A Universal Sequencing System for Unknown Oligomers

David Doran[1], Emma Clarke[1], Graham Keenan[1], Emma Carrick[1], Cole Mathis[1] and Leroy Cronin[1*]

[1]*School of Chemistry, University of Glasgow, Joseph Black Building, University Avenue, Glasgow G12 8QQ, U.K.* [*]*Corresponding author email:* lee.cronin@glasgow.ac.uk

No synthetic chemical system can produce complex oligomers with fidelities comparable to biological systems. To bridge this gap, chemists must be able to characterise synthetic oligomers. Currently there are no tools for identifying synthetic oligomers with sequence resolution. Herein, we present a system that allows us to do omics-level sequencing for synthetic oligomers and use this to explore unconstrained complex mixtures. The system, Oligomer-Soup-Sequencing (OLIGOSS), can sequence individual oligomers in heterogeneous and polydisperse mixtures from tandem mass spectrometry (MS/MS) data. Unlike existing software, OLIGOSS can sequence oligomers with different backbone chemistries. Using an input file format, OLIG, that formalizes the set of abstract properties, any MS/MS fragmentation pathway can be defined. This has been demonstrated on four model systems of linear oligomers. OLIGOSS can screen large sequence spaces, enabling reliable sequencing of synthetic oligomeric mixtures, with false discovery rates (FDRs) of 0-1.1%, providing sequence resolution comparable to bioinformatic tools.

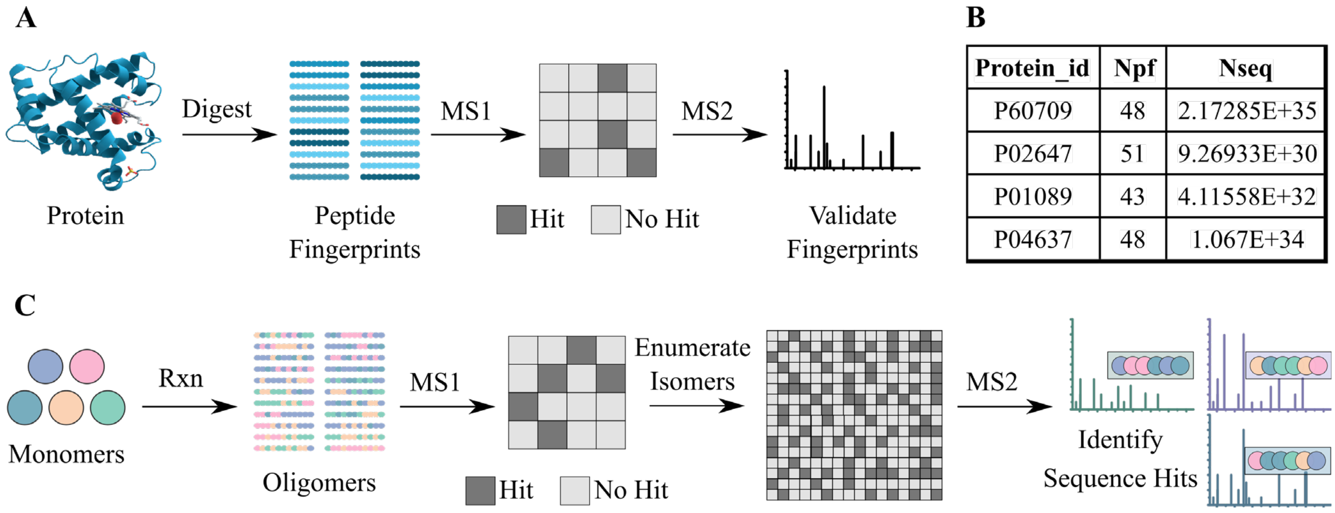
Omics technologies have enabled the rapid, accurate and cheap identification and sequencing of biological oligomers and polymers(*1*). Many of these technologies, particularly in proteomics, rely on mass spectrometry (MS) and tandem mass spectrometry (MS/MS) coupled to pre-measurement chromatographic separation via liquid chromatography (LC-MS, LC-MS/MS)(*2*). Detailed knowledge of the separation of biological oligomers via LC, their MS ionization and fragmentation via MS/MS, is used to identify specific biological oligomers and polymers from very complex, heterogeneous mixtures(*1*, *3*, *4*). Similar knowledge exists for many classes of synthetic oligomers analyzed via LC-MS/MS(*5–9*), but this knowledge has not been leveraged in a systematic manner. Therefore, synthetic oligomer sequencing remains a highly labor-intensive and time-consuming process, and is typically limited to relatively simple oligomer mixtures with 1-3 unique monomers(*6*). Additionally, there are currently no software tools available for omics-like identification and sequencing of synthetic oligomer mixtures, and this is in part because the conceptual basis for this is not yet developed. Some tools that do exist for characterization of synthetic oligomer MS/MS data are either unable to distinguish between isomeric sequences(*10*, *11*) or are bespoke to a single oligomer class(*12–14*).

Conceptually, a tool matching the precision, throughput and resolution of omics-level analysis for sequencing synthetic oligomers in mixtures would fulfill a major unmet need in materials science, oligomer chemistry and other fields, potentially spawning a new field of synthetic "oligomerics"(*5*, *15*). Two areas with possibly the most to benefit from such a tool include: i) identification of oligomeric degradation products and contaminants in polymer materials(*16–18*) and ii) sequence-controlled oligomerization(*5*, *15*). The benefits of identifying degradation products and contaminants include the quick, reliable identification of unwanted and potentially dangerous oligomers that may reduce the structural and functional viability of polymeric materials, or even pose a serious health risk in consumer products. In addition, achieving sequence-controlled oligomerization in multicomponent reactions remains an elusive goal in polymer chemistry and materials science(*15*). Several synthetic strategies have emerged for producing sequence-controlled oligomers and polymers, with stepwise(*19*) and more recently chain growth methods(*20*) enabling sequence-controlled synthesis of specific oligomers and oligomer libraries. However, no synthetic system to date has come close to matching the fidelity, specificity and turnover of biological systems in their production of complex, sequence-encoded polymers(*15*). Sequence-encoding of complex synthetic oligomers cannot be achieved without first having an oligomeromics tool for identifying which synthetic oligomer sequences are produced(*5*, *15*). Despite the potential for oligomeric sequencing, and well-established knowledge of synthetic oligomer fragmentation via MS/MS, a tool has not been developed before due to two key reasons. Firstly, the sequence space is vast as the combinatorial nature of sequence space to be screened as a function of oligomer length and number of unique monomers increases dramatically (Eq 1).

$$N = (1 + n) \sum_{i=1}^{L} m^{i} \qquad \text{(Eq. 1)}$$

where N = total number of unique sequences, n = number of potential terminal modifications, L = maximum sequence length, m = number of unique monomers.

Hence it is challenging to extensively survey the full range of possible products in synthetic oligomer mixtures. Secondly, diversity in unknown systems is incredibly challenging, due to the vast variation in MS/MS fragmentation pathways(*6*). In contrast, biological omics software deals with a very limited set of oligomer and polymer backbone chemistries including peptides(*21*, *22*), nucleic acids(*23*), and lipids(*24*). Additionally, pre-determined constraints on the sequence space for biological oligomers can greatly reduce the search space for sequence assignments(*1*, *25*, *26*), see Figure 1A. Such constraints often do not exist for synthetic oligomer mixtures(*5*, *15*). This means, without pre-defined sequence constraints, screening of similar oligomer mixtures from a synthetic source may require the enumeration and screening of 10^{30}-10^{34} x more unique sequences as compared to a typical proteomics peptide fingerprint identification workflow (Figure 1C). Screening of larger sequence spaces can be achieved by utilizing more computational resources or optimized search algorithms(*27*). However, the diversity in backbone chemistries remains a major challenge for synthetic oligomer sequencing(*5*, *6*). For biological oligomers, there are extensive and rich MS/MS data sets available for most major classes of biomacromolecules(*5*, *28*), enabling database-based sequence searches(*29*). Knowledge of fragmentation pathways also allows for first principles *in silico* fragmentation(*13*). For synthetic oligomers, these extensive data sets do not exist(*5*). Therefore, OLIGOSS relies on *in silico* fragmentation to predict potential MS2 fragments from synthetic oligomer sequences.



Protein_id	Npf	Nseq
P60709	48	2.17285E+35
P02647	51	9.26933E+30
P01089	43	4.11558E+32
P04637	48	1.067E+34

Figure 1: Sequence Space Constraints in Biological and Synthetic Oligomeromics. A: Standard proteomics workflow for identifying full-length proteins. Protein(s) are enzymatically digested to produce peptide fingerprints, the presence of which are then confirmed via identification of precursors at MS1. Fingerprints are then validated compositionally via MS/MS of confirmed MS1 precursors, with no need to screen for isomeric sequences. B: Number of peptide fingerprints (Npf) and total number of potential isomeric sequences for all fingerprints (Nseq) for four proteins identified in the UniProt database, subjected to *in silico* digest with trypsin. In proteomics workflows, only fingerprints are screened, not possible isomeric sequences. For synthetic oligomers screened by OLIGOSS, any and all potential isomers may be present C: Standard OLIGOSS sequencing workflow for identifying sequences in synthetic oligomer mixtures. Compositional screening is carried out on MS1 data. For each composition, any and all possible isomeric sequences could be present in the mixture. All isomers must then be enumerated and screened in the MS2 data to determine which isomeric sequences are present for each composition.

With sufficient knowledge of fragmentation pathways*, in silico* fragmentation of a single oligomer class is just a question of automation, requiring only hard coding of each specific fragmentation pathway. Given the diversity of fragmentation mechanisms for synthetic oligomers(*6*), this rapidly becomes impractical for a universal – or even near universal – sequencing tool. Considering the two examples of polyimines (Schiff base oligomers) and depsipeptides from Figure 2. In the case of polyimines (Figure 2A), fragmentation of a linear oligomer results in three cyclic fragment series, likely the result of complex rearrangements in a CID collision cell. Contrast this with peptides, which form two distinct linear fragment series along with monomer-specific satellite or signature ions (Figure 2B). To create OLIGOSS, a *de novo* universal sequencing tool, these dramatically different fragmentation mechanisms had to be translated into a reasonably simple, generalizable set of properties that could also be used to describe fragmentation of other oligomer classes.

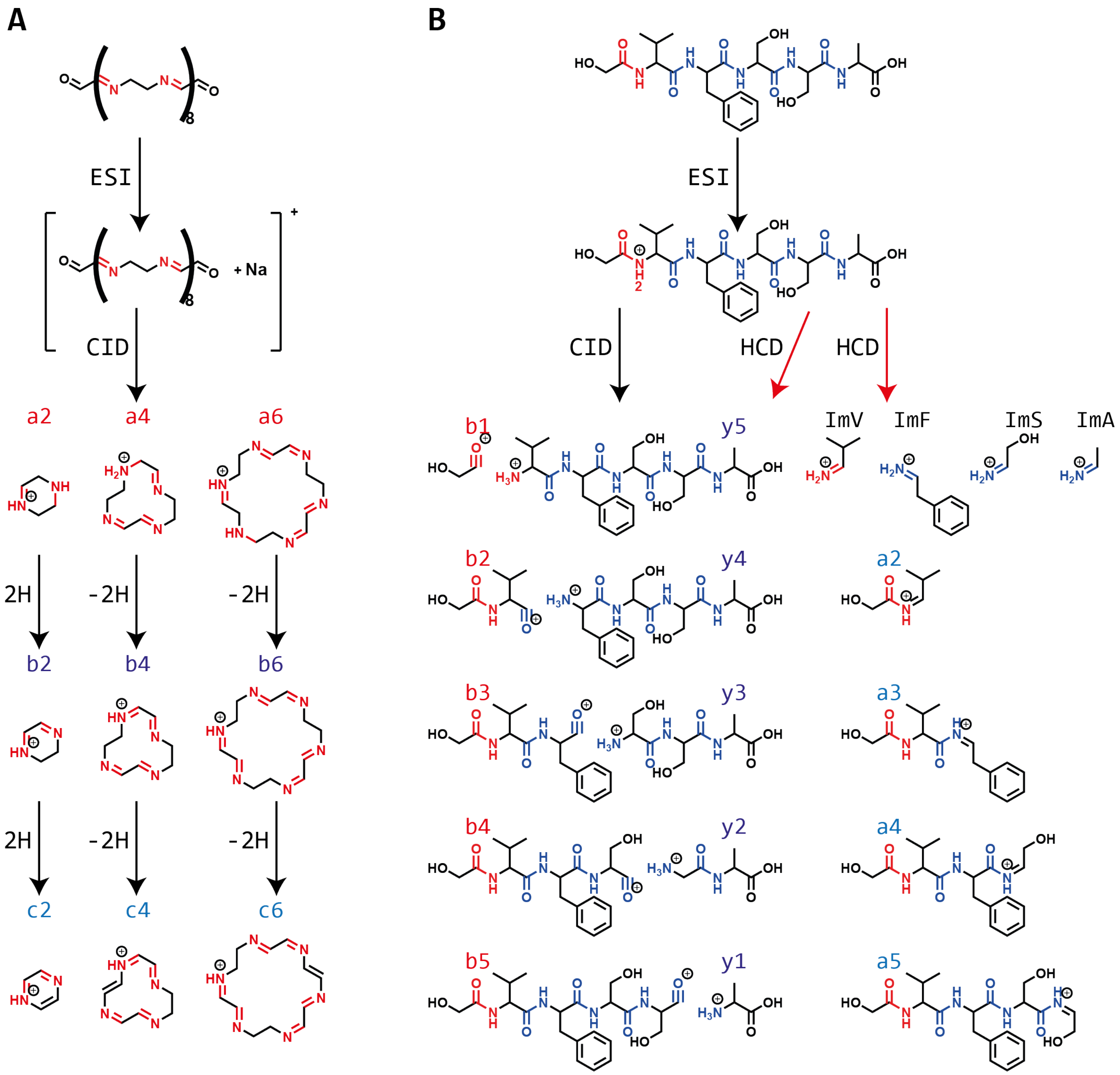


Figure 2: Differences in Fragmentation Mechanism between two Well-Known Oligomer Classes, Peptides and Polyimines. A: Under Collision-Induced Dissociation (CID), three cyclic fragment series are formed from a polyimine precursor. B: The fragmentation of peptides is affected by the fragmentation method used, with Higher-energy C-trap-Induced Dissociation (HCD) producing additional satellite fragments not observed in as high abundance in CID spectra.

OLIG - A Constraints Format for Oligomer MS/MS

Standard MS/MS fragment nomenclatures have been proposed for synthetic oligomers(*6*), analogous to standards that are well-established for peptides(*3*, *22*, *30*) and other biological oligomers(*23*, *31*). However, to our knowledge, never before have a set of generalized principles been used to describe and translate the pathways that produce these fragments into a universal framework for defining oligomer fragmentation. This is precisely what OLIGOSS does, using a set of abstract properties that make up OLIG, a new and very simple set of principles for oligomer MS/MS. Despite the variation in fragmentation mechanisms for synthetic oligomers, OLIGOSS translates MS/MS fragmentation pathways into a set of universal, abstract OLIG properties. These properties can then be used to predict and screen for fragments matching specific sequences in LC-MS/MS data. Fragmentation of a single oligomer class can often occur via different routes(*6*). Each of these can produce qualitatively different fragments. The relative predominance of each pathway will vary depending on instrumental conditions. Therefore, OLIG can be used to describe individual fragmentation pathways. Several of these translated pathways can then be combined to screen for sequences of the target oligomer class, with the flexibility to choose which will predominate for the specific instrumental conditions (e.g. fragmentation method).

Combined, these OLIG properties define the positions along an oligomer backbone at which specific fragments may occur, as well as predicting the exact mass-to-charge ratio (*m/z*) for all MS2 ions corresponding to the fragment and potential interactions with other ions present in the analyte (Figure 3). A basic summary of the major abstract properties is shown in Table 1. For a full list of properties see Supplementary Information Section S3. These can be considered a novel yet simple framework for describing oligomer fragmentation via MS/MS. For translations of the polyimine and depsipeptide fragmentation pathways shown in Figure 2, see Table 2.

Table 1: Basic OLIG Properties for Describing Oligomer Fragmentation. For full details of abstract properties see Supplementary Information.

Property	Description
i_{adduct}	Charge *and* mass due to exchangeable ions that are associated with fragments in a linear series by default (e.g. protonation for positive peptidic y-ions). These ions can be exchanged for extrinsic ions without affecting the overall charge state of fragments.
i_{ion}	Charge (*not* m/z) due to non-exchangeable ions that are associated with fragments in a linear series (e.g. acylium ions for positive peptidic b-ions). Addition of any *extrinsic* ions results in the final charge of fragments being equivalent to the sum of the intrinsic charge plus extrinsic charges.
m_{diff}	The *neutral mass* difference between fragments in a series and their corresponding equivalent intact sequence.
Symmetry	Specifies whether termini are equivalent for the oligomer class. If termini are non-equivalent, each terminus is arbitrarily assigned a value of 0 or -1.
T	"Home" terminus from which a fragment series is indexed. This is only relevant for asymmetric oligomer classes. Remaining non-home terminus is denoted as !T.
F	Fragmentation unit. The number of unit increments between adjacent fragments in a linear series. Used principally for alternating co-oligomers with periodic differences in backbone links. Series with F=1, F=2 and F=3 will occur ever 1, 2 and 3 residues along the backbone, respectively.
s	Position (relative to T) at which a fragment series begins. Series with s=0, s=1, and s=2 will be indexed from 0, 1 and 2 F from T, respectively.
e	Ending position (relative to !T) at which a fragment series ends. Series with e=0, e=1 and e=2 will terminate 0, 1 and 2 F prematurely from !T.
E	Exceptions to the values of the aforementioned properties when fragmenting a particular bond type. This is used only for oligomer classes with irregular mixed backbone links.

Table 2: Depsipeptide and Polyimine Fragmentation Pathways Translated into OLIG. Three depsipeptide series (a, b and c) and three polyimine series (a*, b* and c*) are characterized by properites described in Table 3: mass difference (m_{diff}), intrinsic ion (i_{ion}), instrinsic adduct (i_{adduct}), home terminus (T), starting position (s), end position (e), fragmentation unit (F) and exceptions (E).

Series	m_{diff}	i_{ion}	i_{adduct}	T	s	e	F	E
a	-COOH	0	+H	0	1	1	1	null
b	-OH	1	N/A	0	1	1	1	null
y	+H	0	+H	-1	0	0	1	"ester"
a*	-OH	0	+H	0	0	0	2	null
b*	-H3O	0	+H	0	0	0	2	null
c*	-H5O	0	+H	0	0	0	2	null

Deploying OLIG in a *De Novo* Sequencing Workflow

In OLIGOSS, after translating the fragmentation pathways of an oligomer class into OLIG, these translated pathways can then be saved in a reusable configuration file that can be called on whenever required for a sequencing experiment. As there are often several translated fragmentation pathways in a single OLIG configuration file, which of these pathways are relevant in a particular experiment will vary depending on instrumentation used and other experiment-specific parameters. Therefore, as well as reusable OLIG configuration files, which only must be translated and stored once per oligomer class, OLIGOSS uses run parameter files to execute individual experiments. These run parameter files provide all information necessary for determining which fragmentation pathways are relevant, and for processing and filtering MS1 and MS/MS spectra (Figure 3). Upon execution of a sequencing workflow, the full scope of all possible ionization and fragmentation pathways are read from the appropriate OLIG configuration file (Figure 3). Experimental details parsed from the experimental run file and the instrument(s) used are then used to determine which of the possible fragmentation pathways are likely to predominate. A validation step ensures that the chosen pathways are feasible, before a full MS1 library of all possible precursor permutations is generated and screened. For each precursor hit, all possible isomeric sequences are enumerated along with corresponding MS2 product ions. These are then also screened in MS2 spectra. At the end of this process, sequences are assigned a confidence score on the basis of the number of MS2 product ions confirmed in MS2 spectra (Eq 2) and (optionally) the distribution of confirmed fragments along the sequence backbone (Eq 3, Eq 4). Throughout MS2 screening, checks are in place to ensure the quality of spectral matches for each confirmed MS1 precursor – MS2 fragment combination (see Supplementary Information S5.3).

$$C = \frac{n_c}{n_t - n_{uo}} \qquad \text{(Eq. 2)}$$

Confirmed Fragment Ratio (C). n_c = number of confirmed fragments, n_t = number of theoretical fragments, n_{uo} = number of unconfirmed optional fragments.

$$< \alpha > = \frac{1}{N} \sum_i^N \frac{\alpha_i}{L} \qquad \text{(Eq. 3)}$$

Mean Continuous Fragment Coverage (<α>), a measure of the mean maximum coverage for each fragment series. N = number of individual fragment series, L = length of largest theoretical continuous block of fragments in series.

$$Confidence = < \alpha > S + C(1 - S) \qquad \text{(Eq. 4)}$$

Final Confidence Score. <α> = Mean Continuous Fragment Coverage, S = weighting factor between 0 and 1, C = Confirmed Fragment Ratio.

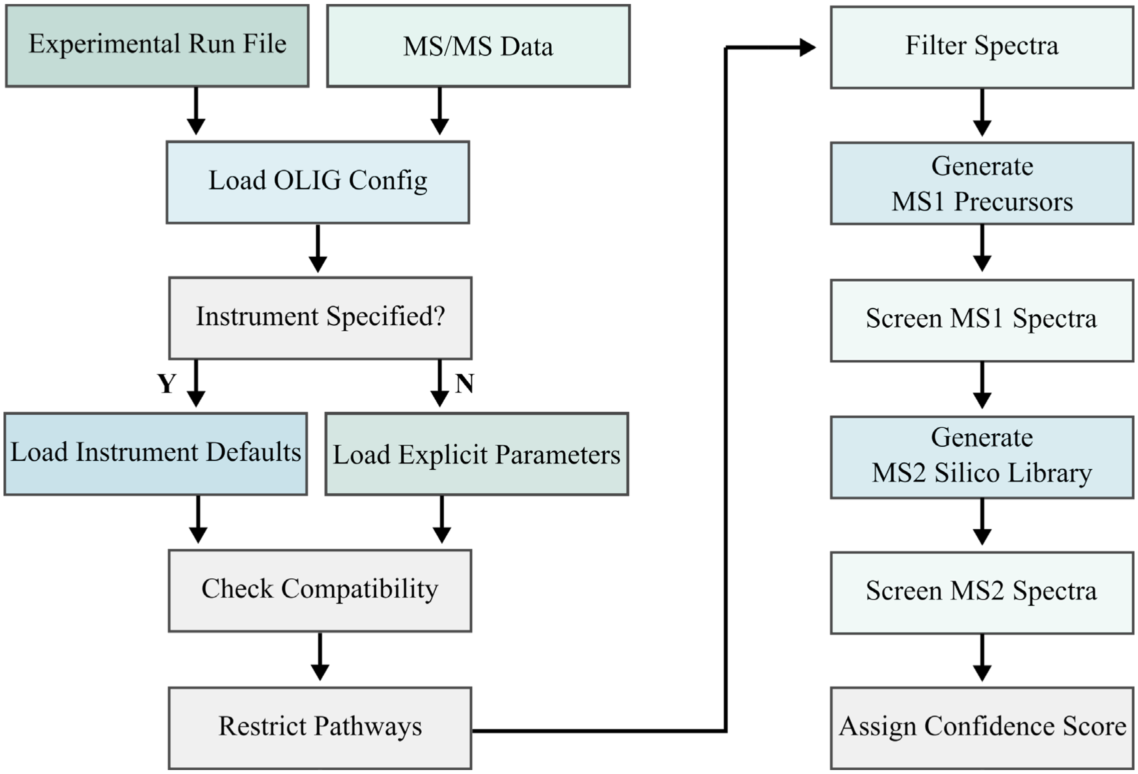


Figure 3: OLIGOSS Exhaustive Screen. Input data and experiment run file are read by OLIGOSS, which then chooses the appropriate OLIG configuration file for the oligomer class. If an instrument model is specified in the run file, instrument-specific defaults for ionization and fragmentation pathways, as well as instrument performance (resolution, sensitivity, retention time ranges), are loaded. If no instrument model is specified, these values must be explicitly stated in the experiment run file. Compatibility of the oligomer class with the chosen run parameters and instrument(s) is then checked and validated, after which input spectra are filtered and the most suitable fragmentation pathways are chosen for screening, A compositional MS1 precursor library is then generated *in silico* and screened in input MS1 spectra. For each composition confirmed at MS1, a full library of all corresponding isomeric sequences and their MS2 fragments is then generated and screened. Individual sequence scores are then assigned a confidence score and ranked.

Model Systems for *De Novo* Sequencing

OLIGOSS was designed for *de novo* sequencing of complex, polydisperse oligomer mixtures, particularly those which cannot be sequenced using current omics software. Therefore, to develop and validate this new tool, product mixtures containing a diversity of products and possible fragmentation pathways were required. Four model systems of oligomer mixtures were used: depsipeptides, N-terminally acylated peptides, polyesters and polyimines (Figure 4). Depsipeptides are peptidic oligomers with a mixture of amide and ester backbone linkages.

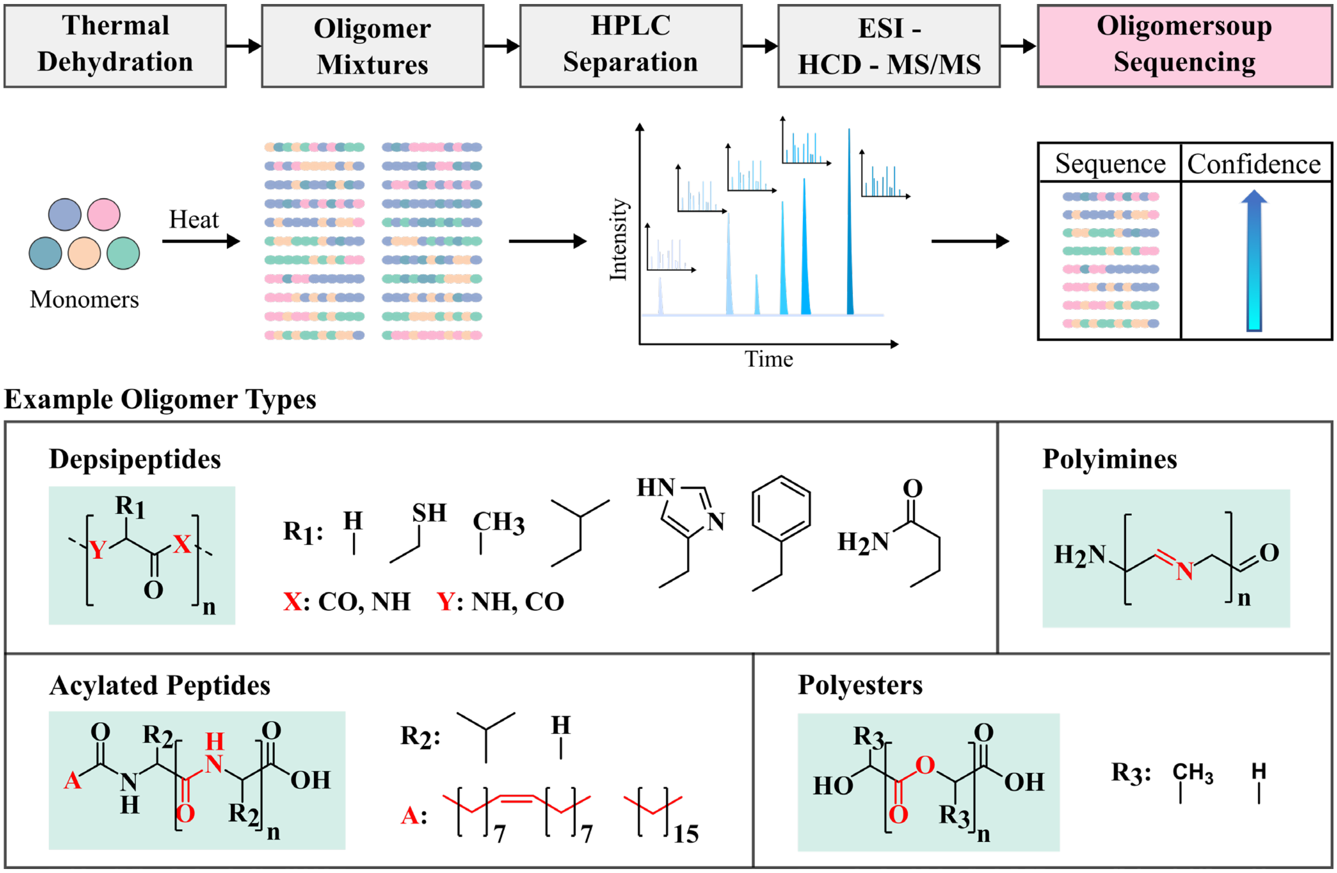


Figure 4: Model Reaction Systems for OLIGOSS. From simple thermal dehydration of monomer solutions, polydisperse oligomer mixtures are produced. These are then separated via HPLC, ionized via electrospray ionization (ESI) and fragmented via high energy C-trap induced dissociation (HCD) using on a UHR-Orbitrap Tribrid Lumos. In each duty cycle, the N most abundant precursors are selected for fragmentation. The result is a series of MS1 and MS2 spectra, which are then converted and passed on to OLIGOSS for sequencing.

Despite having similar MS/MS fragmentation pathways to pure peptides(*32*), and their occurrence in many well-characterized natural products(*34*), current proteomics software tools are unable to sequence depsipeptides(*12*, *34*). Using simple wet-dry cycles of amino acid and α-hydroxy acid monomers, polydisperse depsipeptide mixtures can be produced with high yield(*35*–*37*). Thus, they make for an ideal candidate for testing and validating OLIGOSS’s sequencing capabilities.

Five sets of depsipeptide products, each produced via wet-dry cycling of three amino acids and the α-hydroxy acid glycolic acid, were analysed via LC-MS/MS and sequenced using OLIGOSS. Due to the combinatorial nature of sequence space in these reactions (Eq 1), diverse oligomer sequences of various lengths were produced. OLIGOSS was able to successfully survey this sequence space for all reactions tested (Figure 5), determining the proportion of ester linkages (Figure 5a) and proportion of total sequence space represented (Figure 5b) in each product mixture as a function of oligomer length. An overall trend was observed for increased ester enrichment and decreased representation of sequence space with increasing oligomer unit length.

