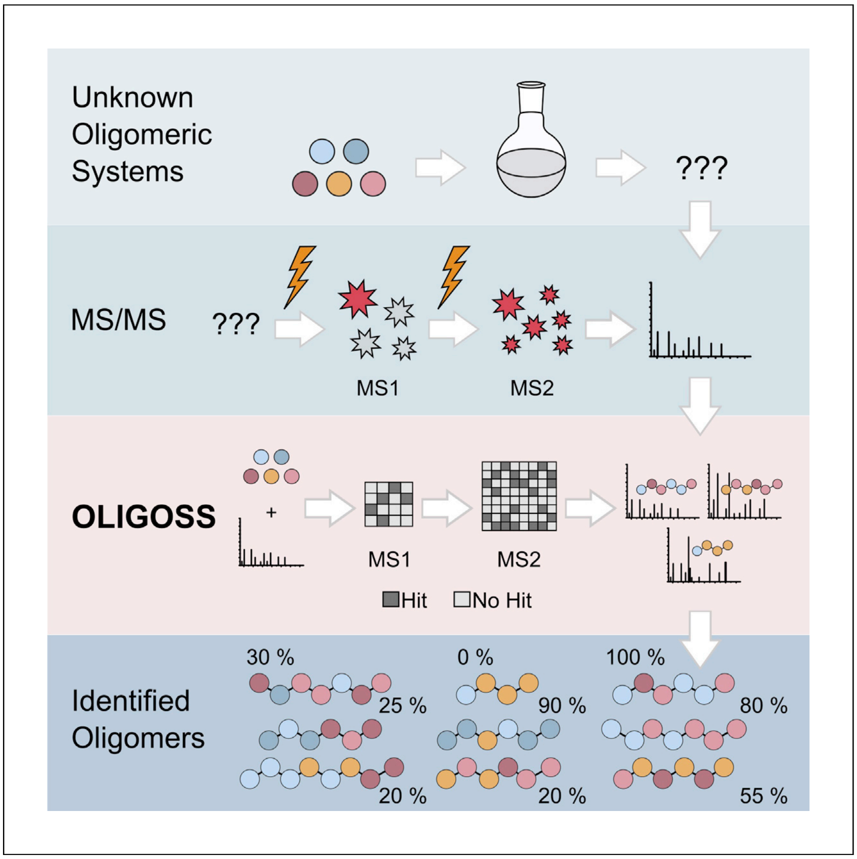
Cell Reports
Physical Science

CellPress
OPEN ACCESS

Article

Exploring the sequence space of unknown oligomers and polymers



Doran et al. present oligomer-soup-sequencing (OLIGOSS), which offers an approach to the sequencing of unknown oligomer systems. Using a set of backbone-agnostic abstract properties to define fragmentation, OLIGOSS is capable of sequencing any linear oligomer class amenable to MS/MS, regardless of backbone chemistry.

David Doran, Emma Clarke, Graham Keenan, Emma Carrick, Cole Mathis, Leroy Cronin

lee.cronin@glasgow.ac.uk

Highlights

New approach to the sequencing of unknown oligomer systems

Sequencing for any linear oligomer classes amenable to MS/MS

Validation using synthetic peptides, polyesters, polyimines, and depsipeptide oligomers

Used to map RNA methylation sites and sequence a ribosomally synthesized thioholgamide

Doran et al., Cell Reports Physical Science *2*, 100685
December 22, 2021 © 2021 The Author(s).
https://doi.org/10.1016/j.xcrp.2021.100685

Cell Reports
Physical Science

CellPress
OPEN ACCESS

Article

Exploring the sequence space of unknown oligomers and polymers

David Doran,[1,2] Emma Clarke,[1,2] Graham Keenan,[1] Emma Carrick,[1] Cole Mathis,[1] and Leroy Cronin[1,3,*]

SUMMARY

The characterization of the chemistry of life on earth has been facilitated by developments in analysis and sequencing of bio-oligomers using tandem mass spectrometry (MS/MS). Bio-oligomers can be identified with sequence-level resolution in analytes more complex than any synthetic mixture, enabled by well-established knowledge of fragmentation properties and extensive MS/MS databases built up over decades. However, unknown oligomer systems remain difficult to characterize, as no comparable databases exist, partly because of the vast chemical diversity and fragmentation pathways. Here, we present oligomer-soup-sequencing (OLIGOSS), a new approach to the sequencing of unknown oligomer systems. Using a novel set of backbone-agnostic abstract properties to define fragmentation, OLIGOSS is capable of sequencing any linear oligomer class amenable to MS/MS, regardless of backbone chemistry. We validated OLIGOSS by sequencing synthetic peptides, polyesters, polyimines and depsipeptide oligomers, mapped RNA methylation sites, and a ribosomally synthesized peptide, thioholgamide, directly from a cell lysate without purification.

INTRODUCTION

Analytical technologies have enabled the rapid, accurate, and cheap analysis of biological oligomers and polymers with sequence-level resolution.[1] This has revolutionized many fields in biology, chemical biology, and medicine.[2] Several of these technologies, particularly proteomics, rely on mass spectrometry (MS) and tandem mass spectrometry (MS/MS) to identify specific bio-oligomer sequences.[1] Because of the complexity of these analytes, oligomers are usually separated by liquid chromatography prior to MS/MS (LC-MS/MS).[3,4] Knowledge of the MS ionization and MS/MS fragmentation properties of oligonucleotides and peptides allows sequencing of these oligomers from very complex, heterogeneous mixtures.[5] Despite the comparative simplicity of synthetic oligomer mixtures, and the fact that many synthetic oligomer classes have well-characterized fragmentation pathways, general tools do not exist for sequencing synthetic oligomer mixtures.[6] In our work exploring prebiotic chemistry through untargeted chemical synthesis,[7] we realized that even the identification of very simple systems would be prohibitively challenging. A tool for matching the precision, throughput, and sequence-level resolution of biological systems for unknown oligomers would not only be important for prebiotic chemistry but also in exploring new developments in materials science, polymer and oligomer chemistry, and other fields.[8] Examples include the analysis of oligomers synthesized via sequence-controlled oligomerization, sequencing of oligomeric natural products, and finally the identification of oligomeric contaminants and degradation products in polymer materials.

