Use of machine learning tools and NIR spectra to estimate residual moisture in freeze-dried products

Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy [ISSN: 0022-3549], 293, Article 122485.
DOI: 10.1016/j.saa.2023.122485
https://www.sciencedirect.com/science/article/pii/S1386142523001701

Use of Machine Learning Tools and NIR Spectra to Estimate Residual Moisture in Freeze-Dried Products

Ambra Massei[1,2], Nunzia Falco[2], Davide Fissore[1]

1. Dipartimento di Scienza Applicata e Tecnologia, Politecnico di Torino, corso Duca degli Abruzzi 25, 10129 Torino
2. Global Pharmaceutical Development Department, Merck Serono SpA, via Luigi Einaudi 11, 00012 Guidonia Montecelio (Roma)

Abstract

Residual Moisture (RM) in freeze-dried products is one of the most important critical quality attributes (CQAs) to monitor, since it affects the stability of the active pharmaceutical ingredient (API). The standard experimental method adopted for the measurements of RM is the Karl-Fischer (KF) titration, that is a destructive and time-consuming technique. Therefore, Near-Infrared (NIR) spectroscopy was widely investigated in the last decades as an alternative tool to quantify the RM. In the present paper, a novel method was developed based on NIR spectroscopy combined with machine learning tools for the prediction of RM in freeze-dried products. Two different types of models were used: a linear regression model and a neural network based one. The architecture of the neural network was chosen so as to optimize the prediction of the residual moisture, by minimizing the root mean square error with the dataset used in the learning step. Moreover, the parity plots and the absolute error plots were reported, allowing a visual evaluation of the results. Different factors were considered when developing the model, namely the range of wavelengths considered, the shape of the spectra and the type of model. The possibility of developing the model using a smaller dataset, obtained with just one product, that could be then applied to a wider range of products was investigated, as well as the performance of a model developed for a dataset encompassing several products. Different formulations were analyzed: the main part of the dataset was characterized by a different percentage of sucrose in solution (3%, 6% and 9% specifically); a smaller part was made up of sucrose-arginine mixtures at different percentages and only one formulation was characterized by another excipient, the trehalose. The product-specific model for the 6% sucrose mixture was found consistent for the prediction of RM in other sucrose containing mixtures and in the one containing trehalose, while failed for the dataset with higher percentage of arginine. Therefore, a global model was developed by including a certain percentage of all the available dataset in the calibration phase. Results presented and discussed in this paper demonstrate the higher accuracy and robustness of the machine learning based model with respect to the linear models.

Keywords

Freeze-drying, Residual moisture, Near-Infrared Spectroscopy, Karl Fischer titration, Machine Learning, Neural Networks

List of abbreviations

ANN	Artificial Neural Networks
API	Active Pharmaceutical Ingredient
CQA	Critical Quality Attribute
CV	Cross Validation
FDA	Food and Drug Administration
KF	Karl Fischer
LM	Levenberg-Marquardt
LR	Linear Regression
ML	Machine Learning
MSE	Mean Squared Error
MVA	Multivariate Analysis
NIR	Near-Infrared
PAT	Process Analytical Technologies
PCA	Principal Component Analysis
PLS	Partial Least Square
QbD	Quality by Design
RM	Residual Moisture
RMSE	Root Mean Square Error
RMSEC	Root Mean Square Error of Calibration
RMSECV	Root Mean Square Error of Cross-Validation
SNV	Standard Normal Variate
SOP	Standard Operating Procedure
SR	Small Range
WR	Wide Range

List of symbols

J	number of wavelengths
M	number of samples
$\mathbf{X}$	matrix of NIR spectra (M x J)
$\mathbf{Y}$	matrix of quality attributes (M x 1)

1 Introduction

Freeze-drying is a crucial step in many drug manufacturing processes as it provides long-term stability to formulations containing an active pharmaceutical ingredient (API). The aim of the freeze-drying process is to remove the water present in a product by sublimation, converting the ice into vapour, by operating at low pressure and temperature. The low operating temperatures make this process particularly suitable for heat-sensitive products, such as pharmaceuticals. [1]

Pharmaceutical companies must meet standards imposed by regulatory agencies and pharmacopeia or selected by the manufacturer, so final products must meet certain Critical Quality Attributes (CQAs) [2,3]. Since water could generate biological and chemical degradation processes, the residual moisture (RM) is one of the main CQA to monitor in order to assess the quality of the product.