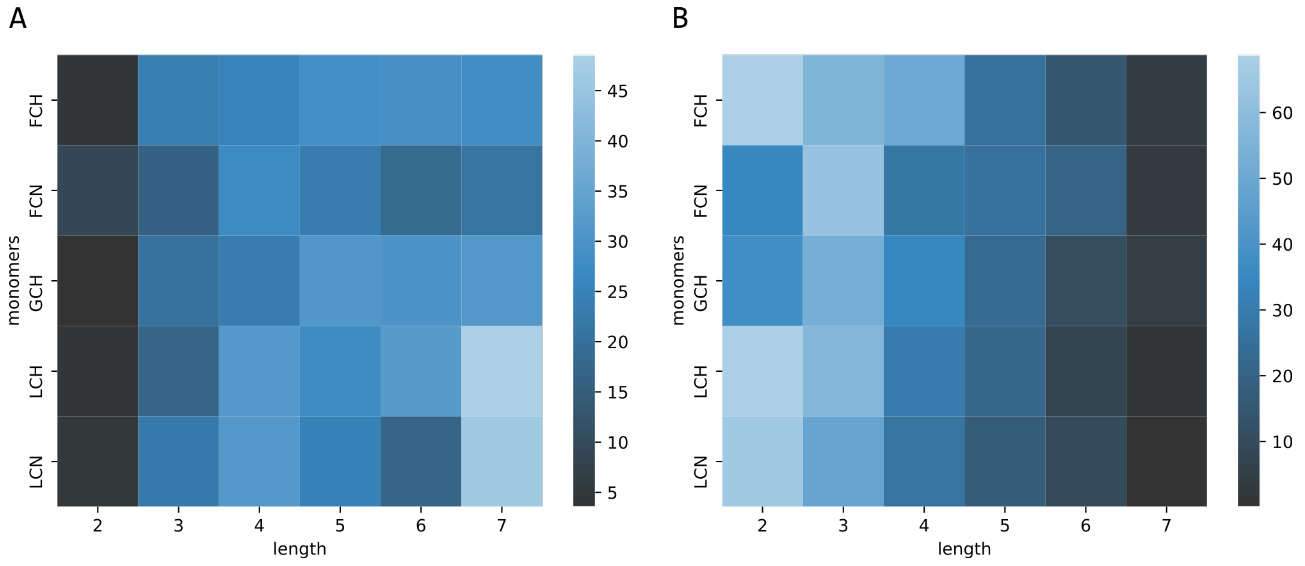


Figure 5: Sequence Diversity of Depsipeptide Mixtures Surveyed by OLIGOSS. Each mixture was produced via wet-dry cycling of three amino acid monomers (denoted by one-letter codes) and glycolic acid. A: Proportion of backbone links comprised of ester bonds (in %). B: Proportion of sequence space covered (in %). Monomer one-letter codes: F (phenylalanine), C (cysteine), H (histidine), G (glycine), L (leucine).

Successful analysis of depsipeptide mixtures demonstrates the ability of OLIGOSS to sequence and characterize an oligomer class with complex fragmentation pathways and heterogeneous backbone linkages that are outside the scope of existing omics software. To test its potential universality in linear oligomer sequencing from MS/MS data, polyimine and pure polyester oligomer mixtures were synthesized, analyzed via LC-MS/MS and sequenced (Figure 4).

To our knowledge, polyimine fragmentation pathways are not well-characterized in CID-like MS/MS. Nonetheless, fragment signatures were identified via mass ladders present in HCD MS2 spectra of polyimine precursors (Figure 6A). Translation of these proposed fragmentation pathways into an OLIG configuration file was used to sequence polyimine oligomers produced from unconstrained diamine-dialdehyde condensation of ethylenediamine (e) and glyoxal (o) (Figure 6A-B, Table 2). Unlike the random depsipeptide backbones, polyimines produced in such a manner from bi-functional non-self-reactive monomers must be alternating co-oligomers. Therefore, the sequence space for screening is relatively constrained. The majority of polyimine sequences confirmed were of relatively low confidence ($\leq$ 30%). Nonetheless, LC-MS/MS data was screened using a false polyimine oligomer library of equal size to the eo library but comprised of monomers not used in the reaction. No false hits were found at any confidence threshold for the false library (Figure 6C). Polyesters have well-characterized fragmentation pathways(*38*). Despite reports of polyester LC-MS/MS in biological(*39*), industrial(*40*) and even prebiotic chemistry(*41*) settings, to our knowledge no tools exist for automated polyester sequencing. A mixture of random oligoesters, produced via unconstrained oligomerization of glycolic acid and lactic acid, was analyzed via LC-MS/MS and successfully sequenced using OLIGOSS (Figure 4D). Despite the complex fragmentation pathways of polyester cations, which are a result of charge-remote fragmentations and association with extrinsic ions(*6*, *38*) (an example [M+Na]+ polyester fragment is shown in Figure 6D), 523 unique polyester sequences were confirmed. No sequences were found for the equivalent false oligomer library control.

Benchmarking OLIGOSS's Performance

Having demonstrated the ability of OLIGOSS to sequence products that cannot be analyzed using proteomics tools, its performance was benchmarked for pure peptide sequencing. Eight peptide standards (ASGNQ, FSGNQ, GSGNQ, ASGNQSGV, FSGNQSGV, GSGNQVGS, FSGNQSGVSA and FSGNQVGSAS) were synthesized, analyzed via LC-MS/MS and then subjected to blind sequencing runs. In each run, only the backbone class, constituent monomers and oligomer length were specified. Standards were chosen due to their neutral loss-prone sidechains(*42*), thus increasing the average number of unique MS1 precursor and MS2 product ions to be screened for each run. Neutral losses were predominant in MS2 product ion spectra (Figure 7A).

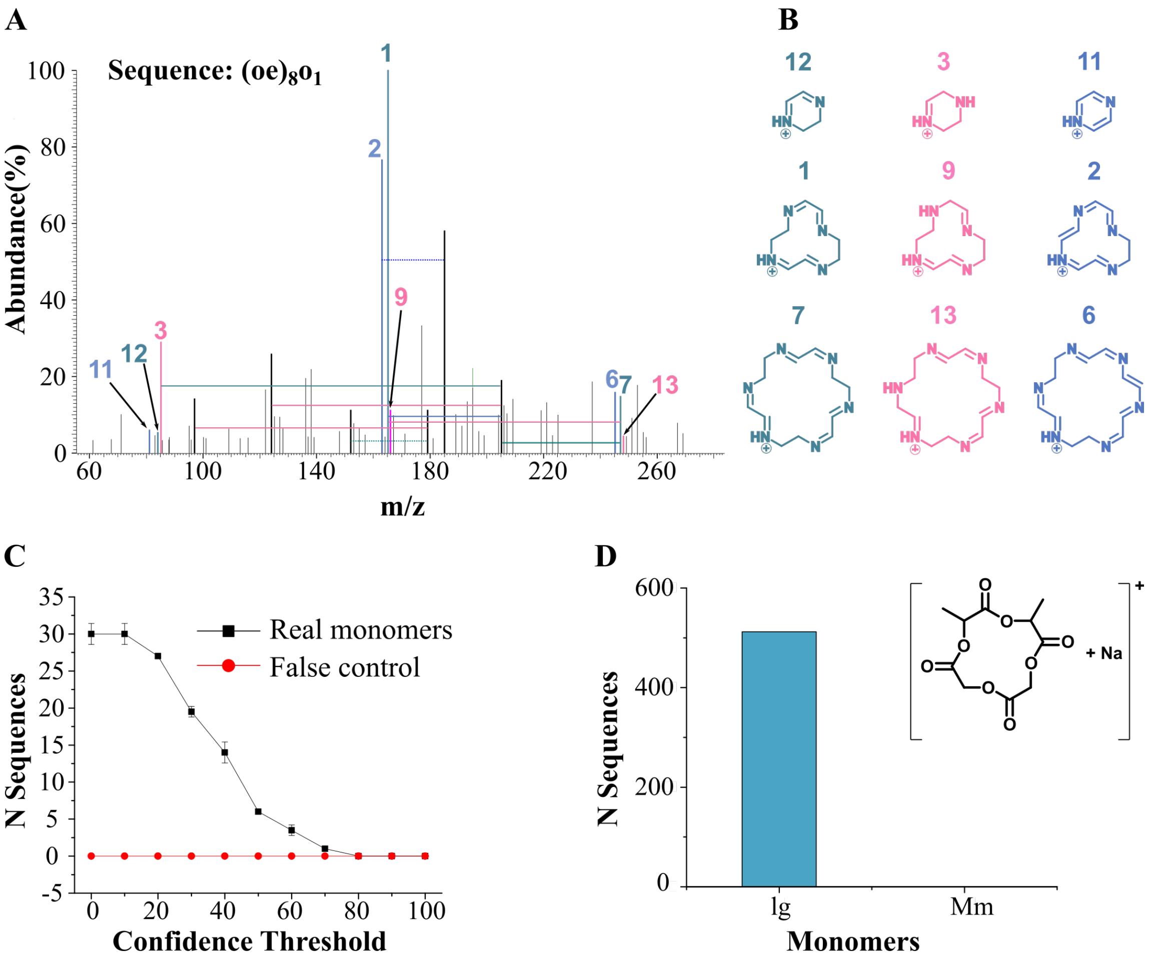


Figure 6: Polyimine and Polyester Sequencing. A: MS2 HCD spectrum of (oe)8o1 alternating co-oligomer polyimine. Peaks assigned to a specific fragment are highlighted. B: Fragments identified in MS2 spectrum from (A). C: Number of confirmed sequences over minimum confidence thresholds for standard run and false monomer control. D: Example polyester cyclic fragment and associated sequencing result of unconstrained lactic acid ("l") and glycolic acid ("g") mixture, with false monomer control ("Mm"). Data in C and D represent mean of 3 measurements ± 1 S.D.

In each of the runs, false assignments were made (Figure 7B-D). However, these represent only a small fraction of sequence space screened for both isomeric and non-isomeric assignments. For octameric sequences ASGNQSGV and FSGNQSGV the sequence space covered in the blind runs was equal to 10,080 isomeric and 1.67 x 10^6 non-isomeric sequences. Isomeric false assignments were below 10 % for all octameric standards. Very few non-isomeric false assignments were made for the octamers, never exceeding 0.1 % (Figure 7C).

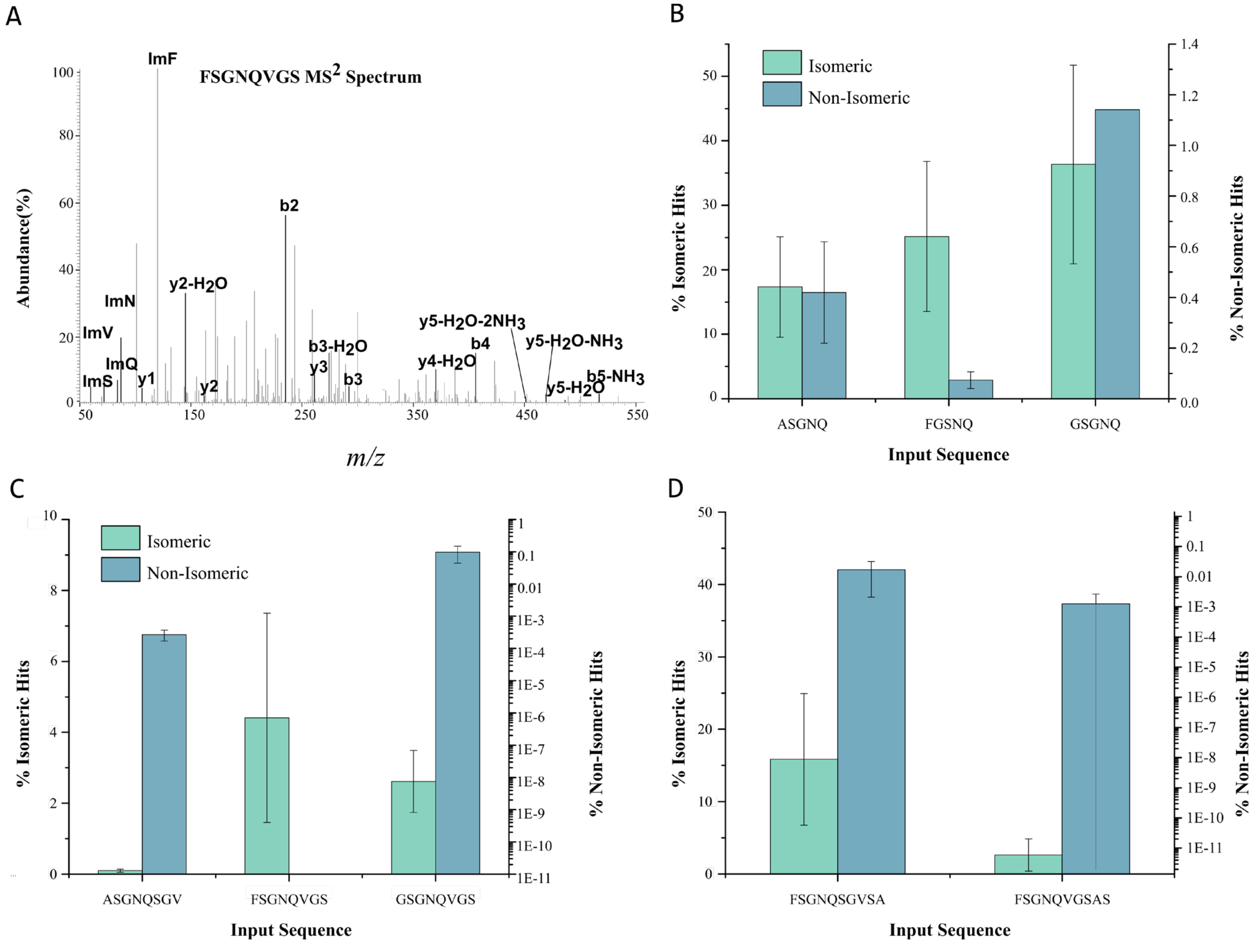


Figure 7: Peptide Sequencing Standards. A: MS2 HCD spectrum of peptide standard FSGNQVGS. b- and y-fragments are annotated along with any associated H_2O and NH_3 neutral losses. Monomer-specific immonium fragments are denoted with an "Im" prefix. B: Isomeric and Non-Isomeric false assignments for pentameric standards ASGNQ, FSGNQ and GSGNQ. C: Isomeric and Non-Isomeric false assignments for octameric standards ASGNQSGV, FSGNQSGV and GSGNQSGV. D: Isomeric and Non-Isomeric false assignments for decameric standards FSGNQSGVSA and FSGNQVGSAS. Data in B, C and D represent mean of 5 measurements ± 1 S.D.

This is superior to typical false discovery rates (FDRs) reported for peptide matches in proteomics workflows(*43*), possibly due to their reliance on spectral matching(*44*) rather than first principles *in silico* fragmentation employed in OLIGOSS. FDRs were slightly higher for the three pentameric standards tested (Figure 7B), ranging from 10 - 35 % and 0.05 – 1.1 % for isomeric and non-isomeric false assignments, respectively. Considering the reduced sequence space for pentamers, this is not unexpected. For decameric standards FSGNQSGVSA and FSGNQVGSAS, the total sequence space screened was 279,138 isomeric and 2.82 x 10^8 non-isomeric sequences. FDRs for non-isomeric sequences were similar to the octameric standards, not exceeding 0.1 % for non-isomeric sequences (Figure 7D). Isomeric sequence FDRs for the decameric standards were comparable to the pentameric standards, ranging from 2.23 – 15.84 % for FSGNQSGVSA and FSGNQVGSAS respectively. The comparably high isomeric FDR for the decameric standards is likely due to gaps in fragment series observed in raw MS2 spectra, which is not atypical for longer oligomers(*6*).

Sequencing in the Presence of End-Groups

Presence of terminal end-groups can dramatically affect fragmentation pathways of oligomers. Thus, OLIGOSS has been designed with the capability to screen for terminal modifications while simultaneously sequencing products. By performing wet-dry cycles of depsipeptides in the presence of fatty acids oleic acid and palmitic acid, mixtures of N-terminally acylated peptides were produced. These were analyzed via LC-MS/MS and sequenced using OLIGOSS, which lead to the successful identification of acylated sequences (Figure 8), with no reductions in performance in sequencing of non-acylated species in the same mixture (Supplementary Information Section S5). Two example spectra of N-terminally acylated peptides identified by OLIGOSS, a palmitated and oleated valine dimer, are shown in Figure 8. Free acylium signature fragments were found for both palmitic acid (Figure 8A) and oleic acid (Figure 8B) moieties.

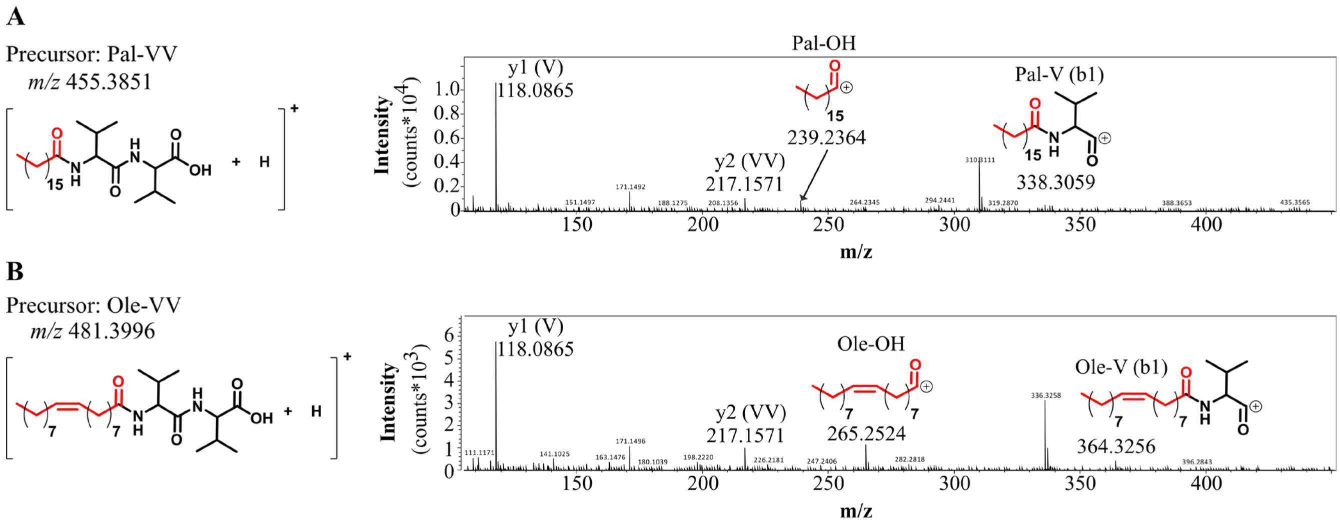


Figure 8: N-Terminally Acylated Peptides Identified by OLIGOSS. CID MS2 spectrum of valine dimer with A: N-terminal palmitic acid acylation, B: N-terminal oleic acid acylation.

Peptidic b- and y- fragments were observed as acylated moieties (in the case of b1 fragments) and fully dissociated (in the case of observed y2 fragments) for both species. Based on these observations, end-group modifications were incorporated into OLIGOSS's OLIG configuration files, with the ability to specify expected behavior for free modification signature ions, specific target sites, and also effects on other fragment series (Supplementary Information Section S7). At least one software tool exists for end-group characterization in oligomer MS/MS spectra(*45*). However, it requires that the oligomer sequence is predetermined and operates on single spectra, thus providing only a limited and extremely low through-put means of analysis. OLIGOSS, in contrast can simultaneously sequence oligomers and screen for end-groups in the thousands of spectra typically acquired for high-resolution LC-MS/MS analyses.

Conclusions

OLIGOSS is the first tool for automated, *de novo* oligomer sequencing from tandem mass spectrometry data with the ability to sequence oligomers with multiple backbone chemistries. We have demonstrated the ability of OLIGOSS to sequence five different sets of synthetic oligomeric mixtures (peptides, acylated peptides, polyimines, polyesters and depsipeptides), the latter three of which current biological omics tools are unable to analyze. Unlike other tools which are bespoke for a single oligomer class(*12*, *13*), or are unable to perform sequencing(*10*, *11*, *37*), OLIGOSS has the potential for expansion to any set of oligomers which are amenable to analysis via MS/MS. All of the code used here will be made available for use and editing subject to a GPLv3 license, in the hope that others may benefit from OLIGOSS and use it to sequence even more classes of oligomers which are still beyond the reach of omics software. Given the increased utility of mass spectrometry in the analysis of oligomers and polymers, this tool has the potential to greatly expand the capabilities of researchers in oligomer and polymer chemistry and related fields. The ability to screen large sequence spaces, combined with the flexibility to handle a diverse range of backbone chemistries, will enable researchers to perform truly omics-level characterization and sequencing of non-biological oligomers and polymers for the first time. The "omics" revolution has led to many great advancements in biology and medicine, thanks in no small part to software tools for automated sequencing of biological oligomers from MS/MS data. OLIGOSS may be the first step towards a second and perhaps similarly fruitful "omics" revolution, the one of oligomeromics.

Methods

Software and Data Analysis

OLIGOSS was written in Python 3.7.5. All figures from OLIGOSS sequencing output were generated using Python package Seaborn 0.11.0 and OriginPro 2016. OLIGOSS source code can be viewed on github: https://github.com/croningp/oligoss.git. pip install oligoss A copy of a test dataset and config file is here along with a very basic tutorial: doi:10.5281/zenodo.4252732

Depsipeptide Model Reactions

To produce a model system of N-terminally acylated and non-acylated (depsi)peptide mixtures (Scheme 1), α-amino acid and α-hydroxy acid monomers were subject to thermal dehydration using methods similar to those described previously in the literature (*36*, *46*). Depsipeptide starting mixtures were made up with 0.1 M amino acid and 0.1 M glycolic acid (Sigma, CAS: 79-14-1) in HPLC-grade H2O and adjusted to desired pH using 2M H3PO4 or 2M NaOH. Immediately prior to heating at 95 oC for 15 hours in open-cap glass vials, 10 mL of pH-adjusted monomer stock was added to 0.33 mL oleic acid (Sigma, CAS: 112-80-1), 0.256 g palmitic acid (TCI, CAS: 112-80-1) or 0.33 mL HPLC-grade H2O for oleated, palmitated and fatty acid-free reactions respectively. Upon dehydration after heating at 95 oC for 15 hours, samples were redissolved in 10 mL HPLC-Grade H2O. Redissolved products were then sonicated at 45 oC for ≥ 15 min. After sonication, 1.2 mL aliquots were harvested and centrifuged at 10,000 rpm for 30 min and the aqueous layer harvested. The harvested aqueous layer was then diluted 1:10 in MS-grade H2O and filtered into a glass HPLC vial through a 0.22 µm nylon syringe filter.

Polyimine Model Reactions

To produce a model system of alternating co-polymers, Schiff base polymers (polyimines) were synthesized via uncontrolled oligomerization of diamine and dialdehyde monomers. Monomer stocks were made up to 0.1 M in appropriate solvent. 2 mL of each monomer stock (one dia-mine and one dialdehyde per reaction) was added to a 10 mL glass vial. 6 mL HPLC-grade MeCN was then added to the vial to give a total reaction volume of 10 mL and starting material concentration of 0.02 M for each diamine and dialdehyde. Mixtures were then stirred at 200 rpm and continuously heat-ed at 70 oC for 30 min. Products were then cooled to 4 oC prior to 50 % dilution in a 1:1 MeCN:MeOH mixture (MS-grade, + 0.01 % formic acid). Diluted products were filtered through 0.22 µM nylon syringe membrane before analysis via the Orbitrap Lumos Tribrid Mass Spectrometer.

Mass Spectrometry Data Acquisition

Unless specified otherwise, all measurements acquired via the Orbitrap Lumos Tribid mass spectrometry were carried out in positive mode using DDA to select the most intense ions for tandem mass spectrometry via HCD. To ensure sufficient acquisition of low abundance products, a 1 min dynamic exclusion window was applied with width of 5 ppm.

Solid Phase Peptide Synthesis

All Fmoc-protected amino acids and Fmoc-protected Wang resins were purchased and used without further purification from NovaBioChem and Sigma-Aldrich. All solvents and reagents were purchased from Sigma-Aldrich. 2 mL reactor vials with frit filters were purchased from Biotage.

Peptide synthesis was performed using the Biotage Syro II automated peptide synthesiser fitted with two 48 reactor blocks. Each 2 mL reactor vial (RV) was loaded with the desired Fmoc-protected Wang resin (0.25 mmol). Each synthesis was repeated in multiple vials across the reactor block to afford a suitable yield. The peptide synthesis proceeded in four stages: swelling, deprotection, coupling and washing.

500 µL Ultrapure DMF was added and each RV was shaken for 1 hour at room temperature. Following the resin swelling, the RVs were drained for 60 seconds using vacuum.

The deprotection was performed in two stages. 500 µL of piperidine solution (20 % v/v in DMF) was added and the RV was shaken at room temperature for 3 minutes. After this first deprotection reaction, the RV was drained and 500 µL of fresh piperidine solution was dispensed into the RV. The second deprotection reaction lasted for 10 minutes, after which the RV was drained. 500 µL Ultrapure DMF was added to the RV and shaken for 60 s, followed by a 60 s drain. The RV was washed this way a further 4 times.

Double coupling was carried out for each amino acid addition. The required amino acid solution (4.0 eq, 0.5 M in DMF) was dispensed into the RV followed by hydroxybenzotriazole (HOBt, 4 eq, 0.5 M in DMF) and N,N′-diisopropylcarbodiimide (DIC, 4 eq, 3 M in DMF). The RV was shaken at room temperature for 1 hour. The reagents were then drained and the resin was washed with Ultrapure DMF (500 µL) as previously described. Cycles of deprotection and coupling were repeated with different amino acids until the peptide was of desired composition.

After a final deprotection of the N-terminus amino acid, the resin-bound peptide was washed five times with Ultrapure DMF, as previously described. Following the DMF washing, the peptides were further washed with DCM (500 µL) for 60 s whilst shaking.

The reactor blocks were removed from the Syro II and placed into a fumehood, all subsequent operations were carried out manually. 2 mL of cleavage cocktail (96 % trifluoroacetic acid, 2 % triisopropyl silane, 2% H2O) was added to each RV and left to shake for approximately 3 hours at room temperature. Following this, the cleaved solution was drained into a 15 mL centrifuge tube. 10 mL of cold diethyl ether was added to the filtrate and the solution was left to precipitate at -20°C overnight. The resulting solid was washed under centrifugation (4.5 minutes, 4000 rpm) three times with 15 mL of cold ether. The ether from the final wash was discarded and the remaining solid was left to dry in a desiccator for at least 15 hours.

REFERENCES

1. J. S. Cottrell, Protein identification using MS/MS data. *J. Proteomics.* **74**, 1842-1841 (2011).

2. Z. Qiu, J. Peng, L. Mou, X. Li, F. Meng, P. Yu, Development and application of a UPLC–MS/MS method for P-glycoprotein quantification in human tumor cells. *J. Chromatogr. B Anal. Technol*. **1084**, 14-22 (2018).

3. I. K. Chu, C. K. Siu, J. K. C. Lau, W. K. Tang, X. Mu, C. K. Lai, X. Guo, X. Wang, N. Li, Y. Xia, X. Kong, H. Bin Oh, V. Ryzhov, F. Tureček, A. C. Hopkinson, K. W. M. Siu, Proposed nomenclature for peptide ion fragmentation. *Int. J. Mass Spectrom*. **390**, 24-27 (2015).

4. Y. Mechref, Use of CID/ETD mass spectrometry to analyze glycopeptides. *Curr. Protoc. Protein Sci*. **12**, 12-24 (2012).

5. E. Altuntaş, U. S. Schubert, "Polymeromics": Mass spectrometry based strategies in polymer science toward complete sequencing approaches: A review. *Anal. Chim. Acta.* **808**, 59-69 (2014).

6. C. Wesdemiotis, N. Solak, M. J. Polce, D. E. Dabney, K. Chaicharoen, B. C. Katzenmeyer, Fragmentation pathways of polymer ions. *Mass Spectrom. Rev*. **30**, 523-559 (2011).

7. T. M. Crescentini, J. C. May, J. A. McLean, D. M. Hercules, Mass spectrometry of polyurethanes. *Polymer (Guildf).* **181**, 1-33 (2019).

8. J. P. Williams, G. R. Hilton, K. Thalassinos, A. T. Jackson, J. H. Scrivens, The rapid characterisation of poly(ethylene glycol) oligomers using desorption electrospray ionisation tandem mass spectrometry combined with novel product ion peak assignment software. *Rapid Commun. Mass Spectrom.* **13**, 888-897 (2007).

9. K. De Bruycker, A. Welle, S. Hirth, S. J. Blanksby, C. Barner-Kowollik, Mass spectrometry as a tool to advance polymer science. *Nat. Rev. Chem*. **4**, 257–268 (2020).

10. K. De Bruycker, T. Krappitz, C. Barner-Kowollik, High Performance Quantification of Complex High Resolution Polymer Mass Spectra. ACS Macro Lett. **7**, 1443–1447 (2018).

11. A. J. Surman, M. Rodriguez-Garcia, Y. M. Abul-Haija, G. J. T. Cooper, P. S. Gromski, R. Turk-MacLeod, M. Mullin, C. Mathis, S. I. Walker, L. Cronin, Environmental control programs the emergence

of distinct functional ensembles from unconstrained chemical reactions. *Proc. Natl. Acad. Sci*. **116**, 5387-5392 (2019).

12. J. G. Forsythe, A. S. Petrov, W. C. Millar, S.-S. Yu, R. Krishnamurthy, M. A. Grover, N. V. Hud, F. M. Fernández, Surveying the sequence diversity of model prebiotic peptides by mass spectrometry. *Proc. Natl. Acad. Sci.* **114**, E7652-E7659 (2017).

13. P. J. Sample, K. W. Gaston, J. D. Alfonzo, P. A. Limbach, RoboOligo: Software for mass spectrometry data to support manual and de novo sequencing of post-transcriptionally modified ribonucleic acids. *Nucleic Acids Res*. **43**, E64-E77 (2015).

14. B. Ma, Novor: Real-Time Peptide de Novo Sequencing Software. *J. Am. Soc. Mass Spectrom*. **26**, 1885-1994 (2015).

15. J. K. Szymański, Y. M. Abul-Haija, L. Cronin, Exploring Strategies to Bias Sequence in Natural and Synthetic Oligomers and Polymers. *Acc. Chem. Res.* **51**, 649-658 (2018).

16. M. Himmelsbach, W. Buchberger, E. Reingruber, Determination of polymer additives by liquid chromatography coupled with mass spectrometry. A comparison of atmospheric pressure photoionization (APPI), atmospheric pressure chemical ionization (APCI), and electrospray ionization (ESI). *Polym. Degrad. Stab.* **94**, 1213-1219 (2009).

17. R. Noguerol-Cal, J. M. López-Vilariño, M. V. González-Rodríguez, L. F. Barral-Losada, Development of an ultraperformance liquid chromatography method for improved determination of additives in polymeric materials. *J. Sep. Sci.* **30**, 2452-2459 (2007).

18. M. Gill, M. J. Garber, Y. Hua, D. Jenke, Development and validation of an HPLC-MS-MS method for quantitating bis(2,2,6,6-tetramethyl-4-piperidyl) sebacate (tinuvin770) and a related substance in aqueous extracts of plastic materials. *J. Chromatogr. Sci*. **48**, 200-207 (2010).

