Supplementary Information for

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]

Table of Contents

Materials and general considerations

Safety. No unexpected or unusually high safety hazards were encountered.

Materials. Fmoc-β-amino acids, including Fmoc-L-β-homoalanine, Fmoc-L-β-homoisoleucine, Fmoc-L-β-homoleucine, Fmoc-L-β-homophenylalanine, Fmoc-(1S,2S)-2-aminocyclopentane carboxylic acid, Nβ-Fmoc-Nω-Boc-L-β-homolysine, Fmoc-O-tert-butyl-L-β-homoserine, and Fmoc-α-amino acids, including Fmoc-glycine, Fmoc-L-alanine, Fmoc-L-isoleucine, Fmoc-L-leucine, Fmoc-L-phenylalanine, Fmoc-O-tert-butyl-L-serine, Fmoc-L-β-homotryptophan, Fmoc-L-aspartic acid β-tert-butyl ester, Fmoc-L-glutamic acid γ-tert-butyl ester, Nα-Fmoc-Nε-Boc-L-lysine were purchased from Chem-Impex International, Inc. (Wood Dale, IL, USA). Fmoc-L-norleucine was purchased from Thermo Scientific Chemicals. Fmoc-L-norvaline was purchased from Santa Cruz Biotechnology. HATU was obtained from Oakwood Chemicals. Tentagel S RAM Fmoc was purchased from Advanced ChemTech (Louisville, KY). Menadione, N,N-Diisopropylethylamine, Mueller Hinton Broth, and dibasic sodium phosphate were obtained from Sigma-Aldrich (St. Louis, MO). 3-(N-Morpholino) propanesulfonic acid (MOPS) was obtained from Fisher Scientific (Pittsburgh, PA). 2,3-Bis(2-methoxy-4-nitro-5-sulfophenyl)-2H-tetrazolium-5-carboxanilide (XTT) was purchased from Invitrogen. Gibco brand RPMI 1640 powder (containing phenol red and L-glutamine and without sodium bicarbonate or HEPES) and Dulbecco's phosphate-buffered saline (DPBS, without calcium or magnesium) were obtained from Thermo Fisher Scientific (Waltham, MA). Water (18.2 MΩ) was purified using a Millipore filtration system. Cell Titer Glo 2.0 assay kits were obtained from Promega (Madison, WI).

General considerations. *C. albicans* strain SC5314, *E. coli* strain 25922, and *S. aureus* strain 3359 were purchased from ATCC. *C. glabrata* 5376, *C. parapsilosis* 5986, and *C. tropicalis* 98-234 are clinical isolates from invasive candidiasis and were generously donated by Dr. David Andes (University of Wisconsin-Madison). For hemolysis experiments, freshly expired human red blood cells were obtained from the University of Wisconsin−Madison Hospital blood bank. All microbial strains were stored as a 50% glycerol stock at −80 °C and grown in a liquid YPD (for fungal cells) or TSB (for bacterial cells) medium. Peptide sequence, activity, SMILES string, and RDKit descriptor data are available online; see '**Data availability' section**.

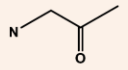
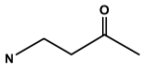
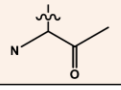
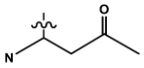
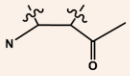
Computational Prediction using Gaussian Process Regression

S1: Conversion of Peptide Sequences to SMILES Strings

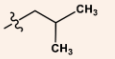
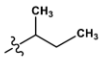
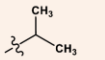
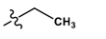
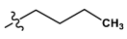
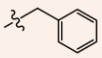
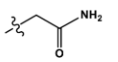
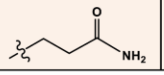
This section details the creation of simplified molecular-input line-entry system (SMILES) strings based on peptide sequences in our workflow. **Figure S1** shows the building blocks for creating a SMILES string for a peptide with β amino acids, which was decomposed into backbone (**Figure S1a**) and side chain (**Figure S1b**) elements. Additionally, all sequences considered in this study contained a protonated N-terminus ([NH3+]) and amidated C-terminus (C(=O)N), as is typical for α/β-peptides studied experimentally.[1,2]

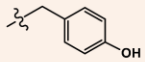
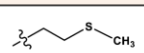
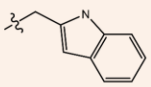
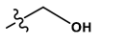
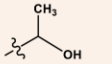
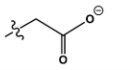
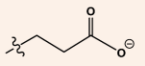
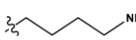
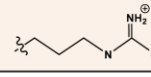
For backbone SMILES segments (**Figure S1a**), the zig-zag line in the Sequence column corresponds to the attachment of a side chain, which involves a corresponding @ symbol as reflected in the SMILES segment which denotes chirality. For protein sequences, @@ refers to L-amino acids while @ refers to D-amino acids; however, the specification of chirality has no effect on calculated 2D RDKit descriptors, so we kept all chiral centers as @ for consistency and ease of visualization. As depicted in **Figure S1c**, the backbone SMILES segments (**Figure S1a**) were first added in the same order as the peptide sequence, and then side chains (**Figure S1b**) were added in reverse order for all non-glycine amino acids.

a

Backbone	Sequence	SMI
N-terminus (protonated)		[NH3+]
α backbone (Glycine)		NCC(=O)
β backbone (Glycine)		NCCC(=O)
α backbone (not Glycine)		N[C@H](C(=O)
β backbone (not Glycine)		N[C@H](CC(=O)
ACPC		N[C@H]#[C@H](C(=O)

b

Sidechain	Sequence	SMI
Ala (A)		C
Leu (L)		CC(C)C
Ile (I)		[C@H](CC)C
Val (V)		C(C)C
Abu		CC
Nva		CCC
Nle		CCCC
Phe (F)		Cc9ccccc9
Asn (N)		CC(=O)N
Gln (Q)		CCC(=O)N

Sidechain	Sequence	SMI
ACPC*		CCC#
Tyr (Y)		Cc8ccc(cc8)O
Met (M)		CCSC
Trp (W)		Cc8c[nH]c9c8cccc9
Ser (S)		CO
Thr (T)		[C@H](C)O
Asp (D)		CC(=O)[O-]
Glu (E)		CCC(=O)[O-]
Lys (K)		CCCC[NH3+]
Arg (R)		CCCNC(=[NH2+])N

c

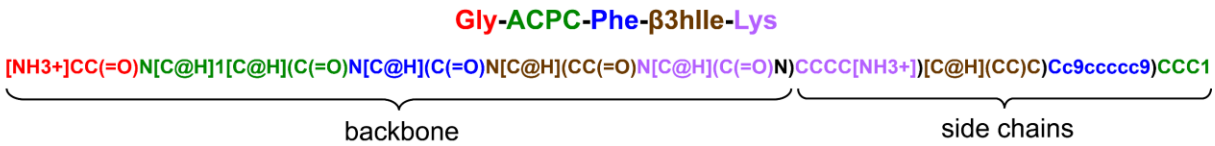


Figure S1. SMILES string creation process involving (a) backbone fragments and (b) side chain fragments. (c) Visualization of addition process for backbone and sidechain fragments for a model 5-amino acid sequence color coded by residue. The highlighted # in (a) and (b) is a counter for the number of ACPC amino acids in the sequence (e.g., a sequence with 3 amino acids would have 1, 2, 3 in place of this #) along the SMILES string for both the (a) backbone and (b) side chain fragment.

S2: Descriptor Preprocessing

For each iterative GPR round, we first calculated all 200 molecular descriptors available with the RDKit Cheminformatics toolkit for all training peptide sequences with the following methodology:

1. Each peptide SMILES string (**Section S1**) was constructed as a molecule interpretable by RDKit ('Mol' object) with the MolFromSmiles() function.
2. The full set of 200 descriptor labels was called as rdkit.Chem.Descriptors._descList
3. The numerical values of these descriptors were calculated for each peptide Mol object using the MoleculeDescriptors() module.

As described in the main text, the first step of descriptor preprocessing was to remove all descriptors that had constant numerical values for all sequences in the training set which largely involved counts for chemical groups not pertinent to peptides (e.g., ketones, halogens, etc.). **Figure S2** shows the number of nonconstant descriptors kept for each prediction round (starting with 105 for Round 1 and ending with 120 for Round 6), highlighting in red new nonconstant descriptors introduced with novel amino acids in Rounds 3 and 4.

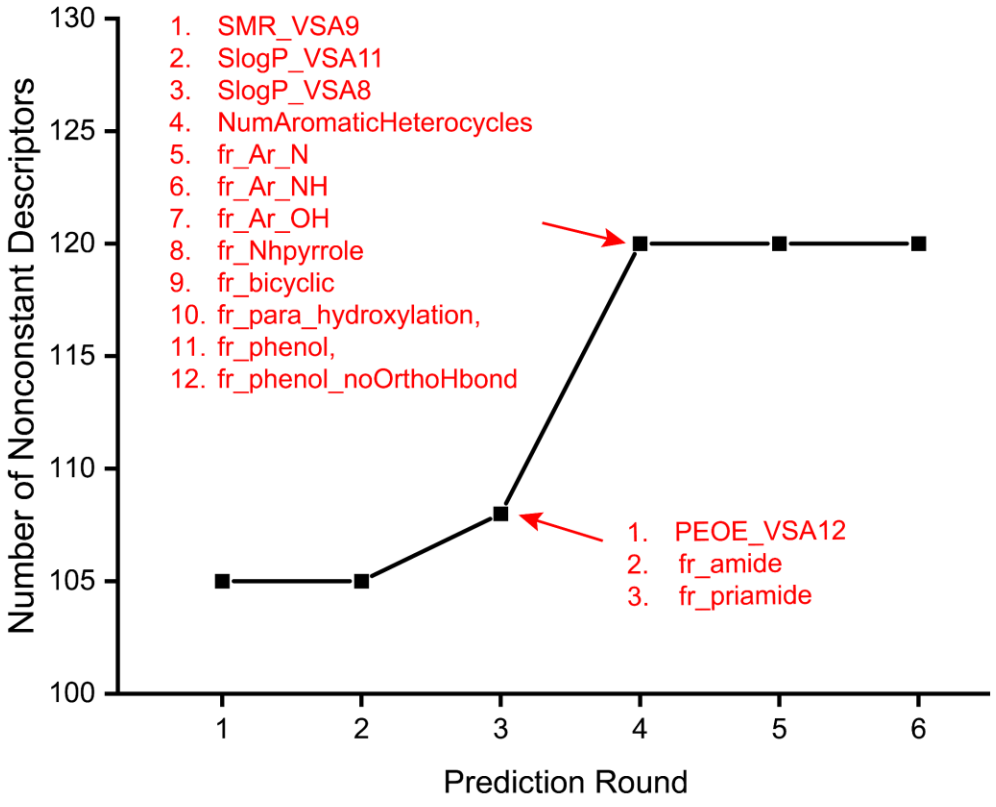


Figure S2. Number of nonconstant descriptors from the original set of 200 RDKit descriptors for each prediction round. Red annotations refer to nonconstant descriptors added due to new peptide sequences that were incorporated in the training set after each prediction round.

The second step of descriptor preprocessing involved further reduction of this nonconstant descriptor space with 10-fold LASSO cross-validation using the LassoCV module available in *sklearn*. The LASSO regression model is based on the linear regression model with an additional L1 regularization term with coefficient α in the cost function that forces small descriptor coefficients (descriptors unimportant for model predictions) to zero as shown in Equation S1:

$$Cost = \sum_{i=1}^{N}\left(y_i - \sum_{j=1}^{M} x_{ij}W_j\right)^2 + \alpha\sum_{j=1}^{M}|W_j| \quad \text{(S1)}$$

y_i is the $\log_2$ scaled experimentally measured HC_{10} or MIC value of the i^{th} peptide in the training set, x_{ij} is the value of the j^{th} descriptor for the i^{th} peptide, W_j is the coefficient of the j^{th} descriptor for the regression model, and α is the regularization parameter. Generally, an increase in α leads to a decrease in the number of descriptors kept by the LASSO model, and the goal is to find the α value that minimizes the Cost function. LASSO cross-validation (LASSO CV) involves splitting up the training data into *n* folds, training a LASSO model on *n – 1 folds*, validating the model on the remaining fold, and then repeating this process until all *n* folds are used as the validation set (see **Figure S5a** for visualization of this process for 10-fold CV). This process is conducted for a range of α values, selecting the α that minimizes the average root mean squared error (RMSE) of the LASSO model across the *n* folds.

As described in **Section S4**, all labels (*i.e.*, experimentally measured $Log_2(HC_{10})$ or $Log_2(MIC)$ values) were proportionately allocated across 10 folds to ensure equal distributions of label values (low to high) per fold. LASSO regression models were trained to predict these labels using 100 different values of α (Log_{10} scaled between 10^{-3} and 1). The model (*i.e.*, value of α) that minimized the average RMSE between experimental and predicted values across all ten folds was selected, and the corresponding set of descriptors associated with that model was used as the reduced descriptor set. **Figure S3** shows plots of the RMSE *vs.* α across the folds for different prediction rounds. The minimum of this black line denotes the minimum of the average RMSE which is marked with a vertical dashed line and the corresponding number of descriptors that are kept with this α value.

The LASSO CV procedure was applied to the initial 147-sequence training set to select 20 descriptors for $Log_2(HC_{10})$ and 13 descriptors for $Log_2(MIC)$ predictions for prediction Rounds 1 to 3; the set of descriptors was unchanged for these three rounds. RMSE *vs.* α plots are shown in **Figure S3a**. For Round 4 onwards, the descriptors were updated each round with the updated methodology described in the **'Iterative GPR Model Training' Section** in the main text to more accurately predict $Log_2(HC_{10})$ and $Log_2(MIC)$ labels for newly introduced peptide sequences. The corresponding RMSE *vs.* α plots are shown in **Figures S3b-d**.

Table S1 shows the reduced set of descriptors used for $Log_2(HC_{10})$ predictions for each prediction round and **Table S2** shows the reduced set of descriptors used for $log_2(MIC)$ predictions for each prediction round. In these tables, the 'Initial' column refers to the descriptors used for Rounds 1 to 3 whereas Rounds 4 to 6 are labeled accordingly. Red numbers are provided in the 'Initial' column to denote the absolute descriptor weights provided by the LASSO model as a signifier of decreasing descriptor importance. These numbers are also cross-referenced in columns for Rounds 4 to 6 to visualize how reimplementing the LASSO CV procedure for later rounds reorders these descriptor importances, where descriptors are also listed in decreasing importance for each round.

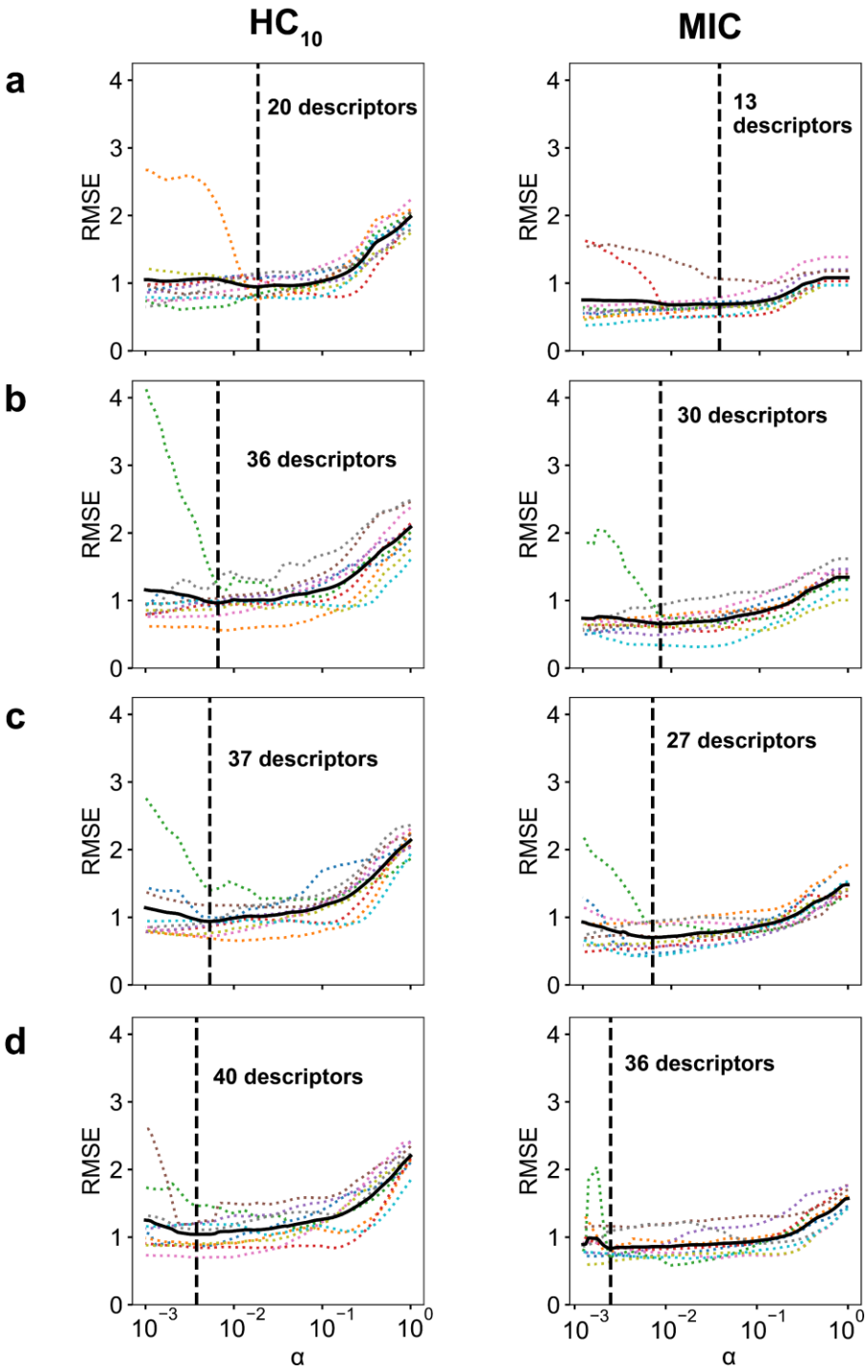


Figure S3. LASSO 10-fold cross validation curves for (a) Round 1, (b) Round 4, (c) Round 5, and (d) Round 6 for both HC_{10} (left column) and MIC (right column). Each dotted line represents the calculated RMSE for one of the ten folds across 100 different α values for the LASSO model, and the solid bold line represents the average of the 10-folds. The dashed vertical line represents the α value that minimizes the average RMSE with the corresponding number of descriptors labeled in bold.

To check if our descriptor selection approach mitigates model overfitting (by minimizing RMSE variance across LASSO CV folds for selected α values) as new data are added across prediction

rounds, **Figure S4** shows the variance of the RMSE vs. α distributions plotted in **Figure S3** across the 10 CV folds. Although the HC_{10} RMSE variance for the selected value of α increases somewhat per round, the MIC RMSE variance is similar for both Rounds 1 and 6 despite an increase in the number of kept descriptors (**Table S2**). Additionally, for both metrics the RMSE variance is relatively steady or increases as α increases compared to the selected value of α (dashed line) per round. We would expect that the variance would sharply decrease with fewer descriptors (larger value of α) if the model were overfit for the selected value of α; indeed, we see that in all cases very low values of α do correspond to a large variance increase, which would be due to overfitting with large numbers of descriptors. The plateau of the variance for the selected values of α instead indicates that our descriptor selection approach effectively minimizes overfitting per round.

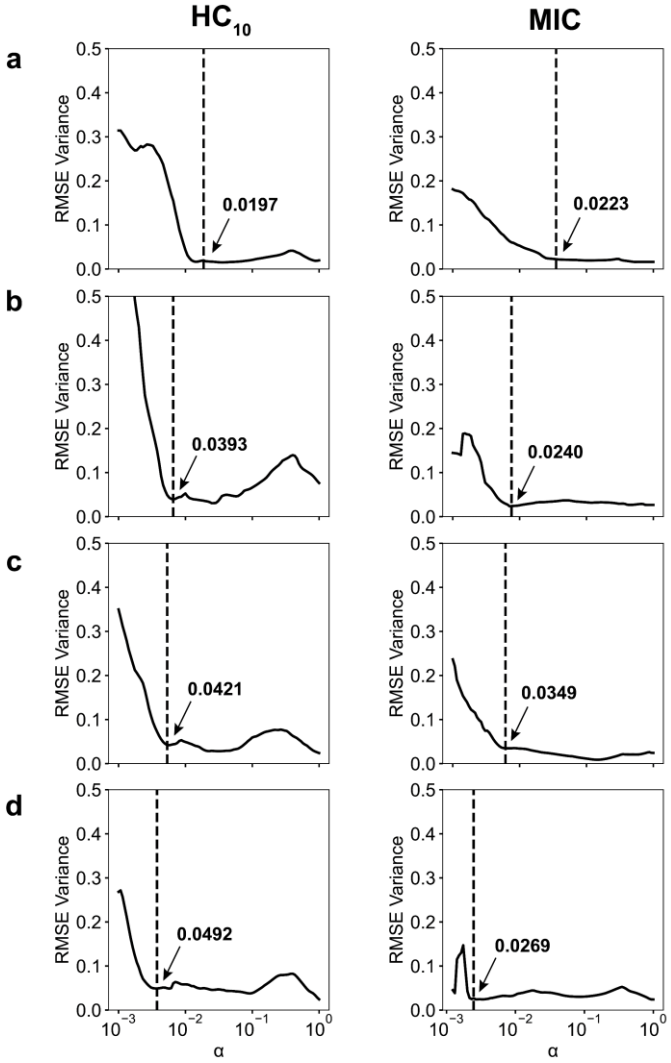


Figure S4: RMSE variance across the 10-folds used for LASSO CV (dotted lines in **Figure S3**) vs. α for (a) Round 1, (b) Round 4, (c) Round 5, and (d) Round 6 for both HC_{10} (left column) and MIC (right column). The dashed vertical line represents the selected α per round (from **Figure S3**) with the corresponding RMSE variance at this α value captioned in bold.

Table S1. RDKit descriptors used for each round of Gaussian process regression model training for $Log_2(HC_{10})$ predictions listed in decreasing descriptor importance. Descriptors were chosen based on a 10-fold LASSO CV approach after removing all constant descriptors (see **Figure S3**). 'Initial' column refers to descriptors chosen based on the initial 147 training sequences and utilized for Rounds 1-3 of experimental evaluation. The remaining columns (Round 4, Round 5, Round 6) refer to their corresponding prediction rounds. Red numbers indicate the rank ordering of descriptors based on coefficient values in relation to the 'Initial' column to visualize how rank orderings change with LASSO cross validation descriptor updates. A value of '---' means there is no reference descriptor to compare to in the original 20-descriptor set.

Initial		Round 4		Round 5		Round 6	
Idx	Desc	Ref Idx	Desc	Ref Idx	Desc	Ref Idx	Desc
1	PEOE_VSA8	---	Chi4n	2	PEOE_VSA6	---	Chi4n
2	PEOE_VSA6	2	PEOE_VSA6	---	Chi4n	2	PEOE_VSA6
3	VSA_EState2	1	PEOE_VSA8	1	PEOE_VSA8	15	Chi3n
4	EState_VSA2	3	VSA_EState2	3	VSA_EState2	1	PEOE_VSA8
5	PEOE_VSA7	---	VSA_EState4	15	Chi3n	---	qed
6	fr_unbrch_alkane	15	Chi3n	---	VSA_EState4	8	EState_VSA5
7	EState_VSA6	8	EState_VSA5	8	EState_VSA5	3	VSA_EState2
8	EState_VSA5	10	SMR_VSA6	4	EState_VSA2	---	VSA_EState4
9	FpDensityMorgan1	4	EState_VSA2	---	qed	---	MaxPartialCharge
10	SMR_VSA6	18	EState_VSA1	10	SMR_VSA6	10	SMR_VSA6
11	EState_VSA4	9	FpDensityMorgan1	---	BalabanJ	18	EState_VSA1
12	SlogP_VSA4	---	PEOE_VSA11	---	PEOE_VSA11	---	BalabanJ
13	MaxEStateIndex	---	BalabanJ	9	FpDensityMorgan1	19	MinEStateIndex
14	MinPartialCharge	---	qed	18	EState_VSA1	---	PEOE_VSA11
15	Chi3n	19	MinEStateIndex	---	MaxPartialCharge	13	MaxEStateIndex
16	Chi2n	---	MaxPartialCharge	19	MinEStateIndex	---	VSA_EState6
17	NumAromaticCarbocycles	---	FpDensityMorgan3	14	MinPartialCharge	---	VSA_EState5
18	EState_VSA1	14	MinPartialCharge	---	SlogP_VSA8	4	EState_VSA2
19	MinEStateIndex	---	VSA_EState7	11	EState_VSA4	---	PEOE_VSA9
20	MaxAbsEStateIndex	---	SlogP_VSA8	---	FpDensityMorgan3	---	SlogP_VSA8
		11	EState_VSA4	13	MaxEStateIndex	14	MinPartialCharge
		---	SlogP_VSA4	---	PEOE_VSA10	11	EState_VSA4
		13	MaxEStateIndex	---	SMR_VSA9	---	PEOE_VSA10
		---	PEOE_VSA12	20	MaxAbsEStateIndex	---	MinAbsEStateIndex
		7	EState_VSA6	7	EState_VSA6	20	MaxAbsEStateIndex
		20	MaxAbsEStateIndex	---	SlogP_VSA4	9	FpDensityMorgan1
		---	SMR_VSA9	---	PEOE_VSA12	---	SlogP_VSA4
		---	MinAbsEStateIndex	---	VSA_EState5	---	NumAromaticHeterocycles
		---	fr_Ar_OH	---	MinAbsEStateIndex	---	fr_Ar_OH
		---	NumAromaticHeterocycles	---	NumAromaticHeterocycles	---	PEOE_VSA12
		---	SlogP_VSA11	---	SlogP_VSA11	5	PEOE_VSA7
		---	fr_amide	---	fr_Ar_N	---	FpDensityMorgan3
		---	fr_Ar_N	---	fr_Ar_OH	7	EState_VSA6
		---	fr_Nhpyrrole	---	fr_amide	---	SMR_VSA9
		---	fr_phenol	---	MinAbsPartialCharge	---	SlogP_VSA11
		---	MinAbsPartialCharge	---	fr_phenol	---	fr_phenol
				---	fr_Nhpyrrole	---	fr_Nhpyrrole
						---	fr_Ar_N
						---	MinAbsPartialCharge
						---	fr_bicyclic

Table S2. RDKit descriptors used for each round of Gaussian process regression model training for Log_2(MIC) predictions listed in decreasing descriptor importance. Descriptors were chosen based on a 10-fold LASSO CV approach after removing all constant descriptors (see **Figure S3**). 'Initial' column refers to descriptors chosen based on the initial 147 training sequences and utilized for Rounds 1-3 of experimental evaluation. The remaining columns (Round 4, Round 5, Round 6) refer to their corresponding prediction rounds. Red numbers indicate the rank ordering of descriptors based on coefficient values in relation to the 'Initial' column to visualize how rank orderings change with LASSO cross validation descriptor updates. A value of '---' means there is no reference descriptor to compare to in the original 13-descriptor set.

