Check for updates

Sequence controlled secondary structure is important for the site-selectivity of lanthipeptide cyclization†

Xuenan Mi, [ID] [a] Emily K. Desormeaux, [ID] [b] Tung T. Le, [ID] [b] Wilfred A. van der Donk [ID] *[b] and Diwakar Shukla [ID] *[acd]

Lanthipeptides are ribosomally synthesized and post-translationally modified peptides that are generated from precursor peptides through a dehydration and cyclization process. ProcM, a class II lanthipeptide synthetase, demonstrates high substrate tolerance. It is enigmatic that a single enzyme can catalyze the cyclization process of many substrates with high fidelity. Previous studies suggested that the site-selectivity of lanthionine formation is determined by substrate sequence rather than by the enzyme. However, exactly how substrate sequence contributes to site-selective lanthipeptide biosynthesis is not clear. In this study, we performed molecular dynamic simulations for ProcA3.3 variants to explore how the predicted solution structure of the substrate without enzyme correlates to the final product formation. Our simulation results support a model in which the secondary structure of the core peptide is important for the final product's ring pattern for the substrates investigated. We also demonstrate that the dehydration step in the biosynthesis pathway does not influence the site-selectivity of ring formation. In addition, we performed simulation for ProcA1.1 and 2.8, which are well-suited candidates to investigate the connection between order of ring formation and solution structure. Simulation results indicate that in both cases, C-terminal ring formation is more likely which was supported by experimental results. Our findings indicate that the substrate sequence and its solution structure can be used to predict the site-selectivity and order of ring formation, and that secondary structure is a crucial factor influencing the site-selectivity. Taken together, these findings will facilitate our understanding of the lanthipeptide biosynthetic mechanism and accelerate bioengineering efforts for lanthipeptide-derived products.

1 Introduction

Natural products are important sources of new drugs, with more than 60% of all new drugs derived from natural products or their derivatives.[1] Ribosomally synthesized and post-translationally modified peptides (RiPPs) are a fast-growing natural product family because recent developments in genome mining algorithms have facilitated their discovery.[2–9] Most RiPPs follow a similar biosynthetic logic: the precursor peptide, encoded by a structural gene, is modified by enzymes to generate the mature natural product. Most precursor peptides are composed of a highly conserved leader peptide at the N-terminus, which is important for recognition by modification enzymes, as well as a highly variable core sequence at the C-terminus that is transformed into the mature product.[10]

Lanthipeptides are the largest group of RiPPs based on currently sequenced genomes[11,12] and have a variety of bioactivities, including antimicrobial,[13] antiviral,[14] antifungal[15] and antiallodynic activities.[16] Lanthipeptide synthetases catalyze the dehydration of Ser/Thr residues to dehydroalanine (Dha)/(Z)-dehydrobutyrine (Dhb) respectively, followed by an intra-molecular Michael-type addition of Cys thiols to the Dha/Dhb to form the class-defining Lan or MeLan linkages (Fig. 1A and B).[17] After modification, the leader peptide is removed by proteases and the mature product is exported. This paper will discuss class II lanthipeptide synthetases which are single bifunctional enzymes capable of catalyzing both the dehydration and the cyclization reactions. Enzymatic modifications catalyzed by many lanthipeptide synthetases demonstrate relaxed substrate specificity, often tolerating significant changes made to the core peptide sequences.[18–21] This feature has been widely utilized for bioengineering of novel lanthipeptides.[22–29] A remarkable

[a]Center for Biophysics and Quantitative Biology, University of Illinois at Urbana-Champaign, Urbana, IL, 61801, USA

[b]Department of Chemistry and Howard Hughes Medical Institute, University of Illinois at Urbana-Champaign, Urbana, IL, 61801, USA. E-mail: vddonk@illinois.edu

[c]Department of Chemical and Biomolecular Engineering, University of Illinois at Urbana-Champaign, Urbana, IL, 61801, USA. E-mail: diwakar@illinois.edu

[d]Department of Bioengineering, University of Illinois at Urbana-Champaign, Urbana, IL, 61801, USA

† Electronic supplementary information (ESI) available: Fig. S1–S18 and Table S1. See DOI: https://doi.org/10.1039/d2sc06546k

example of substrate tolerance is the class II lanthipeptide synthetase ProcM, which was discovered in the marine pico-cyanobacterium *Prochlorococcus* MIT9313.[30] Unlike most other lanthipeptide biosynthetic pathways, the ProcM biosynthetic gene cluster does not encode just a single precursor peptide with one modifying enzyme.[30] Along with ProcM, genes encoding 30 distinct putative precursor peptides were identified in the genome, termed ProcA; the mature products were called prochlorosins (Pcn). A related enzyme SyncM is encoded in *Synechococcus* MITS9509 together with up to 80 putative substrate peptides,[31] and SyncM has also been demonstrated to have very high substrate tolerance.[32] ProcM is a bifunctional enzyme with an N-terminal dehydratase domain and a C-terminal cyclase domain that acts on its 30 distinct substrate sequences, and forms lanthipeptides with highly diverse sequences and ring patterns.[30,33,34] Previous studies have provided a possible explanation for the high substrate tolerance suggesting that the core peptide sequence rather than the enzyme may determine the final ring pattern.[35–37] In this proposed model, the substrate has a propensity towards a specific ring pattern based on its conformational free energy landscape. ProcM would accelerate the cyclization by increasing the nucleophilicity of the Cys thiolate coordinated to the zinc ion in the active site and covalently locking the peptide into such favorable conformations.[36]

In a recent study, Le *et al.* investigated how substrate sequence controls site-selectivity of lanthionine formation by ProcM.[35] The findings with a library of ProcA3.3 precursor peptide variants supported a model in which substrate sequence determines the site-selectivity of lanthionine formation. However, the experimental data could not predict how the core sequence controls the ring pattern of the final product through modulating its conformational free energy landscape. Computational studies using molecular dynamics (MD)

simulations have been an effective approach to investigate conformational equilibria of peptides and proteins.[38–42] In this study we used the core peptide of ProcA3.3 as a model and performed atomic-scale molecular dynamic simulations of linear peptides in solution to explore the factors that may determine the ring pattern. Our results revealed that the secondary structure of the core sequence is an important factor for controlling the final product structure. Furthermore, we investigated whether the order of cyclization is determined by substrate sequence alone or is also influenced by ProcM. Pcn1.1 and Pcn2.8, which contain two non-overlapping rings, are well-suited candidates to study the order of cyclization. Our computational analysis indicated that the C-terminal rings of ProcA1.1 and ProcA2.8 have higher probabilities to form based on energetic stabilities inferred from the simulation-based conformational energy landscape. To validate our results, we performed Liquid Chromatography-Mass Spectrometry (LC-MS) analysis to confirm that ProcA1.1 is cyclized in a C-to-N-terminal fashion. Our work provides computational evidence to explain how substrate sequence determines the prochlorosin site-selectivity and order of cyclization and demonstrates how secondary structure of the core sequence is a crucial factor to control ring patterns of final lanthipeptide products.

