Artificial Intelligence in the Life Sciences 1 (2021) 100009

Contents lists available at ScienceDirect

Artificial Intelligence in the Life Sciences

journal homepage: www.elsevier.com/locate/ailsci

Conceptual Analysis

Chemistry-centric explanation of machine learning models

Raquel Rodríguez-Pérez[1], Jürgen Bajorath[2,*]

Department of Life Science Informatics, B-IT, LIMES Program Unit Chemical Biology and Medicinal Chemistry, Rheinische Friedrich-Wilhelms-Universität, Friedrich-Hirzebruch-Allee 6, Bonn D-53115, Germany

Artificial intelligence (AI) is increasingly being considered across chemical disciplines, just as in many other areas of science, often with high expectations for "revolutionary" advances. Medicinal chemistry and drug design are among the focal points of these developments [1–4] and there is rising interest in deep neural network (DNN) architectures and deep learning (DL) for the generation of novel compounds and prediction of various molecular properties [3,4].

In chemoinformatics, medicinal chemistry, and drug design, the term AI is typically used synonymously with machine learning (ML), which represents only a part of the methodological AI spectrum [5]. However, ML already has a long history in chemoinformatics and medicinal chemistry, spanning more than two decades, and is widely applied for molecular property predictions as well as the search for novel active compounds [6]. Neural networks (NNs) were popular early on in chemoinformatics, but have been increasingly replaced over time with other approaches [7] such as the support vector machine (SVM) [8] and random forest (RF) [9] algorithms or Bayesian modeling [10], mostly due to the intrinsic tendency of NNs to overfit property prediction models trained on moderately sized data sets. With DL, NNs have re-emerged in this field, for the most part as DNN architectures [11].

In computer science, it is often observed that methodological complexity does not scale with predictive performance [12] and the same applies to medicinal chemistry and drug design [6]. Although "big data" trends are also beginning to emerge in medicinal chemistry [13], ML predictions are typically based upon relatively small data regimes and well-defined molecular representations. These conditions do not play into the strengths of DNNs compared to other areas such as image analysis or natural language processing where DL has made a large impact in recent years [14,15]. Accordingly, standard ML methods often perform comparably well or better than DNNs in predicting biological activity and other molecular properties [6].

On the other hand, DL offers new opportunities in chemoinformatics and medicinal chemistry that were hardly possible to address in a comparable way until recently such as, for example, in chemical reaction modeling [1] or large-scale generative *de novo* compound design [3]. In generative design and DNN-based compound repositioning, successful applications charting new territory are beginning to appear [16] but other claimed advances remain often questionable [17], partly due to the lack of generally accepted evaluation criteria and standards for assessing chemical novelty [17]. Even in successful design applications, newly generated or repurposed chemical entities are often viewed controversially from a medicinal chemistry perspective. Clearly, the field is in flux but ML/DL in medicinal chemistry and drug design is still largely dominated by methodological considerations (what could be done?) rather than successful practical applications (what has been done?), which are still far from being routine [18].

One of two particularly relevant aspects concerning ML in medicinal chemistry emerging from the discussion above is that a variety of methods are being employed for property prediction and compound design, ranging from simple decision tree-based algorithms and probabilistic modeling to complex DNNs, often with comparable success. The other important aspect relates to the rationalization of predictions, as further discussed in the following.

Given the often cited "black box" character of most ML models [12,19], their predictions are difficult to rationalize. In the practice of medicinal chemistry, lack of transparency of predictions and model rationalization continue to hinder the acceptance of ML and limit the impact of predictive modeling on experimental programs [6,18], despite the long tradition of ML in this field. Similar considerations apply to other chemical disciplines that are primarily experimentally driven. Ideally, one would like to know how certain and reliable a given prediction is and also understand it in intuitive chemical terms. While a few studies have presented approaches for uncertainty estimation of ML predictions [20,21], robust and generally applicable methods are currently not yet available. However, in computer science and other fields, the potential of explainable or interpretable ML is a much discussed topic [12,22,23], which also is of high relevance for ML applications in chemistry. In computer science, interpretable ML refers to algorithms whose predictions can be directly reconciled (such as decision trees) and explainable ML to methods enabling the rationalization of black box models [12]. However, in chemistry-related publications, these terms have been used more or less interchangeably concerning the rationalization of models and their decisions.