[1] School of Chemistry, University of Glasgow, Joseph Black Building, University Avenue, Glasgow, G12 8QQ, UK

[2] These authors contributed equally

[3] Lead contact

*Correspondence: lee.cronin@glasgow.ac.uk

https://doi.org/10.1016/j.xcrp.2021.100685

Despite the potential for oligomeric sequencing, developing new approaches has been challenging for two key reasons. First, the sequence space is vast because of the combinatorial nature of possible oligomeric sequences that must be screened. Sequence space increases dramatically as a function of oligomer length and number of unique monomers (Equation 1). Hence it is challenging to extensively survey the full range of possible products in synthetic oligomer mixtures.

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

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

A more fundamental conceptual problem is the diversity of chemistries and fragmentation pathways in unknown oligomers.[9] For peptides and nucleotides, extensive databases exist for matching spectra to unknown target sequences. Biological sequencing tools are therefore limited to the two classes of bio-oligomers for which these data exist: peptides and oligonucleotides.[10,11] Biological sequencing tools can also leverage knowledge of peptide and oligonucleotide MS/MS fragmentation. Non-canonical backbones, even those that are similar to peptides and found commonly in natural products, such as depsipeptides, are outside the scope of these tools.[12] Several tools exist for sequencing synthetic oligomers, using rules that are specific to a single class of oligomers.[13–16] However, ideally oligomeric sequencing must not be limited in scope to a specific subset of oligomers. Instead, it should be amenable to sequencing as wide a variety of oligomer classes as possible, regardless of their specific backbone or sidechain chemistries.

A generalized OLIGOSS concept is shown in Figure 1: the user provides mass spectrometry data (in .mzML or .json format) and the constraints of the oligomer space to be screened as an input file in the .json format. This file contains information unique to the sequencing experiment, such as monomer identity, backbone chemistry, possible fragmentation series, and parameters for pre-filtering spectra. Pre-filtering parameters are determined by knowledge of instrument conditions (such as expected signal intensities). An instrument configuration file is used to store and validate instrument-based parameters and restrictions (Supplemental experimental procedures S4; Figure S12). A comprehensive description of all input parameters can be found in Supplemental experimental procedures S5 and Figures S13 and S14. An oligomer configuration file is then selected on the basis of the specified backbone chemistry, and these abstract properties, written using OLIG formalisms, are used to create subsequent silico dictionaries. For guidance on how to construct a polymer configuration file, see Supplemental experimental procedures S3 and Figures S2–S11. To overcome the performance limitations associated with a full exhaustive search of the sequence space, OLIGOSS first pre-filters spectra on the basis of user-defined thresholds and performs a compositional search. For compositions that have been matched to MS1 peaks, all possible isomeric sequences and their corresponding silico dictionaries containing all possible theoretical MS2 peaks are generated. OLIGOSS then searches for these theoretical peaks in MS2 spectra with matching parent ions for a given sequence. Each sequence is then assigned a confidence score, a percentage-based metric that can be summarized to be the number of fragments confirmed compared with all theoretical fragments. Full details of the confidence calculations can be found in Supplemental experimental procedures S5.6.

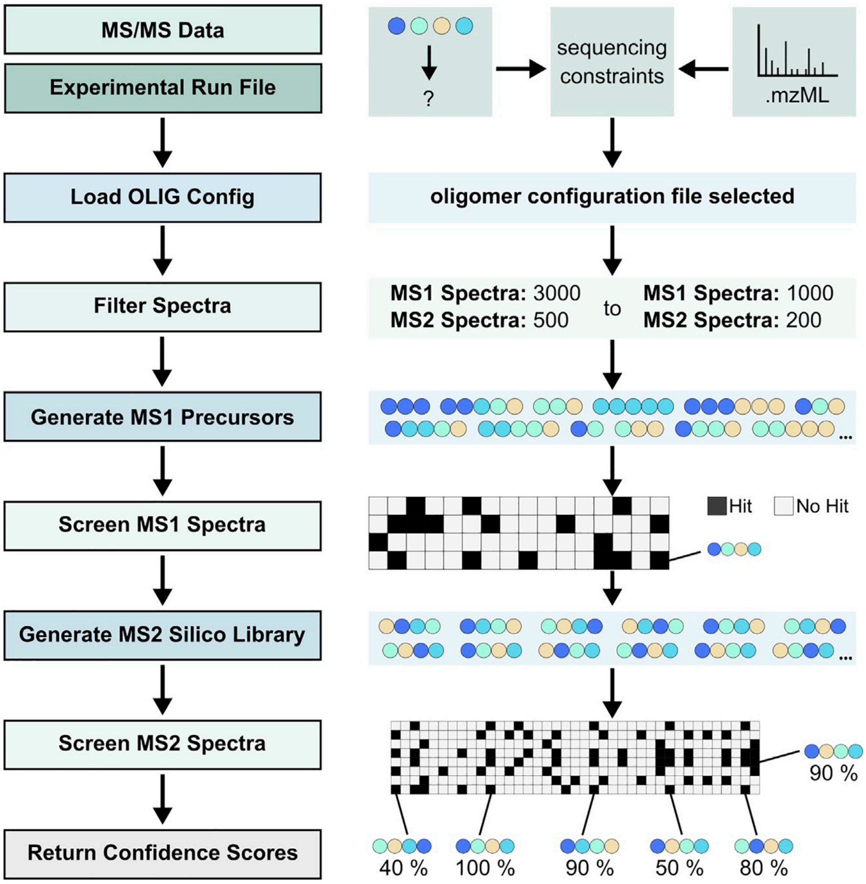


Figure 1. Standard OLIGOSS sequencing workflow

Standard OLIGOSS sequencing workflow for identifying sequences in synthetic oligomer mixtures.

