Genome and Transcriptome Mining of Macrocyclic Plant Peptide Natural Products for Multi-*omics* guided Drug Discovery

by

Derrick Ousley

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Medicinal Chemistry)
in the University of Michigan
2026

Doctoral Committee:

Assistant Professor Roland D. Kersten, Chair
Associate Professor Marcy Balunas
Assistant Professor Matthew O'Meara
Associate Professor Ashootosh Tripathi

Derrick Ousley

dousley@umich.edu

ORCID ID: 0000-0003-0770-5874

© Derrick Ousley 2026

Dedication

I dedicate this dissertation to my family, especially my mother, father, and grandmother. Thank you for continually supporting me, regardless of whatever adventure life has taken me through. From monetary support to making sure I did well physically, academically, and socially, to college visits and casual phone calls when we are far away, you always made sure that I knew that I was capable of whatever I put my mind to, that I am the master of my own destiny, and to always seek knowledge for myself. Thank you for reminding me to be creative, humble, definite, imaginative, practical, kind, and helpful to others in need. To my nephew, I hope I am a good role model for you. To all my friends, thank you for your support, confidence, and courage in helping me achieve my goals. From late-night gaming to pickup basketball and even listening to me rant about life and stocks, you all always lent your kind words of wisdom. To the next generation of scientists and leaders, don't ever stop pursuing your goals, even when the waters are rough, and I hope to be living proof that you can do it too. Finally, to anyone who reads this, I hope you always keep a long-term mindset over short-term headwinds, keep your armor on, and finger to the sky.

Acknowledgements

Firstly, I would like to acknowledge the funding sources that enabled the research in this dissertation. Upon entering graduate school in the medicinal chemistry program I was granted the Rackham Merit Fellowship for the Rackham Graduate School and the Eugene and Shirley Cordes Scholarship from the College of Pharmacy, then later in my graduate tenure I received the NIH T32 Pharmacological Sciences Training Program from the Department of Pharmacology all of which combined allowed me to continually focus on producing high quality scholarly work without interruption. I am deeply grateful to the friends, family, colleagues, and acquaintances I have gained throughout my life who have supported my journey through graduate school. To my research advisor, Dr. Roland Kersten, thank you for adapting your mentorship style to accommodate my work ethic and for providing the independence to pursue my own research aspirations. I believe that, especially in our last 2 ½ years together, we really pieced together high-quality projects that accommodated both of our interests, leading to publications of similarly high impact. To my committee members, Dr. Marcy Balunas, Dr. Matthew O'Meara, and Dr. Ashu Tripathi, I appreciate your professional opinions, support on data analysis, willingness to write letters of recommendation/feedback, and flexibility in getting things scheduled over the last few years, which allowed for healthy scientific growth with reduced imposter syndrome. To Dr. Cherie Dotson, thank you for always looking out for me and supporting me both when you were at the College of Pharmacy and during your time at Rackham with the RMF program. To Dr. Desnor Chigumba, Dr. Lisa Mydy, Dr. Jordan Hungerford, Dr. Khadija Shafiq, Dr. Wenjie Li, and Dr. Xiaofeng Wang for helping me develop core fundamental techniques in natural product chemistry and, conversely, allowing me to teach you all as I learned, reinforcing my learning and building confidence as an independent scientist. To Gabi, Bassma, and future Kersten Lab members, I wish you the best in your future endeavors and hope that you are successful in your undergraduate/graduate/postdoctoral/postgraduate careers. To Kevin Williams, Matthew Francis, and my GSSA/P&G colleagues, thanks for giving me the opportunity to work at one of the largest CPG companies in the world, allowing me to demonstrate my skillset in a unique and accommodating way at the industry level, helping me develop initiative to create industry

partnerships and networks, and just overall making it a very welcoming experience. To all the interns I bonded with during that time, thanks for the conversations and for exploring Cincinnati City; shout-out to the undergrads for including me in your outings and hoop sessions, even though I was “unc status”. To all the friends I have made over the years playing basketball at NCRB, that’s for recreational gameplay. To Dr. Lori Isom and Dr. Yoichi Osawa, thanks for always being willing to write letters of recommendation and provide words of encouragement; and to the rest of the PSTP community, thanks for allowing me to showcase my leadership skills to help collaboratively execute two awesome symposia for the Michigan life sciences community. To Professor Gordon, thanks for being my unofficial mentor in investments, for always encouraging me to find ways to pair my passions for investing and the sciences, and for providing an elite environment to understand venture capital. To my friends I met through WVF, thanks for your lightheartedness. To all the friends I made at the community life center food pantry, thank you for allowing me to assist in your mission to alleviate hunger in the Washtenaw community over the last few years.

Table of Contents

Dedication ii

Acknowledgements iii

Table of Contents v

List of Tables x

List of Figures xiii

Abstract xx

Chapter 1 Introduction 1

1.1 Perspectives of Plant Secondary Metabolites in Drug Discovery 1

1.1.1 Traditional Plant Natural Product Discovery 3
1.1.2 Challenges in Traditional Plant Natural Product Discovery 4
1.1.3 Advancements in Multi-omics and Computational Discovery of Plant Natural Products 5

1.2 Plant Genome Mining of Burpitides from Fused Precursor Peptides 7

1.2.1 Ribosomally synthesized and post-translationally modified peptides (RiPPs) in Plants 8
1.2.1.1 Introduction to Burpitides 14
1.2.1.2 Plant Genome Mining via RepeatFinder 16
1.2.2 Methods and Key Resources 17
1.2.2.1 Bioinformatics 17
1.2.2.2 Plant Cultivation – Nicotiana benthamiana 17
1.2.2.3 Transient peptide expression in N. benthamiana 17
1.2.2.4 Peptide chemotyping 18
1.2.2.5 Peptide purification 18
1.2.2.6 Peptide structure elucidation 19
1.2.3 Genome mining of fused burpitide cyclases 19
1.2.3.1 Identification of candidate burpitide cyclases from plant genomes 19
1.2.3.2 RepeatFinder analysis of candidate burpitide cyclases 20

1.2.4 Detection of Predicted Burpitides 22
1.2.4.1 Detection of an Analyte Matching a Predicted Burpitide Core Peptide 23
1.2.4.2 Biosynthetic Confirmation of the Predicted Burpitide Precursor Peptide 27
1.2.4.3 Characterization of a Plant Source for Burpitide Isolation 34
1.2.5 Characterization of Burpitide Cyclase Product 34
1.2.5.1 Scaled N. benthamiana expression 35
1.2.5.2 Purification 35
1.2.5.3 NMR 39
1.2.5.4 Stereochemical analysis 42
1.3 Dissertation Overview 45

1.4 References 47

Chapter 2 Benchmarking of *de novo* Assemblers for Large-scale Plant Transcriptome Mining ..52

2.1 Statement of contribution 52

2.2 Summary 52

2.3 Introduction 52

2.4 Results 55

2.4.1 RNA-seq Assembly Benchmarking for Discovery of Plant RiPP Precursor Peptides 55
2.4.2 Large-scale Transcriptome Mining of Fused Moroidin Cyclases and Burpitides 58
2.5 Discussion 59

2.6 Methods and Materials 60

2.6.1 RNA-Seq data download and trimming 60
2.6.2 Unassembled RNA-seq data search for stephanotic acid core peptides ... 60
2.6.3 Transcriptome assembly 61
2.6.4 Transcriptome analysis 62
2.6.5 Data Availability 63
2.7 Supplemental Information 63

2.8 Reference 69

Chapter 3 *In Silico* Burpitide Library Generation for Computational Drug Discovery and Cheminformatics 74

3.1 Statement of contribution 74

3.2 Summary 74

3.3 Introduction 75

3.4 Results 81

3.4.1 Comparison of Megahit and SPAdes de novo Assemblers for Known RiPP Scaffold Prediction and Discovery from Large Datasets 81
3.4.2 In Silico Library Generation and Molecular Descriptor Prediction 83
3.4.3 Cheminformatics Analysis Through Principal Component Analysis 84
3.4.4 Preparation of In Silico Library and Docking Against LspA 88
3.4.5 Tobacco Expression and Mass Spectrometry based Characterization of Biosynthetically Feasible Lyciumins 92
3.4.6 Future Experimentation 96

3.5 Discussion 97

3.6 Methods and Materials 100

3.6.1 Materials 100
3.6.2 Plant Cultivation 100
3.6.3 Peptide Chemotyping 101
3.6.4 RNA-Seq data download 101
3.6.5 Transcriptome assembly 101
3.6.6 Transcriptome analysis 102
3.6.7 Preliminary QPFGVFRW Isolation 102
3.6.8 Transient gene expression in Nicotiana benthamiana 103
3.6.9 Descriptor prediction Library prep and docking 104
3.6.10 PCA 105

3.7 Supplemental Information 105

3.8 References 119

Chapter 4 Diversification of Stephanotic-acid type Burpitides and Improved Moroidin Bioactivity against H1437 Lung Adenocarcinoma Cells 123

4.1 Statement of contribution 123

4.2 Summary 123

4.3 Introduction 124

4.4 Results 125

4.4.1 Characterization of stephanotic acid-type burpitides with a Trp-Val-crosslinked second-ring 125
4.4.2 Characterization of stephanotic acid-type burpitides with a Tyr-Phe-crosslinked second-ring 128

4.4.3 Chemotaxonomic expansion and cancer cell cytotoxic improvement of Moroidins 130
4.4.4 RNA-seq of moroidin-QLLVWRNH treated H1437 Lung Adenocarcinoma Cells, Differential Gene Expression (DGE) analysis, and AutoPlex for Mechanism of Action (MoA) Interrogation 133
4.5 Discussion 140
4.6 Methods and Materials 144
4.6.1 Materials 144
4.6.2 Plant Cultivation 144
4.6.3 Peptide Chemotyping 145
4.6.4 Glechomanin Isolation 145
4.6.5 Moroidin-QLLVWRNH isolation 147
4.6.6 Mercurialin Isolation 147
4.6.7 Structure elucidation of glechomanin and mercurialin 148
4.6.8 BURP-domain gene cloning 150
4.6.9 Transient gene expression in Nicotiana benthamiana 150
4.6.10 Protein purification 151
4.6.11 Enzyme assays 153
4.6.12 Bioassays 155
4.6.13 Phylogenetic Tree Construction 156
4.6.14 Dosed Treatment of H1437 Cells with moroidin-QLLVWRNH and RNA-seq analysis 156
4.6.15 Deseq2 for Differential gene expression and heatmap/volcano plot creation in R 157
4.6.16 Seqkit2 and ENTAP analysis on the GreatLakes Cluster 162
4.6.17 Mechanism of Action (MoA) interrogation with AutoPlex and sample output 162
4.6.18 Data Availability 162
4.7 Supplemental Information 163
4.8 Reference 216
Chapter 5 Generation of Multi-omics Plant Burptide Library 220
5.1 Statement of contribution 220
5.2 Summary 220
5.3 Introduction 221
5.4 Results 225
5.4.1 Plant Cultivation of Diverse Seeds for Seedling Growth 225
5.4.2 Workflow for Tissue Extractions and LC-MS/MS Analysis Yields More Available Metabolomes 228

5.4.3 Preliminary Data Analysis to Aid in Discovery of Macrocyclic Burpitides From Seedling Extracts 230
5.5 Discussion 234
5.6 Methods and Materials 236
5.6.1 Materials 236
5.6.2 Sample Preparation of Plant Material and Peptide Chemotyping 236
5.6.3 Seeds and Plant Cultivation 237
5.6.4 Molecular Network Generation 237
5.6.5 Phylogenetic Tree Construction and Bar Graph Construction 238
5.6.6 MassQL and OpenMS Filtering 238
5.6.7 Protein Fasta Downloads and RepeatFinder for Core Peptide Discovery 240
5.6.8 Data Availability via MassIVE 240
5.7 Supplemental Information 241
5.8 References 248
Chapter 6 Dissertation Summary and Conclusion 253
6.1 Dissertation Summary 253
6.2 Dissertation Conclusion 260

List of Tables

Table 1.1 Core peptides defined by RepeatFinder in BURP-domain protein sequences. 16

Table 1.2 Common Burpitide Modifications. 25

Table 1.3 Iminium ion masses of proteinogenic amino acids. 26

Table 1.4 Preparative HPLC gradient. 38

Table 1.5 Semipreparative HPLC gradient (selanine A example) 39

Table 2.1 Benchmarking results of *de novo* transcriptome assembly for plant RiPP precursor discovery. 57

Table 2.2 Benchmarking results of *de novo* and genome-guided transcriptome assembly for plant RiPP precursor discovery. 58

Table 2.3 Plant genomes and plant RNA-seq data in the NCBI sequence read archive. 64

Table 2.4 Benchmarking of read trimming for *de novo* transcriptome assembly of plant RiPP precursor discovery. 64

Table 2.5 Comparison of scaled transcriptome mining for fused moroidin cyclases and moroidin core peptides in published 1kp SOAPdenovo-trans assemblies and SPAdes-assemblies. 64

Table 2.6 RNA-seq datasets and plant species with predicted stephanotic acid-type burpitide cores from fused burpitide cyclases in assembled plant transcriptomes (tblastn and PHI-BLAST hits). 65

Table 2.7 Predicted core peptides of stephanotic acid-type burpitides in assembled plant transcriptomes. 66

Table 2.8 Comparison of burpitide transcriptome mining. 67

Table 3.1 Core peptide predictions by SPAdes and MEGAHIT. 82

Table 3.2 Top-scoring Predicted Peptides from LspA Docking Screen 91

Table 3.3 Predicted RiPP Library Sequences and Descriptors 105

Table 4.1 An overview of the RNA-seq experiments. 134

Table 4.2 RNA-seq datasets and plant species with predicted stephanotic acid-type burpitide cores from fused burpitide cyclases in assembled plant transcriptomes (tblastn and PHI-BLAST hits). 163

Table 4.3 ^{1}H-NMR (800 MHz) and ^{13}C-NMR (200 MHz) data for glechomanin in DMSO-d_6. 164

Table 4.4 ^{1}H-NMR (800 MHz) and ^{13}C-NMR (200 MHz) data for glechomanin in methanol-d_4. 165

Table 4.5 Tandem MS fragmentation of linear tryptic core peptide of GheBURP-1xQLFVWGW. Proteomic data MassIVE ID: MSV000097169 166

Table 4.6 Tandem MS fragmentation of putative bicyclic tryptic core peptide of GheBURP-1xQLFVWGW. Proteomic data MassIVE ID: MSV000097169 167

Table 4.7 ^{1}H-NMR (800 MHz) and ^{13}C-NMR (200 MHz) data for mercurialin in DMSO-d_6. 168

Table 4.8 MS/MS analysis of linear LysC-proteolytic core peptide of ManBURP-1xQLFFWRY. Proteomic data MassIVE ID: MSV000097169 169

Table 4.9 MS/MS analysis of putative bicyclic LysC-proteolytic core peptide of ManBURP-1xQLFFWRY. Proteomic data MassIVE ID: MSV000097169 170

Table 4.10 ^{1}H-NMR (800 MHz) and ^{13}C-NMR (200 MHz) data for moroidin-QLLVWRNH in DMSO-d_6. 171

Table 4.11 Chemotaxonomic distribution of stephanotic acid-type burpitides. 172

Table 4.12 Cloning primers...... 172

Table 5.1 Iminium ion masses of proteinogenic amino acids...... 230

Table 5.2 Known and Candidate Lyciumins Identified Through GNPS Clustering...... 233

Table 5.3 Seedling grown in this study, protein FASTA downloaded, and tissues sampled. 241

List of Figures

Figure 1.1 Traditional Workflow for Plant Natural Product Discovery and Pitfalls. 3

Figure 1.2 Workflow for Genome Mining-based Plant Natural Product Discovery and Pitfalls. 6

Figure 1.3 General depiction of RiPP biosynthetic pathway. 9

Figure 1.4 Burpitide RiPPs from plants. 15

Figure 1.5. Burpitide genome mining covered in this chapter. 17

Figure 1.6 Detection of selanine A in methanolic extract of *S. kraussiana* aerial tissue. 27

Figure 1.7 Metabolic profiling of *S. kraussiana* and heterologous SkrBURP expression. ... 34

Figure 1.8 NMR chemical shifts of burpitide crosslinks. 41

Figure 2.1 Transcriptome mining of plant RiPPs. 54

Figure 2.2 Plant RiPP structures and biosynthetic precursor peptides 68

Figure 2.3 Detection of stephanotic acid-QLLVW in plantMASST-LC-MS2 data and manually analyzed LC-MS1 data. 69

Figure 3.1 Founding members domain structure of the BURP domain family from ref[37]. . 79

Figure 3.2 Primary structures of representatives of the two general types of BURP domain cyclases and corresponding peptide natural products. 79

Figure 3.3 Structures of Drug-like Burpitide RiPPs. 80

Figure 3.4 Structures of Burpitide RiPP Motifs Mined Using RepeatFinder. 81

Figure 3.5 PCA of mined transcriptome BURP-domain-RiPPs. 85

Figure 3.6 Principal component analysis of mined plant RiPPs and know orbitides from Cybase. 86

Figure 3.7 X-ray crystal structure of LspA in complex with the macrocyclic antibiotic globomycin (PDB: 6RYO) 89

Figure 3.8 Docked structure of LspA in complex with QPFGVFRW. 92

Figure 3.9 Lyciumins analyzed by Liquid Chromatography-Mass Spectrometry (LC-MS) for characterization and quantification using extracted ion chromatograms (EICs). 94

Figure 3.10 Quantification of Lyciumin Expression analyzed by Liquid Chromatography-Mass Spectrometry (LC-MS) using extracted ion chromatograms (EICs). 95

Figure 3.11 *LbaLycA* precursor gene expression construct containing a 12x core peptide repeat of QPFGVFRW. 95

Figure 3.12 LspA mechanism 116

Figure 3.13 Principal component analysis of mined plant RiPPs and Macrocyles in clinical trials or approved by the FDA. 116

Figure 3.14 Principal component analysis of mined plant RiPPs and DrugBank's Antibacterial Agents database 117

Figure 3.15 Principal component analysis of mined plant RiPPs and the Coconut Natural Product Database 117

Figure 3.16 Principal component analysis of mined plant RiPPs and the Chembl Macrocycles Database 118

Figure 3.17 Docked QPFGAWRW in LspA active site. 118

Figure 4.1 Transcriptome mining of plant RiPPs. 124

Figure 4.2 Characterization of stephanotic acid-burpitides with a Trp-Val-crosslinked second ring. 126

Figure 4.3 Characterization of stephanotic acid-burpitides with a Tyr-Phe-crosslinked second ring. 129

Figure 4.4 Characterization of moroidin burpitides in pinks and grasses. 132

Figure 4.5 Bioactivity of moroidin-QLLVWRNH. 132

Figure 4.6 DGE Analysis Workflow of RNA-seq Experiments. 135

Figure 4.7 Moroidin-QLLVWRNH versus DMSO Differentially Expressed Genes Across All Time and Concentration Points. 136

Figure 4.8 Volcano plot of top filtered and annotated DGE results from Moroidin-QLLVWRNH vs. DMSO treatment for 6hrs at concentration IC_{50}. 137

Figure 4.9 Heatmap of top filtered and annotated DGE results from Moroidin-QLLVWRNH vs. DMSO treatment for 6hrs at concentration IC_{50}. 138

Figure 4.10 Phylogenetic distribution of characterized stephanotic acid-type burpitides in flowering plants. 142

Figure 4.11 Plant RiPP structures and biosynthetic precursor peptides. 173

Figure 4.12 MS and MS/MS spectrum of glechomanin 174

Figure 4.13 ^{1}H-NMR spectrum of glechomanin in DMSO-d_6 (800 MHz) and in methanol-d_4 (800 MHz). 175

Figure 4.14 ^{1}H-^{1}H COSY-NMR spectrum of glechomanin in DMSO-d_6 and in methanol-d_4. 176

Figure 4.15 ^{1}H-^{1}H-TOCSY-NMR spectrum of glechomanin in DMSO-d_6 and in methanol-d_4. 177

Figure 4.16 HSQC-NMR spectrum of glechomanin in DMSO-d_6 and in methanol-d_4. 178

Figure 4.17 HMBC-NMR spectrum of glechomanin in DMSO-d_6 and in methanol-d_4. 179

Figure 4.18 ROESY-NMR spectrum of glechomanin in DMSO-d_6 and in methanol-d_4. 180

Figure 4.19 ^{13}C-NMR spectrum (200 MHz) of glechomanin in DMSO-d_6. 181

Figure 4.20 Key 2D NMR correlations of glechomanin in DMSO-d_6 181

Figure 4.21 Key 2D NMR correlations of glechomanin in methanol-d_4. 182

Figure 4.22 Stereochemical analysis of glechomanin. 182

Figure 4.23 L-FDAA derivatization of the hydrolysate of glechomanin and amino acid standards. 183

Figure 4.24 Glechomanin geometry optimization and ROESY analysis of Trp7-Cα assignment 184

Figure 4.25 Characterization of glechomanin biosynthesis in *Orchidantha maxillarioides*. 185

Figure 4.26 Glechomanin cyclase purification 186

Figure 4.27 LC-MS analysis of *in vitro* reconstitution of GheBURP-1xQLFVWGW. 186

Figure 4.28 MS and MS/MS analysis of unmodified core peptide from *in vitro* reconstitution assay of GheBURP-1xQLFVWGW. 187

Figure 4.29 MS and MS/MS analysis of modified core peptide from *in vitro* reconstitution of GheBURP-1xQLFVWGW. 188

Figure 4.30 MS and MS/MS spectrum of mercurialin. 189

Figure 4.31 ^{1}H-NMR spectrum of mercurialin in DMSO-d_6 (800 MHz). 190

Figure 4.32 ^{1}H-^{1}H-COSY-NMR spectrum of mercurialin in DMSO-d_6. 191

Figure 4.33 ^{1}H-^{1}H-TOCSY-NMR spectrum of mercurialin in DMSO-d$_6$. 192

Figure 4.34 HSQC-NMR spectrum of mercurialin in DMSO-d$_6$. 193

Figure 4.35 HMBC-NMR spectrum of mercurialin in DMSO-d$_6$. 193

Figure 4.36 NOESY-NMR spectrum of mercurialin in DMSO-d$_6$. 194

Figure 4.37 ^{13}C-NMR (200 MHz) spectrum of mercurialin in DMSO-d$_6$ 195

Figure 4.38 Key 2D NMR correlations of mercurialin. 195

Figure 4.39 Stereochemical analysis of mercurialin. 196

Figure 4.40 Mercurialin geometry optimization and NOESY analysis of Phe4-C β assignment. 197

Figure 4.41 Mercurialin cyclase purification. 198

Figure 4.42 LC-MS analysis of *in vitro* reconstitution of mercurialin cyclase. 199

Figure 4.43 MS and MS/MS analysis of unmodified core peptide from *in vitro* reconstitution of mercurialin cyclase. 200

Figure 4.44 MS and MS/MS analysis of modified core peptide from *in vitro* reconstitution of mercurialin cyclase. 201

Figure 4.45 ^{1}H-NMR spectrum of moroidin-QLLVWRNH in DMSO-d$_6$ (800 MHz). 202

Figure 4.46 ^{1}H-^{1}H-COSY-NMR spectrum of moroidin-QLLVWRNH in DMSO-d$_6$. 202

Figure 4.47 ^{1}H-^{1}H-TOCSY-NMR spectrum of moroidin-QLLVWRNH in DMSO-d$_6$. 203

Figure 4.48 HSQC-NMR spectrum of moroidin-QLLVWRNH in DMSO-d$_6$. 203

Figure 4.49 HMBC-NMR spectrum of moroidin-QLLVWRNH in DMSO-d$_6$. 204

Figure 4.50 NOESY-NMR spectrum of moroidin-QLLVWRNH in DMSO-d$_6$. 204

Figure 4.51 ^{13}C-NMR spectrum of moroidin-QLLVWRNH in DMSO-d$_6$. 205

Figure 4.52 Key 2D NMR correlations of moroidin-QLLVWRNH. 205

Figure 4.53 Comparison of moroidin-QLLVWRNH standard to moroidin product of KjaBURP-1xQLLVWRNH expressed for 6 days transiently in *N. benthamiana*. 206

Figure 4.54 Characterization of moroidin-QLLVWPKH biosynthesis in *Orchidantha maxillarioides*. 207

Figure 4.55 Characterization of moroidin-QLLVWRSH biosynthesis in *Orchidantha maxillarioides.* 208

Figure 4.56 Characterization of moroidin-QLLVWTDH biosynthesis in *Orchidantha maxillarioides* 209

Figure 4.57 Characterization of moroidin-QLFVWPKH biosynthesis in *Orchidantha maxillarioides.* 210

Figure 4.58 *In vitro* anticancer assays of stephanotic acid-type peptides. 211

Figure 4.59 Characterization of stephanotic acid-QLLVW in pinks, mints, grasses, and Lowiaceae. 211

Figure 4.60 Representative peptide natural products with C-C-macrocyclic crosslinks. .. 212

Figure 4.61 Biosynthetic proposals of burpitides from this study. 213

Figure 4.62 Volcano plot of top filtered and annotated DGE results from Moroidin-RNH vs. DMSO treatment for 6hrs at concentration IC$_{25}$. 214

Figure 4.63 Heatmap of top filtered and annotated DGE results from Moroidin-RNH vs. DMSO treatment for 6hrs at concentration IC$_{25}$. 214

Figure 4.64 Volcano plot of top filtered and annotated DGE results from Moroidin-RNH vs. DMSO treatment for 6hrs at concentration IC$_{75}$. 215

Figure 4.65 Heatmap of top filtered and annotated DGE results from Moroidin-RNH vs. DMSO treatment for 6hrs at concentration IC_{75}. 216

Figure 5.1 Examples of subclasses of known burpitides. 224

Figure 5.2 Phylogenetic distribution of seedlings represented in this study. 226

Figure 5.3 Workflow for MassQL and OpenMS data filtering. 229

Figure 5.4 Lyciumin subset of molecular network of MassQL-processed LC-MS-datasets with cardinality command of 1-5 different amino acid iminium ions in a given tandem MS spectrum. 231

Figure 5.5 Candidate BURP-domain-peptide cyclase precursors identified with RepeatFinder. 232

Figure 5.6 Detection of candidate lyciumin derivatives in plant tissues. 233

Figure 5.7 Bar graph distribution of plant families. 247

Figure 5.8 Burpitide subset of molecular network of MassQL-processed LC-MS-datasets with cardinality command of 1-5 different amino acid iminium ions in a given tandem MS spectrum. 248

Figure 5.9 Cyclopeptide Alkaloid subset of molecular network of MassQL-processed LC-MS-datasets with cardinality command of 1-5 different amino acid iminium ions in a given tandem MS spectrum. 248

Abstract

Plants serve as a prolific source of novel structures for drug discovery. The identification of bioactive plant natural products through laboratory-based methods remains challenging and labor-intensive due to the inherent complexity of plant chemistry. Consequently, there is a critical need for effective strategies to enable *in silico* prediction of the structures and bioactivities of plant-derived compounds. With the increasing availability of plant genetic data and structural biology models of drug targets, this project seeks to develop a computational pipeline to predict bioactive compounds from plants by (a) identifying biosynthetically feasible molecules from plant genes in diverse genetic databases and (b) conducting both *in silico and in vitro* evaluation of predicted plant natural products against relevant human diseases. The long-term objective is to establish a plant genome mining strategy that streamlines the prediction of bioactive natural products from plant biosynthetic genes.

Ribosomally synthesized and post-translationally modified peptides (RiPPs) are produced as precursor peptides encoded by structural genes containing a core peptide region that undergoes modification and is subsequently released as the mature peptide. The primary compound class targeted in this project is burpitides, which are side-chain-macrocyclic plant RiPPs derived from BURP-domain-peptide cyclases. Mining plant genomes for peptide cyclases responsible for burpitide biosynthesis enables (a) prediction of chemical structures from precursor gene sequences, (b) *in silico* macrocyclic sampling and screening based on medicinal chemistry profiles bordering Lipinski's rule of 5, and (c) scaffold diversification to facilitate expanded *in vitro* screening. Furthermore, it is hypothesized that the NCBI Sequence Read Archive (SRA) represents an underexplored resource for large-scale, gene-guided discovery of plant bioactive peptides from fused burpitide cyclase pathways. The underlying rationale is that *in silico* analysis of plant genetic resources, when applied to rational drug discovery of diverse plant peptides, will yield new pharmacologically relevant lead structures for plant-derived therapeutics to treat human diseases.

Benchmarking *de novo* RNA-seq assembly tools transformed thousands of public datasets from non-model organisms into a searchable database of 27,224 full-length, non-redundant BURP-

domain sequences for biosynthetic gene discovery and the identification of fused moroidin cyclases, with 46 novel stephanotic acid-type burpitide cores from 17 plant species. This approach enabled the identification of new post-translational modifications in plant RiPP biosynthesis, including the isolation of glechomanin (Val-Trp crosslink) and mercurialin (Phe-Try crosslink), the discovery of moroidin-QLLVWRNH with a 5-fold improvement in bioactivity against H1437 lung adenocarcinoma cells, and the expansion of chemotaxonomic knowledge of moroidins across diverse plant families. RNA-Seq analysis, integrated with knowledge graph-driven agentic artificial intelligence, guided preliminary deconvolution of the mechanism of action for moroidin-QLLVWRNH. *In silico* library curation of 375 two- and three-dimensional structures enabled structure-based virtual screening against lipoprotein signal peptidase II (LspA), a cyclic peptide antimicrobial target. This docking enabled the computational discovery of a lyciumin hit candidate, QPFGVFRW. Expression of BURP-domain precursor genes in *Nicotiana benthamiana* provided a source-plant-independent biosynthesis strategy for accessing macrocyclic peptides. Additionally, the cultivation and extraction of seedling tissues from diverse plant species broadened metabolomic coverage by generating 1169 full MS/MS spectra, enabling the identification of lyciumin metabolites in nightshade stems and beet roots. This will enable future integration with publicly available plant genomes for natural product discovery. This research advances the understanding of plant peptide chemistry and its application in rational drug design through plant genome mining.

Chapter 1 Introduction

1.1 Perspectives of Plant Secondary Metabolites in Drug Discovery

Living organisms produce complex chemicals categorized as either primary or secondary metabolites[1]. Primary metabolites are required for the organism's growth, reproduction, and development. Primary metabolites are molecular building blocks, including nucleic acids, proteins, polysaccharides, and lipids, that fuel conserved metabolic pathways like glycolysis and the citrate cycle[2]. Secondary metabolites are not required for the organism's survival but instead provide it with a competitive edge in its environment[1–3]. Secondary metabolites populate specialized pathways that can be species- or cell-type-specific or even expressed under certain environmental conditions[2]. Secondary metabolites can be isolated from a wide range of species, with plants being prolific producers of these specialized metabolites[93].

Altogether, plants are estimated to produce anywhere from 200,000 to 1,000,000 structurally distinct natural products and a single plant may contain upwards of 5,000 to 10,000 compounds[4–6]. Many of these metabolites serve specific, evolutionarily advantageous roles for their host, as their biosynthetic pathways require substantial energy to regulate[7]. Most of these natural products are known for aiding defense or repair mechanisms triggered by abiotic and/or biotic stress responses, or even assisting in the acquisition of space and nutrients[7]. Additionally, they have been implicated in communication, microbial detection, immune response, and improving plant fitness, with a plethora of roles still being questioned and investigated[8].

Despite many natural products having unknown *in planta* functions, humans have historically used these phytochemicals as an abundant source of natural and derivatized medicines[7,9]. The use of these plant bioactives directly correlates with humanity's increase in life expectancy and quality of life, as they have been documented to fight a variety of pathogens and diseases dating back to 5000-1500 B.C.[7,9–11]. Proceeding the "Synthetic Era", ~80% of the medicines from the early 1900s were obtained from roots, barks, stems, and leaves of plants[9].

Contemporarily, a shift has occurred in large pharmaceutical companies away from natural product drug development towards prioritizing faster, more structurally expansive drug discovery methods such as combinatorial chemistry, high-throughput screening, and fragment-based screening, which can create inexhaustible chemical libraries for rapid screening[7,9]. Natural products, however, continue to play a dominant role in pharmaceuticals, as 60% and 75% of cancer and infectious disease drugs, respectively, are natural products or derivatives[7,9,12,13].

Numerous bioactive plant molecules have significantly influenced modern medicine, including the analgesic morphine from *Papaver somniferum* and the anti-malarial agent artemisinin from *Artemisia annua*[7]. Another notable example is the cytotoxic agent Taxol, which was identified from *Taxus brevifolia* extracts in 1971 and subsequently approved by the FDA in 1992, 1994, and 1999 for the treatment of ovarian, breast, and non-small cell lung cancers, respectively[7,9,14]. These landmark discoveries stimulated increased research interest in natural products throughout the latter half of the twentieth century[7,9].

The potential for discovering novel natural plant products remains vast, given the limited current understanding of the plant kingdom. The Viridiplantae includes an estimated 375,000 to 500,000 plant species; however, only approximately 28,200 species are reportedly used for medicinal purposes, and 4,478 are referenced in publications on medicinal regulatory matters[15–17]. When considering only those plant-derived agents incorporated into Western medicine, this figure decreases by a factor of 20 to approximately 150-200 plants[9,16,18]. This disparity indicates that more than 290,000 plant species remain underexplored and unstudied, especially given the exponentially greater number of secondary metabolites that could be investigated[16].

The fact that many plant species remain understudied demonstrates that currently approved agents do not fully capture the potential diversity of plant natural products for disease targeting or as lead scaffolds for developing more effective treatments. For instance, only 3% of FDA-approved antibacterial drugs from 1930 to 2013 were derived from plants, a figure that declined to 0% as all subsequently approved antibiotics were isolated from microbes[11,19]. It is therefore important to examine the factors contributing to the decline in plant-based drug discovery over the past few decades.

Despite continual advancements in bioassays, separation techniques, and structure elucidation technologies—particularly in high-resolution mass spectrometry and NMR—plant-derived drug leads have not advanced at a comparable rate. Factors such as improved cultivation

of plant species, pharmaceutical globalization, and expanded understanding of phylodiversity are contributing to a resurgence in plant natural product research[9]. However, this renewed interest has not yet bridged the knowledge gaps between underexplored plant species and the identification of viable clinical candidates. Effective translation of botanical extract analysis into the characterization and isolation of novel plant natural product drug leads requires a comprehensive understanding of the methodologies employed in the field[9]. Systematic analysis of workflows, including field collection, plant cultivation, analytical fractionation, bioassays, and structure elucidation, enables researchers to combine and adapt methods to advance novel approaches in plant natural drug discovery pipelines[9].

1.1.1 Traditional Plant Natural Product Discovery

Traditional plant natural product discovery has relied on crude botanical extract mixtures, followed by the isolation of pure chemical entities[20,21] (**Figure 1.1**). The process typically begins with the collection of large quantities of plant material to generate crude extract libraries[21]. These extracts contain complex mixtures of thousands of metabolites at varying concentrations[21,22]. Screening these crude extracts for biological relevance enables the identification of hits that can be prioritized for bioactivity-guided fractionation[20,21]. This approach involves repeatedly partitioning bioactive extracts into simplified fractions based on properties such as polarity, size, or charge[20,21,23]. Only fractions that demonstrate biological activity in subsequent re-screening assays are retained for further purification and compound isolation[20,21].

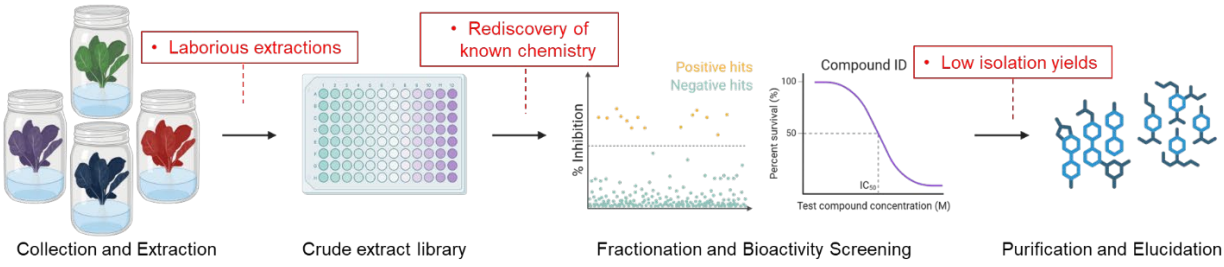


Figure 1.1 Traditional Workflow for Plant Natural Product Discovery and Pitfalls.

Liquid Chromatography-Coupled Tandem Mass Spectrometry (LC-MS/MS) is integral to the analytical characterization of plant chemistry. This technique integrates compound separation with high-sensitivity detection, supporting robust method development[2,22,24]. Advanced Orbitrap

and Q-ToF LC-MS/MS platforms deliver highly accurate molecular mass measurements, enhancing confidence in metabolite identification and enabling the detection of specific metabolites with sub-ppm accuracy[2,22–24]. Additionally, High-Performance Liquid Chromatography (HPLC) and Ultra-High-Performance Liquid Chromatography (UHPLC) offer the resolution required to separate isomers and analogues in plant extracts[2,22,24].

Nuclear Magnetic Resonance (NMR) spectroscopy is fundamental to plant natural product discovery, as it enables the determination of definitive molecular structures. NMR is particularly valuable for confirming the precise regiochemistry of compounds. Two-dimensional NMR techniques, such as COSY, HSQC, HMBC, and NOESY, facilitate the mapping of atomic connectivity both through bonds and through space, thereby supporting the validation of proposed novel structures[2,22,24]. The introduction of higher-field NMR instruments operating at 600 to 1,200 MHz has further improved sensitivity and advanced research on plant natural products[22,24]. Mastery of traditional methodologies and core instrumentation provides researchers with a comprehensive toolkit for natural product medicinal chemistry, supporting novel chemical characterization and isolation[2].

1.1.2 Challenges in Traditional Plant Natural Product Discovery

Plants represent a valuable starting point for natural product discovery; however, researchers must exercise caution and maintain precision due to persistent challenges. Historically, this process has been labor-intensive and further hindered by the metabolic complexity of plants, technical limitations of traditional bioassay screening, and difficulties in sourcing authenticated plant material[21].

The chemical complexity of plant tissues, while central to discovery, also presents significant challenges. Crude extracts may contain thousands of metabolites and unidentified impurities, leading to incorrect characterization of bioactivity during downstream analyses[21]. False positives may result from synergistic effects or from common compounds, such as tannins and flavonoids, that cause non-specific assay interference[25]. Additionally, plant chemistry is governed by enzymatic pathways evolved for ecological survival rather than human therapeutic benefit. Traditional discovery methods tend to prioritize the most abundant or accessible molecules, potentially overlooking highly potent but low-abundance metabolites[7,21,22,24].

The traditional process of plant natural product discovery is characterized by labor-intensive extraction and purification protocols, which are incompatible with the accelerated timelines of contemporary drug discovery and development[2,9,25]. Intensive dereplication is required to prevent the repeated identification of known compounds[22,24]. Crude plant natural product libraries are often unsuitable for high-throughput screening (HTS) or target-based assays due to issues such as poor compound solubility[21,22,24,25]. Pre-fractionated libraries with standardized subfractions offer a partial solution for modern HTS by increasing testing efficiency[21,25]. However, elucidating downstream mechanisms of action remains the most time-consuming step, frequently delaying drug discovery and development pipelines[21,25].

A frequently overlooked challenge in plant natural product discovery is the consistent sourcing of authentic plant material[9,20]. Securing high-quality, phylogenetically diverse material is complicated by environmental stressors, seasonal variability, and complex international regulations such as the Nagoya Protocol[2,20,25]. These factors can cause significant variation in secondary metabolite profiles, hindering the reproducibility of bioactive hits. Even when authenticated material is obtained, limited availability often restricts the structural characterization of low-abundance compounds[2,21,25]. Additionally, source plants may be rare or slow-growing, which constrains sustainable supply for downstream testing and development, and may limit commercial viability if a drug reaches regulatory approval[9]. As a result, project abandonment prior to medicinal chemistry optimization is a common outcome in plant natural product research[9,21,25].

1.1.3 Advancements in Multi-omics and Computational Discovery of Plant Natural Products

Recent methodological advances are propelling plant natural product research beyond traditional limitations. Multi-omics approaches, underpinned by comprehensive knowledge of biosynthetic genes, facilitate the discovery of novel chemical scaffolds in the post-genomic era[2]. By integrating genomics, transcriptomics, and metabolomics, these strategies address significant gaps in biosynthetic logic[94-97]. The field is transitioning from individualized spectral analysis to comprehensive metabolome analysis[2,20,22,24,25]. The Global Natural Products Social molecular networking (GNPS) platform has transformed metabolic analysis by clustering mass spectrometry fragmentation patterns, enabling the identification of structurally related molecules based on similar fragmentation patterns[26,27]. This approach streamlines the annotation of entire

metabolomes[20,22,24,25]. GNPS serves as a robust dereplication tool, minimizing redundancy in known chemical entities and prioritizing unexplored chemical space through competitive fragmentation modeling of experimental and predicted spectra[26,27].

To address the persistent shortage of new lead compounds resulting from low isolation yields and labor-intensive purification processes, researchers are increasingly utilizing plant genetic sequencing in innovative gene-guided discovery strategies (**Figure 1.2**) [2,9,20,22,28]. This shift is facilitated by the rapid expansion of sequencing data available in repositories such as the NCBI Sequence Read Archive (SRA), which provides open-access datasets abundant in biosynthetic potential[29,30].

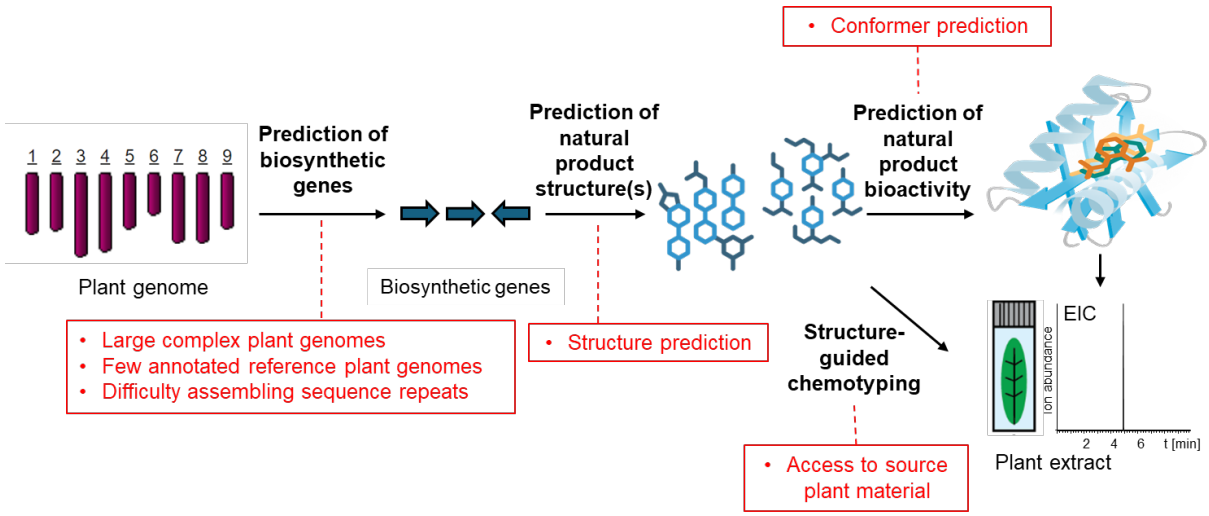


Figure 1.2 Workflow for Genome Mining-based Plant Natural Product Discovery and Pitfalls.

Innovative methods for analyzing large-scale plant genetic sequencing datasets are emerging rapidly. For instance, MetaGraph enables users to efficiently query specific biosynthetic motifs across diverse taxa, offering a search experience analogous to that of major search engines[31]. Additionally, plantiSMASH 2.0 has been developed to identify Biosynthetic Gene Clusters (BGCs) within plant genomes[32]. Unlike microbial genomes, plant genomes often lack clear clustering of BGCs, necessitating advanced algorithms such as plantiSMASH to detect co-localized biosynthetic genes and such as MEANtools to associate co-expressed or loosely linked genes involved in specialized metabolic pathways[32,33,98].

The field is experiencing a shift from reliance on serendipitous compound characterization and isolation toward the rationalized, scalable production of biosynthetically accessible metabolites[2,20,25,28]. Once promising biosynthetic pathways are identified through data mining, the

corresponding genes can be expressed in heterologous plant or microbial hosts, such as *Nicotiana benthamiana or Escherichia coli*, for metabolic engineering[25,34,35]. The feasibility of using *N. benthamiana* for plant metabolite production was recently demonstrated by the Sattely group, who successfully reconstituted a 17-gene and 20-gene biosynthetic pathway for Taxol precursors baccatin III and 3'-*N*-debenzoyl-2'-deoxypaclitaxel, respectively[95]. This work also achieved the production of taxadien-5α-ol and other key intermediates[95]. Additionally, the O'Connor group successfully demonstrated microbial biosynthesis of the chemotherapy precursors vindoline and catharanthine in a 30-step engineered heterologous pathway in yeast[99]. Furthermore, the platform enables the production of over 3,000 natural Monoterpene indole alkaloids (MIAs)[99]. This enables the production of milligram to gram quantities of target compounds, effectively addressing challenges related to limited plant material[2]. Furthermore, enzymatic engineering offers the potential to generate novel scaffolds and analogs with enhanced pharmacological properties compared to existing compounds[2]. The chemistry of predicted pathways can then be elucidated by producing the proposed molecules within controlled genetic systems[2,20,25].

High-resolution mass spectrometry-based metabolomics, molecular networking, and the integration of expanding sequence datasets are becoming foundational in plant natural products research as these approaches are increasingly adopted as standard practice[20,22,25,28]. The identification of novel lead compounds directly from genetic sequencing data addresses the limitations of traditional, labor-intensive methodologies. Importantly, multi-omics strategies are intended to complement, rather than replace, established natural product analytical chemistry techniques[2]. Together, these approaches provide comprehensive frameworks for the discovery, isolation, and scalable production of next-generation plant-derived therapeutics. Continued integration of these methods is expected to generate new knowledge and research opportunities. The combined use of computational mining and heterologous validation is poised to drive significant advances in the post-genomic era, enabling the discovery of plant-derived scaffolds previously considered unattainable for drug development[2].

1.2 Plant Genome Mining of Burpitides from Fused Precursor Peptides

Plants produce a large diversity of natural products while having some of the largest genomes found in nature. With increasing plant genomic resources, genome mining can uncover

new natural product chemistry hidden in complex plant metabolomes guided by biosynthetic gene predictions. A growing class of plant natural products, for which genome mining is an effective discovery strategy, are a recently defined plant peptide class called burpitides. Burpitides are ribosomally synthesized and post-translationally modified peptides (RiPP) which contain crosslinks between tryptophan or tyrosine and other amino acid side chains or the peptide backbone. These class-defining post-translational modifications are formed by BURP-domain-containing proteins (named after their founding members) called burpitide cyclases. Burpitide cyclases can include the core peptide motif, i.e. the sequence of the final RiPP natural product, encoded in the same polypeptide as the BURP domain in so-called fused or autocatalytic burpitide cyclases. Therefore, fused burpitide cyclases enable prediction of candidate burpitide chemotypes from genome-derived BURP-domain gene sequences. Here, we outline a genome mining approach for fused burpitide cyclase products via bioinformatic structural prediction, peptide chemotyping, biosynthetic confirmation and structure elucidation.

1.2.1 Ribosomally synthesized and post-translationally modified peptides (RiPPs) in Plants

Prokaryotic and eukaryotic organisms both possess the biosynthetic machinery to produce small peptides with a wide range of bioactivities, including immunosuppressant, antibiotic, and signaling functions[2,36]. A portion of these peptides are proteolytic fragments that originate from protein precursors. Many species have evolved specific enzymatic machinery for post-translational modifications and proteolytic cleavages, followed by dedicated protein pumps for the export of compounds into the extracellular space or specialized compartments for secretion[2,36]. Resulting small peptides contain morphed sidechains, backbones, and/or terminal connectivity, allowing them to exhibit compound behaviors, such as those of stable small molecules, to overcome protease-labile scaffolds[2,36]. This class of diverse ribosomally generated natural product peptides is known as ribosomally synthesized and post-translationally modified peptides (RiPPs)[2,36]. Current understanding of plant chemistry shows that plants produce their peptidic natural products exclusively through ribosomal pathways, leading to some interesting peptide chemistry[36].

RiPP biosynthesis involves ribosomal formation of a precursor peptide, enzymatic post-translational modification of a core peptide within the precursor peptide, and subsequent proteolytic maturation to yield the final RiPP natural product (**Figure 1.3**)[36,37]. The genetically

encoded precursor peptide(s) are the preliminary substrate for these biosynthetic pathways. Precursor peptides sequences are subcategorized based on functionality. The definition of RiPP precursors was recently expanded to include leader, core, and follower peptides, as well as recognition sequences[2,36,37]. Leader, follower, and recognition sequences are often the regions that attract and engage post-translational modifying enzymes/protein domains and proteases for interaction; however, the presence of these regions is not required for plant RiPP precursor peptides[2,36,37].

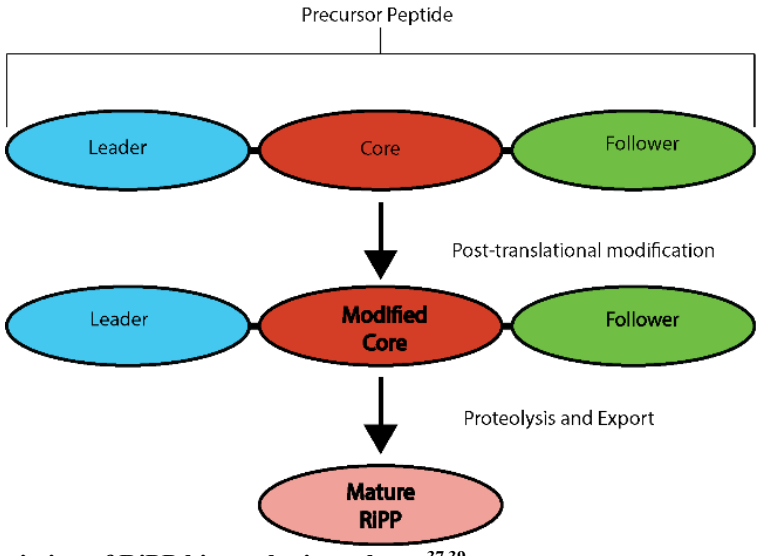


Figure 1.3 General depiction of RiPP biosynthetic pathway[37,39].

Plant RiPPs derive from single- or multi-core precursors, with the vast majority characterized as multi-core precursor peptides, meaning they include more than one peptide embedded in a repetitive sequence[36]. Lyciumins, moroidins, and cyclopeptide alkaloids all exhibit multi-core precursor peptide biosynthetic production pathways[36]. Advantages for plants producing multi-core precursor peptides over single-core precursor peptides include increased, faster, and more efficient transcription and translation of a single gene, yielding multiple RiPP products[36].

The hypothesized evolutionary path of plant multi-core peptides is internal gene duplication and neofunctionalization for core peptide diversification[36]. Plant precursor peptides are typically 100 amino acids or more, making them larger than the precursor peptides typically observed in bacterial RiPP biosynthesis[36]. Plants have larger RiPP precursor peptides because they often include signal peptides pathways for processing and secretion, possess multiple core peptides, have evolved core peptides as storage proteins, and can be directly fused to catalytic

domains[36]. Furthermore, scientists hypothesize that cyclic peptide motifs developed 'hijacking' mechanisms to subvert the plant protein production and processing machinery and to adapt them for biosynthetic production[36]. These cyclic motifs may then function as inhibitor protein loops or RiPP core peptides, enabling downstream post-translational biosynthetic modifications and proteolysis.

Plant RiPPs generally comprise less post-translational modification (PTM) diversity than fungal and prokaryotic RiPPs and they are often macrocyclic[37–40]. This is because bacterial RiPPs incorporate non-proteogenic amino acids and the machinery to modify them, whereas plant RiPPs generally contain only proteogenic amino acids, and only a limited number of them are post-translationally modified[36]. Scientists have isolated and characterized >1,500 modified peptides from plants to date, with sizes ranging from 4 to 66 amino acids[36].

Macrocyclization patterns observed in plant RiPPs include head-to-tail (cyclotides), side-chain-to-side-chain (cyclopeptide alkaloids), or side-chain-to-backbone (lyciumins)[36]. Furthermore, several plant RiPP classes are characterized by post-translational modifications, such as disulfide bonds, pyroglutamate, tyrosine sulfation, and proline hydroxylation, which are also common to other eukaryotic species[36]. Plant RiPPs are biosynthesized via split- or fused-autocatalytic pathways. Cyclopeptide alkaloids are examples of RiPPs produced from split pathways. In split pathways, they are expressed as proteins separated from their respective downstream post-translational modifying enzymes[36]. Characterization of BURP-domain-containing proteins' roles in RiPP biosynthesis provided evidence for autocatalytic pathways in plant biosynthesis, in which the core peptide is attached to the cyclase domain responsible for macrocyclization[36,41]. Finally, plant RiPP biosynthesis uses processing enzymes for secretion, such as the proteolytic asparaginyl endopeptidases (AEPs) that process vacuolar storage proteins[2,36].

Plant RiPPs can be discovered using a variety of methodologies, including bioactivity-guided, structure-guided, and gene-guided approaches. Of these approaches, structure-guided and gene-guided approaches have proven to be effective for targeted discovery of plant RiPPS[36]. Though bioactivity-guided is a well-established discovery strategy in plants, two main downsides to its use for RiPP discovery are time-consuming fractionation and the rediscovery of known bioactive natural products.

Biocheminformatic pipelines have been developed to aid in correlating bioactivity of fractionated plant extracts to LC-MS datasets containing peptide mass signals, such as employing

PepSAVI (statistically-guided bioactive peptides prioritized *via* mass spectrometry) for high-throughput identification of candidate masses of bioactive peptides, enabling rapid downstream dereplication leading to characterization of cyclotide cyO2 from an extract library of *Viola odorata*[42,43]. Existing MS-based strategies, including molecular networking and Dereplicator+, through Global Natural Product Social Molecular Networking (GNPS), also enable rapid dereplication and discovery of potential bioactive RiPP compounds through spectral library comparison or *de novo* structure elucidation for assessment of novel chemotypes[26,27,36,44,45].

Structure-guided chemotyping is a mainstay discovery strategy for plant RiPPs. This strategy heavily relies on understanding and exploiting physicochemical properties during natural product isolation or the utilization of analytical chemistry techniques. Mass spectrometry is a powerful tool for RiPP characterization, providing structural information about the analytes. Macrocyclic post-translational modifications pose challenges that prevent peptide sequencing[36]. The recent introduction of Ultraviolet photodissociation (UVPD) tandem mass spectrometry and the further development of collision-induced dissociation both represent successful advances in methods for increased MS fingerprinting and fragmentation from underivatized extract samples, enabling peptide sequencing from mass spectrometry data[36,46]. Peptide sequencing data from MS/MS directly informs whether an analyte plant peptide is novel and should be used to guide further experimentation.

Predicted amino acids and the target analyte mass are necessary to enable the connection between RiPP core peptides to the target spectrum, and effectively identified lyciumins, stephanotic acid-type burpitides, and cyclopeptide alkaloids[36]. This process of connecting MS/MS data to the respective precursor peptide sequences for RiPP discovery has recently been implemented in HypoRiPPAtlas[47], enabling RiPP structural prediction from candidate core peptide sequences in BURP-domain-containing precursor peptides. Next, Dereplicator+ generates theoretical tandem mass spectra based on predicted RiPP structures and compares them to experimental tandem MS data for discovery[44,45,47]. Proof-of-concept studies using HypoRiPPAtlas identified several lyciumins and cyclopeptide alkaloids[47]. Indeed, the structural information provided by mass spectrometry is a timeless resource for driving new plant RiPP structure-guided discovery approaches, but it must be complemented by NMR or crystallography to determine the true structures of candidate peptides[36].

Access to plant genetic data resources has exploded in recent decades, with nearly 10,000 species transcriptome datasets and 1000s of plant genomes available via the NCBI [29,30]. Pairing that wealth of sequencing knowledge with increased characterization of precursor peptide genes leads to novel RiPP discovery methods through gene-guided bioinformatic prediction of peptide chemotypes. Genome mining and transcriptome analysis successfully identified split and fused precursor peptides involved in the biosynthesis of cyclotides, cysteine-rich peptides (CRPs), lyciumins, cyclopeptide alkaloids, hibispeptins, and moroidins[36]. The preliminary bioinformatic characterization of precursors directly yielded a candidate core peptide sequence for downstream analytical analysis and chemotyping.

One persistent challenge in gene-guided discovery of plant RiPPs is the lower biosynthetic operon clustering compared to bacterial species, which complicates the identification of new precursor peptides and enzymes that catalyze post-translational modifications[36]. Searching for open reading frames (ORFs) that encode core peptides, or identifying scaffold-generating RiPP genes fused to their presumed precursor peptides provides routes to discovery[36]. Furthermore, tissue-specific paired transcriptomics and metabolomics are necessary to create a powerful discovery dynamic, especially when actively employed together to elucidate pathways involving non-co-localized biosynthetic enzymes[36]. In this regard, the power of the discovery dynamic would be quantified by statistical coupling of gene expression with metabolite molecular abundance via Pearson correlations, with correlation coefficients>0.8 across different tissues, providing high-confidence links. Another way to quantify this dynamic is through fold-change enrichment, which shows that both genes and metabolites are highly enriched in the same specialized tissue or at a specific developmental stage. A >10-fold increase in both transcript and peptide concentrations in response to a specific trigger would provide evidence of a functioning biosynthetic pathway in the absence of genomic clustering. Finally, quantified precision between the genetic sequence and the mass spectrum is useful when the mass error is <5 ppm for the predicted metabolite, paired with an MS/MS fragmentation pattern that accounts for >90% of the theoretical b- and y- ions of the genetically encoded sequence.

The robust co-existence and detailed validation of analytical and computational methods fuel the rapid acceleration of plant RiPP research inquiries and discovery opportunities. These methods advance the trajectory of plant natural products in biotechnology, bioengineering, and medicinal applications. These are fueled by collaborative projects such as the 1KP database, the

10KP project, Phytozome, and GNPS, enabling seamless, effortless sharing of high-quality metabolomics and genetic data among researchers[15,26,27,48,49]. The next boundary to push in the identification of plant RiPPs is the automated integration of multi-omics datasets to enable discovery and chemotyping, as observed for microbial RiPPs.

Microbial RiPP discovery has benefited from highly integrated pipelines like antiSMASH and PRISM that link BCGs to predicted chemical structures. The core methodological gap for plants is the disconnection between genomic potential and metabolomic reality. Researchers can identify a candidate BURP-domain or cyclizing enzyme in a genome, find an uncharacterized macrocycle in a tissue extract, or observe a specific biological response in a cell assay. More standardized automation to link these datasets is a computational challenge because plant genomes are large and repetitive, making RiPP BCG identification more complicated, as they are not tightly clustered. There is a speed disconnect between the chemical search space for core peptides from precursor candidates and cross-referencing against MS/MS spectra. Furthermore, plant RiPPs may involve additional on-canonical post-translational modifications (PTMs) that are not directly inferable from sequencing data, creating mismatches between sequencing data and observable chemistry. This may require researchers to make educated guesses about chemical structures based on mass shifts, potentially leading to higher false discovery rates. Algorithms like Met-Miner attempt to link plant specialized metabolites to their biosynthetic genes by looking for correlated patterns of occurrence across hundreds of species. MetMiner demonstrates validated efficiency in data mining and robustness in metabolite annotation, thereby directly matching MS/MS fragmentation patterns to genetic sequences. Overall, the missing piece is an accurate, integrated discovery engine that can link sequencing data to metabolites and then predict biological activity, especially for plant peptide species.

Furthermore, benchmarking and improved *de novo* assembly strategies are crucial for assembly of plant RiPP precursor genes that are inherently repetitive due to core peptides being embedded as sequence repeats. Improved tandem mass spectrometry methods and their implementation in innovative proof-of-concept experiments to generate core peptide sequence data for *de novo* or gene-aided structure characterization are also needed[36]. Structural and mechanistic studies of burpitide macrocyclization catalyzed by BURP-domain cyclase offer a unique area for new research questions and discoveries, as biosynthetic knowledge surrounding downstream proteolytic processing of split and fused pathways, as well as terminal core-peptide modifications,

remains limited[36]. The future of plant RiPP discovery, especially regarding burpitides, is overflowing with uncharacterized bioactive molecules, enzyme reactions, and bioengineering platforms.

1.2.1.1 Introduction to Burpitides

Genome mining of ribosomally synthesized and post-translationally modified peptides (RiPPs) is well-established in microbes but it is still in its infancy for organisms with large genomes[37,39,50,51]. Plants have some of the largest genomes found in nature and produce a great diversity of peptide natural products via the ribosomal pathway[36,52]. This offers unique opportunities for genomics-guided RiPP discovery in times of rapidly growing plant genomic resources[48]. This article describes genome mining of a recent addition to plant RiPP classification called burpitides[36]. The result of our presented burpitide genome mining experiment is a structurally elucidated and biosynthetically verified burpitide based on bioinformatic prediction of a precursor peptide from a plant genome. Burpitides are small cyclic plant RiPPs of 4–8 amino acids with crosslinks between aromatic amino acids such as tryptophan and tyrosine to other amino acids or the peptide backbone[36,41,53–56]. These crosslinks often involve functionalization of C(sp3)-H-bonds by copper-dependent burpitide cyclases comprising a C-terminal BURP domain[57–59]. In addition, burpitide crosslinks can also derive from aromatic cross-coupling like in moroidin[54]. Burpitides are differentiated based on their crosslinks into (a) lyciumins, which have a tryptophan-indole-N-crosslink, (b) stephanotic acids, which have a tryptophan-indole-C-crosslink, (c) cyclopeptide alkaloids, which have a tyrosine-phenol-ether-crosslink, and (d) hibispeptins which are proposed as burpitides to have crosslinks via tyrosine-phenol-C-bonds (**Figure 1.4a**)[41,55,60–62,56,63].

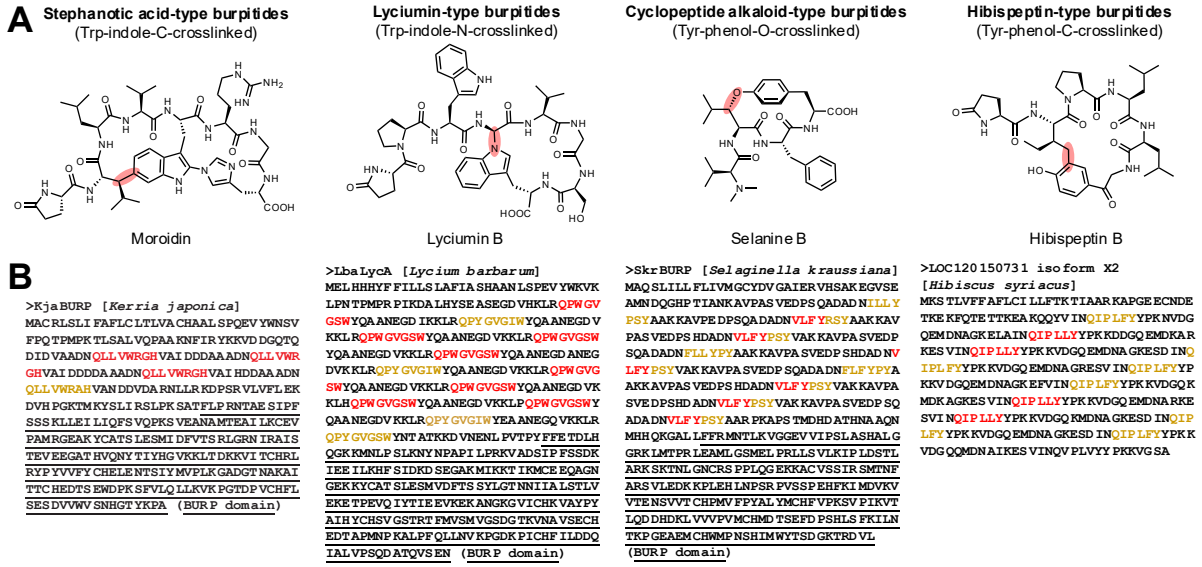


Figure 1.4 Burpitide RiPPs from plants.
(A) Subclasses of burpitides differentiated by crosslinks. Macrocyclic bonds formed by PTM are highlighted in red color. (B) Burpitide precursor peptides of representative burpitides shown in Figure 1.4A. BURP-domain sequences are underlined, core peptides specific to structures in Figure 1.4A are highlighted in red, core peptides for other burpitides are highlighted in orange.

Many precursor peptides of burpitides include multiple core peptides embedded in sequence repeats and are large proteins such as lyciumin precursor peptide LbaLycA (543 amino acids, **Figure 1.4b**)[41,54,55,47,56,64]. Burpitides have two general types of precursor peptides - split and fused precursor peptides[36,41,56]. Fused precursor peptides include core peptides in the same polypeptide as the BURP domain whereas split precursor peptides are encoded in different open reading frames than their peptide cyclases. Plant genomes contain on average 12 BURP-domain proteins and several studies have applied genome mining of fused BURP-domain precursor peptides for the discovery of burpitide chemistry[65,41,47]. Genome mining of fused burpitide precursor peptides has been prolific in the discovery of new plant RiPP chemistry exemplified by the subclass-defining discoveries of lyciumins, cercic acid, stephanotic acid-[LV], and cyclopeptide alkaloids. These compounds were then biosynthetically verified by transient expression in *Nicotiana benthamiana* or *in vitro* reconstitution of corresponding burpitide cyclases[41]. The direct connection of burpitide core peptides and burpitide cyclases in fused precursor pathways enables discovery via genome mining of BURP-domain sequences in a given plant genome. In addition, for one characterized split burpitide pathway from *Coffea arabica,* the separate precursor peptide is colocalized with its burpitide cyclase[56].

As burpitide PTMs include C-N-, C-O-, and C-C-bonds between tryptophan or tyrosine and other amino acid side chains or the peptide backbone, and burpitides can be mono- or bicyclic, a large chemical space is achievable via burpitide pathways. Therefore, most of these plant RiPPs are likely uncharacterized and ripe for picking via genome mining approaches such as the one described in this article.

1.2.1.2 Plant Genome Mining via RepeatFinder

RepeatFinder was developed as an algorithm to characterize sequence repeats and associated burpitide core peptides within BURP-domain protein sequences. RepeatFinder has (a) a module, which characterizes known, i.e. predefined, core peptides in BURP-domain protein sequences and (b) a module which recognizes sequence repeats *de novo* similar to previously published repeat detection algorithms such as RADAR[66,67]. Core peptides of burpitide subclasses predefined in RepeatFinder are listed in **Table 1.1**.

Table 1.1 Core peptides defined by RepeatFinder in BURP-domain protein sequence.

Core peptide (x = any proteinogenic amino acid)	Burpitide	Core peptide in RepeatFinder known_motif.txt (. = any proteinogenic amino acid)	RepeatFinder annotation
QPxxxxxW	Lyciumin	QP…..W	Lyciumin
QPxxxYxW	Legumenin	QP…Y.W	Legumenin
QIxxW	Cercic acid	QI..W	Cca-like peptide
xLxYxxY	Bicyclopeptide alkaloid	.L.Y..Y	Bicyclic cyclopeptide alkaloid
QLxxW	Stephanotic acid	QL..W	Stephanotic acid

A genome mining experiment can be defined as the connection of biosynthetic genes with their natural product chemotype(s) through applied biosynthetic knowledge. In this burpitide genome mining example, a BURP-domain-protein-encoding gene is connected to a burpitide analyte by identification of core peptide motif(s) in a BURP-domain protein. Our general outline of burpitide genome mining includes the following steps (**Figure 1.5**):

1) Identification of BURP-domain protein sequences from a plant genome
2) Prediction of a burpitide precursor peptide and its core peptide(s) from BURP-domain proteins

3) Detection of burpitide analytes matching predicted core peptide sequences
4) Experimental connection of a burpitide chemotype and a BURP-domain precursor peptide
5) Structure elucidation of a target burpitide

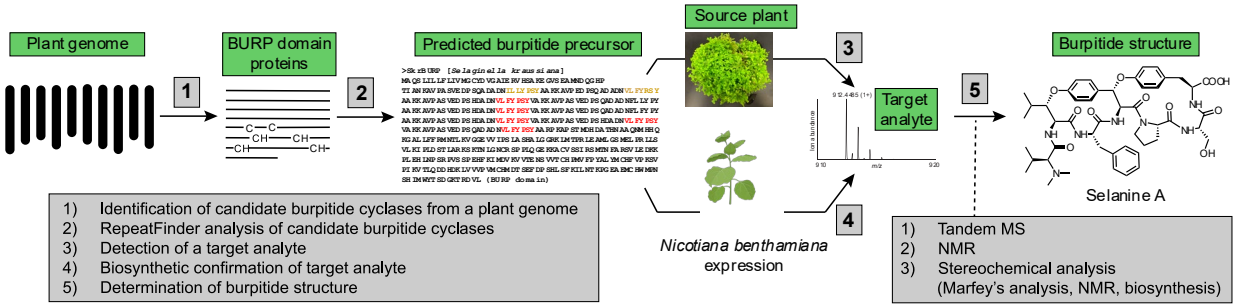


Figure 1.5. Burpitide genome mining covered in this chapter.

The genome mining example in this article is the characterization of bicyclopeptide alkaloid selanine A from Selaginella *kraussiana* (**Figure 1.5**)[41].

1.2.2 Methods and Key Resources

1.2.2.1 Bioinformatics

- NCBI BLAST: https://blast.ncbi.nlm.nih.gov/Blast.cgi
- Ubuntu (e.g. version 24.04.1) – Linux environment
- RepeatFinder: https://github.com/FlorisdeWaal/RepeatFinder_standalone
- Python 2 distribution (RepeatFinder currently only runs on Python 2)
- Selaginella *kraussiana* SequenceServer: https://sk.ccgg.fun/blast

1.2.2.2 Plant Cultivation – Nicotiana benthamiana

- Seeds: USDA Germplasm – Accession: PI 555478
- Soil: Premier Horticulture Pro Mix HP High Porosity with Mycorise
- Growth light: Spider Farmer SF2000

1.2.2.3 Transient peptide expression in N. benthamiana

- *Agrobacterium tumefaciens* LBA4404 (e.g. GoldBio, CC-220-5×50)
- Expression vector pEAQ-HT: Leaf Expression Systems, George Lomonosoff (John Innes Centre, Norwich, UK)[34].
- Electroporator: Biorad® MicroPulser Electroporator (Cat#: 1652100)
- Gene synthesis – Twist Bioscience Inc.®, Gene Fragment without Adapters (Cat#: 107448)

1.2.2.4 Peptide chemotyping

- FastPrep-24™ 5G bead beating grinder and lysis system (MP Biomedicals, Cat#: 116005500)
- Mass spectrometer: Thermo HESI-II-QExactive Orbitrap mass spectrometer coupled to a Thermo Vanquish UHPLC system
- LC column: Phenomenex Kinetex® 2.6 µm C18 reverse phase 100 Å 150 × 3 mm (Cat#: 00F-4462-Y0)

1.2.2.5 Peptide purification

- LH-20 LC - solid phase: Sephadex LH-20, 100 g (Cytiva, Cat#: 17009001)
- Gilson fraction collector (FC 203B model, Cat#: 171011DV)
- Glass column - Chemglass, 8 × 60cm column, 45/50 (Cat#: CG118931)
- LC-MS column: Phenomenex Kinetex® 2.6 µm C18 reverse phase 100 Å 50 × 3 mm (Cat#: 00B-4462-Y0)
- HPLC - Instrument: Shimadzu LC-20AP liquid chromatograph equipped with a SPD-20A UV/VIS detector and a FRC-10A fraction collector
- Preparative column: Phenomenex Kinetex® 5 µm C18 100 Å LC Column 150 × 21.2 mm (Cat#: 00F-4601-P0-AX)
- Semipreparative column: Phenomenex Kinetex® 5 µm C18 100 Å LC Column 250 × 10.0 mm (Cat#: 00G-4601-N0)
- Lyophilizer - Labconco™ FreeZone™ 4.5 L −50 °C Benchtop Freeze Dryer (Cat#: 700401000)

1.2.2.6 Peptide structure elucidation

- NMR - Glassware: Shigemi Tubes (DMSO - Cat#: DMS-005B, Methanol - Cat#: MMS-005B, Chloroform - Cat#: CMS-005B)
- Marfey's analysis - Glassware: Pressure Vessel, Heavy Wall, Cylindrical, 2 mL, #7 GL Thread Capacity 2 mL GL thread size 7, O.D.x Overall 13 × 45 mm (Cat#: P160702)
- Geometry optimization: Gaussian (version Gaussian 16, Revision A.03)[68].

1.2.3 Genome mining of fused burpitide cyclases

In this step, a fused burpitide precursor peptide is predicted from a plant genome.

1.2.3.1 Identification of candidate burpitide cyclases from plant genomes

The first step in burpitide genome mining from fused precursor peptides is the identification of all BURP-domain proteins from a given plant genome. To identify all BURP-domain protein sequences in the target plant *Selaginella kraussiana*, a BURP-domain protein query is searched via blastp in the annotated proteins of the *S. kraussiana* genome[69,70].

- Blastp Sali3-2 sequence (Genbank accession AAB66369.1) against annotated proteins of a target plant genome on NCBI or on a genome-specific BLAST SequenceServer[71].

  Example: The *Selaginella kraussiana* genome is hosted on an internal SequenceServer (version 3.1.3, https://sk.ccgg.fun/blast) for BLAST search. Blastp Sali3–2 in *S. kraussiana* [Selaginella kraussiana_v2] Protein database (Blast version: BLASTP 2.16.0+, blast parameters: evalue 1e-05, matrix BLOSUM62, gap-open 11, gap-extend 1, filter F) on the SequenceServer. The annotated proteins of the *S. kraussiana* genome are not in the NCBI non-redundant database. For other plants, check target plant genome papers for location of annotated protein files for BLAST search.
- Download all blastp hits of BURP-domain proteins into a single protein fasta file.

  Example: Eight BURP-domain proteins were identified in the annotated proteins from the *S. kraussiana* genome on SequenceServer https://sk.ccgg.fun/blast[70,71].

1.2.3.2 RepeatFinder analysis of candidate burpitide cyclases

The second step in burpitide genome mining from fused precursor peptides is the prediction of burpitide core peptides within BURP-domain protein sequences of the *S. kraussiana* genome with RepeatFinder.

- Download Ubuntu (e.g. version 24.04.1). On PC, install Windows Subsystem for Linux (WSL) in command line:

```
#Installation of Windows Subsystem for Linux.
wsl –install
```

- In Ubuntu, clone the git repository for RepeatFinder to your local machine.

```
#Cloning git repository for RepeatFinder to your local machine.
git clone   https://github.com/FlorisdeWaal/RepeatFinder_standalone
```

- Alternatively, master files for RepeatFinder can be downloaded directly to your local machine from: https://github.com/FlorisdeWaal/RepeatFinder_standalone.
- RepeatFinder will not run using Python 3, so that your Linux environment requires switching to Python 2 with the following commands. Further information can be found here: https://linux.how2shout.com/how-to-install-python-2-7-on-ubuntu-24-04-noble-lts-linux/

```
#Switching to python2.
sudo apt update && sudo apt upgrade
sudo apt install -y build-essential checkinstall libncursesw5-dev
libssl-dev libsqlite3-dev tk-dev libgdbm-dev libc6-dev libbz2-dev
libffi-dev
wget https://www.python.org/ftp/python/2.7.18/Python-2.7.18.tgz
tar -xvf Python-2.7.18.tgz
cd Python-2.7.18
./configure --enable-optimizations
make
sudo make install
sudo apt install curl
curl https://bootstrap.pypa.io/pip/2.7/get-pip.py -o get-pip.py
sudo python2.7 get-pip.py
```

- Open the file 'known_motifs.txt' to edit the core peptide motif search query that RepeatFinder will use to highlight known core peptide sequences in your fasta file. Additional search motifs can be added to this file based on experimental needs for genome mining **Table 1.1**.

```
#Open 'known_motifs.txt' file for optional editing.
cd /home/username/RepeatFinder_standalone

nano known_motifs.txt
```

- Run RepeatFinder software locally to analyze the target protein fasta file:

```
#Running RepeatFinder to analyze the target protein fasta file.
cd /home/username/RepeatFinder_standalone

python2 run_rf.py -i
/home/username/RepeatFinder_standalone/protein.fa                    -o
/home/username/RepeatFinder_standalone/Selaginella_kraussiana_BURP.ht
ml
```

- For detection of BURP-domain proteins with bicyclopeptide alkaloid core peptides, open the output file designated in the previous step (formatted as an html file that can be opened in any web browser). In the web browser, click ctrl+F to search the file for "bicyclic cyclopeptide alkaloid".

Example: Among the eight BURP-domain proteins in the *S. kraussiana* genome, RepeatFinder identified a BURP-domain protein (accession SK09G15410) which has eight bicyclic cyclopeptide alkaloid motifs (see output below). Among these motifs, the predicted core peptide VLFYPSY is represented six times.

RepeatFinder hit of SkrBURP in *S. kraussiana* genome[70]:

```
sequence SK09G15410
Pattern ADN[IFV]L[LF]Y[PR][YS][YA][AV][AK][KRG] has been found 9
times
```

Instances:

ADNILFYPSYAAK|ADNVLFYRSYAAK|ADNVLFYPSYVAK|ADNFLLYPYAAKK|ADNVLFYP
SYVAK|ADNVLFYPSYAAR|ADNVLFYPSYAAR|ADNVLFYPSYAAG|ADNVLFYPSYAAR

Known motif bicyclic cyclopeptide alkaloid pattern: .L.Y..Y has been found 8 times

Sequence:

MAQSLILLFLIVMGCYDVGAIEKRLPWHYEHADN**ILFYPSY**AAKAAKKAVPEDPSQADADN**VLF**
YRSYAAKAAKKAVPASVEDPSHDADN**VLFYPSY**VAKVAKKAVPASVEDPSQADADN**FLLYPY**AA
KKAVPASVEDPSHDADN**VLFYPSY**VAKVAKKAVPTSVEDPSQADADN**VLFYPSY**AARAARHKAP
STMDHDATHNAAHNKEDPSQADADN**VLFYPSY**AARAARPKAPSTMDHDATHNAAHNKEDPSQAD
ADN**VLFYPSY**AAGAAGPKAPSTMDHDATHNAAHNKEDPSQADADN**VLFYPSY**AARPKAPSTMDH
DATHNAAQNMHHQKGALLFFRMNTLKVGGEVVIPSLASHALGGRKLMTPRLEAMLGSMELPRLL
SVLKIPLDSTLARKSKTNLGNCRSPPLQGEKKACVSSIRSMTNFARSVLEDKKPLEHLNPSRPV
SSPEHFKIMDVKVVTENSVVTCHPMVFPYALYMCHFVPKSVPIKVTLQDDHDKLVVVPVMCHMD
TSEFDPSHLSFKILNTKPGEAEMCHWMPNSHIMWYTSDGKTRDVL

(**bold** – annotated bicyclopeptide alkaloid core peptides, underlined – annotated sequence repeats detected by RepeatFinder)

1.2.4 Detection of Predicted Burpitides

Following prediction of a candidate precursor peptide, analytes matching its core peptide(s) are analyzed from either the source plant or from a heterologous expression system of the precursor peptide such as *N. benthamiana*. A successful genome mining experiment requires:

(a) detection of an analyte matching a predicted burpitide core peptide within a candidate precursor peptide
(b) biosynthetic confirmation of the predicted burpitide precursor peptide of a target burpitide analyte
(c) characterization of a plant source for isolation of sufficient quantities of a target burpitide

1.2.4.1 Detection of an Analyte Matching a Predicted Burpitide Core Peptide

The detection of analytes matching core peptides within a candidate precursor peptide requires the acquisition of the genome-sequenced source plant. Plant sources should be evaluated for potential scale-up peptide isolation following burpitide detection.

- **USDA germplasm database** (https://www.ars-grin.gov). This resource provides plant material such as seeds to educational institutions free of charge. It is less useful for natural product isolation from provided source plant material due to generally small amounts provided by it.
- **Commercial vendor**. For international vendors, a plant import permit (PPQ-587) from USDA and shipping guidelines such as phytosanitation certificate provided by the vendor are usually required. It can be a good source for both peptide detection and purification depending on available plant material.
- **iNaturalist** (https://www.inaturalist.org/). iNaturalist is a resource to search for plant species in any given country. For collections in public areas, please obtain collection permits as required, e.g. through the local Department of Natural Resources within the U.S. It can be a good source for both detection and purification depending on available plant material at natural habitat. It is recommended to not collect more than 5% of a population (annual plants) or 5% of reproductive structures of trees and shrubs.
- **Botanical gardens.** Check local and national botanical gardens for your target plant in living collections and contact botanical garden scientists for plant access. A great resource within the U.S. is the Missouri botanical garden in St. Louis, MO with more than 8800 species of plants (https://livingcollections.org/home). It is less useful for natural product isolation from provided source plant material due to generally small plant amounts.
- **Herbaria.** In addition to living plant collections, check local herbaria for the presence of your target plant. An example within the Midwest U.S. is the Midwest Herbaria consortium (https://midwestherbaria.org/portal/). Herbaria are less useful for natural product isolation from provided source plant material due to generally small amounts provided by herbaria.

Example: A live plant of *Selaginella kraussiana* was obtained for initial chemotyping and RNA isolation from the Matthaei botanical garden at the University of Michigan, Ann Arbor, and it was then purchased in larger amounts from a commercial vendor (Plants4Home LLC) for peptide isolation.

To analyze a source plant for a target peptide, a methanolic extract of available tissues is generated for liquid chromatography-tandem mass spectrometry analysis.

- Collect a small sample (0.3 g for leaves, 0.2 g for stem, 0.1 g for seeds) from several tissues of the target plant (e.g. root, stem, leaf, seed – for transgenic *Nicotiana benthamiana* only test infiltrated leaves).
- Example: 0.3 g of *S. kraussiana* aerial tissues were collected for chemotyping.
- Add *N. benthamiana* (or other soft) leaves to plastic homogenizer tubes (2 mL, Cat#: 5076200) with 10–15 × 2 mm silica balls (Cat#: NC0419764) and screw-top lids (Cat#: 5068002) or equivalent tubes for tissue-grinding instruments. Add woody and hard tissues to metal homogenizer tubes with a 0.25 in. steel ball (2 mL, Cat#: MP116991006).
- Freeze the sample tubes at −80 °C for>30 min.
- Grind the samples with a homogenizer (for example, MP Biomedical, FastPrep-24™ 5 G, Cat#: 116005500). Load a maximum of 12 metal tubes in the 24-tube FastPrep homogenizer holder for 60 s at 6 m/s.
- Add 1 mL of 80 % methanol (HPLC grade) to the ground plant material.
- Grind the samples with Tissuelyzer for an additional 20 s at 6 m/s.
- Incubate sample tubes at 60 °C in waterbath for 10 min.
- Centrifuge the tubes in a benchtop centrifuge at 16000 x *g* for 5 min.
- Filter the extract with Whatman® Mini-UniPrep® syringeless filters with slit septa (Cat# WHAUS503NPEORG).
- Example: *S. kraussiana* aerial tissue was ground in plastic tissuelyzer tubes.
- Analyze the crude methanol extracts by LC-MS on a high-accuracy mass spectrometer with an electrospray ionization source coupled to a UPLC/HPLC, for example, Thermo HESI-II-QExactive Orbitrap mass spectrometer coupled to a Thermo Vanquish UHPLC system.

- LC parameters: injection volume 5 µL, Phenomenex Kinetex® 2.6 µm C18 reverse phase 100 Å 150 × 3 mm, Solvents: A – 0.1 % formic acid, B – acetonitrile (0.1 % formic acid), gradient: 0 min: 10 % B, 5 min: 60 % B, 5.1 min: 95 % B, 6 min: 95 % B, 6.1 min: 10 % B, 9.9 min: 10 % B, 0.5 mL/min,
- MS – positive ion mode, Full MS: Resolution 70,000, mass range 400-1,500 *m/z*, dd-MS2 (data-dependent MS/MS): resolution 17,500, AGC target 1e5, loop count 5, isolation width 1.0 *m/z*, collision energy 25 eV, dynamic exclusion 0.5 s. Most burpitides peptides are detectable in the mass range 400-1,500 *m/z*.

- Calculate monoisotopic masses of linear core peptides and modify based on possible modifications (**Table 1.2**). For peptides with arginines and lysines, calculate $[M+2H]^{2+}$ species as mass signals with more charges can be more intense for peptides with positively charged amino acids. Peptide masses can be calculated in ChemDraw (Revvity Signals, e.g. version 22.2.0).

Table 1.2 Common Burpitide Modifications.

Post-translational modification	Core peptide mass shift [Da]
Proton	+1.0073
Cyclization	-2.0156
Pyroglutamate formation (N-terminal core peptide Gln)	-17.0266
N-monomethylation	+14.0157
N,N-dimethylation	+28.0310
C-terminal oxidative decarboxylation	-46.0055

Example:

m(calc, VLFYPSY) = 888.4502 [M+H]+
+ (2x14.0157) (2 Me, *N,N*-dimethylation of valine)
- (4x 1.0078) (bicyclization via loss of 2 hydrogens)
912.4504 Da

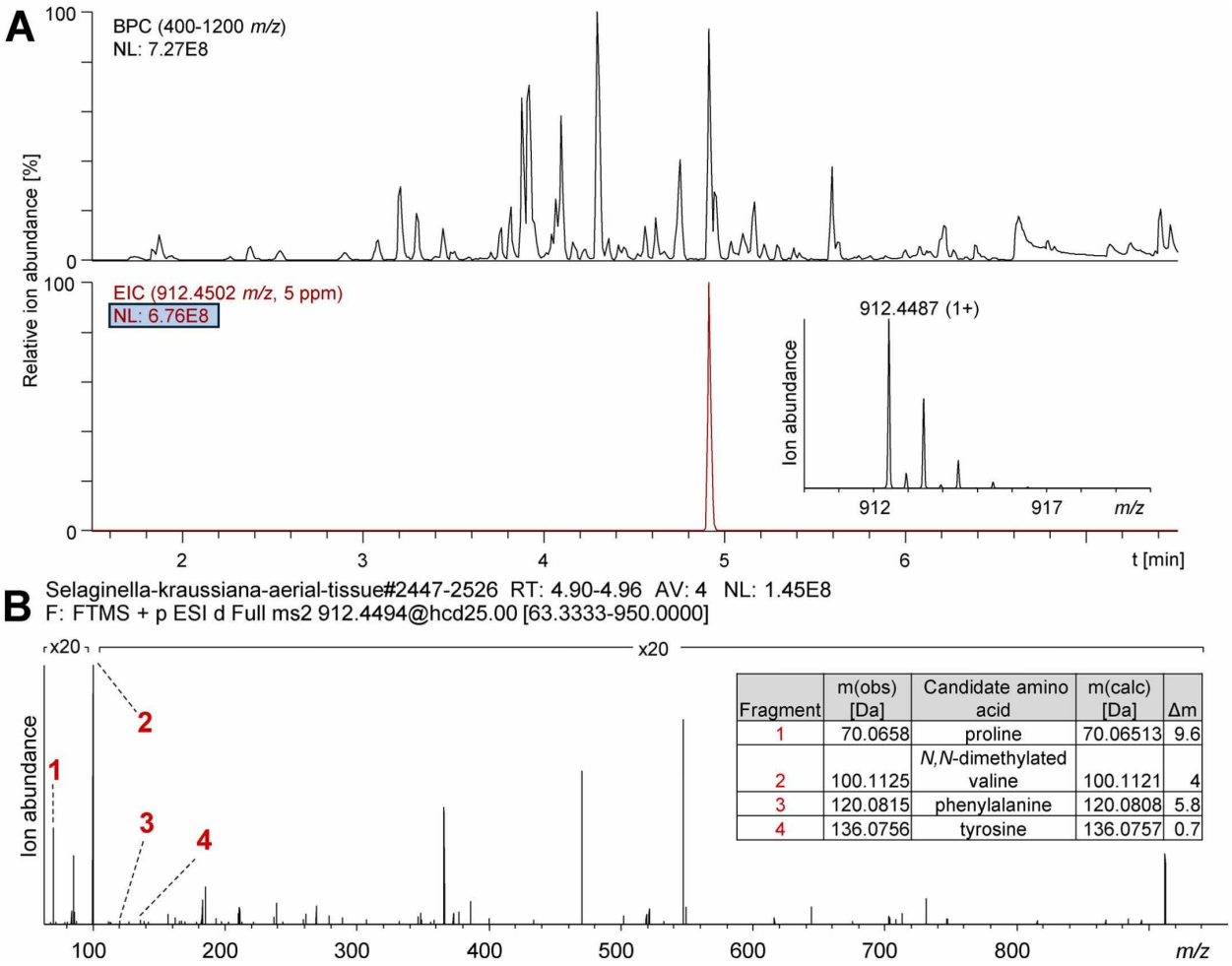
- Search the calculated mass value of candidate peptide structures in your LC-MS data.
- Check the mass spectrum of a candidate peptide EIC peak for correct charge state.

Example: m(obs) = 912.4487 (1+) (**Figure 1.6A**)

- Check the tandem mass spectrum of any detected peptide mass signal for structural feature of its core peptide such as amino acid iminium ion masses (**Table 1.3**). Iminium ions are usually only detected for non-cyclized residues.

Table 1.3 Iminium ion masses of proteinogenic amino acids.

Amino acid	Iminium ion	Monomethylated	Dimethylated
Alanine	44.0495	58.0651	72.0808
Aspartate	88.0393	102.0550	116.0706
Glutamate	102.0550	116.0706	130.0863
Phenylalanine	120.0808	134.0964	148.1121
Glycine	30.0338	44.04948	58.06513
Histidine	110.0713	124.0869	138.1026
Leucine	86.0964	100.1121	114.1277
Isoleucine	86.0964	100.1121	114.1277
Lysine	101.1073	115.1230	129.1386
Methionine	104.0529	118.0685	132.0842
Asparagine	87.0553	101.0709	115.0866
Proline	70.0651	84.0808	98.09643
Glutamine	101.0709	115.0866	129.1022
Arginine	129.1135	143.1291	157.1448
Serine	60.0444	74.0600	88.0757
Threonine	74.0600	88.0757	102.0913
Valine	72.0808	86.0964	100.1121
Tryptophan	159.0917	173.1073	187.1230
Tyrosine	136.0757	150.0913	164.1070
Pyroglutamate	84.0444		



Fragment	m(obs) [Da]	Candidate amino acid	m(calc) [Da]	Δm
1	70.0658	proline	70.06513	9.6
2	100.1125	*N,N*-dimethylated valine	100.1121	4
3	120.0815	phenylalanine	120.0808	5.8
4	136.0756	tyrosine	136.0757	0.7

Figure 1.6 Detection of selanine A in methanolic extract of *S. kraussiana* aerial tissue.
(A) Base peak chromatogram (BPC), extracted ion chromatogram (EIC), and mass spectrum of a putative burpitide corresponding to core peptide VLFYPSY (dimethylated and bicyclic). The blue box shows the ion intensity (NL = normalized level) of the putative burpitide in the extract EIC. The value indicates if a peptide analyte can be isolated. (B) Tandem mass spectrum of putative burpitide corresponding to core peptide VLFYPSY which includes four masses of proteinogenic amino acids. Δm is the mass error in [ppm].

Example: *N,N*-Me$_2$Val, Phe, Tyr, Pro iminium ions detected (**Figure 1.6B**)

1.2.4.2 Biosynthetic Confirmation of the Predicted Burpitide Precursor Peptide

A connection of a burpitide analyte with its predicted fused burpitide cyclase can be confirmed by heterologous expression of the cyclase gene in *Nicotiana benthamiana* via agroinfiltration and subsequent metabolic profiling of the transgenic *N. benthamiana* leaves for the target burpitide[41]. In some cases, further modifications such as N- and C-terminal modifications are required to match a burpitide product to a source plant analyte.

1.2.4.2.1 Tobacco Cultivation

Cultivation of >4-week-old *N. benthamiana* plants is required for transient expression of a target burpitide cyclase for a biosynthetic confirmation experiment. *N. benthamiana* seeds can be obtained from the USDA Germplasm database (Accession: PI 555478).

- Add *N. benthamiana* seeds to damp soil (Premier Horticulture Pro Mix HP High Porosity with Mycorise mixed with Organic Perlite Planting Soil Additive) at ~0.5cm depth in a 3.5" square x 3" deep plastic pots. Cover loosely with soil. Add pots to 10"x20" black plastic propagation tray and cover with plastic humidity domes (10"x20").
- Cultivate plants under 16h light/8h dark cycle at room temperature (20-25 °C) with full spectrum growth lights (e.g. Spider Farmer SF2000).
- After 1-2 weeks, transfer single seedlings at ~1 cm size and two leaf-stage with tweezers to individual 3.5" square x 3" deep plastic pots with damp soil added with Osmocote slow-release fertilizer. Add 18 pots to 10"x20" black plastic propagation tray and cover with plastic humidity domes (10"x20").
- Once seedlings begin to grow in individual pots (~1-2 additional weeks), remove the plastic humidity domes and change the pot number to ~12 pots per tray to improve soil aeration. Water as needed and enable topsoil to dry out before rewatering seedlings. Once plant leaves reach>3 cm diameter, they can be used for agroinfiltration.
- For large *N. benthamiana* cultivation for scaled peptide purification, grow ~50 plants and infiltrate them continuously from ~4-6-week age when stem growth increases.

1.2.4.2.2 Cloning Wild-type Burpitide Cyclase from Source Plant into pEAQ-HT

For biosynthetic confirmation, the target burpitide cyclase gene can be cloned from the source plant tissue, from which the burpitide was detected by LC-MS, into *N. benthamiana* expression vector pEAQ-HT[34].

- Isolate total RNA from target plant tissue with plant-specific RNA isolation kit, for example QIAGEN Plant RNeasy Mini kit (Cat#: 74904). Test both lysing buffers (RLC and RLT) for highest RNA yields and quality. Source plant tissues collected for RNA isolation should be frozen immediately in liquid nitrogen for subsequent RNA isolation or on dry ice before storage at −80 °C. Avoid thawing plant material for RNA isolation.

Example: Total RNA was isolated from aerial tissues of *S. kraussiana.*

- Transform total RNA of target plant tissue into cDNA, for example with Maxima H Minus First Strand cDNA Synthesis Kit (Thermo, Cat#: K1652).
- Clone target burpitide cyclase gene with forward and reverse cloning primers from genome sequence gene into tobacco expression vector pEAQ-HT via Gibson cloning (Gibson et al., 2009). Add Gibson cloning overhangs for pEAQ-HT to primers (fwd-primer-overhang: TGCCCAAATTCGCGACCGGT, rev-primer-overhang: CCAGAGTTAAAGGCCTCGAG). Linearize pEAQ-HT with *Age*I and *Xho*I for Gibson cloning (NEB restriction enzymes)[72].

  Example: The following gene sequence was obtained by tblastn search from the *S. kraussiana* genome (v2, CDS, https://sk.ccgg.fun/blast)[70]:

>SK09G15410.mRNA1

ATGGCTCAATCACTCATTCTATTGTTTCTGATTGTTATGGGATGTTATGATGTTGGAGCGATTGAGAAACGCCTTCC
TTGGCACTATGAACACGCCGATAACATTTTATTCTATCCGAGCTATGCTGCTAAGAAAGCCGTTCCTGAGGATCCAA
GCCAAGCTGATGCCGATAACGTTTTATTCTATCGGAGCTATGCTGCTAAGAAAGCCGTTCCTGCATCAGTCGAGGAT
CCAAGCCATGATGCCGATAACGTTTTATTCTATCCGAGTTATGTTGCTAAGAAAGCCGTTCCTGCATCAGTCGAGGA
TCCAAGCCAAGCTGATGCCGATAACTTTTTACTCTATCCGTATGCTGCTAAGAAAGCCGTTCCTGCATCAGTCGAGG
ATCCAAGCCATGATGCCGATAACGTTTTATTCTATCCGAGTTATGTTGCTAAGAAAGCCGTTCCTACATCAGTCGAG
GATCCAAGCCAAGCTGATGCCGATAACGTTTTATTCTATCCGAGCTATGCTGCCAGACATAAGGCTCCAAGCACCAT
GGATCACGATGCTACACACAATGCAGCACACAACAAGGAGGATCCAAGTCAAGCTGATGCCGATAACGTTTTATTCT
ATCCGAGCTATGCTGCCAGACCTAAGGCTCCAAGCACCATGGATCACGATGCTACACACAATGCAGCACACAACAAG
GAGGATCCAAGTCAAGCTGATGCCGATAACGTTTTATTCTATCCGAGCTATGCTGCCGGACCTAAGGCTCCAAGCAC
CATGGATCACGATGCTACACACAATGCAGCACACAACAAGGAGGATCCAAGTCAAGCTGATGCCGATAACGTTTTAT
TCTATCCGAGCTATGCTGCCAGACCTAAGGCTCCAAGTACCATGGATCACGATGCTACACACAATGCAGCACAAAAC
ATGCATCACCAGAAAGGAGCACTCCTCTTCTTCAGAATGAATACACTAAAGGTTGGTGGTGAAGTAGTTATACCCTC
ACTAGCAAGCCATGCTCTTGGTGGACGAAAGCTGATGACTCCAAGACTAGAGGCCATGCTTGGAAGCATGGAATTGC
CTCGGCTTCTCTCAGTTTTGAAGATTCCTCTAGACTCTACACTCGCTCGCAAAAGCAAGACCAACTTAGGGAACTGC
CGATCTCCACCTTTACAAGGAGAGAAAAAGGCTTGCGTCTCCTCAATTAGAAGCATGACAAACTTTGCACGTTCCGT
CTTGGAAGACAAGAAGCCCCTCGAACATCTTAATCCCTCCCGTCCGGTGTCTTCTCCAGAGCATTTTAAGATAATGG
ACGTGAAAGTGGTCACTGAGAACTCCGTGGTCACATGTCATCCTATGGTGTTCCCCTATGCTCTTTACATGTGCCAC
TTCGTTCCCAAGAGTGTCCCCATCAAAGTCACCTTGCAGGACGATCACGACAAACTGGTTGTAGTACCTGTCATGTG
CCATATGGACACGTCTGAGTTCGACCCCTCCCACCTCTCGTTCAAAATCCTTAACACAAAACCTGGGGAAGCAGAAA
TGTGTCACTGGATGCCCAACTCCCATATCATGTGGTACACTTCGGACGGAAAGACGCGCGATGTTCTTTGA

(Gibson overhangs are underlined)

Example: The following cloning primers were designed to clone the target burpitide cyclase gene into linearized pEAQ-HT via Gibson cloning:

SkrBURP-pEAQ-HT-fwd: TGCCCAAATTCGCGACCGGTATGGCTCAATCACTCATTCTATTG

SkrBURP-pEAQ-HT-rev: CCAGAGTTAAAGGCCTCGAGTCAAAGAACATCGCGCGTC

Example: The following sequence was cloned from cDNA of a *S. kraussiana* plant:

>MW570735.1 Selaginella kraussiana BURP (BURP) mRNA, complete cds

ATGGCTCAATCACTCATTCTATTGTTTCTGATTGTTATGGGATGTTATGATGTTGGAGCGATTGAGCGTGTTCATAG

TGCCAAGGAAGGCGTTTCTGAAGCAATGAATGATCAGGGGCATCCTACCATTGCCAACAAAGCCGTTCCTGCATCAG

TCGAGGATCCAAGCCAAGCTGATGCCGATAACATTTTATTATATCCGAGCTATGCTGCTAAGAAAGCCGTTCCTGAG

GATCCAAGCCAAGCTGATGCCGATAACGTTTTATTCTATCGGAGCTATGCTGCTAAGAAAGCCGTTCCTGCATCAGT

CGAGGATCCAAGCCATGATGCCGATAACGTTTTATTCTATCCGAGTTATGTTGCTAAGAAAGCCGTTCCTGCATCAG

TCGAGGATCCAAGCCAAGCTGATGCCGATAACTTTTTACTCTATCCGTATGCTGCTAAGAAAGCCGTTCCTGCATCA

GTCGAGGATCCAAGCCATGATGCCGATAACGTTTTATTCTATCCGAGTTATGTTGCTAAGAAAGCCGTTCCTGCATC

AGTCGAGGATCCAAGCCAAGCTGATGCCGATAACTTTTTATTCTATCCGTATGCTGCTAAGAAAGCCGTTCCTGCAT

CAGTCGAGGATCCAAGCCATGATGCCGATAACGTTTTATTCTATCCGAGTTATGTTGCTAAGAAAGCCGTTCCTGCA

TCAGTCGAGGATCCAAGCCATGATGCCGATAACGTTTTATTCTATCCGAGTTATGTTGCTAAGAAAGCCGTTCCTGC

ATCAGTCGAGGATCCAAGCCAAGCTGATGCCGATAACGTTTTATTCTATCCGAGCTATGCTGCCAGACCTAAGGCTC

CAAGTACCATGGATCACGATGCTACACACAATGCAGCACAAAACATGCATCACCAGAAAGGAGCACTCCTCTTCTTC

AGAATGAATACACTAAAGGTTGGTGGTGAAGTAGTTATACCCTCACTAGCAAGCCATGCTCTTGGTGGACGAAAGCT

GATGACTCCAAGACTAGAGGCCATGCTTGGAAGCATGGAATTGCCTCGGCTTCTCTCAGTTTTGAAGATTCCTCTAG

ACTCTACACTCGCTCGCAAAAGCAAGACCAACTTAGGGAACTGCCGATCTCCACCTTTACAAGGAGAGAAAAAGGCT

TGCGTCTCCTCAATTAGAAGCATGACAAACTTTGCACGTTCCGTCTTGGAAGACAAGAAGCCCCTCGAACATCTTAA

TCCCTCCCGTCCGGTGTCTTCTCCAGAGCATTTTAAGATAATGGACGTGAAAGTGGTCACTGAGAACTCCGTGGTCA

CATGTCATCCTATGGTGTTCCCCTATGCTCTTTACATGTGCCACTTCGTTCCCAAGAGTGTCCCCATCAAAGTCACC

TTGCAGGACGATCACGACAAACTTGTTGTAGTACCTGTCATGTGCCATATGGACACGTCTGAGTTCGACCCCTCCCA

CCTCTCGTTCAAAATCCTTAACACAAAACCTGGGGAAGCAGAAATGTGTCACTGGATGCCCAACTCCCATATCATGT

GGTACACTTCGGACGGAAAGACGCGCGATGTTCTTTGA

- Sequence cloned burpitide cyclase in pEAQ-HT via Sanger sequencing (for example Azenta/Genewiz Inc., suggested forward sequencing primer: CTTCTGCTTGACGAGGTATTG, suggested reverse sequencing primer: AACCGCTCACCAAACATAG) or Nanopore sequencing (for example: Plasmidsaurus Inc.).

1.2.4.2.3 Synthetic Gene Design and Cloning

If a live source plant is not available or no mRNA can be isolated for burpitide cyclase cloning, the target burpitide cyclase gene can be synthesized as a gene fragment with Gibson overhangs as described above and cloned into linearized pEAQ-HT. For burpitide cyclases with multiple core peptide motifs, design a burpitide cyclase with a single core peptide for biosynthetic confirmation if gene synthesis is unacceptable ('complex') for commercial gene synthesis. For this design, use the 5'-gene sequence to the first core peptide, the sequence of a target core peptide and the 3'-gene sequence from the last core peptide to the 3'-gene end. Clone synthetic genes into pEAQ-HT like wild-type genes above.

Example: The truncated version of SkrBURP with one core peptide of VLFYPSY consisted of the N-terminal protein sequence until the first core peptide was used (bold in SkrBURP protein sequence below), the core peptide VLFYPSY (red core peptide in SkrBURP protein sequence below), and the C-terminal part of SkrBURP after the last core peptide (underlined in SkrBURP protein sequence below). The corresponding nucleotide sequence (black except core peptide in red) with its amino acid sequence (blue except core peptide in red) is shown below as SkrBURP-1xVLFYPSY.

>SkrBURP

MAQSLILLFLIVMGCYDVGAIERVHSAKEGVSEAMNDQGHPTIANKAVPASVEDPSQADADN*ILLYPSYAAKKAVPE*
DPSQADADNVLFYRSYAAKKAVPASVEDPSHDADNVLFYPSYVAKKAVPASVEDPSQADADNFLLYPYAAKKAVPAS
VEDPSHDADNVLFYPSYVAKKAVPASVEDPSQADADNFLFYPYAAKKAVPASVEDPSHDADNVLFYPSYVAKKAVPA
*SVEDPSHDADNVLFYPSYVAKKAVPASVEDPSQADADN***VLFYPSY**AARPKAPSTMDHDATHNAAQNMHHQKGALLFF
RMNTLKVGGEVVIPSLASHALGGRKLMTPRLEAMLGSMELPRLLSVLKIPLDSTLARKSKTNLGNCRSPPLQGEKKA
CVSSIRSMTNFARSVLEDKKPLEHLNPSRPVSSPEHFKIMDVKVVTENSVVTCHPMVFPYALYMCHFVPKSVPIKVT
LQDDHDKLVVVPVMCHMDTSEFDPSHLSFKILNTKPGEAEMCHWMPNSHIMWYTSDGKTRDVL

>SkrBURP-1xVLFYPSY

ATGGCTCAATCACTCATTCTATTGTTTCTGATTGTTATGGGATGTTATGATGTTGGAGCGATTGAGCGTGTTCATAGTGCCAAGGAAGGC
M A Q S L I L L F L I V M G C Y D V G A I E R V H S A K E G
GTTTCTGAAGCAATGAATGATCAGGGGCATCCTACCATTGCCAACAAAGCCGTTCCTGCATCAGTCGAGGATCCAAGCCAAGCTGATGCC
V S E A M N D Q G H P T I A N K A V P A S V E D P S Q A D A
GATAACGTTTTATTCTATCCGAGTTATGCTGCCAGACCTAAGGCTCCAAGTACCATGGATCACGATGCTACACACAATGCAGCACAAAAC
D N V L F Y P S Y A A R P K A P S T M D H D A T H N A A Q N
ATGCATCACCAGAAAGGAGCACTCCTCTTCTTCAGAATGAATACACTAAAGGTTGGTGGTGAAGTAGTTATACCCTCACTAGCAAGCCAT

M H H Q K G A L L F F R M N T L K V G G E V V I P S L A S H

GCTCTTGGTGGACGAAAGCTGATGACTCCAAGACTAGAGGCCATGCTTGGAAGCATGGAATTGCCTCGGCTTCTCTCAGTTTTGAAGATT

A L G G R K L M T P R L E A M L G S M E L P R L L S V L K I

CCTCTAGACTCTACACTCGCTCGCAAAAGCAAGACCAACTTAGGGAACTGCCGATCTCCACCTTTACAAGGAGAGAAAAAGGCTTGCGTC

P L D S T L A R K S K T N L G N C R S P P L Q G E K K A C V

TCCTCAATTAGAAGCATGACAAACTTTGCACGTTCCGTCTTGGAAGACAAGAAGCCCCTCGAACATCTTAATCCCTCCCGTCCGGTGTCT

S S I R S M T N F A R S V L E D K K P L E H L N P S R P V S

TCTCCAGAGCATTTTAAGATAATGGACGTGAAAGTGGTCACTGAGAACTCCGTGGTCACATGTCATCCTATGGTGTTCCCCTATGCTCTT

S P E H F K I M D V K V V T E N S V V T C H P M V F P Y A L

TACATGTGCCACTTCGTTCCCAAGAGTGTCCCCATCAAAGTCACCTTGCAGGACGATCACGACAAACTTGTTGTAGTACCTGTCATGTGC

Y M C H F V P K S V P I K V T L Q D D H D K L V V V P V M C

CATATGGACACGTCTGAGTTCGACCCCTCCCACCTCTCGTTCAAAATCCTTAACACAAAACCTGGGGAAGCAGAAATGTGTCACTGGATG

H M D T S E F D P S H L S F K I L N T K P G E A E M C H W M

CCCAACTCCCATATCATGTGGTACACTTCGGACGGAAAGACGCGCGATGTTCTTTGA

P N S H I M W Y T S D G K T R D V L stop

1.2.4.2.4 Agroinfiltration of pEAQ-HT-burpitide-cyclase

Infiltration of *N. benthamiana* with *Agrobacterium tumefaciens* LBA4404 including cloned burpitide cyclase in pEAQ-HT can be achieved by syringe infiltration or vacuum infiltration. The protocol is well-established, and a brief one is described as follows[34,35]:

- Transform electrocompetent *A. tumefaciens* LBA4404 with pEAQ-HT-plasmid with cloned burpitide cyclase (voltage: 2.4 kV) via electroporation cuvette (Biorad Gene Pulser®/MicroPulser™ Electroporation Cuvettes, 0.1 cm gap: Cat#: 1652089) in Biorad® MicroPulser Electroporator.
- Plate transformant cells on YM-agar and incubate for 48 h at 30 °C. YM-agar is made from 0.4 g yeast extract, 10 g D-mannitol, 0.2 g magnesium sulfate heptahydrate, 0.1 g sodium chloride, 0.5 g potassium phosphate trihydrate, dibasic, add 1 L deionized water, adjust to pH 7, add 15 g agar; autoclave; after autoclaving add kanamycin (50 µg/mL), streptomycin (100 µg/mL) and rifampicin (50 µg/mL).
- Inoculate 5 mL liquid YM medium (same as YM-agar above, however without agar) with kanamycin, streptomycin and rifampicin (same concentration as YM-agar above) with one transformation colony in a 15 mL conical tube. Incubate for 24-36 h at 220 rpm and 30 °C on a shaking incubator. *A. tumefaciens* is a facultative anaerobe and requires aeration.

- Inoculate 25 mL liquid YM medium with kanamycin, streptomycin and rifampicin as before with the 5 mL starter culture in a 50 mL Bio-Reaction tube (CellTreat, Cat#: 229475) or sterile flask and incubate for 24 h at 220 rpm and 30 °C on a shaking incubator.
- Centrifuge *A. tumefaciens* cells for 30 min at 3000 × *g*.
- Resuspend in MMA medium to final OD_{600} of 0.8. MMA medium: 100 mM MES (pH 5.6), 10 mM magnesium chloride, 150 µM acetosyringone (synonym: 3′,5′-dimethoxy-4′-hydroxyacetophenone; dissolve acetosyringone in DMSO before addition to MMA). Incubate resuspension for at least 30 min at room temperature before infiltration.
- Place *N. benthamiana* plants into the shade 1-2 h before infiltration. Infiltrate *N. benthamiana* leaves on the bottom side via a sterile 1 mL syringe. Cultivate *N. benthamiana* plants as described above.

1.2.4.2.5 Metabolic Profiling and Comparison of Source Plant Methanolic Extract

Collect *N. benthamiana* leaves seven days after infiltration and prepare and analyze leaf samples by LC-MS as described above. Compare extracted ion chromatograms (EIC) of methanolic extracts of the source plant and a transgenic *N. benthamiana* leaf expressing the target burpitide cyclase for biosynthetic confirmation of its burpitide (**Figure 1.7**).

Example: An analyte matching the mass of the bicyclic core peptide VLFYPSY but without the predicted two *N*-methyl groups (desmethyl-selanine A in **Figure 1.7**) was detected in methanolic extract of *N. benthamiana* expressing wild-type SkrBURP and SkrBURP-1xVLFYPSY (truncated SkrBURP with one core peptide of VLFYPSY). The desmethyl-selanine A analyte was matched to the VLFYPSY burpitide detected in methanolic extracts (selanine A) by reductive dimethylation of the desmethyl-selanine A analyte after an HPLC concentration step (**Figure 1.7**, see Section 5.2.3.1 for HPLC method)[73,41]. Herein, desmethyl-selanine A-containing 1. preparative HPLC fractions were redissolved in 100 µL 100 mM sodium acetate buffer (pH 6, 100 µL volume). 51 µL peptide solution were added with 3 µL formaldehyde solution (4 % stock solution) and 3 µL sodium cyanoborohydride solution (260 mM stock solution). After 5 min incubation at 37 °C, 3 µL ammonium hydroxide (4 % stock solution) were added and the reaction mixture was analyzed by LC-MS[73].

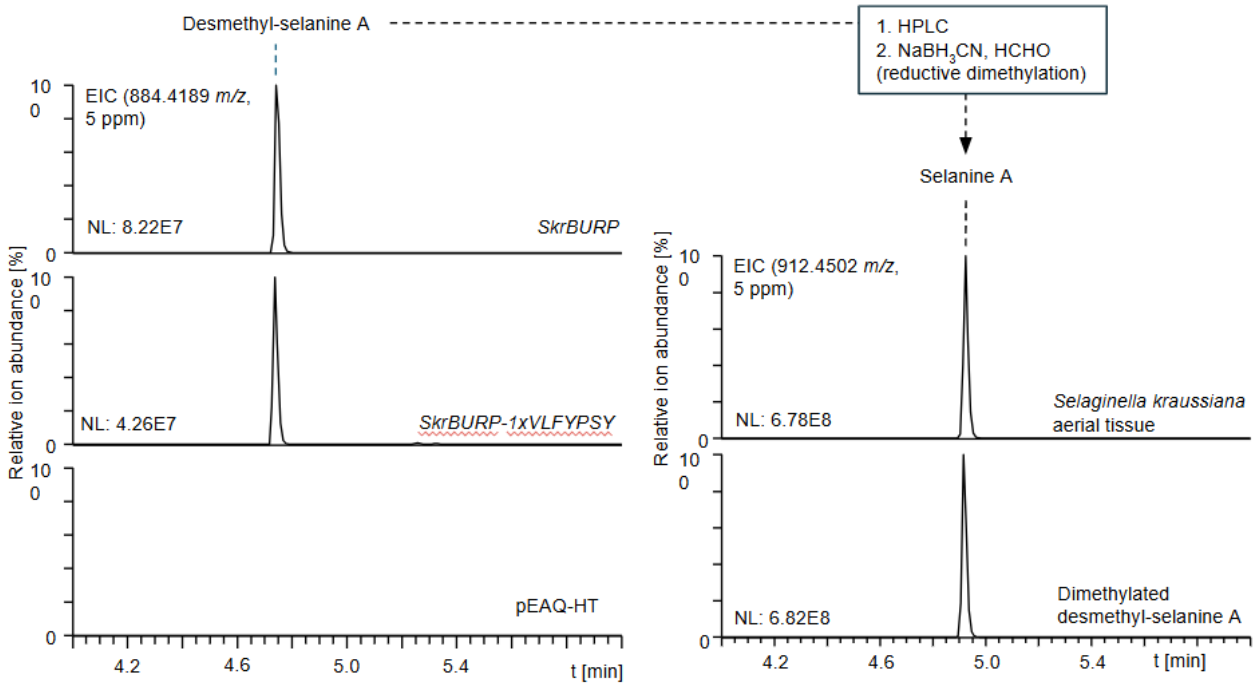


Figure 1.7 Metabolic profiling of *S. kraussiana* and heterologous SkrBURP expression.
Detection of desmethyl-selanine A analyte by transient expression of SkrBURP (wild-type and 1xVLFYPSY) which matches the predicted burpitide with core peptide VLFYPSY from *S. kraussiana* after HPLC purification and reductive dimethylation.

1.2.4.3 Characterization of a Plant Source for Burpitide Isolation

Isolation of a burpitide for structure elucidation requires the identification of a plant source with sufficient quantities and purity of the target compound. It is recommended to purify from plants – either source plant or transgenic *N. benthamiana* – with EIC values of the target peptide analyte >5e8 based on the described LC-MS chemotyping protocol for milligram quantity peptide isolation (**Figure 1.6**).

Example: The EIC of selanine A in source plant *S. kraussiana* by peptide chemotyping is 6.8e8 (**Figure 1.6**). In addition, the plant was commercially available to obtain 1 kg of fresh plant material. Therefore, *S. kraussiana* was chosen as a source for isolation.

1.2.5 Characterization of Burpitide Cyclase Product

After identifying a plant for the isolation of the target burpitide of the genome mining experiment, the peptide is extracted from plant material, purified and analyzed in its structure by NMR and Marfey's analysis.

1.2.5.1 Scaled N. benthamiana expression

For isolation of a target peptide from transgenic *N. benthamiana* expressing a target burpitide cyclase at recommended EIC value (>5e8), 1 kg of *N. benthamiana* leaves should be infiltrated and collected as fresh weight for subsequent peptide isolation.

1.2.5.2 Purification

Burpitide purification from plant material usually comprises extraction, liquid chromatography, and high-performance liquid chromatography. Selanine A purification from source plant *S. kraussiana* is described below as an example towards structure elucidation of a burpitide targeted by genome mining.

1.2.5.2.1 Extraction

The first separation step is extraction of ground source plant material or transient *N. benthamiana*. Extraction can reduce the metabolic complexity of the sample by partitioning with solvents of increasing hydrophilicity such as hexanes, ethyl acetate, and *n*-butanol. The following protocol is adapted from a described literature extraction of moroidin from seeds of source plant Celosia argentea[74].

Grinding of soft plant tissues:

- Freeze the plant material in −80 °C freezer overnight.
- Grind a few pieces of dry ice in a food processor (Hamilton Beach, Cat#: 70730).
- Add frozen plant material to the food processor and top off with a few pieces of dry ice.
- Grind for 1 min at top speed.

Grinding of woody and hard tissues:

- Dry woody or seed tissue in lyophilizer.
- Cut plant material into pieces smaller than 10 cm with a scissor or wood cutter.

- Grind the dry plant material in a spice grinder (Model: YF3-1/110 V, Health Life LTD) in a fume hood.

All plant samples continued:

- Transfer the ground plant material to a 2.8 L Erlenmeyer flask via a large plastic powder funnel. Depending on plant material amount, 1-2 kg of plant material can be distributed among six 2.8 L flasks.
- Add 1-1.5 L of methanol (ACS or HPLC grade) to each flask so that ground plant material is covered.
- Add flasks to shaking incubator at 37 °C and 160 rpm for overnight extraction.
- Vacuum-filter methanolic extract through a silica filter into a 2 L vacuum-filtration flask.
- Rotovaporate crude methanol extract to dryness.
- Resuspend the crude methanol extract in 2 L deionized water.
- Partition the aqueous resuspension with hexanes (ACS or HPLC grade, 1:1, v/v) two to three times until green color is removed from the aqueous phase.
- Partition the aqueous resuspension with ethyl acetate (ACS or HPLC grade, 1:1, v/v) two times.
- Extract the aqueous resuspension with *n*-butanol (ACS or HPLC grade, 1:1, v/v) two times.
- Check the presence of the target burpitide in the ethyl acetate and *n*-butanol extract by LC-MS.

<u>*Example:*</u> *Selanine A* was mainly present in the n-butanol fraction.

- Rotovaporate peptide-containing fraction to dryness.

1.2.5.2.2 LH20 Liquid Chromatography

The second separation step is liquid chromatography with a Sephadex LH20 resin (Cytvia) under gravity flow.

- Resuspend 200 g of Sephadex LH20 resin (Cytvia) in 2 L of 10 % methanol and load a glass column (Chemglass, 8 × 60 cm column, 45/50, Cat#: CG118931).
- Resuspend the peptide fraction extract in ~100 mL of 10 % methanol (ACS or HPLC grade). Use a sonication water bath for resuspension if needed.
- Load the peptide fraction to the LH20 column. Run 10 % methanol (ACS or HPLC grade) as a starting mobile phase and increase the methanol concentration of the mobile phase by 10 % every 500 mL up to 100 %. Collect 10 mL fractions with a Gilson fraction collector (FC 203B model, Cat#: 171011DV) and check fractions for the presence of the target peptide by LC-MS. Suggested method:
    - LC: Injection volume 2 µL, LC – Phenomenex Kinetex® 2.6 µm C18 reverse phase 100 Å 50 × 3 mm LC column, LC gradient: solvent A – 0.1 % formic acid, solvent B – acetonitrile (0.1 % formic acid), 0 min: 5 % B, 2.5 min: min: 95 % B, 3.0 min: 95 % B, 3.1 min: 5 % B, 5.0 min: 5 % B, 0.5 mL/min
    - MS: MS – positive ion mode, Full MS: resolution 35,000, mass range 400-1,250 *m/z*, dd-MS^2: resolution 17,500, loop count 5, collision energy 25 eV, dynamic exclusion 0.5 s
- Combine fractions with the target peptide with EIC ion count peak values>1e8 (better:>1e9) and rotovaporate to dryness.

1.2.5.2.3 HPLC

The third separation step is high-performance liquid chromatography (HPLC) which usually consists of two rounds of preparative HPLC and 1–2 rounds of semipreparative HPLC to yield a pure burpitide.

1.2.5.2.3.1 Preparative HPLC

A general first preparative HPLC method for burpitide purification is:

- Injection: 1-5 mL resuspension in starting condition
- Column: Phenomenex Kinetex® 5 µm C18 100 Å LC Column 150 × 21.2 mm (Cat#: 00F-4601-P0-AX)

- Solvents: solvent A – 0.1 % TFA, solvent B – acetonitrile (0.1 % TFA)
- Gradient:

Table 1.4 Preparative HPLC gradient.

Time [min]	Solvent B [%]
0	10
1	10
36	50
39	95
42	95
42.5	10
60.1	10

- Flow rate: 7.5 mL/min

A second preparative HPLC method with a solvent B range of 20 % is usually applied to improve preparative separation with the retention concentration of solvent B in the middle of the gradient.

Example: Selanine A purification gradient of 2. preparative HPLC: 20–40 % B.

1.2.5.2.3.2 Semipreparative HPLC

A general first semipreparative HPLC method for burpitide purification is*:*

- Injection: 1–5 mL resuspended in starting condition
- Column: Phenomenex Kinetex® 5 µm C18 100 Å LC Column 250 × 10.0 mm (Cat#: 00G-4601-N0)
- Solvents: solvent A – 0.1 % TFA, solvent B – acetonitrile (0.1 % TFA)
- Gradient: The gradient is usually 5 % and starting condition is determined by peptide elution in the middle of the gradient.

Example: Selanine A purification gradient of 1. semipreparative HPLC: 32–37 % B.

