Wavelength (Å)	0.97872
Space group	P1-
a / b / c (Å)	14.7 / 37.3 / 43.2
α / β / γ (º)	107.4 / 96.7 / 97.7
Volume ($Å^3$)	22040
Matthews coefficient ($Å^3$/Da)	1.82
Solvent content (%)	32.3
Molecules in asymmetric unit	2
Resolution range (Å)	18.34-1.40 (1.45-1.40)
Number of Reflections	55889 (2695)
Unique Reflections	14624 (744)
Completeness	86.8% (43.2%)
Redundancy	3.8 (3.6)
Mean I / σ(I)	10.1 (1.8)
CC1/2*	0.998 (0.843)
R_{merge}	0.0656 (0.642)
R_{meas}	0.0766
Wilson B-factor ($Å^2$)	48.4
Refinement statistics	
Refinement program	phenix.refine: 1.9_1692
Non-hydrogen protein atoms refined	420
Resolution range (Å)	18.34-1.40 (1.51-1.40)
No. of reflections used in refinement	14614
Completeness (%)	86.8 (52.0)
Reflections in cross-validation set	731
R-value (work)	0.226
R-value (free)	0.236
R-value (overall)	0.228
Coordinate error (ML, Å)	0.02
Mean ADP ($Å^2$)	27.0
Mean ADP macromolecule ($Å^2$)	20.9
RMSD Bond lengths (Å)	0.009
RMSD Bond angles (°)	1.24

5.4.5 Constrained Non-Linear Regression Analysis for Fitting Heterochiral Coiled Coil Parametric Equations in MATLAB

The parameters in the heterochiral coiled coil parametric equations are moderately correlated; therefore, unconstrained non-linear regression is exceedingly difficult. To overcome this difficulty, constrained non-linear regression is employed with a very reasonable initial parameter guess (restraints are in [brackets]):

{R_0 = 4.6, [4.2, 4.9]; R_1 = 2.3, [1.6, 2.8]; φ_l = 0,[-2π, 2π]; Δz = 0, [-5, 5]; d = 1.5, [1.46, 1.53]; φ_r = 0, [-2π, 2π]}

p0 = [4.6,2.3,1.4,0,0,1.5,0];

lower_bound = [4,2,1.6,-2*pi,-5,1.46,-2*pi];

upper_bound = [4.9,2.8,2*pi,2*pi,5,1.53,2*pi];

CA=load('39A_CA.m');

(39A_CA.m is a three column text file that contains the x,y,z coordinates of each C_α atom extracted from the M2-TM I39A PDB file)

[p1, fval, exitflag, output] = fmincon(@(p) crick_rmsd(CA,p), p,[ ],[ ],[ ],[ ],lower_bound, upper_bound);

```
function rmsd = crick_rmsd(CA,p0)
%CA is a Nx3 matrix, experimental C-alpha coordinates from X-ray structure
%p0 is initial parameter guess
%1) generate parameteric model coordinates from parameter guess
%2) quantify deviation from parametric model and X-ray structure
XYZ = crick(p0);
rmsd = superimpose(XYZ',CA');
end
```

```
function [XYZ] = crick(p0)
%function returns Cartesian coordinates of antiparallel, heterochiral
%coiled coil given a parameter vector
%model params:
%R0    = superhelix radius
%R1    = alpha-helix radius
%w1    = helical frequency
%phi_l = left-handed helix, initial phase angle
```

```
%delz  = vertical offset between L and D-helices (Angstroms)
%d     = vertical rise/residue (Angstroms/residue
%phi_r = right-handed helix, initial phase angle

t = 1:1:22; %residue index for 22 residue L/D-M2
R0 = p0(1);
R1 = p0(2);
w1 = p0(3);
phi_l = p0(4);
del_z = p0(5);
scale = p0(6);
phi_r = p0(7);
xl = R0 + R1*cos(w1*t + phi_l);
yl = R1*sin(w1*t+phi_l);
zl = t*d+del_z;
xr = -R0 - R1*cos((-1*w1*t + phi_r));
yr = R1*sin(-w1*t+phi_r);
zr = t*d;
x = [xl'; xr'];
y = [yl'; yr'];
z = [zl'; flipud(zr')]; %account for antiparallel helices
XYZ = [x y z];

function [rmsd, M, X, Y] = superimpose(X_in, Y_in)
% [rmsd, M] = superimpose(X, Y)
% For two 3-by-N matrices, X and Y, returns the best fit root mean square
% deviation between the two, resulting from an optimal alignment
% (a translation and a rotation by some angle around some axis) of X onto
% Y. The matrix performing this optimal linear transformation is returned
% as well. Also optionally returns the
%
% Implements the SVD method by Kabsch et al. (Acta Crystallogr 1976, A32,
% 922) and taken from Coutsias et al. J Comp Chem 2004, 25(15) 1849.

N = size(X_in, 2);

X = X_in - repmat(mean(X_in, 2), 1, N);
Y = Y_in - repmat(mean(Y_in, 2), 1, N);
R = X*(Y');
[V,S,W] = svd(R);
I = eye(3,3);
sgn = sign(det(R));
I(3,3) = sgn;
M = W*I*V';
rmsd = sqrt((sum(sum(X.^2) + sum(Y.^2)) - 2*(S(1,1) + S(2,2) +
sgn*S(3,3)))/N);

x_ = M*X;
plot3(x_(1,:),x_(2,:),x_(3,:),'linewidth',2)
hold on
plot3(Y(1,:),Y(2,:),Y(3,:),'linewidth',2)

end
```

5.4.6 Non-covalent Layer-Spanning Interactions in Previously Reported Transmembrane (TM) Domain Structures

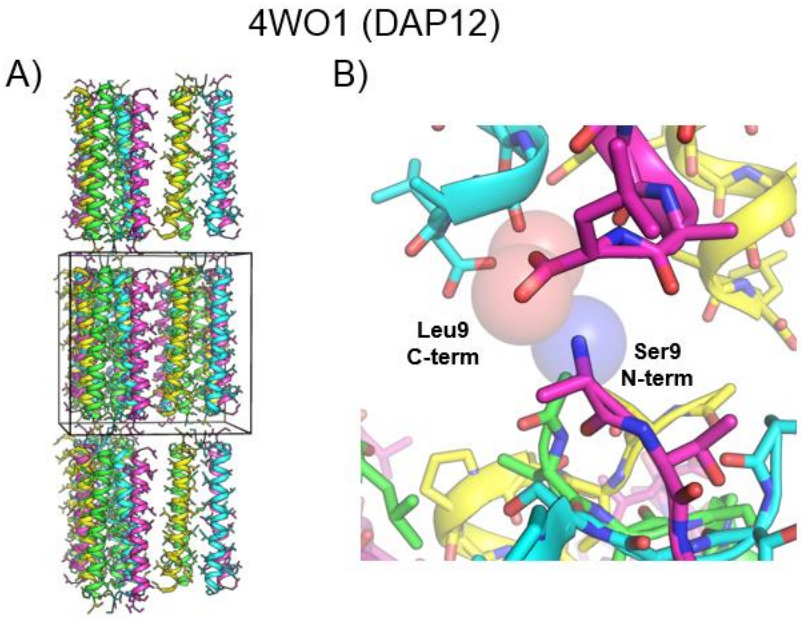


Figure 5.43. A) DAP12 TM domain type I packing in an orthorhombic lattice. B) A Coulombic interaction between N and C-terminus of a symmetrically equivalent chain.

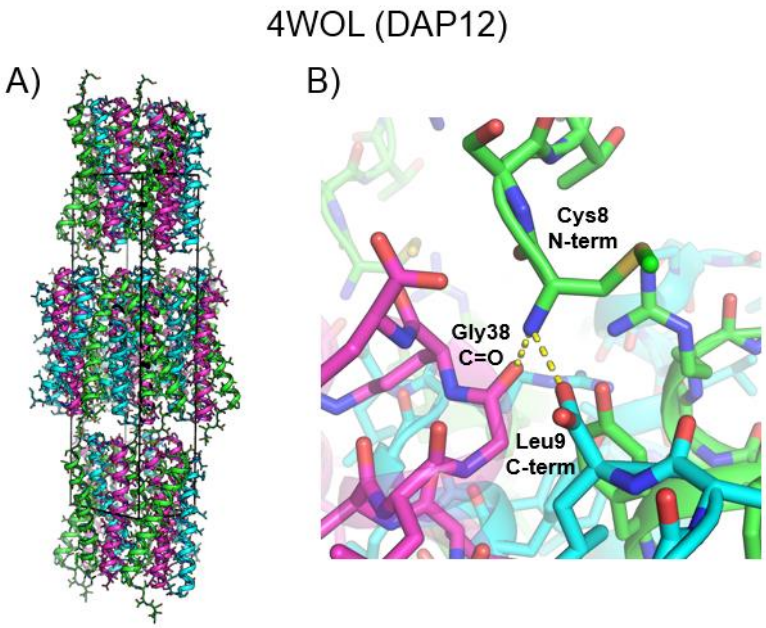


Figure 5.44. A) DAP12 TM type I packing in an orthorhombic lattice. B) A tripartite salt-bridging/H-bonding interaction between the N-terminus of one chain and the C-termini of chains in an adjacent layer.

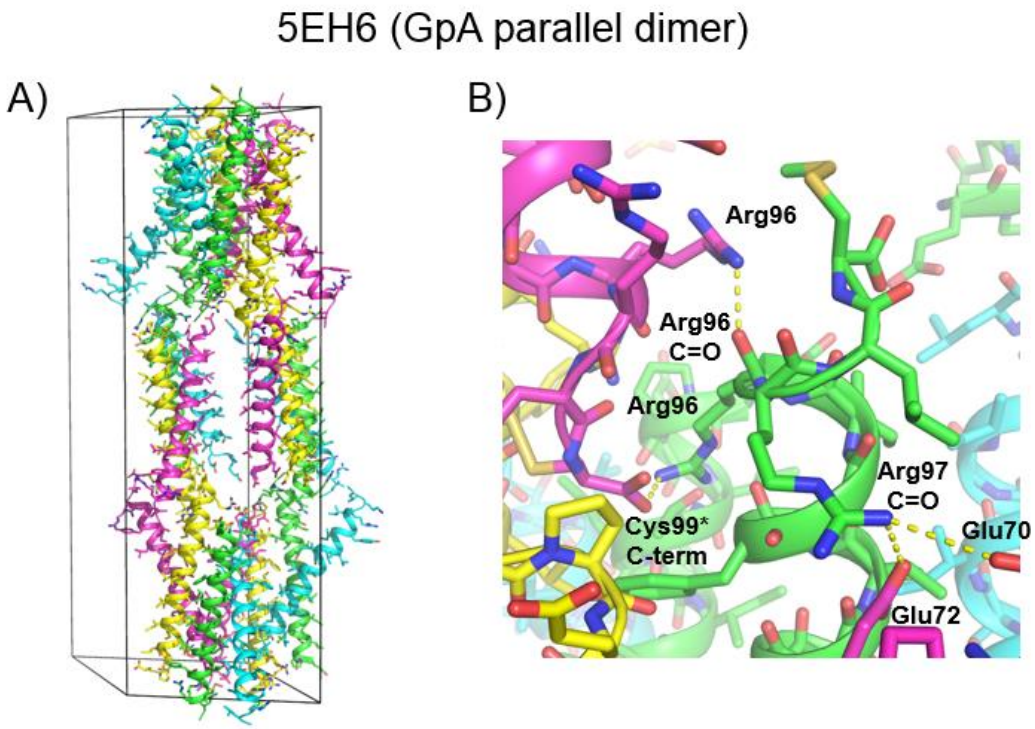


Figure 5.45. A) Two parallel homodimers are related by an antiparallel interface between dimers in an R32 crystal. The two dimers are related by a two-fold symmetry axis (perpendicular to the long edge of the unit cell). B) A network of H-bonding and salt-bridging interactions between layers.

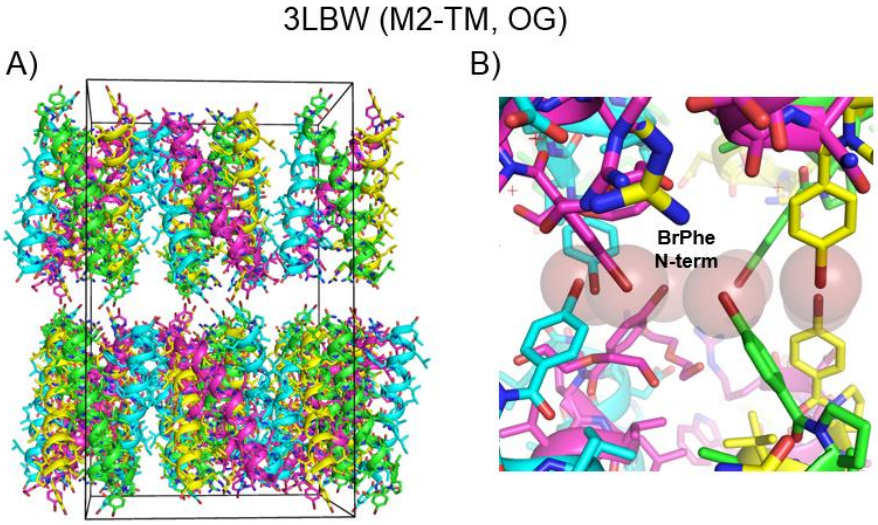


Figure 5.46. A) The first reported TM domain structure displays type I packing in an orthorhombic lattice. B) N-terminal bromophenylalanine was used to experimentally phase the structure (SAD). Packing between layers is mediated by non-covalent contacts between bromine atoms (drawn with 1.8 Å radius), which is a unique feature among all reported TM domain structures.

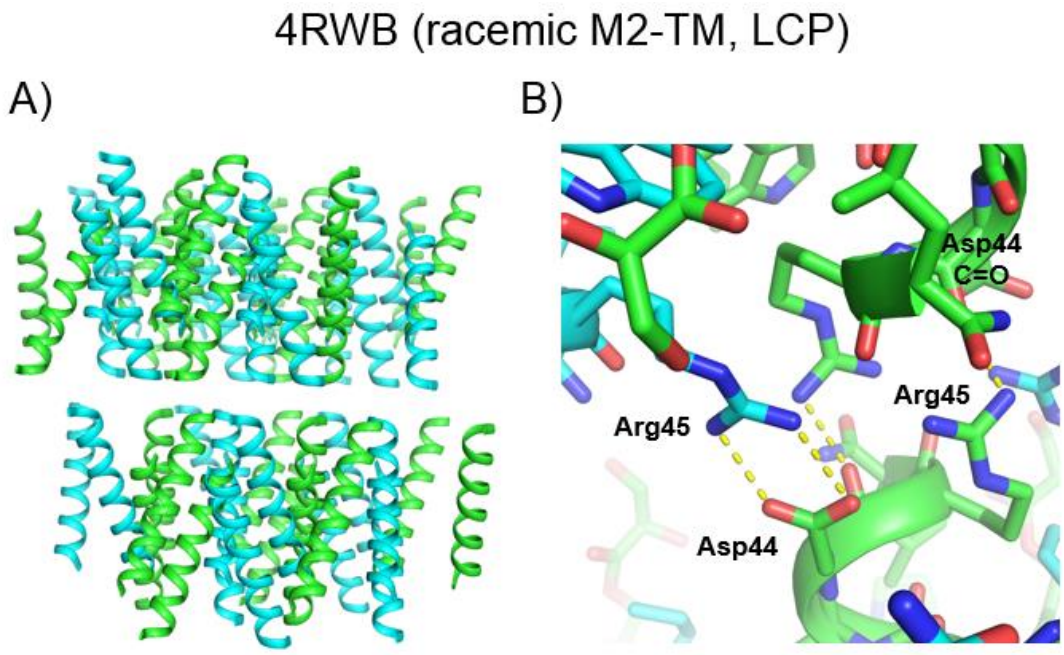


Figure 5.47. A) Racemic M2 crystals grown from LCP display type I packing in a monoclinic lattice, which is unique among all currently reported TM domain structures. B) In contrast to the homochiral LCP derived M2 structures, interlayer contacts are mediated by charged side chains rather than monoolein.

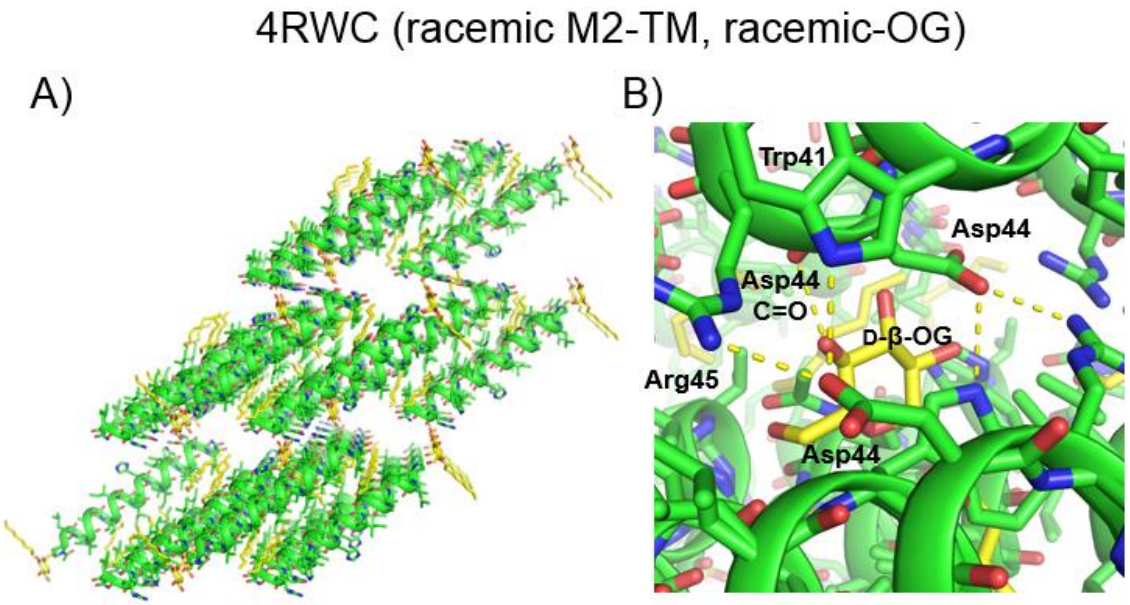


Figure 5.48. A) Racemic M2, crystallized from racemic-L/D-β-octyl glucoside (rac-OG), displays type I packing in a triclinic lattice and features a unique heterochiral interface. B) Interlayer packing interactions are mediated by charged side chains and an L/D- β-octyl glucoside molecule.

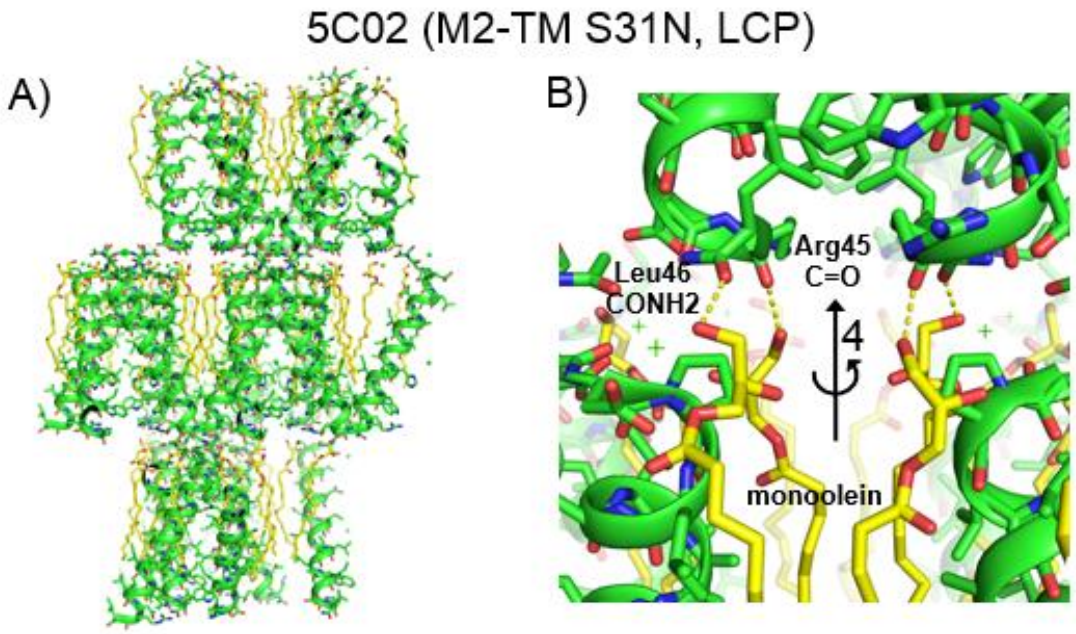


Figure 5.49. A) An amantadine resistant variant of M2 (S31N) crystallizes with type I packing in a tetragonal lattice. B) A channel of monoolein, centered on the four-fold axis, makes H-bonding contacts with backbone carbonyls.This phenomonen is represented in the schematic in Figure 5.8.

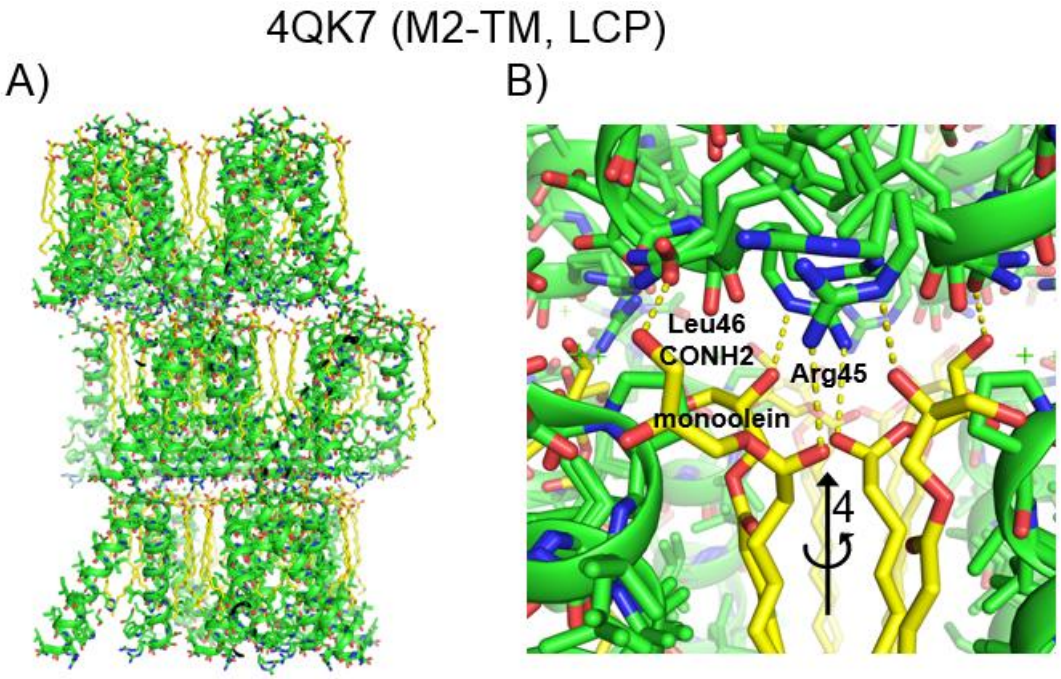


Figure 5.50. A) Homochiral M2-TM crystallizes with type I packing in a tetragonal lattice. B) The monoolein channel, centered on the four-fold axis, makes H-bonding contacts with charged side chains and backbone carbonyls. Similar interactions were observed in PDB: 4QKC.

5.5 References

(1) Weinstock, M. T.; Francis, J. N.; Redman, J. S.; Kay, M. S. *Biopolymers* **2012**, *98* (5), 431–442.

(2) Welch, B. D.; VanDemark, A. P.; Heroux, A.; Hill, C. P.; Kay, M. S. *Proc. Natl. Acad. Sci. U. S. A.* **2007**, *104* (43), 16828–16833.

(3) Miller, S. M.; Simon, R. J.; Ng, S.; Zuckermann, R. N.; Kerr, J. M.; Moos, W. H. *Drug Dev. Res.* **1995**, *35*, 20–32.

(4) Patterson, W. R.; Anderson, D. H.; Degrado, W. F.; Cascio, D.; Eisenberg, D. *Protein Sci.* **1999**, *8*, 1410–1422.

(5) Rodriguez-Granillo, A.; Annavarapu, S.; Zhang, L.; Koder, R. L.; Nanda, V. *J Am Chem Soc* **2011**, *133* (46), 18750–18759.

(6) Shepherd, N. E.; Hoang, H. N.; Abbenante, G.; Fairlie, D. P. *J. Am. Chem. Soc.* **2009**, No. 8, 15877–15886.

(7) Anil, B.; Song, B.; Tang, Y.; Raleigh, D. P. *J. Am. Chem. Soc.* **2004**, *126* (41), 13194–13195.

(8) Uppalapati, M.; Lee, D. J.; Mandal, K.; Li, H.; Miranda, L. P.; Lowitz, J.; Kenney, J.; Adams, J. J.; Ault-Riché, D.; Kent, S. B. H.; Sidhu, S. S. *ACS Chem. Biol.* **2016**, *11* (4), 1058–1065.

(9) Bhardwaj, G.; Mulligan, V. K.; Bahl, C. D.; Gilmore, J. M.; Harvey, P. J.; Cheneval, O.; Buchko, G. W.; Pulavarti, S. V. S. R. K.; Kaas, Q.; Eletsky, A.; Huang, P.-S.; Johnsen, W. A.; Greisen, P. J.; Rocklin, G. J.; Song, Y.; Linsky, T. W.; Watkins, A.; Rettie, S. A.; Xu, X.; Carter, L. P.; Bonneau, R.; Olson, J. M.; Coutsias, E.; Correnti, C. E.; Szyperski, T.; Craik, D. J.; Baker, D. *Nature* **2016**, *538* (7625), 329–335.

(10) Simon, M. D.; Maki, Y.; Vinogradov, A. A.; Zhang, C.; Yu, H.; Lin, Y. S.; Kajihara, Y.; Pentelute, B. L. *J. Am. Chem. Soc.* **2016**, *138* (37), 12099–12111.

(11) Dintzis, H. M.; Symer, D. E.; Dintzis, R. Z.; Zawadzke, L. E.; Berg, J. M. *Proteins Struct. Funct. Bioinforma.* **1993**, *16* (3), 306–308.

(12) Schumacher, T. N. M.; Mayr, L. M.; Minor Jr., D. L.; Milhollen, M. A.; Burgess, M. W.; Kim, P. S. *Science (80-.).* **1996**, *271*, 1854–1857.

(13) Guichard, G.; Benkirane, N.; Zeder-Lutz, G.; van Regenmortel, M. H.; Briand, J. P.; Muller, S. *Proc. Natl. Acad. Sci. U. S. A.* **1994**, *91* (21), 9765–9769.

(14) Li, C.; Pazgier, M.; Li, J.; Li, C.; Liu, M.; Zou, G.; Li, Z.; Chen, J.; Tarasov, S. G.; Lu, W. Y.; Lu, W. *J. Biol. Chem.* **2010**, *285* (25), 19572–19581.

(15) Lupas, A. N.; Gruber, M. *Adv. Protein Chem.* **2005**, *70*, 37–38.

(16) Woolfson, D. N. *Adv. Protein Chem.* **2005**, *70*, 79–112.

(17) Hicks, M. R.; Walshaw, J.; Woolfson, D. N. *J. Struct. Biol.* **2002**, *137*, 73–81.

(18) Moutevelis, E.; Woolfson, D. N. *A Periodic Table of Coiled-Coil Protein Structures*; 2009; Vol. 385.

(19) Oakley, M. G.; Kim, P. S. *Biochemistry* **1998**, *37* (36), 12603–12610.

(20) Apgar, J. R.; Gutwin, K. N.; Keating, A. E. *Proteins Struct. Funct. Genet.* **2008**, *72* (3), 1048–1065.

(21) Crick, F. H. C. *Acta Crystallogr.* **1953**, *6* (8), 689–697.

(22) Huang, P.-S.; Oberdorfer, G.; Xu, C.; Pei, X. Y.; Nannenga, B. L.; Rogers, J. M.; DiMaio, F.; Gonen, T.; Luisi, B.; Baker, D. *Science (80-.).* **2014**, *346* (6208), 481–485.

(23) Joh, N. H.; Wang, T.; Bhate, M. P.; Acharya, R.; Wu, Y.; Grabe, M.; Hong, M.; Grigoryan, G.; DeGrado, W. F. *Science* **2014**, *346* (6216), 1520–1524.

(24) Thomson, A. R.; Thomson, A. R.; Wood, C. W.; Burton, A. J.; Bartlett, G. J.; Sessions, R. B.; Brady, R. L.; Woolfson, D. N. *Science (80-.).* **2014**, *346*, 485–488.

(25) Arnott, S.; Dover, S. D. *J. Mol. Biol.* **1967**, *30* (1), 209–212.

(26) Kükenshöner, T.; Hagemann, U. B.; Wohlwend, D.; Räuber, C.; Baumann, T.; Keller, S.; Einsle, O.; Müller, K. M.; Arndt, K. M. *Biomacromolecules* **2014**, *15* (9), 3296–3305.

(27) Sia, S. K.; Kim, P. S. *Biochemistry* **2001**, *40* (30), 8981–8989.

(28) Hartmann, M. D.; Mendler, C. T.; Bassler, J.; Karamichali, I.; Ridderbusch, O.; Lupas, A. N.; Hernandez Alvarez, B. *Elife* **2016**, *5*.

(29) Hicks, M. R.; Holberton, D. V; Kowalczyk, C.; Woolfson, D. N. *Folding and Design*. 1997, pp 149–158.

(30) Mortenson, D. E.; Steinkruger, J. D.; Kreitler, D. F.; Perroni, D. V; Sorenson, G. P.; Huang, L. *Proc. Natl. Acad. Sci.* **2015**, *112* (43), 13144–13149.

(31) Harbury, P. B.; Zhang, T.; Kim, P. S.; Alber, T. *Science* **1993**, *262* (5138), 1401–1407.

(32) Lyons, J. A.; Shahsavar, A.; Paulsen, P. A.; Pedersen, B. P.; Nissen, P. *Curr. Opin. Struct. Biol.* **2016**, *38*, 137–144.

(33) Trenker, R.; Call, M. E.; Call, M. J. *J. Am. Chem. Soc.* **2015**, *137* (50), 15676–15679.

(34) Knoblich, K.; Park, S.; Lutfi, M.; van ’t Hag, L.; Conn, C. E.; Seabrook, S. A.; Newman, J.; Czabotar, P. E.; Im, W.; Call, M. E.; Call, M. J. *Cell Rep.* **2015**, *11* (8), 1184–1192.

(35) Thomaston, J. L.; Woldeyes, R. A.; Fraser, J. S.; Klein, M. L.; Fiorin, G.; Degrado,

W. F.; Thomaston, J. L.; Alfonso-prieto, M.; Woldeyes, R. A.; Fraser, J. S.; Klein, M. L. *Proc. Natl. Acad. Sci.* **2015**, *112* (18), 14260–14265.

(36) Acharya, R.; Carnevale, V.; Fiorin, G.; Levine, B. G.; Polishchuk, A. L.; Balannik, V.; Samish, I.; Lamb, R. a; Pinto, L. H.; DeGrado, W. F.; Klein, M. L. *Proc. Natl. Acad. Sci. U. S. A.* **2010**, *107* (34), 15075–15080.

(37) Thomaston, J. L.; DeGrado, W. F. *Protein Sci.* **2016**, *25*, 1551–1554.

(38) Cherezov, V. *Curr. Opin. Struct. Biol.* **2011**, *21* (4), 559–566.

(39) Wang, J.; Wu, Y.; Ma, C.; Fiorin, G.; Wang, J.; Pinto, L. H.; Lamb, R. A.; Klein, M. L.; DeGrado, W. F. *Proc. Natl. Acad. Sci.* **2013**, *110* (4), 1315–1320.

(40) Yeates, T. O.; Kent, S. B. H. *Annu. Rev. Biophys.* **2012**, *41*, 41–61.

(41) Ostermeier, C.; Michel, H. *Methods Mol. Biol.* **2013**, *1033*, 697–701.

(42) Dasgupta, S.; Iyer, G. H.; Bryant, S. H.; Lawrence, C. E.; Bell, J. A. *Proteins Struct. Funct. Genet.* **1997**, *28* (4), 494–514.

(43) Schow, E. V.; Freites, J. A.; Cheng, P.; Bernsel, A.; Von Heijne, G.; White, S. H.; Tobias, D. J. *J. Membr. Biol.* **2011**, *239* (1–2), 35–48.

(44) Grigoryan, G.; DeGrado, W. F. *J. Mol. Biol.* **2011**, *405* (4), 1079–1100.

(45) Gernert, K. M.; Surles, M. C.; Labean, T. H.; Richardson, J. S.; Richardson, D. C. *Protein Sci.* **1995**, *4* (11), 2252–2260.

(46) Walshaw, J.; Woolfson, D. N. *J. Mol. Biol.* **2001**, *307* (5), 1427–1450.

(47) Kabsch, W. *Acta Crystallogr. Sect. D Biol. Crystallogr.* **2010**, *66* (2), 125–132.

(48) McCoy, A. J.; Grosse-Kunstleve, R. W.; Adams, P. D.; Winn, M. D.; Storoni, L. C.; Read, R. J. *J. Appl. Crystallogr.* **2007**, *40* (4), 658–674.

(49) Emsley, P.; Cowtan, K. *Acta Crystallogr. Sect. D Biol. Crystallogr.* **2004**, *60* (12 I), 2126–2132.

(50) Adams, P. D.; Afonine, P. V.; Bunkóczi, G.; Chen, V. B.; Davis, I. W.; Echols, N.; Headd, J. J.; Hung, L. W.; Kapral, G. J.; Grosse-Kunstleve, R. W.; McCoy, A. J.; Moriarty, N. W.; Oeffner, R.; Read, R. J.; Richardson, D. C.; Richardson, J. S.; Terwilliger, T. C.; Zwart, P. H. *Acta Crystallogr. Sect. D Biol. Crystallogr.* **2010**, *66* (2), 213–221.

(51) Moriarty, N. W.; Grosse-Kunstleve, R. W.; Adams, P. D. *Acta Crystallogr. Sect. D Biol. Crystallogr.* **2009**, *65* (10), 1074–1080.

Appendix A Benzylic Amide Protecting Group-based Strategies for the Synthesis of Transmembrane Peptides

Acknowledgments:

I am very grateful to Zhihui Yao for acquiring VT-NMR spectra and also to Brian Ferrer for synthesizing Fmoc-(L)-Ala-(Dmb)Gly-OH in addition to his efforts at synthesizing the *P. aeruginosa* Flp pilin N-terminus.

A.1 Synopsis

The transmembrane domains of single-pass integral membrane proteins may play essential roles in signal transduction by mediating higher order associations through specific contacts between membrane-inserted residues.[1,2] Similarly, the transmembrane domains of viral fusion peptides also play a key role in the formation of complex assemblies that drive membrane fusion events between host cell membranes and viral envelopes.[3,4] A further understanding of these interactions that predominantly occur in the membrane will require high resolution atomic models in the form of X-ray crystal structures. A significant obstacle to obtaining crystals of these recalcitrant peptide assemblies is the synthesis and purification of suitable amounts of material. A successful X-ray crystallographic study will require upwards of ~10-50 mg of pure peptide during the course of obtaining an initial lead crystallization condition, optimizing the lead condition, and producing a suitable number of replicate crystals for X-ray diffraction screening.

These hydrophobic transmembrane peptides are prone to aggregation during solid-phase synthesis and/or reverse-phase HPLC purification in organic solvents.[5,6] Beginning in the early 1990s investigators first described the use of N-benzylic amide backbone protecting groups to overcome aggregation of hydrophobic peptides during

solid-phase synthesis (Figure A.1).[5–7] These tertiary amide protecting groups are acid-labile and thus compatible with Fmoc-based solid-phase peptide synthesis. The motivation behind this approach is the notion that aggregation is primarily mediated by inter-strand hydrogen bonding interactions between nascent peptide chains. Installation of a tertiary amide backbone protecting group eliminates a potential hydrogen bond donor and thereby discourages H-bond-mediated aggregation. Early efforts in this area primarily focused on *N,N*-protected glycine, i.e, Fmoc-protected N-2,4-dimethoxybenzylglycine, due to the difficulty associated with the acylation of a somewhat sterically-hindered secondary amine.[5] To circumvent any deletion products that may result from incomplete acylation of a secondary amine, dipeptides of the form Fmoc-Xxx-(Dmb)Gly-OH were introduced.[8] This strategy has been successful enough that an array of Fmoc-Xxx-(Dmb)Gly-OH dipeptides are commercially available from several vendors (Aaptec, Bachem, and EMD Millipore).

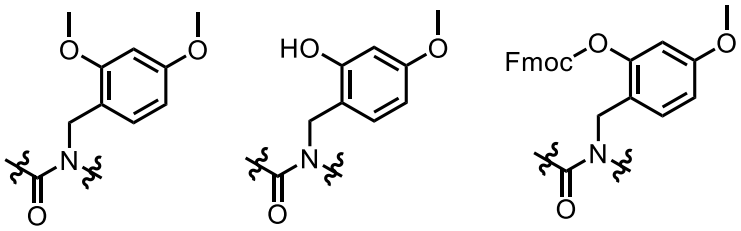


Figure A.1. (left) 2,4-dimethoxybenzyl (Dmb), (center) 2-hydroxy-4-methoxybenzyl, and (right) Fmoc-protected 2-hydroxy-4-methoxybenzyl amide backbone protecting groups.

To extend this strategy to α-amino acids beyond glycine, a process that involves a difficult solution-phase acylation of a Dmb-protected amino acid, I discovered that benzyl esters can be deprotected via Pd-catalyzed hydrogenolysis in the presence of a Dmb-amide (Figure A.2). The benzyl ester protecting group facilitates solubility in organic solvents and silica flash chromatography. This protecting group strategy extends the Dmb

strategy to Dmb-dipeptides that are not commercially available, such as Fmoc-(L)-Leu-(L)-(Dmb)Leu-OH. Furthermore, this approach also facilitates access to Dmb-dipeptides that contain D-amino acids for projects that employ racemic peptide crystallization.

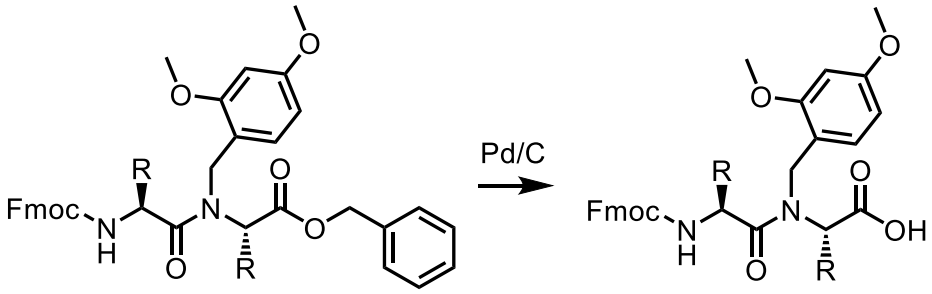


Figure A.2. Selective deprotection of benzyl esters in the presence of Dmb-amides via Pd-catalyzed hydrogenolysis. Presumably the steric hindrance around the Dmb-amide plays a role in the selectivity.

Liu *et al.* has extended the application of these orthogonal benzylic amide protecting groups to include a 2-hydroxy-4-methoxy-3-nitrobenzyl-derived amide protecting group that can be functionalized on resin with a solubility tag, such as Ac-RRRRG.[9,10] These efforts aimed to facilitate the solution-phase segment condensation of unprotected transmembrane peptides via native-chemical ligation, which requires that peptides are soluble in aqueous buffers. A 2-hydroxy-4-methoxy-3-nitrobenzaldehyde (Hmnb) derived protecting group is installed on resin via reductive amination. After completion of the primary sequence and capping with a terminal *N*-Boc-protected or *N*-acetyl amino acid, the nitro group can be reduced with a highly concentrated stannous (II) chloride solution to an aniline that can then be functionalized with an auxiliary tag (Figure A.3). This strategy is extremely useful because it facilitates the downstream purification of these highly aggregation-prone transmembrane peptides.

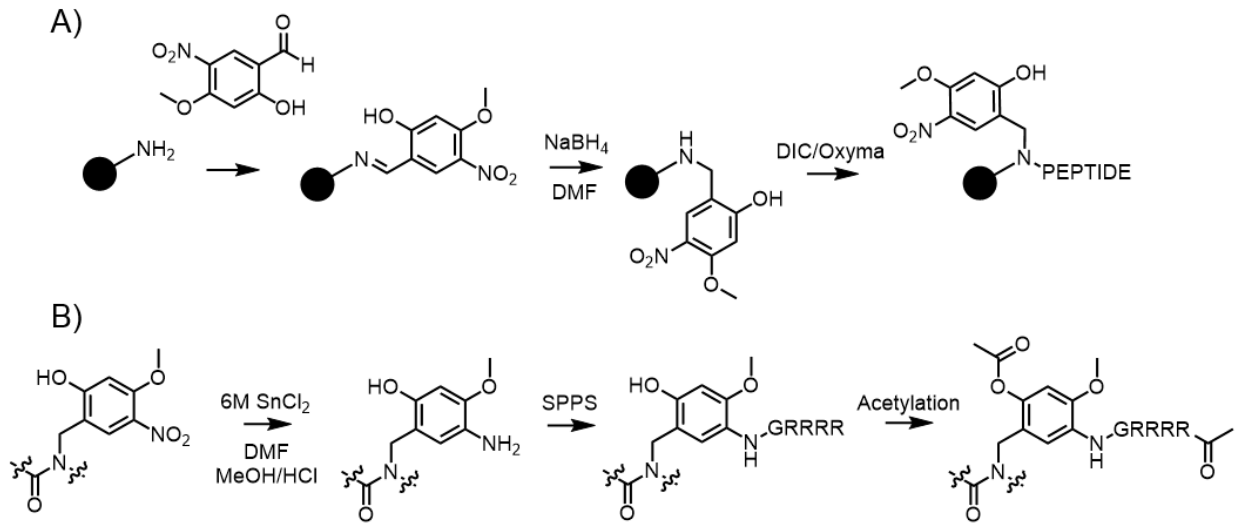
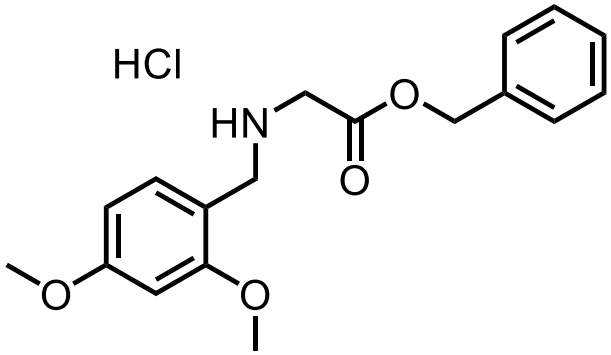


Figure A.3. The functionalizable backbone protecting group strategy reported by Liu et al.[9] A) 2-Hydroxy-4-methoxy-3-nitrobenzaldehyde (Hmnb) is installed on resin via reductive amination. B) After the initial peptide sequence is synthesized the nitro group is reduced to aniline and a solubilizing tag is installed via standard Fmoc-SPPS. Acetylation of the hydroxyl group stabilizes the benzylic amide during cleavage, and after purification the phenyl ester is hydrolyzed and the acid-labile tag is removed via TFA cleavage.

