Supplementary Information for

Deep Learning to Design Nuclear-Targeting Abiotic Miniproteins

Authors: Carly K. Schissel[1†], Somesh Mohapatra[2†], Justin M. Wolfe[1#], Colin M. Fadzen[1‡], Kamela Bellovoda[3], Chia-Ling Wu[3], Jenna A. Wood[3], Annika B. Malmberg[3], Andrei Loas[1], Rafael Gómez-Bombarelli[2*], Bradley L. Pentelute[1,4,5,6*]

Affiliations:

[1]Massachusetts Institute of Technology, Department of Chemistry, 77 Massachusetts Avenue, Cambridge, MA 02139, USA

[2]Massachusetts Institute of Technology, Department of Materials Science and Engineering, 77 Massachusetts Avenue, Cambridge, MA 02139, USA

[3]Sarepta Therapeutics, 215 First Street, Cambridge, MA 02142, USA

[4]The Koch Institute for Integrative Cancer Research, Massachusetts Institute of Technology, 500 Main Street, Cambridge, MA 02142, USA

[5]Center for Environmental Health Sciences, Massachusetts Institute of Technology, 77 Massachusetts Avenue, Cambridge, MA 02139, USA

[6]Broad Institute of MIT and Harvard, 415 Main Street, Cambridge, MA 02142, USA

†These authors contributed equally to this work.

#Current address: Ra Pharmaceuticals, 87 Cambridgepark Drive, Cambridge, MA 02140, USA

‡Current address: Harvard Medical School, 25 Shattuck Street, Boston, MA 02115, USA

*Correspondence to: blp@mit.edu, rafagb@mit.edu

Table of Contents

1 Materials and General Methods

1.1 Reagents and Solvents

H-Rink Amide-ChemMatrix resin was obtained from PCAS BioMatrix Inc. (St-Jean-sur-Richelieu, Quebec, Canada). 1-[Bis(dimethylamino)methylene]-1*H*-1,2,3-triazolo[4,5-b]pyridinium-3-oxid-hexafluorophosphate (HATU), 4-pentynoic acid, 5-azidopentanoic acid, Fmoc-β-Ala-OH, Fmoc-6-aminohexanoic acid, and Fmoc-L-Lys(N_3) were purchased from Chem-Impex International (Wood Dale, IL). PyAOP was purchased from P3 BioSystems (Louisville, KY). Fmoc-protected amino acids (Fmoc-Ala-OHxH_2O, Fmoc-Arg(Pbf)-OH; Fmoc-Asn(Trt)-OH; Fmoc-Asp-(O*t*Bu)-OH; Fmoc-Cys(Trt)-OH; Fmoc-Gln(Trt)-OH; Fmoc-Glu(O*t*Bu)-OH; Fmoc-Gly-OH; Fmoc-His(Trt)-OH; Fmoc-Ile-OH; Fmoc-Leu-OH; Fmoc-Lys(Boc)-OH; Fmoc-Met-OH; Fmoc-Phe-OH; Fmoc-Pro-OH; Fmoc-Ser(But)-OH; Fmoc-Thr(*t*Bu)-OH; Fmoc-Trp(Boc)-OH; Fmoc-Tyr(*t*Bu)-OH; Fmoc-Val-OH), were purchased from the Novabiochem-line from Sigma Millipore. Peptide synthesis-grade *N*,*N*-dimethylformamide (DMF), CH_2Cl_2, diethyl ether, and HPLC-grade acetonitrile were obtained from VWR International (Radnor, PA). All other reagents were purchased from Sigma-Aldrich (St. Louis, MO). Milli-Q water was used exclusively.

1.2 Liquid-chromatography mass-spectrometry

LCMS analyses were performed on either an Agilent 6520 Accurate-Mass Q-TOF LCMS (abbreviated as 6520) or an Agilent 6550 iFunnel Q-TOF LCMS system (abbreviated as 6550) coupled to an Agilent 1260 Infinity HPLC system. Mobile phases were: 0.1% formic acid in water (solvent A) and 0.1% formic acid in acetonitrile (solvent B). The following LCMS methods were used for characterization:

Method A: 1-61% B over 9 min, Zorbax C3 column (6520)

LC: Zorbax 300SB-C3 column: 2.1 × 150 mm, 5 µm, column temperature: 40 °C, gradient: 0-2 min 1% B, 2-11 min 1-61% B, 11-12 min 61-95% B, 12-15 min 95% B; flow rate: 0.8 mL/min.

MS: Positive electrospray ionization (ESI) extended dynamic range mode in mass range 300–3000 *m*/*z*. MS is on from 4 to 11 min.

Method B: 1-61% B over 10 min, Phenomenex Jupiter C4 column (6550)

LC: Phenomenex Jupiter C4 column: 1.0 × 150 mm, 5 μm, column temperature: 40 °C, gradient: 0-2 min 1% B, 2-12 min 1-61% B, 12-16 min 61-90% B; 16-20 min 90% B; flow rate: 0.1 mL/min. MS: Positive electrospray ionization (ESI) extended dynamic range mode in mass range 100–1700 *m/z*. MS is on from 4 to 12 min.

Method C: 1-61% B over 10 min, Agilent EclipsePlus C18 column (6550)

LC: Agilent EclipsePlus C18 RRHD column: 2.1 × 50 mm, 1.8 μm, column temperature: 40 °C, gradient: 0-2 min 1% B, 2-12 min, 1-61% B, 12-13 min, 61% B, 13-16 min, 1% B; flow rate: 0.1 mL/min.

MS: Positive electrospray ionization (ESI) extended dynamic range mode in mass range 300–3000 m/z. MS is on from 4 to 12 min. This method was used exclusively for characterization of the modular library.

All data were processed using Agilent MassHunter software package. Y-axis in all chromatograms shown represents total ion current (TIC) unless noted.

1.3 General method for peptide preparation

Fast-flow Peptide Synthesis: Peptides were synthesized on a 0.1 mmol scale using an automated fast-flow peptide synthesizer. A 100 mg portion of ChemMatrix Rink Amide HYR resin was loaded into a reactor maintained at 90 ºC. All reagents were flowed at 40 mL/min with HPLC pumps through a stainless-steel loop maintained at 90 ºC before introduction into the reactor. For each coupling, 10 mL of a solution containing 0.4 M amino acid and 0.38 M HATU in DMF were mixed with 600 μL diisopropylethylamine and delivered to the reactor. Fmoc removal was accomplished using 10.4 mL of 20% (v/v) piperidine. Between each step, DMF (15 mL) was used to wash out the reactor. For peptides in the modular library, special coupling conditions were used for arginine, in which 10 mL of a solution containing 0.4 M Fmoc-L-Arg(Pbf)-OH and 0.38 M PyAOP in DMF were mixed with 600 μL diisopropylethylamine and delivered to the reactor. For Mach peptides, additional special coupling conditions were used according to the optimized peptide synthesis protocol.[1] To couple unnatural amino acids or to cap the peptide (e.g. with 4-pentynoic acid), the resin was incubated for 30 min at room temperature with amino acid (1 mmol) dissolved in 2.5 mL 0.4 M HATU in DMF with 500 μL diisopropylethylamine. After completion of the synthesis, the resin was washed 3 times with dichloromethane and dried under vacuum.

Peptide Cleavage and Deprotection: Each peptide was subjected to simultaneous global side-chain deprotection and cleavage from resin by treatment with 5 mL of 94% trifluoroacetic acid (TFA), 2.5% 1,2-ethanedithiol (EDT), 2.5% water, and 1% triisopropylsilane (TIPS) (v/v) for 7 min at 60 °C or at room temperature for 2 to 4 hours. For arginine-rich sequences, the resin was treated with a cleavage cocktail consisting of 82.5% TFA, 5% phenol, 5% thioanisole, 5% water, and 2.5% EDT (v/v) for 14 hours at room temperature. For peptides containing azide, EDT was substituted for thioanisole. The cleavage cocktail was first concentrated by bubbling N_2 through the mixture, and cleaved peptide was precipitated and triturated with 40 mL of cold ether (chilled in dry ice). The crude product was pelleted by centrifugation for three minutes at 4,000 rpm and the ether was decanted. This wash step was repeated two more times. After the third wash, the pellet was dissolved in 50% water and 50% acetonitrile containing 0.1% TFA, filtered through a fritted syringe to remove the resin and lyophilized.

Peptide Purification: The peptides were dissolved in water and acetonitrile containing 0.1% TFA, filtered through a 0.22 µm nylon filter and purified by mass-directed semi-preparative reversed-phase HPLC. Solvent A was water with 0.1% TFA additive and Solvent B was acetonitrile with 0.1% TFA additive. A linear gradient that changed at a rate of 0.5% B/min was used. Most of the peptides were purified on an Agilent Zorbax SB C3 column: 9.4 x 250 mm, 5 µm. Extremely hydrophilic peptides, such as the arginine-rich sequences were purified on an Agilent Zorbax SB C18 column: 9.4 x 250 mm, 5 µm. Using mass data about each fraction from the instrument, only pure fractions were pooled and lyophilized. The purity of the fraction pool was confirmed by LC-MS.

Macrocyclization: Mach12 and Mach13 contained cysteine linked macrocycles. Purified unprotected peptide (1 mM) was dissolved in DMF with decafluorobiphenyl (2 mM) and DIEA (50 mM) and incubated at room temperature for 3 h. The solution was then diluted 100-fold in 1% acetonitrile, 2% TFA in water and purified directly by reverse-phase HPLC.

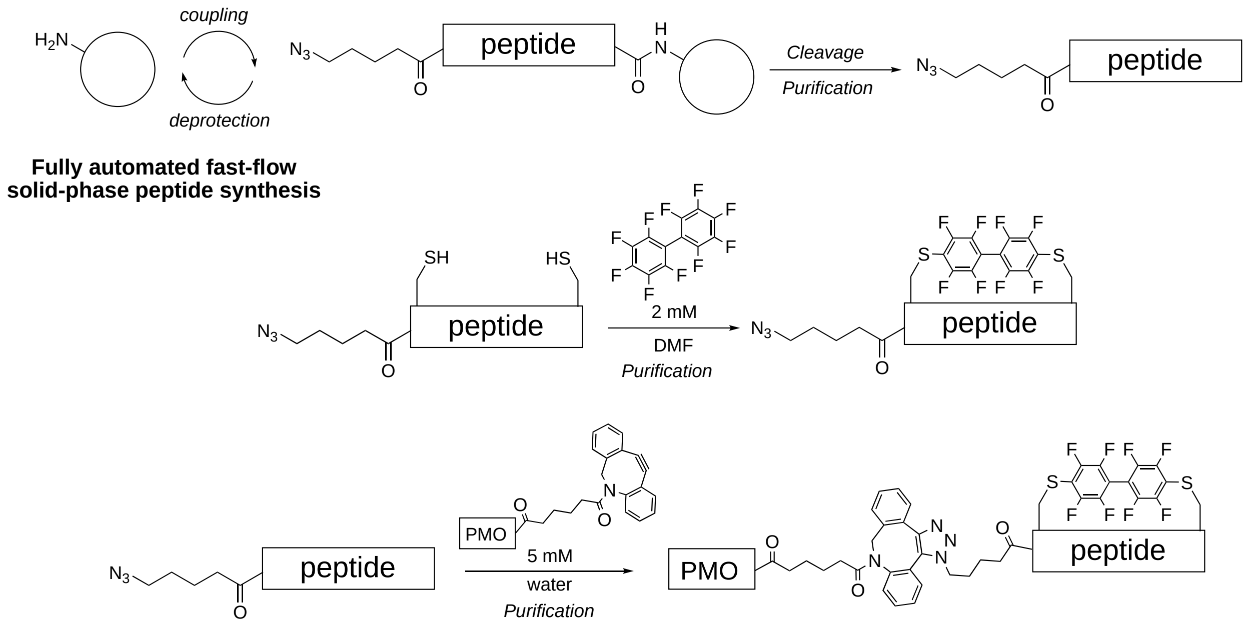


Figure 1. Synthesis route for Mach peptides. Predicted sequences were synthesized by fully automated SPPS, cyclized, and conjugated to PMO. Our synthesis technology can reliably synthesize long peptides in one-shot. If the predicted peptide contains a Cys macrocycle, cleaved and purified peptides were cyclized before being attached to PMO via copper-free click chemistry.

1.4 PMO-DBCO Synthesis

PMO IVS-654 (50 mg, 8 µmol) was dissolved in 150 µL DMSO. To the solution was added a solution containing 2 equivalents of dibenzocyclooctyne acid (5.3 mg, 16 µmol) activated with HBTU (37.5µL of 0.4 M HBTU in DMF, 15 µmol) and DIEA (2.8 µL, 16 µmol) in 40 µL DMF (Final reaction volume = 0.23 mL). The reaction proceeded for 25 min before being quenched with 1 mL of water and 2 mL of ammonium hydroxide. The ammonium hydroxide hydrolyzed any ester formed during the course of the reaction. After 1 hour, the solution was diluted to 40 mL in water/acetonitrile and purified using reverse-phase HPLC (Agilent Zorbax SB C3 column: 21.2 x 100 mm, 5 µm) and a linear gradient from 2 to 60% B (solvent A: water; solvent B: acetonitrile) over 58 min (1% B / min). Using mass data about each fraction from the instrument, only pure fractions were pooled and lyophilized. The purity of the fraction pool was confirmed by LC-MS.

2 Analysis and benchmarking of CNN model

2.1 Benchmarking against regression models

We benchmarked the fingerprint (FP) representation and convolutional neural network (CNN) model against other model architectures and one-hot representation-based models in their ability to predict activities of both library and Mach sequences (Figs. S2, S3; Tables S2, S3). 8 scikit-learn model architectures (ridge, lasso, stochastic gradient descent, gaussian process, random forest, support vector and gradient boosting regression) and extreme gradient boosting regression were evaluated.[2,3]

Relevant hyperparameters for every model were optimized with Bayesian search using scikit-optimize (Table S1).[4] The hyperparameter optimization was done by 3-fold cross-validation, with the objective function as minimization of average root mean-squared error (RMSE) on the randomly held out validation dataset. The metrics have been reported on the test dataset. The train-valid-test dataset split was 70:10:20.

Table 1. Hyperparameters and optimized hyperparameters for regression and classification model architectures, fingerprint and one-hot representations have been noted. The hyperparameters follow a notation – parameter, datatype, values. Datatypes are categorical, integer and real. For categorical datatype, the list of hyperparameters is noted, and for integer and real datatypes, minimum and maximum values are noted.

Model Architecture	Hyperparameters	Optimized parameters			
		Regression - Fingerprint	Regression - One Hot	Classification - Fingerprint	Classification - One Hot

Ridge	fit_intercept, Categorical, [True, False];
	normalize, Categorical, [True, False];
	alpha, Real, [1e-3, 1e3];
	solver, Categorical, ['svd', 'cholesky', 'lsqr', 'sparse_cg', 'sag', 'saga']
		alpha=546.999, fit_intercept=True, normalize=False, solver=sparse_cg	alpha=41.986, fit_intercept=True, normalize=False, solver=sag	alpha=118.608, fit_intercept=False, normalize=True, solver=svd	alpha=96.513, fit_intercept=False, normalize=True, solver=cholesky

Lasso	fit_intercept, Categorical, [True, False];
	normalize, Categorical, [True, False];
	alpha, Real, [1e-3, 1e3];
	precompute, Categorical, [True, False];
	selection, Categorical, [‘selection’, ‘random’]
		alpha=0.014, fit_intercept=False, normalize=True,precompute=False,selection=random	alpha=0.007, fit_intercept=True, normalize=False, precompute=True selection=random	-	-

SGD	loss, Categorical, ['squared_loss', 'epsilon_insensitive', 'huber', 'squared_epsilon_insensitive';
	penalty, Categorical; ['l1', 'l2', 'elasticnet'];
	alpha, Real, [1e-3, 1e3];
	fit_intercept, Categorical, [True, False];
	l1_ratio, Real, [1e-3, 1];
	learning_rate, Categorical, ['invscaling', 'constant', 'optimal', 'adaptive'];
	epsilon, Real, [1e-3, 1e3];
	eta0, Real, [1e-2, 10];
	power_t, Real, [1e-2, 10];
		alpha=0.002, average=True, epsilon=0.004, eta0=1.738, fit_intercept=True, l1_ratio=0.003, learning_rate=invscaling, loss=huber, penalty=l2, power_t=0.544	alpha=0.002, average=True, epsilon=0.004, eta0=1.738, fit_intercept=True, l1_ratio=0.003, learning_rate=invscaling, loss=huber, penalty=l2, power_t=0.544	alpha=0.001, average=True, epsilon=0.077, eta0=0.026, fit_intercept=False, l1_ratio=0.005, learning_rate=optimal, loss=log, penalty=l2, power_t=4.976	alpha=0.023, average=False, epsilon=0.009, eta0=0.01, fit_intercept=True, l1_ratio=0.1, learning_rate=adaptive, loss=hinge, penalty=l2, power_t=0.01

	average, Categorical, [True, False]				

Gaussian Process	kernel, Categorical, [Matern, RBF, DotProduct];
	alpha, Real, [1e-11, 1e-6];
	n_restarts_optimizer, Integer, [0, 10]
		alpha=1.57e-7, kernel=Matern, n_restarts_optimizers=8	alpha=7.320e-9, kernel=Matern, n_restarts_optimizer=10	kernel=Matern, n_restarts_optimizers=5	kernel=Matern, n_restarts_optimizers=5

Random Forest	criterion, Categorical, ['mse', 'mae']-Regression;
	criterion, Categorical, ['mse', 'entropy]-Classification
	n_estimators, Integer, [10, 1000];
	max_depth, Integer, [1, 10]
		criterion=’mse’, max_depth=7, n_estimators=626	criterion=’mse’, max_depth=6, n_estimators=127	criterion=entropy, max_depth=8, n_estimators=908	criterion=entropy, max_depth=10, n_estimators=630

XGBoost	gamma, Real, [1e-6, 10];
	eta,Real, [1e-3, 1];
	max_depth, Integer, [1, 10];
	tree_method, Categorical, ['auto', 'exact', 'approx', 'hist'];
	alpha, Real, [1e-3, 1e3];
	lambda, Real, [1e-3, 1e3];
	sketch_eps,Real, [1e-3, 1]
		alpha=0.035, eta=0.864, gamma=2.285, reg_lambda=3.945, sketch_eps=0.003, tree_method='exact'	eta=0.864, gamma=2.285, reg_alpha=0.035, reg_lambda=3.945, sketch_eps=0.003,tree_method=’exact'	alpha=2.407, eta=0.035, lambda=0.004, max_depth=5, tree_method=exact	alpha=0.835, eta=0.155, lambda=15.360, max_depth=4, tree_method=auto

Support Vector	kernel, Categorical, ['linear', 'poly', 'rbf', 'sigmoid'];
	degree, Integer, [1, 6];
	gamma, Real, [1e-6, 10];
	C, Real, [1e-2, 10],
	epsilon,Real, [1e-3, 10];
	shrinking, Categorical, [True, False]
		C=2.671, epsilon=0.009, gamma=0.002, kernel='linear'	C=0.024, epsilon=0.220, gamma=4.966e-5, kernel='linear'	C=0.030, degree=5, gamma=0.011, kernel=linear, shrinking=False	C=0.288, degree=5, gamma=0.076, kernel='poly', probability=True, shrinking=False

Gradient Boosting	loss, Categorical, ['ls', 'lad', 'huber', 'quantile'];
	learning_rate, Real, [1e-2, 1];
	n_estimators, Integer, [10, 1000];
	criterion, Categorical, ['friedman_mse', 'mse', 'mae'];
	max_depth, Integer, [1, 10]
		criterion='mse', learning_rate=0.018, loss='huber', max_depth=6, n_estimators=250	criterion='mse', learning_rate=0.018, loss='huber', max_depth=6, n_estimators=250	criterion=mse, learning_rate=0.018, loss=deviance, max_depth=6, n_estimators=250	criterion=mse, learning_rate=0.018, loss=deviance, max_depth=6, n_estimators=250

Nearest Neighbors	weights, Categorical, [‘uniform’, ‘distance’];
	leaf_size, Integer, [10, 100];
	n_neighbors, Integer, [2, 20];
	algorithm, Categorical, ['auto', 'ball_tree', 'kd_tree', 'brute'];
	p, Integer, [1, 5]
		algorithm=ball_tree, leaf_size=16, n_neighbors=7, p=5, weights=uniform	algorithm=brute, leaf_size=56, n_neighbors=4, p=3, weights=uniform	algorithm=auto, leaf_size=57, n_neighbors=4, p=4, weights=uniform	algorithm=auto, leaf_size=57, n_neighbors=4, p=4, weights=uniform

We evaluated the CNN-fingerprint (CNN-FP) models against individual models and model ensembles trained with FP and one-hot encoding representations.

RMSE and other metrics between the predicted and experimental activity values were used to compare individual models. For the validation dataset, random forest-FP (RF-FP) has lowest RMSE, but RF models cannot extrapolate outside the range of the training data, so they are limited toward the task of designing more active peptides. For the activity values of the Mach sequences, other models such as CNN one-hot had better RMSE, R^2, Pearson and Spearman's rank correlation coefficients. Based on the RMSE values, the best performing model architecture was one-hot based CNN. In practice, however, we observed that this model was not able to extrapolate activities beyond the training data, whereas the CNN-FP model could. We determined that RMSE and other metrics for the CNN-FP model are significantly affected by outlying predicted activity values. Upon removing the outlier (Mach12 with 140 predicted activity by CNN-FP model), we observed that the CNN-FP model outperforms all other models in terms of RMSE and R^2.

Unlike CNN-FP, none of the simpler models predict the activity of Mach peptides to be above the maximum of training dataset, as apparent in the parity plot and the RMSE and R^2 metrics for the Mach dataset without the outlier (Table S2). This experiment shows that simpler models are limited by the range of the training data, and are unable to extrapolate in the co-domain space. While other models may be able to produce sequences with high (>20-fold) experimental activity, the ability to extrapolate predicted activities is critical for the informed selection of predicted sequences to validate. Extrapolation is a necessary model feature for our goal of designing sequences with activities higher than those in the training set.

To mitigate the role of outliers that impact performance, we evaluated the use of model ensembling. Ensembled CNN one-hot model performed the best amongst all models on the validation dataset, while ensembled CNN-FP outperformed model-feature combinations on RMSE for Mach dataset, with and without the outlier (Table S3). Although the choice of sequences for experimental validation was not based on predictions from ensemble models, we note that ensemble models can robustly extrapolate predictions outside the training data for future studies.

From our analysis, we observe that simpler models can complement the CNN predictions in decision-making, such as ranking of predicted peptides, as noted from the high Pearson's and Spearman's correlation coefficients. The CNN model is necessary to be able to predict peptides with higher activity than the training set.

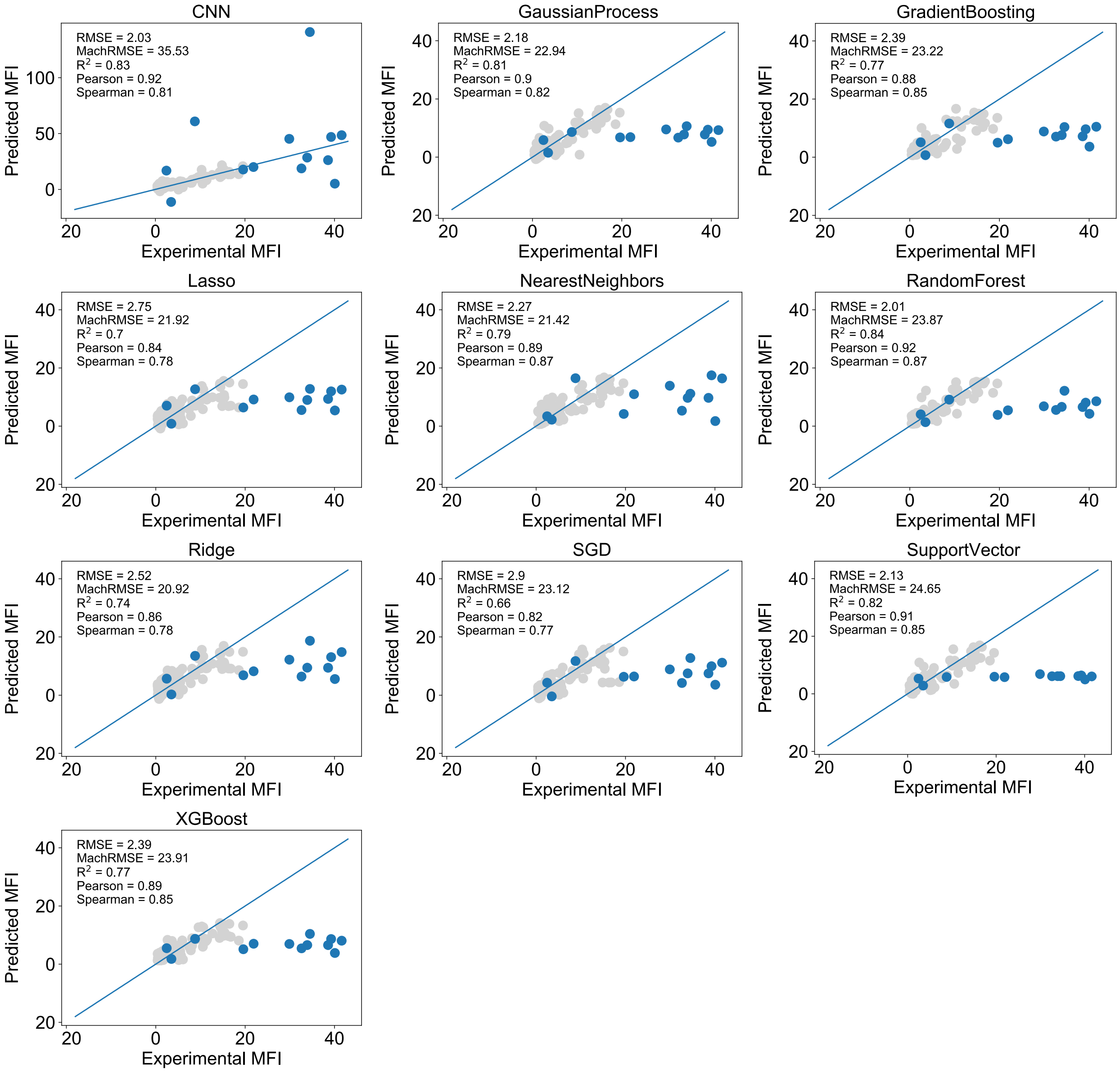


Figure 2. Parity plots for CNN and other models trained using 2048-bit fingerprints are shown. Random forest regressor performs best on the validation set, however is unable to extrapolate for Mach peptides. The second best, and the optimal model is support vector regression. The held-out data for validation of the model is shown in grey, and predictions for Mach peptides are shown in blue. Key evaluation metrics have been noted in the data inset. Only the CNN model shows a range of predicted values above the training data, as do the Mach peptides.

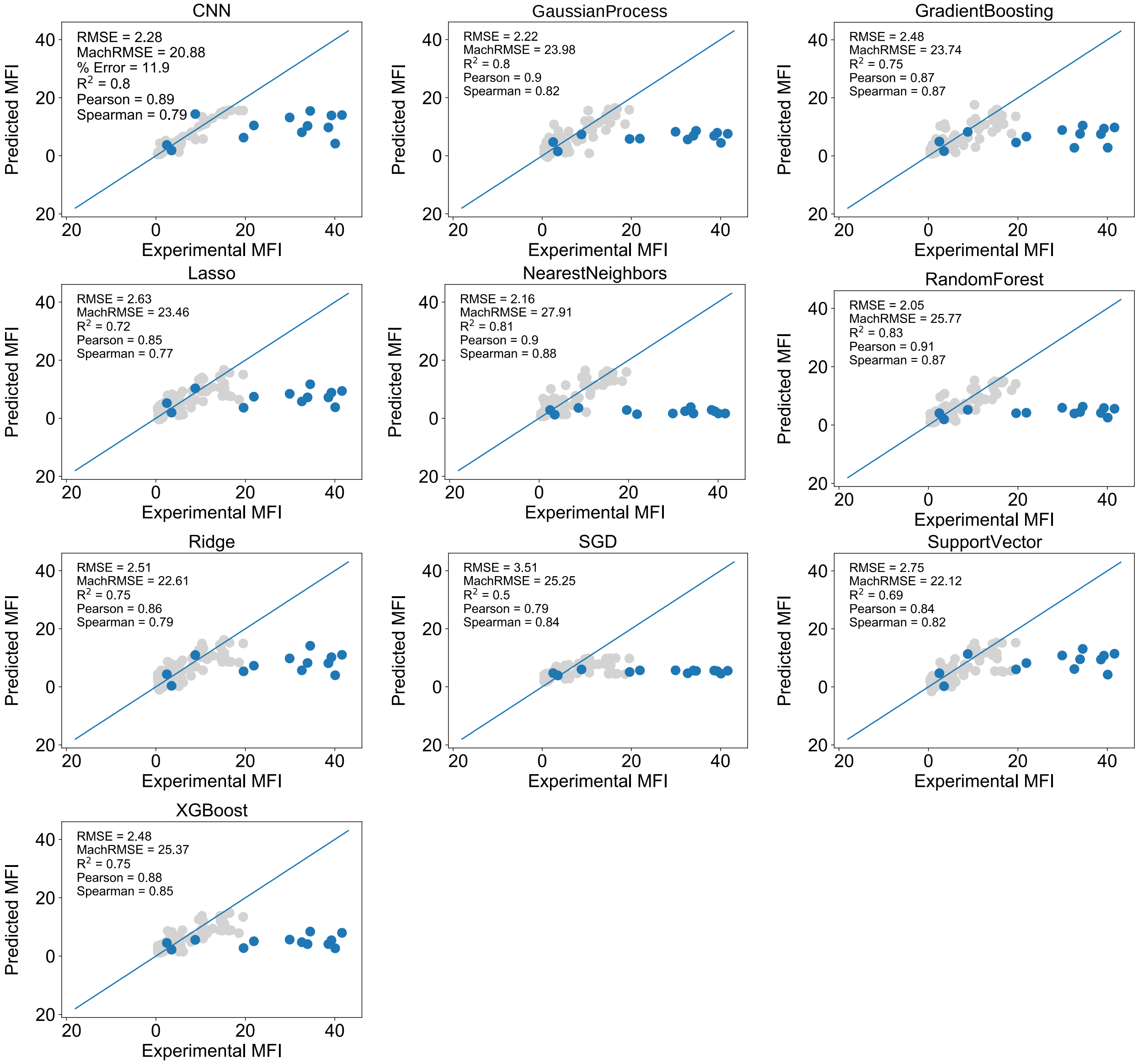


Figure 3. Parity plots for CNN and other models trained using one-hot encodings are shown. The held-out data for validation of the model is shown in grey, and predictions for Mach peptides are shown in blue. Key evaluation metrics have been noted in the data inset. Only the CNN model shows a range of predicted values above the training data, as do the Mach peptides.

Table 2. The CNN model with 2048-bit fingerprint features (CNN-FP) was tested along with other regression models, trained on fingerprints and one-hot encodings. The best values for each metric have been highlighted in red. The RF model slightly outperforms the original CNN in the validation dataset metrics; however, it is known to be limited in predicting within the range of training data only. As regards testing against the Mach dataset, CNN-Onehot model outperforms the CNN-FP model. However, upon removing the outlier sequence (CNN-FP predicted activity: 140), CNN-FP turns out to be the most optimal model. **r** and **ρ** refers to Pearson's and Spearman's correlation, respectively.

Model - Feature	Validation						Mach				Mach, without outlier			
	uRMSE	RMSE	% Error	R2	r	ρ	RMSE	R2	r	ρ	RMSE	R2	r	ρ
CNN-FP	0.41	2.03	10.55	0.83	0.92	0.81	35.53	-5.90	0.29	0.37	20.16	-1.11	0.22	0.35
Ridge-FP	0.51	2.51	13.04	0.75	0.87	0.78	21.23	-1.46	0.46	0.41	21.56	-1.41	0.45	0.39
Lasso-FP	0.55	2.74	14.24	0.70	0.84	0.78	22.23	-1.70	0.46	0.41	22.34	-1.59	0.44	0.38
SGD-FP	0.61	3.03	15.76	0.63	0.80	0.81	23.28	-1.96	0.42	0.35	23.33	-1.83	0.40	0.32
GP-FP	0.44	2.17	11.32	0.81	0.90	0.82	22.93	-1.88	0.55	0.40	22.86	-1.71	0.53	0.38
RF-FP	0.41	2.02	10.51	0.84	0.92	0.87	23.83	-2.10	0.49	0.46	23.93	-1.97	0.50	0.45
XGBoost-FP	0.47	2.34	12.19	0.78	0.88	0.82	23.68	-2.07	0.49	0.47	23.81	-1.94	0.49	0.45
SVR-FP	0.43	2.13	11.06	0.82	0.91	0.85	24.65	-2.32	0.57	0.43	24.32	-2.07	0.56	0.44
GB-FP	0.48	2.37	12.33	0.77	0.88	0.86	23.19	-1.94	0.44	0.33	23.11	-1.77	0.41	0.34
kNN-FP	0.46	2.27	11.81	0.79	0.89	0.87	21.42	-1.51	0.35	0.29	21.25	-1.34	0.34	0.28
CNN-Onehot	0.46	2.28	11.90	0.80	0.89	0.79	20.88	-1.38	0.47	0.38	21.02	-1.29	0.45	0.37
Ridge-Onehot	0.51	2.51	13.08	0.75	0.87	0.79	22.70	-1.82	0.48	0.42	22.86	-1.71	0.46	0.41
Lasso-Onehot	0.53	2.62	13.64	0.72	0.85	0.78	23.28	-1.96	0.39	0.36	23.36	-1.83	0.36	0.33
SGD-Onehot	0.73	3.61	18.78	0.48	0.74	0.80	25.68	-2.61	0.28	0.03	25.33	-2.33	0.27	0.05
GP-Onehot	0.45	2.21	11.51	0.80	0.90	0.82	24.02	-2.16	0.56	0.39	23.85	-1.95	0.54	0.38
RF-Onehot	0.42	2.06	10.74	0.83	0.91	0.86	25.89	-2.67	0.31	0.24	25.62	-2.41	0.28	0.22
XGBoost-Onehot	0.50	2.49	12.96	0.75	0.87	0.78	25.83	-2.65	0.31	0.23	25.80	-2.46	0.27	0.21
SVR-Onehot	0.55	2.74	14.26	0.70	0.84	0.82	22.12	-1.68	0.51	0.40	22.18	-1.55	0.49	0.38
GB-Onehot	0.50	2.49	12.97	0.75	0.87	0.87	23.54	-2.03	0.45	0.41	23.55	-1.88	0.43	0.39
kNN-Onehot	0.44	2.16	11.24	0.81	0.90	0.88	27.91	-3.26	-0.10	-0.07	27.46	-2.92	-0.07	-0.05

Validation loss is defined as unitless root-mean-squared (uRMSE), since the training and validation data are normalized through scaling by the standard deviation of the training data (σ). Re-scaled RMSE is defined as root-mean-squared error in fold-over-PMO units (uRMSE × σ). %Error is defined through the equation below, where range is the difference between the maximum and minimum value in the training data (19.52 and 0.31 respectively) in fold-over-PMO units.

$$\% \, Error = \frac{uRMSE \ \times \boldsymbol{\sigma}}{range} \times 100\%$$

Table 3. 5 models with distinct random initialization seeds were trained for all possible model-feature combinations (models with ‘*’ do not have a random state initialization, in the sklearn implementation). In comparison to metrics obtained from individual models (Table S2), we note that both CNN models, based on FP and one-hot encoding respectively, stand out for the ensemble models. CNN one-hot is the optimal model on the validation dataset. CNN-FP outperforms CNN-One-hot and other models, however, on RMSE for Mach dataset and Mach dataset without the outlier. **r** and **ρ** refers to Pearson’s and Spearman’s correlation, respectively.

Model - Feature	**Validation**						**Mach**				**Mach, without outlier**			
	uRMSE	**RMSE**	**% Error**	**R2**	**r**	**ρ**	**RMSE**	**R2**	**r**	**ρ**	**RMSE**	**R2**	**r**	**ρ**
CNN-FP	0.41	2.03	10.58	0.83	0.91	0.80	19.89	-1.42	0.33	0.34	17.49	-0.59	0.30	0.32
Ridge-FP	0.51	2.51	13.04	0.75	0.87	0.78	21.23	-1.46	0.46	0.41	21.56	-1.41	0.45	0.39
Lasso-FP	0.55	2.74	14.25	0.70	0.84	0.78	22.23	-1.70	0.46	0.41	22.34	-1.59	0.44	0.38
SGD-FP	0.61	3.01	15.64	0.64	0.80	0.79	23.36	-1.98	0.47	0.40	23.34	-1.83	0.44	0.38
GP-FP	0.44	2.17	11.32	0.81	0.90	0.82	22.94	-1.88	0.55	0.40	22.86	-1.71	0.53	0.38
RF-FP	0.41	2.01	10.48	0.84	0.92	0.87	23.97	-2.14	0.47	0.48	24.10	-2.02	0.49	0.47
XGBoost-FP	0.47	2.34	12.19	0.78	0.88	0.82	23.68	-2.07	0.49	0.47	23.81	-1.94	0.49	0.45
SVR-FP*	0.43	2.13	11.06	0.82	0.91	0.85	24.65	-2.32	0.57	0.43	24.32	-2.07	0.56	0.44
GB-FP	0.48	2.38	12.37	0.77	0.88	0.86	23.12	-1.92	0.42	0.34	23.05	-1.76	0.39	0.34
kNN-FP*	0.46	2.27	11.81	0.79	0.89	0.87	21.42	-1.51	0.35	0.29	21.25	-1.34	0.34	0.28
CNN-Onehot	0.34	1.66	8.64	0.89	0.94	0.85	20.29	-1.25	0.48	0.46	20.48	-1.18	0.46	0.45
Ridge-Onehot	0.61	3.03	15.77	0.63	0.82	0.83	25.13	-2.45	0.39	0.20	24.84	-2.20	0.36	0.20
Lasso-Onehot	0.57	2.82	14.66	0.68	0.83	0.81	23.21	-1.95	0.31	0.31	23.10	-1.77	0.27	0.28
SGD-Onehot	0.73	3.60	18.75	0.48	0.74	0.80	25.52	-2.56	0.29	0.01	25.17	-2.29	0.27	0.02
GP-Onehot	0.45	2.21	11.51	0.80	0.90	0.82	23.87	-2.12	0.55	0.39	23.72	-1.92	0.54	0.38
RF-Onehot	0.41	2.01	10.48	0.84	0.92	0.87	25.75	-2.63	0.33	0.29	25.50	-2.38	0.29	0.26
XGBoost-Onehot	0.50	2.49	12.96	0.75	0.87	0.78	25.83	-2.65	0.31	0.23	25.80	-2.46	0.27	0.21
SVR-Onehot*	0.55	2.74	14.26	0.70	0.84	0.82	22.12	-1.68	0.51	0.40	22.18	-1.55	0.49	0.38
GB-Onehot	0.50	2.48	12.91	0.75	0.87	0.87	23.76	-2.09	0.40	0.42	23.77	-1.93	0.37	0.40
kNN-Onehot*	0.44	2.16	11.24	0.81	0.90	0.88	27.91	-3.26	-0.10	-0.07	27.46	-2.92	-0.07	-0.05

2.2 Benchmarking against classification models

We benchmarked our regression models with corresponding classification models for both fingerprint and one-hot encoding representations (Tables S4, S5). The classification model architectures are similar to those reported in the literature for CPP prediction.[5–9] Similar to our earlier work, the classes were obtained by setting fold over PMO activity threshold of 3.0, above which the sequences were classified as active, otherwise inactive.[10] The hyperparameter optimization and train-valid-test split was same as the benchmarking of regression models (SI Section 2.1, Table S1). For the CNN models, the architecture was kept largely the same as regression models, with the only modifications being the activation function of last layer as sigmoid and loss function as binary crossentropy, which are conventional modifications for classification model architecture. In addition to conventional metrics for evaluating classification models, we used Spearman's coefficient to estimate the rank correlation between the predicted probabilities from the classification models and experimental MFI values of the sequences.

