PARAS: high-accuracy machine-learning of substrate specificities in nonribosomal peptide synthetases

Barbara R. Terlouw[1,2], Chuan Huang[3,4,5#], David Meijer[1#], José D. D. Cediel-Becerra[6#], Marlene L. Rothe[3,7], Matthew Jenner[3], Shanshan Zhou[3], Yu Zhang[3], Christopher D. Fage[3†], Yuta Tsunematsu[8], Gilles P. van Wezel[2], Serina L. Robinson[9], Fabrizio Alberti[7], Lona M. Alkhalaf[3], Marc G. Chevrette[6,10]*, Gregory L. Challis[3,4,5]* and Marnix H. Medema[1,2]*

1. Bioinformatics Group, Department of Plant Science, Wageningen University & Research, Wageningen, The Netherlands
2. Institute of Biology, Leiden University, Leiden, The Netherlands
3. Department of Chemistry, University of Warwick, Coventry, United Kingdom
4. Department of Biochemistry and Molecular Biology, Biomedicine Discovery Institute, Monash University, Clayton, Australia
5. ARC Centre of Excellence for Innovations in Peptide and Protein Science, Monash University, Clayton, Australia
6. Department of Microbiology and Cell Science, University of Florida, Gainesville, Florida, USA
7. School of Life Sciences, University of Warwick, Coventry, United Kingdom
8. Graduate School of Bioagricultural Sciences, Nagoya University, Furo-cho, Chikusa, Nagoya, Aichi, Japan
9. Department of Environmental Microbiology, Swiss Federal Institute of Aquatic Science and Technology (EAWAG), Dübendorf, Switzerland
10. University of Florida Genetics Institute, University of Florida, Gainesville, Florida, USA

# These authors contributed equally to this work.
† Current address: Institute for Integrative Biology of the Cell (I2BC), CEA, CNRS, Université Paris-Saclay, 91198 Gif-sur-Yvette, France
* Corresponding authors: mchevrette@ufl.edu, g.l.challis@warwick.ac.uk, marnix.medema@wur.nl

Abstract

Nonribosomal peptides are chemically and functionally diverse natural products with important applications in medicine and agriculture. Bacterial and fungal genomes contain thousands of nonribosomal peptide biosynthetic gene clusters (BGCs) of unknown function, providing a promising resource for peptide discovery. Core structural features of such peptides can be inferred by predicting the substrate(s) of adenylation (A) domains in nonribosomal peptide synthetases (NRPSs). However, existing approaches to A domain prediction rely on limited datasets and often struggle with domains selecting large substrates or from less-studied taxa. Here, we systematically curate and computationally analyse 3,254 A domains and present two new high-accuracy specificity predictors, PARAS and PARASECT. A new type of A domain with unusually high L-tryptophan specificity was identified through the application of PARAS, and intact protein mass spectrometry to the corresponding NRPS showed it to direct the production of tryptopeptin-related metabolites in *Streptomyces* species. Together, these technologies will accelerate the characterisation of novel NRPSs and their metabolic products.

PARAS and PARASECT are available at https://paras.bioinformatics.nl.

Introduction

Bacteria and fungi produce a multitude of specialised metabolites, including numerous structurally and functionally diverse non-ribosomal peptides with important applications in medicine and agriculture. These include the antibiotic daptomycin[1], the immunosuppressant cyclosporine[2], the fungicide UK-2A[3], and the herbicide Bialaphos[4]. Also, nonribosomal peptides such as lugdunin[5], colibactin[6] and thanamycin[7] have been found to be key mediators of (host-)microbiome interactions.

Giant modular multienzymes called nonribosomal peptide synthetases (NRPSs) assemble these natural products. Each NRPS module contains an adenylation (A) domain, which selectively binds an amino (or other) acid and catalyses formation of an aminoacyl adenylate by reaction with ATP, and a peptidyl carrier protein (PCP) domain that is post-translationally modified via attachment of a coenzyme derived phosphopantetheine (PPant) arm to a conserved Ser residue. Nucleophilic attack of the thiol group at the terminus of the PPant arm on the activated carboxyl group of the aminoacyl adenylate results in formation of an aminoacyl thioester. In chain-elongating modules, condensation (C) domains catalyse peptide bond formation between aminoacyl thioesters attached to PCP domains in biosynthetically adjacent modules[8]. While ribosomally-synthesised peptides are limited to 20 proteinogenic amino acids, NRPSs can incorporate over 300 different building blocks[9], including proteinogenic and non-proteinogenic amino acids, alpha-keto and alpha-hydroxy acids, various N-terminal acyl 'caps', and C-terminal amines and alcohols. This, combined with a large array of on- and post-NRPS tailoring domains/enzymes, and the ability of NRPSs to work together with type I modular polyketide synthases (PKSs), results in a wide range of modifications to the peptide scaffold and yields structurally diverse metabolites with immense functional potential.

In the late 1990s, Stachelhaus *et al.* and Challis and coworkers independently identified residues in A domain active sites predicted to control substrate selectivity. Based on the crystal structure of an A domain from the first module of gramicidin S synthetase A with its cognate L-Phe substrate and AMP bound, they converged on a core set of nine selectivity-conferring amino acid residues[10], with Stachelhaus *et al.* proposing an additional tenth[11]. Stachelhaus *et al.* proposed that the selectivity-conferring residues constitute a "nonribosomal code" (sometimes referred to as the Stachelhaus code), likening it to the genetic code used as a blueprint for assembling proteins. Challis and co-workers, however, viewed this as an over-simplification, emphasising that these residues form a complex three-dimensional network of precisely positioned functional groups able to recognise diverse substrates using a combination of electrostatic, hydrophobic, and hydrogen bond interactions. This approach revealed additional nuances, such as distinct motifs for recognition of ornithine, *N*5-hydroxyornitione and *N*5-hydroxy-*N*5-formylornithine.

In a 2004 review, Challis and coworkers highlighted that recognition of substrates with smaller side chains than L-Phe likely requires fewer than nine or ten residues[12]. Models of two types of L-Pro-activating A domains with distinct "nonribosomal codes" were used to illustrate this[12]. Subsequent X-ray crystallographic studies of an A domain with its cognate L-Thr substrate / ATP and the corresponding adenylate bound provided experimental confirmation of <9-10 residues required for selectivity[13]. Conversely, recognition of substrates with larger sides chains than L-Phe would be expected to require more than 10 residues. Taken together, these observations undermined the notion of a one-size-fits-all "nonribosomal code" for prediction of A domain substrate selectivity. Notwithstanding these limitations, Challis and coworkers applied their model for A domain substrate recognition to prediction of the amino

acids selected by a cryptic NRPS encoded by the *S. coelicolor* genome[14]. Subsequent identification of the novel siderophore coelichelin as the product of this NRPS verified the accuracy of these predictions, heralding the dawn of genome mining as a new approach for discovery of novel nonribosomal peptides[15]. With (meta)genome sequencing costs at an all-time low, genome mining has now assumed a key role in natural product discovery[16,17]. Tools like antiSMASH[18] have enabled researchers worldwide to perform millions of analyses to explore sequenced genomes for specialised metabolite BGCs.

Subsequently, other approaches to predict A domain substrate selectivity have been developed, including profile Hidden Markov Model (pHMM)-based methods[15,19] and various machine learning algorithms trained on extended, 34-residue signatures, including support vector machines (SVMs)[20,21], random forests[22,23], and ensemble methods that use a combination of these approaches[24]. While methods using these extended 34-residue signatures have a better chance of capturing all residues that are involved in substrate recognition, the available tools deploying them for substrate prediction have a common limitation: a lack of high-quality training data. The SANDPUMA and AdenPredictor algorithms are based on the largest training set of only ~1000 data points[23,24]. NRPSPredictor2, which is still used by the widely used BGC annotation pipeline antiSMASH, was only trained on ~550 data points, with 9 out of the 30 most common substrates being covered by <10 training data points. Also, evolutionarily independent clades of A domains with very different active site architectures can recognise identical substrates, highlighting a need for a large and diverse dataset that covers as many A domain clades as possible. Furthermore, current predictors only consider sequence features of the A domain and do not explore structural features of either the enzyme or the substrate. As such, these algorithms cannot learn from the three-dimensional configurations that ultimately govern substrate selectivity, nor can they use information about substrate similarity. Finally, current predictors were trained on datasets that contained data points labelled with multiple substrates.

Some A domains are naturally promiscuous, particularly those activating large hydrophobic substrates such as L-Phe, L-Leu, L-Ile, and L-Val. However, the ATP-pyrophosphate exchange assay commonly used to probe substrate selectivity has several shortcomings that misleadingly suggest many A domains are promiscuous. First, it does not permit competition between multiple substrates, as encountered by A domains *in cellulo*, to be assessed; only the relative adenylatability of isolated substrates can be measured. Second, it only measures the adenylation reaction, providing no means of determining whether the subsequent thiolation reaction also contributes to substrate selectivity. NRPS condensation domains catalyse peptide bond formation between PCP-domain-bound aminoacyl (and peptidyl) thioesters. It is thus the overall selectivity of an A domain for formation of different aminoacyl thioesters that primarily dictates the composition of NRPS products. Current state-of-the-art algorithms are not true multi-label predictors, representing the selectivity of "promiscuous" A domains as unique multi-substrate strings corresponding to unique specificities, rather than substrate sets.

Here, we use sequence comparisons and AlphaFold models of thousands of A domains to reveal widespread independent evolution of domains with similar substrate selectivity. Motivated by our observations and powered by large-scale training data curation, we developed two fast and accurate machine learning frameworks: PARAS (Predictive Algorithm for Resolving Adenylation domain Selectivity) and PARASECT (Predictive Algorithm for Resolving Adenylation domain Selectivity by featurising Enzyme and Compound in Tandem). These algorithms overcome challenges with previously developed predictors, using an expanded and manually curated training set three times the size of previous datasets,

providing options for structure-guided feature extraction, integrating substrate features, and accommodating reliable multi-label predictions. We benchmark PARAS and PARASECT against AdenPredictor, SANDPUMA, and NRPSPredictor2, showing a substantial increase in predictive accuracy on independent holdout data (by 20%, 27%, and 22%, respectively). Feature importance analysis shows how different residues are predictive for specificity of A domains accepting different substrates, but also for independently evolved clades of A domains accepting the same substrates. Subsequently, we showcase how PARAS and PARASECT accurately predict L-Trp as the substrate of an unusual A domain in a novel NRPS shown to assemble metabolites related to the immunomodulator tryptopeptin A. Intact protein MS analyses of aminoacyl thioester formation on the downstream *holo*-PCP domain associated with this A domain confirm it has high specificity for L-Trp and highlight the importance of competition experiments for developing an accurate picture of physiologically relevant A domain substrate selectivity.

Results

An expanded and highly curated training set of A domain specificities

The performance of predictive models largely relies on the size of the dataset on which they are trained. For this reason, we collated an A domain dataset from three sources: MIBiG 3.0[25], the SANDPUMA/AdenPredictor (SP/AP) training set[23,24], and the NRPSPredictor2 training set[21]. The resulting combined dataset counted 3,254 A domains from diverse phyla (Figure 1A), recognising 252 different substrates (Figure 1B). These data were extensively curated and validated prior to model training (see Online Methods). 381 domains exclusively recognised substrates for which there were too few data points available (n < 11) to reliably validate model performance. The remaining 2,873 unique A domains (34 substrates for which n >= 11; Figure 1B) were used to train models for validation, which accounts for 2.6 and 5.4 times as many labelled data points as were used for training SP/AP (1089 data points) and NRPSPredictor2 (534 data points), respectively. Upon curation based on detailed phylogenetic and literature analysis, we discovered and corrected misannotations in the SP/AP and NRPSPredictor2 datasets (137 [12.6%] in SP/AP, 31 [5.8%] in NRPSPredictor2), which will likely have impacted the performance of these models.

Metadata were available from the MIBiG database for a subset of these A domains, providing i) the evidence that was used to obtain each substrate annotation, and ii) A domain selectivity classified as major and minor substrate(s). The majority of annotations were based on structural inferences of the peptide scaffolds produced *in vivo*, and less frequently by high levels of activation (>80% activation compared to the predominant substrate) in A domain specificity assays, such as ATP-pyrophosphate exchange[25]. Of the 2,660 A domains with available metadata, 111 (~4%) were elucidated using ATP-pyrophosphate exchange assays. For promiscuous A domains with multiple major substrates, this proportion was substantially higher (17/126; ~13%). For promiscuous A domains with one or more minor substrates in addition to their major substrate, this amounted to an even greater proportion of ~28% (29/105; Figure 1C). These differences suggest that A domain promiscuity is overestimated by the ATP-pyrophosphate exchange assay, which can report false positives (as discussed in the introduction), with many activated substrates never being transferred to the cognate PCP domain *in cellulo* and therefore never being incorporated into the NRP scaffold. Thus, we chose to remove all minor substrates from our dataset.

Large-scale analysis of structural models reveals independent evolution of A domain substrate specificity

We next investigated other shortcomings of current state-of-the-art A domain predictors. We observed that they disproportionately made errors in the classification of large proteinogenic amino acids such as phenylalanine, tryptophan, and lysine[21,23] (F1-score of 0.688, 0.320 and 0.400 in NRPSPredictor2, respectively; most common mispredictions in AdenPredictor). We hypothesised that due to the size and degrees of rotational freedom of the side chains of large substrates, parallel evolution might have given rise to A domain active site pockets that select identical substrates using different active site architectures.

To explore this hypothesis, we constructed AlphaFold models for all 3,254 A domains and explored the diversity of their active site pockets using principal component analysis on their three-dimensional voxel grids (see Online Methods). We observed that A domains selecting substrates for which current predictors perform well (e.g., threonine, and serine), cluster together in one or a few major groups, indicating high active site similarity (Figure 2A). In contrast, A domains selecting large substrates (e.g., phenylalanine, tryptophan, and lysine), were predicted to have divergent active site pockets (Figure 2B), suggesting multiple active site architectures for these substrates, possibly resulting from parallel evolution. This was further supported by phylogenetic analysis of the 34 amino acid extended active site signatures, which demonstrates that A domains recognising large substrates are dispersed across multiple different monophyletic clades, while 41% of Thr-incorporating A domains fall into a single monophyletic clade and 75% of Thr-incorporating A domains fall into the largest five clades (Figure 2C).

