Supporting Information

Retrosynthetic Accessibility Score (RAscore) - Rapid Machine Learned Synthesizability Classification from AI Driven Retrosynthetic Planning

Amol Thakkar,[1 2 *] Veronika Chadimová,[1] Esben Jannik Bjerrum,[1] Ola Engkvist,[1] and Jean-Louis Reymond[2]

[1] Hit Discovery, Discovery Sciences, R&D, AstraZeneca, Gothenburg, 431 50, Sweden.

[2] Department of Chemistry and Biochemistry, University of Bern, Bern, CH-3012, Switzerland.

*Corresponding authors: amol.thakkar@dcb.unibe.ch, jean-louis.reymond@dcb.unibe.ch

Contents

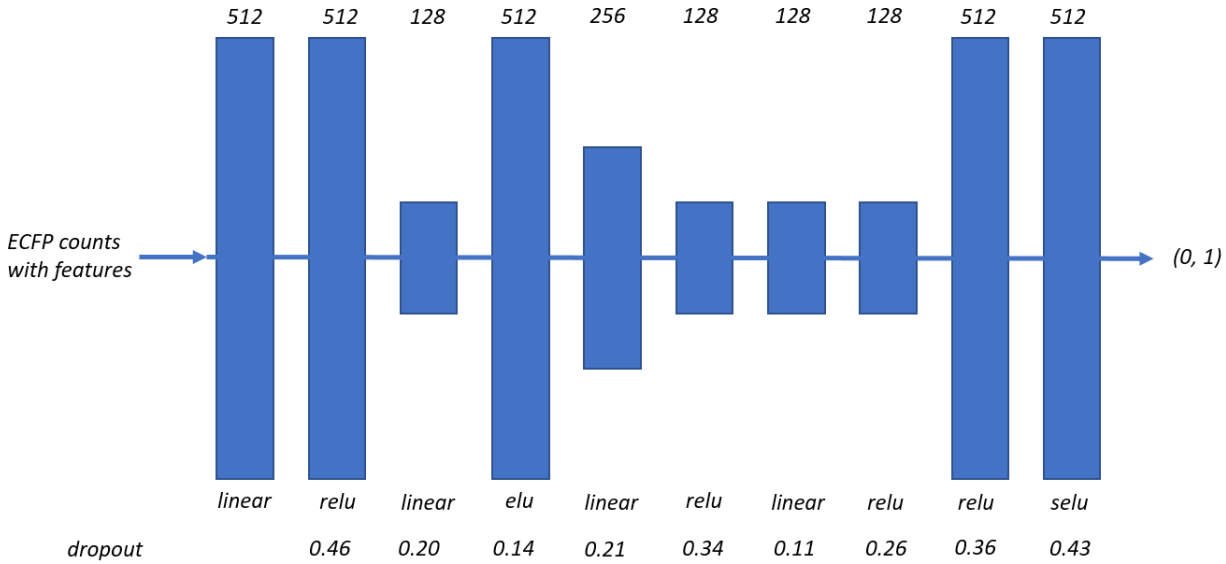
SI-1: Retrosynthesis Prediction for Training Set Generation

Refer to the attached .csv files for the training and test datasets.

Refer to the configuration file found in our GitHub repository for the settings used during the label generation process: https://github.com/reymond-group/RAscore

SI-2: Machine Learning Classifiers for Estimation of Retrosynthetic Accessibility

SI-2.1: Example of the optimal architecture found by hyperparameter optimization for the ChEMBL dataset.



Optimal architecture found for the ChEMBL dataset using Optuna for hyperparameter optimization. The hyperparameters to optimize were, the number of layers, the size of the layers, the activation function, dropout rate, and the learning rate. Full details of the parameters found are given in the SI. As input for the model 2048 dimensional counted ECFPs with features and a radius of 3 were used, and as output the class label, solved or unsolved as obtained via retrosynthetic analysis using our CASP tool.

