

Rapidly evolving aphid gall effector proteins exhibit saposin-like folds

**Fatema Bhinderwala¹, Aishwarya Korgaonkar^{2,3}, Kota Gopalakrishna^{2,4},
Thomas C. Mathers^{5,6}, Shuji Shigenobu⁷, J. Fernando Bazan⁸, Saskia A.
Hogenhout⁵, Guillermo Calero¹, Angela M. Gronenborn^{1,*}, David L.
Stern^{2,3,*}**

1 Department of Structural Biology, 3501 Fifth Avenue, University of Pittsburgh School of Medicine, Pittsburgh, Pennsylvania, 15261 USA

2 Janelia Research Campus, Howard Hughes Medical Institute, 19700 Helix Drive, Ashburn, Virginia 20147 USA

3 Howard Hughes Medical Institute and Stowers Institute for Medical Research, 1000 E 50th St., Kansas City, Missouri 64110 USA

4 Present Address: Meso Scale Diagnostics, 16020 Industrial Drive, Gaithersburg, MD 20877 USA

5 Department of Crop Genetics, John Innes Centre, Norwich, NR4 7UH, Great Britain

6 Tree of Life, Wellcome Sanger Institute, Hinxton, Cambridge, CB10 1SA, Great Britain

7 Laboratory of Evolutionary Genomics, Trans-Scale Biology Center, National Institute for Basic Biology (NIBB), Nishigonaka 38, Myodaiji, Okazaki 444-8585, Japan

8 h bioconsulting llc, Stillwater, MN 55082 USA

* corresponding authors

ORCID

Fatema Bhinderwala: 0000-0002-3033-8438

Aishwarya Korgaonkar: 0009-0004-4244-0140

Kota Gopalakrishna: 0000-0001-9275-7059

Thomas Mathers : 0000-0002-8637-3515

Shuji Shigenobu: 0000-0003-4640-2323

J. Fernando Bazan : 0000-0002-3645-9935

Saskia A. Hogenhout : 0000-0003-1371-5606

Guillermo Calero: 0000-0003-3730-4676

Angela Gronenborn: 0000-0001-9072-3525

David L. Stern: 0000-0002-1847-6483

e-mail

Fatema Bhinderwala: fab47@pitt.edu

Aishwarya Korgaonkar: korgaonkara@hhmi.org

Kota Gopalakrishna: kngopala@gmail.com

Thomas Mathers: tm18@sanger.ac.uk

Shuji Shigenobu: shige@nibb.ac.jp

J. Fernando Bazan: jfbazan@gmail.com

Saskia A. Hogenhout: Saskia.Hogenhout@jic.ac.uk

Guillermo Calero: guc9@pitt.edu

Angela Gronenborn: amg100@pitt.edu

David L. Stern: dstern@stowers.org

Abstract

Many insects manipulate plants by injecting effector proteins. In one extreme example of this molecular “hijacking”, *Hormaphis cornu* aphids inject bicycle proteins into *Hamamelis virginiana* (Witch Hazel), contributing to the development of novel organs called galls. Bicycle proteins share no amino acid sequence similarity with proteins of known function. Here, we report the crystal structures of two divergent bicycle proteins. Both proteins contain saposin-like folds: one with multiple disulfide bonds exhibits a helix swap; the other has no disulfide bonds and possesses two tandem domains. To explore the structural evolution of bicycle proteins, we predicted bicycle protein structures with Alphafold2 (AF2). While AF2 did not recover the two experimental structures using existing databases, it succeeded after we provided multiple sequence alignments (MSAs) containing protein sequences encoded in new genome sequences from closely related aphid species. Using this customized approach at scale, we generated 2400 high-confidence predictions for bicycle proteins from seven aphid species. This dataset revealed that bicycle proteins without cysteines are outliers in fold space and appear to have evolved from ancestral proteins with disulfide-bonded saposin-like folds. While all bicycle proteins contain predicted saposin-like folds, they display a vast diversity of structural and physicochemical properties. While this diversity thwarts prediction of conserved functions encoded in structure, it suggests that bicycle proteins have evolved to target diverse plant processes and/or to evade plant immune surveillance.

Keywords: AlphaFold predictions, Aphid, Plant galls, Effector proteins, Crystal structure, bicycle genes, Evolution, Arms race.

Significance statement

Parasites introduce specialized “effector” proteins into hosts, both to suppress host immunity and to release nutrients. The molecular functions and structures of most effector proteins are unknown. Effector proteins often evolve rapidly and share no similarity with proteins of known function. Here, we demonstrate that machine learning algorithms can accurately predict the structures of aphid “bicycle” effector proteins when supplemented with data from closely related species. We exploit this finding to generate predictions of 2400 bicycle protein structures. These proteins exploit a common motif, yet exhibit diverse structures that form distinct structural clusters. Despite the clustering of these proteins in structure space, they occupy a nearly uniformly physicochemical space, suggesting that they encode a large diversity of molecular functions.

Introduction

Pathogens employ effector proteins to manipulate host physiology, development, and behavior (1–6). In response to these effectors, hosts deploy immune responses. Over evolutionary time, conflict between effectors and immune responses can result in a so-called molecular “arms-race” between the host and pathogen, associated with rapid molecular evolution (7–11). This results in the generation of novel protein families that

share little amino acid sequence similarity with proteins of known function, complicating the inference of their molecular functions from sequence-based comparisons.

Aphids are plant sap-sucking insects that secrete salivary proteins into plant cells to facilitate access to nutrients and inhibit plant defense responses (12). The most extreme examples of plant manipulation by aphids and other insects is the production of novel organs, called galls, which reflect reprogramming of plant cell growth and development (13, 14). Aphids and other insects induce galls to provide nutritious enclosed spaces that protect them from predators, parasites, and environmental vicissitudes (15–17). Bicycle proteins have recently been implicated as aphid-produced effectors that influence gall development (13). *Bicycle* genes have evolved rapidly owing to strong diversifying selection (13), and the genome of the gall-inducing aphid *Hormaphis cornu* encodes more than 650 *bicycle* paralogs. All aphid genomes examined to date encode *bicycle* genes (18), suggesting that many or all aphids inject bicycle proteins while feeding on plants to manipulate plant physiology and development.

Bicycle proteins possess an N-terminal secretion signal sequence, which is likely removed during translocation to the secretory vesicles. A pair of distributed cysteine-tyrosine-cysteine motifs (C-Y-C) is found in most bicycle proteins—giving the family its name (“bi-CYC-like”)—and suggests that these proteins contain two structurally related domains. However, outside the signal sequences, these proteins display no similarity to proteins or domains of known function (13). Here, we report the X-ray structures of two bicycle proteins from *H. cornu*. Both structures include saposin-like folds, but in topologically different arrangements. These experimentally determined structures were

used as positive controls to benchmark AlphaFold2 (AF2) predictions of additional bicycle proteins. Multiple sequence alignments (MSA) of bicycle proteins encoded by newly sequenced genomes of species closely related to *H. cornu* allowed AF2 to predict models that closely resembled the crystal structures, as well as high-confidence models for thousands of bicycle proteins from multiple species. These predicted models exhibit extensive variability in overall amino acid sequence space, especially of surface residues, as well as of physicochemical properties, with no evidence of conserved domains that might define biochemical functions. The extensive diversity of these proteins may be required to allow them to engage with multiple molecular targets in plants and/or to evade plant defense mechanisms.

Results

A bicycle protein with helix swapped saposin-like folds

We selected 18 *bicycle* genes (Table S1) that are highly expressed in *H. cornu* salivary glands based on RNAseq data (13) for heterologous expression. Of these, four proteins were expressed well in *E. coli* and three in *SF9* insect cells and were purified in sufficient quantities for initial crystallization tests. Of these, two proteins—g3873 and g2703—yielded diffraction-quality crystals (Figure S1, Table S1).

The X-ray structure of g3873 was solved at 2.0 Å using sulfur-SAD phasing and revealed an all-helical fold, possessing two sub-domains, each comprising four-helical bundles: helices $\alpha 1$, $\alpha 2$, $\alpha 3$, and $\alpha 4'$ form one bundle, and $\alpha 1'$, $\alpha 2'$, $\alpha 3'$, and $\alpha 4$ form the

second bundle (Figure 1A,C, and S2, Table S2). The first (C30) and second (C191) conserved cysteines of the first and second CYC motif, respectively, form a disulfide bond linking helices $\alpha 1$ and $\alpha 4'$. The second (C92) and first (C120) conserved cysteines of the first and second CYC motif, respectively, form a second disulfide bond linking helices $\alpha 4$ and $\alpha 1'$ (Figure 1A). A third disulfide bond connects C38 in helix $\alpha 1$ and C48 in helix $\alpha 2$, and these cysteines are not strongly conserved across bicycle proteins. The average distance between the sulfur atoms in the C30-C191, C92-C120, and C38-C48 bonds is $2.05 \pm 0.1 \text{ \AA}$, $2.08 \pm 0.1 \text{ \AA}$, and $2.03 \pm 0.05 \text{ \AA}$, respectively. The presence of three disulfide bonds as well as an additional free sulfhydryl group was confirmed by titration using 5,5'-dithiobis-2-nitrobenzoic acid (19, 20) and the sulfur positions were clearly identified in the difference density maps (Figure S3). The two disulfide bonds associated with the CYC motifs form left-handed (LH) spirals, with $\chi 1$ and $\chi 1'$ dihedral angles of 180° and 60° , respectively. The C30-C191 disulfide bond is also an LH spiral, whereas the C92-C120 disulfide bond is a right-handed (RH) staple. Of the 20 different disulfide bond geometries, -LH spirals are the most common and energetically favorable disulfide conformation, with the lowest dihedral strain energy (DSE) (21). The calculated DSE is $9.2 \pm 1.1 \text{ kJ/mol}$ for each of the two LH spirals and $14.9 \pm 1.2 \text{ kJ/mol}$ for the RH staple (Table S3) (22, 23).

The structure of g3873 (Figure 1A, C) is dissimilar from all proteins in the PDB (24), as determined by DALI (25) and FoldSeek (26) (Table S4). Manual inspection revealed, however, that each set of four clustered alpha helices is reminiscent of the saposin fold (27–29). Saposins are conserved proteins that contain a hydrophobic core sandwiched between two pairs of alpha helices, which allows them to bind phospholipids (30). Although the two

four-helical bundles of G3873 resemble the canonical saposin fold, they appear to lack a hydrophobic internal surface that can adopt an alternate open conformation (31) (Figure S2).

