RESEARCH ARTICLE

Interpretable machine learning methods for in vitro pharmaceutical formulation development

Zhuyifan Ye1 | Wenmian Yang2 | Yilong Yang3 | Defang Ouyang1

1 State Key Laboratory of Quality Research in Chinese Medicine, Institute of Chinese Medical Sciences (ICMS), University of Macau, Macau, China
2 State Key Laboratory of Internet of Things for Smart City, University of Macau, Macau, China
3 School of Software, Beihang University, Beijing, China

Correspondence
Defang Ouyang, State Key Laboratory of Quality Research in Chinese Medicine, Institute of Chinese Medical Sciences, University of Macau, Macau 999078, China.
Email: defangouyang@um.edu.mo

Funding information
Macau Science and Technology Development Fund (FDCT), Grant/Award Number: 0029/2018/A1

Abstract

Background: Machine learning has become an alternative approach for pharmaceutical formulation development. However, many machine learning applications in pharmaceutics only focus on model performance rather than model interpretability.

Aim: This study aims to propose an attention-based deep neural network (DNN) for pharmaceutical formulation development.

Methods: An attention-based DNN, AttPharm, was proposed. AttPharm separately handled feature values and feature physical meaning by representation learning to successfully apply the attention mechanism to the pharmaceutical tabular data. Furthermore, the distributions of the attention weights were computed using AttPharm. Two post hoc methods, local interpretable model-agnostic explanation (LIME) and TreeSHAP, were utilized to obtain the post hoc model interpretability for lightGBM.

Results: The results demonstrated that AttPharm significantly improved the model performance of plain neural networks on a pharmaceutical cyclodextrin dataset because the attention mechanism could extract related features and find minute variation. Notably, the attention weights were analyzed, which illustrated global and local feature-level and sample-level model interpretability, thus providing insights for formulation design. Comparing with post hoc methods, AttPharm can be used without the concern of the faithfulness of interpretability.

Conclusion: This is the first step in applying the attention-based DNN to pharmaceutical formulation development. Considering the importance of model interpretability, the proposed approach may have a wide range of applications in pharmaceutics.

KEYWORDS
deep neural networks, formulation design, interpretable machine learning methods, the attention mechanism

1 | INTRODUCTION

Increasingly, machine learning algorithms are being applied to food and pharmaceutical research, such as fruit extracts, natural sweeteners, sustained-release matrix tablets, and oral fast disintegrating tablets (Abhimanyu et al., 2020; Astray et al., 2020; Bhagya Raj & Dash, 2020; Deng et al., 2021; Goel et al., 2021; Han et al., 2019; Yang et al., 2019; Ye et al., 2019; Zhao et al., 2019). The use of state-of-the-art
machine learning algorithms in pharmaceutics and food research has improved the efficiency of drug and food development. However, the current applications of machine learning in pharmaceutics and food mainly focus on model performance. In addition to accurate model prediction, machine learning interpretability is increasingly critical and has gained widespread attention (Bibal & Frénay, 2016). Different dosage forms, foods, and formulations depend on different molecular mechanisms. Although researchers have devoted significant resources, many of them remain unknown. Therefore, there is still room for improvement in machine learning predictions in drug and food research to gain insight into the mechanisms and accelerate their development.

Model interpretability has been increasingly recognized as an integral part of machine learning (Gentiluomo et al., 2019; Sheridan, 2019). Model interpretability helps locate the essential substructures and active sites of compounds. Furthermore, model interpretability helps discover key excipients and processes in medicinal and food formulation development. From 2000 to 2020, some attempts of interpretable methods were also applied to biomedical studies. They could be classified into several methods according to different algorithms, including linear models (Aben et al., 2016; Bergström et al., 2007; Davis & Vasanthi, 2015; García-Sosa et al., 2012; Gavaghan et al., 2007; Helguera et al., 2013; Keles et al., 2004; Lapinsh et al., 2005; Lee et al., 2019; Loryan et al., 2015; Mahlin et al., 2011; Marchese Robinson et al., 2017; Saiz-Urra et al., 2011; Santos-Filho & Cherkasov, 2008; Tabei et al., 2012; Wan et al., 2007; Yoshida & Niwa, 2006; Zhou et al., 2013), tree-based models (Brace et al., 2007; Clare et al., 2006; Freitas et al., 2015; Huang et al., 2007; Krishnan & Westhead, 2003; Kuz’min et al., 2011; Marchese Robinson et al., 2017; Nembri et al., 2016; Newby et al., 2013; Welling et al., 2015; Yuan et al., 2012; Zhang et al., 2016), naïve Bayesian classifiers (Briesemester et al., 2010a, 2010b; Hu et al., 2010; Sun, 2005; Sutherland et al., 2008), support vector machine (SVM) using linear kernel function (Rosenbaum et al., 2011; Tabei et al., 2012), rule-based methods (such as association rule mining and empirical rules; Clare et al., 2006; Gleeson, 2008; Langham & Jain, 2008; Leung et al., 2010; Levatić et al., 2013; Lovering et al., 2009), feature selection algorithms (such as genetic algorithms; Cavill et al., 2009; Danishuddin & Khan, 2016; Ganguly et al., 2006; Huynh-Thu et al., 2012; Saiz-Urra et al., 2011; Santos-Filho & Cherkasov, 2008), neural network techniques (such as the attention mechanism, gradient back-propagation based methods, and visualization of learned features in the middle layer; Bogard et al., 2019; Gentiluomo et al., 2019; Karimi et al., 2019; Li et al., 2019; Liu et al., 2018; Xiong et al., 2019). They can also be classified into premodeling methods, intrinsically interpretable models, and post hoc methods, according to the chronological order in which the interpretations are obtained.

Deep neural networks (DNNs), widely known as deep learning, have been applied to many real-world services and have been shown to outperform humans in many areas of government, academia, and industry. The key aspect of deep learning is the general automatic feature extractor (Abrol et al., 2021; Lecun et al., 2015). Attention-based DNNs can distinguish inputs, leading to more outstanding performance than standard DNNs (Luong et al., 2015; Rush et al., 2015; Vaswani et al., 2017; Wu et al., 2016). Due to the generated attention weights, attention-based DNNs are considered to be a kind of intrinsically interpretable model (Xu et al., 2015). The attention mechanism assigns an attention weight distribution to each sample to extract the relevant information. The mean value of the attention weights for all samples can be considered as an estimate of the global feature importance. The attention weight distribution of a single sample can be considered as the local feature importance (sample level). The attention weights of a feature for all samples can be considered as the feature importance (feature level) that varies with the value of the feature. Recently, several studies have applied attention-based DNNs to biomedical research. They used high-dimensional sequence and graph data as the network input to predict molecular activity, molecular stability, and compound-protein affinity (Karimi et al., 2019; Li et al., 2019; Liu et al., 2018; Xiong et al., 2019). However, the application of attention-based DNNs to two-dimensional tabular data is an unresolved difficulty. In pharmaceutical and food fields, structured tabular data are the most widely used data type rather than sequence, graph, or image high-dimensional data. To our knowledge, no studies apply attention-based DNNs to formulation development.

