Electronic Supplementary Information

Machine Learning the Ropes:

Principles, Applications and Directions in Synthetic Chemistry

Felix Strieth-Kalthoff, Frederik Sandfort, Marwin H. S. Segler* and Frank Glorius*

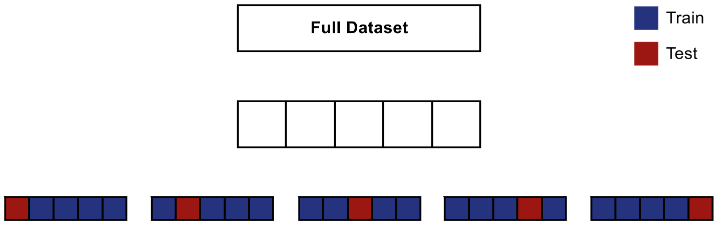
Westfälische Wilhelms-Universität Münster, Organisch-Chemisches Institut, Corrensstr. 40, 48149
Münster (Germany).

1 Fundamentals of Machine Learning

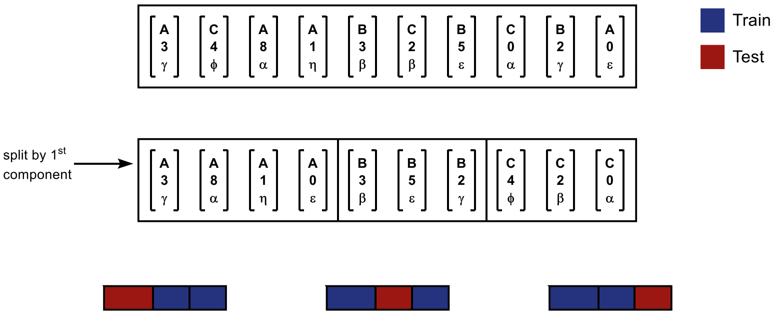
Cross Validation and Correlation Metrics

Cross validation is a statistical procedure to confirm the validity of discovered correlations. As such, the same (or similar) correlation should be observed for different train/test splits.

The most prominent cross validation scheme is **k-fold cross validation**. In this approach, the data set is split into *k* (in the example: *k* = 5) equally sized subsamples. In *k* independent cross validation steps, all but one subsamples are used as training to predict the remaining subsample.



Leave-one-out cross validation is, strictly speaking a borderline case of *k*-fold cross validation, in which *k* equals the number of data points. In a chemical context, especially for multi-component data sets, the concept of leave-one-out cross validation can be applied to selected features (e.g. explicit components) only. The following example shows leave-one-out cross validation by feature 1.



A third approach for cross validation is **random cross validation**, also referred to as **Monte-Carlo cross validation**. Therefore, for each cross validation set, a new random test-train spilt is generated, using a pre-defined split ratio.

For a regression task, correlation between predicted and observed values can be measured through the squared Pearson Correlation coefficient, given as:

$$r^2 = \frac{\left(\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})\right)^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2 \cdot \sum_{i=1}^{n}(y_i - \bar{y})^2}$$

where x_i and y_i describe the predicted and observed values, respectively, and $\bar{x}$ and $\bar{y}$ stand for the mean value of predicted and observed values, respectively.

Similarly, the mean average error (MAE), and the root mean square error (RMSE), can be calculated:

$$\mathrm{MAE} = \frac{1}{n} \cdot \sum_{i=1}^{n} |y_i - x_i|$$

$$\mathrm{RMSE} = \sqrt{\mathrm{MSE}} = \sqrt{\frac{1}{n} \cdot \sum_{i=1}^{n} (y_i - x_i)^2}$$

For classification tasks, a variety of evaluation metrics exists, and their calculation depends on the number of classes. Considering the special case of a binary classification (positive or negative), four different cases are possible: true positive (TP), false positive (FP), true negative (TN), false negative (FN) (Table S1). One of the most essential metrics is the accuracy which is the proportion of correct among all results:

$$\mathrm{accuracy} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{FP} + \mathrm{FN} + \mathrm{TN}}$$

However, using the accuracy as a classification metric is often problematic for imbalanced problems, i.e. problems with only few actual positives and many actual negatives. In such cases, other metrics have to be applied, for example balanced accuracy, precision, recall, Matthew’s correlation coefficient, the Area Under the Receiver Operating Characteristic Curve (ROC AUC), or the Area under the Precision Recall Curve.

