ARTICLE

Machine Learning-Driven Discovery of Highly Selective Antifungal Peptides Containing Non-Canonical β-Amino Acids

Douglas H. Chang,†[a] Joshua D. Richardson,†[a] Myung-Ryul Lee,[a] David M. Lynn,*[ab] Sean P. Palecek,*[a] and Reid C. Van Lehn*[ab]

Antimicrobial peptides (AMPs) are promising compounds for the treatment and prevention of multidrug-resistant infections because of their ability to directly disrupt microbial membranes, a mechanism that is less likely to lead to resistance compared to antibiotics. Unfortunately, natural AMPs are prone to proteolytic cleavage in vivo and have relatively low selectivity for microbial versus human cells, motivating the development of synthetic peptidomimetics of AMPs with improved peptide stability, activity, and selectivity. However, a lack of understanding of structure-activity relationships for peptidomimetics constrains development to rational design or experimental predictors, both of which are cost and time prohibitive, especially when the design space of possible sequences scales exponentially with the number of amino acids. To address these challenges, we developed an iterative Gaussian process regression (GPR) approach to explore a large design space of 336,000 synthetic α/β-peptide analogues of a natural AMP, aurein 1.2, based on an initial training set of 147 sequences and their biological activities against microbial pathogens and selectivity for microbes vs. mammalian cells. We show that the quantification of prediction uncertainty provided by GPR can guide the exploration of this design space via iterative experimental measurements to efficiently discover novel sequences with up to a 52-fold increase in antifungal selectivity compared to aurein 1.2. The highest selectivity peptide discovered using this approach features an unconventional amino-acid substitution strategy that was previously unseen by the model and would be unlikely to be explored by conventional rational design. Overall, this work demonstrates a generalizable approach that integrates computation and experiment to accurately predict the selectivity of AMPs containing synthetic amino acids, which we employed to discover new α/β-peptides that hold promise as selective antifungal agents to combat the antimicrobial resistance crisis.

Introduction

Since the introduction of insulin to treat type 1 diabetes 100 years ago, large strides have been made in the field of peptide therapeutics to treat a range of other diseases such as cancer, multiple sclerosis, and HIV.[1, 2] Antimicrobial peptides (AMPs), a component of the innate immune system, have garnered particular attention as potential agents to combat the alarming increase in multi-drug-resistant bacterial and fungal infections. Many natural AMPs are cationic, α-helical peptides that work via disruption of microbial cell membranes, making them less likely to lead to resistance than antibiotics with a specific molecular target.[3-6] AMPs are typically amphiphilic in nature, with opposing hydrophilic and hydrophobic faces that are important for membrane disruption.[5] However, natural α-helical AMPs typically have low stability *in vivo* because of their susceptibility to proteolytic cleavage[7, 8] and also exhibit low selectivities for microbial versus human cells; these two limitations have served as obstacles to their translation into clinical therapies.[9] Novel approaches to design peptides and their analogues are necessary to tackle these challenges.

One strategy to address these issues is to substitute traditional α-amino acids with noncanonical β-amino acids, which have an additional carbon in the backbone relative to α-amino acids, in specific patterns and positions of the original sequence.[9-13] This class of synthetic peptides, known as α/β-peptides, can exhibit enhanced proteolytic resistance while maintaining side chain presentations similar to those of α-peptides, enabling the templating of sequences on naturally occurring AMPs.[14-18] In addition, using β-amino acids enables the exploration of a much larger array of amino-acid combinations compared to naturally occurring peptides, which could lead to enhanced stability and selectivity profiles. For example, our previous studies developed α/β-peptides templated on aurein 1.2, an AMP found in Australian bell frogs, to enhance their antifungal selectivity against *Candida albicans*.[9, 19] *C. albicans* is the most prevalent hospital-acquired fungal pathogen and is known to cause bloodstream infections,[20] leading to a mortality rate of up to 70% in the case of sepsis.[21] Despite an alarming increase in antifungal

[a] *Department of Chemical and Biological Engineering, University of Wisconsin-Madison, Madison, WI, USA.*
[b] *Department of Chemistry, University of Wisconsin-Madison, Madison, WI, USA.*
† These authors contributed equally. * Corresponding authors.
Electronic Supplementary Information (ESI) available: See DOI: 10.1039/x0xx00000x

resistance, only a handful of antifungal drug classes are approved for use in the clinic, motivating a drive for the development of new treatments, including AMP-based therapeutics.[22-24] Recent rational design efforts by our group have resulted in α/β-peptides that exhibit significantly enhanced antifungal selectivity, with up to a 22-fold improvement over aurein 1.2.[19]

The antifungal selectivities of α/β-peptides in our prior work were quantified using a selectivity index (SI). This parameter is the ratio between mammalian cell toxicity (often defined as 'HC_{10}', or the concentration of peptide required to cause 10% hemolysis, or the lysis of human red blood cells) and antifungal activity (defined as the minimum inhibitory concentration, or 'MIC', needed to inhibit at least 90% growth of fungal cells).[9] The SI is used as a metric to guide the design of highly selective (high-SI) antimicrobial AMPs and mimetics.[9, 25, 26] However, obtaining AMPs and peptidomimetics with high antifungal SI remains challenging compared to the design of highly selective antibacterial AMPs because antifungal toxicity correlates more strongly with hemolysis than does antibacterial toxicity.[19, 27] This challenge is largely attributed to the more similar membrane compositions of eukaryotic fungal and mammalian cells compared to prokaryotic bacterial cells, which affect the membrane-active mechanism of most AMPs and mimetics.[28-30] While many peptides can disrupt membranes, the most valuable therapeutic option would be the ones that are most selective.

Our past studies have shown that experimentally determined peptide physicochemical properties such as helicity and hydrophobicity can be used to accurately predict fungal activity and hemolysis metrics, in addition to providing specific ranges over which high selectivity could be expected.[9, 19] However, experimentally determined physicochemical properties require low-throughput biophysical characterization. In addition, most physicochemical properties of short peptides are highly correlated, rendering rational design a difficult strategy because altering a single amino acid can result in changes in multiple physicochemical properties.[31] Consequently, rational design approaches require educated guesses about the results of a single amino acid substitution followed by α/β-peptide synthesis, an overall approach that is often prohibitive at large scale in terms of time and financial cost. Therefore, a predictive method to navigate the sequence space and discover novel α/β-peptide sequences with improved antifungal selectivity would be useful and could significantly expedite development. Unfortunately, the prediction of novel α/β-peptide sequences is hindered by the relative lack of experimental data available in databases compared to natural α-peptide AMPs (such as APD3,[32] CAMPR3,[33] and DRAMP[34]). This data scarcity limits the application of large database-driven machine learning (ML) methods, such as deep learning,[35, 36] natural language models,[37] and variational autoencoding,[38] which have proven useful in past studies for the prediction of novel canonical AMP sequences. Additionally, sequence-based data representations for α-peptide AMPs involving the 20 canonical amino acids (e.g., one-hot encoding,[39] quantitative matrices[40]) are not directly transferable to AMPs containing synthetic amino acids. Consequently, predicting new sequences of synthetic peptidomimetics remains challenging.

In this study, we demonstrate that highly selective α/β-peptidomimetics can be discovered by iteratively predicting their biological activities *in silico* using an ML model with computationally derived descriptors and then measuring selectivity using *in vitro* experimental measurements to test model predictions and update model parameters. Using an initial dataset of 147 α/β-peptide sequences as a starting point,[9, 19] we developed an iterative Gaussian process regression (GPR) workflow to predict HC_{10} and MIC values of novel α/β-peptide sequences templated on the most selective sequences in this initial training set. As input to the GPR model, we used 2D molecular descriptors that can be quickly computed from peptide chemical structures without relying on large databases or computationally expensive predictive techniques. Peptides were selected for synthesis based upon GPR predictions, with experimentally measured HC_{10} and MIC values then used to update GPR model parameters to guide the next prediction round. Because GPR estimates the uncertainty of regression predictions,[41-45] this approach permits evaluation of promising new sequences with low uncertainty and high predicted SI while also testing sequences with high uncertainty to expand the design space (e.g., sequences with new amino acids) for future prediction rounds. After six rounds, we identified new selective α/β-peptides with up to a 52-fold improvement in SI compared to aurein 1.2. Although these new peptides possess desirable physicochemical properties typically indicative of high antifungal selectivity, their sequences, which largely contain unconventional amino acid substitutions, would be less likely to be identified through rational design. Our results demonstrate the effectiveness of an iterative GPR-guided strategy to screen through a large sequence space and identify multiple new α/β-peptides with significantly enhanced antifungal selectivity. These peptides could, with further development, prove valuable as therapeutic agents to combat the rising incidence of drug-resistant fungal infections.

Results and Discussion

Properties of Training Sequences and Design Space Generation

The initial training set for the iterative GPR predictive workflow included 147 13-amino-acid α/β-peptides templated on aurein 1.2 across four different synthetic α/β-peptide backbone motifs: ααβ, αααβ, ααβαααβ, and αβαβααβ, all of which adopt a helical conformation.[10, 13, 46] Fifteen sequences of the training set are newly reported (**Appendix I**). These peptides were designed in our past studies to maximize antifungal selectivity.[9] As depicted in **Figure 1a**, β-amino acids contain an additional backbone carbon compared to canonical α-amino acids, which is prefixed with a "β" in the amino-acid sequence. The β-amino acid 'ACPC' (trans-2-aminocyclopentane-carboxylic acid, simplified as 'X') is a cyclic 5-membered ring that improves helical folding and stability.[14, 47-49] **Figure 1b** visualizes the occurrence of each α- and β-amino acid in the initial 147 sequence training set (dashed horizontal line) in terms of their polarity: nonpolar (orange), polar (dark blue), charged basic (cyan), and charged acidic (red). Only 13 of the 20 possible