Here I describe the synthesis of several TM peptides that employ Dmb-dipeptides. I have also tabulated my efforts to crystallize these peptides both from detergents and from lipidic cubic phases (Table A.**1**, Table A.**2**). These crystallization experiments were largely unfruitful; however I do report two interesting possible lead crystallization conditions for an EGFR-TM derived peptide both in β-D-octyl glucoside and from the lipid cubic phase. Finally, I describe the successful incorporation of the Liu *et al.* solubilizing tag for the synthesis and purification of Ac-RRRRG-tagged parainfluenza virus 5 (PIV5) transmembrane domain and also the C-terminus of the *Pseudomonas aeruginosa* Flp pilin protein. These syntheses represent preliminary efforts towards the ligation of larger membrane proteins or fusion peptides that may be more amenable to crystallization than isolated single pass TM-domains.

A.2 Synthesis of Dmb-dipeptides



H-(Dmb)Gly-COOH

Formula: $C_{18}H_{21}ClNO_4$

***IUPAC*:** 2-(benzyloxy)-N-(2,4-dimethoxybenzyl)-2-oxoethan-1-aminium chloride

***SMILES*:** O=C(OCC1=CC=CC=C1)C[NH2+]CC2=C(OC)C=C(OC)C=C2.[Cl-]

Procedure: A 1 L round-bottom flask was charged with 2,4-dimethoxybenzaldehyde (8.23 g, 49.6 mmol, 1.0 eq.), glycine benzyl ester hydrochloride (10.0 g, 49.6 mmol, 1.0 eq.), and sodium triacetoxyborohydride (13.7 g, 64.5 mmol, 1.3 eq.). 1,2-Dichloroethane (250 mL) was added to the flask at room temperature. While the contents of the flask were stirring, triethylamine (6.90 mL, 49.6 mmol, 1.0 eq.) was added dropwise, and the solution was stirred overnight. The reaction was quenched with saturated aqueous sodium bicarbonate (150 mL), layers were separated, and the aqueous layer was extracted with DCM (2x100 mL). Organic layers were combined and washed with brine (50 mL), dried over sodium sulfate, and concentrated via rotary evaporation (portion-wise in a 500 mL round-bottom flask) to afford a thick yellow oil. The yellow oil was redissolved in diethyl ether (200 mL; anhydrous) and cooled on ice. HCl gas was generated in a vacuum flask by dripping concentrated H_2SO_4 onto NaCl. HCl gas was transferred to the ethereal amine solution via tygon tubing connected to the

vacuum flask (see Chapter 3, Figure 3.51 for the ammonium hydrochloride salt precipitation apparatus). A yellow oil formed and sank to the bottom of the round bottom flask. The oil was triturated in the ether by gently heating the flask with a steam bath, after which ether was decanted from the flask. The thick oil formed 11.5 g of an off-white foam (66% yield) under hi-vacuum that was scraped out of the flask. After one day the foam powder reverted back to a hard tacky yellow solid that was used directly in the next step.

^{1}H NMR (400 MHz, $CDCl_3$): δ = 9.57 (s, 2H), 7.46-7.32 (m, 6H), 6.65-6.51 (m, 2H), 5.22 (s, 2H), 4.09 (s, 2H), 3.90 (s, 2H), 3.78 (s, 6H),

ESI-MS [M+H]: *calculated* 316.1543, *observed* 316.1539

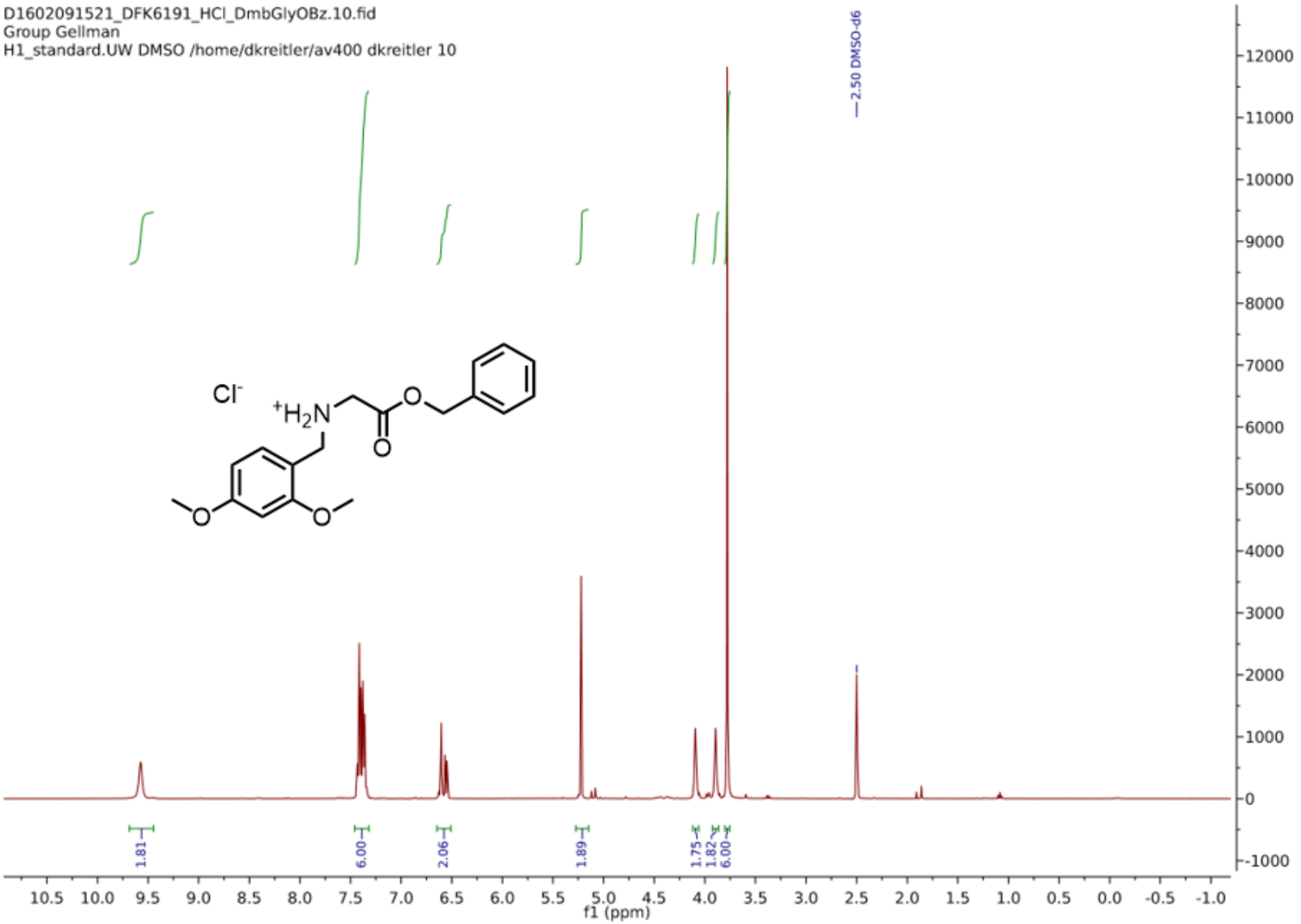
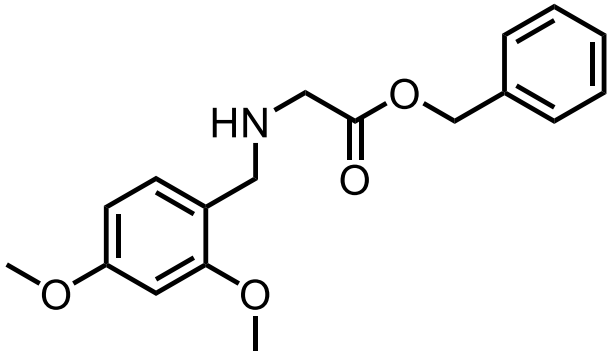


Figure A.4. ^{1}H-NMR spectrum of HCl H-(Dmb)-Gly-OH**.**



H-(Dmb)Gly-COOH

Formula: $C_{18}H_{21}NO_4$

***IUPAC*:** benzyl 2-{[(2,4-dimethoxyphenoxyphenyl)methyl]amino}acetate

***SMILES*:** COc2ccc(CNCC(=O)OCc1ccccc1)c(OC)c2

Procedure: *This procedure employs an alternative purification strategy for H-(Dmb)Gly-OH (silica flash column versus ammonium salt precipitation).* A 250 mL round-bottom flask was charged with 2,4-dimethoxybenzaldehyde (0.87 g, 5.2 mmol, 1.05 eq.), glycine benzyl ester hydrochloride (1.0 g, 5.0 mmol, 1.0 eq.), and sodium triacetoxyborohydride (1.6 g, 7.4 mmol, 1.5 eq.). 1,2-Dichloroethane (50 mL) was added to the flask at room temperature. While the contents of the flask were stirring, triethylamine (0.86 mL, 5.2 mmol, 1.05 eq.) was added dropwise, and the solution was stirred overnight. The reaction was quenched with saturated aqueous sodium bicarbonate (50 mL), layers were separated, and the aqueous layer was extracted with DCM (50 mL). Organic layers were combined and washed with brine (15 mL), dried over sodium sulfate, and concentrated via rotary evaporation (portion-wise in a 500 mL round-bottom flask) to afford a thick yellow oil, which was then purified via silica flash chromatography (R_f=0.2-0.4, 4% v/v MeOH in DCM; DCM was washed with aqueous ammonium hydroxide in a 2 L separatory funnel) to afford 1.3 g of a brown oil (85% yield).

^{1}H NMR (400 MHz, $CDCl_3$): δ = 7.39 (m, 5H), 7.10 (d, *J*=8.0 Hz, 1H), 6.46-6.37 (m, 2H), 5.14 (s, 2H), 3.82-3.77 (m, 6H), 3.76-3.70 (m, 2H), 3.43 (s, 2H), 2.13-1.90 (m, 2H)

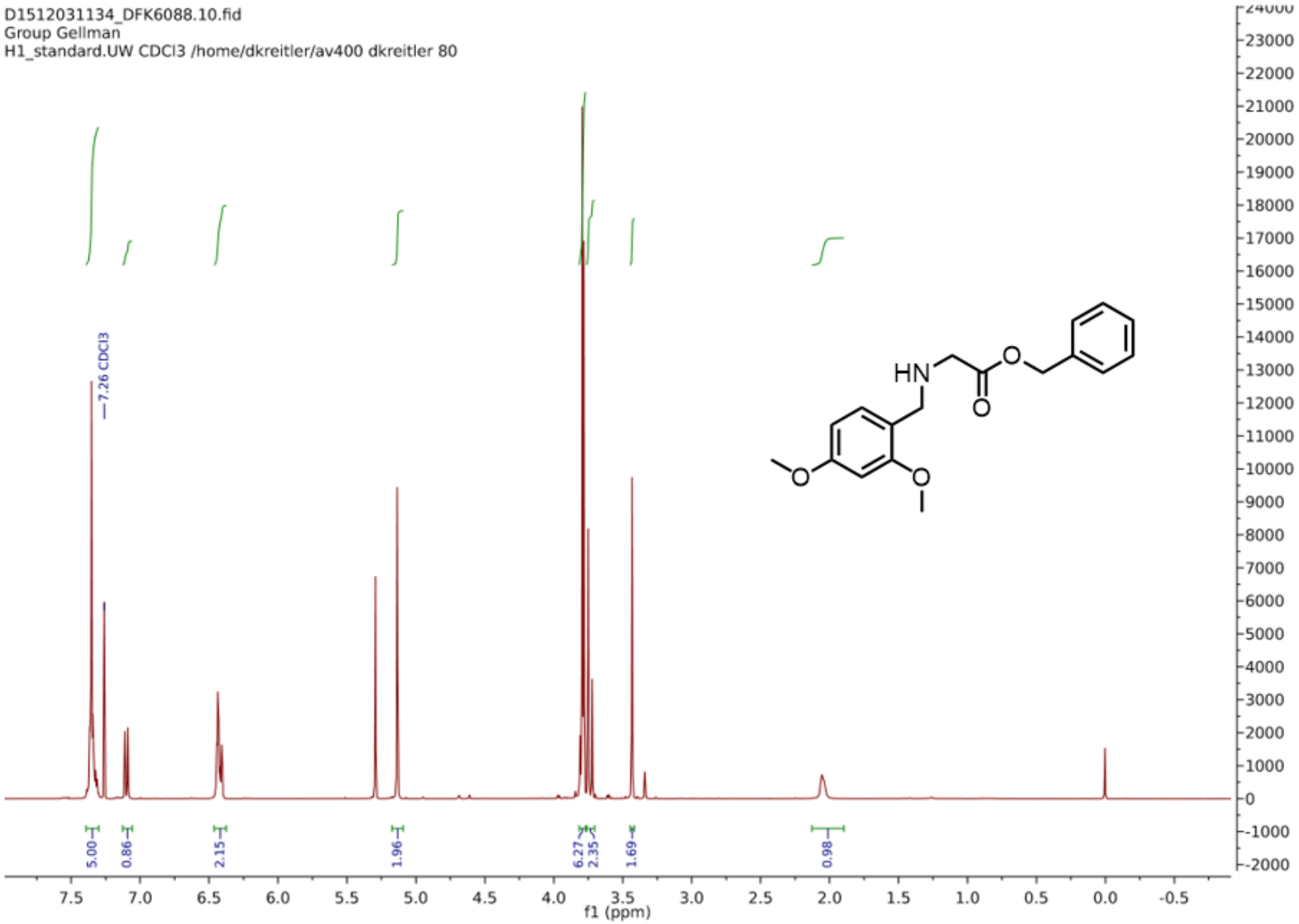
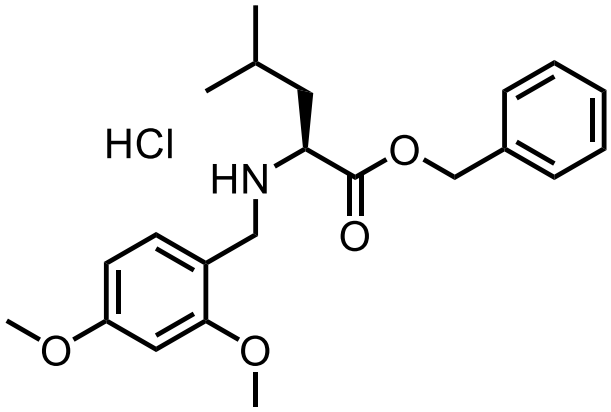


Figure A.5. 1H-NMR spectrum of H-(Dmb)Gly-OH.



H-(L)-(Dmb)Leu-OH

Formula: $C_{22}H_{29}NO_4$

***IUPAC*:** (S)-1-(benzyloxy)-N-(2,4-dimethoxybenzyl)-4-methyl-1-oxopentan-2-aminium chloride

***SMILES*:** O=C(OCC1=CC=CC=C1)[C@H](CC(C)C)[NH2+]CC2=C(OC)C=C(OC)C=C2.[Cl-]

Procedure: A 500 mL round-bottom flask was charged with TsOH·H-(L)-Leu-OBzl (10.0 g, 25.4 mmol, 1.0 eq.), 2,4-dimethoxybenzaldehyde (4.22 g, 26.7 mmol, 1.05 eq.), and sodium triacetoxyborohydride (6.46 g, 30.5 mmol, 1.2 eq.), which were subsequently dissolved in 2,2-dichloroethane (250 mL). Triethylamine (3.54 mL, 25.4 mmol, 1.0 eq.) was added to the stirring solution via pipet, and the solution was stirred at room temperature for 16 h. DCE was then removed via rotary evaporation. Saturated aqueous sodium bicarbonate (200 mL) was added to the flask, and the heterogeneous mixture was extracted with DCM (2x100 mL). Organic layers were combined, washed with brine (50 mL), dried over sodium sulfate, and concentrated via rotary evaporation. The resulting crude oil was dissolved in diethyl ether (anhydrous; 150 mL) in a 250 mL round-bottom flask. The hydrochloride ammonium salt was precipitated from the ethereal solution by passing gaseous HCl (generated with H_2SO_4 and NaCl) over the ethereal, amine-

containing solution, which generated a colloidal suspension. HCl gas production was ceased, and the colloidal suspension was heated with a steam bath until boiling, after which, 5 mL of acetonitrile was added to dissolve any remaining precipitate. The flask was then cooled to -20°C for 2 h, after which, a white crystalline solid had formed. Then, approximately half of the remaining ether was removed via gentle rotary evaporation (no heat bath), and the enthalpy of vaporization was exploited to further cool the flask. Following this concentration, ether was decanted and the white solid was filtered and washed with cold ether (anhydrous; 2x20 mL) to afford 8.7 g of a white solid (85% yield).

^{1}H NMR (400 MHz, MeOD): δ = 7.44-7.34 (m, 5H), 7.26-7.20 (m, 1H), 6.63-6.49 (m, 2H), 5.31-5.15 (m, 2H), 4.24-4.12 (m, 2H), 3.92 (dd, *J*=9.0, 5.5 Hz, 1H), 3.84 (s, 3H), 3.81 (s, 3H), 1.85-1.68 (m, 2H), 1.63 (dh, *J*=8.6, 6.3 Hz), 0.90 (*J*=17.8, 6.4 Hz, 6H)

^{13}C NMR (400 MHz, MeOD): δ = 169.91, 164.36, 136.18, 134.12, 130.02, 129.94, 129.75, 111.57, 111.57, 106.30, 99.39, 69.28, 58.79, 56.16, 56.01, 46.67, 40.00, 25.87, 23.09, 25.87, 23.09, 21.71

ESI-MS [M+H]: *calculated* 371.2097, *observed* 371.2162

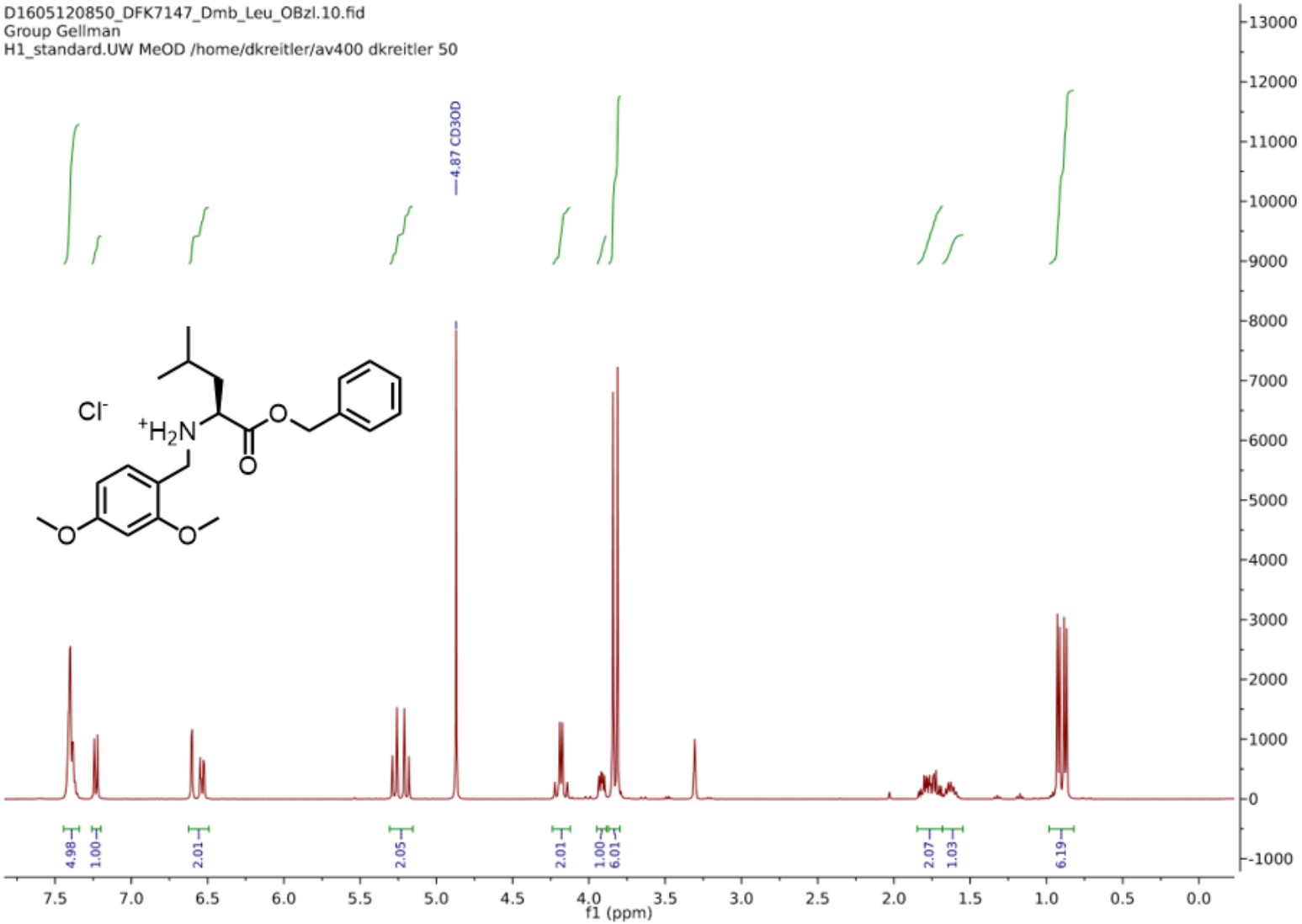


Figure A.6. ^{1}H-NMR spectrum of HCl H-(L)-(Dmb)Leu-OH.

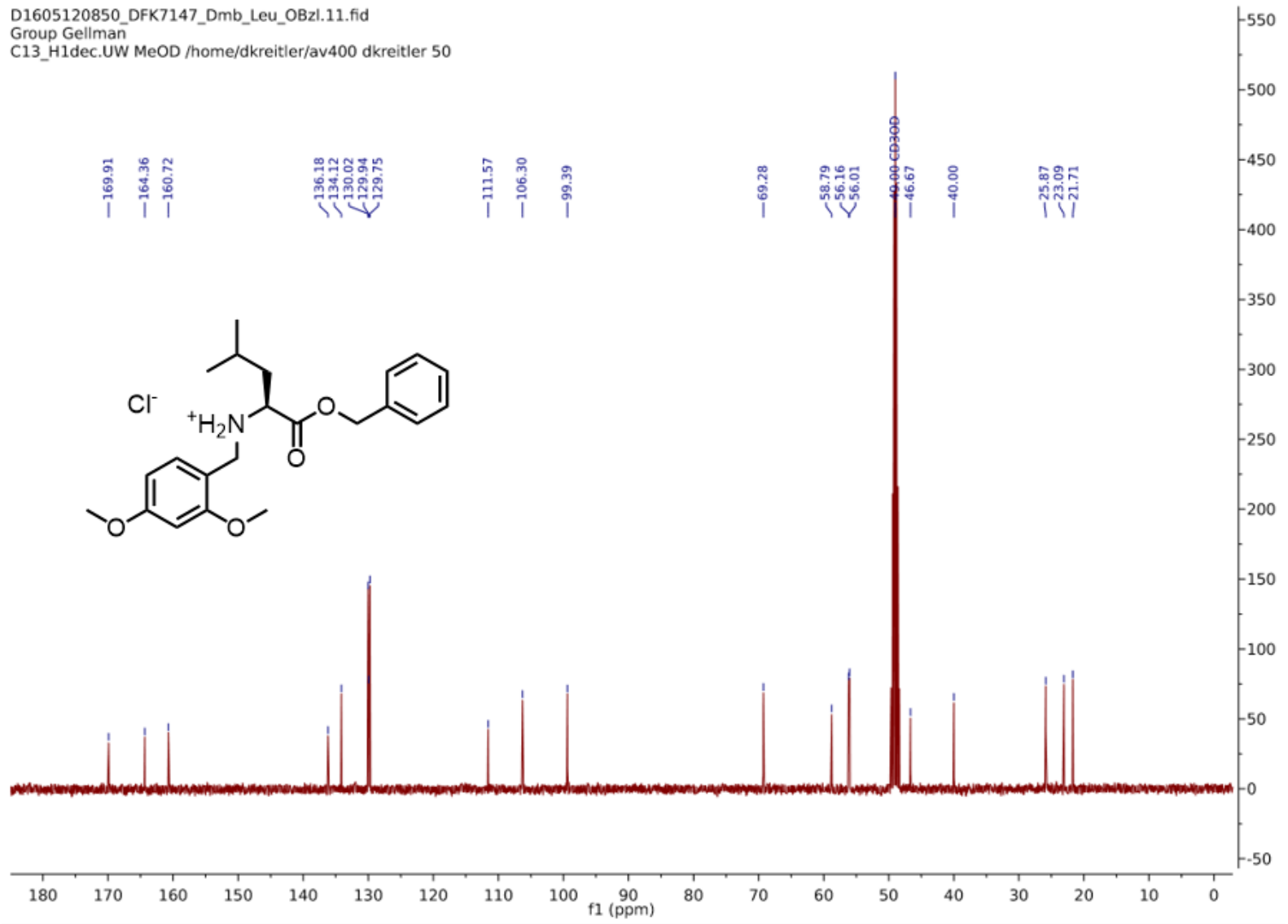
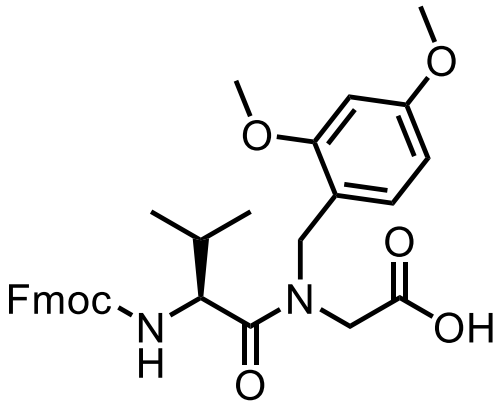


Figure A.7. ^{13}C-NMR spectrum of HCl H-(L)-(Dmb)Leu-OH.



Fmoc-(L)-Val-(Dmb)Gly-OH

Formula: $C_{18}H_{21}NO_4$

***IUPAC*:** 2-[(2S)-N-[(2,4-dimethoxyphenyl)methyl]-2-({[(9H-fluoren-9-yl)methoxy]carbonyl}amino)-3-methylbutanamido]acetic acid

***SMILES*:** COc4ccc(CN(CC(=O)O)C(=O)[C@@H](NC(=O)OCC3c1ccccc1c2ccccc23)C(C)C)c(OC)c4

Procedure: A 250 mL round-bottom flask was charged with Fmoc-(L)-Val-OH (5.2 g, 15.2 mmol, 1.0 eq.) and HATU (5.75 g, 15.2 mmol, 1.0 eq.), which were subsequently dissolved in DMF (150 mL). *N,N*-diisopropylethylamine (5.29 mL, 30.4 mmol, 2.0 eq.) was added to the stirring solution via pipet, and the solution was stirred at room temperature for 5 min. H-(Dmb)Gly-OBzl oil was then dissolved in DMF (10 mL) and added dropwise to the stirring solution. The solution was stirred overnight at room temperature. The solution was then diluted to 500 mL with water (de-ionized). This dilution resulted in a heterogeneous mixture that was extracted with EtOAc (2x500 mL). The organic layers were combined and then washed with: 5% w/v aqueous LiCl (2x250 mL); 10% w/v aqueous citric acid (1x250 mL); and brine (50 mL). Following these washings, the organic layer was dried over $MgSO_4$, filtered (Whatman filter paper), and concentrated via rotary evaporation to afford a crude yellow oil that was then dissolved in

MeOH (100 mL) in a 500 mL round-bottom flask. Pd/C (10% w/w, 640 mg) was subsequently added to the flask. The atmosphere in the flask was then evacuated and replaced with N_2, after which an H_2 balloon was attached to the flask. The heterogeneous mixture was vigorously stirred for 4 h, after which the H_2 balloon was removed, and the mixture was filtered over a celite pad to remove Pd/C. MeOH was evaporated to afford a yellow oil that was purified via silica chromatography (60% v/v EtOAc in hexanes, 1% v/v acetic acid) to afford a pale yellow oil that was subsequently dissolved in toluene (15 mL), which was then removed via rotary evaporation. This toluene dissolution/evaporation procedure was repeated five times to remove residual acetic acid. Finally, the pale yellow oil foamed into a white solid under hi-vacuum. The foam was scraped out of the flask and ground into 3.8 g of a fine white powder (46% yield over the two-steps).

^{1}H NMR (500 MHz, DMSO-d6; 80°C): δ = 12.72 (s, 1H), 8.33-8.23 (m, 2H), 8.20-8.05 (m, 2H), 7.88-7.79 (m, 2H), 7.78-7.65 (m, 3H), 7.63-7.37 (m, 2H), 7.01-6.76 (m, 2H), 5.06-4.85 (m, 2H), 4.77-4.51 (m, 4H), 4.48-4.29 (m, 1H), 4.22-4.02 (m, 6H), 1.70-1.46 (m, 1H), 1.38-1.20 (m, 6H)

^{13}C NMR (500 MHz, DMSO-d6; 80°C): δ = 171.48, 169.58, 160.16, 158.12, 143.56, 143.50, 140.39, 129.58, 128.46, 127.75, 127.18, 126.61, 124.86, 119.57, 116.18, 104.61, 98.41, 65.55, 55.38, 55.11, 54.89, 46.53, 46.11, 45.98, 30.22, 18.84, 17.26

ESI-MS [M+H]: *calculated* 547.2366, *observed* 547.243

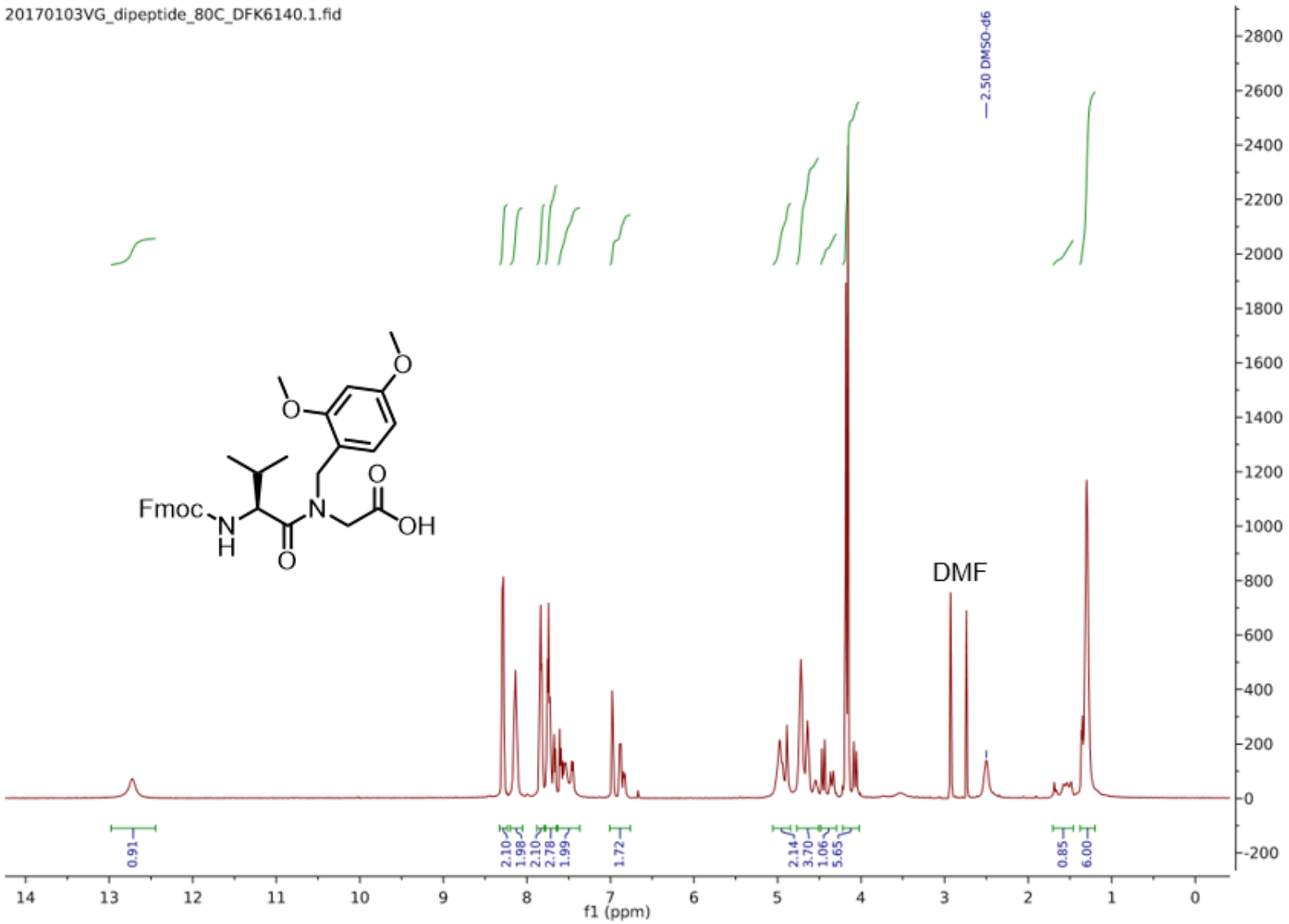


Figure A.8. ^{1}H-NMR spectrum (80°C) of Fmoc-(L)-Val-(Dmb)Gly-OH.

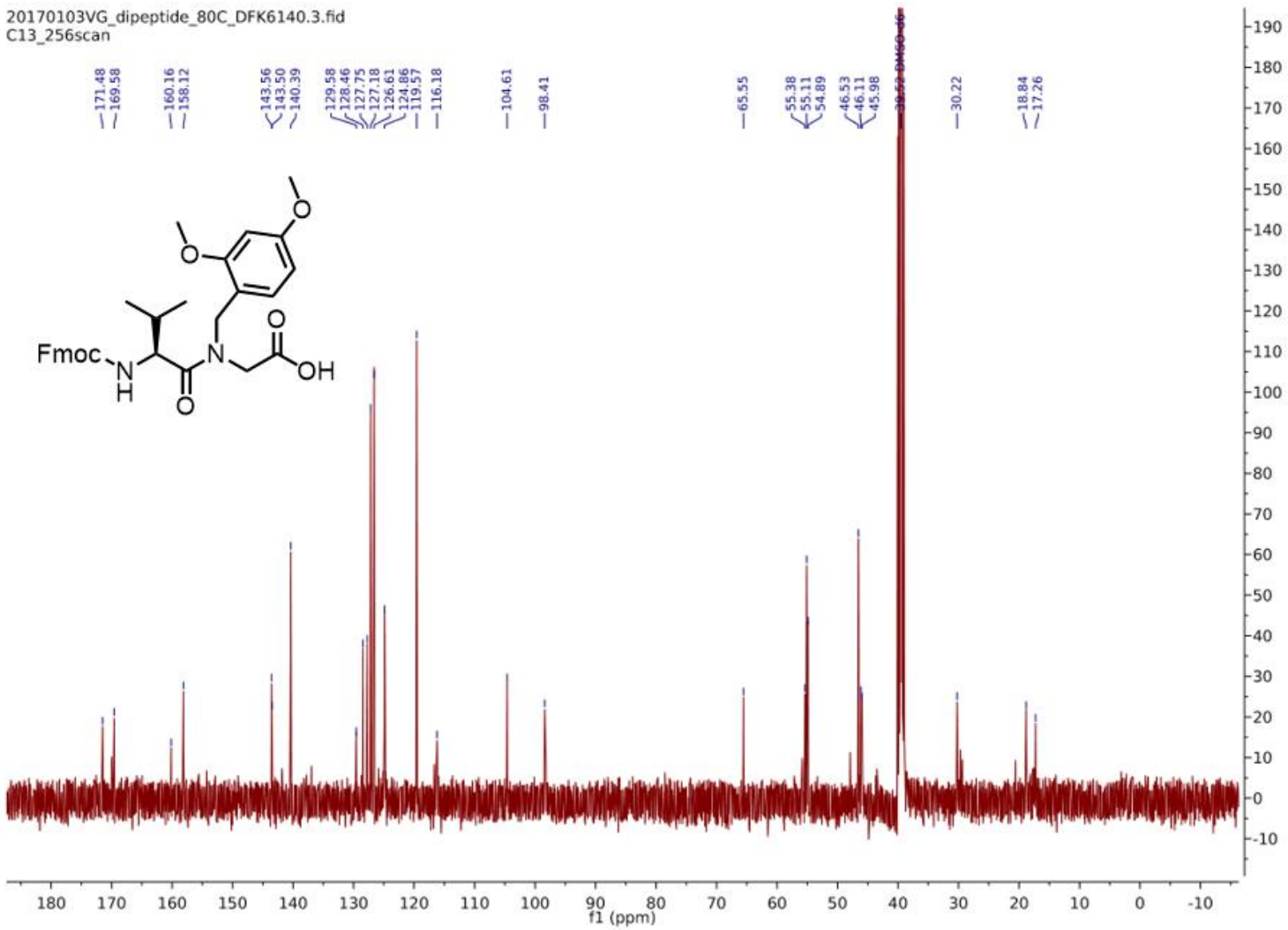
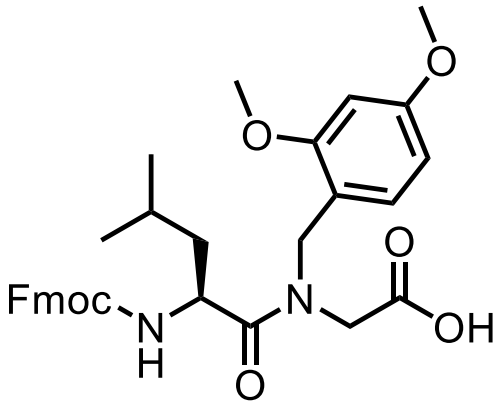


Figure A.9. ^{13}C-NMR spectrum (80°C) of Fmoc-(L)-Val-(Dmb)Gly-OH.



Fmoc-(L)-Leu-(Dmb)Gly-OH

Formula: $C_{36}H_{44}N_2O_7$

***IUPAC*:** *N*-((((9H-fluoren-9-yl)methoxy)carbonyl)-L-leucyl)-*N*-(2,4-dimethoxybenzyl)glycine

***SMILES*:**

O=C(N[C@@H](CC(C)C)C(N(CC1=C(OC)C=C(OC)C=C1)CC(O)=O)=O)OCC2C3=CC=CC=C3C4=CC=CC=C42

Procedure: A 250 mL round-bottom flask was charged with Fmoc-(L)-Leu-OH (5.36 g, 15.2 mmol, 1.0 eq.) and HATU (5.76 g, 15.2 mmol, 1.0 eq.), which were dissolved in DMF (150 mL). *N,N*-Diisopropylethylamine (5.28 mL, 30.4 eq., 2.0 eq.) was added to the stirring solution via pipet, and the solution was stirred for 5 min at room temperature. H-(Dmb)Gly-OH (5.0 g, 15.2 mmol, 1.0 eq.) was added to the stirring solution, and the flask was heated to 40°C and stirred for 16 h. The solution was then diluted to 500 mL with water (de-ionized) and extracted with EtOAc (2x250 mL). Organic layers were combined and washed with: 5% w/v aqueous LiCl (2x250 mL), 10% w/v aqueous citric acid (2x250 mL), and brine (50 mL) in the order listed. The organic solution was then dried over $MgSO_4$ and concentrated to afford a crude brown oil that foamed into a solid

under vacuum. The solid reverted back to an oil when removed from vacuum. The crude oil was then purified via silica flash chromatography (R_f=0.2; 30% v/v EtOAc in hexanes). Product-containing fractions were concentrated to afford a pale yellow oil that foamed into a white solid under vacuum. The oil was then dissolved in MeOH (90 mL) in a 500 mL round-bottom flask. Pd/C (10% w/w, 580 mg) was added to the flask, and the atmosphere in the flask was evacuated and replaced with N_2. The flask was then outfitted with an H_2 balloon and stirred vigorously for 3.5 h. The heterogeneous solution was then filtered over a celite pad, and the resulting solution was concentrated to afford a yellow oil that was purified via silica flash chromatography (50% v/v EtOAc in hexanes with 1% v/v acetic acid). Product-containing fractions were combined and concentrated to afford a yellow oil that was subsequently dissolved in toluene (20 mL) and concentrated via rotary evaporation (bath temperature, 60°C). This procedure was repeated five times (to remove residual acetic acid) to afford a yellow oil that foamed into a white solid. The solid was then ground into a powder to afford 3.0 g of a white solid (37% yield).

^{1}H NMR (500 MHz, DMSO-d6; 80°C): δ = 12.36 (s, 1H), 7.93-7.80 (m, 2H), 7.78-7.64 (m, 2H), 7.47-7.37 (m, 2H), 7.37-7.28 (m, 2H), 7.28-7.00 (m, 2H), 6.65-6.20 (m, 2H), 4.82-3.96 (m, 6H), 3.90-3.52 (m, 6H), 1.73-1.19 (m, 4H), 0.98-0.73 (m, 6H)

^{13}C NMR (500 MHz, DMSO-d6; 80°C): δ = 172.27, 169.65, 160.16, 159.75, 158.05, 143.59, 143.48, 140.42, 140.40, 129.54, 129.22, 128.47, 127.76, 127.74, 127.18, 127.16, 126.62, 126.60, 126.57, 124.85, 119.59, 119.57, 116.10, 104.65, 98.43, 98.23, 65.46, 55.14, 55.12, 54.90, 49.03, 47.98, 46.58, 46.55, 46.32, 45.78, 43.90, 40.61, 30.83, 30.81, 27.86, 23.88, 23.77, 22.70, 21.62, 21.60, 20.97, 13.38

ESI-MS [M-H]: *calculated* 561.2595, *observed* 561.2594

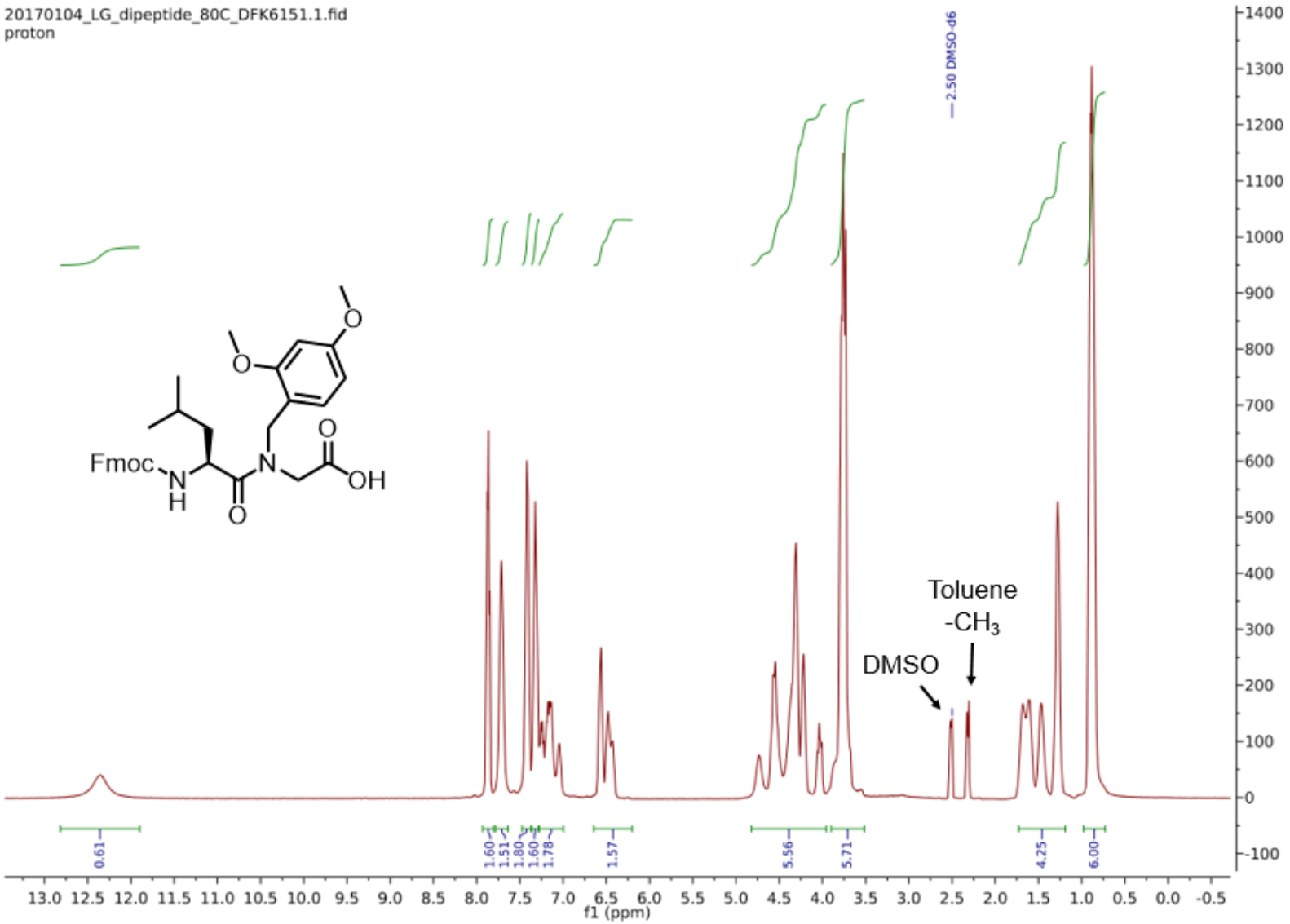


Figure A.10. ^{1}H-NMR spectrum (80°C) of Fmoc-(L)-Leu-(Dmb)Gly-OH.