Table 1.5 Semipreparative HPLC gradient (selanine A example)

Time [min]	Solvent B [%]
0	32
1	32
21	37
22	95
24.5	95
25	32
45	32

- Flow rate: 7.5 mL/min

A second semipreparative HPLC method is usually required for purification. This step can be either an isocratic step, another small gradient with acetonitrile (0.1 % TFA) or a gradient with methanol (0.1 % TFA) as solvent B. Combine all fractions with the target peptide, dry down *in vacuo* and remove residual solvent and water in a lyophilizer before NMR analysis.

Example: Selanine A purification gradient of 2. semipreparative HPLC: 30–37 % B.

1.2.5.3 NMR

NMR analysis is a key step to elucidate the structure of a target burpitide and complete the genome mining experiment.

1.2.5.3.1 NMR Sample Preparation

Common solvents for burpitide NMR analysis are DMSO-d_6 and methanol-d_4. Cyclopeptide alkaloids with decarboxylated C-termini are often analyzed in chloroform-d_3.

- Dissolve the dried burpitide sample in the target NMR solvent and load it into an NMR tube for an available NMR probe.

  Example: Selanine A was dissolved in 270 µL DMSO-d_6 and loaded in a DMSO-Shigemi tube (Cat#: DMS-005B) for analysis on a 5 mm Triple resonance inverse detection TCI cryoprobe.

1.2.5.3.2 NMR Sample Preparation

Analyze burpitide sample by 1D NMR (proton and carbon) and 2D NMR (COSY, TOCSY, HSQC, HMBC, NOESY). If the NOESY signals are weak, acquire ROESY NMR data instead.

Example: For selanine A, ROESY NMR data was acquired.

1.2.5.3.3 NMR Data Analysis

NMR data analysis of a target burpitide is informed by structural information derived from previous genome mining and chemotyping experiments such as core peptide sequence and putative post-translational modifications.

Example: For selanine A, a core peptide VLFYPSY was known and bicyclization and N-terminal dimethylation was predicted as a structural model for NMR analysis (**Section 1.2.4.1**)

Suggested steps:

- Characterize amino acid spin systems first. A good starting point are δH values of α-protons (3.5-5.5 ppm).
- Determine amino acid sequence based on HMBC and NOESY/ROESY data.
- Determine N- and C-terminal modifications, e.g. *N*-methylation, pyroglutamate or hydroxystyrylamine. *N*-methyl groups are characterized by broad singlets in the 1-3 ppm range which usually do not show COSY/TOCSY/HMBC correlations to the other amino acid protons/carbons except to each other. Pyroglutamate is characterized by a high field carbonyl group (176-180 ppm) of the precursor glutamine which formed a lactam with the N-terminal amine. The protons of the enamine double bond of hydroxystyrylamine are generally in the range of 5.8-7 ppm in DMSO-d_6 [75].

Example: The *N*-methyl groups of selanine A were singlets at 1.17 ppm and 1.24 ppm.

- Determine crosslinks of burpitides. Most crosslinks involve unactivated carbons in other amino acid side chains and they can often be identified by unusually high chemical shifts of protons and associated carbons in amino acid side chains compared to literature chemical shifts of unmodified amino acid side chains (**Figure 1.8**). In general, the chemical shift of carbons of the non-aromatic crosslinking residue in burpitides can indicate the bond type: $\delta^{13}C$ = ~75–80 ppm for C-O-bonds where O is from a phenol-hydroxy group, $\delta^{13}C$ = ~60–65 ppm for C-N-bonds where N is from an indole-amine, and $\delta^{13}C$ = ~50–55 ppm for C-C-bonds where the second carbon is an indole carbon (C6 in stephanotic acids) (**Figure 1.8**). A reference for literature NMR values of plant peptides is the review by Zhou and Tan[40].

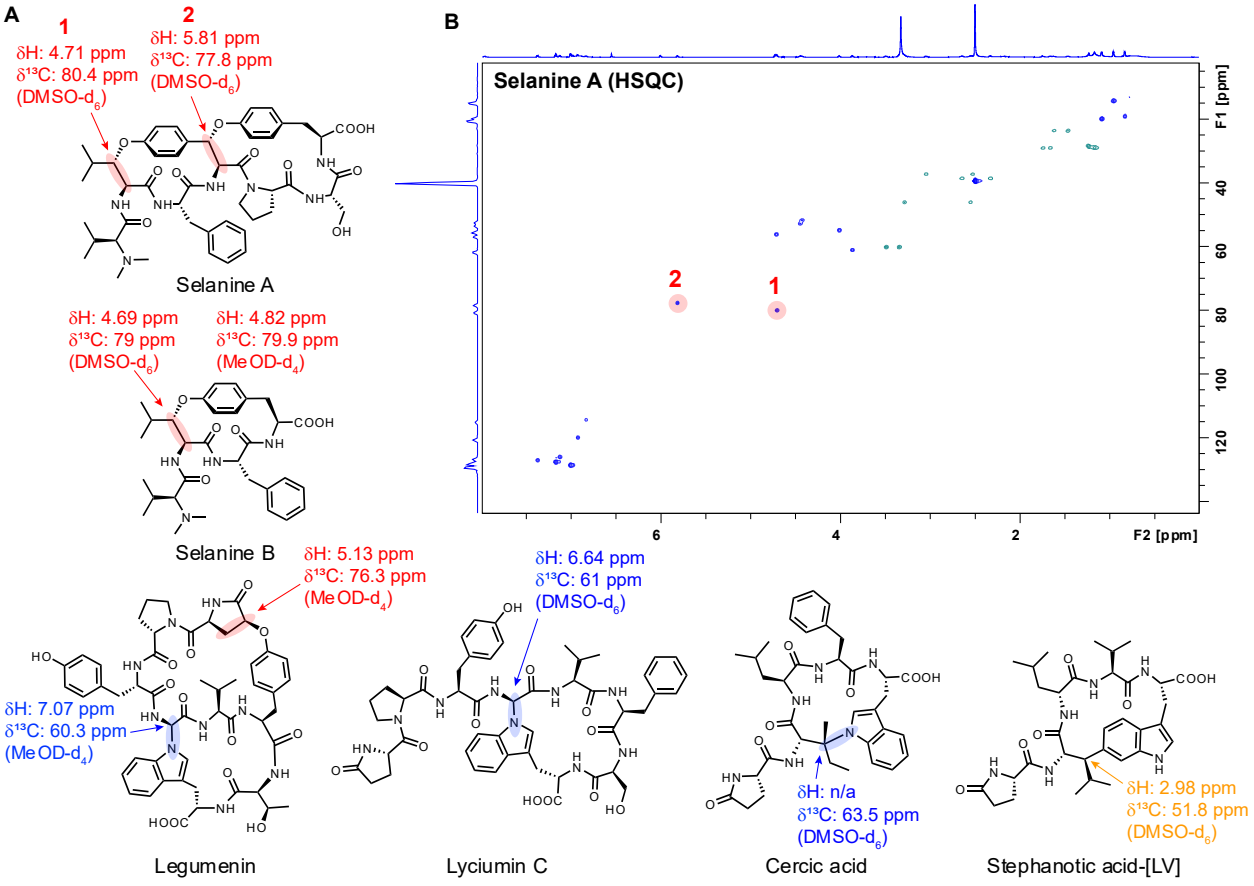


Figure 1.8 NMR chemical shifts of burpitide crosslinks.
(A) Chemical shifts of crosslinking sites in representative burpitide structures. Blue color indicates C-N-crosslinks, red color indicates C-O-crosslinks, orange color indicates C-C-crosslinks. (B) HSQC spectrum of selanine A with HSQC signals of C-O-crosslink sites highlighted in red color and numbered, 1 and 2.

- Example: In selanine A, the β-protons of leucine-2 and tyrosine-4 were shifted downfield (4.71 ppm and 5.81 ppm, respectively) and there was only one β-proton present at each of these residues indicating a connection to a heteroatom. Corresponding β-carbon shifts of leucine-2 and tyrosine-4 were 80.4 ppm and 77.8 ppm suggesting crosslinks via ether bonds at these sites. HMBC signals between leucine-2-Hβ and tyrosine-4-Hβ and tyrosine-4-C4 and tyrosine-7-C4, respectively, and ROESY signals between leucine-2-Hβ and tyrosine-4-Hβ and tyrosine-4-H3/5 and tyrosine-7-H3/5, respectively, established crosslinks between (a) leucine-2-Cβ and tyrosine-4-C4 and (b) tyrosine-4-Cβ and tyrosine-7-C4.

1.2.5.4 Stereochemical analysis

1.2.5.4.1 Marfey's analysis

Marfey's analysis is used to determine the configuration of α-carbons of amino acids in a given burpitide[76,77]. In an acidic environment, the peptide undergoes hydrolysis, resulting in the cleavage of peptide bonds and the formation of amino acid fragments. Amino acid fragments react with Marfey's reagent in a basic sodium bicarbonate environment, producing derivatives through a nucleophilic substitution reaction. D/L-amino acids are reacted with Marfey's reagent to

produce derivatives, which are then used as standards for comparison with peptide hydrolysis derivatives in HPLC analysis.

- Hydrolyze approximately 0.1 mg of purified peptide in 0.6 mL of 6 M hydrochloric acid in a sealed thick-walled reaction vessel (Cat#: P160702).
- Heat the sample at 110 °C and stir for 12 h.
- After cooling to room temperature, concentrate the hydrolysate to dryness under nitrogen gas and redissolve in 100 µL water. Transfer to a 2 mL LC-MS sample glass vial.
- Add 100 µL 1 M sodium bicarbonate and 100 µL of a 1 % (w/v) acetone solution of Marfey's reagent 1-fluoro-2,4 dinitrophenyl-5-l-alanine amide (L-FDAA) to the solution.

- Incubate the reaction mixture at 40 °C for 1 h and then quench the reaction by adding 100 µL of 1 M hydrochloric acid.
- Filter the sample with a syringeless 0.2 µm LC-MS filter (Cytvia).
- Prepare 50 mM L-FDAA-amino acid derivative standards dissolved in water for each candidate amino acid in the target burpitide. 50 µL of each amino acid was treated with 20 µL of 1 M sodium bicarbonate and 100 µL of 1 % l-FDAA (1 % acetone solution) at 40 °C for 1 h, respectively. After the reaction, the solution was quenched with 20 µL of 1 M hydrochloric acid and diluted with 810 µL acetonitrile. Filter the standard sample with syringeless 0.2 µm LC-MS filter.
- Analyze the burpitide hydrolysate and the amino acid standards by LC-MS with the following LC-MS settings:

  LC: injection volume 2 µL; LC, Phenomenex Kinetex® 2.6 µm C18 reverse phase 100 Å 150 × 3 mm LC column; LC gradient: solvent A, 0.1 % formic acid; solvent B, acetonitrile (0.1 % formic acid); 0 min, 10 % B; 40 min, 60 % B; 41 min, 95 % B; 44 min, 95 % B; 45 min, 10 % B, 50 min, 10 % B, 0.5 mL/min;

  MS: positive ion mode; full MS, resolution 70,000, mass range 100–1,000 *m/z*.
- Compare retention times between l-FDAA-amino acid standards of d- and l-amino acids with corresponding l-FDAA-amino acid mass peaks in the hydrolysate to determine Cα-configurations of amino acids in a target burpitide. The mass shift of the amino acid mass compared to FDAA-derivatized amino acid mass is + 252.0494 Da.

1.2.5.4.2 Stereochemical analysis of macrocyclization sites

Stereochemical analysis of burpitides often involves characterization of configurations at cyclization sites because chiral centers are formed via activation of unactivated carbons in amino acid side chains or at glycine-α-carbons. Stereochemical analysis at cyclization sites can be investigated by:

1.2.5.4.2.1 J coupling constants

J coupling constants of hydrogens at burpitide crosslink sites can provide insights into the determination of the relative configurations at these sites in comparison to *J* coupling constants at corresponding positions of similar compounds with characterized configuration. Similar *J* coupling constants at burpitide crosslinks and a characterized compound reflect identical relative configurations including similar dihedral angles between hydrogens at the crosslinking site and neighboring hydrogens.

Example: In selanine A, the crosslink positions of both rings were determined to be at the Cβ positions based on 1D and 2D NMR data. Specifically, ^{3}J(Hα,Hβ) in the Leu2 fragment was found to be 7.8 Hz, and in the Tyr4 fragment, ^{3}J(Hα,Hβ) was 6.5 Hz. Comparison with literature values for similar compounds indicates that the ^{3}J(Hα,Hβ) for the l-erythro configuration typically falls within the range of 6.4–9.6 Hz, whereas for the L-threo configuration, it is in the range of 0–2 Hz[78–81]. Based on this, the crosslink positions in selanine A were assigned as l-erythro. Additionally, the Cα configurations of Leu2 and Tyr4 were both determined to be l-type, leading to the conclusion that both Leu-Cβ and Tyr-Cβ are in the *S*-configuration.

1.2.5.4.2.2 NOE/ROE data in comparison to hydrogen-distances in geometry-optimized models

Another approach to determine the stereochemistry at crosslinking sites is the comparison of geometry-optimized models with NOE/ROE data where a pre-optimized model is generated by MM2/MMFF94 calculations and then with B3LYP-D3/6–31 +G(d,p) function via Gaussian[82,68,83–85]. NOESY/ROESY signals are often assumed to be absent for protons separated by more than 5 Å for the interpretation of these candidate stereoisomer models[85].

1.2.5.4.2.3 Comparison of burpitide cyclase products to authentic standards

The stereochemistry of burpitides can also be inferred by generation of cyclases of target burpitides with core peptides of a known burpitide, for which an authentic standard can be obtained. For example, a stephanotic acid-type burpitide cyclase from *Cercis canadensis,* which had multiple stephanotic acid core peptides (QLxxW, x is any amino acid), was generated as a 1xQLIVW construct, expressed in *N. benthamiana* and its burpitide product in methanolic *N.*

benthamiana extract was compared to an authentic standard of stephanotic acid (core peptide: QLIVW) to confirm the stereochemistry of the target burpitide cyclase products as the stereochemistry of stephanotic acid[41]. The LC-MS method was the same as for peptide chemotyping before.

1.2.5.4.2.4 Small molecule X-ray crystallography

A method to determine absolute burpitide configuration is X-ray crystallography. Several burpitides such as moroidin and a legumenin-type bicyclic peptide Glycnsisitin A were characterized in their 3D structure by crystallization and subsequent diffraction analysis[86,87].

1.2.5.4.3 Perspective

The field of plant natural products is experiencing an explosion of genome sequencing enabling scaled bioinformatic approaches for natural product discovery similar to microbial natural product chemistry in the previous two decades[88,47,89,90,50]. RiPPs are an approachable natural product class of genomics-guided discovery since their structure is directly encoded in the genome sequence within a precursor peptide gene. Among plant RiPPs, burpitides are an exciting target for genome mining due to high structural diversity formed by burpitide cyclases and the co-occurrence of core peptides within fused burpitide cyclases[41]. The protocol outlined in this chapter is a blueprint for burpitide discovery from fused burpitide pathways utilizing annotated plant genomes. Future directions of burpitide genome mining are streamlined platforms for plant genome mining of burpitides and automated prediction of tailoring steps by combination of genomics with metabolomics data[33,91,92].

1.3 Dissertation Overview

This dissertation systematically investigates the discovery of bioactive natural products through plant genome mining to support multi-omics and *in silico*-guided drug discovery. Laboratory-based identification of bioactive plant natural products is challenging and labor-

intensive due to the complexity of the plant metabolome. Accordingly, this body of work addresses the challenge of predicting natural product structures and medicinal functions in silico from plant biosynthetic genes. The long-term objective is to establish a plant genome mining strategy that streamlines the prediction of bioactive natural products from plant biosynthetic genes. The research aims to identify biosynthetically feasible peptides from plant transcriptomes and genomes, followed by the *in silico* and *in vitro* evaluation of peptide function using multi-omic drug discovery approaches. The central thesis posits that plant genome mining for peptide cyclases, such as BURP-domain-containing proteins, enables accurate chemical structure prediction from precursor gene sequences, in silico screening, and expanded in vitro screening. Additionally, it is hypothesized that the NCBI Sequence Read Archive (SRA) constitutes an underexplored resource for large-scale, gene-guided discovery of plant bioactive peptides from fused burpitide cyclase pathways. The rationale is that in silico analysis of plant genetic resources applied to rational drug discovery of diverse plant peptides will yield new pharmacologically relevant plant therapeutics for human diseases.

Chapter 2 presents benchmarking and comparative evaluation of *de novo* assemblers for assembling repetitive BURP-domain-containing protein sequences. This knowledge is then applied to large-scale mining of eukaryotic burpitide ribosomally synthesized and post-translationally modified peptides (RiPPs) in plant transcriptomes, broadening discovery strategies for new macrocyclic structures. Chapter 3 extends this approach by employing cheminformatics-guided generation of an *in silico* library of burpitides and applying it to computational drug discovery through proof-of-concept screening against lipoprotein signal peptidase II (LspA), a cyclic peptide antimicrobial target, using molecular docking. Chapter 4 authenticates and validates the de novo assembler and genome mining investigations from Chapter 2 by discovering diversified post-translational modifications of stephanotic-acid-type burpitides. It further demonstrates the potential of plant peptides as by identifying a moroidin analogue with enhanced bioactivity against H1437 lung cancer cells. Chapter 5 integrates analytical chemistry, LC-MS/MS-based metabolomics, and plant genomics to generate a *multi-omics* burpitide library.

Collectively, this dissertation offers a comprehensive overview of diverse strategies for the discovery of novel plant natural products. The research expands the understanding of plant peptide chemistry and its application in rational drug design through plant genome mining.

1.4 References

1. Elshafie, H. S., Camele, I. & Mohamed, A. A. A Comprehensive Review on the Biological, Agricultural and Pharmaceutical Properties of Secondary Metabolites Based-Plant Origin. *Int. J. Mol. Sci.* **24**, 3266 (2023).
2. Walsh, C. T. & Tang, Y. *Natural Product Biosynthesis: Chemical Logic and Enzymatic Machinery*. (The Royal Society of Chemistry, 2017). doi:10.1039/9781788016612.
3. Craney, A., Ahmed, S. & Nodwell, J. Towards a new science of secondary metabolism. *J. Antibiot. (Tokyo)* **66**, 387–400 (2013).
4. Fang, C., Fernie, A. R. & Luo, J. Exploring the Diversity of Plant Metabolism. *Trends Plant Sci.* **24**, 83–98 (2019).
5. Huang, X.-Q. & Dudareva, N. Plant specialized metabolism. *Curr. Biol.* **33**, R473–R478 (2023).
6. Hart, C. E. *et al.* Defining the limits of plant chemical space: challenges and estimations. *GigaScience* **14**, giaf033 (2025).
7. Van Herwerden, E. F. & Süssmuth, R. D. Sources for Leads: Natural Products and Libraries. in *New Approaches to Drug Discovery* (eds Nielsch, U., Fuhrmann, U. & Jaroch, S.) vol. 232 91–123 (Springer International Publishing, Cham, 2015).
8. Zaynab, M. *et al.* Role of secondary metabolites in plant defense against pathogens. *Microb. Pathog.* **124**, 198–202 (2018).
9. McChesney, J. D., Venkataraman, S. K. & Henri, J. T. Plant natural products: Back to the future or into extinction? *Phytochemistry* **68**, 2015–2022 (2007).
10. Quave, C. L. Antibiotics from nature: Traditional medicine as a source of new solutions for combating antimicrobial resistance. in *AMR Control 2016: Overcoming Global Antimicrobial Resistance* 97–102 (World Alliance Against Antibiotic Resistance (WAAAR), 2016).
11. Established antibacterial drugs from plants. in *Advances in Botanical Research* vol. 106 81–149 (Elsevier, 2023).
12. Garcia Jimenez, D., Poongavanam, V. & Kihlberg, J. Macrocycles in Drug Discovery─Learning from the Past for the Future. *J. Med. Chem.* **66**, 5377–5396 (2023).
13. Asma, S. T. *et al.* Natural Products/Bioactive Compounds as a Source of Anticancer Drugs. *Cancers* **14**, 6203 (2022).
14. Wani, M. C. & Horwitz, S. B. Nature as a remarkable chemist: a personal story of the discovery and development of Taxol. *Anticancer. Drugs* **25**, 482–487 (2014).
15. One Thousand Plant Transcriptomes Initiative. One thousand plant transcriptomes and the phylogenomics of green plants. *Nature* **574**, 679–685 (2019).
16. Chassagne, F. *et al.* A Systematic Review of Plants With Antibacterial Activities: A Taxonomic and Phylogenetic Perspective. *Front. Pharmacol.* **11**, 586548 (2021).
17. Allkin, B. Useful Plants – Medicines: At Least 28,187 Plant Species are Currently Recorded as Being of Medicinal Use. in *State of the World's Plants 2017* (ed. Willis, K. J.) (Royal Botanic Gardens, Kew, London (UK), 2017).
18. Hamilton, A. *Medicinal Plants and Conservation: Issues and Approaches*. (International Plants Conservation Unit, WWF-UK, Godalming, UK, 2003).
19. Patridge, E., Gareiss, P., Kinch, M. S. & Hoyer, D. An analysis of FDA-approved drugs: natural products and their derivatives. *Drug Discov. Today* **21**, 204–207 (2016).

20. Najmi, A., Javed, S. A., Al Bratty, M. & Alhazmi, H. A. Modern Approaches in the Discovery and Development of Plant-Based Natural Products and Their Analogues as Potential Therapeutic Agents. *Molecules* **27**, 349 (2022).
21. Wilson, B. A. P., Thornburg, C. C., Henrich, C. J., Grkovic, T. & O'Keefe, B. R. Creating and screening natural product libraries. *Nat. Prod. Rep.* **37**, 893–918 (2020).
22. Salem, M. A. *et al.* Metabolomics in the Context of Plant Natural Products Research: From Sample Preparation to Metabolite Analysis. *Metabolites* **10**, 37 (2020).
23. Park, J. *et al.* Mass Spectral Data of Primary and Secondary Metabolites Changes in Medicinal Plants by Solvent Polarity. *Sci. Data* **12**, 893 (2025).
24. Gaudêncio, S. P. *et al.* Advanced Methods for Natural Products Discovery: Bioactivity Screening, Dereplication, Metabolomics Profiling, Genomic Sequencing, Databases and Informatic Tools, and Structure Elucidation. *Mar. Drugs* **21**, 308 (2023).
25. Henrich, C. J. & Beutler, J. A. Matching the power of high throughput screening to the chemical diversity of natural products. *Nat. Prod. Rep.* **30**, 1284 (2013).
26. Yu, J. S. *et al.* A versatile toolkit for drug metabolism studies with GNPS2: from drug development to clinical monitoring. *Nat. Protoc.* https://doi.org/10.1038/s41596-025-01237-6 (2025) doi:10.1038/s41596-025-01237-6.
27. Wang, M. *et al.* Sharing and community curation of mass spectrometry data with Global Natural Products Social Molecular Networking. *Nat. Biotechnol.* **34**, 828–837 (2016).
28. Gaudêncio, S. P. *et al.* Advanced Methods for Natural Products Discovery: Bioactivity Screening, Dereplication, Metabolomics Profiling, Genomic Sequencing, Databases and Informatic Tools, and Structure Elucidation. *Mar. Drugs* **21**, 308 (2023).
29. Leinonen, R., Sugawara, H., Shumway, M., & on behalf of the International Nucleotide Sequence Database Collaboration. The Sequence Read Archive. *Nucleic Acids Res.* **39**, D19–D21 (2011).
30. Katz, K. *et al.* The Sequence Read Archive: a decade more of explosive growth. *Nucleic Acids Res.* **50**, D387–D390 (2022).
31. Karasikov, M. *et al.* Efficient and accurate search in petabase-scale sequence repositories. *Nature* **647**, 1036–1044 (2025).
32. Del Pup, E. *et al.* plantiSMASH 2.0: improvements to detection, annotation, and prioritization of plant biosynthetic gene clusters. Preprint at https://doi.org/10.1101/2025.10.28.683968 (2025).
33. Kautsar, S. A., Suarez Duran, H. G., Blin, K., Osbourn, A. & Medema, M. H. plantiSMASH: automated identification, annotation and expression analysis of plant biosynthetic gene clusters. *Nucleic Acids Res.* **45**, W55–W63 (2017).
34. Sainsbury, F., Thuenemann, E. C. & Lomonossoff, G. P. pEAQ: versatile expression vectors for easy and quick transient expression of heterologous proteins in plants. *Plant Biotechnol. J.* **7**, 682–693 (2009).
35. Stephenson, M. J., Reed, J., Brouwer, B. & Osbourn, A. Transient Expression in Nicotiana Benthamiana Leaves for Triterpene Production at a Preparative Scale. *J. Vis. Exp.* 58169 (2018) doi:10.3791/58169.
36. Chekan, J. R., Mydy, L. S., Pasquale, M. A. & Kersten, R. D. Plant peptides – redefining an area of ribosomally synthesized and post-translationally modified peptides. *Nat. Prod. Rep.* **41**, 1020–1059 (2024).

37. Arnison, P. G. *et al.* Ribosomally synthesized and post-translationally modified peptide natural products: overview and recommendations for a universal nomenclature. *Nat Prod Rep* **30**, 108–160 (2013).
38. Kessler, S. C. & Chooi, Y.-H. Out for a RiPP: challenges and advances in genome mining of ribosomal peptides from fungi. *Nat. Prod. Rep.* **39**, 222–230 (2022).
39. Montalbán-López, M. *et al.* New developments in RiPP discovery, enzymology and engineering. *Nat. Prod. Rep.* **38**, 130–239 (2021).
40. Tan, N.-H. & Zhou, J. Plant Cyclopeptides. *Chem. Rev.* **106**, 840–895 (2006).
41. Chigumba, D. N. *et al.* Discovery and biosynthesis of cyclic plant peptides via autocatalytic cyclases. *Nat. Chem. Biol.* **18**, 18–28 (2022).
42. Parsley, N. C. *et al.* PepSAVI-MS reveals anticancer and antifungal cycloviolacins in Viola odorata. *Phytochemistry* **152**, 61–70 (2018).
43. Kirkpatrick, C. L. *et al.* The "PepSAVI-MS" Pipeline for Natural Product Bioactive Peptide Discovery. *Anal. Chem.* **89**, 1194–1201 (2017).
44. Mohimani, H. *et al.* Dereplication of microbial metabolites through database search of mass spectra. *Nat. Commun.* **9**, 4035 (2018).
45. Mohimani, H. *et al.* Dereplication of peptidic natural products through database search of mass spectra. *Nat. Chem. Biol.* **13**, 30–37 (2017).
46. Brodbelt, J. S., Morrison, L. J. & Santos, I. Ultraviolet Photodissociation Mass Spectrometry for Analysis of Biological Molecules. *Chem. Rev.* **120**, 3328–3380 (2020).
47. Lee, Y.-Y. *et al.* HypoRiPPAtlas as an Atlas of hypothetical natural products for mass spectrometry database search. *Nat. Commun.* **14**, 4219 (2023).
48. Cheng, S. *et al.* 10KP: A phylodiverse genome sequencing plan. *GigaScience* **7**, (2018).
49. Goodstein, D. M. *et al.* Phytozome: a comparative platform for green plant genomics. *Nucleic Acids Res.* **40**, D1178–D1186 (2012).
50. Tietz, J. I. *et al.* A new genome-mining tool redefines the lasso peptide biosynthetic landscape. *Nat. Chem. Biol.* **13**, 470–478 (2017).
51. Medema, M. H., De Rond, T. & Moore, B. S. Mining genomes to illuminate the specialized chemistry of life. *Nat. Rev. Genet.* **22**, 553–571 (2021).
52. De Veer, S. J., Kan, M.-W. & Craik, D. J. Cyclotides: From Structure to Function. *Chem. Rev.* **119**, 12375–12421 (2019).
53. Kandy, S. K., Pasquale, M. A. & Chekan, J. R. Aromatic side-chain crosslinking in RiPP biosynthesis. *Nat. Chem. Biol.* **21**, 168–181 (2025).
54. Kersten, R. D. *et al.* Gene-Guided Discovery and Ribosomal Biosynthesis of Moroidin Peptides. *J. Am. Chem. Soc.* **144**, 7686–7692 (2022).
55. Kersten, R. D. & Weng, J.-K. Gene-guided discovery and engineering of branched cyclic peptides in plants. *Proc. Natl. Acad. Sci.* **115**, (2018).
56. Lima, S. T. *et al.* A Widely Distributed Biosynthetic Cassette Is Responsible for Diverse Plant Side Chain Cross-Linked Cyclopeptides**. *Angew. Chem. Int. Ed.* **62**, e202218082 (2023).
57. Hattori, J., Boutilier, K. A., Campagne, M. M. L. & Miki, B. L. A conserved BURP domain defines a novel group of plant proteins with unusual primary structures. *Mol. Gen. Genet. MGG* **259**, 424–428 (1998).
58. Mydy, L. S. *et al.* An intramolecular macrocyclase in plant ribosomal peptide biosynthesis. *Nat. Chem. Biol.* **20**, 530–540 (2024).
59. Noyon, M. R. O. K. & Hematian, S. Peptide crosslinking by a class of plant copper enzymes. *Trends Chem.* **6**, 649–655 (2024).

60. Morita, H., Yoshida, N., Takeya, K., Itokawa, H. & Shirota, O. Configurational and conformational analyses of a cyclic octapeptide, lyciumin A, from Lycium chinense Mill. *Tetrahedron* **52**, 2795–2802 (1996).
61. Yahara, S. *et al.* Cyclic Peptides, Acyclic Diterpene Glycosides and Other Compounds from Lycium chinense MILL. *Chem. Pharm. Bull. (Tokyo)* **41**, 703–709 (1993).
62. Yoshikawa, K., Tao, S. & Arihara, S. Stephanotic Acid, a Novel Cyclic Pentapeptide from the Stem of *Stephanotis f loribunda*. *J. Nat. Prod.* **63**, 540–542 (2000).
63. Yun, B.-S., Ryoo, I.-J., Lee, I.-K. & Yoo, I.-D. Hibispeptin B, a novel cyclic peptide from Hibiscus syriacus. *Tetrahedron* **54**, 15155–15160 (1998).
64. Wang, X. *et al.* Large-scale transcriptome mining enables macrocyclic diversification and improved bioactivity of the stephanotic acid scaffold. *Nat. Commun.* **16**, 4198 (2025).
65. Mydy, L. S., Chigumba, D. N. & Kersten, R. D. Plant Copper Metalloenzymes As Prospects for New Metabolism Involving Aromatic Compounds. *Front. Plant Sci.* **12**, 692108 (2021).
66. Heger, A. & Holm, L. Rapid automatic detection and alignment of repeats in protein sequences. *Proteins Struct. Funct. Genet.* **41**, 224–237 (2000).
67. Ramesh, S. *et al.* Bioinformatics-Guided Expansion and Discovery of Graspetides. *ACS Chem. Biol.* **16**, 2787–2797 (2021).
68. Bulik, I. W., Frisch, M. J. & Vaccaro, P. H. Vibrational self-consistent field theory using optimized curvilinear coordinates. *J. Chem. Phys.* **147**, 044110 (2017).
69. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410 (1990).
70. Liu, W. *et al.* Genome and transcriptome of Selaginella kraussiana reveal evolution of root apical meristems in vascular plants. *Curr. Biol.* **33**, 4085-4097.e5 (2023).
71. Priyam, A. *et al.* Sequenceserver: A Modern Graphical User Interface for Custom BLAST Databases. *Mol. Biol. Evol.* **36**, 2922–2924 (2019).
72. Gibson, D. G. *et al.* Enzymatic assembly of DNA molecules up to several hundred kilobases. *Nat. Methods* **6**, 343–345 (2009).
73. Boersema, P. J., Raijmakers, R., Lemeer, S., Mohammed, S. & Heck, A. J. R. Multiplex peptide stable isotope dimethyl labeling for quantitative proteomics. *Nat. Protoc.* **4**, 484–494 (2009).
74. Morita, H., Shimbo, K., Shigemori, H. & Kobayashi, J. Antimitotic activity of moroidin, a bicyclic peptide from the seeds of Celosia argentea. *Bioorg. Med. Chem. Lett.* **10**, 469–471 (2000).
75. Kang, K. B. *et al.* Jubanines F–J, cyclopeptide alkaloids from the roots of Ziziphus jujuba. *Phytochemistry* **119**, 90–95 (2015).
76. Fujii, K., Ikai, Y., Oka, H., Suzuki, M. & Harada, K. A Nonempirical Method Using LC/MS for Determination of the Absolute Configuration of Constituent Amino Acids in a Peptide: Combination of Marfey's Method with Mass Spectrometry and Its Practical Application. *Anal. Chem.* **69**, 5146–5151 (1997).
77. Vijayasarathy, S. *et al.* C_3 and 2D C_3 Marfey's Methods for Amino Acid Analysis in Natural Products. *J. Nat. Prod.* **79**, 421–427 (2016).
78. Bitchi, M. B. *et al.* Isolation and structure elucidation of cyclopeptide alkaloids from the leaves of Heisteria parvifolia. *Phytochemistry* **167**, 112081 (2019).
79. Lomchoey, N. *et al.* New bioactive cyclopeptide alkaloids with rare terminal unit from the root bark of *Ziziphus cambodiana*. *RSC Adv.* **8**, 18204–18215 (2018).
80. Morel, A. F. *et al.* Antibacterial cyclopeptide alkaloids from the bark of Condalia buxifolia. *Phytochemistry* **61**, 561–566 (2002).

81. Mostardeiro, M. A. *et al.* Cyclopeptide Alkaloids: Stereochemistry and Synthesis of the Precursors of Discarines C and D and Myrianthine A. *J. Nat. Prod.* **76**, 1343–1350 (2013).
82. Chen, B. *et al.* Cytotoxic and Antibacterial Isomalabaricane Terpenoids from the Sponge *Rhabdastrella globostellata*. *J. Nat. Prod.* **85**, 1799–1807 (2022).
83. Grimblat, N., Gavín, J. A., Hernández Daranas, A. & Sarotti, A. M. Combining the Power of *J* Coupling and DP4 Analysis on Stereochemical Assignments: The *J* -DP4 Methods. *Org. Lett.* **21**, 4003–4007 (2019).
84. Schramma, K. R., Bushin, L. B. & Seyedsayamdost, M. R. Structure and biosynthesis of a macrocyclic peptide containing an unprecedented lysine-to-tryptophan crosslink. *Nat. Chem.* **7**, 431–437 (2015).
85. Wyche, T. P., Ruzzini, A. C., Schwab, L., Currie, C. R. & Clardy, J. Tryptorubin A: A Polycyclic Peptide from a Fungus-Derived Streptomycete. *J. Am. Chem. Soc.* **139**, 12899–12902 (2017).
86. Xu, X. *et al.* Moroidin, a Cyclopeptide from the Seeds of *Celosia cristata* That Induces Apoptosis in A549 Human Lung Cancer Cells. *J. Nat. Prod.* **85**, 1918–1927 (2022).
87. Zhou, J. *et al.* Glycnsisitin A: A promising bicyclic peptide against heart failure that facilitates TFRC-mediated uptake of iron in cardiomyocytes. *Acta Pharm. Sin. B* **14**, 3125–3139 (2024).
88. Blin, K. *et al.* antiSMASH 6.0: improving cluster detection and comparison capabilities. *Nucleic Acids Res.* **49**, W29–W35 (2021).
89. Merwin, N. J. *et al.* DeepRiPP integrates multiomics data to automate discovery of novel ribosomally synthesized natural products. *Proc. Natl. Acad. Sci.* **117**, 371–380 (2020).
90. Mohimani, H. *et al.* Automated Genome Mining of Ribosomal Peptide Natural Products. *ACS Chem. Biol.* **9**, 1545–1551 (2014).
91. Schorn, M. A. *et al.* A community resource for paired genomic and metabolomic data mining. *Nat. Chem. Biol.* **17**, 363–368 (2021).
92. Singh, K. S., Van Der Hooft, J. J. J., Van Wees, S. C. M. & Medema, M. H. Integrative omics approaches for biosynthetic pathway discovery in plants. *Nat. Prod. Rep.* **39**, 1876–1896 (2022).
93. Medema, M. H., De Rond, T. & Moore, B. S. Mining genomes to illuminate the specialized chemistry of life. *Nat. Rev. Genet.* **22**, 553–571 (2021).
94. Lau, W. & Sattely, E. S. Six enzymes from mayapple that complete the biosynthetic pathway to the etoposide aglycone. *Science* **349**, 1224–1228 (2015).
95. McClune, C. J. *et al.* Discovery of FoTO1 and Taxol genes enables biosynthesis of baccatin III. *Nature* **643**, 582–592 (2025).
96. Caputi, L. *et al.* Missing enzymes in the biosynthesis of the anticancer drug vinblastine in Madagascar periwinkle. *Science* **360**, 1235–1239 (2018).
97. Xiong, X. *et al.* The Taxus genome provides insights into paclitaxel biosynthesis. *Nat. Plants* **7**, 1026–1036 (2021).
98. Singh, K. S. *et al.* MEANtools integrates multi-omics data to identify metabolites and predict biosynthetic pathways. *PLOS Biol.* **23**, e3003307 (2025).
99. Zhang, J. *et al.* A microbial supply chain for production of the anti-cancer drug vinblastine. *Nature* **609**, 341–347 (2022).

Chapter 2 Benchmarking of *de novo* Assemblers for Large-scale Plant Transcriptome Mining

2.1 Statement of contribution

This chapter is reproduced and adapted from the following: Wang, X., Shafiq, K., Ousley, D.A. *et al.* Large-scale transcriptome mining enables macrocyclic diversification and improved bioactivity of the stephanotic acid scaffold. *Nat Commun* 16, 4198 (2025). Dr. Xiaofeng Wang purified, isolated, and structurally elucidated the peptides. Dr. Khadija Shafiq conducted the bioassays and aided with purification of Glechomanin. Dr. Khadija and I performed large-scale transcriptome assembly and analysis. I completed the comparative assessment of transcriptome mining tools and helped co-author the publication.

2.2 Summary

Nearly 10,000 plant species are represented by RNA-seq datasets in the NCBI sequence read archive, which are difficult to search in unassembled format due to database size. Here, we optimize RNA-seq assembly to transform most of this public RNA-seq data to a searchable database for biosynthetic gene discovery. We test our transcriptome mining pipeline towards the diversification of moroidins, which are plant ribosomally-synthesized and post-translationally-modified peptides (RiPPs) biosynthesized from copper-dependent peptide cyclases. Thus, we demonstrate that large-scale transcriptome mining can broaden the medicinal chemistry toolbox for chemical and biological exploration of eukaryotic RiPP structures.

2.3 Introduction

Plants produce a large diversity of ribosomally-synthesized and post-translationally-modified peptides (RiPPs) with endogenous functions in development and abiotic and biotic stress responses[1–3]. These bioactivities have been translated into pesticides, immunosuppressants, and lead structures for chemotherapy drug candidates[4,5]. RiPPs are natural products biosynthesized by the ribosomal formation of a precursor peptide that is post-translationally modified in a core peptide sequence and proteolytically cleaved to release the modified core as the mature RiPP[6,7]. In prokaryotes, the discovery of RiPPs has been driven over the past two decades by advances in sequencing technologies, the availability of large numbers of genome sequences, and the identification of genes and motifs involved in RiPP biosynthesis for bioinformatic prediction[8–12]. However, genome mining as a strategy for plant RiPP discovery is constrained by the limited number of high-quality plant genomes with annotated gene products in searchable repositories such as the NCBI non-redundant protein database (**Table 2.3**)[13]. In addition, plant genome mining of natural product biosynthetic genes in the context of their metabolic pathways is complicated by eukaryotic gene structures and less biosynthetic gene clustering than in prokaryotes[14]. An alternative to plant genomics-guided discovery is the use of transcriptomes to predict peptide chemistry[15–17].

In the past decade, over 9000 phylodiverse plant species have had their transcriptomes sequenced[18] and deposited as RNA-seq data in the NCBI Sequence Read Archive (SRA)[19] (May 2024 (**Table 2.3**)). However, only a small fraction of this data has been utilized for gene-guided discovery of plant natural products to date. Most transcriptomes are sequenced using next-generation sequencing (NGS) technologies that provide short-read sequence data[20,21], and a key step in the transcriptomics-guided discovery of RiPPs is the assembly of such short-read sequences into longer transcripts that represent full-length coding sequences of expressed genes. For non-model organisms, i.e., those lacking a reference genome, transcriptome assembly is performed de novo using specialized short-read NGS[22]. Three common RNA-seq assemblers are Trinity[23], SPAdes[24,25], and SOAPdenovo-trans[26], which have been described as efficient in multi-species benchmarking[22]. Trinity and SPAdes were applied in RiPP discovery studies from plant RNA-seq data[27–30] and SOAPdenovo-trans was originally employed to assemble transcriptomes of the 1000 plant transcriptomes initiative (1KP)[18]. Another common assembler, MEGAHIT[31], was developed for fast and memory-efficient assembly of metagenomic short-read NGS data, but has limited applications in RNA-seq data assembly to date[32]. *De novo* transcriptome mining of RiPPs requires

assembly of a precursor peptide gene so that the core peptide sequence is connected to a searchable sequence motif for bioinformatic precursor gene identification (**Figure 2.1a**).

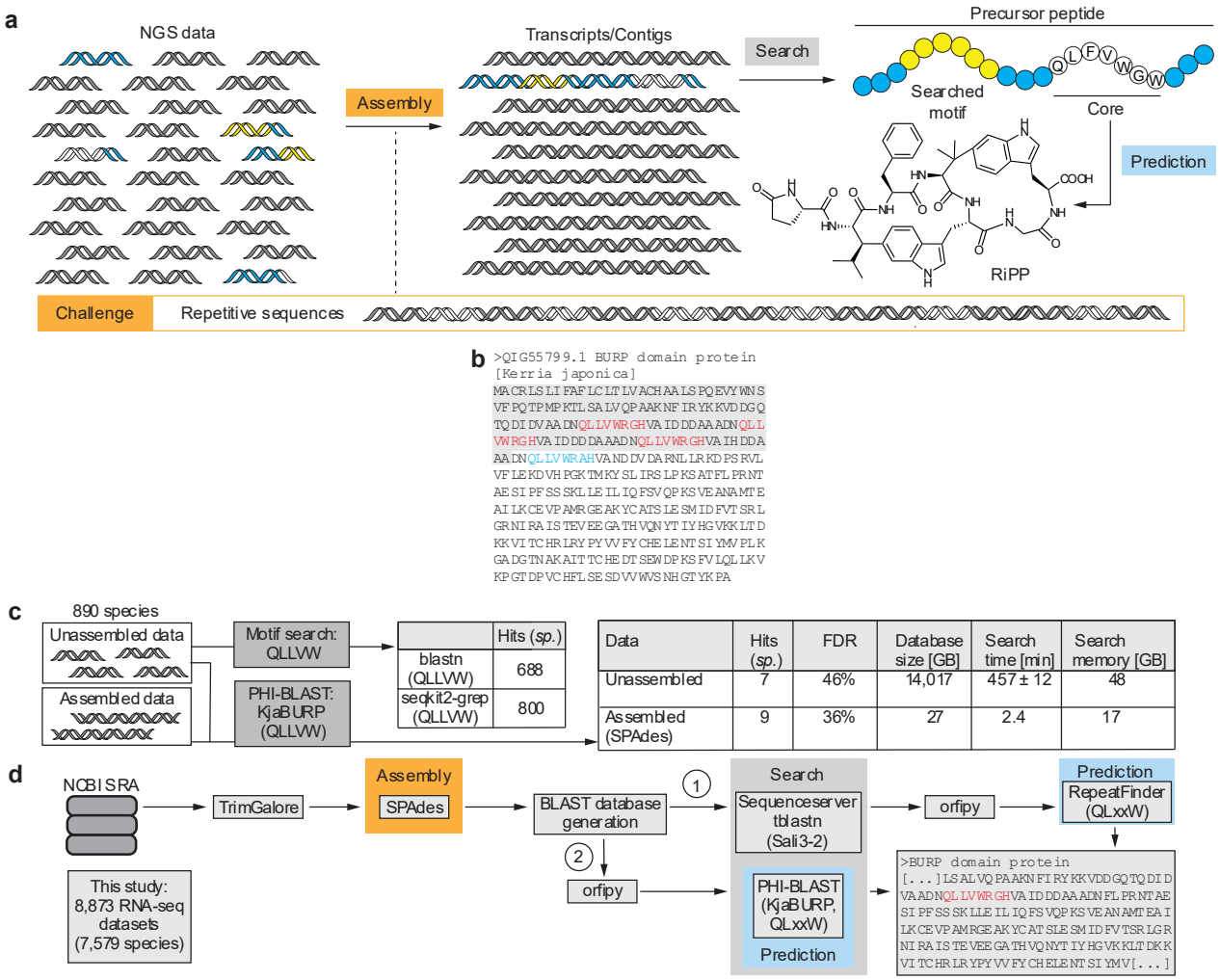


Figure 2.1 Transcriptome mining of plant RiPPs.
a General workflow of transcriptomics-guided discovery of plant RiPPs including *de novo* assembly of NGS data as a challenging step addressed in this work. **b** Moroidin precursor peptide KjaBURP from *Kerria japonica*. Moroidin core peptides are highlighted in red and blue color. Sequence highlighted in gray was not assembled with the BURP domain by SOAPdenovo-trans in 1kp database[18]. BURP-domain sequence is underlined. **c** Comparison of unassembled and assembled RNA-seq data search for discovery of stephanotic acid-type burpitides. **d** Transcriptome mining workflow of this study with either (1) Sequenceserver-based tblastn search or (2) commandline PHI-BLAST search. *sp*. species, FDR false discovery rate. Source data are provided as a Source Data file.

The moroidin burpitide cyclase KjaBURP, discovered in *Kerria japonica* through transcriptome mining, contains four moroidin core peptide sequence repeats within a 380-amino-acid protein domain (**Figure 2.1b**)[15]. Similarly, many plant RiPP precursors feature multiple core peptides embedded in sequence repeats[1,3,29,30,33–35], which pose a challenge for *de novo* assembly as it can result in incomplete assembly of the precursor-defining sequence motifs and its core

peptide(s) (**Figure 2.1b, Figure 2.2**)[29]. An example is the misassembly of moroidin precursor gene *KjaBURP* from *Kerria japonica* transcriptome by SOAPdenovo-trans[26] within the 1kp database which misses the N-terminal repetitive core peptide domain with moroidin cores (QLLVWRGH) required for moroidin discovery in japanese kerria **(Figure 2.1b)**. Given the importance of *de novo* assembly of precursor gene transcripts from RNA-seq data, a systematic comparison of NGS assemblers in the context of scaled plant RiPP discovery from transcriptomes is needed.

Herein, we assess three de novo short-read assemblers to generate precursor peptides from RNA-seq datasets of RiPP-producing plants and apply the most effective assembler towards mining the transcriptomes of 7569 plant species for new bioactive stephanotic acid-type burpitides and their biosynthetic cyclases.

2.4 Results

2.4.1 RNA-seq Assembly Benchmarking for Discovery of Plant RiPP Precursor Peptides

Towards large-scale transcriptome mining of plant RiPPs such as stephanotic acid-type burpitides, we wanted to first test if sequencing data assembly is advantageous for peptide discovery from RNA-seq data. We searched unassembled and SPAdes-*de novo* assembled transcriptomes of 890 plant species for the presence of the stephanotic acid core peptide QLLVW (**Figure 2.1c**). The target plant species had metabolomic data deposited in plantMASST[36], a repository of tandem mass spectrometry-based metabolomic data, and which can thus be tested for the presence of the stephanotic acid-QLLVW analyte. The unassembled RNA-seq data were first searched for the presence of the QLLVW motif as nucleotide or protein queries via blastn or Seqkit2[37], respectively. These searches of the stephanotic acid motif yielded too many false positives (98–99%) for effective peptide discovery (**Figure 2.1c**). Next, we searched homologs of moroidin cyclase KjaBURP with a QLLVW motif in both unassembled and assembled data via PHI-BLAST, a protein BLAST with target motif query[38]. Searching assembled data yielded 9 of 11 stephanotic acid-containing plants based on metabolomic data (**Figure 2.3**) compared to 7 hits in the unassembled data query. In addition, the assembled data search had a lower false discovery rate (36% versus 46%) compared to the unassembled data search. The BLAST database size of

assembled transcriptomes was 0.2% of unassembled RNA-seq data and search time and memory was 0.5% and 33%, respectively, with the same computational resource for assembled RNA-seq data (**Figure 2.1c and Figure 2.3**). Given the smaller database size, lower search cost, and better search results, we focused on applying RNA-seq assembly for plant peptide discovery from short sequence motifs.

As a prerequisite for large-scale plant transcriptome assembly, we aimed to identify a *de novo* NGS assembler that balances accuracy, speed, and computational cost for generating transcripts of published plant RiPP precursor genes. We included non-burpitide precursor peptides to expand the benchmarking data to 27 RNA-seq datasets from RiPP-producing plants. These datasets encompassed 36 precursor genes of (a) burpitides[15,29,30,33], (b) cysteine-rich peptides[39–50], (c) cyclotides[3,51,52], (d) orbitides[35,53,54], (e) PawS-derived RiPPs[55], and (f) linear RiPPs[1,56,57] (**Figure 2.2**). The 36 corresponding precursor peptides encoded 138 total core peptides and 65 unique core peptide sequences. 120 of the total core peptides were localized in sequence repeats within 18 precursor peptides. These datasets allowed us to assess the *de novo* assembly of sequence repeats within precursor peptides for subsequent bioinformatic core peptide detection. Each RNA-seq dataset was trimmed and assembled in triplicate on a computing cluster using the *de novo* NGS assemblers Trinity[23], SPAdes[24,25], or MEGAHIT[31]. MEGAHIT was the fastest of all assemblers and used the lowest total memory (**Table 2.1**). SPAdes assembly of all datasets required on average 19% more time and 15% more memory than MEGAHIT, while Trinity assembly needed on average ~19x more time and ~15x more memory than MEGAHIT (**Table 2.1**). For the detection of core peptides for RiPP discovery, a core should be assembled with a searchable motif, such as a BURP domain, to enable discovery of its RiPP. Some RiPPs require the correct assembly of short, often repetitive core peptide domains in addition to assembly with a searchable motif for successful RiPP discovery. SPAdes assembled more total core peptides with searchable motifs (67%) than Trinity (47%) and MEGAHIT (49%). Furthermore, SPAdes assembled 74% unique core peptides correctly with searchable motifs for RiPP discovery, whereas Trinity (57%) and MEGAHIT (58%) assembled less unique cores correctly. To further assess the accuracy of precursor gene assembly, the sum of sequence similarity and sequence identity values of all assembled precursors were determined by comparison to the database precursor sequences. Herein, SPAdes had the highest similarity and identity scores (81.4% and 80.6%, respectively) (**Table 2.1**).

Finally, SPAdes assembled the most repetitive cores (62%), followed by MEGAHIT (43%) and Trinity (37%). Based on replicates of all de novo benchmarking data assemblies, MEGAHIT and SPAdes assemblies were fully reproducible, whereas differences could be observed for individual Trinity assemblies. Read trimming before assembly resulted in more detected core peptides and more accurate precursor peptide assembly (**Table 2.4**). Ultimately, our benchmarking identified SPAdes as the de novo assembler, which generated precursor peptide transcripts with most unique core peptides at low computational cost and high speed. This assessment was further realized by a scaled *de novo* transcriptome assembly of the 1kp higher plant RNA-seq data with SPAdes and comparison of moroidin precursor search results via PHI-BLAST with the published datasets assembled by de novo transcriptome assembler SOAPdenovo-trans (**Table 2.5**)[18,26]. SPAdes-assembled 1kp data enabled identification of six fused moroidin cyclases with twelve different moroidin core peptides in five species, whereas published 1kp data contained only three fused moroidin cyclases with five different core peptides in three species (**Table 2.5**), showing that *de novo* SPAdes assembly can yield more predicted peptides at scale than existing plant transcriptomes from other assemblers.

Table 2.1 Benchmarking results of de novo transcriptome assembly for plant RiPP precursor discovery.
27 RNA-seq datasets were trimmed and assembled which include 36 precursor genes with 138 core peptides and 65 unique core peptides. Each RNA-seq dataset was assembled in triplicate (n = 3). Displayed values are means with one standard deviation shown as an error.

De novo assembler	Total time [min]	Total memory usage [GB]	Total detected cores (138 total cores)	Detected unique correct cores (65 total unique cores)	Detected repetitive cores (120 total repetitive cores)	Similarity [sum %] (total similarity score: 3600)	Identity [sum %] (total identity score: 3600)
MEGAHIT (v1.2.9)	1371 ± 66	210 ± 0	67 ± 0	38 ± 0	51 ± 0	2741 ± 0	2738 ± 0
SPAdes (v3.15.5)	1627 ± 42	241 ± 3	93 ± 0	48 ± 0	74 ± 0	3012 ± 0	2981 ± 0
Trinity (v2.15.1)	21833 ± 558	813 ± 23	64 ± 0	37 ± 0	44 ± 1	2618 ± 10	2597 ± 7

An alternative to *de novo* RNA-seq assembly is genome-guided transcriptome assembly which requires an annotated reference genome[58]. With increasing numbers of high-quality plant genomes complementing available RNA-seq data, we also compared genome-guided transcriptome assembly with StringTie[59] or Trinity to *de novo* assemblers regarding precursor peptide assembly (**Table 2.2**). Overall, genome-guided assembly via StringTie outperformed *de novo* SPAdes assembly in time, total detected cores, and assembly of cores in sequence repeats.

However, SPAdes assemblies of precursor peptides were superior to genome-guided assembly by StringTie or Trinity in the number of detected unique cores and identity scores (**Table 2.2**). Based on lower number of high-quality plant genomes compared to plant RNA-seq datasets (**Table 2.3**), we proceeded with a *de novo* assembly strategy with SPAdes (**Figure 2.1d**).

Table 2.2 Benchmarking results of de novo and genome-guided transcriptome assembly for plant RiPP precursor discovery.
16 RNA-seq datasets with corresponding annotated genomes were trimmed and assembled which include 20 precursor genes with 98 core peptides and 38 unique core peptides. Each RNA-seq dataset was assembled in triplicate (n = 3). Displayed values are means with one standard deviation shown as an error.

Assembler	Approach	Total time [min]	Total memory usage [GB]	Total detected cores (98 total cores)	Detected unique correct cores (38 total unique cores)	Detected repetitive cores (87 total repetitive cores)	Similarity [sum %] (total similarity score: 2000)	Identity [sum %] (total identity score: 2000)
MEGAHIT (v1.2.9)	De novo	940 ± 44	123 ± 1	42 ± 0	24 ± 0	32 ± 0	1582 ± 0	1587 ± 0
SPAdes (v3.15.5)	De novo	1185 ± 39	180 ± 1	66 ± 0	32 ± 0	56 ± 0	1728 ± 0	1713 ± 0
Trinity (v2.15.1)	De novo	13683 ± 621	485 ± 6	41 ± 0	25 ± 0	30 ± 1	2618 ± 10	2597 ± 7
StringTie (v2.2.1)	Genome-guided	922 ± 621	1635 ± 2	77 ± 0	27 ± 0	65 ± 0	1565 ± 0	1545 ± 0
Trinity (v2.15.1)	Genome-guided	10545 ± 417	1787 ± 4	62 ± 0	28 ± 0	51 ± 0	1795 ± 1	1691 ± 1

2.4.2 Large-scale Transcriptome Mining of Fused Moroidin Cyclases and Burpitides

Given that the second ring of moroidins is required for their lung adenocarcinoma cell cytotoxicity, we aimed to generate and search a large plant transcriptome database bioinformatically for new moroidin chemotypes and underlying burpitide cyclases for bicyclic diversification of the stephanotic acid scaffold for improved bioactivity. To create a large plant transcriptome database, 8893 plant transcriptomes were *de novo* assembled with SPAdes from RNA-seq data from the NCBI SRA representing 7579 plant species from 498 plant families. The assembly database was searched by tblastn[60,61] for the presence of BURP-domain-encoding genes as a starting point for the discovery of stephanotic acid-type burpitides. Burpitide cores can be predicted from BURP-domain protein sequences of fused burpitide peptide cyclases, i.e., their core peptide(s) are encoded in the N-terminal domain (RD22 BURP domains) or within the BURP domain (USP BURP domains)[62,63]. The assembled plant transcriptome database included 27,224 full-length non-redundant BURP-domain sequences, which were then analyzed via RepeatFinder[33]

for BURP-domain sequences with one or more internal stephanotic acid core peptide motif(s). Among the BURP-domain proteins, candidate stephanotic acid-type burpitide cyclases were identified from 37 species, which represented 7 plant families without reported stephanotic acid chemotypes (**Table 2.6**). In addition to Sequenceserver-tblastn search, we also applied PHI-BLAST search with moroidin cyclase KjaBURP and the stephanotic acid core QLxxW as a search motif. PHI-BLAST search yielded the same core peptide hits as Sequenceserver-tblastn search but PHI-BLAST resulted in 28 additional predicted stephanotic acid-type core peptides from transcripts without or with incomplete BURP-domain sequence (**Supplementary 2.6 and 2.7**). Given the focus on fused burpitide cyclases in this study, we excluded these additional core peptides in subsequent RiPP discovery experiments.

Candidate cyclases from 17 species included a stephanotic acid core with a histidine at positions 7–9 from the N-terminal glutamine that identified them as putative moroidin cyclases capable of forming a second-ring via Trp-His-crosslinking (**Table 2.6**). The predicted moroidin cyclases included 46 unique core peptides with most of the sequence diversity occurring in the second-ring, i.e., at core positions 6–9 (**34 unique motifs, Table 2.7**). In addition, several predicted cyclases featured extended stephanotic acid cores with tyrosine or tryptophan at positions 7–9 instead of histidine indicating the possibility of new second-ring crosslinks (**Table 2.6**).

2.5 Discussion

Here, we applied *de novo* plant transcriptome assembly and transcriptome mining for large-scale bioinformatic discovery of bioactive stephanotic acid-type burpitides. Based on our benchmarking analysis, SPAdes emerged as the most effective assembler for scaled transcriptome assembly due to a balance of high speed, high accuracy and low computational cost. MEGAHIT is faster, but not as accurate in *de novo* assembly of precursor transcripts as SPAdes, likely due to its design to maximize contig length during metagenomic assembly. SPAdes was best in assembly of repetitive plant RiPP precursor genes given their high number of detected core peptides in sequence repeats compared to other tested NGS assemblers. SPAdes has been applied in transcriptome mining of burpitides from fused and split precursor peptides but at one order of magnitude smaller data scale and with smaller outcome in terms of characterized burpitide chemistry (**Table 2.8**). Our transcriptome mining of fused burpitide cyclases utilized most plant

species with RNA-seq data in the NCBI SRA and revealed a large diversity of second ring analogs of the stephanotic acid scaffold with the core peptide QLLVWRNH being found in three plant families and therefore representing an interesting candidate for biological evaluation. Thus, our study informs the selection of current de novo NGS assemblers to efficiently address a major bottleneck in gene-guided discovery of plant natural products from RNA-seq data, i.e., the generation of large searchable genetic data. Since plants are eukaryotic organisms spanning a wide range of genome sizes from 61 Mbp to 160 Gbp[64,65] and plant RiPP precursor genes are structurally diverse[62], our results of plant transcriptome assembly should be applicable to other eukaryotic RNA-seq data for transcriptomics-guided peptide discovery[66].

2.6 Methods and Materials

2.6.1 RNA-Seq data download and trimming

Paired-end RNA-seq data (**Table 2.6)** were downloaded from NCBI SRA via fasterq-dump (--split-files) of the SRA Toolkit (v2.10.9)[19] to the Great Lakes High Performance Computing (HPC) Cluster at the University of Michigan, Ann Arbor. The datasets were trimmed via TrimGalore (v0.6.7) with default settings (Phred cutoff: 20, Default pair-cutoff: 20 bp).

2.6.2 Unassembled RNA-seq data search for stephanotic acid core peptides

Unassembled RNA-seq datasets of 890 plant species (**Table 2.5**) were either searched for nucleotide sequence CAACTTCTTGTTTGG encoding stephanotic acid core peptide QLLVW (from the Kerria japonica moroidin cyclase KjaBURP, GenBank ID QIG55799.1) via blastn on the NCBI BLAST webserver (https://blast.ncbi.nlm.nih.gov/Blast.cgi, default parameters: matrix BLOSUM62, gap-open 11, gap-extended 1, Filter L, e-value 0.05) or for protein sequence of core peptide QLLVW via seqkit2 (v2.3.0)[37] on the Great Lakes HPC cluster. For webserver search, each SRA dataset was searched individually for the target nucleotide sequence with a maximum search hit number of 100. For seqkit2 search, the 890 species datasets were downloaded and trimmed in batches of 100 datasets. Duplicate reads were then removed from each dataset via seqkit2 rmdup command and the resulting fastq datasets were converted to fasta datasets via seqkit2 fq2fa

command. The output fasta files were combined into one fasta-file for each 100-transcriptomes-batch and they were 6-frame translated via orfipy (v0.0.4, parameters: --pep --min 90 --between-stops)[67]. The orfipy output pep files were reduced in fasta-header size via awk command and the formatted pep file was searched via seqkit2 grep command for the presence of translated reads with the core peptide QLLVW. The identified reads were extracted from the seqkit2 grep output file for their respective SRA numbers with awk command. Unassembled data was converted to BLAST databases by makeblastdb and searched by commandline PHI-BLAST (BLAST+ v2.16.0, default settings, E-value 10)[38] for the presence of BURP-domain proteins including stephanotic acid core peptide QLLVW with KjaBURP GenBank accession QIG55799.1) as a query sequence. For comparison to assembled data, the same datasets were assembled de novo by SPAdes (v3.15.5, rnaSPAdes), combined and 6-frame translated by orfipy (v0.0.4, parameters: --pep --min 450 --between-stops), converted to BLAST databases by makeblastdb and searched by commandline PHI-BLAST (BLAST+ v2.16.0, default settings, E-value 10) for the presence of BURP-domain proteins including stephanotic acid core peptide QLLVW with KjaBURP. The ORF length of 90 bp was chosen for unassembled data 6-frame translation due to recommended PHI-BLAST database length minimum of 30 amino acids and a common RNA-seq read length of 90 bp. For benchmarking of stephanotic acid prediction via unassembled or assembled RNA-seq data searches, a deposited stephanotic acid-[LV] tandem mass spectrum (GNPS: CCMSLIB00006709935) was searched via plantMASST[36] (v2025.3.25, parameters: PM Tolerance (Da) 0.05, Fragment Tolerance (Da) 0.05, Cosine Threshold 0.7, Minimum Matched Peaks 3) in a metabolomic database including all 890 plant species searched in their RNA-seq data. False discovery rate was calculated as [(number of false positives)/((number of false positives) + (number of true positives))].

2.6.3 Transcriptome assembly

For *de novo* assembly benchmarking, datasets (**Table 2.1**) were trimmed via TrimGalore (v0.6.7) with default settings (Phred cutoff: 20, Default pair-cutoff: 20 bp) and assembled *de novo* with Trinity (v2.15.1)[23], SPAdes (v3.15.5)[24,25], or MEGAHIT (v1.2.9)[31] with the parameters in scripts specified on a GitHub repository (https://github.com/UM-KerstenLab4009/Moroidin-Transcriptomics). The working memory for all benchmarking datasets was 48 GB except for

Trinity assembly of SRR7440026 data (80 GB memory), which failed at 48 GB memory. For large-scale de novo assembly, the datasets were assembled with SPAdes on the Great Lakes HPC cluster with the same parameters as for benchmarking. Assemblies which failed at 48 GB memory were assembled at 180 GB memory (**Table 2.6**). For genome-guided assembly benchmarking, datasets (**Table 2.2**) were trimmed via TrimGalore (v0.6.7) with default settings. For each RNA-seq dataset, reads were aligned and mapped to the genome assembly fasta file with STAR (v2.7.11a) using the genome-specific gtf/gff3 annotation file. STAR alignment of a genome was done with the parameter –sjdbOverhang RNA-seq read-length minus 1. The resulting sam file was converted to bam file via samtools (v1.21). The transcriptome was assembled from the bam file either by StringTie (v2.2.1) or Trinity (v2.15.1). For genome-guided transcriptome assembly with Trinity, the parameter genome_guided_max_intron 50,000 was used. For StringTie transcriptome assembly, the final gtf file was converted with gffread (v0.12.7) to a fasta file.

2.6.4 Transcriptome analysis

For BLAST library preparation, fasta headers of all *de novo* assemblies and genome-guided assemblies were added with their SRA number and reduced in length below 51 letters. Precursor assembly benchmarking was done via Sequenceserver-based BLAST search: *De novo* assembled transcriptomes were searched for their corresponding precursor peptides via tblastn (v2.16.0 +)[60] on Sequenceserver (v3.1.0)[61] with the following BLAST parameters: e-value 1e-5, matrix BLOSUM62, gap-open 11, gap-extend 1, filter L. The top hit of each search was translated into protein sequence and aligned by Needle[68] pairwise alignment with the target precursor peptide. The resulting similarity and identity scores of all 36 precursor peptides were recorded (**Table 2.1**), added and compared to the maximum score (3600 = 36*100%) (**Table 2.1**). Large-scale search of BURP-domain transcripts was performed either via Sequenceserver BLAST and subsequent RepeatFinder analysis or via commandline PHI-BLAST and subsequent RepeatFinder analysis. For Sequenceserver-based BLAST search, all SPAdes *de novo* assembled transcriptomes (**Table 2.6**) were combined into 100 transcriptome-databases and searched via tblastn (BLAST+ v2.16.0 +) for homologs of Sali3-2 as a BURP-domain-containing protein query (GenBank ID AAB66369) on Sequenceserver (v3.1.0) with the following BLAST parameters: E-value 1e-05, matrix BLOSUM62, gap-open 11, gap-extend 1, filter L, max, max_target_seqs 5000. tblastn hits

from all databases were combined into a fasta-file and translated into protein sequences via orfipy (v0.0.4)[67] with the following parameters: --pep --min 450 --between-stops. The resulting protein sequence pep file was converted to a fasta file and analyzed by standalone RepeatFinder[33] on the Great Lakes HPC Cluster at the University of Michigan, Ann Arbor (see GitHub repository for RepeatFinder settings). Candidate stephanotic acid-type burpitide cyclases were characterized in RepeatFinder output as core QLxxW, where x is any proteinogenic amino acid, and presence in an open reading frame, which encoded a BURP-domain motif. For commandline BLAST search, all SPAdes *de novo* assemblies (**Table 2.1**) were combined into a fasta file and then translated into protein sequences via orfipy (v0.0.4)[67] with the following parameters: --pep --min 450 --between-stops. The resulting pep file was searched via PHI-BLAST (BLAST+ v2.16.0 +) with phi_pattern QQL-(x)2-W and KjaBURP (GenBank ID QIG55799.1) on default settings (E-value 10). The resulting hits were analyzed by RepeatFinder. In our searches, a BURP-domain motif was defined by the presence of at least half the BURP-domain-conserved residues (i.e., FF/Fx-C-C-CH of FF/Fx-C-C-CH-CH-CH-CH)[63]. Hits with a single QLxxW core peptide at the first N-terminal α-helix of BURP-domain proteins[69] were removed. PHI-BLAST results (**Table 2.6**) also included predicted stephanotic acid-type core peptides from transcripts without or incomplete BURP-domain sequence. Reduction of the PHI-BLAST E-value 10 to 1e-5 had no effect on search results from fused burpitide cyclases but it reduced hits from incomplete or non-BURP-domain transcripts (**Table 2.7**). In order to search for other core peptide motifs within BURP-domain proteins, the phi_pattern should be changed accordingly and a query BURP-domain protein sequence should be selected which includes the target motif, e.g., for a moroidin (core: QLLVWRGH) PHI-BLAST search, use the phi_pattern QQLLVWRGH and KjaBURP (GenBank ID QIG55799.1) which includes a QLLVWRGH motif in its sequence.

2.6.5 Data Availability

Transcriptome assemblies were deposited in Deep Blue Data repository of the University of Michigan under accession 9306t0217 [https://doi.org/10.7302/z14x-5m94][70,71].

2.7 Supplemental Information

Table 2.3 Plant genomes and plant RNA-seq data in the NCBI sequence read archive.

NCBI Plant genomes (May 2024)	#
Assemblies	1869
Assembly status	
Complete genome	25
Chromosome	868
Scaffold	270
Contig	706
Assemblies with annotation in nr protein database	414
Species	1728
Species with annotation in nr protein database	386
Represented families	225
Rep. families with annotation in nr protein database	112
RNA-seq datasets (SRA) (May 2024)	#
Species	>9000
Transcriptome assemblies (this study)	8897
Species (this study)	7579
Represented families (this study)	498

Table 2.4 Benchmarking of read trimming for de novo transcriptome assembly of plant RiPP precursor discovery.

De novo assembler	Total time [min]	Total memory usage [GB]	Total detected cores (138 total cores)	Detected unique correct cores (65 total unique cores)	Detected repetitive cores (120 total repetitive cores)	Similarity [sum %] (total similarity: 3600%)	Identity [sum %] (total identity: 3600%)	Gaps [sum %] (total gaps: 3600%)
SPAdes (v3.15.5)	2041 ± 273	269.9 ± 6.6	88	44	69	2948	2915	609
TrimGalore (v0.6.7) + SPAdes (v3.15.5)	(439 + 1628 = 2067) ± 241	(63 + 241 = 304) ± 3	93	48	74	3012	2981	545

27 RNA-seq datasets were assembled which include 36 precursor genes with 138 core peptides and 65 unique core peptides. Each RNA-seq dataset was assembled with prior read trimming (TrimGalore) or without read trimming in triplicate (n=3) with the standard deviation shown.

Table 2.5 Comparison of scaled transcriptome mining for fused moroidin cyclases and moroidin core peptides in published 1kp SOAPdenovo-trans assemblies and SPAdes-assemblies.

Assembly	Species	Detected fused moroidin cyclases	Detected moroidin core peptides
SOAPdenovo-trans (1kp study[4])	3 (*Lomandra longifolia, Bauhinia tomentosa, Kerria japonica*)	3	5 (QLLVWREH, QLLVWRSH, QLLFWFGRH, QLLVWFARH, QLLVWRAH)

SPAdes (v3.15.5, this study)	5 (*Kerria japonica, Lomandra longifolia, Orchidantha maxillarioides, Bauhinia tomentosa, Cercis canadensis*)	6	14 (QLLFWFGRH, QLLFWTRH, QLLVWFGRH, QLLVWPKH, QLLVWQKH, QLLVWRAH, QLLVWREH, QLLVWRGH, QLLVWRH, QLLVWRNH, QLLVWRSH, QLLVWSRH, QLLVWTDH, QLLVWTRH)

Table 2.6 RNA-seq datasets and plant species with predicted stephanotic acid-type burpitide cores from fused burpitide cyclases in assembled plant transcriptomes (tblastn and PHI-BLAST hits)

#	SRA#	Organism	Family	Predicted core peptides (stephanotic acid in **bold,** 1x = single core precursor)	Stephanotic acid Core: QLxxW	Moroidin Core: QLxxWx(xx)H	2.ring-Trp Core: QLxxWx(xx)W	2.ring-Tyr Core: QLxxWx(xx)Y
1	SRR12990041	*Aganosma wallichii*	Apocynaceae	**QLLVW**RW (1x)	x		x	
2	SRR6456116	*Stellaria aquatica*	Caryophyl-laceae	**QLLVW**RNH, **QLFVW**RNH, **QLLVW**SKH	x	x		
3	SRR5219180	*Mercurialis huetii*	Euphorbiaceae	**QLFFW**RY, **QLFFW**KY	x			x
4	ERR2597552	*Mercurialis annua*	Euphorbiaceae	**QLFFW**KY	x			x
5	SRR8797417	*Medicago archiducis-nicolai*	Fabaceae	**QLFVW**GGH	x	x		
6	ERR6688729	*Medicago arabica*	Fabaceae	**QLLIW**GGH, **QLLVW**GGH, **QLFVW**GGH	x	x		
7	SRR3141655	*Vigna mungo*	Fabaceae	**QLLVW**RW (1x)	x		x	
8	SRR3938646	*Medicago marina*	Fabaceae	**QLLWW**GGH (1x)	x	x		
9	ERR2021956	*Bauhinia blakeana*	Fabaceae	**QLLVW**RQH, **QLLVW**IKH, **QLLIW**RQH, **QLLVW**RGH, **QLLIW**KGH, **QLLVW**RSH,	x	x		
10	DRR168799	*Bauhinia forficata*	Fabaceae	**QLRVW**TSH, **QLSVW**TSH, **QLRVW**QSH, **QLLIW**MNH, **QLLVW**RKNH	x	x		
11	ERR2021957	*Bauhinia purpurea*	Fabaceae	**QLSFW**, **QLLVW**RNH, **QLLVW**RGH, **QLIVW**RRNH, **QLFVW**RGH, **QLRVW**TKH, **QLLVW**REH, **QLIVW**RNH, **QLLVW**RRGH, **QLLVW**RRNH,	x	x		
12	SRR16068779	*Medicago sativa*	Fabaceae	**QLFVW**SDH, **QLFVW**DDGH, **QLLVW**SSDH, **QLFVW**GGVH, **QLFVW**GGH, **QLLVW**VGH, **QLLVW**GGGH, **QLFVW**GGGH, **QLLVW**GGDH,	x	x		
14	SRR18083643	*Glechoma longituba*	Lamiaceae	**QLLVW**RW, **QLLVW**KW	x		x	
15	SRR5150732	*Glechoma hederacea*	Lamiaceae	**QLFVW**GW, **QLLVW**KW, **QLLVW**RW	x		x	
16	ERR3487362	*Salvia* sp. 1KP	Lamiaceae	**QLFVW**GW	x			
17	ERR2040758	*Orchidantha maxillari- oides*	Lowiaceae	**QLLVW**KW, **QLLVW**RSH, **QLFNW**YTH, **QLLVW**PKH, **QLLVW**TDH	x	x	x	
18	SRR16483051	*Cinna latifolia*	Poaceae	**QLLVW**, **QLLVW**RTY, QL**LLVW**RQY **QLLVW**RSY, **QLFIW**RSY, **QLFIW**RSY, **QLLLW**RSY, **QLFIW**RSY, **QLLVW**RSY, **QLLIW**RTY, **QLLVW**RTY, **QLLVW**RQY, **QLLLW**RRSY,**QLLVW**RSY, **QLFIW**RSY, **QLLIW**RTY, **QLLVW**RTY, **QLFIW**RRSY, **QLLIW**RRTY,	x			x
19	SRR5406165	*Digitaria cuyabensis*	Poaceae	**QLLVW**RNH	x	x		
20	SRR16234948	*Stipa sibirica*	Poaceae	**QLLVW**RKSW, **QLLVW**RESW	x		x	
21	SRR13316951	*Cinnagrostis crispa*	Poaceae	**QLLVW**RDW	x		x	
22	SRR16483054	*Holcus mollis*	Poaceae	**QLLVW**RSY**, QLLIW**RTY, **QLLVW**RTY, **QLLVW**RQY	x			x
23	SRR16243467	*Panicum brevifolium*	Poaceae	**QLLVW**RNH	x	x		
24	SRR13316952	*Cinnagrostis cabrerae*	Poaceae	**QLLVW**RDW	x		x	
25	SRR23873622	*Aegilops uniaristata*	Poaceae	**QLLIW**	x			
26	SRR16299781	*Leptothrium senegalense*	Poaceae	**QLLVW**RSH, **QLLVW**RNH	x	x		
27	SRR16483114	*Neomolinia japonica*	Poaceae	**QLLVW**RSW	x		x	
28	ERR3634426	*Brachypodium boissieri*	Poaceae	**QLLVW**RSW	x		x	
29	SRR12955278	*Bromus catharticus*	Poaceae	**QLLVW**RSW	x		x	
30	SRR6221583, SRR6221613	*Sacciolepis striata*	Poaceae	**QLLVW**RNH	x	x		
31	SRR5406195	*Zuloagaea bulbosa*	Poaceae	**QLLVW**RNH	x	x		
32	ERR9784826	*Dichrosta-chys cinerea*	Fabaceae	**QLLVW**GY	X			X

33	SRR7817914	*Hordeum brevisubula- tum*	Poaceae	**QLRVW**RKY (1x)	x			x
34	SRR23631370	*Amomum smithiae*	Zingiberaceae	**QLLVW** (1x)	x			
		Known producers						
35	ERR2040423	*Kerria japonica*	Rosaceae	**QLLVW**RAH, **QLLVW**RGH	x	x		
36	ERR706821	*Bauhinia tomentosa*	Fabaceae	**QLLVW**RSH, **QLLVW**RNH, **QLLFW**FGRH, **QLLVW**SRH, **QLLVW**TRH, **QLLFW**TRH,	x	x		
37	ERR706845	*Cercis canadensis*	Fabaceae	**QLLVW**, **QLKVW**, **QLLVW**QKH, **QLLVW**RH	x	x		

Table 2.7 Predicted core peptides of stephanotic acid-type burpitides in assembled plant transcriptomes.

#	Moroidin cores	Glechomanin cores	Mercurialin cores	Stephanotic acid core
1	QLLVWRH	QLFVWGW	QLFFWKY	QLFFW
2	QLSFWKH	QLLLWKW	QLFFWRY	QLFIW
3	QLFNWYTH	QLLVWKW	QLLVWGY	QLFNW
4	QLFVWGGH	QLLVWRW	QLFIWRSY	QLFVW
5	QLFVWRGH	QLLVWRDW	QLLIWRTY	QLIVW
6	QLFVWRNH	QLLVWRSW	QLLLWRSY	QLLFW
7	QLFVWSDH	QLLVWRESW	QLLVWRQY	QLLIW
8	QLIVWRNH	*QLLVWGW*	QLLVWRSY	QLLLW
9	QLLFWTRH	*QLLVWCAW*	QLLVWRTY	QLLVW
10	QLLIWGGH	*QLLVWQSW*	QLRVWRKY	QLLWW
11	QLLIWKGH	*QLLVWRAW*	QLFIWRRSY	QLRVW
12	QLLIWMNH	*QLLVWREW*	QLLIWRRTY	QLSFW
13	QLLIWRQH	Fused: 8	QLLLWRRSY	QLSVW
14	QLLVWGGH	*Split/ Incomplete: 5*	QLLVWRRSY	*QLLMW*
15	QLLVWIKH		QLLVWRRTY	*QLPVW*
16	QLLVWPKH		*QLVPWEGY*	Fused: 13
17	QLLVWQKH		Fused: 15 Split/incomplete: 1	Split/ 2 Incomplete:
18	QLLVWRAH			
19	QLLVWREH			
20	QLLVWRGH			
21	QLLVWRNH			
22	QLLVWRQH			
23	QLLVWRSH			
24	QLLVWSKH			
25	QLLVWSRH			
26	QLLVWTDH			
27	QLLVWTRH			
28	QLLVWVGH			
29	QLLWWGGH			
30	QLRVWQSH			
31	QLRVWTKH			
32	QLRVWTSH			
33	QLSVWTSH			
34	QLFVWAAGH			
35	QLFVWDDGH			
36	QLFVWGGGH			
37	QLFVWGGVH			
38	QLIVWRRNH			
39	QLLFWFGRH			
40	QLLVWFGRH			
41	QLLVWGGDH			
42	QLLVWGGGH			

43	QLLVWRKNH
44	QLLVWRRGH
45	QLLVWRRNH
46	QLLVWSSDH
47	*QFLVWRGH*
48	*QLFIWRVH*
49	*QLFVWRSH*
50	*QLLIWPRH*
51	*QLLIWRNH*
52	*QLLMWRGH*
53	*QLLVWHNH*
54	*QLLVWKTH*
55	*QLLVWLNH*
56	*QLLVWPEH*
57	*QLLVWPRH*
58	*QLLVWRTH*
59	*QLPVWRGH*
60	*QLRVWTQH*
61	*QLVPWEGH*
62	*QLLVWPPGH*
63	*QLLVWPPRH*
64	*QLLVWPPSH*
65	*QLLVWRRIH*
66	*QLLVWRRSH*
	Fused: 46 *Split/Incomplete: 20*

Table 2.8 Comparison of burpitide transcriptome mining.

Comparison of RNA-seq studies	Datasets	Detected stephanotic acid core peptides	Characterized new burpitides (NMR, bio- synthesis)	Characterized new PTMs (NMR, bio- synthesis)
Kersten *et a!.* 3	700	3	1	1
Kriger *et a!.* 5	793	5	0	0
This study	8893	55	3	2

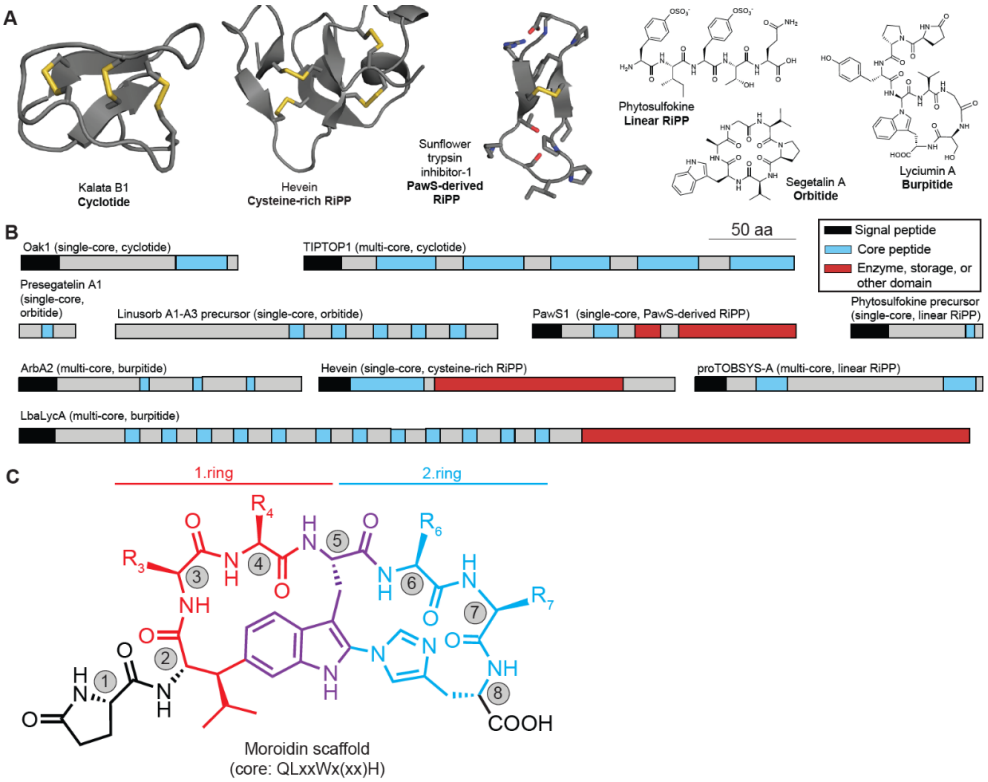


Figure 2.2 Plant RiPP structures and biosynthetic precursor peptides
(**A**) Chemical structures of representative plant RiPPs. (**B**) Primary precursor peptide structures of representative plant RiPPs including single- and multi-core precursor peptides. (**C**) Residue numbering in moroidin scaffold. Residue numbers are in grey circles. The second ring can be contracted or extended by one residue 3. x is any proteinogenic amino acid. Abbreviations: aa- amino acid.

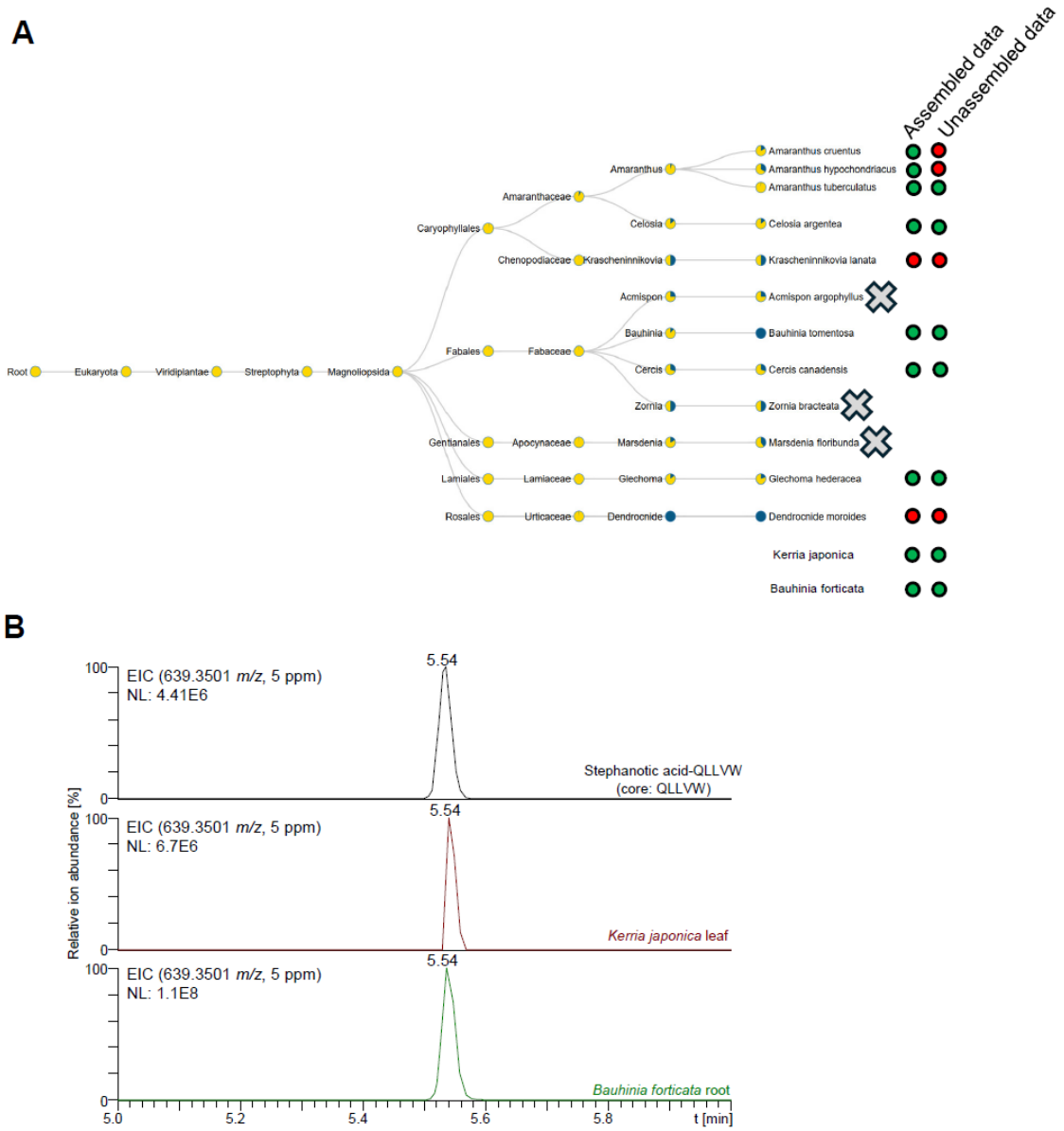


Figure 2.3 Detection of stephanotic acid-QLLVW in plantMASST-LC-MS2 data and manually analyzed LC-MS1 data.

(**A**) plantMASST (v2025.3.25) 6 phylogenetic representation of hits for searching stephanotic acid-QLLVW tandem MS spectrum (CCMSLIB00006709935). A corresponding PHIBLAST hit (KjaBURP homolog with QLLVW motif) in available transcriptomics data (assembled or unassembled) is indicated by a green circle, no PHI-BLAST hit is indicated by a red circle. plantMASST hits without availableRNA-seq data are labeled with a grey cross. (**B**) LC-MS1-based detection of stephanotic acid-QLLVW in methanolic extracts of *Bauhinia forticata* and *Kerria japonica,* which were missed in plantMASST, likely due to the lack of stephanotic acid-QLLVW tandem MS spectra in deposited datasets.

2.8 Reference

1. Pearce, G., Moura, D. S., Stratmann, J. & Ryan, C. A. Production of multiple plant hormones from a single polyprotein precursor. *Nature* **411**, 817–820 (2001).

2. Pearce, G., Moura, D. S., Stratmann, J. & Ryan, C. A. RALF, a 5-kDa ubiquitous polypeptide in plants, arrests root growth and development. *Proc. Natl. Acad. Sci.* **98**, 12843–12847 (2001).
3. Jennings, C., West, J., Waine, C., Craik, D. & Anderson, M. Biosynthesis and insecticidal properties of plant cyclotides: The cyclic knotted proteins from *Oldenlandia affinis*. *Proc. Natl. Acad. Sci.* **98**, 10614–10619 (2001).
4. Gründemann, C., Stenberg, K. G. & Gruber, C. W. T20K: An Immunomodulatory Cyclotide on Its Way to the Clinic. *Int. J. Pept. Res. Ther.* **25**, 9–13 (2019).
5. Oguis, G. K., Gilding, E. K., Jackson, M. A. & Craik, D. J. Butterfly Pea (Clitoria ternatea), a Cyclotide-Bearing Plant With Applications in Agriculture and Medicine. *Front. Plant Sci.* **10**, 645 (2019).
6. Arnison, P. G. *et al.* Ribosomally synthesized and post-translationally modified peptide natural products: overview and recommendations for a universal nomenclature. *Nat Prod Rep* **30**, 108–160 (2013).
7. Montalbán-López, M. *et al.* New developments in RiPP discovery, enzymology and engineering. *Nat. Prod. Rep.* **38**, 130–239 (2021).
8. Chen, I.-M. A. *et al.* IMG/M: integrated genome and metagenome comparative data analysis system. *Nucleic Acids Res.* **45**, D507–D516 (2017).
9. Merwin, N. J. *et al.* DeepRiPP integrates multiomics data to automate discovery of novel ribosomally synthesized natural products. *Proc. Natl. Acad. Sci.* **117**, 371–380 (2020).
10. Mohimani, H. *et al.* Automated Genome Mining of Ribosomal Peptide Natural Products. *ACS Chem. Biol.* **9**, 1545–1551 (2014).
11. Tietz, J. I. *et al.* A new genome-mining tool redefines the lasso peptide biosynthetic landscape. *Nat. Chem. Biol.* **13**, 470–478 (2017).
12. Kloosterman, A. M., Shelton, K. E., Van Wezel, G. P., Medema, M. H. & Mitchell, D. A. RRE-Finder: a Genome-Mining Tool for Class-Independent RiPP Discovery. *mSystems* **5**, e00267-20 (2020).
13. O'Leary, N. A. *et al.* Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res.* **44**, D733–D745 (2016).
14. Jacobowitz, J. R. & Weng, J.-K. Exploring Uncharted Territories of Plant Specialized Metabolism in the Postgenomic Era. *Annu. Rev. Plant Biol.* **71**, 631–658 (2020).
15. Kersten, R. D. *et al.* Gene-Guided Discovery and Ribosomal Biosynthesis of Moroidin Peptides. *J. Am. Chem. Soc.* **144**, 7686–7692 (2022).
16. Kriger, D., Pasquale, M. A., Ampolini, B. G. & Chekan, J. R. Mining raw plant transcriptomic data for new cyclopeptide alkaloids. *Beilstein J. Org. Chem.* **20**, 1548–1559 (2024).
17. Fisher, M. F. *et al.* The genetic origin of evolidine, the first cyclopeptide discovered in plants, and related orbitides. *J. Biol. Chem.* **295**, 14510–14521 (2020).
18. One Thousand Plant Transcriptomes Initiative. One thousand plant transcriptomes and the phylogenomics of green plants. *Nature* **574**, 679–685 (2019).
19. Kodama, Y., Shumway, M., Leinonen, R., & on behalf of the International Nucleotide Sequence Database Collaboration. The sequence read archive: explosive growth of sequencing data. *Nucleic Acids Res.* **40**, D54–D56 (2012).
20. Goodwin, S., McPherson, J. D. & McCombie, W. R. Coming of age: ten years of next-generation sequencing technologies. *Nat. Rev. Genet.* **17**, 333–351 (2016).
21. Hu, T., Chitnis, N., Monos, D. & Dinh, A. Next-generation sequencing technologies: An overview. *Hum. Immunol.* **82**, 801–811 (2021).

22. Hölzer, M. & Marz, M. *De novo* transcriptome assembly: A comprehensive cross-species comparison of short-read RNA-Seq assemblers. *GigaScience* **8**, giz039 (2019).
23. Grabherr, M. G. *et al.* Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat. Biotechnol.* **29**, 644–652 (2011).
24. Bankevich, A. *et al.* SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing. *J. Comput. Biol.* **19**, 455–477 (2012).
25. Bushmanova, E., Antipov, D., Lapidus, A. & Prjibelski, A. D. rnaSPAdes: a *de novo* transcriptome assembler and its application to RNA-Seq data. *GigaScience* **8**, giz100 (2019).
26. Xie, Y. *et al.* SOAPdenovo-Trans: *de novo* transcriptome assembly with short RNA-Seq reads. *Bioinformatics* **30**, 1660–1666 (2014).
27. Park, S. *et al.* Cyclotide Evolution: Insights from the Analyses of Their Precursor Sequences, Structures and Distribution in Violets (Viola). *Front. Plant Sci.* **8**, 2058 (2017).
28. Fisher, M. F. *et al.* A family of small, cyclic peptides buried in preproalbumin since the Eocene epoch. *Plant Direct* **2**, e00042 (2018).
29. Kersten, R. D. & Weng, J.-K. Gene-guided discovery and engineering of branched cyclic peptides in plants. *Proc. Natl. Acad. Sci.* **115**, (2018).
30. Lima, S. T. *et al.* A Widely Distributed Biosynthetic Cassette Is Responsible for Diverse Plant Side Chain Cross-Linked Cyclopeptides**. *Angew. Chem. Int. Ed.* **62**, e202218082 (2023).
31. Li, D., Liu, C.-M., Luo, R., Sadakane, K. & Lam, T.-W. MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct *de Bruijn* graph. *Bioinformatics* **31**, 1674–1676 (2015).
32. Hempel, C. A. *et al.* Metagenomics versus total RNA sequencing: most accurate data-processing tools, microbial identification accuracy and perspectives for ecological assessments. *Nucleic Acids Res.* **50**, 9279–9293 (2022).
33. Chigumba, D. N. *et al.* Discovery and biosynthesis of cyclic plant peptides via autocatalytic cyclases. *Nat. Chem. Biol.* **18**, 18–28 (2022).
34. Jayasena, A. S. *et al.* Stepwise Evolution of a Buried Inhibitor Peptide over 45 My. *Mol. Biol. Evol.* **34**, 1505–1516 (2017).
35. Song, Z., Burbridge, C., Schneider, D. J., Sharbel, T. F. & Reaney, M. J. T. The flax genome reveals orbitide diversity. *BMC Genomics* **23**, 534 (2022).
36. Gomes, P. W. P. *et al.* plantMASST - Community-driven chemotaxonomic digitization of plants. Preprint at https://doi.org/10.1101/2024.05.13.593988 (2024).
37. Shen, W., Sipos, B. & Zhao, L. SeqKit2: A Swiss army knife for sequence and alignment processing. *iMeta* **3**, e191 (2024).
38. Zhang, Z. Protein sequence similarity searches using patterns as seeds. *Nucleic Acids Res.* **26**, 3986–3990 (1998).
39. Shen, Y. *et al.* Potentides: New Cysteine-Rich Peptides with Unusual Disulfide Connectivity from *Potentilla anserina*. *ChemBioChem* **20**, 1995–2004 (2019).
40. Villanueva, J. *et al.* Characterization of the wound-induced metallocarboxypeptidase inhibitor from potato[1]: cDNA sequence, induction of gene expression, subcellular immunolocalization and potential roles of the C-terminal propeptide. *FEBS Lett.* **440**, 175–182 (1998).
41. Colilla, F. J., Rocher, A. & Mendez, E. γ-Purothionins: amino acid sequence of two polypeptides of a new family of thionins from wheat endosperm. *FEBS Lett.* **270**, 191–194 (1990).

42. Broekaert, I., Lee, H. I., Kush, A., Chua, N. H. & Raikhel, N. Wound-induced accumulation of mRNA containing a hevein sequence in laticifers of rubber tree (Hevea brasiliensis). *Proc. Natl. Acad. Sci.* **87**, 7633–7637 (1990).
43. Huang, R.-H. *et al.* Two novel antifungal peptides distinct with a five-disulfide motif from the bark of *Eucommia ulmoides* Oliv. *FEBS Lett.* **521**, 87–90 (2002).
44. Kumari, G. *et al.* Molecular diversity and function of jasmintides from Jasminum sambac. *BMC Plant Biol.* **18**, 144 (2018).
45. Kumari, G. *et al.* Cysteine-Rich Peptide Family with Unusual Disulfide Connectivity from *Jasminum sambac*. *J. Nat. Prod.* **78**, 2791–2799 (2015).
46. Tam, J. P. *et al.* Ginsentides: Cysteine and Glycine-rich Peptides from the Ginseng Family with Unusual Disulfide Connectivity. *Sci. Rep.* **8**, 16201 (2018).
47. Tailor, R. H. *et al.* A Novel Family of Small Cysteine-rich Antimicrobial Peptides from Seed of Impatiens balsaminaIs Derived from a Single Precursor Protein. *J. Biol. Chem.* **272**, 24480–24487 (1997).
48. Andreev, Y. A. *et al.* Genes encoding hevein-like defense peptides in wheat: Distribution, evolution, and role in stress response. *Biochimie* **94**, 1009–1016 (2012).
49. Ghag, S. B., Shekhawat, U. K. S. & Ganapathi, T. R. Petunia Floral Defensins with Unique Prodomains as Novel Candidates for Development of Fusarium Wilt Resistance in Transgenic Banana Plants. *PLoS ONE* **7**, e39557 (2012).
50. Schrader-Fischer, G. & Apel, K. Organ-specific expression of highly divergent thionin variants that are distinct from the seed-specific crambin in the crucifer Crambe abyssinica. *Mol. Gen. Genet. MGG* **245**, 380–389 (1994).
51. Mylne, J. S. *et al.* Cyclic Peptides Arising by Evolutionary Parallelism via Asparaginyl-Endopeptidase–Mediated Biosynthesis. *Plant Cell* **24**, 2765–2778 (2012).
52. Poth, A. G., Colgrave, M. L., Lyons, R. E., Daly, N. L. & Craik, D. J. Discovery of an unusual biosynthetic origin for circular proteins in legumes. *Proc. Natl. Acad. Sci.* **108**, 10127–10132 (2011).
53. Jayasena, A. S. *et al.* Stepwise Evolution of a Buried Inhibitor Peptide over 45 My. *Mol. Biol. Evol.* **34**, 1505–1516 (2017).
54. Condie, J. A. *et al.* The biosynthesis of Caryophyllaceae-like cyclic peptides in *Saponaria vaccaria* L. from DNA-encoded precursors. *Plant J.* **67**, 682–690 (2011).
55. Mylne, J. S. *et al.* Albumins and their processing machinery are hijacked for cyclic peptides in sunflower. *Nat. Chem. Biol.* **7**, 257–259 (2011).
56. Matsubayashi, Y. & Sakagami, Y. Phytosulfokine, sulfated peptides that induce the proliferation of single mesophyll cells of Asparagus officinalis L. *Proc. Natl. Acad. Sci.* **93**, 7623–7627 (1996).
57. Fletcher, J. C., Brand, U., Running, M. P., Simon, R. & Meyerowitz, E. M. Signaling of Cell Fate Decisions by *CLAVATA3* in *Arabidopsis* Shoot Meristems. *Science* **283**, 1911–1914 (1999).
58. Martin, J. A. & Wang, Z. Next-generation transcriptome assembly. *Nat. Rev. Genet.* **12**, 671–682 (2011).
59. Pertea, M. *et al.* StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nat. Biotechnol.* **33**, 290–295 (2015).
60. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410 (1990).

61. Priyam, A. *et al.* Sequenceserver: A Modern Graphical User Interface for Custom BLAST Databases. *Mol. Biol. Evol.* **36**, 2922–2924 (2019).
62. Chekan, J. R., Mydy, L. S., Pasquale, M. A. & Kersten, R. D. Plant peptides – redefining an area of ribosomally synthesized and post-translationally modified peptides. *Nat. Prod. Rep.* **41**, 1020–1059 (2024).
63. Hattori, J., Boutilier, K. A., Campagne, M. M. L. & Miki, B. L. A conserved BURP domain defines a novel group of plant proteins with unusual primary structures. *Mol. Gen. Genet. MGG* **259**, 424–428 (1998).
64. Fleischmann, A. *et al.* Evolution of genome size and chromosome number in the carnivorous plant genus Genlisea (Lentibulariaceae), with a new estimate of the minimum genome size in angiosperms. *Ann. Bot.* **114**, 1651–1663 (2014).
65. Fernández, P. *et al.* A 160 Gbp fork fern genome shatters size record for eukaryotes. *iScience* **27**, 109889 (2024).
66. Lin, Z., Agarwal, V., Cong, Y., Pomponi, S. A. & Schmidt, E. W. Short macrocyclic peptides in sponge genomes. *Proc. Natl. Acad. Sci.* **121**, e2314383121 (2024).
67. Singh, U. & Wurtele, E. S. orfipy: a fast and flexible tool for extracting ORFs. *Bioinformatics* **37**, 3019–3020 (2021).
68. Madeira, F. *et al.* The EMBL-EBI Job Dispatcher sequence analysis tools framework in 2024. *Nucleic Acids Res.* **52**, W521–W525 (2024).
69. Mydy, L. S. *et al.* An intramolecular macrocyclase in plant ribosomal peptide biosynthesis. *Nat. Chem. Biol.* **20**, 530–540 (2024).
70. Kersten, R. D. *et al.* Large-scale transcriptome mining enables macrocyclic diversification and improved bioactivity of the stephanotic acid scaffold. University of Michigan - Deep Blue Data https://doi.org/10.7302/Z14X-5M94 (2025).
71. Wang X., Shafiq K., Ousley D.A. *et al.* Large-scale transcriptome mining enables macrocyclic diversification and improved bioactivity of the stephanotic acid scaffold. *Nat. Commun. 16, 4198* (2025).

Chapter 3 *In Silico* Burpitide Library Generation for Computational Drug Discovery and Cheminformatics

3.1 Statement of contribution

I downloaded plant transcriptomes from the NCBI SRA using applicable, available accession numbers. I *de novo* assembled the subsequent RNA-seq data into the full plant transcriptomes, representing full-length coding sequences. I performed the nucleotide blast of the transcriptome sequences against Sali-3-2 to identify BURP-domain cyclase candidates, translated the resulting hit nucleotide sequences to protein sequences with Orfipy, and performed transcriptome mining analysis of known BURPitide motifs using RepeatFinder. I constructed *in silico* RiPP libraries in 2D, 3D, and SMILES formats, and I predicted molecular descriptors for PCA and chemoinformatic analysis. I performed ligand preparation, macrocyclic sampling, molecular docking, and pose optimization using the Schrödinger suite. I performed the heterologous expression of Lyciumins in *Nicotiana benthamiana* and LC-MS/MS analysis. Dr. Jordan Hungerford and I performed the peptide extraction and preliminary purification of QPFGVFRW.

3.2 Summary

Natural products play a critical role in the development of novel pharmaceuticals. The identification of bioactive plant-derived natural products through laboratory-based methods remains challenging and labor-intensive due to the inherent complexity of plant chemistry. There is a need for robust *in silico* strategies to predict the structures and bioactivity of plant natural products. Advances in the post-genomic era have increased the availability of plant genetic data and structural biology models of drug targets. This chapter characterizes the chemical diversity of

known BURP-domain-derived ribosomally synthesized and post-translationally modified peptide (RiPP) chemotypes, generates an *in silico* library of plant RiPPs, and docks this library against relevant drug targets. The development of a computational pipeline to predict biosynthetically feasible molecules from plant genes, along with the evaluation of the BURP-domain-derived RiPP library, will facilitate the identification of compounds targeting lipoprotein signal peptidase II (LspA), a cyclic peptide antimicrobial target (PDB: 6RYO) implicated in human diseases.

As a proof-of-concept, BURP-domain peptide cyclase genes encoding cyclic peptide motifs were identified from the transcriptomes of approximately 2,500 plant species. The corresponding Burpitide macrocycle three-dimensional structures were predicted, curated into a library, and screened against LspA. LspA is a bacterial lipoprotein (BLP) post-translational processing enzyme required for full virulence in *Methicillin-resistant Staphylococcus aureus* (MRSA) and its enzymatic activity are essential for MRSA survival. Computational screening of the 375-member burpitide library against LspA yielded 12 hits, with the most notable being candidate QPFGAWRW, achieving the highest docking score (Gscore = -12.025). The peptide QPFGVFRW (Gscore = -9.857) was prioritized for biosynthetic production due to its robust interaction network, including an Arginine residue that forms direct salt bridges and hydrogen bonds with the essential Asp-Asp catalytic dyad and higher production titers in tobacco.

Antimicrobial-resistant (AMR) bacterial strains, including MRSA, represent a significant global health threat as resistance rates surpass the pace of new antibiotic discovery. Cyclic peptides demonstrating promising *in silico* interactions with disease targets were produced using transient gene expression in tobacco. This study provides bioinformatic insights into the diversity of side-chain macrocyclic peptides across the plant kingdom.

3.3 Introduction

Antimicrobial resistance (AMR) refers to the capacity of microorganisms to develop defense mechanisms against drugs intended to eliminate them[1,2]. AMR represents a critical global health threat, as the rate of resistance development surpasses the discovery of new antibiotics, thereby compromising the effective treatment of infectious diseases. Globally, AMR-related deaths

are estimated at approximately 750,000 per year and are projected to increase to 10 million by 2050[1].

Staphylococcus aureus (S. aureus) is an opportunistic Gram-positive bacterial pathogen carried by approximately one-third of the human population[2–5]. *S. aureus* is found in the normal human flora on skin and mucous membranes, as well as in the environment, and typically does not cause infections on healthy skin. However, when it enters the bloodstream or internal tissues, it can result in minor skin infections, severe infections, and potentially life-threatening diseases[5]. Approximately 5% of U.S. hospital patients carry methicillin-resistant S. aureus (MRSA) in their noses or on their skin, which can lead to skin and soft tissue infections (SSTIs), bone and joint infections, pneumonia, bacteremia, osteomyelitis, and endocarditis[2,3,5,6]. *S. aureus* is distinguished by its cocci shape, clustered arrangements, and thick peptidoglycan cell walls, contributing to its increasing resistance to first-line antibiotics. Additionally, S. aureus can enter dormant, non-growing states, enabling it to evade antibiotics and complicating efforts to target its pathways[5,6].

S. aureus can acquire resistance mechanisms from other *Staphylococci* species, such as the *mecA* gene from *Staphylococcus sciuri* (*S. sciuri*)[6,7]. The *mecA* gene encodes penicillin-binding protein 2a (PBP2a), which conferred resistance to methicillin shortly after its clinical introduction[5–7]. Methicillin-resistant *S. aureus* (MRSA) accounts for approximately 60% of all *S. aureus* infections in the United States and exhibits resistance to many β-lactam antibiotics[5]. According to the Centers for Disease Control and Prevention (CDC), there are an estimated 300,000 to 400,000 cases annually, resulting in 80,461 severe infections and 11,825 deaths, with a total annual healthcare cost of $3-4 billion in the United States[3,5].

Vancomycin remains the primary treatment for MRSA and is considered a “last resort” antibiotic because resistance develops more slowly. However, increased use of vancomycin has exerted strong selective pressure, facilitating the emergence of vancomycin-intermediate and -resistant strains[5]. Vancomycin is a non-ribosomal macrocyclic glycopeptide derived from *Streptomyces orientalis that* binds to lipid II precursors in the bacterial cell membrane and inhibits penicillin-binding protein (PBP) crosslinking of peptidoglycans[5]. Despite its potent bactericidal activity, vancomycin’s poor pharmacokinetic properties restrict its administration primarily to intravenous routes[5–9].

Daptomycin, a macrocyclic non-ribosomal lipopeptide produced by *Streptomyces roseosporus*, is employed as salvage therapy and as an alternative to vancomycin for MRSA

treatment[10,11]. Like vancomycin, daptomycin's poor pharmacokinetic properties restrict its use primarily to intravenous administration, and additional supportive measures are required to manage associated toxicities[12,13]. Given *S. aureus's* propensity to develop drug-resistant strains, such as MRSA, and the narrow therapeutic windows of current antibiotics, the identification of new compounds capable of effectively treating MRSA is imperative.

Bacterial lipoproteins (BLPs) are found in both Gram-positive and Gram-negative bacteria and perform diverse cellular functions, including cell envelope biogenesis, metal acquisition, and nutrient uptake[14–16]. Post-translational processing enzymes for BLPs are critical for Gram-positive bacterial pathogenesis, virulence, and survival. Lipoprotein signal peptidase II (LspA) is a BLP post-translational processing enzyme necessary for full virulence in MRSA, and its activity is essential for MRSA survival in human blood[14–16]. LspA is a cytoplasmic-transmembrane protein that exists as a monomer with two domains: (1) four transmembrane helices with both N- and C-termini located in the cytoplasm, and (2) a periplasmic domain composed of two subdomains[14,16]. These subdomains include the B-cradle, which extends from the helical protein core, and the periplasmic helix (PH), which is positioned at the top of the inner membrane[14–16].

LspA functions as an aspartyl endopeptidase, utilizing its Asp-Asp catalytic dyad to cleave pro-lipoproteins into a signal peptide and diacylated lipoproteins (**Figure 3.12**). These lipoproteins contribute to securing the membrane and peptidoglycan layers or are secreted into the environment to modulate the host immune response[14–18]. Inhibition of LspA results in the accumulation of lipoprotein precursors in the inner cytoplasmic membrane, ultimately leading to cell death. Targeting LspA by blocking its catalytic Asp-Asp dyad offers a novel therapeutic opportunity to address the increasing incidence of resistant infections[16].

LspA can be inhibited by globomycin, a depsipeptide isolated from S*treptomyces hagronesis,* which utilizes the hydroxyl group on its serine residue to interact with the catalytic dyad and block LspA's activity[14–16]. Globomycin demonstrates activity in gel shift and FRET assays, with IC_{50} values in the nanomolar range against recombinant LspA constructs from *Pseudomonas aeruginosa* and MRSA (LspPae and LspMrs, respectively). Targeting bacterial virulence constitutes a promising strategy for MRSA treatment, as resistance develops more slowly when virulence factors are targeted[17,18]. LspA is a particularly attractive target due to its distinct mechanism compared to existing drugs and the absence of human homologs[14–18].

Natural products are a primary source of approved medicines and lead compounds, particularly in antibiotic drug discovery. Natural products from bacteria and fungi have historically yielded numerous FDA-approved antibiotics. However, large pharmaceutical companies have recently shifted their focus away from anti-infective and natural product development toward more profitable therapeutic areas, resulting in a reduced number of antibiotics in development pipelines[10,19,20]. Consequently, in the current post-antibiotic era, the discovery and development of new compounds capable of effectively treating MRSA is of critical importance.

Antibiotic development pipelines still include several bacterial peptides with side-chain macrocyclizations, such as globomycin, darobactin, and arylomycin analogs (**Figure 3.3B)** [14,15,21,22]. Plants also synthesize diverse side-chain-macrocyclic peptides structurally similar to bacterial macrocyclic peptides with antibiotic properties, and these plant peptide structures can be accurately predicted from precursor genes[23–25]. With the increasing availability of plant genetic data in the post-genomic era, this project seeks to address the need for new antibiotics through a plant transcriptome mining approach, leveraging plants as untapped reservoirs of chemical diversity.

Ribosomally synthesized and post-translationally modified peptides (RiPPs) display a wide range of bioactivities, particularly in modulating and disrupting microbial functions[26]. For example, darobactin inhibits BamA, thereby limiting the assembly of outer membrane proteins in Gram-negative bacteria[21]. RiPPs are produced as precursor peptides encoded by structural genes containing a core peptide region, which is subsequently modified, proteolytically cleaved, and released as the mature peptide[27,28].

Side-chain-macrocyclic plant RiPPs derived from BURP-domain-peptide cyclases, termed burpitides, represent an attractive target compound class for this project because they frequently encode the peptide motif within the same polypeptide as the peptide cyclase[29]-. BURP-domain proteins are plant-specific and named after their four founding members: BNM2, USP, RD22, and PG1β[30–34]. These proteins are characterized by a C-terminal domain with a conserved N-terminal FF motif and a C-X11-12-C-X47-59-CH-X10-CH-X23-27-CH-X23-26-CH motif[25,30,35] (**Figure 3.1**).

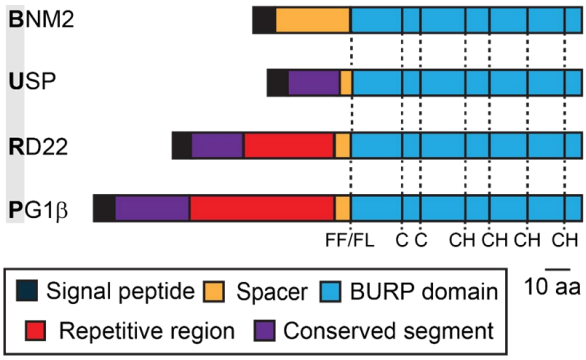


Figure 3.1 Founding members domain structure of the BURP domain family from ref[37].

BURP-domains function as precursor peptides within ribosomally synthesized and post-translationally modified peptide (RiPP) biosynthetic pathways. These domains possess an N-terminal signal peptide and either a core peptide encoded directly in the C-terminal domain or a repetitive core peptide region encoded N-terminally to the BURP domain[23–25,29,35–37] (**Figure 3.2**). BURP-domain proteins act as autocatalytic plant metalloenzymes that catalyze side-chain-to-side-chain macrocyclization reactions in RiPPs, resulting in the formation of C(sp2)-C(sp3), C(sp3)-O, and C(sp3)-N linkages through aromatic-aromatic or aromatic-aliphatic side chains[25,35].

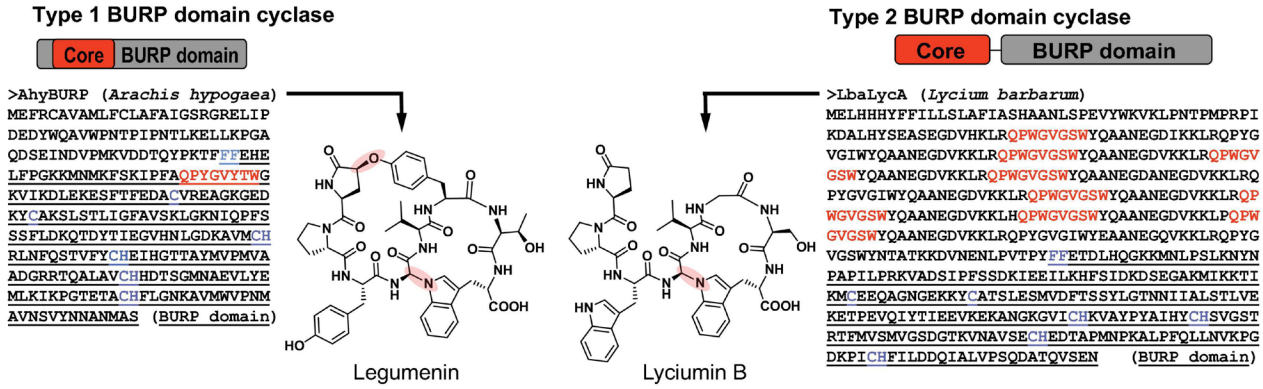


Figure 3.2 Primary structures of representatives of the two general types of BURP domain cyclases and corresponding peptide natural products.
In the BURP domain sequence, core peptides corresponding to the natural product are highlighted in red, the BURP domain sequence is underlined, the BURP domain-defining residues are highlighted in blue from ref[35].

These plant-derived macrocyclic peptides are largely unexplored scaffolds for drug discovery, development, and medicinal chemistry, although they have shown promising bioactivity as potential anti-cancer, hypotensive, anxiolytic, and antiviral therapeutics[23,38–42]. Furthermore, they have similar macrocyclic structures to the antibacterial leads globomycin, darobactin, and arylomycin (**Figure 3.3A and B**).

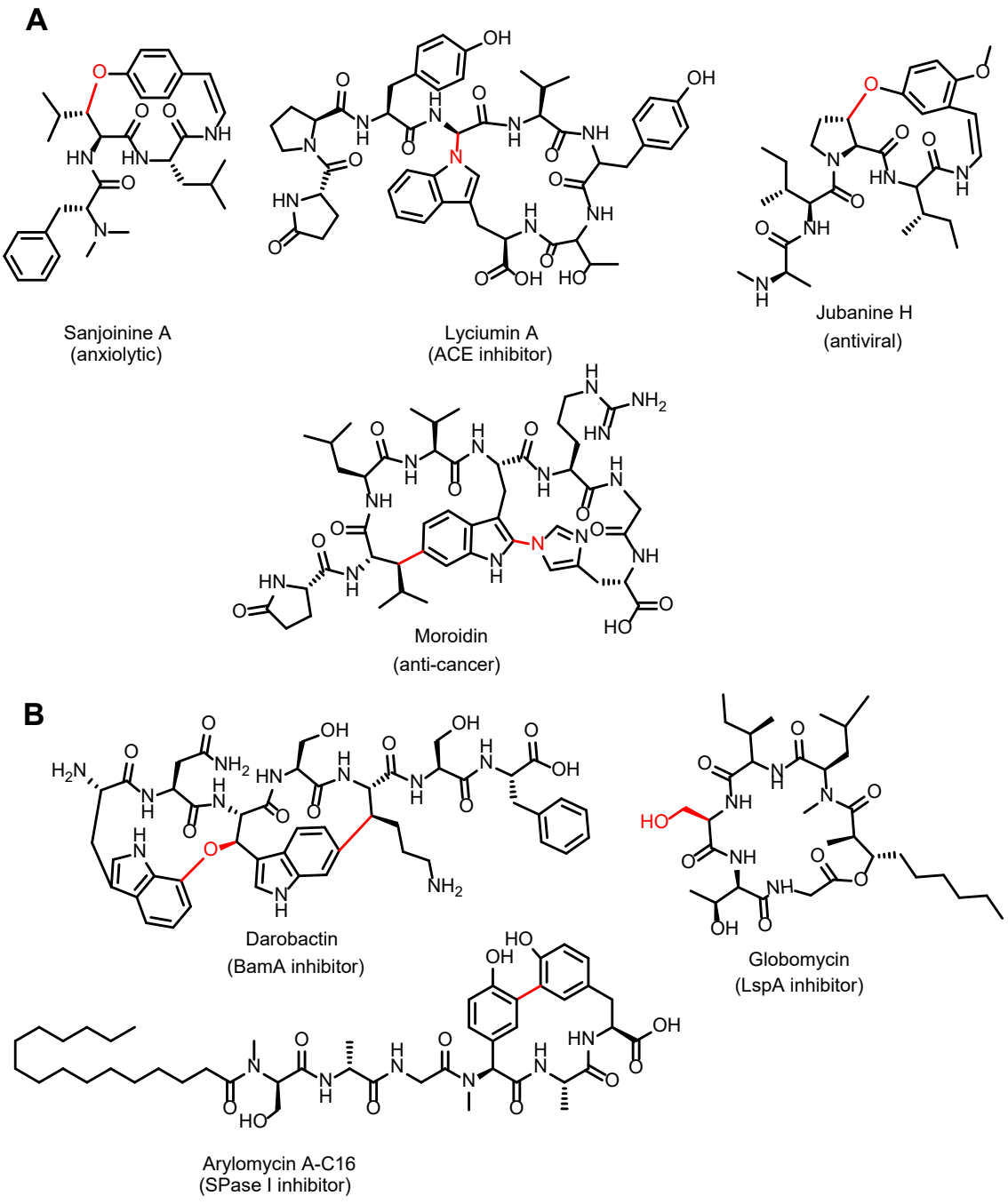


Figure 3.3 Structures of Drug-like Burpitide RiPPs.
(A) and Antibiotic Lead Structures (B) with Sidchain Macrocyclic Crosslinks Highlighted in Red (except Globomycin – red highlights serine reside inhibiting LspA catalytic dyad).

Herein, I assess and assemble RNA-seq datasets with SPAdes to generate precursor peptides from RiPP-producing plants. Furthermore, I apply computational tools to mine the transcriptomes of 2472 plant species for known burpitide chemotypes using the optimized pipeline described in Chapter 2. Finally, this chapter demonstrated the ability to translate theoretical

structures into a library of molecules that are mass spectrometry-characterizable and biosynthetically feasible to produce in a heterologous host, such as tobacco, for potential downstream purification and assay testing.

3.4 Results

3.4.1 Comparison of Megahit and SPAdes de novo Assemblers for Known RiPP Scaffold Prediction and Discovery from Large Datasets

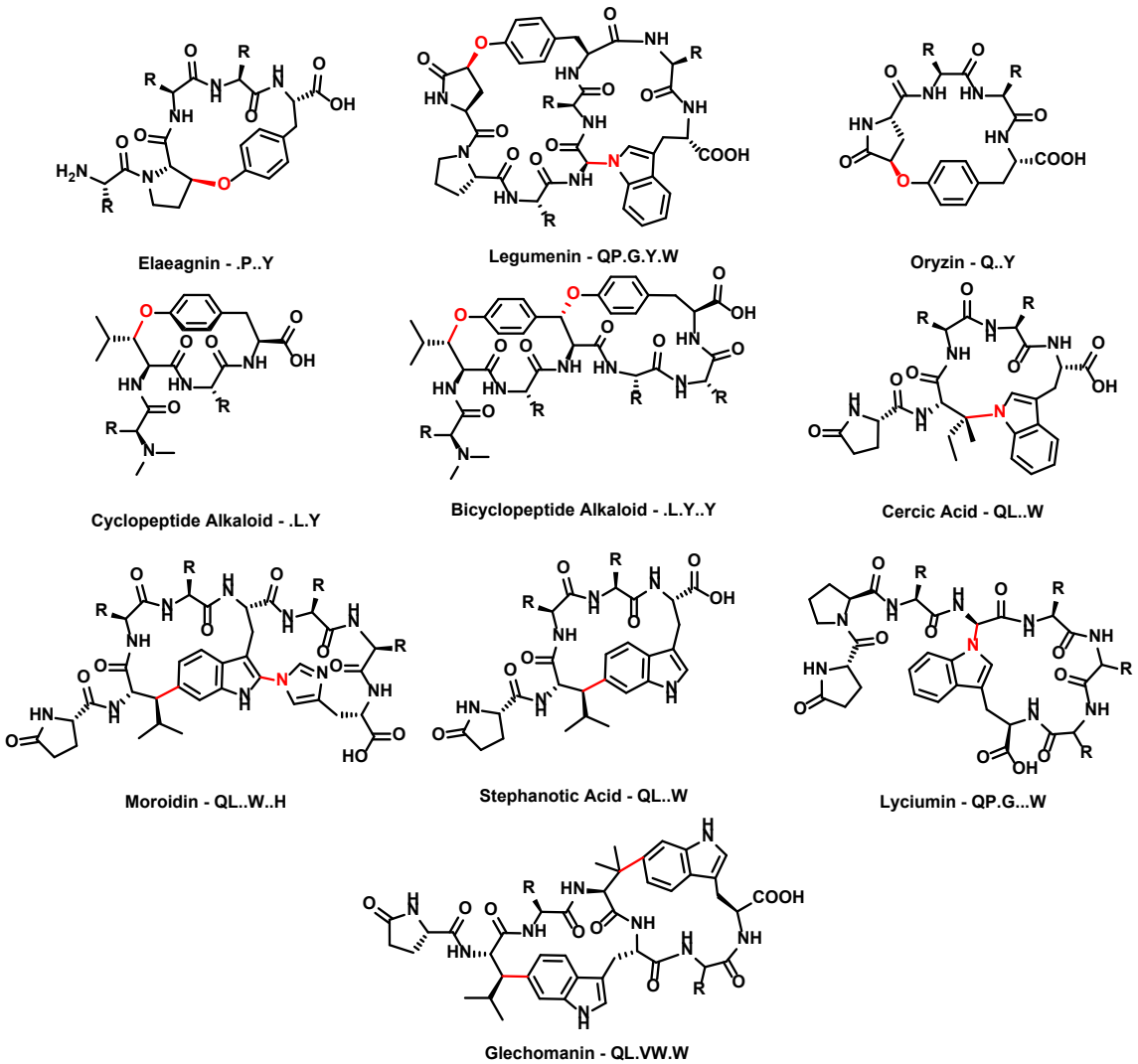


Figure 3.4 Structures of Burpitide RiPP Motifs Mined Using RepeatFinder.

