machine learning & knowledge extraction

Article

VLA-SMILES: Variable-Length-Array SMILES Descriptors in Neural Network-Based QSAR Modeling

Antonina L. Nazarova [1,*,†] and Aiichiro Nakano [2,*]

[1] Department of Quantitative & Computational Biology, Bridge Institute, USC Michelson Center for Convergent Bioscience, University of Southern California, Los Angeles, CA 90089, USA
[2] Collaboratory of Advanced Computing and Simulations, Department of Computer Science, Department of Physics & Astronomy, Department of Quantitative & Computational Biology, University of Southern California, Los Angeles, CA 90089, USA
* Correspondence: nazarova@usc.edu (A.L.N.); anakano@usc.edu (A.N.)
† Past address: Loker Hydrocarbon Research Institute, Bridge Institute, Chemistry Department, University of Southern California, Los Angeles, CA 90089, USA.

Abstract: Machine learning represents a milestone in data-driven research, including material informatics, robotics, and computer-aided drug discovery. With the continuously growing virtual and synthetically available chemical space, efficient and robust quantitative structure–activity relationship (QSAR) methods are required to uncover molecules with desired properties. Herein, we propose variable-length-array SMILES-based (VLA-SMILES) structural descriptors that expand conventional SMILES descriptors widely used in machine learning. This structural representation extends the family of numerically coded SMILES, particularly binary SMILES, to expedite the discovery of new deep learning QSAR models with high predictive ability. VLA-SMILES descriptors were shown to speed up the training of QSAR models based on multilayer perceptron (MLP) with optimized backpropagation (ATransformedBP), resilient propagation (iRPROP$^-$), and Adam optimization learning algorithms featuring rational train–test splitting, while improving the predictive ability toward the more compute-intensive binary SMILES representation format. All the tested MLPs under the same length-array-based SMILES descriptors showed similar predictive ability and convergence rate of training in combination with the considered learning procedures. Validation with the Kennard–Stone train–test splitting based on the structural descriptor similarity metrics was found more effective than the partitioning with the ranking by activity based on biological activity values metrics for the entire set of VLA-SMILES featured QSAR. Robustness and the predictive ability of MLP models based on VLA-SMILES were assessed via the method of QSAR parametric model validation. In addition, the method of the statistical H_0 hypothesis testing of the linear regression between real and observed activities based on the $F_{2,n-2}$-criteria was used for predictability estimation among VLA-SMILES featured QSAR-MLPs (with n being the volume of the testing set). Both approaches of QSAR parametric model validation and statistical hypothesis testing were found to correlate when used for the quantitative evaluation of predictabilities of the designed QSAR models with VLA-SMILES descriptors.

Keywords: machine learning; deep learning; neural networks; SMILES; descriptors; QSAR

Citation: Nazarova, A.L.; Nakano, A. VLA-SMILES: Variable-Length-Array SMILES Descriptors in Neural Network-Based QSAR Modeling. *Mach. Learn. Knowl. Extr.* **2022**, *4*, 715–737. https://doi.org/10.3390/make4030034

Academic Editors: Yoichi Hayashi and Andreas Holzinger

Received: 14 June 2022
Accepted: 1 August 2022
Published: 5 August 2022

1. Introduction

In the rising era of big data and artificial intelligence, machine learning (ML)-based technologies have become one of the key approaches in computer-aided drug discovery, allowing fast processing of large-scale and continuously growing chemical libraries [1,2]. Quantitative structure–activity relationship (QSAR) or quantitative structure–property relationship (QSPR)-based modeling evolved as the leading framework for the development of scalable and versatile methods for in silico activity or property prediction [3–6].

When considering a specific biological target, ML modeling remains an efficient and low-cost choice for activity prediction by training the model on a library of compounds with known biological activities. Prior to utilizing a QSAR model for screening unknown compounds, it should be tested on a variety of externally generated sets and demonstrate consistency and robustness for the chosen biological system [7]. Direct research expenses for experimental testing can be substantially reduced with the increasing predictive ability of computer-assisted screening in the sheer magnitude of the synthetically available chemical space [8,9]. Recent QSAR studies have focused on both nonlinear and linear methods, such as the *k* nearest neighbor method (*k*NN) [10], random forest [11], and artificial neural networks (ANNs) [12] including deep learning neural networks (DNN) as a basis of deep learning methods [13]. In a comparative study of 16 different types of ML algorithms for QSAR, the neural network-based, i.e., principal component analysis (PCA)-ANN and deep neural network (DNN), models were reported to be among the best in terms of prediction abilities within those commonly used in QSAR [14]. The development of ANN-based models to improve predictive ability has gained great attention and provided a new direction in ML-based QSAR studies [15–17]. In particular, multilayer perceptrons (MLPs) have been demonstrated to be promising for structural design and biological activity prediction [18]. Success of ANN models such as MLP is dictated by their ability to establish complex nonlinear relationships among different types of predictors, which is the basis of structure–activity (QSAR) or structure–property (QSPR) modeling [18]. The capability of ANNs to solve these complex correlations is linked to the "learning" potential to adapt parameters to fit the multidimensional space of training samples obtained either experimentally or computationally. A variety of learning algorithms and optimization strategies for ANNs, such as input sequence calibration and weight update initializations, have been reported and can be readily applied in various systems [19].

One of the most time-consuming tasks in the development of QSAR models is the preparation of the modeling and validation datasets that should span the entire scope of potential molecular structures and scaffolds [20]. To improve the predictive ability in ML models, the datasets should be accurately described in numerical, computer-friendly notation. Recent ML studies used various types of molecular descriptors [21], among which the SMILES (Simplified Molecular Input Line Entry System) representation proposed by Weininger in 1988 remains one of the most commonly used and low-space-complexity descriptors [22]. Recent advancements in molecular representations using SMILES-based formats have facilitated the discovery of novel therapeutics [23] and enhanced toxicity prediction [24], driving the development of versatile and easily generated molecular descriptors for QSAR studies [25,26]. Despite these successes, several critical issues remain unsolved in neural network-based QSAR modeling. A recent study demonstrated the importance of numerical coding of SMILES, where decimal and binary SMILES coding schemes were shown to work well for solving structure–property relationships in dielectric polymer motifs [27]. However, it remains to be studied systematically how such encoding influences the speed and accuracy of training in a more general setting. In this work, we propose variable-length-array numerical SMILES-based representations (VLA-SMILES) and apply them as digital input sequences for the structural description of bioactive molecules (Figure 1). Notably, compared to binary SMILES for molecular representation, VLA-SMILES structural descriptors have been shown to be reduced in size by array length encoding (*k* clustered binary numbers, explained in Section 2.2) while preserving the structural peculiarities. Thus, the training convergence time when using VLA-SMILES-based structural representation was shown to be k^2 lower compared to the binary SMILES representation format. Two datasets involving small active molecules targeting the human angiotensin II receptor (ATR) of both type 1 and type 2 (Dataset#1) and human immunodeficiency virus-1 (HIV-1) protease receptor (Dataset#2) were generated and used for validation and testing of the designed models. The candidate molecules with corresponding bioactivity data were extracted from the open-access ChEMBL database, the largest publicly available resource of compound bioactivity data [28,29]. The obtained diverse datasets contain both agonist and antagonist ligands of the ATR and inhibitors of HIV-1 protease, which showed activity in either active or inactive receptor conformation states. The VLA-SMILES were implemented and tested in 199 different types of neural network-based QSAR models, including MLP-based models with one and two hidden layers, as well as deep learning models based on autoencoders. In addition to the varied VLA-SMILES input format of molecular structures, the QSAR models also differed by the rational type of the database train–test split algorithms introduced (ranking by activity and Kennard–Stone-based), activation functions (Sigmoid, Tanh, and ReLU), and the learning approaches implemented (ATransformedBP, iRPROP, and Adam).

To quantitatively validate the predictive ability and robustness of developed QSAR models using VLA-SMILES-based descriptors, we utilized the standard parameter-based QSAR validation approach, as well as a newly developed statistical H_0 hypothesis testing methodology. The standard parametric approach includes the calculation of the coefficient of correlation R (Pearson's), the square of the coefficient of correlation R^2, the determination coefficient R_0^2 ($R_0^{2\prime}$), a slope coefficient k (k') for the linear regression through the origin (ideal regression), and the determination coefficient q^2 for linear regression between real and observed activities in the testing phase [30,31]. On the basis of activity prediction results derived for nearly 160 QSAR models, Golbraikh and Tropsha proposed quantitative criteria for these parameters, where the satisfaction of these criteria signifies good prediction ability of a particular model of interest [30]. While Alexander and Tropsha later reported the root-mean-square error (RMSE) and R^2 parameters to be enough for estimating the practical usefulness of a model, the previously defined standard parameters of model validity were still referred to as relevant for measuring a model's predictive ability if properly applied [31]. As an alternative to the standard QSAR validation approach proposed by Golbraikh and Tropsha, the possibility of $F_{1,n-2}$ (with n being the volume of the testing set) distribution function was reported toward acceptance of the statistical H_0 hypothesis of not-better-than-average activity prediction. In addition to the abovementioned method, herein we propose new criteria of statistical H_0 hypothesis testing of the linear regression between real and observed activities based on 2D probability density distributions for the regression coefficients. The validity proofs, as well as correct implementation conditions of the current criteria, are provided by Kendall and Stuart [32]. Thus, calculated critical values $t_{1-\alpha}$ for the $F_{2,n-2}$-statistics (with two and $n - 2$ degrees of freedom) were found to correlate well with the statistical criteria of the QSAR predictability validation approach, as well as with the root-mean-square error (RMSE) parameter for the testing phase.

As rational approaches for dataset partitioning have been demonstrated to provide more diverse results, we formulated the training and testing sets employing rational train–test splitting approaches, Kennard–Stone-based and ranking by activity [20]. For the entire set of VLA-SMILES-based description strategies, MLP-based QSAR models featuring Kennard–Stone splitting yielded better predictive ability than those based on ranking by activity splitting. In addition to dataset splitting optimization, the MLP models were developed using several learning optimizers including affine transformed backpropagation ATransformedBP [27], resilient backpropagation [33], and Adam optimizer [34].

The entire set of VLA-SMILES-coded MLP QSAR models were developed using a C++ codebase. Such self-developed software facilitates flexible and adaptive ML-based model investigation, particularly inner-parameter variability and optimization, which would be more restricted in plug-in-play modules or library packages.

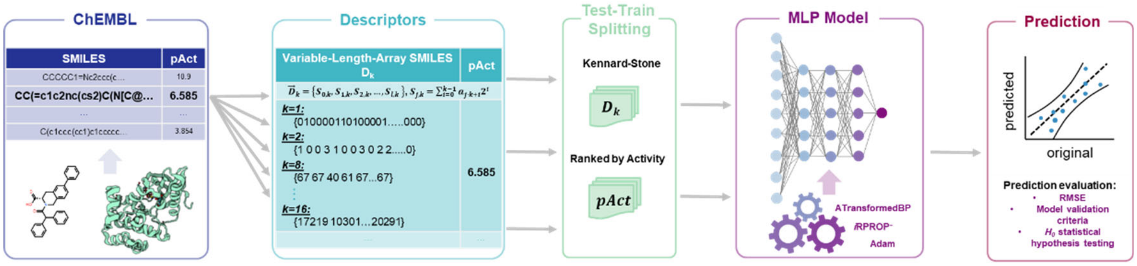

Figure 1. Flowchart of our MLP-based QSAR modeling using variable-length-array numerical SMILES-based descriptors. Batch-trained affine transformed BP (ATransformedBP), epoch trained resilient propagation (iRPROP$^-$), and Adam optimizer learning algorithms form the foundation of the developed MLP models.

2. Materials and Methods

2.1. Dataset Description

The original datasets of the bioactive compounds were obtained from the ChEMBL database (ver. 25) which includes close to 1.8 million chemical structures [28,29]. Both receptor families, angiotensin II and protease, are multifunctional enzymes that play important roles in organism functioning while regulating many biological processes. The ChEMBL pool of compounds targeting the intensively studied human receptor target of angiotensin II receptor type 1 (AT$_1$R) and type 2 (AT$_2$R) consists of 3462 structures. Both AT$_1$R and AT$_2$R play role in the regulation of blood pressure, as well as in sodium excretion [35,36]. Inhibition of AT$_1$R and AT$_2$R reduces the risk of hypertension due to the regulation of cardiovascular and electrolyte homeostasis; activation was recently proposed to be an effective treatment of neurological cognitive disorders, including Alzheimer's disease [37]. The ChEMBL pool of human protease active ligands consists of 1935 structures. Proteases play an essential role in cell behavior and survival, which makes them one of the main drug targets, and they are of interest as prognostic biomarkers of cancer, inflammatory, and cardiovascular diseases [38]. Inhibition of HIV-1 protease is an effective treatment in COVID-19 [39], by stopping the virus's lifecycle, and, for over 30 years, in highly active antiretroviral therapy (HAART) against AIDS [40].

