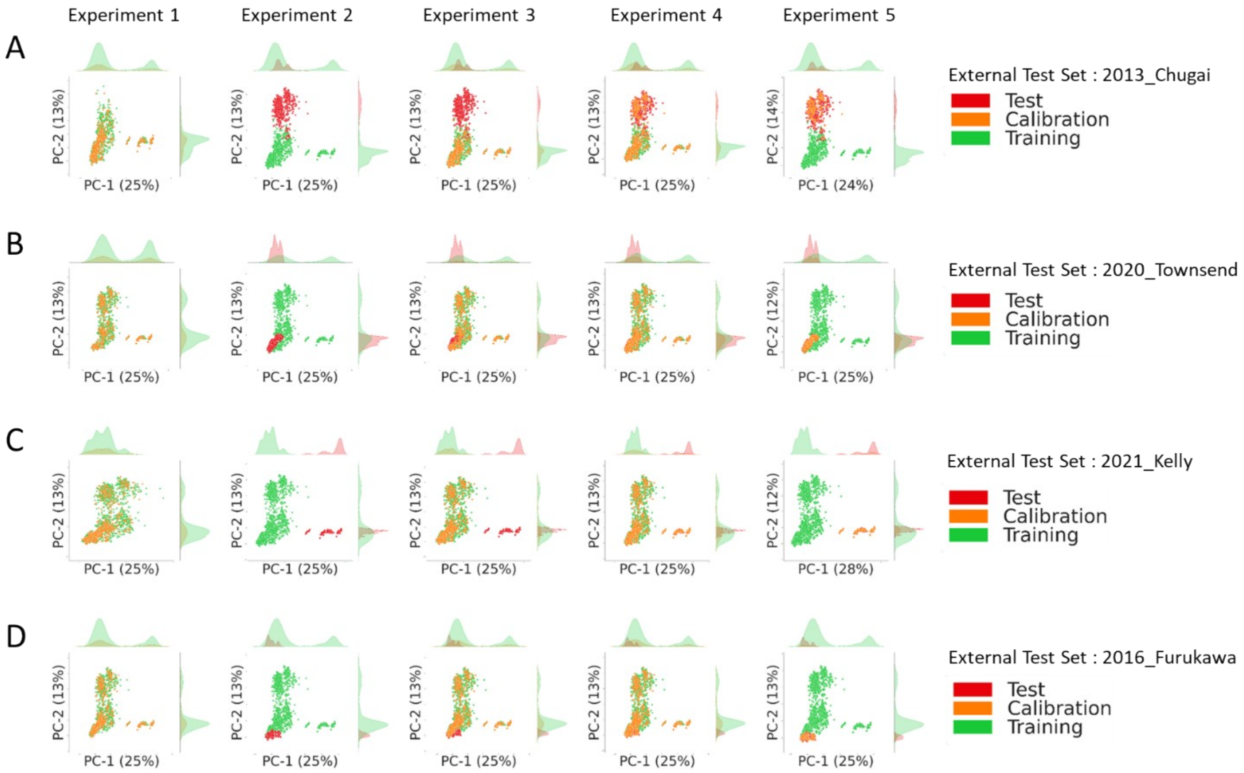
Electronic Supplementary Information (ESI)

A methodology to correctly assess the applicability domain of cell membrane permeability predictors for cyclic peptides

Gökçe Geylan,[*ab] Leonardo De Maria,[c] Ola Engkvist,[ad] Florian David [b] and Ulf Norinder [efg]



Supplementary Figure 1. Chemical space visualizations for experiments 1-5 with external test sets (A) 2013_Chugai, (B) 2020_Townsend, (C) 2021_Kelly, and (D) 2016_Furukawa. Column labelled Experiment 1 shows the chemical space of the training data and the calibration data. The calibration data was used as the internal validation set in this experiment. Column "Experiment 2" shows the chemical space spanned by the training data and the external test set to showcase the difference between the datasets. Column "Experiment 3" shows the training, calibration, and test sets for demonstrating the peptide sets in traditional conformal prediction application, or the baseline experiment. Column "Experiment 4" shows the training, calibration, and test sets for concatenating the 20% of the external set to the baseline calibration set. At last, column "Experiment 5" demonstrating the datasets when the calibration set is created by taking 20% of the external test set.

Supplementary Table 1. Description of Experiments 1-7

Experiment	Description
1	Model building with hyperparameter optimization to evaluate the internal validation set performance
2	Final optimized models from Experiment 1 were used to evaluate the performance of the corresponding external test set
3	The conformal prediction framework is applied to the models from Experiment 2 to evaluate the performance of the corresponding external test set
4	The conformal prediction framework is applied to the models from Experiment 2 to recalibrate the external test set predictions, in a 5-fold cross-validation manner, by augmenting the calibration set with 20% of the holdout external test set fold to evaluate the performance of the corresponding external test set
5	The conformal prediction framework is applied to the models from Experiment 2 by using a subset, 20%, of the external test set to evaluate the performance of the corresponding external test set
6	Comparison of the improvement of model's reliability between Experiment 4 and 5
7	Model building with hyperparameter optimization on the entire data with various cross-validation split strategies

Supplementary Table 2. The model performance metrics on the internal validation data. The models and the performance metrics are labelled with the external data set they will be evaluated on in the next experiment, Experiment 2.