Benchmarking analysis indicated that Megahit is a promising de novo assembler for RiPP discovery, with lower memory usage and assembly time than SPAdes[43–45]. To compare both assemblers for RiPP discovery in larger transcriptome datasets, 2872 plant transcriptomes representing 2427 plant species from 429 plant families were assembled using both Megahit and SPAdes. Each dataset was subsequently searched for homologous BURP-domain cyclase genes with tblastn (e-value cutoff of 1e-5), using Sali3-2, a member of the USP-type BURP-domain proteins, as the query[46–48].

BURP-domain BLAST hits were translated from RNA sequences into protein sequences using Orfipy[49]. The resulting protein sequences were analyzed with RepeatFinder, which identifies core peptide motifs in open reading frames surrounding BURP domains. RepeatFinder facilitates automated mining of core peptide motifs without relying on prior knowledge of RiPP core peptide features, thereby enabling the discovery of both known and novel chemistries in a high-throughput manner[25].

The datasets were searched for core peptides corresponding to the following BURP-domain-derived RiPPs: lyciumin, legumenin, moroidin, stephanotic acid, cercic acid, and bicyclopeptide alkaloids. Using this workflow, SPAdes assemblies revealed 204 BURP-domain-derived RiPP cores, while Megahit assemblies identified 183 (**Table 3.1**). Due to the higher yield of burpitides and computational advantages, subsequent mining with RepeatFinder focused on the SPAdes transcriptomes. The expanded search included peptide types characterized by the Kersten lab: lyciumin, legumenin, bicyclopeptide alkaloid, cyclopeptide alkaloid, cercic acid, stephanotic acid, moroidin, elaeagnin, glechomanin, and oryzin (**Figure 3.4**).

Table 3.1 Core peptide predictions by SPAdes and MEGAHIT.

BURP-domain-derived RiPP	SPAdes (v3.15.5)	MEGAHIT (v1.2.9)
Lyciumin	131	119
Legumenin	12	12
Moroidin	21	19
Stephanotic acid	9	8
Cercic acid	2	2
Bicyclopeptide alkaloid	29	23
Total	204	183

3.4.2 In Silico Library Generation and Molecular Descriptor Prediction

A total of 375 BURP-domain-derived RiPPs were extracted from SPAdes transcriptomes to construct an *in silico* library based on the motifs shown in **Figure 3.4**. For each unique plant RiPP chemotype, a 2D structure and its corresponding SMILES code were manually generated in ChemDraw. The peptide SMILES codes were then imported into Maestro to calculate drug-like property descriptors, including ALogP, molecular weight, number of rotatable bonds, chiral centers, hydrogen bond donors and acceptors, and polar surface area[50] (**Table 3.3**). These descriptors characterize peptide lipophilicity, molecular size, flexibility, chirality, intermolecular bonding capacity, and polarity. Collectively, they define the library's drug-like property space and enable evaluation against Lipinski's Rule of Five and the Beyond Rule of Five (bRo5) criteria, which are commonly applied to macrocycles[20].

These molecules occupy the Beyond Rule of Five (bRo5) chemical space, which permits violation of several of Lipinski's rules while retaining pharmacologically relevant properties[20]. Burpitides display a high number of chiral centers and elevated molecular weights, combined with a reduced number of rotatable bonds, which may stabilize the peptides in bioactive conformations. Macrocycles exhibit high polar surface area due to the peptide backbone, and side-chain-to-backbone macrocyclization of the burpitide scaffold can shield polar groups from solvent exposure. Additionally, macrocycles demonstrate chameleonic properties, enabling conformational shifts that conceal polar groups in lipophilic environments and reveal them in aqueous environments[20].

The peptides possess a high density of functional groups, constrained within the macrocyclic framework, which facilitates the formation of extensive hydrogen-bonding networks across flat or shallow protein surfaces[20]. Descriptor prediction in Maestro enabled a transition from sequence-based to pharmacophore-based analysis, expanding the range of potential targets to include undruggable proteins that are inaccessible to Ro5-compliant compounds. Macrocyclic crosslinks further enhance metabolic stability by limiting proteolytic cleavage[20]. Maestro also supported 3D structure prediction of transcriptome-predicted RiPPs, as well as *in silico* library curation and storage[50].

3.4.3 Cheminformatics Analysis Through Principal Component Analysis

Principal Component Analysis (PCA) based on a correlation matrix was employed to evaluate the structural uniqueness of the library's chemical space. PCA, a multivariate statistical technique, reduces high-dimensional data to two dimensions to reveal clustering patterns and relationships[51]. In this study, PCA was applied to physicochemical descriptors to elucidate structural relationships among diverse peptide motif groups. **Figure 3.5** demonstrates that the first two principal components (PC1 and PC2) account for 90.4% of the total variance, suggesting that the selected descriptors comprehensively represent the library's chemical diversity. The loading vectors indicate that Chiral Centers, Number of Rotatable Bonds (Nrotb), Molecular Weight (MW), and ALogP are negatively correlated with PC2, whereas H-bond donors, H-bond acceptors, and Polar Surface Area (PSA) show a strong positive correlation with PC2. This divergence in PCA chemical space underscores structural trade-offs between lipophilicity and hydrogen-bonding capacity in these macrocycles.

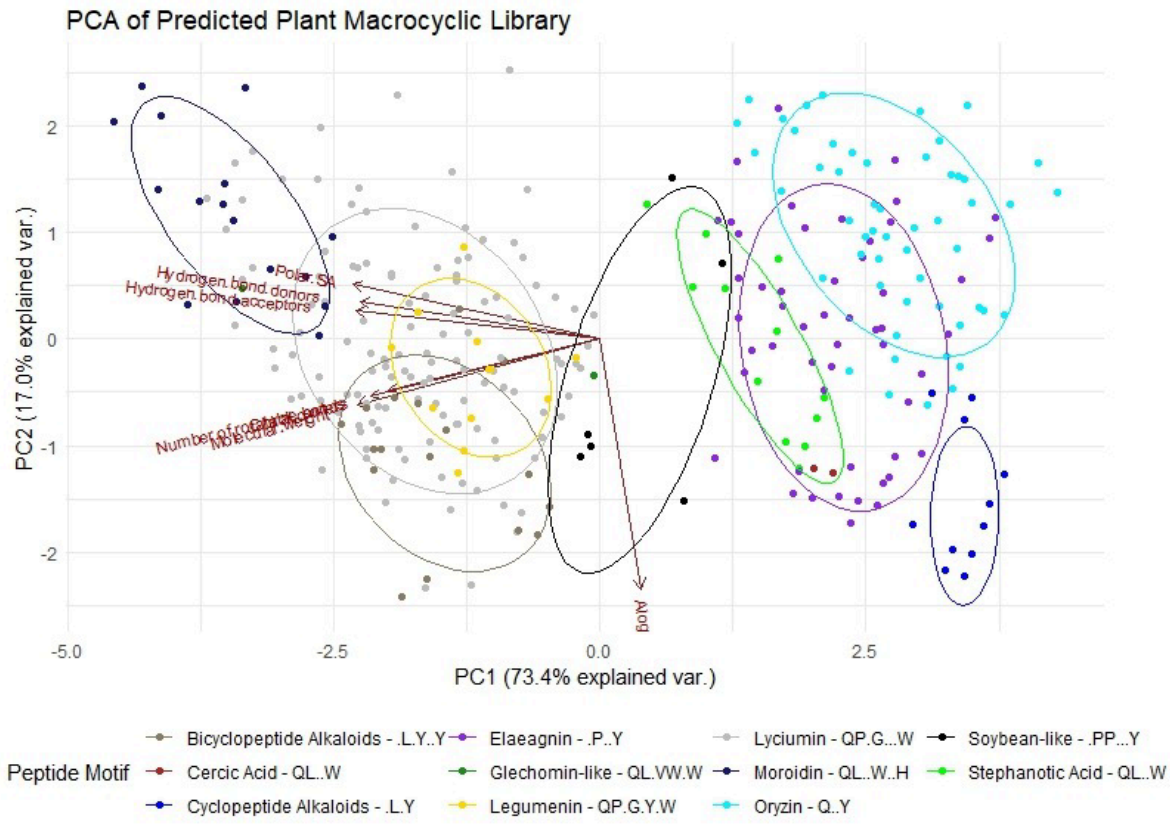


Figure 3.5 PCA of mined transcriptome BURP-domain-RiPPs.
RiPPs are clustered by peptide motif. The percentage variance explained by each principal component is given in brackets next to each axis label. Each variable is shown as a vector. Angles between vectors indicate degree of correlation.

To situate the chemical diversity of the burpitide library within the broader context of peptidic plant-derived secondary metabolites, a comparison was conducted with known orbitides, which are head-to-tail macrocyclic plant ribosomally synthesized and post-translationally modified peptides (RiPPs) sourced from CyBase[52]. PCA accounted for 86.2% of the variance in the dataset **(Figure 3.6A). O**rbitides occupy a wide and dispersed region of chemical space, whereas BURP-derived burpitides cluster at the peripheries of the PCA plot. This clustering likely reflects the more constrained descriptor ranges of burpitide scaffolds, which exhibit greater side-chain-to-backbone macrocyclization rigidity than the more flexible head-to-tail architectures of standard orbitides. The observed clustering further suggests that BURP-domain cyclases impose strict control over the physicochemical properties of their products. Box-and-whisker plots **(Figure 3.6B)** indicate that burpitides are characterized by high chiral density and optimized polar surface areas. The negative correlation between the number of rotatable bonds (Nrotb) and the

second principal component (PC2) illustrates the combined effect of high molecular weight and structural rigidity.

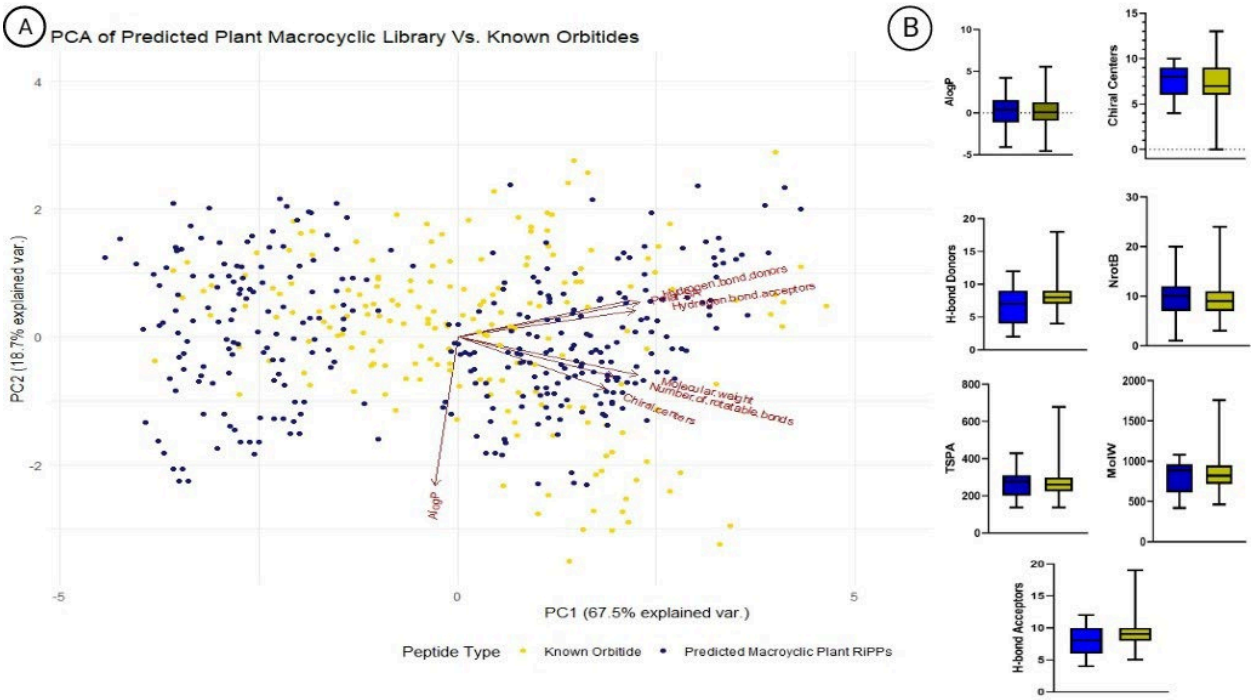


Figure 3.6 Principal component analysis of mined plant RiPPs and know orbitides from Cybase.
The percentage variance explained by each principal componentis given in brackets next to each axis label. Each variable is shown as a vector. Angles between vectors indicate their degree of correlation. (B) Box and whisker plots of chemoinformatic descriptors.

To emphasize the distinct positioning of the burpitide library within medicinal chemistry and natural product research, principal component analysis (PCA) was conducted against established databases, including DrugBank (antibacterials), COCONUT (diverse natural products), ChEMBL (macrocycles), and FDA-approved as well as clinical-stage macrocycles **(Figures 3.13-16)**[20,53–55]. Burpitides occupy a more constrained and chiral-dense chemical space, indicating increased structural rigidity and localized polarity. In contrast, COCONUT and ChEMBL display extensive structural diversity, whereas BURP-domain-derived ribosomally synthesized and post-translationally modified peptides (RiPPs) demonstrate a balance between molecular weight and polar surface area that supports biological stability. Notably, compared with DrugBank's antibacterial compounds, the burpitide library lies within the chemical space of conventional clinical antibiotics, highlighting its potential to address cross-resistance in multidrug-resistant bacteria.

The structural constraint of the library is biologically imperative for the pre-organized nature of the macrocyclic RiPPs. These constraints limit the conformational space to states resembling protein-binding pockets, whereas linear peptides have inherent conformational flexibility, thereby requiring a reduction in entropy to adopt a single bioactive conformation. The rigid scaffolds of the peptides hold functional groups in fixed orientations, allowing increased interactions within the receptor pocket, while the peptide backbone and side-chain macrocyclizations shield the protein from protease degradation. These structural constraints validate the sequence of diversity, as sequence diversity is localized within the constrained framework of the macrocycle, whereas diversity in linear sequence may cause different patterns that may reorient binding in the pocket. The constrained structure of the macrocycle remains intact regardless of sequence variability. Thus, peptides have the potential to be tuned for binding pockets by swapping residues to change physicochemical properties such as AlogP or TPSA without losing the fundamental shape of the ligand that fits the pocket.

This makes the relationship between sequence and structure more predictable and easier to model based on scaffold connectivity. Examples of this connectivity include Elaeagnin and Oryzin, which utilize aryl-ether linkages, while Moroidins and Lyciumins contain tryptophan-centered macrocyclizations, making them more likely to lock into bioactive conformations and thus relevant as computational drug probes. Dual-ring systems create complex shapes that may not be easily replicated by synthetic small molecules with lower chiral density. The cheminformatics analysis shows that the peptides occupy the periphery of the orbitide space of the PCA, with a balance of molecular weight and polar groups, suggesting that nature has evolved to create these peptides and confine them to specific chemical spaces. The comparison between DrugBank antibacterials and FDA-approved/clinical-trial macrocycles is important for drug discovery, as it characterizes the novelty of the burpitides as natural-product structures that balance the size and rigidity found in successful clinical antibiotics. Furthermore, it demonstrates that the structures are distinct from those of small antibacterial molecules and certain linear peptides, thereby enhancing the ability to target binding pockets in multidrug-resistant bacteria.

3.4.4 Preparation of In Silico Library and Docking Against LspA

The burpitide library constitutes a distinct physicochemical space characterized by specialized macrocyclic peptide scaffolds. The molecular weight of the library ranges from 400 to 1070 Da, with a mean of approximately 800 Da. In contrast, Lipinski's "Rule of 5" for oral drugs sets an upper limit at 500 Da. This observation confirms that the library exists "Beyond Rule of 5," indicating an enhanced capacity for selective pocket binding and size-based complementarity. Such properties enable these peptides to effectively target binding pockets with larger surface areas, whereas Ro5-compliant molecules often fail to engage large, flat, or deep protein binding sites. Most structures in the library possess between 8 and 10 rotatable bonds, reflecting a degree of pre-organized rigidity. Structures with a greater number of rotatable bonds may be less suitable for certain binding poses due to potential entropic energy loss. Macrocyclic crosslinking within these peptides can stabilize bioactive conformations upon binding to protein targets. Furthermore, the number of chiral centers per structure ranges from 8 to 10, which may further enhance rigidity, as these peptides are less stereochemically complex than other plant-derived cyclic peptides such as orbitides.

The restricted topological polar surface area (TPSA) range of 137-440 $Å^2$ suggests that these peptides can target polar binding pockets. Their propensity to form hydrogen-bonding networks at charged active sites may facilitate high-affinity binding through intermolecular interactions with residues such as asparagine, glutamine, and serine. The electrostatic profile is reinforced by the presence of 5 to 11 hydrogen bond donors per peptide, suggesting that those with higher donor counts can form electrostatic interactions with the backbone or carboxylate groups of aspartic acid and glutamic acid residues, thereby stabilizing the peptide orientation in an inhibitory binding pose. The presence of 7 to 12 hydrogen bond acceptors indicates that burpitides may anchor to target protein hydrogen bond donors, strengthening bonding networks through salt bridges with arginine or lysine residues. AlogP values spanning from -4.09 to +4.23 demonstrate that these peptides can access both polar hydrophilic and hydrophobic environments, enabling targeting of diverse biological compartments. Candidates with higher AlogP values may preferentially target hydrophobic intramembrane protein environments, whereas those with lower AlogP values are better suited for targeting soluble extracellular loops. Collectively, these descriptors enable the library to be evaluated beyond simple shape matching, supporting physicochemical complementarity to binding pockets based on rigidity, electrostatics, and polarity.

To assess the therapeutic potential of the burpitide library in antibacterial drug discovery, inhibition of Lipoprotein signal peptidase II (LspA) was selected as a proof-of-concept target, given that LspA is a post-translational processing enzyme in *Staphylococcus aureus*. A structure-based virtual screen was conducted using the X-ray crystal structure of LspA in complex with the macrocyclic antibiotic globomycin (PDB: 6RYO; see **Figure 3.7**)[14,16]. LspA is an essential aspartyl endopeptidase that employs an Asp-Asp catalytic dyad to cleave signal peptides from pro-lipoproteins (see **Figure 3.12**). In Methicillin-resistant *S. aureus* (MRSA), inhibition of LspA results in the lethal accumulation of lipoprotein precursors in the cytoplasmic membrane, establishing it as a high-value anti-virulence target for potentially mitigating antibiotic resistance[15–18].

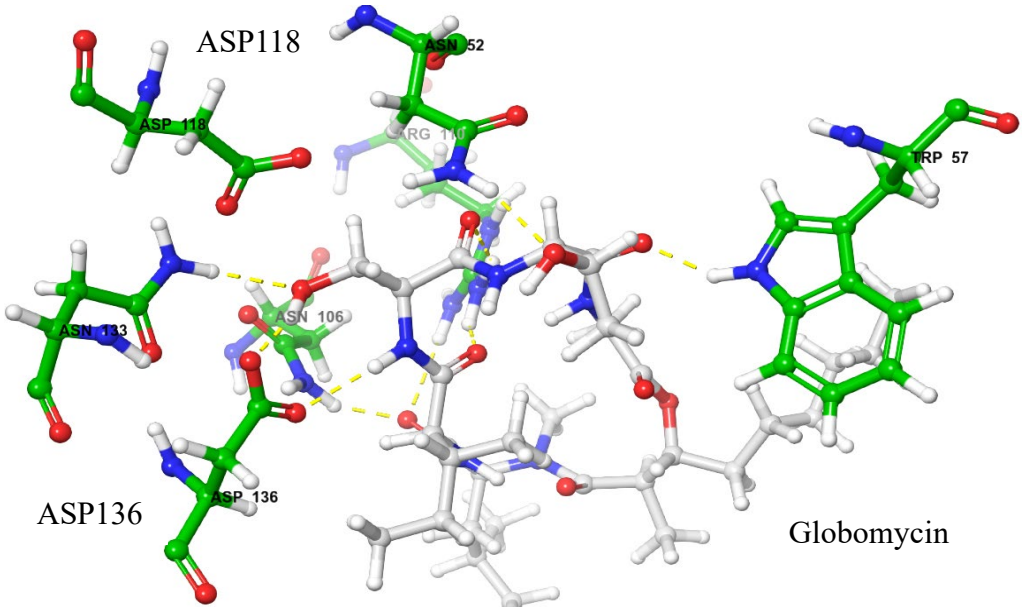


Figure 3.7 X-ray crystal structure of LspA in complex with the macrocyclic antibiotic globomycin (PDB: 6RYO)

The X-ray crystal structure was imported into Maestro, and the active site of LspA, as specified in the Receptor Grid Generation panel, was used to generate a grid file for ligand docking. The cognate ligand, globomycin, was energy-minimized using LigPrep and subsequently redocked into the LspA active site, yielding an RMSD of 1.75 Å relative to the original crystal structure. This result confirmed the accuracy of the docking grid for the library. The predicted burpitide library was also energy-minimized using LigPrep. Docking was performed using the Glide

Standard Precision (SP) module with Prime macrocyclic sampling enabled to capture conformations involving side-chain-to-backbone crosslinks.

After the initial docking, poses were visually inspected to identify specific interactions with the catalytic dyad. Prime MM-GBSA (Molecular Mechanics-Generalized Born Surface Area) was then used to calculate the predicted ligand binding energy (ΔGbind). Results were evaluated using two scoring metrics: (1) Glide score (Gscore), which quantifies the quality of the docking fit, and (2) Prime MM-GBSA (ΔGbind), which estimates the free binding energy of the protein-ligand complex. For both metrics, more negative values indicate stronger predicted affinity. The *in silico* screen identified several hits with superior binding scores compared to globomycin **(Table 3.2).** All twelve hit peptides achieved better Glide scores than the reference compound, suggesting improved geometric and electrostatic fit within the LspA hydrophobic tunnel and catalytic site.

To establish the discriminatory power of the docking pipeline and validate the -9 kcal/mol Gscore threshold, a retrospective screening campaign was conducted using a curated set of benchmark ligands, unrelated negative controls, and property-matched decoys. The known binders globomycin and myxovirescin A were selected as positive controls. While globomycin (the cognate ligand for PDB 6RYO) was successfully redocked, myxovirescin A showed lower Glide and Prime MM-GBSA scores. This is likely because myxovirescin A is co-crystalized with an LspA state in the PDB (6RYP -2.3 Å) that has a ~1.10 Å Cα-RMSD shift difference as compared to the LspA structure co-crystalized with globomycin when they are overlaid through sequence alignment. This suggests that the pocket undergoes subtle rearrangements to accommodate different macrocyclic scaffolds, with the 6RYO-derived receptor grid being optimized for globomycin-like peptidic architectures. The CXCR4 antagonist Plerixafor and the kinase inhibitor Lorlatinib were used as non-binding negative controls. No poses were generated from the docking of these negative controls. Given the structural overlap in chemical space between the orbitides and burpitides, 40 orbitides were selected as property-matched decoys for docking. These molecules share the macrocyclic peptide scaffold but lack certain pharmacophoric requirements for LspA inhibition. None of the orbitides—including the highest-scoring candidate, segetalin C (Gscore = -8.631 kcal/mol)—met the -9 kcal/mol cutoff. While some peptides contacted the ASP-ASP catalytic dyad, none had glide scores below the -9 kcal/mol cutoff and were not carried forward to Prime MM-GBSA pose optimization. This result confirms that the docking pipeline

effectively discriminates between general plant macrocyclic peptides and targeted LspA binders, reinforcing the validity of the -9 kcal/mol threshold for prioritizing hits.

Table 3.2 Top-scoring Predicted Peptides from LspA Docking Screen

Peptide	Glide Score (kcal/mol)	Prime MM-GBSA (ΔG kcal/mol)
QPFGAWRW	-12.025	-91.41
QPFGTRIW	-11.146	-87.67
QPYGVIRW	-10.878	-56.17
QPFGTRVW	-10.854	-80.32
QPFGTKIW	-10.569	-84.70
QPFGAWKW	-10.226	-81.03
QPFGVRFW	-9.886	-70.16
QPFGVFRW	-9.857	-87.90
QPFGQRSW	-9.827	-108.88
QPFGIRIW	-9.603	-76.17
QPFGALRW	-9.285	-89.04
QPFGAAKW	-9.029	-92.43
Globomycin	-7.934	-95.79
Myxovirescin A	-5.379	-64.58
Plerixafor	No pose generated	-
Lorlatinib	No pose generated	-

Within the burpitide library, QPFGAWRW achieved the highest Glide score (-12.025), surpassing globomycin. Prime MM-GBSA analysis for QPFGQRSW indicates stable binding energy, primarily driven by electrostatic and van der Waals interactions between the macrocycle and the hydrophobic surface area of LspA **(Figure 3.17**). **Alt**hough QPFGAWRW demonstrated the highest overall Glide score, QPFGVFRW (Gscore = -9.857) was selected for subsequent biosynthetic production and purification. This decision was based on its favorable combination of docking score, higher initial heterologous expression levels in tobacco, and a protein-ligand interaction network that anchors the peptide within the LspA active site **(Figure 3.8)**.

The Arginine residue of the peptide forms hydrogen bonds and salt bridges directly with the Asp-Asp catalytic dyad, suggesting the potential to inhibit the enzyme's proteolytic activity. The C-terminal Tryptophan residue engages in aromatic hydrogen bonds and a cation-pi interaction with Arginine-110 and Threonine-71 (**Figure 3.8)**. An aromatic hydrogen bond is also established between the N-terminal pyroglutamate and Phenylalanine-66. Additionally, the structure is stabilized by multiple hydrogen bonds between the peptide backbone and other active

site residues (**Figure 3.8)**. The frequent occurrence of C-terminal Tryptophan (W) residues among the top candidates underscores the importance of aromatic crosslinks in stabilizing the peptide-protein interface via cation-pi and pi-stacking interactions with active-site residues such as Arg110 and Trp57. Certain peptides, such as QPFGAAKW, displayed higher predicted binding affinities by MM-GBSA despite lower Glide scores. This observation may indicate potential overfitting during the docking process or limitations of the MM-GBSA force field in capturing all relevant conformational interactions.

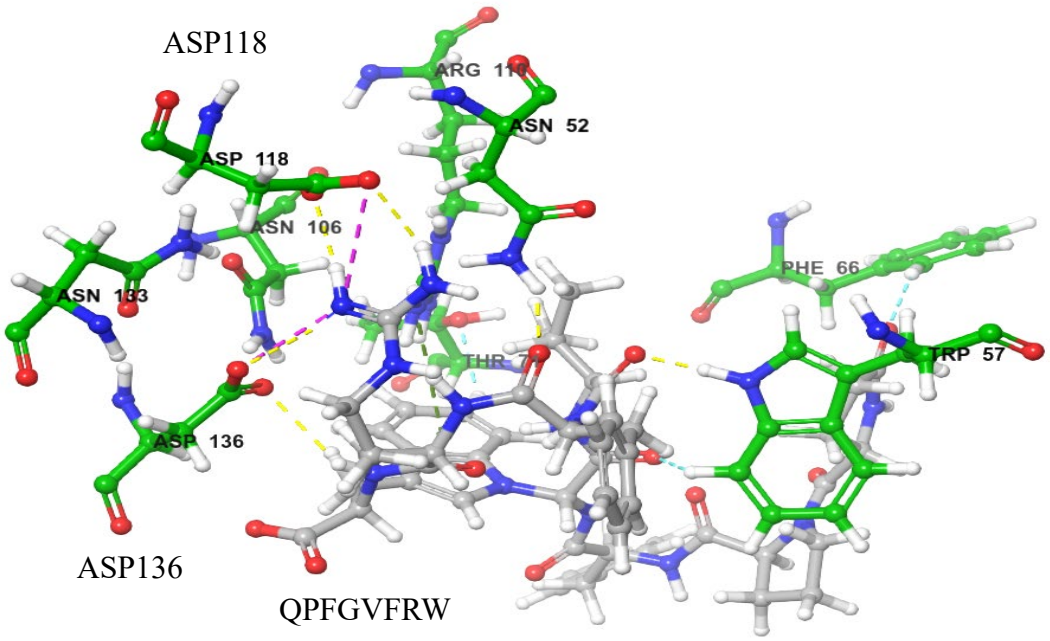


Figure 3.8 Docked structure of LspA in complex with QPFGVFRW.

Prioritizing QPFGVFRW increased the likelihood of successful *in vitro* production. These findings offer computational validation for advancing from sequence analysis to the *in vitro* production phase, thereby bridging the gap between transcriptomic prediction and targeted antibacterial discovery.

3.4.5 Tobacco Expression and Mass Spectrometry based Characterization of Biosynthetically Feasible Lyciumins

To assess the biosynthetic feasibility of the high-scoring docking hits, these candidates were transitioned from computational models to heterologous expression in a plant-based system. The lyciumin candidates were incorporated into the established Lyciumin A precursor (LbaLycA) from *Lycium barbarum,* utilizing the BURP-domain cyclase to ensure that the non-native core peptides were processed within a structural context recognized by the enzymatic machinery[24]. Each candidate was cloned as a triplicate core peptide repeat within the N-terminal domain of the *LbaLycA* precursor gene and inserted into the high-yield pEAQ-HT plasmid vector for Agrobacterium-mediated transient expression in *Nicotiana benthamiana*[24,56,57].

After transformation into *Agrobacterium tumefaciens*, 25mL starter cultures were cultivated and pressure-infiltrated into the surfaces of young tobacco leaves using a 1 mL syringe. Following a 7-day incubation, the leaves were harvested, flash-frozen, and ground into a fine powder. Peptides were extracted with methanol and analyzed by Liquid Chromatography-Mass Spectrometry (LC-MS) for characterization and quantification using extracted ion chromatograms (EICs) (**Figure 3.9)**.

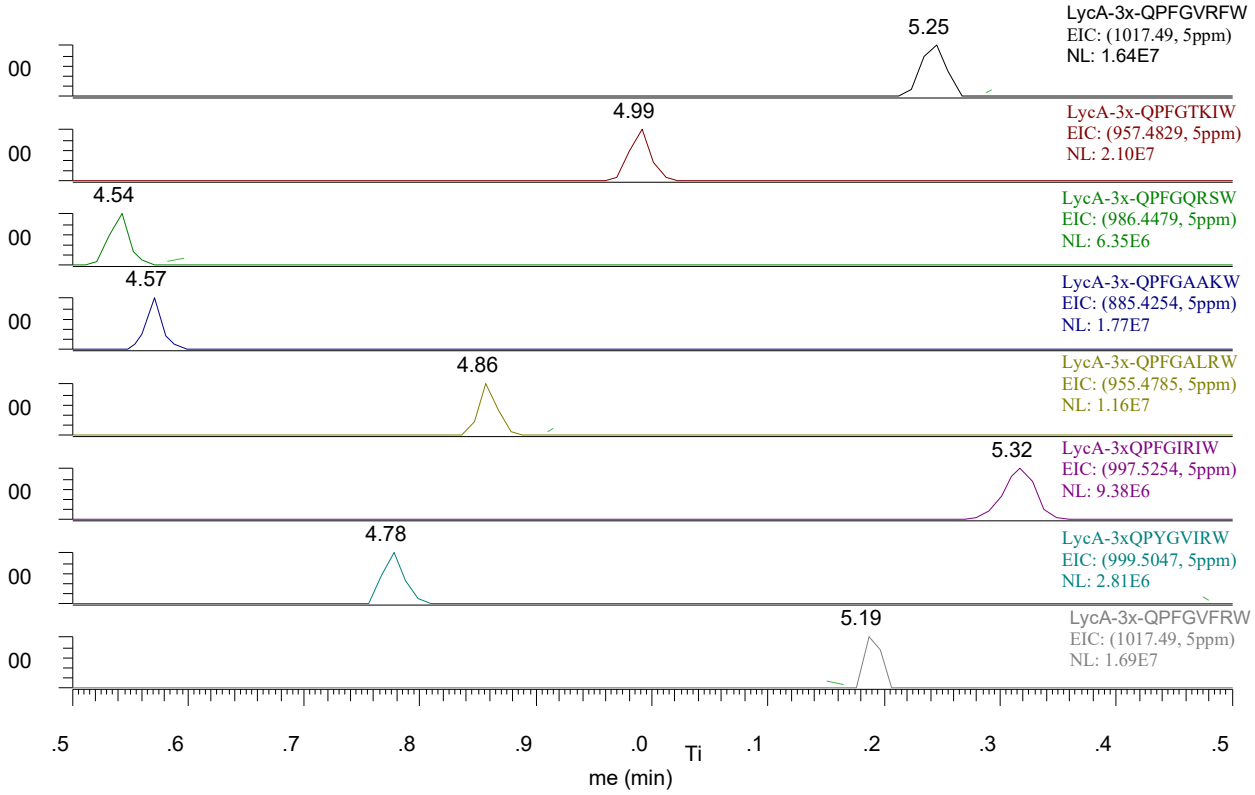


Figure 3.9 Lyciumins' analyzed by Liquid Chromatography-Mass Spectrometry (LC-MS) for characterization and quantification using extracted ion chromatograms (EICs).

The heterologous expression system yielded eight of the twelve lyciumin hits in tobacco. Quantification of expression levels by ion count indicated variation in biosynthetic efficiency among the different core sequences (**Figure 3.10**). The candidates QPFGAAKW and QPFGVFRW exhibited the highest yields within the expressed library. The high expression titer of QPFGVFRW further establishes its status as the primary hit for this scaffold within the library.

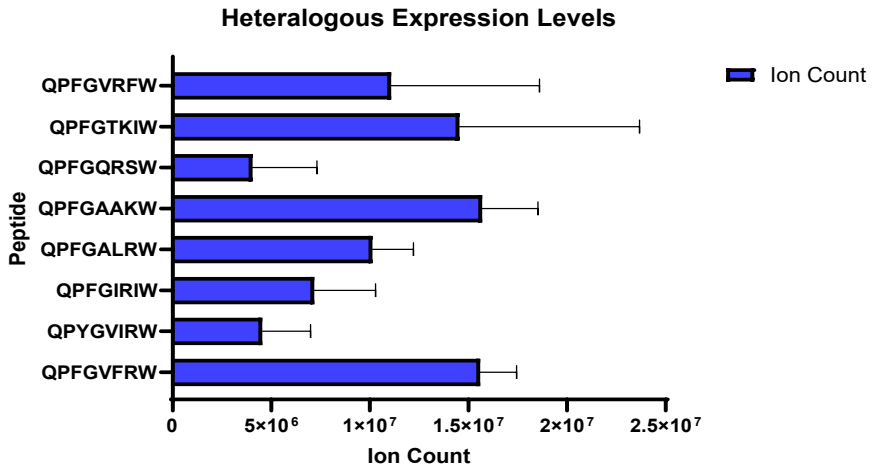


Figure 3.10 Quantification of Lyciumin Expression analyzed by Liquid Chromatography-Mass Spectrometry (LC-MS) using extracted ion chromatograms (EICs).
The bar graph is a representation of small scale lyciumin heterologous expression in tobacco leaves. The blue bars are quantified by NL values from LC-MS analysis in EIC's. The length of the bar is representative of the mean expression level where replicates is n=3. The error bars represent one standard deviation.

Due to the high expression levels of QPFGVFRW, yield optimization was performed to facilitate downstream purification. Additionally, a second *LbaLycA* precursor gene expression construct containing a 12x core peptide repeat of the QPFGVFRW sequence in the pEAQ-HT vector was designed to increase the concentration of the mature macrocycle per gram of leaf tissue (**Figure 3.11**).

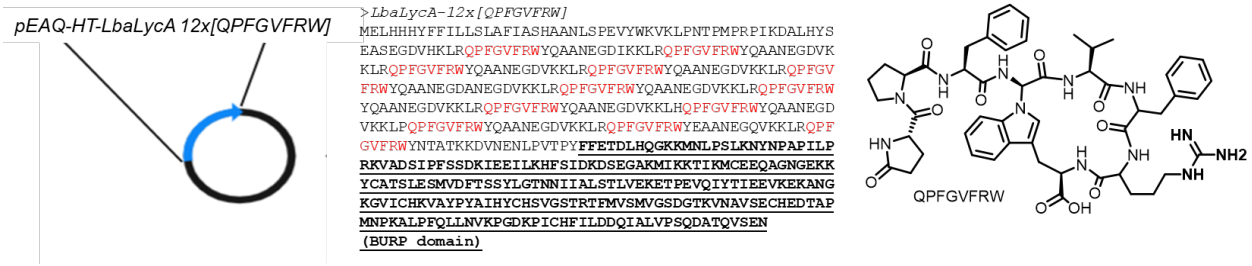


Figure 3.11 ***LbaLycA*** **precursor gene expression construct containing a 12x core peptide repeat of QPFGVFRW.**

After transformation into *Agrobacterium tumefaciens*, 2-liter cultures were cultivated and pressure-infiltrated into tobacco leaf surfaces using a 1 mL syringe. Following a 7-day incubation period, leaves were harvested, flash-frozen, and ground into a fine powder. Peptide extraction was performed using methanol, followed by partitioning with hexane, ethyl acetate, and *n*-butanol. The presence of QPFGVFRW in the *n*-butanol extract was confirmed by LC-MS analysis. The

rotovaped *n*-butanol extract was then resuspended in methanol and subjected to size-exclusion chromatography on Sephadex LH-20 for isolation.

The target peptide was lost during purification, most likely due to irreversible binding or degradation within the stationary phase. If purification had been successful, the workflow would have advanced to high-performance liquid chromatography (HPLC) for final purification, followed by structural characterization using two-dimensional nuclear magnetic resonance (2D-NMR). The intended functional validation of QPFGVFRW included LspA enzymatic gel-shift and FRET activity assays, as well as testing against a clinically relevant MRSA strain to determine minimum inhibitory concentration (MIC) values. Future work would then focus on developing lipopeptides by proteolytic cleavage of solvent-exposed lyciumin residues identified through molecular docking, followed by attachment of aliphatic carbon chains via amide coupling reactions to generate lipopeptide-based therapeutics for further antibacterial testing.

3.4.6 Future Experimentation

To address the challenges of sampling constrained macrocycles to avoid introducing non-physical strain, calculating conformational strain with MacroModel in Schrodinger will highlight the energy penalty that the burpitides incur for adopting specific conformations relative to their respective global minima. Sampling the identified confirmations will reveal energy differences in kcal/mol; strains with energy greater than 10 kcal/mol are considered strained and non-physical. Extracting the docked coordinates of the lyciumin hits and performing the MacroModel sampling energy calculation will indicate how far the poses deviate from their low-energy states. Furthermore, use of the Prime Macrocyclic Sampling module beyond the docking stage will help refine macrocycles, as it relies on a cyclic coordinate descent algorithm to ensure side-chains are sampled while the backbone remains geometrically closed and energetically minimized.

Also, a comparative study is needed between cyclic peptides and their respective linear structures. This would involve generating and modeling the linear peptides and performing conformational searches for both linear and cyclic peptide forms. Calculating the change in cyclization energy by subtracting the linear energy minimum from the cyclic energy minimum would quantify the entropic and enthalpic cost of peptide cyclization. Higher strain in the cyclic form would suggest that the peptide is forced into shapes that the linear peptide cannot reach.

These experiments would help prioritize computationally identified hits that show high-affinity binding and are thermodynamically feasible.

To expand the docking studies, future experiments are also needed. Docking the native biological prolipoprotein signal peptide into the LspA pocket and comparing the binding pose and energetics of the burpitides with those of the natural substrate would be beneficial. Furthermore, expanding the docking of known-binding positive controls to include a series of known synthetic globomycin analogues would help validate the accuracy of the predictions. Further comparison of the burpitides against a background set would be beneficial. This can be done by using Schrödinger's Canvas or RDKit to generate a vast library of property-matched decoys that are comparable in size to the docked library (~500-1000 ligands) and synthetic-like macrocycles to match the burpitide library's molecular weight, AlogP, hydrogen bond donors/acceptors, and rotatable bonds. This background set of decoy matches should be docked using the same Glide SP/Prime workflow to demonstrate that the hit rate and mean Glide scores of the burpitide library are superior to those of decoys showing evidence of ligand enrichment.

To address the trade-off between fidelity and throughput, future work should include a hierarchical funnel with more downstream high-fidelity physics-based modeling. Molecular dynamics simulations of the 12 identified hits run for 100-500ns using Desmond would help evaluate the pose stability and show the ability to maintain Asp-Asp catalytic dyad contacts throughout the simulation. Furthermore, free energy perturbation (FEP) calculations using FEP+ would help determine the absolute binding free energy and provide greater confidence than Glide scores and MM-GBSA calculations. Finally, further wet-lab experiments for in vitro validation will be needed upon successful purification of QPFGVFRW. Testing for LspA enzymatic inhibition using purified LspA in a FRET-based cleavage assay with the fluorogenic prolipoprotein substrate would help measure any decrease in fluorescence as a function of burpitide concentration and determine the IC50. Additionally, testing purified burpitide against a panel of MRSA strains would help determine the minimum inhibitory concentration (MIC) by observing growth inhibition or cessation, thereby validating the antibiotic's efficacy.

3.5 Discussion

This chapter presents a comprehensive pipeline that connects raw transcriptomic data to the rational design of bioactive plant macrocyclic peptides. It demonstrates that BURP-domain-derived ribosomally synthesized and post-translationally modified peptides (RiPPs) can be systematically identified through high-throughput transcriptomic analysis, mapped using advanced cheminformatics, and used for antibacterial drug discovery via structure-based virtual screening. The workflow initiates with *de novo* assembly of Sequence Read Archive (SRA) accessions into full-length coding sequences, followed by targeted bioinformatic RiPP motif searches using RepeatFinder and peptide translation with Orfipy[25,49,58].

This approach identified 375 unique BURP-domain-derived sequences, thereby expanding the known structural diversity of burpitides. In contrast to traditional natural product discovery, which depends on isolating compounds from the biomass of a limited number of plants, this genetic-first strategy captures biosynthetic potential from thousands of diverse plant species. Consequently, the primary limitation of the library shifts from the abundance and availability of source plants to considerations of biosynthetic feasibility[59].

Simplified Molecular Input Line Entry System (SMILES) strings facilitated the translation of peptide sequences into computationally recognizable chemical notation. These strings are integral to the pipeline, enabling rapid conversion of two-dimensional structures derived from amino acid sequences into three-dimensional molecular representations compatible with high-throughput medicinal chemistry screening software for predicting drug-like properties. Principal Component Analysis (PCA) was employed to manage the multidimensional data and to map the burpitide library against drug and chemical space databases, including DrugBank and ChEMBL[20,51–55]. This analysis demonstrated that the macrocycles occupy a privileged chemical space. The library exhibits high chiral density and structural rigidity, characteristics that are advantageous for targeting flat or complex protein interfaces that are typically inaccessible to traditional small molecules. Additionally, the library shows significant overlaps with known antibacterial drug chemical space, indicating its potential utility.

The conformational complexity of side-chain-macrocyclic peptides was addressed by conducting molecular docking of the burpitide library against Lipoprotein signal peptidase II (LspA) using the Schrödinger Maestro suite[50]. The suitability of the 10 × 10 × 10 Å grid size was validated by a redocking experiment using the macrocyclic antibiotic globomycin as a structural control. While larger grids are employed, this targeted approach focuses on the defined LspA

binding pocket. The successful pose prediction in molecular docking demonstrated that the grid was sufficient to encompass all critical intermolecular interactions and pharmacophoric features of the macrocyclic ligand. Thus, since globomycin is a macrocycle similar to the burpitides, a 10 Å grid is more than adequate for your plant peptides. Furthermore, the grid is site-specific to the active site, and a larger grid in the known pocket would decrease efficiency, as energy could be wasted calculating poses in the solvent or on protein surface residues far from the active site. The identification of the peptide QPFGVFRW as a hit underscores the predictive pipeline's capacity to surpass the clinically relevant benchmark, globomycin, in docking performance. The observed salt-bridge and hydrogen-bonding interactions between the peptide's Arginine residue and the LspA Asp-Asp catalytic dyad provide a structural basis for the anti-virulent potential of these scaffolds against LspA in methicillin-resistant Staphylococcus aureus (MRSA)[14–18].

There is precedent for this work within academic research. A study by Oselusi et al. used Schrödinger software to screen a library of natural products against proteins involved in peptidoglycan cell wall biosynthesis[62]. This study establishes the use of target-based screening approaches to identify natural products with bioactivity against bacterial machinery, though it is based on non-peptides. They used a library of 1518 mixed-scaffold natural products from the COCONUT database and identified 64 compounds at a 4.2% hit rate, but mimicked the filtering in this chapter by using Glide scores and Prime MM-GBSA to narrow down hits. Another study by Xu et al. used ~2,000 peptide fragments and medium-sized peptides to address antibiotic resistance by targeting beta-lactamase, yielding a hit rate of ~11%, which is higher because the peptides were derived from known protein-binding surfaces[63]. Mustafa et al. used a library of 11 plant antimicrobial peptides (AMPs) to screen against a panel of bacterial targets, yielding 4 peptide hits for a 36% hit rate, demonstrating the effects of biased libraries of validated plant peptides and justifying the choice of targets[64]. Here, in this chapter, 12 peptides were identified from screening the 375-ligand library, yielding a hit rate of 3.2%, suggesting that the docking pipeline is effective at filtering for shape complementarity.

The successful heterologous expression of eight out of twelve constructs in *Nicotiana benthamiana* confirms the biosynthetic feasibility of in silico predictions[56]. The use of a twelvefold core peptide repeat *LbaLycA* precursor enabled the high expression titers required for extraction, purification, and isolation[24]. Furthermore, it is hypothesized that product yield titers for other lyciumins could be improved by testing other candidate lyciumin cyclases from transcriptomes

mined in chapter 2 The loss of the QPFGVFRW lead during Sephadex LH-20 purification highlights a technical challenge in downstream processing. This research advances the field beyond rational drug discovery and design of head-to-tail cyclic peptides by introducing side-chain-to-backbone macrocyclizations. By demonstrating that bioactive hits can be identified from genetic sequences for computational screening and subsequently optimized through biosynthetic engineering of plant RiPP precursors, this work establishes a new paradigm in plant natural product discovery. The pipeline serves as a tool to transform the genetic diversity of the plant kingdom into a programmable and biosynthetically feasible resource for drug design.

3.6 Methods and Materials

3.6.1 Materials

All chemicals and reagents were from Thermo Fisher Scientific Inc. if not otherwise noted. All LC-MS solvents were Optima LC-MS-grade solvents (Fisher Scientific Inc.). All synthetic gene fragments were from Twist Biosciences Inc. Deoxynucleotide primers were from Integrated DNA Technologies Inc. LC-MS analysis was performed on a Thermo QExactive Orbitrap mass spectrometer coupled to a Vanquish Ultra high performance liquid chromatography instrument. LC-MS data was analyzed with QualBrowser from Thermo Xcalibur software package (v4.3.73.11, Thermo Scientific).

3.6.2 Plant Cultivation

Nicotiana benthamiana were grown from seeds in Premier Horticulture Pro Mix HP High Porosity with Mycorrhizae for plant growth under plant growth lights with a 16 h light/8 h dark cycle for 4–12 weeks at room temperature. *N. benthamiana* seedlings were transferred once to larger pots with the same soil and added Osmocote fertilizer.

3.6.3 Peptide Chemotyping

Leaf and stem plant material (0.2 g fresh weight) was collected from aerial tissues, frozen and ground with a MP Biomedicals FastPrep-24-5G Tissuelyser in 2-mL MP Biomedicals tubes with 2.3-mm Zirconia beads for 60 s at 6 m/s. 1 mL of 80% methanol was added to ground tissue and ground for 20 s at 6 m/s. The methanolic extract was incubated for 10 min in a 60 °C water bath, centrifuged for 5 min at 16,000 × g at room temperature, and filtered through Whatman 0.2 µm syringeless LC-MS filters (UN503NPEORG) before LC-MS analysis. Filtered extracts were transferred to LC-MS vials and analyzed by LC-MS/MS with the following settings: LC-Phenomenex Kinetex® 2.6 µm C18 reverse phase 100 Å 150 × 3 mm LC column; LC gradient, solvent A, 0.1% formic acid; solvent B, acetonitrile (0.1% formic acid); 0 min, 10% B; 5 min, 60% B; 5.1 min, 95% B; 6 min, 95% B; 6.1 min, 10% B; 9.9 min, 10% B; 0.5 mL/min, 30 °C; MS, positive ion mode; full MS, resolution 70,000; mass range 400–1200 m/z; dd-MS2 (data-dependent MS/MS), resolution 17,500; AGC target 1 × 105, loop count 5, isolation width 1.0 m/z, collision energy 25 eV and dynamic exclusion 0.5 s. Sample numbers were n = 3 (biological replicates) for each transient gene expression experiment in N. benthamiana. The negative control of transient gene expression experiment in N. benthamiana was an empty pEAQ-HT vector infiltration experiment.

3.6.4 RNA-Seq data download

Paired-end RNA-seq data were downloaded from NCBI SRA via fasterq-dump (--split-files) of the SRA Toolkit (v2.10.9)[60,61] to the Great Lakes High Performance Computing (HPC) Cluster at the University of Michigan, Ann Arbor.

3.6.5 Transcriptome assembly

Datasets were assembled de novo with SPAdes (v3.15.5)[43,44], or MEGAHIT (v1.2.9)[45] with the parameters in scripts specified on a GitHub repository (https://github.com/UM-KerstenLab4009/Moroidin-Transcriptomics). The working memory for all datasets was 48 GB.

For large-scale de novo assembly, the datasets were assembled with SPAdes on the Great Lakes HPC cluster. Assemblies which failed at 48 GB memory were assembled at 180 GB memory.

3.6.6 Transcriptome analysis

Large-scale search of BURP-domain transcripts was performed either via Sequenceserver BLAST and subsequent RepeatFinder analysis[46,62]. For Sequenceserver-based BLAST search, all SPAdes de novo assembled transcriptomes were combined into 100 transcriptome-databases and searched via tblastn (BLAST+ v2.16.0 +) for homologs of Sali3-2 as a BURP-domain-containing protein query (GenBank ID AAB66369) on Sequenceserver (v3.1.0) with the following BLAST parameters: E-value 1e-05, matrix BLOSUM62, gap-open 11, gap-extend 1, filter L, max, max_target_seqs 5000. tblastn hits from all databases were combined into a fasta-file and translated into protein sequences via orfipy (v0.0.4)[49] with the following parameters: --pep --min 450 --between-stops. The resulting protein sequence pep file was converted to a fasta file and analyzed by standalone RepeatFinder on the Great Lakes HPC Cluster at the University of Michigan, Ann Arbor (see GitHub repository for RepeatFinder settings). Candidate burpitide cyclases were characterized in RepeatFinder output and presence in an open reading frame, which encoded a BURP-domain motif. In our searches, a BURP-domain motif was defined by the presence of at least half the BURP-domain-conserved residues (i.e., FF/Fx-C-C-CH of FF/Fx-C-C-CH-CH-CH-CH)[37]. Hits with a single core peptide at the first N-terminal α-helix of BURP-domain proteins were removed.

3.6.7 Preliminary QPFGVFRW Isolation

Transgenic tobacco leaves expressing LbaLycABURP-12xQPFGVFRW (1 kg) were frozen at −80 °C and ground in a food processor with dry ice and extracted in 8 L of methanol for 16 h at 37 °C with agitation (160 rpm). The methanol extract was filtered over silica and dried *in vacuo*. The dried extract was resuspended in 2 L of water and partitioned sequentially with 2 L of hexane (3x), ethyl acetate (2x), and n-butanol (2x). The n-butanol fraction was dried, resuspended in 10% methanol, and subjected to Sephadex LH20 chromatography (60 × 8 cm) using a methanol

gradient (20%, 40%, 60%, 80%, 100%; 500 mL each). Fractions were analyzed by LC-MS with the following parameters: injection volume, 2 µL; column, Phenomenex Kinetex® 2.6 µm C18 100 Å (50 × 3 mm); solvent A, 0.1% formic acid in water; solvent B, acetonitrile (0.1% formic acid); 0.5 mL/min, 30 °C; gradient: 0 min, 5% B; 2.5 min, 95% B; 3.0 min, 95% B; 3.1 min, 5% B; 5.0 min, 5% B. MS settings: positive ion mode; full MS resolution, 35,000; mass range, 400–1200 *m/z*; dd-MS^2 resolution, 17,500; collision energy, 25 eV; dynamic exclusion, 0.5 s.

3.6.8 Transient gene expression in Nicotiana benthamiana

LbaLycA-3xQPFGVFRW, LbaLycA-3xQPFGAAKW, LbaLycA-3xQPFGAWRW, LbaLycA-3xQPFGTRIW, LbaLycA-3xQPYGVIRW, LbaLycA-3xQPFGTRVW, LbaLycA-3xQPFGTKIW, LbaLycA-3xQPFGQRSW, LbaLycA-3xQPFGAWKW, LbaLycA-3xQPFGVRFW, LbaLycA-3xQPFGIRIW, and LbaLycA-3xQPFGALRW were synthesized with N-terminal overhang TGCCCAAATTCGCGACCGGT and C-terminal overhang CTCGAGGCCTTTAACTCTGG for Gibson cloning into pEAQ-HT linearized by restriction enzymes AgeI and XhoI. Agrobacterium tumefaciens LBA4404 was transformed with synthetic constructs pEAQ-HT-*LbaLycA* synthetic constructs (LbaLycA-3xQPFGVFRW, LbaLycA-3xQPFGAAKW, LbaLycA-3xQPFGAWRW, LbaLycA-3xQPFGTRIW, LbaLycA-3xQPYGVIRW, LbaLycA-3xQPFGTRVW, LbaLycA-3xQPFGTKIW, LbaLycA-3xQPFGQRSW, LbaLycA-3xQPFGAWKW, LbaLycA-3xQPFGVRFW, LbaLycA-3xQPFGIRIW, and LbaLycA-3xQPFGALRW) by electroporation (2.5 kV) and then plated on YM agar plates (0.4 g of yeast extract, 10 g of d-mannitol, 0.1 g of sodium chloride (NaCl), 0.2 g of magnesium sulfate (heptahydrate), 0.5 g of potassium phosphate (dibasic, trihydrate), 15 g of agar, and 1 L of deionized water, adjusted to pH 7) with 100 µg/mL rifampicin, 50 µg/mL kanamycin, and 100 µg/mL streptomycin, and incubated for 2 days at 30 °C. A 5 mL starter culture of the YM medium with 100 µg/mL rifampicin, 50 µg/mL kanamycin, and 100 µg/mL streptomycin was inoculated with a clone of A. tumefaciens LBA4404 pEAQ-HT-*LbaLycA* and incubated for 24–36 h at 30 °C in a shaking incubator at 220 rpm. Subsequently, the starter culture was used to inoculate a 25 mL culture of the YM medium with 100 µg/mL rifampicin, 50 µg/mL kanamycin, and 100 µg/mL streptomycin, which was incubated for 24 h at 30 °C in a shaking incubator at 220 rpm. The cells from the 25 mL culture were centrifuged for 30 min at 3000 × g, the YM medium was discarded, and cells were resuspended in MMA medium

(10 mM MES KOH buffer (pH 5.6), 10 mM magnesium chloride (MgCl2), 100 µM acetosyringone) to give a final optical density (OD600) of 0.8. The Agrobacterium suspension was infiltrated into the bottom of the leaves of N. benthamiana plants (6-8 weeks old) using 1 mL plastic syringes. N. benthamiana plants were placed in the shade 2 h before syringe infiltration. After infiltration, N. benthamiana plants were grown as described above for 6 days. Subsequently, infiltrated leaves were collected and subjected to peptide chemotyping. For peptide purification, infiltrated leaves were collected after 7 days and stored at -80 °C until purification.

3.6.9 Descriptor prediction Library prep and docking

The structure-based virtual screening workflow was executed within the Schrödinger Maestro (v2022-4) suite. The 375-member burpitide library was imported into Maestro via a CSV file containing core peptide sequences and their corresponding SMILES strings. Molecular descriptors were calculated using the Maestro Molecular Descriptors panel. LigPrep generated energy-minimized 3D conformations with desalting, tautomer generation, and retain chirality's using OPLS4 force field and at a target pH of 7.0 ± 2.0 using Epik. The X-ray crystal structure of LspA in complex with globomycin (PDB: 6RYO) was retrieved from the Protein Data Bank. The structure was refined using the Protein Preparation Wizard with the complex split into its constituent protein, ligand (globomycin), and structural water molecules. Pre-processing included: assigning bond orders using the CCD database, creating disulfide bonds, generating heteroatom states using Epik at pH 7.4, predicting pKa values for amino acid side chains using PROPKA at pH 7.0, removing water molecules beyond 5.0 Å, and restrained minimization focused on hydrogen bonds only. The receptor grid was defined using the Receptor Grid Generation panel with the grid center localized at the site centroid of the workspace ligand (globomycin). The docking box dimensions were set to 10 Å x 10 Å x 10 Å. To validate the accuracy of the receptor grid re-docking was performed using the cognate ligand, globomycin. The ligand was extracted from the 6RYO complex, prepared via the LigPrep workflow described above, and docked back into the generated receptor grid. The best-scoring pose of the re-docked globomycin was compared to the original crystallographic coordinates. The protocol achieved a Root Mean Square Deviation (RMSD) of 1.75 Å. Virtual screening was performed using Glide in Standard Precision (SP) mode with prime macrocyclic sampling and flexible docking enabled. The output was set to generate up

to 30 poses per ligand and a total inclusion of 100 poses for post-docking analysis. A two-stage filtering process was applied: a Glide score (Gscore) cutoff of -9.0 kcal/mol and manually inspection for direct contacts with the Asp-Asp catalytic dyad. Complexes that passed these criteria were subjected to Prime MM-GBSA using VSGB solvation model and OPLS4 force field with a 5.0 Å sampling distance. A final visual inspection of the minimized complexes ensured that interactions with the aspartyl dyad were maintained post-minimization.

3.6.10 PCA

Principal Component Analysis (PCA) was performed to visualize and compare the chemical space of databases. All statistical analyses were conducted using R Statistical Software (version 4.2.3). physicochemical descriptors (ALogP, molecular weight (MolW), number of rotatable bonds (Nrotb), chiral centers, hydrogen bond donors (HBD), hydrogen bond acceptors (HBA), and total polar surface area (TPSA)) were curated into structured .csv files. To account for the different units and scales of these descriptors the data were standardized using the prcomp function. Multivariate visualization and biplot generation were executed using the following specialized packages: ggplot2 (v3.4.2), ggbiplot (v0.55), Scales (v1.2.1), and Plyr (v1.8.8). The degree of variance captured by the principal components was assessed using scree plots generated within the base R graphics package, ensuring that the selected components (PC1 and PC2) accounted for a majority of the library's total variance. Additional PCA models were constructed with descriptors sourced from: CyBase, COCONUT Database, ChEMBL, and FDA/Clinical Trial Database.

3.7 Supplemental Information

Table 3.3 Predicted RiPP Library Sequences and Descriptors

RiPP Sequence	Chiral centers	HBA	HBD	Polar SA	NRotB	MW	AlogP	SMILE
QPAGQVYW	8	10	9	329.56	12	929.008	-0.898	O=C1N[C@H](C(N2[C@H](C(N[C@@H](C)C(N[C@H]3C(N[C@@H](CCC(N)=O)C(N[C@@H](C(C)C)C(N[C@@H](CC4=CC=C(O)C=C4)C(N[C@@H](C(O)=O)CC5=CN3C6=C5C=CC=C6)=O)=O)=O)=O)=O)CCC2)=O)CC1
QPATVIGW	8	8	7	266.24	9	807.912	-0.414	O=C1N[C@H](C(N2[C@H](C(N[C@@H](C)C(N[C@H]3C(N[C@@H](C(C)C)C(NC(C(C)CC)C(NCC(N[C@@H](C(O)=O)CC4=CN3C5=C4C=CC=C5)=O)=O)=O)=O)=O)CCC2)=O)CC1
QPAYNYLW	8	10	9	329.56	12	929.008	-0.83	O=C1N[C@H](C(N2[C@H](C(N[C@@H](C)C(N[C@H]3C(N[C@H](C(N[C@@H](CC4=CC=C(O)C=C4)C(N[C@@H](CC(C)C)C(N[C@@H](C(O)=O)CC5=CN3C6=C5C=CC=C6)=O)=O)=O)CC(N)=O)=O)=O)=O)CCC2)=O)CC1
QPDSERKW	8	9	8	428.76	20	996.055	-3.789	O=C1N[C@H](C(N2[C@H](C(N[C@@H](CC(O)=O)C(N[C@H]3C(N[C@H](C(N[C@@H](CCCNC(N)=N)C(N[C@@H](CCCCN)C(N[C@@H](C(O)=O)CC4=CN3C5=C4C=CC=C5)=O)=O)=O)CCC(O)=O)=O)=O)=O)CCC2)=O)CC1

QPFEARTW	9	10	9	348.37	14	943.038	-1.697	O=C1N[C@H](C(N2[C@H](C(N[C@@H](CC3=CC=CC=C3)C(N[C@H]4C(N[C@H](C(N[C@@H](CCCNC(N)=N)C(NC(C(C)O)C(N[C@@H](C(O)=O)CC5=CN4C6=C5C=CC=C6)=O)=O)=O)C)=O)=O)=O)CCC2)=O)CC1
QPFGAAKW	8	8	7	292.26	12	884.998	-0.596	O=C1N[C@H](C(N2[C@H](C(N[C@@H](CC3=CC=CC=C3)C(N[C@H]4C(N[C@H](C(N[C@@H](C)C(N[C@@H](CCCCN)C(N[C@@H](C(O)=O)CC5=CN4C6=C5C=CC=C6)=O)=O)C)=O)=O)=O)CCC2)=O)CC1
QPFGAGIW	8	8	7	266.24	10	855.956	0.298	O=C1N[C@H](C(N2[C@H](C(N[C@@H](CC3=CC=CC=C3)C(N[C@H]4C(N[C@H](C(NCC(NC(C(C)CC)C(N[C@@H](C(O)=O)CC5=CN4C6=C5C=CC=C6)=O)=O)C)=O)=O)=O)CCC2)=O)CC1
QPFGAKVW	8	8	7	292.26	13	913.052	0.247	O=C1N[C@H](C(N2[C@H](C(N[C@@H](CC3=CC=CC=C3)C(N[C@H]4C(N[C@H](C(N[C@@H](CCCCN)C(N[C@@H](C(C)C)C(N[C@@H](C(O)=O)CC5=CN4C6=C5C=CC=C6)=O)=O)=O)C)=O)=O)=O)CCC2)=O)CC1
QPFGALRW	8	9	8	328.14	15	955.093	0.046	O=C1N[C@H](C(N2[C@H](C(N[C@@H](CC3=CC=CC=C3)C(N[C@H]4C(N[C@H](C(N[C@@H](CC(C)C)C(N[C@@H](CCCNC(N)=N)C(N[C@@H](C(O)=O)CC5=CN4C6=C5C=CC=C6)=O)=O)=O)C)=O)=O)=O)CCC2)=O)CC1
QPFGALTW	9	9	8	286.47	11	900.01	0.204	O=C1N[C@H](C(N2[C@H](C(N[C@@H](CC3=CC=CC=C3)C(N[C@H]4C(N[C@H](C(N[C@@H](CC(C)C)C(NC(C(O)C)C(N[C@@H](C(O)=O)CC5=CN4C6=C5C=CC=C6)=O)=O)=O)C)=O)=O)=O)CCC2)=O)CC1
QPFGALVW	8	8	7	266.24	11	898.037	1.558	O=C1N[C@H](C(N2[C@H](C(N[C@@H](CC3=CC=CC=C3)C(N[C@H]4C(N[C@H](C(N[C@@H](CC(C)C)C(NC(C(C)C)C(N[C@@H](C(O)=O)CC5=CN4C6=C5C=CC=C6)=O)=O)=O)C)=O)=O)=O)CCC2)=O)CC1
QPFGARFW	8	9	8	328.14	15	989.11	0.37	O=C1N[C@H](C(N2[C@H](C(N[C@@H](CC3=CC=CC=C3)C(N[C@H]4C(N[C@H](C(N[C@@H](CCCNC(N)=N)C(N[C@@H](CC5=CC=CC=C5)C(N[C@@H](C(O)=O)CC6=CN4C7=C6C=CC=C7)=O)=O)=O)C)=O)=O)=O)CCC2)=O)CC1
QPFGARIW	9	9	8	328.14	15	955.093	0.113	O=C1N[C@H](C(N2[C@H](C(N[C@@H](CC3=CC=CC=C3)C(N[C@H]4C(N[C@H](C(N[C@@H](CCCNC(N)=N)C(NC(C(CC)C)C(N[C@@H](C(O)=O)CC5=CN4C6=C5C=CC=C6)=O)=O)=O)C)=O)=O)=O)CCC2)=O)CC1
QPFGARLW	8	9	8	328.14	15	955.093	0.046	O=C1N[C@H](C(N2[C@H](C(N[C@@H](CC3=CC=CC=C3)C(N[C@H]4C(N[C@H](C(N[C@@H](CCCNC(N)=N)C(N[C@@H](CC(C)C)C(N[C@@H](C(O)=O)CC5=CN4C6=C5C=CC=C6)=O)=O)=O)C)=O)=O)=O)CCC2)=O)CC1
QPFGARTW	9	10	9	348.37	14	943.038	-1.697	O=C1N[C@H](C(N2[C@H](C(N[C@@H](CC3=CC=CC=C3)C(N[C@H]4C(N[C@H](C(N[C@@H](CCCNC(N)=N)C(NC(C(C)O)C(N[C@@H](C(O)=O)CC5=CN4C6=C5C=CC=C6)=O)=O)=O)C)=O)=O)=O)CCC2)=O)CC1
QPFGARVW	8	9	8	328.14	14	941.066	-0.343	O=C1N[C@H](C(N2[C@H](C(N[C@@H](CC3=CC=CC=C3)C(N[C@H]4C(N[C@H](C(N[C@@H](CCCNC(N)=N)C(N[C@@H](C(C)C)C(N[C@@H](C(O)=O)CC5=CN4C6=C5C=CC=C6)=O)=O)=O)C)=O)=O)=O)CCC2)=O)CC1
QPFGAWKW	8	8	8	308.05	14	1000.134	0.738	O=C1N[C@H](C(N2[C@H](C(N[C@@H](CC3=CC=CC=C3)C(N[C@H]4C(N[C@H](C(N[C@@H](CC5=CNC6=C5C=CC=C6)C(N[C@@H](CCCCN)C(N[C@@H](C(O)=O)CC7=CN4C8=C7C=CC=C8)=O)=O)C)=O)=O)=O)CCC2)=O)CC1
QPFGAWRW	8	9	9	343.93	14	1014.12	-0.432	O=C1N[C@H](C(N2[C@H](C(N[C@@H](CC3=CC=CC=C3)C(N[C@H]4C(N[C@H](C(N[C@@H](CC5=CNC6=C5C=CC=C6)C(N[C@@H](CCNC(N)=N)C(N[C@@H](C(O)=O)CC7=CN4C8=C7C=CC=C8)=O)=O)=O)C)=O)=O)=O)CCC2)=O)CC1
QPFGAYYW	8	10	9	306.7	12	1012.098	2.06	O=C1N[C@H](C(N2[C@H](C(N[C@@H](CC3=CC=CC=C3)C(N[C@H]4C(N[C@H](C(N[C@@H](CC5=CC=C(O)C=C5)C(NC(CC6=CC=C(O)C=C6)C(N[C@@H](C(O)=O)CC7=CN4C8=C7C=CC=C8)=O)=O)C)=O)=O)=O)CCC2)=O)CC1
QPFGDYVW	8	9	8	323.77	13	992.064	0.97	O=C1N[C@H](C(N2[C@H](C(N[C@@H](CC3=CC=CC=C3)C(N[C@H]4C(N[C@H](C(N[C@@H](CC5=CC=C(O)C=C5)C(N[C@@H](C(C)C)C(N[C@@H](C(O)=O)CC6=CN4C7=C6C=CC=C7)=O)=O)CC(O)=O)=O)=O)=O)CCC2)=O)CC1
QPFGFLIW	9	8	7	266.24	14	988.163	3.57	O=C1N[C@H](C(N2[C@H](C(N[C@@H](CC3=CC=CC=C3)C(N[C@H]4C(N[C@H](C(N[C@@H](CC(C)C)C(NC(C(C)CC)C(N[C@@H](C(O)=O)CC5=CN4C6=C5C=CC=C6)=O)=O)=O)CC7=CC=CC=C7)=O)=O)=O)CCC2)=O)CC1
QPFGGFAW	7	8	7	266.24	10	889.974	0.555	O=C1N[C@H](C(N2[C@H](C(N[C@@H](CC3=CC=CC=C3)C(N[C@H]4C(NCC(N[C@@H](CC5=CC=CC=C5)C(N[C@@H](C(O)=O)CC6=CN4C7=C6C=CC=C7)=O)=O)=O)=O)=O)CCC2)=O)CC1
QPFGINIW	10	9	8	309.33	14	955.09	0.804	O=C1N[C@H](C(N2[C@H](C(N[C@@H](CC3=CC=CC=C3)C(N[C@H]4C(NC(C(N[C@@H](CC(N)=O)C(NC(C(C)CC)C(N[C@@H](C(O)=O)CC5=CN4C6=C5C=CC=C6)=O)=O)=O)C(CC)C)=O)=O)=O)CCC2)=O)CC1
QPFGINSW	9	10	9	329.56	13	929.008	-1.384	O=C1N[C@H](C(N2[C@H](C(N[C@@H](CC3=CC=CC=C3)C(N[C@H]4C(NC(C(N[C@@H](CC(N)=O)C(NC(CO)C(N[C@@H](C(O)=O)CC5=CN4C6=C5C=CC=C6)=O)=O)=O)C(CC)C)=O)=O)=O)CCC2)=O)CC1
QPFGINTW	10	10	9	329.56	13	943.035	-1.006	O=C1N[C@H](C(N2[C@H](C(N[C@@H](CC3=CC=CC=C3)C(N[C@H]4C(NC(C(N[C@@H](CC(N)=O)C(NC(C(C)O)C(N[C@@H](C(O)=O)CC5=CN4C6=C5C=CC=C6)=O)=O)=O)C(CC)C)=O)=O)=O)CCC2)=O)CC1
QPFGINVW	9	9	8	309.33	13	941.063	0.347	O=C1N[C@H](C(N2[C@H](C(N[C@@H](CC3=CC=CC=C3)C(N[C@H]4C(NC(C(N[C@@H](CC(N)=O)C(NC(C(C)C)C(N[C@@H](C(O)=O)CC5=CN4C6=C5C=CC=C6)=O)=O)=O)C(CC)C)=O)=O)=O)CCC2)=O)CC1
QPFGIRIW	10	9	8	328.14	17	997.174	1.412	O=C1N[C@H](C(N2[C@H](C(N[C@@H](CC3=CC=CC=C3)C(N[C@H]4C(NC(C(N[C@@H](CCCNC(N)=N)C(NC(C(C)CC)C(N[C@@H](C(O)=O)CC5=CN4C6=C5C=CC=C6)=O)=O)=O)C(CC)C)=O)=O)=O)CCC2)=O)CC1
QPFGIRTW	10	10	9	348.37	16	985.119	-0.398	O=C1N[C@H](C(N2[C@H](C(N[C@@H](CC3=CC=CC=C3)C(N[C@H]4C(NC(C(N[C@@H](CCCNC(N)=N)C(NC(C(O)C)C(N[C@@H](C(O)=O)CC5=CN4C6=C5C=CC=C6)=O)=O)=O)C(CC)C)=O)=O)=O)CCC2)=O)CC1
QPFGIRVW	9	9	8	328.14	16	983.147	0.956	O=C1N[C@H](C(N2[C@H](C(N[C@@H](CC3=CC=CC=C3)C(N[C@H]4C(NC(C(N[C@@H](CCCNC(N)=N)C(N[C@@H](C(C)C)C(N[C@@H](C(O)=O)CC5=CN4C6=C5C=CC=C6)=O)=O)=O)C(CC)C)=O)=O)=O)CCC2)=O)CC1
QPFGLRFW	8	9	8	328.14	17	1031.191	1.602	O=C1N[C@H](C(N2[C@H](C(N[C@@H](CC3=CC=CC=C3)C(N[C@H]4C(N[C@H](C(N[C@@H](CCCNC(N)=N)C(N[C@@H](CC5=CC=CC=C5)C(N[C@@H](C(O)=O)CC6=CN4C7=C6C=CC=C7)=O)=O)=O)CC(C)C)=O)=O)=O)CCC2)=O)CC1
QPFGLRTW	9	10	9	348.37	16	985.119	-0.465	O=C1N[C@H](C(N2[C@H](C(N[C@@H](CC3=CC=CC=C3)C(N[C@H]4C(N[C@H](C(N[C@@H](CCCNC(N)=N)C(NC(C(O)C)C(N[C@@H](C(O)=O)CC5=CN4C6=C5C=CC=C6)=O)=O)=O)CC(C)C)=O)=O)=O)CCC2)=O)CC1
QPFGLTIW	10	9	8	286.47	13	942.091	1.504	O=C1N[C@H](C(N2[C@H](C(N[C@@H](CC3=CC=CC=C3)C(N[C@H]4C(N[C@H](C(NC(C(O)C)C(NC(C(C)CC)C(N[C@@H](C(O)=O)CC5=CN4C6=C5C=CC=C6)=O)=O)=O)CC(C)C)=O)=O)=O)CCC2)=O)CC1
QPFGLTVW	9	9	8	286.47	12	928.064	1.047	O=C1N[C@H](C(N2[C@H](C(N[C@@H](CC3=CC=CC=C3)C(N[C@H]4C(N[C@H](C(NC(C(O)C)C(N[C@@H](C(C)C)C(N[C@@H](C(O)=O)CC5=CN4C6=C5C=CC=C6)=O)=O)=O)CC(C)C)=O)=O)=O)CCC2)=O)CC1
QPFGLVAW	8	8	7	266.24	11	898.037	1.558	O=C1N[C@H](C(N2[C@H](C(N[C@@H](CC3=CC=CC=C3)C(N[C@H]4C(N[C@H](C(N[C@@H](C(C)C)C(N[C@@H](C)C(N[C@@H](C(O)=O)CC5=CN4C6=C5C=CC=C6)=O)=O)=O)CC(C)C)=O)=O)=O)CCC2)=O)CC1
QPFGMRTW	9	11	9	373.67	17	1003.156	-1.369	O=C1N[C@H](C(N2[C@H](C(N[C@@H](CC3=CC=CC=C3)C(N[C@H]4C(N[C@H](C(N[C@@H](CCCNC(N)=N)C(NC(C(C)O)C(N[C@@H](C(O)=O)CC5=CN4C6=C5C=CC=C6)=O)=O)=O)CCSC)=O)=O)=O)CCC2)=O)CC1
QPFGMTVW	9	10	8	311.77	13	946.101	0.143	O=C1N[C@H](C(N2[C@H](C(N[C@@H](CC3=CC=CC=C3)C(N[C@H]4C(N[C@H](C(NC(C(C)O)C(N[C@@H](C(C)C)C(N[C@@H](C(O)=O)CC5=CN4C6=C5C=CC=C6)=O)=O)=O)CCSC)=O)=O)=O)CCC2)=O)CC1
QPFGQRSW	8	11	10	391.46	17	986.063	-3.031	O=C1N[C@H](C(N2[C@H](C(N[C@@H](CC3=CC=CC=C3)C(N[C@H]4C(N[C@H](C(NC(CCCNC(N)=N)C(N[C@@H](CO)C(N[C@@H](C(O)=O)CC5=CN4C6=C5C=CC=C6)=O)=O)=O)CCC(N)=O)=O)=O)=O)CCC2)=O)CC1
QPFGQRTW	9	11	10	391.46	17	1000.09	-2.654	O=C1N[C@H](C(N2[C@H](C(N[C@@H](CC3=CC=CC=C3)C(N[C@H]4C(N[C@H](C(NC(CCCNC(N)=N)C(NC(C(C)O)C(N[C@@H](C(O)=O)CC5=CN4C6=C5C=CC=C6)=O)=O)=O)CCC(N)=O)=O)=O)=O)CCC2)=O)CC1
QPFGTKIW	10	9	8	312.49	15	957.106	0.193	O=C1N[C@H](C(N2[C@H](C(N[C@@H](CC3=CC=CC=C3)C(N[C@H]4C(NC(C(NC(CCCCN)C(NC(C(CC)C)C(N[C@@H](C(O)=O)CC5=CN4C6=C5C=CC=C6)=O)=O)=O)C(O)C)=O)=O)=O)CCC2)=O)CC1
QPFGTLLW	9	9	8	286.47	13	942.091	1.436	O=C1N[C@H](C(N2[C@H](C(N[C@@H](CC3=CC=CC=C3)C(N[C@H]4C(NC(C(N[C@@H](CC(C)C)C(N[C@@H](CC(C)C)C(N[C@@H](C(O)=O)CC5=CN4C6=C5C=CC=C6)=O)=O)=O)C(O)C)=O)=O)=O)CCC2)=O)CC1
QPFGTRAW	9	10	9	348.37	14	943.038	-1.697	O=C1N[C@H](C(N2[C@H](C(N[C@@H](CC3=CC=CC=C3)C(N[C@H]4C(NC(C(N[C@@H](CCCNC(N)=N)C(N[C@@H](C)C(N[C@@H](C(O)=O)CC5=CN4C6=C5C=CC=C6)=O)=O)=O)C(O)C)=O)=O)=O)CCC2)=O)CC1
QPFGTRFW	9	10	9	348.37	16	1019.137	-0.141	O=C1N[C@H](C(N2[C@H](C(N[C@@H](CC3=CC=CC=C3)C(N[C@H]4C(NC(C(N[C@@H](CCCNC(N)=N)C(N[C@@H](CC5=CC=CC=C5)C(N[C@@H](C(O)=O)CC6=CN4C7=C6C=CC=C7)=O)=O)C)(O)C)=O)=O)=O)CCC2)=O)CC1

QPFGTRIW	10	10	9	348.37	16	985.119	-0.398	O=C1N[C@H](C(N2[C@H](C(N[C@@H](CC3=CC=CC=C3)C(N[C@H]4C(NC(C(N[C@@H](CCCNC(N)=N)C(NC(C(C)CC)C(N[C@@H](C(O)=O)CC5=CN4C6=C5C=CC=C6)=O)=O)=O)C(O)C)=O)=O)=O)CCC2)=O)CC1
QPFGTRTW	10	11	10	368.6	15	973.064	-2.207	O=C1N[C@H](C(N2[C@H](C(N[C@@H](CC3=CC=CC=C3)C(N[C@H]4C(NC(C(N[C@@H](CCCNC(N)=N)C(NC(C(O)C)C(N[C@@H](C(O)=O)CC5=CN4C6=C5C=CC=C6)=O)=O)=O)C(O)C)=O)=O)=O)CCC2)=O)CC1
QPFGTRVW	9	10	9	348.37	15	971.092	-0.854	O=C1N[C@H](C(N2[C@H](C(N[C@@H](CC3=CC=CC=C3)C(N[C@H]4C(NC(C(N[C@@H](CCCNC(N)=N)C(NC(C(C)C)C(N[C@@H](C(O)=O)CC5=CN4C6=C5C=CC=C6)=O)=O)=O)C(O)C)=O)=O)=O)CCC2)=O)CC1
QPFGVDIW	9	8	7	303.54	13	942.047	0.981	O=C1N[C@H](C(N2[C@H](C(N[C@@H](CC3=CC=CC=C3)C(N[C@H]4C(N[C@H](C(N[C@@H](CC(O)=O)C(NC(C(C)CC)C(N[C@@H](C(O)=O)CC5=CN4C6=C5C=CC=C6)=O)=O)C(C)C)=O)=O)=O)CCC2)=O)CC1
QPFGVFAW	8	8	7	266.24	11	932.055	1.882	O=C1N[C@H](C(N2[C@H](C(N[C@@H](CC3=CC=CC=C3)C(N[C@H]4C(N[C@H](C(N[C@@H](CC5=CC=CC=C5)C(N[C@@H](C)C(N[C@@H](C(O)=O)CC6=CN4C7=C6C=CC=C7)=O)=O)=O)C(C)C)=O)=O)=O)CCC2)=O)CC1
QPFGVFGW	7	8	7	266.24	11	918.028	1.398	O=C1N[C@H](C(N2[C@H](C(N[C@@H](CC3=CC=CC=C3)C(N[C@H]4C(N[C@H](C(N[C@@H](CC5=CC=CC=C5)C(NCC(N[C@@H](C(O)=O)CC6=CN4C7=C6C=CC=C7)=O)=O)=O)C(C)C)=O)=O)=O)CCC2)=O)CC1
QPFGVFRW	8	9	8	328.14	16	1017.164	1.213	O=C1N[C@H](C(N2[C@H](C(N[C@@H](CC3=CC=CC=C3)C(N[C@H]4C(N[C@H](C(N[C@@H](CC5=CC=CC=C5)C(NC(CCCNC(N)=N)C(N[C@@H](C(O)=O)CC6=CN4C7=C6C=CC=C7)=O)=O)=O)C(C)C)=O)=O)=O)CCC2)=O)CC1
QPFGVFSW	8	9	8	286.47	12	948.054	0.994	O=C1N[C@H](C(N2[C@H](C(N[C@@H](CC3=CC=CC=C3)C(N[C@H]4C(N[C@H](C(N[C@@H](CC5=CC=CC=C5)C(N[C@@H](CO)C(N[C@@H](C(O)=O)CC6=CN4C7=C6C=CC=C7)=O)=O)=O)C(C)C)=O)=O)=O)CCC2)=O)CC1
QPFGVFVW	8	8	7	266.24	12	960.109	2.725	O=C1N[C@H](C(N2[C@H](C(N[C@@H](CC3=CC=CC=C3)C(N[C@H]4C(N[C@H](C(N[C@@H](CC5=CC=CC=C5)C(N[C@@H](C(C)C)C(N[C@@H](C(O)=O)CC6=CN4C7=C6C=CC=C7)=O)=O)=O)C(C)C)=O)=O)=O)CCC2)=O)CC1
QPFGVGVW	7	8	7	266.24	10	869.983	0.685	O=C1N[C@H](C(N2[C@H](C(N[C@@H](CC3=CC=CC=C3)C(N[C@H]4C(N[C@H](C(NCC(N[C@@H](C(C)C)C(N[C@@H](C(O)=O)CC5=CN4C6=C5C=CC=C6)=O)=O)=O)C(C)C)=O)=O)=O)CCC2)=O)CC1
QPFGVRFW	8	9	8	328.14	16	1017.164	1.213	O=C1N[C@H](C(N2[C@H](C(N[C@@H](CC3=CC=CC=C3)C(N[C@H]4C(N[C@H](C(N[C@@H](CCCNC(N)=N)C(N[C@@H](CC5=CC=CC=C5)C(N[C@@H](C(O)=O)CC6=CN4C7=C6C=CC=C7)=O)=O)=O)C(C)C)=O)=O)=O)CCC2)=O)CC1
QPFGVWRW	8	9	9	343.93	16	1056.201	0.991	O=C1N[C@H](C(N2[C@H](C(N[C@@H](CC3=CC=CC=C3)C(N[C@H]4C(N[C@H](C(N[C@@H](CC5=CNC6=C5C=CC=C6)C(N[C@@H](CCCNC(N)=N)C(N[C@@H](C(O)=O)CC7=CN4C8=C7C=CC=C8)=O)=O)=O)C(C)C)=O)=O)=O)CCC2)=O)CC1
QPFGVYAW	8	9	8	286.47	11	948.054	1.615	O=C1N[C@H](C(N2[C@H](C(N[C@@H](CC3=CC=CC=C3)C(N[C@H]4C(N[C@H](C(N[C@@H](CC5=CC=C(O)C=C5)C(N[C@@H](C)C(N[C@@H](C(O)=O)CC6=CN4C7=C6C=CC=C7)=O)=O)=O)C(C)C)=O)=O)=O)CCC2)=O)CC1
QPFGYRVW	8	10	9	348.37	16	1033.164	0.945	O=C1N[C@H](C(N2[C@H](C(N[C@@H](CC3=CC=CC=C3)C(N[C@H]4C(N[C@H](C(N[C@@H](CCCNC(N)=N)C(N[C@@H](C(C)C)C(N[C@@H](C(O)=O)CC5=CN4C6=C5C=CC=C6)=O)=O)=O)CC7=CC=C(O)C=C7)=O)=O)=O)CCC2)=O)CC1
QPFKTRKW	9	10	9	374.39	18	1000.134	-1.776	O=C1N[C@H](C(N2[C@H](C(N[C@@H](CC3=CC=CC=C3)C(N[C@H]4C(NC(C(N[C@@H](CCCNC(N)=N)C(N[C@@H](CCCCN)C(N[C@@H](C(O)=O)CC5=CN4C6=C5C=CC=C6)=O)=O)=O)C(O)C)=O)=O)=O)CCC2)=O)CC1
QPFLPIFW	9	8	6	257.45	12	972.12	2.657	O=C1N[C@H](C(N2[C@H](C(N[C@@H](CC3=CC=CC=C3)C(N[C@H]4C(N5[C@@H](CCC5)C(NC(C(CC)C)C(N[C@@H](CC6=CC=CC=C6)C(N[C@@H](C(O)=O)CC7=CN4C8=C7C=CC=C8)=O)=O)=O)=O)=O)CCC2)=O)CC1
QPFRMRIW	9	10	8	353.44	18	1015.211	0.441	O=C1N[C@H](C(N2[C@H](C(N[C@@H](CC3=CC=CC=C3)C(N[C@H]4C(N[C@H](C(N[C@@H](CCCNC(N)=N)C(NC(C(C)CC)C(N[C@@H](C(O)=O)CC5=CN4C6=C5C=CC=C6)=O)=O)=O)CCSC)=O)=O)=O)CCC2)=O)CC1
QPFRSIGW	8	9	8	286.47	11	871.956	-0.59	O=C1N[C@H](C(N2[C@H](C(N[C@@H](CC3=CC=CC=C3)C(N[C@H]4C(N[C@H](C(NC(C(C)CC)C(NCC(N[C@@H](C(O)=O)CC5=CN4C6=C5C=CC=C6)=O)=O)=O)CO)=O)=O)=O)CCC2)=O)CC1
QPFTITAW	10	9	8	286.47	11	900.01	0.272	O=C1N[C@H](C(N2[C@H](C(N[C@@H](CC3=CC=CC=C3)C(N[C@H]4C(NC(C(NC(C(C)O)C(N[C@@H](C)C(N[C@@H](C(O)=O)CC5=CN4C6=C5C=CC=C6)=O)=O)=O)C(CC)C)=O)=O)=O)CCC2)=O)CC1
QPFTVFAW	8	8	7	266.24	11	932.055	1.882	O=C1N[C@H](C(N2[C@H](C(N[C@@H](CC3=CC=CC=C3)C(N[C@H]4C(N[C@H](C(N[C@@H](CC5=CC=CC=C5)C(N[C@@H](C)C(N[C@@H](C(O)=O)CC6=CN4C7=C6C=CC=C7)=O)=O)=O)C(C)C)=O)=O)=O)CCC2)=O)CC1
QPFTVFGW	7	8	7	266.24	11	918.028	1.398	O=C1N[C@H](C(N2[C@H](C(N[C@@H](CC3=CC=CC=C3)C(N[C@H]4C(N[C@H](C(N[C@@H](CC5=CC=CC=C5)C(NCC(N[C@@H](C(O)=O)CC6=CN4C7=C6C=CC=C7)=O)=O)=O)C(C)C)=O)=O)=O)CCC2)=O)CC1
QPFTVFSW	8	9	8	286.47	12	948.054	0.994	O=C1N[C@H](C(N2[C@H](C(N[C@@H](CC3=CC=CC=C3)C(N[C@H]4C(N[C@H](C(N[C@@H](CC5=CC=CC=C5)C(N[C@@H](CO)C(N[C@@H](C(O)=O)CC6=CN4C7=C6C=CC=C7)=O)=O)=O)C(C)C)=O)=O)=O)CCC2)=O)CC1
QPFTVGAW	7	8	7	266.24	9	841.929	-0.158	O=C1N[C@H](C(N2[C@H](C(N[C@@H](CC3=CC=CC=C3)C(N[C@H]4C(N[C@H](C(NCC(N[C@@H](C)C(N[C@@H](C(O)=O)CC5=CN4C6=C5C=CC=C6)=O)=O)=O)C(C)C)=O)=O)=O)CCC2)=O)CC1
QPFTVGSW	7	9	8	286.47	10	857.928	-1.046	O=C1N[C@H](C(N2[C@H](C(N[C@@H](CC3=CC=CC=C3)C(N[C@H]4C(N[C@H](C(NCC(N[C@@H](CO)C(N[C@@H](C(O)=O)CC5=CN4C6=C5C=CC=C6)=O)=O)=O)C(C)C)=O)=O)=O)CCC2)=O)CC1
QPFTVIGW	8	8	7	266.24	11	884.01	1.141	O=C1N[C@H](C(N2[C@H](C(N[C@@H](CC3=CC=CC=C3)C(N[C@H]4C(N[C@H](C(NC(C(CC)C)C(NCC(N[C@@H](C(O)=O)CC5=CN4C6=C5C=CC=C6)=O)=O)=O)C(C)C)=O)=O)=O)CCC2)=O)CC1
QPFTVLAW	8	8	7	266.24	11	898.037	1.558	O=C1N[C@H](C(N2[C@H](C(N[C@@H](CC3=CC=CC=C3)C(N[C@H]4C(N[C@H](C(N[C@@H](CC(C)C)C(N[C@@H](C)C(N[C@@H](C(O)=O)CC5=CN4C6=C5C=CC=C6)=O)=O)=O)C(C)C)=O)=O)=O)CCC2)=O)CC1
QPFTVLGW	7	8	7	266.24	11	884.01	1.074	O=C1N[C@H](C(N2[C@H](C(N[C@@H](CC3=CC=CC=C3)C(N[C@H]4C(N[C@H](C(N[C@@H](CC(C)C)C(NCC(N[C@@H](C(O)=O)CC5=CN4C6=C5C=CC=C6)=O)=O)=O)C(C)C)=O)=O)=O)CCC2)=O)CC1
QPFTVLRW	8	9	8	328.14	16	983.147	0.889	O=C1N[C@H](C(N2[C@H](C(N[C@@H](CC3=CC=CC=C3)C(N[C@H]4C(N[C@H](C(N[C@@H](CC(C)C)C(N[C@@H](CCCNC(N)=N)C(N[C@@H](C(O)=O)CC5=CN4C6=C5C=CC=C6)=O)=O)=O)C(C)C)=O)=O)=O)CCC2)=O)CC1
QPFTVSAW	8	9	8	286.47	10	871.956	-0.562	O=C1N[C@H](C(N2[C@H](C(N[C@@H](CC3=CC=CC=C3)C(N[C@H]4C(N[C@H](C(N[C@@H](CO)C(N[C@@H](C)C(N[C@@H](C(O)=O)CC5=CN4C6=C5C=CC=C6)=O)=O)=O)C(C)C)=O)=O)=O)CCC2)=O)CC1
QPFTVTAW	9	9	8	286.47	10	885.983	-0.184	O=C1N[C@H](C(N2[C@H](C(N[C@@H](CC3=CC=CC=C3)C(N[C@H]4C(N[C@H](C(NC(C(C)O)C(N[C@@H](C)C(N[C@@H](C(O)=O)CC5=CN4C6=C5C=CC=C6)=O)=O)=O)C(C)C)=O)=O)=O)CCC2)=O)CC1
QPFTVVGW	7	8	7	266.24	10	869.983	0.685	O=C1N[C@H](C(N2[C@H](C(N[C@@H](CC3=CC=CC=C3)C(N[C@H]4C(N[C@H](C(N[C@@H](C(C)C)C(NCC(N[C@@H](C(O)=O)CC5=CN4C6=C5C=CC=C6)=O)=O)=O)C(C)C)=O)=O)=O)CCC2)=O)CC1
QPFTVYAW	8	9	8	286.47	11	948.054	1.615	O=C1N[C@H](C(N2[C@H](C(N[C@@H](CC3=CC=CC=C3)C(N[C@H]4C(N[C@H](C(N[C@@H](CC5=CC=C(O)C=C5)C(N[C@@H](C)C(N[C@@H](C(O)=O)CC6=CN4C7=C6C=CC=C7)=O)=O)=O)C(C)C)=O)=O)=O)CCC2)=O)CC1
QPFTVYGW	7	9	8	286.47	11	934.027	1.131	O=C1N[C@H](C(N2[C@H](C(N[C@@H](CC3=CC=CC=C3)C(N[C@H]4C(N[C@H](C(N[C@@H](CC5=CC=C(O)C=C5)C(NCC(N[C@@H](C(O)=O)CC6=CN4C7=C6C=CC=C7)=O)=O)=O)C(C)C)=O)=O)=O)CCC2)=O)CC1
QPFTVYSW	8	10	9	306.7	12	964.054	0.727	O=C1N[C@H](C(N2[C@H](C(N[C@@H](CC3=CC=CC=C3)C(N[C@H]4C(N[C@H](C(N[C@@H](CC5=CC=C(O)C=C5)C(N[C@@H](C)C(N[C@@H](C(O)=O)CC6=CN4C7=C6C=CC=C7)=O)=O)=O)C(C)C)=O)=O)=O)CCC2)=O)CC1
QPGRKSTW	8	10	9	332.72	12	840.898	-4.035	O=C1N[C@H](C(N2[C@H](C(NCC(N[C@@H]3C(N[C@H](C(N[C@@H](CO)C(NC(C(C)O)C(N[C@@H](C(O)=O)CC4=CN3C5=C4C=CC=C5)=O)=O)=O)CCCCN)=O)=O)=O)CCC2)=O)CC1
QPIILPIW	10	8	6	257.45	12	904.085	2.076	O=C1N[C@H](C(N2[C@H](C(NC(C(C)CC)C(N[C@H]3C(N[C@H](C(N4[C@@H](CCC4)C(NC(C(C)CC)C(N[C@@H](C(O)=O)CC5=CN3C6=C5C=CC=C6)=O)=O)=O)CC(C)C)=O)=O)=O)CCC2)=O)CC1
QPLGGRTW	8	10	9	348.37	14	894.993	-2.505	O=C1N[C@H](C(N2[C@H](C(N[C@@H](CC(C)C)C(N[C@H]3C(NCC(N[C@@H](CCCNC(N)=N)C(NC(C(C)O)C(N[C@@H](C(O)=O)CC4=CN3C5=C4C=CC=C5)=O)=O)=O)=O)=O)=O)CCC2)=O)CC1
QPLGTWIW	10	9	9	302.26	13	981.128	1.423	O=C1N[C@H](C(N2[C@H](C(N[C@@H](CC(C)C)C(N[C@H]3C(NC(C(N[C@@H](CC4=CC(C=CC=C5)=C5N4)C(NC(C(C)CC)C(N[C@@H](C(O)=O)CC6=CN3C7=C6C=CC=C7)=O)=O)=O)C(C)O)=O)=O)=O)CCC2)=O)CC1

QPLGVRIW	9	9	8	328.14	16	949.129	0.632	O=C1N[C@H](C(N2[C@H](C(N[C@@H](CC(C)C)C(N[C@H]3C(N[C@H](C(N[C@@H](CCCNC(N)=N)C(NC(C(C)CC)C(N[C@@H](C(O)=O)CC4=CN3C5=C4C=CC=C5)=O)=O)=O)C(C)C)=O)=O)=O)CCC2)=O)CC1
QPLQSTTW	10	11	10	326.93	11	869.937	-2.75	O=C1N[C@H](C(N2[C@H](C(N[C@@H](CC(C)C)C(N[C@H]3C(N[C@H](C(NC(C(O)C)C(NC(C(C)O)C(N[C@@H](C(O)=O)CC4=CN3C5=C4C=CC=C5)=O)=O)=O)CO)=O)=O)=O)CCC2)=O)CC1
QPLRARKW	8	9	8	354.16	17	936.09	-1.589	O=C1N[C@H](C(N2[C@H](C(N[C@@H](CC(C)C)C(N[C@H]3C(N[C@H](C(N[C@@H](CCCNC(N)=N)C(N[C@@H](CCCCN)C(N[C@@H](C(O)=O)CC4=CN3C5=C4C=CC=C5)=O)=O)=O)C)=O)=O)=O)CCC2)=O)CC1
QPLRSTTW	10	11	10	326.93	11	869.937	-2.75	O=C1N[C@H](C(N2[C@H](C(N[C@@H](CC(C)C)C(N[C@H]3C(N[C@H](C(N[C@@H]([C@@H](O)C)C(N[C@@H]([C@H](C)O)C(N[C@@H](C(O)=O)CC4=CN3C5=C4C=CC=C5)=O)=O)=O)CO)=O)=O)=O)CCC2)=O)CC1
QPMDNQLW	8	11	9	377.72	16	954.08	-2.748	O=C1N[C@H](C(N2[C@H](C(N[C@@H](CCSC)C(N[C@H]3C(N[C@H](C(N[C@@H](CCC(N)=O)C(N[C@@H](CC(C)C)C(N[C@@H](C(O)=O)CC4=CN3C5=C4C=CC=C5)=O)=O)=O)CC(N)=O)=O)=O)=O)CCC2)=O)CC1
QPRPFTYW	9	11	10	368.6	16	1035.136	-0.408	O=C1N[C@H](C(N2[C@H](C(N[C@@H](CCCNC(N)=N)C(N[C@H]3C(N[C@H](C(NC(C(O)C)C(N[C@@H](CC4=CC=C(O)C=C4)C(N[C@@H](C(O)=O)CC5=CN3C6=C5C=CC=C6)=O)=O)=O)CC7=CC=CC=C7)=O)=O)=O)CCC2)=O)CC1
QPSDDGAW	7	9	8	323.77	9	797.785	-4.089	O=C1N[C@H](C(N2[C@H](C(N[C@@H](CO)C(N[C@H]3C(N[C@H](C(NCC(N[C@@H](C)C(N[C@@H](C(O)=O)CC4=CN3C5=C4C=CC=C5)=O)=O)CC(O)=O)=O)=O)=O)CCC2)=O)CC1
QPSFFHRW	8	11	10	377.05	17	1009.101	-2.503	O=C1N[C@H](C(N2[C@H](C(N[C@@H](CO)C(N[C@H]3C(N[C@H](C(N[C@@H](CC4=CNC=N4)C(N[C@@H](CCCCNC(N)=N)C(N[C@@H](C(O)=O)CC5=CN3C6=C5C=CC=C6)=O)=O)=O)CC7=CC=CC=C7)=O)=O)=O)CCC2)=O)CC1
QPSGLAYW	8	10	9	306.7	11	901.982	-0.441	O=C1N[C@H](C(N2[C@H](C(N[C@@H](CO)C(N[C@H]3C(N[C@H](C(N[C@@H](C)C(N[C@@H](CC4=CC=C(O)C=C4)C(N[C@@H](C(O)=O)CC5=CN3C6=C5C=CC=C6)=O)=O)=O)CC(C)C)=O)=O)=O)CCC2)=O)CC1
QPSPEVYW	8	10	9	344	13	945.992	-1.152	O=C1N[C@H](C(N2[C@H](C(N[C@@H](CO)C(N[C@H]3C(N[C@H](C(N[C@@H](C(C)C)C(N[C@@H](CC4=CC=C(O)C=C4)C(N[C@@H](C(O)=O)CC5=CN3C6=C5C=CC=C6)=O)=O)=O)CCC(O)=O)=O)=O)=O)CCC2)=O)CC1
QPTLAACW	9	9	9	325.27	8	812.886	-2.5	O=C1N[C@H](C(N2[C@H](C(NC(C(C)O)C(N[C@H]3C(N[C@H](C(N[C@@H](C)C(N[C@@H](CS)C(N[C@@H](C(O)=O)CC4=CN3C5=C4C=CC=C5)=O)=O)=O)C)=O)=O)=O)CCC2)=O)CC1
QPTTAVGW	8	9	8	286.47	8	795.857	-2.224	O=C1N[C@H](C(N2[C@H](C(NC(C(C)O)C(N[C@H]3C(N[C@H](C(N[C@@H](C(C)C)C(NCC(N[C@@H](C(O)=O)CC4=CN3C5=C4C=CC=C5)=O)=O)=O)C)=O)=O)=O)CCC2)=O)CC1
QPTTVVGW	8	9	8	286.47	9	823.911	-1.381	O=C1N[C@H](C(N2[C@H](C(NC(C(C)O)C(N[C@H]3C(N[C@H](C(N[C@@H](C(C)C)C(NCC(N[C@@H](C(O)=O)CC4=CN3C5=C4C=CC=C5)=O)=O)=O)C(C)C)=O)=O)=O)CCC2)=O)CC1
QPVTVIGW	8	8	7	266.24	10	835.966	0.429	O=C1N[C@H](C(N2[C@H](C(N[C@@H](C(C)C)C(N[C@H]3C(N[C@H](C(NC(C(CC)C)C(NCC(N[C@@H](C(O)=O)CC4=CN3C5=C4C=CC=C5)=O)=O)=O)C(C)C)=O)=O)=O)CCC2)=O)CC1
QPVTVLGW	7	8	7	266.24	10	835.966	0.361	O=C1N[C@H](C(N2[C@H](C(N[C@@H](C(C)C)C(N[C@H]3C(N[C@H](C(N[C@@H](CC(C)C)C(NCC(N[C@@H](C(O)=O)CC4=CN3C5=C4C=CC=C5)=O)=O)=O)C(C)C)=O)=O)=O)CCC2)=O)CC1
QPVTVTGW	8	9	8	286.47	9	823.911	-1.381	O=C1N[C@H](C(N2[C@H](C(N[C@@H](C(C)C)C(N[C@H]3C(N[C@H](C(NC(C(C)O)C(NCC(N[C@@H](C(O)=O)CC4=CN3C5=C4C=CC=C5)=O)=O)=O)C(C)C)=O)=O)=O)CCC2)=O)CC1
QPVTVTSW	9	10	9	306.7	10	853.937	-1.785	O=C1N[C@H](C(N2[C@H](C(N[C@@H](C(C)C)C(N[C@H]3C(N[C@H](C(NC(C(C)O)C(N[C@@H](CO)C(N[C@@H](C(O)=O)CC4=CN3C5=C4C=CC=C5)=O)=O)=O)C(C)C)=O)=O)=O)CCC2)=O)CC1
QPWEVFSW	8	9	9	302.26	12	987.091	0.772	O=C1N[C@H](C(N2[C@H](C(N[C@@H](CC3=CNC4=C3C=CC=C4)C(N[C@H]5C(N[C@H](C(N[C@@H](CC6=CC=CC=C6)C(N[C@@H](CO)C(N[C@@H](C(O)=O)CC7=CN5C8=C7C=CC=C8)=O)=O)=O)C(C)C)=O)=O)=O)CCC2)=O)CC1
QPWGAYYW	8	10	10	322.49	12	1051.135	1.838	O=C1N[C@H](C(N2[C@H](C(N[C@@H](CC3=CNC4=C3C=CC=C4)C(N[C@H]5C(N[C@H](C(N[C@@H](CC6=CC=C(O)C=C6)C(N[C@@H](CC7=CC=C(O)C=C7)C(N[C@@H](C(O)=O)CC8=CN5C9=C8C=CC=C9)=O)=O)=O)C)=O)=O)=O)CCC2)=O)CC1
QPWGIHSW	9	10	10	330.94	13	991.082	-1.213	O=C1N[C@H](C(N2[C@H](C(N[C@@H](CC3=CNC4=C3C=CC=C4)C(N[C@H]5C(NC(C(N[C@@H](CC6=CNC=N6)C(N[C@@H](CO)C(N[C@@H](C(O)=O)CC7=CN5C8=C7C=CC=C8)=O)=O)=O)C(C)CC)=O)=O)=O)CCC2)=O)CC1
QPWGVASW	8	9	9	302.26	10	910.993	-0.784	O=C1N[C@H](C(N2[C@H](C(N[C@@H](CC3=CNC4=C3C=CC=C4)C(N[C@H]5C(N[C@H](C(N[C@@H](C)C(N[C@@H](CO)C(N[C@@H](C(O)=O)CC6=CN5C7=C6C=CC=C7)=O)=O)=O)C(C)C)=O)=O)=O)CCC2)=O)CC1
QPWGVFAW	8	8	8	282.03	11	971.092	1.66	O=C1N[C@H](C(N2[C@H](C(N[C@@H](CC3=CNC4=C3C=CC=C4)C(N[C@H]5C(N[C@H](C(N[C@@H](CC6=CC=CC=C6)C(N[C@@H](C)C(N[C@@H](C(O)=O)CC7=CN5C8=C7C=CC=C8)=O)=O)=O)C(C)C)=O)=O)=O)CCC2)=O)CC1
QPWGVGAW	7	8	8	282.03	9	880.966	-0.38	O=C1N[C@H](C(N2[C@H](C(N[C@@H](CC3=CNC4=C3C=CC=C4)C(N[C@H]5C(N[C@H](C(NCC(N[C@@H](C)C(N[C@@H](C(O)=O)CC6=CN5C7=C6C=CC=C7)=O)=O)=O)C(C)C)=O)=O)=O)CCC2)=O)CC1
QPWGVGSW	7	9	9	302.26	10	896.965	-1.268	O=C1N[C@H](C(N2[C@H](C(N[C@@H](CC3=CNC4=C3C=CC=C4)C(N[C@H]5C(N[C@H](C(NCC(N[C@@H](CO)C(N[C@@H](C(O)=O)CC6=CN5C7=C6C=CC=C7)=O)=O)=O)C(C)C)=O)=O)=O)CCC2)=O)CC1
QPWGVGTW	8	9	9	302.26	10	910.993	-0.891	O=C1N[C@H](C(N2[C@H](C(N[C@@H](CC3=CNC4=C3C=CC=C4)C(N[C@H]5C(N[C@H](C(NCC(NC(C(C)O)C(N[C@@H](C(O)=O)CC6=CN5C7=C6C=CC=C7)=O)=O)=O)C(C)C)=O)=O)=O)CCC2)=O)CC1
QPWGVNSW	8	10	10	345.35	12	954.018	-2.062	O=C1N[C@H](C(N2[C@H](C(N[C@@H](CC3=CNC4=C3C=CC=C4)C(N[C@H]5C(N[C@H](C(N[C@@H](CC(N)=O)C(NC(CO)C(N[C@@H](C(O)=O)CC6=CN5C7=C6C=CC=C7)=O)=O)=O)C(C)C)=O)=O)=O)CCC2)=O)CC1
QPWGVVAW	8	8	8	282.03	10	923.047	0.947	O=C1N[C@H](C(N2[C@H](C(N[C@@H](CC3=CNC4=C3C=CC=C4)C(N[C@H]5C(N[C@H](C(N[C@@H](C(C)C)C(N[C@@H](C)C(N[C@@H](C(O)=O)CC6=CN5C7=C6C=CC=C7)=O)=O)=O)C(C)C)=O)=O)=O)CCC2)=O)CC1
QPWTAVSW	8	9	9	302.26	10	910.993	-0.784	O=C1N[C@H](C(N2[C@H](C(N[C@@H](CC3=CNC4=C3C=CC=C4)C(N[C@H]5C(N[C@H](C(N[C@@H](C(C)C)C(N[C@@H](CO)C(N[C@@H](C(O)=O)CC6=CN5C7=C6C=CC=C7)=O)=O)=O)C)=O)=O)=O)CCC2)=O)CC1
QPWTVFSW	8	9	9	302.26	12	987.091	0.772	O=C1N[C@H](C(N2[C@H](C(N[C@@H](CC3=CNC4=C3C=CC=C4)C(N[C@H]5C(N[C@H](C(N[C@@H](CC6=CC=CC=C6)C(N[C@@H](CO)C(N[C@@H](C(O)=O)CC7=CN5C8=C7C=CC=C8)=O)=O)=O)C(C)C)=O)=O)=O)CCC2)=O)CC1
QPWTVGSW	7	9	9	302.26	10	896.965	-1.268	O=C1N[C@H](C(N2[C@H](C(N[C@@H](CC3=CNC4=C3C=CC=C4)C(N[C@H]5C(N[C@H](C(NCC(N[C@@H](CO)C(N[C@@H](C(O)=O)CC6=CN5C7=C6C=CC=C7)=O)=O)=O)C(C)C)=O)=O)=O)CCC2)=O)CC1
QPWTVSLW	8	9	9	302.26	12	953.074	0.448	O=C1N[C@H](C(N2[C@H](C(N[C@@H](CC3=CNC4=C3C=CC=C4)C(N[C@H]5C(N[C@H](C(N[C@@H](CO)C(N[C@@H](CC(C)C)C(N[C@@H](C(O)=O)CC6=CN5C7=C6C=CC=C7)=O)=O)=O)C(C)C)=O)=O)=O)CCC2)=O)CC1
QPWTVVSW	8	9	9	302.26	11	939.047	0.059	O=C1N[C@H](C(N2[C@H](C(N[C@@H](CC3=CNC4=C3C=CC=C4)C(N[C@H]5C(N[C@H](C(N[C@@H](C(C)C)C(N[C@@H](CO)C(N[C@@H](C(O)=O)CC6=CN5C7=C6C=CC=C7)=O)=O)=O)C(C)C)=O)=O)=O)CCC2)=O)CC1
QPYAVFAW	8	9	8	286.47	11	948.054	1.615	O=C1N[C@H](C(N2[C@H](C(N[C@@H](CC3=CC=C(O)C=C3)C(N[C@H]4C(N[C@H](C(N[C@@H](CC5=CC=CC=C5)C(N[C@@H](C)C(N[C@@H](C(O)=O)CC6=CN4C7=C6C=CC=C7)=O)=O)=O)C(C)C)=O)=O)=O)CCC2)=O)CC1
QPYGVEYW	8	10	9	344	14	1022.091	1.024	O=C1N[C@H](C(N2[C@H](C(N[C@@H](CC3=CC=C(O)C=C3)C(N[C@H]4C(N[C@H](C(N[C@@H](CCC(O)=O)C(N[C@@H](CC5=CC=C(O)C=C5)C(N[C@@H](C(O)=O)CC6=CN4C7=C6C=CC=C7)=O)=O)=O)C(C)C)=O)=O)=O)CCC2)=O)CC1
QPYGVFAW	8	9	8	286.47	11	948.054	1.615	O=C1N[C@H](C(N2[C@H](C(N[C@@H](CC3=CC=C(O)C=C3)C(N[C@H]4C(N[C@H](C(N[C@@H](CC5=CC=CC=C5)C(N[C@@H](C)C(N[C@@H](C(O)=O)CC6=CN4C7=C6C=CC=C7)=O)=O)=O)C(C)C)=O)=O)=O)CCC2)=O)CC1
QPYGVFGW	7	9	8	286.47	11	934.027	1.131	O=C1N[C@H](C(N2[C@H](C(N[C@@H](CC3=CC=C(O)C=C3)C(N[C@H]4C(N[C@H](C(N[C@@H](CC5=CC=CC=C5)C(NCC(N[C@@H](C(O)=O)CC6=CN4C7=C6C=CC=C7)=O)=O)=O)C(C)C)=O)=O)=O)CCC2)=O)CC1
QPYGVFPW	8	9	7	277.68	11	974.093	1.933	O=C1N[C@H](C(N2[C@H](C(N[C@@H](CC3=CC=C(O)C=C3)C(N[C@H]4C(N[C@H](C(N[C@@H](CC5=CC=CC=C5)C(N6[C@@H](CCC6)C(N[C@@H](C(O)=O)CC7=CN4C8=C7C=CC=C8)=O)=O)=O)C(C)C)=O)=O)=O)CCC2)=O)CC1
QPYGVFSW	8	10	9	306.7	12	964.054	0.727	O=C1N[C@H](C(N2[C@H](C(N[C@@H](CC3=CC=C(O)C=C3)C(N[C@H]4C(N[C@H](C(N[C@@H](CC5=CC=CC=C5)C(N[C@@H](CO)C(N[C@@H](C(O)=O)CC6=CN4C7=C6C=CC=C7)=O)=O)=O)C(C)C)=O)=O)=O)CCC2)=O)CC1