19. S. Martens, J. Van Den Begin, A. Madder, F. E. Du Prez, P. Espeel, Automated Synthesis of Monodisperse Oligomers, Featuring Sequence Control and Tailored Functionalization. *J. Am. Chem. Soc*. **138**, 14182-14185 (2016).

20. R. Liu, L. Zhang, Z. Huang, J. Xu, Sequential and alternating RAFT single unit monomer insertion: model trimers as the guide for discrete oligomer synthesis. *Polym. Chem*. **11**, 4557-4567 (2020).

21. P. Roepstorff, J. Fohlman, Letter to the editors - Proposal for a common nomenclature for sequence ions in mass spectra of peptides. *Biol. Mass Spectrom*. **11**, 601-601 (1984).

22. K. Biemann, Nomenclature for peptide fragment ions (positive ions). *Methods Enzymol.* **193**, 886-887 (1990).

23. S. A. Mcluckey, G. J. Van Berkel, G. L. Glish, Tandem Mass Spectrometry of Small, Multiply Charged Oligonucleotides. *J. Am. Soc. Mass Spectrom*. **3**, 60-70 (1992).

24. M. Claeys, L. Nizigiyimana, H. Van Den Heuvel, P. J. Derrick, Mechanistic aspects of charge-remote fragmentation in saturated and mono-unsaturated fatty acid derivatives. Evidence for homolytic cleavage. *Rapid Commun. Mass Spectrom*. **10**, 770-774 (1996).

25. B. Thiede, W. Höhenwarter, A. Krah, J. Mattow, M. Schmid, F. Schmidt, P. R. Jungblut, Peptide mass fingerprinting. *Methods*. **35**, 237-247 (2005).

26. D. J. C. Pappin, P. Hojrup, A. J. Bleasby, Rapid identification of proteins by peptide-mass fingerprinting. *Curr. Biol.* **3**, 327-332 (1993).

27. F. T. Zohora, M. Z. Rahman, N. H. Tran, L. Xin, B. Shan, M. Li, DeepIso: A Deep Learning Model for Peptide Feature Detection from LC-MS map. *Sci. Rep.* **9**, 17168-17181 (2019).

28. A. Bateman *et al.*, UniProt: A hub for protein information. *Nucleic Acids Res.* **43**, 204-212 (2015).

29. N. H. Tran, R. Qiao, L. Xin, X. Chen, C. Liu, X. Zhang, B. Shan, A. Ghodsi, M. Li, Deep learning enables de novo peptide sequencing from data-independent-acquisition mass spectrometry. *Nat. Methods.* **16**, 63-66 (2019).

30. K. Biemann, Contributions of mass spectrometry to peptide and protein structure. *Biol. Mass Spectrom*. **16**, 99-111 (1988).

31. B. Domon, C. E. Costello, A systematic nomenclature for carbohydrate fragmentations in FAB-MS/MS spectra of glycoconjugates. *Glycoconj. J.* **5**, 397–409 (1988).

32. J. M. Gurevich-Messina, S. L. Giudicessi, M. C. Martínez-Ceron, G. Acosta, R. Erra-Balsells, O. Cascone, F. Albericio, S. A. Camperi, A simple protocol for combinatorial cyclic depsipeptide libraries sequencing by matrix-assisted laser desorption/ionisation mass spectrometry. *J. Pept. Sci.* **21**, 40-45 (2015).

33. M. M. Sheil, G. W. Kilby, J. M. Curtis, C. D. Bradley, P. J. Derrick, Low-energy tandem mass spectra of the cyclic depipeptide valinomycin—a comparison with four-sector tandem mass spectra. *Org. Mass Spectrom.* **28**, 574-576 (2005).

34. S. M. Williams, J. S. Brodbelt, MSn characterization of protonated cyclic peptides and metal complexes. J. Am. Soc. Mass Spectrom. **15**, 1039-1054 (2004).

35. J. G. Forsythe, S.-S. S. Yu, I. Mamajanov, M. A. Grover, R. Krishnamurthy, F. M. Fernandez, N. V Hud, F. M. Fern??ndez, N. V Hud, Ester-Mediated Amide Bond Formation Driven by Wet–Dry Cycles: A Possible Path to Polypeptides on the Prebiotic Earth. *Angew Chem Int Ed Engl*. **54**, 9871-9875 (2015).

36. S. S. Yu, M. D. Solano, M. K. Blanchard, M. T. Soper-Hopper, R. Krishnamurthy, F. M. Fernández, N. V. Hud, F. J. Schork, M. A. Grover, Elongation of Model Prebiotic Proto-Peptides by Continuous Monomer Feeding. *Macromolecules*. **50**, 9286–9294 (2017).

37. D. Doran, Y. M. Abul-Haija, L. Cronin, Emergence of Function and Selection from Recursively Programmed Polymerisation Reactions in Mineral Environments. *Angew. Chemie - Int. Ed*. **58**, 11253-11256 (2019).

38. M. A. Arnould, R. Vargas, R. W. Buehner, C. Wesdemiotis, Tandem mass spectrometry characteristics of polyester anions and cations formed by electrospray ionization. *Eur. J. Mass Spectrom*. **11**, 243-256 (2005).

39. G. Adamus, W. Sikorska, M. Kowalczuk, M. Montaudo, M. Scandola, Sequence distribution and fragmentation studies of bacterial copolyester macromolecules: characterization of PHBV macroinitiator by electrospray ion-trap multistage mass spectrometry. *Macromolecules*. **33**, 5797–5802 (2000).

40. J. Song, A. Šišková, M. G. Simons, W. J. Kowalski, M. M. Kowalczuk, O. F. Van Den Brink, LC-multistage mass spectrometry for the characterization of poly(Butylene adipate-co-butylene terephthalate) copolyester. *J. Am. Soc. Mass Spectrom*. **22**, 641-648 (2011).

41. I. Mamajanov, G. D. Cody, Protoenzymes: The case of hyperbranched polyesters. *Philos. Trans. R. Soc. A Math. Phys. Eng. Sci.* **375**, 20160357 (2017).

42. D. B. Martin, J. K. Eng, A. I. Nesvizhskii, A. Gemmill, R. Aebersold, Investigation of neutral loss during collision-induced dissociation of peptide ions. *Anal. Chem*. **77**, 4870–4882 (2005).

43. G. S. Omenn, Data management and data integration in the HUPO plasma proteome project. *Methods Mol. Biol.* **696**, 247-257 (2011).

44. S. L. Hubler, P. Kumar, S. Mehta, C. Easterly, J. E. Johnson, P. D. Jagtap, T. J. Griffin, Challenges in Peptide-Spectrum Matching: A Robust and Reproducible Statistical Framework for Removing Low-Accuracy, High-Scoring Hits. *J. Proteome Res*. **19**, 161–173 (2020).

45. K. Thalassinos, A. T. Jackson, J. P. Williams, G. R. Hilton, S. E. Slade, J. H. Scrivens, Novel Software for the Assignment of Peaks from Tandem Mass Spectrometry Spectra of Synthetic Polymers. *J. Am. Soc. Mass Spectrom*. **18**, 1324-1331 (2007).

46. I. Mamajanov, P. J. Macdonald, J. Ying, D. M. Duncanson, G. R. Dowdy, C. A. Walker, A. E. Engelhart, F. M. Fernández, M. A. Grover, N. V. Hud, F. J. Schork, Ester formation and hydrolysis during wet-dry cycles: Generation of far-from-equilibrium polymers in a model prebiotic reaction. *Macromolecules*. **47**, 1334–1343 (2014).

Acknowledgments. The authors gratefully acknowledge financial support from the John Templeton Foundation (Grant 60025), EPSRC (Grant Nos EP/L023652/1, EP/R01308X/1, EP/J015156/1, EP/P00153X/1), the Breakthrough Prize Foundation and NASA (Agnostic Biosignatures award #80NSSC18K1140), MINECO (project CTQ2017-87392-P) and ERC (project 670467 SMART-POM).

Author Contributions: LC conceived of the initial theory and hypothesis, designed the project, and coordinated the efforts of the research team. DD developed the concept, built the algorithm, software and collected data with coding help from EClarke who also undertook the synthesis of the polymers and mass spec. data collection and processing with ECarrick. GK helped with coding and CM helped coordinate the team and gave advice on the code and the manuscript. DD, EClarke, LC together wrote the manuscript with input from all the authors.

Competing Interests. The authors declare no competing interests.

Supplementary Information is available.

SUPPLEMENTARY INFORMATION

A Universal Sequencing System for Unknown Oligomers

David Doran, Emma Clarke, Graham Kennan, Emma Carrick, Cole Mathis and Leroy Cronin*

School of Chemistry, University of Glasgow, Glasgow G12 8QQ, UK

*Corresponding author: Lee.Cronin@glasgow.ac.uk

Contents

1. OLIG: Abstraction of Polymer Properties

The properties of linear polymers that are relevant to identifying and characterising oligomers from mixtures can be broken down into three categories:

1. Intrinsic chemical / cross-reactivity properties.
2. MS ionization properties.
3. MS^n fragmentation pathways.

1.1 Polymer Properties: Chemistry and Cross-Reactivity

Information on polymer-specific chemical properties is essential to build and constrain *in silico* libraries for screening of heterogeneous product mixtures. The following chemical properties are required for construction of screening libraries:

1. Monomer neutral masses.
2. Monomer cross-reactivities.
3. Elongation unit: what is the minimum number of monomer units added to each elongating oligomer chain.
4. Elongation mass difference: the mass gained or lost upon addition of a single monomer to an elongating chain, relative to the sum of all individual free monomers in a sequence.
5. Maximum and minimum possible sequence length.
6. Side chain covalent modifications.
7. Terminal covalent modifications.

1.2 Polymer Properties: MS / Precursor Ionization

Having enumerated all possible sequences formed from inherent physical and chemical restraints for a polymer class, parameters relevant to how products may ionize must then be defined. This is dependent on not only the properties of the polymer, but also the ion source used:

1. Polymer intrinsic ions: *non-exchangeable* ions that are universal to all products of a given polymer class and are not sidechain-specific.

2. Polymer adducts: *exchangeable* ions associated with sequences in a given polymer class that are also not side sidechain-specific.
3. Side chain intrinsic ions: *non-exchangeable* ions formed at specific monomer sidechain functional groups.
4. Side chain adducts: *exchangeable* ions associated with specific monomer sidechain functional groups.
5. Polymer symmetry: are termini equivalent?

1.3 Polymer Properties: MS^n Fragmentation

In order to distinguish between isomeric precursors via tandem mass spectrometry, properties relevant to MS^n fragmentation pathways must be defined. Like MS ionization, these are dependent on both polymer class and instrumentation (in this case ion source, mass analyser(s) and fragmentation method). A single polymer class will often undergo competing fragmentation pathways, the relative predominance of which will vary between sequences and instruments. Oligomersoup defines each fragmentation pathway individually with the following properties:

1. Fragmentation mass difference: the mass gained or lost after a single fragmentation event, relative to the neutral mass of the equivalent intact sequence.
2. Indexing: the terminus from which the fragment series is indexed. This is only relevant for fragmentation of asymmetric polymers.
3. Start index: the position along the polymer backbone at which the first detectable fragmentation event occurs.
4. End index: the position along the polymer backbone at which the final detectable fragmentation event occurs.
5. Fragmentation unit: the number of monomers dissociated from the fragmenting backbone after each fragmentation event.
6. Intrinsic fragment charge: the default charge state of fragments produced via a given fragmentation pathway that are due to *non-exchangeable* ions.
7. Intrinsic fragment adducts: the default charge state and additional mass produced via a given fragmentation pathway that are due to *exchangeable* ions.

8. Permissible adducts: specific, non-intrinsic *exchangeable* ions that may be associated with fragment ions. These can displace intrinsic adducts, but their effects on charge state and mass are cumulative with intrinsic fragment charge.

2. OLIG: Application and Experimental Constraints

The aforementioned abstract properties determine the possibilities available for ionization and fragmentation of a polymer. For many polymer classes, there are a large number of possibilities for ionization and fragmentation. However, which of these possibilities is more likely – and most relevant to confidently identifying specific sequences – will usually depend on experimental constraints. Oligomersoup therefore enables a wide range of experimental parameters to be defined, depending on instrumentation used to obtain data and the properties of specific analytes:

1. Available adducts: *exchangeable* ions that may be associated with precursor and / or fragmented product ions. Which specific adducts are available will depend on what ions are present in the analyte.
2. Mass and charge range: the *m/z* range detectable will be directly dependent on the instrument used and its configuration, which will often vary between experiments.
3. Minimum and maximum sequence length: this will depend on the source of the products and their solubility / ease of introduction into the ion source.
4. Constituent monomers and sequence distribution: screening libraries are constrained by monomers present in analyte and can also be further constrained by searching for only a pre-defined list of specific sequences.
5. Permissible fragmentation pathways: predominant fragmentation pathways will vary depending on instrumental set-up.
6. Core fragmentation pathways: which of the permissible fragmentation pathways to use for confidence assignment.
7. Product abundance: depending on product ionization and mass analyser sensitivity, constraints can be placed on the minimum expected abundance for detected ions that correspond to products.
8. Product separation: retention times of products can be predicted and constrained for experiments, particularly when pre-measurement chromatographic separation is used.

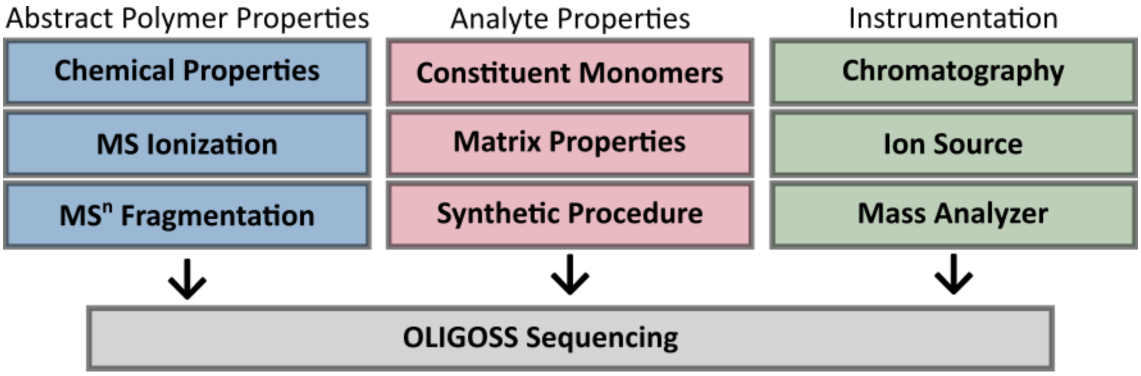


Figure 1: Properties Relevant to De Novo Sequencing in Oligomersoup. Abstract Polymer Properties define the possible bounds of sequence space, MS ionization and MS^n fragmentation for a polymer class. Analyte Properties constrain these possibilities in analyses by specifying which monomers and adducts may be present, and expected sequence distributions. Finally, separation of products and favoured ionization and fragmentation pathways are determined by Instrumentation Properties. Combining all of these properties enables full de novo sequencing of linear oligomer mixtures.

3. Defining Abstract Polymer Properties: Constructing a Polymer Configuration File

3.1 General Chemical Properties

Monomer neutral masses, functional groups and cross-reactivities are essential for constructing *in silico* libraries for screening. Each monomer is assigned a one-letter code, with its neutral monoisotopic mass defined. The type and number of reactive functional groups must also be defined (Figure 2).

```
"MONOMERS": {
       "A": [89.04768, [["amine", 1], ["carboxyl", 1]]],
       "C": [121.01975, [["amine", 1], ["carboxyl", 1]]]
```

Figure 2: Example Monomer Definitions. Two monomers, the amino acids alanine and cysteine, are assigned one-letter codes, with associated monoisotopic neutral masses and functional groups.

To constrain sequence space and rule out chemically infeasible sequences, functional group cross-reactivities must also be defined. Each functional group is assigned an identifier string, and cross-reactive groups are specified along with one-letter codes of monomers containing at least one of the functional group available for reaction (Figure 3).

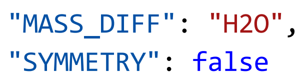
```
"REACTIVITY_CLASSES": {
    "amine": [["carboxyl", "hydroxyA"], ["A", "C", "D", "E", "F", "G", "H",
        "I", "K", "L", "M", "N", "P", "Q", "R", "S", "T", "V", "W", "Y"]],

    "carboxyl": [["amine", "hydroxyA"], ["A", "C", "D", "E", "F", "G", "H",
        "I", "K", "L", "M", "N", "P", "Q", "R", "S", "T", "V", "W", "Y"]],

    "hydroxyA": [["amine", "carboxyl", "hydroxyA"], ["g"]]
}
```

Figure 3: Example Reactivity Classes. Three reactive functional groups (amine, carboxyl and hydroxy acid) are assigned identification strings. Cross-reactive groups are defined, with list of one-letter codes corresponding to monomers containing one or more functional groups.

For polymer classes with equivalent termini, palindromic sequences are equivalent. This is not the case for polymer classes with non-equivalent termini. Therefore, to avoid screening for redundant palindromic sequences when dealing with symmetric linear polymers or discarding non-equivalent palindromic sequences for asymmetric linear polymers, polymer class symmetry must also be defined (Figure 4). Finally, to calculate the neutral mass of each sequence, the mass gained – or lost – upon addition of a monomer must be defined (Figure 4).



```
"MASS_DIFF": "H2O",
"SYMMETRY": false
```

Figure 4: Example Mass_Diff and Symmetry. In this case, the polymer class is an asymmetric condensation polymer

3.2 MS Ionization Properties

General chemical properties of a polymer class are used to constrain what sequences could be present in analytes. However, MS ionization properties must be defined in order to screen for these sequences in MS data. It is assumed by default that products are compatible with all potential adducts (exchangeable ions) present in an analyte, and that all oligomers can ionize either via intrinsically charged monomers (if present) or via adduct association. However, specific adducts can be excluded if required. In addition, side chain adducts can be specified if adduct formation is expected to occur at specific monomer side chains (Figure 4).

```
    "IONIZABLE_SIDECHAINS": {
        "K": {
            "pos": ["H", 1, 1],
            "neg": null},
        "R": {
            "pos": ["H", 1, 1],
            "neg": null},
        "E": {
            "pos": null,
            "neg": ["-H", 1, 1]},
        "D": {
            "pos": null,
            "neg": ["H", 1, 1]},
        "H": {
            "pos": ["H", 1, 1],
            "neg": null}
    }
```

Figure 5: Example Side Chain Ionization Events. Potentially cationic side chains (K, R, H) can ionize via a proton adduct. Potentially anionic side chains (D, E) can ionize via proton abstraction.

In tandem mass spectrometry of ions produced in soft ionization sources (ESI, MALDI), it is assumed that minimal fragmentation occurs upon formation and detection of precursors.[1] However, in some sources in-source fragmentation can occur prior to fragmentation via tandem MS.[2] The most common form of such fragmentation events leads to loss of neutral species, with the resulting ions of reduced mass being referred to as neutral loss products.[3] As the likelihood of these events and the expected mass losses is dependent on what functional groups are present in an ion,[3] neutral losses can be predicted and incorporated into MS screening libraries. Sidechain-specific neutral losses can be specified for each monomer (Figure 6).

```
    "LOSS_PRODUCTS": {
        "S" : ["H2O"],
        "T": ["H2O"],
        "E": ["H2O"],
        "D": ["H2O"],
        "N" : ["NH3", "H2O"],
        "R": ["NH3"],
        "K": ["NH3"],
        "Q" : ["NH3"]
}
```

Figure 6: Example Neutral Losses for Monomer Side Chains. Lists of potential neutral losses are specified for each monomer.

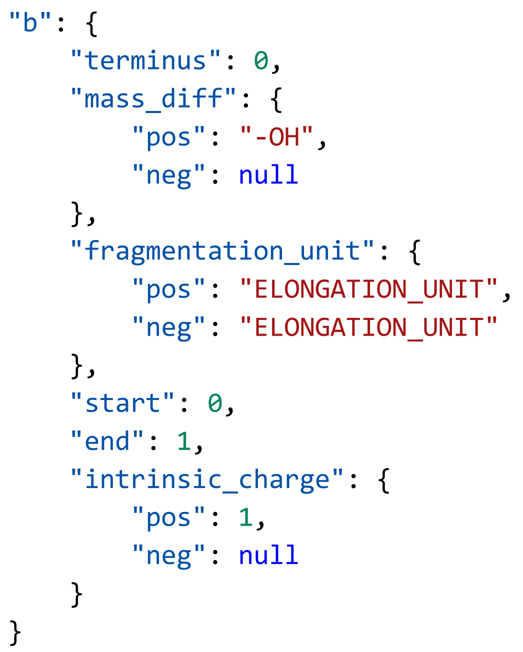
3.3 MS^n Fragmentation Properties

MS ionization properties are often sufficient for compositional characterization of oligomers.[4] However, as isomeric sequences will usually share a common precursor *m/z*, this is insufficient for identifying which isomeric sequences are present in the analyte. In many cases, isomeric precursors can be identified via distinct MS$^{2\text{-}n}$ fragments. To screen for such fragments, potential fragmentation pathways for a polymer class must be well defined. Polymer fragmentation events generally fall into one of two categories: backbone fragmentation and sidechain fragmentation.[5]

3.3.1 Backbone Fragmentation Pathways

Fragmentation along the polymer backbone is usually most informative for sequencing. Therefore, many fragmentation pathways can be represented by fragment series which are indexed along the backbone. For asymmetric polymer classes, each fragment series is indexed from a specific “home” terminus. In some cases, the relative positions of where the fragment series initiates and terminates must also be specified, as many backbone fragmentation events will only occur as internal cleavages under specific conditions, rather than running the full length of the backbone.

In addition, the shift in mass (***not*** shift in *m/*z) of fragments (relative to the intact sequence equivalent) must be specified for each fragment series. This should include any neutral species lost in the fragmentation event, as well as any intrinsic ions or adducts. It is important to note that intrinsic ions and adducts, where applicable, must also be specified separately for a fragment series. These will usually differ depending on precursor charge. Examples of two configuration file entries for peptidic fragment series are shown for b fragments (Figure 7) and y fragments (Figure 9).



```
"b": {
    "terminus": 0,
    "mass_diff": {
        "pos": "-OH",
        "neg": null
    },
    "fragmentation_unit": {
        "pos": "ELONGATION_UNIT",
        "neg": "ELONGATION_UNIT"
    },
    "start": 0,
    "end": 1,
    "intrinsic_charge": {
        "pos": 1,
        "neg": null
    }
}
```

Figure 7: Peptide b fragment configuration

3.3.2 Sidechain Fragmentation Pathways

Many polymer classes also have side chain structures which are prone to fragmentation (e.g. peptides and proteins). In addition to side chain-specific neutral losses, many monomer side chains have unique signature ions.[6] The favourability of the fragmentation pathways that produce these signature ions can vary dramatically between both polymer classes and individual monomers within the same polymer class, meaning some signatures are much more likely to appear with high abundance than others. Monomers with predominant side chain signatures can added to a list of “dominant” signatures, the presence of which can then be used to validate – or refute – sequencing and compositional assignments (Figure 8).

```
"Im": [
            ["F", 120.0813], ["D",88.0399], ["E",102.0555],
            ["I", 86.0969], ["L", 86.0969], ["H", 110.0718],
            ["C", 76.0221, 133.0436, 134.0276, 147.0772], ["K", 101.1079],
            ["S", 60.0449], ["Y", 136.0762], ["V", 72.08133],
            ["T", 74.06059], ["A", 44.05003], ["M", 104.0534, 120.0483],
            ["Q", 101.0715], ["P", 70.06568], ["N", 87.05584, 70.02864]
        ],

        "dominant": ["L", "I", "F", "P", "H", "Y"]

    }
```

Figure 8: Example Signature Ions.

3.3.3 Exceptions to Standard Fragmentation Rules: Mixed Backbones

The backbone fragmentation events described above depend on the type of backbone linkage being fragmented. For polymer classes with mixed backbone linkages this can naturally lead to some properties of a backbone fragment series varying at specific indices, depending on the type of bond present at that index. This must be accounted for when constructing a configuration file for mixed backbone polymer classes. In cases where the backbone linkages vary in a consistent and repeating pattern (e.g. every 2nd backbone link is of a different type) this can be accounted for by updating the fragmentation unit of individual series. However, for random copolymers with unpredictable linkage positions, exceptions must be hard coded into the fragment series configuration. It is important when constructing exception properties for fragment series that specified monomer functional groups are correct in the configuration file. This is because Oligomersoup uses monomer functional groups to work out which bonds are relevant for fragmentation exceptions.

Oligomersoup can apply exceptions to fragmentation mass difference, intrinsic adducts and intrinsic ions. This will depend on whether the final monomer in the fragment is linked by a specific functional group (e.g. ester linkages for depsipeptides). The range of indices at which the exceptions apply must also be specified, as these may not be relevant for the whole length of the fragmenting backbone. An example of a fragment series with exceptions is the peptidic y fragment, which fragments differently at ester linkages caused by the presence of hydroxy acid (“hydroxyA”) functional groups (Figure 9, Figure 10).

```
"y": {
           "terminus": -1,
           "mass_diff": {
               "pos": "H",
               "neg": "-H"
           },
           "fragmentation_unit": {
               "pos": "ELONGATION_UNIT",
               "neg": "ELONGATION_UNIT"
           },
           "start": 1,
           "end": 0,
           "intrinsic_adducts": {
               "pos": "H",
               "neg": "-H"
           },
           "exceptions": {
               "mass_diff": {
                   "pos": {
                       "-1": {
                           "hydroxyA": {
                               "mass_diff_value": 26.98709,
                               "start": 0,
                               "end": 1
                           }
                       }
                   }
               },
               "intrinsic_ions": {
                   "pos": {
                       "-1": {
                           "hydroxyA": {
                               "intrinsic_adduct": null,
                               "intrinsic_charge": 1,
                               "start": 0,
                               "end": 1
                           }
                       }
                   }
               }
           }
       }
```

Figure 9: Peptide y fragment configuration.

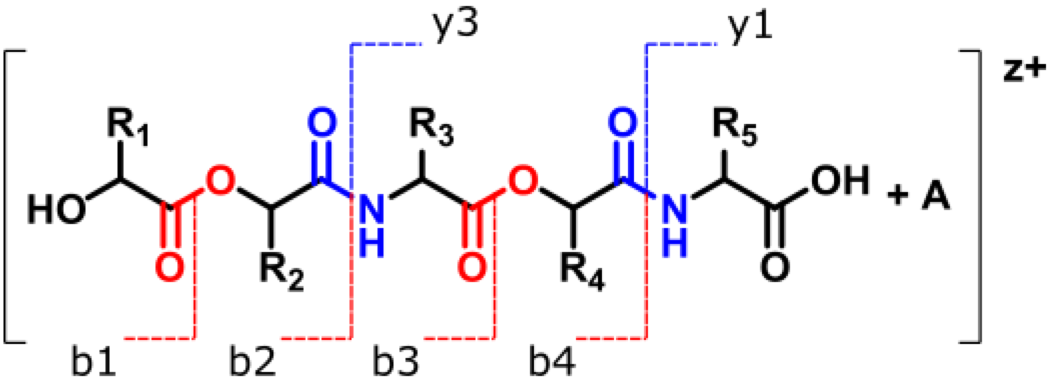


Figure 10: MS2 Fragment of an Example Cationic Depsipeptide. Backbone linkages are shown in blue (amide) and red (ester). Due to the exception applied for fragmentation of ester bonds for y-fragments, only y fragments that result from fragmentation of an amide (y1 and y3) are produced. No exceptions apply to b fragments for ester fragmentation.

3.4 Covalent Modifications

Covalent modifications, such as end-groups and side chain attachments can be screened using Oligomersoup. As these modifications can potentially affect MS ionization and MS$^{2\text{-}n}$ fragmentation, they must be added to the polymer class configuration file. Each configuration file is denoted a three-letter code, with termini and side chains it can target specified (*Figure 11*).

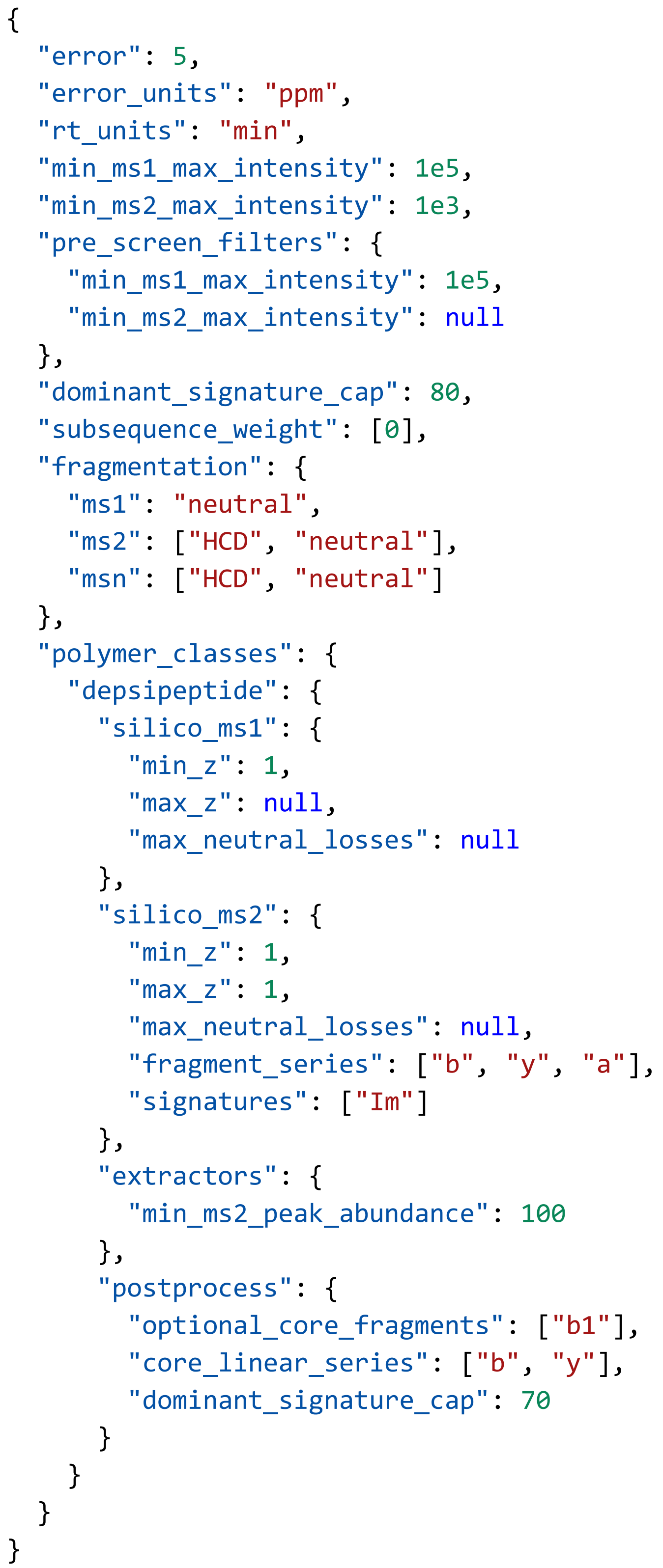
```
"Ole": {
            "mass": 282.25589,
            "termini": [0],
            "side_chains_attachments": ["K", "R"],
            "free_mod_fragments": {
                "pos": [265.2532],
                "neg": [281.2181]
            },
            "mass_diff": {
                "ms1": "H2O",
                "ms2": "H2O"
            },
            "universal_ms2_shift": false
        }
```

Figure 12: Example Covalent Modification for Oleic Acid (“Ole”).