To visually illustrate the nature of active site pocket differences that may be observed between A domains in phylogenetically distinct branches, we compared the predicted active sites of two Trp-incorporating A domains. We chose the A domain from Qui18 (GenPept accession AET98916.1; Qui18-A1), an NRPS involved in quinomycin biosynthesis in *Streptomyces griseovariabilis subsp. bandungensis*[26]; and the third A domain from BreC (GenPept accession ATY37608.1; BreC-A3), an NRPS involved in the biosynthesis of brevicidine in *Brevibacillus laterosporus*[27]. The 10-residue active site signatures of these A domains are highly divergent: DAWTVTGVGK for Qui18-A1 and DPTQAGEVVK for BreC-A3, only sharing three common residues, two of which are the highly conserved Asp and Lys residues that form salt bridges with the backbone ammonium and carboxylate groups in the majority of A domains. Their extended 34-residue active site signatures, which are used by NRPSPredictor2, SANDPUMA, and AdenPredictor, are also highly dissimilar, with only 10 out of 34 residues in common and a pairwise alignment bitscore of 31.0, compared to a score of 184.0 for the two most similar tryptophan-recognising A domains in our dataset. The distance between both domains in a principal components analysis (PCA; Figure 2B), as well as docking of Trp-adenylate (TYM) to the AlphaFold models for both A domains predicts that these differences are manifest in their 3D structures, further suggesting A domains can bind the sides chains of large substrates using architecturally diverse pockets (Supplementary Discussion 1; Figure 2D, 2E). This underscores the need for a sufficiently varied training dataset to cover A domains spanning different phylogenetic branches.

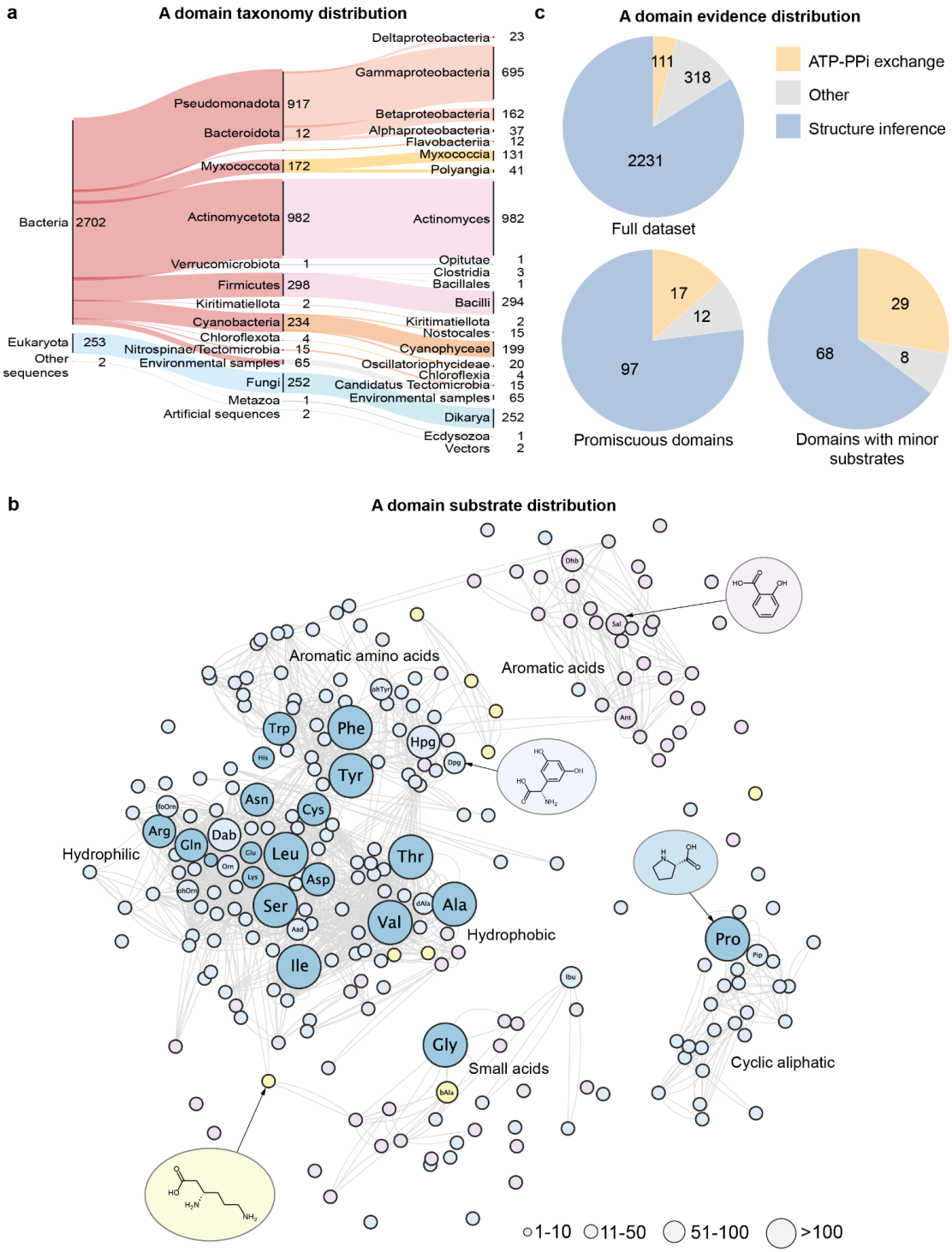

Figure 1. Overview of A domains used for training PARAS and PARASECT. a. Alluvial diagram showing the taxonomic distribution of all A domains for which taxonomy was obtained. **b.** Tanimoto similarity network (cutoff=0.46) of substrates activated by the A domains in the dataset. Node size represents the number of times a substrate is present in the dataset. Blue: α-amino acids. Dark blue: proteinogenic amino acids. Yellow: β-amino acids. Pink: Other acids. The 34 substrates labelled with abbreviations are the substrates used for training validation and benchmarking models. **c.** Proportion of A domains annotated using ATP-pyrophosphate exchange assays across the full dataset, A domains with multiple major substrate specificity annotations, and A domains with minor substrate specificity annotations.

Aad: 2-aminoadipic acid. Dab: 2,4-diaminobutyric acid. Ibu: 2-aminoisobutyric acid. bAla: β-alanine. Dhb: 2,3-dihydroxybenzoic acid. dAla: D-alanine. Orn: ornithine. foOrn: *N5*-formyl-*N5*-hydroxyornithine. ohOrn: *N5*-hydroxyornithine. Hpg: 4-hydroxyphenylglycine. Dpg: 3,5-dihydroxyphenylglycine. ohTyr: (*R*)-β-hydroxytyrosine. Pip: pipecolic acid. Ant: anthranilic acid. Sal: salicylic acid.

Improving A domain substrate prediction: structure-based approaches and substrate featurization

In addition to increasing the size of our training set, extensive data curation, and eliminating unreliable annotations of substrate promiscuity, we explored two avenues for improving A domain substrate prediction: including featurising the A domain and substrate in tandem to allow multi-label predictions (Figure 3a,c) and using protein-structure-based features (Figure 3b).

Prior to training all models, we made two train-test splits: one by stratifying on substrate class, and the other by stratifying on phylogeny (as defined by bitscore clustering). The latter enables us to assess how well our algorithms will perform on sequences that are phylogenetically distinct from any training data. The average protein percent identity of a test data point to its closest relative in the training set was 75%±17% for the class-based split and 58%±16% for the phylogeny-based split. In both cases, the ratio of training data to test data was approximately 3:1 (Figure S1). With our expanded training set, we explored how different feature selection and extraction methods affect model performance.

For feature selection, we investigated the amino acid sequence features also used by NRPSPredictor2, AdenPredictor, and SANDPUMA (Online Methods); structural A domain features derived from 3D AlphaFold 2 models (Online Methods); and Morgan fingerprints[28] of the substrates, which convert chemical structures into bitvectors representing their substructures. For both class-based and phylogeny-based train-test splits, we trained random forest models based on all possible feature combinations (Figure S2). These fell into two model categories: (i) those that only use an A domain as input and predict its most likely substrate (PARAS), and (ii) those that use both an A domain and a putative substrate as input and predict the probability that the A domain selects that substrate (PARASECT).

Importantly, PARAS is only trained on the first-listed substrate, which always corresponds to the substrate incorporated into its validated natural product. Therefore, PARAS does not suffer from potential inaccuracies in the annotation of promiscuous A domains that arise from ATP-pyrophosphate exchange assay data. In contrast, PARASECT was trained on all major substrates listed for each A domain, and is therefore more sensitive to these inaccuracies. We limited the likelihood of such misannotations impacting PARASECT model performance by excluding minor substrates, which reduced the percentage of promiscuous annotations relying on ATP-pyrophosphate exchange data to 13%.

Structure-based feature selection and extraction provide slight improvements at the expense of speed

Because the shape of an A domain's active site determines its substrate selectivity [10,11,29], we expected that models based on structural features would perform well. However, we observed that models trained on sequence features substantially outperformed those trained on structural features. Using both sequence and structural features did not significantly improve performance compared to using sequence features only (Figure S2), suggesting that there is little extra information contained in the structural features that the model cannot interpret from sequence features. However, feature importance analyses show that in models combining sequence and structural features, many of structural features were among the most frequently used in the model, indicating that certain structural information can be informative (Supplementary Discussion).

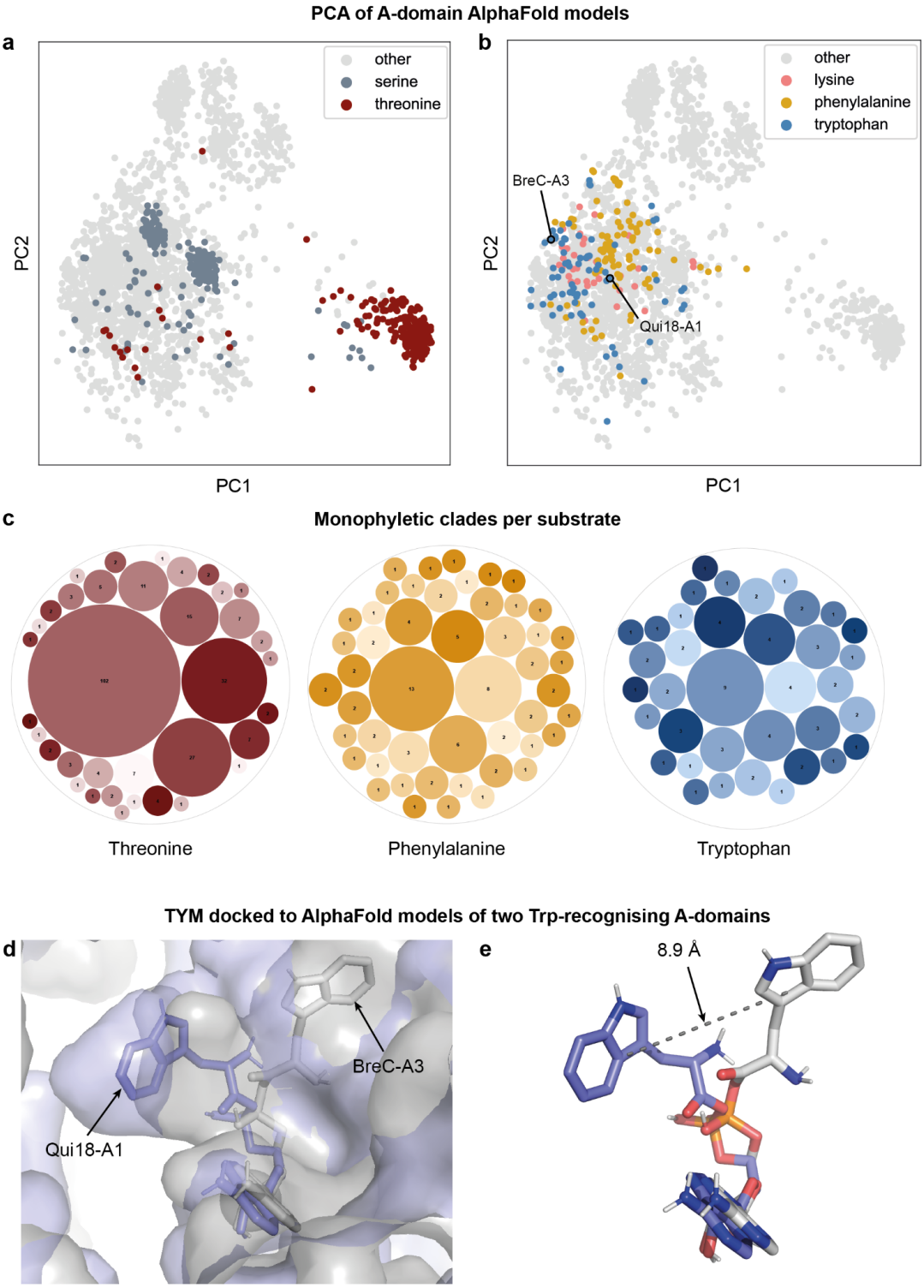

Figure 2. Large-scale structure and sequence analysis of A domain active site diversity. a. PCA of 3D A domain active sites based on 3254 AlphaFold models, with those incorporating the polar substrates Ser and Thr highlighted, and **b.** A domains incorporating the large substrates Phe, Trp, and Lys highlighted. Trp-incorporating domains Qui18-A1 and BreC-A3 are indicated. **c.** Bubble representation of the number per monophyletic clade for Thr, Trp, and Phe-incorporating A domains, based on a phylogenetic tree of the 34 amino acid extended active site signatures. Each bubble represents one monophyletic clade, with the size of each bubble proportional to the number of A domains in the monophyletic clade. **d.** Comparison of TYM docked to AlphaFold-modelled Qui18-A1 (blue) and BreC-A3 (grey). The left pocket is smaller in BreC-A3 (grey) than in Qui18-A1 (blue), making it unlikely for the tryptophan substrate to fit into this pocket in BreC-A3. Instead, the tryptophan moiety of TYM docked to BreC-A3 localises to a pocket on the right, leading to a 8.9 Å shift (d; dashed line) of the indole C3 atom in the bound TYM relative to the Qui18-A1 docked structure.

In addition to model accuracy, runtime is another important consideration for an A domain predictor: SANDPUMA was implemented in antiSMASH 4.0 and subsequently removed in antiSMASH 5.0 due to a computational cost of 2-5 minutes per A domain. The inclusion of structural features in PARAS and PARASECT would increase their runtime by ~20 minutes per A domain, as AlphaFold models would have to be constructed prior to running the model. As structural features did not substantially improve sequence-based PARAS and PARASECT models, which take only a few milliseconds per prediction, we opted for sequence-based models for benchmarking.