QPYGVFTW	9	10	9	306.7	12	978.081	1.104	O=C1N[C@H](C(N2[C@H](C(N[C@@H](CC3=CC=C(O)C=C3)C(N[C@H]4C(N[C@H](C(N[C@@H](CC5=CC=CC=C5)C(NC(C(C)O)C(N[C@@H](C(O)=O)CC6=CN4C7=C6C=CC=C7)=O)=O)=O)C(C)C)=O)=O)=O)CCC2)=O)CC1
QPYGVFVW	8	9	8	286.47	12	976.109	2.458	O=C1N[C@H](C(N2[C@H](C(N[C@@H](CC3=CC=C(O)C=C3)C(N[C@H]4C(N[C@H](C(N[C@@H](CC5=CC=CC=C5)C(N[C@@H](C(C)C)C(N[C@@H](C(O)=O)CC6=CN4C7=C6C=CC=C7)=O)=O)=O)C(C)C)=O)=O)=O)CCC2)=O)CC1
QPYGVGAW	7	9	8	286.47	9	857.928	-0.425	O=C1N[C@H](C(N2[C@H](C(N[C@@H](CC3=CC=C(O)C=C3)C(N[C@H]4C(N[C@H](C(NCC(N[C@@H](C)C)C(N[C@@H](C(O)=O)CC5=CN4C6=C5C=CC=C6)=O)=O)=O)C(C)C)=O)=O)=O)CCC2)=O)CC1
QPYGVGFW	7	9	8	286.47	11	934.027	1.131	O=C1N[C@H](C(N2[C@H](C(N[C@@H](CC3=CC=C(O)C=C3)C(N[C@H]4C(N[C@H](C(NCC(N[C@@H](CC5=CC=CC=C5)C(N[C@@H](C(O)=O)CC6=CN4C7=C6C=CC=C7)=O)=O)=O)C(C)C)=O)=O)=O)CCC2)=O)CC1
QPYGVGIW	8	9	8	286.47	11	900.01	0.874	O=C1N[C@H](C(N2[C@H](C(N[C@@H](CC3=CC=C(O)C=C3)C(N[C@H]4C(N[C@H](C(NCC(NC(C(C)CC)C(N[C@@H](C(O)=O)CC5=CN4C6=C5C=CC=C6)=O)=O)=O)C(C)C)=O)=O)=O)CCC2)=O)CC1
QPYGVGSW	7	10	9	306.7	10	873.928	-1.313	O=C1N[C@H](C(N2[C@H](C(N[C@@H](CC3=CC=C(O)C=C3)C(N[C@H]4C(N[C@H](C(NCC(N[C@@H](CO)C(N[C@@H](C(O)=O)CC5=CN4C6=C5C=CC=C6)=O)=O)C(C)C)=O)=O)=O)CCC2)=O)CC1
QPYGVIGW	8	9	8	286.47	11	900.01	0.874	O=C1N[C@H](C(N2[C@H](C(N[C@@H](CC3=CC=C(O)C=C3)C(N[C@H]4C(N[C@H](C(NC(C(C)CC)C(NCC(N[C@@H](C(O)=O)CC5=CN4C6=C5C=CC=C6)=O)=O)=O)C(C)C)=O)=O)=O)CCC2)=O)CC1
QPYGVLFW	8	9	8	286.47	13	990.136	2.847	O=C1N[C@H](C(N2[C@H](C(N[C@@H](CC3=CC=C(O)C=C3)C(N[C@H]4C(N[C@H](C(N[C@@H](CC(C)C)C(N[C@@H](CC5=CC=CC=C5)C(N[C@@H](C(O)=O)CC6=CN4C7=C6C=CC=C7)=O)=O)=O)C(C)C)=O)=O)=O)CCC2)=O)CC1
QPYGVLYW	8	10	9	306.7	13	1006.135	2.579	O=C1N[C@H](C(N2[C@H](C(N[C@@H](CC3=CC=C(O)C=C3)C(N[C@H]4C(N[C@H](C(N[C@@H](CC(C)C)C(N[C@@H](CC5=CC=C(O)C=C5)C(N[C@@H](C(O)=O)CC6=CN4C7=C6C=CC=C7)=O)=O)=O)C(C)C)=O)=O)=O)CCC2)=O)CC1
QPYGVSIW	9	10	9	306.7	12	930.036	0.47	O=C1N[C@H](C(N2[C@H](C(N[C@@H](CC3=CC=C(O)C=C3)C(N[C@H]4C(N[C@H](C(N[C@@H](CO)C(NC(C(CC)C)C(N[C@@H](C(O)=O)CC5=CN4C6=C5C=CC=C6)=O)=O)=O)C(C)C)=O)=O)=O)CCC2)=O)CC1
QPYGVWGW	7	9	9	302.26	11	973.064	0.908	O=C1N[C@H](C(N2[C@H](C(N[C@@H](CC3=CC=C(O)C=C3)C(N[C@H]4C(N[C@H](C(N[C@@H](CC5=CNC6=C5C=CC=C6)C(NCC(N[C@@H](C(O)=O)CC7=CN4C8=C7C=CC=C8)=O)=O)=O)C(C)C)=O)=O)=O)CCC2)=O)CC1
QPYGVYAW	8	10	9	306.7	11	964.054	1.347	O=C1N[C@H](C(N2[C@H](C(N[C@@H](CC3=CC=C(O)C=C3)C(N[C@H]4C(N[C@H](C(N[C@@H](CC5=CC=C(O)C=C5)C(N[C@@H](C)C(N[C@@H](C(O)=O)CC6=CN4C7=C6C=CC=C7)=O)=O)=O)C(C)C)=O)=O)=O)CCC2)=O)CC1
QPYGVYFW	8	10	9	306.7	13	1040.153	2.903	O=C1N[C@H](C(N2[C@H](C(N[C@@H](CC3=CC=C(O)C=C3)C(N[C@H]4C(N[C@H](C(N[C@@H](CC5=CC=C(O)C=C5)C(N[C@@H](CC6=CC=CC=C6)C(N[C@@H](C(O)=O)CC7=CN4C8=C7C=CC=C8)=O)=O)=O)C(C)C)=O)=O)=O)CCC2)=O)CC1
QPYGVYLW	8	10	9	306.7	13	1006.135	2.579	O=C1N[C@H](C(N2[C@H](C(N[C@@H](CC3=CC=C(O)C=C3)C(N[C@H]4C(N[C@H](C(N[C@@H](CC5=CC=C(O)C=C5)C(N[C@@H](CC(C)C)C(N[C@@H](C(O)=O)CC6=CN4C7=C6C=CC=C7)=O)=O)=O)C(C)C)=O)=O)=O)CCC2)=O)CC1
QPYGVYSW	8	11	10	326.93	12	980.053	0.459	O=C1N[C@H](C(N2[C@H](C(N[C@@H](CC3=CC=C(O)C=C3)C(N[C@H]4C(N[C@H](C(N[C@@H](CC5=CC=C(O)C=C5)C(N[C@@H](CO)C(N[C@@H](C(O)=O)CC6=CN4C7=C6C=CC=C7)=O)=O)=O)C(C)C)=O)=O)=O)CCC2)=O)CC1
QPYGVYTW	9	11	10	326.93	12	994.08	0.837	O=C1N[C@H](C(N2[C@H](C(N[C@@H](CC3=CC=C(O)C=C3)C(N[C@H]4C(N[C@H](C(N[C@@H](CC5=CC=C(O)C=C5)C(NC(C(C)O)C(N[C@@H](C(O)=O)CC6=CN4C7=C6C=CC=C7)=O)=O)=O)C(C)C)=O)=O)=O)CCC2)=O)CC1
QPYSVFAW	8	9	8	286.47	11	948.054	1.615	O=C1N[C@H](C(N2[C@H](C(N[C@@H](CC3=CC=C(O)C=C3)C(N[C@H]4C(N[C@H](C(N[C@@H](CC5=CC=CC=C5)C(N[C@@H](C)C(N[C@@H](C(O)=O)CC6=CN4C7=C6C=CC=C7)=O)=O)=O)C(C)C)=O)=O)=O)CCC2)=O)CC1
QPYTALSW	8	10	9	306.7	11	901.982	-0.44	O=C1N[C@H](C(N2[C@H](C(N[C@@H](CC3=CC=C(O)C=C3)C(N[C@H]4C(N[C@H](C(N[C@@H](CC(C)C)C(N[C@@H](CO)C(N[C@@H](C(O)=O)CC5=CN4C6=C5C=CC=C6)=O)=O)=O)C(C)C)=O)=O)=O)CCC2)=O)CC1
QPYTVFAW	8	9	8	286.47	11	948.054	1.615	O=C1N[C@H](C(N2[C@H](C(N[C@@H](CC3=CC=C(O)C=C3)C(N[C@H]4C(N[C@H](C(N[C@@H](CC5=CC=CC=C5)C(N[C@@H](C)C(N[C@@H](C(O)=O)CC6=CN4C7=C6C=CC=C7)=O)=O)=O)C(C)C)=O)=O)=O)CCC2)=O)CC1
QPYTVFGW	7	9	8	286.47	11	934.027	1.131	O=C1N[C@H](C(N2[C@H](C(N[C@@H](CC3=CC=C(O)C=C3)C(N[C@H]4C(N[C@H](C(N[C@@H](CC5=CC=CC=C5)C(NCC(N[C@@H](C(O)=O)CC6=CN4C7=C6C=CC=C7)=O)=O)=O)C(C)C)=O)=O)=O)CCC2)=O)CC1
QPYTVFLW	8	9	8	286.47	13	990.136	2.847	O=C1N[C@H](C(N2[C@H](C(N[C@@H](CC3=CC=C(O)C=C3)C(N[C@H]4C(N[C@H](C(N[C@@H](CC5=CC=CC=C5)C(N[C@@H](CC(C)C)C(N[C@@H](C(O)=O)CC6=CN4C7=C6C=CC=C7)=O)=O)=O)C(C)C)=O)=O)=O)CCC2)=O)CC1
QPYTVFNW	8	10	9	329.56	13	991.08	0.337	O=C1N[C@H](C(N2[C@H](C(N[C@@H](CC3=CC=C(O)C=C3)C(N[C@H]4C(N[C@H](C(N[C@@H](CC5=CC=CC=C5)C(N[C@@H](CC(N)=O)C(N[C@@H](C(O)=O)CC6=CN4C7=C6C=CC=C7)=O)=O)=O)C(C)C)=O)=O)=O)CCC2)=O)CC1
QPYTVFSW	8	10	9	306.7	12	964.054	0.727	O=C1N[C@H](C(N2[C@H](C(N[C@@H](CC3=CC=C(O)C=C3)C(N[C@H]4C(N[C@H](C(N[C@@H](CC5=CC=CC=C5)C(N[C@@H](CO)C(N[C@@H](C(O)=O)CC6=CN4C7=C6C=CC=C7)=O)=O)=O)C(C)C)=O)=O)=O)CCC2)=O)CC1
QPYTVGAW	7	9	8	286.47	9	857.928	-0.425	O=C1N[C@H](C(N2[C@H](C(N[C@@H](CC3=CC=C(O)C=C3)C(N[C@H]4C(N[C@H](C(NCC(N[C@@H](C)C(N[C@@H](C(O)=O)CC5=CN4C6=C5C=CC=C6)=O)=O)=O)C(C)C)=O)=O)=O)CCC2)=O)CC1
QPYTVGSW	7	10	9	306.7	10	873.928	-1.313	O=C1N[C@H](C(N2[C@H](C(N[C@@H](CC3=CC=C(O)C=C3)C(N[C@H]4C(N[C@H](C(NCC(N[C@@H](CO)C(N[C@@H](C(O)=O)CC5=CN4C6=C5C=CC=C6)=O)=O)C(C)C)=O)=O)=O)CCC2)=O)CC1
QPYTVIAW	9	9	8	286.47	11	914.037	1.358	O=C1N[C@H](C(N2[C@H](C(N[C@@H](CC3=CC=C(O)C=C3)C(N[C@H]4C(N[C@H](C(NC(C(CC)C)C(N[C@@H](C)C(N[C@@H](C(O)=O)CC5=CN4C6=C5C=CC=C6)=O)=O)=O)C(C)C)=O)=O)=O)CCC2)=O)CC1
QPYTVLHW	8	10	9	315.15	13	980.1	0.405	O=C1N[C@H](C(N2[C@H](C(N[C@@H](CC3=CC=C(O)C=C3)C(N[C@H]4C(N[C@H](C(N[C@@H](CC(C)C)C(N[C@@H](CC5=CNC=N5)C(N[C@@H](C(O)=O)CC6=CN4C7=C6C=CC=C7)=O)=O)=O)C(C)C)=O)=O)=O)CCC2)=O)CC1
QPYTVMAW	8	10	8	311.77	12	932.074	0.386	O=C1N[C@H](C(N2[C@H](C(N[C@@H](CC3=CC=C(O)C=C3)C(N[C@H]4C(N[C@H](C(N[C@@H](CCSC)C(N[C@@H](C)C(N[C@@H](C(O)=O)CC5=CN4C6=C5C=CC=C6)=O)=O)=O)C(C)C)=O)=O)=O)CCC2)=O)CC1
QPYTVMGW	7	10	8	311.77	12	918.047	-0.098	O=C1N[C@H](C(N2[C@H](C(N[C@@H](CC3=CC=C(O)C=C3)C(N[C@H]4C(N[C@H](C(N[C@@H](CCSC)C(NCC(N[C@@H](C(O)=O)CC5=CN4C6=C5C=CC=C6)=O)=O)=O)C(C)C)=O)=O)=O)CCC2)=O)CC1
QPYTVSAW	8	10	9	306.7	10	887.955	-0.829	O=C1N[C@H](C(N2[C@H](C(N[C@@H](CC3=CC=C(O)C=C3)C(N[C@H]4C(N[C@H](C(N[C@@H](CO)C(N[C@@H](C)C(N[C@@H](C(O)=O)CC5=CN4C6=C5C=CC=C6)=O)=O)=O)C(C)C)=O)=O)=O)CCC2)=O)CC1
QPYTVSSW	8	11	10	326.93	11	903.954	-1.717	O=C1N[C@H](C(N2[C@H](C(N[C@@H](CC3=CC=C(O)C=C3)C(N[C@H]4C(N[C@H](C(N[C@@H](CO)C(N[C@@H](CO)C(N[C@@H](C(O)=O)CC5=CN4C6=C5C=CC=C6)=O)=O)=O)C(C)C)=O)=O)=O)CCC2)=O)CC1
QPYTVTAW	9	10	9	306.7	10	901.982	-0.452	O=C1N[C@H](C(N2[C@H](C(N[C@@H](CC3=CC=C(O)C=C3)C(N[C@H]4C(N[C@H](C(NC(C(C)O)C(N[C@@H](C)C(N[C@@H](C(O)=O)CC5=CN4C6=C5C=CC=C6)=O)=O)=O)C(C)C)=O)=O)=O)CCC2)=O)CC1
QPYTVWGW	7	9	9	302.26	11	973.064	0.908	O=C1N[C@H](C(N2[C@H](C(N[C@@H](CC3=CC=C(O)C=C3)C(N[C@H]4C(N[C@H](C(N[C@@H](CC5=CNC6=C5C=CC=C6)C(NCC(N[C@@H](C(O)=O)CC7=CN4C8=C7C=CC=C8)=O)=O)=O)C(C)C)=O)=O)=O)CCC2)=O)CC1
QPYTVYSW	8	11	10	326.93	12	980.053	0.459	O=C1N[C@H](C(N2[C@H](C(N[C@@H](CC3=CC=C(O)C=C3)C(N[C@H]4C(N[C@H](C(N[C@@H](CC5=CC=C(O)C=C5)C(N[C@@H](CO)C(N[C@@H](C(O)=O)CC6=CN4C7=C6C=CC=C7)=O)=O)=O)C(C)C)=O)=O)=O)CCC2)=O)CC1
QPAYNYLW	9	10	8	318.56	5	926.992	-1.132	O=C([C@@H]1C2)N[C@@H]2C(N3[C@@H](CCC3)C(N[C@@H](C)C(N[C@H]4C(N[C@@H](CC(N)=O)C(N[C@@H](CC5=CC=C(O1)C=C5)C(N[C@@H](CC(C)C)C(N[C@@H](CC6=CN4C7=C6C=CC=C7)C(O)=O)=O)=O)=O)=O)=O)=O
QPFGAYYW	9	10	8	295.7	5	1010.082	1.758	O=C([C@@H]1C2)N[C@@H]2C(N3[C@@H](CCC3)C(N[C@@H](CC4=CC=CC=C4)C(N[C@H]5C(N[C@@H](C)C(N[C@@H](CC6=CC=C(O1)C=C6)C(N[C@@H](CC7=CC=C(O)C=C7)C(N[C@@H](CC8=CN5C9=C8C=CC=C9)C(O)=O)=O)=O)=O)=O)=O)=O)=O
QPFGDYVW	9	9	7	312.77	6	990.048	0.668	O=C([C@@H]1C2)N[C@@H]2C(N3[C@@H](CCC3)C(N[C@@H](CC4=CC=CC=C4)C(N[C@H]5C(N[C@@H](CC(O)=O)C(N[C@@H](CC6=CC=C(O1)C=C6)C(N[C@@H](C(C)C)C(N[C@@H](CC7=CN5C8=C7C=CC=C8)C(O)=O)=O)=O)=O)=O)=O)=O)=O
QPFGVYAW	9	9	7	275.47	4	946.038	1.313	O=C([C@@H]1C2)N[C@@H]2C(N3[C@@H](CCC3)C(N[C@@H](CC4=CC=CC=C4)C(N[C@H]5C(N[C@@H](C(C)C)C(N[C@@H](CC6=CC=C(O1)C=C6)C(N[C@@H](C)C(N[C@@H](CC7=CN5C8=C7C=CC=C8)C(O)=O)=O)=O)=O)=O)=O)=O)=O

QPFTVYAW	9	9	7	275.47	4	946.038	1.313	O=C([C@@H]1C2)N[C@@H]2C(N3[C@@H](CCC3)C(N[C@@H](CC4=CC=CC=C4)C(N[C@H]5C(N[C@@H](C(C)C)C(N[C@@H](CC6=CC=C(O1)C=C6)C(N[C@@H](C)C(N[C@@H](CC7=CN5C8=C7C=CC=C8)C(O)=O)=O)=O)=O)=O)=O)=O
QPFTVYGW	8	9	7	275.47	4	932.011	0.829	O=C([C@@H]1C2)N[C@@H]2C(N3[C@@H](CCC3)C(N[C@@H](CC4=CC=CC=C4)C(N[C@H]5C(N[C@@H](C(C)C)C(N[C@@H](CC6=CC=C(O1)C=C6)C(NCC(N[C@@H](CC7=CN5C8=C7C=CC=C8)C(O)=O)=O)=O)=O)=O)=O)=O
QPFTVYSW	9	10	8	295.7	5	962.038	0.425	O=C([C@@H]1C2)N[C@@H]2C(N3[C@@H](CCC3)C(N[C@@H](CC4=CC=CC=C4)C(N[C@H]5C(N[C@@H](C(C)C)C(N[C@@H](CC6=CC=C(O1)C=C6)C(N[C@@H](CO)C(N[C@@H](CC7=CN5C8=C7C=CC=C8)C(O)=O)=O)=O)=O)=O)=O)=O
QPWGAYYW	9	10	9	311.49	5	1049.119	1.678	O=C([C@@H]1C2)N[C@@H]2C(N3[C@@H](CCC3)C(N[C@@H](CC4=CC(C=CC=C5)=C5N4)C(N[C@H]6C(N[C@@H](C)C(N[C@@H](CC7=CC=C(O1)C=C7)C(N[C@@H](CC8=CC=C(O)C=C8)C(N[C@@H](CC9=CN6C%10=C9C=CC=C%10)C(O)=O)=O)=O)=O)=O)=O)=O
QPYGVYAW	9	10	8	295.7	4	962.038	1.045	O=C([C@@H]1C2)N[C@@H]2C(N3[C@@H](CCC3)C(N[C@@H](CC4=CC=C(O)C=C4)C(N[C@H]5C(N[C@@H](C(C)C)C(N[C@@H](CC6=CC=C(O1)C=C6)C(N[C@@H](C)C(N[C@@H](CC7=CN5C8=C7C=CC=C8)C(O)=O)=O)=O)=O)=O)=O)=O
QPYGVYFW	9	10	8	295.7	6	1038.137	2.601	O=C([C@@H]1C2)N[C@@H]2C(N3[C@@H](CCC3)C(N[C@@H](CC4=CC=C(O)C=C4)C(N[C@H]5C(N[C@@H](C(C)C)C(N[C@@H](CC6=CC=C(O1)C=C6)C(N[C@@H](CC7=CC=CC=C7)C(N[C@@H](CC8=CN5C9=C8C=CC=C9)C(O)=O)=O)=O)=O)=O)=O)=O
QPYGVYLW	9	10	8	295.7	6	1004.119	2.277	O=C([C@@H]1C2)N[C@@H]2C(N3[C@@H](CCC3)C(N[C@@H](CC4=CC=C(O)C=C4)C(N[C@H]5C(N[C@@H](C(C)C)C(N[C@@H](CC6=CC=C(O1)C=C6)C(N[C@@H](CC(C)C)C(N[C@@H](CC7=CN5C8=C7C=CC=C8)C(O)=O)=O)=O)=O)=O)=O)=O
QPYGVYSW	9	11	9	315.93	5	978.037	0.157	O=C([C@@H]1C2)N[C@@H]2C(N3[C@@H](CCC3)C(N[C@@H](CC4=CC=C(O)C=C4)C(N[C@H]5C(N[C@@H](C(C)C)C(N[C@@H](CC6=CC=C(O1)C=C6)C(N[C@@H](CO)C(N[C@@H](CC7=CN5C8=C7C=CC=C8)C(O)=O)=O)=O)=O)=O)=O)=O
QPYGVYTW	10	11	9	315.93	5	992.064	0.535	O=C([C@@H]1C2)N[C@@H]2C(N3[C@@H](CCC3)C(N[C@@H](CC4=CC=C(O)C=C4)C(N[C@H]5C(N[C@@H](C(C)C)C(N[C@@H](CC6=CC=C(O1)C=C6)C(N[C@@H]([C@H](O)C)C(N[C@@H](CC7=CN5C8=C7C=CC=C8)C(O)=O)=O)=O)=O)=O)=O)=O
QPYTVYSW	9	11	9	315.93	5	978.037	0.157	O=C([C@@H]1C2)N[C@@H]2C(N3[C@@H](CCC3)C(N[C@@H](CC4=CC=C(O)C=C4)C(N[C@H]5C(N[C@@H](C(C)C)C(N[C@@H](CC6=CC=C(O1)C=C6)C(N[C@@H](CO)C(N[C@@H](CC7=CN5C8=C7C=CC=C8)C(O)=O)=O)=O)=O)=O)=O)=O
QILIW	7	5	5	187.73	8	652.798	2.331	O=C1CC[C@@H](C(N[C@@H](C(N[C@@H](CC(C)C)C(NC(C(C)CC)C(N[C@H](C(O)=O)CC2=CN3C4=C2C=CC=C4)=O)=O)=O)[C@@]3(C)CC)=O)N1
QILFW	6	5	5	187.73	8	686.815	2.588	O=C1CC[C@@H](C(N[C@@H](C(N[C@@H](CC(C)C)C(N[C@@H](CC2=CC=CC=C2)C(N[C@@H](C(O)=O)CC3=CN4C5=C3C=CC=C5)=O)=O)=O)[C@@]4(C)CC)=O)N1
QLFFW	6	5	6	198.59	8	720.833	2.595	O=C1CC[C@@H](C(N[C@@H](C(N[C@@H](CC2=CC=CC=C2)C(N[C@@H](CC3=CC=CC=C3)C(N[C@H](C(O)=O)CC4=CNC5=C4C=CC6=C5)=O)=O)=O)[C@@H]6C(C)C)=O)N1
QLFVW	6	5	6	198.59	7	672.788	1.882	O=C1CC[C@@H](C(N[C@@H](C(N[C@@H](CC2=CC=CC=C2)C(N[C@@H](C(C)C)C(N[C@H](C(O)=O)CC3=CNC4=C3C=CC5=C4)=O)=O)=O)[C@@H]5C(C)C)=O)N1
QLKFW	6	5	6	224.61	10	701.83	0.96	O=C1CC[C@@H](C(N[C@@H](C(N[C@@H](CCCCN)C(N[C@@H](CC2=CC=CC=C2)C(N[C@H](C(O)=O)CC3=CNC4=C3C=CC5=C4)=O)=O)=O)[C@@H]5C(C)C)=O)N1
QLKVW	6	5	6	224.61	9	653.785	0.247	O=C1CC[C@@H](C(N[C@@H](C(N[C@@H](CCCCN)C(N[C@@H](C(C)C)C(N[C@H](C(O)=O)CC2=CNC3=C2C=CC4=C3)=O)=O)=O)[C@@H]4C(C)C)=O)N1
QLLFW	6	5	6	198.59	8	686.815	2.271	O=C1CC[C@@H](C(N[C@@H](C(N[C@@H](CC(C)C)C(N[C@@H](CC2=CC=CC=C2)C(N[C@H](C(O)=O)CC3=CNC4=C3C=CC5=C4)=O)=O)=O)[C@@H]5C(C)C)=O)N1
QLLIW	7	5	6	198.59	8	652.798	2.014	O=C1CC[C@@H](C(N[C@@H](C(N[C@@H](CC(C)C)C(NC(C(CC)C)C(N[C@H](C(O)=O)CC2=CNC3=C2C=CC4=C3)=O)=O)=O)[C@@H]4C(C)C)=O)N1
QLLVW	6	5	6	198.59	7	638.771	1.558	O=C1CC[C@@H](C(N[C@@H](C(N[C@@H](CC(C)C)C(N[C@@H](C(C)C)C(N[C@H](C(O)=O)CC2=CNC3=C2C=CC4=C3)=O)=O)=O)[C@@H]4C(C)C)=O)N1
QLNFW	6	6	7	241.68	8	687.759	-0.239	O=C1CC[C@@H](C(N[C@@H](C(N(C@@H](CC(N)=O)C(N[C@@H](CC2=CC=CC=C2)C(N[C@H](C(O)=O)CC3=CNC4=C3C=CC5=C4)=O)=O)=O)[C@@H]5C(C)C)=O)N1
QLRAW	6	6	7	260.49	9	653.744	-1.186	O=C1CC[C@@H](C(N[C@@H](C(N[C@@H](CCCNC(N)=N)C(N[C@@H](C)C(N[C@H](C(O)=O)CC2=CNC3=C2C=CC4=C3)=O)=O)=O)[C@@H]4C(C)C)=O)N1
QLRDW	6	6	7	297.79	11	697.754	-1.83	O=C1CC[C@@H](C(N[C@@H](C(N[C@@H](CCCNC(N)=N)C(N[C@@H](CC(O)=O)C(N[C@H](C(O)=O)CC2=CNC3=C2C=CC4=C3)=O)=O)=O)[C@@H]4C(C)C)=O)N1
QLRVW	6	6	7	260.49	10	681.799	-0.343	O=C1CC[C@@H](C(N[C@@H](C(N[C@@H](CCCNC(N)=N)C(N[C@@H](C(C)C)C(N[C@H](C(O)=O)CC2=CNC3=C2C=CC4=C3)=O)=O)=O)[C@@H]4C(C)C)=O)N1
QLSVW	6	6	7	218.82	6	612.689	-0.562	O=C1CC[C@@H](C(N[C@@H](C(N[C@@H](CO)C(N[C@@H](C(C)C)C(N[C@H](C(O)=O)CC2=CNC3=C2C=CC4=C3)=O)=O)=O)[C@@H]4C(C)C)=O)N1
QLKVWQEH	9	10	10	410.12	15	1046.16	-2.859	O=C1CC[C@@H](C(N[C@@H](C(N(C@@H](CCCCN)C(N[C@@H](C(C)C)C(N[C@H](C(N[C@@H](CCC(N)=O)C(N[C@@H](CCC(O)=O)C(N[C@@H](CC2=CN3C=N2)C(O)=O)=O)=O)=O)CC4=C3NC5=C4C=CC6=C5)=O)=O)=O)[C@@H]6C(C)C)=O)N1
QLKVWQKH	9	10	10	398.84	16	1045.219	-2.615	O=C1CC[C@@H](C(N[C@@H](C(N[C@@H](CCCCN)C(N[C@@H](C(C)C)C(N[C@H](C(N[C@@H](CCC(N)=O)C(N[C@@H](CCCCN)C(N[C@@H](CC2=CN3C=N2)C(O)=O)=O)=O)=O)CC4=C3NC5=C4C=CC6=C5)=O)=O)=O)[C@@H]6C(C)C)=O)N1
QLLFWTRH	10	11	11	385.84	14	1079.236	-0.735	O=C1CC[C@@H](C(N[C@@H](C(N[C@@H](CC(C)C)C(N[C@@H](CC2=CC=CC=C2)C(N[C@H](C(NC(C(O)C)C(N[C@@H](CCCNC(N)=N)C(N[C@@H](CC3=CN4C=N3)C(O)=O)=O)=O)=O)CC5=C4NC6=C5C=CC7=C6)=O)=O)=O)[C@@H]7C(C)C)=O)N1
QLLIWMNH	10	11	10	372.1	13	1033.226	-0.762	O=C1CC[C@@H](C(N[C@@H](C(N[C@@H](CC(C)C)C(NC(C(CC)C)C(N[C@H](C(N[C@@H](CCSC)C(N[C@@H](CC(N)=O)C(N[C@@H](CC2=CN3C=N2)C(O)=O)=O)=O)=O)CC4=C3NC5=C4C=CC6=C5)=O)=O)=O)[C@@H]6C(C)C)=O)N1
QLLVWQKH	9	10	10	372.82	14	1030.204	-1.304	O=C1CC[C@@H](C(N[C@@H](C(N[C@@H](CC(C)C)C(N[C@@H](C(C)C)C(N[C@H](C(N[C@@H](CCC(N)=O)C(N[C@@H](CCCCN)C(N[C@@H](CC2=CN3C=N2)C(O)=O)=O)=O)=O)CC4=C3NC5=C4C=CC6=C5)=O)=O)=O)[C@@H]6C(C)C)=O)N1
QLLVWQLH	9	10	10	346.8	12	1015.189	0.007	O=C1CC[C@@H](C(N[C@@H](C(N(C@@H](CC(C)C)C(N[C@@H](C(C)C)C(N[C@H](C(N[C@@H](CCC(N)=O)C(N[C@@H](CC(C)C)C(N[C@@H](CC2=CN3C=N2)C(O)=O)=O)=O)=O)CC4=C3NC5=C4C=CC6=C5)=O)=O)=O)[C@@H]6C(C)C)=O)N1
QLLVWRAH	9	10	10	365.61	12	1001.165	-0.937	O=C1CC[C@@H](C(N[C@@H](C(N[C@@H](CC(C)C)C(N[C@@H](C(C)C)C(N[C@H](C(N[C@@H](CCCNC(N)=N)C(N[C@@H](C)C(N[C@@H](CC2=CN3C=N2)C(O)=O)=O)=O)=O)CC4=C3NC5=C4C=CC6=C5)=O)=O)=O)[C@@H]6C(C)C)=O)N1
QLLVWRGH	8	10	10	365.61	12	987.138	-1.421	O=C1CC[C@@H](C(N[C@@H](C(N[C@@H](CC(C)C)C(N[C@@H](C(C)C)C(N[C@H](C(N[C@@H](CCCNC(N)=N)C(NCC(N[C@@H](CC2=CN3C=N2)C(O)=O)=O)=O)=O)CC4=C3NC5=C4C=CC6=C5)=O)=O)=O)[C@@H]6C(C)C)=O)N1
QLLVWRNH	9	11	11	408.7	14	1044.19	-2.215	O=C1CC[C@@H](C(N[C@@H](C(N[C@@H](CC(C)C)C(N[C@@H](C(C)C)C(N[C@H](C(N[C@@H](CCCNC(N)=N)C(N[C@@H](CC(N)=O)C(N[C@@H](CC2=CN3C=N2)C(O)=O)=O)=O)=O)CC4=C3NC5=C4C=CC6=C5)=O)=O)=O)[C@@H]6C(C)C)=O)N1
QLLVWRSH	9	11	11	385.84	13	1017.164	-1.825	O=C1CC[C@@H](C(N[C@@H](C(N[C@@H](CC(C)C)C(N[C@@H](C(C)C)C(N[C@H](C(N[C@@H](CCCNC(N)=N)C(N[C@@H](CO)C(N[C@@H](CC2=CN3C=N2)C(O)=O)=O)=O)=O)CC4=C3NC5=C4C=CC6=C5)=O)=O)=O)[C@@H]6C(C)C)=O)N1
QLLVWSRH	9	11	11	385.84	13	1017.164	-1.825	O=C1CC[C@@H](C(N[C@@H](C(N(C@@H](CC(C)C)C(N[C@@H](C(C)C)C(N[C@H](C(N[C@@H](CO)C(N[C@@H](CCCNC(N)=N)C(N[C@@H](CC2=CN3C=N2)C(O)=O)=O)=O)=O)CC4=C3NC5=C4C=CC6=C5)=O)=O)=O)[C@@H]6C(C)C)=O)N1
QLRVWQSH	9	12	12	428.93	14	1032.135	-4.014	O=C1CC[C@@H](C(N[C@@H](C(N[C@@H](CCCNC(N)=N)C(N[C@@H](C(C)C)C(N[C@H](C(N[C@@H](CCC(N)=O)C(N[C@@H](CO)C(N[C@@H](CC2=CN3C=N2)C(O)=O)=O)=O)=O)CC4=C3NC5=C4C=CC6=C5)=O)=O)=O)[C@@H]6C(C)C)=O)N1
QLRVWTKH	10	11	11	411.86	15	1046.206	-2.759	O=C1CC[C@@H](C(N[C@@H](C(N[C@@H](CCCNC(N)=N)C(N[C@@H](C(C)C)C(N[C@H](C(NC(C(C)O)C(N[C@@H](CCCCN)C(N[C@@H](CC2=CN3C=N2)C(O)=O)=O)=O)=O)CC4=C3NC5=C4C=CC6=C5)=O)=O)=O)[C@@H]6C(C)C)=O)N1
QLRVWTQH	10	12	12	428.93	14	1046.162	-3.636	O=C1CC[C@@H](C(N[C@@H](C(N[C@@H](CCCNC(N)=N)C(N[C@@H](C(C)C)C(N[C@H](C(NC(C(C)O)C(N[C@@H](CCC(N)=O)C(N[C@@H](CC2=CN3C=N2)C(O)=O)=O)=O)=O)CC4=C3NC5=C4C=CC6=C5)=O)=O)=O)[C@@H]6C(C)C)=O)N1

QLRVWTSH	10	12	12	406.07	12	1005.11	-3.568	O=C1CC[C@@H](C(N[C@H](C(N[C@@H](CCCNC(N)=N)C(N[C@@H](C(C)C)C(N[C@H](C(NC(C(C)O)C(N[C@@H](CO)C(N[C@@H](CC2=CN3C=N2)C(O)=O)=O)=O)=O)=O)CC4=C3NC5=C4C=CC6=C5)=O)=O)=O)[C@@H]6C(C)C)=O)N1
QLSVWTSH	10	12	12	364.4	8	936	-3.786	O=C1CC[C@@H](C(N[C@H](C(N[C@@H](CO)C(N[C@@H](C(C)C)C(N[C@H](C(NC(C(C)O)C(N[C@@H](CO)C(N[C@@H](CC2=CN3C=N2)C(O)=O)=O)=O)=O)CC4=C3NC5=C4C=CC6=C5)=O)=O)=O)[C@@H]6C(C)C)=O)N1
ILLYPSY	10	9	6	245.04	10	892.071	2.781	CC(C)[C@@H]1OC2=CC=C(C[C@H](OC3=CC=C(C[C@H](NC([C@H](CO)NC4=O)=O)C(O)=O)C=C3)[C@@H](C(N5[C@H]4CCC5)=O)NC([C@@H](NC([C@H]1NC([C@@H](N(C)C)C(CC)C)=O)=O)CC(C)C)=O)C=C2
ILLYQSY	10	10	8	296.92	13	923.085	1.506	CC(C)[C@@H]1OC2=CC=C(C[C@H](OC3=CC=C(C[C@H](NC([C@H](CO)NC4=O)=O)C(O)=O)C=C3)[C@@H](C(N[C@H]4CCC(N)=O)=O)NC([C@@H](NC([C@H]1NC([C@@H](N(C)C)C(CC)C)=O)=O)CC(C)C)=O)C=C2
LLSYQGY	8	10	8	296.92	11	866.977	-0.278	CC(C)[C@@H]1OC2=CC=C(C[C@H](OC3=CC=C(C[C@H](NC(CNC4=O)=O)C(O)=O)C=C3)[C@@H](C(N[C@H]4CCC(N)=O)=O)NC([C@@H](NC([C@H]1NC([C@@H](N(C)C)CC(C)C)=O)=O)CO)=O)C=C2
PLFYHRY	9	10	8	324.18	13	1005.153	0.933	CC(C)[C@@H]1OC2=CC=C(C[C@H](OC3=CC=C(C[C@H](NC([C@H](CCCNC(N)=N)NC4=O)=O)C(O)=O)C=C3)[C@@H](C(N[C@H]4CC5=CNC=N5)=O)NC([C@@H](NC([C@H]1NC(C6N(C)CCC6)=O)=O)CC7=CC=CC=C7)=O)C=C2
PLFYRSY	9	10	8	315.73	12	955.09	0.93	CC(C)[C@@H]1OC2=CC=C(C[C@H](OC3=CC=C(C[C@H](NC([C@H](CO)NC4=O)=O)C(O)=O)C=C3)[C@@H](C(N[C@H]4CCCNC(N)=N)=O)NC([C@@H](NC([C@H]1NC(C5N(C)CCC5)=O)=O)CC6=CC=CC=C6)=O)C=C2
QLFYPKY	9	9	5	267.9	10	951.099	1.855	CC(C)[C@@H]1OC2=CC=C(C[C@H](OC3=CC=C(C[C@H](NC([C@H](CCCCN)NC4=O)=O)C(O)=O)C=C3)[C@@H](C(N5[C@H]4CCC5)=O)NC([C@@H](NC([C@H]1NC([C@H]6N(C)CCC6=O)=O)=O)CC7=CC=CC=C7)=O)C=C2
QLFYPRY	9	10	6	303.78	11	979.112	1.265	CC(C)[C@@H]1OC2=CC=C(C[C@H](OC3=CC=C(C[C@H](NC([C@H](CCCNC(N)=N)NC4=O)=O)C(O)=O)C=C3)[C@@H](C(N5[C@H]4CCC5)=O)NC([C@@H](NC([C@H]1NC([C@H]6N(C)CCC6=O)=O)=O)CC7=CC=CC=C7)=O)C=C2
QLLYPNY	9	10	6	284.97	8	903.01	0.332	CC(C)[C@@H]1OC2=CC=C(C[C@H](OC3=CC=C(C[C@H](NC([C@H](CC(N)=O)NC4=O)=O)C(O)=O)C=C3)[C@@H](C(N5[C@H]4CCC5)=O)NC([C@@H](NC([C@H]1NC([C@H]6N(C)CCC6=O)=O)=O)CC(C)C)=O)C=C2
VLFYPSY	9	9	6	245.04	9	912.062	2.649	CC(C)[C@@H]1OC2=CC=C(C[C@H](OC3=CC=C(C[C@H](NC([C@H](CO)NC4=O)=O)C(O)=O)C=C3)[C@@H](C(N5[C@H]4CCC5)=O)NC([C@@H](NC([C@H]1NC([C@@H](N(C)C)C(C)C)=O)=O)CC6=CC=CC=C6)=O)C=C2
VLFYRSY	9	10	8	315.73	14	971.133	1.661	CC(C)[C@@H]1OC2=CC=C(C[C@H](OC3=CC=C(C[C@H](NC([C@H](CO)NC4=O)=O)C(O)=O)C=C3)[C@@H](C(N[C@H]4CCCNC(N)=N)=O)NC([C@@H](NC([C@H]1NC([C@@H](N(C)C)C(C)C)=O)=O)CC5=CC=CC=C5)=O)C=C2
FLFY	5	4	3	137.07	9	614.748	4.234	CC(C)[C@@H]1OC2=CC=C(C[C@@H](C(O)=O)NC([C@@H](NC([C@H]1NC([C@@H](N(C)C)CC3=CC=CC=C3)=O)=O)CC4=CC=CC=C4)=O)C=C2
FLLY	5	4	3	137.07	9	580.731	3.91	CC(C)[C@@H]1OC2=CC=C(C[C@@H](C(O)=O)NC([C@@H](NC([C@H]1NC([C@@H](N(C)C)CC3=CC=CC=C3)=O)=O)CC(C)C)=O)C=C2
ILLY	6	4	3	137.07	9	546.713	3.653	CC(C)[C@@H]1OC2=CC=C(C[C@@H](C(O)=O)NC([C@@H](NC([C@H]1NC([C@@H](N(C)C)C(C)CC)=O)=O)CC(C)C)=O)C=C2
LLSY	5	5	4	157.3	8	520.631	1.466	CC(C)[C@@H]1OC2=CC=C(C[C@@H](C(O)=O)NC([C@@H](NC([C@H]1NC([C@@H](N(C)C)CC(C)C)=O)=O)CO)=O)C=C2
PLFY	5	4	3	137.07	6	550.661	2.79	CC(C)[C@@H]1OC2=CC=C(C[C@@H](C(O)=O)NC([C@@H](NC([C@H]1NC(C3N(C)CCC3)=O)=O)CC4=CC=CC=C4)=O)C=C2
QLFY	5	5	3	154.14	6	564.644	1.919	CC(C)[C@@H]1OC2=CC=C(C[C@@H](C(O)=O)NC([C@@H](NC([C@H]1NC([C@H]3N(C)CCC3=O)=O)=O)CC4=CC=CC=C4)=O)C=C2
QLLY	5	5	3	154.14	6	530.627	1.595	CC(C)[C@@H]1OC2=CC=C(C[C@@H](C(O)=O)NC([C@@H](NC([C@H]1NC([C@H]3N(C)CCC3=O)=O)=O)CC(C)C)=O)C=C2
VLFY	5	4	3	137.07	8	566.704	3.521	CC(C)[C@@H]1OC2=CC=C(C[C@@H](C(O)=O)NC([C@@H](NC([C@H]1NC([C@@H](N(C)C)C(C)C)=O)=O)CC3=CC=CC=C3)=O)C=C2
APASY	6	6	4	200.39	3	505.532	-2.295	C[C@@H](C(N(CC1)[C@H](C(N[C@@H](C)C(N[C@H]2CO)=O)=O)[C@H]1OC(C=C3)=CC=C3C[C@@H](C(O)=O)NC2=O)=O)N
APFGY	5	5	3	180.16	4	551.605	-0.335	C[C@@H](C(N(CC1)[C@H](C(N[C@@H](CC2=CC=CC=C2)C(NC3)=O)=O)[C@H]1OC(C=C4)=CC=C4C[C@@H](C(O)=O)NC3=O)=O)N
APFLY	6	5	3	180.16	6	607.713	1.381	C[C@@H](C(N(CC1)[C@H](C(N[C@@H](CC2=CC=CC=C2)C(N[C@H]3CC(C)C)=O)=O)[C@H]1OC(C=C4)=CC=C4C[C@@H](C(O)=O)NC3=O)=O)N
DPQIY	7	6	4	260.55	9	632.676	-1.709	N[C@H](C(N(CC1)[C@H](C(N[C@@H](CCC(N)=O)C(N[C@H]2[C@H](CC)C)=O)=O)[C@H]1OC(C=C3)=CC=C3C[C@@H](C(O)=O)NC2=O)=O)CC(O)=O
EPFPY	6	5	2	208.67	7	649.707	0.144	N[C@H](C(N(CC1)[C@H](C(N[C@@H](CC2=CC=CC=C2)C(N3[C@H]4CCC3)=O)=O)[C@H]1OC(C=C5)=CC=C5C[C@@H](C(O)=O)NC4=O)=O)CCC(O)=O
EPYPY	6	6	3	228.9	7	665.706	-0.123	N[C@H](C(N(CC1)[C@H](C(N[C@@H](CC2=CC=C(O)C=C2)C(N3[C@H]4CCC3)=O)=O)[C@H]1OC(C=C5)=CC=C5C[C@@H](C(O)=O)NC4=O)=O)CCC(O)=O
FPDIY	7	5	3	217.46	8	651.723	0.804	N[C@H](C(N(CC1)[C@H](C(N[C@@H](CC(O)=O)C(N[C@H]2[C@H](CC)C)=O)=O)[C@H]1OC(C=C3)=CC=C3C[C@@H](C(O)=O)NC2=O)=O)CC4=CC=CC=C4
FPFPY	6	5	2	171.37	6	667.769	2.023	N[C@H](C(N(CC1)[C@H](C(N[C@@H](CC2=CC=CC=C2)C(N3[C@H]4CCC3)=O)=O)[C@H]1OC(C=C5)=CC=C5C[C@@H](C(O)=O)NC4=O)=O)CC6=CC=CC=C6
FPGGY	4	5	3	180.16	4	537.577	-0.819	N[C@H](C(N(CC1)[C@H](C(NCC(NC2)=O)=O)[C@H]1OC(C=C3)=CC=C3C[C@@H](C(O)=O)NC2=O)=O)CC4=CC=CC=C4
FPGVY	5	5	3	180.16	5	579.659	0.508	N[C@H](C(N(CC1)[C@H](C(NCC(N[C@H]2C(C)C)=O)=O)[C@H]1OC(C=C3)=CC=C3C[C@@H](C(O)=O)NC2=O)=O)CC4=CC=CC=C4
FPLFY	6	5	3	180.16	8	683.812	2.937	N[C@H](C(N(CC1)[C@H](C(N[C@@H](CC(C)C)C(N[C@H]2CC3=CC=CC=C3)=O)=O)[C@H]1OC(C=C4)=CC=C4C[C@@H](C(O)=O)NC2=O)=O)CC5=CC=CC=C5
FPLIY	7	5	3	180.16	8	649.794	2.68	N[C@H](C(N(CC1)[C@H](C(N[C@@H](CC(C)C)C(N[C@H]2[C@H](CC)C)=O)=O)[C@H]1OC(C=C3)=CC=C3C[C@@H](C(O)=O)NC2=O)=O)CC4=CC=CC=C4
FPLLY	6	5	3	180.16	8	649.794	2.613	N[C@H](C(N(CC1)[C@H](C(N[C@@H](CC(C)C)C(N[C@H]2CC(C)C)=O)=O)[C@H]1OC(C=C3)=CC=C3C[C@@H](C(O)=O)NC2=O)=O)CC4=CC=CC=C4
FPLVY	6	5	3	180.16	7	635.767	2.224	N[C@H](C(N(CC1)[C@H](C(N[C@@H](CC(C)C)C(N[C@H]2C(C)C)=O)=O)[C@H]1OC(C=C3)=CC=C3C[C@@H](C(O)=O)NC2=O)=O)CC4=CC=CC=C4
FPNTY	7	7	5	243.48	7	638.683	-1.639	N[C@H](C(N(CC1)[C@H](C(N[C@@H](CC(N)=O)C(N[C@H]2C(C)O)=O)=O)[C@H]1OC(C=C3)=CC=C3C[C@@H](C(O)=O)NC2=O)=O)CC4=CC=CC=C4
FPSVY	6	6	4	200.39	6	609.685	0.104	N[C@H](C(N(CC1)[C@H](C(N[C@@H](CO)C(N[C@H]2C(C)C)=O)=O)[C@H]1OC(C=C3)=CC=C3C[C@@H](C(O)=O)NC2=O)=O)CC4=CC=CC=C4
FPVLY	6	5	3	180.16	7	635.767	2.224	N[C@H](C(N(CC1)[C@H](C(N[C@@H](C(C)C)C(N[C@H]2CC(C)C)=O)=O)[C@H]1OC(C=C3)=CC=C3C[C@@H](C(O)=O)NC2=O)=O)CC4=CC=CC=C4
FPYIY	7	6	4	200.39	8	699.811	2.737	N[C@H](C(N(CC1)[C@H](C(N[C@@H](CC2=CC=C(O)C=C2)C(N[C@H]3[C@@H](C)CC)=O)=O)[C@H]1OC(C=C4)=CC=C4C[C@@H](C(O)=O)NC3=O)=O)CC5=CC=CC=C5

FPYNY	6	7	5	243.48	8	700.755	0.16	N[C@H](C(N(CC1)[C@H](C(N[C@@H](CC2=CC=C(O)C=C2)C(N[C@H]3CC(N)=O)=O)=O)[C@H]1OC(C=C4)=CC=C4C[C@@H](C(O)=O)NC3=O)=O)CC5=CC=CC=C5
FPYTY	7	7	5	220.62	7	687.756	0.927	N[C@H](C(N(CC1)[C@H](C(N[C@@H](CC2=CC=C(O)C=C2)C(N[C@H]3C(C)O)=O)=O)[C@H]1OC(C=C4)=CC=C4C[C@@H](C(O)=O)NC3=O)=O)CC5=CC=CC=C5
GPGLY	4	5	3	180.16	4	503.56	-1.143	NCC(N(CC1)[C@H](C(NCC(N[C@H]2CC(C)C)=O)=O)[C@H]1OC(C=C3)=CC=C3C[C@@H](C(O)=O)NC2=O)=O
IPFFY	7	5	3	180.16	8	683.812	3.005	NC(C(N(CC1)[C@H](C(N[C@@H](CC2=CC=CC=C2)C(N[C@H]3CC4=CC=CC=C4)=O)=O)[C@H]1OC(C=C5)=CC=C5C[C@@H](C(O)=O)NC3=O)=O)C(C)CC
IPFSY	7	6	4	200.39	7	623.712	0.56	NC(C(N(CC1)[C@H](C(N[C@@H](CC2=CC=CC=C2)C(N[C@H]3CO)=O)=O)[C@H]1OC(C=C4)=CC=C4C[C@@H](C(O)=O)NC3=O)=O)C(C)CC
IPHHY	7	7	5	237.52	8	663.74	-1.878	NC(C(N(CC1)[C@H](C(N[C@@H](CC2=CNC=N2)C(N[C@H]3CC4=CNC=N4)=O)=O)[C@H]1OC(C=C5)=CC=C5C[C@@H](C(O)=O)NC3=O)=O)C(C)CC
IPMSY	7	7	4	225.69	8	607.732	-0.668	NC(C(N(CC1)[C@H](C(N[C@@H](CCSC)C(N[C@H]2CO)=O)=O)[C@H]1OC(C=C3)=CC=C3C[C@@H](C(O)=O)NC2=O)=O)C(C)CC
LPFFY	6	5	3	180.16	8	683.812	2.937	N[C@H](C(N(CC1)[C@H](C(N[C@@H](CC2=CC=CC=C2)C(N[C@H]3CC4=CC=CC=C4)=O)=O)[C@H]1OC(C=C5)=CC=C5C[C@@H](C(O)=O)NC3=O)=O)CC(C)C
LPFSY	6	6	4	200.39	7	623.712	0.493	N[C@H](C(N(CC1)[C@H](C(N[C@@H](CC2=CC=CC=C2)C(N[C@H]3CO)=O)=O)[C@H]1OC(C=C4)=CC=C4C[C@@H](C(O)=O)NC3=O)=O)CC(C)C
LPIIY	8	5	3	180.16	8	615.777	2.424	N[C@H](C(N(CC1)[C@H](C(N[C@@H]([C@@H](C)CC)C(N[C@H]2[C@H](CC)C)=O)=O)[C@H]1OC(C=C3)=CC=C3C[C@@H](C(O)=O)NC2=O)=O)CC(C)C
LPQIY	7	6	4	223.25	9	630.748	0.168	N[C@H](C(N(CC1)[C@H](C(N[C@@H](CCC(N)=O)C(N[C@H]2C(CC)C)=O)=O)[C@H]1OC(C=C3)=CC=C3C[C@@H](C(O)=O)NC2=O)=O)CC(C)C
LPSAY	6	6	4	200.39	5	547.613	-1.063	N[C@H](C(N(CC1)[C@H](C(N[C@@H](CO)C(N[C@H]2C)=O)=O)[C@H]1OC(C=C3)=CC=C3C[C@@H](C(O)=O)NC2=O)=O)CC(C)C
LPSGY	5	6	4	200.39	5	533.586	-1.547	N[C@H](C(N(CC1)[C@H](C(N[C@@H](CO)C(NC2)=O)=O)[C@H]1OC(C=C3)=CC=C3C[C@@H](C(O)=O)NC2=O)=O)CC(C)C
LPYTY	7	7	5	220.62	7	653.739	0.603	N[C@H](C(N(CC1)[C@H](C(N[C@@H](CC2=CC=C(O)C=C2)C(N[C@H]3C(C)O)=O)=O)[C@H]1OC(C=C4)=CC=C4C[C@@H](C(O)=O)NC3=O)=O)CC(C)C
MPKKY	6	6	3	257.5	13	663.843	-1.237	N[C@H](C(N(CC1)[C@H](C(N[C@@H](CCCCN)C(N[C@H]2CCCCN)=O)=O)[C@H]1OC(C=C3)=CC=C3C[C@@H](C(O)=O)NC2=O)=O)CCSC
PPEDY	6	5	3	240.77	7	617.618	-1.83	O=C(O)[C@H]1NC([C@H](CC(O)=O)NC([C@H](CCC(O)=O)NC([C@H]([C@H]2OC(C=C3)=CC=C3C1)N(CC2)C([C@@H]4CCCN4)=O)=O)=O)=O
PPELY	6	5	3	203.47	7	615.689	0.046	O=C(O)[C@H]1NC([C@H](CC(C)C)NC([C@H](CCC(O)=O)NC([C@H]([C@H]2OC(C=C3)=CC=C3C1)N(CC2)C([C@@H]4CCCN4)=O)=O)=O)=O
PPFDY	6	5	3	203.47	6	635.68	0.049	O=C(O)[C@H]1NC([C@H](CC(O)=O)NC([C@H](CC2=CC=CC=C2)NC([C@H]([C@H]3OC(C=C4)=CC=C4C1)N(CC3)C([C@@H]5CCCN5)=O)=O)=O)=O
PPFNY	6	6	4	209.26	6	634.695	-0.585	O=C(O)[C@H]1NC([C@H](CC(N)=O)NC([C@H](CC2=CC=CC=C2)NC([C@H]([C@H]3OC(C=C4)=CC=C4C1)N(CC3)C([C@@H]5CCCN5)=O)=O)=O)=O
PPYTY	7	7	5	206.63	5	637.696	-0.085	O=C(O)[C@H]1NC([C@H](C(C)O)NC([C@H](CC2=CC=C(O)C=C2)NC([C@H]([C@H]3OC(C=C4)=CC=C4C1)N(CC3)C([C@@H]5CCCN5)=O)=O)=O)=O
SPFGY	5	6	4	200.39	5	567.604	-1.223	N[C@H](C(N(CC1)[C@H](C(N[C@@H](CC2=CC=CC=C2)C(NC3)=O)=O)[C@H]1OC(C=C4)=CC=C4C[C@@H](C(O)=O)NC3=O)=O)CO
SPFMY	6	7	4	225.69	8	641.749	-0.411	N[C@H](C(N(CC1)[C@H](C(N[C@@H](CC2=CC=CC=C2)C(N[C@H]3CCSC)=O)=O)[C@H]1OC(C=C4)=CC=C4C[C@@H](C(O)=O)NC3=O)=O)CO
SPFPY	6	6	3	191.6	5	607.669	-0.421	N[C@H](C(N(CC1)[C@H](C(N[C@@H](CC2=CC=CC=C2)C(N3[C@H]4CCC3)=O)=O)[C@H]1OC(C=C5)=CC=C5C[C@@H](C(O)=O)NC4=O)=O)CO
SPFSY	6	7	5	220.62	6	597.63	-1.627	N[C@H](C(N(CC1)[C@H](C(N[C@@H](CC2=CC=CC=C2)C(N[C@H]3CO)=O)=O)[C@H]1OC(C=C4)=CC=C4C[C@@H](C(O)=O)NC3=O)=O)CO
SPQAY	6	7	5	243.48	6	562.585	-3.252	N[C@H](C(N(CC1)[C@H](C(N[C@@H](CCC(N)=O)C(N[C@H]2C)=O)=O)[C@H]1OC(C=C3)=CC=C3C[C@@H](C(O)=O)NC2=O)=O)CO
TPEDY	7	6	4	274.99	8	621.606	-2.885	NC(C(N(CC1)[C@H](C(N[C@@H](CCC(O)=O)C(N[C@H]2CC(O)=O)=O)=O)[C@H]1OC(C=C3)=CC=C3C[C@@H](C(O)=O)NC2=O)=O)C(C)O
TPFGY	6	6	4	200.39	5	581.631	-0.845	NC(C(N(CC1)[C@H](C(N[C@@H](CC2=CC=CC=C2)C(NC3)=O)=O)[C@H]1OC(C=C4)=CC=C4C[C@@H](C(O)=O)NC3=O)=O)C(C)O
TPYGY	6	7	5	220.62	5	597.63	-1.113	NC(C(N(CC1)[C@H](C(N[C@@H](CC2=CC=C(O)C=C2)C(NC3)=O)=O)[C@H]1OC(C=C4)=CC=C4C[C@@H](C(O)=O)NC3=O)=O)C(C)O
VPAFY	6	5	3	180.16	5	593.686	0.992	N[C@H](C(N(CC1)[C@H](C(N[C@@H](C)C(N[C@H]2CC3=CC=CC=C3)=O)=O)[C@H]1OC(C=C4)=CC=C4C[C@@H](C(O)=O)NC2=O)=O)C(C)C
VPFFY	6	5	3	180.16	7	669.785	2.548	N[C@H](C(N(CC1)[C@H](C(N[C@@H](CC2=CC=CC=C2)C(N[C@H]3CC4=CC=CC=C4)=O)=O)[C@H]1OC(C=C5)=CC=C5C[C@@H](C(O)=O)NC3=O)=O)C(C)C
VPFLY	6	5	3	180.16	7	635.767	2.224	N[C@H](C(N(CC1)[C@H](C(N[C@@H](CC2=CC=CC=C2)C(N[C@H]3CC(C)C)=O)=O)[C@H]1OC(C=C4)=CC=C4C[C@@H](C(O)=O)NC3=O)=O)C(C)C
WPFVY	6	5	4	195.95	7	708.822	2.468	N[C@H](C(N(CC1)[C@H](C(N[C@@H](CC2=CC=CC=C2)C(N[C@H]3C(C)C)=O)=O)[C@H]1OC(C=C4)=CC=C4C[C@@H](C(O)=O)NC3=O)=O)CC5=CC(C=CC=C6)=C6N5
WPYYY	6	7	6	236.41	8	788.865	2.646	N[C@H](C(N(CC1)[C@H](C(N[C@@H](CC2=CC=C(O)C=C2)C(N[C@H]3CC4=CC=C(O)C=C4)=O)=O)[C@H]1OC(C=C5)=CC=C5C[C@@H](C(O)=O)NC3=O)=O)CC6=CC(C=CC=C7)=C7N6
YPFFY	6	6	4	200.39	8	733.829	2.994	N[C@H](C(N(CC1)[C@H](C(N[C@@H](CC2=CC=CC=C2)C(N[C@H]3CC4=CC=CC=C4)=O)=O)[C@H]1OC(C=C5)=CC=C5C[C@@H](C(O)=O)NC3=O)=O)CC6=CC=C(O)C=C6
YPILY	7	6	4	200.39	8	665.794	2.413	N[C@H](C(N(CC1)[C@H](C(N[C@@H]([C@@H](C)CC)C(N[C@H]2CC(C)C)=O)=O)[C@H]1OC(C=C3)=CC=C3C[C@@H](C(O)=O)NC2=O)=O)CC4=CC=C(O)C=C4
QAAY	5	5	4	162.93	1	432.437	-1.161	C[C@H](NC([C@@H](N1)C2)=O)C(N[C@@H](C)C(N[C@H](C(O)=O)CC3=CC=C(C=C3)O[C@H]2C1=O)=O)=O
QADY	5	5	4	200.23	3	476.447	-1.805	C[C@H](NC([C@@H](N1)C2)=O)C(N[C@@H](CC(O)=O)C(N[C@H](C(O)=O)CC3=CC=C(C=C3)O[C@H]2C1=O)=O)=O
QAFY	5	5	4	162.93	3	508.536	0.395	C[C@H](NC([C@@H](N1)C2)=O)C(N[C@@H](CC3=CC=CC=C3)C(N[C@H](C(O)=O)CC4=CC=C(C=C4)O[C@H]2C1=O)=O)=O

QAGY	4	5	4	162.93	1	418.41	-1.645	C[C@H](NC([C@@H](N1)C2)=O)C(NCC(N[C@H](C(O)=O)CC3=CC=C(C=C3)O[C@H]2C1=O)=O)=O
QAKY	5	5	4	188.95	5	489.533	-1.24	C[C@H](NC([C@@H](N1)C2)=O)C(N[C@@H](CCCCN)C(N[C@H](C(O)=O)CC3=CC=C(C=C3)O[C@H]2C1=O)=O)=O
QDFY	5	5	4	200.23	5	552.546	-0.249	O=C([C@@H](NC([C@@H](NC([C@@H](N1)C2)=O)CC(O)=O)=O)CC3=CC=CC=C3)N[C@H](C(O)=O)CC4=CC=C(C=C4)O[C@H]2C1=O
QEFY	5	5	4	200.23	6	566.573	0.072	O=C([C@@H](NC([C@@H](NC([C@@H](N1)C2)=O)CCC(O)=O)=O)CC3=CC=CC=C3)N[C@H](C(O)=O)CC4=CC=C(C=C4)O[C@H]2C1=O
QFLY	5	5	4	162.93	5	550.617	1.627	O=C([C@H](CC(C)C)NC([C@@H](NC([C@@H](N1)C2)=O)CC3=CC=CC=C3)=O)N[C@H](C(O)=O)CC4=CC=C(C=C4)O[C@H]2C1=O
QGPY	4	5	3	154.14	1	444.448	-1.327	O=C(C(CCC1)N1C(CNC([C@@H](N2)C3)=O)=O)N[C@H](C(O)=O)CC4=CC=C(C=C4)O[C@H]3C2=O
QGTY	5	6	5	183.16	2	448.436	-2.156	O=C(C(NC(CNC([C@@H](N1)C2)=O)=O)C(C)O)N[C@H](C(O)=O)CC3=CC=C(C=C3)O[C@H]2C1=O
QHFY	5	6	5	191.61	5	574.598	-0.49	O=C([C@@H](NC([C@@H](NC([C@@H](N1)C2)=O)CC3=CNC=N3)=O)CC4=CC=CC=C4)N[C@H](C(O)=O)CC5=CC=C(C=C5)O[C@H]2C1=O
QIFY	6	5	4	162.93	5	550.617	1.694	O=C([C@@H](NC(C(NC([C@@H](N1)C2)=O)C(C)CC)=O)CC3=CC=CC=C3)N[C@H](C(O)=O)CC4=CC=C(C=C4)O[C@H]2C1=O
QIVY	6	5	4	162.93	4	502.572	0.981	O=C([C@@H](NC(C(NC([C@@H](N1)C2)=O)C(C)CC)=O)C(C)C)N[C@H](C(O)=O)CC3=CC=C(C=C3)O[C@H]2C1=O
QKAY	5	5	4	188.95	5	489.533	-1.24	O=C([C@@H](NC([C@@H](NC([C@@H](N1)C2)=O)CCCCN)=O)C)N[C@H](C(O)=O)CC3=CC=C(C=C3)O[C@H]2C1=O
QKGY	4	5	4	188.95	5	475.506	-1.724	O=C(CNC([C@@H](NC([C@@H](N1)C2)=O)CCCCN)=O)N[C@H](C(O)=O)CC3=CC=C(C=C3)O[C@H]2C1=O
QKHY	5	6	5	217.63	7	555.596	-2.125	O=C([C@@H](NC([C@@H](NC([C@@H](N1)C2)=O)CCCCN)=O)CC3=CNC=N3)N[C@H](C(O)=O)CC4=CC=C(C=C4)O[C@H]2C1=O
QKKY	5	5	4	214.97	9	546.629	-1.319	O=C([C@@H](NC([C@@H](NC([C@@H](N1)C2)=O)CCCCN)=O)CCCCN)N[C@H](C(O)=O)CC3=CC=C(C=C3)O[C@H]2C1=O
QKNY	5	6	5	232.04	7	532.558	-2.518	O=C([C@@H](NC([C@@H](NC([C@@H](N1)C2)=O)CCCCN)=O)CC(N)=O)N[C@H](C(O)=O)CC3=CC=C(C=C3)O[C@H]2C1=O
QKQY	5	6	5	232.04	8	546.585	-2.197	O=C([C@@H](NC([C@@H](NC([C@@H](N1)C2)=O)CCCCN)=O)CCC(N)=O)N[C@H](C(O)=O)CC3=CC=C(C=C3)O[C@H]2C1=O
QKRY	5	6	5	250.85	10	574.642	-1.909	O=C([C@@H](NC([C@@H](NC([C@@H](N1)C2)=O)CCCCN)=O)CCCNC(N)=N)N[C@H](C(O)=O)CC3=CC=C(C=C3)O[C@H]2C1=O
QKSY	5	6	5	209.18	6	505.532	-2.128	O=C([C@@H](NC([C@@H](NC([C@@H](N1)C2)=O)CCCCN)=O)CO)N[C@H](C(O)=O)CC3=CC=C(C=C3)O[C@H]2C1=O
QLFY	5	5	4	162.93	5	550.617	1.627	O=C([C@@H](NC([C@@H](NC([C@@H](N1)C2)=O)CC(C)C)=O)CC3=CC=CC=C3)N[C@H](C(O)=O)CC4=CC=C(C=C4)O[C@H]2C1=O
QLHY	5	6	5	191.61	5	540.581	-0.815	O=C([C@@H](NC([C@@H](NC([C@@H](N1)C2)=O)CC(C)C)=O)CC3=CNC=N3)N[C@H](C(O)=O)CC4=CC=C(C=C4)O[C@H]2C1=O
QLLY	5	5	4	162.93	5	516.599	1.303	O=C([C@@H](NC([C@@H](NC([C@@H](N1)C2)=O)CC(C)C)=O)CC(C)C)N[C@H](C(O)=O)CC3=CC=C(C=C3)O[C@H]2C1=O
QLNY	5	6	5	206.02	5	517.543	-1.207	O=C([C@@H](NC([C@@H](NC([C@@H](N1)C2)=O)CC(C)C)=O)CC(N)=O)N[C@H](C(O)=O)CC3=CC=C(C=C3)O[C@H]2C1=O
QLPY	5	5	3	154.14	3	500.556	0.389	O=C(C(CCC1)N1C([C@@H](NC([C@@H](N2)C3)=O)CC(C)C)=O)N[C@H](C(O)=O)CC4=CC=C(C=C4)O[C@H]3C2=O
QLSY	5	6	5	183.16	4	490.518	-0.817	O=C([C@@H](NC([C@@H](NC([C@@H](N1)C2)=O)CC(C)C)=O)CO)N[C@H](C(O)=O)CC3=CC=C(C=C3)O[C@H]2C1=O
QLYY	5	6	5	183.16	5	566.616	1.359	O=C([C@@H](NC([C@@H](NC([C@@H](N1)C2)=O)CC(C)C)=O)CC3=CC=C(O)C=C3)N[C@H](C(O)=O)CC4=CC=C(C=C4)O[C@H]2C1=O
QMFY	5	6	4	188.23	6	568.654	0.722	O=C([C@@H](NC([C@@H](NC([C@@H](N1)C2)=O)CCSC)=O)CC3=CC=CC=C3)N[C@H](C(O)=O)CC4=CC=C(C=C4)O[C@H]2C1=O
QNFY	5	6	5	206.02	5	551.561	-0.883	O=C([C@@H](NC([C@@H](NC([C@@H](N1)C2)=O)CC(N)=O)=O)CC3=CC=CC=C3)N[C@H](C(O)=O)CC4=CC=C(C=C4)O[C@H]2C1=O
QNRY	5	7	6	267.92	8	560.571	-3.108	O=C([C@@H](NC([C@@H](NC([C@@H](N1)C2)=O)CC(N)=O)=O)CCCNC(N)=N)N[C@H](C(O)=O)CC3=CC=C(C=C3)O[C@H]2C1=O
QPDY	5	5	3	191.44	3	502.485	-1.487	O=C([C@@H](NC(C(CCC1)N1C([C@@H](N2)C3)=O)=O)CC(O)=O)N[C@H](C(O)=O)CC4=CC=C(C=C4)O[C@H]3C2=O
QPGY	4	5	3	154.14	1	444.448	-1.327	O=C(CNC(C(CCC1)N1C([C@@H](N2)C3)=O)=O)N[C@H](C(O)=O)CC4=CC=C(C=C4)O[C@H]3C2=O
QPKY	5	5	3	180.16	5	515.571	-0.922	O=C([C@@H](NC(C(CCC1)N1C([C@@H](N2)C3)=O)=O)CCCCN)N[C@H](C(O)=O)CC4=CC=C(C=C4)O[C@H]3C2=O
QPNY	5	6	4	197.23	3	501.5	-2.121	O=C([C@@H](NC(C(CCC1)N1C([C@@H](N2)C3)=O)=O)CC(N)=O)N[C@H](C(O)=O)CC4=CC=C(C=C4)O[C@H]3C2=O
QPSY	5	6	4	174.37	2	474.475	-1.731	O=C([C@@H](NC(C(CCC1)N1C([C@@H](N2)C3)=O)=O)CO)N[C@H](C(O)=O)CC4=CC=C(C=C4)O[C@H]3C2=O
QQFY	5	6	5	206.02	6	565.588	-0.562	O=C([C@@H](NC([C@@H](NC([C@@H](N1)C2)=O)CCC(N)=O)=O)CC3=CC=CC=C3)N[C@H](C(O)=O)CC4=CC=C(C=C4)O[C@H]2C1=O
QQGY	4	6	5	206.02	4	475.462	-2.602	O=C(CNC([C@@H](NC([C@@H](N1)C2)=O)CCC(N)=O)=O)N[C@H](C(O)=O)CC3=CC=C(C=C3)O[C@H]2C1=O
QQSY	5	7	6	226.25	5	505.489	-3.006	O=C([C@@H](NC([C@@H](NC([C@@H](N1)C2)=O)CCC(N)=O)=O)CO)N[C@H](C(O)=O)CC3=CC=C(C=C3)O[C@H]2C1=O
QQTY	6	7	6	226.25	5	519.516	-2.628	O=C(C(NC([C@@H](NC([C@@H](N1)C2)=O)CCC(N)=O)=O)C(C)O)N[C@H](C(O)=O)CC3=CC=C(C=C3)O[C@H]2C1=O
QRFY	5	6	5	224.83	8	593.645	-0.274	O=C([C@@H](NC([C@@H](NC([C@@H](N1)C2)=O)CCCNC(N)=N)=O)CC3=CC=CC=C3)N[C@H](C(O)=O)CC4=CC=C(C=C4)O[C@H]2C1=O

QRNY	5	7	6	267.92	8	560.571	-3.108	O=C([C@@H](NC([C@@H](NC([C@@H](N1)C2)=O)CCCNC(N)=N)=O)CC(N)=O)N[C@H](C(O)=O)CC3=CC=C(C=C3)O[C@H]2C1=O
QRPY	5	6	4	216.04	6	543.584	-1.512	O=C(C(CCC1)N1C([C@@H](NC([C@@H](N2)C3)=O)CCCNC(N)=N)=O)N[C@H](C(O)=O)CC4=CC=C(C=C4)O[C@H]3C2=O
QRQY	5	7	6	267.92	9	574.599	-2.787	O=C([C@@H](NC([C@@H](NC([C@@H](N1)C2)=O)CCCNC(N)=N)=O)CCC(N)=O)N[C@H](C(O)=O)CC3=CC=C(C=C3)O[C@H]2C1=O
QRTY	6	7	6	245.06	7	547.573	-2.341	O=C(C(NC([C@@H](NC([C@@H](N1)C2)=O)CCCNC(N)=N)=O)C(C)O)N[C@H](C(O)=O)CC3=CC=C(C=C3)O[C@H]2C1=O
QSFY	5	6	5	183.16	4	524.535	-0.493	O=C([C@@H](NC([C@@H](NC([C@@H](N1)C2)=O)CO)=O)CC3=CC=CC=C3)N[C@H](C(O)=O)CC4=CC=C(C=C4)O[C@H]2C1=O
QSGY	4	6	5	183.16	2	434.409	-2.533	O=C(CNC([C@@H](NC([C@@H](N1)C2)=O)CO)=O)N[C@H](C(O)=O)CC3=CC=C(C=C3)O[C@H]2C1=O
QSLY	5	6	5	183.16	4	490.518	-0.817	O=C([C@@H](NC([C@@H](NC([C@@H](N1)C2)=O)CO)=O)CC(C)C)N[C@H](C(O)=O)CC3=CC=C(C=C3)O[C@H]2C1=O
QSRY	5	7	6	245.06	7	533.546	-2.718	O=C([C@@H](NC([C@@H](NC([C@@H](N1)C2)=O)CO)=O)CCCNC(N)=N)N[C@H](C(O)=O)CC3=CC=C(C=C3)O[C@H]2C1=O
QTDY	6	6	5	220.46	4	506.473	-2.316	O=C([C@@H](NC(C(NC([C@@H](N1)C2)=O)C(C)O)=O)CC(O)=O)N[C@H](C(O)=O)CC3=CC=C(C=C3)O[C@H]2C1=O
QTFY	6	6	5	183.16	4	538.562	-0.116	O=C([C@@H](NC(C(NC([C@@H](N1)C2)=O)C(C)O)=O)CC3=CC=CC=C3)N[C@H](C(O)=O)CC4=CC=C(C=C4)O[C@H]2C1=O
QTNY	6	7	6	226.25	4	505.489	-2.95	O=C([C@@H](NC(C(NC([C@@H](N1)C2)=O)C(C)O)=O)CC(N)=O)N[C@H](C(O)=O)CC3=CC=C(C=C3)O[C@H]2C1=O
QTRY	6	7	6	245.06	7	547.573	-2.341	O=C([C@@H](NC(C(NC([C@@H](N1)C2)=O)C(C)O)=O)CCCNC(N)=N)N[C@H](C(O)=O)CC3=CC=C(C=C3)O[C@H]2C1=O
QVVY	5	5	4	162.93	3	488.545	0.525	O=C([C@@H](NC([C@@H](NC([C@@H](N1)C2)=O)C(C)C)=O)C(C)C)N[C@H](C(O)=O)CC3=CC=C(C=C3)O[C@H]2C1=O
QWIY	6	5	5	178.72	5	589.654	1.614	O=C(C(NC([C@@H](NC([C@@H](N1)C2)=O)CC3=CC(C=CC=C4)=C4N3)=O)C(C)CC)N[C@H](C(O)=O)CC5=CC=C(C=C5)O[C@H]2C1=O
QWPY	5	5	4	169.93	3	573.611	0.633	O=C(C(CCC1)N1C([C@@H](NC([C@@H](N2)C3)=O)CC4=CC(C=CC=C5)=C5N4)=O)N[C@H](C(O)=O)CC6=CC=C(C=C6)O[C@H]3C2=O
QYPY	5	6	4	174.37	3	550.573	0.446	O=C(C(CCC1)N1C([C@@H](NC([C@@H](N2)C3)=O)CC4=CC=C(O)C=C4)=O)N[C@H](C(O)=O)CC5=CC=C(C=C5)O[C@H]3C2=O
QYSY	5	7	6	203.39	4	540.534	-0.761	O=C([C@@H](NC([C@@H](NC([C@@H](N1)C2)=O)CC3=CC=C(O)C=C3)=O)CO)N[C@H](C(O)=O)CC4=CC=C(C=C4)O[C@H]2C1=O
QYWY	5	6	6	198.95	5	639.671	1.603	O=C([C@@H](NC([C@@H](NC([C@@H](N1)C2)=O)CC3=CC=C(O)C=C3)=O)CC4=CC(C=CC=C5)=C5N4)N[C@H](C(O)=O)CC6=CC=C(C=C6)O[C@H]2C1=O
IPPLSFY	9	8	5	249.8	10	833.991	1.198	O=C(O)[C@@H]1CC2=CC=C(O[C@@H]3[C@@H](C(N[C@@H](CC(C)C)C(N[C@@H](CO)C(NC(C(N1)=O)CC4=CC=CC=C4)=O)=O)=O)N(CC3)C([C@@H]5CCCN5C([C@H](C(CC)C)N)=O)=O)C=C2
SPPPPVY	8	8	3	232.22	5	753.86	-1.409	O=C(O)[C@@H]1CC2=CC=C(O[C@@H]3C(C(N4[C@@H](CCC4)C(N5[C@@H](CCC5)C(NC(C(N1)=O)C(C)C)=O)=O)=O)N(CC3)C([C@@H]6CCCN6C([C@H](CO)N)=O)=O)C=C2
GPPGQFY	6	8	5	272.66	8	762.827	-2.37	O=C(O)[C@@H]1CC2=CC=C(O[C@@H]3[C@@H](C(NCC(N[C@@H](CCC(N)=O)C(NC(C(N1)=O)CC4=CC=CC=C4)=O)=O)=O)N(CC3)C([C@@H]5CCCN5C(CN)=O)=O)C=C2
VPPLFVY	8	7	4	229.57	9	832.018	2.473	O=C(O)[C@@H]1CC2=CC=C(O[C@@H]3[C@@H](C(N[C@@H](CC(C)C)C(N[C@@H](CC4=CC=CC=C4)C(NC(C(N1)=O)C(C)C)=O)=O)=O)N(CC3)C([C@@H]5CCCN5C([C@H](C(C)C)N)=O)=O)C=C2
VPPFFTY	9	8	5	249.8	9	868.008	1.444	O=C(O)[C@@H]1CC2=CC=C(O[C@@H]3[C@@H](C(N[C@@H](CC4=CC=CC=C4)C(N[C@@H](CC5=CC=CC=C5)C(NC(C(N1)=O)[C@@H](C)O)=O)=O)=O)N(CC3)C([C@@H]6CCCN6C([C@H](C(C)C)N)=O)=O)C=C2
QLLVWRSW	9	10	12	383.81	12	1066.237	-0.558	O=C1CC[C@@H](C(N[C@H](C(N[C@@H](CC(C)C)C(N[C@@H](C(C)C(C=C2)=CC3=C2C4=CN3)C)C(N[C@H](C(NC(CCCNC(N)=N)C(N[C@@H](C(N[C@H](C(O)=O)C4)=O)CO)=O)=O)CC5=CNC6=C5C=CC7=C6)=O)=O)=O)[C@@H]7C(C)C)=O)N1
QPDKSSYW	8	11	10	364.23	12	919.91	-3.205	O=C1N[C@H](C(N2[C@H](C(N[C@@H](CC(O)=O)C(N[C@H]3C(N[C@H](C(N[C@@H](CO)C(N[C@@H](CC4=CC=C(O)C=C4)C(N[C@@H](C(O)=O)CC5=CN3C6=C5C=CC=C6)=O)=O)=O)CO)=O)=O)=O)CCC2)=O)CC1
QPFGFWAW	8	8	8	282.03	12	1019.137	2.373	O=C1N[C@H](C(N2[C@H](C(N[C@@H](CC3=CC=CC=C3)C(N[C@H]4C(N[C@H](C(N[C@@H](CC5=CNC6=C5C=CC=C6)C(N[C@@H](C)C(N[C@@H](C(O)=O)CC7=CN4C8=C7C=CC=C8)=O)=O)=O)CC9=CC=CC=C9)=O)=O)=O)CCC2)=O)CC1
QPFGFWIW	9	8	8	282.03	14	1061.218	3.672	O=C1N[C@H](C(N2[C@H](C(N[C@@H](CC3=CC=CC=C3)C(N[C@H]4C(N[C@H](C(N[C@@H](CC5=CNC6=C5C=CC=C6)C(NC(C(CC)CC)C(N[C@@H](C(O)=O)CC7=CN4C8=C7C=CC=C8)=O)=O)=O)CC9=CC=CC=C9)=O)=O)=O)CCC2)=O)CC1
QPFGLCIW	9	8	8	305.04	13	943.121	2.097	O=C1N[C@H](C(N2[C@H](C(N[C@@H](CC3=CC=CC=C3)C(N[C@H]4C(N[C@H](C(N[C@@H](CS)C(NC(C(C)CC)C(N[C@@H](C(O)=O)CC5=CN4C6=C5C=CC=C6)=O)=O)=O)CC(C)C)=O)=O)=O)CCC2)=O)CC1
GPVFY	5	5	3	180.16	5	579.659	0.508	NCC(N(CC1)[C@H](C(N[C@@H](C(C)C)C(N[C@H]2CC3=CC=CC=C3)=O)=O)[C@H]1OC(C=C4)=CC=C4C[C@@H](C(O)=O)NC2=O)=O
QPDNVVLW	8	8	7	303.54	12	894.003	0.201	O=C1N[C@H](C(N2[C@H](C(N[C@@H](CC(O)=O)C(N[C@H]3C(N[C@H](C(N[C@@H](C(C)C)C(N[C@@H](CC(C)C)C(N[C@@H](C(O)=O)CC4=CN3C5=C4C=CC=C5)=O)=O)=O)C(C)C)=O)=O)=O)CCC2)=O)CC1
QPLANVAW	8	9	8	309.33	11	864.964	-1.276	O=C1N[C@H](C(N2[C@H](C(N[C@@H](CC(C)C)C(N[C@H]3C(N[C@H](C(N[C@@H](C(C)C)C(N[C@@H](C)C(N[C@@H](C(O)=O)CC4=CN3C5=C4C=CC=C5)=O)=O)=O)CC(N)=O)=O)=O)=O)CCC2)=O)CC1
QPLTNIGW	8	9	8	309.33	12	864.964	-1.304	O=C1N[C@H](C(N2[C@H](C(N[C@@H](CC(C)C)C(N[C@H]3C(N[C@H](C(NC(C(C)CC)C(NCC(N[C@@H](C(O)=O)CC4=CN3C5=C4C=CC=C5)=O)=O)=O)CC(N)=O)=O)=O)=O)CCC2)=O)CC1
QPYGVFAW	8	9	8	286.47	11	948.054	1.615	O=C1N[C@H](C(N2[C@H](C(N[C@@H](CC3=CC=C(O)C=C3)C(N[C@H]4C(N[C@H](C(N[C@@H](CC5=CC=CC=C5)C(N[C@@H](C)C(N[C@@H](C(O)=O)CC6=CN4C7=C6C=CC=C7)=O)=O)=O)C(C)C)=O)=O)=O)CCC2)=O)CC1
QPYGVGAW	7	9	8	286.47	9	857.928	-0.425	O=C1N[C@H](C(N2[C@H](C(N[C@@H](CC3=CC=C(O)C=C3)C(N[C@H]4C(N[C@H](C(NCC(N[C@@H](C)C(N[C@@H](C(O)=O)CC5=CN4C6=C5C=CC=C6)=O)=O)=O)C(C)C)=O)=O)=O)CCC2)=O)CC1
QPYGVGSW	7	10	9	306.7	10	873.928	-1.313	O=C1N[C@H](C(N2[C@H](C(N[C@@H](CC3=CC=C(O)C=C3)C(N[C@H]4C(N[C@H](C(NCC(N[C@@H](CO)C(N[C@@H](C(O)=O)CC5=CN4C6=C5C=CC=C6)=O)=O)=O)C(C)C)=O)=O)=O)CCC2)=O)CC1
QPYGVIRW	9	10	9	348.37	16	999.146	0.689	O=C1N[C@H](C(N2[C@H](C(N[C@@H](CC3=CC=C(O)C=C3)C(N[C@H]4C(N[C@H](C(NC(C(C)CC)C(N[C@@H](CCCNC(N)=N)C(N[C@@H](C(O)=O)CC5=CN4C6=C5C=CC=C6)=O)=O)=O)C(C)C)=O)=O)=O)CCC2)=O)CC1
QPYGVYSW	8	11	10	326.93	12	980.053	0.459	O=C1N[C@H](C(N2[C@H](C(N[C@@H](CC3=CC=C(O)C=C3)C(N[C@H]4C(N[C@H](C(N[C@@H](CC5=CC=C(O)C=C5)C(N[C@@H](CO)C(N[C@@H](C(O)=O)CC6=CN4C7=C6C=CC=C7)=O)=O)=O)C(C)C)=O)=O)=O)CCC2)=O)CC1
QPYRNIYW	9	11	10	349.79	14	1021.106	0.526	O=C1N[C@H](C(N2[C@H](C(N[C@@H](CC3=CC=C(O)C=C3)C(N[C@H]4C(N[C@H](C(NC(C(C)CC)C(N[C@@H](CC5=CC=C(O)C=C5)C(N[C@@H](C(O)=O)CC6=CN4C7=C6C=CC=C7)=O)=O)=O)CC(N)=O)=O)=O)=O)CCC2)=O)CC1

QPYSVFAW	8	9	8	286.47	11	948.054	1.615	O=C1N[C@H](C(N2[C@H](C(N[C@@H](CC3=CC=C(O)C=C3)C(N[C@H]4C(N[C@H](C(N[C@@H](CC5=CC=CC=C5)C(N[C@@H](C)C(N[C@@H](C(O)=O)CC6=CN4C7=C6C=CC=C7)=O)=O)=O)C(C)C)=O)=O)=O)CCC2)=O)CC1
QPYGVYSW	9	11	9	315.93	5	978.037	0.157	O=C([C@@H]1C2)N[C@@H]2C(N3[C@@H](CCC3)C(N[C@@H](CC4=CC=C(O)C=C4)C(N[C@H]5C(N[C@@H](C(C)C)C(N[C@@H](CC6=CC=C(O1)C=C6)C(N[C@@H](CO)C(N[C@@H](CC7=CN5C8=C7C=CC=C8)C(O)=O)=O)=O)=O)=O)=O)=O)=O
FLIYQSY	10	10	8	296.92	13	957.103	1.83	CC(C)[C@@H]1OC2=CC=C([C@H](OC3=CC=C(C[C@H](NC([C@H](CO)NC4=O)=O)C(O)=O)C=C3)[C@@H](C(N[C@H]4CCC(N)=O)=O)NC(C(NC([C@H]1NC([C@@H](N(C)C)CC5=CC=CC=C5)=O)=O)C(C)CC)=O)C=C2
FLLYPSY	9	9	6	245.04	10	926.089	3.037	CC(C)[C@@H]1OC2=CC=C([C@H](OC3=CC=C(C[C@H](NC([C@H](CO)NC4=O)=O)C(O)=O)C=C3)[C@@H](C(N5[C@H]4CCC5)=O)NC([C@@H](NC([C@H]1NC([C@@H](N(C)C)CC6=CC=CC=C6)=O)=O)CC(C)C)=O)C=C2
FLVYTSY	10	10	8	274.06	10	916.05	1.82	CC(C)[C@@H]1OC2=CC=C([C@H](OC3=CC=C(C[C@H](NC([C@H](CO)NC4=O)=O)C(O)=O)C=C3)[C@@H](C(NC4C(C)O)=O)NC([C@@H](NC([C@H]1NC([C@@H](N(C)C)CC5=CC=CC=C5)=O)=O)C(C)C)=O)C=C2
ILFYSYY	10	10	8	274.06	12	992.149	4.075	CC(C)[C@@H]1OC2=CC=C([C@H](OC3=CC=C(C[C@H](NC([C@H](CC4=CC=C(O)C=C4)NC5=O)=O)C(O)=O)C=C3)[C@@H](C(N[C@H]5CO)=O)NC([C@@H](NC([C@H]1NC([C@@H](N(C)C)C(CC)C)=O)=O)CC6=CC=CC=C6)=O)C=C2
ILLYPSY	10	9	6	245.04	10	892.071	2.781	CC(C)[C@@H]1OC2=CC=C([C@H](OC3=CC=C(C[C@H](NC([C@H](CO)NC4=O)=O)C(O)=O)C=C3)[C@@H](C(N5[C@H]4CCC5)=O)NC([C@@H](NC([C@H]1NC([C@@H](N(C)C)C(CC)C)=O)=O)CC(C)C)=O)C=C2
LLFYSYY	9	10	8	274.06	12	992.149	4.008	CC(C)[C@@H]1OC2=CC=C([C@H](OC3=CC=C(C[C@H](NC([C@H](CC4=CC=C(O)C=C4)NC5=O)=O)C(O)=O)C=C3)[C@@H](C(N[C@H]5CO)=O)NC([C@@H](NC([C@H]1NC([C@@H](N(C)C)CC(C)C)=O)=O)CC6=CC=CC=C6)=O)C=C2
LLVYQSY	9	10	8	296.92	12	909.058	1.049	CC(C)[C@@H]1OC2=CC=C([C@H](OC3=CC=C(C[C@H](NC([C@H](CO)NC4=O)=O)C(O)=O)C=C3)[C@@H](C(N[C@H]4CCC(N)=O)=O)NC([C@@H](NC([C@H]1NC([C@@H](N(C)C)CC(C)C)=O)=O)C(C)C)=O)C=C2
LLYYQSY	9	11	9	317.15	13	973.102	1.495	CC(C)[C@@H]1OC2=CC=C([C@H](OC3=CC=C(C[C@H](NC([C@H](CO)NC4=O)=O)C(O)=O)C=C3)[C@@H](C(N[C@H]4CCC(N)=O)=O)NC([C@@H](NC([C@H]1NC([C@@H](N(C)C)CC(C)C)=O)=O)CC5=CC=C(O)C=C5)=O)C=C2
VLFYPSY	9	9	6	245.04	9	912.062	2.649	CC(C)[C@@H]1OC2=CC=C([C@H](OC3=CC=C(C[C@H](NC([C@H](CO)NC4=O)=O)C(O)=O)C=C3)[C@@H](C(N5[C@H]4CCC5)=O)NC([C@@H](NC([C@H]1NC([C@@H](N(C)C)C(C)C)=O)=O)CC6=CC=CC=C6)=O)C=C2
VLFYRSY	9	10	8	315.73	14	971.133	1.661	CC(C)[C@@H]1OC2=CC=C([C@H](OC3=CC=C(C[C@H](NC([C@H](CO)NC4=O)=O)C(O)=O)C=C3)[C@@H](C(N[C@H]4CCCNC(N)=N)=O)NC([C@@H](NC([C@H]1NC([C@@H](N(C)C)C(C)C)=O)=O)CC5=CC=CC=C5)=O)C=C2
FLIY	6	4	3	137.07	9	580.731	3.977	CC(C)[C@@H]1OC2=CC=C(C[C@H](C(O)=O)NC(C(NC([C@H]1NC([C@@H](N(C)C)CC3=CC=CC=C3)=O)=O)C(C)CC)=O)C=C2
FLLY	5	4	3	137.07	9	580.731	3.91	CC(C)[C@@H]1OC2=CC=C(C[C@H](C(O)=O)NC([C@@H](NC([C@H]1NC([C@@H](N(C)C)CC3=CC=CC=C3)=O)=O)CC(C)C)=O)C=C2
FLVY	5	4	3	137.07	8	566.704	3.521	CC(C)[C@@H]1OC2=CC=C(C[C@H](C(O)=O)NC([C@@H](NC([C@H]1NC([C@@H](N(C)C)CC3=CC=CC=C3)=O)=O)C(C)C)=O)C=C2
ILFY	6	4	3	137.07	9	580.731	3.977	CC(C)[C@@H]1OC2=CC=C(C[C@H](C(O)=O)NC([C@@H](NC([C@H]1NC([C@@H](N(C)C)C(C)CC)=O)=O)CC3=CC=CC=C3)=O)C=C2
ILLY	6	4	3	137.07	9	546.713	3.653	CC(C)[C@@H]1OC2=CC=C(C[C@H](C(O)=O)NC([C@@H](NC([C@H]1NC([C@@H](N(C)C)C(C)CC)=O)=O)CC(C)C)=O)C=C2
LLFY	5	4	3	137.07	9	580.731	3.91	CC(C)[C@@H]1OC2=CC=C(C[C@H](C(O)=O)NC([C@@H](NC([C@H]1NC([C@@H](N(C)C)CC(C)C)=O)=O)CC3=CC=CC=C3)=O)C=C2
LLVY	5	4	3	137.07	8	532.686	3.197	CC(C)[C@@H]1OC2=CC=C(C[C@H](C(O)=O)NC([C@@H](NC([C@H]1NC([C@@H](N(C)C)CC(C)C)=O)=O)C(C)C)=O)C=C2
LLYY	5	5	4	157.3	9	596.73	3.642	CC(C)[C@@H]1OC2=CC=C(C[C@H](C(O)=O)NC([C@@H](NC([C@H]1NC([C@@H](N(C)C)CC(C)C)=O)=O)CC3=CC=C(O)C=C3)=O)C=C2
VLFY	5	4	3	137.07	8	566.704	3.521	CC(C)[C@@H]1OC2=CC=C(C[C@H](C(O)=O)NC([C@@H](NC([C@H]1NC([C@@H](N(C)C)C(C)C)=O)=O)CC3=CC=CC=C3)=O)C=C2
FPALY	6	5	3	180.16	6	607.713	1.381	N[C@H](C(N(CC1)[C@H](C(N[C@@H](C)C(N[C@H]2CC(C)C)=O)=O)[C@H]1OC(C=C3)=CC=C3C[C@@H](C(O)=O)NC2=O)=O)CC4=CC=CC=C4
LPFFY	6	5	3	180.16	8	683.812	2.937	N[C@H](C(N(CC1)[C@H](C(N[C@@H](CC2=CC=CC=C2)C(N[C@H]3CC4=CC=CC=C4)=O)=O)[C@H]1OC(C=C5)=CC=C5C[C@@H](C(O)=O)NC3=O)=O)CC(C)C
NPLFY	6	6	4	223.25	8	650.738	0.103	N[C@H](C(N(CC1)[C@H](C(N[C@@H](CC(C)C)C(N[C@H]2CC3=CC=CC=C3)=O)=O)[C@H]1OC(C=C4)=CC=C4C[C@@H](C(O)=O)NC2=O)=O)CC(N)=O
SPFFY	6	6	4	200.39	7	657.73	0.817	N[C@H](C(N(CC1)[C@H](C(N[C@@H](CC2=CC=CC=C2)C(N[C@H]3CC4=CC=CC=C4)=O)=O)[C@H]1OC(C=C5)=CC=C5C[C@@H](C(O)=O)NC3=O)=O)CO
VPFFY	6	5	3	180.16	7	669.785	2.548	N[C@H](C(N(CC1)[C@H](C(N[C@@H](CC2=CC=CC=C2)C(N[C@H]3CC4=CC=CC=C4)=O)=O)[C@H]1OC(C=C5)=CC=C5C[C@@H](C(O)=O)NC3=O)=O)C(C)C
QQRY	5	7	6	267.92	9	574.599	-2.787	O=C([C@@H](NC([C@@H](NC([C@@H](N1)C2)=O)CCC(N)=O)=O)CCCNC(N)=N)N[C@H](C(O)=O)CC3=CC=C(C=C3)O[C@H]2C1=O
QLLVWPRH	9	10	9	356.82	12	1027.203	-0.619	O=C1CC[C@@H](C(N[C@H](C(N[C@@H](CC(C)C)C(N[C@@H](C(C)C)C(N[C@H](C(N2[C@@H](CCC2)C(N[C@@H](CCCNC(N)=N)C(N[C@@H](CC3=CN4C=N3)C(O)=O)=O)=O)=O)=O)CC5=C4NC6=C5C=CC7=C6)=O)=O)=O)[C@@H]7C(C)C)=O)N1
QLFVWGW	7	7	9	272.58	6	914.04	1.296	O=C1CC[C@@H](C(N[C@H](C(N[C@@H](CC2=CC=CC=C2)C(N[C@@H](C(C)(C(C=C3)=CC4=C3C5=CN4)C)C(N[C@H](C(NCC(N[C@@H](C5)C(O)=O)=O)=O)CC6=CNC7=C6C=CC8=C7)=O)=O)=O)[C@@H]8C(C)C)=O)N1

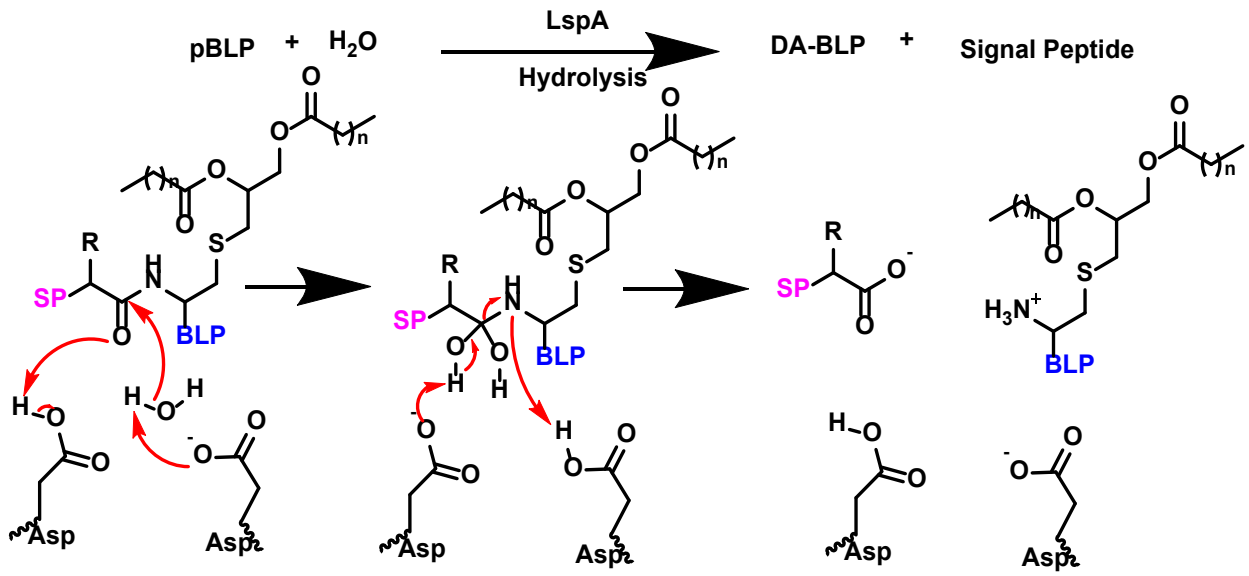


Figure 3.12 ***LspA mechanism***

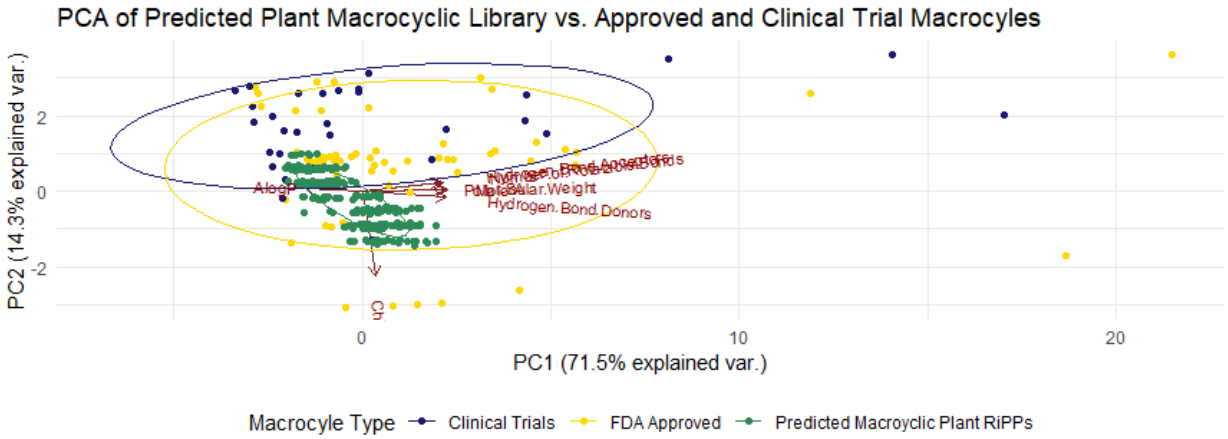


Figure 3.13 Principal component analysis of mined plant RiPPs and Macrocyles in clinical trials or approved by the FDA.

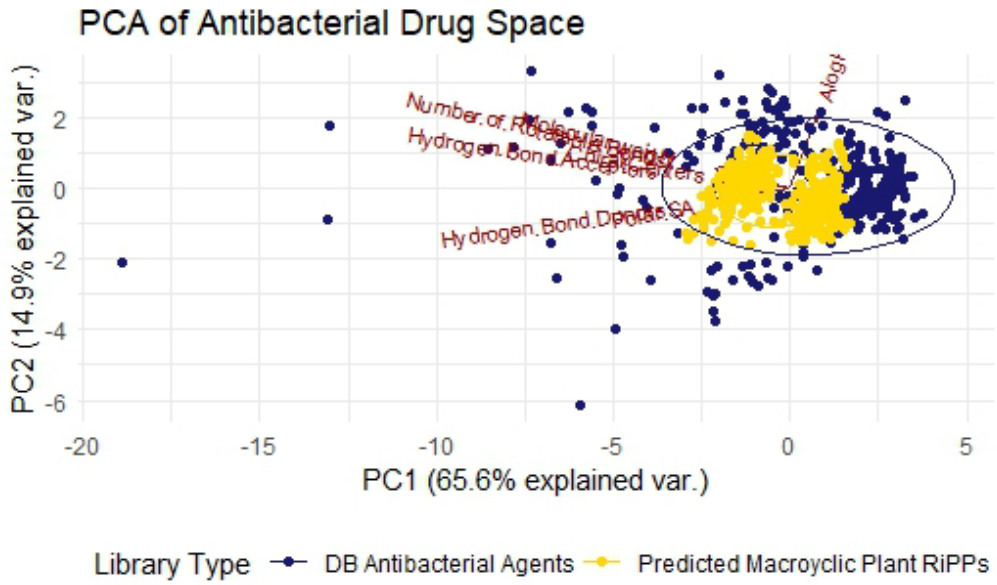


Figure 3.14 Principal component analysis of mined plant RiPPs and DrugBank's Antibacterial Agents database



Figure 3.15 Principal component analysis of mined plant RiPPs and the Coconut Natural Product Database

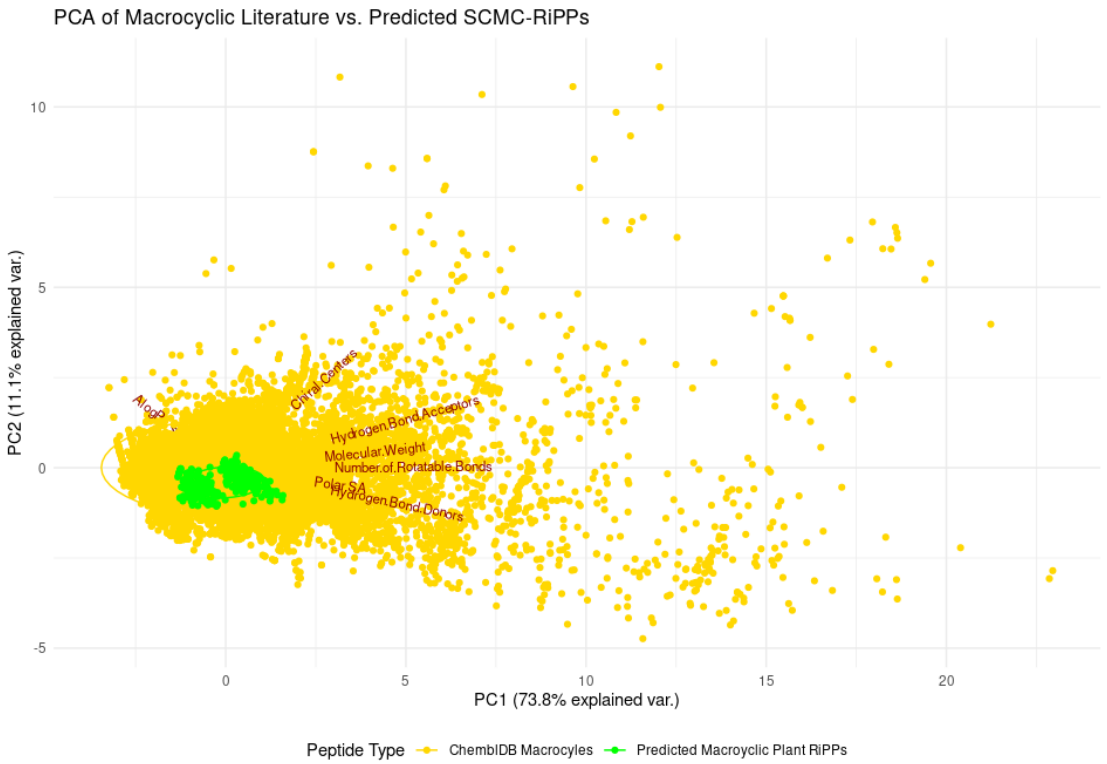


Figure 3.16 Principal component analysis of mined plant RiPPs and the Chembl Macrocycles Database

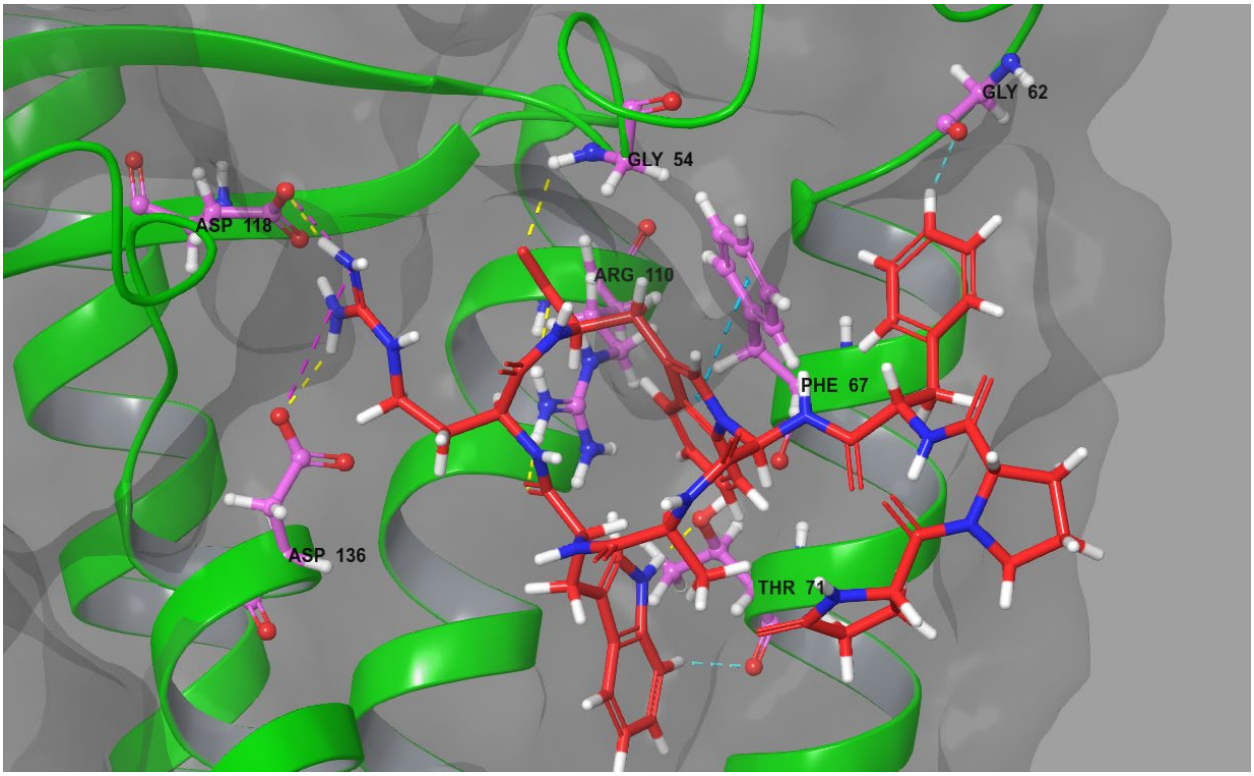


Figure 3.17 Docked QPFGAWRW in LspA active site.

3.8 References

1. Chassagne, F. *et al.* A Systematic Review of Plants With Antibacterial Activities: A Taxonomic and Phylogenetic Perspective. *Front. Pharmacol.* **11**, 586548 (2021).
2. Centers for Disease Control and Prevention (U.S.). Antibiotic resistance threats in the United States, 2019. (2019).
3. General Information | MRSA | CDC.
4. Zaynab, M. *et al.* Role of secondary metabolites in plant defense against pathogens. *Microb. Pathog.* **124**, 198–202 (2018).
5. Kim, W., Hendricks, G. L., Tori, K., Fuchs, B. B. & Mylonakis, E. Strategies against Methicillin-Resistant *Staphylococcus Aureus* Persisters. *Future Med. Chem.* **10**, 779–794 (2018).
6. Tobin, E. H., Jogu, P. & Koirala, J. Methicillin-Resistant Staphylococcus aureus. in *StatPearls* (StatPearls Publishing, Treasure Island (FL), 2025).
7. Peacock, S. J. & Paterson, G. K. Mechanisms of Methicillin Resistance in *Staphylococcus aureus*. *Annu. Rev. Biochem.* **84**, 577–601 (2015).
8. Vancomycin infusion guidelines - DoseMeRx.
9. Intravenous Vancomycin | Dosing Guidelines, Trough Levels, Side Effects | DoseMeRx.
10. Van Herwerden, E. F. & Süssmuth, R. D. Sources for Leads: Natural Products and Libraries. in *New Approaches to Drug Discovery* (eds Nielsch, U., Fuhrmann, U. & Jaroch, S.) vol. 232 91–123 (Springer International Publishing, Cham, 2015).
11. Choo, E. J. & Chambers, H. F. Treatment of Methicillin-Resistant *Staphylococcus aureus* Bacteremia. *Infect. Chemother.* **48**, 267 (2016).
12. Daptomycin.
13. Estes, K. S. & Derendorf, H. Comparison of the pharmacokinetic properties of vancomycin, linezolid, tigecyclin, and daptomycin. *Eur. J. Med. Res.* **15**, 533–543 (2010).
14. Vogeley, L. *et al.* Structural basis of lipoprotein signal peptidase II action and inhibition by the antibiotic globomycin. *Science* **351**, 876–880 (2016).
15. Kiho, T. *et al.* Structure–activity relationships of globomycin analogues as antibiotics. *Bioorg. Med. Chem.* **12**, 337–361 (2004).
16. Olatunji, S. *et al.* Structures of lipoprotein signal peptidase II from Staphylococcus aureus complexed with antibiotics globomycin and myxovirescin. *Nat. Commun.* **11**, 140 (2020).
17. Nguyen, M. T. & Götz, F. Lipoproteins of Gram-Positive Bacteria: Key Players in the Immune Response and Virulence. *Microbiol. Mol. Biol. Rev. MMBR* **80**, 891–903 (2016).
18. Nguyen, M.-T., Matsuo, M., Niemann, S., Herrmann, M. & Götz, F. Lipoproteins in Gram-Positive Bacteria: Abundance, Function, Fitness. *Front. Microbiol.* **11**, 582582 (2020).
19. McChesney, J. D., Venkataraman, S. K. & Henri, J. T. Plant natural products: Back to the future or into extinction? *Phytochemistry* **68**, 2015–2022 (2007).
20. Garcia Jimenez, D., Poongavanam, V. & Kihlberg, J. Macrocycles in Drug Discovery─Learning from the Past for the Future. *J. Med. Chem.* **66**, 5377–5396 (2023).
21. Kaur, H. *et al.* The antibiotic darobactin mimics a β-strand to inhibit outer membrane insertase. *Nature* **593**, 125–129 (2021).
22. Liu, J. *et al.* Synthesis and Characterization of the Arylomycin Lipoglycopeptide Antibiotics and the Crystallographic Analysis of Their Complex with Signal Peptidase. *J. Am. Chem. Soc.* **133**, 17869–17877 (2011).

23. Kersten, R. D. *et al.* Gene-Guided Discovery and Ribosomal Biosynthesis of Moroidin Peptides. *J. Am. Chem. Soc.* **144**, 7686–7692 (2022).
24. Kersten, R. D. & Weng, J.-K. Gene-guided discovery and engineering of branched cyclic peptides in plants. *Proc. Natl. Acad. Sci.* **115**, (2018).
25. Chigumba, D. N. *et al.* Discovery and biosynthesis of cyclic plant peptides via autocatalytic cyclases. *Nat. Chem. Biol.* **18**, 18–28 (2022).
26. Cao, L., Do, T. & Link, A. J. Mechanisms of action of ribosomally synthesized and post-translationally modified peptides (RiPPs). *J. Ind. Microbiol. Biotechnol.* **48**, kuab005 (2021).
27. Arnison, P. G. *et al.* Ribosomally synthesized and post-translationally modified peptide natural products: overview and recommendations for a universal nomenclature. *Nat Prod Rep* **30**, 108–160 (2013).
28. Montalbán-López, M. *et al.* New developments in RiPP discovery, enzymology and engineering. *Nat. Prod. Rep.* **38**, 130–239 (2021).
29. Chekan, J. R., Mydy, L. S., Pasquale, M. A. & Kersten, R. D. Plant peptides – redefining an area of ribosomally synthesized and post-translationally modified peptides. *Nat. Prod. Rep.* **41**, 1020–1059 (2024).
30. Hattori, J., Boutilier, K. A., Campagne, M. M. L. & Miki, B. L. A conserved BURP domain defines a novel group of plant proteins with unusual primary structures. *Mol. Gen. Genet. MGG* **259**, 424–428 (1998).
31. Bass□ner, R. *et al.* Abundant embryonic mRNA in field bean (Vicia faba L.) codes for a new class of seed proteins: cDNA cloning and characterization of the primary translation product. *Plant Mol. Biol.* **11**, 321–334 (1988).
32. Treacy, B. K. *et al.* Bnm1, a Brassica pollen-specific gene. *Plant Mol. Biol.* **34**, 603–611 (1997).
33. Zheng, L., Heupel, R. C. & DellaPenna, D. The beta subunit of tomato fruit polygalacturonase isoenzyme 1: isolation, characterization, and identification of unique structural features. *Plant Cell* **4**, 1147–1156 (1992).
34. Yamaguchi-Shinozaki, K. & Shinozaki, K. The plant hormone abscisic acid mediates the drought-induced expression but not the seed-specific expression of rd22, a gene responsive to dehydration stress in Arabidopsis thaliana. *Mol. Gen. Genet. MGG* **238–238**, 17–25 (1993).
35. Mydy, L. S., Chigumba, D. N. & Kersten, R. D. Plant Copper Metalloenzymes As Prospects for New Metabolism Involving Aromatic Compounds. *Front. Plant Sci.* **12**, 692108 (2021).
36. Mydy, L. S. *et al.* An intramolecular macrocyclase in plant ribosomal peptide biosynthesis. *Nat. Chem. Biol.* **20**, 530–540 (2024).
37. Hattori, J., Boutilier, K. A., Campagne, M. M. L. & Miki, B. L. A conserved BURP domain defines a novel group of plant proteins with unusual primary structures. *Mol. Gen. Genet. MGG* **259**, 424–428 (1998).
38. Han, H. *et al.* Anxiolytic-like effects of sanjoinine A isolated from Zizyphi Spinosi Semen: Possible involvement of GABAergic transmission. *Pharmacol. Biochem. Behav.* **92**, 206–213 (2009).
39. Morita, H., Shimbo, K., Shigemori, H. & Kobayashi, J. Antimitotic activity of moroidin, a bicyclic peptide from the seeds of Celosia argentea. *Bioorg. Med. Chem. Lett.* **10**, 469–471 (2000).

40. Kang, K. B. *et al.* Jubanines F–J, cyclopeptide alkaloids from the roots of Ziziphus jujuba. *Phytochemistry* **119**, 90–95 (2015).
41. Yahara, S. *et al.* Structures of anti-ace and -renin peptides from lycii radicis cortex. *Tetrahedron Lett.* **30**, 6041–6042 (1989).
42. Morita, H., Yoshida, N., Takeya, K., Itokawa, H. & Shirota, O. Configurational and conformational analyses of a cyclic octapeptide, lyciumin A, from Lycium chinense Mill. *Tetrahedron* **52**, 2795–2802 (1996).
43. Bankevich, A. *et al.* SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing. *J. Comput. Biol.* **19**, 455–477 (2012).
44. Bushmanova, E., Antipov, D., Lapidus, A. & Prjibelski, A. D. rnaSPAdes: a *de novo* transcriptome assembler and its application to RNA-Seq data. *GigaScience* **8**, giz100 (2019).
45. Li, D., Liu, C.-M., Luo, R., Sadakane, K. & Lam, T.-W. MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct *de Bruijn* graph. *Bioinformatics* **31**, 1674–1676 (2015).
46. Priyam, A. *et al.* Sequenceserver: A Modern Graphical User Interface for Custom BLAST Databases. *Mol. Biol. Evol.* **36**, 2922–2924 (2019).
47. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410 (1990).
48. Tang, Y. *et al.* Expression of a Vacuole-Localized BURP-Domain Protein from Soybean (SALI3-2) Enhances Tolerance to Cadmium and Copper Stresses. *PLoS ONE* **9**, e98830 (2014).
49. Singh, U. & Wurtele, E. S. orfipy: a fast and flexible tool for extracting ORFs. *Bioinformatics* **37**, 3019–3020 (2021).
50. Release 2022-3 | Schrödinger.
51. Jolliffe, I. T. & Cadima, J. Principal component analysis: a review and recent developments. *Philos. Transact. A Math. Phys. Eng. Sci.* **374**, 20150202 (2016).
52. Wang, C. K. L., Kaas, Q., Chiche, L. & Craik, D. J. CyBase: a database of cyclic protein sequences and structures, with applications in protein discovery and engineering. *Nucleic Acids Res.* **36**, D206–D210 (2007).
53. Wishart, D. S. *et al.* DrugBank: a knowledgebase for drugs, drug actions and drug targets. *Nucleic Acids Res.* **36**, D901-906 (2008).
54. Gaulton, A. *et al.* ChEMBL: a large-scale bioactivity database for drug discovery. *Nucleic Acids Res.* **40**, D1100–D1107 (2012).
55. Sorokina, M., Merseburger, P., Rajan, K., Yirik, M. A. & Steinbeck, C. COCONUT online: Collection of Open Natural Products database. *J. Cheminformatics* **13**, 2 (2021).
56. Stephenson, M. J., Reed, J., Brouwer, B. & Osbourn, A. Transient Expression in Nicotiana Benthamiana Leaves for Triterpene Production at a Preparative Scale. *J. Vis. Exp.* 58169 (2018) doi:10.3791/58169.
57. Sainsbury, F., Thuenemann, E. C. & Lomonossoff, G. P. pEAQ: versatile expression vectors for easy and quick transient expression of heterologous proteins in plants. *Plant Biotechnol. J.* **7**, 682–693 (2009).
58. Katz, K. *et al.* The Sequence Read Archive: a decade more of explosive growth. *Nucleic Acids Res.* **50**, D387–D390 (2022).
59. Walsh, C. T. & Tang, Y. *Natural Product Biosynthesis: Chemical Logic and Enzymatic Machinery*. (The Royal Society of Chemistry, 2017). doi:10.1039/9781788016612.

60. Kodama, Y., Shumway, M., Leinonen, R., & on behalf of the International Nucleotide Sequence Database Collaboration. The sequence read archive: explosive growth of sequencing data. *Nucleic Acids Res.* **40**, D54–D56 (2012).
61. O'Leary, N. A. *et al.* Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res.* **44**, D733–D745 (2016).
62. Oselusi, S. O. *et al. A.* Computational target-based screening of anti-MRSA natural products reveals potential multitarget mechanisms of action through peptidoglycan synthesis proteins. ACS Omega **7**, 37896–37906 (2022).
63. Xu, X. *et al.* In silico screening of protein-binding peptides with an application to developing peptide inhibitors against antibiotic resistance. *PNAS Nexus* **3**, 541 (2024).
64. Mustafa, G. *et al.* Investigation of plant antimicrobial peptides against selected pathogenic bacterial species using a peptide-protein docking approach. BioMed Res. Int. **2022**, 1077814 (2022).

Chapter 4 Diversification of Stephanotic-acid type Burpitides and Improved Moroidin Bioactivity against H1437 Lung Adenocarcinoma Cells

4.1 Statement of contribution

This chapter is partially reproduced and adapted from the following: Wang, X., Shafiq, K., Ousley, D.A. *et al.* Large-scale transcriptome mining enables macrocyclic diversification and improved bioactivity of the stephanotic acid scaffold. *Nat Commun* 16, 4198 (2025). Dr. Xiaofeng Wang purified, isolated, and structurally elucidated the peptides. Dr. Khadija Shafiq conducted the bioassays, performed the RNA-seq assays and extraction, and aided with the purification of Glechomanin. I performed RNA-seq analysis of treated H1437 cells, constructed a phylogenetic tree, and co-authored the publication. Max Garcia (Sexton Lab) performed the AutoPlex analysis to help analyze the mechanism of action.

4.2 Summary

Moroidins are bicyclic compounds with a conserved stephanotic acid scaffold, which becomes cytotoxic to non-small cell lung adenocarcinoma cells with an additional C-terminal macrocycle. We discover moroidin analogs with second ring structures diversified at the crosslink and the non-crosslinked residues including a moroidin analog from water chickweed, which exhibits higher cytotoxicity against lung adenocarcinoma cells than moroidin. Our study expands stephanotic acid-type peptides to grasses, Lowiaceae, mints, pinks, and spurges. Finally, it explores moroidin-QLLVWRNH's mechanism of action through RNA-seq analysis and differential gene expression.

4.3 Introduction

Burpitides are a recent addition to the diversity of plant RiPPs (**Figure 4.11**)[1–3], which consist of 4-8 amino acids and feature crosslinks between aromatic amino acids, such as tryptophan or tyrosine, and other amino acids or the peptide backbone[4–7]. These crosslinks often involve functionalization of C(sp^3)-H-bonds by copper-dependent burpitide cyclases comprising a C-terminal BURP domain[8–12]. The stephanotic acid-type burpitide scaffold is derived from a QLxxW core peptide (where x is any proteinogenic amino acid), and defined by a C-C-bond between a Leu-Cβ and a tryptophan-indole-C6[13,14]. These peptides can be monocyclic such as stephanotic acid-[LV] from *Cercis canadensis*, or bicyclic such as moroidin from *Dendrocnide moroides* (**Figure 4.1a**)[15,16]. In moroidins, the stephanotic acid scaffold forms the first ring and the second ring is formed by a C-N-crosslink between the tryptophan-indole-C2 and the imidazole-N1 of a C-terminal histidine (**Figure 4.1b**). Reported moroidins are 7–8 amino acids-long with the second ring comprising 2–4 amino acids[4,15,16].

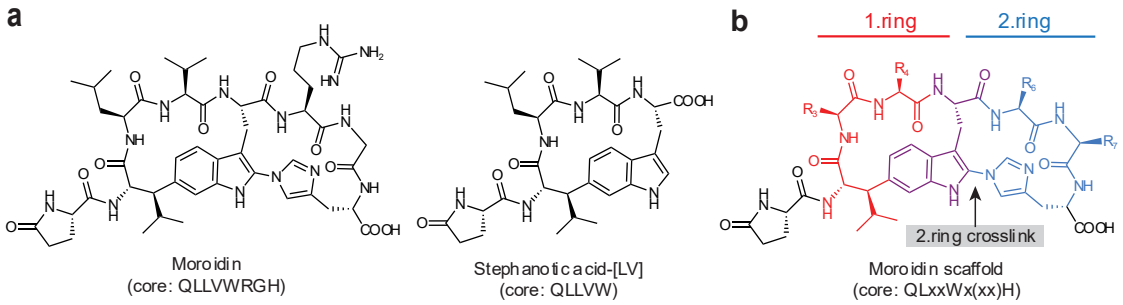


Figure 4.1 Transcriptome mining of plant RiPPs.
a Chemical structures of stephanotic acid-type burpitides. **b** The moroidin scaffold.

Moroidin and its analog celogentin C exhibit moderate *in vitro* cytotoxicity against non-small cell lung cancer whereas stephanotic acid-[LV] has no such detectable cytotoxicity, indicating that the second ring of moroidin is required for its bioactivity[4,7]. Herein, we characterize two new post-translational modifications in stephanotic acid-type burpitides and their associated biosynthetic cyclase's. Furthermore, we assess the bioactivity of a moroidin analog and establish an expanded chemotaxonomic landscape for accessing stephanotic acid-type burpitides in diverse plant families. All peptides studied in this chapter were previously mined from the transcriptomes of 7569 plant species studied in chapter 2.

4.4 Results

4.4.1 Characterization of stephanotic acid-type burpitides with a Trp-Val-crosslinked second-ring

We first set out to search for a stephanotic acid-type burpitide with a second-ring-crosslink via a C-terminal tryptophan. One corresponding burpitide cyclase with a core including a C-terminal tryptophan (**Table 4.2**) was encoded in a transcriptome of *Glechoma hederacea* (**Lamiaceae, Figure 4.2a**), encoding the core peptides QLFVWGW and QLLVWGW (**Table 4.2**). An analyte with a mass value of 914.419 Da, which matched the core peptide QLFVWGW with mass shifts corresponding to pyroglutamate (PyroGlu) formation and two macrocyclizations, was detected in the methanolic extract of *G. hederacea* leaves (**Figure 4.2b, and Figure 4.12**). The candidate burpitide cyclase gene, named *GheBURP*, of this analyte was cloned from the leaf transcriptome of *G. hederacea* and included four core peptides, three QLFVWGW and one QLLVWGW. Transient expression of GheBURP in Nicotiana benthamiana via the pEAQ-HT expression vector[17] through Agrobacterium tumefaciens LBA4404 infiltration yielded the same QLFVWGW analyte in methanolic leaf extract of transgenic tobacco six days after infiltration (**Figure 4.2b**). To elucidate the structure of burpitide 1 (**Figure 4.2c**), the analyte was isolated from G. hederacea collected from the University of Michigan Arboretum due to higher peptide yields in the source plant than transgenic N. benthamiana expressing GheBURP, and subsequently subjected to NMR and Marfey's analysis.

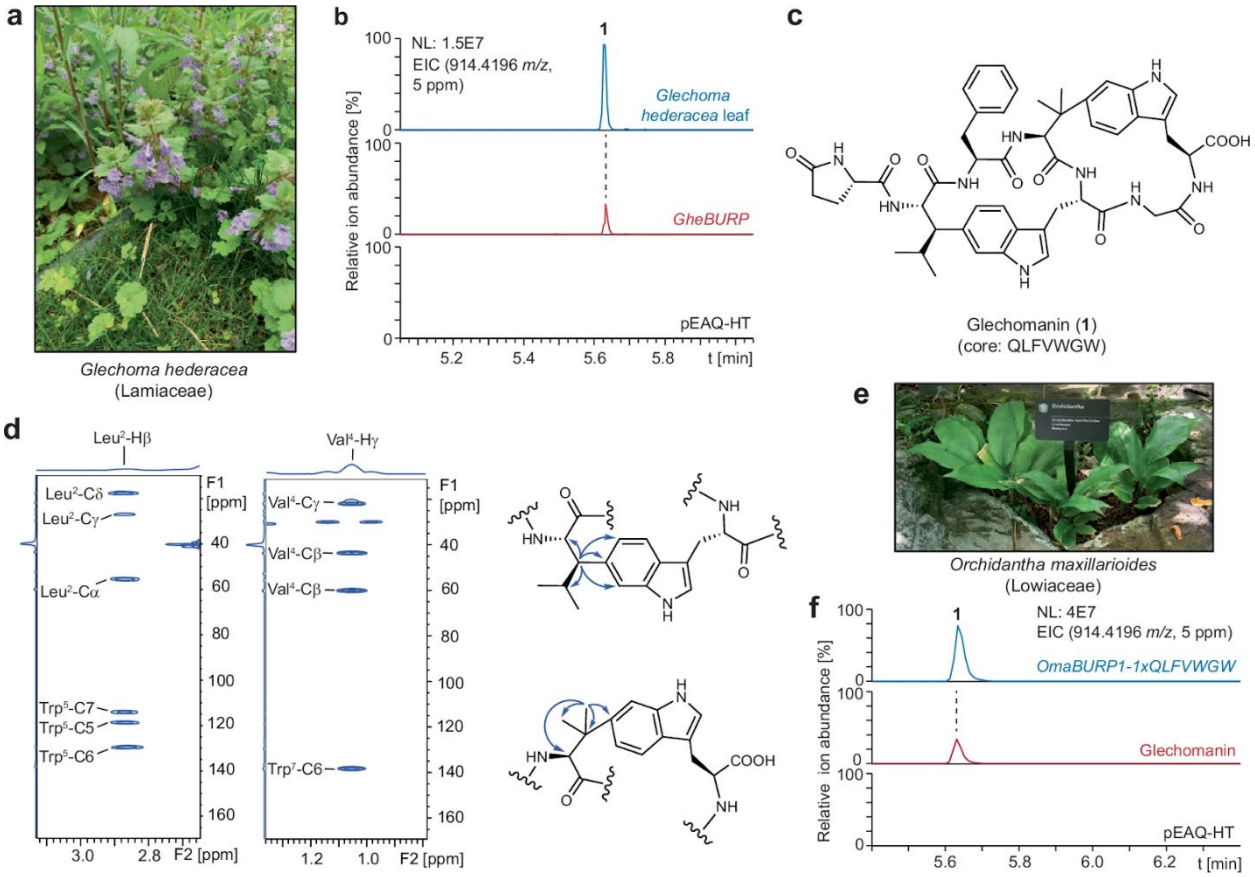


Figure 4.2 Characterization of stephanotic acid-burpitides with a Trp-Val-crosslinked second ring.
a *Glechoma hederacea* in the University of Michigan Arboretum. **b** Glechomanin detection in methanolic leaf extracts of *G. hederacea* and of transgenic *N. benthamiana* 6 days after infiltration with *Agrobacterium tumefaciens* LBA4404 pEAQ-HT-*GheBURP*. **c** Glechomanin structure. **d** Key HMBC signals of glechomanin side-chain-crosslinks. **e** *Orchidantha maxillarioides* plant in Missouri botanical garden. **f** Glechomanin detection in methanolic leaf extract of transgenic *N. benthamiana* 6 days after infiltration with *A. tumefaciens* LBA4404 pEAQ-HT-*OmaBURP1-1xQLFVWGW*. EIC: extracted ion chromatogram. Source data are provided as a Source Data file.