Table S1: Possible results of a binary classification.

		Actual	
		Positive	**Negative**
Predicted	**Positive**	True positive (TP)	False Positive (FP)
	Negative	False negative (FN)	True negative (TN)

Machine Learning Algorithms

For a detailed discussion of machine learning algorithms, including the concept, mathematical background, and computational implementations, the reader is referred to textbooks of statistical learning / machine learning (references 51–54), as well as educative blog entries (reference 55, for introduction only!). Table S2 provides an overview of some of the most relevant algorithms for supervised machine learning, along with the general concepts, critical aspects and hyperparameters, as well as additional illustrative literature references. However, it should be noted that this exemplified overview does not claim to be comprehensive.

Table S2: Overview of algorithms for supervised machine learning tasks, as used in the context of (molecular) chemistry.

Algorithm	Fundamental Concept	Additional References
(Multivariate) Linear Regression	target variable is expressed as linear combination of features	
Logistic Regression	classification based on linear regression, using a logistic function	
Bayesian Classifiers	probabilistic classification based on prior knowledge, using Bayes' theorem	56
Decision Tree Learning	flowchart-like diagrams with branching decision nodes, which predict the output as a sequence of queries	
Bagging (Bootstrap Aggregating)	ensemble of predictors, e.g. decision trees, each one trained on a bootstrap sample of the whole training data	
Random Forest	special case of bagging, in which tree branches are generated using a random subset of features	57
Gradient Tree Boosting	ensemble of sequentially generated decision trees, each tree focusing on the errors of previous trees	
Feedforward Artificial Neural Networks	fully interconnected layers of neurons, each neuron receives signals from neurons of previous layer and processes them through non-linear activation functions	8,58
Recurrent Artificial Neural Networks	neural network in which recursive node connections (within a layer, from a node to itself) are possible, possess time-dependent behavior	
k-Nearest Neighbours	instance-based algorithm, target of a data point is predicted based on its *k* nearest neighbours in feature space, distance-based	
Kernel Ridge Regression	linear regression using Kernel functions which map the data into a higher-dimensional space	59
Support Vector Machine	separation of classes of data points through linear hyperplane, uses Kernel functions for mapping into higher-dimensional space, where linear separation can be achieved	60

References

51 I. H. Witten, E. Frank and M. A. Hall, *Data Mining: Practical Machine Learning Tools and Techniques*, Morgan Kaufmann, Burlington (MA), 2011.

52 C. M. Bishop, *Pattern Recognition and Machine Learning*, Springer, New York, 2006.

53 K. P. Murphy, *Machine Learning: A Probabilistic Perspective*, MIT Press, Cambridge (MA), 2012.

54 S. L. Brunton and J. N. Kutz, *Data Driven Science and Engineering*, MIT Press, Cambridge (MA), 2019.

55 A series of useful tutorials for getting started with machine learning algorithms can be found online at https://towardsdatascience.com (accessed: March 2020).

56 A. Bender, *Methods Mol. Biol.*, 2011, **672**, 175–196.

57 V. Svetnik, A. Liaw, C. Tong, J. C. Culberson, R. P. Sheridan and B. P. Feuston, *J. Chem. Inf. Comput. Sci.*, 2003, **43**, 1947–1958.

58 J. Gasteiger and J. Zupan, *Angew. Chem. Int. Ed. Engl.*, 1993, **32**, 503–527.

59 K. Vu, J. C. Snyder, L. Li, M. Rupp, B. F. Chen, T. Khelif, K.-R. Müller and K. Burke, *Int. J. Quantum Chem.*, 2015, **115**, 1115–1128.

60 W. S. Noble, *Nat. Biotechnol.*, 2006, **24**, 1565–1567.

2 Representing Organic Molecules

For a more detailed introduction to molecular representations, the reader is referred to Chemoinformatics textbooks (references 61, 62).