The best metrics for the performance of the classification models against the held-out validation and Mach datasets varied across different model architectures. The CNN, support vector, random forest and stochastic gradient descent models were the optimal models across both representations. This benchmarking experiment further confirms that CNN model architecture is optimal at both classification and regression tasks.

Table 4. The CNN model with 2048-bit fingerprint features was benchmarked against classification models. Spearman correlation coefficient is calculated using predicted probabilities from the classification model and the experimental MFI values. The best values for each metric have been highlighted in red.

	Validation Dataset Metrics						Mach Dataset Metrics					
	Accu-racy	Preci-sion	F1	Recall	ROC-AUC	Spear-man	Accu-racy	Preci-sion	F1	Recall	ROC-AUC	Spear-man
CNN-FP	0.85	0.81	0.85	0.92	0.92	0.81	0.46	0.92	0.66	0.82	0.84	1.00
Gaussian Process	0.84	0.69	0.77	0.87	0.85	0.82	0.54	0.50	0.67	1.00	0.57	0.01
Gradient Boosting	0.78	0.67	0.70	0.73	0.77	0.81	0.77	0.83	0.87	0.91	0.45	0.21
Nearest Neighbors	0.88	0.76	0.82	0.90	0.88	0.79	0.69	0.75	0.82	0.90	0.45	0.28
Random Forest	0.87	0.82	0.82	0.83	0.86	0.85	0.77	0.83	0.87	0.91	0.45	0.02
Ridge	0.90	0.86	0.87	0.88	0.89	0.75	0.85	0.92	0.92	0.92	0.46	0.39
SGD	0.89	0.86	0.86	0.86	0.88	0.72	0.77	0.83	0.87	0.91	0.45	0.00
Support Vector	0.91	0.88	0.88	0.88	0.90	0.75	0.77	0.83	0.87	0.91	0.45	0.00
XGBoost	0.84	0.73	0.77	0.82	0.83	0.84	0.46	0.50	0.63	0.86	0.43	0.22

Table 5. The CNN model with one-hot representation was benchmarked against classification models. Spearman correlation coefficient is calculated using predicted probabilities from the classification model and the experimental MFI values. ROC-AUC for Mach dataset could not be calculated for nearest neighbors and support vector model architectures as all predicted probabilities were less than 0.50. The best values for each metric have been highlighted in red.

	Validation Dataset Metrics						Mach Dataset Metrics					
	Accuracy	**Precision**	**F1**	**Recall**	**ROC-AUC**	**Spearman**	**Accuracy**	**Precision**	**F1**	**Recall**	**ROC-AUC**	**Spearman**
CNN-One-Hot	0.88	0.83	0.77	0.87	0.83	0.62	0.45	0.91	0.58	0.82	0.87	0.62
Gaussian Process	0.87	0.73	0.81	0.90	0.88	0.83	0.77	0.75	0.86	1.00	0.63	0.01
Gradient Boosting	0.86	0.82	0.82	0.82	0.85	0.83	0.77	0.75	0.86	1.00	0.63	-0.27
Nearest Neighbors	0.88	0.80	0.84	0.89	0.88	0.82	0.08	0.00	0.00	0.00	N/A	0.27
Random Forest	0.90	0.82	0.86	0.91	0.90	0.84	0.77	0.75	0.86	1.00	0.63	-0.07
Ridge	0.87	0.80	0.82	0.85	0.86	0.69	0.77	0.83	0.87	0.91	0.45	0.00
SGD	0.90	0.88	0.87	0.86	0.89	0.76	0.77	0.83	0.87	0.91	0.45	0.00
Support Vector	0.84	0.78	0.78	0.79	0.83	0.77	0.08	0.00	0.00	0.00	N/A	0.12
XGBoost	0.88	0.80	0.83	0.87	0.87	0.83	0.46	0.42	0.59	1.00	0.56	0.09

2.3 Benchmarking against CPP webservers

We compared our model to currently available CPP prediction tools by evaluating predictions for Mach peptides (accessed on September 3, 2020). (doi: 10.1186/1479-5876-11-74, doi: 10.1186/s12864-017-4128-1, doi: 10.1021/acs.jproteome.7b00019, doi: 10.1021/acs.jproteome.8b00148, doi: 10.1093/bib/bby091) Of note, these prediction tools do not allow for unnatural residues, therefore when testing the Mach sequences, B (β-alanine) and X (aminohexanoic acid) were replaced by A (alanine) and L (leucine) respectively. Macrocyclic peptides were treated as linear peptides. (Table S6).

All the webservers are generic (do not differentiate between different cargo) binary classifiers and provide the classification probability of the sequence being a CPP, and uptake probability. Most webservers, with the exception of CellPPD, classified all Mach peptides, including the negative control Mach11, as CPP. This result indicates that the webservers are not robust enough to differentiate between highly active and poorly active CPPs. The current work of training a quantitative model (regressor) over a standard dataset with consistent cargo and experimentation is necessary to achieve this distinction.

Table 6. Online webservers that were accessible (as of September 3, 2020) were used to benchmark the Mach peptides. The row labeled 'Model' notes the name of the webserver, and a brief summary has been given in the row 'Summary' (format: year of publication - model architecture - sequence size limitation). The unabbreviated forms of the model architectures are – SVM: Support Vector Machine, RF: Random Forest, and ERT: Extremely Randomized Trees. For CPPred-FL (*), there is no limitation on sequence size, however the prediction is done for a window of 40 residues. Pred denotes prediction, upt denotes uptake and conf denotes confidence.

Model →	**CellPPD**		**SkipCPP-Pred**		**CPPred-RF**				**MLCPP**				**CPPred-FL**		**Current Work**
Summary →	2013 - SVM - 1-50		2017 - RF - >10		2017 - RF - N/A				2018 - ERT and RF - 5-30				2018 - RF - N/A*		2020 - CNN - N/A
Sequence ↓	Pred	SVM Score	Pred	Pred Conf	Pred	Pred Conf	Upt	Upt Conf	Pred	Pred Conf	Upt	Upt Conf	Pred	Pred Conf	Activity
Mach 1	Non-CPP	-0.03	CPP	0.93	CPP	0.73	High	0.50	CPP	0.86	High	0.51	CPP	0.69	29.68
Mach 2	Non-CPP	-0.13	CPP	0.81	CPP	0.67	High	0.52	CPP	0.77	Low	0.31	CPP	0.85	20.82
Mach 3	CPP	0.43	CPP	0.79	CPP	0.73	High	0.52	CPP	0.85	Low	0.45	CPP	0.69	19.52
Mach 4	CPP	0.20	CPP	0.81	CPP	0.79	High	0.56	CPP	0.87	Low	0.39	CPP	0.61	18.48
Mach 5	CPP	0.07	CPP	0.71	CPP	0.69	High	0.59	CPP	0.58	Low	0.41	CPP	0.76	17.47
Mach 6	Non-CPP	-0.03	CPP	0.94	CPP	0.72	High	0.53	CPP	0.86	Low	0.50	CPP	0.82	27.25
Mach 7	Non-CPP	-0.09	CPP	0.93	CPP	0.73	High	0.61	CPP	0.86	High	0.51	CPP	0.77	5.30
Mach 8	Non-CPP	-0.02	CPP	0.91	CPP	0.68	High	0.53	CPP	0.74	High	0.50	CPP	0.78	50.55
Mach 9	CPP	0.18	CPP	0.90	CPP	0.78	High	0.56	CPP	0.89	High	0.51	CPP	0.85	48.90
Mach 10	Non-CPP	-0.56	CPP	0.92	CPP	0.61	High	0.52	CPP	0.68	High	0.52	CPP	0.67	63.43
Mach 11	Non-CPP	-0.40	CPP	0.70	CPP	0.56	High	0.57	CPP	0.56	Low	0.35	CPP	0.77	-11.75
Mach 12	CPP	0.07	CPP	0.90	CPP	0.73	High	0.55	CPP	0.79	High	0.52	CPP	0.75	140.24
Mach 13	CPP	0.48	CPP	0.84	CPP	0.78	High	0.55	CPP	0.86	High	0.62	CPP	0.83	44.68

2.4 Training CNN model on sequences with lower activity (fold over PMO)

In order to determine which factors of the training set are required for accurate prediction, we tested the predictor's accuracy when trained on datasets with increasing activity. We trained CNN models with the same architecture as the optimized model, with activity thresholds increasing by 0.5, starting at 0.5 until 19 (Fig. S4, Table S7). Trained models are able to extrapolate activities beyond the training data only once the training data reaches activity around 8-fold over PMO. The performance then continues to increase with subsequently increasing activity thresholds. RMSE for held out test dataset consisting of sequences having higher activity than threshold increases at first, and then decreases, indicating a barrier for learning. RMSE for Mach sequences continues to decrease as the number of training data points and threshold increases, indicating that extrapolation for high activity sequences (such as Mach sequences) requires a wide range of training data. Including even higher activity sequences in future rounds of model training would potentially lead to more accurate predictions in future work. The experiment notes that activity has a sequence dependence that the model is able to learn, once meeting an activity threshold in the training data.

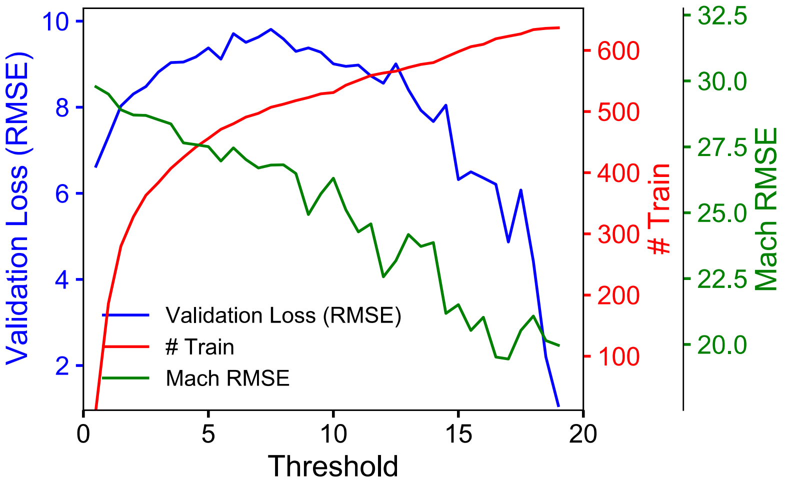


Figure 4.Threshold for extrapolation. The validation loss (RMSE), in blue, for the held-out dataset of sequences having higher activity than the ones used for training the model goes through a barrier. RMSE for Mach sequences, in green, is decreasing with increasing threshold. The number of training datapoints is indicated in red.

Table 7. The evaluation metrics for models trained with sequences filtered by activity threshold have been noted. The original CNN model has been highlighted in grey color.

Threshold	#Train	#Test	uRMSE	RMSE	MachRMSE	% Error	R^2	Pearson
0.5	13	627	1.34	6.63	29.78	34.53	-0.79	-0.08
1	186	454	1.48	7.32	29.5	38.08	-1	-0.09
1.5	280	360	1.62	8.03	28.91	41.82	-1.43	-0.33
2	328	312	1.68	8.31	28.71	43.28	-1.72	0.03
2.5	363	277	1.71	8.48	28.69	44.14	-1.98	-0.04
3	384	256	1.78	8.82	28.53	45.91	-2.38	-0.26
3.5	407	233	1.83	9.04	28.37	47.08	-2.8	-0.19
4	425	215	1.83	9.05	27.65	47.11	-3.07	-0.08
4.5	442	198	1.85	9.17	27.58	47.73	-3.53	0.02
5	456	184	1.9	9.38	27.51	48.84	-4.16	-0.03
5.5	471	169	1.84	9.12	26.96	47.48	-4.46	0.11
6	480	160	1.96	9.71	27.46	50.56	-5.72	0.04
6.5	491	149	1.92	9.51	27.02	49.49	-6.29	-0.04
7	497	143	1.95	9.63	26.7	50.15	-7.08	0.1
7.5	507	133	1.98	9.81	26.81	51.05	-8.61	-0.07
8	512	128	1.94	9.59	26.82	49.95	-8.99	-0.04
8.5	518	122	1.88	9.3	26.49	48.43	-9.34	0.11
9	523	117	1.9	9.38	24.93	48.85	-10.38	0.14
9.5	529	111	1.88	9.28	25.72	48.32	-11.27	0.12
10	531	109	1.82	9.01	26.31	46.92	-10.92	0.14
10.5	543	97	1.81	8.95	25.11	46.6	-13.17	0.22
11	551	89	1.82	8.98	24.27	46.73	-16	0.17
11.5	559	81	1.76	8.73	24.58	45.42	-19	0.11
12	563	77	1.73	8.56	22.57	44.57	-21.05	0.13
12.5	566	74	1.82	9.01	23.18	46.9	-25.47	0
13	572	68	1.7	8.41	24.17	43.78	-26.83	0.03
13.5	577	63	1.6	7.93	23.72	41.26	-27.35	0.23
14	580	60	1.55	7.67	23.87	39.92	-27.54	0.27
14.5	589	51	1.63	8.05	21.18	41.92	-36.2	0.24
15	598	42	1.28	6.32	21.51	32.89	-27.91	0.26
15.5	606	34	1.31	6.5	20.53	33.83	-38.67	0.27
16	610	30	1.29	6.36	21.03	33.11	-41.49	0.32
16.5	619	21	1.26	6.21	19.52	32.33	-57.81	-0.04
17	623	17	0.99	4.87	19.45	25.38	-45.61	0.65
17.5	627	13	1.23	6.08	20.53	31.66	-76.13	0.59
18	634	6	0.89	4.42	21.08	23.03	-75.21	0.66
18.5	636	4	0.45	2.2	20.15	11.47	-39.37	0.88
19	637	3	0.22	1.07	19.97	5.57	-44.72	-0.99
CNN-FP	512*	128*	0.41	2.03	10.55	35.53	0.83	0.92

2.5 Training CNN model on sequences by leaving one residue/linker out

In order to demonstrate the advantage of the fingerprint representations in giving the model flexibility in predictions, we trained CNN models by leaving one residue out then having it predict activity of all sequences (Table S8). Models trained with sequences leaving one residue/linker out are able to compensate for the missing residue to a certain extent. Models were trained with sequences without a particular residue/linker and evaluated against the test dataset with sequences containing the residue/linker. This demonstrates that the model is able to infer chemical rules, such as the similarities between glutamic acid and glutamine, and aspartic acid and asparagine. It is hypothesized that the performance in some cases may be marred by a lack of sufficient training data points for the maximum similarity residue. The absence of similar sequence motifs, in the case of replacement by maximum similarity residue, may also be a contributing factor in the performance of the models. These results suggest that the model is able to use information learned from other amino acids.

Table 8. The evaluation metrics for models trained with sequences without a particular residue/linker have been noted, in decreasing order of RMSE. The residue/linker in the dataset with maximum chemical similarity, evaluated using Tanimoto similarity over the fingerprints, has been noted in the column Chem. Simil.

Residue/ Linker	Chem. Simil.	#Train	#Test	uRMSE	RMSE	Mach RMSE	% Error	R^2	Pearson
I	V, 0.70	114	526	1.41	6.95	20.79	36.17	-2	0.3
E	Q, 0.66	456	184	1.39	6.86	24.49	35.68	-0.35	0.54
A	S, 0.76	150	490	1.24	6.14	17.93	31.95	-0.99	0.42
R	Q, 0.51	12	628	1.16	5.76	28.62	29.96	-0.34	-0.12
L	I, 0.63	74	566	1.14	5.66	22.77	29.44	-0.23	0.12
K	L, 0.62	25	615	1.14	5.64	19.86	29.35	-0.26	0.24
Q	E, 0.66	221	419	1.14	5.64	21.74	29.34	-1.19	0.31
S	A, 0.76	186	454	1.1	5.44	27.79	28.32	-0.09	0.18
G	A, 0.62	172	468	1.1	5.42	25.35	28.21	-0.02	0.39
P	Q, 0.47	143	497	1.07	5.29	24.8	27.52	-0.09	0.28
H	F, 0.27	402	238	1.05	5.2	22.31	27.07	-1.2	0.18
3	W, 0.18	64	576	1.05	5.17	27.17	26.91	-0.01	0.33
V	I, 0.70	44	596	0.99	4.88	25.93	25.42	0.08	0.56
2	C, 0.55	64	576	0.97	4.79	26.96	24.94	0.13	0.44
B	A, 0.55	592	48	0.92	4.57	24.62	23.8	0.55	0.87
X	K, 0.61	592	48	0.89	4.4	25.02	22.92	0.59	0.87
Y	F, 0.82	387	253	0.83	4.13	18.74	21.48	0.15	0.5
N	D, 0.54	267	373	0.79	3.91	21.33	20.37	0.09	0.38
C	2, 0.55	495	145	0.7	3.47	22.95	18.07	0.63	0.8
F	Y, 0.82	245	395	0.67	3.31	25.38	17.23	0.36	0.62
W	F, 0.35	323	317	0.65	3.23	22.65	16.81	0.36	0.66
M	V, 0.48	445	195	0.63	3.14	24.05	16.34	0.27	0.55
T	S, 0.57	343	297	0.62	3.07	22.13	15.99	0.46	0.7
D	N, 0.54	399	241	0.6	2.96	22.58	15.42	0.42	0.71
CNN-FP	-	512*	128*	**0.41**	**2.03**	**10.55**	35.53	**0.83**	**0.92**

3 Analysis of the role of generator

In order to investigate the role and advantages of generator, we conducted five in silico experiments comparing the generator to other methods of generating seed sequences (Fig. S5, Table S9). The predictor-optimizer loop was seeded with 50 sequences sampled from random sequences from the predictor training dataset (CPP Library), the 50 most active sequences from the predictor training dataset (CPP Library Top50), the CPP thesaurus (CPPsite 2.0), randomly generated sequences with equal likelihood for all amino acids at all sites, and sequences sampled using the generator as reported in the main text.

Our three criteria for optimized sequences are: high predicted activity, low similarity, and low Arg content. The optimized sequences have varying ranges of these characteristics. The top 50 sequences from the predictor dataset receive a head start in terms of activity, resulting in the highest predicted activity, followed by the generator-sampled, CPP library, and CPP thesaurus and random sequences. On the other hand, the maximum and mean similarities for sequences optimized using the seeds from the Top50 and full CPP Library are higher than for the sequences optimized using generator-sampled, CPP thesaurus, and randomly generated seeds. Finally, the generator-sampled, CPP thesaurus, and random sequences resulted in optimized sequences with lower Arg content than sequences from Top50 and full CPP library. Taken together, sampling seeds from the generator is a more favorable option for meeting our three criteria. However, we note that with appropriate diversity constraints and predicted activity thresholds, it is possible to sample sequences using other routes and still predict sequences with the desired characteristics. We note that this comparison uses predicted activity, and robust comparison of these strategies would require experimental validation.

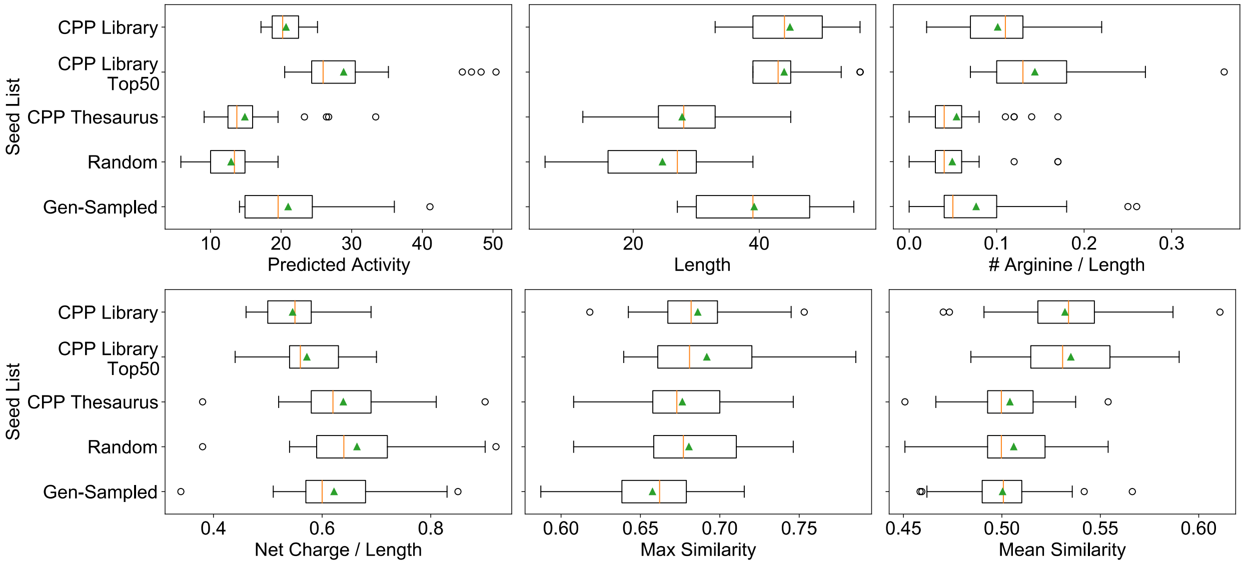


Figure 5. Box plots comparing optimized sequences comparing random seeds, randomly selected sequences from the predictor dataset, and seed sequences sampled from the generator dataset. For the box plot, the box marks the interquartile range (IQR), Q1 and Q3; the whiskers are at Q1-1.5*IQR and Q3+1.5*IQR; the orange line is the median; the green triangle is the mean, and outliers, if outside the whiskers, are marked as dots, N = 50 sequences.

Table 9. Statistical metrics comparing optimization of seed sequences sampled from different lists.

Statistic	Seed List	Predicted Activity	Length	# Arginine/ Length	Net Charge/ Length	Maximum Similarity	Mean Similarity
Mean	**CPP Library**	20.70	44.88	0.10	0.55	0.69	0.53
	CPP Library Top 50	28.86	43.95	0.14	0.57	0.69	0.54
	CPP Thesaurus	14.87	27.76	0.05	0.64	0.68	0.50
	Random	12.92	24.63	0.05	0.66	0.68	0.51
	Gen-Sampled	21.00	39.2	0.08	0.62	0.66	0.50
Median	**CPP Library**	20.23	44	0.11	0.55	0.68	0.53
	CPP Library Top 50	25.94	43	0.13	0.56	0.68	0.53
	CPP Thesaurus	13.74	28	0.04	0.62	0.67	0.50
	Random	13.40	27	0.04	0.64	0.68	0.50
	Gen-Sampled	19.58	39	0.05	0.60	0.66	0.50
Minimum	**CPP Library**	17.16	33	0.02	0.46	0.62	0.47
	CPP Library Top 50	20.51	39	0.07	0.44	0.64	0.48
	CPP Thesaurus	9.11	12	0.00	0.38	0.61	0.45
	Random	5.81	6	0.00	0.38	0.61	0.45
	Gen-Sampled	14.12	27	0.00	0.34	0.59	0.46
Maximum	**CPP Library**	25.15	56	0.22	0.69	0.75	0.61
	CPP Library Top 50	50.42	56	0.36	0.70	0.79	0.59
	CPP Thesaurus	33.40	45	0.17	0.90	0.75	0.55
	Random	19.57	39	0.17	0.92	0.75	0.55
	Gen-Sampled	41.07	55	0.26	0.85	0.72	0.57
Q1	**CPP Library**	18.74	39	0.07	0.50	0.67	0.52
	CPP Library Top 50	24.34	39	0.10	0.54	0.66	0.51
	CPP Thesaurus	12.47	24	0.03	0.58	0.66	0.49
	Random	10.01	16	0.03	0.59	0.66	0.49
	Gen-Sampled	14.89	30	0.04	0.57	0.64	0.49
Q3	**CPP Library**	22.47	50	0.13	0.58	0.70	0.55
	CPP Library Top 50	30.51	45	0.18	0.63	0.72	0.55
	CPP Thesaurus	15.98	33	0.06	0.69	0.70	0.52
	Random	14.89	30	0.06	0.72	0.71	0.52
	Gen-Sampled	24.42	48	0.1	0.68	0.68	0.51

4 Analysis of the optimization approach

4.1 Analyzing the role of constraints in the optimizer

We tested the model with varying combinations of constraints to probe the role of each constraint in sequence optimization (Table S10). 5 seed sequences with variable length (10, 20, 30, 40, 50) were used to seed the predictor-optimizer loop, where the optimizer had none to all constraints – maximization of predicted activity, minimization of similarity, minimization of Arg content, minimization of length, and maintenance of net charge for water solubility. Removal of a constraint leads to clear change in the optimized sequences. For instance, when minimization of Arg content is not a constraint, the sequences have a high degree of net Arg. Such sequences have been shown to be toxic in vivo and are already a known cell-penetrating motif. A goal of this work was to generate unique high-activity sequences that do not rely on Arg for activity.

The constraints are also necessary to shift away from the bias in the training dataset, since without them, the predicted sequences would appear to be very similar to the sequences in the training dataset. This dataset was created using a combinatorial approach with the currently known cell-penetrating peptides. The peptides in the training dataset inherently have a net high Arg content and longer length than desired. The optimizer also makes mutations to the seed sequences using motifs from the training data, so without minimizing similarity, the end sequences would resemble the training sequences. The constraints help to reduce the bias present in the dataset, and optimize sequences towards desired properties and away from those in the training dataset such that we are more likely to discover new sequences and motifs.

Table 10. List of top sequences obtained from the optimization without additional constraints.

Constraints	Sequences	Intensity	Length	Relative Arg	Relative Charge
Seeds	KHAPRRESSW	2.50	10.00	0.20	0.28
	RWTAWTLRRIAKAVGPIVRR	2.71	20.00	0.25	0.30
	SCRRPQRKDVLTIAHRSRNRIRGAHARPNR	4.60	30.00	0.30	0.35
	GKEKQSWRRFQRKTPRSAAQMRAKRALARARLQLSRSQRR	3.89	40.00	0.28	0.35
	RSSHHGCARSPRLRRHKRRKPIKVRLRRRMKLELKKTARKRKSRRRGLHC	2.78	50.00	0.32	0.52
MFI	RRRRRQRRRRRR	11.53	12.00	0.92	0.92
	RKRRRQRKRRRRWPXRXIPQYDQXF	14.40	25.00	0.40	0.44
	KKKRPQLKRRRRGPMRXCSEFDFHFPRPTK	14.45	30.00	0.23	0.36
	GKKRRSRRRRRRGPKGGVPQPSQGYPKYSBNRXRRRRRX	28.31	39.00	0.36	0.46
	RRRRRLLKRRRRKGKKXLPKFREGYPLGLKPRKRRQRRRYRWGRGKHRTWW	26.90	51.00	0.37	0.53
MFI, Length	RRRRRQRRRRRR	11.53	12.00	0.92	0.92
	RRRRRQGKRRRRGPRGKVPEPPQHSPKY	15.53	28.00	0.36	0.46
	RKKRRQRKRRRRGPMGKRSRPSQGYALYLK	16.03	30.00	0.33	0.50

	BKKKNSBBKRRRWRGKNAPQPKAKYPLWILRRRRRQRGRYRR	21.28	42.00	0.31	0.48
	XRRRRLLRRLRRNPGRGRLRVIFGRKRGAANRXRRMRXRGPWARKRHXRW	33.74	50.00	0.42	0.48
MFI, Arg	RRRRRQRRRRRRWPMG	12.50	16.00	0.69	0.69
	RRRRRQRKRRRRWCKKGIPE	13.14	20.00	0.50	0.60
	RRKRPQERRRRRGLNRXCSEPPQHYAIYCK	14.16	30.00	0.30	0.32
	GKRKKSBLKKRRALKXRRBKAKAGQRQYALKRXRRQRXRLPRWR	25.83	44.00	0.30	0.50
	RRNRRENKRRRRGLBMALPRPAEGYLLRLINKRRLQRRRGPWARXRKXRW	32.84	50.00	0.36	0.38
MFI, Charge	RRRRRQRRRRRR	11.53	12.00	0.92	0.92
	RRRRRQRKRRRRGPMGXCPRP	14.19	21.00	0.52	0.57
	RRRRRQRRRRRRGPGGGNPRPEQHVPDFLBG	16.10	31.00	0.39	0.35
	GKKRRQRRRRRRGPNNKNPQFSQKYPQPPRXKRRRRRRR	24.81	39.00	0.41	0.54
	RRLRRLRLRRRRYLRGKLLQKKVKYKQGLRBRRRRQRXRAPRKPRRRRKRWCR	38.53	53.00	0.45	0.58
MFI, Length, Arg	RRKRRQ	6.66	6.00	0.67	0.83
	RRKRRQRKRRRRGPKKGVPQ	14.32	20.00	0.45	0.65
	RRRRRQRPKKRRGPLRGCPQFRQHFLQYL	15.43	29.00	0.34	0.44
	BRRRRQRKRRRRYRBKGIPQPREKYLQYLIRXXKRQRXRRRR	26.89	42.00	0.43	0.50
	RRGRRLRKLRRRWRGRRRAKPRLGYPRYADRRRRRERRRRRYWRQKHXRW	29.25	50.00	0.52	0.56
MFI, Arg, Charge	RRRRRQRRRRRRWP	12.32	14.00	0.79	0.79
	BKKRRQRRRRNRWRGKNCPQPSLSYAMY	14.31	28.00	0.29	0.39
	WRKRPQRKRRRRWPKKADPQPAQBVAQPLBGRX	17.67	33.00	0.24	0.33
	GRKKRQRLKRRRGPMRGKPQPSSKYPRYSKKXRRLQRRX	23.06	39.00	0.31	0.49
	RRRRRRLRRRRRRPGNALARADQDYLAYVLNRGRRRRRXACBCRXLHW	26.00	48.00	0.42	0.39
MFI, Length, Charge	RRRRRQRRRRRR	11.53	12.00	0.92	0.92
	RRRRRQEKLRRRGPNKGIPQPSQHYPIYLLG	16.94	31.00	0.26	0.32
	RRRRRQRKKRRRGPLGGGLQFKEGVPQYVQNRX	18.48	33.00	0.30	0.36
	RKRRKSRRRRRRRRPNGGRSQPEQXYLLPTBXRRRRRKXXRRRW	27.35	43.00	0.44	0.49
	RRRRRKALKLLRYPKKINLQPREKQPQWLAKKRRRRRRXRRRWRXRRWRWXCGRXM	53.09	56.00	0.38	0.48
MFI, Length, Arg, Charge	HRKRRQ	6.70	6.00	0.50	0.80
	RRRRRQRKRRRRWRGGGVPRPSQBQPV	14.73	27.00	0.44	0.48
	RRKRRQRRRRRRGGKBGBPIPIQXVPQYLIRXXRRBR	20.80	37.00	0.38	0.43
	KKRRKQAKKRRRNPKKNNPQFDFHFPRPTLXRGRRKGRX	23.29	39.00	0.26	0.46

	RRLRRSGLKSRRGLLGKSSQPSKGRRLPSKKRGKLL KGXGLWGRGKRXTWWCM	30.39	53.00	0.21	0.36

4.2 Optimizing for sequences with all canonical residues

We determined that sequences with comparable activity could not be achieved using only canonical residues. We optimized peptides containing only canonical residues by constraining the optimizer to use only canonical residues for mutations (Fig. S6, Table S11). 50 seed sequences sampled from generator were used to seed the predictor-optimizer loop. While we can predict fully canonical peptides, the predicted activities of these peptides are significantly lower than those containing noncanonical residues. Given the constraints in the optimizer (minimization of length and Arg content), we observe diminishing returns of length versus Arg content, where shorter sequences have more Arg in order to have a high predicted MFI.

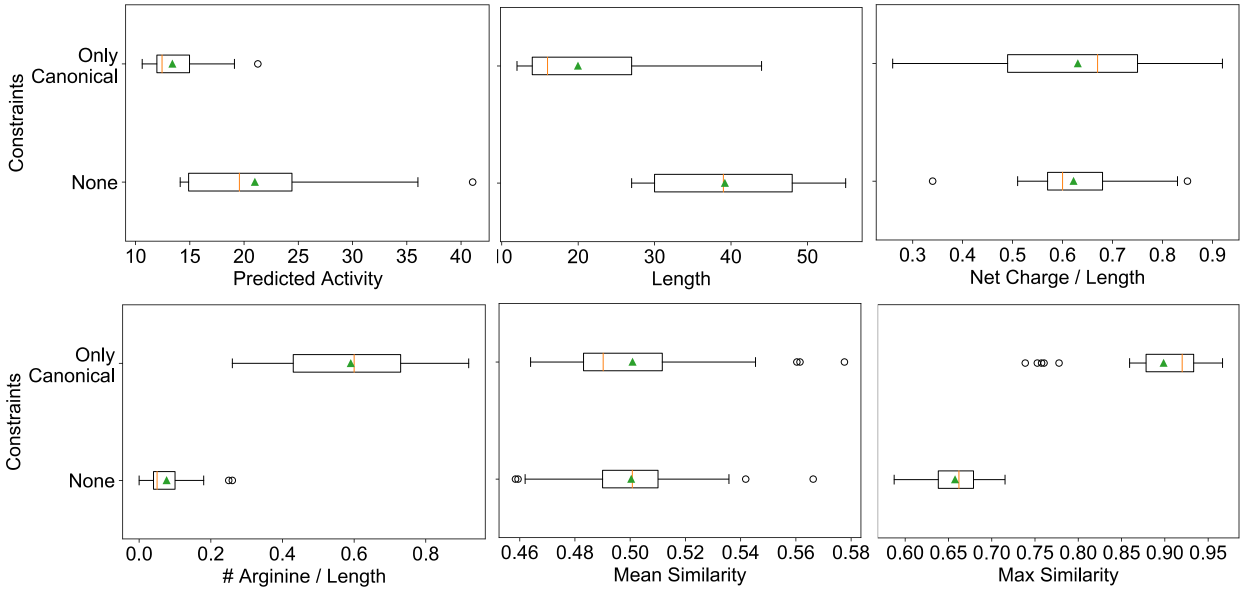


Figure 6. Box plots comparing optimized sequences with and without the constraint of only being able to use canonical residues for the genetic algorithm mutations. For the box plot, the box marks the interquartile range (IQR), Q1 and Q3; the whiskers are at Q1-1.5*IQR and Q3+1.5*IQR; the orange line is the median; the green triangle is the mean, and outliers, if outside the whiskers, are marked as dots, N = 50 sequences.

Table 11. Optimized sequences in decreasing order of predicted MFI.

Sequence	Predicted MFI	Length	Net Arg	Net Charge
PRKKRRSRRRRRRRLRGDPQPPQGRKIYVLGTRRLQRRRGPWRPRRRGRR	21.62	50	0.46	0.5
KRRRRQIRRRRRYRLRNVLQPEQMRKQGLLGRRRRQLRRYPYRR	21.27	44	0.45	0.48
ARKRRQRKRRRRWPMGANLVFSLHYAQYTKGRRRRRR	19.12	37	0.38	0.48
RRRRRQRKRRRRGPGGGDPEPAQGYPI	16.18	27	0.37	0.33
KAKRRQRRRRRRGPNGGDPRPSQHYPD	16	27	0.33	0.36
RRRRRQRRRRRRGPQKPCPQPSQKYA	15.85	26	0.42	0.5
RRRRRQGKRRRRWRNRGCPQPDQKYPDYC	15.51	29	0.38	0.38
RRRRRQRRRRRRWPQRPLPQPRQHILDYVN	15.36	30	0.43	0.43
RRRRREEKRRRRGPGGPCLQFSLSAPQYSK	15.2	30	0.3	0.3
RRRRRQRRRRRRWPMGKMPQPSQ	15.12	23	0.48	0.52
KKKRRQRRRRRRWRAKGIPEPSFKYKQPPHGR	15.04	32	0.31	0.49
RRRRRQRRRRRRWRGGPCPRPIQHIPQ	14.97	27	0.48	0.51
RRRRRQRRRRRRGRGGPRSQFSQHYPQ	14.6	27	0.48	0.51
RRRRRLGKRRTRGPLGPCPQFDEGILI	14.28	27	0.3	0.26
RRRRRQRPLRRRGPNKPCPEPDQ	14.17	23	0.39	0.35
KRRRREEKKKKKWRRGGCPRPRQHYPQYPKG	13.87	31	0.26	0.44
RRRRRLRKRRRRGPMGKCSD	13.02	20	0.5	0.55
RRRRRQRKRRRRGCNGNCPD	12.92	20	0.5	0.5
RRRRRQRKRRRRGPMGPC	12.79	18	0.56	0.61
RRRRRQRKKRRRGPMGPC	12.52	18	0.5	0.61
RRRRRQRRRRRRWPMG	12.5	16	0.69	0.69
RRRRRQRRRRRRGPM	12.44	15	0.73	0.73
RRRRRQRRRRRRWPM	12.41	15	0.73	0.73
RRRRRQRRRRRRWPG	12.36	15	0.73	0.73
RKRRRQRRRRRRGPM	12.35	15	0.67	0.73
RRRRRQRRRRRRGPG	12.35	15	0.73	0.73
RRKRRQRRRRRRWP	12.27	14	0.71	0.79
RRRRRQRRRRRRGP	12.27	14	0.79	0.79
RRRRRQRNRRRRWPM	12.23	15	0.67	0.67
RRRRRQRPRRRRWP	12.2	14	0.71	0.71
RRRRRQRKKRRRWPM	12.08	15	0.6	0.73
RRRRRERKRRRRWPMG	11.97	16	0.62	0.62
RRRRRQRKKRRRWP	11.96	15	0.6	0.73

RRRRRQRRRRRR	11.53	13	0.85	0.85
RRRRRQRRRRRR	11.53	12	0.92	0.92
RRKRRQRRRRRR	11.49	12	0.83	0.92
RRRRRQRKRRRR	11.48	12	0.83	0.92
RRKRRQRKRRRR	11.45	12	0.75	0.92
RKRRRQRRRRRR	11.44	12	0.83	0.92
RRRRRQRPRRRR	11.39	12	0.83	0.83
RRRRRQERRRRR	11.15	12	0.83	0.75
RRKRQQRRRRRR	10.61	12	0.75	0.83
RRRRRLRRRRRR	10.6	12	0.92	0.92
RRRRRQRRRKRR	10.56	12	0.83	0.92
RRKRRQRRRRHR	10.51	12	0.75	0.9
HRKRRQ	6.7	7	0.43	0.68
HRKRRQ	6.7	6	0.5	0.8
RRRRRQ	6.61	6	0.83	0.83
HRKRRE	5.91	6	0.5	0.63
RRKRRM	5.42	6	0.67	0.83

5 Attribution analysis

5.1 Interpretability

Using the conceptual attribution framework developed to understand activation of neural networks for image classification, we developed a toolkit to visualize the decision making process of the CNN model.[11] We chose the first convolution layer of the model to access the fingerprint indices. Taking the first derivative of the model output (normalized fluorescence intensity) with respect to the input representation (row matrix of fingerprints), produces a Jacobian matrix of partial derivatives. We performed element-wise multiplication of the Jacobian with the input representation to zero out the activation of absent chemical features, and clipped negative values, to focus on features that drive high MFI. We analyzed the role of individual activated fingerprints by visualizing the corresponding chemical substructure, and also obtained the average activation over the residue positions and fingerprint indices (Fig. S7 and S8).