A bicycle protein with tandem saposin-like folds

The second bicycle protein we crystallized, g2703, has an amino acid sequence highly divergent from other bicycle proteins (Figure S4). Although g2703 contains no cysteines, and was not recognized as a bicycle protein in our initial study (13), it has a similar intron-exon structure to other *bicycle* genes and was identified as a member of the bicycle gene family using a gene structure classifier (18). This protein has an N-terminal secretion signal and is highly expressed in the salivary glands of *H. cornu* (Figure S5), supporting the inference that it serves as an effector protein. Thus, while *bicycle* genes were named originally based on the presence of conserved C-Y-C motifs, the sequence of g2703 suggests that at least some *bicycle* genes do not fit neatly within this sequence-based categorization.

The X-ray structure of the g2703 protein (Figure 1B, D) was solved at 1.4 Å using selenomethionine-labeled protein (Figure S1) and multiwavelength anomalous diffraction (MAD) phasing. The structure of g2703 contains two saposin-like domains in tandem (Figure 1D), connected by a long alpha helix. This long central helix appears to have evolved by a merger of the last helix of the first saposin-like unit and the first helix of the second unit (Figure S4). Similar to g3873, neither of the two saposin-like folds exhibits strong amphipathic properties that might facilitate protein-lipid interaction (32).

The saposin-like domains of g3873 and g2703 share limited structural similarity (33, 34) with the 52 saposin structures deposited in the PDB (24) (Figure S6, S7, and S8, Table S5 and S6). Most similar to g3873 is the saposin domain of a proteolytically processed human acyloxyacyl hydroxylase saposin (35) (PDB ID: 5W78, TM_score = 0.44 Å, backbone RMSD for 68 residues = 3.65Å) (Figure S6). The N-terminal saposin-like domain of g2703 showed modest similarity to the saposin domain of prophytepsin (36) from barley, characterized as a vacuolar aspartic proteinase (PDB ID: 1QDM, TM_score = 0.495, backbone RMSD for 71 residues = 4.03 Å) (Figure S7).

The presence of saposin-like domains in the crystal structures of both g3873 and g2703 suggests that they evolved from a common ancestral protein (see also Figure S4), despite the lack of sequence similarity between these two proteins. This observation provides independent support for the hypothesis that the bicycle gene-structure classifier (18) can identify *bicycle* homologs that cannot be recognized by sequence similarity.

AlphaFold2 can predict bicycle protein structures using custom multiple sequence alignments

Since these two bicycle proteins exhibited no significant sequence similarity to others in Genbank (37) and the solved structures were different from all structures in the PDB (24), we submitted our structures as targets for the 15th Critical Assessment of Techniques for Protein Structure Prediction (CASP15) challenge (38, 39). CASP is a community-wide experiment held every two years that assesses computational methods for determining protein structures from amino acid sequences. At the CASP15 competition,

in the single protein and domain-modeling category, 162 groups submitted predictions for g3873 (CASP ID: T1130) and g2703 (CASP ID: T1131). Final metric GDT_TS scores were used to evaluate the accuracy of the predictions relative to the experimental structures (40). As expected, given the breakthrough success of AlphaFold in CASP14 (41, 42), most prediction teams used deep learning-based algorithms, resulting in >90 GDT_TS scores for most targets. The two groups that accurately predicted the structure of g3873 (T1130) (GDT_TS scores >50) in CASP15 incorporated MSAs of bicycle proteins reported in one of our earlier studies (18). In contrast, no group correctly predicted g2703 (mean GDT_TS score of 12, range 11.65-25.31), making g2703 (CASP ID: T1131) the lowest-scoring target (43). This may be because homologs of g2703 were not present in any previously released genome sequences and thus could not be used to generate MSAs. Consistent with the results from CASP15, we also failed to generate accurate predictions for both g2703 or g3873 using AF2, AF3, and ESM-Fold in-house with publicly available sequence databases (44) (Figure 2 and S9).

Given these observations, we hypothesized that MSAs from homologous proteins are required for accurate reconstruction using current machine learning algorithms for structure prediction. We therefore sequenced and annotated the genomes of three species closely related to *H. cornu*: *H. betulae*, *H. hamamelidis*, and *Hamamelistes spinosus* (Figure S10c). In addition, we re-annotated genomes from several aphid species distantly related to *H. cornu* (13), *Schlechtendalia chinensis* (45), *Tetraneura akinire* (18), and *Acyrtosiphon pisum* (46), guided by publicly available RNA-seq data (www.ncbi.nlm.nih.gov/sra), and predicted *bicycle* genes using a gene-structure classifier (18). Searching these translated

aphid proteomes with *phmmer* (47) identified 55 and six homologs of g3873 and g2703, respectively. When provided with these custom MSAs, AF2 accurately predicted the experimentally determined X-ray structures of both proteins (Figure 2B, 2D, and S11). Thus, despite the fact that bicycle proteins were not part of the training set used to parameterize AF2, this model was able to recover bicycle protein structures when provided with MSAs of homologous proteins.

Bicycle proteins exhibit extensive structural diversity and surface variability

Given the success of predicting the experimentally solved bicycle proteins with AF2, we generated MSAs and predicted structures of 4924 bicycle proteins from seven aphid species (Table S7). We retained only protein structures for which $\geq 80\%$ of amino acids exhibited a per-residue measure of local confidence (pLDDT) ≥ 60 over a minimum length of 70 residues (Figure S10). Filtering resulted in a dataset of 2400 high confidence structures, allowing length-controlled cross-species comparisons of both protein structure and physicochemical space.

We used this large dataset to explore four questions: Is the saposin-like fold conserved across aphid bicycle protein structures? Are the tandem and helix-swapped permuted structures related? How conserved is the bicycle protein structural and physicochemical space? Finally, does the conformational space and the associated physicochemical signature of bicycle proteins permit inference of a putative function for bicycle proteins?

Saposin-like folds are found in all aphid bicycle proteins

For all aphid bicycle proteins in our curated dataset, alpha-helical saposin-like domains are predicted to occur in a wide variety of arrangements, including one, two, three, four, or six saposin-like domains (Figure 3 and S12). Many of these protein structures were predicted with high confidence (Figure S10) and all high-confidence predicted folds possess at least one saposin-like domain (e.g. Figure 3). Nonetheless, these domains could not be perfectly superimposed and instead displayed a wide variety of quantitative variations in saposin-like domain shape, size, and arrangement, as we discuss in more detail below.

The tandem-linked bicycle proteins likely evolved from helix-swapped proteins

To explore the evolutionary relationship between tandem-linked and helix-swapped saposin-like domain bicycle proteins, we identified distantly related paralogs of g2703 using *jackhmmer* (48). We aligned the amino acid sequences of remote paralogs to estimate the phylogeny of these proteins and mapped the protein models predicted by AF2 onto the phylogeny (Figure 3). We observed that most proteins closely related to g2703 lack disulfide-bonding cysteines and are predicted to adopt tandem-linked saposin-like domains, similar to the crystal structure of g2703. (As we found for g3873 and g2703, AF2/3 and ESMFold produced low-confidence structures for these paralogs without custom-curated MSAs (examples shown in Figures S13 and S14)). g2703 and related proteins with tandem-linked domains are phylogenetically nested within clades of bicycle proteins that contain two, four, or six helix-swapped saposin-like domains with two or more disulfide bonds (Figure 3). Thus, it is likely that bicycle proteins adopting the

unswapped saposin-like domains evolved from proteins with helix-swapped saposin-like domains.

Bicycle proteins exhibit diverse sequences and structures

Given this evidence that bicycle proteins can evolve diverse structures, we were motivated to further explore bicycle protein diversity. We first characterized protein sequence divergence using pairwise length-normalized Levenshtein distances (48) (lower half of Figure 4A). Almost all sequences are highly divergent from each other, with most pairwise distances > 0.75 . The dendrogram adjacent to the distance matrix (Figure 4A) reveals that some sequences cluster together within species, especially for *S. chinensis*, but even these sequences do not exhibit clear clusters of similarity that might define subfamilies. Thus, bicycle proteins display little evidence of sequence similarity, extending our initial observations of the extreme diversity of bicycle protein sequences within species (13, 18).

We then explored AF2 predicted structural diversity by calculating Levenshtein distances using the ABEGO classification (49) of structural motifs of amino acids (upper half of Figure 4A). The ABEGO-based map exhibits a wider dynamic range than the raw amino-acid sequence-based map, with distances saturating at approximately 0.5 rather than near 0.8 (Figure 4A), suggesting that bicycle proteins do share similarities in ABEGO space. Also, the ABEGO distance matrix displays some clustering, although little of this clustering is by species (dendrogram on top of Figure 4A).

To further explore the clustering observed in the ABEGO-based map, we performed t-SNE (50) embedding of all-vs-all TM-align comparisons for all 2400 proteins (Figure 4B). Many distinct structural neighborhoods were identified in this embedding. The medoid models for each cluster are shown in Figure S15. Substantial species partitioning was observed for some clusters (Figure S16, S17). This is most striking for the phylogenetically isolated species in our study: *A. pisum*, *T. akinire*, and *S. chinensis* (Figure S10c). Nonetheless, bicycle proteins from *Hormaphis* and *Hamamelistes* occupy most of these clusters, indicating that the relative arrangements of the alpha helices in the saposin-like domains vary extensively, even within species. Thus, the bicycle protein family has evolved to occupy a large and discontinuous region of structural space within the context of the saposin-like four-helix domains.

Bicycle proteins occupy diverse physicochemical spaces

In an effort to identify features that may be conserved among the bicycle proteins, we assessed the physicochemical features of each predicted protein. Five complementary properties were examined: surface chemistry, electrostatic charge, amphipathic character, surface hydrophobic patchiness, and surface texture (Tables S7 and S8). We embedded these features using UMAP (52) and attempted to identify groups of similar proteins using Leiden clustering (48) (Figure 5). Even though UMAP tends to “aggressively” cluster points (51, 52), we observe an almost uniform distribution of bicycle proteins characterized by these five properties, suggesting that bicycle proteins sample relatively uniformly across this physicochemical space.

We observed only a few interpretable subgroups in this physicochemical space, defined primarily by distinct combinations of surface hydrophobicity, charge balance, and amphipathicity. One subset of clusters is defined primarily by hydrophobicity and patch morphology: clusters 1, 6, and 3, which combine hydrophobicity with minimal amphipathicity. A second subset was dominated by electrostatics and charge composition, being either positively charged (cluster 2), negatively charged (cluster 4), or possessing a pronounced surface dipole (cluster 5). The third subset possesses amphipathic helices and high hydrophobic moment (clusters 7 and 9) or low hydrophobicity with amphipathic signatures (cluster 8). Medoids taken as representatives of each of these clusters and their Poisson-Boltzmann surface electrostatics are superimposed on the structures to aid visualization of these physicochemical subtypes (Figure S18). The physicochemical subtypes are not species-partitioned (Figure S19), suggesting that the variability in physicochemical features extends across the bicycle protein family (Figures S20 and S21).