Molecular Fingerprints

With the ascent of compound and reaction databases, different fingerprint-based models have been developed. Most algorithms have focused either on substructure keys fingerprints, which encode given molecular fragments as queries, or topological fingerprints, which try to encode the topology of the molecular graph. Selected examples include:

Substructure Keys Fingerprints:

- MACCS Keys Fingerprint (reference 63)
- PubChem Fingerprint (reference 64)
- BCI Fingerprint (reference 65)

Topological Fingerprints:

- Morgan Fingerprints / Extended-Connectivity Fingerprints (references 12, 66)
- Daylight Fingerprints (reference 67)
- Atom Pairs Fingerprint (reference 68)

Jaeger *et al.* recently introduced a fingerprint-like feature vector, which was trained by unsupervised learning. For further details, see reference 69.

The concept of extended-connectivity fingerprints was recently extended to 3D structure encoding, see reference 70.

3D Representations

Building on the concept of coulomb matrices, further 3D representations have been developed. The classification into bags of (similar) bonds is described in reference 71, the concept was later extended to many-body potentials, in order to include angle and dihedral potentials (reference 72). Moreover, histogram-based methods, which build on the distribution of certain bonds, angles or dihedrals within the training data set, have been described (see reference 14).

References

61 J. Gasteiger, T. Engel, *Chemoinformatics: A Textbook*, Wiley-VCH, Weinheim, 2003.

62 A. R. Leach, V. J. Gillet, *An Introduction To Chemoinformatics*, Springer, Dordrecht, 2007.

63 J. L. Durant, B. A. Leland, D. R. Henry and J. G. Nourse, *J. Chem. Inf. Comput. Sci.*, 2002, **42**, 1273–1280.

64 E. E. Bolton, Y. Wang, P. A. Thiessen and S. H. Bryant, *Annu. Rep. Comput. Chem.*, 2008, **4**, 217–241.

65 J. M. Barnard and G. M. Downs, *J. Chem. Inf. Model.*, 1997, **37**, 141–142.

66 H. L. Morgan, *J. Chem. Doc.*, 1965, **5**, 107–113.

67 Daylight Chemical Information Systems. https://www.daylight.com/ (accessed: March 2020).

68 R. E. Carhart, D. H. Smith and R. Venkataraghavan, *J. Chem. Inf. Comput. Sci.*, 1985, **25**, 64–73.

69 S. Jaeger, S. Fulle and S. Turk, *J. Chem. Inf. Model.*, 2018, **1**, 27–35.

70 S. D. Axen, X.-P. Huang, E. L. Cáceres, L. Gendelev, B. L. Roth and M. J. Keiser, *J. Med. Chem.*, 2017, **60**, 7393–7409.

71 K. Hansen, F. Biegler, R. Ramakrishnan, W. Pronobis, O. A. von Lilienfeld, K.-R. Müller, A. Tkatchenko, *J. Phys. Chem. Lett.*, 2015, **6**, 2326–2331.

72 B. Huang and O. A. von Lilienfeld, *J. Chem. Phys.*, 2016, **145**, 161102.

3 Property Prediction, Activity Prediction, and Molecular Design

Quantitative Prediction of Properties and Activities

Quantitative structure activity- and structure-property prediction models have had a longstanding history within chemistry, for the beginnings and recent reviews, see references 73–74 and 75–77, respectively.

Molecular Design

For further reading on virtual screening and the need for molecule libraries, see references 78–82.

Early approaches towards *de-novo* design can be found in reference 82 and references therein.

The scoring and benchmarking of generative modelling approaches is an ongoing challenge (reference 84), albeit examples of prospective usage in the context of drug discovery have been proven (references 85–86).

References

73 C. Hansch, P. P. Maloney, T. Fujita and R. M. Muir, *Nature*, 1962, **194**, 178–180.

74 C. Hansch, *Acc. Chem. Res.*, 1969, **2**, 232–239.

75 R. D. Cramer, *J. Comput. Aided Mol. Des.*, 2012, **26**, 35–38.

76 A. Cherkasov, E. N. Muratov, D. Fourches, A. Varnek, I. I. Baskin, M. Cronin, J. Dearden, P. Gramatica, Y. C. Martin, R. Todeschini, V. Consonni, V. E. Kuz'min, R. Cramer, R. Benigni, C. Yang, J. Rathman, L. Terfloth, J. Gasteiger, A. Richard and A. Tropsha, *J. Med. Chem.*, 2014, **57**, 4977–5010.