The logarithmic values of the activity parameter in the original human angiotensin II (ATR) receptor database, pAct, span from unknown and 0 to 10.9. Compounds with unknown pAct and pAct = 0, as well as duplicates, oligomers, or structures with high molecular weight, were excluded in the data curation phase. The final Dataset#1 consisted of 1005 ATR active drug-like compounds with a pAct value in the range 3.8–10.9. The same data curation procedures were applied for the human HIV-1 protease receptor pool giving final Dataset#2 of protease active structures with known affinities, consisting of 1378 drug-like compounds with a pAct value in the range 2.7–13.6. Since a compound is considered to be active if revealing pAct > 6, both generated Dataset#1 and Dataset#2 preserved high diversity, while containing ligands with low activity and high activity of pAct being in a range from ~2 to ~10 and higher [41]. While the database also reported other types of activity data (IC50, K_d, K_i, and K_b) for some structures, only pAct values were available for all the compounds. An example compound, its chemical structure, and its SMILES and VLA-SMILES representations, together with the corresponding activity value, pAct, is depicted in Figure 1.

2.2. Data Encoding: Variable-Length-Array (VLA) SMILES-Based Descriptors

SMILES, a single-line spaceless representation, is the most common machine-readable format due to the reversibility and generality of features [22]. The SMILES notation is a conventional form describing chemical structures and is widely used in computer modeling and ML. Thus, one-dimensional SMILES representation has been utilized in predicting the structure–activity or structure–property relationships in the fields of material science [42], biochemistry [43], polymers [44,45], and drug discovery [46]. While the classical SMILES notation is based on a fixed alphabet and follows a set of rules achieving a linear string format, variations of SMILES-type syntax representations demonstrated good performance in structure–activity relationship [47] and generative model [47,48] studies. SELFIES (self-referencing embedded strings) as a string-based molecular representation approach improved memory storage capabilities while retaining the robustness and user-friendliness of SMILES [49]. Variable dictionary-featured text-based representations have shown to be useful when mapping chemical structure information, e.g., CUSTODI (custom tokenization dictionary) [50]. However, implementation of CUSTODI requires nontrivial preprocessing of the dataset. Other expansions of the SMILES string language include dot-separated CurlySMILES [51] (Curly-braces enhanced Smart Material Input Line Entry Specification), eclectic-featured quasi-SMILES [52], and substructure-extended SMARTS [53].

In this work, we designed and used a variety of numerical representations of machine-readable SMILES notation. Initially, all molecular structures in the dataset were defined in the canonical SMILES notation, and the largest string was found to consist of 234 characters. We denote the length of the longest string as L_{max} = 234. Subsequently, all SMILES entries in the dataset were padded with zeros to ensure a consistent length of strings (L = 234). The obtained vectors were then mapped with ASCII decoding tables to represent the atomic composition in SMILES byte-type numerical format of range 0–255. Figure 2 depicts an example flowchart of all steps in the variable-length-array encoding of a molecular structure. This two-step conversion is illustrated in the example of the methoxy group, a common structural motif readily available in the majority of chemical libraries. A methoxy group in SMILES is defined as representation (1) (Figure 2, step (B)).

$$S_{SMILE} = \{C\ o\ c\}. \tag{1}$$

In the first step of the VLA encoding, the SMILES string of the methoxy-group (representation (1)) is converted to the SMILES numerical format using ASCII tables with the following transformation into the binary SMILES having a length of 24 digits (Figure 2, step (C)):

$$D_{bin} = D_1 = \{01000011\ 01{,}001{,}111\ 01100011\}. \tag{2}$$

The SMILES format in binary representation in ASCII codes can be clustered by two, three, and more binary values to be represented in other numerical formats. In the case of Dataset#1 and Dataset#2, the lengths of the resulting vectors in binary representation were equal to d_1 = 1872 and d_2 = 1192 correspondingly. Clustering by the k-th sequenced binary symbols produces numerical sequences with a length of d/k, where k is an array length and an integer factor of d (herein, for Dataset#1 with d_1 = 1872, we used k = 1, 2, 4, 6, 8, 12, and 16, whereas, for Dataset#2 with d_2 = 1192, only the values of k = 1, 2, 4, and 8 were possible). A common representation for the resulting vector unit is defined as

$$S_{j,k} = \sum_{i=0}^{k-1} a_{j\cdot k+i} 2^i, \tag{3}$$

where $a_{j\cdot k+i}$ is a binary unit (either "0" or "1") in the j-th array with j (= 0, 1, 2, …, d/k–1) being an index of an array-based element in the VLA-featured SMILES notation.

For the example of the SMILES-encoded methoxy-group (1), representation (3) can be defined as D_2 variable-length-array-based SMILES representation (array length k = 2).

$$D_2 = \{100310331203\}. \tag{4}$$

Thus, any chemical structure can be described in the VLA-based binary SMILES format using Equation (5).

$$\vec{D}_k = \{S_{0,k}, S_{1,k}, S_{2,k}, \dots, S_{l,k}\}, \tag{5}$$

where $l = 8L_{max}/k - 1$, and $0 \le S_{j,k} \le 2^k - 1$.

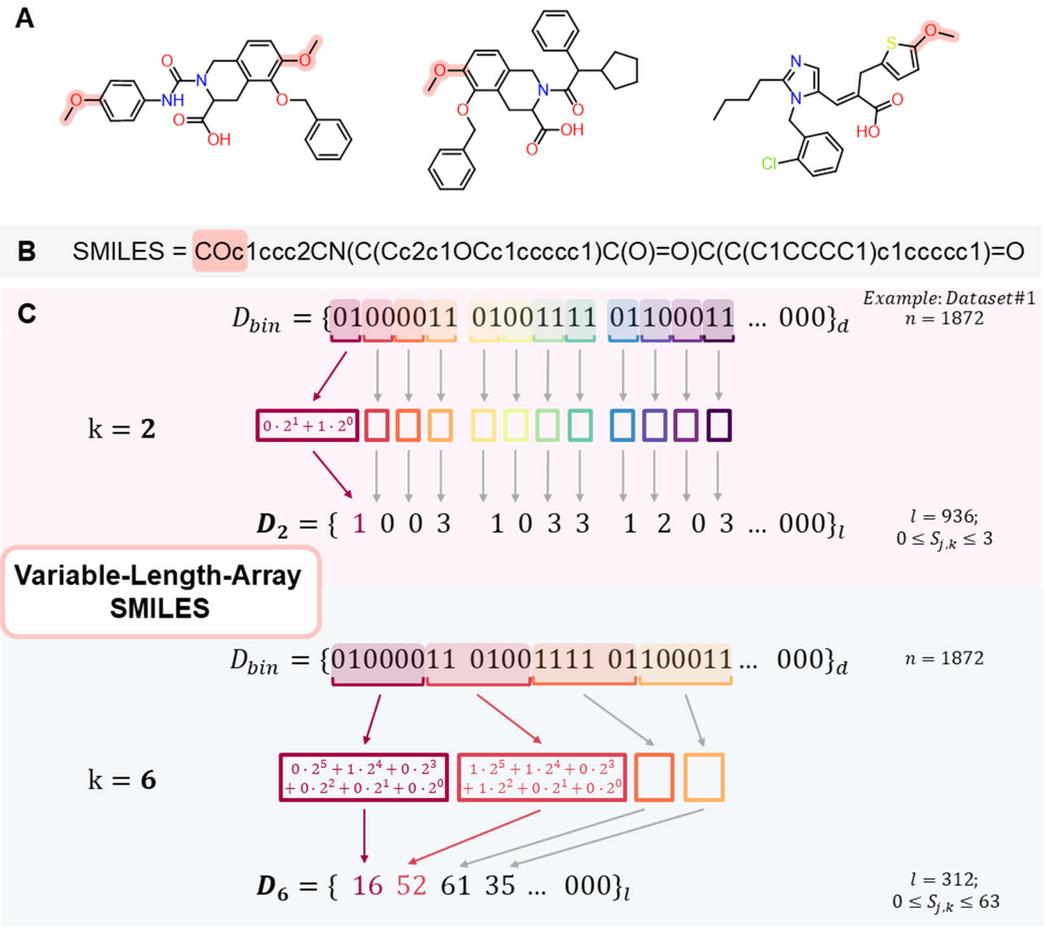
A
B
SMILES = COc1ccc2CN(C(Cc2c1OCc1ccccc1)C(O)=O)C(C(C1CCCC1)c1ccccc1)=O
C
D_{bin} = {01000011 01001111 01100011 ... 000}$_d$
Example: Dataset#1
n = 1872
k = 2
$0\cdot2^1+1\cdot2^0$
D_2 = { 1 0 0 3 1 0 3 3 1 2 0 3 ... 000}$_l$
l = 936; $0 \le S_{j,k} \le 3$
Variable-Length-Array SMILES
D_{bin} = {01000011 01001111 01100011 ... 000}$_d$
n = 1872
k = 6
$0\cdot2^5+1\cdot2^4+0\cdot2^3+0\cdot2^2+0\cdot2^1+0\cdot2^0$
$1\cdot2^5+1\cdot2^4+0\cdot2^3+1\cdot2^2+0\cdot2^1+0\cdot2^0$
D_6 = { 16 52 61 35 ... 000}$_l$
l = 312; $0 \le S_{j,k} \le 63$

Figure 2. Data encoding using the variable-length-array-featured SMILES notation (VLA-SMILES). (**A**) Examples of compounds carrying a methoxy functional group, a common structural motif in the dataset. (**B**) SMILES representation of one example. The methoxy group and its SMILES code are highlighted in red. (**C**) Binary mapped SMILES representation with zero padding at the end of the sequence. Transformation of the obtained arrays D_2 ($k = 2$) and D_6 ($k = 6$).

One of the promising characteristics of the VLA-featured representation is its advanced intermolecular bond description with increasing value of the array k-group. The modeling studies revealed that array-based VLA-SMILES representation enhances the information content of the chemical substructure features compared to SMILES byte-type numerical format. The VLA-type representation can, thus, improve the predictive ability of MLP models that utilize text-based notation as input data. The obtained $S_{j,k}$ values in $\vec{D}_k$ were normalized by a range-scaling procedure to distribute their values within the range −0.5 to 0.5 as

$$\hat{S}_{j,k} = \frac{(S_{j,k} - 0.5 \cdot (S_{j,k(max)} + S_{j,k(min)}))}{(S_{j,k(max)} - S_{j,k(min)})}, \tag{6}$$

where $S_{j,k(\max)}$ and $S_{j,k(\min)}$ are the maximum and minimum values of elements of $\vec{D}_k$.

2.3. Theoretical Background: Multilayer Perceptron and Statistical Metrics of the Model Prediction Ability

The multilayer perceptron (MLP) architecture with classical gradient learning algorithms (e.g., backpropagation [54], RPROP [33], Adam [34]) is considered the basis of variable ANNs, e.g., recurrent (RNN) [27], convolutional (CNN) [55], and graph (GNN) [56]. Here, various MLPs with one, two, and three hidden layers and the abovementioned learning procedures were designed to solve QSARs for a set of active ATR (Dataset#1) and

HIV-1 protease ligands (Dataset#2). A typical schematic representation of the MLP architecture is depicted in Figure 3. The block scheme of the MLP includes an input layer, hidden layers with various activation functions $F(Y)$, and an output layer. Input and hidden layers comprise l neurons with l weighing parameters. The output signals from the first hidden and second layer are determined as

$$\boldsymbol{Y}^{(1)}_{(\boldsymbol{m})\boldsymbol{i}} = \overrightarrow{\boldsymbol{W}}^{(1,\boldsymbol{m})}\vec{\boldsymbol{S}}^{\boldsymbol{T}}{}_{(\boldsymbol{k})\boldsymbol{i}} + \boldsymbol{T}^{(1)}_{(\boldsymbol{m})\boldsymbol{i}}, \tag{7}$$

$$\boldsymbol{Y}^{(2)}_{(\boldsymbol{m})\boldsymbol{i}} = \overrightarrow{\boldsymbol{W}}^{(2,m)}\vec{\boldsymbol{X}}^{\boldsymbol{T}}{}^{(1)}_{(\boldsymbol{m})\boldsymbol{i}} + \boldsymbol{T}^{(2)}_{(\boldsymbol{m})\boldsymbol{i}}, \tag{8}$$

where $\vec{\boldsymbol{S}}_{(k)i} = (\boldsymbol{S}_{(\boldsymbol{0},\boldsymbol{k})\boldsymbol{i}}, \boldsymbol{S}_{(\boldsymbol{1},\boldsymbol{k})\boldsymbol{i}}, \ldots, \boldsymbol{S}_{(\boldsymbol{l},\boldsymbol{k})\boldsymbol{i}})$ is an input row vector with the dimension l for the *i*-th ligand, $\boldsymbol{X}^{(1)}_{(\boldsymbol{m})\boldsymbol{i}} = \boldsymbol{F}(\boldsymbol{Y}^{(1)}_{(\boldsymbol{m})\boldsymbol{i}})$ and $\boldsymbol{X}^{(2)}_{(\boldsymbol{m})\boldsymbol{i}} = \boldsymbol{F}(\boldsymbol{Y}^{(2)}_{(\boldsymbol{m})\boldsymbol{i}})$ are the outputs of an activation function for the first and second hidden layers, $\overrightarrow{\boldsymbol{W}}^{(1,m)} = (\omega_0^{(1,m)}, \omega_1^{(1,m)}, \ldots, \omega_l^{(1,m)})$ and $\overrightarrow{\boldsymbol{W}}^{(2,m)} = (\omega_0^{(2,m)}, \omega_1^{(2,m)}, \ldots, \omega_l^{(2,m)})$ are row vectors of the weighting parameters (m=0,1,…,l), and $\boldsymbol{T}^{(2)}_{(\boldsymbol{m})\boldsymbol{i}}$ and $\boldsymbol{T}^{(1)}_{(\boldsymbol{m})\boldsymbol{i}}$ are biases (not shown in Figure 3).