2 Methods

2.1 Molecular dynamics (MD) simulation

Atomistic MD simulations were performed starting from the linear structure of the peptides. All peptides were constructed using PyMol.[43] The peptide was solvated in a TIP3P water box, and the system was neutralized by Na^+ ions and Cl^- ions using Packmol.[44] All MD simulations were performed using the Amber18 software package employing the Amber ff14SB force
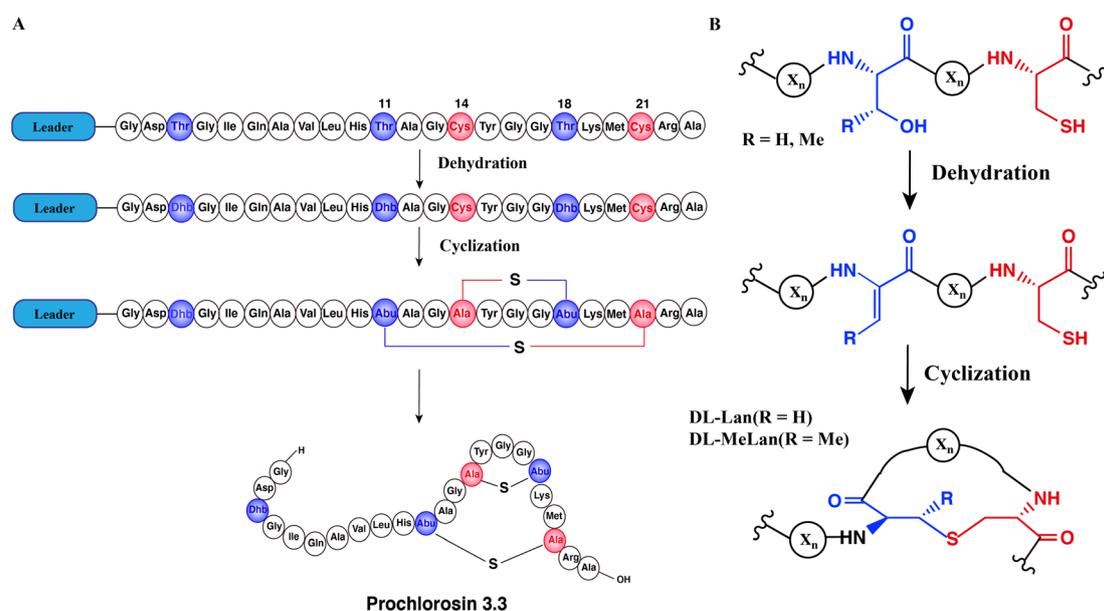

Fig. 1 Schematic biosynthetic pathway of prochlorosin 3.3. (A) Biosynthetic route to Pcn3.3 and cartoon representation of the sequence and structure. (B) Post-translational modifications carried out by ProcM during Pcn3.3 biosynthesis. X_n represents the peptide chain.
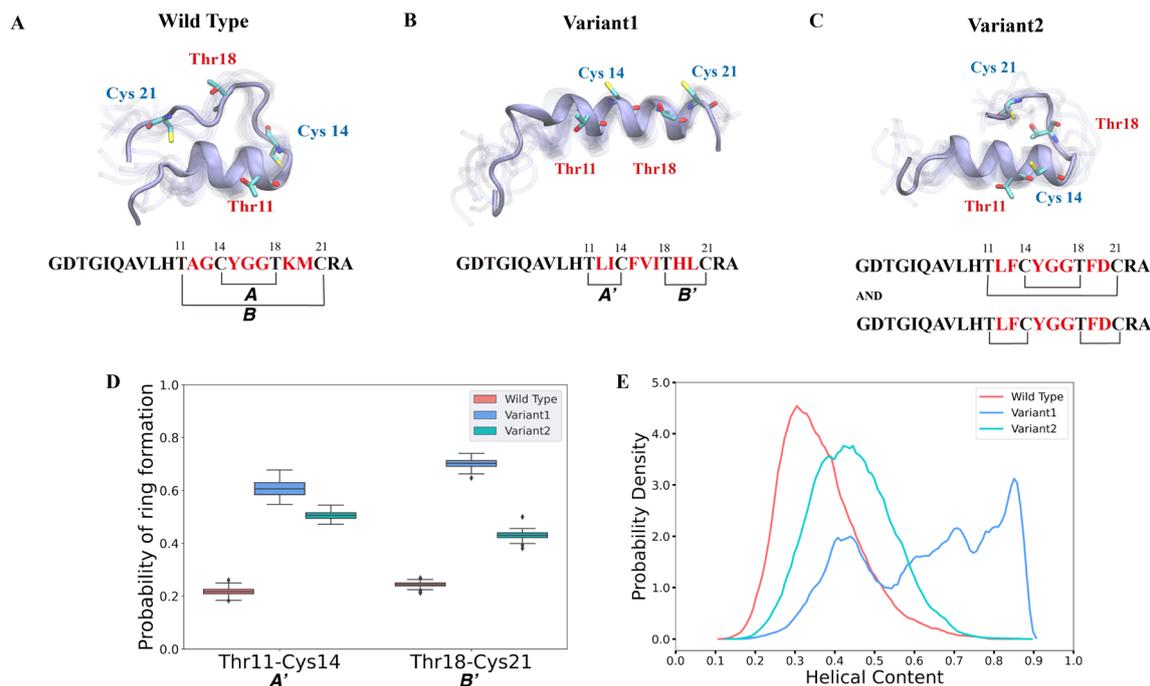
Fig. 2 Different secondary structures lead to distinct ring patterns of ProcA3.3 variants. (A)–(C) Cartoon representation of the most populated state in the free energy landscape of ProcA3.3 variants (for the entire free energy landscape of all states, see Fig. S3A–C†). In each state, 10 snapshots were randomly selected and aligned together. One snapshot is shown in ice blue, while the others are shown in transparent ice blue. (D) Probability of formation of ring A' (Thr11/Cys14) and B' (Thr18/Cys21) for ProcA3.3 variants. Error bar was calculated by 200 bootstrap samples with 80% of total trajectories. (E) Distribution of helical content of ProcA3.3 variants.

field.[45] Parameters for dehydro amino acids were generously provided by Gonzalo Jiménez-Osés.[35] They were generated originally using the AMBER gaff2 force field and with partial charges set to fit the electrostatic potential generated with HF/6-31G(d) using the RESP method.[46] The charges were calculated according to the Merz–Singh–Kollman scheme using Gaussian 16.[47] Each MD system was first minimized for 50 000 cycles using steepest descent for the first 5000 cycles and conjugate

gradient for the remaining 45 000 cycles. The systems were heated from 0 K to 300 K under NVT ensemble. The heating step was conducted for 3 ns using Langevin thermostat with a collision frequency of 2 ps^{-1}.[48] The systems were equilibrated in NPT ensemble for 2 ns with a pressure of 1 bar using Monte Carlo barostat.[49] The systems were further equilibrated in NPT ensemble (300 K and 1 bar) for 50 ns and then underwent production runs.[50] The SHAKE algorithm was used to constrain
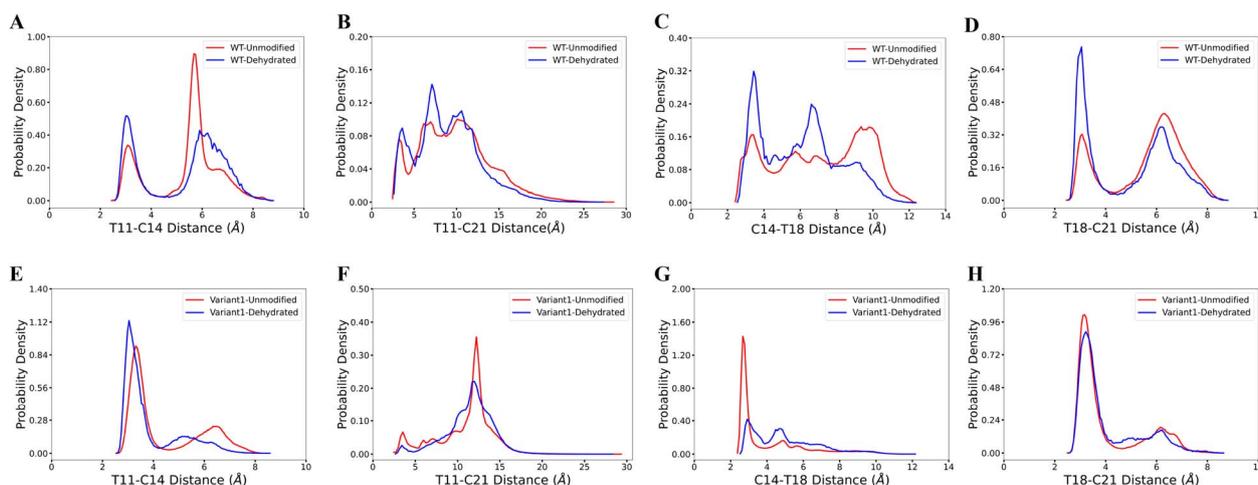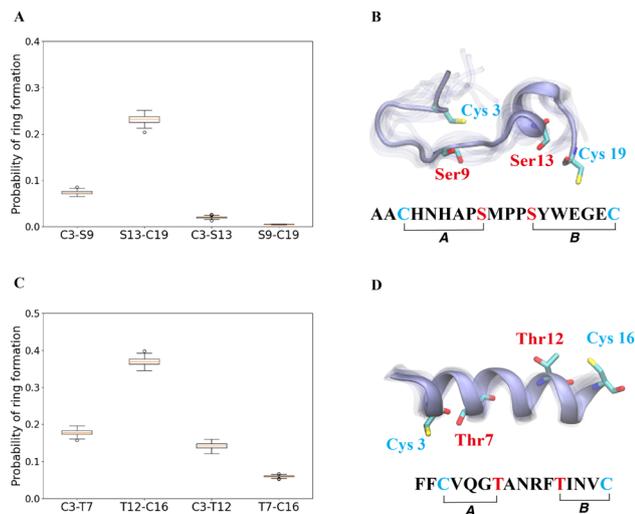

Fig. 3 Comparison of unmodified and dehydrated ProcA3.3 variants. (A)–(D) Distance distribution of four important features for unmodified and fully dehydrated WT ProcA3.3. (E)–(H) Distance distribution of four important features for unmodified and fully dehydrated ProcA3.3 Variant1.