77 J. Ma, R. P. Sheridan, A. Liaw, G. E. Dahl and V. Svetnik, *J. Chem. Inf. Model.*, 2015, **55**, 263–274.

78 J.-L. Reymond, L. Ruddigkeit, L. Blum and R. van Deursen, *WIRES Comput. Mol. Sci.*, 2012, **2**, 717–733.

79 B. K. Shoichet, *Nature*, 2004, **432**, 862–865.

80 W. P. Walters, *J. Med. Chem.*, 2019, **62**, 1116–1124.

81 N. van Hilten, F. Chevillard and P. Kolb, *J. Chem. Inf. Model.*, 2019, **59**, 644–651.

82 M. Hartenfeller and G. Schneider, *WIREs Comput. Mol. Sci.*, 2011, **1**, 742–759.

83 X. Yang, Y. Wang, R. Byrne, G. Schneider and S. Yang, *Chem. Rev.*, 2019, **119**, 10520–10594.

84 N. Brown, M. Fiscato, M. H. S. Segler and A. C. Vaucher, *J. Chem. Inf. Model.*, 2019, **59**, 1096–1108.

85 D. Merk, L. Friedrich, F. Grisoni and G. Schneider, *Mol. Inf.*, 2018, **37**, 1700153.

86 Y. Yang, R. Zhang, Z. Li, L. Mei, S. Wan, H. Ding, Z. Chen, J. Xing, H. Feng, J. Han, H. Jiang, M. Zheng, C. Luo and B. Zhou, *J. Med. Chem.*, 2020, **63**, 1337–1360

4 Computer-Aided Synthesis Planning

Reaction Representation, Assessment and Classification

For a pioneering study on automated reaction role assignment and atom-to-atom mapping, see reference 87.

Reaction classification using machine learning and reaction fingerprints is described in reference 88.

Concept of Computer-Aided Synthesis Planning

For reviews on computer-aided synthesis planning, please refer to references 89–90. The underlying concepts as well as early computational approaches to CASP programs are given in references 91–95.

Prediction of Reaction Products / Retrosynthetic Disconnections

Rule-based methodologies for product / precursor prediction have been the most commonly used approach. More information on automated rule extraction is given in reference 96. For a rule-based product prediction study on textbook reactions of alkyl halides, see reference 97. A recent example of rule ranking for retrosynthetic disconnections using graph convolutional networks can be found in reference 98.

Rule-based product prediction on the mechanistic level has been pioneered by the Baldi group (reference 35,99): The initial study combined two neural networks, the first identifying electron sources and sinks and the second ranking all possible combinations of these. A recent report picks up the concept of reaction product prediction by learning electron movements but circumvents the need for hand-coded elementary reactions. The model is currently limited to predicting linear electron flow (LEF) reactions, i.e. reactions that can be represented by alternating bond removing and bond forming events. Based on learnable graph representations, the model was trained to identify the atom which is most likely to react. The bond addition and removing sequence is eventually terminated by a second neural network which learned to identify the end of a reaction. This approach was found to outperform other solely data-driven approaches for predicting LEF reactions (reference 100).

More examples of predicting reaction products and retrosynthetic disconnections without the use of reaction templates are given in reference 101 (graph-edit based) and references 102–105 (sequence-to-sequence models).

An early example of predicting reaction conditions using machine learning for the Michael addition can be found in reference 106. For examples of more general condition prediction tools (as applied by Coley *et al.*, reference 34), please refer to references 107,108.

Tree Search Algorithms

Further details regarding reinforcement learning and Monte Carlo tree search as used in CASP programs are depicted in references 109,110. Besides MCTS, policy optimization via reinforcement learning has been applied to tree search in CASP. Most likely retrosynthetic disconnections were generated for a large library of compounds, using a defined policy. Every route was assigned a cost by a user-defined cost function (i.e. synthesis length, synthesis complexity), and the disconnection policy is updated in order to minimize the synthetic cost. By iteratively repeating this learning procedure, the choice of retrosynthetic disconnection could be significantly improved (reference 111).