SI-2.2: Optimal Model Hyperparameters

The following boundaries where used to train the classifiers:

SI-2.2.1: Logistic Regression

"algorithm": {"LogisticRegression": {

"solver": ["newton-cg", "lbfgs", "sag", "saga"],

"C": {

"low": 0.1,

"high": 1.5 }

SI-2.2.2: Random Forest

"algorithm": {"RandomForestClassifier": {

"max_depth": {

"low": 10,

"high": 20

},

"n_estimators": {

"low": 10,

"high": 100

},

"max_features": ["sqrt", "log2"]

}

SI-2.2.3 XGB Classifier

"algorithm": {"XGBClassifier": {

"max_depth": {

"low": 10,

"high": 20

},

"n_estimators": {

"low": 10,

"high": 100

},

"learning_rate": {

"low": 0.05,

"high": 0.2

}

}

SI-2.2.4 Neural Network

"algorithm": {

"DNNClassifier": {

"layer_1": [128, 256, 512],

"activation_1": ["relu", "elu", "selu", "linear"],

"dropout_1": 0.1,

"max_layers": 10,

"layer_size": [128, 256, 512],

"layer_activations": ["relu", "elu", "selu", "linear"],

"layer_droput": {"low": 0,

"high": 0.5},

"learning_rate": {"low": 1e-5,

"high": 1e-1}

}

}

SI-2.2.5 Parameters for NN ecfp counts with features

{"layer_1": 512, "activation_1": "linear", "num_layers": 10, "units_2": 512, "activation_2": "relu", "dropout_2": 0.45834579304621176, "units_3": 128, "activation_3": "linear", "dropout_3": 0.20214636121010582, "units_4": 512, "activation_4": "elu", "dropout_4": 0.13847113009081813, "units_5": 256, "activation_5": "linear", "dropout_5": 0.21312873496871235, "units_6": 128, "activation_6": "relu", "dropout_6": 0.33530504087548707, "units_7": 128, "activation_7": "linear", "dropout_7": 0.11559123444807062, "units_8": 128, "activation_8": "relu", "dropout_8":

0.2618908919792556, "units_9": 512, "activation_9": "relu", "dropout_9": 0.3587291059530903, "units_10": 512, "activation_10": "selu", "dropout_10": 0.43377277017943133, "learning_rate": 1.5691774834712003e-05}

SI-2.2.6: Parameters for NN ecfp counts
{"layer_1": 256, "activation_1": "selu", "num_layers": 2, "units_2": 128, "activation_2": "relu", "dropout_2": 0.15578695546915372, "learning_rate": 2.632240761263429e-05}

SI-2.2.7: Parameters for XGBoost ecfp counts
{"max_depth": 19, "n_estimators": 97, "learning_rate": 0.19984033197055842}

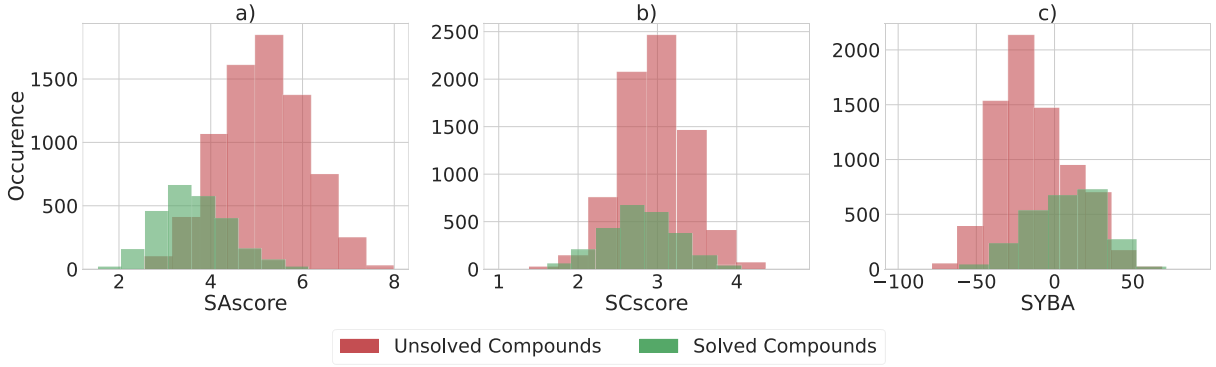
SI-2.2.8 Parameters for SA Score Logistic Regression
{"C": 0.18582521970918675, "solver": "lbfgs"}