Currently, CQAs are measured through laboratory testing of samples collected from a batch. The most used method to measure RM in freeze-dried products is the Karl Fisher (KF) titration, which has a lot of disadvantages, such as it is a destructive method, so the samples analyzed are wasted, turning out as an economic loss for the company income. It is time-consuming, since handling of the sample is required, and the instrument must be calibrated before each analytical session. Moreover, safety issues for operators are not negligible, since polluting reactants (formamide and methanol) are involved [4,5].

For improving pharmaceutical developments and manufacturing, new technologies have been encouraged by the regulatory authorities in the last years. In particular, FDA published in 2002 the *Pharmaceutical cGMPs for the 21st Century: A Risk-Based Approach*, and in 2004 the *PAT – A Framework for Innovative Pharmaceutical Development Manufacturing and Quality Assurance* [6,7]. This new approach introduces the concept of *Quality-by-Design* (QbD) according to which the quality of the product has to be embedded in its production process and not just tested at the end of the manufacturing. In this framework, Near-Infrared Spectroscopy (NIR Spectroscopy) has been investigated a lot as one of the most powerful Process Analytical Technologies (PAT) tool in many fields, such as the agricultural, food and pharmaceutical industries. In fact, it is a rapid, non-invasive method that requires minimal sample pretreatments. Moreover, it allows to verify the RM on majority of the vials within a batch, instead of a fraction of it, for demonstration of batch homogeneity and uniformity. Due to the strong absorption of water around 5150 cm^{-1}, NIR spectroscopy was widely used for the determination of RM [8,9,10,11,12]. The main challenging issue is finding a reliable model that allows to predict the RM value from the NIR spectra. This can be done through the chemometric analysis.

Chemometrics is the method to extract chemical relevant information from the available data. Most

of the published applications for RM determination are focused on the application of Partial Least Squares (PLS) as regression method used in the model development step. It turned out to be a powerful tool, but it considers only the linear dependence between the spectra and the residual moisture content [13,14,15,16].

To address this issue and to try to improve the accuracy of the models, machine learning tools could be a suitable alternative for dealing with complex data.

Machine Learning techniques differ from traditional algorithms because they have also the ability to learn as well as to apply pre-programmed decisions. Traditional software receives input data and codes written by the user and generate an output. Machine Learning algorithms, on the other hand, are able to find the functional relationship that binds the input data with the desired output. Their formulation does not require *a priori* knowledge of the physics governing the system or the relationships that link input and output variables [17]. Therefore, the main advantages of machine learning tools are the capability of nonlinear modelling and to give very good results with little knowledge and little training data [18].

Most of the published studies using machine learning tools refer to the food and agriculture field. For example, Parastar *et al*. demonstrated the possibility to discriminate fresh from thawed meat by using different machine learning algorithms based on NIR spectra [19]. Coronel-Reyes *et al.* determined the egg storage time at room temperature using an artificial neural network (ANN) [20]. Richter *et al.* used machine learning techniques to establish the geographical origin of white asparagus [21]. Martins *et al.* presented a deep learning architecture for the prediction of the soluble solids content of fruit. [22] However, very few studies have investigated the coupling between NIR spectroscopy and machine learning tools in the pharmaceutical industry. An attempt in this direction was made by Zhao *et al.* They developed prediction models, coupling machine learning approaches with NIR spectroscopy, for a rapid quantification of three active pharmaceutical ingredients (API). They also compared the performances with the traditional PLS algorithm [23]. Akbar *et al.* explained strategies toward machine learning-based mAb design and the computational and experimental steps required. However, the most challenging issue for the application of machine learning in biopharmaceutical manufacturing is to demonstrate the robustness of the models for GMP use [24].