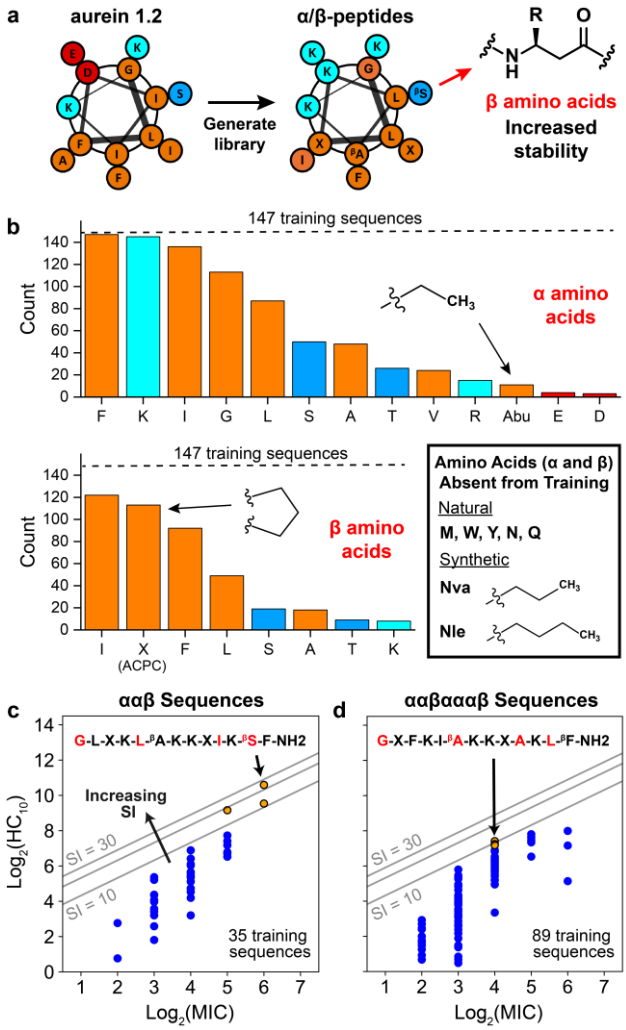


Figure 1. Overview of α/β-peptides and training design space for the Gaussian process regression workflow. (a) Helical wheel representation of aurein 1.2, which was used as the template sequence to generate a library of 147 α/β-peptide sequences (used as the initial training set in our study) in previous rational design studies. Amino acid types are color-coded as follows: nonpolar (orange), polar (dark blue), basic (cyan), and acidic (red). "β" refers to amino acids with an additional backbone carbon to increase stability compared to traditional α-amino acids and R denotes side chain. (b) Number of training sequences (out of 147, denoted with dashed line) that contain at least one of each α- (top) and β- (bottom) amino acid. Amino acids (α and β) absent from the training peptides but utilized in this study are boxed at lower right. Amino acids are color coded in the same manner as in the previous panel. (c-d) Distribution of training sequences for the (c) αaβ and (d) αaβαααβ backbones according to their hemolysis ($Log_2(HC_{10})$) and antifungal activity ($Log_2(MIC)$). Grey lines indicate constant SI values equal to 10, 20, and 30. The top three SI template peptides for each backbone are shown as orange points. The corresponding amino-acid template sequence is shown with positions corresponding to varying amino acids in red. "—NH2 refers to an amidated C-terminus, whereas the N-terminus is +1-charged (these properties are shared for all peptide sequences). All other training sequences are shown as blue points. $Log_2(HC_{10})$ vs. $Log_2(MIC)$ distributions for the other two backbones used for model training (αααβ and αβαβααβ) are shown in Figure S1.

canonical α-amino acids were present in the 147-peptide training dataset. Additionally, only five α-amino acids (phenylalanine 'F', lysine 'K', isoleucine 'I', glycine 'G', leucine 'L') and three β-amino acids (β-homoisoleucine '$^{\beta}$I', ACPC 'X', β-homophenylalanine '$^{\beta}$F') were represented in at least half of the training sequences. Therefore, a large sequence design space remains unexplored, even for short 13-amino-acid peptides.

To reduce this design space, we focused on templating and improving on high selectivity (high-SI) peptides in the training set based on amino-acid positions that vary amongst them (red amino acids in **Figures 1c-d**). The SI is defined as the ratio of experimentally determined peptide HC_{10} to MIC measurements as shown in Equation 1:

$$SI = \frac{HC_{10}}{MIC} \qquad (1)$$

From two-fold serial dilutions, HC_{10} was interpolated continuously, whereas MIC was discretely determined to stay consistent with conventional methodologies.[9, 19] All HC_{10} and MIC data were Log_2 scaled to accommodate the serially diluted nature of the experimental measurements and higher-SI peptides trend towards the top left of the plots in **Figures 1c-d**.

The three highest-SI peptides of the 147 training sequences consisted of the αaβ backbone (indices **#241, #239, #231**) with SI values ranging from 11.6 to 24.1; we selected these three peptides as template peptides for the αaβ backbone. The sequences of these peptides varied in amino-acid positions 5, 10, and 12, and we additionally chose to vary amino-acid position 1 during the iterative GPR workflow to investigate the effects of N-terminal substitutions on peptide hemolysis and antifungal activity. Previous experimental structure-activity studies have found that substituting the N-terminal 'G' in aurein 1.2 with 'A' significantly decreased activity against *C. albicans*,[50] whereas a 'G' to 'A' substitution increased broad-spectrum activity against bacteria.[51] The three template αaβ-peptide sequences are shown as orange points in **Figure 1c** along with the other 32 training sequences on the αaβ backbone (dark blue points) for reference. The 4th-6th highest SI peptides in the training set (**#133, #029, #131**) consisted of the αaβαααβ backbone with SI values ranging from 9.1 to 10.6 (orange points in **Figure 1d**); we selected these three peptides as template peptides for the αaβαααβ backbone. We chose to consider amino acid variations in these three peptides (at positions 1, 6, 10, 12) given their high antifungal activity, albeit with lower selectivity compared to the αaβ-peptides. **Figure 1c** and **Figure 1d** shows the two resulting template sequences (for the αaβ and αaβαααβ backbones, respectively) with the four amino-acid positions in each sequence that were varied during the GPR approach indicated in red.

For α positions in the template sequences, we considered the 13 amino acids included in the training data (top histogram in **Figure 1b**) plus five natural amino acids (methionine 'M', tryptophan 'W', tyrosine 'Y', asparagine 'N', glutamine 'Q') and two hydrophobic synthetic amino acids (norvaline 'Nva', norleucine 'Nle') for a total of 20 possible substitutions. Nva and Nle are extensions of the α-aminobutyric acid (Abu) amino acid present in training peptides and were included based on

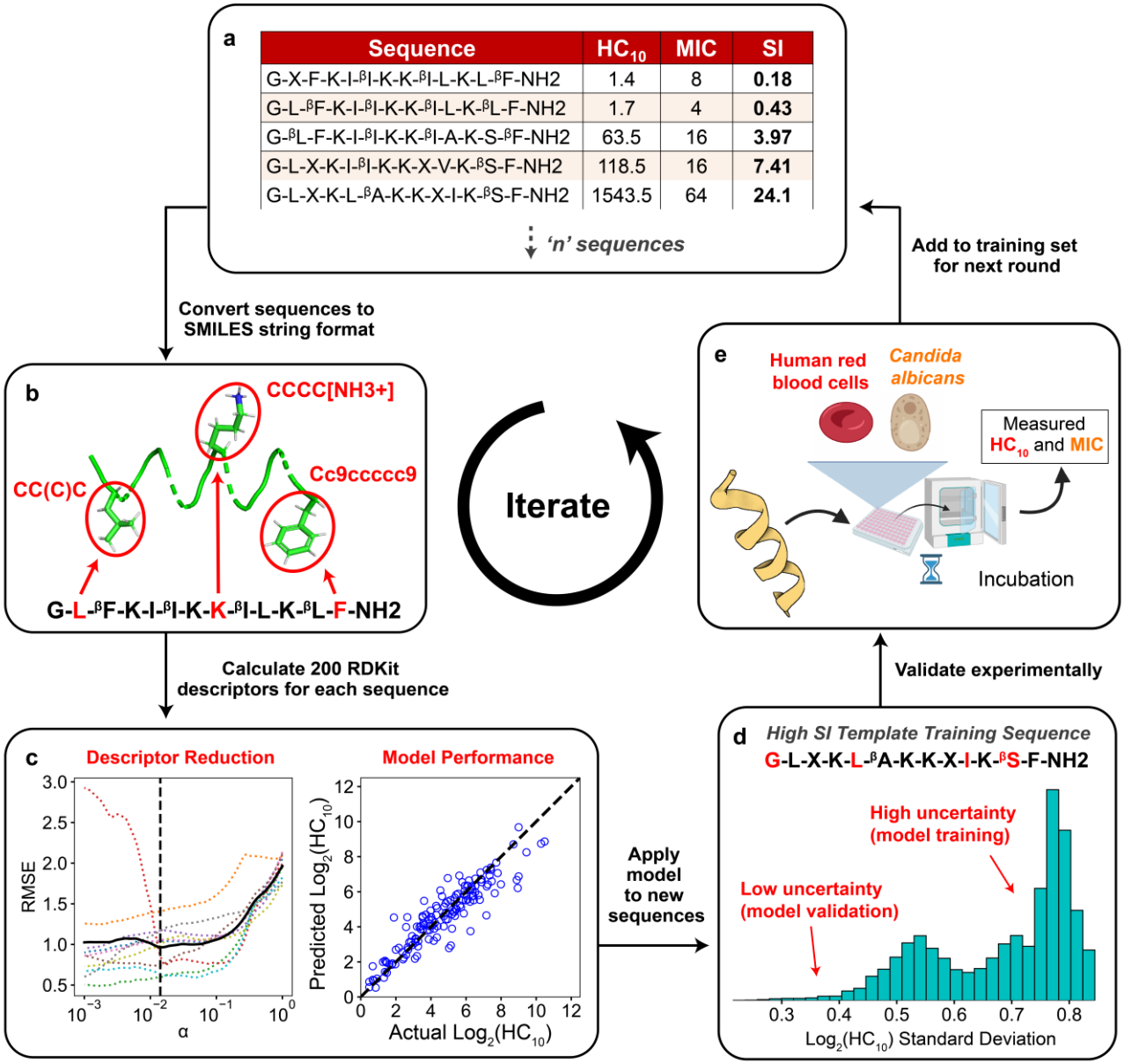


Figure 2. Schematic demonstrating the integrated computational and experimental iterative approach used in this study. (a) Training set of 'n' sequences with corresponding HC_{10} and MIC labels for Gaussian process regression (GPR). (b) All sequences were converted to SMILES string representations (**Section S2**), in which selected amino acids are highlighted in red for visualization. (c) After calculating 200 descriptors for each sequence and removing all constant descriptors, LASSO cross validation was used to reduce the set of descriptors. The reduced set of descriptors was then used to train separate GPR models for HC_{10} and MIC using data in the training set (right plot, see **Table S4** for model parameters). (d) The trained GPR model was applied to predict HC_{10} and MIC for all peptides in the design space generated from the template sequence. New sequences were selected for experimental synthesis based on a balance between sequences with low uncertainty predictions for model validation and high uncertainty predictions for further model training. (e) HC_{10} and MIC values for new sequences were quantified experimentally, compared to model predictions, and added to the training set (a) for the next round. This figure includes graphics imported from Biorender.com.

previous studies suggesting that these amino acids increase membrane binding by promoting positive membrane curvature[52] and improve antimicrobial activity by increasing peptide hydrophobicity.[53] Three natural amino acids were not considered: proline 'P' due to destabilization of AMP folding by inducing kinks in helices,[54, 55] cysteine 'C' due to its capability to form disulfide bridges,[56] and histidine 'H' due to its sensitivity to physiological pH, leading to changes in charge.[57] For β positions in the template sequences, the β version of the 20 α-amino acids described above plus ACPC 'X' were considered for a total of 21 possible substitutions. Given the three α- and one β-amino acid positions varied for each backbone type (red amino acids in **Figures 1c-d**), two different design spaces of 168,000 (20^3×21) possible sequences were considered for this study for a total of 336,000 sequences.