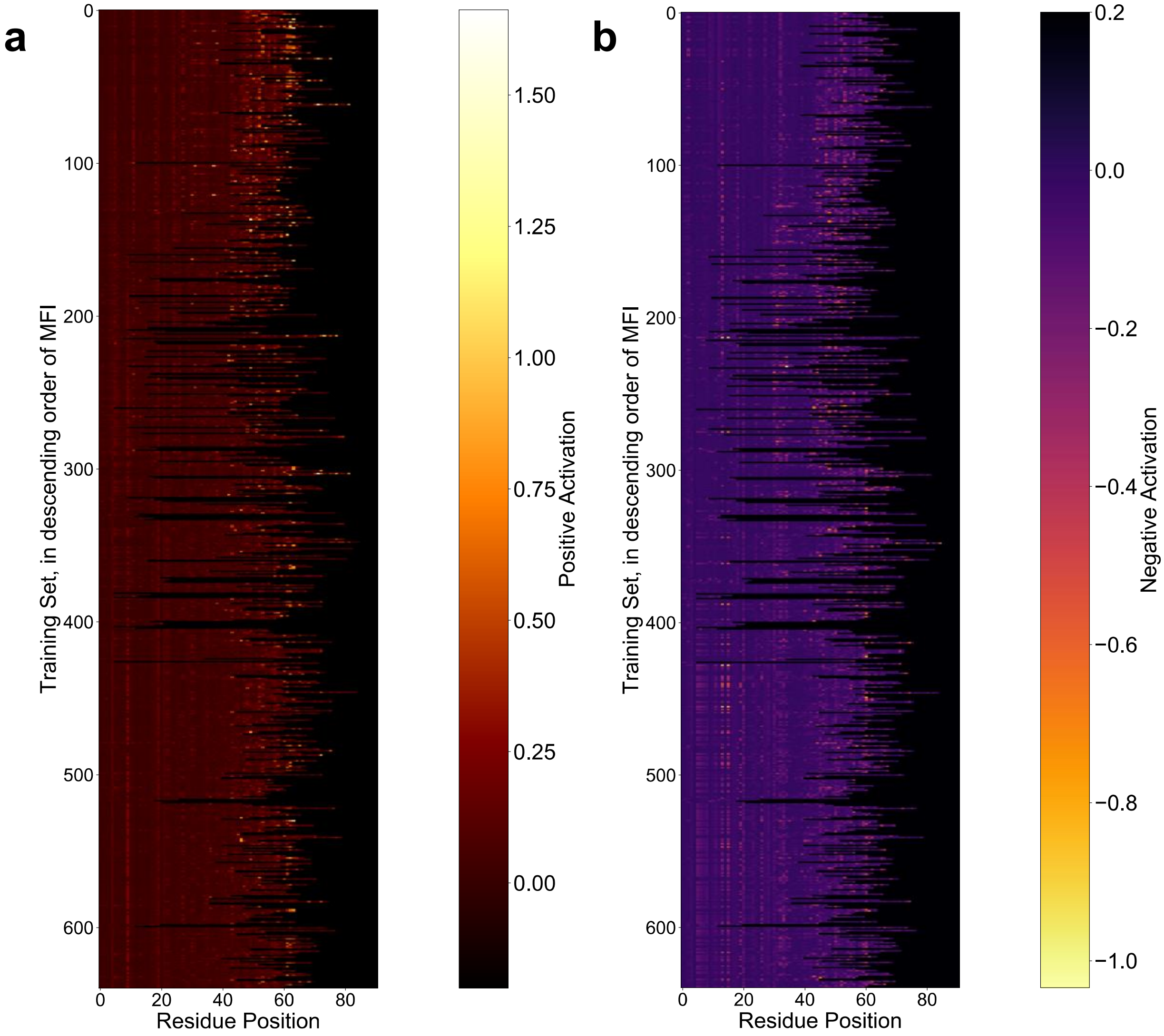


Figure 7 Activation map of predictor training set relative to amino acid position. Gradient activations for sequences are arranged in descending order of experimental normalized MFI for **(a)** positive and **(b)** negative activation averaged over residue position. The positive activation for C-terminal residues decreases with decrease in normalized MFI values. The most active sequences have a highly positively activated C-terminus and a sparsely negatively activated C-terminus.

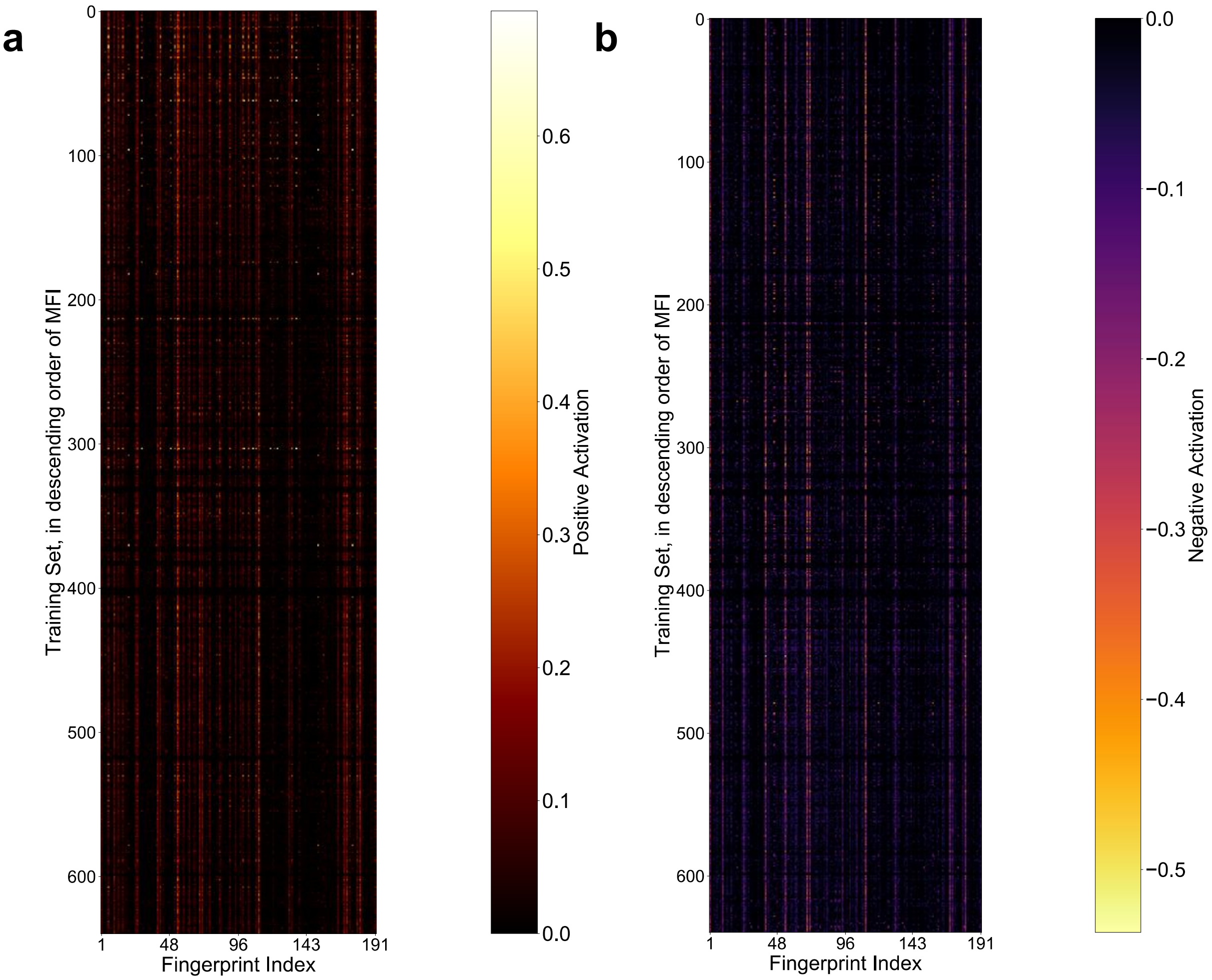


Figure 8 Activation map of predictor training dataset relative to fingerprint index. Gradient activations for sequences are arranged in descending order of normalized MFI for **(a)** positive and **(b)** negative activation averaged over fingerprint index. The most positively activated substructures by residue are for aminohexanoic acid, β-alanine, aspartic acid, threonine and serine.

5.2 Robustness of Attribution Analysis

We evaluated the robustness of gradient-based attribution by analyzing residue-activations for mutated Mach3 sequences (Fig. S9). We mutated each active Ahx residue individually with Ala, followed by both with beta-alanine and aminoundecanoic acid. This experiment is analogous to the method reported for validating attributions for protein-ligand binding by designing an adversarial ligand, which is a modified version of a correctly predicted ligand.[12] In this report, the modified ligand, present in a database of useful decoys[13], should have been predicted as non-binding, however, the model owing to other substructures inaccurately predicts this as a binding ligand.

Consistent with our earlier findings, we observed that the most activated residue was the lone Ahx in single Ala mutations, followed by Arg when both Ahx are mutated to Ala or β-Ala, residues with a shorter alkyl backbone. The most activated residue reverts back to undecanoic acid, a residue with a longer alkyl backbone than Ahx, for the undecanoic acid mutation. This experiment validates the robustness of attribution analysis, both in terms of activated residues which conform to known biochemical principles and experimental validation of the mutations (SM Section 8).

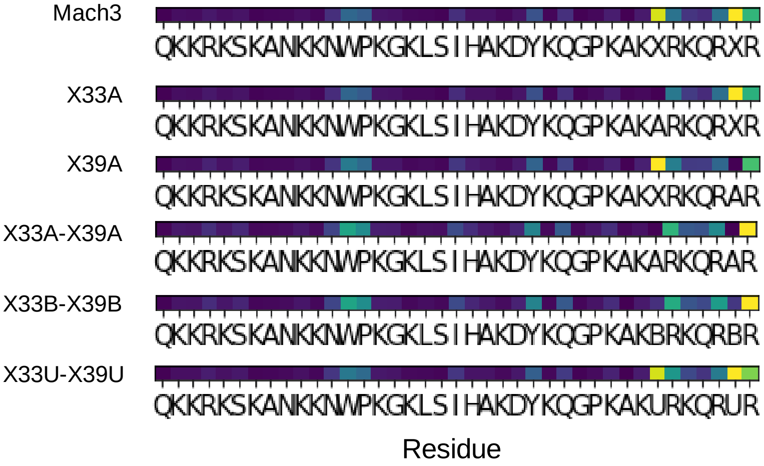


Figure 9. Attribution analysis of mutations. Alanine and β-Alanine mutations of the most active residue(s) shows a fall in the predicted MFI and a corresponding change in the positive activation heatmap. Ahx remains the most active residue for single Ala mutations, that changes to Arg when both C-terminus Ahx are mutated to Ala and β-Ala. However, when Ahx is mutated to undecanoic acid (U), both U are the most positively activated residues.

6 Similarity and immunogenicity analysis of predicted sequences

6.1 Similarity of Training and Validation Sequences

Similarity among sequences in each training and validation dataset was analyzed using Jaro-Winkler distance metric (Fig. S10).[14] Each sequence was compared with the rest of the library to evaluate the string similarity.

The sequences used to train the generator have a mean similarity of 47%, indicating that we capture a combinatorial chemical space of cell-penetrating peptide sequences. For the sequences used to train the predictor, composed mostly of the modular library, the mean similarity is 66%. The modularity of the sequences from the library can be seen clearly in the visualization of sequence similarity (Fig S10). The four highlighted squares along the diagonal correspond to module 2 of the sequences. Similarly, the four lighter colored boxes correspond to module 3. The non-modular sequences, which are dissimilar from one another, are on the bottom of the visualization.

Finally, Mach sequences were confirmed to be unique by comparing to library sequences as well as a protein database. Similarity of Mach sequences was first compared to the library using mean Jaro-Winkler distance (Fig. S11). All Mach sequences had a mean similarity less than 60% when compared to the training dataset. Then to compare Mach peptides to the existing proteome, we used BLASTp on the online server.[15] The search was done using default values to search the UniProt database. There was no sequence homology between Mach sequences and known proteins for significant E-values less than 0.01. For the unnatural residues, B (β-alanine) and X (aminohexanoic acid) were replaced by A (alanine) and L (leucine) respectively for the search operation. Mach sequences containing cysteine macrocycles were excluded from the search.

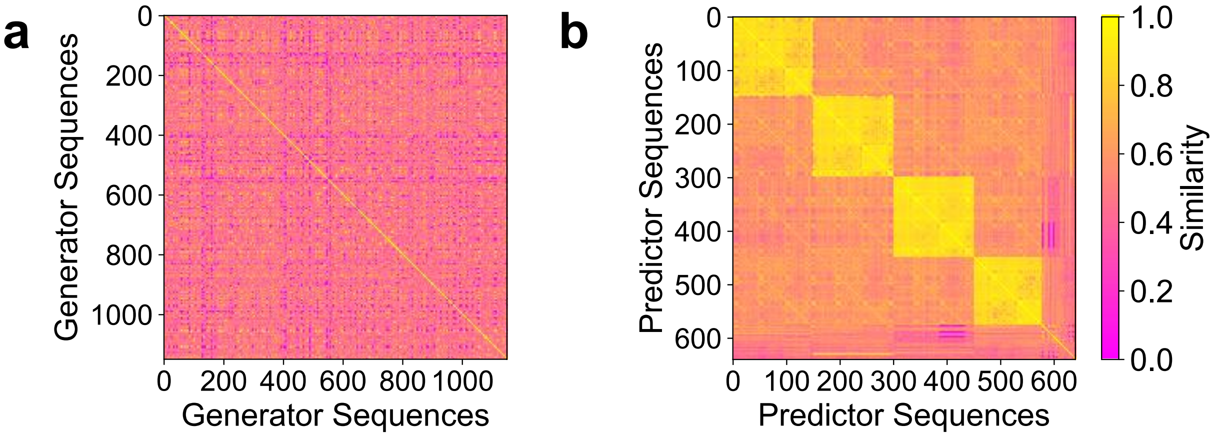


Figure 10 Similarity of sequences used in the training of generator and predictor. Each sequence used in training of **(a)** generator (Nested LSTM) and **(b)** predictor (Convolutional Neural Network based model) is compared with the rest of respective training dataset. The mean similarities of the sequences are 47% and 66% for the generator and predictor respectively. The heatmap for the predictor sequences have a modular pattern owing to the combinatorial nature of the library. Jaro-Winkler distance was used as the metric to assess the similarity between two sequences

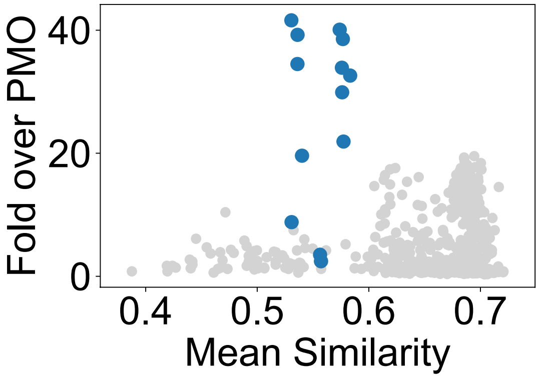


Figure 11. Similarity and experimental normalized MFI of Mach and training sequences. Mach sequences (blue) are novel and high-performing in comparison to the sequences used in the training of the predictor (grey). For each Mach sequence, the Jaro-Winkler distance with the predictor training dataset was averaged. For the rest of the training dataset, the mean similarity was calculated by averaging over the similarity with rest of the library. The mean similarities for all Mach sequences is less than 60%.

6.2 Immunogenicity Score of Predicted and Library Sequences

The likelihood of being a T-cell epitope was calculated for all sequences using an online server.[16] The score is an arbitrary number, where a higher positive value indicates a higher probability of the peptide to be immunogenic and vice-versa. For the unnatural residues, B (β-alanine) and X (aminohexanoic acid) were replaced by A (alanine) and L (leucine) respectively for the search operation. The Mach sequences were compared to the sequences used in the training of the predictor (Fig. S12).

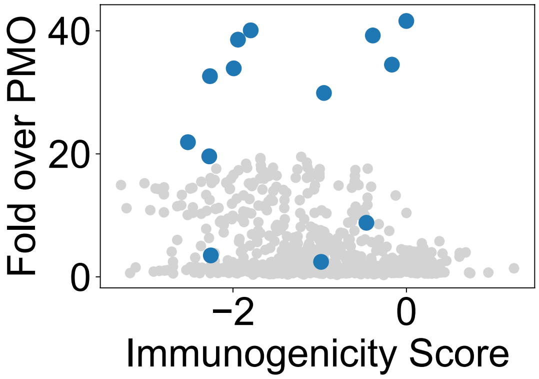


Figure 12. In silico immunogenicity score for Mach and training sequences. Predicted immunogenicity for Mach sequences is within the range of the predicted immunogenicity for the sequences used in training of predictor. Mach sequences have a substantially higher experimental normalized MFI within the same range of immunogenicity, in comparison to the sequences used in the training of the predictor. The immunogenicity scores are the likelihood of being a T-cell epitope. The values are calculated using an online predictor.[16]

7 Evaluation of PMO-Mach constructs

7.1 Endocytosis Inhibition Assays

Chemical endocytosis inhibitors were used to probe the mechanism of delivery of PMO in a pulse-chase format. For the PMO constructs, HeLa 654 cells were preincubated with various chemical inhibitors or incubated at 4 °C for 30 minutes before treatment with PMO-Mach constructs for three hours. Treatment media was then replaced with fresh media for 22 hours. Cells were then lifted as previously described and EGFP synthesis was measured by flow cytometry (Fig. S13).

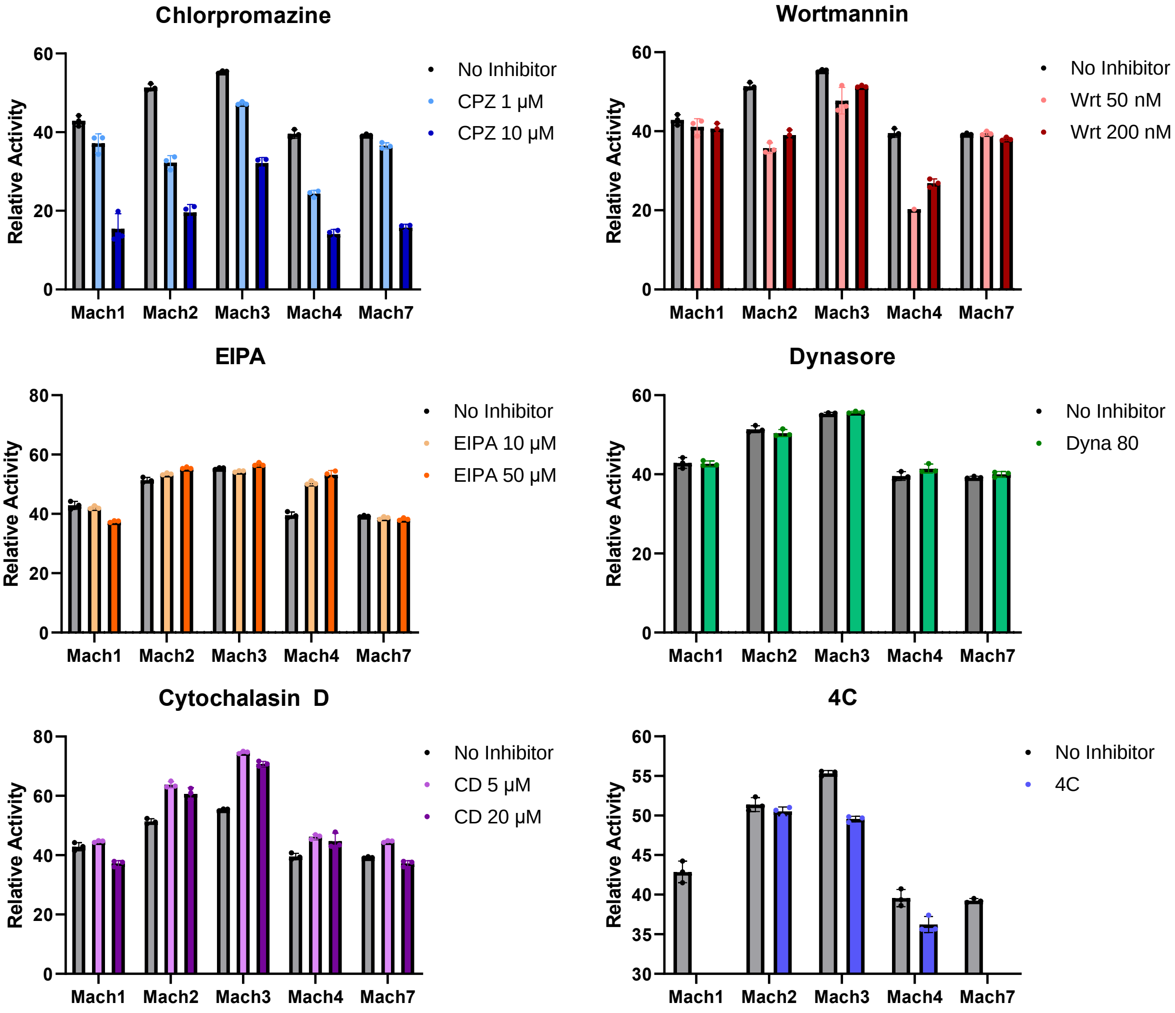


Figure 13. PMO-Mach peptides enter cells by energy-dependent endocytosis. PMO activity of Mach constructs when treated with various endocytosis inhibitors. Chlorpromazine (CPZ) has a dose-dependent effect on PMO activity for each of the Mach constructs, indicating that constructs may enter via clathrin-mediated endocytosis. Each bar represents group mean ± SD, n = 3, with the exception of Mach4 Wrt 50 nM condition in which n=1 due to Wrt toxicity to cells.

7.2 Circular Dichroism

Peptides were dissolved in PBS buffer to obtain stock solutions of 1 mM. The circular dichroism (CD) spectra was obtained from 195 to 250 nm using an AVIV 420 circular dichroism spectrometer with a 1 mm path length quartz cuvette. Peptides in PBS buffer at 20 µM, with or without 10 mM sodium dodecyl sulfate (SDS) were used in the measurement (Fig. S14).

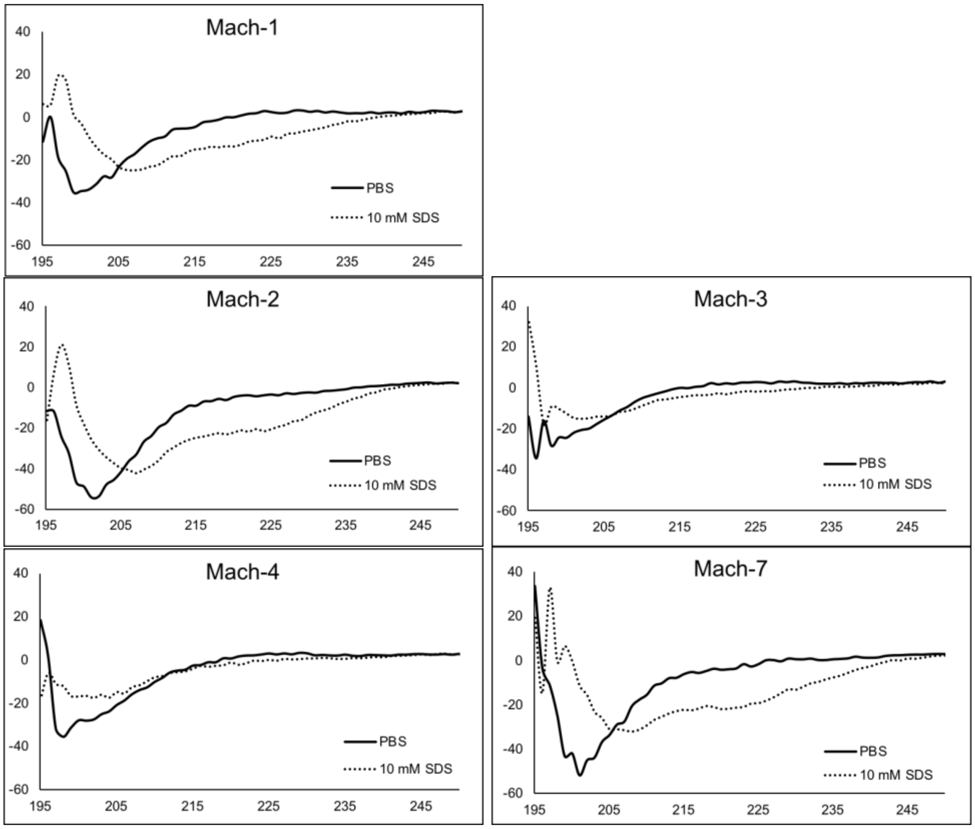
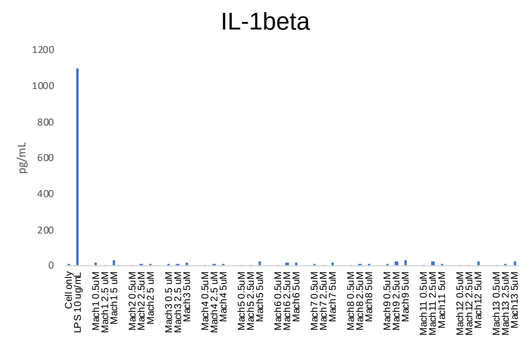
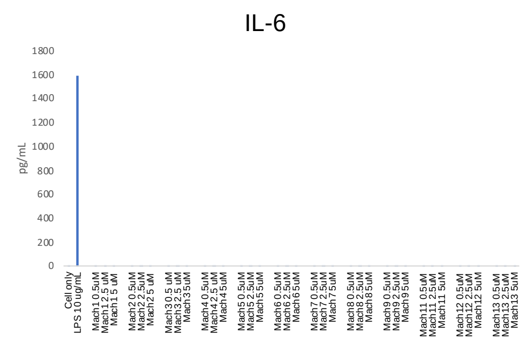


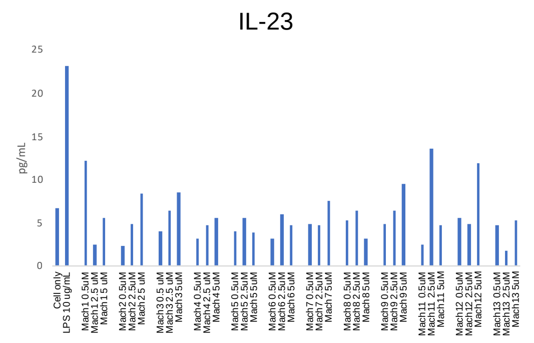
Figure 14. Circular dichroism of azide-Mach peptides. 20 µM Mach peptides were either incubated in PBS or 10 mM SDS before analysis using circular dichroism. In buffer, these peptides do not exhibit secondary structure. In a lipid environment, Mach1, Mach2, and Mach7 exhibit partial alpha helicity.

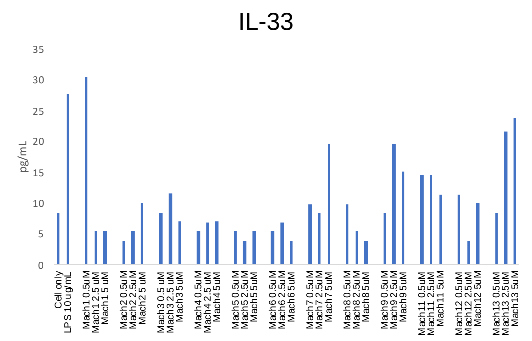
7.3 Inflammation panel

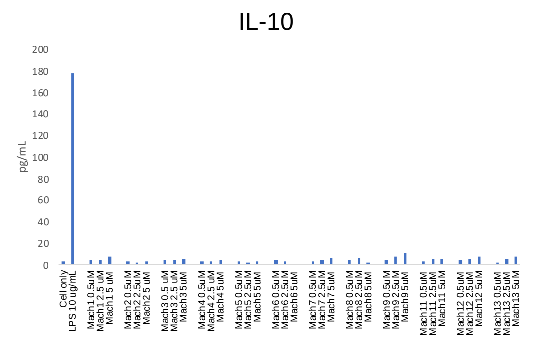
THP-1 cells (ATCC TIB-202) were grown in RPMI 1640 media supplemented with 10% (v/v) FBS, 1% (v/v) penicillin-streptomycin, L-glutamine, non-essential amino acids, sodium pyruvate at 37 °C and 5% CO_2. THP-1 cells (450k/mL) were treated with 25 nM phorbol 12-myristate 13-acetate (PMA) at 37 °C and 5% CO_2 for 24 h to trigger differentiation into macrophages. Then, media was replaced with fresh RPMI media and the cells were incubated for another 24 h. Cells were then collected, spun down, and brought up in complete RPMI media to a cell density of 500k/mL. 100k cells were plated in each well of a 96-well plate, leaving the first two columns empty. Duplicate wells were treated with varying concentrations of the PMO-peptide conjugates at 37 °C and 5% CO_2 for 2 h. Media-only and no treatment wells were used as negative controls, and 10 µg/mL bacterial lipopolysaccharide (LPS) treatment was used as a positive control. Following treatment, each well was washed three times, given fresh media, and incubated for 12 h. Supernatant was transferred to a V-bottom plate. Inflammatory cytokines in the supernatant were assayed using LEGENDplex Human Inflammation panel (BioLegend). Analysis was carried out on a BD LSRII flow cytometer and data was analyzed using BioLegend's accompanying software (Fig S15).

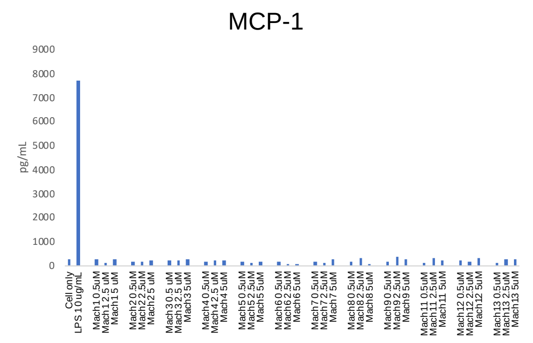

IL-1beta
pg/mL
1200
1000
800
600
400
200
0
Cell only
LPS 10 ug/mL
Mach1 0.5uM
Mach1 2.5 uM
Mach1 5 uM
Mach2 0.5uM
Mach2 2.5uM
Mach2 5 uM
Mach3 0.5 uM
Mach3 2.5 uM
Mach3 5uM
Mach4 0.5uM
Mach4 2.5 uM
Mach4 5uM
Mach5 0.5uM
Mach5 2.5uM
Mach5 5uM
Mach6 0.5uM
Mach6 2.5uM
Mach6 5uM
Mach7 0.5uM
Mach7 2.5uM
Mach7 5uM
Mach8 0.5uM
Mach8 2.5uM
Mach8 5uM
Mach9 0.5uM
Mach9 2.5uM
Mach9 5uM
Mach11 0.5uM
Mach11 2.5uM
Mach11 5uM
Mach12 0.5uM
Mach12 2.5uM
Mach12 5uM
Mach13 0.5uM
Mach13 2.5uM
Mach13 5uM

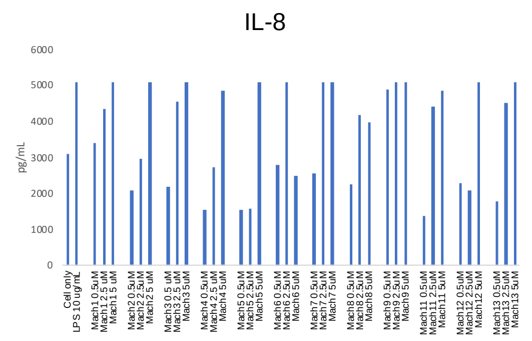

IL-6
pg/mL
1800
1600
1400
1200
1000
800
600
400
200
0
Cell only
LPS 10 ug/mL
Mach1 0.5uM
Mach1 2.5 uM
Mach1 5 uM
Mach2 0.5uM
Mach2 2.5uM
Mach2 5 uM
Mach3 0.5 uM
Mach3 2.5 uM
Mach3 5uM
Mach4 0.5uM
Mach4 2.5 uM
Mach4 5uM
Mach5 0.5uM
Mach5 2.5uM
Mach5 5uM
Mach6 0.5uM
Mach6 2.5uM
Mach6 5uM
Mach7 0.5uM
Mach7 2.5uM
Mach7 5uM
Mach8 0.5uM
Mach8 2.5uM
Mach8 5uM
Mach9 0.5uM
Mach9 2.5uM
Mach9 5uM
Mach11 0.5uM
Mach11 2.5uM
Mach11 5uM
Mach12 0.5uM
Mach12 2.5uM
Mach12 5uM
Mach13 0.5uM
Mach13 2.5uM
Mach13 5uM

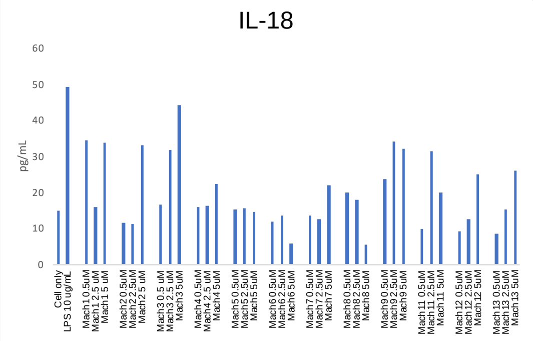

IL-23
pg/mL
25
20
15
10
5
0
Cell only
LPS 10 ug/mL
Mach1 0.5uM
Mach1 2.5 uM
Mach1 5 uM
Mach2 0.5uM
Mach2 2.5uM
Mach2 5 uM
Mach3 0.5 uM
Mach3 2.5 uM
Mach3 5uM
Mach4 0.5uM
Mach4 2.5 uM
Mach4 5uM
Mach5 0.5uM
Mach5 2.5uM
Mach5 5uM
Mach6 0.5uM
Mach6 2.5uM
Mach6 5uM
Mach7 0.5uM
Mach7 2.5uM
Mach7 5uM
Mach8 0.5uM
Mach8 2.5uM
Mach8 5uM
Mach9 0.5uM
Mach9 2.5uM
Mach9 5uM
Mach11 0.5uM
Mach11 2.5uM
Mach11 5uM
Mach12 0.5uM
Mach12 2.5uM
Mach12 5uM
Mach13 0.5uM
Mach13 2.5uM
Mach13 5uM


IL-33
pg/mL
35
30
25
20
15
10
5
0
Cell only
LPS 10 ug/mL
Mach1 0.5uM
Mach1 2.5 uM
Mach1 5 uM
Mach2 0.5uM
Mach2 2.5uM
Mach2 5 uM
Mach3 0.5 uM
Mach3 2.5 uM
Mach3 5uM
Mach4 0.5uM
Mach4 2.5 uM
Mach4 5uM
Mach5 0.5uM
Mach5 2.5uM
Mach5 5uM
Mach6 0.5uM
Mach6 2.5uM
Mach6 5uM
Mach7 0.5uM
Mach7 2.5uM
Mach7 5uM
Mach8 0.5uM
Mach8 2.5uM
Mach8 5uM
Mach9 0.5uM
Mach9 2.5uM
Mach9 5uM
Mach11 0.5uM
Mach11 2.5uM
Mach11 5uM
Mach12 0.5uM
Mach12 2.5uM
Mach12 5uM
Mach13 0.5uM
Mach13 2.5uM
Mach13 5uM


IL-10
pg/mL
200
180
160
140
120
100
80
60
40
20
0
Cell only
LPS 10 ug/mL
Mach1 0.5uM
Mach1 2.5 uM
Mach1 5 uM
Mach2 0.5uM
Mach2 2.5uM
Mach2 5 uM
Mach3 0.5 uM
Mach3 2.5 uM
Mach3 5uM
Mach4 0.5uM
Mach4 2.5 uM
Mach4 5uM
Mach5 0.5uM
Mach5 2.5uM
Mach5 5uM
Mach6 0.5uM
Mach6 2.5uM
Mach6 5uM
Mach7 0.5uM
Mach7 2.5uM
Mach7 5uM
Mach8 0.5uM
Mach8 2.5uM
Mach8 5uM
Mach9 0.5uM
Mach9 2.5uM
Mach9 5uM
Mach11 0.5uM
Mach11 2.5uM
Mach11 5uM
Mach12 0.5uM
Mach12 2.5uM
Mach12 5uM
Mach13 0.5uM
Mach13 2.5uM
Mach13 5uM


MCP-1
pg/mL
9000
8000
7000
6000
5000
4000
3000
2000
1000
0
Cell only
LPS 10 ug/mL
Mach1 0.5uM
Mach1 2.5 uM
Mach1 5 uM
Mach2 0.5uM
Mach2 2.5uM
Mach2 5 uM
Mach3 0.5 uM
Mach3 2.5 uM
Mach3 5uM
Mach4 0.5uM
Mach4 2.5 uM
Mach4 5uM
Mach5 0.5uM
Mach5 2.5uM
Mach5 5uM
Mach6 0.5uM
Mach6 2.5uM
Mach6 5uM
Mach7 0.5uM
Mach7 2.5uM
Mach7 5uM
Mach8 0.5uM
Mach8 2.5uM
Mach8 5uM
Mach9 0.5uM
Mach9 2.5uM
Mach9 5uM
Mach11 0.5uM
Mach11 2.5uM
Mach11 5uM
Mach12 0.5uM
Mach12 2.5uM
Mach12 5uM
Mach13 0.5uM
Mach13 2.5uM
Mach13 5uM