An N-terminal pyroglutamate was characterized by COSY, TOCSY and HMBC and the spin systems of all other amino acids of the predicted core peptide QLFVWGW were determined by COSY, TOCSY, HMBC and ROESY. The core peptide sequence QLFVWGW was confirmed by HMBC and ROESY (**Figure 4.13-4.19, and Table 4.3 and 4.4**). The two macrocyclizations were defined as Leu2-Cβ to Trp5-C6 such as in stephanotic acid-[LV] and Val4-Cβ to Trp7-C6. The Leu2-Trp5-crosslink was characterized by (a) HMBC of Leu-Hβ to Trp5-C6, Trp5-C5, Trp5-C7 (**Figure 4.2d**) and ROESY from Leu-Hβ to Trp5-H7, from Leu-Hα to Trp5-H5 and from Leu-Hδ1 to Trp5-H7 (**Figure 4.17, 4.18, 4.20, and 4.21**) and (b) the detection of stephanotic acid-[LV] (core: QLLVW) after transient expression of GheBURP, which encodes a QLLVWGW core, in *N. Benthamiana* (**Figure 4.22**). The Val4-Trp7-crosslink was characterized by (a) HMBC of Val4-Hγ1 to Val4-Cβ and Trp7-C6 (**Figure 4.2d**) and (b) ROESY of Val4-Hγ1 with Trp7-H7 and of

Val4-Hγ2 with Trp7-H5 and Trp7-H7 and of Val4-Hα with Trp7-H7 (**Figure 4.17, 4.18, 4.20, and 4.21**). In addition, no Val4-Hβ signal was detected. Furthermore, two indole-NH-signals were detected excluding a crosslink via an indole-nitrogen as seen in lyciumins[5]. Marfey's analysis of glechomanin revealed PyroGlu1 and Phe3 are both l-configuration (**Figure 4.23**). Moroidin features the same macrocyclic Leu-Trp-connection in the first ring as glechomanin and we observed that the chemical shift value and coupling constants of the β-position hydrogen in the Leu2 fragment of moroidin and glechomanin were comparable (δH 3.06 (dd, J = 3.0, 11.5 Hz)[4] and δ_H 2.87 (dd, *J* = 3.1, 11.6 Hz), respectively) (**Table 4.3**). Therefore, we deduce that the β-carbon of Leu2 in glechomanin shares the same relative configuration as observed in moroidin. In addition, the stereochemistry of the first five glechomanin residues (QLFVW) was assigned as stephanotic acid-[LV] because transient expression of WT GheBURP, which has a QLLVWGW core, and GheBURP-1xQLLVW, which has a stephanotic acid-[LV] core peptide, yielded stephanotic acid-[LV] (**Figure 4.22**). The stereochemistry of the C-terminal tryptophan was determined as L-Trp7 (Trp7-S-Cα) by comparison of glechomanin NOE data to two geometry-optimized structures of glechomanin with R- or S-configuration at Trp7-Cα and the coupling constant of the hydrogen atoms at the α and β1 positions (Jαβ1 = 11.2 Hz) (**Figure 4.24**). In addition to glechomanin from *G. hederacea*, we detected a glechomanin analyte with the core peptide QLLVWKW in *Orchidantha maxillarioides* (Lowiaceae) (**Figure 4.2e**) based on a predicted glechomanin cyclase in its transcriptome (**Table 4.2**) Cloning and heterologous expression of the corresponding burpitide cyclase gene OmaBURP1 in *N. benthamiana* yielded the putative glechomanin-QLLVWKW analyte as observed in *O. maxillarioides* leaf extract (**Figure 4.25**). The structure of this analyte was confirmed as a glechomanin scaffold by expression of a synthetic OmaBURP1 with one core peptide of QLFVWGW, i.e., the glechomanin core, in transgenic *N. benthamiana* which resulted in the detection of glechomanin from ground ivy (**Figure 4.2f**).

GheBURP belongs to the RD22-type burpitide cyclases, which are autocatalytic and contain repetitive N-terminal core peptide domains[1,7]. To characterize GheBURP as a glechomanin cyclase, GheBURP was synthesized as an *E. coli*-codon-optimized gene construct with one core peptide of QLFVWGW and expressed heterologously as a His-tagged protein in *E. coli* BL21(DE3) (**Figure 4.26**). GheBURP-1xQLFVWGW was subsequently purified via denature-refolding from inclusion bodies and immobilized nickel-affinity chromatography (**Figure 4.26**)[18]. LC-MS/MS-based bottom-up proteomic analysis of GheBURP-1xQLFVWGW

with trypsin yielded an analyte with the mass of 2398.134 Da matching the sequence QVTTYNAASDNGNQLFVWGWK and, thus, including the glechomanin core peptide (**Figure 4.27 and 4.28**, **Table 4.5**). Incubation of GheBURP-1xQLFVWGW with 1 mM cupric chloride at pH 7 for 24 h and subsequent tryptic digest resulted in the detection of a tryptic core peptide species with a loss of four hydrogens in the peptide motif QLFVWGWK (**Figure 4.27**). The characterized MS/MS fragmentation of these tryptic core peptides indicated the formation of two macrocyclization bonds in the QLFVWGW core peptide (**Figure 4.29**, **Table 4.6**). To confirm the mass loss of four hydrogens in the core peptide as the crosslinks of the glechomanin structure, the tryptic core peptide species was purified by HPLC from a scaled GheBURP-1xQLFVWGW enzyme assay with cupric chloride followed by trypsin proteolysis. The purified modified GheBURP peptide was then incubated with an aminopeptidase and a carboxypeptidase for 24 h. LC-MS analysis of the exopeptidase assay detected an analyte matching glechomanin (**Figure 4.27**) which shows that the mass loss observed after GheBURP enzyme assay represents the bicyclic bonds of glechomanin. Our chemotyping and biochemical analysis revealed that stephanotic acid-type burpitide pathways yielding bicyclic scaffolds via a C-terminal tryptophan-crosslink exist in plants.

4.4.2 Characterization of stephanotic acid-type burpitides with a Tyr-Phe-crosslinked second-ring

Next, we aimed to identify a stephanotic acid pathway yielding a bicyclic scaffold with a C-terminal tyrosine from candidate cyclases with corresponding core peptides in our transcriptome database (**Table 4.2**). A transcriptome of annual mercury (Mercurialis annua, Euphorbiaceae) encoded RD22-type BURP-domain proteins with putative stephanotic acid-type cores QLFFWRY and QLFFWKY with C-terminal tyrosines. The full-length sequence of one such BURP-domain protein, named ManBURP, was obtained from the recently released annual mercury genome (GenBank ID XP_050232080.1)[19]. Transient expression of a full-length ManBURP in *N. benthamian*a yielded an analyte with the mass 1038.483 Da, which matched the core peptide sequence QLFFWRY after loss of four hydrogens corresponding to two putative macrocyclizations and loss of ammonia due to N-terminal pyroglutamate formation (**Figure 4.30**). The same analyte

was detectable in an herbarium sample of M. annua (**Figure 4.3a, b**) but we could not obtain this source plant in sufficient quantities for peptide isolation. Therefore, we transiently expressed a truncated ManBURP-1xQLFFWRY construct via pEAQ-HT expression system in N. benthamiana at scale and purified 5 mg of the target peptide from methanolic extracts of 1.3 kg of transgenic tobacco leaves.

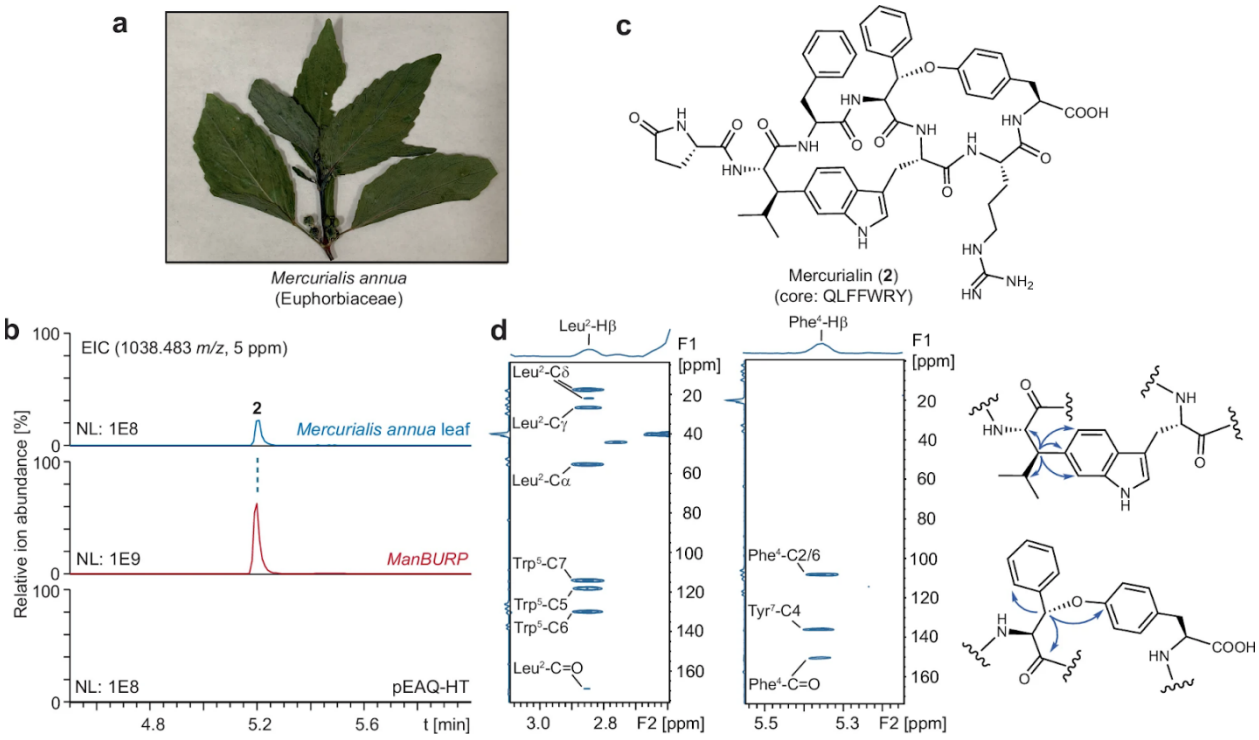


Figure 4.3 Characterization of stephanotic acid-burpitides with a Tyr-Phe-crosslinked second ring.
a *Mercurialis annua* from University of Michigan herbarium. **b** Mercurialin detection in methanolic leaf extracts of *M. annua* and of transgenic *N. benthamiana* 6 days after infiltration with *A. tumefaciens* LBA4404 pEAQ-HT-ManBURP. **c** Mercurialin structure. **d** Key HMBC correlations of mercurialin side-chain-crosslinks. EIC: extracted ion chromatogram.

The pure peptide, named mercurialin, was characterized by analysis of 1D and 2D NMR spectroscopic data (**Figure 4.3c**, **Table 4.7**, **Figure 31-38**). All its amino acids, including an N-terminal pyroglutamate, were determined by COSY, TOCSY, HMBC and NOESY. The core peptide sequence QLFFWKY was confirmed by HMBC and ROESY. Mercurialin has the characteristic stephanotic acid crosslink between Leu2-Cβ and Trp5-C6, evidenced by HMBC correlations of Leu-Hβ with Trp5-C5, C6 and C7 as well as NOESY correlations from Leu2-Hβ to Trp5-H7, from Leu2-Hα to Trp5-H5, and from Leu2-Hδ1 to Trp5-H7. The crosslinking position of the second identified ring is established between Phe4-Cβ and Tyr7-O, as unambiguously supported by HMBC correlations of Phe4-Hβ with Tyr7-C4 and NOESY correlations from Phe4-

Hβ to Tyr7-H3 and Tyr7-H5 (**Figure 4.3d**, **Table 4.7**, **Figure 35 and 36**). To determine the stereochemistry of mercurialin, we performed Marfey's analysis, which indicates that PyroGlu, Phe, and Arg are all present in the L-configuration. The stereochemistry of the stephanotic acid ring (QLFFW) was inferred as the stephanotic acid stereochemistry based on (a) the similar coupling constants of Leu-Hβ in mercurialin and glechomanin (**Tables 4.3 and 4.7**) and (b) the generation of stephanotic acid-[LV] after transient expression of ManBURP-1xQLLVWRY in *N. Benthamiana* (**Figure 39**). Given that the amino acid backbone synthesized by the ribosome exhibits an L-configuration, the stereochemistry of the C-terminal tyrosine is assigned as L-Tyr. The second macrocyclization site at the Cβ of phenylalanine in core residue 4 was determined to be S by comparing NOE correlations of mercurialin with two geometry-optimized structures featuring R- or S-configuration at Phe4-Cβ (**Figure 40**).

A truncated construct of mercurialin cyclase ManBURP with one core of QLFFWRY was expressed and purified like glechomanin cyclase construct GheBURP-1xQLFVWGW (**Figure 41**). Bottom-up proteomic analysis of the ManBURP-1xQLFFWRY with LysC digestion identified an analyte with a mass of 4123.068 Da, corresponding to the sequence IPQNYNFQPEPNQLFFWRYQNRGNDIIRLPSMK, which includes the mercurialin core peptide (**Figure 42 and 43**, **Table 4.8**). Incubation of ManBURP-1xQLFFWRY with 1 mM cupric chloride at pH 7 for 24 h, followed by LysC digestion, resulted in a loss of four hydrogens in the detected LysC-proteolytic core peptide (**Figure 42**). MS/MS fragmentation confirmed two macrocyclic bonds in the QLFFWRY core peptide (**Figure 44**, **Table 4.9**). To verify the 4H loss as crosslinks in the mercurialin structure, the LysC-proteolytic core peptide was purified by HPLC and incubated with aminopeptidase and carboxypeptidase for 24 h. LC-MS detected a product consistent with mercurialin (**Figure 42**), confirming the 4H mass loss reflects the two macrocyclic crosslinks of mercurialin. Mercurialin and its cyclase from *M. annua* show that the stephanotic acid pathway has also evolved bicyclic scaffolds with a C-terminal tyrosine-crosslink in plants.

4.4.3 Chemotaxonomic expansion and cancer cell cytotoxic improvement of Moroidins

Bioinformatic analysis of moroidin core peptides within our predicted moroidin cyclases showed that the core QLLVWRGH representing moroidin (**Figure 4.1b**) only appeared twice in

nineteen candidate moroidin cyclases whereas the core peptide QLLVWRNH was present in nine of the candidate moroidin cyclases from three plant families (**Figure 4.4a, Table 4.2**). Motivated by a rationale that evolution of a single core in multiple plant families might suggest potent bioactivity, we targeted moroidin-QLLVWRNH for isolation and biological evaluation[20]. A candidate moroidin-QLLVWRNH burpitide cyclase was identified in the transcriptome of water chickweed (*Stellaria aquatica*, Caryophyllaceae) (**Figure 4.4b**). The cloned candidate cyclase from *S. aquatica* contained 18 QLLVWRNH cores and a corresponding moroidin-QLLVWRNH analyte was detected in the methanolic extract of water chickweed from a park close to Lake Erie, Michigan (**Figure 4.4c**). Moroidin-QLLVWRNH was purified (32 mg) and characterized by LC-MS/MS, NMR and transient tobacco expression of Stellaria aquatica burpitide cyclase SaqBURP and reported moroidin cyclase KjaBURP with an updated core peptide (KjaBURP-1xQLLVWRNH), as a moroidin natural product (**Figure 4.4c, 4.4d, Figures 4.45-53**, **Table 4.10).** High-content screening of moroidin-QLLVWRNH against H1437, H1299, and U2OS cancer cell lines revealed significant cytotoxicity in H1437 non-small cell lung adenocarcinoma cells (IC_{50} = 1.4 μM, **Figure 4.5a**), exhibiting a five-fold lower IC50 than celogentin C, the most potent previously reported moroidin analog[21]. To evaluate selectivity and bioactivity, the peptide was orthogonally screened in three additional cancer cell lines (A549, LNCaP, and HUH-7) and three non-cancerous fibroblast lines (IMR-90, HFF, and MEF, Figure **4.5a, b**). Moroidin-QLLVWRNH was found inactive in H1299, U2OS, and LNCaP cells, but exhibited moderate dose-responsive cytotoxicity in A549 (IC50 = 12.7 μM) and HUH-7 (IC_{50} = 26.3 μM) cells. Notably, no significant dose-responsive cytotoxicity was observed in any non-cancerous fibroblast lines, underscoring its promising cancer cell specificity-an essential factor for clinical translation.

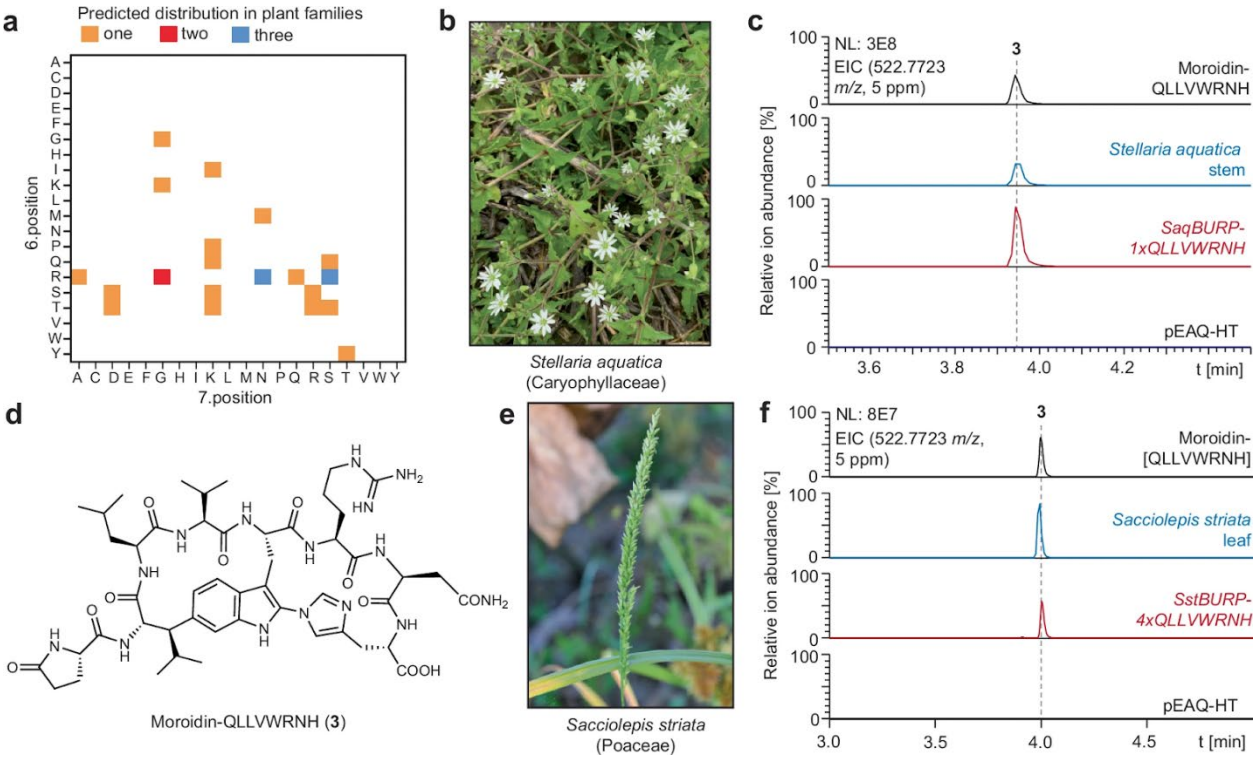
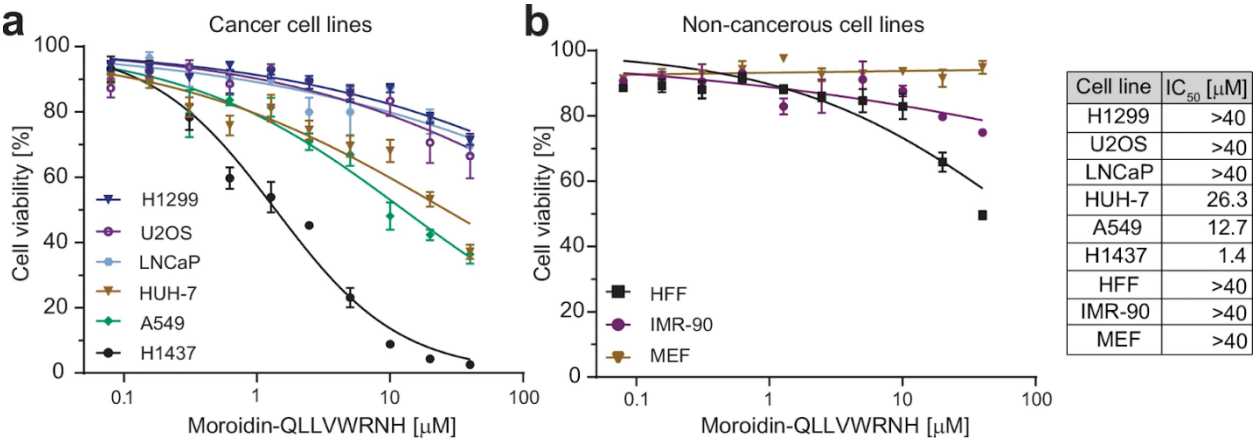


Figure 4.4 Characterization of moroidin burpitides in pinks and grasses.
a Second-ring sequence diversity of moroidin cores in transcriptomics-derived fused burpitide cyclases. **b** *Stellaria aquatica*. **c** LC-MS characterization of moroidin-QLLVWRNH in *S. aquatica* and transgenic *N. benthamiana* after 6 days of expression of SaqBURP-1xQLLVWRNH. **d** Moroidin-QLLVWRNH structure. **e** *Sacciolepis striata*. Photo credit: Larry Allain, U.S. Geological Survey. f LC-MS characterization of moroidin-QLLVWRNH in S. striata and in transgenic *N. benthamiana* after 6 days of expression of SstBURP-4xQLLVWRNH. EIC: extracted ion chromatogram.



Cell line	IC_{50} [μM]
H1299	>40
U2OS	>40
LNCaP	>40
HUH-7	26.3
A549	12.7
H1437	1.4
HFF	>40
IMR-90	>40
MEF	>40

Figure 4.5 Bioactivity of moroidin-QLLVWRNH.
a *In vitro* anticancer assays of moroidin-QLLVWRNH. Cell viability in six human cancer cell lines is plotted for the peptide across ten concentrations: H1299 (non-small cell lung cancer), U2OS (osteosarcoma), LNCaP (prostate cancer), HUH-7 (hepatoma), A549 (lung carcinoma), and H1437 (non-small cell lung adenocarcinoma). The data points represent the mean of triplicate samples (n = 3) and error bars represent one standard error of the mean. **b** *In vitro* counterscreen of moroidin-QLLVWRNH in three non-cancerous cell lines is plotted for the peptide across ten concentrations: HFF (human foreskin fibroblast), IMR-90 (human lung fibroblast), and MEF (mouse embryonic

fibroblast). The data points represent the mean of triplicate samples (n = 3) and error bars represent one standard error of the mean. Corresponding IC_{50} values of (a and b) are listed.

We identified additional moroidin analogs in Sacciolepis striata (Poaceae) and *Orchidantha maxillarioides* (Lowiaceae), two plant families without reported moroidin chemotypes. A fused burpitide cyclase with core peptides for moroidin-QLLVWRNH was predicted in the *S. striata* transcriptome (**Figure 4.4e**) and moroidin-QLLVWRNH was confirmed in the methanolic extract from a herbarium S. striata specimen, matching the chemotype of moroidin-QLLVWRNH from *S. aquatica* and KjaBURP-1xQLLVWRNH (**Figure 4.4f**). Similarly, four moroidin core peptides predicted in a transcript from the *O. maxillarioides* transcriptome and sequenced from cloned burpitide cyclase gene OmaBURP2 were verified as analytes in methanolic leaf extracts of *O. maxillarioides* by LC-MS/MS (**Figures 4.54-57**). Two of these analytes were confirmed as moroidins through LC-MS comparison with methanolic extracts of transgenic N. benthamiana expressing OmaBURP2 or KjaBURP with one core peptide of QLLVWRSH or QLLVWPKH (**Figures 4.54-57**).

4.4.4 RNA-seq of moroidin-QLLVWRNH treated H1437 Lung Adenocarcinoma Cells, Differential Gene Expression (DGE) analysis, and AutoPlex for Mechanism of Action (MoA) Interrogation

The H1437 lung adenocarcinoma cell line, established from a 60-year-old male with stage 1 lung adenocarcinoma and pleural effusion metastasis, exhibits several genetic aberrations that contribute to its oncogenic properties[22]. This cell line serves as a clinically relevant model for evaluating the cytotoxic potential of the stephanotic acid scaffold. Key pathway mutations present in H1437 cells include TP53, MAP2K1, and MET/PI3K-AKT, which disrupt cell cycle regulation and promote persistent survival signaling[23–27]. Additionally, various adenocarcinomas commonly express marker genes such as NKX2-1, NAPSA (involved in the functional development of the brain, lungs, and thyroid gland), cytokeratin 7 (an epithelial intermediate filament protein), and MUC1 (a transmembrane glycoprotein that when overexpressed facilities involvement in accelerated growth and metastases of cancer cells). Investigating the effects of moroidin-QLLVWRNH peptide treatment on gene expression in H1437 cells will clarify critical steps in elucidating the mechanism of action (MoA) and identifying relevant target genes and proteins.

This knowledge will also inform rational drug design and facilitate the discovery of novel moroidin peptides for lung cancer therapy.

The transcriptome of untreated H1437 cells was used as a baseline to establish thresholds for gene expression levels. In contrast, H1437 cells treated with moroidin-QLLVWRNH peptide were analyzed to identify differentially expressed genes in non-small cell lung adenocarcinoma. Moroidin-QLLVWRNH or a normalized DMSO concentration was administered at three concentrations (IC_{75}, IC_{50}, and IC_{25}), followed by incubation with H1437 cells. RNA was extracted after 3, 6, or 24 hours of treatment for subsequent quality assessment, RNA sequencing, and de novo transcriptomic assembly to capture gene expression changes (**Table 4.1**).

Table 4.1 An overview of the RNA-seq experiments.

Treatment	Dose (IC_x)	Conc. (uM)	Replicates	RNA Extraction (time)
1 - Vehicle Control	N/A	DMSO	3	3, 6, 24hr
2A - QLLVWRNH	IC_{25}	0.33	3	3, 6, 24hr
2B - QLLVWRNH	IC_{50}	1.25	3	3, 6, 24hr
2C - QLLVWRNH	IC_{75}	4.8	3	3, 6, 24hr

RNA sequencing (RNA-seq) is frequently used for differential expression analysis, which identifies genes with varying expression levels across experimental conditions[28]. Such genes provide biological insights into the processes influenced by the conditions under investigation. Differential Gene Expression (DGE) analysis aims to quantify the contribution of these effects and to distinguish biologically relevant genes from those with negligible changes[28]. The count data utilized in this analysis reflect the number of sequence reads mapped to each gene[28]. A higher count indicates a greater number of reads associated with a gene, supporting the inference that the gene was expressed at a higher level in the sample[28].

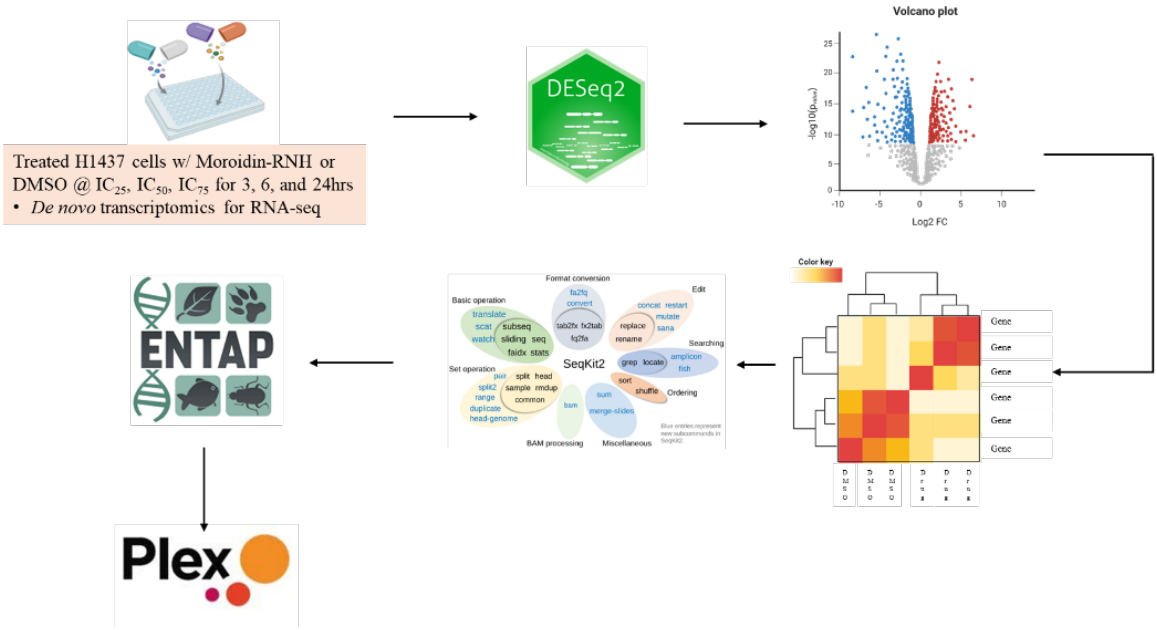


Figure 4.6 DGE Analysis Workflow of RNA-seq Experiments.

Differential gene expression (DGE) analysis was conducted using the DESeq2 package within the RStudio environment on the University of Michigan's Great Lakes High-Performance Computing cluster[28,29]. DESeq2 normalized count data to account for differences in library sizes and RNA composition between samples, followed by differential expression analysis[28,29]. The resulting dataset comprised genes from nine pairwise concentration comparisons of moroidin-QLLVWRNH versus the corresponding DMSO control. To ensure biological relevance, genes were filtered using an adjusted p-value (p-adj) < 0.05 and a 5% False Discovery Rate (FDR) alpha value, and the remaining genes were retained for downstream analysis. Volcano plots and heatmaps were generated in RStudio using the EnhancedVolcano and ComplexHeatmap R packages to visualize gene regulation patterns and assess clustering among replicates[30,31]. SeqKit2 was employed to segregate genes of interest for subsequent annotation with EnTAP[32,33]. AutoPlex facilitated the interpretation of annotation results and enabled rapid analytical comparison of the potential moroidin-QLLVWRNH mechanism of action with those of known drugs[34] (**Figure 4.6**).

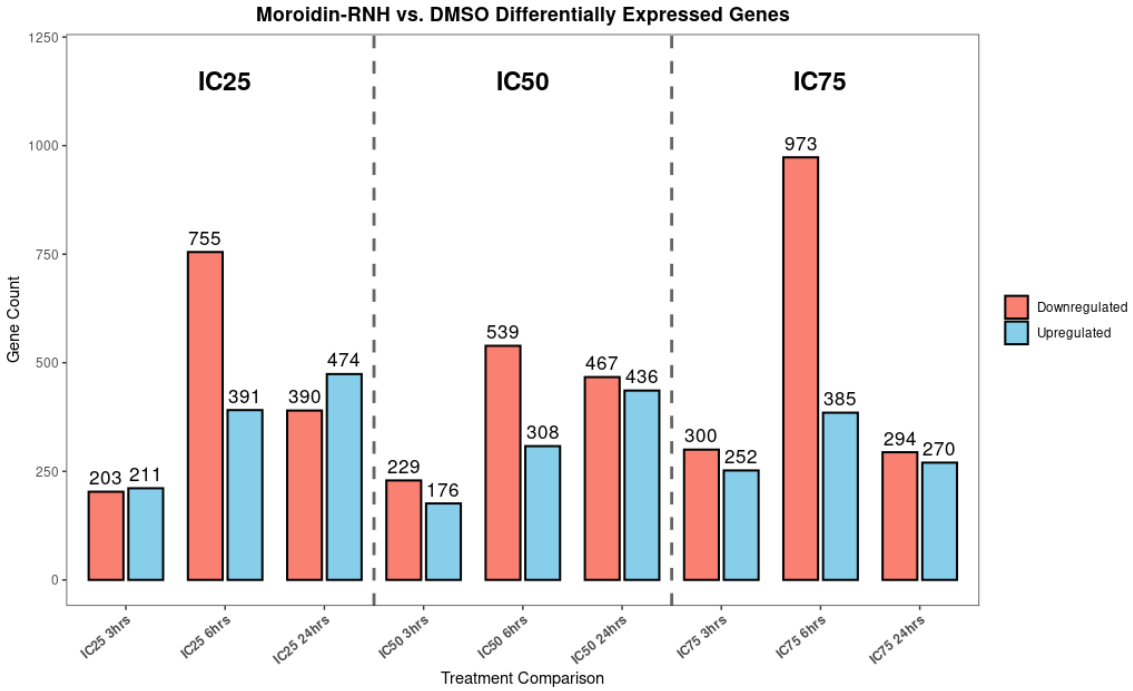


Figure 4.7 Moroidin-QLLVWRNH versus DMSO Differentially Expressed Genes Across All Time and Concentration Points.

Across all concentrations, 2,903 upregulated and 4,150 downregulated differentially expressed genes (DEGs) serve as quantitative indicators of the compound's potency and peptide-induced stress, reflecting the extent of gene dysregulation at different concentrations and exposure times (**Figure 4.7**). This section focuses on the analysis of the half-maximal inhibitory concentration (IC_{50}) under the 6-hour peptide treatment condition, which represents the point at which half the population responds to lethal stress without evidence of non-specific metabolic collapse. Additional heatmaps and volcano plots for the IC_{25} and IC_{75} concentrations at the 6-hour time point are provided as Figures to demonstrate the dose-dependent progression of cellular responses (**Figures 4.62-4.65**).

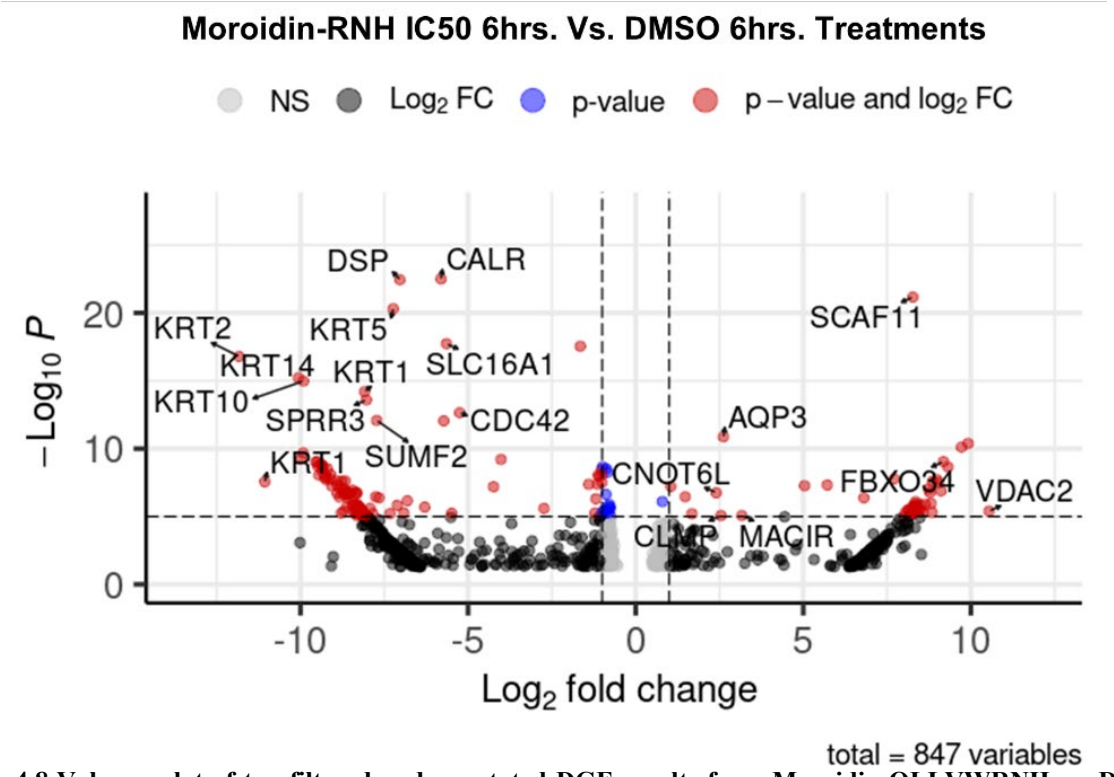


Figure 4.8 Volcano plot of top filtered and annotated DGE results from Moroidin-QLLVWRNH vs. DMSO treatment for 6hrs at concentration IC_{50}.

The volcano plot indicates that the moroidin-QLLVWRNH response involves the reprogramming of specific cellular pathways, rather than representing transcriptional noise. All genes shown in the plot satisfied the statistical threshold of an adjusted p-value ≤ 0.05. Upregulated genes may serve as candidates for compensatory survival mechanisms, while downregulated genes could contribute to core oncogenic pathways in H1437 cells (**Figure 4.8**).

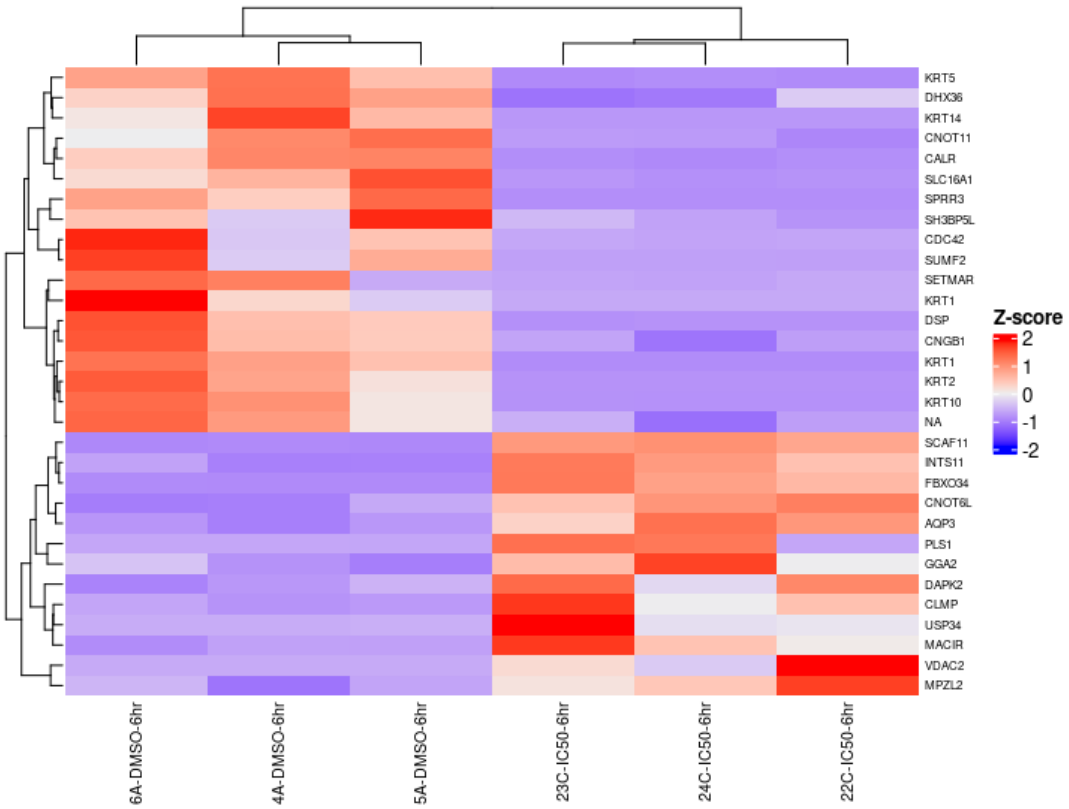


Figure 4.9 Heatmap of top filtered and annotated DGE results from Moroidin-QLLVWRNH vs. DMSO treatment for 6hrs at concentration IC_{50}.

The complementary heatmap for this treatment condition demonstrates clustering among biological replicates. Gene clustering indicates that moroidin-QLLVWRNH may influence multi-protein complexes or interconnected signaling cascades rather than isolated enzymes (**Figure 4.9**). The AutoPlex platform was employed to interrogate the genes of interest visualized in the heatmap and volcano plot[34]. AutoPlex is a knowledge graph platform designed for drug discovery and mechanism-of-action (MoA) deconvolution[34]. It operates by taking a query gene signature and performing extensive cross-referencing against integrated RNA profiles, proteomic datasets, Perturb-seq data, and curated pathway databases[34]. A Plex search involves inputting a gene set and searching across these categories to identify biological similarities between the novel compound and established biological mechanisms[34].

The AutoPlex platform and analysis are based on a preprint, thus highlighting the preliminary nature of this work. The AutoPlex framework was developed using curated LINCS L1000 and PCL (Perturbation Connectivity Library) databases. The training data covers perturbations documented through 2021-2022, with 2022 as the cutoff date, ensuring that reference

perturbations and queries are processed using consistent computational pipelines. The analysis itself relies on transcriptomic profiles, morphological profiling, and drug compound metadata. AutoPlex makes predictions using a deep neural network (DNN) architecture designed for multi-label classification and pattern matching. Upon input of differential expression data vectors, the model compresses signatures in which treatment conditions and known drugs are co-localized based on biological effect. Then it computes the connectivity between the query and thousands of known drug-target clusters, and if the cellular response to the treatment mirrors that of a known drug, AutoPlex assigns a high probability to that signature. The signatures are unique transcriptomic fingerprints elicited by a compound and are context-dependent, depending on the baseline expression of the target protein in that tissue. AutoPlex test for consensus signatures that persist across multiple cell lines, indicating biological targets. Prediction sensitivity is tied to the broad cell line baseline biology, and if a compound targets a pathway that is silent in a specific cell line, the resulting signature will be noisy and reduce prediction confidence. Thus, AutoPlex utilizes a "weighted consensus" to prioritize targets that appear consistently across cell lines. The model is trained on drugs, primarily FDA-approved small molecules, where the target is validated. The training set includes ~20,000 small molecules and ~5,000 genetic perturbations, which AutoPlex uses to identify a chemical signature that matches a specific protein's absence.

Only moroidin-QLLVWRNH was analyzed to ensure high-fidelity target identification. Although future studies may expand this approach to a comparative library of moroidin analogs for transcriptomic-based structure-activity relationships (SAR), the current investigation focuses on preliminarily resolving the mechanism of action (MoA) for this compound. AutoPlex analysis reveals a multi-pathway stress signature indicative of the transition from adaptive survival to committed cell death. The platform characterizes the data using similarity rankings and network communities.

The analysis identified interconnected activation of apoptotic pathways, mitochondrial stress, and autophagy dysfunction, contributing to multimodal cell death. AutoPlex identified GSK3B, PGAM5, and VDAC2 as central regulatory nodes within cellular stress networks. The involvement of VDAC2 and other mitochondrial outer membrane proteins suggests that moroidin-QLLVWRNH disrupts bioenergetic homeostasis and metabolic regulation. GSDMB indicates the potential for moroidin-QLLVWRNH to engage in pyroptosis pathways in addition to apoptosis. Furthermore, SQSTM1 (p62) and PDCD7 suggest that moroidin-QLLVWRNH may inhibit

autophagic flux, leading to the accumulation of cytotoxic protein aggregates. Collectively, AutoPlex findings indicate that moroidin-QLLVWRNH induces a systemic failure of cellular maintenance mechanisms. Integrating predictive algorithms such as AutoPlex with computational pipelines for natural product drug discovery in plants provides an effective approach to probing and profiling novel therapeutic mechanisms of action in this chemical space, which may be challenging to achieve with traditional synthetic combinatorial chemistry.

Further experiments to follow up on targets identified with AutoPlex should include computational and wet-lab validation. Computationally, performing *in silico* reverse docking with moroidin-QLLVWRNH and the AutoPlex-predicted protein target will identify protein-binding targets. Additionally, conducting molecular dynamics simulations of moroidin-QLLVWRNH in complex with the target will evaluate the binding stability and identify critical H-bonds. In the wet lab, thermal proteome profiling could be performed by treating cell lysates with moroidin-QLLVWRNH and then analyzing them by mass spectrometry to identify heat-induced shifts in thermal stability. Another option is a cellular thermal shift assay to perform a more targeted analysis of the predicted target in living cells. Finally, functional bioassays would test the ability of moroidin-QLLVWRNH to inhibit substrate conversion by the identified proteins and to monitor substrate accumulation following treatment across a range of concentrations and time points. These experiments would help to identify what moroidin-QLLVWRNH binds to and how it modulates the pathway upon binding.

4.5 Discussion

Our transcriptome mining of fused burpitide cyclases utilized most plant species with RNA-seq data in the NCBI SRA and revealed a large diversity of second ring analogs of the stephanotic acid scaffold with the core peptide QLLVWRNH being found in three plant families and therefore representing an interesting candidate for biological evaluation. Moroidin-QLLVWRNH demonstrated selective dose-responsive cytotoxicity against H1437 lung cancer cells and exhibited limited toxicity in non-cancerous fibroblasts. These findings propose moroidin-QLLVWRNH as a structure with a specific mechanism-of-action in non-small cell lung cancer, offering a putative target for rational drug design based on the stephanotic acid scaffold.

Furthermore, our study showed that other stephanotic acid-type burpitides exist in nature which utilize a tryptophan or tyrosine as the C-terminal residue for crosslinking in the second ring. Although glechomanin and mercurialin did not exhibit in vitro dose-responsive cytotoxicity (**Supplementary Fig. 4.58**), their characterized cyclases could increase the number of biosynthetically feasible second-ring analogs of moroidin in the future. The minimal cytotoxicity of glechomanin and mercurialin against H1437 lung adenocarcinoma cells could be due to the deviation from the QLLVW sequence of the moroidin first ring, a 7 aa-core peptide, or no second ring crosslink between the C-terminal residue and the central tryptophan-5-indole like in moroidins. Screening additional stephanotic acid-type cyclases with 8 aa-core peptides and C-terminal tyrosine or tryptophan residues could enable access to these stephanotic acid scaffolds.

Our study expands stephanotic acid-type burpitide chemistry by introduction of two bicyclic scaffolds and we uncover a broader chemotaxonomic distribution of stephanotic acid-type burpitides (**Figure 4.10**, **Table 4.11**) such as stephanotic acids in pinks, mints, spurges, grasses and Lowiaceae (**Figures 4.22, 4.39, and 4.59**); moroidins in pinks, grasses and Lowiaceae (**Figure. 4.4c, f, and Supplementary Figs. 4.54-57**); glechomanins in mints and Lowiaceae (**Figure. 4.2b, f),** and mercurialins in spurges (**Figure. 4.3b**). The Val-Trp- and Phe-Tyr-crosslinks of glechomanin and mercurialin represent new post-translational modifications in RiPP biosynthesis.

C-C-crosslinks between amino acid side chains, such as the Leu-Trp-bond in the stephanotic acid scaffold and the Val-Trp-bond in glechomanin, are a common occurrence in cyclic peptide natural products. Glechomanin highlights that burpitide C-C-crosslinks usually include an unactivated carbon[4,6,7] which is similar to C-C-macrocyclic bonds in bacterial radical SAM-derived RiPPs[35–39]. In contrast, bacterial CYP450-peptide cyclases generate peptide macrocycles largely by C(sp^2)-C(sp^2)-bonds[40–47], if C-C-crosslinks are formed, with few C(sp^3)-C(sp^2)-exceptions[47,48] (**Figure 4.60**). Analogous to moroidin[4], a biosynthetic proposal for glechomanin and mercurialin peptides in *G. hederacea* and *M. annua* could be established based on transient expression studies of GheBURP and ManBURP in *N. benthamiana*, as well as in vitro reconstitution of GheBURP and ManBURP (**Figure 4.61**).

Initially, GheBURP is translated by the ribosome, yielding the precursor peptide with an N-terminal domain containing three repeated core peptides for glechomanin. Subsequently, the BURP domain catalyzes the bicyclization of the core peptides within the N-terminal domain,

serving as a substrate, in the presence of copper ions. The modified N-terminus will then undergo proteolytic cleavage by endopeptidases, resulting in the formation of a glechomanin analog with an N-terminal glutamine. This residue can cyclize either spontaneously or through a glutamine cyclotransferase to pyroglutamate[5]. Finally, exopeptidase cleavage facilitates the maturation of the C-terminus of glechomanin. Mercurialin biosynthesis occurs most likely through a similar mechanism from ManBURP in *M. annua* which features the formation of a cyclopeptide alkaloid crosslink, i.e., a tyrosine-phenol ether bond, after formation of the stephanotic acid scaffold.

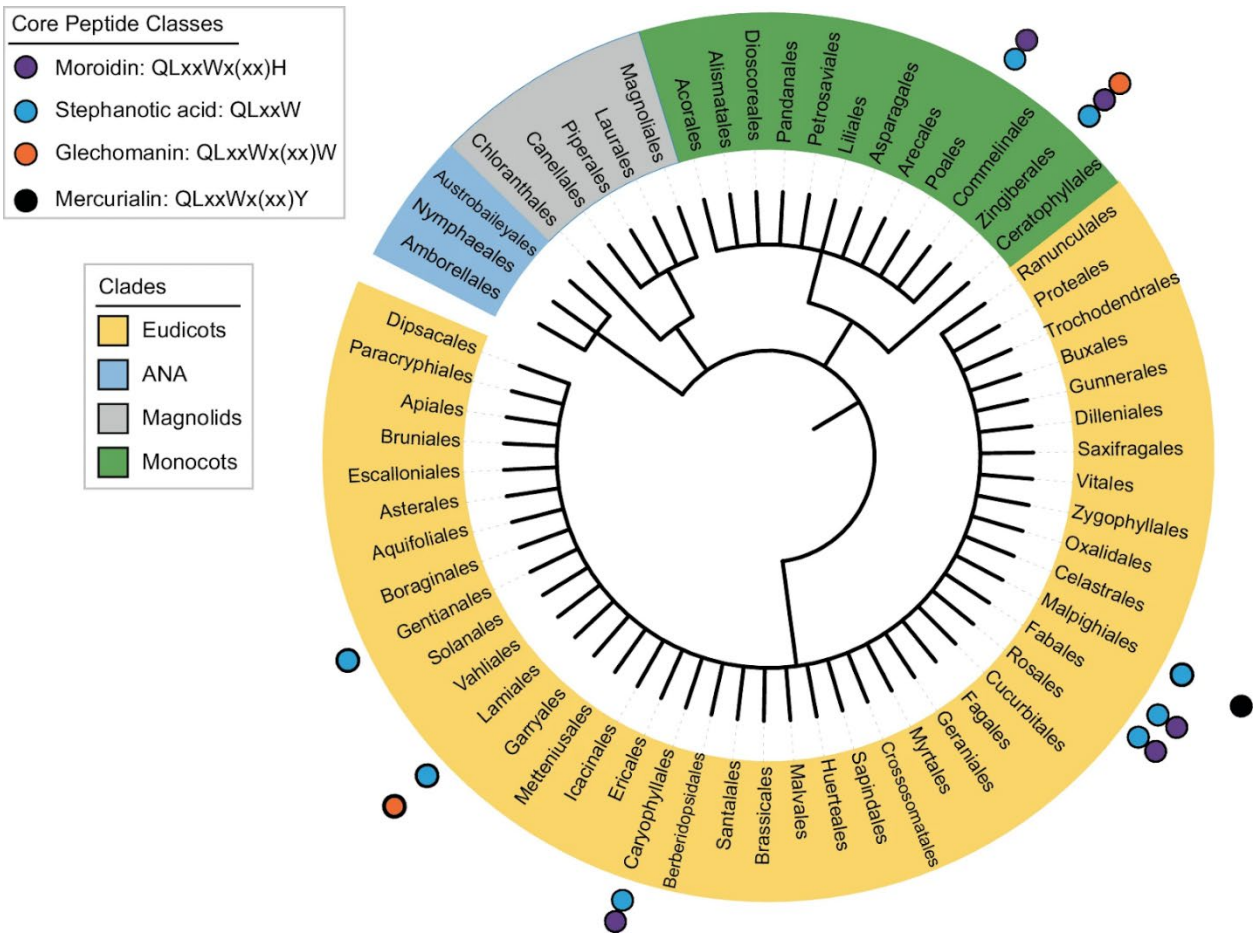


Figure 4.10 Phylogenetic distribution of characterized stephanotic acid-type burpitides in flowering plants. Displayed phylogenetic resolution are plant orders. See Table 4.11 for species information. ANA: Amborellales, Nymphaeales, and Austrobaileyales clade; x: any proteinogenic amino acid.

Furthermore, our study can inspire next steps in development of stephanotic acid-type burpitides through discovery of a bioactive moroidin analog for mechanism-of-action studies and moroidin cyclases for further second-ring diversification.

This study presents a preliminary analysis of transcriptomic RNA-seq data from H1437 non-small cell lung cancer cells treated with the potent moroidin analogue Moroidin-QLLVWRNH. The R packages EnhancedVolcano and ComplexHeatmap were essential for generating volcano plots and heatmaps, and for managing the high-dimensional data produced by DESeq2[29–31]. These tools enabled the identification of statistically significant transcripts based on log2-fold change and p-value distributions, as well as biological replicate clustering. EnTAP, a functional annotation tool, assigned biological meaning to significant transcripts by cross-referencing multiple databases (UniProt/SwissProt), thereby enhancing mechanistic insights[33].

AutoPlex was employed to transition from a list of differentially expressed genes (DEGs) toward elucidating a more definitive mechanism of action (MoA)[34]. AutoPlex, a knowledge-graph search engine platform based on integrated systems biology, evaluates the signature of Moroidin-QLLVWRNH and cross-references it against an extensive library of experimental perturbations[34]. The platform facilitates the identification of transcriptional differences and synthesizes data to support hypothesis generation, enabling rapid deconvolution of the drug's effects. This approach identified a central regulatory node cluster comprising GSK3B, GSDMB, PGAM5, and VDAC2 in the IC_{50} 6-hour dataset.

AutoPlex analysis across nine pairwise comparisons reveals a time-dependent transition in the drug's mechanism. At the early 3-hour mark, adaptive stress responses are observed in IC_{25} and IC_{50} concentrations through HSPA5-mediated endoplasmic reticulum stress and DNA damage checkpoints. By the 6-hour mark at IC_{50} and IC_{75} concentrations, evidence indicates multi-pathway stress. AutoPlex also identifies overlaps with established chemotherapeutics. At the IC_{25} 6-hour mark, Moroidin-QLLVWRNH exhibits similarities to DNA-damaging agents such as Doxorubicin, Etoposide, and Cyclophosphamide[49–51]. The platform further identifies the proteotoxic potential of Moroidin-QLLVWRNH by disrupting proteostasis and epigenetic regulation. At the IC25 24-hour mark, the drug profile shares features with the protease inhibitor Carfilzomib and the HDAC inhibitor Vorinostat[52,53]. AutoPlex uses similarity ranks; a rank greater than 0.6 indicates high-confidence biological overlap. These preliminary findings suggest that Moroidin-QLLVWRNH is a multimodal agent affecting DNA damage, proteasome inhibition, and mitochondrial dysfunction, potentially preventing H1437 cells from relying on a single resistance mechanism. These insights support the potential of Moroidin-QLLVWRNH as a robust peptide for non-small cell lung cancer and highlight areas for future investigation into its mechanism of action.

4.6 Methods and Materials

4.6.1 Materials

All chemicals and reagents were from Thermo Fisher Scientific Inc. if not otherwise noted. All LC-MS solvents were Optima LC-MS-grade solvents (Fisher Scientific Inc.). All synthetic gene fragments were from Twist Biosciences Inc. Deoxynucleotide primers were from Integrated DNA Technologies Inc. LC-MS analysis was performed on a Thermo QExactive Orbitrap mass spectrometer coupled to a Vanquish Ultra high performance liquid chromatography instrument. LC-MS data was analyzed with QualBrowser from Thermo Xcalibur software package (v4.3.73.11, Thermo Scientific). Preparative and semi-preparative high performance liquid chromatography (HPLC) was performed on a Shimadzu HPLC with two binary LC20-AP pumps, a DGU-403 Degasser, an SPD-20A ultraviolet-visible light detector, and a FRC-10A fraction collector. NMR analysis was done on a Bruker Ascend 800 MHz NMR spectrometer with a 5 mm Triple Resonance Inverse Detection TCI CryoProbe. NMR data was analyzed with MestReNova (v14.0.0, Mestrelab Research).

4.6.2 Plant Cultivation

Nicotiana benthamiana and *Glechoma hederacea* were grown from seeds in Premier Horticulture Pro Mix HP High Porosity with Mycorrhizae for plant growth under plant growth lights with a 16 h light/8 h dark cycle for 4–12 weeks at room temperature. *N. benthamiana* seedlings were transferred once to larger pots with the same soil and added Osmocote fertilizer. *G. hederacea* seeds were obtained from the Seed Corner Inc. (Cat #: GROUNDIVYS100). *Sacciolepis striata* was collected from a sample in the University of Michigan Herbarium (Cat # 1423220). *Glechoma hederacea* for glechomanin isolation was collected from the University of Michigan Arboretum (42°16'32.7“N 83°43'27.2“W, June 2022/2023). *Stellaria aquatica* for moroidin-QLLVWRNH isolation and RNA isolation was collected from Pointe Mouillee State

Game Area (41°59'54.8“N 83°11'47.7“W) with Department of Natural Resources permit 24–25 (June 2024). *Mercurialis annua* was collected from a sample in the University of Michigan Herbarium (Cat # 1642848). *Orchidantha maxillarioides* (accession 1989-4177-1) for peptide chemotyping and RNA isolation was collected from the Missouri Botanical Garden (June 2024). *Bauhinia forticata* for stephanotic acid-QLLVW chemotyping was purchased from Plant Delights Nursery Inc. *Kerria japonica* for stephanotic acid-QLLVW chemotyping was purchased as a mature plant from Green Promise Farms™.

4.6.3 Peptide Chemotyping

Leaf and stem plant material (0.2 g fresh weight) was collected from aerial tissues, frozen and ground with a MP Biomedicals FastPrep-24-5G Tissuelyser in 2-mL MP Biomedicals tubes with 2.3-mm Zirconia beads for 60 s at 6 m/s. 1 mL of 80% methanol was added to ground tissue and ground for 20 s at 6 m/s. The methanolic extract was incubated for 10 min in a 60 °C water bath, centrifuged for 5 min at 16,000 × g at room temperature, and filtered through Whatman 0.2 µm syringeless LC-MS filters (UN503NPEORG) before LC-MS analysis. Filtered extracts were transferred to LC-MS vials and analyzed by LC-MS/MS with the following settings: LC-Phenomenex Kinetex® 2.6 µm C18 reverse phase 100 Å 150 × 3 mm LC column; LC gradient, solvent A, 0.1% formic acid; solvent B, acetonitrile (0.1% formic acid); 0 min, 10% B; 5 min, 60% B; 5.1 min, 95% B; 6 min, 95% B; 6.1 min, 10% B; 9.9 min, 10% B; 0.5 mL/min, 30 °C; MS, positive ion mode; full MS, resolution 70,000; mass range 400–1200 m/z; dd-MS2 (data-dependent MS/MS), resolution 17,500; AGC target 1 × 105, loop count 5, isolation width 1.0 m/z, collision energy 25 eV and dynamic exclusion 0.5 s. Sample numbers were n = 3 (biological replicates) for each source plant extraction and each transient gene expression experiment in *N. benthamiana* and n = 1 for *Mercurialis annua* chemotyping from Herbarium specimen. The negative control of transient gene expression experiment in *N. benthamiana* was an empty pEAQ-HT vector infiltration experiment.

4.6.4 Glechomanin Isolation

Mature *Glechoma hederacea* (root, leaves, stem, flowers, 10 kg total) was collected and frozen at −80 °C and ground with dry ice in a food processor to a fine powder. The plant powder was extracted with 1 L methanol (HPLC grade) per 500 g plant powder for 16 h at 37 °C at 160 rpm. The crude extract was vacuum-filtered through a silica filter and then dried in vacuo. The dried methanol extract was resuspended in 2 L of water. The resuspension was partitioned 1:1 (v/v) with hexanes (3x) and ethyl acetate (2x) and then extracted with n-butanol (1:1, v/v, 2x). The n-butanol extract was dried in vacuo, resuspended in 20% methanol and applied to a Sephadex LH20 (Sigma-Aldrich, LH20100) liquid chromatography (LC) solid phase column (60 × 8 cm) with a mobile phase gradient of 20%, 40%, 60%, 80% and 100% methanol (500 mL each). LH20 LC fractions were analyzed for the presence of the target analyte with the following LC-MS settings: injection volume 2 µL; LC, Phenomenex Kinetex 2.6 µm C18 reverse phase 100 Å 50 × 3 mm LC column; LC gradient. solvent A, 0.1% formic acid; solvent B, acetonitrile (0.1% formic acid); 0 min, 5% B; 2.5 min, 95% B; 3.0 min, 95% B; 3.1 min, 5% B; 5.0 min, 5% B, 0.5 mL/min, 30 °C; MS, positive ion mode; full MS, resolution 35,000, mass range 400–1200 m/z; dd-MS2, resolution 17,500; loop count 5, collision energy 25 eV and dynamic exclusion 0.5 s. LH20 LC fractions with glechomanin were combined and dried *in vacuo*.

Glechomanin extracts were separated by preparative HPLC with the following LC conditions in two rounds: 1. separation: Phenomenex Kinetex® 5 µm C18 100 Å LC Column 150 × 21.2 mm, LC gradients: solvent A - 0.1% TFA, solvent B - acetonitrile (0.1% TFA), 7.5 mL/min, 1. separation: 0 min: 10% B, 1 min: 10% B, 36 min: 50% B, 39 min: 95% B, 42 min: 95% B, 42.5 min: 10% B, 60.1 min: 10% B. 2. separation: Phenomenex Kinetex® 5 µm C18 100 Å LC Column 150 × 21.2 mm, 7.5 mL/min, 0 min: 15% B, 1 min: 15% B, 36 min: 35% B, 39 min: 95% B, 42 min: 95% B, 42.5 min: 15% B, 60.1 min: 15% B. Preparative HPLC fractions were analyzed for glechomanin described above. Subsequently, glechomanin fractions were combined, dried in vacuo and dissolved in 20% acetonitrile. The glechomanin fraction was separated by semipreparative HPLC with the following conditions: 1. separation: Kinetex® 5 µm C18 100 Å 250 × 10 mm column, solvent A - 0.1% TFA, solvent B - acetonitrile (0.1% TFA), 1.5 mL/min, 0 min - 20% B, 1 min - 20% B, 21 min - 30% B, 22 min - 95% B, 25 min - 95% B, 25.1 min - 20% B, 45 min - 20% B. 2. separation: solvent A - 0.1% TFA, solvent B - methanol (0.1% TFA), 37 min - 65% B, 47 min - 65% B. Subsequently, glechomanin fractions were combined and dried *in vacuo*. Glechomanin (2 mg) was a white powder.

4.6.5 Moroidin-QLLVWRNH isolation

Stellaria aquatica aerial tissue (1.5 kg) was frozen at −80 °C and ground in a food processor with dry ice and extracted in 6 L of methanol for 16 h at 37 °C with agitation (160 rpm). The methanol extract was filtered over silica and dried under reduced pressure. The dried extract was resuspended in 2 L of water and partitioned sequentially with 2 L of hexanes (3x), ethyl acetate (2x), and n-butanol (2x). The n-butanol fraction was dried, resuspended in 10% methanol, and subjected to Sephadex LH20 chromatography (60 × 8 cm) using a methanol gradient (10%, 20%, 30%, 40%, 50%; 500 mL each). Fractions were analyzed by LC-MS with the following parameters: injection volume, 2 µL; column, Phenomenex Kinetex® 2.6 µm C18 100 Å (50 × 3 mm); solvent A, 0.1% formic acid in water; solvent B, acetonitrile (0.1% formic acid); 0.5 mL/min, 30 °C; gradient: 0 min, 5% B; 2.5 min, 95% B; 3.0 min, 95% B; 3.1 min, 5% B; 5.0 min, 5% B. MS settings: positive ion mode; full MS resolution, 35,000; mass range, 400–1200 m/z; dd-MS2 resolution, 17,500; collision energy, 25 eV; dynamic exclusion, 0.5 s. Moroidin-QLLVWRNH-containing fractions were combined, dried in vacuo, and then separated by preparative HPLC with the following LC conditions in two rounds: Phenomenex Kinetex® 5 µm C18 100 Å LC column 150 × 21.2 mm; 1. separation: LC gradients: solvent A - 0.1% TFA in water, solvent B - acetonitrile (0.1% TFA), 7.5 mL/min, 1. separation: 0 min: 10% B, 1 min: 10% B, 36 min: 50% B, 39 min: 95% B, 42 min: 95% B, 42.5 min: 10% B, 60.1 min: 10% B. 2. separation: 7.5 mL/min, 0 min: 15% B, 1 min: 15% B, 36 min: 35% B, 39 min: 95% B, 42 min: 95% B, 42.5 min: 15% B, 60.1 min: 15% B. Fractions were analyzed for moroidin-QLLVWRNH by LC-MS as for LH20 LC and combined. Final purification was achieved by semipreparative HPLC (Kinetex® 5 µm C18 100 Å 250 × 10 mm, solvent A: 0.1% TFA in water, solvent B: acetonitrile, 0.1% TFA, 1.5 mL/min: 1. Separation: 0 min - 20% B, 1 min - 23% B, 45 min - 23% B), 2. Separation: 0 min - 23% B, 30 min - 23% B. Moroidin-QLLVWRNH (32 mg) was obtained as a white powder.

4.6.6 Mercurialin Isolation

Transgenic tobacco leaves expressing ManBURP-1xQLFFWRY (1.3 kg) were frozen at −80 °C and ground in a food processor with dry ice and extracted in 8 L of methanol for 16 h at 37 °C with agitation (160 rpm). The methanol extract was filtered over silica and dried *in vacuo*. The dried extract was resuspended in 2 L of water and partitioned sequentially with 2 L of hexane (3x), ethyl acetate (2x), and n-butanol (2x). The n-butanol fraction was dried, resuspended in 10% methanol, and subjected to Sephadex LH20 chromatography (60 × 8 cm) using a methanol gradient (20%, 40%, 60%, 80%, 100%; 500 mL each). Fractions were analyzed by LC-MS with the following parameters: injection volume, 2 µL; column, Phenomenex Kinetex® 2.6 µm C18 100 Å (50 × 3 mm); solvent A, 0.1% formic acid in water; solvent B, acetonitrile (0.1% formic acid); 0.5 mL/min, 30 °C; gradient: 0 min, 5% B; 2.5 min, 95% B; 3.0 min, 95% B; 3.1 min, 5% B; 5.0 min, 5% B. MS settings: positive ion mode; full MS resolution, 35,000; mass range, 400–1200 *m/z*; dd-MS^2 resolution, 17,500; collision energy, 25 eV; dynamic exclusion, 0.5 s. Mercurialin-containing fractions were combined, dried *in vacuo*, and then separated by preparative HPLC with the following LC conditions in two rounds: Phenomenex Kinetex® 5 µm C18 100 Å LC Column 150 × 21.2 mm, solvent A - 0.1% TFA in water, solvent B - acetonitrile (0.1% TFA); 1. separation: LC gradients:, 7.5 mL/min, 1. separation: 0 min: 10% B, 1 min: 10% B, 36 min: 50% B, 39 min: 95% B, 42 min: 95% B, 42.5 min: 10% B, 60.1 min: 10% B. 2. separation: 7.5 mL/min, 0 min: 20% B, 1 min: 20% B, 36 min: 40% B, 39 min: 95% B, 42 min: 95% B, 42.5 min: 20% B, 60.1 min: 20% B. Fractions were analyzed for mercurialin by LC-MS and combined. Final purification was achieved by 2 rounds of semi-preparative HPLC (Kinetex® 5 µm C18 100 Å 250 x 10 mm column, solvent A: 0.1% TFA in water, solvent B: acetonitrile, 0.1% TFA, 1.5 mL/min; 1. and 2. separation: 0 min - 38% B, 35 min - 38% B. Mercurialin (5 mg) was obtained as a white powder.

4.6.7 Structure elucidation of glechomanin and mercurialin

For 1D and 2D NMR analysis, glechomanin was dissolved in DMSO-d_6 or methanol-d_4 (Cambridge Isotope Laboratories, Inc.), whereas mercurialin and moroidin-QLLVWRNH were prepared in DMSO-d_6. All data were acquired using corresponding Shigemi NMR tubes (Wilson Glass). To compare ROESY/NOESY data with geometry-optimized structures of glechomanin and

mercurialin, geometry optimizations were performed for the Trp7-Cα stereoisomers of glechomanin and the Phe-Cβ stereoisomers of mercurialin. Initially, the four stereoisomer structures were energy-minimized using MM2 followed by MMFF94 calculations in Chem3D (v.22.2.0, Revvity Signals, maximum iterations set to 10,000). The resulting conformers were subsequently geometry-optimized using the B3LYP-D3/6-31 + G(d,p) method in Gaussian 16 (revision A.03)[54] on the Great Lakes HPC cluster and analyzed in Chem3D.

For Marfey's analysis, ~400 μg of purified glechomanin or mercurialin were hydrolyzed in 600 μL of 6 M hydrochloric acid (HCl) in a sealed thick-walled reaction vessel. The sample was heated at 110 °C and stirred overnight for 12 h. Subsequently, the hydrolysate was concentrated to dryness under nitrogen gas and redissolved in 100 μL water. Then, 100 μL of 1 M sodium bicarbonate ($NaHCO_3$) and 100 μL of a 1% acetone solution of Marfey's reagent, 1-fluoro-2,4-dinitrophenyl-5-l-alanine amide (l-FDAA), were added to the solution. The reaction mixture was incubated at 40 °C for 1 h and then quenched by adding 100 μL of 1 M HCl. For the preparation of l-FDAA-amino acid standard derivatives, 50 mM of each amino acid (d/l-Phe, d/l-Glu, d/l-Leu, d/l-Trp, d/l-Val, d/l-Phe, d/l-Arg and d/l-Tyr) dissolved in water (50 μL) was treated with 1 M $NaHCO_3$ (20 μL) and 1% l-FDAA (100 μL) at 40 °C for 1 h, respectively. After the reaction, the solution was quenched with 1 M HCl (20 μL) and diluted with acetonitrile (810 μL) for LC-MS analysis. LC-MS/MS analysis parameters were as follows: injection volume 2 μL; LC, Phenomenex Kinetex 2.6 μm C18 reverse phase 100 Å 150 × 3 mm LC column; LC gradient: solvent A, 0.1% formic acid; solvent B, acetonitrile (0.1% formic acid); 0 min, 10% B; 40 min, 60% B; 41 min, 95% B; 44 min, 95% B; 45 min, 10% B, 50 min, 10% B, 0.5 mL/min, 30 °C; MS, positive ion mode; full MS, resolution 70,000, mass range 100–1000 *m/z*. For the determination of d/l-Arg, the following LC-MS/MS parameters were employed: injection volume 5 μL; LC, Higgins Analytical PROTO300 C4 5 μm 250 × 4.6 mm LC column; LC gradient: solvent A, 0.1% formic acid; solvent B, acetonitrile (0.1% formic acid); 0 min, 10% B; 38 min, 20% B; 39 min, 95% B; 45 min, 95% B; 46 min, 10% B, 56 min, 10% B, 0.5 mL/min, 30 °C; MS, positive ion mode; full MS, resolution 70,000, mass range 100–1000 *m/z*. The configurations of the amino acid residues were determined by comparing their retention times and elution orders with those of FDAA derivatives of amino acid standards.

4.6.8 BURP-domain gene cloning

Leaves of 10-week-old *Glechoma hederacea*, stem tissue of *Stellaria aquatica*, and leaves of *Orchidantha maxillarioides* were collected and total RNA was extracted from each plant with QIAGEN RNeasy Plant Mini Kit. Complementary DNA was prepared from *G. hederacea* leaf total RNA, *S. aquatica* stem total RNA, and *O. maxillarioides* leaf total RNA with Thermo Scientific™ Maxima H Minus First Strand cDNA Synthesis Kit. Cloning primers for *GheBURP*, *SaqBURP, and OmaBURP1/2* were designed from candidate *GheBURP* transcripts, *SaqBURP* transcripts, and *OmaBURP1/2* transcripts, respectively (**Table 4.12**). GheBURP and OmaBURP1/2 were amplified with Phusion High-Fidelity DNA polymerase (New England Biolabs) and cloned by Gibson cloning assembly[55] (New England Biolabs) into pEAQ-HT[17] linearized by restriction enzymes AgeI and XhoI. Cloned *GheBURP* were sequenced by Sanger sequencing. Similarly, *SaqBURP* was cloned into pHis8 (pET28a with N-terminal 8xHis-tag) linearized by restriction enzyme NcoI. Cloned *SaqBURP* was sequenced by Nanopore sequencing (Plasmidsaurus Inc).

4.6.9 Transient gene expression in Nicotiana benthamiana

SstBURP-1xQLLVWRNH, *SstBURP-4xQLLVWRNH, SaqBURP-1xQLLVWRNH, OmaBURP1-1xQLFVWGW, ManBURP-1xQLFFWRY, ManBURP-1xQLLVWRY*, *GheBURP-1xQLLVW, KjaBURP-1xQLLVWRGH*, *KjaBURP-1xQLLVWRNH, KjaBURP-1xQLLVWRSH*, and *KjaBURP-1xQLLVWPKH* were synthesized with N-terminal overhang TGCCCAAATTCGCGACCGGT and C-terminal overhang CTCGAGGCCTTTAACTCTGG for Gibson cloning into pEAQ-HT[17] linearized by restriction enzymes AgeI and XhoI. Agrobacterium tumefaciens LBA4404 was transformed with pEAQ-HT-GheBURP wildtype or synthetic constructs (GheBURP-1xQLLVW), pEAQ-HT-KjaBURP synthetic constructs (KjaBURP-1xQLLVWRNH, KjaBURP-1xQLLVWRSH, KjaBURP-1xQLLVWPKH), pEAQ-HT-SstBURP synthetic constructs (SstBURP-1xQLLVWRNH, SstBURP-4xQLLVWRNH), pEAQ-HT-SaqBURP synthetic construct (SaqBURP-1xQLLVWRNH), pEAQ-HT-OmaBURP1 wild-type and synthetic constructs (OmaBURP1-1xQLFVWGW), pEAQ-HT-OmaBURP2 wildtype

construct and pEAQ-HT-ManBURP wildtype and synthetic constructs (ManBURP-1xQLFFWRY, ManBURP-1xQLLVWRY) by electroporation (2.5 kV) and then plated on YM agar plates (0.4 g of yeast extract, 10 g of d-mannitol, 0.1 g of sodium chloride (NaCl), 0.2 g of magnesium sulfate (heptahydrate), 0.5 g of potassium phosphate (dibasic, trihydrate), 15 g of agar, and 1 L of deionized water, adjusted to pH 7) with 100 µg/mL rifampicin, 50 µg/mL kanamycin, and 100 µg/mL streptomycin, and incubated for 2 days at 30 °C. A 5 mL starter culture of the YM medium with 100 µg/mL rifampicin, 50 µg/mL kanamycin, and 100 µg/mL streptomycin was inoculated with a clone of A. tumefaciens LBA4404 pEAQ-HT-GheBURP or other precursor gene constructs and incubated for 24–36 h at 30 °C in a shaking incubator at 220 rpm. Subsequently, the starter culture was used to inoculate a 25 mL culture of the YM medium with 100 µg/mL rifampicin, 50 µg/mL kanamycin, and 100 µg/mL streptomycin, which was incubated for 24 h at 30 °C in a shaking incubator at 220 rpm. The cells from the 25 mL culture were centrifuged for 30 min at 3000 × g, the YM medium was discarded, and cells were resuspended in MMA medium (10 mM MES KOH buffer (pH 5.6), 10 mM magnesium chloride (MgCl2), 100 µM acetosyringone) to give a final optical density (OD600) of 0.8. The Agrobacterium suspension was infiltrated into the bottom of the leaves of N. benthamiana plants (6-8 weeks old) using 1 mL plastic syringes. N. benthamiana plants were placed in the shade 2 h before syringe infiltration. After infiltration, N. benthamiana plants were grown as described above for 6 days. Subsequently, infiltrated leaves were collected and subjected to peptide chemotyping. For peptide purification, infiltrated leaves were collected after 7 days and stored at -80 °C until purification.

4.6.10 Protein purification

E. coli-codon-optimized *GheBURP-1xQLFVWGW* and *ManBURP-1xQLFFWRY* were cloned via Gibson cloning into NcoI-linearized pHis8 expression vector (pET28a with an N-terminal 8xHis-tag) with 5'-overhang GAAAACTTGTACTTCCAGGCCCATGGC and 3'-overhang CCATGGCGGATCCGAATTCGAG. They were transformed into E. coli BL21(DE3) respectively. Each colony of E. coli BL21(DE3) was inoculated to a 10 mL Luria-Bertani medium with 50 µg/mL kanamycin for 16 h incubation at 37 °C and 200 rpm. The starter culture was used to inoculate 1 L Terrific Broth liquid medium with 50 µg/mL kanamycin and the culture was incubated at 37 °C and 200 rpm until OD600 reached 0.7-1. The culture was then added with

Isopropyl-β-d-1-thiogalactopyranoside to a final concentration of 50 µg/mL and cooled to 18 °C for incubation at 200 rpm for 16 h.

Cells were centrifugated at 3500 × g and 4 °C. The cell pellet was resuspended in solution buffer (50 mM Tris-hydrochloride (Tris-HCl) (pH 8), 1 mM ethylenediaminetetraacetic acid (EDTA, pH 8), 1 mM dithiothreitol (DTT), 25% sucrose (w/v)) with 5 mL per g wet cell pellet. Resuspended cells were sonicated on ice to break cells open for 2 min 57 sec at 80% amplitude (59 s on, 59 s off). Subsequently, 1 mg lysozyme and 1 mg DNAse I is added per mL lysate and 25 µL divalent cation solution (400 mM MgCl2, 400 mM calcium chloride (CaCl2)) is added per mL lysate. The lysate was then stirred at room temperature for 20 min. Next, the same volume of lysis buffer (50 mM Tris-HCl (pH 8), 1 mM EDTA (pH 8), 1 mM DTT, 200 mM NaCl, 1.0 % Triton X-100 was added to the lysate and mixed for 1 h at 4 °C. The lysate was sonicated again on ice for 2 min 57 sec and 60% amplitude (59 s on, 59 s off) and then added with EDTA solution (pH 8, final concentration 20 mM). The inclusion bodies were subsequently checked for full resuspension. The resulting lysate was centrifugated for 30 min at 6000 × g and 4 °C. The supernatant was discarded and the inclusion bodies were resuspended in the same volume as the discarded supernatant of Wash Buffer with Triton X-100 (50 mM Tris-HCl (pH 8), 1 mM EDTA (pH 8), 1 mM DTT, 100 mM NaCl, 0.5% Triton X-100). The inclusion body resuspension was sonicated on ice for 2 min 57 sec and 60% amplitude (59 s on, 59 s off). The inclusion bodies are checked for full resuspension and centrifugated for 30 min at 6000 × g and 4 °C. The inclusion bodies were washed one more time with Wash Buffer with Triton X-100, before the inclusion bodies were washed twice with Wash Buffer without Triton X-100 (50 mM Tris-HCl (pH 8), 1 mM EDTA (pH 8), 1 mM DTT, 100 mM NaCl). After the final centrifugation, the supernatant was discarded and the inclusion body pellet was resuspended in TE buffer (10 mM Tris-HCl (pH 8), 1 mM EDTA (pH 8)) at 5 mL per g wet inclusion bodies and then aliquoted as 1 mL inclusion body resuspension for storage at -80 °C before further protein purification.

Unless otherwise specified, all subsequent procedures were conducted at room temperature. The denaturing solution was prepared in a 15 mL conical tube by mixing solid urea with 4 mL of 50 mM Tris-HCl (pH 8) and 100 mM NaCl, achieving a final concentration of 7.5–8.0 M, followed by the addition of 8 mM β-mercaptoethanol. Thawed inclusion bodies (1 mL aliquot) were then added to the denaturing solution and gently mixed. The mixture was inverted or pipetted every 10 min for 60 min until complete dissolution was achieved. Subsequently, the

solution was filtered through a 0.45 µm syringe filter into a conical tube. The dissolved and filtered GheBURP-1xQLFVWGW/ManBURP-1xQLFFWRY was diluted into 500 mL of 50 mM Tris-HCl (pH 8), 500 mM NaCl, and 2.5 M urea to prevent aggregation during refolding. Before use, a 4 L stock of refolding buffer (50 mM Tris-HCl pH 8, 500 mM NaCl, 0.5 mM tris(2-carboxyethyl)phosphine (GoldBio), 10% glycerol) was cooled to 4 °C. The diluted GheBURP-1xQLFVWGW/ManBURP-1xQLFFWRY (each 500 mL) was then introduced into 3.5 kDa molecular weight cut-off (MWCO) dialysis tubing and dialyzed overnight at 4 °C in 4 L of refolding buffer. Following dialysis, the protein was removed from the tubing and centrifuged at 4000 × g at 4 °C for 30 min to eliminate aggregates prior to purification. Subsequently, 2.5 mL of Ni-NTA resin was added to a 20 mL column, washed with deionized water, and equilibrated with 10 column volumes (CV) of refolding buffer containing 30 mM imidazole at pH 8. The refolded protein was then applied onto the resin, and the flow-through was discarded. The resin was washed with 10 CV of refolding buffer containing 30 mM imidazole at pH 8, followed by elution of the protein with 10 CV of refolding buffer containing 600 mM imidazole at pH 8. A mixture containing 1 mg of TEV protease for every 10 mg of eluted protein was subjected to overnight dialysis in 3.5 kDa MWCO dialysis tubing at 4 °C with stirring in 4 L of TEV protease buffer (50 mM Tris-HCl (pH 8), 100 mM NaCl, 3 mM reduced glutathione (GSH), 0.3 mM oxidized glutathione (GSSG)). Subsequently, GheBURP-1xQLFVWGW/ManBURP-1xQLFFWRY was separated from uncleaved His8-GheBURP-1xQLFVWGW/His8-ManBURP-1xQLFFWRY, respectively, and the His-tagged TEV protease using a second 2.0 mL Ni-NTA column, which had been pre-equilibrated with 10 CV of TEV protease buffer in a disposable 20 mL column. The flow-through was then concentrated using Amicon Ultra-15mL 10 kDa MWCO centrifugal concentrators, subjected to SDS-PAGE analysis, and utilized for enzyme assays.

4.6.11 Enzyme assays

For core peptide analysis, enzyme assays were performed using 50 µL volumes containing either copper(II) chloride ($CuCl_2$) from JT Baker Chemical Co. or water. Assays for BURP-domain proteins comprised 300 µM GheBURP-1xQLFVWGW/100 µM ManBURP-1xQLFFWRY, 1 mM $CuCl_2$, and a citrate:phosphate buffer (pH 7) with final concentrations of 18 mM citrate and 165 mM Na_2HPO_4. The enzyme reactions were incubated for 24 h at 20 °C. Subsequently, 1 µg of

trypsin (Sigma-Aldrich, T7575-1KT) in 20 µL of 1 M Tris-HCl (pH 8)/1 µg of LysC (NEB, P8109S) in 10 µL of 10 mM Tris-HCl (pH 8) buffer was added to GheBURP/ManBURP assay, respectively, and further incubated at 37 °C for 24 h. Each digest was centrifuged, added to ChromacolTM LC-MS vial, and subjected to LC–MS/MS analysis. LC-MS/MS settings: Injection volume 15 µL, LC-Higgins Analytical PROTO300 C4 5 µm 250 × 4.6 mm, solvent A-0.1% formic acid, solvent B-acetonitrile (0.1% formic acid), 0.7 mL/min, 30 °C, LC gradient: 0 min: 10% B, 10 min: 50% B, 11 min: 95% B, 13 min: 95% B, 13.1 min: 10% B, 20 min: 10% B, MS - Positive ion mode, Full MS: Resolution 35,000, mass range: 700-1250 m/z for GheBURP digest and 700-1450 m/z for ManBURP digest, dd-MS2: resolution 17,500, AGC target 1e5, loop count 5, isolation window 0.4 m/z, collision energy 25 eV, dynamic exclusion 2 s. Scaled enzyme assays for exoproteolytic cyclic peptide formation were conducted on a 2 mL scale using purified GheBURP-1xQLFVWGW/ManBURP-1xQLFFWRY obtained from a 4 L/2 L expression and purification process. These assays included 130–350 µM of the BURP domain cyclase and 1 mM CuCl2 in a citrate:phosphate buffer at pH 7. The assays were incubated for 24 h at 20 °C. Following incubation, trypsin/LysC was introduced into the GheBURP/ManBURP reaction mixtures at a protease-substrate molar ratio of 1:250/1:100, with the pH adjusted to 8 using a 1 M Tris-HCl (pH 8) buffer and incubated at 37 °C for 24 h. Two digests were subjected to preparative HPLC, respectively (LC settings, Phenomenex Kinetex 5 µm C18 100 Å LC Column 150 x 21.2 mm, 7.5 mL/min; solvent A, 0.1% TFA; solvent B, acetonitrile (0.1% TFA); LC gradient, 0 min, 10% B; 1 min, 10% B; 36 min, 50% B; 39 min, 95% B; 42 min, 95% B; 42.5 min, 10% B; 60.1 min, 10% B). HPLC fractions were analyzed by LC–MS analysis for target masses of modified core peptides with the following LC–MS parameters: injection volume 2 µL; LC, Phenomenex Kinetex 2.6 µm C18 reverse phase 100 Å 50 × 3 mm LC column; LC gradient. solvent A, 0.1% formic acid; solvent B, acetonitrile (0.1% formic acid); 0 min, 5% B; 2.5 min, 95% B; 3.0 min, 95% B; 3.1 min, 5% B; 5.0 min, 5% B, 0.5 mL/min, 30 °C; MS, positive ion mode; full MS, resolution 35,000, mass range 400–1200 m/z; dd-MS2, resolution 17,500; loop count 5, collision energy 25 eV and dynamic exclusion 0.5 s. LC fractions containing target peptides were dried and resuspended in 30 µL DMSO for further analysis. Exopeptidase assays were conducted with ~25–50 µg of peptides (with 4% DMSO), carboxypeptidase Y (from S. cerevisiae, Sigma-Aldrich), and aminopeptidase N (from Rat, Sigma Aldrich) at a peptidase:peptides ratio of 1:10 (w/w). These assays were performed in 100 mM Tris-HCl buffer at pH 7 and 37 °C for 24 h. Following

incubation, the assays were centrifuged and analyzed by LC-MS/MS as follows: Injection volume 15 µL, LC-Higgins Analytical PROTO300 C4 5 µm 250 × 4.6 mm, all gradients: solvent A-0.1% formic acid, solvent B-acetonitrile (0.1% formic acid), 0.7 mL/min, 30 °C, LC gradient: 0 min: 15% B, 60 min: 40% B, 61 min: 95% B, 63 min: 95% B, 63.1 min: 15% B, 80 min: 15% B, MS - Positive ion mode, Full MS: Resolution 35,000, mass range: 800–1300 m/z for GheBURP and 700–1450 m/z for ManBURP, dd-MS2: resolution 17,500, AGC target 1e5, loop count 5, isolation width 0.4 m/z, collision energy 25 eV, dynamic exclusion 0.6 s. The analysis was compared to purified glechomanin or mercurialin standards.

4.6.12 Bioassays

Cancer cell lines H1437 (ATCC® CRL-5872), H1299 (ATCC® CRL-5803), A549 (ATCC® CCL-185), U2OS (ATCC® HTB-96), LNCaP (ATCC® CRL-1740), HUH-7 (JCRB® JCRB0403), and noncancerous cell lines IMR-90 (ATCC® CCL-186), HFF (ATCC® CRL-2088), and MEFs were maintained at 37 °C and 5% CO_2 atmosphere. Standard isolation and culture of primary mouse embryonic fibroblasts from pregnant female mice at the University of Michigan established the MEF cell line[56] H1437, H1299 and LNCaP were grown in Roswell Park Memorial Institute 1640 Medium (Fisher Scientific, 11875119), U2OS were grown in McCoy's 5 A Medium (Thermo Fisher, 16600082), and the remaining lines grown in Dulbecco's Modified Eagle's medium (Fisher Scientific, 11995065). A549 cell media were supplemented with 15% Fetal Bovine Serum (FBS; Thermo Fisher, A5670701), while the remaining media were supplemented with 10% FBS and 1x penicillin-streptomycin (Gibco, 15140122).

384-well plates (Perkin Elmer, 6057300) were seeded with H1437, H1299, A549, U2OS, LNCaP, HUH-7, IMR-90, HFF, and MEF at 5000 cells per well in 50 µL of their respective complete media. Cell seeding densities were optimized such that the cells would be 90–100% confluent in DMSO vehicle control wells after 72 h. Cells were allowed to attach to the wells for 24 h at 37 °C and 5% CO_2. Compounds (10 mM) were added using the HPD300e digital compound dispenser in randomized, ten-point, two-fold dose response ($n = 3$) from a starting concentration of 40 µM. Compounds were normalized to the highest class DMSO volume of 1%, then incubated at 37 °C and 5% CO_2 for 48 h. Each cell line had $n = 16$ positive controls (10 µM taxol) and vehicle

(DMSO). Cells were fixed with 4% paraformaldehyde for 20 min, permeabilized with 0.03% Triton X-100 for 20 min and rinsed twice with PBS between each step. Cells were then stained for 30 min with a 1:1000 PBS dilution of Hoechst 33340 (Thermo Scientific, H1399) for nuclei staining and 1:10,000 PBS dilution of CellMask Orange (Thermo Fisher, H32713) for cell delineation. Fixed and stained plates were washed twice with PBS, then left in a final volume of 25 µL per well for imaging using a Thermo Fisher CX5 high content imaging microscope with LED excitation (386/23 nm, 560/25 nm) at 10x magnification. Cells were imaged at a single z-plane as optimized through image-based autofocus for each channel, and the exposure times were selected to maximize the signal-to-background ratio. Cellpose[57] was used to segment and count cell nuclei for each cell line, a percent viability score for compounds was generated by normalizing the cell count to the average well-level cell counts observed in the negative control (DMSO), and dose-response curves were fit using the GraphPad Prism software (v10.1.0, Dotmatics).

4.6.13 Phylogenetic Tree Construction

The tree was constructed in the web server Interactive Tree of Life[58] displaying angiosperm orders distinguished as ANA grade, Magnolids, Monocots and Eudicots[59].

4.6.14 Dosed Treatment of H1437 Cells with moroidin-QLLVWRNH and RNA-seq analysis

RNA sequencing (RNA-seq) was used to compare gene expressions profiles in peptide-treated versus untreated H1437 cells. 6-well plates were seeded with 300,000 H1437 cells per well in 2 mL of complete RPMI media, with seeding densities optimized to reach 90-100% confluence in control wells after 48 hours. After 24 hours, peptides or the normalized DMSO concentration were added in three-point doses (IC_{75}, IC_{50}, and IC_{25}), followed by incubation. Cells underwent RNA extraction after 3, 6, or 24 hrs. treatment to capture gene expression changes using the Direct-zol RNA Miniprep kit, and RNA quality assessed with 260/280 absorbance ratios (detect protein contamination) or 260/230 ratio (detect contamination with organic compounds)[60]. Yields were >100 ng/mL of eluted RNA in 25-30 µL which was sufficient for RNA-seq, which was outsourced to Genewiz.

4.6.15 Deseq2 for Differential gene expression and heatmap/volcano plot creation in R

Below is an example script used for *Differential gene expression using* DESeq2 in R[29]. Furthermore, it features the script for how to generate bar graph(s) using ggplot, volcano plot(s) using EnhancedVolcano, and the heatmap(s) using ComplexHeatmap[30,31].

```
#DMSO 6hr vs Ic25, ic50, ic75
#Load in the Packages
          if (!requireNamespace("BiocManager", quietly = TRUE))
            +     +     +     +     install.packages("BiocManager", force = TRUE)
          BiocManager::install("tximport", force = TRUE)
          BiocManager::install("DESeq2", force = TRUE)
          BiocManager::install("tximportData", force = TRUE)
          library(tximport)

#Load in the data
filesIC256hrvsDMSO               <-              c("/scratch/rkersten_root/rkersten1/shared_data/MoA_RNAseq/MoroidinMOA/4A.tsv",
"/scratch/rkersten_root/rkersten1/shared_data/MoA_RNAseq/MoroidinMOA/5A.tsv",
"/scratch/rkersten_root/rkersten1/shared_data/MoA_RNAseq/MoroidinMOA/6A.tsv",
"/scratch/rkersten_root/rkersten1/shared_data/MoA_RNAseq/MoroidinMOA/13B.tsv",
"/scratch/rkersten_root/rkersten1/shared_data/MoA_RNAseq/MoroidinMOA/14B.tsv")
filesIC506hrvsDMSO               <-              c("/scratch/rkersten_root/rkersten1/shared_data/MoA_RNAseq/MoroidinMOA/4A.tsv",
"/scratch/rkersten_root/rkersten1/shared_data/MoA_RNAseq/MoroidinMOA/5A.tsv",
"/scratch/rkersten_root/rkersten1/shared_data/MoA_RNAseq/MoroidinMOA/6A.tsv",
"/scratch/rkersten_root/rkersten1/shared_data/MoA_RNAseq/MoroidinMOA/22C.tsv",
"/scratch/rkersten_root/rkersten1/shared_data/MoA_RNAseq/MoroidinMOA/23C.tsv",
"/scratch/rkersten_root/rkersten1/shared_data/MoA_RNAseq/MoroidinMOA/24C.tsv")
          filesIC756hrvsDMSO          <-          c("/scratch/rkersten_root/rkersten1/shared_data/MoA_RNAseq/MoroidinMOA/4A.tsv",
"/scratch/rkersten_root/rkersten1/shared_data/MoA_RNAseq/MoroidinMOA/5A.tsv",
"/scratch/rkersten_root/rkersten1/shared_data/MoA_RNAseq/MoroidinMOA/6A.tsv",
"/scratch/rkersten_root/rkersten1/shared_data/MoA_RNAseq/MoroidinMOA/32D.tsv",
"/scratch/rkersten_root/rkersten1/shared_data/MoA_RNAseq/MoroidinMOA/33D.tsv")

#Name the Files
names(filesIC256hrvsDMSO) <- c("4A-DMSO-6hr", "5A-DMSO-6hr", "6A-DMSO-6hr", "13B-IC25-6hr", "14B-IC25-6hr")
names(filesIC506hrvsDMSO) <- c("4A-DMSO-6hr", "5A-DMSO-6hr", "6A-DMSO-6hr", "22C-IC50-6hr", "23C-IC50-6hr", "24C-IC50-6hr")
names(filesIC756hrvsDMSO) <- c("4A-DMSO-6hr", "5A-DMSO-6hr", "6A-DMSO-6hr", "32D-IC75-6hr", "33D-IC75-6hr")

          #Load in Kallisto Files
          txi.kallistoIC256hrvsDMSO <- tximport(filesIC256hrvsDMSO, type = "kallisto", txOut = TRUE)
          txi.kallistoIC506hrvsDMSO <- tximport(filesIC506hrvsDMSO, type = "kallisto", txOut = TRUE)
          txi.kallistoIC756hrvsDMSO <- tximport(filesIC756hrvsDMSO, type = "kallisto", txOut = TRUE)

          #Name the samples
          sample_namesIC256hrvsDMSO <- c("DMSO-6hr", "DMSO-6hr", "DMSO-6hr", "IC25-6hr", "IC25-6hr")
          sample_namesIC506hrvsDMSO <- c("DMSO-6hr", "DMSO-6hr", "DMSO-6hr", "IC50-6hr", "IC50-6hr", "IC50-6hr")
          sample_namesIC756hrvsDMSO <- c("DMSO-6hr", "DMSO-6hr", "DMSO-6hr", "IC75-6hr", "IC75-6hr")

          # Names the rows
          row_namesIC256hrvsDMSO <- c("4A", "5A", "6A", "13B", "14B")
          row_namesIC506hrvsDMSO <- c("4A", "5A", "6A", "22C", "23C", "24C")
          row_namesIC756hrvsDMSO <- c("4A", "5A", "6A", "32D", "33D")

          #make sure row lenght = sample table lenght
          if(length(row_namesIC256hrvsDMSO) != length(sample_namesIC256hrvsDMSO)) {stop("The number of row names and conditions
must be the same.")}
          if(length(row_namesIC506hrvsDMSO) != length(sample_namesIC506hrvsDMSO)) {stop("The number of row names and conditions
must be the same.")}
          if(length(row_namesIC756hrvsDMSO) != length(sample_namesIC756hrvsDMSO)) {stop("The number of row names and conditions
must be the same.")}

          #make the sample table
          sampleTableIC256hrvsDMSO     <-     data.frame(condition     =     factor(sample_namesIC256hrvsDMSO),     row.names     =
row_namesIC256hrvsDMSO)
          sampleTableIC506hrvsDMSO     <-     data.frame(condition     =     factor(sample_namesIC506hrvsDMSO),     row.names     =
row_namesIC506hrvsDMSO)
          sampleTableIC756hrvsDMSO     <-     data.frame(condition     =     factor(sample_namesIC756hrvsDMSO),     row.names     =
row_namesIC756hrvsDMSO)

          View(sampleTableIC256hrvsDMSO)
          View(sampleTableIC506hrvsDMSO)
          View(sampleTableIC756hrvsDMSO)

          #make count matirx table
          countmatrixIC256hrvsDMSO <- txi.kallistoIC256hrvsDMSO$counts
          countmatrixIC506hrvsDMSO <- txi.kallistoIC506hrvsDMSO$counts
          countmatrixIC756hrvsDMSO <- txi.kallistoIC756hrvsDMSO$counts
```

```
          View(countmatrixIC256hrvsDMSO)
          View(countmatrixIC506hrvsDMSO)
          View(countmatrixIC756hrvsDMSO)

          #make sure the rownames in the count matrix are same
          rownames(countmatrixIC256hrvsDMSO) <- rownames(countmatrixIC256hrvsDMSO)
          rownames(countmatrixIC506hrvsDMSO) <- rownames(countmatrixIC506hrvsDMSO)
          rownames(countmatrixIC756hrvsDMSO) <- rownames(countmatrixIC756hrvsDMSO)

          #round the counts to the nearest 0
          countmatrixIC256hrvsDMSO <- round(countmatrixIC256hrvsDMSO, 0)
          countmatrixIC506hrvsDMSO <- round(countmatrixIC506hrvsDMSO, 0)
          countmatrixIC756hrvsDMSO <- round(countmatrixIC756hrvsDMSO, 0)

          #Load in the package
          library(DESeq2)
          rownames(sampleTableIC256hrvsDMSO) <- colnames(countmatrixIC256hrvsDMSO)
          rownames(sampleTableIC506hrvsDMSO) <- colnames(countmatrixIC506hrvsDMSO)
          rownames(sampleTableIC756hrvsDMSO) <- colnames(countmatrixIC756hrvsDMSO)

          View(countmatrixIC256hrvsDMSO)
          View(countmatrixIC506hrvsDMSO)
          View(countmatrixIC756hrvsDMSO)

          View(sampleTableIC256hrvsDMSO)
          View(sampleTableIC506hrvsDMSO)
          View(sampleTableIC756hrvsDMSO)

          #Build Matrix
          model.matrix(~condition,data=sampleTableIC256hrvsDMSO)
          model.matrix(~condition,data=sampleTableIC506hrvsDMSO)
          model.matrix(~condition,data=sampleTableIC756hrvsDMSO)

          #Set Up Deseq2 to run
          deseq2IC256hrvsDMSO.dds     <-     DESeqDataSetFromMatrix(countData     =     countmatrixIC256hrvsDMSO,     colData     =
sampleTableIC256hrvsDMSO, design = ~ condition)
          deseq2IC506hrvsDMSO.dds     <-     DESeqDataSetFromMatrix(countData     =     countmatrixIC506hrvsDMSO,     colData     =
sampleTableIC506hrvsDMSO, design = ~ condition)
          deseq2IC756hrvsDMSO.dds     <-     DESeqDataSetFromMatrix(countData     =     countmatrixIC756hrvsDMSO,     colData     =
sampleTableIC756hrvsDMSO, design = ~ condition)

          #Run Deseq2
          View(deseq2IC256hrvsDMSO.dds)
          View(deseq2IC506hrvsDMSO.dds)
          View(deseq2IC756hrvsDMSO.dds)

          deseq2IC256hrvsDMSO.dds <- DESeq(deseq2IC256hrvsDMSO.dds)
          deseq2IC506hrvsDMSO.dds <- DESeq(deseq2IC506hrvsDMSO.dds)
          deseq2IC756hrvsDMSO.dds <- DESeq(deseq2IC756hrvsDMSO.dds)

          #Get the Deseq2 Results
          deseq2IC256hrvsDMSO.res <- results(deseq2IC256hrvsDMSO.dds, contrast = c("condition", "IC25-6hr", "DMSO-6hr"))
          deseq2IC506hrvsDMSO.res <- results(deseq2IC506hrvsDMSO.dds, contrast = c("condition", "IC50-6hr", "DMSO-6hr"))
          deseq2IC756hrvsDMSO.res <- results(deseq2IC756hrvsDMSO.dds, contrast = c("condition", "IC75-6hr", "DMSO-6hr"))

          #Check the summary results
          summary(deseq2IC256hrvsDMSO.res)
          summary(deseq2IC506hrvsDMSO.res)
          summary(deseq2IC756hrvsDMSO.res)

          #Filter out the rows with N/A and apply a padjusted filter of 0.05 (5% FDR)
          sigsIC256hrvsDMSO <-na.omit(deseq2IC256hrvsDMSO.res)
          sigsIC506hrvsDMSO <-na.omit(deseq2IC506hrvsDMSO.res)
          sigsIC756hrvsDMSO <-na.omit(deseq2IC756hrvsDMSO.res)

          sigsIC256hrvsDMSO <- sigsIC256hrvsDMSO[sigsIC256hrvsDMSO$padj < 0.05,]
          sigsIC506hrvsDMSO <- sigsIC506hrvsDMSO[sigsIC506hrvsDMSO$padj < 0.05,]
          sigsIC756hrvsDMSO <- sigsIC756hrvsDMSO[sigsIC756hrvsDMSO$padj < 0.05,]

          summary(sigsIC256hrvsDMSO)
          summary(sigsIC506hrvsDMSO)
          summary(sigsIC756hrvsDMSO)

          ###BAR CHARTS

          gene_data <- data.frame(
            regulation = c("Upregulated", "Downregulated"),
            count = c(391, 755)
          )

          # Create the bar chart with enhancements
          ggplot(gene_data, aes(x = regulation, y = count, fill = regulation)) +
            geom_bar(stat = "identity", color = "black", size = 0.7, width = 0.6) +  # Adjust bar width for spacing
            geom_text(aes(label = count), vjust = -0.5, size = 5) +  # Add number counts on top of the bars
            labs(
              title = "Moroidin-RNH IC25 6hrs. vs. DMSO 6hrs. Treatment",
              x = "",
              y = "Gene Count"
            ) +
            scale_fill_manual(values = c("Upregulated" = "skyblue", "Downregulated" = "salmon")) +
            theme_bw() +  # Apply a clean theme
            theme(
              plot.title = element_text(hjust = 0.5, face = "bold"), # Center and bold the title
              legend.position = "none",  # Remove the legend
              panel.grid.major = element_blank(),  # Remove major grid lines
```

```
    panel.grid.minor = element_blank()   # Remove minor grid lines
  ) +
  expand_limits(y = max(gene_data$count) * 1.1)

gene_data_ic50 <- data.frame(
  regulation = c("Upregulated", "Downregulated"),
  count = c(308, 539)  # New counts for IC50
)

# Create the second bar chart
ggplot(gene_data_ic50, aes(x = regulation, y = count, fill = regulation)) +
  geom_bar(stat = "identity", color = "black", size = 0.7, width = 0.6) +  # Adjust bar width for spacing
  geom_text(aes(label = count), vjust = -0.5, size = 5) +  # Add number counts on top of the bars
  labs(
    title = "Moroidin-RNH IC50 6hrs. vs. DMSO 6hrs. Treatment",  # New title
    x = "",
    y = "Gene Count"
  ) +
  scale_fill_manual(values = c("Upregulated" = "skyblue", "Downregulated" = "salmon")) +
  theme_bw() +  # Apply a clean theme
  theme(
    plot.title = element_text(hjust = 0.5, face = "bold"), # Center and bold the title
    legend.position = "none",  # Remove the legend
    panel.grid.major = element_blank(),  # Remove major grid lines
    panel.grid.minor = element_blank()   # Remove minor grid lines
  ) +
  expand_limits(y = max(gene_data_ic50$count) * 1.1)

gene_data_ic75 <- data.frame(
  regulation = c("Upregulated", "Downregulated"),
  count = c(385, 973)  # New counts for IC75
)

# Create the bar chart for IC75
ggplot(gene_data_ic75, aes(x = regulation, y = count, fill = regulation)) +
  geom_bar(stat = "identity", color = "black", size = 0.7, width = 0.6) +  # Adjust bar width for spacing
  geom_text(aes(label = count), vjust = -0.5, size = 5) +  # Add number counts on top of the bars
  labs(
    title = "Moroidin-RNH IC75 6hrs. vs. DMSO 6hrs. Treatment",  # New title for IC75
    x = "",
    y = "Gene Count"
  ) +
  scale_fill_manual(values = c("Upregulated" = "skyblue", "Downregulated" = "salmon")) +
  theme_bw() +  # Apply a clean theme
  theme(
    plot.title = element_text(hjust = 0.5, face = "bold"), # Center and bold the title
    legend.position = "none",  # Remove the legend
    panel.grid.major = element_blank(),  # Remove major grid lines
    panel.grid.minor = element_blank()   # Remove minor grid lines
  ) +
  expand_limits(y = max(gene_data_ic75$count) * 1.1)


#Set up Volcano Plot
BiocManager::install('EnhancedVolcano', force = TRUE)
install.packages("textshaping")

#All RNA seq results
RNADeseqresultsIC256hrvsDMSO <- as.data.frame(deseq2IC256hrvsDMSO.res)
RNADeseqresultsIC506hrvsDMSO <- as.data.frame(deseq2IC506hrvsDMSO.res)
RNADeseqresultsIC756hrvsDMSO <- as.data.frame(deseq2IC756hrvsDMSO.res)

#Filtered RNA-seq results
SigRNADeseqresultsIC256hrvsDMSO <- as.data.frame(sigsIC256hrvsDMSO)
SigRNADeseqresultsIC506hrvsDMSO <- as.data.frame(sigsIC506hrvsDMSO)
SigRNADeseqresultsIC756hrvsDMSO <- as.data.frame(sigsIC756hrvsDMSO)

#Library package
library(EnhancedVolcano)

#Add blank row
RNADeseqresultsIC256hrvsDMSO$blank <- ""
RNADeseqresultsIC506hrvsDMSO$blank <- ""
RNADeseqresultsIC756hrvsDMSO$blank <- ""
SigRNADeseqresultsIC256hrvsDMSO$blank <- ""
SigRNADeseqresultsIC506hrvsDMSO$blank <- ""
SigRNADeseqresultsIC756hrvsDMSO$blank <- ""

#Add numbered row
RNADeseqresultsIC256hrvsDMSO$numbered <- 1:nrow(RNADeseqresultsIC256hrvsDMSO)
RNADeseqresultsIC506hrvsDMSO$numbered <- 1:nrow(RNADeseqresultsIC506hrvsDMSO)
RNADeseqresultsIC756hrvsDMSO$numbered <- 1:nrow(RNADeseqresultsIC756hrvsDMSO)
SigRNADeseqresultsIC256hrvsDMSO$numbered <- 1:nrow(SigRNADeseqresultsIC256hrvsDMSO)
SigRNADeseqresultsIC506hrvsDMSO$numbered <- 1:nrow(SigRNADeseqresultsIC506hrvsDMSO)
SigRNADeseqresultsIC756hrvsDMSO$numbered <- 1:nrow(SigRNADeseqresultsIC756hrvsDMSO)

#make Volcano Plots The default cut-off for log2FC is >|2|; the default cut-off for P value is 10e-6.
EnhancedVolcano(RNADeseqresultsIC256hrvsDMSO,     x     =     "log2FoldChange",     y     =     "padj",     lab     =
RNADeseqresultsIC256hrvsDMSO$blank)
EnhancedVolcano(SigRNADeseqresultsIC256hrvsDMSO,     x     =     "log2FoldChange",     y     =     "padj",     lab     =
SigRNADeseqresultsIC256hrvsDMSO$blank)
EnhancedVolcano(SigRNADeseqresultsIC506hrvsDMSO,     x     =     "log2FoldChange",     y     =     "padj",     lab     =
SigRNADeseqresultsIC506hrvsDMSO$blank)
EnhancedVolcano(SigRNADeseqresultsIC756hrvsDMSO,     x     =     "log2FoldChange",     y     =     "padj",     lab     =
SigRNADeseqresultsIC756hrvsDMSO$blank)
```

```
          EnhancedVolcano(SigRNADeseqresultsIC256hrvsDMSO,     x     =     "log2FoldChange",     y     =     "padj",     lab     =
SigRNADeseqresultsIC256hrvsDMSO$numbered)
          EnhancedVolcano(SigRNADeseqresultsIC506hrvsDMSO,     x     =     "log2FoldChange",     y     =     "padj",     lab     =
SigRNADeseqresultsIC506hrvsDMSO$numbered)
          EnhancedVolcano(SigRNADeseqresultsIC756hrvsDMSO,     x     =     "log2FoldChange",     y     =     "padj",     lab     =
SigRNADeseqresultsIC756hrvsDMSO$numbered)

          #Get the significant points
          volcano_genesSigRNADeseqresultsIC256hrvsDMSO <- subset(SigRNADeseqresultsIC256hrvsDMSO, abs(log2FoldChange) > 2 & padj
< 1e-6)
          volcano_genesSigRNADeseqresultsIC506hrvsDMSO <- subset(SigRNADeseqresultsIC506hrvsDMSO, abs(log2FoldChange) > 2 & padj
< 1e-6)
          volcano_genesSigRNADeseqresultsIC756hrvsDMSO <- subset(SigRNADeseqresultsIC756hrvsDMSO, abs(log2FoldChange) > 2 & padj
< 1e-6)

          volcano_gene_namesSigRNADeseqresultsIC256hrvsDMSO <- rownames(volcano_genesSigRNADeseqresultsIC256hrvsDMSO)
          volcano_gene_namesSigRNADeseqresultsIC506hrvsDMSO <- rownames(volcano_genesSigRNADeseqresultsIC506hrvsDMSO)
          volcano_gene_namesSigRNADeseqresultsIC756hrvsDMSO <- rownames(volcano_genesSigRNADeseqresultsIC756hrvsDMSO)

          volcano_genesSigRNADeseqresultsIC256hrvsDMSO_df <- as.data.frame(volcano_genesSigRNADeseqresultsIC256hrvsDMSO)
          volcano_genesSigRNADeseqresultsIC506hrvsDMSO_df <- as.data.frame(volcano_genesSigRNADeseqresultsIC506hrvsDMSO)
          volcano_genesSigRNADeseqresultsIC756hrvsDMSO_df <- as.data.frame(volcano_genesSigRNADeseqresultsIC756hrvsDMSO)

          #Write Full file for pathway all up and downregulated genes that are filtered
          output_fileSigRNADeseqresultsIC256hrvsDMSO                          <-                          "/home/dousley/RNA-
Seq_Analysis/Volcano_Plots/SigRNADeseqresultsIC256hrvsDMSOFullforPathway.csv"
          output_fileSigRNADeseqresultsIC506hrvsDMSO                          <-                          "/home/dousley/RNA-
Seq_Analysis/Volcano_Plots/SigRNADeseqresultsIC506hrvsDMSOFullforPathway.csv"
          output_fileSigRNADeseqresultsIC756hrvsDMSO                          <-                          "/home/dousley/RNA-
Seq_Analysis/Volcano_Plots/SigRNADeseqresultsIC756hrvsDMSOFullforPathway.csv"

          write.csv(SigRNADeseqresultsIC256hrvsDMSO, file = output_fileSigRNADeseqresultsIC256hrvsDMSO, row.names = TRUE)
          write.csv(SigRNADeseqresultsIC506hrvsDMSO, file = output_fileSigRNADeseqresultsIC506hrvsDMSO, row.names = TRUE)
          write.csv(SigRNADeseqresultsIC756hrvsDMSO, file = output_fileSigRNADeseqresultsIC756hrvsDMSO, row.names = TRUE)

          #Write file for only significant genes in the volcano plot
          output_filevolcano_genesSigRNADeseqresultsIC256hrvsDMSO_df                  <-                  "/home/dousley/RNA-
Seq_Analysis/Volcano_Plots/volcano_genesSigRNADeseqresultsIC256hrvsDMSO_df.csv"
          output_filevolcano_genesSigRNADeseqresultsIC506hrvsDMSO_df                  <-                  "/home/dousley/RNA-
Seq_Analysis/Volcano_Plots/volcano_genesSigRNADeseqresultsIC506hrvsDMSO_df.csv"
          output_filevolcano_genesSigRNADeseqresultsIC756hrvsDMSO_df                  <-                  "/home/dousley/RNA-
Seq_Analysis/Volcano_Plots/volcano_genesSigRNADeseqresultsIC756hrvsDMSO_df.csv"

          write.csv(volcano_genesSigRNADeseqresultsIC256hrvsDMSO_df,                          file                          =
output_filevolcano_genesSigRNADeseqresultsIC256hrvsDMSO_df, row.names = TRUE)
          write.csv(volcano_genesSigRNADeseqresultsIC506hrvsDMSO_df,                          file                          =
output_filevolcano_genesSigRNADeseqresultsIC506hrvsDMSO_df, row.names = TRUE)
          write.csv(volcano_genesSigRNADeseqresultsIC756hrvsDMSO_df,                          file                          =
output_filevolcano_genesSigRNADeseqresultsIC756hrvsDMSO_df, row.names = TRUE)

          #Load in Heatmap Packages
          library(ComplexHeatmap)
          BiocManager::install('ComplexHeatmap', force = TRUE)

          #Set Signifigant data frames back to padjusted file w/ FDR of 0.05 (5%) and adding back in blank columns
          SigRNADeseqresultsIC256hrvsDMSO <- as.data.frame(sigsIC256hrvsDMSO)
          SigRNADeseqresultsIC506hrvsDMSO <- as.data.frame(sigsIC506hrvsDMSO)
          SigRNADeseqresultsIC756hrvsDMSO <- as.data.frame(sigsIC756hrvsDMSO)
          SigRNADeseqresultsIC256hrvsDMSO$numbered <- 1:nrow(SigRNADeseqresultsIC256hrvsDMSO)
          SigRNADeseqresultsIC506hrvsDMSO$numbered <- 1:nrow(SigRNADeseqresultsIC506hrvsDMSO)
          SigRNADeseqresultsIC756hrvsDMSO$numbered <- 1:nrow(SigRNADeseqresultsIC756hrvsDMSO)
          SigRNADeseqresultsIC256hrvsDMSO$blank <- ""
          SigRNADeseqresultsIC506hrvsDMSO$blank <- ""
          SigRNADeseqresultsIC756hrvsDMSO$blank <- ""

          #Filter the results for the heatmap by basemean and logfold2change
          HeatmapSigRNADeseqresultsIC256hrvsDMSO <- SigRNADeseqresultsIC256hrvsDMSO[(SigRNADeseqresultsIC256hrvsDMSO$baseMean >
100) & (abs(SigRNADeseqresultsIC256hrvsDMSO$log2FoldChange ) > 2),]
          HeatmapSigRNADeseqresultsIC506hrvsDMSO <- SigRNADeseqresultsIC506hrvsDMSO[(SigRNADeseqresultsIC506hrvsDMSO$baseMean >
100) & (abs(SigRNADeseqresultsIC506hrvsDMSO$log2FoldChange ) > 2),]
          HeatmapSigRNADeseqresultsIC756hrvsDMSO <- SigRNADeseqresultsIC756hrvsDMSO[(SigRNADeseqresultsIC756hrvsDMSO$baseMean >
100) & (abs(SigRNADeseqresultsIC756hrvsDMSO$log2FoldChange ) > 2),]

          #Create the Matrix for the heatmap
          MatrixIC256hrvsDMSO                 <-counts(deseq2IC256hrvsDMSO.dds,                 normalized                 =
T)[rownames(HeatmapSigRNADeseqresultsIC256hrvsDMSO),]
          MatrixIC506hrvsDMSO                 <-counts(deseq2IC506hrvsDMSO.dds,                 normalized                 =
T)[rownames(HeatmapSigRNADeseqresultsIC506hrvsDMSO),]
          MatrixIC756hrvsDMSO                 <-counts(deseq2IC756hrvsDMSO.dds,                 normalized                 =
T)[rownames(HeatmapSigRNADeseqresultsIC756hrvsDMSO),]


          #Apply scale to the matrix
          MatrixIC256hrvsDMSO.Zscore <- t(apply(MatrixIC256hrvsDMSO, 1, scale))
          MatrixIC506hrvsDMSO.Zscore <- t(apply(MatrixIC506hrvsDMSO, 1, scale))
          MatrixIC756hrvsDMSO.Zscore <- t(apply(MatrixIC756hrvsDMSO, 1, scale))

          #make sure the column names for the z score are = to the rownames in the samples
          colnames(MatrixIC256hrvsDMSO.Zscore) <- rownames(sampleTableIC256hrvsDMSO)
          colnames(MatrixIC506hrvsDMSO.Zscore) <- rownames(sampleTableIC506hrvsDMSO)
          colnames(MatrixIC756hrvsDMSO.Zscore) <- rownames(sampleTableIC756hrvsDMSO)

          #Heatmap w/ numbers
```

```
          Heatmap(MatrixIC256hrvsDMSO.Zscore,     cluster_rows     =     T,     cluster_columns     =     T,     column_labels     =
colnames(MatrixIC256hrvsDMSO.Zscore),             name             =             "Z-score",             row_labels             =
HeatmapSigRNADeseqresultsIC256hrvsDMSO[rownames(MatrixIC256hrvsDMSO.Zscore),]$numbered)
          Heatmap(MatrixIC256hrvsDMSO.Zscore,     cluster_rows     =     T,     cluster_columns     =     T,     column_labels     =
colnames(MatrixIC256hrvsDMSO.Zscore),             name             =             "Z-score",             row_labels             =
HeatmapSigRNADeseqresultsIC256hrvsDMSO[rownames(MatrixIC256hrvsDMSO.Zscore),]$blank)

          Heatmap(MatrixIC506hrvsDMSO.Zscore,     cluster_rows     =     T,     cluster_columns     =     T,     column_labels     =
colnames(MatrixIC506hrvsDMSO.Zscore),             name             =             "Z-score",             row_labels             =
HeatmapSigRNADeseqresultsIC506hrvsDMSO[rownames(MatrixIC506hrvsDMSO.Zscore),]$numbered)
          Heatmap(MatrixIC506hrvsDMSO.Zscore,     cluster_rows     =     T,     cluster_columns     =     T,     column_labels     =
colnames(MatrixIC506hrvsDMSO.Zscore),             name             =             "Z-score",             row_labels             =
HeatmapSigRNADeseqresultsIC506hrvsDMSO[rownames(MatrixIC506hrvsDMSO.Zscore),]$blank)

          Heatmap(MatrixIC756hrvsDMSO.Zscore,     cluster_rows     =     T,     cluster_columns     =     T,     column_labels     =
colnames(MatrixIC756hrvsDMSO.Zscore),             name             =             "Z-score",             row_labels             =
HeatmapSigRNADeseqresultsIC756hrvsDMSO[rownames(MatrixIC756hrvsDMSO.Zscore),]$numbered)
          Heatmap(MatrixIC756hrvsDMSO.Zscore,     cluster_rows     =     T,     cluster_columns     =     T,     column_labels     =
colnames(MatrixIC756hrvsDMSO.Zscore),             name             =             "Z-score",             row_labels             =
HeatmapSigRNADeseqresultsIC756hrvsDMSO[rownames(MatrixIC756hrvsDMSO.Zscore),]$blank)

          #Get the significant points
          heatmap_genesSigRNADeseqresultsIC256hrvsDMSO                                                                                <-
subset(SigRNADeseqresultsIC256hrvsDMSO[(SigRNADeseqresultsIC256hrvsDMSO$baseMean             >             100)             &
(abs(SigRNADeseqresultsIC256hrvsDMSO$log2FoldChange ) > 2),])
          heatmap_genesSigRNADeseqresultsIC506hrvsDMSO                                                                                <-
subset(SigRNADeseqresultsIC506hrvsDMSO[(SigRNADeseqresultsIC506hrvsDMSO$baseMean             >             100)             &
(abs(SigRNADeseqresultsIC506hrvsDMSO$log2FoldChange ) > 2),])
          heatmap_genesSigRNADeseqresultsIC756hrvsDMSO                                                                                <-
subset(SigRNADeseqresultsIC756hrvsDMSO[(SigRNADeseqresultsIC756hrvsDMSO$baseMean             >             100)             &
(abs(SigRNADeseqresultsIC756hrvsDMSO$log2FoldChange ) > 2),])

          heatmap_gene_namesSigRNADeseqresultsIC256hrvsDMSO <- rownames(heatmap_genesSigRNADeseqresultsIC256hrvsDMSO)
          heatmap_gene_namesSigRNADeseqresultsIC506hrvsDMSO <- rownames(heatmap_genesSigRNADeseqresultsIC506hrvsDMSO)
          heatmap_gene_namesSigRNADeseqresultsIC756hrvsDMSO <- rownames(heatmap_genesSigRNADeseqresultsIC756hrvsDMSO)

          heatmapSigRNADeseqresultsIC256hrvsDMSO_df <- as.data.frame(heatmap_genesSigRNADeseqresultsIC256hrvsDMSO)
          heatmapSigRNADeseqresultsIC506hrvsDMSO_df <- as.data.frame(heatmap_genesSigRNADeseqresultsIC506hrvsDMSO)
          heatmapSigRNADeseqresultsIC756hrvsDMSO_df <- as.data.frame(heatmap_genesSigRNADeseqresultsIC756hrvsDMSO)

          #Write file for only significant genes in the heatmap
          output_fileheatmapSigRNADeseqresultsIC256hrvsDMSO_df                        <-                        "/home/dousley/RNA-
Seq_Analysis/Heatmap/Heatmap_genesSigRNADeseqresultsIC256hrvsDMSO_df.csv"
          output_fileheatmapSigRNADeseqresultsIC506hrvsDMSO_df                        <-                        "/home/dousley/RNA-
Seq_Analysis/Heatmap/Heatmap_genesSigRNADeseqresultsIC506hrvsDMSO_df.csv"
          output_fileheatmapSigRNADeseqresultsIC756hrvsDMSO_df                        <-                        "/home/dousley/RNA-
Seq_Analysis/Heatmap/Heatmap_genesSigRNADeseqresultsIC756hrvsDMSO_df.csv"

          write.csv(heatmapSigRNADeseqresultsIC256hrvsDMSO_df,   file   =   output_fileheatmapSigRNADeseqresultsIC256hrvsDMSO_df,
row.names = TRUE)
          write.csv(heatmapSigRNADeseqresultsIC506hrvsDMSO_df,   file   =   output_fileheatmapSigRNADeseqresultsIC506hrvsDMSO_df,
row.names = TRUE)
          write.csv(heatmapSigRNADeseqresultsIC756hrvsDMSO_df,   file   =   output_fileheatmapSigRNADeseqresultsIC756hrvsDMSO_df,
row.names = TRUE)


          #####BAR Graphs All Conditions#######
          library(ggplot2)

          gene_data <- data.frame(
            Condition = c(
              "IC25 3hrs", "IC25 3hrs",
              "IC25 6hrs", "IC25 6hrs",
              "IC25 24hrs", "IC25 24hrs",
              "IC50 3hrs", "IC50 3hrs",
              "IC50 6hrs", "IC50 6hrs",
              "IC50 24hrs", "IC50 24hrs",
              "IC75 3hrs", "IC75 3hrs",
              "IC75 6hrs", "IC75 6hrs",
              "IC75 24hrs", "IC75 24hrs"
            ),
            Regulation = rep(c("Upregulated", "Downregulated"), 9),
            Count = c(
              211, 203,     # IC25 3hr
              391, 755,     # IC25 6hr
              474, 390,     # IC25 24hr
              176, 229,     # IC50 3hr
              308, 539,     # IC50 6hr
              436, 467,     # IC50 24hr
              252, 300,     # IC75 3hr
              385, 973,     # IC75 6hr
              270, 294      # IC75 24hr
            )
          )

          gene_data$Condition <- factor(
            gene_data$Condition,
            levels = c(
              "IC25 3hrs", "IC25 6hrs", "IC25 24hrs",
              "IC50 3hrs", "IC50 6hrs", "IC50 24hrs",
              "IC75 3hrs", "IC75 6hrs", "IC75 24hrs"
            )
          )

          # y-position for the group label (change as needed for aesthetics)
```

```
label_y <- max(gene_data$Count) * 1.18

p <- ggplot(gene_data, aes(x = Condition, y = Count, fill = Regulation)) +
  geom_bar(stat = "identity", position = position_dodge(width = 0.8),
           color = "black", width = 0.7, size = 0.7) +
  geom_text(
    aes(label = Count),
    position = position_dodge(width = 0.8),
    vjust = -0.5,
    size = 4.5
  ) +
  scale_fill_manual(values = c("Upregulated" = "skyblue", "Downregulated" = "salmon")) +
  labs(
    title = "Moroidin-RNH vs. DMSO Differentially Expressed Genes",
    x = "Treatment Comparison",
    y = "Gene Count"
  ) +
  theme_bw() +
  theme(
    plot.title = element_text(hjust = 0.5, face = "bold"),
    axis.text.x = element_text(angle = 40, hjust = 1, face = "bold"),
    legend.title = element_blank(),
    panel.grid.major = element_blank(),
    panel.grid.minor = element_blank()
  ) +
  expand_limits(y = max(gene_data$Count) * 1.23) +
  geom_vline(xintercept = 3.5, linetype = "dashed", size = 1, color = "grey40") +
  geom_vline(xintercept = 6.5, linetype = "dashed", size = 1, color = "grey40") +
  # Group labels
  annotate("text", x = 2, y = label_y, label = "IC25", size = 6, fontface = "bold") +
  annotate("text", x = 5, y = label_y, label = "IC50", size = 6, fontface = "bold") +
  annotate("text", x = 8, y = label_y, label = "IC75", size = 6, fontface = "bold")
```

4.6.16 Seqkit2 and ENTAP analysis on the GreatLakes Cluster

Top transcripts were generated in a FASTA file with seqkit2 and annotated with EnTAP (v1.0.0) using UniProt/SwissProt83 as a database[32,33].