We also examined structure-informed methods that do not require the modelling of A domains that the user wants to query, including structure-guided profile alignment for active site extraction. Compared to extraction using sequence-based profile alignments, as used by SANDPUMA, this method yielded active sites with fewer gaps (Figure S3), which in turn led to models with slightly increased performance (Table S1). However, extracting the active sites using the A domain pHMM from NRPSPredictor2 is much faster (milliseconds rather than seconds per domain), and led to even fewer gaps overall (163 vs 673). Comparison of different active site extraction methods is discussed in more detail in our Supplementary Discussion.

For 35 A domains, structure-guided profile alignment yielded active sites with fewer gaps than pHMM-based extraction. Of these, 17 recognise non-amino acid substrates, 2 tether aspartate to the PCP domain by the side chain carboxyl group, and 14 belong to kingdoms and phyla that are underrepresented in our dataset, including Fungi, Myxococcota, Cyanobacteriota, and various uncultured bacteria from environmental samples. This suggests that structure-guided profile alignment might be better suited for extracting the active sites of A domains recognising non-amino acid substrates and of A domains that are phylogenetically distinct from those in our dataset. For example, for fungal A domains, Heard and Winter demonstrated that structural modeling enhanced substrate preference prediction[30]. The limitations of current predictors for fungal A domains or non-amino acids are understandable, as the pHMM developed by Rausch *et al.*[20] was based on a phylogenetically restricted set of A domains that recognise amino acid substrates. Therefore, we added an option to our command line models and our web-page to select either method for sequence feature extraction (Figure S4).

PARASECT allows predicting multiple or previously unseen A domain substrates

Our PARASECT models provide a novel method of A domain selectivity prediction: rather than solely using A domain features to predict a substrate, PARASECT considers both the A domain and the structure of the substrate and predicts if the domain is compatible with that substrate. This has a number of advantages.

First, the model can learn from the similarity and differences between substrates during training, e.g., how the presence of an aromatic ring in the substrate combined with hydrophobic residues in an A domain active site affects selectivity. We observe that PARASECT's top predictions (i.e., those substrates with the highest predicted interaction probabilities with an A domain) are often structurally similar, in contrast to PARAS, indicating that the PARASECT model indeed incorporates substrate similarity to make classifications. We confirmed this with feature inference analysis of our cross-validation models, which showed six substrate features as the most informative in the model (Figure S5). Also, as multiple top predictions may be biologically relevant, PARASECT provides a means to obtain a range of substrate predictions for an A domain among which the true substrate is likely present.

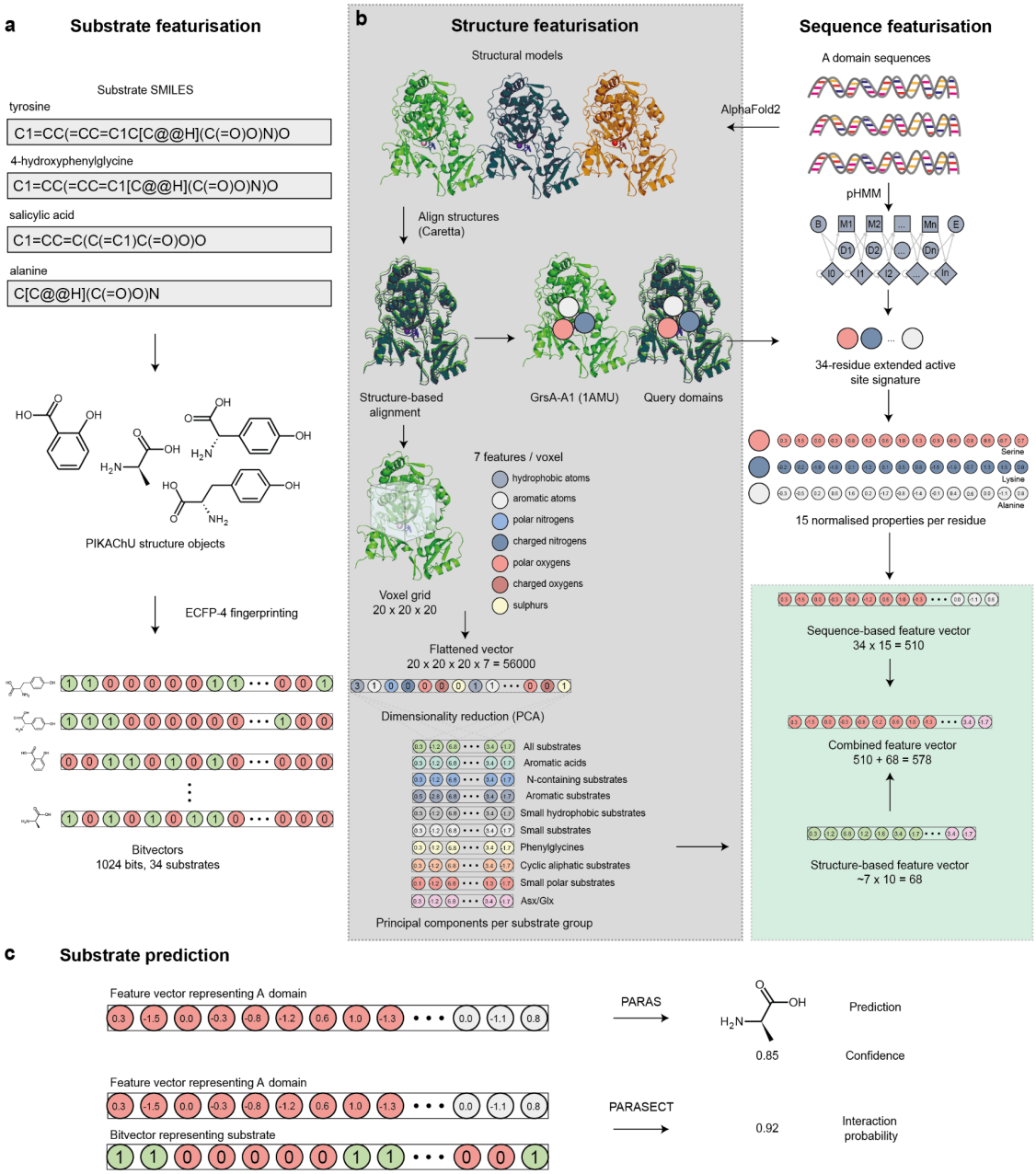

Figure 3. Overview of PARAS and PARASECT workflow. **a.** Feature extraction for substrates, used in PARASECT models. SMILES strings of substrates are first processed and converted into PIKAChU structure graphs. Then, Extended Connectivity Fingerprinting (ECFP-4) is used to generate bit vectors that represent the substrate structure. **b.** Feature extraction for sequence-based, structure-based and combined models. In structure featurisation, structural models are generated for A domain sequences using AlphaFold2, and the resulting structural models are aligned using structure-based alignment. Then, either sequence-based features can be extracted from this structure-guided alignment based on the alignment with the reference GrsA-A1 crystal structure (then proceeding again with the sequence-based workflow), or structure-based features can be extracted using a 3D voxel grid, where physicochemical properties of coordinates across the substrate-binding pocket are encoded before reducing their dimensionality with principal component analysis. In sequence-based featurisation, sequence-based features are directly extracted through alignment to a profile Hidden Markov Model (pHMM) along with the GrsA-A1 reference sequence to extract a 34-residue active site signature. For each of the 34 residues, 15 normalised physicochemical properties are computed and encoded in a feature vector. Sequence-based feature vectors can either be used on their own (as by default, to optimise speed) or combined with structure-based feature vectors. **c.** PARAS and PARASECT input and output. The PARAS algorithm is a multi-label classifier that predicts a single substrate for a query A domain, with a given confidence level. In contrast, PARASECT computes interaction probabilities for each and every substrate separately, and these can then be ranked to identify the most likely substrate(s).

Second, PARASECT can be queried with substrates not present in the model's training data. For example, we used PARASECT to predict the interaction between the A domain in module 6 of the calcium-dependent antibiotic (CDA) NRPS (MIBiG accession BGC0000315), which canonically incorporates 4-hydroxyphenylglycine. This domain is also able to incorporate 4-fluorophenylglycine in mutasynthesis experiments[31]. PARASECT predicted a high interaction probability of 75.5%, returning 4-fluorophenylglycine as the second most likely substrate to be incorporated after 4-hydroxyphenylglycine (Figure S7). Notably, PARASECT's training set does not include fingerprints of substrates containing F atoms. This demonstrates a potential application of PARASECT in targeting A domains for variant creation via precursor-directed biosynthesis. To make this feature available to users, we have included an option to upload substrates in SMILES format[32] in our web application.

Finally, PARASECT can train on and make predictions for promiscuous A domains that incorporate multiple substrates, because the model is trained on and returns predictions for A domain/substrate pairs. The general applicability of this feature greatly depends on the quality of annotations for promiscuous A domains. While we aimed to reduce the number of false positives by excluding minor substrates from our dataset, generating less biased and higher quality data in the future will facilitate more reliable A domain promiscuity predictions.

PARAS and PARASECT outperform existing predictors

To demonstrate the improvement of PARAS and PARASECT over other A domain selectivity predictors, we benchmarked both tools against NRPSPredictor2, AdenPredictor, and SANDPUMA. As SANDPUMA is an ensemble of multiple predictive algorithms, we also assessed these individually. In these tests, we used data that did not occur in the training sets of either SP/AP or PARAS/PARASECT (470 and 506 A domains for class-based and phylogeny-based stratification, respectively).

PARAS and PARASECT substantially outperformed all other algorithms for both class-based (Figure 4a,b) and phylogeny-based (Figure S8) train-test splits. For class-based splits, we observed accuracies of ~91% for PARAS and ~87% for PARASECT compared to the next best algorithm NRPSPredictor2, which has an accuracy of ~59% when 'no call' is considered an incorrect prediction. While PARAS/PARASECT models trained on phylogeny-based splits performed ~12% worse, PARAS still outperforms the next-best algorithm AdenPredictor with +~20% accuracy on this dataset (Figure S8).

While many current A domain specificity predictors enforce a cutoff to only return highly confident predictions, we chose not to implement such a cutoff for PARAS. As PARAS already performs very well at low confidence thresholds and many correct predictions would be missed when implementing a threshold (Figure 4c), we felt the community would benefit more from our raw model. Still, PARAS and PARASECT always return a confidence score alongside their predictions so that the user can consider this in the context of the biological system they are studying.

When querying PARASECT with sequences distantly related to the training data (e.g., from less-well-studied organisms), it is important to note that its feature to provide prediction probabilities for multiple possible substrates can come in very useful: in the phylogeny-stratified dataset, when the top two predictions are considered, the accuracy of the model improves by nearly 10% (Figure 4d). However, PARASECT frequently gives high interaction scores to substrates structurally similar to the actual substrate, even when these other substrates are not selected by the A domain (Figure S6), leading to a slight drop in overall performance compared to PARAS (Figure 4a). At the same time, PARASECT does also allow

correct identification of real substrate promiscuity: out of 19 promiscuous A domains in our class-stratified test set, for 12 the top 'x' PARASECT predictions exactly matched the set of substrates incorporated by the A domain, where 'x' is the number of different substrates incorporated. For the remaining 7, there was at least 1 substrate overlap between the top 'x' PARASECT predictions and the actual substrate selectivities (Figure 4e).

We also assessed the substrate-specific prediction accuracy for each of the 34 substrates that PARAS and PARASECT were trained on. As our computational modelling and docking indicated that active site architectures of domains recognising large substrates may vary substantially across phylogenetic clades, we were particularly interested in performance differences for large substrates. Comparing the test sets obtained through class-based and phylogeny-based stratification (i.e., testing on sequences relatively closely versus distantly related to the training data), we indeed observed a large difference in performance across these two datasets for various large substrates, including L-Arg, L-Trp, and L-Phe. While PARASECT correctly classified 80-85% of the data points for all three substrates in the class-based test set, this dropped to around 50% in the phylogeny-stratified test set (Figure S9). Other hard-to-predict substrates were underrepresented ones (such as pipecolic acid, D-Ala, and 2-aminoadipic acid) as well as L-Val and L-Ile, whose active sites frequently clade together. Similar trends are visible when comparing substrate-specific performance of PARAS and NRPSPredictor2. Yet, particularly for L-Lys, L-Trp and L-Phe, PARAS, with its much larger and curated dataset, predicts these substrates with considerably better accuracy (Figure 4f). This illustrates the advantages of a large and diverse training set to cover as much phylogenetic breadth as possible.

After model validation, cross-validation, and benchmarking, we used our full dataset to train three sequence-based models that we made available to the community on our web application (paras.bioinformatics.nl): a PARAS model and a PARASECT model trained on all 34 substrates for which at least 11 examples exist in our dataset; and an additional PARAS model trained on all 3,254 A domains. These models will also be available through direct outlinks in the version 8 release of antiSMASH.

Characterisation of an unusual tryptophan-incorporating A domain in the tryptopeptin NRPS

To further validate PARAS and PARASECT, particularly in the context of A domains incorporating amino acids with large hydrophobic side chains, we selected tryptopeptin A[33], a biosynthetically uncharacterised member of the peptidyl epoxyketone family of proteasome inhibitors with an unusual tryptophan-derived α, β-epoxyketone pharmacophore. We hypothesised that tryptopeptin is assembled following similar biosynthetic logic to eponemycin and TMC-86A[34,35]. Thus, an NRPS would catalyse the successive condensation of a 3-methylbutyryl thioester with L-Val, L-*allo*-Thr and L-Trp to give rise to an N-acyl-tripeptidyl thioester, followed by polyketide synthase-catalysed net elongation of the Trp residue with propanoate, and conversion of the resulting α-methyl-β-ketoacid to the corresponding epoxyketone by a flavin-dependent synthase.