Rather than hard-code rules for each individual fragment for each of these oligomer classes, OLIGOSS uses a set of abstract properties that have been designed to represent all possible oligomer fragmentation events, regardless of specific backbone chemistries. These abstract properties are represented in a set of formalisms known as OLIG, a new and very simple set of principles for defining oligomer MS/MS. We propose that any linear oligomer amenable to ionization and fragmentation via MS/MS will be amenable to sequencing using these OLIG formalisms. Standard MS/MS fragment nomenclatures have been proposed for synthetic oligomers,[9] analogous to standards that are well established for peptides[5,17,18] and other bio-oligomers.[19,20] However, to our knowledge, a set of generalized principles to describe and translate the pathways that produce these fragments into a universal framework for defining oligomer fragmentation has never been produced before. OLIGOSS achieves this by using the set of abstract properties that make up OLIG. 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.

There are often several fragmentation pathways for a single oligomer class,[9] and each of these can produce qualitatively different fragments; the relative predominance of each pathway will vary depending on instrumental conditions. OLIG can be used to define individual fragmentation pathways, and several of these translated

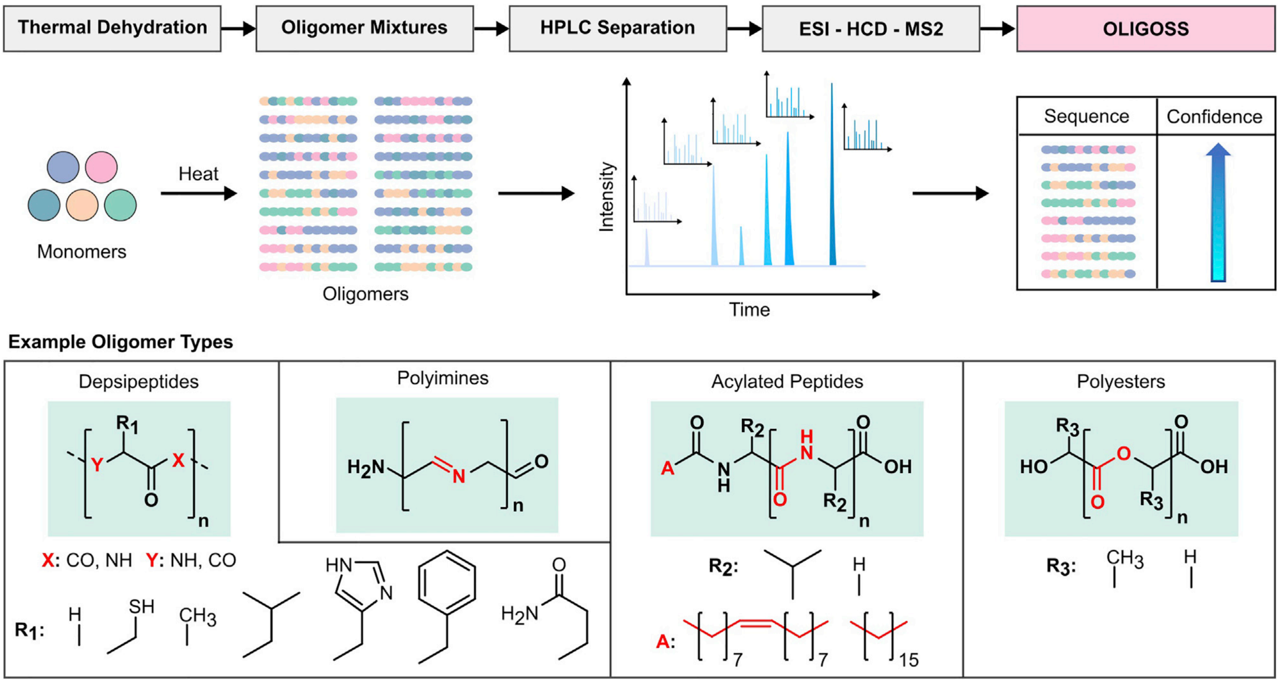


Figure 2. Example experimental workflow

Example experimental workflow before OLIGOSS analysis and reaction systems previously sequenced by OLIGOSS.

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). 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. For a full list of abstract properties, see Supplemental experimental procedures S1 and S2, Figure S1, and Tables S1 and S2, which can be considered a novel yet simple framework for describing oligomer fragmentation via MS/MS.

To test the universality of OLIGOSS, several model oligomer systems were used to validate it as a sequencing tool for both single standards and mixtures: peptides, depsipeptides, polyesters, acylated peptides, polyimines, and RNA (Figure 2). The variation in backbone chemistries is beyond the capabilities of any existing biological omics tool or other sequencing software.[6]

RESULTS AND DISCUSSION

Sequencing of peptide standards

First, to benchmark OLIGOSS's sequencing abilities, and its deployment of abstract OLIG properties in representing well-known fragmentation mechanisms, peptide standards were analyzed in blind sequencing runs, with the only input to OLIGOSS being potential constituent monomers and length of target sequences. Typical backbone fragmentation as well as sidechain-specific fragmentation events were successfully translated into OLIG configuration files. Eight peptide standards with proteinogenic monomers (ASGNQ, FSGNQ, GSGNQ, ASGNQSGV, FSGNQSGV, GSGNQVGS, FSGNQSGVSA, and FSGNQVGSAS) as well as three with non-proteinogenic monomers (cvevvveG, GevcvvvG, and vvecveeG) were synthesized,