The message function for the predicted activity can be expressed as

$$\widehat{\boldsymbol{pAct}}_{\boldsymbol{i}} = \overrightarrow{\boldsymbol{W}}^{(3)}\vec{\boldsymbol{X}}^{\boldsymbol{T}}{}^{(2)}_{(\boldsymbol{m})\boldsymbol{i}} + \boldsymbol{T}^{(3)}_{\boldsymbol{i}}, \tag{9}$$

where $\overrightarrow{\boldsymbol{W}}^{(3)}$ is a one-dimensional row vector, and $\boldsymbol{T}^{(3)}_{\boldsymbol{i}}$ is the bias for the output layer (not shown in Figure 3).

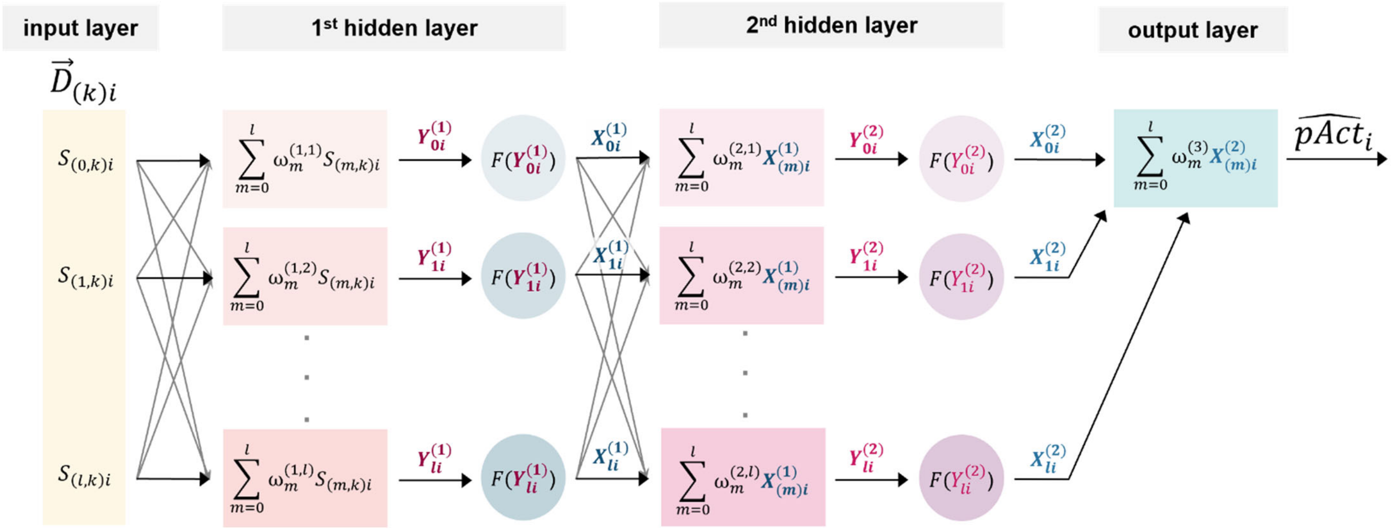

Figure 3. Schematic representation of the multilayer perceptron (MLP) neural network architecture (with two hidden layers) utilizing the VLA-featured binary SMILES input $\vec{D}_{(k)}$.

The activation function $F(Y)$ is critical for achieving a high prediction ability in QSAR. We used three commonly applied activation functions:

$$\text{Hyperbolic tangent, Tanh: } F(Y) = Tanh(Y), \tag{10}$$

$$\text{Sigmoid: } F(Y) = 1/(1 + e^{-Y}), \tag{11}$$

$$\text{Rectified Linear Unit, ReLU: } F(Y) = \begin{cases} 0, Y \le 0 \\ Y, Y > 0 \end{cases}. \tag{12}$$

To determine the prediction ability of MLP models, one needs to evaluate the accuracy of the nonlinear mapping of the variable-length-array SMILES into the predicted values of the activity $\widehat{pAct}$. We used the error parameter, root-mean-square error (RMSE), as evaluation criteria for predictive ability, as well as the learning efficiency, which is referred to as the loss function.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=0}^{n-1}(pAct_i - \widehat{pAct}_i)^2}. \tag{13}$$

As addition error measures, the relative standard deviation parameter (RSD) with the maximum and average values were used. They are defined in Equations (14) and (15), respectively.

$$RSD_{max} = (\max_{i} |(1 - {\widehat{pAct}_i}/{pAct_i})|) \times 100\ (\%), \tag{14}$$

$$RSD_{av} = \frac{1}{n}\sum_{i=0}^{n-1} |(1 - \widehat{pAct}_i/pAct_i)| \times 100\ (\%). \tag{15}$$

2.4. Formation of Training and Testing Sets: Method of Rational Splitting

To evaluate the predictive ability of the QSAR model, one needs to split the original dataset into training and testing sets. For smaller datasets, the leave-one-out cross-validation splitting procedure can be used [57]. Larger databases require the use of random or rational partitioning methods. QSAR modeling relying on random train–test split methodologies tends to perform less effectively if the molecules in the testing set are structurally very different from those in the training set, which can be easily achieved for a random split [58]. More rational division algorithms, i.e., rational selection of the training and test sets, are crucial for developing accurate and robust models that lead to superior generalization characteristics [59–61]. Multiple algorithms have been reported for intelligent dataset division including sphere exclusion [62], Kennard–Stone [63], and ranking by activity methods [64]. The Kennard–Stone algorithm allows the selection of a very diverse subset of compounds in terms of the Euclidean distance between the descriptors. It constitutes the basis of various clustering methods and training set generation for the validation of QSAR models. The ranking of compounds by activity with subsequent separation into equal-sized groups is another commonly used algorithm for intelligent training/testing set generation [64]. Ranking methods are based on cluster preprocessing of the input descriptors. The evolution of the data clustering theory and clustering algorithms covers an independent area of research [65–67]. For this work, we performed and tested two types of data partitioning methods (Figure 4):

- Splitting 1: train–test split using Kennard–Stone algorithm [63],
- Splitting 2: train–test split using ranking by activity [64].

In the case of the Kennard–Stone protocol, VLA-based SMILES representations were used as structure-based descriptors. First, two reference samples $\vec{D}_m$ and $\vec{D}_n$, which had the largest distance between corresponding descriptors, were selected. These reference samples should exhibit $\max_{m,n} \rho(\vec{D}_m, \vec{D}_n)$ in Euclidean metrics are defined as

$$\rho(\vec{D}_m, \vec{D}_n) = \sqrt{\sum_{j=0}^{l-1}(S_{j,m} - S_{j,n})^2}, \tag{16}$$

where *n* and *m* are selected indices of the entries in the original dataset. These two entries are automatically selected for the training set. The next sample *i* needs to be chosen to satisfy the maximal criteria, $\max_{i}(\min(\rho(\vec{D}_i, \vec{D}_m), \rho(\vec{D}_i, \vec{D}_n)))$. The remaining samples are ranked in the descending order of conformity with $\max_{i}(\min_{j} \rho(\vec{D}_i, \vec{D}_j))$, where *j* refers to the samples that have already been ranked. The first 95% samples of the generated pool were selected as the training (learning) set, while the remaining 5% constituted the testing set.

For the ranking-based train–test splitting method, we used the activity parameter pAct as descriptor values for clustering [64]. The samples were sorted in descending order of pAct. The sorted list was then divided into 50 groups, each containing 20 samples sorted in descending order of pAct (Figure 4). The first 19 samples of each group were assigned to the training set (95% of the dataset), whereas the remaining samples formed the testing set (5% of the dataset).

```
Require: $\vec{D}$ preprocessing of the input sequence of chemical descriptors

Splitting 1: Kennard-Stone Splitting Method
  Step 1: $n_1 = 0$ (sorted list), $n = 1000$ (Database#1)
                                   $n = 1370$ (Database#2)
  Finding $\vec{D}_{k,}\vec{D}_m$ with the max Euclidian distance between
  structural descriptors (cluster-featured SMILES):
          $\rho(\vec{D}_{k,}\vec{D}_m) = \max_{i,j \in n} \rho(\vec{D}_i, \vec{D}_j)$
          $\vec{D}_{k,}\vec{D}_m \rightarrow n_1$
          $n_1 = 2$
  Step 2: Finding $\vec{D}_i$ with the maximum minimum value of the
  $\rho(\vec{D}_{i,}\vec{D}_j)$ , where $\vec{D}_j \in n_1$
          while $(m \notin n_1)$
          for all $\vec{D}_j \in n_1$
            $\rho(\vec{D}_{m,}\vec{D}_j)$
          if $\vec{D} = \max_m(\min_j \rho(\vec{D}_m, \vec{D}_j))$
              $\vec{D} \rightarrow n_1$ (sorted list)
              $n_1 + +$
  Step 3: In the obtained sorted list $n_1 = n$, first 95% goes to
  training set, remaining go to the testing set (5% of the
  dataset).

Splitting 2: Ranking by Activity-based Splitting Method
  Step 1: Samples are sorted by activity parameter descriptor pAct
  in the descent order:
          $n_1 = 0$
          while $(m \notin n_1)$
            if $\vec{D} = \max_m (\mathrm{pAct}(\vec{D}_m))$
              $\vec{D} \rightarrow n_1$ (sorted list)
              $n_1 + +$
  Step 2: Division of $n_1 = n$ by N groups. Each group contains
  number of samples equals to the bin size:
          $bin = n_1/N$
  Step 3: Coherent selection of samples from each decently
  ordered $bin$, first 95% goes to training set, remaining go to the
  testing set (5% of the dataset).
```

Figure 4. Pseudocode of the dataset partitioning splitting methods. (**A**) Train–test splitting using Kennard–Stone algorithm. (**B**) Ranking by activity train–test splitting algorithm. Due to the 5%/95% partitioning ratio with regard to the original dataset, we elaborated with the first $n = 1000$ (Dataset#1) and $n = 1370$ (Dataset#2) samples to have a feasible round number for the train–test splitting.

2.5. Training Algorithms

Optimized backpropagation (ATransformedBP), resilient propagation RPROP, and Adam optimizer learning algorithms were used for the training of MLP models. These recurrent learning methods comprise the family of gradient-based ML algorithms [68].

The backpropagation (BP) algorithm is one of the most widely used supervised learning algorithms in neural networks [54]. It continuously updates the weighting parameters for the row-vectors $\overrightarrow{\boldsymbol{W}}^{(\boldsymbol{p},\boldsymbol{m})}$ and biases, e.g., for MLP with p hidden layers as illustrated in Figure 3. Overtraining and local minimum trapping are the most common issues of the learning phase [69]. Overtraining remains a challenge when a model's generalizability becomes substantially lower when achieving excessive accuracy during training [70]. Thus, the decreased RMSE (Equation (13)) for a larger number of epochs during the training phase does not guarantee a similar performance in the testing phase due to overtraining. At the same time, a large RMSE value in the learning phase inevitably results in a high value of RMSE for the testing set, indicating a poor predictive ability of the model. The problematic overtraining can be overcome using a regularization method, particularly early stopping, where the learning is terminated at the minimum of the loss function in the test phase and the corresponding weighting parameters are recorded [70–72]. On the other hand, trapping in a local minimum may be mitigated utilizing resilient backpropagation (RPROP), Adam optimizer, and ATransformedBP (a modification of the BP algorithm with input affine transformation).

