Supplementary Materials for

Biosynthesis of guanitoxin enables global environmental detection in freshwater cyanobacteria

Stella T. Lima, Timothy R. Fallon‡, Jennifer L. Cordoza‡, Jonathan R. Chekan, Endrews Delbaje, Austin R. Hopiavuori, Danillo O. Alvarenga, Steffaney M. Wood, Hanna Luhavaya, Jackson T. Baumgartner, Felipe A. Dörr, Augusto Etchegaray, Ernani Pinto, Shaun M. K. McKinnie,* Marli F. Fiore,* Bradley S. Moore*

‡These authors contributed equally to this work

*Corresponding authors. Email: bsmoore@ucsd.edu (B.S.M); fiore@cena.usp.br (M.F.F); smckinnie@ucsc.edu (S.M.K.M)

This PDF file includes:

Table of Contents

General materials and methods

Sphaerospermopsis torques-reginae **ITEP-024 culture**

A non-axenic culture of *Sphaerospermopsis torques-reginae* strain ITEP-024, a known producer of guanitoxin ((5*S*)-5-[(dimethylamino)methyl]-1-{[hydroxy(methoxy)phosphoryl]oxy}-4,5-dihydro-1*H*-imidazol-2-amine) was a gift from V. R. Werner (Museum of Natural Sciences, Porto Alegre, Brazil), and was maintained in conditions similar to previous descriptions (*1*). In brief, ITEP-024 cultures were grown in 50 mL autoclaved ASM-1 medium (*2*), excepting the final ASM-1 ZnCl and $CuCl_2$ concentrations were 2.5 µM and 0.01 µM, respectively, and the pH was 7.0-7.4. Cultures were grown in 125 mL borosilicate glass Erlenmeyer flasks, sealed with gas-permeable waxed paper. Cultures were either maintained under ambient laboratory light cycle, light intensity, and temperature on the benchtop, or in lighted incubators under previously described conditions (*1*). Cultures were harvested either by centrifugation, or by GF/F (Whatman, Cytiva) glass fiber filtration of trichomes. All culture manipulations were performed in biosafety or chemical fume hoods as appropriate. Cultures were biologically and chemically inactivated by 10% bleach treatment before disposal.

Transformation into *E. coli*

The plasmids were transformed into *E. coli* DH10B chemically competent cells for storage and BL21(DE3) for expression. The transformation by heat shock proceeded according to the following protocol: 0.5 µL of plasmid was added to the chemical competent cells and maintained into ice for 30 minutes. After this, the cells were heated to 42 °C for 45 seconds and placed in the ice again for 3 minutes; 900 µL of LB medium was added in the tube and the cells were incubated for 50 minutes at 37 °C and 200 rpm of agitation. After this step, the cells were plated on LB agar plates supplemented with the corresponding antibiotics. The plates were incubated at 37 °C, overnight. An inoculum with colonies was grown overnight in the same condition of LB agar plates at 37 °C and 200 rpm of agitation and purified following the protocol of Plasmid DNA Purification QIAprep Spin Miniprep Kit (QIAGEN). After plasmid purification, concentrations were measured by NanoDrop UV-vis spectrophotometry and stored at –20 °C.

Extraction and LC-MS analyses

Lyophilized culture was extracted with 5 mL of ethanol/acetic acid 0.1M (20:80 v/v), sonicated for 1 minute on ice and centrifuged at 5,000 x g for 15 minutes. The supernatant was lyophilized and resuspended in methanol and filtered with a syringe into an autosampler vial. The Hydrophilic Interaction Liquid Chromatography (HILIC) separation was carried out on a SeQuant® ZIC-HILIC, 150 x 2.1 mm, 5 µm, 200 Å (Merck) column similar to the method described in (*3*).

HILIC Method: Separation was achieved under gradient elution at 0.2 mL/min where elution A was 5 mM ammonium formate containing 0.01% (v/v) formic acid, and elution B was acetonitrile/water (90:10 v/v) with 0.01% (v/v) formic acid. Elution started with a linear gradient of 90% B to 20% until 35 min, second isocratic gradient of 20% B until 37.50 min and a third isocratic gradient of 90% B until 45 min.

RP Method: For this analysis a Synergi Polar-RP 4µ 250 x 4.6 mm column (Phenomenex) used at 0.75 mL/min with the following method: 0% B (12 min), 0 to 100% B (5 min), 100% B (3 min),

100 to 0% B (2.5 min), 0% B (2.5 min), wherein A = 0.1% aqueous formic acid, and B = 0.1% formic acid in acetonitrile.

General LC-MS measurements were measured on a Bruker Elute UHPLC system coupled with a Bruker amaZon SL ESI-Ion Trap mass spectrometer in positive mode. Compounds were separated via reversed-phase chromatography on a Bruker Intensity Solo C18(2), 2 µm- 2 x 100 mm column with the eluents water + 0.1% formic acid (Solvent A) and acetonitrile + 0.1% formic acid (Solvent B). The LC method uses a flow rate of 0.300 mL/min and the following gradient: 10% to 20% B over 3 minutes, 20% to 45% B over 3 minutes, 45% - 100% B over 2 minutes, hold at 100% B for 2 minutes, 100% - 10% B over 1 minute, hold at 10% B for 2 minutes.

Semi-preparative HPLC purification was performed using a Shimadzu Prominence preparative HPLC system with a SPD-20A model UV/Vis detector. Analytical HPLC purification was performed using an Agilent 1260 Infinity system with a G1314B model VWD. The monitored wavelength was 210 nm for all runs.

***Sphaerospermopsis torques-reginae* ITEP-024 genome sequencing and assembly**

For Illumina genome sequencing: Illumina data for ITEP-024 that had been previously quality and adaptor trimmed, was prepared in a previously described project (*1*). In brief, a Nextera XT (Illumina) sequencing library was prepared and sequenced on a MiSeq instrument to a depth of ~26M reads with 300x300 bp paired-end (PE) sequencing. In total, 13.4 Gbp (~2400x coverage) of quality and adaptor trimmed data was used for downstream analyses. The library insert size of was ~125-400 bp with a tail up to 800 bp, as determined by Qualimap v2.2.1 (*4*) analysis of reads aligned to the ITEP-024 genome assembly with bowtie2 v2.4.2 (*5*). These quality and adaptor trimmed reads are available on NCBI SRA as accession (SRR15608978).

For Nanopore genome sequencing: 50 mL of *S. torques-reginae* ITEP-024 was harvested via centrifugation, and the pellet was resuspended in 500 µL of 10 mM Tris pH 8.0. 7 mg of lysozyme was added, and the suspension incubated at 37 °C for 30 minutes. 1.5 mL of cetrimonium bromide (CTAB) buffer (3% CTAB, 1.4 M NaCl, 20 mM EDTA, 100 mM Tris pH 8.0, 3% polyvinylpolypyrrolidone (PVPP), 0.2% β-mercaptoethanol) was added, followed by 10 µL of proteinase K at 20 mg/mL and a 2-hour 50˚C incubation with mixing every 15 minutes. The sample was centrifuged at 15,000 x g for 15 minutes at room temperature, and the supernatant transferred to a new tube and added 400 µL of 5M potassium acetate pH 8.0, the tube was then kept on ice for 30 minutes to precipitate polysaccharides. The tube was centrifuged at 15,000 x g for 15 minutes at 4 °C. Half the volume of the supernatant was transferred to a new tube and extracted with 1 equivalent of phenol:chloroform:isoamyl alcohol (25:24:1). The tube was centrifuged at 12,000 x g for 5 minutes at 4 °C, and supernatant and precipitate DNA were transferred to a new tube with 1 equivalent in volume of ice-cold isopropanol. The sample was harvested at 12,000 x g for 15 minutes at 4 °C. After discarding the supernatant, the DNA pellet was washed 2 times with 0.5 mL of 75% ethanol. The DNA was air dried for 30 minutes at 37 °C and resuspended in 40 µL of 10 mM Tris pH 8.5. Sample was maintained at -20 °C until sequencing. A Nanopore ligation sequencing library (P/N SQK-LSK109) was prepared from this DNA following the manufacturer's instructions, and the resulting library sequenced via a Flongle flowcell on an Oxford Nanopore MinION sequencer. The resulting dataset was checked via NanoPlot v1.27.0 (*6*), and had an N50 of ~5.5 Kbp and a yield of ~0.2 Gbp. A systematic error was noted with an unexpectedly high

(>25%) proportion of palindromic reads. We speculate that these errors may be due to the Flongle system being a new product from Oxford Nanopore at the time of the experiment.

S. torques-reginae **ITEP-024 genome assembly & annotation**

For the Illumina MiSeq dataset, read quality was checked using FastQC v0.10.1 (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/). Reads were additionally quality filtered to trim bases with a Phred score below 25 using Prinseq (*7*). To help reduce the effect of the aforementioned Nanopore Flongle palindromic errors, palindromic sequences from the Nanopore dataset were consensus-corrected and trimmed with Canu (*8*). These quality and palindrome trimmed reads are available on NCBI SRA as accession (SRR15608977). A second correction was applied on these Canu trimmed sequences by mapping the Illumina reads with CoLoRMap (*9*). Two *de novo* hybrid assemblies were produced with Unicycler (*10*) and MaSuRCA (*11*) using both the Illumina and Nanopore datasets as inputs. The generated contigs from both assemblers were improved using SSPACE (*12*) to merge scaffolds, Pilon (*13*) for polishing and GapFiller (*14*) for filling of gaps. 26 total iterations of scaffolding, gap-filling, and polishing were performed. Subsequently, an overall assembly was generated by merging both assemblies with Flye using the subassembly feature (*15*). The quality assessment of the assemblies was done with Quast 5.0.2 (*16*) and lineage-expected gene completeness statistics and lineage-incongruent gene contamination statistics were measured with CheckM (*17*) (Fig. S3B). Two ~185 Kbp plasmids and some limited sequences from presumed contaminating heterotrophic bacteria were detected in some of the assemblies. They however did not have clear specialized metabolism related genes and could not be clearly linked as ITEP-024 sequences, so were not analyzed further. The post-assembly circularization tool Circlator version (*18*) was used to check for circularization of the *S. torques-reginae* ITEP-024 chromosomal genome. However, the software could not provide clear conclusions about the linearity or circularization of the genome. Since linear contigs produced by genome assemblers can contain the absence of short sequences that would join the contig ends, it is difficult to ensure whether the *S. torques-reginae* ITEP-024 genome is linear or circular, or if it was not circularized due to the absence of sequences in both ends. We have left the determination of a linear or circular chromosomal genome for ITEP-024 unresolved, but as circular chromosomes are most common in the cyanobacterial phylum, we speculate that the genome would be circular. The intermediate working gene annotation was performed with Prokka (*19*), while the final gene annotation on the NCBI submitted chromosomal assembly (NCBI accession CP080598.1) was produced via the NCBI Prokaryotic Gene Annotation Pipeline (PGAP).

Molecular Biology/Biochemical Methods

gnt cloning

The enzyme coding sequences were optimized for expression in *E. coli* and synthesized by GenScript Inc. Synthetic guanitoxin biosynthetic genes were sub cloned into the pET28a(+) kanamycin resistant expression vector containing an N-terminal hexahistidine (His_6) tag. The pET28a(+) plasmids containing synthetic *gnt* genes were resuspended in 100 µL of sterilized ultrapure water and transformed into *E. coli* DH10B chemically competent cells for plasmid storage and BL21(DE3) for protein expression as previously described.

pCOLADuet-*gntB-gntC* vector assembly and transformation

Using their individual pET28a(+) plasmids as templates, *gntB* and *gntC* were amplified by PCR using the primers listed in Table S2 and following amplification conditions: For primer set gntB-F/R, the following program was used: an initial denaturation at 98 °C (30 s); 30 cycles of 98 °C (10 s), 70 °C (30 s), and 72 °C (30 s); and a final extension at 72 °C (2 min). For primer set gntC-F/R, the following program was used: an initial denaturation at 98 °C (30 s); 30 cycles of 98 °C (10 s), 62 °C (30 s), and 72 °C (30 s); and a final extension at 72 °C (2 min). For primer set pCOLADuet-F/R, the following program was used: an initial denaturation at 98 °C (30 s); 30 cycles of 98 °C (10 s), 61 °C (30 s), and 72 °C (2 min); and a final extension at 72 °C (2 min). PCR-amplified *gntB* (957 bp) and *gntC* (1113 bp) were individually and sequentially added into multiple cloning sites 1 and 2 of pCOLADuet-1 respectively following Gibson Assembly Master Mix protocols (New England Biolabs).

Gnt protein expression

A general method was followed for each of the guanitoxin pathway enzymes: GntA*, GntC, GntD, GntE, GntF, GntG, GntI, and GntJ. A 20 mL starter culture of LB media containing 50 µg/mL kanamycin was inoculated with *E. coli* BL21(DE3) containing the appropriate Gnt-containing expression plasmid from glycerol stocks and shaken overnight at 37 °C and 200 rpm. The next day, 10 mL of starter culture was used to inoculate 1L of TB media containing 50 µg/mL kanamycin in a 2.8 L flask and was shaken at 37 °C and 200 rpm until the OD_{600} reached 0.9. The incubation temperature was decreased to 18 °C for one hour, followed by the addition of 100 µM isopropyl-β-D-thiogalactopyranoside (IPTG) to induce protein expression. The cultures were incubated at 18 °C and shook at 200 RPM for 20 hours. Cultures were harvested by centrifuging at 2500 x g and 4 °C for 30 minutes. Pellets were resuspended in 30 mL lysis buffer (20 mM Tris-HCl pH 8.0, 1 M NaCl, 20 mM imidazole, and 10% glycerol) and stored at -70 °C until purification.

*In the case of GntA (heme-dependent *N*-hydroxylase), a slightly modified growth protocol was used to supply heme required for proper folding of GntA. Once the OD_{600} had reached 0.4, 7 µM porcine hemin (Chem-Impex) was added to the culture and was then allowed to continue growing until an OD_{600} of 0.9 was reached.

Gnt protein purification

E. coli cell pellets containing *gnt* genes were thawed and sonicated on ice (FisherBrand Model 505 Sonic Dismembrator, 3.2 mm microtip, 40% amplitude, 15 s pulse on/45 s pulse off for a total of 7 minutes). The lysate was centrifuged at 16,000 x g for 30 minutes at 4 °C or until the supernatant had clarified. Each protein was purified using an AKTAGo FPLC system at 4 °C and buffers that had been filtered through a 0.22 µM nitrocellulose membrane. Clarified lysate was loaded onto a 5 mL HisTrap FF Column (GE Healthcare Life Sciences) that had been equilibrated with at least 25 mL of Buffer A (20 mM Tris-HCl pH 8.0, 1 M NaCl, and 20 mM imidazole) at a maximum flow rate of 2 mL/min. After loading, the column was rinsed with Buffer A until UV absorbance had returned to baseline. The column was then washed with 10% Buffer B (20 mM Tris-HCl pH 8.0, 1 M NaCl, and 250 mM imidazole) to remove weakly bound protein with at least 25 mL buffer or until UV absorbance returned to baseline. His_6-tagged protein was eluted with a linear gradient from 100% Buffer A to 100% Buffer B over 60 mL, while collecting 5 mL fractions. Fractions

were assessed for purity through SDS-PAGE (10% or 12% acrylamide, depending on protein size) and were combined if they were at least 90% pure. Protein was concentrated to a volume of 2.5 mL or less using a 10 kDa or 30 kDa cutoff (based on each protein's size) Amicon Ultra-15 concentrator. Protein was buffer exchanged into GF Buffer (50 mM HEPES pH 8.0 and 300 mM KCl) using a pre-equilibrated PD-10 gravity flow column, or further purified by size exclusion chromatography using a HiLoad 16/60 Superdex 75 column or HiLoad 16/60 Superdex 200 column, based on protein sizes and possibility of dimers (GE Healthcare Life Sciences) using a 20 mM HEPES (pH 7.5) and 300 mM KCl buffer. Protein concentration was estimated using the Bradford assay based on a Bovine Serum Albumin standard; if necessary, protein was further concentrated after this exchange. Each protein was aliquoted and stored at -70 °C for future use. The following yields of pure protein were obtained: GntA (60 mg/L), GntC (18 mg/L), GntD (35 mg/L), GntE[^] (13 mg/L), GntF (225 mg/L), GntG (29 mg/L), GntI (70 mg/L), and GntJ[#] (180 mg/L).

[^]For GntE (PLP-dependent aminotransferase), the protein eluted in the absence of PLP and upon fractionation began to aggregate in a concentration-dependent manner. 500 μM PLP and 20% glycerol was added to the combined fractions to stabilize GntE for concentration and buffer exchange. Prior to buffer exchange, any aggregate was removed by centrifuging at 20,000 x g for 10 minutes at 4 °C. After desalting into GF buffer, GntE retained PLP and was not concentrated further. No further aggregation was observed before aliquots were made.

[#]For GntJ (*O*-methyltransferase) the hexahistidine (His_6) tag was cut off using 10 units of thrombin per mg of recombinant protein incubated overnight, retaining some uncleaved protein as a control. The extent of His_6 tag cleavage was assessed by SDS-PAGE (12% acrylamide) and size exclusion chromatography using a HiLoad 16/60 Superdex 75 column was used to purify GntJ as described previously.

General procedure for Gnt enzyme assays: Marfey's Analysis and UPLC-MS conditions

A 50 μL aliquot of the Gnt reaction mixture (typically 100 μL) was removed and added to 20 μL of a saturated sodium bicarbonate solution. The addition of 100 μL of freshly-prepared 1% w/v of 1-fluoro-2,4-dinitrophenyl-5-L-alanine amide (L-FDAA, Marfey's reagent) in acetone began the derivatization reaction, which was incubated at 37 °C for 90 minutes. After incubation, reactions were quenched by the addition of 25 μL of 1N HCl, before centrifuging (15000 rpm for 5 minutes). The clarified supernatant was extracted and subjected to RP-UPLC-MS (Bruker Intensity Solo C18(2), 2.1 x 100 mm) at a flow rate of 0.300 mL/min using the following method: 10 - 20% B (3 min), 20 - 45% B (3 min), 45 - 100% B (2 min), 100% B (2 min), 100 - 10% B (1 min), 10% B (2 min), where A = 0.1% aqueous formic acid, and B = 0.1% formic acid in acetonitrile.

In vivo GntB/GtnC enzyme assay

The pCOLA-Duet-1 pET28a plasmids containing no insert (empty pET28a vector), only *gntB* (pET28a vector), and *gntB* and *gntC* in tandem, were transformed into BL21(DE3) *E. coli*. 5 mL LB starter cultures containing 50 μg/mL kanamycin of BL21(DE3) colonies grew overnight. Cells were centrifuged (3400 rpm, 5 min, 4 °C), decanted, and resuspended in 5 mL M9 minimal media. Resuspended BL21(DE3) cells (1 mL) were inoculated into 50 mL aliquots of M9 minimal media containing 30 μg/mL kanamycin. Cultures were incubated (200 rpm, 37 °C) until reaching an

OD_{600} of 0.7, then were cooled and incubated for one additional hour (200 rpm, 18 °C). Cultures were induced with 1 mM IPTG to induce GntB and GntB/C protein expression. After 5 days of incubation (200 rpm, 18 °C), 1 mL aliquots of each culture were extracted. Aliquots were sonicated (40% amplitude, 10 seconds on, 10 seconds off, for 1.5 minutes) to lyse the cells, and then cell lysates were centrifuged for 10 minutes (15000 rpm, 4 °C). From the centrifuged cell lysate, 50 µL of the supernatant was aliquoted for Marfey's derivatization and subsequent UPLC-MS analysis.

In vitro GNT enzyme assays

GntC functional assays

GntC activity assays were conducted in 50 mM K_2HPO_4 buffer (pH 8.0) using 1 mM substrate, 100 µM PLP, and 50 µM of purified GntC enzyme. Total reaction volumes were brought to 100 µL of total volume with MilliQ water. Assays were incubated at room temperature overnight, then a 50 µL aliquot was extracted for Marfey's derivatization before subsequent LC-MS analysis.

GntD functional assay

GntD activity assays were conducted in 50 mM K_2HPO_4 buffer (pH 8.0) using 1 mM substrate, 100 µM $FeSO_4$, 2.5 mM α-ketoglutarate, 50 µM L-ascorbic acid and 50 µM of purified GntD enzyme. Total reaction volumes were brought to a total volume of 100 µL using MilliQ water. Assays were incubated at room temperature for 6 hours and 50 µL aliquots were extracted after 90 minutes and 6 hours, derivatized via Marfey's reagent and analyzed by LC-MS.

GntD scale up reactions

Scaled up reactions for NMR characterization were run using the conditions previously described in 1 mL aliquots and 43 total reactions were set up. Scaled reactions ran for 14 hours at room temperature. The reactions were pooled, quenched with an equivalent amount of HPLC-grade methanol, then centrifuged at 11000 rpm for 15 min. The supernatant was concentrated *in vacuo* to remove methanol and the remaining aqueous reaction was frozen and lyophilized overnight.

HPLC Purification of GntD product β-hydroxy-L-enduracididine (**4**)

The lyophilized aqueous layer was resuspended in MilliQ water and purified by semi-preparative RP-HPLC (Phenomenex Synergi 4µm Polar RP 80 Å, 10 x 250mm) at a flow rate of 3.5 mL/min using the following method: 0.5% B (5 min), 0.5-95% B (1 min), 95% B (4 min), 95%-0.5% B (1 min), 0.5% B (4 min), where A = 0.1% aqueous formic acid, and B = 0.1% formic acid in acetonitrile. The peak containing **4** was manually collected (retention time ~3.90 min), concentrated *in vacuo* and lyophilized. The sample was then further purified by analytical RP-HPLC (Phenomenex Synergi 4µm Polar RP 80 Å, 4.6 x 250 mm) at a flow rate of 1.0 mL/min. Compound **4** eluted in 0.5% B (retention time ~2.85 min) and was manually collected, concentrated *in vacuo* and lyophilized to afford the product as a white solid. ^{1}H NMR (500 MHz, D_2O + 0.1% MeOH): δ 4.34 (ddd, J = 9.7, 6.8, 5.5 Hz, 1H), 4.06 (dd, J = 6.7, 3.1 Hz, 1H), 3.90 (d, J = 3.1 Hz, 1H), 3.83 (dd, J = 10.0, 10.0 Hz, 1H), 3.65 (dd, J = 10.3, 5.5 Hz, 1H); ^{13}C NMR (500 MHz, D_2O + 0.1% MeOH): δ 171.4, 159.4, 71.5, 56.5, 55.9, 44.9; HRMS (TOF) Calculated for $C_6H_{13}N_4O_3$ 189.0982, found 189.0982 $(M+H)^+$

GntCDGE dependence assay

The one pot assay to assess if **2** could be converted to **6** was conducted in 50 mM K_2HPO_4 buffer (pH 8.0) using 1 mM **2**. Cofactors used in this assay included 100 μM PLP, 1 mM αKG, 50 μM L-ascorbic acid, 100 μM $FeSO_4$, and 5 mM L-glutamate, with 40 μM of purified GntC, and 25 μM of purified GntD, GntG, and GntE. Assays were incubated at room temperature overnight and then 50 μL aliquots were extracted for Marfey's derivatization prior to LC-MS analysis.

GntE/G functional assay

GntE/G activity assays were performed in 50 mM K_2HPO_4 buffer (pH 8.0) using 1 mM substrate **6**, 100 μM PLP, 1 mM α-ketoglutarate, 1 mM glycine, and 50 μM purified GntG enzyme and 50 μM purified GntE enzyme. Total reaction volumes were brought to 100 μL using MilliQ water. Assays were incubated at room temperature overnight and then a 50 μL aliquot was extracted for Marfey's derivatization before subsequent LC-MS analysis.

GntF: *N*-methyltransferase

GntF activity assays were performed in 50 mM Tris buffer (pH 7.4) using 0.1 mM substrate **6**, 1 mM *S*-adenosylmethionine (SAM), and 20 μM purified GntF enzyme. Total reaction volumes were brought to 500 μL with MilliQ water and incubated at 27 °C for 18 hours. The reaction was quenched with one volume of acetonitrile on ice and filtered at 14000 x g, 4 °C for 10 minutes using both 3 kDa cutoff filters and 0.2 μm filters. The supernatant was then removed and subjected to LC-MS analysis.

GntA: *N*-hydroxylase

GntA activity assays were performed in 50 mM Tris buffer (pH 7.4) using 0.1 mM substrate **7**, 5 mM NADPH, and 20 μM purified GntA enzyme. Total reaction volumes were brought to 500 μL with MilliQ water and incubated at 27 °C for 18 hours. The reaction was quenched with one volume of acetonitrile on ice and filtered at 14000 x g, 4 °C for 10 minutes using 0.2 μm filters. The supernatant was then removed and subjected to LC-MS analysis.

GntI: kinase

GntI activity assays were performed using the filtrate of the GntA enzymatic assay, sequentially adding 2 mM ATP, 100 mM NaCl, 2 mM $MgCl_2$ and 20 μM purified GntI enzyme. The reaction volume was brought to 200 μL with 50 mM Tris buffer pH 7.4 and incubated at 37 °C for 30 minutes. The reaction mixture was quenched with one volume of acetonitrile on ice and filtered at 14000 x g, 4 °C for 10 minutes using 0.2 μm filters. Supernatant was then removed and subjected to LC-MS analysis.

GntJ: *O*-methyltransferase

GntJ activity assays were performed using the filtrate of the GntI enzymatic assay, sequentially adding 1 mM of SAM and 20 μM of GntJ enzyme without the His_6 tag. The reaction volume was brought to 100 μL with 50 mM Tris buffer pH 7.4 and incubated at 27 °C for 18 hours. The reaction mixture was quenched with one volume of acetonitrile on ice and filtered at 14000 x g, 4 °C for 10 minutes using 0.2 μm filters. Supernatant was then removed and subjected to LC-MS analysis.

Guanitoxin biosynthetic gene cluster (BGC) in publicly available environmental sequencing

Initial searches for guanitoxin BGC

Using the SearchSRA tool (*5*, *20–24*) 108,465 SRA metagenomic and metatranscriptomic datasets were searched for the guanitoxin biosynthetic gene cluster using a translated nucleotide (diamond) search with the GNT A-J peptide sequences. Hits within the 'results.zip' file provided by SearchSRA were filtered to those reads with a >= 90% protein sequence identity using a custom script (https://github.com/photocyte/gnt_paper/searchSRA_workflow).

Full sensitivity search of metatranscriptomic data for the guanitoxin BGC

All freshwater Illumina sequencing metatranscriptomic samples on SRA were selected via the following search on the NCBI web interface: "(lake[All Fields] OR pond[All Fields] OR reservoir[All Fields] OR river[All Fields] OR stream[All Fields] OR bay[All Fields] or bog[All Fields] or cyano[All Fields]) AND "ecological metagenomes"[Organism] NOT "marine metagenome"[Organism] NOT "seawater metagenome"[Organism] NOT "marine sediment metagenome"[Organism] AND "biomol rna"[Properties] AND "platform illumina"[Properties]". These search terms were empirically derived, as there was no single metadata field which applied to all freshwater metatranscriptomic samples. Metadata for these datasets was exported using the NCBI SRA Run Selector tool in CSV format. A brief script was used to filter this CSV file to a list of SRA target accession ids. A custom Nextflow (*25*) workflow (https://github.com/photocyte/gnt_paper/metaT_search_workflow) was used to download and search these datasets. The workflow was run with parameters "--sra_ids_file target_SRAs.txt --fasta ./target_genes/gnt_cds.fna -resume -with-trace -with-report". In brief, SRA reads were downloaded using fasterq-dump. Datasets resulting in .fastq files with a filesize of <160 KB were discarded and not analyzed further. Next, reads were aligned to the *gnt* BGC genes (*gntA-J*) coding nucleotide (CDS) regions, using bowtie2 (*5*) with parameters "--very-sensitive-local --no-unal". Mapped reads were summarized and processed from the resulting .bam files using a custom Jupyter (*26*) notebook (https://github.com/photocyte/gnt_paper/metaT_search_workflow/jupyter). In brief, reads were parsed into and manipulated using pandas (*27*, *28*) dataframes, while extensive metadata from the SRA database was re-added using pysradb (*29*). The results of this analysis are shown in Table S1 and Figure 3B.

Metatranscriptomic *de novo* assembly

We *de novo* assembled *gnt* BGCs from the identified metatranscriptomic SRA data, using a custom Nextflow (*25*) workflow. The workflow was run with parameters "--targets SRR5249014;SRR5249015;SRR1601415;SRR1601417;SRR1601416;SRR5834679;SRR5834616;SRR5194978 --peps gnt_peps.faa --nucl gnt_cds.fna -with-trace -resume". In brief, the workflow downloads SRA reads in .fastq format from each of the SRA datasets which had reads mapping to all 8 *gnt* CDS sequences (see Full sensitivity search of metatranscriptomic data for the guanitoxin BGC methods; Table S1) using fasterq-dump (NCBI sra-tools v2.10.8) via a bioconda environment (*30*). *De novo* transcriptome assembly was performed on the downloaded .fastq files using rnaSPAdes v3.14.1 (*31*) via a bioconda/quay.io Singularity container (*32*). The resulting *de novo* transcriptome assemblies were filtered to just the *gnt* BGC contigs using similarity seaches with the *gnt* CDSs and GNT peptides (externally provided to the workflow) via tblastn/BLAST+ (*33*) via a Singularity container. Lastly, Prokka (*19*) via a Singularity container was used to annotate the selected *gnt* BGCs. Externally provided GNT peptides were provided to Prokka to propagate our standardized *gnt* BGC naming scheme.

Metagenome Assembled Genomes & taxonomic identification of GNT BGC source

For metagenomic datasets, the read adaptors were trimmed, and quality filtered with Cutadapt (*34*). The metagenome sequences were assembled with metaSPAdes (*35*) and the resulting scaffolds were clustered into bins using the automated binning by MetaBAT (*36*). The Metagenome Assembled Genomes (MAGs) were filtered based on its sequence completeness above 70% and contamination below 10% measured by CheckM (*17*). The quality assessment of the assemblies was done with Quast 5.0.2 (*16*). GNT gene clusters were screened with BLAST+ (*33*) using an in-house script.

Taxonomic classification and phylogenomic analyses were performed with GTDB-Tk (*37*), using the Genome Taxonomy Database (GTDB) (*38*). The pipeline generated the tree through the identification and alignment of 120 bacterial single-copy conserved marker genes, then inferred the phylogeny of the concatenated sequences using FastTree (*39*) with the WAG+GAMMA models and maximum likelihood algorithm. Drawing of trees and annotations were done with iTOL v4 (*40*) and Inkscape (https://inkscape.org/). The MAGs and its closest related genomes extracted from the NCBI Genbank Database (*41*) had their average nucleotide identities calculated with OrthoANI v1.4 (*42*).

Supplementary Figures

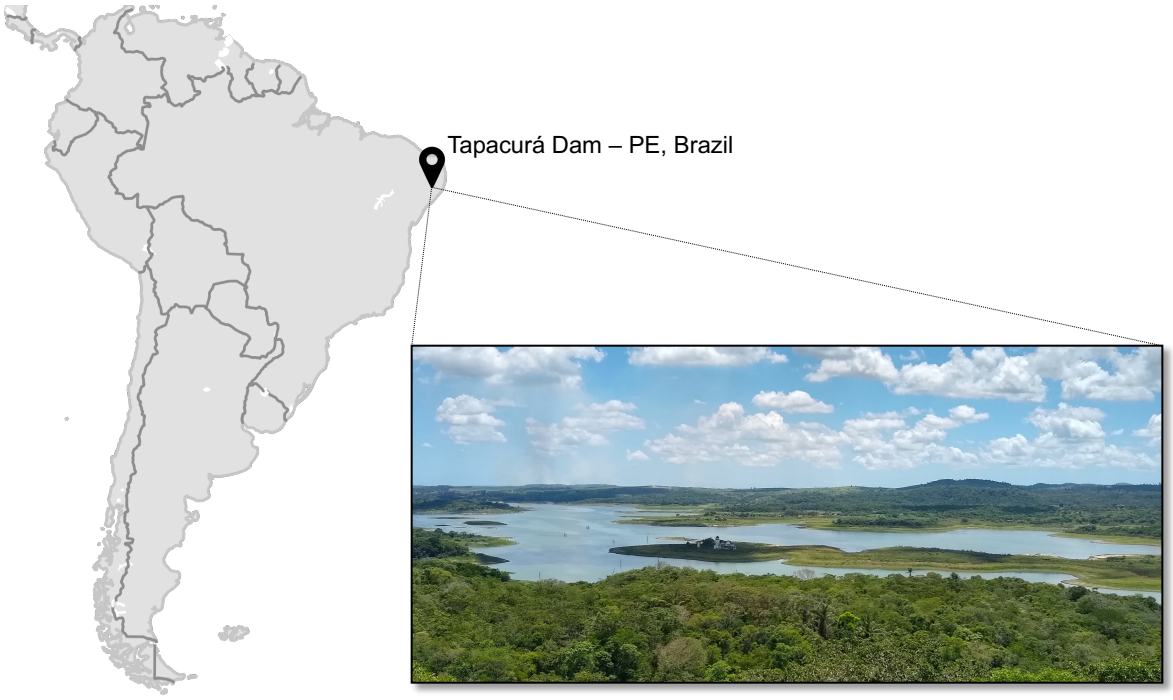


Figure S1.
Geographic location of guanitoxin-producing *Sphaerospermopsis torques-reginae* ITEP-024 cyanoHAB bloom. *S. torques-reginae* ITEP-024 bloom localization during the end of summer and beginning of fall 2002, Tapacurá Reservoir, Recife, PE – Brazil (8.037° S 35.169° W) (*43*). Photo of the Tapacurá Reservoir used with permission from Dr. Cihelio Alves Amorim (Department of Biological Sciences, Middle East Technical University, Ankara, Turkey).

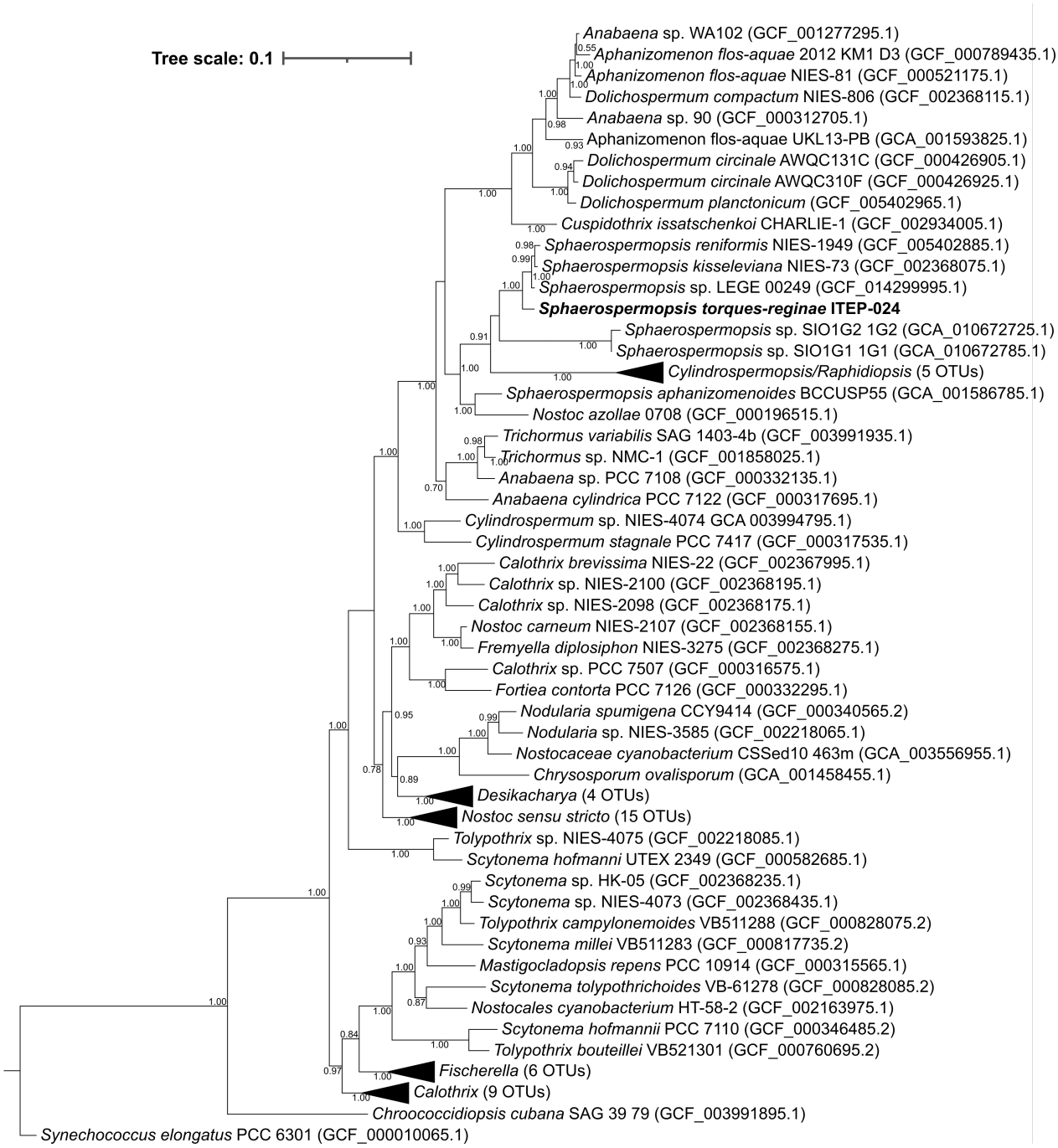


Figure S2.

Genome phylogeny for *Sphaerospermopsis torques-reginae* ITEP-024. The genome tree is inferred using GTDB-Tk (*37*) by approximately-maximum-likelihood phylogenetic analysis from an aligned concatenated set of 120 single copy marker proteins for Bacteria using Genome Taxonomy Database (GTDB) (*38*). The robustness of the phylogenetic tree was estimated via bootstrap analysis using 1000 replications. Bar: 0.1 changes per nucleotide position.

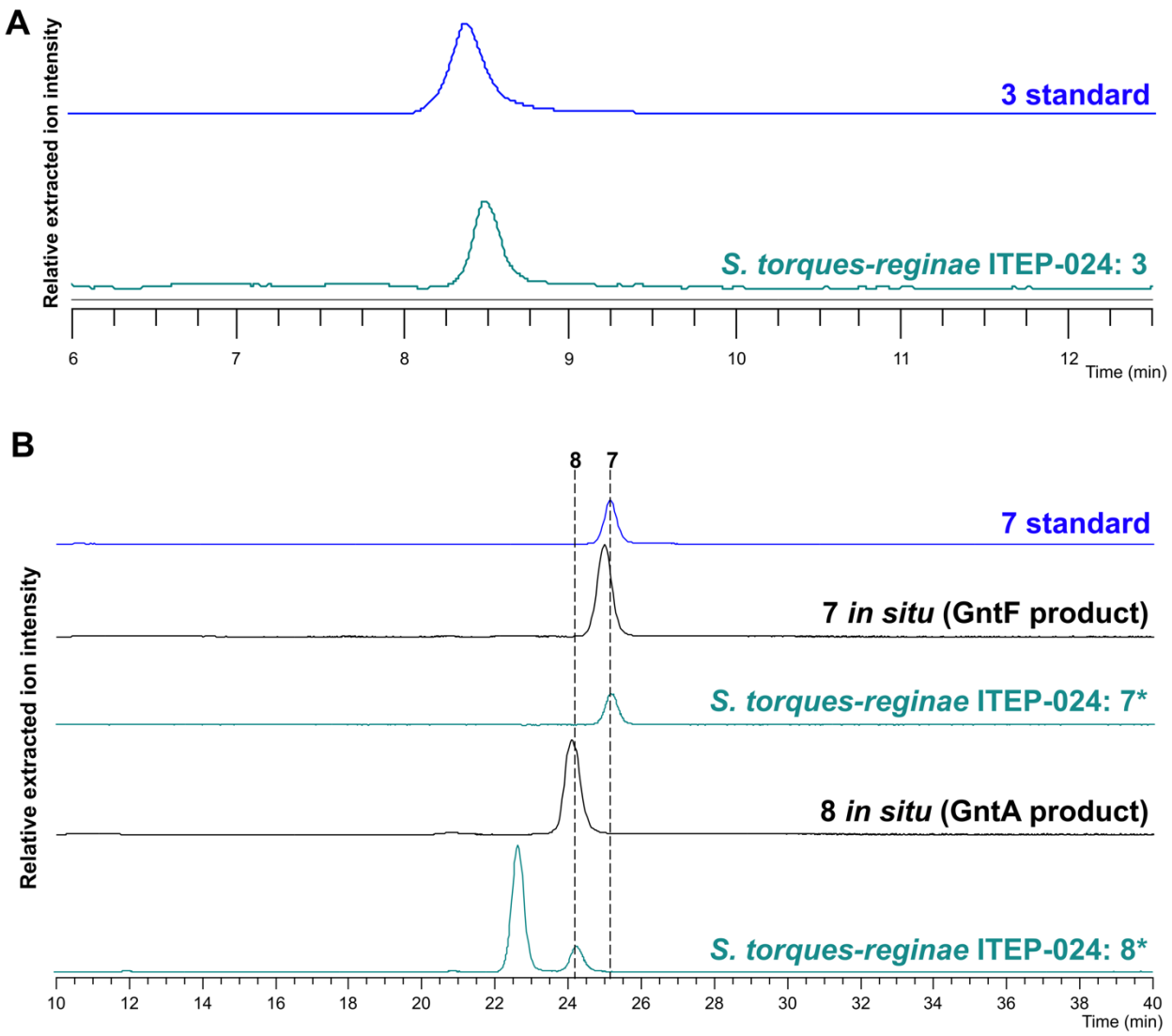
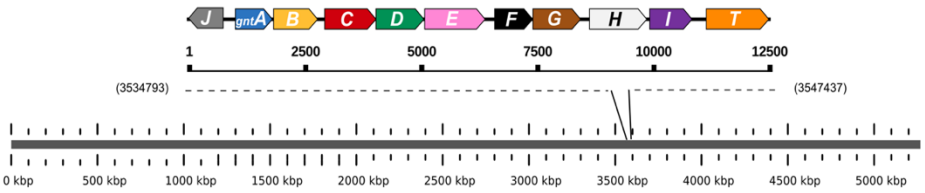


Figure S3.
Identification of guanitoxin biosynthetic intermediates **3**, **7**, and **8** in *S. torques-reginae* ITEP-024 methanolic culture extracts. Positive mode HILIC-MS chromatograms identify the presence of **3**, **7**, and **8** in the cyanobacteria culture extract as compared with synthetic standards **3** and **7**, as well as *in situ* enzyme-generated intermediates **7** and **8**. (A) Positive mode extracted ion chromatogram (EIC ± 0.0010 *m/z*) comparison for **3** ($[M+H]^+$ 173.1033 *m/z*) from the synthetic standard (blue) and cyanobacterial culture extracts (green). (B) Positive mode extracted ion chromatogram (EIC ± 0.0100 *m/z*) comparisons for **7** and **8** ($[M+H]^+$ 143.1291, 159.1240 *m/z* respectively) from synthetic standard **7** (blue), enzyme-generated **7** and **8** (black), and cyanobacterial culture extracts (green).

A



B

Genome data of *S. torques-reginae ITEP-024*

Features	*Sphaerospermopsis torques-reginae* ITEP-024
Assembly Size	5,254,486
Scaffolds	1
GC (%)	37.45
CDS	4863
Completeness (%)	99.44
Contamination (%)	0
Coverage	1916x

C

Proposed functions and similarity for proteins in the GNT BGC

Protein	Amino Acids	Proposed Function	Protein Homolog and Origin	AA Similarity (%)	Accession number
GntA	267	N-hydroxylase	YqcI/YcgG family protein [*Calothrix*]	77%	WP_096685072.1
GntB	318	Arginine hydroxylase	Fatty acid desaturase [*Calothrix* sp. NIES-2098]	52%	WP_096589832.1
GntC	370	Hydroxyarginine PLP-dependent cyclase	Aminotransferase class I/II-fold pyridoxal phosphate-dependent enzyme [*Calothrix* sp. NIES-2098]	62%	WP_096589833.1
GntD	347	Enduracididine b-hydroxylase	arginine b-hydroxylase Fe(II)/a-ketoglutarate dependent enzyme [*Calothrix desertica*]	45%	WP_127087020.1
GntE	436	Aminotransferase PLP-dependent	Aminotransferase class III-fold pyridoxal phosphate-dependent enzyme [*gamma proteobacterium BW-2*]	80%	WP_149684313.1
GntF	274	N-methyltransferase SAM-dependent	N-methyltransferase [*Leptolyngbya* sp. PCC 7375]	30%	WP_006513898.1
GntG	344	Aldolase	Low-specificity L-threonine aldolase [*Ardenticatena maritima*]	63%	WP_060687126.1
GntH	413	Phosphatase	MBL fold metallo-hydrolase [*Mesorhizobium* sp. F7]	45%	WP_052224966.1
GntI	249	Kinase	Choline/Ethanolamine kinase [*Chloroflexi bacterium OLB14*]	31%	KXK14598.1
GntJ	249	O-methyltransferase SAM-dependent	SAM-dependent methyltransferase [*Chloroflexi bacterium*]	44%	PZC47808.1
GntT	452	Transporter	MATE family efflux transporter [Scytonema hofmannii PCC 7110]	57%	KYC36324.1

Figure S4.
Sphaerospermopsis torques-reginae ITEP-024 genome and guanitoxin biosynthetic gene cluster information. (A) Genome representation with 5.2 Mbp of length and guanitoxin biosynthetic gene cluster position in the genome. (B) Genome data of *S. torques-reginae* ITEP-024 cyanobacterium. (C) Proposed functions and similarity for proteins in the Gnt biosynthetic gene cluster (BGC). Gnt protein sequences were compared by Basic Local Alignment Search Tool (BLAST) against publicly available data.

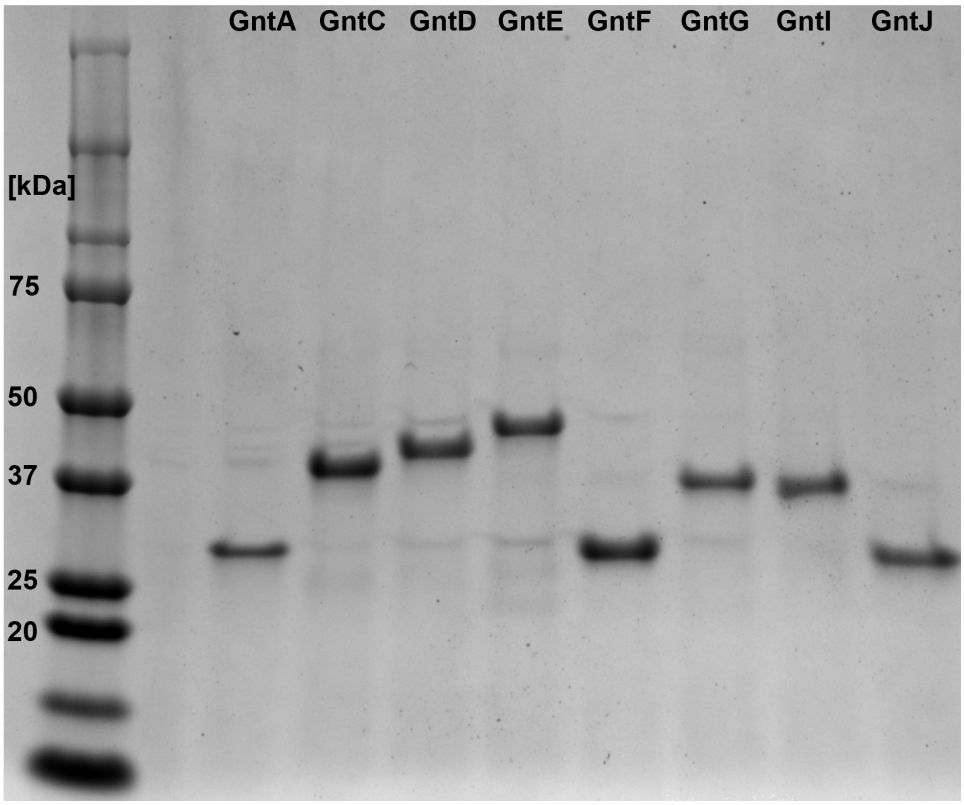


Figure S5.

SDS-PAGE analysis of GntA/C/D/E/F/G/I/J proteins. 4–15% Mini-PROTEAN® TGX™ Precast Protein Gels (Bio-Rad) loaded with Precision Plus Protein Dual Color Standards (Bio-Rad) and purified soluble Gnt pathway proteins (2 μg).

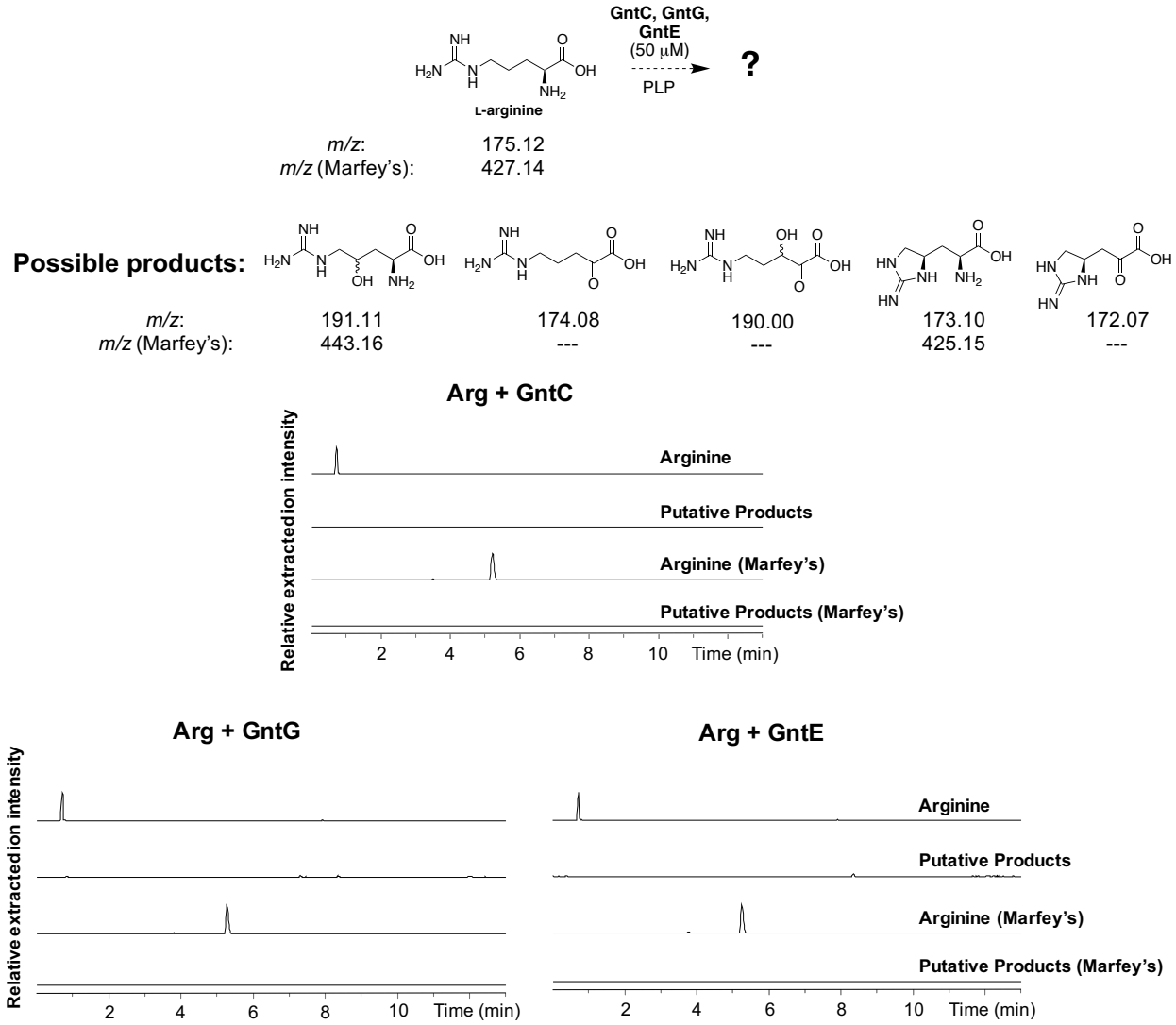


Figure S6.

Guanitoxin PLP-dependent enzymes do not react with L-arginine. GntC/GntG/GntE reactions were set up as previously described for 5 hours at room temperature. Half of each assay was methanol-quenched, while the other half was derivatized with Marfey's reagent for optimized retention times and diastereomer separation prior to UPLC-MS analysis. Relative intensities of positive mode extracted ion chromatograms were extracted from UPLC-MS traces (EIC ± 0.30 *m/z*) for non-derivatized arginine ($[M+H]^+$ 175.12 *m/z*) and derivatized arginine ($[M+H]^+$ 427.14 *m/z*), or all putative non-derivatized products ($[M+H]^+$ 191.11, 174.08, 190.00, 173.10, 172.10 *m/z*) or and all putative derivatized products ($[M+H]^+$ 443.16, 425.15 *m/z*).

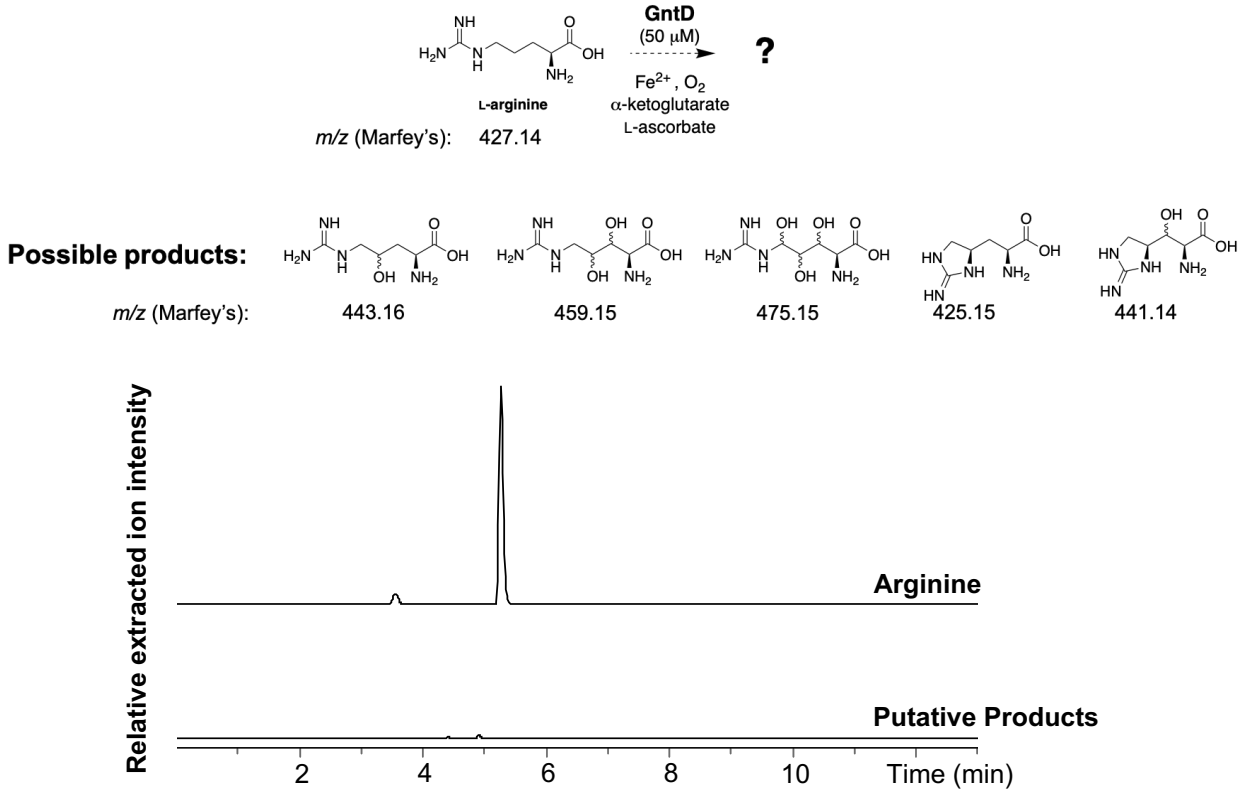


Figure S7.
GntD does not hydroxylate L-arginine. GntD reactions were set up as previously described and incubated for 4 hours at room temperature. Reactions were derivatized via Marfey's analysis for optimized retention times prior to UPLC-MS analysis. Relative intensities of positive mode extracted ion chromatograms were extracted from UPLC-MS traces (EIC ± 0.30 *m/z*) for either derivatized L-arginine ($[M+H]^+$ 427.14 *m/z*) or all putative derivatized products ($[M+H]^+$ 443.16, 459.15, 475.15, 425.15, 441.14 *m/z*).