In industrial applications and data competitions, the gradient boosting decision tree (GBDT) framework is competitive among machine learning models, especially on structured datasets (Friedman, 2001). For example, lightGBM is an ensemble learning algorithm based on GBDT (Ke et al., 2017). lightGBM is further optimized using exclusive feature bundling and histogram-based learning, which has the advantages of good performance and high efficiency. However, GBDT framework models do not provide estimates of local feature importance (sample level and feature level). Additional post hoc interpretable methods can be used to provide local interpretability for the trained models (Gilpin et al., 2019; Johansson et al., 2011). Local interpretable model-agnostic explanation (LIME) and Shapley additive explanation (SHAP) are two approaches (Lundberg & Lee, 2017; Ribeiro et al., 2016). LIME uses a linear model as a surrogate model. LIME obtains the predictions of the original model by slightly changing the inputs and uses the surrogate model to fit the original model locally and linearly. However, there are concerns about the faithfulness of the surrogate model interpretation (Gilpin et al., 2019). In addition, SHAP estimates the difference in model performance when each feature is eliminated. TreeSHAP is an efficient implementation of SHAP proposed for the GBDT models, which truncates the nodes containing the features that need to be eliminated (Lundberg et al., 2018, 2019). TreeSHAP uses an elimination-based approach to estimate feature importance. However, accurate estimation requires the assumption that the input features are independent of each other, which is a strong assumption in practice (Lundberg et al., 2019). For these reasons, we believe that Attention-based DNNs can be an alternative approach.

In this study, an attention-based DNN, AttPharm, was proposed. AttPharm handled separately the feature values and the feature physical meaning by representation learning to apply the attention mechanism to the structured tabular data. AttPharm was available in
three versions, including AttPharm, AttPharm-formula, and AttPharm-formula-ResNet. By integrating the attention mechanism, the normalization formula, and residual blocks, AttPharm-formula-ResNet gave accurate predictions with interpretable attention weight distributions. The proposed methods were also compared with other commonly used algorithms, including SVM, plain neural network, and lightGBM. Furthermore, two post hoc interpretable methods, TreeSHAP and LIME, were applied to trained lightGBM to get post-hoc interpretability for comparison purposes. The pros and cons of numerous interpretable methods were discussed for their further applications in pharmaceutics and food research.

## METHODS

### 2.1 Dataset and data splitting strategy

Recently, the pharmaceutical formulation data for the binding-free energy of inclusion complexes between 8 types of cyclodextrins (CD) and 1320 diverse guest molecules were extracted from research literature published from 1990 to 2018 by the author’s lab (Zhao et al., 2019). The literature was recognized to belong to the binary CD complexation system. The original dataset was further checked and later used as a model dataset in this study. Five formulations were deleted due to incorrect descriptors, and three formulations were deleted due to unconvertible SMILES. The predicted logS and logP of guest molecules were obtained from the ALOGPS 2.1 program (http://www.vcclab.org) and added to the dataset (Tetko et al., 2005). Finally, a tabular dataset including 2992 data and 42 features was generated for machine learning modeling. The features were composed of 19 molecular descriptors of small guest molecules, 21 molecular descriptors of CD, 2 process parameters of pH, and temperatures. The range of the predicted target, the binding-free energy, was from $-40$ to $0$ kJ/mol.

A three dataset splitting strategy was carried out. This strategy has been widely applied in machine learning fields. The dataset was split into three subsets of training (2318 data), validation (339 data), and test (335 data) subsets. Stratified random sampling was used to keep the balanced proportion of each CD type (Zhao et al., 2019). The training subset was used for constructing models, the validation subset was for tuning the hyperparameters, and the test subset was for testing model generalizability on unseen data.

### 2.2 Evaluation criteria for machine learning models

Mean absolute error (MAE), mean squared error (MSE), and determination coefficient ($R^2$) have been commonly used as the metrics for evaluating model performance in regression problems. MAE shows the MAE between real values and predicted values, MSE measures the MSE between real values and predicted values, and $R^2$ indicates the correlation between real values and predicted values. They are defined as

$$\text{MAE} = \frac{1}{n} \sum_{i=1}^{n} |\hat{y}_i - y_i|$$  \hspace{1cm} (1)

$$\text{MSE} = \frac{1}{n} \sum_{i=1}^{n} (\hat{y}_i - y_i)^2$$  \hspace{1cm} (2)

$$R^2 = 1 - \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2}$$  \hspace{1cm} (3)

where $n$ is the number of formulations, $y_i$ is the ith real value, and $\hat{y}_i$ is the ith predicted value.

### 2.3 Hyperparameters of SVM, plain neural network, and lightGBM

Recent development and applications of machine learning have been driven both by data availability and advances in learning algorithms. A variety of machine learning algorithms have been proposed to approximate different mapping types (Abhimanyu et al., 2020; Bishnupriya et al., 2019). Traditional machine learning processes rely on feature engineering, the process of extracting features using domain knowledge. DNNs are regarded as universal feature extractors that automatically extract critical features from raw data through multi-layer architectures and gradient back-propagation (Abrol et al., 2021). In this study, SVM, a plain neural network, and lightGBM were introduced to predict the binding-free energy of CD inclusion complexes. The plain neural network was used as the baseline. Specifically, SVM was constructed by Scikit-learn version 0.21.2 in Python (Pedregosa et al., 2011). The plain neural network was developed by Keras version 2.2.4 with the backhand of TensorFlow version 1.11 in Python (Abadi et al., 2016). lightGBM was established by the lightGBM package version 2.2.3 in Python (Ke et al., 2017).

The hyperparameters of machine learning models have impacts on model performance. The diversity of machine learning algorithms determines different hyperparameter search methods, such as random search and grid search. Specifically, a random search was used for SVM and lightGBM because their hyperparameters were continuous, and random search methods searched hyperparameter spaces more efficiently than grid search. We adjusted the hyperparameters of the plain neural network and the AttPharm models manually. Due to DNNs’ large hyperparameter search space, DNNs trained using random search may suffer from underfitting or overfitting. The hyperparameter configurations of SVM (the kernel function, the penalty parameter C, and the $\gamma$), the plain neural network (the number of hidden layers, the number of neurons in each of the hidden layers, the learning rate, the optimization algorithm, the number of epochs, the batch size, and the coefficient lambda of the $l_2$ regularization), lightGBM (the learning rate, the number of trees, the subsample ratio, the subsample ratio of columns, the regularization terms of the number of leaves, and the minimum number of samples in a child leaf) are listed in Table 1.
TABLE 1  Model hyperparameter configurations of machine learning models

| Machine learning algorithms       | Hyperparameter configurations                        |
|----------------------------------|-----------------------------------------------------|
| SVM                              | Radial basis function; 70; 0.025                      |
| Plain neural network             | 5; 512, 128, 64, 64; 0.002; Adam with $\beta_1$, $\beta_2$ of 0.9, 0.999; 500; 500; 0.005; 0.2 |
| lightGBM                         | 0.03; 700; 0.6; 0.6; 56; 20                           |
| AttPharm                         | 8; 4; 128, 32, 32, 8; 5; 128, 32, 32, 32; 0.006; Adam with $\beta_1$, $\beta_2$ of 0.9, 0.999; 220; 500; 0.001 |
| AttPharm-formula                 | 16; 4; 512, 128, 64, 16; 9; 512, 128, 64, 64, 64, 64, 64, 64, 64; 0.003; Adam with $\beta_1$, $\beta_2$ of 0.9, 0.999; 300; 500; 0 |
| AttPharm-formula-ResNet          | 16; 3; 256, 64, 16; 5; 512, 128, 32, 32, 32; 0.004; Adam with $\beta_1$, $\beta_2$ of 0.9, 0.999; 219; 500; 0 |

Note: ";" separates different hyperparameters; ";" means the hyperparameter is composed of more than one element.
Abbreviation: SVM, support vector machine.