Iterative GPR Workflow Implementation and Model Training

Figure 2 summarizes the iterative workflow developed and used in this study. The goal of this approach was to guide the selection of peptide sequences for experimental testing, and

thus model training and AMP design, by quantifying the uncertainty of model predictions obtained from GPR, building on studies that have successfully implemented iterative Gaussian-based techniques.[58, 59] For our approach, each sequence in the training set (**Figure 2a**; *i.e.*, the sequences with corresponding experimentally measured HC_{10} and MIC) was first transformed into its simplified molecular-input line-entry system (SMILES) representation (**Figure 2b**), which uniquely represents the bonding, branching, ring groups, and aromaticity of each peptide[60] through the definition of both backbone and sidechain fragments (see **Section S2 and Figure S2** for more details). This choice was motivated by the ability of the SMILES data representation to capture important molecular properties of non-canonical amino acids in α/β-peptide sequences, unlike other popular sequence encodings that are only applicable to α-amino acids.[39, 40] We utilized these SMILES strings as input to the RDKit Cheminformatics toolkit to calculate 200 molecular descriptors for each sequence.[61] Descriptors are numerical values associated with each sequence that quantify properties such as atomic van der Waals surface areas (VSA), partial charges, bonding and branching complexity, and chemical group counts (**Section S3**).

The total number of descriptors was large compared to the number of training sequences, indicating that model training could potentially lead to overfitting, which would not generate robust predictions. Therefore, the total number of descriptors used during model training in each prediction round was reduced in two steps. First, all descriptors that had constant numerical values for all sequences in the training set were excluded because they cannot distinguish between properties of different sequences (**Figure S3**). Second, we performed 10-fold LASSO cross validation (CV) (left plot in **Figure 2c**) to minimize the average root-mean-square error of model predictions compared to the measured HC_{10} and MIC values (further details in **Section S3 and Figure S4**). This approach quantifies the trade-off between the number of descriptors and prediction accuracy to identify a reduced set of descriptors that yields acceptable prediction accuracy. This procedure resulted in between 13 and 40 descriptors used for model training for each round. Utilizing this reduced descriptor set per round (**Tables S1-S2**), the highest accuracy (R^2) GPR model (right plot in **Figure 2c**) from a hyperparameter search was then selected (see **Table S4**) for HC_{10} and MIC predictions of novel α/β-peptide sequences. We note that separate models, using separate sets of descriptors, were trained to independently predict HC_{10} and MIC values. GPR model selection criteria are described in **Section S4**. For all regression steps, HC_{10} and MIC values were Log_2 scaled (**Figure S6**) to simplify comparisons to experimental values of hemolysis and antifungal activity, which are typically assessed, by convention, using 2-fold serial dilutions of peptide concentrations.

The two trained GPR models (for HC_{10} and MIC) were then used to assess peptides from the design space of 168,000 possible sequences (for a single template sequence), which we refer to as test sequences. Test sequences were converted to SMILES representations and corresponding descriptors were calculated. Due to the large range of descriptor values for the test sequences (up to six times larger than the upper bound of descriptor values for training sequences, see **Figures S7-S8**), we only used the GPR models to predict HC_{10} and MIC values for test sequences for which all descriptor values were within the bounds of the values computed for training sequences to prevent large prediction inaccuracies due to extrapolation (**Table S5**). We then utilized the ability of GPR to quantify prediction uncertainty through standard deviations (σ) to inform the selection of test sequences for experimental synthesis. To compare prediction uncertainties across prediction rounds by accounting for the changes in standard deviation distributions that resulted from updating the descriptors (**Section S3**) and the GPR models (**Section S4**) every round, we define the normalized standard deviation (NSD), which varies from 0 to 1, in Equation 2:

$$NSD = \frac{\sigma - \sigma_{min}}{\sigma_{max} - \sigma_{min}} \qquad (2)$$

σ is the standard deviation of a prediction for a test sequence and σ_{min} and σ_{max} are the minimum and maximum standard deviations of all predicted values for a prediction round (considering only test sequences with all descriptors within the lower and upper bounds of training descriptors, see **Section S6**). Test sequences with low NSD were selected for experimental synthesis to validate the model's predictive capabilities, whereas sequences with high NSD were chosen to further train the model (**Figure 2d**). GPR predictions of HC_{10} and MIC for these new sequences per prediction round were compared to experimental hemolysis and antifungal activity assays, respectively (**Figure 2e**), and then added to the training set (**Figure 2a**) for subsequent rounds. The overall approach was performed in six prediction rounds starting from the initial 147-sequence training set. Across these six rounds, this approach led to a total of 23 newly discovered peptides (*i.e.*, peptides selected for experimental synthesis and measurement of HC_{10} and MIC).

This iterative GPR approach demonstrates stable accuracy per prediction round for both HC_{10} and MIC (black lines in **Figure 3a**), and R^2 values for the comparison of GPR-predicted and experimentally-determined HC_{10} and MIC were comparable to previous efforts to predict MIC for a variety of antimicrobial compounds using simple sequence- or SMILES-based input data representations.[62-64] As expected, MIC accuracy was consistently lower than HC_{10} given the nature of antifungal activity assays, where, by convention, the maximum of three replicates from serially diluted experiments, rather than the average, is taken as the MIC [9, 65, 66] (**Section S9**). Therefore, continuous Log_2(MIC) GPR predictions were rounded up to the nearest whole number to match this experimental methodology. For Rounds 1 to 3, the selected set of descriptors remained the same (see 'Initial' column in **Tables S1-S2**). However, given a large increase in the maximum error between experimentally measured HC_{10} and MIC versus GPR predictions for both quantities in Round 3 (red curves in **Figure 3a**) after the introduction of a large HC_{10} and MIC sequence to the training data (sequence **2-4**, see **Figure 4a** and **Figures S9-S10**), we

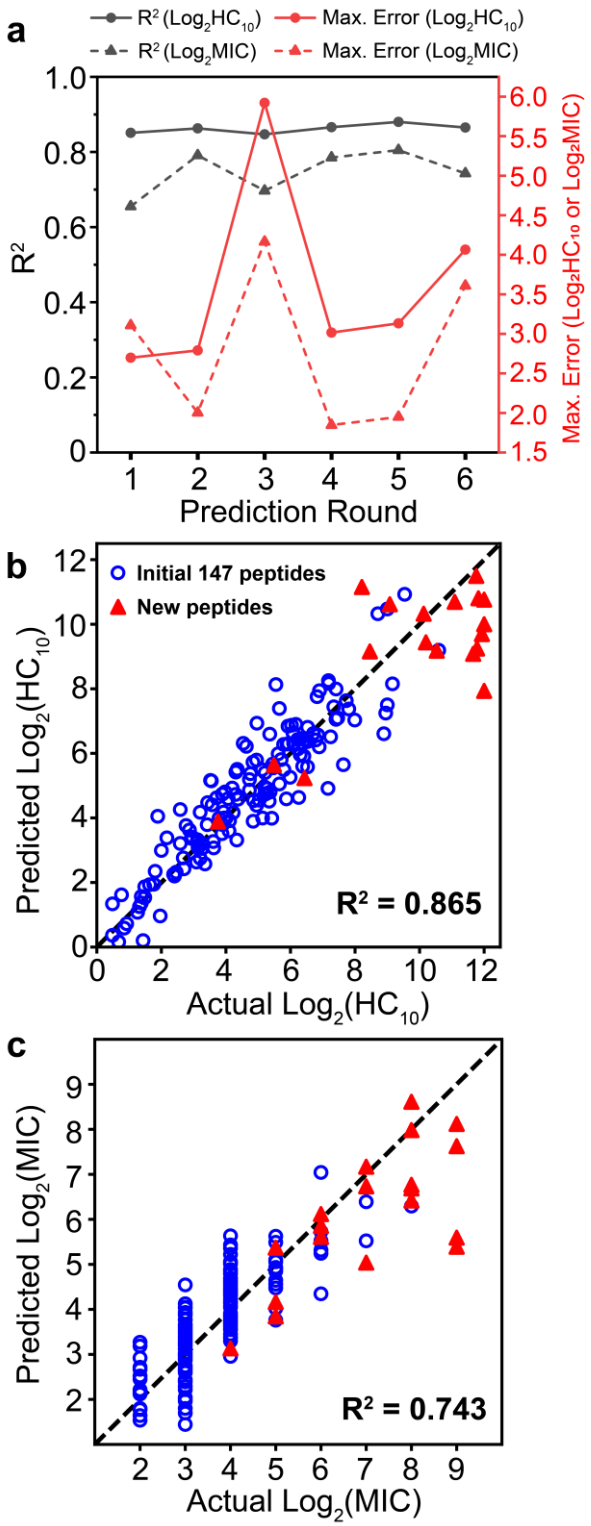


Figure 3. (a) R^2 (black lines) and maximum error (red lines) metrics calculated per prediction round for separate HC_{10} (solid) and MIC (dashed) prediction workflows. (b-c) Comparison of GPR-predicted vs. actual experimentally measured values of (b) $Log_2(HC_{10})$ and (c) $Log_2(MIC)$. GPR predictions used the final trained GPR model for the 6th prediction round. Points are validation set predictions from 10-fold cross-validation; that is, each point is the predicted value from when the corresponding sequence is not used for model training. The original 147 training peptides are shown as open blue circles and new peptides discovered with GPR are shown as red triangles. Coefficients of determination (R^2) are included in bold.

revised the workflow to update the selected set of descriptors through LASSO CV during each subsequent round from Round 4 onwards as described above. This approach better predicted high HC_{10} and MIC values for newly introduced peptides, particularly compared to the consistent underprediction of high-value MIC labels in Rounds 1 to 3 (**Figure S10**).

Figure 3b-c shows parity plots for $Log_2(HC_{10})$ and $Log_2(MIC)$ predictions for the final GPR model used for Round 6, which not only visualize the final prediction accuracy, but also highlight the novel sequences introduced (red triangles) and their distributions of $Log_2(HC_{10})$ and $Log_2(MIC)$ values compared to the original 147 training sequences (blue open circles). As part of our evaluation of model robustness across prediction rounds (**Figure S11-S13**), we also probed the predictive power of the GPR models through y-randomization (**Figure S13**),[67, 68] which compares randomly shuffled labels with the ones used for model development. Overall, these results verified that the predictions made by the model were not from random chance, thereby validating the rigor of our descriptor and GPR model selection approach.

GPR Guides the Discovery of α/β-peptide Sequences with Novel Amino Acids and Motifs

Figure 4a shows the results of the GPR workflow for the 23 newly discovered peptides experimentally tested during the six prediction rounds (**Figures S17, S18**), which are plotted as stars in **Figures 4b-c** and compared to the initial training sequences in blue (**Figure 1c-d**). In general, NSD values aligned well with the difference in actual (Act.) vs. predicted (Pred.) labels, and for newly introduced amino acids or motifs, there was an initial prediction inaccuracy that was resolved when reintroduced in later prediction rounds. For instance, sequences **1-1** (new Nle) and **2-1** (new $^{\beta}$F-F C-terminal motif) were underpredicted by a factor of approximately two for both $Log_2(HC_{10})$ and $Log_2(MIC)$. Sequences **2-3** (containing Nle) and **3-1** (containing $^{\beta}$F-F), however, were predicted with excellent accuracy relative to experimental measurements. Additionally, NSD values correlated with how chemically disparate newly introduced amino acids were from amino acids present in training sequences. For instance, sequence **3-1** was predicted with relatively low uncertainty (NSD < 0.30 for both $Log_2(HC_{10})$ and $Log_2(MIC)$ predictions) despite the new Nva amino acid. This is consistent with Nva having linear alkyl side chain that is intermediate in length compared to Abu and Nle (reference chemical structures in **Figure 1b**), which were present in the training set by Round 3. In contrast, peptides **2-4**, **3-2**, and **3-4**, each of which each introduced amino acids '**Q**,' '$^{\beta}$Y,' and 'W,' containing amine, phenol, and indole chemical groups, respectively, had higher prediction uncertainty (NSD > 0.70), even with a single amino-acid substitution relative to training sequences.