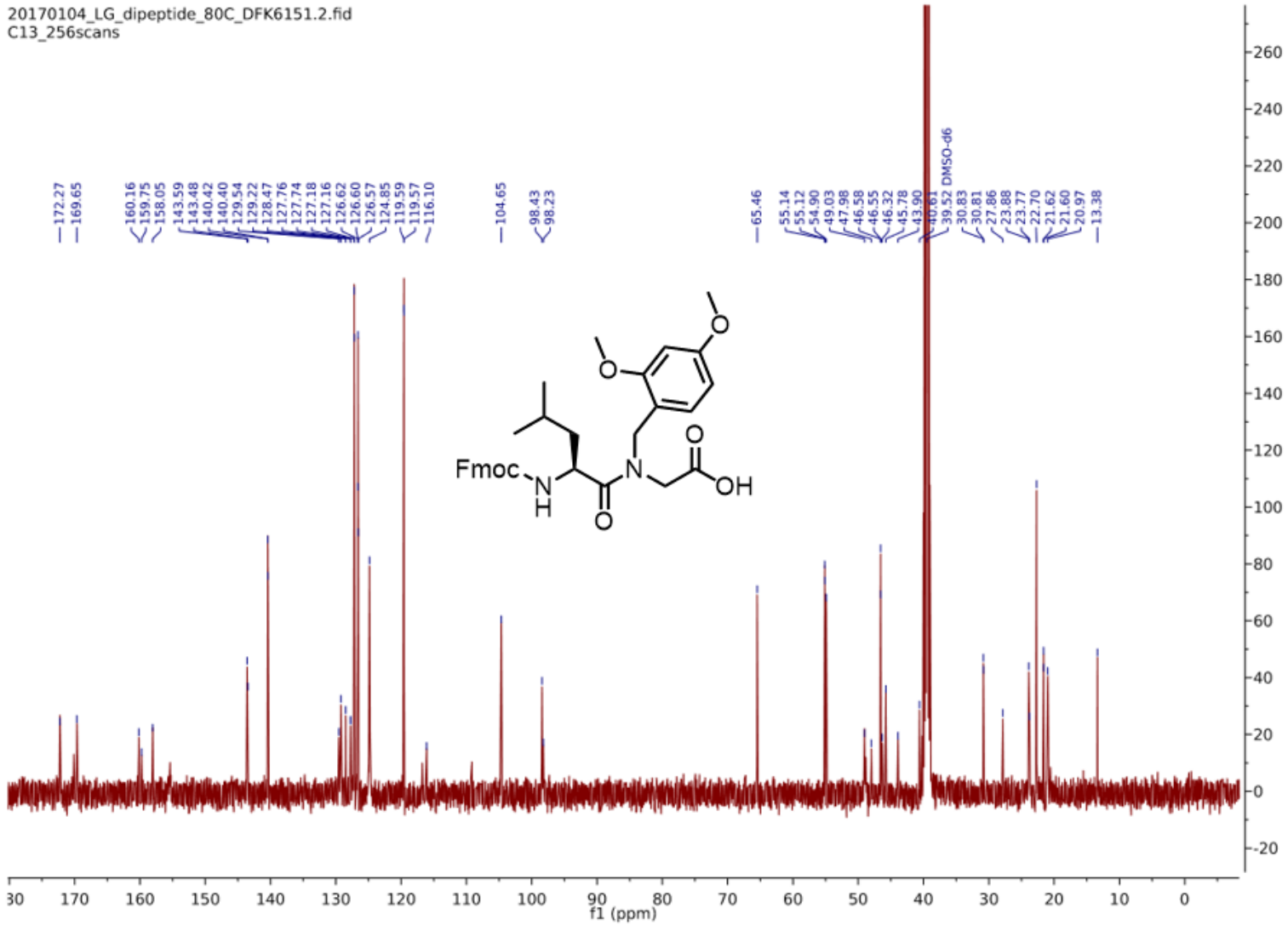
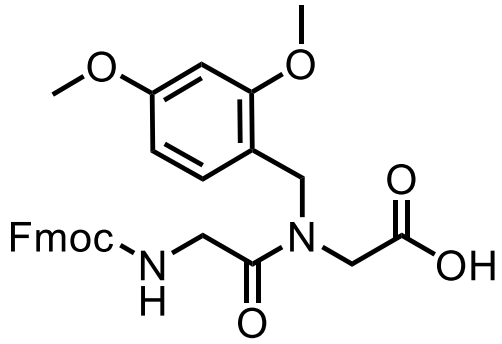


Figure A.11. ^{13}C-NMR spectrum (80°C) of Fmoc-(L)-Leu-(Dmb)Gly-OH.



Fmoc-Gly-(Dmb)Gly-OH

Formula: $C_{28}H_{28}N_2O_7$

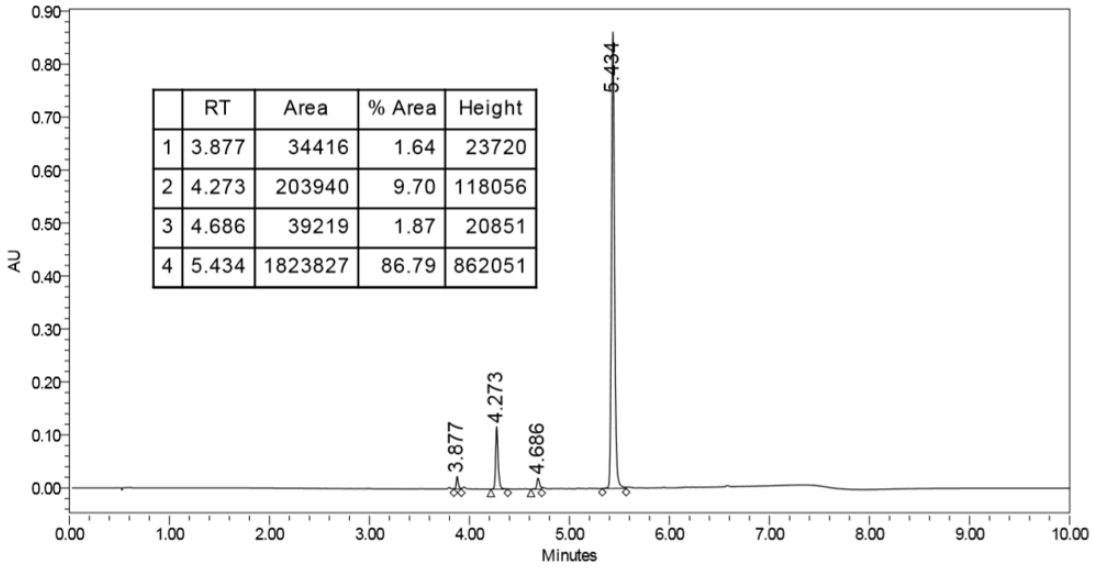
***IUPAC*:** N-((((9H-fluoren-9-yl)methoxy)carbonyl)glycyl)-N-(2,4-dimethoxybenzyl)glycine

***SMILES*:**

O=C(NCC(N(CC1=C(OC)C=C(OC)C=C1)CC(O)=O)=O)OCC2C3=CC=CC=C3C4=CC=CC=C42

Procedure: A 250 mL round-bottom flask was charged with Oxyma (2.02 g, 14.2 mmol, 1.0 eq.) and Fmoc-Gly-OH (4.23 g, 14.2 mmol, 1.0 eq.), which were then dissolved in DMF (35 mL). *N,N'*-Diisopropylcarbodiimide (DIC; 2.2 mL, 14.2 mmol, 1.0 eq.) was added to the solution via pipet. The solution was then stirred at room temperature for 15 minutes (to generate the active ester). In a separate 50 mL conical tube, the hydrochloride salt of H-(Dmb)Gly-OBzl was dissolved in DMF (35 mL) with DIEA (2.48 mL, 14.2 mmol, 1.0 eq.). The H-(Dmb)Gly-OBzl solution was added to the original Fmoc-Gly-OH/Oxyma solution dropwise via pipet, and the resulting solution was heated to 60°C and stirred for 2.5 h. The solution was then diluted to 250 mL with water (de-ionized) and extracted with EtOAc (2x150 mL). Organic layers were combined and washed with: 5% w/v aqueous LiCl (3x50 mL); 10% w/v aqueous citric acid (3x50 mL); aqueous saturated sodium bicarbonate (12x50 mL; until the yellow hue due to Oxyma was removed); and washed with brine (50 mL). The solution was then dried over sodium

sulfate and concentrated via rotary evaporation to afford 6.3 g of a crude yellow oil that foamed into an off-white solid under vacuum. The crude oil was dissolved in MeOH (100 mL). Pd/C (10% w/w, 600 mg) was added, and the atmosphere of the flask was evacuated and replaced with N_2. The flask was then outfitted with an H_2 balloon, and the solution was stirred vigorously for 3 h. After this time the balloon was removed, and the solution was filtered over a celite pad to remove residual Pd/C. The celite pad was rinsed with an additional portion of MeOH (100 mL), and the organic fractions were concentrated via rotary evaporation to afford a yellow oil. The oil was dissolved in EtOAc (40 mL) and precipitated via dropwise addition of the EtOAc solution to a constantly stirred heptane solution. The precipitate-laden flask was then cooled to -20°C for 2 h, after which the precipitate was filtered to afford 4.7 g of an off-white solid (65% yield).



	RT	Area	% Area	Height
1	3.877	34416	1.64	23720
2	4.273	203940	9.70	118056
3	4.686	39219	1.87	20851
4	5.434	1823827	86.79	862051

Figure A.12. UPLC purity analysis of Fmoc-Gly-(Dmb)Gly-OH, 10-95%B over 6 min (A: H_2O, 0.1% v/v TFA, B: ACN, 0.1% v/v TFA; column: Waters Acquity CSH C18 1.7 µm, 2.1x100 mm; flow rate: 0.25 mL/min; monitored at 254 nm; purity 86.8%).

^{1}H NMR (500 MHz, DMSO-d6; 80°C): δ = 7.91-7.78 (m, 2H), 7.75-7.64 (m, 2H), 7.45-7.37 (m, 2H), 7.36-7.26 (m, 2H), 7.12-7.02 (m, 1H), 6.67-6.38 (m, 2H), 5.37 (s, 1H), 4.50-4.36

(m, 2H), 4.32-4.19 (m, 3H), 4.09-4.02 (m, 1H), 3.95-3.84 (m, 2H), 3.82-3.73 (m, 6H), 3.71-3.61 (m, 1H), 1.05-1.01 (m, 3H)

^{13}C NMR (500 MHz, DMSO-d6; 80°C): δ = 158.13, 156.56, 143.58, 140.38, 139.15, 137.18, 129.77, 128.48, 127.17, 126.83, 120.87, 119.59, 119.57, 119.51, 108.78, 104.67, 98.55, 65.59, 55.15, 54.92, 46.51, 40.41, 30.80, 22.8

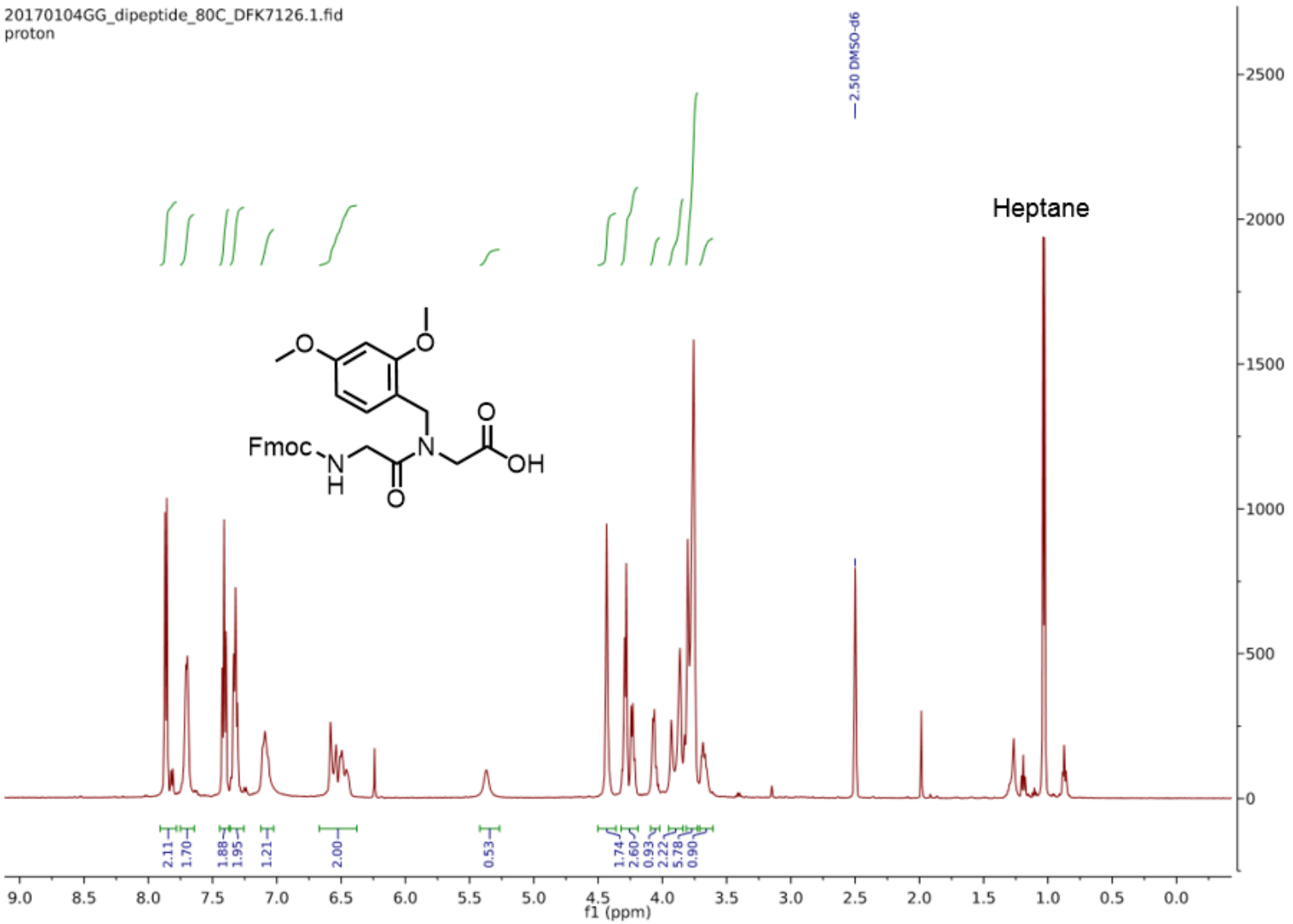


Figure A.13. ^{1}H-NMR spectrum (80°C) of Fmoc-Gly-(Dmb)Gly-OH.

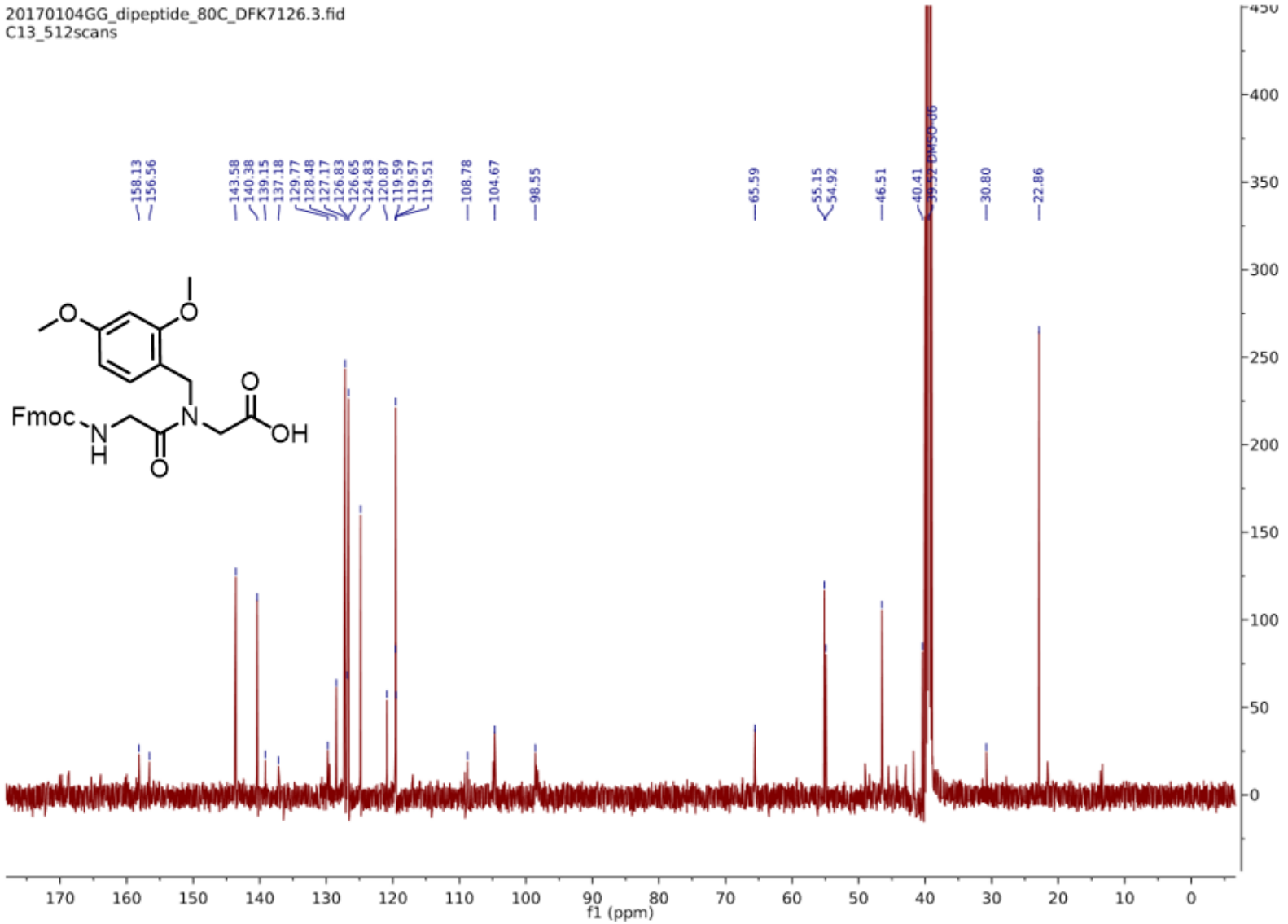
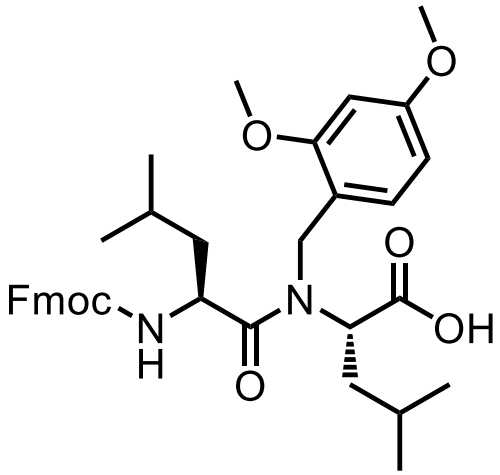


Figure A.14. ^{13}C-NMR spectrum (80°C) of Fmoc-Gly-(Dmb)Gly-OH.



Fmoc-(L)-Leu-(L)-(Dmb)Leu-OH

Formula: $C_{22}H_{29}NO_4$

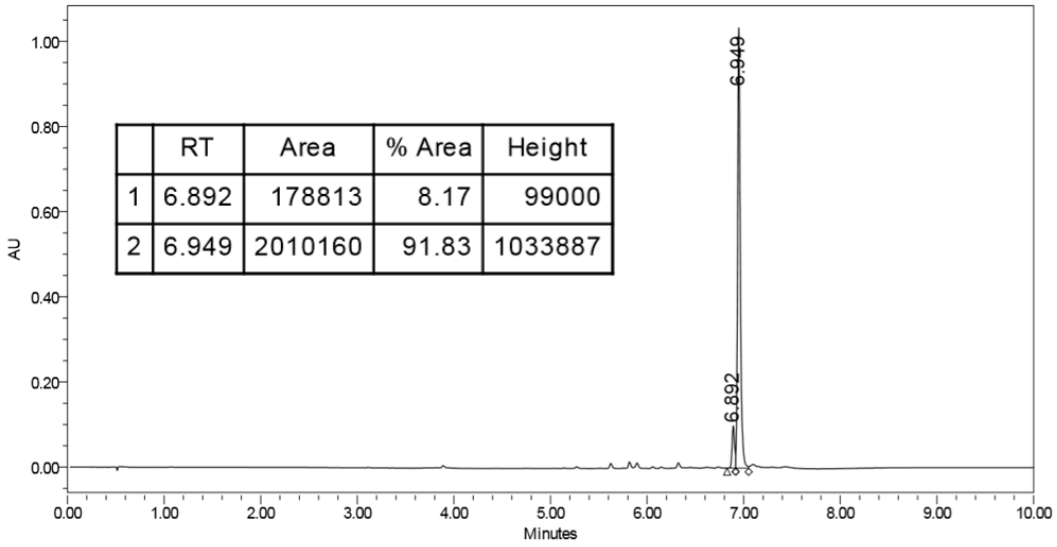
***IUPAC*:** N-((((9H-fluoren-9-yl)methoxy)carbonyl)-L-leucyl)-N-(2,4-dimethoxybenzyl)-L-leucine

***SMILES*:**

O=C(N[C@@H](CC(C)C)C(N(CC1=C(OC)C=C(OC)C=C1)[C@@H](CC(C)C)C(O)=O)=O)OCC2C3=CC=CC=C3C4=CC=CC=C42

Procedure: A 250 mL round-bottom flask was charged with Fmoc-(L)-Leu-OH (6.06 g, 17.2 mmol, 1.0 eq.) and Oxyma (2.44 g, 17.2 mmol, 1.0 eq.), which were dissolved in DMF (40 mL). *N,N'*-Diisopropylcarbodiimide (2.66 mL, 17.2 mmol, 1.0 eq.) was added via pipet to the solution, which was stirred at room temperature for 15 min. In a separate 50 mL conical tube, H-(L)-(Dmb)Leu-OH (7.0 g, 17.2 mmol, 1.0 eq.) was dissolved in DMF (40 mL) with *N,N*-diisopropylethylamine (DIEA; 3.00 mL, 17.2 mmol, 1.0 eq.). The amine/base solution was added dropwise via pipet to the initial flask, the resulting solution was heated to 60°C and stirred for 3.5 h. The reaction vessel was then cooled to room temperature and diluted to 250 mL with water (de-ionized). The heterogeneous solution was extracted with EtOAc (2x150 mL), and organic layers were combined and washed

with: 5% w/v aqueous LiCl (3x50 mL), 10% w/v aqueous citric acid (3x50 mL), saturated aqueous sodium bicarbonate (10x50 mL), and brine (50 mL). The organic phase was then dried over sodium sulfate and concentrated to afford a pale yellow oil that was purified via silica flash chromatography (R_f=0.25, 254 nm; 30% v/v EtOAc in hexanes). Product-containing fractions were concentrated and dissolved in MeOH (120 mL) in a 500 mL round-bottom flask. Pd/C (10% w/w, 0.8 g) was added, and the atmosphere in the flask was evacuated and replaced with N_2. The flask was then outfitted with an H_2 balloon, and the mixture was stirred vigorously for 4 h, after which the heterogeneous mixture was filtered over a celite pad and concentrated to afford a yellow oil that was purified via silica flash chromatography (4% v/v MeOH in DCM). Product-containing fractions were concentrated to afford a pale yellow oil that foamed into a white solid under vacuum. The white foam was then ground into a fine powder to afford 2.8 g of a white solid (25% yield) that was 91.8% pure by UPLC (Figure A.15).



	RT	Area	% Area	Height
1	6.892	178813	8.17	99000
2	6.949	2010160	91.83	1033887

Figure A.15. UPLC purity analysis of Fmoc-(L)-Leu-(L)-(Dmb)Leu-OH, 10-95%B over 6 min (A: H_2O, 0.1% v/v TFA, B: ACN, 0.1% v/v TFA; column: Waters Acquity CSH C18 1.7 µm, 2.1x100 mm; flow rate: 0.25 mL/min; monitored at 254 nm; purity 91.8%).

^{1}H NMR (500 MHz, DMSO-d6; 80°C): δ = 12.12 (s, 1H), 7.92-7.79 (m, 2H), 7.77-7.63 (m, 2H), 7.44-7.37 (m, 2H), 7.36-7.26 (m, 2H), 7.23-7.06 (m, 1H), 6.59-6.41 (m, 2H), 4.77-4.13 (m, 6H), 3.78 (s, 3H), 3.72 (s, 3H), 1.78-1.18 (m, 6H), 1.02-0.48 (m, 12H)

^{13}C NMR (500 MHz, DMSO-d6; 80°C): δ = 172.83, 171.90, 143.45, 140.37, 128.80, 128.47, 127.73, 127.16, 126.82, 126.59, 126.56, 124.82, 120.86, 119.56, 104.44, 98.21, 65.41, 55.11, 54.91, 46.53, 24.33, 23.80, 22.71, 21.98, 21.71, 20.69

ESI-MS [M-H]: *calculated* 615.3149, *observed* 615.3073

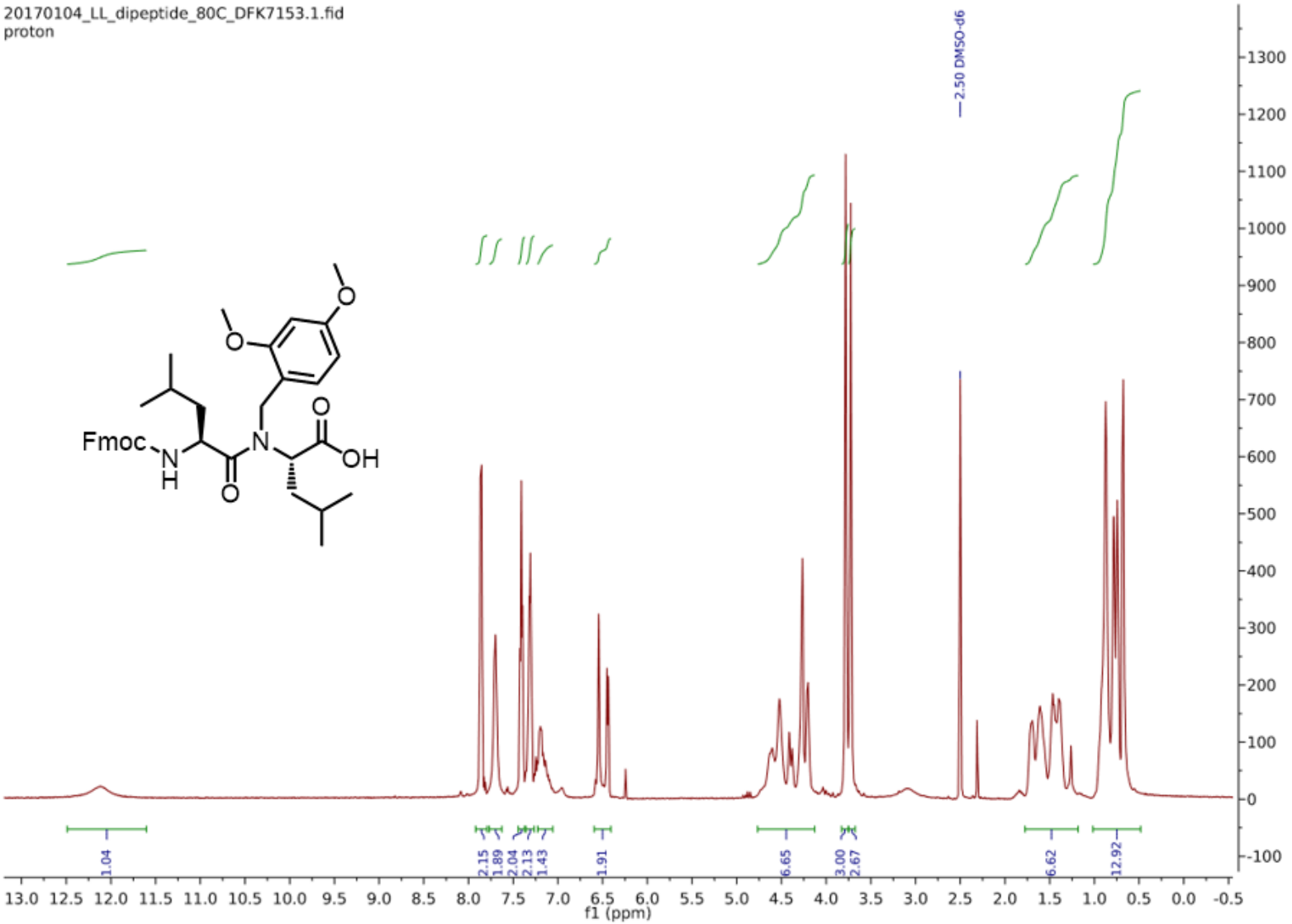


Figure A.16. ^{1}H-NMR spectrum (80°C) of Fmoc-(L)-Leu-(L)-(Dmb)Leu-OH.

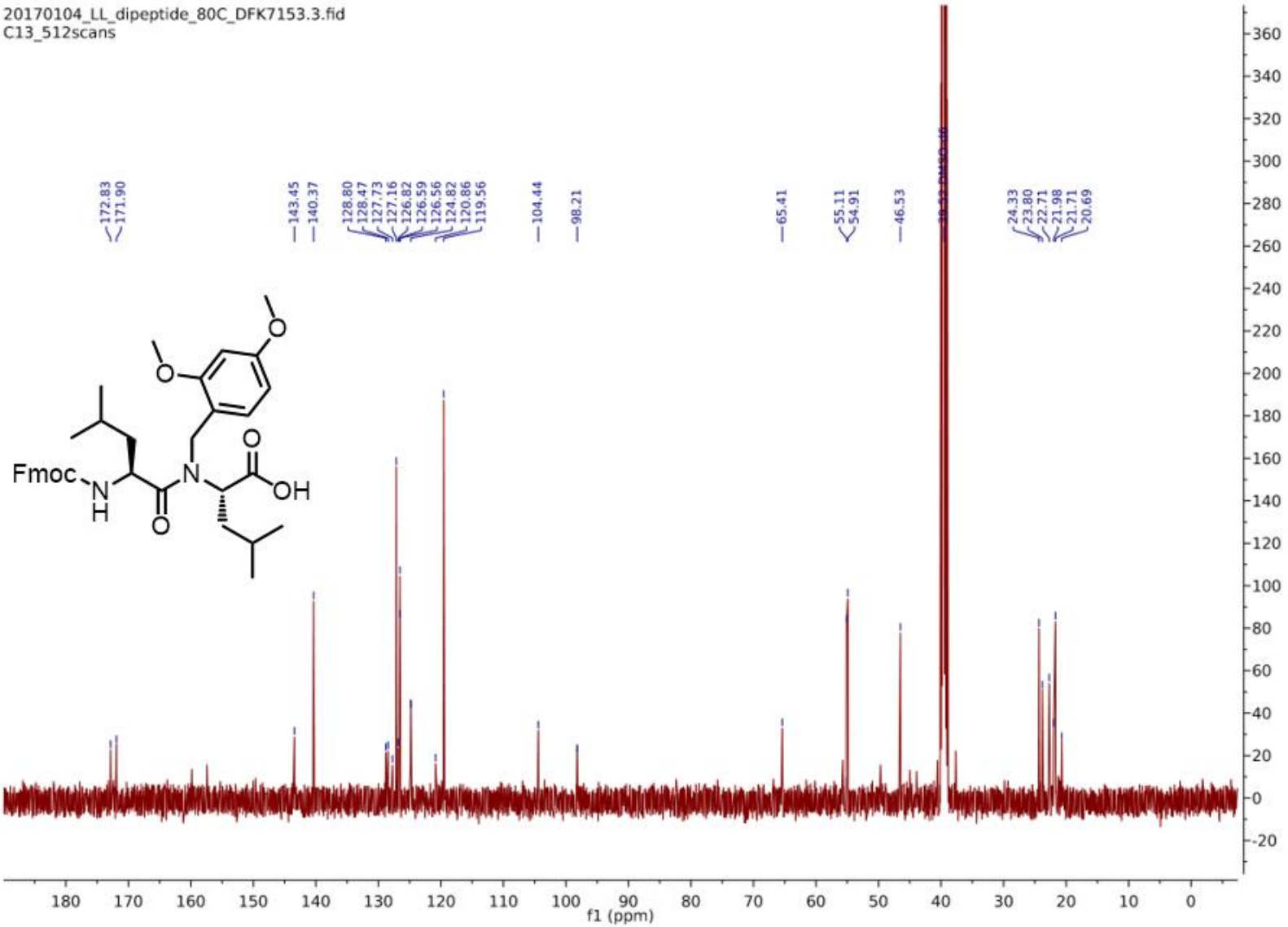
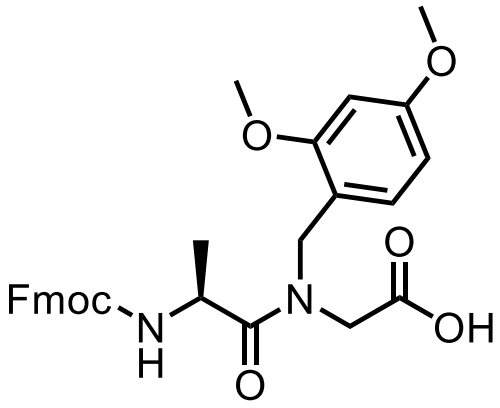


Figure A.17. ^{13}C-NMR spectrum (80°C) of Fmoc-(L)-Leu-(L)-(Dmb)Leu-OH.



Fmoc-(L)-Ala-(Dmb)Gly-OH

Formula: $C_{29}H_{30}N_2O_7$

***IUPAC*:** *N*-((((9H-fluoren-9-yl)methoxy)carbonyl)-L-alanyl)-*N*-(2,4-dimethoxybenzyl)glycine

***SMILES*:**

O=C(N[C@@H](C)C(N(CC1=C(OC)C=C(OC)C=C1)CC(O)=O)=O)OCC2C3=CC=CC=C3C4=CC=CC=C42

Procedure:

A 250 mL round-bottom flask was charged with Fmoc-(L)-Ala-OH (5.92 g, 19.0 mmol, 1.0 eq.) and Oxyma (2.70 g, 19.0 mmol, 1.0 eq.) which were dissolved in DMF (150 mL). *N,N'*-Diisopropylcarbodiimide (2.94 mL, 19.0 mmol, 1.0 eq.) was then added via pipet and the solution was stirred for 15 min at room temperature. H-(Dmb)Gly-OH (5.92 g, 19.0 mmol, 1.0 eq.) was added to the solution, and the flask was heated to 60°C for 2 h. The solution was then diluted to 250 mL with water (de-ionized) and extracted with EtOAc (3x100 mL). Organic layers were combined and washed with: 5% w/v aqueous LiCl (3x100 mL), saturated aqueous sodium bicarbonate (7x100 mL), 10% w/v aqueous citric acid (3x100 mL), and brine (100 mL) in the order listed. The organic solution was then

dried over sodium sulfate and concentrated via rotary evaporation to afford a crude brown oil that was purified via silica flash chromatography (30% EtOAc in hexanes). Product-containing fractions were combined and concentrated via rotary evaporation to afford a pale yellow oil that was subsequently dissolved in MeOH (100 mL). Pd/C (10% w/v, 700 mg) was added to the solution, and the atmosphere in the flask was evacuated and replaced with N_2. The flask was then outfitted with an H_2 balloon, and the solution was stirred vigorously for 3 h. The solution was then filtered over a celite pad and concentrated via rotary evaporation to afford a pale yellow oil that foamed into a white solid under vacuum. The foam was ground into a fine powder to afford 5.2 g of a white solid (54% yield over the two steps).

^{1}H NMR (500 MHz, DMSO-d6; 80°C): δ = 7.90-7.79 (m, 2H), 7.75-7.64 (m, 2H), 7.45-7.37 (m, 2H), 7.36-7.27 (m, 2H), 7.18-6.98 (m, 1H), 6.60-6.38 (m, 2H), 4.79-4.17 (m, 6H), 4.04-3.59 (m, 8H), 1.28-1.14 (m, 3H)

^{13}C NMR (500 MHz, DMSO-d6; 80°C): δ = 172.36, 169.60, 166.73, 165.13, 158.06, 143.58, 143.52, 140.38, 139.15, 137.17, 129.65, 129.42, 128.48, 127.17, 126.83, 126.63, 126.59, 126.52, 124.85, 124.78, 123.66, 120.87, 119.56, 119.51, 119.44, 116.08, 108.79, 104.95, 104.66, 98.43, 98.20, 65.51, 55.27, 55.13, 54.94, 54.89, 49.84, 49.06, 46.52, 46.17, 45.79, 43.14, 18.73, 17.74

ESI-MS [M+H]: *calculated* 519.2126, *observed* 519.2125, **[M+NH$_4$]:** *calculated* 536.2391, *observed* 536.2390

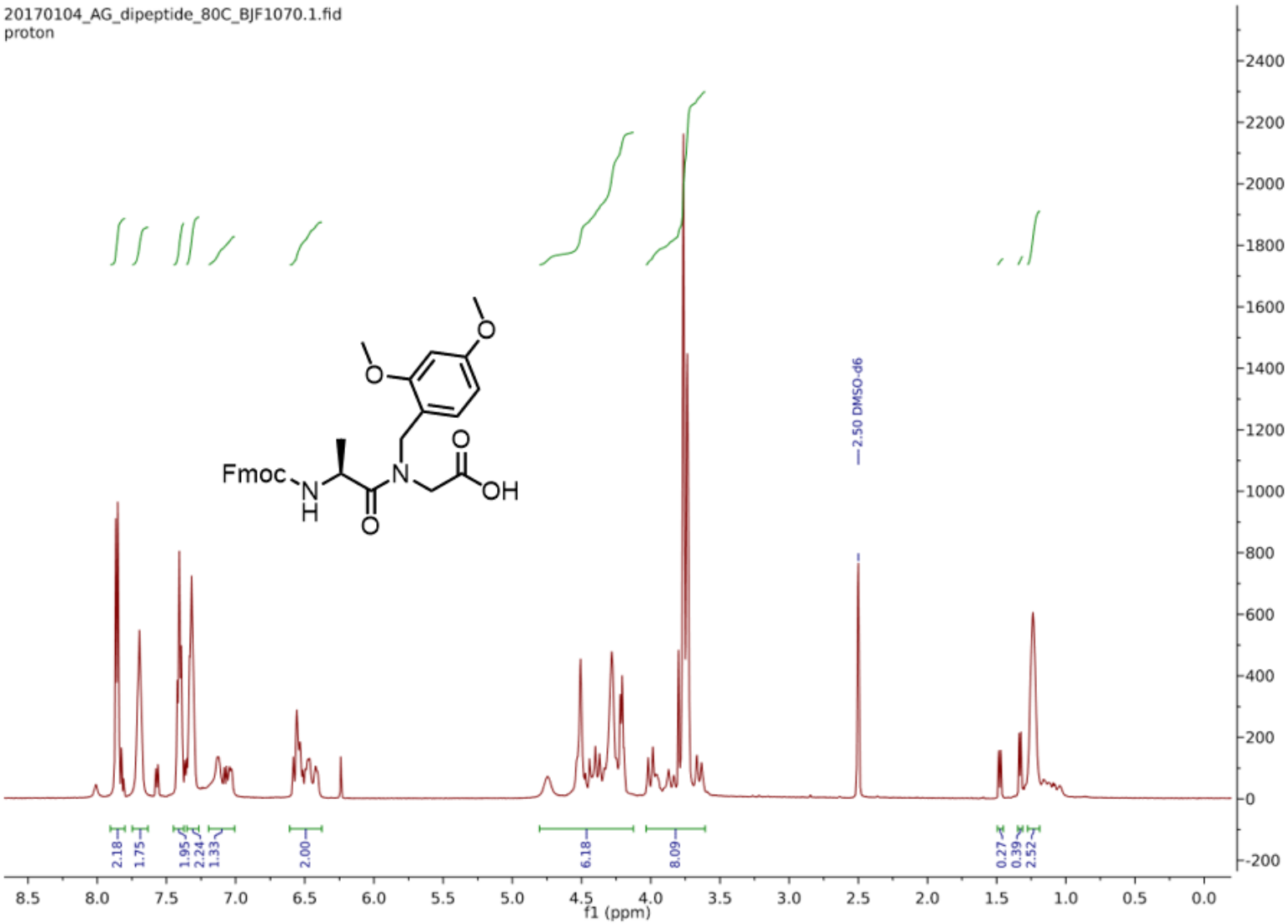


Figure A.18. ^{1}H-NMR spectrum of Fmoc-(L)-Ala-(Dmb)Gly-OH.

