Large scale ab initio modeling of structurally uncharacterized antimicrobial peptides reveals known and novel folds

Mara Kozic1,2 | Stephen J. Fox2 | Jens M. Thomas1 | Chandra S. Verma2,3,4 | Daniel J. Rigden1

1Institute of Integrative Biology, University of Liverpool, Liverpool, L69 7ZB, U.K.
2Agency for Science, Technology and Research (A*STAR), Bioinformatics Institute, Singapore
3Department of Biological Sciences, National University of Singapore, Singapore
4School of Biological Sciences, Nanyang Technological University, Singapore

Correspondence
Daniel J. Rigden, Institute of Integrative Biology, Room 101, Biosciences Building, University of Liverpool, Crown St., Liverpool L69 7ZB, U.K.
E-mail: drigden@liverpool.ac.uk

Abstract
Antimicrobial resistance within a wide range of infectious agents is a severe and growing public health threat. Antimicrobial peptides (AMPs) are among the leading alternatives to current antibiotics, exhibiting broad spectrum activity. Their activity is determined by numerous properties such as cationic charge, amphipathicity, size, and amino acid composition. Currently, only around 10% of known AMP sequences have experimentally solved structures. To improve our understanding of the AMP structural universe we have carried out large scale ab initio 3D modeling of structurally uncharacterized AMPs that revealed similarities between predicted folds of the modeled sequences and structures of characterized AMPs. Two of the peptides whose models matched known folds are Lebocin Peptide 1A (LP1A) and Odorranain M, predicted to form β-hairpins but, interestingly, to lack the intramolecular disulfide bonds, cation-π or aromatic interactions that generally stabilize such AMP structures. Other examples include Ponericin Q42, Latarcin 4a, Kassinatuerin 1, Ceratotoxin D, and CPF-B1 peptide, which have α-helical folds, as well as mixed αβ folds of human Histatin 2 peptide and Garvicin A which are, to the best of our knowledge, the first linear αββ fold AMPs lacking intramolecular disulfide bonds. In addition to fold matches to experimentally derived structures, unique folds were also obtained, namely for Microcin M and Ipomicin. These results help in understanding the range of protein scaffolds that naturally bear antimicrobial activity and may facilitate protein design efforts towards better AMPs.

KEYWORDS
ab initio modeling, antimicrobial peptide, antimicrobial resistance, protein structure-function, structure prediction

1 INTRODUCTION

Antimicrobial resistance within a wide range of infectious agents is a severe and growing public threat.1 A 2013 report from the American Centers for Disease Control estimated that over 23,000 deaths and more than 2 million cases of infections were caused by drug-resistant bacteria in the USA alone in 2013.2 Antimicrobial proteins and peptides are among the leading alternatives to current antibiotics, exhibiting activity against a wide variety of bacteria and other microbes3 and are of particular interest since they have maintained their effectiveness over hundreds of millions of years demonstrating that definitive resistance to them is not readily acquired by bacteria. Proteins with antimicrobial activity typically contain fewer than 200 residues, with most much shorter—12–100 residues—and hence are commonly known as antimicrobial peptides (AMPs).4,5 They are produced by the immune systems of species from all domains of life. Most of them are cationic at physiological pH, with net positive charge ranging from +2 to +9, and hydrophobic with an amphipathic structure.6 Another property...
important in targeting bacterial membranes is amino acid composition. Trp residues are frequently found in AMPs and multiple studies have highlighted their importance in interactions with biological membranes. Peptides containing only Arg and Trp residues can be highly antimicrobial.7 Trp residues are critical for anchoring and insertion of peptides into the membrane8,9 and their removal can have drastic effects on the antimicrobial activity of peptides.10 Simulations have been used extensively to probe these interactions.11 Trp is stabilized by hydrogen bond interactions with water molecules and headgroups at the interface.12,13 However, the Trp residues can equally lie inside the membranes where their bulky sidechains can disrupt the packed lipid chains.9 Similar behavior is also seen for Tyr and to some extent for Phe side chains.14,15 It is common to see the insertion of Trp residues in efforts to design AMPs.16,17

AMPs can contain secondary structures of all kinds—helices, β-sheets, extended, and loop regions. Generally, AMPs can be divided in 4 structural groups: α, β, αβ, and non-αβ.18 The most abundant structural group of AMPs are amphipathic α-helices, followed by αβ and all-β structures.19,20 Aside from short, linear α-helical peptides, more complex all-α folds have also been found. These include helix hairpins and helical bundles, commonly found in class II bacteriocins such as the well-known food preservative nisin. AMPs with αβ structure often have disulfide bonds, such as those seen in plant defensins’ cysteine-stabilized αβ (CSαβ) motif. All-β AMPs have structures comprised of multiple β-strands, for instance a simple β-hairpin stabilized by a circular backbone and disulfide bonds (as seen in θ-defensins of non-human primates) or the cysteine-stabilized triple-stranded β-sheet seen in human defensins. Unlike conventional antibiotics, which generally target metabolic enzymes, AMPs act mainly by membrane-targeting mechanisms and are selective due to the difference in charge of prokaryotic and eukaryotic cell membranes. Furthermore, AMPs have faster antimicrobial activity than conventional antibiotics.21,22

Generally speaking, AMPs can be divided into two mechanistic classes: membrane disruptive and non-membrane disruptive (acting on intracellular targets). Disruption of the negatively charged prokaryotic membrane is the predominant mode of action of AMPs, with three main mechanisms proposed: the barrel stave, toroidal, and carpet model23,24.

AMPs have therapeutic potential as bioactive coatings for needles, catheters, implants, surgical tools, bandages, and even contact lenses. However, only a few have been approved for clinical use, and only for topical application, mainly due to their toxic properties.25,26 The main difficulty in AMP drug development is our lack of understanding of modes of action.27 The availability of structural information is crucial in facilitating AMP design efforts to predict, understand and implement knowledge-based enhancement of activities yet the pace of structural determination lags far behind AMP discovery; currently, there are over 2000 AMP sequences known, but only about 10% of them are structurally characterized. Researchers have used different methods in order to optimize antimicrobial activity on known protein scaffolds. Quantitative structure-activity relationship (QSAR), Regression models and Machine Learning approaches such as Artificial Neural Network (ANN), Support Vector Machine (SVM), Random Forests (RF) and Hidden Markov Models (HMM) are some of the approaches employed.28,29 However, most of these studies are sequence-based, and design efforts based on the structural properties of more complex folds, such as studies on β-hairpins by Edwards et al.30 or Yang et al.,31 are less common. Notably, sequence information on its own is not sufficient to determine relevant properties of folds, such as the amphipathicity or dipole moment of the molecule.

Due to coevolution with pathogens, AMP sequences are exceptionally diverse.32 AMP genes have been found to evolve rapidly in both vertebrates and invertebrates as a result of rapid gene duplication, diversification, and positive selection. This has been documented for mammal, bird, amphibian, and insect AMPs. Positive selection in AMPs seems to be highest immediately after gene duplication, although there could be a limit on observing high number of nonsynonymous substitutions in distantly related sequences.33 It is known that AMP and immune genes evolve much faster than non-immune genes34,35 with other work showing that AMPs can evolve 3 times faster than other proteins.36 Due to this rapid evolution, reconstruction of the evolutionary history of AMPs can be a challenging task.37 Moreover, this limits the possibilities and scope of homology modeling that can be performed for known sequences: not only is the number of available templates limited (see above) but evolutionary relationships between targets and templates are often hard to discern. In such cases, where no experimentally solved homologous structures exist, or exist but cannot be identified, models have to be constructed from scratch by performing ab initio modeling. Successful ab initio modeling of proteins without structurally characterized homologs with RMSD values around 2–5 Å has been reported for sequences shorter than 100–120 residues,38–40 with the most recent CASP free modeling experiments showing this limitation to be at 150 residues.41,42 AMPs are generally small in size which makes them particularly suitable for ab initio modeling. Furthermore, they often contain disulfide bonds which, if their connectivity can be predicted, provide valuable additional data to guide modeling. Most often, cysteines in extracellular proteins come in even numbers.43 In an overview of disulfide-containing AMPs, Lehrer44 discusses peptides with intermolecular as well as intramolecular disulfide bonds. AMPs with one cysteine are quite rare and have been found to form hetero- or homodimers. While redox status is known to have an effect on antimicrobial activity,45,46 Lehrer’s overview46 does not give examples of AMPs containing two reduced cysteines. Interestingly, when its cysteines are reduced by the host, human antimicrobial peptide β-defensin 1 shows increased activity.46

Here, we carry out a large scale ab initio modeling of the structures of structurally uncharacterized AMPs with Rosetta47 aiming to improve our understanding of the AMP structural universe. Although it is well-known that membrane-active AMPs can undergo structural changes when adopting a functional conformation at the membrane,48 prediction of structures in aqueous solution can be expected to illuminate non-obvious evolutionary relationships and shed light on structural determinants of initial membrane interaction. We assembled a protocol to create a representative set of AMP sequences which have no predicted homologs of known 3D structure, and predicted disulfide bonds in order to facilitate their modeling. Following ab initio modeling, we tested their stability, compared their 3D structures against characterized AMPs and found fold matches as well as several unique folds.
2 MATERIALS AND METHODS