To further support the ability of the GPR model to learn physicochemical properties of new amino acids, we calculated the NSD distribution of each newly introduced amino acid in Rounds 1 to 3 since the set of descriptors for HC_{10} and MIC predictions remained the same for these prediction rounds

a

Round	Idx	Dif	New AA	Sequence	$Log_2(HC_{10})$ Act.	$Log_2(HC_{10})$ Pred.	$Log_2(HC_{10})$ Dif.	$Log_2(HC_{10})$ NSD	$Log_2(MIC)$ Act.	$Log_2(MIC)$ Pred.	$Log_2(MIC)$ Dif.	$Log_2(MIC)$ NSD	Actual SI
1	1	1	yes	Nle-L-X-K-I-βA-K-K-X-I-K-βA-F-NH2	10.1	7.7	2.4	0.42	7	5	2	0.62	8.7
	2	1	no	A-L-X-K-I-βA-K-K-X-I-K-βA-F-NH2	11.8	8.6	3.3	0.30	8	6	2	0.53	14.2
	3	1	no	Abu-L-X-K-I-βA-K-K-X-I-K-βA-F-NH2	11.8	8.6	3.1	0.32	8	6	2	0.58	13.5
2	1	1	no	G-L-X-K-I-βA-K-K-X-I-K-βF-F-NH2	6.4	4.2	2.3	0.18	5	3	2	0.07	2.7
	2	1	no	G-L-X-K-I-βA-K-K-X-L-K-βS-F-NH2	10.5	9.8	0.7	0.01	6	6	0	0.10	23.1
	3	1	no	G-L-X-K-Nle-βA-K-K-X-I-K-βS-F-NH2	10.2	9.9	0.3	0.28	6	6	0	0.21	18.3
	4	1	yes	Q-L-X-K-I-βA-K-K-X-I-K-βA-F-NH2	12.0	6.2	5.8	0.85	9	5	4	0.78	8.0
3	1	1	yes	G-L-X-K-Nva-βA-K-K-X-I-K-βF-F-NH2	5.5	5.7	-0.2	0.27	5	4	1	0.21	1.4
	2	1	yes	G-L-X-K-Nle-βA-K-K-X-I-K-βY-F-NH2	3.8	4.4	-0.7	0.88	4	4	0	0.72	0.8
	3	1	no	A-L-X-K-Nle-βA-K-K-X-I-K-βS-F-NH2	9.1	10.8	-1.7	0.40	8	6	2	0.62	2.1
	4	1	yes	Abu-L-X-K-W-βA-K-K-X-I-K-βA-F-NH2	8.2	7.5	0.7	0.82	7	6	1	0.81	2.3
4	1	2	no	G-L-X-K-V-βA-K-K-X-I-K-βK-F-NH2	11.8	9.5	2.3	0.17	7	5	2	0.20	27.5
	2	3	no	G-L-X-K-V-βA-K-K-X-V-K-βK-F-NH2	12.0	10.3	1.7	0.19	8	5	3	0.24	16.0
	3	2	yes	K-L-X-K-L-βA-K-K-X-L-K-βG-F-NH2	11.7	8.6	3.0	0.23	8	5	3	0.23	12.6
	4	4	no	R-L-X-K-V-βA-K-K-X-V-K-βK-F-NH2	12.0	11.5	0.5	0.35	9	6	3	0.38	8.0
5	1	1	no	G-X-F-K-I-βA-K-K-X-V-K-A-βF-NH2	8.5	8.6	-0.2	0.14	5	5	0	0.10	11.0
	2	3	no	I-X-F-K-I-βA-K-K-X-T-K-R-βF-NH2	11.9	8.8	3.1	0.39	9	5	4	0.25	7.6
	3	4	no	I-X-F-K-I-βS-K-K-X-T-K-Abu-βF-NH2	12.0	7.6	4.4	0.36	9	4	5	0.39	8.0
	4	1	no	K-X-F-K-I-βA-K-K-X-V-K-T-βF-NH2	11.1	10.2	0.9	0.18	6	6	0	0.16	34.0
6	1	1	no	G-X-F-K-I-βA-K-K-X-V-K-S-βF-NH2	11.1	10.7	0.3	0.13	6	7	-1	0.08	33.6
	2	1	no	K-X-F-K-I-βA-K-K-X-S-K-L-βF-NH2	10.2	10.8	-0.6	0.16	6	7	-1	0.09	18.8
	3	2	no	K-X-F-K-I-βA-K-K-X-T-K-V-βF-NH2	11.8	10.9	0.8	0.16	7	7	0	0.09	27.8
	4	1	no	K-X-F-K-I-βA-K-K-X-V-K-R-βF-NH2	11.8	10.6	1.3	0.28	6	7	-1	0.16	57.1

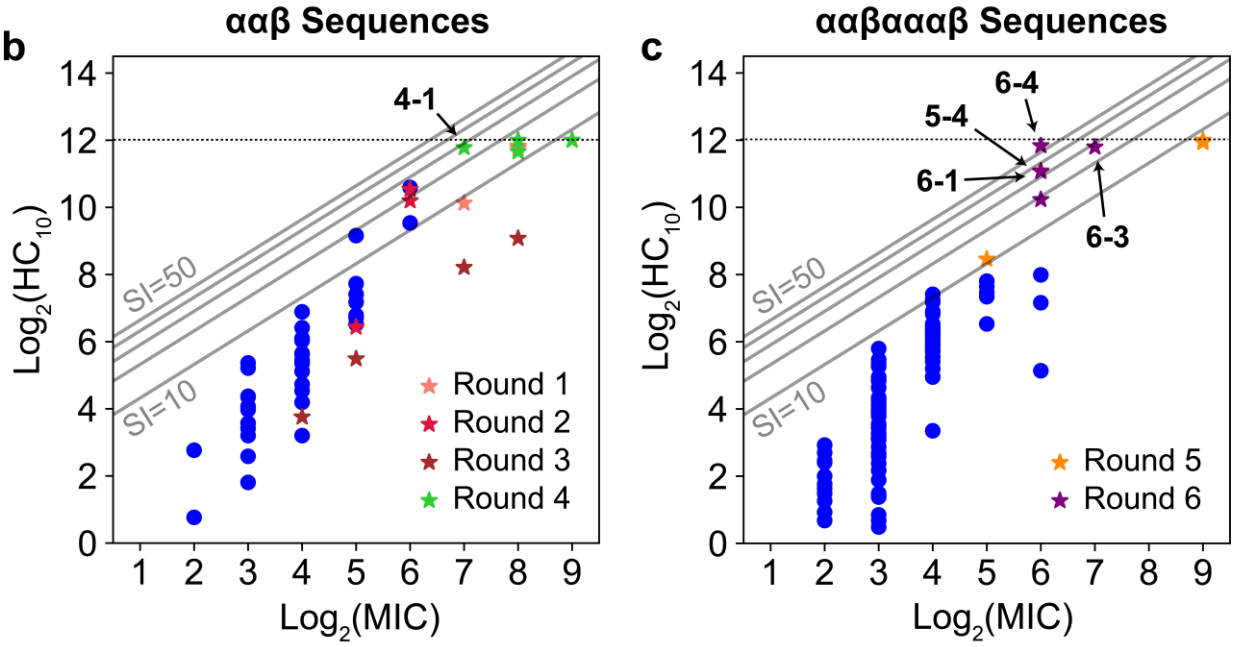


Figure 4. (a) Summary of GPR prediction results. Amino-acid substitutions relative to any training sequence are in red. The 'Dif' column quantifies the number of these substitutions, and the 'New AA' column denotes if any substitution is an amino acid not present in the training data for that round. Experimentally measured (Act.), GPR-predicted (Pred.), and their difference are shown for both $Log_2(HC_{10})$ and $Log_2(MIC)$. NSD refers to the normalized GPR standard deviation and ranges from 0 (dark green; low uncertainty) to 1 (dark red; high uncertainty). SI computed from the experimentally measured HC_{10} and MIC is also shown. (b-c) $Log_2(HC_{10})$ vs. $Log_2(MIC)$ distributions comparing original training (dark blue dots) vs. new peptides added during the iterative workflow (stars) for prediction rounds (b) 1 to 4 and (c) 5 to 6. Grey diagonal lines denote SI bands with values of 10, 20, 30, 40, and 50. All peptides discovered with larger SI values than the original highest-SI peptide from the initial training set (SI=24.1) are indicated in bold. The horizontal dotted line at $Log_2(HC_{10}) = 12$ denotes the upper limit of the hemolysis assay, corresponding to a peptide concentration of 4096 μg/mL. The test design space and corresponding standard deviation ranges for all GPR prediction results are visualized in Figure S15.

(**Tables S1-S2**). Bar and whisker plots of these NSD distributions are plotted in **Figure S14** for the Nle, 'Q', Nva, 'βY,' and 'W' amino acids, which show that, in general, there was a large decrease in the average NSD (therefore, a decrease in prediction uncertainty) for these amino acids in prediction rounds following their introduction. Interestingly, Nva (**Figure S14c**) largely followed the same trends as Nle (**Figure S14a**), despite not being introduced prior to the changes observed in