Fig. 4 (A) Probability of formation of all possible rings for ProcA2.8, with ring formation defined as the distance between the β-carbon of Thr and the sulfur of Cys being within 7.5 Å. (B) Cartoon representation of the state leading to rings A and B in the free energy landscape for ProcA2.8 (for the entire free energy landscape of all states, see Fig. S8A†). In each state, 10 snapshots were randomly selected and aligned together. One snapshot is shown in ice blue, while the others are shown in transparent ice blue. (C) Probability of formation of all possible rings for ProcA1.1. (D) Cartoon representation of the state in which ring A and B are formed in the free energy landscape for ProcA1.1 (for the entire free energy landscape of all states, see Fig. S8C†). Error bar was calculated by 200 bootstrap samples with 80% of total trajectories.

hydrogen-containing bonds.[51] All systems were subject to hydrogen mass repartitioning (HMR),[52] which redistributes mass between hydrogen atoms and covalently bonded atoms of the peptide to allow the time step of the simulation to be increased to 4 fs.
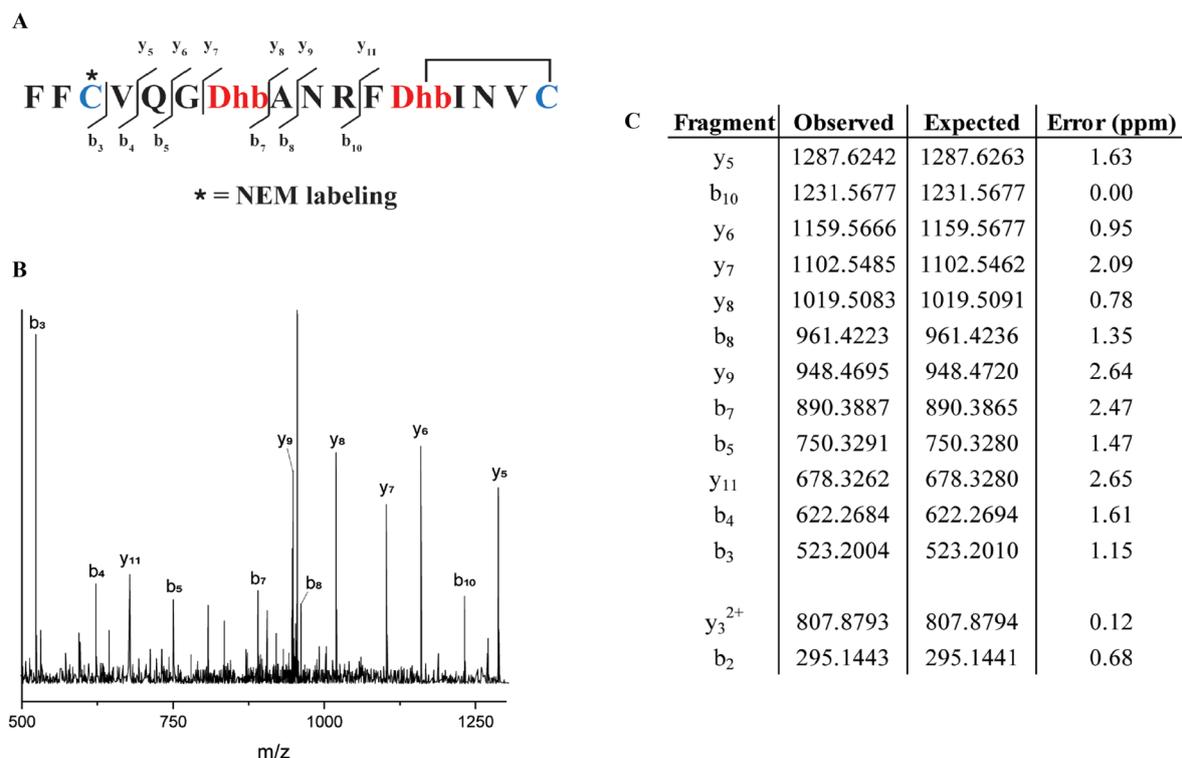
2.2 Adaptive sampling

All simulation data were obtained by applying the adaptive sampling method, that was used to efficiently sample the conformational space of the peptide.[53] The adaptive sampling approach has been applied in sampling protein-ion binding, protein-ligand binding and protein folding processes.[54–57] In this study, the least count based adaptive sampling was used to find the new conformational states quickly.[53] Adaptive sampling was performed as follows:

(1) Run a series of short MD simulations from a collection of starting structures.

(2) Cluster all collected simulation data using a K-means algorithm.[58] The distances of all pairs of residues separated by two or more residues were used as features to generate 100 clusters.

(3) Randomly pick one state from each of 25 clusters with the least population as seeds to start a new simulation.

(4) Repeat steps 1–3 until the sampling reaches convergence.

2.3 Markov state model (MSM) construction

In the adaptive sampling, we generated many short trajectories to capture the dynamic process of the system. MSM was built to

Fragment	Observed	Expected	Error (ppm)
y_5	1287.6242	1287.6263	1.63
b_{10}	1231.5677	1231.5677	0.00
y_6	1159.5666	1159.5677	0.95
y_7	1102.5485	1102.5462	2.09
y_8	1019.5083	1019.5091	0.78
b_8	961.4223	961.4236	1.35
y_9	948.4695	948.4720	2.64
b_7	890.3887	890.3865	2.47
b_5	750.3291	750.3280	1.47
y_{11}	678.3262	678.3280	2.65
b_4	622.2684	622.2694	1.61
b_3	523.2004	523.2010	1.15
y_3^{2+}	807.8793	807.8794	0.12
b_2	295.1443	295.1441	0.68

Fig. 5 (A) Sequence of ProcA1.1 intermediate after leader peptide cleavage with observed fragmentation indicated. (B) Tandem-MS fragmentation of ProcA1.1. (C) Table of observed and expected fragment ion masses and the error calculated (ppm).

connect these independent trajectories thermodynamically and kinetically, and remove bias introduced by the least count based adaptive sampling.[55,59,60] The distances between residue pairs separated by two or more residues were used to featurize the simulation data. The time-lagged independent component analysis (tICA) was used to reduce dimensions of featurized data. Each time-lagged independent component (tIC) is a linear combination of features. Using tICs, we could determine the discrete states of the system, which have different slowest timescales. The discrete states were further discretized based on the K-means algorithm.[58] Then MSM was used to model the entire dynamic system through the corresponding transition probability matrix (T) between discretized states. Each element (T_{ij}) in the matrix represents the probability of transitioning from state i to state j at lag time τ, which is long enough to validate Markovian behavior.[60] The lag time τ was estimated from the implied time scale, the minimum τ at which the implied time scale converged was selected as the MSM lag time (Fig. S10–S16A†). The optimized hyperparameters (the number of clusters and the number of tICs) of the MSM were selected by maximizing the VAMP-2 score (Fig. S10–S16B†), calculated by the sum of squared eigenvalues from the transition probability matrix,[61,62] which maximizes the kinetic variance contained in the features.[63,64] The first six eigenvalue were used to calculate the VAMP-2 score and 10-fold cross-validation was calculated to obtain the average score using pyEMMA.[65]

2.4 Trajectory analysis

CPPTRAJ module in AMBER 18 (ref. 45) and the python package MDTraj 1.9.0 (ref. 66) were used to analyze trajectory data, and VMD 1.9.3 (ref. 67) and PyMol[43] were used to visualize MD snapshots. The python package Matplotlib[68] was used to generate the 2-D plot of the free energy landscape.