Lack of conserved domains thwarts the prediction of bicycle protein functions

The prior analysis suggests that there is limited conservation of features that might provide clues about the molecular function of bicycle proteins. Therefore, we examined whether surface residues displayed any patterns of conservation. Mapping Shannon entropy of residues derived from a multiple sequence alignment of bicycle proteins, which provides a quantitative measure of per-residue conservation. Sequence entropy was mapped onto the crystal structure of g3873 (Figure S22), as a representative for the most common fold among predicted bicycle proteins. All conserved residues are buried within the interior of the protein, likely required for structural packing, whereas all surface

residues show extreme variability. This absence of any specific conserved sequence signature on the exterior of the protein precludes an inference of bicycle protein functions.

Discussion

Pathogens frequently introduce effector proteins into hosts for multiple purposes, including driving physiological and developmental programs beneficial to the parasite, securing access to nutrients, and suppressing the host immune response (53, 54). These proteins often alter host cell structure, gene expression, or normal signaling pathways. Many effector proteins do not exhibit domains of known function, and the putative functions of most effector proteins are challenging to decipher (12, 55, 56). Protein structure can provide one path toward insight into effector functions.

Here we report the crystal structures of two *H. cornu* aphid effector proteins that belong to the bicycle protein family. The CYC motifs present in the g3873 structure contribute to two saposin-like folds, in a helix swapped arrangement, secured by disulfide linkages. In contrast, the g2703 structure includes tandem-linked saposin-like folds. It is striking that although g2703 includes no cysteines, and thus no CYC motifs, it also generates saposin-like folds. This shared structural motif suggests that g3873 and g2703 shared a common ancestor. Indeed, phylogenetic (Figure 3) and sequence (Figure S4) analyses suggest that proteins exhibiting tandem saposin-like folds evolved from proteins with helix-swapped saposin-like folds.

Despite the presence of saposin-like folds in these proteins, current evidence suggests that bicycle proteins do not possess functions similar to canonical saposins. Saposins are found across eukaryotes and participate in membrane recycling (27, 31). These proteins contain a strongly hydrophobic core that can be exposed to capture lipids. The structures of the bicycle proteins solved in this study do not display features that would allow a hydrophobic core to be exposed. Instead, the saposin-like fold apparently provides a rigid backbone for a rapidly evolving solvent-exposed surface. Other proteins without sequence similarity to saposins contain domain-swapped saposin-like folds (57), although in this case, pairs of helices are swapped, which is different from the single helix swap observed in g3873. Saposin-like folds have also been observed for other proteins, distinct from those that possess lipid or membrane-modifying properties (58, 59). Thus, the saposin-like fold is likely an example of convergent evolution in multiple independent protein families and may provide a stable backbone that allows diversification in sequence space for multiple functions (59). Another example of this kind of convergent evolution is found in the lipocone family of proteins that act in several functional contexts, including regulation of membrane lipid composition, extracellular polysaccharide biosynthesis, and biogenesis of lipopolysaccharides (60).

While AF2 and other deep learning and protein language models failed to accurately predict bicycle protein structures with default settings and existing databases, we found that AF2 can accurately predict these structures when provided with MSAs of homologous proteins identified from the genomes of closely related species. By sequencing several new aphid genomes and re-annotating previously available genomes to optimize the annotation

of *bicycle* genes, we generated 2,400 high-confidence bicycle protein structures across seven aphids species. This large dataset allowed us to explore bicycle protein structure evolution and to search for sequence, structural, and physicochemical patterns that may provide clues about bicycle protein function.

We observed high levels of sequence diversity amongst the bicycle proteins, (Figure 4A) with no clear clusters of similar sequences, even within a single species. In contrast, the predicted structure space, either ABEGO encoded (Figure 4A) or measured by structural similarity with TM scores (Figure 4B), displayed clusters of similar backbone structures. Individual clusters were mostly not associated with individual species, indicating that a broad diversity of protein shapes are encoded in bicycle proteins from multiple species.

In contrast to the clusters in shape space, we observed a nearly continuous distribution of bicycle proteins in physicochemical space (Figure 5), which is also not strongly partitioned by species. This variability is consistent with the lack of sequence conservation, particularly on the surface of bicycle proteins (Figure S22). There are thus no surface regions of these proteins that exhibit conservation in sequence or physicochemical space, even amongst subsets of proteins, that might guide inference of the molecular functions of these proteins.

The proteins within sub-clusters of the structural space (Figure 4B) do not tend to cluster within the physicochemical property groups shown in Figure 5, indicating that structure and physicochemical properties have evolved independently within this protein family. As a result, the diversity of protein structures and surface physicochemical

properties highlight the challenge of inferring effector protein function strictly from structural data. Nonetheless, the presence of structural and amino acid sequence diversity exhibited by bicycle proteins, combined with our previous finding that *bicycle* genes have evolved in response to strong positive natural selection (13), suggests that individual bicycle proteins may perform distinct functions in plant cells (13) and/or that they have evolved to evade plant defense recognition systems.

Acknowledgements

We thank Goran Ceric, Robert Lines, Donald Olbris, and Tom Dolafi for assistance in installing, running, and maintaining the software and databases used in this study. For help generating the *Hormaphis betulae* genome, we thank Dr. Katsushi Yamaguchi for next-generation sequencing and bioinformatics support, Dr. Shunta Yorimoto for aphid collection, and Dr. Asahi Akita for next-generation sequencing support, including genomic DNA extraction and library preparation. We thank members of David Clapham's lab for use of their Crystal Gryphon instrument and Teresa Brosenitsch for critical reading and editing of the manuscript. We acknowledge use of the Stanford Synchrotron Radiation Lightsource, SLAC National Accelerator Laboratory, which is supported by the U.S. Department of Energy, Office of Science, Office of Basic Energy Sciences under Contract No. DE-AC02-76SF00515. The SSRL Structural Molecular Biology Program is supported by the DOE Office of Biological and Environmental Research and by the National Institutes of Health (NIH Grant P30GM133894). This article is subject to HHMI's Open Access to Publications policy. HHMI lab heads have previously granted a nonexclusive CC BY 4.0 license to the public and

a sub-licensable license to HHMI in their research articles. Pursuant to those licenses, the author-accepted manuscript of this article can be made freely available under a CC BY 4.0 license immediately upon publication.

Author contributions

Protein expression and crystallization was performed by KG, AK, and FBhinderwala. X-ray structures were solved by FBhinderwala, GC. New genome sequences were generated by TM, SS, SH, and DLS. RNA sequencing and genome annotation was performed by DLS. Computational analyses and figure preparation were performed by FBhinderwala and DLS. The project was conceived by AK, FBhinderwala, FBazan, AMG, and DLS. The project was funded by AMG and DLS. The manuscript was written by FBhinderwala, AK, AMG, and DLS. All authors contributed to manuscript revisions.

Methods and Materials

Tests for protein expression: Seven of the eighteen tested bicycle proteins could be expressed in sufficient amounts in either bacterial or Baculovirus-mediated insect cell expression systems for structural studies. Poor expression of the remaining bicycle proteins may have resulted from improper folding and/or degradation of these cysteine-rich proteins.

Constructs: A synthetic gene sequence encoding the predicted secreted g2703 protein was codon optimized for expression in *E. coli* and purchased from IDT DNA Technologies. This

DNA was inserted between NotI and BamHI restriction sites into the pET51b vector.

(MilliporeSigma/Novagen) The final protein construct contains an N-terminal Strep-Tag II and a C-terminal His₁₀ tag. The DNA encoding g3873 was purchased from GenScript and inserted into the pET28a vector (MilliporeSigma/Novagen) using NdeI and XhoI restriction sites. The resulting protein construct contains an N-terminal His₆ tag followed by a TEV cleavage site.

Protein Expression and Purification: For protein expression, *E. coli* BL21(DE3) was transformed with the pET51 vector encoding *g2703*. Cells were grown for 8 hours at 37°C in 5mL of LB medium containing 100 µg/ml carbenicillin. This starter culture was added to 50mL of modified M9 medium, containing 4g/L ¹⁵NH₄Cl and 0.2 g/L glucose as nitrogen and carbon sources, respectively. Cells were grown overnight and diluted to an A₆₀₀ of 0.25 into 1L of fresh modified M9 medium and grown to an A₆₀₀ of 0.8. Expression was induced with 500 µM IPTG overnight at 18°C. For selenomethionine labeling, 100 mg/L of selenomethionine was added to the culture at an A₆₀₀ of 0.5, cells were grown at 18°C for 1 hour, induced with 500 µM IPTG, and grown further overnight. Cells were harvested by centrifugation at 4000g at 4°C and resuspended in lysis buffer (50 mM NaH₂PO₄, pH 8, 250 mM NaCl, 1x EDTA-free protease inhibitor cocktail) and lysed on ice by sonication (for a total process time of 10 min with 5 sec on and 10 sec off at 50% power). Cell debris was removed by centrifugation at 16,000 g for 45 min at 4°C, and the Strep-tagged protein was purified over a StrepTrap HP column (Cytiva) in 50 mM NaH₂PO₄, pH 8, 250 mM NaCl, pH 8.0 (affinity buffer). The column was washed with 10 volumes of affinity buffer, and the bound protein was eluted in affinity buffer containing 250 mM desthiobiotin. Protein-

containing fractions were pooled and concentrated using an Amicon-type centrifugal device with a 10kDa cutoff filter. 5mL of the concentrated sample was further purified by size exclusion chromatography (SEC) over a Sephadex 75 column (Cytiva) in SEC buffer (50 mM NaH₂PO₄, pH 7, 150mM NaCl). Following SEC, a final purification step was performed by ion exchange over a MonoQ column with a thirty-column volume gradient from 0-0.5M NaCl in 50 mM NaH₂PO₄, pH 7 buffer. Protein fractions at 0.15 M NaCl corresponding to the purified g2703 were pooled.

For protein expression of the His-tagged g3873 protein, Shuffle T7 cells (New England Biolabs) were transformed with the pET28a vector containing the g3873 gene. Cultures were grown as detailed above for g2703 expression. Harvested cells were resuspended in His-affinity buffer (50 mM NaH₂PO₄, pH 8, 250 mM NaCl, 30 mM imidazole) with 1x EDTA-free protease inhibitor cocktail and lysed by sonication as described above. Since most of the protein was in the insoluble fraction, the pellet was further sonicated in denaturing affinity buffer (50 mM NaH₂PO₄, pH 8, 250 mM NaCl, 30 mM imidazole, 8M Urea) and the supernatant was clarified by centrifugation. The cleared supernatant was loaded on two tandem 5 mL HisTrap FF columns (Cytiva), and the columns were washed with five column volumes of denaturing affinity buffer. The bound protein was refolded on the column by washing with 100 column volumes of a linear urea gradient from 8M urea to 0M urea in the affinity buffer. After a wash with five column volumes of 6% elution buffer (50 mM NaH₂PO₄, pH 8, 150 mM NaCl, 500 mM imidazole):94% affinity buffer, the refolded G3873 protein was eluted in 100% elution buffer. The N-terminal His tag was cleaved by adding 10:1 protein: TEV protease to the protein and incubation at room temperature for 4 hours.