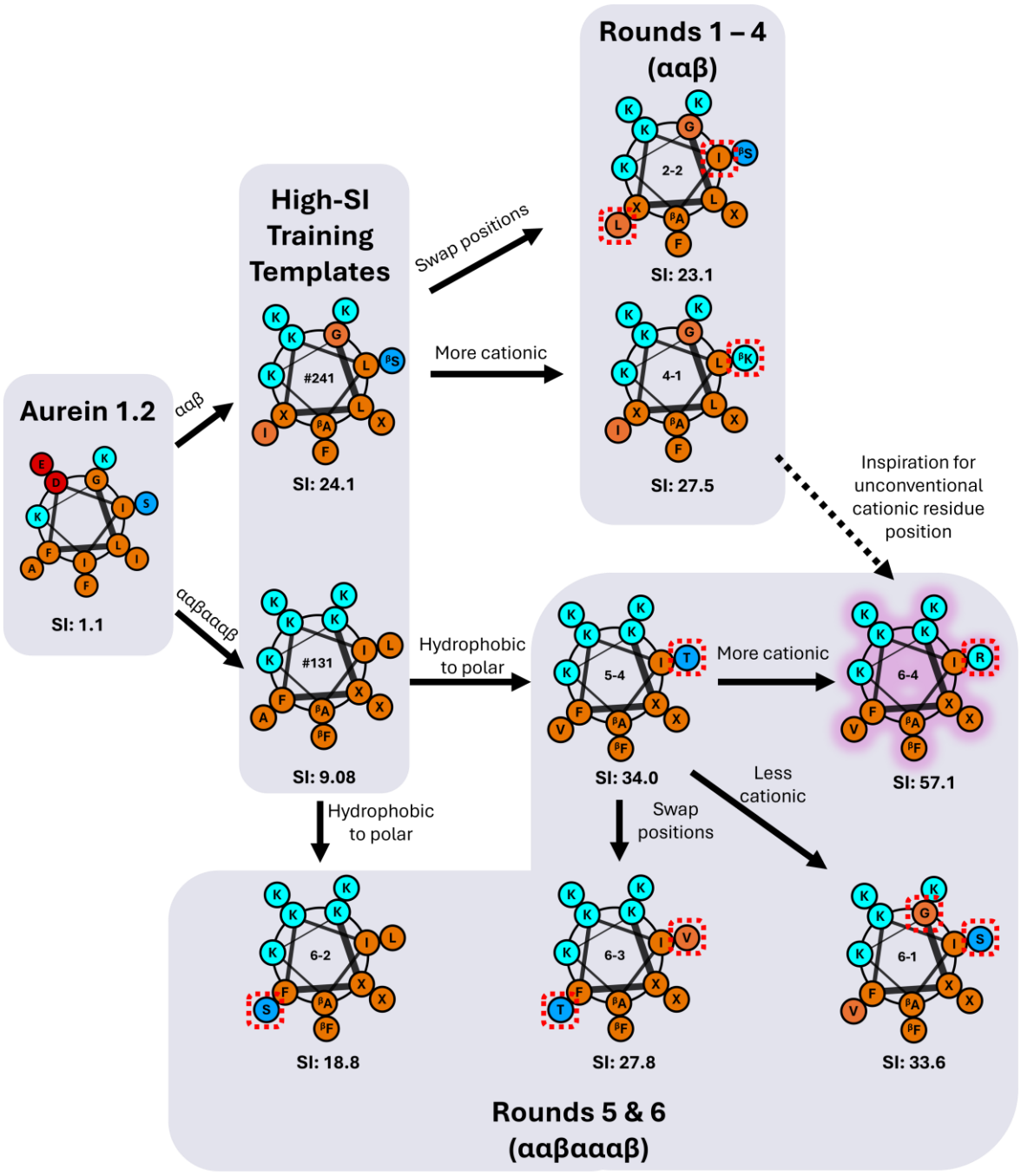


Figure 5. Flowchart of the iterative GPR-guided sequence-substitutions of α/β-peptides templated on aurein 1.2 that resulted in high selectivity against *C. albicans*. Hydrophobic amino acids are shown in orange, cationic amino acids in cyan, anionic amino acids in red, and neutral polar amino acids in blue. β-Amino acids are visualized with a subscript and ACPC is represented as X. Red dashed boxes represent single amino-acid substitutions between each peptide. Each peptide is labelled with its experimentally measured SI. The highest-SI peptide discovered by the iterative GPR approach is highlighted in purple.

NSD. This observation supports the model's ability to transfer learned information from amino acids present in the training sequences to new amino acids that have similar chemical structures, and highlights the advantage of 2D molecular descriptors over more traditional data representations for peptide sequences such as one-hot or orthogonal encoding,[39, 69] which would distinguish Nva and Nle as separate amino acids with no intuition for their chemical similarity.

GPR Approach Yields Multiple Highly Selective α/β-Peptides and Identifies an Unconventional Cationic Amino Acid Substitution

Figure 5 visualizes amino acid substitutions of aurein 1.2 that resulted in the highest-SI peptides discovered through the iterative GPR approach. We first note that the highest-SI template peptides (**#241,** SI = 24.1; **#131**, SI = 9.08) previously developed through rational design exhibited improved selectivity compared to aurein 1.2.[9, 19] In the initial prediction rounds (1 to 4), peptides were synthesized with the focus on model training by introducing new amino acids and previously

utilized amino acids in novel positions. In this process, two high-SI peptides were discovered, one of which featured an unconventional cationic amino acid (**4-1**, '$^{\beta}$K' substitution) on the hydrophobic face, a substitution that was not explored in our prior rational design attempts because of concerns that reductions in amphiphilicity and hydrophobicity would result in excessive losses in antifungal activity.[19, 70] As expected, these newly discovered high-SI peptides did not exhibit increased antifungal activity. Instead, the decrease in hemolytic activity was significant, which drove an increase in selectivity. Therefore, the initial rounds demonstrate that peptides discovered by the model improved selectivity by reducing hemolysis rather than by increasing antifungal activity.

In Rounds 5 and 6, the more antifungal ααβαααβ backbone was explored as the template (**Figure 1d**) using the same workflow as in Rounds 1 to 4, but by switching the design space to both evaluate the transferability of our approach and to obtain higher-SI peptides. This change enabled us to use the large amount of low hemolysis (high $Log_2(HC_{10})$) training data from Rounds 1 to 4 (**Figure 4a**) to probe whether the model could be applied to a different peptide backbone with inherently higher antifungal activities (lower Log_2(MIC)) to find new high-SI peptides. Interestingly, this approach was successful in discovering four additional high-SI peptides (**5-4**, **6-1**, **6-3**, **6-4**; 27.8 < SI < 57.1) with high prediction accuracy and low uncertainty (low NSD) while drawing on learned trends from previous prediction rounds on the ααβ backbone (**Figure 4c**). For instance, as seen in peptides **4-3** and **4-4**, 1st position cationic (K, R) amino acids tend to produce peptides with low hemolysis (high $Log_2(HC_{10})$) and good selectivity (high-SI), which was reflected by many high-SI ααβαααβ candidate sequences in Rounds 5 and 6 (**5-4**, **6-2**, **6-3**, **6-4**) with a 1st position 'K.' Additionally, drawing on the ability of peptide **4-1** to increase selectivity with a cationic amino-acid substitution on the hydrophobic face of the ααβ backbone, an arginine (R) substitution on the hydrophobic face of a high-SI ααβαααβ template (**#131**) was predicted and experimentally validated, resulting in an over 6-fold improvement in selectivity (**6-4** in **Figure 5**). With an SI of 57.1, peptide **6-4** was the highest-SI peptide identified in this study and showcases a substantial increase in antifungal selectivity compared to the training data. Overall, this approach was able to discover 13 novel high-SI peptides (11.0 < SI < 57.1) with comparable or higher SIs than the template peptides (9.08 < SI < 24.1) (**Figure 4b-c**).

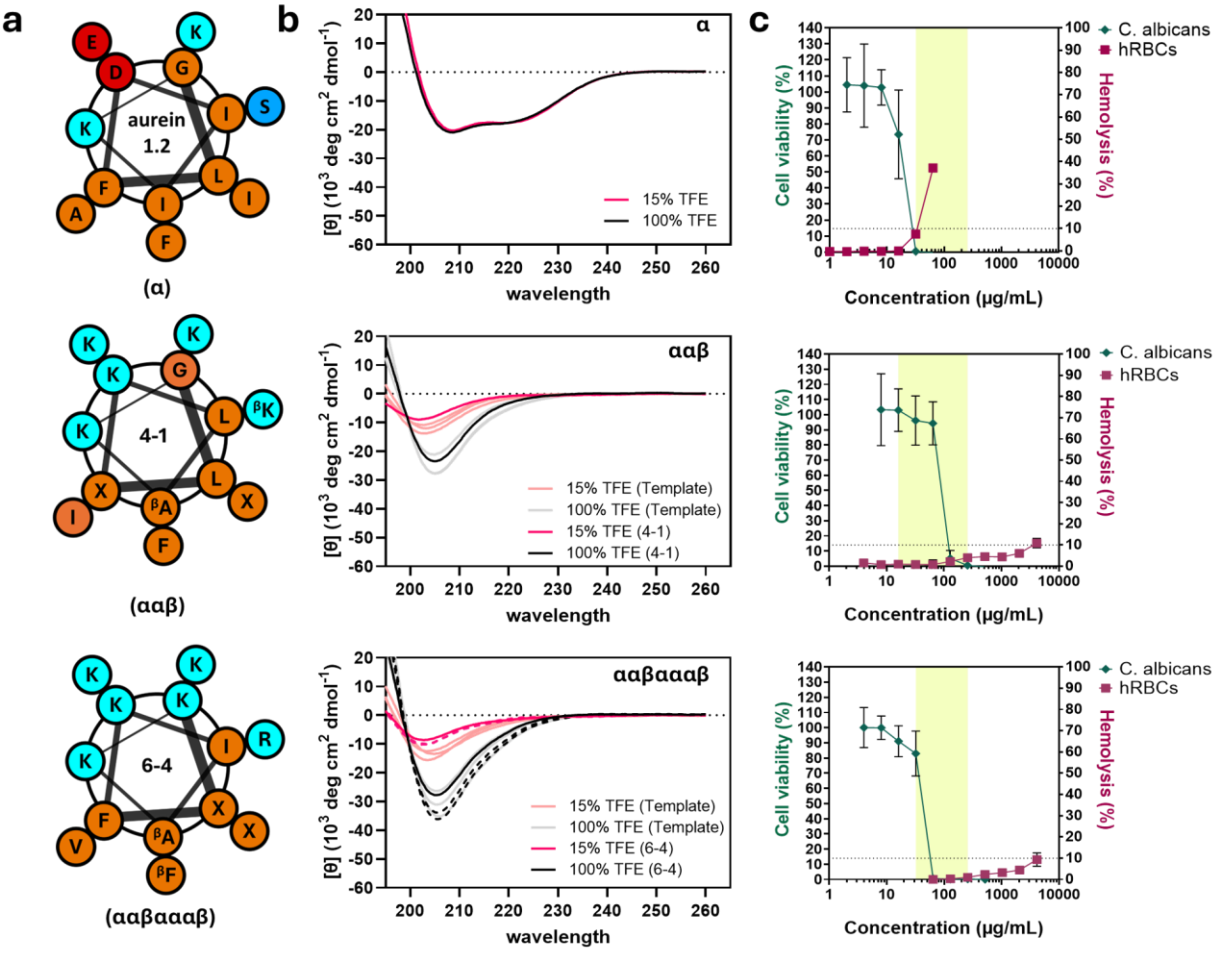
These results demonstrate that an iterative ML approach can successfully discover high-SI peptides through substitutions that would not have been rationally anticipated. Furthermore, the advantage of the SMILES representation for peptide sequences is that it is agnostic to the backbone type (combination of α and β amino acids), thereby promoting the transferability of the model across backbones in this study and, potentially, for other synthetic motifs (e.g., γ-amino acids with two additional backbone hydrocarbons compared to α-amino acids[71]) in the future.

Improved Antifungal Selectivity in GPR-Screened α/β-Peptides is Dependent on Charge, Hydrophobicity, and Helical Rigidity

To better understand the relationship between sequence, structure, and activity for the newly discovered high-SI α/β-peptides in comparison to aurein 1.2 and the high-SI template peptides, we experimentally measured their physicochemical properties (**Table 1**) and visualized representative comparisons

Table 1. Characterization of the physicochemical properties of aurein 1.2, template peptides, and high-SI newly discovered peptides. Peptide sequences are presented with one-letter amino acid codes. β-amino acids are labeled with a superscript, and ACPC amino acids are labeled as X. Amino-acid substitutions relative to any training sequence are in red. The average retention time was obtained from three independent analytical RP-HPLC measurements. Molar ellipticities (θ) at 100% TFE and 15% TFE were measured using circular dichroism in three independent experiments in triplicate (Figure S19, S20). Helical rigidity was calculated as the ratio between molar ellipticity in 15% TFE and in 100% TFE (Section S9). Measured molecular weight was obtained from ESI-MS (Table S7).