IL-8
pg/mL
6000
5000
4000
3000
2000
1000
0
Cell only
LPS 10 ug/mL
Mach1 0.5uM
Mach1 2.5 uM
Mach1 5 uM
Mach2 0.5uM
Mach2 2.5uM
Mach2 5 uM
Mach3 0.5 uM
Mach3 2.5 uM
Mach3 5uM
Mach4 0.5uM
Mach4 2.5 uM
Mach4 5uM
Mach5 0.5uM
Mach5 2.5uM
Mach5 5uM
Mach6 0.5uM
Mach6 2.5uM
Mach6 5uM
Mach7 0.5uM
Mach7 2.5uM
Mach7 5uM
Mach8 0.5uM
Mach8 2.5uM
Mach8 5uM
Mach9 0.5uM
Mach9 2.5uM
Mach9 5uM
Mach11 0.5uM
Mach11 2.5uM
Mach11 5uM
Mach12 0.5uM
Mach12 2.5uM
Mach12 5uM
Mach13 0.5uM
Mach13 2.5uM
Mach13 5uM


IL-18
pg/mL
60
50
40
30
20
10
0
Cell only
LPS 10 ug/mL
Mach1 0.5uM
Mach1 2.5 uM
Mach1 5 uM
Mach2 0.5uM
Mach2 2.5uM
Mach2 5 uM
Mach3 0.5 uM
Mach3 2.5 uM
Mach3 5uM
Mach4 0.5uM
Mach4 2.5 uM
Mach4 5uM
Mach5 0.5uM
Mach5 2.5uM
Mach5 5uM
Mach6 0.5uM
Mach6 2.5uM
Mach6 5uM
Mach7 0.5uM
Mach7 2.5uM
Mach7 5uM
Mach8 0.5uM
Mach8 2.5uM
Mach8 5uM
Mach9 0.5uM
Mach9 2.5uM
Mach9 5uM
Mach11 0.5uM
Mach11 2.5uM
Mach11 5uM
Mach12 0.5uM
Mach12 2.5uM
Mach12 5uM
Mach13 0.5uM
Mach13 2.5uM
Mach13 5uM

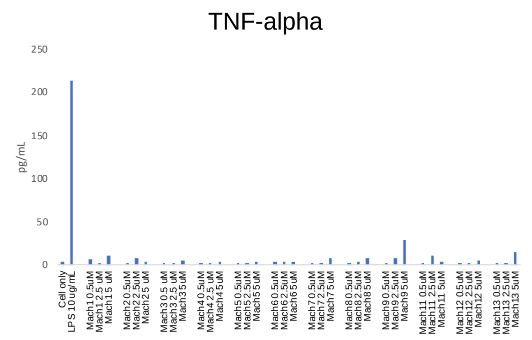
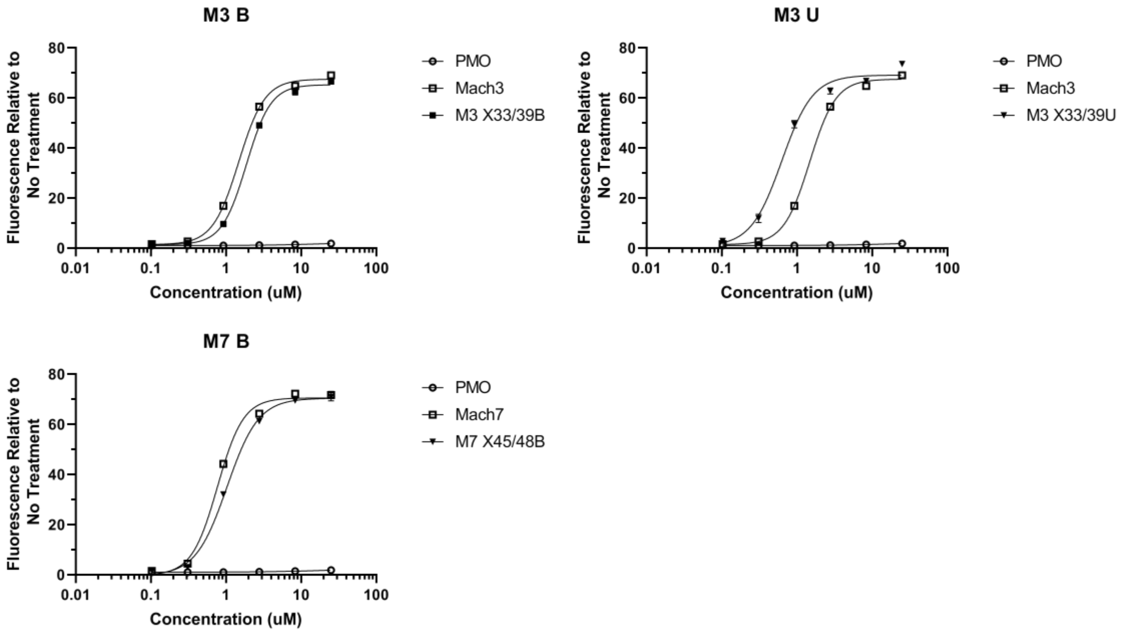

TNF-alpha
pg/mL
250
200
150
100
50
0
Cell only
LPS 10 ug/mL
Mach1 0.5uM
Mach1 2.5 uM
Mach1 5 uM
Mach2 0.5uM
Mach2 2.5uM
Mach2 5 uM
Mach3 0.5 uM
Mach3 2.5 uM
Mach3 5uM
Mach4 0.5uM
Mach4 2.5 uM
Mach4 5uM
Mach5 0.5uM
Mach5 2.5uM
Mach5 5uM
Mach6 0.5uM
Mach6 2.5uM
Mach6 5uM
Mach7 0.5uM
Mach7 2.5uM
Mach7 5uM
Mach8 0.5uM
Mach8 2.5uM
Mach8 5uM
Mach9 0.5uM
Mach9 2.5uM
Mach9 5uM
Mach11 0.5uM
Mach11 2.5uM
Mach11 5uM
Mach12 0.5uM
Mach12 2.5uM
Mach12 5uM
Mach13 0.5uM
Mach13 2.5uM
Mach13 5uM

Figure 15. PMO-Mach constructs are nonimmunogenic in vitro. Inflammation panel results of cytokines that were detected in human monocyte-derived macrophages. IL-1B, TNF-a, IL-6, IL-10, and MCP-1 are all released after treatment with lipopolysaccharide (LPS), but exhibit no significant increase after treatment with PMO-Mach constructs. Each bar represents group mean, n = 2.

8 Post-hoc mutations of PMO-Mach miniprotein

Attribution analysis provides opportunity for post-hoc experimentation with peptide sequences. Given that Ahx is highly activated in Mach3, and reports that extended alkyl backbone chain amino acids have a large effect on CPP activity, we hypothesized that mutating these residues to residues with a longer chain may increase activity. Also observed is that the C-terminus of Mach3 and Mach7 are the highest activated regions on the sequence, we hypothesized that the 10 C-terminal residues may retain some CPP activity.

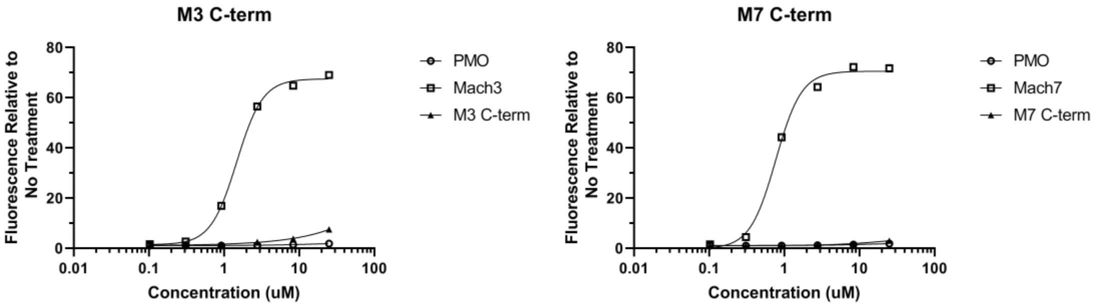
We made several point mutations and truncations to Mach3 and Mach7 to investigate our hypotheses. Mutating to undecanoic acid indeed enhanced the activity of Mach3, decreasing the EC50 to 0.6 µM from 1.5 µM. Mutation from X to B decreased activity only slightly in both Mach3 and Mach7 (Fig S16). Finally, the 10 C-terminal residues of each miniprotein do not retain the activity of the full-length sequence (Fig. S17).



Mach3	QKKRKSKANKKNWPKGKLSIHAKDYKQGPKAKXRKQRXR
Mach3 X33,38B	QKKRKSKANKKNWPKGKLSIHAKDYKQGPKAKBRKQRBR
Mach3 X33,38U	QKKRKSKANKKNWPKGKLSIHAKDYKQGPKAKURKQRUR
Mach7	XKHPXAVQBAARAWKVPAAALWKKKRLKKSSKQKKKWLWKARSAXKYXRLI
Mach7 X45,48B	XKHPXAVQBAARAWKVPAAALWKKKRLKKSSKQKKKWLWKARSABKYBRLI

Figure 16. Mutations of Mach peptides can affect activity. Shown are dose-response curves in HeLa 654 after testing with PMO-Mach analogs, along with their sequences. B = beta-alanine, X = aminohexanoic acid, U = aminoundecanoic acid. Mutation to beta-alanine decreases activity

slightly. Mutation to aminoundecanoic acid increases activity of Mach3 significantly. Activity is shown as fluorescence relative to untreated cells, with the curve corresponding to PMO alone also shown. Points and error bars represent mean and standard deviation, respectively. n = 3.



Mach3	QKKRKSKANKKNWPKGKLSIHAKDYKQGPKAKXRKQRXR
Mach7	XKHPXAVQBAARAWKVPAAALWKKKRLKKSSKQKKKWLWKARSAXKYXRLI
M3 C-term	KAKXRKQRXR
M7 C-term	RSAXKYXRLI

Figure 17. Truncation of Mach peptides ablates PMO activity. Shown are dose-response curves in HeLa 654 after testing with PMO-Mach analogs, along with their sequences. B = beta-alanine, X = aminohexanoic acid. The 10 C-terminal residues of Mach3 and Mach7 do not retain the activity of the parent sequence. Activity is shown as fluorescence relative to untreated cells, with the curve corresponding to PMO alone also shown. Points and error bars represent mean and standard deviation, respectively. n = 3.

9 Recombinant Expression & Purification

His_6-SUMO-G_5-DTA(C186S), His_6-SUMO-G_5-DTA(C186S, E148S) and G_5-EGFP-His_6 were overexpressed in E. coli BL21 (DE3) cells. Approximately 10 g of cell pellet was lysed by sonication in 50 mL of 20 mM Tris, 150 mM NaCl, pH 7.5 buffer containing 30 mg lysozyme, 2 mg DNAase I, and 1 tablet of cOmplete^{TM} Protease Inhibitor Cocktail. The suspension was centrifuged at 16,000 rpm for 30 min to remove cell debris. The supernatant was loaded onto a 5 mL HisTrap FF Ni-NTA column (GE Healthcare, UK) and washed with 30 mL of 100 mM imidazole in 20 mM Tris, 150 mM NaCl, pH 8.5. Protein was eluted from the column with buffer containing 300 mM imidazole in 20 mM Tris, 150 mM NaCl, pH 8.5. Imidazole was removed from protein via centrifugation in Millipore centrifugal filter unit (10K).

For the DTA constructs, the His_6-SUMO tag was then cleaved from the protein with SUMO protease (previously recombinantly expressed) by incubating a 1:1000 protease:protein ratio in 20 mM Tris, 150 mM NaCl, pH 7.5 overnight at 4 °C. Desired protein was separated from His_6-

SUMO tag by flowing the mixture through a 5 mL HisTrap FF Ni-NTA column. Finally, purified protein was isolated by size exclusion chromatography using HiLoad 26/600 Superdex 200 prep grade size exclusion chromatography column (GE Healthcare, UK) in 20 mM Tris, 150 mM NaCl, pH 7.5 buffer.

For the EGFP construct, purified protein was isolated by anion exchange chromatography using HiTrap Q HP anion exchange chromatography column (GE Healthcare, UK) in (0-40 %B over 20 CV) where A: 20 mM Tris, pH 8.5 buffer and B: 1 M NaCl, 20 mM Tris, pH 8.5 buffer.

Proteins were analyzed using an SDS-Page gel. In addition, proteins were analyzed by ESI-QTOF LCMS to confirm molecular weight and purity. The protein charge-state envelope was deconvoluted using Agilent MassHunter Bioconfirm using maximum entropy.

Supplemental References

1. Hartrampf, N. *et al.* Synthesis of Proteins by Automated Flow Chemistry. *ChemRxiv Prepr.* (2020) doi:10.26434/chemrxiv.11833503.v1.
2. Pedregosa, F. *et al.* Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).
3. Chen, T. & Guestrin, C. XGBoost: A Scalable Tree Boosting System. in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 785–794 (Association for Computing Machinery, 2016). doi:10.1145/2939672.2939785.
4. Head, T., Kumar, M., Nahrstaedt, H., Louppe, G. & Shcherbatyi, I. *scikit-optimize/scikit-optimize*. (Zenodo, 2020). doi:10.5281/zenodo.4014775.
5. Gautam, A. *et al.* In silico approaches for designing highly effective cell penetrating peptides. *J. Transl. Med.* **11**, 74 (2013).
6. Wei, L., Tang, J. & Zou, Q. SkipCPP-Pred: an improved and promising sequence-based predictor for predicting cell-penetrating peptides. *BMC Genomics* **18**, 742 (2017).
7. Wei, L. *et al.* CPPred-RF: A Sequence-based Predictor for Identifying Cell-Penetrating Peptides and Their Uptake Efficiency. *J. Proteome Res.* **16**, 2044–2053 (2017).
8. Manavalan, B., Subramaniyam, S., Shin, T. H., Kim, M. O. & Lee, G. Machine-Learning-Based Prediction of Cell-Penetrating Peptides and Their Uptake Efficiency with Improved Accuracy. *J. Proteome Res.* **17**, 2715–2726 (2018).
9. Qiang, X. *et al.* CPPred-FL: a sequence-based predictor for large-scale identification of cell-penetrating peptides by feature representation learning. *Brief. Bioinform.* **21**, 11–23 (2020).
10. Wolfe, J. M. *et al.* Machine Learning To Predict Cell-Penetrating Peptides for Antisense Delivery. *ACS Cent. Sci.* **4**, 512–520 (2018).
11. Selvaraju, R. R. *et al.* Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. *Proc. IEEE Int. Conf. Comput. Vis.* 618–626 (2017) doi:10.1109/ICCV.2017.74.

12. McCloskey, K., Taly, A., Monti, F., Brenner, M. P. & Colwell, L. J. Using attribution to decode binding mechanism in neural network models for chemistry. *Proc. Natl. Acad. Sci.* **116**, 11624–11629 (2019).

13. Mysinger, M. M., Carchia, M., Irwin, John. J. & Shoichet, B. K. Directory of Useful Decoys, Enhanced (DUD-E): Better Ligands and Decoys for Better Benchmarking. *J. Med. Chem.* **55**, 6582–6594 (2012).

14. Winkler, W. E. String Comparator Metrics and Enhanced Decision Rules in the Fellegi-Sunter Model of Record Linkage. (1990).

15. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410 (1990).

16. Calis, J. J. A. *et al.* Properties of MHC Class I Presented Peptides That Enhance Immunogenicity. *PLoS Comput. Biol.* **9**, (2013).

Supplemental Figures and Tables

Table 12: List of peptides used for each module in the 600-member library*

PMO	Name	Sequence
	PMO IVS2-654	GCT ATT ACC TTA ACC CAG
Peptide 1	Name	Sequence
	Penetratin	RQIKIWFQNRRMKWKK
	pVec	LLIILRRRIRKQAHAHSK
	TP10	AGYLLGKINLKALAALAKKIL
	DPV6	GRPRESGKKRKRKRLKP
Peptide 2	Name	Sequence
	KRVK (NLS)	KRVK
	SV40 (NLS)	PKKKRKV
	AAV-PHP.eB	SDGTLAVPFKA
Peptide 3	Name	Sequence
	DPV6	ZGRPRESGKKRKRKRLKP
	PPC3	ZKKYRGRKRHPR
	PPC5	ZGRKAARAPGRRKQ
	R12	ZRRRRRRRRRRRR
	R12 full cycle	Z**C**RRRRRRRRRRRR**C**
	R12 N-cycle	Z**C**RRRRRR**C**RRRRRR
	R12 C-cycle	ZRRRRRR**C**RRRRRR**C**
	R12 benzyl bicycle	Z*C*RRRRRR*C*RRRRRR*C*
	R12 double cycle	Z**C**RRRRRR**CC**RRRRRR**C**
	Bpep	ZRXRRBRRXRRBR
	Bpep full cycle	Z**C**RXRRBRRXRRBR**C**
	Bpep C-cycle	ZRXRRBR**C**RXRRBR**C**
	Penetratin (nle)	ZRQIKIWFQNRRMKWKK
	Engrailed N-cycle	Z**C**QIKIWF**C**NKRAKIKK
	Engrailed C-cycle	ZSQIKIWFQ**C**KRAKIK**C**
	Engrailed full cycle	Z**C**SQIKIWFQNKRAKIKK**C**
	pVEC	ZLLIILRRRIRKQAHAHSK
	pVEC-Bpep	ZLLIILRRRIRKQAHAHSKRXRRBRRXRRBR
	AIP6 full cycle	Z**C**RLRWR**C**
	Melittin-Bpep	ZGIGAVLKVLTTGLPALISWIKRKRQQRXRRBRRXRRBR
	Bh3 helix	ZIWIAQELRRIGDEFNAYYARR
	Bac7	ZRRIRPRPPRLPRPRPRPLPFPRPG

Buforin 2	ZTRSSRAGLQWPVGRVHRLLRK
Melittin	ZGIGAVLKVLTTGLPALISWIKRKRQQ
SynB1	ZRGGRLSYSRRRFSTSTGR
S413-PVrev	ZALWKTLLKKVLKAPKKKRKV
Ribotoxin2 L3	ZKLIKGRTPIKFGKADCDRPPKHSQNGMGK
PreS2-TLM	ZPLSSIFSRIGDP
MAP	ZKLALKALKALKAALKLA
W/R	ZRRWWRRWRR
MAP12	ZLKTLTETLKELTKTLTEL
SAP	ZVRLPPPVRLPPPVRLPPP
SVM1	ZFKIYDKKVRTRVVKH
SVM3	ZKGTYKKKLMRIPLKGT
SVM4	ZLYKKGPAKKGRPPLRGWFH
YTA4	ZIAWVKAFIRKLRKGPLG
439a	ZGSPWGLQHHPPRT
HoxA13 serine2	ZRQVTIWSQNRRVKSKK
Bip	ZVSALK
PPR3	ZPPRPPRPPR
PPR4	ZPPRPPRPPRPPR
AIP6	ZRLRWR
DPV15b	ZGAYDLRRRERQSRLRRRERQSR
TAT	ZRKKRRQRRR
Penetratin	ZRQIKIWFQNRRMKWKK
R9	ZRRRRRRRRR
HoxA13 serine1	ZRSVTIWFQSRRVKEKK
KRVK TP10	ZKRVKAGYLLGKINLKALAALAKKIL
TP10 KRVK	ZAGYLLGKINLKALAALAKKILKRVK
SV40 TP10	ZPKKKRKVAGYLLGKINLKALAALAKKIL

*Z refers to 4-pentynoyl. "**C**" refers to cysteines that are linked with decafluorobiphenyl. "*C*" refers to cysteines that are linked with 1,3,5-trisbromomethylbenzene. Module 4 included fifty CPPs, including a mixture of chimeric peptides, cyclic peptides, and bicyclic peptides that we have previously reported to improve PMO delivery.

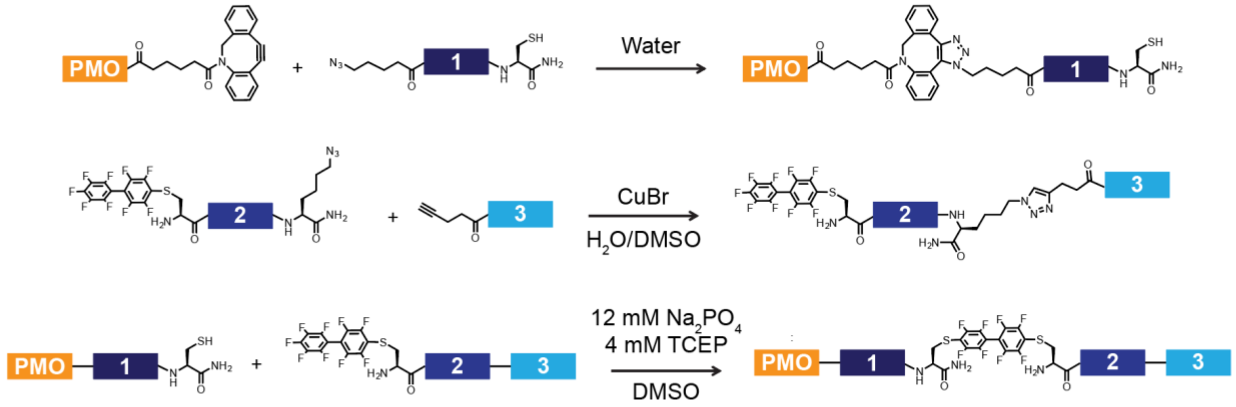


Figure 18. Synthesis of modular library.

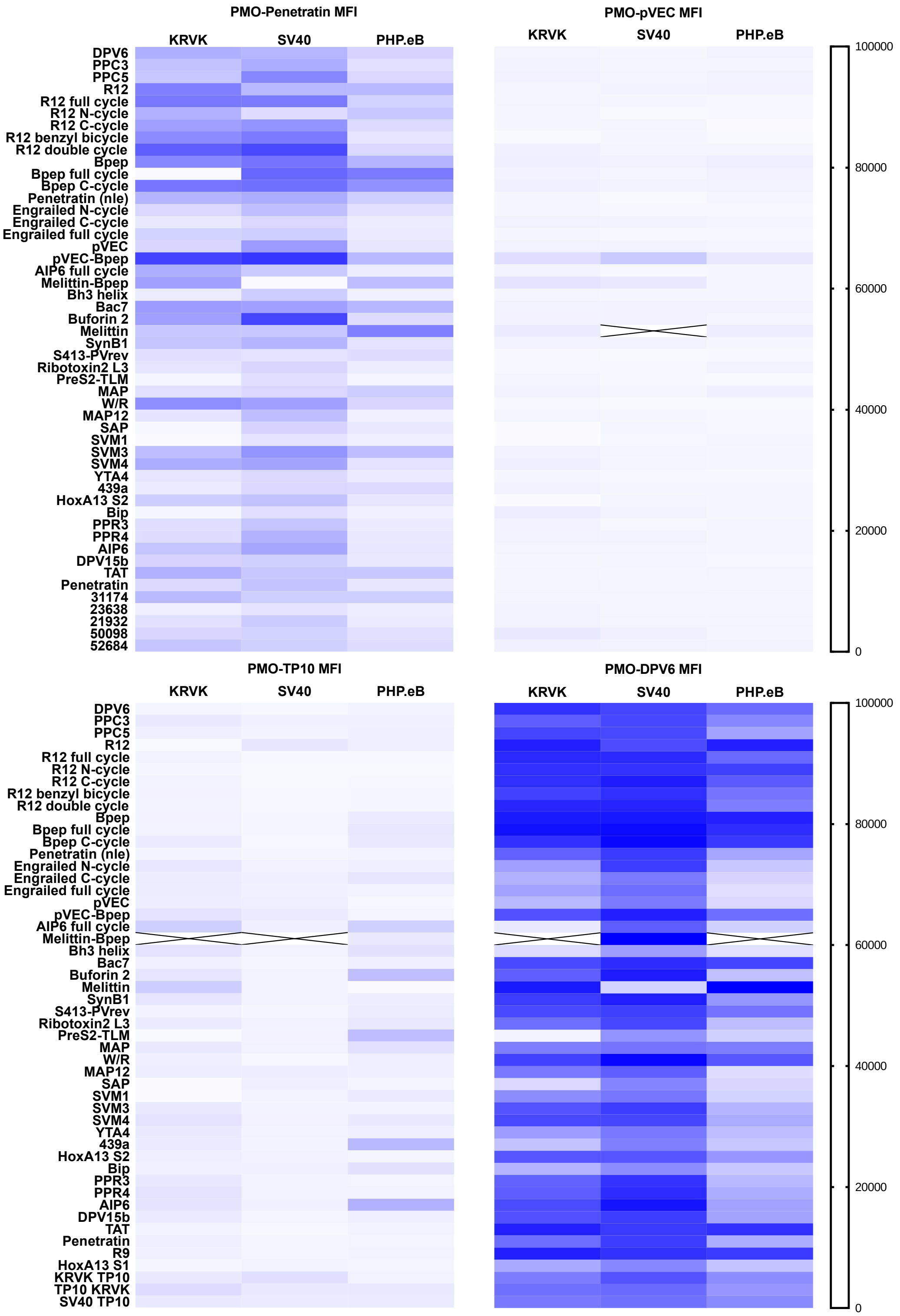


Figure 19. Mean fluorescence intensity of 600-member library. The heat maps show the mean fluorescence intensity of the 600 constructs tested in the HeLa-654 assay (n=1 replicate well). Boxes marked with an "X" are constructs in which the gated cell count was zero. The complete list of sequences and activities can be found in Supplemental Table 1.

Table 13. List of Mach peptides.*

		Fold over PMO	% Arg	PPMO MW	Net charge
Mach 1	ALKBRSAAKAVRWPKKKIKQASK KVAKYALXXXRKKKAASKXWLQ LHWPRW	45	8	12,645	18
Mach 2	PPLRNAKKKNLKNNLKMDPKFTK KVKQGALKLNRRKKNRGPKGPX KHWTT	27	8	12,499	18
Mach 3	QKKRKSKANKKNWPKGKLSIHAK DYKQGPKAKXRKQRXR	39	10	11,324	17
Mach 4	KKGKKQNKKKHRWPKKKVPQPK KMFKQGABXRX	25	6	10,622	16
Mach 5	AKKKIAKAKKHRGPNBGIHAPVS KIKDPLKXXX	3	8	10,222	11
Mach 6	ALKBRSAAKAVRWPKKAIKQASK KVAKYALKXXRKKKAASKXWLQ LHWPRW	43	8	12,603	18
Mach 7	XKHPXAVQBAARAWKVPAAALW KKKRLKKSSKQKKKWLWKARSA XKYXRLI	36	8	12,645	18
Mach 8	BKGKNLLAKIRRGPNGGNBQGSQ GYLLYLLXRXRRQRXXYPWWRX KHXRWXXRXRGHXRRRRQXLKP DRXRGGKGSVS	39	21	15,929	22
Mach 9	KKKKNLNBKSRRGPNGGALQPSQ GYLQPLNXRXRRQRXXYPWWRX KHXRWRXRYHXRRRRQXLKPG	38	21	14,845	22
Mach 11	TSNLKLHLAPPVKKKALKKPLYK AKKKKKVVSPTWXTDQEW	4	0	11,423	11
Mach 12	KGGKNLAKKIRRGPNGGALQPSQ GYLLYLBXRXRRQRXXGPXWRX KHXRWXXXXXRPTHXRRRRQXL **C**PGRXRP**C**RGSVS	40	20	16,285	22
Mach 13	AKKKKLGBKALRWPNGK**C**PQPK EK**C**PKYLLGRXRRKRXRYPWWR XKHRRW	30	18	13,228	20

*Peptide 10 was found to degrade in solution, so its analysis was discontinued. ‘X’ is 6-amino hexanoic acid, and ‘B’ is β-alanine. **C** residues are linked through decafluorobiphenyl.

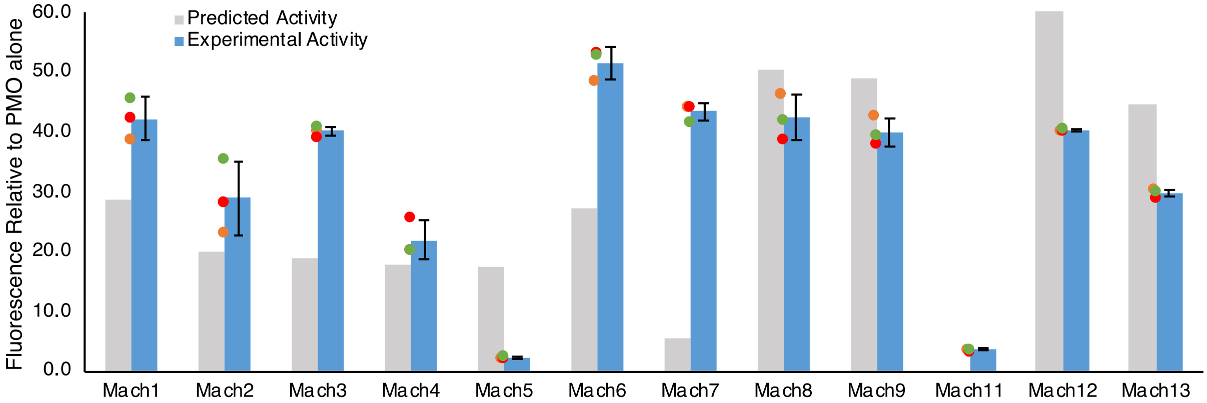


Figure 20. Experimental vs Predicted activity of Mach peptides. Mach peptides enhance delivery of PMO by 40-50 fold as determined by the HeLa 654 assay. Experimental activity (blue) is comparable to predicted activity (grey). Mach12 predicted activity is off the scale, at 140. Each bar represents group mean ± SD, N = 3.

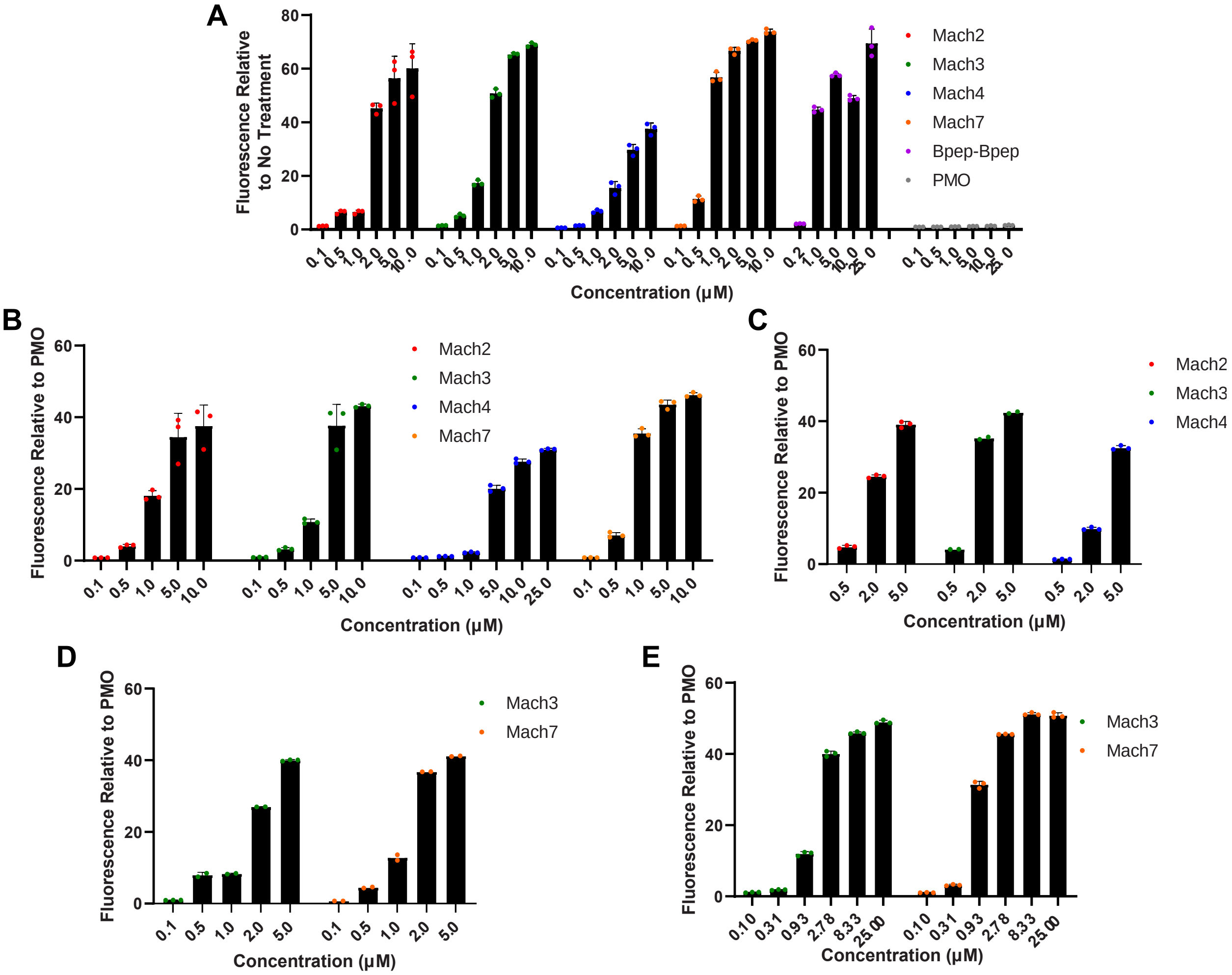


Figure 21. Dose-response in HeLa 654 cells (Activity). PMO-Mach constructs elicit a dose-dependent increase in EGFP fluorescence. (A) Shown is GFP fluorescence relative to the no treatment condition, including PMO alone. Also included here is chimera PMO-Bpep-Bpep, a previously reported high-performing PMO-peptide. Each bar represents group mean ± SD, N = 3 distinct samples. (B-E) Shown is GFP fluorescence relative to the PMO alone condition. PMO-Mach peptides were tested at different concentration ranges with similar results. Each bar represents group mean ± SD, N = 3 distinct samples, except for (C) Mach3 and (D) Mach7 in which bar represents group mean, N = 2.

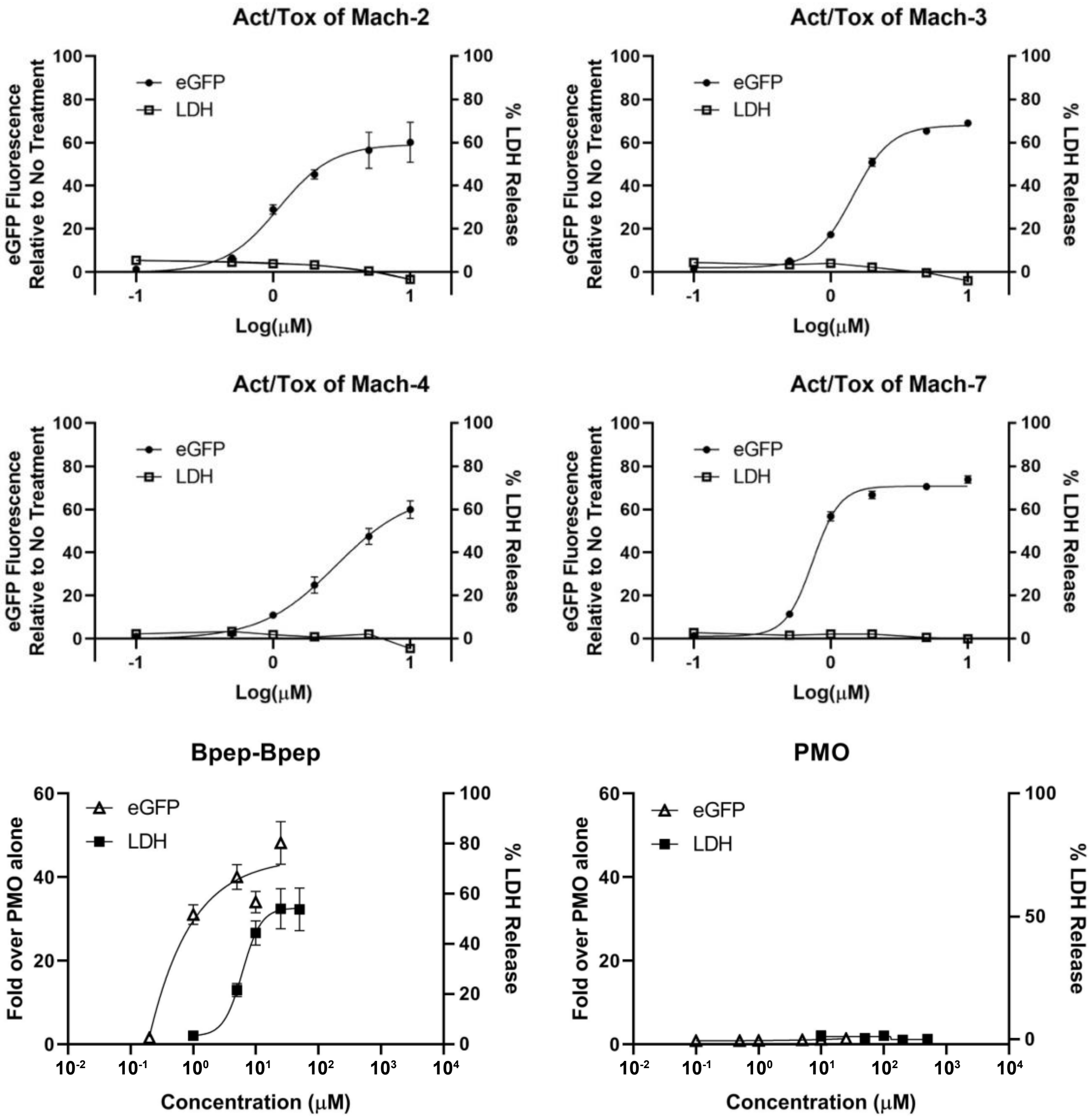


Figure 22. Dose response curves corresponding to activity and toxicity in HeLa 654. HeLa 654 cells were treated with varying concentrations of PMO-Mach constructs or PMO-Bpep-Bpep for 22 h. RPTEC cells were also treated with the same concentrations of PMO-Bpep-Bpep, as in Figure 23. Following treatment, cell supernatant was removed and tested for LDH release, reported as % LDH release relative to full lysis control (LDH, square). Toxicity is compared to activity (EGFP, triangle) in HeLa 654 from Figure 21. Each point represents group mean ± SD, N = 3 distinct samples.