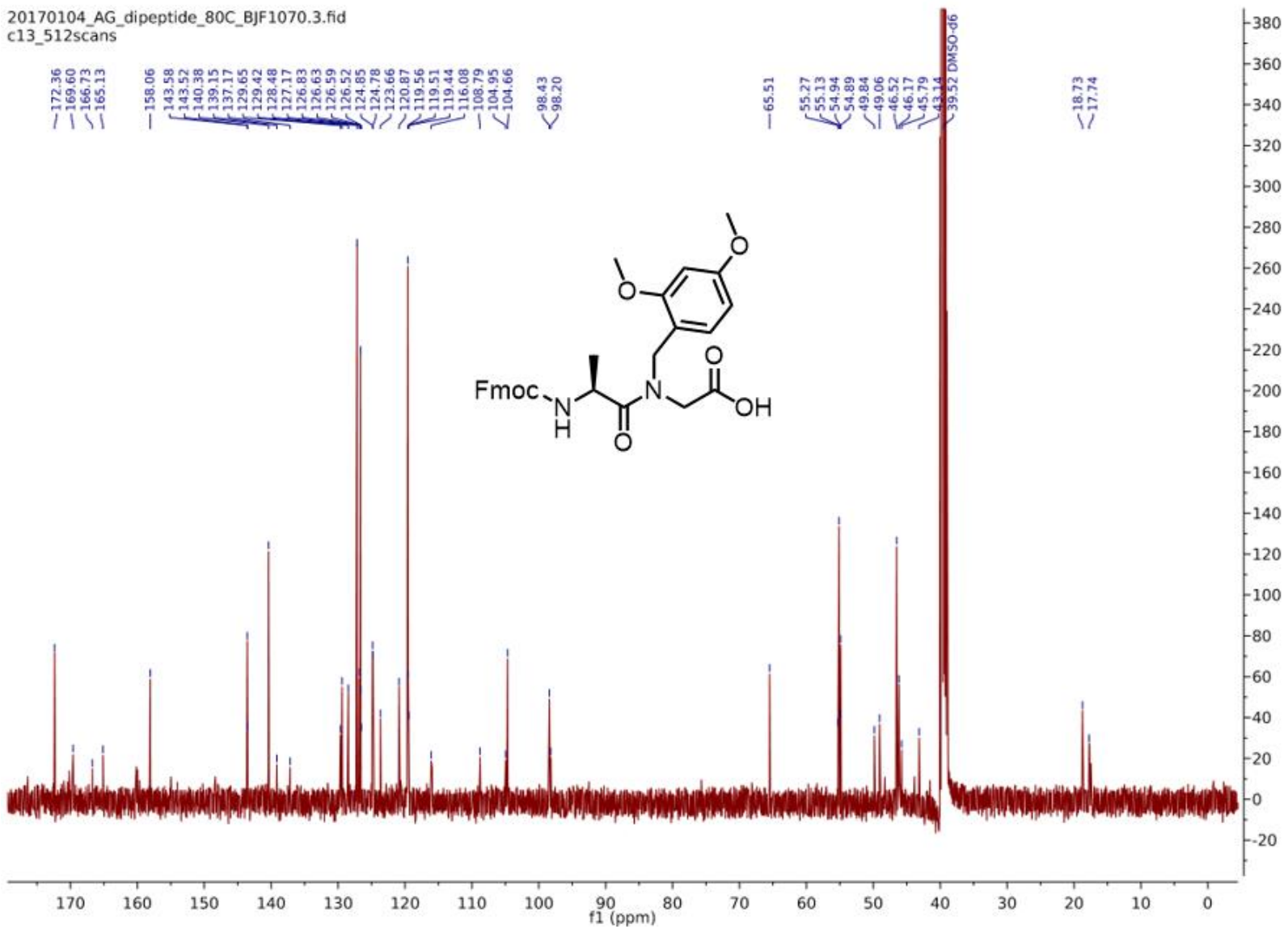
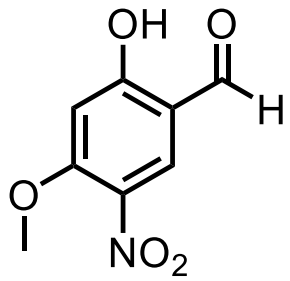


Figure A.19. ^{13}C-NMR spectrum of Fmoc-(L)-(Dmb)Gly-OH.



2-hydroxy-4-methoxy-5-nitrobenzaldehyde (Hmnb)[9,10]

Formula: $C_8H_7NO_5$

***IUPAC*:** 2-hydroxy-4-methoxy-5-nitrobenzaldehyde

***SMILES*:** [H]C(C1=C(O)C=C(OC)C([N+]([O-])=O)=C1)=O

Procedure: 1:1 AcOH:HNO_3 (40 mL) was added to a 250 mL round-bottom flask which was cooled on ice. 2-Hydroxy-4-methoxybenzaldehyde (5.0 g, 32.9 mmol, 1.0 eq.) was added portion-wise to the AcOH:HNO_3 solution. The resultant solution was then stirred for 16 h while the flask remained on ice. The solution was then diluted with ice water (200 mL), which caused a light-brown precipitate to form. The precipitate was collected via filtration (Buchner funnel, Whatman filter paper, and water aspirator) and dried under vacuum for 2 h to afford a light-brown powder ,which was re-crystallized in ethanol to afford 1.2 g of a light-yellow, feathery powder (18% yield).

^{1}H NMR (400 MHz, DMSO-d6): δ = 12.04 (s, 1H), 10.12 (s, 1H), 8.30 (s, 1H), 3.97 (s, 3H)

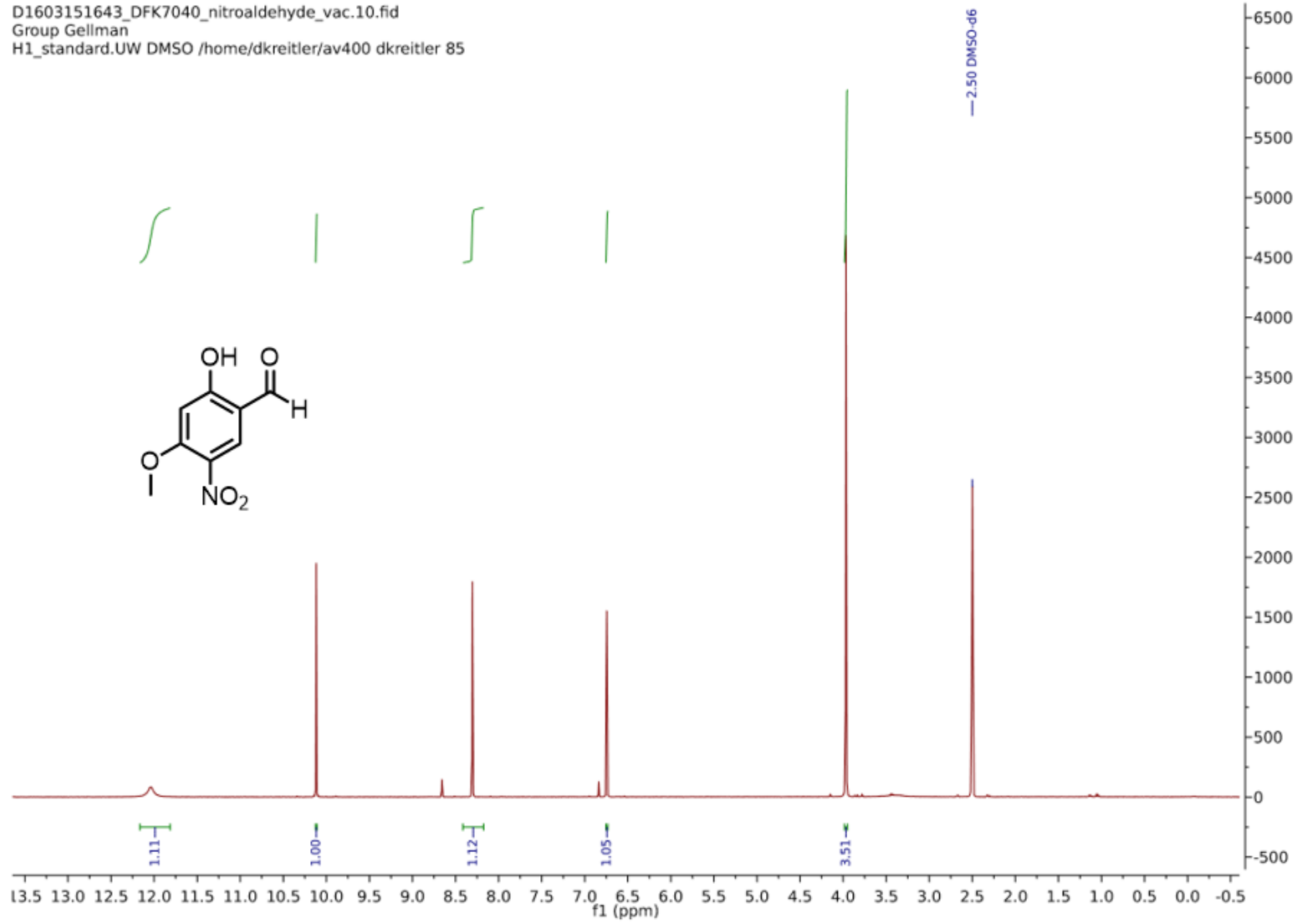


Figure A.20. ^{1}H-NMR spectrum of 2-hydroxy-4-methoxy-5-nitrobenzaldehyde (Hmnb).

Table A.1. Sequences and purification conditions of screened TM-peptides.

Transmembrane domain sequence	Notebook reference	HPLC gradient solvents
H-IPSIATGMVGALLLLLVVA**LG**IGLFMRRRHIVR-$CONH_2$	EGFR-TM-A (L) Synthesis: DFK6108 Purification:DFK6113-6, DFK6160	A: EtOH/H_2O, 0.05% v/v TFA B: iPrOH, 0.05% v/v TFA
H-IPSIATGMVGALLLLLVVA**LG**IGLFMRRRHIV-$CONH_2$	EGFR-TM-B (L) Synthesis: DFK6108 Purification:DFK6113-6	A: EtOH/H_2O, 0.05% v/v TFA B: iPrOH, 0.05% v/v TFA
H-IPSIATGMVGALLLLLVVA**LG**IGLFMRRRHI-$CONH_2$	EGFR-TM-C (L) Synthesis: DFK6108 Purification:DFK6113-6	A: EtOH/H_2O, 0.05% v/v TFA B: iPrOH, 0.05% v/v TFA
H-IPSIATGMVGALLLLLVVA**LG**IGLFMRRRHIV-COOH	EGFR-TM-D (L) Synthesis: DFK6108 Purification:DFK6113-6	A: EtOH/H_2O, 0.05% v/v TFA B: iPrOH, 0.05% v/v TFA
H-IPSIATGMVGALLLLLVVA**LG**IGLFMRRRHIVR-COOH	EGFR-TM-E (L) Synthesis: DFK6125 Purification: DFK6160	A: EtOH/H_2O, 0.05% v/v TFA B: iPrOH, 0.05% v/v TFA
H-IPSIATGMVGALL**LL**LVVA**LG**IGLFMRRRHIVR-$CONH_2$	EGFR-TM-F (L) Synthesis: DFK6174 Purification: DFK6182	A: EtOH/H_2O, 0.05% v/v TFA B: iPrOH, 0.05% v/v TFA
H-IPSIATGM**VG**ALL**LL**LVVA**LG**IGLFMRRRHIVR-$CONH_2$	EGFR-TM-G (L) Synthesis: DFK6194 Purification: DFK7016	A: EtOH/H_2O, 0.05% v/v TFA B: iPrOH, 0.05% v/v TFA
Ac-PSIATGMVGALLLLLVVALGIGLFMR-$CONH_2$	EGFR-TM-H (L) Synthesis: DFK5151 Purification:DFK5162	A: EtOH/H_2O, 0.05% v/v TFA B: iPrOH, 0.05% v/v TFA
Ac-PSIATGMVGALLLLLVVALGIGLFMR-$CONH_2$	EGFR-TM-H (D) Synthesis: DFK5146 Purification: DFK5165	A: EtOH/H_2O, 0.05% v/v TFA B: iPrOH, 0.05% v/v TFA
Ac-SIATGMVGALLLLLVVALGIGLFM-$CONH_2$	EGFR-TM-I (L) Synthesis: DFK3196 Purification: DFK5133	A: EtOH/H_2O, 0.05% v/v TFA B: iPrOH, 0.05% v/v TFA

Ac-PLTSIISAVVGILLVVVLGVVFGILIK-$CONH_2$	ErbB2-TM (L) Synthesis: DFK5137 Purification: DFK5148, DFK5174	A: EtOH/H_2O, 0.05% v/v TFA B: iPrOH, 0.05% v/v TFA
Ac-PLTSIISAVVGILLVVVLGVVFGILIK-$CONH_2$	ErbB2-TM (D) Synthesis: DFK5146 Purification: DFK5148, DFK5174	A: EtOH/H_2O, 0.05% v/v TFA B: iPrOH, 0.05% v/v TFA
Ac-VLSIIAIALGSLGLILIILLSVVVWK-$CONH_2$	PIV5-TM (D) Synthesis: DFK5146 Purification: DFK5156-7, DFK5178	A: EtOH/H_2O, 0.05% v/v TFA B: iPrOH, 0.05% v/v TFA
Ac-VLSIIAIALGSLGLILIILLSVVVWK-$CONH_2$	PIV5-TM (L) Synthesis: DFK5110 Purification: DFK5128	A: EtOH/H_2O, 0.05% v/v TFA B: iPrOH, 0.05% v/v TFA
Ac-GEIVAVIFGLLLGAALLLGILVFRSR-$CONH_2$	EphA1-TM (L) Synthesis: DFK5114 Purification: DFK5126-7	A: EtOH/H_2O, 0.05% v/v TFA B: iPrOH, 0.05% v/v TFA
Ac-GEIVAVIFGLLLGAALLLGILVFRSR-$CONH_2$	EphA1-TM (D) Synthesis: DFK5146	A: EtOH/H_2O, 0.05% v/v TFA B: iPrOH, 0.05% v/v TFA
Ac-LAVIGGVAVGVVLLLVLAGVGFFIHRR-$CONH_2$*	EphA2-TM Synthesis: DFK5116 L	A: EtOH/H_2O, 0.05% v/v TFA B: iPrOH, 0.05% v/v TFA
Ac-PSIATGMVGALLLLLVVALGIGLFMRRRHIVRKRTLRRLLQERE-$CONH_2$	EGFR-TM-JMA Synthesis: DFK5083 Purification: DFK5099	A: H_2O, 0.1% v/v TFA B: ACN, 0.1% v/v TFA
Ac-PLVVAASIIAILHLILWILDRL-$CONH_2$**	M2-TM Synthesis: DFK5106 Purification: DFK5124	A: 9:1 H_2O: ACN, 0.1% v/v TFA B: 9:3:1 iPrOH:CAN:H_2O, 0.1% v/v TFA
H-CFFDGVGEKVGGLAPTAN-$CONH_2$ Solubility tag: Ac-RRRRG	Flp pilin C-terminal thiol with solubilizing tag Synthesis: DFK7133-4 Purification: DFK7134-5	A: H_2O, 0.1% v/v TFA B: ACN, 0.1% v/v TFA
Ac-ANAIEYAVIAGLIAVALIAVLSPTDSGIVGGLK-$CONHNH_2$***	Flp pilin N-terminal hydrazide with solubilizing tag Synthesis: BJF1112, DFK9030 Purification: DFK9030-2	A: H_2O, 0.1% v/v TFA B: ACN, 0.1% v/v TFA
H-CLGLGLILIILLSVVVWKLL-CONH2** Solubility tag: Ac-RRRRRRβA	PIV5-TM with solubilizing tag	A: H_2O, 0.1% v/v TFA B: ACN, 0.1% v/v TFA

	Synthesis: DFK7082 Purification: DFK7090,DFK7120	
H-CLGSL**GL**ILII**LL**SVVVWKLL-$CONH_2$ Solubility tag: Ac-RRRRRR(βAla)	PIV5-TM with solubilizing tag Synthesis: DFK7058 Purification: DFK7064	A: H_2O, 0.1% v/v TFA B: ACN, 0.1% v/v TFA
H-CLGSL**GL**ILII**LL**SVVVWKLL-$CONH_2$[$] Solubility tag: Ac-RRRRG	PIV5-TM with solubilizing tag Synthesis: DFK9002 Purification: DFK9010	A: H_2O, 0.1% v/v TFA B: ACN, 0.1% v/v TFA
Ac-R(Nle)KQIEDKIEEILSKIYHIENEIARIKKLLSIIAI-CONHNH2	GCN4-pII/PIV5-TM C-terminal hydrazide	A: H_2O, 0.1% v/v TFA B: ACN, 0.1% v/v TFA

* not soluble in 2,2,2-trifluoroethanol (TFE), **microwave-assisted synthesis was employed, TFA=2,2,2-trifluoroacetic acid, ACN=acetonitrile, **XX**=2,4-dimethoxybenzyl amide protecting group, **XX**=solubilizing removable backbone amide linker (installed on-resin as Hmnb aldehyde), *** could not resolve impurities, but detected ion corresponding to tagged peptide in MALDI-TOF spectrum (DFK9031). [$]Synthesized at room temperature manually with 0.5-1 h double couplings (4 equivalent 100 mM Oxyma/COMU).
All Dmb-dipeptides were coupled with HATU/DIEA, even when another coupling reagent was used for the rest of the synthesis

Table A.2. LCP and detergent-based TM-peptide crystallization efforts.

TM-sequence	Crystallization media*	Peptide reference	Crystallization reference	Condition(s)
ErbB2	racemic (LDAO)	DFK5148/5158	DFK5175	Membfac 1-48
ErbB2	racemic (OG)	DFK5148/5158	DFK5180	Membfac 1-48
EGFR-TM-A	LCP	DFK6113A	DFK6121	Memgold 1-96, Membfac 1-48
EGFR-TM-G	LCP	DFK7016	DFK7020	Memgold 1-96, Membfac 1-48, Morpheus 1-75
EGFR-TM-H	racemic (OG)	DFK5162/5165	DFK5176	Membfac 1-48, Crystal screen II 1-48
PIV5	racemic (TFE)	DFK5110/5146	DFK5178	Membfac 1-48, Crystal screen II 1-48
M2	racemic (LDAO)	DFK3066	DFK6015	Memgold 1-96
KR12	racemic (TFE)	BJF1025	DFK6025	Memgold 1-96
EGFR-TM-B	LCP	DFK6113B	DFK6168	Memgold 42, 69 replicates
EGFR-TM-F	LCP	DFK6182	DFK7008	Memgold 1-48
EGFR-TM-G	LCP	DFK7016	DFK7020	Membfac 1-48, Morpheus 1-75, Memgold 1-96
EGFR-TM-F	OG	DFK6182	DFK9008	Crystal screen II, 1-48

Memgold, Morpheus (Molecular Dimensions, Altamonte Springs, FL, USA); Membfac, Crystal screen II (Hampton Research, Aliso Viejo, CA, USA)

* OG = β-D-octylglucoside, LDAO = lauryldimethylamine *N*-oxide, LCP = lipid cubic phase, TFE = 2,2,2-trifluoroethanol

LCP experiments were set-up as described by Steinkruger

Figure A.21. Potential crystallization hit for EGFR-TM-F (Hampton research crystal screen II #47; crystallization conditions: DFK9008). Hanging drop vapor diffusion 2% w/v β-D-octyl glucoside. Crystal dissolved when pulled from the drop and mounted at room temperature in a mitogen sleeve.

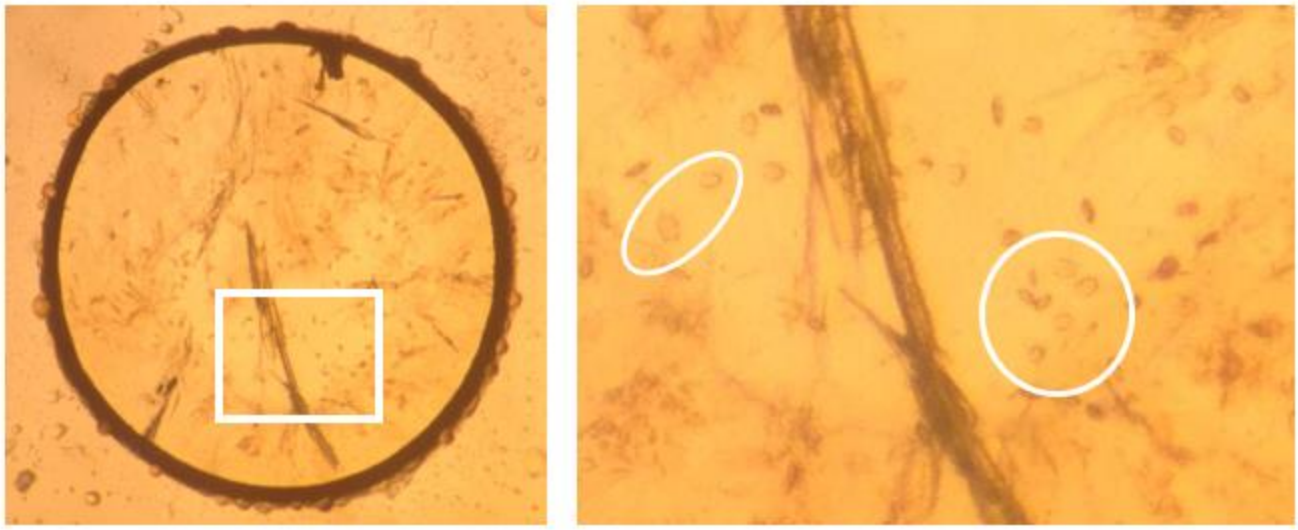


Figure A.22. Potential crystallization hit for EGFR-TM-G in the lipid cubic phase, Molecular Dimensions Memgold I #94 (II-46). Crystals observed 8 months after the initial condition was set-up (Peptide homochiral EGFR-TM; crystallization set-up described on DFK7020). The drop has dried out and only the lipid bollus remains. An attempt to mount these extremely small crystals was unsuccessful.

A.3 Automated Synthesis of Transmembrane Peptides

The following room temperature, automated synthesis procedure was found to afford useful quantities of transmembrane peptides (>5 mg for 50 µmol scale) with decent purity (>85% by UPLC peak area; 220 nm):

50 umol scale, ChemMatrix rink amide resin (0.4-0.5 mmol/g)

Coupling: 2x1 h coupling solution (N_2 mixing; 4 mL 1:1 A:B; A=0.1 M Fmoc-amino acid, B=0.1 M HATU/0.4 M *N*-methylmorpholine, both A and B solutions are in biotechnology grade DMF) followed by 5x2 mL DMF (ACS reagent grade) washes with 30 second mixing time. In cases where a Dmb-dipeptide is employed, only a single coupling reaction is performed.

Deprotection: 1x5min, 15 min deprotection solution (N_2 mixing; 2 mL 20% v/v piperidine in ACS reagent grade DMF) followed by 6x2 mL DMF (ACS reagent grade) washes with 30 second mixing time.

Acetylation: Following the final deprotection and washing step, resin is transferred to a fritted polypropylene syringe and free amino termini are acetylated with 4 mL of 8/1/1 DMF/Ac_2O/DIEA for 15 min (mixing is performed by capping the syringe and placing it on a rocker). After this time the solution is drained from the syringe and the resin is washed with 3x4 mL DMF followed by 2x2 mL DCM. Finally, the resin is dried on a vacuum manifold.

Cleavage: 95/2.5/2.5 TFA/H2O/TIPS (4 mL) is added to the resin bed in the fritted syringe, and the syringe is placed on a rocker for approximately 3 h (cleavage solutions of sequences that contain one or multiple Dmb-dipeptides tend to take on a deep purple hue). The cleavage solution is then drained into a 24 mL glass vial, and TFA is removed

by a gentle N_2 stream to afford a crude, viscous oil. The vial is then connected to a vacuum manifold outfitted with a KOH trap, and residual TFA is removed under vacuum for an additional 1-2 h. The crude peptide oil is precipitated from cold, anhydrous diethyl ether (-20°C; 2x20 mL; occasionally the crude oil will remain adhered to the bottom of the glass vial when ether is added; this adherence can be alleviated by sonicating the vial for 10-20 seconds). The precipitate is then allowed to settle to the bottom of the glass vial, and ether is gently removed via pipet. Residual ether is then removed by a gentle N_2 stream. Finally, the glass vial is connected to a vacuum manifold, and the peptide is dried for an additional 1-2 h under vacuum.

Application Note: A glass vial is employed here for precipitation and drying steps rather than a 50 mL centrifuge tube. This switch is made for two reasons. The first reason is to avoid the introduction of 'ghost peaks' in HPLC chromatograms that may arise from preparing HPLC samples in fluorinated solvents, which can possibly extract UV-active compounds from plastic centrifuge tubes. The second reason is that a glass vial can be readily connected to a vacuum manifold outfitted with an oil pump that can remove residual TFA and ether. The 'complete' removal of TFA and ether is important because residual solvent can affect solubility during HPLC sample preparation.

HPLC sample preparation: The dry peptide powder (in a glass vial) is dissolved in 5 mL of 2,2,2-trifluoroethanol (TFE) and filtered through a 0.2 µm syringe filter.

HPLC purification overview: The employed HPLC solvent system can be a major factor in whether or not purification of a hydrophobic peptide sequence is successful. No solvent system is a panacea for hydrophobic peptides. Furthermore, solvent systems that require elevated temperature or fluorinated solvents in the mobile phase can have deleterious

effects on preparative, reverse-phase column performance and longevity. Work on the M2 ion channel has shown that some proportion of acetonitrile and isopropanol is an effective eluting solvent for the hydrophobic M2 TM peptide.[11–13] Yao *et al.* have reported an effective solvent combination for eluting the transmembrane domain of parainfluenza virus 5 (PIV5).[3] The Yao A solvent comprises 50% v/v EtOH/H_2O, and the B solvent comprises isopropanol, both with 0.1% v/v TFA. I have found that 0.05% v/v TFA with these solvents is also effective.

The HPLC purification conditions for EGFR, ErbB2, PIV5, and EphA1 transmembrane domains are reported here in Figures 23-26 respectively. In each case the Yao solvent system with 0.05% v/v TFA was employed at room temperature. Anhydrous reagent grade ethanol was used with HPLC grade isopropanol and millipore filtered water. Thorough degassing of water/organic solvent mixtures is essential. These mixtures tend to foam more readily, and this foaming can result in the accumulation of air in reciprocating HPLC pump heads, which is highly detrimental to peak resolution.

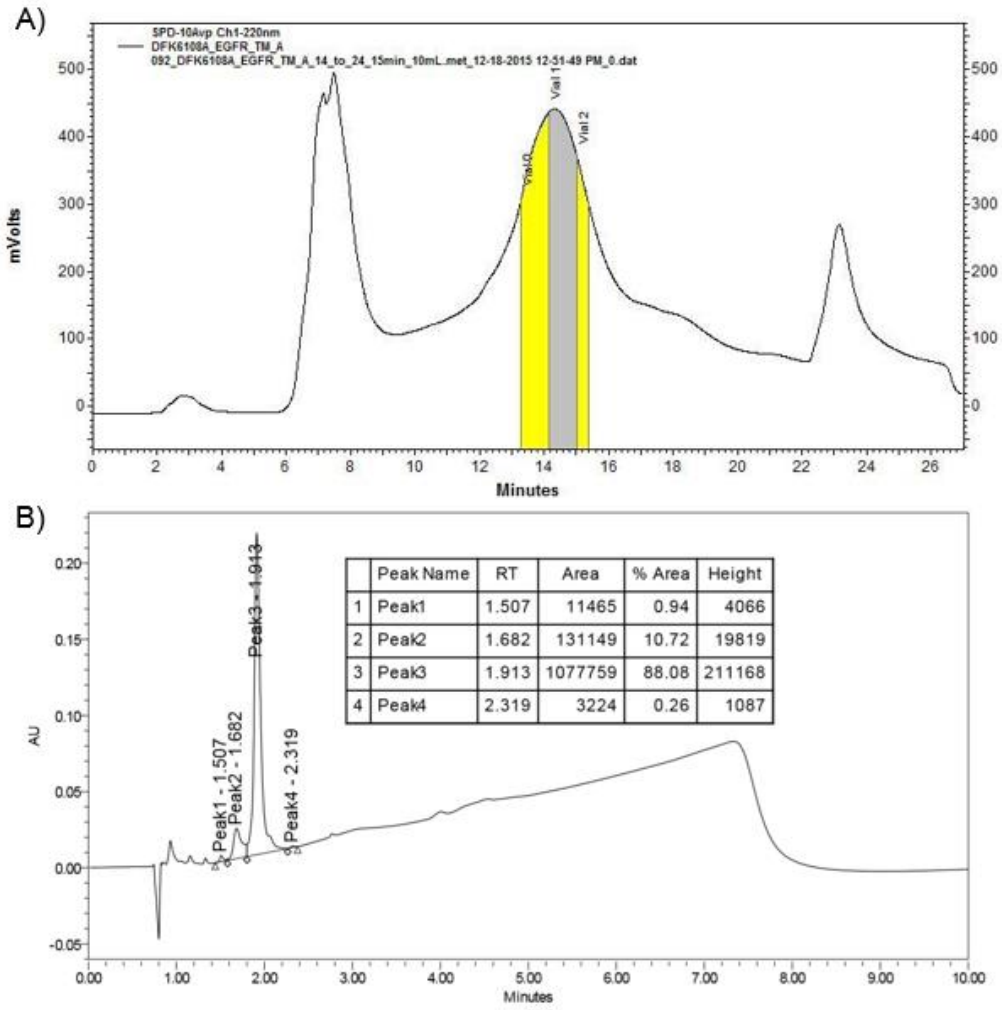


Figure A.23. A) Crude preparatory HPLC (yellow and gray shading represent peak areas that were collected during purification) and B) UPLC purity analysis chromatograms for the EGFR-TM (sequence: EGFR-TM-A in Table A.**1**). Preparatory HPLC: column, Supelco C5 (21.2 x 250 mm), flow rate 10 mL/min, gradient 14-24%B over 15 minutes, A solvent (50% v/v EtOH/H_2O, 0.05% v/v TFA), B solvent (iPrOH, HPLC grade, 0.05% v/v TFA), injection volume 250 µL. Analytical UPLC: column, Waters-Acquity Protein BEH C4 (2.1 x 100 mm, 1.7 µm, 50°C), flow rate 0.35 mL/min, gradient 10-65%B over 6 minutes, A solvent (50% v/v EtOH/H_2O, 0.05% v/v TFA), B solvent (iPrOH, HPLC grade, 0.05% v/v TFA). **$[M+H]_{mono}$** calculated 3556.2, observed 3555.2.

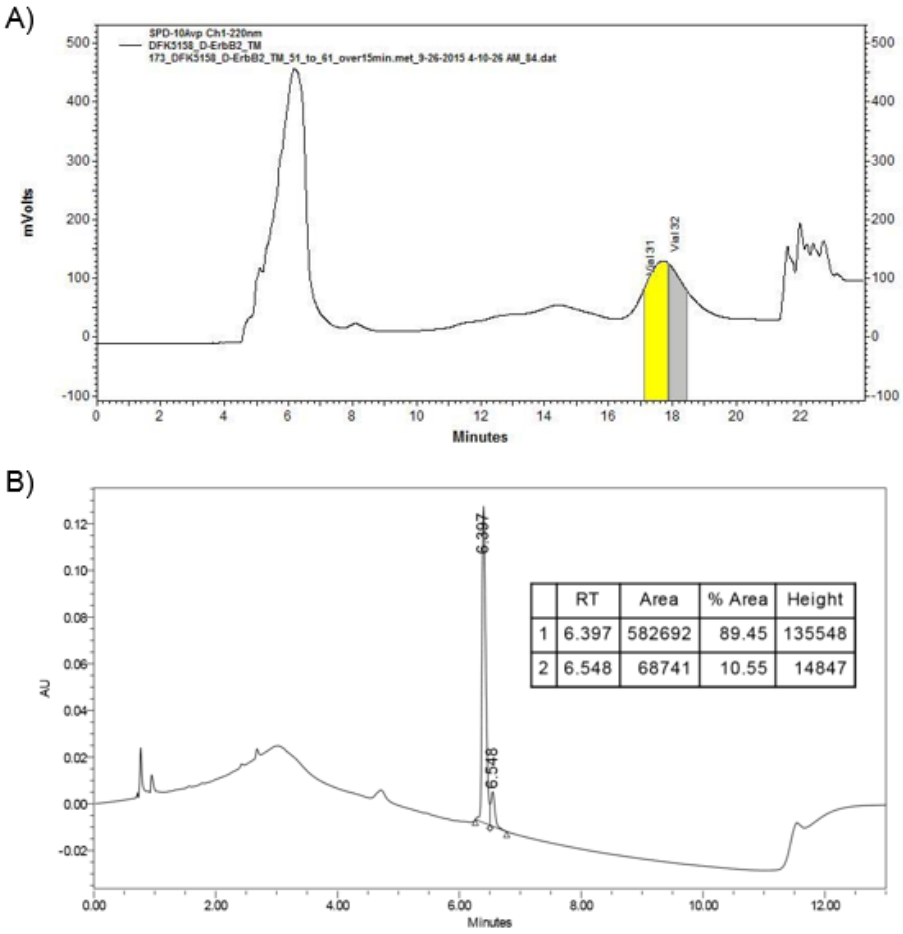


Figure A.24. A) Crude preparatory HPLC (yellow and gray shading represent peak areas that were collected during purification) and B) UPLC purity analysis chromatograms for (D)-ErbB2-TM (Table A**.1**). Preparatory HPLC: column, Supelco C5 (21.2 x 250 mm), flow rate 10 mL/min, gradient 51-61%B over 15 minutes, A solvent (50% v/v EtOH/H_2O, 0.05% v/v TFA), B solvent (iPrOH, HPLC grade, 0.05% v/v TFA), injection volume 250 µL. Analytical UPLC: column, Waters-Acquity Protein BEH C4 (2.1 x 100 mm, 1.7 µm, 50°C), flow rate 0.35 mL/min, gradient 10-95%B over 9.5 minutes, A solvent (50% v/v EtOH/H_2O, 0.05% v/v TFA), B solvent (iPrOH, HPLC grade, 0.05% v/v TFA). **[M+Na]$_{mono}$** calculated 2795.9, observed 2796.1.

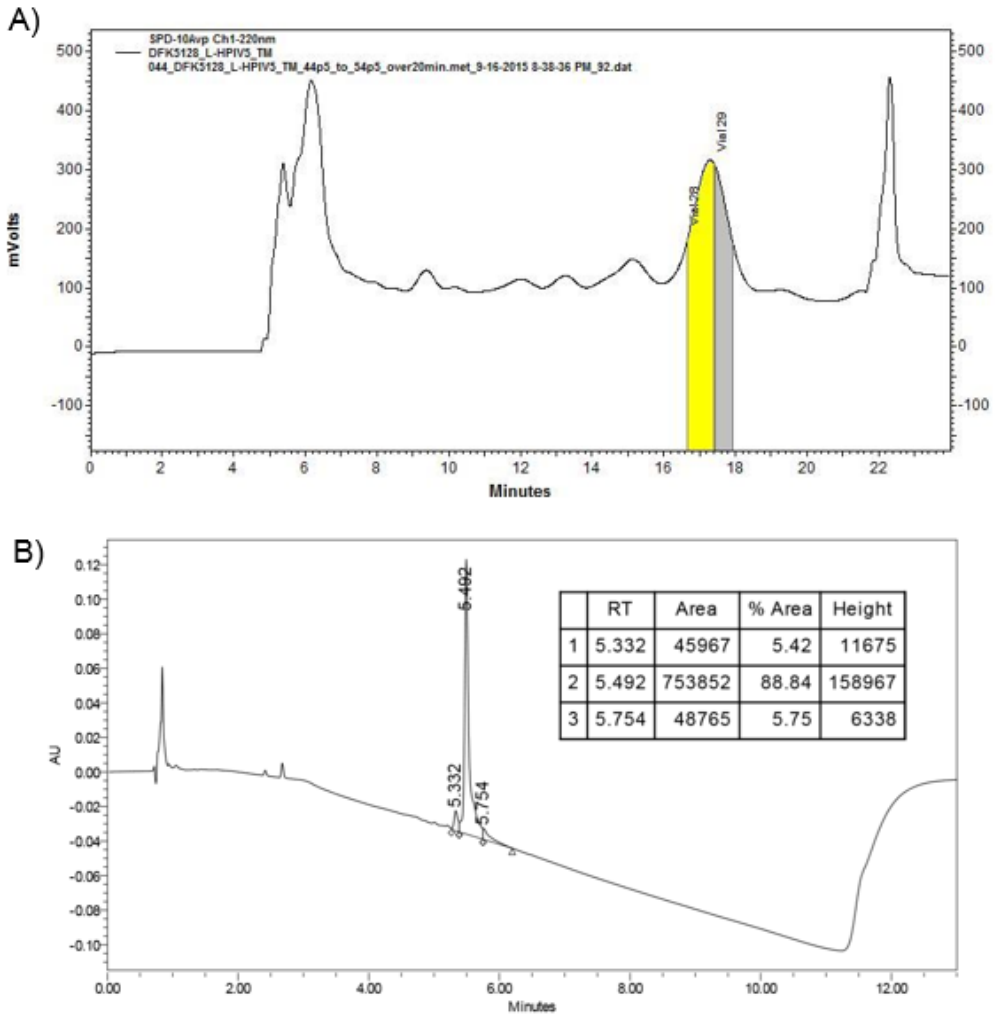
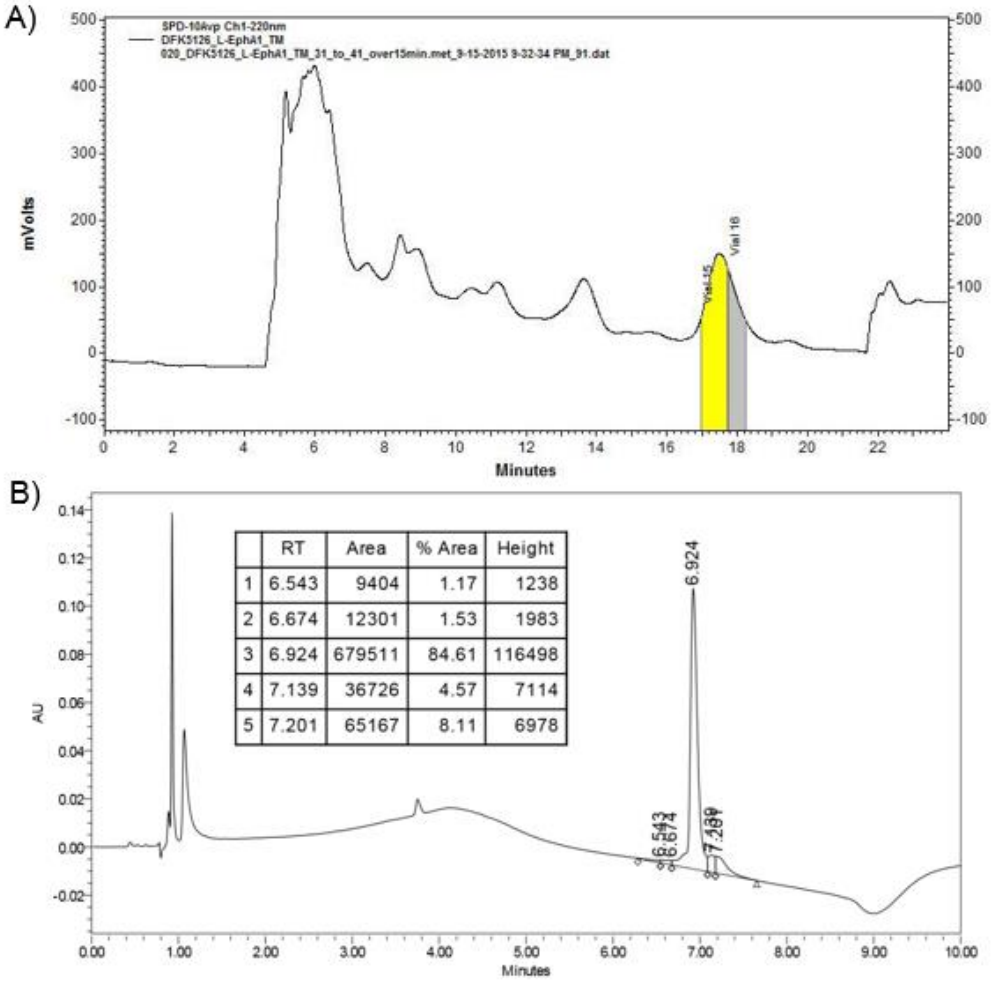


Figure A.25. A) Crude preparatory HPLC (yellow and gray shading represent peak areas that were collected during purification) and B) UPLC purity analysis chromatograms for the PIV5-TM (Table A.**1**). Preparatory HPLC: column, Supelco C5 (21.2 x 250 mm), flow rate 10 mL/min, gradient 44.5-54.5%B over 20 minutes, A solvent (50% v/v EtOH/H_2O, 0.05% v/v TFA), B solvent (iPrOH, HPLC grade, 0.05% v/v TFA), injection volume 250 µL. Analytical UPLC: column, Waters-Acquity Protein BEH C4 (2.1 x 100 mm, 1.7 µm, 50°C), flow rate 0.35 mL/min, gradient 10-95%B over 9.5 minutes B solvent, A solvent (50% v/v EtOH/H_2O, 0.05% v/v TFA), B solvent (iPrOH, HPLC grade, 0.05% v/v TFA). **[M+Na]$_{mono}$** calculated 2779.8, observed 2779.2.



	RT	Area	% Area	Height
1	6.543	9404	1.17	1238
2	6.674	12301	1.53	1983
3	6.924	679511	84.61	116498
4	7.139	36726	4.57	7114
5	7.201	65167	8.11	6978

Figure A.26. A) Crude preparatory HPLC (yellow and gray shading represent peak areas that were collected during purification) and B) UPLC purity analysis chromatograms for (L)-EphA1-TM (Table A**.1**). Preparatory HPLC: column, Supelco C5 (21.2 x 250 mm), flow rate 10 mL/min, gradient 31-41%B over 15 minutes, A solvent (50% v/v EtOH/H_2O, 0.05% v/v TFA), B solvent (iPrOH, HPLC grade, 0.05% v/v TFA), injection volume 250 µL. Analytical UPLC: column, Waters-Acquity Protein BEH C4 (2.1 x 100 mm, 1.7 µm, 50°C), flow rate 0.30 mL/min, gradient 5-60%B over 7 minutes, A solvent (50% v/v EtOH/H_2O, 0.05% v/v TFA), B solvent (iPrOH, HPLC grade, 0.05% v/v TFA). **$[M+H]_{mono}$** calculated 2751.7, observed 2753.2.