Here, a linear regression model and a neural network based model were developed to predict the RM values in freeze-dried products starting from NIR spectra. The performances of the two models were compared by calculating the RMSE value and plotting the parity diagrams. The aim of the present work is to demonstrate the feasibility of the application of neural networks for this purpose and their better performance with respect to the linear model. Moreover, the proof of the robustness and the ability of the neural network based model to predict RM values of products not involved in the

calibration step is another goal of this work. Lastly, the effect of reducing the size of the training step on the model performance was investigated.

2 Materials and Methods

2.1 *Machine Learning Techniques*

The interest in Machine Learning (ML) technique has been continuously growing in recent years. Machine learning algorithms use computational methods to "learn" information directly from data, without relying on predetermined equations, e.g. a first-principle based model [25,26,27,28].

A typical workflow for the building of a machine learning model consists of the following steps [29]. Firstly, the starting dataset must contain several input properties, called *feature*, and the outputs, called labels or *targets*. In the present case study, the initial dataset is made up of the matrix **X** of the NIR spectra (input) and the vector **Y** of residual moisture values measured by KF titration of a specific formulation (output). The dimensions of the two variables are respectively *M x J* and *M x 1*. In the **X** matrix, each column represents a wavelength (*J*), so a variable, and each row is a measurement, so a spectrum (*M*). The latter parameter can also be interpreted as the number of samples analyzed. Machine learning algorithms are not smart enough to understand the difference between noise and structured information contained into the data. Therefore, the next step is the preprocessing of the dataset to identify potentially outlier and remove the noise of measurements. In this step a scaling or a normalization of the data could be necessary, as the Standard Normal Variate (SNV) method [25]. In this preliminary phase, the Principal Component Analysis conducted by Bobba et *al.* was exploited for the wavelength range selection, as deeply described in their work [30]. Briefly, by looking at the spectra of samples with different residual moisture it can be seen that the greatest variation occurs at about 5150 cm^{-1}, which corresponds to the specific peak of water. Water is a component that all the datasets used have in common. Therefore, the range of wavelengths of interest should encompass this value and to obtain a robust model, mostly influenced by water content, it might be effective to focus on a narrow range of wavelengths that encompassed the water peak. In this way, specific peaks due to product specific characteristics were less considered. All these statements are in line with the findings of the previous works of Bobba *et al.* [30]. This step was a sort of feature selection, allowing the identification of the most relevant variables for building the model, in this case the range of wavelengths specific to water. Then, the dataset was split into two parts: 70% of the dataset was used as training set ($\mathbf{X}_{\mathbf{train}}$ and $\mathbf{Y}_{\mathbf{train}}$) and the remaining 30% as test set ($\mathbf{X}_{\mathbf{test}}$ and $\mathbf{Y}_{\mathbf{test}}$). These percentages were also varied during the study, decreasing the size of the dataset up to 40% for training purposes, aiming to identify the minimum size of the dataset needed for training. The training set was used in

the calibration phase of the model, where it processed the spectra with the values of RM obtained by KF to find out the model. In this phase the choice of the learning algorithm was really important. In the present work two different algorithms were compared: a linear regression model and a neural network. There is no best method or one size fits. Finding the right algorithm is partly just trial and error [18]. Then, the test set ($\mathbf{X_{test}}$), that the model has never seen before, was projected into the trained model to obtain the predicted values of residual moisture ($\mathbf{Y_{pred}}$). To assess the predictivity of the model, the Root Mean Squared Error (RMSE) value was calculated with Eq. 1 between the RM value measured by KF (y_i) and the one predicted by the model ($y_{pred,i}$):

$$RMSE = \left[\frac{\sum_{i=1}^{M}\left(y_i - y_{pred,i}\right)^2}{M}\right]^{0.5} \quad (1)$$