2.5 Ring formation probability calculation

MD calculations do not result in chemical transformation due to the fixed atom connectivity of the molecules. Conformations that potentially lead to ring formation were defined as the distance between the β-carbon of threonine and the sulfur of cysteine being within 7.5 Å. This distance is usually used to define residue–residue interactions that lead to protein folding.[69,70] Use of alternative distance did not result in significant differences in probabilities (Fig. S1†). The distance between atoms was calculated using MDtraj 1.9.0.[66] The equilibrium ring formation probability was calculated as the product of raw ring formation probability within each MSM state multiplied by the equilibrium probability of the MSM state, as follow:

$$P_{\text{ring-formation,eq}} = \sum_{i}^{N_{\text{states}}} P_{\text{ring-formation,raw},i} \times \pi_i \qquad (1)$$

2.6 Error analysis

Errors on thermodynamics calculations were generated by bootstrapping.[71] For each bootstrap sample, 80% of the total number of trajectories were randomly selected. We kept the original state index and built MSM for each sample. We generated N bootstrap samples and used the standard deviation of N samples as error bar.

2.7 Purification of ProcA peptides

The genes for His$_6$-tagged ProcA peptides were cloned into a pET15b vector and overexpressed as previously described.[72] Cell paste was resuspended in 20 mL of LanA B1 buffer (6.0 M guanidine HCl, 20 mM NaH$_2$PO$_4$, 0.5 mM imidazole, 1 mM TCEP, pH 7.5) for 1 L of cell culture and lysed via sonication at 60% amplitude for 5 min with a 2.0 s on and 6.0 s off pulse. The lysate was clarified via centrifugation at 50 000×g for 1 h and the supernatant was filtered using 0.45 μm syringe filters. The filtered lysate was loaded onto a gravity flow Ni-NTA column with 1 mL of resin preequilibrated with 6 column volumes (CV) of LanA B1 buffer. The column was washed with 10 CV of LanA B1 buffer, 5 CV of LanA B2 buffer (6.0 M guanidine HCl, 20 mM NaH$_2$PO$_4$, 30 mM imidazole, 1 mM TCEP, pH 7.5), and 5 CV of LanA wash buffer (20 mM NaH$_2$PO$_4$, 30 mM imidazole, 300 mM NaCl, 1 mM TCEP, pH 7.5). The peptide was then eluted using 10 CV of LanA elution buffer (20 mM NaH$_2$PO$_4$, 1 M imidazole, 100 mM NaCl, 1 mM TCEP, pH 7.5). The buffer of the purified peptide was then exchanged into a thrombin digestion buffer (50 mM HEPES, 100 mM NaCl, 1 mM TCEP, pH 8.0) using a 3 kDa Amicon centrifugation filter and the sample was concentrated to approximately 5 mL. Bovine thrombin high purify grade (MP Biomedicals) was added to the sample (100 units) and the sample was left overnight at 4 °C for removal of the His$_6$-tag. Thrombin-digested peptides were purified via reversed phase HPLC using a Phenomenex Luna C5 semiprep column (250 × 10 mm, 10 μm, 100 Å) at a flow rate of 8 mL min^{-1} and with the following gradient over 32 min: 2% B for 10 min, 2–30% B over 2 min, 30–60% B over 15 min, 60–100% B over 2 min, and hold at 100% B for 3 min (A: 0.1% TFA in H$_2$O, B: 100% MeCN, 0.1% TFA). HPLC-purified peptide was collected and lyophilized, and the peptides were stored as a powder at −20 °C until use.

2.8 Purification of ProcM

ProcM was overexpressed as reported earlier utilizing a pRSFDuet vector.[72,73] All purification steps were carried out in a cold room (4 °C) or on an ice bath. Cell paste was resuspended in 20 mL of LanM start buffer (20 mM Tris, 500 mM NaCl, 10% glycerol, pH 7.6) per L of expression and the cell mixture was allowed to nutate with protease inhibitor (Pierce™), lysozyme (50 mg L^{-1} culture), and Benzonase (12 μL L^{-1} cell culture) for 1 h. Cell lysis was performed via sonication at 35% amplitude for 15 min with a 4.0 s on, 9.9 s off pulse. The lysate was clarified via centrifugation at 50 000×g for 1 h and the supernatant was filtered using 0.45 μm syringe filters. The filtered lysate was loaded onto a Ni-Hi-Trap column equilibrated with 6 CV of ProcM start buffer. The loaded column was manually washed with 6 CV of ProcM wash buffer (20 mM Tris, 500 mM NaCl, 30 mM imidazole, 10% glycerol, pH 7.6) before being attached to an Akta fast protein liquid chromatography (FPLC) system to

complete the elution using ProcM wash and elution buffers (20 mM Tris, 500 mM NaCl, 1 M imidazole, 10% glycerol, pH 7.6) and monitoring elution at 280 nm with the following gradient: 0–2% over 2 CV, 2–20% over 10 CV, 20–100% over 0.5 CV. The protein elution fractions were analyzed via gel electrophoresis and fractions containing ProcM were concentrated using 50 kDa Amicon centrifugal filtration. The ProcM was then further purified/desalted using a FPLC gel filtration column (Superdex 200, 1.5 × 60 cm, GE healthcare) using a 1 mL min^{-1} flow rate and an isocratic elution using ProcM storage buffer (50 mM HEPES, 500 mM KCl, 5% glycerol, pH 7.6). ProcM eluted in three peaks (aggregate, then oligomer, followed by monomer). The monomeric peaks were collected, concentrated via 50 kDa Amicon centrifugal filtration, aliquoted into single use portions, flash frozen in liquid nitrogen, and stored at −80 °C until use.

2.9 ProcM reactions

Prior to initiation of reactions, 4 μM ProcM and 80 μM ProcA peptide were preincubated separately at 25 °C for 1 h in 750 μL of reaction mixture (5 mM ATP, 0.17 mM ADP, 5 mM MgCl$_2$, 100 mM HEPES, 0.1 mM TCEP, pH 7.5). The two samples were mixed thoroughly to initiate a reaction with a final concentration of 2 μM ProcM and 40 μM ProcA peptide. An 80 μL aliquot was removed at desired time points and quenched into 900 μL of quench buffer (111 mM citrate, 1.11 mM EDTA, pH 3.0). Each reaction was initially quenched at 15, 30, 45, and 60 min. Reaction times were adjusted as needed to obtain monocyclized intermediate (30 min for ProcA1.1). After quenching, 100 μL of 100 mM TCEP was added to each aliquot and samples were incubated at 25 °C for 10 min. The pH of the aliquots was adjusted to approximately 6.3 via the addition of 40–45 μL of 5 M NaOH. The free thiols were then alkylated by the addition of 11 μL of 1 M N-ethylmaleimide (NEM) in EtOH and the reaction was incubated at 37 °C for 10 min before the reaction was quenched by the addition of 11 μL of formic acid.[74] The buffer of the time points was then exchanged to water using 3 kDa Amicon centrifugal filters. Samples were then lyophilized and stored dry until used for analysis.