* Corresponding author.
E-mail address: bajorath@bit.uni-bonn.de (J. Bajorath).
[1] Present address: Novartis Institutes for Biomedical Research, Novartis Campus, CH-4002 Basel.
[2] Given his role as Editor in Chief, Jürgen Bajorath had no involvement in the peer-review of this article and has no access to information regarding its peer-review. Full responsibility for the editorial process for this article was delegated to Mingyue Zheng.

https://doi.org/10.1016/j.ailsci.2021.100009
Received 20 September 2021; Received in revised form 28 September 2021; Accepted 28 September 2021
Available online 30 September 2021

In medicinal chemistry, there is an urgent need for ML model rationalization. Although the reluctance of medicinal chemists to rely on black box predictions in their compound design and optimization efforts is a widespread phenomenon, only a limited number of investigations have addressed the explanation or interpretation of ML predictions in the context of structure-activity relationship (SAR) analysis [24]. With increasing use of complex DNN architectures in the field [3,4], this widens the gap between model availability and acceptance and further limits the impact of ML.

For activity prediction models, available rationalization strategies typically aim to identify molecular representation features that determine individual predictions, mostly in a model-dependent manner. These approaches predominantly employ feature weighting techniques to rationalize predictions of kernel-based [25,26] or Bayesian [27] classification models. In addition, weight gradients across different NN layers can also be determined to trace determinants of predictions and explain (D)NN models. However, these gradients tend to be unstable, frequently resulting in different interpretations of very similar predictions [28].

In addition to model-dependent techniques, model-independent methods can also be considered, which are principally preferred because they are applicable to any ML model, regardless of the complexity of the underlying algorithm. Moreover, model-independent approaches typically do not require balancing performance and interpretability across different models [29]. However, with sensitivity analysis [30], only one model-independent approach has until recently been applied to property predictions in chemoinformatics. This methodology was adapted about a decade ago to study the influence of systematic feature value changes on activity predictions using ML models [31,32]. Sensitivity analysis generally relies on feature perturbation to study ensuing effects. In particular, partial derivatives were used to assess the impact of local perturbations of fragment descriptors on model predictions [32]. Sensitivity analysis becomes rapidly infeasible with increasing model dimensionality, which limits its applicability. Hence, this methodology has been more or less abandoned in the field or substituted by approximations that better summarize perturbations effects.

Our group has considered alternative approaches with potential for ML model-independent rationalization from different viewpoints including broad applicability, quantitative assessment of predictions, and ease of visual interpretation. In light of these criteria, the concept of *Shapley values* [33,34] from cooperative game theory [34,35] has become our methodology of choice. Shapley values were originally introduced in 1953 [33] to quantitatively account for contributions of individual players forming a team. These values provide a quantitative assessment of cooperative contributions to a team's ultimate success (total gain). Accordingly, they specify a partition of merit among individual players by calculating the average of all contributions made by a given player in different team constellations.

The Shapley value concept is readily transferable to ML by applying the following analogies: The game a team engages in can perceived as a prediction task for a single instance (e.g., a compound). The merit for this task is given by the difference between its prediction and the average prediction of all instances. The players participating in the game are features values of the instance that cooperate (act in concert) to obtain the merit for a given prediction. The resulting Shapley value of a given feature is then obtained as the average contribution of a feature over all possible feature combinations. Accordingly, Shapley values account for the partition of contributions over individual features comprising a feature vector or set (such as a molecular representation). A key aspect of the Shapley value concept that sets it apart from model-dependent feature weighting methods is that not only the contribution of feature presence to a given prediction can be quantified, but also the contribution of feature absence.

For feature sets of increasing size, systematic calculations of Shapley values on the basis of all possible feature combinations become computationally increasingly expensive. Therefore, to render the approach practical for ML, a *locally interpretable explanatory model* is derived that limits required Shapely value calculations. This methodology is termed *Shapley Additive exPlanations* (*SHAP*) [36] and can be perceived as an extension of the *Local Interpretable Model-agnostic Explanations* (LIME) approach [37]. A brief account of the underlying theory is presented in the **Appendix** below.