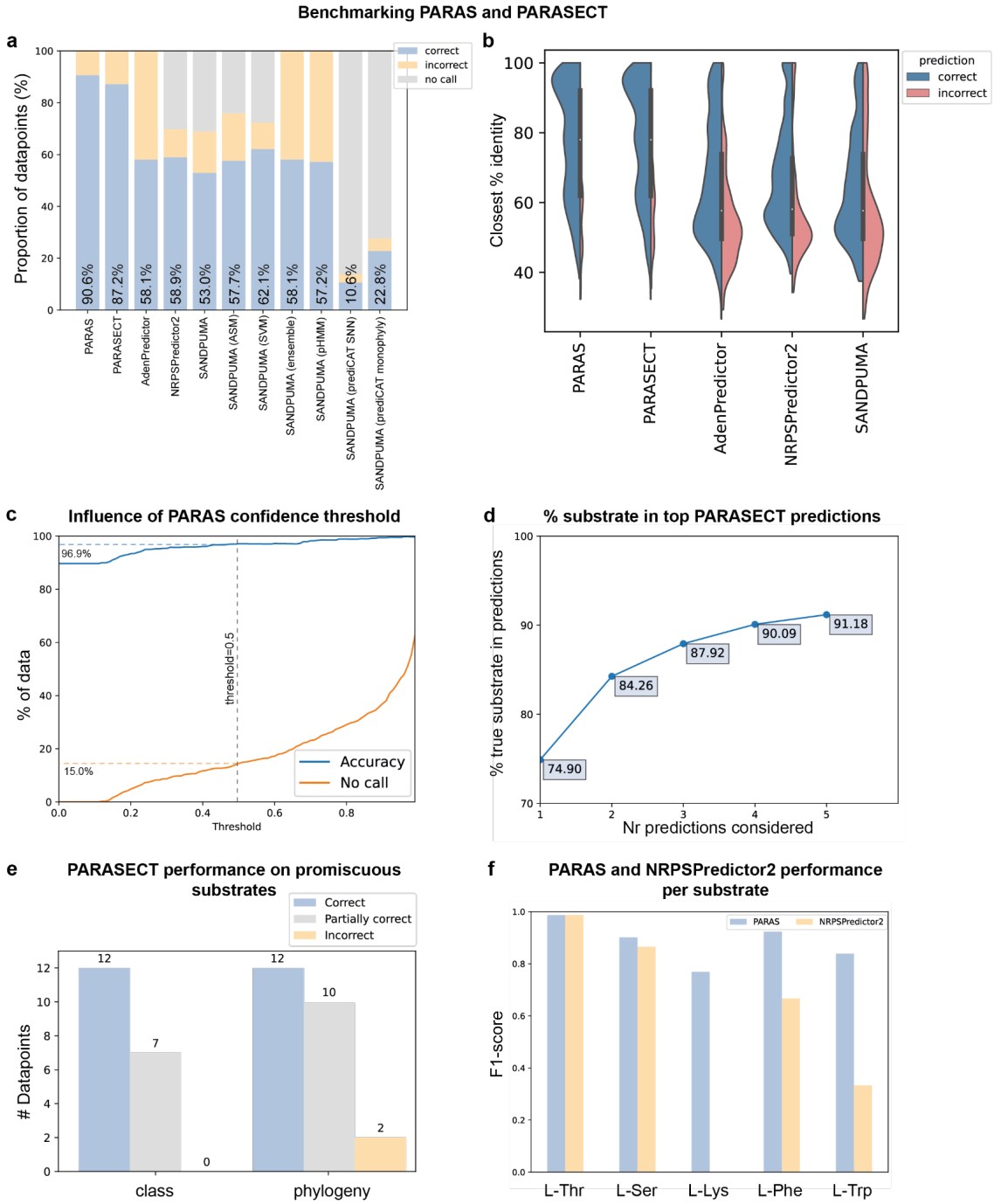

Figure 4. Benchmarking PARAS and PARASECT. a. Overall performance of sequence-based PARAS and PARASECT models trained on class-stratified training data compared to other algorithms. PARAS performs ~32% better than the next best performing algorithm, NRPSPredictor. **b.** Violin plot comparing performance of PARAS and PARASECT to other algorithms for data points at different levels of sequence identity to the training set. **c.** PARAS accuracy and number of data points for which no prediction was made across different confidence threshold cutoffs. **d.** % of times PARASECT trained on a phylogeny-stratified dataset predicts the correct substrate among its top 1-5 predictions. **e.** PARASECT performance on promiscuous substrates. For data points labelled as 'correct', the substrates selected by the A domain and the top x substrates predicted for that domain, where x is the number of substrates selected by that domain, match exactly. 'Partially correct' data points are A domains for which there is an overlap of at least one substrate between the actual x substrates and the top x predicted substrates. For data points labelled 'incorrect', there is no overlap between the actual substrates and the top x predicted substrates. **f.** Per-substrate performance of PARAS compared to NRPSPredictor2, using the F1-score as metric (i.e., the harmonic mean of precision and recall). No or marginal improvement is observed for substrates that clade together, such as L-Thr and L-Ser, while performance increases substantially for large substrates such as L-Lys, L-Phe, and L-Trp.

Using ClusterTools[36], we searched the Prokaryotic RefSeq Representative Genomes repository at NCBI for gene clusters containing a homologue of the gene encoding the eponemycin epoxyketone synthase EpnF (protein ID AHB38508.1) and genes encoding at least one A, C, and ketosynthase domain. Among the hits, a BGC (GenBank accession: NZ_MAXF01000131) in the draft genome sequence of *Streptomyces sparsogenes* ATCC 25498 (DSM 40356) contained genes encoding a putative epoxyketone synthase (TtpC), a trimodular NRPS (TtpD), a PKS module (TtpE) containing a putative C-methyl transferase domain, and a proteasome β-subunit (Figures 5A and 5B). Detailed manual annotation of the enzymes encoded by this BGC indicated that it likely directs the biosynthesis of tryptopeptin-like metabolite(s) because (i) the 9-residue active site signatures of the A domains in modules 1 and 2 of the NRPS in the BGC were predicted to select L-Val and L-Thr, respectively, and (ii) although the substrate of the A domain in module 3 of the NRPS could not be predicted from the 9-residue active site signature, the relative compactness and hydrophobicity of the residues suggested the preferred substrate is a large hydrophobic amino acid (Figure 5B).
No tryptopeptin production was detected in *S. sparsogenes* DSM 40356. We therefore sequenced and assembled the genome of a previously identified tryptopeptin A producer *Streptomyces sp.* maeda85. Analysis of the genome sequence revealed a BGC (GenBank accession: PQ740899) with high similarity to the putative tryptopeptin BGC we identified in *S. sparsogenes* DSM 40356 (Figure 5A). UHPLC-ESI-Q-TOF-MS/MS analyses confirmed that *Streptomyces sp.* maeda85 produces tryptopeptin A (Figure S10).

To verify that the highly similar BGCs identified in the genomes of *S. sparsogenes* DSM 40356 and *Streptomyces* sp. Maeda85 direct the production of tryptopeptin, both were cloned from genomic DNA using our recently reported high efficiency yeast transformation-associated recombination capture vector pCAP1000[35]. The captured BGCs were transferred via conjugation from *E. coli* into *Streptomyces coelicolor* M1154[37]. Analyses of culture extracts from the transconjugants using UHPLC-ESI-Q-TOF-MS/MS identified five tryptopeptin-related metabolites, all containing an *N*-methyl-L-Val-L-(*allo*)-Thr-L-Trp core scaffold (Figure 5C; Figure S11, Supplementary Discussion). Notably, these metabolites were produced in similar relative quantities by strains expressing the two BGCs, neither of which yielded production of tryptopeptin A in *S. coelicolor*.

To confirm that the A domain in module 3 of TtpD incorporates L-tryptophan, we overproduced the corresponding A-PCP di-domain in *E. coli* BL21 and *E. coli* BAP1 as N-terminal His_6 fusion proteins. These strains are expected to produce the *apo* and *holo* forms of the protein, respectively[38]. Production of the *apo*-A-PCP di-domain by *E. coli* BL21 was confirmed by UHPLC-ESI-Q-ToF-MS analysis of the purified protein (Figure S12, top row). The *apo* protein was converted to its *holo* form using Sfp and coenzyme A (Figure S12, second row) and the ability of the A domain to load L-Trp, L-Phe, L-Leu, L-Val, and L-His onto the downstream PCP domain was examined using UHPLC-ESI-Q-ToF-MS (Figure S12, rows 3-7). When the *holo*-A-PCP di-domain was incubated with the substrates in isolation, high levels of L-Trp and L-Phe loading were observed, along with some loading of L-Leu, very little loading of L-His and no loading of L-Val. However, when incubated with a mixture of L-Trp, L-Phe and L-Leu, the *holo*-A-PCP di-domain exclusively loaded L-Trp (Figure 5D), suggesting that this is the preferred substrate. We validated this by purifying the A-PCP di-domain from *E. coli* BAP1. UHPLC-ESI-Q-ToF-MS analysis showed this is a mixture of the unloaded and Trp-loaded *holo*-protein (Figure 5D), confirming L-Trp is the sole substrate loaded onto the PCP domain *in cellulo*.

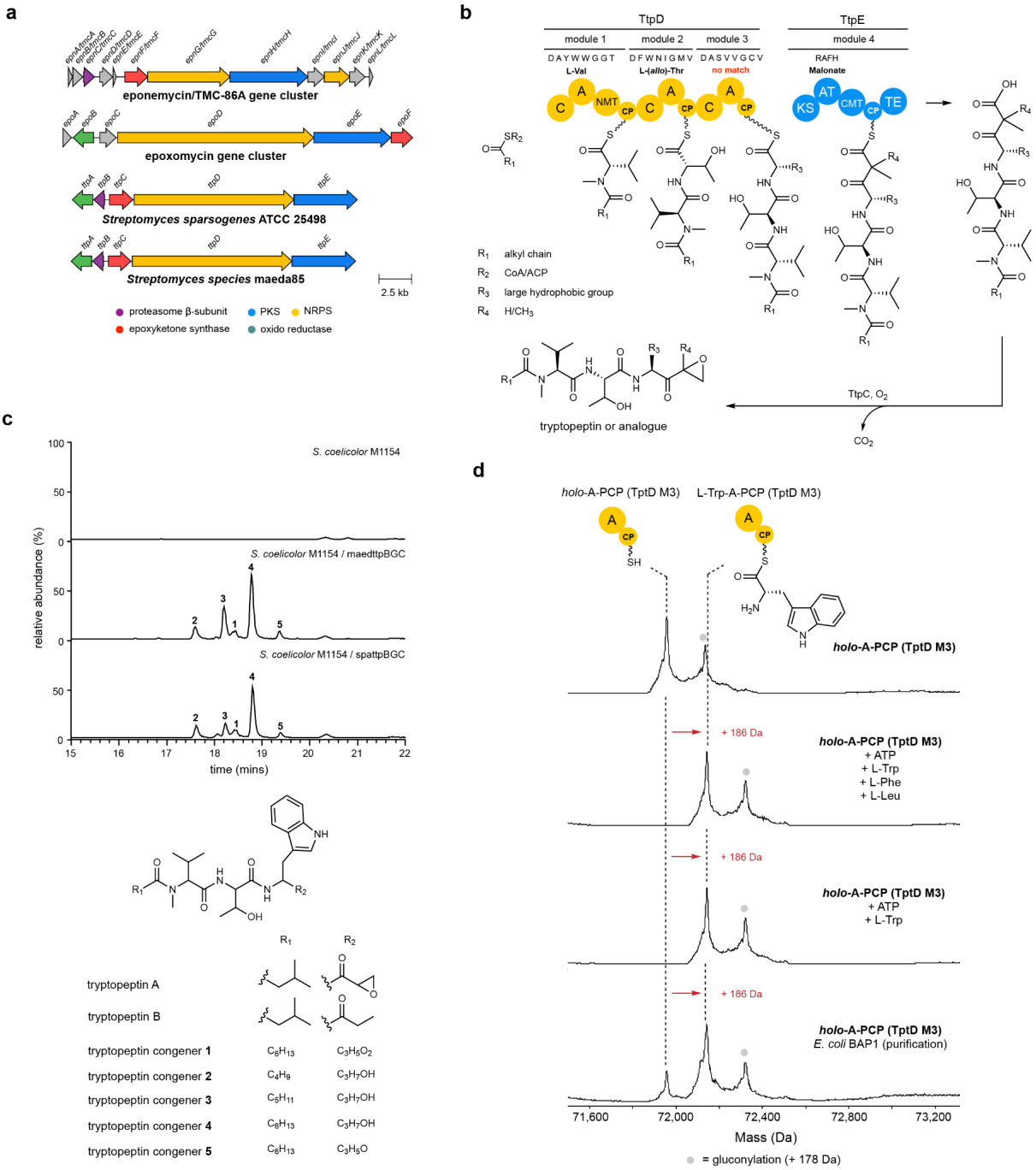

Figure 5. Experimental validation of the tryptopeptin biosynthetic gene cluster. a. Comparison of putative tryptopeptin BGCs in *Streptomyces sparsogenes* and *Streptomyces* sp. maeda85 with the eponemycin and epoxomycin BGCs. **b.** Proposed biosynthetic pathway for tryptopeptin A. **c.** Extracted ion chromatograms (EICs) from UHPLC-ESI-Q-TOF-MS analyses of culture extracts from *S. coelicolor* M1154, *S. coelicolor* M1154 containing pCAP1000maedttpBGC and pCAP1000spattpBGC corresponding to $[M+H]^+$ for tryptopeptin-related metabolites **1**, **2**, **3**, **4**, and **5**. The proposed planar structures are consistent with MS/MS analyses (Fig S11). **d.** Intact protein MS analysis of the purified *holo*-A-PCP di-domain from module 3 of TtpD incubated with an amino acid mixture, L-Trp, or loaded *in cellulo* using *E. coli* BAP1 cells, demonstrating that TtpD-A3 preferentially loads L-Trp *in vitro* and *in vivo*.

While we could use the broad selectivity hypothesised by manual inspection of the active site of the A domain in module 3 of TtpD to link the BGC in both strains to production of tryptopeptin-related metabolites, this example demonstrates the need for selectivity predictors that can handle A domains with active site signatures that are divergent from previously characterised examples. Because PARAS and PARASECT were developed with these types of domains in mind, we were curious to see how our models performed on the A domain in module 3 of TtpD. We therefore predicted the substrate preference of the A domains in TtpD from *S. sparsogenes* DSM 40356 with PARAS and PARASECT, in addition to AdenPredictor, NRPSPredictor2 and SANDPUMA. PARAS and PARASECT both correctly predicted that this A domain incorporates L-Trp, while in contrast, AdenPredictor and NRPSPredictor2 were not able to identify L-Trp as the substrate, instead predicting L-Asp and L-Leu, respectively. The only other algorithm that correctly identified tryptophan was the SVM submodel of SANDPUMA, with the other submodels of this algorithm not able to give a confident prediction. Importantly, the A domain from module 3 of TtpD does not occur in the PARAS and PARASECT training sets. Indeed, PCA-based comparison of the AlphaFold model of this domain to models of other L-Trp-incorporating A domains shows that it is structurally distinct from domains in the training set, possibly indicating yet another binding mode for this substrate (Figure S13). This demonstrates the added value of PARAS and PARASECT over existing predictors and non-machine learning methods for the discovery and de-orphaning of natural products.