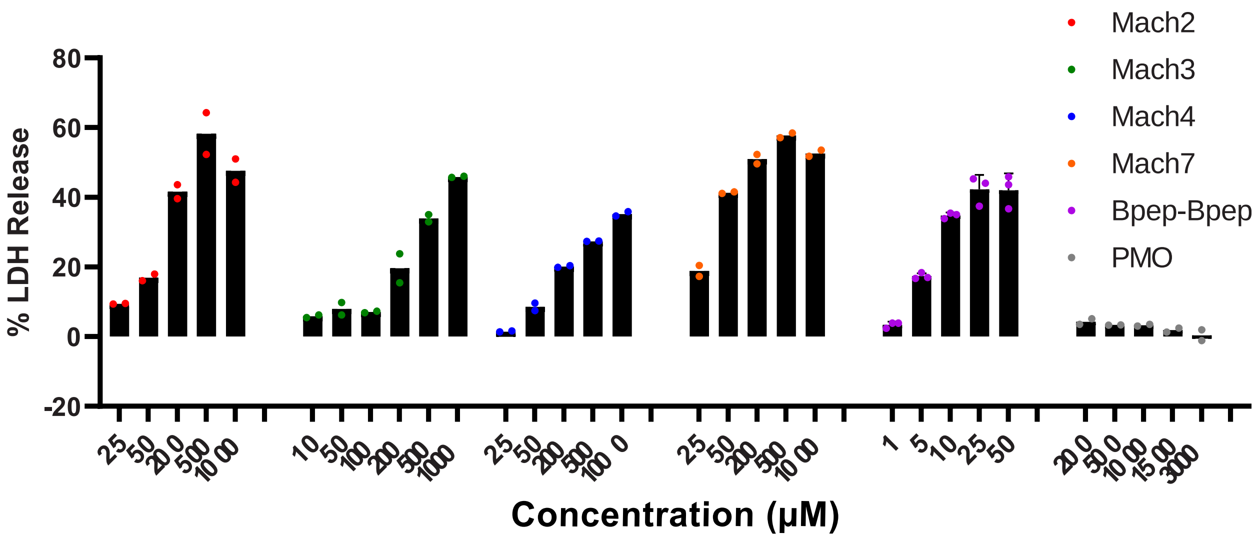


Figure 23. Dose-response in RPTEC (Toxicity). PMO-Mach constructs elicit a dose-dependent increase in membrane toxicity as measured by LDH release assay. Data shown here for Mach3, 4, and 7 are what is shown in main text Figure 4A-C. LC50 of PMO-Mach constructs are between 100-200 µM, in contrast to PMO-Bpep-Bpep, which has a significantly lower LC50 near 10 µM. Each bar represents group mean ± SD, N = 2, except for Bpep-Bpep N = 3 distinct samples.

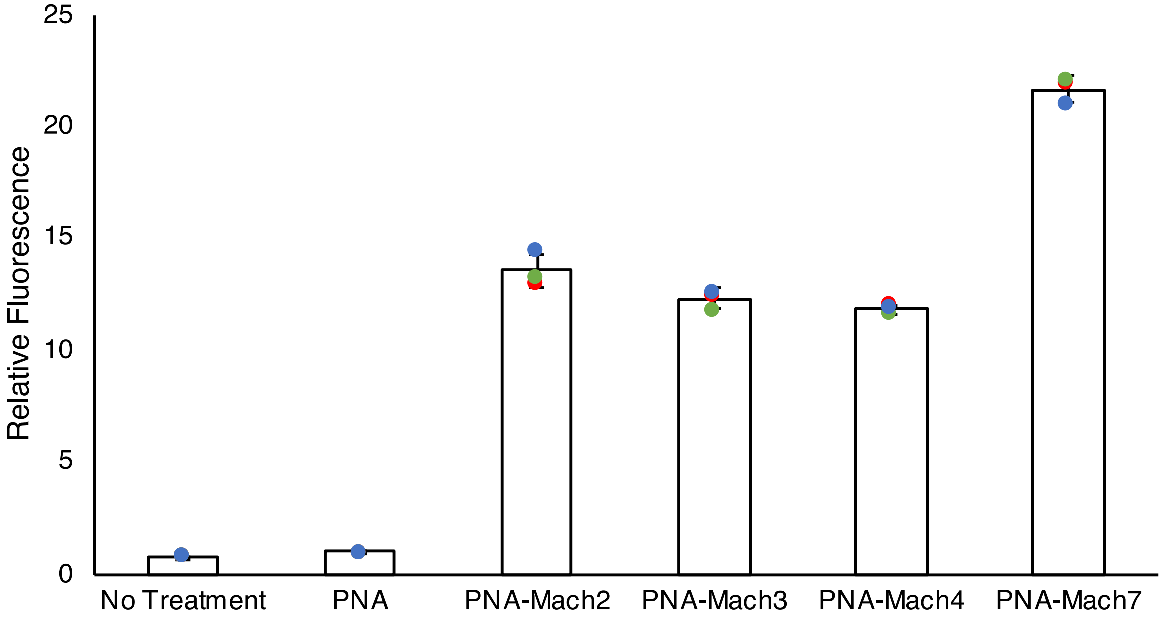


Figure 24. Mach peptides enhance delivery of peptide nucleic acid (PNA). PNA-Mach constructs were evaluated at 5 µM in the HeLa EGFP 654 assay. Each bar represents group mean ± SD, N = 3 distinct samples.

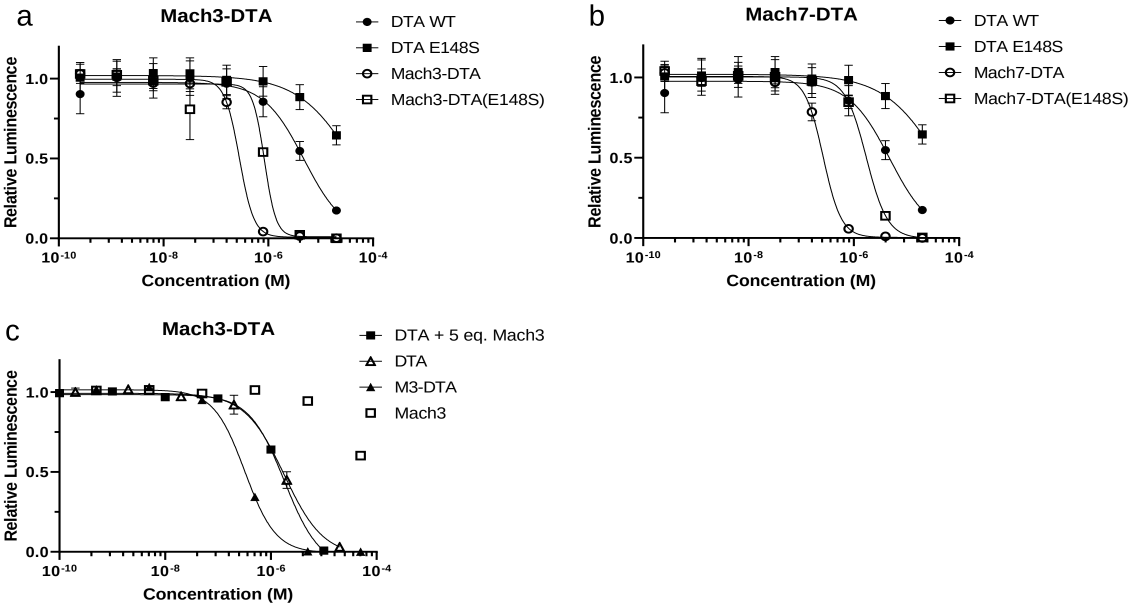


Figure 25. Mach-DTA conjugates produce dose-dependent toxicity in HeLa cells. Attachment of WT DTA to (a) Mach3 or (b) Mach7 produces significantly greater activity than attachment to DTA (E148S) which has 300-fold lower activity than wild-type. (c) Covalent attachment of Mach3 is required for DTA constructs to be delivered to the cytosol. DTA alone has the same toxicity as DTA co-incubated with 5 equivalents of Mach3 peptide. Each point represents group mean ± SD, N = 3 distinct samples, with the exception of Mach3-DTA(E148S) and Mach7-DTA(E148S) in which n = 2. Experiments, excluding Mach3-DTA(E148S) and Mach7-DTA(E148S), were repeated at slightly different concentrations with similar results.

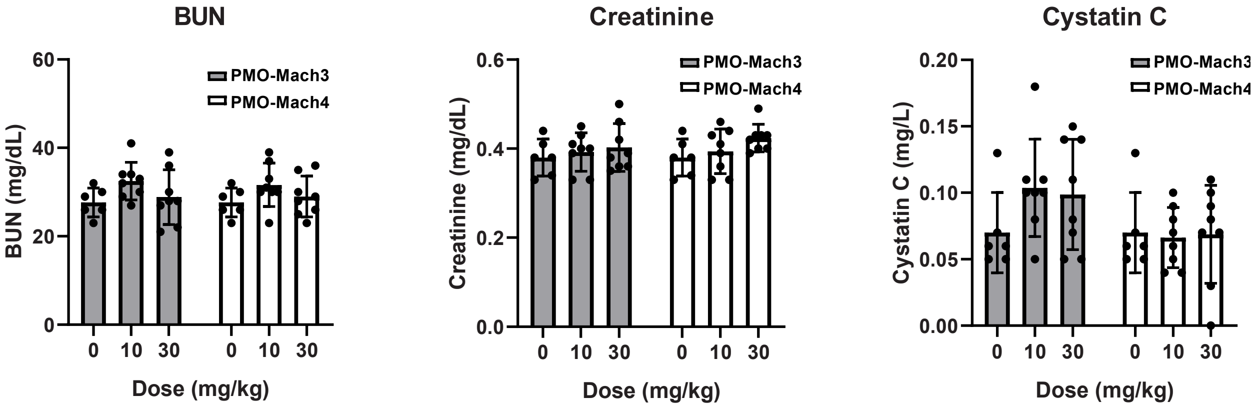
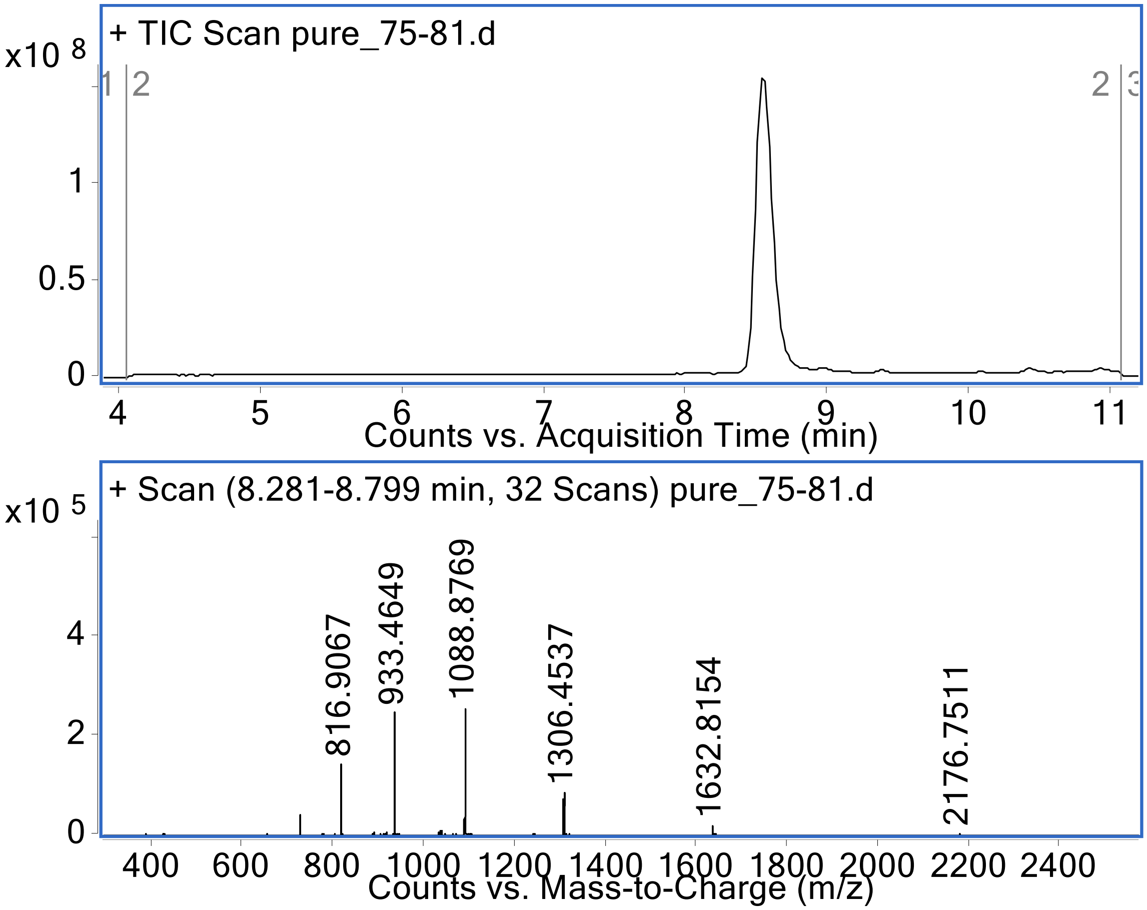


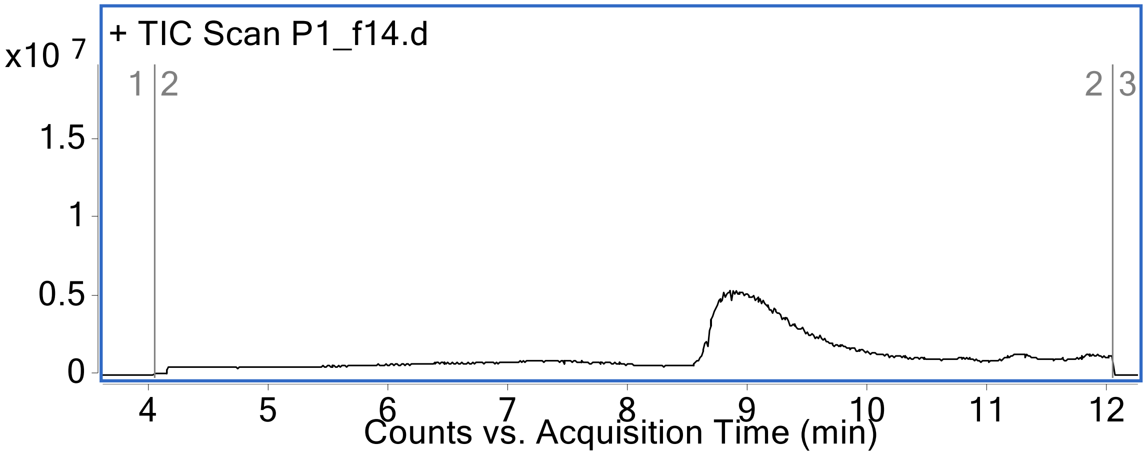
Figure 26. PMO-Mach constructs do not induce kidney toxicity in mice. In EGFP 654 mice, levels of (a) blood urea nitrogen (BUN), (b) creatinine, and (c) cystatin C remained unchanged. Each bar represents group mean ± SD. Saline (n = 6), Mach3 and Mach4 at 5 mg/kg (n = 4), all other n = 8 mice. A two-tailed Mann-Whitney U test showed no significant difference between groups.

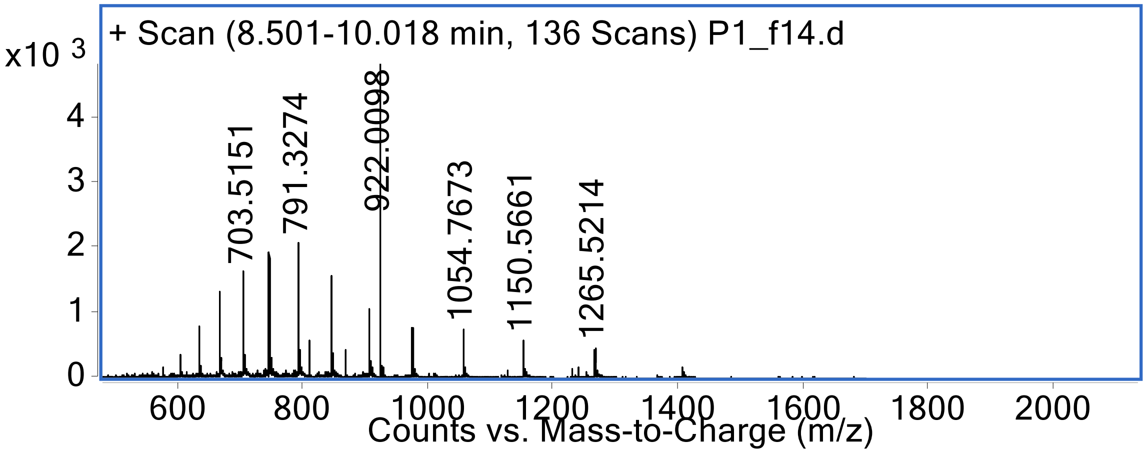
Appendix 1 LC-MS Characterization

PMO-DBCO (Method A)
Mass Expected: 6527.9 Da
Mass Observed: 6527.9 Da
PMO sequence: GCT ATT ACC TTA ACC CAG

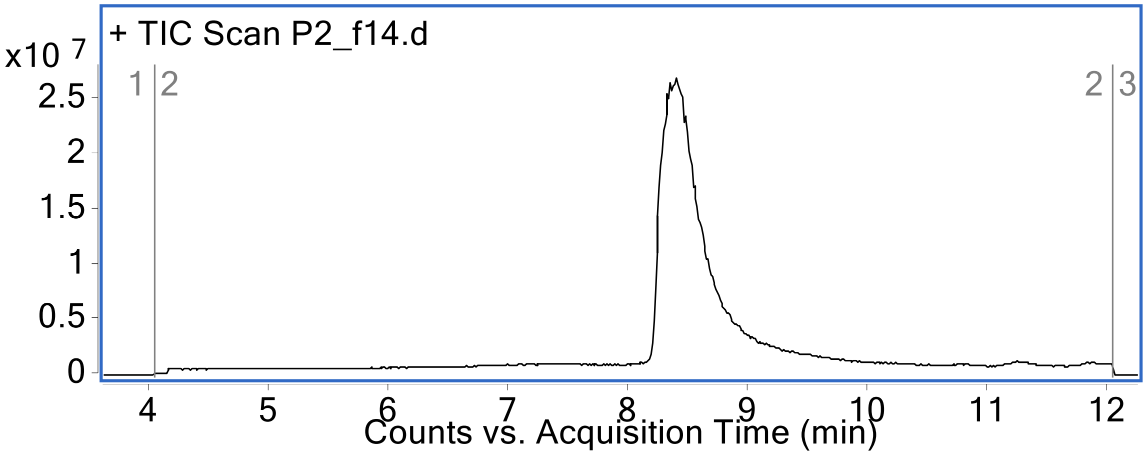


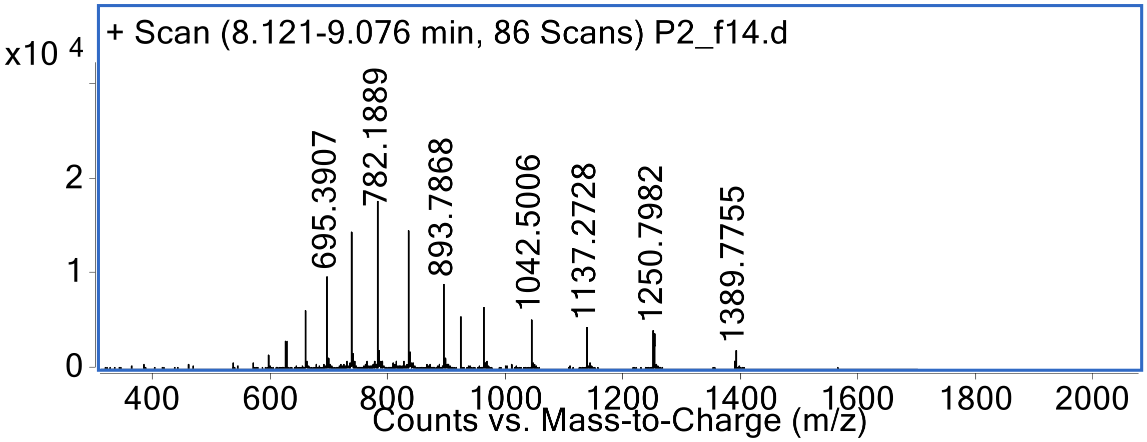
PMO-Mach1 (Method B)
Mass Expected: 12645.4 Da
Mass Observed: 12645.6 Da
Peptide sequence:
ALKBRSAAKAVRWPKKKIKQASKKVAKYALXXXRKKKAASKXWLQLHWPRW



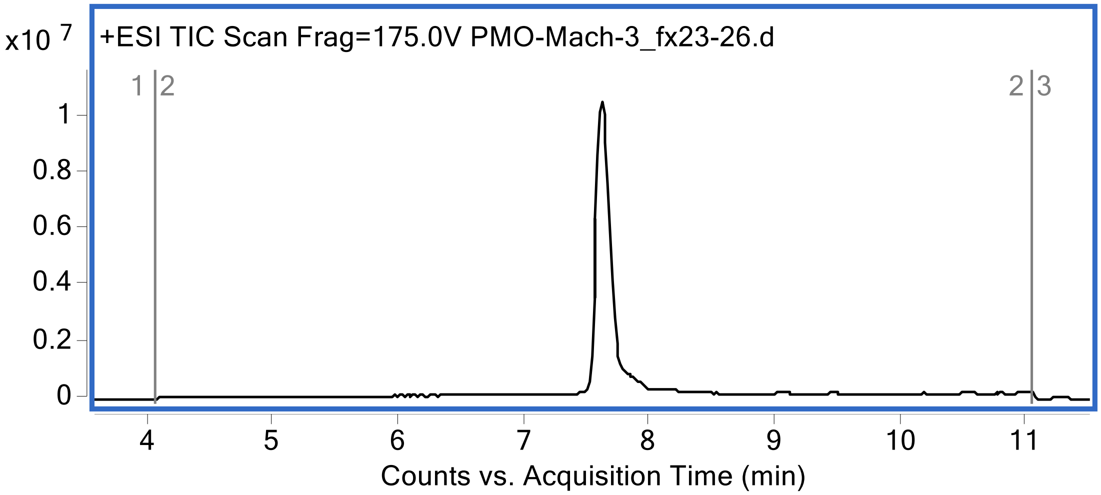


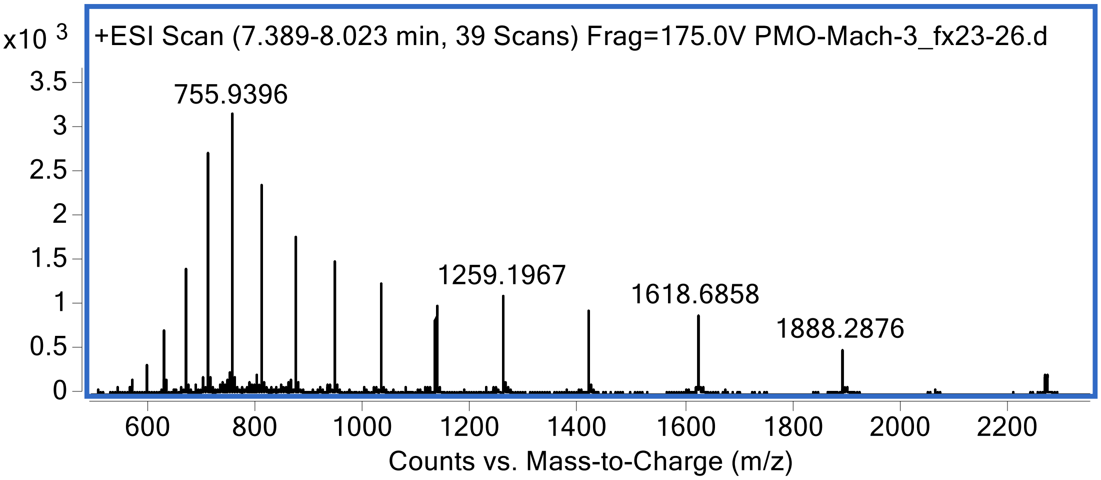
PMO-Mach2 (Method B)
Mass Expected: 12499.1 Da
Mass Observed: 12499.2 Da
Peptide sequence:
PPLRNAKKKNLKNNLKMDPKFTKKVKQGALKLNRRKKNRGPKGPXKHWTT



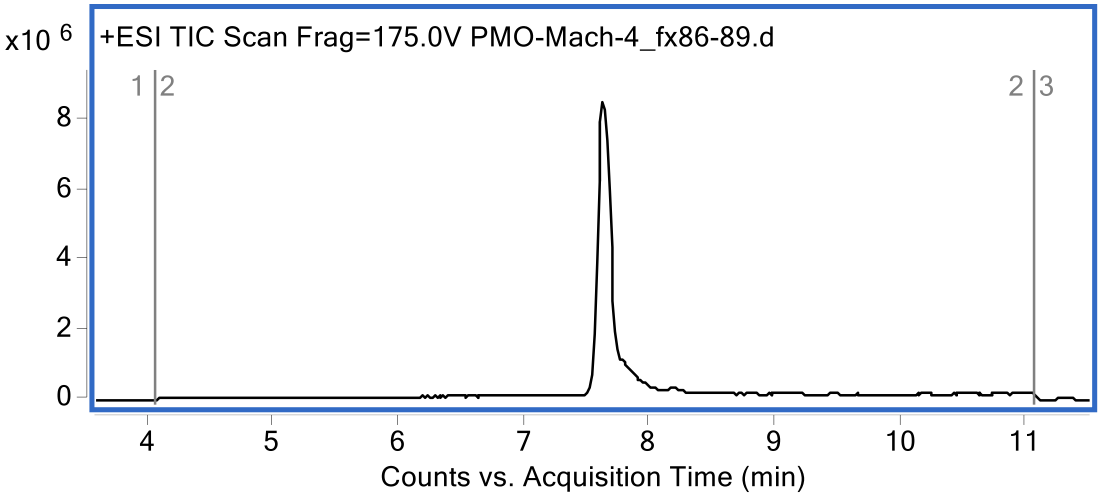


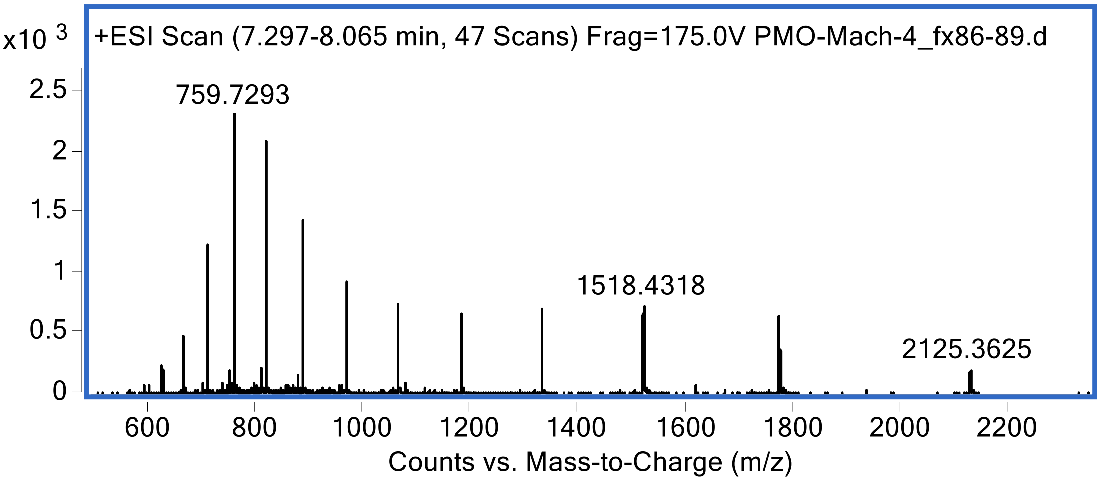
PMO-Mach3 (Method A)
Mass Expected: 11323.6 Da
Mass Observed: 11324.3 Da
Peptide sequence: QKKRKSKANKKNWPKGKLSIHAKDYKQGPKAKXRKQRXR



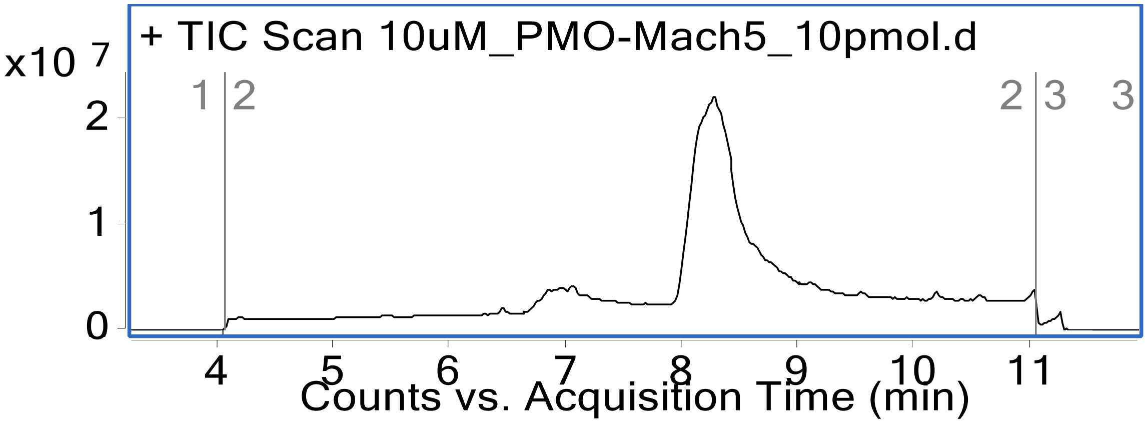


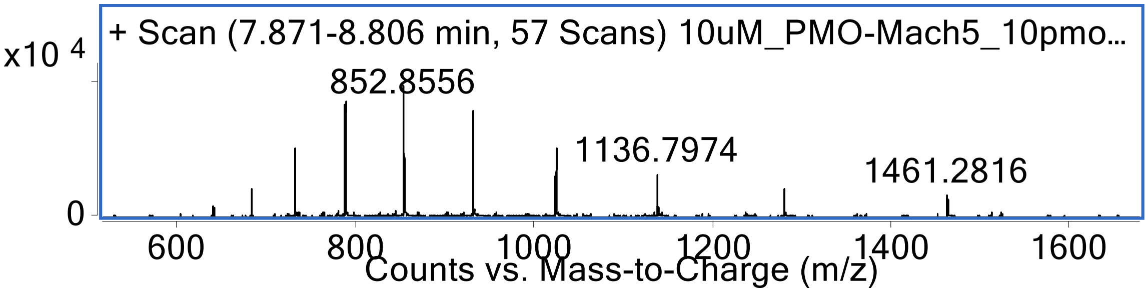
PMO-Mach4 (Method A)
Mass Expected: 10622.0 Da
Mass Observed: 10622.5 Da
Peptide sequence: KKGKKQNKKKHRWPKKKVPQPKKMFKQGABXRX



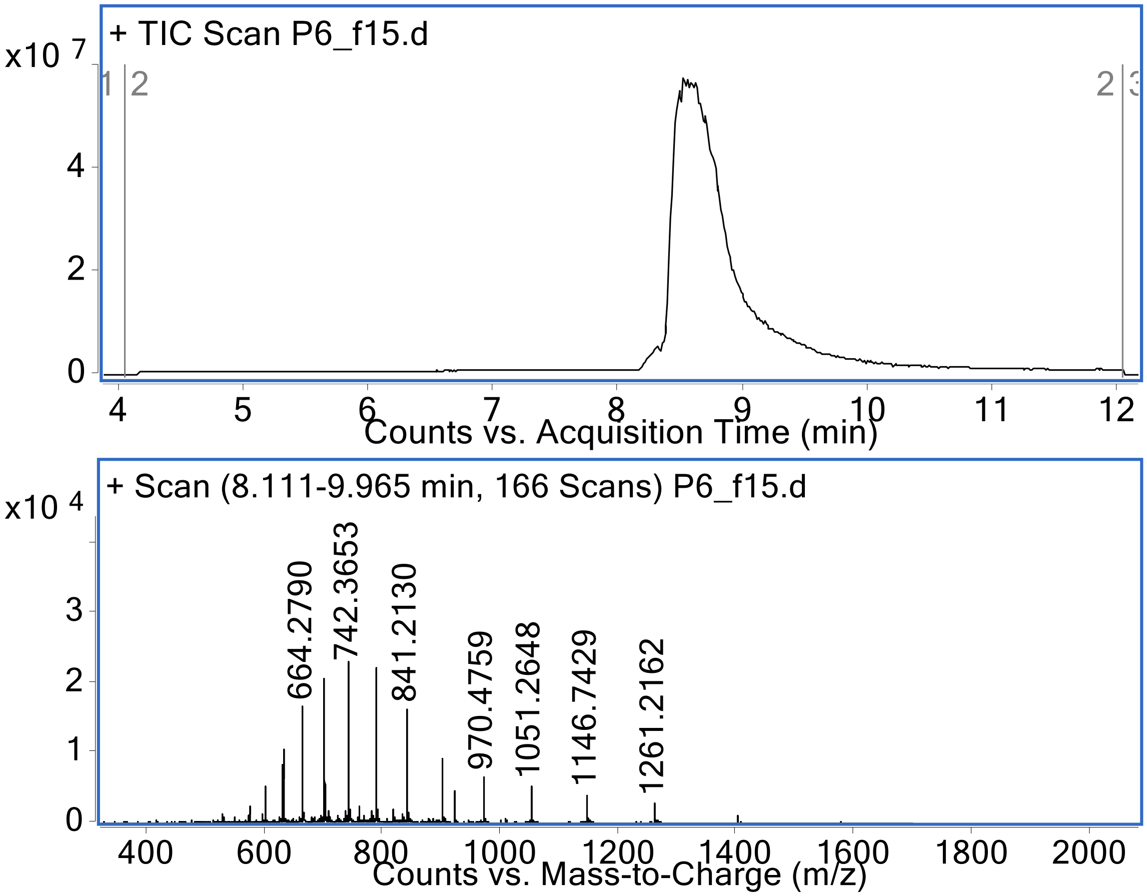


PMO-Mach5 (Method A)
Mass Expected: 10222.5 Da
Mass Observed: 10222.5 Da
Peptide sequence: AKKKIAKAKKHRGPNBGIHAPVSKIKDPLKXXX

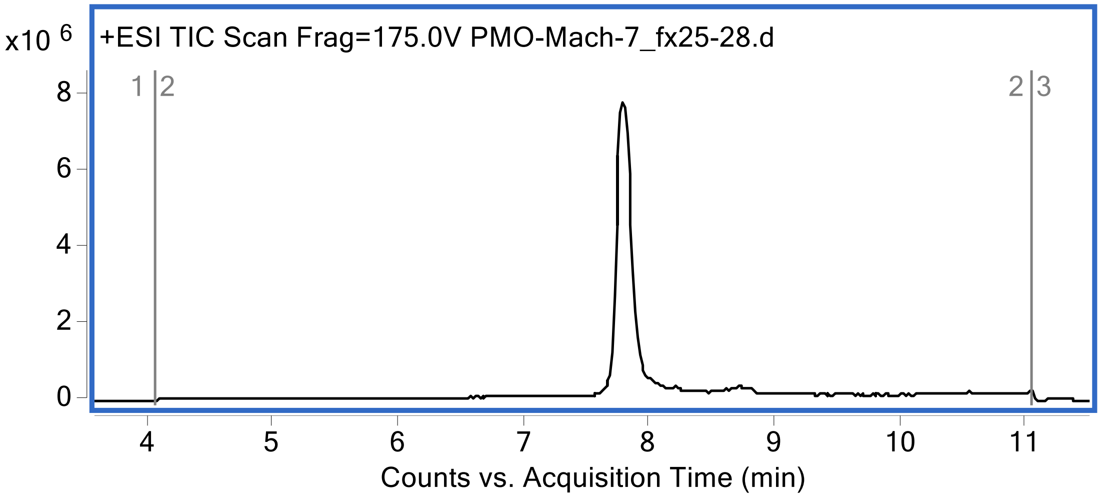


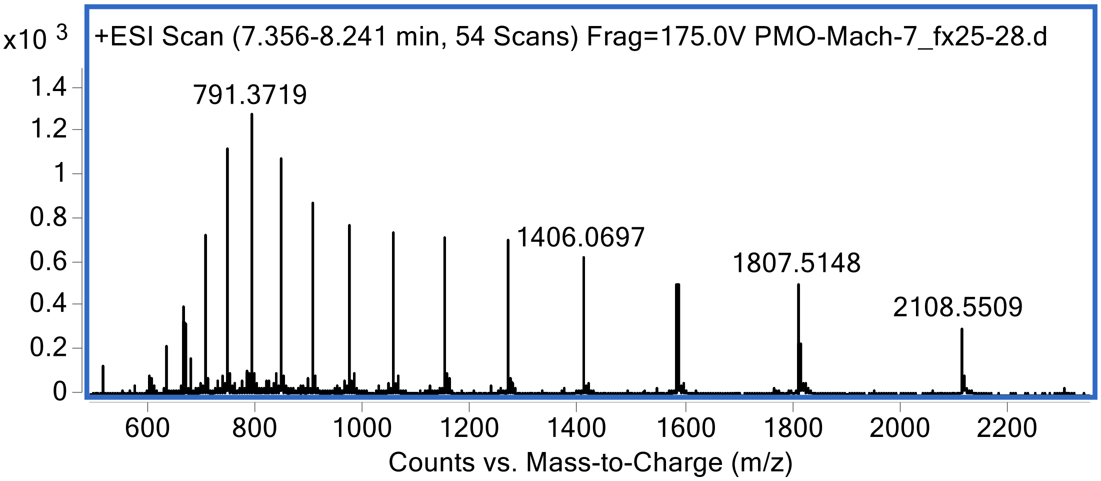


PMO-Mach6 (Method B)
Mass Expected: 12603.4 g/mol
Mass Observed: 12603.4 g/mol
Peptide sequence:
ALKBRSAAKAVRWPKKAIKQASKKVAKYALKXXRKKKAASKXWLQLHWPRW