Initial		Round 4		Round 5		Round 6	
Idx	Desc	Ref Idx	Desc	Ref Idx	Desc	Ref Idx	Desc
1	PEOE_VSA8	1	PEOE_VSA8	10	FpDensityMorgan1	12	Chi4n
2	EState_VSA7	10	FpDensityMorgan1	1	PEOE_VSA8	1	PEOE_VSA8
3	PEOE_VSA6	7	VSA_EState7	12	Chi4n	---	Chi3n
4	EState_VSA2	12	Chi4n	---	VSA_EState4	5	VSA_EState2
5	VSA_EState2	---	FpDensityMorgan3	3	PEOE_VSA6	---	VSA_EState4
6	PEOE_VSA11	---	SMR_VSA6	---	FpDensityMorgan2	8	PEOE_VSA7
7	VSA_EState7	---	EState_VSA1	---	BalabanJ	---	qed
8	PEOE_VSA7	---	SlogP_VSA4	---	EState_VSA3	---	Kappa3
9	MaxEStateIndex	5	VSA_EState2	---	EState_VSA5	---	BalabanJ
10	FpDensityMorgan1	3	PEOE_VSA6	---	EState_VSA1	---	PEOE_VSA9
11	VSA_EState6	---	EState_VSA5	---	SMR_VSA6	---	EState_VSA1
12	Chi4n	13	EState_VSA6	---	PEOE_VSA9	---	fr_C_O
13	EState_VSA6	---	MinEStateIndex	---	qed	3	PEOE_VSA6
		---	EState_VSA9	---	NumAromaticRings	---	fr_Al_COO
		---	NumAromaticRings	---	SlogP_VSA4	---	MinEStateIndex
		---	PEOE_VSA12	2	EState_VSA7	---	SlogP_VSA5
		---	SlogP_VSA8	5	VSA_EState2	---	EState_VSA5
		---	PEOE_VSA10	---	FpDensityMorgan3	---	VSA_EState5
		---	EState_VSA4	---	SlogP_VSA8	10	FpDensityMorgan1
		---	fr_NH2	---	MinEStateIndex	---	MaxPartialCharge
		---	NumAromaticHeterocycles	---	fr_amide	11	VSA_EState6
		---	qed	---	MinAbsEStateIndex	---	FpDensityMorgan3
		4	EState_VSA2	7	VSA_EState7	---	PEOE_VSA10
		---	VSA_EState4	---	SlogP_VSA11	---	PEOE_VSA1
		---	VSA_EState5	---	PEOE_VSA12	---	fr_quatN
		---	fr_amide	---	NumAromaticHeterocycles	---	EState_VSA4
		---	fr_Ar_OH	---	SMR_VSA9	---	MinAbsEStateIndex
		---	SlogP_VSA11			9	MaxEStateIndex
		---	fr_phenol			---	SMR_VSA6
		---	SMR_VSA9			---	MaxAbsEStateIndex
						---	SlogP_VSA8
						13	EState_VSA6
						---	MinPartialCharge
						---	NumAromaticHeterocycles
						---	MinAbsPartialCharge
						---	fr_Nhpyrrole

Tables S1-S2 show that there are many similar 'high importance' descriptors kept after LASSO CV between the $Log_2(HC_{10})$ and Log_2(MIC) prediction workflows, indicating that similar physiochemical properties are important for predicting both red blood cell hemolysis and antifungal activity. For descriptors with an underscore separating two physiochemical properties with a terminal number (e.g., PEOE_VSA6), the atomic contributions of both properties are calculated in a molecule, atoms are assigned to bins that indicate ranges of the first property (chosen bin for descriptor denoted by terminal number), and then atoms that fall into the bin range selected for the first property are summed up to calculate the contributions of the second property. For instance, in calculating the PEOE_VSA6 descriptor, the van der Waals (VSA) surface area of

all atoms that have a partial charge (PEOE) between -0.10 and -0.05 (range 6) are summed up. Definitions of all 200 2D RDKit descriptors considered in this study and corresponding references can be found in the RDKit documentation[1], but in general, important molecular properties for both the $Log_2(HC_{10})$ and $Log_2(MIC)$ LASSO CV descriptor selection (**Tables S1-S2**) workflows include:

1. **PEOE:** Atomic partial charge calculated with the 'Partial Equalization of Orbital Electronegativities' method[2]
2. **VSA:** van der Waals surface area
3. **EState:** 'Electrotopological-State' index that encodes electronic and topological for each atom to depict their accessibility to interact with neighboring atoms through the calculation of contributing electrons and hydrogen atoms[3]
4. **Chi:** Captures the complexity of molecular connectivity in a molecule (e.g., branching, rings, etc.)[4]
5. **FpDensityMorgan:** Captures chemical and connectivity attributes of atoms based on the definition of 'similarity fingerprints' [5]
6. **SMR:** molar refractivity
7. **SlogP:** octanol-water partition coefficient (logP value)
8. **qed:** 'quantitative estimation of drug-likeness' [6]

Additionally, there was an increase in the number of descriptors required to adequately describe the training data for newly introduced amino acids. While the $Log_2(HC_{10})$ prediction workflow maintained many of the same descriptors with the 10-highest weights even with the new descriptor update procedure implemented for Round 4 onwards, there was a large redistribution and introduction of new descriptors for the MIC workflow, indicating that the initial 13-descriptor set was not sufficient as the training data increased in size. Additionally, many new descriptors introduced in Rounds 3 and 4 (**Figure S2**) were utilized in Rounds 5 and 6 (e.g., fr_amide, NumAromaticHeterocycles), albeit not with large weights (i.e., they were of low importance to model predictions).

S3: Gaussian Process Regression Model Selection Criteria

This section details the GPR model selection procedure for each prediction round (in support of the right plot in **Figure 2c**), which was implemented independently for both $Log_2(HC_{10})$ and $Log_2(MIC)$ predictions. We utilized the GaussianProcessRegressor() module in *sklearn* for all model training and predictions. For each round of experimental evaluation, the kernel and corresponding hyperparameters were optimized through a 10-fold grid search cross validation (CV) approach to maximize the coefficient of determination (R^2) when comparing predicted and experimentally measured $Log_2(HC_{10})$ and $Log_2(MIC)$ values across 10 proportionately allocated folds (see **Section S4**). The 4 kernels considered with corresponding hyperparameters are shown in **Table S3**:

Table S3. Overview of GPR kernels and hyperparameters considered for grid search cross validation. For the Radial Basis Function, Rational, Quadratic, and Exponentiated Sine Squared kernels, $d(x_i,x_j)$ refers to the Euclidean distance between two feature vectors in the training set x_i and x_j. For the Dot Product kernel, $x_i \cdot x_j$ denotes the dot product between two feature vectors in the training set.

Kernel	Equation	Hyperparameters
Radial Basis Function (RBF)	$\exp\left(-\frac{d(x_i,x_j)^2}{2l^2}\right)$	l = length scale
Rational Quadratic	$\left(1+\frac{d(x_i,x_j)^2}{2\alpha l^2}\right)^{-\alpha}$	l = length scale α = alpha
Exponentiated Sine Squared	$\exp\left(-\frac{2sin^2(\pi d(x_i,x_j)/p)}{l^2}\right)$	l = length scale p = periodicity
Dot Product	${\sigma_0}^2 + x_i \cdot x_j$	σ_0 = inhomogeneity

We considered 5 values (0.01, 0.1, 1, 10, 100) for all hyperparameters (length scale, alpha, periodicity, inhomogeneity) during the grid search process. An additional hyperparameter κ, which is the value added to the kernel diagonal to prevent numerical issues during fitting, varied among the values [1×10^{-20}, 1×10^{-15}, 1×10^{-10}, 1×10^{-5}, 0.01, 0.1, 1.0, 5.0, 10.0]. Therefore, a total of 540 combinations of kernels and hyperparameters were considered per round of model training, and the model with the highest average R^2 across the 10 folds was chosen for test sequence predictions. **Table S4** shows the combination of κ and kernel with associated hyperparameters that was chosen for each prediction round.

Table S4. Gaussian kernel and hyperparameter set selected per round for (a) $Log_2(HC_{10})$ and (b) $Log_2(MIC)$ predictions. κ refers to the value added to the kernel diagonal during fitting to prevent numerical issues. Selected kernels listed with corresponding hyperparameters in parenthesis, which are defined in **Table S3**. The 'R^2' column denotes the 10-fold cross validation accuracy of the GPR model with the associated kernel and hyperparameters per round and is identical to the R^2 values tabulated in **Table S6** and plotted in **Figure 4a**.

a

Round	κ	Kernel	R^2
1	**0.1**	Exponentiated Sine Squared (l= **0.1**, p = **100**)	0.851
2	**0.1**	Exponentiated Sine Squared (l= **0.01**, p = **100**)	0.863
3	**0.1**	Exponentiated Sine Squared (l= **0.1**, p = **100**)	0.847
4	**1**	Dot Product (σ_0 = **1**)	0.866
5	**1**	Dot Product (σ_0 = **1**)	0.880
6	**1**	Dot Product (σ_0 = **0.01**)	0.865

b

Round	κ	Kernel	R^2
1	**0.1**	Rational Quadratic (l = **100**, α = **0.1**)	0.655
2	**0.1**	Exponentiated Sine Squared (l= **0.1**, p = **100**)	0.791
3	**0.1**	Rational Quadratic (l = **100**, α = **100**)	0.697
4	**1**	Dot Product (σ_0 = **1**)	0.785
5	**1**	Dot Product (σ_0 = **1**)	0.805
6	**0.1**	Dot Product (σ_0 = **0.1**)	0.743

S4: HC_{10} and MIC Label Preprocessing

For both 10-fold cross-validation (CV) steps in the GPR workflow discussed in the main text - descriptor reduction with LASSO CV (see **Section S2**), and GPR model and hyperparameter selection with Gridsearch CV (see **Section S3**) - we implemented a 'proportionate allocation' procedure on labels (*i.e.*, $Log_2(HC_{10})$ and $Log_2(MIC)$ values) in each round. The goal of proportionate allocation is to ensure equal distributions of label numerical values (i.e., low to high) in each fold used during 10-fold cross validation for all model training steps in the workflow. Proportionate allocation of both HC_{10} and MIC labels was implemented with an in-house python function that performed the following steps, which were also shown schematically in **Figure S5**:

1. The target label (either HC_{10} or MIC) is appended as an additional column to a pandas DataFrame of all RDKit descriptors.
2. This DataFrame is sorted from low to high values of the label column using the sort_values() method in pandas (*sorted_df*). **Figure S5c** shows the increasing label magnitude for this sorted DataFrame.
3. The first 'n' rows of this sorted DataFrame (*sorted_df*) divisible by 10 (e.g., for initial 147 training this would be the first 140 labels) are distributed to a new empty DataFrame (*new_df*) to proportionately allocate them across 10 folds. For instance, for 140 sequences, rows 1-10 in *sorted_df* (first 10 entries in **Figure S5c**) become rows 1, 15, 29, … , 113, 127 in *new_df*; rows 11-20 in sorted_df become rows 2, 16, 30, …, 114, 128 in *new_df*, etc.
4. The remaining 'm' rows with the highest label values not included in step 3 (e.g., last 7 rows of *sorted_df* for initial 147 training) are added to the first m folds. **Figure S5d** demonstrates the final proportionately allocated 10 folds of HC_{10} labels of the initial 147 training peptides.

Overall, we found this proportionate allocation approach leads to consistently higher prediction accuracies on test folds during cross-validation (**Figure S14**), particularly given the introduction of a large number of sequences with large values of HC_{10} and MIC during the 6 prediction rounds (**Figure S11-S12**).

Additionally, given the serially diluted nature of experimental assays to measure both HC_{10} labels for human red blood cells and MIC labels for *Candida albicans*[1,2], where concentrations of peptide in µg/mL are doubled between measurements, we log_2 scaled all HC_{10} and MIC labels before any regression steps (e.*g.*, LASSO CV, GPR). **Figure S6** shows the distributions of both HC_{10} and MIC labels both before (**Figure S6a**) and after (**Figure S6b**) implementing this log_2 scaling, which was critical in preventing large outliers in peptide concentration for model training.

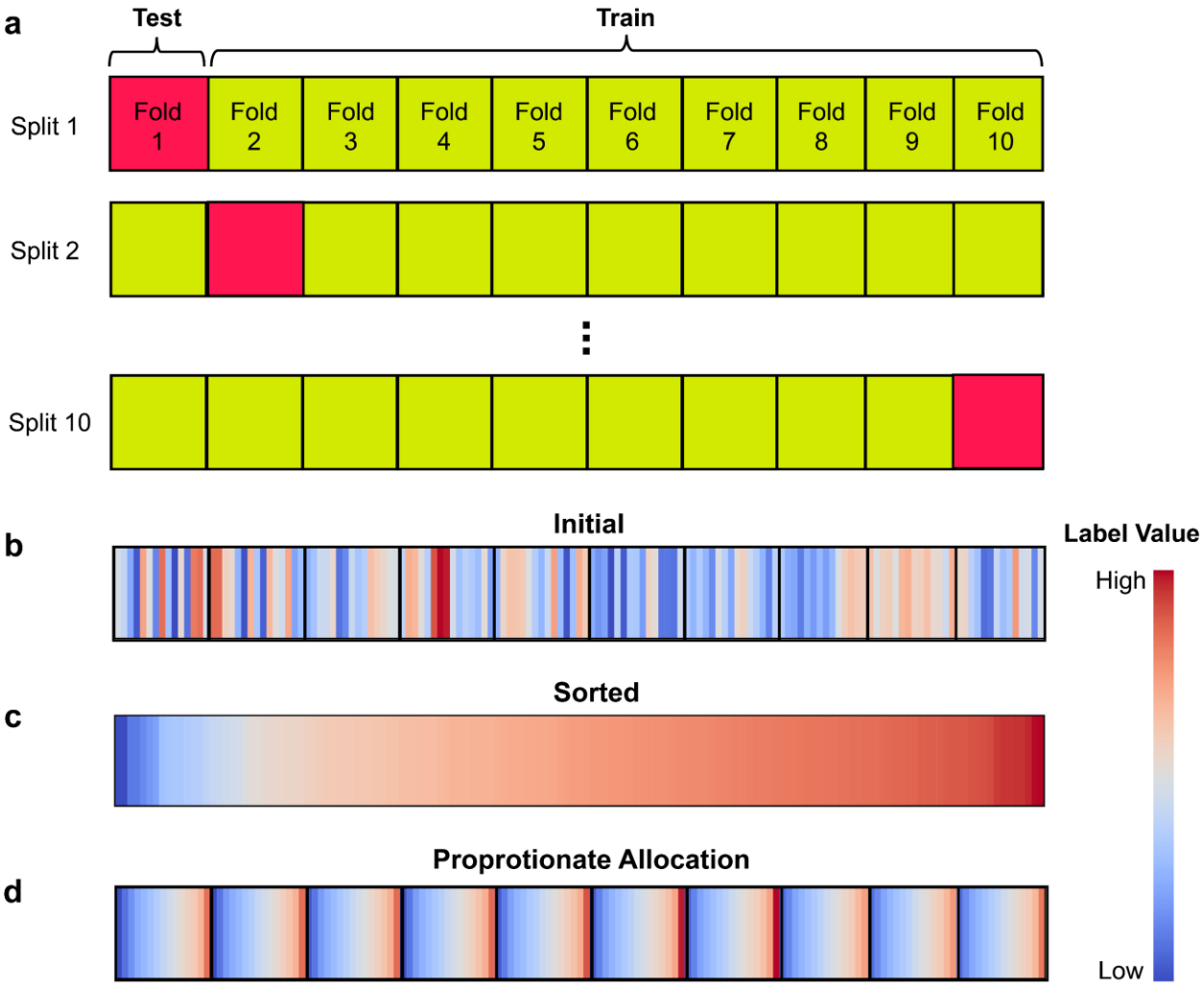


Figure S5. Demonstrating the proportionate allocation procedure for all cross-validation (CV) steps in the GPR workflow. (a) General schematic for 10-fold cross validation where training data (yellow) were used to fit a regression model that was then utilized to predict labels for a held-out test set (red). This procedure was repeated over 10 folds (*i.e.*, unique splits of the data), where training metrics (e.g., R^2, RMSE, etc.) were then calculated as an average across the 10 folds. (b-d) Steps involved in the implementation of the proportionate allocation procedure: (b) initial label values of all training peptides (blue = low value and red = high value), (c) sorting all labels in increasing order, (d) proportionately allocating sorted labels across the 10 folds for cross validation as demonstrated in (a).

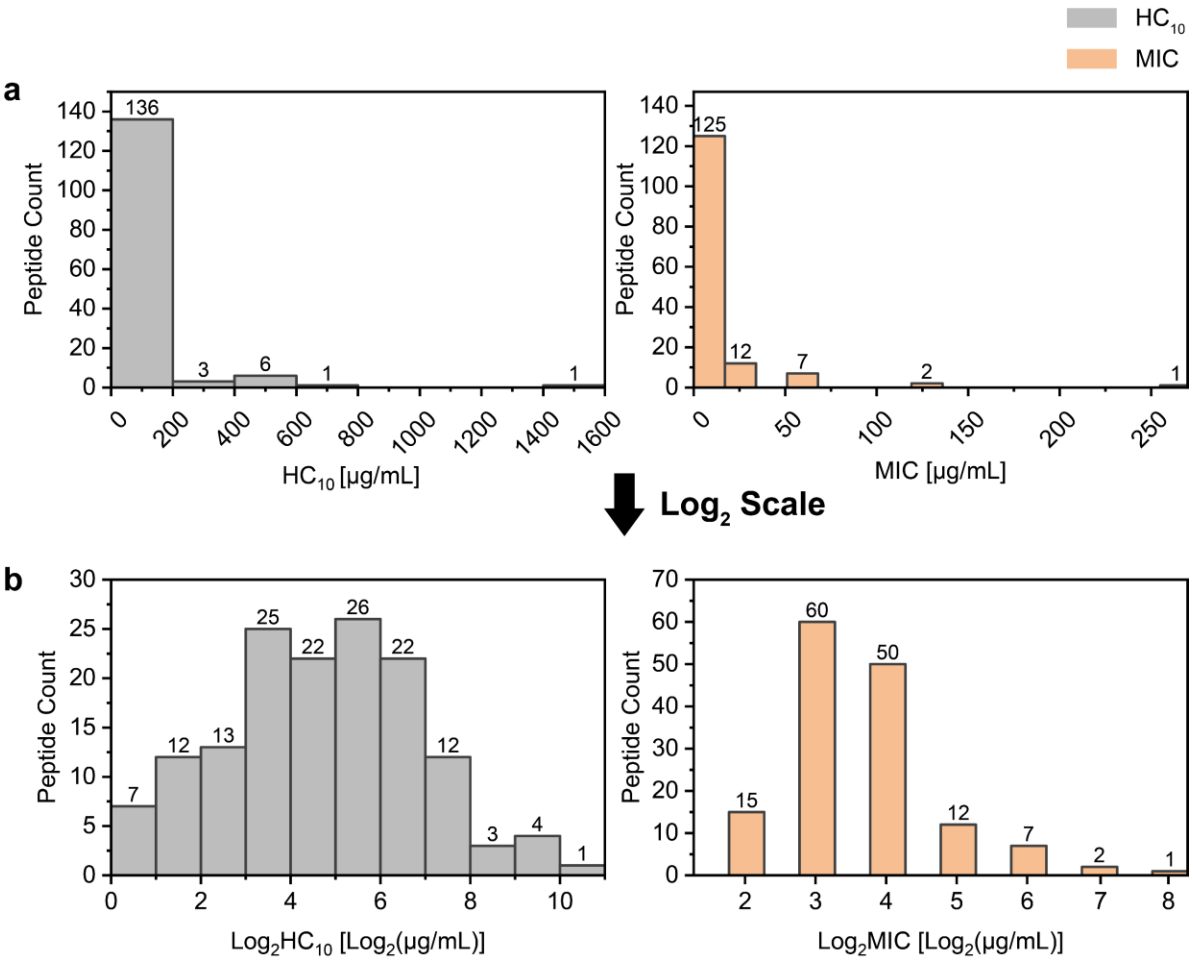


Figure S6. HC_{10} and MIC label distributions for the initial 147-peptide training set based on total peptide counts both (a) before and (b) after Log_2 scaling.

S5: Test Design Space Generation and Truncation

This section details the generation and reduction of the test sequence design space considered for each prediction round, starting from an initial set of 168,000 test sequences templated on either the ααβ backbone for prediction rounds 1 to 4 (**Figure 3c**) or αααβαααβ backbone for rounds 5 to 6 (**Figure 3d**). As described in the **'Iterative GPR Workflow Implementation' Section** in the main text, each 168,000-test sequence space introduced a large number of low and high descriptor values compared to values of descriptors computed for the training set (here, we refer only to values of descriptors from the reduced set obtained after the LASSO CV procedure detailed in **Section S2**). Therefore, to prevent potential prediction errors due to large extrapolation of descriptors values compared to values for the training set, we reduced the 168,000-sequence design spaces to include only test sequences for which each descriptor value fell within the minimum and maximum bounds of the values for that same descriptor in the training set (noting that the training set is updated before each prediction round with experimental values from the preceding round).

Figure S7 summarizes the range of test sequence descriptor values for the 168,000-test sequence space generated from the ααβ template for the 10 highest-weight descriptors kept after LASSO CV for prediction Round 1. **Figure S8** shows similar information for the 168,000-test sequence space generated from the ααβαααβ template for the 10 highest-weight descriptors kept after LASSO CV for prediction Round 5. Values are normalized such that the minimum and maximum values of each descriptor computed from the training set are equal to -1 and 1, respectively. For both backbone templates, there are descriptors with values over 6x higher than values in the training set (e.g., EState_VSA5 (**Figures S7b**, **S8b**, **S8c**) and SMR_VSA6 (**Figures S7b**, **S8b**)), highlighting the importance of reducing the set of test sequences to avoid significant extrapolation.

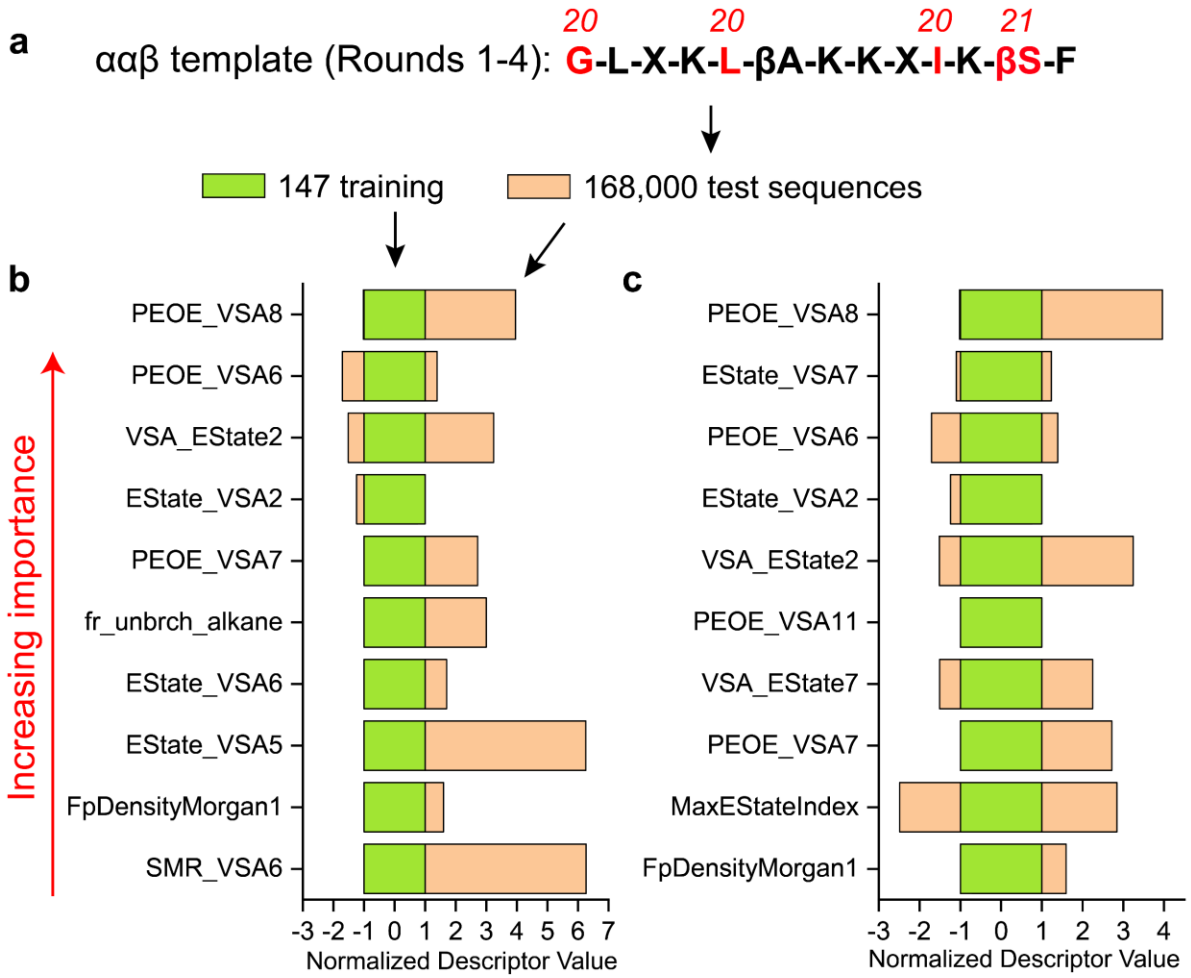


Figure S7. Ranges of descriptor values computed for the initial 147 training sequences (green) vs. the 168,000 test sequences generated using the ααβ template (tan). (a) Highest-SI ααβ template sequence shown with red amino acids where substitutions were made. Ranges of descriptor values for the 10 highest-weight descriptors (based on absolute value of coefficient) kept after LASSO cross validation (see Initial column in **Tables S1**-**S2**) are shown for both (b) $\log_2(HC_{10})$ and (c) $\log_2(MIC)$. Descriptor values are normalized such that such that the minimum and maximum values of each descriptor computed from the training set are equal to -1 and 1, respectively.