2.1 Sequence assembling and processing

Sequences longer than 20 and shorter than 120 amino acids were collected from UniProt\textsuperscript{49} and APD2\textsuperscript{19} on 17 March 2015. APD2 was chosen from the several AMP databases available since it is manually curated, and comparatively large and up-to-date. The UniProt release at the time was 2015.03. Sequence redundancy was reduced to a threshold of 45\% using CD-HIT\textsuperscript{50} and its global alignment option. HHpred\textsuperscript{51} was used to detect sequences with structurally characterized homologs in the PDB70 database as at 6 September 2014. PDB70 is a version of PDB that is redundancy-reduced to 70\% sequence identity.\textsuperscript{52} Upon inspection of the results, three conditions were required to be satisfied before a given AMP could be considered to have a homolog: (1) HHpred fold match probability higher than 90\%, (2) alignment coverage of query sequence higher than 40\%, and (3) absence of any mismatch greater than 2-fold in length of query and hit. Sequences with >35\% residues with a IUpred\textsuperscript{53} score of 0.5 or above were considered to be intrinsically disordered AMPs. Since ab initio modeling is not suitable for these proteins, they were not considered further.

2.2 Disulfide bond prediction, ab initio modeling, ab initio benchmarking and clustering

Disulfide bond predictions were made using Disulfind, Dianna, and Dinosolve.\textsuperscript{54-56} Disulfind and Dianna were run at their respective servers while Dinosolve was run locally (and used the nr database\textsuperscript{57} from 17 March 2015 and PSI-BLAST v.2.2.26\textsuperscript{58}). Consensus predictions deriving from 2 independent methods were obtained: there were no cases where all three programs agreed on all disulfides. For AMPs with 2 cysteines, it was assumed that the disulfide bond exists.\textsuperscript{44} For those AMPs with 3 or 4 cysteines, we ran Rosetta\textsuperscript{47} modeling with all possible combinations. AMPs with 5 and more cysteines were run with any consensus and without disulfide constraints. It should be noted, however, that although intermolecular disulfide bonds in AMPs are considered to be rare,\textsuperscript{59} peptides with odd number of cysteines, could form dimers or oligomerize, such as recently seen in rodent \(\alpha\)-defensin-related AMPs.\textsuperscript{60}

Ab initio modeling of 184 AMPs was performed with Rosetta software using the fix_disulf and relax flags. Use of the nohoms flag to exclude homologous fragments was unnecessary since modeling was only done for targets which HHpred bore no obvious homology with PDB entries. Where an AMP target contained a modified amino-acid, it was modeled using the natural, unmodified version. Where the predicted structure for such an AMP proved to be of interest, the likely consequences of the unmodified modification were considered later. For each AMP, 1000 models were made. Models defined as successful by Rosetta (meaning they passed the filters that eliminate models with non-protein like features) were clustered into 10 groups using SPICKER v. 2.0\textsuperscript{61} in order to identify the likely near-native structure for each AMP. Larger and more homogeneous clusters indicate more reliable fold candidates. We performed Rosetta benchmarking on AMPs of known structure (fold matches of our modeled AMPs, see Tables S1-S3 and Figures S1-S3) to detect a threshold value for cluster size to refer to when inspecting our models. For the benchmarking, the nohoms flag was used when running Rosetta in order to exclude fragments from the target and homologous structures. We chose to include modeled structures where the largest cluster size was at least 25\% of the total number of successfully modeled structures. However, structures with lower percentages were scrutinized and considered if the centroids of three largest clusters had similar folds (Figures S4-S7).

2.3 Fold matching and visual representation of matches

We assembled a database of experimentally solved AMP structures to compare models against. To ensure that we collected as many structures as possible, PDB codes were collected from both APD2 and UniProt in the following manner: first, an APD2 (March 2015) search was run with filter Original Location: PDB, which resulted in 229 PDB entries. Then, a UniProt (April 2015) search was run with keywords keyword: "Antimicrobial [KW-0929]" annotation:(type:peptide) database:(type:pdb). The results of this run were additionally filtered: only one PDB structure was taken for each UniProt entry, prioritized such that: (1) an X-ray structure was chosen if possible, (2) the X-ray structure with the highest resolution and matching start and end positions, compared to the AMP, was used, (3) if an X-ray structure was not available, the first NMR structure to match the start and end residue was chosen, and (4) the first chain was chosen. Structures with non-matching start and end residues were omitted if the mismatch was greater than 5 amino acids.

For each modeled AMP, a structure similarity search was carried out with GESAMT\textsuperscript{62} to compare the three largest clusters’ centroid models against the local database of AMP experimental structures. The results were then filtered so that only those modeled structures meeting all of the following conditions were left: (1) Q-score (a measure of structural similarity ranging from 0 to 1) \(\geq 0.3\), (2) query (AMP model) sequence length \(< 1.5 \times \) match (experimental structure) sequence length, that is, query can’t be more than 50\% larger than match, (3) match sequence length \(< 1.5 \times \) model sequence length, that is, match cannot be more than 50\% larger than query, and (4) number of aligned residues \(\geq 0.7 \times \) query sequence length, that is, the alignment covers at least 70\% of the query. Additionally, after filtering out results, fold matches were inspected manually: a model was considered to have a fold match if the number and type of secondary structure elements was similar. Ab initio models with tertiary structure matching at least one of the top 3 filtered fold matches were considered further.

In order to visualize the structural similarity, GESAMT was run in all-vs-all fashion on a set of AMPs from the PDB90 and matching modeled structures. The resulting similarity matrix was used as input for CLANS software\textsuperscript{63} to cluster the structures. The CLANS software was used to visualize clusters of modeled structures of AMPs and the matching folds in the PDB,\textsuperscript{52} 56 structures in total. 100 152 rounds of convex clustering using a value of 0.4 for standard deviation cutoff and requiring a minimum of 2 sequences per cluster initially clustered 49 out of 56 structures into eight clusters.
Since GESAMT employs a topology-dependent algorithm we additionally used the topology-independent superposition method CLICK to search AMP models not matching by GESAMT against the same local database of experimentally determined AMP structures. Again the results were filtered so that the query couldn’t be more than 50% larger than the match, and the match could not be more than 50% larger than query. For models where no matches with AMPs were found, we ran additional CLICK database search on all protein chains from the PDB90 (not just AMP structures) and filtered results again in a similar manner, after which matches with Z-score values higher than 3 were taken forward. In cases where all of the Z-score values were lower than this threshold, we lowered this value to 2. Finally, all of the matches that were left after the filtering were visually inspected.

### 2.4 | Stability of peptides

Molecular Dynamics (MD) simulations were performed using the AMBER package and AMBER FF14SB force field. Simulations with explicit solvent were performed using TIP3P water molecules with a 12 Å buffer between peptide atoms and the edge of a rectangular box. For each simulation, 10 000 steps of minimization were performed, with the first 5 000 using the steepest descent algorithm followed by 5 000 steps of conjugate gradient. The system was heated to 300 K in two steps; first heating from 0 to 100 K for 5 ps followed by heating from 100 to 300 K for another 100 ps, both using the Langevin thermostat. In the production step, we simulated the system at 300 K and 1 atm using the Berendsen barostat for 100 ns. Simulations with implicit solvent were run for 1 m.

For the last 50 ns of each simulation, structural alignment was performed on C\text{a} atoms of residues that formed regular secondary structure in the Rosetta model. RMSD clustering was carried out using MMTSB Tool Set based on those C\text{a} atoms. The structure closest to the centroid model was taken as a representative for each highly populated cluster.

### 3 | RESULTS AND DISCUSSION

In order to select and process AMPs, a workflow was implemented (Figure 1) to collect a non-redundant set of AMP sequences, eliminating those whose fold could be reliably inferred by homology detection and those predicted to be largely intrinsically disordered. Ab initio modeling of the resulting set, with or without predicted disulfide bonding as an additional constraint, was carried out using Rosetta. Clustering of the results determined candidate fold predictions, which were then compared to known AMP structures. Due to evolutionary constraints, protein folds can remain conserved even when there is an apparent lack of homology. Similarity between our models and known AMP structures could therefore result from distant, unsuspected homology. Similar folds can also arise as a product of convergent evolution; the best example among AMPs are defensins, which are taxonomically widespread over insects, mammals and plants, and are found to adopt a variety of folds such as β-sheets (triple-stranded β-sheet of Human Neutrophil α-defensin HNP-3 is an example), cyclic backbone...
results. Among 216 modeling runs, 48% of peptides had largest cluster containing less than 25% of the total number of successful models, 20% of peptides between 25 and 39% of the total, 10% of peptides between 40 and 49% and 22% had largest cluster containing between 50 and 100% of the total number of successful models (Figure 2). All-α structures were predominant in the most reliably modeled categories (Figure 2). This meant that, initially, 52% of structures were taken as reliable. Where the largest cluster was not larger than 25%, a comparison of the top, second and third largest cluster centroids was carried out. In two cases where these matched visually, the prediction was also considered of interest.