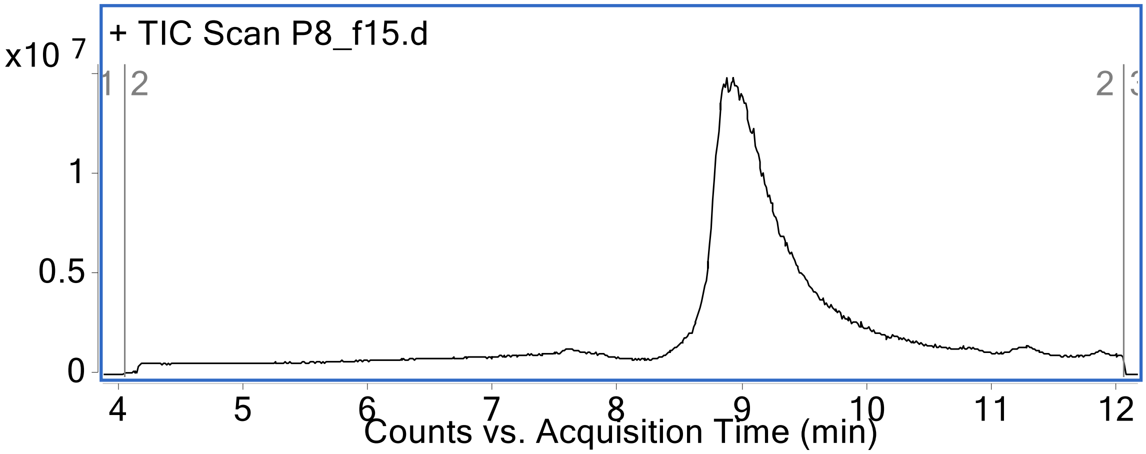


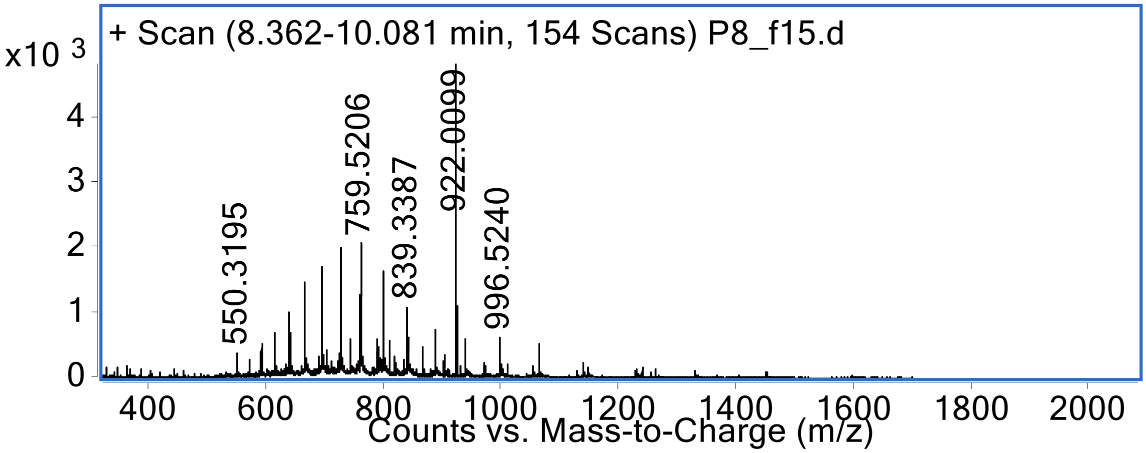
PMO-Mach7 (Method A)
Mass Expected: 12645.4 Da
Mass Observed: 12645.9 Da
Peptide sequence:
XKHPXAVQBAARAWKVPAAALWKKKRLKKSSKQKKKWLWKARSAXKYXRLI



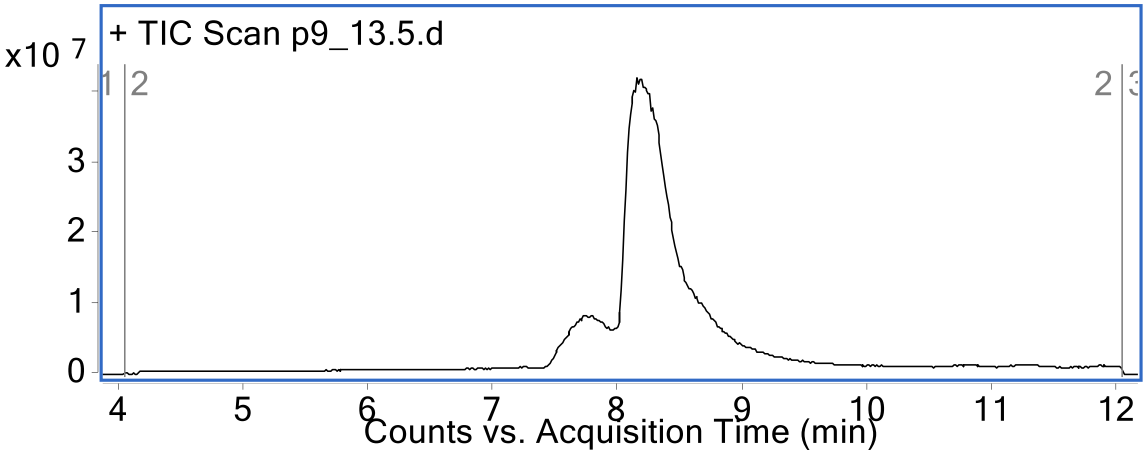


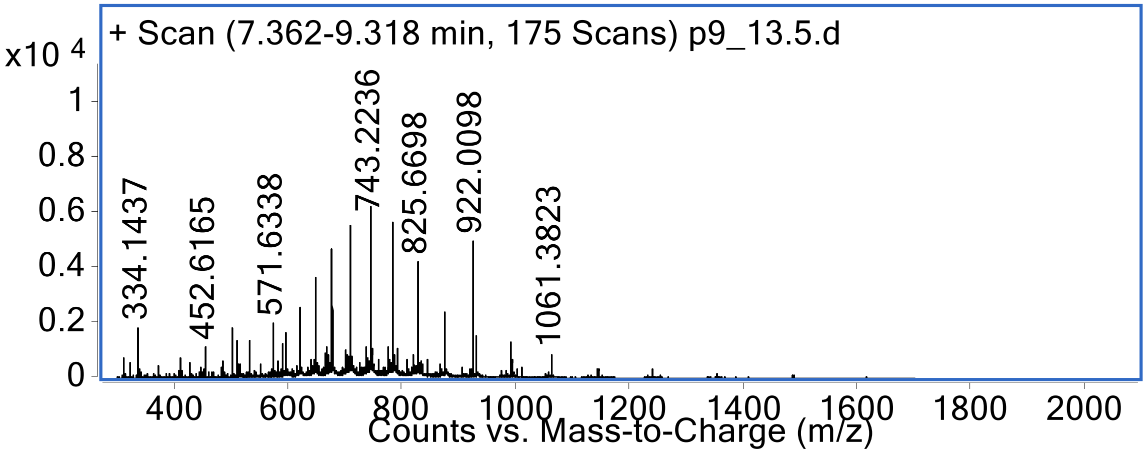
PMO-Mach8 (Method B)
Mass Expected: 15929.1 Da
Mass Observed: 15929.3 Da
Peptide sequence:
BKGKNLLAKIRRGPNGGNBQGSQGYLLYLLXRXRRQRXXYPWWRXKHXRWXXRXRG
HXRRRRQXLKPDRXRGGKGSVS



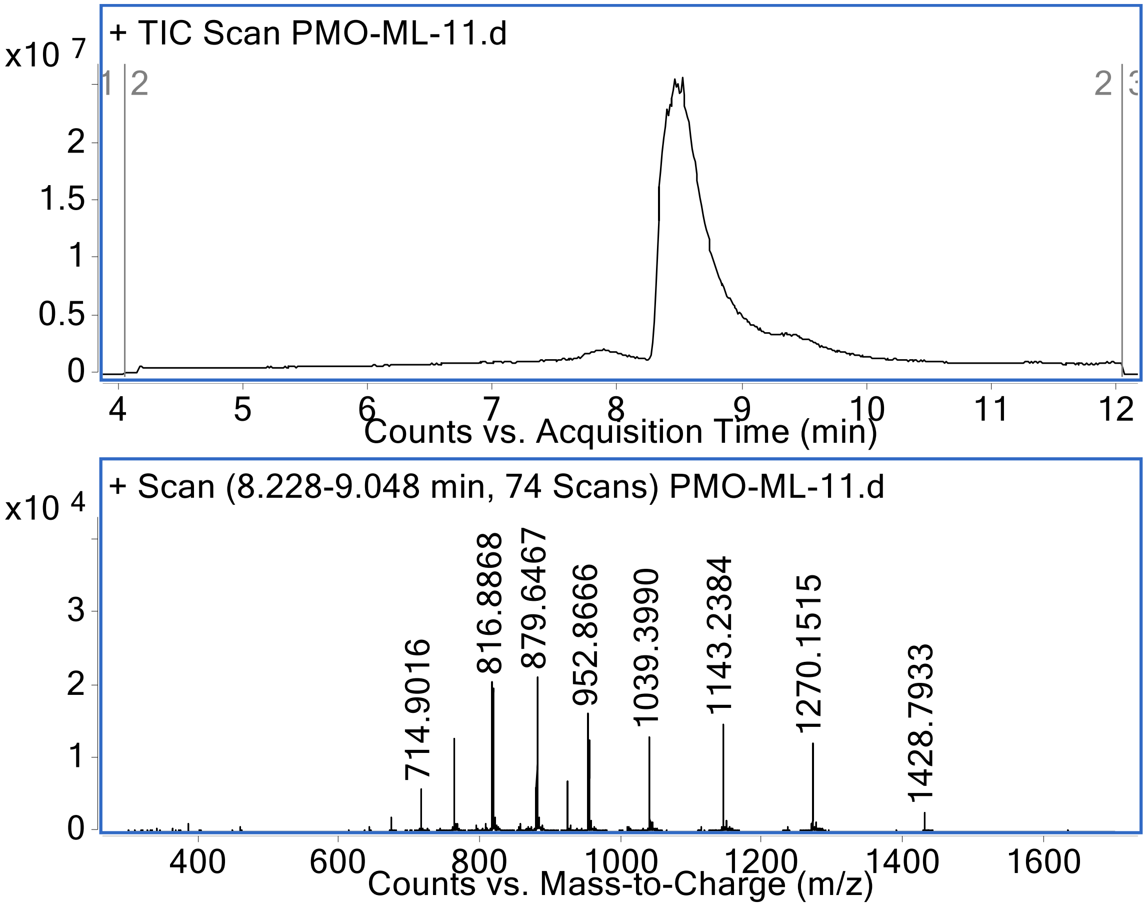


PMO-Mach9 (Method B)
Mass Expected: 14844.8 Da
Mass Observed: 14845.0 Da
Peptide sequence:
KKKKNLNBKSRRGPNGGALQPSQGYLQPLNXRXRRQRXXYPWWRXKHXRWRXRYH
XRRRRQXLKPG

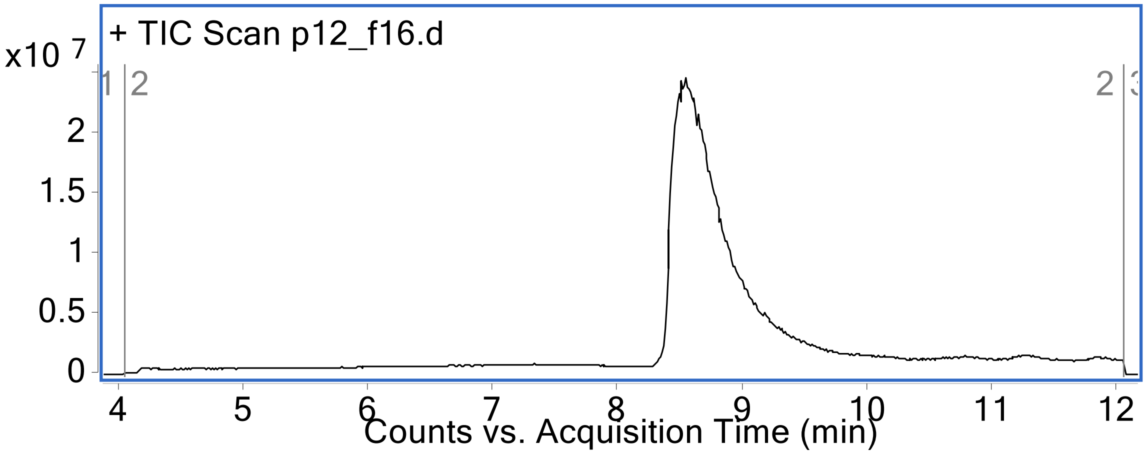


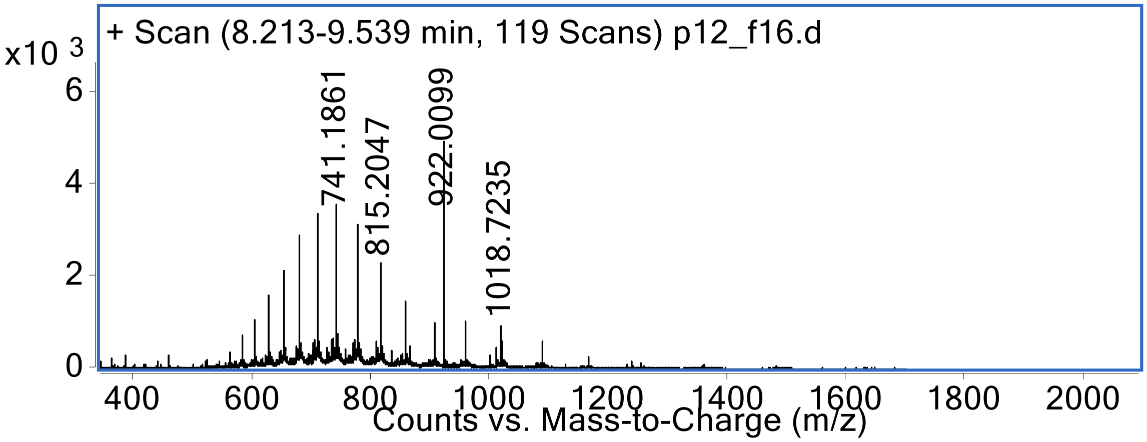


PMO-Mach11 (Method B)
Mass Expected: 11422.8 Da
Mass Observed: 11422.8 Da
Peptide sequence: TSNLKLHLAPPVKKKALKKPLYKAKKKKKVVSPTWXTDQEW

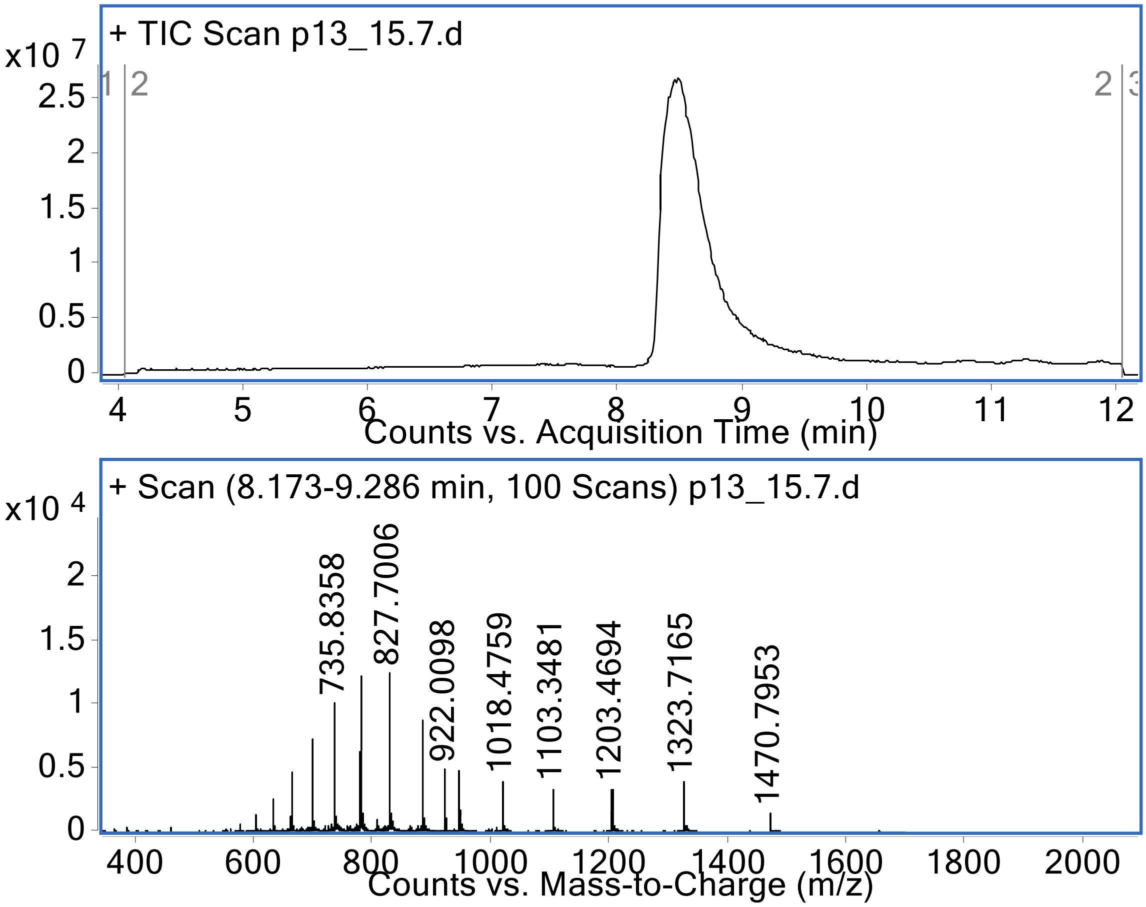


PMO-Mach12 (Method B)
Mass Expected: 16284.5 Da
Mass Observed: 16284.7 Da
Peptide sequence:
KGGKNLAKKIRRGPNGGALQPSQGYLLYLBXRXRRQRXXGPXWRXKHXRWXXXXXXR
PTHXRRRRQXL**C**PGRXRP**C**RGSVS

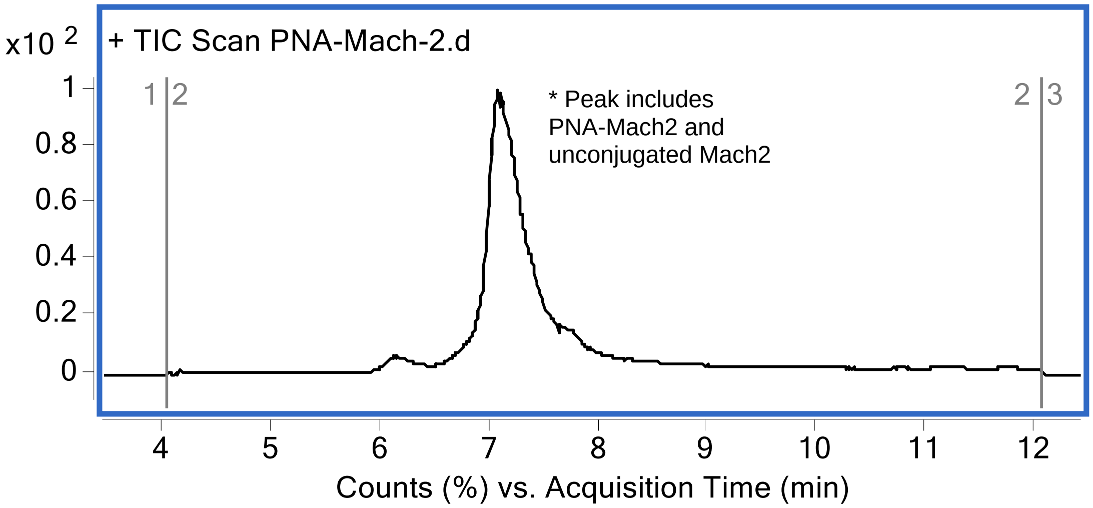




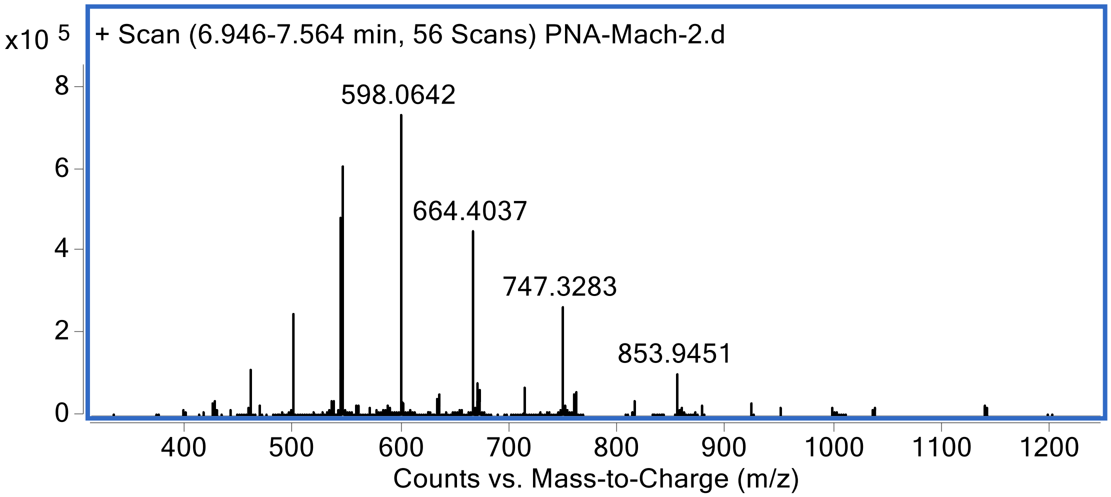
PMO-Mach13 (Method B)
Mass Expected: 13227.8 Da
Mass Observed: 13228.0 Da
Peptide sequence:
AKKKKLGBKALRWPNGK**C**PQPKEK**C**PKYLLGRXRRKRXRYPWWRXKHRRW



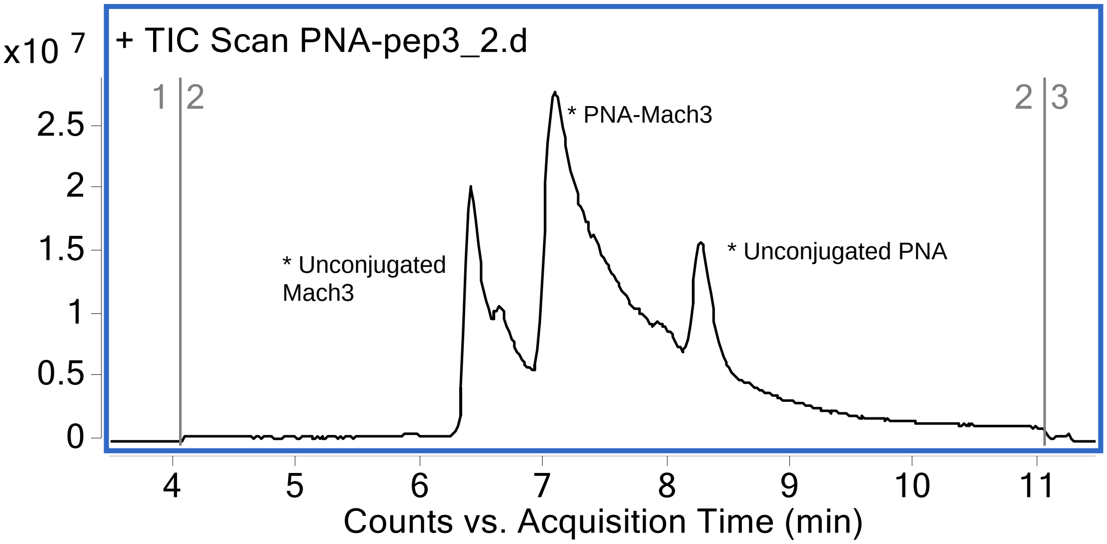
PNA-Mach2 (Method B)
Mass Expected: 11375.5 Da
Mass Observed: 11374.9 Da

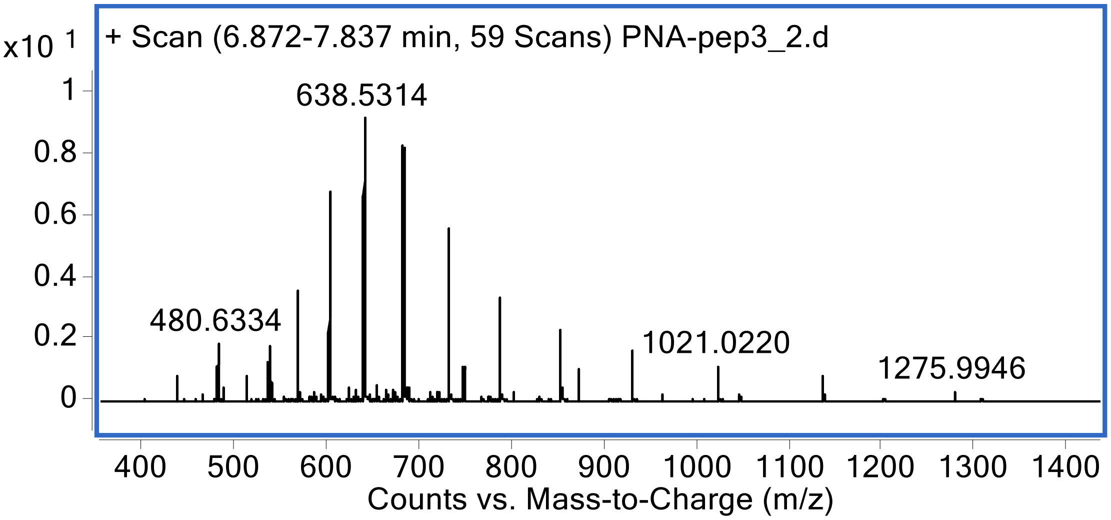


* Peak includes PNA-Mach2 and unconjugated Mach2

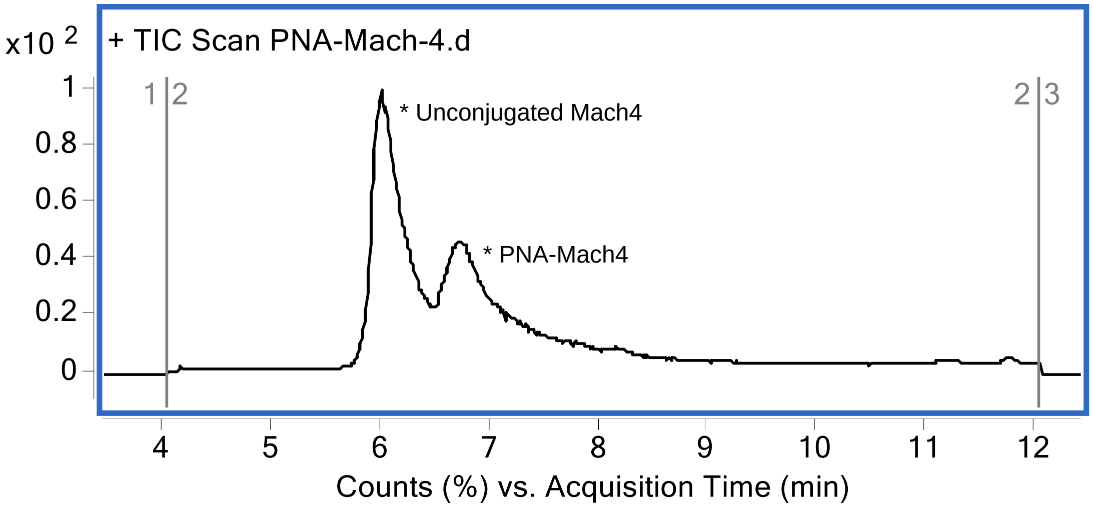


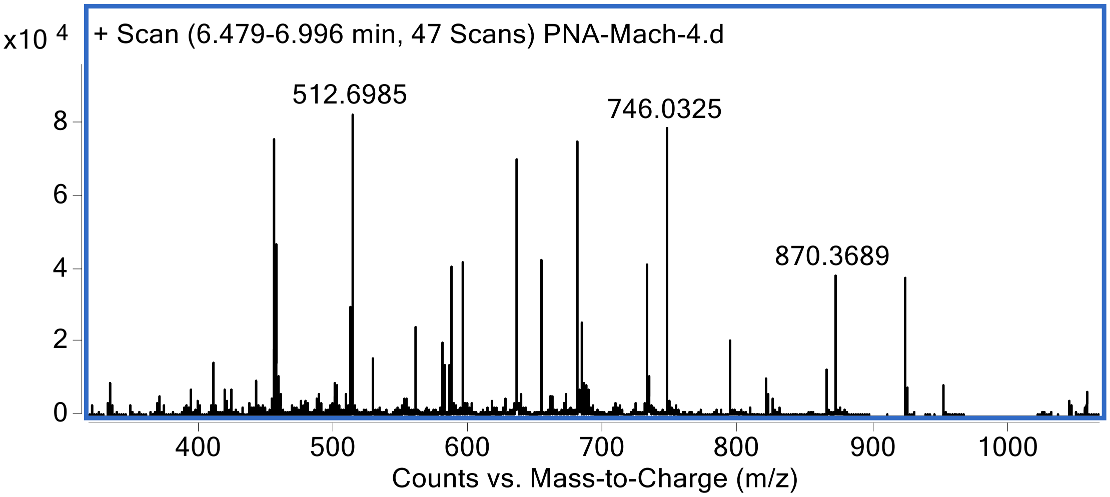
PNA-Mach3 (Method A)
Mass Expected: 10200.0 Da
Mass Observed: 10200.6 Da



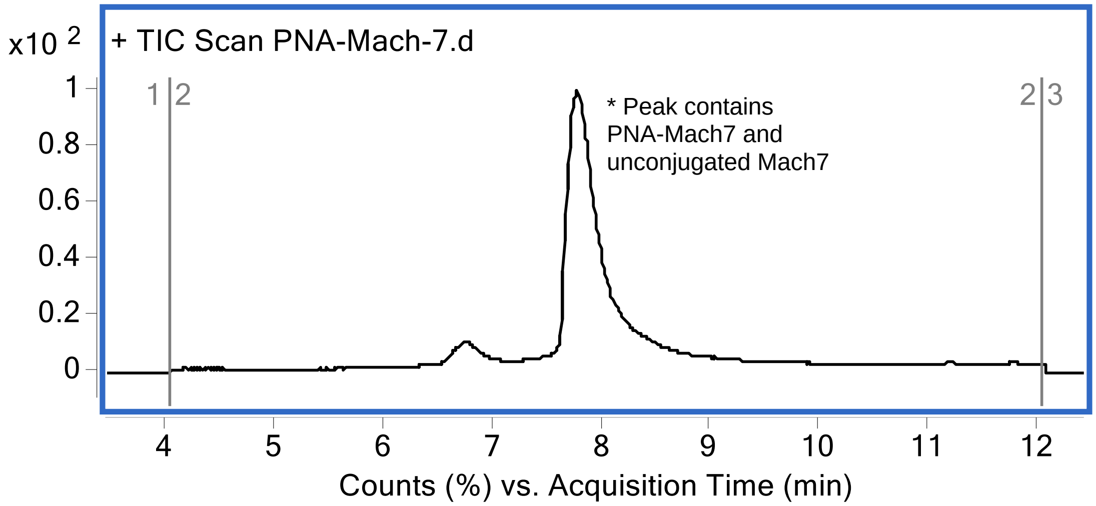


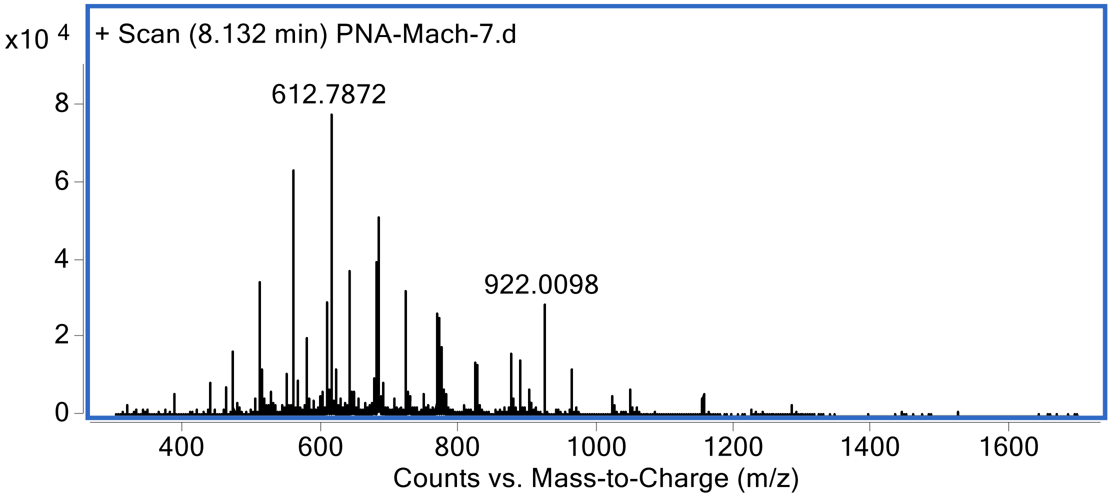
PNA-Mach4 (Method B)
Mass Expected: 9498.4 Da
Mass Observed: 9497.8 Da



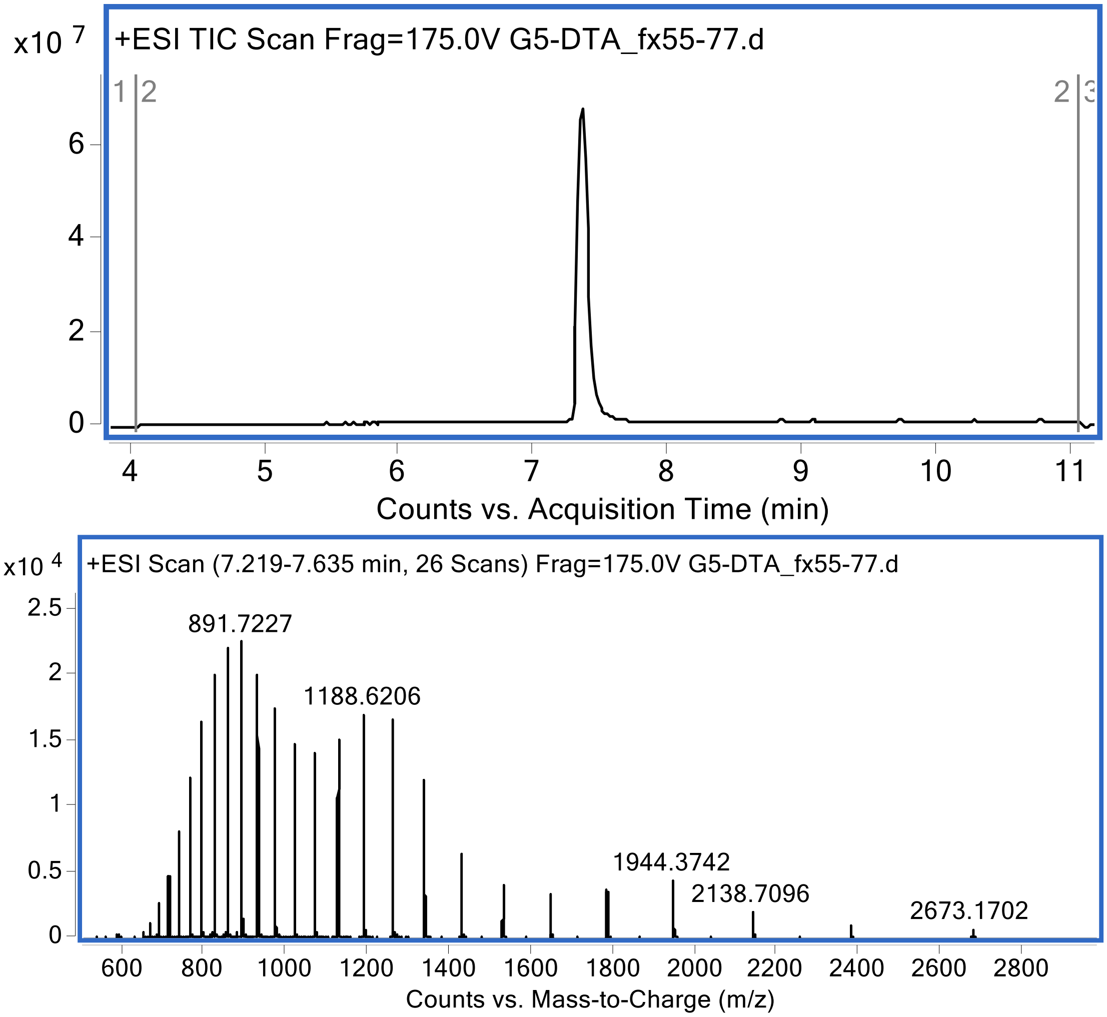


<u>PNA-Mach7 (Method B)</u>
Mass Expected: 11521.8 Da
Mass Observed: 11521.3 Da





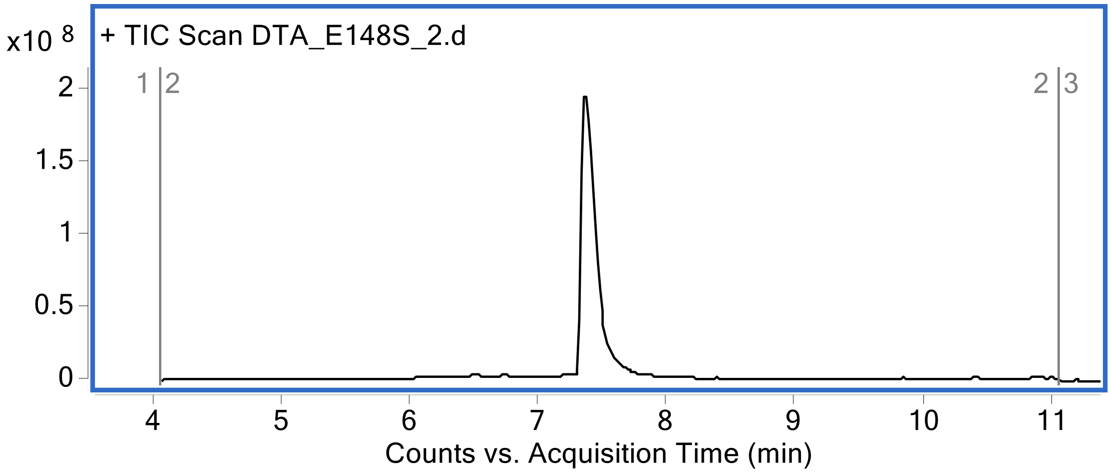
G5-DTA(C186S) (Method A)
Mass Expected: 21376.8 Da
Mass Observed: 21377.2 Da