4. Instrument Configuration Files

Polymer-specific configuration files define the full scope of possible ionization and / or fragmentation pathways for sequences of a specific polymer class. However, the relative predominance of these pathways will usually depend on the capabilities and operating mode of the mass spectrometer, as well as other experiment-specific constraints (see section *Running Oligomersoup: Input Parameters*).

Oligomersoup therefore enables pre-configuration files for specific instruments. Default values for silico, extractor and postprocess parameters can be set for specific instruments and also instrument-polymer combinations. An example of instrument-specific configuration file is shown in Figure 12 for the Orbitrap Lumos Tribrid used in our group. Non-polymer specific parameters such ass error, intensity thresholds and fragmentation modes can be set, as well as polymer-specific parameters which govern default values for certain inputs when this instrument is used in conjunction with products of a specific polymer class (in this case depsipeptides).



```
{
  "error": 5,
  "error_units": "ppm",
  "rt_units": "min",
  "min_ms1_max_intensity": 1e5,
  "min_ms2_max_intensity": 1e3,
  "pre_screen_filters": {
    "min_ms1_max_intensity": 1e5,
    "min_ms2_max_intensity": null
  },
  "dominant_signature_cap": 80,
  "subsequence_weight": [0],
  "fragmentation": {
    "ms1": "neutral",
    "ms2": ["HCD", "neutral"],
    "msn": ["HCD", "neutral"]
  },
  "polymer_classes": {
    "depsipeptide": {
      "silico_ms1": {
        "min_z": 1,
        "max_z": null,
        "max_neutral_losses": null
      },
      "silico_ms2": {
        "min_z": 1,
        "max_z": 1,
        "max_neutral_losses": null,
        "fragment_series": ["b", "y", "a"],
        "signatures": ["Im"]
      },
      "extractors": {
        "min_ms2_peak_abundance": 100
      },
      "postprocess": {
        "optional_core_fragments": ["b1"],
        "core_linear_series": ["b", "y"],
        "dominant_signature_cap": 70
      }
    }
  }
}
```

Figure 12: Instrument Configuration File for Orbitrap Lumos Tribrid Mass Spectrometer.

5. Running Oligomersoup: Input Parameters

There is a total of 47 unique parameters that can be passed in to Oligomersoup to execute a sequencing workflow. These parameters can be broken down into four categories:

1. Core parameters
2. Silico parameters
3. Extractor parameters
4. Postprocessing parameters

In each Oligomersoup run, an input parameters JSON and MS/MS data files in mzmlripper format must be passed in to execute. The input parameters (experiment run file) then calls on pre-configured instrument- and polymer-specific configuration files to execute a sequencing workflow.

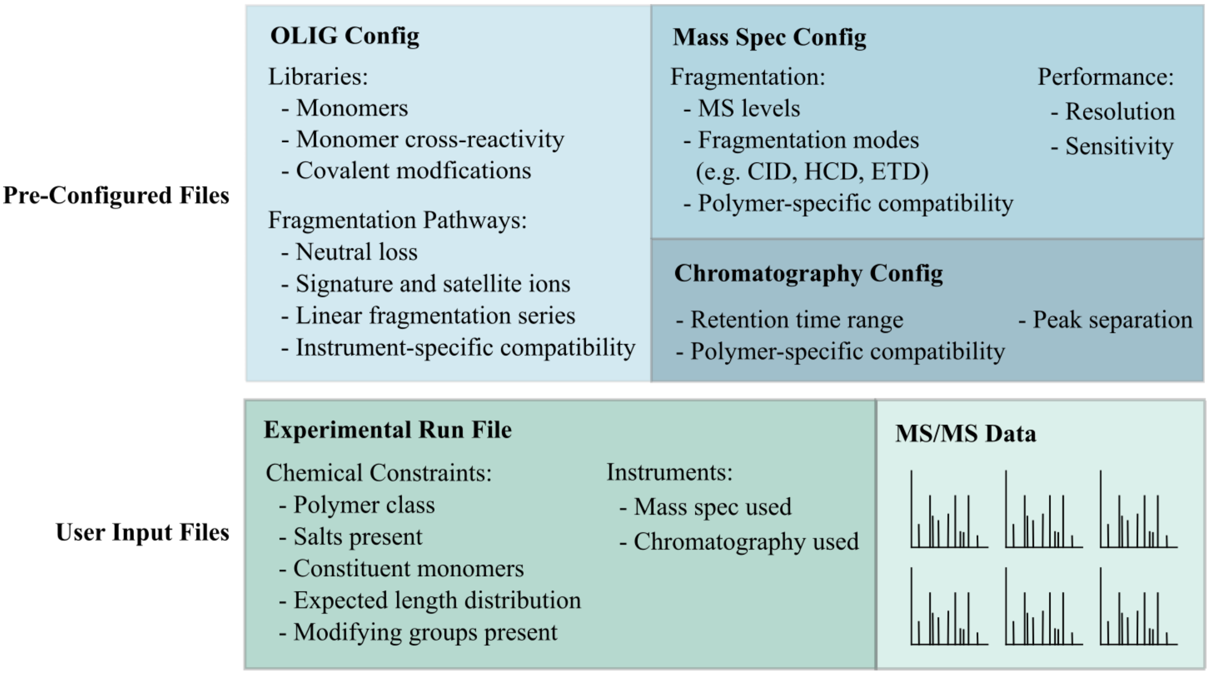


Figure 13: Oligomersoup Input files.

5.1 Core Parameters

The six core parameters are required in every sequencing run. They define essential properties used by all Oligomersoup modules:

1. “mode”: defines overall charge of ions in data. Either “pos” or “neg” for positive and negative mode, respectively.
2. “monomers”: list of monomer codes, detailing which constituent monomers (or suspected monomers) are present in products.
3. “polymer_class”: defines the type of polymer class to which oligomeric products belong.
4. “instrument”: specifies the mass spectrometer used to obtain data. This is essential for retrieving instrument-specific default values for products of the chosen polymer class. If this is not defined, all parameters must be explicitly stated.
5. “chromatography”: specifies pre-chromatographic separation method used in experiment.
6. “screening_method”: this determines which sequencing workflow will be executed by Oligomersoup.

5.2 Silico Parameters

Silico parameters are used by the silico module to build theoretical MS1 precursor and MS2 product ion libraries. There are three kinds of silico parameter: general, ms1 and ms2, which relate to general, ms1- and ms2-specific silico operations respectively.

General silico parameters:

1. “max_length”: determines maximum length (in monomer units) of potential sequences.
2. “min_length”: determines minimum length (in monomer units) of potential sequences.

3. “isomeric_targets”: specifies list of target sequences. If a value for this parameter is supplied, only sequences isomeric to one or more targets will be screened for in sequencing workflows.

4. “modifications”: specifies any covalent modifications that may target specific termini and / or sidechains.

MS1 silico parameters:

1. “adducts”: specifies which extrinsic ions (e.g. H^+, Na^+) may be present in analytes.

2. “min_z”: specifies minimum charge for MS1 precursors.

3. “max_z”: specifies maximum charge for MS1 precursors.

4. “max_neutral_losses”: specifies maximum number of sidechain-specific MS1 neutral loss fragmentation events per sequence.

5. “universal_sidechain_modifications”: specifies whether, in the presence of covalent modifications (specified by “modifications” parameter in general silico parameters) all sidechain targets are universally modified by one or more modifying group.

6. “universal_terminal_modifications”: specifies whether, in the presence of covalent modifications, all terminal targets are modified by one or more modifying group.

MS2 silico parameters:

1. “fragment_series”: specifies which linear fragment series to include in theoretical MS2 product ion libraries.

2. “adducts”: specifies extrinsic ions expected to be associated with product ions (defaults to silico.ms1.adducts if not specified).

3. “max_neutral_losses”: specifies maximum number of sidechain-specific MS2 neutral loss fragmentation events per sequence.

4. “signatures”: specifies which signature ion types may be present in MS2 product ion spectra (e.g. immonium ions for amine-containing fragments in HCD).

5. “min_z”: specifies minimum charge of MS2 product ions.

6. “max_z”: specifies maximum charge of MS2 product ions.

5.3 Extractor Parameters

Extractor parameters are used by the extractors module to filter and screen data for MS1 precursor and MS2-product ion associations. Extractor parameters:

1. “error”: specifies error threshold for matching theoretical *m/z* values to observed data. This can be supplied in relative units of parts per million (ppm) or absolute mass units (u).
2. “error_units”: specifies the units for error thresholding (“ppm” or “u” for relative and absolute error, respectively).
3. “min_rt”: minimum retention time (in minutes) for peaks associated with target sequences.
4. “max_rt”: maximum retention time (in minutes) for peaks associated with target sequences.
5. “rt_units”: specifies retention time units in raw mzML file. This is vendor-specific and should be set in instrument configuration files.
6. “min_ms2_peak_abundance”: specifies the minimum relative intensity of the most intense peak associated with the target sequence in MS2 spectra.
7. “min_ms1_total_intensity”: specifies minimum total MS1 ion current associated with a target sequence’s precursors.
8. “min_ms2_total_intensity”: specifies minimum total MS2 ion current associated with a target sequence’s MS2 product ions
9. “min_ms1_max_intensity”: specifies minimum peak of intensity for a sequence’s MS1 precursor ions.
10. “min_ms2_max_intensity”: specifies minimum peak of intensity for a sequence’s MS2 product ions.

11. "pre_screen_filters": specifies any criteria that spectra must meet to be considered for screening:

    a. "min_ms1_max_intensity": specifies minimum absolute intensity of MS1 spectra.

    b. "min_ms2_max_intensity": specifies minimum absolute or relative intensity of MS2 spectra.

    c. "min_ms1_total_intensity": specifies minimum total ion current for individual MS1 spectra.

    d. "min_ms2_total_intensity": specifies minimum total ion current for individual MS2 spectra.

    e. "min_rt": specifies minimum retention time for valid MS1 and MS2 spectra.

    f. "max_rt": specifies maximum retention time for valid MS1 and MS2 spectra.

    g. "precursors": specifies specific precursor *m/z* values. If specified, MS2 spectra will be filtered for one more target precursors before screening.

5.4 Postprocessing Parameters

Postprocessing parameters are used by the postprocessing module to carry out final data manipulation and assign confidence values to confirmed sequences. Postprocessing parameters:

1. "exclude_fragments": list of specific fragment ids to exclude from confidence calculations under all circumstances.

2. "optional_core_fragments": list of specific fragment ids to exclude from confidence calculations **if** they have not been confirmed for target sequences.

3. "dominant_signature_cap": upper limit on confidence assignments for sequences with one or more dominant signature ion missing from observed data (e.g. aromatic immonium ions).

4. "essential_fragments": list of specific fragment ids that must be confirmed for target sequences.

5. “subsequence_weight”: weighing factor (S) to assign to mean continuous fragment coverage.

6. “core_linear_series”: list of fragment series types to use in confidence calculations. An equal weighting will be applied to all core series.

7. “rt_bin”: specifies minimum gap between peaks (in minutes) for MS1 EICs.

8. “ms2_rt_bin”: specifies binning region for MS2 spectra relative to associated MS1 precursor peak(s).

9. “spectral_assignment_plots”: specifies whether to output annotated spectra for individual sequencing hits.

10. “min_plot_confidence”: specifies minimum confidence score for sequence hits to be plotted.

6. Linear Peptide Standards

To test the ability of Oligomersoup to distinguish between isomeric sequences, pure peptide standards of various lengths were synthesised using an automated peptide synthesiser. Six standards were chosen: ASGNQ, FSGNQ, GSGNQ, ASGNQSGV, FSGNQVGS, and GSGNQVGS. These sequences were selected due to the presence of loss product-prone sidechains and prominent signature ions. A blind sequencing run was performed in which both the minimum and maximum sequence length was set to the length of the standard sequence and no limit was placed on the number of possible neutral loss products. Confidence assignments were calculated at a range of subsequence weights (S = 0, 0.25, 0.5, 0.75, 1). The mean confidence assignment for the target sequence was inversely proportional to S for all six standards (Figure 11).

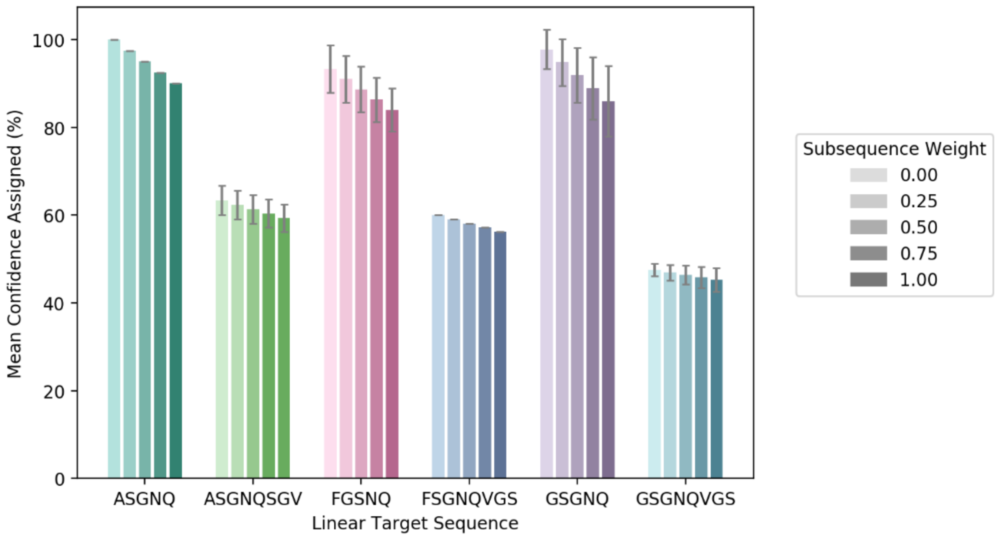


Figure 14: Effect of Subsequence Weighting on Target Sequence Confidence Assignment. Data show mean of 5 replicates ± 1 S.D.

For all the standards tested, false assignments were made for both isomeric and non-isomeric sequences. Whilst in some cases, the number of false assignments was numerically large, these false assignments represented a tiny proportion of the possible sequence space. For example, Oligomersoup identified 184 isomeric sequences and around 795 non-isomeric sequences for the linear standard GSGNQVGS (subsequence weight = 0.5) but the number of possible isobaric and non-isobaric sequences was 10, 080 (1.83 %) and 1.67 x 10^6 (0.05 %) respectively. This demonstrated the ability of Oligomersoup to successfully identify specific sequences from large product pools.

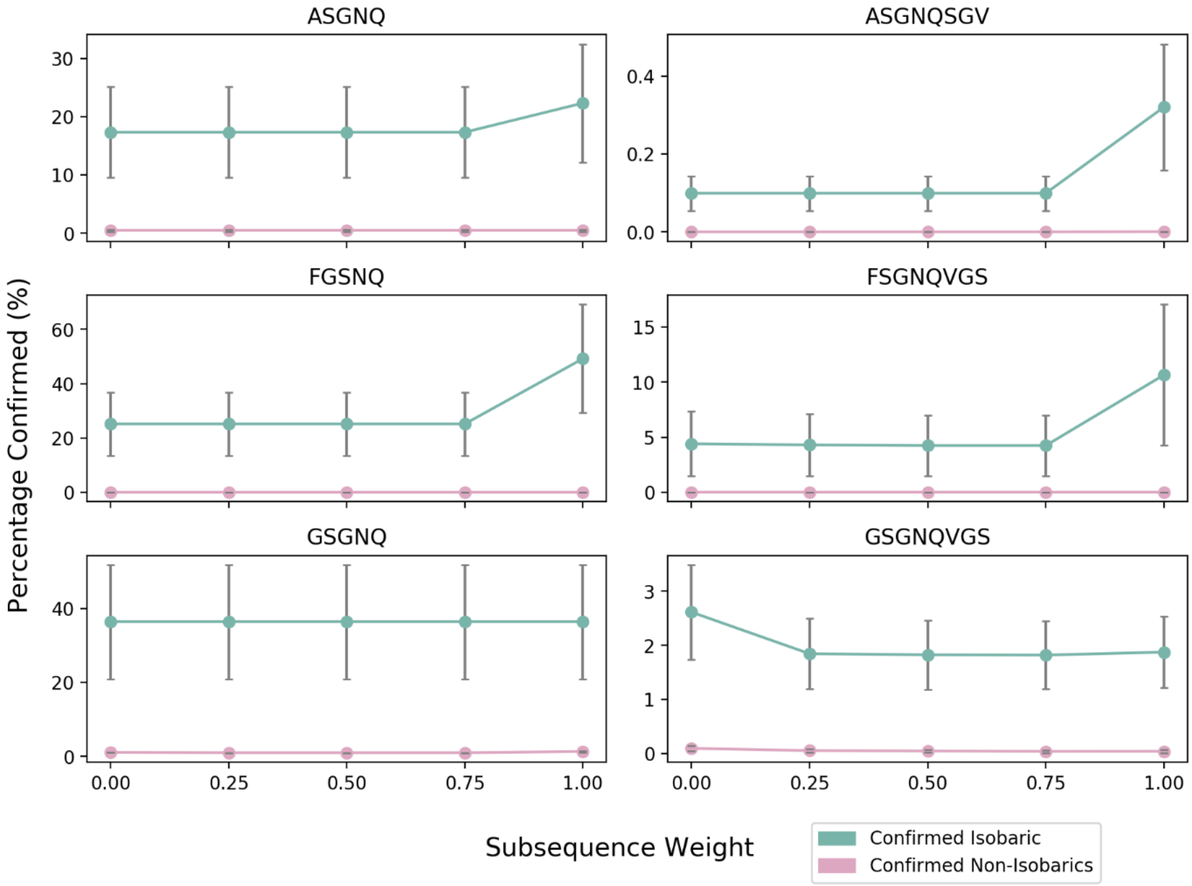


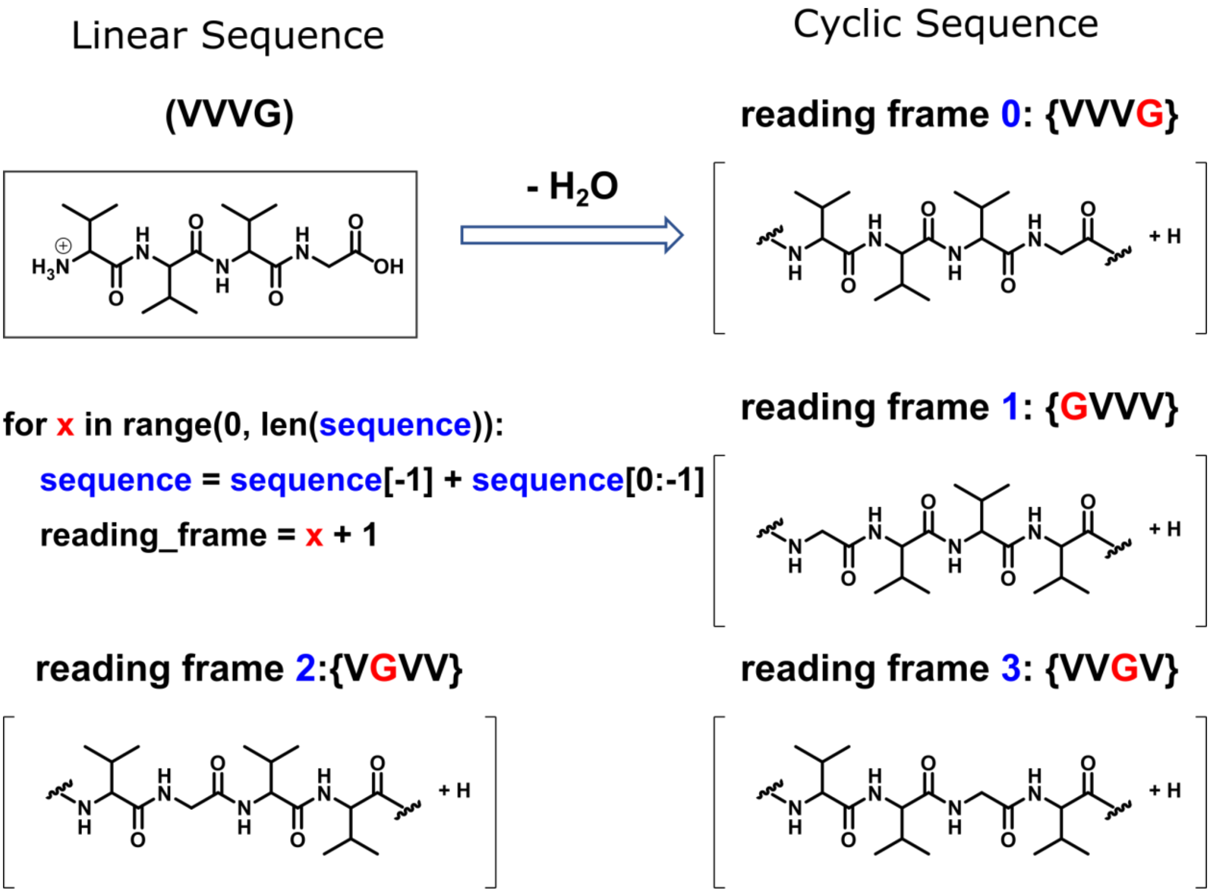
Figure 15: Effect of Subsequence Weighting Factor (S) on Confidence Assignment of Linear Standards. The percentage potential isomeric and non-isomeric sequences confirmed with a confidence greater than or equal to the target sequence as a function of S in a blind sequencing run. Data represent mean of 5 replicates ± 1 S.D.

7.Limitations and Edge Cases

7.1 Head-to-Tail Cyclized Sequences

Having demonstrated the capabilities of Oligomersoup for *de novo* sequencing of linear oligomers, attempts were made towards characterisation and sequencing of branched and cyclic sequences. Previous attempts at identifying cyclic depsipeptide oligomers in mixtures have relied heavily on pre-measurement separation via ion mobility.[7]. However, cyclic sequences have potential signatures in standard ESI-MS and CID.[8]

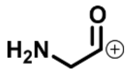
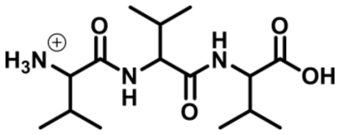
For any given cyclic sequence, assuming little or no in-source fragmentation, the precursor mass is expected to be equivalent to the full linear equivalent mass with an additional mass_diff subtracted (an extra condensation in the case of peptides, depsipeptides and other condensation polymer classes). Discounting internal cleavage events, our rationale for assigning potentially cyclic sequences relies on the principal of reading frame shifts. As there is no way to predict the position of ring opening, a linear fragment series can begin from any point in the cyclic sequence. Therefore, in our workflow, in the assignment of potentially cyclic fragments, a series of reading frame shifts are carried out when fragmenting the proposed cyclic precursor. A reading frame is defined as the equivalent linear sequence of a cyclic target, starting from a particular point in the cyclic sequence. To carry out a reading frame shift, the final monomer of a sequence is 'shifted' to the first index of the sequence, creating a new reading frame. Each reading frame has a linear sequence equivalent, with unique fragments assigned in a previous step in the workflow. If the reading frame shift has any unique fragments, these are assigned as 'shifted' or cyclic fragments, with annotation defined by the number of reading frame shifts from the starting sequence. For example, if the y3 fragment of the first reading frame shift is unique to that shifted sequence, the cyclic or 'shifted' fragment is assigned to a cyclic precursor in the following notation: 'y3-1'. This process is illustrated with an example proposed cyclic peptide sequence in Figure 29, with cyclic fragments for this sequence given in Table 3.



```
for x in range(0, len(sequence)):
    sequence = sequence[-1] + sequence[0:-1]
    reading_frame = x + 1
```

Figure 16: Cyclic Sequence and Corresponding Reading Frame Shifts. Target cyclic sequence has 4 unique reading frames. Each reading frame shift is carried out by shifting the final monomer of a sequence to the first monomer.

Table 3: Unique Reading Frame Shift Fragments for Proposed Cyclic Sequence {VVVG}.

Shifted Fragment	**Structure**
b1-1	
y3-1	

To test this reading frame shift method for identification of cyclic sequences, a standard of the linear depsipeptide valinomycin was used in a blind sequencing run. An *in silico* library containing valinomycin and all of its 18,400 isomeric sequences was screened in a blind run.

Of four unique valinomycin reading frames, none had any fragments that were unique to all sequences within that frame (Figure 30). Therefore, the requirement of a shifted fragments to be unique to a particular reading frame was relaxed, and instead all reading frame shifts were permissible provided that one or more shifted fragment is not found in the equivalent linear sequence. In a blind run screening all 18,400 valinomycin isomers and 18,400 linear equivalents, no positive hits were found for linear sequences and only 93 were found amongst the cyclic isomers, including that target valinomycin. Thus, it was established that Oligomersoup has some ability to distinguish between linear and cyclic depsipeptide sequences.

However, further investigation will be required to determine whether the reading frame shift method is suitable for identification of cyclic sequences for other polymer classes. The goal of Oligomersoup is to provide a near-universal sequencing tool for linear oligomers, with the only requirement being products that can ionize under a soft ionization source (e.g. ESI or MALDI) and be induced to fragment under MS/MS. Therefore, we have opted not to include the reading frame shift in any of the available Oligomersoup sequencing workflows until it has been demonstrated to work for other polymer classes.

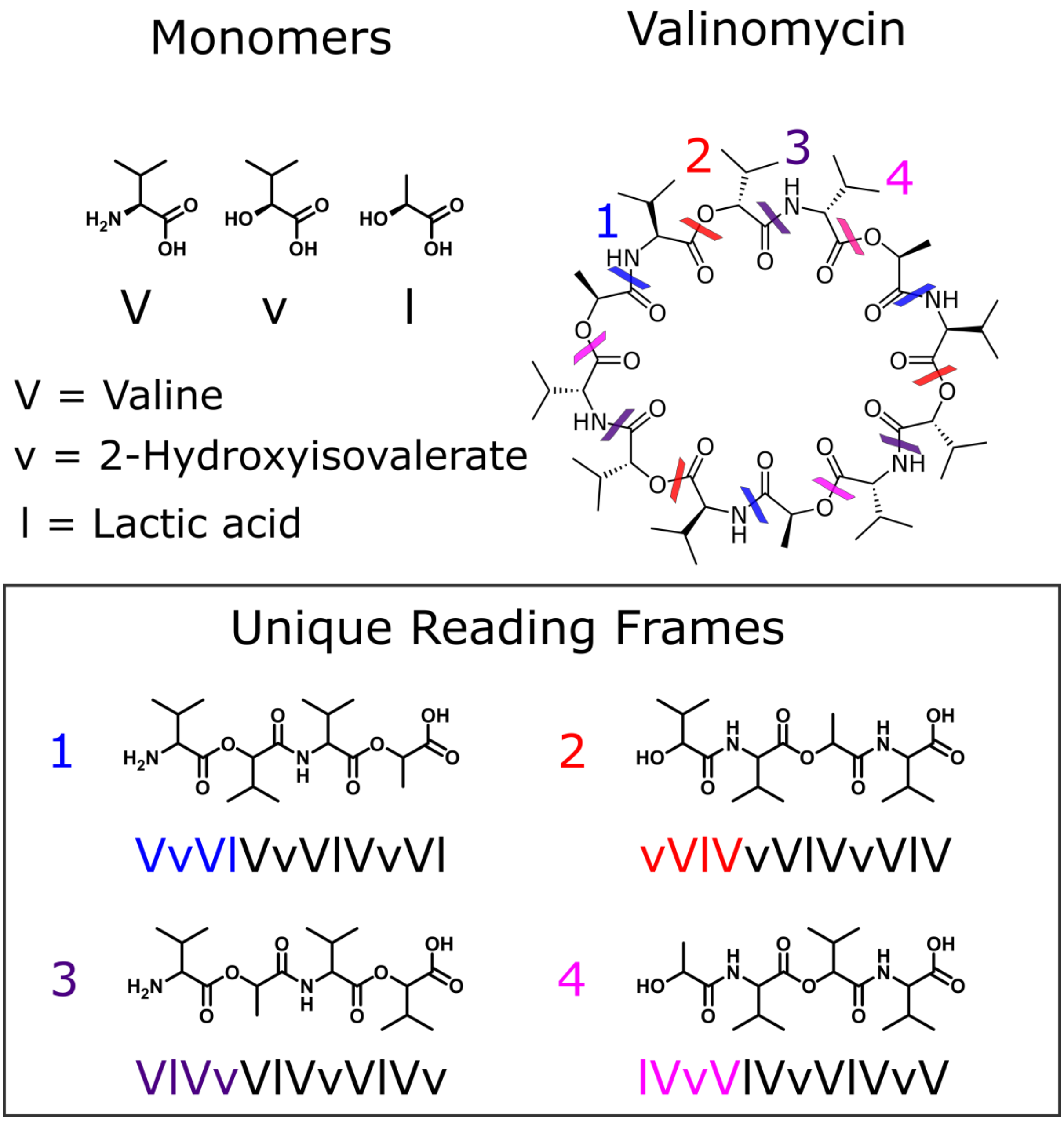


Figure 17: In Silico Fragmentation of Cyclic Depsipeptide Valinomycin. Sequence is split into 4 unique reading frames, each of which is made up of a repeating tetramer sub-sequence.

As mentioned previously, the reading frame shift method for assigning cyclic sequences relies partly on an extra precursor mass_diff relative to the intact linear equivalent sequence. For diverse oligomers with a range of sidechain functionalities (such as peptides), in-source neutral loss fragmentation events can have the same effect. In the case of depsipeptides, residues with a hydroxyl sidechain (S, T, D, E) can undergo a water loss, equivalent to an extra mass_diff for condensation polymers such as depsipeptides. For monomers with amine and / or carboxylate sidechain functional groups, the potential for forming extra links provide an additional source of loss products as well as alternative, branching backbone architectures that may be falsely assigned as cyclic.

As can be seen in Figure 31, additional neutral losses (which may be a result of head-to-tail cyclization, head-to-sidechain cyclization, sidechain branching or in-source fragmentation) were observed for products of unconstrained oligomerization of amino acids with hydroxyl, carboxylate and amine sidechains (Figure 31).