2.4  AttPharm model

2.4.1  A normalization formula

In machine learning, z-score normalization (also known as standardization) and min–max normalization are usually adopted as feature scaling methods for data processing. Z-score normalization subtracts the mean from each feature and then divides the result by its standard deviation. Min–max normalization scale the range of features in [0, 1]. However, the z-score normalization and the min–max normalization cannot properly represent the physicochemical properties of features. In pharmaceutics and food, some features have opposite physicochemical properties when taken to different values. For example, pH less than 7 means acidic and more than 7 means basic. Thus, specific normalization suitable for pharmaceutics and food should be introduced to scale the range of features.

Following the min–max normalization target range using the range in [0, 1] to rescale the range of features, the range [−1, 1] is used to scale the range of features for representing the physicochemical properties. For example, the normalized pH values of acid and base are less than and more than 0. The normalization formula (Equation 4) is first subtracting the shift from each feature and then dividing the values of each feature by the difference between the shift and maximum values:

$$x' = \frac{x - \text{shift}}{\max|x| - \text{shift}}$$  (4)

where $x$ is an original value, $x'$ is the normalized value, and shift is the shift constant. For AttPharm, the usually used standardization was adopted. For AttPharm-formula and AttPharm-formula-ResNet, the normalization formula was adopted.

2.4.2  The attention mechanism

A new network was developed for applying the attention mechanism to handling the tabular data. Figure 1 shows the improved attention-based DNN structure with the feature positional encoding and residual blocks (AttPharm). The AttPharm model contains five steps. In step 1, the features’ absolute positions go through an embedding layer to generate a position matrix (Vaswani et al., 2017). The physical meaning of features is considered to be implicit in the positions, the physical meaning is projected into a vector space, and the physical meaning and values of features are separately processed. In step 2, the feed-forward residual blocks take in the active pharmaceutical ingredient (API) features, generate a drug vector, and use the drug vector as the attention mechanism query vector. In step 3, the dot-product attention mechanism is adopted (Bahdanau et al., 2014). The drug vector is used as the query vector to query each position vector in the position matrix to compute the similarity values, the similarity values are normalized by softmax to obtain the final attention weights. In step 4, each position vector is multiplied by the corresponding feature value to compute the input feature vectors. In step 5, the weighted sum of the attention weights and the input feature vectors is computed as a formulation vector and used as the input to the output layers with residual blocks. All network parameters were end-to-end optimized on the data by the back-propagation gradient descent optimization algorithm automatically.

Different from the original attention mechanism, often combined the top of the attention mechanism with convolutional neural network and recurrent neural network (Karimi et al., 2019; Vaswani et al., 2017; Wu et al., 2018), tabular data were neither sequence data nor image data, a feed-forward network was a good choice. The new attention-based DNN model not only includes representation learning and positional encoding to handle the tabular data but also contains residual blocks to improve the model performance. The attention
weight is
\[ \alpha_i = \frac{\exp e_i}{\sum_{n=1}^{N} \exp e_n} \]  
(5)
\[ e_i = p_i \cdot d, \]  
(6)
where \( \alpha_i \) is an attention weight, \( e_i \) is a similarity, \( N \) is the number of features, \( p_i \) is a position vector, and \( d \) is a drug vector.

The formulation vector is
\[ \text{formulation} = \sum_{i=1}^{N} \alpha_i i_i, \]  
(7)
where \( \text{formulation} \) is a formulation vector, \( N \) is the number of features, \( \alpha_i \) is an attention weight, and \( i_i \) is an input feature vector.

Finally, the cost function (C) and the objective function (J) are
\[ C(\hat{y}, y) = \frac{1}{n} \sum_{i=1}^{n} (\hat{y}_i - y_i)^2 \]  
(8)
\[ J(w) = C(\hat{y}, y) + \lambda \|w\|^2, \]  
(9)
where \( \hat{y} \) are predicted values, \( y \) are real values, \( n \) is the number of data, \( w \) are network weights, and \( \lambda \) is the coefficient of l2 regularization.

### 2.4.3 Residual blocks

The residual blocks were applied to boosting the AttPharm model generalization. The residual blocks solve the network degradation problem, which occurs when the depth of the networks increased (He et al., 2015). The residual blocks bridge shortcut connections to skip one or more layers, thus the network is easy to learn identity mappings, go deeper, and perform better than DNNs without residual blocks. In this study, the residual blocks were used, and if the input dimension does not match the output dimension, a linear projection is performed.

### 2.5 Hyperparameters of AttPharm

Three versions of AttPharm models were tested, namely AttPharm (base), AttPharm-formula, and AttPharm-formula-ResNet. They were established by Keras version 2.2.4 with the backhand of TensorFlow version 1.11 in Python (Abadi et al., 2016).

For AttPharm and AttPharm-formula, the hyperparameters of (the number of dimensions of the embedding layer, the number of hidden layers in the fully connected neural network (FCNN), the number of neurons in each of the hidden layers in the FCNN, the number of hidden layers in the output layers, the number of neurons in each of the output layers, the learning rate, the optimization algorithm, the number of epochs, the batch size, and the learning rate decay) are listed in Table 1.

For AttPharm-formula-ResNet, the hyperparameters (the number of dimensions of the embedding layer, the number of residual blocks, the shape of each of the residual blocks, the number of residual blocks followed by a dense layer of 16 units in the output layers, the shape of each of the residual blocks in the output layers, the learning rate, the optimization algorithm, the number of epochs, the batch size, and the learning rate decay) are listed in Table 1.