2.10 Liquid chromatography-mass spectrometry

Samples were resuspended in 50 μL of water, and the leader peptide was removed via the addition of 50 μL of LahT$_{150}$ solution[75] (2.0 mg mL^{-1}) and left overnight. The protease LahT$_{150}$ was removed from samples via addition of 100 μL of MeOH and 5 μL of formic acid. The sample was centrifuged to pellet any precipitated protein in the sample and 10 μL aliquots were injected onto an Agilent Infinity1260 Liquid Chromatography instrument coupled to an Agilent 6454 QTOF mass spectrometer using a Kinetex® 2.6 μm C8 100 Å, LC column (150 × 2.1 mm). The samples were run with a flow rate of 0.4 mL min^{-1} at 45 °C. The peptides were eluted using the following gradient: 0–5% B over 3 min, 5–50% B over 10 min, and 50–95% B over 1 min for a total run time of 14 min (A: 0.1% formic acid in water, B: 100% MeCN, 0.1% formic acid). QTOF data was collected with the following settings: polarity =

positive, data storage = centroid, acquisition range (m/z) = 100–1700, gas temperature = 325 °C, gas flow = 13 L min^{-1}, nebulizer = 35 psig, sheath gas temp = 275 °C, sheath gas flow = 12 L min^{-1}, Vcap = 4000, nozzle voltage = 500 V, fragmentor = 175 V, skimmer = 65 V, octupole RF peak = 750, scan rate = 5 spectra per s, acquisition mode = target MS, and collision energy = 20/30 V. Fragmentation at fixed collision energy (20/30 V) was performed on the +2/+3 charge species of the target precursors. The data were processed with Qualitative Analysis 10.0 (Agilent).

3 Results and discussion

3.1 Secondary structure may determine ring patterns of ProcA3.3 variants

Previous studies suggested that it is not the enzyme ProcM but rather its substrate sequences that determine the site-selectivity of lanthionine formation.[35,36] Recent work applied trapped ion mobility spectrometry-tandem mass spectrometry (TIMS-MS/MS) to discern that wild type (WT) ProcA3.3 contains an overlapping ring pattern, whereas ProcA3.3 Variant1 displays a non-overlapping ring pattern and ProcA3.3 Variant2 displays a mixture of both ring patterns (the ratio of overlapping and non-overlapping is 40 : 60) (sequences are shown in Fig. 2). However, experimental results could not predict how the core sequence controls the ring pattern of the final product. We hypothesized that the solution structure of the core peptide may be a key factor to control ring pattern. ProcM processing of ProcA3.3 has been shown previously to be an irreversible process under assay conditions, and thus the first ring that is formed determines the ultimate ring pattern.[76]

In our study, we first ran 40 μs MD simulations for the core peptide of WT ProcA3.3 and the two representative variants to provide the general energy landscape and identify any secondary structures. To evaluate the important residue–residue contacts, we performed time-lagged independent components analysis (tICA) on the simulation data. tICA is a method of dimension reduction to identify slow processes by finding coordinates of maximal autocorrelation at a given lag time.[63] Each time-lagged independent component (tIC) is a linear combination of different features. Identification of slow conformational processes (~0.8–2 μs) is important as it indicates that all dynamics processes have been captured. In addition, slow structural movements have been shown to often be linked to functionally important processes in protein folding[77] and conformational changes.[78] The first two tICs (tIC1 and tIC2) are the slowest components of the system during the MD simulation. We evaluated all pairwise residue–residue interactions in the peptide, and as shown in Fig. S2,† of the pairwise interactions that lead to ring formation the distance between Thr11 and Cys14 has the highest correlation with tIC1, especially for Variant1. Such a high correlation suggests that the contact between Thr11 and Cys14 is one of the slowest dynamic processes during the conformational dynamics. We projected the MD simulation data along the Thr11-Cys14 distance and α – helix content of ProcA3.3 to compare ProcA3.3 WT and variants.

Fig. 2A shows the most populated conformation from the MSM weighted free energy landscape (Fig. S3A†) for WT ProcA3.3. We observed that WT ProcA3.3 forms an α-helix spanning residues 4 to 12, with Cys14, Thr18 and Cys21 all located within a flexible loop region. This structure is conducive for formation of rings A (Cys14/Thr18) and B (Thr11/Cys21). Conversely, the MSM weighted free energy landscape (Fig. S3B†) for ProcA3.3 Variant1 shows that the α – helix spans residues 8 to 22 (Fig. 2B). Thr11, Cys14, Thr18 and Cys21 are all located within the helix, which renders the Thr11-Cys14 residue pair closer and facilitates formation of rings A′ (Thr11/Cys14) and B′ (Thr18/Cys21). For ProcA3.3 Variant2, Thr11 and Cys14 are both within the α – helix, but Thr18 and Cys21 are in the flexible loop region (Fig. 2C). The α – helix structure enables ring A′ and B′ formation; simultaneously, there is a chance to form rings A and B. Therefore, ProcA3.3 Variant2 can form products with both overlapping and non-overlapping ring patterns.

We next compared probabilities of ring formation quantitatively. The ring formation was defined as conformers in which the distance between the β-carbon of Thr and the sulfur of Cys was within 7.5 Å (see Methods). When comparing the probabilities of forming rings A′ and B′ in ProcA3.3 WT and variants, the two variants clearly have a much higher probability than the WT (Fig. 2D). When comparing the probabilities of forming ring A, Variant1 has a higher probability of forming this ring than WT (Fig. S4A†), because the α – helix also facilitates Cys14 and Thr18 to be close. We do note that WT has a relatively higher probability of forming ring B (Fig. S4B†). The higher probability to form ring A′ and B′ in the two variants would enhance their contribution to the conformational selection mechanism toward determining the ring pattern (Fig. 2D). Therefore, these data support a model in which the substrate sequence plays a determinate role in the site-selectivity of ProcA3.3 variants. But the simulations do not explain the preference of ProcM to form the A ring over the A′ and B′ rings for WT ProcA3.3, because our simulations predict preferential formation of the A′ and B′ rings for all three peptides (Fig. S6†). We note, however, that the difference between the probability of forming ring A′/B' and ring A in the WT is much smaller than the difference in Variant1 and Variant2. This observation indicates the simulations correctly predict the trend towards more non-overlapping ring formation for the variants than the WT ProcA. One limitation of our study is that we are not investigating the enzyme-peptide binding which could shed light on the reason why the enzyme prefers to form the overlapping ring pattern in the WT ProcA3.3.

Here, the distance between the β-carbon of Thr and the sulfur of Cys was used as a proxy for the likelihood of ring formation, because the ring is formed between these two atoms. We also calculated the predicted probabilities of ring formation if any of their heavy atoms is within a cutoff of 4.5 Å. Using this definition, we observed the same trend of probabilities of different rings formation among variants of the ProcA3.3 core peptide (Fig. S4C, S4D and S5†). In addition, we quantitatively compared the α-helix content in WT and variant ProcA3.3 core peptide. Fig. 2E shows that Variant1 has a significantly higher α-helix content compared to WT and Variant2.

3.2 Dehydration does not influence ring pattern predictions

Our computational analysis supports the model that substrate sequence determines the cyclization process of ProcA and suggests this occurs at least in part through controlling secondary structure. As mentioned, ProcM processing of ProcA3.3 is an irreversible process under assay conditions, and therefore the first ring that is formed determines the ultimate ring pattern.[76] In the biosynthetic pathway of lanthipeptides, before the cyclization step, the first step is dehydration of Ser/Thr residues in the core peptide to generate dehydroalanine (Dha from Ser) or (Z)-dehydrobutyrine (Dhb from Thr).[17] Therefore, we next ran MD simulations starting from both unmodified peptides and dehydrated peptides to investigate whether the dehydration process affects the conformational preference of the peptides. Four possible initial rings can be formed from the core peptide of ProcA3.3, connecting Thr11/Cys14, Thr11/Cys21, Cys14/Thr18 and Thr18/Cys21 residue pairs. Fig. 3 shows the distance distribution of these four features for unmodified and fully dehydrated ProcA3.3 WT and Variant1. Unmodified and dehydrated peptide models demonstrate consistent distance distributions of all four features, indicating that the dehydration process does not greatly influence the predicted solution structure of the peptide in terms of favored conformations, although some changes in probability densities are observed. Dehydration did slightly decrease the helical content of the peptide (Fig. S7†).