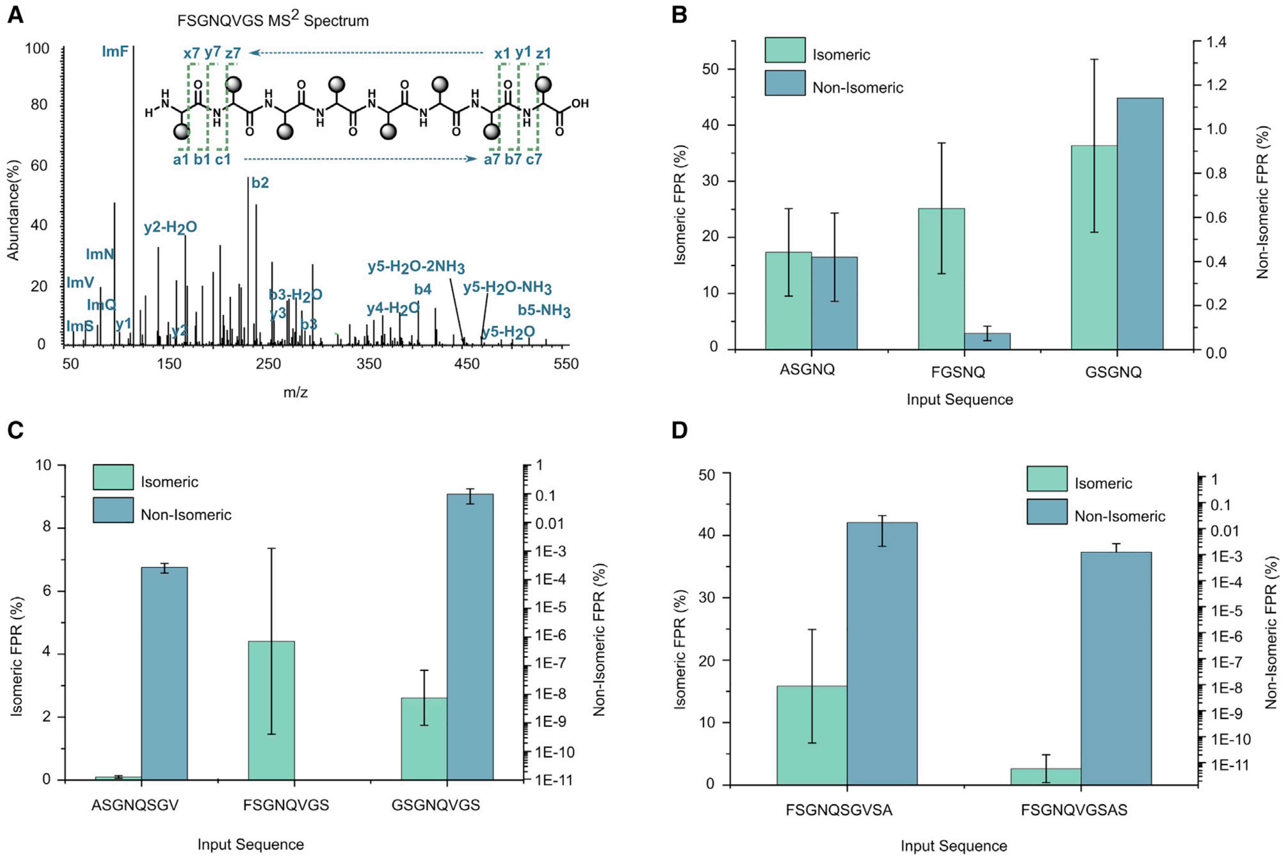


Figure 3. Annotated MS2 spectrum and benchmarking

(A–D) Annotated MS2 spectrum for FSGNQVGS (A) and OLIGOSS benchmarking on peptide standards (B–D). Data in (B), (C), and (D) represent the mean of five measurements ± 1 SD.

analyzed using electrospray ionization with MS2 fragmentation via higher energy C-trap dissociation (ESI-HCD MS/MS), and then subjected to blind OLIGOSS sequencing runs (see Supplemental experimental procedures S6, Figure S15, and Table S3 for the non-proteinogenic standards data). This gave a total sequence space to screen for standards of up to 2.82×10^8 unique sequences (Equation 1). Proteinogenic standards were chosen because of their neutral loss-prone side-chains, thus increasing the average number of unique MS1 precursor and MS2 product ions to be screened for each run.[21] Neutral losses were predominant in MS2 product ion spectra for these standards (see Figure 3A for an example spectrum).

Sequence assignments, covering the entirety of isomeric and non-isomeric sequence space, were ranked by confidence score (see Supplemental experimental procedures S5.6 for confidence scoring calculation). False positive rates (FPRs) were defined as the percentage sequence hits that were not the correct sequence with a confidence score equal to or greater than the correct sequence. Despite variations in the absolute confidence scores of the various standards, FPRs were low for pentameric standards (Figure 3B). For octameric sequences ASGNQSGV, FSGNQSGV, and GSGNQSGV, the sequence space covered in the blind runs was equal to 10,080 isomeric and 1.67×10^6 non-isomeric sequences. Isomeric FPRs were less than 10% for all octameric standards. Very few non-isomeric false assignments were made for the octamers, never exceeding an FPR of 0.1% (Figure 3C). For decameric standards FSGNQSGVSA and FSGNQVGSAS, the total sequence space screened was 279,138 isomeric and 2.82×10^8 non-isomeric sequences (Figure 3D). FPRs for non-isomeric sequences were similar to the octameric standards, not exceeding 0.1% for non-isomeric sequences (Figure 3C).

Isomeric sequence FPRs for the decameric standards were comparable with the pentameric standards (Figure 3B), ranging from 2.23% to 15.84% for FSGNQSGVSA and FSGNQVGSAS, respectively. The comparably high isomeric FPR for the decameric standards is likely due to gaps in fragment series observed in raw MS2 spectra (Figure 3A), which is not atypical for longer oligomers.[9] To demonstrate that OLIGOSS can handle multiple MS2 fragmentation methods, two peptides fragmented via electron-transfer dissociation (ETD) were also successfully sequenced with similar FPRs to peptides fragmented via HCD (see Supplemental experimental procedures S6). At the time of writing, this decameric sequence space is the largest tackled by OLIGOSS. For the blind runs, there were 282 million sequences possible, and for the five MS data files screened (several gigabytes of data), the total run time was about 20 h, with negligible memory use. For more complex samples, run time would scale according to the number of monomers present, the number of spectra to screen, and how many MS1 hits are identified. OLIGOSS allows the user to control the number of processor cores used as well as the maximum memory use. These parameters can also be used to decrease run time.