The number of samples in the dataset analyzed was reported as *M*. The lower the RMSE value, the better the model performances. Also, the Root Mean Squared Error of Calibration (RMSEC) and of Cross-Validation (RMSECV) were calculated to assess the accuracy of the model. They were defined as the squared difference between the measured and the predicted values of residual moisture divided by the number of samples forming, respectively, the calibration set and the cross-validation one. If similar values between the two quantities are obtained, it means that the model is performing well.
Moreover, the parity plots were reported, which correlate the RM values calculated by the model (y-axis) with the RM values measured by KF (x-axis). Obviously, the best situation is that all points lie on the bisector, i.e. the line corresponding to $y = x$, meaning that the value of residual moisture measured through KF is equal to the calculated one from NIR spectra, while the farther the point is from the bisector, the worse the agreement of the ANN model.
It has been also useful to report the diagrams that correlate the absolute error (%) with the RM values measured by KF. An arbitrary threshold for this absolute error (difference between residual moisture calculated and measured by KF) equal to 0.5% (indicated by a red line in the following graphs) was chosen. Ideally the value of absolute error should be zero, but this is (obviously) quite impossible. Therefore, a slightly little discrepancy between the two instruments (NIR and KF) had to be set, and the value of 0.5% was considered adequate in this study. In this framework it has to be considered that also KF titration is affected by an error, that may be considered equal to 0.3%. Therefore, when using NIRs for RM in line estimation we do not know if the error is due just to the ANN model or also to KF: we can take into account this occurrence by setting a lower target value of RM.

2.2 *Linear Regression Model*

Linear models make a prediction using a linear function of the input features. These models are simple to interpret and easy to fit, so usually they are used as a baseline for evaluating other, more complex, regression models [25].

In a linear model the output, the RM value in this case, is the weighted sum of the input features, so the matrix of NIR spectra, plus the bias term (that is the constant intercept term). For multivariate regression, the general prediction formula for a linear model is reported in Eq. 2:

$$y_{pred,i} = \theta_0 \cdot x_0^i + \theta_1 \cdot x_1^i + \cdots + \theta_p \cdot x_p^i + b \quad (2)$$

The input variables are denoted with x_0 to x_p and *i* refers to the rows of the **X** matrix. θ and b are parameters of the model that are learned and $y_{pred,i}$ is the prediction the model makes. Training the model means setting its parameters so that it best fits the training set. Therefore, the aim of the model is to find the parameters that minimize the cost function, that is the Mean Squared Error (MSE), defined as the sum of the squares of the residuals [27].

Spyder (Python 3.9) was used to create the linear regression model by using machine learning exploiting the module Scikit-learn. This module includes a number of libraries that must be installed, such as *NumPy*, *SciPy* and *matplotlib*. *NumPy* is essential because any data will have to be converted to a *NumPy* array. *SciPy* provides advanced linear algebra routines, mathematical function optimization and statistical distribution. *Matplotlib* allows the visualization of the data, line charts, histograms or scatter plots. All these libraries were installed and imported in Spyder. The command *train_test_split* was used to split the dataset into the training and the test set [31].

2.3 *Neural Network*

Neural networks, also known as artificial neural network (ANN), are used in machine learning field and are at the heart of deep learning [17]. The main limitation of the linear models is the resolution of nonlinear problems, so characterized by very different features as input. Neural networks can be a powerful tool to overcome this issue. Their structures are inspired by the human brain, mimicking the interaction of biological neurons with each other [28].