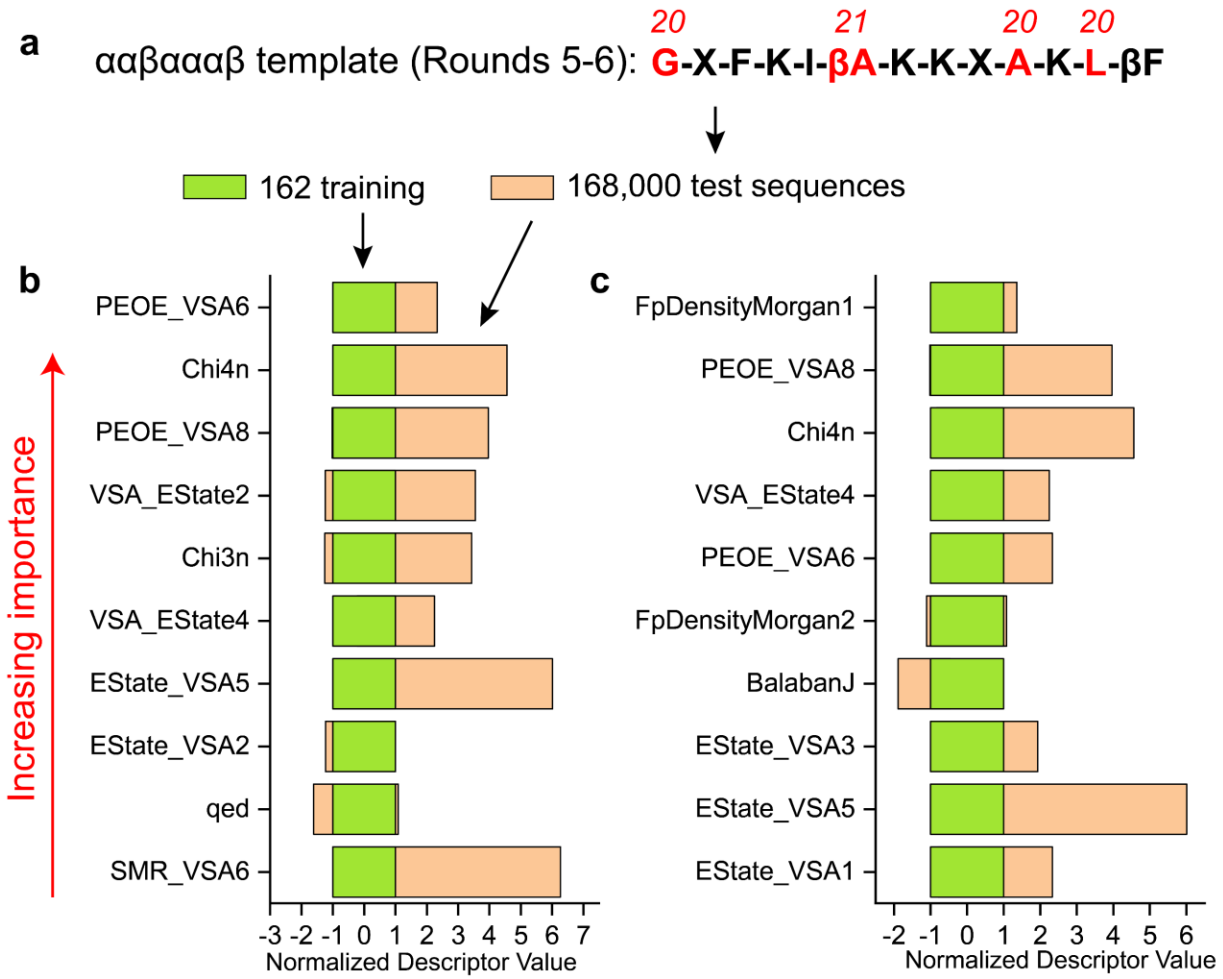


Figure S8. Ranges of descriptor values computed for the 162 training sequences after prediction Round 4 (green) vs. the 168,000 test sequences generated using the αααβαααβ template (tan). (a) Highest-SI αααβαααβ template sequence shown with red amino acids where substitutions were made. Ranges of descriptor values for the 10 highest-weight descriptors (based on absolute value of coefficient) kept after LASSO cross validation (see Round 5 column in **Tables S1-S2**) shown for both (b) $\log_2(HC_{10})$ and (c) $\log_2(MIC)$. Descriptor values are normalized such that such that the minimum and maximum values of each descriptor computed from the training set are equal to -1 and 1, respectively.

The number of test sequences of the full 168,000-sequence design space for which all values of the reduced set of descriptors were within the training descriptor bounds (*i.e.*, all descriptor values were within the green bars in **Figures S7-S8**) are tabulated in **Table S5** for both the $\text{Log}_2(HC_{10})$ and $\text{Log}_2(MIC)$ prediction workflows; the intersection of these sequences is in the 'Total' column and represent the number of sequences that could be potentially selected for experimental synthesis during the GPR workflow. As discussed in the **'Iterative GPR Model Training' Section** in the main text, Rounds 1-3 all had the same 14,137-test sequence space since the descriptors for $\log_2(HC_{10})$ and $\log_2(MIC)$ predictions remained constant across these rounds ('Initial' column in **Tables S1-S2**).

Table S5: Number of test sequences per round (out of 168,000) for which all values of the reduced set of descriptors were within the training descriptor bounds. Columns labeled "$Log_2(HC_{10})$", "$Log_2(MIC)$", "Total" indicate all test sequences within these bounds for the $Log_2(HC_{10})$ descriptors (**Table S1**), $Log_2(MIC)$ descriptors (**Table S2**), and the intersection of these sequences, respectively.

Round Number	$Log_2(HC_{10})$	$Log_2(MIC)$	Total
1	15712	21989	**14137**
2	15712	21989	**14137**
3	15712	21989	**14137**
4	17418	24280	**17238**
5	11347	15058	**10716**
6	10111	19057	**8764**

S6: Training Set Backbones and Templating Considerations

To further support the selection of the high SI template peptides to generate the test sequence design space (**Figure 3c-d**), **Figure S9** shows the $log_2(HC_{10})$ *vs*. $log_2(MIC)$ distribution for training peptides for the αααβ and αβαβααβ backbones that were present in the training data but not used as templates. As demonstrated by these plots, these backbones were less practical for templating to discover new high SI sequences compared to the ααβ and ααβαααβ backbones.

For the αααβ backbone (**Figure S9a**), there were 5 positions to consider (2 α and 3 β) amongst the three highest-SI peptides which would result in an initial design space of 20^2 x 21^3 or approximately 3,700,000 test sequences, over 20 times greater the 168,000-test design space considered for each backbone template in the main text. Additionally, these highest-SI αααβ sequences (#059, #135, #018) had 3.7<SI<7.4, which was a much lower range of SI values than the ααβ backbone template (11.6<SI<24.1) without providing any advantage in increasing antifungal activity.

For the αβαβααβ backbone (**Figure S9b**), the three potential template sequences (#053, #024, #023) similarly had a low SI (4<SI<7.7), and two of these sequences had relatively low antifungal activity against *C. albicans* ($log_2(MIC) > 5$). Additionally, only three positions varied amongst these three sequences (1 α and 2 β) resulting in a test design space of only 8820 sequences. Therefore, the αβαβααβ backbone was also less promising for discovering new high SI sequences with our iterative GPR methodology.

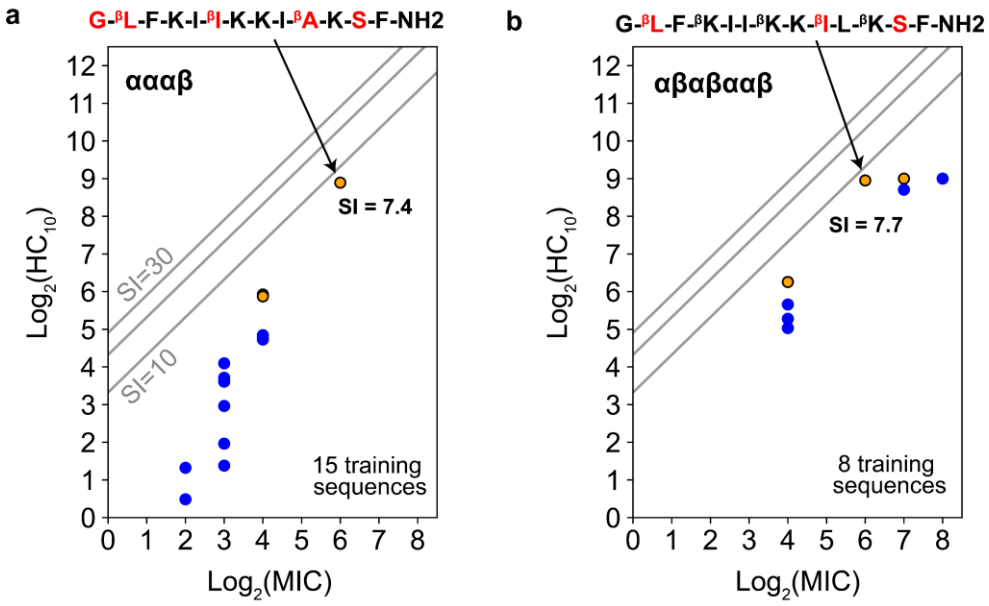


Figure S9. $Log_2(HC_{10})$ vs. $Log_2(MIC)$ distributions for the (a) αααβ and (b) αβαβααβ backbone in the initial 147-training sequence space. Analogous to **Figure 3c-d**, orange dots represent the three highest-SI peptides for each backbone with residue positions that vary amongst these three peptides highlighted in the sequence at the top. Remaining sequences are shown as blue dots. SI bands with values of 10, 20, and 30 are shown as grey lines.

Figure S10 shows the experimental Log_2HC_{10} vs. Log_2MIC distribution of all 147 AMPs used in the initial training set in this study (combining information from **Figure 3c-d** and **Figure S9**). Points are colored coded by each of the 4 backbone types – ααβ = blue, αααβ = orange, ααβααβ = green, αβαβααβ = red. In general, AMPs across the different backbone types follow similar Log_2HC_{10} value ranges at each Log_2MIC value where large amounts of experimental data are available ($2 < Log_2MIC < 5$), suggesting that there is not a direct structure-activity relationship for predicting SI based on backbone type alone. Additionally, this plot highlights that there is a significant increase in AMP hemolytic potential relative to antimicrobial activity (large decrease in SI) as AMPs reach low Log_2MIC values (e.g., $Log_2MIC = 2$). Few AMPs display SI > 10 (ααβααβ at $Log_2MIC = 4$, ααβ at $Log_2MIC = 5$-6), and this supports the selection of initial high-SI templates across these 2 backbone types for mid-range Log_2MIC experimental values in our study to further increase AMP selectivity over iterative GPR rounds (more discussion in the **'Properties of Training Sequences and Design Space Generation' Section** in the main text).

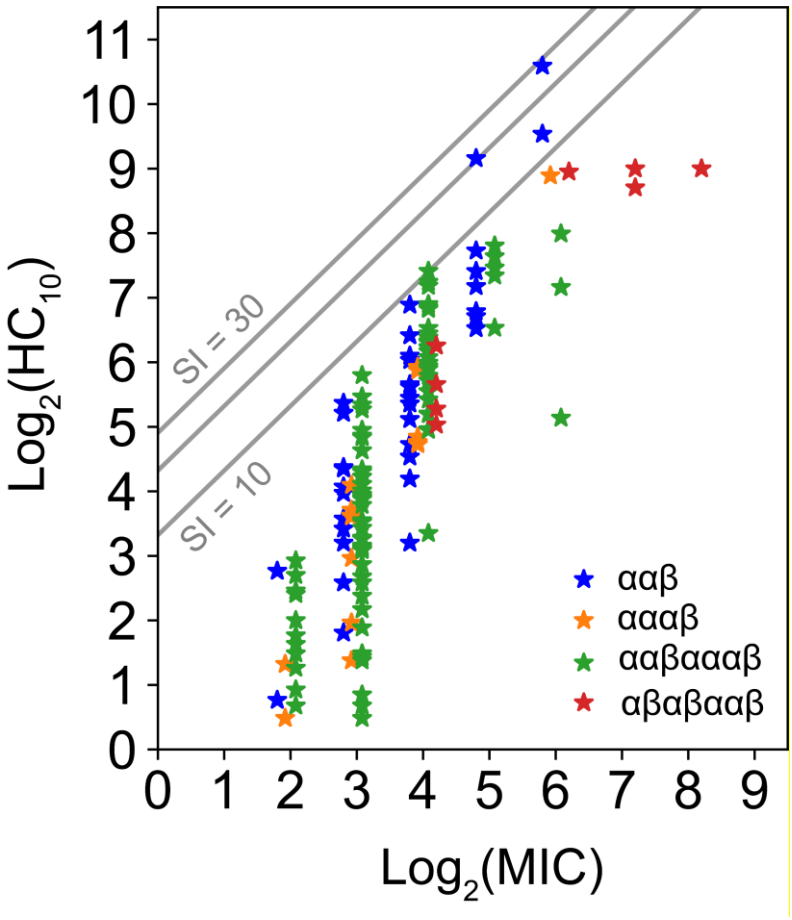


Figure S10. $Log_2(HC_{10})$ vs. $Log_2(MIC)$ distribution of all 147 initial training set peptides used in this study colored coded by backbone type – αaβ = blue, αaaβ = orange, αaβaaaβ = green, αβαβaaβ = red. Grey lines indicate constant SI values equal to 10, 20, and 30. All points are plotted for constant $Log_2(MIC)$ values in the range [2,8] and are slightly offset around each $Log_2(MIC)$ value for visualization purposes.

S7: Model Prediction Metrics and Robustness Checks

Parity plots for all prediction rounds are shown in **Figure S11** for $log_2(HC_{10})$ predictions and **Figure S12** for $log_2(MIC)$ predictions to support the evolution of the GPR model accuracy (R^2) in **Figure 4a** in the main text. Matching the visualization in **Figures 4b-c**, the initial 147 training sequences are shown as open blue circles and newly discovered sequences are shown as red triangles, and the R^2 per round is labelled in the bottom right corner of each plot. These R^2 values are the same as the black lines in **Figure 4a**. All points are plotted as the test set prediction from 10-fold cross validation on proportionately allocated labels, and therefore the blue circle vs. red triangle distinction is for visualization only to demonstrate the addition of peptides with high $log_2(HC_{10})$ and $log_2(MIC)$ values to the training set as iteration proceeds.

Additionally, the large underprediction of test sequence **2-4** is shown as a circled red triangle in the Round 3 plot for both the $log_2(HC_{10})$ and $log_2(MIC)$ workflows in **Figures S11** and **S12** respectively, which leads to the large increase in Maximum Error in the main text (red lines in

Figure 4a). This observation motivated the choice to update the set of descriptors with LASSO CV each round starting with Round 4 (as described in **Section S2**) which leads to better predictions for sequences with large $\log_2(HC_{10})$ and $\log_2(MIC)$ values (and therefore lower maximum errors), as supported by the reorganization of the most important descriptors kept with the LASSO model that better describe new amino acids and motifs (**Tables S1-S2**).

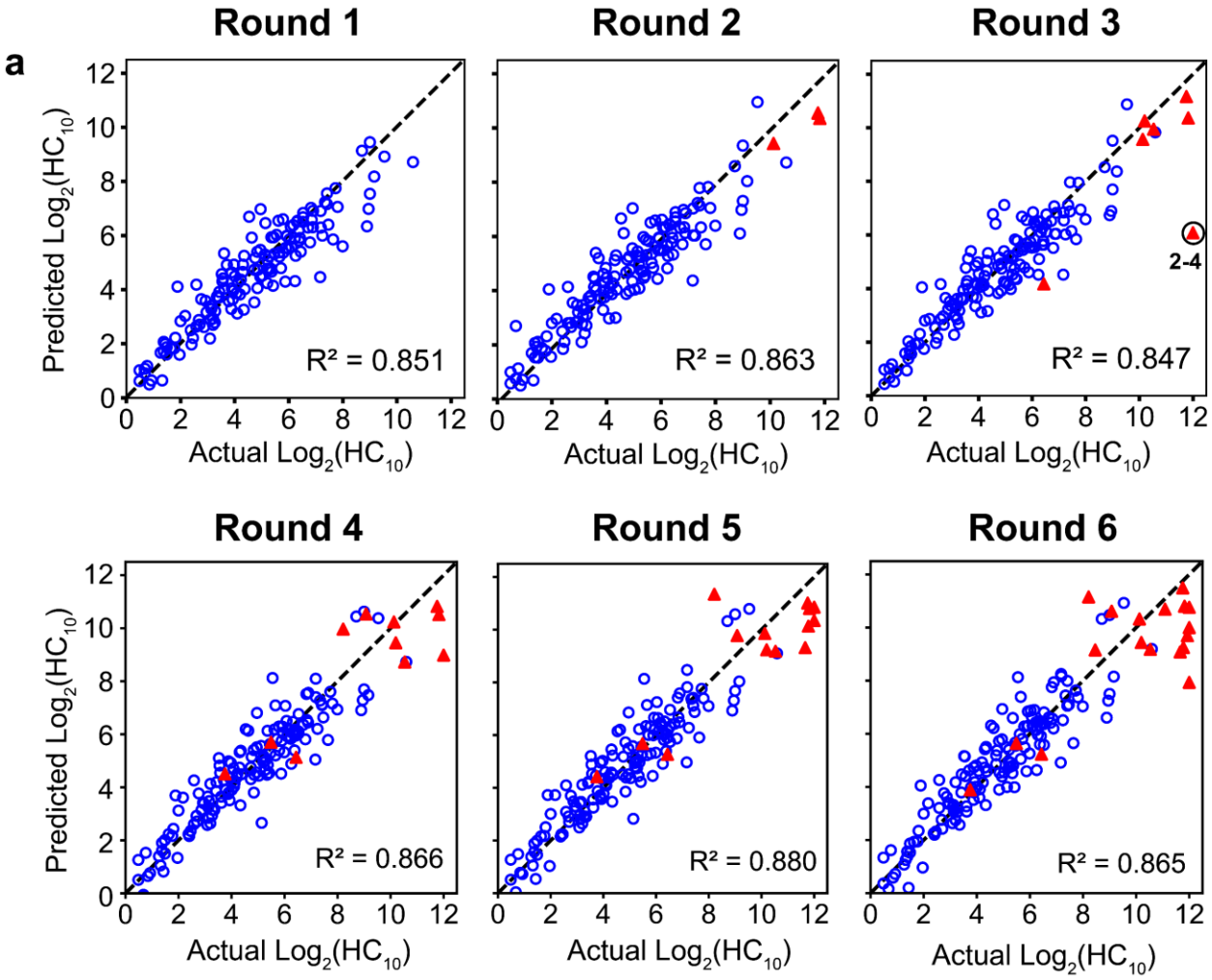


Figure S11. Parity plots for predicted $Log_2(HC_{10})$ values versus experimentally measured (actual) $Log_2(HC_{10})$ values for all rounds of model training. Open blue circles denote the initial 147 training sequences while red solid triangles denote new sequences discovered during the iterative GPR workflow. Points report test set predictions from 10-fold cross-validation (*i.e.*, predictions for when the corresponding sequence is in the test set and hence not used for model training). Coefficient of determination (R^2) values are also listed in each plot which match the solid black line in **Figure 4a**. Sequence 2-4 is circled in black in the parity plot for Round 3. The Round 6 parity plot is the same as **Figure 4b** in the main text.

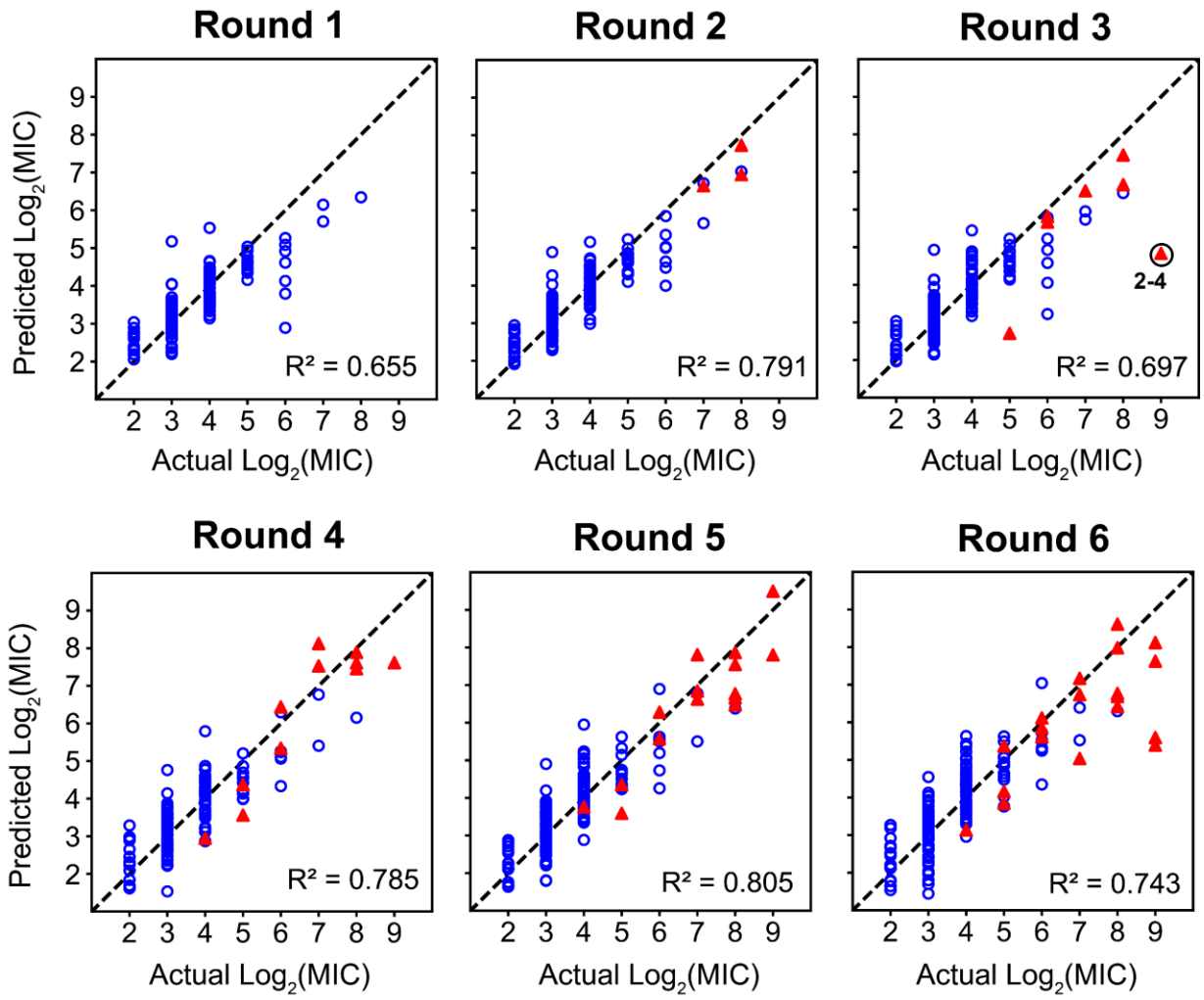


Figure S12. Parity plots for predicted Log_2(MIC) values versus experimentally measured (actual) Log_2(MIC) values for all rounds of model training. Open blue circles denote the initial 147 training sequences while red solid triangles denote new sequences discovered during the iterative GPR workflow. Points report test set predictions from 10-fold cross-validation (*i.e.*, predictions for when the corresponding sequence is in the test set and hence not used for model training). Coefficient of determination (R^2) values are also provided in each plot which match the dashed black line in **Figure 4a**. Sequence 2-4 is circled in black in the parity plot for Round 3. Round 6 parity plot is the same as **Figure 4c** in the main text.

Additionally, the following prediction metrics per round are tabulated in **Table S6** for both $log_2(HC_{10})$ and log_2(MIC) predictions: R^2 (also shown as black lines in **Figure 4a**), root mean squared error (RMSE), Pearson's r, and Maximum Error (also shown as red lines in **Figure 4a**). All metrics were calculated with 10-fold cross validation on proportionately allocated folds.

Table S6. Various model metrics across 6 rounds of GPR training for both the HC_{10} and MIC label prediction workflows: R^2 (coefficient of determination), RMSE (root mean squared error), Pearson's r (linear correlation coefficient), and Max. Error (maximum residual error).