References

87 N. Schneider, N. Stiefl and G. A. Landrum, *J. Chem. Inf. Model.*, 2016, **56**, 2336–2346.

88 N. Schneider, D. M. Lowe, R. A. Sayle and G. A. Landrum, *J. Chem. Inf. Model.*, 2015, **55**, 39–53.

89 M. H. Todd, *Chem. Soc. Rev.*, 2005, **34**, 247–266.

90 C. W. Coley, W. H. Green and K. F. Jensen, *Acc. Chem. Res.*, 2018, **51**, 1281 – 1289.

91 E. J. Corey and W. T. Wipke, *Science*, 1969, **166**, 178–192.

92 J. Bauer, E. Fontain, D. Forstmeyer and I. Ugi, *Tetrahedron Comput. Methodol.*, 1988, **1**, 129–132.

93 I. Ugi, E. Fontain and J. Bauer, *Anal. Chim. Acta*, 1990, **235**, 155–161.

94 P. Röse and J. Gasteiger, *Anal. Chim. Acta*, 1990, **235**, 163–168.

95 H. Gelernter, J. R. Rose and C. Chen, *J. Chem. Inf. Comput. Sci.*, 1990, **30**, 492–504.

96 J. Law, Z. Zsoldos, A. Simon, D. Reid, Y. Liu, S. Y. Khew, A. P. Johnson, S. Major, R. A. Wade and H. Y. Ando, *J. Chem. Inf. Model.*, 2009, **49**, 593–602.

97 J. N. Wei, D. Duvenaud and A. Aspuru-Guzik, *ACS Cent. Sci.*, 2016, **2**, 725–732.

98 S. Ishida, K. Terayama, R. Kojima, K. Takasu and Y. Okuno, *J. Chem. Inf. Model.*, 2019, **59**, 5026–5033.

99 D. Fooshee, A. Mood, E. Gutman, M. Tavakoli, G. Urban, F. Liu, N. Huynh, D. Van Vranken and P. Baldi, *Mol. Syst. Des. Eng.*, 2018, **3**, 442–452.

100 J. Bradshaw, M. J. Kusner, B. Paige, M. H. S. Segler and J. M. Hernández-Lobato, 2019, arXiv:1805.10970.

101 K. Do, T. Tran and S. Venkatesh, 2018, arXiv:1812.09441.

102 J. Nam and J. Kim, 2016, arXiv:1612.09529.

103 P. Schwaller, T. Gaudin, D. Lányi, C. Bekas and T. Laino, *Chem. Sci.*, 2018, **9**, 6091–6098.

104 A. A. Lee, Q. Yang, V. Sresht, P. Bolgar, X. Hou, J. L. Klug-McLeod and C. R. Butler, *Chem. Commun.*, 2019, **55**, 12152–12155.

105 P. Karpov, G. Godin and I. Tetko, *ChemRxiv Preprint*, 2019, DOI: 10.26434/chemrxiv.8058464.v1.

106 G. Marcou, J. A. de Sousa, D. A. R. S. Latino, A. de Luca, D. Horvath, V. Rietsch and A. Varnek, *J. Chem. Inf. Model.*, 2015, **55**, 239–250.

107 H. Gao, T. J. Struble, C. W. Coley, Y. Wang, W. H. Green and K. F. Jensen, *ACS Cent. Sci.*, 2018, **4**, 1465–1476.

108 E. Walker, J. Kammeraad, J. Goetz, M. T. Robo, A. Tewari and P. M. Zimmerman, *J. Chem. Inf. Model.*, 2019, **59**, 3645–3654.

109 R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction*, MIT Press, Cambridge (MA), 2015.

110 C. B. Browne et al., *IEEE Trans. Comput. Intell. AI Games*, 2012, **4**, 1–43.

111 J. S. Schreck, C. W. Coley and K. J. M. Bishop, *ACS Cent. Sci.*, 2019, **5**, 970–981.

5 Quantitative Prediction of Reaction Outcomes

The quantitative assessment of reaction outcomes (particularly selectivities) has been a focus for almost a century, originating from Hammett's seminal works (see reference 112). In recent years, the group of Sigman has driven multivariate linear regression techniques forward, for further reading, see references 38,113–118.