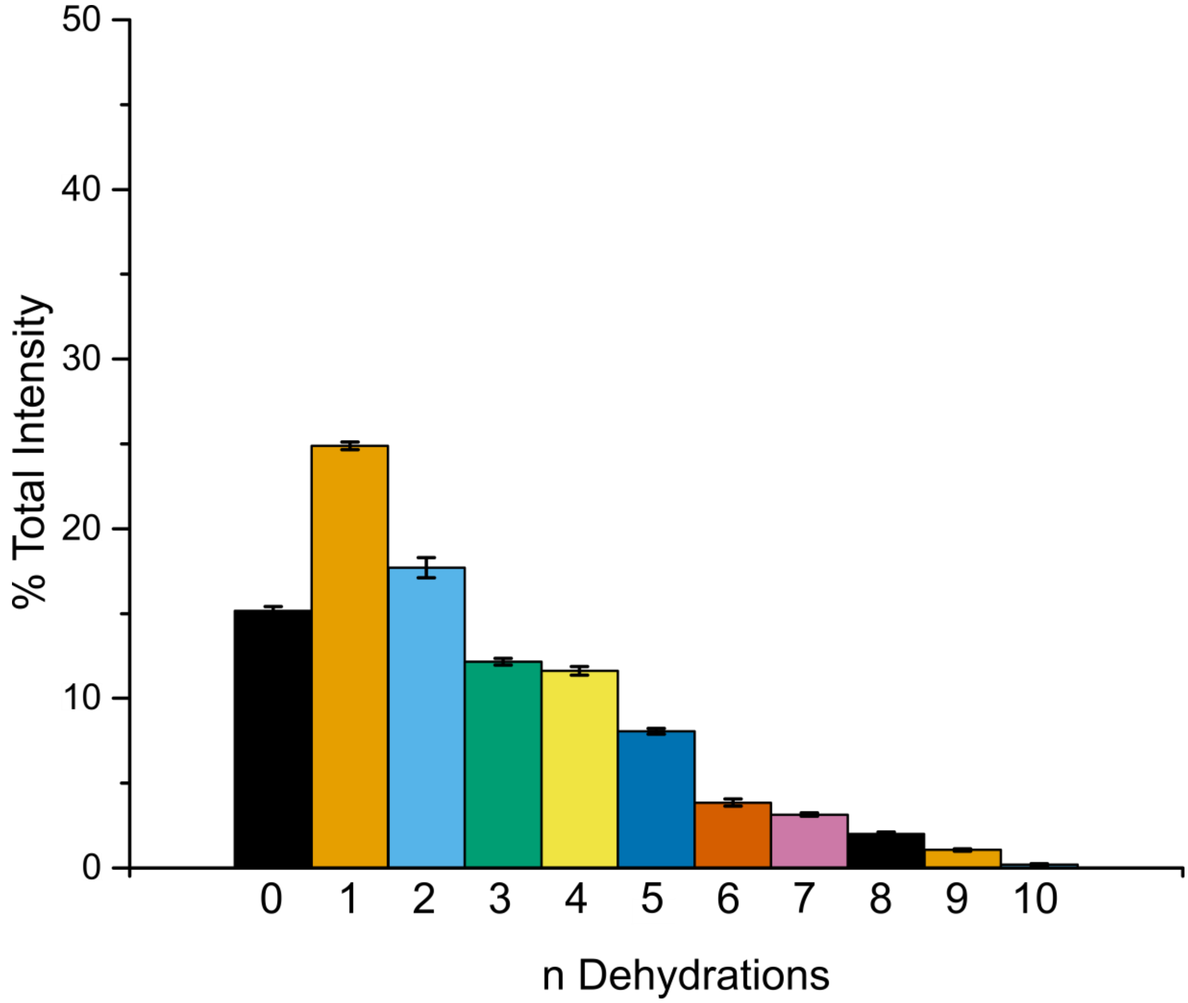


Figure 18: MS1 Intensity Distribution of Dehydrated Peptides. Products of D, H and S polymerisation were screened for multiply dehydrated products at MS1. Data represent mean of 3 measurements ± 1 S.D.

In-source neutral loss fragmentation can be minimised in ESI sources by careful adjustment of acceleration voltages.[2] However, even if in-source neutral losses could be disregarded as a source of extra mass_diffs, sidechain branching sequences may still be falsely assigned as cyclic. This is demonstrated by one proposed cyclic sequence identified by Oligomersoup, the tetrameric peptide DDHD (Figure 32). Extra mass_diffs were observed for this sequence in both MS1 precursors (Figure 33) and MS2 product ions (Figure 34). Only one shifted fragment (“y1-1”) was proposed for this sequence in the *in silico* library. However, for every proposed fragment that could match the cyclic sequence undergoing reading frame shift, an equivalent isomeric fragment of a singly branched species could match the MS2 spectra equally well (Figure 4, Table 4).

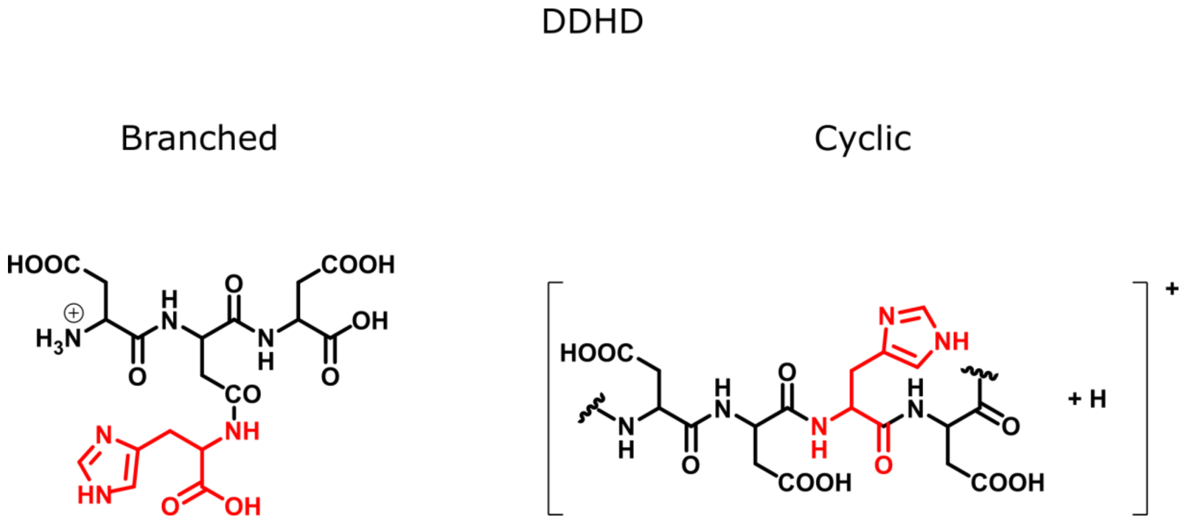


Figure 19: Precursor Ions for Branched and Cyclic DDH. Histidine residue, which is linked to the D2 residue by a β-peptide bond in the branched sequence, is marked in red.

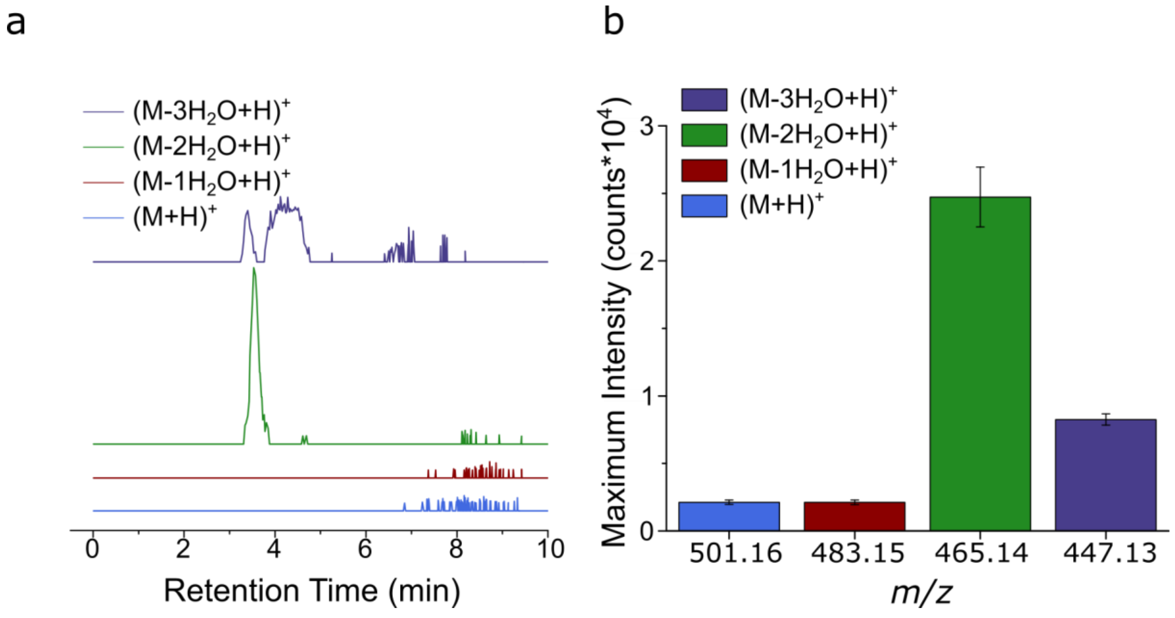


Figure 20: Intensity Distribution of DDHD MS1 Precursors. (a) Extracted Ion Chromatograms (EICs) of DDHD MS1 ions with 0, 1, 2 and 3 dehydrations relative to the

standard linear $(M+H)^+$ ion; (b) maximum intensity of each ion. Data in b represent mean of 3 measurements ± 1 S.D. EIC error tolerance was set to an absolute value of 0.01 from target m/z.

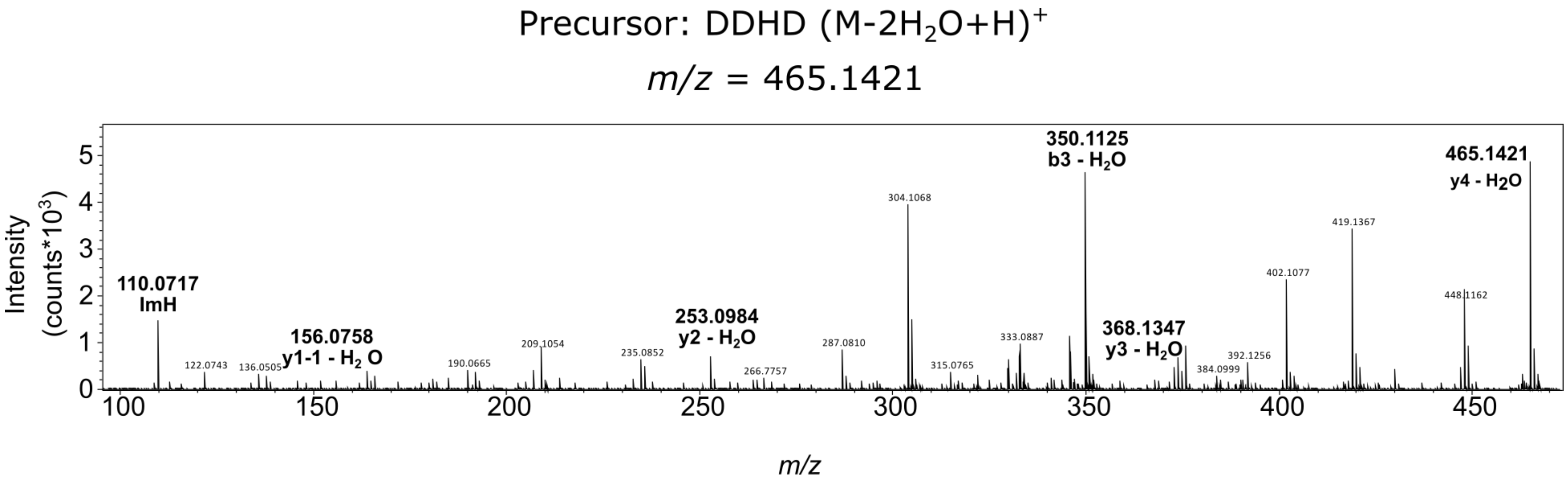
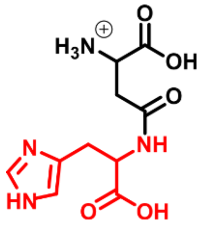
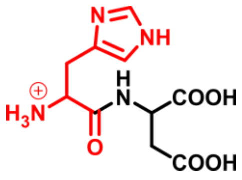
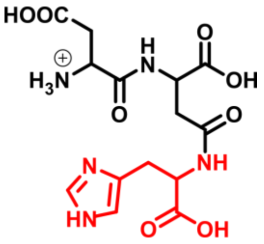
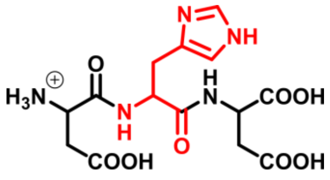
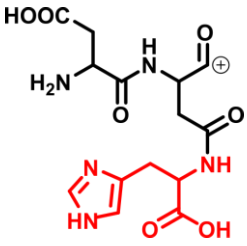
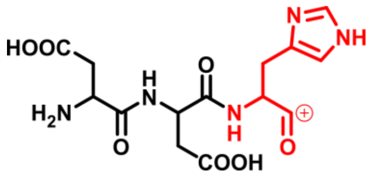
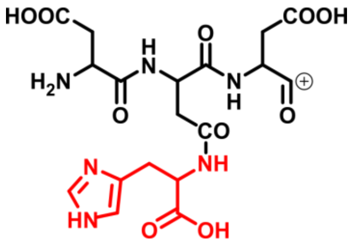
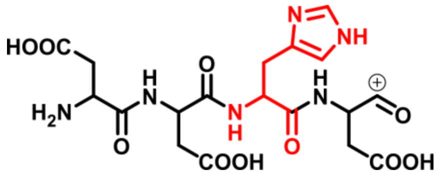
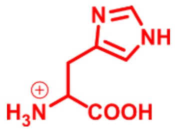
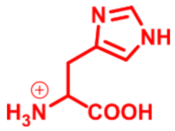


Figure 21: MS2 Spectrum of DDHD Precursor. Precursor = DDHD (M-2H$_2$O+H)$^+$ where M = neutral mass of standard, linear sequence. Relative abundance for confirmed fragments: y4 (100%), b3 (95.21%), ImH (30.12%), y2 (12.87%), y3 (4.37%), y1-1 (3.88%). ‘ImH’ = H immonium fragment.

Table 4: Isomeric MS2 Fragments for Cyclic and Branched DDHD Generated from Precursors in Figure 31.

Fragment	Branched Equivalent	Linear / Cyclic Equivalent
y2		
y3		
b3		
b4		
y1-1		

7.2 Sidechain Crosslinked Sequences

In addition to branched and head-to-tail cyclized sequences, characterisation of sidechain-crosslinked peptide sequences has been attempted by Oligomersoup. Like standard linear oligomers, the goal here is to propose a set of generalizable parameters sufficient to define crosslinks that are agnostic with regard to polymer class. This is straightforward for MS1 precursor ions: any sidechain-sidechain crosslinked can be defined by the following:

1. Target monomers for crosslinking
2. Crosslinking mass_diff
3. Disruption of neutral losses by crosslinking
4. Disruption of sidechain-specific ionization by crosslinking

These four parameters are all that is required to generate and screen for MS1 compositional libraries. This was attempted for disulfide-containing peptide sequences produced via unconstrained oligomerisation of cysteine with other amino acids. The characteristic -2H crosslinking mass_diff (Figure 21) was used to identify crosslinked precursors under a variety of conditions (Figure 22).

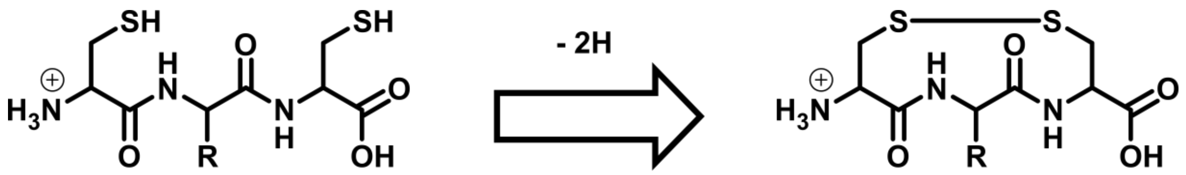


Figure 22: IntraPeptide Disulfide MS1. Mass shift upon formation of disulfide = - 2H (2 x 1.0078 u).

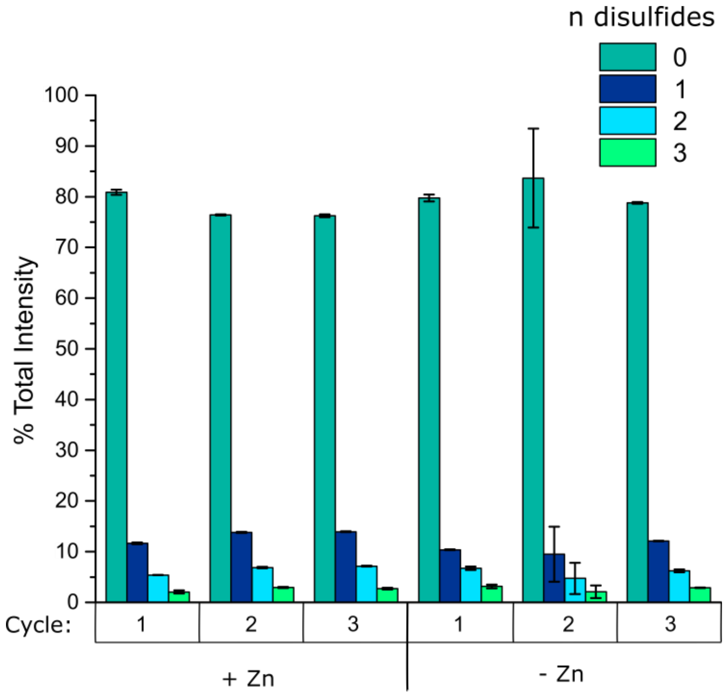


Figure 23: Disulfide Bond Screening at MS1 for Reactions of C + H + L at pH 2.5, 130^{o}C. in the Presence and Absence of Zinc Acetate. Data represent mean of 3 measurements ± 1 S.D.

Under some circumstances, disulphide peptides can fragment at the disulphide bond to produce a signature hydroxyalanine and disulfohydryl fragment (Figure 23). However, the fragmentation methods required to produce such a fragmentation event are not amenable to the standard, linear fragmentation pathways that Oligomersoup relies on for sequencing.[9] To date, all Oligomersoup sequencing workflows have treated individual data sets in isolation, with no option to compare between treatment groups that is found in most proteomics software packages.[10] However, in future versions of Oligomersoup, comparative workflows will be available to aid in the identification of non-standard sequences (e.g. comparison of the same products fragmented via HCD and ETD to produce complimentary data sets).

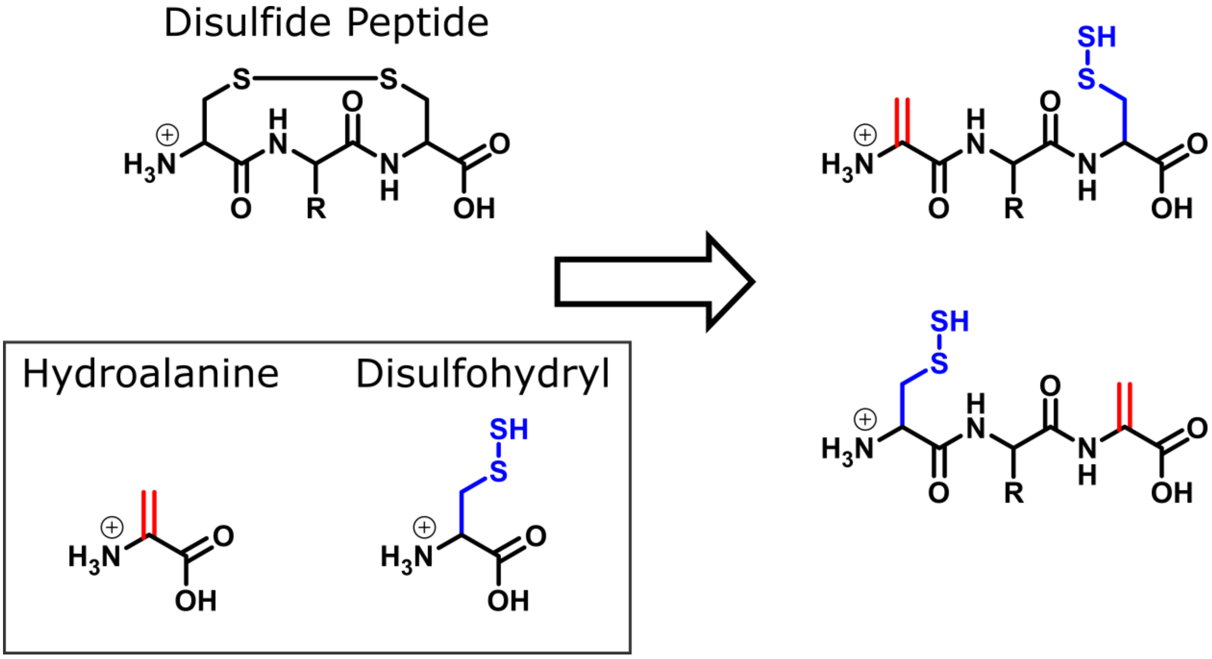


Figure 24: Fragmentation of IntraPeptide Disulfide Bonds. Fragmentation of the S-S bond in the disulfide-containing peptide precursor ion results in hydroalanine (red) and disulfohydryl (blue) fragments). Standard fragmentation of the backbone to produce y- and b- linear fragment series then takes place.

The sequence string conventions currently being used by Oligomersoup for crosslinked and covalently modified sequences is shown in Figure 24 with an example depsipeptide sequence.

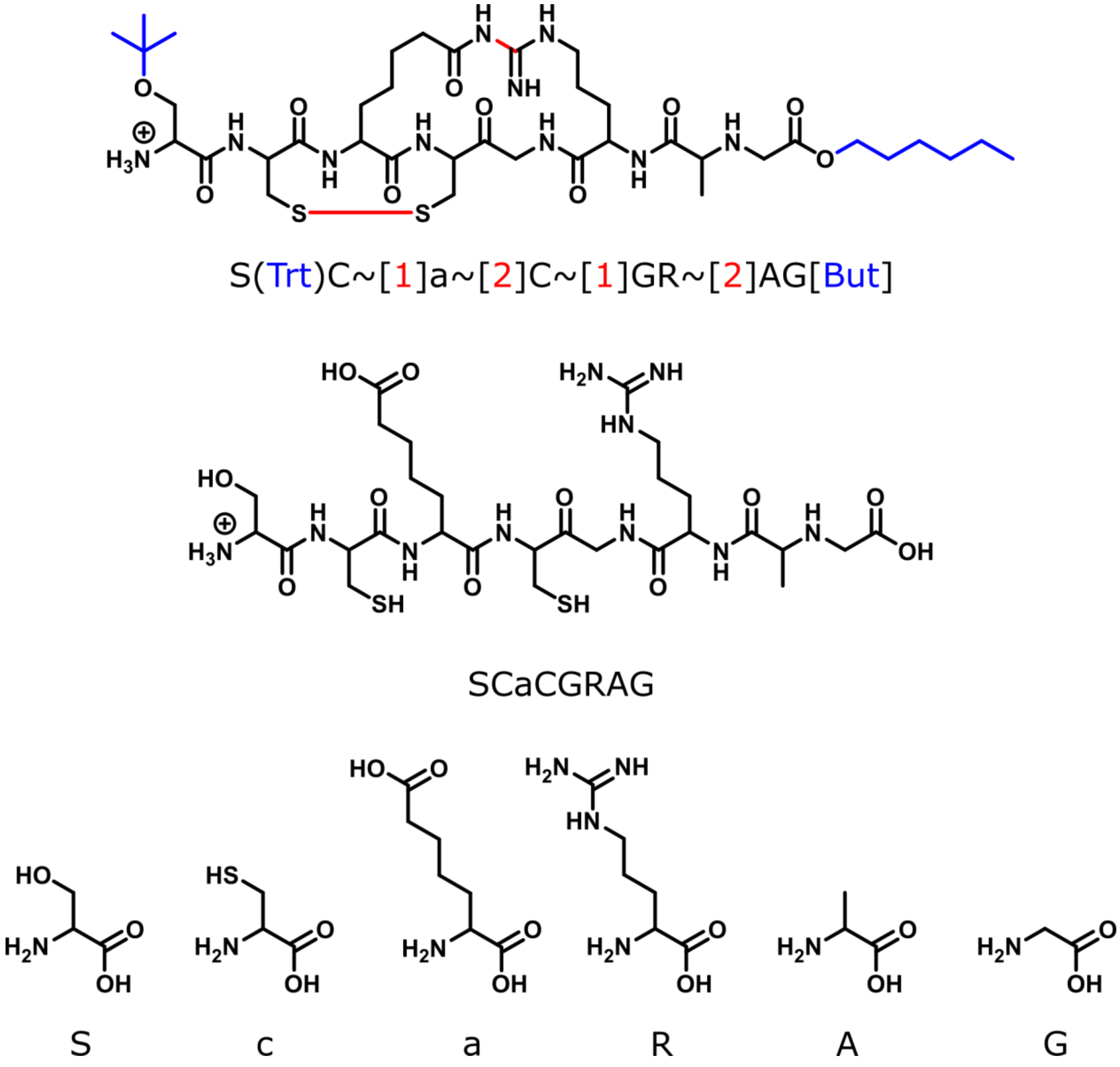


Figure 25: Sequence String Annotation. Side chain modifications are denoted by regular brackets immediately adjacent to the modified monomer one letter code. Internal crosslink pairs are denoted by ~[n] where n = number of crosslinking pair, starting from n = 1. Non-side chain terminal modifications are denoted by regular square brackets [mod], where mod = modification code. In the example above, monomer one letter codes are: S (serine), C (cysteine), a (2-aminopimelic acid), R (arginine), A (alanine), G (glycine). Side chain and terminal modification codes are: Trt (trityl), But (butanoic acid).

8. Oligomersoup Package Structure

Oligomersoup utilises five modules (Figure 46):

1. silico: responsible for all operations related to building theoretical ("*in silico*") sequence libraries
2. extractors: responsible for filtering and screening MS data
3. postprocessing: responsible for final data manipulation and confidence assignments
4. utils: responsible for formatting, error-checking and storage of pre-configured files (i.e. polymer- and instrument-specific settings)
5. workflows: responsible for combining modules 1-4 and executing full Oligomersoup workflows

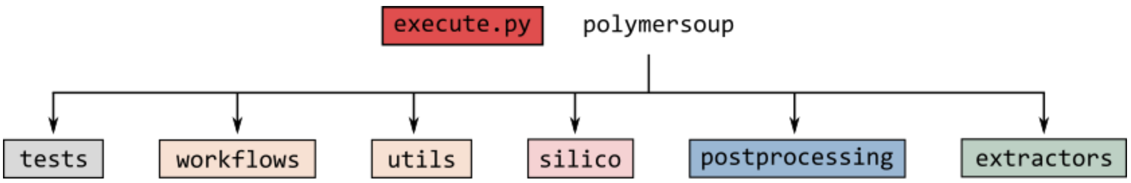


Figure 26: Oligomersoup module / file structure. All Oligomersoup runs are executed from execute.py. Five modules (workflows, utils, silico, postprocessing and extractors) are called from execute, with an additional testing module to be used only by developers.

8.1 Silico Module

The silico module is responsible for generating all *in silico* sequence libraries of both MS1 precursors and MS2 product ions. There are five submodules in the silico module (Figure 48):

1. ms1_silico: responsible for silico operations related to generating theoretical precursor ions.
2. ms2_silico: responsible for silico operations related to generating theoretical MS2 / product ions.

3. polymer: responsible for generating Polymer objects from experimental run parameters. This defines local scope of silico operations for MS1 ionization and MS2 product ion fragmentation.
4. silico_helpers: responsible for basic silico operations (such as sequence string handling) that are required by all other silico modules.
5. silico_handler: responsible for high-level silico operations called directly in sequencing workflows.

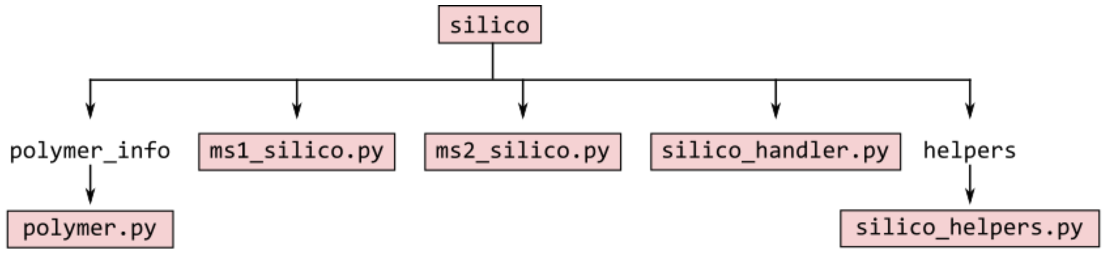


Figure 27: Silico Module Structure.

For the silico module to generate sequence libraries, the scope of silico operations must be narrowed from experimental run parameters to create an instance of the Polymer class. This Polymer object then defines the specific properties required to generate all theoretical ions for the sequences to be screened (Figure 45).

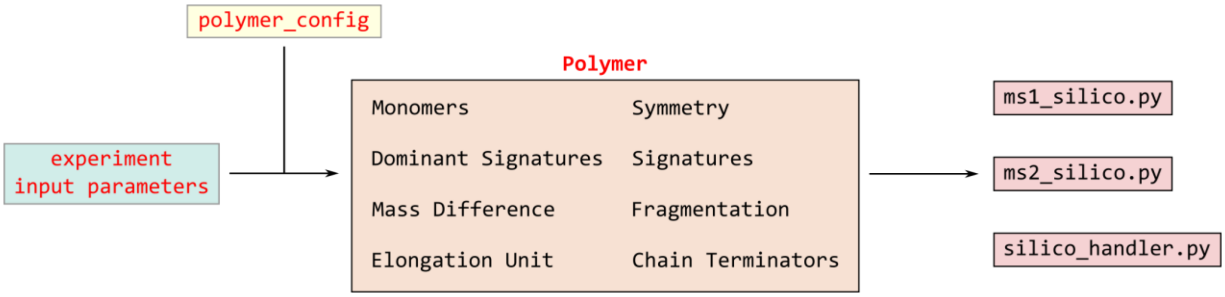


Figure 28: Polymer Object for Narrowing Scope of *In Silico* Fragmentation. Input parameters are passed in from an experiment run file (via a Parameters object). The “polymer_class” input in the input parameters determines which polymer-specific configuration file is used to define the full scope of reactivity, ionization and fragmentation for sequences in *in silico* libraries. Other input parameters (e.g. specific monomers used, instrumentation and matrix

properties) are then used to build a Polymer object, which contains all relevant information for the scope of ionization and fragmentation for sequences in the current run.