4.6.17 Mechanism of Action (MoA) interrogation with AutoPlex and sample output

The top 100 gene names for each condition were sorted by Log2FC from highest to lowest in Excel. The top 100 gene names for each condition were copied and the AutoPlex queries were run.

4.6.18 Data Availability

Gene sequences of BURP-domain precursor peptides were deposited in GenBank (*GheBURP*-PP316100 [https://www.ncbi.nlm.nih.gov/nuccore/PP316100.1/], *SstBURP*-PP316102 [https://www.ncbi.nlm.nih.gov/nuccore/PP316102], *SaqBURP*-PQ435901 [https://www.ncbi.nlm.nih.gov/nuccore/PQ435901], *OmaBURP1*-PQ435900 [https://www.ncbi.nlm.nih.gov/nuccore/PQ435900], *OmaBURP2*-PQ463684

[https://www.ncbi.nlm.nih.gov/nuccore/PQ463684]). LC-MS/MS datasets (*Glechoma hederacea* leaf extract, *Sacciolepis striata* leaf extract, *Mercurialis annua* stem extract, *Stellaria aquatica* stem extract, *Orchidantha maxillarioides* stem and leaf extract) were deposited in GNPS-MassIVE[61] under accession MSV000096083 [https://massive.ucsd.edu/ProteoSAFe/dataset.jsp?task=20959bf587854448850f0ae82d834343]. Proteomic datasets of burpitide cyclase characterization were deposited in GNPS-MassIVE[61] under accession MSV000097169 [https://massive.ucsd.edu/ProteoSAFe/dataset.jsp?task=811455e7894e47659396e322f5a0e5f4]. NMR data was deposited in NP-MRD[62] (glechomanin-NP0350903 [https://np-mrd.org/natural_products/NP0350903], mercurialin-NP0350904 [https://np-mrd.org/natural_products/NP0350904], moroidin-QLLVWRNH-NP0350905 [https://np-mrd.org/natural_products/NP0350905]). Source data are provided with this paper.

4.7 Supplemental Information

Table 4.2 RNA-seq datasets and plant species with predicted stephanotic acid-type burpitide cores from fused burpitide cyclases in assembled plant transcriptomes (tblastn and PHI-BLAST hits).

#	SRA#	Organism	Family	Predicted core peptides (stephanotic acid in **bold,** 1x = single core precursor)	Stephanoticacid Core: QLxxW	Moroidin Core: QLxxWx(xx)H	2.ring-Trp Core: QLxxWx(xx)W	2.ring-Tyr Core: QLxxWx(xx)Y
1	SRR12990041	*Aganosma wallichii*	Apocynaceae	**QLLVW**RW (1x)	x		x	
2	SRR6456116	*Stellaria aquatica*	Caryophyl-laceae	**QLLVW**RNH, **QLFVW**RNH, **QLLVW**SKH	x	x		
3	SRR5219180	*Mercurialis huetii*	Euphorbiaceae	**QLFFW**RY, **QLFFW**KY	x			x
4	ERR2597552	*Mercurialis annua*	Euphorbiaceae	**QLFFW**KY	x			x
5	SRR8797417	*Medicago archiducis-nicolai*	Fabaceae	**QLFVW**GGH	x	x		
6	ERR6688729	*Medicago arabica*	Fabaceae	**QLLIW**GGH, **QLLVW**GGH, **QLFVW**GGH	x	x		
7	SRR3141655	*Vigna mungo*	Fabaceae	**QLLVW**RW (1x)	x		x	
8	SRR3938646	*Medicago marina*	Fabaceae	**QLLWW**GGH (1x)	x	x		
9	ERR2021956	*Bauhinia blakeana*	Fabaceae	**QLLVW**RQH, **QLLVW**IKH, **QLLIW**RQH, **QLLVW**RGH, **QLLIW**KGH, **QLLVW**RSH,	x	x		
10	DRR168799	*Bauhinia forficata*	Fabaceae	**QLRVW**TSH, **QLSVW**TSH, **QLRVW**QSH, **QLLIW**MNH, **QLLVW**RKNH	x	x		
11	ERR2021957	*Bauhinia purpurea*	Fabaceae	**QLSFW**, **QLLVW**RNH, **QLLVW**RGH, **QLIVW**RRNH, **QLFVW**RGH, **QLRVW**TKH, **QLLVW**REH, **QLIVW**RNH, **QLLVW**RRGH, **QLLVW**RRNH,	x	x		
12	SRR16068779	*Medicago sativa*	Fabaceae	**QLFVW**SDH, **QLFVW**DDGH, **QLLVW**SSDH, **QLFVW**GGVH, **QLFVW**GGH, **QLLVW**VGH, **QLLVW**GGGH, **QLFVW**GGGH, **QLLVW**GGDH,	x	x		
14	SRR18083643	*Glechoma longituba*	Lamiaceae	**QLLVW**RW, **QLLVW**KW	x		x	
15	SRR5150732	*Glechoma hederacea*	Lamiaceae	**QLFVW**GW, **QLLVW**KW, **QLLVW**RW	x		x	
16	ERR3487362	*Salvia* sp. 1KP	Lamiaceae	**QLFVW**GW	x			
17	ERR2040758	*Orchidantha maxillari- oides*	Lowiaceae	**QLLVW**KW, **QLLVW**RSH, **QLFNW**YTH, **QLLVW**PKH, **QLLVW**TDH	x	x	x	
18	SRR16483051	*Cinna latifolia*	Poaceae	**QLLVW**, **QLLVW**RTY, **QLLVW**RQY **QLLVW**RSY, **QLFIW**RSY, **QLFIW**RSY, **QLLLW**RSY, **QLFIW**RSY, **QLLVW**RSY, **QLLIW**RTY, **QLLVW**RTY, **QLLVW**RQY, **QLLLW**RRSY, **QLLVW**RSY, **QLFIW**RSY, **QLLIW**RTY, **QLLVW**RTY, **QLFIW**RRSY, **QLLIW**RRTY,	x			x

19	SRR5406165	*Digitaria cuyabensis*	Poaceae	**QLLVW**RNH	x	x		
20	SRR16234948	*Stipa sibirica*	Poaceae	**QLLVW**RKSW, **QLLVW**RESW	x		x	
21	SRR13316951	*Cinnagrostis crispa*	Poaceae	**QLLVW**RDW	x		x	
22	SRR16483054	*Holcus mollis*	Poaceae	**QLLVW**RSY, **QLLIW**RTY, **QLLVW**RTY, **QLLVW**RQY	x			x
23	SRR16243467	*Panicum brevifolium*	Poaceae	**QLLVW**RNH	x	x		
24	SRR13316952	*Cinnagrostis cabrerae*	Poaceae	**QLLVW**RDW	x		x	
25	SRR23873622	*Aegilops uniaristata*	Poaceae	**QLLIW**	x			
26	SRR16299781	*Leptothrium senegalense*	Poaceae	**QLLVW**RSH, **QLLVW**RNH	x	x		
27	SRR16483114	*Neomolinia japonica*	Poaceae	**QLLVW**RSW	x		x	
28	ERR3634426	*Brachypodium boissieri*	Poaceae	**QLLVW**RSW	x		x	
29	SRR12955278	*Bromus catharticus*	Poaceae	**QLLVW**RSW	x		x	
30	SRR6221583, SRR6221613	*Sacciolepis striata*	Poaceae	**QLLVW**RNH	x	x		
31	SRR5406195	*Zuloagaea bulbosa*	Poaceae	**QLLVW**RNH	x	x		
32	ERR9784826	*Dichrosta-chys cinerea*	Fabaceae	**QLLVW**GY	X			X
33	SRR7817914	*Hordeum brevisubula-tum*	Poaceae	**QLRVW**RKY (1x)	x			x
34	SRR23631370	*Amomum smithiae*	Zingiberaceae	**QLLVW** (1x)	x			
		Known producers						
35	ERR2040423	*Kerria japonica*	Rosaceae	**QLLVW**RAH, **QLLVW**RGH	x	x		
36	ERR706821	*Bauhinia tomentosa*	Fabaceae	**QLLVW**RSH, **QLLVW**RNH, **QLLFW**FGRH, **QLLVW**SRH, **QLLVW**TRH, **QLLFW**TRH,	x	x		
37	ERR706845	*Cercis canadensis*	Fabaceae	**QLLVW**, **QLKVW**, **QLLVW**QKH, **QLLVW**RH	x	x		

Table 4.3 ^{1}H-NMR (800 MHz) and ^{13}C-NMR (200 MHz) data for glechomanin in DMSO-d$_6$

Residue	C	$\delta(^{13}C)$ [ppm]	H	$\delta(^{1}H)$ (int, m, *J*) [ppm]	COSY	TOCSY	HMBC	ROESY
PyroGlu^{1}	*α*	55.2	*α*	3.98 (^{1}H, dd, 7.9, 4.4)	PyroGlu^{1}: H*β1,* H*β2*, NH	PyroGlu^{1}: H*β1,* H*β2*, H*γ*, NH	PyroGlu^{1}: C*β*	PyroGlu^{1}: H*β1*, H*β2,* NH, Leu^{2}: NH
	β	24.8	*β1*	2.10 [a] (^{1}H)	PyroGlu^{1}: H*α,* H*β2*, H*γ*	PyroGlu^{1}: H*α*, H*β2*, H*γ*, NH	PyroGlu^{1}: C*α*, C*γ*, C*δ*, C=O	PyroGlu^{1}: H*α*, H*β2*
	γ	29.3	*β2*	1.61 (^{1}H, br s)	PyroGlu^{1}: H*α,* H*β1*, H*γ*	PyroGlu^{1}: H*α*, H*β1*, H*γ*, NH,	n/a	PyroGlu^{1}: H*α*, H*β1*, H*γ*
	δ	177.2	*γ*	2.10 [a] (2H)	PyroGlu^{1}: H*α,* H*β2*	PyroGlu^{1}: H*α*, H*β2*, NH	PyroGlu^{1}: C*α*, C*β*, C*δ*	PyroGlu^{1}: H*β2*
	C=O	171.2	NH	7.74 (^{1}H, s)	PyroGlu^{1}: H*α*	PyroGlu^{1}: H*α*, H*β1,* H*β2*, H*γ*	PyroGlu^{1}: C*α*, C*β*, C*γ*, C*δ*	PyroGlu^{1}: H*α*, Leu^{2}: NH
Leu^{2}	*α*	55.1	*α*	4.72 (^{1}H, br t, 10.3)	Leu^{2}: NH, H*β*	Leu^{2}: H*β*, H*γ*, H*δ1*, H*δ2*, NH	n/a	Leu^{2}: H*β*, H*γ*, H*δ2*, NH, Phe^{3}: NH, Trp^{5}: H5
	β	52.2	*β*	2.87 (^{1}H, dd, 11.6, 3.1)	Leu^{2}: H*α,* H*γ*	Leu^{2}: H*α*, H*γ*, H*δ1*, H*δ2*, NH	Leu^{2}: C*α*, C*γ*, C*δ1*, C*δ2*, C=O, Trp^{5}: C5, C6, C7	Leu^{2}: H*α*, H*γ*, H*δ1*, H*δ2*, NH, Phe^{3}: NH, Trp^{5}: H7
	γ	26.6	*γ*	2.05[a] (^{1}H)	Leu^{2}: H*β,* H*δ1,* H*δ2*	Leu^{2}: H*α*, H*β*, H*δ1*, H*δ2*, NH	n/a	Leu^{2}: H*α*, H*β*, H*δ1*, H*δ2*, NH, Trp^{5}: H7
	δ1	21.8	*δ1*	0.66 (3H, d, 6.7)	Leu^{2}: H*γ*	Leu^{2}: H*α*, H*β*, H*γ*, H*δ2*, NH	Leu^{2}: C*β*, C*γ*, C*δ2*	Leu^{2}: H*β*, H*γ*, H*δ2*, Trp^{5}: H7
	δ2	17.3	*δ2*	0.82 (3H, br s)	Leu^{2}: H*γ*	Leu^{2}: H*α*, H*β*, H*γ*, H*δ1*, NH	n/a	Leu^{2}: H*α*, H*β*, H*γ*, H*δ1*, Trp^{5}: H5
	C=O	169.1	NH	8.09 (IH, d, 9.4)	Leu^{2}: H*α*	Leu^{2}: H*α*, H*β*, H*γ*, H*δ1*, H*δ2*	PyroGlu^{1}: C=O	PyroGlu^{1}: H*α*, H*β2*, Leu^{2}: H*α*, H*β*, H*γ*
Phe^{3}	*α*	55.6	*α*	4.26 (^{1}H, br t, 10.5)	Phe^{3}: H*β1,* H*β2*, NH	Phe^{3}: H*β1,* H*β2*, NH	Phe^{3}: C*β*, C=O	Phe^{3}: H*β1*, H*β2*, H2, H6, NH, Val^{4}: NH
	β	38.9	*β1*	2.78 (^{1}H, br d, 11.6)	Phe^{3}: H*α*	Phe^{3}: H*α*, NH	n/a	Phe^{3}: H*α*, H2, H6, NH, Val^{4}: H*γ1*, NH
			β2	2.85[a] (^{1}H)	Phe^{3}: H*α*	Phe^{3}: H*α*, NH	n/a	Phe^{3}: H*α*, H2, H6, NH, Val^{4}: H*γ1*, NH
	C1	137.5	H2	7.11 (2H, d, 7.2)	Phe^{3}: H3	Phe^{3}: H3, H4, H5, H6	Phe^{3}: C*α*, C4, C6	Phe^{3}: H*α*, H*β1*, H*β2*, H3
	C2	128.9	H3	7.16 (2H, t, 7.2)	Phe^{3}: H2, H4	Phe^{3}: H2, H4, H5, H6	Phe^{3}: C1, C5	Phe^{3}: H2, H4
	C3	127.9	H4	7.13 (^{1}H, t, 7.2)	Phe^{3}: H3, H5	Phe^{3}: H2, H3, H5, H6	Phe^{3}: C2, C6	Phe^{3}: H3, H5
	C4	126.0	H5	7.16 (2H, t, 7.2)	Phe^{3}: H4, H6	Phe^{3}: H2, H3, H4, H6	Phe^{3}: C1, C3	Phe^{3}: H4, H6
	C5	127.9	H6	7.11 (2H, d, 7.2)	Phe^{3}: H5	Phe^{3}: H2, H3, H4, H5	Phe^{3}: C*α*, C2, C4	Phe^{3}: H*α*, H*β1*, H*β2*, H5
	C6	128.9	NH	8.35 (^{1}H, s)	Phe^{3}: H*α*	Phe^{3}: H*α,* H*β1,* H*β2*	n/a	Phe^{3}: H*α*, H*β2*, Leu^{2}: H*α*, Val^{4}: NH, Trp^{5}: H5
	C=O	169.8						
Val^{4}	*α*	59.7	*α*	4.41 (^{1}H, s)	Val^{4}: NH	Val^{4}: NH	n/a	Val^{4}: H*γ1*, Trp^{5}: NH, Trp^{7}: H5
	β	43.2	*γ1*	1.05 (3H, s)	n/a	Val^{4}: H*γ2*	Val^{4}: C*α*, C*β*, C*γ2*, Trp^{7}: C6	Val^{4}: H*α*, H*γ2*, Trp^{7}: H7
	γ1	29.6	*γ2*	1.45 (3H, s)	n/a	Val^{4}: H*γ1*	Val^{4}: C*α*	Val^{4}: H*γ1*, NH, Trp^{7}: H5, H7
	γ2	21.5			Val^{4}: H*α*	Val^{4}: H*α*		Val^{4}: H*α*, H*γ2*, Phe^{3}: H*β2*, NH

	C=O	170.3[b]	NH	7.56 (^{1}H, s)			n/a	
Trp[5]	*α*	51.2	*α*	5.02 (^{1}H, br s)	Trp[5]: H*β1,* NH	Trp[5]: H*β1,* H*β2* NH	n/a	Trp[5]: H*β1*, H4
	β	27.9	*β1*	2.71 (^{1}H, dd, 14.0, 11.4)	Trp[5]: H*α*, H*β2*	Trp[5]: H*α,* H*β2*,NH	n/a	Trp[5]: H*α*, H*β2*, H2, NH
	C2	123.8	*β2*	3.29[a] (^{1}H)	Trp[5]: H*α*	Trp[5]: NH, H*α*	n/a	Trp[5]: H*β1*
	C3	108.8	(NH)1	10.52 (^{1}H, s)	Trp[5]: H2	Trp[5]: H2	Trp[5]: C3	Trp[5]: H2, H7
	C4	118.3	H2	6.89 (^{1}H, s)	Trp[5]: (NH)1	Trp[5]: (NH)1	n/a	Trp[5]: H*β1*, (NH)1
	C5	118.2	H4	7.37 (^{1}H, d, 8.4)	Trp[5]: H5	Trp[5]: H5	Trp[5]: C3, C6, C8	Trp[5]: H*α*, H*β2*, H5
	C6	129.3	H5	6.86 (^{1}H, br d, 8.4)	Trp[5]: H4	Trp[5]: H4	Trp[5]: C7, C9	Leu[2]: H*α,* H*δ2*, Trp[5]: H4
	C7	113.8	H7	6.77 (^{1}H, s)	n/a	Trp[5]: H4	Leu[2]: C*β*, Trp[5]: C5, C9	Leu[2]: H*β*, H*γ*, H*δ1*, Trp[5]: (NH)1
	C8	136.2	NH	7.64 (^{1}H, s)	Trp[5]: H*α*	Trp[5]: H*α*, H*β1,* H*β2*	n/a	Val[4]: H*α*, Trp[5]: H*β1*
	C9	124.9						
	C=O	170.9[b]						
Gly[6]	*α*	41.7	*α1*	3.96[a] (^{1}H)	Gly[6]: H*α2*	Gly[6]: H*α2*, NH	n/a	Gly[6]: H*α2*
	C=O	167.3[b]	*α2*	3.30[a] (^{1}H)	Gly[6]: H*α1*	Gly[6]: NH	n/a	Gly[6]: H*α1*
			NH	7.86[a] (^{1}H)	n/a	Gly[6]: H*α1,* H*α2*	n/a	Gly[6]: H*α2*
Trp[7]	*α*	54.8	*α*	4.52 (^{1}H, br s)	Trp[7]: H*β1*, H*β2*, NH	Trp[7]: H*β1*, H*β2*, NH	n/a	Trp[7]: H*β1*, H*β2*, NH
	β	29.2	*β1*	3.22 (^{1}H, br d, 13.2)	Trp[7]: H*α*, H*β2*	Trp[7]: H*α*, H*β2*, NH	Trp[7]: C2, C3, C9	Trp[7]: H*α*, H*β2*, H2, H4
	C2	124.1	*β2*	2.84[a] (^{1}H)	Trp[7]: H*α*, H*β1*	Trp[7]: H*α*, H*β1*, NH	n/a	Trp[7]: H*α*, H*β1*, H4
	C3	110.8	(NH)1	10.69 (^{1}H, s)	Trp[7]: H2	Trp[7]: H2	n/a	Trp[7]: H2, H7
	C4	117.8	H2	7.19 (^{1}H, s)	Trp[7]: (NH)1	Trp[7]: (NH)1	n/a	Trp[7]: H*β1*, (NH)1
	C5	115.6	H4	7.72[a] (^{1}H)	Trp[7]: H5	Trp[7]: H5, H7	n/a	Trp[7]: H*β1*, H*β2*, H5, NH
	C6	137.8	H5	6.79[a] (^{1}H)	Trp[7]: H4	Trp[7]: H4, , H7	n/a	Val[4]: H*α*, H*γ1*, H*γ2*, Trp[7]: H4,
	C7	110.6	H7	7.42 (^{1}H, s)	n/a	Trp[7]: H4, H5	n/a	Val[4]: H*γ1*, H*γ2*, Trp[7]: (NH)1
	C8	136.5	NH	8.49 (^{1}H, s)	Trp[7]: H*α*	Trp[7]: H*α*, H*β1*, H*β2*	n/a	Gly[6]: H*α2*, Trp[7]: H*α*, H4
	C9	126.7						

a – overlapping signals, b – signals cannot be assigned by 2D NMR, br s – broad singlet, d – doublet, dd – double doublet, t – triplet, m – multiplet.

Table 4.4 ^{1}H-NMR (800 MHz) and ^{13}C-NMR (200 MHz) data for glechomanin in methanol-d$_4$.

Residue	C	*δ*(^{13}C) [ppm]	H	*δ*(^{1}H) (int, m, *J*) [ppm]	COSY	TOCSY	HMBC	ROESY
PyroGlu[1]	*α*	56.4	*α*	4.16 (^{1}H, m)	PyroGlu[1]: H*β1,* H*β2*	PyroGlu[1]: H*β1,* H*β2*, H*γ*	n/a	PyroGlu[1]: H*β1*, H*β2,* Leu[2]: NH
	β	25.3	*β1*	1.75 (^{1}H, m)	PyroGlu[1]: H*α,* H*β2*, H*γ*	PyroGlu[1]: H*α*, H*β2*, H*γ*	n/a	PyroGlu[1]: H*α*, H*β2*
	γ	29.2	*β2*	2.34 [a] (^{1}H)	PyroGlu[1]: H*α,* H*β1*, H*γ*	PyroGlu[1]: H*α*, H*β1*, H*γ*	PyroGlu[1]: C*γ*	PyroGlu[1]: H*α*, H*β1*
	δ	n/a	*γ*	2.35 [a] (2H)	PyroGlu[1]: H*α,* H*β1*	PyroGlu[1]: H*α*, H*β1*	n/a	PyroGlu[1]: H*β1*
	C=O	n/a	NH	n/a	n/a	n/a	n/a	n/a
Leu[2]	*α*	56.6	*α*	4.78 (^{1}H, t, 10.5)	Leu[2]: NH, H*β*	Leu[2]: H*β*, H*γ*, H*δ1*, H*δ2*, NH	Leu[2]: C*β*	Leu[2]: H*β*, H*δ2*, NH, Phe[3]: NH, Trp[5]: H5
	β	52.7	*β*	3.00 (^{1}H, br d, 11.7)	Leu[2]: H*α,* H*γ*	Leu[2]: H*α*, H*γ*, H*δ1*, H*δ2*, NH	n/a	Leu[2]: H*α*, H*γ*, H*δ1*, H*δ2*, NH, Phe[3]: NH, Trp[5]: H7
	γ	27.6	*γ*	2.18 (^{1}H, m)	Leu[2]: H*β,* H*δ1,* H*δ2*	Leu[2]: H*α*, H*β*, H*δ1*, H*δ2*, NH	Leu[2]: C*δ2*	Leu[2]: H*α*, H*β*, H*δ1*, H*δ2*, NH, Trp[5]: H7
	δ1	20.7	*δ1*	0.76 (3H, d, 6.4)	Leu[2]: H*γ*	Leu[2]: H*α*, H*β*, H*γ*, H*δ2*, NH	n/a	Leu[2]: H*β*, H*γ*, Trp[5]: H7
	δ2	16.9	*δ2*	0.98 (3H, d, 6.8)	Leu[2]: H*γ*	Leu[2]: H*α*, H*β*, H*γ*, H*δ1*, NH	Leu[2]: C*δ1*	Leu[2]: H*α*, H*β*, H*γ*, Trp[5]: H5
	C=O	n/a	NH	8.06 [a]	Leu[2]: H*α*	Leu[2]: H*α*, H*β*, H*γ*, H*δ1*, H*δ2*	n/a	PyroGlu[1]: H*α*, H*β2*, Leu[2]: H*α*, H*β*, H*γ*
Phe[3]	*α*	55.9	*α*	4.64 (^{1}H, m)	Phe[3]: H*β1,* H*β2*, NH	Phe[3]: H*β1,* H*β2*, NH	n/a	Phe[3]: H*β1*, H*β2*, H2, H6, NH,
	β	39.7	*β1*	3.00 (^{1}H, br d, 11.4)	Phe[3]: H*α*, H*β2*	Phe[3]: H*α*, H*β2*, NH	n/a	Phe[3]: H*α*, H*β2*, H2, H6, NH, Val[4]: H*γ1*
			β2	2.76 (^{1}H, t, 12.4)	Phe[3]: H*α*, H*β1*	Phe[3]: H*α*, H*β1*, NH	n/a	Phe[3]: H*α*, H*β1*, H2, H6, NH, Val[4]: H*γ1*, H*γ2*
	C1	n/a	H2	7.24 [a] (^{1}H)	Phe[3]: H3	Phe[3]: H3, H4, H5, H6	Phe[3]: C4	Phe[3]: H*α*, H*β1*, H*β2*, H3
	C2	128.9	H3	7.23 [a] (^{1}H)	Phe[3]: H2, H4	Phe[3]: H2, H4, H5, H6	Phe[3]: C5	Phe[3]: H2, H4
	C3	128.0	H4	7.18 [a] (^{1}H)	Phe[3]: H3, H5	Phe[3]: H2, H3, H5, H6	Phe[3]: C2, C6	Phe[3]: H3, H5
	C4	126.3	H5	7.23 [a] (^{1}H)	Phe[3]: H4, H6	Phe[3]: H2, H3, H4, H6	Phe[3]: C3	Phe[3]: H4, H6
	C5	128.0	H6	7.24 [a] (^{1}H)	Phe[3]: H5	Phe[3]: H2, H3, H4, H5	Phe[3]: C4	Phe[3]: H*α*, H*β1*, H*β2*, H5
	C6	128.9	NH	8.08 [a] (^{1}H)	Phe[3]: H*α*	Phe[3]: H*α,* H*β1,* H*β2*	n/a	Leu[2]: H*α,* Phe[3]: H*α*, H*β1*, H*β2*
	C=O	n/a						
Val[4]	*α*	60.7	*α*	4.45 (^{1}H, d, 8.5)	Val[4]: NH	Val[4]: NH	n/a	Val[4]: H*γ1*, Trp[5]: NH, Trp[7]: H5
	β	42.7	*γ1*	1.10 (3H, s)	n/a	Val[4]: H*γ2*	Val[4]: C*β,* C*γ2*, Trp[7]: C6	Val[4]: H*α*, Trp[7]: H7
	γ	28.9	*γ2*	1.47 (3H, s)	n/a	Val[4]: H*γ1*	Val[4]: C*γ2*	Val[4]: H*γ1*, NH, Trp[7]: H5, H7
	γ	20.1		7.25 [a] (^{1}H)	Val[4]: H*α*	Val[4]: H*α*		Val[4]: H*α*, Phe[3]: H*β1*, H*β2*, NH

	C=O	n/a	NH				n/a	
Trp[5]	α	51.2	α	5.23 (^{1}H, br s)	Trp[5]: Hβ*1,* NH	Trp[5]: Hβ*1,* Hβ*2* NH	n/a	Trp[5]: Hβ*1*, H4
	β	27.5	β_1	2.86 (^{1}H, t, 13.5)	Trp[5]: Hα, Hβ*2*	Trp[5]: Hα, Hβ*2*,NH	n/a	Trp[5]: Hα, Hβ*2*, H2, NH
	C2	124.2	β_2	3.48 (^{1}H, dd, 14.3, 4.8)	Trp[5]: Hα, Hβ*1*	Trp[5]: Hα, Hβ*1*,NH	Trp[5]: Cα	Trp[5]: Hα, Hβ*1*, H4
	C3	107.6	(NH)1	n/a	n/a	n/a	n/a	n/a
	C4	118.8	H2	6.86 (^{1}H, s)	n/a	Trp[5]: Hβ*1,* Hβ*2*	Trp[5]: C3	Trp[5]: Hβ*1*
	C5	117.8	H4	7.46 (^{1}H, d, 8.3)	Trp[5]: H5	Trp[5]: H5	Trp[5]: C3, C6, C8	Trp[5]: Hα, Hβ*2*
	C6	127.2	H5	6.97 (^{1}H, d, 8.3)	Trp[5]: H4	Trp[5]: H4	Trp[5]: C7, C9	Leu[2]: Hα, Hβ, Hδ*2*, Phe[3]: NH, Trp[5]: H4
	C7	114.0	H7	6.92 (^{1}H, s)	n/a	Trp[5]: H4	Trp[5]: C5	Leu[2]: Hβ, Hδ*1*
	C8	145.8	NH	7.19 [a] (^{1}H)	Trp[5]: Hα	Trp[5]: Hα, Hβ*1,* Hβ*2*	n/a	Val[4]: Hα, Trp[5]: Hβ*1*
	C9	125.3						
	C=O	n/a						
Gly[6]	α	41.2	α*1*	3.35	Gly[6]: Hα*2*	Gly[6]: Hα*2*	n/a	Gly[6]: Hα*2*, Trp[7]: Hα,
	C=O	n/a	α*2*	4.31	Gly[6]: Hα*1*	Gly[6]: Hα*1*	n/a	Gly[6]: Hα*1*
			NH	n/a	n/a	n/a	n/a	n/a
Trp[7]	α	55.5	α	4.73 (^{1}H, br d, 11.2)	Trp[7]: Hβ*1*, Hβ*2*	Trp[7]: Hβ*1*, Hβ*2*,	n/a	Gly[6]: Hα*1*, Trp[7]: Hβ*2*
	β	29.3	β_1	3.00 (^{1}H, m)	Trp[7]: Hα, Hβ*2*	Trp[7]: Hα, Hβ*2*	n/a	Trp[7]: Hα, Hβ*2*, H2, H4
	C2	124.3	β_2	3.41[a] (^{1}H)	Trp[7]: Hα, Hβ*1*	Trp[7]: Hα, Hβ*1*,	Trp[7]: Cα	Trp[7]: Hβ*1*, H2
	C3	n/a	(NH)1	n/a	n/a	n/a	n/a	n/a
	C4	117.6	H2	7.21[a]	n/a	n/a	n/a	Trp[7]: Hβ*1*, Hβ*2*
	C5	115.5	H4	7.64 (^{1}H, d, 8.0)	Trp[7]: H5	Trp[7]: H5, H7	Trp[7]: C7	Trp[7]: Hβ*1*, Hβ*2*, H5
	C6	138.1	H5	6.97 (^{1}H, d, 8.0)	Trp[7]: H4	Trp[7]: H4, H7	n/a	Val[4]: Hγ*1*, Hγ*2*, Trp[7]: Hα, H4
	C7	110.6	H7	7.55 (^{1}H, s)	n/a	Trp[7]: H4, H5	Trp[7]: C5	Val[4]: Hα, Hγ*1*, Hγ*2*
	C8	n/a	NH	n/a	n/a	n/a	n/a	n/a
	C9	n/a						
	C=O	n/a						

a – overlapping sig nals, b – signals cannot be assigned by 2D NMR, br s – broad singlet, d – doublet, dd – double doublet, t – triplet, m – multiplet.

Table 4.5 Tandem MS fragmentation of linear tryptic core peptide of GheBURP-1xQLFVWGW. Proteomic data MassIVE ID: MSV000097169

Fragment (z)	m(obs)	m(calc)	Δm [ppm]
b2 (1+)	228.1342	228.1343	0.4
b3 (1+)	329.1817	329.1820	0.9
y2 (1+)	333.1919	333.1922	0.9
y3 (1+)	390.2133	390.2136	0.8
b4 (1+)	430.2295	430.2297	0.5
y4 (1+)	576.2935	576.2929	1
b5 (1+)	593.2931	593.2930	0.2
y5 (1+)	675.3613	675.3614	0.1
b6 (1+)	707.3354	707.3359	0.7
b7 (1+)	778.3722	778.3730	1
y6 (1+)	822.4287	822.4298	1.3
b8 (1+)	849.4096	849.4102	0.7
y7 (1+)	935.5138	935.5138	0
b9 (1+)	936.4405	936.4422	1.8
b10 (1+)	1051.4697	1051.4691	0.6
y8 (1+)	1063.5707	1063.5724	1.6
b11 (1+)	1165.5113	1165.5120	0.7
y9 (1+)	1177.6152	1177.6153	0.1
y10 (1+)	1234.6355	1234.6368	1.1
b13 (1+)	1336.5736	1336.5764	2.1

y11 (1+)	1348.6780	1348.6797	1.3
y12 (1+)	1463.7054	1463.7067	0.9
b14 (1+)	1464.6520	1464.6350	11.6
y13 (1+)	1550.7360	1550.7387	1.7
b15 (1+)	1577.7193	1577.7191	0.1
y14 (1+)	1621.7744	1621.7758	0.9
y15 (1+)	1692.8098	1692.8129	1.8
b16 (1+)	1724.7881	1724.7875	0.3
y16 (1+)	1806.8535	1806.8559	1.3
b17 (1+)	1823.8510	1823.8559	2.7
y17 (1+)	1969.9164	1969.9192	1.4
b18 (1+)	2009.9360	2009.9352	0.4
y18 (1+)	2070.9632	2070.9669	1.8
y19 (1+)	2172.0123	2172.0145	1
b21 (2+)	2381.1300	2381.1310	0.4
y21 (2+)	2399.1370	2399.1415	1.9

Table 4.6 Tandem MS fragmentation of putative bicyclic tryptic core peptide of GheBURP-1xQLFVWGW. Proteomic data MassIVE ID: MSV000097169

Fragment (z)	m(obs)	m(calc)	Δm [ppm]
b2 (1+)	228.1341	228.1343	0.9
b3 (1+)	329.1816	329.1820	1.2
b5 (1+)	593.2915	593.293	2.5
b6 (1+)	707.3357	707.3359	0.3
y7* (1+)	931.4817	931.4826	1.0
b7 (1+)	778.3717	778.3730	1.7
y8* (1+)	1059.5388	1059.5411	2.2
b8 (1+)	849.4096	849.4102	0.7
y9* (1+)	1173.5830	1173.5841	0.9
b9 (1+)	936.4404	936.4422	1.9
y10* (1+)	1230.6054	1230.6055	0.1
b10 (1+)	1051.4704	1051.4691	1.2
y11* (1+)	1344.6453	1344.6484	2.3
b11 (1+)	1165.5071	1165.512	4.2
y12* (1+)	1459.6713	1459.6754	2.8
y13* (1+)	1546.7044	1546.7074	1.9
b13 (1+)	1336.5758	1336.5764	0.4
y14* (1+)	1617.7390	1617.7445	3.4
b14 (1+)	1464.6286	1464.6350	4.4
y15* (1+)	1688.7785	1688.7816	1.8
y16* (1+)	1802.8227	1802.8246	1.1
y17* (1+)	1965.8819	1965.8879	3.1
y18* (1+)	2066.9320	2066.9356	1.7

y19* (1+)	2167.9759	2167.9833	3.4
b21* (2+)	2377.0933	2377.0997	2.7

Table 4.7 ^{1}H-NMR (800 MHz) and ^{13}C-NMR (200 MHz) data for mercurialin in DMSO-d_6.

Residue	C	δ(^{13}C) [ppm]	H	δ(^{1}H) (int, m, *J*) [ppm]	COSY	HMBC	NOESY
PyroGlu^{1}	*α*	55.1	*α*	3.92 (^{1}H, br d, 7.5)	PyroGlu^{1}: NH, H*β1*, H*β2*	PyroGlu^{1}: C*β*, C*δ*	PyroGlu^{1}: NH, H*β1*, H*β2*, H*γ*, Leu^{2}: NH
	β	24.9	*β1*	1.49 (^{1}H, m)	PyroGlu^{1}: H*α*, H*β2*, H*γ*	PyroGlu^{1}: C*α*, C*γ*, C*δ*, C=O	PyroGlu^{1}: H*α*, H*β2*, H*γ*, NH
	γ	29.2	*β2*	2.04 (^{1}H, m)	PyroGlu^{1}: H*α*, H*β1*, H*γ*	PyroGlu^{1}: C*α*, C*γ*, C*δ*	PyroGlu^{1}: H*α*, H*β1*, H*γ*, NH
	δ	177.2	*γ*	2.04 (2H, m)	PyroGlu^{1}: H*β1*, H*β2*	PyroGlu^{1}: C*β*, C*δ*	PyroGlu^{1}: H*β1*, H*β2*
	C=O	171.5	NH	7.70 (^{1}H, s)	PyroGlu^{1}: H*α*	PyroGlu^{1}: C*α*, C*β*, C*γ*, C*δ*	PyroGlu^{1}: H*α*, Leu^{2}: NH, H*β2*
Leu^{2}	*α*	55.4	*α*	4.57 (^{1}H, t, 10.5)	Leu^{2}: NH, H*β*	PyroGlu^{1}: C=O, Leu^{2}: C*β*, C=O	Leu^{2}: NH, H*β*, H*γ*, H*δ2*, Phe^{3}: NH, Trp^{5}: H5
	β	52.4	*β*	2.85 (^{1}H, dd, 2.9, 11.6)	Leu^{2}: H*α*, H*γ*	Leu^{2}: C*α*, C*γ*, C*δ1*, C*δ2*, C=O, Trp^{5}: C5, C6, C7	Leu^{5}: NH, H*α*, H*γ*, H*δ1*, H*δ2*, Trp^{5}: H7
	γ	26.7	*γ*	2.04 (^{1}H, m)	Leu^{2}: H*β*, H*δ1*, H*δ2*	n/a	Leu^{2}: NH, H*β*, H*δ1*, H*δ2*
	δ1	21.8	*δ1*	0.66 (3H, d, 6.7)	Leu^{2}: H*γ*	Leu^{2}: C*β*, C*γ*, C*δ2*	Leu^{2}: NH, H*β*, H*γ*, H*δ2*, Trp^{5}: H7
	δ2	17.6	*δ2*	0.85 (3H, d, 6.7)	Leu^{2}: H*γ*	Leu^{2}: C*β*, C*γ*, C*δ1*	Leu^{2}: H*α*, H*γ*, H*δ1*, Trp^{5}: H4, H5
	C=O	169	NH	8.11 (^{1}H, d, 9.1)	Leu^{2}: H*α*	PyroGlu^{1}: C=O, Leu^{2}: C*α*	PyroGlu^{1}: H*α*, Leu^{2}: H*α*, H*β*, H*γ*
Phe^{3}	*α*	54.3	*α*	3.86 (^{1}H, t, 9.5)	Phe^{3}: NH, H*β1*, H*β2*	Leu^{2}: C=O, Phe^{3}: C*β*, C=O	Phe^{3}: NH, H*β1*, H*β2*, Trp^{5}: H2, Arg^{6}: NH
	β	37.2	*β1*	1.66 (^{1}H, t, 13.7)	Phe^{3}: H*α*, H*β2*	Phe^{3}: C*α*, C1, C2, C6	Phe^{3}: NH, H*α*, H*β2*, H2/6, Phe^{4}: H2/6, Trp^{5}: H4
	C1	137.7	*β2*	1.18 (^{1}H, d, 13.7)	Phe^{3}: H*β1*	Phe^{3}: C2, C6	Phe^{3}: H*α*, H*β1*, H2/6
	C2	128.3	H2	6.74 (^{1}H, m)	Phe^{3}: H3	Phe^{3}: C1, C3	Phe^{3}: H*α*, H*β1/2*, H3
	C3	127.8	H3	7.04 (^{1}H, br s)	Phe^{3}: H2, H4	Phe^{3}: C2, C4	PyroGlu^{1}: H*β1/2*, Phe^{3}: H2
	C4	125.8	H4	7.04 (^{1}H, br s)	Phe^{3}: H3, H5	Phe^{3}: C3, C5	Phe^{3}: H3/5
	C5	127.8	H5	7.04 (^{1}H, br s)	Phe^{3}: H4, H6	Phe^{3}: C4, C6	PyroGlu^{1}: H*β1/2*, Phe^{3}: H2
	C6	128.3	H6	6.74 (^{1}H, m)	Phe^{3}: H5	Phe^{3}: C1, C5	Phe^{3}: H*α*, H*β1/2*, H5
	C=O	169.2	NH	7.95 (^{1}H, br d, 8.11)	Phe^{3}: H*α*	Leu^{2}: C=O	Leu^{2}: H*α*, Phe^{3}: H*α*, H*β1*, Trp^{5}: H5
Phe^{4}	*α*	56.4	*α*	4.44 (^{1}H, dd, 8.8,7.0)	Phe^{4}: NH, H*β*	Phe^{3}: C=O, Phe^{4}: C*β*, C=O	Phe^{4}: NH, H*β*, H2/6, Trp^{5}: NH, H4
	β	83.2	*β*	5.35 (^{1}H, d, 7.0)	Phe^{4}: H*α*	Phe^{4}: C=O, C1, C2, C6, Tyr^{7}: C4	Phe^{4}: H*α*, H2/6, Tyr^{7}: H2, H3, H5, H6
	C1	138					
	C2	128.2	H2	7.31 (^{1}H, d, 7.4)	Phe^{4}: H3	Phe^{4}: C*β*, C1, C3, C4	Phe^{3}: NH, H*β1/2*, Phe^{4}: H*α*, H*β*, Tyr^{7}: H3/5
	C3	127.8	H3	7.26 (^{1}H, t, 7.4)	Phe^{4}: H2, H4	Phe^{4}: C1, C2, C4	Phe^{3}: H*β1/2*, Phe^{4}: H*β*, H2/6, H4
	C4	127.9	H4	7.16 (^{1}H, t, 7.4)	Phe^{4}: H3, H5	Phe^{4}: C2, C3, C5, C6	Phe^{4}: H3/5
	C5	127.8	H5	7.26 (^{1}H, t, 7.4)	Phe^{4}: H4, H6	Phe^{4}: C1, C2, C4	Phe^{3}: H*β1/2*, Phe^{4}: H*β*, H2/6, H4
	C6	128.2	H6	7.31 (^{1}H, d, 7.4)	Phe^{4}: H5	Phe^{4}: C*β*, C1, C4, C5	Phe^{3}: NH, H*β1/2*, Phe^{4}: H*α*, H*β*, Tyr^{7}: H3/5
	C=O	170.2	NH	7.50 (^{1}H, br d, 8.8)	Phe^{4}: H*α*	n/a	Phe^{3}: NH, H*β1*, Phe^{4}: H*α*
Trp^{5}	*α*	53.2	*α*	5.39 (^{1}H, br s)	Trp^{5}: NH, H*β1*, H*β2*	n/a	Trp^{5}: NH, H*β1*, H*β2*, H4, Arg^{6}: NH
	β	29.8	*β1*	3.03 (^{1}H, t, 14.2)	Trp^{5}: H*α*, H*β2*	Trp^{5}: C*α*, C3, C=O	Trp^{5}: NH, H*α*, H*β2*, H2, Arg^{6}: H*β*
	C2	124.1	*β2*	3.19 (^{1}H, br d, 14.2)	Trp^{5}: H*α*, H*β1*	n/a	Trp^{5}: H*α*, H*β1*, H2, H4, Arg^{6}: H*β*
	C3	108.5	(NH)1	10.58 (^{1}H, s)	Trp^{5}: H2	Trp^{5}: C2, C3, C8	Trp^{5}: H2, H7
	C4	118.1	H2	6.98 (^{1}H, m)	Trp^{5}: (NH)1	Trp^{5}: C3, C8, C9	Trp^{5}: (NH)1, H*β1*, Arg^{6}: H*β*
	C5	118.3	H4	7.46 (^{1}H, d, 8.3)	Trp^{5}: H5	Trp^{5}: C3, C5, C6, C8, C9	Leu^{2}: H*δ2*, Trp^{5}: NH, H*α*, H*β1/2*, H5
	C6	130.2	H5	6.85 (^{1}H, d, 8.3)	Trp^{5}: H4	Trp^{5}: C9	Leu^{2}: H*α*, H*β*, H*δ2*, Phe^{3}: NH, Phe^{4}: H2/6, Trp^{5}: H*α*, H4
	C7	114.1	H7	6.77 (^{1}H, s)	n/a	Trp^{5}: C5	Leu^{2}: NH, H*α*, H*β*, Trp^{5}: (NH)1
	C8	136.4					
	C9	124.8	NH	8.43 (^{1}H, d, 7.0)	Trp^{5}: H*α*	n/a	Phe^{4}: H*α*, Trp^{5}: H*α*, H*β1*, Arg^{6}: NH
	C=O	171.2					
Arg^{6}	*α*	52.2	*α*	4.34 (^{1}H, dd, 6.9, 13.6)	Arg^{6}: NH, H*β*	Trp^{5}: C=O	Arg^{6}: NH, H*α*, H*β1*, H*β2*, H*γ*, H*δ*, Tyr^{7}: NH
	β	30.1	*β*	1.49 (2H, m)	Arg^{6}: H*α*, H*γ1/2*	Arg^{6}: C*α*	Arg^{6}: NH, H*α*, H*γ1/2*, H*δ*
	γ	24.9	*γ1*	1.35 (^{1}H, m)	Arg^{6}: H*β*, H*γ2*, H*δ*	n/a	Arg^{6}: NH, H*α*, H*β*, H*γ2*, H*δ*
	δ	40.4	*γ2*	1.43 (^{1}H, m)	Arg^{6}: H*β*, H*γ1*, H*δ*	Arg^{6}: C*β*	Arg^{6}: NH, H*α*, H*β*, H*γ1*, H*δ*
	ε	157.8	*δ*	3.10 (2H, m)	Arg^{6}: *ε*(NH), H*γ1/2*	n/a	Arg^{6}: H*α*, H*β*, H*γ1/2*, H*ε*
	C=O	169.1	*ε(NH)*	7.54 (^{1}H, br s)	Arg^{6}: H*δ*	n/a	Arg^{6}: H*γ*, H*δ*
			NH	6.92 (^{1}H, br s)	Arg^{6}: H*α*	n/a	Trp^{5}: NH, H*α*, H*β1*, H*β2*, Arg^{6}: H*α*, H*β*, H*γ*, Tyr^{7}: H*α*
Tyr^{7}	*α*	53.2	*α*	4.41 (^{1}H, dd,12.3, 11.5)	Tyr^{7}: H*β1/2*, NH	Tyr^{7}: C=O	Arg^{6}: NH, Tyr^{7}: NH, H*β1/2*, H2, H6
	β	37.7	*β1*	3.13 (^{1}H, d, 12.3)	Tyr^{7}: H*α*, H*β2*	Tyr^{7}: C*α*, C1, C2, C6	Tyr^{7}: NH, H*α*, H*β2*, H2, H6
	C1	129.9	*β2*	2.46 (^{1}H, t, 12.3)	Tyr^{7}: H*α*, H*β1*	Tyr^{7}: C*α*, C1, C2, C6	Tyr^{7}: NH, H*α*, H*β1*, H2, H6
	C2	130.6	H2	6.98 (^{1}H, br s)	Tyr^{7}: H3	n/a	Phe^{4}: H*β*, Tyr^{7}: NH, H*α*, H*β1/2*, H3/5
	C3	130.6	H3	6.65 (^{1}H, d, 7.6)	Tyr^{7}: H2	Tyr^{7}: C1, C2, C4	Phe^{4}: H*β*, Tyr^{7}: H*α*, H*β1/2*, H2, H6
	C4	156.1	H5	6.65 (^{1}H, d, 7.6)	Tyr^{7}: H6	Tyr^{7}: C1, C4, C6	Phe^{4}: H*β*, Tyr^{7}: H*α*, H*β1/2*, H2, H6

	C5	130.6	H6	6.98 (^{1}H, br s)	Tyr^7: H5	n/a	Phe^4: Hβ, Tyr^7: NH, Hα, Hβ1/2, H3/5
	C6	130.6	NH	8.34 (^{1}H, br s)	Tyr^7: Hα	n/a	Arg^6: Hα, Tyr^7: Hα, Hβ, H2, H6
	C=O	172.5					

a – overlapping sig nals, b – signals cannot be assigned by 2D NMR, br s – broad singlet, d – doublet, dd – double doublet, t – triplet, m – multiplet**.**

Table 4.8 MS/MS analysis of linear LysC-proteolytic core peptide of ManBURP-1xQLFFWRY. Proteomic data MassIVE ID: MSV000097169

Fragment (z)	m(obs) [Da]	m(calc) [Da]	Δm [ppm]
b3 (1+)	339.2023	339.2027	1.2
b4 (1+)	453.2458	453.2457	0.2
y4 (1+)	462.2379	462.2381	0.4
y5 (1+)	575.3219	575.3222	0.5
b5 (1+)	616.3091	616.3090	0.1
b6 (1+)	730.3513	730.3519	0.8
y7 (1+)	844.5042	844.5074	3.8
b7 (1+)	877.4199	877.4203	0.5
y8 (1+)	957.5915	957.5914	0.1
b8 (1+)	1005.4786	1005.4789	0.3
y9 (1+)	1072.6192	1072.6184	0.7
b9 (2+)	1102.5309	1102.5317	0.7
b10 (1+)	1231.5740	1231.5743	0.2
y11 (1+)	1243.6860	1243.6828	2.6
y12 (2+)	1399.7827	1399.7839	0.8
y13 (2+)	1513.8281	1513.8268	0.9
b13 (2+)	1570.7383	1570.7285	6.3
b14 (2+)	1683.8045	1683.8126	4.8
y16 (3+)	1961.0489	1961.0498	0.4
b17 (2+)	2164.0193	2164.0287	4.3
y18 (3+)	2294.1956	2294.1975	0.8
y19 (3+)	2441.2646	2441.2660	0.6
y20 (3+)	2554.3457	2554.3500	1.7
y21 (3+)	2682.4079	2682.4086	0.2
y22 (3+)	2796.4493	2796.4515	0.8
y23 (3+)	2893.5032	2893.5043	0.4
y24 (3+)	3022.5446	3022.5469	0.7
y25 (3+)	3119.5973	3119.5996	0.7
b25 (3+)	3167.4905	3167.4871	1.1
y26 (3+)	3247.6580	3247.6582	0.1
y27 (3+)	3394.7237	3394.7266	0.9
y28 (3+)	3508.7684	3508.7696	0.3
b28 (4+)	3549.7446	3549.7564	3.3
y29 (3+)	3671.8289	3671.8329	1.1

b31 (4+)	3846.9230	3846.9252	0.6
b32 (4+)	3977.9670	3977.9657	0.3
y32 (5+)	4010.9844	4010.9872	0.7
b33 (5+)	4106.0564	4106.0607	1.1
y33 (5+)	4124.0704	4124.0712	0.2

Table 4.9 MS/MS analysis of putative bicyclic LysC-proteolytic core peptide of ManBURP-1xQLFFWRY. Proteomic data MassIVE ID: MSV000097169

Fragment (z)	m(obs) [Da]	m(calc) [Da]	Δm [ppm]
b3 (1+)	339.2027	339.2027	0.0
b4 (1+)	453.2456	453.2457	0.2
y4 (1+)	462.2379	462.2381	0.4
y5 (1+)	575.3225	575.3222	0.5
b5 (1+)	616.3088	616.3090	0.3
b6 (1+)	730.3512	730.3519	1.0
y6 (1+)	731.4170	731.4233	8.6
b7 (1+)	877.4197	877.4203	0.7
y8 (1+)	957.5915	957.5914	0.1
b8 (1+)	1005.4785	1005.4789	0.4
b9 (2+)	1102.5309	1102.5317	0.7
b10 (1+)	1231.5736	1231.5743	0.6
y11 (1+)	1243.6853	1243.6828	2.0
b11 (1+)	1328.6351	1328.6270	6.1
y12 (2+)	1399.7859	1399.7839	1.4
b13 (2+)	1570.7179	1570.7285	6.7
y20* (3+)	2550.3152	2550.3187	1.4
y21* (3+)	2678.3783	2678.3773	0.4
y22* (3+)	2792.4203	2792.4202	0.0
y23* (3+)	2889.4718	2889.4730	0.4
y24* (3+)	3018.5141	3018.5156	0.5
y25* (3+)	3115.5680	3115.5683	0.0
b25* (3+)	3163.4567	3163.4558	0.3
y26* (3+)	3243.6257	3243.6269	0.4
y27* (3+)	3390.6962	3390.6953	0.3
y28* (3+)	3504.7283	3504.7382	2.8
b31* (4+)	3842.8850	3842.8939	2.3
b32* (4+)	3973.9362	3973.9344	0.4
y32* (5+)	4006.9534	4006.9558	0.6
b33* (5+)	4102.0264	4102.0293	0.7

Table 4.10. ^{1}H-NMR (800 MHz) and ^{13}C-NMR (200 MHz) data for moroidin-QLLVWRNH in DMSO-d$_6$.

Residue	C	δ(^{13}C) [ppm]	H	δ(^{1}H) (int, m, *J*) [ppm]	COSY correlations	HMBC correlations	NOESY correlations
PyroGlu[1]	*α*	55.1	*α*	4.10 (^{1}H, dd, 3.0, 8.6)	PyroGlu[1]: NH, H*β1*, H*β2*	PyroGlu[1]: C*β*, C*δ*	PyroGlu[1]: NH, H*β1*, H*β2*, Hγ, Leu[2]: NH
	β	25.4	*β1*	2.25 (^{1}H, m)	PyroGlu[1]: H*α*, H*β2*, H*γ*	PyroGlu[1]: C*α*, C*γ*, C=O	PyroGlu[1]: H*α*, H*β2*, H*γ*
	γ	28.9	*β2*	1.70 (^{1}H, m)	PyroGlu[1]: H*α*, H*β1*, H*γ*	n/a	PyroGlu[1]: H*α*, H*β1*, H*γ*
	δ	177.3	*γ*	2.09 (2H, t, 8.7)	PyroGlu[1]: H*β1*, H*β2*	PyroGlu[1]: C*α*, C*β*, C*δ*	PyroGlu[1]: H*β1*, H*β2*
	C=O	172.2	NH	7.88 (^{1}H, s)	PyroGlu[1]: H*α*	PyroGlu[1]: C*α*, C*β*, C*γ*, C*δ*	PyroGlu[1]: H*α*
Leu[2]	*α*	54.7	*α*	4.85 (^{1}H, t, 10.3)	Leu[2]: NH, H*β*	PyroGlu[1]: C=O, Leu[2]: C*β*, C=O	Leu[2]: NH, H*γ*, H*δ2*, Leu[3]: NH, Val[4]: NH, Trp[5]: H5
	β	51.5	*β*	3.06 (^{1}H, dd, 3.3, 11.1)	Leu[2]: H*α*, H*γ*	Leu[2]: C*α*, C*γ*, C*δ1*, C*δ2*, Trp[5]: C5, C6, C7	Leu[2]: NH, H*α*, H*γ*, H*δ1*, Trp[5]: H7
	γ	26.6	*γ*	2.14 (^{1}H, m)	Leu[2]: H*β*, H*δ1*, H*δ2*	Leu[2]: C*δ1*	Leu[2]: NH, H*β*, H*δ1*, H*δ2*
	δ1	21.7	*δ1*	0.75 (3H, m)	Leu[2]: H*γ*	Leu[2]: C*β*, C*γ*, C*δ2*	Leu[2]: H*β*, H*γ*, H*δ2*, Trp[5]: H7
	δ2	17.1	*δ2*	0.83 (3H, d, 6.4)	Leu[2]: H*γ*	Leu[2]: C*β*, C*γ*, C*δ1*	Leu[2]: H*α*, H*γ*, H*δ1*, Trp[5]: H5
	C=O	170.3	NH	8.55 (^{1}H, d, 8.8)	Leu[2]: H*α*	PyroGlu[1]: C=O, Leu[2]: C*α*	PyroGlu[1]: H*α*, Leu[2]: H*α*, H*β*, H*γ*
Leu[3]	*α*	52	*α*	3.96 (^{1}H, br t, 10.7)	Leu[3]: NH, H*β*	Leu[2]: C=O, Leu[3]: C*β*, C=O	Leu[3]: NH, H*β1*, H*β2*, H*δ1*, Val[4]: NH
	β	41.8	*β*	1.18/1.46 (ea ^{1}H, dd, 5.1, 9.7)	Leu[3]: H*α*, H*γ*	Leu[3]: C*α*, C*γ*, C*δ1*, C*δ2*	Leu[3]: H*α*, H*γ*, H*δ1*, H*δ2*, Val[4]: NH
	γ	23.8	*γ*	1.42 (^{1}H, m)	Leu[3]: H*β*, H*δ1*, H*δ2*	Leu[3]: C*δ1*, C*δ2*	Leu[3]: H*β*, H*δ1*, H*δ2*
	δ1	20.8	*δ1*	0.68 (3H, d, 6.0)	Leu[3]: H*γ*	Leu[3]: C*β*, C*γ*, C*δ2*	Leu[3]: H*α*, H*β*, H*γ*, H*δ2*
	δ2	23	*δ2*	0.78 (3H, d, 6.4)	Leu[3]: H*γ*	Leu[3]: C*β*, C*γ*, C*δ1*	Leu[3]: H*β*, H*γ*, H*δ1*
	C=O	171.5	NH	8.38 (^{1}H, d, 9.2)	Leu[3]: H*α*	Leu[3]: C=O	Leu[2]: H*α*, Leu[3]: H*α*, H*β*, Val[4]: NH
Val[4]	*α*	57.5	*α*	3.67 (^{1}H, t, 7.7)	Val[4]: NH, H*β*	Val[4]: C*β*, C*γ1*, C*γ2*, C=O	Val[4]: NH, H*β*, H*γ1*, H*γ2*, Trp[5]: NH
	β	30.8	*β*	1.84 (^{1}H, m)	Val[4]: H*α*, H*γ*	Val[4]: C*α*, C*γ1*, C*γ2*	Val[4]: NH, H*α*, H*γ1*, H*γ2*
	γ1	18.2	*γ1*	0.78 (3H, d, 5.8)	Val[4]: H*β*	Val[4]: C*α*, C*β*, C*γ2*	Val[4]: NH, H*α*, H*β*
	γ2	18.3	*γ2*	0.79 (3H, d, 5.8)	Val[4]: H*β*	Val[4]: C*α*, C*β*, C*γ1*	Val[4]: NH, H*α*, H*β*
	C=O	169.2	NH	6.92 [a] (^{1}H)	Val[4]: H*α*	Leu[4]: C=O	Leu[2]: H*α*, Leu[3]: NH, H*β*, Val[4]: H*α*, H*γ*
Trp[5]	*α*	48.8	*α*	5.51 (^{1}H, dd, 8.7, 14.8)	Trp[5]: NH, H*β1*, H*β2*	n/a	Trp[5]: NH, H*β1*, H*β2*, H4, Arg[6]: NH
	β	27.7	*β1*	2.80 (^{1}H, dd, 11.3, 15.3)	Trp[5]: H*α*, H*β2*	Trp[5]: C2	Trp[5]: NH, H*α*, H*β2*, Arg[6]: NH, His[8]: H2
	C2	124.6	*β2*	3.64 (^{1}H, dd, 4.2, 15.3)	Trp[5]: H*α*, H*β1*	n/a	Trp[5]: H*α*, H*β1*, C4 Arg[6]: NH, Asn[7]: NH, His[8]: H2
	C3	103.4	(NH)1	11.80 (^{1}H, s)	n/a	Trp[5]: C2, C3	Trp[5]: H7, His[8]: H4
	C4	119.4	H4	7.55 (^{1}H, d, 8.2)	Trp[5]: H5	Trp[5]: C2, C3, C6, C8	Leu[3]: H*δ*, Trp[5]: H*α*, H*β*, H5, Arg[6]: NH
	C5	119.8	H5	7.01 (^{1}H, d, 8.2)	Trp[5]: H4	Trp[5]: C2, C7	Leu[2]: H*α*, Hδ2, Leu[3]: NH, Trp[5]: H4
	C6	131.7	H7	6.93 (^{1}H, s)	n/a	Leu[2]: C*β*, Trp[5]: C2, C4, C5	Leu[2]: H*β*, H*γ*, H*δ1*, Trp[5]: H5, (NH)1
	C7	114.2	NH	8.15 (^{1}H, s)	Trp[5]: H*α*	n/a	Val[4]: H*α*, H*γ*, Trp[5]: H*α*, H*β*
	C8	132.7					
	C9	127					
	C=O	171.6					
Arg[6]	*α*	52.8	*α*	4.35 (^{1}H, dd, 7.5, 13.1)	Arg[6]: NH, H*β1*, H*β2*	n/a	Arg[6]: NH, H*β1*, H*β2*, H*γ*, H*δ*, Asn[7]: NH
	β	28.5	*β1*	1.65 (^{1}H, m)	Arg[6]: H*α*, H*β2*, H*γ*	n/a	Arg[6]: NH, H*α*, H*β2*, H*γ*, H*δ*
	γ	24.8	*β2*	1.82 (^{1}H, m)	Arg[6]: H*α*, H*β1*, H*γ*	n/a	Arg[6]: NH, H*α*, H*β1*, H*γ*, H*δ*
	δ	40.3	*γ*	1.54 [a] (^{1}H)	Arg[6]: H*β1*, H*β2*, H*δ*	n/a	Arg[6]: NH, H*α*, H*β1*, H*β2*, H*δ*, H*ε*
	ε	156.8	*δ*	3.12 (2H, dd, 6.1, 12.2)	Arg[6]: *ε*(NH), H*γ*	n/a	Arg[6]: H*α*, H*β1*, H*β2*, H*γ*, H*ε*
	C=O	171.3 [b]	*ε*(NH)	7.69 (^{1}H, m)	Arg[6]: H*δ*	n/a	Arg[6]: H*γ*, H*δ*
			NH	8.67 (^{1}H, d, 7.5)	Arg[6]: H*α*	Trp[5]: C=O	Trp[5]: H*α*, H*β1*, H*β2*, Arg[6]: H*α*, H*β1*, H*β2*, H*γ*, Asn[7]: NH, His[8]: H2
Asn[7]	*α*	49.5	*α*	4.54 (^{1}H, m)	Asn[7]: NH, H*β1*, H*β2*	Asn[7]: C*γ*	Asn[7]: NH, His[8]: NH, H2
	β	36.5	*β1*	2.42 (^{1}H, d, 14.5)	Asn[7]: H*α*, *β2*	n/a	Trp[5]: H4, Asn[7]: H*α*, H*β2*, NH
	γ	172.6	*β2*	2.86 (^{1}H, d, 14.5)	Asn[7]: H*α*, *β1*	n/a	Trp[5]: H4, Asn[7]: H*α*, H*β2*, NH
	C=O	169.4 [b]	NH	7.79 (^{1}H, br s)	Gly[7]: H*α*	n/a	Trp[5]: H*β1*, Arg[6]: H*α*, NH Asn[7]: H*α*
His[8]	*α*	51.1	*α*	4.66 (^{1}H, br s)	His[8]: NH, H*β1*, H*β2*	n/a	His[8]: NH, H*β1*, H*β2*, H2
	β	27.2	*β1*	3.12 [a] (^{1}H)	His[8]: H*α*, H*β2*	His[8]: C*α*	His[8]: H*α*, H*β2*
	C1	131.7	*β2*	3.28 (^{1}H, d, 14.5)	His[8]: H*α*, H*β1*	n/a	His[8]: H*α*, H*β1*, H2, NH
	C2	120.8	H2	7.85 (^{1}H, s)	n/a	His[8]: C*β*	Trp[5]: H*β1*, H*β2*, His[8]: NH, H*α*, H*β1*, H*β2*
	C4	136.5	H4	9.02 (^{1}H, br s)	n/a	n/a	Trp[5]: (NH)1, Arg[6]: H*α*
	C=O	172.1 [b]	NH	7.68 [a] (^{1}H)	His[8]: H*α*	n/a	Asn[7]: NH, H*α*, His[8]: H*α*, H*β*

a – overlapping sig nals, b – signals cannot be assigned by 2D NMR, br s – broad singlet, d – doublet, dd – double doublet, t – triplet, m – multiplet.

Figure 4. Table 4.11 Chemotaxonomic distribution of stephanotic acid-type burpitides.

Species	Family	Order	Stephanotic acid QLxxW	Moroidin QLxxWx(xx)H	Glechomanin QLxxWx(xx)W	Mercurialin QLxxWx(xx)Y
Stellaria aquatica	Caryophyllaceae	Caryophyllales	x*	x*		
Bauhinia tomentosa	Fabaceae	Fabales	x	x		
Stephanotis floribunda	Apocynaceae	Gentianales	x			
Glechoma hederacea	Lamiaceae	Lamiales	x*		x*	
Mercurialis annua	Euphorbiaceae	Malpighiales	x*			x*
Sacciolepis striata	Poaceae	Poales	x*	x*		
Kerria japonica	Rosaceae	Rosales	x	x		
Dendrocnide moroides	Urticaceae	Rosales		x		
Orchidantha maxillari-oides	Lowiaceae	Zingiberales	x*	x*	x*	

x - Characterized chemotype & experimentally verified precursor peptide (x* - this study)

x - Characterized chemotype & no experimentally verified precursor peptide

Table 4.12 Cloning primers.

Name	Sequence
GheBURP-pEAQ-HT- fwd	TGCCCAAATTCGCGACCGGTATGGAATTCAAGATTCTCCCC
GheBURP-pEAQ-HT- rev	CCAGAGTTAAAGGCCTCGAGTTAATTTTTGGAAACCCACAC
SaqBURP-pH8-fwd	GAAAACTTGTACTTCCAGGCCCATGGCATGAGGTTCAAAATGCTTATTAGTTATATT
SaqBURP-pH8-rev	CTCGAATTCGGATCCGCCATGGTTACATAGCATGAGGAGAAGTATTAGC
OmaBURP1-pEAQ- HT-fwd	TGCCCAAATTCGCGACCGGTATGGATCGCCTCACCTCTG
OmaBURP1-pEAQ- HT-rev	CCAGAGTTAAAGGCCTCGAGTTACACAGTGCTAGTCGGTG
OmaBURP2-pEAQ- HT-fwd	TGCCCAAATTCGCGACCGGTATGGATCGCCTAACCTCTGC
OmaBURP2-pEAQ- HT-rev	CCAGAGTTAAAGGCCTCGAGCTACAGAATGCTACGCGGTGC

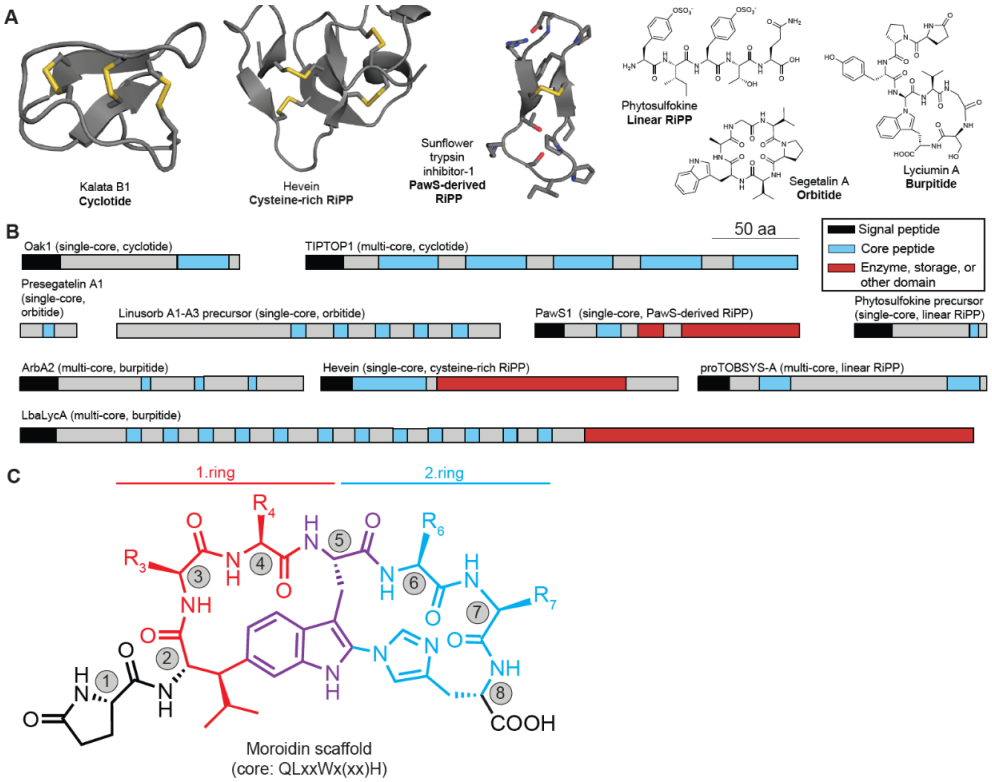


Figure 4.11 Plant RiPP structures and biosynthetic precursor peptides.

(**A**) Chemical structures of representative plant RiPPs. (**B**) Primary precursor peptide structures of representative plant RiPPs including single- and multi-core precursor peptides. (**C**) Residue numbering in moroidin scaffold. Residue numbers are in grey circles. The second ring can be contracted or extended by one residue 3. x is any proteinogenic amino acid. Abbreviations: aa- amino acid.

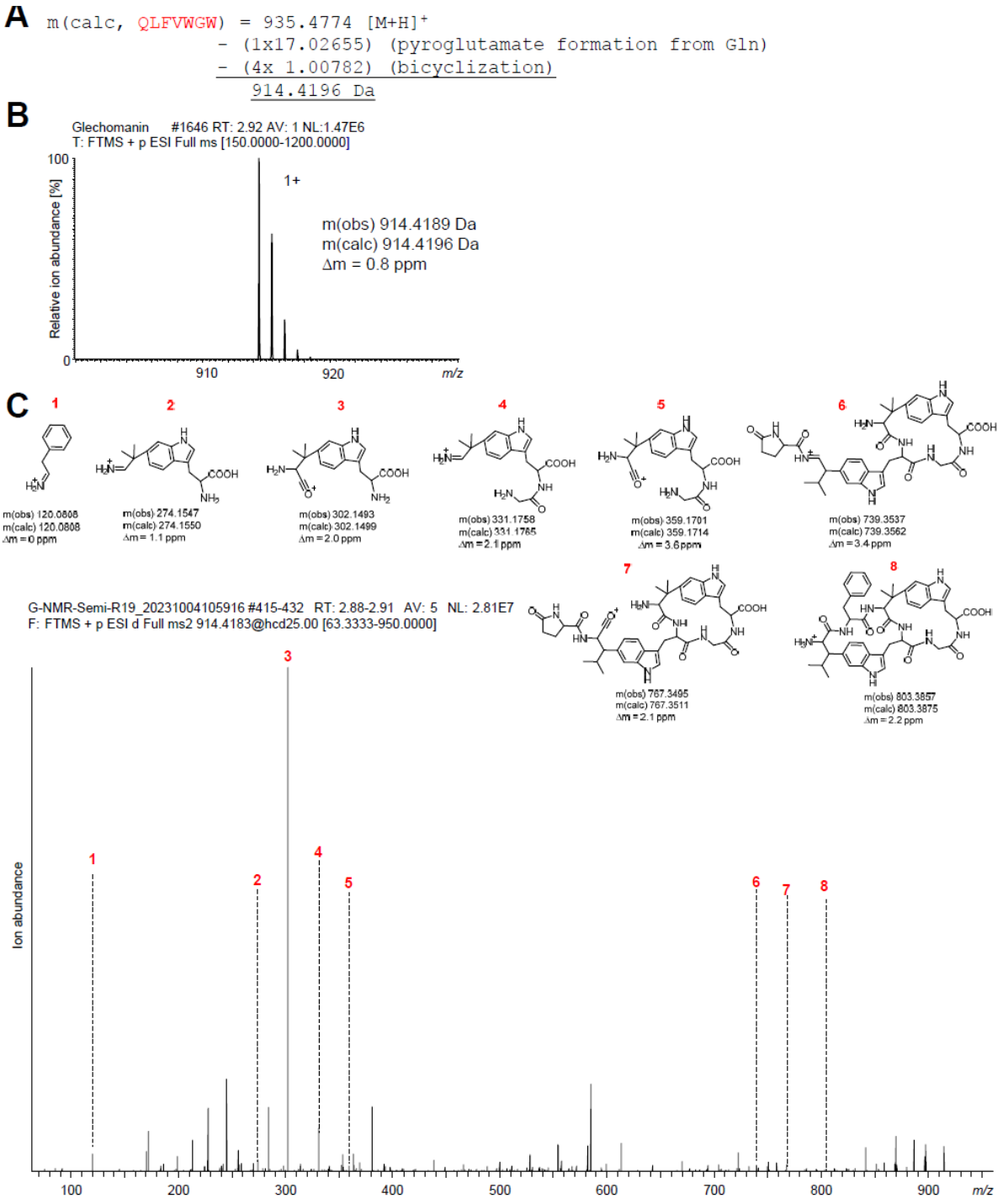


Figure 4.12 MS and MS/MS spectrum of glechomanin.

(**A**) Mass calculation of glechomanin from predicted core peptide sequence and post-translational modifications. (**B**) MS1 spectrum of glechomanin.(**C**) MS/MS spectrum of glechomanin and predicted fragments.

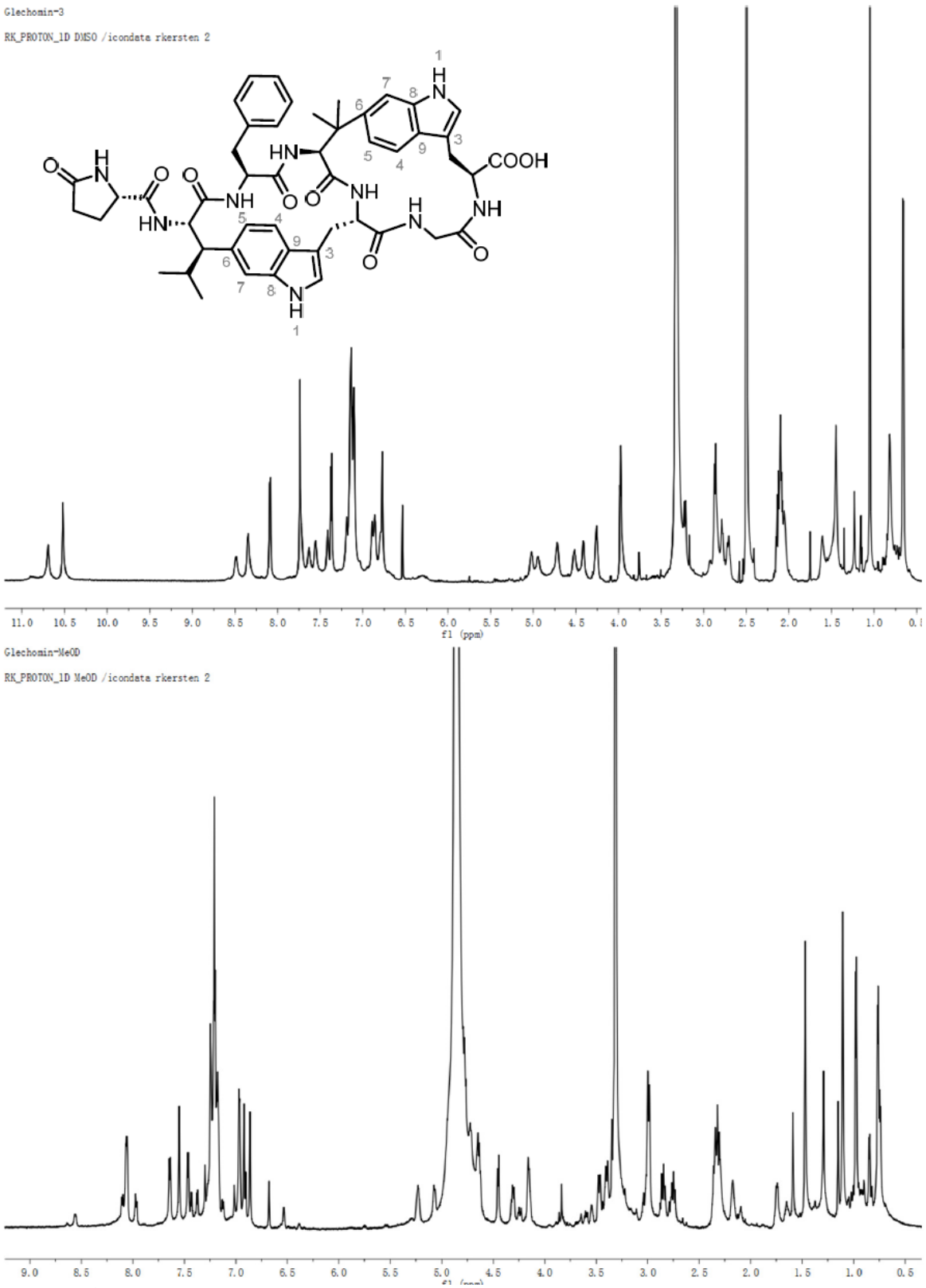


Figure 4.13 ^{1}H-NMR spectrum of glechomanin in DMSO-d_6 (800 MHz) and in methanol-d_4 (800 MHz).

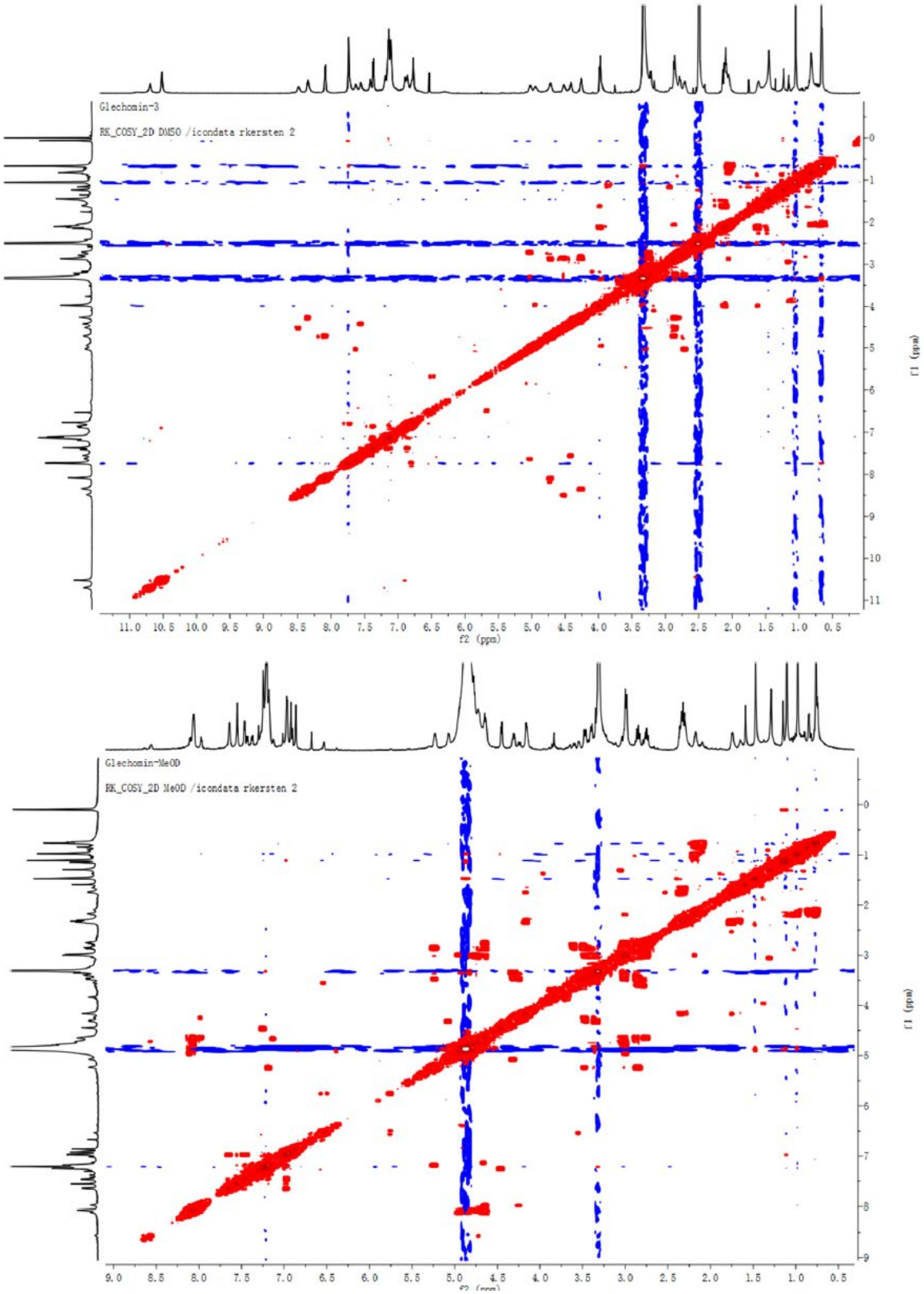


Figure 4.14 ^{1}H-^{1}H COSY-NMR spectrum of glechomanin in DMSO-d_6 and in methanol-d_4.

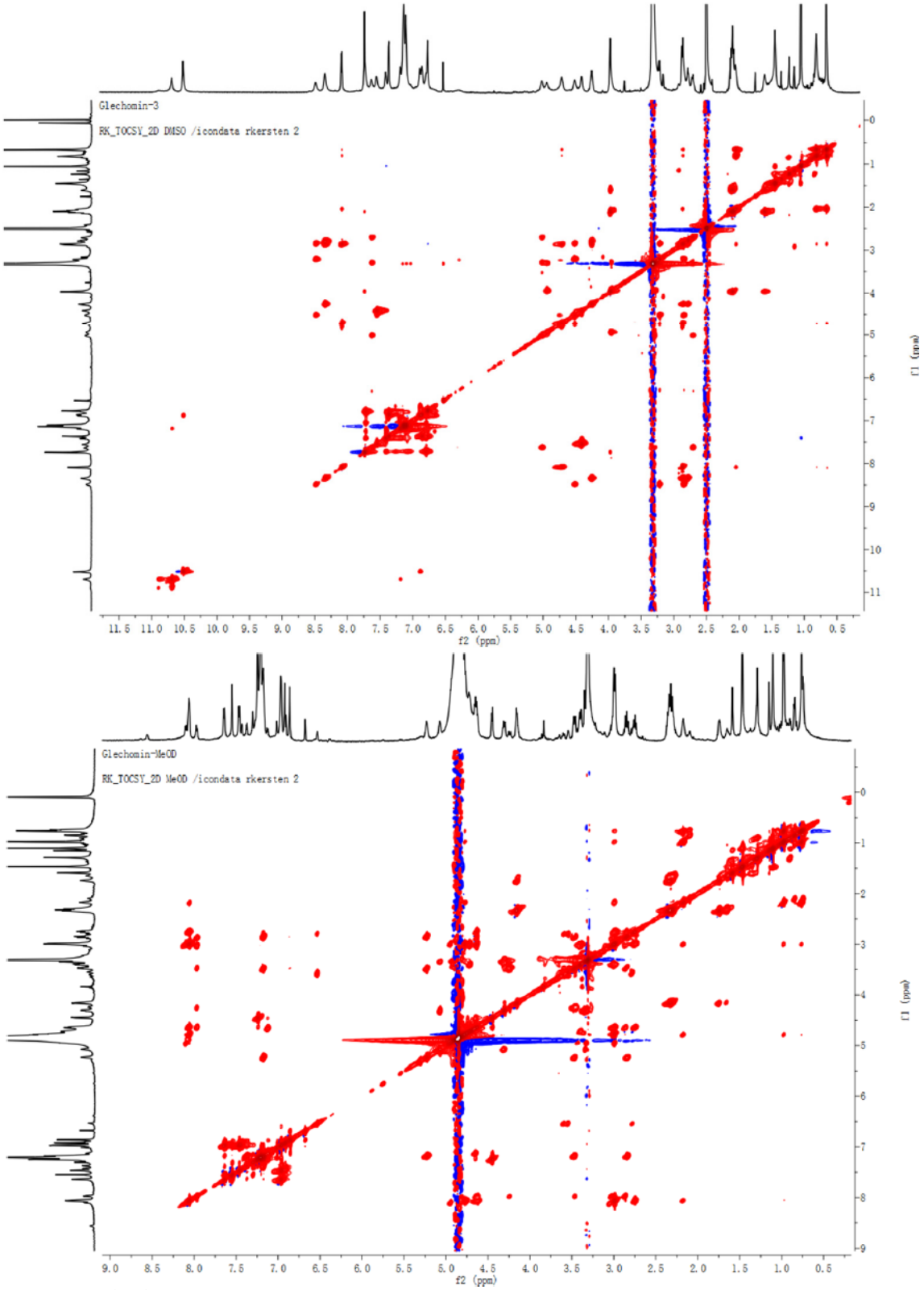


Figure 4.15 ^{1}H-^{1}H-TOCSY-NMR spectrum of glechomanin in DMSO-d_6 and in methanol-d_4.

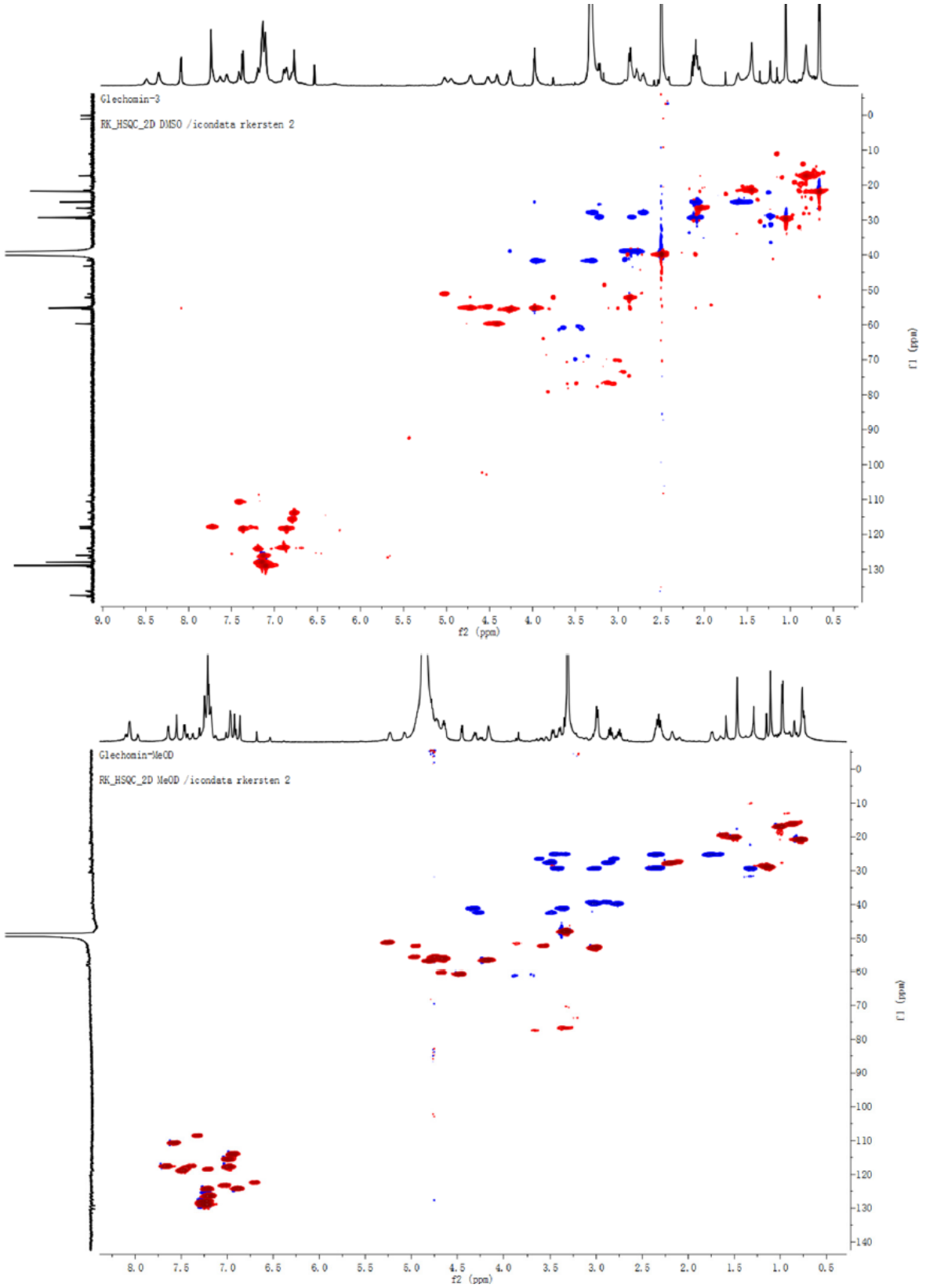


Figure 4.16 HSQC-NMR spectrum of glechomanin in DMSO-d_6 and in methanol-d_4.

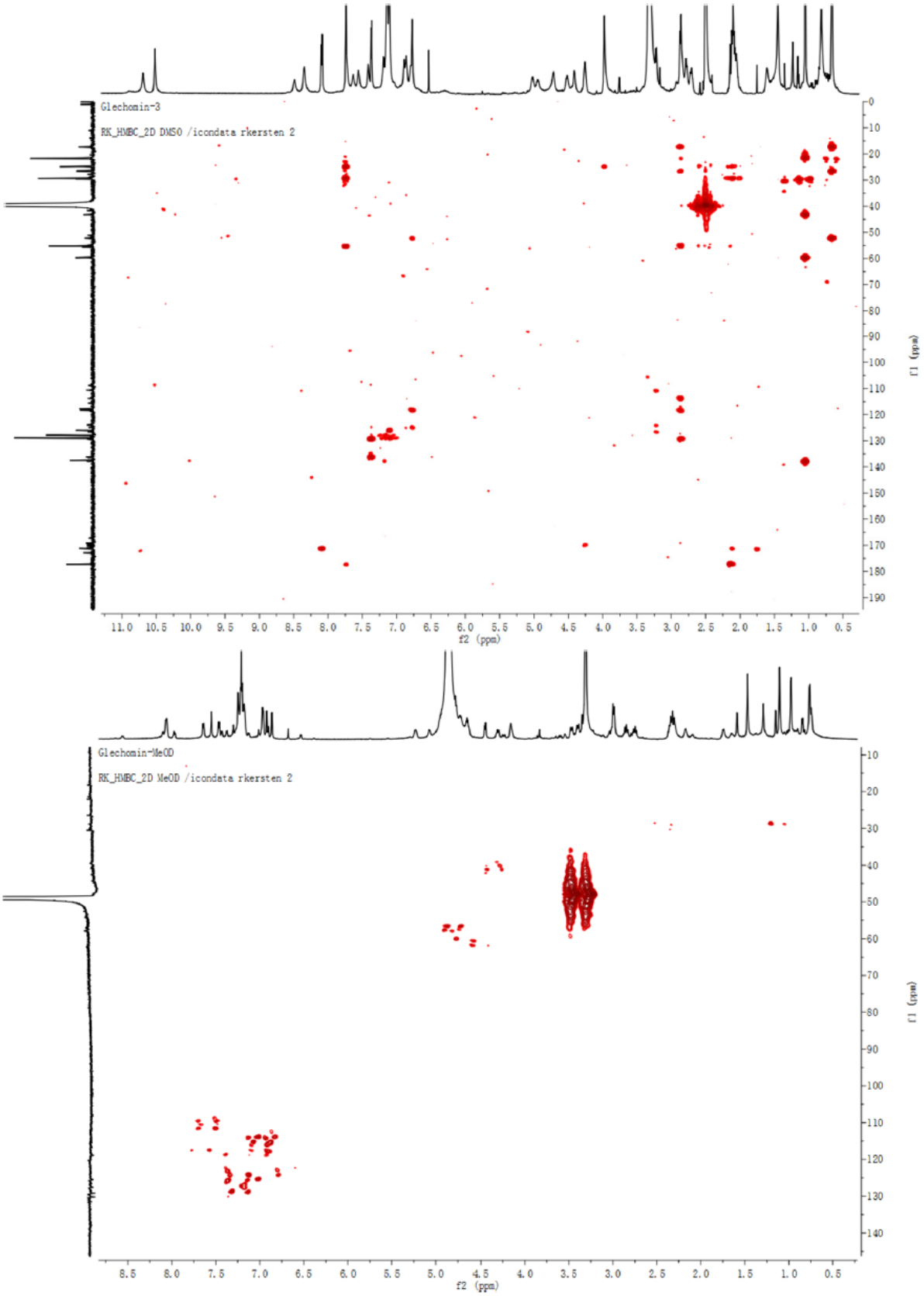


Figure 4.17 HMBC-NMR spectrum of glechomanin in DMSO-d_6 and in methanol-d_4.

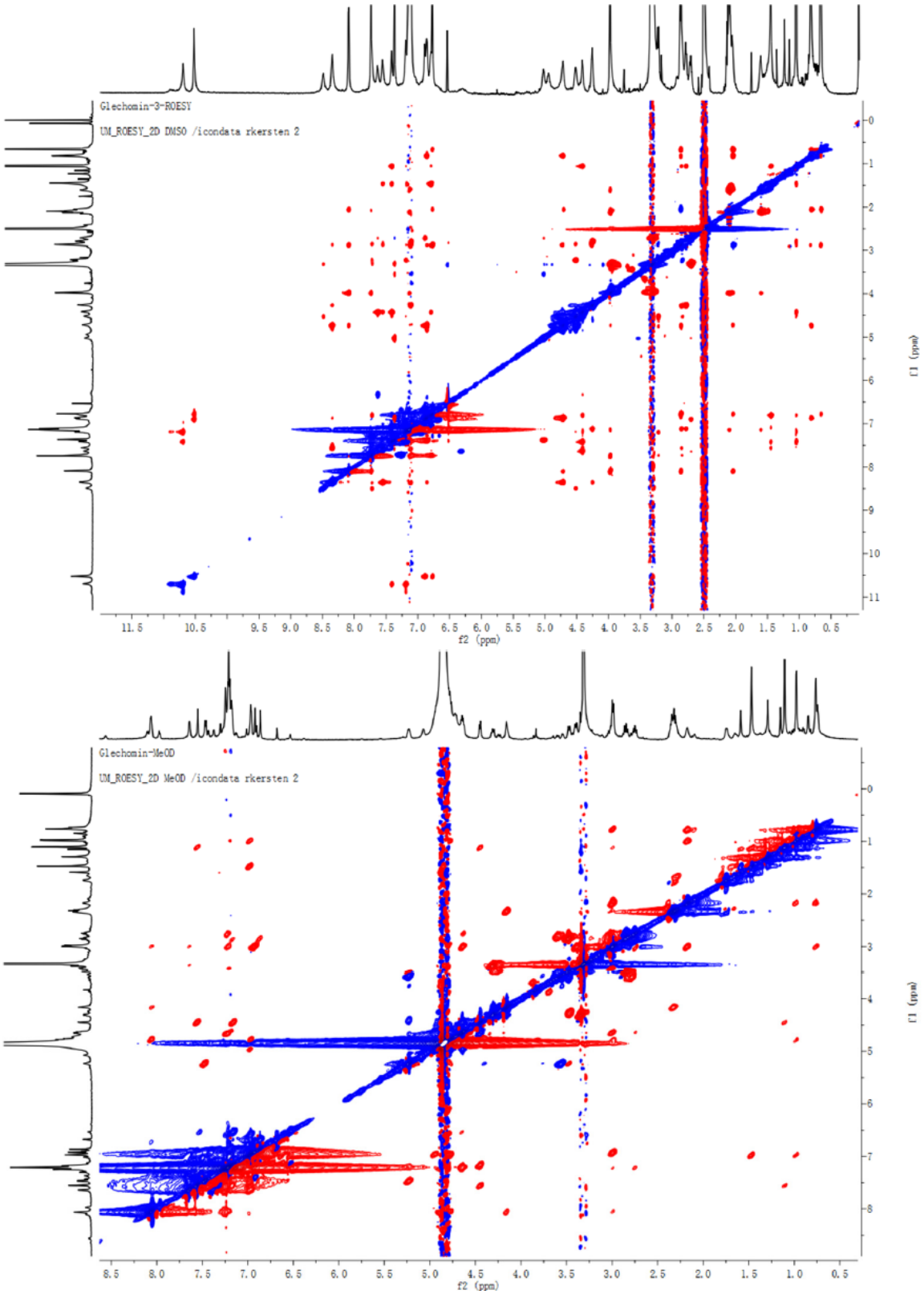


Figure 4.18 ROESY-NMR spectrum of glechomanin in DMSO-d_6 and in methanol-d_4.

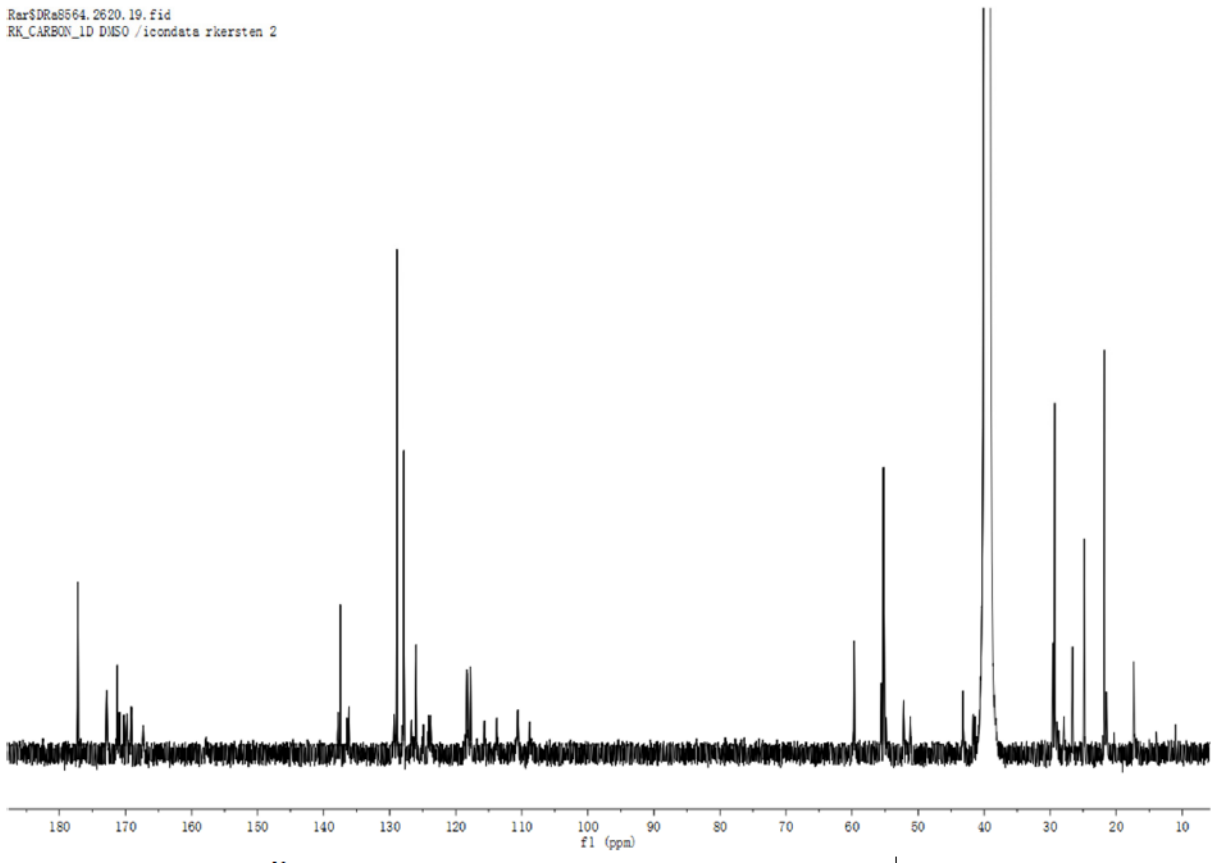


Figure 4.19 ^{13}C-NMR spectrum (200 MHz) of glechomanin in DMSO-d$_6$.

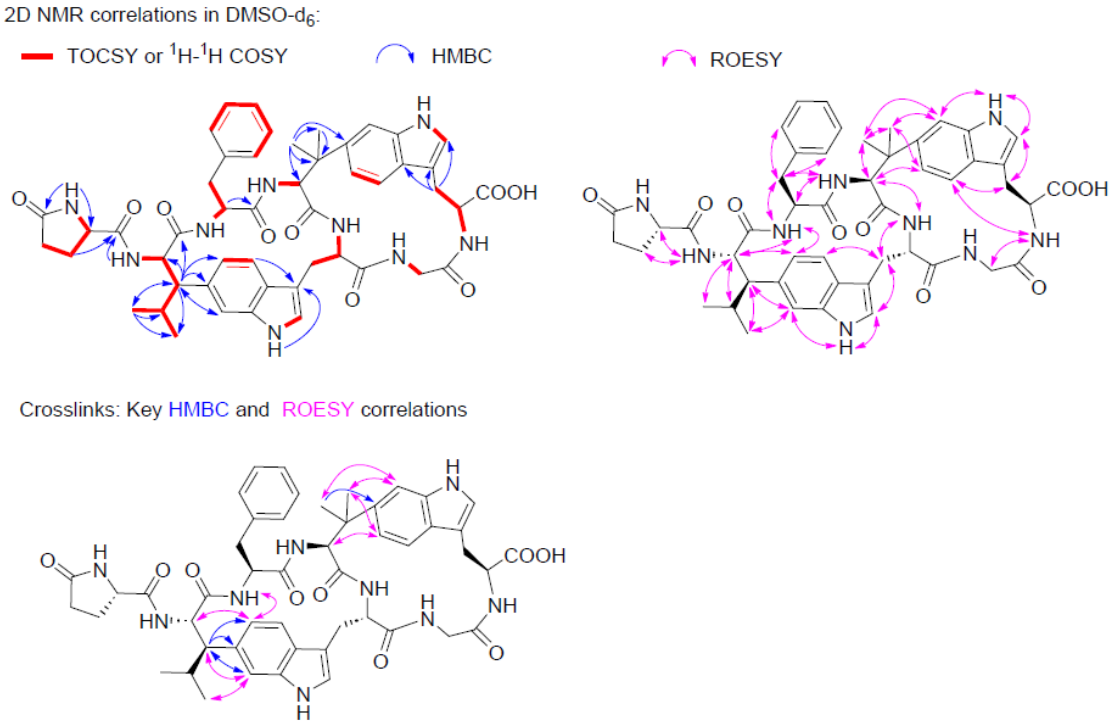


Figure 4.20 Key 2D NMR correlations of glechomanin in DMSO-d$_6$

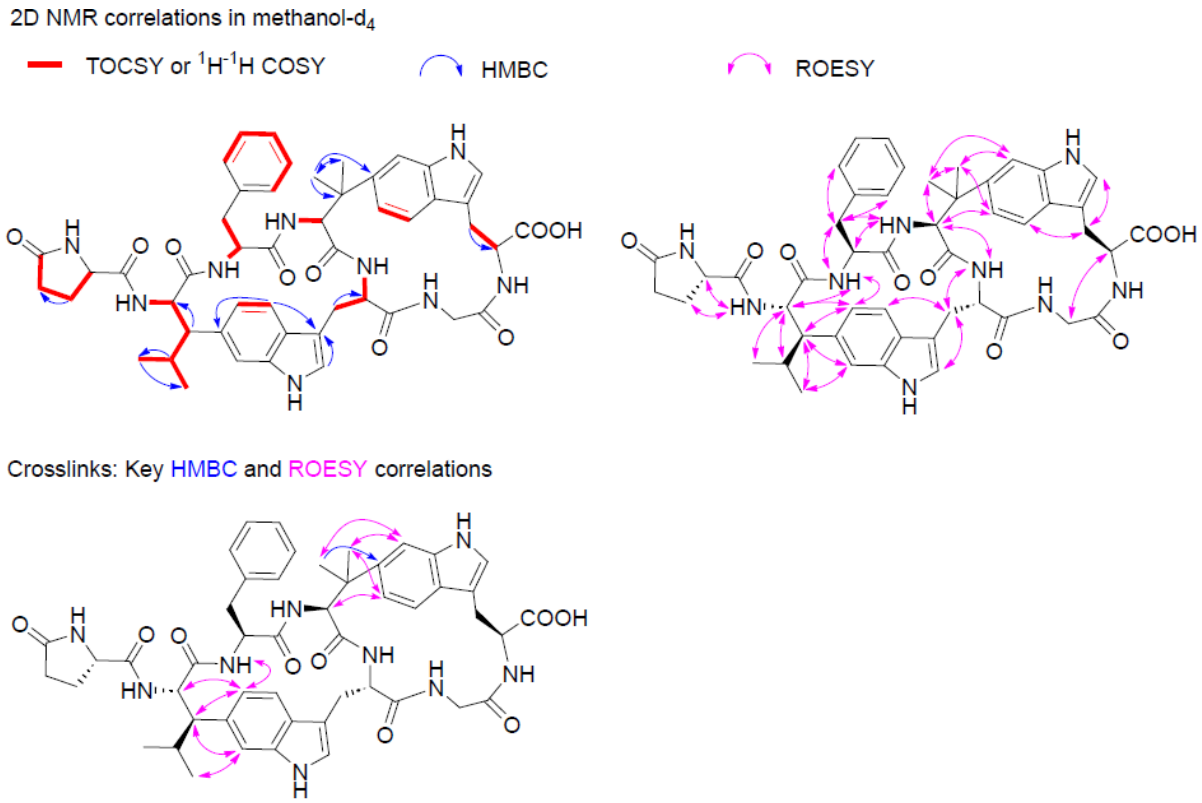


Figure 4.21 Key 2D NMR correlations of glechomanin in methanol-d_4.

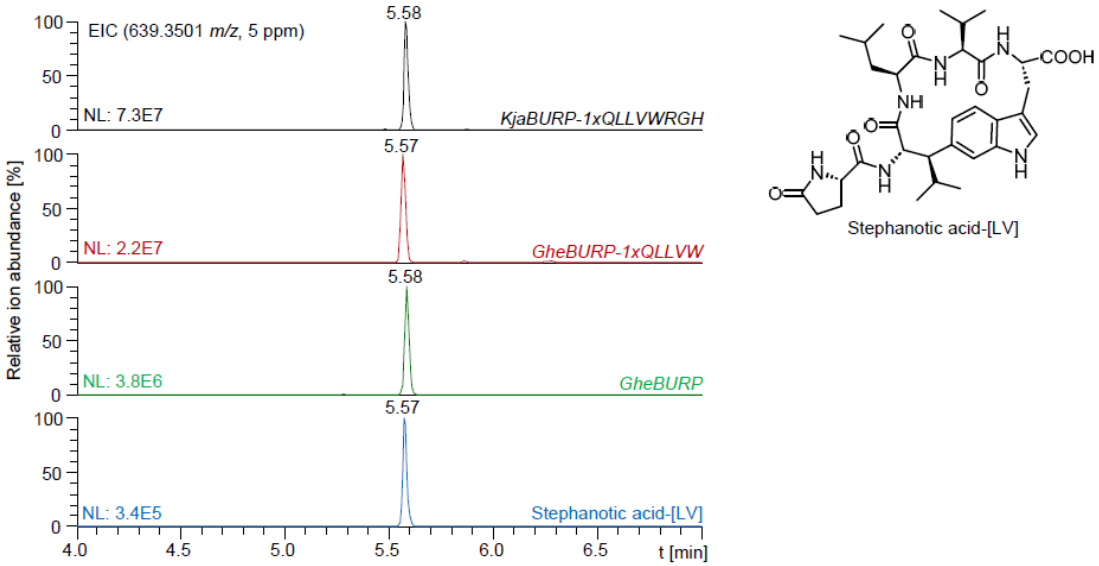


Figure 4.22 Stereochemical analysis of glechomanin.

LC-MS-based comparison of glechomanin firstring analytes compared to authentic standard stephanotic acid-[LV] after transient expression of KjaBURP-1xQLLVWRGH (moroidin cyclase), GheBURP (glechomanin cyclase including a QLLVW core peptide) and GheBURP-1xQLLVW (glechomanin cyclase with truncated substrate domain

to include one QLLVW core peptide) for 6 days in *N. benthamiana* via expression vector pEAQ-HT and *Agrobacterium tumefaciens* LBA4404 infiltration. Extracted ion chromatogram (EIC) traces of methanolic *N. benthamiana* leaf extracts are compared to authentic standard of stephanotic acid-[LV] (100 ng/mL).

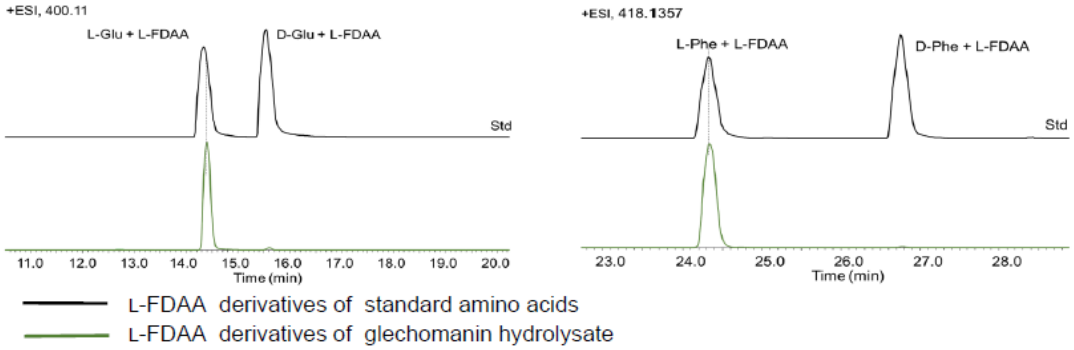


Figure 4.23 L-FDAA derivatization of the hydrolysate of glechomanin and amino acid standards.

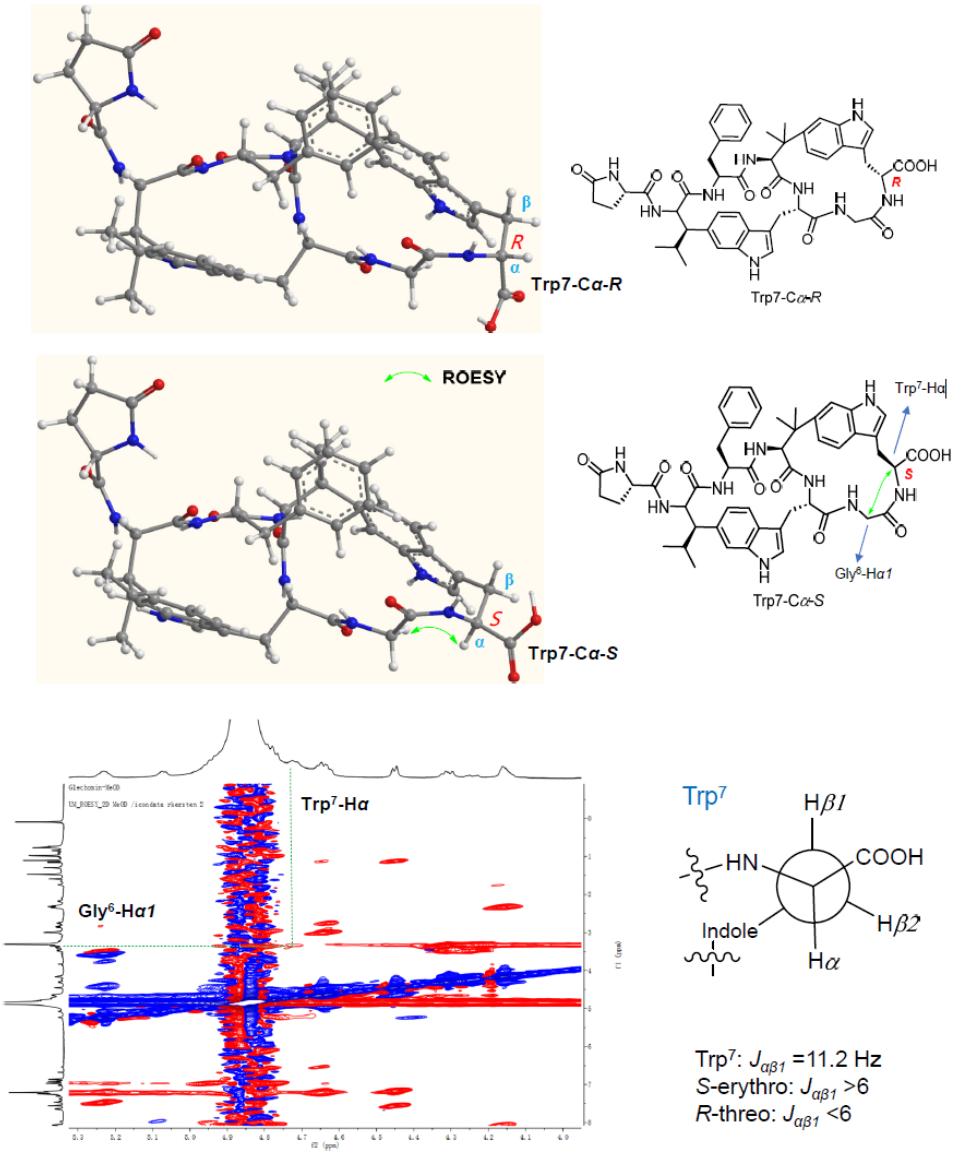


Figure 4.24 Glechomanin geometry optimization and ROESY analysis of Trp7-C*α* assignment

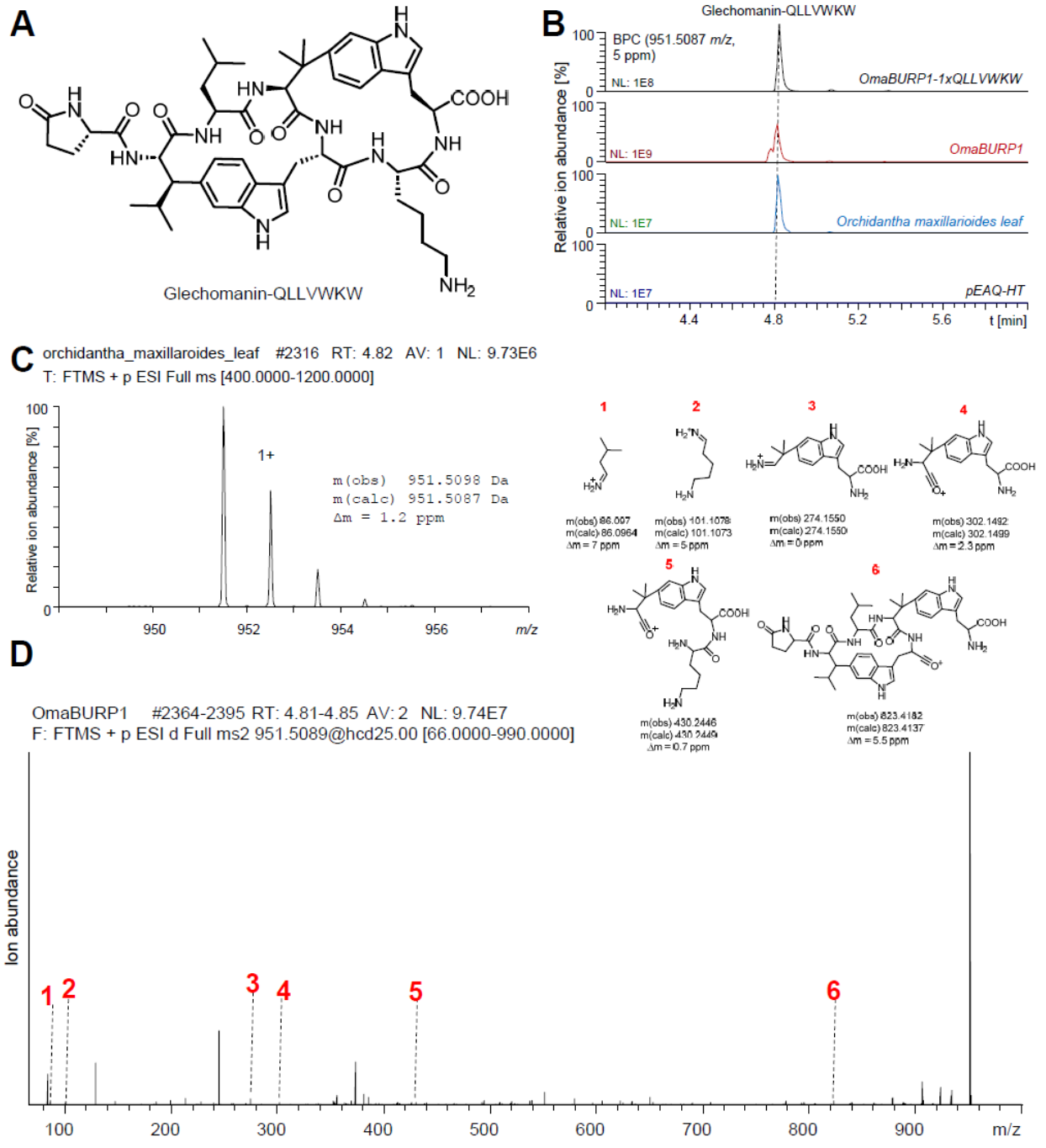


Figure 4.25 Characterization of glechomanin biosynthesis in *Orchidantha maxillarioides.*

(**A**) Predicted glechomanin-QLLVWKW chemotype. (**B**) LC-MS characterization of predicted glechomanin-QLLVWKW in methanolic extracts of *O. maxillarioides* leaf and transgenic *N. benthamiana* after 6 days of expression of OmaBURP1. (**C**) MS analysis of predicted glechomanin-QLLVWKW. (**D**) MS/MS analysis of predicted glechomanin-QLLVWKW.

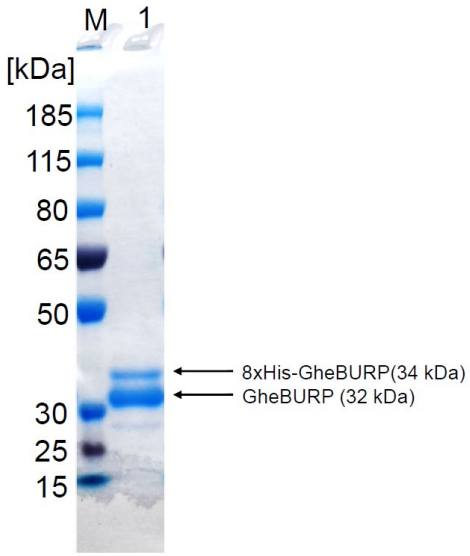


Figure 4.26 Glechomanin cyclase purification

SDS-PAGE of GheBURP-1xQLFVWGW after denature-refold purification. Source data are provided as a Source Data file.