A neural network consists of neurons connected between each other that relate the inputs to the desired outputs [32]. They are characterized by three main parts: the input layer, one or more hidden layers and the output layer. It can be seen as a generalization of linear models that performs multiple stages of processing to come to a decision. In fact, each layer is made up of a certain number of nodes or neurons that have an associated weight [25]. The input layer is characterized by a number of neurons equal to the number of inputs: so, in the presented case, the values of absorbances at specific

wavelengths. The number of neurons of the hidden layer were chosen by trials and errors, since there is no specific rule to follow. The best combination was found to be two hidden layers with respectively 10 and 5 neurons, found by trials and errors. The output layer consists of one single neuron since it is a regression task and the aim was the determination of the residual moisture content in the final product. The values of the input layer are collected and sent to each neuron of the next hidden layer. Here, all inputs are multiplied by their respective weights and then summed up. These weights help the network to determine which variables contribute in a more significant way to the output. To make this model more powerful than a linear one, a nonlinear function is applied to the result, also known as activation function. Afterward, the result is used in the weighted sum that computes the output. This allows neural networks to learn much more complex functions than a linear model could. The main activation functions are the logistic regression, the rectified linear unit (ReLU) and the hyperbolic tangent. It has been found in literature that logistic regression is primarily used for classification tasks. Making several attempts it was evident that the hyperbolic tangent was the function that generates the best results. Neural networks models have a lot of coefficients (weights and bias) to learn with respect to the linear model: one between every input and every hidden layer and one between every neuron in the hidden layer and the output [25,27]. The output of each neuron in each hidden layer is given by Eq. 4:

$$out_j^{(r)} = f\left(b_{r-1} + \sum_{i=1}^{n_{r-1}} w_{i,j}^{(r-1)} out_i^{(r-1)}\right) \quad (4)$$

The first term, $out_j^{(r)}$, is the output of the j^{th} neuron in the r layer and $out_i^{(r-1)}$ represents the output of the i^{th} neuron of the previous layer. The weights are reported as $w_{i,j}$ and are referred to the neuron of the previous layer, as the bias[29,33]. The activation function is indicated with f and is given by $f = tanh()$.

MATLAB (R2019b) was used to create the neural networks used in the present work exploiting the *Neural Net Fitting* toolbox. It is a toolbox allowing the building of the neural network by setting several parameters. Firstly, it was essential to define the percentages of training and test set of the initial dataset. After the construction of the network a training function must be selected. In this work the Levenberg-Maqruardt algorithm was used, by selecting *trainlm* in the associated setting, which has been found to be much more efficient than other techniques. As transfer function, the hyperbolic tangent was used by selecting *tansig* for all the hidden layers. The preprocessing of the data was applied both to the input and the output data by processing PCA with the command *processpca* and

by normalizing them to fall in the range [-1,1] with the command *mapminmax.* Further details on the Levenberg-Marquardt algorithm may be found in [34,35,36].

During the training step, neural networks learn by repeating the process of forward and backpropagation for every input variable several epochs updating the values of the parameters. The training of the neural network stops when the maximum number of epochs (equal to 1000) is reached, the performance gradient falls below the minimum value, or the validation performance has increased more than the maximum failures selected (equal to 6). The most challenging step is the validation of the network with the test set. Once the neural network has fit the data, it forms a generalization of the input-output relationship and is ready to predict new input variables that has never been seen [32].

2.4 *Experimental Procedures*

Freeze-drying cycles were conducted in the laboratories of the Guidonia Montecelio (Italy) site of Merck Serono S.p.A using a lab-scale freeze-drier (Lyostar3, SP Scientific, Warminster, USA). Specifically, the dataset acquired by Bobba *et al*. was expanded carrying out additional tests in the same operating conditions. Other experimental data were acquired from the previously published study [30]. Seven different aqueous solutions, freeze-dried into 2R glass vial (Nuova Ompi, Piombino Dese, Italy) with a filling volume of 1 mL, were considered for samples preparation:

- sucrose 6%$_w$ aqueous solution, labelled as S6;
- sucrose 3%$_w$ aqueous solution, labelled as S3;
- sucrose 9%$_w$ aqueous solution, labelled as S9;
- sucrose 6%$_w$ + arginine 0.5%$_w$ aqueous solution, labelled as SA05;
- sucrose 6%$_w$ + arginine 1%$_w$ aqueous solution, labelled as SA1;
- sucrose 3%$_w$ + arginine 3%$_w$ aqueous solution, labelled as SA3;
- trehalose 6%$_w$ aqueous solution, labelled as T6.