Sequencing of depsipeptide mixtures

OLIGOSS was designed for *de novo* sequencing of complex, polydisperse oligomer mixtures, particularly those that cannot be sequenced using current targeted approaches. Therefore, to develop and validate this new tool, product mixtures containing a diversity of species and possible fragmentation pathways were required. Four model systems of oligomer mixtures were used: depsipeptides, N-terminally acylated peptides, polyesters, and polyimines. Depsipeptides are peptidic oligomers with a mixture of amide and ester backbone linkages. Despite having similar MS/MS fragmentation pathways to pure peptides, and their occurrence in many well-characterized natural products, current proteomics software tools are unable to sequence depsipeptides.[12,22] Using simple wet-dry cycles of amino acid and α-hydroxy acid monomers, polydisperse depsipeptide mixtures can be produced with high yield.[12,23] 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 analyzed using LC-MS/MS and sequenced using OLIGOSS. Because of the combinatorial nature of sequence space in these reactions (Equation 1), diverse oligomer sequences of various lengths were produced. OLIGOSS was able to successfully survey this sequence space for all reactions tested, determining the proportion of ester linkages and proportion of total sequence space represented (Figure 4) 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. 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. Depsipeptides represent a well-known class of natural products.[24]

RNA methylation site mapping and natural product identification

Another biological use case for OLIGOSS is the mapping of RNA methylation sites, which are of direct relevance to epigenetics and disease and are difficult to detect