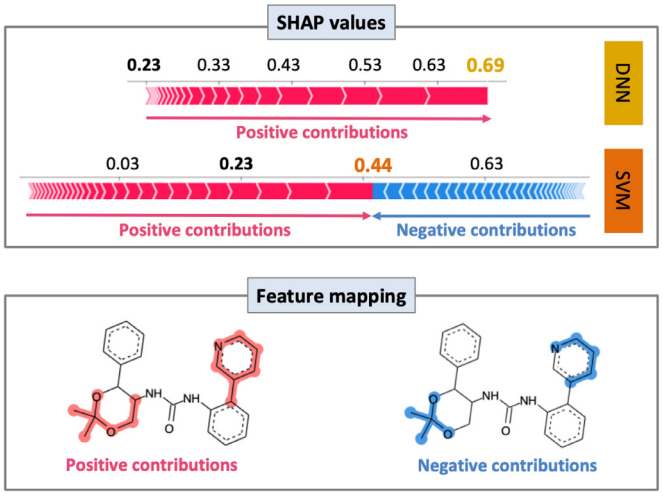

Fig. 1. Interpretation of predictions. SHAP feature importance values are represented as sequential arrows recording positive (red) and negative (blue) contributions to the prediction of activity, resulting in a cumulative output probability. For ML, the given compound was represented using a topological fingerprint comprising overlapping atom environment features. Top-ranked features making positive or negative contributions are mapped onto the structure. The activity of the compound was only correctly predicted using a DNN but not an SVM model. The figure was adapted from Rodríguez-Pérez and Bajorath [39]. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article).

We have adapted the SHAP methodology for explaining compound activity predictions and SARs [38]. To these ends, SHAP was combined with molecular feature mapping and feature classification according to quantitative contributions to correct or incorrect ML model predictions [38] and successfully applied to different compound activity prediction tasks including classification, regression, or multi-task learning [38,39]. Fig. 1 shows exemplary results of molecular SHAP analysis.

Shown is an active compound whose biological activity that was correctly predicted by a DNN model whereas an SVM classifier predicted this compound to be inactive. The SHAP contribution plots of these predictions reveal that multiple features were assigned comparable importance for predictions with both models. However, the SVM model yielded negative contributions for a number of features that were not prioritized by the DNN. Mapping of highly weighted positive (red) and negative (blue) features from the SVM prediction on the compound structure revealed that these features formed overlapping substructures, thus detecting and explaining a model error.

Recently, an algorithm was introduced for the exact calculation of local SHAP values specifically with decision tree-based methods, taking advantage of the tree distribution [40]. By comparing predictions using SHAP and this algorithm directly, high correlation of greater than 80% was observed between approximated and exactly determined local feature importance values, both for decision tree-based classification and regression models [39]. These findings further supported the reliability of SHAP's local kernel approximation.

We also note that there are conceptually distinct yet complementary approaches to SHAP for model interpretation. For example, the *Similarity Maps* approach depends on the removal of individual atoms from test compounds and estimation of their importance by quantifying the

resulting change of a prediction [41]. This is in contrast to SHAP that does not rely on atom removal (i.e., introduction of a perturbation at the structural level), but assesses collaborative (i.e., combination-based) contributions of representation features. The Similarity Maps approach does not require feature mapping but cannot quantify contributions of feature absence. In addition, single-atom perturbations might not significantly modify a model's prediction but affect structural integrity.

Another conceptually distinct approach involves feature weighting of multiple models generated with the same ML algorithm, providing the basis for the determination of *feature importance correlation* [42]. For target-based compound activity classes, this measure yielded model-internal data set signatures and revealed similar compound binding characteristics of proteins as well as functional relationships. Strong feature importance correlation between compound activity prediction models indicated functional similarities between different targets that were not related to compound binding [42]. For RF models derived for more than 200 targets, the Gini impurity (GI) criterion [43] served as a measure of node-based recursive partitioning quality. GI is a metric from information theory defined in Eq. (1):

$$\mathrm{GI} = \sum_{i=1}^{n} p_i \left(1 - p_i\right) \tag{1}$$

Here, p_i is the frequency for class i at a given node, and n is 2 for binary classification. Accordingly, GI for a given feature is equivalent to the mean decrease in GI, i.e., the normalized sum of all impurity decrease values for nodes in the RF where splitting was based on that feature. Thus, increasing values indicate increasing feature importance for the RF model [43]. Correlation or statistical association across feature importance values from different models is then quantified using correlation coefficients.