### 3 RESULTS AND DISCUSSION

#### 3.1 Feature selection

lightGBM feature importance ranking was used as the feature selection method due to its more robust results than stability selection LASSO and recursive feature elimination strategies. Six datasets were constructed, containing the most important 43, 40, 35, 30, 25, 20, and 15 features, respectively. The MAE results of lightGBM and plain neural network were obtained (Figure 2), which showed that the MAE values declined as the number of the features decreased from 43 to 25 and went up from 15 to 10. Therefore, the 25 features were utilized for the following modeling process. The 25 features included 19 descriptors of small molecules, 4 descriptors of CDs, and 2 process parameters. For example, small molecules were represented by the descriptors of molecular weight, \( \log_S \), \( \log_P \), \( X\logP3 \) (a calculation of \( \log_P \) by machine
Table 2 The model performance on the training, validation and test subsets

| Algorithms            | Training set | Validation set | Test set |
|-----------------------|--------------|----------------|----------|
|                       | MAE  | MSE  | \(R^2\) | MAE  | MSE  | \(R^2\) | MAE  | MSE  | \(R^2\) |
| lightGBM              | 1.29 | 3.28 | 0.88    | 1.23 | 2.61 | 0.87    | 1.34 | 3.31 | 0.86 |
| AttPharm-formula-ResNet (ours, final) | 1.50 | 4.55 | 0.84    | 1.60 | 4.27 | 0.79    | 1.58 | 4.58 | 0.81 |
| AttPharm-formula (ours) | 1.47 | 4.42 | 0.84    | 1.73 | 5.47 | 0.74    | 1.70 | 6.21 | 0.74 |
| AttPharm (ours)       | 1.66 | 5.47 | 0.80    | 1.79 | 5.62 | 0.73    | 1.88 | 6.25 | 0.73 |
| Plain neural network (baseline) | 1.51 | 4.48 | 0.84    | 1.74 | 5.18 | 0.75    | 1.90 | 6.88 | 0.71 |
| SVM                   | 1.89 | 9.16 | 0.67    | 1.72 | 5.44 | 0.74    | 1.99 | 7.73 | 0.67 |

Abbreviations: MAE, mean absolute error; MSE, mean squared error; SVM, support vector machine.

Learning model), complexity, rotatable bond count, heavy atom count, topological polar surface area, hydrogen bond acceptor count, hydrogen bond donor count, ring count, aromatic ring count, molar refractivity, polarizability, Van der Waals volume, Van der Waals surface area, solvent accessible surface area, minimum projection radius, and maximum projection radius. And the CDs were represented by the descriptors of molecular weight, logP, minimum projection radius, and maximum projection radius. The process parameters included pH and temperatures.

### Performance Comparison of Different Machine Learning Models

The MAE, MSE, and \(R^2\) results of the machine learning models were obtained on the training, validation, and test subsets (Table 2). The SVM and the plain neural network were used as the baseline models. On the final test subset, the SVM and the plain neural network got similar MAE and the \(R^2\) results. In addition, AttPharm demonstrated a better result than that of SVM and the plain neural network. AttPharm-formula further enhanced the performance of AttPharm. Furthermore, AttPharm-formula-ResNet and lightGBM models had comparable and better results. AttPharm-formula-ResNet significantly improved the model performance of the plain neural network from the MAE of 1.90–1.58 kJ/mol, from the MSE of 6.88–4.58, and from the \(R^2\) of 0.71–0.81. The results in Table 2 indicate that the AttPharm-formula-ResNet and lightGBM had a comparable and superior prediction of the \(R^2\) over 0.81 and the MAE under 1.59 kJ/mol, which met the formulation development requirements.

lightGBM is a tree-based ensemble learning algorithm (Ke et al., 2017). Tree-based ensemble models combining tree structure base learners usually demonstrate good performance on tabular data with pre-engineered features (Abrol et al., 2021; Friedman, 2001). However, traditional models are unable to learn new representations from raw data and rely on informative low-dimensional features. On the other hand, the models need post hoc interpretable approaches, such as LIME and TreeSHAP, to provide local interpretability (feature level and sample level). The post hoc interpretable methods use a simple linear model as the surrogate model. The surrogate model locally approximates the complex nonlinear model and then reports the interpretability of the surrogate model. There are concerns about the faithfulness of the interpretation using surrogate model methods (Gilpin et al., 2019). SHAP’s accurate estimation requires the assumption that the input features are independent of each other, which is a strong assumption in practice (Lundberg et al., 2019).

Deep learning is able to both learn task-specific representations from raw data and determine decision boundaries (Abrol et al., 2021; Lecun et al., 2015). Powerful feature extraction capabilities contribute to the superior performance of deep learning models. On structured tabular data, despite depriving deep learning of its key aspect of representation learning, AttPharm-formula-ResNet also has outstanding performance similar to the cutting-edge tree-based ensemble model (\(R^2 > 0.8\)). The method has the potential to scale well because the neural networks may benefit from both learning new representational patterns and the increasing data. Furthermore, the attention mechanism based deep learning models allow for reliable correlation estimates. Attention-based DNNs do not depend on additional interpretable methods. The current limitation is the small domain data in the food and pharmaceutics fields. However, it is safe to expect that these samples will grow much larger in the coming years, due to the development of high-throughput screening and computational pharmaceutics. For these reasons, we believe that attention-based DNNs can be an alternative approach.

### Results of the Attention Weights

AttPharm-formula-ResNet predicted the CD binding-free energy, provided the attention weight distributions that could be recognized as global and local (feature level and sample level) interpretability. For CD inclusion complexes, although the impact factors such as CD types, pH, and temperatures were studied in experiments, it is still hard to investigate the influence of substance innate features, such as molecular weight, logP, and logS. Although the global feature importance ranking was studied in the previous research by ensemble models (Zhao et al., 2019), feature-level and sample-level interpretability and a general machine learning architecture are still necessary.
First, the mean attention weights over the dataset are computed. The features containing the API and CD descriptors and process parameters were ranked (Figure 3) according to the overall mean attention weights. The mean attention weights globally estimated the importance of the features, which helped determine key factors that influenced CD inclusion complexes. Preparation into the CD inclusion complexes can improve the solubility and bioavailability of insoluble drugs, enhance drug stability, cover up the bad smell and reduce drug irritation and side effects (Carrier et al., 2007). The preparation process started after adding API into CD saturated solution (in given pH) at a proper temperature. The inclusion complexes formed when the API molecule entered the cavity of CD to form a noncovalent bond by stirring, milling, crystallizing, filtering, and drying (Zhao et al., 2018).

According to the feature list shown in Figure 3, the top six critical features included “XLogP3” (a prediction of logP by machine learning techniques) and “logP” of API, “ring count” of API, “temperature,” “complexity” of API, and “maximum projection radius” of CD. Among these features, the “XLogP3” and “logP” of API were well-recognizable impact factors on the binding-free energy of inclusion complexes. Generally, the binding-free energy decreased with the “XLogP3” and “logP” of API increasing. Apart from the “XLogP3” and “logP” of API, other substance innate features contained “ring count,” “complexity” of API, and “maximum projection radius” of CD, which were difficult to be studied in pharmaceutical experiments and exhibited a significant impact on the binding-free energy. Because the forming process of the inclusion complexes is a self-assembly process between host and guest molecules, “ring count” and “complexity” of API would increase the steric hindrance, leading to a high energy barrier against inclusion complex preparation. As CD molecules are cylindrical, the drugs with specific structures inserted into the hydrophobic cavity to form the inclusion complexes, “maximum projection radius” of CD can enhance the degree of the insertion, thus reducing the binding-free energy and improve the inclusion complex stability. Besides, “temperature” can affect the binding-free energy. Generally, in most of the CD inclusion complex forming processes, the enthalpy change ($\Delta\text{H}$) is negative.