Case		**Performance Metrics on Validation Set**					
Holdout Case	Model Name	Balanced Accuracy	ROC-AUC	MCC	Sensitivity	Specificity	Precision
2016 Furukawa	RF	0.78	0.78	0.59	0.90	0.66	0.85
	XGBoost	0.78	0.78	0.59	0.90	0.67	0.86
	LightGBM	0.69	0.69	0.5	0.97	0.41	0.78
	SVM	0.78	0.78	0.58	0.89	0.67	0.86
2013 Chugai	RF	0.77	0.77	0.57	0.89	0.66	0.82
	XGBoost	0.79	0.79	0.59	0.88	0.70	0.84
	LightGBM	0.72	0.72	0.51	0.95	0.48	0.77
	SVM	0.78	0.78	0.57	0.85	0.72	0.85
2021 Kelly	RF	0.77	0.77	0.56	0.90	0.63	0.85
	XGBoost	0.50	0.50	0.00	1.00	0.00	0.69
	LightGBM	0.61	0.61	0.37	0.98	0.24	0.74
	SVM	0.50	0.50	0.00	1.00	0.00	0.69
2020 Townsend	RF	0.82	0.82	0.64	0.89	0.74	0.87
	XGBoost	0.86	0.86	0.73	0.92	0.80	0.90
	LightGBM	0.63	0.63	0.42	0.99	0.27	0.72
	SVM	0.76	0.76	0.57	0.94	0.58	0.81

Supplementary Table 3. The model performance metrics on the external test sets.

Case		Performance Metrics on External Test Set					
Holdout Case	Model Name	Balanced Accuracy	ROC-AUC	MCC	Sensitivity	Specificity	Precision
2016 Furukawa	RF	0.50	0.50	-0.02	0.99	0.01	0.56
	XGBoost	0.50	0.50	0.00	0.94	0.06	0.56
	LightGBM	0.50	0.50	0.00	1.00	0.00	0.56
	SVM	0.45	0.45	-0.11	0.60	0.29	0.53
2013 Chugai	RF	0.51	0.51	0.02	0.95	0.07	0.90
	XGBoost	0.57	0.57	0.09	0.69	0.46	0.92
	LightGBM	0.50	0.50	-0.02	1.00	0.00	0.90
	SVM	0.45	0.45	-0.06	0.54	0.37	0.88
2021 Kelly	RF	0.56	0.56	0.27	1.00	0.12	0.64
	XGBoost	0.50	0.50	0.00	1.00	0.00	0.61
	LightGBM	0.58	0.58	0.32	1.00	0.17	0.65
	SVM	0.50	0.50	0.00	1.00	0.00	0.61
2020 Townsend	RF	0.6	0.60	0.20	0.76	0.44	0.76
	XGBoost	0.55	0.55	0.13	0.84	0.27	0.73
	LightGBM	0.50	0.50	0.00	1.00	0.00	0.70
	SVM	0.50	0.50	0.04	1.00	0.00	0.70

Supplementary Table 4. The conformal prediction metrics of efficiency and validity were calculated on the external test sets predicted by the calibrated models. The metrics were computed at significance level=0.2, mandating the model to produce predictions with 80% confidence. The efficiency and validity metrics were calculated for both the "Permeable" and "Non-permeable" classes separately and these classes were labelled as "1" and "0" respectively for these metrics.

Case		Significance Level = 0.2			
Holdout Case	Model Name	Efficiency 0	Efficiency 1	Validity 0	Validity 1
2016 Furukawa	RF	0.42	0.42	0.79	0.94
	XGBoost	0.56	0.58	0.62	0.92
	LightGBM	0.5	0.49	0.65	0.96
	SVM	0.48	0.43	0.96	0.67
2013 Chugai	RF	0.21	0.24	1.0	0.76
	XGBoost	0.38	0.38	1.0	0.68
	LightGBM	0.4	0.39	1.0	0.63
	SVM	0.47	0.48	0.8	0.69
2021 Kelly	RF	0.68	0.3	1.0	0.7
	XGBoost	0.63	0.36	1.0	0.67
	LightGBM	0.64	0.31	0.98	0.77
	SVM	0.0	0.0	1.0	1.0
2020 Townsend	RF	0.68	0.56	0.93	0.74
	XGBoost	0.74	0.84	0.64	0.85
	LightGBM	0.52	0.61	0.82	0.91
	SVM	0.09	0.06	0.99	0.99

Supplementary Table 5. The conformal prediction metrics for the recalibration of the models by expanding the calibration set with a portion of the external test set (Experiment 4). The metrics were calculated at significance level=0.2. The 5-fold recalibration process was reported by aggregating as the mean (± standard deviation). The efficiency and validity metrics were calculated for both the "Permeable" and "Non-permeable" classes separately and these classes were labelled as "1" and "0" respectively for these metrics.