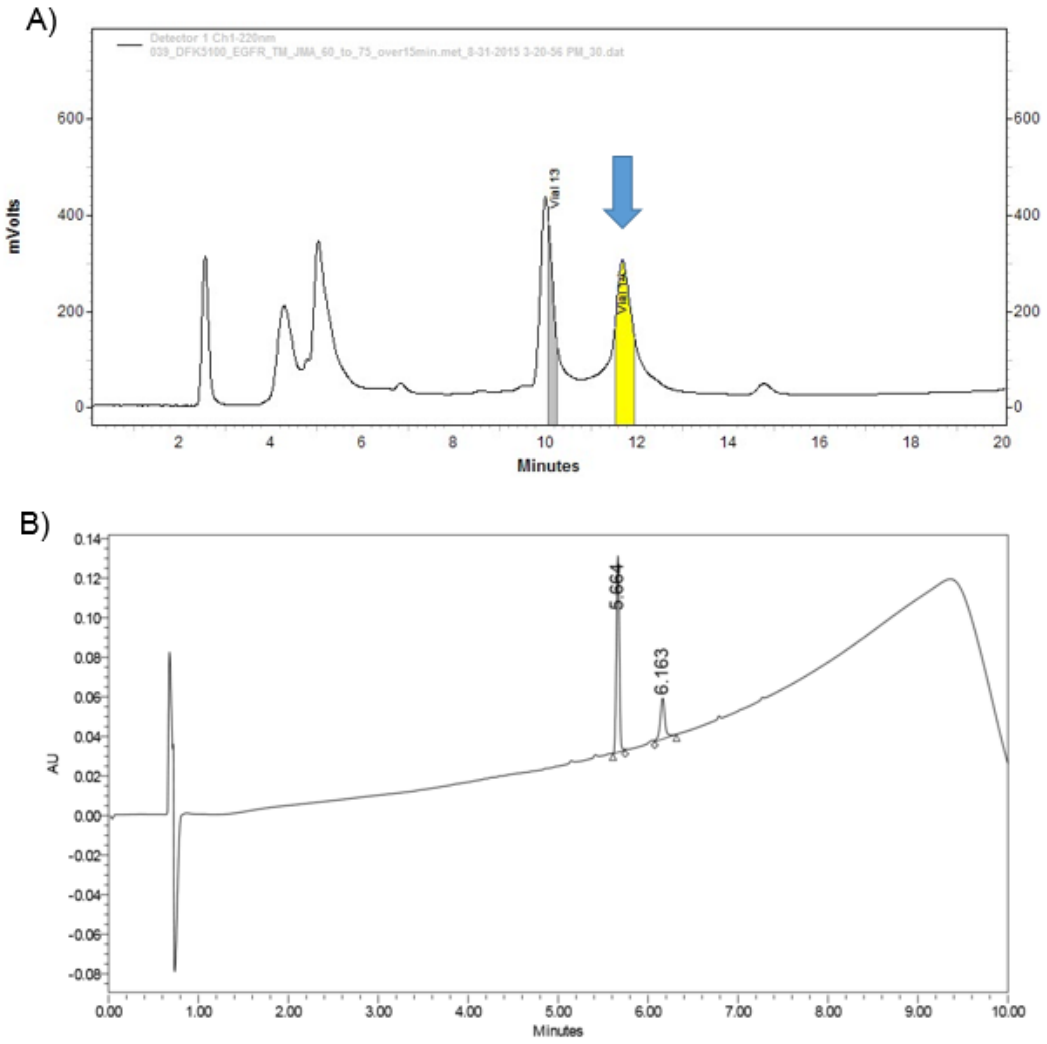


Figure A.27. A) Crude preparatory HPLC (isolated peak indicated by blue arrow and yellow shading) and B) UPLC purity analysis chromatograms for EGFR-TM-JMA (Table A**.1**). Preparatory HPLC: column, Supelco C5 (21.2 x 250 mm), flow rate 15 mL/min, gradient 60-75%B over 15 minutes, A solvent (H_2O, 0.1% v/v TFA), B solvent (ACN, HPLC grade, 0.1% v/v TFA). Analytical UPLC: column, Waters-Acquity Protein BEH C4 (2.1 x 100 mm, 1.7 µm, 50°C), flow rate 0.40 mL/min, gradient 10-95%B over 8 minutes, A solvent (H_2O, 0.1% v/v TFA), B solvent (ACN, HPLC grade, 0.1% v/v TFA). **$[M+H]_{avg}$** calculated 5066.2, observed 5066.5 (Important technical note: this peptide does not ionize well enough during MALDI-TOF MS to be detected in settings that employ a reflectron, successful detection requires instrument settings found in the parameter file LP_Pepmix.par). Furthermore, the isolated yield was very low (<1 mg from 100 µmol scale synthesis).

A.4 A Removable Backbone Linker (RMB) to Assist the Synthesis and Purification of Transmembrane (TM) Peptides

The aim here was to synthesize the N and C-terminal segments of the *Pseudomonas aeruginosa* Flp pilin, both tagged with a solubilizing tag at the underlined positions (*vide infra*), and then subsequently ligate, desulfurize, and remove the solubility tag using the procedure described by Liu *et al.* to afford the full-length protein (Figure A***.3***).[9] In the N-terminal segment, I placed the tag in the middle of the sequence, to facilitate synthesis of the more difficult latter half of the peptide (Figure A**.**28). I chose a position immediately before a relatively unhindered residue (Ala), to facilitate acylation following the tagged Leu residue. In the C-terminal peptide, I placed the solubility tag at a DG motif to prevent aspartimide formation during synthesis. DBU, a strong base that would typically promote aspartimide formation, was used in the deprotection solution; no detectable aspartimide formation occurred. I was able to obtain ~2.5 mg (100 µmol scale) of the solubility tag C-terminal peptide. The N-terminal peptide proved more difficult; I was able to detect an ion that corresponded to the [M+H] ion of the desired product during the attempted HPLC purification, but there were many unresolvable impurities (Figure A***.31***). The solubility-tagged N-terminal peptide could likely be obtained in sufficient yield after more optimization of the synthesis.

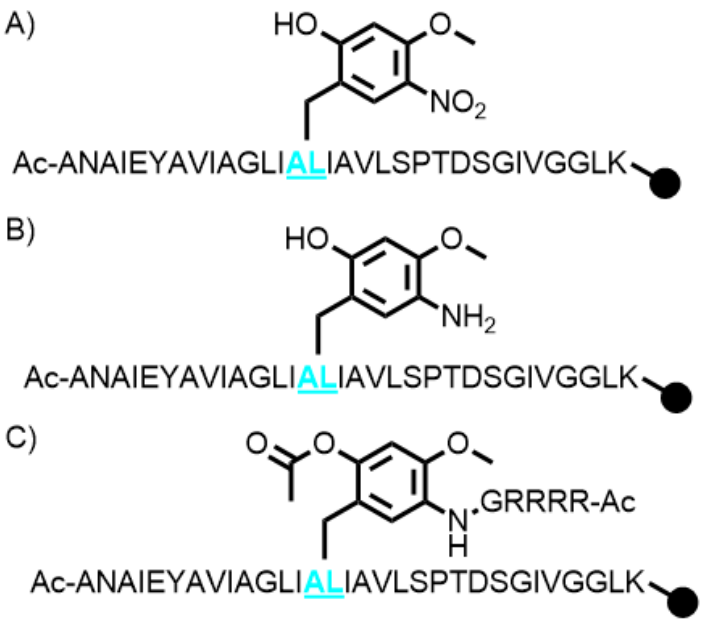


Figure A.28. A) The unreduced Hmnb-derived solubility tag. B) The nitro group is reduced to an aniline via $SnCl_2$/HCl in DMF. C) Finally, solubilizing residues can be introduced to the aniline functionality via standard Fmoc-SPPS, followed by acetylation. The acetylated hydroxyl group stabilizes the solubilizing linker to TFA cleavage. In each case the sequence is derived from the N-terminal segment of the P. aeruginosa Flp pilin.

N-term: Ac-ANAIEYAVIAGLIAVALIAVLSPTDSGIVGGLK-CONHNH$_2$

C-term: CFFDGVGEKVGGLAPTAN-CONH$_2$

Full-length protein:

Ac-ANAIEYAVIAGLIAVALIAVLSPTDSGIVGGLK**A**FFDGVGEKVGGLAPTAN-CONH$_2$

The following procedure was used to synthesize the *Pseudomonas aeruginosa* Flp pilin C-terminus containing a Ac-RRRRG solubility tag:

100 µmol scale synthesis, µwave-assisted

Coupling reagent: (1-Cyano-2-ethoxy-2-oxoethylidenaminooxy)dimethylamino-morpholino-carbenium hexafluorophosphate (COMU)

Base: 2,4,6-trimethylpyridine (TMP)

Coupling solvent: DMF (biotechnology grade; Sigma-Aldrich)

Deprotection solution: 5% w/v piperazine and 2% v/v DBU in DMF (ACS reagent grade; Sigma-Aldrich)

Coupling reaction: 2x2 min ramp to 70°C, 10 min hold at 70°C; 100 mM Fmoc-amino acid (4 equivalents), COMU (4 equivalents), TMP (8 equivalents). Coupling solutions were preactivated for >1 min prior to addition to resin.

Resin loading: NovaPEG amino resin (0.49 mmol/g) was loaded with 1.5 equivalents of Fmoc-peptide amide linker (Fmoc-PAL; Chem-Impex, Wood Dale, IL, USA), 1.5 equivalents of COMU, and 3 equivalents of TMP in DMF (biotechnology grade). The reaction was performed with a single extended coupling reaction (2 min ramp to 70°C, 10 min hold at 70°C).

Solubility tag linker installation: Reductive amination of the Hmnb aldehyde after Fmoc-Leu-OH deprotection, 2x25 min, 2 mL 50 mM Hmnb followed by 2x5 min 4 mL 100 mM $NaBH_4$ in DMF (ACS reagent grade). The resin was rinsed 2-3 times with DMF between each step.

Reduction of Hmnb nitro group to aniline: Following the final coupling of Boc-Cys(Trt)-OH the resin was washed with DMF and 4 mL of a 6 M $SnCl_2$,10 mM HCl solution was added to the resin (HCl was added from a 3 N MeOH/HCl solution). The resin was stirred vigorously while nitrogen was bubbled through. The reaction was repeated twice (1x3 h, 1 h).

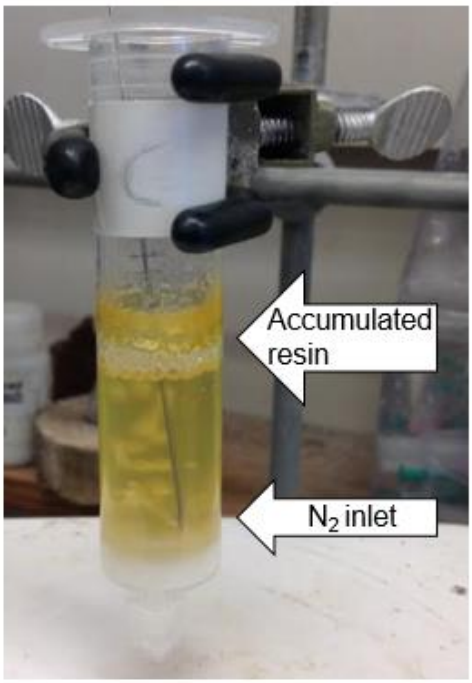


Figure A.29. $SnCl_2$/HCl-mediated reduction of nitro to aniline. Vigorous agitation of the resin bed is important for achieving complete reduction of the nitro group. The 6 M $SnCl_2$ solution is denser than the NovaPEG resin so the resin is buoyant in the solution. Vigorous stirring creates a vortex in the solution that prevents the resin from floating to the top of the vessel. Periodically, resin that accumulates at the top of the vessel at the air/solution interface must be rinsed or scraped down the side of the vessel. N_2 is introduced at the bottom of the vessel via a canula.

Solubility tag synthesis: Following nitro group reduction the sequence Ac-RRRRG was coupled onto the aniline functionality using the coupling procedure (*vide supra*). Liu *et al*. emphasize the importance of an initial, sterically unhindered residue (Gly) to facilitate coupling onto the aniline group.

Final deprotection and acetylation: After the final deprotection reaction, the free Arg amino group was acetylated by adding 8/1/1 DMF/Ac_2O/DIEA (4 mL) to the resin and stirring for 15 min. The free hydroxyl group on the tag linker was also acetylated at this step.

Cleavage: 6.8 mL TFA, 400 µL thioanisole, 400 µL H2O/ 200 µL EDT/ 600 mg phenol (2 h). The cleavage solution was then filtered into a glass vial and concentrated with an N_2

stream and dried for 1-2 h under vacuum to remove residual TFA. The peptide was then precipitated in diethyl ether (2x20 mL) and dried under N_2 and then under vacuum.

HPLC Sample preparation: The crude peptide powder was then dissolved in 15 mL TFE, 5 mL 1:1 H_2O/ACN. This solution was filtered with a 0.2 µm syringe filter and then purified via HPLC.

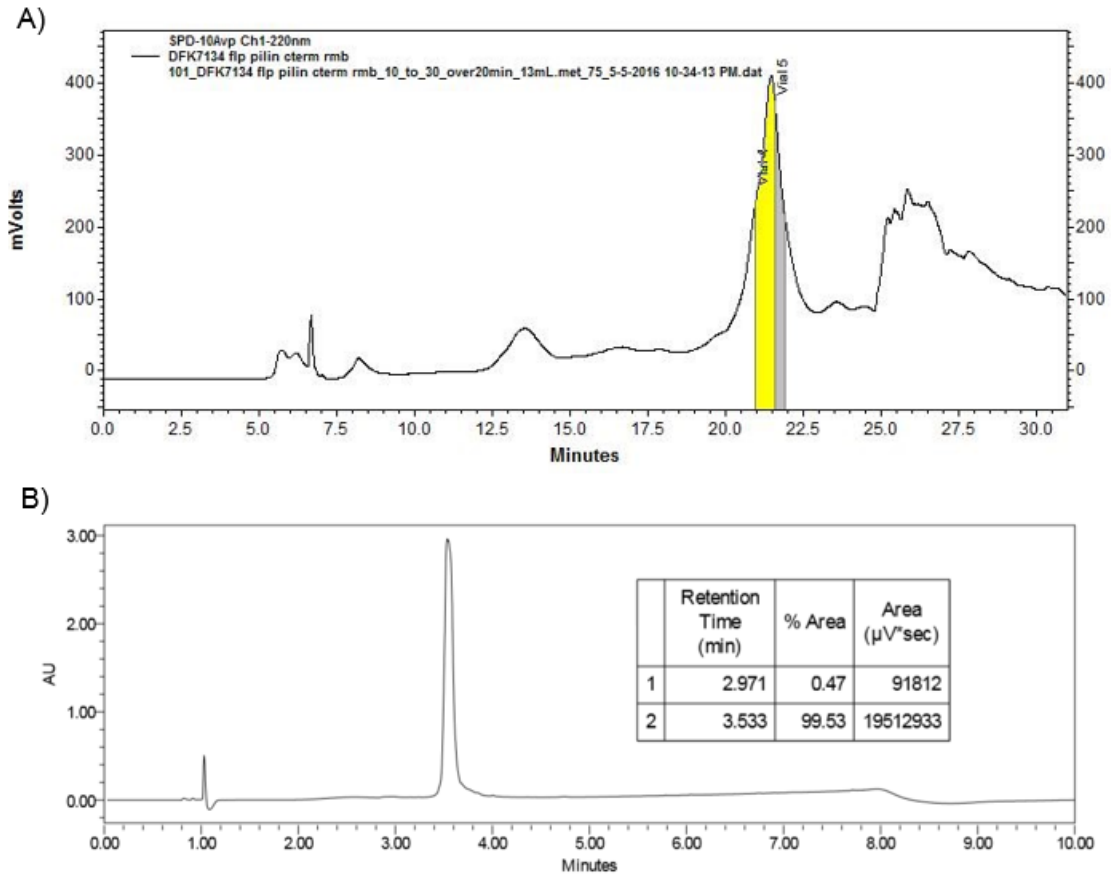


Figure A.30. A) Crude preparatory HPLC (yellow and gray shading represent peak areas that were collected during purification) and B) UPLC purity analysis chromatograms for the solubility tagged P. aeruginosa Flp pilin C-terminus (Table A**.1**). Preparatory HPLC: column, Supelco C5 (21.2 x 250 mm), flow rate 13 mL/min, gradient 10-30%B over 20 minutes, A solvent (H_2O, 0.1% v/v TFA), B solvent (ACN, HPLC grade, 0.1% v/v TFA). Analytical UPLC: column, Waters-Acquity Peptide BEH C18 (2.1 x 100 mm, 1.7 µm, 50°C), flow rate 0.30 mL/min, gradient 10-95%B over 6 minutes, A solvent (H_2O, 0.1% v/v TFA), B solvent (ACN, HPLC grade, 0.1% v/v TFA). **$[M+H]_{mono}$** calculated 2698.4, observed 2698.0.

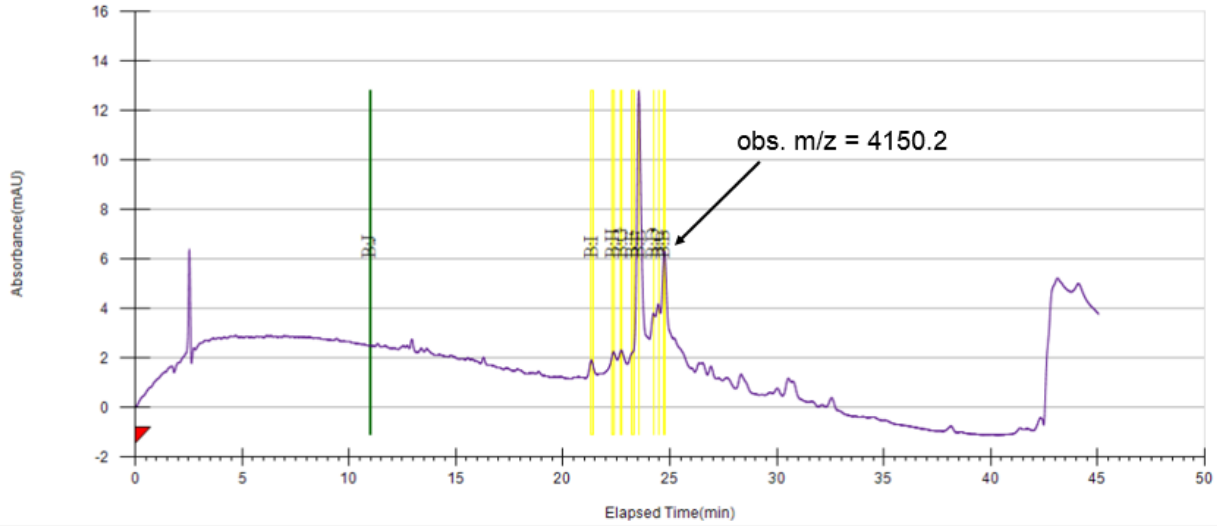


Figure A.31. Crude preparatory HPLC chromatogram for the solubility tagged P. aeruginosa Flp pilin N-terminus with a C-terminal hydrazide (Table A**.1**). The synthetic procedure employed for the C-terminal Flp pilin was used here. The C-terminal hydrazide was prepared with 2-chlorotrityl chloride resin as described by Liu et al.[9] Preparatory HPLC: column, Waters XBridge C8 (30.0 x 150 mm, 5 µm), flow rate 35 mL/min, gradient 10-60%B over 40 minutes, A solvent (H_2O, 0.1% v/v TFA), B solvent (ACN, HPLC grade, 0.1% v/v TFA). **[M+H]$_{mono}$** calculated 4151.3, observed 4093.2, 4036.3, 4150.2 (fraction corresponding to peak indicated by black arrow).

In addition to the Flp pilin, I also optimized the synthesis of the PIV5-TM domain with a solubilizing tag. The first iteration employed microwave-assisted synthesis and a 6xArg residue tag (Ac-RRRRRRβA). The purity and yield of this first attempt was poor, and subsequent attempts involving microwave-assisted synthesis also resulted in poor crude purity and a low yield. Ultimately I found that manual room temperature synthesis with 0.5-1 h double couplings allowed me to obtain a useable quantity of pure material (6.5 mg from a 200 µmol scale). The aim here was to generate a fusion peptide via native chemical ligation with GCN4-pII. The putative mechanism of PIV5 membrane fusion involves the formation of a PIV5-TM trimer in the pre-fusion complex. Previous work by

Bissonnette *et al.* has shown convincing biochemical evidence for the identity of residues that comprise the PIV5-TM trimer interface.[4] I endeavored to generate a fusion protein between PIV5-TM and the more soluble GCN4-pII (also a trimer).[4,14] In this hypothetical fusion protein the core Ile residues that comprise the GCN4-pII interface would be in register with the PIV5-TM core residues identified by Bissonnette *et al.* I believe this fusion peptide represents a promising candidate for a synthetic membrane fusion peptide with an enhanced propensity to crystallize. Furthermore, the PIV5-TM sequence has a native Cys residue; therefore, no post-ligation desulfurization is required. I was able to synthesize the GCN4-pII with a C-terminal hydrazide with a very high yield (~35 mg for a 200 µmol scale). In an initial ligation attempt, the C-terminal GCN4-pII hydrazide was readily oxidized to the thioester, but the 'solubility' tagged PIV5-TM was marginally soluble in the aqueous buffered ligation conditions (clear light scattering and precipitate). Further optimization may include dissolving the solubility tagged PIV5-TM domain in a minimal amount of organic solvent prior to dilution with the ligation buffer or varying the length and/or position of the solubility tag.

GCN4-pII hydrazide (includes N-terminal fragment of PIV5-TM sequence):

Ac-R(Nle)KQIEDKIEEILSKIYHIENEIARIKKLLSIIAI-CONHNH2

PIV5-TM (tag position underlined):

H-CLGSL**GL**ILII**LL**SVVVWKLL-CONH$_2$

Full length hypothetical fusion GCN4-pII fusion protein (heptad register bold/underlined):

Ac-R**(Nle)**KQ**I**EDK**I**EE**I**LSK**I**YH**I**ENE**I**AR**I**KKI**L**SI**I**AIC**L**GS**L**GLI**L**II**L**LSV**V**VWKLL-CONH$_2$

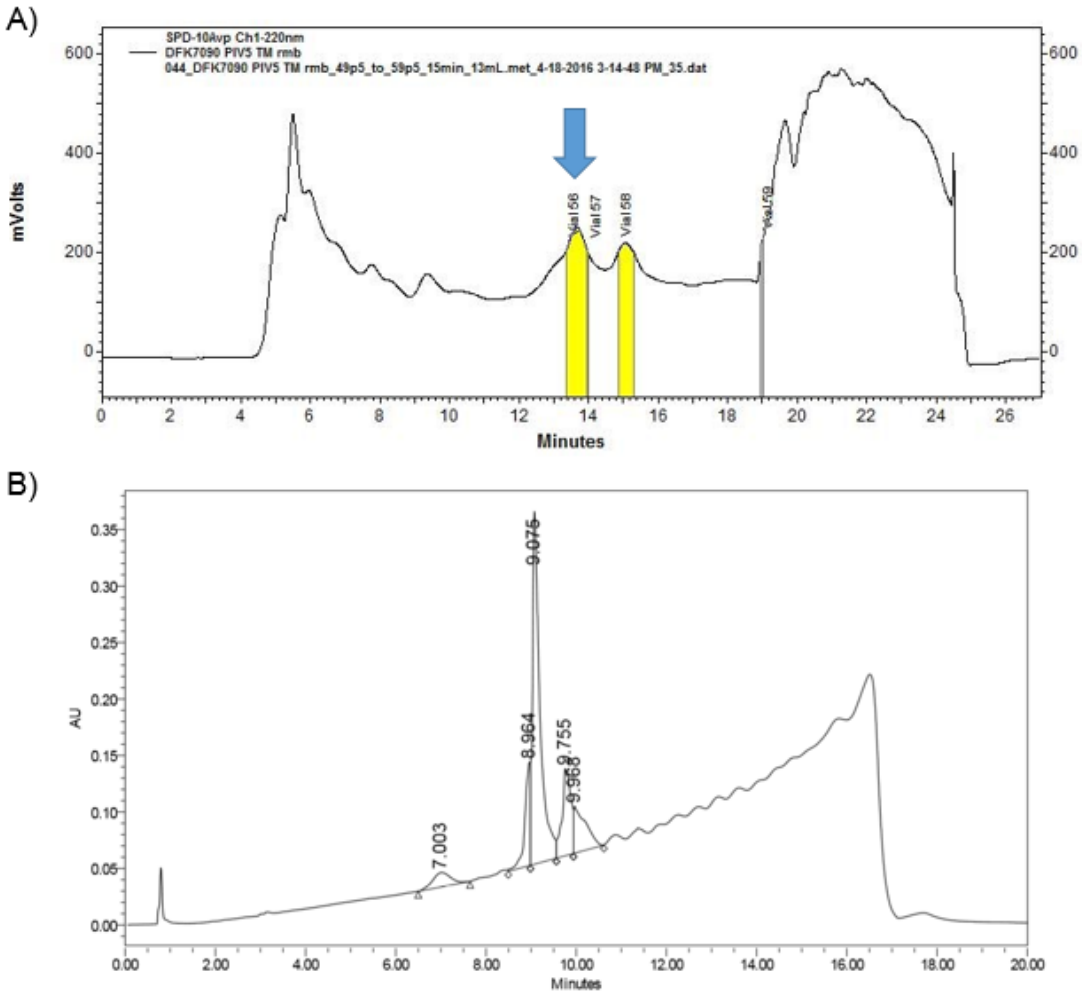
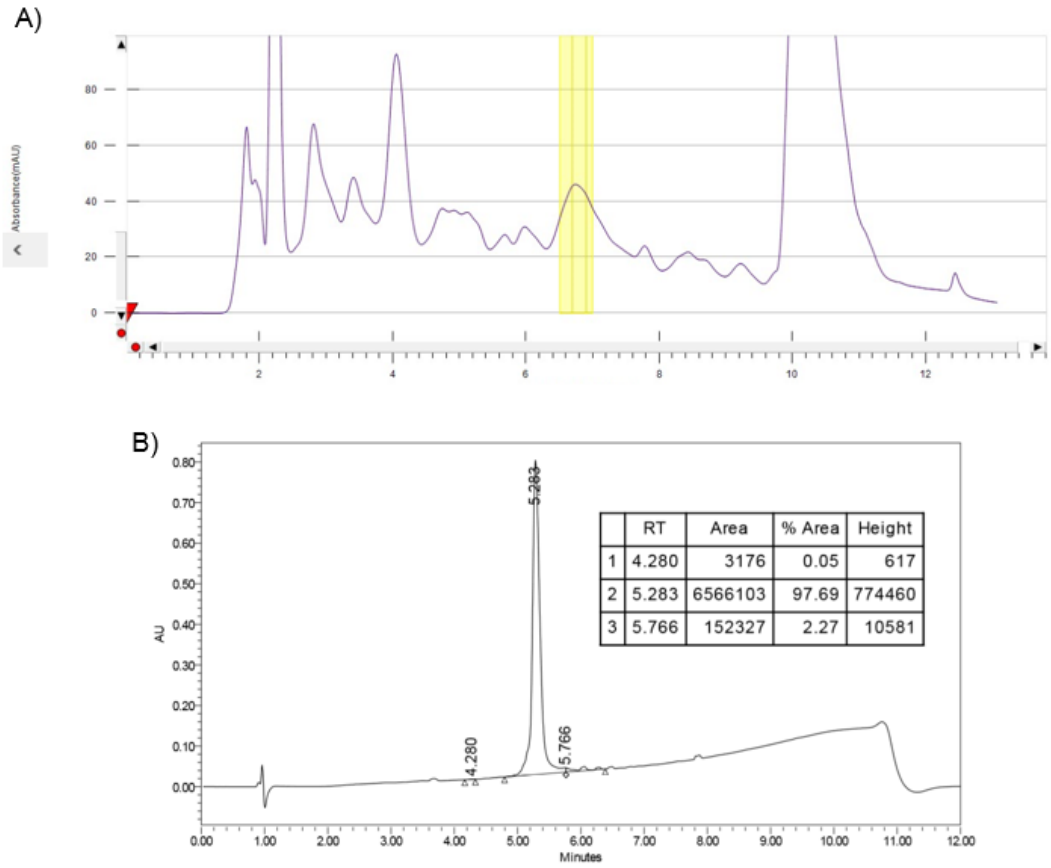


Figure A.32. A) Crude preparatory HPLC and B) UPLC purity analysis chromatograms for the solubility tagged PIV5-TM domain (Table A**.1**) synthesized under microwave-assisted conditions (same conditions as employed for the Flp pilin C-terminus except single couplings were used and the deprotection solution comprised 20% v/v piperidine in DMF). The solubility tag comprised Ac-RRRRRR(βAla) Preparatory HPLC: column, Supelco C5 (21.2 x 250 mm), flow rate 13 mL/min, gradient 49.5-59.5%B over 15 minutes, A solvent (H_2O, 0.1% v/v TFA), B solvent (ACN, HPLC grade, 0.1% v/v TFA). Analytical UPLC: column, Waters-Acquity Peptide BEH C18 (2.1 x 100 mm, 1.7 µm), flow rate 0.25 mL/min, gradient 10-95%B over 6 minutes, A solvent (H_2O, 0.1% v/v TFA), B solvent (ACN, HPLC grade, 0.1% v/v TFA). **$[M+H]_{mono}$** calculated 3664.2, observed 3662.0.



	RT	Area	% Area	Height
1	4.280	3176	0.05	617
2	5.283	6566103	97.69	774460
3	5.766	152327	2.27	10581

Figure A.33. A) Crude preparatory HPLC (yellow shading indicates peak area that was collected during purification) and B) UPLC purity analysis chromatograms for the solubility tagged PIV5-TM domain (Table A.**1**) synthesized with manual 0.5-1 h room temperature double couplings (Oxyma/COMU/TMP; DFK9002). The solubility tag comprised Ac-RRRRG. Preparatory HPLC: column, Waters Xbridge C8 (30 x 150 mm, 5 µm), flow rate 35 mL/min, gradient 35-43%B over 8 minutes, A solvent (H_2O, 0.1% v/v TFA), B solvent (ACN, HPLC grade, 0.1% v/v TFA). Analytical UPLC: column, Waters-Acquity Peptide BEH C18 (2.1 x 100 mm, 1.7 µm), flow rate 0.25 mL/min, gradient 10-95%B over 6 minutes, A solvent (H_2O, 0.1% v/v TFA), B solvent (ACN, HPLC grade, 0.1% v/v TFA). **[M+H]$_{mono}$** calculated 3181.0, observed 3181.0.

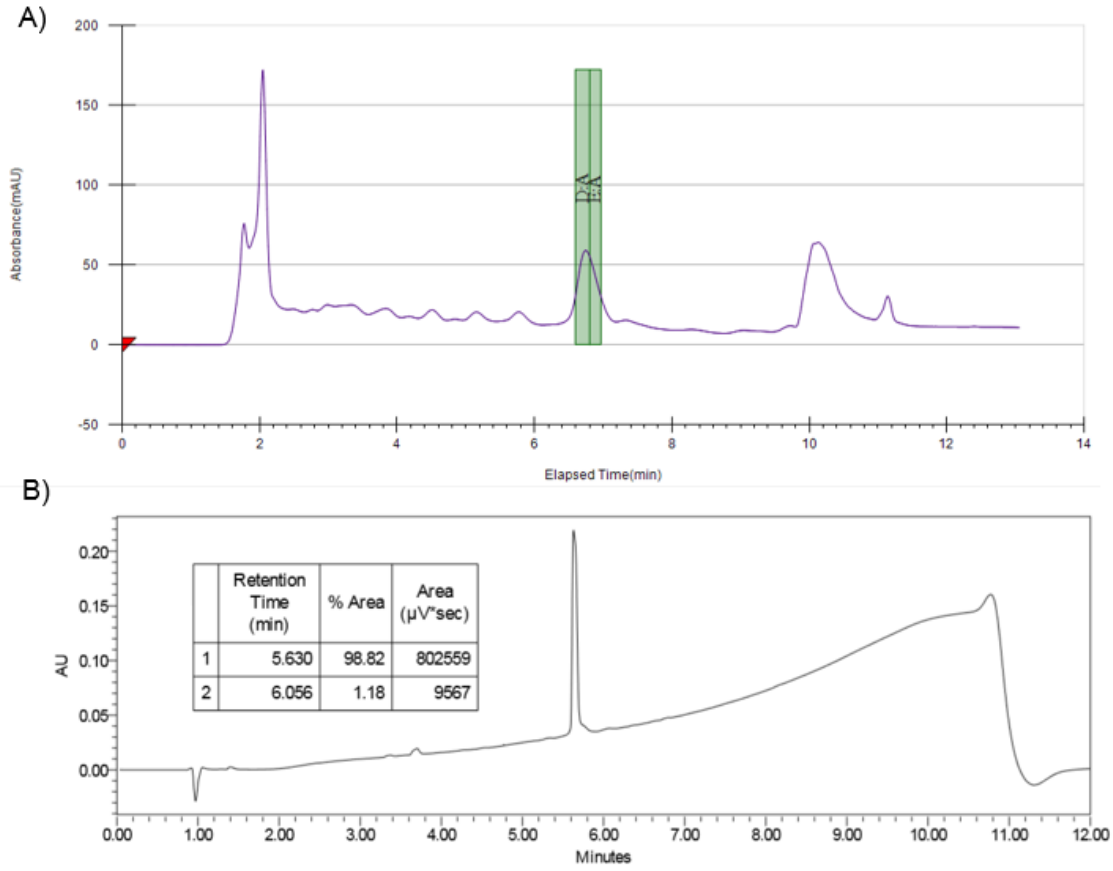


Figure A.34. A) Crude preparatory HPLC and B) UPLC purity analysis chromatograms for the solubility tagged GCN4-pII hydrazide (Table A**.1**) synthesized with manual 30 min coupling reactions (60°C oil bath; DIC/oxyma; DFK9027). Preparatory HPLC: column, Agilent C8 (30 x 150 mm), flow rate 35 mL/min, gradient 42-50%B over 8 minutes, A solvent (H_2O, 0.1% v/v TFA), B solvent (ACN, HPLC grade, 0.1% v/v TFA). Analytical UPLC: column, Waters-Acquity Peptide BEH C18 (2.1 x 100 mm, 1.7 µm), flow rate 0.25 mL/min, gradient 10-95%B over 6 minutes, A solvent (H_2O, 0.1% v/v TFA), B solvent (ACN, HPLC grade, 0.1% v/v TFA). **$[M+H]_{mono}$** calculated 4229.5, observed 4230.4.

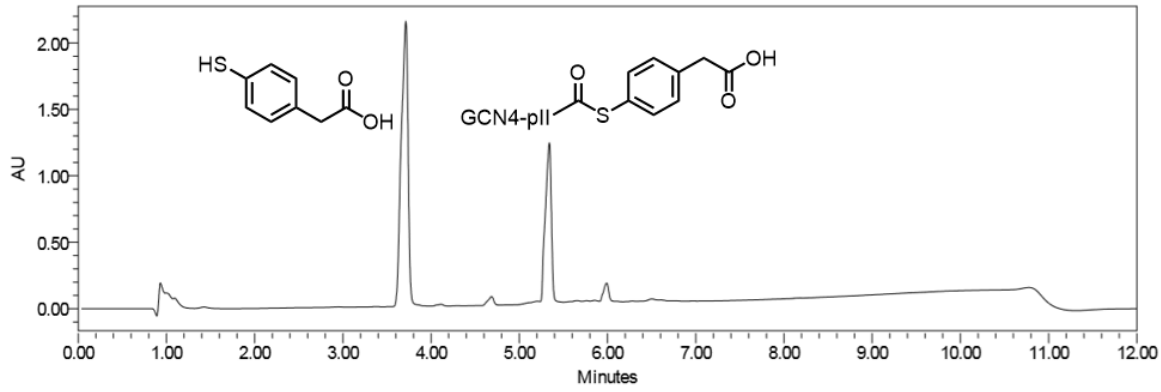


Figure A.35. Attempted ligation between PIV5-TM with an Ac-RRRRG solubility tag and GCN4-pII hydrazide (3 h time point). After 3 h there is no peak corresponding to the PIV5-TM peptide (indicating marginal solubility). After 5 h the only ion detectable via MALDI-TOF MS corresponds to the GCN4-pII/MPAA thioester (experimental details, DFK9033). **[M+H]**$_{mono}$ calculated 4665.5, observed 4666.3.

References:

(1) Lyons, J. A.; Shahsavar, A.; Paulsen, P. A.; Pedersen, B. P.; Nissen, P. *Curr. Opin. Struct. Biol.* **2016**, *38*, 137–144.

(2) Ostermeier, C.; Michel, H. *Methods Mol. Biol.* **2013**, *1033*, 697–701.

(3) Yao, H.; Lee, M. W.; Waring, A. J.; Wong, G. C. L.; Hong, M. *Proc. Natl. Acad. Sci. U. S. A.* **2015**, *112* (35), 10926–10931.

(4) Bissonnette, M. L. Z.; Donald, J. E.; DeGrado, W. F.; Jardetzky, T. S.; Lamb, R. A. *J. Mol. Biol.* **2009**, *386* (1), 14–36.

(5) Johnson, T.; Quibell, M.; Owen, D.; Sheppard, R. C. *J. Chem. Soc. Chem. Commun.* **1993**, 369–372.

(6) Johnson, T.; Quibell, M. *Tetrahedron Lett.* **1994**, *35* (3), 463–466.

(7) Quibell, M.; Turnell, W.; Johnson, T. *J. Org. Chem.* **1994**, *59* (16), 1745–1750.

(8) Cardona, V.; Eberle, I.; Barthelemy, S.; Beythien, J.; Doerner, B.; Schneeberger, P.; Keyte, J.; White, P. D. *Int. J. Pept. Res. Ther.* **2008**, *14* (4), 285–292.

(9) Zheng, J. S.; He, Y.; Zuo, C.; Cai, X. Y.; Tang, S.; Wang, Z. A.; Zhang, L. H.; Tian, C. L.; Liu, L. *J. Am. Chem. Soc.* **2016**, *138* (10), 3553–3561.

(10) Abdel-Aal, A. B. M.; Papageorgiou, G.; Raz, R.; Quibell, M.; Burlina, F.; Offer, J.

J. Pept. Sci. **2016**, *22* (5), 360–367.

(11) Steinkruger, J. D. Helical, Exploration of Sequence-Stability Relationships in Model Proteins Using Thioester Exchange and Lipid Cubic Mesophase Crystallization Attempts Using, 2012.

(12) Acharya, R.; Carnevale, V.; Fiorin, G.; Levine, B. G.; Polishchuk, A. L.; Balannik, V.; Samish, I.; Lamb, R. a; Pinto, L. H.; DeGrado, W. F.; Klein, M. L. *Proc. Natl. Acad. Sci. U. S. A.* **2010**, *107* (34), 15075–15080.

(13) Mortenson, D. E.; Steinkruger, J. D.; Kreitler, D. F.; Perroni, D. V; Sorenson, G. P.; Huang, L. *Proc. Natl. Acad. Sci.* **2015**, *112* (43), 13144–13149.

(14) Harbury, P. B.; Zhang, T.; Kim, P. S.; Alber, T. *Science* **1993**, *262* (5138), 1401–1407.

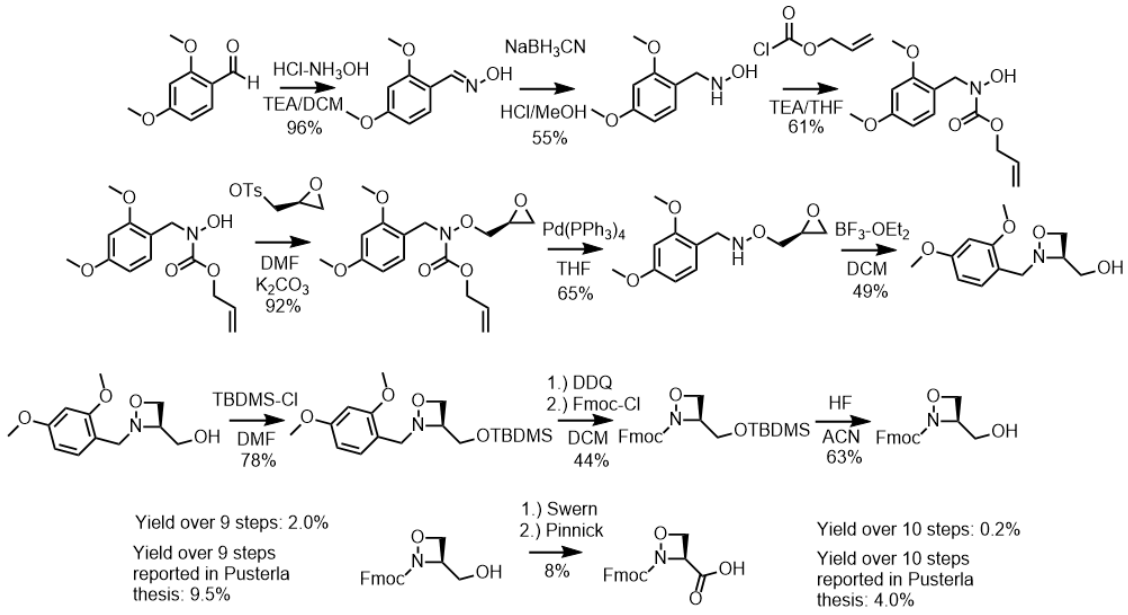
Appendix B Synthesis of Fmoc-oxazetidine Carboxylic Acid and the Impact of Fmoc-oxazetidine Incorporation on the HPLC Purification of the Parainfluenza Virus 5 Transmembrane Domain

Synopsis

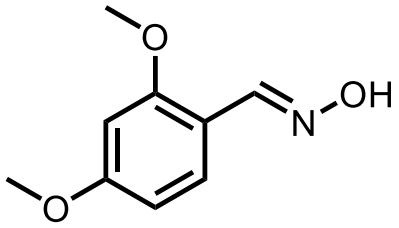
A ketoacid hydroxylamine (KAHA) ligation strategy has been reported that employs a reactive oxazetidine monomer, which results in a serine residue upon native amide bond formation.[1,2] I hypothesized that this ligation strategy would be useful for obtaining transmembrane peptides fused with soluble peptide domains to enhance the likelihood of transmembrane peptide crystallogenesis. At face value this approach is attractive for two reasons. The first reason is that the reactivity of the oxazetidine monomer, presumably due to ring strain, would allow for facile ligation reactions at low concentration. This reactivity at low concentration would be useful for membrane peptides that are marginally soluble (even in fluorinated organic solvents such as trifluoroethanol or hexafluorisopropanol). The second reason is that this iteration of the KAHA ligation strategy requires no covalent post-ligation modification of the final product. Ultimately, I determined that the synthetic effort required to obtain the necessary quantities of the Fmoc-protected oxazetidine monomer for solid-phase peptide synthesis was too great to make the oxazetidine KAHA ligation a viable strategy for obtaining the quantity of peptide (>10 mg) required for crystallization studies. Furthermore, I found that in one case, the parainfluenza virus 5 transmembrane domain (PIV5-TM), incorporation of the Fmoc-oxazetidine monomer significantly decreased peptide solubility in trifluoroethanol, thus nullifying one of the main attractive features of this ligation strategy.

Here, I report my efforts to reproduce the 10 step synthesis of the Fmoc-oxazetidine carboxylic acid initially described by Pusterla and Bode (Scheme B.**4**).[1,2] I

made modifications to purification steps reported by Pusterla and Bode that employ cyclohexane, which is an economically prohibitive solvent for the scale required to generate useful quantities of the Fmoc-oxazetidine monomer. In addition, I have altered the last purification step from an HPLC method to an MPLC method that could likely be reproduced with the proper stationary phase and a standard flash column. The disparity in final yields between my efforts and those of Pusterla and Bode are predominately due to the second and third steps to obtain the Dmb-hydroxamic acid and the final two-step oxidation. In steps where I modified intermediate purification conditions, the yields were mostly reproducible, with the exception of the final purification. The diminished yield on the final step relative to the yield reported by Pusterla appeared to be unrelated to the purification and mainly due to side-reactions that occurred during the two-step oxidation.



Scheme B.4. The 10 step synthesis of the Fmoc-protected oxazetidine monomer reported by Pusterla and Bode.[1,2]



Dmb-aldoxime

Formula: $C_9H_{11}NO_3$

***IUPAC*:** N-[(2,4-dimethoxyphenyl)methylidene]hydroxylamine

***SMILES*:** COc1ccc(C=NO)c(OC)c1

Procedure: A 500 mL round bottom flask was charged with 2,4-dimethoxybenzaldehyde (50.0 g, 301 mmol, 1.0 eq.) and hydrochloride hydroxylamine (25.1 g, 361 mmol, 1.2 eq.). Prolonged storage of hydrochloride hydroxylamine can result in a clumpy solid. Therefore to ensure proper dissolution, hydrochloride hydroxylamine clumps were ground into a fine powder with a steel spatula. The contents of the flask were stirred in DCM (180 mL) to generate a heterogeneous suspension. The flask was placed on ice while stirring and allowed to cool for 10 minutes. Next, triethylamine (92 mL, 660 mmol, 2.2 eq.) was added to the flask dropwise while the suspension was stirring (the suspension became viscous upon addition of each portion of triethylamine, but fluidity was maintained with continuous stirring). After complete addition of triethylamine the flask was allowed to warm to room temperature, sealed with a rubber septum, and stirred for approximately 17 h. At the conclusion of this time, the flask was diluted to 500 mL with DCM and transferred to a 1 L separatory funnel. The organic phase was washed with saturated aqueous ammonium chloride (100 mL). The organic phase was subsequently washed with brine (100 mL); dried over sodium sulfate and concentrated in a 500 mL round bottom flask to form an

off-white solid. A reflux condenser was attached to the flask and the solid was then triturated in boiling pentanes (300 mL) for 1 h, vacuum filtered, and washed with room temperature pentanes. The filtrate was collected and dried for 2 h at 50°C under vacuum to afford 52.7 g of a white powder (96% yield).