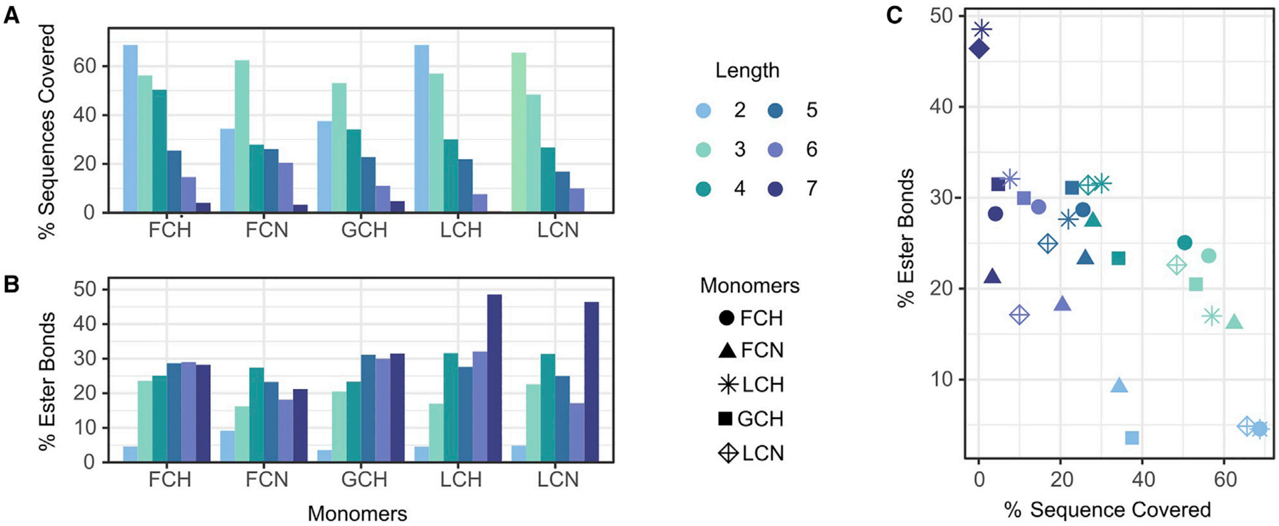


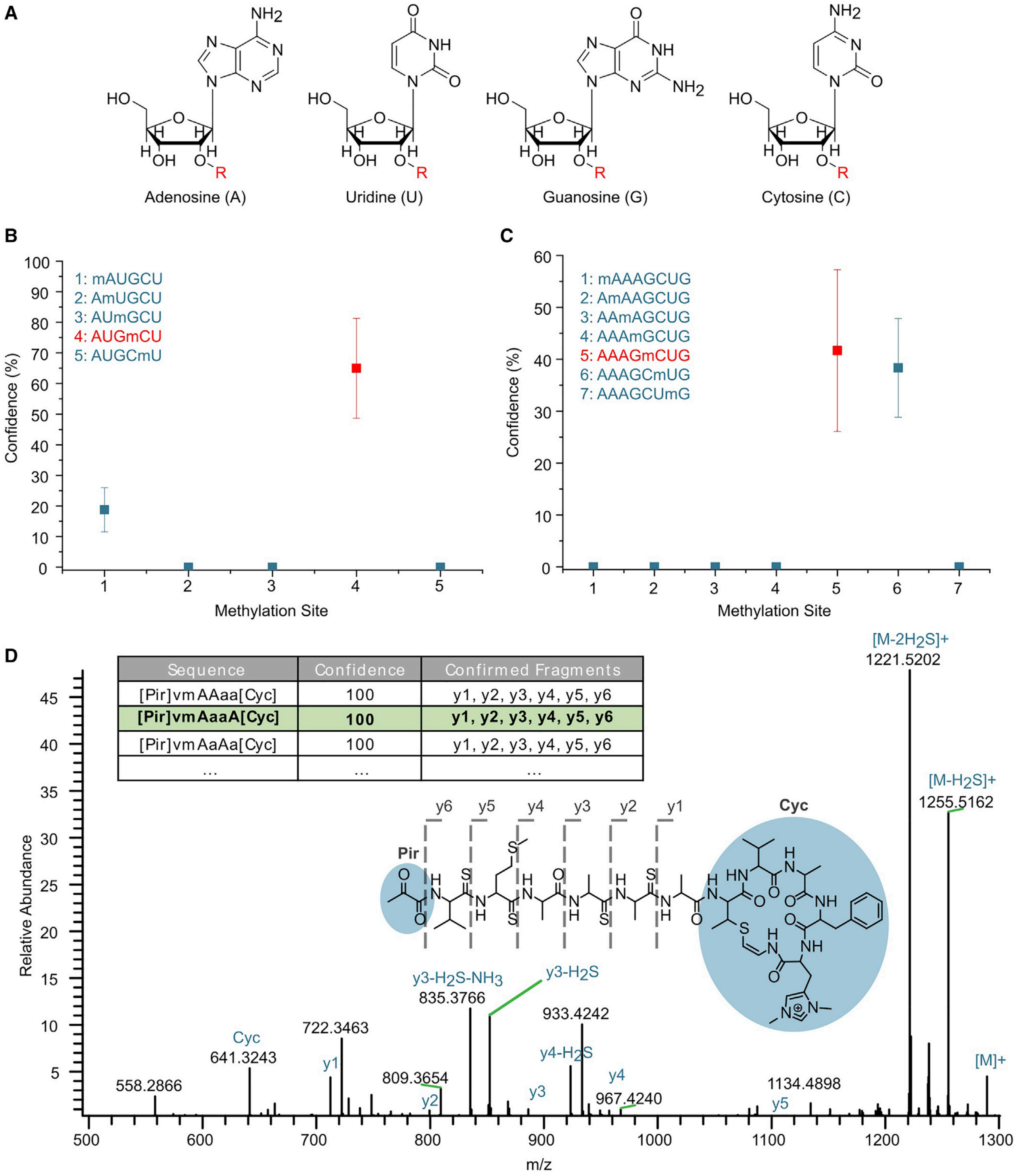
Figure 4. Sequence diversity

Sequence diversity of depsipeptide mixtures surveyed by OLIGOSS.

using standard next-generation RNA sequencing methods. To this end, two RNA standards with one methylated monomer each were sequenced in blind runs in the same manner as the peptide standards (Figure 5). OLIGOSS was able to successfully map the methylation sites in standards AUGmCU (Figure 5B) and AAAGmCUG (Figure 5C), with the correct methylation sites being assigned the highest confidence score in each of the blind runs. Unmethylated RNA standard AUGCUGG was also successfully sequenced in a blind run, with mean isomeric and non-isomeric FPRs of 26.2% and 1.67%, respectively.

Finally, we decided to see if OLIGOSS could help sequence an example of a thioholgamide in a whole-cell lysate. Thioholgamides are ribosomally synthesized and post-translationally modified peptides (RiPPs) and belong to the thioamitide family.[25] They are potent cytotoxins that display anti-proliferative activity against a range of cancer cell lines and were originally isolated from *Streptomyces malaysiense* MUSC 136 (DSM 100712). Targeted knockout studies of the biosynthetic gene cluster led to the production of a new thioholgamide derivative lacking the β-hydroxy group on the N1, N3-dimethylhistidinium moiety found in these compounds.[26] To test the ability of OLIGOSS to assign the correct sequence to this class of natural products, a cell extract containing a thioholgamide with four unique monomers was successfully sequenced using OLIGOSS (see Figure 5D). Of 180 potential isomers, 3 (including the correct sequence) were assigned with 100% confidence, giving an isomeric FPR of 1.67%. An isomeric FPR of 5.0% was also obtained for the hydroxylated form of this thioholgamide, detected in the same cell lysate (see Supplemental experimental procedures S9; Figures S24 and S25).

OLIGOSS is the first tool for automated, *de novo* oligomer sequencing from tandem mass spectrometry data with the ability to sequence oligomers regardless of backbone chemistries. We have demonstrated the ability of OLIGOSS to sequence four oligomer classes of direct biological relevance: peptides, depsipeptides, methylated RNA, and a thioholgamide in cell extract, the latter three of which current biological omics tools are unable to sequence reliably. In addition, OLIGOSS has also been tested on mixtures of polyimines and polyesters (see Supplemental experimental procedures S7; Figure S16). Unlike other tools that are bespoke for a single



Sequence	Confidence	Confirmed Fragments
[Pir]vmAAaa[Cyc]	100	y1, y2, y3, y4, y5, y6
[Pir]vmAaaA[Cyc]	**100**	**y1, y2, y3, y4, y5, y6**
[Pir]vmAaAa[Cyc]	100	y1, y2, y3, y4, y5, y6
...	...	...

Figure 5. Mapping methylation sites and spectral assignments

(A–D) Mapping methylation sites for 2′-o-methylated RNA standards (A–C) and manual spectrum assignments versus OLIGOSS sequencing for thioholgamide sequence vmAaaA (D). Data in (B) and (C) represent the mean of five measurements ± 1 SD.

oligomer class,[12,27] or are unable to perform sequencing,[7,28] OLIGOSS has the potential for expansion to any set of oligomers that are amenable to analysis via MS/MS. All the code used here will be made available for use and editing subject to a GNU General Public License version 3 (GPLv3) license, in the hope that others may benefit from OLIGOSS and use it to sequence even more classes of oligomers that 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 aims to be a first step toward of the exploration of oligomeromics.

EXPERIMENTAL PROCEDURES

Resource availability

Lead contact

Further information and requests for resources should be directed to and will be fulfilled by the lead contact, Leroy Cronin (lee.cronin@glasgow.ac.uk).

Materials availability

Suppliers for chemicals used in our experiments are detailed alongside the reagent in the corresponding experimental sections below.