Peptide type	Peptide Idx	Sequence	αβ motif	Retention Time (min ± SD)	Charge	[θ] 100% TFE ± SD	[θ] 15% TFE ± SD	Helical rigidity %± SD	Measured Molecular Weight (Da)
Wild type	aurein 1.2	G-L-F-D-I-I-K-K-I-A-E-S-F-NH2	α	23.93 ± 0.01	1	-16.8 ± 0.4	-17.1 ± 0.1	102 ± 2	1479.79
Template peptides	#241	G-L-X-K-L-$^{\beta}$A-K-K-X-I-K-$^{\beta}$S-F-NH2	ααβ	15.45 ± 0.01	5	-27.2 ± 1.7	-9.1 ± 1.1	34 ± 5	1481.94
	#239	G-L-X-K-L-$^{\beta}$A-K-K-X-L-K-$^{\beta}$S-F-NH2	ααβ	16.11 ± 0.01	5	-27.4 ± 2.9	-10.5 ± 0.2	38 ± 4	1481.94
	#231	G-L-X-K-I-$^{\beta}$A-K-K-X-I-K-$^{\beta}$A-F-NH2	ααβ	14.95 ± 0.01	5	-20.5 ± 1.3	-12.3 ± 1.6	60 ± 9	1465.94
	#133	G-X-F-K-I-$^{\beta}$A-K-K-X-A-K-L-$^{\beta}$F-NH2	ααβαααβ	16.61 ± 0.01	5	-30.9 ± 1.1	-14.2 ± 2.0	46 ± 7	1499.96
	#131	K-X-F-K-I-$^{\beta}$A-K-K-X-V-K-L-$^{\beta}$F-NH2	ααβαααβ	15.55 ± 0.01	6	-26.2 ± 0.8	-11.6 ± 1.4	44 ± 5	1599.13
	#29	G-X-F-K-I-$^{\beta}$I-K-K-X-A-K-S-$^{\beta}$F-NH2	ααβαααβ	18.17 ± 0.01	5	-35.2 ± 1.6	-13.1 ± 0.9	37 ± 3	1515.96
Newly discovered peptides	4-1	G-L-X-K-**V**-$^{\beta}$A-K-K-X-I-K-$^{\beta}$**K**-F-NH2	ααβ	13.50 ± 0.02	6	-23.1 ± 0.8	-7.2 ± 0.6	31 ± 3	1509.01
	5-4	K-X-F-K-I-$^{\beta}$A-K-K-X-V-K-**T**-$^{\beta}$F-NH2	ααβαααβ	14.02 ± 0.04	6	-36.0 ± 0.6	-8.5 ± 1.6	24 ± 4	1587.06
	6-1	G-X-F-K-I-$^{\beta}$A-K-K-X-V-K-**S**-$^{\beta}$F-NH2	ααβαααβ	15.38 ± 0.08	5	-33.7 ± 1.0	-8.5 ± 0.9	25 ± 3	1501.96
	6-4	K-X-F-K-I-$^{\beta}$A-K-K-X-V-K-**R**-$^{\beta}$F-NH2	ααβαααβ	13.33 ± 0.04	7	-27.6 ± 2.2	-7.3 ± 1.7	27 ± 6	1642.10



d

Peptide idx	Conventional MIC	Smaller-interval MIC (sMIC)	HC_{10}	Conventional SI	**sSI**
Aurein 1.2	32	30	35.0	1.1	**1.2**
4-1	128	80	3516.7	27.5	**44.0**
5-4	64	50	2178.2	34.0	**43.6**
6-1	64	50	2147.8	33.6	**43.0**
6-4	64	60	3655.6	57.1	**60.9**

Figure 6. Characterization of the highest-SI newly discovered peptides for each backbone. Aurein 1.2 (top), αaβ-backbone peptide 4-1 (middle), and αaβαααβ-backbone peptide 6-4 (bottom) are shown. (a) Helical wheel representations of sequences and expected amphiphilic amino acid arrangement. Hydrophobic amino acids are shown in orange, cationic amino acids in cyan, anionic amino acids in red, and non-charged polar amino acids in blue. β-amino acids are visualized with a subscript and ACPC is represented as X. (b) Circular dichroism spectra of the three peptides in 15% (red) and 100% trifluoroethanol (black). The CD curves of other high-SI peptides with a αaβαααβ backbone are represented as dashed lines. (c) Comparisons between antifungal (*C. albicans*, green) and hemolytic (hRBC, red) activities of peptides along with their broad-spectrum MIC range (highlighted). Highlighted regions represent the effective range of MICs of each peptide against all fungal and bacterial cells tested. Data points are the averages of three independent experiments with at least two technical duplicates each and error bars represent the standard deviation. (d) Table of *C. albicans* MIC, smaller-interval MIC (sMIC; see text), hRBC HC_{10}, SI and smaller-interval SI (sSI) for the four highest SI peptides in comparison to aurein 1.2. sMIC was measured at 5 μg/mL intervals and sSI was calculated by HC_{10}/sMIC.

in **Figure 6**. Notably, the highest-SI peptides discovered by the iterative GPR approach (**4-1, 5-4, 6-1, 6-4**; 27.5 < SI < 57.1) exhibited lower hydrophobicity (i.e., smaller HPLC-retention times; RT) and reduced helical rigidity compared to the highest-SI template peptides (**#029-#241**; 9.08 < SI < 24.1) (**Table 1, Figure 6**). Peptide net charge either remained the same or increased by one compared to the template peptides, demonstrating the usefulness of such substitutions in certain positions. In general, all high-SI peptides in this study possessed lower hydrophobicity and helical rigidity, while possessing a higher cationic charge compared to aurein 1.2. These are well-documented characteristics of highly selective antifungal peptides.[26, 72]

Generally, decreases in hydrophobicity are correlated with decreased mammalian cell toxicity,[73-75] which, in turn, leads to higher selectivity. Supporting these trends, newly discovered high-SI peptides exhibited significantly decreased hydrophobicity in all cases compared to their respective high-SI

templates (i.e., RT of 13-15 min compared to 15-18 min for ααβαααβ backbones). When compared to aurein 1.2, the differences were even greater, with an at least 10.5-minute reduction in RT, illustrating the importance of low hydrophobicity for antifungal selectivity.

Similarly, the helical rigidities of newly discovered high-SI peptides were lower than those of the template peptides and aurein 1.2 (**Table 1**, **Figure 6b**). Quantified as the ratio between the molar ellipticity of peptide (minimum amplitude [θ]) in 15% and in 100% trifluoroethanol (TFE), helical rigidity describes the ability of a peptide to remain helical in aqueous conditions in comparison to its maximum helical conformation, which TFE is known to induce.[76] TFE is often used to mimic the microbial cell membrane environment,[26, 72, 77] while 15% TFE in water has been used to obtain quantifiable differences in low-helicity peptides in aqueous conditions.[9, 19] Similar to hydrophobicity, high helical rigidity results in toxicity against human cells.[19, 26, 78-85] For example, aurein 1.2 presented classic α-helical CD spectra in both aqueous and organic solvent conditions (15% and 100% TFE, **Figure 6b**), with a high helical rigidity of 102 ± 2% (**Table 1**). Meanwhile, both template and newly discovered high-SI peptides exhibited CD curves characteristic of helical α/β-peptides,[9, 13, 17] with lower helicity in aqueous conditions (red line, $[\theta]_{15}$) compared to those in organic conditions (black line, $[\theta]_{100}$) (**Table 1**, **Figure 6b**, **Figure S20**). Given that these highly selective α/β-peptides exhibited similar folding patterns to aurein 1.2, but with significantly reduced helical rigidity, these results support the use of α/β-peptides as a strategy to enhance antimicrobial selectivity of naturally sourced AMPs without compromising their side-chain presentation pattern.

Newly discovered peptides from our approach were also less helical than the template peptides regardless of backbone type. For example, peptide **4-1** (ααβ-motif) exhibited around 31% helical rigidity while ααβ template peptides had 34-60%; similarly, peptides **5-4**, **6-1**, and **6-4 (**ααβαααβ-motif) exhibited helical rigidities between 24-27%, while ααβαααβ template peptides had 37-46%. This decrease in helical rigidity can be attributed largely to the significant reduction in the helicity of peptides in aqueous conditions ($[\theta]_{15}$) rather than changes in helicity in organic conditions ($[\theta]_{100}$) (**Figure 6b**, **Figure S21**). Similar trends have been observed for homogeneous α-amino acid-containing antifungal[26] and antibacterial AMPs,[78-81] where the ability of the peptides to adopt a non-helical-to-helical transition upon exposure to helix-inducing conditions (such as at the cell membrane interface, or in TFE) was a critical contributor to selectivity, predominantly driven by the high correlation between helicity and hemolysis.[82-85] Overall, these results suggest that the iterative GPR approach discovers higher selectivity peptides largely by decreasing their hydrophobicity and increasing their helical flexibility.

α/β-Peptides with High Antifungal Selectivity Induce Significantly Reduced Hemolysis and Kill other Fungal and Bacterial Microbes

Figure 6c summarizes cell viability differences, highlighting how variations in sequence, secondary structure, and other physicochemical properties characterized in **Table 1** influence selectivity. In addition, the peptides were evaluated against other clinical fungal (*Candida glabrata, Candida tropicalis, Candida parapsilosis*) and model bacterial strains (*S. aureus, E. coli*) to assess their potential applications against other microbial species in comparison to aurein 1.2 (**Table S8, S9**). Aurein 1.2 demonstrated activity against *C. albicans* at an MIC of 32 μg/mL (green line), with other microbial MICs ranging from 16 to 256 μg/mL (highlighted region, **Figure 6c**) (**Table S8**, **Figure S22**). However, aurein 1.2 exhibited 37.2 ± 1.8% hemolysis at 64 μg/mL (red line), resulting in a narrow SI of just 1.1 against *C. albicans*, and was unable to kill other microbial cells at the concentrations tested. In comparison, peptides **4-1** (ααβ backbone) and **6-4** (ααβαααβ backbone) exhibited comparable MICs (128 μg/mL and 64 μg/mL against *C. albicans*, respectively, and 16 - 256 μg/mL and 32 - 256 μg/mL broad-spectrum, **Table S8**, **Figure S22**) to aurein 1.2, while only exhibiting hemolytic activities of 10.9 ± 2.2% and 9.4 ± 3.1% at a concentration of 4096 μg/mL, respectively. This represents an almost 128-fold reduction in hemolysis while maintaining similar antifungal and antibacterial activity (highlighted region, **Figure 6c**). Other newly discovered high-SI peptides (**5-4**, **6-1**) demonstrated similar trends (**Figure S18, S22** and **Table S8, S9**). Notably, some peptides also exhibited a significant improvement in antibacterial selectivity, such as peptide **4-1,** with an *E. coli*-specific SI of 219.8 compared to an SI of 1.1 using aurein 1.2, representing a 200-fold increase. These results not only indicate that our model effectively discovers peptides with increased *C. albicans* selectivity compared to human red blood cells (hRBCs), but also demonstrate that these peptides are active against other microbial cells.