8.2 Extractors Module

The extractors module is responsible for all conversion, filtering and screening of mass spectrometry data. There are five submodules the extractors module:

1. spectra_processing: responsible for spectral manipulation and processing independent of any silico data for screening (e.g. dynamic exclusion of observed peaks, generation of BPCs and TICs).
2. ripper_handler: responsible for converting mzml ripper data into RipperDict objects that can be handled by the other extractors submodules and the postprocessing module.
3. filters: responsible for filtering ripper data for parameters such as retention time ranges, precursor *m/z*, intensity thresholds, and presence of specific target ions in spectra.
4. data_extraction: responsible for screening ripper data, searching for theoretical MS1 precursors and MS2 product ions and matching to observed data.
5. extractor_helpers: this submodule contains basic functions that are essential to the operation of all other extractors submodules.

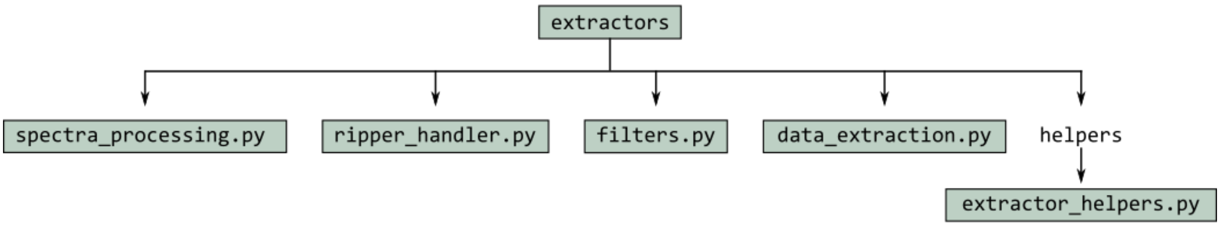


Figure 29: Extractors Module Structure.

8.3 Postprocessing Module

The postprocessing module is responsible for manipulation and final processing of extracted data, after screening of observed data for silico sequence libraries, including assigning final confidence values to confirmed sequences. The postprocessing module has just two submodules (Figure 47):

1. postprocess: responsible for performing postprocessing operations on whole ripper data sets.
2. postprocess_helpers: responsible for individual confidence calculations and plotting of spectra.

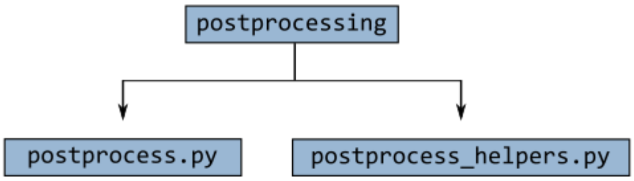


Figure 30: Postprocessing Module Structure.

8.4 Utils Module

The utils (“utilities”) module is responsible for several functions, including:

1. Storing polymer- and instrument-specific configuration files.
2. Setting and handling permissible inputs and types for experimental run parameters.
3. Logging.
4. Oligomersoup custom errors.

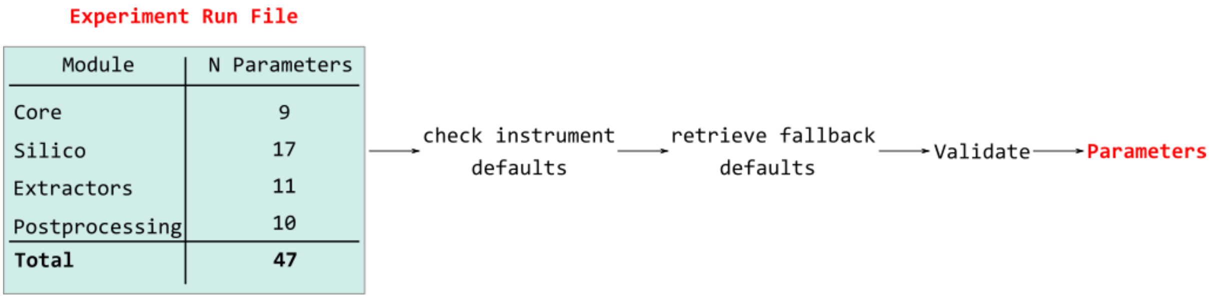
The utils module has twelve submodules (Figure 48):

1. instrument_standards.fragmentation_methods: defines default behaviours of polymer classes for particular fragmentation methods (e.g. HCD, CID, ETD).
2. general_functions: contains very basic functions such as file handling, used throughout all other Polynersoup modules.
3. logger_utils: responsible for setting up logging configuration for Oligomersoup.
4. errors: this submodule contains custom error classes.
5. type_dicts: this module contains two submodules (parameter_fallbacks and parameter_type_dicts), and is responsible for validating input parameters upon execution of a Oligomersoup sequencing workflow. Submodules of type_dicts are described in more detail, below.
6. parameter_handlers: this submodule is responsible for converting input parameters to a Parameters object to be passed on to other modules, and for retrieving default values for parameters, including instrument- and polymer-specific defaults. Submodules of parameter_handlers are described in more detail, below.
7. global_chemical_constants: this submodule contains important chemical constants, including common neutral species and ions. Full details of global_chemical_constants are given in more detail, below.

8.4.1 Parameter Handlers

The utils.parameter_handlers and utils.type_dicts submodules are responsible for reading user input parameters upon execution of Oligomersoup, retrieving any default values for parameters not explicitly stated, and validating all input parameters (including checks for feasibility of parameter values for instrumentation and other experimental conditions).

If an input parameter is not explicitly stated in the run file, the utils.parameter_handlers submodule will check instrument-specific configuration file and attempt to retrieve a default value that best matches experimental conditions. If this cannot be done, the utils.type_dicts submodule will attempt to retrieve a last resort general default value for the parameter. If no such value exists, a custom error will be raised and the user will need to update either the instrument settings or explicitly state a valid value for the missing parameter. After all parameters have been accounted for, utils.type_dicts will validate all parameters in the final Parameters object before it is passed on to other modules for use in sequencing workflows.



Module	N Parameters
Core	9
Silico	17
Extractors	11
Postprocessing	10
Total	**47**

Figure 31: Input Parameter Fallbacks and Instrument Defaults. There are 47 unique input parameters that can be passed in from a run file to execute a Oligomersoup sequencing workflow. Oligomersoup attempts to set values for any missing parameters by first checking instrument-dependent default values, which depend on the model of mass spectrometer being used as well as the chosen polymer class. If no instrument-specific default values are found, last resort “fallback” values are retrieved. After this, all parameter values are validated and passed on to other modules via a Parameters object.

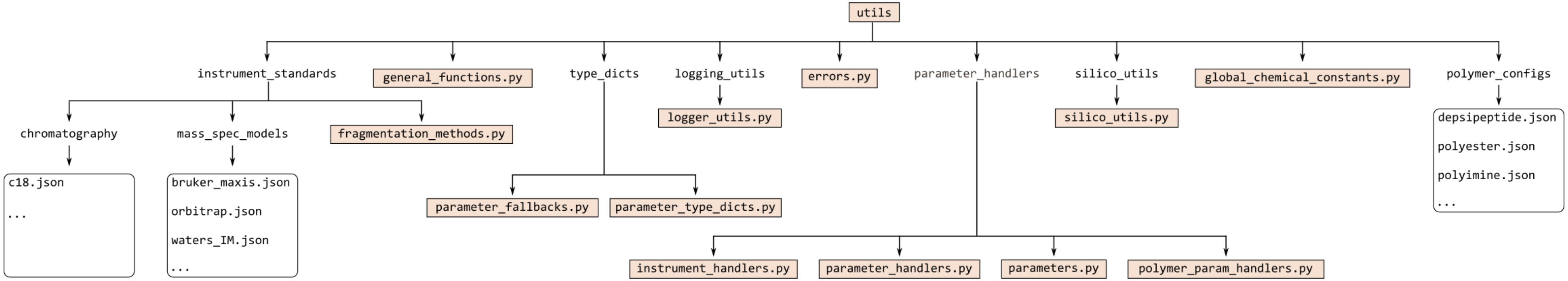


Figure 33: Utils Module Structure.

9. Soup Data

Three model systems of unconstrained oligomerization reactions were used to develop and test Oligomersoup:

1. One-pot depsipeptide oligomerization and N-terminal acylation
2. Non-acylated peptide and depsipeptide oligomerization
3. Amine-aldehyde condensation to produce polyimines

Products of model system 1 were sequenced using a Bruker Maxis Impact II UHR-QqTOF (Ultra-High Resolution Qq-Time-Of-Flight) mass spectrometer. Unless stated otherwise, all mass spectrometry data for model systems 2 and 3 were acquired using a Thermo Scientific Orbitrap Fusion Lumos Tribrid mass spectrometer.

In this section, these model systems will be described as well as how they aided in the development and validation of Oligomersoup.

9.1 Model System 1: One-Pot Depsipeptide Oligomerization and N-Terminal Acylation

The initial soup mixtures used to develop and test Oligomersoup were N-terminally acylated (depsi)peptides synthesized via thermal dehydration of amino acid, hydroxy acid and fatty acids. In each of these reactions one non-polar amino acid (valine or glycine) was reacted with glycolic acid ± one fatty acid (either oleic acid or palmitic acid) and ± the polar amino acid asparagine. Glycine and valine were chosen for their potential to form peptides in high yield from wet-dry cycling reactions,[11] while asparagine was chosen to enhance solubility of products. Despite containing a free primary amine, carboxamide side-chains such as asparagine are far less prone to branching via β and γ-linkages at the side-chain than lysine and arginine, removing the difficulty inherent in analysing mixtures of multiply branched products.

Due to the well-documented pH-dependence of (depsi)peptide oligomerization under thermal dehydration, reactions were carried out at a range of starting pH values. Pure peptide oligomerization in such reactions proceeds almost exclusively at high (9.0-9.8) and low (2.5) pH. Elongation of depsipeptides via thermal wet-dry cycling is also highly pH-dependent, with optimal yield occurring from a starting pH of 3.0. The low efficacy of unactivated amino- and α-hydroxy acid polymerisation at moderate pH is unsurprising, given the zwitterionic nature of standard amino acids under these conditions. At pH 9.0-9.8, deprotonated amines are more nucleophilic, and thus more reactive with carbonyl groups. Peptide elongation at extreme alkaline pH ($\geq$ 10) is hindered by base-catalysed peptide hydrolysis. Prior to work published in our group, unactivated amino acid polymerisation had not been explored under very acidic conditions (< pH 3), likely due to the poor nucleophilicity of protonated amines.

By carrying out thermal dehydrations at a pH range, Oligomersoup was provided with a variety of product mixtures for testing, with stark differences expected in both yield and diversity of products.

As outlined above, ideal pH ranges for peptide and depsipeptide polymerisation are well-established. However, the effect of pH on a system incorporating one-pot fatty acid acylation was unknown. Indeed, it was unknown whether acylation would occur at all without activation of either the amino acids or fatty acids by e.g. N-carboxyanhydrides or acyl chlorides. Thus, reactions were screened at a wide range of pH values (2.5, 3, 4, 5, 7, 9, 10) at 95 ^{o}C in open-cap, thermal dehydrations from a starting concentration of 0.1 M amino acid, 0.1 M glycolic acid and 0.1 M fatty acid. Reactions incorporating asparagine were carried out with 0.01 M asparagine + 0.09 M valine or glycine, giving a fixed overall amino acid concentration of 0.1 M in all reactions and a 9:1 non-polar/polar amino acid ratio in the hetero-polymerisation reactions. Starting reagent solutions had an initial pH of approximately 3.0-3.5, depending on the amino acids used, and were adjusted accordingly via the addition of either H_3PO_4 or NaOH.

9.1.1 Monomer Conversion

Due to the potential diversity of products, and the fact that Oligomersoup was still in development with its sequencing abilities not yet established, a method for characterising products that did not rely on sequencing was required.

Hydrophilic liquid interaction chromatography (HILIC) was used to separate individual amino acids used in the polymerisation reactions. Consumption of each amino acid in the reactions was estimated by measuring the concentration of free amino acid remaining in product mixtures. As most underivatized amino acids are not suitable for detection via UV absorbance due to their lack of strong chromophores, charged aerosol detection (CAD) was chosen as the detector in the quantitative HPLC used for reaction screening. Free amino acid concentrations remaining in product mixtures were calculated by reading signal intensity of resolved peaks (Figure 32c, d) off a standard curve (Figure 32b), with conversion calculated from a known starting concentration. Conversion of both valine and glycine monomers was measured across the full pH range tested.

Glycine showed high conversion across a much wider pH range compared to valine. However, it is worth noting that unactivated glycine is notorious for forming the cyclic dimer 2,5-diketopiperazine, a thermodynamic “dead end” that may preclude further elongation under some circumstances. Therefore, conversion of glycine monomer may not in itself be indicative of high yield of linear (depsi)peptide products. Consequently, valine conversion was used as the primary means of screening ideal reaction conditions (Figure 32c, d).

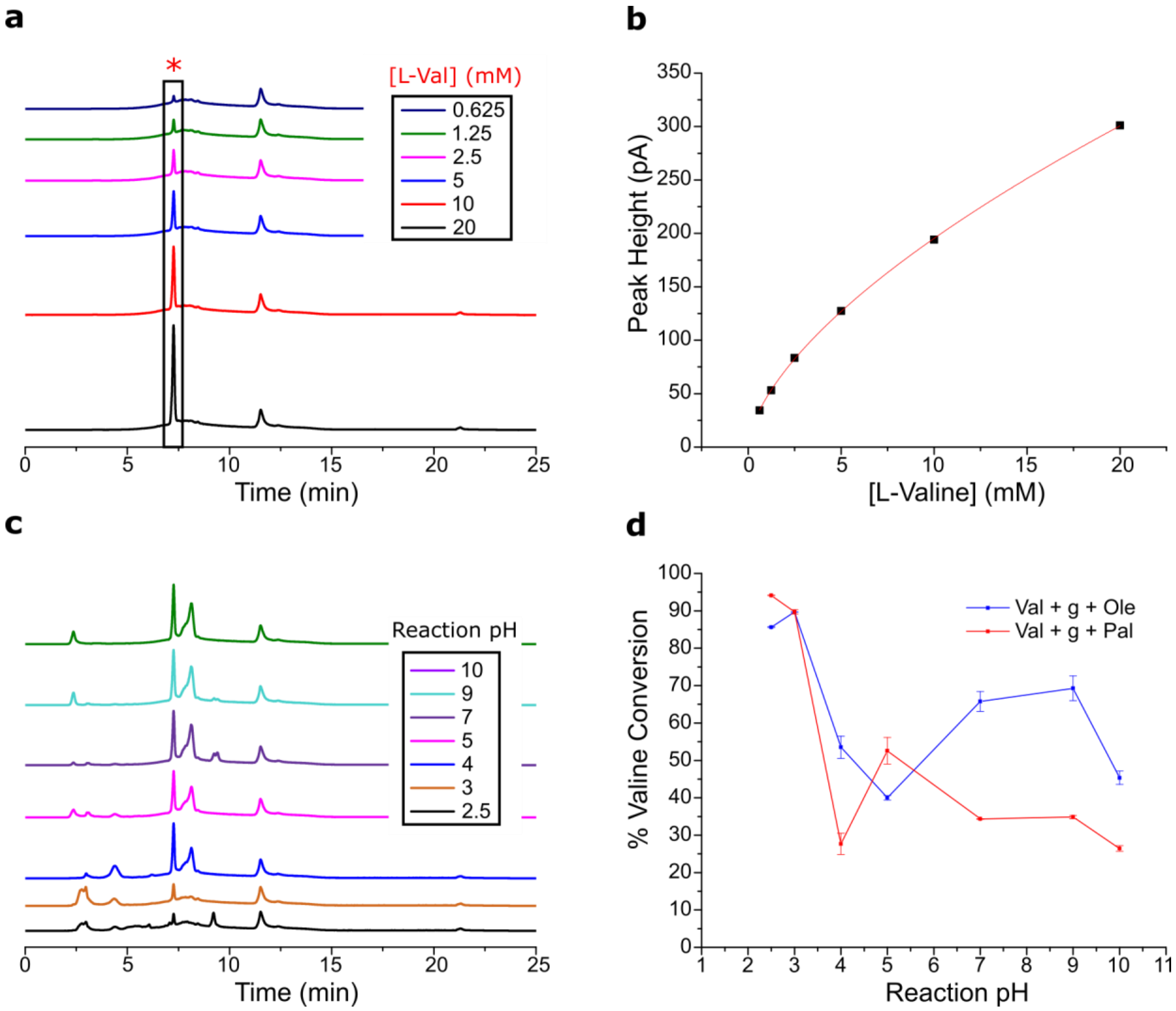


Figure 33: Hydrophilic Interaction Liquid Chromatography with zic-HILIC HPLC column and Charged Aerosol Detector of Valine Lipid-Depsipeptide Hybrid Reactions at pH Range. A) zic-HILIC chromatograms of L-Valine standards; B) L-Valine standard curve; C) chromatograms of reaction products with V, g and Ole; D) calculated valine monomer conversion for reactions of V, g and Ole / Pal. Data in b and d represent 3 measurements ± 1 S.D.

As can be seen from Figure 32d, maximum valine conversion occurred at low pH (2.5-3.0), consistent with previous literature reports of depsipeptide systems. Valine conversion in the absence of asparagine was similar in both the oleic and palmitic acid reactions, with both exhibiting maximum conversion at pH 2.5-3.0 (Figure 32d). This is unsurprising given the increased reactivity of glycolic acid in unactivated thermal dehydration reactions at this pH range. Interestingly, valine conversion in the oleic acid reactions also exhibited high conversion at pH 7-9, with values approaching those observed for the extremely efficacious reactions at pH 2.5 and 3.0 (Figure 32d). A less dramatic increase in conversion was observed for the palmitic acid reactions at pH 5 (Figure 32d). As both oleic and palmitic acid have a typical pKa of 4.9-5.1, no obvious explanation could be offered for the differences observed in the pH dependency of these reactions. However, it is pertinent to note that the pKa of single chain fatty acids can vary widely depending on a variety of factors, including vesicle packing, pH of surrounding solution and presence of salts.

To confirm monomer conversion as measured via zic-HILIC HPLC, quantitative ^{1}H NMR with an internal standard of maleic acid (Figure 33). ^{1}H NMR peaks of free valine in product mixtures were assigned by comparison to an L-Valine standard, which was assigned via 1H NMR, 13C DEPT-Q and 1H-13C{1H} HSQC. Free valine concentration in product mixtures was estimated by calculating the relative integrals of valine ^{1}H peaks to ^{1}H peaks of maleic acid at a known concentration (Figure 33). Products were re-dissolved at 2.5x initial concentration in deuterated pH 7.4 50 mM Na2HPO4 buffer and given only minimal time (≤ 10 min) for dissolution at room temperature before filtration through a 0.22 μm nylon membrane. This was to ensure minimal dissolution of valine-containing peptide products, which are less soluble in aqueous solution than the free monomer.

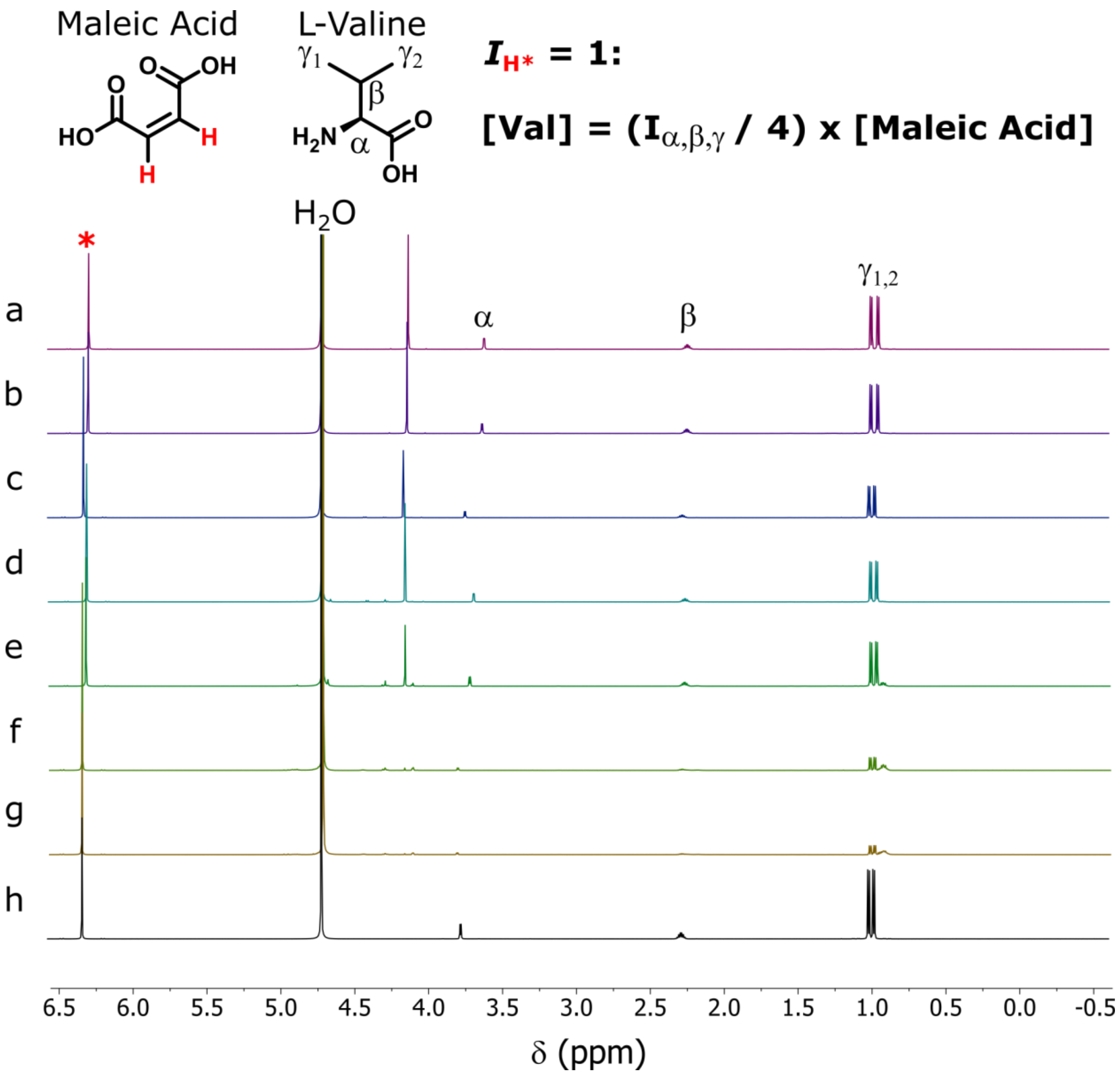


Figure 34: Quantitative 1H NMR Measurement of Free Valine in V + g + Pal Reaction Products at pH Range with Maleic Acid. Samples were run in D_2O with 10 mM internal maleic acid standard, with integral of maleic acid peak (IH*) arbitrarily set to 1 for ease of calculation. 1H NMR spectra of 10 mM maleic acid with reaction products at pH 10 (a), 9 (b), 7, (c), 5 (d), 4 (e), 3 (f), 2.5 (g) and 10 mM L-Valine standard (h).

As illustrated in Figure 34, conversion values obtained via HPLC and NMR were in almost perfect agreement between analytical replicates of the same product mixtures. To assess the reproducibility of these measurements between experiments, conversion was measured via HPLC on a set of repeat reactions carried out on different days under otherwise identical conditions. Conversion values between these experimental repeats were in agreement at all pH values, with the only possible exception being pH 7 (Figure 34).

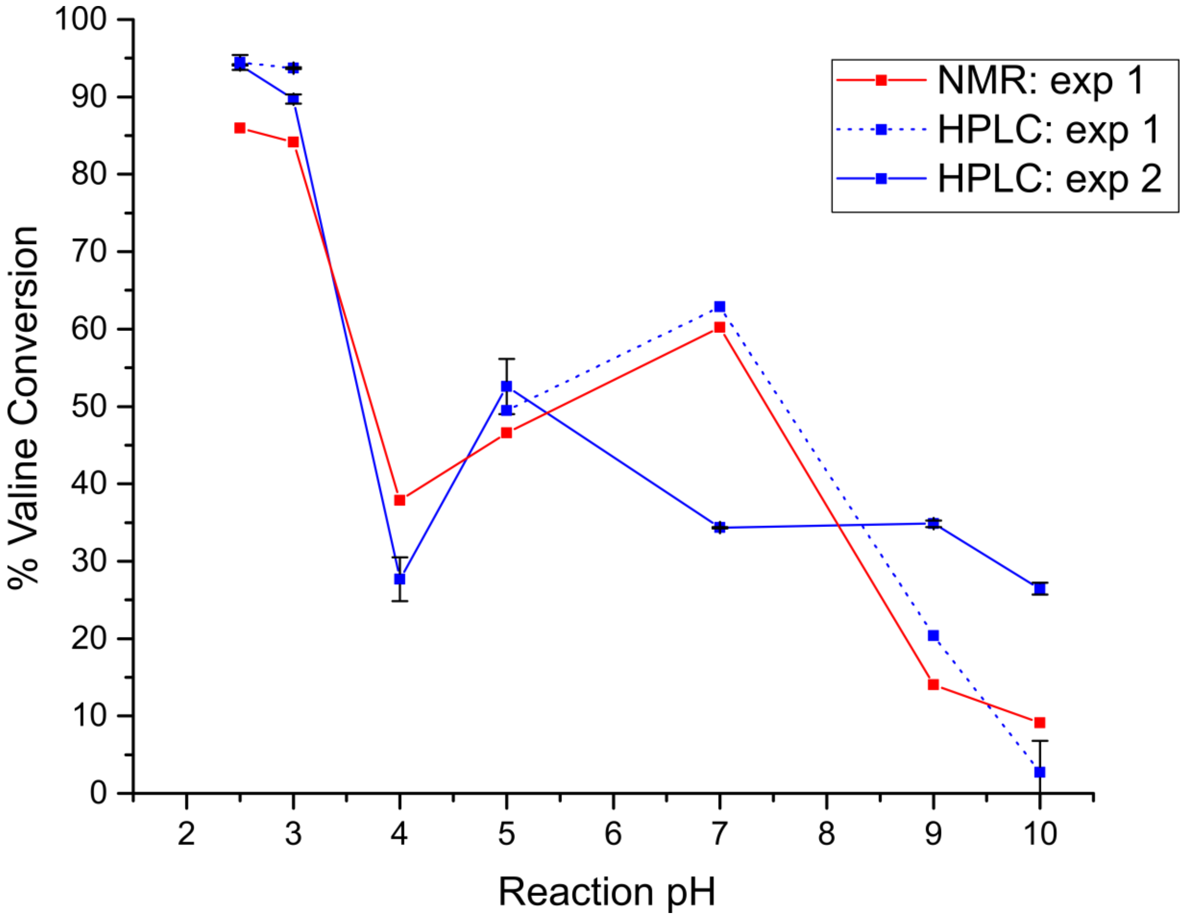


Figure 35: Validation of Valine Conversion Measurement via 1H NMR and HPLC. Free Valine monomer concentration was measured for two analytical replicates of the same product mixtures via 1H NMR and HPLC (NMR: exp1, HPLC: exp1) and via HPLC for experimental replicates of product mixtures produced in separate reactions (HPLC: exp2). Data represent mean of 3 measurements ± 1 S.D.

9.1.2 Product Characterisation

Having established differences in monomer conversion between conditions using two independent, quantitative methods, a means of separation and detection of products via MS/MS was required to provide data for testing Oligomersoup.

HPLC is an indispensable tool for separation and detailed analysis of polymer mixtures in both biological and non-biological settings. To ensure optimal conditions for separation of mixtures of acylated and non-acylated depsipeptides comprised of monomers with a wide range of side-chain polarities, the AdvanceBio Peptide Plus HPLC column was chosen. Unlike standard reverse-phase columns, this column is suitable for retention of both hydrophobic and hydrophilic peptides, owing to its hybrid end-capped C18 stationary phase with a charged surface. Furthermore, separation is possible using MS-friendly additives such as formic acid.

Having identified a suitable method for separating potential products via HPLC, remaining product mixtures were screened using the AdvanceBio Peptide Plus HPLC column with a CAD detector. In broad agreement with the monomer conversion results, the number of peaks observed was highly pH-dependent, with the highest number of peaks being detected in low pH products for the majority of reactions tested, for both valine (Figure 35) and glycine (Figure 36a-d) reactions. An intense peak with late retention time (25-27 min) was observed in pH 9-10 products of oleic and palmitic acid reactions, indicating formation of acylated species (Figure 36a-d).