Feature inference reveals residues involved in substrate recognition

In addition to being a suitable machine learning solution for small datasets, the random forest models used by PARAS and PARASECT have another advantage: their simple architecture based on combining multiple decision trees make their models interpretable. Therefore, we were able to use feature inference to gather clues about the biological basis for A domain substrate selection. We first looked at features that are used most frequently to make predictions in general (Figure 6A), and then calculated substrate-specific feature importances to assess whether there is a difference in coordinating residues between A domains that recognise different substrates (Figure 6B-F). Using our sequence-based PARAS model for feature inference, we visualised feature importance three-dimensionally by highlighting the residues in AlphaFold models of representative A domains with manually docked substrates.

Overall, we found that the residues close to the active site were the most informative (Figure 6A). The six most informative residues (residues 7, 10, 13, 17, 22, and 31 of the extended active site signature) are all also part of the 10 residue active site signature as defined by Stachelhaus *et al.*[11], and all but one (residue 31) belong to the 9 residues described by Challis *et al.*[10]. These residues are positioned more closely to the active site pocket than other residues in our analysis, providing some evidence that the model is looking at biologically relevant features to make predictions. Still, the model uses many features derived from residues outside of the 10 residue active site signature, indicating that these residues are also informative. Possibly, the residues just outside the active site pocket may influence the position of residues that point inside of it. For example, an A domain that selects L-Glu may have aromatic ring stacking of one phenylalanine and two tryptophan residues. While only one of the tryptophans directly borders the active site pocket, all three residues are highly predictive for A domains that select glutamic acid (Figure S14). Also, residues not directly bordering the active site could carry a phylogenetic signal that the model uses for its predictions.

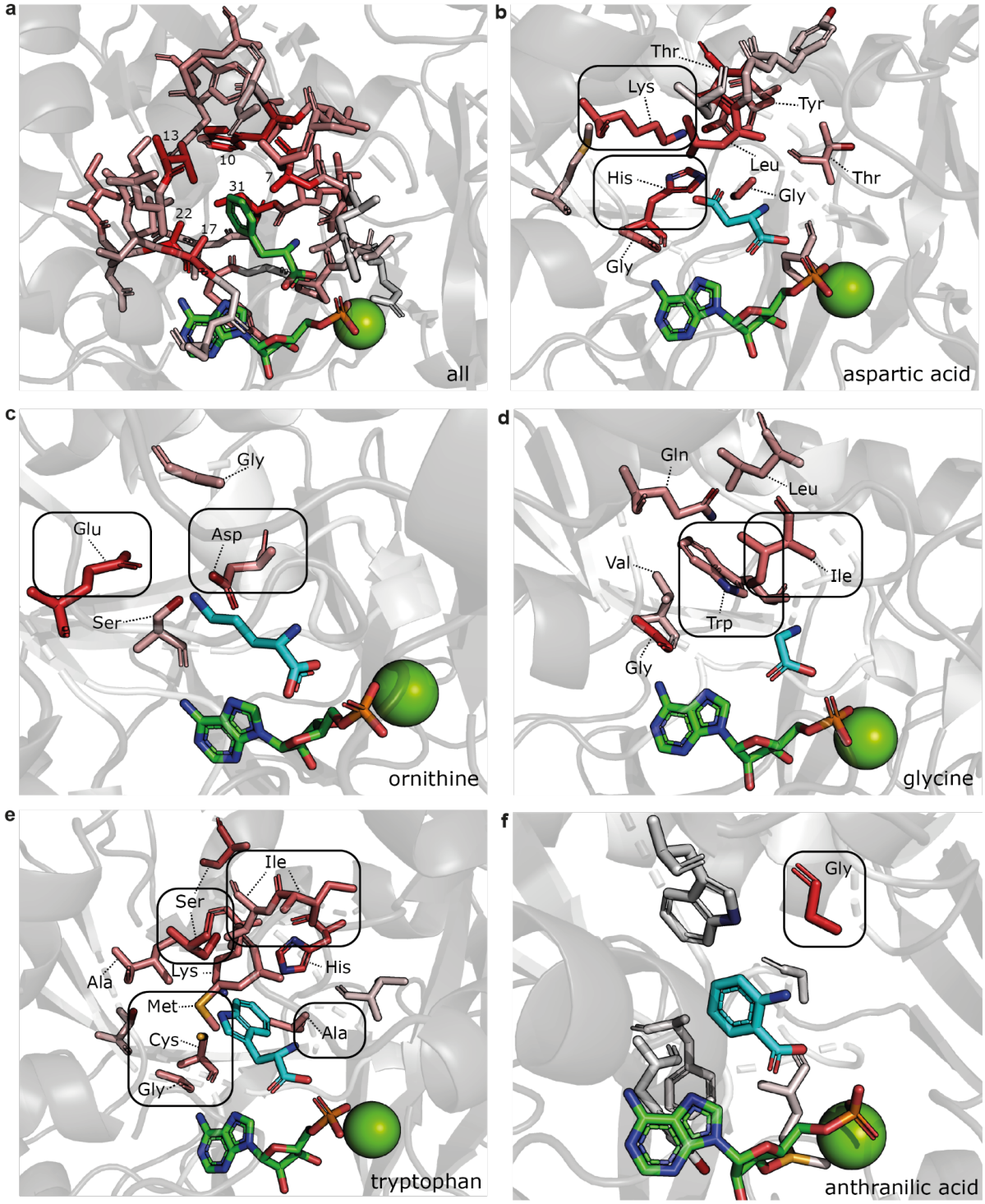

Figure 6. Three-dimensional visualisation of PARAS feature importance analysis. Feature importances are visualised on AlphaFold models of representative A domains. Residues with an importance greater than zero are shown. The redder the residue, the more informative the features that describe that residue. Substrate carbon atoms are shown in cyan; AMP carbon atoms and Mg^{2+} are shown in green. **a.** Overall per-residue feature importance. The six most informative residues are labelled with their extended active site residue number. **b.** Feature importance for aspartic acid-selecting A domains, with positively charged lysine and histidine residues indicated. **c.** Feature importance for ornithine-selecting A domains, with negatively charged aspartic acid and glutamic acid residues indicated. **d.** Feature importance for glycine-selecting A domains, with large hydrophobic isoleucine and tryptophan residues indicated. **e.** Feature importance for tryptophan-selecting A domains, with hydrophobic and aromatic residues indicated. **f.** Feature importance for anthranilic acid-selecting A domains, highlighting the highly conserved residue 6, which is an aspartic acid in most NRPS A domains.

Substrate-specific feature importance analysis revealed that the residues that are important for making an A domain selectivity prediction depend on the substrate that an A domain recognises, suggesting that the residues involved in the active site are different for each specificity. Largely, the residues deemed important by the model are congruent with what one would expect biologically: positively charged lysine and histidine residues for aspartic acid (Figure 6B); negatively charged aspartic acid and glutamic acid residues for ornithine (Figure 6C); large hydrophobic residues restricting the size of the active site pocket for glycine (Figure 6D); small hydrophobic residues leaving open a large pocket for tryptophan (Figure 6E); and a non-aspartic acid residue at the highly conserved position 6 for anthranilic acid, which lacks an amino group in its backbone and therefore does not require aspartic acid for substrate stabilisation (Figure 6F). Figures showing feature importances for all 34 different substrates can be found in the Supplementary Material (Figures S14j-S19). While some of the informative residues overlap between substrates, many do not: for example, the negative charge required in active site pockets selecting glutamic acid and aspartic acid is conferred by a different set of residues for each (Figure S14). As such, simple sequence similarity measures against a one-size-fits-all 'specificity-conferring code' as once hypothesised cannot accurately describe or predict A domain selectivity, demonstrating why machine learning is necessary for solving this challenge.

Our feature importance analysis also shed light on limitations of previous algorithms in predicting large substrates: to make predictions for A domains recognising large substrates, PARAS uses more features on average than for most other substrates (Figure 6E, Figure S16). The same was true for smaller substrates which do not clade together. This supports our hypothesis that A domain specificities have evolved multiple times, and explains why our expanded dataset greatly improved model performance for these substrates.

Conclusions and Future Perspectives

A domain specificity prediction has historically been challenging for those A domains that do not clade according to the substrates that they recognise. In this work, we used large-scale analysis of over 3000 AlphaFold models to illustrate that this challenge is especially prominent for large substrates, likely due to a higher degree of rotational freedom in their respective A domain active sites.

To tackle this issue, we leveraged a larger and better curated dataset, evaluating structure-based approaches and introducing substrate featurization methods to build PARAS and PARASECT: two novel A domain selectivity predictors which offer substantial improvements compared to current state-of-the-art tools, showing a performance increase of 20%-32%. A cross-comparison of our new models and currently available tools, which leverage a wide variety of featurization methods and machine learning algorithms, respectively, clearly demonstrates that data is crucial. AdenPredictor and our best performing PARAS model use identical featurization methods and highly similar machine learning models to make predictions, yet PARAS outperformed AdenPredictor by ~33% on independent holdout data. By far the most substantial difference is the increased size and quality of the PARAS dataset, which is over three times as large as the AdenPredictor dataset and was extensively curated prior to model training. As such, we recommend focusing machine

learning efforts in low-data regimes on data collation, acquisition, and curation, whenever this is feasible.

While the addition of substrate features to PARASECT did not lead to improved model performance, it has the benefit that the model accounts for substrate similarity in its decisions, leading to more informative secondary and tertiary predictions. This makes PARASECT especially useful for obtaining substrate approximations for A domains that are phylogenetically distant from A domains in the training set. Structure-based feature extraction methods have similar use cases, showing improved A domain active site extraction for clades of A domains underrepresented in the training data.

By combining computational and experimental methods, we identified a BGC that directs the production of tryptopeptin-related metabolites in two *Streptomyces* strains. Importantly, the tryptopeptin A scaffold, which includes a critical tryptophan substrate in position 3, was only accurately predicted by PARAS and PARASECT, and not by other models and methods. This, together with feature inference analysis of our models that shows a greater quantity of important features for large substrates, demonstrates that our models successfully address the challenge of large-substrates in A domain selectivity prediction. Also, it is noteworthy that the L-Trp-incorporating A domain of TtpD was structurally different from all other L-Trp-incorporating A domains in our dataset, indicating that A domain diversity remains underexplored, even in well-represented phyla like the Actinomycetota.

To our knowledge, heterologous expression of A domains in *E. coli* BAP1 cells followed by MS to approximate *in vivo* A domain selectivity is a novel and important innovation over existing A domain assaying methods. Critically, this method assays both the adenylation and thiolation half-reactions that are required for A domain substrate selection, in stark contrast to the ATP-PPi exchange assays that are still a popular method for substrate determination. In addition, the bias that is introduced by researchers choosing which substrates to assay is removed; instead, the assayed A domain has access to the full diversity of substrates that are naturally available in *E. coli* BAP1, providing a means to perform large-scale competition assays including many substrates at once. Assaying A domains for substrates that are not naturally available to *E. coli* BAP1 can be achieved by altering the growth medium in which the cells are grown. We therefore recommend that this approach be widely adopted by those in the natural products community that study NRPS systems. This will generate high-quality data to be integrated into future iterations of PARAS and particularly PARASECT, which can leverage these data to distinguish between substrate-tolerant A domains and truly promiscuous A domains that incorporate multiple substrates *in vivo*. Combined with feature inference approaches such as those demonstrated here, this can provide critical insights into molecular determinants of A domain promiscuity.

Already, PARAS and PARASECT have been used to discover natural products in the form of novel tridecaptin derivatives[39]. There, PARAS was used to identify two novel lipopeptide antibiotics, tridecaptin A_5 and tridecaptin D, correctly predicting the substrates of all A domains in the NRPS enzymes TriD and TriE, while NRPSPredictor2 and AdenPredictor made various errors (Table S2). By making both models freely available on an easily navigable web application, and by integrating an outlink to PARAS into antiSMASH v8.0, we hope that these models will be widely used by the natural products community to de-orphan and discover novel compounds.

Online methods

Computational methods

Data collection and curation

To train PARAS and PARASECT, we collated a dataset of 3254 unique A domains from three different sources: the NRPSPredictor2 dataset[21], the SANDPUMA dataset[24], and MIBiG 3.0[25]. We also added various novel data entries from literature. A domains were labelled by their GenBank, RefSeq[40,41] or UniProt protein ID[42] and a number corresponding to the index of the A domain in the NRPS enzyme. For A domains from the MIBiG dataset and novel data entries, we downloaded the sequence of their protein from GenBank and extracted the sequences of N-terminal and C-terminal A subdomains with HMMer3[43], using the AMP-binding (PF00501.29) and AMP-binding_C (PF13193.7) Pfam pHMMs[44], respectively. We were able to extract both subdomains for 2617 A domains in our dataset, and only extracted the N-terminal subdomain for the other 683. For A domains from SANDPUMA and NRPSPredictor2, we directly used the sequences from the datasets made available by the authors online. We also deduplicated our dataset through pairwise identity searches, allowing for truncations on either end of the A domain that might have occurred due to the different extraction methods used by different authors. For each A domain, we kept all protein labels.

As errors are unavoidable in human-collated datasets, we set out to curate our dataset in three steps. First, during deduplication, we flagged any domains with identical sequences but different selectivities and consulted the literature to fix their annotations. We repeated this process for A domains with identical 10 amino acid active site signatures, asserting that any differences in specificity were backed up by literature, fixing annotations where they were incorrect or omitting data points when we suspected experimental data to be unreliable. For instance, the active site signatures reported in literature for the first A domains in the proteins AAZ03554.1 and AAZ03552.1 were drastically different from the active site signatures we extracted[45]. Finally, we phylogenetically grouped the A domains based on their 34 amino acid extended active site signature with FastTree[46] and similarly verified the specificity of obvious outliers.

Structural modelling

Prior to structural modelling, we aligned all adenylation domains with MAFFT v7.508[47] (default settings). We used this alignment and the reference crystal structure of the GrsA A domain (1AMU[29]) to extract the N-terminal and C-terminal subdomains for each A domain. We then obtained predicted structures for each A subdomain domain using ColabFold[48] in batch mode (default settings) without side chain relaxation. We only acquired structures of the C-terminal subdomain for 2617 domains in our dataset. As the C-terminal subdomain does not play a large role in substrate recognition, with only the highly conserved K517 residue involved in stabilising the amino acid backbone of the substrate[10,11,21], we decided to generate full structural models by combining the AlphaFold model of each native N-terminal subdomain with the C-terminal subdomain of 1AMU. We then aligned all full structure models to 1AMU in PyMol[49] (open-source, v2.5.0) and saved all models in identical orientations in Protein Data Bank[50] (PDB) format. Finally, we copied the Mg^{2+} and AMP cofactors from 1AMU to the PDB

files of all A domains to reflect the state of the active site pocket without the amino acid substrate bound.