Sucrose and arginine were supplied by Merck Life Science (Darmstadt, Germany), while trehalose by Sigma-Aldrich (Saint Louis, USA). Ultra-pure water was obtained by a Millipore water system (IQ 7000, Merck Millipore, Burlington, USA). Vials were placed in a honeycomb layout and surrounded by metal frames, in direct contact with the shelves of the freeze-dryer. The process conditions of the freeze-drying cycle conducted were the same used by Bobba *et al.* [30] and here reported for the sake of clarity:

- freezing at -45°C for 6 h, with an annealing step at -15°C for 2 h;
- primary drying at -25°C and 5 Pa for 30 h;
- secondary drying at 35°C and 5 Pa for 10 h.

All the cooling / heating rates were set at ± 2°C/min, except for the heating rate in the transition from

the primary to the secondary drying, set at +1°C/min. In order to explore a wide range of moisture in the samples, an already implemented manual humidification was made to get a range of residual moisture in the sample between 1 – 5 %. The amount of water to be added (in the order of μL) in each vial was calculated by multiplying the weight of the cake by the target moisture content to reach, by assuming an initial RM of 0.5 %. Then, the small amount of water was inserted into the stopper and vials were closed upside-down. They were left in the upside position over-night, leading the diffusion of water [37]. The number of samples making up each dataset is summarized in Table 1.

Table 1: Number of samples in each dataset and description of the corresponding formulation.

Data set	Formulation	N° Samples
S6	Sucrose 6%$_w$	91
S3	Sucrose 3%$_w$	63
S9	Sucrose 9%$_w$	36
SA05	Sucrose 6%$_w$ + arginine 0.5%$_w$	30
SA1	Sucrose 6%$_w$ + arginine 1%$_w$	28
SA3	Sucrose 3%$_w$ + arginine 3%$_w$	31
T6	Trehalose 6%$_w$	45

2.5 *NIR Spectra Acquisition*

After freeze-drying and humidification, all the samples were analyzed by a Fourier Transform NIR spectrometer (Antaris MX FT-NIR, Thermo Fischer Scientific, Waltham, USA), equipped with an InGaAs detector and a halogen NIR source. The acquisition of the spectra was in diffuse reflectance mode in the full wavelength range 10000 – 4000 cm^{-1}. The spectrum of each sample was the result of the average between 96 scans to reduce the noise of measurement and increase its quality. The NIR probe pointed on the side of the freeze-dried cake [30].

2.6 *Karl-Fisher titration*

After the acquisition of the spectra, all the samples were analyzed by KF titrations following the Standard Operative Procedure (SOP) of the company. A coulometric titrator was employed (C30S Mettler Toledo, Columbus, USA). The equipment was calibrated before each analytical session with the titration of a standard solution (Honeywell HYDRANAL Water Standard 1.0, Fisher Scientific, Milano, Italy). The solvent used included formamide and methanol and the percentage of water content was calculated [30]. The KF titration was used as reference method to build the regression

model. Being a standardized analytical procedure, the reference values contain an analytical error and a limit of detection. The expected error from a KF analysis may be up to $\pm$0.3% [38, 39, 40]. A such high error with respect to the NIR measurements is justified by the handling and variability of the sample preparation.