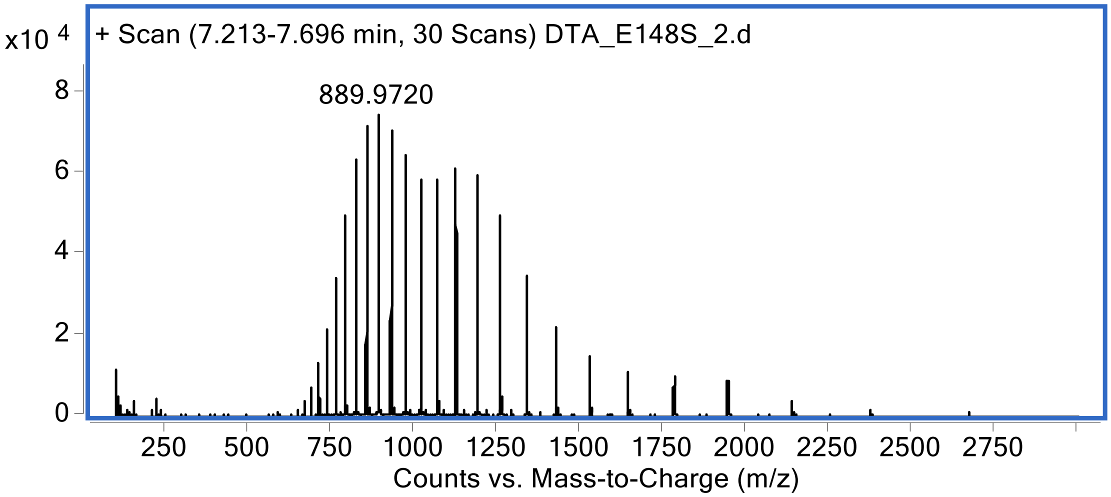


G5-DTA(C186S, E148S) (Method A)

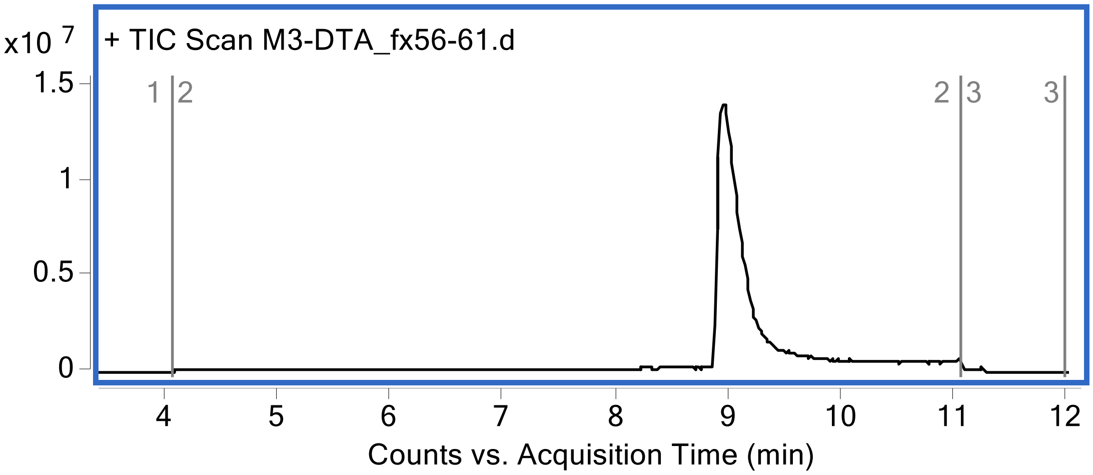
Mass Expected: 21334.6 Da

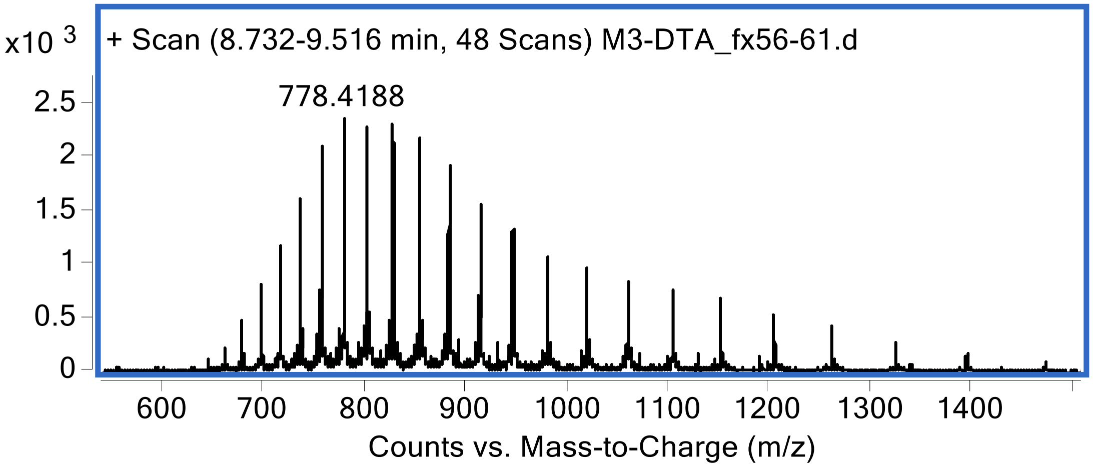
Mass Observed: 21335.3 Da



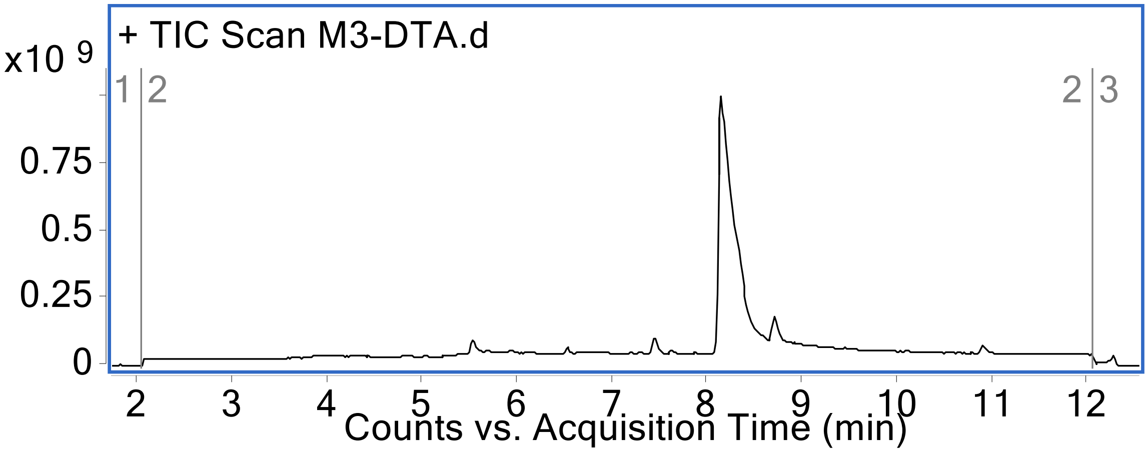


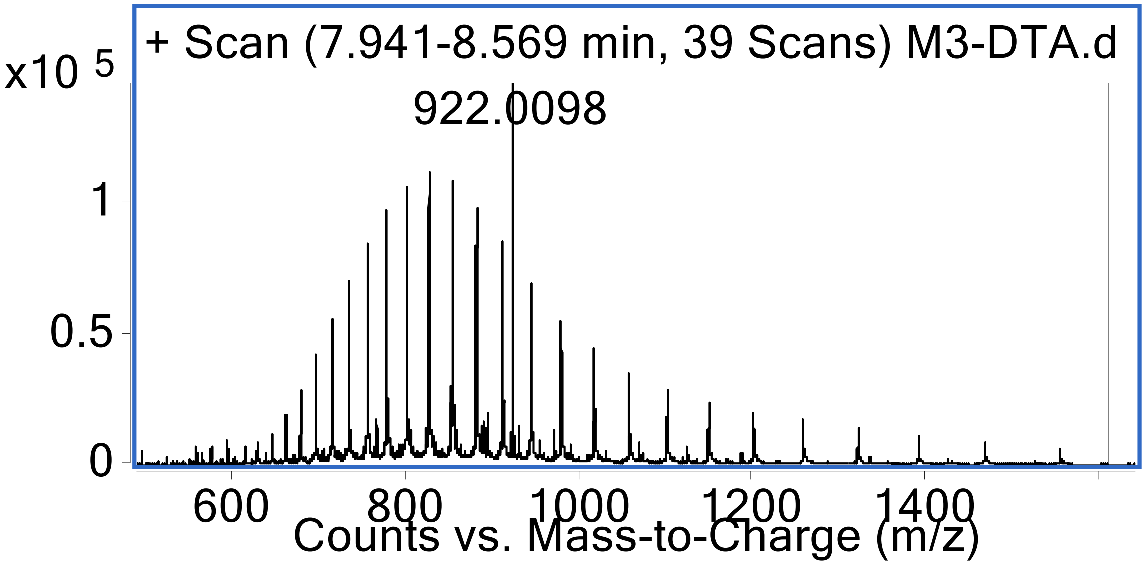
<u>Mach3-DTA(C186S) (Method A)</u>
Mass Expected: 26428.7 Da
Mass Observed: 26432.0 Da



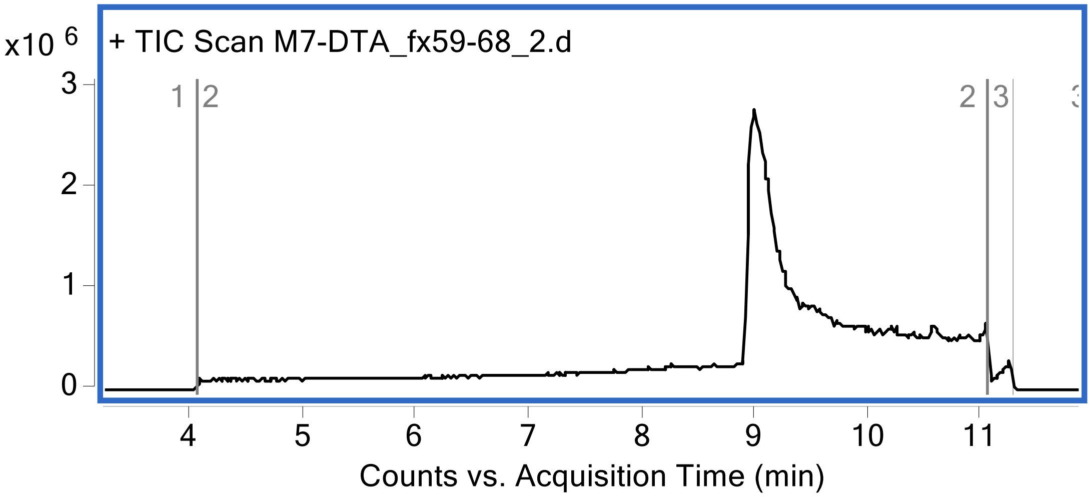


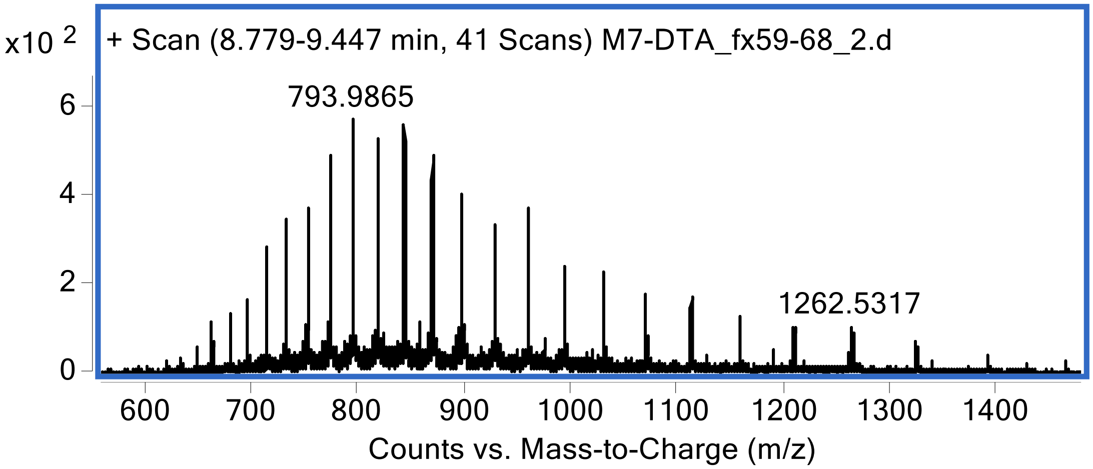
Mach3-DTA(C186S, E148S) (Method B)
Mass Expected: 26386.7 Da
Mass Observed: 26388.2 Da



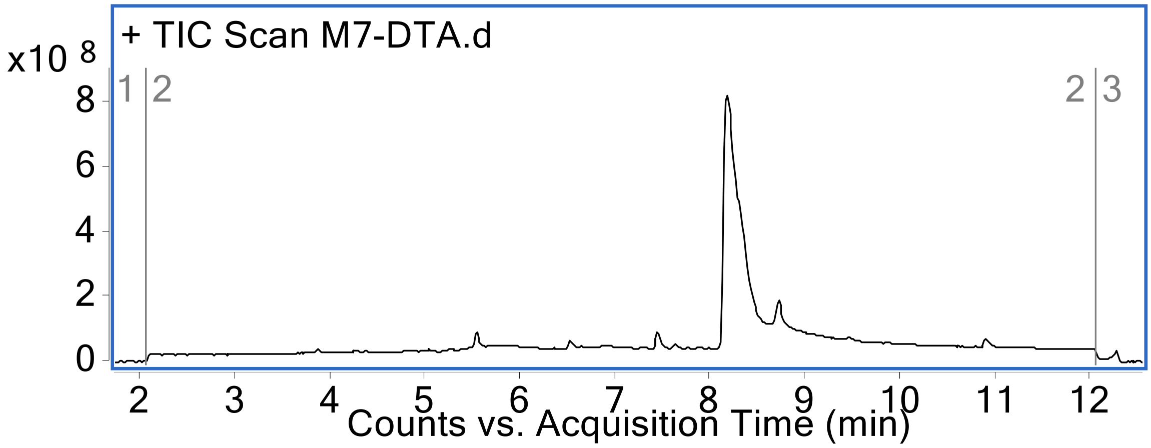


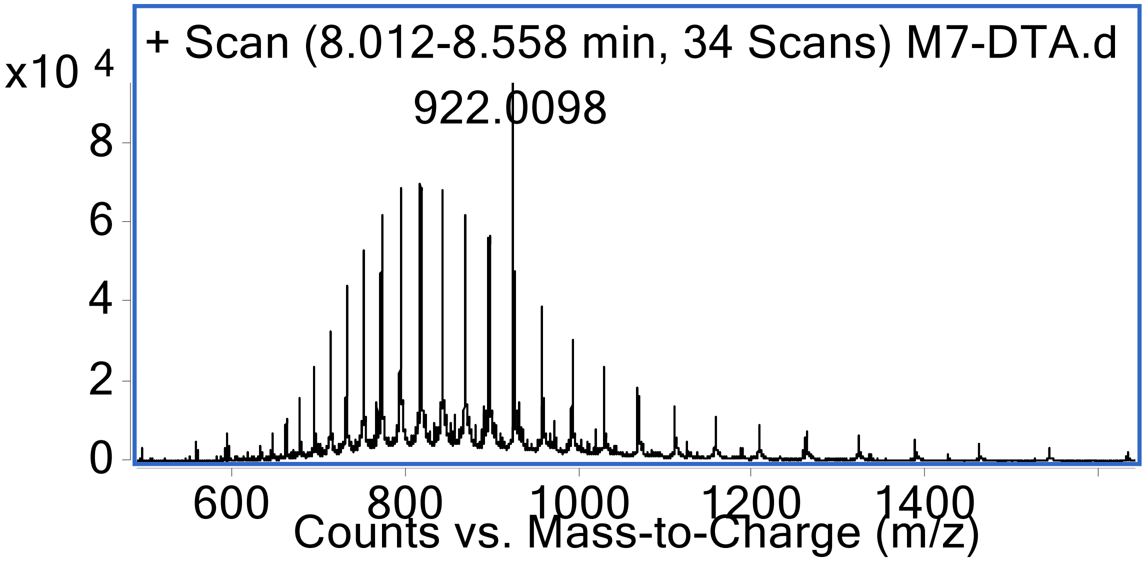
Mach7-DTA(C186S) (Method A)
Mass Expected: 27750.5 Da
Mass Observed: 27755.1 Da



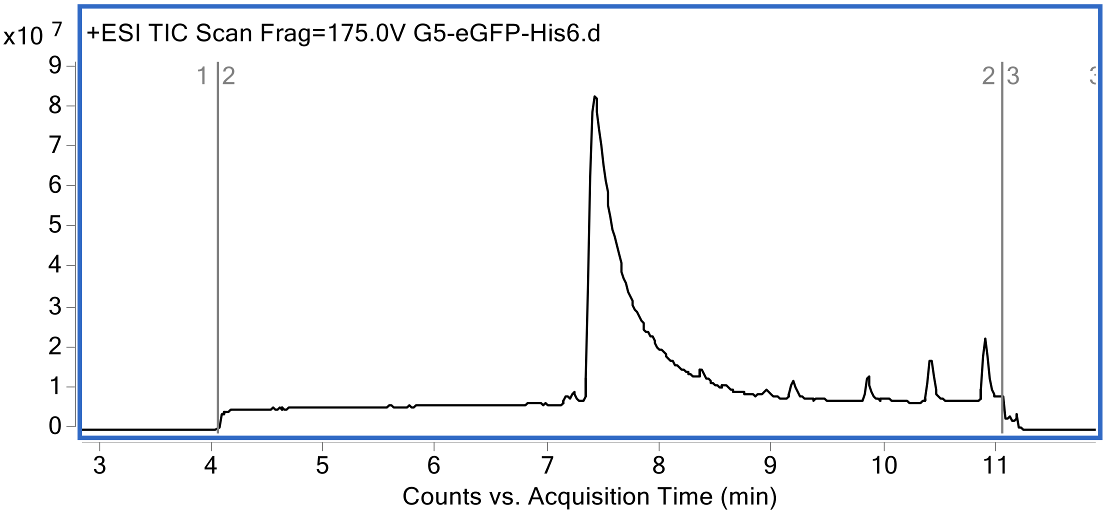


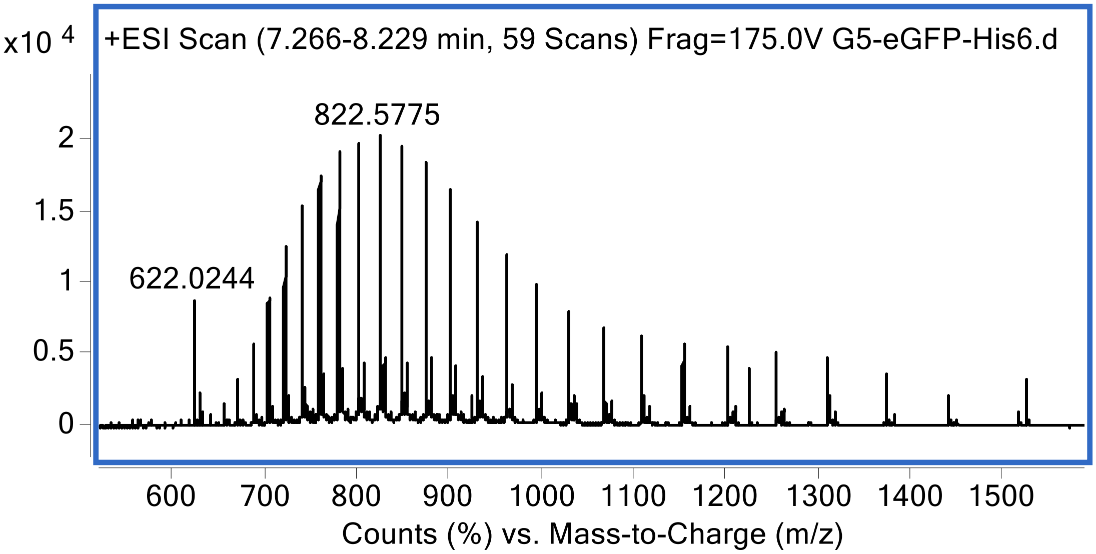
Mach7-DTA(C186S, E148S) (Method B)
Mass Expected: 27708.5 Da
Mass Observed: 27710.1 Da



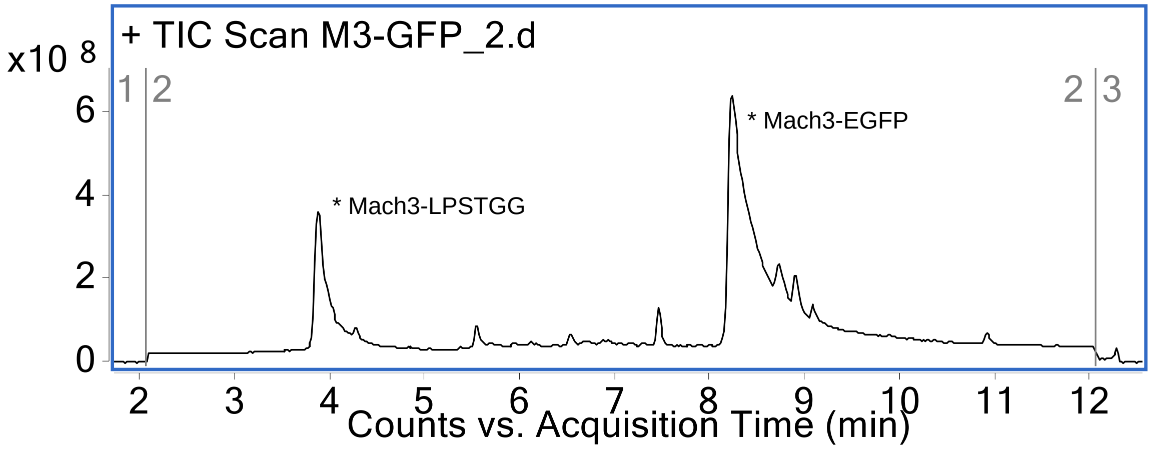


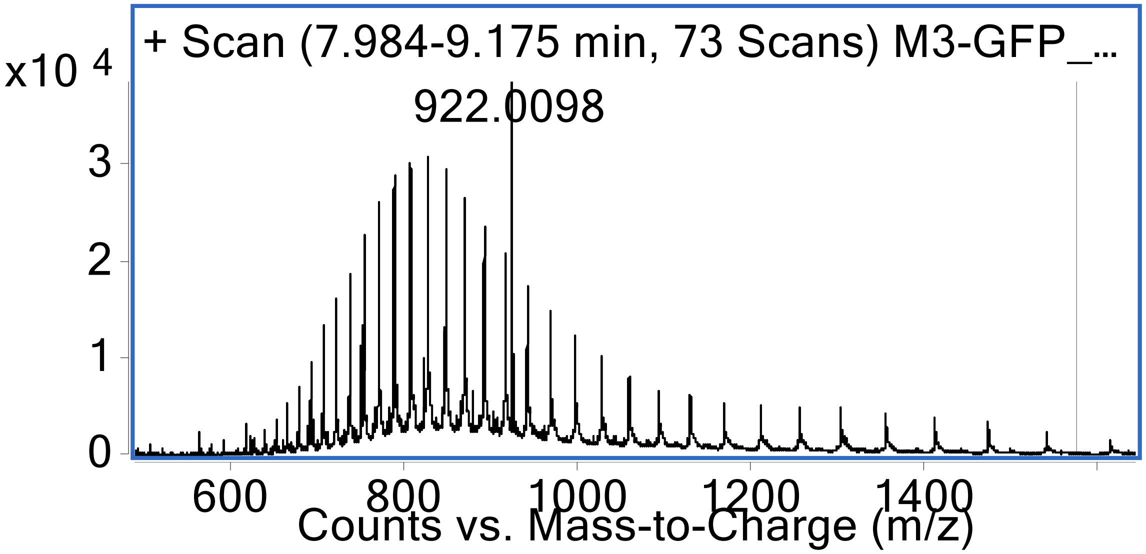
G5-EGFP (Method A)
Mass Expected: 28754.4 Da
Mass Observed: 28754.8 Da



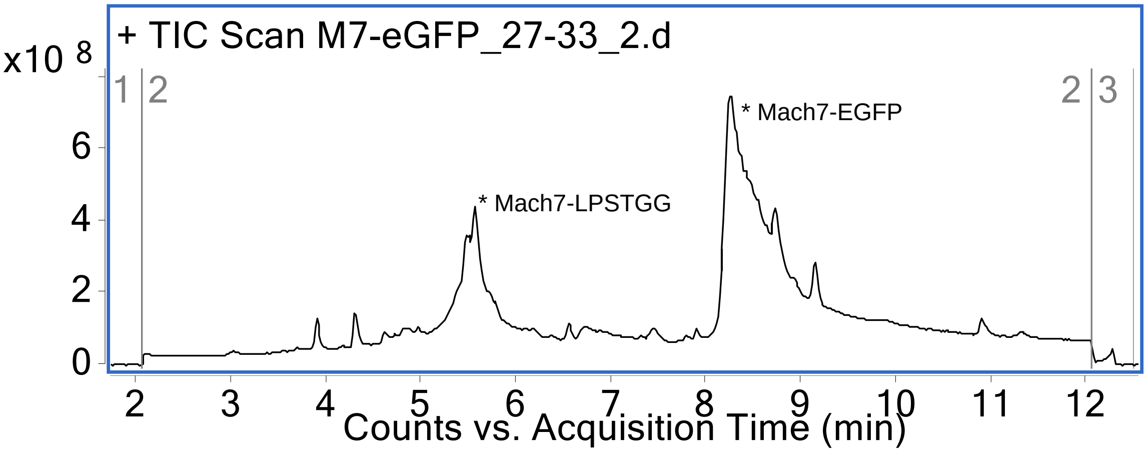


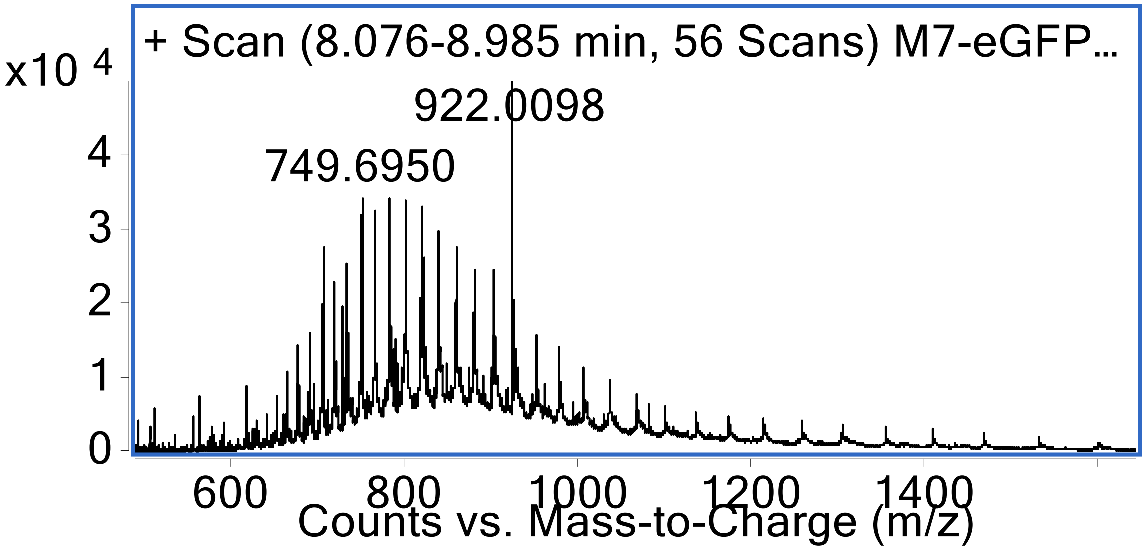
Mach3-EGFP (Method B)
Mass Expected: 33806.5 Da
Mass Observed: 33807.3 Da





Mach7-EGFP (Method B)
Mass Expected: 35128.3 Da
Mass Observed: 35130.3 Da





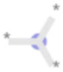
Appendix 2 Topological fingerprints

CB_Index = Condensed Bit-vector index, used in the figures in Manuscript and SI
TF_Index = Topological Fingerprint index, for the corresponding CB_index

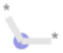
All ON bits out of the 2048-bits have been represented in the following set of figures. The radius of exploration goes from 0 (atom, itself) to 3 nearest neighbors. The coloring scheme denotes the node atom in blue, atoms which are a part of an aromatic ring in yellow, connected neighbors as a part of the topological exploration in black, and the unexplored neighboring atoms and nodes in gray.

CB_Index	TF_Index	CB_Index	TF_Index	CB_Index	TF_Index	CB_Index	TF_Index
1	1	**51**	585	**101**	1114	**151**	1693
2	11	**52**	623	**102**	1117	**152**	1719
3	22	**53**	625	**103**	1127	**153**	1731
4	27	**54**	650	**104**	1139	**154**	1736
5	32	**55**	667	**105**	1141	**155**	1737
6	67	**56**	671	**106**	1143	**156**	1750
7	70	**57**	680	**107**	1145	**157**	1751
8	74	**58**	708	**108**	1152	**158**	1752
9	79	**59**	713	**109**	1158	**159**	1754
10	80	**60**	724	**110**	1171	**160**	1758
11	119	**61**	727	**111**	1185	**161**	1773
12	132	**62**	739	**112**	1199	**162**	1778
13	140	**63**	742	**113**	1213	**163**	1783
14	150	**64**	745	**114**	1221	**164**	1785
15	173	**65**	759	**115**	1226	**165**	1791
16	197	**66**	776	**116**	1258	**166**	1794
17	204	**67**	784	**117**	1259	**167**	1805
18	220	**68**	785	**118**	1267	**168**	1840
19	222	**69**	786	**119**	1268	**169**	1844
20	227	**70**	806	**120**	1283	**170**	1847
21	229	**71**	807	**121**	1287	**171**	1849
22	231	**72**	831	**122**	1290	**172**	1873
23	272	**73**	857	**123**	1301	**173**	1876
24	280	**74**	878	**124**	1307	**174**	1879
25	283	**75**	889	**125**	1313	**175**	1882
26	289	**76**	894	**126**	1325	**176**	1898
27	293	**77**	900	**127**	1349	**177**	1910
28	294	**78**	926	**128**	1357	**178**	1911
29	295	**79**	931	**129**	1380	**179**	1912
30	305	**80**	955	**130**	1388	**180**	1917
31	310	**81**	966	**131**	1427	**181**	1926
32	321	**82**	971	**132**	1431	**182**	1928
33	328	**83**	981	**133**	1451	**183**	1937
34	329	**84**	983	**134**	1452	**184**	1946
35	362	**85**	989	**135**	1459	**185**	1947
36	364	**86**	1014	**136**	1462	**186**	1969
37	368	**87**	1017	**137**	1507	**187**	1970
38	376	**88**	1019	**138**	1517	**188**	2006
39	378	**89**	1022	**139**	1544	**189**	2013
40	389	**90**	1027	**140**	1547	**190**	2022
41	394	**91**	1028	**141**	1558	**191**	2042
42	412	**92**	1031	**142**	1564		
43	420	**93**	1034	**143**	1573		
44	425	**94**	1057	**144**	1601		
45	473	**95**	1066	**145**	1602		
46	482	**96**	1072	**146**	1607		
47	545	**97**	1082	**147**	1633		
48	553	**98**	1088	**148**	1656		
49	561	**99**	1104	**149**	1661		
50	575	**100**	1110	**150**	1685		