Feature importance correlation analysis only requires model-internal information, but does not depend on model explanations. Feature weights or importance values can be extracted using multiple strategies, depending on the underlying ML algorithm. Moreover, in contrast to SHAP-based analysis of individual predictions, feature importance correlation analysis is based upon global model assessment, taking many compounds into account. While SHAP and feature importance correlation analysis are conceptually distinct, these methods are complementary and can be combined for ML model assessment. It is anticipated that increasing efforts will be expanded in chemoinformatics to develop methods for model explanation in order to further increase the acceptance of ML in the practice of medicinal chemistry.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A

SHAP theory

The principal goal of an explanation model g is to locally approximate and thereby simplify a complex model f that is difficult to understand. Additive feature attribution methods generate an explanation model via a linear function of binary variables, given by Eq. (2):

$$g\left(x'\right) = \phi_0 + \sum_{i=1}^{M} \phi_i x'_i \tag{2}$$

where $x' \in \{0,1\}^M$, M is the number of input features, and $\phi_i \in \mathbb{R}$. The presence or absence of a feature value impacting the prediction represents a feature contribution (ϕ_i). Accordingly, a weight must be assigned to each variable for which the LIME methodology [37] can be applied and further extended.

LIME generates the explanation ξ of an instance x according to Eq. (3):

$$\xi(x) = \underset{g \in G}{\operatorname{argmin}}\, \mathcal{L}\left(f, g, \pi_x\right) + \Omega(g) \tag{3}$$

where G is a class of interpretable (linear) models, $\mathcal{L}$ is the loss function to minimize, π_x the proximity measure between an instance z and x (kernel defining locality), and $\Omega(g)$ an optional regularization term to limit model complexity.

For the explanation of a given test instance x, the following procedure is applied:

(i) Artificial samples are obtained by permuting features of the test instance x.
(ii) These samples are weighted by the value of a kernel calculated for them and x.
(iii) A model g is trained to predict $f(x)$ with coefficients estimating feature importance.

Hence, LIME builds a linear model g in a feature region proximal to the test instance, with ML model f typically being non-linear. The LIME approach provides the basis for the development of the kernel SHAP methodology, as explained in the following.

Shapley values account for the distribution of feature contributions to a model's prediction for a given test instance. To determine the contribution of a feature i, all operations by which a feature might be added to the set ($N!$) and a summation over all possible sets (S) must be carried out. For any feature sequence, the marginal contribution by adding feature i is given by $[f(S \cup \{i\}) - f(S)]$. The resulting quantity is weighted by the number of combinations available to form the set prior to addition of feature I, i.e., ($|S|!$), and the order in which remaining features might be added, i.e., (($|N| - |S| - 1$)!). Hence, the importance of a given feature i is defined by Eq. (4):

$$\phi_i = \frac{1}{N!} \sum_{S \subseteq N \setminus \{i\}} |S|!(|N| - |S| - 1)![f(S \cup \{i\}) - f(S)] \tag{4}$$

It follows that Shapley values represent a unique way of dividing a model's output among feature contributions satisfying three axioms: *local accuracy* (or additivity), *consistency* (or symmetry), and *nonexistence* (or null effect).

Additive feature attribution methods typically do not consider two properties that are of high relevance for assessing feature importance, i.e., *local accuracy* and *consistency*. The SHAP formalism was specifically devised to take these axiomatic properties into account [36]. The property *local accuracy* ensures that the sum of individual feature attributions is equal to the original prediction because SHAP allocates the model prediction across contributing features. Furthermore, *consistency* ensures that feature importance correctly accounts for different models on a relative scale. Hence, if a change in a feature value has larger impact on model A than model B, feature importance should be larger in A. These properties can be accounted for by representing feature importance as Shapley values [36].

A weighting procedure for artificial samples is a key aspect for connecting Shapley values to the LIME approach. In LIME, heuristic choices are made to select $\mathcal{L}$, $\Omega(g)$, and π_x. By contrast, the SHAP method introduces a special kernel function that is related to the Shapley value definition, assuming that feature weights follow the two axioms of interpretability. Specifically, SHAP uses the following procedure for interpreting an instance x:

(i) Training data are organized by k-means clustering and the k samples are weighted by the number of training instances they represent. These samples constitute a background data set of given feature values.
(ii) Artificial samples are obtained by replacing features of the test instance x with the values from the background data set.
(iii) These artificial samples are weighted by the value of the SHAP kernel calculated for each sample and x.