		Prediction Round					
Metric		**1**	**2**	**3**	**4**	**5**	**6**
R^2	**$Log_2(HC_{10})$**	0.851	0.863	0.847	0.866	0.880	0.865
	Log_2(MIC)	0.655	0.791	0.697	0.785	0.805	0.743
RMSE	**$Log_2(HC_{10})$**	0.830	0.859	0.956	0.897	0.916	1.014
	Log_2(MIC)	0.640	0.556	0.718	0.630	0.658	0.803
Pearson's r	**$Log_2(HC_{10})$**	0.924	0.931	0.922	0.930	0.938	0.930
	Log_2(MIC)	0.813	0.891	0.839	0.886	0.898	0.863
Max. Error	**$Log_2(HC_{10})$**	2.698	2.790	5.921	3.016	3.133	4.065
	Log_2(MIC)	3.108	2.002	4.165	1.847	1.950	3.608

Figure S13 visualizes the prediction RMSE per round of newly discovered AMPs across the 6 prediction rounds ('Actual' vs 'Predicted' columns in **Figure 5a**). This figure captures several key changes and observations in the iterative workflow, including:

(1) The decrease in prediction RMSE from Rounds 1-3 using the same descriptor set (**Table S5**), illustrating model improvement.
(2) The increase in RMSE for Rounds 4 and 5 due to updating the set of descriptors based on LASSO CV (in Round 4) and switching from the ααβ to αααβαααβ backbone (in Round 5).
(3) The expected minimum value of the RMSE observed for Round 6, which results from the selection of test sequences with low uncertainty (low predicted standard deviation) compared to earlier rounds.

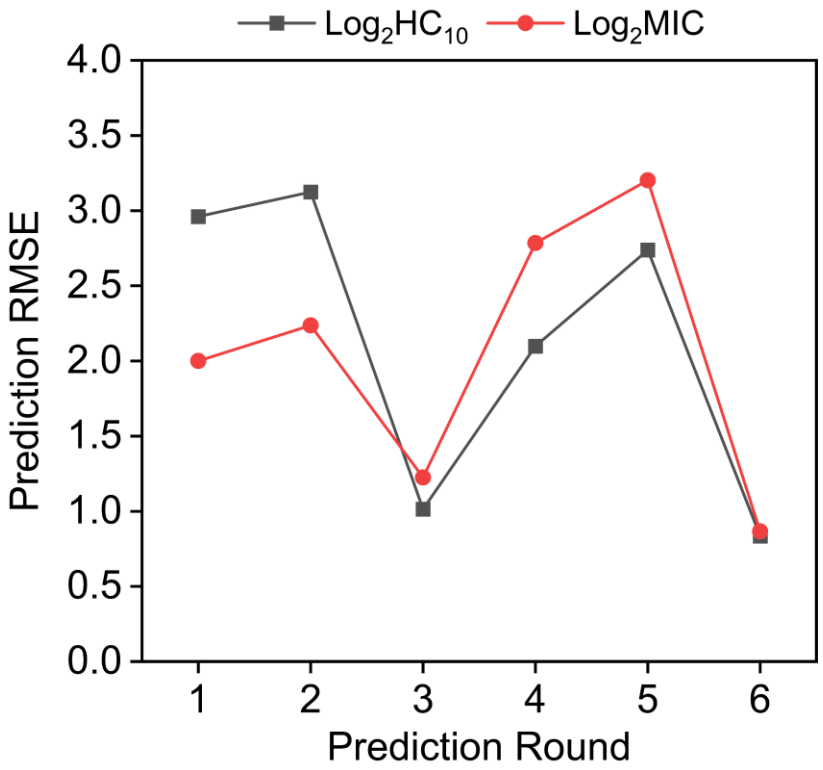


Figure S13. Prediction RMSE per prediction round for the iterative GPR workflow for both Log_2HC_{10} (black) and Log_2MIC (red) predictions. RMSE values are calculated based on the difference between experimentally measured ('Act.') and GPR predicted ('Pred.') values in **Figure 5a**.

As discussed in the **'Iterative GPR Model Training' Section** in the main text, we conducted a set of additional robustness checks on the GPR model to ensure maximum accuracy on new test sequence predictions.

First, we compared the 10-fold CV R^2 per round with and without the proportionate allocation procedure discussed in **Section S4** to validate that this methodology leads to consistently better model accuracy compared to a random allocation of training labels based on the order they appear in the training set (example for initial 147 training for prediction Round 1 in **Figure S5b**) as is implemented with the KFold() module in *sklearn*. These results are plotted per round for both the proportionately allocated model used in the main text (solid lines) and the random allocation robustness check (dashed lines) in **Figure S14**. For consistency with the methodology in the main text, the robustness check also applied the random allocation for descriptor selection for LASSO CV (see **Section S2**). Notably, log_2(MIC) predictions were more heavily impacted by the differences in these procedures, and model accuracy even decreases to below $R^2 = 0.4$ in Rounds 5 to 6.

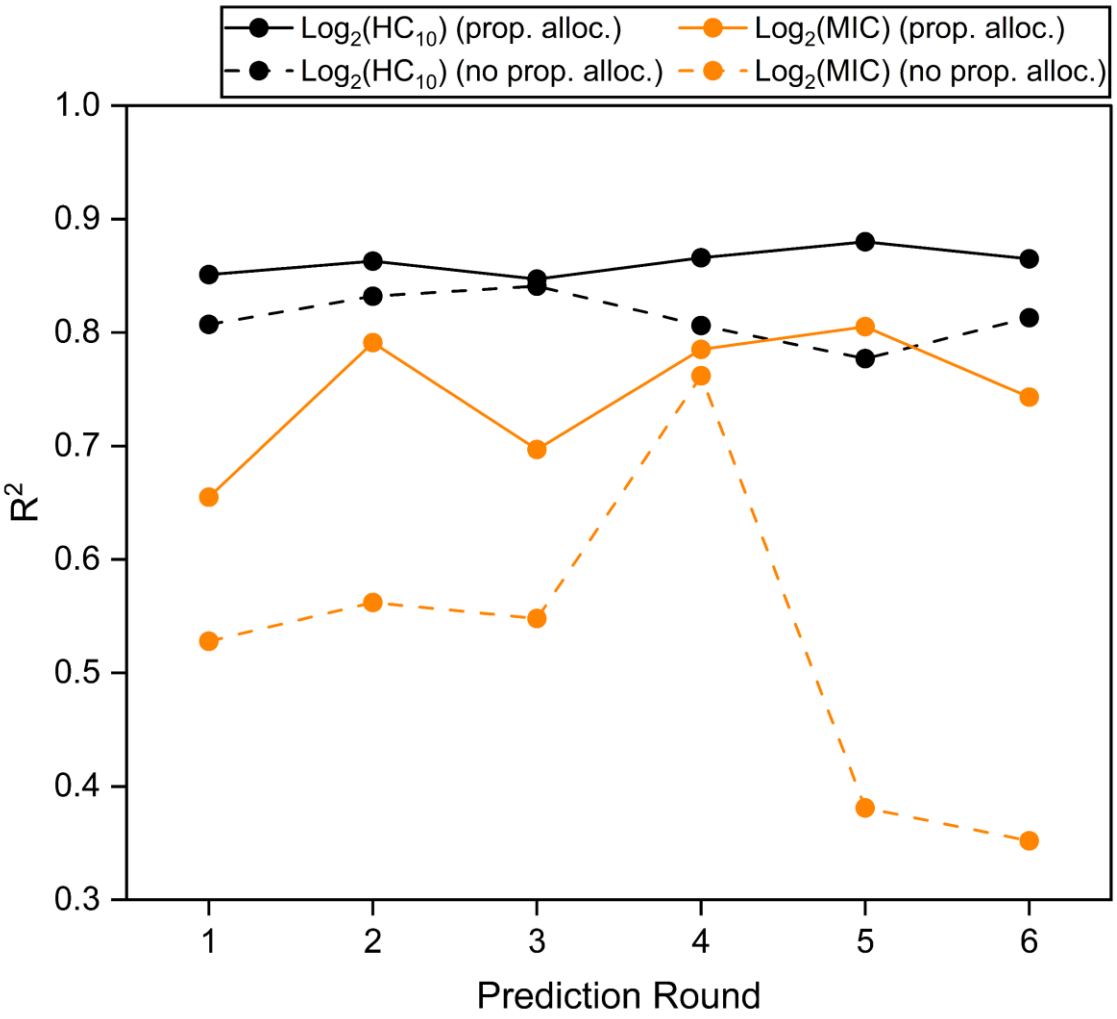


Figure S14. Model performance (average R^2 across 10-fold CV) comparing effect of proportionate allocation (prop. alloc.) of labels per round (see **Section S4**). $\log_2(HC_{10})$ values are black and $\log_2(MIC)$ values are orange. Solid lines are the same as in main text (**Figure 4a**) where labels are proportionately allocated (low to high magnitude) to each of the 10 folds before CV. Dashed lines indicate random allocation based on the order in which labels appear in training set.

An additional robustness check was to compare the procedure used in the main text, in which the GPR kernel and hyperparameters are updated each round (**Section S3**), to predictions obtained if the kernel and hyperparameters were instead kept consistent across all rounds. For this robustness check, models were trained using the Exponentiated Sine Squared(l=0.1, p=100) model with κ=0.1 for $\log_2(HC_{10})$ predictions and Rational Quadratic(l=100, α=0.1) with κ=0.1 for $\log_2(MIC)$ predictions for all 6 rounds since these were the kernel and hyperparameters selected for Round 1. Labels are proportionately allocated, and therefore the same descriptors were used for both the original methodology and this robustness check.

Figure S15 shows selecting optimal GPR model parameters each round (solid lines) led to a marginal increase or no change in R^2 compared to if the kernel and hyperparameters were kept

constant for Rounds 1 to 3. This result can be attributed to the consistent set of 20 descriptors for $\log_2(HC_{10})$ and 13 descriptors for $\log_2(MIC)$ that were used for these rounds (**Tables S1-S2**) which also led to small changes in the GPR model parameters when updated each round (**Table S4**). However, there was a large decrease in R^2 if GPR parameters were kept constant for Round 4 onwards, particularly for $\log_2(MIC)$ for which the GPR model lost all predictive capability for Rounds 5 and 6 (dashed orange curve). We attribute this behavior to the updated set of descriptors associated with these rounds. **Table S4** shows that the Dot Product kernel better predicted $\log_2(HC_{10})$ and $\log_2(MIC)$ labels once the number of descriptors increased. Therefore, **Figure S15** highlights the importance of the GridSearch CV procedure to maximize model accuracy each round after the set of descriptors is updated with LASSO CV.

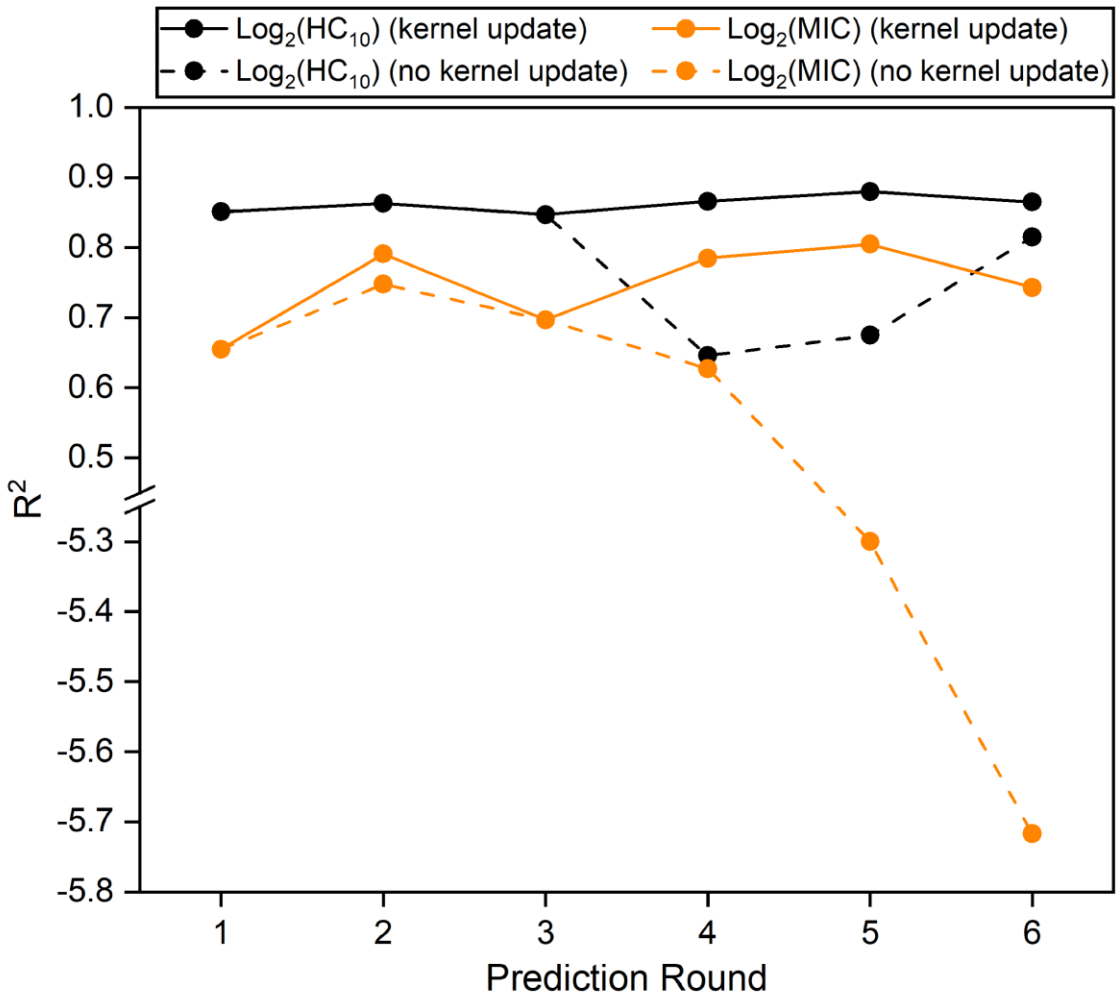


Figure S15. Model performance (average R^2 across 10-fold CV with proportionate allocation) comparing the effect of updating GPR kernel and hyperparameters for each prediction round. Solid lines are the same as in the main text (**Figure 4a**). Dashed lines indicate model predictions if the GPR kernel and hyperparameters are kept constant from Round 1 onwards.

Finally, we implemented a y-randomization procedure[7, 8] to check that GPR predictions were not the result of chance. In this approach, we compared distributions in the GPR model RMSE (computed as the average 10-fold CV RMSE) over 100 trials in which models were trained using

labels for both $log_2(HC_{10})$ and $log_2(MIC)$ that were randomly shuffled and proportionately allocated. The average RMSE across the 100 randomized trials is plotted as a red bar for each trial in **Figure S16**, and the minimum RMSE for these 100 trials is denoted as a horizontal dashed line with a caption. For reference, the average 10-fold RMSE for the model used in the main text is shown as green bars, and these values are also provided in the RMSE row in **Table S6**. These results demonstrate that the trained GPR model used for test sequence predictions (green bar) not only has a lower RMSE than the average randomized RMSE (red bar) but also the minimum RMSE across the 100 randomized trials (horizontal dashed line on red bar), thereby validating the robustness and accuracy of our GPR approach.

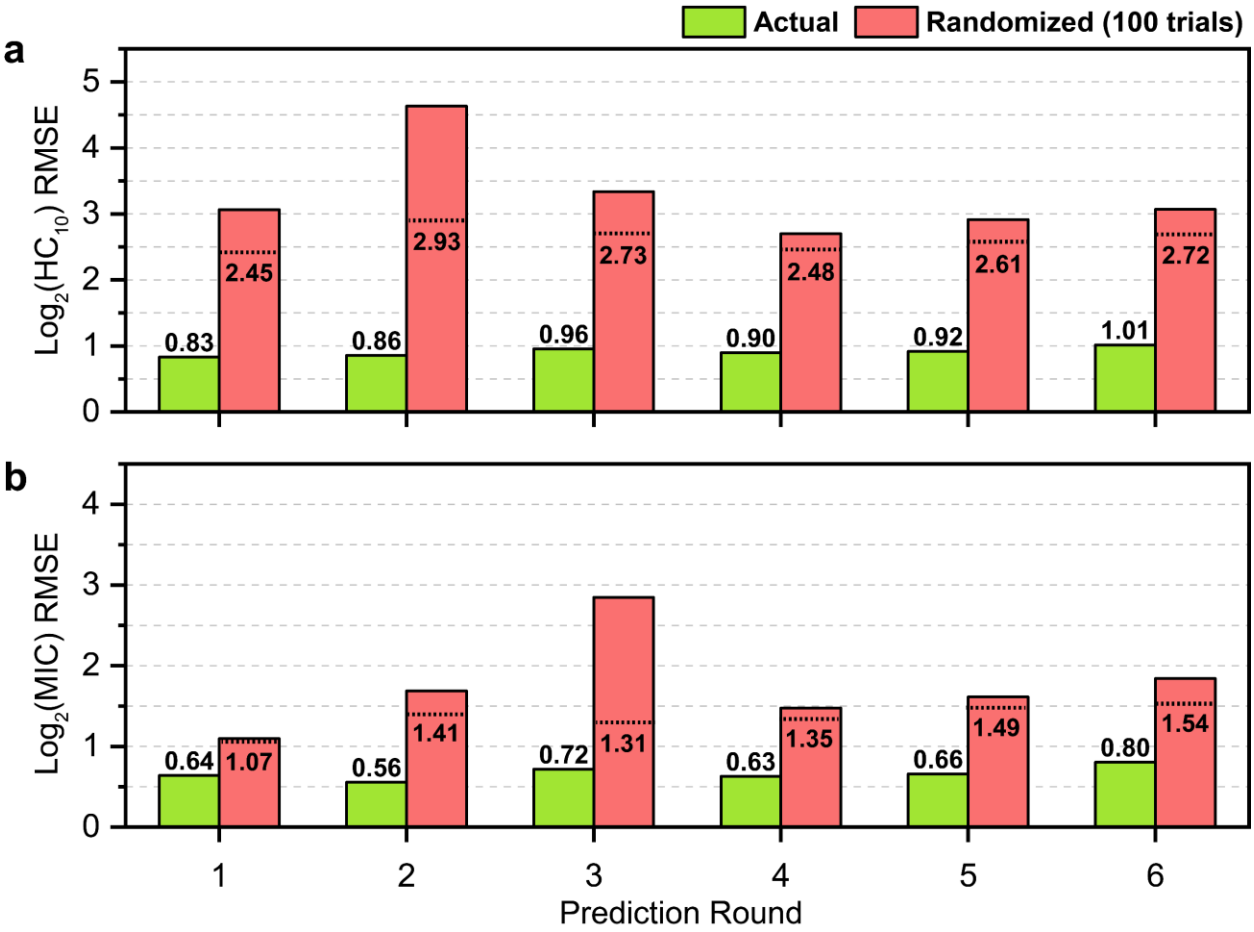


Figure S16. Testing model robustness per prediction round with y-randomization. The actual RMSE for each round based on GPR predictions for models trained using correct labels is shown in green. The average RMSE computed for 100 trials in which GPR models were trained using labels that were randomly shuffled before calculating RMSE in red. Comparisons are made for (a) $Log_2(HC_{10})$ and (b) $Log_2(MIC)$ labels. The actual RMSE and the minimum RMSE of the 100 trials for Randomized RMSE are labelled in bold for each round.

S8: Test Predictions and Uncertainty Analysis

To support discussion in the **'GPR Guides the Discovery of α/β-peptide Sequences with Novel Amino Acids and Motifs' Section** in the main text, we utilized the constant 14,137-sequence test

design space from Rounds 1 to 3 ('Total' column in **Table S5**) to probe if GPR model predictions for an amino acid that is newly introduced in a prediction round improve in future prediction rounds (based on decreases in the normalized standard deviation; NSD). **Figure S17** shows bar and whisker plots of NSDs for all sequences that contain at least 1 instance of newly introduced amino acids from Rounds 1 to 3 (Nle, Q, Nva, βY, W) for both the $\log_2(HC_{10})$ and $\log_2(MIC)$ prediction workflows. The sequences that introduce new amino acids relative to the initial 147-training set are labeled in blue to indicate the round in which they are introduced. Rounds 1-3 are shown as well as an additional hypothetical 4th round (labeled as '4*') in which descriptors are kept the same as rounds 1-3 to maintain the same 14,137-sequence space for direct comparison of impacts on the NSD distribution for new amino acids introduced in Round 3 (Nva, βY, W). Additionally, the plot shows Round 4 considered in the main text (labeled as '4') in which descriptors are updated via LASSO CV (**Tables S1-S2**). NSD distributions for Round 4 are included to analyze the impact of updating descriptors on NSD distributions, although the test sequence spaces vary slightly between Rounds 1-3 (14,137) and Round 4 (17,238) as shown in **Table S5** because of this descriptor update.

The NSD distributions for newly introduced amino acids in Rounds 1-3 have the following trends for $\log_2(HC_{10})$ predictions:

(1) Nle does not exhibit a strong increase or decrease in the NSD distribution as Nle-containing sequences are reintroduced across the rounds. This is most likely attributed to the large number of sequences in the initial 147-sequence training set containing Ala and Abu (**Figure 3b**), which are chemically similar to Nle.
(2) Q, βY, and W exhibit a sharp decrease in the mean and median NSD as well as a widening of the NSD distribution towards lower NSD values.
(3) Nva (**Figure S17c**) followed very similar trends to Nle (**Figure S17a**) due to similarities in chemical structure as discussed in the main text.

Conversely, for $\log_2(MIC)$ distributions, a consistent decrease in the NSD distribution generally resulted from the updated set of descriptors in Round 4, further supporting the need to use an increased number of descriptors (see **Table S2**) after Round 3 to more accurately describe newly introduced amino acids.