SI-2.2.9 Parameters for SC Score Logistic Regression
{"C": 0.22237611805770982, "solver": "saga"}

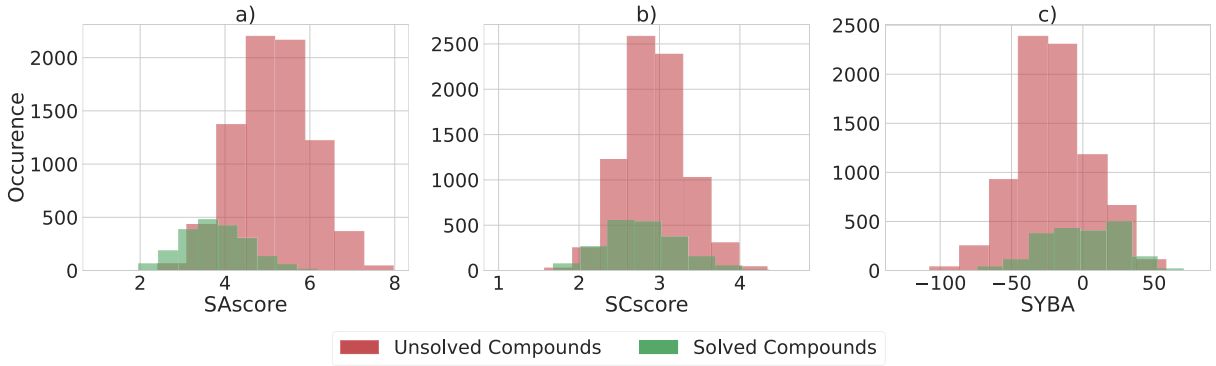
SI-2.2.10 Parameters for SYBA Logistic Regression
{"C": 0.39649336266446344, "solver": "newton-cg"}

SI-3: Attempts at using SAscore, SCscore and SYBA

SI-3.1: GDBChEMBL



SI-3.2: GDBMedChem

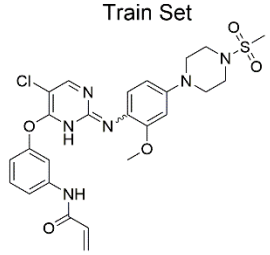
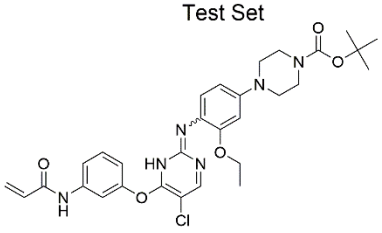
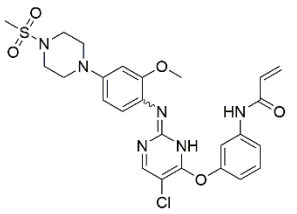
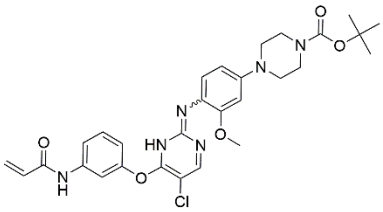
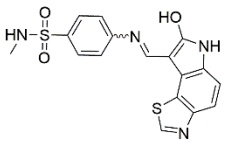
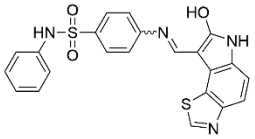
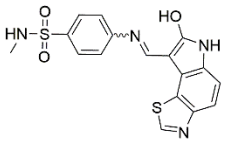
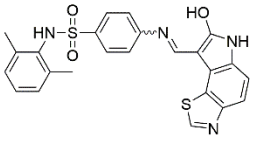
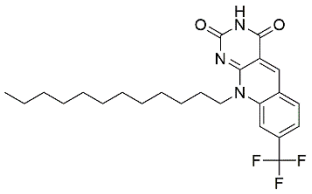
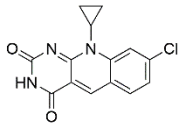
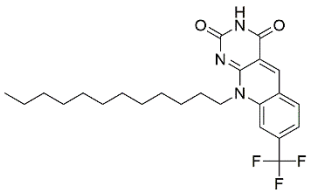
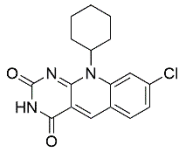
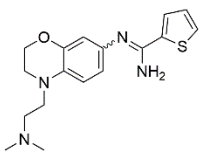
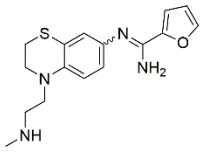
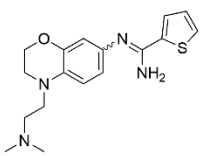