Data and code availability

OLIGOSS source code can be viewed on GitHub (https://github.com/croningp/oligoss). A copy of a test dataset and a configuration file along with a very basic tutorial can be found at https://zenodo.org/record/4252732#.YZxOtdDMJhE.

Software and data analysis

OLIGOSS was written in Python 3.8.7. A description of the OLIGOSS package structure is detailed in Supplemental experimental procedures S8 and Figures S17–S23. All figures from OLIGOSS sequencing output were generated using Python package Seaborn 0.11.0 and OriginPro 2016.

Depsipeptide model reactions

To produce a model system of N-terminally acylated and non-acylated (depsi)peptide mixtures (Scheme S1), α-amino acid and α-hydroxy acid monomers were subject to thermal dehydration using methods similar to those described previously in the literature (12, 23). Depsipeptide starting mixtures were made up with 0.1 M amino acid and 0.1 M glycolic acid (Sigma; CAS: 79-14-1) in high-performance liquid chromatography (HPLC)-grade H_2O and adjusted to desired pH using 2M H_3PO_4 or 2M NaOH. Immediately prior to heating at 95°C for 15 h in open-cap glass vials, 10 mL 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°C for 15 h, samples were redissolved in 10 mL HPLC-grade H_2O. Redissolved products were then sonicated at 45°C for $\geq$15 min. After sonication, 1.2 mL aliquots were harvested and centrifuged at 10,000 rpm for 30 min, and the aqueous was 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.

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. Two milliliters of each monomer stock (one diamine and one dialdehyde per reaction) was added to a 10 mL glass vial. HPLC-grade MeCN (6 mL) 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°C for 30 min. Products were then cooled to 4°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.

Mass spectrometry data acquisition

Unless specified otherwise, all measurements acquired using the Orbitrap Lumos Tribrid mass spectrometer 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 peptide standards except FSGNQSGV, FSGNQSGVSA, and FSGNQVGSAS were synthesized using the Biotage Syro II automated peptide synthesizer fitted with two 48-reactor blocks. For synthesis of aforementioned "FSGNQ" peptides, see Supplemental experimental procedures S11.4.

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. Two milliliter reactor vials (RVs) with frit filters were purchased from Biotage. Each 2 mL reactor vial 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.

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

The deprotection was performed in two stages. Piperidine solution (500 μL, 20% v/v in DMF) was added, and the RV was shaken at room temperature for 3 min. After this first deprotection reaction, the RV was drained, and 500 μL fresh piperidine solution was dispensed into the RV. The second deprotection reaction lasted for 10 min, after which the RV was drained. Ultrapure DMF (500 μL) was added to the RV and shaken for 60 s, followed by a 60 s drain. The RV was washed this way a further four 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 h. 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-terminal amino acid, the resin-bound peptide was washed five times with Ultrapure DMF, as previously described. Following DMF washing, the peptides were further washed with DCM (500 μL) for 60 s while shaking.

The reactor blocks were removed from the Syro II and placed into a fume hood, and all subsequent operations were carried out manually. Two milliliters cleavage cocktail (96% trifluoroacetic acid, 2% triisopropyl silane, 2% H_2O) was added to each RV and left to shake for approximately 3 h at room temperature. Following this, the cleaved solution was drained into a 15 mL centrifuge tube. Cold diethyl ether (10 mL) 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 min, 4,000 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 h.

RNA standards

Lyophilized RNA standards were purchased from Integrated DNA Technologies (Coralville, IA). Standards were dissolved in 80/20 v/v mass spectrometry-grade water and acetonitrile prior to analysis. Samples were infused directly into a heated ESI source (Thermo Fisher Scientific) with a −3.3 kV voltage applied. Fragmentation was carried out using HCD at a fixed energy of 35%.

Thioholgamide synthesis

The thioholgamide was produced by *S. lividans* YA8 from a heterologously expressed thioholgamide biosynthetic gene cluster[26,27]. The main thioholgamide species was thioholgamide A (*m*/*z* 1305), and the crude extract was used for sequencing.

Mass spectrometry data acquisition

All mass spectrometry methods for acquisition and data handling are detailed in Supplemental experimental procedures S10, Scheme S1, and Table S4.

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.xcrp.2021.100685.

ACKNOWLEDGMENTS

We gratefully acknowledge financial support from the John Templeton Foundation (grant 60025), the Engineering and Physical Sciences Research Council (EPSRC) (grants EP/L023652/1, EP/R01308X/1, EP/J015156/1, and EP/P00153X/1), the Breakthrough Prize Foundation and NASA (Agnostic Biosignatures award 80NSSC18K1140), MINECO (project CTQ2017-87392-P), and the European Research Council (ERC) (project 670467 SMART-POM). We thank Dr. Jesko Koehnke of the University of Glasgow for a sample of a thioholgamide.

AUTHOR CONTRIBUTIONS

L.C. conceived the initial theory and hypothesis, designed the project, and coordinated the efforts of the research team. D.D. developed the concept and along with E. Clarke built the algorithm and collected data. D.D. synthesized and analyzed acylated peptide and polyimine mixtures. E. Clarke synthesized and analyzed peptide standards and depsipeptide mixtures. D.D., E. Clarke, and G.K. wrote the software. E. Clarke, D.D., and E. Carrick carried out mass spectrometry data collection and processing. C.M. helped coordinate the team and gave advice on the code, concept, and manuscript. D.D., E. Clarke, and L.C. together wrote the manuscript with input from all authors.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: August 19, 2021
Revised: September 30, 2021
Accepted: November 18, 2021
Published: December 8, 2021

REFERENCES

1. Cottrell, J.S. (2011). Protein identification using MS/MS data. J. Proteomics *74*, 1842–1851.

2. Amiri-Dashatan, N., Koushki, M., Abbaszadeh, H.-A., Rostami-Nejad, M., and Rezaei-Tavirani, M. (2018). Proteomics applications in health: biomarker and drug discovery and food industry. Iran. J. Pharm. Res. *17*, 1523–1536.