3.3 Order of cyclization of prochlorosins

The cyclization process of prochlorosins is an attractive topic for investigation because for each substrate a single enzyme with one cyclization active site catalyzes the formation of a specific ring pattern out of many possible patterns. The formation of the final ring pattern is directly related to the order of cyclization since the first cyclization sets the final pattern. In the prochlorosin family, the three-dimensional structures of Pcn1.1, 2.1, 2.8, 2.10, 2.11 were determined by nuclear magnetic resonance (NMR) spectroscopy.[34] Among them, both Pcn1.1 and 2.8 contain two non-overlapping rings, but the rings are of different sizes (Fig. 4). Previous studies of ProcA2.8 have shown ring B (Ser13/Cys19) is formed first,[73] where the correct formation of ring B is important for subsequent formation of ring A (Cys3/Ser9) via preorganization of the peptide structure to facilitate cyclization.[79] To explore whether substrate sequence can be utilized to predict the order of cyclization, we ran 25 μs MD simulations for ProcA2.8 in solution and calculated the probability of the formation of the four possible rings (Cys3/Ser9, Ser13/Cys19, Cys3/Ser13, Ser9/Cys19). Our result shows that ring B has a significantly higher probability to form (Fig. 4A) and is more energetically favorable than ring A (Fig. S8A†). Thus, MD simulations mirror the experimental observation of ring B forming first.

Similarly, from the MSM weighted free energy landscape of ProcA1.1 (Fig. S8C†), we concluded that the conformation leading to ring B (Thr12/Cys16) is more stable than that leading to ring A (Cys3/Thr7) and has a higher probability of formation (Fig. 4C). To verify the order of cyclization predicted from

simulation, ProcM was incubated with ProcA1.1 and the assay was quenched at various time points as described in the Methods section. After quenching, the samples were derivatized *via* NEM alkylation of free thiols to identify partially cyclized intermediates, and the leader peptide was removed using the protease LahT$_{150}$ to obtain the core peptide.[75] Assay times and conditions were optimized to allow buildup of doubly dehydrated, singly cyclized intermediates. This peptide intermediate was then analyzed by LC-ESI MS/MS to obtain fragmentation data and determine the location(s) of NEM alkylation and thus cyclization.

The fragmentation data obtained of the monocyclized intermediate indicates Cys3 is alkylated and that ring B had been formed (Fig. 5). These findings indicate that cyclization of Cys16 occurs before that of Cys3 and that the cyclization order of ProcA1.1 is similar to that of ProcA2.8.[72] Notably, MD simulations correctly predicted the cyclization outcome. Importantly, C-to-N directionality is not uniform with ProcA substrates. For example, for WT ProcA3.3 the inner ring is formed first.[73] Hence, the directionality is not enforced by the enzyme but is likely a result of the inherent reactivity of each Cys-Dha/Dhb pair, which in turn is governed by the conformational energy landscape.

4 Conclusion

In this work, we ran molecular dynamics simulation for the wild type substrate peptide ProcA3.3 and its variants in solution. Without the lanthipeptide synthetase, ProcM, as a catalyst, we predicted the same ring pattern for the core peptides of two ProcA3.3 variants as the experimental results, with distance between reacting residues used as a proxy for the likelihood of a chemical reaction. Results from MD simulation provide an orthogonal line of evidence to support and explain how the core sequence can determine the final product's ring pattern, rather than the enzyme. Our simulations also predicted the correct trend for increased formation of non-overlapping ring patterns as a function of core peptide sequence. However, the simulations could not predict the preferred formation of the larger A-ring in WT ProcA3.3 over the smaller A′ and B′ rings. Formation of smaller rings is typically preferred based on entropic considerations, and the enzyme overcomes these inherent preferences. In a recent study, Habibi *et al.* showed that during the biosynthesis of another class II lanthipeptide, haloduracin β, the secondary structure in the core peptide may direct preferential interactions with the active sites of its synthetase and alter the post-translational modification pattern.[80] It is possible that similar interactions are also involved in A-ring formation in ProcA3.3.

In addition to correctly predicting trends in cyclization patterns, we demonstrate that in the biosynthetic pathway to prochlorosins, the dehydration step does not considerably alter the predicted solution structure of the precursor peptide. This finding suggests that even if the core sequence is not dehydrated, it can be used to determine the site-selectivity of the final products, an important finding for future prediction of ring patterns from sequence. Furthermore, we show that the

secondary structure of the core peptide sequence is a crucial factor to control formation of the final product by determining the conformational energy landscape in solution.

Another question we investigated in our work is the order of cyclization. We chose two candidates, ProcA1.1 and ProcA2.8 which both form two non-overlapping rings of different sizes, to see if we could utilize our simulations to predict the order of ring formation. Based on our MD simulation analysis of these two prochlorosins, the C-terminal ring has a higher probability to form and remain in a more energetically stable conformation, which suggests that the C-terminal ring will be first to form for both peptides. Previous study of ProcA2.8 indeed showed the C-terminal ring is the first ring to be formed during the biosynthesis by ProcM.[73] In this study, we also experimentally verified the prediction from the MD simulation that ProcA1.1 is also cyclized in a C-to-N terminal order. Thus, for both ProcA1.1 and ProcA2.8, we see the same directionality of cyclization with and without ProcM as a catalyst, which provides further evidence to suggest substrate sequence controls the order of cyclization. However, it is noteworthy that the core peptide sequence may not be the only factor that determines the order of cyclization. In addition to possible interactions with active site residues mentioned above, two studies have suggested that the leader peptide when physically attached to the core peptide affects the order of cyclization for lanthipeptide synthetases that have only one substrate peptide.[80,81]

Overall, we utilized molecular dynamic simulation to demonstrate that the substrate core peptide sequence likely controls the final ring pattern of the prochlorosins and the order of the ring formation through determining secondary structure. This study therefore aids in the understanding of the mechanism of ring formation in the ProcM system (and likely the SyncM system[32]) and the relationship between sequence and final product. We emphasize that our studies are specific for systems in which one enzyme modifies a large number of substrates with highly diverse sequences in which co-evolution of enzyme and substrates appears unlikely.[37] Our approach to correlate molecular dynamics of peptide in solution with final lanthipeptide ring pattern will most likely not be valid for systems in which the substrate peptide and the lanthipeptide synthetase have co-evolved to make one specific ring pattern that displays biological activity for which the enzyme likely actively governs the formation of those rings that are required for bioactivity. In addition to explaining order of ring formation and trends of ring patterns, our work may provide an effective strategy to facilitate the prediction of the ring pattern of lanthipeptides discovered by genome mining by analyzing the secondary structure of the core sequence and the conformational energy landscape. Compared to experimental investigations, the computational methodology is much faster and may accelerate bioengineering efforts and the design of novel lanthipeptides.

Data availability

The data and code can be found at the following GitHub repository: **https://github.com/ShuklaGroup/Lanthipeptide**.

Author contributions

Conflicts of interest

There are no conflicts to declare.

Acknowledgements

Notes and references

1 D. J. Newman and G. M. Cragg, *J. Nat. Prod.*, 2020, **83**, 770–803.

2 K. J. Hetrick and W. A. van der Donk, *Curr. Opin. Chem. Biol.*, 2017, **38**, 36–44.

3 J. I. Tietz, C. J. Schwalen, P. S. Patel, T. Maxson, P. M. Blair, H.-C. Tai, U. I. Zakai and D. A. Mitchell, *Nat. Chem. Biol.*, 2017, **13**, 470–478.

4 S. Luo and S.-H. Dong, *Molecules*, 2019, **24**, 1541.

5 K. Scherlach and C. Hertweck, *Nat. Commun.*, 2021, **12**, 3864.

6 M. H. Medema, T. de Rond and B. S. Moore, *Nat. Rev. Genet.*, 2021, **22**, 553–571.

7 L. Albarano, R. Esposito, N. Ruocco and M. Costantini, *Mar. Drugs*, 2020, **18**, 199.

8 R. H. Baltz, *J. Ind. Microbiol. Biotechnol.*, 2021, **48**, kuab044.

9 E. Kenshole, M. Herisse, M. Michael and S. J. Pidot, *Curr. Opin. Chem. Biol.*, 2021, **60**, 47–54.

10 P. G. Arnison, M. J. Bibb, G. Bierbaum, A. A. Bowers, T. S. Bugni, G. Bulaj, J. A. Camarero, D. J. Campopiano, G. L. Challis, J. Clardy, P. D. Cotter, D. J. Craik, M. Dawson, E. Dittmann, S. Donadio, P. C. Dorrestein, K.-D. Entian, M. A. Fischbach, J. S. Garavelli, U. Göransson, C. W. Gruber, D. H. Haft, T. K. Hemscheidt, C. Hertweck, C. Hill, A. R. Horswill, M. Jaspars, W. L. Kelly, J. P. Klinman, O. P. Kuipers, A. J. Link, W. Liu, M. A. Marahiel, D. A. Mitchell, G. N. Moll, B. S. Moore, R. Müller, S. K. Nair, I. F. Nes, G. E. Norris, B. M. Olivera, H. Onaka, M. L. Patchett, J. Piel, M. J. T. Reaney, S. Rebuffat, R. P. Ross, H.-G. Sahl, E. W. Schmidt, M. E. Selsted, K. Severinov, B. Shen, K. Sivonen, L. Smith, T. Stein, R. D. Süssmuth, J. R. Tagg, G.-L. Tang, A. W. Truman, J. C. Vederas, C. T. Walsh, J. D. Walton, S. C. Wenzel, J. M. Willey and W. A. van der Donk, *Nat. Prod. Rep.*, 2013, **30**, 108–160.