^{1}H NMR (400 MHz, $CDCl_3$): δ = 9.12 (s, 1H), 8.38 (s, 1H), 7.56 (d, *J*=8.5 Hz, 1H), 6.49 (m, 2H), 3.84 (s, 3H), 3.83 (s, 3H)

^{13}C NMR (400 MHz, $CDCl_3$): δ = 162.35, 158.95, 146.53, 128.58, 113.66, 105.25, 98.46, 55.54, 55.46

ESI-MS [M+H]: *calculated* 182.0739, *observed* 182.0811

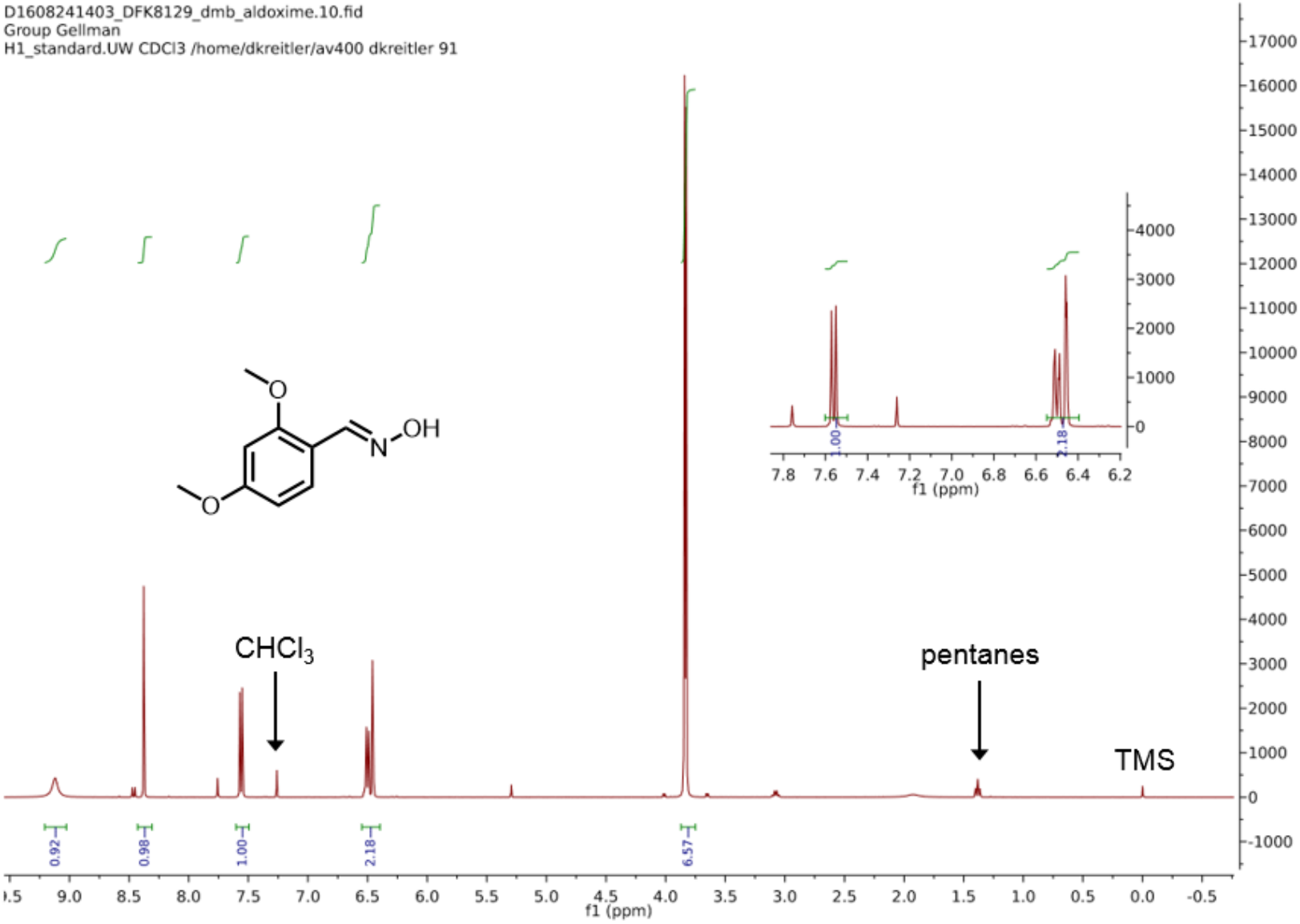


Figure B.1. ^{1}H NMR spectrum of Dmb-aldoxime.

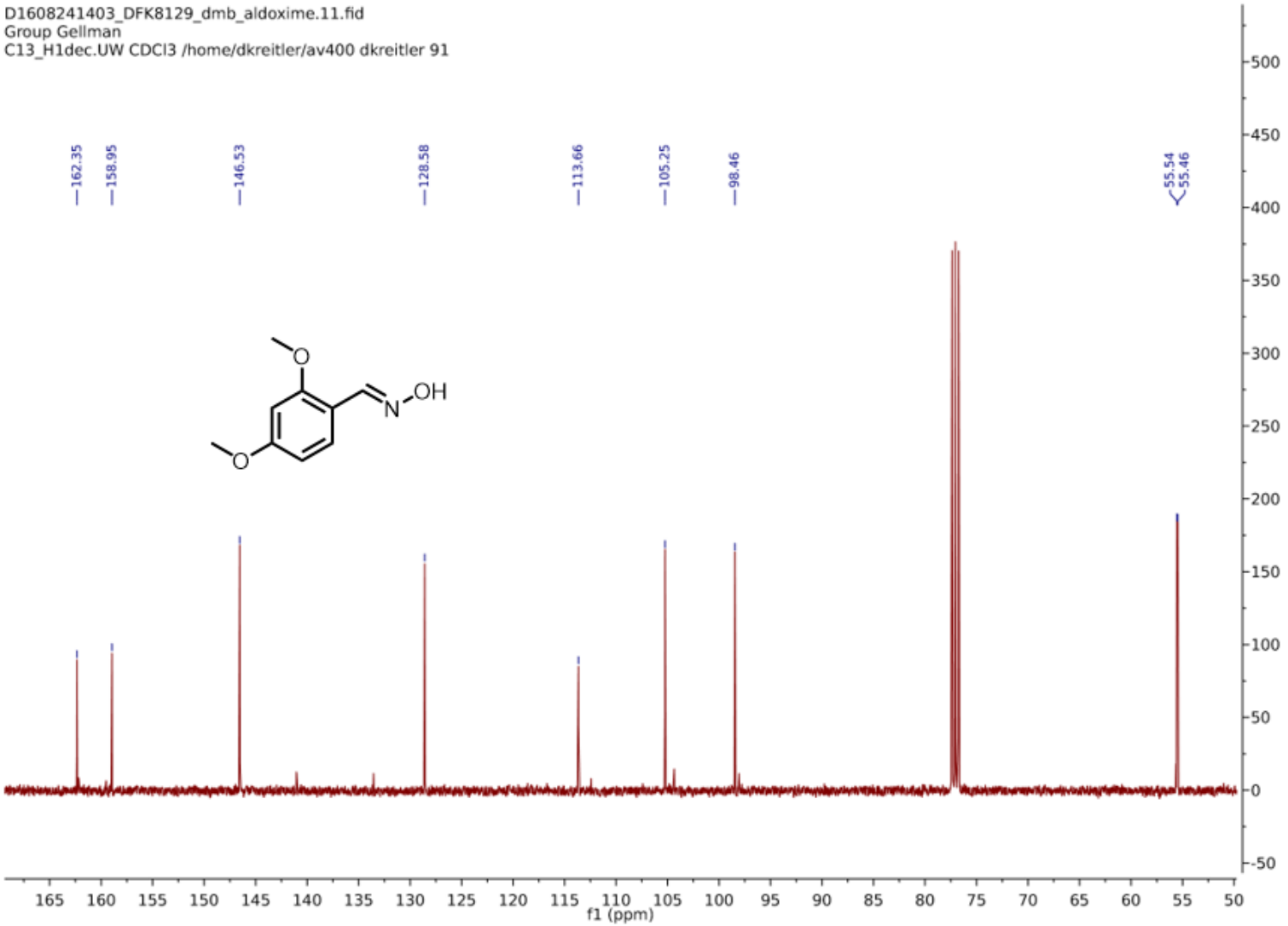
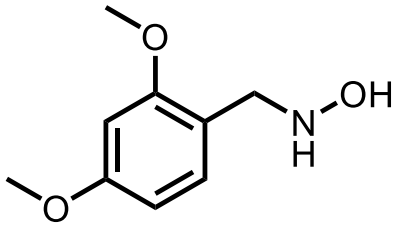


Figure B.2. ^{13}C NMR spectrum of Dmb-aldoxime.



Dmb-hydroxylamine

Formula: $C_9H_{13}NO_3$

***IUPAC*:** N-[(2,4-dimethoxyphenyl)methyl]hydroxylamine

***SMILES*:** COc1ccc(CNO)c(OC)c1

Procedure: A 1 L round bottom flask was charged with N-[(2,4-dimethoxyphenyl)methylidene]hydroxylamine (52.7 g, 290 mmol, 1.0 eq.). Anhydrous MeOH (540 mL, 0.55 M) was added to the flask and the aldoxime was dissolved with vigorous stirring and sonication. Approximately 20 mg of bromocresol green was added and the flask was positioned in a room temperature water bath. A 250 mL addition funnel was placed on top of the flask and filled with a 2N HCl/MeOH solution*. The 2N HCl/MeOH solution was added to the stirring contents of the flask dropwise at a rate of approximately 1 drop/second. Initially, the flask contents were blue (indicating an alkaline pH) but after several minutes of dropwise addition of methanolic HCl the color gradually changed to green and then yellow, after which dropwise addition of the HCl solution was slowed to approximately 0.5 drops/second to maintain the yellow color. After approximately 200 mL of the methanolic solution had been added, dropwise addition was halted and the yellow color of the solution remained for 15 minutes. At this point the addition funnel was removed and replaced with a rubber septum vented with a nitrogen stream, and the reaction was stirred for an additional 3 h at room temperature. After 3 h the pH of the

solution was adjusted to approximately 9-10 (as indicated by pH paper) with a 5 M aq. NaOH solution. The MeOH was then evaporated off with a rotary evaporator to afford a blue oil (due to residual bromocresol green). The oil was then partitioned between DCM (300 mL) and an aqueous saturated K_2CO_3 solution (300 mL). The phases were separated and the aqueous layer was extracted with DCM (2x200 mL). The combined organic layers were then washed with brine (200 mL); dried over sodium sulfate; filtered through celite to remove any residual bromocresol green; and concentrated via rotary evaporation to afford a yellow oil that formed a yellow solid under vacuum. The yellow solid was scraped from the flask and transferred to a mortar where it was ground into a fine powder. The yellow powder was then triturated in anhydrous diethyl ether at room temperature (the heterogeneous suspension was stirred vigorously with a steel spatula). Particulates were allowed to settle for 1 h at -20°C after which the ether was decanted and the resulting product allowed to dry under a nitrogen stream to afford 29.4 g of a yellow solid (55% yield).

*Methanolic HCl was generated by dripping approximately 100 mL of H_2SO_4 from a 250 mL addition funnel into 100 mL of concentrated aq. HCl in a 250 mL round bottom flask. A vacuum adapter was used to connect the addition funnel to the round bottom flask. A length of tygon tubing connected to the adapter and containing a Pasteur pipet in the end was used to bubble HCl gas through MeOH.

^{1}H NMR (400 MHz, $CDCl_3$): δ = 7.17 (d, *J*=7.9 Hz, 1H), 6.46 (m, 2H), 4.01 (s, 2H), 3.82 (s, 3H), 3.81 (s, 3H)

^{13}C NMR (400 MHz, $CDCl_3$): δ = 160.82, 159.01, 132.04, 117.33, 103.99, 98.65, 55.50, 55.38, 53.74

ESI-MS [M+H]: *calculated* 184.09, *observed* 184.0965

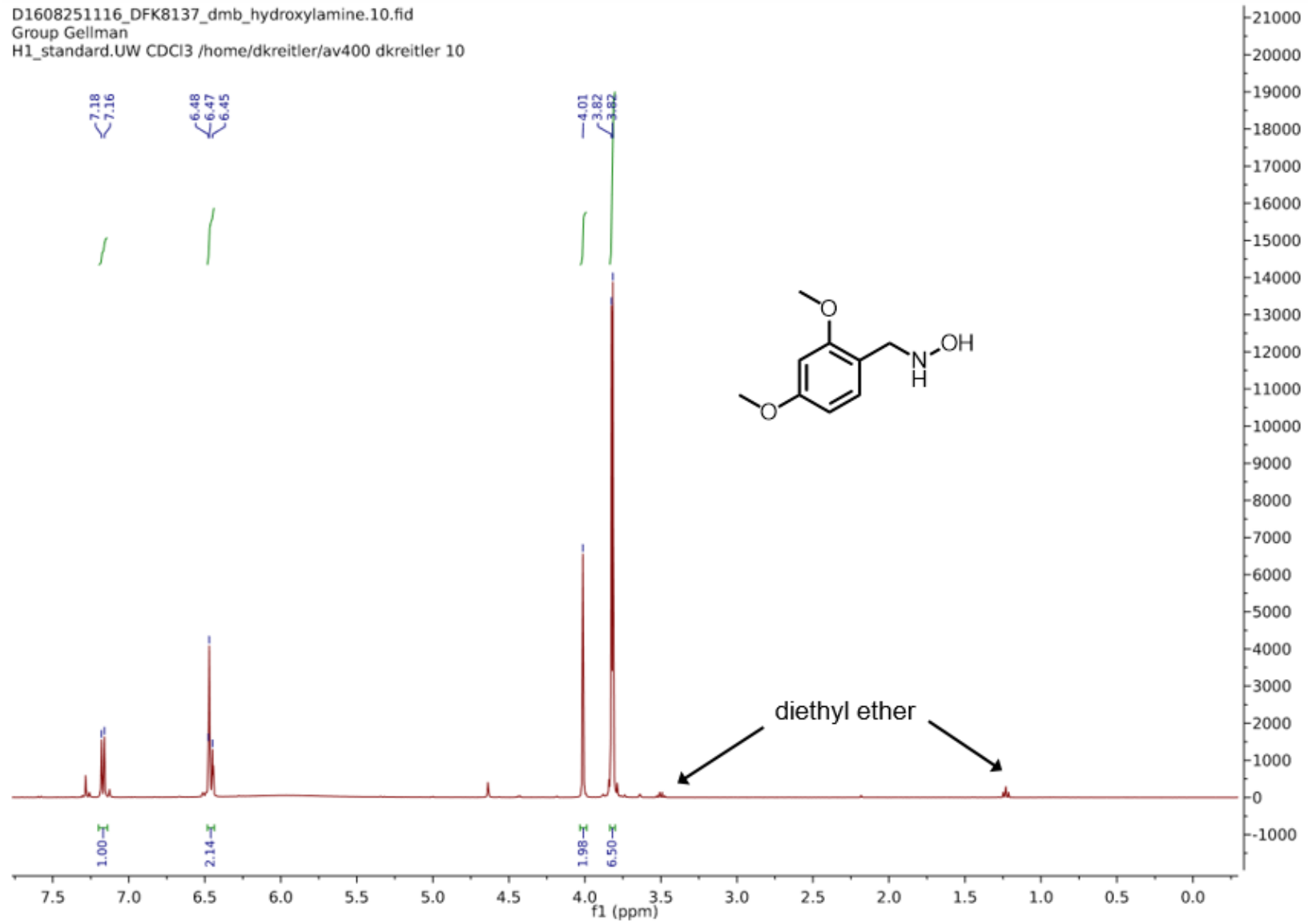


Figure B.3. ^{1}H NMR spectrum of Dmb-hydroxylamine.

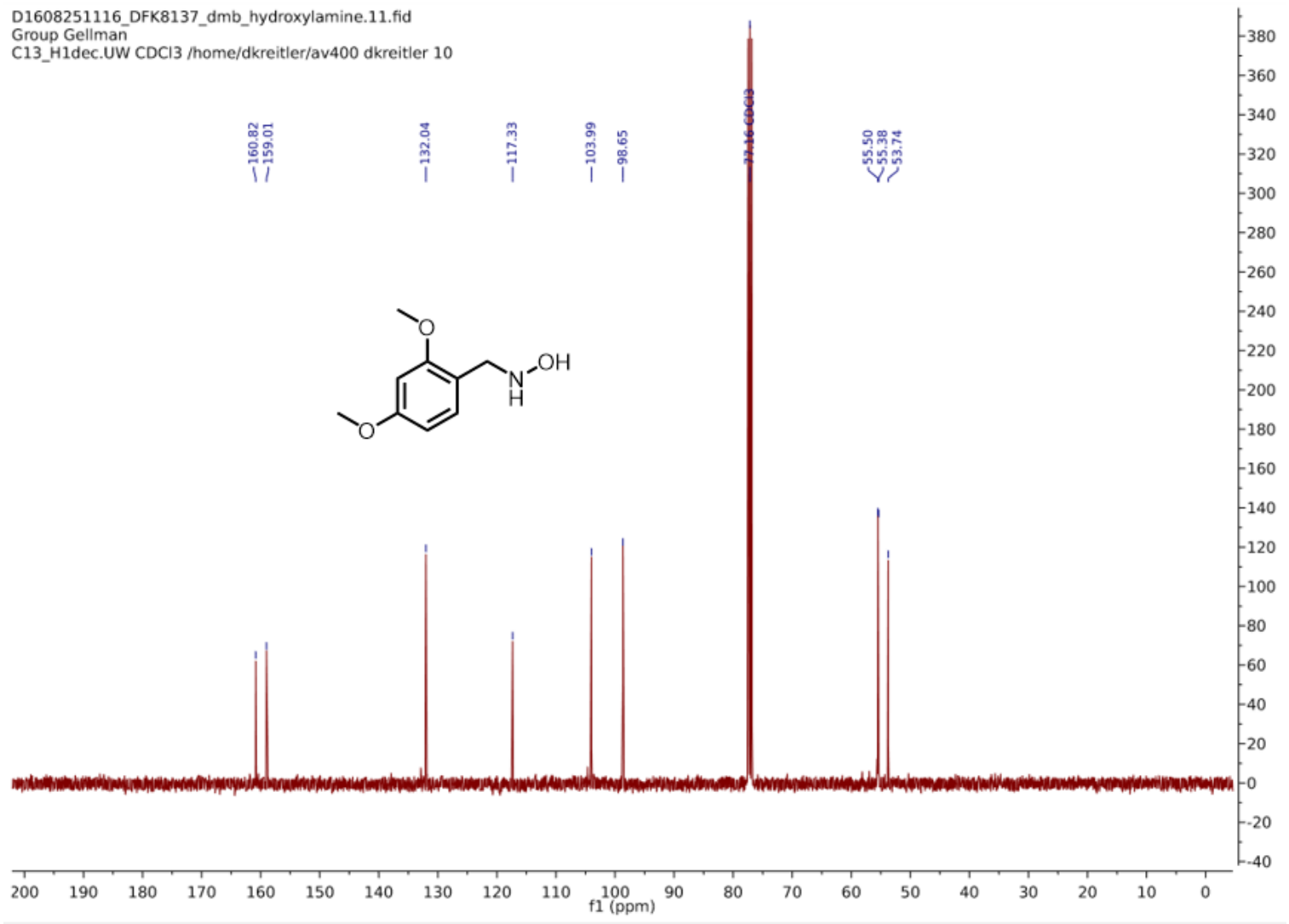
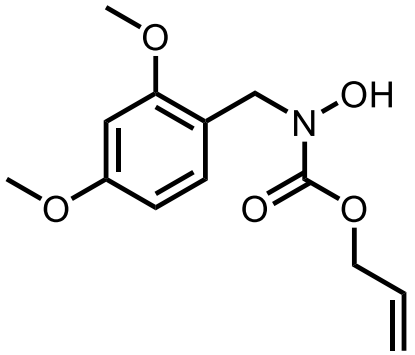


Figure B.4. ^{13}C NMR spectrum of Dmb-hydroxylamine.



Dmb-hydroxamic acid

Formula: $C_{13}H_{17}NO_5$

***IUPAC*:** prop-2-en-1-yl N-[(2,4-dimethoxyphenyl)methyl]-N-hydroxycarbamate

***SMILES*:** C=CCOC(=O)N(O)Cc1ccc(OC)cc1OC

Procedure: A 1 L round-bottom flask was charged with N-[(2,4-dimethoxyphenyl)methyl]hydroxylamine (29.4 g, 160.5 mmol, 1.0 eq.). The solid was suspended in dry THF (0.5 M, 320 mL, 3 Å sieves). While stirring, triethylamine (33.5 mL, 240.8 mmol, 1.5 eq.) was added dropwise and the flask was cooled on ice for 15 min. Allyl chloroformate (17.9 mL, 168.5 mmol, 1.05 eq.) was diluted in dry THF (50 mL, 3 Å sieves) in a 250 mL addition funnel which was then connected to the top of the stirring round bottom flask. The allyl chloroformate solution was added to the flask dropwise over 15 min, after which the addition funnel was removed and replaced with a rubber septum connected to a nitrogen stream. The ice bath was removed and the flask was stirred at room temperature for 1.5 h, after which 5% w/v aq. NaOH solution (75 mL) was added to the flask and which was stirred for an additional 30 min. The reaction was then quenched with saturated aqueous ammonium chloride (100 mL). The phases were separated and the aqueous layer was extracted with DCM (2x200 mL). The organic layers were then combined; washed with brine; dried over sodium sulfate; and concentrated to afford a

thick brown oil. A 6 cm inner diameter column was packed with 500 mL silica P60 and equilibrated with 25% v/v EtOAc in hexanes. Initially, 1 L of the equilibration solvent was used to pack the column and the excess solvent was collected and recycled. The crude oil was loaded directly onto the silica bed via 12 mL polypropylene syringe. After which the recycled equilibration solvent was flowed through the column followed by 1.5 L of 40% v/v EtOAc in hexanes. Fractions containing the product (R_f=0.35, 1:1 EtOAc/hexanes, $KMnO_4$ stain) were pooled and concentrated to afford 26.2 g of a thick yellow oil (61% yield).

^{1}H NMR (400 MHz, $CDCl_3$): δ = 7.18 (m, 1H), 6.60 (s, 1H), 6.46 (m, 2H), 5.93 (ddt, *J*=17.3, 10.8, 5.6 Hz, 1H), 5.32 (dd, *J*=17.1, 1.6 Hz, 1H), 5.23 (dd, *J*=10.5, 1.3 Hz, 1H), 4.01 (s, 2H), 4.71 (s, 2H), 4.64 (dt, *J*=5.6, 1.5 Hz, 2H), 3.81 (s, 3H), 3.80 (s, 3H)

^{13}C NMR (400 MHz, $CDCl_3$): δ = 160.64, 158.43, 156.82, 132.40, 130.61, 118.05, 116.63, 104.11, 98.62, 66.80, 55.51, 55.41, 49.48

ESI-MS [M+H]: *calculated* 268.1180, *observed* 268.1175

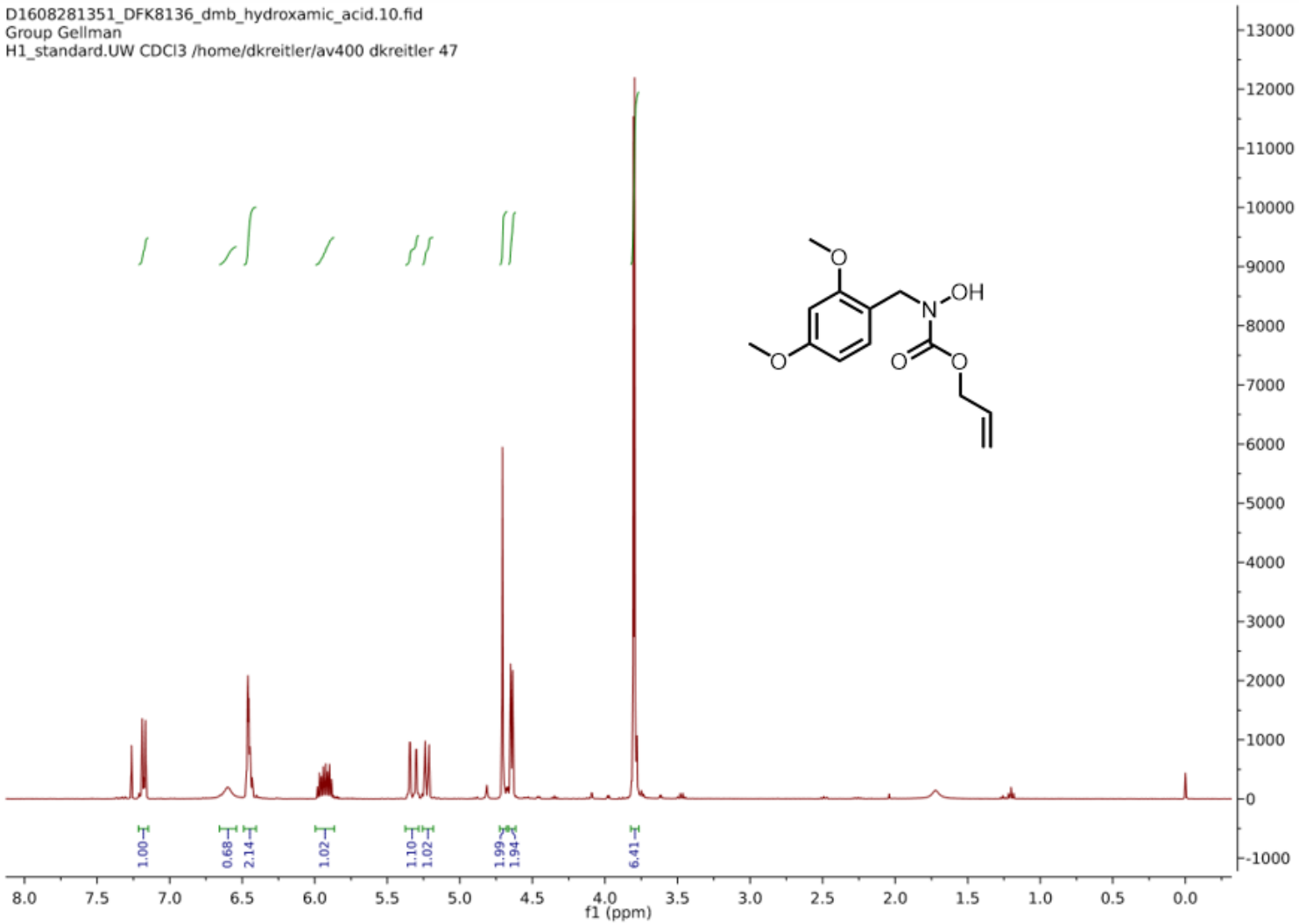


Figure B.5. ^{1}H NMR spectrum of Dmb-hydroxamic acid.

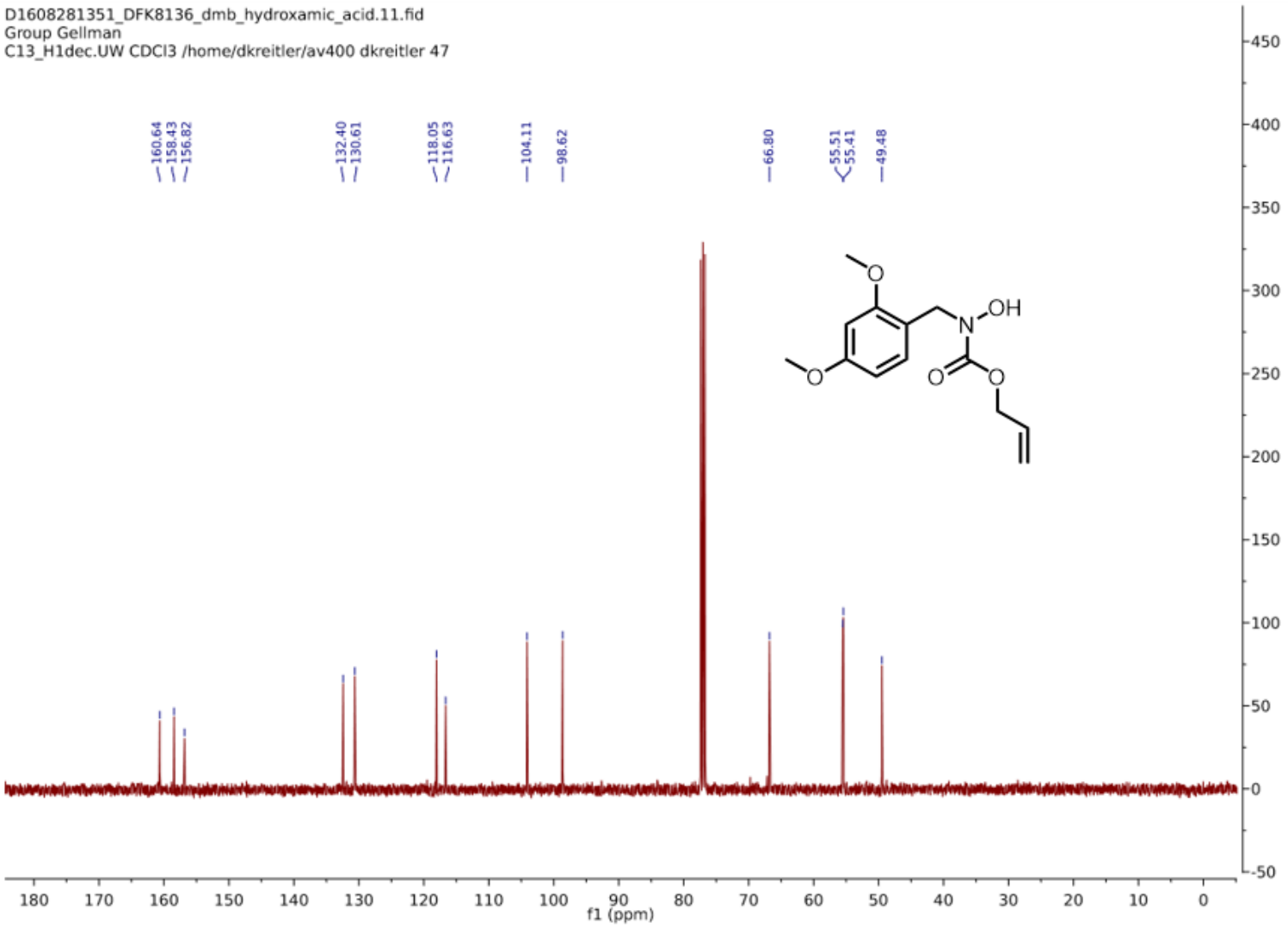
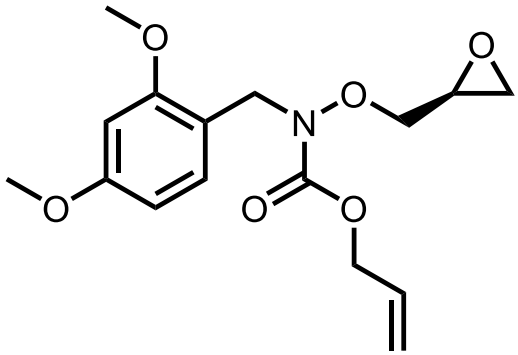


Figure B.6. ^{13}C NMR spectrum of Dmb-hydroxamic acid.



Dmb-hydroxylamine(alloc) epoxide

Formula: **$C_{16}H_{21}NO_6$**
***IUPAC*: prop-2-en-1-yl N-[(2,4-dimethoxyphenyl)methyl]-N-[(2S)-oxiran-2-ylmethoxy]carbamate**

***SMILES*: C=CCOC(=O)N(Cc1ccc(OC)cc1OC)OC[C@@H]2CO2**

Procedure: Prop-2-en-1-yl N-[(2,4-dimethoxyphenyl)methyl]-N-hydroxycarbamate (26.2 g, 1.0 eq., 98.0 mmol) was dissolved in dry DMF (3 Å molecular sieves, 0.65 M, 150 mL) in a flame-dried 500 mL round bottom flask. While stirring at room temperature solid K_2CO_3 (28.5 g, 2.1 eq., 205.0 mmol) and (2*S*)-glycidyl tosylate (Chem-Impex Intl., 21.2 g, 0.95 eq., 93.1 mmol) were added in the order listed. The flask was then sealed with a rubber septum and stirred for 36 h at 60°C (oil bath). The reaction was monitored over the course of the 36 h time period with ^{1}H-NMR (*vide infra*; the tosylate exhibits a signal at 2.45 ppm that can be conveniently integrated to determine consumption of starting material. Similarly, the oxirane protons are shifted slightly upfield once alkylation occurs and can also be used to monitor reaction progress). The tosylate was used as the limiting reagent due to the fact that the tosylate and the product are extremely difficult to resolve with silica chromatography. At the conclusion of 36 h the heterogeneous reaction mixture was partitioned between brine (300 mL; aqueous, saturated) and diethyl ether (300 mL). The phases were separated and the aqueous layer was extracted with diethyl

ether (2x200 mL). The organic layers were combined and washed with aqueous 5% w/v LiCl (5x100 mL), water (2x100 mL), and brine (2x100 mL). The organic layer was then dried over sodium sulfate and evaporated to afford 29.1 g of a thick, yellow oil (92% yield) which was used directly in the next step.

***crude NMR sample preparation:** Approximately 0.5 mL of the reaction mixture was transferred via pipet into a 25 mL glass vial that contained 5 mL diethyl ether and 5 mL brine (saturated, aqueous). The ether layer was then filtered through a glass pipet containing a sodium sulfate and glass wool plug. The ether was evaporated and the crude oil was dissolved in $CDCl_3$ and re-filtered through a fresh sodium sulfate plug. The resulting sample was then subjected to ^{1}H-NMR analysis.

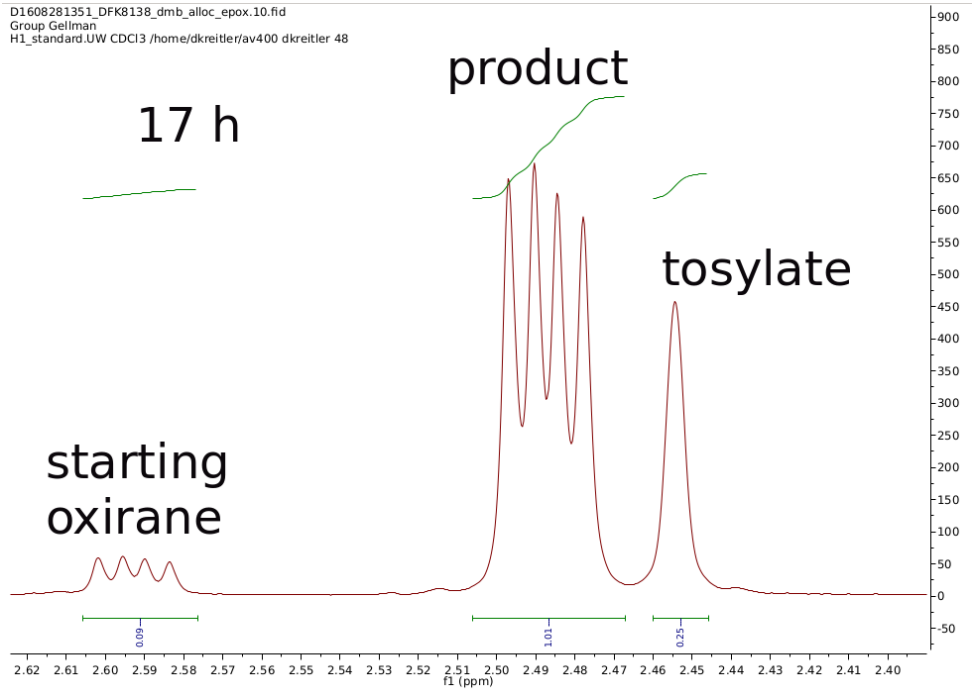


Figure B.7. Monitoring reaction progress as quantified by (2*S*)-glycidyl tosylate consumption.

^{1}H NMR (400 MHz, $CDCl_3$): δ = 7.24-7.14 (m, 1H), 6.50-6.38 (m, 2H), 5.95 (ddt, *J*=17.4, 10.7, 5.5 Hz, 1H), 5.34 (d, *J*=17.3 Hz), 5.23 (d, *J*=10.5 Hz), 4.73-4.61 (m, 4H), 3.81 (s, 3H), 3.80 (s, 3H), 3.09 (m, 1H), 2.73 (t, *J*=4.6 Hz, 1H), 2.48 (dd, *J*=5.1, 2.6 Hz)

^{13}C NMR (400 MHz, $CDCl_3$): δ = 160.65, 158.64, 157.16, 132.34, 130.84, 117.93, 116.68, 104.01, 98.42, 76.42, 66.64, 55.39, 55.35, 49.00, 48.18, 44.63

ESI-MS [M+H]: *calculated* 324.1442, *observed* 324.1437

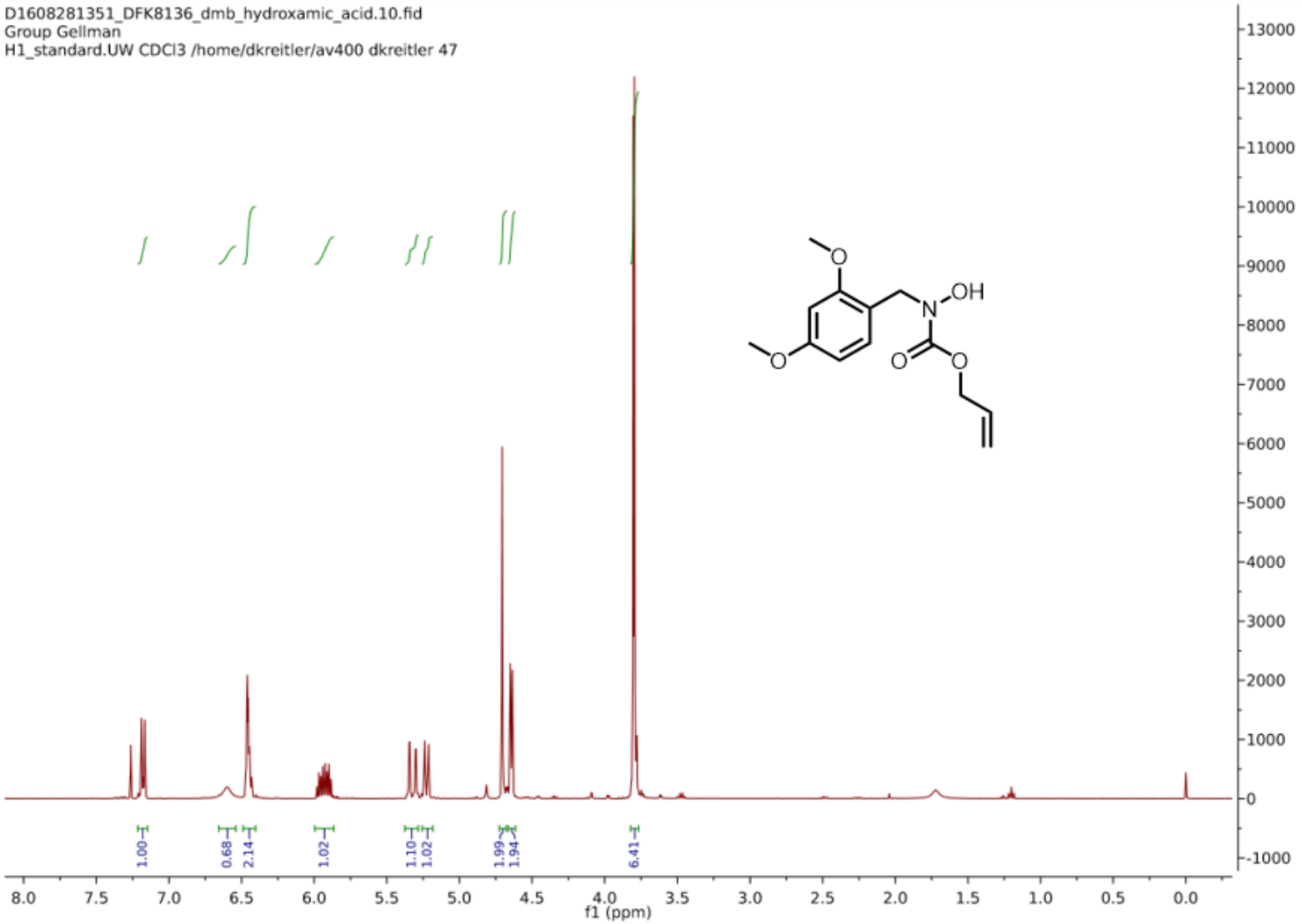


Figure B.8. ^{1}H NMR spectrum of Dmb-hydroxylamine(alloc) epoxide.