We showed that the ATransformedBP approach paired with an affine transform optimization strategy for the input sequence exhibits superior performance for the prediction of polymer dielectric constants using traditional Elman-type recurrent neural networks (RNNs) [27]. The ATransformedBP approach is based on the preprocessing of the SMILES-based input, in our case, the variable-length-array-featured $\vec{D}_k$ vector,

$$\vec{D}'_{k_i} = \gamma \cdot (\vec{D}_{k_i} - \langle \vec{D}_{k_i} \rangle), 1 \leq i \leq n, \tag{17}$$

where γ is the affine-transformation factor.

The optimal value of γ for each of the VLA-based binary SMILES representations was found to satisfy the minimum criteria of the loss function in the testing set,

$$\gamma_{opt} = \underset{\alpha}{min} \langle E(\widehat{pAct}_i - pAct_i) \rangle. \tag{18}$$

The input vector data of the variable-length-array SMILES was used under the optimal value of the hyperparameter α before testing the prediction abilities of the MLP-based models using ATransformedBP (Table S1).

Resilient backpropagation (RPROP) was also implemented to solve the convergence and local minimum problems. Previous comparative studies based on RNN modeling using iRPROP, one of the four types of RPROP learning methods, showed superior results in comparison to backpropagation [27,73–76].

Another learning method that is frequently used in ML-based models is the Adam stochastic gradient-based optimization strategy [34]. This method performs weight updates recursively via calculation of the bias-corrected first and second adaptive moments estimations.

In addition, we designed a deep neural network (DNN)-based QSAR model based on an MLP Autoencoder [77]. The Autoencoder implements phases of the rough setting of the weighting parameters and fine-tuning. Our model consisted of three hidden layers using Adam optimizer as a learning algorithm for the first and second phases of an Autoencoder realization.

The internal parameters of all QSAR models for the structure–activity studies are summarized in Table S1. The design of MLP-based QSAR models involved optimization of several hyperparameters, such as learning rate, γ affine transform parameter, and the number of the epochs, which depended on the array-featured molecular representation, learning algorithm, and NN architecture.

2.6. Statistical Criteria for Predictive Ability of QSAR Models

A model is considered robust and of high predictive capability when the abovementioned quantities satisfy the following criteria for a testing set: $q^2 > 0.5$, $R^2 > 0.6$, $|R_0^2 - R_0'^2| < 0.3$, $0.85 \leq k \leq 1.15$ or $0.85 \leq k' \leq 1.15$ [30]. Here, R_0^2 and k are the determination coefficients and slope values for linear regression through the origin between the actual and predicted, whereas $R_0'^2$ and k' are the corresponding determination coefficients between predicted and actual activities for the testing phase. These conditions are determined on the basis of a linear regression assumption between the observed and predicted values of a specific parameter, in our case, the biological activity pAct. The values of the parameters for the QSAR model predictive ability validation described above were found to be correlated with the RMSE and were used for a comprehensive analysis of model performance.

As a more rigorous criterion to analyze the model's predictive ability, we used statistical hypothesis testing [32]. The H_0 hypothesis assumes a resemblance of the linear regression $\widehat{pAct} = \hat{a} + \hat{b} \cdot pAct$ to the ideal linear regression with values $\hat{a} \approx 0, \hat{b} \approx 1$. The point estimations of the intercept $\hat{a}$ and the slope $\hat{b}$ for the predicted vs. actual data regression are defined as follows [78]:

$$\hat{a} = \frac{\sum_{i=1}^{n} pAct_i^2 \cdot \sum_{i=1}^{n} \widehat{pAct}_i - \sum_{i=1}^{n} pAct_i \sum_{i=1}^{n} pAct_i \cdot \widehat{pAct}_i}{n \cdot \sum_{i=1}^{n} pAct_i^2 - \left(\sum_{i=1}^{n} pAct_i\right)^2}, \tag{19}$$

$$\hat{b} = \frac{n \cdot \sum_{i=1}^{n} pAct_i \cdot \widehat{pAct}_i - \sum_{i=1}^{n} pAct_i \sum_{i=1}^{n} \widehat{pAct}_i}{n \cdot \sum_{i=1}^{n} pAct_i^2 - \left(\sum_{i=1}^{n} pAct_i\right)^2}, \tag{20}$$

where n is the size of the testing set (in our case, $n = 50$ for Dataset#1 and $n = 70$ for Dataset#2).

Validity of the H_0 hypothesis is defined by the confidence intervals $y_0^{(\pm)}(pAct)$ with the significance level α. It is based on the fact that ideal linear regression is within the upper and lower limit for all the predicted $\widehat{pAct}$ values of the testing set. To accept the H_0 hypothesis, the regression values $\widehat{pAct}$ should be within the range $[y_0^{-}(pAct), y_0^{+}(pAct)]$ [32].

$$y_0^{(-)}(pAct) \le \hat{a} + \hat{b} \cdot pAct \le y_0^{(+)}(pAct). \tag{21}$$

where, $y_0^{(+)}(pAct)$ and $y_0^{(-)}(pAct)$ are upper and lower curve limits, defined as follows [32]:

$$y_0^{(\pm)}(pAct) = \hat{a} + \hat{b} \cdot pAct \pm \sqrt{2 \cdot t_{1-\alpha}}\{\frac{s^2}{n} + \frac{s^2 pAct^2}{\sum_{i=1}^{n} pAct_i^2}\}^2, \tag{22}$$

where $t_{1-\alpha}$ determines a critical value for the F_{m_1,m_2} distribution with $m_1 = 2$ and $m_2 = n - 2$ degrees of freedom, for $\alpha = 0.001$, $t_{1-\alpha} = 8.01$ [78], and $s^2 = \frac{1}{n-2}\sum_{i=1}^{n}(\widehat{pAct}_i - (\hat{a} + \hat{b} \cdot pAct_i))^2$. A lower $t_{1-\alpha}$ value for the H_0 hypothesis to be valid (Equation (22)) indicates a higher predictive ability of the QSAR model.

3. Results

RMSE for the learning phase tends to decrease with the increasing values of epochs, whereas the testing set RMSE minimum values depend on the epoch number and defined regularization approach for QSAR modeling. Thus, the RMSE values corresponding to the last epoch of the training phase, as well as the minimum RMSE of the testing phase, were recorded for each prediction model. Depending on the k value of the length-array-featured SMILES representation, the duration of training depended on the epoch referred to as the characteristic minimum of the testing RMSE (Supplementary Materials, Tables S2–S10). The predictive ability comparison included the models with D_1 VLA-SMILES encoding where the clustering was made by $k = 1$ (D_1) sequenced binary symbols, which is commonly referred to as binary SMILES representation.

3.1. Comparison of Kennard–Stone and Ranking by Activity Splitting Methodologies

To compare the Kennard–Stone and Ranking by Activity train–test splitting algorithms, prediction results of QSAR models using both splitting types were evaluated using Dataset#1 (Tables S2, S3, and S6 (Kennard–Stone-based MLPs), Tables S4 and S7 (ranking by activity-based MLPs). The applied learning procedures included iRPROP$^-$ and Adam optimizer for QSAR models based on one hidden layer MLP with various activation functions and VLA-SMILES descriptors. When evaluating the evolution of loss function in testing and training phases, developed QSAR models can be separated into two groups. The first group includes MLPs with RMSE values not exceeding 0.85. Such QSAR models are considered of good predictive ability and are applicable for further QSAR analysis with external datasets. The second group included models with lower RMSE values and are considered of less predictive ability, whereby their application in the QSAR analysis is less efficient.

Table 1 shows the minimum RMSE for testing sets for MLP-based models with iRPROP$^-$ learning and *Sigmoid(Y)* activation with different train–test splitting for Dataset#1. For the Kennard–Stone train–test splitting, the models with length-array-featured SMILES representations D_1, D_2, and D_6 are referred to as the first group of good predictive ability with the RMSE value not exceeding 0.85. When implementing ranking by activity partitioning, all the designed models were in the second group of mild predictive ability, revealing the RMSE minimum of the testing set to exceed 0.85. Albeit of low prediction ability, the D_{12} VLA-SMILES descriptor-based MLP was the only model from the modeling set using ranking by activity for the train–test splitting that outperformed the D_{12} VLA-SMILES descriptors-based MLP model but with Kennard–Stone dataset partitioning.

Table 1. RMSE values for testing sets of MLPs with one hidden layer and iRPROP^{-} learning procedure: Kennard–Stone vs. ranking by activity-based train–test split (*Sigmoid* activation function was used), Dataset#1.

Kennard-Stone-Based Train-Test Splitting							
VLA-SMILES format	D_1 $k = 1$	D_2 $k = 2$	D_4 $k = 4$	D_6 $k = 6$	D_8 $k = 8$	D_{12} $k = 12$	D_{16} $k = 16$
iRPROP ($\boldsymbol{Sigmoid}(\boldsymbol{S})$)							
Minimum RMSE for testing set	0.77	0.77	0.88	0.84	0.94	0.95	0.89
Adam ($\boldsymbol{Sigmoid}(\boldsymbol{S})$)							
Minimum RMSE for testing set	0.82	0.79	0.84	0.94	0.93	0.99	0.93
Ranking by Activity-Based Train-Test Splitting							
VLA-SMILES format	D_1 $k = 1$	D_2 $k = 2$	D_4 $k = 4$	D_6 $k = 6$	D_8 $k = 8$	D_{12} $k = 12$	D_{16} $k = 16$
iRPROP ($\boldsymbol{Sigmoid}(\boldsymbol{S})$)							
Minimum RMSE for testing set	0.87	0.95	0.88	0.87	1.02	0.87	0.94
Adam ($\boldsymbol{Sigmoid}(\boldsymbol{S})$)							
Minimum RMSE for testing set	1.01	1.18	1.14	1.21	1.29	1.11	1.26

Prognosis and training results for the QSAR models with *ReLU* and *Tanh* activations are shown in Tables S3 and S4. The entire family of VLA-SMILES descriptors-based models with ranking by activity partitioning and *ReLU* and *Tanh* activations belonged to the second group of mild predictive ability models. Yet, an interesting observation was that MLPs with D8 VLA-SMILES descriptors revealed the best prediction accuracy compared to other VLA-SMILES-based MLPs for *Tanh* and *RELU* activation sets. With Kennard–Stone-based splitting, models using variable-length-array-based descriptors D_1, D_2, and D_4 (for *ReLU* activation) and D_1, D_2 (for *Tanh* activation) exhibited good predictive ability in terms of RMSE minimum criteria.

The epoch-dependent loss function evolutions in the testing and training phases generated with Kennard–Stone and ranking by activity partitioning for single-layered MLPs with variable-length-array-featured SMILES and *ReLU*, *Sigmoid,* and *Tanh* activations, as well as iRPROP^{-} learning, are shown in Figures S1–S6, respectively.

The same trend in RMSE evolution between the two rational splitting methods was observed for MLP models using the Adam learning procedure (Table S6 and S7). Here, QSAR modeling with ranking by activity splitting and *Sigmoid* activation showed low activity prediction with RMSE values above 1.0 for the entire family of the VLA-based SMILES representations (Table 1). For the MLP with Kennard–Stone partitioning, the models with D_1, D_2, and D_4 featured array length representations of SMILES descriptors were the only ones satisfying RMSE minimum criteria for the good prediction ability modeling group.

Graphical examples of RMSE evolution for the training and testing phases for the QSARs with D_2-featured array length SMILES representations using Kennard–Stone and ranking by activity-based splits are shown in Figure 5. Here, testing progress reached 0.77 and 0.79 minimum values for Kennard–Stone-based partitioning with the iRPROP^{-} and Adam optimizer-based learnings, respectively (Figure 5A). For the models with ranking by activity train–test splitting, the minimum RMSE for those with the VLA-SMILES D_2 reached 0.95 and 1.18 when implementing iRPROP^{-} and Adam learning algorithms, respectively (Figure 5B).

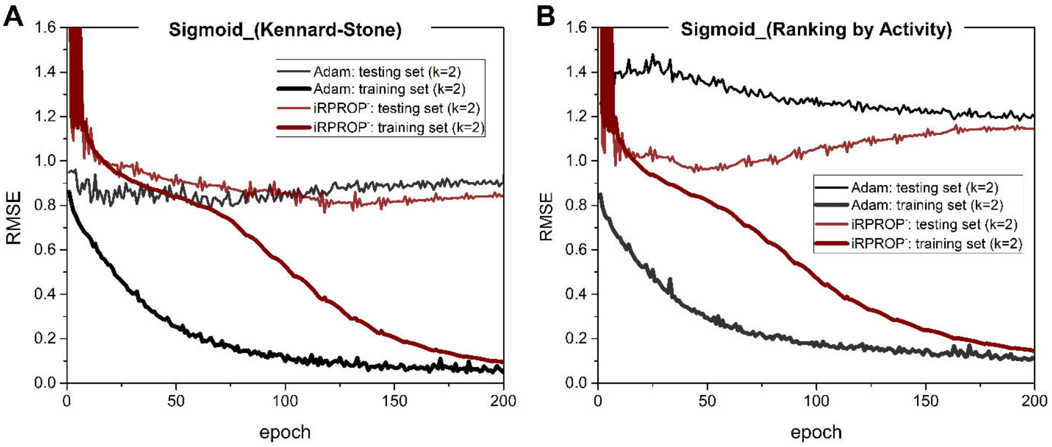

Figure 5. RMSE parameters for training and testing sets for MLPs with one hidden layer with *i*RPROP and Adam learning procedures and *Sigmoid* activation function using (**A**) Kennard–Stone and (**B**) ranking by activity-based train–test splits, Dataset#1.