3. Li, G., Cai, C., Ren, T., and Tang, X. (2014). Development and application of a UPLC-MS/MS method for the pharmacokinetic study of 10-hydroxy camptothecin and hydroxyethyl starch conjugate in rats. J. Pharm. Biomed. Anal. *88*, 345–353.

4. Mechref, Y. (2012). Use of CID/ETD mass spectrometry to analyze glycopeptides. Curr. Protoc. Protein Sci. *Chapter 12*. Unit 12.11.11–12.11.11.

5. Chu, I.K., Siu, C.-K., Lau, J.K.-C., Tang, W.K., Mu, X., Lai, C.K., Guo, X., Wang, X., Li, N., Xia, Y., et al. (2015). Proposed nomenclature for peptide ion fragmentation. Int. J. Mass Spectrom. *390*, 24–27.

6. Altuntaş, E., and Schubert, U.S. (2014). "Polymeromics": Mass spectrometry based strategies in polymer science toward complete sequencing approaches: a review. Anal. Chim. Acta *808*, 56–69.

7. Surman, A.J., Rodriguez-Garcia, M., Abul-Haija, Y.M., Cooper, G.J.T., Gromski, P.S., Turk-MacLeod, R., Mullin, M., Mathis, C., Walker, S.I., and Cronin, L. (2019). Environmental control programs the emergence of distinct functional ensembles from unconstrained chemical reactions. Proc. Natl. Acad. Sci. U S A *116*, 5387–5392.

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

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

10. Tran, N.H., Zhang, X., Xin, L., Shan, B., and Li, M. (2017). De novo peptide sequencing by deep learning. Proc. Natl. Acad. Sci. U S A *114*, 8247–8252.

11. Wein, S., Andrews, B., Sachsenberg, T., Santos-Rosa, H., Kohlbacher, O., Kouzarides, T., Garcia, B.A., and Weisser, H. (2020). A computational platform for high-throughput analysis of RNA sequences and modifications by mass spectrometry. Nat. Commun. *11*, 926.

12. Forsythe, J.G., Yu, S.-S., Mamajanov, I., Grover, M.A., Krishnamurthy, R., Fernández, F.M., and Hud, N.V. (2015). 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.

13. Burel, A., Carapito, C., Lutz, J.-F., and Charles, L. (2017). MS-DECODER: milliseconds sequencing of coded polymers. Macromolecules *50*, 8290–8296.

14. Celasun, S., Remmler, D., Schwaar, T., Weller, M.G., Du Prez, F., and Börner, H.G. (2019). Digging into the sequential space of thiolactone precision polymers: a combinatorial strategy to identify functional domains. Angew. Chem. Int. Ed. Engl. *58*, 1960–1964.

15. Martens, S., Landuyt, A., Espeel, P., Devreese, B., Dawyndt, P., and Du Prez, F. (2018). Multifunctional sequence-defined macromolecules for chemical data storage. Nat. Commun. *9*, 4451.

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

17. Biemann, K. (1990). Appendix 5. Nomenclature for peptide fragment ions (positive ions). Methods Enzymol. *193*, 886–887.

18. Biemann, K. (1988). Contributions of mass spectrometry to peptide and protein structure. Biomed. Environ. Mass Spectrom. *16*, 99–111.

19. McLuckey, S.A., Van Berkel, G.J., and Glish, G.L. (1992). Tandem mass spectrometry of small, multiply charged oligonucleotides. J. Am. Soc. Mass Spectrom. *3*, 60–70.

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

21. Xia, Q., Lee, M.V., Rose, C.M., Marsh, A.J., Hubler, S.L., Wenger, C.D., and Coon, J.J. (2011). Characterization and diagnostic value of amino acid side chain neutral losses following electron-transfer dissociation. J. Am. Soc. Mass Spectrom. *22*, 255–264.

22. Williams, S.M., and Brodbelt, J.S. (2004). MS(n) characterization of protonated cyclic peptides and metal complexes. J. Am. Soc. Mass Spectrom. *15*, 1039–1054.

23. Yu, S.-S., Solano, M.D., Blanchard, M.K., Soper-Hopper, M.T., Krishnamurthy, R., Fernández, F.M., Hud, N.V., Schork, F.J., and Grover, M.A. (2017). Elongation of model prebiotic proto-peptides by continuous monomer feeding. Macromolecules *50*, 9286–9294.

24. Curtis, J.M., Bradley, C.D., Derrick, P.J., and Sheil, M.M. (1992). Four-sector tandem mass spectrometry: a comparison of the molecular and quasi-molecular ions of the cyclic depsipeptide valinomycin formed using electron impact, chemical ionization, fast atom bombardment, field desorption and electrospray Ionizatfion. Org. Mass Spectrom. *27*, 502–507.

25. Kjaerulff, L., Sikandar, A., Zaburannyi, N., Adam, S., Herrmann, J., Koehnke, J., and Müller, R. (2017). Thioholgamides: thioamide-containing cytotoxic RiPP natural products. ACS Chem. Biol. *12*, 2837–2841.

26. Sikandar, A., Lopatniuk, M., Luzhetskyy, A., and Koehnke, J. (2020). Non-Heme Monooxygenase ThoJ catalyzes thioholgamide β-hydroxylation. ACS Chem. Biol. *15*, 2815–2819.

27. Sample, P.J., Gaston, K.W., Alfonzo, J.D., and Limbach, P.A. (2015). RoboOligo: software for mass spectrometry data to support manual and de novo sequencing of post-transcriptionally modified ribonucleic acids. Nucleic Acids Res. *43*, e64.

28. De Bruycker, K., Krappitz, T., and Barner-Kowollik, C. (2018). High performance quantification of complex high resolution polymer mass spectra. ACS Macro Lett. *7*, 1443–1447.