3.1 | Visual representation of fold matches

Modeled structures that matched AMP folds in the PDB were clustered using CLANS in a semi-automated manner along with the corresponding fold matches (31 models and 25 fold matches making a total of 56 structures). Several modes of clustering were trialled but none proved capable of results that were fully in accord with expert assignment based on visual examination. For example, proteins with mixed αβ topologies sometimes allied more closely with β-hairpins, through a good fit of that portion, rather than with proteins with the same αβ overall topology but more poorly matching β-structure. Therefore, some manual (re)assignments were made to fine-tune an initial clustering of 8 groups for presentation purposes (Figure S8) and for discussion below. Three clusters contained β-hairpins and were combined and joined with a single modeled β-hairpin structure (Odorranain M1, AP01300) left unassigned by the original clustering. One experimental structure (Mytilin B, PDB code 2EEM) from this group was reassigned to a group of two combined clusters containing αβ folds joined by two modeled structures left unassigned, one αβ (Garvicin A, AP02402) the other ββ (Rattusin, AP02178). Four helix hairpins that were left unassigned, namely experimental structures of Sublancin 16B (PDB code 2MUJ), Thurincin H (2LZI), Thurincin CD (2LZX), and EcAMP1 (2L2R) were joined with two clusters containing a continuum of v-shaped, helix-kink-helix and helix hairpin structures. As a result, the structures were finally clustered into the four groups shown in Figure S8: (1) v-shaped, helix-kink-helix and helix hairpins which form a continuum of structures shown in red, (2) αββ folds and βαβ folds shown in blue, (3) β-hairpins shown in magenta, and (4) helix bundles shown in green. These groups contain 32, 10, 11, and 3 structures, respectively. In Figure S8, structures with greater similarity (higher Q-scores) are positioned at shorter separations. We next discuss the results in each fold family.

3.2 | Fold matches

All the fold matches shown here had a Q-score ≥0.3 and were additionally manually screened so that matches were considered only when the AMP model and the experimentally determined matching structure were not too dissimilar in length and aligned over a majority of the model structure (see Methods).
3.2.1 | Helix bundles

A single confidently modeled AMP, the bacteriocin Lacticin Q, was predicted to fold as a helix bundle. Its ab initio modeling resulted in a large cluster comprising more than 30% of the successfully modeled structures, and suggested the presence of four helices, spanning residues 2–14, 16–25, 28–35, and 38–49, respectively. PSIPRED structure prediction for Lacticin Q identified two helices (residues 4–13 and 38–49) with high scores; and moderately high helical propensity for the other two helices predicted by Rosetta. GESAMT resulted in only two fold matches for our modeled structure of Lacticin Q with our local AMP structure database: Enterocins 7A and 7B, both also bacteriocins. The fold matches had Q-scores of 0.589 and 0.558, RMSD values of 1.8 and 1.9 Å on Cα atoms, and sequence identities of 16% and 17% for Enterocins 7A and 7B, respectively. Similarly, the Dali server gave Z scores of 4.4 and 4.2, a slightly higher RMSD of 1.9 and 2.3 Å on Cα atoms and sequence identities of 16% and 19% for Enterocins 7A and 7B, respectively.

Bacteriocins are antimicrobial peptides synthesized by the ribosomes of a variety of bacteria (both Gram-positive and Gram-negative). Cotter et al. categorized bacteriocins into three classes: class I, also known as lantibiotics, are post-translationally modified peptides containing amino acids called lanthionines; class II are a heterogeneous group of small heat-stable non-lanthionine containing peptides which may have disulfide/thioether bonds; and class III are large, heat-labile, lytic proteins called bacteriolysins. Lacticin Q belongs to class II and shows selectivity for Gram-positive bacteria at the strain level suggesting that membrane lipid composition might not be the only determinant of its antimicrobial activity. It is also known that the peptide causes accumulation of hydroxyl radicals.

It has been suggested that circular bacteriocins share a common overall structural motif of a saposin fold, that is, four helices surrounding a hydrophobic core, regardless of low shared sequence identity and our results are consistent with this (Figure 3). Tryptophan residues are known to be involved in protein folding as well as to have a tendency for burial at the bilayer interface. Another common feature of circular and leaderless bacteriocins is the presence of solvent-exposed tyrosine or tryptophan residues that are likely to facilitate membrane penetration. A comparison of the modeled structure of Lacticin Q with the NMR structure of Enterocin 7A showed that the aromatic side chains located at different positions in the sequences were seen at structurally equivalent locations in the two structures. Out of three surface tryptophans in Enterocin 7A, two, namely W13 and W31, are in close proximity and have the same orientation as seen for W23 and W32, respectively, in Lacticin Q. Strikingly, these Trp pairs are found in corresponding positions in 3D space despite not aligning in the structure-based sequence alignment (Figure 3). Although Lohans et al. had tentatively proposed a relationship between Lacticin Q and Enterocins 7A and 7B, based on weak sequence similarity in the N- and C-terminal helical regions, the striking structural similarity and structural correspondence between likely functional aromatic residues are strongly indicative of homology over the whole protein despite the low overall degree of sequence identity.

Lacticin Q shows selective antimicrobial activity against various Gram-positive bacteria. It is frequently compared to another class II
bacteriocin, nisin, for its nanomolar range antimicrobial activity, pore size and ATP efflux. However, compared to nisin, lacticin Q is a leaderless bacteriocin—the peptide is synthesized without the N-terminal leader sequence that is otherwise removed when exporting from cells. Lacticin Q, Enterocin 7A and 7B are unmodified leaderless N-formylated bacteriocins that adopt helical conformations in solution. They all have an overall net charge of +6 which induces binding to negatively charged lipids although a comparison of the electrostatic properties of the Lacticin Q model and Enterocin crystal structures shows no strong similarity (not shown). The authors suggest a huge properties of the Lacticin Q model and Enterocin crystal structures as in cyclic, while reducing disulfides in Arenicin-1 led to decreased activity. Interestingly, it has been shown by Ma et al. that disulfides were not only dispensable for Thanatin activity and toxicity, but that the secondary structure was maintained in their absence as well. The other fold matches are \( \beta \)-hairpins lacking disulfide bonds. Although a Conibear et al. study showed that a cyclic \( \beta \)-hairpin structures, our hypothesis that Tachyplesin I forms anion-selective pores and translocates, Androctonin seems to act via a detergent-like mechanism. While Matsuzaki et al. suggested that Tachyplesin I shows sequence similarities among the PDB fold matches to these 3 peptides, we find three \( \beta \)-defensins and structures with linear backbones, such as protegrins, thanatin, gomesin, tachyplesins, polyphemusins, and arenicins. Among the PDB fold matches to these 3 peptides, we find three \( \beta \)-defensins: BTD-2, RTD-1, and HTD2 (Retrocyclin-2). These are backbone cyclic \( \beta \)-hairpin AMPs containing three parallel disulfide bonds also known as the cystine ladder motif. A study on 18 residue long BTD-2 \( \beta \)-defensin analogues by Conibear et al. showed that a cyclic backbone appears to be essential for membrane activity resulting in antibacterial effects, as was also reported earlier by Tang et al. for RTD-1 \( \beta \)-defensin. However, the disulfide bonds have been shown to be essential for stability of these AMPs, as well as for resistance to the action of proteolytic enzymes. Disulfides can be either essential and dispensable for the activity of \( \beta \)-hairpins: Protegrin-1 was as active in linear form as in cyclic, while reducing disulfides in Arenicin-1 led to decreased activity. Interestingly, it has been shown by Ma et al. that disulfides were not only dispensable for Thanatin activity and toxicity, but that the secondary structure was maintained in their absence as well.

The other fold matches are \( \beta \)-hairpins with linear backbones: Androctonin (PDB code 1CZ6), Tachyplesin I (1WO1) and pLR (2M3N). Androctonin and Tachyplesin I show sequence similarities and both contain two disulfide bonds with the same connectivity: 1–2, 2–3. Even though they are both active against Gram-positive and Gram-negative bacteria, Hetru et al. suggest that a proper comparison of their modes of action cannot be made due to the different number of residues separating the cysteines. While Matsuzaki et al. hypothesized that Tachyplesin I forms anion-selective pores and translocates, Androctonin seems to act via a detergent-like mechanism. pLR \( \beta \)-hairpin fold match is a peptide with a linear backbone which has one disulfide and flexible terminal regions. It is likely that, as with other peptides, pLR forms oligomers in solution whose large size prevents effective traversing of the membrane in Gram-negatives.