The cleaved g3873 protein was passed over a His-trap affinity column to remove the remaining TEV protease and the His-tag fragment. The column flow-through was further purified by size exclusion over a Sephadex 75 column in the SEC buffer. Protein-containing fractions were pooled and concentrated using an Amicon device with 10 kDa cut-off filters. Final protein concentrations were estimated by measuring A_{280} (NanoDrop Instruments), using an extinction coefficient of $15,600 \text{ M}^{-1} \text{ cm}^{-1}$ for g3873 and $49,800 \text{ M}^{-1} \text{ cm}^{-1}$ for g2703. Protein purity and labeling efficiencies were assessed by mass spectrometry. Pure protein samples were flash-frozen in liquid nitrogen and stored at -80°C until further use.

Crystallization and data collection: Both g2703 and g3873 proteins were buffer exchanged into 25 mM HEPES buffer at pH 7.0 and used for crystal screening at 4°C and 20°C by the hanging drop procedure. Pure protein was mixed with an equal volume of reservoir solution. Crystallization hits obtained from initial screens were optimized using a range of pH, protein concentration, and varying buffer, salt, and precipitant composition. Diffraction-quality crystals were harvested, flash frozen in liquid nitrogen with cryoprotection in 20% (v/v) glycerol containing reservoir solution. Diffraction datasets were collected at the Stanford Synchrotron Radiation Lightsource (SSRL) BL12-2 beamline. The g2703 and the Se-Met g2703 crystals were in the space group of $P2_12_12_1$ and diffracted to $1.4 \text{ \AA}$ resolution. The crystals of g3873 diffracted to $2.0 \text{ \AA}$ and were phased using sulfur-based anomalous diffraction. Crystallographic datasets were integrated and scaled using CCPN. Crystal data collection parameters and all structure statistics are summarized in Table 1.

Genome sequencing:

High molecular weight genomic DNA was prepared from samples of aphids collected from a single gall for each of the species *Hormaphis hamamelidis*, *Hormaphis betulae*, and *Hamamelistes spinosus* (Table S11) according to the manufacturer's protocol (CG000145_SamplePrepDemonstratedProtocol -DNAExtractionSingleInsects.pdf, <https://support.10xgenomics.com/genome-exome/sample-prep/doc/demonstrated-protocol-dna-extraction-from-single-insects>). HMW DNA was quantified using a Qubit fluorometer (Thermo Fisher Scientific, Cat #Q32866) with the Qubit™ dsDNA BR Assay Kit (Thermo Fisher Scientific, Cat #Q32850) and fragment size was assessed by pulsed-field gel electrophoresis (PFGE). The DNA was run on a 1% agarose gel (Seakem Gold Agarose, Lonza, Rockland, ME, USA, Cat #50150) in 0.5x TBE buffer using the BioRad CHEF Mapper system (BioRad, Hercules, CA, USA, Cat #M1703650) for 15 hours. The gel was then stained with SYBR Gold dye (Thermo Fisher Scientific, Waltham, MA, USA, Cat #11494). PFGE results indicated that the aphid DNA ranged in size from 50 to 300 kbp

For all three species, 10× Genomics Chromium linked-read libraries were generated using the Chromium Genome Library Kit & Gel Bead Kit v2 (10× Genomics, San Francisco, CA, USA, Cat #120258), following the manufacturer's protocol. The Illumina reads were assembled using the Supernova assembler (ver. 2.1.1).

For *Hamamelistes spinosus*, HiC reads were acquired to facilitate scaffolding of this genome. A single insect was frozen in liquid nitrogen in an Eppendorf tube, ground with a plastic pestle in 1 mL 1% formaldehyde for 20 min. Formaldehyde was quenched by adding 110 µL 1.25 M glycine for 15 min. The sample was spun at maximum speed for 15 min in a tabletop centrifuge. The liquid was removed and replaced with phosphate buffered saline,

then spun again. The supernatant was removed, and the sample was stored at -80°C prior to shipping to Phase Genomics (Seattle, Washington, USA) for Hi-C library preparation and sequencing.

We used a 3D-DNA assembly pipeline (63) in “haploid mode” to detect misjoins in the *H. spinosus* Supernova assembly and generate chromosome-scale super scaffolds. HiC alignments for the 3D-DNA pipeline were generated with Juicer v1.6.2 (64) using default settings. The scaffolded assembly was manually reviewed based on inspection of the HiC contact map using Juicebox Assembly Tools (65). The assembly was screened for contamination with BlobTools v1.0.1 (66, 67) using read coverage from the 10x Genomics linked reads and taxonomy information from BLASTN v2.2.31 (68) searches against the National Center for Biotechnology Information (NCBI) nucleotide database. We assessed the quality and completeness of the final assembly using BUSCO v3.0 (69, 70) with the arthropoda_odb9 gene set (71) and by comparing K-mer content of the 10X Genomics linked reads to the assembly with KAT comp v2.3.1 (72).

RNA sequencing:

For all three species, salivary glands were dissected from individual insects by placing insects in phosphate buffered saline and then grasping insects with forceps at the pronotum and anterior abdomen and pulling apart. The salivary glands remain attached to the brain on the anterior portion. The glands were isolated from other tissue using Minutien insect pins, with their tips bent $\sim 30^{\circ}$, mounted in micro dissecting needle holders. Separated glands and the remaining carcass were collected separately into 100 μL

of Arcturus PicoPure Extraction buffer. Total RNA was prepared using the Arcturus PicoPure RNA Isolation kit including the optional DNase step. Barcoded RNASeq libraries were prepared for Illumina NextSeq 550 sequencing (150 bp PE) with a method described previously (73).

Annotation of bicycle genes

The new genome assemblies were re-annotated existing genome assemblies with RNAseq evidence using BRAKER (74). We then predicted *bicycle* genes using a gene-structure classifier (18) and manually annotated all predicted *bicycle* genes within Apollo (75) using RNAseq evidence collected from salivary glands. GFF files containing BRAKER annotation plus manual annotations are provided as supplementary material.

Phylogeny estimation

An unrooted phylogeny was estimated for the seven species used in this study (Figure S9c) using *getphylo* (76) using all default parameters and starting with FASTA files of all predicted proteins from our new genome annotations for all species.

AlphaFold2 protein structure prediction

AlphaFold2 (77) was installed on the Janelia Compute Cluster. Initial predictions for G3873 and G2703 were carried out using default sequence databases. Since these predictions did not match the crystal structures, we generated custom MSAs for bicycle proteins using predicted protein sequences for previously published aphid genomes and our new aphid genome sequences using a custom bash shell script (Supplementary Material). We then ran

AlphaFold2 with the flag `--use_precomputed_msas`. These predictions, generated using custom MSAs, had higher confidence and resulted in models that were similar to the crystal structures for both proteins.

We used the above procedure to predict models of all potential bicycle proteins in *H. cornu*, *A. pisum*, and *S. chinensis*. In addition, we also predicted bicycle protein models for all proteins identified by *jackhmmer* search (78) with G2703.

AF2 predictions and filtering: The initial set of all predicted models, was filtered according to the following criteria: (1) low-quality AF2 prediction were removed, using a generous cut-off of the median pLDDT score of 60 for at least 80% of the length of the protein sequence, and (2) the polypeptide had to contain at least 70 residues and more than one alpha helix. (Supplementary Figure S10). From the initial set of 4093 proteins, applying the requirement of a median pLDDT >60 for at least 80% of the polypeptide length eliminated 1530 proteins, and the length and helical content filter eliminated another 140 models. The remaining 2423 protein models were inspected manually, and an additional 20 models were removed since these proteins exhibited >99.0% sequence identity to another bicycle protein; they were essentially duplicates. Details regarding the number of proteins per aphid species predicted by subsequent analyses are provided in the SI (Table S7)

Sequence variability across the bicycle proteins

The overall amino acid sequence space that is sampled by the bicycle proteins was queried as to variability using the length-normalized Levenshtein (48) distance. This provides a

measure for the difference between two sequences. For each sequence pair, a and b , the Levenshtein distance $d(a, b)$ was computed, which provides the minimum number of substitutions, insertions, and deletions required to transform a into b . Since the dataset includes proteins with up to six CYC motifs, it is essential to normalize this distance by the protein max length:

$$D(a, b) = \frac{d(a, b)}{\max(|a|, |b|)}$$

Where a and b denote lengths of the two protein sequences in the pair. This results in a symmetric distance between 0-1 for sequences of comparable length (with $D = 0$ for identical sequences) and permits us to compare divergence across the dataset. All pairwise distances were computed in an all-vs-all manner and used for the heatmap visualization. Identification of the most ‘isolated’ sequences within the bicycle proteins was carried out as follow: for each bicycle protein a , the fraction of other bicycle proteins b whose normalized distance from a exceeds a threshold T is calculated:

$$frac_{\geq T}(a) = \frac{1}{N-1} \sum_{b \neq a} 1\{D(a, b) \geq T\}$$

where N is the total number of bicycle proteins and $1\{D(a, b) \geq T\}$ is the indicator function. Large values of $frac_{\geq T}(a)$ indicate that a is distant from a large fraction of members in the dataset (i.e., an isolated sequence), whereas smaller values indicate that a is relatively close to most sequences.

We also generated an ABEGO Levenshtein distance map for all 2403 bicycle proteins in the filtered dataset. Comparing all 2403 bicycle proteins demonstrates that the ABEGO sequences within the family are highly diverse. We clustered the ABEGO sequences to illustrate sequence variation across the family and the lack of conserved sequence patterns within each aphid species. The dendrogram was colored at the lowest level using the species to which the corresponding bicycle protein belonged (Figure 4B)

Evaluating physicochemical diversity across bicycle proteins: Physicochemical properties of all predicted bicycle models were computed based on the location and nature of each amino acid in the AF2 predicted structures. A total of 22 different features were evaluated (Table S8, 9). The solvent accessible surface area (SASA) was estimated using the FreeSASA algorithm (79), where all atom-level SASA were summed per residue to obtain the total SASA. Charged and hydrophobic SASA partitions were derived by defining hydrophobic (I, L, V, F, C, M, A, W, Y, P), positively charged (K, R, and H at pH 7.0), and negatively charged amino acids (D and E). $f_{\text{SASA hydrophobic}}$ was calculated as $\text{SASA}_{\text{hydrophobic}} / \text{SASA}_{\text{total}}$.