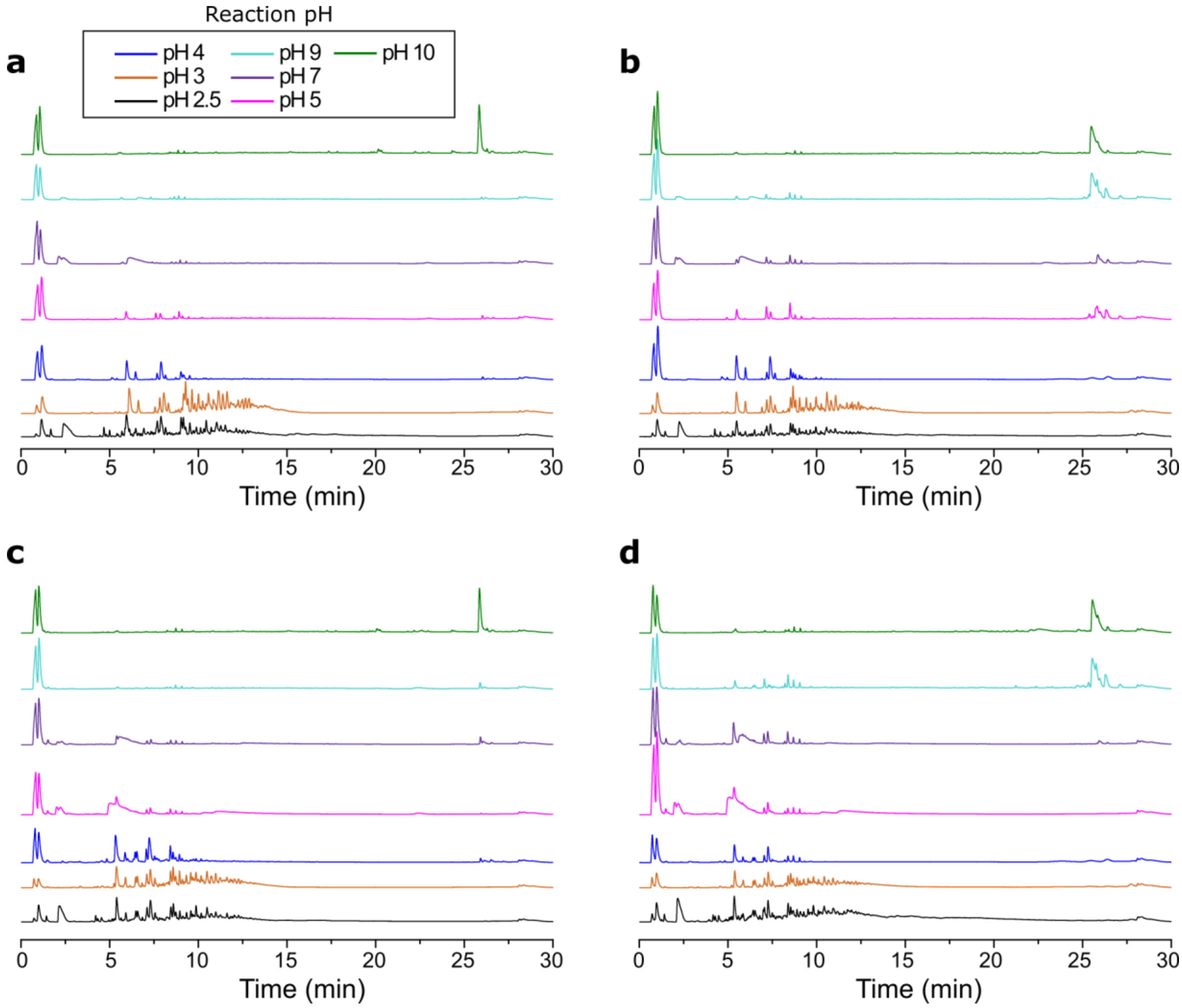


Figure 36: AdvanceBioPeptide Plus HPLC-CAD of Valine Lipid-Depsipeptide Hybrid Reactions at pH Range. Reactions products of Val, glycolic acid plus: A) Ole; B) Asn and Ole; C) Pal; D) Asn and Pal.

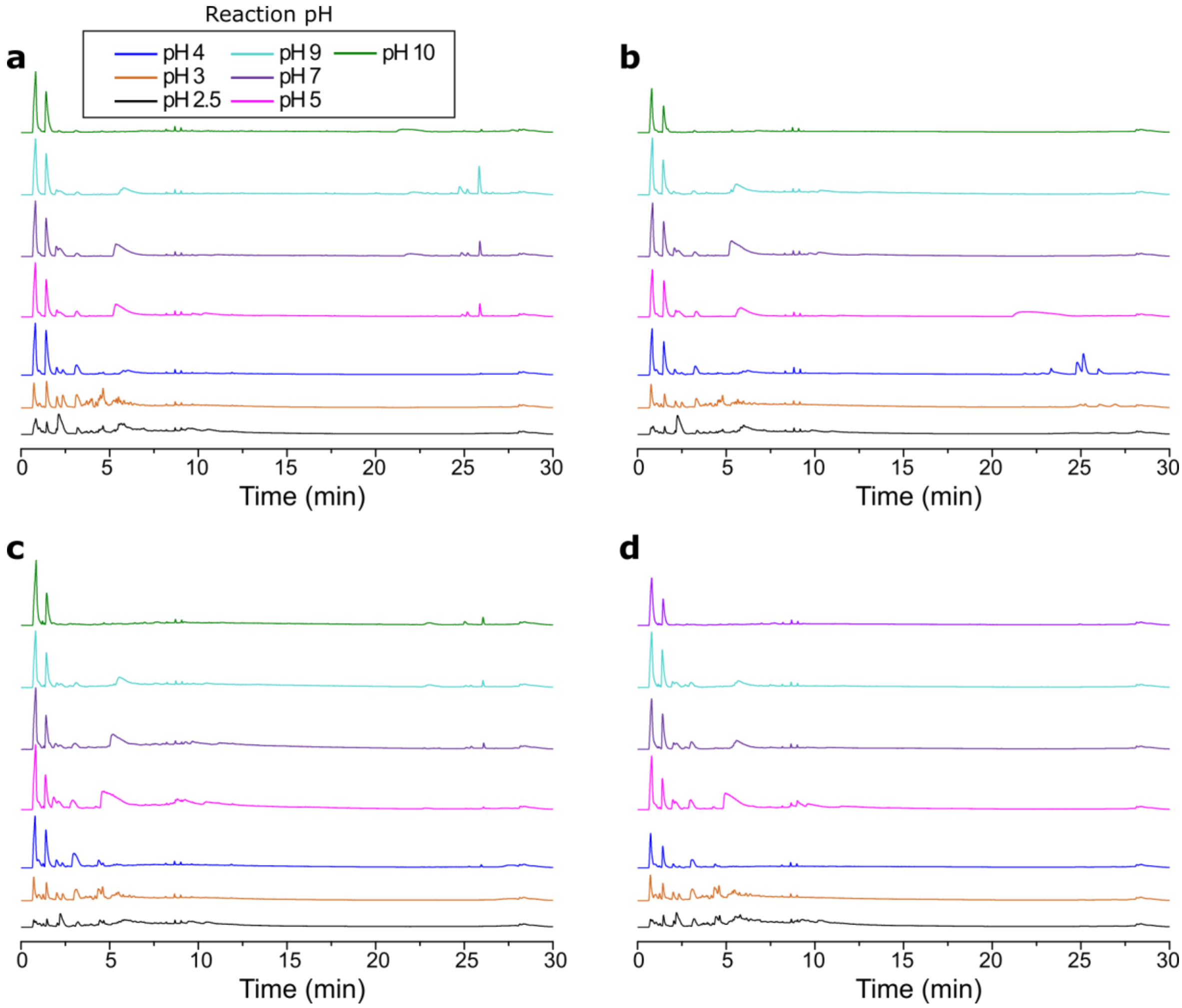


Figure 37: AdvanceBioPeptide Plus HPLC-CAD of Glycine Lipid-Depsipeptide Hybrid Reactions at pH Range. Reactions products of Gly, glycolic acid plus: A) Ole; B) Asn and Ole; C) Pal; D) Asn and Pal.

9.1.2.1 Separation and Characterisation of N-Terminally Acylated Sequences

Two acylated valine dimers were identified via MS1 EICs of acylated precursors (Figure 37), one with an N-terminal oleic acid moiety (Figure 37 a,b) and the other palmitic acid (Figure 37 c,d). Abundance of both species was highly pH-dependent, with intensity peaking in the high pH reactions (pH 9 and 10 for the oleic- and palmitic acid-modified dimer, respectively).

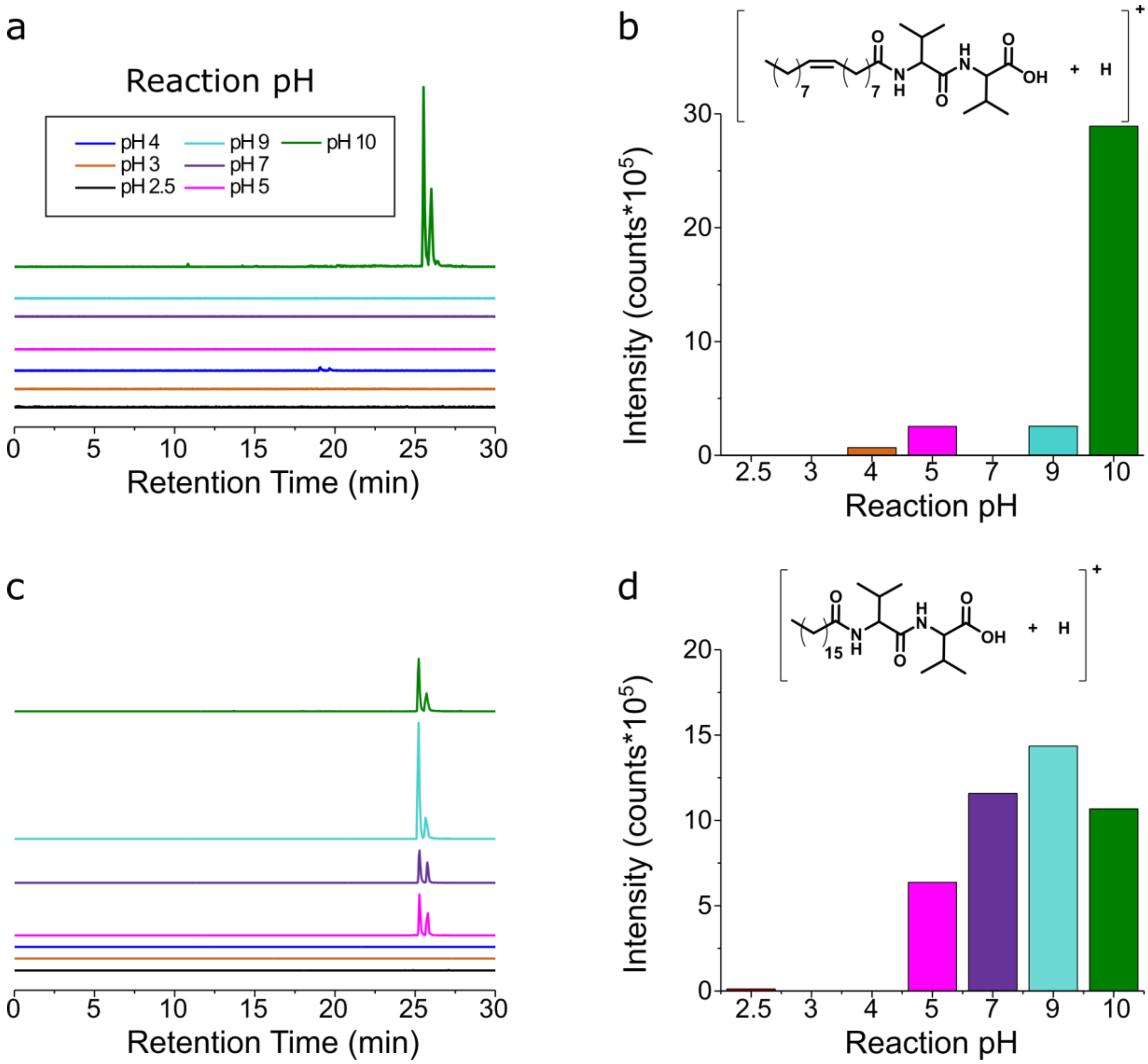


Figure 38: LC-MS of Acylated Valine Dimers from Reaction Products of V + g + Ole (a, b) and Pal (c, d). MS1 EICs and total EIC intensities are shown for oleated and palmitated valine dimer in reaction products at pH 2.5, 3, 4, 5, 7, 9 and 10 from reagents V + g + Ole (a: EICs, b: total intensity of Ole-VV); V + g + Pal (a: EICs, b: total intensity Pal-VV). m/z of Ole-VV and Ole-Pal = 481.3994 and 455.3851, respectively.

Both species were isolated and fragmented via CID in standard acquisition using DDA on the Bruker Maxis Impact II. The initial assumption prior to analysis was that the fatty acid moiety would completely dissociate from the peptides, leading to standard peptide fragment series with an additional acylium signature ion corresponding to the dissociated fatty acid (Figure 38). However, acylated b1 fragments were observed in the CID MS^2 spectra of both precursors as well as the free fatty acid acylium ions (Figure 39, Figure 40). This indicated only partial dissociation of the acylating moieties during fragmentation via CID. This led to the introduction of the MS2 Silico parameter “universal_ms2_shift” for covalent modifications, a parameter which can now be set in polymer- and instrument-specific configuration files.

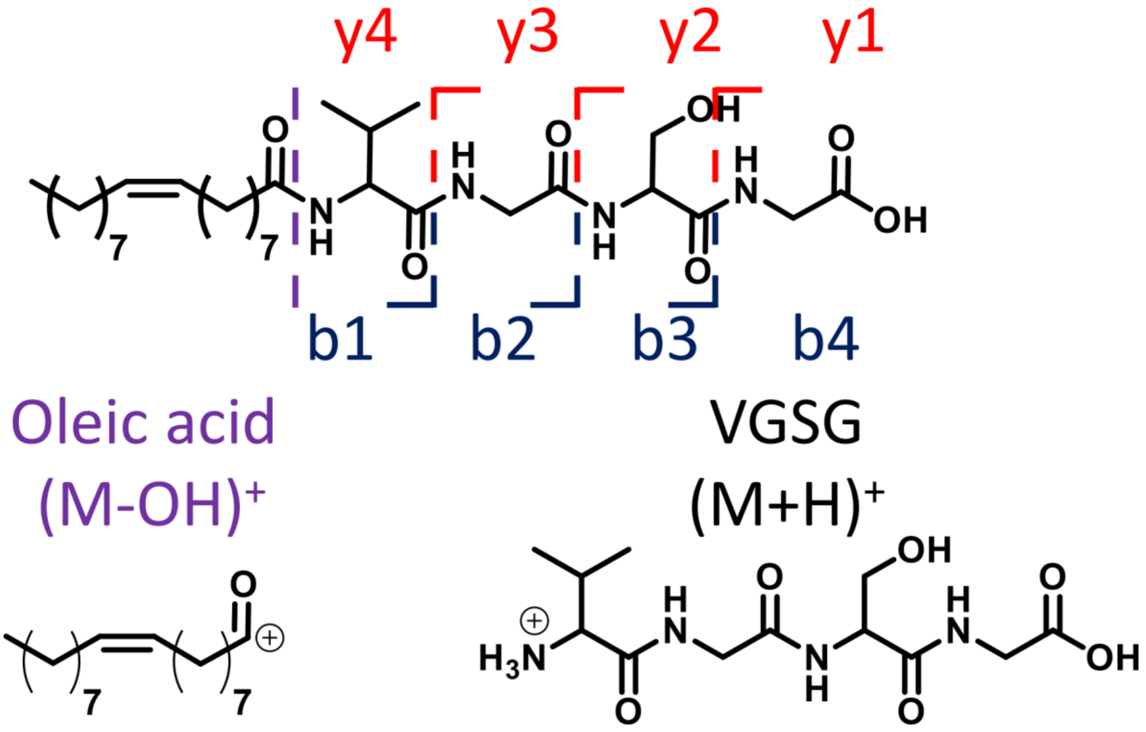


Figure 39: Standard Precursor and Free Acyl Fragment for Example N-terminally Oleated Sequence VGSG

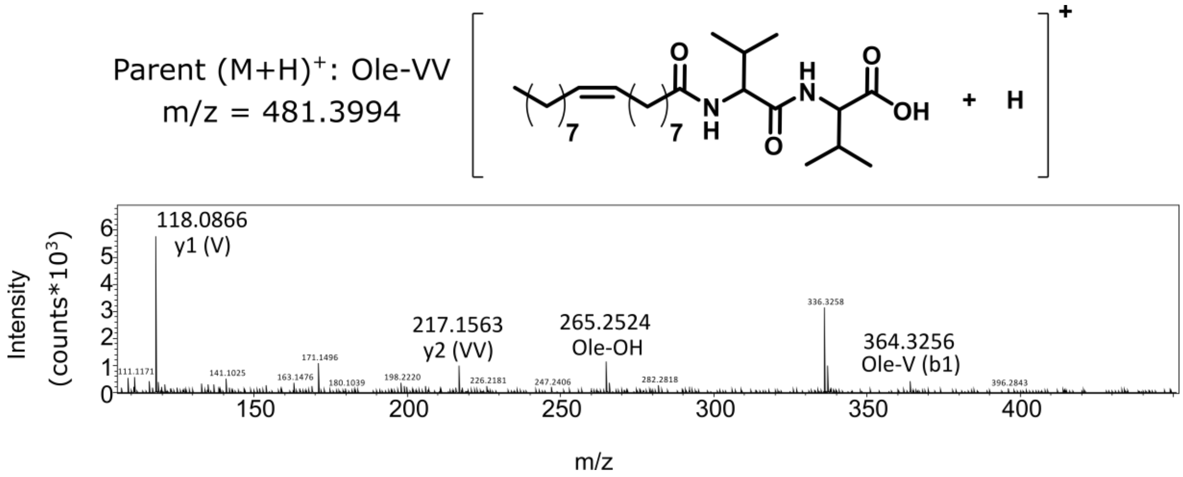
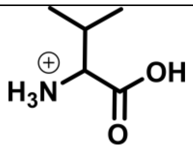
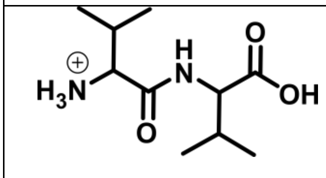
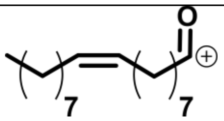
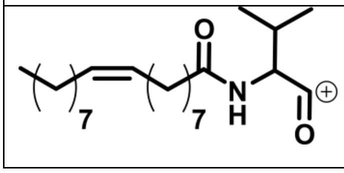


Figure 40: MS2 Spectrum of Oleated Valine Dimer with Peak Assignments. Assigned MS2 fragments of Ole-VV from reaction products of V + g + N + Ole (pH 10). See Table 3 for fragment structures

Table 3: Observed MS2 Fragments of Oleated Valine Dimer

Fragment	Observed *m/z*	Theoretical *m/z*	Error (ppm)
	118.0866	118.0868	1.69
	217.1563	217.1553	4.60
	265.2524	265.2532	3.02
	364.3256	364.3216	10.98

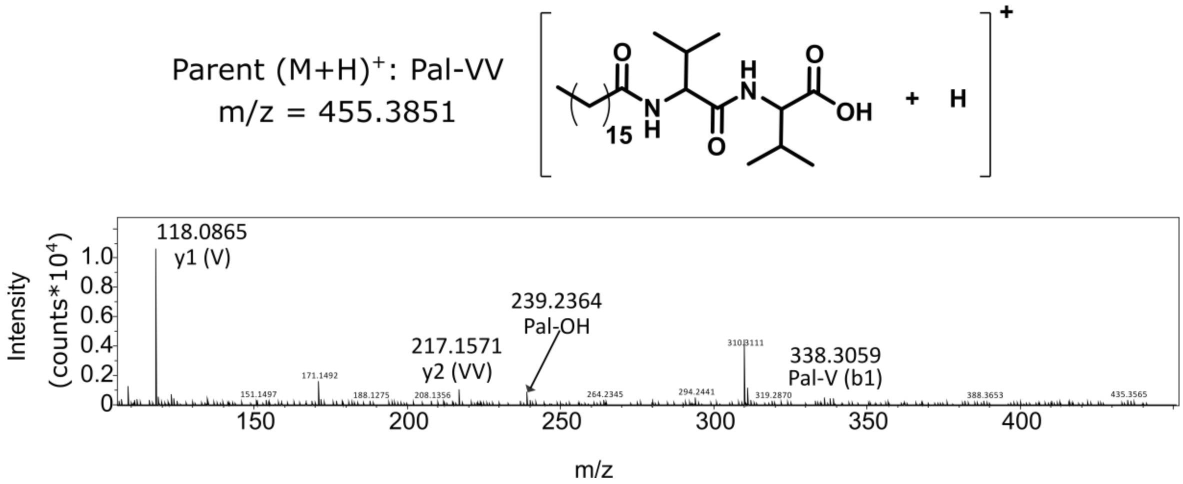
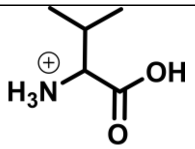
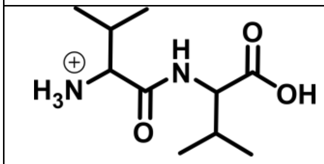
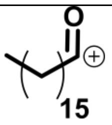
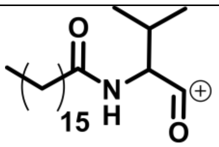


Figure 41: MS2 Spectrum of Palmitated Valine Dimer with Peak Assignments. Assigned MS2 fragments of Pal-VV from reaction products of V + g + N + Pal (pH 10). See Table 4 for fragment structures.

Table 4: Observed MS2 Fragments for Palmitated Valine Dimer

Fragment	**Observed *m/z***	**Theoretical *m/z***	**Error (ppm)**
	**118.0865**	**118.0868**	**2.54**
	**217.1571**	**217.1553**	**8.29**
	**239.2364**	**239.2375**	**4.60**
	**338.3059**	**338.3060**	**0.30**

The goal of Oligomersoup is to provide a tool for *de novo* sequencing and characterisation of heterogeneous mixtures of oligomers from data acquired via LCMS and LCMS/MS. To serve as a model system for initial development of this tool, product mixtures would ideally consist of a variety of both acylated and non-acylated species. MS1 compositional screening of non-acylated depsipeptides produced from reactions of V + g ± Ole confirmed presence of non-acylated products even in the presence of oleic acid (Figure 41).

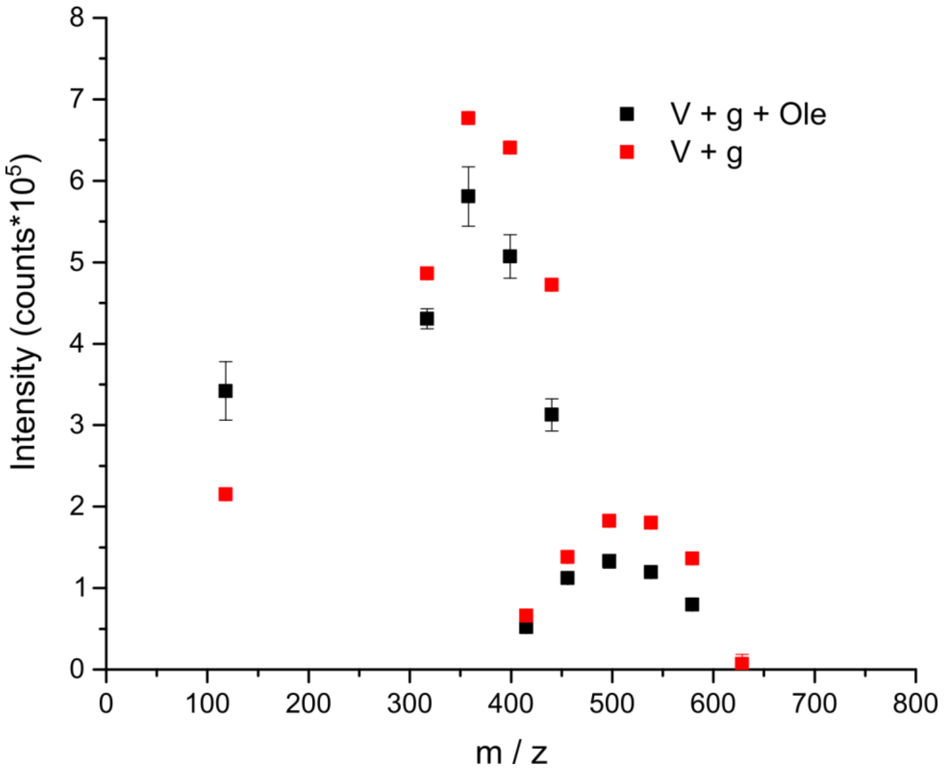


Figure 42: Effect of Oleic Acid on Detection of Non-Acylated Depsipeptides. Measured intensities of non-acylated products are shown for reactions of V + g (red) and V + g + Ole (black) at pH 2.5. Data represent mean of 3 measurements ± 1 S.D.

9.1.2.2 Sequencing of Acylated and Non-Acylated Depsipeptides

A variety of sequences were identified in product mixture, with the majority confirmed with confirmed fragment ratio ≥ 60% for the best samples (Figure 42). Sequencing of all product mixtures across the full range of pHs and input monomer combinations tested confirmed that the pH-dependence of these reactions observed via HPLC was consistent with the number of unique sequences confirmed with a confidence ≥ 60% (Figure 43, Figure 44).

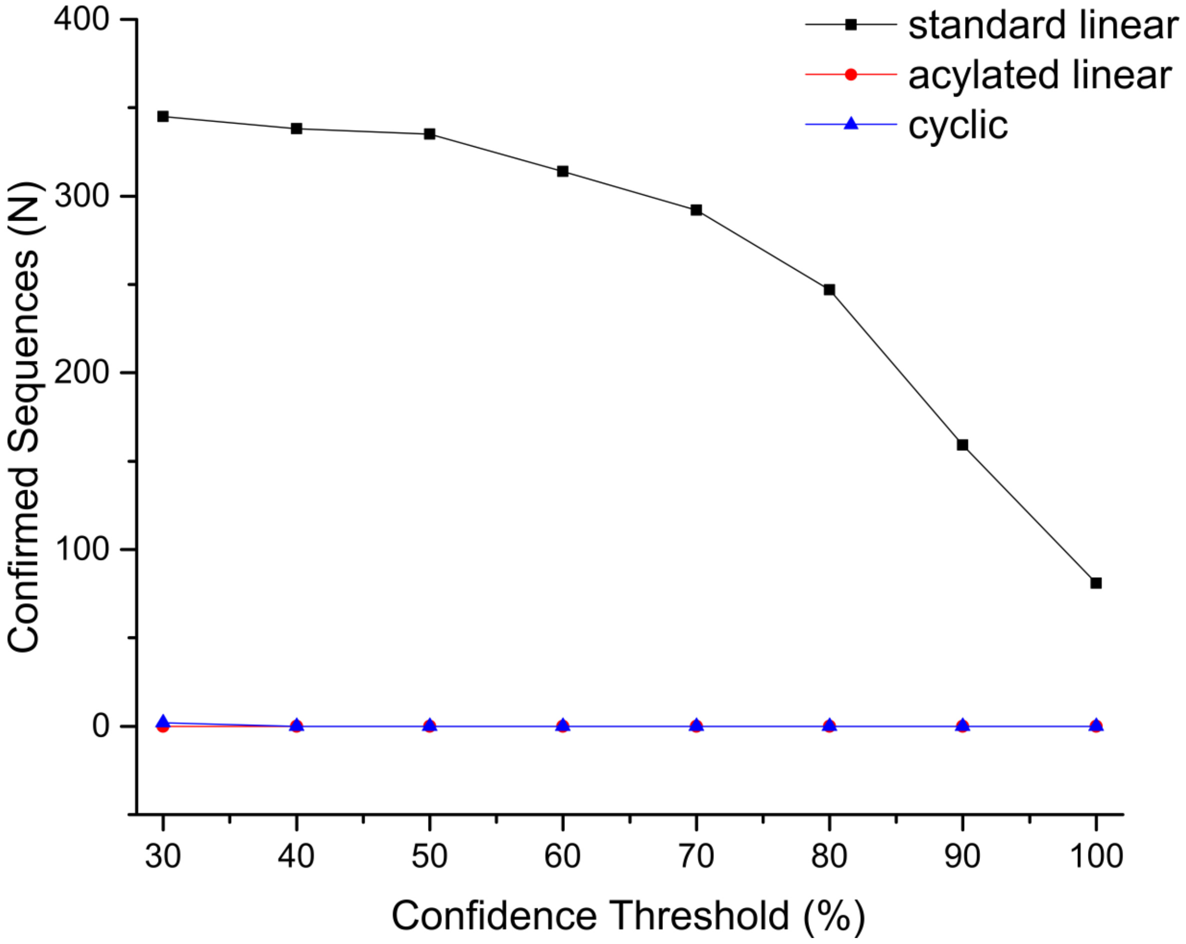


Figure 43: Number of Confirmed Sequences as a Function of Minimum MS2 Fragment Assignment for Products of V + g + Pal (pH 2.5).