The other Adam-based MLPs with *ReLU* or *Tanh* activation and Kennard–Stone-based train–test splitting showed superiority over the corresponding models with ranking by activity split on the example of Dataset#1 (Tables S6 and S7). For QSAR implementing the Kennard–Stone algorithm and *ReLU* activation, R_2 and R_4 length-array-featured models had RMSE lower than 0.82, satisfying the requirement for the first group of high-accuracy models. With *Tanh* activation, MLPs models based on VLA-SMILES D_1, D_2, and D_4 also accomplished criteria of good predictive ability modeling. Regardless of the activation function, the entire family of prediction models with the VLA-based SMILES representations using the ranking by activity splitting showed a minimum RMSE value of nearly 1.0 and above. Hence, they were in the second group of models with low predictive ability. The general observation was a reduction of the predictive ability in some VLA-SMILES descriptor-based models with the increasing *k* number of sequenced binary symbols during VLA-encoding. This can be explained by the fact that a higher number of clustered binary numbers leads to increasing of structural peculiarity mismatching during encoding to VLA-SMILES from Binary SMILES. The epoch-dependent RMSE in the testing and training phases of MLPs derived with Kennard–Stone-based and ranking by activity splitting methods for variable-length-array-featured SMILES using *ReLU*, *Sigmoid,* and *Tanh* activation functions and Adam-based learning are presented in Figures S10–15, respectively.

A notable observation was that the training convergence speed increased with increasing values of *k* in length-array-featured representations D_k. Compared with D_1 (k = 1), the traditional binary SMILES-featured MLP, the convergence time for the MLPs with $k > 1$ *VLA*-SMILES notation was proportional to α/k^2, where the prefactor α depends on the CPU (processor made and model, clock speed, number of cores, etc.), as well as the NN model architecture (number of training epochs, learning algorithm, etc.).

$$CPU(time) = \alpha/k^2. \tag{23}$$

Table 2 lists the CPU convergence times for the implemented VLA-SMILES-based MLP with one hidden layer and Adam optimizer. The theoretical CPU times according to Equation (23) (with α = 66.47 × 256 = 17,016.32) were within 4% with regard to the experimentally observed values.

Table 2. CPU times for the training convergence of the models with VLA-SMILES representations (single-layer MLP with Adam learning, Kennard–Stone-based rational train–test splitting, and *Sigmoid* activation). The single prefactor in the theoretical CPU time was determined to reproduce the observed CPU for the MLP with D_{16} (k = 16) cluster-featured SMILES, Dataset#1.

Adam ($\boldsymbol{Sigmoid}$ ($\boldsymbol{S}$))	VLA-SMILES Representation						
	$\boldsymbol{D_1}$ $\boldsymbol{k = 1}$	$\boldsymbol{D_2}$ $\boldsymbol{k = 2}$	$\boldsymbol{D_4}$ $\boldsymbol{k = 4}$	$\boldsymbol{D_6}$ $\boldsymbol{k = 6}$	$\boldsymbol{D_8}$ $\boldsymbol{k = 8}$	$\boldsymbol{D_{12}}$ $\boldsymbol{k = 12}$	$\boldsymbol{D_{16}}$ $\boldsymbol{k = 16}$
CPU time (theor), s	17,016.32	4254.08	1063.52	472.67	265.88	118.17	66.47
CPU time (exp), s	17,651.90	4668.33	1199.30	503.09	268.03	116.56	66.47

ATransformedBP based MLP modeling was implemented only with Kennard–Stone-based partitioning, taking into account its improved performance over ranking by activity. For activation function *Tanh*, the evolution of the loss function for the training and testing phases showed VLA-SMILES D_1, D_2, and D_4 to satisfy the criteria of the first group of models with good predictive ability (Table 3). For the *Sigmoid* activation-built models, the D_1 and D_2 length-array SMILES based models fitted the group of models with the high prediction ability, yet minimum RMSE values for these models were the same or lower than those for *Tanh*-based models. The affine transform γ parameters spanning in the range of 1–4 for each of the designed models were tested to find an optimal one for best prediction results. A full set of γ parameters, as well as epochs corresponding to the testing and training set minimum RMSE values, are reported in Table S2.

Epoch-dependent RMSE values in the testing and training phases with Kennard–Stone-based and ranking by activity splitting methods for different length-array-featured SMILES using *Sigmoid*, *Tanh*, and *ReLU* activation functions and ATransformed-based learning QSAR are shown in Figures S16–18, respectively.

When applying the Kennard–Stone-based train–test partitioning strategy for Dataset#2, the MLP models with D_2 and D_4-based VLA-SMILES descriptors outperformed D_1 length-array SMILES for all three activation functions implemented (Table S5). Thus, single-layered MLPs with D_2 and D_4 VLA-SMILES structural representation and either *ReLU* or *Sigmoid* activation were referred to the first group of models with high predictive ability. The epoch-dependent RMSE for the testing and training phases of MLPs derived with the Kennard–Stone-based splitting method and variable-length-array-featured SMILES using *ReLU*, *Sigmoid,* and *Tanh* activation functions and *i*RPROP$^-$-based learning for Dataset#2 are presented in Figures S7–S9, respectively.

Table 3. RMSE values for testing sets for MLP with one hidden layer using ATransformedBP learning (*Tanh* and *Sigmoid* activation functions, Kennard–Stone based train–test split), Dataset#1.

	Kennard–Stone-Based Train–Test Splitting						
VLA-SMILES format	D_1 k = 1	D_2 k = 2	D_4 k = 4	D_6 k = 6	D_8 k = 8	D_{12} k = 12	D_{16} k = 16
	iRPROP ($\boldsymbol{Sigmoid}$($\boldsymbol{S}$))						
Minimum RMSE for Testing set	0.81	0.80	0.85	0.96	0.90	0.98	0.91
	iRPROP$^-$ ($\boldsymbol{Tanh}$($\boldsymbol{S}$))						
Minimum RMSE for Testing set	0.84	0.80	0.84	0.93	0.90	0.96	0.95
	iRPROP$^-$ ($\boldsymbol{ReLU}$($\boldsymbol{S}$))						
Minimum RMSE for Testing set	0.84	0.82	0.84	0.93	0.93	1.02	0.90

The key findings can be summarized as follows:

1. The Kennard–Stone-based train–test splitting was found to be more efficient than ranking by activity for the investigated QSAR models.
2. The models built on variable-length-array SMILES D_1, D_2, D_4, or D_6 showed equivalent prediction when implemented together with the Kennard–Stone partitioning and were in the first group of models with high predictive ability with RMSE not exceeding 0.85. All types of VLA-featured SMILES-based models with ranking by activity partitioning were in the second group of models of low prediction ability.

3.2. Analysis of Predictive Ability Concerning Activation Functions

MLP models with *Sigmoid* activation exhibited lower RMSE for the majority of variable-length-array SMILES descriptors regardless of the learning algorithm using Dataset#1 (Tables S3 and S6). For iRPROP$^-$ learning and *ReLU* activated series of single-layer MLPs, only two models with the VLA-SMILES-featured representations D_1, and D_2 belonged to the group of models with high predictive ability, having a minimum RMSE of 0.79 and 0.80, respectively. For the models with *Tanh* activation, the only MLP model in the first group of models with high predictive ability was the D_1 length-array SMILES-based one. For comparison, three VLA-based representations D_1, D_2, and D_6 satisfied the rule of the first group of models when using *Sigmoid* activation. A similar trend within single-layer MLPs was observed for models with Adam- and ATransformedBP-based learnings. Thus, for the majority of the QSAR models with variable-length-array SMILES representations, *Sigmoid* activation demonstrated superior prediction results over the *ReLU* or *Tanh*.

3.3. MLP Prediction Models with Two Hidden Layers

The predictive abilities of the MLP models with two hidden layers based on iRPROP$^-$ and Adam optimizer learning algorithms were evaluated (Tables S8 and S9). Both types of two-hidden-layer models led to similar activity prediction results to the prior described results from single-layered MLPs. Table 4 shows the dependency of the RMSE on variable-length-array SMILES for double-layered MLP models with *Sigmoid* activation for the Dataset#1. When considering iRPROP$^-$-based MLPs, only one out of seven VLA-based descriptors model, D_1-based, was in the first group of models with high predictive ability. The models with D_4 and D_8 VLA-featured descriptors resulted in borderline RMSE, thus assigned to the second group of models with low predictive ability. For Adam-based MLP architectures, two models with length-array-based SMILES representations D_1 and D_2 demonstrated compatibility with the first group of models with high predictive ability, having RMSE not exceeding 0.81.

Table 4. RMSE values for testing sets for MLP with two hidden layers using iRPROP$^-$ and Adam optimizer learning (*Sigmoid* activation function, Kennard–Stone-based train–test split), Dataset#1.

MLP	**Two Hidden Layers**						
VLA-SMILES format	D_1 $k = 1$	D_2 $k = 2$	D_4 $k = 4$	D_6 $k = 6$	D_8 $k = 8$	D_{12} $k = 12$	D_{16} $k = 16$
	iRPROP ($\boldsymbol{Sigmoid(S)}$)						
Minimum RMSE for Testing set	0.81	0.87	0.85	0.89	0.85	0.98	0.90
	Adam ($\boldsymbol{Sigmoid(S)}$)						
Minimum RMSE for Testing set	0.81	0.80	0.86	1.01	0.94	0.98	0.90

Figure 6 shows the evolution of RMSE as a function of epoch in the training and testing phases for one- and two-hidden-layer MLPs with iRPROP^{-} and Adam optimizer and the D_4 VLA-featured SMILES descriptors. When using the iRPROP^{-} learning procedure, the testing set's loss function reached a similar minimum RMSE of 0.88 and 0.84 for one- and two-layered MLPs, respectively.

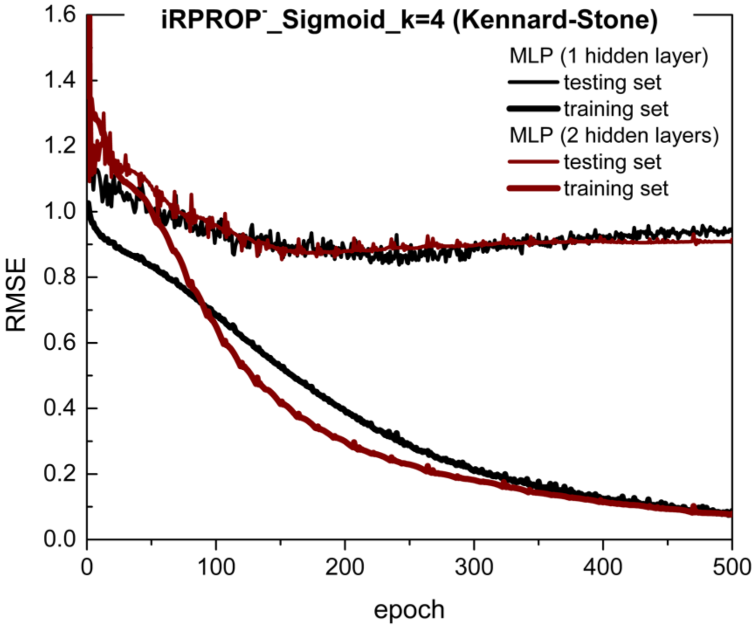

Figure 6. RMSE values as a function of epoch in training and testing sets for iRPROP^{-}-based learning MLP with one and two hidden layers (*Sigmoid* activation function, Kennard–Stone-based train–test splitting), Dataset#1.

MLPs with two hidden layers and iRPROP^{-} but *Tanh* activation showed D_2 and D_4 variable-length-arrays *SMILES* to satisfy criteria of the first group of models with high prediction ability (Table S8). Adam-based MLPs with *Tanh* activation allowed only the D_2-based descriptor model to be in the group of high accuracy QSAR (Table S9). The entire set of epoch-dependent RMSE in the testing and training phases for MLPs with two hidden layers, variable-length-array-featured SMILES, using iRPROP^{-} and Adam optimizer learning procedures is provided in Figures S19–S22, respectively. The key finding is the similarity of the single and two-hidden-layer MLPs models in terms of the prediction for all types of VLA-featured SMILES representations involved.