Sequence-based domain feature extraction

To obtain sequence-based features for our models, we first extracted the 34 residues that correspond to the 34 residues within 8Å of the phenylalanine substrate in the A domain of GrsA, as also done by the authors of NRPSPredictor2[21]. We tested three approaches: two used Muscle[51] (v3.8.1551) profile alignments, using the sequence of the GrsA (BAA00406.1) A domain as a reference. As a guide alignment, we used either sequence-based or structure-based sequence alignments of a subset of our dataset (923 A domains). We obtained the sequence-based guide alignment with Muscle (v3.8.1551), and the structure-based guide alignment by modelling the N-terminal A subdomains with Modeller[52] (v10.0, default: 250 models per sequence, no loop refinement) and aligning the resulting homology models with Caretta-shape[53] (v1.0, default settings). As sequence extraction time scales exponentially with the size of the guide alignment, we chose to use a subset of our data rather than the full dataset. The final sequence extraction approach, and the one we ended up using for our published models, uses a previously used NRPS-specific AMP-binding domain pHMM[20] with HMMer2[43] (v.2.3.2).

Each of the resulting 34 amino acid residues was featurised as a list of 15 physicochemical properties describing hydrophobicity (WOLS870101[54], NEU1, NEU2, NEU3 (Neumaier *et al.* accessed 2024), size (WOLS870102[54], TSAJ990101[55]), electron state (WOLS870103[54]), hydrogen bond donors (FAUJ880109[56]), polarity (GRAR740102[57], RADA880108 [58], ZIMJ680103 [59], occurrence in alpha-helices (CHOP780201), beta-sheets (CHOP780202) and beta-turns (CHOP780203)[60], and isoelectric point (ZIMJ680104 [59])[61], as previously described for NRPSPredictor and NRPSPredictor2[20,21]. These physicochemical features were then concatenated into feature vectors of length 510 (34 x 15; Figure 3B). Physicochemical properties and guide alignments can be found at https://github.com/BTheDragonMaster/parasect/tree/master/paras/data.

Comparing structure-based and sequence-based guide alignments

To compare structure-based and sequence-based guide alignments for feature extraction, we first split a subset of 923 domains into ten cross-validation sets, creating structure-based alignments (Caretta-shape[53] v1.0) and sequence-based alignments (Muscle[51] v3.8.1551) for each training set, stratifying on substrate class and using only the first reported substrate as response. Then, we featurised the 34 active site residues that we extracted from the structure-based and sequence-based alignments (as described above) and trained random forest models (scikit-learn[62] v1.1.3, default settings) on those same cross-validation sets and compared the test accuracy of the resulting two models. To ensure fair comparison, sequence-informed and structure-informed models were trained on the same data points using the same random seed for each cross-validation set.

Structure-based domain feature extraction

To convert our AlphaFold models into feature lists, we first placed a cubic voxel grid of 20x20x20 voxels centred around the position of the β-carbon of the phenylalanine substrate in the 1AMU reference structure. Next, for each A domain, we determined how many atoms overlapped with each voxel, recording counts for 7 atom types separately: aromatic,

hydrophobic, charged, and non-charged oxygens, charged and non-charged nitrogens, and sulphurs. We then concatenated these 7 numbers for all 8000 voxels into a single vector of length 56000 representing the A domain active site. As this is too large a number of features to work with considering our dataset size, we decided to reduce dimensionality through principal component analysis (PCA)[63]. To ensure that the resulting principal components captured variation between the active sites of highly similar substrates, we first assigned each data point to one of 9 substrate categories: aromatic amino acids, aromatic acids, Asx/Glx, cyclic aliphatic amino acids, substrates with nitrogen-containing side chains, phenylglycines, small amino acids, small hydrophobic amino acids, and small polar amino acids (Figure 3B). Data points that recognise substrates belonging to different groups were assigned to multiple groups. We only included substrates for which we have 11 or more examples in our dataset (Figure 2B). Then, we performed PCA on each of the 9 substrate groups as well as the full dataset using scikit-learn[62] (v1.2.0) and stored the transformative models. For each PCA, we made a scree plot and determined the optimal number of components by using the kneedle package[64] (v0.8.3) to automatically locate the knee in the plot. This yielded between 5-8 PCs for each substrate group and 11 PCs for the full dataset, cumulating to 68 PCs in total. We then applied each of the 10 transformative models on the full dataset to obtain all 68 PCs for all data points. The resulting PCs can be found in the 'structure data' folder of our online GitHub repository. These 68 PCs were consequently used as features for training PARAS and PARASECT (Figure 3B).

Substrate feature extraction

As PARASECT predicts interaction between an A domain and a substrate based on features of both the domain and the substrate's molecular structure, we also needed a way to featurise molecules. To this purpose, we collected isomeric SMILES strings[32] for each substrate and used PIKAChU[65] to extract ECFP-4 molecular fingerprints[28] of length 1024 for each compound. These vectors were directly used for model training (Figure 3C, D).

Data stratification

Prior to model training, we stratified our data in two different ways to ensure we could properly assess how well our model will perform on unseen sequences afterwards. One split was done based on phylogeny such that each substrate was represented in both the training and test set with at least one example; and the other split was based on substrate class alone.

To split based on phylogeny, we used the PairwiseAligner module from Biopython[66] (v1.81) to calculate bit-scores for all A domain pairs in our dataset. For each domain, we then concatenated the bit-scores for all pairs that domain was involved in (including a pair with itself) into a vector, keeping the order of domains constant for each iteration. Therefore, domains that are phylogenetically similar should have similar vectors. Next, we used the KMeans module from scikit-learn[62] (v1.2.0) to cluster our domains based on these bit-score vectors. We automatically detected the optimal number of clusters using kneedle[64] (v0.8.3), which yielded 4 clusters. We then distributed these clusters across our training and test sets such that both clusters had at least one example for each substrate in each. There was exactly one way to distribute the clusters that met this requirement, with 2 clusters assigned to each. We chose the larger dataset, accounting for about 75% of the data, as our training set, and used the remaining data points for our test set. We later used this train-test split to assess how well our models are likely to perform on A domains that are phylogenetically very different from A domains in our training set.

As many data points have multiple substrate labels, we used an iterative stratification method that is suitable for multi-label datasets[67]. This strategy assures that data points are divided across train and test sets such that each substrate class is proportionally represented. We used the MultilabelStratifiedShuffleSplit module from the iterstrat package (v0.1.6; https://pypi.org/project/iterative-stratification/) to implement iterative stratification in python, taking into consideration only those substrates for which there were at least 11 examples in our dataset. We chose a train-test ratio of 3:1 such that the training set size would be comparable to the size of the training set that we obtained by splitting on phylogeny. Exact train-test splits can be found in the 'train test splits' folder on our GitHub page.

Model training and evaluation

During model development, we used the MultiLabelStratifiedKFold module from the iterstrat package (v0.1.6) to perform 5-fold cross-validation on our training sets to assess intermediate model performance and tune model parameters. Both PARAS and PARASECT are random forest models implemented with the RandomForestClassifier module from scikit-learn[62] (v1.2.0). For each train-test split, we trained three models for both PARAS and PARASECT, each using different approaches to featurising A domains: sequence features only, structure features only, or a combination of both. As parameter tuning barely affected model performance, we decided to train all our models with 1000 trees, using default settings otherwise.

We trained PARAS as a single-label classifier, choosing the first listed label (which corresponds to the predominant substrate) as response for training purposes. PARAS output is therefore always a single substrate (Figure 3D). For assessing model performance, we designated a prediction as correct if the predicted substrate was among the substrates listed for that domain. We used balanced sampling to account for the differences in substrate counts for each class and prevent the model from prioritising the correct prediction of overrepresented substrates.

In contrast, PARASECT was trained on domain-substrate pairs, with a floating-point number as response, representing the probability of interaction between the domain and the substrate (Figure 3D). As this leads to highly imbalanced datasets, with a ratio of positive to negative data points around 1:30, we randomly under-sampled our data to achieve a 1:1 ratio using the RandomUnderSampler module from imblearn[68] (v0.10.1). Model performance was subsequently assessed in two different ways: by looking at metrics that are commonly used for evaluating binary data, including precision, recall, and F1-score; and by determining if the substrate with the highest probability of interaction for a domain corresponds to one of the substrates that domain recognises. We additionally estimated a 'cut-off' interaction probability by obtaining and comparing mean and median interaction probabilities for true positives and false positives. For both PARAS and PARASECT, we also gauged model performance for each different substrate class.

Benchmarking

To benchmark PARAS and PARASECT against current state-of-the-art A domain predictors, we compared their performance against AdenPredictor[23], SANDPUMA[24], and NRPSPredictor2[21]. To this purpose, we ran these three predictors on all A domains that did not occur in SANDPUMA's training set, which includes most A domains from the AdenPredictor and NRPSPredictor2 training sets as well. For a fair comparison, we only assessed model performance on A domains in the test sets of PARAS and PARASECT. For

SANDPUMA, NRPSPredictor2, and AdenPredictor, a prediction was labelled as correct if there was overlap between the true substrates of an A domain and the predicted substrates. For PARAS and PARASECT, a prediction was only labelled correct if the top substrate was among the true substrates of an A domain. All benchmarking and speed assessment were performed on a Macbook Pro from 2020 with M1 chip.

Structure modelling and substrate docking

To obtain structural models for the tryptophan-recognising A domains of Qui18-A1 and BreC-A3, we separately modelled their N-terminal and C-terminal subdomains with AlphaFold with amber relaxation[48,69] (Google Colab v1.3.0, default settings). We then aligned both subdomains to the 1AMU reference structure to obtain full A domain models.

Next, we performed two dockings for each A domain: one with two separate ligands: tryptophan (PDB:TRP) and AMP; and the other with a reaction intermediate: tryptophanyl-adenylate (PDB:TYM). First, we prepared the ligands with the SDMolSupplier module of RDKit[70] (v2022.03.5) by moving them into the rough vicinity of the active site, and subsequently converted them to .pdbqt format with the Meeko package (v0.2; https://pypi.org/project/meeko/). We converted the A domains to .pdbqt format with MGLTools. We manually added the Mg^{2+} atom into the .pdbqt files after. Finally, we performed the docking with Autodock Vina[71] (v1.2.3; exhaustiveness: 64; number of poses: 40), saving the best 20 poses.

Feature inference

We performed two types of feature inference on our models: one that looks at overall feature importance in the entire model (PARAS and PARASECT); and one that assesses substrate-specific feature inference (PARAS only). For the former, feature importances were directly extracted from the RandomForestClassifier instances, and summed across all features describing the same residue to yield a measure of residue-level importance (Figure 6A). For substrate-specific feature inference, we iterated over the nodes of each individual tree in the forest, and determined the information gain per node for each substrate that was seen by that node. Information gain is defined as:

$$Information\ gain\ =\ H(t)\ -\ H(s,t)$$

Where $H(t)$ is the entropy of the parent node, and $H(s,t)$ is the average entropy of the two child nodes. Entropy is calculated as follows:

$$H(t)\ =\ -[p_{S,t}log_2(p_{S,t})\ +p_{NS,t}log_2(p_{NS,t})]$$

Where the probability of selecting substrate 'S' at node t, $p_{S,t}$= $n(t,S)$ / $n(t)$
And the probability of not selecting substrate 'S' at node t, $p_{NS,t}$= $n(t,NS)$ / $n(t)$
Where $n(t)$ is the total number of domains seen by node t, $n(t,S)$ is the number of domains recognising substrate S seen node t, and $n(t,NS)$ is the number of domains not recognising substrate S seen by node t.

The average entropy of the two child nodes, $H(s,t)$, is given by:

$$H(s,t) = P_L H(t_L) + P_R H(t_R)$$

Where $H(t_L)$ and $H(t_R)$ are the entropy of the left and right child nodes of parent node t, calculated in the same way as $H(t)$; the probability of a domain at the left child node $P_L = n(t_L)/n(t)$; and the probability of a domain at the right child node $P_R = n(t_R)/n(t)$.
Where $n(t_L)$ and $n(t_R)$ are the number of domains seen by the left and right child nodes of node t, respectively.

As nodes are split on a single feature, information gain is a measure of entropy decrease for a specific substrate when the data is split on that feature. As such, the higher the information gain for a node, the more important the feature linked to that node.

To obtain an average substrate-specific information gain per feature, we can calculate the average gain of a feature across all its occurrences in the forest. An artefact of this method is that the information gain of features that are typically only used a few times in the forest but drastically decrease entropy when they occur is disproportionally inflated. This was especially problematic for features resolving leaf nodes for substrates that do not clade well together, such as tryptophan and alanine. Therefore, for each substrate we set the information gain for features that were used throughout the forest fewer than x times to 0, where we varied x from 0 to 500 at intervals of 50. We chose inclusion thresholds for each substrate individually, typically selecting higher inclusion thresholds for substrates whose recognising domains do not form monophyletic clades. Then, we obtained per-residue information gain by averaging across the information gain for all 15 features, Finally, we normalised the residue-level information gain across all substrates to highlight the differential contribution of certain residues to classifying specific substrates, and automatically visualised normalised per-residue information gain in Pymol[49] for each substrate (Figure 6B-D, Figures Sj-Sq).

Web portal development

To facilitate and promote the usage of PARAS and PARASECT, we have developed a web-based graphical user interface (v1.0.0). The web portal's frontend is developed using JavaScript React (v18.1.0), which communicates with a Python Flask (v3.0.2) backend application. The web portal can be accessed at https://paras.bioinformatics.nl/. The user is able to upload a file containing protein sequences in FASTA format or a GenBank file, or input protein sequences in FASTA format directly. Settings for the user to customise are the model to use (PARAS substrate specificity prediction for all substrates, PARAS substrate specificity prediction for common substrates, or PARASECT) and if to use profile-guided structure alignment for active site extraction. Additionally, if users pick PARASECT they are able to upload their own list of substrates to predict substrate specificity for. A short guide accompanied by screenshots from the v1.0.0 version of the web application can be found in Table S3 in the Supplementary Information.

Experimental methods

Bacterial Strains, Plasmids, and Culture Conditions.