In addition, a set of features describing the charge density and net charge was calculated. The fraction of positive, negative, and hydrophobic residues was calculated by dividing individual values by the length of the sequence. Surface hydrophobicity, surface roughness, hydrophobic patchiness, amphipathic helical fraction, and max hydrophobic moment were also tabulated. All properties and the formulas used to compute them are described in detail in Supplementary methods.

The presence of correlations between these features was evaluated. Correlation will be present since some features contribute to the same overall property. We checked the range of each variable in order to prevent skewed assessment of the overall physicochemical nature of the proteins. The calculated parameters for all 2400 bicycle proteins are available in Table S9.

Using the reduced set of uncorrelated features, our analysis used a custom Python script within the *sklearn* toolkit (67) and the data were plotted using *matplotlib* (68). Non-linear dimensionality reduction was performed with t-SNE (51) and UMAP (52, 69) for all non-protein size-dependent parameters, i.e., 11 out of the 22 original and a reduced set of physicochemical parameters (9 out of 11), with a correlation threshold set at 0.85. After assessing various combinations of parameters, we opted to visualize the UMAP projection using 30 neighbors and a distance of 0.05, with a Leiden resolution of clusters at 1.2. This allowed us to preserve the global structure of the map and maximize the consistency of cluster memberships. Medoids representing each of the clusters were identified and used to highlight the physicochemical properties of members in each cluster. The overall distribution of each feature across aphid species and Leiden clusters is provided in the supplementary data (Figure S19).

Mapping the diversity of structural space across bicycle proteins: Using the 2400 AF2 predicted bicycle protein models, we performed an all-versus-all comparison of TM scores (70, 71). This all-versus-all tm-score matrix was reduced to a two-dimensional space using t-SNE (51), with subsequent Leiden clustering (49). This resulted in a global map of predicted bicycle protein structures (provided in Figure 5). Medoids representative of each

cluster were identified, and their model structures were compared to the experimental X-ray structure of G3873 from *H. cornu*. Clustering was solely based on structural features, since no cluster over-represented a particular aphid species or a defined polypeptide chain length.

References

1. I. Kaloshian, L. L. Walling, Hemipteran and dipteran pests: Effectors and plant host immune regulators. *J. Integr. Plant Biol.* **58**, 350–361 (2016).
2. J. Stuart, Insect effectors and gene-for-gene interactions with host plants. *Curr. Opin. Insect Sci.* **9**, 56–61 (2015).
3. C. A. Schneider, E. Calvo, K. E. Peterson, Arboviruses: How Saliva Impacts the Journey from Vector to Host. *Int. J. Mol. Sci.* **22**, 9173 (2021).
4. D. Giron, E. Huguet, G. N. Stone, M. J. A. Body, Insect-induced effects on plants and possible effectors used by galling and leaf-mining insects to manipulate their host-plant. *J. Insect Physiol.* **84**, 70–89 (2016).
5. M. B. Vogt, *et al.*, Mosquito saliva alone has profound effects on the human immune system. *PLoS Negl. Trop. Dis.* **12**, e0006439 (2018).
6. D. A. Lefteri, *et al.*, Mosquito saliva enhances virus infection through sialokinin-dependent vascular leakage. *Proc. Natl. Acad. Sci.* **119**, e2114309119 (2022).
7. P. Li, Y.-J. Lu, H. Chen, B. Day, The Lifecycle of the Plant Immune System. *Crit. Rev. Plant Sci.* **39**, 72–100 (2020).
8. M. Vinkler, *et al.*, Understanding the evolution of immune genes in jawed vertebrates. *J. Evol. Biol.* **36**, 847–873 (2023).
9. G.-Z. Han, Origin and evolution of the plant immune system. *New Phytol.* **222**, 70–83 (2019).
10. T. Nyman, F. Bokma, J.-P. Kopelke, Reciprocal diversification in a complex plant-herbivore-parasitoid food web. *BMC Biol.* **5**, 49–49 (2007).
11. D. J. Obbard, F. M. Jiggins, D. L. Halligan, T. J. Little, Natural selection drives extremely rapid evolution in antiviral RNAi genes. *Curr. Biol.* **16**, 580–585 (2006).
12. D. A. Elzinga, G. Jander, The role of protein effectors in plant-aphid interactions. *Curr. Opin. Plant Biol.* **16**, 451–456 (2013).
13. A. Korgaonkar, *et al.*, A novel family of secreted insect proteins linked to plant gall development. *Curr. Biol.* **31**, 1836–1849.e12 (2021).
14. J. C. Schultz, P. P. Edger, M. J. A. Body, H. M. Appel, A galling insect activates plant reproductive programs during gall development. *Sci. Rep.* **9**, 1833 (2019).

15. G. N. Stone, K. Schönrogge, The adaptive significance of insect gall morphology. *Trends Ecol. Evol.* **18**, 512–522 (2003).
16. M. Rostás, D. Maag, M. Ikegami, M. Inbar, Gall volatiles defend aphids against a browsing mammal. *BMC Evol. Biol.* **13** (2013).
17. M. Inbar, A. Eshel, D. Wool, Interspecific competition among phloem-feeding insects mediated by induced host-plant sinks. *Ecology* **76**, 1506–1515 (1995).
18. D. L. Stern, C. Han, Gene Structure-Based Homology Search Identifies Highly Divergent Putative Effector Gene Family. *Genome Biol. Evol.* **14**, evac069 (2022).
19. C. K. Riener, G. Kada, H. J. Gruber, Quick measurement of protein sulfhydryls with Ellman’s reagent and with 4,4'-dithiodipyridine. *Anal. Bioanal. Chem.* **373**, 266–276 (2002).
20. G. L. Ellman, Tissue sulfhydryl groups. *Arch. Biochem. Biophys.* **82**, 70–77 (1959).
21. J. Richardson, “The Anatomy and Taxonomy of Protein Structure” in *Advances in Protein Chemistry*, (Academic Press, 1981), pp. 167–339.
22. B. Schmidt, L. Ho, P. J. Hogg, Allosteric disulfide bonds. *Biochemistry* **45**, 7429–7433 (2006).
23. P. M. Harrison, M. J. Sternberg, The disulphide beta-cross: from cystine geometry and clustering to classification of small disulphide-rich protein folds. *J. Mol. Biol.* **264**, 603–623 (1996).
24. H. M. Berman, *et al.*, The Protein Data Bank. *Nucleic Acids Res.* **28**, 235–242 (2000).
25. L. Holm, A. Laiho, P. Törönen, M. Salgado, DALI shines a light on remote homologs: One hundred discoveries. *Protein Sci. Publ. Protein Soc.* **32**, e4519 (2023).
26. M. van Kempen, *et al.*, Fast and accurate protein structure search with Foldseek. *Nat. Biotechnol.* 1–4 (2023). <https://doi.org/10.1038/s41587-023-01773-0>.
27. M. Rossmann, *et al.*, Crystal Structures of Human Saposins C and D: Implications for Lipid Recognition and Membrane Interactions. *Structure* **16**, 809–817 (2008).
28. H. Bruhn, A short guided tour through functional and structural features of saposin-like proteins. *Biochem. J.* **389**, 249–257 (2005).
29. Y. Kishimoto, M. Hiraiwa, J. O’Brien, Saposins: structure, function, distribution, and molecular genetics. *J. Lipid Res.* **33**, 1255–1267 (1992).

30. E. Liepinsh, M. Andersson, J. M. Ruyschaert, G. Otting, Saposin fold revealed by the NMR structure of NK-lysin. *Nat. Struct. Biol.* **4**, 793–795 (1997).
31. V. E. Ahn, K. F. Faull, J. P. Whitelegge, A. L. Fluharty, G. G. Privé, Crystal structure of saposin B reveals a dimeric shell for lipid binding. *Proc. Natl. Acad. Sci. U. S. A.* **100**, 38–43 (2003).
32. R. Gautier, D. Douguet, B. Antonny, G. Drin, HELIQUEST: a web server to screen sequences with specific alpha-helical properties. *Bioinforma. Oxf. Engl.* **24**, 2101–2102 (2008).
33. Y. Zhang, J. Skolnick, Scoring function for automated assessment of protein structure template quality. *Proteins* **57**, 702–710 (2004).
34. J. Xu, Y. Zhang, How significant is a protein structure similarity with TM-score = 0.5? *Bioinformatics* **26**, 889–895 (2010).
35. A. Gorelik, K. Illes, B. Nagar, Crystal structure of the mammalian lipopolysaccharide detoxifier. *Proc. Natl. Acad. Sci. U. S. A.* **115**, E896–E905 (2018).
36. J. Kervinen, *et al.*, Crystal structure of plant aspartic proteinase prophytepsin: inactivation and vacuolar targeting. *EMBO J.* **18**, 3947–3955 (1999).
37. D. A. Benson, *et al.*, GenBank. *Nucleic Acids Res.* **41**, D36–42 (2013).
38. A. Kryshchak, T. Schwede, M. Topf, K. Fidelis, J. Moult, Critical assessment of methods of protein structure prediction (CASP)-Round XV. *Proteins* **91**, 1539–1549 (2023).
39. L. T. Alexander, *et al.*, Protein target highlights in CASP15: Analysis of models by structure providers. *Proteins* **91**, 1571–1599 (2023).
40. A. Zemla, null Venclovas, J. Moult, K. Fidelis, Processing and evaluation of predictions in CASP4. *Proteins Suppl* **5**, 13–21 (2001).
41. E. Callaway, “It will change everything”: DeepMind’s AI makes gigantic leap in solving protein structures. *Nature* **588**, 203–204 (2020).
42. J. Jumper, *et al.*, Applying and improving AlphaFold at CASP14. *Proteins* **89**, 1711–1721 (2021).
43. A. Kryshchak, T. Schwede, M. Topf, K. Fidelis, J. Moult, Critical assessment of methods of protein structure prediction (CASP)-Round XIV. *Proteins* **89**, 1607–1617 (2021).
44. A. J. Simpkin, *et al.*, Tertiary structure assessment at CASP15. *Proteins Struct. Funct. Bioinforma.* **91**, 1616–1635 (2023).