Raising the temperature is not suitable for the inclusion complex forming. However, on the other hand, it is beneficial to increase the solubility of CD to obtain the inclusion complexes.

Feature-level and sample-level interpretability are two local model interpretabilities. The feature-level interpretability for the “XLogP3” of API was obtained by plotting the attention weights of the “XLogP3” of API versus the feature values of the “XLogP3” of API, which eventually illustrated the change of model attention with the drug lipophilicity increasing (Figure 4). The “XLogP3” of API was the most influenced feature ranked by the attention weights in the CD dataset. Here, when the “XLogP3” of API had a negative value, the model paid relatively high attention to this feature. This was possible because of the unstable binding of polar drugs caused by the hydrophobic interaction between host and guest molecules. As the cavity of CD was hydrophobic, when a guest molecule entered the cavity, its hydrophilic group was far away from the cavity, and a polar molecule was not stable in the nonpolar cavity. When the “XLogP3” of API had a positive value, the attention was relatively stable. The API concentrated between 0 and 5 “XLogP3” because typically insoluble and nonpolar drugs were preferred by researchers to get the improved CD inclusion complexes. This can also suggest that when the “XLogP3” of API is greater than 5, the confidence of the prediction is insufficient due to the sparse data distribution.

The sample-level interpretability for a specific example formulation was collected by representing an attention weight distribution for the example formulation, which locally evaluated the impact of the features on this example formulation (Figure 5). Similar to the global feature importance list shown above, the key points for achieving this CD inclusion complex included the “XLogP3” and “logP” of API, closely related to the lipophilic interaction between host and guest molecules in the system. The “complexity” of API also played an important role in forming the inclusion complex prepared by the example formulation. Additionally, the “maximum projection radius” and “minimum projection radius” of CD can influence the preparation of the CD inclusion complex by the different cavity radiiuses of CD because CD selectively holds the guest molecules with the molecular volume matching their cavity that leads to stable inclusion complexes. Apart from these features, the “logP” of CD reflects the lipophilicity of CD. This feature
3.4 Results of SHAP and LIME

The TreeSHAP method was applied to lightGBM to get the feature level post hoc interpretability. The SHAP values for the “XLogP3” of API are shown in Figure 6. A high SHAP value indicated that the feature of the instance had a high impact on the prediction made by lightGBM. The negative values of the “XLogP3” of API displayed the correlation with the relatively high SHAP values, which is roughly consistent with the above attention results.

The LIME method was applied to lightGBM to obtain the sample level post hoc interpretability. The LIME interpretation for the example formulation is illustrated in Figure 7. Using a linear model as the surrogate model, LIME interpreted the linear model’s behaviors as the underlying mechanisms of lightGBM. For this formulation, the “logP” of API, the “minimum projection radius” of CD, the “logS,” “hydrogen bond donor count,” and “aromatic ring count” of API jointly made lightGBM to give out the prediction of a relatively low binding-free energy. This also demonstrated a unique advantage of LIME that LIME can provide the direction of feature importance.

The global mean SHAP values over the dataset are computed, and the features containing both the API and CD innate features and experimental parameters were ranked (Figure 8). As expected, there were inconsistencies between AttPharm and TreeSHAP when comparing the top six features. For the first, second, and sixth features, the same conclusion was made. “logP” and “XLogP3” of APIs impacted the binding-free energy of the inclusion complexes. Generally, the binding-free energy decreased with the “XLogP3” and “logP” of APIs increasing. The same sixth feature was “maximum projection radius” of CD because CD molecules are cylindrical, the drugs with specific structures are inserted into the hydrophobic cavity to form the inclusion complexes. However, the third, fourth, and fifth features predicted by AttPharm were API properties and process, including “Ring count” of APIs, “Temperature” and “Complexity” of APIs. TreeSHAP focused on “logS” of APIs, “minimum projection radius,” and “Aromatic ring count” of APIs.

3.5 Pros and cons of current interpretable methods

Interpretable machine learning methods have some advantages for pharmaceutical formulation development: (a) First, because training data were collected from the real world, biases and subjective prejudice would exist, thus misleading the model. Model interpretability helped researchers identify and correct the biases in machine learning models. (b) Second, model interpretability improved model generalization. By leveraging experts’ knowledge and experience, model interpretability can speed up the analysis of incorrect predictions and unexpected behavior. (c) Third, model interpretability extracted the knowledge that was originally embedded in model structure and parameters. It also provided a new way for people to learn from big data from a knowledge acquisition perspective. (d) Fourth, model interpretability assisted in getting the trust of pharmaceutical and food experts. Reasonable interpretation increased user confidence in using the models in practice. In particular, model interpretability was valuable for high-risk tasks. (e) Fifth, model interpretability facilitates communication between machine learning engineers and pharmaceutical and food scientists. Interpretable methods translated functions into understandable figures and tables. These benefits fuel further applications of machine learning techniques in pharmaceutical and food research.

The discussion about model interpretability is still open. Some challenges existed and possibly limited the further development of the interpretable methods: (a) First, model interpretations may not faithfully reflect the underlying working mechanism of the model. (b) Second, general criteria to quantitatively evaluate model interpretability were lacking. Thus, it was difficult to compare the model interpretability of different models and interpretable methods. (c) Third, the
criterion to quantitatively estimate the interpretability faithfulness were also unavailable. These limitations of model interpretability may affect the user confidence in adopting interpretable methods.

Nowadays, various interpretable machine learning techniques were proposed (Aben et al., 2016; Bilal & Frénay, 2016; Briesemeister et al., 2010a, 2010b; Du et al., 2020; Gentiluomo et al., 2019; Gilpin et al., 2019; Guidotti et al., 2018; Karimi et al., 2019; Kuz’min et al., 2011; Li et al., 2019; Liu et al., 2018; Lundberg & Lee, 2017; Lundberg et al., 2018; Marchese Robinson et al., 2017; Ribeiro et al., 2016; Rosenbaum et al., 2011; Xu et al., 2015). The recent progress of the interpretable methods was summarized (Table 3). According to the time when model interpretability was obtained, these methods were divided into three types of premodeling methods, intrinsically interpretable models and post hoc methods. Premodeling methods started after data cleaning and before model training, mainly by dimensionality reduction and visualization methods. Intrinsically interpretable models happened when making predictions by directly illustrating the internal decision-making process of models. Post hoc methods split the complete modeling process into two parts of training and interpreting. On the basis of the application scope of model interpretability, model interpretability was divided into global interpretability and local interpretability. Global interpretability referred to a whole dataset and an entire model. Local interpretability referred to a sample and a feature.