Streptomyces sparsogenes DSM 40356, *Streptomyces species* maeda85, *Streptomyces coelicolor* M1154 and its mutants were grown on either TSB medium (Becton Dickinson) or SFM medium at 30°C for routine cultivation. For tryptopeptin production, 20µl spore

suspension was inoculated on SY agar medium (starch 2.4%, glucose 0.1%, peptone 0.3%, beef extract 0.3%, yeast extract 0.5%, calcium carbonate 0.4%) and incubated at 30°C for 7 days.

Saccharomyces cerevisiae VL6−48N was grown on YPD medium (glucose 2%, yeast extract 1%, peptone 2%) supplemented with 100 mg/L adenine at 30 °C. For TAR cloning, tryptophan-deficient medium (sorbitol 18.2%, glucose 2.2%, yeast nitrogen base without amino acid and ammonium sulfate 0.17%, yeast synthetic drop-out medium supplements without tryptophan 0.19%, ammonium sulfate 0.5%, and agar 2%) was used with 5-fluoroorotic acid (1mg/ml) (12).

TAR cloning and *Streptomyces*/*E. coli* conjugation

To construct capture vectors for cloning the two putative tryptopeptin BGCs, paired homologous arms were amplified from genomic DNA of corresponding microorganism with 20-30 bp primer-introduced overhangs for GeneArt™ Seamless Cloning and assembled into SpeI/XhoI-digested pCAP1000, together with the counter-selectable cassette *pADH-URA3* amplified from pCAP1000 which is placed between the paired homologous arms. The integrity of resulting plasmids pCAP1000maedttp and pCAP1000spattp was confirmed by restriction digestion and sequencing.

To produce genomic DNA fragments containing tryptopeptin BGC region for TAR cloning, genomic DNA isolated from original hosts was digested with restriction enzymes that has at least one cutting site outside of target region (XbaI for *Streptomyces* sp. maeda85 and BswI for *Streptomyces sparsogenes*) and purified by ethanol precipitation. 0.5-1.5 μg of digested genomic DNA and 0.05-0.08 μg of the corresponding capture vector linearized by PmeI were used for yeast spheroplast transformation carried out using a previously reported protocol. The resulting yeast colonies were picked and screened by PCR (Table S4) using a previously reported protocol with primers designed to amplify fragments from the middle and each end of the target BGC[72].

Plasmids were isolated from PCR-positive yeast colonies and used to transform *E. coli* Top10 for propagation. After verification by restriction digestion (XhoI and PstI for pCAP1000spattpBGC, SrfI and PstI for pCAP1000maedattpBGC), plasmids harboring tryptopeptin BGC were introduced into *S. coelicolor* M1154 by triparental mating from *E. coli* ET12567 with the helper strain *E. coli* ET12567/pUB307 using a standard protocol. Kanamycin resistance was used for exconjugant screening.

Growth of *Streptomyces* strains and analysis of epoxyketone production

After incubation, the agar was cut into small chunks and transferred into glass vials before adding 15 mL of ethyl acetate to each vial. After 2h, the organic extracts were decanted and concentrated by rotary evaporation. The residues were separately dissolved in 1 mL of methanol, and 2 μl of each sample was analysed by UHPLC-ESI-Q-TOF-MS/MS on a Bruker MaXis II mass spectrometer (or Bruker MaXis Impact mass spectrometer) coupled to a Dionex UltiMate 3000 UHPLC fitted with an Agilent Zorbax Eclipse Plus C18 column (100 × 2.1 mm, 1.8 μm). Using a flow rate of 0.2 ml/min, the column was eluted with a combination of water and acetonitrile as follows: 5% (v/v) acetonitrile for 5 min, 5−100% (v/v) acetonitrile over 21 min, 100% (v/v) acetonitrile for 3 min, 100−5% (v/v) acetonitrile over 2 min and 5% (v/v) acetonitrile for 3 min. We ran the mass spectrometer in positive ion mode (scan range: 200-3000 m/z). Source conditions were: end plate offset at −500 V; capillary at −4500 V; nebulizer gas (N_2) at 1.6 bar; dry gas (N_2) at 8 L min−1; dry temperature at 180 °C. Ion transfer conditions

were: ion funnel RF at 200 Vpp; multiple RF at 200 Vpp; quadrupole low mass at 55 m/z; collision energy at 5.0 eV; collision RF at 600 Vpp; ion cooler RF at 50–350 Vpp; transfer time at 121 s; pre-pulse storage time at 1 s. Calibration was performed with 1 mM sodium formate through a loop injection of 20 µL at the start of each run. The resulting spectra were analysed with Bruker's DataAnalysis software (v4.4).

A domain selectivity determination

To determine the substrate selectivity of the third A domain of TtpD (from here on referred to as TtpD-A3), we first transformed *Escherichia coli* with a plasmid containing DNA encoding the His-tagged TtpD-A3-PCP didomain, and then overexpressed and purified the protein. We opted to express the didomain rather than the standalone A domain so that we could determine if both reactions that the adenylation domain catalyses took place: the adenylation reaction, which catalyses the conversion of ATP and the amino acid substrate to an amino-acyl-AMP intermediate; and the thiolation reaction, which transfers the amino acid to the phosphopantetheine arm on the PCP domain. After protein purification, we converted the didomain to its *holo* form *in vitro* by incubating it with CoA-SH and Sfp enzyme, a phosphopantetheinyl transferase which activates the didomain by transferring a phosphopantetheine arm onto the PCP domain from CoA. Then, we assessed the ability of *holo*-TtpD-A3-PCP to load various substrates by incubating the didomain with ATP and a tryptophan, phenylalanine, leucine, histidine and/or valine substrate, either separately for individual substrate assessment, or in tandem to observe which substrate was preferred. We also performed an *in vivo* competition assay by expressing the TtpD-A3-PCP didomain in *BAP1* cells[22], a strain of *E. coli* that expresses Sfp, which allows *in vivo* domain activation and substrate loading. In all cases, substrate activation was measured with UHPLC-ESI-Q-TOF-MS analysis.

PCR amplification

As *Streptomyces sparsogenes* is very GC-rich, primer design with sufficiently long overlaps that still have reasonable annealing temperatures can be challenging, especially if the design requires overhangs with restriction sites for later incorporation of the DNA fragment into a plasmid. For this reason, we did two steps of PCR amplification to obtain PCR fragments containing the DNA encoding the TtpD-A3-PCP didomain: the first to amplify fragments from the GC-rich *Streptomyces sparsogenes* genomic DNA; and the second to amplify fragments with overhangs containing restriction sites (NdeI on the forward primer; EcoRI for the reverse). Detailed PCR protocols for each PCR can be found in Table S5. Primers were ordered from Merck. Resulting PCR reactions were run on an agarose gel (1g agarose in 100 ml 1x TBE buffer) for 70m at 130V, the bands were excised, and the DNA was extracted from the gel slices using the ThermoScientific Gel Extraction Kit.

Plasmid preparation and cloning

Next, we digested both our amplified PCR product (6 µl, ~400 ng; 0.5 µl EcoRI; 0.5 µl NdeI; 11.7 µl dH_2O; 2 µl 10X Buffer O) and the plasmid pET28A (28.3 µl, ~2 µg; 1 µl EcoRI; 1 µl NdeI; 14.7 µl dH_2O; 5 µl 10X Buffer O), which contains a T7 promoter and terminator, a kanamycin resistance cassette, an N-terminal His-tag, an N-terminal thrombin cleavage site, and EcoRI and NdeI restriction sites. Enzymes and buffers were acquired from ThermoFisher scientific. Reactions were incubated for 2h at 37°C and subsequently inactivated for 20m at 65°C.

We then set up a ligation reaction to insert the digested PCR product into the digested pET28A plasmid (2 µl ligase buffer, 1 µl T4 ligase, 4 µl 5ng/µl digested pET28A, 13 µl 5ng/µl digested insert). The reaction was incubated for 5h at 16°C and subsequently stored at 4°C. Buffer and enzyme were acquired from ThermoFisher scientific.
Next, we transformed our ligation reaction into chemically competent *E. coli* One Shot™ TOP10 cells (ThermoFisher Scientific). We added 5 µl of our ligation reaction to one vial of competent cells and mixed gently, after which we placed the reaction on ice for 30m. Then, we heat-shocked the cells for 30s at 42°C and directly transferred them back to ice for 2m. We added 250 µl of LB medium to each vial and shook the vials horizontally at 37°C for 1h (225 rpm). Then, we spread 20 µl transformed cells onto LB plates with kanamycin (50 µg/ml).

To ensure that the plasmids had incorporated the sequence encoding the TtpD-A3-PCP didomain correctly, we inoculated 15 ml LB medium + kanamycin (50 µg/ml) with the resulting colonies, incubated them overnight at 37°C, and isolated the plasmids using ThermoFisher Scientific's GeneJET Plasmid Miniprep Kit. We then checked the plasmids by restriction digestion and gel electrophoresis and sent plasmids that showed bands of the right sizes for sequencing. Cultures that harboured a correct plasmid were stored at -80°C.
Finally, we transformed 55.3 ng of plasmid into both chemically competent BL21* cells and chemically competent *E. coli* BAP1 cells, the latter of which express the Sfp protein responsible for activating the A-PCP didomain as described above. We incubated the reaction on ice for 15 minutes, heat-shocked the cells for 40s at 42°C and directly transferred them back to ice and added 250 µl of LB medium. We shook the vials horizontally at 37°C for 1h (225 rpm), plated 25 µl of our transformed cells onto LB plates with kanamycin (50 µg/ml), and incubated overnight at 37°C.

Protein overexpression and purification

To obtain sufficient protein for analysis, we inoculated 1L of LB + 50 µg/ml kanamycin with either transformed BL21* or BAP1 cells and left them to grow at 37°C to an OD of ~1.0. Then, we added 1ml 0.5M IPTG to induce protein production. Cells were left to produce protein overnight at 15°C. Cells were spun down for 20m at 5000 rpm, 4°C. Pellets were resuspended in ~15ml Tris washing buffer (20mM Imidazole, 20 mM Tris-HCl, 100 mM NaCl, 1-% glycerol). Next, we lysed the cells in a cell disruptor. The lysed cells were spun down for 15m at 17000 rpm, 4°C. Supernatant was transferred to a fresh tube and spun down for another 15m at 17000 rpm, 4°C. Then, the supernatant was filtered (0.45 µl filter) and loaded onto a 1ml Cytiva HisTrap-FF column. The column was washed twice with Tris washing buffer (20mM Imidazole, 20 mM Tris-HCl, 100 mM NaCl, 1-% glycerol), and the protein was eluted with elution buffers containing increasing concentrations of Imidazole (50 µM – 300 µM). Wash fractions were also collected. Fractions were denatured and run on an 8% SDS-PAGE gel for ~30m at 180V. Fractions containing protein of the expected size were concentrated to 0.5ml (5000 MWCO, 4000 rpm, 4°C), the buffer was replaced with storage buffer (20 mM Tris-HCl, 100 mM NaCl, 1-% glycerol) four times. The resulting protein solutions had a concentration of 47.673 mg/ml for *apo*-TtpD-A3-PCP (expressed from TOP10 cells) and 32.05 mg/ml for *holo*-TtpD-A3-PCP (expressed from BAP1 cells). We analysed the proteins by UHPLC-ESI-Q-TOF-MS to assert that the proteins had the correct weights and to check which substrate was loaded by *holo*-TtpD-A3-PCP in BAP1 cells.

In vitro activity assay

To assess which substrate is loaded by TtpD-A3-PCP, we first converted *apo*-TtpD-A3-PCP purified from TOP10 cells to *holo*-TtpD-A3-PCP by incubating 42 µl 200 µM *apo*-TtpD-A3-PCP with 5 µl 100 mM $MgCl_2$ (as Mg^{2+} is required by Sfp and by the A domain itself to stabilise the negative charge of ATP in the active site), 2 µl 20 mM CoA-SH (which provides the phosphopantetheine arm), and 1 µl 400 µM Sfp enzyme (isolated as described above). We incubated the reaction for 1h at room temperature, and then loaded the *holo*-enzyme with substrate by adding 50 µl 170 µM *holo*-TtpD-A3-PCP to 1 µl 100 mM ATP (required for the adenylation reaction) and 1 µl 50 µM substrate dissolved in dH_2O. We tested the substrates tryptophan, phenylalanine, histidine, leucine, and valine. *apo*-TtpD-A3-PCP, *holo*-TtpD-A3-PCP, and loaded *holo*-TtpD-A3-PCP were analysed by UHPLC-ESI-Q-TOF-MS.

UHPLC-ESI-Q-TOF-MS analysis

We analysed intact *apo*-TtpD-A3-PCP, *holo*-TtpD-A3-PCP, and loaded *holo*-TtpD-A3-PCP on a Bruker MaXis II ESI-Q-TOF-MS connected to a Dionex 3000 RS UHPLC (equipped with an ACE C_4-300 RP column (100 x 2.1 mm, 5 µm, 30°C); controlled using Bruker Otof control 4.0). The column was eluted with 0.1% formic acid and 5-100% MeCN in a linear gradient for 30m. We ran the mass spectrometer in positive ion mode (scan range: 200-3000 m/z). We used the following source settings: end plate offset: −500 V; capillary: -4500 V; nebulizer gas (N_2): 1.8 bar; dry gas (N_2): 9.0 L/min; dry temperature: 200 °C. The following ion transfer conditions were used: ion funnel RF: 400 Vpp; multiple RF: 200 Vpp; quadrupole low mass: 200 m/z; collision RF: 2000 Vpp; transfer time: 110.0 µs; pre-pulse storage time: 10.0 µs. The resulting spectra were analysed with Bruker's DataAnalysis software (v4.4).

Data availability

PARAS and PARASECT are available for Python versions 3.9 and up. The source code for the described analyses and training the substrate specificity models are available on GitHub at https://github.com/BTheDragonMaster/parasect under the GNU Affero General Public License Version 3 licence. The source code for the web application (https://paras.bioinformatics.nl) can be found in the webapp branch (https://github.com/BTheDragonMaster/parasect/tree/webapp) of the aforementioned GitHub repository. Trained models for PARAS (for predicting substrate specificity for all substrates, and on the set of common substrates) and PARASECT are archived with Zenodo (https://zenodo.org/records/13165500).