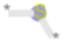
Linker 2



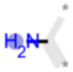
1



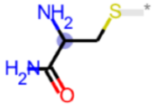
80



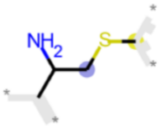
116



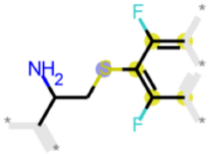
140



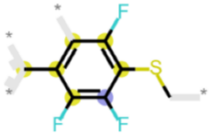
210



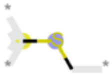
222



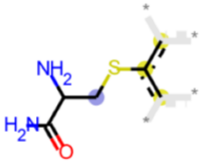
302



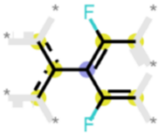
319



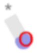
341



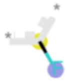
480



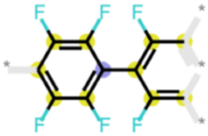
551



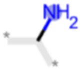
650



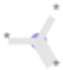
699



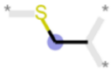
773



786

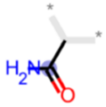


807

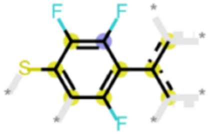


1093

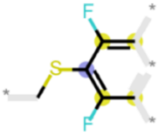
1171



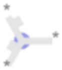
1212



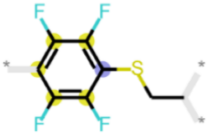
1234



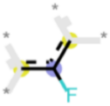
1305



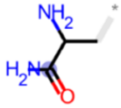
1380



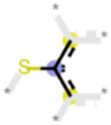
1436



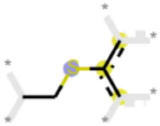
1445



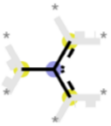
1654



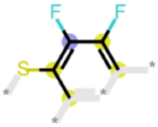
1716



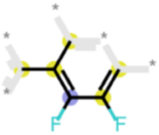
1733



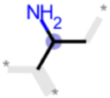
1747



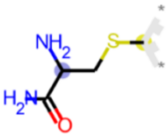
1821



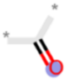
1848



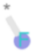
1849



1855

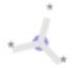


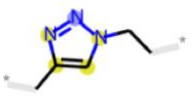
1917

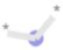


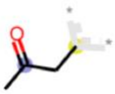
1928

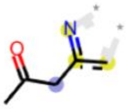
Linker 3

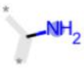

1

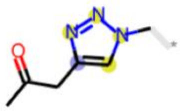

72

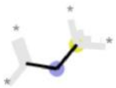

80

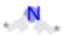

109

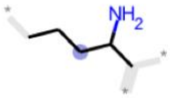

124

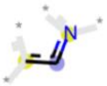

140

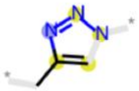

186

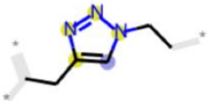

325


378

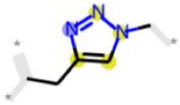

412

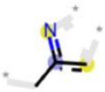

464

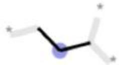

488

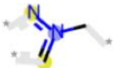

567

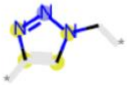
650

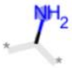

669

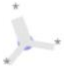

724

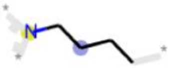

739

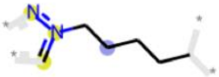

747

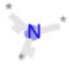

781

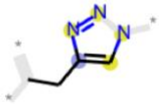

786

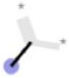

807


893

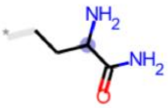

913

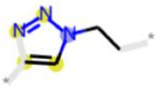

935

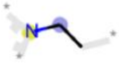

980

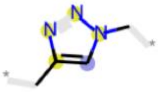

1017

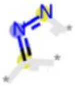
1057


1127

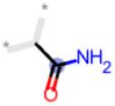

1141

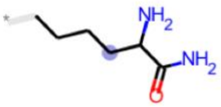

1145

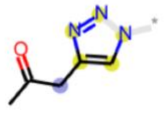

1155


1164

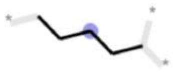
1171


1212

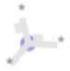

1214



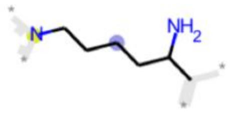
1259



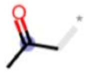
1283



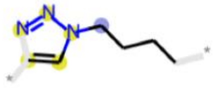
1380



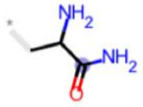
1536



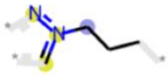
1550



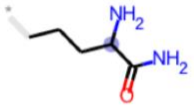
1582



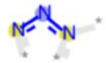
1654



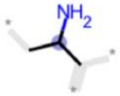
1689



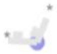
1728



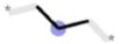
1783



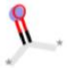
1849



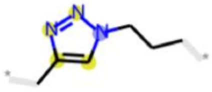
1873



1911

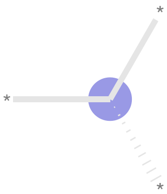


1917

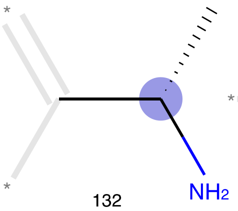


1971

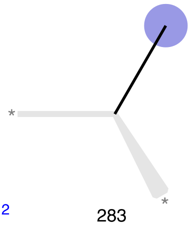
Alanine



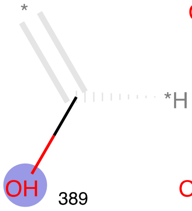
1



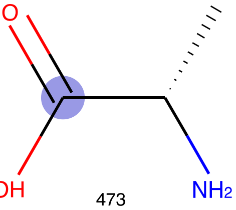
132



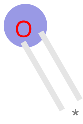
283



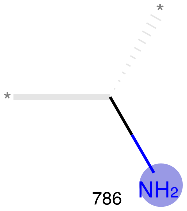
389



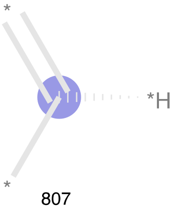
473



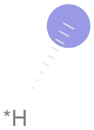
650



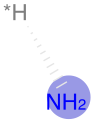
786



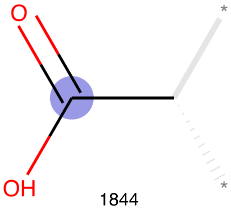
807



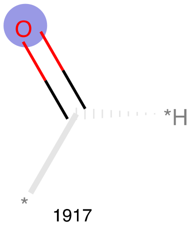
1057



1171

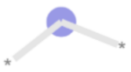


1844

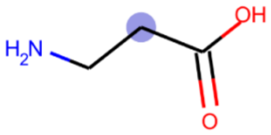


1917

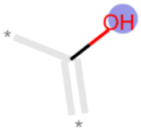
Beta-Alanine



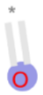
80



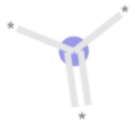
173



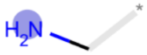
389



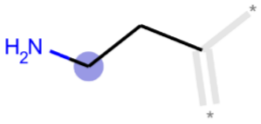
650



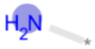
807



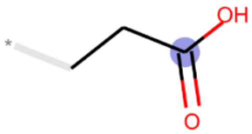
981



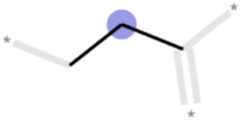
1088



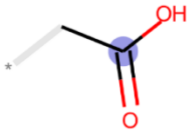
1171



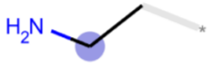
1287



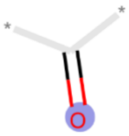
1564



1737

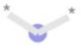


1840

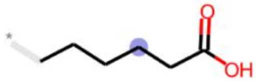


1917

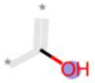
Aminohexanoic acid



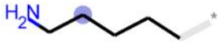
80



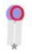
295



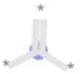
389



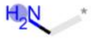
561



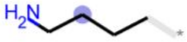
650



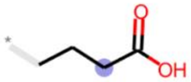
807



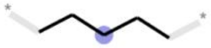
981



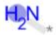
1082



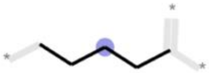
1110



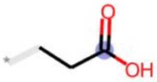
1143



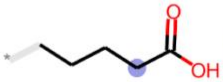
1171



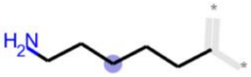
1267



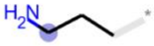
1287



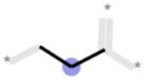
1301



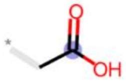
1462



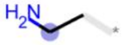
1517



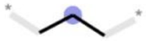
1564



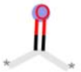
1737



1840

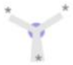


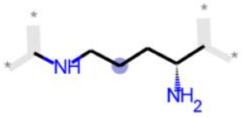
1911

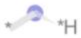


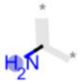
1917

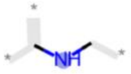
Arginine

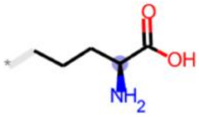

1

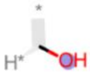

67

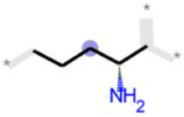

80


140


197

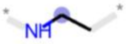

289

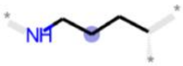

389

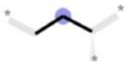

412

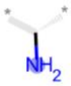
623

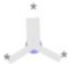
650

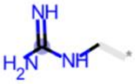

667

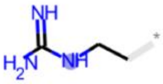

708

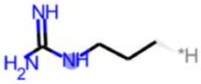

739

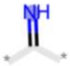

786

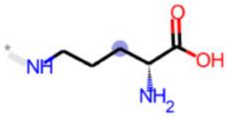

807

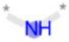

894

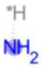

983

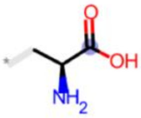

1027

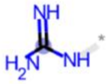

1104

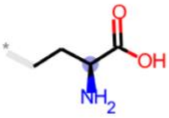

1127

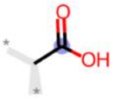

1152

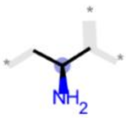

1171

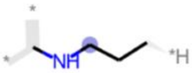

1427

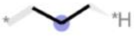

1451

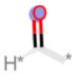

1791

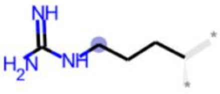

1844


1849

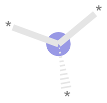

1876


1911

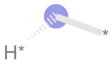

1917


2042

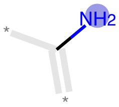
Asparagine



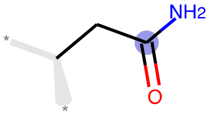
1



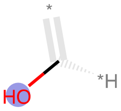
80



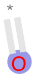
140



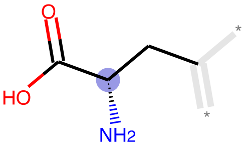
376



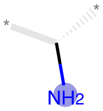
389



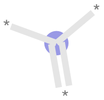
650



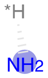
776



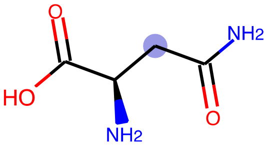
786



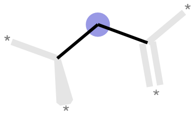
807



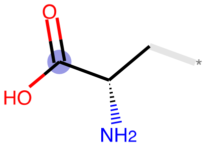
1171



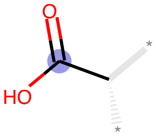
1213



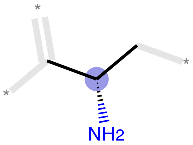
1290



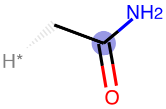
1427



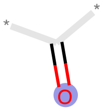
1844



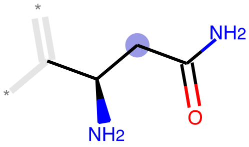
1849



1898

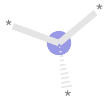


1917

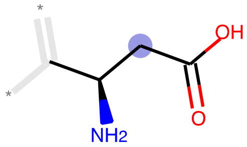


1946

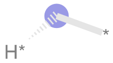
Aspartic acid



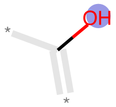
1



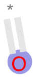
70



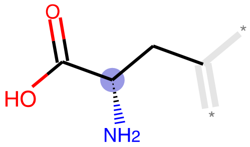
80



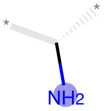
389



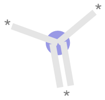
650



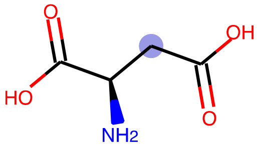
776



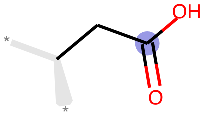
786



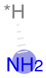
807



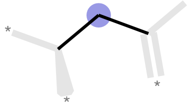
989



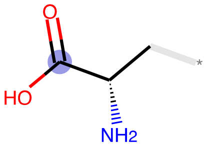
1141



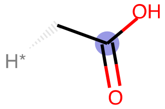
1171



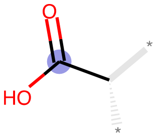
1290



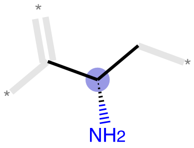
1427



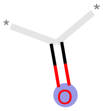
1737



1844

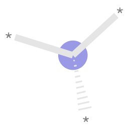


1849

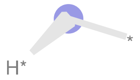


1917

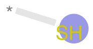
Cysteine



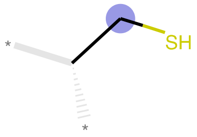
1



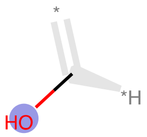
80



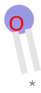
229



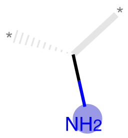
321



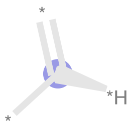
389



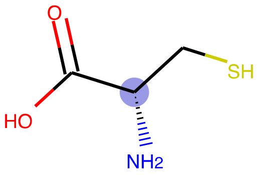
650



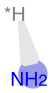
786



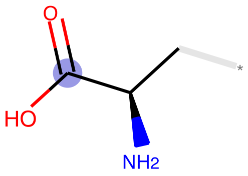
807



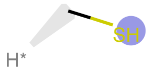
1072



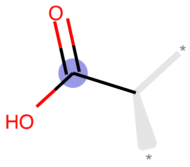
1171



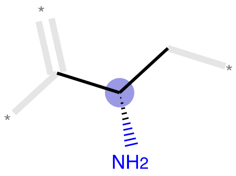
1427



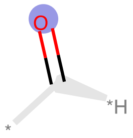
1459



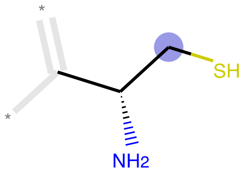
1844



1849

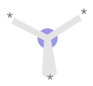


1917

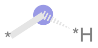


1926

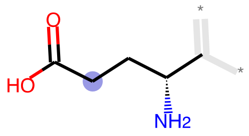
Glutamic acid



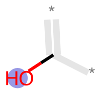
1



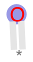
80



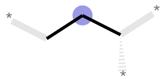
293



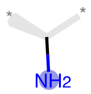
389



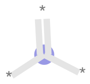
650



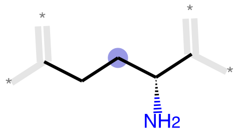
739



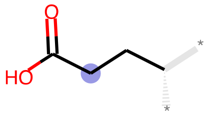
786



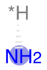
807



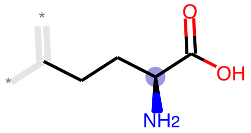
900



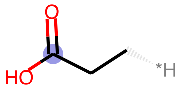
955



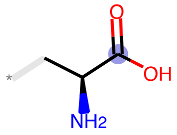
1171



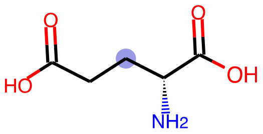
1258



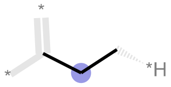
1287



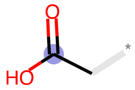
1427



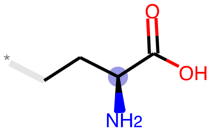
1547



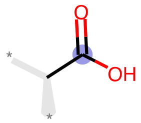
1564



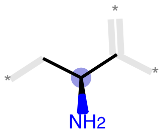
1737



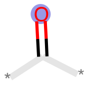
1791



1844

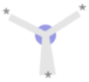


1849

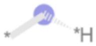


1917

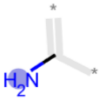
Glutamine



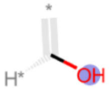
1



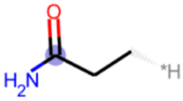
80



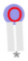
140



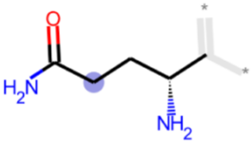
389



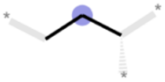
545



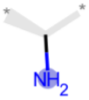
650



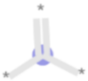
671



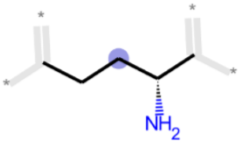
739



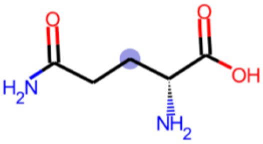
786



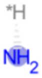
807



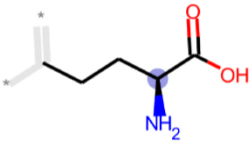
900



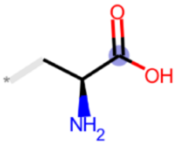
971



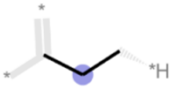
1171



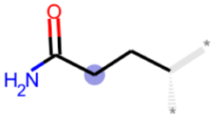
1258



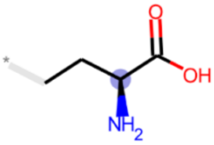
1427



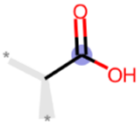
1564



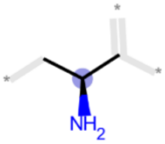
1752



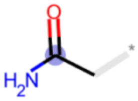
1791



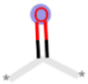
1844



1849

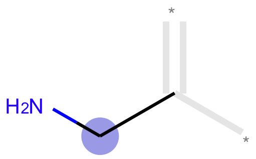


1898

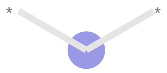


1917

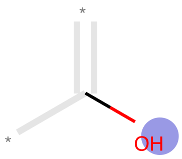
Glycine



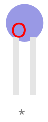
27



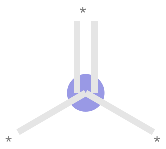
80



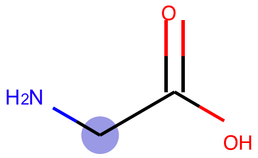
389



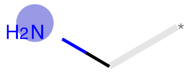
650



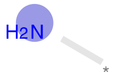
807



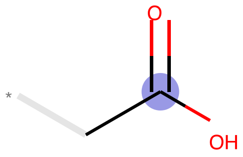
966



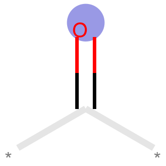
981



1171

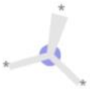


1737

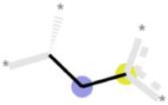


1917

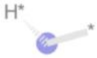
Histidine



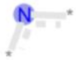
1



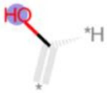
79



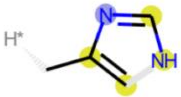
80



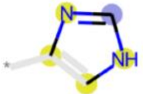
378



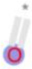
389



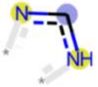
575



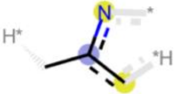
625



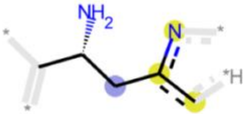
650



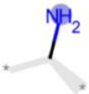
713



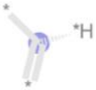
724



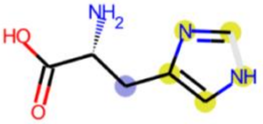
785



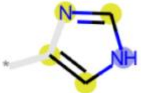
786



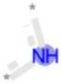
807



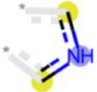
889



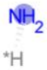
931



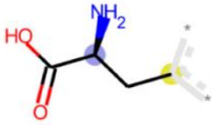
1114



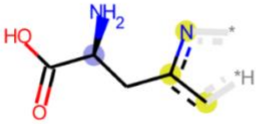
1145



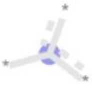
1171



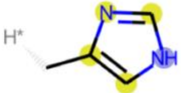
1221



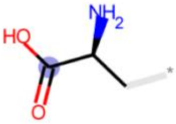
1259



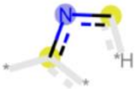
1380



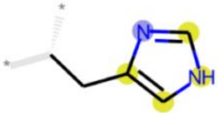
1388



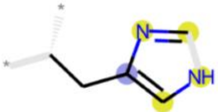
1427



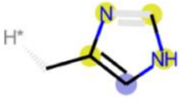
1452



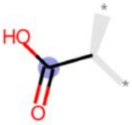
1601



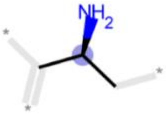
1633



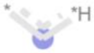
1758



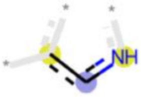
1844



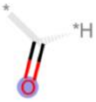
1849



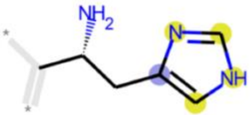
1873



1879

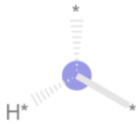


1917

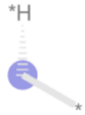


1969

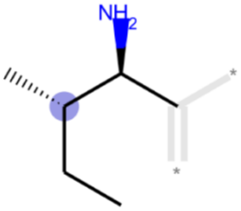
Isoleucine



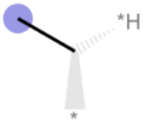
1



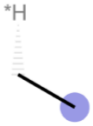
80



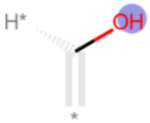
280



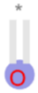
283



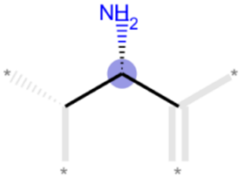
294



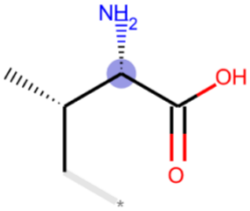
389



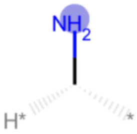
650



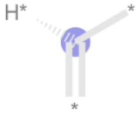
727



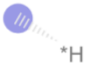
784



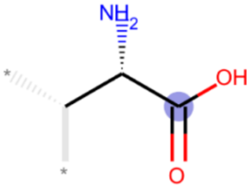
786



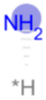
807



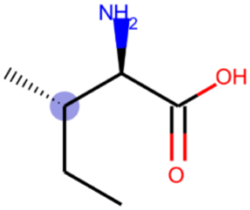
1057



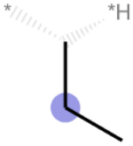
1139



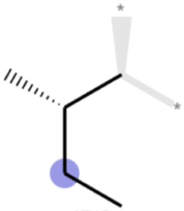
1171



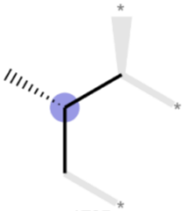
1185



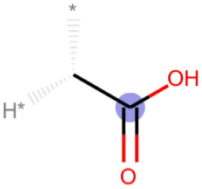
1544



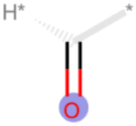
1719



1785

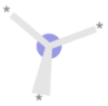


1844

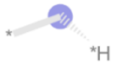


1917

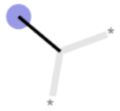
Leucine



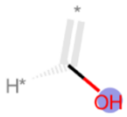
1



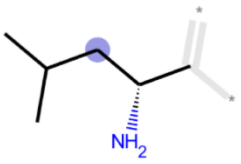
80



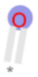
283



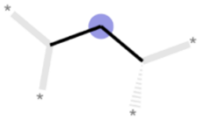
389



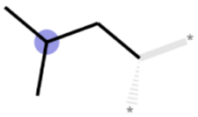
425



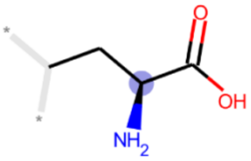
650



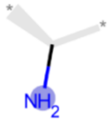
680



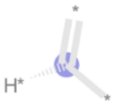
759



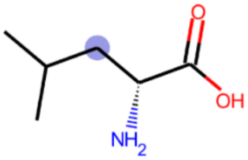
776



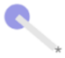
786



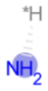
807



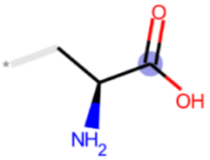
878



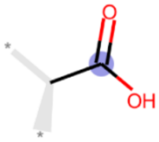
1057



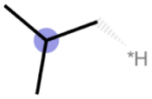
1171



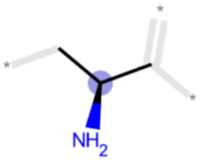
1427



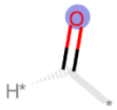
1844



1847

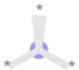


1849

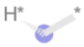


1917

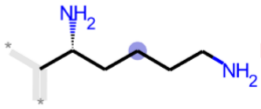
Lysine



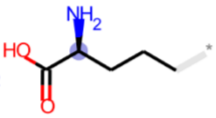
1



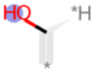
80



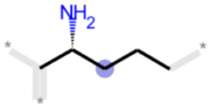
220



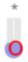
289



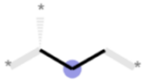
389



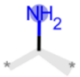
412



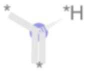
650



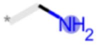
739



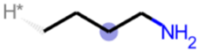
786



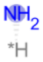
807



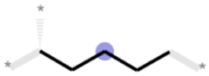
981



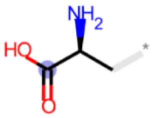
1082



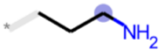
1171



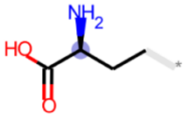
1283



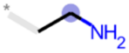
1427



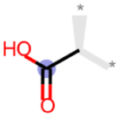
1517



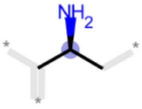
1791



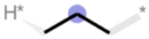
1840



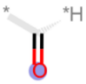
1844



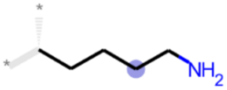
1849



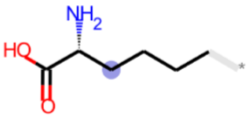
1911



1917

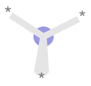


1928

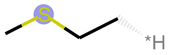


2006

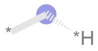
Methionine



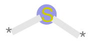
1



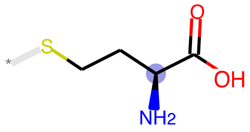
11



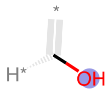
80



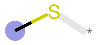
116



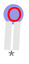
368



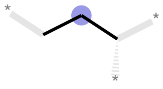
389



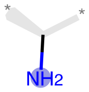
394



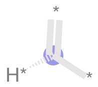
650



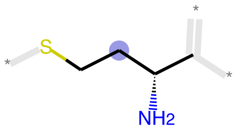
739



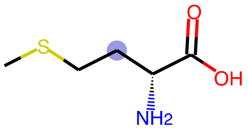
786



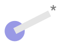
807



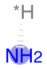
1021



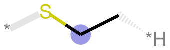
1034



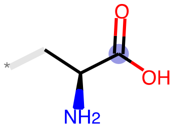
1057



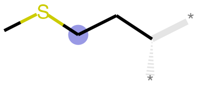
1171



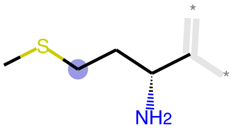
1182



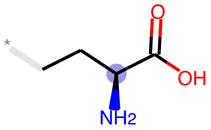
1427



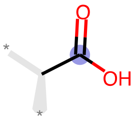
1558



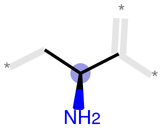
1661



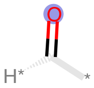
1791



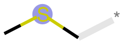
1844



1849

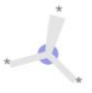


1917

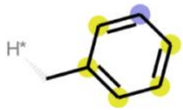


1942

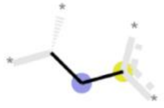
Phenylalanine



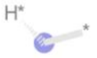
1



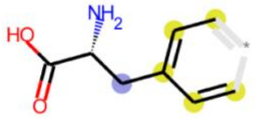
32



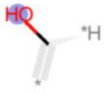
79



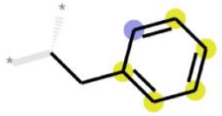
80



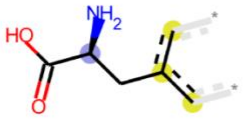
150



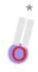
389



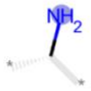
420



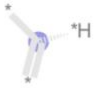
585



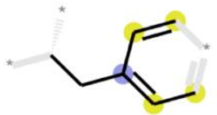
650



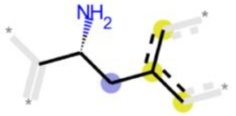
786



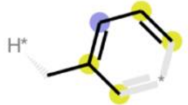
807



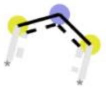
857



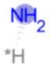
1017



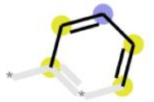
1066



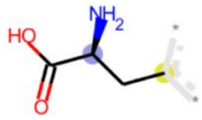
1088



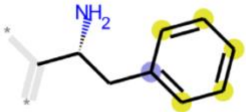
1171



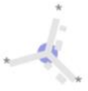
1199



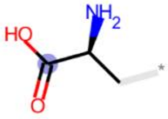
1221



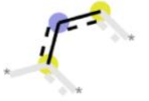
1307



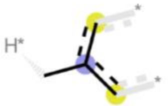
1380



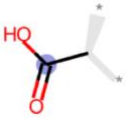
1427



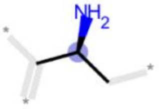
1750



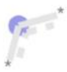
1754



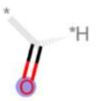
1844



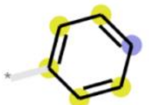
1849



1873

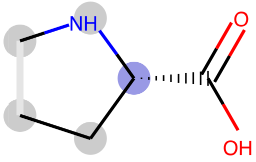


1917

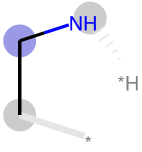


1947

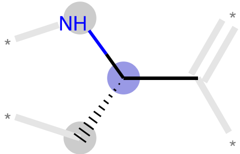
Proline



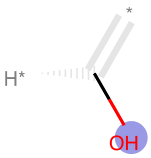
74



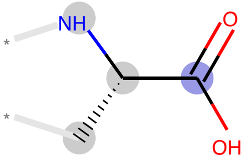
305



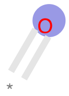
362



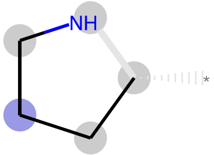
389



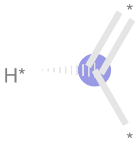
553



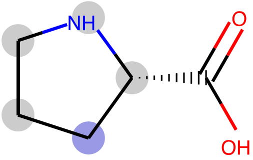
650



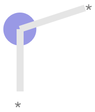
742



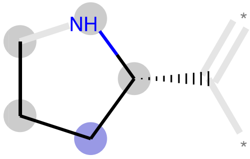
807



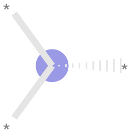
831



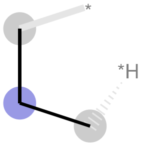
926



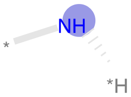
1014



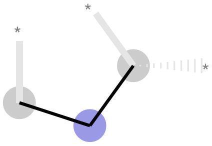
1019



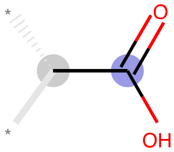
1028



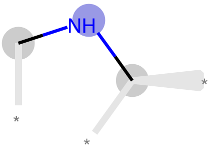
1114



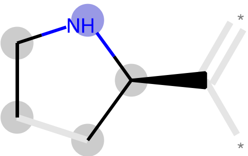
1325



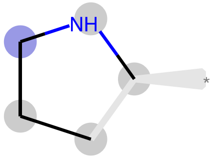
1431



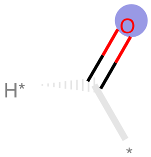
1507



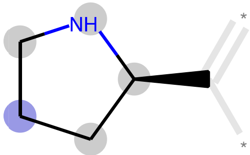
1736



1912

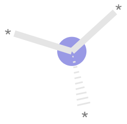


1917

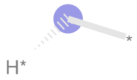


2022

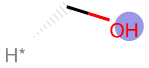
Serine



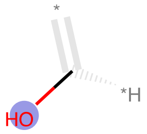
1



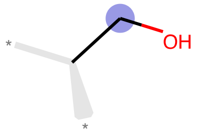
80



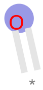
222



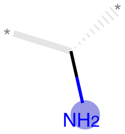
389



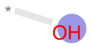
482



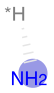
650



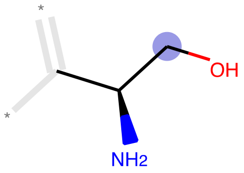
786



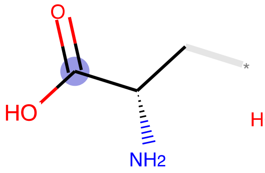
807



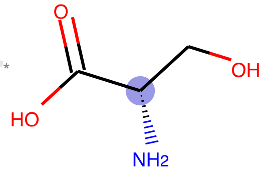
1171



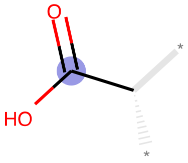
1268



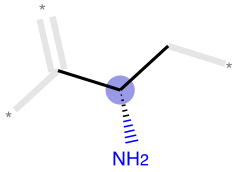
1427



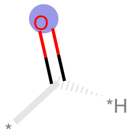
1751



1844

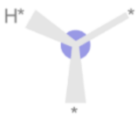


1849

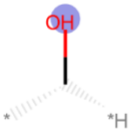


1917

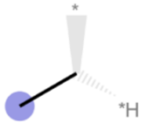
Threonine



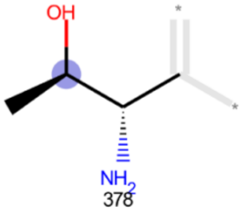
1



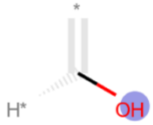
227



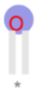
283



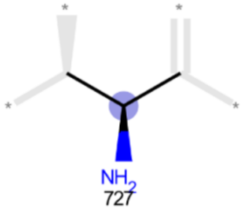
378



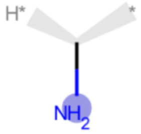
389



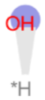
650



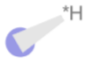
727



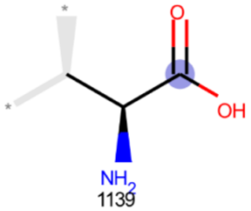
786



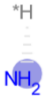
807



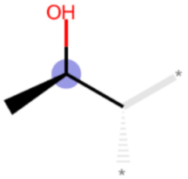
1057



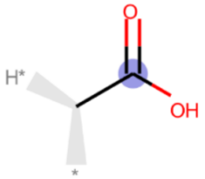
1139



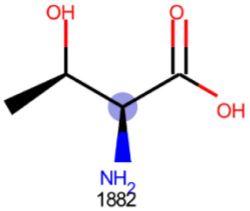
1171



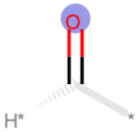
1693



1844

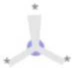


1882

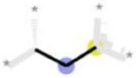


1917

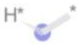
Tryptophan



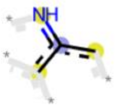
1



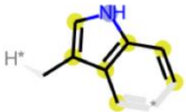
79



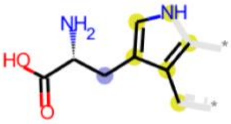
80



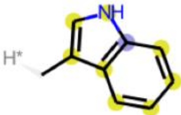
119



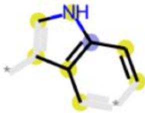
140



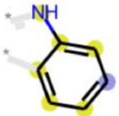
204



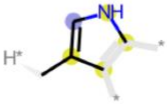
272



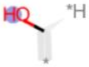
328



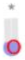
329



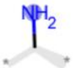
364



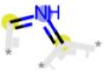
389



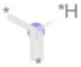
650



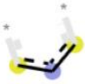
786



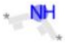
806



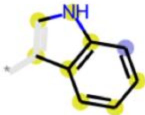
807



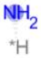
1088



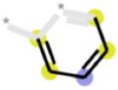
1114



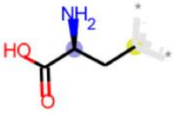
1158



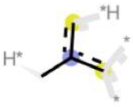
1171



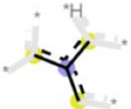
1199



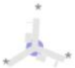
1221



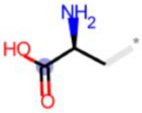
1349



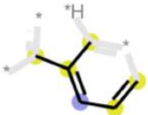
1357



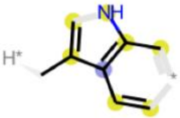
1380



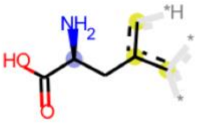
1427



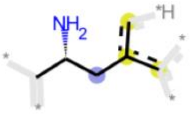
1573



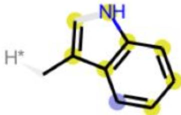
1607



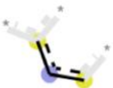
1656



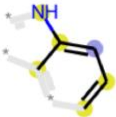
1685



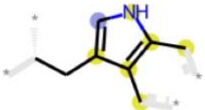
1731



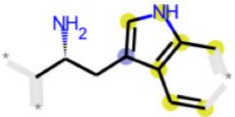
1750



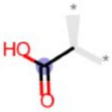
1783



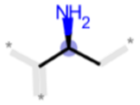
1794



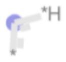
1805



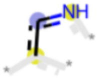
1844



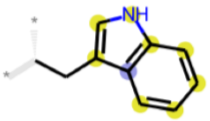
1849



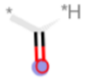
1873



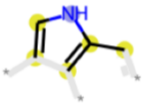
1879



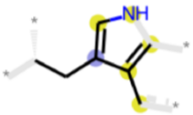
1910



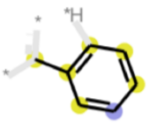
1917



1937

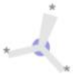


1970

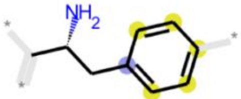


2013

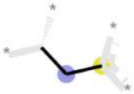
Tyrosine



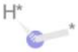
1



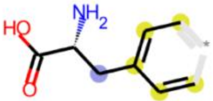
22



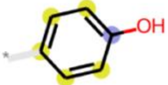
79



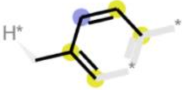
80



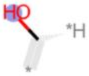
150



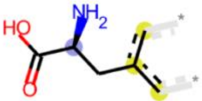
231



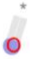
310



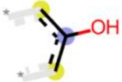
389



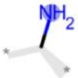
585



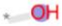
650



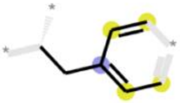
745



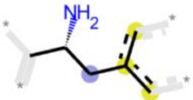
786



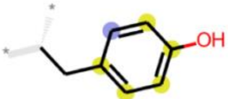
807



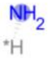
857



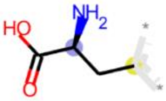
1017



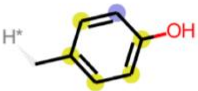
1117



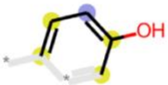
1171



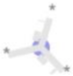
1221



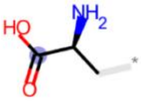
1226



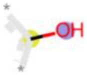
1313



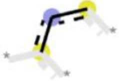
1380



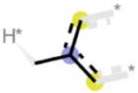
1427



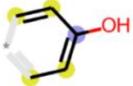
1602



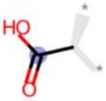
1750



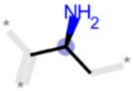
1754



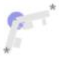
1778



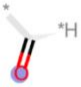
1844



1849

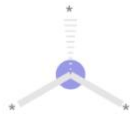


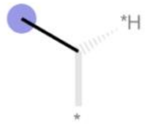
1873



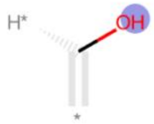
1917

Valine

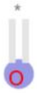

1



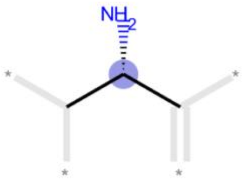
283



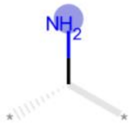
389



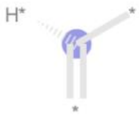
650



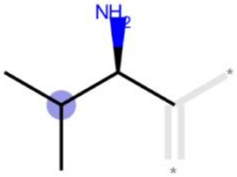
727



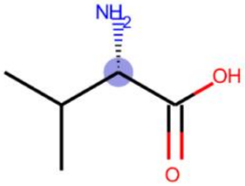
786



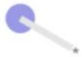
807

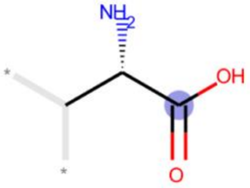


1022

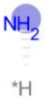


1031

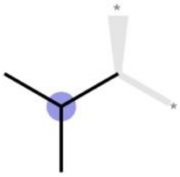

1057

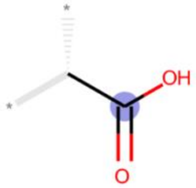


1139

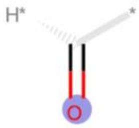


1171


1773



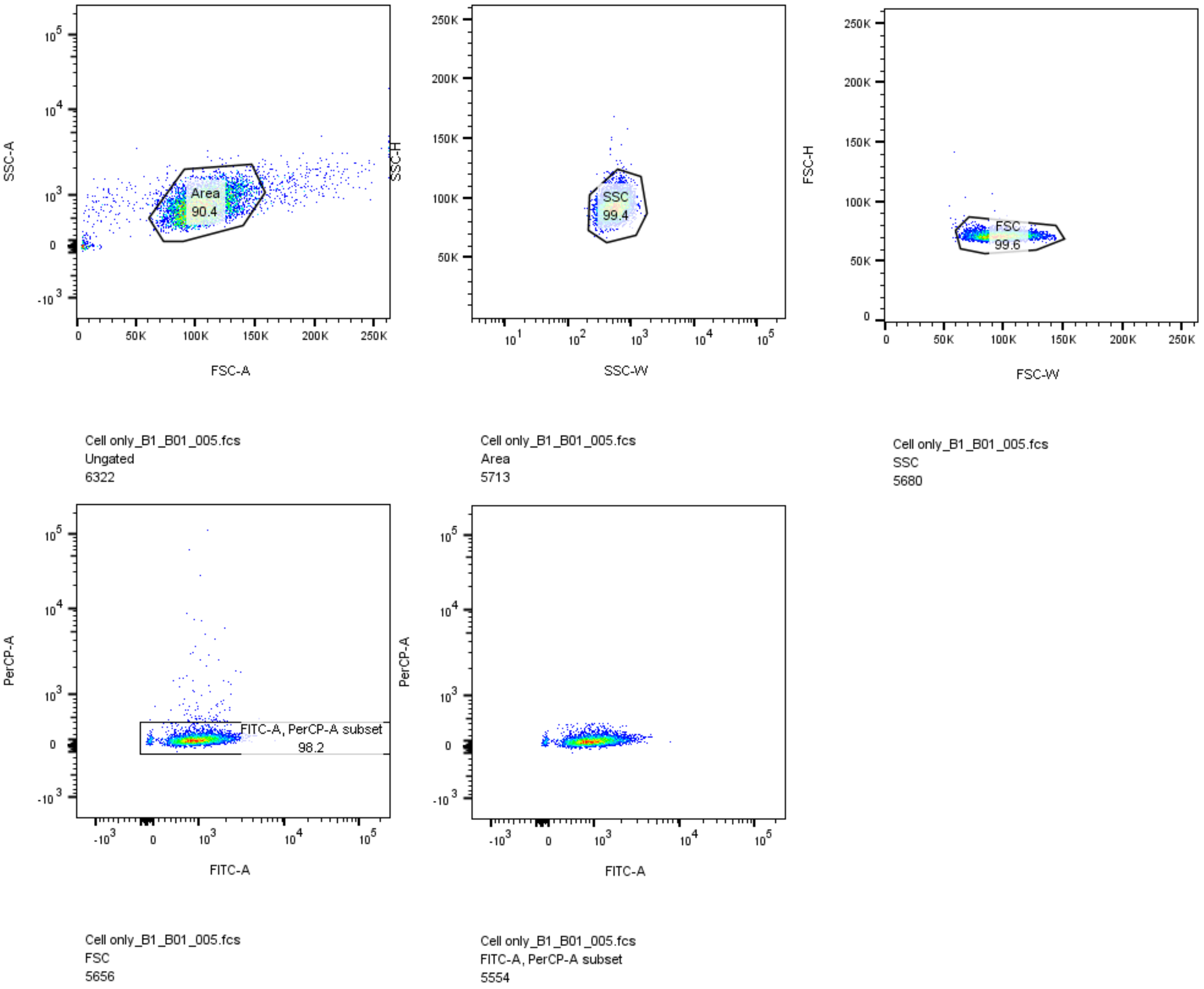
1844



1917

Appendix 3: Example of flow cytometry gating strategy

Gates were created using the cell-only control, and applied to experimental samples. Gates were applied to the main cell population on the SSC vs FSC density plot, and then to the SSC and FSC density plots, respectively, in order to exclude outlying cells. Finally, outlying cells labeled by PerCP were excluded, and the mean fluorescence intensity of FITC was obtained.


SSC-A
FSC-A
Area
Cell only_B1_B01_005.fcs
Ungated
SSC-H
SSC-W
SSC
Cell only_B1_B01_005.fcs
Area
FSC-H
FSC-W
FSC
Cell only_B1_B01_005.fcs
SSC
PerCP-A
FITC-A
FITC-A, PerCP-A subset
Cell only_B1_B01_005.fcs
FSC
PerCP-A
FITC-A
Cell only_B1_B01_005.fcs
FITC-A, PerCP-A subset