Additionally, some post hoc interpretable methods were only designed for specific machine learning models, so-called model-specific methods. Model-agnostic methods were the post hoc methods that were available for any models. Considering the characteristics of specific tasks and the unique advantages of different interpretable methods, it was profitable to combined use multiple interpretable methods in a series of continuous modeling steps.

4 | CONCLUSION

In this study, AttPharm, a DNN architecture based on the attention mechanism, was proposed for formulation development. The results showed that AttPharm surpassed the plain neural network and achieved results comparable to lightGBM on the structured tabular dataset. Furthermore, the attention weight distributions
### TABLE 3  Recent progress of the interpretable machine learning methods

| Types                          | Methods                              | Implementations                                      | Global/local | Model specific/model agnostic |
|-------------------------------|--------------------------------------|-----------------------------------------------------|--------------|-------------------------------|
| Premodeling methods           | Dimensionality reduction and visualization | PCA                                                   | Global       | \                             |
|                               |                                      | t-SNE                                               | Global       | \                             |
|                               |                                      | VAE                                                 | Global       | \                             |
|                               | Cluster and representative sample analysis | k-means                                             | Local        | \                             |
|                               | Feature importance ranking           | Pearson’s correlation coefficient                   | Global       | \                             |
|                               |                                      | Features redundancy                                | Global       | \                             |
|                               |                                      | Information gain                                   | Global       | \                             |
|                               |                                      | Euclidean distance                                 | Global       | \                             |
| Intrinsically interpretable models | Transparent models                | Linear regression                                   | Global       | \                             |
|                               |                                      | Decision tree                                       | Global       | \                             |
|                               |                                      | Rule-based models                                   | Global       | \                             |
|                               |                                      | Naïve Bayesian classification                      | Global       | \                             |
|                               | Adding constraints for interpretability | Sparsity and monotonicity constraints               | Global       | \                             |
|                               |                                      | Pruning decision tree                               | Global       | \                             |
|                               |                                      | Regularization loss                                 | Global       | \                             |
|                               |                                      | Capsule networks                                    | Global       | \                             |
|                               |                                      | Attention mechanism                                 | Local        | \                             |
| Post hoc methods              | Feature importance ranking           | Permutation method                                   | Global       | Model agnostic                |
|                               |                                      | Perturbation method                                 | Local        | Model agnostic                |
|                               |                                      | Tree-based ensemble models                          | Global       | Model specific                |
|                               |                                      | Gradient-based method                               | Local        | Model specific                |
|                               |                                      | Linear regression                                   | Global       | Model agnostic                |
|                               |                                      | Decision tree                                       | Global       | Model agnostic                |
|                               |                                      | Rule-based models                                   | Global       | Model agnostic                |
|                               |                                      | LIME                                                | Local        | Model agnostic                |
|                               |                                      | SHAP                                                | Local        | Model agnostic                |
|                               |                                      | CNN layer representations                           | Global       | Model specific                |
|                               |                                      | RNN layer representations                           | Global       | Model specific                |
|                               |                                      | CNN representations of inputs                       | Local        | Model specific                |
| Abbreviations: CNN, convolutional neural network; LIME, local interpretable model-agnostic explanation; PCA, principal component analysis; RNN, recurrent neural network; SHAP, Shapley additive explanation; t-SNE, t-distributed stochastic neighbor embedding; VAE, variational autoencoder. |

Illustrated both global and local (feature level and sample level) model interpretability. For the pharmaceutical dataset, the results indicated that the hydrophobic interaction between host and guest molecules, steric hindrance, experimental temperature, and hydrogen bond highly impact the forming of the CD inclusion complexes. This is the first step in applying attention-based DNNs to pharmaceutical formulation development. The proposed approach can be further used in other pharmaceutical and food tasks.

**ACKNOWLEDGMENT**

The current research was supported by the Macau Science and Technology Development Fund (FDCT) (grant no. 0029/2018/A1).

**CONFLICT OF INTEREST**

The authors declare that they have no competing interest.

**ORCID**

Zhuyifan Ye [https://orcid.org/0000-0002-9270-0949](https://orcid.org/0000-0002-9270-0949)

**REFERENCES**

Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., Devin, M., Ghemawat, S., Irving, G., Isard, M., Kudlur, M., Levenberg, J., Monga, R., Moore, S., Murray, D. G., Steiner, B., Tucker, P., Vasudevan, V., Warden, P., … Xiaoqiang Z. (2016). Tensorflow: A system for large-scale machine learning. 12th USENIX Symposium on Operating Systems Design and Implementation (OSDI’16), Savannah, GA, November 2–4.
Aben, N., Vis, D. J., Michaut, M., & Wessels, L. F. A. (2016). TANDEM: A two-stage approach to maximize interpretability of drug response models based on multiple molecular data types. Bioinformatics, 32(17), i413–i420.

Abhimanyu, T., Mishra, A. P., Panda, B., Rodriguez, D. C. S., Gaurav, I., & Majhi, B. (2020). Application of artificial intelligence in pharmaceutical and biomedical studies. Current Pharmaceutical Design, 26(29), 3569–3578.

Abhimanyu, T., Mishra, A. P., Panda, B., Sweta, K., & Majhi, B. (2020). Detection of disease-specific parent cells via distinct population of nanoparticles by machine learning. Current Pharmaceutical Design, 26(32), 3985–3996.

Abrol, A., Fu, Z., Salman, M., Silva, R., Du, Y., Plis, S., & Calhoun, V. (2021). Deep learning encodes robust discriminative neuroimaging representations to outperform standard machine learning. Nature Communications, 12(1), 353.

Astray, G., Albuquerque, B. R., Prieto, M. A., Simal-Gandara, J., Ferreira, I. C. F. R., & Barros, L. (2020). Stability assessment of extracts obtained from Arbutus unedo L. fruits in powder and solution systems using machine-learning methodologies. Food Chemistry, 333, 127460.

Bahdanau, D., Cho, K., & Bengio, Y. (2014). Neural machine translation by jointly aligning to align and translate. https://paperswithcode.com/paper/neural-machine-translation-by-jointly.

Bergström, C. A. S., Wassvik, C. M., Johansson, K., & Hubatsch, I. (2007). Poorly soluble marketed drugs display solvation limited solubility. Journal of Medicinal Chemistry, 50(23), 5858–5862.

Bhagya Raj, G. V. S., & Dash, K. K. (2020). Comprehensive study on applications of artificial neural network in food process modeling. Critical Reviews in Food Science and Nutrition, 1–28. 10.1080/10408398.2020.1858398

Bibal, A., & Frénay, B. (2016). Interpretability of machine learning models and representations: An introduction. ESANN 2016 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning, Bruges, Belgium, April 27–29, 2016. i6doc.com publication.

Bishnupriya, P., Babita, M., & Abhimanyu, T. (2019). An integrated-OFFT model for the prediction of protein secondary structure class. Current Computer-Aided Drug Design, 15(1), 45–54.

Bogard, N., Linder, J., Rosenberg, A. B., & Seelig, G. (2019). A deep neural network for predicting and engineering alternative polyadenylation. Cell, 178(1), 91–106.