3.4. Deep Learning, MLP Autoencoder

This section contains experimental evaluations of prediction results derived with MLP Autoencoder modeling. For comparison purposes, the QSAR model based on MLP with one hidden layer, Adam optimizer learning, and *Sigmoid(Y)* activation was taken as a standard MLP method (Table 1). MLP Autoencoder with three hidden layers and 20 iterations of the first phase of rough estimates of the weighting parameters and subsequent fine-tuning emerged as the deep neural network QSAR model (Table 5). Both standard MLP and DNN-built models implemented Kennard–Stone-based train–test splitting methodology. Following comparison studies of RMSE parameters from Tables 1 and 5, the addition of constituting hidden layers did not improve the predictive ability shown by single-layer MLP models. MLP Autoencoder models with D_1, D_2, and D_4 length-array SMILES revealed RMSE minimum satisfying the first group of models with high predictive ability. The same VLA-SMILES-based models also showed high predictive ability among single- and two-hidden-layer MLPs.

Table 5. RMSE values for testing sets of deep learning, MLP Autoencoder with three hidden layers, and Adam learning procedure (Kennard–Stone-based train-test split, *Sigmoid* activation), Dataset#1.

VLA-SMILES	D_1 k = 1	D_2 k = 2	D_4 k = 4	D_6 k = 6	D_8 k = 8	D_{12} k = 12	D_{16} k = 16
Minimum RMSE for testing set	0.85	0.84	0.88	0.94	0.99	1.03	0.92

Graphical examples of RMSE evolution for the training and testing phases for the single-layer MLP and DNN with D_4 length-array-based SMILES are shown in Figure 7. Here, the single-layered MLP testing progress reached minimum RMSE at the value of 0.84, which is nearly the same as for the DNN-based model (minimum RMSE = 0.88). The convergence rate of training for MLP Autoencoder was dependent on the number of iterations in the first phase of tuning and, in this case, was higher than that for the single-layer MLP.

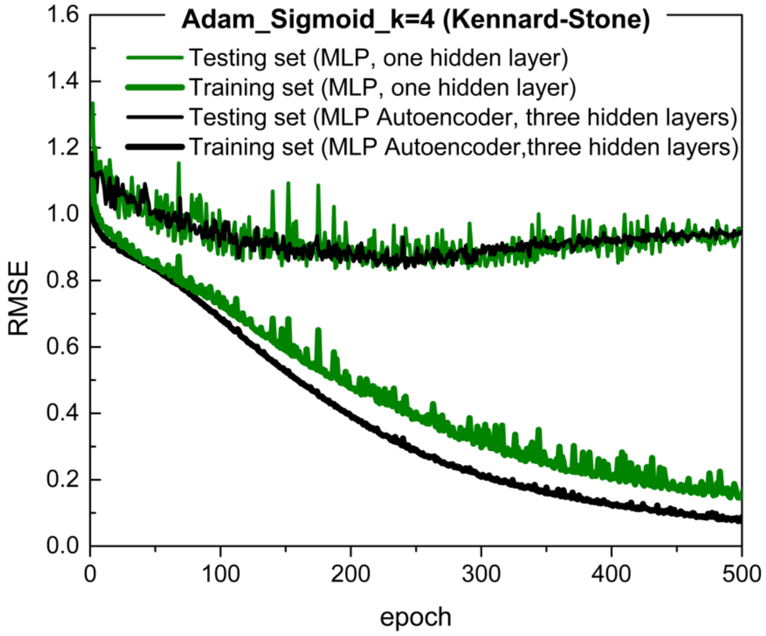

Figure 7. Evolution of RMSE as a function of epoch for training and testing sets for Adam optimizer-based learning MLP with one hidden layer and Autoencoder with three hidden layers (*Sigmoid* activation function, Kennard–Stone-based train–test splitting, and D_4 VLA-SMILES), Dataset#1.

Epoch-dependent RMSE values in the testing and training phases derived with Kennard–Stone-based splitting method for variable-length-array SMILES MLP Autoencoder models using *Sigmoid (S)* activation function are presented in Figure S23. A full set of RMSE parameters, as well as epochs corresponding to the testing phase minimum RMSE and training set final RMSE values for the three-hidden-layer MLP Autoencoder model, are reported in Table S10.

Hence, the increase in the number of hidden layers in the architecture of the proposed VLA-SMILES-based MLP models for QSAR, as well as transition to deep learning, did not lead to substantial improvement in the activity prediction.

3.5. Statistical Analysis of QSAR Model Prediction Ability

In this section, we report the results of the statistical analysis of the model prediction ability using (1) common criteria of QSAR model predictive ability, such as determination coefficients q^2 [64] and R_0^2 ($R_0^{2\prime}$), the square of the Pearson's coefficient of correlation R^2, and slope parameters $k\ (k')$ [30], and (2) method of statistical hypothesis H_0 testing using $F_{2,n\text{-}2}$-statistics (Equation (22)).

Predictive ability evaluation using these criteria was accomplished for MLP models with one hidden layer and the iRPROP$^-$ learning procedure. Following the regularization approach, the predictive ability was evaluated at the epoch corresponding to the minimum of the loss function of the testing phase. Common statistical parameters defining the acceptability of QSAR model with iRPROP$^-$ learning and *Sigmoid (Y)* activation were calculated and are presented in Table 6. As a reminder, when taking into account only the minimum RMSE parameter of the testing phase, the MLPs with D_1, D_2, and D_6 VLA-SMILES were shown to satisfy the first group of models with high accuracy prediction (Table 1). According to the results of the parametric analysis of the QSAR predictive ability, the values of the $\boldsymbol{R^2}$ were found to be similar to either $\boldsymbol{R_0'^2}$ or $\boldsymbol{R_0^2}$ for the entire spectra of the variable-length-array SMILES based MLPs, satisfying the strict condition of the QSAR high predictive ability. Additionally, the slope coefficients $\boldsymbol{k'}$ and $\left|\boldsymbol{R_0^2 - R_0'^2}\right|$ of the linear regression for the testing sets satisfied statistical criteria of model validation for all types of VLA-SMILES representation-based QSAR. Interestingly, the model with the D_4 length-array SMILES showed a minimum RMSE of 0.88, which was slightly higher than a threshold value of a good prediction efficiency, suggested by the authors. Thus, on the basis of only the minimum RMSE criteria, the addition of D_4 VLA-SMILES-based MLP to the group of models with high predictive power was questionable. Meanwhile, following the statistical criteria of $\boldsymbol{q^2}$ and $\boldsymbol{R^2}$, MLPs based on R_4 length-array-featured SMILES (as well as D_8, D_{12}, and D_{16}) did not satisfy requirements for QSAR models with high predictive ability. This remains in agreement with the results of the minimum RMSE analysis reported beforehand (Table 1). The error measure parameter RSD$_{\text{max}}$ for the predicted activity for MLPs with D_1, D_2, and D_6 VLA-featured SMILES reached a level of 23.5%, 24.5%, and 50.5%, respectively, whereas RSD$_{\text{av}}$ did not exceed the level of 8.3%. Overall, the method of prediction ability estimation via the minimum RMSE criteria was found to be in correlation with the model validation parametrical approach.

Table 6. Statistical parameters derived for model predictive ability assessment (single-layered MLP with iRPROP learning and *Sigmoid* activation), Dataset#1.

iRPROP $(\boldsymbol{Sigmoid(S)})$	**VLA-SMILES-Based Representation**						
	$\boldsymbol{D_1}$ $k = 1$	$\boldsymbol{D_2}$ $k = 2$	$\boldsymbol{D_4}$ $k = 4$	$\boldsymbol{D_6}$ $k = 6$	$\boldsymbol{D_8}$ $k = 8$	$\boldsymbol{D_{12}}$ $k = 12$	$\boldsymbol{D_{16}}$ $k = 16$
$\boldsymbol{q^2}$	0.58	0.58	0.44	0.57	0.48	0.47	0.47
$\boldsymbol{R^2}$	0.58	0.61	0.40	0.58	0.51	0.48	0.55
$\boldsymbol{k}$	0.60	0.62	0.46	0.65	0.59	0.47	0.48
$\boldsymbol{k'}$	0.95	0.93	0.98	0.89	0.85	1.00	0.98
$\left|\boldsymbol{R_0^2 - R_0'^2}\right|$	0.02	0.018	0.039	0.01	0.01	0.02	0.01
$\boldsymbol{R_0^2}$	0.57	0.58	0.40	0.58	0.50	0.47	0.46
$\boldsymbol{R_0'^2}$	0.56	0.56	0.36	0.57	0.49	0.45	0.45
$\boldsymbol{RSD_{max}}$, %	23.46	24.47	33.53	50.54	50.47	42.88	40.11
$\boldsymbol{RSD_{av}}$, %	7.46	7.60	8.75	8.28	9.45	8.96	9.44
$\boldsymbol{\hat{\alpha}}$	−0.07	−0.17	0.036	−0.04	−0.09	−0.11	−0.28
$\boldsymbol{\hat{\beta}}$	0.62	0.65	0.44	0.66	0.60	0.48	0.53
$\boldsymbol{t_{1-\alpha}}$	13.27	13.41	23.62	8.69	11.06	26.07	30.49

The results of the method of statistical H_0 hypothesis testing of the linear regression between real and observed activities for the QSAR model with iRPROP$^-$ learning and *Sigmoid* activation are reported in Table 6. Herein, estimations of the linear regression parameters, intercept $\boldsymbol{\hat{\alpha}}$ and the slope $\boldsymbol{\hat{\beta}}$ values, for the predicted vs. original activity data, are reported. Derived slope parameters $\boldsymbol{\hat{\beta}}$ demonstrated MLPs based on D_1, D_2, D_6, and D_8 variable-length-array SMILES to be the closest ones to ideal regression criteria revealing

$\hat{\boldsymbol{\beta}} \geq 0.6$. Thus, none of the models exhibited closeness of $\hat{\boldsymbol{\beta}}$ to 1, yet models with D_4, D_{12}, and D_{16} VLA-SMILES-based descriptors possessed slope values that significantly deviated from the ideal linear regression ($\hat{\boldsymbol{\beta}} < 0.6$), signalizing poor predictive ability. In addition, critical values $t_{1-\alpha}$ for $F_{2,l-2}$-statistics were calculated using Equation (22). MLP models with D_1, D_2, D_6, and D_8 variable-length-array-featured SMILES had $t_{1-\alpha} < 13.5$, whereas MLPs with D_4, D_{12}, and D_{16} length-array SMILES representations demonstrated $t_{1-\alpha} > 23$. Thus, the method of statistical hypothesis testing revealed single-layer iRPROP$^-$-based MLP with VLA-featured SMILES descriptors D_8 to have high predictive power, despite being previously considered a model the low predictive ability with minimum RMSE > 0.85. *F*-statistics analysis also confirmed D_1, D_2, and D_6 VLA-SMILES-featured models to show high prediction power, whereas D_{12} and D_{16} length-array SMILES MLPs were indeed members of the second group of models with poor predictive ability. The parity plots, as well as the limit curves for the testing phase of QSAR models with D_1, D_2, D_6, and D_8 VLA-SMILES formats, are depicted in Figure 8. The dotted line corresponds to the linear regression curve with an intercept $\hat{\boldsymbol{\alpha}}$ and the slope $\hat{\boldsymbol{\beta}}$, the red line signalizes ideal regression, and the upper and lower limit curves resemble $F_{2,l-2}$-statistics with $t_{1-\alpha}$ values that set the requirement for H_0 hypothesis satisfaction. Thus, the statistical method of hypothesis testing was found to be in correlation with the RMSE minimum criteria and parametric model validation methods.