45. H.-Y. Wei, *et al.*, Chromosome-level genome assembly for the horned-gall aphid provides insights into interactions between gall-making insect and its host plant. *Ecol. Evol.* **12**, e8815 (2022).
46. T. C. Mathers, *et al.*, Chromosome-Scale Genome Assemblies of Aphids Reveal Extensively Rearranged Autosomes and Long-Term Conservation of the X Chromosome. *Mol. Biol. Evol.* **38**, 856–875 (2021).
47. S. R. Eddy, Accelerated profile HMM searches. *PLoS Comput. Biol.* **7** (2011).
48. V. A. Traag, L. Waltman, N. J. van Eck, From Louvain to Leiden: guaranteeing well-connected communities. *Sci. Rep.* **9**, 5233 (2019).
49. K. Sakuma, *et al.*, Design of complicated all- α protein structures. *Nat. Struct. Mol. Biol.* **31**, 275–282 (2024).
50. L. van der Maaten, G. Hinton, Visualizing Data using t-SNE. *J. Mach. Learn. Res.* **9**, 2579–2605 (2008).
51. L. McInnes, J. Healy, J. Melville, UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. *ArXiv180203426 Cs Stat* (2020).
52. L. McInnes, J. Healy, J. Melville, UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. (2018).
53. P. N. Dodds, J. P. Rathjen, Plant immunity: towards an integrated view of plant-pathogen interactions. *Nat. Rev. Genet.* **11**, 539–548 (2010).
54. M. C. Giraldo, B. Valent, Filamentous plant pathogen effectors in action. *Nat. Rev. Microbiol.* **11**, 800–814 (2013).
55. M. Franceschetti, *et al.*, Effectors of Filamentous Plant Pathogens: Commonalities amid Diversity. *Microbiol. Mol. Biol. Rev. MMBR* **81**, e00066-16 (2017).
56. K. Seong, K. V. Krasileva, Prediction of effector protein structures from fungal phytopathogens enables evolutionary analyses. *Nat. Microbiol.* **8**, 174–187 (2023).
57. B. C. Bryksa, *et al.*, Structure and Mechanism of the Saposin-like Domain of a Plant Aspartic Protease. *J. Biol. Chem.* **286**, 28265–28275 (2011).
58. B. C. Bornhauser, P.-A. Olsson, D. Lindholm, MSAP is a novel MIR-interacting protein that enhances neurite outgrowth and increases myosin regulatory light chain. *J. Biol. Chem.* **278**, 35412–35420 (2003).

59. R. S. Munford, P. O. Sheppard, P. J. O'Hara, Saposin-like proteins (SAPLIP) carry out diverse functions on a common backbone structure. *J. Lipid Res.* **36**, 1653–1663 (1995).
60. A. M. Burroughs, G. G. Nicastro, L. Aravind, The lipocone superfamily, a unifying theme in metabolism of lipids, peptidoglycan and exopolysaccharides, inter-organismal conflicts and immunity. *eLife* **14**, RP108061 (2025).
61. A. N. Hodder, *et al.*, Insights into Duffy binding-like domains through the crystal structure and function of the merozoite surface protein MSPDBL2 from *Plasmodium falciparum*. *J. Biol. Chem.* **287**, 32922–32939 (2012).
62. A. Ranjan, C. E. Chitnis, Mapping regions containing binding residues within functional domains of *Plasmodium vivax* and *Plasmodium knowlesi* erythrocyte-binding proteins. *Proc. Natl. Acad. Sci. U. S. A.* **96**, 14067–14072 (1999).
63. O. Dudchenko, *et al.*, De novo assembly of the *Aedes aegypti* genome using Hi-C yields chromosome-length scaffolds. *Science* **356**, 92–95 (2017).
64. N. C. Durand, *et al.*, Juicer Provides a One-Click System for Analyzing Loop-Resolution Hi-C Experiments. *Cell Syst.* **3**, 95–98 (2016).
65. O. Dudchenko, *et al.*, The Juicebox Assembly Tools module facilitates de novo assembly of mammalian genomes with chromosome-length scaffolds for under \$1000. [Preprint] (2018). Available at: <https://www.biorxiv.org/content/10.1101/254797v1> [Accessed 17 March 2026].
66. D. R. Laetsch, M. L. Blaxter, BlobTools: Interrogation of genome assemblies. *F1000Research* **6**, 1287 (2017).
67. S. Kumar, M. Jones, G. Koutsovoulos, M. Clarke, M. Blaxter, Blobology: exploring raw genome data for contaminants, symbionts and parasites using taxon-annotated GC-coverage plots. *Front. Genet.* **4**, 237 (2013).
68. C. Camacho, *et al.*, BLAST+: Architecture and applications. *BMC Bioinformatics* **10**, 1–9 (2009).
69. F. A. Simao, R. M. Waterhouse, P. Ioannidis, E. V. Kriventseva, E. M. Zdobnov, BUSCO: Assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* **31**, 3210–3212 (2015).
70. R. M. Waterhouse, *et al.*, BUSCO Applications from Quality Assessments to Gene Prediction and Phylogenomics. *Mol. Biol. Evol.* **35**, 543–548 (2018).
71. E. M. Zdobnov, *et al.*, OrthoDB v9.1: cataloging evolutionary and functional annotations for animal, fungal, plant, archaeal, bacterial and viral orthologs. *Nucleic Acids Res.* **45**, D744–D749 (2017).

72. D. Mapleson, G. Garcia Accinelli, G. Kettleborough, J. Wright, B. J. Clavijo, KAT: a K-mer analysis toolkit to quality control NGS datasets and genome assemblies. *Bioinformatics* **33**, 574–576 (2017).
73. M. S. Cembrowski, *et al.*, Dissociable Structural and Functional Hippocampal Outputs via Distinct Subiculum Cell Classes. *Cell* **173**, 1280-1292.e18 (2018).
74. K. J. Hoff, A. Lomsadze, M. Borodovsky, M. Stanke, *Whole-genome annotation with BRAKER* (2019).
75. N. A. Dunn, *et al.*, Apollo: Democratizing genome annotation. *PLOS Comput. Biol.* **15**, e1006790 (2019).
76. T. J. Booth, S. Shaw, P. Cruz-Morales, T. Weber, getphylo: rapid and automatic generation of multi-locus phylogenetic trees. *BMC Bioinformatics* **26**, 21 (2025).
77. J. Jumper, *et al.*, Highly accurate protein structure prediction with AlphaFold. *Nature* 1–11 (2021). <https://doi.org/10.1038/s41586-021-03819-2>.
78. L. S. Johnson, S. R. Eddy, E. Portugaly, Hidden Markov model speed heuristic and iterative HMM search procedure. *BMC Bioinformatics* **11** (2010).
79. S. Mitternacht, FreeSASA: An open source C library for solvent accessible surface area calculations. *F1000Research* **5**, 189 (2016).
80. F. Pedregosa, *et al.*, Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).
81. J. D. Hunter, Matplotlib: A 2D graphics environment. *Comput. Sci. Eng.* **9**, 90–95 (2007).

Figures

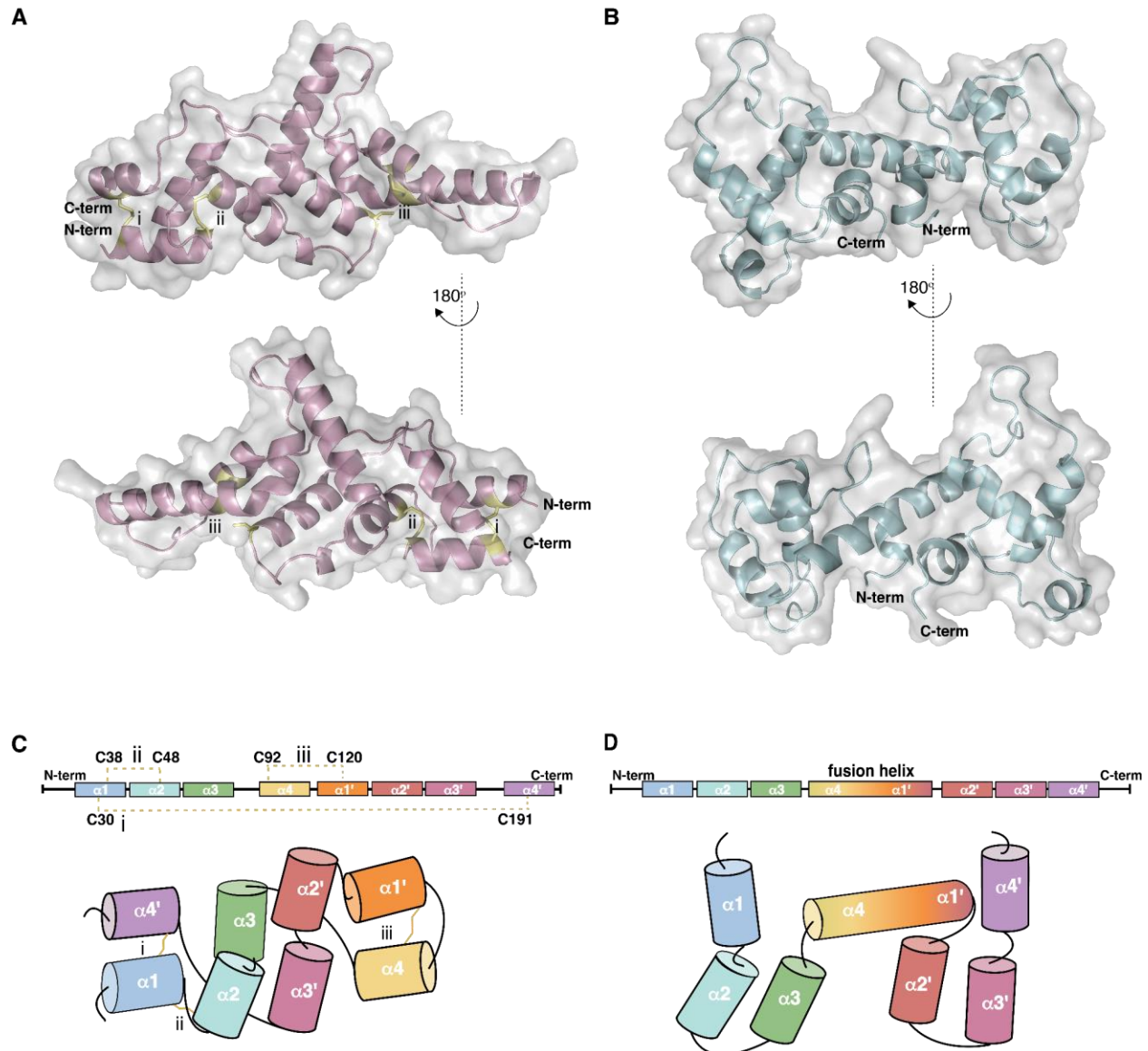


Figure 1: X-ray structures of two *H. cornu* bicycle proteins. (A) 2.0 Å X-ray structure of G3873 in ribbon representation (pink) with the three disulfide bonds and the free cysteine residue shown in stick representation (yellow), embedded in a gray surface representation. (B) 1.4 Å X-ray structure of G2703 in ribbon representation (aqua) embedded in a gray

surface representation. (C and D) Schematic representation of the secondary structure elements in the G3873 (C) and G2703 (D) polypeptide chains. Helical elements are represented by colored boxes or cylinders, labeled $\alpha 1$ through $\alpha 4'$. In G3873, the left and right side four-helical bundles include $\alpha 1$, $\alpha 2$, $\alpha 3$, and $\alpha 4'$ and $\alpha 1'$, $\alpha 2'$, $\alpha 3'$, and $\alpha 4$ helices, respectively. In G2703, the left- and right-side four-helical bundles include $\alpha 1$, $\alpha 2$, $\alpha 3$, and the $\alpha 4$ portions of the $\alpha 4$ - $\alpha 1'$ fused long central helix, and the $\alpha 1'$ portion of the fused long central helix, $\alpha 2'$, $\alpha 3'$, and $\alpha 4'$ helices, respectively. Color coding runs from the N-terminus (aqua blue) to the C-terminus (purple). Disulfide bonds are indicated by yellow lines (dashed; top; zig-zag bottom)