(iv) A weighted linear regression model g is trained to predict $f(x)$. The model coefficients are Shapley values corresponding to feature importance estimates.

Sampling all possible feature subsets is avoided through permutation of the feature vector by setting features on and off. A feature is assigned a large weight if its replacement with an artificial value leads to a significant change in model output. Weights of artificial samples are determined as the number of feature addition sequences of a given subset by the SHAP kernel. Coefficients from local linear regression provide feature weights as Shapley values, which indicate how important a feature is for a given prediction including the direction (sign) of feature influence. The expected explanatory value is calculated as the mean of the model output probability (numerical value) over training set instances. For a given instance, the model output is then calculated as the sum of the expected (base) value and all SHAP feature values.

References

[1] Struble TJ, Alvarez JC, Brown SP, Chytil M, Cisar J, DesJarlais RL, Engkvist O, Frank SA, Greve DR, Griffin DJ, Hou X, Johannes JW, Kreatsoulas C, Lahue B, Mathea M, Mogk G, Nicolaou CA, Palmer AD, Price DJ, Robinson RI, Salentin S, Xing L, Jaakkola T, Green WH, Barzilay R, Coley CW, Jensen KF. Current and future roles of artificial intelligence in medicinal chemistry synthesis. J Med Chem 2020;63:8667–82.

[2] Yang X, Wang Y, Byrne R, Schneider G, Yang S. Concepts of artificial intelligence for computer-assisted drug discovery. Chem Rev 2019;119:10520–94.

[3] Walters WP, Barzilay R. applications of deep learning in molecule generation and molecular property prediction. Acc Chem Res 2020;54:263–70 2020.

[4] Mater AC, Michelle LC. Deep learning in chemistry. J Chem Inf Model 2019;59:2545–59.

[5] Pannu A. artificial intelligence and its application in different areas. Artif Intell 2015;4:79–84.

[6] Bajorath J. State-of-the-art of artificial intelligence in medicinal chemistry. Future Sci OA 2021;7:FSO702.

[7] Varnek A, Baskin I. Machine learning methods for property prediction in chemoinformatics: quo vadis? J Chem Inf Model 2012;52:1413–37.

[8] Vapnik VN. The nature of statistical learning theory. 2nd editor. New York: Springer; 2000.

[9] Breiman L. Random forests. Mach Learn 2001;45:5–32.

[10] Alpaydin E. Introduction to machine learning. 2nd ed. Cambridge, USA: MIT Press; 2010.

[11] Baskin I, Winkler D, Tetko IV. A renaissance of neural networks in drug discovery. Expert Opin Drug Discov 2016;11:785–95.

[12] Rudin C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nat Mach Intell 2019;1:206–15.

[13] Bajorath J. Foundations of data-driven medicinal chemistry. Future Sci OA 2018;4:FSO320.

[14] Shen D, Wu G, Suk HI. Deep learning in medical image analysis. Ann Rev Biomed Eng 2017;19:221–48.

[15] Deng L, Liu Y. Deep learning in natural language processing (editors). New York: Springer; 2018.

[16] Stokes JM, Yang K, Swanson K, Jin W, Cubillos-Ruiz A, Donghia NM, McNair CR, French S, Carfrae LA, Bloom-Ackermann Z, Tran VM, Chiappino-Pepe A, Badran AH, Andrews IW, Chory EJ, Church GM, Brown ED, Jaakkola TS, Barzilay R, Collins JJ. A deep learning approach to antibiotic discovery. Cell 2020;180:688–702.

[17] Walters WP, Murcko M. Assessing the impact of generative AI on medicinal chemistry. Nat Biotechnol 2020;38:143–5.

[18] Bajorath J, Kearnes S, Walters WP, Meanwell NA, Georg GI, Wang S. Artificial intelligence in drug discovery: into the great wide open. J Med Chem 2020;63:8651–2.

[19] Castelvecchi D. Can we open the black box of AI? Nature 2016;538:20–3.

[20] Hirschfeld L, Swanson K, Yang K, Barzilay R, Coley CW. Uncertainty quantification using neural networks for molecular property prediction. J Chem Inf Model 2020;60:3770–80.