Brace, C. L., Melville, J. L., Pickett, S. D., & Hirst, J. D. (2007). Contemporary QSAR classifiers compared. Journal of Chemical Information and Modeling, 47(1), 219–227.

Briesemeister, S., Rahnenführer, J., & Kohlbacher, O. (2010a). Going from where to why—Interpretable prediction of protein subcellular localization. Bioinformatics, 26(9), 1232–1238.

Briesemeister, S., Rahnenführer, J., & Kohlbacher, O. (2010b). YLoc—an interpretable web server for predicting subcellular localization. Nucleic Acids Research, 38(Suppl. 2), W497–W502.

Carrier, R. L., Miller, L. A., & Ahmed, I. (2007). The utility of cyclodextrins for enhancing oral bioavailability. Journal of Controlled Release, 123(2), 78–99.

Cavill, R., Keun, H. C., Holmes, E., Lindon, J. C., Nicholson, J. K., & Ebbels, T. M. D. (2009). Genetic algorithms for simultaneous variable and sample selection in metabolomics. Bioinformatics, 25(1), 112–118.

Clare, A., Karwath, A., Ougham, H., & King, R. D. (2006). Functional bioinformatics for Arabidopsis thaliana. Bioinformatics, 22(9), 1130–1136.

Danishuddin, & Khan, A. U. (2016). Descriptors and their selection methods in QSAR analysis: Paradigm for drug design. Drug Discovery Today, 21(8), 1291–1302.

Davis, G. D. J., & Vasanthi, A. H. R. (2015). QSAR based docking studies of marine algal anticancer compounds as inhibitors of protein kinase B (PKB). European Journal of Pharmaceutical Sciences, 76, 110–118.

Deng, X., Cao, S., & Horn, A. L. (2021). Emerging applications of machine learning in food safety. Annual Review of Food Science and Technology, 12(1), 513–538.

Du, M., Liu, N., & Hu, X. (2018). Techniques for interpretable machine learning. https://doi.org/10.1145/3359786.

Freitas, A. A., Limbu, K., & Ghaforian, T. (2015). Predicting volume of distribution with decision tree-based regression methods using predicted tissue/plasma partition coefficients. Journal of Cheminformatics, 7(1), 6.

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. Annals of Statistics, 29(5), 1189–1232.

Ganguly, M., Brown, N., Schuffenhauer, A., Ertl, P., Gillet, V. I., & Greenidge, P. A. (2006). Introducing the consensus modeling concept in genetic algorithms: Application to interpretable discriminant analysis. Journal of Chemical Information and Modeling, 46(5), 2110–2124.

García-Sosa, A. T., Oja, M., Hetényi, C., & Maran, U. (2012). DrugLogit: Logistic discrimination between drugs and nondrugs including disease specificity by assigning probabilities based on molecular properties. Journal of Chemical Information and Modeling, 52(8), 2165–2180.

Gavaghan, C. L., Arny, C. H., Blomberg, N., Strandlund, G., & Boeyr, S. (2007). Development, interpretation and temporal evaluation of a global QSAR of HERG electrophysiology screening data. Journal of Computer-Aided Molecular Design, 21(4), 189–206.

Gentiluomo, L., Roessner, D., Augustijn, D., Svilinov, H., Kulakova, A., Mahapatra, S., Winter, G., Streicher, W., Rinnan, Å., Peters, G. H. J., Harris, P., & Frièß, W. (2019). Application of interpretable artificial neural networks to early monoclonal antibodies development. European Journal of Pharmaceutics and Biopharmaceutics, 141, 81–89.

Glipin, L. H., Bau, D., Yuan, B. Z., Bajwa, A., Specter, M., & Kagal, L. (2019). Explaining explanations: An overview of interpretability of machine learning. Institute of Electrical and Electronics Engineers Inc.

Gleeson, M. P. (2008). Generation of a set of simple, interpretable ADMET rules of thumb. Journal of Medicinal Chemistry, 51(4), 817–834.

Goel, A., Gajula, K., Gupta, R., & Rai, B. (2021). In-silico screening of database for finding potential sweet molecules: A combined data and structure based modeling approach. Food Chemistry, 343, 128538.

Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., & Pedreschi, D. (2018). A survey of methods for explaining black box models. ACM Computing Surveys, 1–42, 51(5). https://doi.org/10.1145/3326009

Han, R., Xiong, H., Ye, Z., Yang, Y., Huang, T., Jing, Q., Lu, J., Pan, H., Ren, F., & Ouyang, D. (2019). Predicting physical stability of solid dispersions by machine learning techniques. Journal of Controlled Release, 311–312, 16–25.

He, K., Zhang, X., Ren, S., & Sun, J. (2015). Deep residual learning for image recognition. http://arxiv.org/abs/1512.03385

Helguera, A. M., Pérez-Garrido, A., Gaspar, A., Reis, J., Cagide, F., Vina, D., Cordeiro, M. N. D. S., & Borges, F. (2013). Combining QSAR classification models for predictive modeling of human monoamine oxidase inhibitors. European Journal of Medicinal Chemistry, 59, 75–90.

Hu, Y., Unwalla, R., Denny, R. A., Blikker, J., Di, L., & Humblet, C. (2010). Development of QSAR models for microsomal stability: Identification of good and bad structural features for rat, human and mouse microsomal stability. Journal of Computer-Aided Molecular Design, 24(1), 23–35.

Huang, L. T., Gromiha, M. M., & Ho, S. Y. (2007). iPTREE-STAB: Interpretable decision tree based method for predicting protein stability changes upon mutations. Bioinformatics, 23(10), 1292–1293.

Huynh-Thu, V. A., Saeyes, Y., Wenhkel, L., & Geurts, P. (2012). Statistical interpretation of machine learning-based feature importance scores for biomarker discovery. Bioinformatics, 28(13), 1766–1774.

Johansson, U., Sönströd, C., Norinder, U., & Boström, H. (2011). Trade-off between accuracy and interpretability for predictive in silico modeling. Future Medicinal Chemistry, 3(6), 647–663.

Karimi, M., Wu, D., Wang, Z., & Shen, Y. (2019). DeepAffinity: Interpretable deep learning of compound-protein affinity through unified recurrent and convolutional neural networks. Bioinformatics, 35, 3329–3338.
Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., & Liu, T.-Y. (2017). LightGBM: A highly efficient gradient boosting decision tree. Neural Information Processing Systems Foundation. https://paperswithcode.com/paper/lightgbm-a-highly-efficient-gradient-boosting

Keles, S., van der Laan, M. J., & Vulpe, C. (2004). Regulatory motif finding by logic regression. Bioinformatics, 20(16), 2799–2811.

Krishnan, V. G., & Westhead, D. R. (2003). A comparative study of machine-learning methods to predict the effects of single nucleotide polymorphisms on protein function. Bioinformatics, 19(17), 2199–2209.