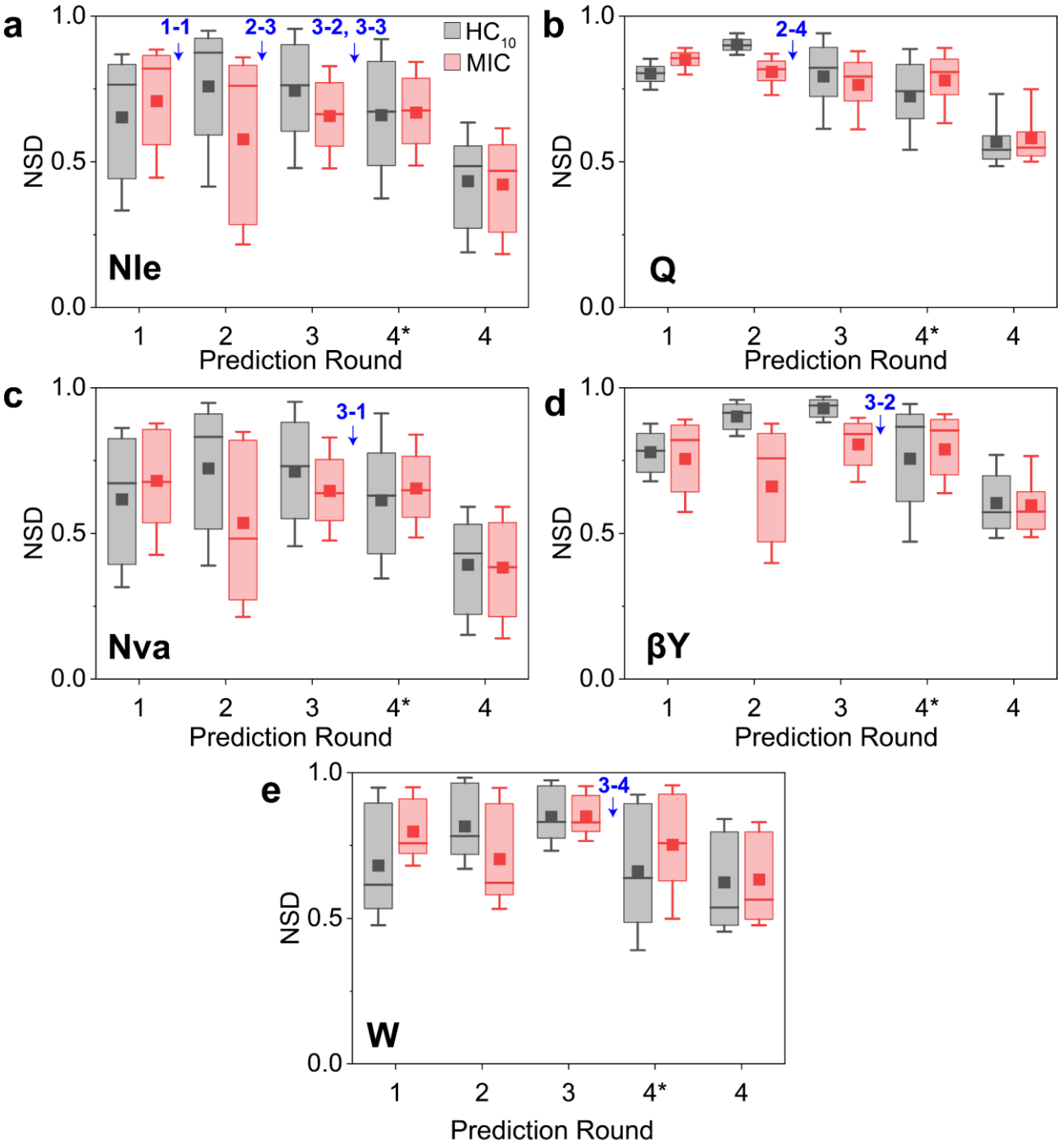
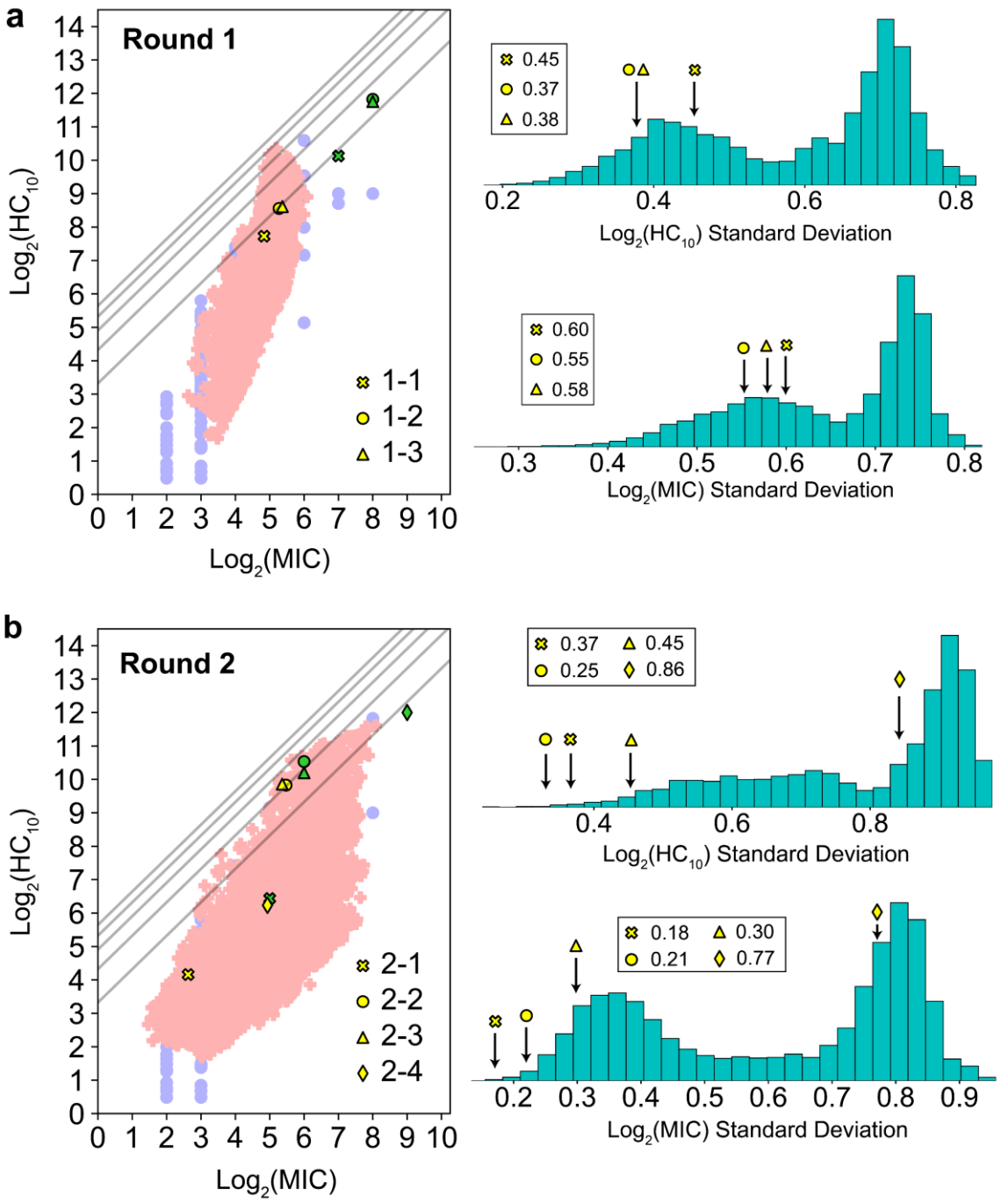
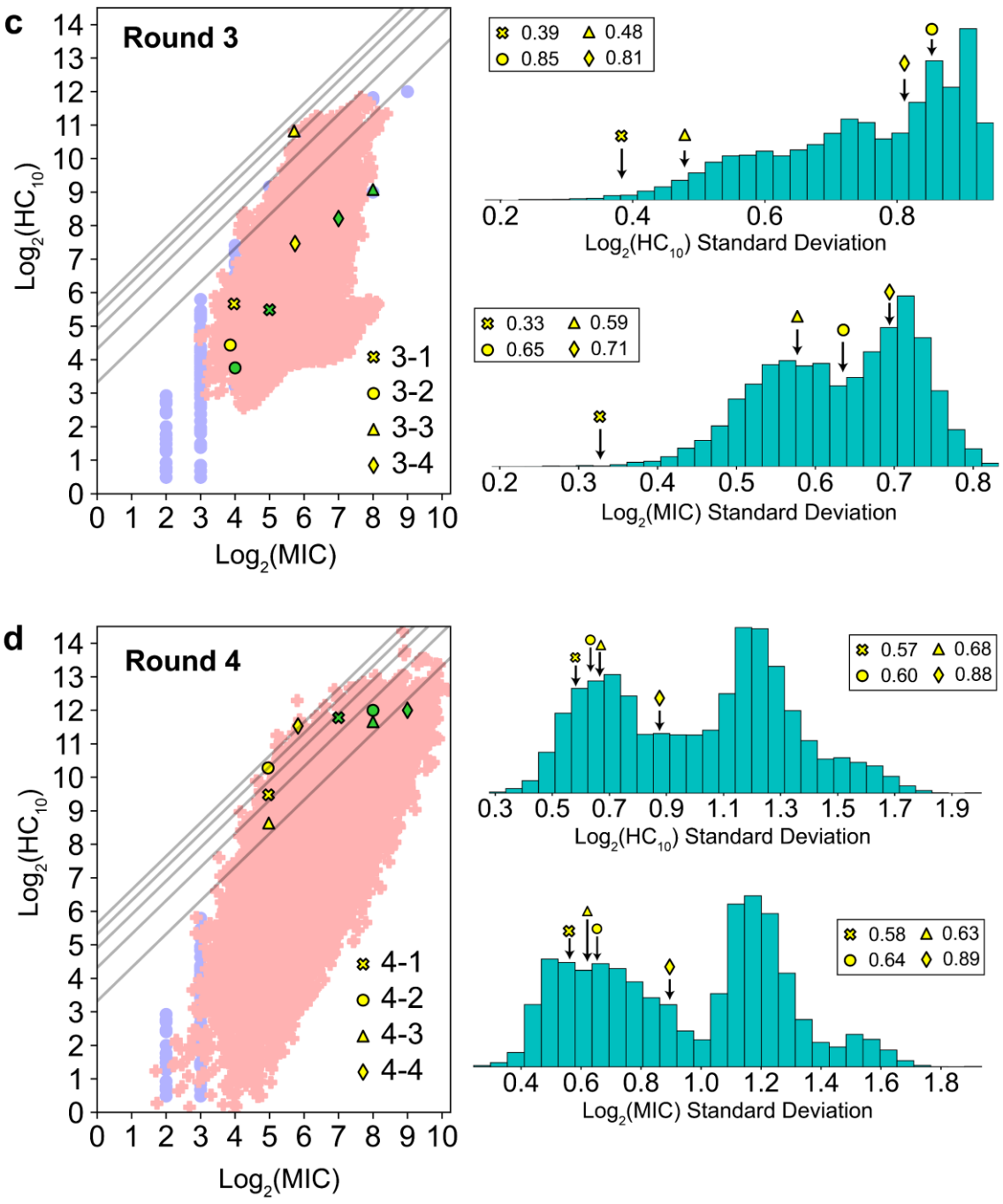


Figure S17. Normalized standard deviation (NSD) distributions for all test sequences containing at least 1 of the amino acids newly introduced in Rounds 1-3 (see **Figure 5a**): (a) Nle, (b) Q, (c) Nva, (d) βY, and (e) W. Distributions for $\log_2(HC_{10})$ and $\log_2(MIC)$ predictions are plotted in black and red, respectively. Rounds 1-4 are plotted as well as a hypothetical Round 4 (4*) in which the set of descriptors was the same as Rounds 1-3 for direct comparison of NSD distributions for the same 14,137 test sequences (see **Table S5**). Mean NSD is plotted as a solid square with horizontal lines indicating the 10th, 25th, 50th, 75th, and 90th percentiles. Rounds in which sequences with new amino acids are introduced into the training data are labeled in blue (with labels corresponding to the labels in **Figure 5a**).

Figure S18 shows GPR predictions for the entire test sequence design space considered per round (from Round 1 in **Figure S18a** to Round 6 in **Figure S18f**), with all predicted $Log_2(HC_{10})$ and $Log_2(MIC)$ values shown as red crosses. Predicted $Log_2(HC_{10})$ and $Log_2(MIC)$ values for test sequences that were chosen for experimental synthesis are indicated by yellow points ('Pred' columns in **Figure 5a**) with corresponding experimentally determined values indicated by green points ('Act' columns in **Figure 5a**). For each $Log_2(HC_{10})$ *vs*. $Log_2(MIC)$ plot, the corresponding standard deviation ranges for test sequence predictions are shown as blue-green histograms at right. Yellow points above these histograms indicate the standard deviations for the sequences selected for experimental synthesis.

Figure S18 shows that, in general, test sequences with low standard deviations that were selected for experimental synthesis led to more accurate model predictions based on proximity to the corresponding green point. In general, overall model accuracy increased (leading to a general decrease in distance between yellow predictions and green measured values) as the criteria for selecting test sequences shifted from probing new amino acids and low certainty (high standard deviation) test sequences in early rounds to selecting high certainty (low standard deviation), high selectivity (large SI) test sequences in later rounds.


a
Round 1
Log2(HC10)
Log2(MIC)
1-1
1-2
1-3
0.45
0.37
0.38
Log2(HC10) Standard Deviation
0.60
0.55
0.58
Log2(MIC) Standard Deviation
b
Round 2
Log2(HC10)
Log2(MIC)
2-1
2-2
2-3
2-4
0.37
0.45
0.25
0.86
Log2(HC10) Standard Deviation
0.18
0.30
0.21
0.77
Log2(MIC) Standard Deviation


c
Round 3
Log2(HC10)
Log2(MIC)
3-1
3-2
3-3
3-4
0.39 0.48
0.85 0.81
Log2(HC10) Standard Deviation
0.33 0.59
0.65 0.71
Log2(MIC) Standard Deviation
d
Round 4
4-1
4-2
4-3
4-4
0.57 0.68
0.60 0.88
Log2(HC10) Standard Deviation
0.58 0.63
0.64 0.89
Log2(MIC) Standard Deviation

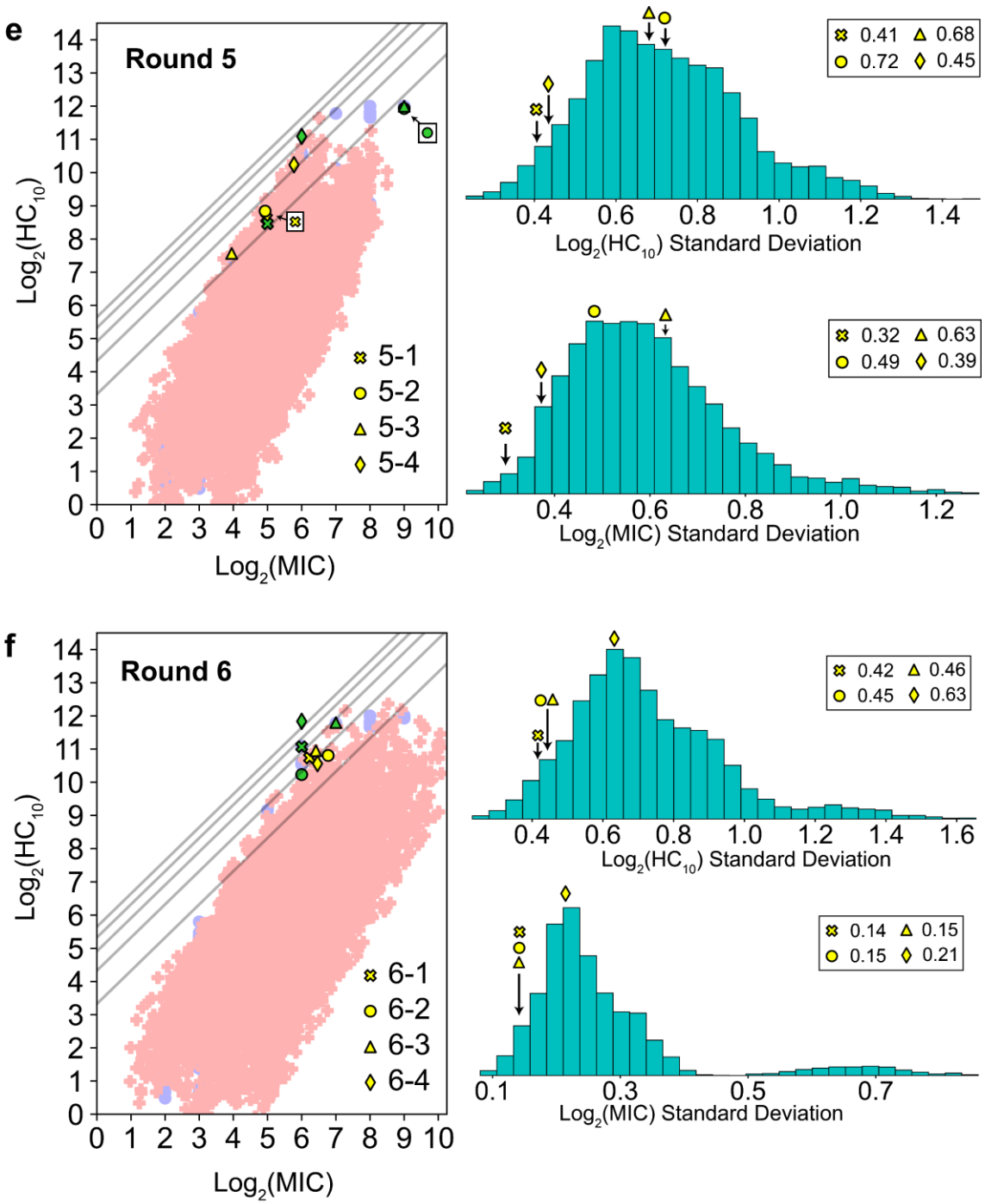


Figure S18. GPR predictions for the entire test sequence design space and corresponding standard deviation ranges per round from (a) Round 1 to (f) Round 6. Red crosses indicate predictions for test sequences, blue circles indicate values for the training set, yellow points indicate predicted values for sequences selected for experimental evaluation ('Pred.' column in **Figure 5a**), and green points indicate corresponding experimentally determined values for the selected sequences ('Act.' column in **Figure 5a**). SI bands of 10, 20, 30, 40, and 50 are grey lines. The standard deviation range of all test sequences (red crosses) are plotted as blue-green histograms on the right. Standard deviations for test sequences selected for experimental evaluation are also labeled in yellow with arrows indicating their locations in the histograms.

S9: Descriptor Importance with Shapley Analysis

To better understand the molecular descriptors that contribute most to model predictions, we conducted additional descriptor importance analysis on Round 6 results through the calculation of Shapley values. The Shapley value calculated for each peptide descriptor value quantifies its relative impact on overall GPR model predictions (either $Log_2(HC_{10})$) or $Log_2(MIC)$): positive Shapley values indicate that the descriptor pushes model predictions to higher values while negative Shapley values indicate that the descriptor pushes model prediction to lower values.[9] To quantify the most impactful descriptors for both $Log_2(HC_{10})$) or $Log_2(MIC)$ predictions, we provide beeswarm plots in **Figure S19** of the top 10 descriptors in decreasing order of the mean absolute Shapley value across all peptides (y-axis) and corresponding Shapley values for all peptide descriptor values (x-axis) as points on these plots. By color coding each of these points by the relative descriptor values (low descriptor values in blue and high descriptor values in red) we can additionally visualize if increases in descriptor values are directly or inversely correlated to their Shapley values. For instance, the descriptor that contributes most to $Log_2(HC_{10})$ predictions is 'Chi4n', and an increase in descriptor value with decreasing Shapley value indicates that peptide sequences with large Chi4n will tend to be predicted with low $Log_2(HC_{10})$.

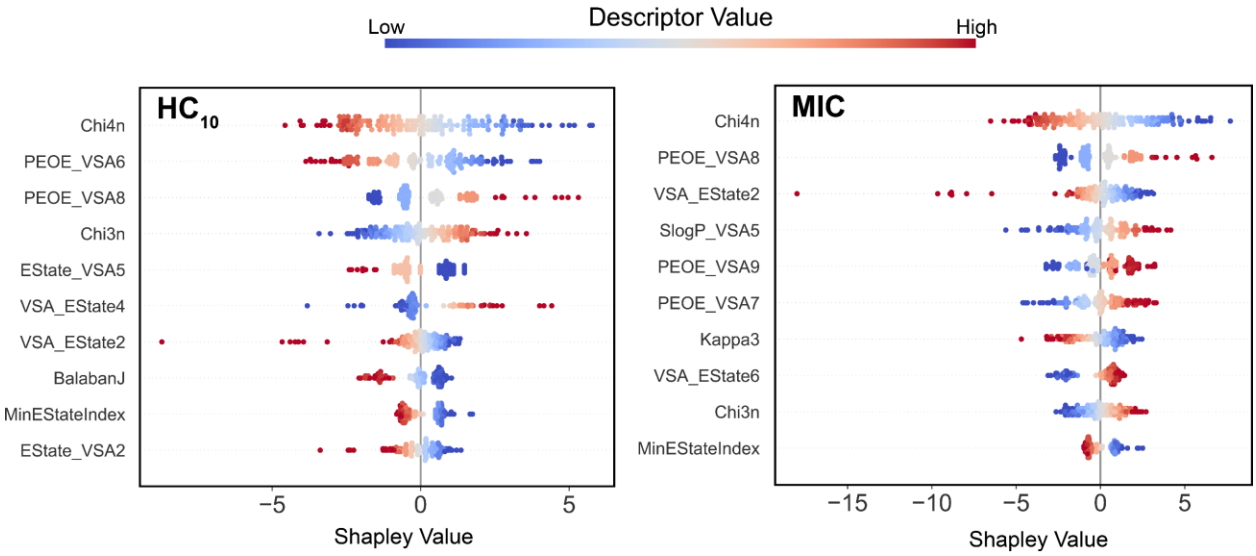


Figure S19: Beeswarm plots visualizing the top 10 most impactful descriptors for round 6 predictions for HC_{10} (left) and MIC (right) experimental labels. Descriptors are ordered by decreasing mean absolute Shapley value in each plot. Each point represents a unique peptide in the training set and is color coded by relative descriptor value to the full training set (blue = low, red = high).

As shown in these plots, the most impactful descriptors for HC_{10} and MIC predictions are:

- Chi4n: This metric quantifies the complexity of higher order connectivities in heavy atom molecular graph representations[4] of peptide sequences (for example the presence of rings or branched sidechains compared to linear segments).
- PEOE_VSA6/8: These metrics quantify van der Waals surface area (VSA) atomic contributions normalized by atomic partial charge (PEOE).[2]

- VSA_EState2: This metric quantifies 'Electrotopological-State' (EState) indices[3] that encode electronic atomic contributions of the peptide normalized by its VSA.

While these descriptors provide some qualitative insight into what appears important in model predictions, they overall lack physical interpretability (reference discussion in **Section S2**) and have no clear relation to either hydrophobicity or helical flexibility. Hydrophobicity is most closely related to 'MolLogP' (octanol-water partition coefficient) descriptor which is not kept during LASSO CV for model predictions for any round (**Table S1-S2**) and is therefore not included in this Shapley analysis. Moreover, we expect the overall hydrophobicity of the entire sequence to be a complex function of sequence descriptors that may not be clearly related to a single descriptor.[10, 11] Helical flexibility is not captured by RDKit since SMILES strings only encode the 2D connectivity, branching, etc. of input AMP sequences.[1]

Experimental evaluation and characterization

S10: Experimental methods

Peptide synthesis. Peptides were synthesized via microwave-assisted Fmoc solid-phase peptide synthesis on a TentaGel S RAM resin (20-40 μmol scale) as reported previously.[12-14] Briefly, solutions of Fmoc-α-amino acid or Fmoc-β-amino acid, coupling reagent (HATU), and base (DIPEA) in DMF were mixed before coupling. Microwave (CEM Discover) irradiation methods were used for coupling of Fmoc amino acids (600 W maximum power, 70 °C, ramp 2 min, hold 8 min) and deprotection of Fmoc (600 W maximum power, 80 °C, ramp 2 min, hold 4 min). After each coupling and deprotection step, the resin was thoroughly washed with DMF and CH_2Cl_2. If sequences did not contain tryptophan or methionine, the peptide was cleaved from the resin by TFA containing 2% H_2O and 2% triisopropylsilane, while peptides with methionine or tryptophan were cleaved using TFA containing 5% H_2O, 5% thioanisole, and 2.5% 1,2-ethanethiol for 1 to 2 h. The crude product was purified by preparative RP-HPLC with a gradient of 25%–73% CH_3CN in water containing 0.1% TFA. Electrospray ionization (ESI) mass spectrometry (Thermo Q Exactive Plus with quadrupole ion trap or Bruker impact II) was used to determine α- and α/β-peptide masses. The calculated and measured peptide masses from ESI mass spectrometry are shown in **Table S7** (newly discovered peptides) and **Section A1** (previously unreported peptides). Full ESI mass spectra for newly discovered peptides are provided in S**ection A2.** The determined purity of peptides was over 95% by subsequent analytical RP-HPLC analysis (representative curves shown in **Figure S20**, and full HPLC curves are provided in **Sections A1** and **A2**).

Antifungal minimum inhibitory concentration (MIC) characterization. Antifungal MIC assays were conducted as previously described.[12, 13] *Candida spp.* cells were streaked on a yeast peptone dextrose (YPD) agar plate from a frozen stock solution and grown overnight at 30 °C. For each assay, a colony was collected from the YPD plate and grown overnight in autoclaved test tubes at 30 °C with shaking in liquid YPD broth and cells were then washed, resuspended, and prepared for subsequent experiments. The antifungal activities of the compounds were determined in 96-well plates according to planktonic broth microdilution susceptibility testing assay guidelines provided by the Clinical and Laboratory Standards Institute. The assay was modified to include a quantitative XTT assessment of cell viability. From DMSO stock solutions of peptides, two-fold

serial dilutions (100 μL) of compounds (resulting in a total of 2% DMSO at the highest concentration tested) in RPMI (pH adjusted to 7.4 with MOPS) were mixed with 100 μL of fungal cell suspension (grown for 24 h at 30 °C and concentration adjusted to 5×10^3 cells/mL) and the plates were incubated at 37 °C for 48 h. Wells lacking compound (cell controls) and wells lacking both compounds and cells (medium sterility controls) were included in every plate. After 48 h, 100 μL of XTT solution (0.5 g L^{-1} in DPBS, pH 7.4, containing 3 μM menadione in acetone) was added to all wells, and plates were incubated at 37 °C in the dark for 1.5 hours and absorbance measurements at 490 nm were recorded using a plate reader (Tecan Infinite M200 PRO, Tecan Life Sciences, Inc). Cell viability was plotted as a function of compound concentration. Percent cell viability was calculated by the below equation, where A_{490}, $A_{490}^{\text{cell control}}$, and $A_{490}^{\text{background}}$ are the average absorbance values of the supernatant (at 490 nm) from wells containing a specific concentration of compound, wells with positive cell control, and medium sterility control wells, respectively.

$$\text{cell viability (\%)} = \frac{\left(A_{490} - A_{490}^{\text{background}}\right)}{A_{490}^{\text{cell control}} - A_{490}^{\text{background}}} \times 100$$

Experiments were performed in a minimum of two technical replicates per concentration and repeated in at least three independent experiments. After averaging, the lowest assayed concentration of compound that resulted in a decrease in normalized absorbance of at least 90% of the mean was taken as the minimum inhibitory concentration (MIC) of that compound. For determination of smaller-interval MIC of a selected panel of peptides, the assay was modified to test peptides in intervals of 5 μg/mL. Briefly, starting from a DMSO stock, peptides were prepared as 0.2 mg/mL in RPMI with <1% total DMSO concentration. Then peptides were further diluted in intervals of 10 μg/mL in a 96-well plate with a total volume of 100 uL in each well. Then 100uL RPMI solution containing 5×10^3 cells/mL were added in each well to obtain 5 μg/mL intervals. The average cell viability curves of all fungal strains for each peptide are shown in **Section S12** and **S14**. For all experiments, DMSO vehicle controls were added as comparisons to ensure that DMSO concentrations used did not affect cell viability.

Hemolysis assay. Hemolysis assays were performed as previously described.[12-15] Human red blood cells (RBCs) were washed three times with tris-buffered saline (TBS, 10 mM Tris-HCl, 100 mM NaCl, pH 7.5), and then diluted 50-fold in TBS to obtain 2% RBCs relative to total RBCs in whole blood. Two-fold serial dilutions (50 μL) of peptides prepared in TBS were mixed with 50 μL of a 2% RBC suspension in a 96-well plate and then incubated at 37 °C for 1 h. Melittin served as a positive lysis control and TBS was used as a negative lysis control. Plates were then centrifuged at 3000 rpm for 5 min, 75 μL of the supernatant was transferred into a fresh plate, and absorbance was measured at 405 nm using a plate reader. The percent hemolysis was calculated as:

$$Hemolysis\ (\%) = \frac{A_{405} - A_{405}^{negative\ control}}{A_{405}^{positive\ control} - A_{405}^{negative\ control}} \times 100$$

where A_{405}, $A_{405\ negative\ control}$, and $A_{405\ positive\ control}$ are the average absorbance values at 405 nm of the supernatant of RBCs treated with peptides, RBCs in TBS lacking peptides, and melittin treated RBCs, respectively. Experiments were performed in duplicate and repeated on at least three

different days. The concentration of peptide at which 10% hemolysis occurred (HC_{10}) was calculated for each experiment and arithmetic averages of HC_{10} values were calculated and reported as shown in **Section S12**. For curve fitting and determination of HC_{10}, a "neutcurve" package in Python 3 was used, which was written and distributed by the Bloom Lab at the Fred Hutch Cancer Center (https://github.com/jbloomlab/neutcurve).

Antibacterial MIC assays. Bacterial cells were streaked from frozen stocks in a similar fashion to fungal cells, but on trypsin-soy agar plates. The antibacterial activities of peptides were determined in 96-well plates according to the planktonic broth microdilution susceptibility testing assay guidelines provided by the Clinical and Laboratory Standards Institute. The assay was modified to include a quantitative XTT assessment of cell viability. From DMSO stock solutions of peptides, two-fold serial dilutions (100 μL) of compounds (resulting in a total of 2% DMSO at the highest concentration tested) in Mueller-Hinton Broth were mixed with 100 μL of bacterial cell suspension (grown for 24 h at 37 °C and concentration adjusted to 1×10^6 cells/mL based on solution optical density at 600 nm), and the plates were incubated at 37 °C for 24 h. Wells lacking compound (cell controls) and wells lacking both compounds and cells (medium sterility controls) were included in every plate. After 24 hours, 100 μL of XTT solution was added in the same manner as for antifungal MIC assays described above and the MIC was determined based on 90% reduction of normalized absorbance, consistent with the antifungal assays. The average cell viability curves for each peptide are shown in **Section S14**. DMSO vehicle controls were added as comparisons to ensure that DMSO concentrations used did not affect cell viability.