A recent review by Panteleev et al. classified \( \beta \)-hairpins into subgroups by the number of disulfides but did not report any AMPs without this feature. Interestingly, none of our modeled structures contain intramolecular disulfides (Figure 4)—they are the first reported \( \beta \)-hairpins lacking disulfide bonds. Although a Conibear et al. study on BTD-2 cyclic hairpin analogues speculates that the absence of disulfide bonds places a limit on the lengths of \( \beta \)-hairpin structures, our models suggest otherwise. Tachyplesin analogs including replacements

### TABLE 1

| Modeled AMP and the corresponding APD2 IDs | C1 size % | Fold matches | Fold matches' corresponding PDB and APD2/UniProt IDs | Identity % |
|------------------------------------------|-----------|--------------|-----------------------------------------------------|------------|
| Lebocin Peptide 1A (AP00030)             | 50        | RTD-1        | 2LYF_A, AP00445                                      | 11         |
|                                          |           | BTD-2        | 2LYE_A, AP00156                                      | 6          |
|                                          |           | Retrocyclin-2| 2LZI_A, AP01208                                      | 0          |
| Odorrainain M1 (AP01300)                 | 29        | Androctonin  | 1CZ6_A, AP00153                                      | 5          |
| Silkworm 001 (AP01974) \(^{a}\)          | 38        | BTD-2pLR     | 2LYE_A, AP00156                                      | 6          |
|                                          |           | Tachyplesin I| 2M3N_A, Q90WP7                                       | 0          |
|                                          |           |               | 1WO1_A, AP00214                                      | 13         |

Identity percentages were obtained through GESAMT. C1 size %—size of the largest cluster compared to the overall number of models.

\(^{a}\)Data unpublished.
of Cys residues with Tyr, Phe, or Ala maintain the fold as a result of aromatic stacking, but our peptides are not spatially constrained by aromatic interactions either.

In order to assess how unusual the absence of disulfides in an AMP-β-hairpin is, the PDB was mined using the mmCIF Keyword Search (Classification) “antimicrobial” and the hits were scanned visually for structural similarity. Shown in Figures S10D and S10E are 2 “β-hairpin-like” AMPs that were found: entries 2LM8 (Cysteine Deleted analog of β-hairpin AMP Tachyplesin I in LPS, CDT-LPS) and 2MQ4 (R11 peptide bound to LPS). Both AMPs are of synthetic origin and very short, with 13 and 11 residues, respectively. CDT-LPS retains a β-hairpin-like structure due to cation-π interactions between residues W2 and R11, but such interactions are not present in our modeled AMPs. RR11 is an N-terminal truncated variant of Cys deleted protegrin-1, CDP-1. Both CDT-LPS and R11 are largely unstructured when free in aqueous solutions. NMR experiments of CDP-1 bound to LPS showed aromatic interactions. While the structure of RR11 bound to LPS somewhat resembles our β-hairpins, it lacks the characteristic hydrogen bonds to form the overall β-hairpin structure and is almost half the size of our modeled β-hairpins.

Since there are no intramolecular disulfides, cation-π or aromatic interactions to keep these structures stable, we looked for hydrogen bonds in the set of trajectory snapshots on which RMSD clustering was performed. Our Rosetta model of Lebocin Peptide 1A showed two β-strands extending from residues 7 to 10 and 14 to 17. This feature was commonly present during the trajectory. For example, two hydrogen bonds between residues Phe8 and Thr16 were present ~81% and ~63% of the time. Similarly, the Odorranain M1 hairpin has residues 2–4 and 14–16 in the β-strand conformation. Here, two hydrogen bonds between Ala3 and Arg15 were present ~75% and ~53% of the time. Finally, the Silkworm 001 AMP (unpublished; APD identifier 01974) Rosetta model showed two β-strands extending from residues 8 to 11 and from 14 to 17. Two hydrogen bonds between Ile9 and Tyr16 were present ~95% and ~86% of the time, and two hydrogen bonds between Ala11 and Tyr14 were present ~76% and ~74% of the time. Together, these findings suggest that LP1A and Odorranain-M1 are the first β-hairpin folds lacking intramolecular disulfide bonds, cation-π or aromatic interactions reported so far, stable only due to hydrogen bonding.

Although the β-fold and charge both appear to be key determinants of antimicrobial activity, their relative importance is not clear. For example, Mohanram and Bhattacharjya associate the loss of charge associated with truncating the RGGR N-terminus in RR11 with the attenuation of activity rather than the loss of β-strands. However, Mani et al find for Protegrin-1 mutants that the β-hairpin fold is more important for activity than the cationicity.

Experimental structures of CDT-LPS and R11 peptides both have significant positive charge spatially arranged on one face of the molecule (Figures S11A and S11B). However, this feature was not seen in our modeled peptides. Silkworm 001 has a single negative charge at its C-terminus tail, and the overall charge of the molecule is zero. Odorranain M1 and Lebocin 1A are positively charged but, in contrast to the crystal structures, cluster positively charged residues (4 or 3 or 5, respectively) at one end, toward the bottom in the orientation shown in Figures S11C and S11D. However, it must be noted that these regions are very flexible and are therefore less reliably predicted.
Finally, Panteleev et al.\textsuperscript{89} report that among several insect AMPs, only 1 was found to adopt a $\beta$-hairpin conformation. This was Thana-tin, derived from the spined soldier bug \textit{Podisus maculiventris}. Here, we predict that both insect peptides, Lebocin Peptide 1A (LP1A) and Silk-worm 001, are also $\beta$-hairpins.

### 3.2.3 $\alpha\beta\beta$ and $\beta\alpha\beta$ folds

This group contains three confidently modeled AMPs. Human Histatin 2\textsuperscript{104} and Garvicin A\textsuperscript{73} were predicted to have $\alpha\beta\beta$ folds, while for Rattusin,\textsuperscript{74} a $\beta\alpha\beta$ fold was predicted. The first largest cluster for Human Histatin 2 peptide contained 35.4\% of the total number of models but fold matches were found for the centroid structures of the second and the third largest clusters only. Nevertheless, the centroid structure of the largest cluster is highly similar to that of the second, and since the third cluster centroid has particularly pronounced secondary structure elements, this model was taken forward as a representative structure (Figure S4). Garvicin A modeling resulted in a small top cluster relative to the number of successfully modeled structures, and Rattusin fold was on the lower limit of 25\% (Table 2, see Methods). Upon inspection of centroids of the remaining two largest clusters for Garvicin A and Rattusin, we found that these models displayed similar structures, confirming their reliability. In addition, we tested the stability of our modeled Rattusin fold by running 100 ns Molecular Dynamics simulations in explicit solvent and performing clustering as described in Methods section. Structure representatives of highly populated clusters superimposed are shown in Figure S12.

Human Histatin 2, similar to its fold matches Termicin (PDB code 1MM0) and AgaDef (2NY8) (Figure 5A),\textsuperscript{52} shows antifungal activity. However, unlike the modeled peptide, its fold matches have three disulfide bonds: starting from the N-terminus, the first one is formed between the helix and first $\beta$-strand Cys residues, and the second and the third disulfide bonds are formed between the helix and second $\beta$-strand Cys residues.

The second modeled structure in this group is Rattusin, an $\alpha$-defensin-related peptide with 5 cysteines. The peptide was modeled without disulfide constraints as there was no consensus on connectivity (see Methods). Here, we predict a $\beta\alpha\beta$ fold which matches the $\alpha\beta\beta$ folds such as found in Palicourein (1R1F) and Cycloviolacin (1NBJ), AMPs with circular backbones, as well as Charybdotoxin (2CRD), a gamma-motif peptide targeting K\textsubscript{1} channels (Figure 5B).

Disulfide bonds formed by Rattusin have been a subject of discussion. Whilst the newest NMR study by Min et al.\textsuperscript{60} shows refolded Rattusin forming homodimers, each protomer being a $\beta$-hairpin, stabilized by five intermolecular disulfide bonds, Patil et al.\textsuperscript{74} suggested that an odd number of cysteines may result in a fraction of Rattusin peptide forming dimers or multimers, which can

### TABLE 2 Fold match results for $\alpha\beta\beta$ and $\beta\alpha\beta$ AMPs

| Modeled AMP and the corresponding APD2 or UniProt IDs | C1 size % | Fold matches | Fold matches' corresponding PDB and APD2/UniProt IDs | Identity % |
|-----------------------------------------------------|-----------|--------------|-------------------------------------------------------|------------|
| Human Histatin 2 (AP00799)                           | 8\textsuperscript{a} | Mytilin B    | 2EEM_A. AP00333                                       | 8          |
|                                                     |           | Termicin     | 1MM0_A. AP00403                                       | 4          |
|                                                     |           | AgaDef       | 2NY8_A. AP01363                                       | 0          |
| Rattusin (AP02178)\textsuperscript{b}              | 25        | Palicourein  | 1R1F_A. AP01034                                       | 0          |
|                                                     |           | Charybdotoxin| 2CRD_A. AP00437                                       | 11         |
|                                                     |           | Cycloviolacin| 1NBJ_A. AP01035                                       | 14         |
| Garvicin A (AP02402)                                 | 17        | Mytilin B    | 2EEM_A. AP00333                                       | 6          |
|                                                     |           | Termicin     | 1MM0_A. AP00403                                       | 0          |
|                                                     |           | Eurocin      | 2LT8_A. AP02119                                       | 9          |

Identity percentages were obtained through GESAMT. C1 size %—size of the largest cluster compared to the overall number of models.