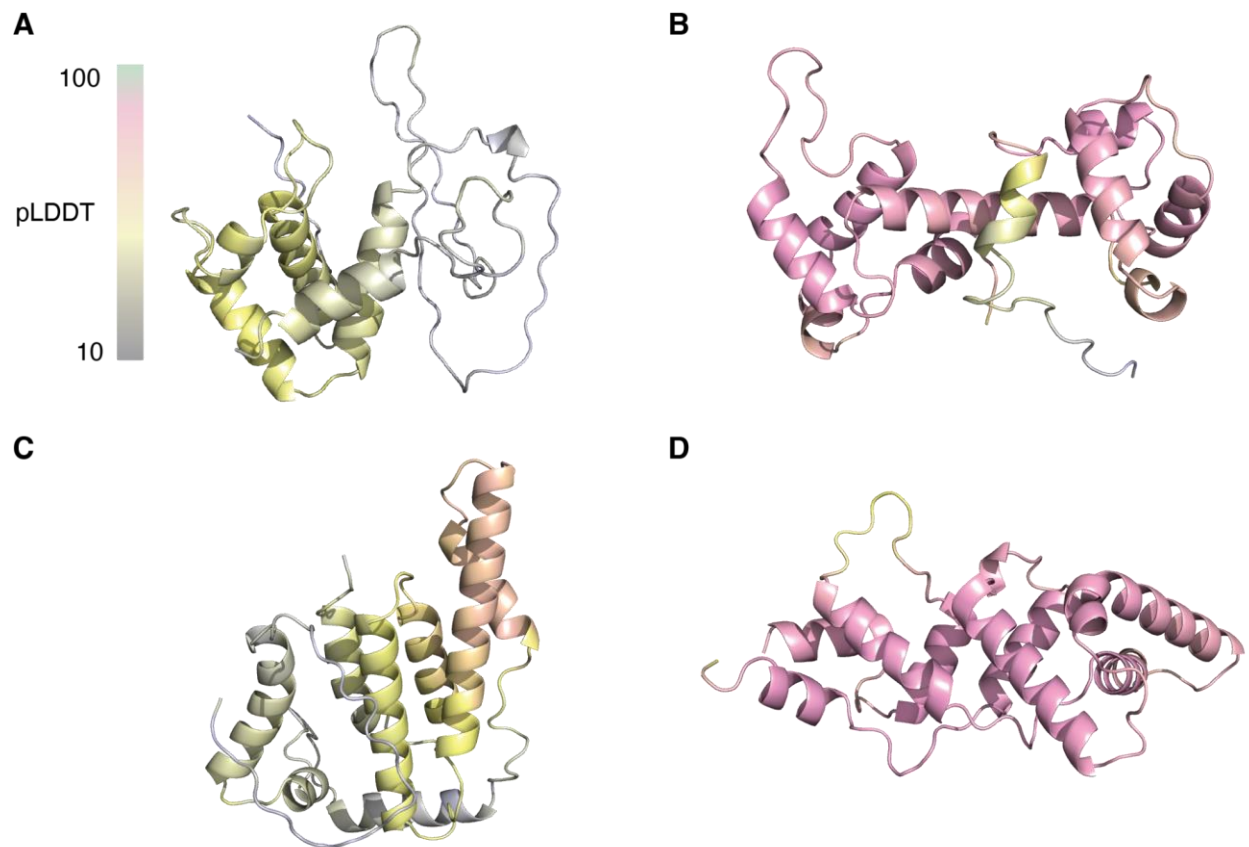


Figure 2: AF2 predicted models for G2703 (A and B) and G3873 (C and D). In (A) and (C), no MSAs were used because the default databases used by AF2 for homolog searches do not contain any homologs. In (B) and (D) custom MSAs were generated with homologous proteins from closely related species. The AF2 predicted models are colored by per-residue measures of local confidence (bar) in the predicted local distance difference test (pLDDT). Grey and pink represent ~10 and ~85 confidence levels, respectively.

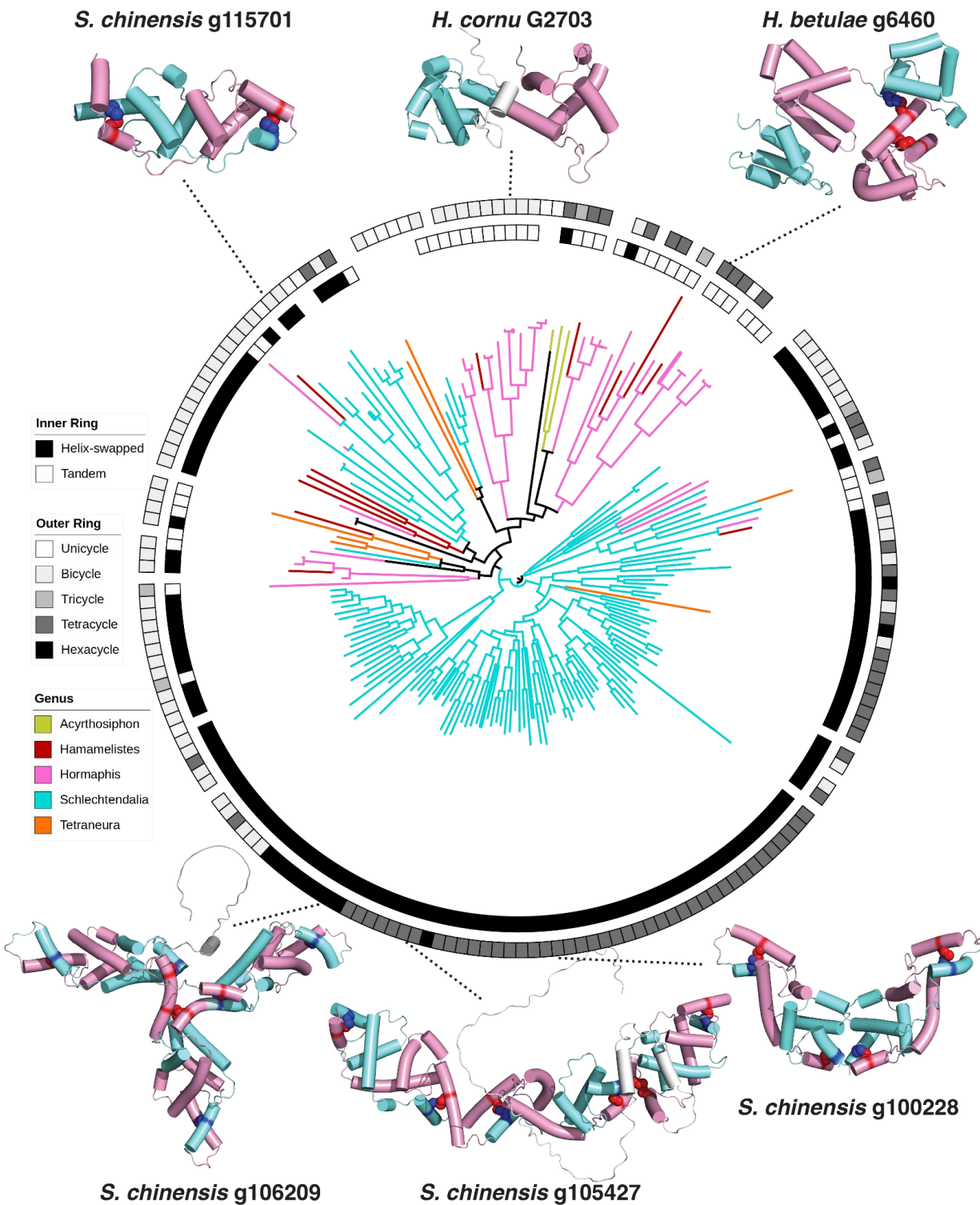
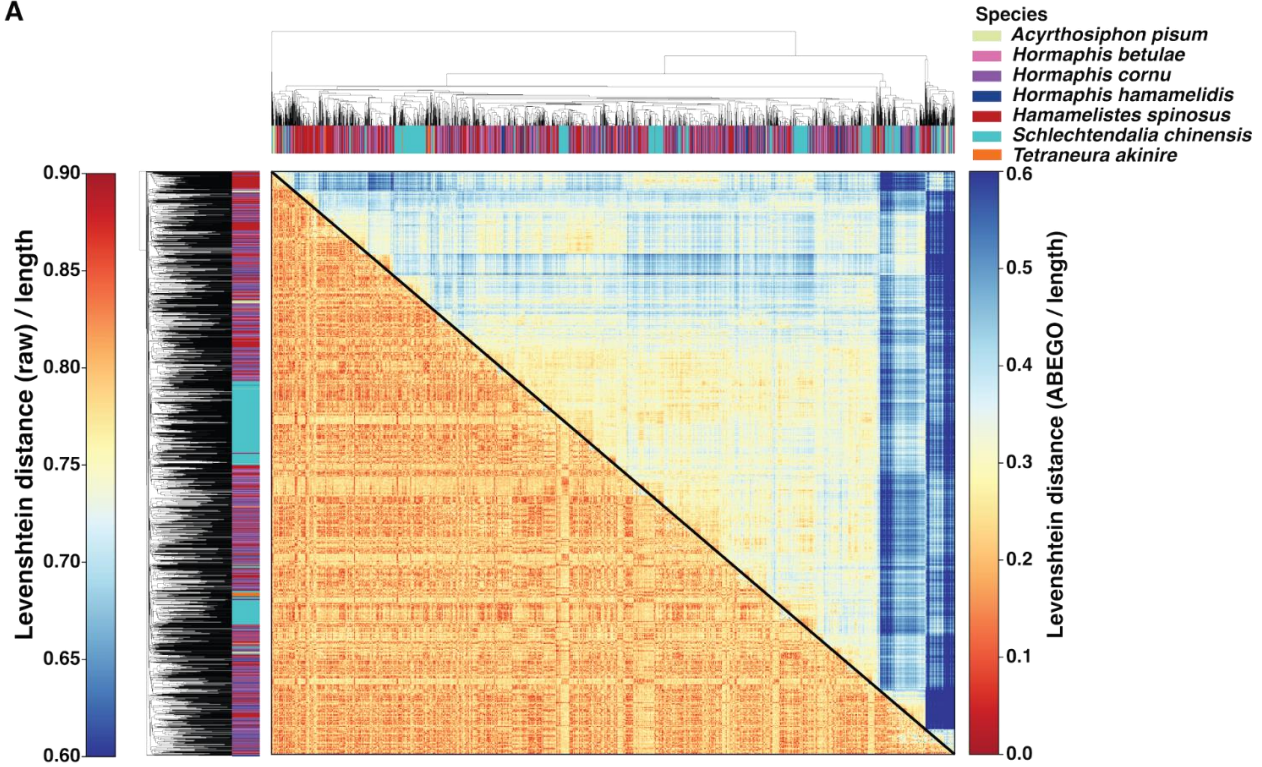