Characterization of hydrophobicity of α- and α/β-peptides. The hydrophobicity of α- and α/β-peptides was measured as described previously.[12, 13, 15] Briefly, retention times of peptides in analytical RP-HPLC using a C18 column (Waters, XBridge) were recorded by dissolving peptides to a concentration of 0.5 to 1 mg/mL in deionized H_2O containing 20 – 30% ACN and 0.1% TFA and then 50 μL of the peptide solution was injected into the HPLC. Retention time was quantified in triplicate with a gradient of 20-80% CH_3CN in water containing 0.1% TFA over 5-35 min.

Helicity and helical rigidity of α- and α/β-peptides. Stock solutions (0.2 mg/mL) in deionized water were prepared, aliquoted, and then lyophilized to obtain the desired amounts of peptides. Peptides were then dissolved in either trifluoroethanol (TFE) or 15% TFE in deionized water to yield a final peptide concentration of 0.1 mM. The 15% TFE concentration condition was used as a substitute to a fully aqueous solvent to obtain quantifiable differences in helicity with changes in sequence due to the inherently low helicity of lead compounds. Circular dichroism (CD) was measured in triplicate using a JASCO-1500 Circular Dichroism Spectrophotometer at 25 °C with a 1 mm path length cell and 4 second digital integration times. We used the CD minimum at 222 nm to estimate the α-helicity of α-peptides and the CD minimum at 206 nm to estimate the helicity of α/β-peptides at 206 nm. In each case, the intensity of the CD minimum measured in pure trifluoroethanol, which increases or saturates oligopeptide hydrophobicity, was used to estimate the signature for the maximum helicity achievable by the peptide. Results are shown in **Section S13**. The estimated relative percent helical rigidity was calculated using the following equation:

For α-peptides:

$$\text{Helical rigidity (\%)} = \frac{[\theta]^{222nm}_{15\%\ trifluoroethanol}}{[\theta]^{222nm}_{trifluoroethanol}} \times 100$$

For α/β peptides:

$$\text{Helical rigidity (\%)} = \frac{[\theta]^{206nm}_{15\%\ trifluoroethanol}}{[\theta]^{206nm}_{trifluoroethanol}} \times 100$$

Statistical analysis. Statistical analysis to compare the molar ellipticities of top-SI test peptides with averages of those of their corresponding backbone templates were performed using Graphpad Prism 10 (version 10.2.3 for Windows, GraphPad Software, San Diego, California, USA). One-way analysis of variance (ANOVA) followed by post-hoc Tukey's test was used after normality testing of sample data. Results were considered statistically significant at $p < 0.05$.

S11: Peptide mass and purity information

Table S7 – Mass spectrometry data for all new test set α/β-peptides introduced through iterative GPR. Ionized mass spec data was obtained from multiple mass spec equipment (Bruker Impact II or Q Exactive Plus Orbitrap) and measured masses at different ionization states were converted to full mass values. Mass spectra are provided in **Section A2**.

Round	Idx	MW. Calc	MW. Found	# of ionization	Total mass found
1	1	508.0251	508.0263	[M+3H]+3	1522.065
	2	494.0095	494.0084	[M+3H]+3	1480.011
	3	498.6813	498.6816	[M+3H]+3	1494.031
2	1	771.5184	771.5140	[M+2H]+2	1542.021
	2	741.5002	741.4985	[M+2H]+2	1481.990
	3	494.6692	494.6705	[M+3H]+3	1481.995
	4	513.0166	513.0176	[M+3H]+3	1537.039
3	1	510.0095	510.0091	[M+3H]+3	1528.013
	2	520.0130	520.0118	[M+3H]+3	1558.021
	3	499.3411	499.3412	[M+3H]+3	1496.009
	4	523.0131	523.0120	[M+3H]+3	1567.022
4	1	503.685	503.6866	[M+3H]+3	1509.048
	2	499.0131	499.0130	[M+3H]+3	1495.027
	3	508.3569	508.3563	[M+3H]+3	1523.056
	4	532.0396	532.0405	[M+3H]+3	1594.109
5	1	743.4871	743.4873	[M+2H]+2	1485.969
	2	543.6958	543.6960	[M+3H]+3	1629.074
	3	787.5133	787.5146	[M+2H]+2	1574.024
	4	794.0291	794.0300	[M+2H]+2	1587.055
6	1	751.4845	751.4845	[M+2H]+2	1501.963
	2	529.6885	529.6881	[M+3H]+3	1587.049
	3	794.0291	794.0292	[M+2H]+2	1587.051
	4	548.0396	548.0395	[M+3H]+3	1642.103

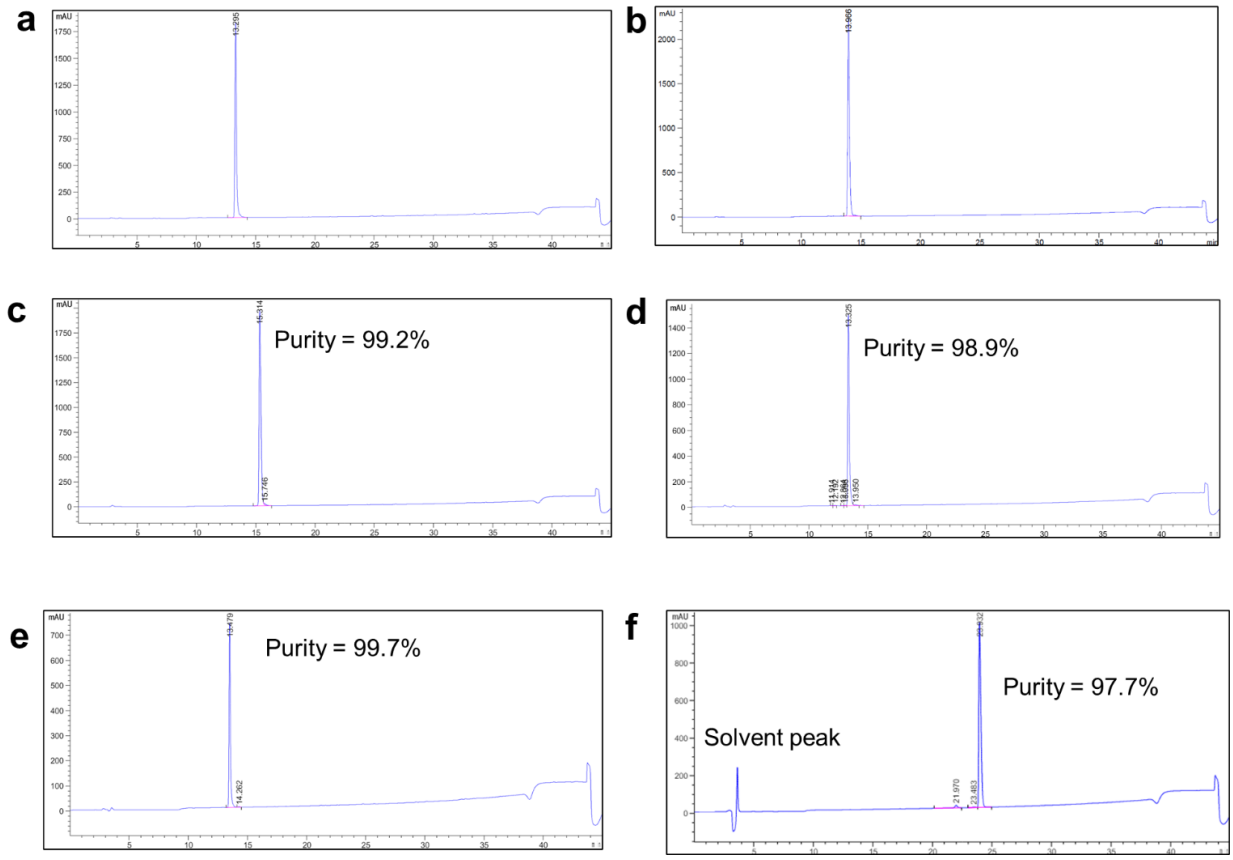


Figure S20. Representative analytical RP-HPLC profiles used to quantify the purity of α/β-aurein analogues. RP-HPLC traces of the highest selectivity α/β-peptides (a) 6-4, (b) 5-4, (c) 6-1, (d) 6-3, (e) 4-1, and (f) aurein 1.2-NH_2 are shown. All peptides reported in the study were > 95% pure. Full HPLC curve data included in **Section A2.**

S12: Experimental evaluation of *C. albicans* MIC assays and hemolysis

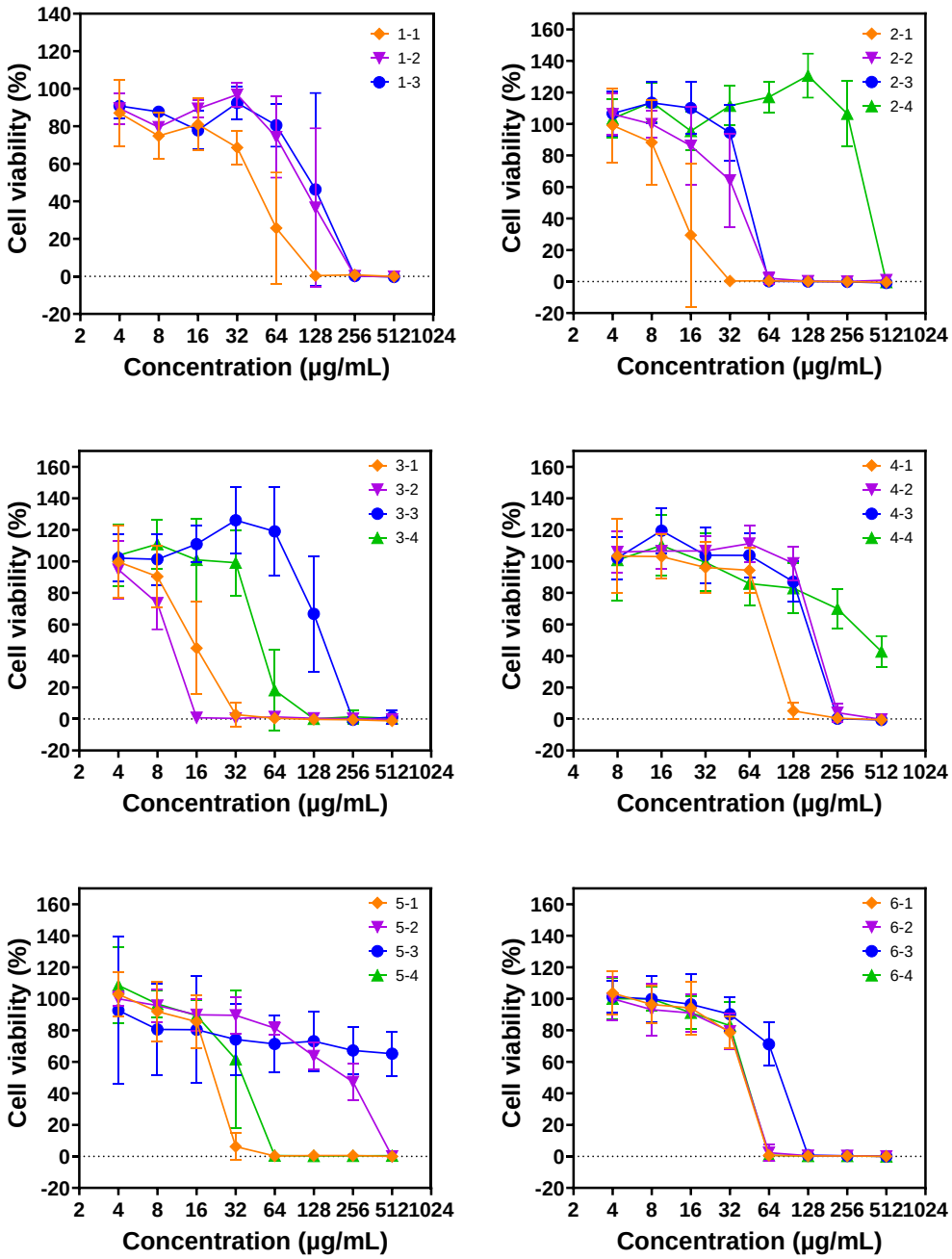


Figure S21. MIC curves for *C. albicans* cells over 6 iterative rounds. Fungal cells (5×10^3 cells/mL) were incubated with compounds for 48 h and susceptibility was assessed using an XTT reduction assay to compare the absorbance at 490 nm for compound-treated samples and untreated samples. Data points are the average of at least three independent experiments with two technical duplicates each or more and error bars represent the standard deviation. Round 1 involved two independent experiments instead.

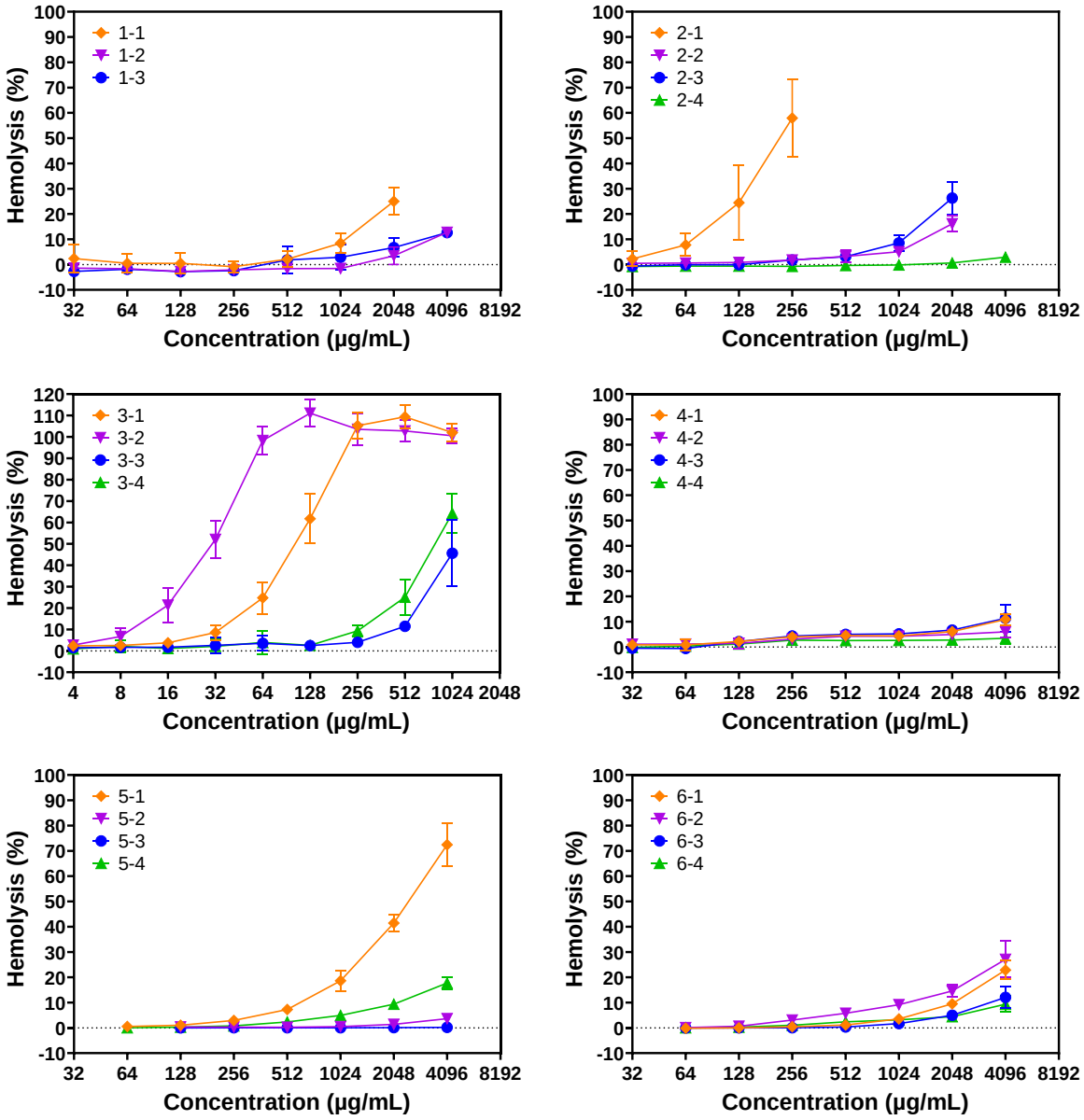


Figure S22. Hemolysis curves of peptides. Peptides were incubated with human RBCs for 1 h, and the absorbance of the supernatant was measured at 405 nm and normalized to melittin-treated RBCs, corresponding to 100% hemolysis. Data points are the average of at least three independent experiments with two technical replicates each or more and error bars represent the standard deviation. Round 1 involved two independent experiments instead. HC_{10} values were determined using a Hill function calculation package in Python (Neutcurve, Bloom Lab)[16] for each experimental curve and the geometric mean was used for the average.

S13: Helicity characterization

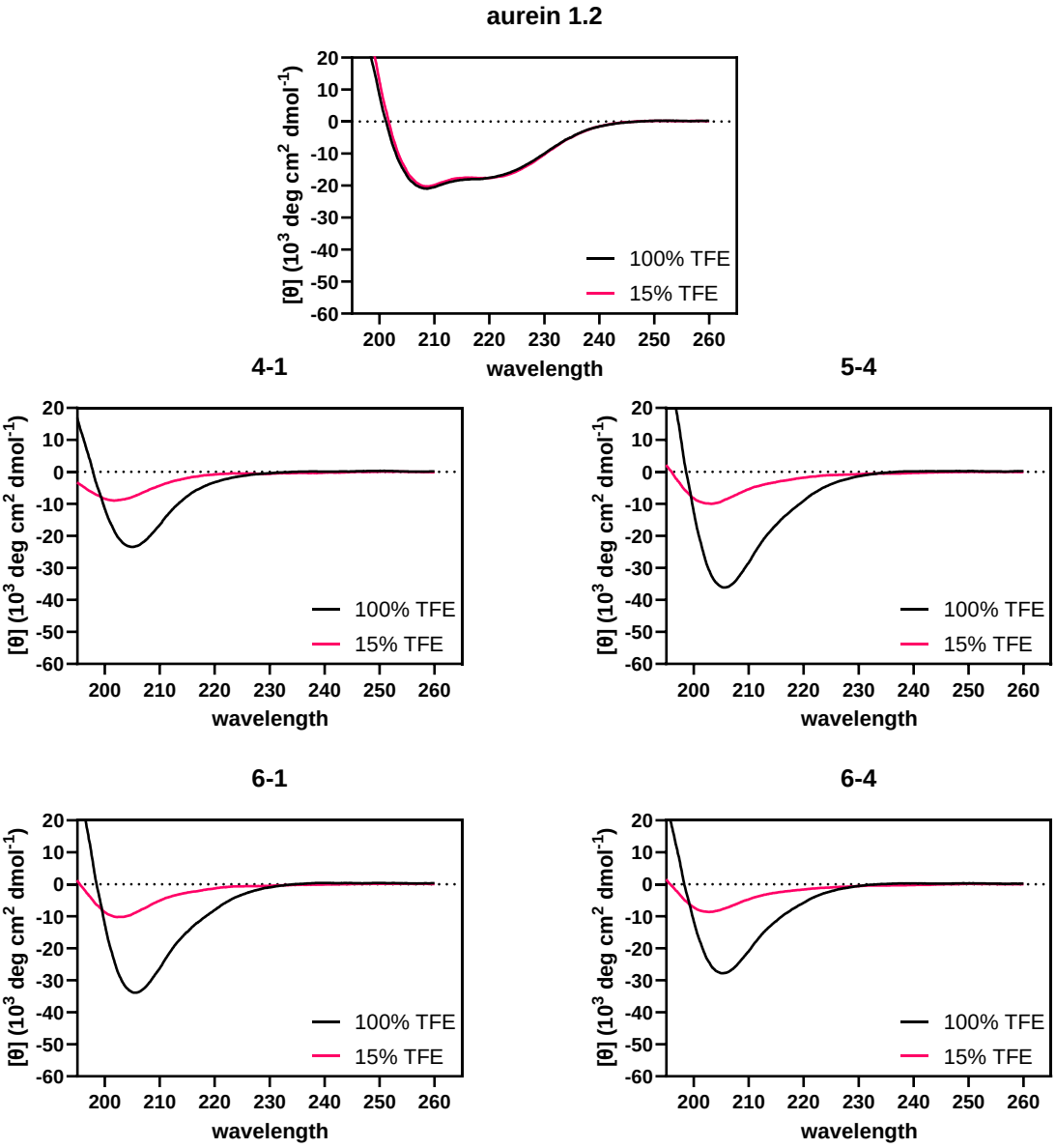


Figure S23. Circular dichroism spectra of high-SI peptides in comparison to aurein 1.2. CD spectra are shown in 15% (red) and 100% trifluoroethanol (black). The molar ellipticity values in each solvent system at 206 nm (α/β-peptides) or 222 nm (α-peptides) were quantified in each solvent system. Data was collected in triplicate and repeated in three independent experiments.

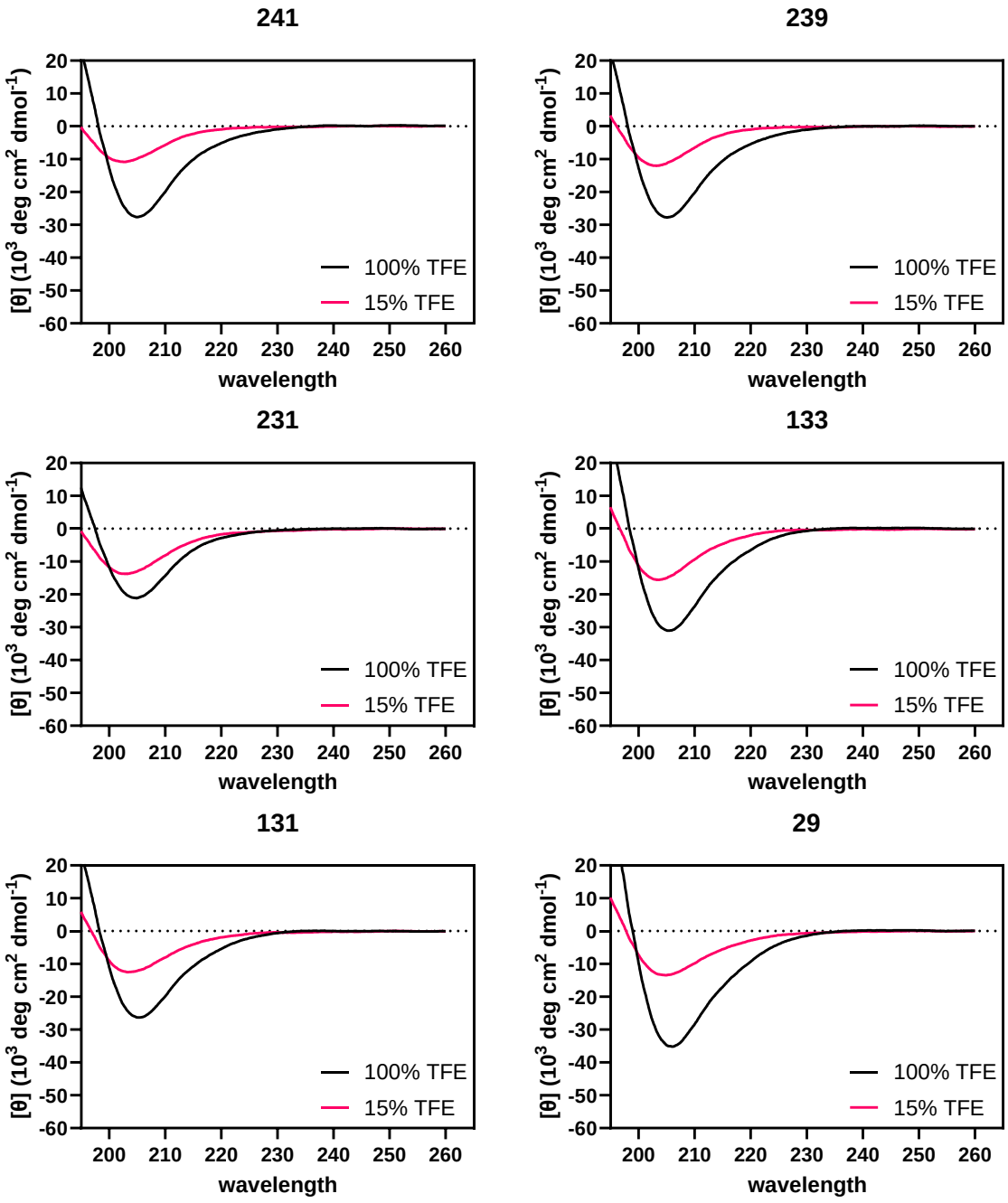


Figure S24. Circular dichroism spectra of training set peptides in 15% (red) and 100% trifluoroethanol (black). The molar ellipticity values in each solvent system at 206 nm (α/β-peptides) or 222 nm (α-peptides) were quantified in each solvent system. Data was collected in triplicate and repeated in three independent experiments.