11 M. C. Walker, S. M. Eslami, K. J. Hetrick, S. E. Ackenhusen, D. A. Mitchell and W. A. Van Der Donk, *BMC Genomics*, 2020, **21**, 387.

12 M. Montalbán-López, T. A. Scott, S. Ramesh, I. R. Rahman, A. J. van Heel, J. H. Viel, V. Bandarian, E. Dittmann, O. Genilloud, Y. Goto, M. J. G. Burgos, C. Hill, S. Kim, J. Koehnke, J. A. Latham, A. J. Link, B. Martínez, S. K. Nair, Y. Nicolet, S. Rebuffat, H.-G. Sahl, D. Sareen, E. W. Schmidt, L. Schmitt, K. Severinov, R. D. Süssmuth, A. W. Truman, H. Wang, J.-K. Weng, G. P. van Wezel, Q. Zhang, J. Zhong, J. Piel, D. A. Mitchell, O. P. Kuipers and W. A. van der Donk, *Nat. Prod. Rep.*, 2021, **38**, 130–239.

13 J. Lubelski, R. Rink, R. Khusainov, G. N. Moll and O. P. Kuipers, *Cell. Mol. Life Sci.*, 2007, **65**, 455–476.

14 T. E. Smith, C. D. Pond, E. Pierce, Z. P. Harmer, J. Kwan, M. M. Zachariah, M. K. Harper, T. P. Wyche, T. K. Matainaho, T. S. Bugni, L. R. Barrows, C. M. Ireland and E. W. Schmidt, *Nat. Chem. Biol.*, 2018, **14**, 179–185.

15 K. I. Mohr, C. Volz, R. Jansen, V. Wray, J. Hoffmann, S. Bernecker, J. Wink, K. Gerth, M. Stadler and R. Müller, *Angew. Chem., Int. Ed.*, 2015, **54**, 11254–11258.

16 K. Meindl, T. Schmiederer, K. Schneider, A. Reicke, D. Butz, S. Keller, H. Gühring, L. Vértesy, J. Wink, H. Hoffmann, M. Brönstrup, G. Sheldrick and R. Süssmuth, *Angew. Chem., Int. Ed.*, 2010, **49**, 1151–1154.

17 L. M. Repka, J. R. Chekan, S. K. Nair and W. A. van der Donk, *Chem. Rev.*, 2017, **117**, 5457–5520.

18 D. Field, B. Collins, P. D. Cotter, C. Hill and R. P. Ross, *J. Mol. Microbiol. Biotechnol.*, 2007, **13**, 226–234.

19 T. Caetano, J. M. Krawczyk, E. Mösker, R. D. Süssmuth and S. Mendo, *Chem. Biol.*, 2011, **18**, 90–100.

20 S. Chen, S. Wilson-Stanford, W. Cromwell, J. D. Hillman, A. Guerrero, C. A. Allen, J. A. Sorg and L. Smith, *Appl. Environ. Microbiol.*, 2013, **79**, 4015–4023.

21 A. A. Vinogradov, J. S. Chang, H. Onaka, Y. Goto and H. Suga, *ACS Cent. Sci.*, 2022, **8**, 814–824.

22 D. Field, P. D. Cotter, C. Hill and R. P. Ross, *Front. Microbiol.*, 2015, **6**, 1363.

23 M. Montalbán-López, A. J. van Heel and O. P. Kuipers, *FEMS Microbiol. Rev.*, 2016, **41**, 5–18.

24 A. Kuthning, P. Durkin, S. Oehm, M. G. Hoesl, N. Budisa and R. D. Süssmuth, *Sci. Rep.*, 2016, **6**, 33447.

25 B. J. Burkhart, N. Kakkar, G. A. Hudson, W. A. van der Donk and D. A. Mitchell, *ACS Cent. Sci.*, 2017, **3**, 629–638.

26 K. J. Hetrick, M. C. Walker and W. A. van der Donk, *ACS Cent. Sci.*, 2018, **4**, 458–467.

27 X. Yang, K. R. Lennard, C. He, M. C. Walker, A. T. Ball, C. Doigneaux, A. Tavassoli and W. A. van der Donk, *Nat. Chem. Biol.*, 2018, **14**, 375–380.

28 S. Schmitt, M. Montalbán-López, D. Peterhoff, J. Deng, R. Wagner, M. Held, O. P. Kuipers and S. Panke, *Nat. Chem. Biol.*, 2019, **15**, 437–443.

29 G. N. Moll, A. Kuipers, R. Rink, T. Bosma, L. de Vries and P. Namsolleck, *Biochem. Soc. Trans.*, 2020, **48**, 2195–2203.

30 B. Li, D. Sher, L. Kelly, Y. Shi, K. Huang, P. J. Knerr, I. Joewono, D. Rusch, S. W. Chisholm and W. A. van der Donk, *Proc. Natl. Acad. Sci. U. S. A.*, 2010, **107**, 10430–10435.

31 A. Cubillos-Ruiz, J. W. Berta-Thompson, J. W. Becker, W. A. van der Donk and S. W. Chisholm, *Proc. Natl. Acad. Sci. U. S. A.*, 2017, **114**, E5424–E5433.

32 P. Arias-Orozco, M. Inklaar, J. Lanooij, R. Cebrián and O. P. Kuipers, *ACS Synth. Biol.*, 2021, **10**, 2579–2591.

33 W. Tang and W. A. van der Donk, *Biochemistry*, 2012, **51**, 4271–4279.

34 S. C. Bobeica, L. Zhu, J. Z. Acedo, W. Tang and W. A. van der Donk, *Chem. Sci.*, 2020, **11**, 12854–12870.

35 T. Le, K. J. D. Fouque, M. Santos-Fernandez, C. D. Navo, G. Jiménez-Osés, R. Sarksian, F. A. Fernandez-Lima and W. A. van der Donk, *J. Am. Chem. Soc.*, 2021, **143**, 18733–18743.

36 Y. Yu, Q. Zhang and W. A. van der Donk, *Protein Sci.*, 2013, **22**, 1478–1489.

37 T. Le and W. A. van der Donk, *Trends Chem.*, 2021, **3**, 266–278.

38 K. Lindorff-Larsen, S. Piana, R. O. Dror and D. E. Shaw, *Science*, 2011, **334**, 517–520.

39 S. Piana, J. L. Klepeis and D. E. Shaw, *Curr. Opin. Struct. Biol.*, 2014, **24**, 98–105.

40 Y. Kuroda, A. Suenaga, Y. Sato, S. Kosuda and M. Taiji, *Sci. Rep.*, 2016, **6**, 19479.

41 U. R. Shrestha, P. Juneja, Q. Zhang, V. Gurumoorthy, J. M. Borreguero, V. Urban, X. Cheng, S. V. Pingali, J. C. Smith, H. M. O'Neill, *et al.*, *Proc. Natl. Acad. Sci. U. S. A.*, 2019, **116**, 20446–20452.