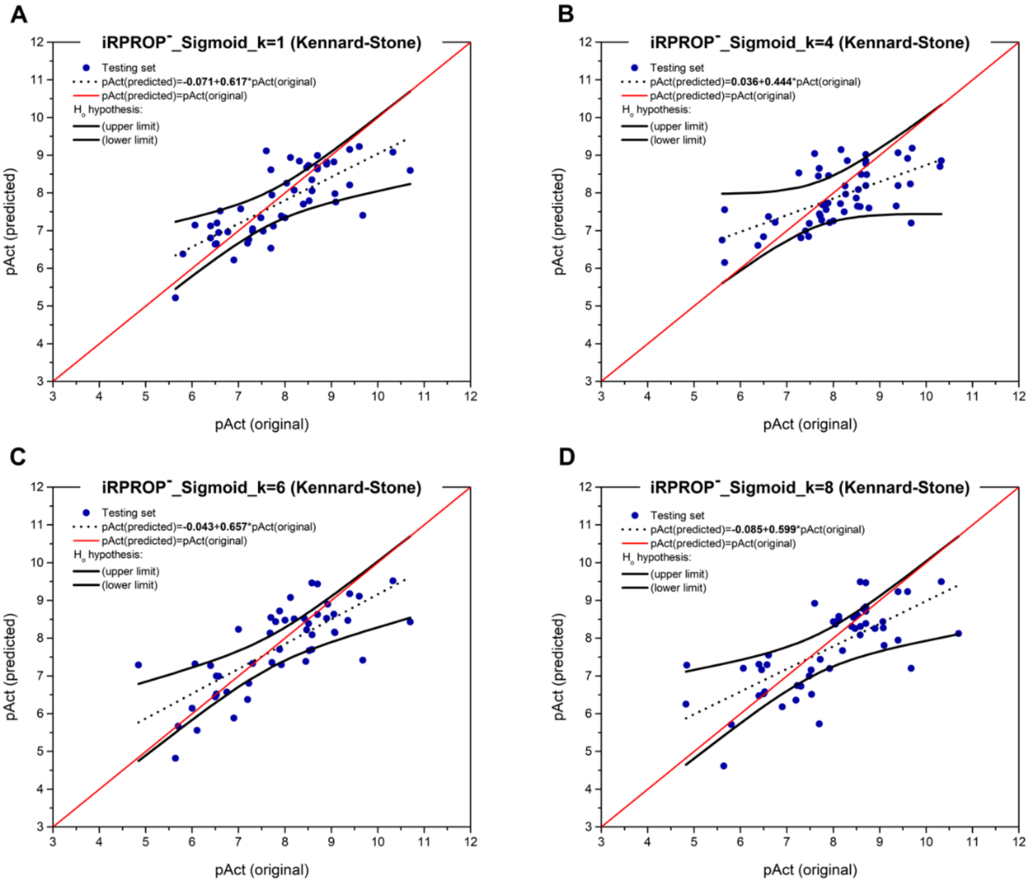

Figure 8. Linear regression parameters for the testing set, and upper and lower limit curves for statistical hypothesis H_0 verification. MLPs with one hidden layer, iRPROP$^-$ optimizer, and *Sigmoid* activation for (**A**) $k = 1$ (D_1), (**B**) $k = 4$ (D_4), (**C**) $k = 6$ (D_6), (**D**) $k = 8$ (D_8) VLA-based SMILES representation, Dataset#1.

Related parity plots, regression, and the upper and lower limit curves for the *F*-statistics for the testing phase of the corresponding cluster-based single-layered MLPs with iRPROP$^-$ learning are reported in Figure S24.

4. Conclusions

We developed a variable-length-array SMILES that allows a robust and straightforward description of the molecular structure contributing to information about intermolecular connectivity. The essence of VLA-featured SMILES is a combination of a sequence (two, three, and more) of SMILES symbols in binary representation to be encoded in other digital formats. The VLA-SMILES descriptors were used for activity prediction with deep learning models, particularly MLP-based QSAR models. Predictive ability was found to increase once the optimal length of the VLA-based SMILES was found.

The developed QSAR MLP models built on VLA-SMILES were based on Adam optimizer, ATransformedBP, and iRPROP$^-$ learning algorithms and various activation functions. The rational splitting procedures for training and testing set generation were implemented for validation of the obtained MLPs. For Dataset#1, the models built with D_1, D_2, D_4, or D_6 VLA-SMILES descriptors (sequences of $k = 1$, $k = 2$, $k = 4$, and $k = 6$ binary SMILES digits, respectively) were the most effective in terms of prediction ability when implemented together with Kennard–Stone partitioning, achieving average RSD_{av} within 8.3%. For dataset#2, the models with D_2, and D_4 VLA-SMILES showed the best prediction ability. Thus, the testing of all possible VLA-SMILES representations for the dataset of interest is required to discover the best variable-length-array encoding in terms of prediction accuracy and training convergence rate.

All types of VLA-SMILES representation-based models with alternative partitioning, i.e., ranking by activity, exhibited lower prediction ability.

Predictive ability was evaluated for MLP models with one and two hidden layers, as well as for MLP Autoencoder with three hidden layers. In comparison with a single-layer MLP, addition of the second and third hidden layers did not improve the activity prediction significantly. When comparing QSAR outcome between MLP with one hidden layer and deep learning with MLP Autoencoder with three hidden layers, again no substantial improvement in the activity prediction was observed.

Based on the calculations and statistical analysis presented in this paper, we conclude that parametric analysis of model validation, as well as the error measure parameter of minimum RMSE, correlate well with the results of the statistical analysis based on H_0 hypothesis testing of an ideal regression verification.

Supplementary Materials: The following supporting information can be downloaded at https://www.mdpi.com/article/10.3390/make4030034/s1: Figures S1–23: RMSE as a function of epoch for learning (training) and testing sets; Figure S24: Parity plot and H_0 hypothesis testing; Table S1: Internal parameters of designed neural network-based QSAR models; Tables S2–10: Training and prediction results in terms of RMSE.

Author Contributions: Conceptualization, A.L.N.; methodology, A.L.N.; software, A.L.N.; validation, A.L.N.; formal analysis, A.L.N.; investigation, A.L.N.; resources, A.L.N. and A.N.; data curation, A.L.N.; writing—original draft preparation, A.L.N.; writing—review and editing, A.N.; visualization, A.L.N.; supervision, A.N.; project administration, A.N.; funding acquisition, A.N. All authors have read and agreed to the published version of the manuscript.

Funding: This study was funded by the National Science Foundation, CyberTraining grant OAC-2118061.

Institutional Review Board Statement: Not applicable

Informed Consent Statement: Not applicable

Data Availability Statement: The following program and data files are available free of charge at DOI:10.5281/zenodo.6361815:

- MLP_Adam_1l.cpp file: MLP model with one hidden layer using Adam optimizer learning algorithm,
- MLP_Adam_2l.cpp file: MLP model with two hidden layers using Adam optimizer learning algorithm,

- DNN_Adam_AutoEncoder.cpp file: MLP model with three hidden layers and AutoEncoder using Adam optimizer learning algorithm,
- MLP_iRPROP-_1l.cpp file: MLP model with one hidden layer using resilient iRPROP learning algorithm,
- MLP_iRPROP-_2l.cpp file: MLP model with two hidden layers using resilient iRPROP learning algorithm,
- MLP_ATransformedBP.cpp file: MLP model with one hidden layer using resilient affine transformed backpropagation learning algorithm ATransformedBP.
- SMILES_LIG.dat, BINARY_SMILES.dat, and pAct.dat: Input files containing ligand structure information (SMILES) and ligand structure information in binary.

Acknowledgments: This work was supported by the National Science Foundation, Award OAC 2118061. Some calculations were performed at the Center for Advanced Research Computing (CARC) of the University of Southern California. Figure 1 and the TOC figure were created using graphical objects from BioRender.com.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Vamathevan, J.; Clark, D.; Czodrowski, P.; Dunham, I.; Ferran, E.; Lee, G.; Li, B.; Madabhushi, A.; Shah, P.; Spitzer, M.; et al. Applications of machine learning in drug discovery and development. *Nat. Rev. Drug Discov.* **2019**, *18*, 463–477.
2. Ekins, S.; Puhl, A.C.; Zorn, K.M.; Lane, T.R.; Russo, D.P.; Klein, J.J.; Hickey, A.J.; Clark, A.M. Exploiting machine learning for end-to-end drug discovery and development. *Nat. Mater.* **2019**, *18*, 435–441.
3. Senior, A.W.; Evans, R.; Jumper, J.; Kirkpatrick, J.; Sifre, L.; Green, T.; Qin, C.; Žídek, A.; Nelson, A.W.R.; Bridgland, A.; et al. Improved protein structure prediction using potentials from deep learning. *Nature* **2020**, *577*, 706–710.
4. Yasonik, J. Multiobjective de novo drug design with recurrent neural networks and nondominated sorting. *J. Cheminform.* **2020**, *12*, 14.
5. Sakai, M.; Nagayasu, K.; Shibui, N.; Andoh, C.; Takayama, K.; Shirakawa, H.; Kaneko, S. Prediction of pharmacological activities from chemical structures with graph convolutional neural networks. *Sci. Rep.* **2021**, *11*, 525.
6. Tsou, L.K.; Yeh, S.-H.; Ueng, S.-H.; Chang, C.-P.; Song, J.-S.; Wu, M.-H.; Chang, H.-F.; Chen, S.-R.; Shih, C.; Chen, C.-T.; et al. Comparative study between deep learning and QSAR classifications for TNBC inhibitors and novel GPCR agonist discovery. *Sci. Rep.* **2020**, *10*, 16771.
7. Cherkasov, A.; Muratov, E.N.; Fourches, D.; Varnek, A.; Baskin, I.I.; Cronin, M.; Dearden, J.; Gramatica, P.; Martin, Y.C.; Todeschini, R.; et al. QSAR Modeling: Where Have You Been? Where Are You Going To? *J. Med. Chem.* **2014**, *57*, 4977–5010.
8. Reymond, J.-L.; Ruddigkeit, L.; Blum, L.; van Deursen, R. The enumeration of chemical space. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* **2012**, *2*, 717–733.
9. Wong, C.H.; Siah, K.W.; Lo, A.W. Estimation of clinical trial success rates and related parameters. *Biostatistics* **2019**, *20*, 273–286.
10. Itskowitz, P.; Tropsha, A. kNearest Neighbors QSAR Modeling as a Variational Problem: Theory and Applications. *J. Chem. Inf. Modeling* **2005**, *45*, 777–785.
11. Svetnik, V.; Liaw, A.; Tong, C.; Culberson, J.C.; Sheridan, R.P.; Feuston, B.P. Random Forest: A Classification and Regression Tool for Compound Classification and QSAR Modeling. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1947–1958.
12. Strieth-Kalthoff, F.; Sandfort, F.; Segler, M.H.S.; Glorius, F. Machine learning the ropes: Principles, applications and directions in synthetic chemistry. *Chem. Soc. Rev.* **2020**, *49*, 6154–6168.
13. Jiménez-Luna, J.; Grisoni, F.; Schneider, G. Drug discovery with explainable artificial intelligence. *Nat. Mach. Intell.* **2020**, *2*, 573–584.
14. Wu, Z.; Zhu, M.; Kang, Y.; Leung, E.L.-H.; Lei, T.; Shen, C.; Jiang, D.; Wang, Z.; Cao, D.; Hou, T. Do we need different machine learning algorithms for QSAR modeling? A comprehensive assessment of 16 machine learning algorithms on 14 QSAR data sets. *Brief. Bioinform.* **2021**, *22*, bbaa321.
15. Baskin, I.I.; Palyulin, V.A.; Zefirov, N.S. Neural Networks in Building QSAR Models. In *Artificial Neural Networks: Methods and Applications*; Livingstone, D.J., Ed.; Humana Press: Totowa, NJ, USA, 2009; pp. 133–154.
16. Hisaki, T.; Aiba née Kaneko, M.; Yamaguchi, M.; Sasa, H.; Kouzuki, H. Development of QSAR models using artificial neural network analysis for risk assessment of repeated-dose, reproductive, and developmental toxicities of cosmetic ingredients. *J. Toxicol. Sci.* **2015**, *40*, 163–180.
17. Žuvela, P.; David, J.; Wong, M.W. Interpretation of ANN-based QSAR models for prediction of antioxidant activity of flavonoids. *J. Comput. Chem.* **2018**, *39*, 953–963.
18. Muratov, E.N.; Bajorath, J.; Sheridan, R.P.; Tetko, I.V.; Filimonov, D.; Poroikov, V.; Oprea, T.I.; Baskin, I.I.; Varnek, A.; Roitberg, A.; et al. QSAR without borders. *Chem. Soc. Rev.* **2020**, *49*, 3525–3564.
19. Wilamowski, B. Neural network architectures and learning algorithms. *IEEE Ind. Electron. Mag.* **2009**, *3*, 56–63.
20. Golbraikh, A.; Tropsha, A. Predictive QSAR modeling based on diversity sampling of experimental datasets for the training and test set selection. *J. Comput.-Aided Mol. Des.* **2002**, *16*, 357–369.