SI-4 Machine Learning Classifiers for Estimation of Retrosynthetic Accessibility

Dataset	Model	Descriptor	ROC-AUC	Accuracy	Precision	Recall	Average Linkage
ChEMBL	NN (RAscore)	ECFP6 counts with features	0.93	0.90	0.92	0.95	0.69
ChEMBL	NN	ECFP6 counts	0.94	0.90	0.92	0.95	0.68
ChEMBL	XGB	ECFP6 counts	0.95	0.91	0.92	0.96	0.65
ChEMBL	XGB	ECFP6 counts with features	0.95	0.90	0.91	0.96	0.62
ChEMBL	RF	ECFP6 counts	0.90	0.83	0.82	0.99	0.30
ChEMBL	RF	ECFP6 counts with features	0.89	0.82	0.82	0.99	0.28
ChEMBL	NN	SA score	0.85	0.81	0.84	0.92	0.37
ChEMBL	Logistic	SA score	0.85	0.81	0.83	0.94	0.36
ChEMBL	Logistic	SC score	0.61	0.75	0.75	1.00	0.27
ChEMBL	NN	SC score	0.61	0.75	0.61	1.00	0.27
ChEMBL	Logistic	SYBA score	0.74	0.78	0.79	0.96	0.25
ChEMBL	NN	SYBA score	0.74	0.78	0.78	0.97	0.21
ChEMBL	-	SAscore	0.15	-	-	-	0.17
ChEMBL	-	SCscore	0.39	-	-	-	0.22
ChEMBL	-	SYBA	0.74	-	-	-	0.17
GDBChEMBL	NN (GDBscore)	ECFP6 counts	0.93	0.87	0.76	0.73	0.64
GDBChEMBL	NN	ECFP6 counts with features	0.94	0.88	0.78	0.74	0.63
GDBChEMBL	XGB	ECFP6 counts	0.94	0.89	0.81	0.73	0.61
GDBChEMBL	XGB	ECFP6 counts with features	0.94	0.88	0.80	0.71	0.61
GDBChEMBL	RF	ECFP6 counts with features	0.89	0.81	0.76	0.40	0.36
GDBChEMBL	RF	ECFP6 counts	0.88	0.81	0.81	0.31	0.32
GDBChEMBL	-	SAscore	0.11	-	-	-	0.26
GDBChEMBL	-	SCscore	0.38	-	-	-	0.14
GDBChEMBL	-	SYBA	0.72	-	-	-	0.17
GDBMedChem	NN	ECFP6 counts	0.93	0.88	0.75	0.64	0.64
GDBMedChem	NN	ECFP6 counts with features	0.94	0.89	0.77	0.66	0.63
GDBMedChem	XGB	ECFP6 counts	0.94	0.89	0.78	0.64	0.61
GDBMedChem	XGB	ECFP6 counts with features	0.94	0.89	0.79	0.64	0.61
GDBMedChem	RF	ECFP6 counts with features	0.89	0.83	0.80	0.27	0.36
GDBMedChem	RF	ECFP6 counts	0.88	0.81	0.91	0.10	0.32
GDBMedChem	-	SAscore	0.13	-	-	-	0.22
GDBMedChem	-	SCscore	0.39	-	-	-	0.14
GDBMedChem	-	SYBA	0.70	-	-	-	0.17

SI -5: Limitations of Template Based CASP Tools

SI-5.1: Example Compound Similarity to Training Set

Train Set	Solved	Tanimoto	Test Set
	1	0.77	
	1	0.83	
	1	0.74	
	1	0.73	
	0	0.46	
	0	0.45	
	0	0.70	
	0	0.61