42 J. Damjanovic, J. Miao, H. Huang and Y.-S. Lin, *Chem. Rev.*, 2021, **121**, 2292–2324.

43 L. Schrödinger and W. DeLano, *PyMOL*, http://www.pymol.org/pymol.

44 L. Martínez, R. Andrade, E. Birgin and J. M. Martínez, *J. Comput. Chem.*, 2009, **30**, 2157–2164.

45 D. A. Case, I. Y. Ben-Shalom, S. R. Brozell, D. S. Cerutti, T. E. III, V. W. D. Cruzeiro, T. A. Darden, R. E. Duke, D. Ghoreishi, M. K. Gilson, H. Gohlke, A. W. Goetz, D. Greene, R. Harris, N. Homeyer, Y. Huang, S. Izadi, A. Kovalenko, T. Kurtzman, T. S. Lee, S. LeGrand, P. Li, C. Lin, J. Liu, T. Luchko, R. Luo, D. J. Mermelstein, K. M. Merz, Y. Miao, G. Monard, C. Nguyen, H. Nguyen, I. Omelyan, A. Onufriev, F. Pan, R. Qi, D. R. Roe, A. Roitberg, C. Sagui, S. Schott-Verdugo, J. Shen, C. L. Simmerling, J. Smith, R. Salomon-Ferrer, J. Swails, R. C. Walker, J. Wang, H. Wei, R. M. Wolf, X. Wu, L. Xiao, D. M. York and P. A. Kollman, *AMBER 2018*, University of California, San Francisco, 2018.

46 C. I. Bayly, P. Cieplak, W. Cornell and P. A. Kollman, *J. Phys. Chem.*, 1993, **97**, 10269–10280.

47 M. J. Frisch, G. W. Trucks, H. B. Schlegel, G. E. Scuseria, M. A. Robb, J. R. Cheeseman, G. Scalmani, V. Barone, G. A. Petersson, H. Nakatsuji, X. Li, M. Caricato, A. V. Marenich, J. Bloino, B. G. Janesko, R. Gomperts, B. Mennucci, H. P. Hratchian, J. V. Ortiz, A. F. Izmaylov, J. L. Sonnenberg, D. Williams-Young, F. Ding, F. Lipparini, F. Egidi, J. Goings, B. Peng, A. Petrone, T. Henderson, D. Ranasinghe, V. G. Zakrzewski, J. Gao, N. Rega, G. Zheng, W. Liang, M. Hada, M. Ehara, K. Toyota, R. Fukuda, J. Hasegawa, M. Ishida, T. Nakajima, Y. Honda, O. Kitao, H. Nakai, T. Vreven, K. Throssell, J. A. Montgomery Jr, J. E. Peralta, F. Ogliaro, M. J. Bearpark, J. J. Heyd, E. N. Brothers, K. N. Kudin, V. N. Staroverov, T. A. Keith, R. Kobayashi, J. Normand, K. Raghavachari, A. P. Rendell, J. C. Burant, S. S. Iyengar, J. Tomasi, M. Cossi, J. M. Millam, M. Klene, C. Adamo, R. Cammi, J. W. Ochterski, R. L. Martin, K. Morokuma, O. Farkas, J. B. Foresman and D. J. Fox, *Gaussian 16 Revision C.01*, Gaussian Inc. Wallingford CT, 2016.

48 R. J. Loncharich, B. R. Brooks and R. W. Pastor, *Biopolymers*, 1992, **32**, 523–535.

49 J. Aqvist, P. Wennerström, M. Nervall, S. Bjelic and B. Brandsdal, *Chem. Phys. Lett.*, 2004, **384**, 288–294.

50 E. Braun, J. Gilmer, H. B. Mayes, D. L. Mobley, J. I. Monroe, S. Prasad and D. M. Zuckerman, *Living J. Comp. Mol. Sci.*, 2019, **1**, 5957.

51 J.-P. Ryckaert, G. Ciccotti and H. J. Berendsen, *J. Comput. Phys.*, 1977, **23**, 327–341.

52 C. W. Hopkins, S. L. Grand, R. C. Walker and A. E. Roitberg, *J. Chem. Theory Comput.*, 2015, **11**, 1864–1874.

53 G. R. Bowman, D. L. Ensign and V. S. Pande, *J. Chem. Theory Comput.*, 2010, **6**, 787–794.

54 M. Lawrenz, D. Shukla and V. S. Pande, *Sci. Rep.*, 2015, **5**, 7918.

55 T. J. Lane, D. Shukla, K. A. Beauchamp and V. S. Pande, *Curr. Opin. Struct. Biol.*, 2013, **23**, 58–65.

56 S. Dutta, B. Selvam and D. Shukla, *ACS Chem. Neurosci.*, 2022, **13**, 379–389.

57 S. Dutta, B. Selvam, A. Das and D. Shukla, *J. Biol. Chem.*, 2022, **298**, 101764.

58 D. Sculley, *Proceedings of the 19th international conference on World wide web – WWW '10*, 2010.

59 D. Shukla, C. X. Hernández, J. K. Weber and V. S. Pande, *Acc. Chem. Res.*, 2015, **48**, 414–422.

60 B. E. Husic and V. S. Pande, *J. Am. Chem. Soc.*, 2018, **140**, 2386–2396.

61 H. Sidky, W. Chen and A. L. Ferguson, *J. Phys. Chem. B*, 2019, **123**, 7999–8009.

62 H. Wu and F. Noé, *J. Nonlinear Sci.*, 2019, **30**, 23–66.

63 G. Pérez-Hernández, F. Paul, T. Giorgino, G. D. Fabritiis and F. Noé, *J. Chem. Phys.*, 2013, **139**, 015102.

64 F. Noé and C. Clementi, *J. Chem. Theory Comput.*, 2015, **11**, 5002–5011.

65 M. K. Scherer, B. Trendelkamp-Schroer, F. Paul, G. Pérez-Hernández, M. Hoffmann, N. Plattner, C. Wehmeyer, J.-H. Prinz and F. Noé, *J. Chem. Theory Comput.*, 2015, **11**, 5525–5542.

66 R. T. McGibbon, K. A. Beauchamp, M. P. Harrigan, C. Klein, J. M. Swails, C. X. Hernández, C. R. Schwantes, L.-P. Wang, T. J. Lane and V. S. Pande, *Biophys. J.*, 2015, **109**, 1528–1532.

67 W. Humphrey, A. Dalke and K. Schulten, *J. Mol. Graphics*, 1996, **14**, 33–38.

68 J. D. Hunter, *Comput. Sci. Eng.*, 2007, **9**, 90–95.

69 C. Yuan, H. Chen and D. Kihara, *BMC Bioinf.*, 2012, **13**, 292.

70 H. Kamisetty, S. Ovchinnikov and D. Baker, *Proc. Natl. Acad. Sci. U. S. A.*, 2013, **110**, 15674–15679.

71 M. M. Sultan, G. Kiss and V. S. Pande, *Nat. Chem.*, 2018, **10**, 903–909.

72 C. J. Thibodeaux, T. Ha and W. A. van der Donk, *J. Am. Chem. Soc.*, 2014, **136**, 17513–17529.

73 S. Mukherjee and W. A. van der Donk, *J. Am. Chem. Soc.*, 2014, **136**, 10450–10459.

74 S. C. Bobeica and W. A. van der Donk, *Methods in Enzymology*, Elsevier, 2018, pp. 165–203.

75 S. C. Bobeica, S.-H. Dong, L. Huo, N. Mazo, M. I. McLaughlin, G. Jiménez-Osés, S. K. Nair and W. A. van der Donk, *eLife*, 2019, **8**, e42305.

76 Y. Yu, S. Mukherjee and W. A. van der Donk, *J. Am. Chem. Soc.*, 2015, **137**, 5140–5148.

77 C. R. Schwantes and V. S. Pande, *J. Chem. Theory Comput.*, 2013, **9**, 2000–2009.

78 S. Dutta and D. Shukla, *Communications Biology*, 2023, **6**, 485.

79 J. D. Hegemann, S. C. Bobeica, M. C. Walker, I. R. Bothwell and W. A. van der Donk, *ACS Synth. Biol.*, 2019, **8**, 1204–1214.

80 Y. Habibi, N. W. Weerasinghe, K. A. Uggowitzer and C. J. Thibodeaux, *J. Am. Chem. Soc.*, 2022, **144**, 10230–10240.

81 C. J. Thibodeaux, J. Wagoner, Y. Yu and W. A. van der Donk, *J. Am. Chem. Soc.*, 2016, **138**, 6436–6444.