21. Mauri, A.; Consonni, V.; Todeschini, R. Molecular Descriptors. In *Handbook of Computational Chemistry*; Springer: Cham, Switzerland, 2016; pp. 1–29.
22. Weininger, D. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *J. Chem. Inf. Modeling* **1988**, *28*, 31–36.
23. Ponzoni, I.; Sebastián-Pérez, V.; Martínez, M.J.; Roca, C.; De la Cruz Pérez, C.; Cravero, F.; Vazquez, G.E.; Páez, J.A.; Díaz, M.F.; Campillo, N.E. QSAR Classification Models for Predicting the Activity of Inhibitors of Beta-Secretase (BACE1) Associated with Alzheimer's Disease. *Sci. Rep.* **2019**, *9*, 9102.
24. Zhang, J.; Norinder, U.; Svensson, F. Deep Learning-Based Conformal Prediction of Toxicity. *J. Chem. Inf. Modeling* **2021**, *61*, 2648–2657.
25. Winter, R.; Montanari, F.; Noé, F.; Clevert, D.-A. Learning continuous and data-driven molecular descriptors by translating equivalent chemical representations. *Chem. Sci.* **2019**, *10*, 1692–1701.
26. David, L.; Thakkar, A.; Mercado, R.; Engkvist, O. Molecular representations in AI-driven drug discovery: A review and practical guide. *J. Cheminform.* **2020**, *12*, 56.
27. Nazarova, A.L.; Yang, L.; Liu, K.; Mishra, A.; Kalia, R.K.; Nomura, K.-I.; Nakano, A.; Vashishta, P.; Rajak, P. Dielectric Polymer Property Prediction Using Recurrent Neural Networks with Optimizations. *J. Chem. Inf. Modeling* **2021**, *61*, 2175–2186.
28. Mendez, D.; Gaulton, A.; Bento, A.P.; Chambers, J.; De Veij, M.; Félix, E.; Magariños, M.P.; Mosquera, J.F.; Mutowo, P.; Nowotka, M.; et al. ChEMBL: Towards direct deposition of bioassay data. *Nucleic Acids Res.* **2019**, *47*, D930–D940.
29. Davies, M.; Nowotka, M.; Papadatos, G.; Dedman, N.; Gaulton, A.; Atkinson, F.; Bellis, L.; Overington, J.P. ChEMBL web services: Streamlining access to drug discovery data and utilities. *Nucleic Acids Res.* **2015**, *43*, W612–W620.
30. Golbraikh, A.; Tropsha, A. Beware of q2! *J. Mol. Graph. Model.* **2002**, *20*, 269–276.
31. Alexander, D.L.J.; Tropsha, A.; Winkler, D.A. Beware of R2: Simple, Unambiguous Assessment of the Prediction Accuracy of QSAR and QSPR Models. *J. Chem. Inf. Modeling* **2015**, *55*, 1316–1322.
32. Kendall, M.G.; Stuart, A. The Advanced Theory of Statistics. *Volume 2: Inference Relatsh*; Hafner Publishing Company: New York, NY, USA, 1961.
33. Riedmiller, M.; Braun, H. A direct adaptive method for faster backpropagation learning: The RPROP algorithm. In Proceedings of the 1993 IEEE International Conference on Neural Networks, Nagoya, Japan, 25–29 October 1993; pp. 586–591.
34. Kingma, D.P.; Ba, J. Adam: A Method for Stochastic Optimization. *arXiv* **2014**, arXiv: 1412.6980.
35. Van Haaster, M.C.; McDonough, A.A.; Gurley, S.B. Blood pressure regulation by the angiotensin type 1 receptor in the proximal tubule. *Curr. Opin. Nephrol. Hypertens.* **2018**, *27*, 1–7.
36. Fatima, N.; Patel, S.N.; Hussain, T. Angiotensin II Type 2 Receptor: A Target for Protection Against Hypertension, Metabolic Dysfunction, and Organ Remodeling. *Hypertension* **2021**, *77*, 1845–1856.
37. Royea, J.; Lacalle-Aurioles, M.; Trigiani, L.J.; Fermigier, A.; Hamel, E. AT2R's (Angiotensin II Type 2 Receptor's) Role in Cognitive and Cerebrovascular Deficits in a Mouse Model of Alzheimer Disease. *Hypertension* **2020**, *75*, 1464–1474.
38. Bond, J.S. Proteases: History, discovery, and roles in health and disease. *J. Biol. Chem.* **2019**, *294*, 1643–1651.
39. Sagawa, T.; Inoue, K.-I.; Takano, H. Use of protease inhibitors for the prevention of COVID-19. *Prev. Med.* **2020**, *141*, 106280.
40. Wang, Y.; Lv, Z.; Chu, Y. HIV protease inhibitors: A review of molecular selectivity and toxicity. *HIV/AIDS–Res. Palliat. Care* **2015**, *7*, 95.
41. Patel, N.; Huang, X.P.; Grandner, J.M.; Johansson, L.C.; Stauch, B.; McCorvy, J.D.; Liu, Y.; Roth, B.; Katritch, V. Structure-based discovery of potent and selective melatonin receptor agonists. *eLife* **2020**, *9*, e53779.
42. Sun, W.; Zheng, Y.; Yang, K.; Zhang, Q.; Shah, A.A.; Wu, Z.; Sun, Y.; Feng, L.; Chen, D.; Xiao, Z.; et al. Machine learning–assisted molecular design and efficiency prediction for high-performance organic photovoltaic materials. *Sci. Adv.* **2019**, *5*, eaay4275.
43. Remington, J.M.; Ferrell, J.B.; Zorman, M.; Petrucci, A.; Schneebeli, S.T.; Li, J. Machine Learning in a Molecular Modeling Course for Chemistry, Biochemistry, and Biophysics Students. *Biophys.* **2020**, *1*, 11.
44. Doan Tran, H.; Kim, C.; Chen, L.; Chandrasekaran, A.; Batra, R.; Venkatram, S.; Kamal, D.; Lightstone, J.P.; Gurnani, R.; Shetty, P.; et al. Machine-learning predictions of polymer properties with Polymer Genome. *J. Appl. Phys.* **2020**, *128*, 171104.
45. Arabnia, H.R.; Deligiannidis, L.; Grimaila, M.R.; Hodson, D.D.; Joe, K.; Sekijima, M.; Tinetti, F.G. *Advances in Parallel & Distributed Processing, and Applications*; Includes all accepted papers of PDPTA, CSC, MSV, GCC 2020; Springer: Berlin/Heidelberg, Germany, 2020.
46. Segler, M.H.S.; Kogej, T.; Tyrchan, C.; Waller, M.P. Generating Focused Molecule Libraries for Drug Discovery with Recurrent Neural Networks. *ACS Cent. Sci.* **2017**, *4*, 120–131.
47. Li, X.; Fourches, D. SMILES Pair Encoding: A Data-Driven Substructure Tokenization Algorithm for Deep Learning. *J. Chem. Inf. Modeling* **2021**, *61*, 1560–1569.
48. O'Boyle, N.; Dalke, A. DeepSMILES: An Adaptation of SMILES for Use in Machine-Learning of Chemical Structures. *ChemRxiv* **2018**. https://doi.org/10.26434/chemrxiv.7097960.v1.
49. Krenn, M.; Häse, F.; Nigam, A.; Friederich, P.; Aspuru-Guzik, A. Self-referencing embedded strings (SELFIES): A 100% robust molecular string representation. *Mach. Learn. Sci. Technol.* **2020**, *1*, 045024.
50. Fite, S.; Nitecki, O.; Gross, Z. Custom Tokenization Dictionary, CUSTODI: A General, Fast, and Reversible Data-Driven Representation and Regressor. *J. Chem. Inf. Modeling* **2021**, *61*, 3285–3291.
51. Drefahl, A. CurlySMILES: A chemical language to customize and annotate encodings of molecular and nanodevice structures. *J. Cheminform.* **2011**, *3*, 1.

52. Toropova, A.P.; Toropov, A.A.; Veselinović, A.M.; Veselinović, J.B.; Leszczynska, D.; Leszczynski, J. Quasi-SMILES as a Novel Tool for Prediction of Nanomaterials' Endpoints. In *Multi-Scale Approaches in Drug Discovery: From Empirical Knowledge to In Silico Experiments and Back*; Speck-Planche, A., Ed.; Elsevier: Amsterdam, The Netherlands, 2017; pp. 191–221.
53. Ropp, P.J.; Kaminsky, J.C.; Yablonski, S.; Durrant, J.D. Dimorphite-DL: An open-source program for enumerating the ionization states of drug-like small molecules. *J. Cheminform.* **2019**, *11*, 14
54. Rumelhart, D.E.; Hinton, G.E.; Williams, R.J. Learning representations by back-propagating errors. *Nature* **1986**, *323*, 533–536.
55. Desai, M.; Shah, M. An anatomization on breast cancer detection and diagnosis employing multi-layer perceptron neural network (MLP) and Convolutional neural network (CNN). *Clin. eHealth* **2021**, *4*, 1–11.
56. Xu, K.; Hu, W.; Leskovec, J.; Jegelka, S. How Powerful are Graph Neural Networks? *arXiv* **2018**, arXiv: 1810.00826.
57. Rácz, A.; Bajusz, D.; Héberger, K. Effect of Dataset Size and Train/Test Split Ratios in QSAR/QSPR Multiclass Classification. *Molecules* **2021**, *26*, 1111.
58. Tan, J.; Yang, J.; Wu, S.; Chen, G.; Zhao, J. A critical look at the current train/test split in machine learning. *arXiv* **2021**, arXiv: 2106.04525.
59. Puzyn, T.; Mostrag-Szlichtyng, A.; Gajewicz, A.; Skrzyński, M.; Worth, A.P. Investigating the influence of data splitting on the predictive ability of QSAR/QSPR models. *Struct. Chem.* **2011**, *22*, 795–804.
60. Martin, T.M.; Harten, P.; Young, D.M.; Muratov, E.N.; Golbraikh, A.; Zhu, H.; Tropsha, A. Does Rational Selection of Training and Test Sets Improve the Outcome of QSAR Modeling? *J. Chem. Inf. Modeling* **2012**, *52*, 2570–2578.
61. Ng, W.; Minasny, B.; Malone, B.; Filippi, P. In search of an optimum sampling algorithm for prediction of soil properties from infrared spectra. *PeerJ* **2018**, *6*, e5722.
62. Snarey, M.; Terrett, N.K.; Willett, P.; Wilton, D.J. Comparison of algorithms for dissimilarity-based compound selection. *J. Mol. Graph. Model.* **1997**, *15*, 372–385.
63. Kennard, R.W.; Stone, L.A. Computer Aided Design of Experiments. *Technometrics* **1969**, *11*, 137–148.
64. Golbraikh, A.; Shen, M.; Xiao, Z.; Xiao, Y.-D.; Lee, K.-H.; Tropsha, A. Rational selection of training and test sets for the development of validated QSAR models. *J. Comput.-Aided Mol. Des.* **2003**, *17*, 241–253.
65. Puggina Bianchesi, N.M.; Romao, E.L.; Lopes, M.F.B.P.; Balestrassi, P.P.; De Paiva, A.P. A Design of Experiments Comparative Study on Clustering Methods. *IEEE Access* **2019**, *7*, 167726–167738.
66. Gobbi, A.; Giannetti, A.M.; Chen, H.; Lee, M.-L. Atom-Atom-Path similarity and Sphere Exclusion clustering: Tools for prioritizing fragment hits. *J. Cheminform.* **2015**, *7*, 11.
67. Jain, A.K.; Murty, M.N.; Flynn, P.J. Data clustering. *ACM Comput. Surv.* **1999**, *31*, 264–323.
68. Pojas, R. *Neural Networks*; Springer: Berlin/Heidelberg, Germany, 1996.
69. Van Ooyen, A.; Nienhuis, B. Improving the convergence of the back-propagation algorithm. *Neural Netw.* **1992**, *5*, 465–471.
70. Hagiwara, K. Regularization learning, early stopping and biased estimator. *Neurocomputing* **2002**, *48*, 937–955.
71. Zur, R.M.; Jiang, Y.; Pesce, L.L.; Drukker, K. Noise injection for training artificial neural networks: A comparison with weight decay and early stopping. *Med. Phys.* **2009**, *36*, 4810–4818.
72. Yao, Y.; Rosasco, L.; Caponnetto, A. On Early Stopping in Gradient Descent Learning. *Constr. Approx.* **2007**, *26*, 289–315.
73. Reed, R.; Marksil, R.J. *Neural Smithing*; MIT Press: Cambridge, MA, USA, 1999.
74. Igel, C.; Hüsken, M. Empirical evaluation of the improved Rprop learning algorithms. *Neurocomputing* **2003**, *50*, 105–123.
75. Xinxing, P.; Lee, B.; Chunrong, Z. A comparison of neural network backpropagation algorithms for electricity load forecasting. In Proceedings of the 2013 IEEE International Workshop on Inteligent Energy Systems (IWIES), Vienna, Austria, 14 November 2013; pp. 22–27.
76. Avan, E.; Sartono, B. Comparison of Backpropagation and Resilient Backpropagation Algorithms in Non-Invasive Blood Glucose Measuring Device. *Int. J. Eng. Res.* **2017**, *8*, 153–157.
77. Yu, S.; Príncipe, J.C. Understanding autoencoders with information theoretic concepts. *Neural Netw.* **2019**, *117*, 104–123.
78. Sachs, L. *Applied Statistics. A Handbook of Techniques*; Springer: Berlin/Heidelberg, Germany, 1984; p. 349.