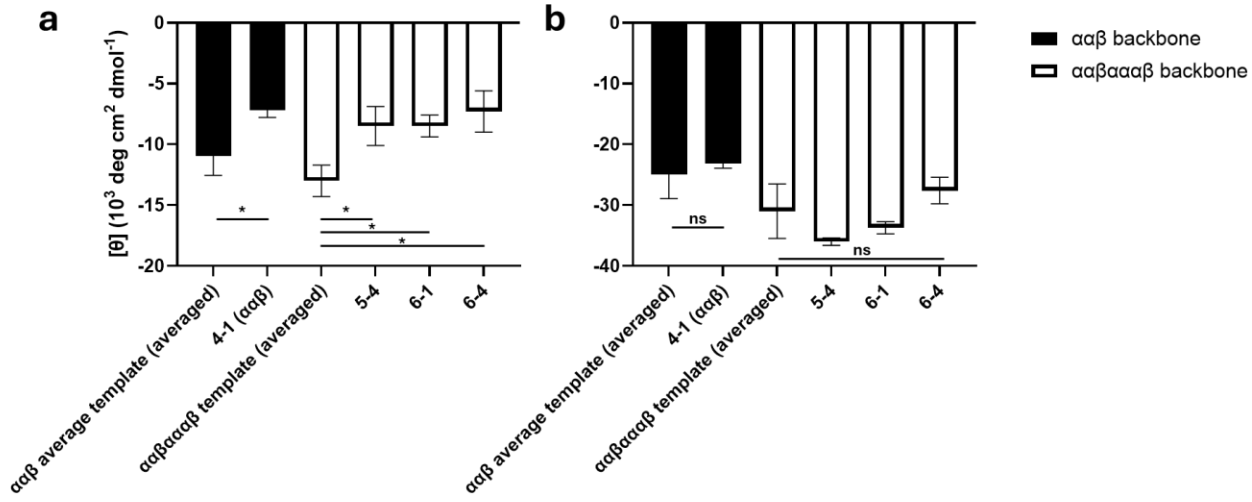


Figure S25. Comparison of molar ellipticities of top SI peptides with averages of those of corresponding templates. "Averaged" denotes the average of template peptides (ααβ: #241, 239, 231; ααβαααβ: #133, 131, 29) in (a) 15% TFE and (b) 100% TFE. Error bars represent standard deviations of three or more independent experiments, with each done in triplicate. Statistical analysis was conducted by one-way ANOVA with Tukey's multiple comparisons test. Statistical significance is represented as asterisks (*; $p < 0.05$) or 'not significant' ('ns'; $p > 0.05$).

S14: MIC assays against other microbial cells and smaller-interval *C. albicans* MIC

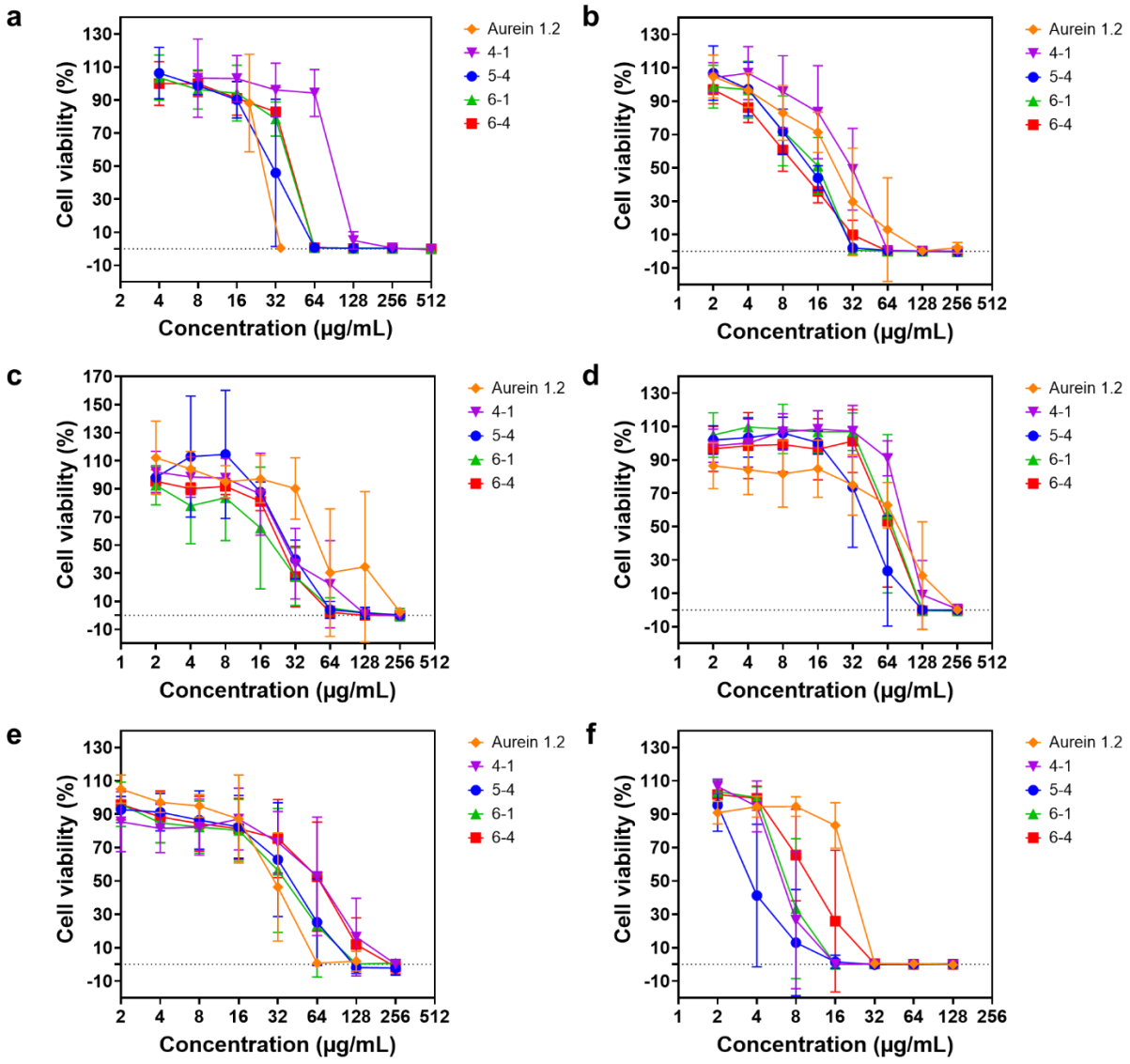


Figure S26. Broad-spectrum MIC curves against (a) *C. albicans*, (b) *C. tropicalis*, (c) *C. parapsilosis*, (d) *C. glabrata*, (e) *S. aureus*, and (f) *E. coli*. Fungal cell (5×10^3 cells/mL) and bacterial cell (1×10^6 cells/mL) solutions were incubated with 1:1 volume of compounds for 48 h and 24 h, respectively. Susceptibility was assessed using an XTT reduction assay to compare the absorbance at 490 nm according to CLSI MIC guidelines. Data points are the average of at least three independent experiments with two technical duplicates each and error bars represent the standard deviation.

Table S8. Antifungal and antibacterial activity of lead peptides. MIC values are the lowest assayed concentration of compound that resulted in a decrease in normalized absorbance of at least 90% of the mean. *C.a. = C. albicans; C. t. = C. tropicalis; C. p. = C. parapsilosis; C. g. = C. glabrata; S. a. = S. aureus; E. c. = E. coli*

	Antifungal MIC				Antibacterial MIC	
Peptide idx	***C. a.***	***C. t.***	***C. p.***	***C. g.***	***S. a.***	***E. c.***
Aurein 1.2	32	128	256	256	64	32
4-1	128	64	128	256	256	16
5-4	64	32	64	128	128	16
6-1	64	32	64	128	128	16
6-4	64	64	64	128	256	32

Table S9. Selectivity indices of lead peptides. SI values are calculated as the ratio of hemolysis over the antimicrobial activity (HC_{10} / MIC). The hemolysis values used are reported (as $\log_2 HC_{10}$) in Figure 5a. *C.a. = C. albicans; C. t. = C. tropicalis; C. p. = C. parapsilosis; C. g. = C. glabrata; S. a. = S. aureus; E. c. = E. coli*

	Antifungal SI				Antibacterial SI	
Peptide idx	***C. a.***	***C. t.***	***C. p.***	***C. g.***	***S. a.***	***E. c.***
Aurein 1.2	1.1	0.3	0.1	0.1	0.5	1.1
4-1	27.5	54.9	27.5	13.7	13.7	219.8
5-4	34.0	68.1	34.0	17.0	17.0	136.1
6-1	33.6	67.1	33.6	16.8	16.8	134.2
6-4	57.1	57.1	57.1	28.6	14.3	114.2

a

Peptide idx	sMIC	sSI
Aurein 1.2	30	1.2
4-1	80	44.0
5-4	50	43.6
6-1	50	43.0
6-4	60	60.9

b

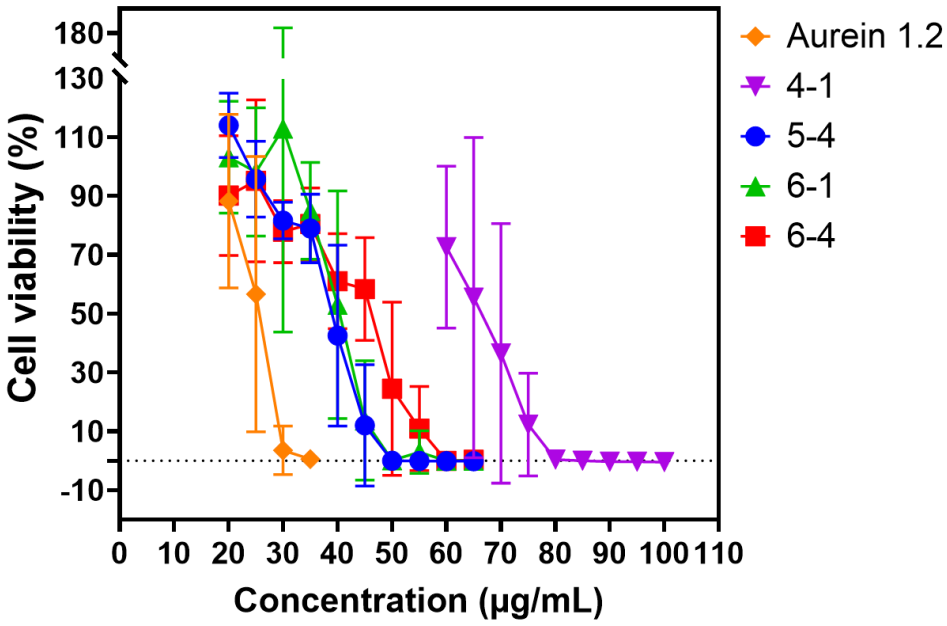


Figure S27. Smaller-interval (a) MIC (sMIC) and SI (sSI) values and (b) cell viability curves for *C. albicans* cells using peptides generated over 6 iterative rounds. Fungal cells (5×10^3 cells/mL) were incubated with compounds for 48 h and susceptibility was assessed using an XTT reduction assay to compare the absorbance at 490 nm for compound-treated samples and untreated samples. (a) "SI" indicates *C. albicans* selectivity over hemolysis (MIC/HC_{10}). (b) Data points are the average of at least three independent experiments with three technical duplicates each and error bars represent the standard deviation.

References

1. "RDKit: Open-source cheminformatics.

2. J. Gasteiger and M. Marsili, Iterative partial equalization of orbital electronegativity—a rapid access to atomic charges, *Tetrahedron*, 1980, **36**, 3219-3228.

3. L. B. Kier and L. H. Hall, An Electrotopological-State Index for Atoms in Molecules, *Pharmaceutical Research*, 1990, **7**, 801-807.

4. L. H. Hall and L. B. Kier, in *Reviews in Computational Chemistry*, 1991, pp. 367-422.

5. S. Riniker and G. A. Landrum, Similarity maps - a visualization strategy for molecular fingerprints and machine-learning methods, *Journal of Cheminformatics*, 2013, **5**, 43.

6. G. R. Bickerton, G. V. Paolini, J. Besnard, S. Muresan and A. L. Hopkins, Quantifying the chemical beauty of drugs, *Nature Chemistry*, 2012, **4**, 90-98.

7. P. Király, R. Kiss, D. Kovács, A. Ballaj and G. Tóth, The Relevance of Goodness-of-fit, Robustness and Prediction Validation Categories of OECD-QSAR Principles with Respect to Sample Size and Model Type, *Molecular Informatics*, 2022, **41**, 2200072.

8. C. Rücker, G. Rücker and M. Meringer, y-Randomization and Its Variants in QSPR/QSAR, *Journal of Chemical Information and Modeling*, 2007, **47**, 2345-2357.

9. H. T. Hsueh, R. T. Chou, U. Rai, W. Liyanage, Y. C. Kim, M. B. Appell, J. Pejavar, K. T. Leo, C. Davison, P. Kolodziejski, A. Mozzer, H. Kwon, M. Sista, N. M. Anders, A. Hemingway, S. V. K. Rompicharla, M. Edwards, I. Pitha, J. Hanes, M. P. Cummings and L. M. Ensign, Machine learning-driven multifunctional peptide engineering for sustained ocular drug delivery, *Nat Commun*, 2023, **14**, 2509.

10. S. Amrhein, S. A. Oelmeier, F. Dismer and J. Hubbuch, Molecular Dynamics Simulations Approach for the Characterization of Peptides with Respect to Hydrophobicity, *The Journal of Physical Chemistry B*, 2014, **118**, 1707-1714.

11. M. A. Cherry, S. K. Higgins, H. Melroy, H.-S. Lee and A. Pokorny, Peptides with the Same Composition, Hydrophobicity, and Hydrophobic Moment Bind to Phospholipid Bilayers with Different Affinities, *The Journal of Physical Chemistry B*, 2014, **118**, 12462-12470.

12. D. H. Chang, M.-R. Lee, N. Wang, D. M. Lynn and S. P. Palecek, Establishing Quantifiable Guidelines for Antimicrobial α/β-Peptide Design: A Partial Least-Squares Approach to Improve Antimicrobial Activity and Reduce Mammalian Cell Toxicity, *ACS Infectious Diseases,* 2023, **9**, 2632-2651.

13. M. R. Lee, N. Raman, S. H. Gellman, D. M. Lynn and S. P. Palecek, Incorporation of β-Amino Acids Enhances the Antifungal Activity and Selectivity of the Helical Antimicrobial Peptide Aurein 1.2, *ACS Chemical Biology*, 2017, **12**, 2975-2980.

14. A. J. Karlsson, W. C. Pomerantz, K. J. Neilsen, S. H. Gellman and S. P. Palecek, Effect of sequence and structural properties on 14-helical β-peptide activity against *Candida albicans* planktonic cells and biofilms, *ACS Chemical Biology*, 2009, **4**, 567-579.

15. M. R. Lee, N. Raman, S. H. Gellman, D. M. Lynn and S. P. Palecek, Hydrophobicity and helicity regulate the antifungal activity of 14-helical β-peptides, *ACS Chemical Biology*, 2014, **9**, 1613-1621.

16. A. N. Loes, R. A. L. Tarabi, J. Huddleston, L. Touyon, S. S. Wong, S. M. S. Cheng, N. H. L. Leung, W. W. Hannon, T. Bedford, S. Cobey, B. J. Cowling and J. D. Bloom, High-throughput sequencing-based neutralization assay reveals how repeated vaccinations impact titers to recent human H1N1 influenza strains, *bioRxiv*, 2024, 2024.2003.2008.584176.

Appendix

A1. Training set peptide purity and activity data

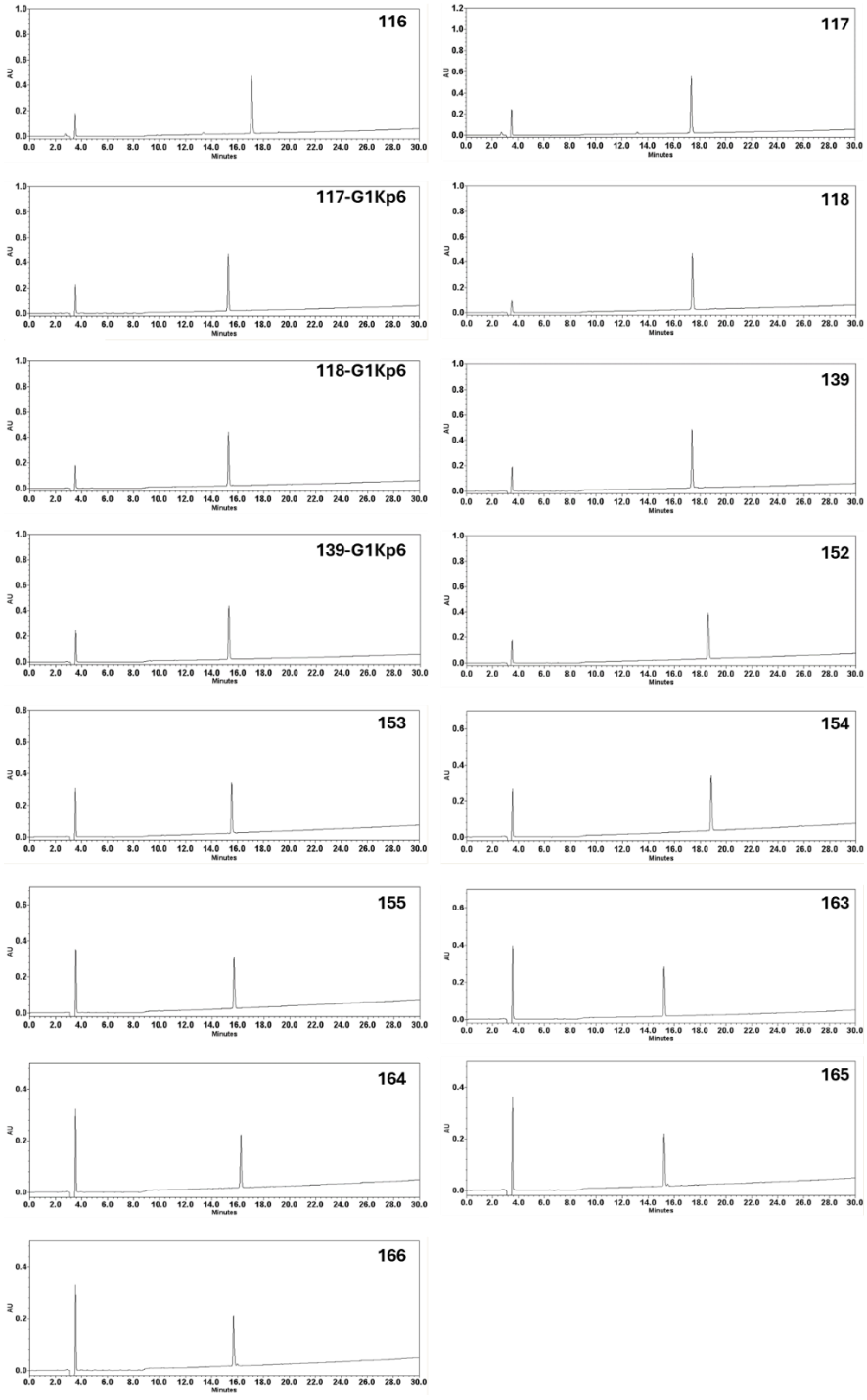


Figure S28. Analytical RP-HPLC profiles to measure retention time and purity of newly reported α/β-aurein 1.2 analogues in the training set.

Table S10. Mass spectrometry data for the subset of training set α/β-peptides that was newly introduced in this study and not previously characterized. Ionized mass spec data was obtained from multiple mass spec equipment (Bruker Impact II or Q Exactive Plus Orbitrap) and measured masses at different ionization states were converted to full mass values.

Round	Idx	MW. Calc	MW. Found	# of ionization	Total mass found
Training	#116	524.6681	524.6675	[M+3H]+3	1571.989
	#117	524.6681	524.6681	[M+3H]+3	1571.989
	#117_G1Kp6	822.0352	822.0324	[M+2H]+2	1643.057
	#118	407.7560	407.7560	[M+4H]+4	1628.001
	#118_G1Kp6	850.0414	850.0389	[M+2H]+2	1699.070
	#139	786.4985	786.5024	[M+2H]+2	1571.997
	#139_G1Kp6	822.0352	822.0387	[M+2H]+2	1643.070
	#152	790.5298	790.5280	[M+2H]+2	1580.048
	#153	826.0665	826.0649	[M+2H]+2	1651.122
	#154	818.5359	818.5336	[M+2H]+2	1636.059
	#155	569.7175	569.7167	[M+3H]+3	1707.137
	#163	539.0239	539.0239	[M+3H]+3	1615.056
	#164	548.3676	548.3678	[M+3H]+3	1643.087
	#165	546.3833	546.3837	[M+3H]+3	1637.134
	#166	541.7114	541.7116	[M+3H]+3	1623.119

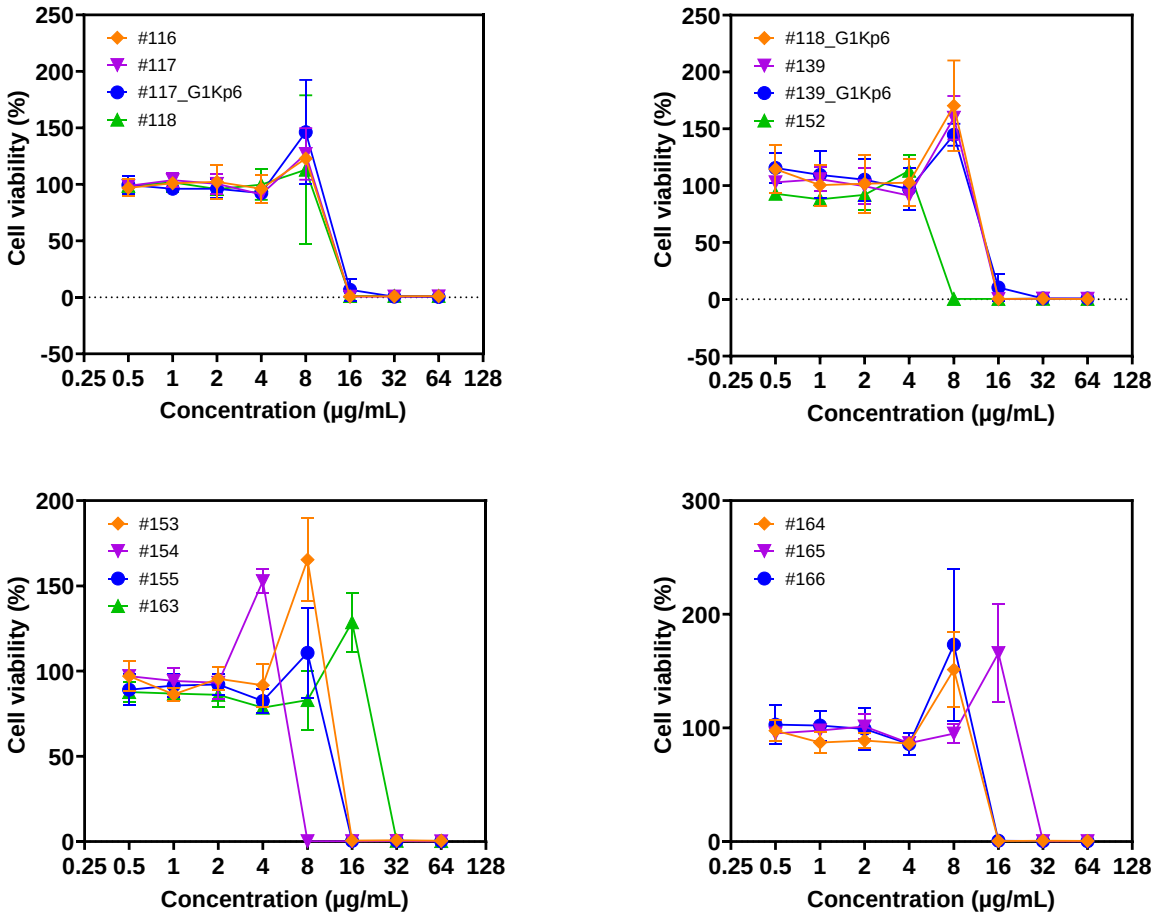


Figure S29. *C. albicans* MIC curves for the subset of training set α/β-peptides that was newly introduced in this study and not previously characterized. Fungal cells (5×10^3 cells/mL) were incubated with compounds for 48 h and susceptibility was assessed using an XTT reduction assay to compare the absorbance at 490 nm for compound-treated samples and untreated samples. Data points are the average of three independent experiments with three technical replicates each and error bars represent the standard deviation.

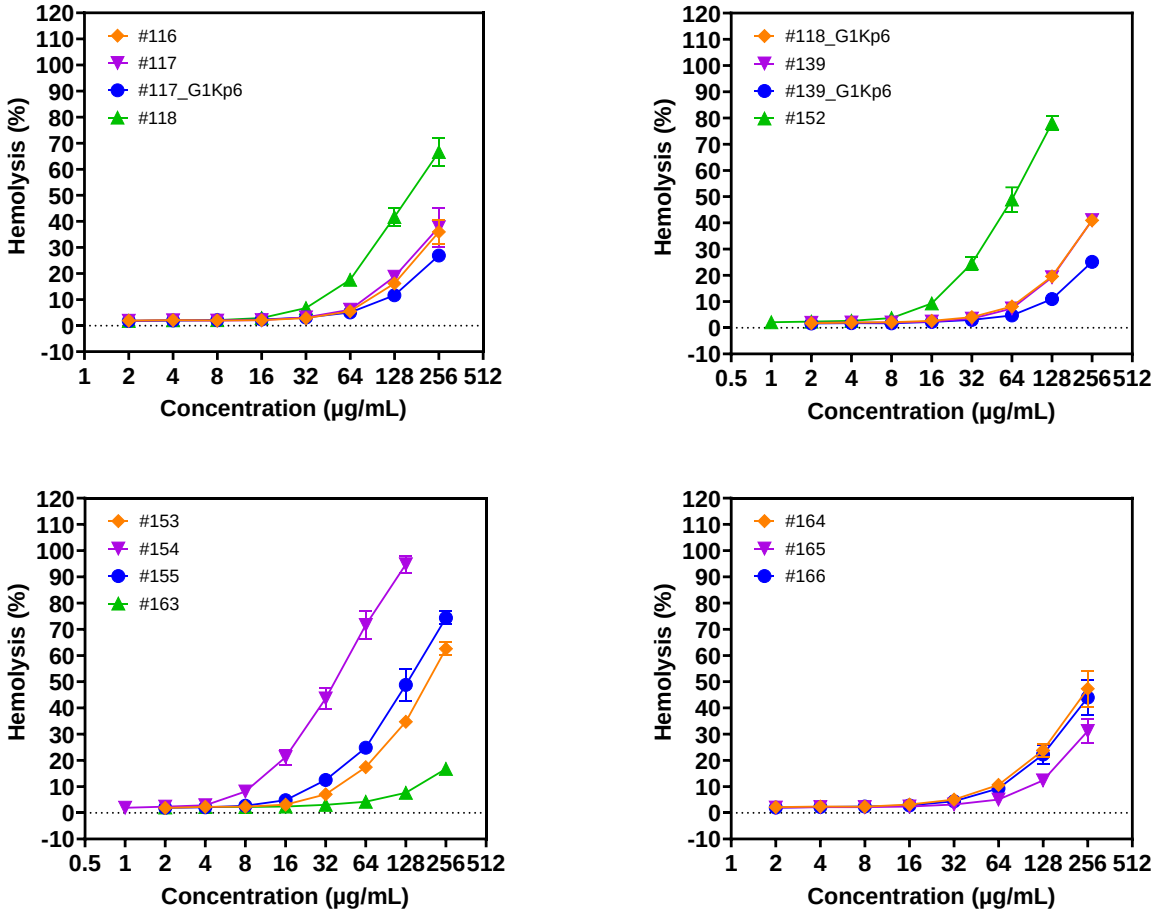


Figure S30. Hemolysis curves for the subset of training set α/β-peptides that was newly introduced in this study and not previously characterized. Peptides were incubated with human RBCs for 1 h, and the absorbance of the supernatant was measured at 405 nm and normalized to melittin-treated RBCs, corresponding to 100% hemolysis. Data points are the average of at least three independent experiments with two technical replicates each or more and error bars represent the standard deviation.