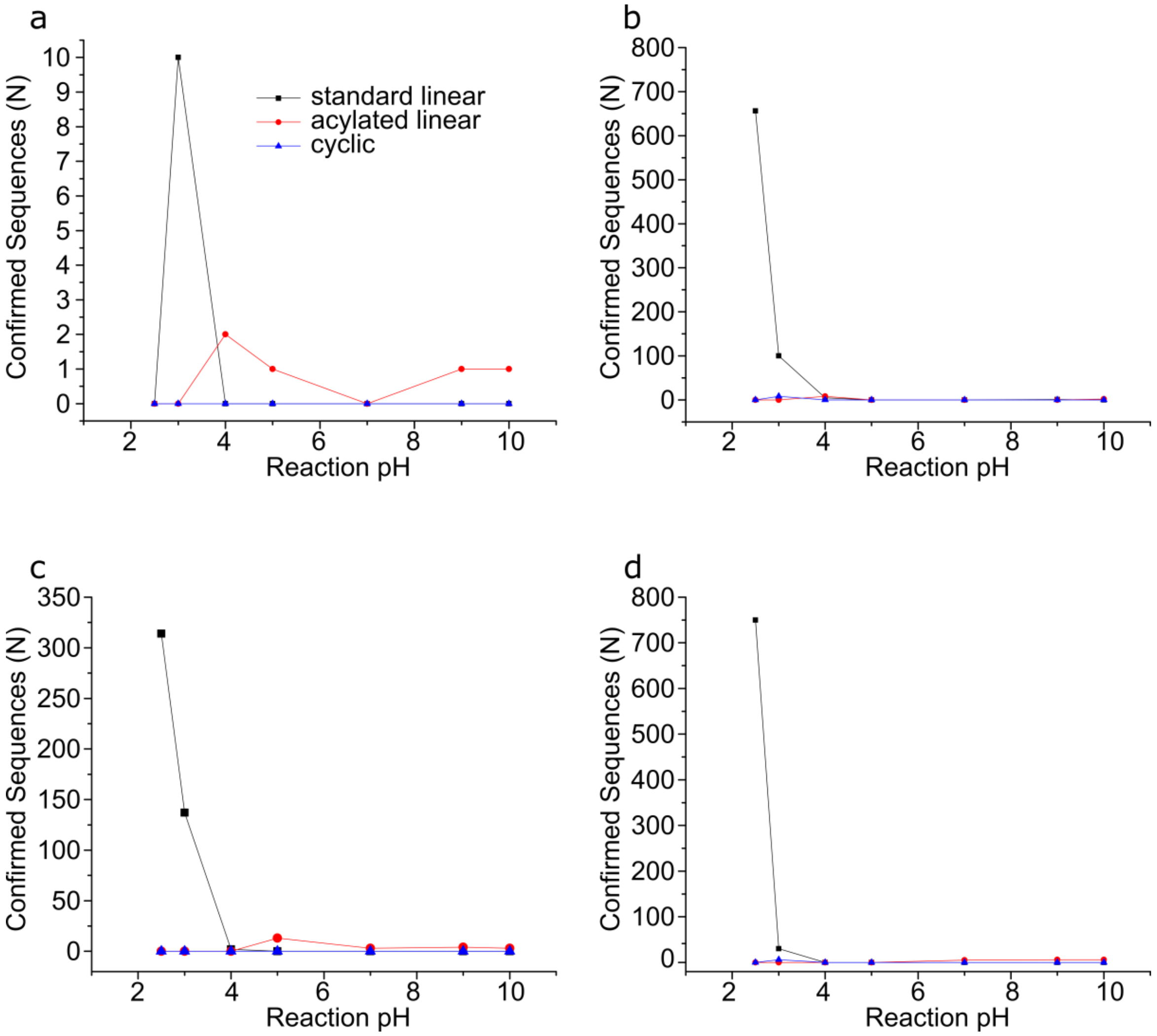


Figure 45: Confirmed Sequences at Confidence Threshold of 60% for Reaction products of V + g with: Ole (a), N + Ole (b), Pal (c), N + Pal (d) at pH Range

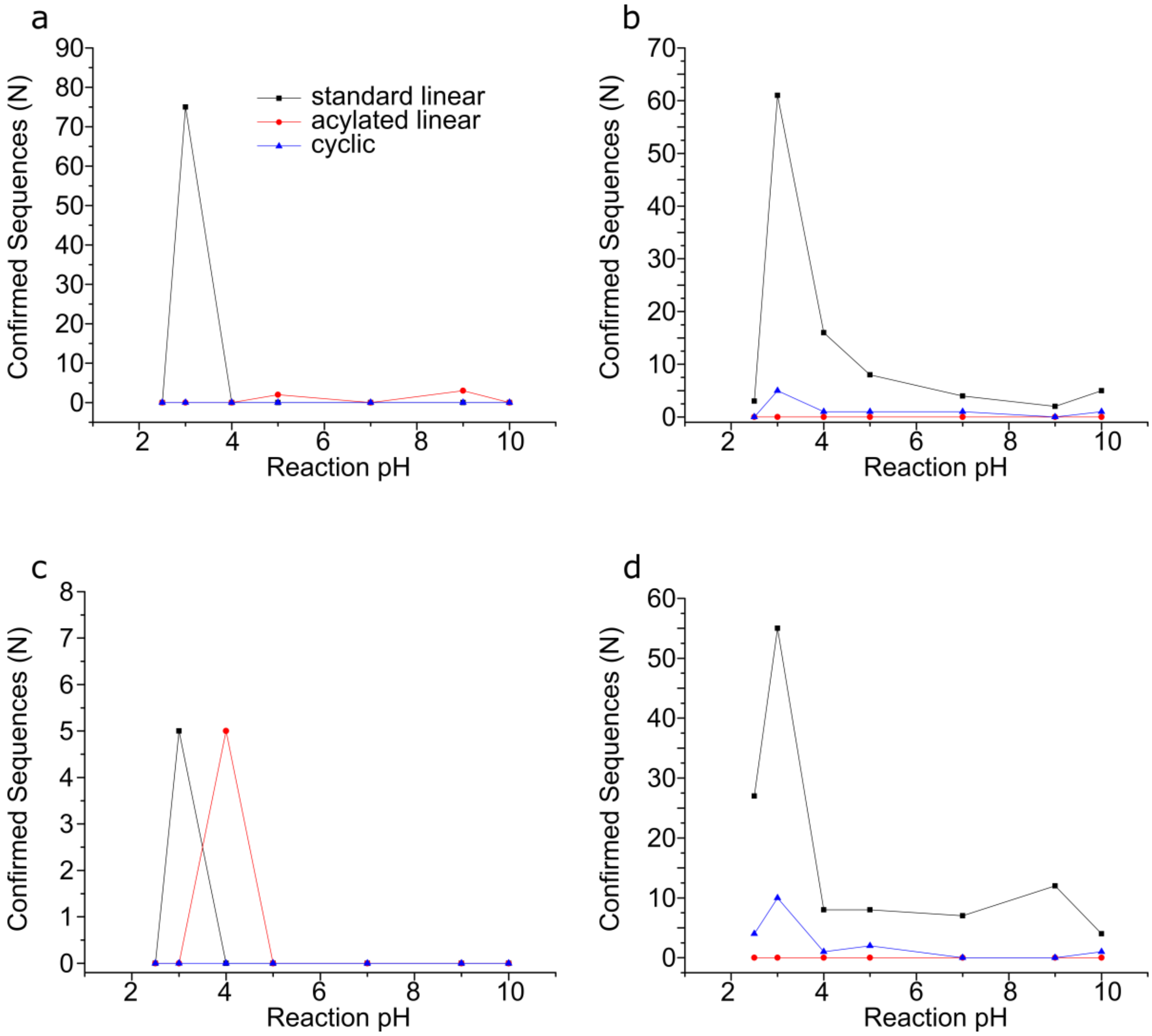


Figure 45: Confirmed Sequences at Confidence Threshold of 60% for Reaction products of G + g with: Ole (a) N + Ole (b), Pal (c), N + Pal (d) at pH Range.

Combined EICs of all non-acylated and acylated sequences confirmed with confidence ≥ 60% also confirmed the late retention times of acylated products (Figure 45). This further validates assignment of these sequences as retention time on a reverse phase column is expected to be greater for products with N-terminally linked large, hydrophobic fatty acid moieties.

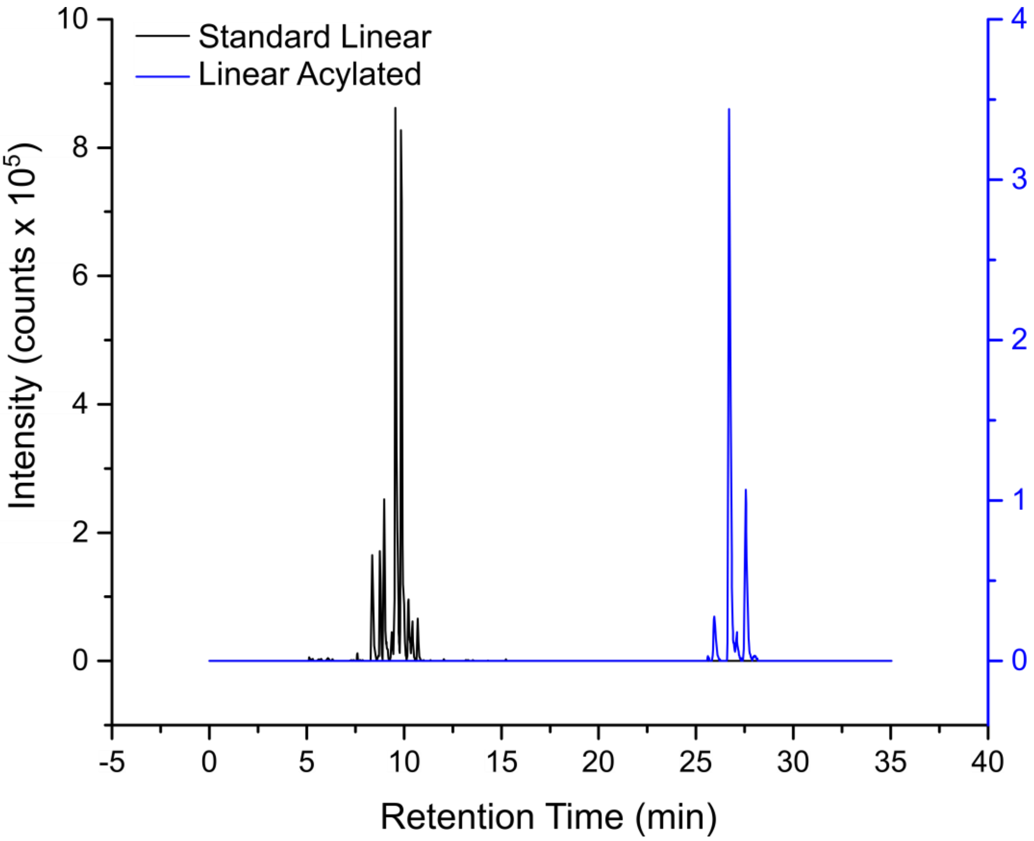


Figure 46: Virtual Base Peak Chromatograms of Confirmed Standard and Acylated Linear Sequences. Summed intensities of MS1 EICs of sequences confirmed at 60% confidence from reaction products of V + g + N + Ole (pH 4) are shown for standard linear sequences (black) and acylated sequences (blue).

9.2 Model System 2: Non-Acylated Peptide and Depsipeptide Oligomerization

We have utilised Oligomersoup to analyse a variety of complex depsipeptide and peptide ‘soups’. In one approach using SPPS methodology, we synthesised peptide mixtures of differing lengths. By using automated SPPS methodologies, peptide elongation was restricted to one monomer per coupling cycle and excess reagents were washed away after each cycle. This restricted the possible lengths of the product, ensured ensures there was a fresh supply of reagents and controlled the combinatorial explosion.

We synthesised a series of peptide mixtures using two approaches. The first involved using a mixture of four different resins and a mixture of the same four amino acids for each coupling reaction. The second synthesis method used only one resin. We chose Fmoc-Phe-Wang for this approach as Phenylalanine has a prominent and characteristic immonium ion, making it easily identifiable. At each coupling stage, a mixture of four amino acids were used. The amino acid present in the resin was not used in this Fmoc-protected amino acid mixture. By only using Phenylalanine at the C-terminus, it enabled us to reduce our sequencing time as it could be set as a ‘terminal group’. By having a residue fixed at one position in the peptide sequence, the potential number of sequences is exponentially reduced. Following synthesis, we used Polyersoup to investigate the changes in sequence preference with increasing length.

HGSFX is used to denote the four resin peptides, and HGSLX is used to denote the peptides synthesised using one resin, where X represents the number of amino acid coupling reactions performed and all other letters are one letter codes of the amino acids used in the reactions. The products of each sequencing run were ranked by intensity (Figure 47), that is the product with the most intense precursor peak was given the rank of 1 and sequentially lower intensity sequences were ranked accordingly. As expected, the more coupling cycles performed, the more sequences confirmed in the product mixture. From the ranked intensity plots we can see that the coupling cycles on one resin were more successful than when there was four as the products are of a more consistent length. Following investigation of the product sequence composition, we noticed that Histidine was heavily represented in the product sequences. This may be because Histidine, as a charged monomer, flies better in the MS and more sequences with histidine are ionised and appear more abundant than other sequences that are less histidine rich. Another trend observed is the sequential repetition of histidine monomers in sequences.

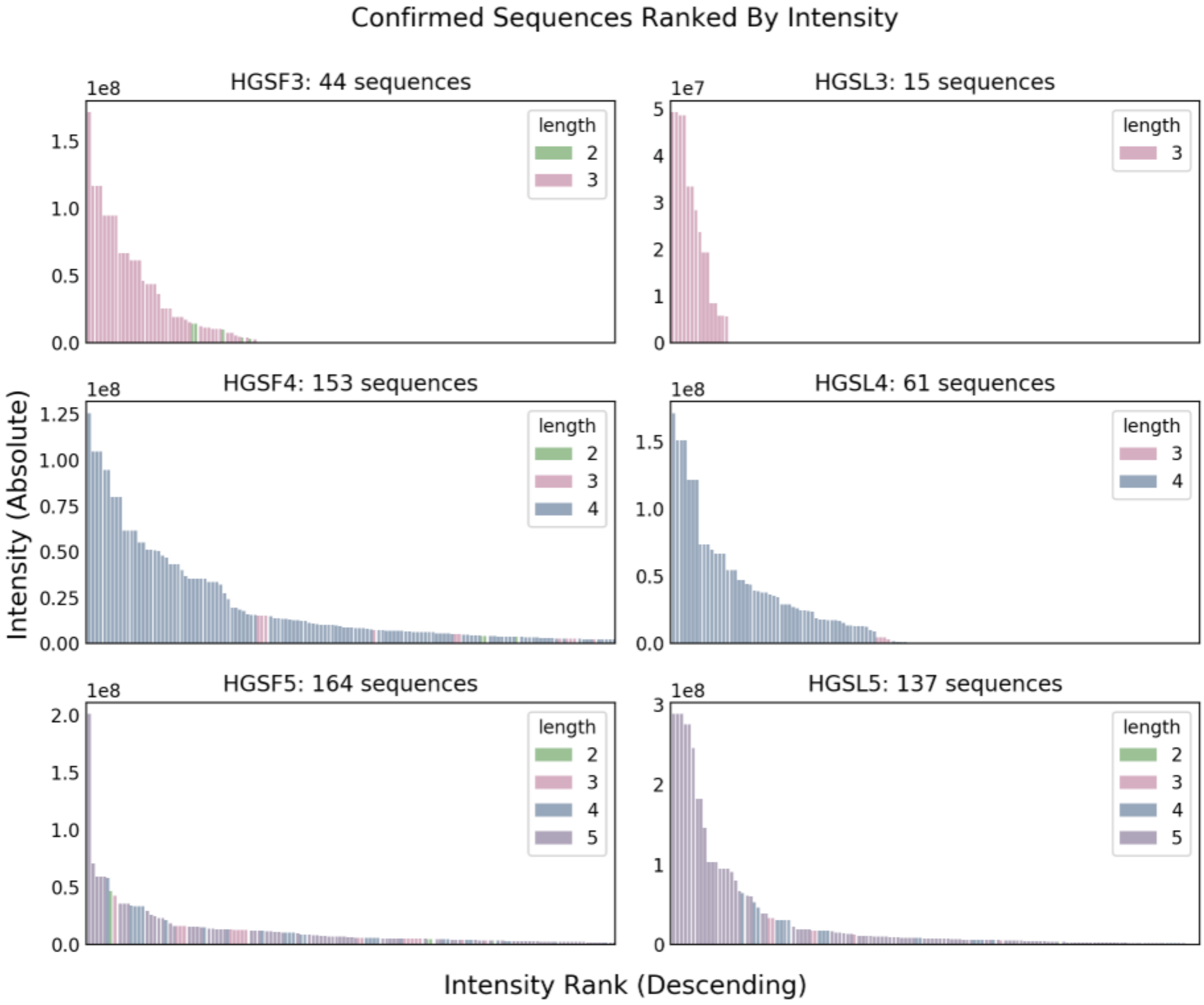


Figure 47: SPPS Soups Intensity Ranks Shown as a Function of Length. Sequences above 1E6 intensity. Title codes represent the four monomers present in the amino acid mix of the coupling stage, followed by the number of coupling cycles carried out.

In different experiments, we used different combinations of three amino acids with glycolic acid to generate depsipeptide mixtures using a single wet-dry cycle (described in Model System 1). The amino acids were chosen based on varying functional group types and shared commonality in biological motifs: glycine (G), cysteine (C), leucine (L), histidine (H), phenylalanine (F) and asparagine (N). Using Oligomersoup we were able to quantify the number of sequences present and identify starting material combinations which produced longer sequences (Figure 48). For example, G+C+H yielded the highest number of confirmed sequences with a product length of up to seven observed.

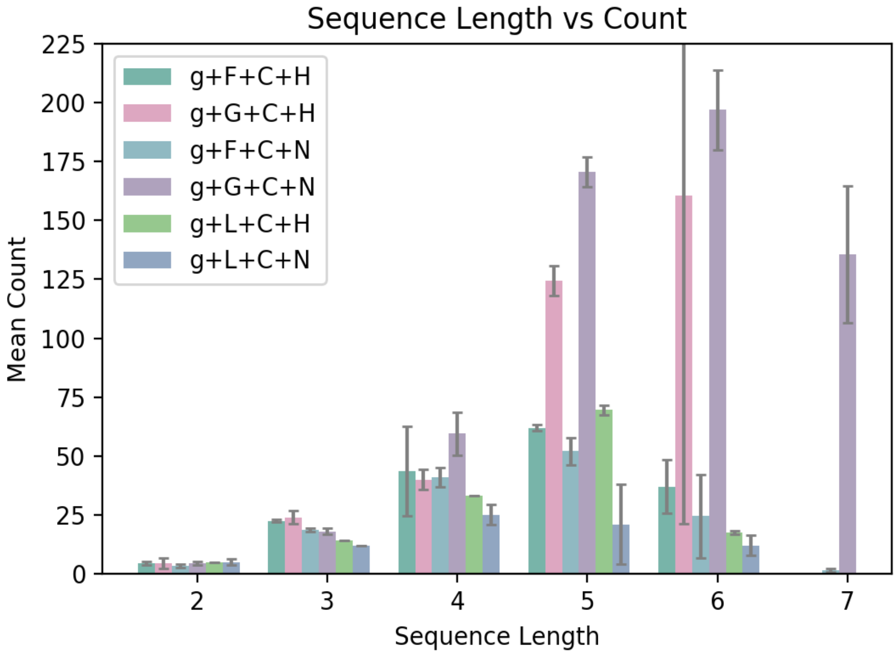


Figure 48: Sequence Length vs. Count of Confirmed Sequences in Depsipeptide Mixtures. Error bars represent the standard deviation between experimental repeats. Minimum confidence = 40 %, subsequence weight = 0.5.

We further analysed the product data obtained by Oligomersoup by looking at sequence space coverage (Figure 49). As seen previously with the standard linear peptides, a large number of confirmed sequences can represent a very small percentage of all possible sequences. Here we quantified the total number of sequences confirmed above 40 % confidence, compared to the number of sequences possible at each length. We observed a trend with all samples, an increase in sequence selectivity with increased sequence length. Using the tool in-house, we have already demonstrated Oligomersoup's ability to be a versatile tool for the analysis of complex mixtures.

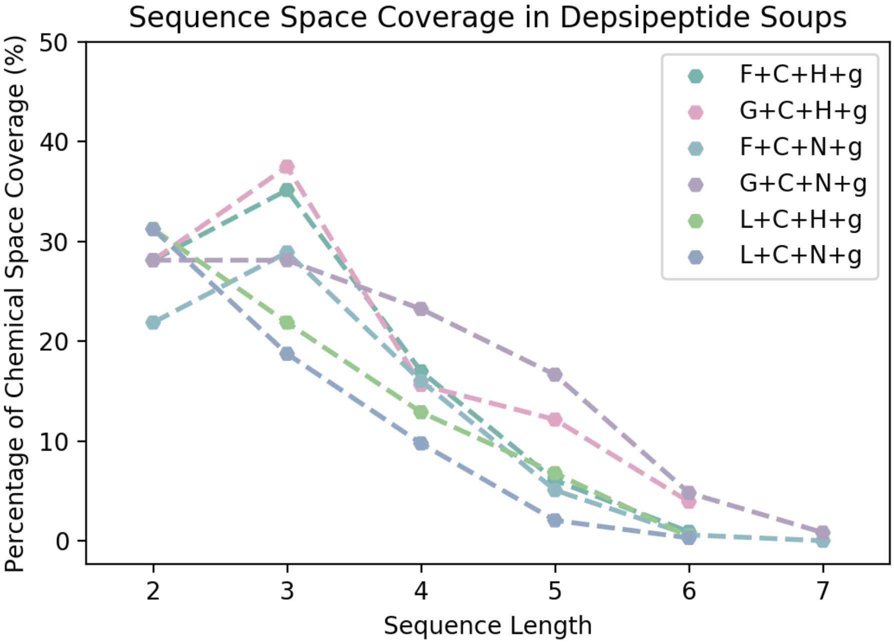
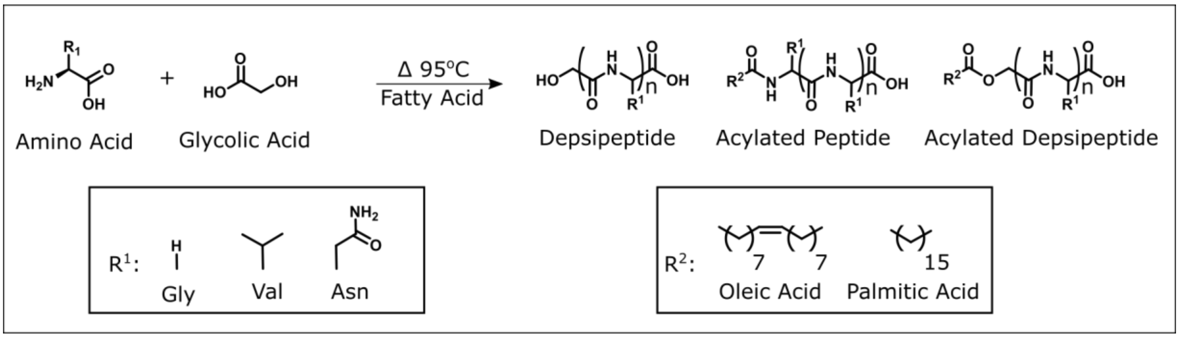


Figure 49: Sequence Length vs. Percentage of Chemical Space Coverage in Depsipeptide Mixtures. Minimum confidence = 40 %, subsequence weight = 0.5.

10. Materials and Methods

10.1 Depsipeptide Model Reactions

To produce a model system of N-terminally acylated and non-acylated (depsi)peptide mixtures (Scheme 1), α-amino acid and α-hydroxy acid monomers were subject to thermal dehydration using methods similar to those described previously in the literature.[11,12]



Scheme 1: Acylated, Non-Acylated Peptide and Depsipeptide Products in a One-Pot Depsipeptide Elongation and Terminal Acylation Reactions.

Depsipeptide starting mixtures were made up with 0.1 M amino acid and 0.1 M glycolic acid (Sigma, CAS: 79-14-1) in HPLC-grade H_2O and adjusted to desired pH using 2M H_3PO_4 or 2M NaOH. Immediately prior to heating at 95 ^oC for 15 hours in open-cap glass vials, 10 mL of pH-adjusted monomer stock was added to 0.33 mL oleic acid (Sigma, CAS: 112-80-1), 0.256 g palmitic acid (TCI, CAS: 112-80-1) or 0.33 mL HPLC-grade H_2O for oleated, palmitated and fatty acid-free reactions respectively. Upon dehydration after heating at 95 ^oC for 15 hours, samples were redissolved in 10 mL HPLC-Grade H_2O. Redissolved products were then sonicated at 45 ^oC for ≥ 15 min. After sonication, 1.2 mL aliquots were harvested and centrifuged at 10,000 rpm for 30 min and the aqueous layer harvested. The harvested aqueous layer was then diluted 1:10 in MS-grade H_2O and filtered into a glass HPLC vial through a 0.22 µm nylon syringe filter.

10.2 Polyimine Model Reactions

To produce a model system of alternating co-polymers, Schiff base polymers (polyimines) were synthesized via uncontrolled oligomerization of diamine and dialdehyde monomers (Table 1).

Monomer stocks were made up to 0.1 M in appropriate solvent (see). 2 mL of each monomer stock (one diamine and one dialdehyde per reaction) was added to a 10 mL glass vial. 6 mL HPLC-grade MeCN was then added to the vial to give a total reaction volume of 10 mL and starting material concentration of 0.02 M for each diamine and dialdehyde. Mixtures were then stirred at 200 rpm and continuously heated at 70 ^{o}C for 30 min. Products were then cooled to 4 ^{o}C prior to 50 % dilution in a 1:1 MeCN:MeOH mixture (MS-grade, + 0.01 % formic acid). Diluted products were filtered through 0.22 µM nylon syringe membrane before analysis via the Orbitrap Lumos Tribrid Mass Spectrometer.

Table 5: Diamine and Dialdehyde Monomers Used for Polyimine Condensation Reactions. Solvent abbreviations are as follows: methanol (MeOH), acetonitrile (MeCN), ethanol (EtOH). All solvents were of HPLC-grade, with the exception of EtOH, which was analytical grade.

Monomer	Solvent
ethylenediamine (e)	1:1 MeOH:MeCN
p-phenylenediamine (p)	70% MeOH + 30% EtOH
p-xylylenediamine (x)	Methanol
glyoxal (o)	1:1 MeOH:MeCN
phthaldialdehyde (h)	MeOH
glutaraldehyde (t)	1:1 MeOH:MeCN

10.3 Mass Spectrometry Data Acquisition

10.3.1 Bruker Maxis Impact II

Unless specified otherwise, all measurements acquired via the Bruker Maxis Impact II Measurements were taken in positive ion mode, with the instrument calibrated to a range of 50-2000 *m/z* using sodium formate calibrant solution. Voltage of the capillary tip was set to 4800 V, end plate offset at -500 V, funnel 1 RF and funnel 2 RF at 400 Vpp, hexapole RF at 100 Vpp, ion energy at 5.0 eV, collision energy at 5 eV, collision cell RF at 200 Vpp, transfer time at 100.0 µs and pre-pulse storage time at 1.0 µs. MS/MS acquisition was carried out using a data-dependent auto-selection of 20 most intense precursors per cycle (minimum precursor intensity = 20000 counts) with a CID collision energy of 35 eV.

Unless specified otherwise, all LC-MS and 'direct injection' analyses with mass spectrometry detection were conducted using a Dionex Ultimate 3000^{TM} system. Direct injection analyses – i.e. analyses without pre-measurement chromatographic separation – were carried out by injecting 10 µl analyte under an isocratic flow of 0.4 ml min^{-1} for a total acquisition time of 3 min. Unless specified otherwise, 95 % MS-grade H_2O + 5 % MS-grade MeCN + 0.1 % formic acid was used the mobile phase for all direct injection analyses. Acquisition time for LC-MS analyses is method-specified, and unless specified otherwise is equal to the total run time of the (U)HPLC method.

Raw Bruker '.baf' files were converted to mzML file format using Bruker Compass Xport 3.0.13.1 called via subprocess from Python 3.7. mzML files were converted to JSON format using unpublished, in-house 'mzml_ripper' package.

10.3.2 Orbitrap Lumos Tribrid

Unless specified otherwise, all measurements acquired via the Orbitrap Lumod Tribid mass spectrometry were carried out in positive mode using DDA to select the most intense ions for tandem mass spectrometry via HCD. To ensure sufficient acquisition of low abundance products, a 1 min dynamic exclusion window was applied with width of 5 ppm.

10.4 Solid Phase Peptide Synthesis

All Fmoc-protected amino acids and Fmoc-protected Wang resins were purchased and used without further purification from NovaBioChem and Sigma-Aldrich. All solvents and reagents were purchased from Sigma-Aldrich. 2 mL reactor vials with frit filters were purchased from Biotage.

Peptide synthesis was performed using the Biotage Syro II automated peptide synthesiser fitted with two 48 reactor blocks. Each 2 mL reactor vial (RV) was loaded with the desired Fmoc-protected Wang resin (0.25 mmol). Each synthesis was repeated in multiple vials across the reactor block to afford a suitable yield. The peptide synthesis proceeded in four stages: swelling, deprotection, coupling and washing.

500 μL Ultrapure DMF was added and each RV was shaken for 1 hour at room temperature. Following the resin swelling, the RVs were drained for 60 seconds using vacuum.

The deprotection was performed in two stages. 500 μL of piperidine solution (20 % v/v in DMF) was added and the RV was shaken at room temperature for 3 minutes. After this first deprotection reaction, the RV was drained and 500 μL of fresh piperidine solution was dispensed into the RV. The second deprotection reaction lasted for 10 minutes, after which the RV was drained. 500 μL Ultrapure DMF was added to the RV and shaken for 60 s, followed by a 60 s drain. The RV was washed this way a further 4 times.

Double coupling was carried out for each amino acid addition. The required amino acid solution (4.0 eq, 0.5 M in DMF) was dispensed into the RV followed by hydroxybenzotriazole (HOBt, 4 eq, 0.5 M in DMF) and N,N′-diisopropylcarbodiimide (DIC, 4 eq, 3M in DMF). The RV was shaken at room temperature for 1 hour. The reagents were then drained and the resin was washed with Ultrapure DMF (500 μL) as previously described. Cycles of deprotection and coupling were repeated with different amino acids until the peptide was of desired composition.

After a final deprotection of the N-terminus amino acid, the resin-bound peptide was washed five times with Ultrapure DMF, as previously described. Following the DMF washing, the peptides were further washed with DCM (500 μL) for 60 s whilst shaking.

The reactor blocks were removed from the Syro II and placed into a fumehood, all subsequent operations were carried out manually. 2 mL of cleavage cocktail (96 % trifluoroacetic acid, 2 % triisopropyl silane, 2% H_2O) was added to each RV and left to shake for approximately 3 hours at room temperature. Following this, the cleaved solution was drained into a 15 mL centrifuge tube. 10 mL of cold diethyl ether was added to the filtrate and the solution was left to precipitate at -20°C overnight. The resulting solid was washed under centrifugation (4.5 minutes, 4000 rpm) three times with 15 mL of cold ether. The ether from the final wash was discarded and the remaining solid was left to dry in a desiccator for at least 15 hours.

Supplementary References

1. Fenn, J. B., Mann, M., Meng, C. K., Wong, S. F. & Whitehouse, C. M. Electrospray ionization for mass spectrometry of large biomolecules. *Science* (1989) doi:10.1126/science.2675315.

2. Parcher, J. F., Wang, M., Chittiboyina, A. G. & Khan, I. A. In-source collision-induced dissociation (IS-CID): Applications, issues and structure elucidation with single-stage mass analyzers. *Drug Test. Anal.* (2018) doi:10.1002/dta.2249.

3. Martin, D. B., Eng, J. K., Nesvizhskii, A. I., Gemmill, A. & Aebersold, R. Investigation of neutral loss during collision-induced dissociation of peptide ions. *Anal. Chem.* (2005) doi:10.1021/ac050701k.

4. Thiede, B. *et al.* Peptide mass fingerprinting. *Methods* (2005) doi:10.1016/j.ymeth.2004.08.015.

5. Wesdemiotis, C. *et al.* Fragmentation pathways of polymer ions. *Mass Spectrom. Rev.* (2011) doi:10.1002/mas.20282.

6. Hohmann, L. J. *et al.* Quantification of the compositional information provided by immonium ions on a quadrupole-time-of-flight mass spectrometer. *Anal. Chem.* (2008) doi:10.1021/ac8006076.

7. Niedermeyer, T. H. J. & Strohalm, M. mMass as a Software Tool for the Annotation of

Cyclic Peptide Tandem Mass Spectra. *PLoS One* (2012) doi:10.1371/journal.pone.0044913.

8. Mohimani, H. *et al.* Sequencing cyclic peptides by multistage mass spectrometry. *Proteomics* (2011) doi:10.1002/pmic.201000697.

9. Gorman, J. J., Wallis, T. P. & Pitt, J. J. Protein disulfide bond determination by mass spectrometry. *Mass Spectrom. Rev.* (2002) doi:10.1002/mas.10025.

10. Liu, F., van Breukelen, B. & Heck, A. J. R. Facilitating Protein Disulfide Mapping by a Combination of Pepsin Digestion, Electron Transfer Higher Energy Dissociation (EThcD), and a Dedicated Search Algorithm SlinkS. *Mol. Cell. Proteomics* (2014) doi:10.1074/mcp.o114.039057.

11. Rodriguez-Garcia, M. *et al.* Formation of oligopeptides in high yield under simple programmable conditions. *Nat. Commun.* (2015) doi:10.1038/ncomms9385.

12. Forsythe, J. G. *et al.* Ester-Mediated Amide Bond Formation Driven by Wet-Dry Cycles: A Possible Path to Polypeptides on the Prebiotic Earth. *Angew. Chemie - Int. Ed.* **54**, 9871–9875 (2015).