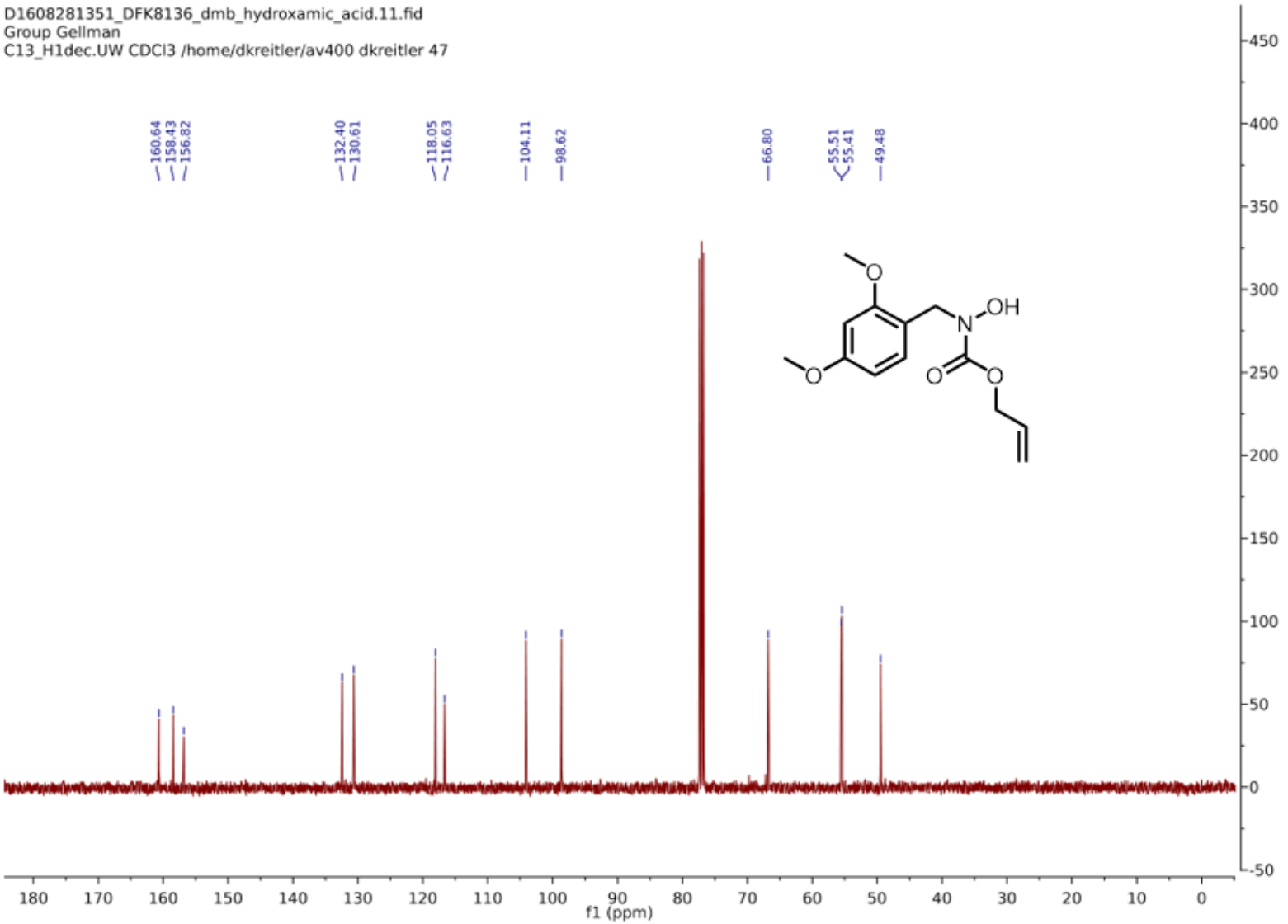
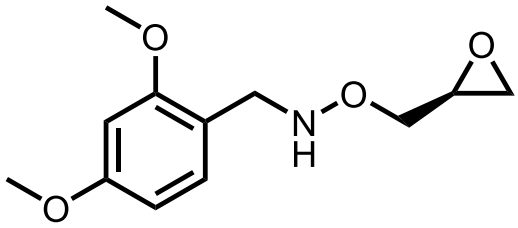


Figure B.9. ^{13}C NMR spectrum of Dmb-hydroxylamine(alloc) epoxide.



Dmb-hydroxylamine epoxide

Formula: $C_{12}H_{17}NO_4$

***IUPAC*:** [(2,4-dimethoxyphenyl)methyl][(2S)-oxiran-2-ylmethoxy]amine

***SMILES*:** COc2ccc(CNOC[C@@H]1CO1)c(OC)c2

Procedure: prop-2-en-1-yl N-[(2,4-dimethoxyphenyl)methyl]-N-[(2S)-oxiran-2-ylmethoxy]carbamate (29.1 g, 90.0 mmol, 1.0 eq.) was dissolved in dry THF (3 Å molecular sieves, 0.2 M, 450 mL) in a 1 L round bottom flask. While stirring, morpholine (15.5 mL, 180 mmol, 2.0 eq.) was added and dry N_2 was bubbled through the solution for 30 min. $Pd(PPh_3)_4$ (312 mg, 0.27 mmol, 0.003 eq.) was added to the flask in one lot and the solution was stirred for 1 h. The flask containing the reaction was then concentrated by rotary evaporation (bath temperature set at 30°C) to afford a crude yellow oil. The crude oil was then divided into six aliquots and purified via dry column vacuum chromatography (*vide infra*) with 75% v/v EtOAc in hexanes to resolve the product from a more polar impurity. Fractions containing the product were concentrated to afford 14.2 g of a yellow oil (65% yield).

Dry column vacuum chromatography: Sand was layered in a 300 mL fritted glass funnel in an amount sufficient to cover the frit. Next, the funnel was placed on top of a 1 L vacuum filter flask connected to a water aspirator and silica P60 was layered in the funnel, leaving

enough room at the top for approximately 150 mL of solvent. Solvent was stirred in with the silica and the silica allowed to settle to form an even surface. Solvent was then pulled through the funnel with the vacuum line. Then, the crude oil was layered on the dry silica surface and covered with sand. Solvent was then gently poured on top of the sand bed to fill the remaining void and a fraction was collected by connecting the vacuum line. This process was repeated and the purity of fractions was established with TLC (50% EtOAc in hexanes, $KMnO_4$ stain).

^{1}H NMR (400 MHz, $CDCl_3$): δ = 7.19-7.13 (m, 1H), 6.46-6.41 (m, 2H), 6.07 (s, 1H), 4.04(s, 2H), 3.90 (dd, *J*=12.1, 3.4 Hz, 1H), 3.81 (s, 3H), 3.79 (s, 3H), 3.64 (dd, *J*=12.1, 6.0 Hz, 1H), 3.18 (ddt, *J*=6.2, 4.2, 3.0, 3.0, 1H), 2.79 (dd, *J*=5.1, 4.2 Hz, 1H), 2.59 (dd, *J*=5.0, 2.7 Hz, 1H),

^{13}C NMR (400 MHz, $CDCl_3$): δ = 160.59, 158.77, 131.65, 117.59, 103.94, 98.54, 74.58, 55.37, 55.32, 51.73, 50.30, 44.76

ESI-MS [M+H]: *calculated* 240.1230, *observed* 240.1226

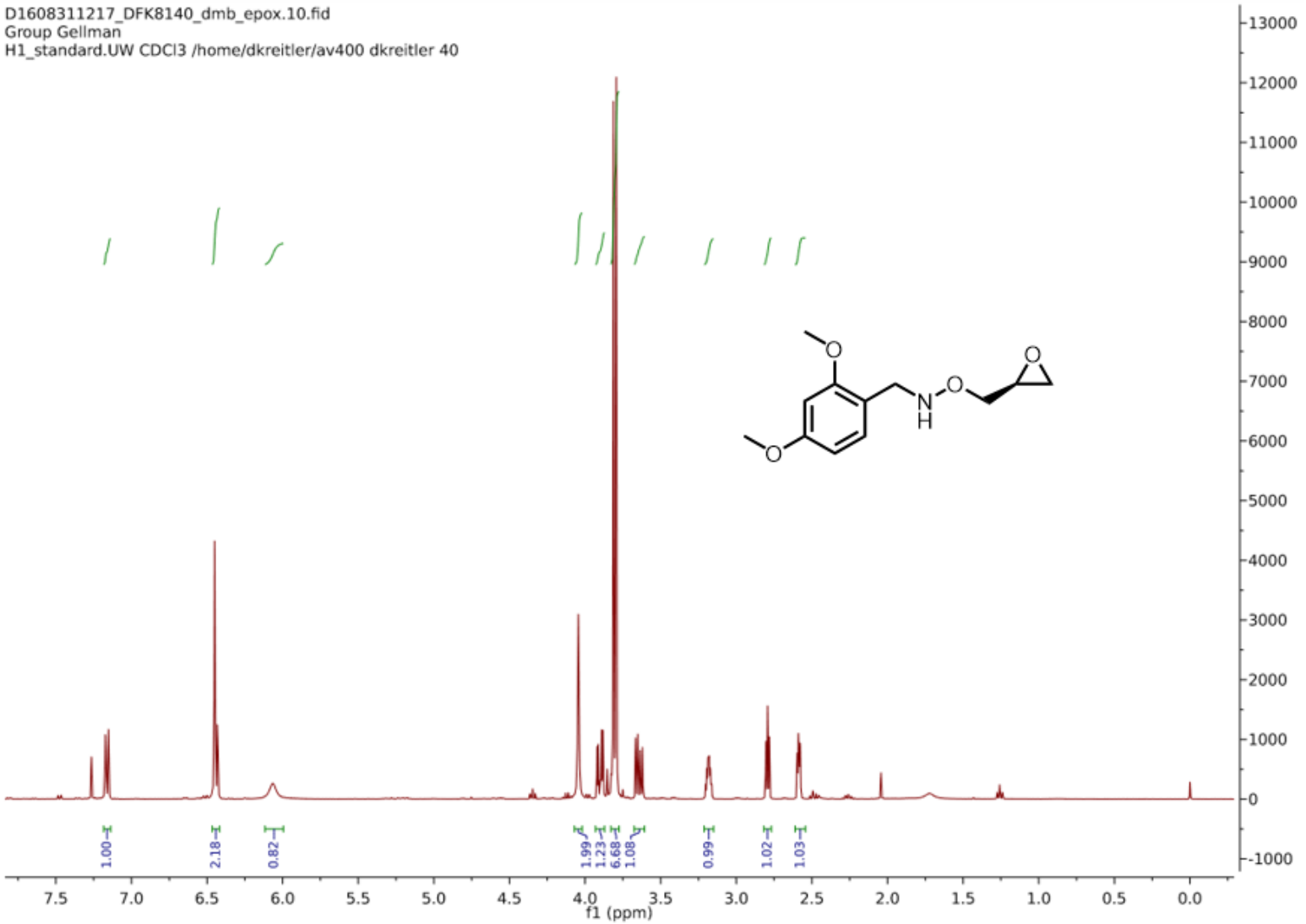


Figure B.10. ^{1}H NMR spectrum of Dmb-hydroxylamine epoxide.

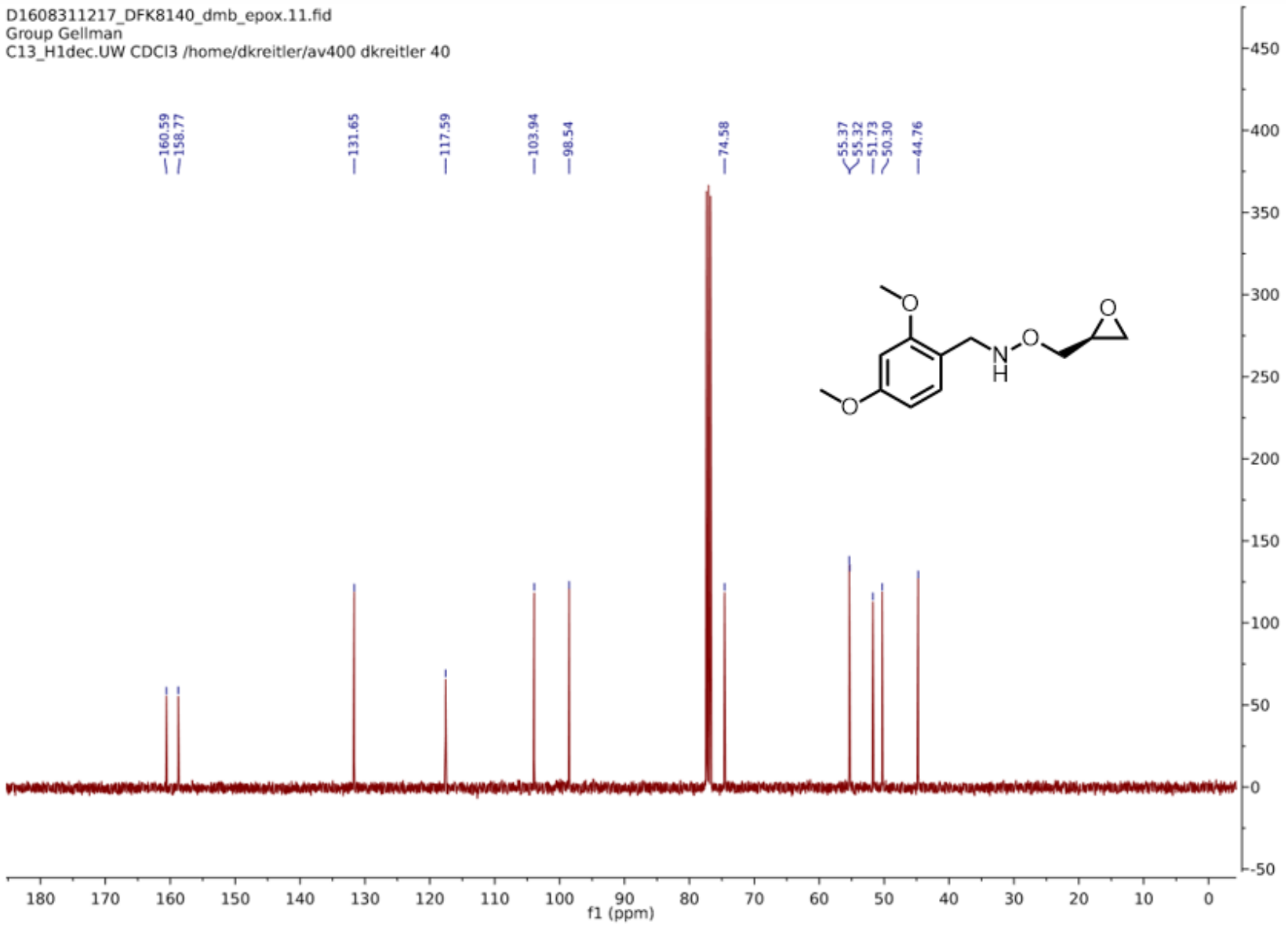
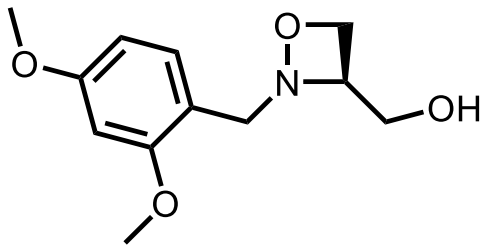


Figure B.11. ^{13}C NMR spectrum of Dmb-hydroxylamine epoxide.



Dmb-oxazetidine alcohol

Formula: $C_{12}H_{17}NO_4$

***IUPAC*:** [(3R)-2-[(2,4-dimethoxyphenyl)methyl]-1,2-oxazetidin-3-yl]methanol

***SMILES*:** COc2ccc(CN1OC[C@H]1CO)c(OC)c2

Procedure: prop-2-en-1-yl N-[(2,4-dimethoxyphenyl)methyl]-N-[(2S)-oxiran-2-ylmethoxy]carbamate (14.2 g, 59.3 mmol, 1.0 eq.) was dissolved in dry DCM (3 Å molecular sieves, 0.22 M, 300 mL) in a 500 mL round bottom flask. The flask was put under a dry N_2 stream and placed in a dry ice/acetonitrile bath maintained at -40°C. After the flask was allowed to cool for 20 min, distilled BF_3-OEt_2* (*vide infra*; 7.70 mL, 62.3 mmol, 1.05 eq.) was added dropwise via 1 mL syringe over 45 minutes. The resulting solution was then allowed to warm gradually from -40°C to -5°C over 9 h. Portions of dry ice were periodically added to the dry ice/acetonitrile bath to decrease the rate of warming over this time period. The solution was then diluted to approximately 400 mL with DCM and washed with a saturated aqueous sodium bicarbonate solution (250 mL). The organic and aqueous layers were separated, and the aqueous layer was extracted with DCM (2x200 mL). The organic layers were then combined, dried over sodium sulfate, and concentrated via rotary evaporation (bath temperature, 25°C) to afford a crude brown oil. The crude oil was then purified via silica flash chromatography (silica P60) 3%→5% v/v MeOH in DCM** (*vide infra*) to afford 7.0 g of a thick, dark-yellow oil (49% yield).

*BF_3-OEt_2 was distilled over CaH_2 through a short-path distillation apparatus under vacuum (water aspirator). The receiving flask was cooled in a dry ice/acetone bath.

** The column was packed with a 600 mL bed volume. Initial packing and equilibration was performed with 1 L 3% v/v MeOH in DCM. Excess solvent was collected and recycled. The crude oil was dissolved in 10 mL DCM and loaded onto the column where the remaining equilibration solvent (3% v/v MeOH in DCM) was used to collect the initial fractions (fraction volumes were approximately 150 mL).

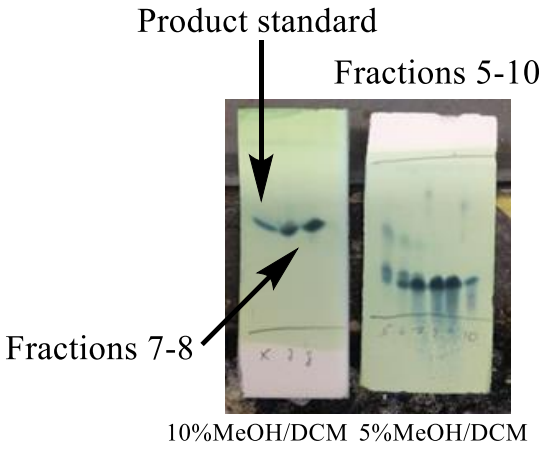


Figure B.12. Fraction purity was established with TLC (PMA stain).

^{1}H NMR (400 MHz, $CDCl_3$): δ = 7.25-7.20 (m, 1H), 6.47-6.43 (m, 2H), 4.86 (dd, *J*=7.9, 5.4 Hz, 1H), 4.75 (dd, *J*=8.6, 5.4 Hz, 1H), 4.22 (ABq, 2H, $\Delta\delta_{AB}$=0.03, J_{AB}=12.6 Hz), 4.11 (dddd, *J*=8.0, 8.0, 3.8, 3.8 Hz), 3.81 (s, 3H), 3.79 (s, 3H), 3.63-3.43 (m, 2H), 2.51 (s, 1H)

^{13}C NMR (400 MHz, $CDCl_3$): δ = 160.74, 158.71, 131.81, 115.25, 104.25, 98.61, 67.70, 66.83, 62.08, 59.61, 55.47, 55.37

ESI-MS [M+H]: *calculated* 240.1230, *observed* 240.1226

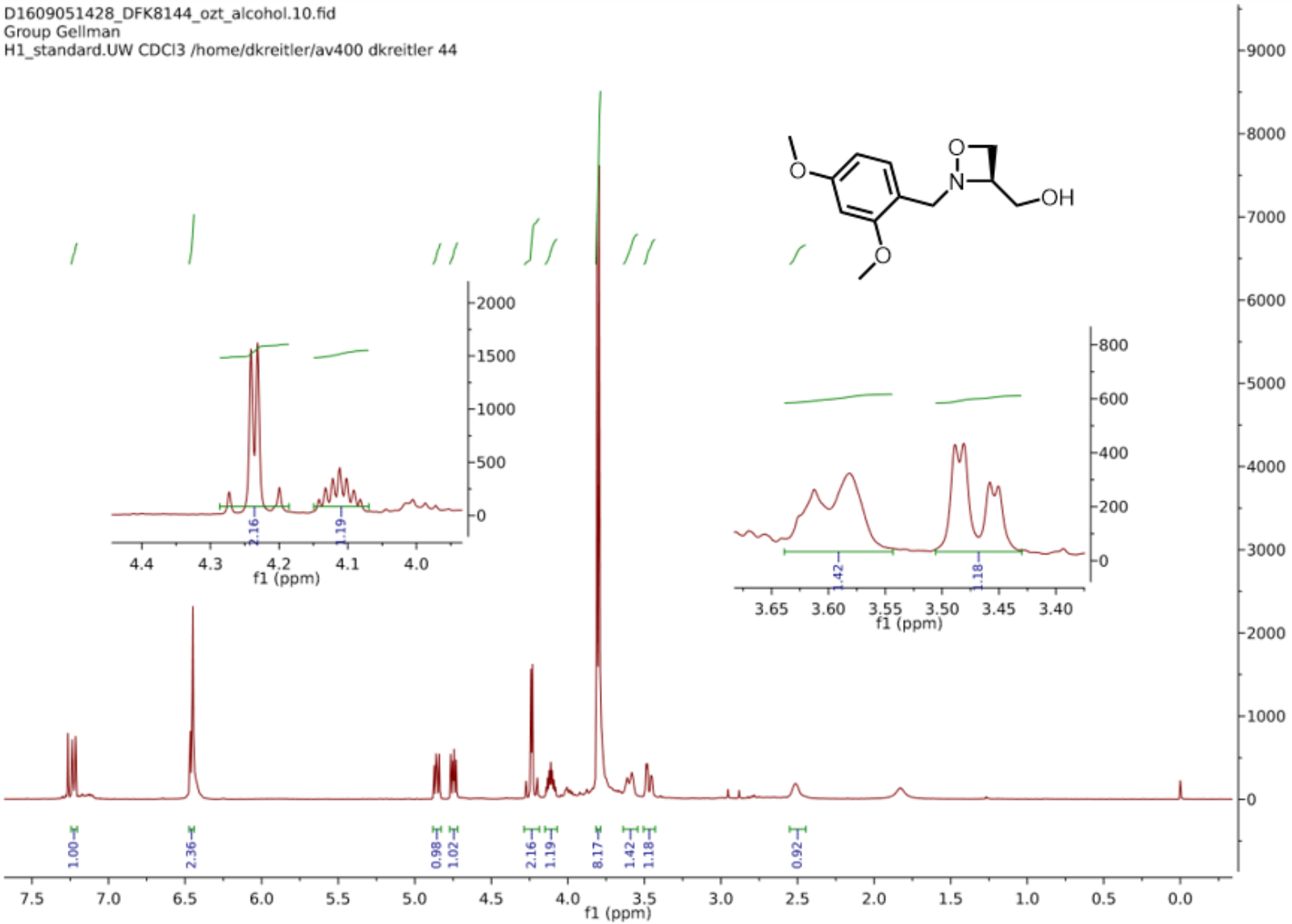


Figure B.13. ^{1}H NMR spectrum of Dmb-oxazetidine alcohol.

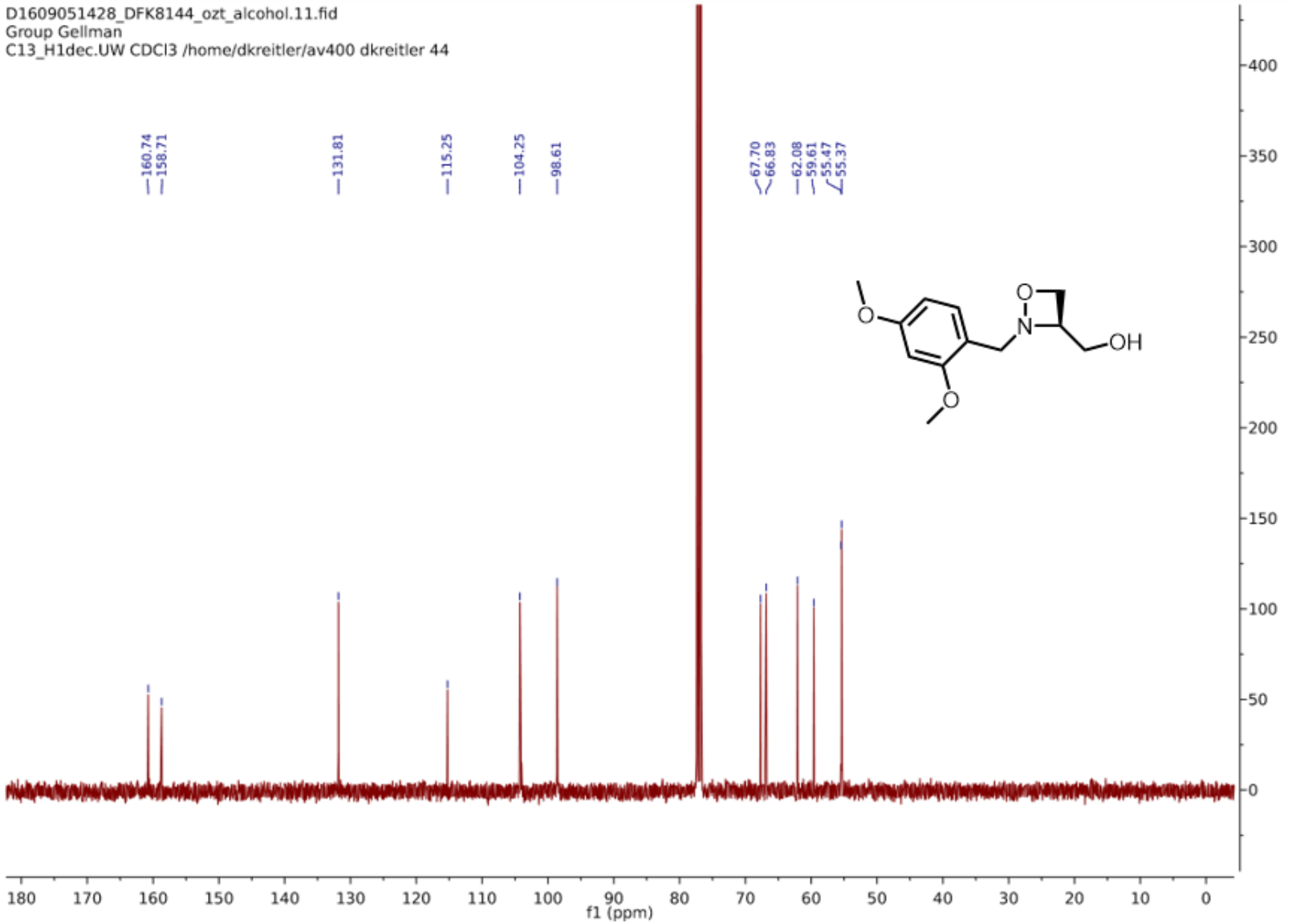
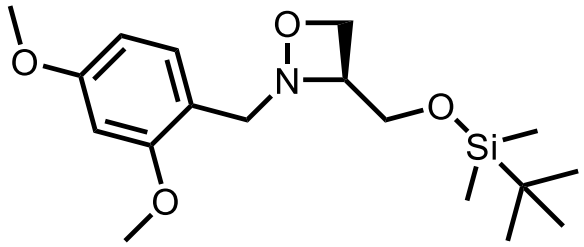


Figure B.14. ^{13}C NMR spectrum of Dmb-oxazetidine alcohol.



Dmb-oxazetidine tbdms ether

Formula: $C_{18}H_{31}NO_4Si$

***IUPAC*:** (3S)-3-{[(tert-butyldimethylsilyl)oxy]methyl}-2-[(2,4-dimethoxyphenyl)methyl]-1,2-oxazetidine

***SMILES*:** COc2ccc(CN1OC[C@H]1CO[Si](C)(C)C(C)(C)C)c(OC)c2

Procedure: [(3R)-2-[(2,4-dimethoxyphenyl)methyl]-1,2-oxazetidin-3-yl]methanol (7.0 g, 29.2 mmol, 1.0 eq.) was dissolved in DMF (3 Å molecular sieves, 0.44 M, 65 mL) in a flame dried 250 mL round bottom flask. While stirring, the flask was cooled to 0°C for 10 min. Imidazole (2.39 g, 35.1 mmol, 1.2 eq.) and *tert*-butyldimethylsilyl chloride (4.4 g, 29.2 mmol, 1.0 eq.) were each added all at once in the order listed. The reaction flask was placed under an N_2 stream and stirred for 2 h while the ice bath was allowed to gradually warm to room temperature. After 2 h, TLC indicated consumption of the primary alcohol starting material (10% v/v MeOH in DCM; PMA stain). The reaction mixture was then partitioned between brine (250 mL) and diethyl ether (250 mL). The layers were separated and the aqueous layer was extracted with diethyl ether (2x200 mL). The organic layers were combined, dried over sodium sulfate and concentrated via rotary evaporation (bath temperature, 25°C) to afford a crude brown oil. The brown oil was subjected to silica flash chromatography (20% v/v EtOAc in hexanes; Silica P60; 600 mL packed bed volume; 6 cm column inner diameter). Product containing fractions were

concentrated to afford 7.8 g of a colorless oil (75% yield; R_f=0.55 in 1:1 EtOAc/cyclohexane, PMA stain).

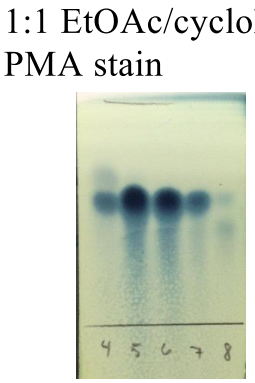


Figure B.15. Product containing fractions of the TBDMS protected Dmb-oxazetidine alcohol.

^{1}H NMR (400 MHz, $CDCl_3$): δ = 7.23-7.19 (m, 1H), 6.44-6.39 (m, 2H), 4.73 (dd, *J*=8.3, 5.5 Hz, 1H), 4.64 (dd, *J*=8.0, 5.5 Hz, 1H), 4.18 (ABq, 2H, $\Delta\delta_{AB}$=0.03, J_{AB}=12.8 Hz), 4.01 (m, 1H), 3.77 (s, 3H), 3.76 (s, 3H), 3.81-3.73 (m, 1H), 3.72-3.66 (s, 1H), 0.83 (s, 9H), 0.00 (s, 3H), -0.01 (s, 3H)

^{13}C NMR (400 MHz, $CDCl_3$): δ = 160.61, 158.73, 131.83, 116.18, 104.19, 98.60, 69.87, 67.61, 64.94, 60.50, 55.53, 55.50, 25.99, -5.25, -5.27

ESI-MS [M+H]: *calculated* 354.2089, *observed* 354.2090

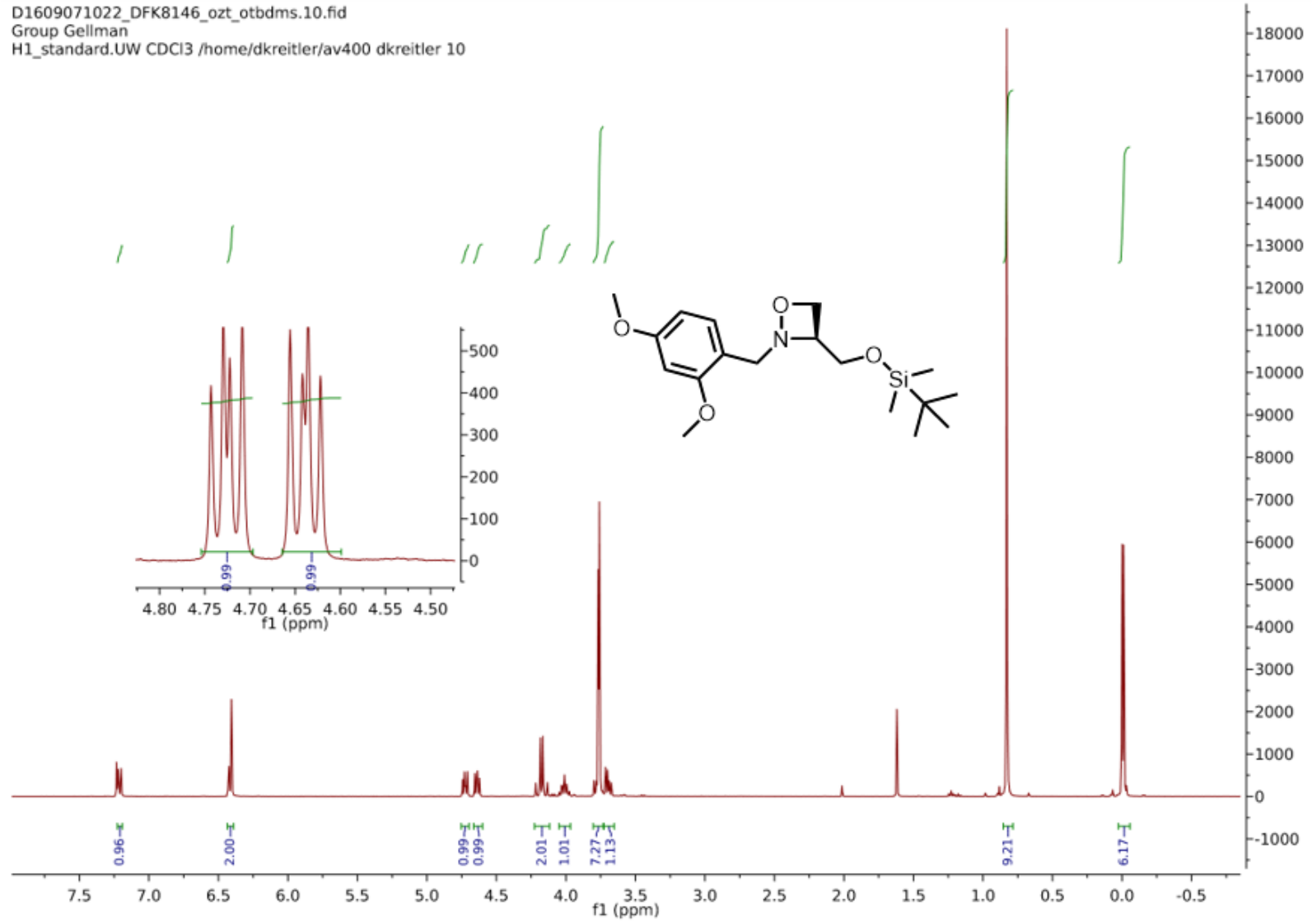


Figure B.16. ^{1}H NMR spectrum of Dmb-oxazetidine tbdms ether.

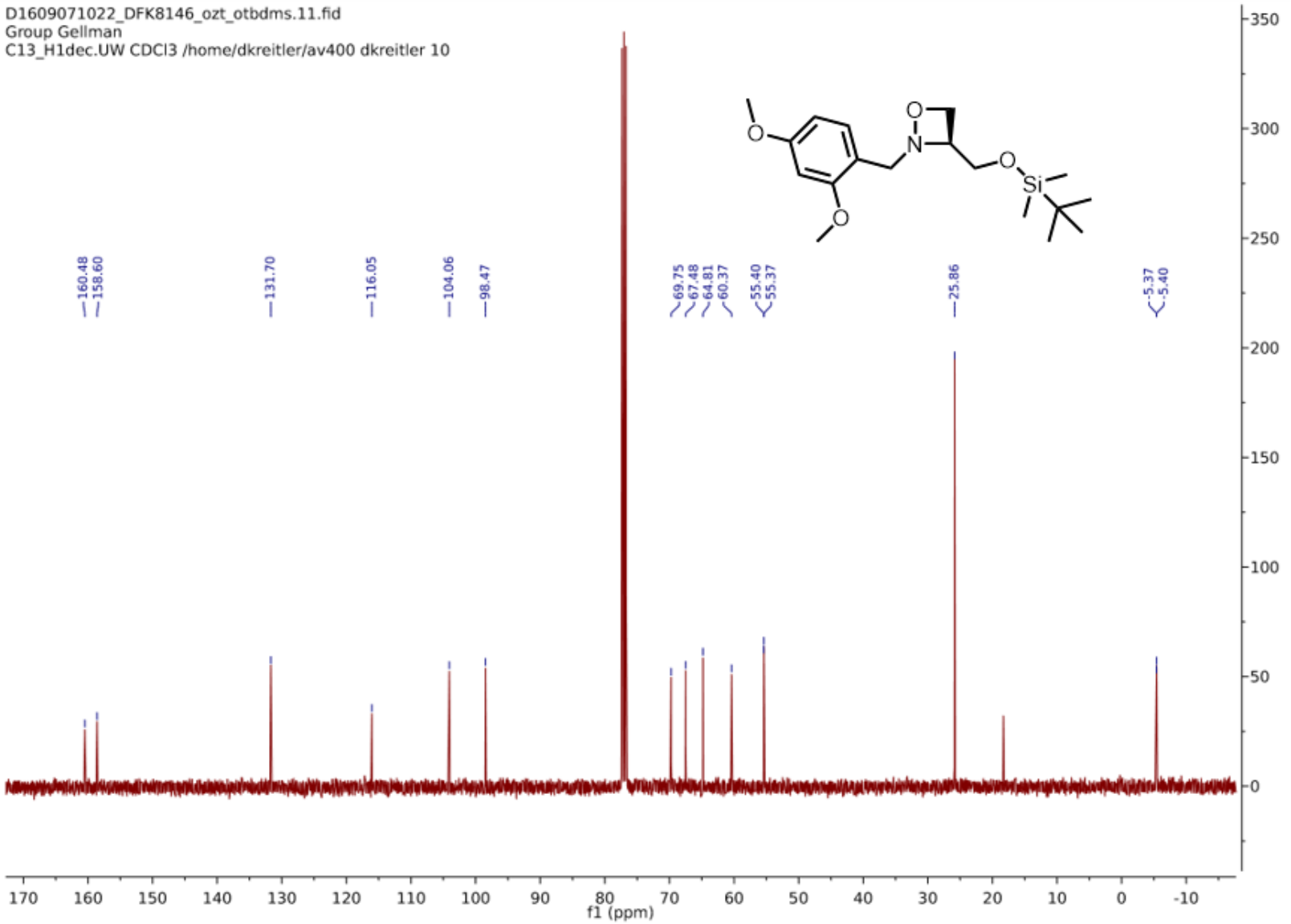
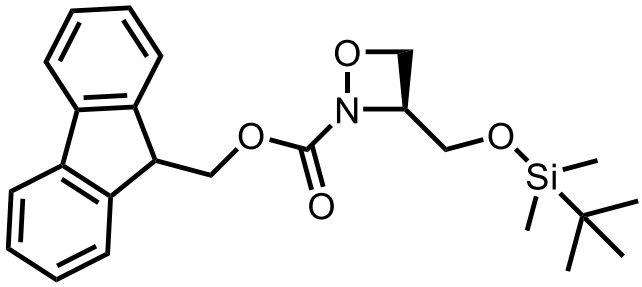


Figure B.17. ^{13}C NMR spectrum of Dmb-oxazetidine tbdms ether.



Fmoc-oxazetidine tbdms ether

Formula: $C_{24}H_{31}NO_4Si$

***IUPAC*:** (9H-fluoren-9-yl)methyl(3S)-3-{[(tert-butyldimethylsilyl)oxy]methyl}-1,2-oxazetidine-2-carboxylate

***SMILES*:** CC(C)(C)[Si](C)(C)OC[C@@H]1CON1C(=O)OCC4c2ccccc2c3ccccc34

Procedure: (3S)-3-{[(tert-butyldimethylsilyl)oxy]methyl}-2-[(2,4-dimethoxyphenyl)methyl] 1,2-oxazetidine (6.30 g, 17.8 mmol, 1.00 eq.) was dissolved in wet DCM (0.20 M, 90 mL) in a 250 mL round bottom flask. While stirring, the flask was placed under an N_2 stream (to avoid excessive condensation) and cooled to -78°C with a dry ice/acetone bath. 2,3-dichloro-5,6-dicyano-1,4-benzoquinone* (DDQ) (4.10 g, 17.8 mmol, 1.00 eq.) was added in small portions over a period of 20 min, after which the flask was sealed with a rubber septum and stirred at -78°C for 10 h. Fmoc chloride (5.74 g, 17.8 mmol, 1.00 eq.) was added in one lot followed by dropwise addition of diisopropylethylamine (DIEA) (3.10 mL, 17.8 mmol, 1.00 eq.). The solution was then allowed to stir 12 h and gradually warm to room temperature. A gentle air stream was placed over the dry ice/acetone bath to facilitate warming. The red solution was diluted to 250 mL with DCM and washed with saturated aqueous ammonium chloride (200 mL). The layers were separated and the aqueous layer was extracted with DCM (2x150 mL). The organic layers were combined; washed with brine; dried over sodium sulfate; and adsorbed onto approximately 15 g of

dry silica via rotary evaporation (bath temperature, 25°C) to afford a crude rust-colored powder. The powder was divided into two portions where each portion was purified via silica flash chromatography (Silica P60, 600 mL packed bed volume, 6 cm inner diameter, 20% v/v EtOAc in hexanes). Product containing fractions (R_f=0.4, 25% v/v EtOAc in hexanes) were combined and concentrated via rotary evaporation (bath temperature, 25°C). During concentration a polar (relative to the product) impurity precipitated from solution. The heterogeneous mixture was diluted with anhydrous diethyl ether and filtered over glass wool to remove the solid impurity. Finally, the solution was concentrated to afford 3.30 g of a yellow oil that foamed into a white, oily solid under vacuum (44% yield).

*DDQ was recrystallized from 4:1 chloroform:benzene. DDQ (8.0 g) was stirred in 4:1 chloroform:benzene (250 mL) in a 500 mL Erlenmeyer flask. The flask was gently heated while stirring over a heat gun, and then the heterogeneous solution was filtered through glass wool into a 500 mL round bottom flask. The flask was then connected to a vacuum source outfitted with a condenser and vacuum was applied to the stationary flask. Gradually the solvent evaporated, depositing a layer of crystals upon the side of the flask. When approximately 10% of the original volume of the mother liquor was remaining the flask was disconnected from the vacuum source, the remaining mother liquor was carefully pipetted out of the flask so as not to disturb the crystals on the side of the flask, and then the flask was placed under hi-vacuum. The wet crystals slowly transitioned from red to a deep yellow color as the residual solvent evaporated. The dry yellow crystals were collected by scraping from the side of the flask and used directly in the reaction.

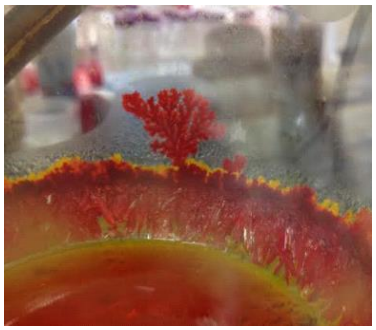


Figure B.18. Recrystallized DDQ deposited on the side of a round-bottom flask under vacuum.

^{1}H NMR (400 MHz, $CDCl_3$): δ = 7.79-7.73 (m, 2H), 7.66-7.56 (m, 2H), 7.43-7.38 (m, 2H), 7.35-7.29 (m, 2H), 4.91-4.83 (m, 2H), 4.77-4.69 (m, 1H), 4.57-4.43 (m, 2H), 3.99 (dd, *J*=11.5, 4.9 Hz, 1H), 3.78 (dd, *J*=11.5, 4.9 Hz, 1H) 0.92 (s, 9H), 0.088 (s, 3H), 0.074 (s, 3H)

^{13}C NMR (400 MHz, $CDCl_3$): δ = 161.56, 143.54, 143.52, 141.47, 141.42, 128.00, 127.34, 127.32, 127.24, 125.34, 125.29, 120.13, 70.18, 68.23, 66.96, 62.41, 47.02, 25.96, -5.19, -5.29

ESI-MS [M+H]: *calculated* 426.2095, *observed* 426.2087, **[M+NH_4]:** *calculated* 443.2361, *observed* 443.2352, **[M+Na]:** *calculated* 448.1915, *observed* 448.1905

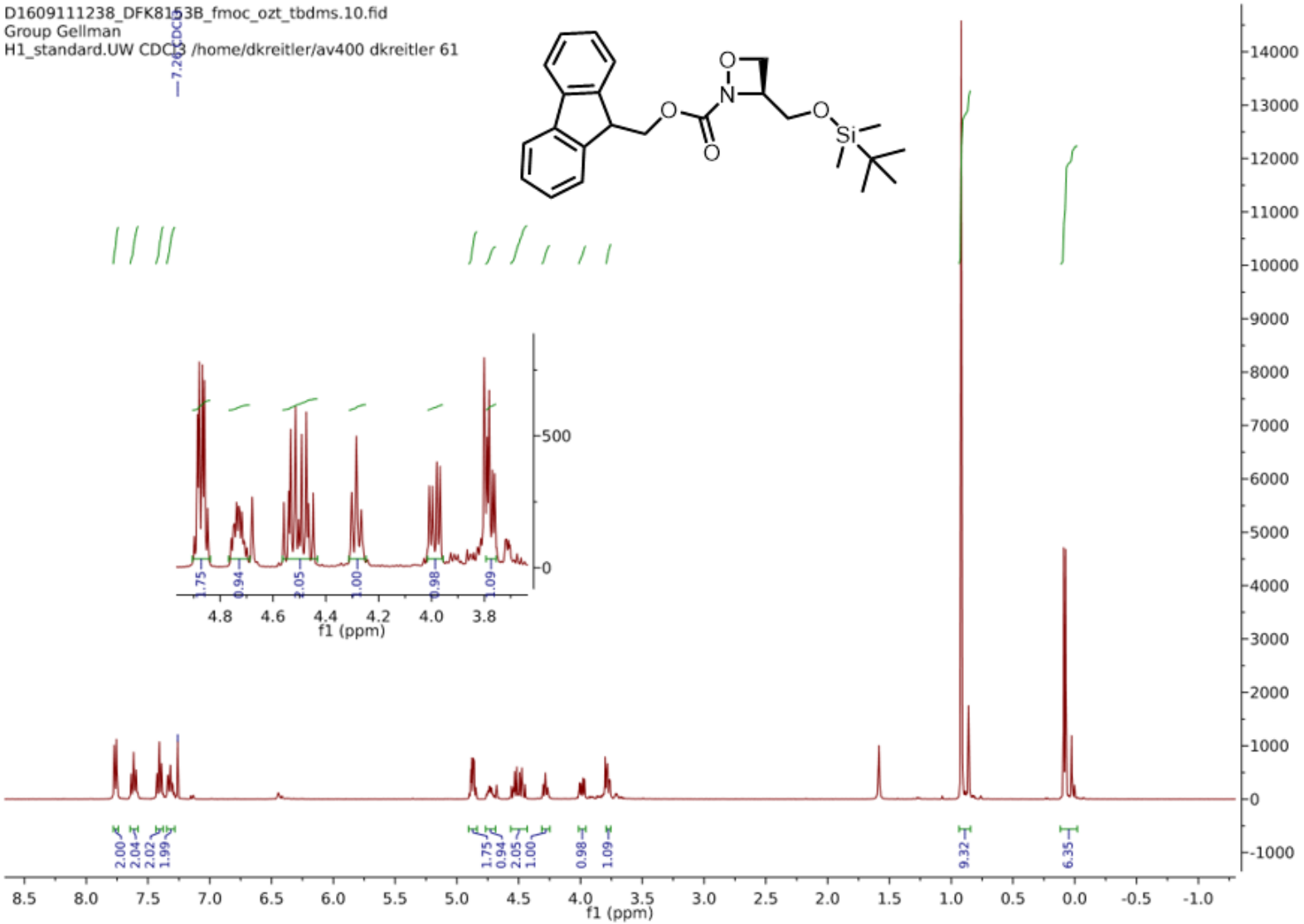


Figure B.19. ^{1}H NMR spectrum of Fmoc-oxazetidine tbdms ether.