2.7 *Pretreatment of spectra*

Preprocessing is a general term for methods to go from "raw" instrumental data to "clean" data for the processing step. In fact, due to the working principle of instruments of measurement, many physical and chemical phenomena can cause a deviation from the linear relationship given by the Beer's law. This results in noise and offset in the plot of the spectra and the aim is to remove it and make all the samples comparable with each other [41,42]. The must-have preprocessing technique used with NIR spectra is the Standard Normal Variate (SNV) correction. It is a row-wise method that allows to highlight better the actual differences between samples. By looking at the full wavelength's spectra, the trend appeared flat for wavenumber values higher than $7000\ cm^{-1}$. Therefore, the wavelength range used for model development was reduced to $7000 - 4250\ cm^{-1}$, according to the findings of Bobba *et al.* [30]. Moreover, some spectra were very noisy and different from the expected ones and so they were removed by PCA or manually.

2.8 *Models developed*

Three models were developed in two different wavelengths ranges. The signals specific to water were observed in the band of O-H stretching and H-O-H bending at around 5150 cm^{-1}. Also, an overtone of O-H stretch was observed at around 6900 cm^{-1}. As pointed out above, most of the spectral information was observed in the region between 4250 and 7000 cm^{-1}. Here, signals specific to the analyzed product were noticed. For instance, in the region between 4000 and 4500 cm^{-1} the sucrose presented a peak corresponding to the C-H stretching. The same situation was almost noticed for the trehalose. Instead, by looking at the spectra of the sucrose-arginine mixtures it appears that as the concentration of arginine increases (in terms of percentage of the total solid fraction), the peak in the water region tends to decrease and another peak, specific to the arginine product, appeared at around 4900 cm^{-1}. The spectra of the different formulations in the smaller region were reported in Fig. S1 and in Fig. S2 of the Supplementary Information.

Taking into account all these considerations, two different wavelengths ranges were examined:

- Small Range (*SR*): it considered the 5290 – 4787 cm^{-1} region and focused on the most significant peak of water at around 5150 cm^{-1}, the one with the highest loading value. According to previous literature works, in this way specific peaks in product characteristics

were less considered and the model would be more robust and generalized [30].

- Wide Range (*WR*): it considered the 7100 – 4250 cm^{-1} region and also included the peaks characteristics of the different products contained in the individual formulations.

In this framework, three different models were developed:

1) *Model S6*: dataset S6 was used (as reference product) as to build and internally validate the model and all the remaining datasets (S3, S9, SA05, SA1, SA3 and T6) were used as external validation set. The percentages of the training and test set were changed during the analysis from 70% up to 40%. The purpose of this model was to obtain a robust model in the perspective of reducing the experimental effort for model development and laboratory testing. The term "robust" refers to the ability of the model to predict with good accuracy the RM value of formulations not included in the calibration step. This may be possible if the inputs given to the model are similar, that is, if the formulations have comparable spectra. In addition, the use of neural networks greatly improved the analysis and can handle even small differences in the input data. The dataset used as reference product (S6 in this case) to build the model was chosen arbitrarily, based on the fact that a larger amount of data was available for this product, although, in principle, any other dataset could have been chosen.
2) *Ad – Hoc Models:* two *ad hoc* models were developed for the sucrose-arginine mixture at different percentages and for the trehalose solution. This step was done to assess the quality of the dataset containing an amino-acid (arginine) and a different excipient (trehalose) with respect to the dataset used for the calibration step in the development of the model. In fact, very bad results were obtained with sucrose-arginine mixture as external dataset with the Model S6. The key reasons will be explained in the following section.
3) *Global Model (GM):* a single dataset, including a certain percentage of all the datasets shown in Table 1, was used in the calibration phase. Also, in this analysis, the percentages of the training set were changed from 70% to 40% to see if any worsening in performance occurred. Many trials were done in this direction by previous literature works, but very few have used neural networks. Being nonlinear tools, they have the advantage to better handle the huge and so different dataset, used as input, with respect to the linear model.