Acknowledgements

This research was funded by the Novel Antibacterial Compounds and Therapies Antagonising Resistance program (NACTAR) from the Dutch Research Council (NWO) [project number 16440 to MHM], an ERC Starting Grant [948770-DECIPHER to MHM], the BBSRC (Grant refs BBSRC BB/M017982/1, BB/T017163/1 and BB/R010218/1 to GLC) and ARC (Grant Ref CE200100012 to GLC). SR was supported by the Swiss National Science Foundation (project no. 501100001711–209124) and the Helmut Horten Foundation. JDDCB and MGC were supported by the Department of Microbiology and Cell Science at University of Florida. MLR and YZ were supported by a PhD studentship from the EPSRC (Grant Ref EP/L016494/1) and a Chancellor's International Scholarship from the University of Warwick, respectively.

Competing Interests

G.L.C. is non-executive director, consultant and shareholder of ErebaGen Ltd. M.H.M. is a member of the scientific advisory board of Hexagon Bio.

References

1. Fleming, A. Penicillin. *Br. Med. J.* **2**, 386 (1941).
2. Dittmann, J., Wenger, R. M., Kleinkauf, H. & Lawen, A. Mechanism of cyclosporin A biosynthesis. Evidence for synthesis via a single linear undecapeptide precursor. *J. Biol. Chem.* **269**, 2841–2846 (1994).
3. Usuki, Y. *et al.* UK-2A, B, C and D, novel antifungal antibiotics from Streptomyces sp. 517-02 VI (2). Structure-activity relationships of UK-2A. *J. Antibiot.* **55**, 607–610 (2002).
4. Blodgett, J. A. V., Zhang, J. K., Yu, X. & Metcalf, W. W. Conserved biosynthetic pathways for phosalacine, bialaphos and newly discovered phosphonic acid natural products. *J. Antibiot.* **69**, 15–25 (2016).
5. Zipperer, A. *et al.* Human commensals producing a novel antibiotic impair pathogen colonization. *Nature* **535**, 511–516 (2016).
6. Wilson, M. R. *et al.* The human gut bacterial genotoxin colibactin alkylates DNA. *Science* **363**, (2019).
7. Mendes, R. *et al.* Deciphering the rhizosphere microbiome for disease-suppressive bacteria. *Science* **332**, 1097–1100 (2011).
8. Süssmuth, R. D. & Mainz, A. Nonribosomal Peptide Synthesis-Principles and Prospects. *Angew. Chem. Int. Ed Engl.* **56**, 3770–3821 (2017).
9. Caboche, S., Leclère, V., Pupin, M., Kucherov, G. & Jacques, P. Diversity of monomers in nonribosomal peptides: towards the prediction of origin and biological activity. *J. Bacteriol.* **192**, 5143–5150 (2010).
10. Challis, G. L., Ravel, J. & Townsend, C. A. Predictive, structure-based model of amino acid recognition by nonribosomal peptide synthetase adenylation domains. *Chem. Biol.* **7**, 211–224 (2000).
11. Stachelhaus, T., Mootz, H. D. & Marahiel, M. A. The specificity-conferring code of adenylation domains in nonribosomal peptide synthetases. *Chem. Biol.* **6**, 493–505 (1999).

12. Scaglione, A. *et al.* Structure of the adenylation domain Thr1 involved in the biosynthesis of 4-chlorothreonine in Streptomyces sp. OH-5093-protein flexibility and molecular bases of substrate specificity. *FEBS J* **284**, 2981–2999 (2017).
13. Lautru, S. & Challis, G. L. Substrate recognition by nonribosomal peptide synthetase multi-enzymes. *Microbiology (Reading)* **150**, 1629–1636 (2004).
14. Challis, G. L. & Ravel, J. Coelichelin, a new peptide siderophore encoded by the Streptomyces coelicolor genome: structure prediction from the sequence of its non-ribosomal peptide synthetase. *FEMS Microbiol Lett* **187**, 111–114 (2000).
15. Skinnider, M. A. *et al.* Comprehensive prediction of secondary metabolite structure and biological activity from microbial genome sequences. *Nat. Commun.* **11**, 6058 (2020).
16. Ziemert, N., Alanjary, M. & Weber, T. The evolution of genome mining in microbes - a review. *Nat Prod Rep* **33**, 988–1005 (2016).
17. Medema, M. H., de Rond, T. & Moore, B. S. Mining genomes to illuminate the specialized chemistry of life. *Nat Rev Genet* **22**, 553–571 (2021).
18. Blin, K. *et al.* antiSMASH 7.0: new and improved predictions for detection, regulation, chemical structures and visualisation. *Nucleic Acids Res.* **51**, W46–W50 (2023).
19. Khayatt, B. I., Overmars, L., Siezen, R. J. & Francke, C. Classification of the adenylation and acyl-transferase activity of NRPS and PKS systems using ensembles of substrate specific hidden Markov models. *PLoS One* **8**, e62136 (2013).
20. Rausch, C., Weber, T., Kohlbacher, O., Wohlleben, W. & Huson, D. H. Specificity prediction of adenylation domains in nonribosomal peptide synthetases (NRPS) using transductive support vector machines (TSVMs). *Nucleic Acids Res.* **33**, 5799–5808 (2005).
21. Röttig, M. *et al.* NRPSpredictor2--a web server for predicting NRPS adenylation domain specificity. *Nucleic Acids Res.* **39**, W362–7 (2011).
22. Robinson, S. L. *et al.* Global analysis of adenylate-forming enzymes reveals β-lactone biosynthesis pathway in pathogenic. *J. Biol. Chem.* **295**, 14826–14839 (2020).
23. Mongia, M. *et al.* AdenPredictor: accurate prediction of the adenylation domain

specificity of nonribosomal peptide biosynthetic gene clusters in microbial genomes. *Bioinformatics* **39**, i40–i46 (2023).

24. Chevrette, M. G., Aicheler, F., Kohlbacher, O., Currie, C. R. & Medema, M. H. SANDPUMA: ensemble predictions of nonribosomal peptide chemistry reveal biosynthetic diversity across Actinobacteria. *Bioinformatics* **33**, 3202–3210 (2017).
25. Terlouw, B. R. *et al.* MIBiG 3.0: a community-driven effort to annotate experimentally validated biosynthetic gene clusters. *Nucleic Acids Res.* **51**, D603–D610 (2023).
26. Zhang, C. *et al.* In vitro characterization of echinomycin biosynthesis: formation and hydroxylation of L-tryptophanyl-S-enzyme and oxidation of (2S,3S) β-hydroxytryptophan. *PLoS One* **8**, e56772 (2013).
27. Li, Y.-X., Zhong, Z., Zhang, W.-P. & Qian, P.-Y. Discovery of cationic nonribosomal peptides as Gram-negative antibiotics through global genome mining. *Nat. Commun.* **9**, 3273 (2018).
28. Rogers, D. & Hahn, M. Extended-connectivity fingerprints. *J. Chem. Inf. Model.* **50**, 742–754 (2010).
29. Conti, E., Stachelhaus, T., Marahiel, M. A. & Brick, P. Structural basis for the activation of phenylalanine in the non-ribosomal biosynthesis of gramicidin S. *EMBO J.* **16**, 4174–4183 (1997).
30. Heard, S. C. & Winter, J. M. Structural, biochemical and bioinformatic analyses of nonribosomal peptide synthetase adenylation domains. *Nat Prod Rep* **41**, 1180–1205 (2024).
31. Hojati, Z. *et al.* Structure, biosynthetic origin, and engineered biosynthesis of calcium-dependent antibiotics from Streptomyces coelicolor. *Chem. Biol.* **9**, 1175–1187 (2002).
32. Anderson, E. *SMILES, a Line Notation and Computerized Interpreter for Chemical Structures*. (1987).
33. Tsunematsu, Y. *et al.* Isolation, Structure Elucidation, and Total Synthesis of Tryptopeptins A and B, New TGF-β Signaling Modulators from Streptomyces sp. (2015) doi:10.1021/ol503340k.

34. Zabala, D. *et al.* A Flavin-Dependent Decarboxylase-Dehydrogenase-Monooxygenase Assembles the Warhead of α,β-Epoxyketone Proteasome Inhibitors. *J Am Chem Soc* **138**, 4342–4345 (2016).
35. Huang, C. *et al.* Parallelized gene cluster editing illuminates mechanisms of epoxyketone proteasome inhibitor biosynthesis. *Nucleic Acids Res* **51**, 1488–1499 (2023).
36. Lorenzo de los Santos, E. & Challis, G. clusterTools: functional element identification for the in silico prioritization of biosynthetic gene clusters. *Access Microbiol.* **1**, (2019).
37. Gomez-Escribano, J. P. & Bibb, M. J. Engineering Streptomyces coelicolor for heterologous expression of secondary metabolite gene clusters. *Microb Biotechnol* **4**, 207–215 (2011).
38. Pfeifer, B. A., Admiraal, S. J., Gramajo, H., Cane, D. E. & Khosla, C. Biosynthesis of complex polyketides in a metabolically engineered strain of E. coli. *Science* **291**, 1790–1792 (2001).
39. Machushynets, N. V. *et al.* Discovery and Derivatization of Tridecaptin Antibiotics with Altered Host Specificity and Enhanced Bioactivity. *ACS Chem. Biol.* **19**, 1106–1115 (2024).
40. Sayers, E. W. *et al.* Database resources of the National Center for Biotechnology Information in 2023. *Nucleic Acids Res.* **51**, D29–D38 (2023).
41. Benson, D. A. *et al.* GenBank. *Nucleic Acids Res.* **46**, D41–D47 (2018).
42. UniProt Consortium, T. UniProt: the universal protein knowledgebase. *Nucleic Acids Res.* **46**, 2699 (2018).
43. Mistry, J., Finn, R. D., Eddy, S. R., Bateman, A. & Punta, M. Challenges in homology search: HMMER3 and convergent evolution of coiled-coil regions. *Nucleic Acids Res.* **41**, e121 (2013).
44. Mistry, J. *et al.* Pfam: The protein families database in 2021. *Nucleic Acids Res.* **49**, D412–D419 (2021).
45. Tooming-Klunderud, A., Rohrlack, T., Shalchian-Tabrizi, K., Kristensen, T. & Jakobsen,

K. S. Structural analysis of a non-ribosomal halogenated cyclic peptide and its putative operon from : implication for evolution of cyanopeptolins. *Microbiology* **155**, 2106–2108 (2009).

46. Price, M. N., Dehal, P. S. & Arkin, A. P. FastTree 2--approximately maximum-likelihood trees for large alignments. *PLoS One* **5**, e9490 (2010).
47. Katoh, K. & Standley, D. M. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.* **30**, 772–780 (2013).
48. Mirdita, M. *et al.* ColabFold: making protein folding accessible to all. *Nat. Methods* **19**, 679–682 (2022).
49. *PyMol*. (2008).
50. Berman, H. M. *et al.* The Protein Data Bank. *Acta Crystallogr. D Biol. Crystallogr.* **58**, 899–907 (2002).
51. Edgar, R. C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* **32**, 1792–1797 (2004).
52. Webb, B. & Sali, A. Comparative Protein Structure Modeling Using MODELLER. *Curr. Protoc. Protein Sci.* **86**, 2.9.1–2.9.37 (2016).
53. Akdel, M., Durairaj, J., de Ridder, D. & van Dijk, A. D. J. Caretta - A multiple protein structure alignment and feature extraction suite. *Comput. Struct. Biotechnol. J.* **18**, 981–992 (2020).
54. Wold, S. & National Research Council Canada. *Principal Property Values for Six Non-Natural Amino Acids and Their Application to a Structure-Activity Relationship for Oxytocin Peptide Analogues: Supplementary Unpublished Material*. (1987).
55. Tsai, J., Taylor, R., Chothia, C. & Gerstein, M. The packing density in proteins: standard radii and volumes. *J. Mol. Biol.* **290**, 253–266 (1999).
56. Fauchère, J. L., Charton, M., Kier, L. B., Verloop, A. & Pliska, V. Amino acid side chain parameters for correlation studies in biology and pharmacology. *Int. J. Pept. Protein Res.* **32**, 269–278 (1988).
57. Grantham, R. Amino acid difference formula to help explain protein evolution. *Science*

185, 862–864 (1974).

58. Radzicka, A. & Wolfenden, R. Comparing the polarities of the amino acids: side-chain distribution coefficients between the vapor phase, cyclohexane, 1-octanol, and neutral aqueous solution. *Biochemistry* **27**, 1664–1670 (1988).
59. The characterization of amino acid sequences in proteins by statistical methods. *Journal of Theoretical Biology* **21**, 170–201 (1968).
60. Chou, P. Y. & Fasman, G. D. *Secondary Structural Prediction of Proteins from Their Amino Acid Sequence*. (1977).
61. Kawashima, S. *et al.* AAindex: amino acid index database, progress report 2008. *Nucleic Acids Res.* **36**, D202–5 (2008).
62. Garreta, R. & Moncecchi, G. *Learning Scikit-Learn: Machine Learning in Python*. (Packt Pub Limited, 2013).
63. Jolliffe, I. T. *Principal Component Analysis*. (Springer Science & Business Media, 2013).
64. Satopaa, V., Albrecht, J., Irwin, D. & Raghavan, B. Finding a 'kneedle' in a haystack: Detecting knee points in system behavior. in *2011 31st International Conference on Distributed Computing Systems Workshops* (IEEE, 2011). doi:10.1109/icdcsw.2011.20.
65. Terlouw, B. R., Vromans, S. P. J. M. & Medema, M. H. PIKAChU: a Python-based informatics kit for analysing chemical units. *J. Cheminform.* **14**, 34 (2022).
66. Cock, P. J. A. *et al.* Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics* **25**, 1422–1423 (2009).
67. Sechidis, K., Tsoumakas, G. & Vlahavas, I. On the Stratification of Multi-label Data. *Machine Learning and Knowledge Discovery in Databases* 145–158 (2011).
68. Website. https://doi.org/10.48550/arXiv.1609.06570 doi:10.48550/arXiv.1609.06570.
69. Baltzis, A. *et al.* Highly significant improvement of protein sequence alignments with AlphaFold2. *Bioinformatics* **38**, 5007–5011 (2022).
70. Website. https://www.rdkit.org/.
71. Eberhardt, J., Santos-Martins, D., Tillack, A. F. & Forli, S. AutoDock Vina 1.2.0: New Docking Methods, Expanded Force Field, and Python Bindings. *J. Chem. Inf. Model.* **61**,

3891–3898 (2021).

72. Kouprina, N. & Larionov, V. Selective isolation of genomic loci from complex genomes by transformation-associated recombination cloning in the yeast Saccharomyces cerevisiae. *Nat Protoc* **3**, 371–377 (2008).