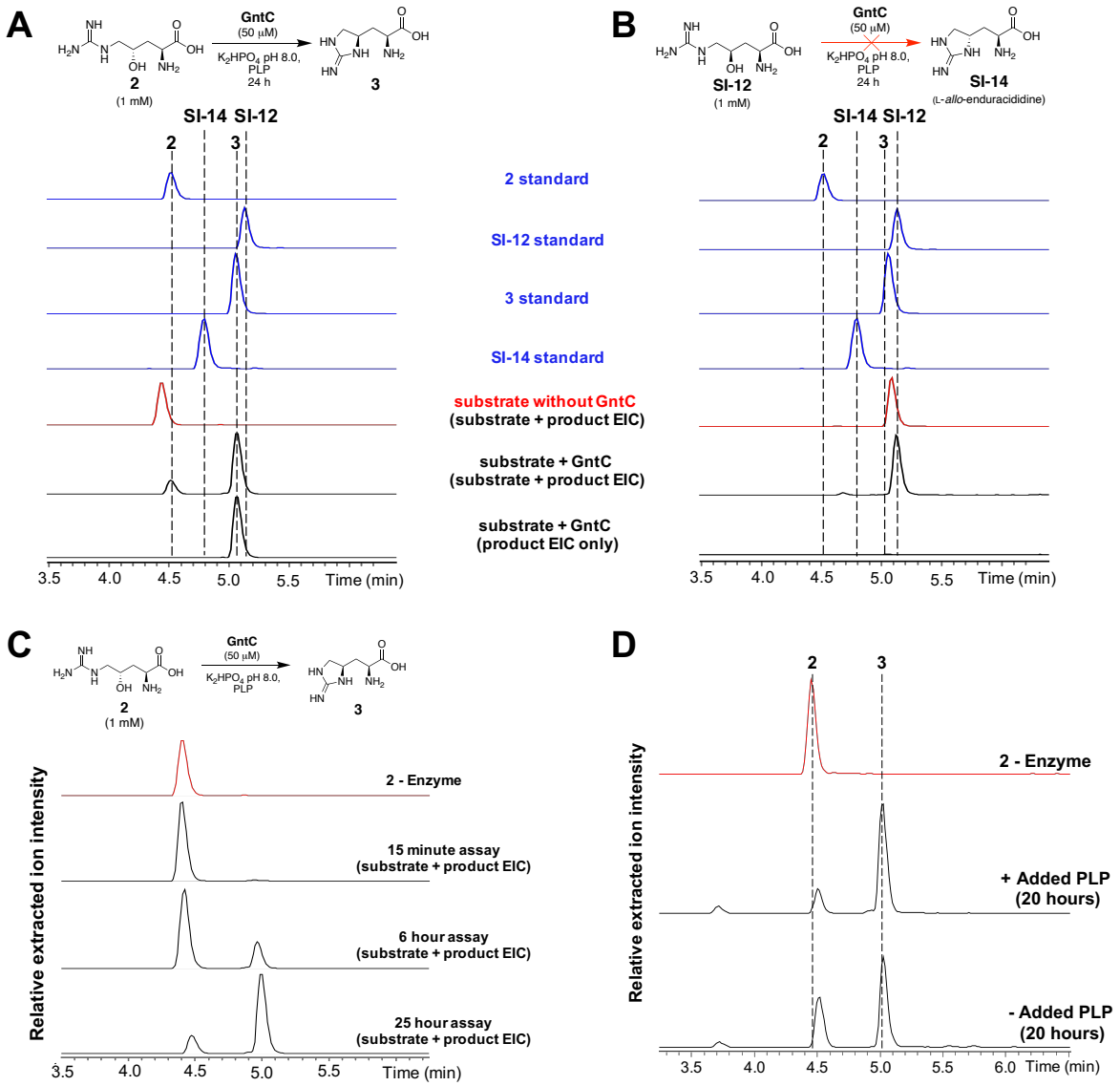


Figure S8.

GntC substrate specificity, time course, and PLP-dependence experiments. GntC assays were set up as previously described and incubated at room temperature and aliquots were taken at the time points listed on the figures (panels A and B: 25 h; panel C: 15 min, 6 h, 25 h; panel D: 20 h). Reactions were derivatized with Marfey's reagent for optimized retention times and diastereomer separation prior to UPLC-MS analysis. Relative intensities of positive mode extracted ion chromatograms were extracted from UPLC-MS traces (EIC ± 0.30 *m/z*) for Marfey derivatized starting material **2** and product **3** ($[M+H]^+$ 443.16, 425.15 *m/z* respectively) in all traces unless otherwise listed (ie the bottom traces in A and B). (**A**) GntC catalyzes the cyclodehydration of **2** *in vitro* to produce **3.** (**B**) GntC shows negligible activity towards epimerized substrate SI-12 and does not produce epimer SI-14. (**C**) GntC shows a time-dependent increase in activity over the course of the 25 h assay, and (**D**) does not need exogenous PLP for catalysis due to co-purifying with this cofactor.

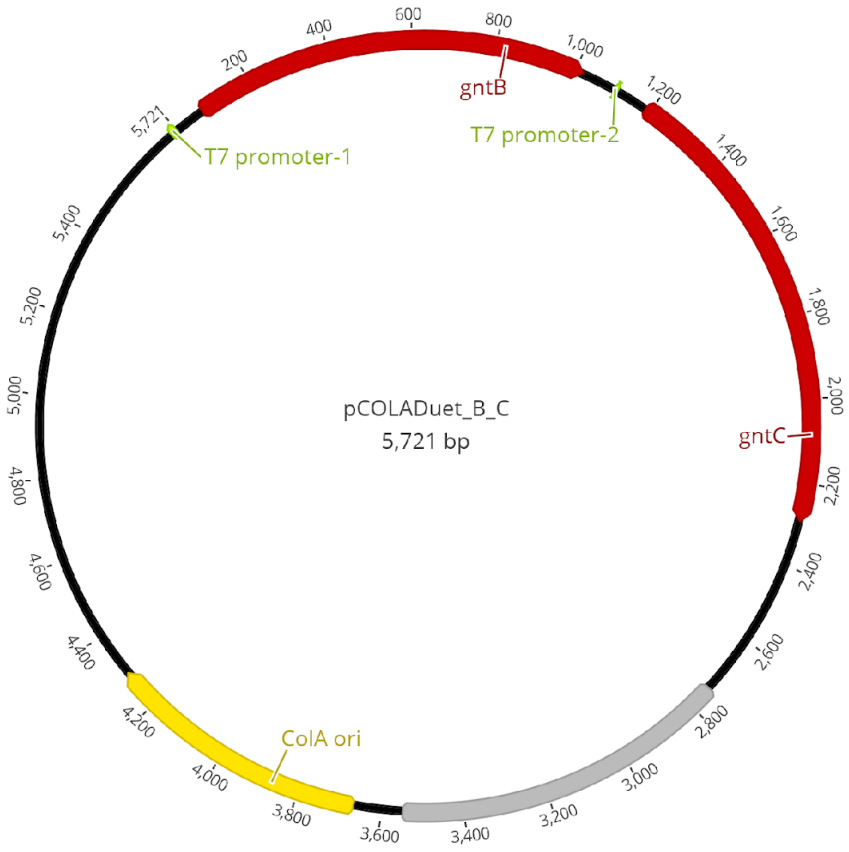


Figure S9.

gntB-*gntC*-pCOLADuet-1 vector map. pCOLADuet-1 vector assembled with *gntB* and *gntC* genes (red) from the guanitoxin pathway for enduracididine production. The vector is designed for the coexpression of two target genes from a single plasmid, which encodes two multiple cloning sites (MCS) each of which is preceded by a T7 promoter (green), *lac* operon and ribosome binding site. The vector has the COLA replicon from ColA *ori* (yellow) and kanamycin resistance gene (gray).

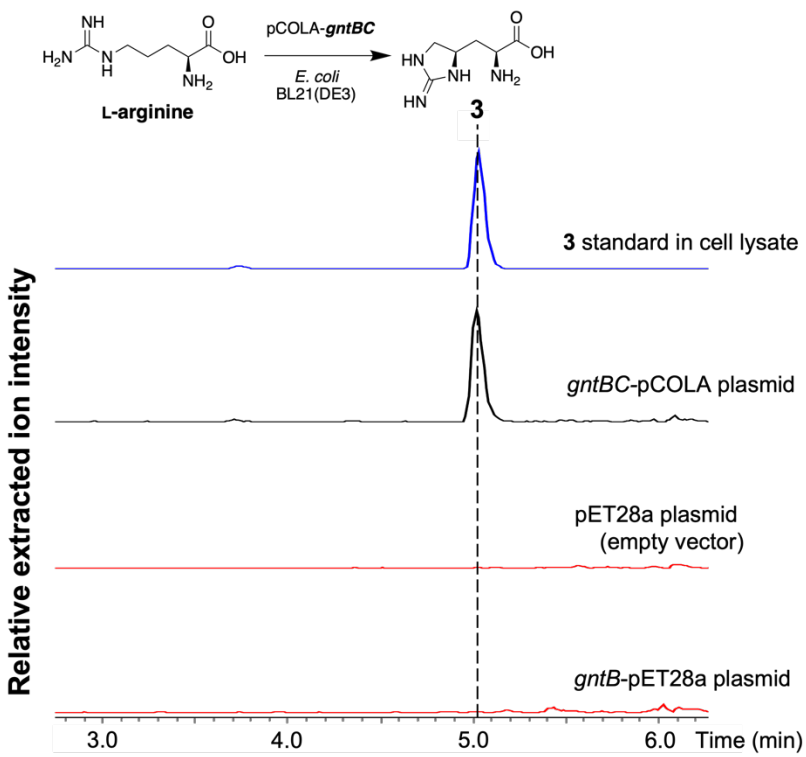
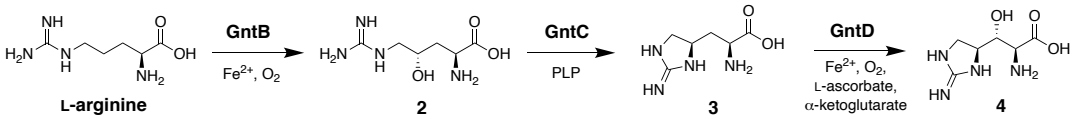


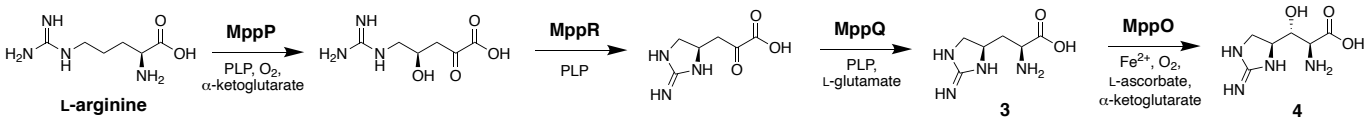
Figure S10.
GntBC produces L-enduracididine (**3**) *in vivo* in *E. coli.* The GntBC *in vivo* assay was set up as previously described and incubated at 18 °C for five days. A 100 μM internal standard of synthetic **3** was added to pET28a cell lysate (blue trace) to correct for variations in retention time based on media components. *In vivo* reactions were derivatized with Marfey's reagent prior to UPLC-MS analysis. Relative intensities of positive mode extracted ion chromatograms were extracted from UPLC-MS traces (EIC ± 0.30 *m/z*) for Marfey-derivatized GntC product **3** ($[M+H]^+$ 425.15 *m/z*). The *in vivo* production of **3** was dependent on the presence of both *gntB* and *gntC* genes (black trace) but was not observed in the *gntB*-pET28a or empty vector pET28a incubations (red traces).

guanitoxin biosynthesis – *Sphaerospermopsis torques-reginae* (cyanobacteria)



this work

mannopeptimycin biosynthesis – *Streptomyces hygroscopicus* (actinobacteria)

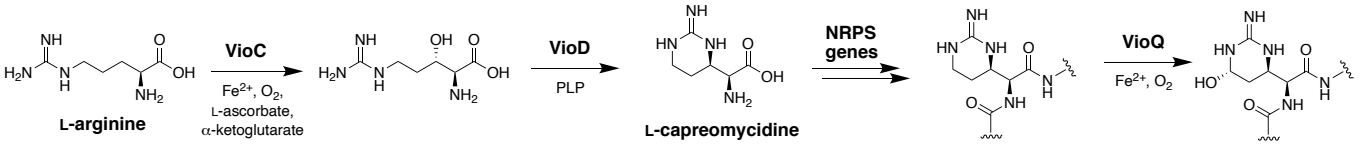


Han et al., *Biochemistry,* **2015**, *54*, 7029;
Han et al., *Biochemistry,* **2018**, *57*, 3252

Burroughs et al., *Biochemistry,* **2013**, *52*, 4492

Haltli et al., *Chem. Biol.,* **2005**, *12*, 1163

viomycin biosynthesis – *Streptomyces puniceus* and other sp. (actinobacteria)

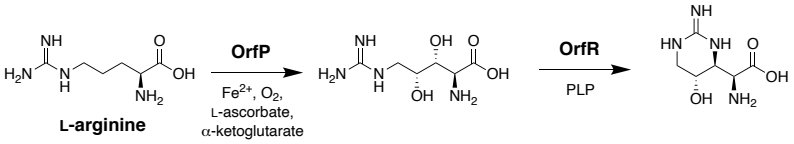


Yin et al., *ChemBioChem,* **2004**, *5*, 1274;
Ju et al., *ChemBioChem,* **2004**, *5*, 1281

Yin et al., *ChemBioChem,* **2004**, *5*, 1278;
Ju et al., *ChemBioChem,* **2004**, *5*, 1281

Fei et al., *J. Nat. Prod.,* **2007**, *70*, 618;
Barkei et al., *ChemBioChem,* **2009**, *10*, 366

streptolidine biosynthesis – *Streptomyces lavendulae* (actinobacteria)



Chang et al., *Angew. Chem. Int. Ed.,* **2014**, *53*, 1943

Figure S11.
Divergent cyclic arginine amino acid biosyntheses that use PLP-dependent enzymology. Comparison of guanitoxin and previously characterized actinobacterial biosynthetic pathways.

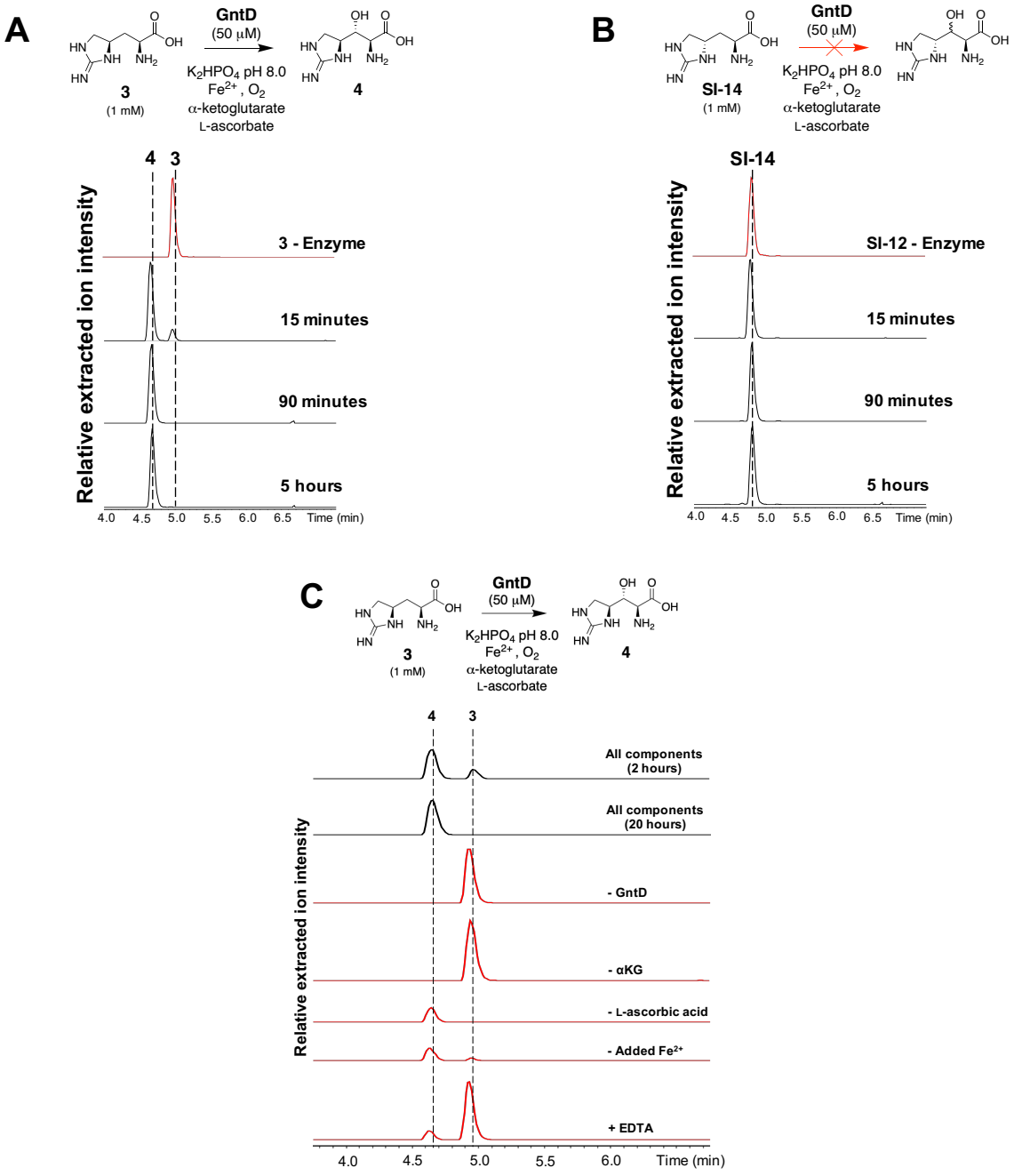


Figure S12.
GntD substrate specificity, time course and dependence experiments. GntD assays were set up as previously described at room temperature, and aliquots were taken at the time points listed on the figures (panels A and B: 15 min., 90 min., 5 h; panel C: 2 h, 20 h). Reactions were derivatized with Marfey's reagent for optimized retention times and diastereomer separation prior to UPLC-MS analysis. Relative intensities of positive mode extracted ion chromatograms were extracted from UPLC-MS traces (EIC ± 0.50 *m/z*) for both Marfey derivatized substrate **3** and product **4** ($[M+H]^+$ 424.15, 441.14 *m/z*, respectively) in all traces. (**A**) GntD rapidly hydroxylates substrate **3** *in vitro* but (**B**) shows negligible activity towards epimer SI-14. (**C**) GntD dependence assay.

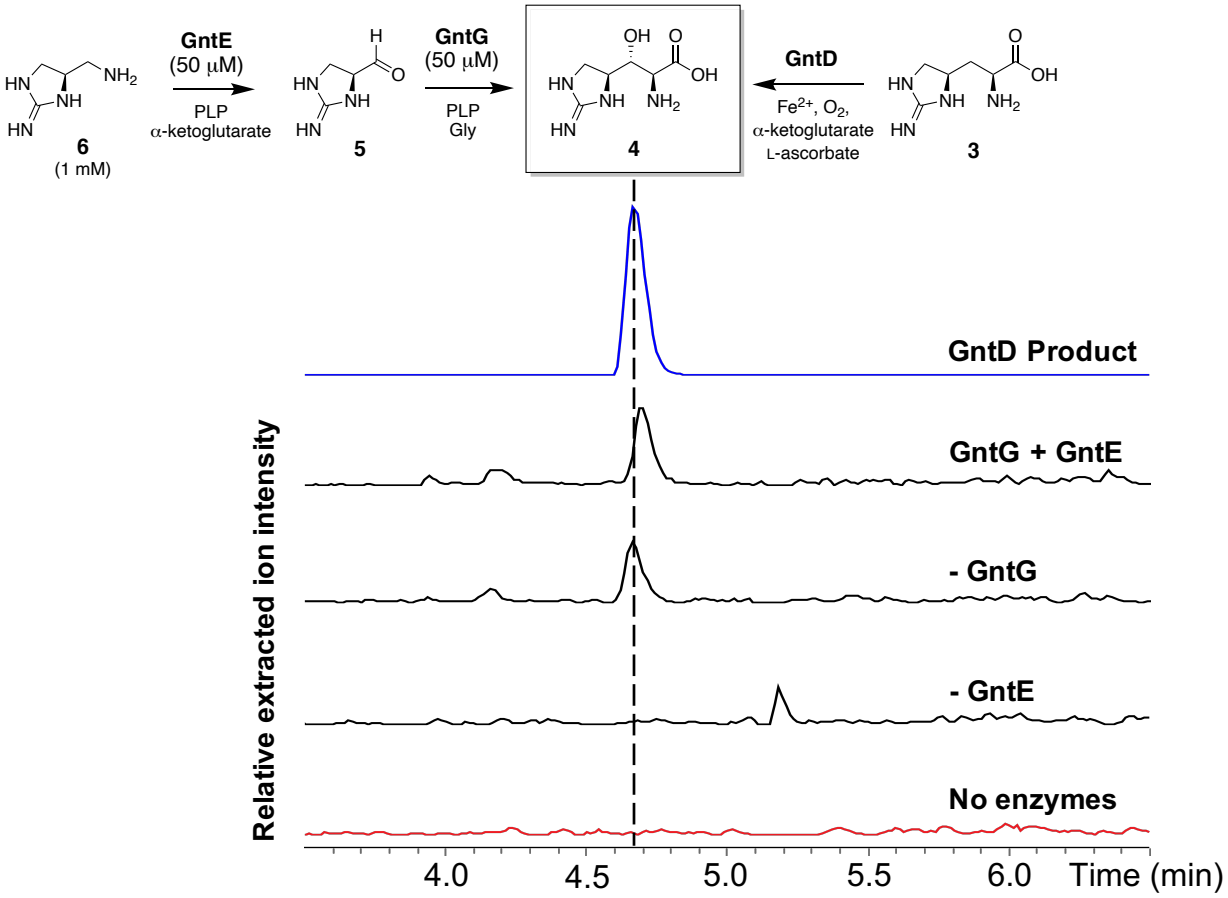


Figure S13.
GntE and GntG forward aldol assay. GntE and GntG *in vitro* aldol reaction dependence assays were set up as previously described and incubated at room temperature for 18 hours. Reactions were derivatized with Marfey's reagent for optimized retention times prior to UPLC-MS analysis. Relative intensities of positive mode extracted ion chromatograms were extracted from UPLC-MS traces (EIC ± 0.50 *m/z*) for Marfey derivatized **4** ($[M+H]^+$ 441.14 *m/z*). The enzymatically isolated GntD reaction product **4** (blue trace) was compared to the incubation of **6** with one or both GntE/G enzymes (black traces), and a no enzyme control (red trace). The inclusion of only GntE (- GntG trace) showed production of **4**, indicating that GntE may be capable of performing both aldolase and transamination chemistries. In contrast, the presence of GntG only (- GntE trace) shows no **4** production, indicating that this functional promiscuity is limited to GntE.

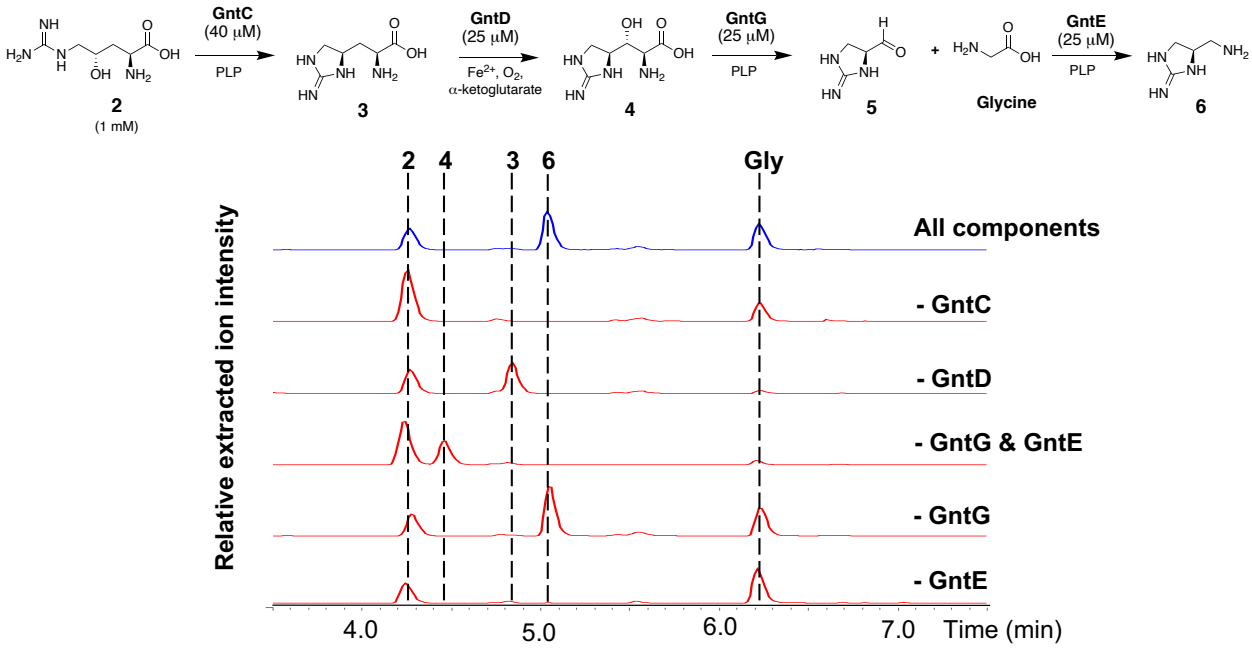


Figure S14.

GntCDGE *in vitro* one pot dependence assay. Enzyme assays were set up as previously described and incubated at room temperature for 18 hours. One condition included all enzymes (blue), while other conditions omitted one or more enzymes (red). Reactions were derivatized with Marfey's reagent prior to UPLC-MS analysis for improved retention times. Relative intensities of positive mode extracted ion chromatograms were extracted from UPLC-MS traces (EIC ± 0.20 *m/z*) for all potential products **2**, **3**, **4**, glycine, and **6** for all traces ($[M+H]^+$ 443.16, 425.15, 441.14, 328.08, and 367.14 *m/z* respectively). The reaction progression from **2** to **6** was halted depending on the omission of particular enzymes that corresponded to their native biosynthetic roles. Analogous to the results obtained in figure S10, the omission of aldolase GntG (- GntG trace) showed no significant deviation from the full assay (all component trace) due to the dual functionality of GntE as both an aldolase and transaminase. However, the omission of aminotransferase GntE (- GntE trace) supports the aldolase activity of GntG via conversion of **4** into glycine and **5** (not detected).

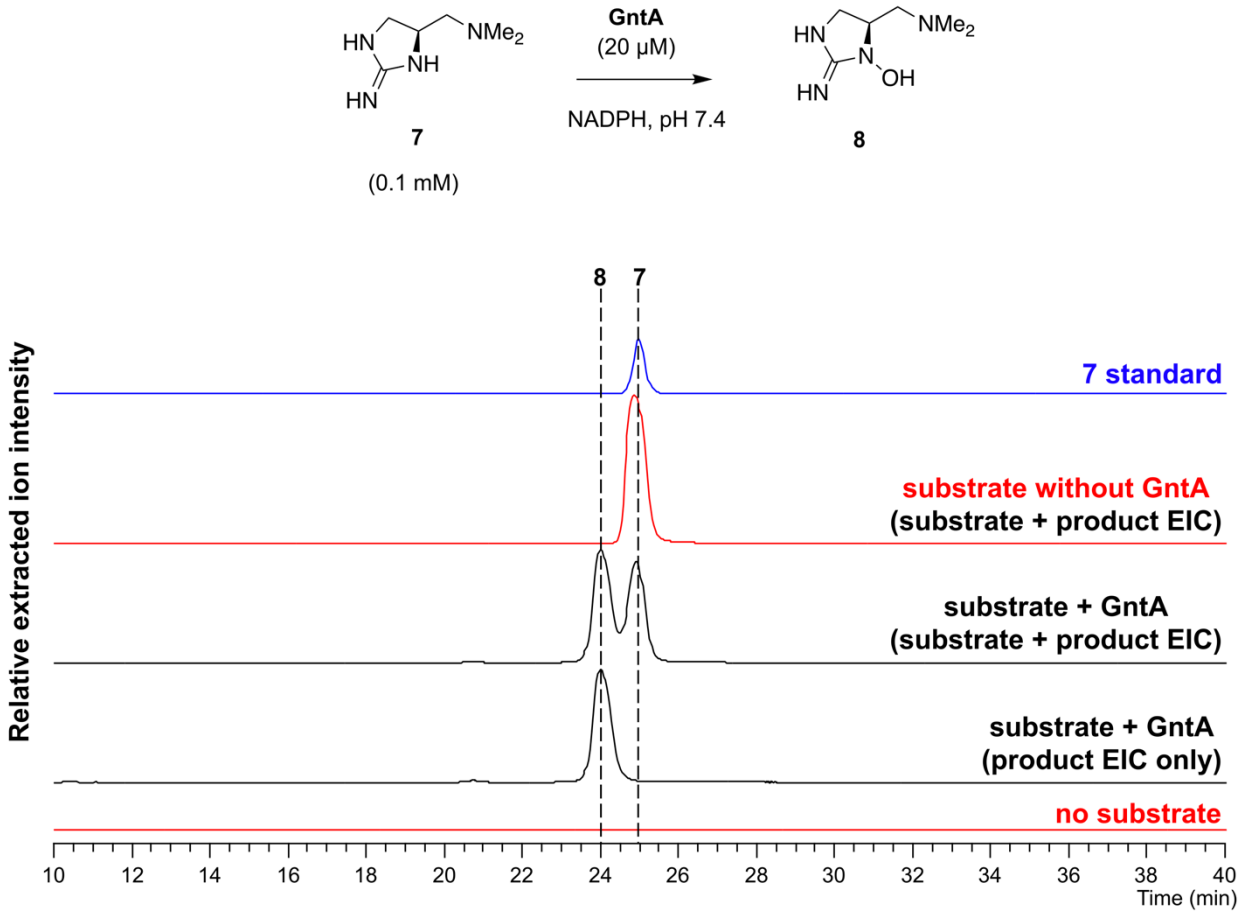


Figure S15.
GntA hydroxylates cyclic guanidine substrate **7**. GntA reactions were set up as previously described and incubated overnight at 27 ºC. Reactions were quenched with acetonitrile and subjected to HILIC-MS analysis. Relative intensities of positive mode extracted ion chromatograms were extracted from HILIC-MS traces (EIC ± 0.0100 *m/z*) for the GntA product **8** and substrate **7** masses as appropriate ($[M+H]^+$ 159.1240 and 143.1291 *m/z* respectively)

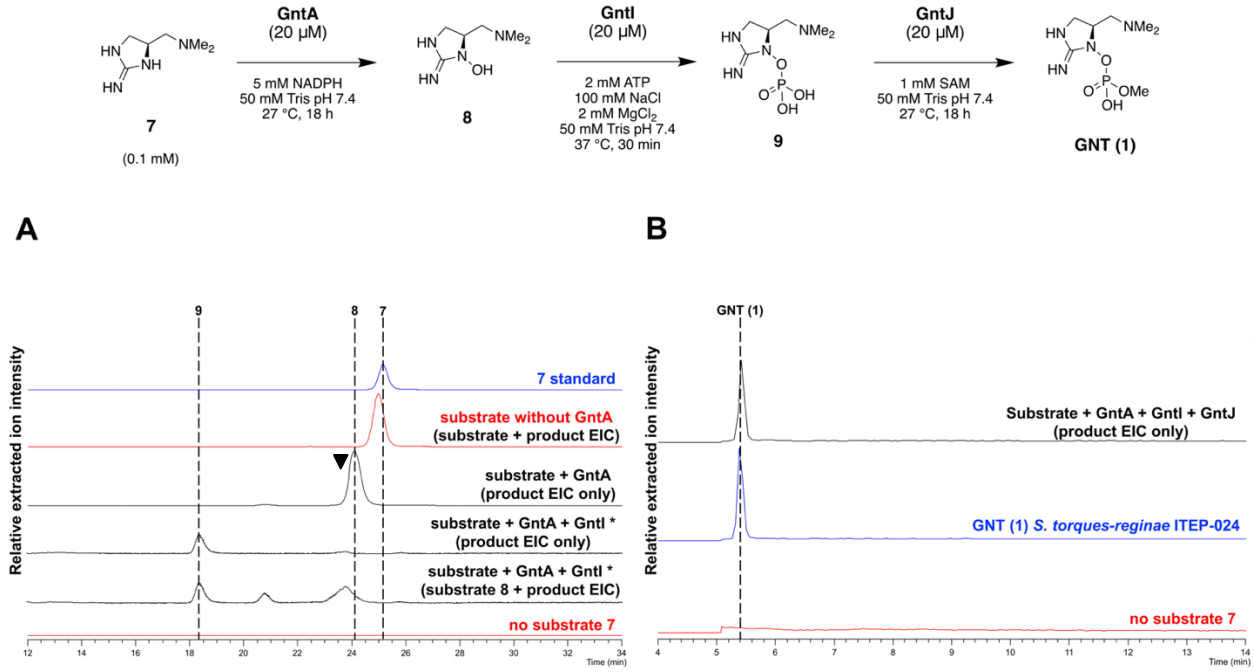


Figure S16.
GntAIJ produce guanitoxin *in situ* from synthetic substrate **7**. The GntA, GntI, and GntJ reactions were set up as previously described beginning with synthetic substrate **7**. Reactions were quenched with acetonitrile and subjected to LC-MS analysis. (A) GntA and GntI reactions were analyzed using the HILIC Method, and relative intensities of positive mode extracted ion chromatograms were extracted from HILIC-MS traces (EIC ± 0.0100 *m/z*) for substrate **7**, GntA product **8**, and GntI product **9** as appropriate ($[M+H]^+$ 143.1291, 159.1240, and 239.0904 *m/z* respectively). (B) The GntAIJ coupled reaction (black trace), no substrate control (red trace), and *S. torques-reginae* ITEP-024 extract (blue trace) were analyzed using the RP method. Relative intensities of positive mode extracted ion chromatograms were extracted from RP-LC-MS traces (EIC ± 0.0100 *m/z*) for the guanitoxin (**1**) mass ($[M+H]^+$ 253.1060 *m/z*). Asterisks indicate that the MS intensities are increased 25-fold relative to other traces for improved visualization.

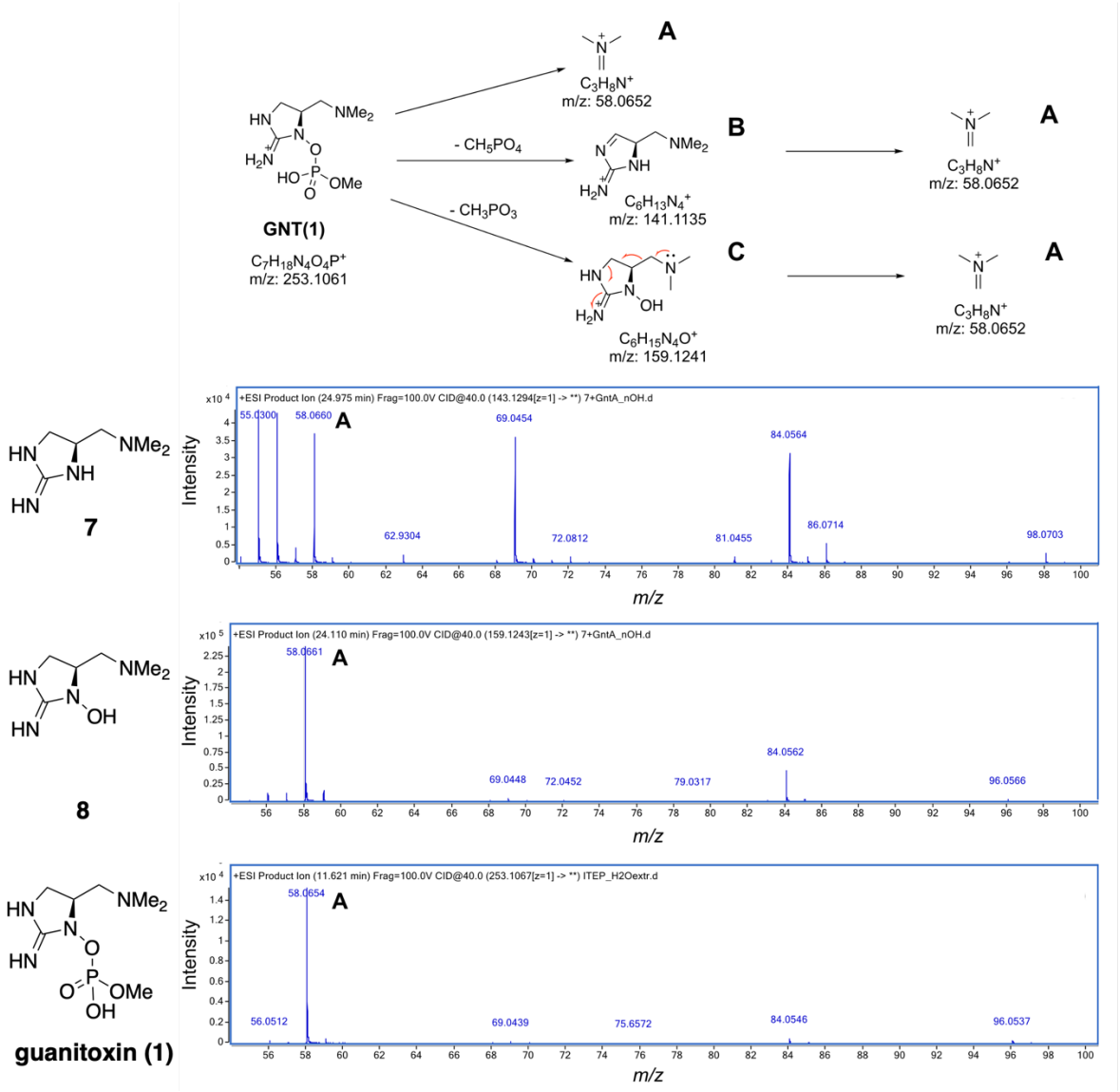


Figure S17
Mass spectrometry-mass spectrometry analyses of guanitoxin biosynthetic intermediates from *Sphaerospermopsis torques-reginae* ITEP-024. Intermediates **7**, **8**, and guanitoxin (**1**) showed diagnostic fragment A (58.0652 *m/z*) following HILIC-MS/MS analyses.

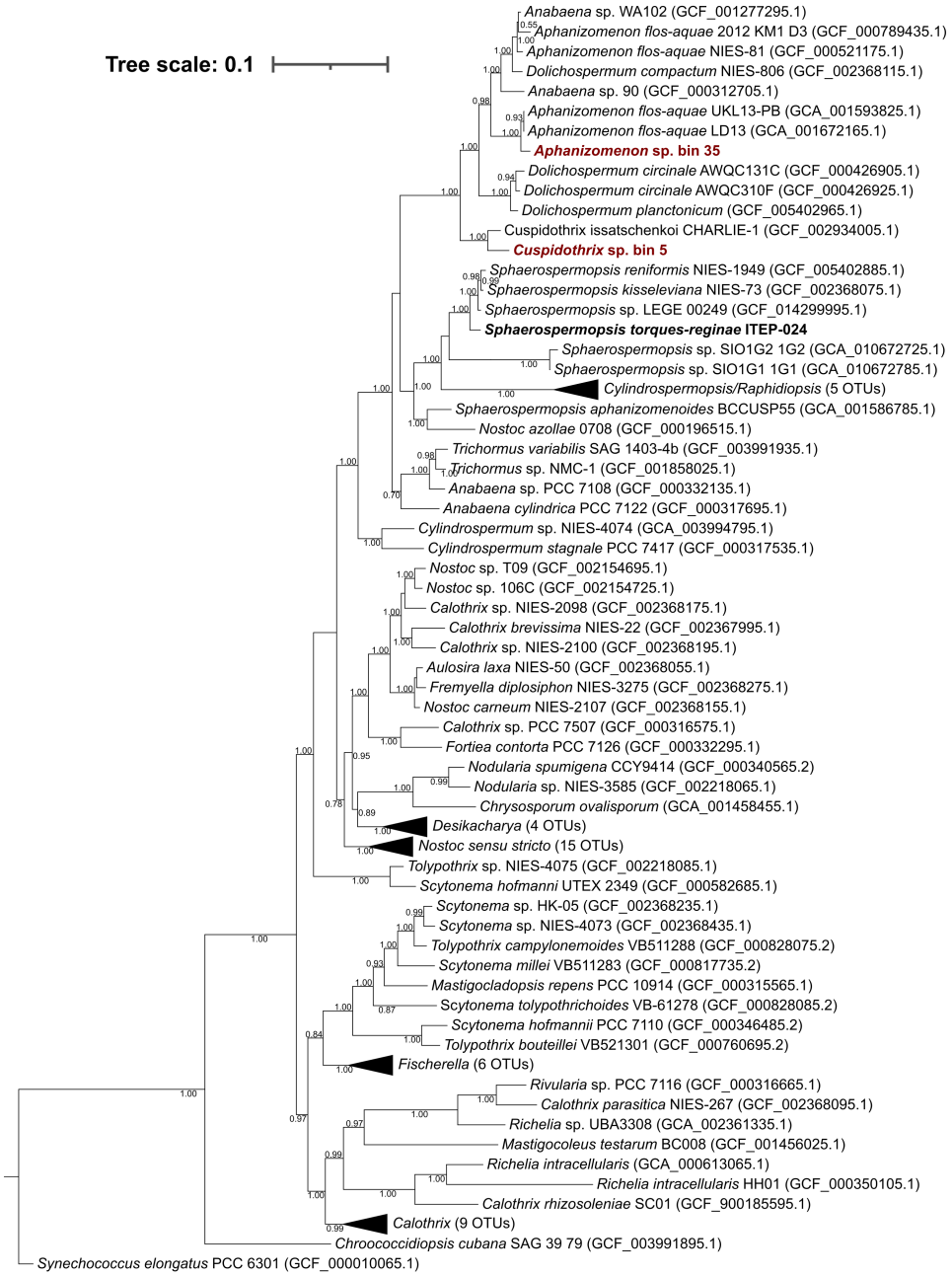


Figure S18.

Phylogenomic tree for taxonomic classification of MAGs based on Genome Taxonomy Database (GTDB) (*38*). *Cuspidothrix* bin 5 belongs to Amazon River and *Aphanizomenon* bin 35 belongs to Lake Mendota. The genome tree is generated using GTDB-Tk (*37*) by the identification and alignment of 120 bacterial single-copy conserved marker genes, then inferred the phylogeny of the concatenated sequences with the WAG+GAMMA models and maximum likelihood algorithm. The robustness of the phylogenetic tree was estimated via bootstrap analysis using 1000 replications. Bar: 0.1 changes per nucleotide position.

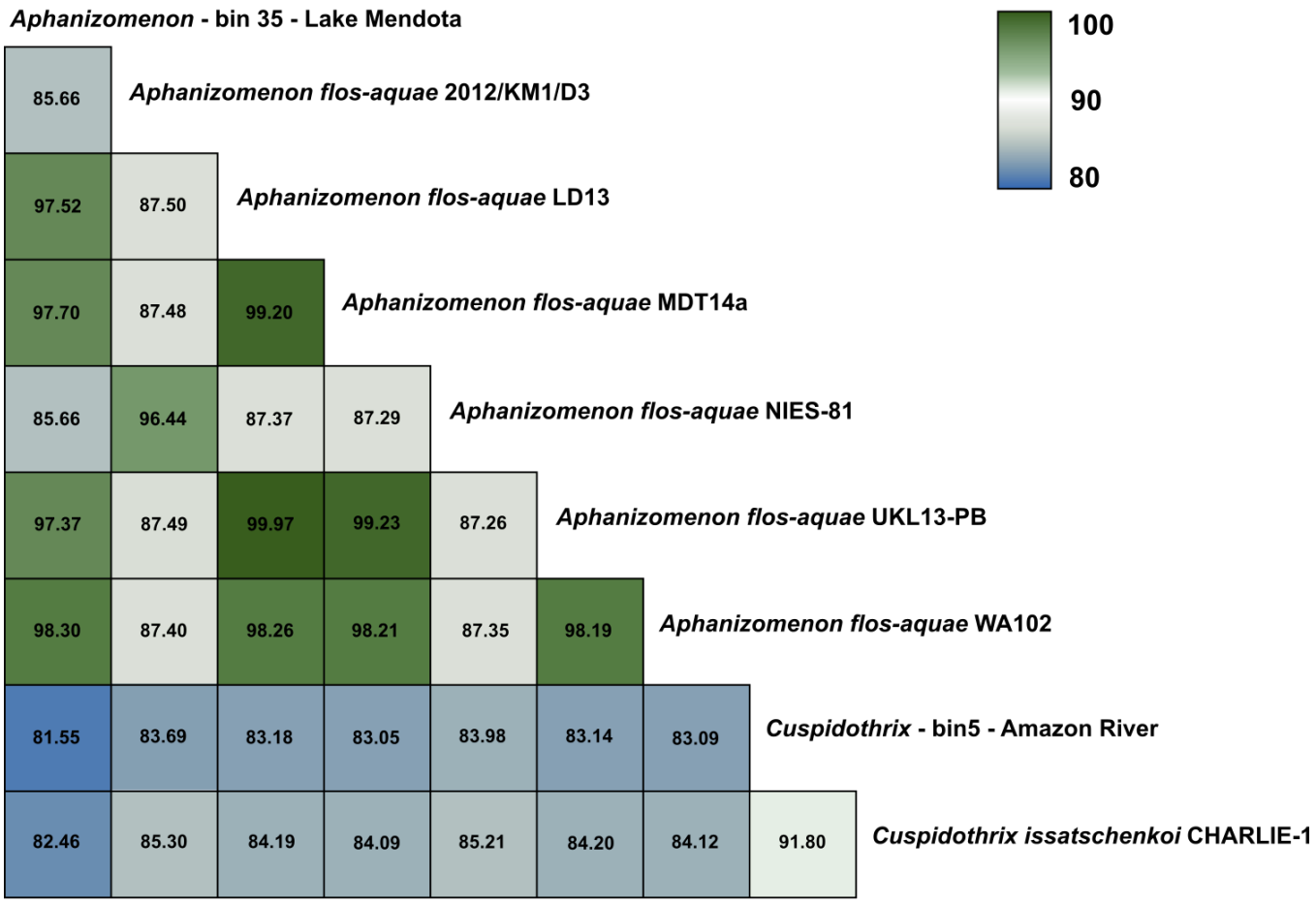
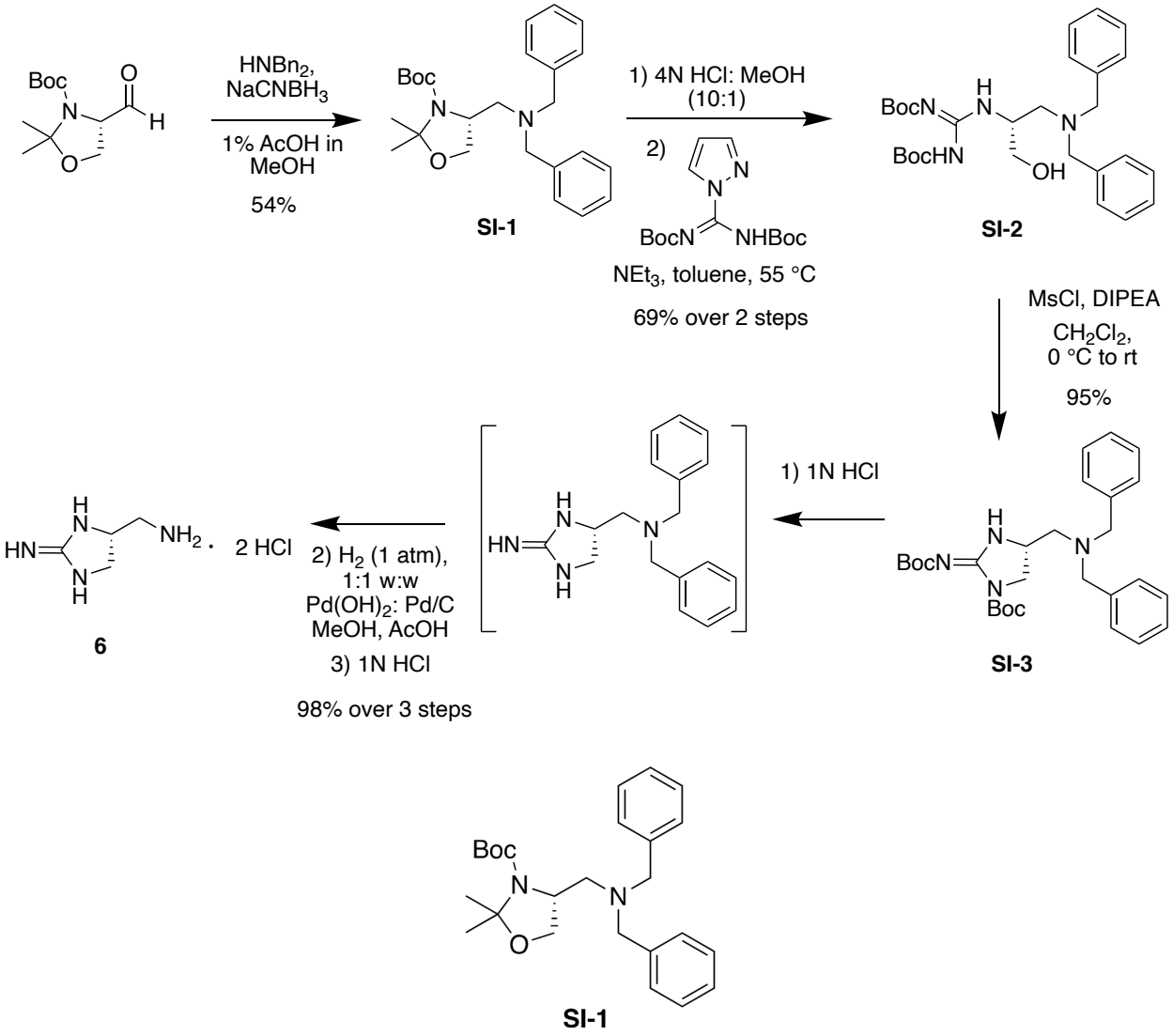


Figure S19.
Genome similarity matrix of MAG-assembled *gnt*-containing cyanobacteria. Similarity between Lake Mendota bin 35 (*Aphanizomenon*) and all *Aphanizomenon* available genomes, and Amazon River bin 5 (*Cuspidothrix*) with the only available *Cuspidothrix* genome. Average nucleotide identities were calculated with OrthoANI v1.4 (*42*).

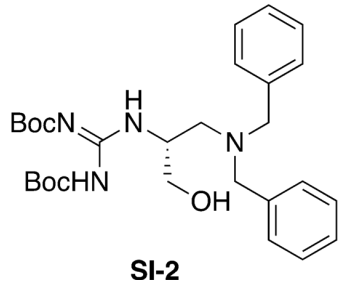
Chemical Synthesis

Synthesis of primary amine intermediate **6**



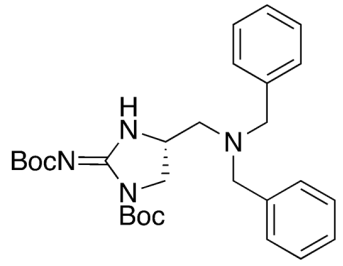
A solution of (*S*)-Garner's aldehyde (0.50 mL, 2.31 mmol) and dibenzylamine (0.47 mL, 2.43 mmol) in methanol (10 mL) had acetic acid (0.1 mL) added and was stirred for 15 minutes at room temperature. To this solution, sodium cyanoborohydride (0.145 g, 2.31 mmol) was added over a few minutes and stirred at room temperature under an argon atmosphere for 20 h. The reaction mixture was cooled to 0 °C and an aqueous solution of saturated sodium bicarbonate (25 mL) was added, followed by ethyl acetate (50 mL). The layers were separated and the aqueous layer was further extracted with ethyl acetate (2 x 50 mL). Pooled organic layers were washed with brine (50 mL), dried over magnesium sulfate, filtered and concentrated *in vacuo*. The crude reaction mixture was purified by silica flash chromatography and eluted over a gradient of 9:1 to 4:1 hexanes:ethyl acetate + 0.1% triethylamine. Pooled fractions were concentrated *in vacuo*, yielding the desired product as a clear light yellow oil (0.514 g, 54%). Rf 0.9 (2:1 hexane:ethyl acetate); IR (CH_2Cl_2 cast) 3062, 2977, 2935, 2877, 2802, 1695, 1454, 1386, 1253, 1170, 1087, 1027 cm^{-1}; ^{1}H-NMR (CD_3OD; 500 MHz, mixture of 2 rotamers) δ 7.36 – 7.28 (m, 8H), 7.26 – 7.20 (m, 2H), 4.04 – 3.99

(m, 0.4H), 3.89 – 3.84 (m, 0.6H), 3.84 – 3.69 (m, 4H), 3.40 (d, *J* = 13.5 Hz, 1H), 3.32 (d, *J* = 11.2 Hz, 1H), 2.56 (d, *J* = 9.8 Hz, 1H), 2.49 (d, *J* = 12.0 Hz, 1H), 1.50 – 1.46 (m, 10H), 1.40 – 1.36 (m, 5H), 1.31 (s, 1H); ^{13}C-NMR (CD_3OD; 125 MHz, mixture of 2 rotamers) δ 153.4, 140.8, 130.2, 130.0, 129.3, 129.3, 128.2, 128.1, 94.8, 81.6, 81.2, 79.5, 67.4, 67.1, 60.3, 59.9, 57.4, 57.1, 56.6, 28.8, 28.7, 28.0, 27.1, 24.6, 23.2; HRMS (ESI) Calculated for $C_{25}H_{35}N_2O_3$ 411.2642, found 411.2637 $(M+H)^+$.



SI-2

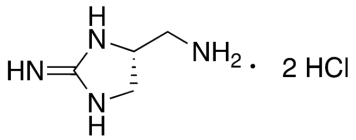
A solution of **SI-1** (0.328 g, 0.80 mmol) in 4 N aqueous HCl (10 mL) and methanol (1 mL) was stirred at room temperature for 2 h. The reaction mixture was concentrated *in vacuo*, using water and toluene co-evaporations to remove additional HCl and water respectively. The crude material was resuspended in toluene (15 mL) and had triethylamine (0.56 mL, 3.99 mmol) and *N*,*N'*-di-Boc-1*H*-pyrazole-1-carboxamidine (0.273 g, 0.88 mmol) sequentially added. The reaction mixture was heated to 55 °C and stirred for 12 h, then was cooled to room temperature, diluted with ethyl acetate (50 mL) and washed with water (2 x 25 mL). The organic layer was dried over magnesium sulfate, filtered, and concentrated *in vacuo*. The crude reaction mixture was purified by silica flash chromatography and eluted over a gradient of 9:1 to 4:1 hexanes:ethyl acetate + 0.1% triethylamine. Pooled fractions were concentrated *in vacuo*, yielding the desired product as a sticky white foam (0.283 g, 69%). Rf 0.65 (2:1 hexane:ethyl acetate); IR (CH_2Cl_2 cast) 3416, 2974, 1725, 1638, 1416, 1357, 1147 cm^{-1}; ^{1}H-NMR (CD_3OD; 500 MHz) δ 7.90 (s, 1H), 7.38 (d, *J* = 7.1 Hz, 4H), 7.27 (t, *J* = 7.5 Hz, 4H), 7.20 (t, *J* = 7.3 Hz, 2H), 5.49 (s, 1H), 4.14 (dq, *J* = 8.7, 4.3 Hz, 1H), 3.77 (d, *J* = 13.6 Hz, 2H), 3.58 (dd, *J* = 11.5, 3.6 Hz, 1H), 3.47 (dd, *J* = 11.5, 4.5 Hz, 1H), 3.42 (d, *J* = 13.5 Hz, 2H), 2.73 (dd, *J* = 13.0, 10.3 Hz, 1H), 2.51 (dd, *J* = 13.1, 4.8 Hz, 1H), 1.61 (s, 9H), 1.42 (s, 9H); ^{13}C-NMR (CD_3OD; 125 MHz) δ 164.2, 157.6, 154.0, 140.5, 130.1, 129.3, 128.1, 84.5, 80.3, 79.5, 64.2, 59.8, 55.7, 52.5, 28.5, 28.3; HRMS (ESI) Calculated for $C_{28}H_{41}N_4O_5$ 513.3071, found 513.3068 $(M+H)^+$.