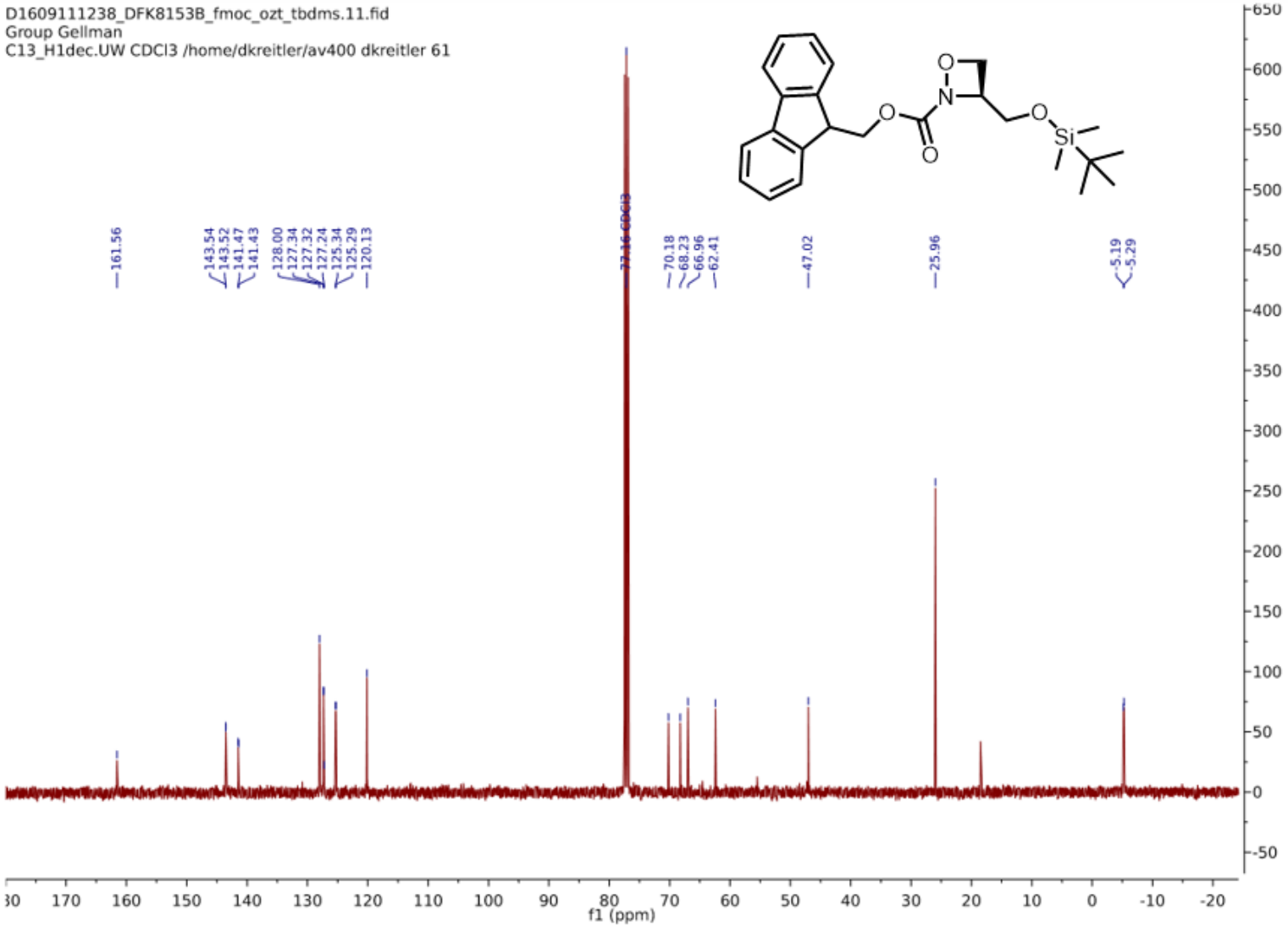
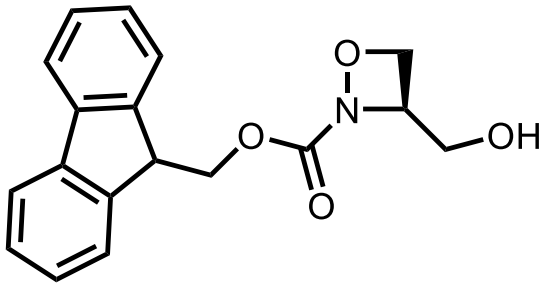


Figure B.20. ^{13}C NMR spectrum of Fmoc-oxazetidine tbdms ether.



Fmoc-oxazetidine alcohol

Formula: $C_{18}H_{17}NO_4$

***IUPAC*:** (9H-fluoren-9-yl)methyl(3R)-3-(hydroxymethyl)-1,2-oxazetidine-2-carboxylate

SMILES: O=C(OCC3c1ccccc1c2ccccc23)N4OC[C@H]4CO

Procedure: (9H-fluoren-9-yl)methyl(3S)-3-{[(tert-butyldimethylsilyl)oxy]methyl}-1,2-oxazetidine-2-carboxylate (3.3 g, 7.8 mmol, 1.0 eq.) was dissolved in acetonitrile (70 mL, 0.11 M) in a 250 mL round bottom flask. The flask containing the resultant solution was cooled to 0°C with an ice/rock salt bath for 15 min while the flask contents were stirring. Then, aqueous HF solution (48%) was added dropwise to the stirring solution via plastic pipet. The solution was stirred for 2.5 h at 0°C, after which the solution was diluted to approximately 200 mL with ethyl acetate. The solution was carefully transferred to a 500 mL separatory funnel and slowly quenched with aqueous sat. sodium bicarbonate (200 mL). Once effervescence had ceased, layers were separated and the aqueous layer was extracted with ethyl acetate (2x100 mL). Organic layers were combined, washed with brine, dried over sodium sulfate, and concentrated via rotary evaporation (bath temperature, 25°C) to afford a thick brown oil.

Flash Column Chromatography: A 5 cm inner diameter flash column was packed with a 300 mL packed bed volume of silica P60. Initially, the column was equilibrated with 1:1 EtOAc:hexanes. The crude brown oil was dissolved in a minimal amount of DCM/EtOAc

and adsorbed onto silica P60 which was then loaded dry onto the column head. A stepwise gradient of 50% v/v EtOAc → 66% v/v EtOAc in hexanes was employed to elute the desired product (500 mL 50% v/v EtOAc in hexanes; 500 mL 60% v/v EtOAc in hexanes; 500 mL 66% EtOAc in hexanes). Fractions were pooled and concentrated via rotary evaporation (bath temperature, 25°C) to afford 1.5 g of a thick yellow oil (63% yield).

^{1}H NMR (400 MHz, $CDCl_3$): 7.80-7.74 (m, 2H), 7.65-7.57 (m, 2H), 7.46-7.38 (m, 2H), 7.37-7.29 (m, 2H), 4.95-4.72 (m, 3H), 4.63-4.49 (m, 2H), 4.29 (t, *J*=6.9 Hz, 1H), 3.85-3.69 (m, 2H), 2.19 (dd, *J*=7.5, 5.4 Hz, 1H)

^{13}C NMR (400 MHz, $CDCl_3$): δ = 143.37, 141.50, 141.47, 128.12, 128.10, 127.40, 125.27, 125.21, 120.18, 69.09, 68.41, 67.71, 47.03

ESI-MS [M+H]: *calculated* 312.1230, *observed* 312.1223, **[M+NH_4]:** *calculated* 329.1496, *observed* 329.1490

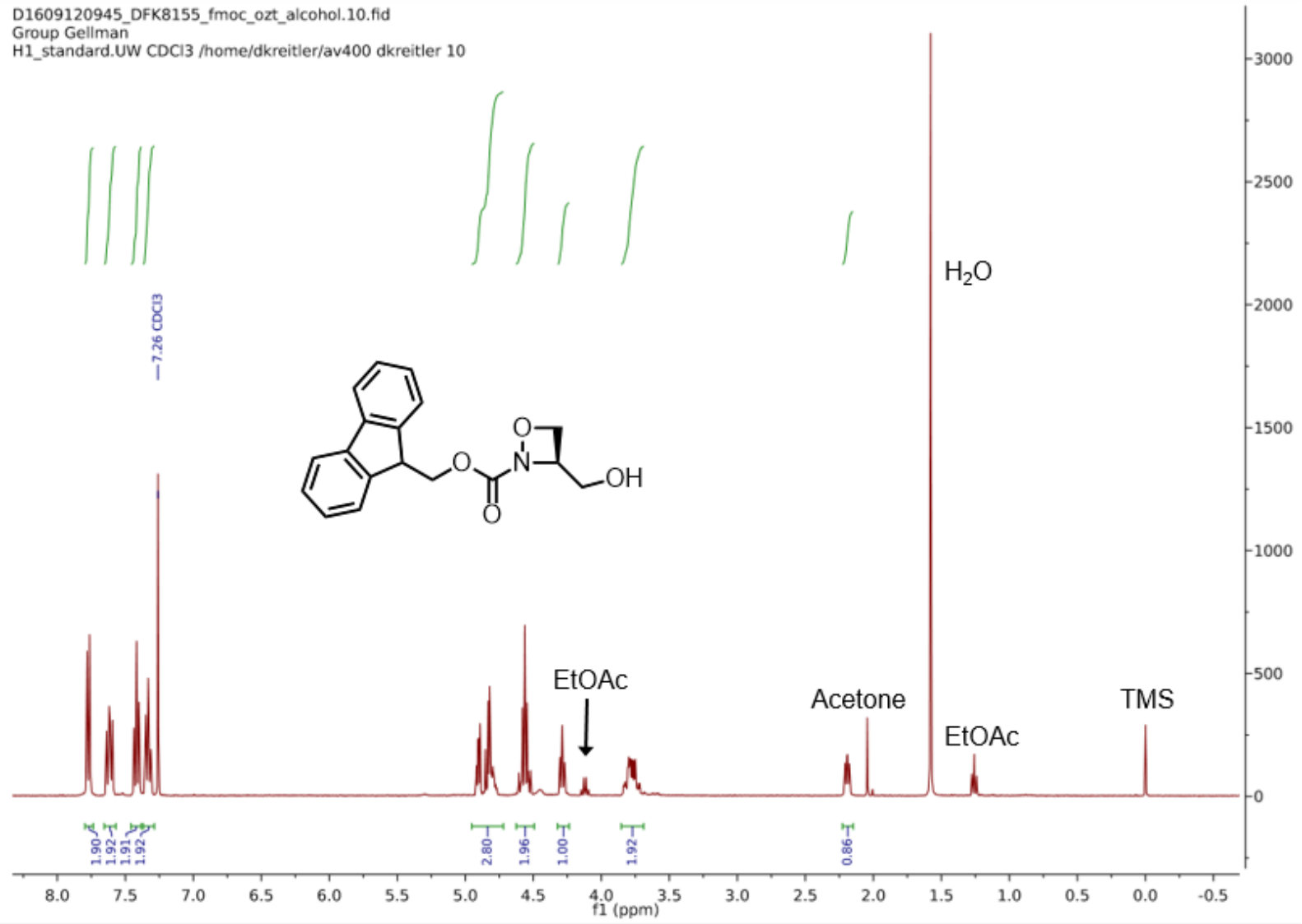


Figure B.21. ^{1}H NMR spectrum of Fmoc-oxazetidine alcohol.

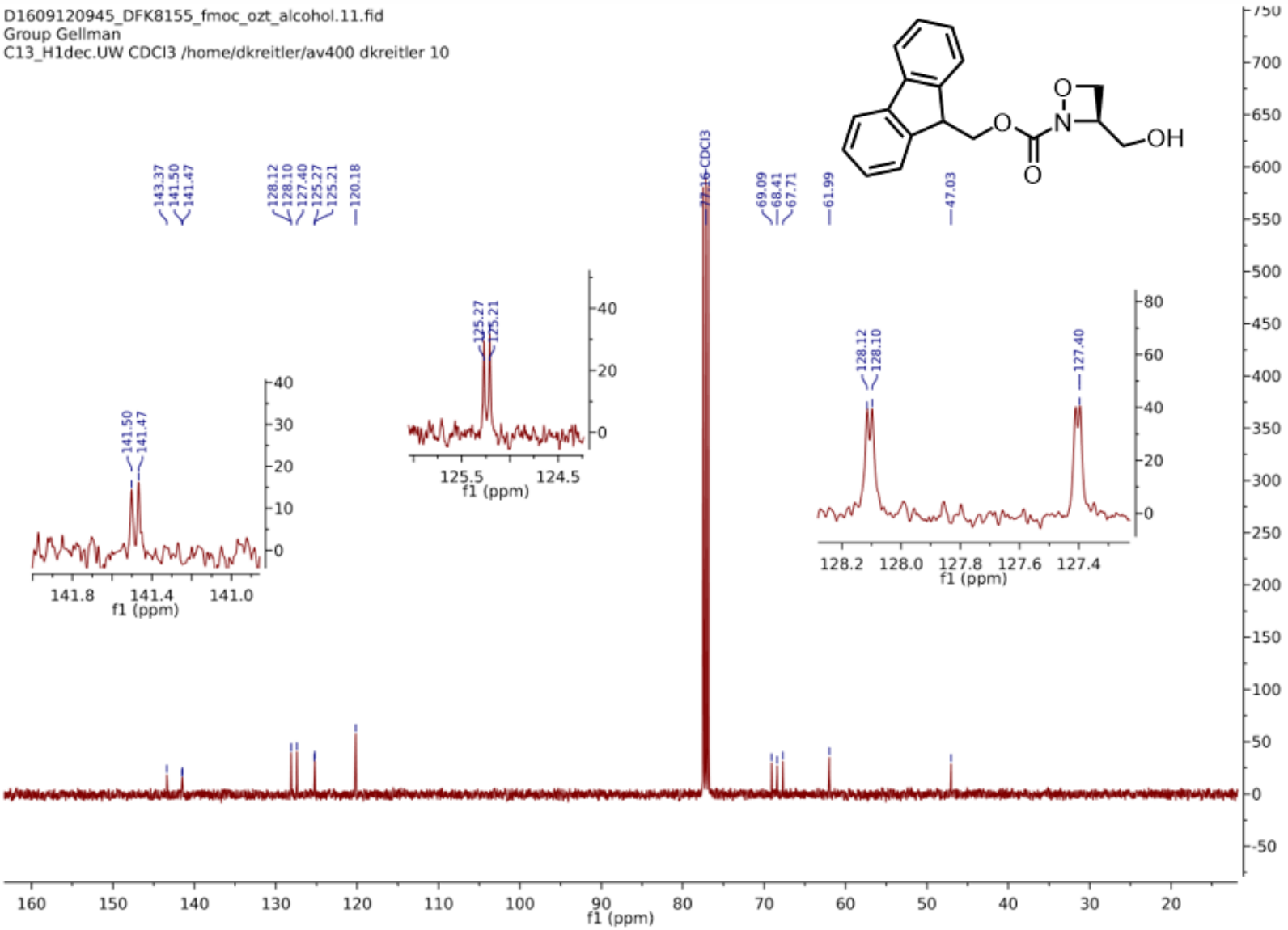
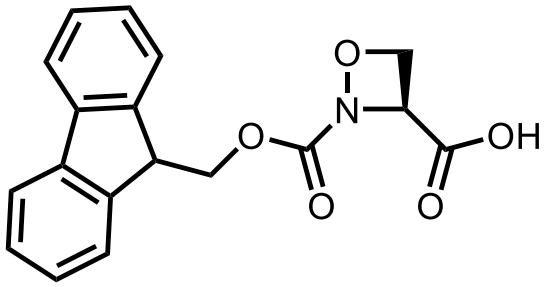


Figure B.22. ^{13}C NMR spectrum of Fmoc-oxazetidine alcohol.



Fmoc-oxazetidine carboxylic acid

Formula: $C_{18}H_{15}NO_5$

IUPAC: (9H-fluoren-9-yl)methyl (3R)-3-(hydroxymethyl)-1,2-oxazetidine-2-carboxylate

SMILES**:** O=C(OCC3c1ccccc1c2ccccc23)N4OC[C@H]4CO

Procedure: This procedure comprises a Swern oxidation followed immediately by a Pinnick (or Lindgren) oxidation with no intermediate purification. Presumably, this sequence of reactions is necessary to prevent oxazetidine ring opening during oxidation. To ensure that the starting material was dry, the oxazetidine alcohol was purified via flash chromatography immediately prior to the initial Swern oxidation.

Swern oxidation: Oxalyl chloride (163 µL, 1.93 mmol, 1.20 eq.) was transferred via syringe through a rubber septum to a flame-dried 50 mL flask under N_2. Dry DCM (3.0 mL, 3 Å sieves) was also added to the flask via syringe under N_2. The flask was cooled to -78°C (dry ice/acetone bath) while the contents of the flask were stirred. Then dry DMSO (125 µL, 1.76 mmol, 1.10 eq., 3 Å sieves) was added to the flask via syringe. The resulting solution was stirred for 10 min at -78°C after which a solution of Fmoc-oxazetidine alcohol (500 mg, 1.60 mmol, 1.00 eq.) in DCM (6.0 mL, 0.27 M) was added to the flask dropwise over 12 min. The resulting solution was stirred for 20 min at -78°C, after which *N,N*-diisopropylethylamine (DIEA; 1.40 mL, 8.00 eq., 5.00 eq.) in dry DCM (5 mL) was added dropwise over 15 min. The resulting solution was stirred at -78°C for

an additional 2 h. (Prior to addition of the DIEA/DCM solution the flask contents were a deep red-orange. Upon addition of DIEA/DCM the solution gradually changed to a pale yellow). The flask was allowed to slowly warm to 0°C over 1 h. After this time period, the solution was transferred to a separatory funnel (500 mL) and diluted with DCM (100 mL). The solution was washed with a 10% w/v aqueous citric acid solution (100 mL, aq.). The layers were separated and the aqueous layer was extracted with DCM (2x50 mL). The organic layers were combined, washed with brine, and dried over sodium sulfate. The solution was then carefully concentrated via rotary evaporation (bath temperature set at room temperature) to afford a pale yellow oil that foamed into an off-white solid under hi-vacuum.

Figure B.23. Presumed crude Fmoc-oxazetidine aldehyde after Swern oxidation.

Pinnick Oxidation: The aldehyde was dissolved in 1:1 THF/*t*-BuOH (10 mL) and cooled to 0°C, after which H_2O (5 mL) and 2-methyl-2-butene (1 mL) were added. While stirring, $NaH_2PO_4 \cdot H_2O$ (1.34 g, 8.64 mmol, 5.40 eq.) and $NaClO_2$ (1.45 g, 16.0 mmol, 10.0 eq.) were added in the order listed. The resulting solution was stirred at 0°C for 2 h, during

which the solution turned bright yellow. The solution was then diluted to approximately 100 mL with EtOAc and washed with 1 M HCl (50 mL aq.). The layers were separated and the aqueous layer extracted with EtOAc (2x50 mL). The organic layers were combined and washed with brine and dried over sodium sulfate. The solution was then carefully concentrated via rotary evaporation to afford a light-brown, oily residue (bath temperature set to room temperature; during concentration a bright yellow distillate was removed from the solution). The crude residue was dissolved in acetonitrile (5 mL, HPLC grade) and subjected to MPLC purification with reverse phase silica (Figure B**.24**).

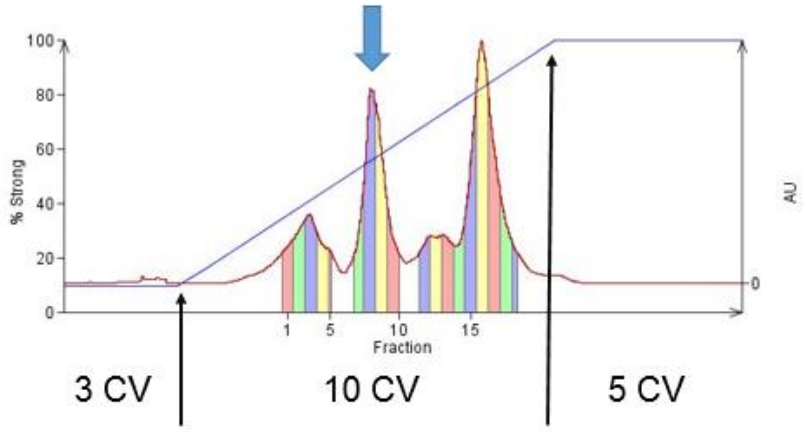


Figure B.24. Column dimensions (inner diameter: 2.8 cm, length: 8.3 cm, 50 mL total volume). Stationary phase, C8 reverse phase silica (40-63 µm, 1.46 mmol/g). The employed MPLC method (10 mL/min) includes an isocratic hold for 3 column volumes (10%B), followed by a gradient (10→100%B) over 10 column volumes, and finally an isocratic flush (100%B) for 5 column volumes. Absorbance was measured at 254 nm and the peak corresponding to Fmoc-oxazetidine-COOH is indicated by the blue arrow.

MPLC Purification: C8 functionalized reverse phase silica (40-63 µm, 1.46 mmol/g) was purchased from Silicycle (Montreal, Quebec, CA) and packed into a fritted 2.8 x 8.3 cm column cartridge. The column was packed by running a blank injection with the MPLC method described in Figure B**.24**. Binary gradients comprised H_2O (A solvent; millipore

filtered) and acetonitrile (B solvent; HPLC grade). Fractions containing the product were pooled together and lyophilized to afford 42 mg of a white powder (8% yield).

^{1}H NMR (400 MHz, $CDCl_3$): δ = 7.8-7.72 (m, 2H), 7.66-7.56 (m, 2H), 7.45-7.36 (m, 2H), 7.36-7.28 (m, 2H), 5.27-5.18 (m, 1H), 5.09-5.03 (m, 2H), 4.59-4.43 (m, 2H), 4.31-4.24 (m, 1H)

^{13}C NMR (400 MHz, $CDCl_3$): δ = 143.26, 143.14, 141.46, 128.13, 125.42, 125.34, 120.22, 69.91, 69.24, 64.29, 46.84

ESI-MS: [M+H] *calculated* 324.0878, *observed* 324.0875

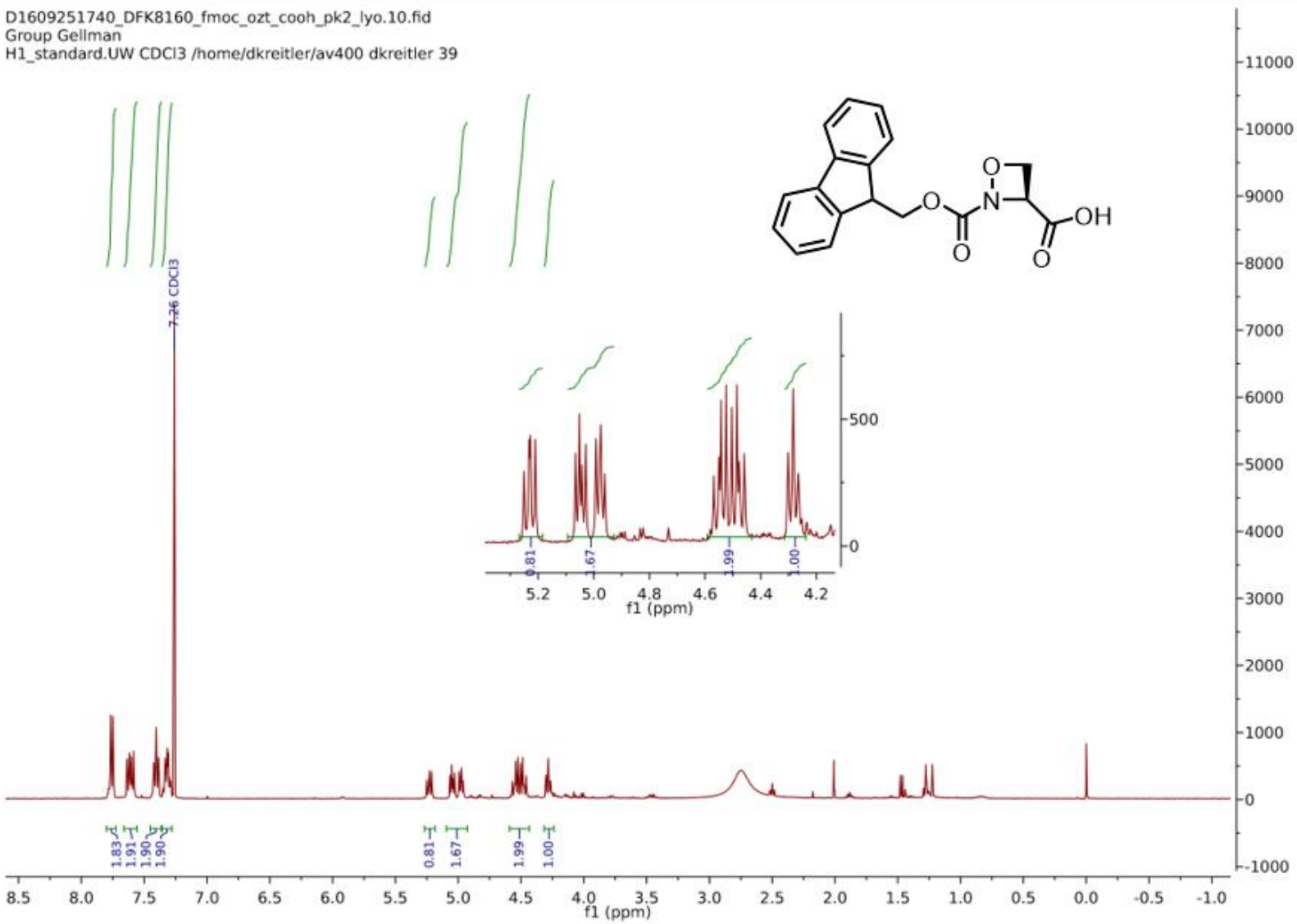


Figure B.25. ^{1}H NMR spectrum of Fmoc-oxazetidine acid.

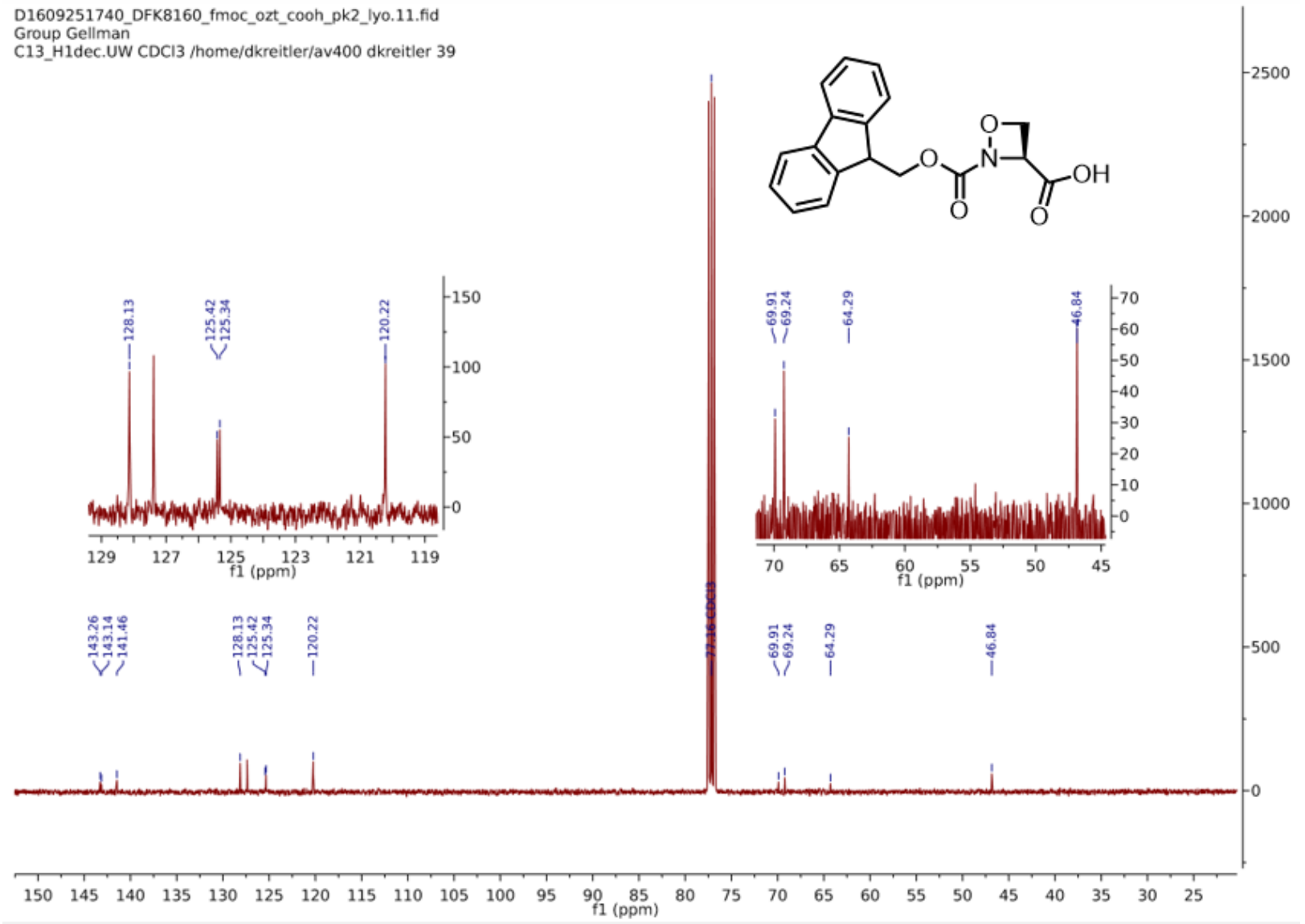


Figure B.26. ^{13}C NMR spectrum of Fmoc-oxazetidine acid.

Incorporation of Fmoc-Ozt-COOH into the Parainfluenza 5 Transmembrane Domain

The parainfluenza virus 5 transmembrane domain (PIV5-TM) was synthesized with microwave-assisted solid phase synthesis (100 µmol scale, Rink amide resin, 0.5 mmol/g, 4 eq. COMU/Fmoc-amino acid, 8 eq. 2,4,6-trimethylpyridine in 4 mL DMF; 2x12 min coupling reactions at 70°C, 1x4 min 20% v/v piperidine/DMF deprotection reactions at 80°C) through the following residues: IIAICLGSLGLILIILLSVVVWKLL. Following deprotection of the N-terminal Ile residue, the resin was then partitioned into 25 µmol and 75 µmol aliquots. Fmoc-Ozt-COOH was coupled onto the 25 µmol aliquot (2 eq. Fmoc-Ozt-COOH, 2 eq. COMU, 4 eq. 2,4,6-trimethylpyridine in 1 mL DMF; room temperature 1 h, magnetic stirring). Following Fmoc-Ozt-COOH coupling, the peptide was cleaved with 95/2.5/2.5 TFA/H_2O/EDT (Pusterla and Bode employed DODT,[1] a less volatile/malodorous dithiol than EDT). The peptide was precipitated in diethyl ether and centrifuged, after which the ether was decanted and the crude peptide pellet dried under a gentle N_2 stream. The crude peptide was dissolved in TFE (2 mL) and subjected to HPLC purification (Agilent Zorbax C3 stable bond, 21.2 x 250 mm; A solvent: 50% EtOH/H_2O, 0.05% v/v TFA; B solvent: iPrOH, 0.05% v/v TFA).

Conclusions: The ion corresponding to the [M+Na] adduct of Fmoc-Ser-PIV5 TM was detected (Figure B.27). This result implies that the Fmoc-oxazetidine monomer was successfully incorporated via solid phase peptide synthesis, but that oxazetidine ring opening occurred prior to or during ionization in MALDI-TOF MS. These data do not rule out the possibility that ring opening occurred prior to ionization in MALDI-TOF MS (Figure B.28). Furthermore, addition of the Fmoc-oxazetidine monomer appears to significantly

decrease the solubility of PIV5 TM in TFE. Although, switching to HFIP appeared to slightly increase solubility, as determined by HPLC peak shape (Figure B.29). Ultimately, the oxazetidine-mediated ketoacid hydroxylamine ligation strategy does not appear to be a viable strategy for obtaining the multi-milligram quantities of transmembrane peptides required for crystallization studies.

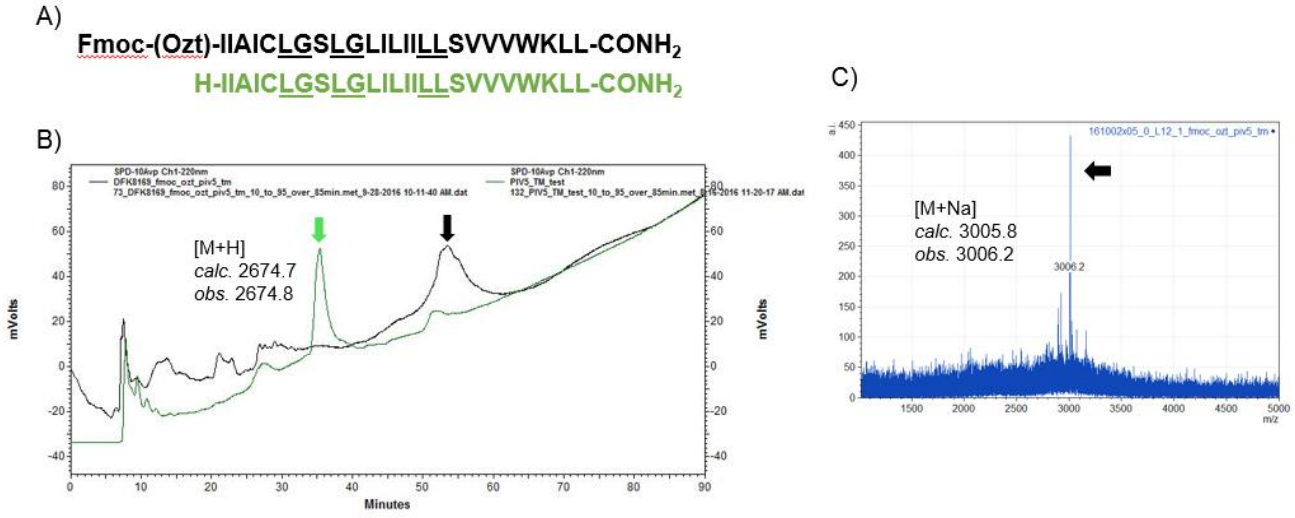


Figure B.27. A) To ensure the successful synthesis of the C-terminal portion of PIV5 TM (underlined positions indicate the incorporation of a Dmb-protected dipeptide) B) a microcleavage was prepared under the reported cleavage and HPLC conditions (green trace), the identity of the peptide lacking N-terminal Fmoc-oxazetidine was established by MALDI-TOF. Fmoc-oxazetidine appeared to increase the hydrophobicity of the PIV5 TM domain (black trace). C) The Na adduct corresponding to Fmoc-oxazetidine-PIV5 TM was detected via MALDI-TOF. However, the mass of the [M+Na] ion suggests that oxazetidine ring opening may occur upon ionization.

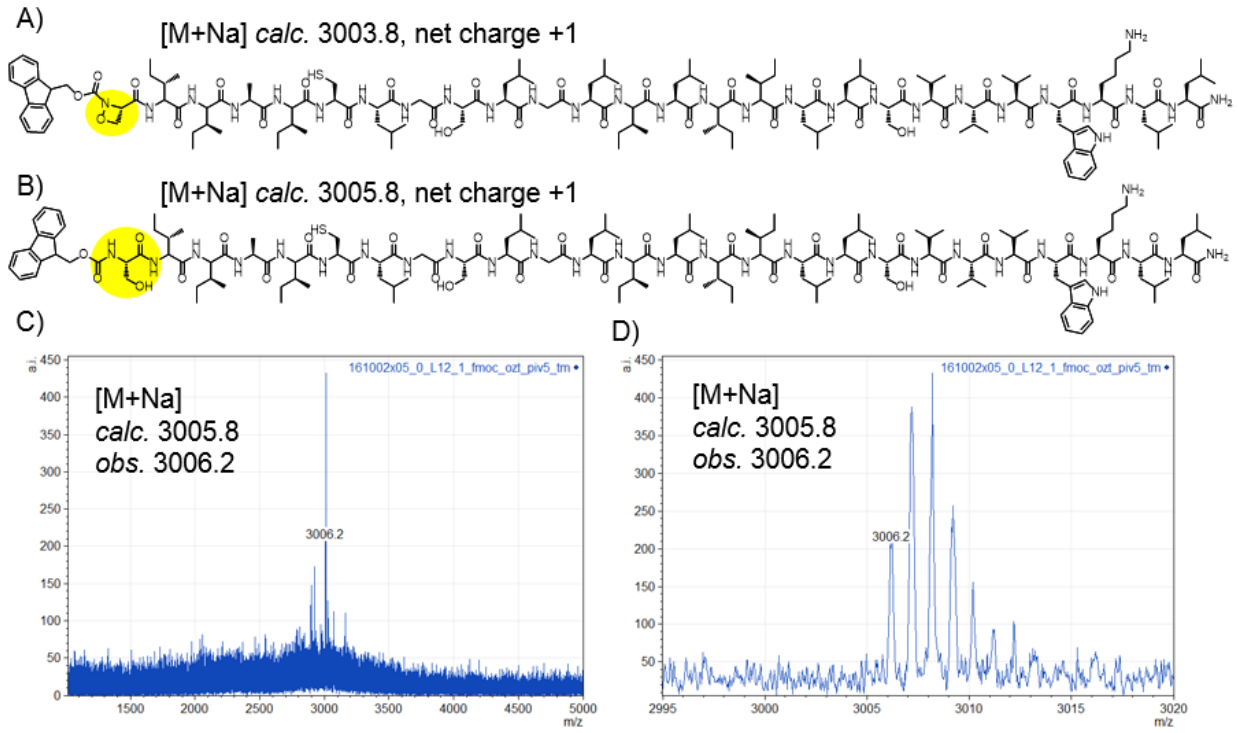


Figure B.28. Evidence for oxazetidine ring opening in MALDI-TOF MS. A) The mass of the [M+Na] ion corresponding to the unopened oxazetidine ring is 3003.8 Da. B) The mass of the [M+Na] ion corresponding to Fmoc-Ser-PIV5 TM is 3005.8 which more closely agrees with C, D) the observed [M+Na] ion. These data support the hypothesis that oxazetidine ring opening occurs upon ionization in MALDI-TOF. Pusterla and Bode reportedly employed ESI-MS for quantifying the mass of oxazetidine-containing peptides.

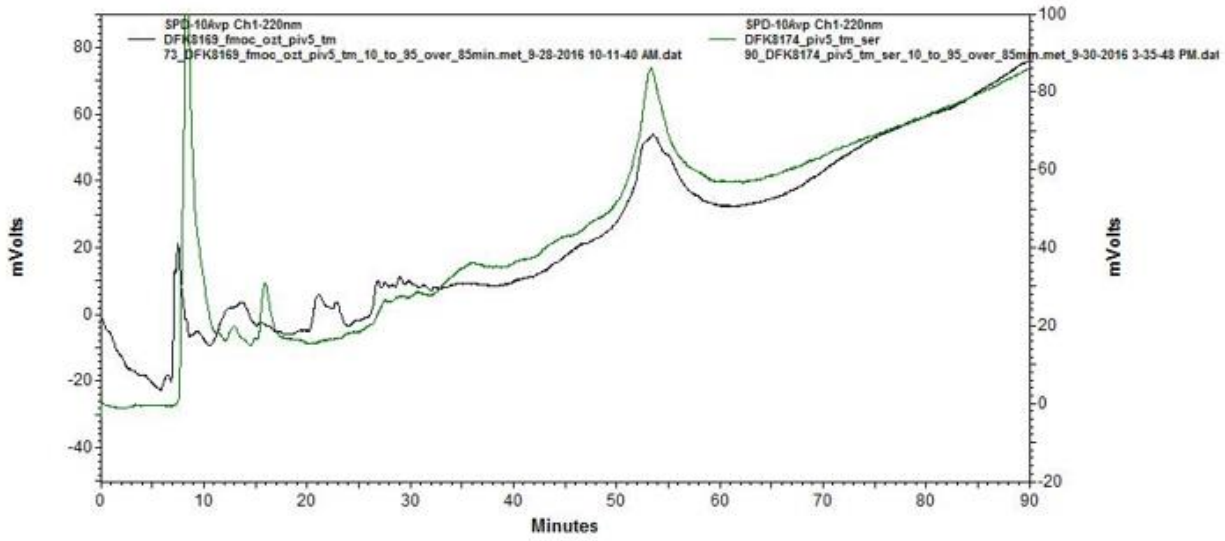


Figure B.29. A 150 µL portion of the Fmoc-oxazetidine-PIV5 TM sample dissolved in TFE was diluted with water, lyophilized, and redissolved in hexafluoroisopropanol (HFIP). HFIP (green trace) appears to partially enhance the solubility of Fmoc-oxazetidine-PIV5 TM relative to TFE (black trace).

References

(1) Pusterla, I.; Bode, J. W. *Nat. Chem.* **2015**, *7*, 668–672.

(2) Pusterla, I. Synthesis and reactivity of an oxazetidine monomer for alpha-ketoacid-hydroxylamine ligation, ETH Zurich, 2014.