C3 size %.

\textsuperscript{b}Modeled structure matching an experimental structure with different topology.

\textsuperscript{556} KOZIC ET AL.

FIGURE 5 Rosetta models aligned with their fold matches. A, Human Histatin 2 peptide (green) aligned with Mytilin B (PDB code 2EEM, shown in cyan), Termicin (1MM0, magenta) and AgaDef peptide (2NY8, purple), B, Rattusin (green) aligned with Palicourein (PDB code 1R1F, shown in cyan), Charybdotoxin (2CRD, magenta), and Cycloviolacin (1NBJ, purple) peptides, and C, Garvicin A (green) aligned with Mytilin B (2EEM, cyan), Termicin (1MM0, magenta), and Eurocin (2LT8, purple) peptide. Disulfide bonds are shown as sticks in blue, pink, and purple, respectively. Alignments were made in GESAMT [Color figure can be viewed at wileyonlinelibrary.com]
in turn enhance its antibacterial effects. However, the peptide remains active even with the disulfides reduced, justifying a prediction of monomeric structure as performed here. It is well known that disulfide reduction can have a wide spectrum of effects from enhancement to total loss of activity (as seen in \( \alpha \) and \( \beta \)-defensins, and \( \theta \)-defensins, respectively), alteration of selectivity (as seen in cryptidin-related sequences) and so forth.\(^{46,74,91,105}\)

Class II bacteriocins are a heterogeneous group of small heat-stable non-lanthionine containing peptides. One of the four modeled \( \alpha\beta \) folds with positive fold matches in the PDB was Garvicin A\(^73\) isolated from Lactococcus garvieae, a class II bacteriocin. Although modeling produced a comparatively small largest cluster, the largest four clusters contained the same overall fold and exhibited pairwise RMSD on C\( \alpha \) atoms of 2.5 Å at most (Figure S5). These considerations suggest a reliably predicted fold. Garvicin A shares a similar fold with several cysteine-rich proteins and defensins, namely Mytilin B (PDB code 2EEM), Termcin (1MMM), Eurocin (2LT8), DEF-AAA (2NY8), Micasin (2LRS), and MGD-1 (1FJN) (Figure 5C).

Compared to Garvicin A, which is active only against L. garvieae strains, its fold matches are active against a wide spectrum of Gram-positive, Gram-negative bacteria and fungi. Interestingly, while Garvicin A has no cysteine residues, its matches are cysteine-rich peptides with 3 or 4 disulfide bonds stabilizing the structures. We mined the PDB with mmCIF Keyword Search (Classification) “antimicrobial” to see whether there was an \( \alpha\beta \) fold entry that didn’t contain disulfides. Out of 339 structures, there was only one \( \alpha\beta \) AMP without disulfide bonds, namely Subtilisin A (1PXQ). This structure has a cyclized peptide backbone (with an amide between the N- and C-termini) and D-amino acids which have unusual sulfur to C\( \alpha \) cross-links. This suggests that Garvicin A and Human Histatin 2 are the first linear \( \alpha\beta \) fold AMPs lacking disulfide bonds.

### 3.2.4 Helix-break-helix and helix-kink-helix continuum

By far the largest group of folds obtained was a continuum of v-shaped, helix-turn-helix and helix hairpins (Figure 6 and Table 3). Attempts to consistently subdivide these into separate helix-break-helix (containing helix-hairpins and v-shaped structures) and helix-kink-helix groups (Figure 6), were unsuccessful.

![FIGURE 6](image_url) Simple helical folds. A, Helix-hairpins Ceratotoxin D and its fold match Thurincin H (PDB code 2LB2). B, Helix-kink-helix Hymenochirin 2B and its v-shaped helix-break-helix fold match Latarcin 2a (PDB code 2G9P). Modeled structures are shown in green and the corresponding fold matches are shown in cyan [Color figure can be viewed at wileyonlinelibrary.com]

For all of the structures in this group, the largest cluster contained high percentages of the total number of models, except for Andropin and Hymenochirin 2B. For these two peptides, we inspected the centroid structures of the remaining two largest clusters and found them to be similar to the largest cluster’s centroid, which gave us confidence that all folds are modeled correctly. The stability of selected folds was tested by running 100 ns simulations in explicit solvent and performing clustering as described in Methods. Structure representatives of highly populated clusters superimposed are shown in Figure 7.

All of the peptides were modeled as monomers, including Cythaurin peptide, which contains a single cysteine residue.\(^{106}\) Cythaurin is believed to predominantly form homodimers: however, both monomer and dimer are active against bacteria, whilst the monomer is non-hemolytic. Cythaurin, along with Ponerincin Q42\(^{107}\), Ceratotoxin D,\(^{108}\) and CPF-B1 peptides,\(^{109}\) shown in Figures 7A, 7D, and 7E, respectively, has a helix-hairpin fold. Although the orientation of helices fluctuated somewhat, they were stable throughout the simulation (Figure 7). Latarcin 4a\(^{110}\) (Figure 7B) showed 2 stable conformations, a single \( \alpha \)-helix and a helix-break-helix fold. Interestingly, the short N-terminus helix of Kassinaturin 1 peptide\(^{111}\) (Figure 7C) rapidly unfolded in solution, while the C-terminal helix remained stable.