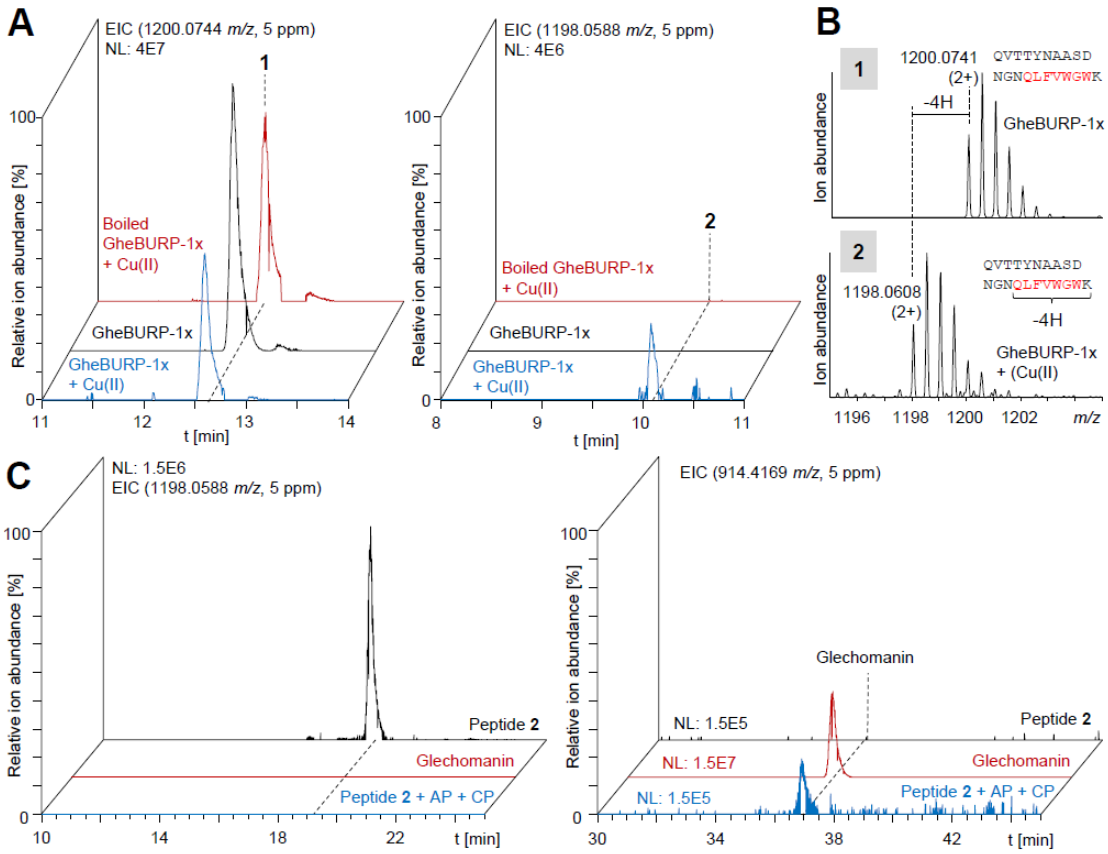


Figure 4.27 LC-MS analysis of *in vitro* reconstitution of GheBURP-1xQLFVWGW.

(**A**) Detection of bicyclic core peptide modifications in purified GheBURP-1xQLFVWGW after incubation with Cu(II) at pH 7. EIC traces of linear and modified tryptic core peptide QVTTYNAASDNGNQLFVWGWK. (**B**) MS

spectra of linear and modified tryptic core peptide. Core peptides are highlighted in red. (**C**) Detection of glechomanin after incubation of modified bicyclic tryptic core peptide QVTTYNAASDNGNQLFVWGWK (peptide 2) with aminopeptidase (AP) and carboxypeptidase (CP).

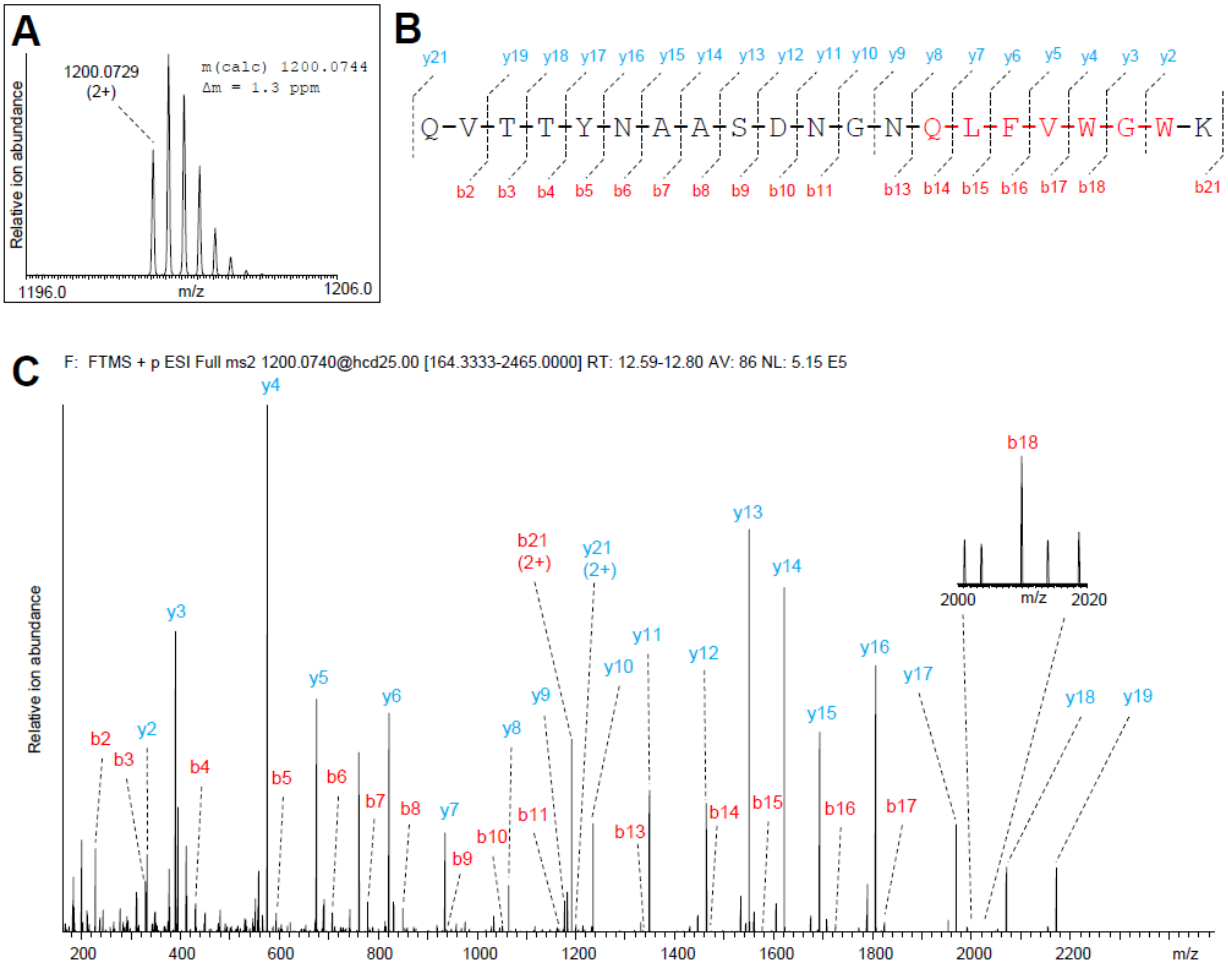


Figure 4.28 MS and MS/MS analysis of unmodified core peptide from *in vitro* reconstitution assay of GheBURP-1xQLFVWGW.

(A) MS1 signal of linear tryptic core peptide QVTTYNAASDNGNQLFVWGWK. Isotope selected for fragmentation labeled by grey box. (B) Detected tandemMS fragments of linear tryptic core peptide. (C) Tandem MS spectrum of linear tryptic core peptide with labeled fragment peaks.

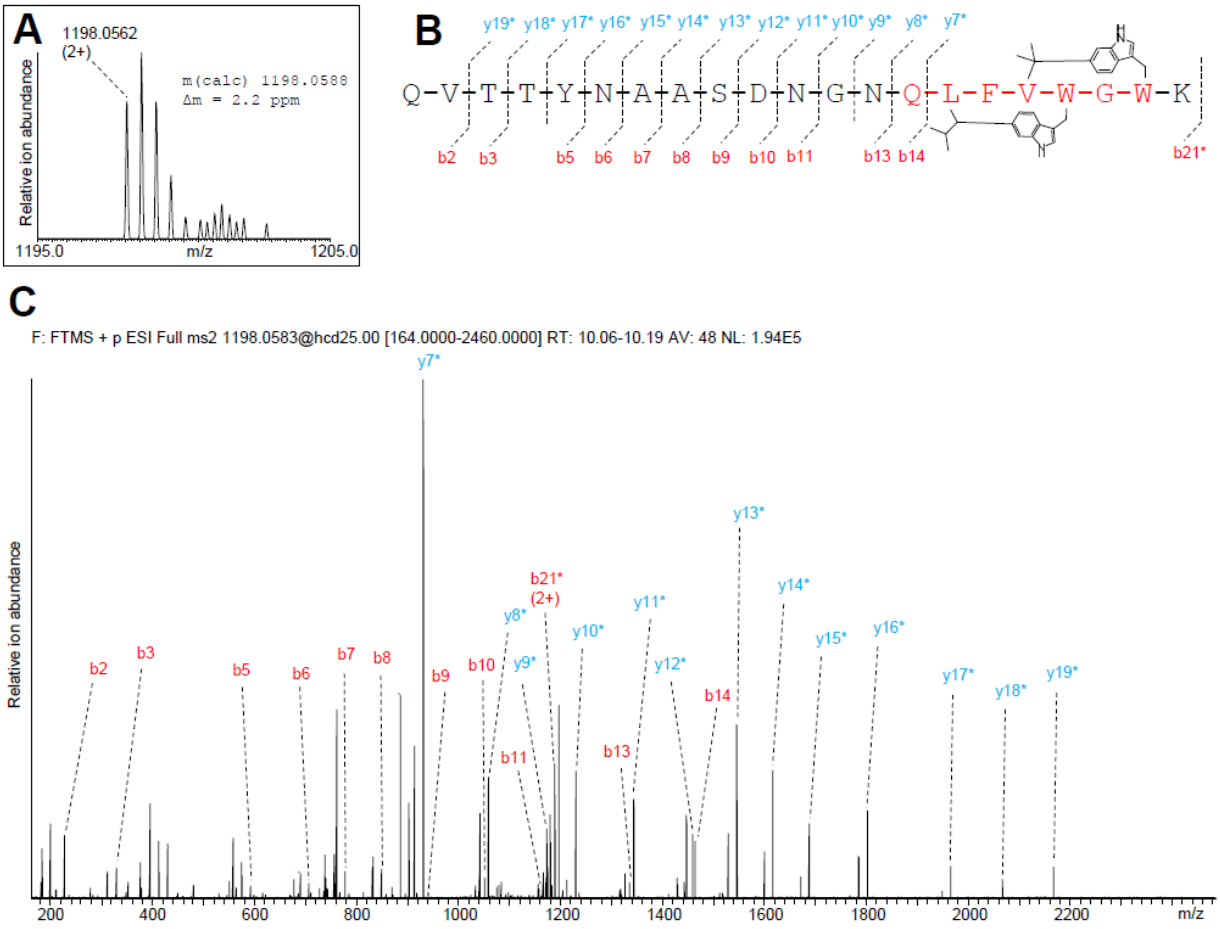


Figure 4.29 MS and MS/MS analysis of modified core peptide from *in vitro* reconstitution of GheBURP-1xQLFVWGW.

(**A**) MS1 signal of putative bicyclic tryptic core peptide QVTTYNAASDNGNQLFVWGWK. Isotope selected for fragmentation labeled by grey box. (**B**) Detected tandem MS fragments of putative bicyclic tryptic core peptide. (**C**) Tandem MS spectrum of putative bicyclic tryptic core peptide with labeled fragment peaks.

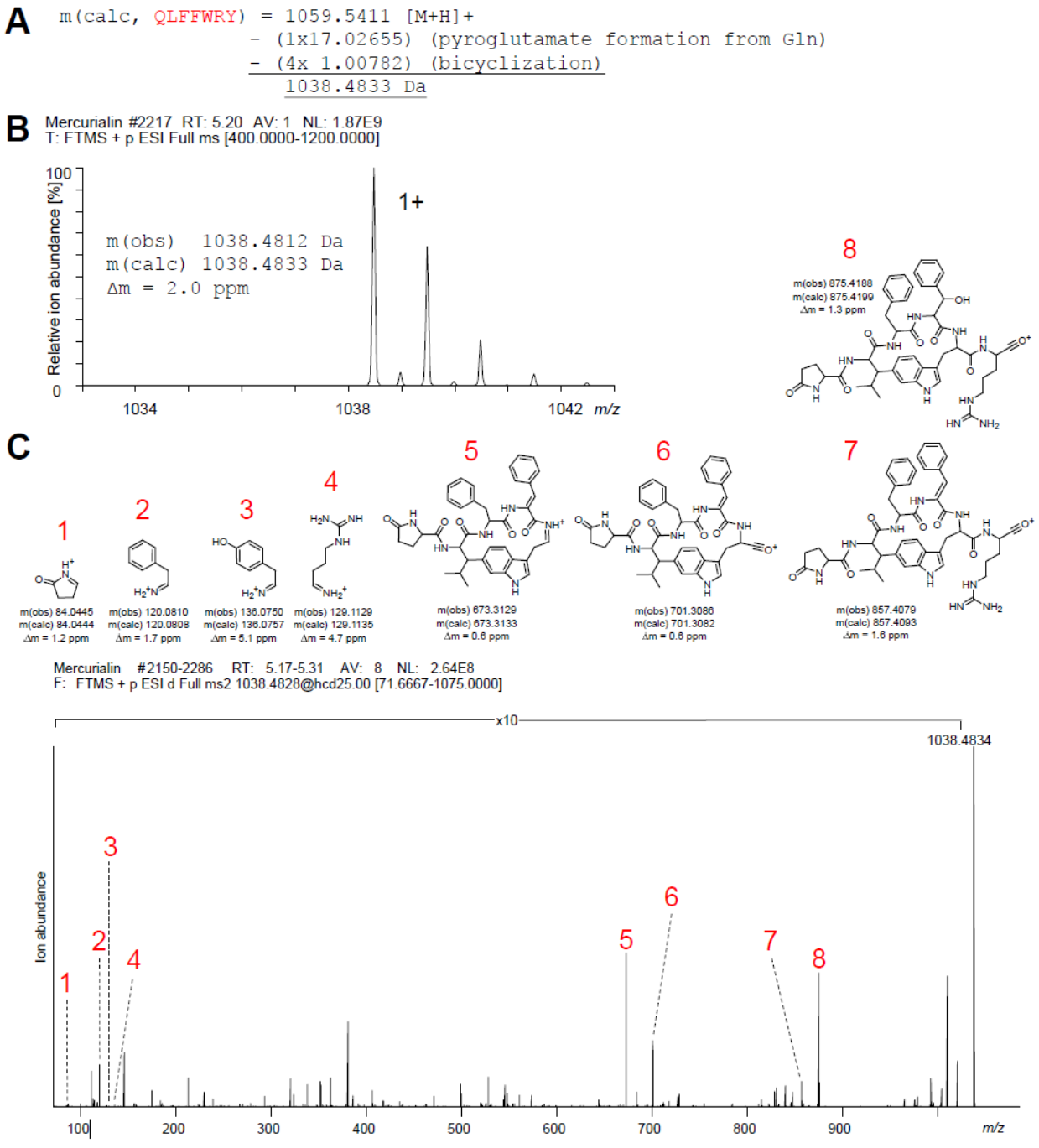


Figure 4.30 MS and MS/MS spectrum of mercurialin.

(**A**) Mass calculation of mercurialin from predicted core peptide sequence and post-translational modifications. (**B**) MS1 spectrum of mercurialin. (**C**) MS/MS spectrum of mercurialin and predicted fragments.

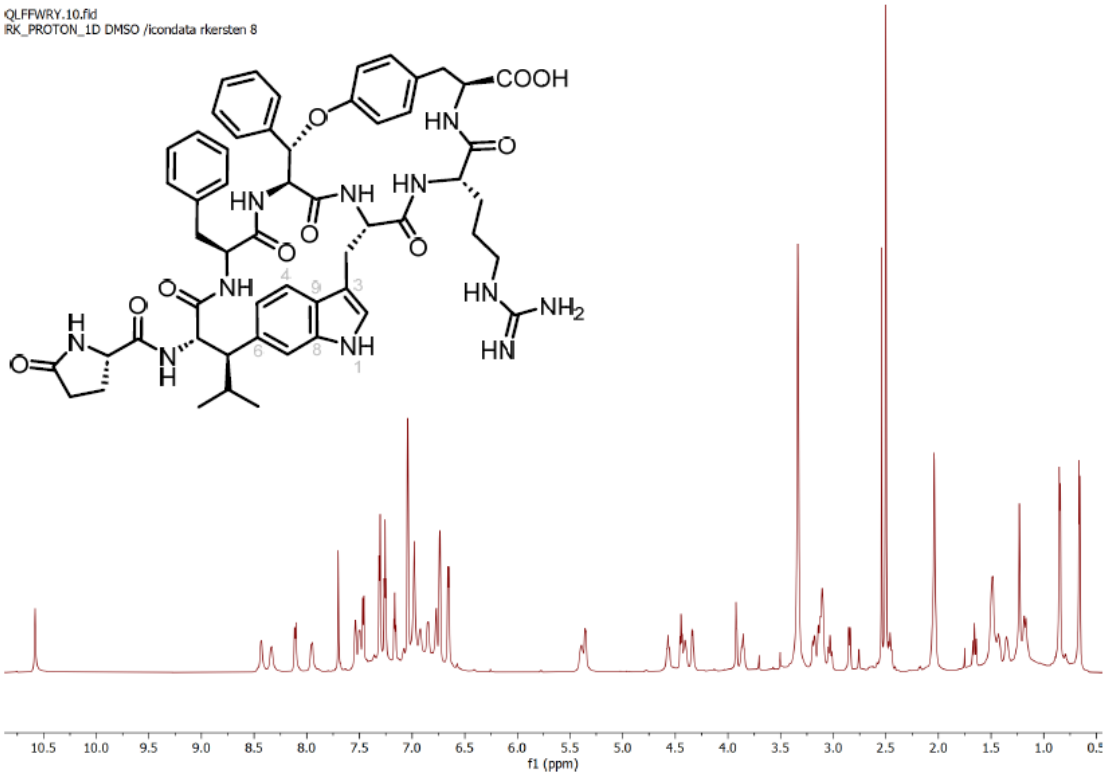


Figure 4.31 ^{1}H-NMR spectrum of mercurialin in DMSO-d_6 (800 MHz).

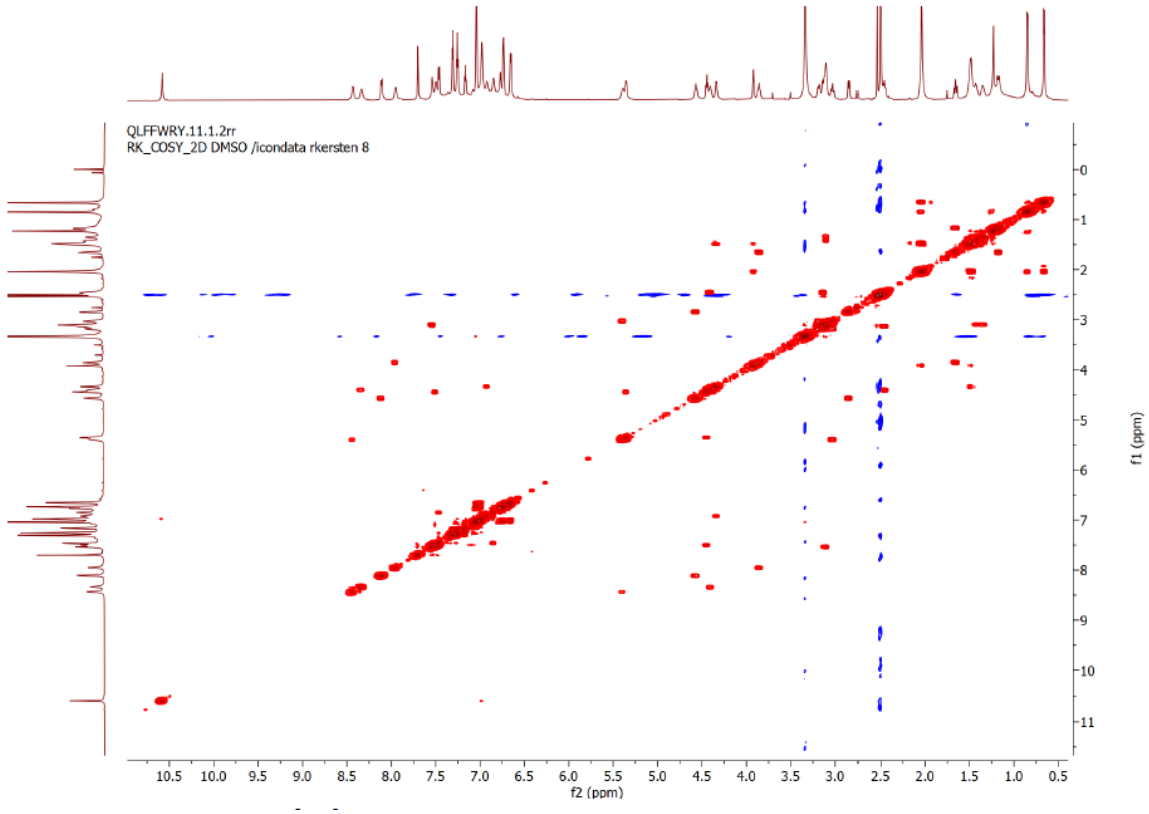


Figure 4.32 ^{1}H-^{1}H-COSY-NMR spectrum of mercurialin in DMSO-d_6.

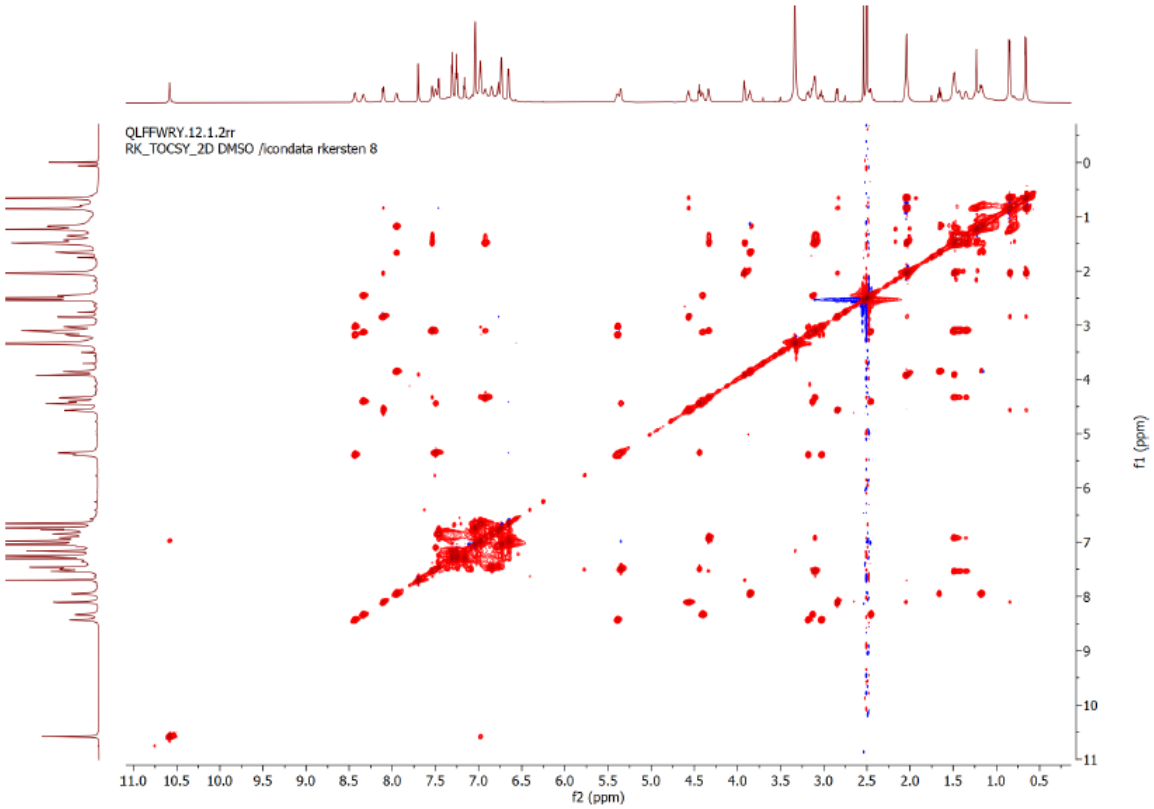


Figure 4.33 ^{1}H-^{1}H-TOCSY-NMR spectrum of mercurialin in DMSO-d$_6$.

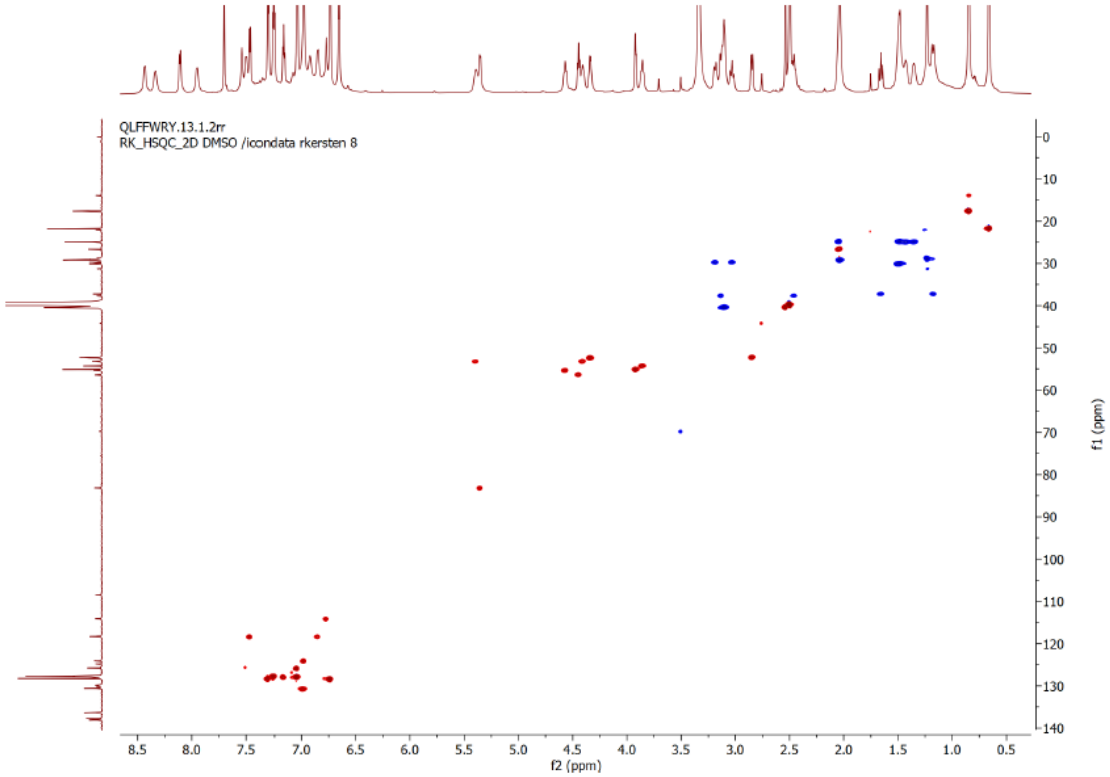


Figure 4.34 HSQC-NMR spectrum of mercurialin in DMSO-d_6.

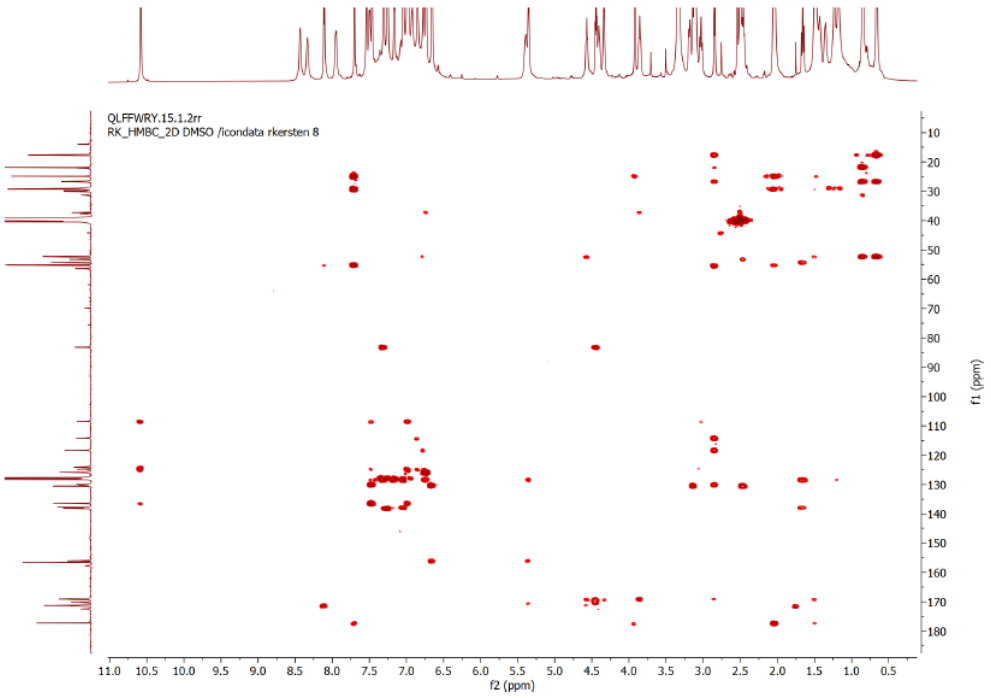


Figure 4.35 HMBC-NMR spectrum of mercurialin in DMSO-d_6.

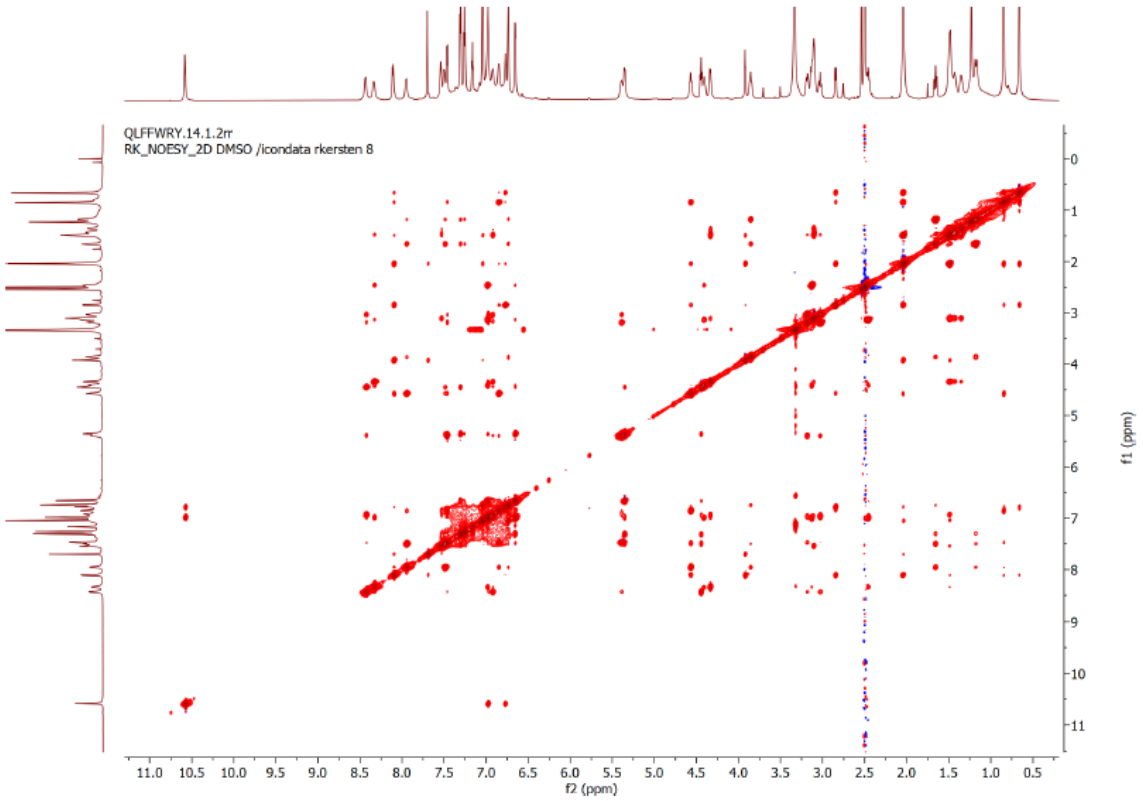


Figure 4.36 NOESY-NMR spectrum of mercurialin in DMSO-d_6.

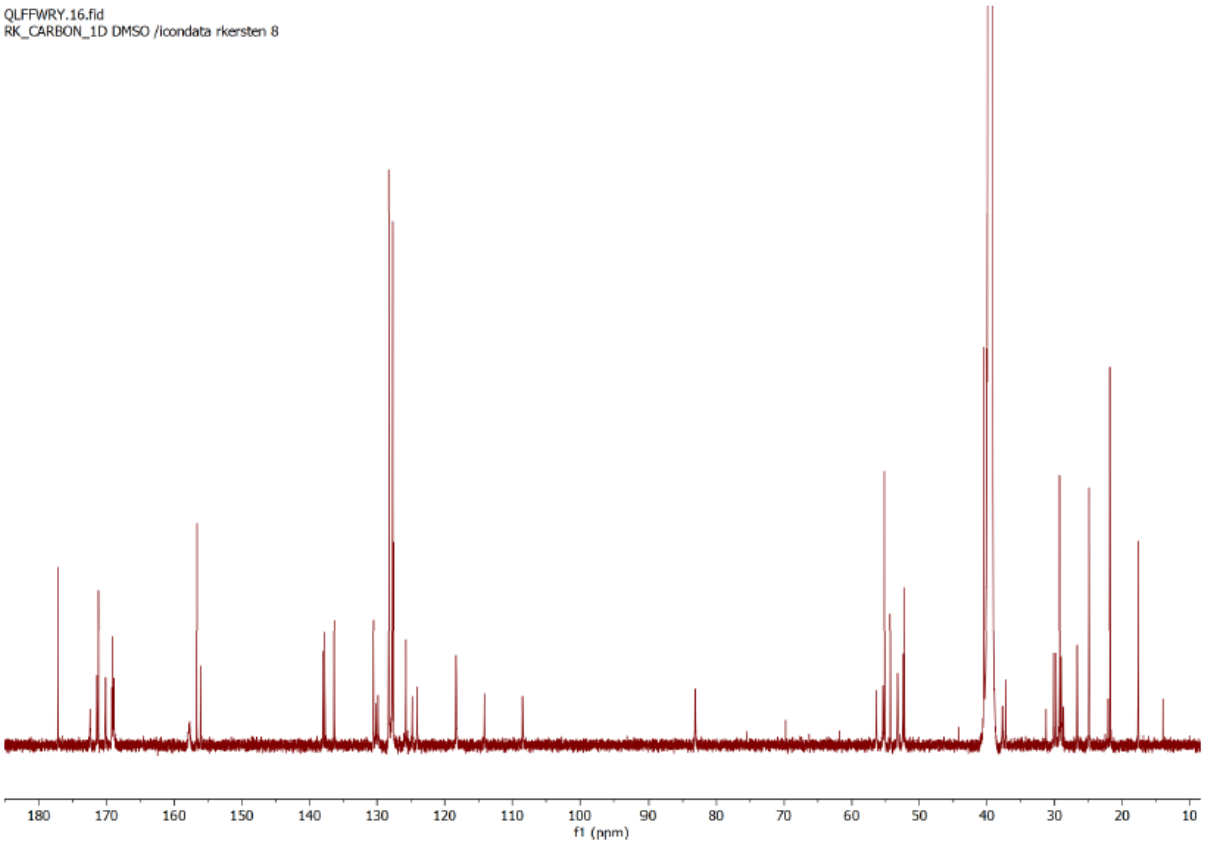


Figure 4.37 ^{13}C-NMR (200 MHz) spectrum of mercurialin in DMSO-d$_6$

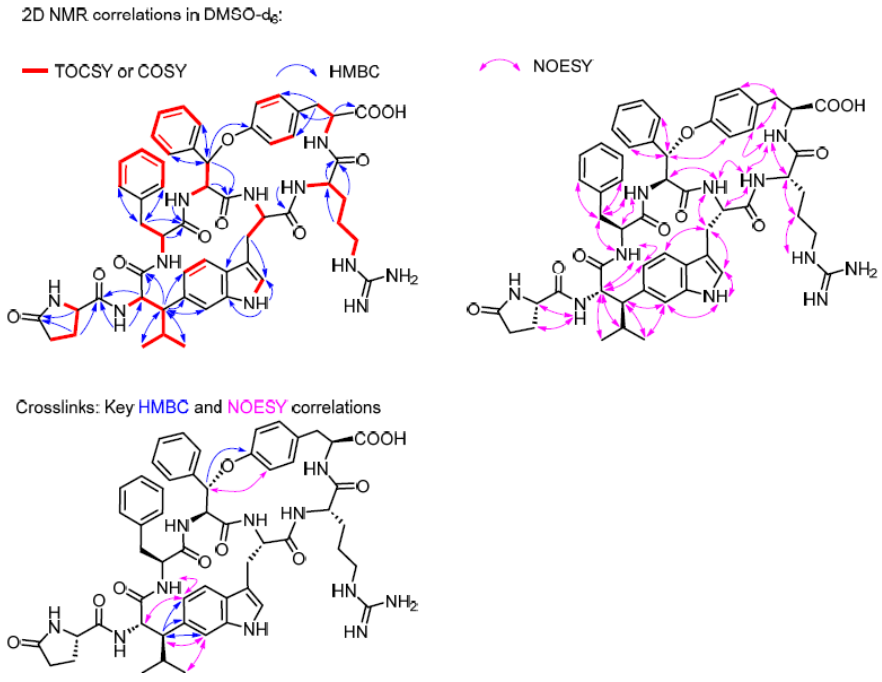


Figure 4.38 Key 2D NMR correlations of mercurialin.

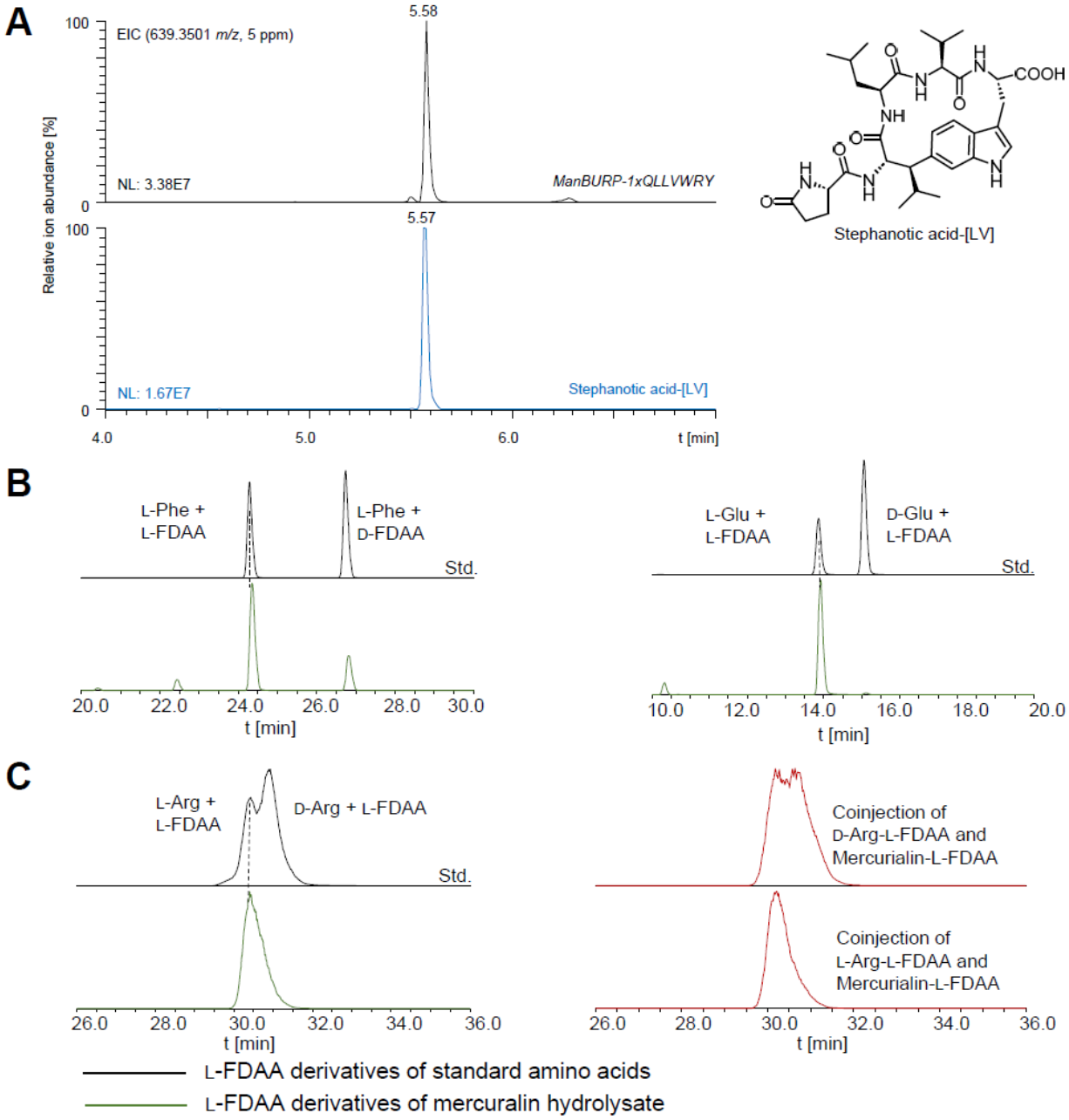


Figure 4.39 Stereochemical analysis of mercurialin.
(**A**) LC-MS-based comparison of mercurialin first-ring analyte compared to authentic standard stephanotic acid-[LV] after transient expression of ManBURP-1xQLLVWRY (mercurialin cyclase with truncated substrate domain to include one QLLVWRY core peptide) for 6 days in *N. benthamiana* via expression vector pEAQ-HT and *Agrobacterium tumefaciens* LBA4404 infiltration. Extracted ion chromatogram (EIC) traces of methanolic *N. benthamiana* leaf extract is compared to authentic standard of stephanotic acid-[LV] (100 ng/mL). (**B**) L-FDAA derivatization of the hydrolysate of mercurialin and amino acid standards. Analyzed by using Phenomenex Kinetex® 2.6 µm C18 reverse phase 100 Å 150 x 3 mm LC column. (**C**) L-FDAA derivatization of the hydrolysate of mercuralin and amino acid standards. Analyzed by using Higgins Analytical PROTO300 C4 5 µm 250 x 4.6 mm LC column.

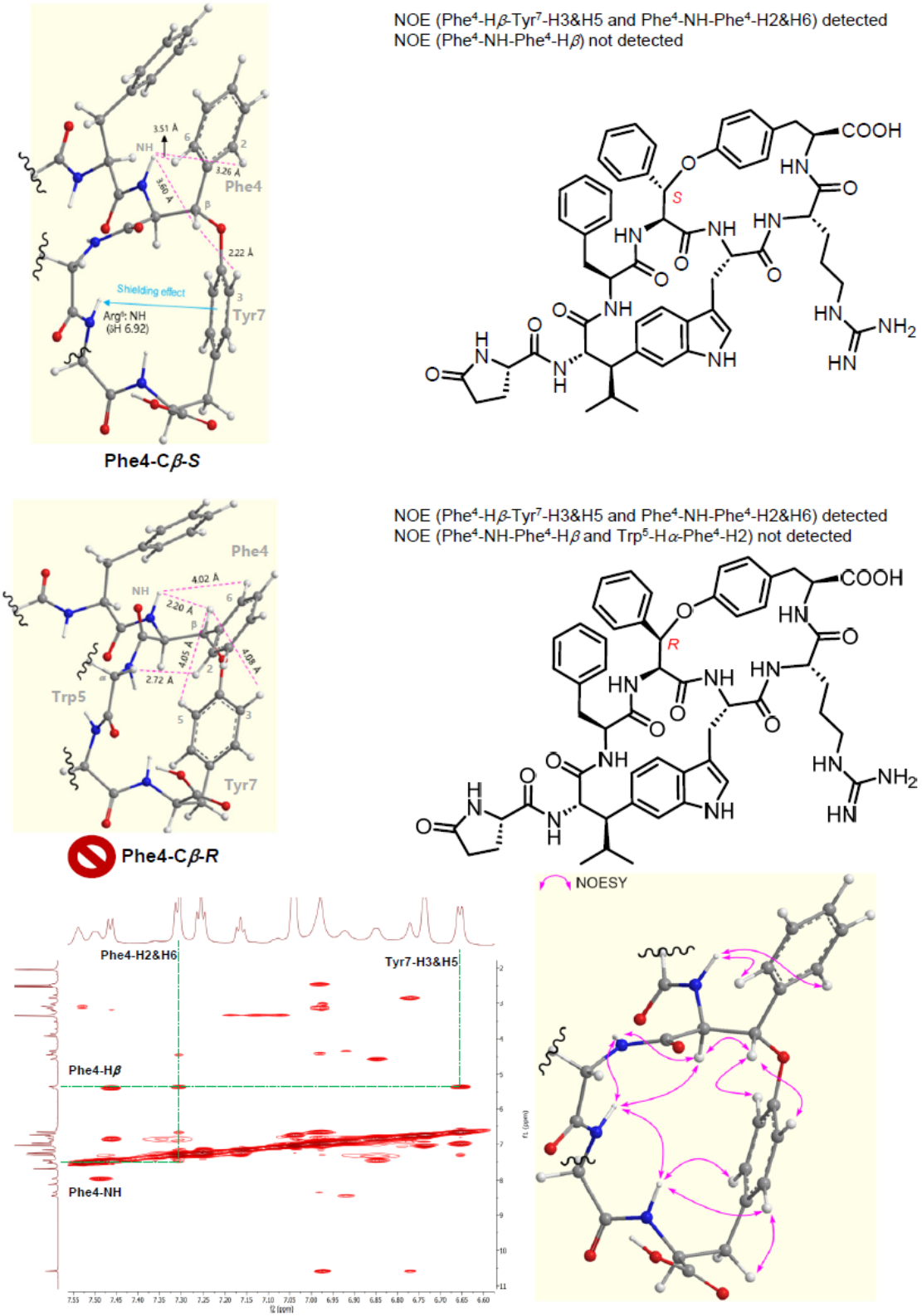


Figure 4.40 Mercurialin geometry optimization and NOESY analysis of Phe4-C β assignment.

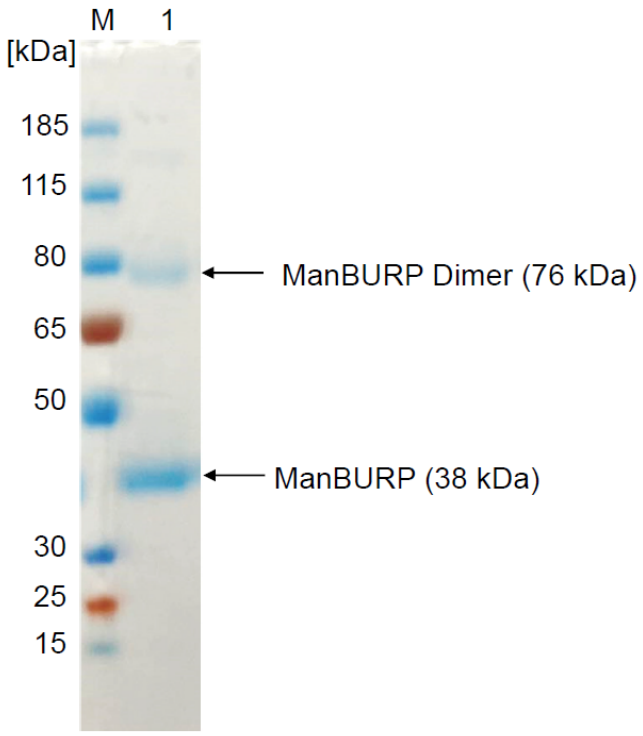


Figure 4.41 Mercurialin cyclase purification.

Non-reducing SDS-PAGE of ManBURP-1xQLFFWRY after denature-refolding purification. Source data are provided as a Source Data file.

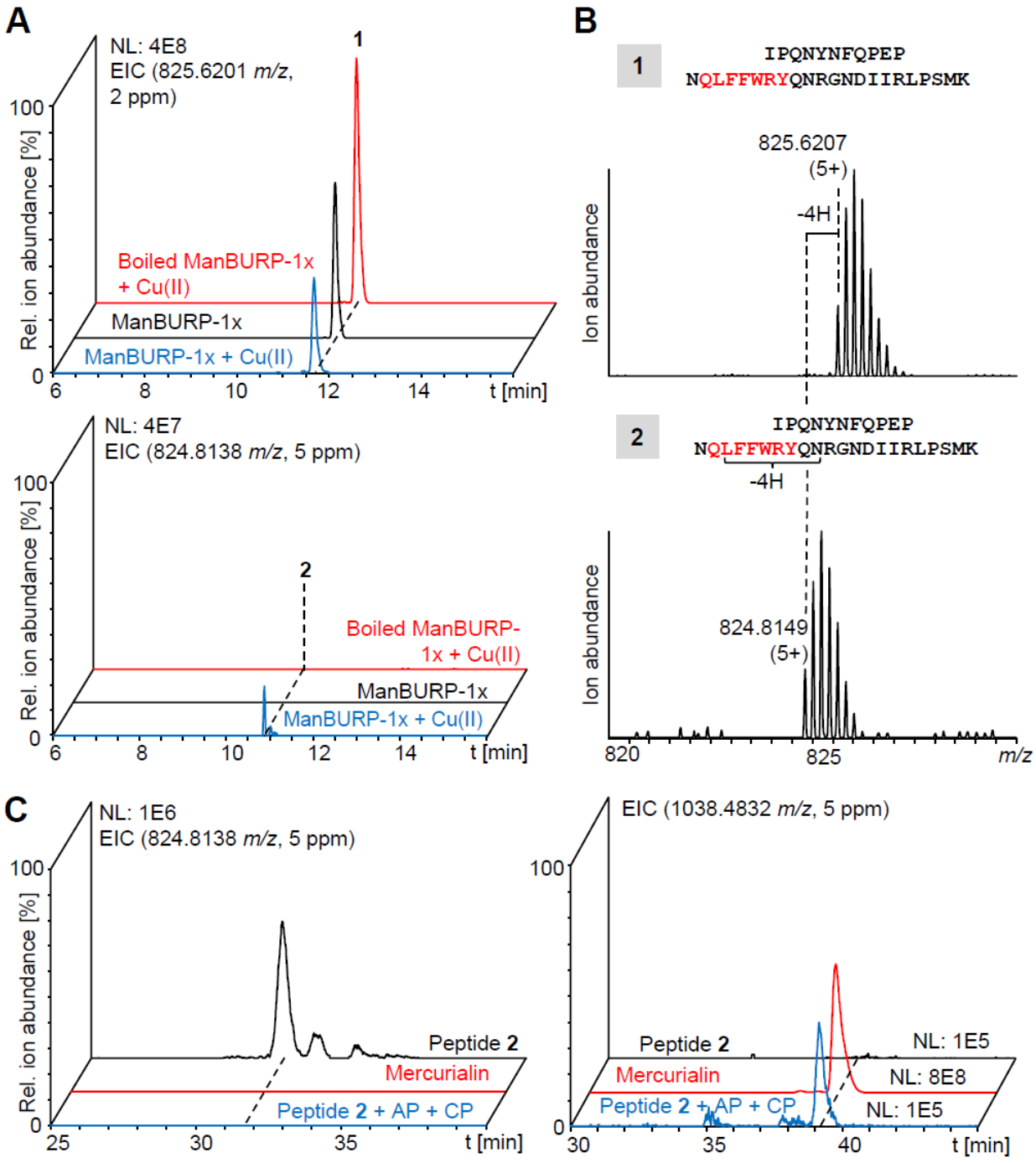


Figure 4.42 LC-MS analysis of *in vitro* reconstitution of mercurialin cyclase.
Detection of bicyclic core peptide modifications in purified ManBURP-1xQLFFWRY after incubation with Cu(II) at pH 7. (**A**) EIC traces of linear and modified LysC-digested core peptide IPQNYNFQPEPNQLFFWRYQNRGNDIIRLPSMK. (**B**) MS spectra of linear and modified LysC-digested core peptide. Core peptide are highlighted in red. (**C**) Detection of mercuralin after incubation of modified bicyclic LysC-digested core peptide (peptide **2**) with aminopeptidase (AP) and carboxypeptidase (CP).

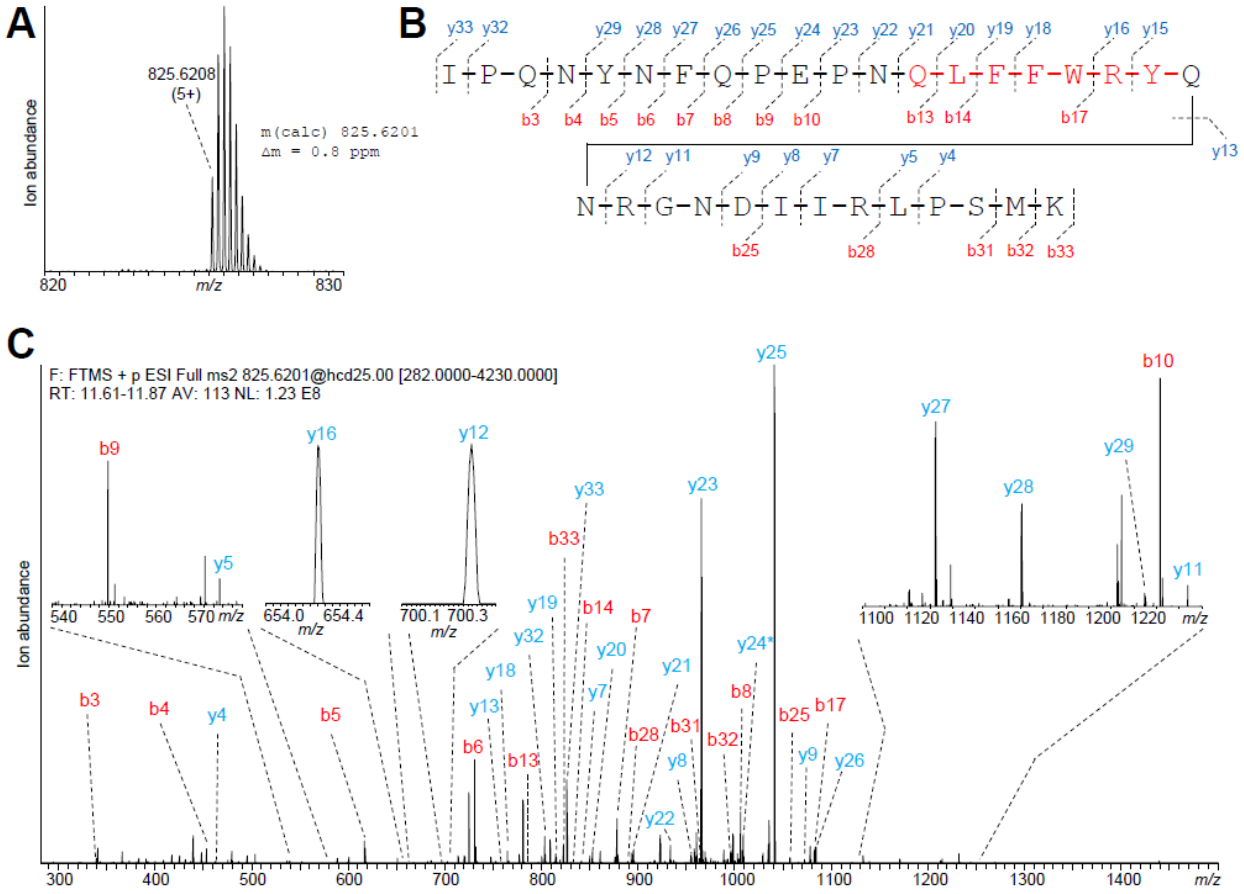


Figure 4.43 MS and MS/MS analysis of unmodified core peptide from *in vitro* reconstitution of mercurialin cyclase.

(A) MS1 signal of linear LysC-proteolytic core peptide. (B) Detected tandem MS fragments of linear LysC-proteolytic core peptide. (C) MS/MS analysis of LysC-proteolytic core peptide.

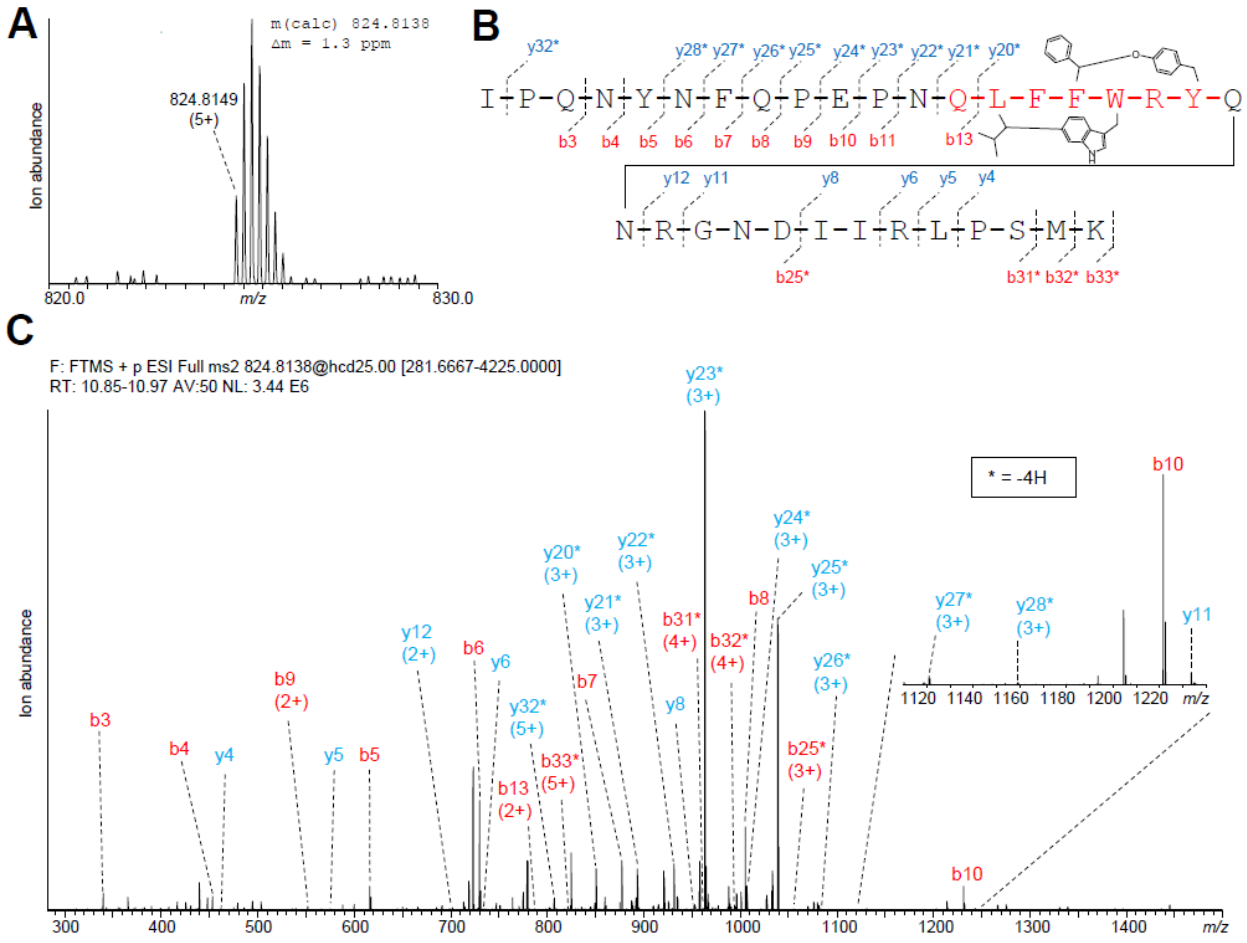


Figure 4.44 MS and MS/MS analysis of modified core peptide from *in vitro* reconstitution of mercurialin cyclase.

(A) MS1 signal of putative bicyclic LysC-proteolytic core peptide. (**B**) Detected tandem MS fragments of putative bicyclic LysC-proteolytic core peptide. (**C**) MS/MS analysis of putative bicyclic LysCproteolytic core peptide.

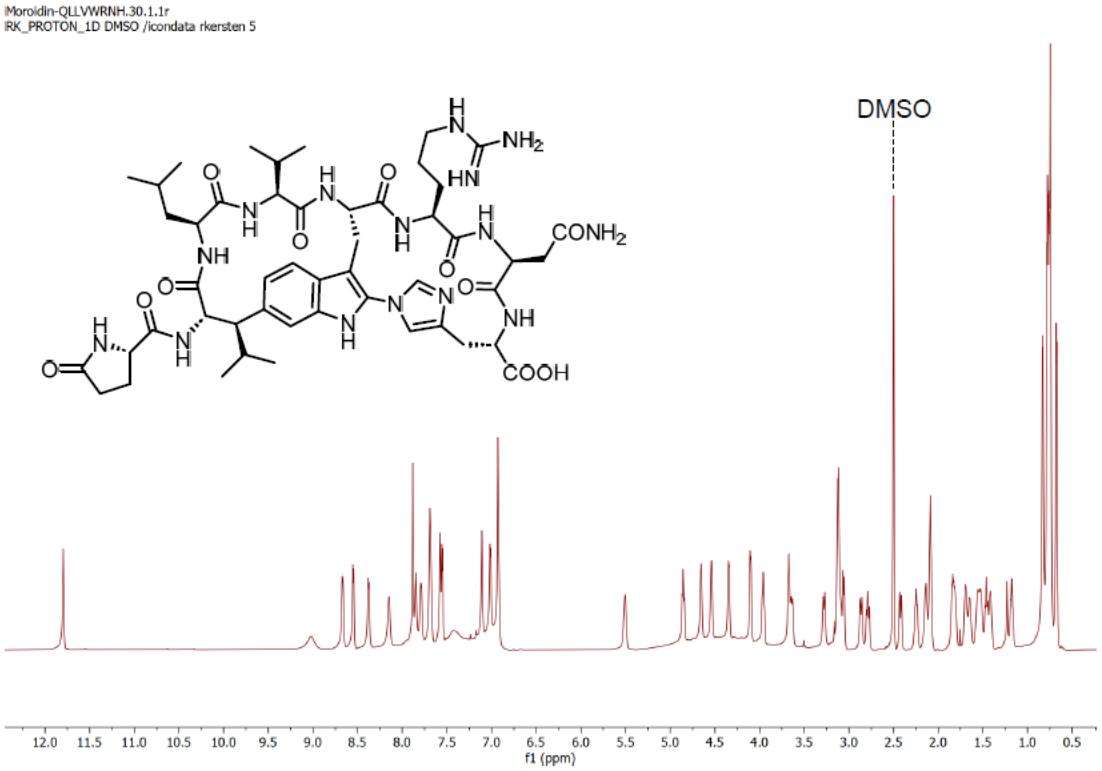


Figure 4.45 ^{1}H-NMR spectrum of moroidin-QLLVWRNH in DMSO-d$_6$ (800 MHz).

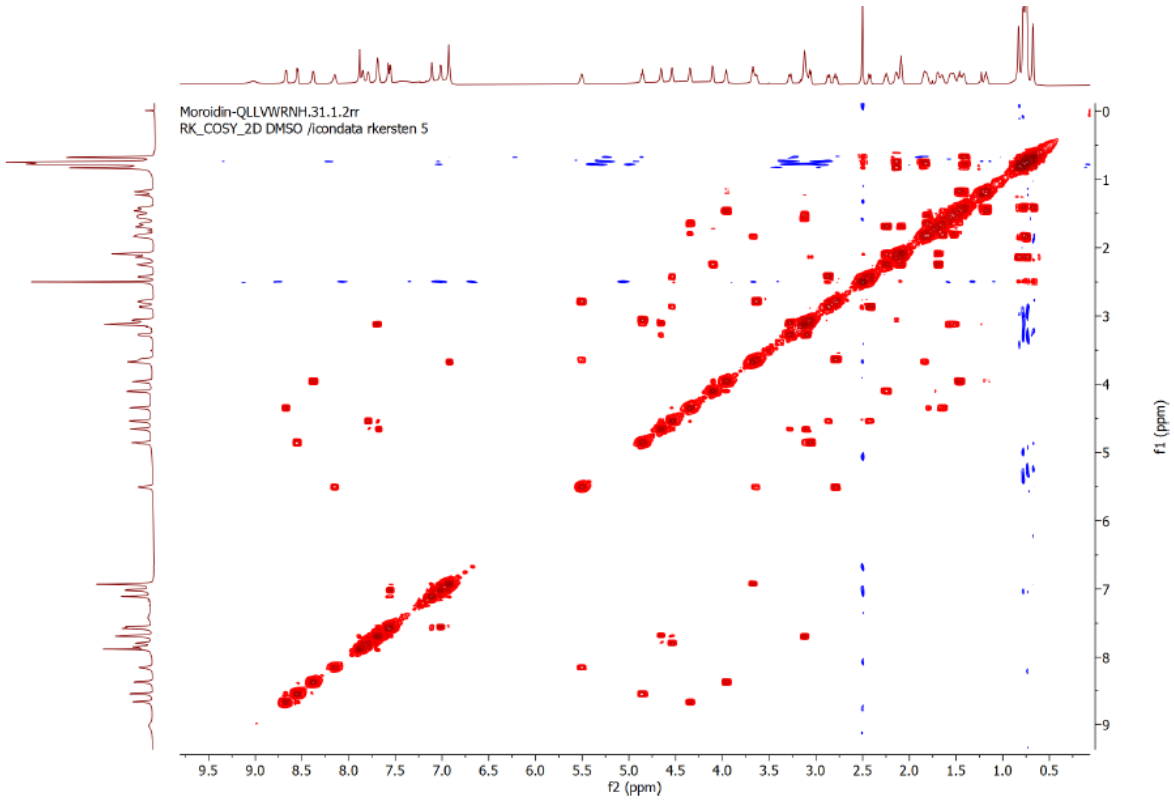


Figure 4.46 ^{1}H-^{1}H-COSY-NMR spectrum of moroidin-QLLVWRNH in DMSO-d$_6$.

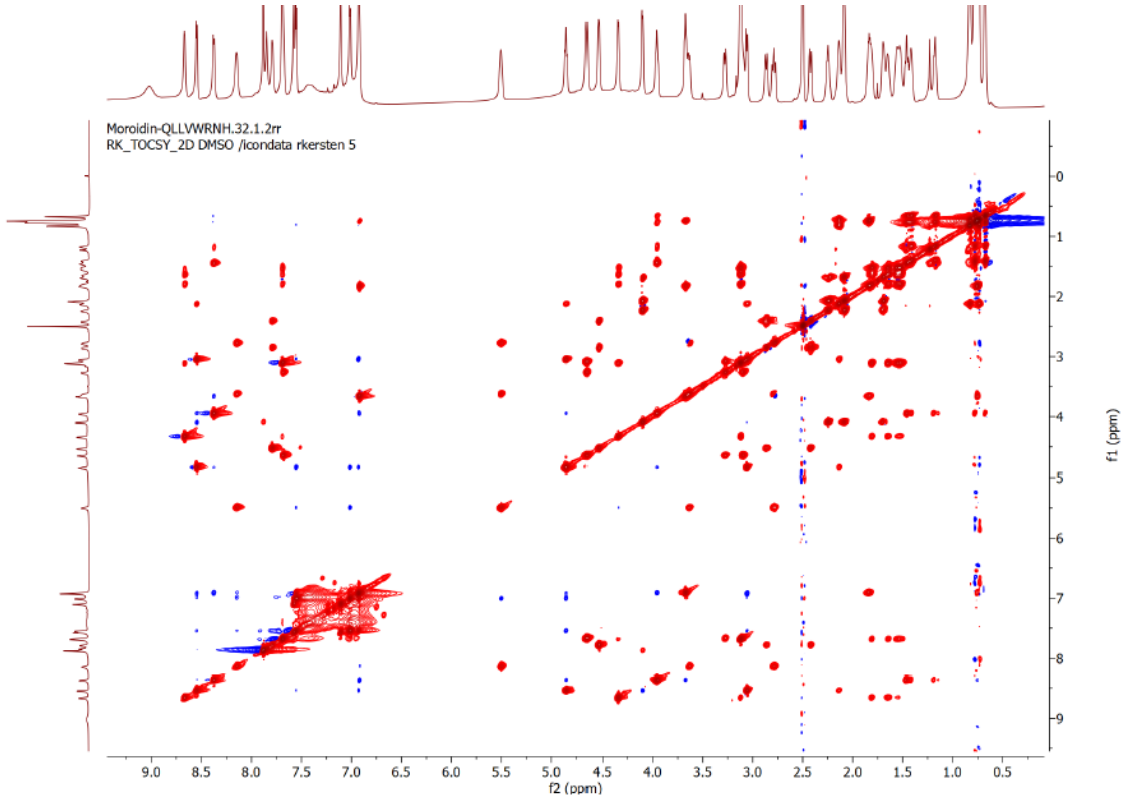


Figure 4.47 ^{1}H-^{1}H-TOCSY-NMR spectrum of moroidin-QLLVWRNH in DMSO-d$_6$.

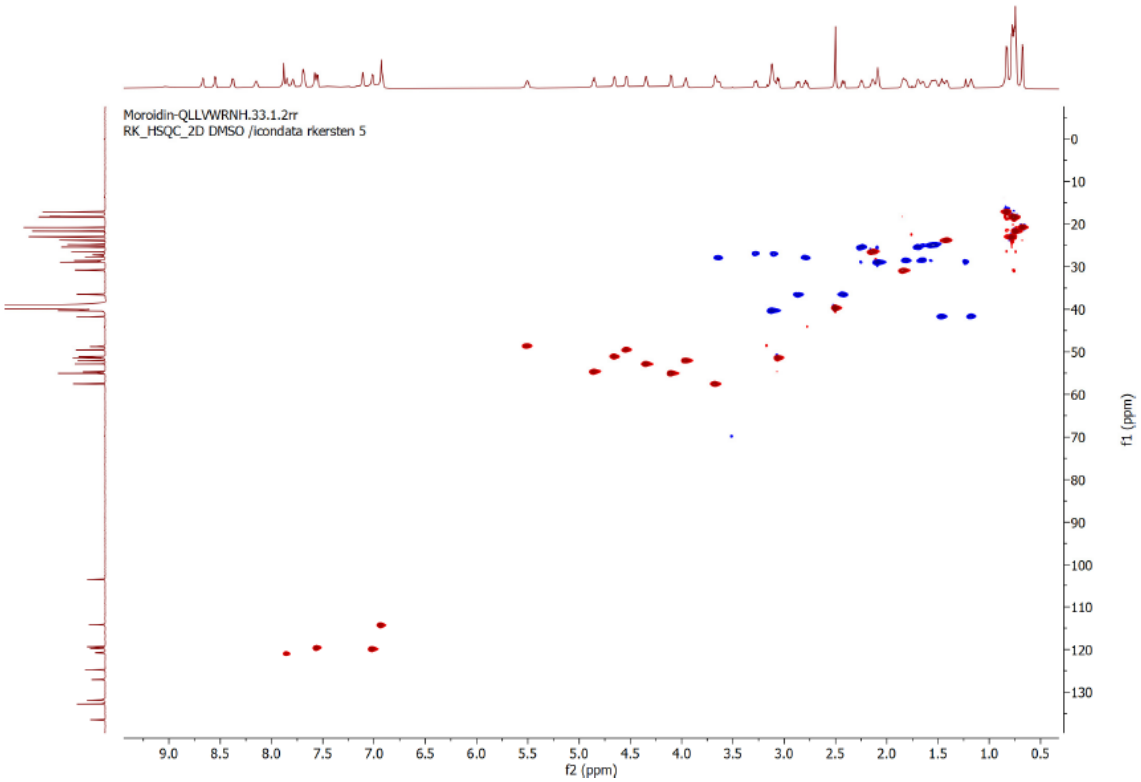


Figure 4.48 HSQC-NMR spectrum of moroidin-QLLVWRNH in DMSO-d$_6$.

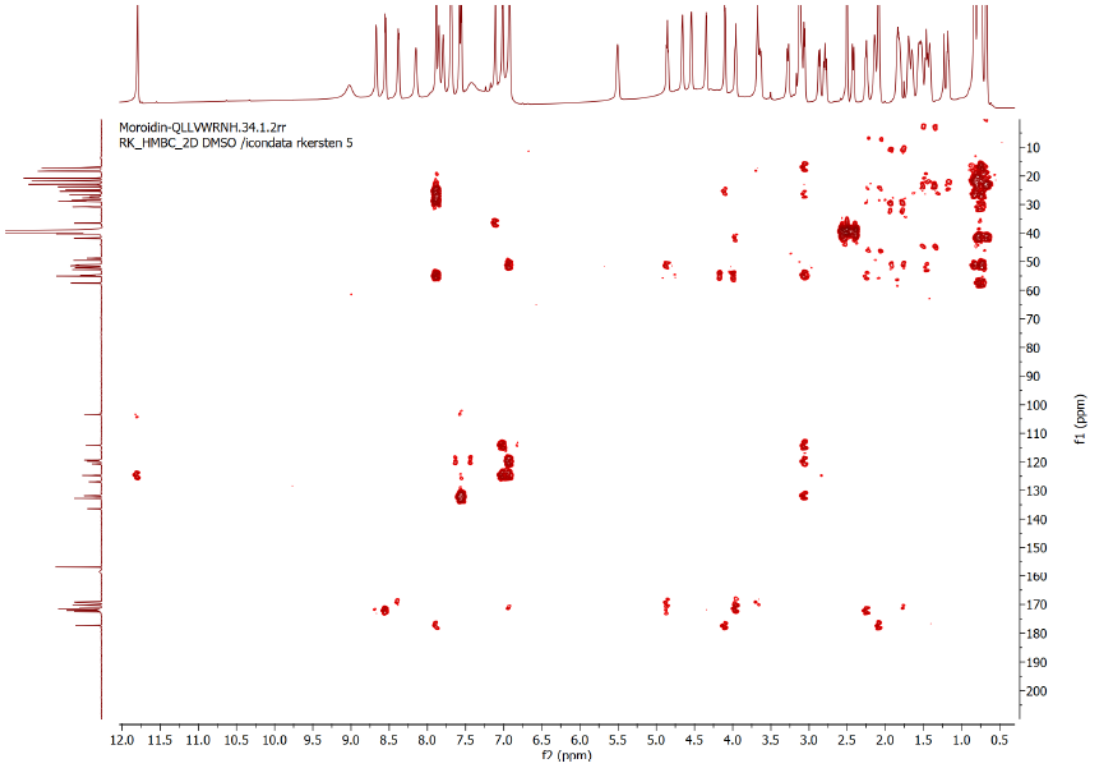


Figure 4.49 HMBC-NMR spectrum of moroidin-QLLVWRNH in DMSO-d_6.

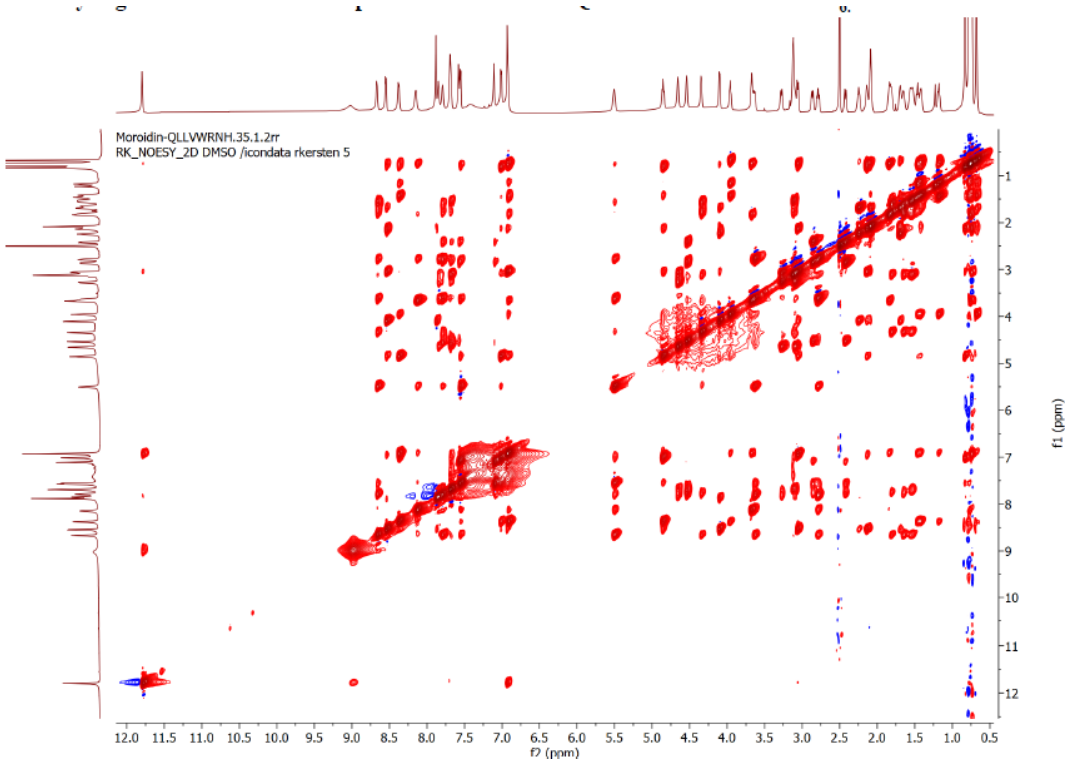


Figure 4.50 NOESY-NMR spectrum of moroidin-QLLVWRNH in DMSO-d_6.

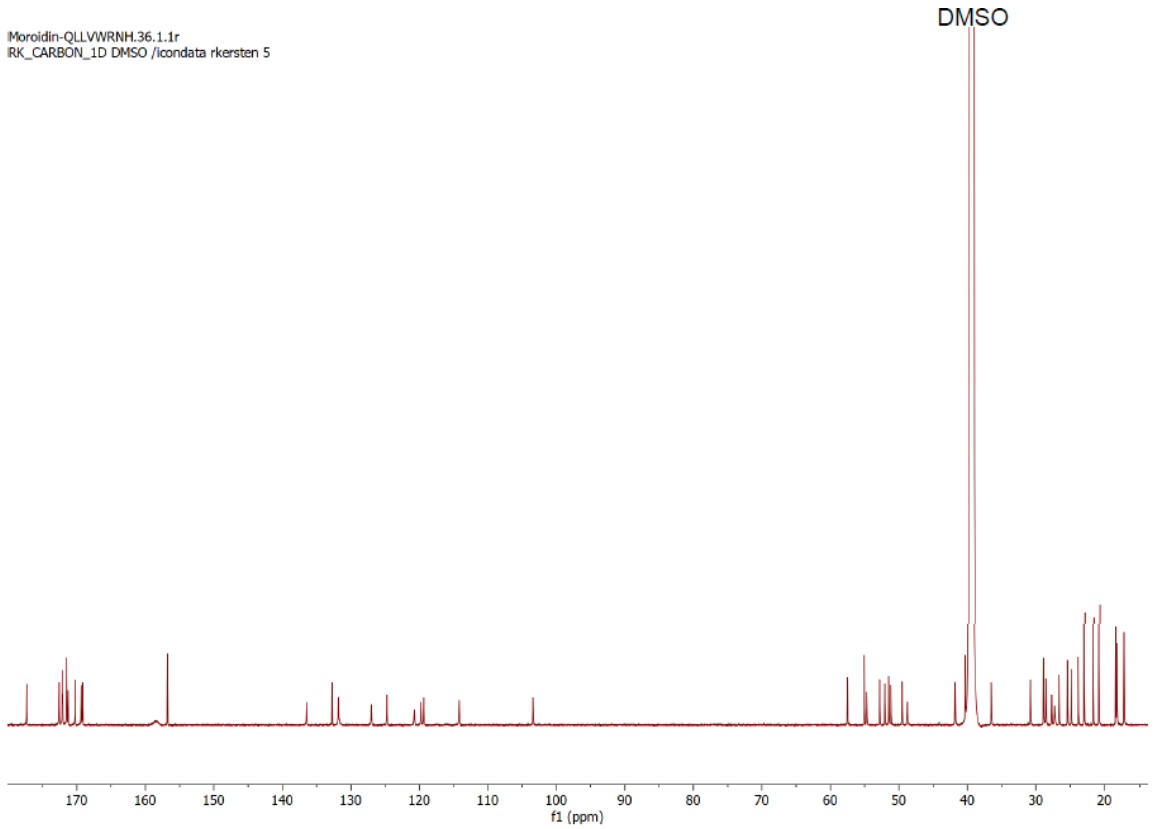


Figure 4.51 ^{13}C-NMR spectrum of moroidin-QLLVWRNH in DMSO-d$_6$.

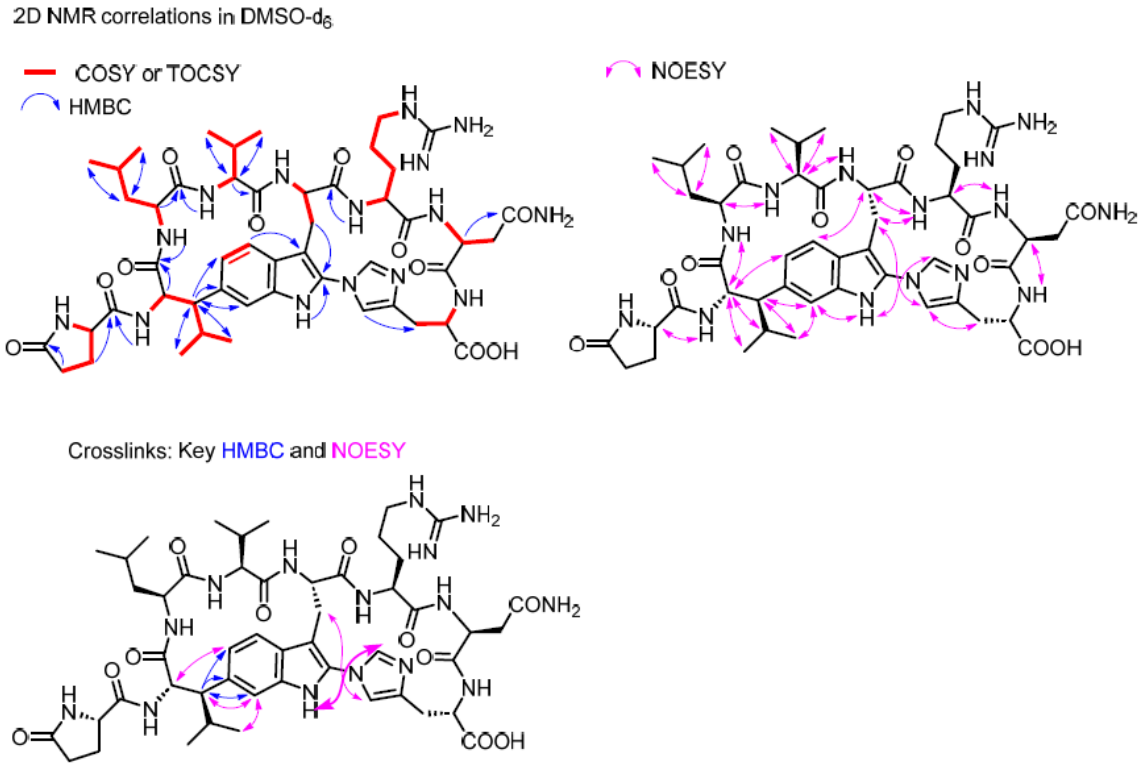


Figure 4.52 Key 2D NMR correlations of moroidin-QLLVWRNH.

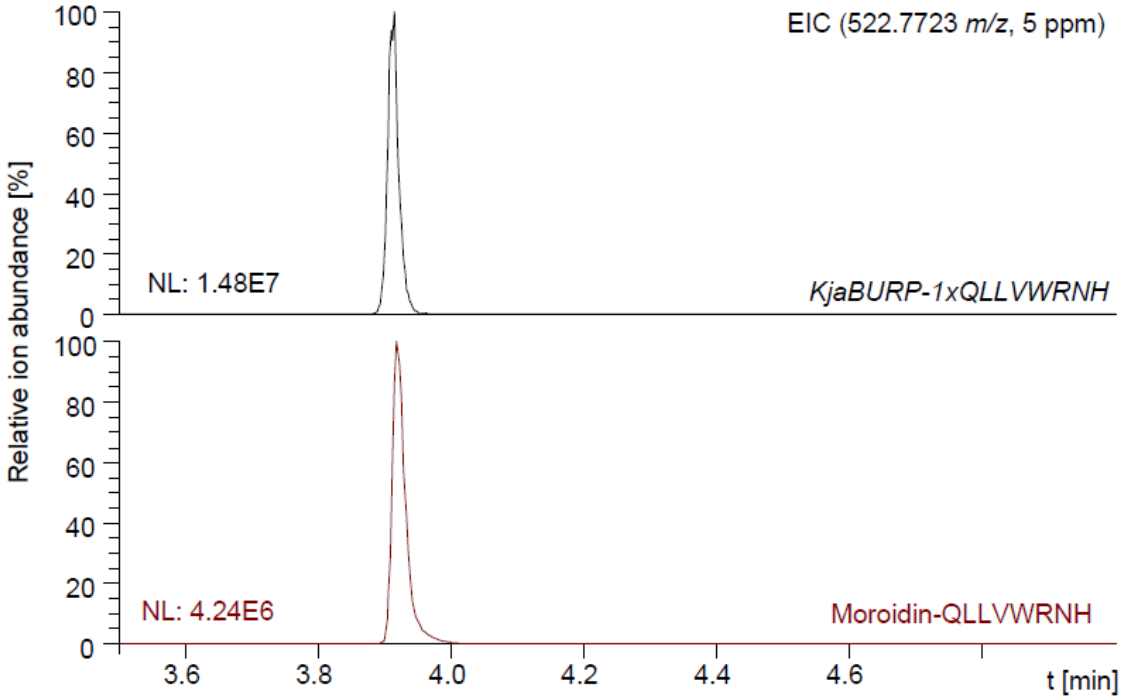


Figure 4.53 Comparison of moroidin-QLLVWRNH standard to moroidin product of KjaBURP-1xQLLVWRNH expressed for 6 days transiently in *N. benthamiana*.

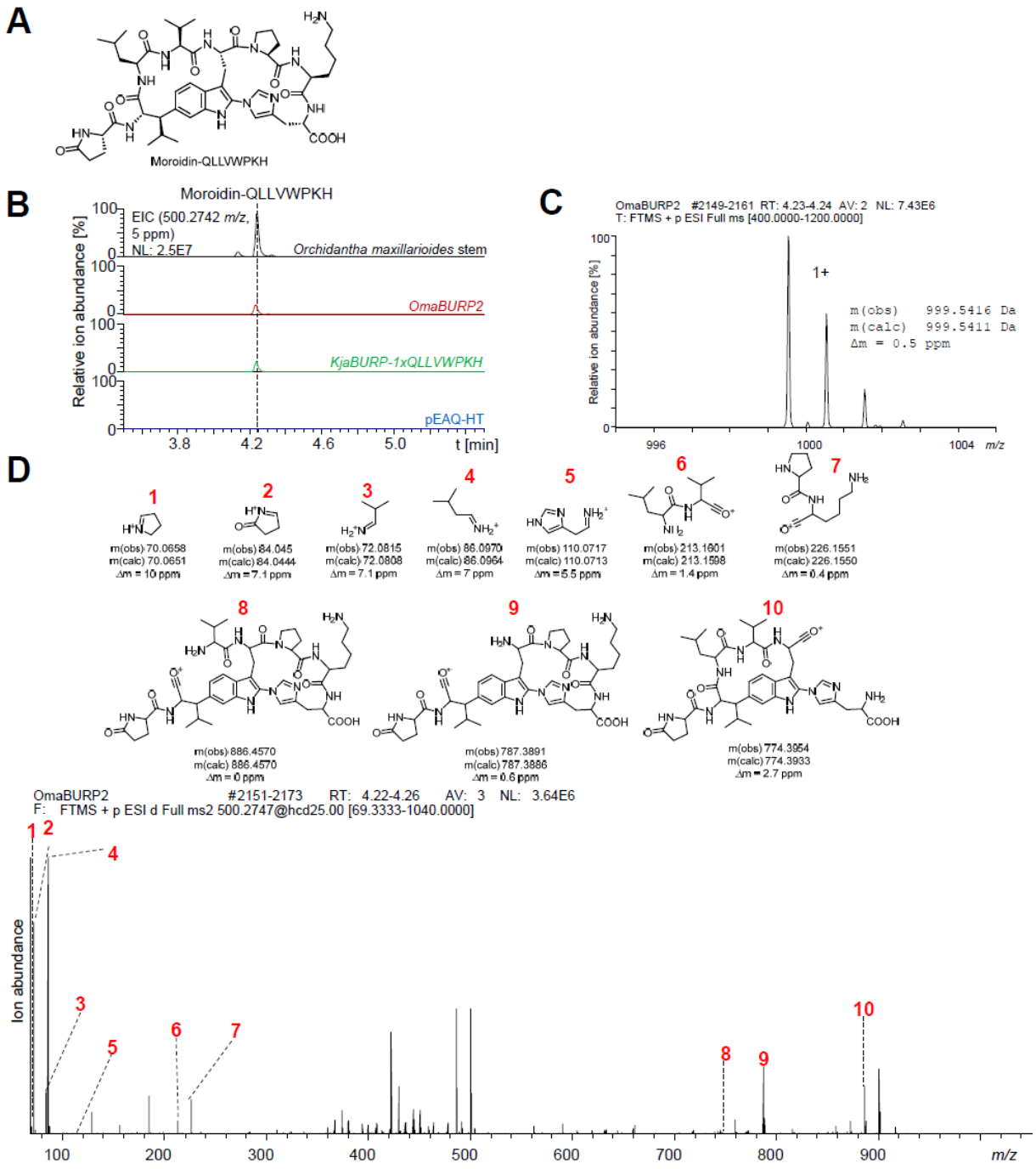


Figure 4.54 Characterization of moroidin-QLLVWPKH biosynthesis in *Orchidantha maxillarioides*.
(A) Predicted OmaBURP2 moroidin-QLLVWPKH product. (B) LC-MS characterization of predicted moroidin-QLLVWPKH in methanolic extracts of *O. maxillarioides* leaf and transgenic *N. benthamiana* after six days of expression of OmaBURP2. Moroidin-QLLVWPKH was compared to KjaBURP-1xQLLVWPKH moroidin product, derived from methanolic extract of transgenic *N. benthamiana* expressing the KjaBURP construct for 6 days. (C) MS1 spectrum and (D) MS/MS spectrum of moroidin-QLLVWPKH with predicted MS/MS fragment ions from transgenic *N. benthamiana* leaves *e*xpressing OmaBURP2.

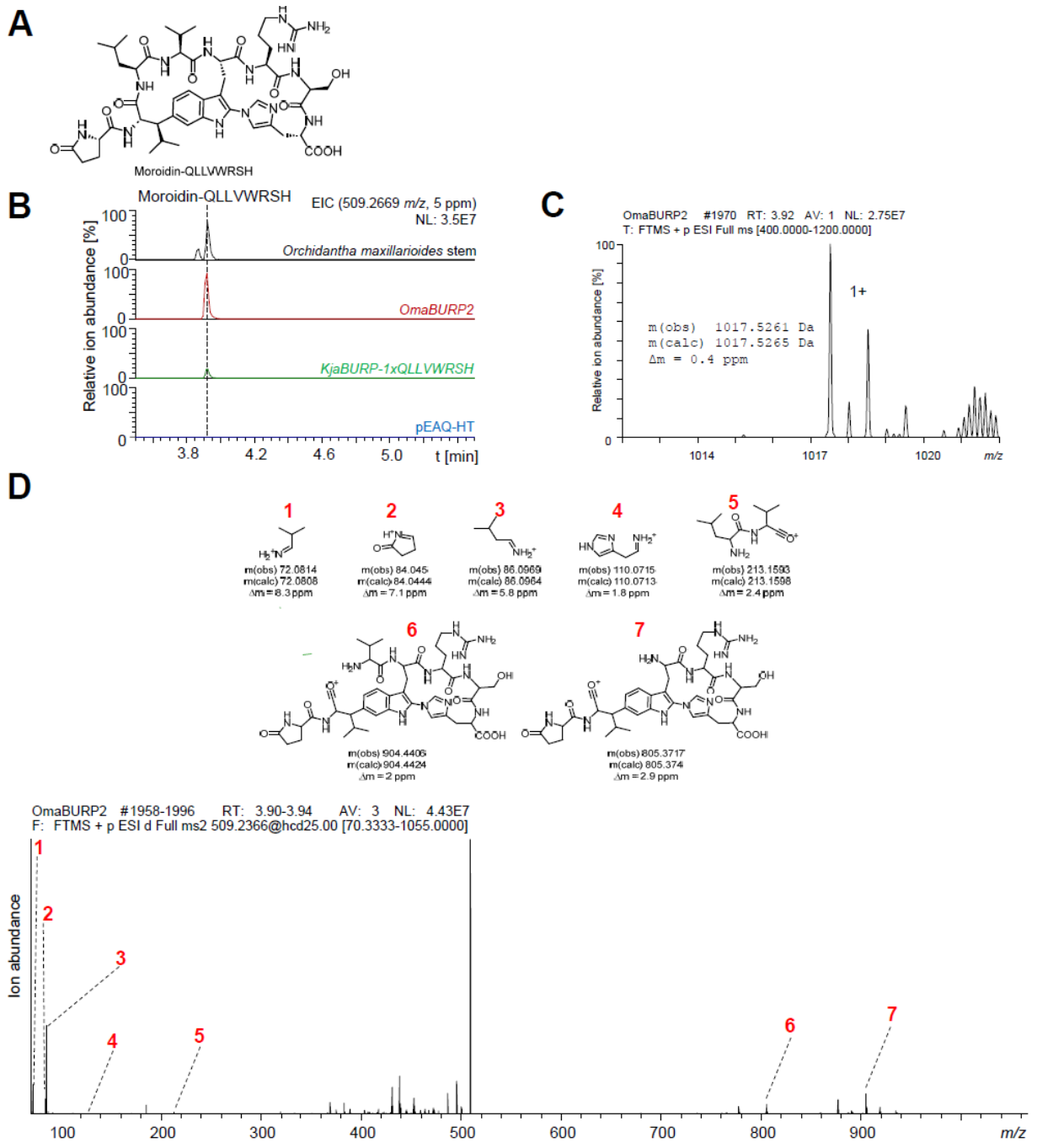


Figure 4.55 Characterization of moroidin-QLLVWRSH biosynthesis in *Orchidantha maxillarioides*.
(A) Predicted OmaBURP2 moroidin-QLLVWRSH product. (B) LC-MS characterization of predicted moroidin-QLLVWRSH in methanolic extracts of *O. maxillarioides* leaf and transgenic *N. benthamiana* after six days of expression of OmaBURP2. Moroidin-QLLVWRSH was compared to KjaBURP-1xQLLVWRSH moroidin product, derived from methanolic extract of transgenic *N. benthamiana* expressing the KjaBURP construct for 6 days. (C) MS1 spectrum and (D) MS/MS spectrum of predicted moroidin-QLLVWRSH with predicted MS/MS fragment ions from transgenic *N. benthamiana* leaves *e*xpressing OmaBURP2.

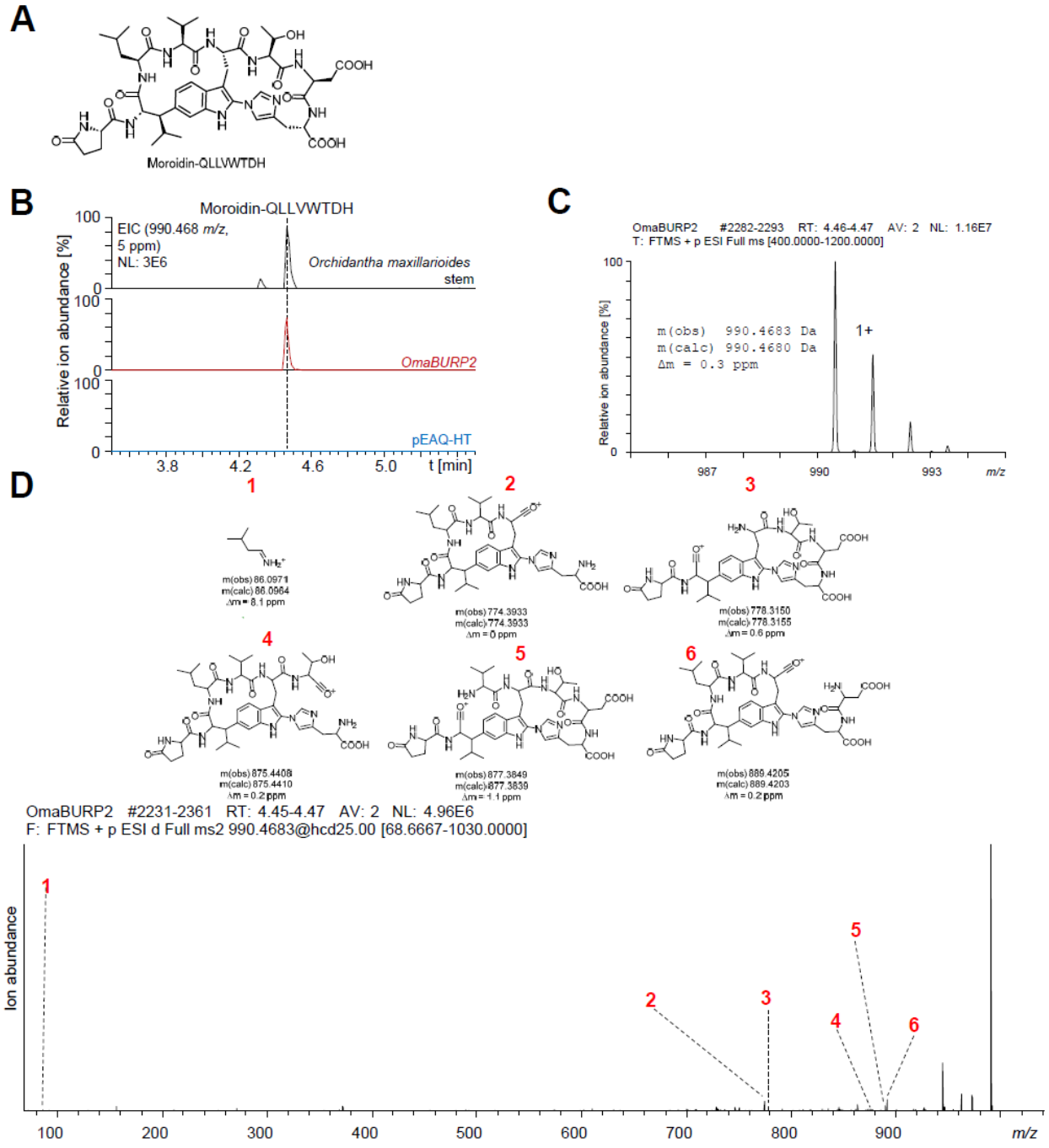


Figure 4.56 Characterization of moroidin-QLLVWTDH biosynthesis in *Orchidantha maxillarioides*
(A) Predicted OmaBURP2 moroidin-QLLVWTDH product. (B) LC-MS characterization of predicted moroidin-QLLVWTDH in methanolic extracts of *O. maxillarioides* leaf and transgenic *N. benthamiana* after six days of expression of OmaBURP2. Moroidin-QLLVWTDH was compared to KjaBURP-1xQLLVWTDH moroidin product, derived from methanolic extract of transgenic *N. benthamiana* expressing the KjaBURP construct for 6 days.(C) MS1 spectrum and (D) MS/MS spectrum of moroidin- QLLVWTDH with predicted MS/MS fragment ions from transgenic *N. benthamiana* leaves *e*xpressing OmaBURP2.

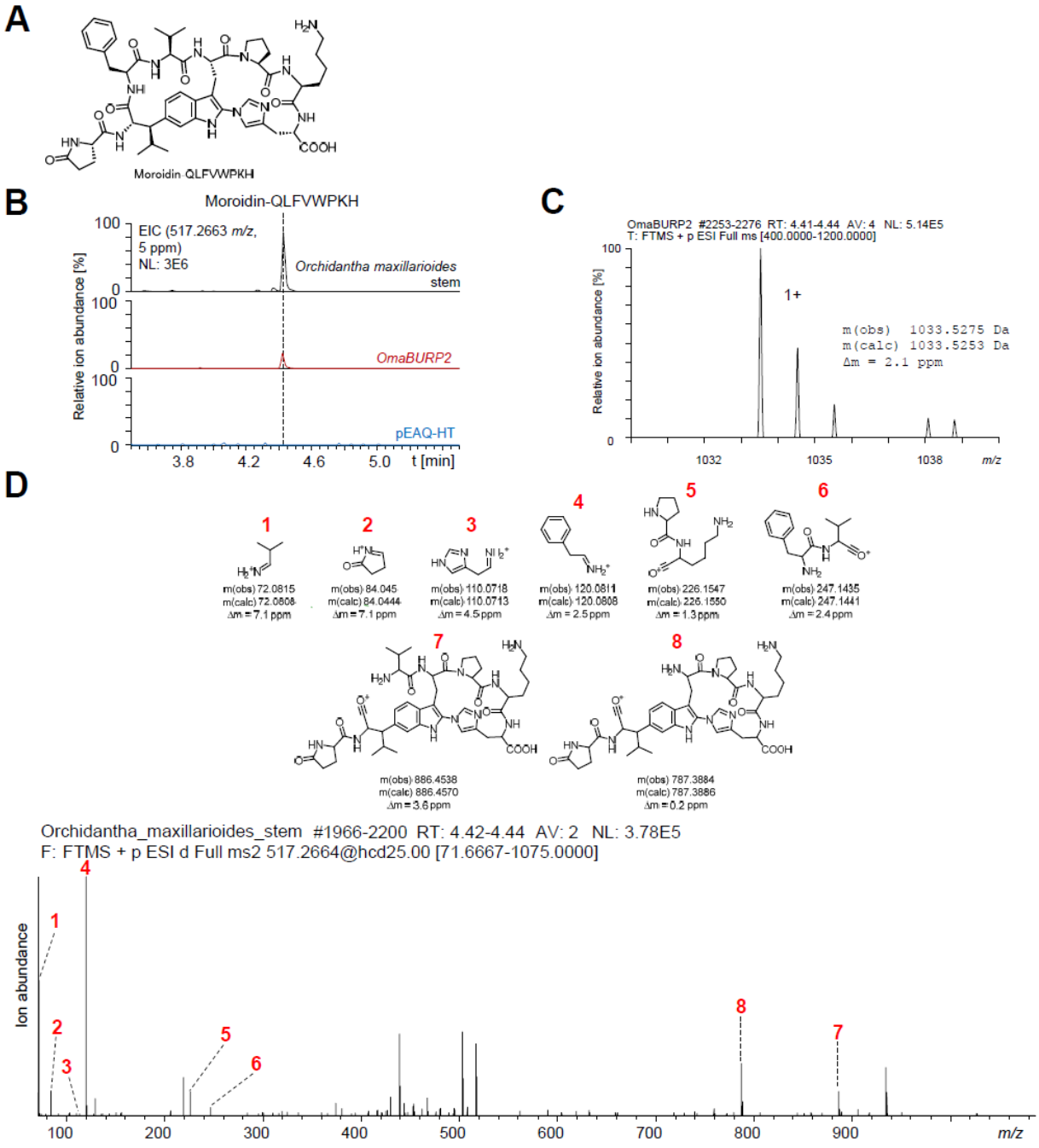


Figure 4.57 Characterization of moroidin-QLFVWPKH biosynthesis in *Orchidantha maxillarioides.*
(A) Predicted OmaBURP2 moroidin- QLFVWPKH product. (B) LC-MS characterization of predicted moroidin-QLFVWPKH in methanolic extracts of *O. maxillarioides* leaf and transgenic *N. benthamiana* after six days of expression of OmaBURP2. Moroidin- QLFVWPKH was compared to KjaBURP-1x QLFVWPKH moroidin product, derived from methanolic extract of transgenic *N. benthamiana* expressing the KjaBURP construct for 6 days.(C) MS1 spectrum and (D) MS/MS spectrum of moroidin-QLLVWPKH with predicted MS/MS fragment ions from transgenic *N. benthamiana* leaves *e*xpressing OmaBURP2 and *O. maxillarioides* stem extract, respectively.

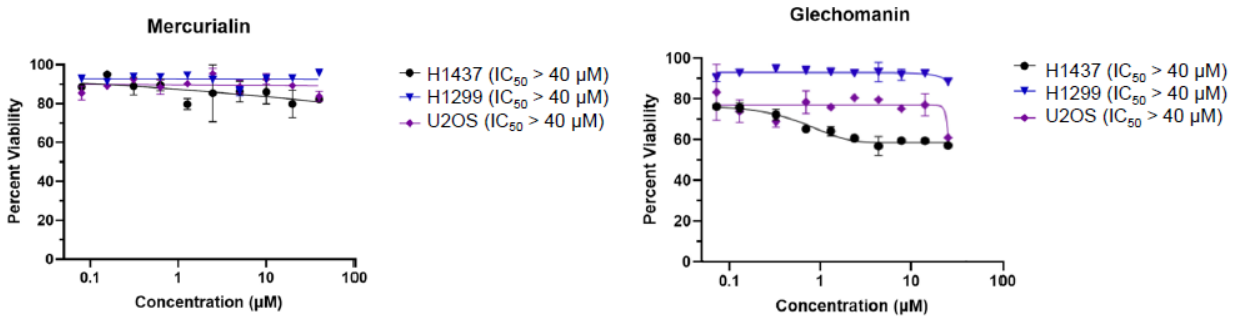


Figure 4.58 ***In vitro*** **anticancer assays of stephanotic acid-type peptides.**
The dose-response cytotoxicity of stephanotic acid-type peptides mercurialin and glechomanin was evaluated in three distinct cancer cell lines: H1437 (non-small cell lung adenocarcinoma), H1299 (non-small cell lung cancer), and U2OS (osteosarcoma). For all assays, the normalized cell viability was assessed across ten-point, two-fold serial dilutions, starting from a top concentration of 40 μM, in triplicate. Data points represent mean cell viability (n=3), with error bars denoting one standard error of the mean.

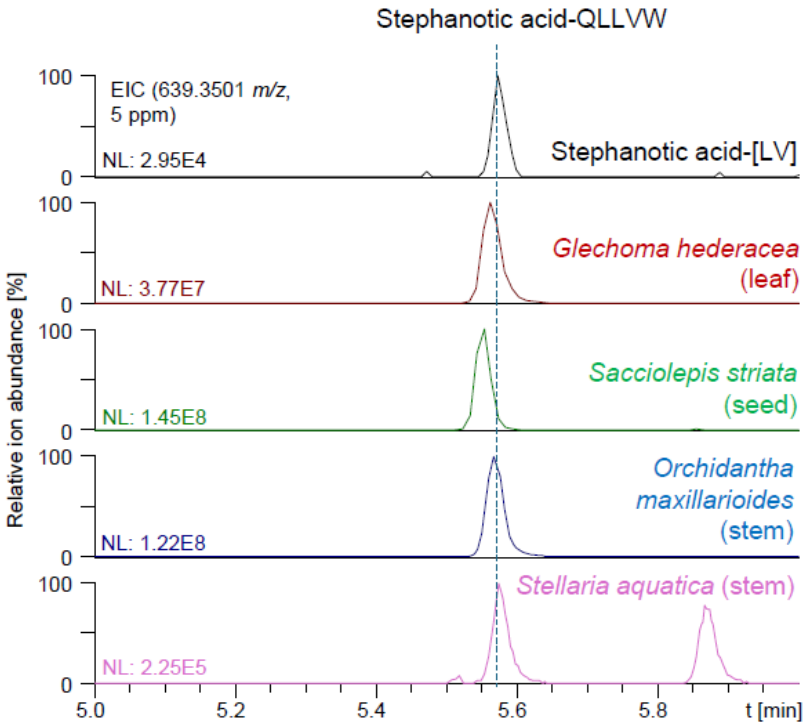


Figure 4.59 Characterization of stephanotic acid-QLLVW in pinks, mints, grasses, and Lowiaceae.

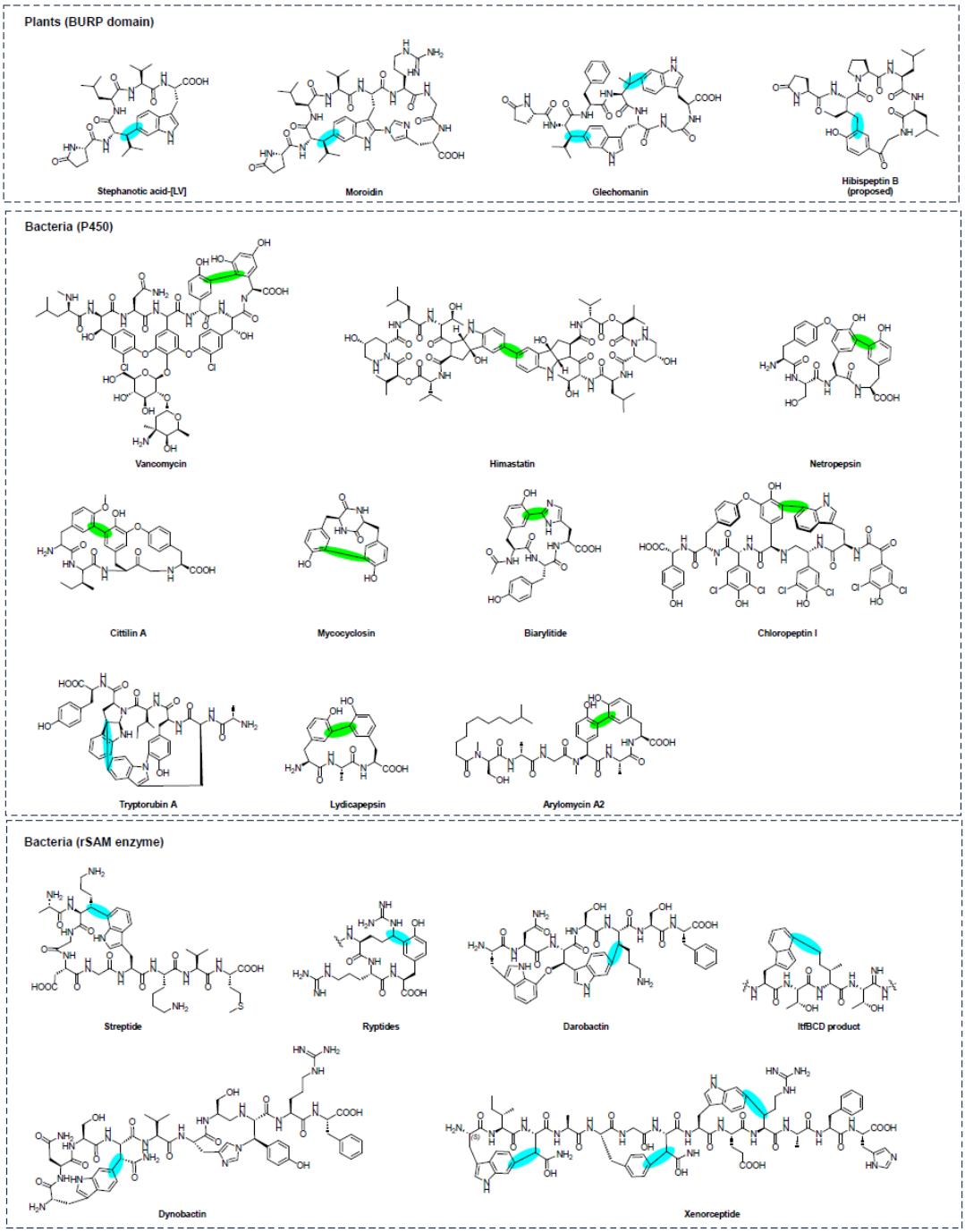


Figure 4.60 Representative peptide natural products with C-C-macrocyclic crosslinks.
Macrocyclic bonds are highlighted in blue (C(*sp3*)-C(*sp2*)) and green (C(*sp2*)-C(*sp2*)), respectively.

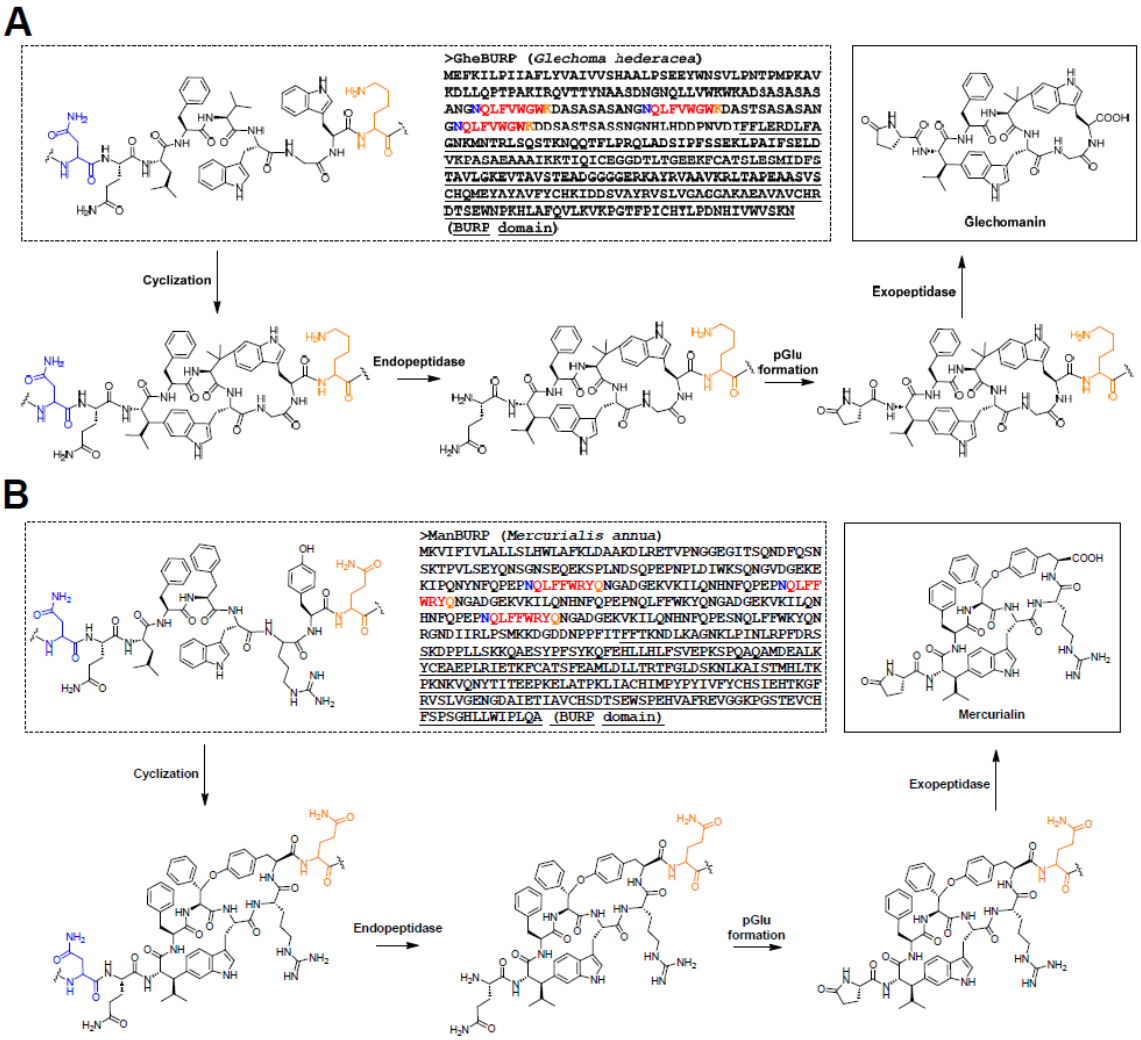


Figure 4.61 Biosynthetic proposals of burpitides from this study.
(**A**) Glechomanin biosynthesis in *Glechoma hederacea* and (**B**) Mercurialin biosynthetic proposal in *Mercurialis annua*.

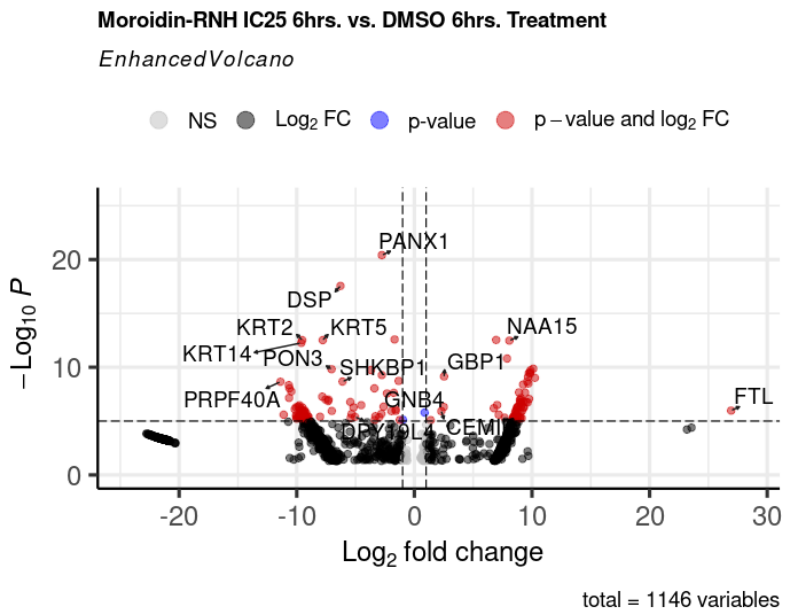


Figure 4.62 Volcano plot of top filtered and annotated DGE results from Moroidin-RNH vs. DMSO treatment for 6hrs at concentration IC_{25}.

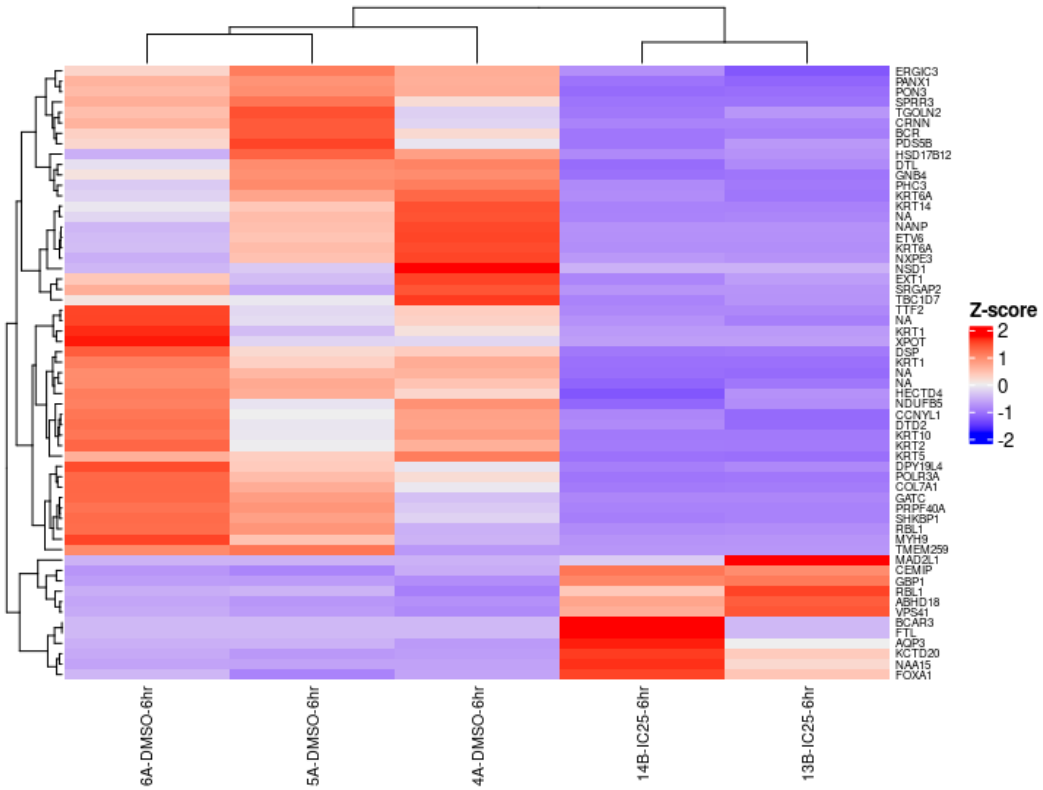


Figure 4.63 Heatmap of top filtered and annotated DGE results from Moroidin-RNH vs. DMSO treatment for 6hrs at concentration IC_{25}.

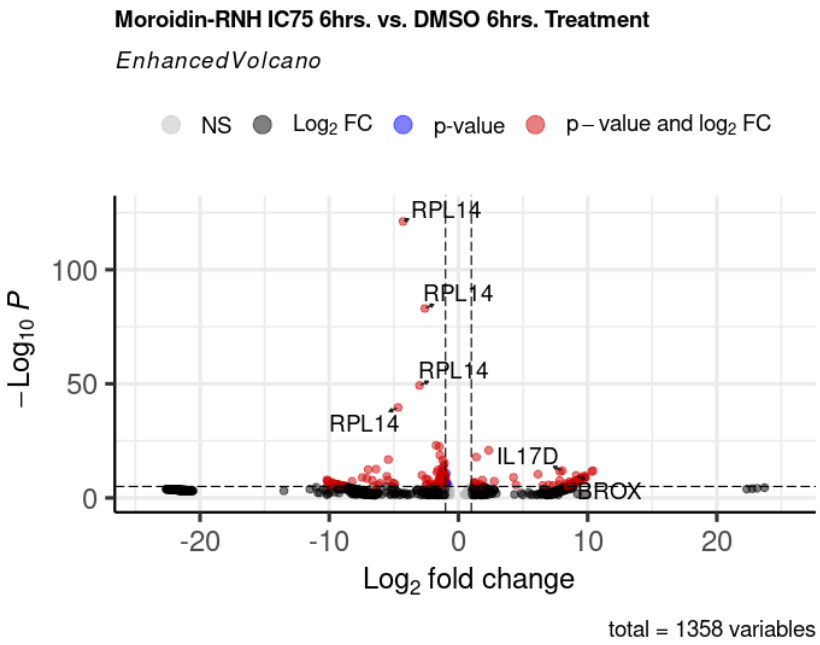


Figure 4.64 Volcano plot of top filtered and annotated DGE results from Moroidin-RNH vs. DMSO treatment for 6hrs at concentration IC_{75}.