Case		**Significance Level = 0.2**			
Holdout Case	Model Name	Efficiency 0	Efficiency 1	Validity 0	Validity 1
	RF	0.76 (± 0.01)	0.65 (± 0.06)	0.92 (± 0.03)	0.7 (± 0.04)
2016	XGBoost	0.8 (± 0.05)	0.69 (± 0.07)	0.9 (± 0.04)	0.68 (± 0.07)
Furukawa	LightGBM	0.8 (± 0.07)	0.63 (± 0.09)	0.92 (± 0.04)	0.65 (± 0.09)
	SVM	0.69 (± 0.05)	0.58 (± 0.05)	0.88 (± 0.04)	0.79 (± 0.01)
	RF	0.61 (± 0.12)	0.94 (± 0.01)	0.61 (± 0.06)	0.98 (± 0.01)
2013	XGBoost	0.67 (± 0.11)	0.94 (± 0.01)	0.59 (± 0.04)	0.98 (± 0.01)
Chugai	LightGBM	0.66 (± 0.18)	0.96 (± 0.01)	0.56 (± 0.09)	0.98 (± 0.01)
	SVM	0.58 (± 0.07)	0.75 (± 0.03)	0.71 (± 0.03)	0.96 (± 0.01)
	RF	0.85 (± 0.02)	0.78 (± 0.03)	0.89 (± 0.01)	0.82 (± 0.02)
2021	XGBoost	0.75 (± 0.06)	0.62 (± 0.12)	0.91 (± 0.05)	0.7 (± 0.05)
Kelly	LightGBM	0.79 (± 0.04)	0.73 (± 0.04)	0.9 (± 0.03)	0.79 (± 0.03)
	SVM	0.55 (± 0.03)	0.46 (± 0.03)	0.92 (± 0.01)	0.88 (± 0.01)
	RF	0.82 (± 0.01)	0.85 (± 0.02)	0.81 (± 0.01)	0.86 (± 0.02)
2020	XGBoost	0.81 (± 0.03)	0.84 (± 0.03)	0.82 (± 0.02)	0.85 (± 0.01)
Townsend	LightGBM	0.8 (± 0.03)	0.83 (± 0.02)	0.8 (± 0.02)	0.86 (± 0.02)
	SVM	0.74 (± 0.03)	0.81 (± 0.03)	0.81 (± 0.01)	0.87 (± 0.02)

Supplementary Table 6. The conformal prediction metrics for the recalibration of the models by employing on a fraction of the external test set (Experiment 5). The metrics were calculated at significance level=0.2. The 5-fold recalibration process was reported by aggregating as the mean (± standard deviation). The efficiency and validity metrics were calculated for both the “Permeable” and “Non-permeable” classes separately and these classes were labelled as “1” and “0” respectively for these metrics.

Case		**Significance Level = 0.2**			
Holdout Case	Model Name	Efficiency 0	Efficiency 1	Validity 0	Validity 1
2016 Furukawa	RF	0.59 (± 0.11)	0.52 (± 0.12)	0.86 (± 0.07)	0.91 (± 0.02)
	XGBoost	0.59 (± 0.12)	0.51 (± 0.16)	0.86 (± 0.07)	0.91 (± 0.03)
	LightGBM	0.51 (± 0.11)	0.44 (± 0.13)	0.9 (± 0.05)	0.9 (± 0.03)
	SVM	0.5 (± 0.12)	0.45 (± 0.13)	0.88 (± 0.07)	0.9 (± 0.03)
2013 Chugai	RF	0.86 (± 0.04)	0.81 (± 0.12)	0.89 (± 0.06)	0.85 (± 0.03)
	XGBoost	0.85 (± 0.06)	0.82 (± 0.06)	0.89 (± 0.1)	0.83 (± 0.03)
	LightGBM	0.92 (± 0.04)	0.91 (± 0.05)	0.82 (± 0.12)	0.85 (± 0.02)
	SVM	0.87 (± 0.1)	0.73 (± 0.26)	0.87 (± 0.09)	0.86 (± 0.01)
2021 Kelly	RF	0.88 (± 0.04)	0.85 (± 0.04)	0.85 (± 0.01)	0.85 (± 0.03)
	XGBoost	0.66 (± 0.1)	0.61 (± 0.12)	0.86 (± 0.05)	0.87 (± 0.04)
	LightGBM	0.8 (± 0.07)	0.79 (± 0.05)	0.85 (± 0.04)	0.82 (± 0.05)
	SVM	0.53 (± 0.08)	0.49 (± 0.08)	0.9 (± 0.03)	0.88 (± 0.03)
2020 Townsend	RF	0.83 (± 0.03)	0.84 (± 0.03)	0.83 (± 0.01)	0.84 (± 0.02)
	XGBoost	0.8 (± 0.04)	0.81 (± 0.05)	0.85 (± 0.02)	0.85 (± 0.01)
	LightGBM	0.85 (± 0.04)	0.85 (± 0.05)	0.82 (± 0.04)	0.82 (± 0.02)
	SVM	0.75 (± 0.05)	0.77 (± 0.05)	0.85 (± 0.03)	0.85 (± 0.02)