Helix-hairpins are formed by two antiparallel \( \alpha \)-helices connected by a loop (Figure 6A). The helices interact through hydrophobic side-chain interactions at the interface. V-shaped structures can be defined as either: (1) two non-parallel \( \alpha \)-helices whose angles intersect at angles from around 45° to 120° connected with a loop region (Figure 6B, bottom panel) or (2) a helical structure extending throughout the peptide but containing a kink. Usually, cationic peptides longer than ~20 amino acids contain a flexible hinge in the middle part promoted by helix-breaking residues such as glycine and/or proline.\(^{112}\)
| Modeled AMP and the corresponding APD2 or UniProt IDs | C1 size % | Fold matches | Fold matches' corresponding PDB and APD2/UniProt IDs | Identity % |
|------------------------------------------------------|-----------|--------------|-------------------------------------------------|------------|
| Citropin 2.1.3. (AP00639)                            | 46        | EcAMP1       | 2L2R_A, AP01760                                  | 12         |
|                                                      |           | Thurincin CD | 2L9X_A, AP01570                                  | 8          |
|                                                      |           | Sublancin 168| 2MIJ_A, AP01606                                 | 4          |
|                                                      |           | Maximin 4    | 2MHW_A, AP00061                                 | 5          |
| Grammistin Gs A (P69845)                            | 52        | Maximin 4    | 2MHW_A, AP00061                                 | 12         |
|                                                      |           | Ltc2a        | 2G9P_A, AP01011                                 | 4          |
| Andropin (Q8WSV4)                                   | 17        | EcAMP1       | 2L2R_A, AP01760                                  | 0          |
|                                                      |           | Sublancin 168| 2MIJ_A, AP01606                                 | 4          |
|                                                      |           | Thurincin CD | 2L9X_A, AP01570                                  | 4          |
|                                                      |           | Sublancin 168| 2L8_A, AP02196                                  | 3          |
| Casocidin I (P02663)                                | 27        | Thurincin H  | 2LBZ_A, AP02394                                 | 7          |
|                                                      |           | EcAMP1       | 2L2R_A, AP01760                                  | 0          |
| Dermaseptin 8 (P84928)                              | 65        | Thurincin CD | 2L9X_A, AP01570                                  | 15         |
|                                                      |           | EcAMP1       | 2L2R_A, AP01760                                  | 0          |
| Ctri10033 (P0DME4)                                  | 69        | Ltc2a        | 2G9P_A, AP01011                                 | 0          |
|                                                      |           | Maximin 4    | 2MHW_A, AP00061                                 | 10         |
|                                                      |           | Fowlcidin 1  | 2AMN_A, AP00557                                 | 13         |
| Ponericin Q42 (AP02435)                             | 29        | Thurincin CD | 2L9X_A, AP01570                                  | 4          |
|                                                      |           | EcAMP1       | 2L2R_A, AP01760                                  | 4          |
|                                                      |           | Maximin 4    | 2MHW_A, AP00061                                 | 15         |
| Latarcin 4a (AP01014)                               | 56        | Ltc2a        | 2G9P_A, AP01011                                 | 22         |
|                                                      |           | Fowlcidin 1  | 2AMN_A, AP00557                                 | 8          |
|                                                      |           | Maximin 4    | 2MHW_A, AP00061                                 | 6          |
| RV 23 (AP01264)                                     | 78        | Ltc2a        | 2G9P_A, AP01011                                 | 20         |
|                                                      |           | Maximin 4    | 2MHW_A, AP00061                                 | 11         |
| Lycotoxin I (AP00516)                               | 91        | EcAMP1       | 2L2R_A, AP01760                                  | 4          |
|                                                      |           | Ltc2a        | 2G9P_A, AP01011                                 | 20         |
|                                                      |           | Maximin 4    | 2MHW_A, AP00061                                 | 14         |
| Kassinatuerin 1 (AP00556)                           | 64        | Maximin 4    | 2MHW_A, AP00061                                 | 6          |
|                                                      |           | Thurincin H  | 2LBZ_A, AP02394                                 | 0          |
| Ceratotoxin D (AP00417)                             | 65        | Thurincin H  | 2LBZ_A, AP02394                                 | 0          |
|                                                      |           | Sublancin 168| 2MIJ_A, AP01606                                 | 0          |
|                                                      |           | Thurincin CD | 2L9X_A, AP01570                                  | 8          |
|                                                      |           | EcAMP1       | 2L2R_A, AP01760                                  | 3          |
| Pilosulin 5 monomer (AP00893)                       | 88        | Thurincin H  | 2LBZ_A, AP02394                                 | 4          |
|                                                      |           | Sublancin 168| 2MIJ_A, AP01606                                 | 4          |
|                                                      |           | EcAMP1hIAPP  | 2L2R_A, AP01760                                  | 3          |
|                                                      |           | Ltc2a        | 2L8_A, AP02196                                  | 3          |
|                                                      |           | Fowlcidin 1hIAPP | 1XC0, AP00644                              | 3          |
| Hymenochirin 2B (AP01965)                           | 24        | Ltc2a        | 2G9P_A, AP01011                                 | 33         |
|                                                      |           | Fowlcidin 1hIAPP | 2AMN_A, AP00557               | 19         |
|                                                      |           | Pardaxin P4  | 2L8_A, AP02196                                  | 8          |
|                                                      |           | Ltc2a        | 1XC0, AP00644                                  | 27         |
| Clavaspirin (AP00502)                               | 65        | Fowlcidin 1  | 2AMN_A, AP00557                                 | 20         |
|                                                      |           | Maximin 4    | 2MHW_A, AP00061                                 | 15         |
|                                                      |           | Ltc2a        | 2G9P_A, AP01011                                 | 18         |
| Cryptonin (AP00722)                                 | 87        | Maximin 4    | 2MHW_A, AP00061                                 | 17         |
|                                                      |           | Ltc2a        | 2G9P_A, AP01011                                 | 25         |
|                                                      |           | Fowlcidin 1  | 2AMN_A, AP00557                                 | 10         |
| Ocellatin 4 (AP00894)                               | 98        | Maximin 4    | 2MHW_A, AP00061                                 | 15         |
|                                                      |           | Thurincin CD | 2L9X_A, AP01570                                  | 11         |
|                                                      |           | Ltc2a        | 2G9P_A, AP01011                                 | 6          |
|                                                      |           | Fowlcidin 1  | 2AMN_A, AP00557                                 | 20         |
| Grammistin Pp 3 (P69847)                            | 55        | Ltc2a        | 2G9P_A, AP01011                                 | 14         |
|                                                      |           | Maximin 4    | 2MHW_A, AP00061                                 | 13         |
|                                                      |           | Fowlcidin 1  | 2AMN_A, AP00557                                 | 21         |
|                                                      |           | Pardaxin P4  | 1XC0, AP00644                                  | 9          |

(Continues)
It has been speculated that v-shaped helix-break-helix structures with strongly amphiphilic α-helix at the N-terminus only are likely to be functional through the carpet mechanism, while structures with N- and C- termini that are both strongly amphiphilic are more likely to act via the pore-forming mechanism. For the first set, helix-break-helix AMPs with a hydrophobic gradient spanning from N- to C-terminus, it has been suggested that the amphipathic N-terminal helix is responsible for interaction with the membrane, while the C terminus, because of its lack of amphipathicity, lies on the membrane and only has a minimal interaction with it.112 For this particular type of structures, Dubovskii et al.112 proposed molecular hydrophobic potential (MHP) plots to be effective in sorting structures by mechanism. However, due to a lack of experimental data on structure and mechanisms of action of helix-kink-helix peptides, this method has not been particularly useful in gaining a clear picture for our modeled peptides and, for now, the structure-function relationship for these AMPs remains unclear.

3.3 | Potential novel AMP folds

The results discussed hitherto relate to fold matches between the AMPs we modeled and others of already determined structures. This revealed cases of likely cryptic homology, such as between Lacticin Q87 and Enterocins 7A and 7B,77 as well examples of recurrent simple folds that are more likely to be examples of structural analogy: a β-hairpin Lebocin Peptide 1A from a moth Manduca sexta88 and θ-defensins BTD-2, RTD-2, and Recrocycin-2 from baboon90 would be likely examples. However, it is unlikely that currently determined AMP structures sample all naturally existing AMP fold space. We therefore examined results that appeared reliable, that is, with a large cluster of similar models and which represented well-packed protein structures, including considering whether they resembled any known folds (not just AMPs) in a topology-independent fashion that could be seen using the CLICK structure database search algorithm.64 Here they are dealt with according to broad fold class and their specificities compared to those of AMPs with similar folds.

### 3.3.1 | All-β folds

A total of three all-β folds modeled (Figure 8) were scanned for fold matches using CLICK database search. The first, Cypemycin113 (Figure 8A) is a bacteriocin active against Gram-positive bacteria. Although it has no lanthionine bridges present, it has some of the structural features of lantibiotics such as dehydrated amino acids, two L-allo-isoleucines, and a modified C-terminal D-cysteine that forms a ring structure with an L-cysteine. Only the disulfide bond constraint, set for the two modeled isoleucines (which were modeled as the closest available representation of L-allo-isoleucine) are solvent exposed, and the D-cysteine is the
terminal residue, we expect these modifications to have limited effect on the overall tertiary structure. The largest cluster of Cypemycin peptide contained 49% of the overall number of modeled structures. The centroid structure (Figure 8A) matched other β-hairpin AMPs such as Protegrin-2 (PDB code 2MUH), Tachyplein I (1WO1, and wild-type 1MA2), ε-defensin HTD-2 (2LZI), Thanatin (8TFV) and Polyphemusin I peptide (1RRK), which are all, with the exception of antiviral HTD-2, active against both Gram-positives and Gram-negatives.52 Due to strict thresholds upon filtering GESAMT results, β-hairpin top results such as Tachyplein I (1WO1), Cyclic L27-11 (2M7I) and Retrocyclin-2 (2LZI) were filtered out and so were not considered with the other β-hairpin matches above.

Scolopendin 1114 (Figure 8B) was found to have a fold similar to bovine neutrophil β-defensin 12 (PDB code 1BNB), chicken AvBD2 defensin (2L5G), human α-defensin HNP1 (3GNY), as well as Spheniscin-2 (1UT3) and Tricyclon A (1YP8). The largest cluster contained 24.2% of the overall number of models. Scolopendin 1 is active against a wide spectrum of Gram-positive and Gram-negative bacteria, as well as fungi, and in Candida albicans it was shown to induce reactive oxygen species (ROS) accumulation. With the exception of antiviral, negatively charged Tricyclon A, all of the fold matches exhibit activity against Gram-positives and Gram-negatives.

Microcin M115,116 (Figure 8C) is a class Ibb microcin, which contains a siderophore moiety C-terminal post-translational modification, that enables the uptake of peptide but was not included in our modeling. The largest cluster contained 15.5% of the overall number of modeled structures. It predicts Microcin M to contain a four-stranded β-sheet but no fold matches with known AMP structures or structures from the PDB90 were found.

### 3.3.2 Mixed αβ folds

Upon performing CLICK database search, five mixed αβ folds were found to have fold matches that met the size and Z-score criteria (see Methods). The first, Propionicin-F117 (Figure 9A) is an unmodified, heat-stable, negatively charged bacteriocin with a ββα fold. Although only 17.9% successfully modeled structures were clustered into the largest cluster, this peptide matched many αβ folds, such as mussel Mytilin (PDB code 2EEM), Plectasin (12FU), Termicin (1MMO), Cg-Def defensin (2B68), Micasin (2LRS), and Eurocin (2L8T). Interestingly, only a single βα fold was found among the filtered results, namely Leuco- cin A (1CW6).52 However, the short β-sheet does not pack against the helix in the same way as seen in our Rosetta model, leading to a much more elongated structure. In contrast, the topologically-distinct αβ folds share a similar mode of interaction between helix and β-sheet as Propionicin-F suggesting that they may be structural analogues. Like many bacteriocins, Propionicin-F has a nanomolar activity against strains of its producer organism, Propionibacterium freudenreichii.