Figure 3: An unrooted tree of remote bicycle protein homologs of *H. cornu* G2703 reveals that the tandem form of bicycle proteins likely evolved from helix-swapped proteins. The

topology of predicted bicycle proteins is shown in the inner ring, either as a helix-swapped saposin-fold resembling the crystal structure of G3873 or a tandem saposin-fold resembling the crystal structure of G2703 from *H. cornu*. The AF2 predicted bicycle models across different aphid species depict examples of a single or multiple saposin-fold architectures, defined by the numbers of CYC motifs (unicycle; one CYC motif, bicycle; two CYC motifs, tricycle; three CYC motifs, etc). At the top, the tandem saposin-fold from the experimentally determined structure of G2703 from *H. cornu* is shown in the center, and an example of a tetracycle g6460 from *H. betulae* at the right. For *S. chinensis*, we show four examples of AF2-predicted models. A bicycle (g115701) and tetracycle (g100228) flank G2703 and different arrangements of hexacycles (g106209 and g105457) are depicted at the bottom. The alpha helices that form the saposin-fold are shown as pink or aqua cylinders to highlight the tandem or helix-swapped domain in each structure.

A



B

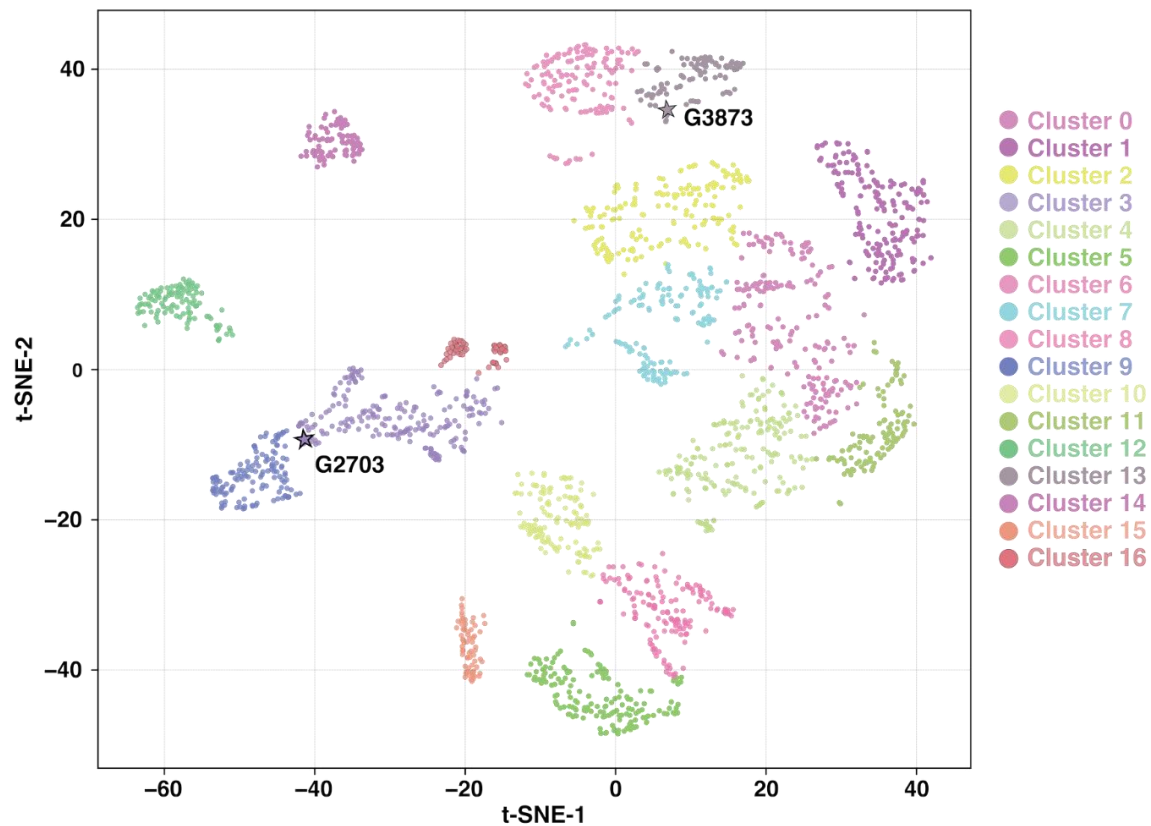


Figure 4: Bicycle proteins are predicted to adopt a wide diversity of structures. (A)

Heatmap of pairwise normalized Levenshtein distances for 2400 bicycle proteins from seven species computed using amino-acid sequences (lower) and ABEGO-encoded backbone-state strings derived from each AF2-predicted model (upper). In both heatmaps, colors represent distance (0 = identical strings; 1 = maximally different, normalized to the length). Bicycle amino acid sequences display extreme diversity (heatmap is colored blue to red, ranging from 0.6 to 1.0), even within species, whereas ABEGO strings retain intermediate similarity (heatmap distances colored from blue to red from 0.0 to 0.6), even across species. These patterns suggest that bicycle proteins exhibit some conserved higher-order backbone structural grammar despite low primary sequence identity. The dendrograms were calculated by hierarchical clustering the distance measures shown in the heatmaps. Dendrogram tips are colored by species (see key), revealing that amino acid sequences and protein structures are divergent across all aphid species sampled. (B) t-SNE embedding of all-vs-all TM scores for 2400 bicycle proteins from seven species. Each point represents one protein and is colored according to its cluster assignment ($k_{\text{cluster}} = 17$). Structurally similar proteins (higher TM scores) are located close to each other.

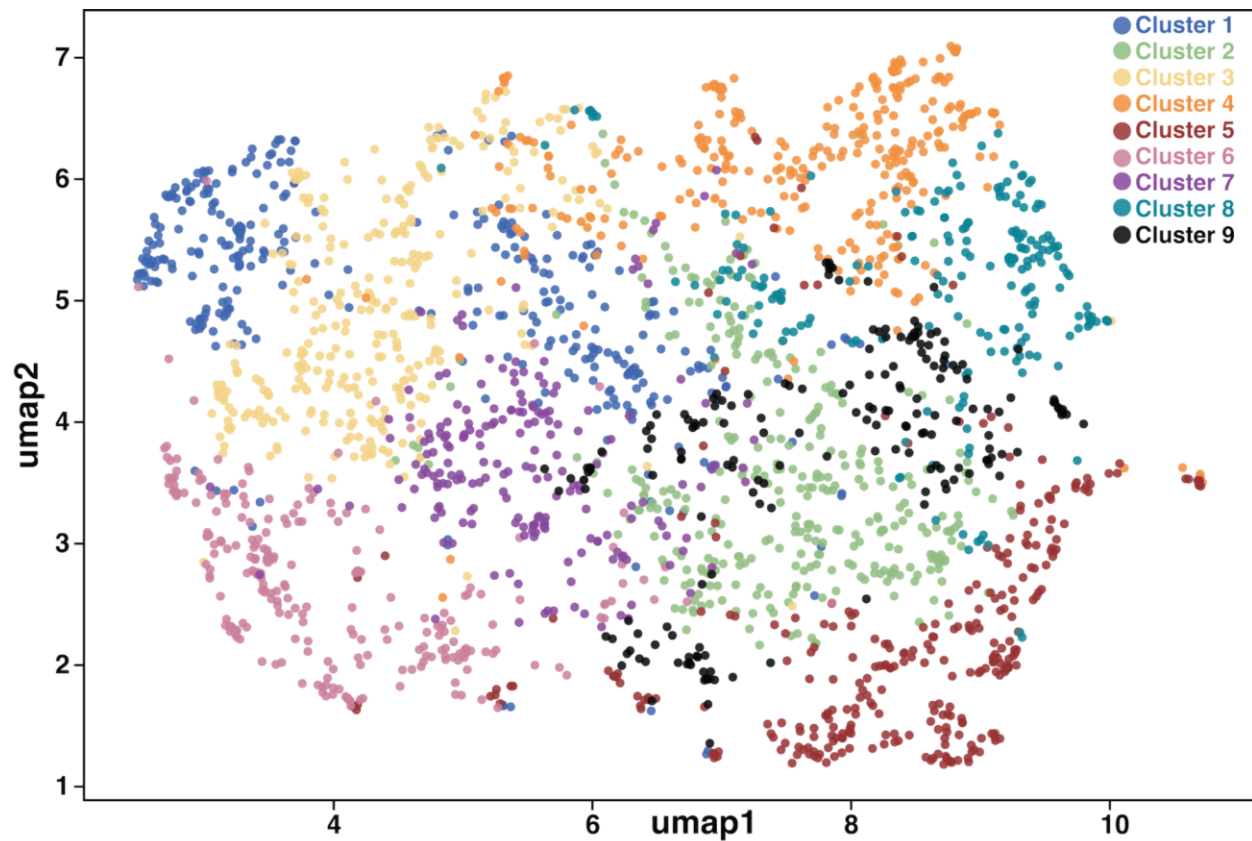


Figure 5: Bicycle proteins display a wide and almost continuous diversity of physicochemical features in the first two dimensions of a UMAP embedding. Each point represents one protein. Different Leiden clusters are shown in different colors and contain sets of proteins primarily with different quantitative combinations of surface hydrophobicity, charge balance, and amphipathicity