[21] Soleimany AP, Amini A, Goldman S, Rus D, Bhatia SN, Coley CW. Evidential deep learning for guided molecular property prediction and discovery. ACS Cent Sci 2021;7:1356–67.

[22] Molnar C, Casalicchio G, Bischl B. Interpretable machine learning – a brief history, state-of-the-art and challenges. In: Proceedings of the Joint European Conference on machine learning and knowledge discovery in databases. Springer; 2020. p. 417–31.

[23] Rudin, C.; Chen, C.; Chen, Z.; Huang, H.; Semenova, L.; Zhong, C. Interpretable machine learning: fundamental principles and 10 grand challenges. arXiv preprint arXiv:2103.11251, 2021.

[24] Polishchuk P. Interpretation of quantitative structure-activity relationship models: past, present, and future. J Chem Inf Model 2017;57:2618–39.

[25] Hansen K, Baehrens D, Schroeter T, Rupp M, Müller KR. Visual interpretation of kernel-based prediction models. Mol Inf 2011;30:817–26.

[26] Balfer J, Bajorath J. Visualization and interpretation of support vector machine activity predictions. J Chem Inf Model 2015;55:1136–47.

[27] Balfer J, Bajorath J. Introduction of a methodology for visualization and graphical interpretation of Bayesian classification models. J Chem Inf Model 2014;54:2451–68.

[28] Ghorbani A, Abid A, Zou J. Interpretation of neural networks is fragile. In: Proceedings of the AAAI Conference on artificial intelligence, 33; 2019. p. 3681–8.

[29] Johansson U, Sönströd C, Norinder U, Boström H. Trade-off between accuracy and interpretability for predictive *in silico* modeling. Future Med Chem 2011;3:647–63.

[30] Iooss B, Saltelli A. Introduction to sensitivity analysis. In: Ghanem R, Higdon D, Owhadi H, editors. Handbook of uncertainty quantification. Cham: Springer; 2016. p. 1–20.

[31] Baskin II, Ait AO, Halberstam NM, Palyulin VA, Zefirov NS. An approach to the interpretation of backpropagation neural network models in QSAR studies. SAR QSAR Environ Res 2002;13:35–41.

[32] Marcou G, Horvath D, Solov'ev V, Arrault A, Vayer P, Varnet A. Interpretability of SAR/QSAR models of any complexity by atomic contributions. Mol Inf 2012;31:639–42.

[33] Shapley LS. A value for N-person games. Contributions to the theory of games. Kuhn HW, Tucker AW, editors, Princeton: Princeton University Press; 1953. Annals of Mathematical Studies, pp. 307-317.

[34] Young HP. Monotonic solutions of cooperative games. Int J Game Theory 1985;14:65–72.

[35] Osborne MJ, Rubinstein A. A course in game theory. Cambridge: MIT Press; 1994.

[36] Lundberg SM, Lee S. A unified approach to interpreting model predictions. Adv Neural Inf Process Syst (NIPS) 2017;30:4766–75.

[37] Ribeiro MT, Singh S, Guestrin C. "Why should I trust you?" Explaining the predictions of any classifier. In: Proceedings of the 22nd ACM SIGKDD International Conference on knowledge discovery and data mining; 2016. p. 1135–44.

[38] Rodríguez-Pérez R, Bajorath J. Interpretation of compound activity predictions from complex machine learning models using local approximations and Shapley values. J Med Chem 2020;63:8761–77.

[39] Rodríguez-Pérez R, Bajorath J. Interpretation of machine learning models using shapley values: application to compound potency and multi-target activity predictions. J Comput Aided Mol Des 2020;34:1013–26.

[40] Lundberg SM, Erion G, Chen H, DeGrave A, Prutkin JM, Nair B, Katz R, Himmelfarb J, Bansal N, Lee S. From local explanations to global understanding with explainable AI for trees. Nat Mach Intell 2020;2:56–67.

[41] Riniker S, Landrum GA. Similarity maps – a visualization strategy for molecular fingerprints and machine-learning methods. J Cheminf 2013;5:43.

[42] Rodríguez-Pérez R, Bajorath J. Feature importance correlation from machine learning indicates functional relationships between proteins and similar compound binding characteristics. Sci Rep 2021;11:14245.

[43] Zwillinger D, Kokoska S. Standard probability and statistics tables and formulae. New York: CRC Chapman & Hall; 2000.