ABP-118a118 shown in Figure 9B is a type Ibb, unmodified, heat-stable bacteriocin, active against both Gram-positives and Gram-negatives. These bacteriocins are comprised of two peptide chains, and the overall activity is obtained by the complementary activity of the two peptides. Interestingly, the α component inhibits bacterial growth on its own, and while the β component has no activity, it complements the activity of the α chain.118 Here, we modeled the α chain, with disulfide bond connectivity predicted to be Cys6-Cys44. 26% of successfully modeled structures were clustered in the top cluster. Its predicted αβ fold matches several AMPs, all containing multiple disulfide bonds: Termicin (1MMO), Beta-purothionin (1BHP) and Viscoctin A3 (1EDO). The latter two peptides show activity against Gram-positives, Gram-negatives and fungi, while Termicin is active against Gram-positive bacteria and fungi.

Mutacin IV119 (Figure 9C) shows activity against the Mitis group of oral streptococci. The peptide was thought to be a type Ibb bacteriocin,119 but this has not been demonstrated unambiguously.120 We have found the A chain of Mutacin IV to match Mytilin (PDB code 2EEM) and Termicin (1MMO), as well as the Mediterranean mussel defensin MGD-1 (1FJN) AMP. The disulfide connectivity for Mutacin IV A was predicted to be Cys14-Cys34 and largest cluster contained 20.2% of the successfully modeled structures.

This group also contains more complex folds, such as chain A Blp1 peptide—another type Ibb bacteriocin (Figure 9D). The highly populated largest cluster (28.1% of models) obtained with disulfide connectivity Cys2-Cys63 suggests a successful fold prediction, but it gave no fold matches with AMPs. A CLICK database search against the PDB90

---

**FIGURE 9** Plausible mixed αβ folds. A, Propionicin-F (Q6E3K9, C1% 17.9); B, ABP-118a (AP01172, C1% 26.0, disulfide bond connectivity was predicted to be Cys6-Cys44). C, Mutacin IV chain A (AP01174, C1% 20.2, disulfide bond connectivity was predicted to be Cys14-Cys34), D, Blp1 chain A (AP01995, C1% 28.1, disulfide bond connectivity was predicted to be Cys2-Cys63), and E, Ipomisin (AP01604, C1% 8.48, disulfide bond connectivity was predicted to be Cys32-Cys63). Structures were color-coded from N-terminus (blue) to C-terminus (red). Disulfide bonds are shown in magenta. C1%—size of the largest cluster compared to the overall number of models [Color figure can be viewed at wileyonlinelibrary.com]
database revealed similarity with Pseudopilin GspI, a chain from enterotoxigenic Escherichia coli secretion system (PDB code 3CI0, chain I), supporting the biophysical plausibility of our modeled fold. However, the two are likely to have evolved independently: Pseudopilins are part of the type 2 secretion system found in Gram-negative bacteria, and GspI is known to be located at the pseudopilus base, interacting with the inner membrane components, while Blp1 AMP is produced by the human oral strain Lactobacillus salivarius and is active against Gram-positive bacteria.

Our most complex fold, Ipomicin peptide (Figure 9E), is active against strains of the producer organism, Streptomyces ipomoeae. With its six secondary structure elements it did not match any known AMP folds. Although only 8.48% of successfully modeled structures were clustered in the top cluster, results of CLICK database search ran on structures from the PDB90 database revealed striking similarity to 2 known folds: BRICHOS domain of Lung Surfactant Protein C (2YAD) and an Uracil-DNA glycosylase protein (1UGI). Both of these are topologically different from Ipomicin, yet the secondary structure elements are positioned in a similar manner (Figure 10). We suggest that this packing is favorable, irrespective of the topological connection of the secondary structure elements, supporting the reliability of the Ipomicin fold prediction. This in turn suggests that the Ipomicin fold prediction adds a new and relatively complex architecture to the array of folds sampled by AMPs.

### 3.3.3 | All-\(\alpha\) folds

Overall ten all-\(\alpha\) folds modeled (Figure 11) were scanned for fold matches using the CLICK database search, and only one was found to have fold matches, namely Vejovine peptide (Figure 11I). The largest cluster contained 18.9% of the total number of models, and the top three clusters all showed similar structure, supporting the reliability of the model.

Vejovine peptide, found in scorpion venom of Vaejovis mexicanus, is active against Gram-negative bacteria and hemolytic to human erythrocytes. The peptide is helical in 60% TFE, and the first 8 residues at N-terminus are crucial for its activity. Our Rosetta model suggests a structure containing four helices spanning residues 2–14, 17–38, 42–47, and 50–58, respectively. Neither GESAMT nor CLICK database searches against characterized AMPs returned results. However, an acyl carrier protein fold from Thermus thermophilus, with similar topology, was found to match Vejovine peptide through CLICK database search against PDB90 (Figure 12B).
CONCLUSIONS

We have modeled the structures of a large set of previously structurally uncharacterized AMPs using a variety of computational methods to ensure robustness. The set was obtained from the APD2 database and a literature search, and contained 184 sequences of between 20 and 120 residues. The validity of the ab initio modeling was supported by both benchmarking against known AMP structures and fold stability testing using Molecular Dynamics experiments.

The newly mapped structural landscape reveals AMPs with similarities to existing folds across different classes, and also predicts new folds for several AMPs. So, for example, familiar α-helical folds were predicted for Ponericin Q42, Latarcin 4a, Kassinaturein 1, Ceratotoxin D, and CPF-B1 peptide and β-hairpins for Lebocin Peptide 1A (LP1A), Odorrana in M1, and Silkworm 001. Interestingly, LP1A and amphibian Odorrana in M1 are the first β-hairpin folds lacking intramolecular disulfide bonds, cation–π or aromatic interactions reported so far, stable only due to hydrogen bonding. Until now, tigerinins, which adopt a beta-turn fold due to a disulfide bridge between two cysteine residues forming a nonapeptide ring, were the only examples of non-helical amphibian antimicrobial peptides. Moreover, a single insect-derived β-hairpin AMP has been reported so far, and we predicted that both insect peptides, Lebocin Peptide 1A (LP1A), and Silkworm 001, are β-hairpins. Examples of mixed αβ folds are Garvicin A and Human Histatin 2, which contain no disulphide bonds, usually found in such AMPs as part of the cysteine-stabilized αβ (CSαβ) motif, making the modeled peptides the first linear αβ fold AMPs lacking intramolecular disulfide bonds. Novel folds were predicted for Microcin and Ipomicin, the latter resembling the BRICHOs domain of Lung Surfactant Protein C, but with different topology between the secondary structure elements. These findings expand our knowledge of the AMP structural universe and may contribute to the structure-based development of more potent AMPs.

ACKNOWLEDGMENT

We acknowledge BMSI (A*STAR) and National Supercomputing Center Singapore for computational support.