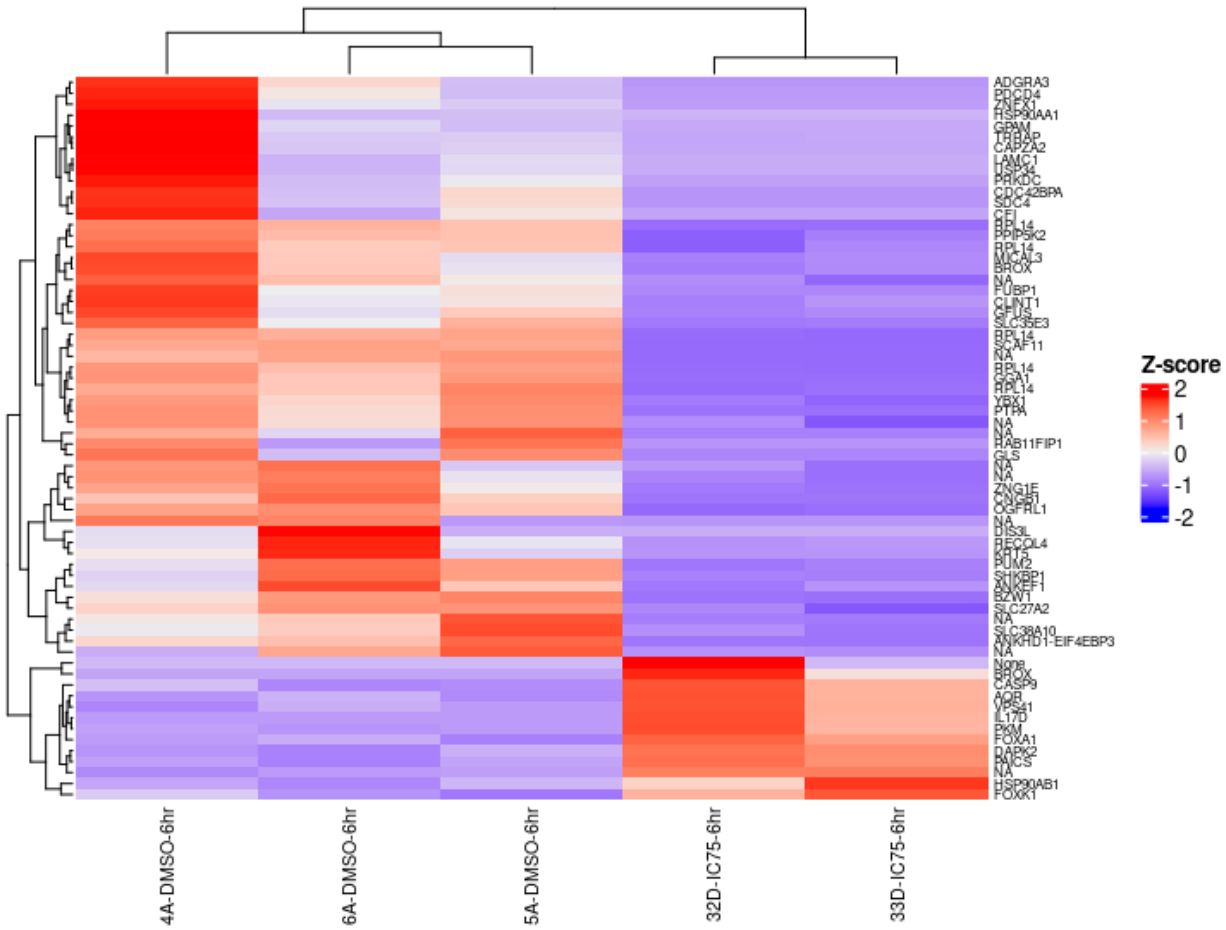


Figure 4.65 Heatmap of top filtered and annotated DGE results from Moroidin-RNH vs. DMSO treatment for 6hrs at concentration IC_{75}.

4.8 Reference

1. Chekan, J. R., Mydy, L. S., Pasquale, M. A. & Kersten, R. D. Plant peptides – redefining an area of ribosomally synthesized and post-translationally modified peptides. *Nat. Prod. Rep.* **41**, 1020–1059 (2024).
2. Kessler, S. C. & Chooi, Y.-H. Out for a RiPP: challenges and advances in genome mining of ribosomal peptides from fungi. *Nat. Prod. Rep.* **39**, 222–230 (2022).
3. Tan, N.-H. & Zhou, J. Plant Cyclopeptides. *Chem. Rev.* **106**, 840–895 (2006).
4. Kersten, R. D. *et al.* Gene-Guided Discovery and Ribosomal Biosynthesis of Moroidin Peptides. *J. Am. Chem. Soc.* **144**, 7686–7692 (2022).
5. Kersten, R. D. & Weng, J.-K. Gene-guided discovery and engineering of branched cyclic peptides in plants. *Proc. Natl. Acad. Sci.* **115**, (2018).
6. Lima, S. T. *et al.* A Widely Distributed Biosynthetic Cassette Is Responsible for Diverse Plant Side Chain Cross-Linked Cyclopeptides**. *Angew. Chem. Int. Ed.* **62**, e202218082 (2023).

7. Chigumba, D. N. *et al.* Discovery and biosynthesis of cyclic plant peptides via autocatalytic cyclases. *Nat. Chem. Biol.* **18**, 18–28 (2022).
8. Hattori, J., Boutilier, K. A., Campagne, M. M. L. & Miki, B. L. A conserved BURP domain defines a novel group of plant proteins with unusual primary structures. *Mol. Gen. Genet. MGG* **259**, 424–428 (1998).
9. Treacy, B. K. *et al.* Bnm1, a Brassica pollen-specific gene. *Plant Mol. Biol.* **34**, 603–611 (1997).
10. Bassner, R. *et al.* Abundant embryonic mRNA in field bean (Vicia faba L.) codes for a new class of seed proteins: cDNA cloning and characterization of the primary translation product. *Plant Mol. Biol.* **11**, 321–334 (1988).
11. Yamaguchi-Shinozaki, K. & Shinozaki, K. The plant hormone abscisic acid mediates the drought-induced expression but not the seed-specific expression of rd22, a gene responsive to dehydration stress in Arabidopsis thaliana. *Mol. Gen. Genet. MGG* **238–238**, 17–25 (1993).
12. Zheng, L., Heupel, R. C. & DellaPenna, D. The beta subunit of tomato fruit polygalacturonase isoenzyme 1: isolation, characterization, and identification of unique structural features. *Plant Cell* **4**, 1147–1156 (1992).
13. Christina Leung, T.-W., Williams, D. H., C J Barna, J., Foti, S. & B.Oelrichs, P. Structural studies on the peptide moroidin from laportea moroides. *Tetrahedron* **42**, 3333–3348 (1986).
14. Yoshikawa, K., Tao, S. & Arihara, S. Stephanotic Acid, a Novel Cyclic Pentapeptide from the Stem of *Stephanotis f loribunda*. *J. Nat. Prod.* **63**, 540–542 (2000).
15. Morita, H., Shimbo, K., Shigemori, H. & Kobayashi, J. Antimitotic activity of moroidin, a bicyclic peptide from the seeds of Celosia argentea. *Bioorg. Med. Chem. Lett.* **10**, 469–471 (2000).
16. Kobayashi, J., Suzuki, H., Shimbo, K., Takeya, K. & Morita, H. Celogentins A−C, New Antimitotic Bicyclic Peptides from the Seeds of *Celosia a rgentea*. *J. Org. Chem.* **66**, 6626–6633 (2001).
17. Sainsbury, F., Thuenemann, E. C. & Lomonossoff, G. P. pEAQ: versatile expression vectors for easy and quick transient expression of heterologous proteins in plants. *Plant Biotechnol. J.* **7**, 682–693 (2009).
18. Mydy, L. S. *et al.* An intramolecular macrocyclase in plant ribosomal peptide biosynthesis. *Nat. Chem. Biol.* **20**, 530–540 (2024).
19. Christenhusz, M. J. M. *et al.* The genome sequence of the Annual Mercury, Mercurialis annua L., 1753 (Euphorbiaceae). *Wellcome Open Res.* **9**, 102 (2024).
20. Negin, B. & Jander, G. Convergent and divergent evolution of plant chemical defenses. *Curr. Opin. Plant Biol.* **73**, 102368 (2023).
21. One Thousand Plant Transcriptomes Initiative. One thousand plant transcriptomes and the phylogenomics of green plants. *Nature* **574**, 679–685 (2019).
22. American Type Culture Collection (ATCC). H1437 [Cell line].
23. Vaughan, C. A. *et al.* Allele specific gain-of-function activity of p53 mutants in lung cancer cells. *Biochem. Biophys. Res. Commun.* **428**, 6–10 (2012).
24. Chougoni, K. K. *et al.* Oncogenic Mutant *p53* Sensitizes Non–Small Cell Lung Cancer Cells to Proteasome Inhibition via Oxidative Stress–Dependent Induction of Mitochondrial Apoptosis. *Cancer Res. Commun.* **4**, 2685–2698 (2024).

25. Gannon, H. S. *et al.* Identification of an "Exceptional Responder" Cell Line to MEK1 Inhibition: Clinical Implications for MEK-Targeted Therapy. *Mol. Cancer Res.* **14**, 207–215 (2016).
26. Tanizaki, J. *et al.* MET Tyrosine Kinase Inhibitor Crizotinib (PF-02341066) Shows Differential Antitumor Effects in Non-small Cell Lung Cancer According to MET Alterations. *J. Thorac. Oncol.* **6**, 1624–1631 (2011).
27. Whithaus, K., Fukuoka, J., Prihoda, T. J. & Jagirdar, J. Evaluation of Napsin A, Cytokeratin 5/6, p63, and Thyroid Transcription Factor 1 in Adenocarcinoma Versus Squamous Cell Carcinoma of the Lung. *Arch. Pathol. Lab. Med.* **136**, 155–162 (2012).
28. Differential gene expression workshop.
29. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15**, 550 (2014).
30. Gu, Z., Eils, R. & Schlesner, M. Complex heatmaps reveal patterns and correlations in multidimensional genomic data. *Bioinformatics* **32**, 2847–2849 (2016).
31. Blighe, K., Rana, S. & Lewis, M. EnhancedVolcano: Publication-ready volcano plots with enhanced colouring and labeling. (2018).
32. Shen, W., Sipos, B. & Zhao, L. SeqKit2: A Swiss army knife for sequence and alignment processing. *iMeta* **3**, e191 (2024).
33. Hart, A. J. *et al.* EnTAP : Bringing faster and smarter functional annotation to non-model eukaryotic transcriptomes. *Mol. Ecol. Resour.* **20**, 591–604 (2020).
34. Selinger, D. W. *et al.* A Framework for Autonomous AI-Driven Drug Discovery. Preprint at https://doi.org/10.1101/2024.12.17.629024 (2024).
35. Schramma, K. R., Bushin, L. B. & Seyedsayamdost, M. R. Structure and biosynthesis of a macrocyclic peptide containing an unprecedented lysine-to-tryptophan crosslink. *Nat. Chem.* **7**, 431–437 (2015).
36. Nguyen, H. *et al.* Characterization of a Radical SAM Oxygenase for the Ether Crosslinking in Darobactin Biosynthesis. *J. Am. Chem. Soc.* **144**, 18876–18886 (2022).
37. Caruso, A., Martinie, R. J., Bushin, L. B. & Seyedsayamdost, M. R. Macrocyclization via an Arginine-Tyrosine Crosslink Broadens the Reaction Scope of Radical *S* -Adenosylmethionine Enzymes. *J. Am. Chem. Soc.* **141**, 16610–16614 (2019).
38. Nguyen, T. Q. N. *et al.* Post-translational formation of strained cyclophanes in bacteria. *Nat. Chem.* **12**, 1042–1053 (2020).
39. Ma, S. *et al.* Post-Translational Formation of Aminomalonate by a Promiscuous Peptide-Modifying Radical SAM Enzyme. *Angew. Chem. Int. Ed.* **60**, 19957–19964 (2021).
40. Forneris, C. C. & Seyedsayamdost, M. R. In Vitro Reconstitution of OxyC Activity Enables Total Chemoenzymatic Syntheses of Vancomycin Aglycone Variants. *Angew. Chem. Int. Ed.* **57**, 8048–8052 (2018).
41. Zhao, B. *et al.* Binding of Two Flaviolin Substrate Molecules, Oxidative Coupling, and Crystal Structure of Streptomyces coelicolor A3(2) Cytochrome P450 158A2. *J. Biol. Chem.* **280**, 11599–11607 (2005).
42. Hug, J. J. *et al.* Biosynthesis of Cittilins, Unusual Ribosomally Synthesized and Post-translationally Modified Peptides from *Myxococcus xanthus*. *ACS Chem. Biol.* **15**, 2221–2231 (2020).
43. Belin, P. *et al.* Identification and structural basis of the reaction catalyzed by CYP121, an essential cytochrome P450 in *Mycobacterium tuberculosis*. *Proc. Natl. Acad. Sci.* **106**, 7426–7431 (2009).

44. Zdouc, M. M. *et al.* A biaryl-linked tripeptide from Planomonospora reveals a widespread class of minimal RiPP gene clusters. *Cell Chem. Biol.* **28**, 733-739.e4 (2021).
45. Nam, H. *et al.* Exploring the Diverse Landscape of Biaryl-Containing Peptides Generated by Cytochrome P450 Macrocyclases. *J. Am. Chem. Soc.* **145**, 22047–22057 (2023).
46. Hu, Y. L. *et al.* P450-Modified Ribosomally Synthesized Peptides with Aromatic Cross-Links. *J. Am. Chem. Soc.* **145**, 27325–27335 (2023).
47. He, B. *et al.* Bacterial Cytochrome P450 Catalyzed Post-translational Macrocyclization of Ribosomal Peptides**. *Angew. Chem. Int. Ed.* **62**, e202311533 (2023).
48. Nanudorn, P. *et al.* Atropopeptides are a Novel Family of Ribosomally Synthesized and Post-translationally Modified Peptides with a Complex Molecular Shape. *Angew. Chem. Int. Ed.* **61**, e202208361 (2022).
49. Ogino, M. H. & Tadi, P. Cyclophosphamide. in *StatPearls* (StatPearls Publishing, Treasure Island (FL), 2025).
50. Kciuk, M. *et al.* Doxorubicin-An Agent with Multiple Mechanisms of Anticancer Activity. *Cells* **12**, 659 (2023).
51. Montecucco, A. & Biamonti, G. Cellular response to etoposide treatment. *Cancer Lett.* **252**, 9–18 (2007).
52. Bubna, A. K. Vorinostat-An Overview. *Indian J. Dermatol.* **60**, 419 (2015).
53. Rajkumar, S. V. & Dispenzieri, A. Multiple Myeloma and Related Disorders. in *Abeloff's Clinical Oncology* 1884-1910.e7 (Elsevier, 2020). doi:10.1016/B978-0-323-47674-4.00101-8.
54. Frisch, M. J. *et al.* Gaussian 16 Rev. A.03. (2016).
55. Gibson, D. G. *et al.* Enzymatic assembly of DNA molecules up to several hundred kilobases. *Nat. Methods* **6**, 343–345 (2009).
56. Durkin, M., Qian, X., Popescu, N. & Lowy, D. Isolation of Mouse Embryo Fibroblasts. *BIO-Protoc.* **3**, (2013).
57. Stringer, C., Wang, T., Michaelos, M. & Pachitariu, M. Cellpose: a generalist algorithm for cellular segmentation. *Nat. Methods* **18**, 100–106 (2021).
58. Letunic, I. & Bork, P. Interactive Tree Of Life (iTOL) v5: an online tool for phylogenetic tree display and annotation. *Nucleic Acids Res.* **49**, W293–W296 (2021).
59. Cole, T. C. H., Hilger, H. H. & Stevens, P. Angiosperm phylogeny poster (APP) – Flowering plant systematics, 2019. Preprint at https://doi.org/10.7287/peerj.preprints.2320v6 (2019).
60. Xiu, W. *et al.* Inhibition of EREG/ErbB/ERK by Astragaloside IV reversed taxol-resistance of non-small cell lung cancer through attenuation of stemness via TGFβ and Hedgehog signal pathway. *Cell. Oncol.* **47**, 2201–2215 (2024).
61. Wang, M. *et al.* Sharing and community curation of mass spectrometry data with Global Natural Products Social Molecular Networking. *Nat. Biotechnol.* **34**, 828–837 (2016).
62. Wishart, D. S. *et al.* The Natural Products Magnetic Resonance Database (NP-MRD) for 2025. *Nucleic Acids Res.* **53**, D700–D708 (2025).

Chapter 5 Generation of Multi-omics Plant Burptide Library

5.1 Statement of contribution

This chapter represents a preliminary exploration intended for future publication. Dr. Roland Kersten and I jointly conceived the research idea, designed the experiments, and supervised the experimental workflow. I was responsible for planting seeds, cultivating them into seedlings, and performing methanol (MeOH) extractions from plant tissues. Dr. Wenjie Li and I conducted the LC-MS/MS analysis of crude plant samples, converted the raw data files, and uploaded them to the MassIVE repository. Together with Gabi May, I constructed the MassQL query to filter mzML files for iminium ions and analyzed the resulting data using Global Natural Product Social Molecular Networking (GNPS). I further analyzed molecular network data using Cytoscape. Additionally, I performed all genomic analyses of successfully grown seedlings and analyzed these data using RepeatFinder.

5.2 Summary

Burpitide-type ribosomally synthesized and post-translationally modified peptides (RiPPs), which originate from BURP-domain-containing proteins, represent an underexplored source of chemical diversity with significant potential for the discovery of therapeutic and bioactive compounds. The identification of burpitides has been limited by the complexity of plant genomes, low tissue abundance, and the absence of integrated multi-omics datasets. To address these challenges, this chapter describes the development of a multi-omics plant peptide natural products library designed to connect genomic potential with metabolomic detection using burpitides as a model. Cultivating more than 200 diverse plant species to the seedling stage enabled comprehensive coverage of metabolic flux stages while minimizing the need for large-scale

facilities and extensive plant material. A standardized 80% methanol extraction and LC-MS/MS workflow generated over 1,150 raw mass spectrometry files, which have been deposited in the MassIVE repository. The bioinformatic pipeline incorporated MassQL and OpenMS for iminium ion filtering, GNPS for molecular networking, and RepeatFinder for genomic mining of 128 available plant protein FASTAs. The approach preliminarily identified known lyciumins and candidate lyciumins with high mass accuracy. Additionally, the analysis revealed a widespread presence of cyclopeptide alkaloids (CPAs). Overall, this work establishes a scalable framework for future automated discovery pipelines and enhances the public availability of plant metabolomics data that can be integrated with existing genomic resources.

5.3 Introduction

Plants generate a wide variety of natural peptidic products with diverse structures and biological activities, facilitating the identification of novel compounds for therapeutic and pesticide development[1–3]. These peptidic natural products are synthesized exclusively via ribosomal pathways. The macrocyclization chemistry involved is of increasing interest because it imparts properties that enhance stability and cell permeability. Macrocycles are highly diversifiable, as macrocyclization can occur through several mechanisms, including head-to-tail, side-chain-to-side-chain, disulfide bonding, and side-chain-to-backbone linkages. Notably, side-chain-to-backbone and side-chain-to-side-chain macrocyclization represent a largely unexplored source of chemical diversity in plants[1,4–8]. Additionally, advances in plant genome sequencing have enabled the high-throughput association of ribosomally produced peptide products with their corresponding biosynthetic genes[7,9–13].

An expanding class of macrocyclic plant peptides, identified through enhanced characterization of biosynthetic genes, consists of ribosomally synthesized and post-translationally modified peptides (RiPPs)[1,4–6]. RiPP biosynthesis initiates with ribosomal synthesis of a precursor peptide, followed by enzymatic post-translational modification of the core peptide(s), and concludes with proteolytic cleavage to produce the mature RiPP[1,4–6]. Plant RiPPs are present in both single- and multi-core precursors, with most containing multiple peptides embedded within the precursor sequence[4]. BURP-domain-containing proteins play a direct role in plant RiPP

biosynthesis, functioning as cyclase domains that catalyze peptide macrocyclization via either an autocatalytic split or a fused pathway[4,7,10–12,14–16].

BURP-domain proteins are plant-specific metalloenzymes defined by a conserved sequence motif: C-X11-12-C-X47-59-CH-X10-CH-X23-27-CH-X23-26-CH[4,11,14,17]. They are named after their four founding members: BNM2, USP, RD22, and PG1β[17]. BNM2, a microspore-derived embryo protein from *Brassica napus, features a BURP domain that is separated* from its precursor peptide[18]. USP, an unidentified seed protein from *Vicia faba, contains* core peptides that may reside within the BURP domain itself[19,20]. RD22 is a drought-responsive protein from *Arabidopsis thaliana, while* PG1β is a β-subunit of polygalacturonase 1 involved in fruit ripening in *Solanum lycopersicum*[21–23]. *T*he RD22 and PG1β subfamilies are distinguished by multiple repeats of 4- to 8-amino-acid core peptides[14,17]. Burpitides, which are side-chain-macrocyclic plant RiPPs, are encoded either within the same polypeptide as the BURP-domain-peptide cyclase or as a repetitive core peptide region N-terminal to the BURP domain[4,7,10–12]. Macrocyclization of burpitides leads to the formation of C(sp2)-C(sp3), C(sp3)-O, and C(sp3)-N linkages via aromatic-aromatic or aromatic-aliphatic side chains[4,10,11,14,15].

Burpitides encompass several subclasses, including lyciumin, cyclopeptide alkaloid, and stephanotic acid[4] (**Figure 5.1**). Lyciumins, isolated from the roots of Chinese wolfberry (*Lycium barbarum), a traditional Chinese medicine for hypertension,* serve as inhibitors of angiotensin-converting enzyme (ACE) and renin[7,24–26]. These compounds are present in various plant tissues and developmental stages, such as seeds, roots, and sprouts[4,7]. Legumenin, a bicyclic octapeptide and lyciumin-type burpitide, was isolated from alfalfa seeds (*Medicago sativa*) and recently identified as a product of a Sali3-2 homolog in the peanut plant, AhyBURP, although its bioactivity remains unknow[11,12,15]. Cercic acid, another lyciumin-type burpitide from *Cercis canadensis, also lacks characterized* bioactivity[11,27].

Cyclopeptide alkaloids (CPAs), which are small peptides composed of 4 to 6 amino acids, represent one of the largest classes of burpitides and exhibit diverse bioactivities[4,16,28–30]. For example, Sanjoinine A from jujube demonstrates GABA-adrenergic anxiolytic activity, while adouetine X from *Ziziphus oxyphylla show*s analgesic effects in mouse models and inhibits Ca2+-ATPase and Na+/K+ ATPase *in vitro*[31–35]. Elaeagnin, a CPA derived from fused burpitide pathways in *Elaeagnus pungens, was* identified using HypoRiPPAtlas as a proof-of-concept[9]. CPAs have

also demonstrated antiplasmodial, insecticidal, and antiviral activities[36–38]. Additionally, CPAs can form bicyclic structures, such as Selanine A from African clubmoss[11,39].

Stephanotic acids constitute another class of burpitides, initially isolated from *Stephanotis floribunda* and later biosynthetically characterized from *Cercis canadensis*[11,12,40]. These compounds did not exhibit bioactivity in in vitro cytotoxicity assays against lung cancer cells[12]. Moroidins, classified as a bicyclic form of stephanotic-acid-type burpitides, were originally isolated from the Australian stinging nettle *Dendrocnide moroides* during investigations of pain-inducing agents[11,12,41]. Analogs have also been identified in *Celosia argentea* (celogentin A–K)*, Kerria japonica,* and *Bauhinia tomentosa*[41–45]. Subsequent studies revealed that moroidin and celogentin C exhibit in vitro anti-tubulin polymerization activity and cytotoxicity against lung adenocarcinoma cells, suggesting that a second ring is critical for cancer cell cytotoxicity[12,42].

Recently, a moroidin analogue (moroidin-QLLVWRNH) was characterized in three plant families and demonstrated selective, dose-dependent cytotoxicity against H1437 lung cancer cells, with minimal toxicity toward non-cancerous fibroblasts[10]. This finding positions the compound as promising for rational drug design targeting non-small cell lung cancer. The study also identified additional stephanotic acid-type burpitides in nature, featuring diverse Val-Trp and Phe-Tyr crosslinks in the second ring, such as glechomanin and mercurialin, which did not display in vitro dose-dependent cytotoxicity[10]. Furthermore, the research expanded the known chemotaxonomic distribution of stephanotic acid-type burpitides across the plant kingdom, documenting stephanotic acids in pinks, mints, spurges, grasses, and Lowiaceae; moroidins in pinks, grasses, and Lowiaceae; glechomanins in mints and Lowiaceae; and mercurialins in spurges[10].

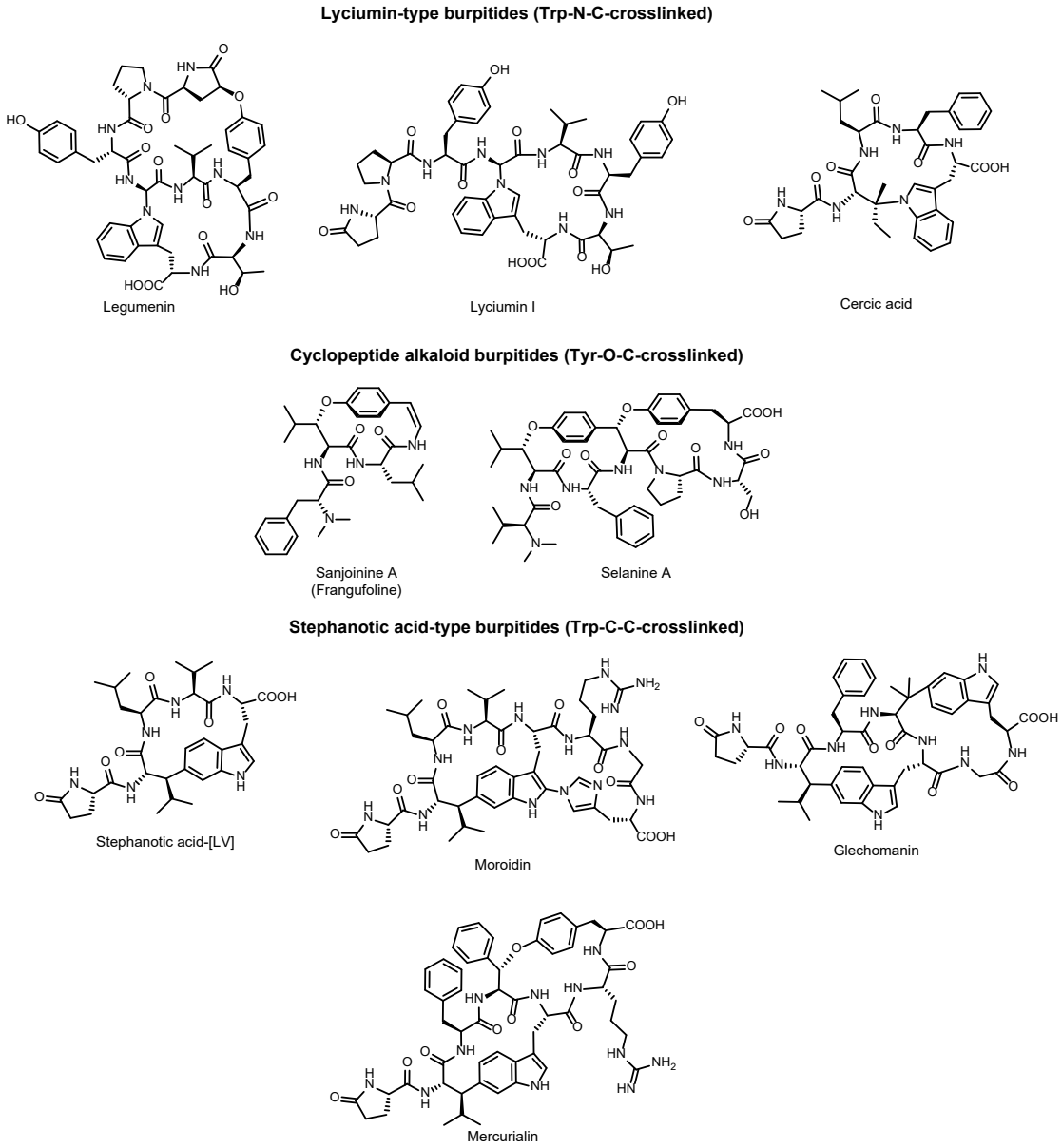


Figure 5.1 Examples of subclasses of known burpitides.
Top - lyciumin-type burpitides, middle – cyclopeptide alkaloid-type burpitides, bottom – stephanotic acid-type burpitides.

The structural and functional diversity of burpitides underscores the importance of identifying uncharacterized precursor peptides that contribute to both novel and established biosynthetic pathways[46]. This process expands chemotaxonomy, facilitates bioactivity investigations, and clarifies endogenous locations and functions. However, the discovery of burpitides is impeded by the complexity of plant genomes, which are often long and repetitive, exhibit low clustering of biosynthetic genes, and suffer from poor annotation[7,10–12,16,28]. Additional challenges include limited isolation from source plant tissues, complex metabolomes, and

difficulties in structural prediction due to potential post-translational modifications by uncharacterized downstream tailoring enzymes[28]. As burpitides originate from ribosomally synthesized and post-translationally modified peptide (RiPP) precursors, they can be identified in plant genomic databases with associated protein sequencing data and through structure-guided metabolic profiling of plant tissues[46]. Recent advances in mass spectrometry, genomic mining, and metabolomic identification tools have enabled systematic classification of natural products from LC-MS/MS data obtained from plant extracts[1,46]. Nevertheless, a significant knowledge gap persists in the availability of integrated plant metabolomics, genomics, and protein-sequencing data necessary for the discovery of cyclic peptide chemistry. This issue reflects a broader challenge in the field of plant natural products: understanding the distribution of the vast diversity of plant metabolites across the plant kingdom[47].

This chapter addresses the challenge of predicting cyclic peptide chemistry in plant genomes by integrating plant metabolomic profiling with bioinformatic mining of biosynthetic genes to identify peptide macrocycles. Additionally, the study expands the publicly available coverage of plant metabolomes, thereby supporting future research on automated linkages among plant metabolites, genes, and transcripts.

5.4 Results

5.4.1 Plant Cultivation of Diverse Seeds for Seedling Growth

The challenge of expanding plant metabolomics availability from diverse tissue samples for linkage to genetic and protein biosynthetic precursors was addressed by (1) utilizing hundreds of diverse plant seeds that are readily cultivated in laboratory settings, (2) implementing sample preparation protocols compatible with both untargeted and peptide-specific targeted metabolomics data acquisition across low and high mass-to-charge scanning environments, and (3) conducting preliminary burpitide analyte analysis from metabolomics data using MassQL, OpenMS, and GNPS, as well as profiling corresponding plant genomes for candidate precursor-peptide genes with RepeatFinder[11,48–51]. The study encompassed a broad phylogenetic spectrum by sourcing

plant seeds from various commercial suppliers and the USDA, resulting in a collection of over 300 distinct plant species.

Planting the seeds according to the specific packaging instructions yielded 204 seedlings (61% germination rate, **Table 5.3**), which were subsequently used for methanol extractions and LC-MS/MS analysis to maximize chemical diversity. Seedlings were selected for study due to their high metabolic flux of secondary metabolites, which is attributed to reduced defensive mechanisms and may promote the production of RiPPs, as some RiPPs function as signaling agents[5,6,11]. Additionally, seedling tissues are soft and easily homogenized, facilitating consistent extractions, and their growth can be controlled and standardized to minimize environmental variability in metabolite profiles. Biosynthetic gene expression may also be elevated during early developmental stages. The use of seedlings also addresses practical constraints, such as limited seed availability and lengthy maturation times, by enabling chemical profiling from small quantities of starting material.

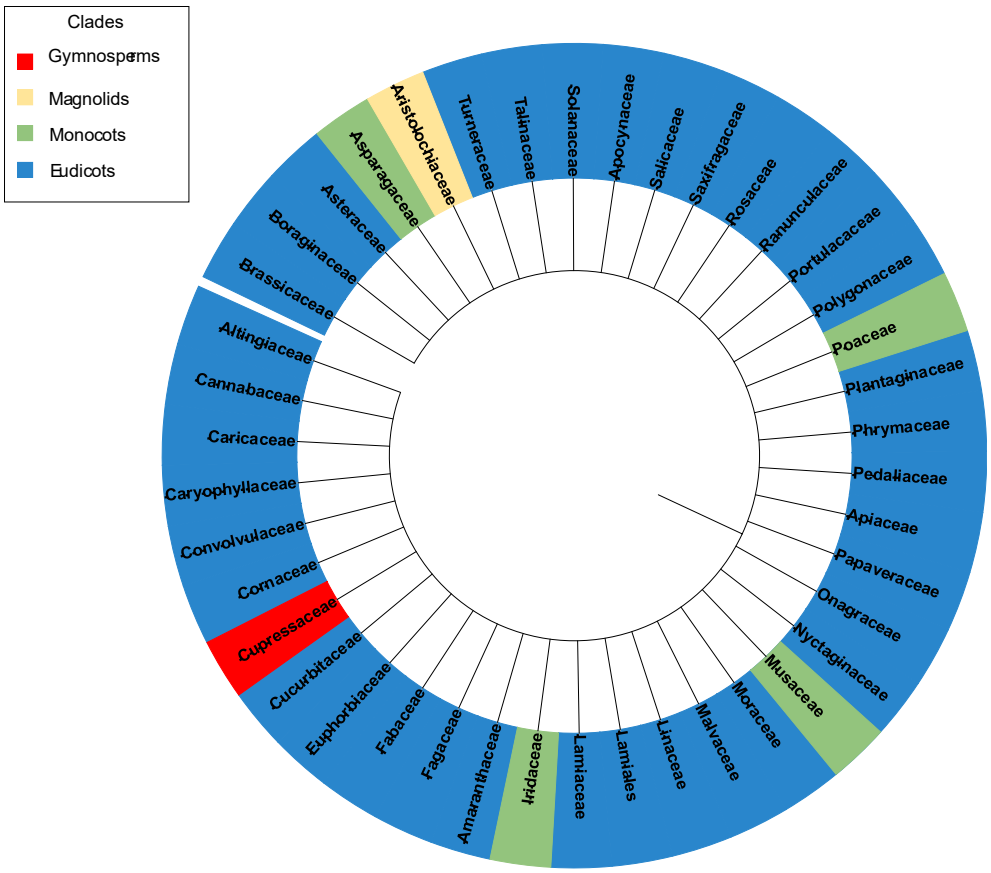


Figure 5.2 Phylogenetic distribution of seedlings represented in this study.
Displayed phylogenetic resolutions are plant orders.

The legume family (Fabaceae) represented the most abundant group of successfully cultivated plants. This family often thrives in nitrogen-deficient soils due to a symbiotic relationship with soil bacteria that fix atmospheric nitrogen, while the plants supply sugars to the bacteria[52] (**Figure 5.2)**. Successfully cultivated species in this family included peanuts (*Arachis hypogaea*, *Arachis stenosperma*), soybeans (*Glycine max, Glycine soja), pigeon p*ea (*Cajanus cajan*), and various clovers (*Trifolium*). The grass family (Poaceae) also demonstrated high yields (**Figure 5.2)**. This family requires substantial sunlight and nitrogen for optimal growth and includes both drought-tolerant and moisture-dependent species, such as rice (*Oryza sativa*), African rice (*Oryza glaberrima*), wheat (*Triticum aestivum*), foxtail millet (Setaria italica), and switchgrass (*Panicum virgatum*)[53].

The nightshade family (Solanaceae), which thrives in warm conditions and requires nutrient-rich soil and consistent moisture, was also prominent in the dataset. Cultivated species included tobacco (*Nicotiana tabacum*), several wild tobaccos (*N. attenuata*, *N. occidentalis, N. megalosiphon), t*omatoes (*Solanum lycopersicum, S. pennellii, S. chilense), and p*eppers (*Capsicum chinense, C. annuum*)[54] (**Figure 5.2**). The *Brassicaceae* family, which grows optimally during cool seasons in well-drained, fertile soil with direct sunlight and produces nutrient-dense vegetables, was represented by thale cress (*Arabidopsis thaliana*), mustards (*Brassica juncea, B. carinata, B. cretica), and a*rugula (*Eruca vesicaria*)[55] (**Figure 5.2**). Numerous species from the Malvaceae family, recognized for their resilience and adaptability to diverse growth conditions, were also present, including emerald okra (*Abelmoschus esculentus*), baobab tree (*Adansonia digitata*), and various cottons (*Gossypium hirsutum, G. barbadense*)[56] (**Figure 5.2**). Several notable plants with documented medicinal value were identified, such as Sweet Annie *(Artemisia annua) from* the Asteraceae family, which is the source of the anti-malarial agent artemisinin; goji berry (Lycium barbarum), a Solanaceous shrub used in Chinese herbal medicine for hypertension; and Madagascar periwinkle (*Catharanthus roseus*), which produces the chemotherapeutic agent's vinblastine and vincristine[3,7,25,26] (**Figure 5.2**). Additionally, the findings indicate that seedlings of rare or difficult-to-cultivate species can be grown more readily in controlled environments, including African baobab (*Adansonia digitata*) from the mallow family, mimosa tree (*Albizia julibrissin*) from the legume family, and giant sequoia (*Sequoiadendron giganteum*) from the Cupressaceae family (**Figure 5.2**).

5.4.2 Workflow for Tissue Extractions and LC-MS/MS Analysis Yields More Available Metabolomes

An adapted sample preparation strategy was implemented to increase the availability of plant metabolomics data from diverse tissue samples, accommodating both untargeted and peptide-specific targeted metabolomics data acquisition in low- and high-mass-to-charge (m/z) scanning environments. The extraction protocol was standardized to maximize chemical diversity across the tissues studied. Root (0.1 g fresh weight), leaf (0.3 g fresh weight), and stem (0.2 g fresh weight) tissue samples were harvested from plants grown from seed to seedling to obtain tissue-specific metabolite profiles that might be overlooked in single-organ screens. Softer tissues, such as leaves and roots, were processed in plastic tubes with glass beads, while harder tissues, such as stems, were processed in metal tubes with metal beads. A tissuelyzer was used to homogenize the flash-frozen samples and ensure complete cellular rupture across all samples. Plant metabolites were extracted in 80% methanol (HPLC grade) during a 10-minute incubation at 60 °C in a water bath to solubilize a broad range of compounds with varying polarities, including hydrophobic burpitides.

To protect sample integrity and ensure the longevity of the Phenomenex Kinetex C18 column and mass spectrometer, supernatants were clarified by high-speed centrifugation to pellet cellular debris and filtered through a 0.2-micron filter (Cytiva syringeless filters) to remove micro-particles that could interfere with ionization during data acquisition. Samples were analyzed using an ESI–Quadrupole–Orbitrap mass spectrometer with C18 reverse-phase chromatography and Data Dependent Acquisition (DDA) in positive ion mode. To maximize metabolite coverage, initial scans were performed in a high m/z range (450-1500 m/z) to target peptides and in a low m/z range to capture small molecules, both of which facilitated comprehensive metabolic profiling of young plant samples. The subsequent analysis in this section focuses exclusively on the high-mass-to-charge data, as it contains the macrocyclic burpitides relevant to this study.



MassQL Commandline Workflow

Step	What it Does	Output
ftpdownload.sh	Move 100s of .mzML files automatically	file transfer
massQL.py	Automatically runs massQL query on entire folder of files. "QUERY MS2DATA WHERE MS2PROD=..."	A .csv of all MS2 hits and their corresponding MS data pulled from the original .mzML
openms_write.py	Writes filtered .mzMLs using original and .csv of hits.	An .mzML consisting only of the hits from the massQL query

Figure 5.3 Workflow for MassQL and OpenMS data filtering.

>1150 raw mass spectrometry files were converted to the open-source .mzML format and uploaded to the MassIVE (Mass Spectrometry Interactive Virtual Environment) repository to facilitate centralized access for further analysis by the research community[57]. A persistent challenge in the identification of side-chain-macrocyclic ribosomally synthesized and post-translationally modified peptides (RiPPs) is the limited b- and y-ion fragmentation observed; however, these compounds consistently produce at least two characteristic amino acid iminium ions in their MS/MS spectra[11,48]. The data were processed on the University of Michigan high-performance computing cluster (HPC) using MassQL (Mass Spectrometry Query Language) to execute an iminium ion search query across all high m/z files, resulting in filtered MS data and MS/MS spectra containing one to five proteinogenic amino acid iminium ions from the original .mzML files[48] (**Figure 5.3**). Implementing a requirement of at least one diagnostic iminium ion per spectrum, with a 10 parts per million (PPM) tolerance, substantially expanded the pool of candidate peptide molecular clusters (**Table 5.1**). Additionally, within the Great Lakes HPC environment, OpenMS was employed to generate filtered .mzML files from MassQL hits[49] (**Figure 5.3**). These curated datasets enabled more precise targeted identification of peptide-specific clusters throughout the seedling library using molecular networking.

Table 5.1 Iminium ion masses of proteinogenic amino acids.

Amino acid	Iminium ion	Monomethylated	Dimethylated
Alanine	44.0495	58.0651	72.0808
Aspartate	88.0393	102.0550	116.0706
Glutamate	102.0550	116.0706	130.0863
Phenylalanine	120.0808	134.0964	148.1121
Glycine	30.0338	44.04948	58.06513
Histidine	110.0713	124.0869	138.1026
Leucine	86.0964	100.1121	114.1277
Isoleucine	86.0964	100.1121	114.1277
Lysine	101.1073	115.1230	129.1386
Methionine	104.0529	118.0685	132.0842
Asparagine	87.0553	101.0709	115.0866
Proline	70.0651	84.0808	98.09643
Glutamine	101.0709	115.0866	129.1022
Arginine	129.1135	143.1291	157.1448
Serine	60.0444	74.0600	88.0757
Threonine	74.0600	88.0757	102.0913
Valine	72.0808	86.0964	100.1121
Tryptophan	159.0917	173.1073	187.1230
Tyrosine	136.0757	150.0913	164.1070
Pyroglutamate	84.0444		

5.4.3 Preliminary Data Analysis to Aid in Discovery of Macrocyclic Burpitides From Seedling Extracts

The resultant molecular network was interrogated for the presence of 19 known burpitide plant RiPP reference spectra, which were incorporated during the GNPS workflow[51]. The broader occurrence of cyclic plant peptides was predicted by clustering with reference peptides in the network and compared to the total number of nodes. The MassQL-cardinality-search network encompassed all reference cyclic peptides, including Adouetine X, Americine, Ceanothine B, Ceanothine C, cyclo-VPPIFY, cyFLLY, cyFLLY-DC (decarboxylated), Deoxybouvardin, Elaeagnin, Lyciumin A, Lyciumin B, Lyciumin C, Lyciumin D, Moroidin, RA-XII, Sanjonine A, Stephanotic Acid, Segetalin B, and Segetalin A. The predicted occurrence of burpitides in the network was approximately 1.2% (345 burpitide nodes out of 26,621 total nodes) (**Figure 5.8**). Analysis of burpitide-specific nodes indicated a broad distribution of side-chain macrocyclic RiPPs, with Lyciumins (77 nodes) and Cyclopeptide Alkaloids (CPAs) (218 nodes) representing the most dominant and intense clusters **(Figure 5.4 and Figure 5.9**). The topography of the network is defined by edges between the nodes to connect them, which represent spectral

alignment and structural similarities between metabolites. These connections are quantified by a cosine similarity score based on tandem mass spectrometry (MS/MS) fragmentation patterns; thus the edge denotes shared chemical sub-structures based on conserved fragmentation.

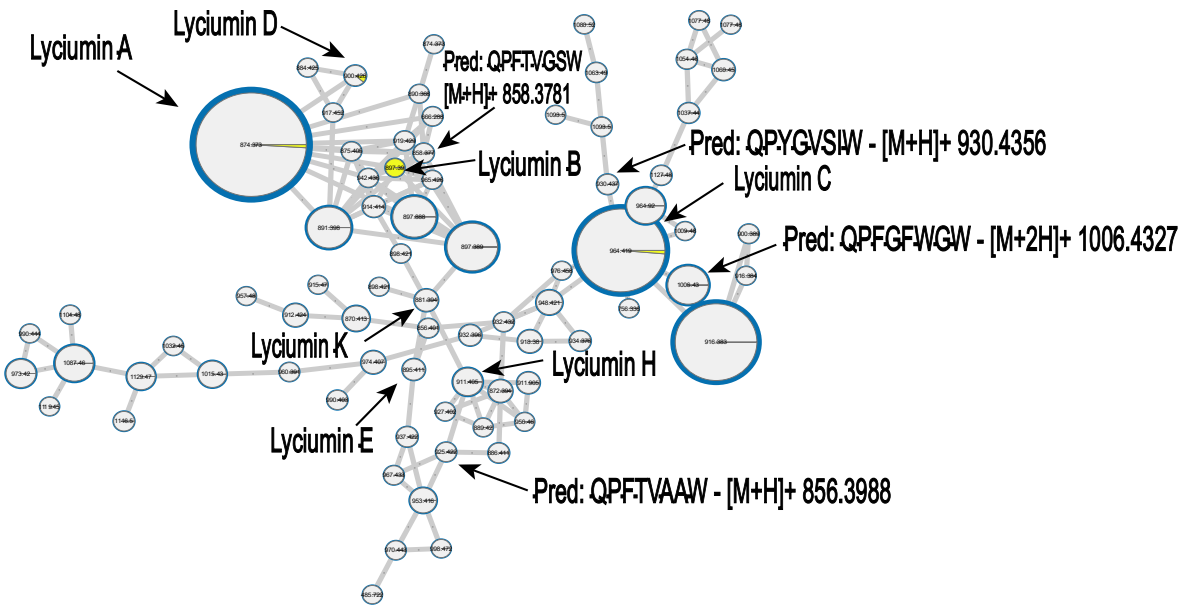


Figure 5.4 Lyciumin subset of molecular network of MassQL-processed LC-MS-datasets with cardinality command of 1-5 different amino acid iminium ions in a given tandem MS spectrum.
Reference peptides are highlighted in yellow. Known lyciumin and candidate peptides are denoted above.

The networking data demonstrates significant clustering with reference spectra for Lyciumin-A-D, sourced from the roots of *Amaranthus* species and *Lycium barbarum* **(Figure 5.4)**. Lyciumin-A (874.37 Da) exhibited the highest precursor ion sum intensity ($3.05x10^9$) within the lyciumin cluster (**Table 5.2**). Lyciumins A, B, and C were identified with 0.00 ppm error, underscoring the robustness of the extraction workflow and MassQL filtering (**Figure 5.4, Table 5.2**). Lyciumin B was detected in multiple *Solanaceous species* and in the root of *Aristolochia californica* (**Table 5.2**). A literature review enabled the identification of nodes for Lyciumins E, H, and K within the molecular network[7] (**Figure 5.4, Table 5.2**). Lyciumins E and H were confirmed in *Beta vulgaris* subsp. vulgaris, and Lyciumin H was also identified in *Glycine max* (**Table 5.2**). Lyciumin K was identified in *Solanum pennellii* (**Table 5.2**). Additionally, the cyclopeptide alkaloids clustered with Ceanothine B are strongly chemotaxonomically associated with the *Fabaceae* family, particularly within the genera *Trifolium*, *Cicer*, and *Glycine* (**Figure 5.9**). This extensive molecular network highlights the potential for discovering novel biosynthetic architectures or split-cyclase systems with uncharacterized biosynthetic machinery (**Figure 5.8**).

The final objective of this project was to link plant cyclic peptide metabolites to the core peptides encoded by BURP-domain cyclases. Of the 204 plants analyzed, 128 had FASTA files for protein sequences associated with their genomic sequencing data in the NCBI Gene Datasets, enabling the direct linkage of biosynthetic precursor peptide sequences to analytes (**Table 5.3**). The availability of these sequences eliminated the need for sequence assembly, computational translation, and open reading frame identification. The datasets were queried for homologous BURP-domain cyclase genes using BLASTp (e-value cutoff of 1e-5), with Sali3-2, a USP-type BURP-domain protein, as the query[58,59]. The resulting protein sequences were analyzed with RepeatFinder, which identifies core peptide motifs in open reading frames adjacent to BURP domains[11]. The RepeatFinder search included peptide types characterized by the Kersten lab: lyciumin, legumenin, bicyclopeptide alkaloid, cyclopeptide alkaloid, cercic acid, stephanotic acid, moroidin, elaeagnin, glechomanin, mercurialin, and oryzin. Due to the prevalence of the lyciumin cluster in the data, several candidate precursors were identified in *Solanum dulcamara, Beta vulgaris subsp. vulgaris, and Amaranthus tricolor* (**Figure 5.5**).



Figure 5.5 Candidate BURP-domain-peptide cyclase precursors identified with RepeatFinder.
Top left – sequence from *Solanum dulcamara*, bottom left - sequence from *Amaranthus tricolor*, Top right and bottom right - sequences from *Beta vulgaris subsp. Vulgaris*. Core peptides highlighted in Figure 5.4/Table 5.2 are highlighted in green and other candidate peptides highlighted in red.

To provide structural context for nodes within the generated molecular networks, predicted [M+H]+ and [M+2H]+ masses for over 440 core peptides identified by RepeatFinder were

generated using ChemDraw and stored as 2D structures and SMILES representations for cross-referencing with experimental LC-MS/MS parent ion data. Cross-referencing the *in silico* SMILES library enabled the profiling of several candidate lyciumin sequences that matched metabolite masses with high mass accuracy. Candidate QPYGVSIW (930.4361 Da) was identified in *Solanum dulcamara* with a mass error of 0.537 ppm (**Figure 5.4-6, Table 5.2**). The *Beta vulgaris* subsp. v*ulgaris* metabolome contained two candidate peptides, QPFTVAAW (856.401 Da) and QPWTISAW (925.422 Da), with mass errors of 2.57 ppm and 1.84 ppm, respectively (**Figure 5.4-6, Table 5.2**). Additionally, candidate peptide QPFTISAW (886.411 Da, 1.13 ppm) was identified serendipitously during the search for QPWTISAW in the *Beta vulgaris subsp. vulgaris* root metabolome, and this peptide was also present in a corresponding protein sequence (**Figure 5.4-6, Table 5.2**).

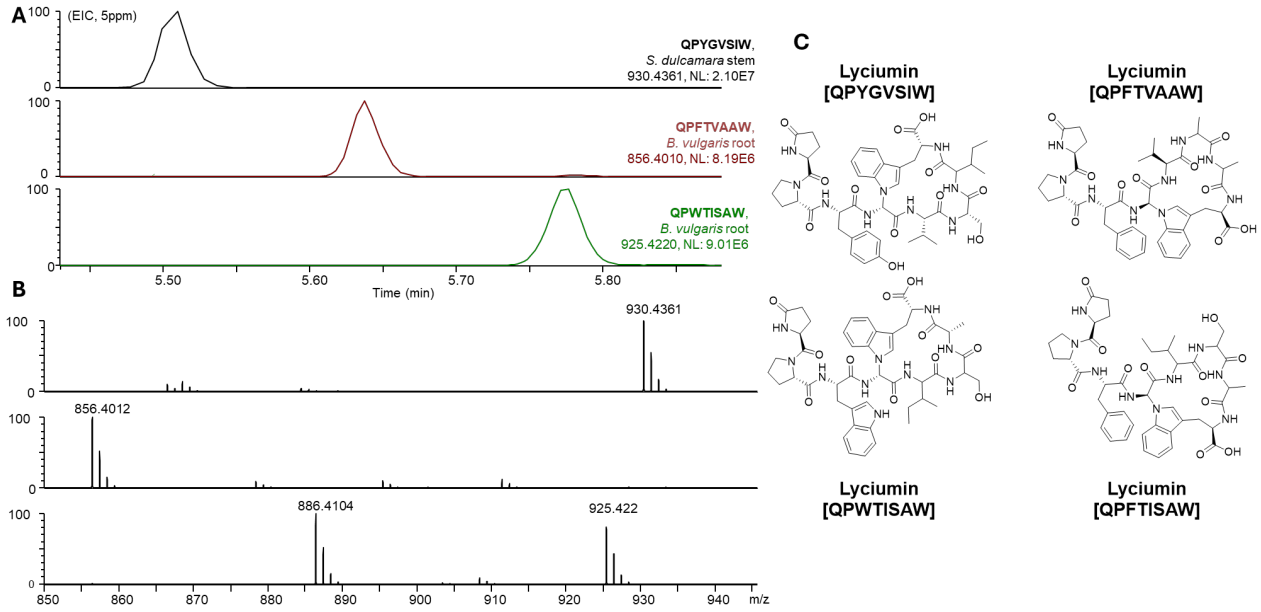


Figure 5.6 Detection of candidate lyciumin derivatives in plant tissues.
(A) LC-MS chemotyping of predicted lyciumin derivative chemotypes in plant extracts. (B) MS analysis of predicted lyciumin derivative chemotypes. (C) Predicted lyciumin derivative chemotypes (core peptides).

Table 5.2 Known and Candidate Lyciumins Identified Through GNPS Clustering.

Predicted Sequence	Predicted Mass (Da)	Observed mass (Da)	Mass Error (PPM)	Plant Species	Ion intensity
Lyciumin A (Ref) - QPYTVGSW	874.373 $[M+H]^+$	874.373	0.00	*Amaranthus hybridus, Amaranthus retroflexus, Amaranthus cruentus, Amaranthus albus,*	3.05E+09

				Amaranthus tricolor, Amaranthus hypochondriacus, Amaranthus blitum, Lycium barbarum	
Lyciumin B (Ref) - QPWGVGSW	897.389 $[M+H]^+$	897.389	0.00	*Solanum dulcamara, Aristolochia californica, Lycium barbarum, Solanum pennellii, Solanum dulcamara*	1.05E+09
Lyciumin C (Ref) - QPYTVFSW	964.42 $[M+H]^+$	964.42	0.00	*Amaranthus hybridus, Amaranthus retroflexus, Amaranthus cruentus, Solanum dulcamara*	2.31E+09
Lyciumin D (Ref) - QPYGVGIW	900.4251 $[M+H]^+$	900.426	1.00	*Solanum dulcamara*	8.08E+07
Lyciumin K - QPWGVGAW	881.3941 $[M+H]^+$	881.394	0.00	*Solanum pennellii*	1.37E+08
Lyciumin E – QPWTVAAW	895.4097 $[M+H]^+$	895.411	1.45	*Beta vulgaris* subsp. *vulgaris*	1.08E+08
Lyciumin H - QPWGVGTW	911.4047 $[M+H]^+$	911.405	0.33	*Beta vulgaris* subsp. *Vulgaris, Glycine max, Solanum pennellii*	3.34E+08
QPYGVSIW	930.4356 $[M+H]^+$	930.4361	0.537	*Solanum dulcamara*	6.43E+07
QPFGFWGW	1006.4327 $[M+2H]^{2+}$	1006.43	0.00	*Amaranthus hybridus, Amaranthus retroflexus, Amaranthus cruentus*	7.16E+08
QPFTVAAW	856.3988 $[M+H]^+$	856.401	2.57	*Beta vulgaris subsp. vulgaris*	3.17E+07
QPFTVGSW	858.3781 $[M+H]^+$	858.377	1.28	*Amaranthus tricolor, Amaranthus hypochondriacus*	3.93E+07
QPWTISAW	925.4203 $[M+H]^+$	925.422	1.84	*Beta vulgaris subsp. vulgaris*	3.85E+07
QPFTISAW	886.4094 $[M+H]^+$	886.411	1.13	*Beta vulgaris subsp. vulgaris*	5.44E+07

5.5 Discussion

The integration of high-resolution metabolomics of diverse seedlings with plant RiPP genomic mining in this chapter advances the discovery of plant-derived side-chain macrocyclic RiPPs and significantly expands the availability of plant metabolomes for future research. Utilizing seedlings in this study eliminated the need for extensive greenhouse facilities and circumvented the long maturation cycles typical in natural habitats, thereby conserving both financial and temporal resources. This approach also demonstrated that the discovery workflow is effective with limited or rare plant material. Seedling cultivation enabled high-throughput metabolic profiling during developmental stages characterized by elevated metabolic flux, thereby facilitating the identification of RiPPs. The project sought to bridge the gap between genomic prediction and chemical detection of burpitide structures. Identification of the candidate lyciumin QPFGFWGW sequence underscores the need for further paired omics studies (**Table 5.2**). Although the peptide was detected in *Amaranthus* species via metabolomics, the corresponding protein FASTA files were not available in the NCBI genomic datasets for subsequent genome mining. Additionally, RepeatFinder analysis identified this core motif in *Pisum sativum* (syn. *Lathyrus oleraceus*), underscoring the potential for missing phylogenetic distributions when relying solely on existing annotations. A persistent challenge in this project and in the processing of RiPP peptidic molecular networking data is the interpretation of mass differences between peptides with identical core sequences[11]. While the workflow demonstrated initial success in matching predicted masses to core peptide sequences, future iterations should focus on enhancing automated MS/MS fragmentation to distinguish structural isomers. Although nodes were not further analyzed for MS/MS fragmentation patterns, the undefined cyclic peptide nodes clustered with reference spectra provide a promising starting point for characterizing burpitide chemistry. Prioritizing analyte spectra with sum precursor ion intensities greater than 1×10^7 is expected to increase the likelihood of successful downstream characterization and peptide isolation. Furthermore, automated prediction of post-translational modifications (PTMs), particularly for cyclopeptide alkaloids (CPAs), will be essential to incorporate into this pipeline. CPA clusters accounted for a substantial portion of the data, but analysis was complicated by limited FASTA availability, a higher likelihood of encountering split-biosynthetic BURP-domain cyclase systems, and uncharacterized downstream enzymes, such as decarboxylases and N-methyltransferases, that can introduce modifications altering the observed mass from the predicted core peptide[16,28]. Incorporating predictive algorithms to account for downstream tailoring enzymes will refine and

expand the predictive capacity of the SMILES library. Future iterations of this workflow, aiming to automate RiPP discovery with exact mass matches from genetic data, must achieve a mass error of less than 5 ppm. Ultimately, expanded seedling cultivation and metabolome profiling, combined with increased sequencing of plant tissues and improved bioinformatic and cheminformatic search models, will provide essential linkages for future plant natural product discovery pipelines, enabling more rapid and efficient discovery.

5.6 Methods and Materials

5.6.1 Materials

All chemicals and reagents were from Thermo Fisher Scientific Inc. if not otherwise noted. All LC-MS solvents were Optima LC-MS-grade solvents (Fisher Scientific Inc.). LC-MS analysis was performed on a Thermo QExactive Orbitrap mass spectrometer coupled to a Vanquish Ultra high performance liquid chromatography instrument. LC-MS data was analyzed with QualBrowser from Thermo Xcalibur software package (v4.3.73.11, Thermo Scientific).

5.6.2 Sample Preparation of Plant Material and Peptide Chemotyping

For all plants samples, 0.3g of leaf, 0.2g of stem, and 0.1g of root material (fresh weight) was collected, frozen at –80 °C, and ground with a MP Biomedicals FastPrep-24-5G Tissuelyser with a Coolprep cryogenic adapter in 2-ml MP Biomedicals tubes with 2.3-mm Zirconia beads (Fisher Scientific) for three 20 s cycles at 6 ms^{-1} speed. 1mL of 80/20% MeOH/Genpure H_20 (v/v) was added to the sample vials prior to a second round of lysing in the conditions described above. Samples were incubated in a 60 °C water bath for 10 minutes, then centrifuged at 16,000 × g for 5 minutes to pellet insolubles. Glass pasteur pipettes were used to transfer supernatants to Whatman syringeless filters (0.2 µm) for filtration, then to LC–MS/MS vials with 0.5mL glass inserts for analysis. Crude MeOH samples were subjected to LC–MS/MS analysis with the following parameters: injection volume 5 µl; LC, Phenomenex Kinetex 2.6 µm C18 reverse phase 100 Å

150 ×3 mm LC column; LC gradient, solvent A, 0.1% formic acid; solvent B, acetonitrile (0.1% formic acid); 0 min, 10% B; 5 min, 60% B; 5.1 min, 95% B; 6 min, 95% B; 6.1 min, 10% B; 9.9 min, 10% B; 0.5 ml min−1; MS, positive ion mode; full MS, resolution 70,000; mass range 400–1,200 *m/z; dd-MS2 (data-dependent MS/MS), resolution 17,500; AGC target 1 × 105, loop count 5, isolation width 1.0 m/z, collision energy 25 eV and dynamic exclusion 0.5 s.* LC–MS data were analyzed with QualBrowser in the Thermo Xcalibur software package (v.4.3.73.11, Thermo Scientific).

5.6.3 Seeds and Plant Cultivation

All plants were grown from seed in Premier Horticulture Pro Mix HP High Porosity with Mycorrhizae for plant growth under plant growth lights with a 16 h light/8 h dark cycle for 4–12 weeks at room temperature or according to specific seed packaging instructions. Seeds were grown from seeds acquired from the USDA, Seedville, and various sources.

5.6.4 Molecular Network Generation

A molecular network was generated using the online workflow available through the GNPS website through filtered removal of all MS/MS fragment ions within +/- 17 Da of the precursor m/z, window filtering to choose only the top 6 fragment ions in the +/- 50Da window throughout the spectrum, setting the precursor ion mass tolerance to 0.05 Da, and the MS/MS fragment ion tolerance of 0.25 Da[51]. The resultant edges were filtered to have a cosine score above 0.7 and more than 5 matched peaks edges between two nodes were kept in the network if nodes appeared in each other's respective top 10 similar nodes. The maximum molecular family size was limited to 500 with lower scoring edges continually removed from the molecular families until it breached below the threshold. The queried spectra included in the network were filtered against GNPS' spectral libraires. Finally, 19 burpitide plant RiPPs reference library spectra filtered the same way as the input data and all matches were kept between reference peptide spectra and network spectra if they scored above 0.7 with at least 6 matched peaks. Molecular networking data were analyzed with

Cytoscape (v.3.8.0)[60]. Candidate peptide clusters were identified by the presence of reference plant peptide spectra.

5.6.5 Phylogenetic Tree Construction and Bar Graph Construction

The tree was constructed in the web server Interactive Tree of Life. The bar graph was generated using Rstudio.

5.6.6 MassQL and OpenMS Filtering

RAW.files from the seedling samples were converted to centroided mzXML-files with Msconvert[61]. Files were transferred from MassIVE to the Great Lakes High Performance Computing (HPC) Cluster at the University of Michigan, Ann Arbor via an FTP client[57]. Datasets were filtered using MassQL and OpenMS was using to write out the hits in .mzML format[48,49]. The scripts are below:

```
# With SLURM, you can load your modules in the SBATCH script
module load Bioinformatics
module load python/3.12.1
pip install massql

#  Put your job commands after this line
python DO_Seedling_massql.py


# Contents of DO_Seedling_massql.py
import os
import subprocess
import csv

input_folder = "/scratch/rkersten_root/rkersten1/dousley/Seedling_MzMlData_DO_2025/Plant_metabolomics_seedlings"
output_folder = 
"/scratch/rkersten_root/rkersten1/dousley/Seedling_MzMlData_DO_2025/Plant_metabolomics_seedlings_output_min1_iminium"
massql_query = "QUERY MS2DATA WHERE MS2PROD=(58.06513 OR 60.04439 OR 70.06513 OR 72.08078 OR 74.06004 OR 84.04439 OR 84.08078 OR 
86.09643 OR 87.05529 OR 88.0393 OR 88.07569 OR 100.11208 OR 101.07094 OR 101.10732 OR 102.05495 OR 102.09134 OR 104.05285 OR 110.07127 OR 
114.12773 OR 115.08659 OR 115.12297 OR 116.0706 OR 118.0685 OR 120.08078 OR 124.08692 OR 129.10224 OR 129.11347 OR 129.13862 OR 130.08625 OR 
132.08415 OR 134.09643 OR 136.07569 OR 138.10257 OR 143.12912 OR 148.11208 OR 150.09134 OR 157.14477 OR 159.09167 OR 164.10699 OR 173.10732 
OR 187.12297):CARDINALITY=range(min=1,max=5):TOLERANCEPPM=10:INTENSITYPERCENT=5"

files_with_hits = 0
total_unique_scans = 0
failed_files = []

for filename in os.listdir(input_folder):
    if filename.endswith(".mzML"):
        input_path = os.path.join(input_folder, filename)
        output_csv = os.path.join(output_folder, f"{os.path.splitext(filename)[0]}_results.csv")

        cmd = [
            "massql",
            input_path,
            massql_query.strip(),
            "--output_file", output_csv
        ]

        print(f"Running query on {filename}")
        try:
            subprocess.run(cmd, check=True)
        except subprocess.CalledProcessError as e:
            print(f"[ERROR] MassQL failed on {filename}: {e}")
            failed_files.append(filename)
            continue  # Skip this file and continue with the next
```

```
                    # Process the CSV for unique scan counting
                    if os.path.exists(output_csv):
                        with open(output_csv, "r") as f:
                            reader = csv.DictReader(f)
                            unique_scans = set()

                            for row in reader:
                                scan = row.get("scan") or row.get("scanNum") or row.get("scan_id")
                                if scan:
                                    unique_scans.add(scan)

                            num_unique = len(unique_scans)
                            if num_unique > 0:
                                files_with_hits += 1
                                total_unique_scans += num_unique

            print("All queries completed.")
            print(f"Files with >0 results: {files_with_hits}")
            print(f"Total unique scan numbers across all results: {total_unique_scans}")

            if failed_files:
                print(f"Failed files include: {', '.join(failed_files)}")


            # With SLURM, you can load your modules in the SBATCH script
            module load Bioinformatics

            #  Put your job commands after this line
            source /home/dousley/miniconda3/etc/profile.d/conda.sh
            conda activate openms_env
            echo "Using Python from: $(which python)"
            python -c "import pyopenms; print('pyOpenMS version:', pyopenms.__version__)"
            cd /scratch/rkersten_root/rkersten1/dousley/massQL
            python openms_writing1.py


            # Contents of openms_writing1.py
            import os
            import pandas as pd
            from pyopenms import MSExperiment, MzMLFile
            import glob

            # === CONFIGURATION ===
            mzml_folder = "/scratch/rkersten_root/rkersten1/dousley/Seedling_MzMlData_DO_2025/Plant_metabolomics_seedlings"
            csv_folder                                                                                                                                  =
"/scratch/rkersten_root/rkersten1/dousley/Seedling_MzMlData_DO_2025/Plant_metabolomics_seedlings_output_min1_iminium_highonly"
            output_folder = "/scratch/rkersten_root/rkersten1/dousley/Seedling_MzMlData_DO_2025/Filterered_Seedling_Data_1iminium"
            combined_output_path                                                                                                                        =
"/scratch/rkersten_root/rkersten1/dousley/Seedling_MzMlData_DO_2025/filter_seedling_combined/AllHighSeedling1iminium.mzML"

            os.makedirs(output_folder, exist_ok=True)

            # === FILTER INDIVIDUAL FILES ===
            for csv_path in glob.glob(os.path.join(csv_folder, "*_results.csv")):
                basename = os.path.basename(csv_path).replace("_results.csv", "")
                mzml_path = os.path.join(mzml_folder, f"{basename}.mzML")
                output_path = os.path.join(output_folder, f"{basename}_filtered.mzML")

                if not os.path.isfile(mzml_path):
                    print(f"⚠️ No MZML file for {basename}, skipping.")
                    continue

                print(f"\n Processing: {basename}")
                print(f" → CSV: {csv_path}")
                print(f" → MZML: {mzml_path}")

                try:
                    df = pd.read_csv(csv_path, sep=',', engine='python')
                    df.columns = [col.strip().lower() for col in df.columns]
                    print(f"✅ Columns in {csv_path}: {df.columns.tolist()}")

                    if 'scan' in df.columns:
                        ms2_scans = set(df['scan'].dropna().astype(int))
                    else:
                        raise ValueError(f"CSV {basename} missing 'scan' column")

                    if 'ms1scan' in df.columns:
                        ms1_scans = set(df['ms1scan'].dropna().astype(int))
                    else:
                        ms1_scans = set()

                    all_scans = ms2_scans.union(ms1_scans)
                    print(f"🪄 MS2 scans: {len(ms2_scans)}, MS1 scans: {len(ms1_scans)}")

                    exp = MSExperiment()
                    MzMLFile().load(mzml_path, exp)

                    filtered_spectra = []
                    for spectrum in exp.getSpectra():
                        native_id = spectrum.getNativeID()
                        try:
                            parts = native_id.strip().split()
                            for part in parts:
                                if part.startswith("scan="):
                                    scan_num = int(part.split('=')[1])
                                    if scan_num in all_scans:
                                        filtered_spectra.append(spectrum)
                                        print(f"✅ Matched scan={scan_num} in native ID: {native_id}")
                                    break
                        except Exception as e:
                            print(f"⚠️ Could not parse scan from: {native_id} → {e}")
```

```
            print(f"✅ Retained {len(filtered_spectra)} out of {len(exp.getSpectra())} spectra")

            new_exp = MSExperiment()
            new_exp.setSpectra(filtered_spectra)
            MzMLFile().store(output_path, new_exp)

            print(f"✅ Written filtered MZML: {output_path}")

        except Exception as e:
            print(f"❌ Failed to process {basename}: {e}")

# === MERGE ALL FILTERED MZML FILES ===
print("\n🔀 Merging all filtered MZML files...")

combined_exp = MSExperiment()
for filtered_path in sorted(glob.glob(os.path.join(output_folder, "*_filtered.mzML"))):
    print(f" Adding: {filtered_path}")
    temp_exp = MSExperiment()
    MzMLFile().load(filtered_path, temp_exp)
    combined_exp.getSpectra().extend(temp_exp.getSpectra())

print(f"✅ Total combined spectra: {len(combined_exp.getSpectra())}")
MzMLFile().store(combined_output_path, combined_exp)
print(f"🎉 Combined MZML written to: {combined_output_path}")
```

5.6.7 Protein Fasta Downloads and RepeatFinder for Core Peptide Discovery

128 available protein FASTA files associated with their genomic sequencing data accessible through the genome section of the NCBI Datasets were downloaded and uploaded to a local Sequenceserver (**Table 5.3**). Searching for BURP-domain protein sequences was performed via Sequenceserver BLAST and subsequent RepeatFinder analysis[11,59]. For Sequenceserver-based BLAST search, all downloaded fastas were searched via BLASTp (BLAST+ v2.16.0+) for homologs of Sali3-2 as a BURP-domain-containing protein query (GenBank ID AAB66369) on Sequenceserver (v3.1.0) with the following BLAST parameters: E-value 1e-05, matrix BLOSUM62, gap-open 11, gap-extend 1, filter L, max, max_target_seqs 5000. BLASTp hits from all databases were combined into a fasta-file. The resulting protein sequence fasta file was analyzed by standalone RepeatFinder on the Great Lakes HPC Cluster at the University of Michigan, Ann Candidate burpitide cyclases were characterized in RepeatFinder output and presence in an open reading frame, which encoded a BURP-domain motif. In our searches, a BURP-domain motif was defined by the presence of at least half the BURP-domain-conserved residues (i.e., FF/Fx-C-C-CH of FF/Fx-C-C-CH-CH-CH-CH)[17]. Hits with a single core peptide at the first N-terminal α-helix of BURP-domain proteins were removed. The predicted [M+H]+ and [M+2H]+ masses for the core peptides mined via RepeatFinder were generated in ChemDraw (Revvity Signals, e.g. version 22.2.0).

5.6.8 Data Availability via MassIVE

LC-MS/MS datasets were deposited into a private GNPS-MassIVE server under accession MSV000099792 [https://massive.ucsd.edu/ProteoSAFe/dataset.jsp?task=6f093b73f0a4435bb1dfc04cc2926981][57].

5.7 Supplemental Information

Table 5.3 Seedling grown in this study, protein FASTA downloaded, and tissues sampled.

Species	Family	Protein Fasta	Root	Leaf	Stem
Abelmoschus Esculentus	*Malvaceae*		✓	✓	✓
Adansonia Digitata	*Malvaceae*		✓	✓	✓
Aegilops speltoides	*Poaceae*		✓	✓	✓
Agrostis gigantea	*Poaceae*		✓		✓
Albizia Julibrissin	*Fabaceae*		✓	✓	✓
Amaranthus albus	*Amaranthaceae*		✓	✓	✓
Amaranthus Blitum	*Amaranthaceae*		✓	✓	✓
Amaranthus Cruentus	*Amaranthaceae*		✓	✓	✓
Amaranthus hybridus	*Amaranthaceae*		✓	✓	✓
Amaranthus Hypochondriacus	*Amaranthaceae*		✓	✓	✓
Amaranthus retroflexus	*Amaranthaceae*		✓	✓	✓
Amaranthus Tricolor	*Amaranthaceae*	✓	✓	✓	✓
Ambrosia Artemisiifolia	*Asteraceae*		✓	✓	✓
Apium Graveolens	*Umbellifers*	✓	✓	✓	✓
Arabidopsis Thaliana	*Brassicaceae*	✓	✓		✓
Arachis Hypogaea	*Fabaceae*	✓	✓	✓	✓
Arachis Ipaensis	*Fabaceae*	✓	✓	✓	✓
Arachis Stenosperma	*Fabaceae*	✓	✓	✓	✓
Arctium Lappa	*Asteraceae*	✓	✓	✓	✓
Artemisia Annua	*Asteraceae*		✓	✓	✓
Artemisia Princeps	*Asteraceae*		✓	✓	✓
Asclepias incarnata	*Apocynaceae*		✓	✓	✓
Asparagus officinalis	*Asparagaceae*	✓	✓	✓	✓
Arabis Alpina	*Brassicaceae*		✓	✓	✓

Avena sativa	*Poaceae*		✓	✓	✓
Barrel medic	*Fabaceae*	✓	✓	✓	✓
Bauhinia variegata	*Fabaceae*	✓	✓	✓	✓
Benincasa Hispida	*Cucurbitaceae*	✓	✓	✓	✓
Beta vulgaris	*Amaranthaceae*		✓	✓	✓
Beta vulgaris subsp. vulgaris	*Amaranthaceae*	✓	✓	✓	✓
Brachypodium Distachyon	*Poaceae*	✓	✓	✓	✓
Brassica Carinata	*Brassicaceae*		✓	✓	✓
Brassica Cretica	*Brassicaceae*	✓	✓	✓	✓
Brassica Juncea	*Brassicaceae*		✓	✓	✓
Brassica Napus	*Brassicaceae*		✓	✓	✓
Brassica Rapa Subsp Triolocularis	*Brassicaceae*	✓	✓	✓	✓
Cajanus Cajan	*Fabaceae*	✓	✓	✓	✓
Camelina Sativa	*Brassicaceae*	✓	✓	✓	✓
Canavalia Gladiata	*Fabaceae*	✓	✓	✓	✓
Cannabis Sativa	*Cannabaceae*	✓	✓	✓	✓
Capsicum Annuum	*Solanaceae*	✓	✓	✓	✓
Capsicum Baccatum	*Solanaceae*		✓	✓	✓
Capsicum Chinense	*Solanaceae*		✓	✓	✓
Carica Papaya	*Caricaceae*	✓	✓	✓	✓
Catharanthus Roseus	*Apocynaceae*		✓	✓	✓
Cheiranthus Cheiri	*Brassicaceae*		✓	✓	✓
Chenopodium Quinoa	*Amaranthaceae*		✓	✓	✓
Cicer arietinum	*Fabaceae*	✓	✓	✓	✓
Cichorium Endivia	*Asteraceae*	✓	✓	✓	✓
Cichorium Intybus	*Asteraceae*	✓	✓	✓	✓
Citrullus Colocynthis	*Cucurbitaceae*		✓	✓	✓
Clitoria Ternatea	*Fabaceae*		✓	✓	✓
Corchorus Capsularis	*Malvaceae*	✓	✓	✓	✓
Corchorus Olitorius	*Malvaceae*	✓	✓	✓	✓
Cornus Florida	*Cornaceae*	✓	✓	✓	✓
Crambe hispanica subsp. abyssinica	*Brassicaceae*		✓	✓	✓
Crotalaria Pallida	*Fabaceae*	✓	✓	✓	✓

Cryptomeria Japonica	*Cupressaceae*	✓	✓	✓	✓
Cucumis melo	*Cucurbitaceae*	✓	✓	✓	✓
Cucumis Sativus	*Cucurbitaceae*	✓	✓	✓	✓
Cucurbita Argyrosperma	*Cucurbitaceae*	✓	✓	✓	✓
Cucurbita maxima	*Cucurbitaceae*	✓	✓	✓	✓
Cucurbita Moschata	*Cucurbitaceae*	✓	✓	✓	✓
Cucurbita pepo	*Cucurbitaceae*	✓	✓	✓	✓
Cynara Cardunculus Var. Scolymus	*Asteraceae*	✓	✓	✓	✓
Dactylis Glomerata	*Poaceae*		✓	✓	✓
Datura Stramonium	*Solanaceae*	✓	✓	✓	✓
Daucus Carota	*Apiaceae*	✓	✓	✓	✓
Daucus Carota Var. Sativus	*Umbellifers*	✓	✓	✓	✓
Dianthus Caryophyllus	*Caryophyllaceae*		✓	✓	✓
Digitaria exilis	*Poaceae*	✓	✓	✓	✓
Dovyalis Caffra	*Salicaceae*	✓	✓	✓	✓
Aristolochia Californica	*Aristolochiaceae*	✓	✓	✓	✓
echinochloa crus-galli	*Poaceae*			✓	✓
Ensete Ventricosum	*Musaceae*	✓	✓	✓	✓
Eragrostis Curvula	*Poaceae*	✓	✓	✓	✓
Erigeron Canadensis	*Asteraceae*	✓	✓	✓	✓
Eruca vesicaria subsp. sativa.	*Brassicaceae*	✓		✓	✓
Erythanthe Guttata	*Phrymaceae*	✓	✓	✓	✓
Eschscholzia californica	*Papaveraceae*		✓	✓	✓
Euphorbia Lathyris	*Euphorbiaceae*	✓	✓	✓	✓
Fagopyrum esculentum	*Polygonaceae*		✓	✓	✓
Ficus carica	*Moraceae*		✓	✓	✓
Fragaria Vesca	*Rosaceae*	✓	✓	✓	✓
glycine Max	*Fabaceae*	✓	✓	✓	✓
Glycine Soja	*Fabaceae*	✓	✓	✓	✓
gossypium arboreum L.	*Malvaceae*	✓	✓	✓	✓
Gossypium Armourianum Kearney	*Malvaceae*	✓	✓	✓	✓
Gossypium Barbadense	*Malvaceae*		✓	✓	✓
Gossypium Hirsutum L.	*Malvaceae*	✓	✓	✓	✓

Gossypium lobatum	*Malvaceae*	✓	✓	✓	✓
Gossypium Mustelinum Miers ex G. Watt	*Malvaceae*		✓	✓	✓
Gossypium Raimondii Ulbr.	*Malvaceae*	✓	✓	✓	✓
Gossypium Tomentosum Nutt. ex Seem	*Malvaceae*		✓	✓	✓
gypsophila vaccaria	*Caryophyllaceae*	✓	✓	✓	✓
Helianthus annuus	*Asteraceae*	✓	✓	✓	✓
Heuchera sanguinea	*Saxifragaceae*		✓		✓
Hibiscus Cannabinus	*Malvaceae*	✓	✓	✓	✓
Hibiscus sabdariffa	*Malvaceae*	✓	✓	✓	✓
Hibiscus Syriacus	*Malvaceae*	✓	✓	✓	✓
hondroanthus impetiginosus	*Lamiales*	✓	✓	✓	✓
Hordeum Vulgare	*Poaceae*	✓	✓	✓	✓
Humulus Lupulus	*Cannabaceae*	✓	✓	✓	✓
Humulus Lupulus Var. Neomexicanus	*Cannabaceae*		✓	✓	✓
Ipomoea Nil	*Convolvulaceae*	✓	✓	✓	
Ipomoea purpurea	*Convolvulaceae*		✓	✓	✓
Ipomoea tricolor	*Convolvulaceae*		✓	✓	✓
Iris Sibirica	*Iridaceae*			✓	✓
Lablab Purpureus	*Fabaceae*		✓	✓	✓
Lactuca Saligna	*Asteraceae*		✓	✓	✓
Lactuca Sativa	*Asteraceae*	✓	✓	✓	✓
Lactuca Virosa	*Asteraceae*		✓	✓	✓
Lens Culinaris	*Fabaceae*		✓	✓	✓
Linum Trigynum	*Linaceae*	✓	✓	✓	
Liquidambar Formosana	*Altingiaceae*	✓	✓	✓	✓
Lolium Multiflorum	*Poaceae*	✓	✓	✓	✓
Lolium Perenne	*Poaceae*	✓	✓	✓	✓
Lolium rigidum	*Poaceae*	✓	✓	✓	✓
lotus tetragonolobus (Psophocarpus tetragonolobus)	*Fabaceae*	✓	✓	✓	✓
Luffa Cylindrica (Luffa aegyptiaca)	*Cucurbitaceae*		✓	✓	✓
Lupinus Albus Var Albus	*Fabaceae*	✓	✓	✓	✓
Lycium barbarum	*Solanaceae*	✓	✓	✓	✓
medicago sativa subsp falcata	*Fabaceae*		✓	✓	✓

Mercurialis Annua	*Euphorbiaceae*	✓	✓	✓	✓
Mirabilis Jalapa	*Nyctaginaceae*		✓	✓	✓
Momordica Charantia	*Cucurbitaceae*	✓	✓	✓	✓
Nemophila menziesii	*Boraginaceae*		✓	✓	✓
nicotiana alata	*Solanaceae*		✓	✓	✓
Nicotiana Megalosiphon	*Solanaceae*		✓	✓	✓
Nicotiana occidentalis	*Solanaceae*		✓	✓	✓
Nicotiana Sylvestris	*Solanaceae*	✓	✓	✓	✓
nicotiana tabacum	*Solanaceae*	✓	✓	✓	✓
Nicotiana Tomentosiformis	*Solanaceae*	✓	✓	✓	✓
Nigella Sativa	*Ranunculaceae*		✓	✓	✓
Oenothera Lamarckiana	*Onagraceae*		✓	✓	✓
oryza barthii	*Poaceae*		✓	✓	✓
oryza glaberrima	*Poaceae*	✓	✓	✓	✓
oryza sativa	*Poaceae*	✓	✓	✓	✓
Pachyrhizus Erosus	*Fabaceae*		✓	✓	✓
Panicum Hallii (Hall's Panicgrass)	*Poaceae*	✓	✓	✓	
Panicum Violaceum	*Poaceae*		✓	✓	✓
Panicum Virgatum	*Poaceae*	✓	✓	✓	✓
Papaver Armeniacum	*Papaveraceae*	✓	✓	✓	✓
Papaver atlanticum flore pleno	*PAPAVERACEAE*	✓	✓	✓	✓
Papaver Nudicaule	*Papaveraceae*	✓	✓	✓	✓
Papaver rhoeas	*PAPAVERACEAE*		✓	✓	✓
Papaver Somniferum	*Papaveraceae*	✓	✓	✓	✓
Paspalum Notatum	*Poaceae*		✓	✓	✓
Penstemon eatonii	*Plantaginaceae*		✓	✓	✓
Perilla Frutescens Var, Frutescens	*Lamiaceae*	✓	✓	✓	✓
Petroselinum Crispum	*Apiaceae*		✓	✓	✓
Petunia Hybrida	*Solanaceae*		✓	✓	✓
Phaseolus Coccineus	*Fabaceae*	✓	✓	✓	✓
Phaseolus Vulgaris	*Fabaceae*	✓	✓	✓	✓
Phragmites	*Poaceae*	✓	✓	✓	
pisum sativum (Lathyrus oleraceus)	*Fabaceae*	✓	✓	✓	✓

Portulaca Oleracea	*Portulacaceae*		✓		✓
Portulaca oleracea subsp. sativa	*Portulacaceae*		✓	✓	✓
Prunus Dulcis	*Rosaceae*	✓		✓	✓
Quercus Rubra	*Fagaceae*	✓	✓	✓	✓
Quercus Suber	*Fagaceae*	✓	✓	✓	✓
raphanus sativus	*Brassicaceae*	✓	✓	✓	✓
Ricinus Communis	*Euphorbiaceae*	✓	✓	✓	✓
Rosmarinus officinalis	*Lamiaceae*		✓	✓	✓
Rudbeckia Hirta	*Asteraceae*		✓	✓	✓
S. Splendens	*Lamiaceae*	✓	✓	✓	✓
Salvia hispanica	*Lamiaceae*	✓	✓	✓	✓
Salvia Miltiorrhiza	*Lamiaceae*	✓	✓	✓	✓
Sequoiadendron giganteum	*Cupressaceae*		✓	✓	✓
Sesamum Indicum	*Pedaliaceae*	✓	✓	✓	✓
Sesamum Radiatum	*Pedaliaceae*	✓	✓	✓	✓
Sesbania Bispinosa	*Fabaceae*	✓	✓	✓	✓
Setaria Italica	*Poaceae*	✓	✓	✓	✓
Setaria Viridis	*Poaceae*	✓	✓	✓	✓
sinapis alba subsp. mairei	*Brassicaceae*	✓	✓	✓	✓
Solanum Chilense	*Solanaceae*		✓	✓	✓
Solanum Dulcamara	*Solanaceae*	✓	✓	✓	✓
Solanum Lycopersicum	*Solanaceae*	✓	✓	✓	✓
Solanum Pennellii	*Solanaceae*	✓	✓	✓	✓
Sorghum bicolor	*Poaceae*	✓	✓	✓	✓
Sorghum bicolor L. Moench	*Poaceae*	✓	✓	✓	✓
Spinacia oleracea	*Amaranthaceae*		✓	✓	✓
Steinchisma Laxum	*POACEAE*		✓	✓	
Stevia rebaudiana	*Asteraceae*	✓	✓	✓	
Tagetes Erecta	*Asteraceae*		✓	✓	✓
Talinum Paniculatum	*Talinaceae*	✓	✓	✓	✓
Tanacetum Coccineum	*Asteraceae*	✓	✓	✓	✓
Trifolium medium	*Fabaceae*	✓	✓	✓	✓
Trifolium pratense	*Fabaceae*	✓	✓	✓	✓

Trifolium repens	*Fabaceae*		✓	✓	✓
Trifolium Subterraneum	*Fabaceae*	✓	✓	✓	✓
triticum aestivum	*Poaceae*		✓	✓	✓
Triticum turgidum	*Poaceae*	✓	✓	✓	✓
Triticum Urartu	*Poaceae*	✓	✓	✓	✓
Turnera Subulata	*Turneraceae*	✓	✓	✓	✓
Vicia Faba	*Fabaceae*	✓	✓	✓	✓
Vicia Villosa	*Fabaceae*	✓	✓	✓	✓
Vigna Angularis	*Fabaceae*		✓	✓	✓
Vigna Angularis Var Nipponensis	*Fabaceae*	✓	✓	✓	✓
Vigna Mungo	*Fabaceae*	✓	✓	✓	✓
Vigna Radiata	*Fabaceae*	✓	✓	✓	✓
Vigna Umbellata	*Fabaceae*	✓	✓	✓	✓
Vigna Unguiculata	*Fabaceae*	✓	✓	✓	✓
Zea Mays	*Poaceae*		✓	✓	✓

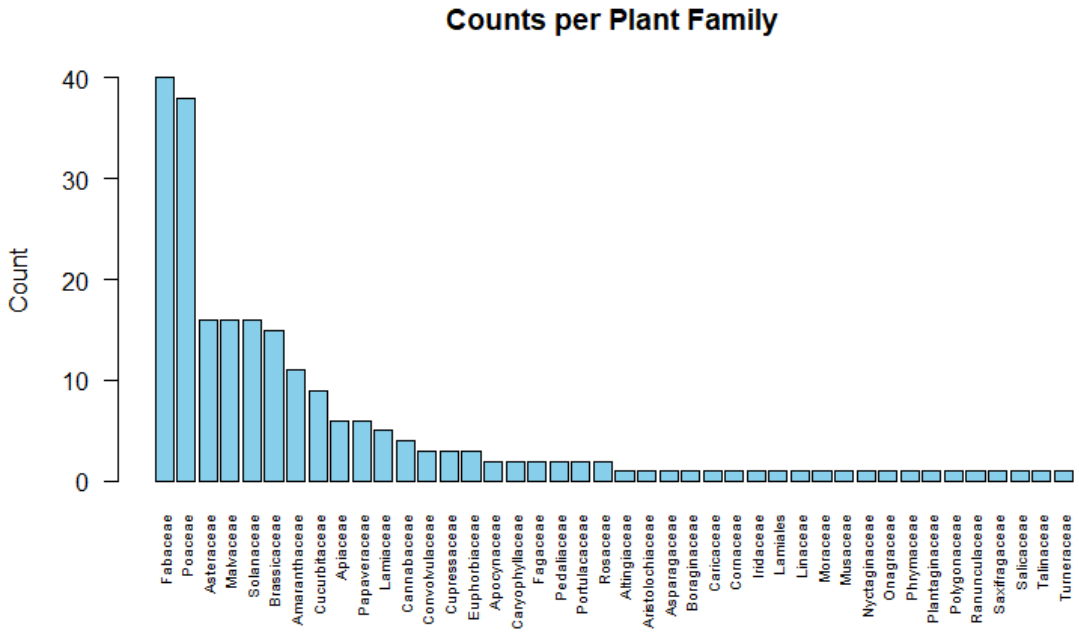


Figure 5.7 Bar graph distribution of plant families.

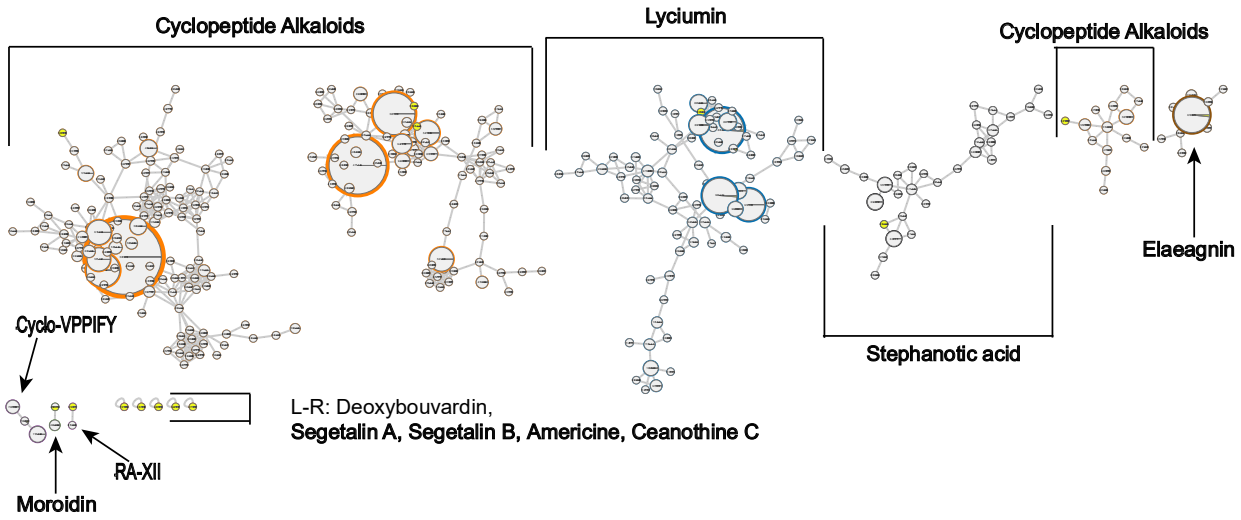


Figure 5.8 Burpitide subset of molecular network of MassQL-processed LC-MS-datasets with cardinality command of 1-5 different amino acid iminium ions in a given tandem MS spectrum.

Reference peptides are highlighted in yellow. Known burpitides are denoted above.

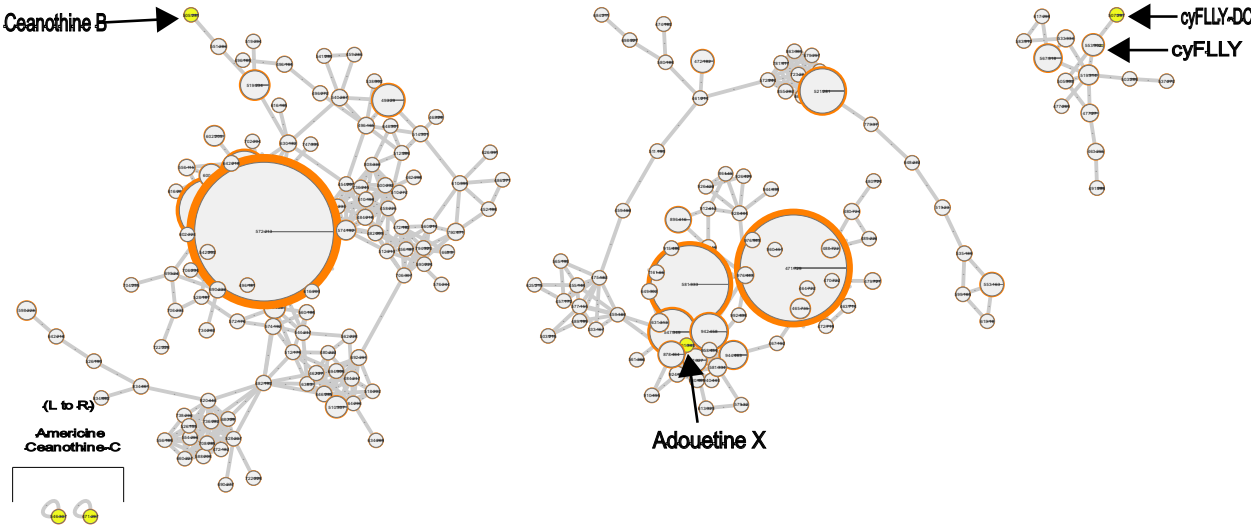


Figure 5.9 Cyclopeptide Alkaloid subset of molecular network of MassQL-processed LC-MS-datasets with cardinality command of 1-5 different amino acid iminium ions in a given tandem MS spectrum.
Reference peptides are highlighted in yellow. Known cyclopeptide alkaloids are denoted above. Note, Sanjoinine A is clustered above with Adouetine X, but hidden behind larger nodes.

5.8 References

1. Walsh, C. T. & Tang, Y. *Natural Product Biosynthesis: Chemical Logic and Enzymatic Machinery*. (The Royal Society of Chemistry, 2017). doi:10.1039/9781788016612.
2. McChesney, J. D., Venkataraman, S. K. & Henri, J. T. Plant natural products: Back to the future or into extinction? *Phytochemistry* **68**, 2015–2022 (2007).

3. Van Herwerden, E. F. & Süssmuth, R. D. Sources for Leads: Natural Products and Libraries. in *New Approaches to Drug Discovery* (eds Nielsch, U., Fuhrmann, U. & Jaroch, S.) vol. 232 91–123 (Springer International Publishing, Cham, 2015).
4. Chekan, J. R., Mydy, L. S., Pasquale, M. A. & Kersten, R. D. Plant peptides – redefining an area of ribosomally synthesized and post-translationally modified peptides. *Nat. Prod. Rep.* **41**, 1020–1059 (2024).
5. Arnison, P. G. *et al.* Ribosomally synthesized and post-translationally modified peptide natural products: overview and recommendations for a universal nomenclature. *Nat Prod Rep* **30**, 108–160 (2013).
6. Montalbán-López, M. *et al.* New developments in RiPP discovery, enzymology and engineering. *Nat. Prod. Rep.* **38**, 130–239 (2021).
7. Kersten, R. D. & Weng, J.-K. Gene-guided discovery and engineering of branched cyclic peptides in plants. *Proc. Natl. Acad. Sci.* **115**, (2018).
8. De Veer, S. J., Kan, M.-W. & Craik, D. J. Cyclotides: From Structure to Function. *Chem. Rev.* **119**, 12375–12421 (2019).
9. Lee, Y.-Y. *et al.* HypoRiPPAtlas as an Atlas of hypothetical natural products for mass spectrometry database search. *Nat. Commun.* **14**, 4219 (2023).
10. Wang, X. *et al.* Large-scale transcriptome mining enables macrocyclic diversification and improved bioactivity of the stephanotic acid scaffold. *Nat. Commun.* **16**, 4198 (2025).
11. Chigumba, D. N. *et al.* Discovery and biosynthesis of cyclic plant peptides via autocatalytic cyclases. *Nat. Chem. Biol.* **18**, 18–28 (2022).
12. Kersten, R. D. *et al.* Gene-Guided Discovery and Ribosomal Biosynthesis of Moroidin Peptides. *J. Am. Chem. Soc.* **144**, 7686–7692 (2022).
13. One Thousand Plant Transcriptomes Initiative. One thousand plant transcriptomes and the phylogenomics of green plants. *Nature* **574**, 679–685 (2019).
14. Mydy, L. S., Chigumba, D. N. & Kersten, R. D. Plant Copper Metalloenzymes As Prospects for New Metabolism Involving Aromatic Compounds. *Front. Plant Sci.* **12**, 692108 (2021).
15. Mydy, L. S. *et al.* An intramolecular macrocyclase in plant ribosomal peptide biosynthesis. *Nat. Chem. Biol.* **20**, 530–540 (2024).
16. Lima, S. T. *et al.* A Widely Distributed Biosynthetic Cassette Is Responsible for Diverse Plant Side Chain Cross-Linked Cyclopeptides**. *Angew. Chem. Int. Ed.* **62**, e202218082 (2023).
17. Hattori, J., Boutilier, K. A., Campagne, M. M. L. & Miki, B. L. A conserved BURP domain defines a novel group of plant proteins with unusual primary structures. *Mol. Gen. Genet. MGG* **259**, 424–428 (1998).
18. Teerawanichpan, P., Xia, Q., Caldwell, S. J., Datla, R. & Selvaraj, G. Protein storage vacuoles of Brassica napus zygotic embryos accumulate a BURP domain protein and perturbation of its production distorts the PSV. *Plant Mol. Biol.* **71**, 331–343 (2009).
19. Van Son, L. *et al.* The BURP domain protein AtUSPL1 of Arabidopsis thaliana is destined to the protein storage vacuoles and overexpression of the cognate gene distorts seed development. *Plant Mol. Biol.* **71**, 319–329 (2009).
20. BäUmlein, H. *et al.* A novel seed protein gene from Vicia faba is developmentally regulated in transgenic tobacco and Arabidopsis plants. *Mol. Gen. Genet. MGG* **225**, 459–467 (1991).

21. Watson, C. F., Zheng, L. & DellaPenna, D. Reduction of tomato polygalacturonase beta subunit expression affects pectin solubilization and degradation during fruit ripening. *Plant Cell* **6**, 1623–1634 (1994).
22. Yamaguchi-Shinozaki, K. & Shinozaki, K. The plant hormone abscisic acid mediates the drought-induced expression but not the seed-specific expression of rd22, a gene responsive to dehydration stress in Arabidopsis thaliana. *Mol. Gen. Genet. MGG* **238–238**, 17–25 (1993).
23. Phillips, K. & Ludidi, N. Drought and exogenous abscisic acid alter hydrogen peroxide accumulation and differentially regulate the expression of two maize RD22-like genes. *Sci. Rep.* **7**, 8821 (2017).
24. Yahara, S. *et al.* Structures of anti-ace and -renin peptides from lycii radicis cortex. *Tetrahedron Lett.* **30**, 6041–6042 (1989).
25. Morita, H., Yoshida, N., Takeya, K., Itokawa, H. & Shirota, O. Configurational and conformational analyses of a cyclic octapeptide, lyciumin A, from Lycium chinense Mill. *Tetrahedron* **52**, 2795–2802 (1996).
26. Yahara, S. *et al.* Cyclic Peptides, Acyclic Diterpene Glycosides and Other Compounds from Lycium chinense MILL. *Chem. Pharm. Bull. (Tokyo)* **41**, 703–709 (1993).
27. Morita, H., Suzuki, H. & Kobayashi, J. Celogenamide A, a New Cyclic Peptide from the Seeds of *Celosia a rgentea*. *J. Nat. Prod.* **67**, 1628–1630 (2004).
28. Hungerford, J. *et al.* Biosynthesis of 14-Membered Cyclopeptide Alkaloids via Nonheme Iron- and 2-Oxoglutarate-Dependent Oxidative Decarboxylation. *J. Am. Chem. Soc.* **148**, 7983–8001 (2026).
29. Kang, K. B. *et al.* Jubanines F–J, cyclopeptide alkaloids from the roots of Ziziphus jujuba. *Phytochemistry* **119**, 90–95 (2015).
30. Lomchoey, N. *et al.* New bioactive cyclopeptide alkaloids with rare terminal unit from the root bark of *Ziziphus cambodiana*. *RSC Adv.* **8**, 18204–18215 (2018).
31. Han, H. *et al.* Anxiolytic-like effects of sanjoinine A isolated from Zizyphi Spinosi Semen: Possible involvement of GABAergic transmission. *Pharmacol. Biochem. Behav.* **92**, 206–213 (2009).
32. Ma, Y. *et al.* Sanjoinine A Isolated from Zizyphi Spinosi Semen Augments Pentobarbital-Induced Sleeping Behaviors through the Modification of GABA-ergic Systems. *Biol. Pharm. Bull.* **30**, 1748–1753 (2007).
33. Trevisan, G. *et al.* Antinociceptive Effects of 14-Membered Cyclopeptide Alkaloids. *J. Nat. Prod.* **72**, 608–612 (2009).
34. Kaleem, W. A. *et al.* Antinociceptive activity of cyclopeptide alkaloids isolated from Ziziphus oxyphylla Edgew (Rhamnaceae). *Fitoterapia* **91**, 154–158 (2013).
35. Hwang, K. H., Han, Y. N. & Han, B. H. Inhibition of calmodulin-dependent Calcium-ATPase and phosphodiesterase by various cyclopeptides and peptide alkaloids from the zizyphus species. *Arch. Pharm. Res.* **24**, 202–206 (2001).
36. Tuenter, E. *et al.* Antiplasmodial Activity, Cytotoxicity and Structure-Activity Relationship Study of Cyclopeptide Alkaloids. *Molecules* **22**, 224 (2017).
37. Tuenter, E. *et al.* Cyclopeptide Alkaloids from *Hymenocardia acida*. *J. Nat. Prod.* **79**, 1746–1751 (2016).
38. Sugawara, F. *et al.* Insecticidal Peptide from Mungbean: A Resistant Factor against Infestation with Azuki Bean Weevil. *J. Agric. Food Chem.* **44**, 3360–3364 (1996).

39. Liu, W. *et al.* Genome and transcriptome of Selaginella kraussiana reveal evolution of root apical meristems in vascular plants. *Curr. Biol.* **33**, 4085-4097.e5 (2023).
40. Yoshikawa, K., Tao, S. & Arihara, S. Stephanotic Acid, a Novel Cyclic Pentapeptide from the Stem of *Stephanotis f loribunda*. *J. Nat. Prod.* **63**, 540–542 (2000).
41. Christina Leung, T.-W., Williams, D. H., C J Barna, J., Foti, S. & B.Oelrichs, P. Structural studies on the peptide moroidin from laportea moroides. *Tetrahedron* **42**, 3333–3348 (1986).
42. Xu, X. *et al.* Moroidin, a Cyclopeptide from the Seeds of *Celosia cristata* That Induces Apoptosis in A549 Human Lung Cancer Cells. *J. Nat. Prod.* **85**, 1918–1927 (2022).
43. Suzuki, H., Morita, H., Shiro, M. & Kobayashi, J. Celogentin K, a new cyclic peptide from the seeds of Celosia argentea and X-ray structure of moroidin. *Tetrahedron* **60**, 2489–2495 (2004).
44. Kobayashi, J., Suzuki, H., Shimbo, K., Takeya, K. & Morita, H. Celogentins A−C, New Antimitotic Bicyclic Peptides from the Seeds of *Celosia a rgentea*. *J. Org. Chem.* **66**, 6626–6633 (2001).
45. Suzuki, H., Morita, H., Iwasaki, S. & Kobayashi, J. New antimitotic bicyclic peptides, celogentins D–H, and J, from the seeds of Celosia argentea. *Tetrahedron* **59**, 5307–5315 (2003).
46. Ousley, D. A., Wang, X., Mydy, L. S. & Kersten, R. D. Plant genome mining of burpitides from fused precursor peptides. in *Methods in Enzymology* S0076687925001958 (Elsevier, 2025). doi:10.1016/bs.mie.2025.04.007.
47. Gomes, P. W. P. *et al.* plantMASST - Community-driven chemotaxonomic digitization of plants. Preprint at https://doi.org/10.1101/2024.05.13.593988 (2024).
48. Damiani, T. *et al.* A universal language for finding mass spectrometry data patterns. *Nat. Methods* **22**, 1247–1254 (2025).
49. Röst, H. L. *et al.* OpenMS: a flexible open-source software platform for mass spectrometry data analysis. *Nat. Methods* **13**, 741–748 (2016).
50. Yu, J. S. *et al.* A versatile toolkit for drug metabolism studies with GNPS2: from drug development to clinical monitoring. *Nat. Protoc.* https://doi.org/10.1038/s41596-025-01237-6 (2025) doi:10.1038/s41596-025-01237-6.
51. Wang, M. *et al.* Sharing and community curation of mass spectrometry data with Global Natural Products Social Molecular Networking. *Nat. Biotechnol.* **34**, 828–837 (2016).
52. Liu, A., Contador, C. A., Fan, K. & Lam, H.-M. Interaction and Regulation of Carbon, Nitrogen, and Phosphorus Metabolisms in Root Nodules of Legumes. *Front. Plant Sci.* **9**, 1860 (2018).
53. Ji, Y., Hewavithana, T., Sharpe, A. G. & Jin, L. Understanding grain development in the Poaceae family by comparing conserved and distinctive pathways through omics studies in wheat and maize. *Front. Plant Sci.* **15**, 1393140 (2024).
54. The Nightshade Plant Family.
55. Murat, F. *et al.* Understanding Brassicaceae evolution through ancestral genome reconstruction. *Genome Biol.* **16**, 262 (2015).
56. Walker, C. C. Malvaceae. in *Dicotyledons: Rosids* (eds Eggli, U. & Nyffeler, R.) 801–806 (Springer International Publishing, Cham, 2023). doi:10.1007/978-3-030-93492-7_74.
57. Choi, M. *et al.* MassIVE.quant: a community resource of quantitative mass spectrometry-based proteomics datasets. *Nat. Methods* **17**, 981–984 (2020).

58. Tang, Y. *et al.* Expression of a Vacuole-Localized BURP-Domain Protein from Soybean (SALI3-2) Enhances Tolerance to Cadmium and Copper Stresses. *PLoS ONE* **9**, e98830 (2014).
59. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410 (1990).
60. Shannon, P. *et al.* Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks. *Genome Res.* **13**, 2498–2504 (2003).
61. Kessner, D., Chambers, M., Burke, R., Agus, D. & Mallick, P. ProteoWizard: open source software for rapid proteomics tools development. *Bioinformatics* **24**, 2534–2536 (2008).

Chapter 6 Dissertation Summary and Conclusion

6.1 Dissertation Summary

This dissertation provides a comprehensive overview of transcriptome mining and multi-omics guided discovery strategies designed to streamline the computational identification of macrocyclic plant peptide natural products. The projects presented collectively integrate transcriptomic, genomic, and metabolic data to develop workflows that enable rational and scalable drug discovery of plant peptides.

Chapter 1 introduces plant natural products and examines the potential of gene-guided discovery of burpitide ribosomally synthesized and post-translationally modified peptides (RiPPs). The chapter contextualizes the limited exploration of plants for medicinal purposes, despite their vast chemical diversity in secondary metabolites. Industrial trends favoring synthetic combinatorial chemistry over natural product research have created significant gaps in drug discovery pipelines, even as natural products and their derivatives remain central to treatments for cancer and infectious diseases. The chapter further highlights burpitides as a highly promising, undervalued, and underexplored class of plant RiPPs with significant potential for bioactive peptide discovery.

Burpitides possess unique aromatic cross-links (C-N, C-O, and C-C bonds) formed by copper-dependent BURP-domain cyclases, which render their drug-like scaffolds accessible through plant pathway bioengineering. The chapter addresses challenges in mining RiPP natural products from plants, such as large genome sizes and non-clustered biosynthetic genes. To overcome these obstacles, the chapter provides a detailed explanation of genome mining, using fused burpitide precursor peptides as a case study, and outlines a systematic approach for the identification, prediction, detection, and structural elucidation of burpitides. The latter portion of

Chapter 1 is reproduced from Ousley DA *et al.* Genome Mining of Plant Peptides. Methods in Enzymology 717, 141-174 (2025).

Chapter 2 is reformatted, reprinted, and expanded from Wang, X., Shafiq, K., Ousley, D.A. *et al.* Large-scale transcriptome mining enables macrocyclic diversification and improved bioactivity of the stephanotic acid scaffold. *Nat Commun* 16, 4198 (2025). This chapter focuses on benchmarking and computational assessment of three *de novo* short-read assemblers (Trinity, SPAdes, and MEGAHIT) for generating precursor peptides from RNA-seq datasets of RiPP-producing plants. The assemblers were evaluated for computational efficiency in assembling plant transcriptomes to discover known RiPPs, using metrics including time, memory usage, total and unique core detection, repetitive core detection, and sequence similarity and identity scores.

SPAdes achieved an effective balance between computational performance and scalability by assembling sequences that produced more predicted peptides at scale, with higher similarity and identity scores, though it required more time and memory than the more memory-efficient MEGAHIT. Additionally, assembled transcriptomes offered greater utility for discovery compared to querying raw unassembled sequence reads. Assembly converted the data into coherent full-length gene sequences, facilitating more efficient bioinformatic extraction of precursor peptides.

Utilizing SPAdes for the *de novo* assembly strategy, large-scale mining of fused moroidin cyclases was conducted using thousands of publicly available plant transcriptome datasets from the NCBI Sequence Read Archive (SRA). This approach generated a transcriptome database of 27,224 non-redundant sequences, which were bioinformatically screened for BURP-domain sequences containing one or more internal stephanotic acid core peptide motifs. The chapter demonstrates the effectiveness of transcriptomics-guided approaches combined with underutilized plant genetic databases, such as NCBI SRA, for RiPP discovery.

Chapter 3 explores all of the bioinformatically predicted core peptide sequences from Chapter 2, generated using SPAdes, and integrates them into a computational drug screening pipeline. The chapter establishes a proof-of-concept *in silico* library of theoretical burpitide structures for rational drug design. Chemoinformatics analyses verified the drug potential of the predicted RiPP structures using molecular descriptors and principal component analysis (PCA). Docking simulations with the *in silico* RiPP library were used to predict the potential bioactivity of the peptides against the validated antimicrobial target, lipoprotein signal peptidase II (LspA).

The primary contribution of this chapter is the demonstration of a workflow that transitions from raw sequence data to theoretical structures, followed by bio-functional predictions. Additionally, the chapter illustrates methods for prioritizing promising candidates for subsequent bioengineering and production in tobacco. In silico libraries are presented as a rapid and effective approach for screening drug-like RiPPs and for supporting downstream experimental prioritization of proposed peptide bioactives.

Chapter 4 is reformatted, reprinted, and expanded from Wang, X., Shafiq, K., Ousley, D.A. *et al.* Large-scale transcriptome mining enables macrocyclic diversification and improved bioactivity of the stephanotic acid scaffold. *Nat Commun* 16, 4198 (2025). This chapter builds upon the work in Chapter 2 by developing an efficient large-scale transcriptome mining strategy and applying it to discover and diversify post-translational modifications within the stephanotic acid scaffold. The results are significant for characterizing novel stephanotic acid-type burpitides featuring unique Trp-Val and Tyr-Phe cross-linked second rings, which have not been previously observed in plants.

From a plant therapeutic discovery perspective, this chapter not only characterizes novel moroidin burpitides with enhanced bioactivity but also establishes a broader chemotaxonomic distribution of these burpitides in pinks and grasses. Bioassays in H1437 human lung cancer cell lines demonstrated that the moroidin analog exhibits greater cytotoxicity than the original moroidin. Preliminary investigations into the mechanism of action of the moroidin analog were initiated, including the use of an AI tool to synthesize and analyze data on potential moroidin drug targets. Collectively, the chapter exemplifies the translational potential of gene-guided discovery strategies in plants for generating potent candidates for cancer therapeutics development.

Chapter 5 presents a project currently in progress, with the goal of future publication upon completion. This exploratory project aims to validate a multi-omics discovery strategy in plants. The approach integrates analytical chemistry, LC-MS/MS-based metabolomics, and plant genomics to expand burpitide discovery. Over 200 phylogenetically diverse species were cultivated in the Kersten Lab for natural product extraction and analyzed to correlate gene presence with metabolite production in roots, leaves, and stems.

The resulting multi-omics library serves as a valuable resource that connects genetic potential to the accessible natural chemical space in plants. Refinement of the pipeline, expansion of the plant cohort, and further exploration of molecular network classes are expected to enhance

its predictive power. Ultimately, the development of this pipeline aims to bridge the gap between bioinformatic gene prediction and experimental natural product analytical chemistry for RiPP discovery.

Taken together, the chapters in the thesis work to bridge large-scale genomic sequencing with the prediction of structural candidates for peptide drug discovery in plants. Chapter 2 provides a strong foundation and starting point for the dissertation. Benchmarking of *de novo* assemblers justifies using assembled data as the starting point for discovering plant RiPPs, rather than attempting to interrogate noisy unassembled data. Furthermore, the benchmarking justified the use of the de novo assembly strategy, as most plants are non-model organisms and do not have a quality reference genome, which limited the use of gene-guided assemblers such as StringTie, even though it outperformed the de novo assembly strategy in all areas except memory consumption, unique core peptide identification, and the identity and similarity scores of assembled precursor peptides.

These categories are crucial for plant RiPP discovery because they ensure that novel peptides can be found from the data, helping to limit redundant discovery of previously known core peptides and identify similarities. Furthermore, it ensures that the precursor sequences are assembled to be as complete as possible, helping to ensure no repetitive encoded sequencing of the core peptide regions is missed, which can limit discovery. Thus, the SPAdes *de novo* assembly strategy enabled high-fidelity assembly of short-read RNA-seq data into coherent precursor peptide sequences for downstream mining and database interrogation to identify unique core peptides.

The computational rigor of Chapter 2 supports the assembly of a non-redundant database of 27,224 sequences, which directly informs the work in Chapter 3 to build the *in silico* library of plant cyclic peptides. The non-redundant database served as the basis for discovery, enabling the conversion of theoretical sequencing data into computational probes for drug discovery and genetic bioengineering in tobacco. The computational bridge here is the transition from sequencing data containing core peptides to 1D SMILES, 2D structures, and 3D models of those peptides generated via physics-based simulations.

The docking simulations against LspA served to both computationally validate the proof-of-concept *in silico* library and prioritize core peptide sequences for interrogation of biosynthetically feasible production. Chapter 3 also helps lay the groundwork for bridging the computational discovery of these natural products with the characterization of the peptides

analytically and biosynthetically. Thus, this shows that the sequences from chapter 2 are more than just prediction vehicles for theoretical peptides and computational probes; they can be accessed in wet-lab-based tobacco production without the need for chemical synthesis, which is difficult for these peptides given their macrocyclic crosslinks. While the failure to fully purify and isolate lyciumin peptides from large-scale infiltration of tobacco leaves limits this portion of the work to strictly experimental proof-of-concept, analysis of small-scale infiltration for heterologous production of these peptides aided decision-making on which peptides are worth the time and cost of bioengineering in tobacco.

The transition from computational-based discoveries to analytical characterization, peptide purification, biosynthetically feasible production, and natural product isolation is fully realized in Chapter 4, which is a direct extension of the large-scale transcriptome mining work completed in Chapter 2. In chapter 2, the assembly of a non-redundant database of 27,224 sequences yielded sequences from 17 plant species with 46 unique moroidin core peptides that were connected directly to a cyclase domain, ultimately aiding in the phylogeny expansion of these specialized macrocyclic plant peptides within the plant kingdom and *in vitro* interrogation of isolatable peptides for their medicinal value in chapter 4. Thus, the stephanotic acid scaffolds identified during mining are confirmed to exist in nature and are a feasible candidate for biosynthetic production and analytical structural validation.

This is supported by the discovery of novel Trp-Val and Tyr-Phe aromatic cross-links biosynthesized by the plant enzymatic machinery and by improved cytotoxicity in H1437 lung cancer cells. In essence, an Assembly-Predict-Biosynthesize-Test-Analyze cycle is established in this work, similar to the Design-Make-Test-Analyze cycle used in drug discovery and optimization campaigns. NCBI SRA interrogation led to a potent candidate that could be potentially translated for future clinical interrogation, given the peptide's low cytotoxicity in non-cancerous fibroblast cells.

Chapter 5 preliminarily builds on the unified predictive system created in previous chapters, especially the mining workflow and the ability to characterize individual burpitides analytically. It broadens the integration into a multi-omics workflow to interrogate a cohort of 200+ seedling species. This chapter correlated the genetic presence of metabolites with their specialized production in specific plant species and tissues, thus closing the loop on accurate prediction and accessibility of the plant peptide natural product chemical space. It helps transition

the growing relationship between code-to-chemistry-based discoveries, validating predicted chemical existence, reasonable biosynthetic production, and the medicinal value of peptide natural products across the plant kingdom.

The foundation of this dissertation in natural product drug discovery could be expanded by future researchers based on the pivotal points of failure and success in hypothesis testing throughout this work. While Chapter 2 was an overall success in generating a large-scale transcriptome database for mining plant sequencing data, it can be further improved to aid in the accurate prediction of new post-translational cross-links. Upon ideation that the discovery of new post-translational in RiPP macrocyclization chemistry is enabled by transcriptome mining, it was predicted that new second-ring crosslinking in stephtanotic acid-type burpitides would follow known moroidin cyclization logic.

Therefore, it was predicted that the Val-Trp crosslink in Glechomanin was initially a His-Trp crosslink. It was not until the experimental isolation, characterization, and structural elucidation of Glechomanin's second ring chemistry that the crosslink was confirmed to exist as a Val-Trp macrocyclization. Thus, the hypothesis was expanded for mercurialin to include additional macrocyclizations arising from aromatic Phe-Tyr crosslinking, as inferred from sequencing data, which was experimentally validated by structural characterization.

Thus, I propose that a future direction for this work is to integrate automated structural prediction checkpoints into the context of computational mining and prediction. Such automated structure predictions could be done in a way that integrates knowledge of post-translational modifications to aid in the discovery of new RiPP macrocyclization chemistry in addition to *a priori* methods. Such automated predictions could have aided cyclopeptide alkaloid discovery in Chapter 5, which proved difficult because these compounds can undergo additional post-translational modifications beyond macrocyclization by BURP-domain cyclases. These modifications included c-terminal decarboxylation and N-terminal methylation. Though candidate cyclopeptide alkaloid nodes were identified with c-terminal decarboxylation, which could aid in the discovery of new decarboxylase enzymes involved in RiPP PTM pathways, the linkage between metabolites and sequencing data was not successful because cyclopeptide-alkaloid-type burpitide can exist and be processed in pathways where the repetitive core peptide domain is separated from the BURP-domain cyclases. Thus, further exploration of the linkages between core

peptides and cyclases in split burpitide pathways is also crucial and will help resolve the identity of peptide nodes in molecular networks in Chapter 5.

An automated method for converting core peptide sequences to SMILES strings is needed. This automation would reduce the time required to draw each core peptide in ChemDraw and convert it to SMILES. This process would have been useful in Chapter 3 for more rapid conversion of sequences into 3D peptide models and in Chapter 5 for generating peptide [M+H] and [M+2H] adducts, thereby enabling more rapid discovery and deconvolution of the molecular network.

Although the work in Chapter 3 does not discuss this, an attempt was made to enumerate peptide scaffolds to generate a library of over 300,000 burpitide ligands that do not exist in nature, demonstrating computational scalability in library creation. Furthermore, these ligands were used in a large-scale virtual screen against anti-bacterial LspA on the Great Lakes cluster using the Maestro virtual screening workflow. This process yields additional lyciumin hits that did not exist in natural plant sequences. Following up on these hits to determine their ability to be biosynthetically produced in tobacco via bioengineering of the LycA precursor gene would demonstrate the ability to expand hit diversification and biosynthetic production. Identification of lyciumin metabolite nodes in chapter 5 and their associated cyclases could aid in the production of additional natural and non-natural core peptides via bioengineering of novel lyciumin cyclases.

Furthermore, the work in Chapter 3 does not discuss or highlight the attempted validation of additional targets for computational docking to characterize the potential bioactivity of these peptides. Additional docking targets included proteins involved in proteostasis and degradation (Keap1 - PDB: 4IJF and ClpP proteolytic subunit - PDB: 3MT6), bacterial cell envelope integrity (SPase I – PDB: 3IIQ and BamA beta-barrel – PDB: 7NRE), as well as cytoskeletal and genetic machinery (Tubulin – PDB: 1Z2B, DnaN sliding clamp – PDB: 5AGU, and double stranded DNA – PDB: 5YTZ). These expanded docking simulation campaigns did not yield candidates with computational binding profiles comparable to or better than the structures' respective native ligands or inhibitors. Several targets had complex binding pockets, preventing the generation of a validated receptor grid that could accurately redock native ligands with an RMSD < 2.0 Å, thereby lacking a reliable benchmark for screening the burpitide library. In structures where benchmarking and receptor grid generation were successful, the resulting glide scores for the burpitide ligands remained equivalent to or lower than the scores of the native co-crystallized ligands, showing that they lacked pharmacophoric fit into certain protein binding sites. This demonstrates that

evolutionary selection of burpitides may have specific targeted biological functions rather than generalized affinity for broad protein surfaces.

Finally, the failure to isolate the prioritized lyciumin candidate for downstream testing in LspA inhibitory assays from Chapter 3 underscores that technical challenges remain in peptide purification. Alternative purification strategies beyond Sephadex LH-20 Size-Exclusion Chromatography should be explored, including C18 Reverse-Phase Solid-Phase Extraction, solvent system changes, or direct HPLC purification after tobacco leaf extractions. If further investigation of this peptide's antibacterial properties is pursued, it offers the opportunity to validate lyciumins as antibacterial agents through wet-lab experimentation.

6.2 Dissertation Conclusion

The transition from labor-intensive traditional methods to multi-omics approaches for natural product discovery is ongoing. This dissertation provides robust, validated methods designed to leverage the rapid expansion of research in this field. Collectively, the work presents multiple demonstrations of computational strategies, including genome mining and large-scale transcriptome analysis, to identify novel burpitide precursors and mature core peptide sequences. The developed methods establish a foundation for comprehensive computational pipelines in future research on plant natural products. Both *in silico* and *in vitro* results underscore the growing potential of plant-derived macrocyclic peptides as novel structures for addressing significant human diseases. It is my hope that this work will inspire the scientific community to further develop efficient, scalable strategies for the rational design of plant-derived secondary metabolites.