Additional reports on the quantitative prediction of reaction outcomes using machine learning are given in references 119–121. In particular, classification of reactions into high/low-yielding (ref. 119), yield prediction for a deoxyfluorination reaction (ref. 120) as well as site selectivity in C – H functionalization reactions (ref. 121) have been investigated.

112 L. P. Hammett, *J. Am. Chem. Soc.*, 1937, **59**, 96–103.
113 M. S. Sigman, K. C. Harper, E. N. Bess and A. Milo, *Acc. Chem. Res.*, 2016, **49**, 1292–1301.
114 A. Milo, E. N. Bess and M. S. Sigman, *Nature*, 2014, **507**, 210–214.
115 K. C. Harper and M. S. Sigman, *Science*, 2011, **333**, 1875–1878.
116 A. Milo, A. J. Neel, F. D. Toste and M. S. Sigman, *Science*, 2015, **347**, 737–743.
117 E. N. Bess, A. J. Bischoff and M. S. Sigman, *Proc. Natl. Acad. Sci. USA*, 2014, **111**, 14698–14703.
118 J. P. Reid and M. S. Sigman, *Nature*, 2019, **571**, 343–348.
119 G. Skoraczyński, P. Dittwald, B. Miasojedow, S. Szymkuć, E. P. Gajewska, B. A. Grzybowski and A. Gambin, *Sci. Rep.*, 2017, **7**, 3582.
120 M. K. Nielsen, D. T. Ahneman, O. Riera and A. G. Doyle, *J. Am. Chem. Soc.*, 2018, **140**, 5004–5008.
121 T. J. Struble, C. W. Coley and K. F. Jensen, *ChemRxiv Preprint*, 2019, DOI: 10.26434/chemrxiv.9735599.v1.

6 Active Learning

Active learning strategies have become popular for both, continuous and discrete optimization problems.

For a review on active learning in drug discovery, please refer to reference 122. Further applications of active learning in QSAR and reaction optimization are given in references 123 and 124. An approach to reaction optimization based on reinforcement learning using the result of one reaction as the feedback function can be found in reference 125. Here, a policy is trained to decide which changes have to be made to the experimental conditions in order to obtain the optimum yield. The model is however limited, as it is based on reaction outcomes of consecutive single experiments and was thus only applied to microdroplet reactions.

[122] D. Reker and G. Schneider, *Drug Discov. Today*, 2015, **20**, 458–465.
[123] D. Reker, P. Schneider and G. Schneider, *Chem. Sci.*, 2016, **7**, 3919–3927.
[124] D. Reker, G. J. L. Bernardes and T. Rodrigues, *ChemRxiv Preprint*, 2018, DOI: 10.26434/chemrxiv.7291205.v1.
[125] Z. Zhou, X. Li and R. N. Zare, *ACS Cent. Sci.*, 2017, **3**, 1337–1344.

7 Data Acquisition and Data Quality

For further reading on contemporary high-throughput experimentation in batch and flow, see references 126–128. More details on reaction assessment via condition- or additive-based screening approaches can be found in references 129–131.

126 A. Buitrago Santanilla, E. L. Regalado, T. Pereira, M. Shevlin, K. Bateman, L.-C. Campeau, J. Schneeweis, S. Berritt, Z.-C. Shi, P. Nantermet, Y. Liu, R. Helmy, C. J. Welch, P. Vachal, I. W. Davies, T. Cernak and S. D. Dreher, *Science*, 2015, **347**, 49–53.

127 D. Perera, J. W. Tucker, S. Brahmbhatt, C. J. Helal, A. Chong, W. Farrell, P. Richardson and N. W. Sach, *Science*, 2018, **359**, 429–434.

128 A.-C. Bédard, A. Adamo, K. C. Aroh, M. G. Russell, A. A. Bedermann, J. Torosian, B. Yue, K. F. Jensen and T. F. Jamison, *Science*, 2018, **361**, 1220–1225.

129 L. Pitzer, F. Schäfers and F. Glorius, *Angew. Chem. Int. Ed.*, 2019, **58**, 8572–8576.

130 T. Gensch, M. Teders and F. Glorius, *J. Org. Chem.*, 2017, **82**, 9154–9159.

131 T. Gensch and F. Glorius, *Science*, 2016, **352**, 294–295.