ORCID

Daniel J. Rigden http://orcid.org/0000-0002-7565-8937

REFERENCES

[1] World Health Organization. Antimicrobial Resistance: Global Report on Surveillance [Internet]. World Health Organization; 2014. http://apps.who.int/iris/bitstream/10665/112642/1/9789241564748_eng.pdf?ua=1[TQ]
[2] Centers for Disease Control and Prevention. Antibiotic Resistance Threats in the United States, 2013 [Internet]. Centers for Disease Control and Prevention; 2013. https://www.cdc.gov/drugresistance/threat-report-2013/pdf/ar-threats-2013-508.pdf
[3] Narayana JL, Chen J-Y. Antimicrobial peptides: possible anti-infective agents. Peptides. 2015;72:88–94.
[4] Moreno-Habel DA, Biglang-Awa IM, Dulce A, Luu DD, Garcia P, Weers PMM. Inactivation of the budded virus of Autographa californica M nucleopolyhedrovirus by gloverin. J Invertebr Pathol. 2012;110(1):92–101.
[5] Sang Y, Blecha F. Antimicrobial peptides and bacteriocins: alternatives to traditional antibiotics. Anim Health Res Rev. 2008;9(02):227–235.
[6] Teixeira V, Feio MJ, Bastos M. Role of lipids in the interaction of antimicrobial peptides with membranes. Prog Lipid Res. 2012;51(2):149–177.
[7] Li J, Koh J-J, Liu S, Lakshminarayanan R, Verma CS, Beuerman RW. Membrane active antimicrobial peptides: translating mechanistic insights to design. Front Neurosci. 2017;11:73
[8] Yau WM, Winley WC, Gawrisch K, White SH. The preference of tryptophan for membrane interfaces. Biochemistry. 1998;37(42):14713–14718.
[9] Vogel HJ, Schibli DJ, Jing W, Lohmeier-Vogel EM, Epand RF, Epand RM. Towards a structure-function analysis of bovine lactoferrin and related tryptophan- and arginine-containing peptides. Biochem Cell Biol. 2002;80(1):49–63.
[10] Fleury Y, Dayem MA, Montagne JJ, et al. Covalent structure, synthesis, and structure-function studies of mesentericin Y 105(37), a defensive peptide from gram-positive bacteria Leuconostoc mesenteroides. J Biol Chem. 1996;271(24):14421–14429.
[11] Shepherd CM, Vogel HJ, Tieleman DP. Interactions of the designed antimicrobial peptide MB21 and truncated dermaseptin S3 with lipid bilayers: molecular-dynamics simulations. Biochem J. 2003;370(1):233–243.
[12] Aliste MP, MacCallum JL, Peter Tieleman D. Molecular dynamics simulations of pentapeptides at interfaces: salt bridge and cation–π interactions. Biochemistry. 2003;42(30):8976–8987.
[13] Petersen FNR, Jensen MØ, Nielsen CH. Interfacial tryptophan residues: a role for the cation–π effect? Biophys J. 2005;89(6):3985–3996.
[14] Gleason NJ, Vostrikov VV, Greathouse DV, Grant CV, Opella SJ, Koepe RE. 2nd. Tyrosine replacing tryptophan as an anchor in GWALP peptides. Biochemistry. 2012;51(10):2044–2053.
[15] Sparks KA, Gleason NJ, Gist R, Langston R, Greathouse DV, Koepe RE. 2nd., Comparisons of interfacial Phe, Tyr, and Trp residues as determinants of orientation and dynamics for GWALP transmembrane peptides. Biochemistry. 2014;53(22):3637–3645.
[16] Chan DI, Prener EJ, Vogel HJ. Tryptophan- and arginine-rich antimicrobial peptides: structures and mechanisms of action. Biochim Biophys Acta. 2006;1758(9):1184–1202.
[17] Bi X, Wang C, Ma L, Sun Y, Shang D. Investigation of the role of tryptophan for membrane interfaces. Biochem J. 2015;468:233–243.
[18] Wang G. Antimicrobial Peptides: Discovery, Design and Novel Therapeutic Strategies. CABI, Wallingford, UK: 2010:230.
[19] Wang G, Li X, Wang Z. APD2: the updated antimicrobial peptide database and its application in peptide design. Nucleic Acids Res. 2009;37(Database issue):D933–D937.
[20] Dathe M, Wieprecht T. Structural features of helical antimicrobial peptides: their potential to modulate activity on model membranes and biological cells. Biochim Biophys Acta. 1999;1462(1–2):71–87.
[21] Marr AK, Gooderham WJ, Hancock RE. Antibacterial peptides for therapeutic use: obstacles and realistic outlook. Curr Opin Pharmacol. 2006;6(5):468–472.
[22] Bai Y, Liu S, Li J, et al. Progressive structuring of a branched antimicrobial peptide on the path to the inner membrane target. J Biol Chem. 2012;287(32):26606–26617.
[23] Sani M-A, Separovic F. How membrane-active peptides get into lipid membranes. Acc Chem Res. 2016;49(6):1130–1138.
lipopolysaccharide micelle: Mechanistic insights into outer-membrane permeabilization and endotoxin neutralization. Biochim Biophys Acta. 2012;1818(7):1613–1624.

[102] Mohanram H, Bhattacharya S. Cysteine deleted protegrin-1 (CDP-1): anti-bacterial activity, outer-membrane disruption and selectivity. Biochim Biophys Acta. 2014;1840(10):3006–3016.

[103] Mani R, Waring AJ, Lehrer RI, Hong M. Membrane-disruptive abilities of beta-hairpin antimicrobial peptides correlate with formation and activity: a 31P and 1H NMR study. Biochim Biophys Acta. 2005;1716(1):11–18.

[104] Xu L, Lai K, Pollock JJ. Histatins 2 and 4 are autoproteolytic degradation products of human parotid saliva. Oral Microbiol Immunol. 1992;7(2):127–128.

[105] Wu Z, Hoover DM, Yang D, et al. Engineering disulfide bridges to dissect antimicrobial and chemotactic activities of human beta-defensin 3. Proc Natl Acad Sci U S A. 2003;100(15):8880–8885.

[106] Lee IH, Lee YS, Kim CH, et al. Dicynthaurin: an antimicrobial peptide from hemocytes of the solitary tunicate, Halocynthia auran-tide from Hemocytes of the Solitary Tunicate, Halocynthia auran-tide. Biochim Biophys Acta. 2001;1527(3):141–148.

[107] Pluzhnikov KA, Kozlov SA, Vassilevski AA, Vorontsova OV, Feofanov AV. Linear antimicrobial peptides from Ectotomma quadridens ant venom. Biochimie. 2014;107(Pt B):211–215.

[108] Rosetto M, De Filippis T, Manetti AG, Marchini D, Baldari CT, Dal-la R. The genes encoding the antibacterial sex-specific peptides ceratotoxins are clustered in the genome of the medfly Ceratitis capitata. Insect Biochem Mol Biol. 1997;27(12):1039–1046.

[109] Mechkarska M, Ahmed E, Coquet L, et al. Antimicrobial peptides with therapeutic potential from skin secretions of the Marsabit clawed frog Xenopus borealis (Pipidae). Comp Biochem Physiol C Toxicol Pharmacol. 2010;152(4):467–472.

[110] Kozlov SA, Vassilevski AA, Feofanov AV, Surovoy AY, Karpunin DV, Grishin EV. Latacrins, antimicrobial and cytolytic peptides from the venom of the spider Lachesana tarabaevi (Zodariidae) that exemplify biomolecular diversity. J Biol Chem. 2006;281(30):20983–20992.

[111] Mattute B, Knoop FC, Conlon JM. Kassinatuerin-1: a peptide with broad-spectrum antimicrobial activity isolated from the skin of the hyperolid frog, Kassina senegalensis. Biochem Biophys Res Commun. 2000;268(2):433–436.

[112] Dubovskii PV, Volynsky PE, Poliansky AA, Chupin VV, Efremov RG, Arseniev AS. Spatial structure and activity mechanism of a novel spider antimicrobial peptide. Biochemistry. 2006;45(35):10759–10767.

[113] Minami Y, Yoshida K-I, Azuma R, et al. Structure of cyphemycin, a new peptide antibiotic. Tetrahedron Lett. 1994;35(43):8001–8004.

[114] Choi H, Hwang J-S, Lee DG. Identification of a novel antimicrobial peptide, sclopendin 1, derived from centipede Scolopendra subspini-pes mutilans and its antifungal mechanism. Insect Mol Biol. 2014;23(6):788–799.

[115] Duquesne S, Destoumieux-Garzón D, Peduzzi J, Rebuffat S. Microcins, gene-encoded antibacterial peptides from enterobacteria. Nat Prod Rep. 2007;24(4):708–734.

[116] Vassiliadis G, Destoumieux-Garzón D, Lombard C, Rebuffat S, Peduzzi J. Isolation and characterization of two members of the siderophore-microcin family, microcins M and H47. Antimicrob Agents Chemother. 2010;54(1):288–297.

[117] Brede DA, Faye T, Johnsborg O, Odegård I, Nes IF, Holø H. Molecular and genetic characterization of propionicin F, a bacteriocin from Propionibacterium freudenreichii. Appl Environ Microbiol. 2004;70(12):7303–7310.

[118] Flynn S, van Sinderen D, Thornton GM, Holø H, Nes IF, Collins JK. Characterization of the genetic locus responsible for the production of ABP-118, a novel bacteriocin produced by the probiotic bacterium Lactobacillus salivarius subsp. salivarius UCC118. Microbiology. 2002;148(4):973–984.

[119] Qi F, Chen P, Caulfield PW. The group I strain of Streptococcus mutans, UA140, produces both the lantibiotic mutacin I and a non-lantibiotic bacteriocin, mutacin IV. Appl Environ Microbiol. 2001;67(1):15–21.

[120] Hale JDF, Ting Y-T, Jack RW, Tagg JR, Heng NCK. Bacteriocin (mutacin) production by Streptococcus mutans genome sequence reference strain UA159: elucidation of the antimicrobial repertoire by genetic dissection. Appl Environ Microbiol. 2005;71(11):7613–7617.

[121] Korotkov KV, Hol WGJ. Structure of the GspK-GspL-GspJ complex from the enterotoxigenic Escherichia coli type 2 secretion system. Nat Struct Mol Biol. 2008;15(5):462–468.

[122] Michel GPF, Voulhoux R. The type II secretory system (T2SS) in Gram-negative bacteria: a molecular nanomachine for secretion of sec and tat-dependent extracellular proteins. In: Wooldridge K, editor. Bacterial Secreted Proteins: Secretory Mechanisms and Role in Pathogenesis. Norfolk, UK: Caister Academic Press; 2009.

[123] Hernández-Aponte CA, Silva-Sanchez J, Quintero-Hernández V, et al. Vejovine, a new antibiotic from the scorpion venom of Vaejo-vis mexicanus. Toxicon. 2011;57(1):84–92.

SUPPORTING INFORMATION
Additional Supporting Information may be found online in the supporting information tab for this article.

How to cite this article: Kozic M, Fox SJ, Thomas JM, Verma CS, Rigden DJ. Large scale ab initio modeling of structurally uncharacterized antimicrobial peptides reveals known and novel folds. Proteins. 2018;86:548–565. https://doi.org/10.1002/prot.25473