SI-3

To a 0 °C solution of **SI-2** (0.091 g, 0.18 mmol) in dry CH_2Cl_2 (10 mL) was sequentially added DIPEA (62 µL, 0.36 mmol) and methanesulfonyl chloride (15 µL, 0.20 mmol). The reaction mixture was slowly warmed to room temperature over 21 h, then quenched by the addition of

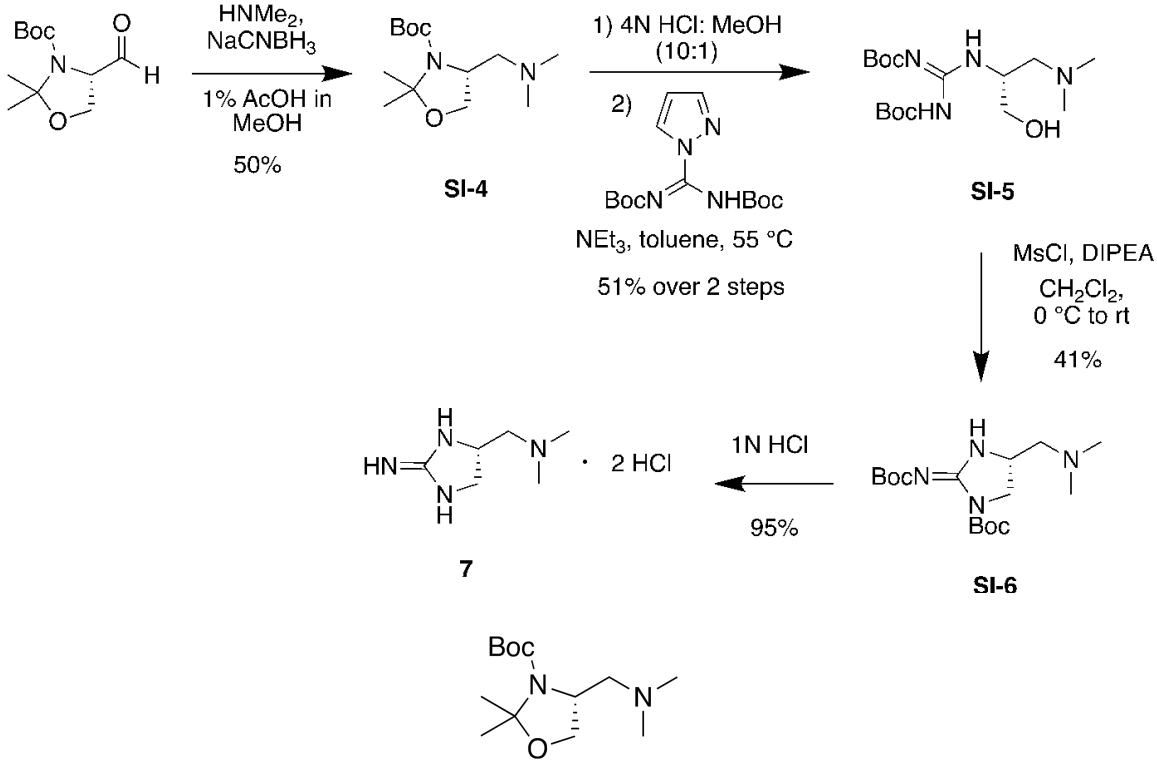
saturated NH_4Cl (20 mL) and diluted with CH_2Cl_2 (10 mL). The layers were separated and the organic layer was further washed with water (20 mL), and brine (20 mL), dried over magnesium sulfate, filtered and concentrated *in vacuo*. The crude reaction mixture was purified by silica flash chromatography and eluted over a stepwise gradient of 99:1 to 49:1 chloroform:methanol + 0.1% triethylamine. Pooled fractions were concentrated *in vacuo*, yielding the desired product as a white solid (0.083 g, 95%). Rf 0.35 (49:1 chloroform:methanol + 1 drop triethylamine); IR (CH_2Cl_2 cast) 3412, 2977, 2809, 1754, 1702, 1649, 1604, 1537, 1446, 1373, 1311, 1144 cm^{-1}; ^{1}H-NMR (CD_3OD; 500 MHz) δ 7.30 – 7.23 (m, 8H), 7.20 – 7.15 (m, 2H), 3.88 (dd, *J* = 9.8, 5.5 Hz, 1H), 3.59 – 3.48 (m, 5H), 3.26 – 3.21 (m, 1H), 2.52 (dd, *J* = 13.0, 5.7 Hz, 1H), 2.41 (dd, *J* = 12.7, 7.1 Hz, 1H), 1.45 (s, 9H), 1.41 (s, 9H); ^{13}C-NMR (CD_3OD; 125 MHz) δ 165.3, 152.1, 151.8, 140.6, 130.1, 130.0, 129.4, 128.2, 84.1, 81.2, 60.5, 60.2, 58.5, 53.2, 46.3, 28.7, 28.4, 28.3, 28.2; HRMS (ESI) Calculated for $C_{28}H_{39}N_4O_4$ 495.2966, found 495.2958 $(M+H)^+$.



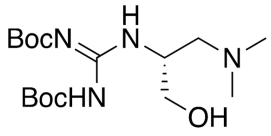
6

A solution of **SI-3** (0.045 g, 0.091 mmol) in 1N aqueous HCl (10 mL) was stirred at room temperature for 16 h, then concentrated *in vacuo*, using water and methanol co-evaporations to remove residual solvents. The crude reaction mixture was resuspended in methanol (10 mL) and had 20% $Pd(OH)_2$ (15 mg, wet) and 10% Pd/C (30 mg) sequentially added. The reaction mixture was bubbled with hydrogen gas using a balloon (1 atm) and monitored by LCMS (C18 RP-HPLC) for consumption of the di-benzylated starting material (consumed after 30 minutes) and the mono-benzylated intermediate (consumed after 22 h overnight incubation). The reaction mixture was filtered through a pad of Celite, rinsed with methanol (30 mL), then 0.1% aqueous acetic acid (30 mL) and concentrated *in vacuo*. 1 N aqueous HCl (5 mL) was added to the filtrate to obtain the HCl salt, then concentrated *in vacuo* and lyophilized. The desired product was obtained as a light yellow solid (0.017 g, 99%). ^{1}H-NMR (D_2O + 0.1% CH_3OH; 500 MHz): δ 4.42 (dddd, *J* = 10.9, 5.7, 5.7, 5.7 Hz, 1H), 3.94 (t, *J* = 10.2 Hz, 1H), 3.54 (dd, *J* = 10.5, 6.0 Hz, 1H), 3.31 – 3.19 (m, 2H); ^{13}C-NMR (D_2O + 0.1% CH_3OH; 500 MHz): δ 159.8, 52.3, 46.0, 42.1; HRMS (ESI) Calculated for $C_4H_{11}N_4$ 115.0978, found 115.0976 $(M+H)^+$.

Synthesis of dimethylamine intermediate **7**

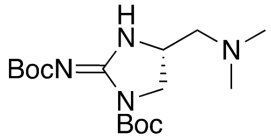


A solution of (*S*)-Garner's aldehyde (0.560 g, 2.44 mmol) and dimethylamine (1.28 mL, [2.0 M solution in THF], 2.56 mmol) in methanol (10 mL) had acetic acid (0.1 mL) added and was stirred for 5 minutes at room temperature. To this solution, sodium cyanoborohydride (0.169 g, 2.69 mmol) was added over a few minutes and stirred at room temperature under an argon atmosphere for 24 h. The reaction mixture was cooled to 0 °C and an aqueous solution of saturated sodium bicarbonate (25 mL) was added, followed by ethyl acetate (50 mL). The layers were separated and the aqueous layer was further extracted with ethyl acetate (2 x 50 mL). Pooled organic layers were washed with brine (50 mL), dried over magnesium sulfate, filtered and concentrated *in vacuo*. The crude reaction mixture was purified by silica flash chromatography and eluted over a gradient of 50:1 to 10:1 ethyl acetate:methanol + 0.1% triethylamine. Pooled fractions were concentrated *in vacuo*, yielding the desired product as a clear colorless oil (0.313 g, 50%). Rf 0.18 (10:1 ethyl acetate:methanol + 0.1% triethylamine); IR (CH_2Cl_2 cast) 3433, 2977, 2825, 2773, 1697, 1641, 1461, 1386, 1254, 1173, 1074 cm^{-1}; ^{1}H-NMR (CD_3OD; 500 MHz, mixture of 2 rotamers) δ 4.01 – 3.89 (m, 3H), 2.53 – 2.44 (m, 1H), 2.42 – 2.27 (m, 7H), 1.52 (s, 3H), 1.50 – 1.44 (m, 12H); ^{13}C-NMR (CD_3OD; 125 MHz, mixture of 2 rotamers) δ 153.2, 94.8, 94.6, 81.8, 81.3, 79.5, 67.7, 67.3, 62.8, 62.0, 56.8, 56.6, 46.3, 46.1, 28.8, 28.7, 28.1, 27.3, 24.5, 23.2; HRMS (ESI) Calculated for $C_{13}H_{27}N_2O_3$ 259.2016, found 259.2014 $(M+H)^+$.



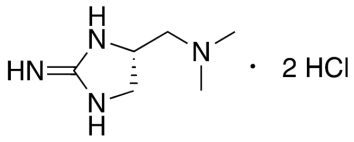
SI-5

A solution of **SI-4** (0.245 g, 0.95 mmol) in 4 N aqueous HCl (10 mL) and methanol (1 mL) was stirred at room temperature for 2.5 h. The reaction mixture was concentrated *in vacuo*, using water and toluene co-evaporations to remove additional HCl and water respectively. The crude material was resuspended in toluene (15 mL) and had triethylamine (0.66 mL, 4.74 mmol) and *N*,*N'*-di-Boc-1*H*-pyrazole-1-carboxamidine (0.324 g, 1.04 mmol) sequentially added. The reaction mixture was heated to 55 °C and stirred for 17 h, then was cooled to room temperature, diluted with ethyl acetate (50 mL) and washed with water (2 x 25 mL). The organic layer was dried over magnesium sulfate, filtered, and concentrated *in vacuo*. The crude reaction mixture was purified by silica flash chromatography and eluted over a stepwise gradient of 50:1 to 33:1 to 20:1 to 9:1 ethyl acetate:methanol + 0.1% triethylamine. Pooled fractions were concentrated *in vacuo*, yielding the desired product as a light yellow oil (0.175 g, 51%). Rf 0.15 (9:1 ethyl acetate:methanol + 0.1% triethylamine); IR (CH_2Cl_2 cast) 3324, 3139, 2978, 1726, 1639, 1565, 1462, 1415, 1320, 1257, 1161, 1053 cm^{-1}; ^{1}H-NMR (CD_3OD; 500 MHz) δ 7.90 (s, 1H), 4.26 – 4.20 (m, 1H), 3.62 (d, J = 4.1 Hz, 2H), 2.58 (dd, J = 12.8, 9.0 Hz, 1H), 2.46 (dd, J = 12.8, 5.3 Hz, 1H), 2.31 (s, 6H), 1.53 (s, 9H), 1.46 (s, 9H); ^{13}C-NMR (CD_3OD; 125 MHz) δ 164.5, 157.7, 154.1, 84.4, 80.3, 63.5, 61.3, 51.6, 46.0, 28.5, 28.2; HRMS (ESI) Calculated for $C_{16}H_{33}N_4O_5$ 361.2445, found 361.2440 $(M+H)^+$.



SI-6

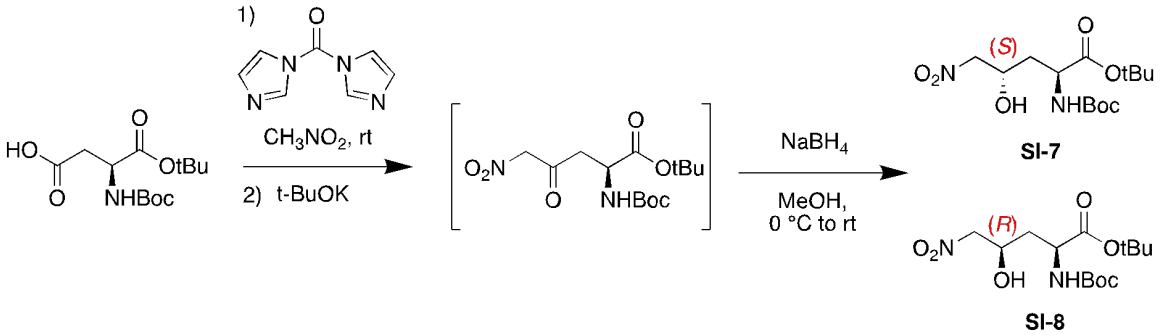
To a 0 °C solution of **SI-5** (0.122 g, 0.34 mmol) in dry CH_2Cl_2 (10 mL) was sequentially added DIPEA (118 µL, 0.68 mmol) and methanesulfonyl chloride (29 µL, 0.37 mmol). The reaction mixture was slowly warmed to room temperature over 18 h, then quenched by the addition of saturated NH_4Cl (20 mL) and diluted with CH_2Cl_2 (10 mL). The layers were separated and the organic layer was further washed with water (20 mL), and brine (20 mL), dried over magnesium sulfate, filtered and concentrated *in vacuo*. The crude reaction mixture was purified by silica flash chromatography and eluted over a gradient of 20:1 to 9:1 chloroform:methanol + 0.1% triethylamine. Pooled fractions were concentrated *in vacuo*, yielding the desired product as a white solid (0.047 g, 41%). Rf 0.4 (9:1 chloroform:methanol + 0.1% triethylamine); IR (CH_2Cl_2 cast) 3411, 2980, 1643, 1600, 1512, 1474, 1340, 1257, 1164, 1106 cm^{-1}; ^{1}H-NMR (CD_3OD; 500 MHz) δ 4.10 (dddd, J = 9.4, 6.7, 6.7, 6.7 Hz, 1H), 3.92 (dd, J = 10.7, 9.4 Hz, 1H), 3.55 (dd, J = 10.6, 6.4 Hz, 1H), 2.60 (dd, J = 12.4, 6.4 Hz, 1H), 2.49 (dd, J = 12.4, 7.1 Hz, 1H), 2.37 (s, 6H), 1.53 (s, 9H), 1.49 (s, 9H); ^{13}C-NMR (CD_3OD; 125 MHz) δ 154.7, 152.8, 152.5, 84.5, 81.8, 64.7, 50.0, 49.8, 45.9, 28.4, 28.3; HRMS (ESI) Calculated for $C_{22}H_{45}N_8O_4$ 485.3558, found 485.3557 (2M-2Boc+H)$^+$.



7

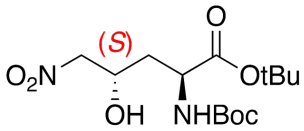
A solution of **SI-6** (0.032 g, 0.093 mmol) in 1N aqueous HCl (5 mL) was stirred at room temperature for 14 h, then concentrated *in vacuo* and lyophilized. The desired product was obtained as a light yellow solid (0.019 g, 95%). ^{1}H-NMR (D_2O + 0.1% CH_3OH; 500 MHz): δ 4.53 (dddd, J = 9.6, 9.6, 5.3, 5.3 Hz, 1H), 3.96 (dd, J = 10.2, 10.2 Hz, 1H), 3.52 – 3.42 (m, 2H), 3.36 (dd, J = 13.5, 4.8 Hz, 1H), 2.95 (s, 6H); ^{13}C-NMR (D_2O + 0.1% CH_3OH; 500 MHz): δ 159.9, 60.4, 50.3, 46.9, 44.5, 42.8; HRMS (ESI) Calculated for $C_6H_{15}N_4$ 143.1291, found 143.1291 $(M+H)^+$.

Synthesis of γ-hydroxy-L-arginine diastereomers **SI-7** and **SI-8**



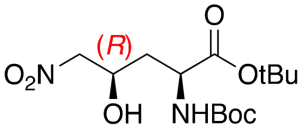
This procedure was adapted from a literature reference (*44*). Briefly, a solution of Boc-L-Asp-OtBu (2.008 g, 6.94 mmol) and 1,1'-carbonyldiimidazole (1.490 g, 6.95 mmol) in nitromethane (40 mL) was stirred at room temperature under Ar gas for 45 minutes. Potassium *tert*-butoxide (1.558 g, 13.88 mmol) was added at once and the resulting solution was stirred for an additional 4.5 hours at room temperature. The reaction mixture was quenched by the addition of 50% aqueous glacial acetic acid (50 mL), then extracted with ethyl acetate (3 x 60 mL). Pooled organic layers were washed with water (50 mL), saturated aqueous sodium bicarbonate (50 mL), water (50 mL), then brine (50 mL). The organic layer was dried over magnesium sulfate, filtered, and concentrated *in vacuo*. The crude material was resuspended in methanol (20 mL) then stirred and cooled to 0 °C. Sodium borohydride (0.262 g, 6.94 mmol) was added portion-wise to this solution over 2 minutes at 0 °C, stirred at this temperature for 45 minutes, then slowly warmed to room temperature over 16 hours. The reaction mixture was quenched by the addition of 1N aqueous HCl until a pH of 3, concentrated *in vacuo*, then resuspended in water (25 mL) and extracted with EtOAc (3 x 25 mL). Pooled organic layers were washed with brine (25 mL), dried over magnesium sulfate, filtered, and concentrated *in vacuo*. The crude reaction mixture was purified by silica flash chromatography and the two desired diastereomers were eluted over a slow gradient of 50:1 to 50:3 dichloromethane:diethyl ether. If additional silica flash chromatography was needed to further purify the two diastereomers, an isocratic 18:1:1 toluene:tetrahydrofuran:ethyl acetate

eluant system was used. Pooled fractions were concentrated *in vacuo*, yielding both desired products as flaky white solids.



SI-7

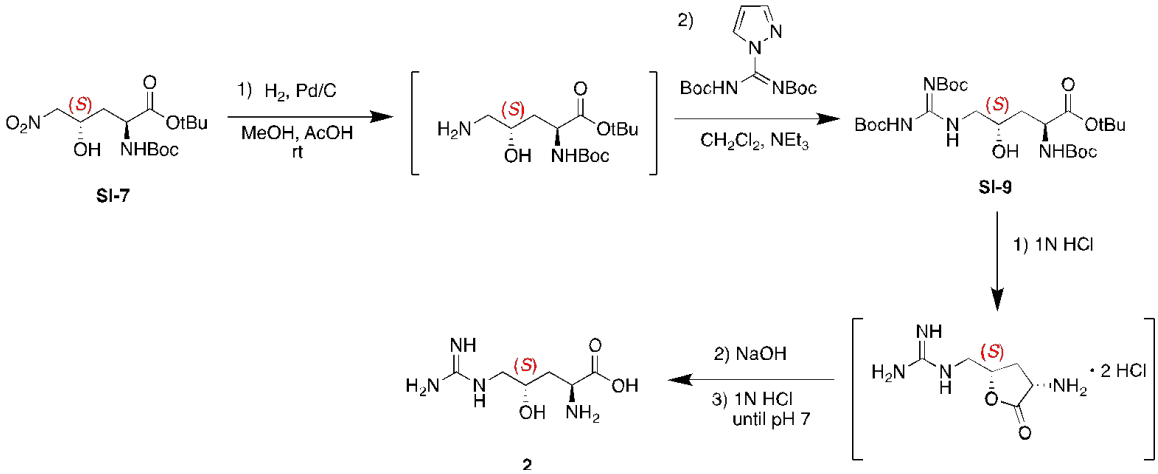
SI-7 (*S*)-diastereomer (0.628 g, 27%). Rf 0.65 (9:1 CH_2Cl_2:Et_2O); IR (CH_2Cl_2 cast) 3412, 2980, 2932, 1695, 1558, 1506, 1369, 1253, 1155 cm^{-1}; ^1H NMR (500 MHz, $CDCl_3$) δ 5.42 (d, *J* = 7.5 Hz, 1H), 4.72 (d, *J* = 4.3 Hz, 1H), 4.50 (dd, *J* = 12.0, 8.3 Hz, 1H), 4.45 – 4.30 (m, 3H), 1.95 (ddd, *J* = 13.9, 11.0, 3.1 Hz, 1H), 1.57 – 1.50 (m, 1H), 1.47 (s, 9H), 1.46 (s, 9H); ^{13}C NMR (126 MHz, $CDCl_3$) δ 171.1, 157.3, 83.2, 81.4, 80.0, 65.2, 50.7, 38.7, 28.4, 28.1; HRMS (ESI) Calculated for $C_{14}H_{27}N_2O_7$ 335.1813, found 335.1831 $(M+H)^+$.



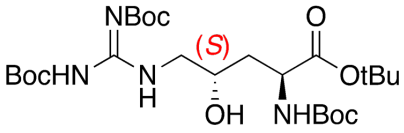
SI-8

SI-8 (*R*)-diastereomer (0.579 g, 25%). Rf 0.5 (9:1 CH_2Cl_2:Et_2O); IR (CH_2Cl_2 cast) 3419, 2979, 2918, 1695, 1557, 1384, 1369, 1252, 1155 cm^{-1}; ^1H NMR (500 MHz, $CDCl_3$) δ 5.42 (s, 1H), 4.54 (dt, *J* = 9.5, 4.8 Hz, 1H), 4.45 (d, *J* = 5.9 Hz, 2H), 4.27 (d, *J* = 6.3 Hz, 1H), 3.36 (s, 1H), 2.08 (d, *J* = 14.2 Hz, 1H), 1.90 (ddd, *J* = 14.6, 9.0, 6.0 Hz, 1H), 1.48 (s, 9H), 1.45 (s, 9H); ^{13}C NMR (126 MHz, CDCl3) δ 171.0, 155.9, 83.1, 80.8, 80.3, 66.3, 51.5, 37.1, 28.4, 28.1; HRMS (ESI) Calculated for $C_{14}H_{27}N_2O_7$ 335.1813, found 335.1828 $(M+H)^+$.

Synthesis of (*S*)-γ-hydroxy-L-arginine **2**

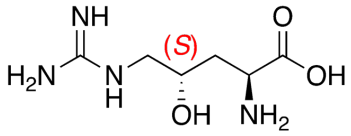


SI-7 SI-9 2



SI-9

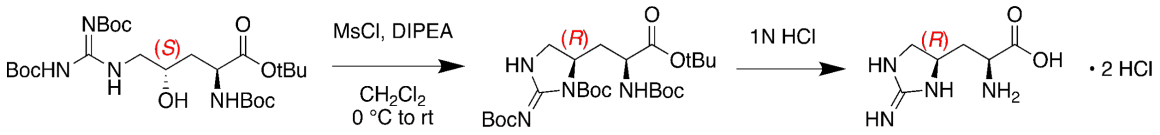
This procedure was adapted from a literature reference (*44*). A solution of **SI-7** (0.454 g, 1.36 mmol) in MeOH (15 mL) and acetic acid (0.078 mL, 1.36 mmol) had 10% Pd/C (0.145 g, 0.136 mmol) added to it and was sparged with Ar gas, then left stirring under a H_2 atmosphere for 15 hours. The reaction mixture was filtered through a pad of Celite and concentrated *in vacuo*. The crude reaction mixture was resuspended in toluene (20 mL) and had *N*,*N*'-di-Boc-1H-pyrazole-1-carboxamidine (0.464 g, 1.49 mmol) and triethylamine (0.95 mL, 6.79 mmol) added sequentially. The reaction mixture was heated to 55 °C and stirred for 17 hours, then quenched by the addition of a saturated aqueous NH_4Cl solution (25 mL). Organic components were extracted with EtOAc (3 x 25 mL), dried over magnesium sulfate, filtered and concentrated *in vacuo*. The crude reaction mixture was purified by silica flash chromatography using an eluent of 4:1 hexanes:ethyl acetate + 0.1% triethylamine. Pooled fractions were concentrated *in vacuo*, yielding the desired product as a clear light yellow oil (0.374 g, 50%). Rf 0.6 (2:1 hexane:ethyl acetate); IR (CH_2Cl_2 cast) 3337, 2980, 2932, 1728, 1645, 1621, 1368, 1155, 1055, 1027 cm^{-1}; ^1H NMR (500 MHz, $CDCl_3$) δ 11.44 (s, 1H), 8.70 (t, *J* = 5.3 Hz, 1H), 5.46 (d, *J* = 8.0 Hz, 1H), 4.99 – 4.77 (m, 1H), 4.46 – 4.26 (m, 1H), 3.84 – 3.73 (m, 1H), 3.73 – 3.64 (m, 1H), 3.28 (ddd, *J* = 13.0, 7.7, 4.1 Hz, 1H), 1.88 (ddd, *J* = 13.9, 10.7, 3.4 Hz, 1H), 1.59 – 1.53 (m, 1H), 1.49 (s, 9H), 1.48 (s, 9H), 1.46 (s, 9H), 1.44 (s, 9H); ^{13}C NMR (126 MHz, $CDCl_3$) δ 171.8, 163.4, 156.9, 156.8, 153.1, 83.3, 82.5, 80.5, 79.5, 77.4, 67.1, 51.3, 46.5, 39.0, 28.4, 28.4, 28.2, 28.1; HRMS (ESI) Calculated for $C_{25}H_{47}N_4O_9$ 547.3338, found 547.3329 $(M+H)^+$.



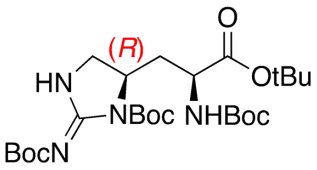
2

A solution of **SI-9** (0.037 g, 0.068 mmol) in 1N aqueous HCl (5 mL) was stirred for 3 hours then concentrated *in vacuo*, using additional water washes to remove excess HCl. The crude reaction mixture was resuspended in water, and an aqueous solution of sodium hydroxide (0.014 g, 0.34 mmol) was added and stirred for 2 hours. The reaction mixture was neutralized by the addition of 1N HCl until pH 7.0, then lyophilized. The desired product was obtained as a white solid alongside sodium chloride and was quantified using ^1H NMR with a MeOH internal standard (0.012 g, 93%). ^1H NMR (500 MHz, D_2O + 0.1% MeOH) δ 4.19 (dd, *J* = 5.9, 5.9 Hz, 1H), 3.99 (dddd, *J* = 8.2, 8.2, 4.1, 4.0 Hz, 1H), 3.34 (dd, *J* = 14.7, 3.8 Hz, 1H), 3.23 (dd, *J* = 14.6, 7.3 Hz, 1H), 2.15 – 2.04 (m, 2H); ^{13}C-NMR (126 MHz, D_2O + 0.1% MeOH) δ 172.2, 157.4, 66.6, 51.2, 46.9, 33.0; HRMS (ESI) Calculated for $C_6H_{15}N_4O_3$ 191.1139, found 191.1136 $(M+H)^+$.

Synthesis of L-enduracididine (**3**)

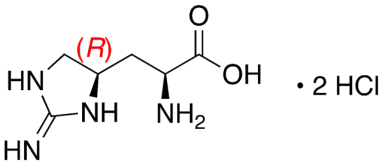


SI-9 **SI-10** **3**



SI-10

To a 0 °C solution of **SI-9** (0.246 g, 0.45 mmol) in dry CH_2Cl_2 (10 mL) was sequentially added DIPEA (0.235 mL, 1.35 mmol), then methanesulfonyl chloride (0.038 mL, 0.50 mmol). The resulting solution was stirred at 0 °C then slowly warmed to room temperature over 17 hours. The reaction mixture was diluted with additional CH_2Cl_2 (10 mL), quenched by the addition of saturated aqueous NH_4Cl (20 mL), then the layers were separated. Organic materials were extracted using subsequent washes with CH_2Cl_2 (2 x 25 mL), then pooled organic layers were washed with brine (25 mL), dried with magnesium sulfate, filtered, and concentrated *in vacuo*. The crude reaction mixture was purified using silica flash chromatography over a gradient of 2:1 to 1:1 to 1:2 hexanes:EtOAc + 0.1% triethylamine. Pooled fractions were concentrated *in vacuo*, yielding the desired product as a sticky clear colorless oil (0.083 g, 35%). Rf 0.2 (1:1 hexane:ethyl acetate + 1 drop triethylamine); IR (CH_2Cl_2 cast) 3349, 2979, 2933, 1756, 1713, 1653, 1606, 1532, 1384, 1370, 1142 cm^{-1}; ^1H NMR (500 MHz, CD_3OD) δ 4.29 (t, J = 9.4 Hz, 1H), 4.05 (d, J = 8.5 Hz, 1H), 3.81 (dd, J = 12.6, 9.1 Hz, 1H), 3.54 (dd, J = 12.7, 3.1 Hz, 1H), 2.12 (ddd, J = 11.3, 9.4, 3.6 Hz, 1H), 1.95 (dd, J = 12.3, 12.0 Hz, 1H), 1.56 (s, 9H), 1.49 (s, 9H), 1.47 (s, 9H), 1.44 (s, 9H); ^{13}C-NMR (126 MHz, CD_3OD) δ 172.6, 157.8, 152.3, 152.2, 152.2, 84.9, 83.1, 81.6, 80.6, 55.6, 52.3, 49.9, 37.0, 28.8, 28.4, 28.4, 28.2; HRMS (ESI) Calculated for $C_{25}H_{45}N_4O_8$ 529.3232, found 529.3261 $(M+H)^+$.

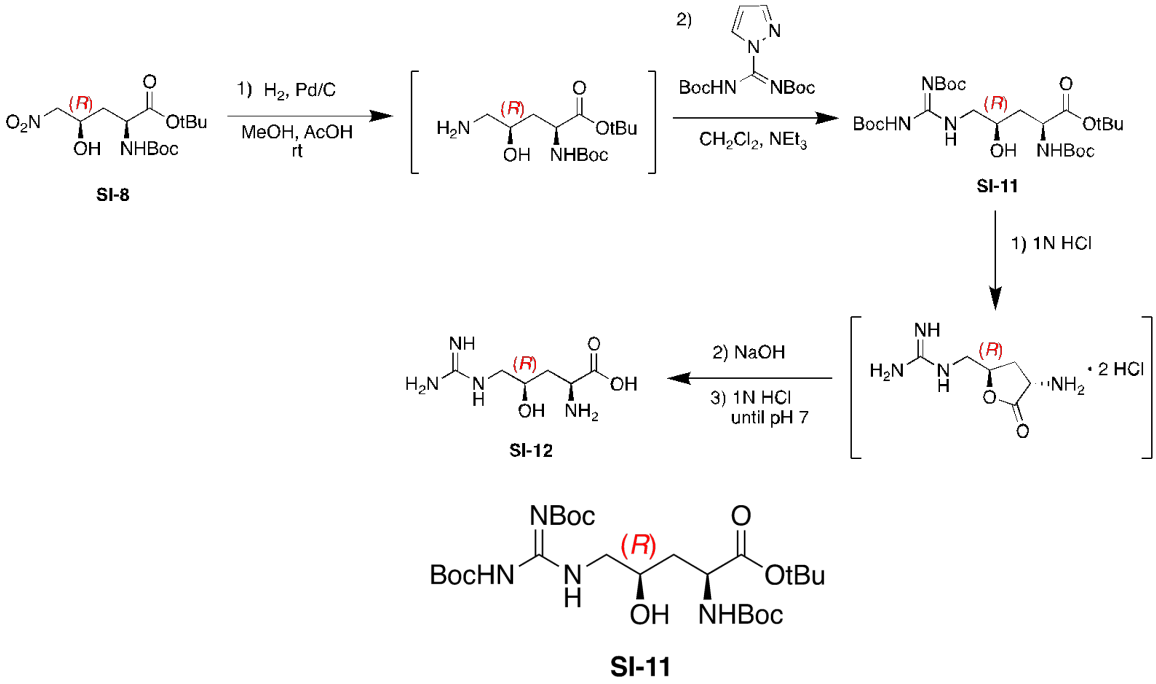


3

A solution of **SI-10** (0.067 g, 0.13 mmol) in 1N HCl (10 mL) was stirred for 3 hours and then concentrated *in vacuo*, using additional water washes to remove excess HCl, then lyophilized. The desired product was obtained as a white solid (0.032 g, quantitative). ^1H NMR (500 MHz, D_2O + 0.1% MeOH) δ 4.34 (dddd, J = 6.8, 6.8, 6.1, 6.1 Hz, 1H), 4.08 (dd, J = 9.2, 5.6 Hz, 1H), 3.88 (dd, J = 9.7, 9.7 Hz, 1H), 3.42 (dd, J = 10.1, 6.0 Hz, 1H), 2.26 (ddd, J = 14.8, 8.8, 5.8 Hz,

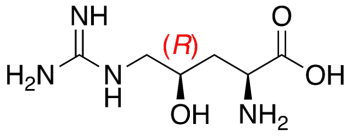
1H), 2.18 (ddd, J = 13.9, 6.7, 6.7 Hz, 1H); ^{13}C-NMR (126 MHz, D_2O + 0.1% MeOH) δ 171.4, 159.6, 52.0, 50.0, 48.3, 35.5; HRMS (ESI) Calculated for $C_6H_{13}N_4O_2$ 173.1033, found 173.1030 $(M+H)^+$.

Synthesis of (*R*)-γ-hydroxy-L-arginine (**SI-12**)



SI-11

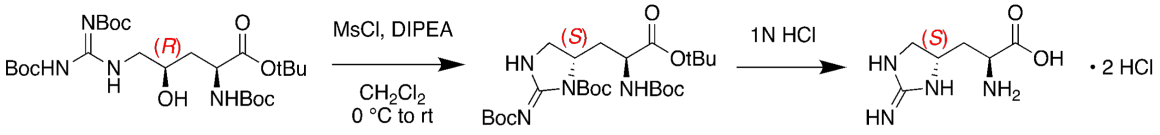
This procedure was adapted from a literature reference (*44*). A solution of **SI-8** (0.200 g, 0.60 mmol) in MeOH (10 mL) and acetic acid (0.034 mL, 0.60 mmol) had 10% Pd/C (0.064 g, 0.060 mmol) added to it and was sparged with Ar gas, then left stirring under a H_2 atmosphere for 15 hours. The reaction mixture was filtered through a pad of Celite and concentrated *in vacuo*. The crude reaction mixture was resuspended in toluene (20 mL) and had *N*,*N*'-di-Boc-1H-pyrazole-1-carboxamidine (0.204 g, 0.66 mmol) and triethylamine (0.42 mL, 2.99 mmol) added sequentially. The reaction mixture was heated to 55 °C and stirred for 17 hours, then quenched by the addition of a saturated aqueous NH_4Cl solution (25 mL). Organic components were extracted with EtOAc (3 x 25 mL), dried over magnesium sulfate, filtered and concentrated *in vacuo*. The crude reaction mixture was purified by silica flash chromatography using an eluent of 4:1 hexanes:ethyl acetate + 0.1% triethylamine. Pooled fractions were concentrated *in vacuo*, yielding the desired product as a clear light yellow oil (0.142 g, 43%). Rf 0.6 (2:1 hexane:ethyl acetate); IR (CH_2Cl_2 cast) 3339, 2980, 2920, 1726, 1645, 1621, 1368, 1157, 1054, 1027 cm^{-1}; ^{1}H NMR (500 MHz, $CDCl_3$) δ 11.42 (s, 1H), 8.69 (s, 1H), 5.48 – 5.38 (m, 1H), 5.22 (s, 1H), 4.21 (d, J = 7.9 Hz, 1H), 3.92 (d, J = 9.0 Hz, 1H), 3.59 (dd, J = 15.5, 5.9 Hz, 1H), 3.43 (dd, J = 13.5, 6.7 Hz, 1H), 1.99 – 1.92 (m, 1H), 1.86 (ddd, J = 14.4, 8.7, 6.4 Hz, 1H), 1.49 (s, 9H), 1.46 (s, 18H), 1.44 (s, 9H); ^{13}C NMR (126 MHz, $CDCl_3$) δ 171.7, 162.9, 157.7, 155.8, 153.1, 83.7, 82.2, 80.0, 79.7, 69.2, 52.1, 47.5, 37.6, 28.5, 28.3, 28.2, 28.1; HRMS (ESI) Calculated for $C_{25}H_{47}N_4O_9$ 547.3338, found 547.3328 $(M+H)^+$.



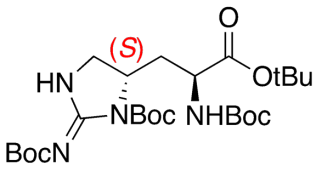
SI-12

A solution of **SI-11** (0.034 g, 0.062 mmol) in 1N aqueous HCl (5 mL) was stirred for 3 hours then concentrated *in vacuo*, using additional water washes to remove excess HCl. The crude reaction mixture was resuspended in water, and an aqueous solution of sodium hydroxide (0.012 g, 0.31 mmol) was added and stirred for 2 hours. The reaction mixture was neutralized by the addition of 1N HCl until pH 7.0, then lyophilized. The desired product was obtained as a white solid alongside sodium chloride and was quantified using ^{1}H NMR with a MeOH internal standard (0.010 g, 88%). ^{1}H NMR (500 MHz, D_2O + 0.1% MeOH) δ 4.08 (ddd, *J* = 7.4, 7.2, 3.6 Hz, 1H), 3.83 (dd, *J* = 8.1, 5.3 Hz, 1H), 3.33 (dd, *J* = 14.9, 3.1 Hz, 1H), 3.21 (dd, *J* = 14.6, 7.4 Hz, 1H), 2.10 (ddd, *J* = 14.9, 5.3, 2.9 Hz, 1H), 1.83 (dt, *J* = 14.8, 9.3 Hz, 1H); ^{13}C-NMR (126 MHz, D_2O + 0.1% MeOH) δ 174.3, 157.3, 68.4, 53.5, 47.0, 34.1; HRMS (ESI) Calculated for $C_6H_{15}N_4O_3$ 191.1139, found 191.1135 $(M+H)^+$.

Synthesis of L-*allo*-enduracididine (**SI-14**)



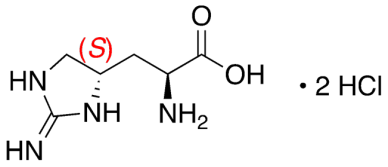
SI-11 **SI-13** **SI-14**



SI-13

To a 0 °C solution of **SI-11** (0.124 g, 0.23 mmol) in dry CH_2Cl_2 (10 mL) was sequentially added DIPEA (0.12 mL, 0.68 mmol), then methanesulfonyl chloride (0.019 mL, 0.25 mmol). The resulting solution was stirred at 0 °C for 2 hours, then slowly warmed to room temperature over 17 hours. The reaction mixture was diluted with additional CH_2Cl_2 (10 mL), quenched by the addition of saturated aqueous NH_4Cl (20 mL), then the layers were separated. Organic materials were extracted using subsequent washes with CH_2Cl_2 (2 x 25 mL), then pooled organic layers were washed with brine (25 mL), dried with magnesium sulfate, filtered, and concentrated *in vacuo*. The crude reaction mixture was purified using silica flash chromatography over a gradient of 4:1 to 1:1 hexanes:EtOAc + 0.1% triethylamine. Pooled fractions were concentrated *in vacuo*, yielding the desired product as a sticky clear colorless oil (0.094 g, 78%). Rf 0.2 (1:1 hexane:ethyl acetate + 1 drop triethylamine); IR (CH_2Cl_2 cast) 3366, 2979, 2929, 1747, 1714, 1606, 1539, 1384, 1369, 1311, 1252, 1149 cm^{-1}; ^{1}H NMR (500 MHz, CD_3OD) δ 4.30 (dddd, *J* = 12.2, 8.3, 3.3, 3.1 Hz, 1H), 4.09 (dd, *J* = 8.9, 5.4 Hz, 1H), 3.84 (dd, *J* = 12.8, 9.2 Hz, 1H), 3.55 (dd, *J* = 12.8, 3.5 Hz, 1H), 2.23

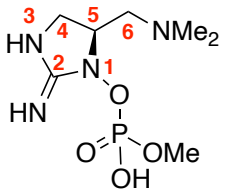
(ddd, J = 13.9, 5.4, 3.0 Hz, 1H), 1.90 (ddd, J = 13.9, 8.8, 8.8 Hz, 1H), 1.56 (s, 9H), 1.49 (s, 9H), 1.48 (s, 9H), 1.45 (s, 9H); ^{13}C-NMR (126 MHz, CD_3OD) δ 172.7, 157.8, 152.4, 152.4, 152.4, 85.0, 83.1, 81.8, 80.6, 56.4, 53.3, 37.2, 28.8, 28.4, 28.4, 28.3; HRMS (ESI) Calculated for $C_{25}H_{45}N_4O_8$ 529.3232, found 529.3266 $(M+H)^+$.



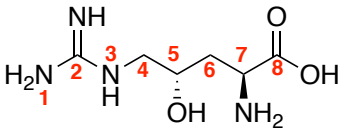
SI-14

A solution of **SI-13** (0.069 g, 0.13 mmol) in 1N HCl (10 mL) was stirred for 3 hours and then concentrated *in vacuo*, using additional water washes to remove excess HCl, then lyophilized. The desired product was obtained as a white solid (0.032 g, quantitative). ^{1}H NMR (500 MHz, D_2O + 0.1% MeOH) δ 4.31 (ddd, J = 13.4, 6.6 Hz, 1H), 4.09 (dd, J = 7.0 Hz, 1H), 3.89 (dd, J = 9.7 Hz, 1H), 3.44 (dd, J = 10.0, 6.4 Hz, 1H), 2.35 (ddd, J = 14.5, 7.2 Hz, 1H), 2.14 (ddd, J = 14.0, 6.5 Hz, 1H); ^{13}C-NMR (126 MHz, D_2O + 0.1% MeOH) δ 171.5, 159.5, 52.2, 50.2, 48.0, 35.2; HRMS (ESI) Calculated for $C_6H_{13}N_4O_2$ 173.1033, found 173.1038 $(M+H)^+$.

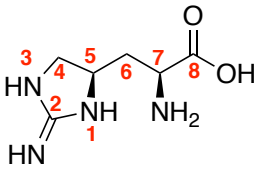
NMR and Compound Characterization



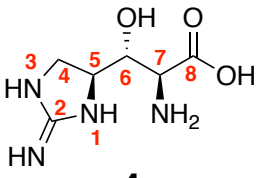
GNT (1)



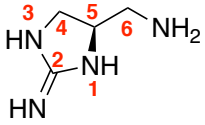
2



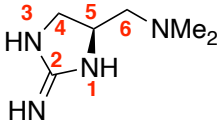
3



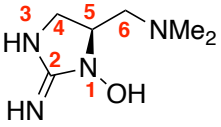
4



6



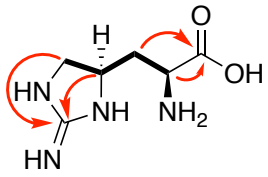
7



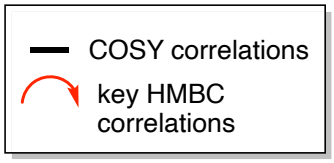
8

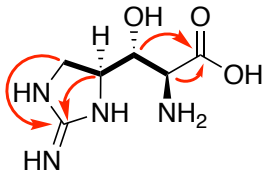
Numbering scheme for isolated, synthetic and enzymatic compounds **1 – 4**, **6 – 8** adapted from the original guanitoxin numbering scheme (*45*).

NMR correlations summary for enzymatic compound **4**



L-enduracididine
L-end, **3**
(synthetic)





β-hydroxy-L-enduracididine
β-OH-L-end, **4**
(enzymatic)

^{1}H NMR table for compounds 1 – 4, 6 – 8

^{1}H	1* (isolated)	8* (isolated)	7* (synthetic)	7 (synthetic)	6 (synthetic)	4 (enzymatic)	3 (synthetic)	2 (synthetic)
4a	4.01, dd (10.1, 9.4)	3.93, dd (10.0, 8.5)	3.98, dd (10.1, 9.7)	3.96, dd (10.2, 10.2)	3.94, dd (10.2, 10.2)	3.83, dd (10.0, 10.0)	3.88, dd (9.7, 9.7)	3.34, dd (14.7, 3.8)
4b	3.51, dd (10.1, 9.7)	3.46, dd (10.0, 8.0)	3.50, dd (10.1, 5.4)	3.48, m	3.54, dd (10.5, 6.0)	3.65, dd (10.3, 5.5)	3.42, dd (10.1, 6.0)	3.23, dd (14.6, 7.3)
5	4.71, m	4.48, dddd	4.55, dddd	4.53, dddd (9.6, 9.6, 5.3, 5.3)	4.42, dddd (10.9, 5.7, 5.7, 5.7)	4.34, ddd (9.7, 6.8, 5.5)	4.34, m	3.99, dddd (8.2, 8.2, 4.1, 4.0)
6a	3.75, dd (13.9, 9.3)	3.75, dd (13.9, 6.8)	3.48, dd (13.4, 8.1)	3.45, m	3.25, m	4.06, dd (6.7, 3.1)	2.26, ddd (14.8, 8.8, 5.8)	2.09, m
6b	3.47, dd (13.9, 2.9)	3.47, dd (13.9, 4.5)	3.38, dd (13.4, 4.8)	3.36, dd (13.5, 4.8)	3.25, m	-	2.18, ddd (13.9, 6.7, 6.7)	2.09, m
7	-	-	-	-	-	3.90, d (3.1)	4.08, dd (9.2, 5.6)	4.19, dd (5.9, 5.9)
NMe_2	3.00, s	3.00, s	2.96, s	2.95, s	-	-	-	-
OMe	3.79, d (11.0)	-	-	-	-	-	-	-

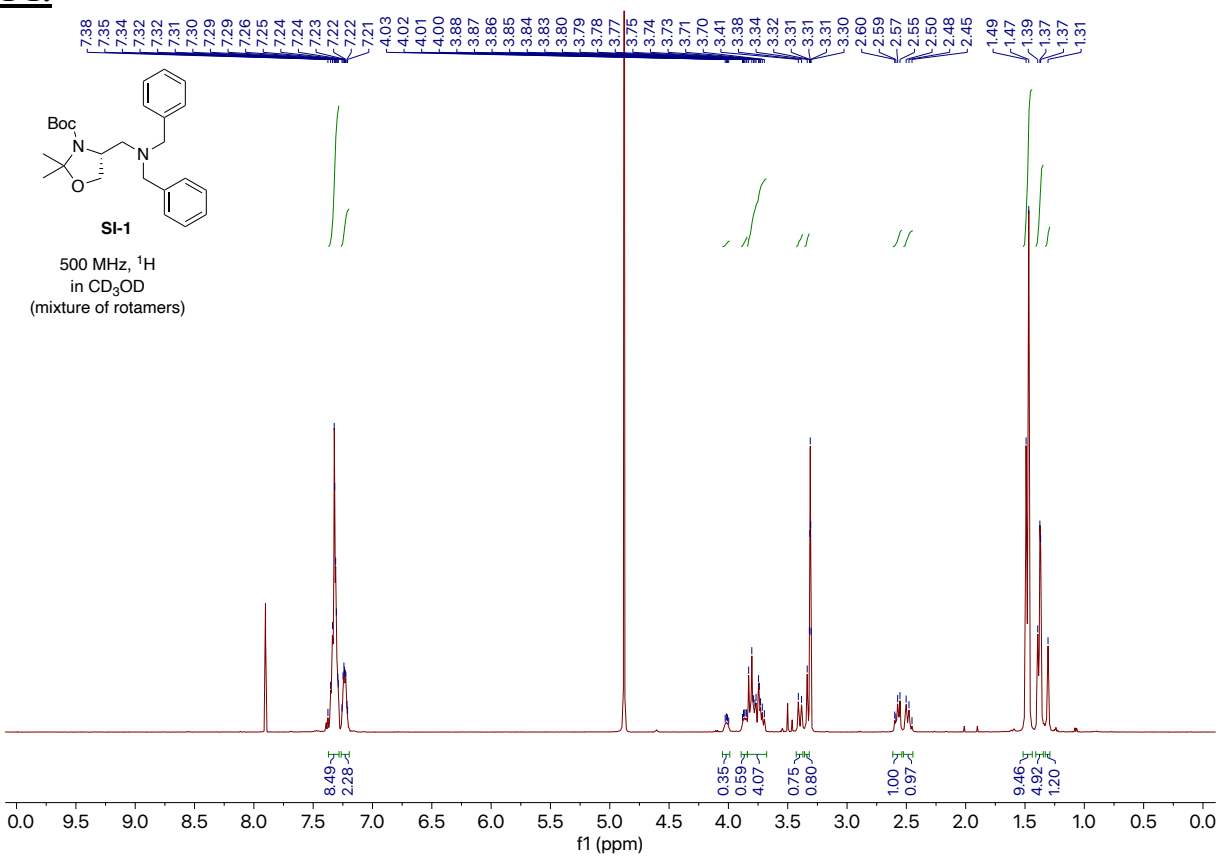
* from (*45*); spectra ran in D_2O + 1% acetic acid-d_4
all other spectra were collected in D_2O + 0.1 % CH_3OH and referenced to methanol (δ 3.34)

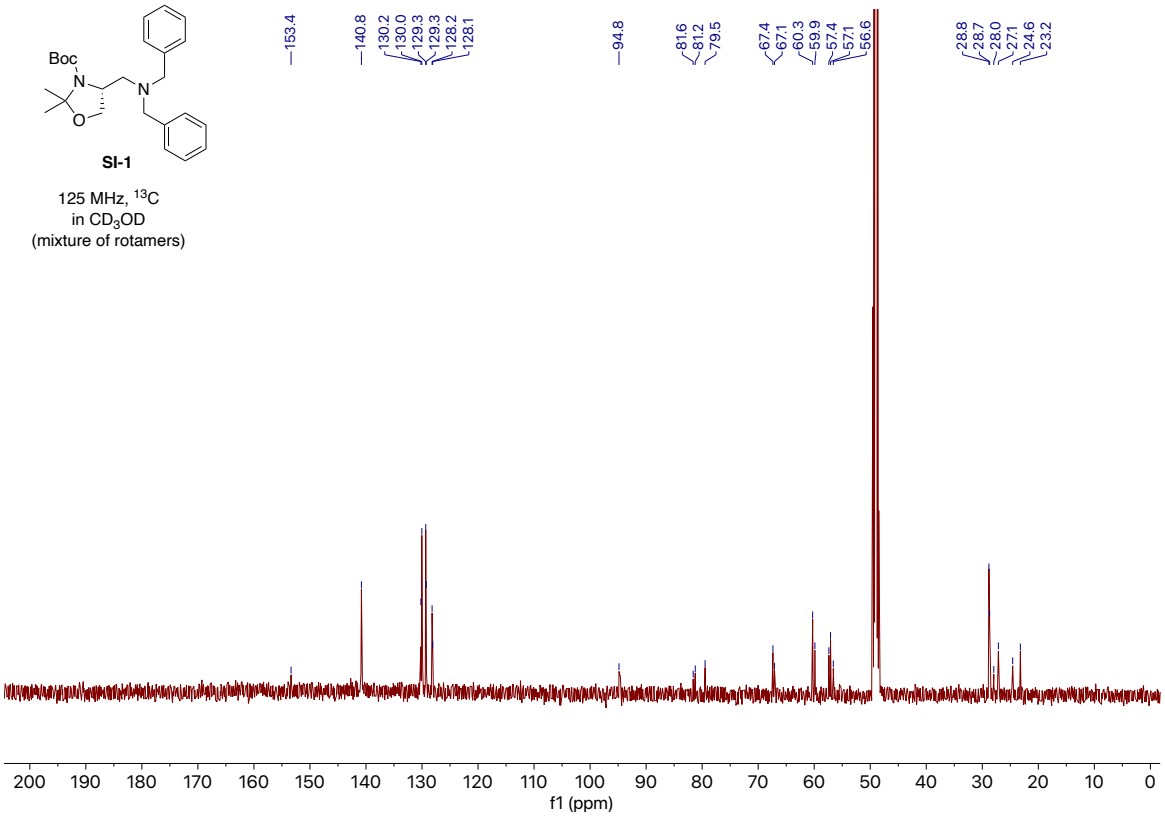
^{13}C NMR table for compounds 1 – 4, 6 – 8

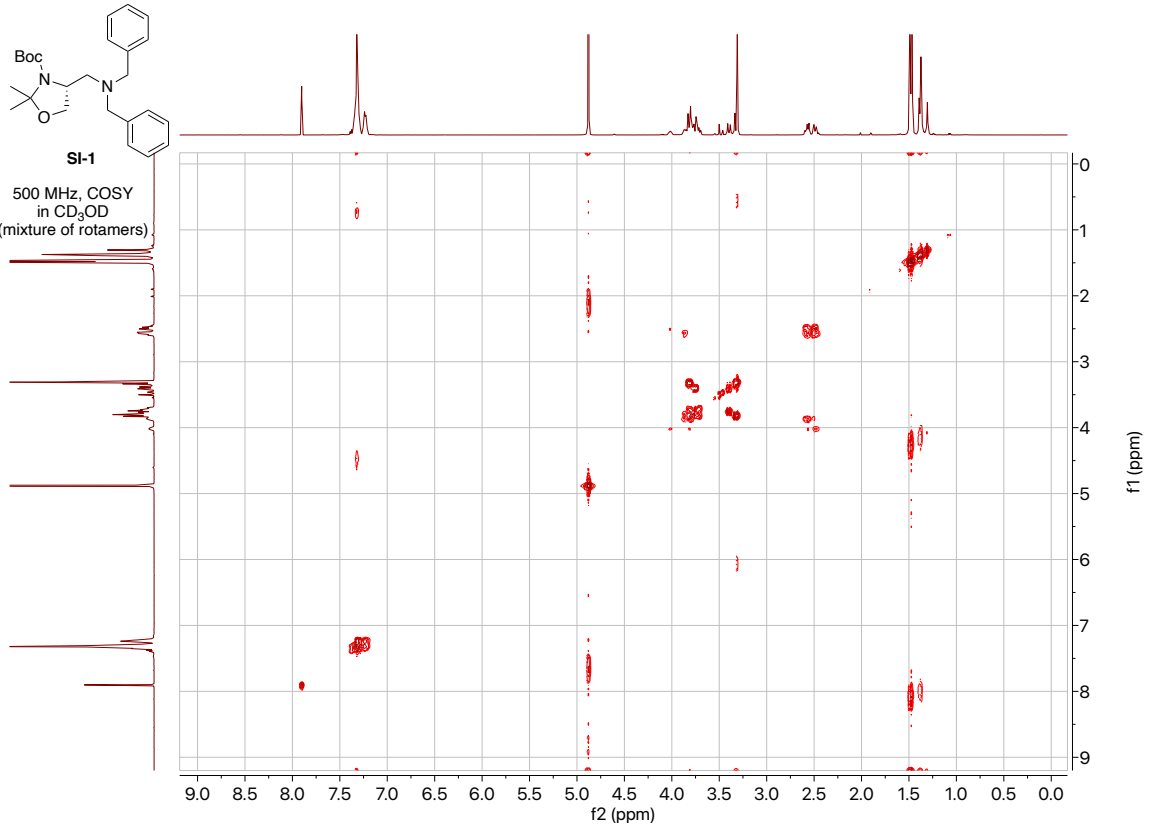
^{13}C	1* (isolated)	8* (isolated)	7* (synthetic)	7 (synthetic)	6 (synthetic)	4 (enzymatic)	3 (synthetic)	2 (synthetic)
2	163.7	162.5	160.6	159.9	159.8	159.4	159.5	157.4
4	56.1	44.9	47.6	46.9	46.0	44.9	48.3	46.9
5	60.3	58.8	51.0	50.3	52.3	56.5	52.1	66.6
6	58.7	58.7	61.1	60.4	42.1	71.5	35.5	33.0
7	-	-	-	-	-	55.9	50.0	51.2
8	-	-	-	-	-	171.4	171.3	172.2
NMe_2	43.9	44.7	44.3	42.8	-	-	-	-
OMe	56.1	-	-	-	-	-	-	-