Finally, we sought to quantify more precisely the antifungal activities of lead peptides against *C. albicans*, the original target of our model, to better understand the extent of improvement in selectivity achieved using our approach. Here, we note that, in all the determinations of MIC in the studies above, we used two-fold serial dilutions to determine the lowest concentration inhibiting over 90% of microbial growth. While conventional, this method yields less accurate representations of true antimicrobial activity that may lie in between the discrete concentrations that are actually tested and is exacerbated at higher ranges (e.g., two compounds with MICs of 64 versus 128 μg/mL misses more precise activity comparisons in contrast to MICs between 1 μg/mL and 2 μg/mL). Because MIC is used to calculate SI, reported SI values may underestimate the improvement in selectivity resulting from our approach. Therefore, we measured the MICs of aurein 1.2 and high-SI test peptides at smaller intervals of 5 μg/mL (sMIC), as opposed to two-fold serial dilutions (**Figure 6d**, **Figure S23**) and calculated corresponding 'smaller interval SIs' (sSI = HC_{10}/sMIC). As expected, all peptides tested including aurein 1.2 exhibited higher sSI than SI, with **4-1** showing the largest increase. Compared to aurein 1.2, which had an sSI of 1.2, all four newly discovered high-SI peptides demonstrated significantly improved selectivities, ranging from 43.0 to 60.9, indicating up to a 51-fold enhancement in selectivity.

Peptides **4-1** and **6-4**, which contain cationic amino acids on their hydrophobic faces, demonstrated the highest sSI despite

different backbones (ααβ for **4-1**, ααβαααβ for **6-4**). These results further support the idea that the unconventional cationic amino acid substitution in the hydrophobic face of α/β-peptides could be a productive strategy to enhance their antifungal selectivity. While unintuitive from a rational design standpoint, similar substitutions, denoted as "specificity determinants" by Hodges and coworkers,[86-88] have been reported previously to increase the selectivity of AMPs[29] and D-enantiomer analogues[86-88] for antibacterial applications, where (1) increases in charge, (2) decreases in hydrophobicity, and (3) decreases in helical rigidity led to a drastic reduction of hemolysis and, in turn, enhanced antibacterial selectivity.[86-88] These studies mainly attributed the enhanced selectivity of these AMPs to their altered membrane-active action, citing differences between prokaryotic and eukaryotic membranes, such as net charge, in the context of reduced helical rigidity of peptides. We observed in this study that *C. albicans* selectivity was also enhanced through a comparable amino-acid substitution, even though *C. albicans* is eukaryotic. A potential explanation for this is that the higher content of anionic phospholipids in *C. albicans* cell membranes compared to those of red blood cells (around 20% difference)[89, 90] may have similarly contributed to the observed increase in selectivity. However, considering that *C. albicans* cells contain many other components different from human red blood cells, such as the presence of a cell wall[89] and different sterols in cell membranes (ergosterol in fungi vs. cholesterol in hRBCs),[89] further mechanistic studies on interactions between α/β-peptides and cell membranes are needed. Nevertheless, these results suggest that substitutions of cationic amino acids into the more hydrophobic face of α/β-peptides can be a strategy to improve their antifungal selectivity. A potential next step to enhance the antifungal selectivity of highly antifungal but hemolytic α/β-peptides could involve exploring cationic-amino acid substitutions at other positions in their hydrophobic faces, which can be greatly accelerated by iterative GPR.

Conclusions

In this work, we demonstrate the advantages of an iterative approach that combines model prediction with experimental evaluation to discover highly selective antifungal peptides containing noncanonical amino acids. By utilizing a Gaussian process regression model capable of quantifying prediction uncertainty, we explored a vast sequence space of peptidomimetic α/β-peptides (336,000 sequences) based on a relatively small initial training set (147 sequences) to enhance their antifungal selectivity. The uncertainties provided by the model enabled us to make informed decisions on the selection of novel α/β-peptide sequences to synthesize and evaluate experimentally each round. After six prediction rounds, in which 23 new peptide sequences were characterized, we discovered peptides with high selectivity (sSI between 43 to 61), which had up to a 52-fold improvement compared to the wild-type aurein 1.2 (sSI = 1.2). In addition, these peptides also showed antimicrobial activity against other microbial species including bacteria (SI of 220 against *E. coli*), further broadening their potential applications.

Experimental characterization of newly discovered high-SI peptides demonstrated that they exhibit physicochemical properties typical of other highly selective natural AMPs, such as high cationic charge, low hydrophobicity, and low helical rigidity. Furthermore, the most selective α/β-peptides for both α/β-backbones contain unconventional cationic amino acid substitutions near the hydrophobic face, suggesting that incorporating such substitutions could be a potential strategy to enhance antifungal selectivity. This confirms that our approach, based on only sequence and activity data, successfully discovers highly selective peptides that would be far less likely to be identified through conventional rational design. Further understanding the impact of these substitutions on interactions with cell membranes to provide mechanistic insight into peptide activity (e.g., by modelling AMP-induced membrane disruption with computational modelling techniques[91]) will be a subject of future work.

Overall, our approach circumvents the need for large sets of low-throughput and costly experimental physicochemical data while maintaining good prediction accuracy. Additionally, since the SMILES data representation captures the structural complexities of both backbone and side chain elements in a compound-agnostic manner, such a workflow is transferrable to other classes of antimicrobial peptides and mimetics, such as, but not limited to, those of peptides with γ-amino acids[71] or with hydrocarbon-staples.[92] We anticipate that our approach, which enabled the discovery of highly selective α/β-peptides with potential for antifungal therapy, can similarly stimulate the discovery of other highly selective peptidomimetics.

Author contributions

Conceptualization: DHC, JDR, RCV, SPP, DML; computational methodology: JDR; experimental methodology: DHC; sequence selection: JDR, DHC; peptide synthesis: DHC, ML; investigation: JDR, DHC; funding acquisition: DML, SPP, RCV, supervision: DML, SPP, RCV; writing: DHC, JDR, RCV, DML, SPP.

Conflicts of interest

There are no conflicts to declare.

Data availability

A PDF of all computational and experimental methodology for the iterative GPR workflow is included in the ESI. Python scripts to run GPR model training and analysis as well as all peptide sequences, SMILES strings, calculated RDKit descriptors, and antimicrobial and hemolytic activity curve datasheets have been uploaded to a GitHub repository associated with this manuscript (github.com/jdrichardson97/Peptide-GPR) to facilitate reproducibility of the results in this work.

Acknowledgements

This study was supported by National Institutes of Health grant R33AI127442 to SPP and DML and by the National Science

Foundation under award DMR-2044997 to RCV. DHC was supported in part by the UW–Madison NIH Biotechnology Training Program (5T32GM135066-02). The authors thank Prof. Samuel Gellman (UW–Madison) for providing resources for α/β-peptide synthesis, purification, and characterization. The authors gratefully acknowledge the use of facilities and instrumentation supported by the National Science Foundation through the University of Wisconsin Materials Research Science and Engineering Center (DMR-2309000). The purchase of the Bruker Impact II and Thermo Q Exactive Plus, which were used to quantify peptide molecular weight, was funded by the Bender gift to the Department of Chemistry and NIH Award 1S10 OD020022-1 to the Department of Chemistry, respectively.

References

1. M. Muttenthaler, G. F. King, D. J. Adams and P. F. Alewood, *Nature Reviews Drug Discovery*, 2021, **20**, 309-325.
2. H. Jenssen, P. Hamill and R. E. W. Hancock, *Clinical microbiology reviews*, 2006, **19**, 491-511.
3. A. H. Benfield and S. T. Henriques, *Frontiers in Medical Technology*, 2020, **2**.
4. M. Zasloff, *Nature*, 2002, **415**, 389-395.
5. L. Yu, K. Li, J. Zhang, H. Jin, A. Saleem, Q. Song, Q. Jia and P. Li, *ACS Applied Bio Materials*, 2022, **5**, 366-393.
6. T. Rozek, K. L. Wegener, J. H. Bowie, I. N. Olver, J. A. Carver, J. C. Wallace and M. J. Tyler, *European Journal of Biochemistry*, 2000, **267**, 5330-5341.
7. Y. Zhu, C. Shao, G. Li, Z. Lai, P. Tan, Q. Jian, B. Cheng and A. Shan, *Journal of Medicinal Chemistry*, 2020, **63**, 9421-9435.
8. C. G. Starr and W. C. Wimley, *Biochim Biophys Acta Biomembr*, 2017, **1859**, 2319-2326.
9. M. R. Lee, N. Raman, S. H. Gellman, D. M. Lynn and S. P. Palecek, *ACS Chemical Biology*, 2017, **12**, 2975-2980.
10. L. M. Johnson and S. H. Gellman, *Methods in enzymology*, 2013, **523**, 407-429.
11. D. F. Hook, P. Bindschädler, Y. R. Mahajan, R. Šebesta, P. Kast and D. Seebach, *Chemistry and Biodiversity*, 2005, **2**, 591-632.
12. D. L. Steer, R. A. Lew, P. Perlmutter, A. I. Smith and M.-I. Aguilar, *Letters in Peptide Science*, 2001, **8**, 241-246.
13. W. S. Horne and S. H. Gellman, *Accounts of chemical research*, 2008, **41**, 1399-1408.
14. W. S. Horne, L. M. Johnson, T. J. Ketas, P. J. Klasse, M. Lu, J. P. Moore and S. H. Gellman, *Proceedings of the National Academy of Sciences*, 2009, **106**, 14751-14756.
15. J. W. Checco, E. F. Lee, M. Evangelista, N. J. Sleebs, K. Rogers, A. Pettikiriarachchi, N. J. Kershaw, G. A. Eddinger, D. G. Belair, J. L. Wilson, C. H. Eller, R. T. Raines, W. L. Murphy, B. J. Smith, S. H. Gellman and W. D. Fairlie, *Journal of the American Chemical Society*, 2015, **137**, 11365-11375.
16. J. L. Price, W. S. Horne and S. H. Gellman, *Journal of the American Chemical Society*, 2010, **132**, 12378-12387.
17. W. S. Horne, J. L. Price and S. H. Gellman, *Proceedings of the National Academy of Sciences*, 2008, **105**, 9151-9156.
18. D. H. Appella, L. A. Christianson, I. L. Karle, D. R. Powell and S. H. Gellman, *Journal of the American Chemical Society*, 1996, **118**, 13071-13072.
19. D. H. Chang, M.-R. Lee, N. Wang, D. M. Lynn and S. P. Palecek, *ACS Infectious Diseases*, 2023, **9**, 2632-2651.
20. S. Haider, C. Rotstein, D. Horn, M. Laverdiere and N. Azie, *Canadian Journal of Infectious Diseases and Medical Microbiology*, 2014, **25**, 308169.
21. P. G. Pappas, M. S. Lionakis, M. C. Arendrup, L. Ostrosky-Zeichner and B. J. Kullberg, *Nature Reviews Disease Primers*, 2018, **4**.
22. R. E. Lewis, *Mayo Clin Proc*, 2011, **86**, 805-817.
23. L. E. Cowen, J. B. Anderson and L. M. Kohn, *Annu Rev Microbiol*, 2002, **56**, 139-165.
24. L. E. Cowen, *FEMS Microbiol Lett*, 2001, **204**, 1-7.
25. R. N. Murugan, B. Jacob, M. Ahn, E. Hwang, H. Sohn, H.-N. Park, E. Lee, J.-H. Seo, C. Cheong, K.-Y. Nam, J.-K. Hyun, K.-W. Jeong, Y. Kim, S. Y. Shin and J. K. Bang, *PLOS ONE*, 2013, **8**, e80025.
26. Y. Lyu, Y. Yang, X. Lyu, N. Dong and A. Shan, *Scientific reports*, 2016, **6**, 27258.
27. Z. Jiang, B. J. Kullberg, H. Van Der Lee, A. I. Vasil, J. D. Hale, C. T. Mant, R. E. W. Hancock, M. L. Vasil, M. G. Netea and R. S. Hodges, *Chemical Biology & Drug Design*, 2008, **72**, 483-495.
28. R. F. Epand, M. A. Schmitt, S. H. Gellman and R. M. Epand, *Biochim Biophys Acta*, 2006, **1758**, 1343-1350.
29. A. Hawrani, R. A. Howe, T. R. Walsh and C. E. Dempsey, *J Biol Chem*, 2008, **283**, 18636-18645.
30. M. Kodedová, M. Valachovič, Z. Csáky and H. Sychrová, *Cell Microbiol*, 2019, **21**, e13093.
31. S. Bobone and L. Stella, *Adv Exp Med Biol*, 2019, **1117**, 175-214.
32. G. Wang, X. Li and Z. Wang, *Nucleic acids research*, 2016, **44**, D1087-D1093.
33. F. H. Waghu, R. S. Barai, P. Gurung and S. Idicula-Thomas, *Nucleic Acids Res*, 2016, **44**, D1094-1097.
34. G. Shi, X. Kang, F. Dong, Y. Liu, N. Zhu, Y. Hu, H. Xu, X. Lao and H. Zheng, *Nucleic Acids Res*, 2022, **50**, D488-D496.
35. G. Cordoves-Delgado and C. R. García-Jacas, *J Chem Inf Model*, 2024, **64**, 4310-4321.
36. D. Veltri, U. Kamath and A. Shehu, *Bioinformatics*, 2018, **34**, 2740-2747.
37. W. Dee, *Bioinform Adv*, 2022, **2**, vbac021.
38. S. N. Dean and S. A. Walper, *ACS Omega*, 2020, **5**, 20746-20754.
39. H. ElAbd, Y. Bromberg, A. Hoarfrost, T. Lenz, A. Franke and M. Wendorff, *BMC Bioinformatics*, 2020, **21**, 235.
40. S. Lata, B. Sharma and G. P. Raghava, *BMC bioinformatics*, 2007, **8**, 1-10.
41. M. Krynski and M. Rossi, *npj Computational Materials*, 2021, **7**, 169.
42. P. A. Romero, A. Krause and F. H. Arnold, *Proc Natl Acad Sci U S A*, 2013, **110**, E193-201.
43. J. Parkinson and W. Wang, *Journal of Chemical Information and Modeling*, 2023, **63**, 4589-4601.
44. V. L. Deringer, A. P. Bartók, N. Bernstein, D. M. Wilkins, M. Ceriotti and G. Csányi, *Chemical Reviews*, 2021, **121**, 10073-10141.