A2. Newly discovered peptide mass spectra and RP-HPLC curves

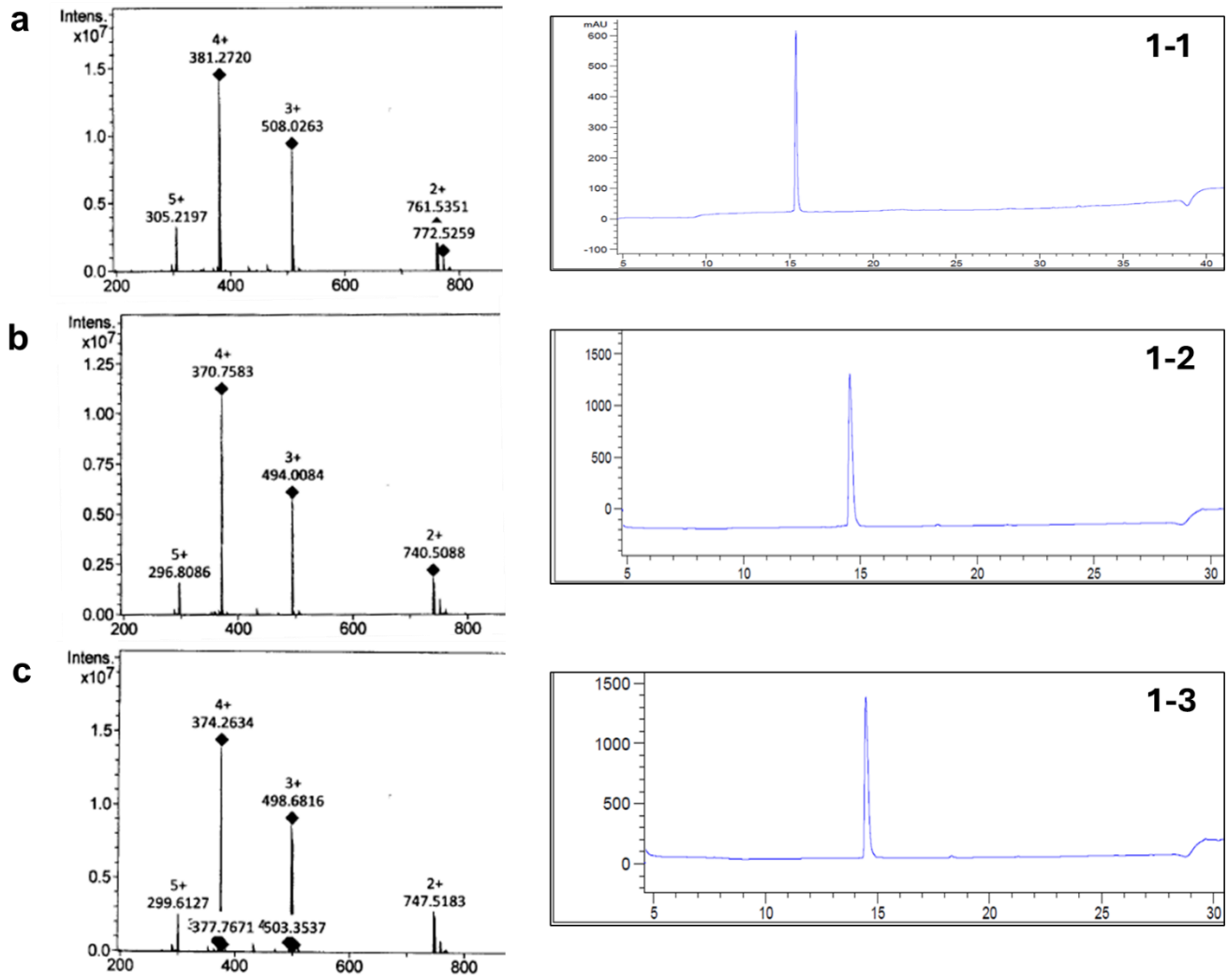


Figure S31. ESI+ mass spectra and analytical RP-HPLC profiles for each newly discovered peptide in round 1. (a) Peptide 1-1, (b) Peptide 1-2, (c) Peptide 1-3.

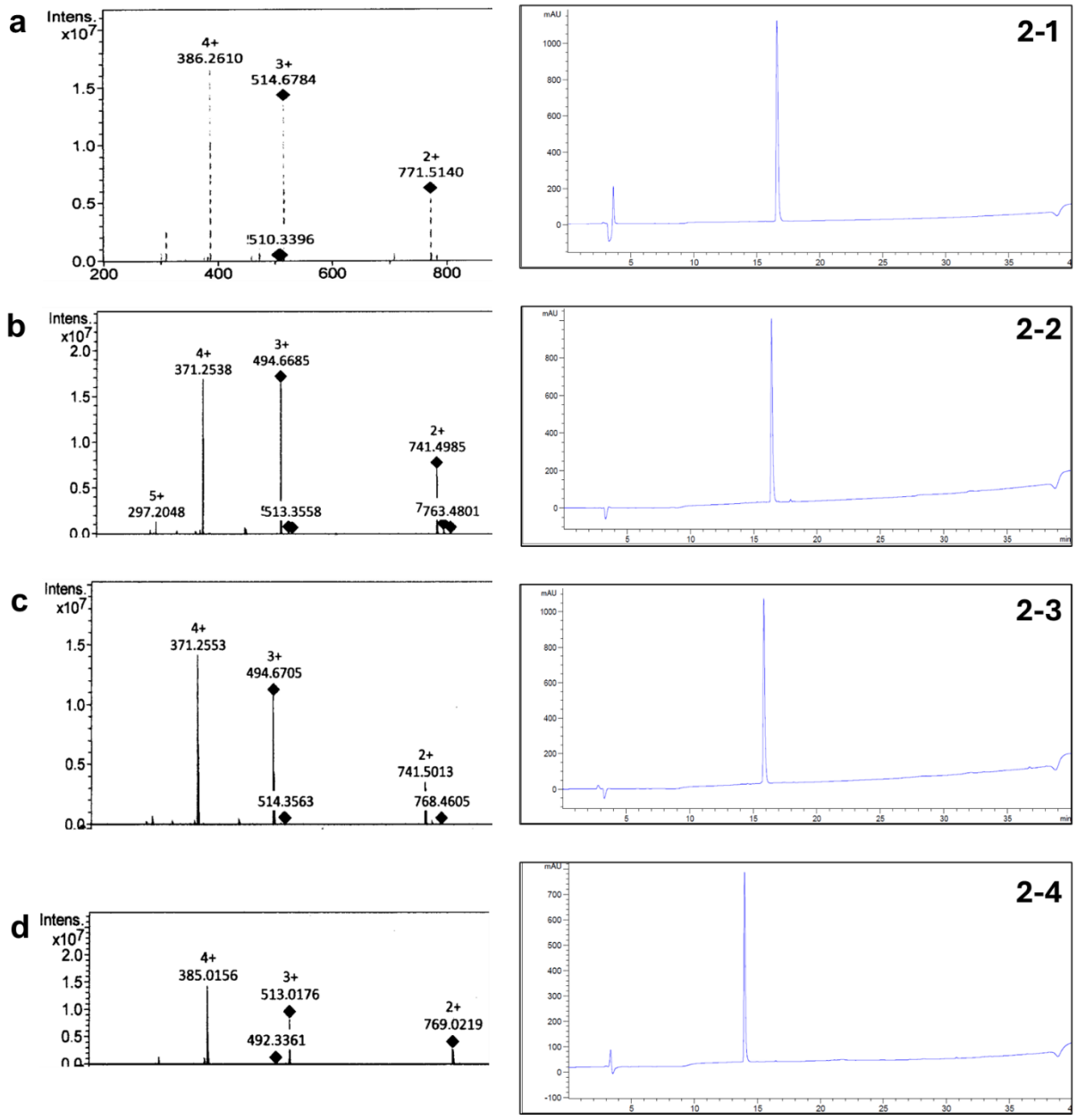


Figure S32. ESI+ mass spectra and analytical RP-HPLC profiles for each newly discovered peptide in round 2. The HPLC peak before 5 minutes is the solvent peak. (a) Peptide 2-1, (b) Peptide 2-2, (c) Peptide 2-3, (d) Peptide 2-4.

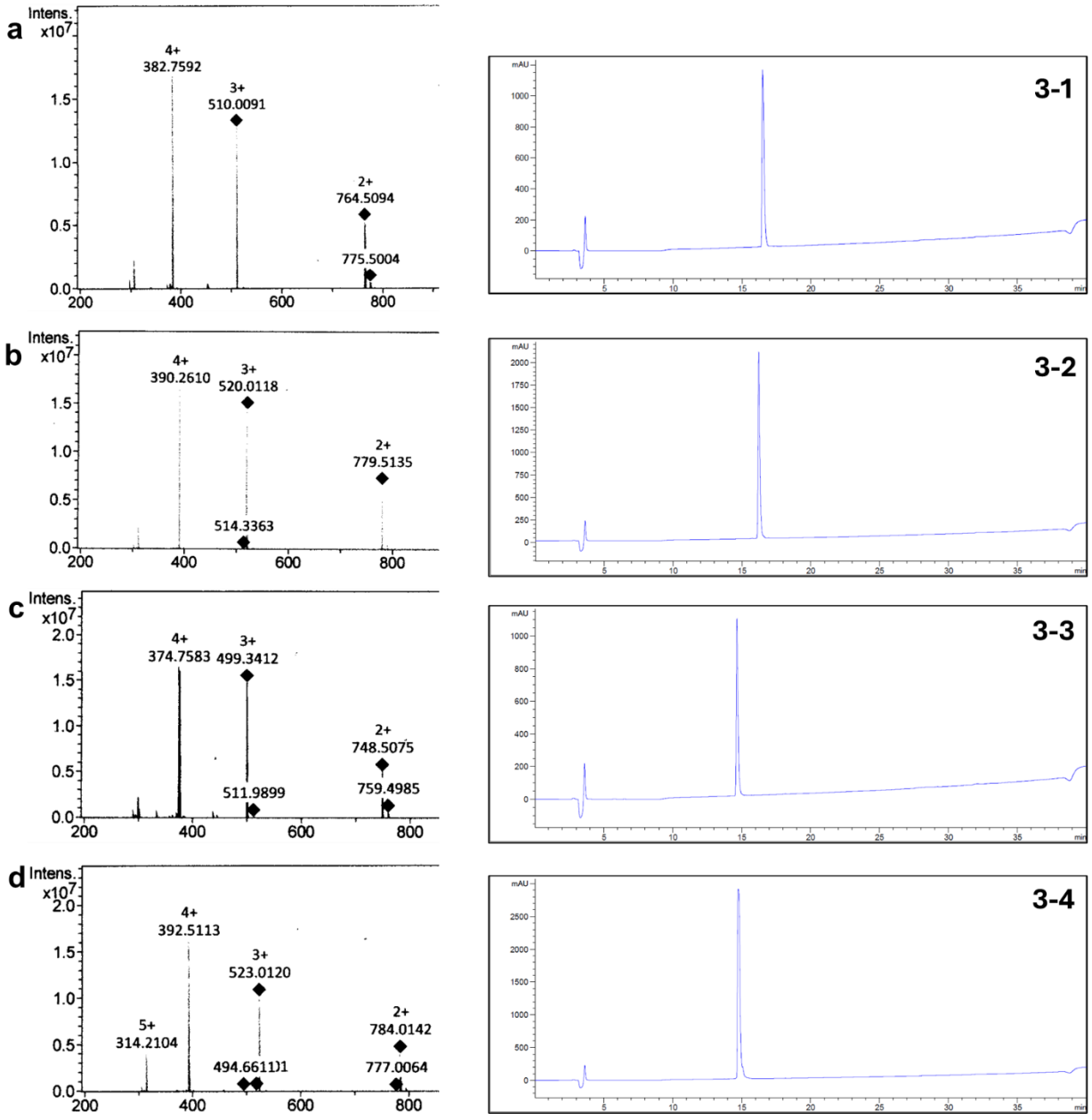


Figure S33. ESI+ mass spectra and analytical RP-HPLC profiles for each newly discovered peptide in round 3. The HPLC peak before 5 minutes is the solvent peak. (a) Peptide 3-1, (b) Peptide 3-2, (c) Peptide 3-3, (d) Peptide 3-4.

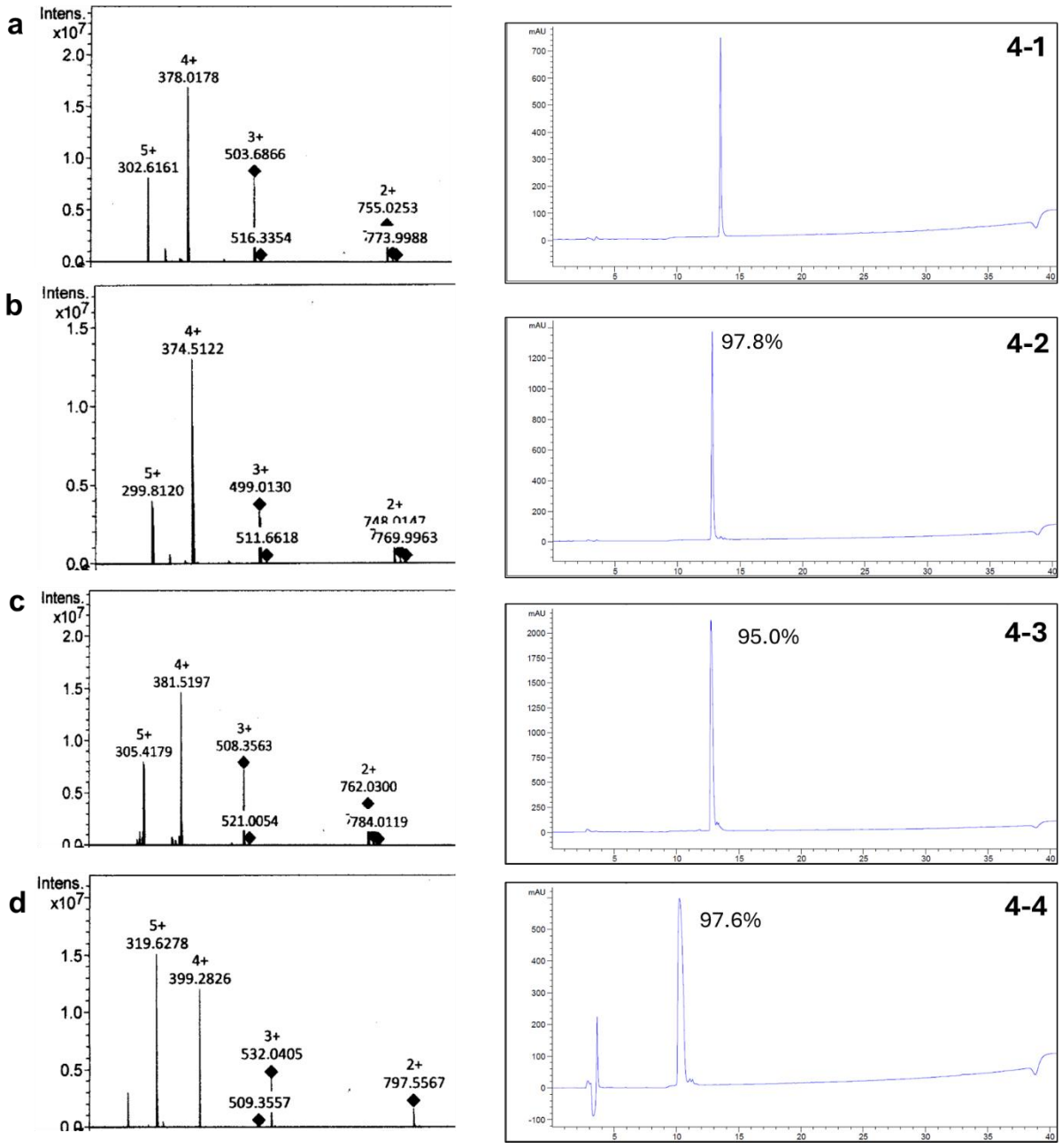


Figure S34. ESI+ mass spectra and analytical RP-HPLC profiles for each newly discovered peptide in round 4. The HPLC peak before 5 minutes is the solvent peak. (a) Peptide 4-1, (b) Peptide 4-2, (c) Peptide 4-3, (d) Peptide 4-4.

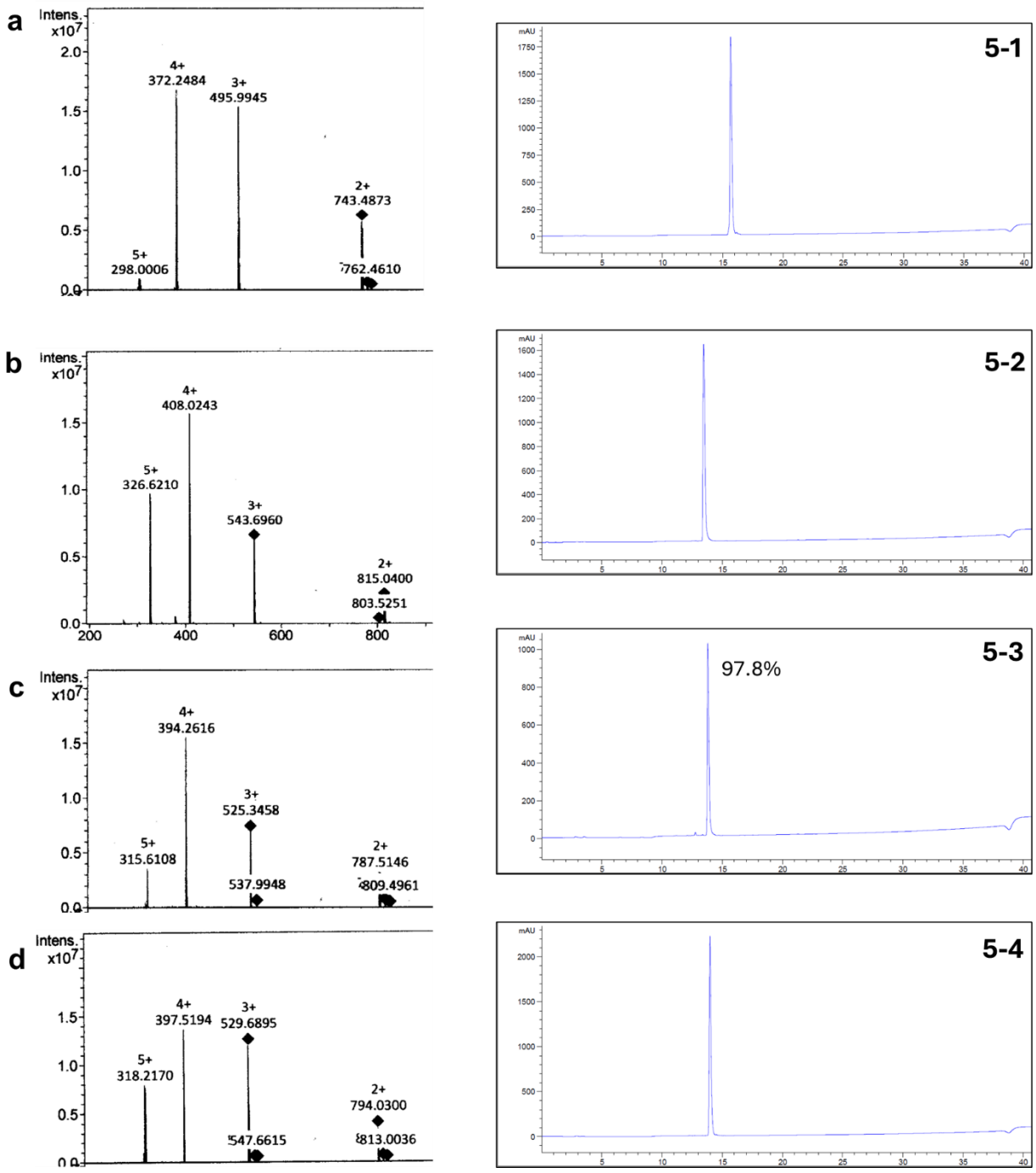


Figure S35. ESI+ mass spectra and analytical RP-HPLC profiles for each newly discovered peptide in round 5. The HPLC peak before 5 minutes is the solvent peak. (a) Peptide 5-1, (b) Peptide 5-2, (c) Peptide 5-3, (d) Peptide 5-4.

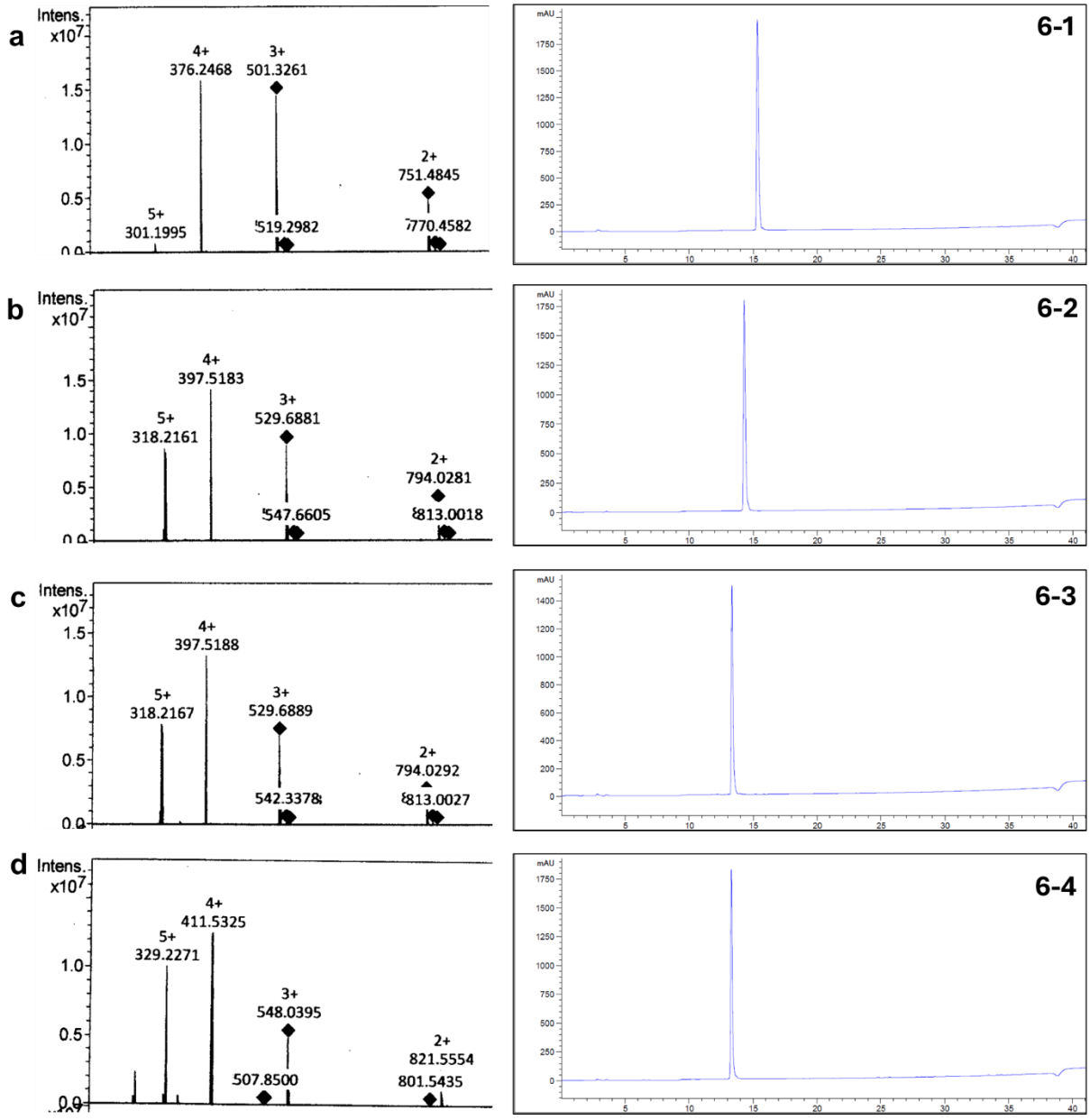


Figure S36. ESI+ mass spectra and analytical RP-HPLC profiles for each newly discovered peptide in round 6. The HPLC peak before 5 minutes is the solvent peak. (a) Peptide 6-1, (b) Peptide 6-2, (c) Peptide 6-3, (d) Peptide 6-4.