Kuz’min, V. E., Polischuk, P. G., Artemenko, A. G., & Andronati, S. A. (2011). Interpretation of QSAR models based on random forest methods. Molecular Informatics, 30(6–7), 593–603.

Langham, J. J., & Jain, A. N. (2008). Accurate and interpretable computational modeling of chemical mutagenicity. Journal of Chemical Information and Modeling, 48(9), 1833–1839.

Lapinsh, M., Prustis, P., Uhlen, S., & Wikberg, J. E. S. (2013). Accurate and interpretable computational modeling of chemical mutagenicity. Journal of Chemical Information and Modeling, 53(10), 2730–2742.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Müller, A., Nothman, J., Louppe, G., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, É. (2011). Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 12, 2825–2830.

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). “Why should I trust you?” Explaining the predictions of any classifier. Association for Computing Machinery.

Rosenbaum, L., Hinselmann, G., John, A., & Zell, A. (2011). Interpreting linear support vector machine models with heat map molecule coloring. Journal of Cheminformatics, 3(3), 11.

Rush, A. M., Chopra, S., & Weston, J. (2015). A neural attention model for abstractive sentence summarization. Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, Lisbon (pp. 379–389), Portugal. Association for Computational Linguistics.

Saiz-Urra, L., Pérez, M. A. C., Helguera, A. M., & Froeyen, M. (2011). Combining molecular docking and Qsar studies for modelling the antigen-activity of cyclophiline derivatives. European Journal of Medicinal Chemistry, 46(7), 2736–2747.

Santos-Filho, O. A., & Cherkasov, A. (2008). Using molecular docking, 3D-Qsar, and cluster analysis for screening structurally diverse data sets of pharmacological interest. Journal of Chemical Information and Modeling, 48(10), 2054–2065.

Sheridan, R. P. (2019). Interpretation of Qsar models by coloring atoms according to changes in predicted activity: How robust is it? Journal of Chemical Information and Modeling, 59(4), 1324–1337.

Sun, H. (2005). A naive Bayes classifier for prediction of multidrug resistance reversal activity on the basis of atom typing. Journal of Medicinal Chemistry, 48(12), 4031–4039.

Sutherland, J. J., Higgs, R. E., Watson, I., & Vieth, M. (2008). Chemical fragment foundations as understandings for target space and activity prediction. Journal of Medicinal Chemistry, 51(19), 2689–2700.

Tabei, Y., Pauwels, E., Stoven, V., Takemoto, K., & Yamanishi, Y. (2012). Identification of chemogenomic features from drug-target interaction networks using interpretable classifiers. Bioinformatics, 28(18), i487–i494.

Tetko, I. V., Gasteiger, J., Todeschini, R., Mauri, A., Livingstone, D., Ertl, P., Palyulin, V. A., Radchenko, E. V., Zefirov, N. S., Makarenko, A. S., Tanchuk, V. Y., & Prokopenko, V. V. (2005). Virtual computational chemistry laboratory—Design and description. Journal of Computer-Aided Molecular Design, 19(6), 453–463.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A., Kaiser, L., & Polosukhin, I. (2017). Attention is all you need. 31st conference on neural information processing systems (NIPS 2017), Long Beach, CA. Neural Information Processing Systems Foundation.

Wan, H., Rehngren, M., Giordano, F., Bergström, F., & Tunek, A. (2007). High-throughput screening of drug-brain tissue binding and in silico prediction for assessment of central nervous system drug delivery. Journal of Medicinal Chemistry, 50(19), 4606–4615.

Welling, S. H., Clemmensen, L. K. H., Buckley, S. T., Hovgaard, L., Brockhoff, P. B., & Refsgaard, H. H. F. (2015). In silico modelling of permeation enhancement potency in Caco-2 monolayers based on molecular descriptors and random forest. European Journal of Pharmacology and Biopharmaceutics, 94, 152–159.
Wu, Y., Schuster, M., Chen, Z., Le, Q. V., Norouzi, M., Macherey, W., Krikun, M., Cao, Y., Gao, Q., Macherey, K., Klingner, J., Shah, A., Johnson, M., Liu, X., Kaiser, Ł., Gouws, S., Kato, Y., Kudó, T., Kazawa, H., ... Dean, J. (2016). Google’s neural machine translation system: Bridging the gap between human and machine translation. CoRR, vol.abs/1609.08144

Wu, Z., Ramsundar, B., Feinberg, E. N., Gomes, J., Geniesse, C., Pappu, A. S., Leswing, K., & Pande, V. (2018). MoleculeNet: A benchmark for molecular machine learning. Chemical Science, 9(2), 513–530.

Xiong, Z. et al. (2019). Pushing the boundaries of molecular representation for drug discovery with the graph attention mechanism. Journal of Medicinal Chemistry, 63, 8749–8760.

Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhutdinov, R., Zemel, R., & Bengio, Y. (2015). Show, attend and tell: Neural image caption generation with visual attention. Proceedings of the International Machine Learning Society (IMLS) (pp. 2048–2057).

Yang, Y., Ye, Z., Su, Y., Zhao, Q., Li, X., & Ouyang, D. (2019). Deep learning for in vitro prediction of pharmaceutical formulations. Acta Pharmaceutica Sinica B, 9(1), 177–185.

Ye, Z., Yang, Y., Li, X., Cao, D., & Ouyang, D. (2019). An integrated transfer learning and multitask learning approach for pharmacokinetic parameter prediction. Molecular Pharmaceutics, 16(2), 533–541.

Yoshida, K., & Niwa, T. (2006). Quantitative structure—Activity relationship studies on inhibition of HERG potassium channels. Journal of Chemical Information and Modeling, 46(3), 1371–1378.

Yuan, Y., Chipman, H. A., & Welch, W. J. (2012). Harvesting classification trees for drug discovery. Journal of Chemical Information and Modeling, 52(12), 3169–3180.

Zhang, Y. Y., Liu, H., Summerfield, S. G., Luscombe, C. N., & Sahi, J. (2016). Integrating in silico and in vitro approaches to predict drug accessibility to the central nervous system. Molecular Pharmaceutics, 13(5), 1540–1550.

Zhao, Q., Miriyala, N., Su, Y., Chen, W., Gao, X., Shao, L., Yan, R., Li, H., Yao, X., Cao, D., Wang, Y., & Ouyang, D. (2018). Computer-aided formulation design for a highly soluble lutein-cyclodextrin multiple-component delivery system. Molecular Pharmaceutics, 15(4), 1664–1673.

Zhou, P., Wang, C., Tian, F., Ren, Y., Yang, C., & Huang, J. (2013). Biomacromolecular quantitative structure-activity relationship (BioQSAR): A proof-of-concept study on the modeling, prediction and interpretation of protein-protein binding affinity. Journal of Computer-Aided Molecular Design, 27(1), 67–78.

How to cite this article: Ye, Z., Yang, W., Yang, Y., & Ouyang, D. (2021). Interpretable machine learning methods for in vitro pharmaceutical formulation development. Food Frontiers, 2, 195–207. https://doi.org/10.1002/fft2.78