45. T. Vornholt, M. Mutný, G. W. Schmidt, C. Schellhaas, R. Tachibana, S. Panke, T. R. Ward, A. Krause and M. Jeschek, *ACS Central Science*, 2024, **10**, 1357–1370.
46. W. S. Horne, M. D. Boersma, M. A. Windsor and S. H. Gellman, *Angew Chem Int Ed Engl*, 2008, **47**, 2853-2856.
47. J. L. Price, W. S. Horne and S. H. Gellman, *J Am Chem Soc*, 2010, **132**, 12378-12387.
48. E. A. Porter, X. Wang, H.-S. Lee, B. Weisblum and S. H. Gellman, *Nature*, 2000, **404**, 565-565.
49. M. Szefczyk, K. Ożga, M. Drewniak-Świtalska, E. Rudzińska-Szostak, R. Hołubowicz, A. Ożyhar and Ł. Berlicki, *RSC Advances*, 2022, **12**, 4640-4647.
50. D. Migoń, M. Jaśkiewicz, D. Neubauer, M. Bauer, E. Sikorska, E. Kamysz and W. Kamysz, *Probiotics Antimicrob Proteins*, 2019, **11**, 1042-1054.
51. S. Soufian and L. Hassani, *Pak J Biol Sci*, 2011, **14**, 729-735.
52. D. S. Radchenko, S. Kattge, S. Kara, A. S. Ulrich and S. Afonin, *Biochimica et Biophysica Acta (BBA) - Biomembranes*, 2016, **1858**, 2019-2027.
53. M. Rajabi, B. Ericksen, X. Wu, E. de Leeuw, L. Zhao, M. Pazgier and W. Lu, *Journal of Biological Chemistry*, 2012, **287**, 21615-21627.
54. G. von Heijne, *Journal of Molecular Biology*, 1991, **218**, 499-503.
55. J. Y. Suh, Y. T. Lee, C. B. Park, K. H. Lee, S. C. Kim and B. S. Choi, *Eur J Biochem*, 1999, **266**, 665-674.
56. R. d. O. Dias and O. L. Franco, *Peptides*, 2015, **72**, 64-72.
57. M. A. Hitchner, L. E. Santiago-Ortiz, M. R. Necelis, D. J. Shirley, T. J. Palmer, K. E. Tarnawsky, T. D. Vaden and G. A. Caputo, *Biochimica et Biophysica Acta (BBA) - Biomembranes*, 2019, **1861**, 182984.
58. A. S. Kelkar, B. C. Dallin and R. C. Van Lehn, *The Journal of Chemical Physics*, 2022, **156**, 024701.
59. Y. Khalak, G. Tresadern, D. F. Hahn, B. L. de Groot and V. Gapsys, *Journal of Chemical Theory and Computation*, 2022, **18**, 6259-6270.
60. D. Weininger, *Journal of Chemical Information and Computer Sciences*, 1988, **28**, 31-36.
61. *RDKit: Open-source cheminformatics. https://www.rdkit.org*
62. K. Nesmerak, A. Toropov and I. Yildiz, *Frontiers in Bioscience-Landmark*, 2022, **27**.
63. M. Yasir, A. M. Karim, S. K. Malik, A. A. Bajaffer and E. I. Azhar, *Saudi Journal of Biological Sciences*, 2022, **29**, 3687-3693.
64. A. Szabóová, O. Kuželka, F. Železný and J. Tolar, *BMC Bioinformatics,* 2012, **13**, S3.
65. N. Raman, M. R. Lee, D. M. Lynn and S. P. Palecek, *Pharmaceuticals*, 2015, **8**, 483-503.
66. M.-R. Lee, N. Raman, S. H. Gellman, D. M. Lynn and S. P. Palecek, *ACS Chemical Biology*, 2014, **9**, 1613-1621.
67. P. Király, R. Kiss, D. Kovács, A. Ballaj and G. Tóth, *Molecular Informatics*, 2022, **41**, 2200072.
68. C. Rücker, G. Rücker and M. Meringer, *Journal of Chemical Information and Modeling*, 2007, **47**, 2345-2357.
69. K. Lin, A. C. W. May and W. R. Taylor, *Journal of Theoretical Biology*, 2002, **216**, 361-365.
70. Y. Akkam, *Jordan Journal of Pharmaceutical Sciences*, 2016, **9**.
71. Y.-H. Shin and S. H. Gellman, *Journal of the American Chemical Society*, 2018, **140**, 1394-1400.
72. J. Wang, S. Chou, Z. Yang, Y. Yang, Z. Wang, J. Song, X. Dou and A. Shan, *J Med Chem*, 2018, **61**, 3889-3907.
73. S. E. Blondelle and R. A. Houghten, *Biochemistry*, 1991, **30**, 4671-4678.
74. L. H. Kondejewski, M. Jelokhani-Niaraki, S. W. Farmer, B. Lix, C. M. Kay, B. D. Sykes, R. E. Hancock and R. S. Hodges, *Journal of Biological Chemistry*, 1999, **274**, 13181-13192.
75. T. Tachi, R. F. Epand, R. M. Epand and K. Matsuzaki, *Biochemistry*, 2002, **41**, 10723-10731.
76. M. K. Luidens, J. Figge, K. Breese and S. Vajda, *Biopolymers*, 1996, **39**, 367-376.
77. Z. Lai, X. Yuan, W. Chen, H. Chen, B. Li, Z. Bi, Y. Lyu and A. Shan, *J Med Chem*, 2024, **67**, 10891-10905.
78. A. J. Beevers and A. M. Dixon, *Chemical Society Reviews*, 2010, **39**, 2146-2157.
79. L. Chen, X. Li, L. Gao and W. Fang, *The Journal of Physical Chemistry B*, 2015, **119**, 850-860.
80. B. Deslouches, S. M. Phadke, V. Lazarevic, M. Cascio, K. Islam, R. C. Montelaro and T. A. Mietzner, *Antimicrobial agents and chemotherapy*, 2005, **49**, 316-322.
81. A. A. Strömstedt, M. Pasupuleti, A. Schmidtchen and M. Malmsten, *Antimicrobial agents and chemotherapy*, 2008, **53**, 593-602.
82. Y. Shai and Z. Oren, *J Biol Chem*, 1996, **271**, 7305-7308.
83. Y. Shai and Z. Oren, *Peptides*, 2001, **22**, 1629-1641.
84. Y. Huang, L. He, G. Li, N. Zhai, H. Jiang and Y. Chen, *Protein Cell*, 2014, **5**, 631-642.
85. M. A. Cherry, S. K. Higgins, H. Melroy, H. S. Lee and A. Pokorny, *J Phys Chem B*, 2014, **118**, 12462-12470.
86. Z. Jiang, C. T. Mant, M. Vasil and R. S. Hodges, *Chem Biol Drug Des*, 2018, **91**, 75-92.
87. Z. Jiang, A. I. Vasil, M. L. Vasil and R. S. Hodges, *Pharmaceuticals (Basel)*, 2014, **7**, 366-391.
88. Z. Jiang, A. I. Vasil, L. Gera, M. L. Vasil and R. S. Hodges, *Chem Biol Drug Des*, 2011, **77**, 225-240.
89. M. Rautenbach, A. M. Troskie and J. A. Vosloo, *Biochimie*, 2016, **130**, 132-145.
90. A. Singh, T. Prasad, K. Kapoor, A. Mandal, M. Roth, R. Welti and R. Prasad, *OMICS: A Journal of Integrative Biology*, 2010, **14**, 665-677.
91. J. D. Richardson and R. C. Van Lehn, *The Journal of Physical Chemistry B*, 2024, **128**, 8737-8752.
92. Y. You, H. Liu, Y. Zhu and H. Zheng, *Amino Acids*, 2023, **55**, 421-442.