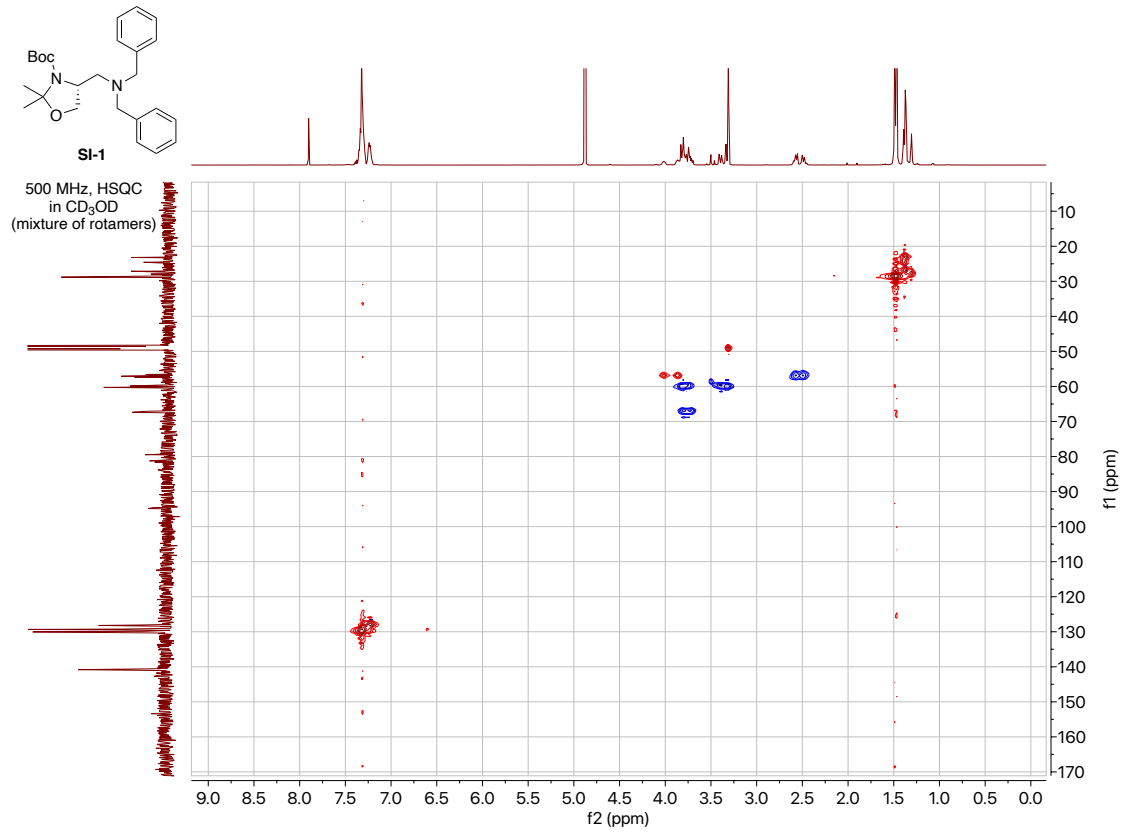
* from (*45*); spectra ran in D_2O + 1% acetic acid-d_4
all other spectra were collected in D_2O + 0.1 % CH_3OH and referenced to methanol (δ 49.0)

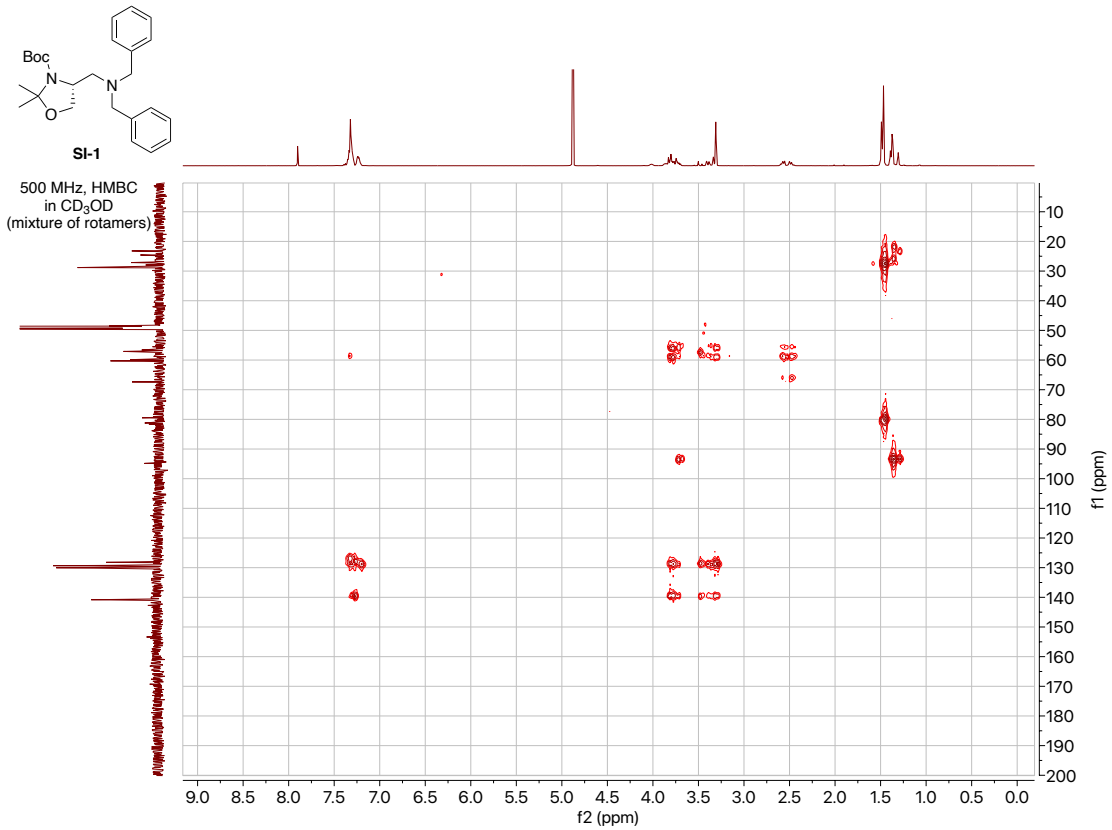
SI-1:


Boc
N
O
N
SI-1
500 MHz, ^{1}H
in CD_3OD
(mixture of rotamers)
f1 (ppm)

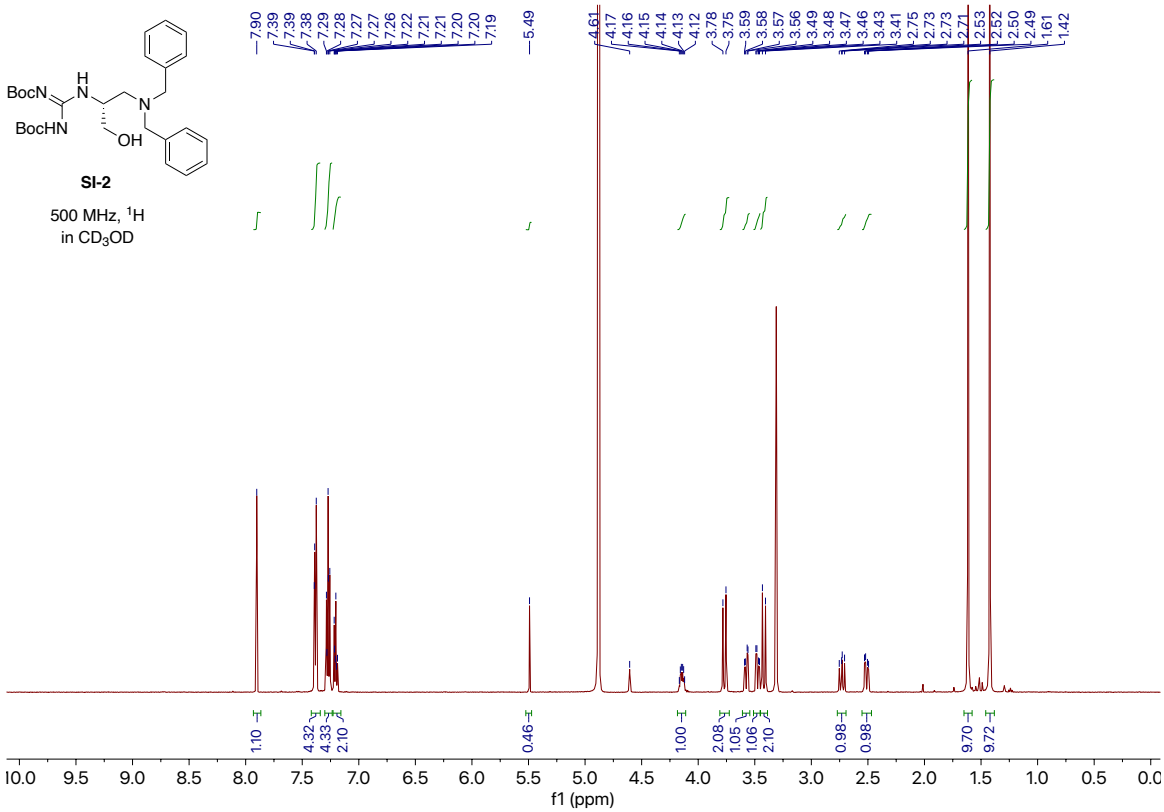

Boc
N
O
N
SI-1
125 MHz, ^{13}C
in CD_3OD
(mixture of rotamers)
f1 (ppm)

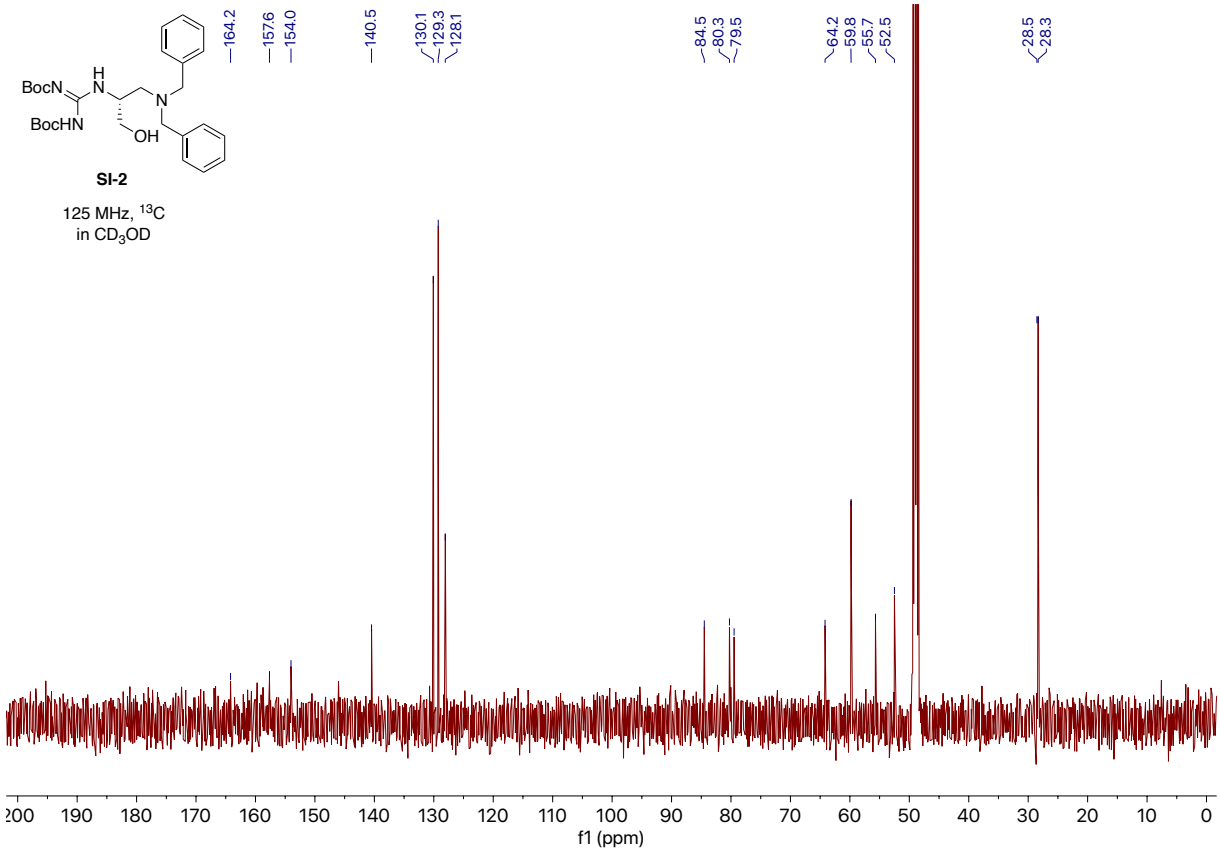

Boc
N
N
O
SI-1
500 MHz, COSY
in CD3OD
(mixture of rotamers)
9.0 8.5 8.0 7.5 7.0 6.5 6.0 5.5 5.0 4.5 4.0 3.5 3.0 2.5 2.0 1.5 1.0 0.5 0.0
f2 (ppm)
0 1 2 3 4 5 6 7 8 9
f1 (ppm)

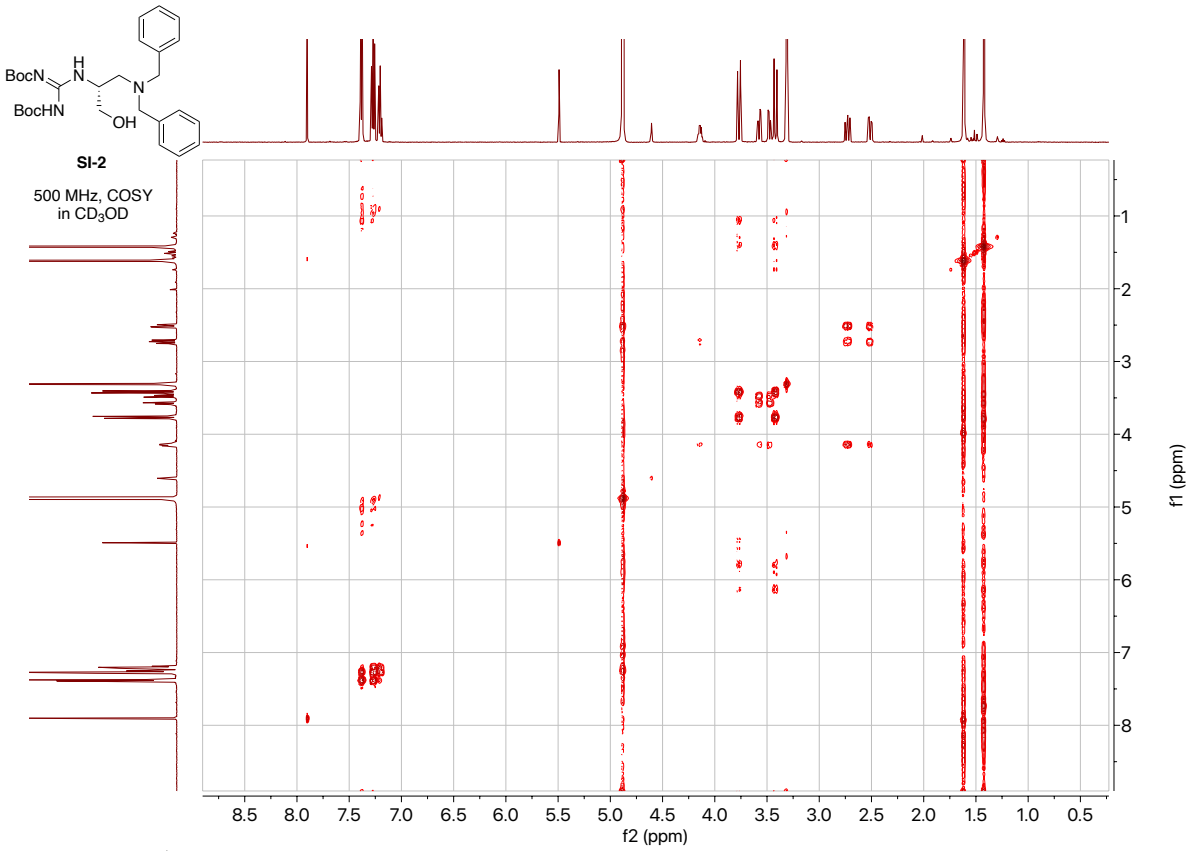

Boc
N
N
O
SI-1
500 MHz, HSQC
in CD3OD
(mixture of rotamers)
9.0 8.5 8.0 7.5 7.0 6.5 6.0 5.5 5.0 4.5 4.0 3.5 3.0 2.5 2.0 1.5 1.0 0.5 0.0
f2 (ppm)
10 20 30 40 50 60 70 80 90 100 110 120 130 140 150 160 170
f1 (ppm)

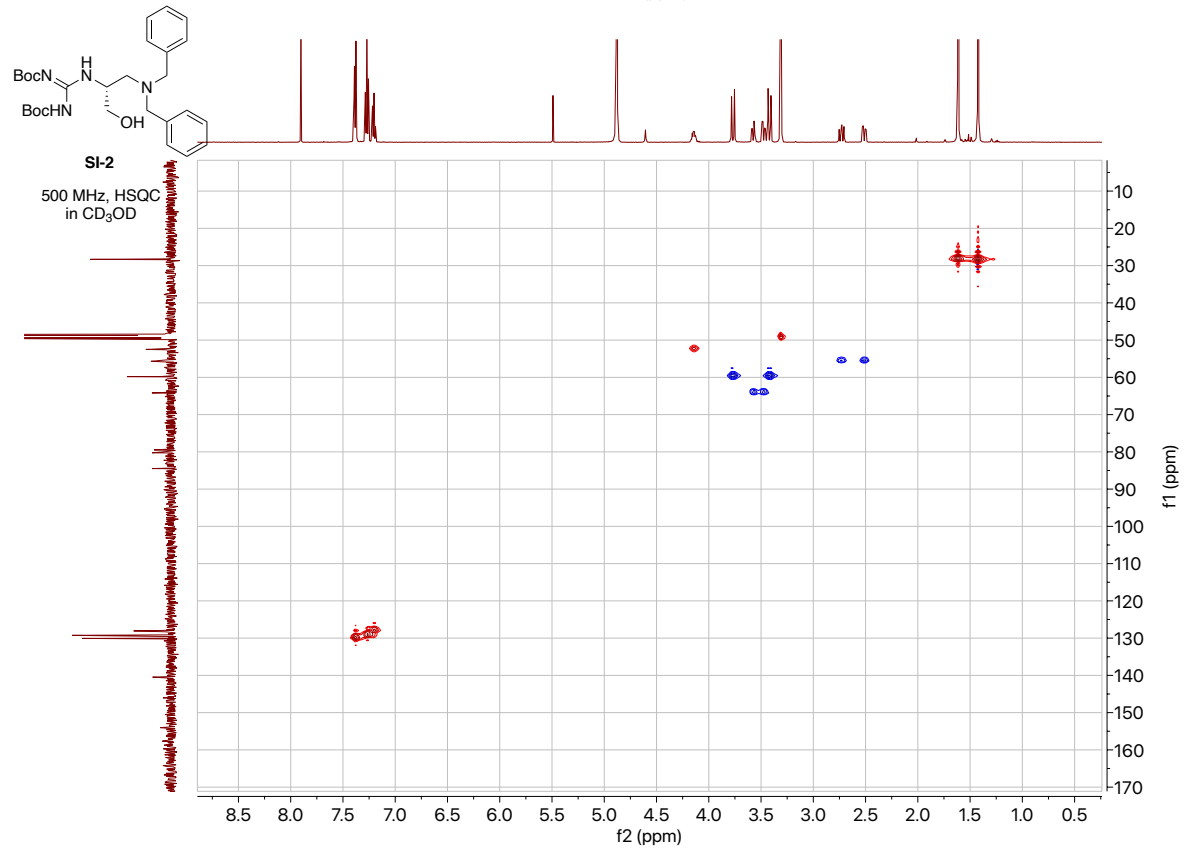

Boc
N
N
O
SI-1
500 MHz, HMBC
in CD3OD
(mixture of rotamers)
f2 (ppm)
f1 (ppm)

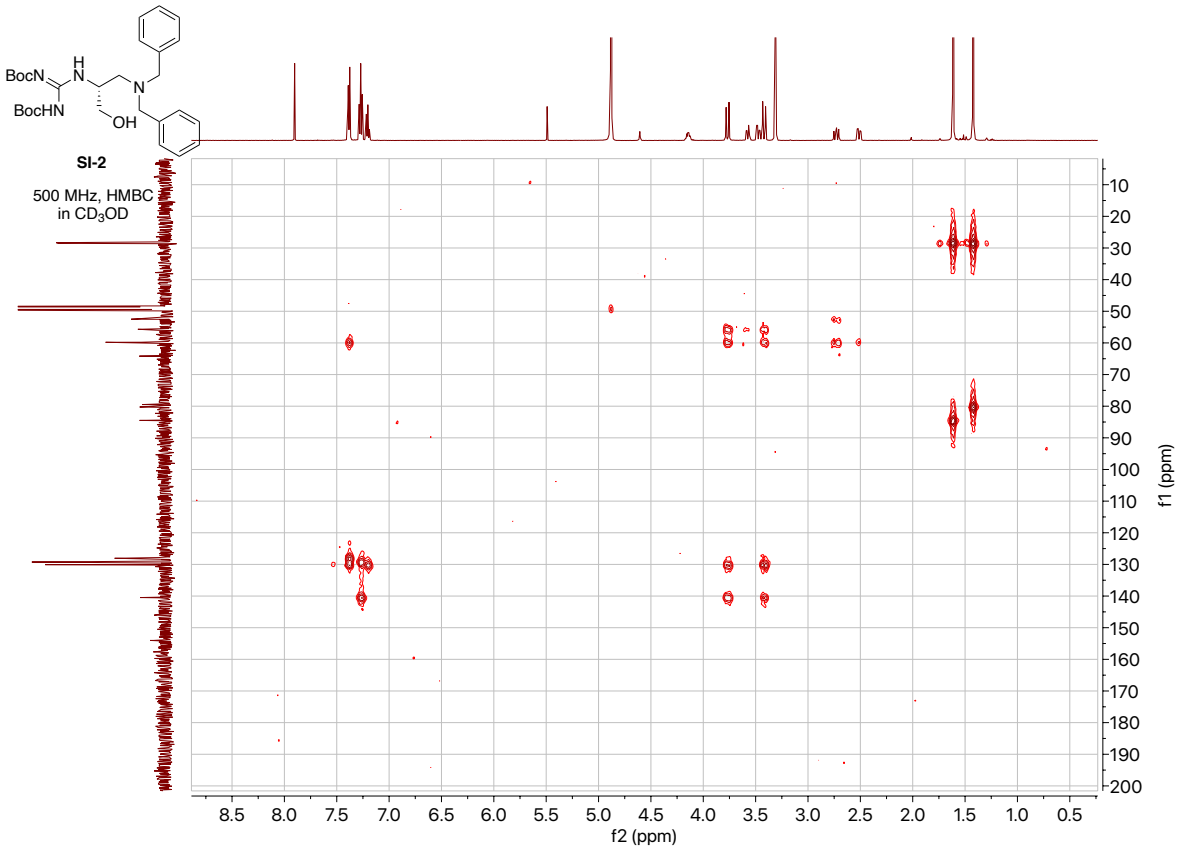
SI-2:


BocN
H
N
N
BocHN
OH
SI-2
500 MHz, ^{1}H
in CD_3OD
7.90
7.39
7.39
7.38
7.29
7.28
7.27
7.27
7.26
7.22
7.21
7.21
7.20
7.20
7.19
5.49
4.61
4.17
4.16
4.15
4.14
4.13
4.12
3.78
3.75
3.59
3.58
3.57
3.56
3.49
3.48
3.47
3.46
3.43
3.41
2.75
2.73
2.73
2.71
2.53
2.52
2.50
2.49
1.61
1.42
1.10
4.32
4.33
2.10
0.46
1.00
2.08
1.05
1.06
2.10
0.98
0.98
9.70
9.72
f1 (ppm)

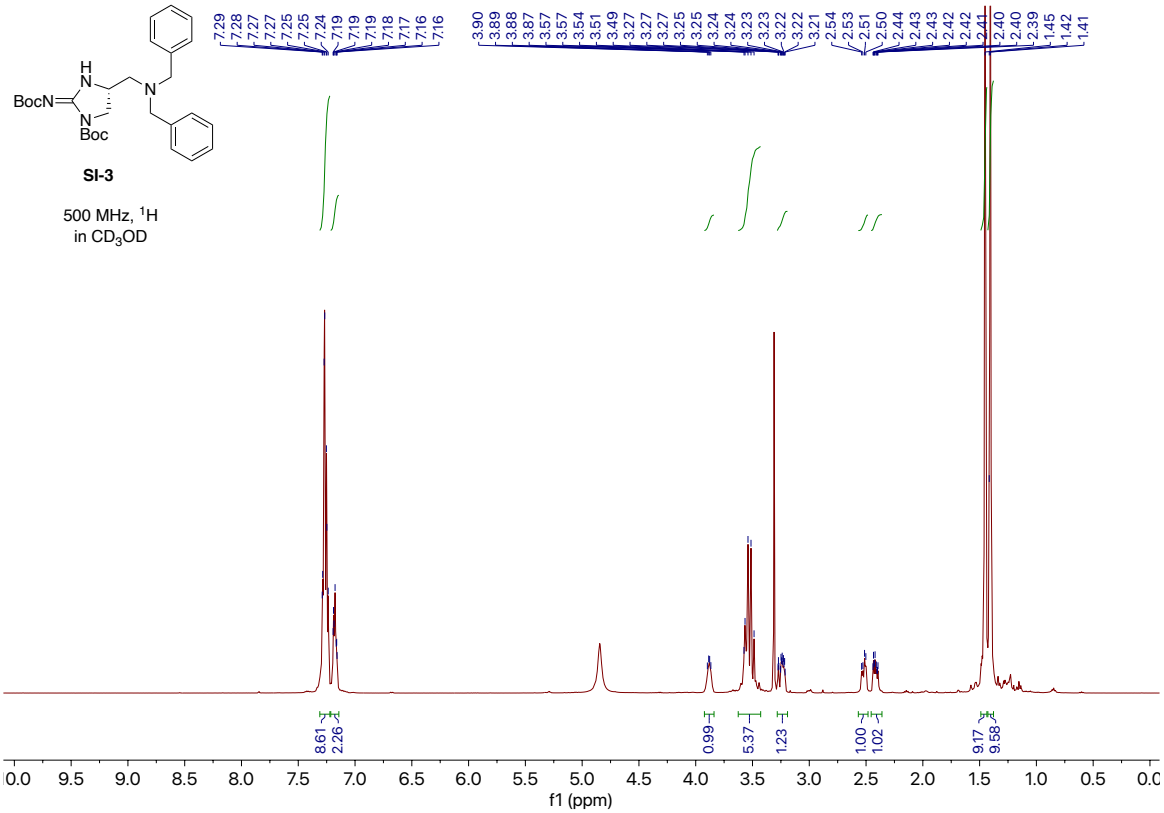

BocN
H
N
N
BocHN
OH
SI-2
125 MHz, ^{13}C
in CD_3OD
164.2
157.6
154.0
140.5
130.1
129.3
128.1
84.5
80.3
79.5
64.2
59.8
55.7
52.5
28.5
28.3
f1 (ppm)

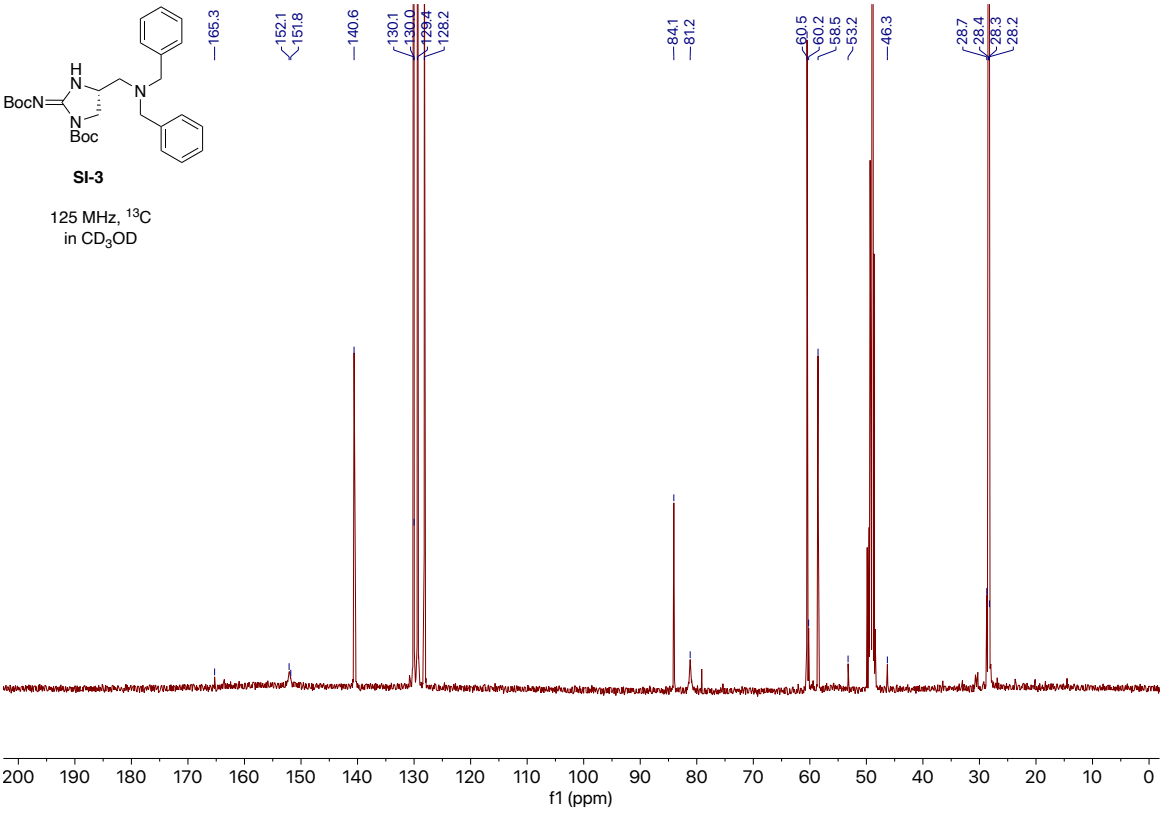

BocN
H
N
BocHN
N
OH
SI-2
500 MHz, COSY
in CD3OD
f2 (ppm)
f1 (ppm)

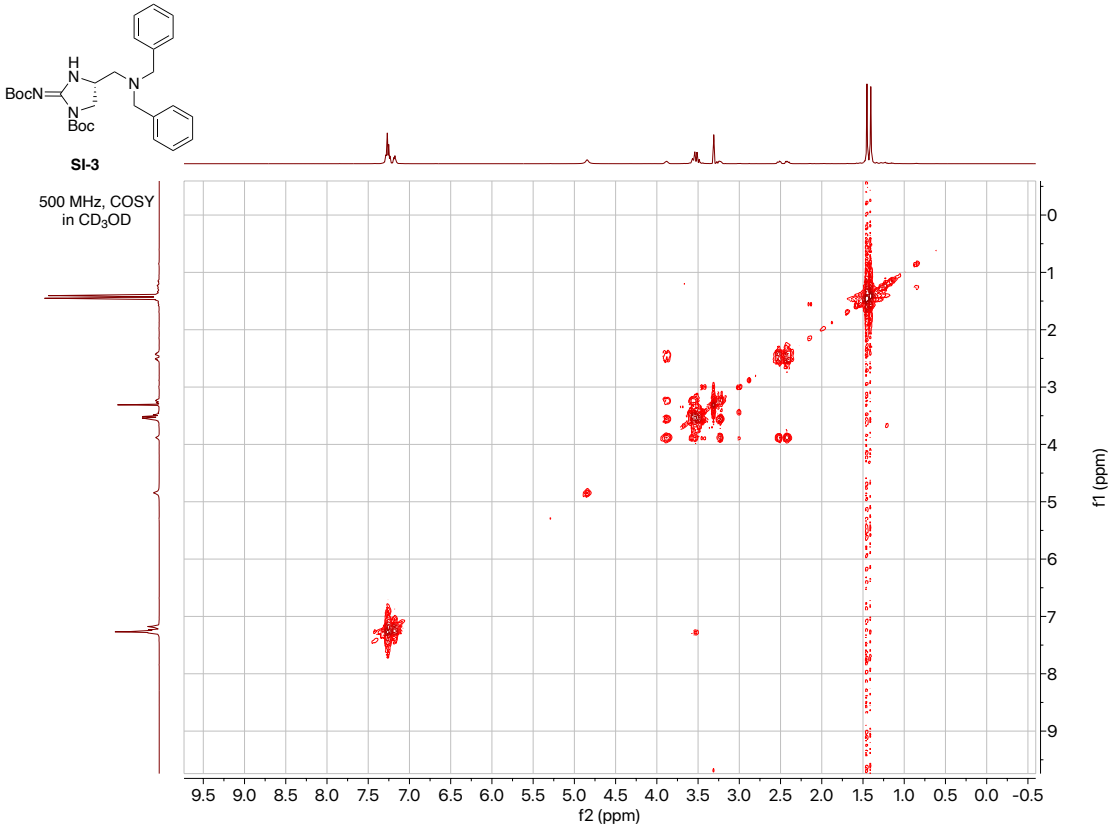

BocN
H
N
BocHN
N
OH
SI-2
500 MHz, HSQC
in CD3OD
f2 (ppm)
f1 (ppm)

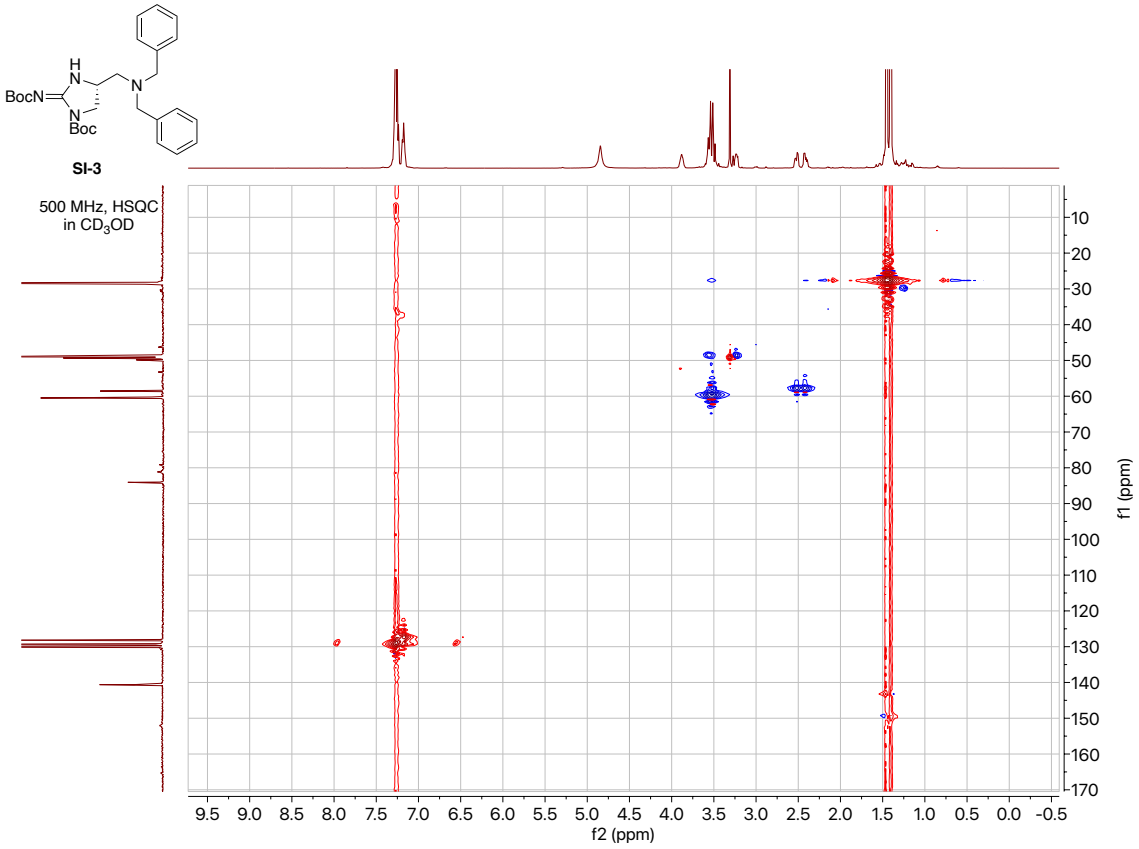

BocN
H
N
BocHN
N
OH
SI-2
500 MHz, HMBC
in CD3OD
8.5
8.0
7.5
7.0
6.5
6.0
5.5
5.0
4.5
4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.5
f2 (ppm)
10
20
30
40
50
60
70
80
90
100
110
120
130
140
150
160
170
180
190
200
f1 (ppm)

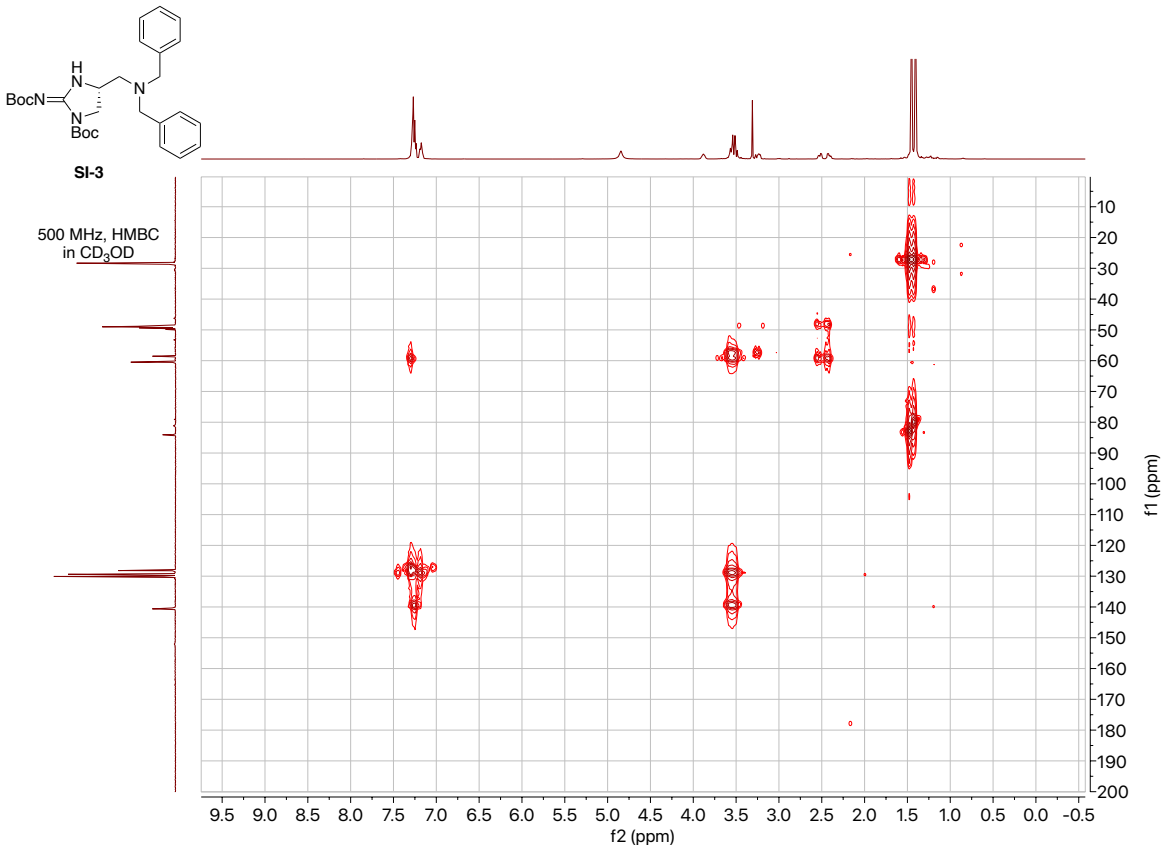
SI-3


BocN
H
N
N
Boc
N
SI-3
500 MHz, ^{1}H
in CD_3OD
7.29 7.28 7.27 7.27 7.25 7.25 7.24 7.19 7.19 7.19 7.18 7.17 7.16 7.16
3.90 3.89 3.88 3.87 3.57 3.57 3.54 3.51 3.49 3.27 3.27 3.27 3.25 3.25 3.24 3.24 3.23 3.23 3.22 3.22 3.21
2.54 2.53 2.51 2.50 2.44 2.43 2.43 2.42 2.42 2.41 2.40 2.40 2.39 1.45 1.42 1.41
8.61 2.26 0.99 5.37 1.23 1.00 1.02 9.17 9.58
f1 (ppm)

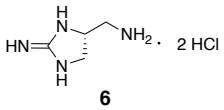

BocN
H
N
N
Boc
N
SI-3
125 MHz, ^{13}C
in CD_3OD
165.3 152.1 151.8 140.6 130.1 130.0 129.4 128.2
84.1 81.2
60.5 60.2 58.5 53.2 46.3
28.7 28.4 28.3 28.2
f1 (ppm)


BocN
H
N
N
Boc
N
SI-3
500 MHz, COSY
in CD3OD
f2 (ppm)
f1 (ppm)


BocN
H
N
N
Boc
N
SI-3
500 MHz, HSQC
in CD3OD
f2 (ppm)
f1 (ppm)

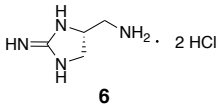

BocN
H
N
N
Boc
N
SI-3
500 MHz, HMBC
in CD3OD
f2 (ppm)
f1 (ppm)

6:



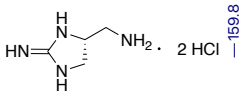
6

500 MHz, ^{1}H
in D_2O + 0.1% MeOH



6

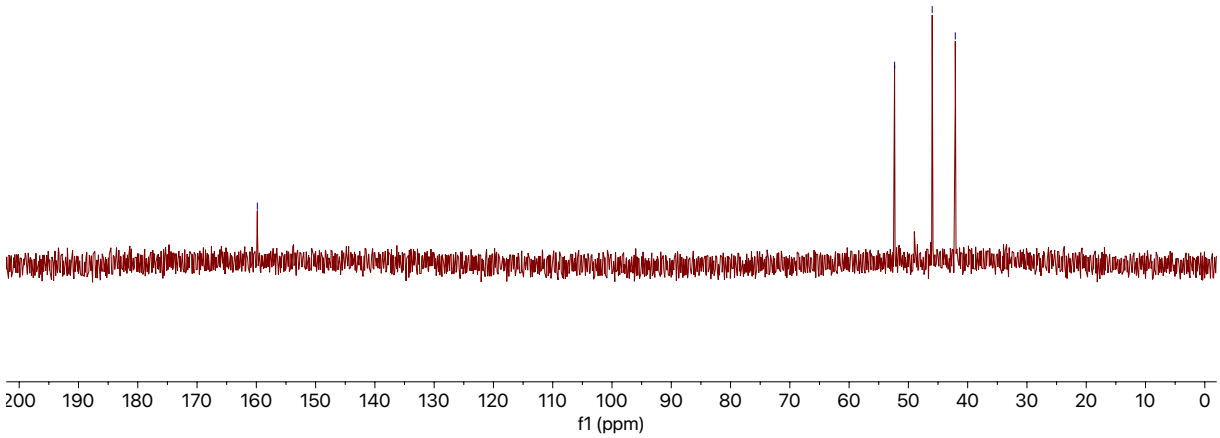
500 MHz, ^{1}H
in D_2O + 0.1% MeOH

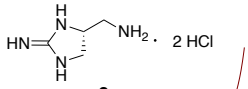


6

125 MHz, ^{13}C
in D_2O + 0.1% MeOH

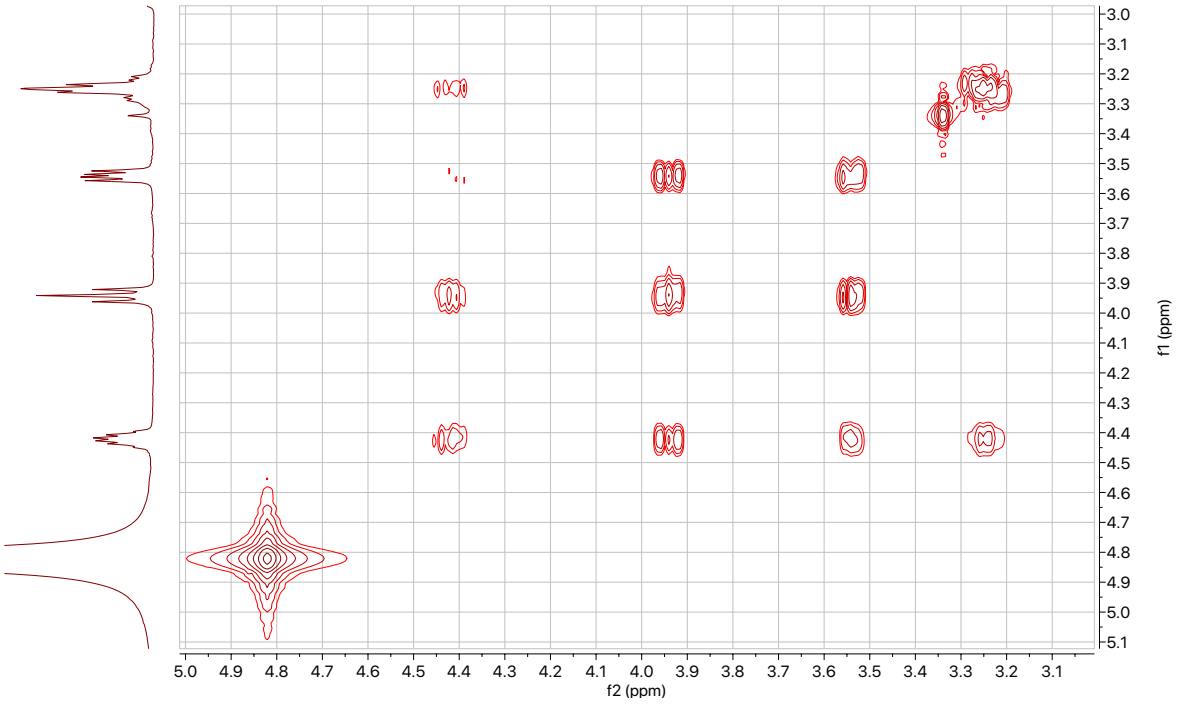
52.3
46.0
42.1

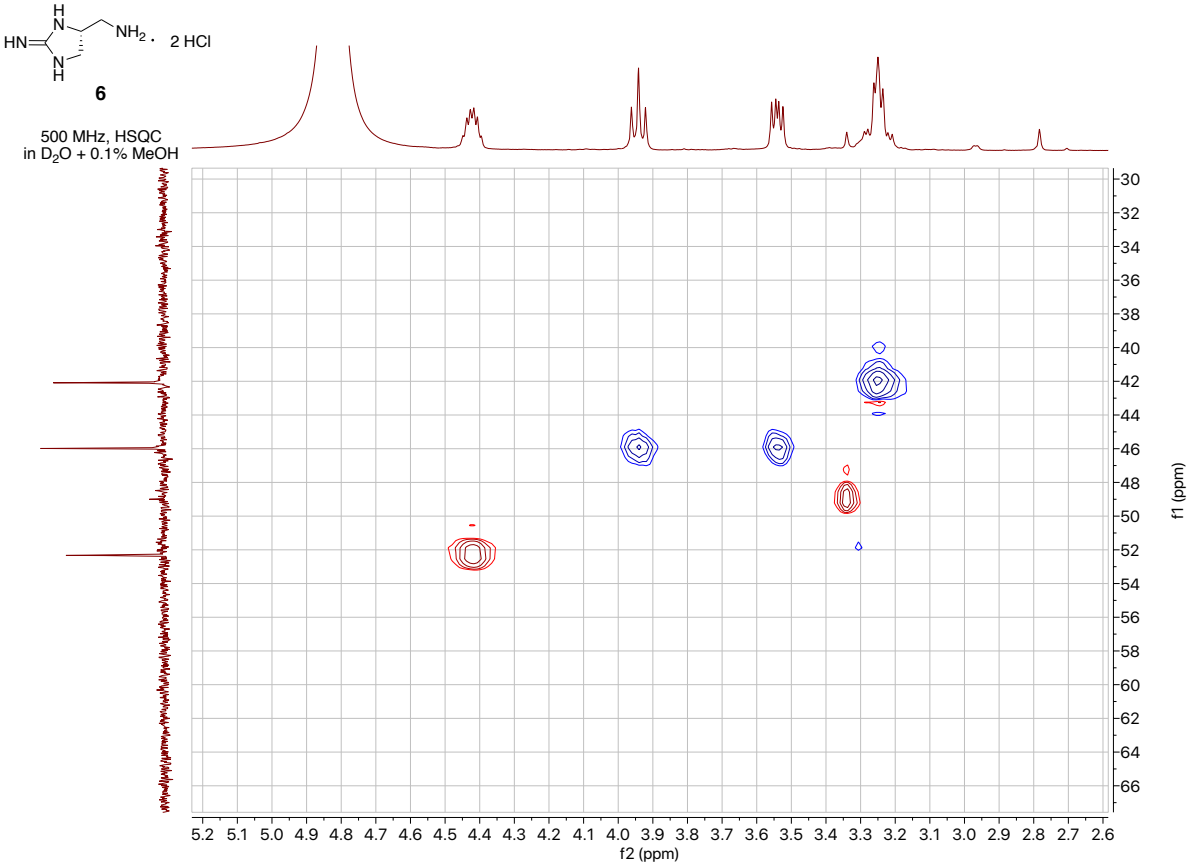


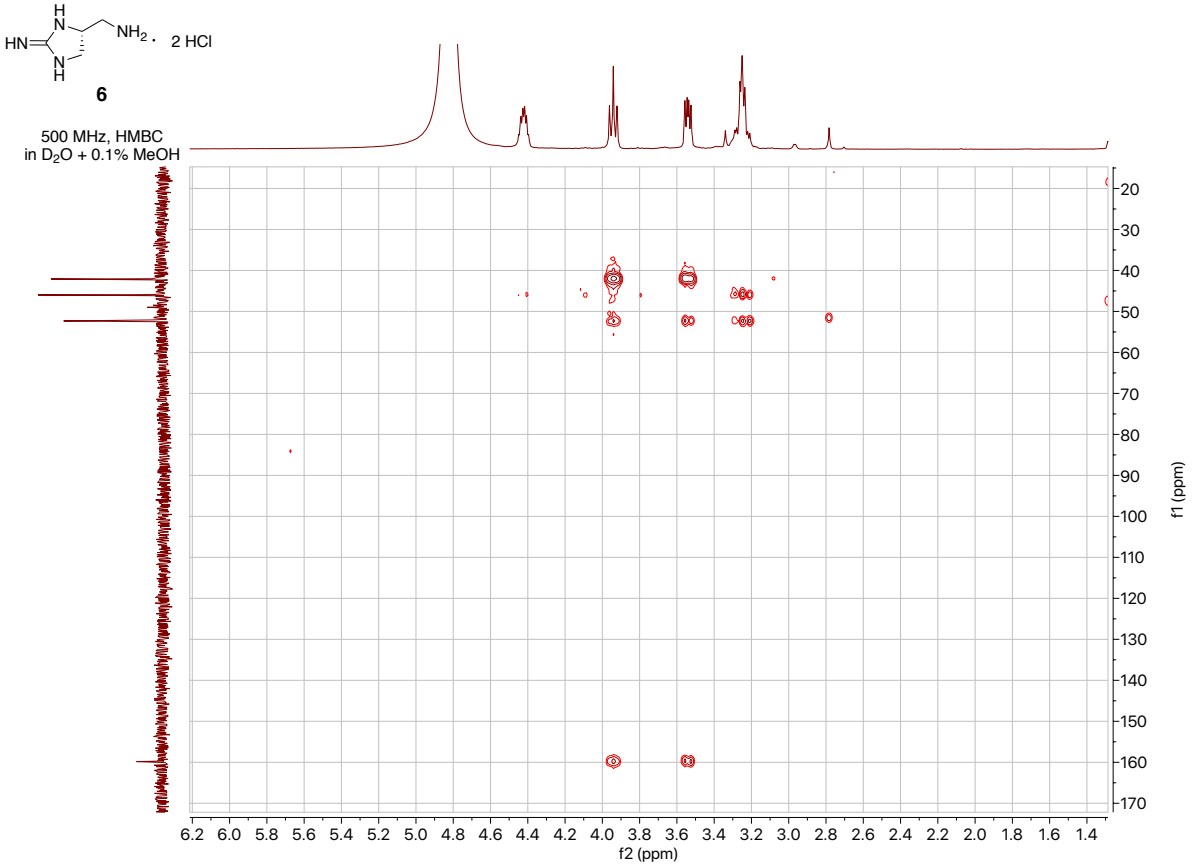


6

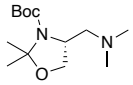
500 MHz, COSY
in D_2O + 0.1% MeOH




6
500 MHz, HSQC
in D2O + 0.1% MeOH
2 HCl
f2 (ppm)
f1 (ppm)

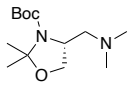

6
500 MHz, HMBC
in D2O + 0.1% MeOH
2 HCl
f2 (ppm)
f1 (ppm)

SI-4:



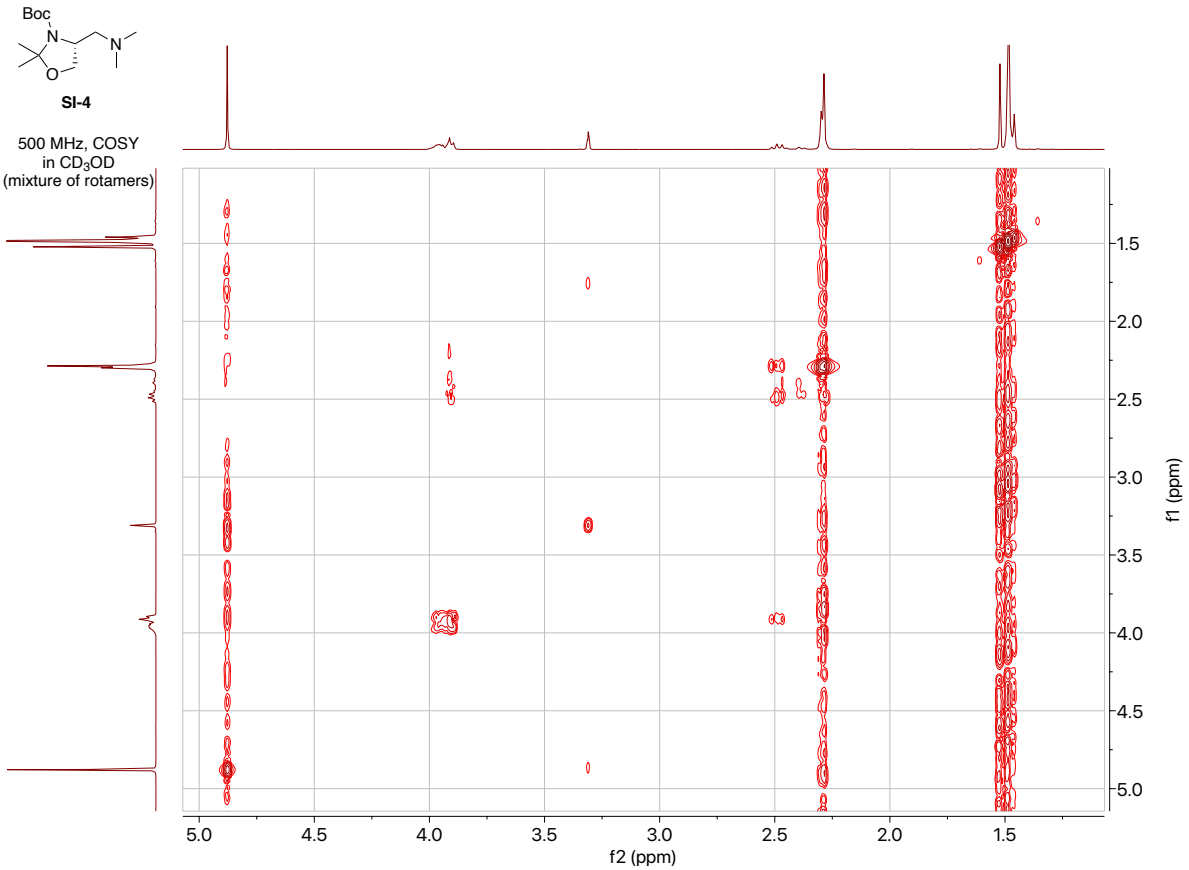
SI-4

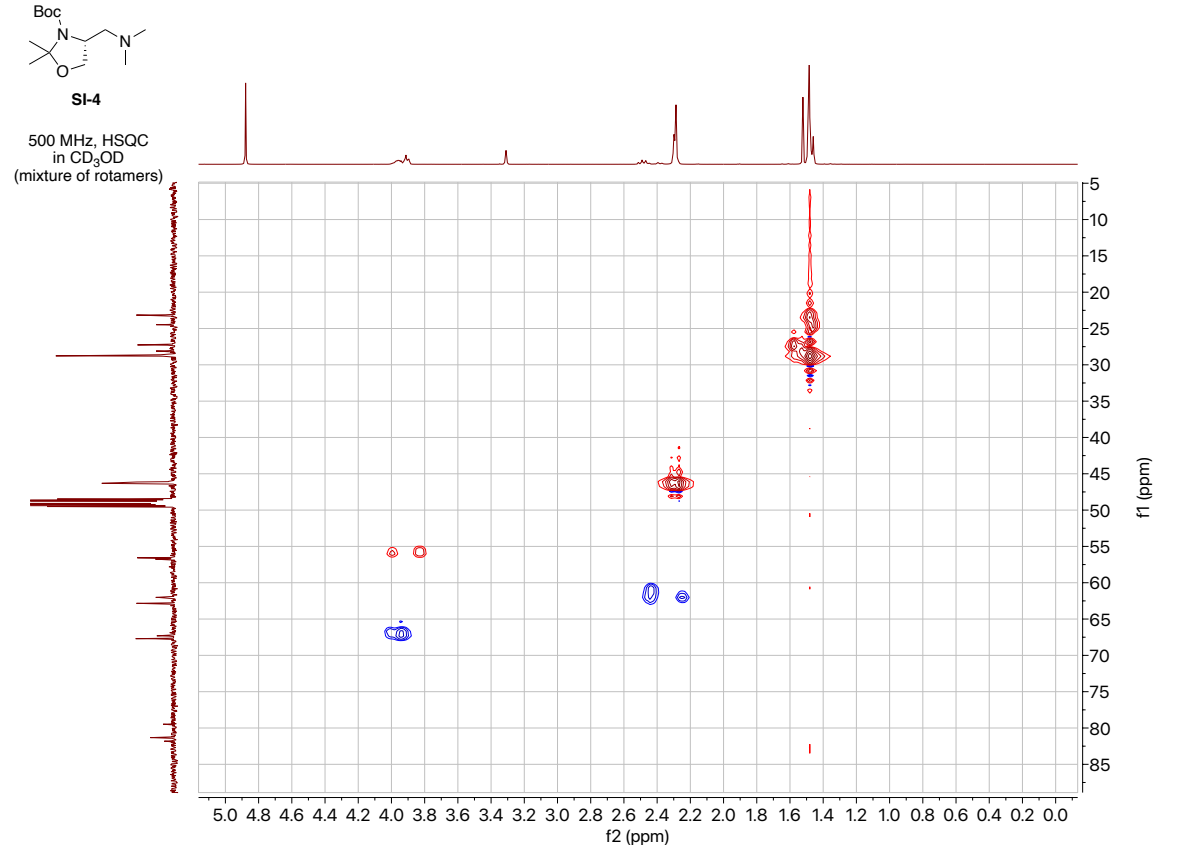
500 MHz, ^{1}H
in CD_3OD
(mixture of rotamers)

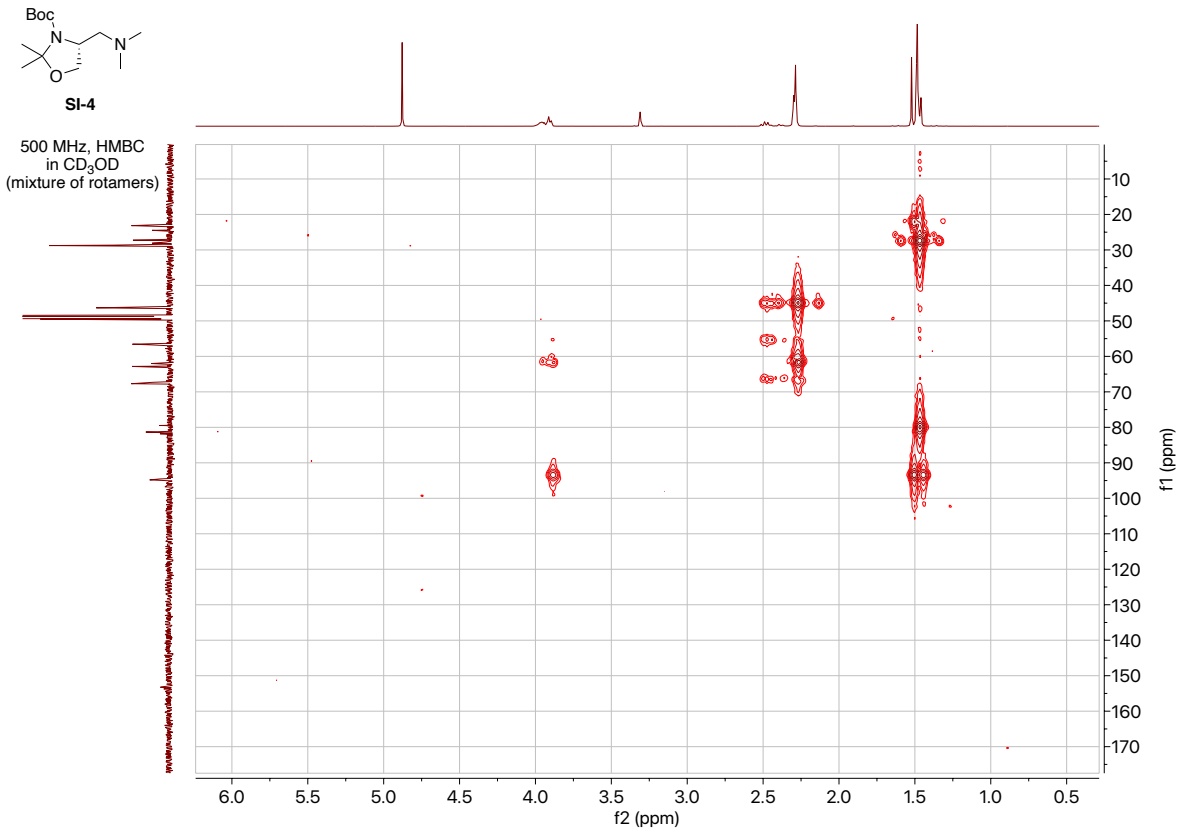


SI-4

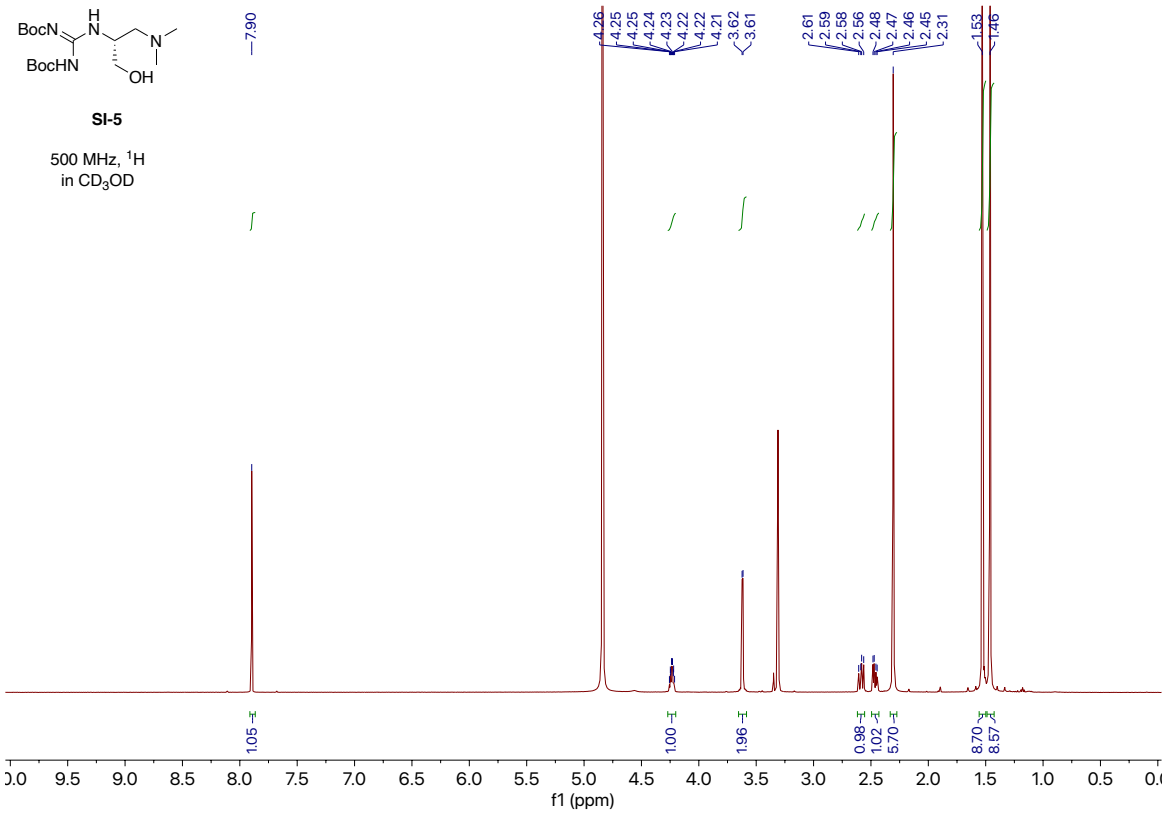
125 MHz, ^{13}C
in CD_3OD
(mixture of rotamers)

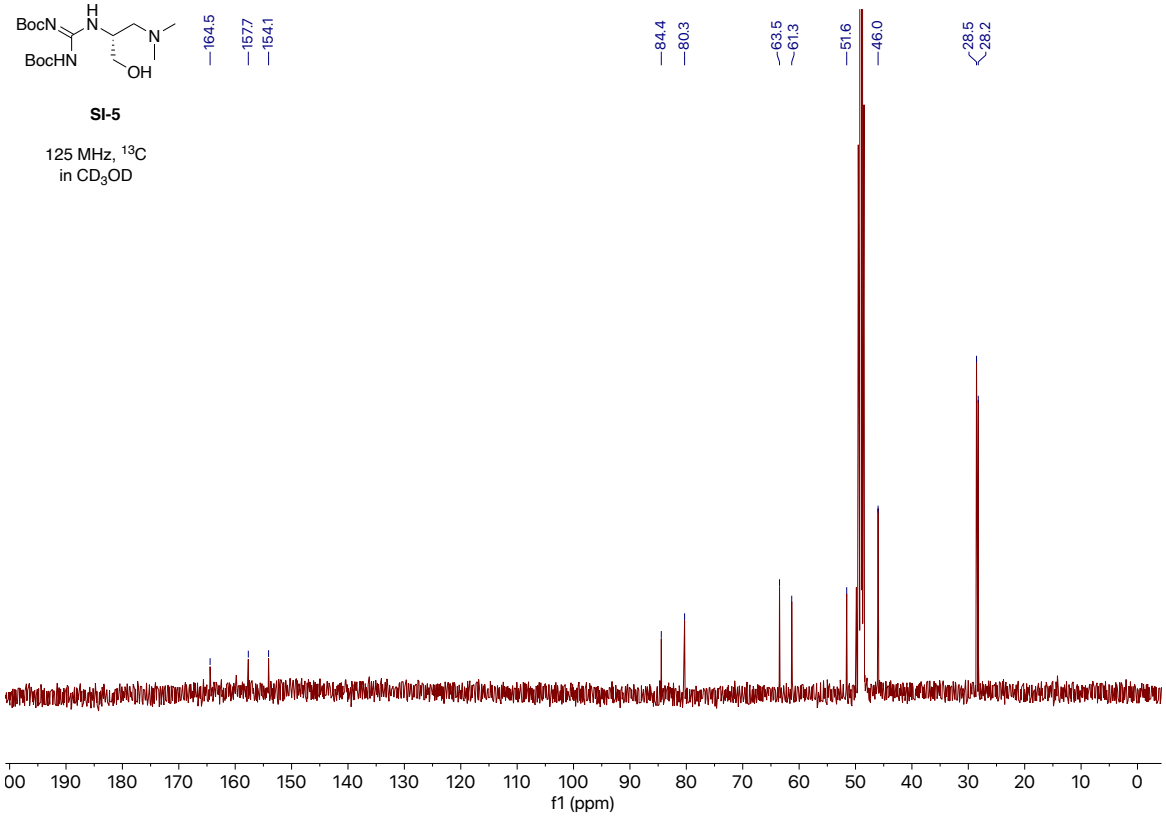

Boc
N
N
O
SI-4
500 MHz, COSY
in CD3OD
(mixture of rotamers)
5.0
4.5
4.0
3.5
3.0
2.5
2.0
1.5
f2 (ppm)
f1 (ppm)


Boc
N
N
O
SI-4
500 MHz, HSQC
in CD3OD
(mixture of rotamers)
5.0 4.8 4.6 4.4 4.2 4.0 3.8 3.6 3.4 3.2 3.0 2.8 2.6 2.4 2.2 2.0 1.8 1.6 1.4 1.2 1.0 0.8 0.6 0.4 0.2 0.0
f2 (ppm)
f1 (ppm)

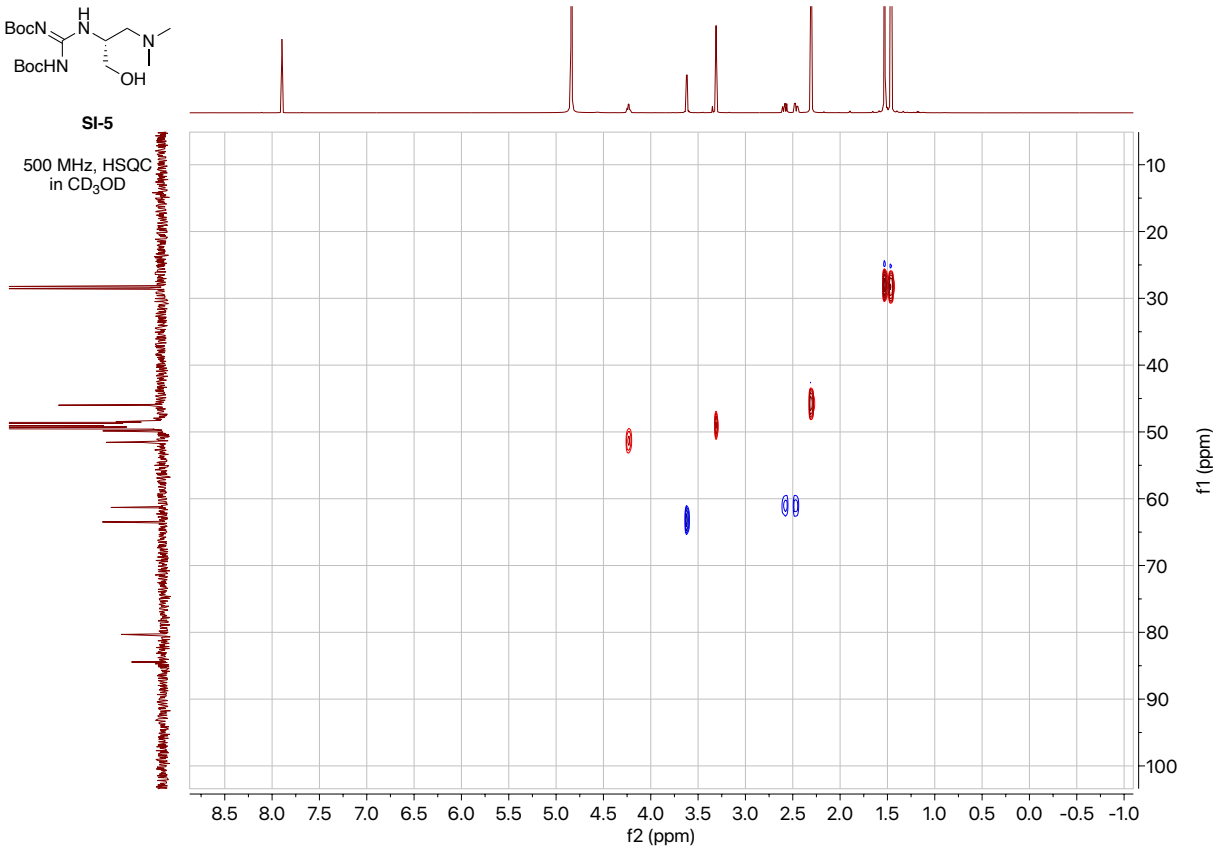

Boc
N
N
O
SI-4
500 MHz, HMBC
in CD3OD
(mixture of rotamers)
6.0
5.5
5.0
4.5
4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.5
f2 (ppm)
10
20
30
40
50
60
70
80
90
100
110
120
130
140
150
160
170
f1 (ppm)

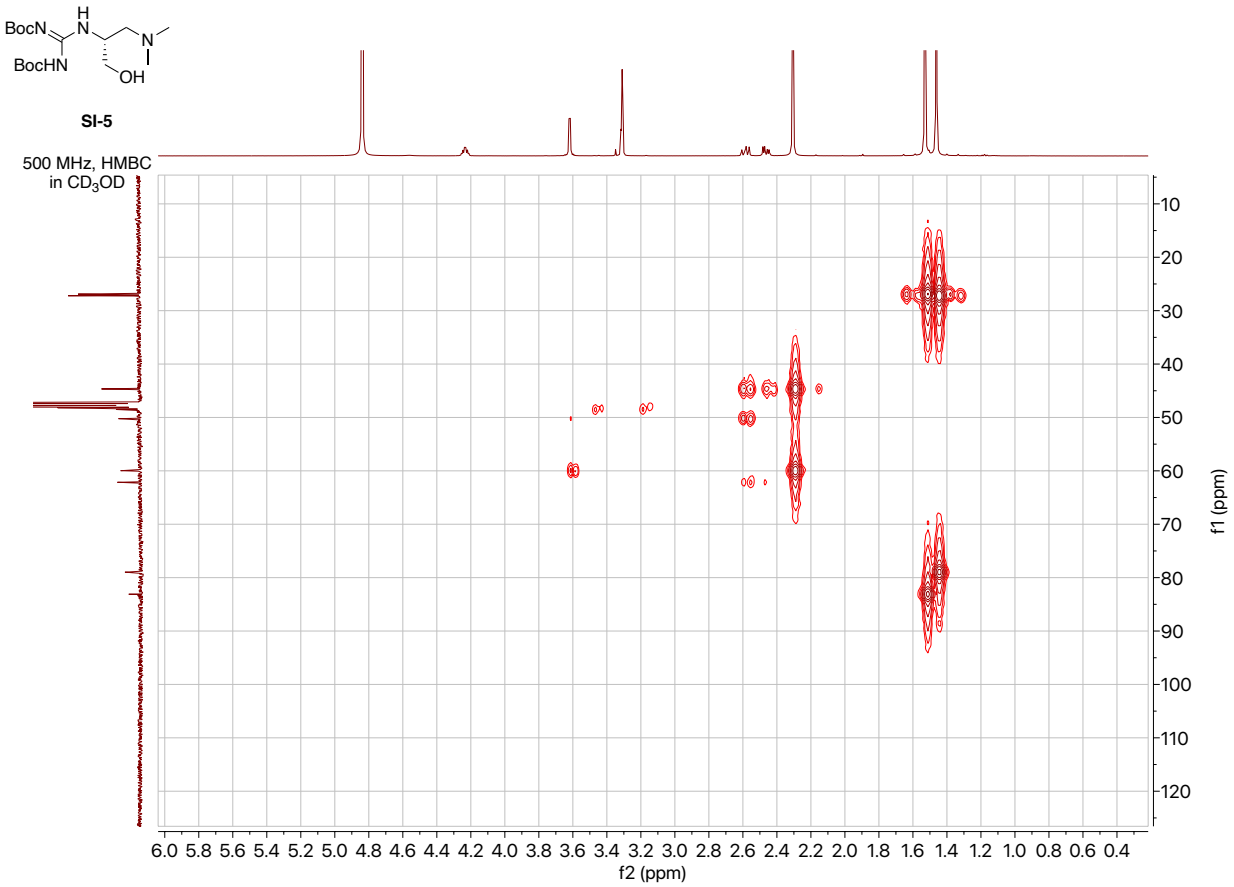
SI-5:


BocN
H
N
N
BocHN
OH
SI-5
500 MHz, 1H
in CD3OD
7.90
4.26
4.25
4.25
4.24
4.23
4.22
4.22
4.21
3.62
3.61
2.61
2.59
2.58
2.56
2.48
2.47
2.46
2.45
2.31
1.53
1.46
1.05
1.00
1.96
0.98
1.02
5.70
8.70
8.57
f1 (ppm)

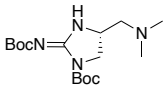

BocN
H
N
N
BocHN
OH
SI-5
125 MHz, 13C
in CD3OD
164.5
157.7
154.1
84.4
80.3
63.5
61.3
51.6
46.0
28.5
28.2
f1 (ppm)


BocN
BocHN
H
N
N
OH
SI-5
500 MHz, COSY
in CD3OD
f1 (ppm)
f2 (ppm)


BocN
BocHN
H
N
N
OH
SI-5
500 MHz, HSQC
in CD3OD
f1 (ppm)
f2 (ppm)

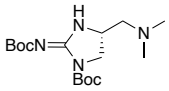

BocN
H
N
N
BocHN
OH
SI-5
500 MHz, HMBC
in CD3OD
f1 (ppm)
f2 (ppm)

SI-6:



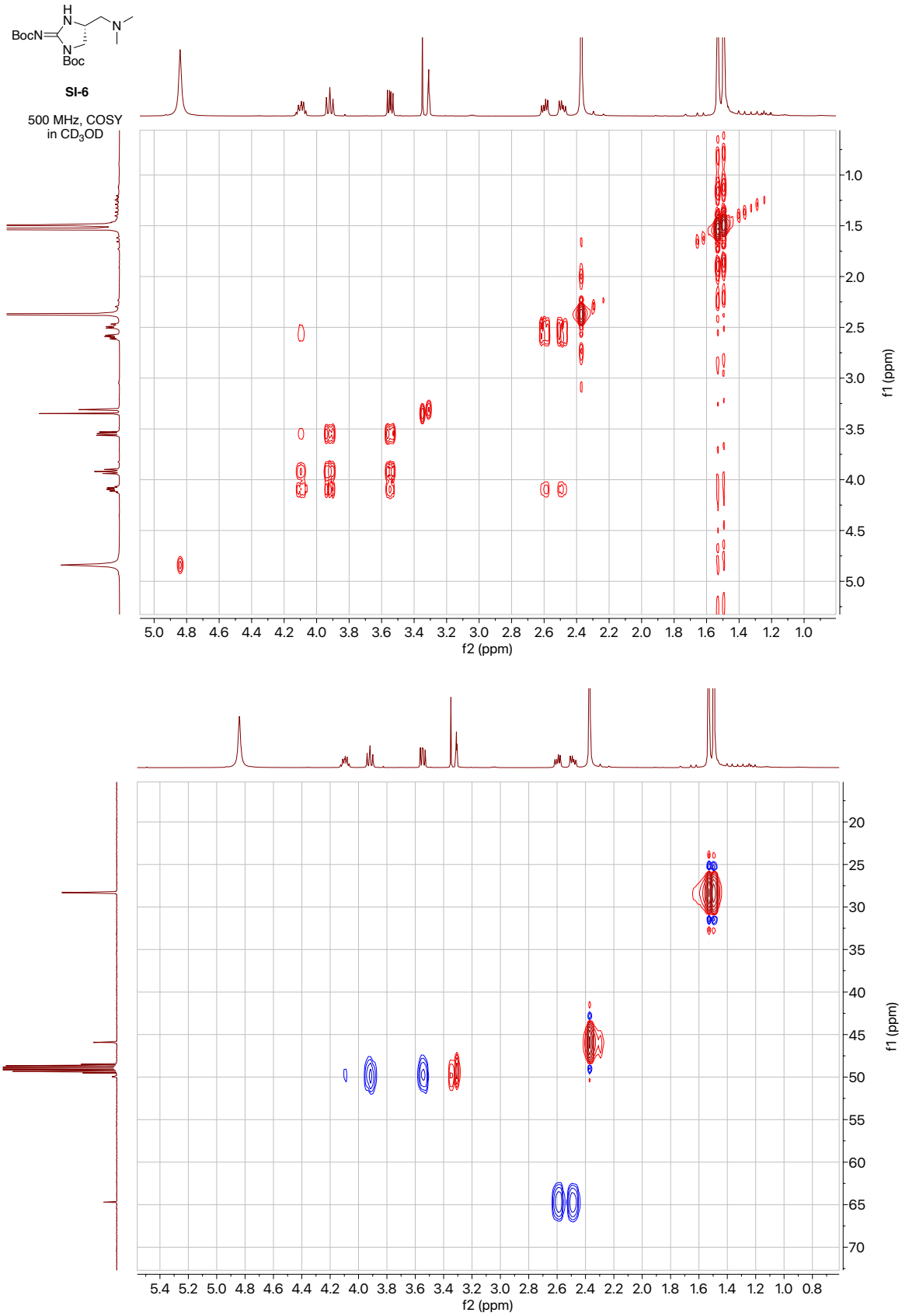
SI-6

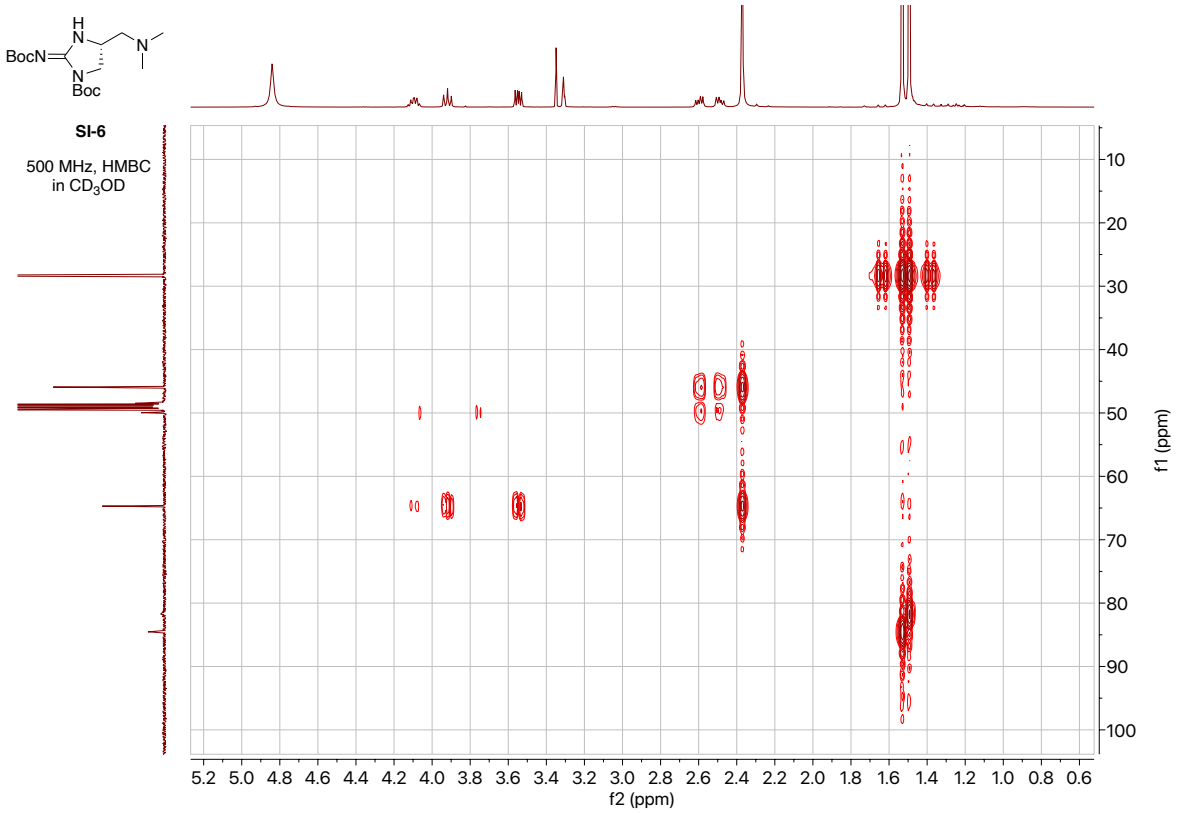
500 MHz, ^{1}H
in CD_3OD



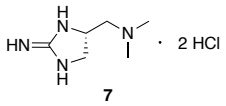
SI-6

125 MHz, ^{13}C
in CD_3OD


BocN
H
N
N
Boc
N
SI-6
500 MHz, COSY
in CD3OD
f2 (ppm)
f1 (ppm)
f2 (ppm)
f1 (ppm)

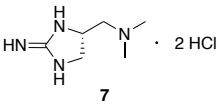

BocN
H
N
N
Boc
N
SI-6
500 MHz, HMBC
in CD3OD
5.2 5.0 4.8 4.6 4.4 4.2 4.0 3.8 3.6 3.4 3.2 3.0 2.8 2.6 2.4 2.2 2.0 1.8 1.6 1.4 1.2 1.0 0.8 0.6
f2 (ppm)
10 20 30 40 50 60 70 80 90 100
f1 (ppm)

7:



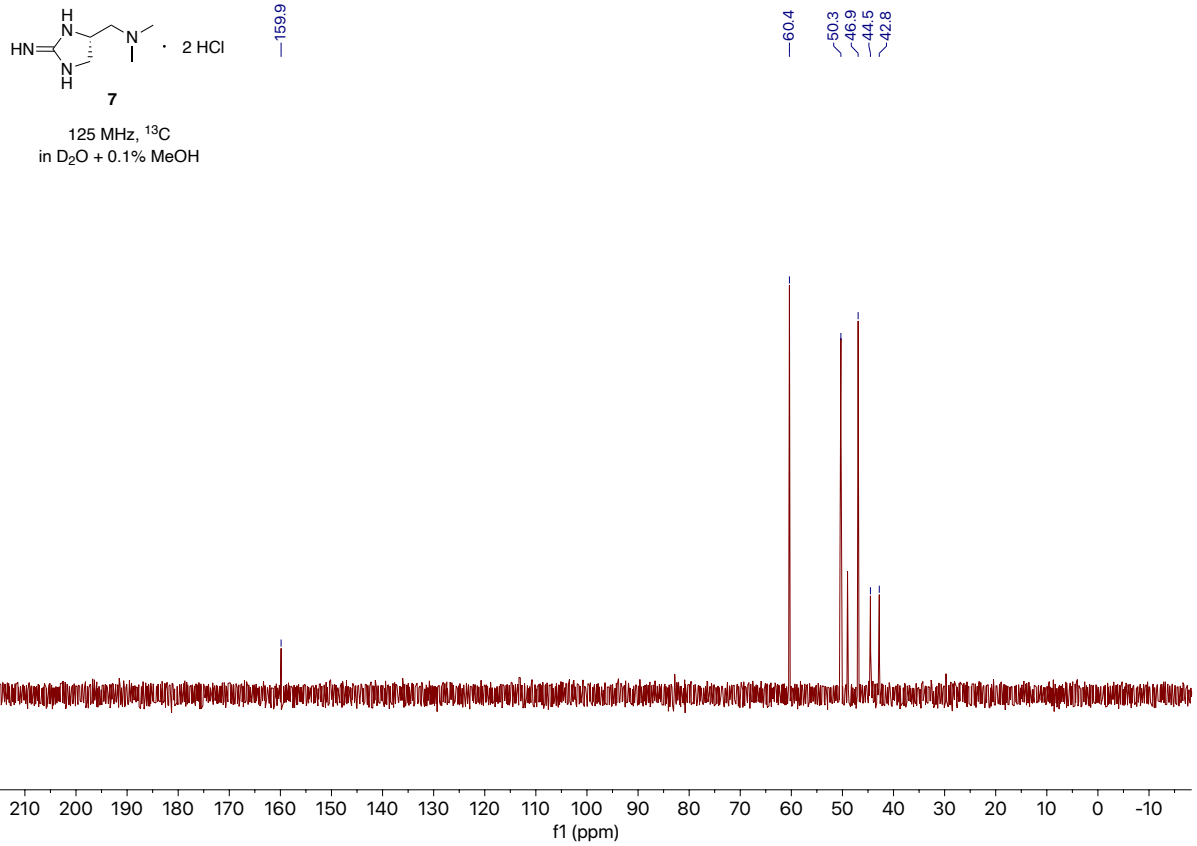
7

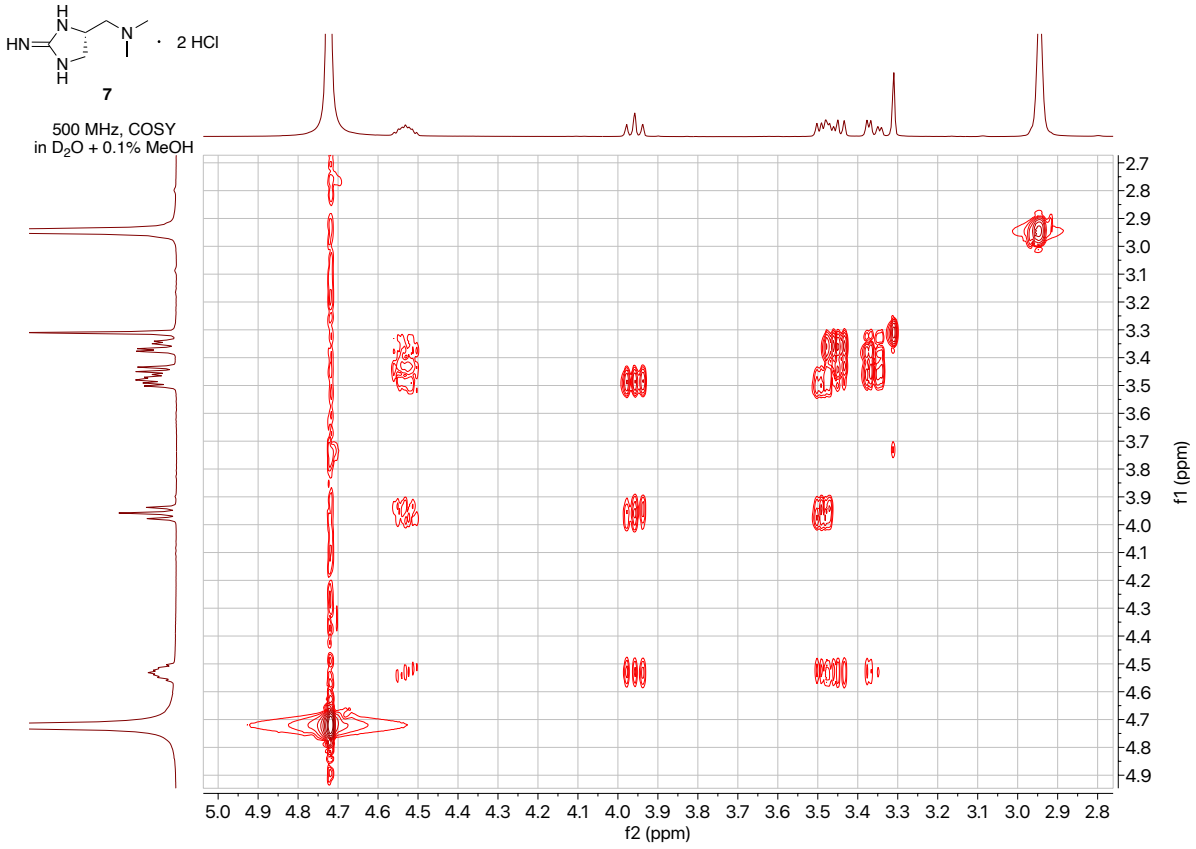
500 MHz, ^{1}H
in D_2O + 0.1% MeOH

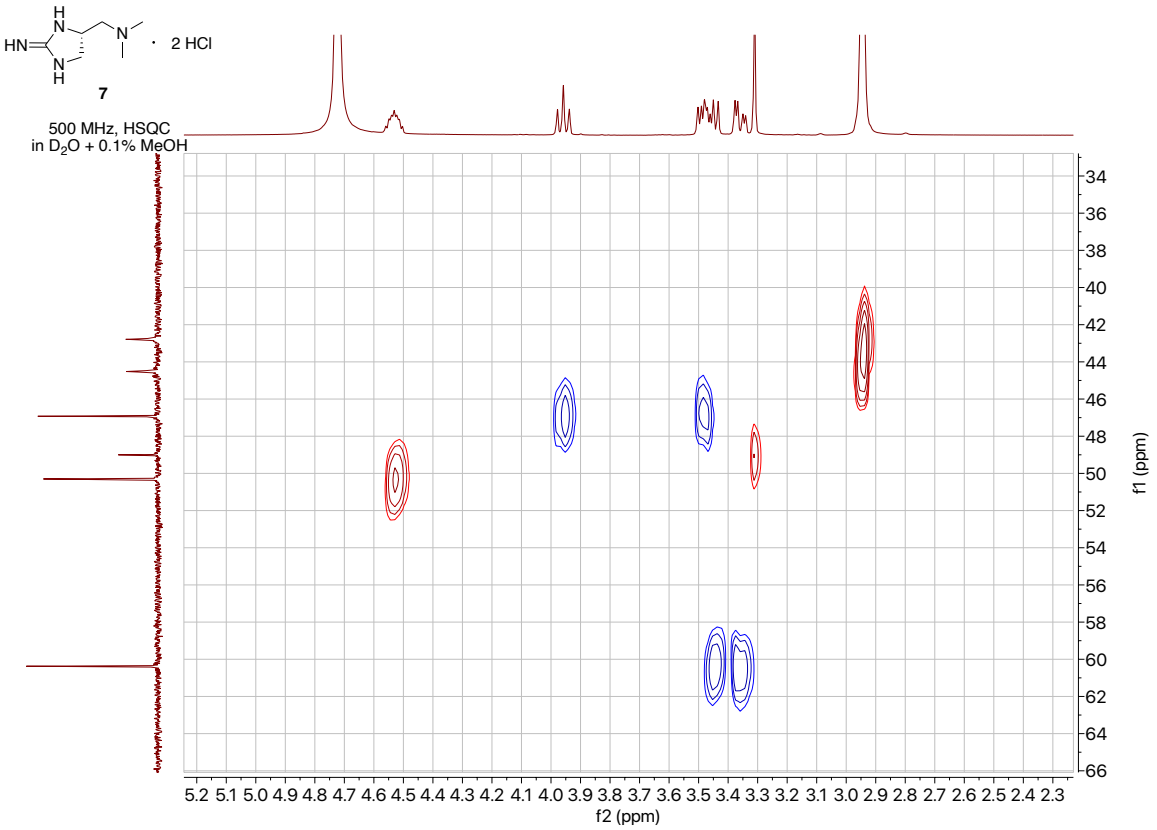


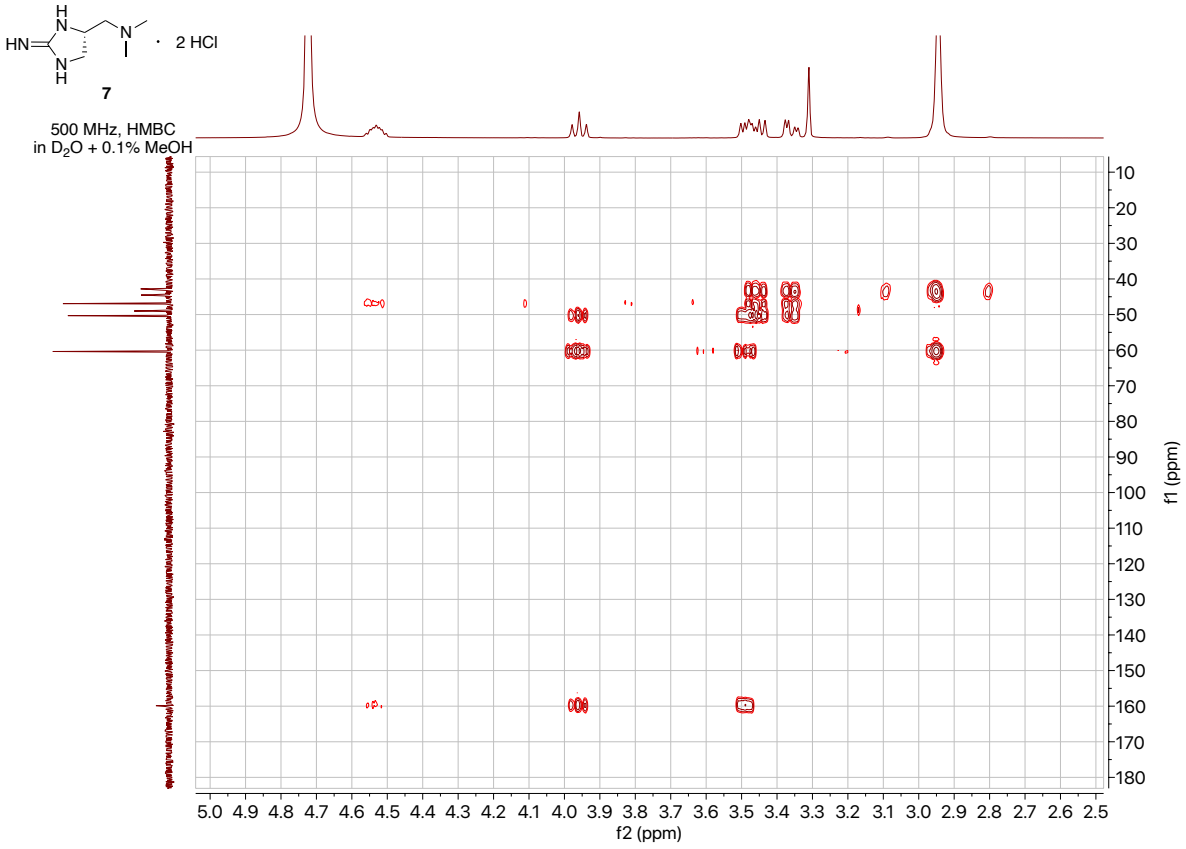
7

500 MHz, ^{1}H
in D_2O + 0.1% MeOH

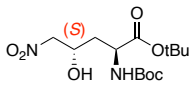

HN
N
N
H
H
· 2 HCl
7
125 MHz, 13C
in D2O + 0.1% MeOH
159.9
60.4
50.3
46.9
44.5
42.8
210 200 190 180 170 160 150 140 130 120 110 100 90 80 70 60 50 40 30 20 10 0 -10
f1 (ppm)


HN
N
N
H
H
· 2 HCl
7
500 MHz, COSY
in D2O + 0.1% MeOH
5.0 4.9 4.8 4.7 4.6 4.5 4.4 4.3 4.2 4.1 4.0 3.9 3.8 3.7 3.6 3.5 3.4 3.3 3.2 3.1 3.0 2.9 2.8
f2 (ppm)
2.7 2.8 2.9 3.0 3.1 3.2 3.3 3.4 3.5 3.6 3.7 3.8 3.9 4.0 4.1 4.2 4.3 4.4 4.5 4.6 4.7 4.8 4.9
f1 (ppm)


7
· 2 HCl
500 MHz, HSQC
in D2O + 0.1% MeOH
f2 (ppm)
f1 (ppm)


7
· 2 HCl
500 MHz, HMBC
in D2O + 0.1% MeOH
f2 (ppm)
f1 (ppm)

SI-7:

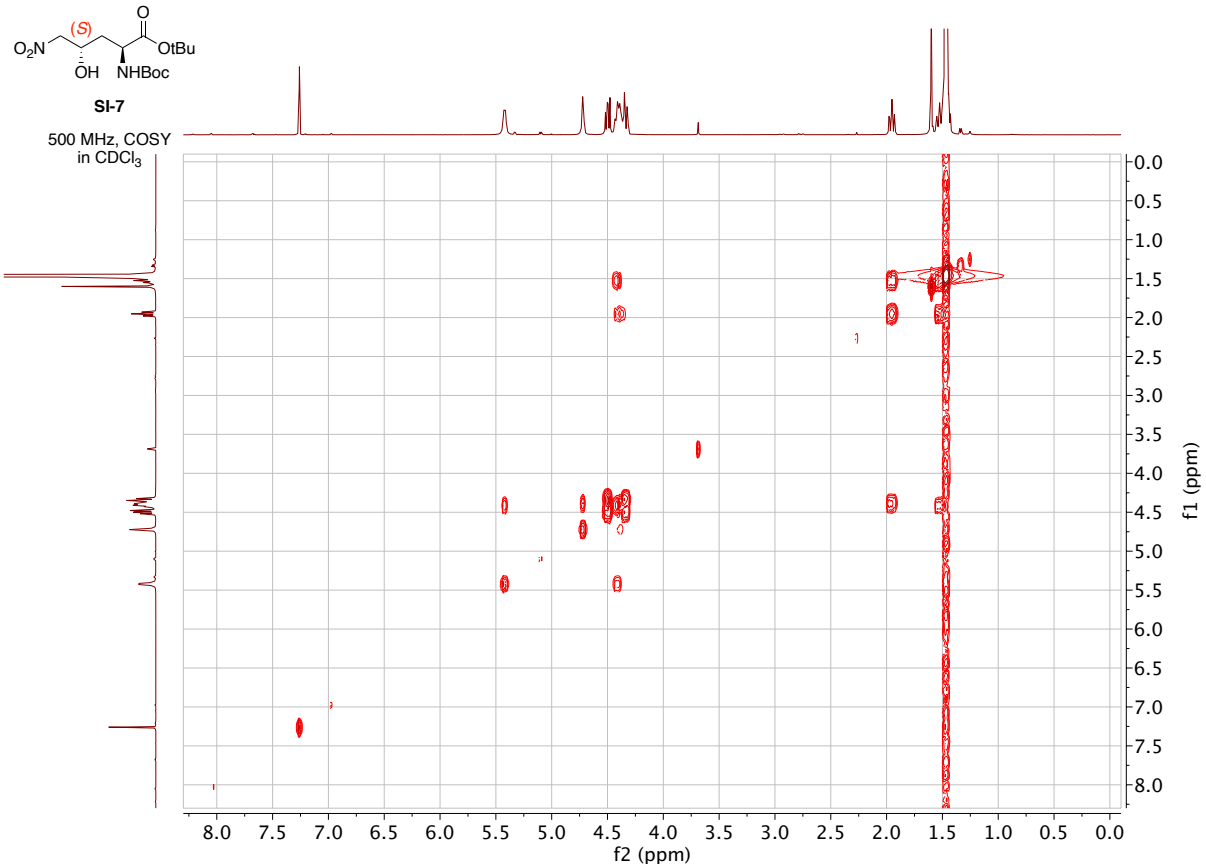


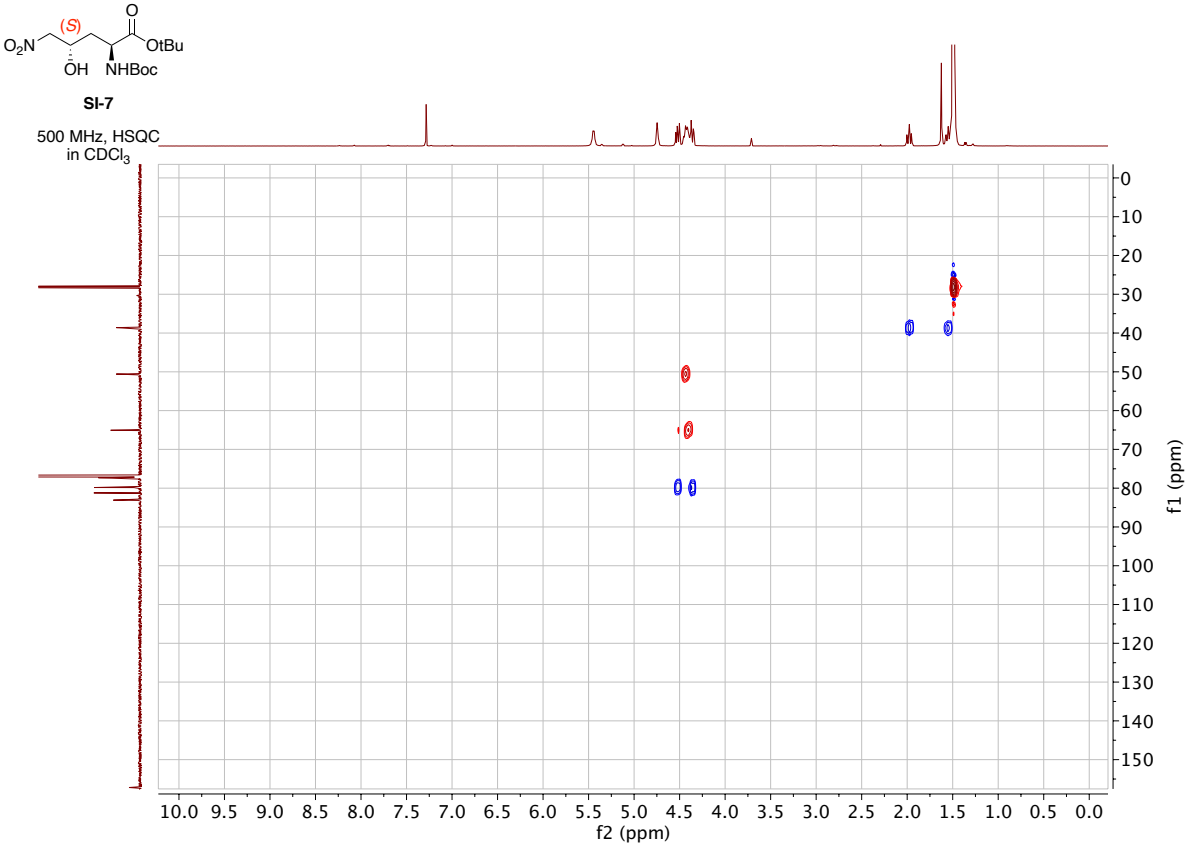
SI-7

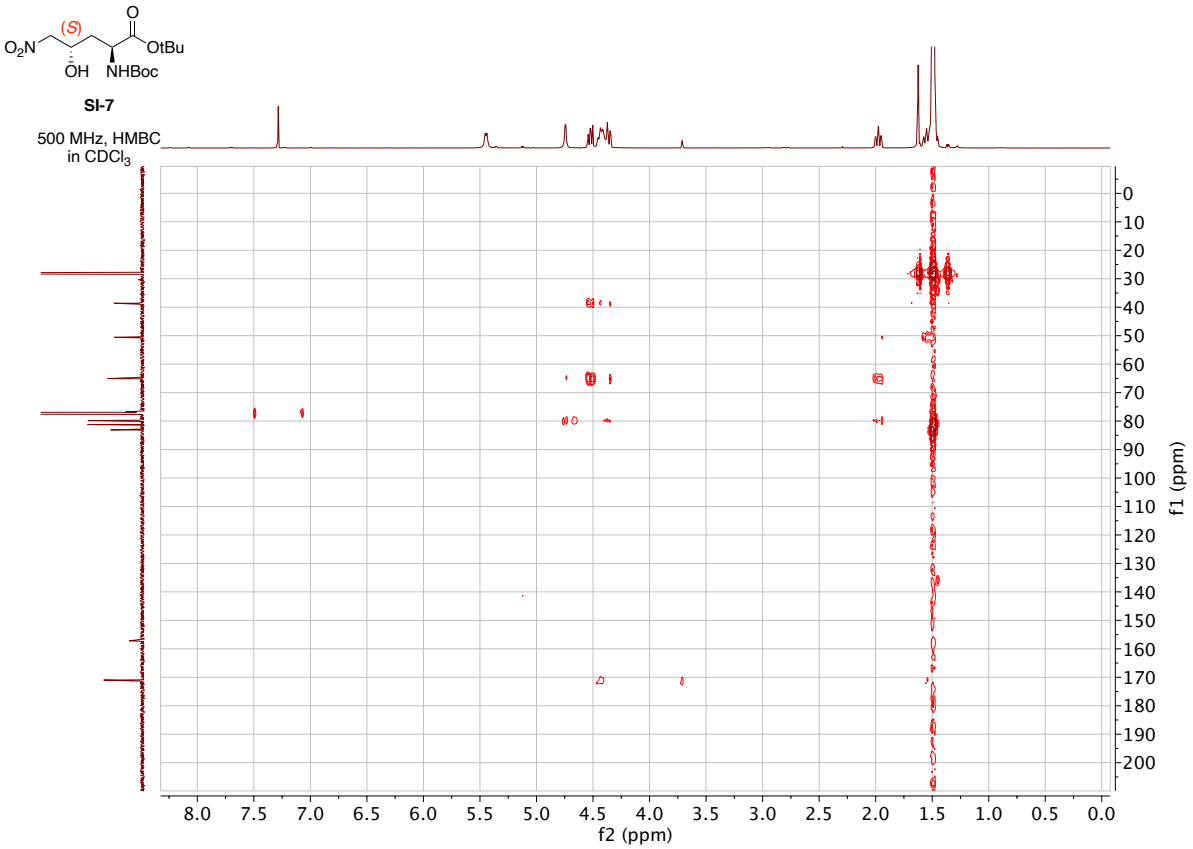
500 MHz, ^{1}H
in $CDCl_3$

SI-7

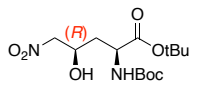
125 MHz, ^{13}C
in $CDCl_3$


O2N
(S)
OH
NHBoc
O
OtBu
SI-7
500 MHz, COSY
in CDCl3
f2 (ppm)
f1 (ppm)


O2N
(S)
OH
NHBoc
O
OtBu
SI-7
500 MHz, HSQC
in CDCl3
f2 (ppm)
f1 (ppm)

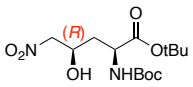

O
(S)
O2N
OtBu
OH
NHBoc
SI-7
500 MHz, HMBC
in CDCl3
f2 (ppm)
f1 (ppm)

SI-8:



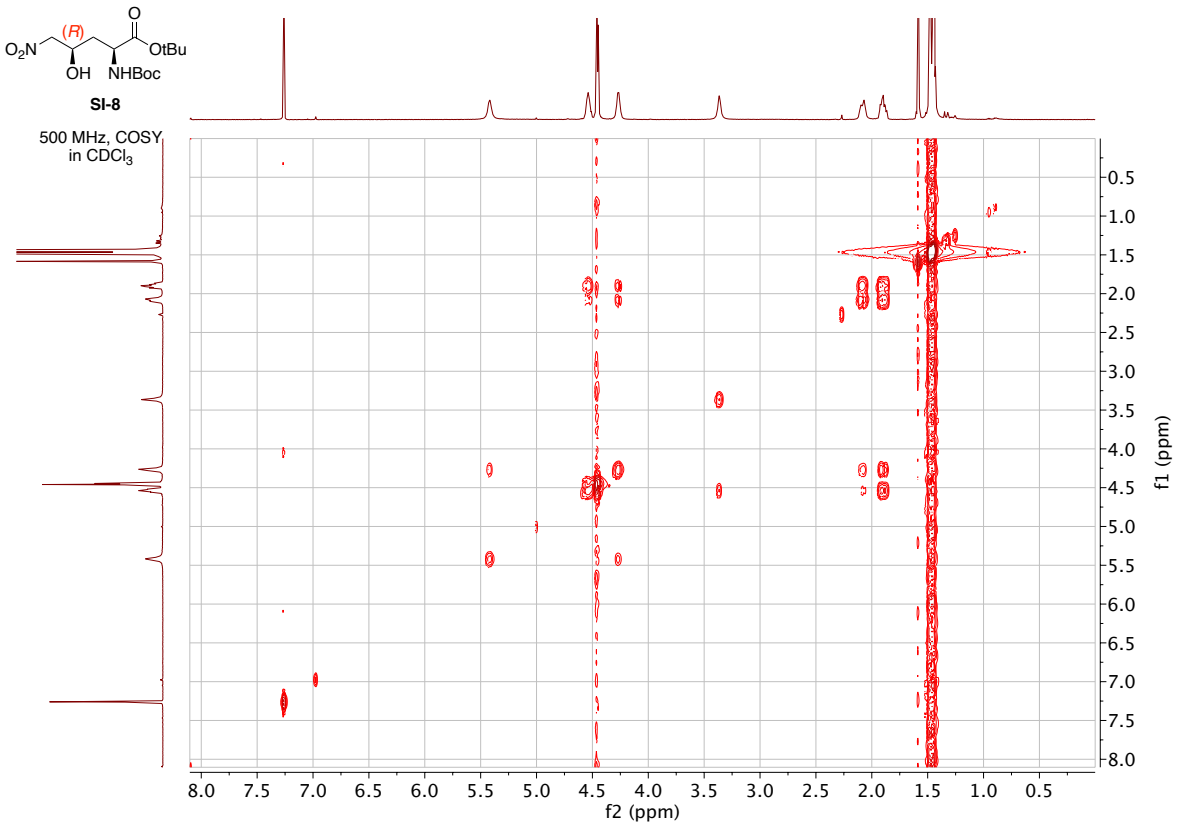
SI-8

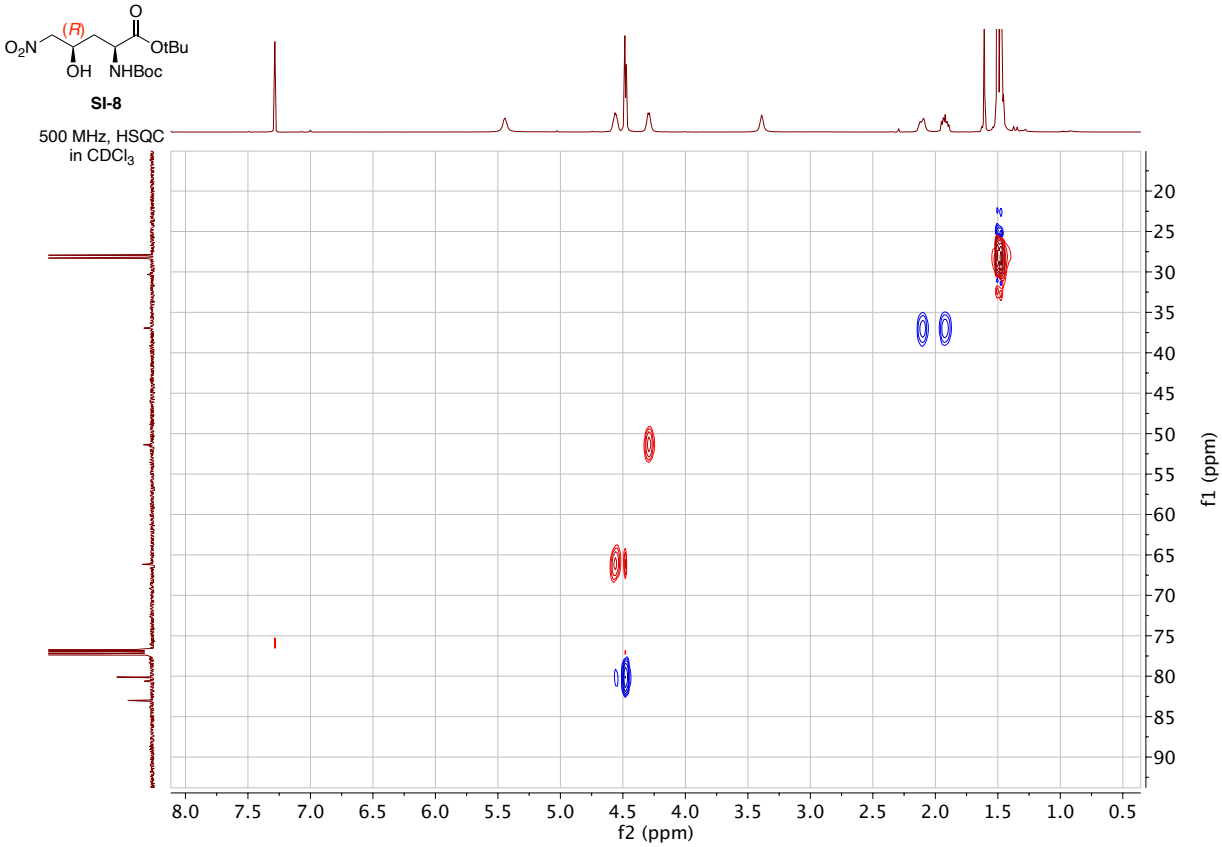
500 MHz, ^{1}H
in $CDCl_3$

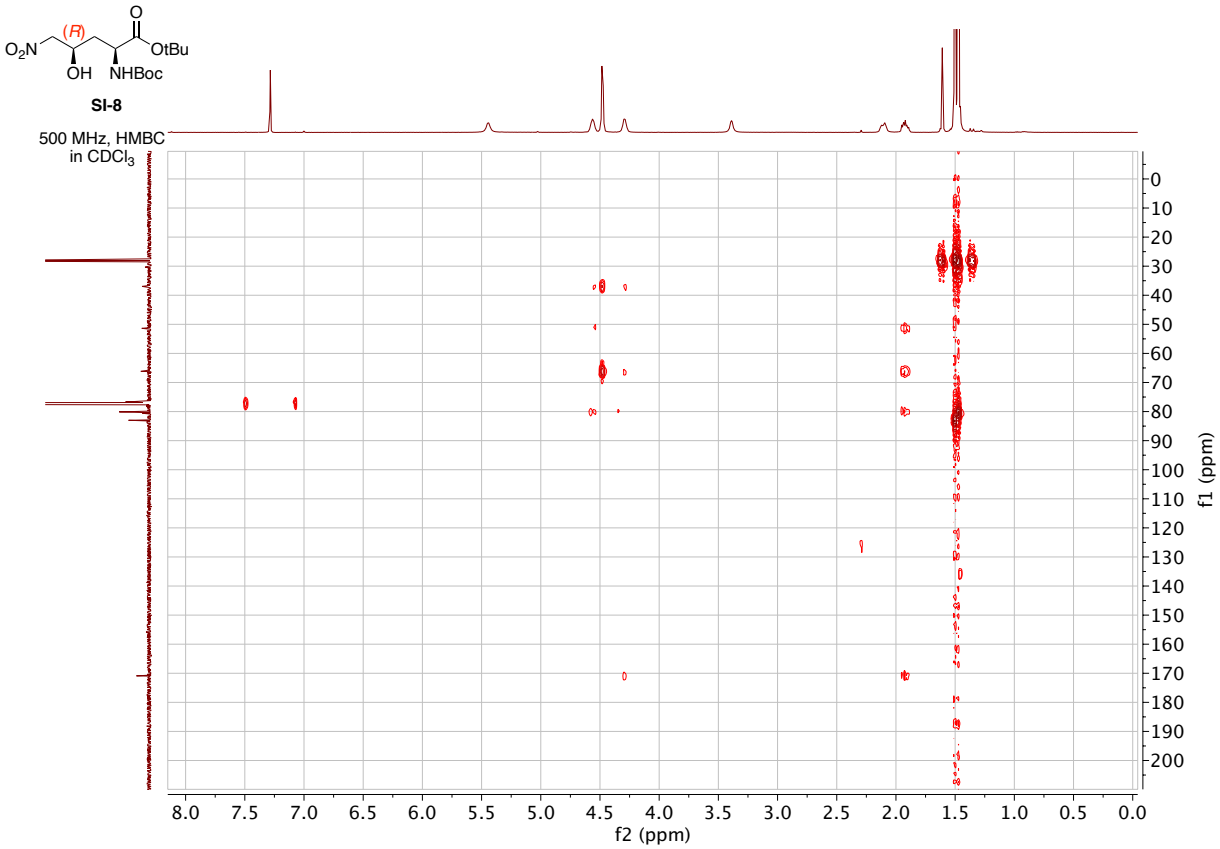


SI-8

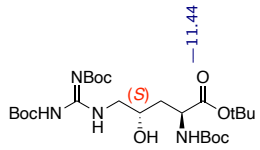
125 MHz, ^{13}C
in $CDCl_3$


O
O2N
(R)
OH
NHBoc
OtBu
SI-8
500 MHz, COSY
in CDCl3
f1 (ppm)
f2 (ppm)


O
O2N
(R)
OH
NHBoc
OtBu
SI-8
500 MHz, HSQC
in CDCl3
f1 (ppm)
f2 (ppm)

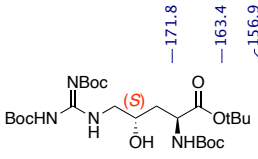

O
(R)
O2N
OtBu
OH
NHBoc
SI-8
500 MHz, HMBC
in CDCl3
f2 (ppm)
f1 (ppm)

SI-9:



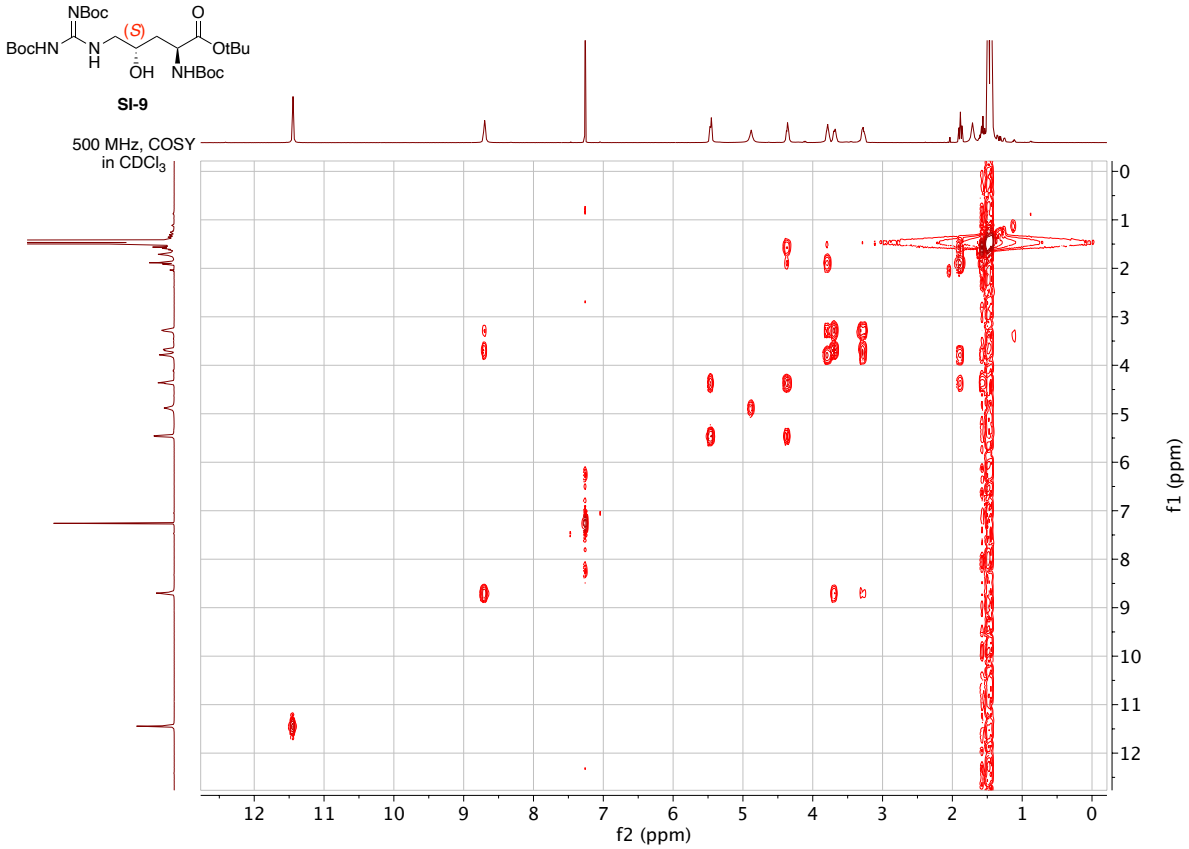
SI-9

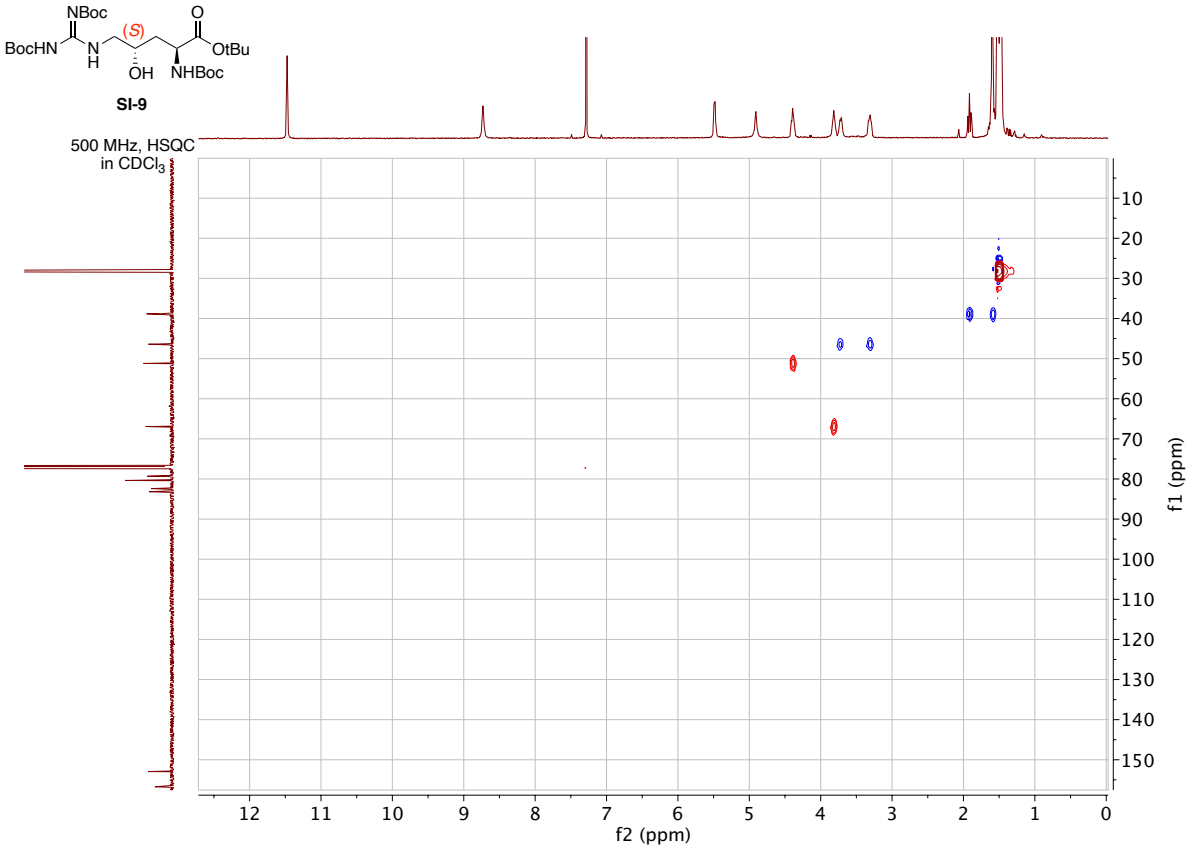
500 MHz, ^{1}H in $CDCl_3$

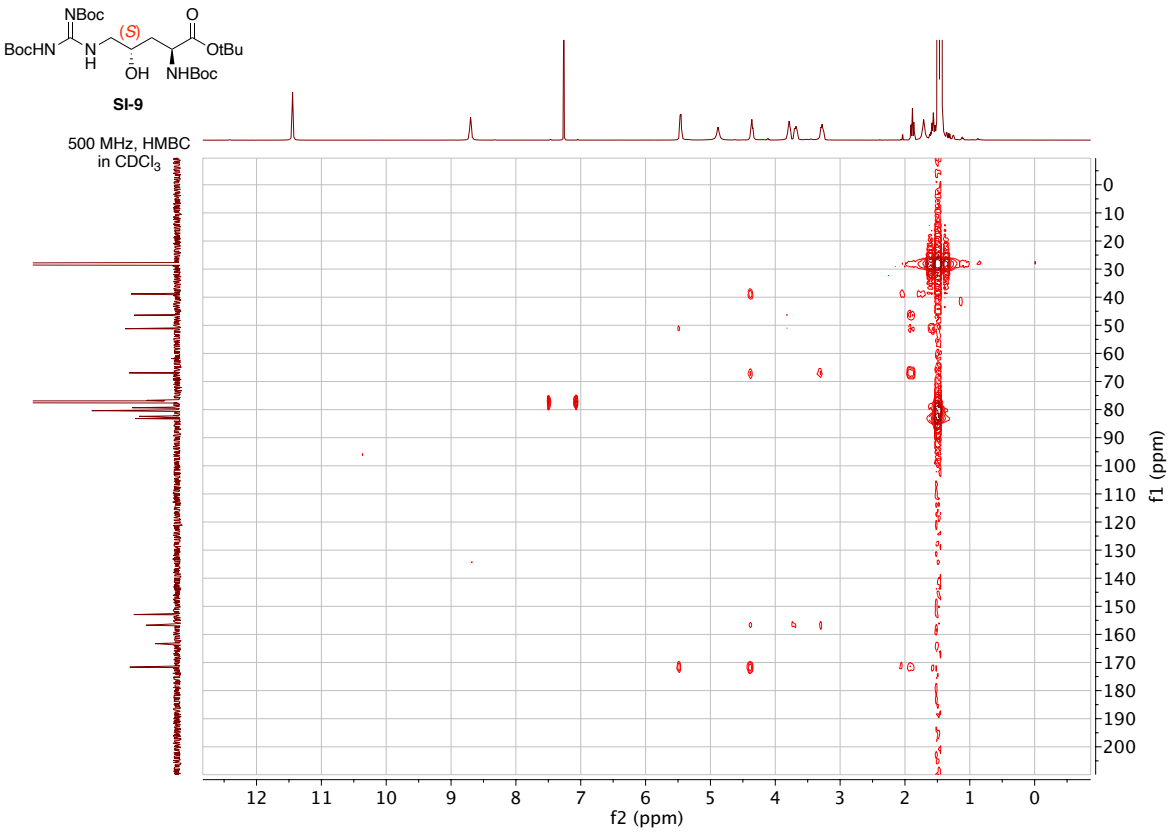


SI-9

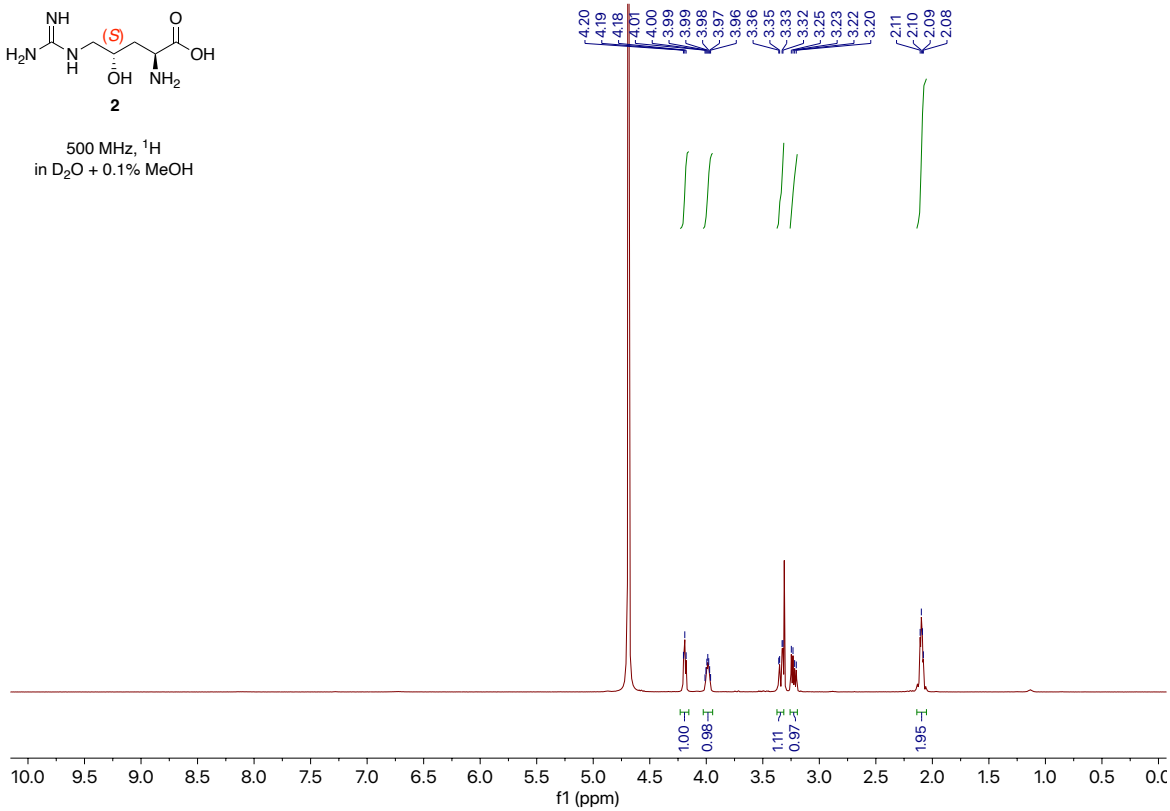
125 MHz, ^{13}C in $CDCl_3$

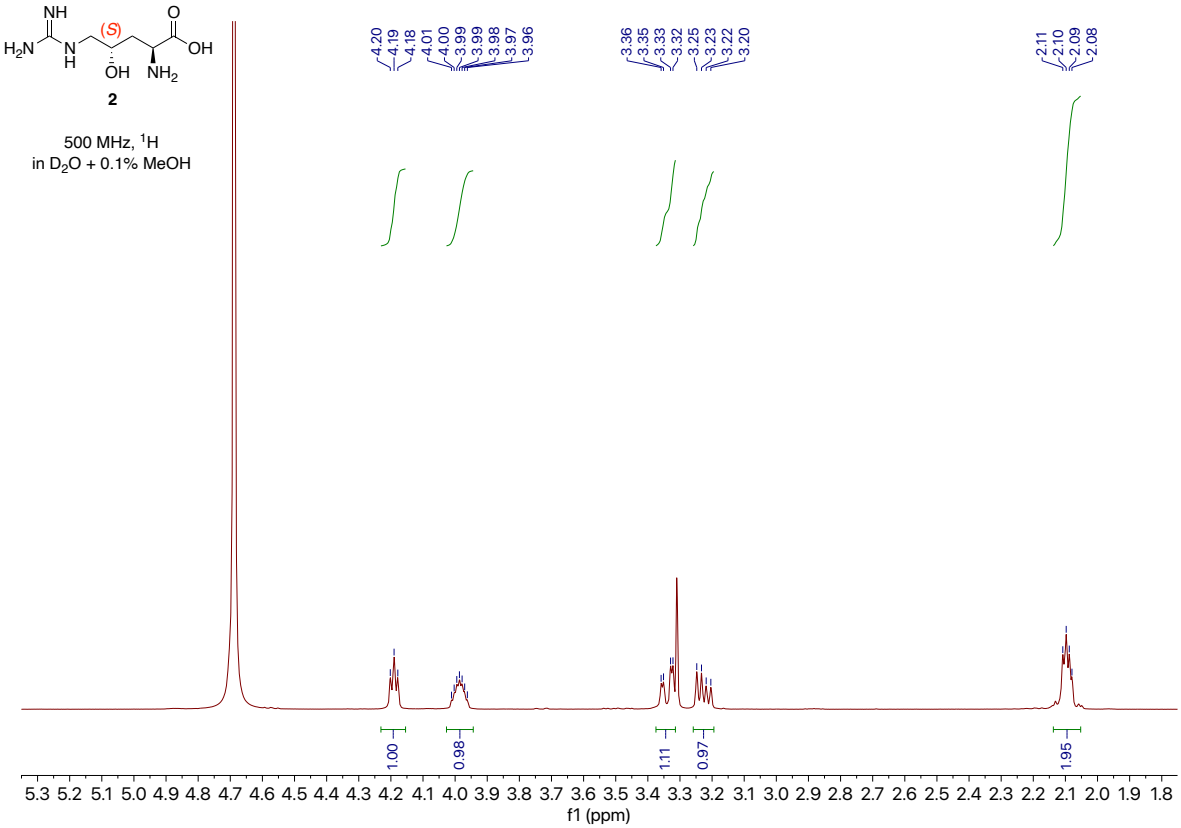

NBoc
BocHN
N
H
(S)
OH
O
OtBu
NHBoc
SI-9
500 MHz, COSY
in CDCl3
f2 (ppm)
f1 (ppm)

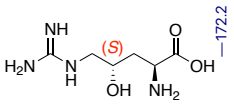

NBoc
BocHN
N
H
(S)
OH
O
OtBu
NHBoc
SI-9
500 MHz, HSQC
in CDCl3
f2 (ppm)
f1 (ppm)


NBoc
O
(S)
BocHN
N
H
OH
NHBoc
OtBu
SI-9
500 MHz, HMBC
in CDCl3
f2 (ppm)
f1 (ppm)

2:

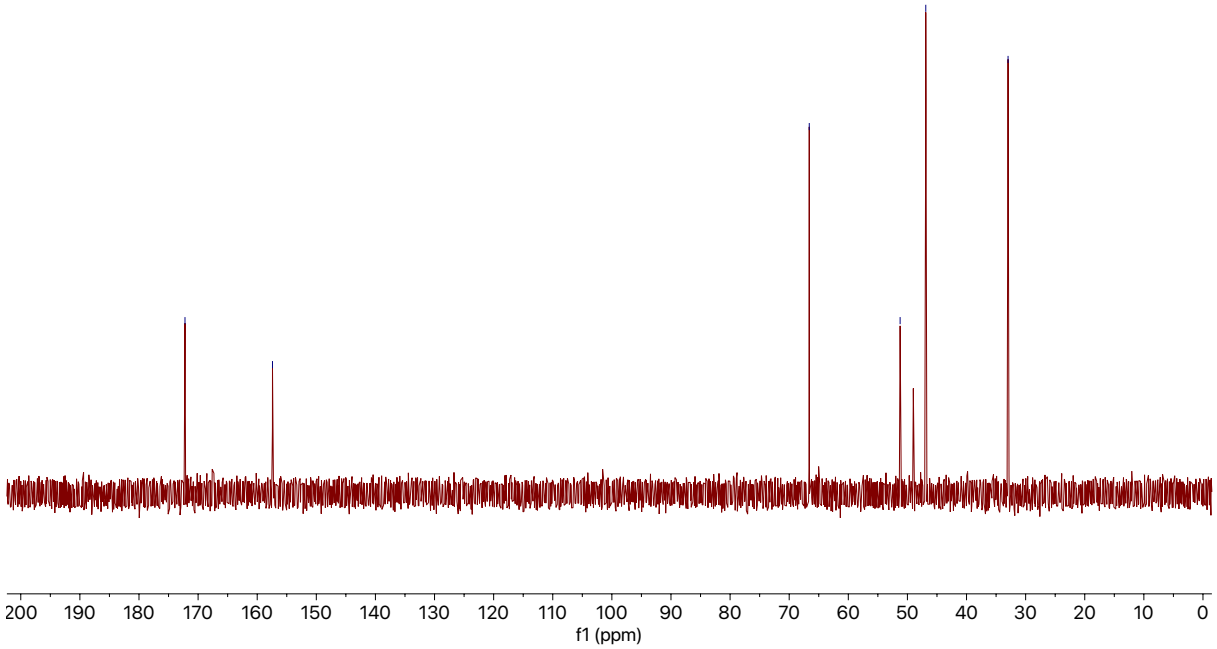

2
500 MHz, ^{1}H
in D_2O + 0.1% MeOH
4.20 4.19 4.18 4.01 4.00 3.99 3.99 3.98 3.97 3.96 3.36 3.35 3.33 3.32 3.25 3.23 3.22 3.20 2.11 2.10 2.09 2.08
1.00 0.98 1.11 0.97 1.95
f1 (ppm)

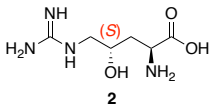

2
500 MHz, ^{1}H
in D_2O + 0.1% MeOH
4.20 4.19 4.18 4.01 4.00 3.99 3.99 3.98 3.97 3.96 3.36 3.35 3.33 3.32 3.25 3.23 3.22 3.20 2.11 2.10 2.09 2.08
1.00 0.98 1.11 0.97 1.95
f1 (ppm)



2

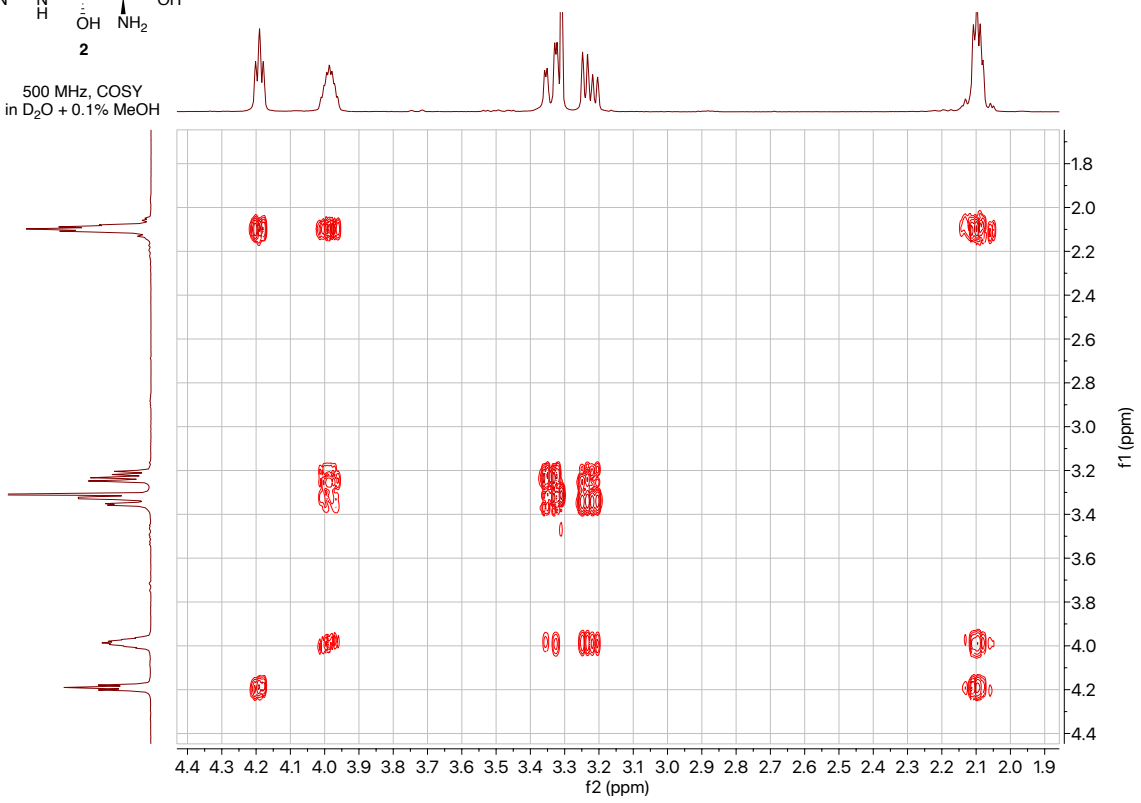
125 MHz, ^{13}C
in D_2O + 0.1% MeOH

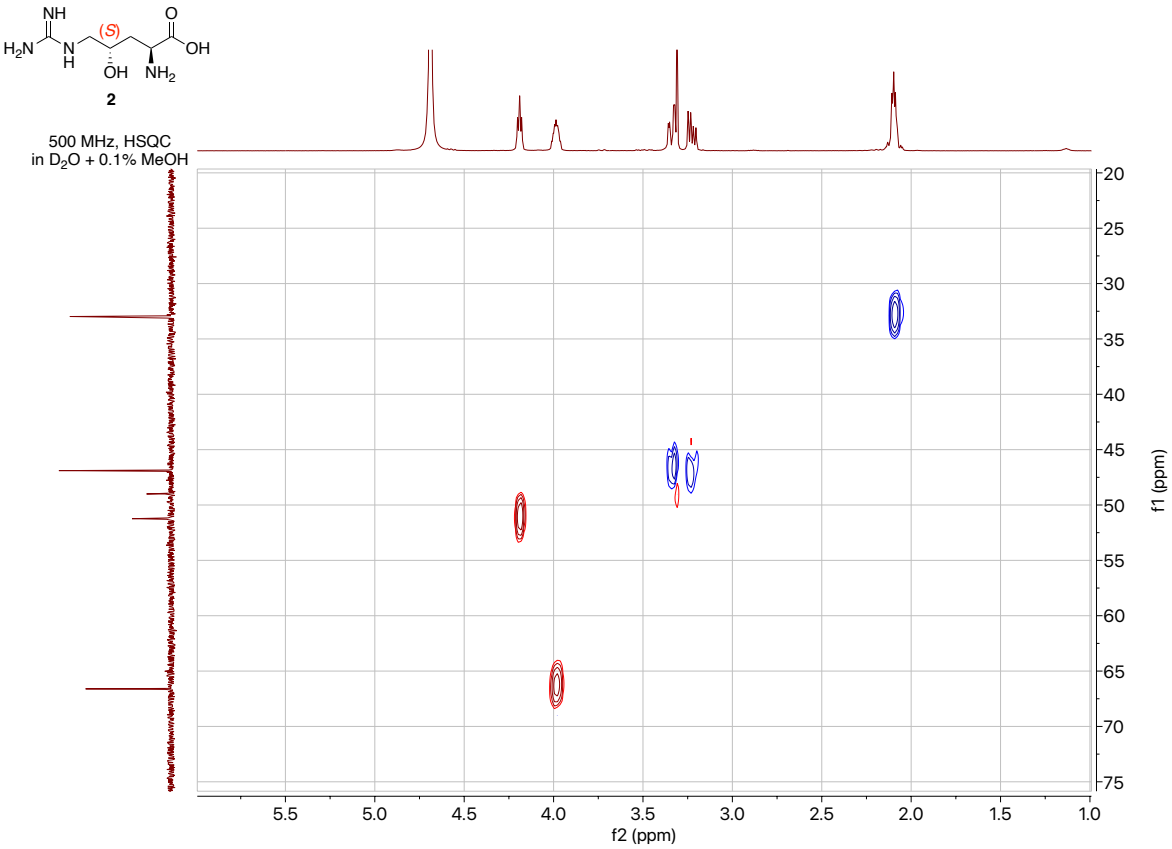


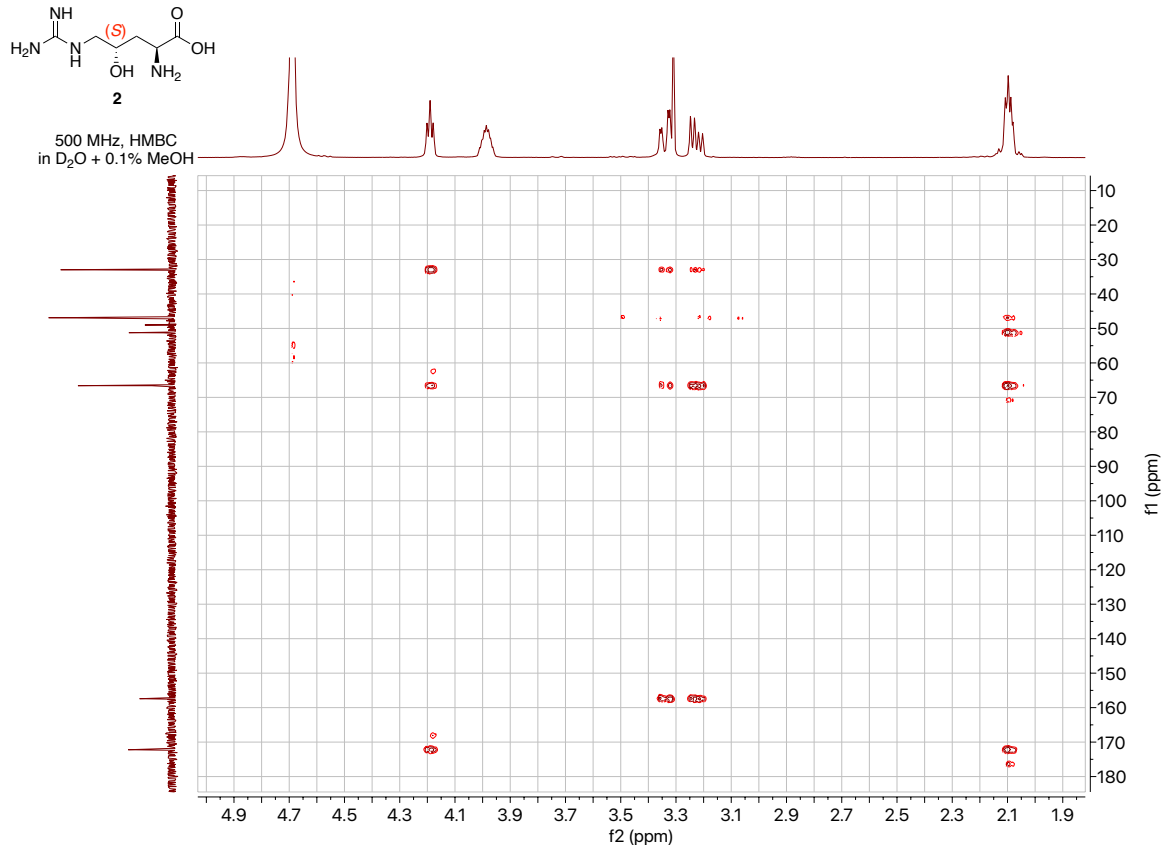


2

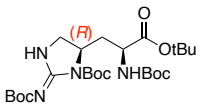
500 MHz, COSY
in D_2O + 0.1% MeOH




NH
O
(S)
H2N
N
H
OH
OH
NH2
2
500 MHz, HSQC
in D2O + 0.1% MeOH
f1 (ppm)
f2 (ppm)

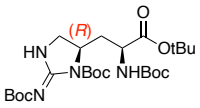

NH
O
(S)
H2N
N
H
OH
OH
NH2
2
500 MHz, HMBC
in D2O + 0.1% MeOH
f1 (ppm)
f2 (ppm)

SI-10:



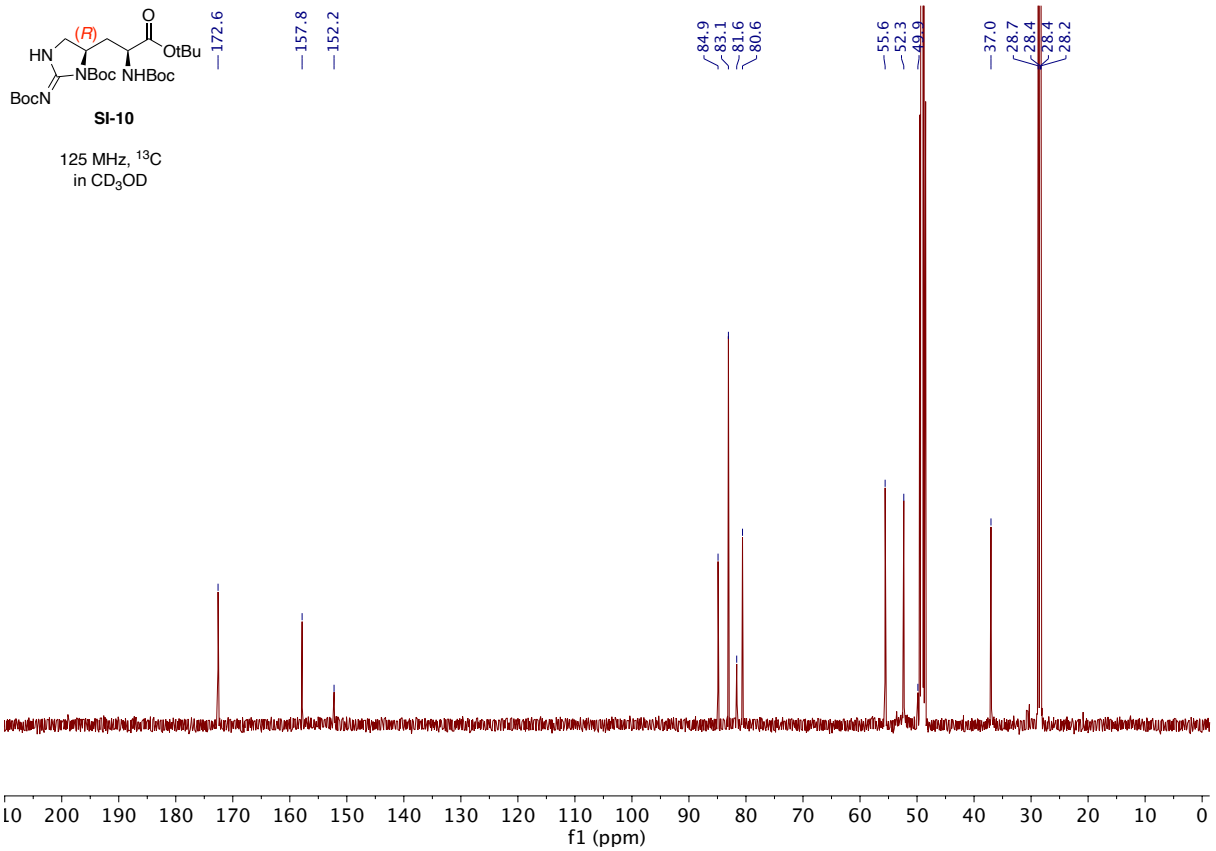
SI-10

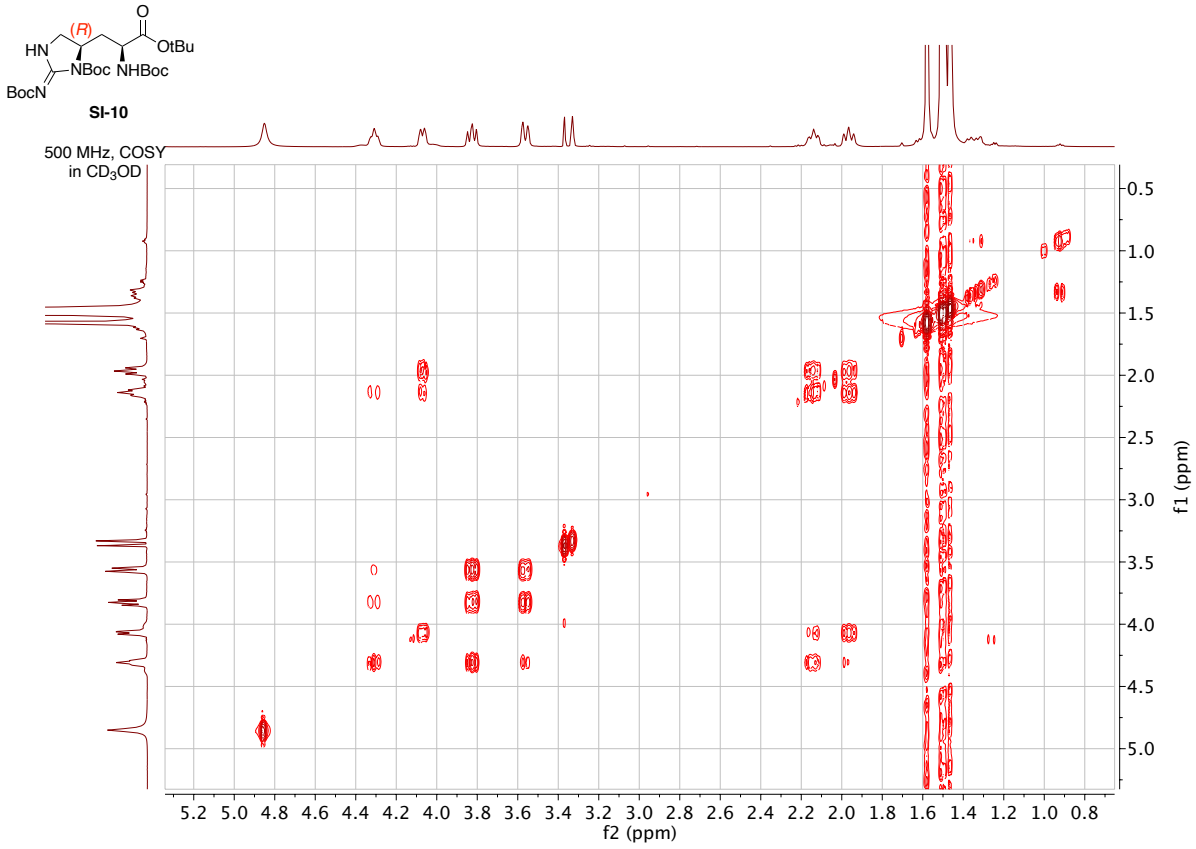
500 MHz, ^{1}H
in CD_3OD

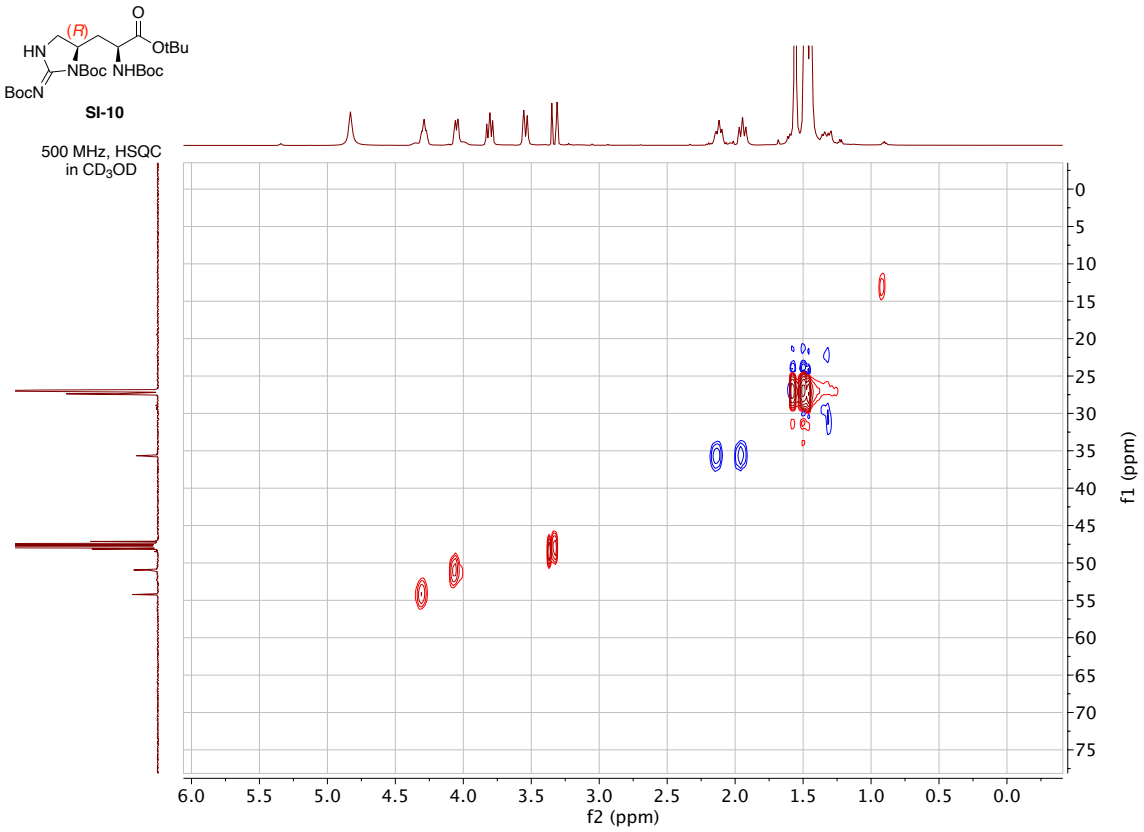


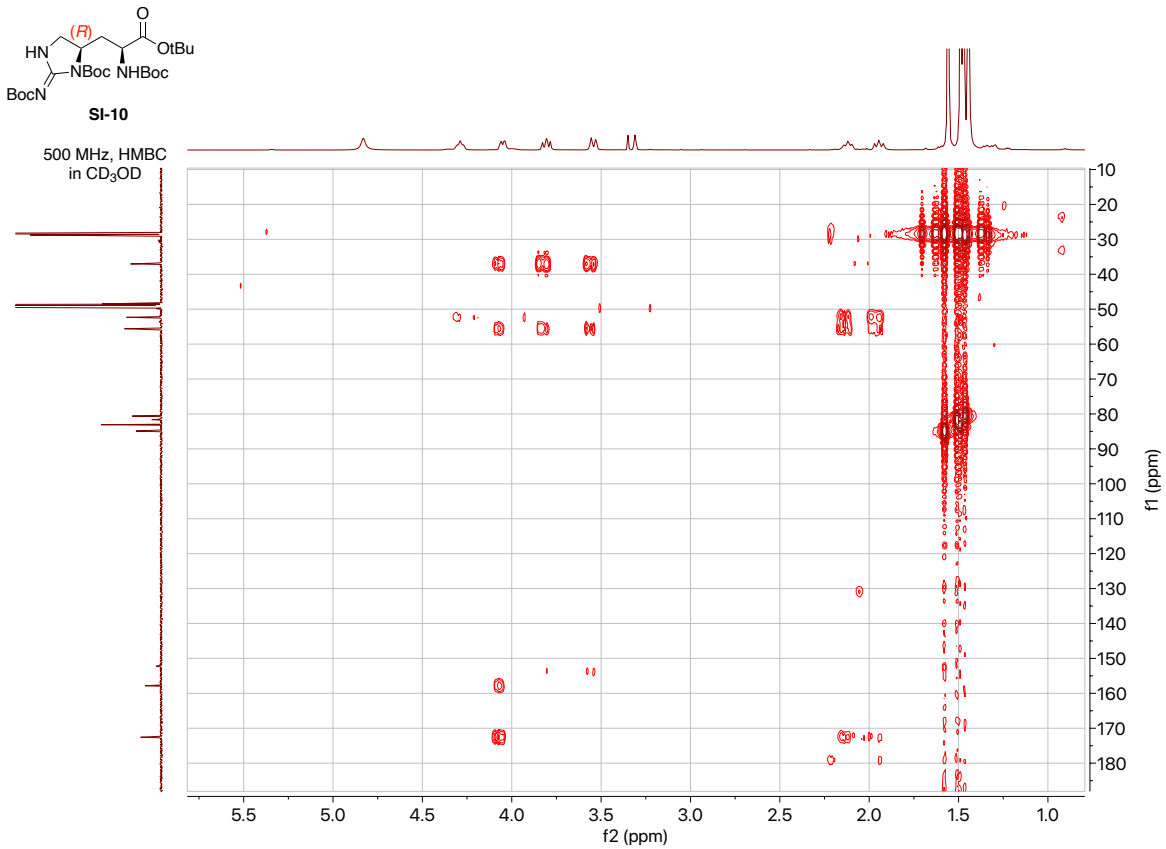
SI-10

500 MHz, ^{1}H
in CD_3OD

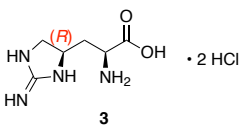

SI-10
125 MHz, ^{13}C
in CD_3OD
172.6
157.8
152.2
84.9
83.1
81.6
80.6
55.6
52.3
49.9
37.0
28.7
28.4
28.4
28.2
f1 (ppm)


SI-10
500 MHz, COSY
in CD_3OD
f1 (ppm)
f2 (ppm)


(R)
HN
NBoc
NHBoc
OtBu
O
BocN
SI-10
500 MHz, HSQC
in CD3OD
f2 (ppm)
f1 (ppm)

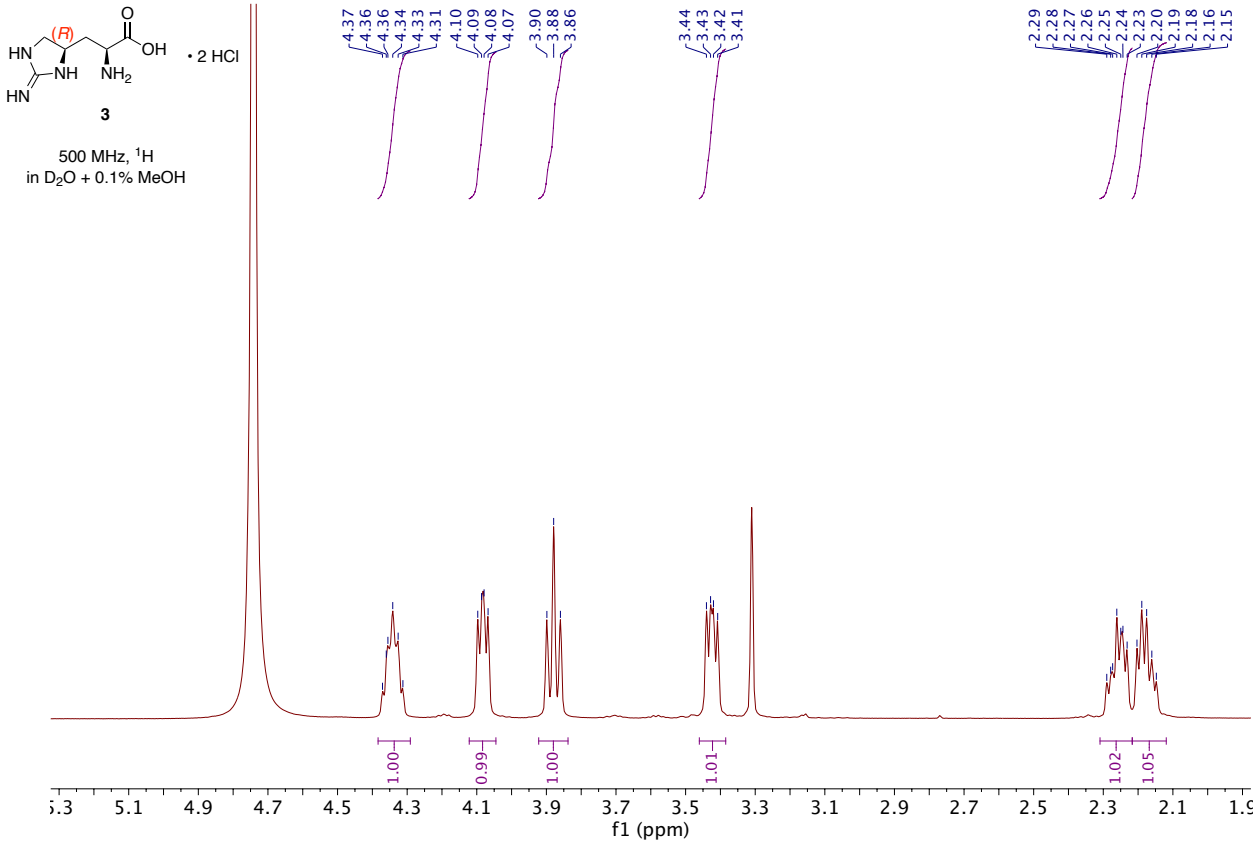

(R)
HN
NBoc
NHBoc
OtBu
O
BocN
SI-10
500 MHz, HMBC
in CD3OD
f2 (ppm)
f1 (ppm)

3:



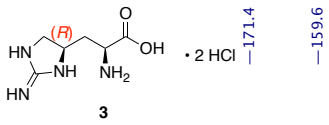
3

500 MHz, ^{1}H
in D_2O + 0.1% MeOH



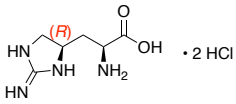
3

500 MHz, ^{1}H
in D_2O + 0.1% MeOH



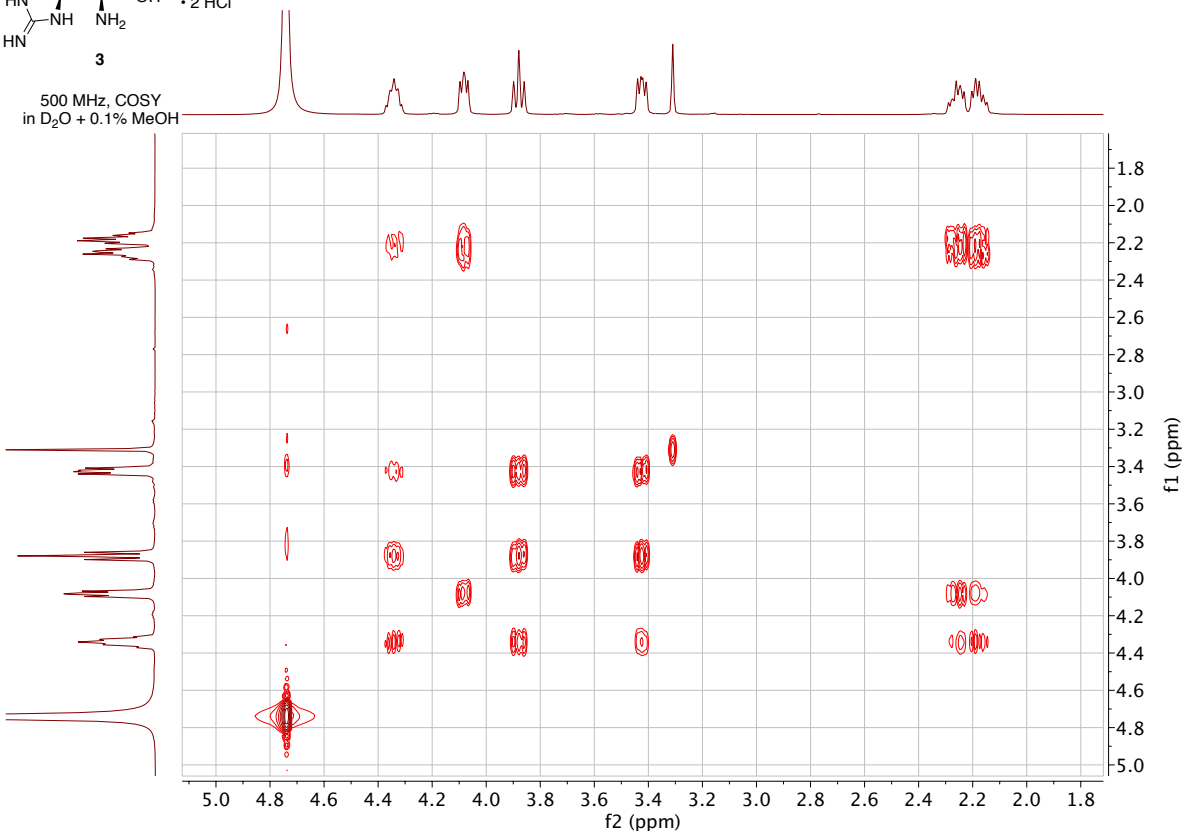
3

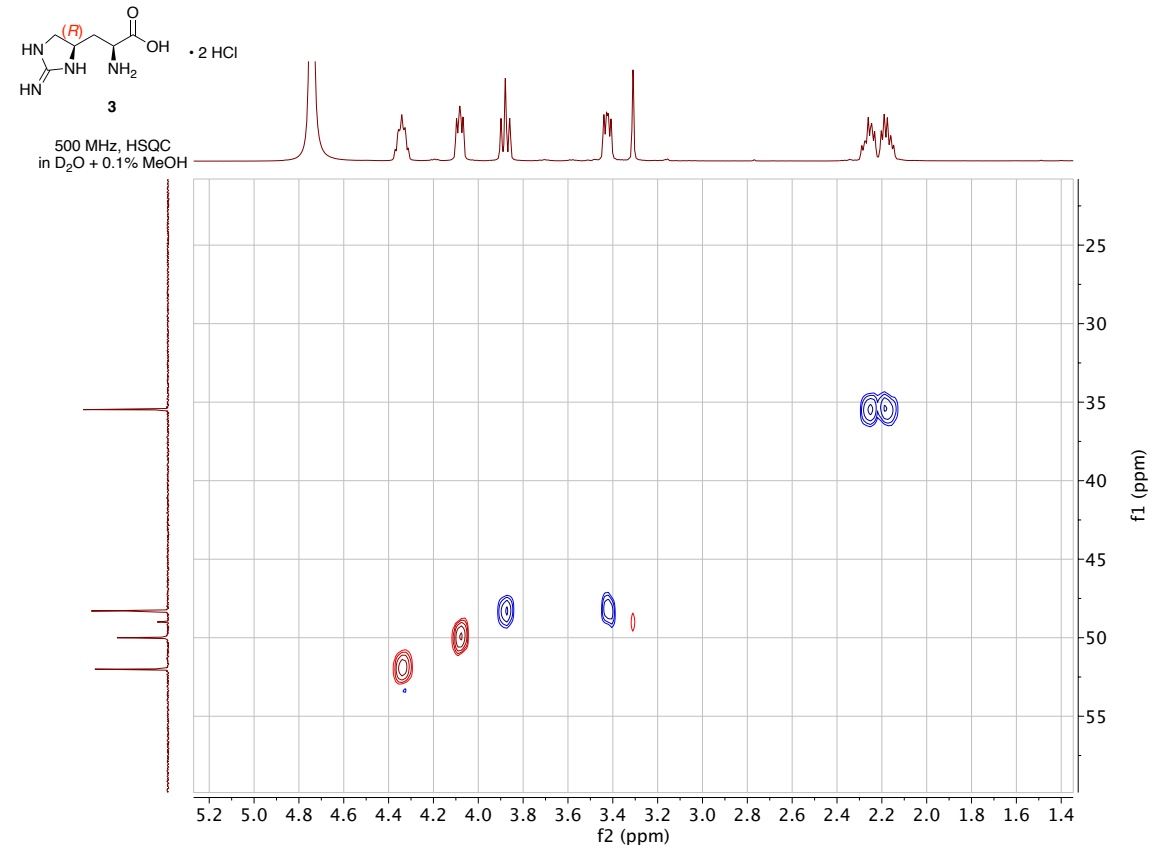
125 MHz, ^{13}C
in D_2O + 0.1% MeOH

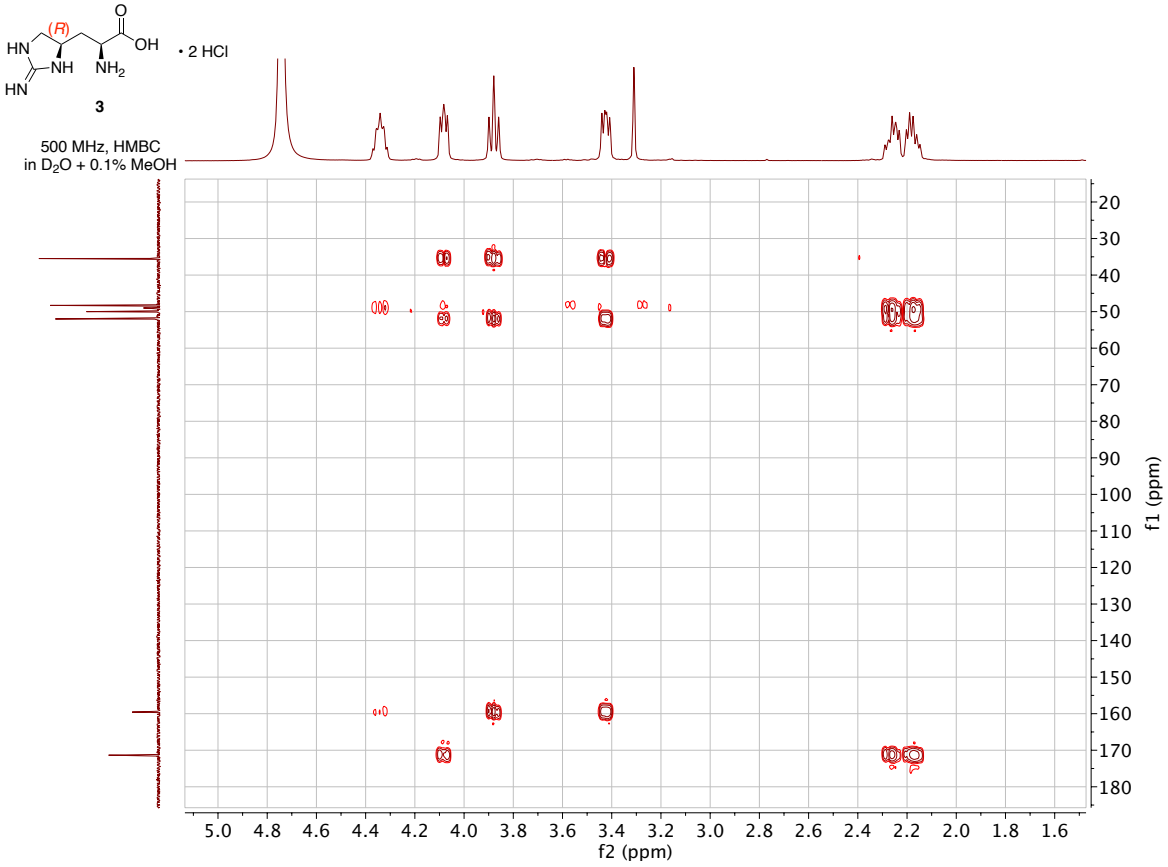


3

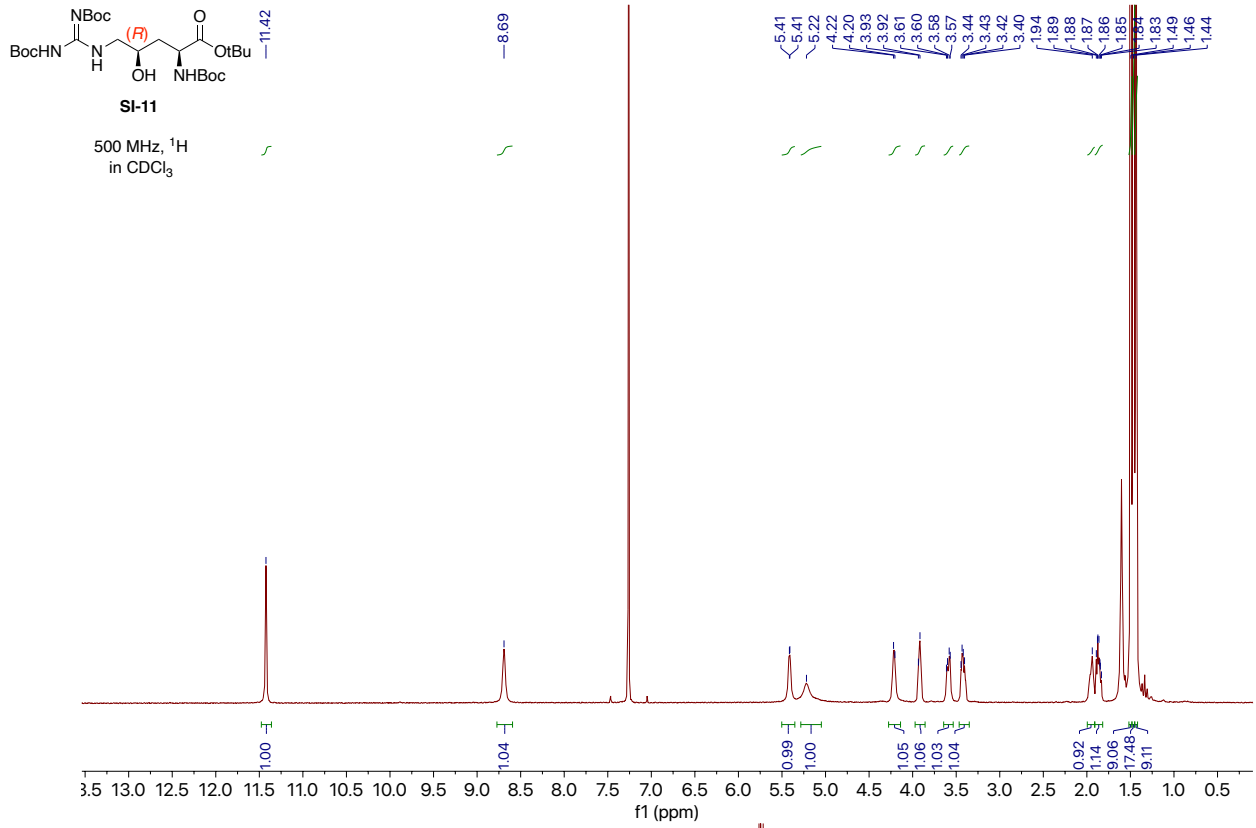
500 MHz, COSY
in D_2O + 0.1% MeOH

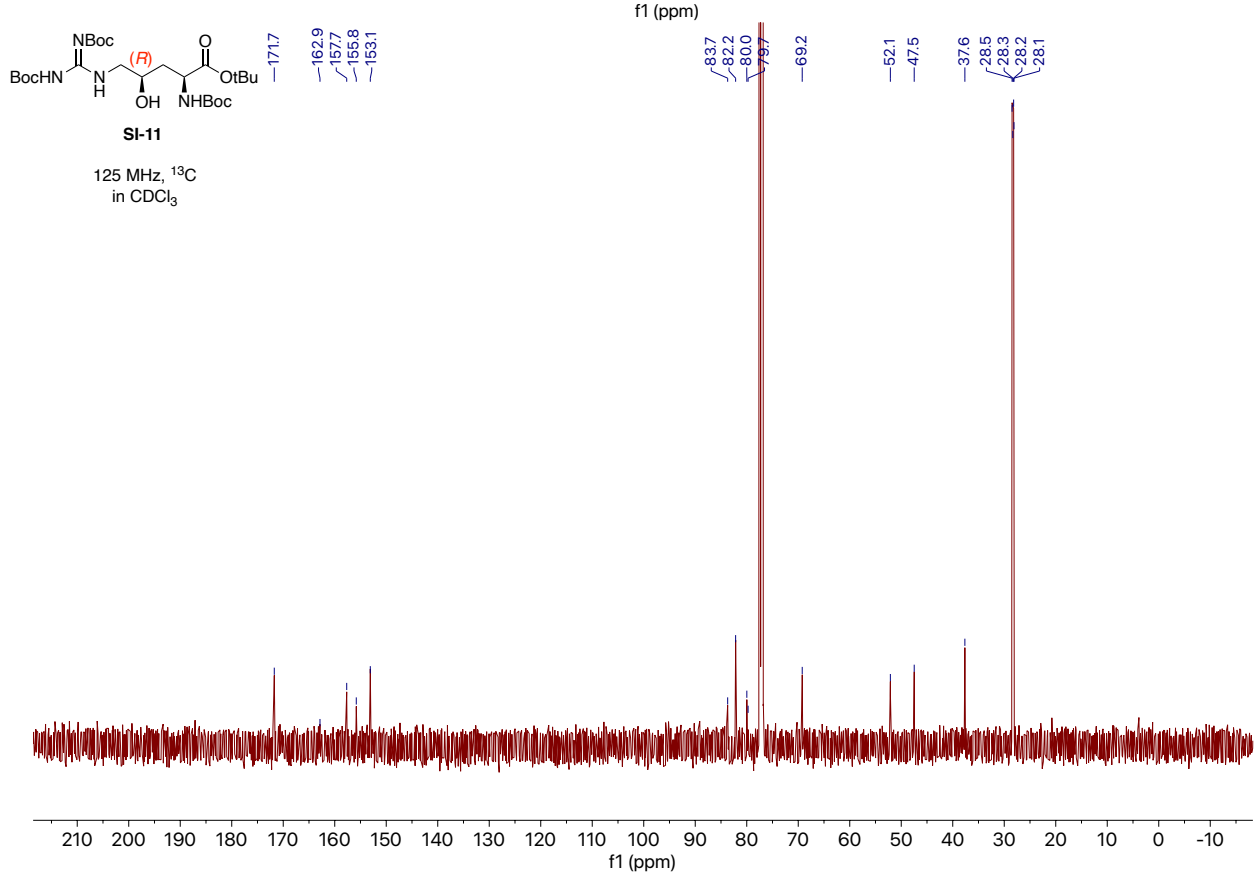


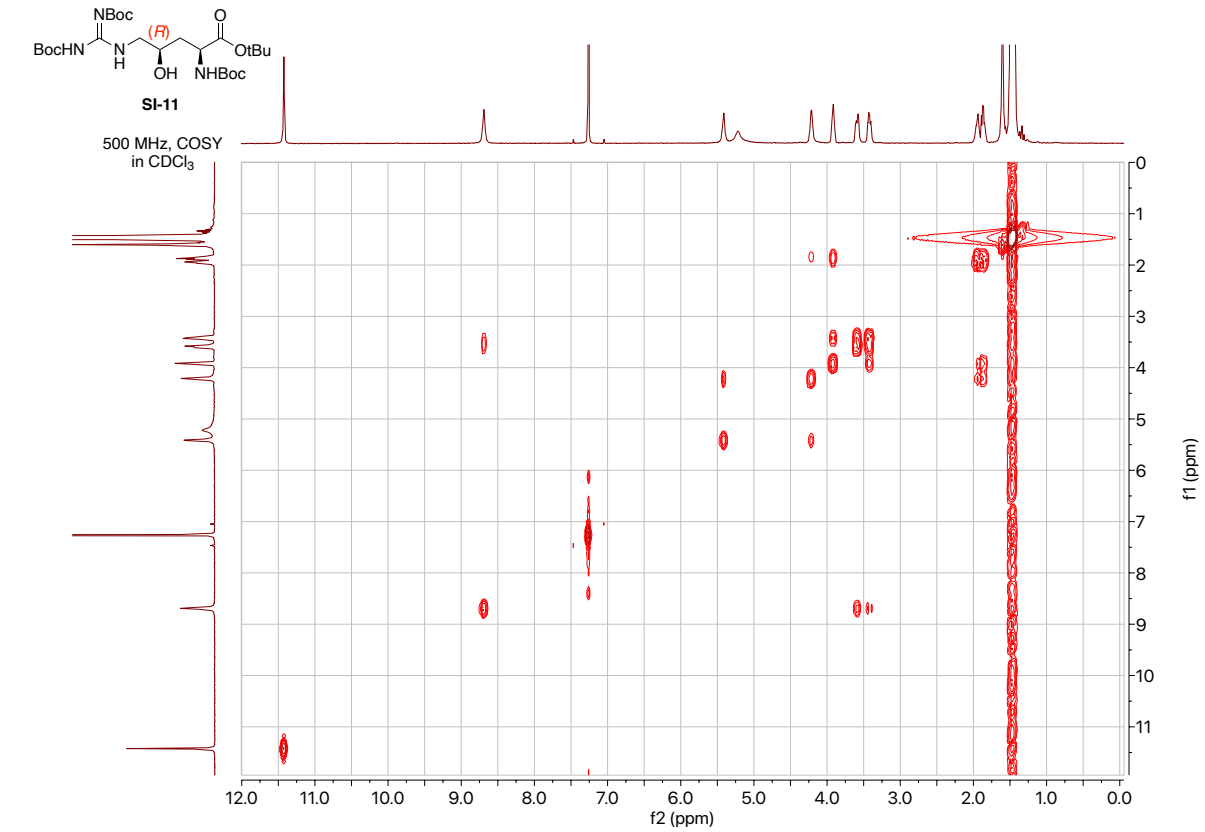

(R)
HN
NH
HN
NH2
OH
O
• 2 HCl
3
500 MHz, HSQC
in D2O + 0.1% MeOH
f1 (ppm)
f2 (ppm)

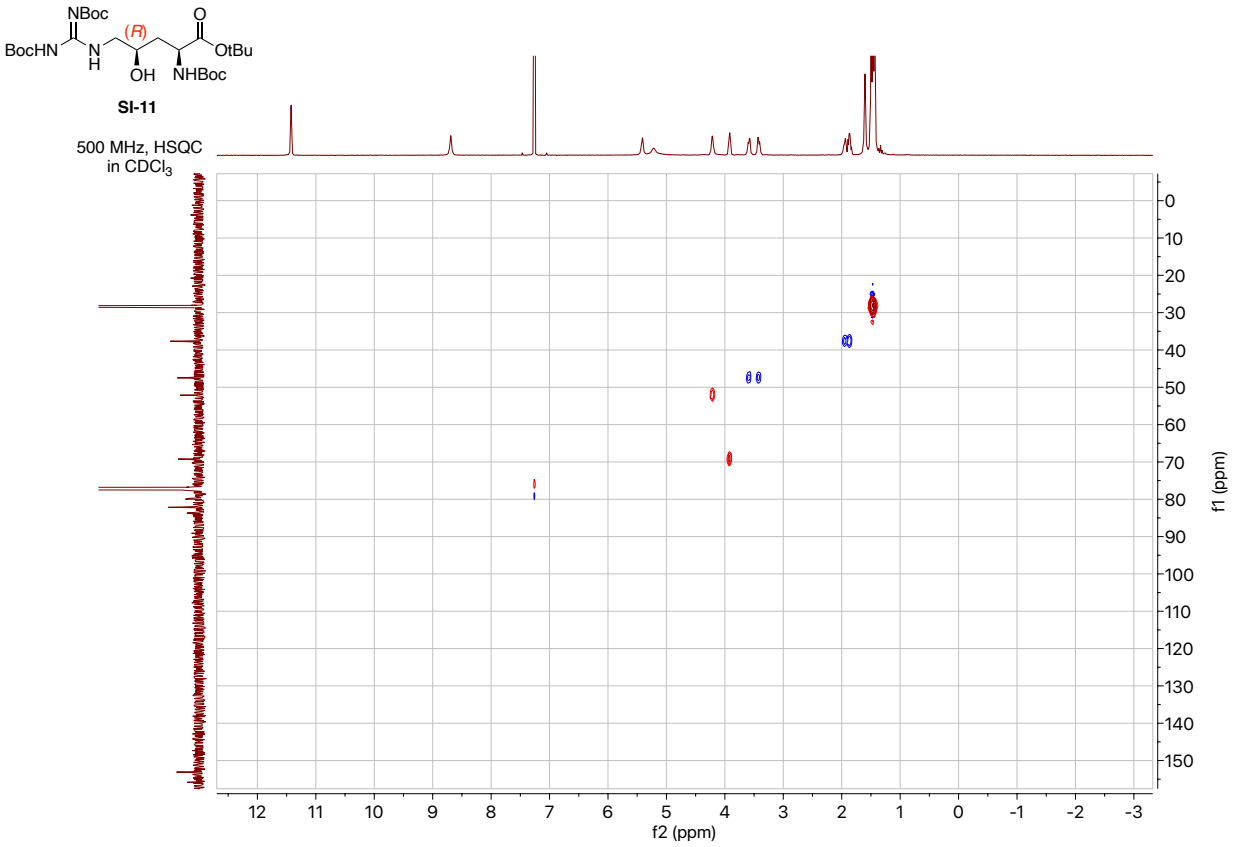

(R)
HN
NH
HN
NH2
OH
O
• 2 HCl
3
500 MHz, HMBC
in D2O + 0.1% MeOH
f1 (ppm)
f2 (ppm)

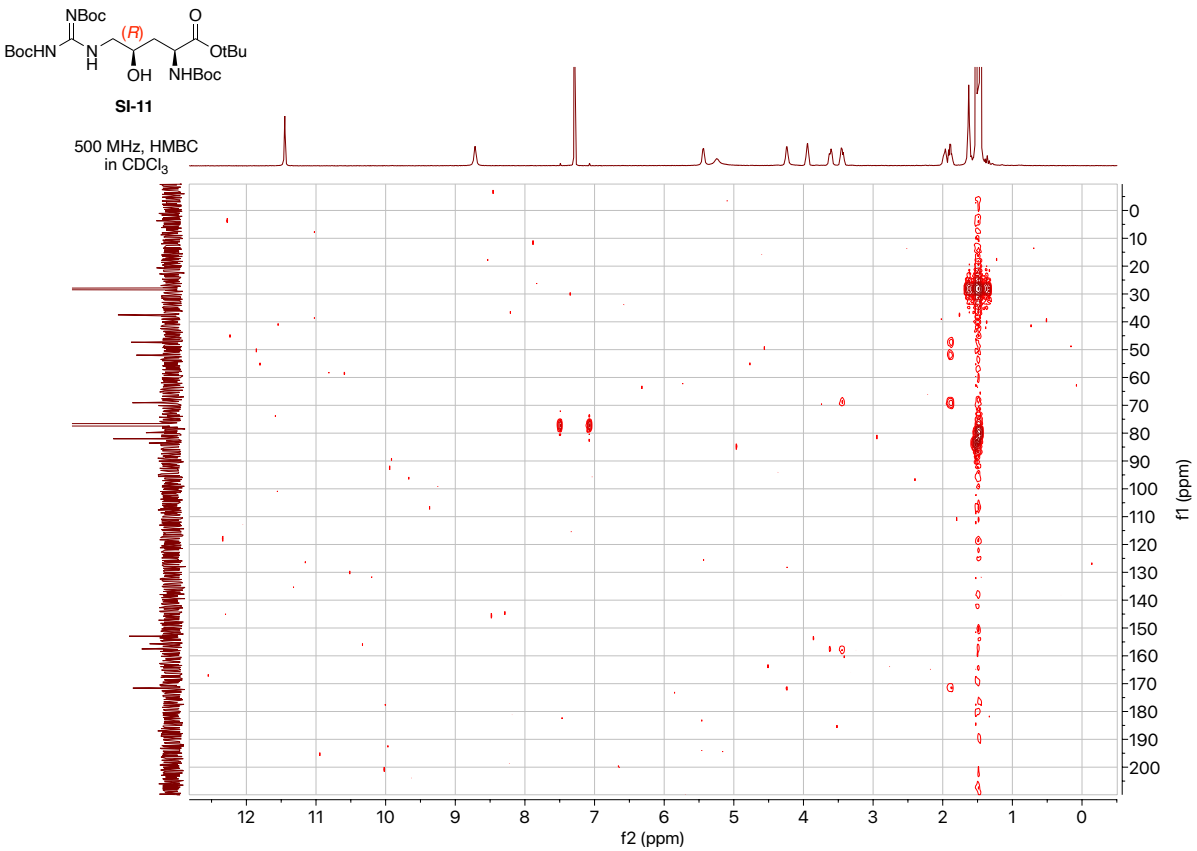
SI-11:


NBoc
BocHN
N
H
(R)
OH
NHBoc
O
OtBu
SI-11
500 MHz, 1H
in CDCl3
11.42
8.69
5.41
5.41
5.22
4.22
4.20
3.93
3.92
3.61
3.60
3.58
3.57
3.44
3.43
3.42
3.40
1.94
1.89
1.88
1.87
1.86
1.85
1.83
1.49
1.46
1.44
1.00
1.04
0.99
1.00
1.05
1.06
1.03
1.04
0.92
1.14
9.06
17.48
9.11
f1 (ppm)

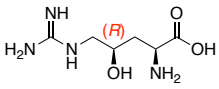

NBoc
BocHN
N
H
(R)
OH
NHBoc
O
OtBu
SI-11
125 MHz, 13C
in CDCl3
171.7
162.9
157.7
155.8
153.1
83.7
82.2
80.0
79.7
69.2
52.1
47.5
37.6
28.5
28.3
28.2
28.1
f1 (ppm)


NBoc
BocHN
N
H
(R)
OH
NHBoc
O
OtBu
SI-11
500 MHz, COSY
in CDCl3
f2 (ppm)
f1 (ppm)


NBoc
BocHN
N
H
(R)
OH
NHBoc
O
OtBu
SI-11
500 MHz, HSQC
in CDCl3
f2 (ppm)
f1 (ppm)

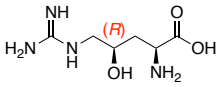

NBoc
O
(R)
BocHN
N
H
OH
NHBoc
OtBu
SI-11
500 MHz, HMBC
in CDCl3
f2 (ppm)
f1 (ppm)

SI-12:



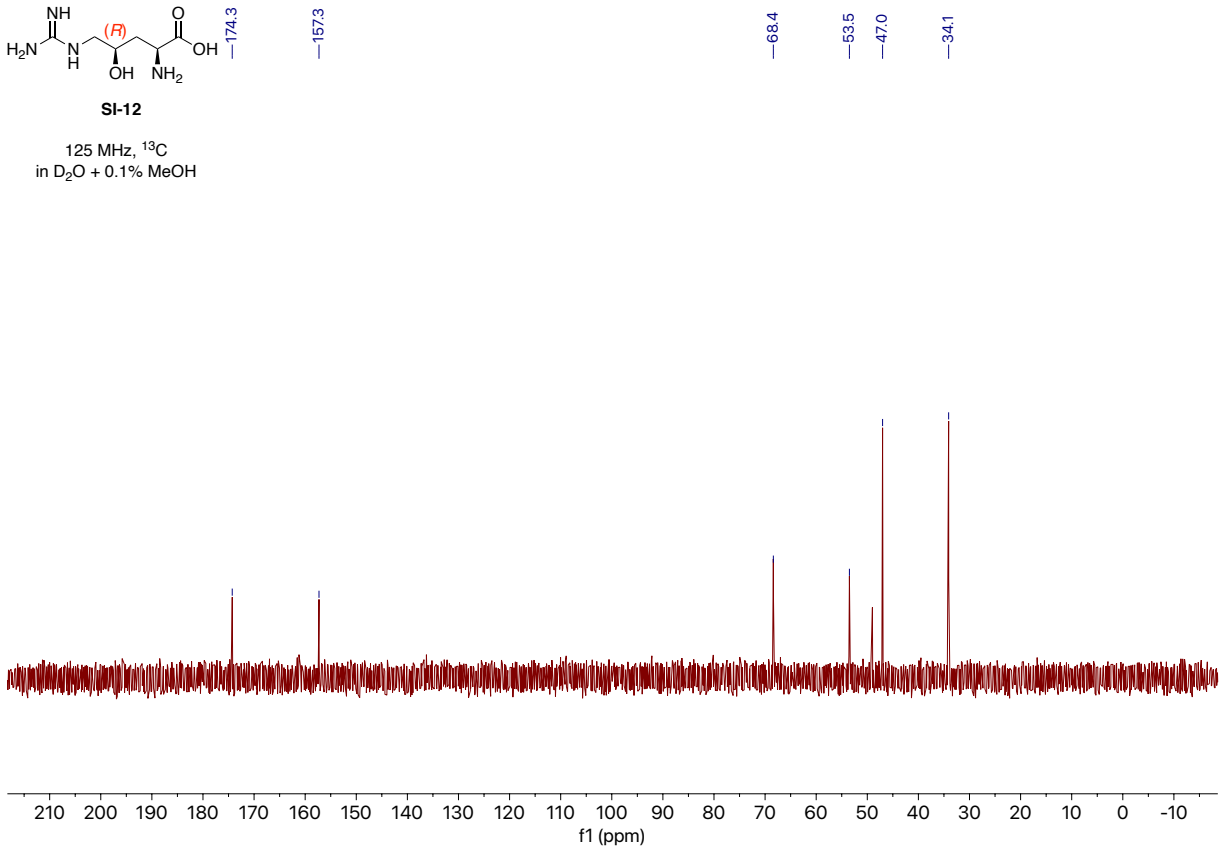
SI-12

500 MHz, ^{1}H
in D_2O + 0.1% MeOH



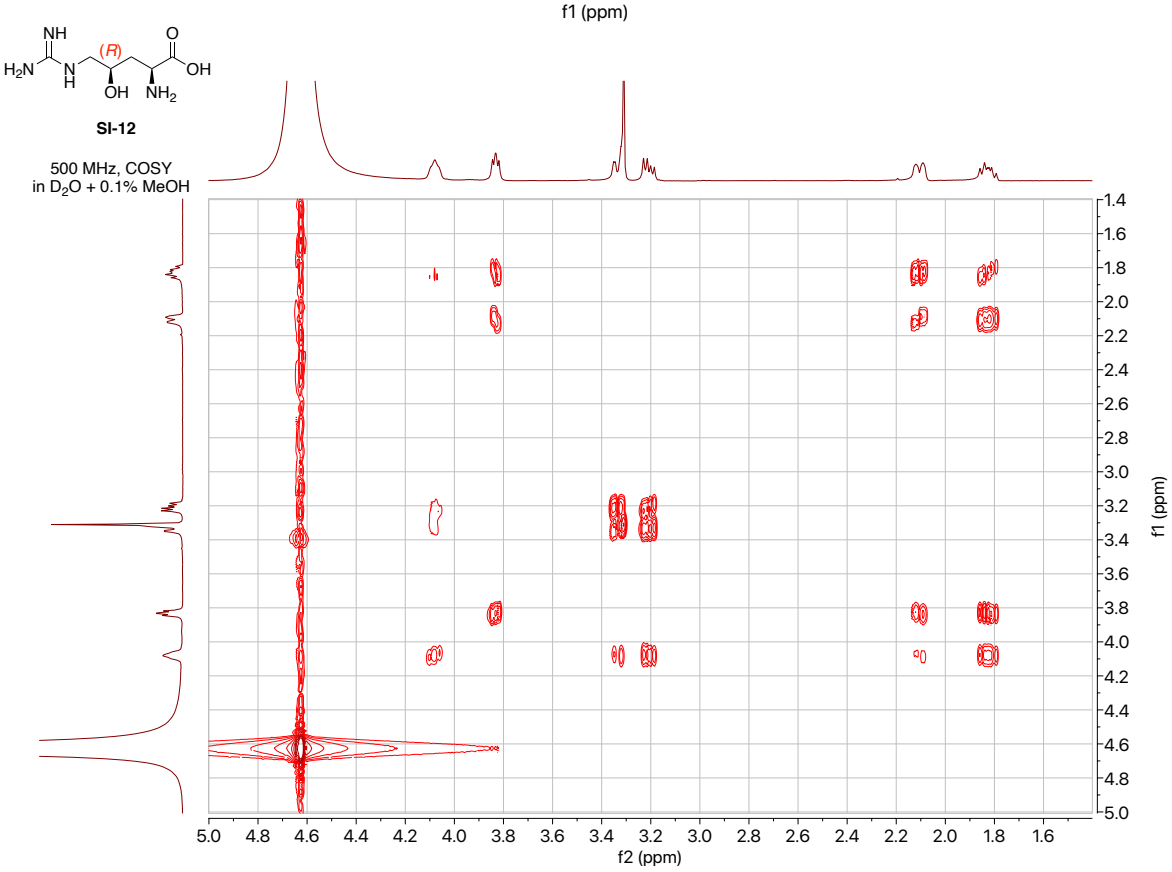
SI-12

500 MHz, ^{1}H
in D_2O + 0.1% MeOH



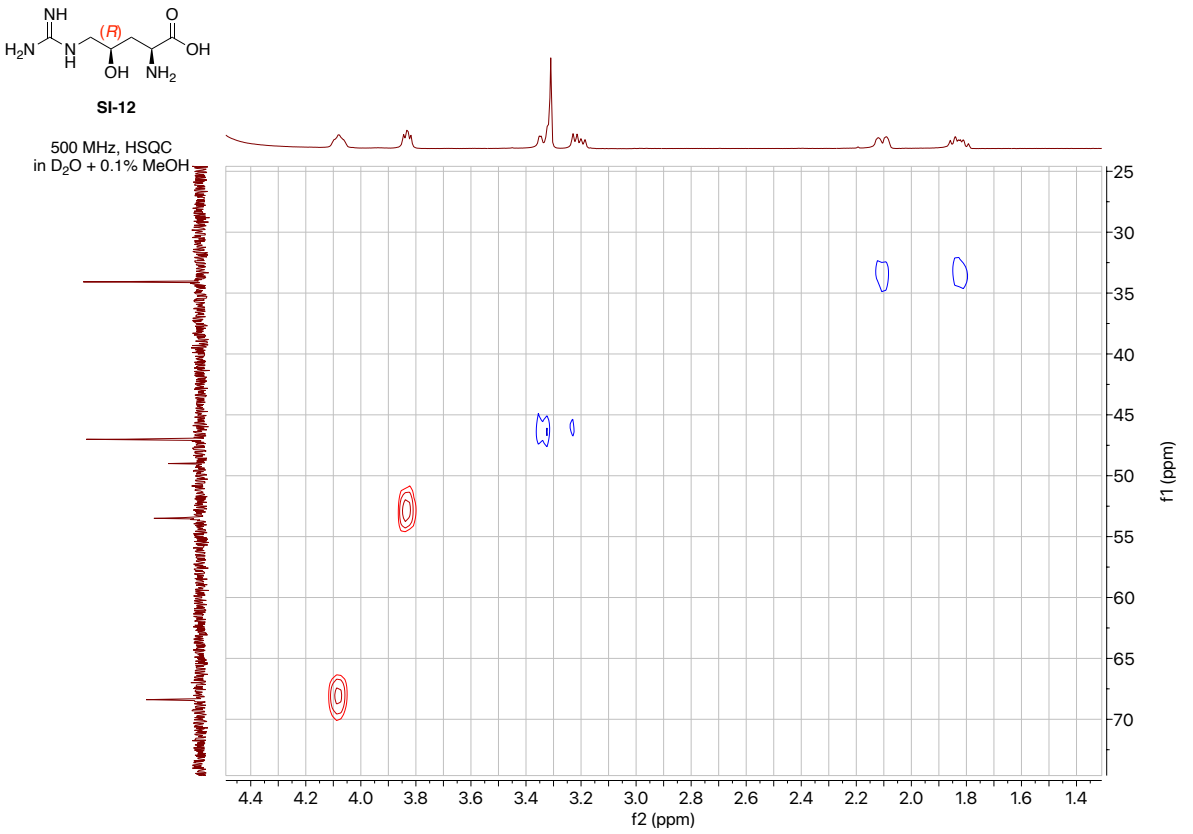
SI-12

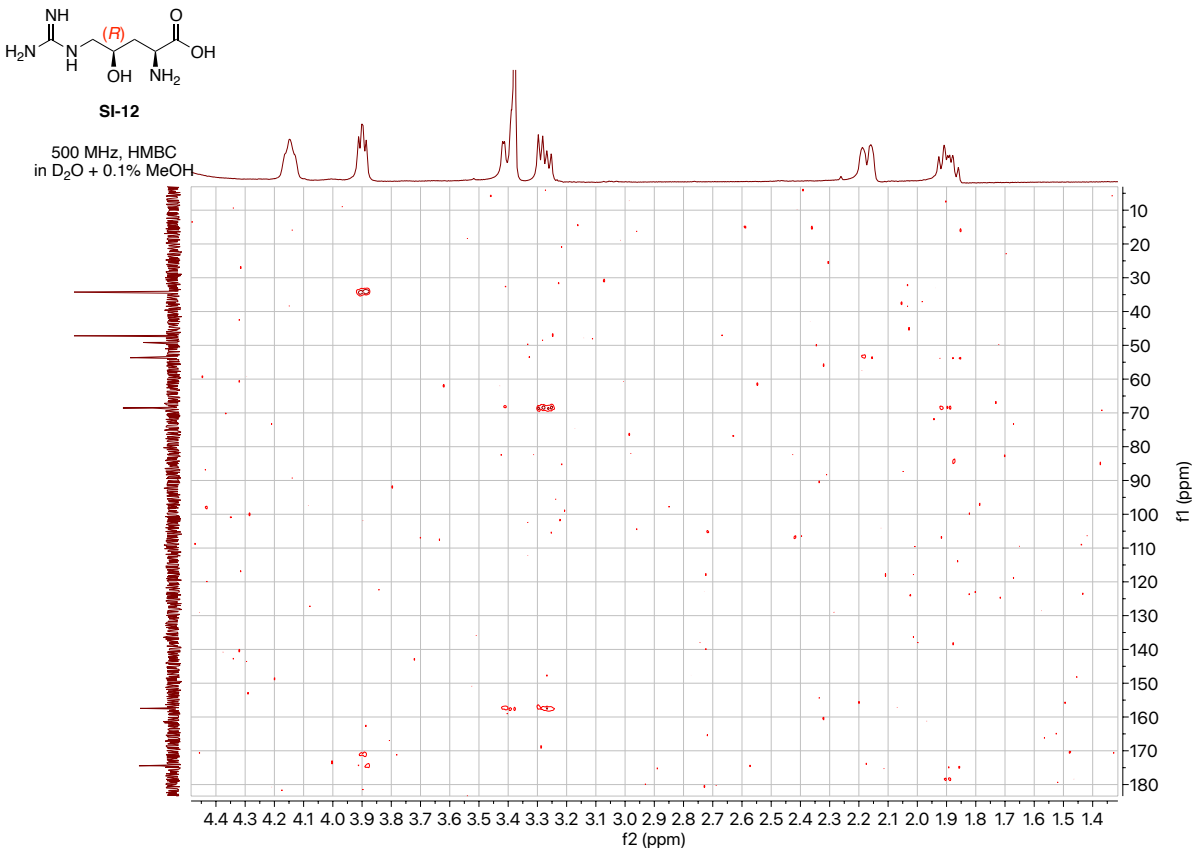
125 MHz, ^{13}C
in D_2O + 0.1% MeOH



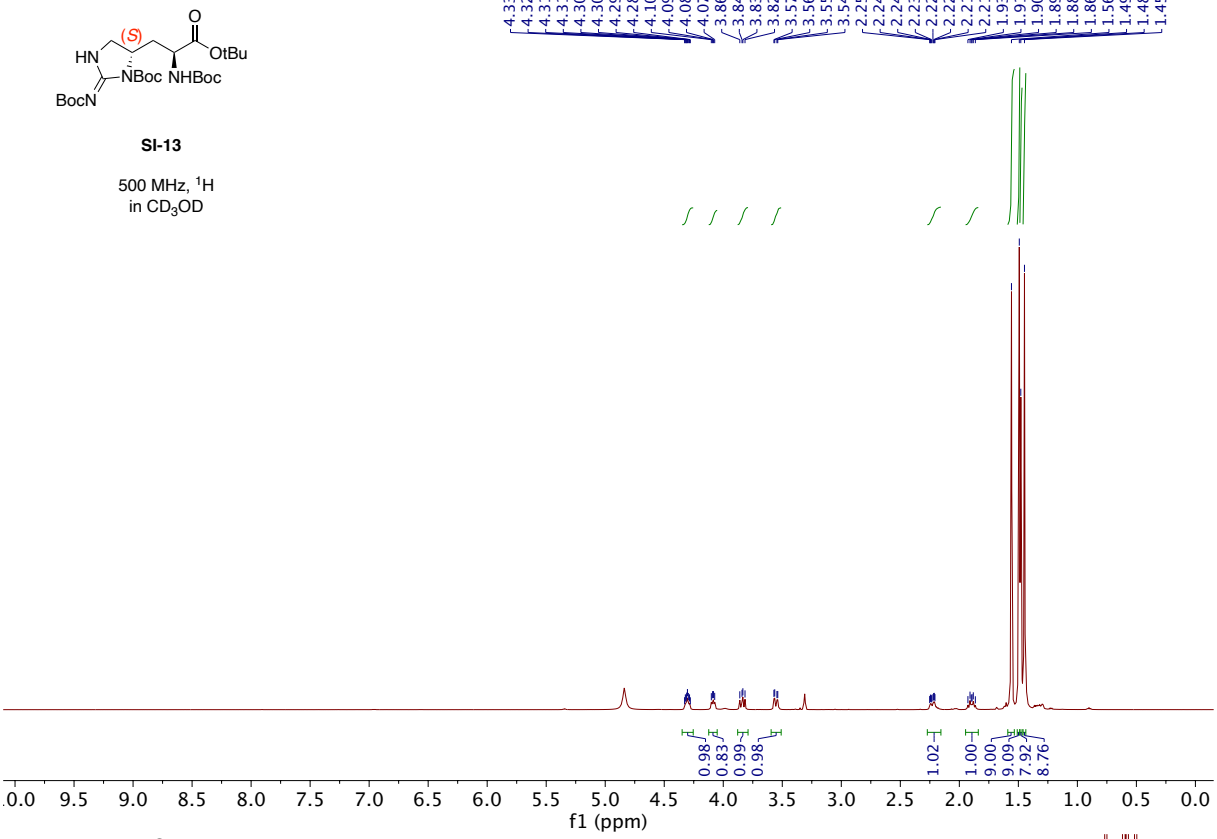
SI-12

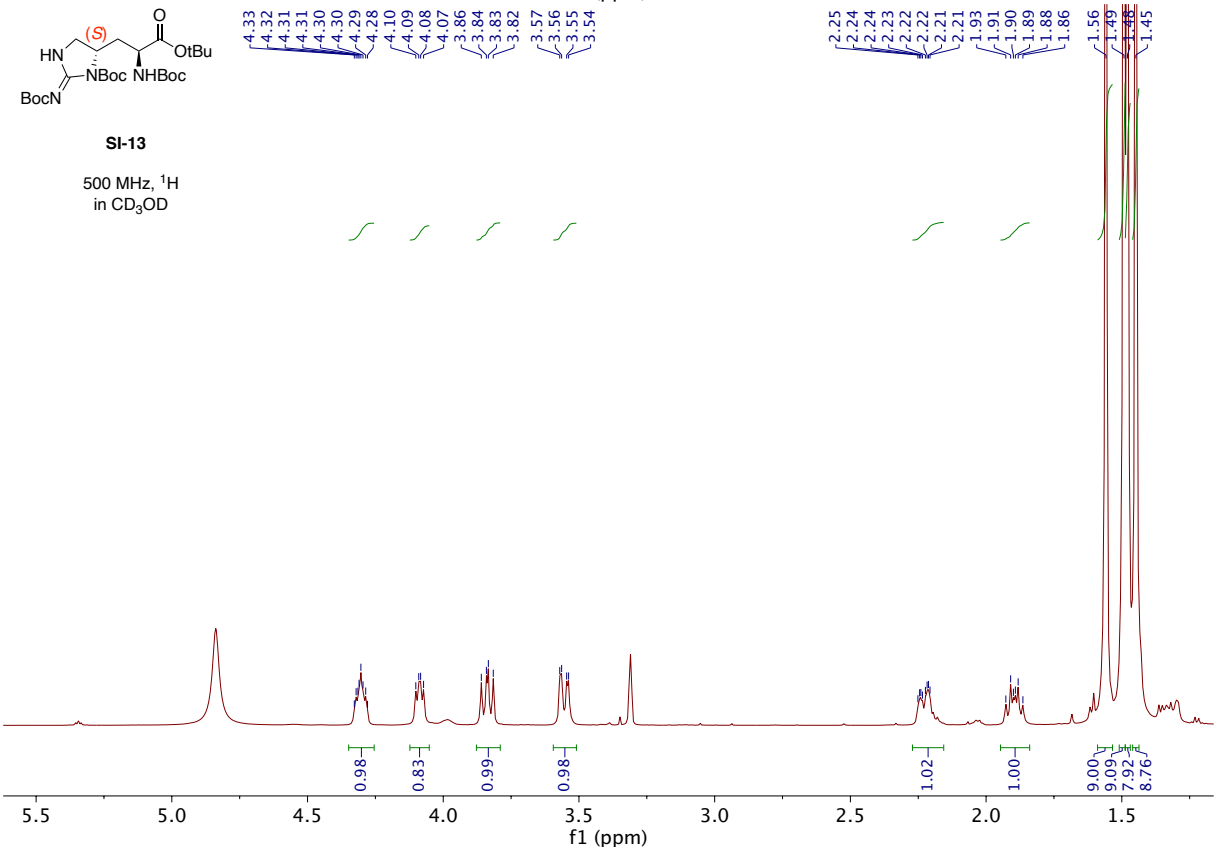
500 MHz, COSY
in D_2O + 0.1% MeOH

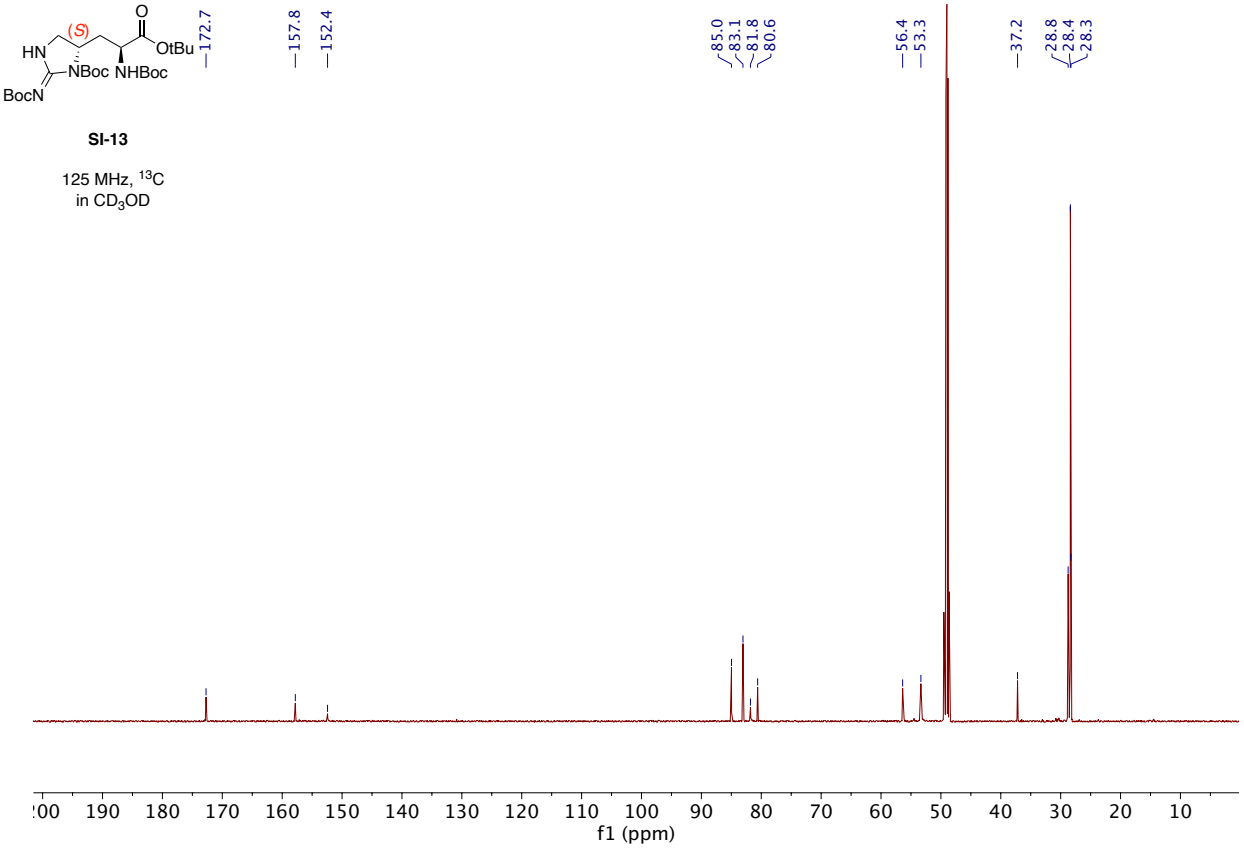

NH
O
(R)
H2N
N
H
OH
OH
NH2
SI-12
500 MHz, HSQC
in D2O + 0.1% MeOH
25
30
35
40
45
50
55
60
65
70
f1 (ppm)
4.4
4.2
4.0
3.8
3.6
3.4
3.2
3.0
2.8
2.6
2.4
2.2
2.0
1.8
1.6
1.4
f2 (ppm)


NH
O
(R)
H2N
N
H
OH
OH
NH2
SI-12
500 MHz, HMBC
in D2O + 0.1% MeOH
10
20
30
40
50
60
70
80
90
100
110
120
130
140
150
160
170
180
f1 (ppm)
4.4 4.3 4.2 4.1 4.0 3.9 3.8 3.7 3.6 3.5 3.4 3.3 3.2 3.1 3.0 2.9 2.8 2.7 2.6 2.5 2.4 2.3 2.2 2.1 2.0 1.9 1.8 1.7 1.6 1.5 1.4
f2 (ppm)

SI-13:

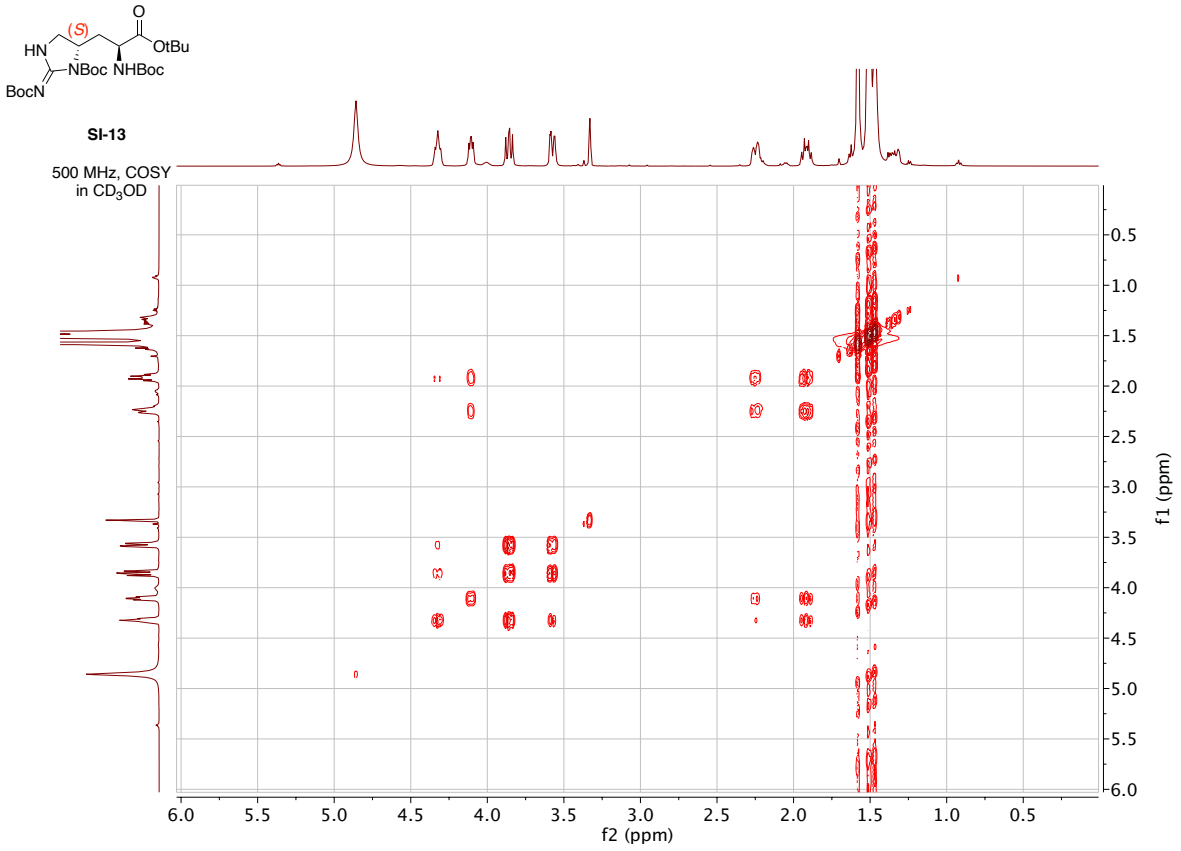






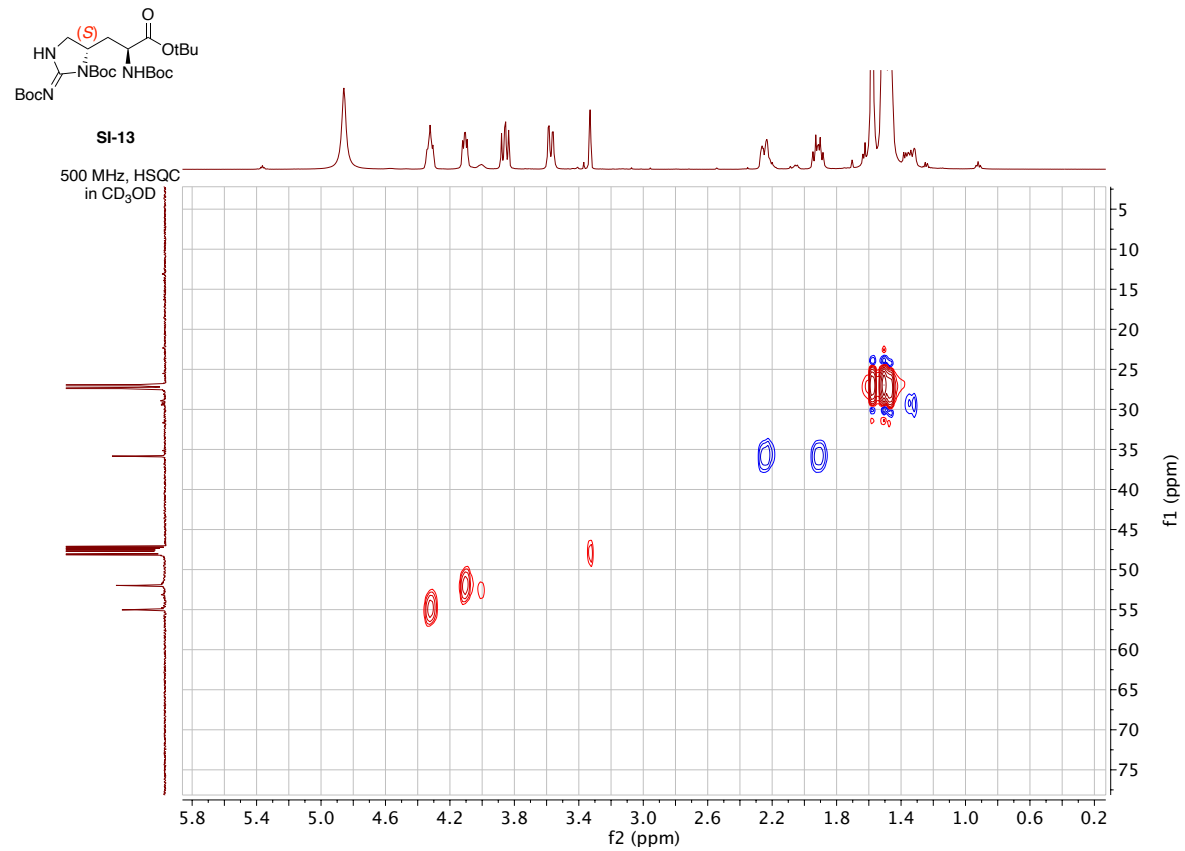
SI-13

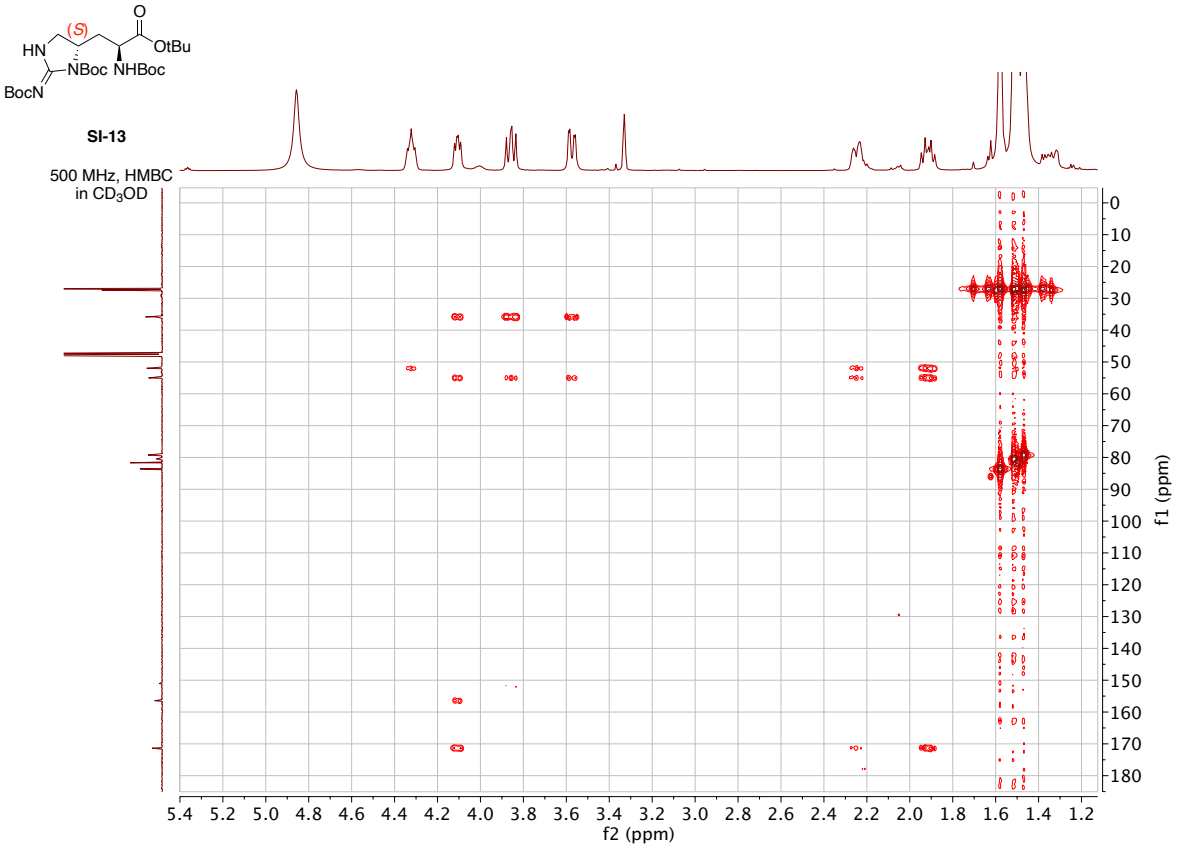
125 MHz, ^{13}C
in CD_3OD



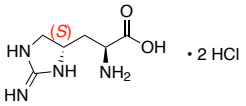
SI-13

500 MHz, COSY
in CD_3OD


SI-13
500 MHz, HSQC
in CD3OD
f2 (ppm)
f1 (ppm)

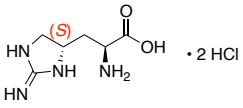

SI-13
500 MHz, HMBC
in CD3OD
f2 (ppm)
f1 (ppm)

SI-14:



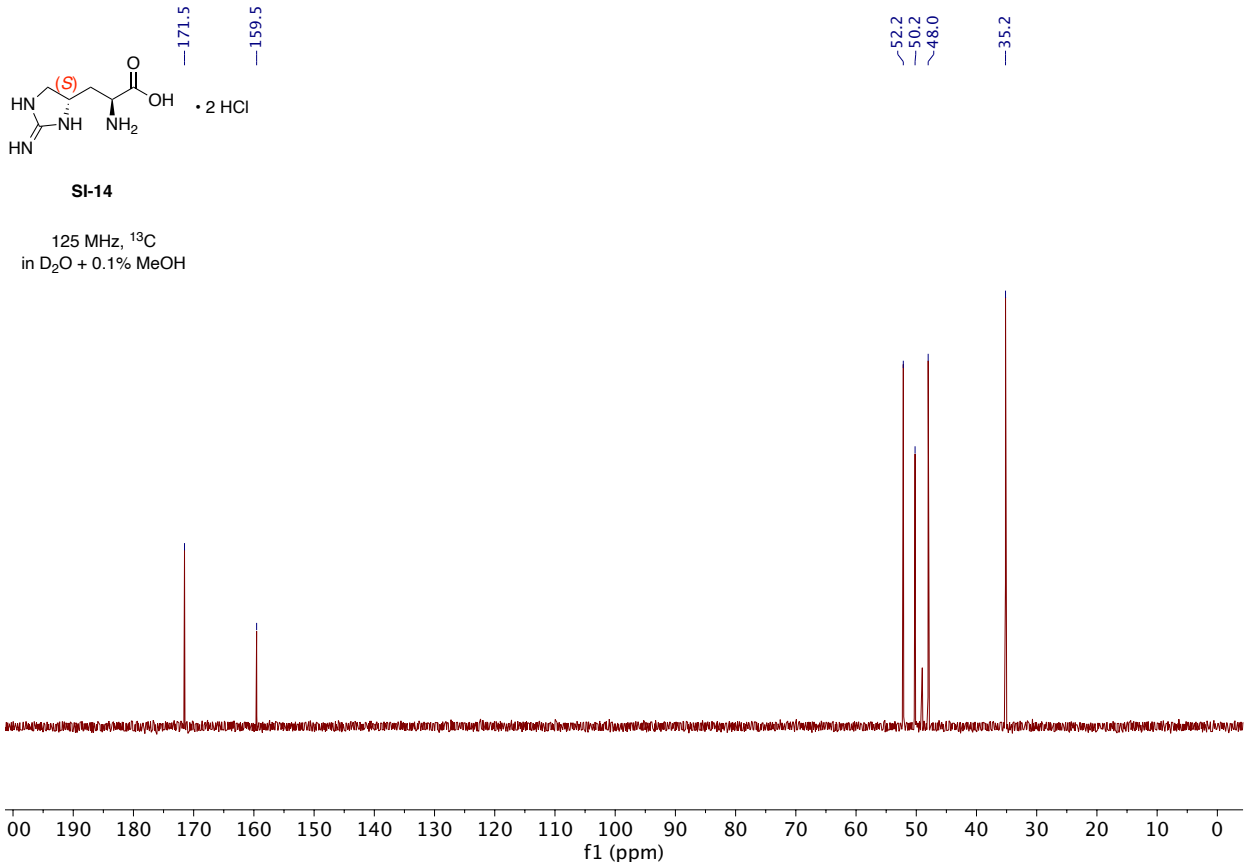
SI-14

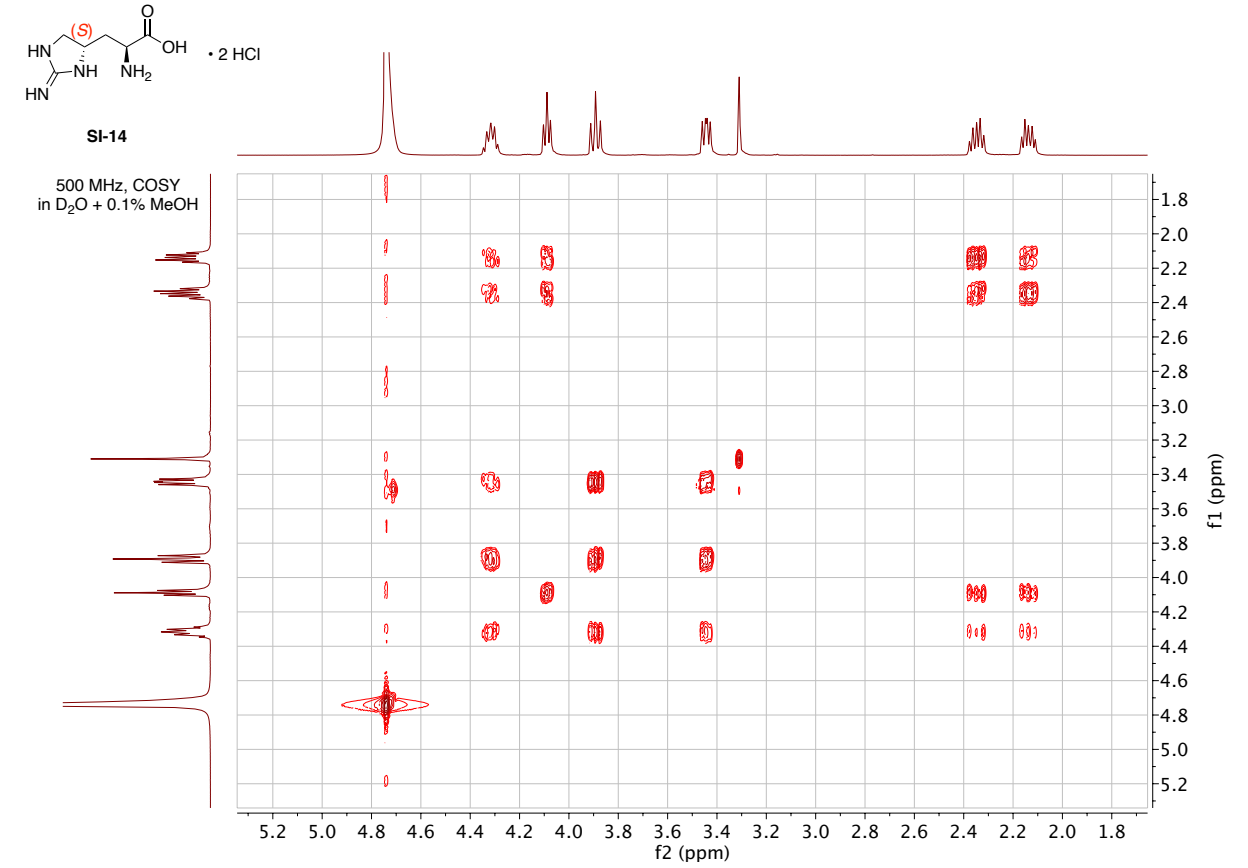
500 MHz, ^{1}H
in D_2O + 0.1% MeOH

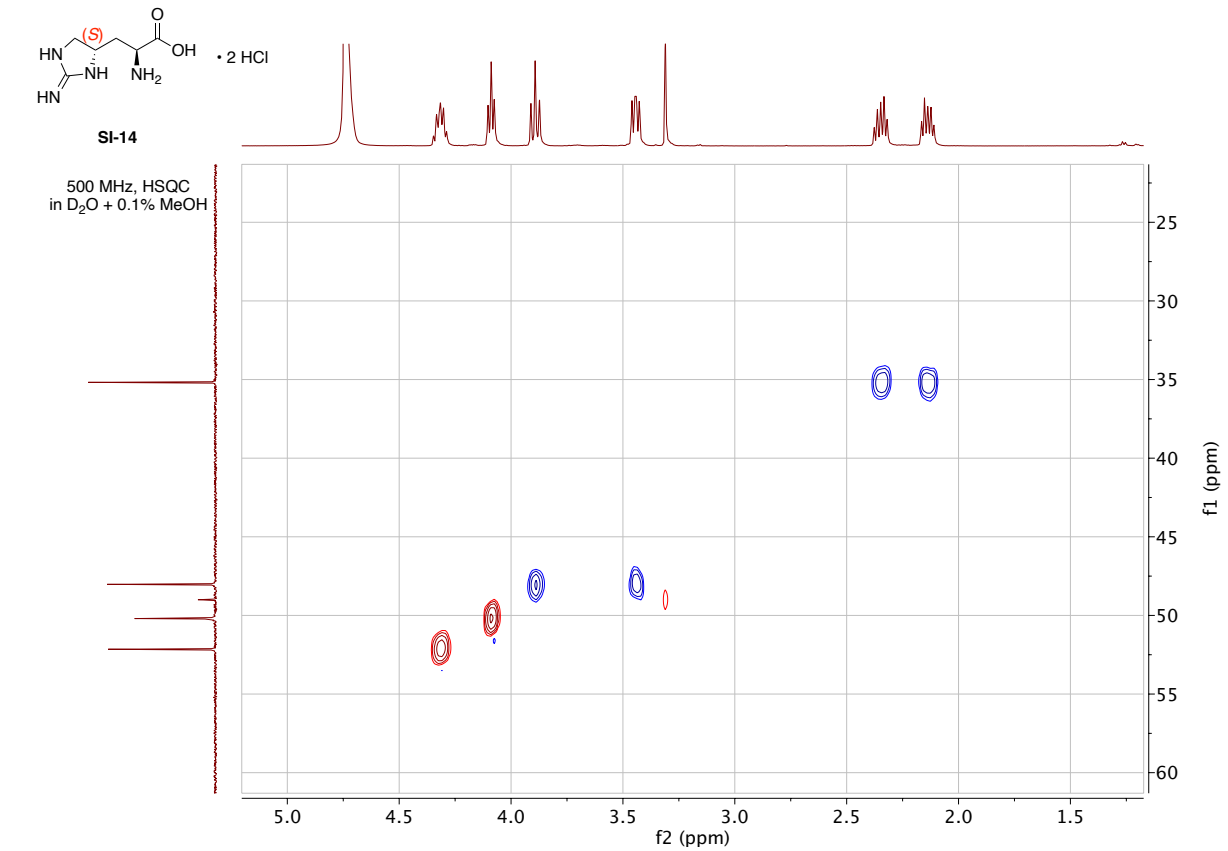


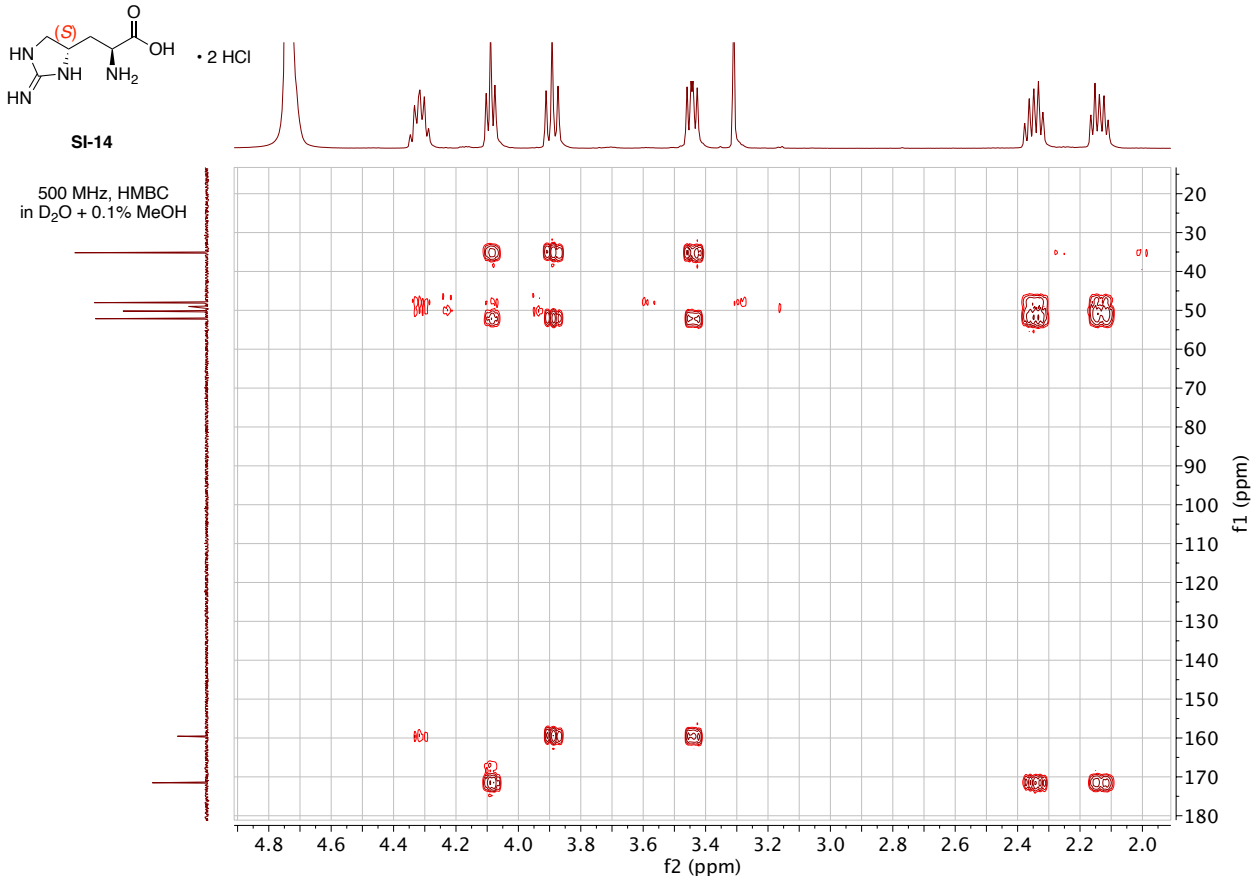
SI-14

500 MHz, ^{1}H
in D_2O + 0.1% MeOH

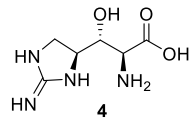

171.5
159.5
52.2
50.2
48.0
35.2
• 2 HCl
SI-14
125 MHz, 13C
in D2O + 0.1% MeOH
f1 (ppm)


• 2 HCl
SI-14
500 MHz, COSY
in D2O + 0.1% MeOH
f1 (ppm)
f2 (ppm)


(S)
HN
NH
HN
NH2
O
OH
• 2 HCl
SI-14
500 MHz, HSQC
in D2O + 0.1% MeOH
f1 (ppm)
f2 (ppm)


(S)
HN
NH
HN
NH2
O
OH
• 2 HCl
SI-14
500 MHz, HMBC
in D2O + 0.1% MeOH
f1 (ppm)
f2 (ppm)

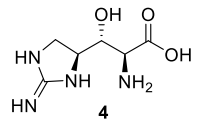
4 (enzymatic)**:**



4

(enzymatic)

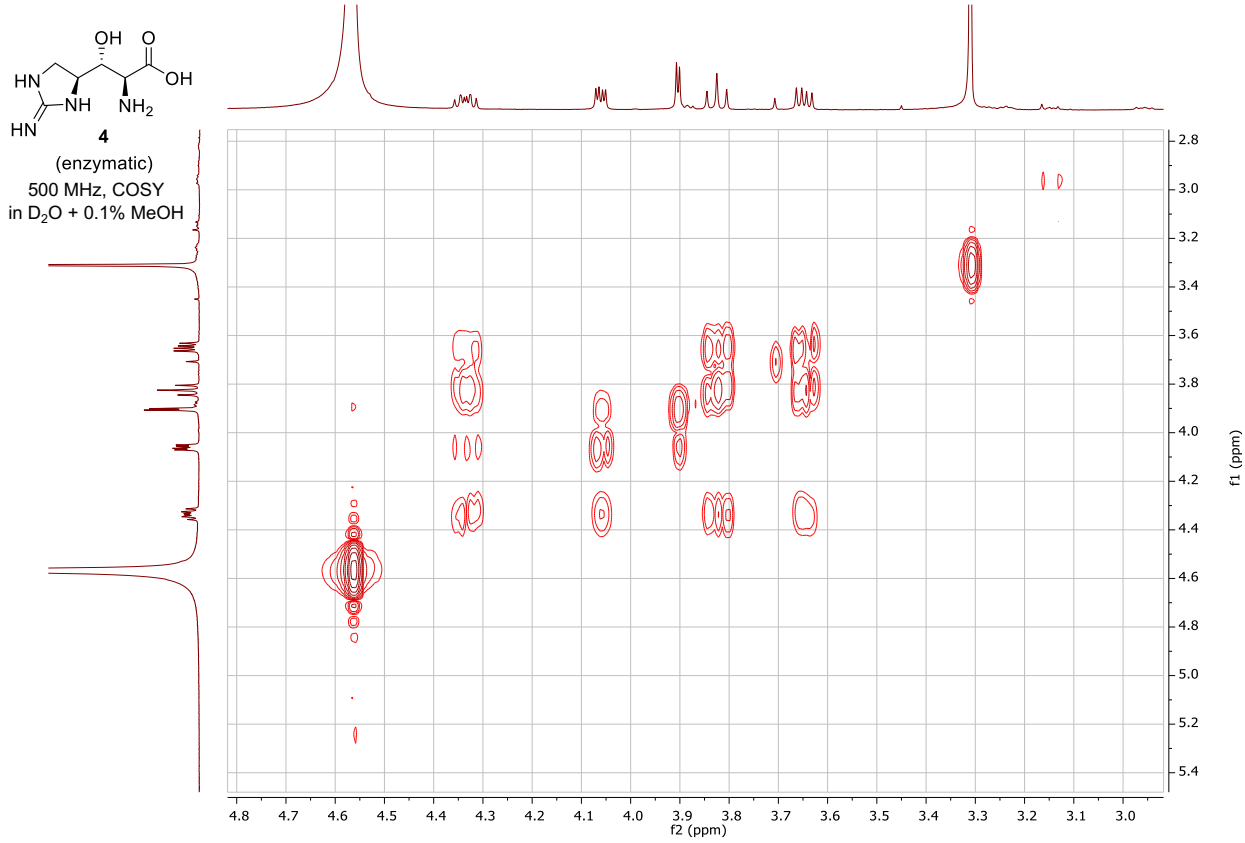
500 MHz, ^{1}H
in D_2O + 0.1% MeOH

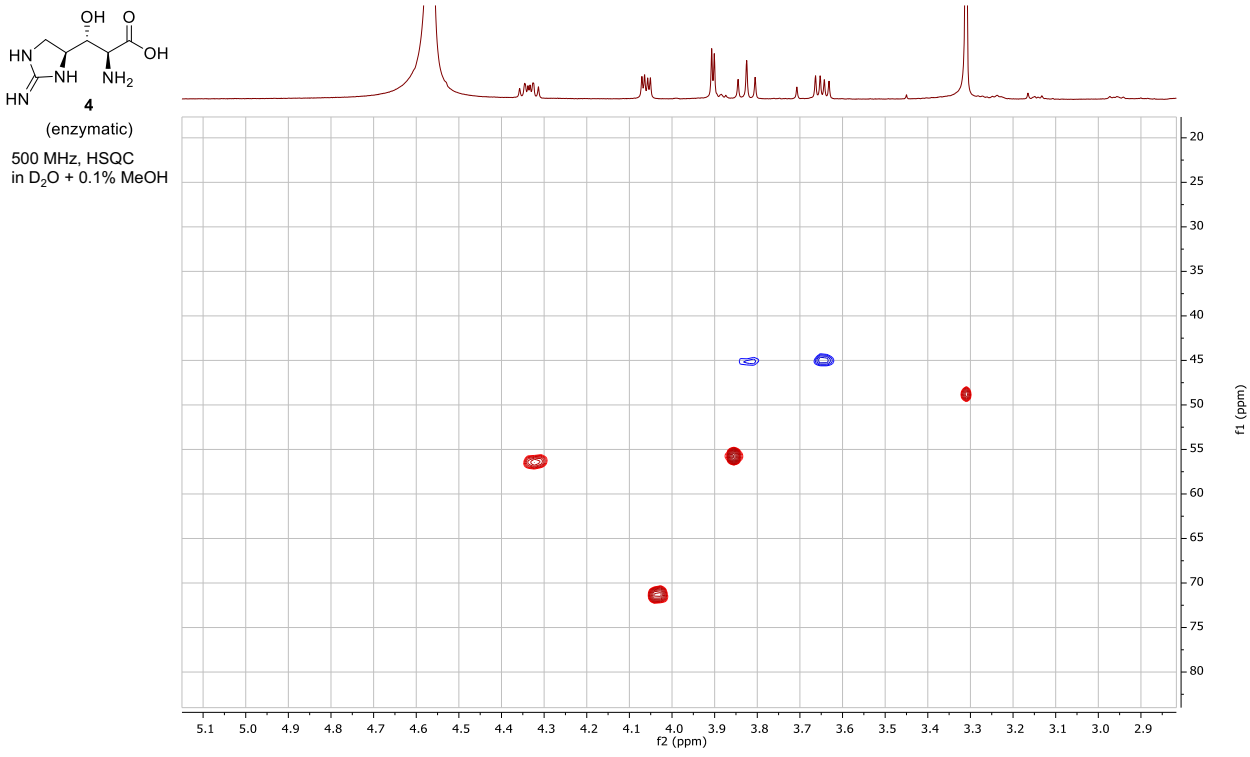


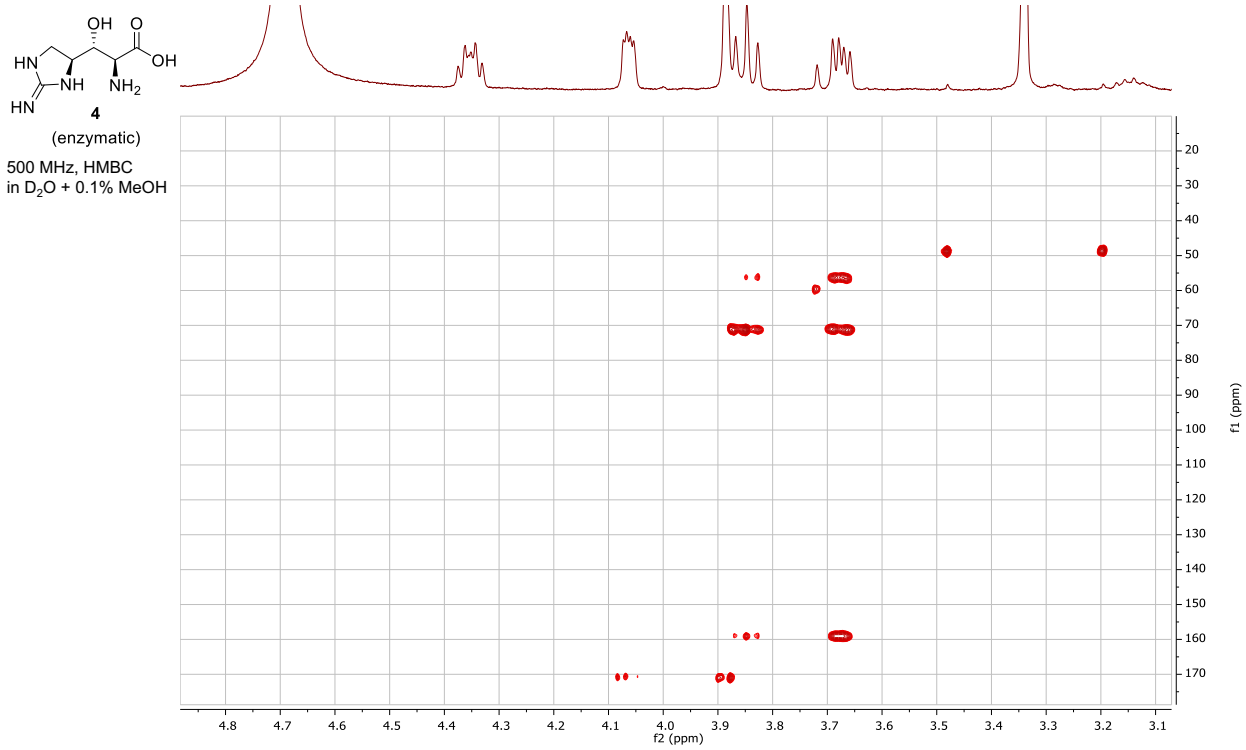
4

(enzymatic)

500 MHz, ^{1}H
in D_2O + 0.1% MeOH


OH
O
HN
OH
NH
NH2
HN
4
(enzymatic)
500 MHz, COSY
in D2O + 0.1% MeOH
f2 (ppm)
f1 (ppm)


OH
O
HN
OH
NH
NH2
HN
4
(enzymatic)
500 MHz, HSQC
in D2O + 0.1% MeOH
f2 (ppm)
f1 (ppm)


OH
O
HN
OH
NH
NH2
HN
4
(enzymatic)
500 MHz, HMBC
in D2O + 0.1% MeOH
f1 (ppm)
f2 (ppm)

Genomic coordinates	Region size	BGC type	Most similar to
CP080598.1: 29917-120960	91,043 bp	NRPS, betalactone, microviridin	**CP080598.1:**43683-105015, 100% alignment to prev. characterized KX788858.1: anabaenopeptin and spumigin BGCs (*1*).
CP080598.1: 286410-335840	49,430 bp	T1-PKS, hglE-KS	
CP080598.1: 875865-896726	20,861 bp	terpene	
CP080598.1: 1152826-1177857	25,031 bp	cyanobactin	anacyclamide D8P
CP080598.1: 1195859-1247974	52,115 bp	hglE-KS, T1-PKS	heterocyte glycolipids
CP080598.1: 1518085-1560328	42,243 bp	T1-PKS	
CP080598.1: 1983890-2012306	28,416 bp	terpene, oligosaccharide	
CP080598.1: 2335145-2376757	41,612 bp	lanthipeptide-class-V	
CP080598.1: 2990394-3011326	20,932 bp	terpene	
CP080598.1: 4204921-4247326	42,405 bp	lanthipeptide-class-V	
CP080598.1: 4412989-4486641	73,652 bp	NRPS	

Table S1.
antiSMASH annotation of the *Sphaerospermopsis torques-reginae* ITEP-024 genome. In total, 11 candidate BGC clusters were detected by antiSMASH. Notably, one location overlapped with the previously characterized anabaenopeptin and spumigin BGCs (*1*). T1-PKS = type-1 PKS, hglE-KS = heterocyte glycolipid synthase-like PKS

SRA dataset	# reads (AS >= 150)	Matched genes	geo_loc_name	env_biome	env_feature	lat_lon	collection_date
SRR5249014	1912	gntA,gntB,gntC,gntD,gntE,gntF,gntG,gntH,gntI,gntJ	USA: Sandusky Bay, Ohio	N/A	N/A	41.474889 N 82.854137 W	6/11/13
SRR5249015	1887	gntA,gntB,gntC,gntD,gntE,gntF,gntG,gntH,gntI,gntJ	USA: Sandusky Bay, Ohio	N/A	N/A	41.474889 N 82.854137 W	6/11/13
SRR1601415	1274	gntA,gntB,gntC,gntD,gntE,gntF,gntG,gntH,gntI,gntJ	USA	Lake Erie	Western Basin	41.7394 N 83.3750 W	8-Oct-13
SRR1601417	1188	gntA,gntB,gntC,gntD,gntE,gntF,gntG,gntH,gntI,gntJ	USA	Lake Erie	Western Basin	41.6989 N 83.4589 W	8-Oct-13
SRR1601416	938	gntA,gntB,gntC,gntD,gntE,gntF,gntG,gntH,gntI,gntJ	USA	Lake Erie	Western Basin	41.6989 N 83.4589 W	8-Oct-13
SRR5834679	160	gntA,gntB,gntC,gntD,gntE,gntF,gntG,gntH,gntI,gntJ	USA: Wisconsin	N/A	N/A	43.099 N 89.405 W	7/14/16
SRR6435700	148	gntA,gntC,gntD,gntE,gntF,gntG,gntH,gntI,gntJ	USA: Wisconsin	N/A	N/A	43.099 N 89.405 W	7/15/16
SRR5194976	111	gntC,gntD,gntE,gntF,gntG,gntH,gntI,gntJ	USA:Wisconsin,Dane,Madison,Lake Mendota	freshwater biome	freshwater lake	43.082834 N 89.409982 W	8/21/15
SRR6435855	100	gntA,gntB,gntC,gntD,gntE,gntF,gntG,gntH,gntJ	USA: Wisconsin	N/A	N/A	43.099 N 89.405 W	7/15/16
SRR6435865	94	gntA,gntC,gntD,gntE,gntF,gntG,gntH,gntI,gntJ	USA: Wisconsin	N/A	N/A	43.099 N 89.405 W	7/16/16
SRR5834683	78	gntC,gntD,gntE,gntF,gntG,gntH,gntI,gntJ	USA: Wisconsin	N/A	N/A	43.099 N 89.405 W	7/15/16
SRR5834616	73	gntA,gntB,gntC,gntD,gntE,gntF,gntG,gntH,gntI,gntJ	USA: Wisconsin	N/A	N/A	43.099 N 89.405 W	7/14/16
SRR5194978	72	gntA,gntB,gntC,gntD,gntE,gntF,gntG,gntH,gntI,gntJ	USA:Wisconsin,Dane,Madison,Lake Mendota	freshwater biome	freshwater lake	43.082834 N 89.409982 W	8/20/15
SRR5834684	48	gntA,gntC,gntD,gntE,gntF,gntG,gntH,gntI,gntJ	USA: Wisconsin	N/A	N/A	43.099 N 89.405 W	7/15/16
SRR5194977	47	gntA,gntC,gntD,gntE,gntF,gntG,gntH,gntI,gntJ	USA:Wisconsin,Dane,Madison,Lake Mendota	freshwater biome	freshwater lake	43.082834 N 89.409982 W	8/20/15
SRR6435857	46	gntA,gntC,gntD,gntE,gntF,gntG,gntJ	USA: Wisconsin	N/A	N/A	43.099 N 89.405 W	7/15/16
SRR6435866	44	gntA,gntC,gntD,gntE,gntF,gntG,gntH,gntJ	USA: Wisconsin	N/A	N/A	43.099 N 89.405 W	7/16/16
SRR5194979	40	gntB,gntC,gntE,gntF,gntG,gntI,gntJ	USA:Wisconsin,Dane,Madison,Lake Mendota	freshwater biome	freshwater lake	43.082834 N 89.409982 W	8/20/15

SRR5834613	34	gntA,gntB,gntC,gntD,gntE,gntG,gntH,gntI,gntJ	USA: Wisconsin	N/A	N/A	43.099 N 89.405 W	7/14/16
SRR5834391	34	gntC,gntD,gntE,gntF,gntG,gntH,gntJ	USA: Wisconsin	N/A	N/A	43.099 N 89.405 W	7/14/16
SRR5834678	34	gntC,gntD,gntE,gntF,gntG,gntI,gntJ	USA: Wisconsin	N/A	N/A	43.099 N 89.405 W	7/14/16
SRR6435860	27	gntE,gntF,gntG,gntH,gntJ	USA: Wisconsin	N/A	N/A	43.099 N 89.405 W	7/15/16
SRR5834691	26	gntA,gntC,gntD,gntE,gntF,gntG,gntH,gntI	USA: Wisconsin	N/A	N/A	43.099 N 89.405 W	7/15/16
SRR5834392	26	gntC,gntE,gntF,gntG,gntI	USA: Wisconsin	N/A	N/A	43.099 N 89.405 W	7/14/16
SRR5834609	25	gntC,gntE,gntG,gntH,gntJ	USA: Wisconsin	N/A	N/A	43.099 N 89.405 W	7/14/16
SRR6435858	25	gntC,gntE,gntF,gntI,gntJ	USA: Wisconsin	N/A	N/A	43.099 N 89.405 W	7/15/16
SRR6435861	24	gntA,gntB,gntC,gntD,gntE,gntF,gntG,gntJ	USA: Wisconsin	N/A	N/A	43.099 N 89.405 W	7/16/16
SRR5834493	24	gntC,gntE,gntF,gntG,gntJ	USA: Wisconsin	N/A	N/A	43.099 N 89.405 W	7/14/16
SRR5834393	22	gntC,gntD,gntE,gntF,gntG,gntI,gntJ	USA: Wisconsin	N/A	N/A	43.099 N 89.405 W	7/14/16
SRR5834688	19	gntC,gntD,gntE,gntF,gntG,gntI	USA: Wisconsin	N/A	N/A	43.099 N 89.405 W	7/15/16
SRR6823695	19	gntC,gntD,gntE,gntF,gntG,gntJ	USA: Madison	N/A	N/A	43.0990 N 89.4050 W	7/16/16
SRR5834395	17	gntE,gntG,gntH,gntI,gntJ	USA: Wisconsin	N/A	N/A	43.099 N 89.405 W	7/14/16
SRR6823701	16	gntA,gntC,gntD,gntE,gntF	USA: Madison	N/A	N/A	43.0990 N 89.4050 W	7/16/16
SRR5834682	16	gntA,gntC,gntD,gntE,gntF,gntG	USA: Wisconsin	N/A	N/A	43.099 N 89.405 W	7/14/16
SRR5834614	12	gntC,gntE,gntG,gntH	USA: Wisconsin	N/A	N/A	43.099 N 89.405 W	7/14/16
SRR1781811	12	gntE,gntF,gntI,gntJ	Brazil: Amazon River	river	Amazon River	1.519367 S 48.917950 W	13-May-11
SRR7990772	12	gntJ	USA: Washington	N/A	N/A	46.1840 N 123.1820 W	8/4/10
SRR5834394	12	gntC,gntD,gntE,gntF,gntI	USA: Wisconsin	N/A	N/A	43.099 N 89.405 W	7/14/16
SRR5834492	12	gntC,gntE,gntG,gntH,gntI,gntJ	USA: Wisconsin	N/A	N/A	43.099 N 89.405 W	7/14/16
SRR5834686	12	gntC,gntE,gntF,gntG,gntI,gntJ	USA: Wisconsin	N/A	N/A	43.099 N 89.405 W	7/15/16

SRR5834612	10	gntC,gntE,gntF,gntG,gntJ	USA: Wisconsin	N/A	N/A	43.099 N 89.405 W	7/14/16
SRR1781915	10	gntC,gntI	Brazil: Amazon River	river	Amazon River	1.519367 S 48.917950 W	13-May-11
SRR12458608	8	gntD,gntG	USA: Harsha Lake	N/A	N/A	39.0 N 84.1 W	8/12/15
SRR5834687	8	gntD,gntG,gntJ	USA: Wisconsin	N/A	N/A	43.099 N 89.405 W	7/15/16
SRR6823694	8	gntE,gntF,gntI	USA: Wisconsin	N/A	N/A	46.0410 N 89.6860 W	7/8/16
SRR12458606	8	gntA,gntD,gntJ	USA: Harsha Lake	N/A	N/A	39.0 N 84.1 W	8/19/15
SRR5834396	7	gntD,gntF,gntG,gntJ	USA: Wisconsin	N/A	N/A	43.099 N 89.405 W	7/14/16
SRR8269088	6	gntJ	USA: Oregon	N/A	N/A	46.234 N 123.909 W	8/1/10
SRR6823687	6	gntC,gntJ	USA: Madison	N/A	N/A	43.0990 N 89.4050 W	7/16/16
SRR5860223	6	gntB,gntC,gntJ	USA:Delaware River, USA	N/A	N/A	39.283 N 75.3633 W	3/19/14
SRR6823698	5	gntE,gntH	USA: Madison	N/A	N/A	43.0990 N 89.4050 W	7/16/16
SRR12458605	4	gntF,gntH	USA: Harsha Lake	N/A	N/A	39.0 N 84.1 W	9/2/15
SRR6823501	4	gntC,gntD,gntI	USA: Madison	N/A	N/A	43.0990 N 89.4050 W	7/14/16
SRR6466476	4	gntJ	USA: Ohio, Lake Erie	N/A	N/A	41.69957 N 83.29410 W	missing
SRR6048585	2	gntJ	USA: Ohio, Lake Erie	N/A	N/A	missing	missing
SRR12458604	2	gntJ	USA: Harsha Lake	N/A	N/A	39.0 N 84.1 W	9/2/15
SRR6823494	2	gntD	USA: Madison	N/A	N/A	43.0990 N 89.4050 W	7/15/16
SRR9599516	2	gntE	USA	N/A	N/A	39.0367 N 84.1381 W	6/25/16
SRR6987382	2	gntD,gntE	USA:Ohio	N/A	N/A	39.0367 N 84.1381 W	May-15
SRR1601412	2	gntA,gntJ	USA	Lake Erie	Western Basin	41.7667 N 83.3086 W	8-Oct-13
SRR9599519	2	gntE	USA	N/A	N/A	39.0367 N 84.1381 W	6/25/16
SRR6466487	2	gntC	USA: Ohio, Lake Erie	N/A	N/A	41.69957 N 83.29410 W	missing

SRR5834685	2	gntJ	USA: Wisconsin	N/A	N/A	43.099 N 89.405 W	7/15/16
SRR7084784	2	gntJ	USA: Oregon	N/A	N/A	46.2320 N 123.8890 W	8/21/07
SRR7085878	2	gntJ	USA: Oregon	N/A	N/A	46.2340 N 123.9090 W	8/1/10
SRR12458595	2	gntE	USA: Harsha Lake	N/A	N/A	39.0 N 84.1 W	6/25/15
SRR12458617	2	gntA	USA: Harsha Lake	N/A	N/A	39.0 N 84.1 W	7/13/15
SRR6987381	2	gntD,gntE	USA:Ohio	N/A	N/A	39.0367 N 84.1381 W	May-15
SRR5830085	2	gntC	USA: Chesapeake Bay	N/A	N/A	39.2637 N 76.0017 W	8/17/15
SRR12458607	2	gntJ	USA: Harsha Lake	N/A	N/A	39.0 N 84.1 W	8/19/15
SRR12458618	2	gntE	USA: Harsha Lake	N/A	N/A	39.0 N 84.1 W	6/30/15
SRR9599514	2	gntE	USA	N/A	N/A	39.0367 N 84.1381 W	6/25/16
SRR8438318	1	gntG	USA: Louisiana	N/A	N/A	29.8571 N 89.9778 W	7/12/16
SRR5209160	1	gntH	USA:Delaware River, USA	N/A	N/A	39.283 N 75.3633 W	missing
SRR6987346	1	gntH	USA:Ohio	N/A	N/A	39.0367 N 84.1381 W	May-15
SRR5834496	1	gntG	USA: Wisconsin	N/A	N/A	43.099 N 89.405 W	7/14/16
SRR6987351	1	gntH	USA:Ohio	N/A	N/A	39.0367 N 84.1381 W	May-15

Table S2.
Sequence Read Archive (SRA) metagenomic and metatranscriptomic datasets with reads that match the *gnt* BGC. AS=alignment score. geo_loc_name, env_biome, env_feature, lat_lon, collection_date, represents data shown in the NCBI SRA metadata field of identical name.

gntB-F	TTGTTTAACTTTAATAAGGAGATATACCATGAAGAAAAACATCAAGA AATACCGTTTC
gntB-R	GCATTATGCGGCCGCAAGCTTGTTAGCTGCTAACTTGGGTCAGA
gntC-F	GTTAAGTATAAGAAGGAGATATACATATGAAGATCCAGCCGGCGCTG
gntC-R	CCGATATCCAATTGAGATCTGCCATATGTTAGCTGGTTTCAATAACGA TCGCCAG
pCOLA-F	CAAGCTTGCGGCCGCATAATGC
pCOLA-R	CATGGTATATCTCCTTATTAAAGTTAAACAA

Table S3.
Primers used in this study.

References

1. S. T. Lima, D. O. Alvarenga, A. Etchegaray, D. P. Fewer, J. Jokela, A. M. Varani, M. Sanz, F. A. Dörr, E. Pinto, K. Sivonen, M. F. Fiore, Genetic organization of anabaenopeptin and spumigin biosynthetic gene clusters in the cyanobacterium *Sphaerospermopsis torques-reginae* ITEP-024. *ACS Chem. Biol.* **12**, 769–778 (2017).

2. P. R. Gorham, J. McLachlan, U. T. Hammer, W. K. Kim, Isolation and culture of toxic strains of *Anabaena flos-aquae* (Lyngb.) de Bréb. *Int. Ver. Für Theor. Angew. Limnol. Verhandlungen*. **15**, 796–804 (1964).

3. F. A. Dörr, V. Rodríguez, R. Molica, P. Henriksen, B. Krock, E. Pinto, Methods for detection of anatoxin-a(s) by liquid chromatography coupled to electrospray ionization-tandem mass spectrometry. *Toxicon*. **55**, 92–99 (2010).

4. K. Okonechnikov, A. Conesa, F. García-Alcalde, Qualimap 2: advanced multi-sample quality control for high-throughput sequencing data. *Bioinformatics*. **32**, 292–294 (2016).

5. B. Langmead, S. L. Salzberg, Fast gapped-read alignment with Bowtie 2. *Nat Methods*. **9**, 357–359 (2012).

6. W. De Coster, S. D'Hert, D. T. Schultz, M. Cruts, C. Van Broeckhoven, NanoPack: visualizing and processing long-read sequencing data. *Bioinformatics*. **34**, 2666–2669 (2018).

7. R. Schmieder, R. Edwards, Quality control and preprocessing of metagenomic datasets. *Bioinformatics*. **27**, 863–864 (2011).

8. S. Koren, B. P. Walenz, K. Berlin, J. R. Miller, N. H. Bergman, A. M. Phillippy, Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. *Genome Res.* **27**, 722–736 (2017).

9. E. Haghshenas, F. Hach, S. C. Sahinalp, C. Chauve, CoLoRMap: Correcting long reads by mapping short reads. *Bioinforma. Oxf. Engl.* **32**, i545–i551 (2016).

10. R. R. Wick, L. M. Judd, C. L. Gorrie, K. E. Holt, Unicycler: Resolving bacterial genome assemblies from short and long sequencing reads. *PLOS Comput. Biol.* **13**, e1005595 (2017).

11. A. V. Zimin, D. Puiu, M.-C. Luo, T. Zhu, S. Koren, G. Marcais, J. A. Yorke, J. Dvorak, S. L. Salzberg, Hybrid assembly of the large and highly repetitive genome of *Aegilops tauschii*, a progenitor of bread wheat, with the MaSuRCA mega-reads algorithm. *Genome Res.*, **27**, 787-792 (2017).

12. M. Boetzer, C. V. Henkel, H. J. Jansen, D. Butler, W. Pirovano, Scaffolding pre-assembled contigs using SSPACE. *Bioinforma. Oxf. Engl.* **27**, 578–579 (2011).

13. B. J. Walker, T. Abeel, T. Shea, M. Priest, A. Abouelliel, S. Sakthikumar, C. A. Cuomo, Q. Zeng, J. Wortman, S. K. Young, A. M. Earl, Pilon: An integrated tool for comprehensive

microbial variant detection and genome assembly improvement. *PLOS ONE*. **9**, e112963 (2014).

14. M. Boetzer, W. Pirovano, Toward almost closed genomes with GapFiller. *Genome Biol.* **13**, R56 (2012).

15. M. Kolmogorov, J. Yuan, Y. Lin, P. A. Pevzner, Assembly of long, error-prone reads using repeat graphs. *Nat. Biotechnol.* **37**, 540–546 (2019).

16. A. Gurevich, V. Saveliev, N. Vyahhi, G. Tesler, QUAST: quality assessment tool for genome assemblies. *Bioinformatics*. **29**, 1072–1075 (2013).

17. D. H. Parks, M. Imelfort, C. T. Skennerton, P. Hugenholtz, G. W. Tyson, CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Res.* **25**, 1043–1055 (2015).

18. M. Hunt, N. D. Silva, T. D. Otto, J. Parkhill, J. A. Keane, S. R. Harris, Circlator: automated circularization of genome assemblies using long sequencing reads. *Genome Biol.* **16**, 294 (2015).

19. T. Seemann, Prokka: rapid prokaryotic genome annotation. *Bioinformatics*. **30**, 2068–2069 (2014).

20. K. Levi, M. Rynge, E. Abeysinghe, R. A. Edwards, in *Proceedings of the Practice and Experience on Advanced Research Computing* (ACM, Pittsburgh PA USA, 2018; https://dl.acm.org/doi/10.1145/3219104.3229278), pp. 1–7.

21. P. J. Torres, R. A. Edwards, K. A. McNair, PARTIE: a partition engine to separate metagenomic and amplicon projects in the Sequence Read Archive. *Bioinformatics*. **33**, 2389–2391 (2017).

22. J. Towns, T. Cockerill, M. Dahan, I. Foster, K. Gaither, A. Grimshaw, V. Hazlewood, S. Lathrop, D. Lifka, G. D. Peterson, R. Roskies, J. R. Scott, N. Wilkins-Diehr, XSEDE: Accelerating scientific discovery. *Comput. Sci. Eng.* **16**, 62–74 (2014).

23. C. A. Stewart, T. M. Cockerill, I. Foster, D. Hancock, N. Merchant, E. Skidmore, D. Stanzione, J. Taylor, S. Tuecke, G. Turner, M. Vaughn, N. I. Gaffney, in *Proceedings of the 2015 XSEDE Conference: Scientific Advancements Enabled by Enhanced Cyberinfrastructure* (ACM, 2015; https://dl.acm.org/citation.cfm?doid=2792745.2792774), p. 29.

24. B. Buchfink, C. Xie, D. H. Huson, Fast and sensitive protein alignment using DIAMOND. *Nat. Methods*. **12**, 59–60 (2015).

25. P. Di Tommaso, M. Chatzou, E. W. Floden, P. P. Barja, E. Palumbo, C. Notredame, Nextflow enables reproducible computational workflows. *Nat. Biotechnol.* **35**, 316–319 (2017).

26. T. Kluyver, B. Ragan-Kelley, F. Pérez, B. Granger, M. Bussonnier, J. Frederic, K. Kelley, J. Hamrick, J. Grout, S. Corlay, P. Ivanov, D. Avila, S. Abdalla, C. Willing, J. development team, in *Positioning and Power in Academic Publishing: Players, Agents and Agendas*, F. Loizides, B. Scmidt, Eds. (IOS Press, 2016; https://eprints.soton.ac.uk/403913/), pp. 87–90.

27. The pandas development team, *pandas-dev/pandas: Pandas 1.0.3* (Zenodo, 2020; https://zenodo.org/record/3715232).

28. W. McKinney, in *Proceedings of the 9th Python in Science Conference*, S. van der Walt, J. Millman, Eds. (2010), pp. 56–61.

29. S. Choudhary, pysradb: A Python package to query next-generation sequencing metadata and data from NCBI Sequence Read Archive (2019), , doi:10.12688/f1000research.18676.1.

30. B. Grüning, R. Dale, A. Sjödin, B. A. Chapman, J. Rowe, C. H. Tomkins-Tinch, R. Valieris, J. Köster, Bioconda: sustainable and comprehensive software distribution for the life sciences. *Nat. Methods*. **15**, 475–476 (2018).

31. E. Bushmanova, D. Antipov, A. Lapidus, A. D. Przhibelskiy, rnaSPAdes: a de novo transcriptome assembler and its application to RNA-Seq data. *GigaScience*. **8**, giz100 (2019).

32. G. M. Kurtzer, V. Sochat, M. W. Bauer, Singularity: Scientific containers for mobility of compute. *PLOS ONE*. **12**, e0177459 (2017).

33. C. Camacho, G. Coulouris, V. Avagyan, N. Ma, J. Papadopoulos, K. Bealer, T. L. Madden, BLAST+: architecture and applications. *BMC Bioinformatics*. **10**, 421 (2009).

34. M. Martin, Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal*. **17**, 10–12 (2011).

35. S. Nurk, D. Meleshko, A. Korobeynikov, P. A. Pevzner, metaSPAdes: a new versatile metagenomic assembler. *Genome Res.* **27**, 824–834 (2017).

36. D. D. Kang, J. Froula, R. Egan, Z. Wang, MetaBAT, an efficient tool for accurately reconstructing single genomes from complex microbial communities. *PeerJ*. **3**, e1165 (2015).

37. P.-A. Chaumeil, A. J. Mussig, P. Hugenholtz, D. H. Parks, GTDB-Tk: a toolkit to classify genomes with the Genome Taxonomy Database. *Bioinforma. Oxf. Engl.*, btz848 (2019).

38. D. H. Parks, M. Chuvochina, D. W. Waite, C. Rinke, A. Skarshewski, P.-A. Chaumeil, P. Hugenholtz, A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life. *Nat. Biotechnol.* **36**, 996–1004 (2018).

39. M. N. Price, P. S. Dehal, A. P. Arkin, FastTree 2 – Approximately maximum-likelihood trees for large alignments. *PLOS ONE*. **5**, e9490 (2010).

40. I. Letunic, P. Bork, Interactive Tree Of Life (iTOL) v4: recent updates and new developments. *Nucleic Acids Res.* **47**, W256–W259 (2019).

41. K. Clark, I. Karsch-Mizrachi, D. J. Lipman, J. Ostell, E. W. Sayers, GenBank. *Nucleic Acids Res.* **44**, D67–D72 (2016).

42. I. Lee, Y. Ouk Kim, S.-C. Park, J. Chun, OrthoANI: An improved algorithm and software for calculating average nucleotide identity. *Int. J. Syst. Evol. Microbiol.* **66**, 1100–1103 (2016).

43. R. J. R. Molica, E. J. A. Oliveira, P. V. V. C. Carvalho, A. N. S. F. Costa, M. C. C. Cunha, G. L. Melo, S. M. F. O. Azevedo, Occurrence of saxitoxins and an anatoxin-a(s)-like anticholinesterase in a Brazilian drinking water supply. *Harmful Algae*. **4**, 743–753 (2005).

44. A. M. Giltrap, L. J. Dowman, G. Nagalingam, J. L. Ochoa, R. G. Linington, W. J. Britton, R. J. Payne, Total synthesis of teixobactin. *Org. Lett.* **18**, 2788–2791 (2016).

45. S. Matsunaga, R. E. Moore, W. P. Niemczura, W. W. Carmichael, Anatoxin-a(s), a potent anticholinesterase from *Anabaena flos-aquae*. *J. Am. Chem. Soc.* **111**, 8021–8023 (1989).