3 Results and Discussion

The performances of the two algorithms, based on the linear model and on neural network, were compared for the prediction of residual moisture in freeze-dried products. The performances are summarized in Table 2 and in Table 3, where the RMSE is shown for Model S6 and the Global model

(Table 2) and for the Ad-Hoc Models (Table 3), both in the small range and in the wide range, used for the various datasets.

Neural network and linear regression turned out in comparable performances in the case of the two *ad hoc* models, as confirmed by the RMSE values in Table 3. For example, in the case of SA05 dataset the RMSEs were respectively 0.288 (LR) and 0.233 (NN). On the contrary, neural network was more accurate in the prediction of dataset different from the ones used in the calibration phase to develop the model, i.e. Model S6. In fact, by considering for example the T6 dataset the RMSE of the neural network was much lower than the one calculated by the linear model (0.199 against 0.334). Also in the Global Model, neural network was better performing thanks to its non-linearity. Especially, its main advantage is the capacity of dealing well with the presence of arginine or trehalose. The RMSE values in Table 2 confirmed these findings. A detailed discussion is provided in the following sections.

Table 2: RMSE obtained when using the Model S6 and the Global Model in the wavelength range SR or WR, and using the linear model (LR) or the neural network (NN).

	Model S6 SR		**Global Model SR**		**Model S6 WR**		**Global Model WR**	
Dataset	*RMSE LR*	*RMSE NN*	*RMSE LR*	*RMSE NN*	*RMSE LR*	*RMSE NN*	*RMSE LR*	*RMSE NN*
S6	0.122	0.172	0.358	0.213	0.066	0.156	0.133	0.167
S3	0.519	0.446	0.493	0.295	0.509	0.681	0.269	0.298
S9	0.774	0.305	0.444	0.309	0.539	0.296	0.249	0.259
SA05	0.873	0.554	0.658	0.356	0.765	0.643	0.426	0.367
SA1	0.707	0.351	0.436	0.387	0.583	0.803	0.269	0.573
SA3	0.519	0.572	0.352	0.173	0.514	0.963	0.367	0.318
T6	0.334	0.199	0.325	0.181	0.773	0.338	0.283	0.249

Table 3: RMSE obtained when using the LR and the NN models in case of *ad hoc* models for arginine-sucrose and for trehalose formulations. All the results reported were obtained for SR.

	Ad-hoc Model for Arginine-Sucrose		**Ad-hoc Model for Trehalose**	
Dataset	*RMSE LR*	*RMSE NN*	*RMSE LR*	*RMSE NN*
SA05	0.288	0.233	-	-
SA1	0.336	0.229	-	-
SA3	0.227	0.115	-	-
T6	-	-	0.334	0.199

3.1 *Model S6*

Model S6 was developed and tested considering both wavelength ranges, SR and WR. The model was developed splitting the S6 dataset into two portions: 70% was used as training set (64 samples) and the remaining 30% as test set (27 samples) in a random way, so in a way that all RM values would be explored by both sets. The dataset S3, S9, SA05, SA1, SA3 and T6 were used as external dataset to assess the robustness of the model. The parity plots are reported in Fig. 1 and compare the RM values calculated by the model with the ones measured by the reference method, the KF titration. In the upper part of the graph, on the left, results obtained using the linear model are reported (graph a) and on the right, those obtained with the neural network (graph b). Obviously, the best situation is that all points lie on the bisector, but small deviations are permitted for technical limitations of the analytical technique and issues with the models. In this case, since S6 dataset was included in the calibration step, the trend fitted perfectly the points as shown in Fig. 1, graphs a and b. Also, the graphs of the absolute error as a function of the RM value measured by KF were reported in Fig. 1, graphs c and d. The absolute error was calculated as the absolute difference between the calculated and the measured value of RM (in %). The red line at 0.5% was assumed as the limit to assure the applicability of the model. In Fig. 1, graphs c and d, it is clearly visible that both the developed models are characterized by a very good accuracy, with 100% of the points lying below the acceptance threshold (red line).

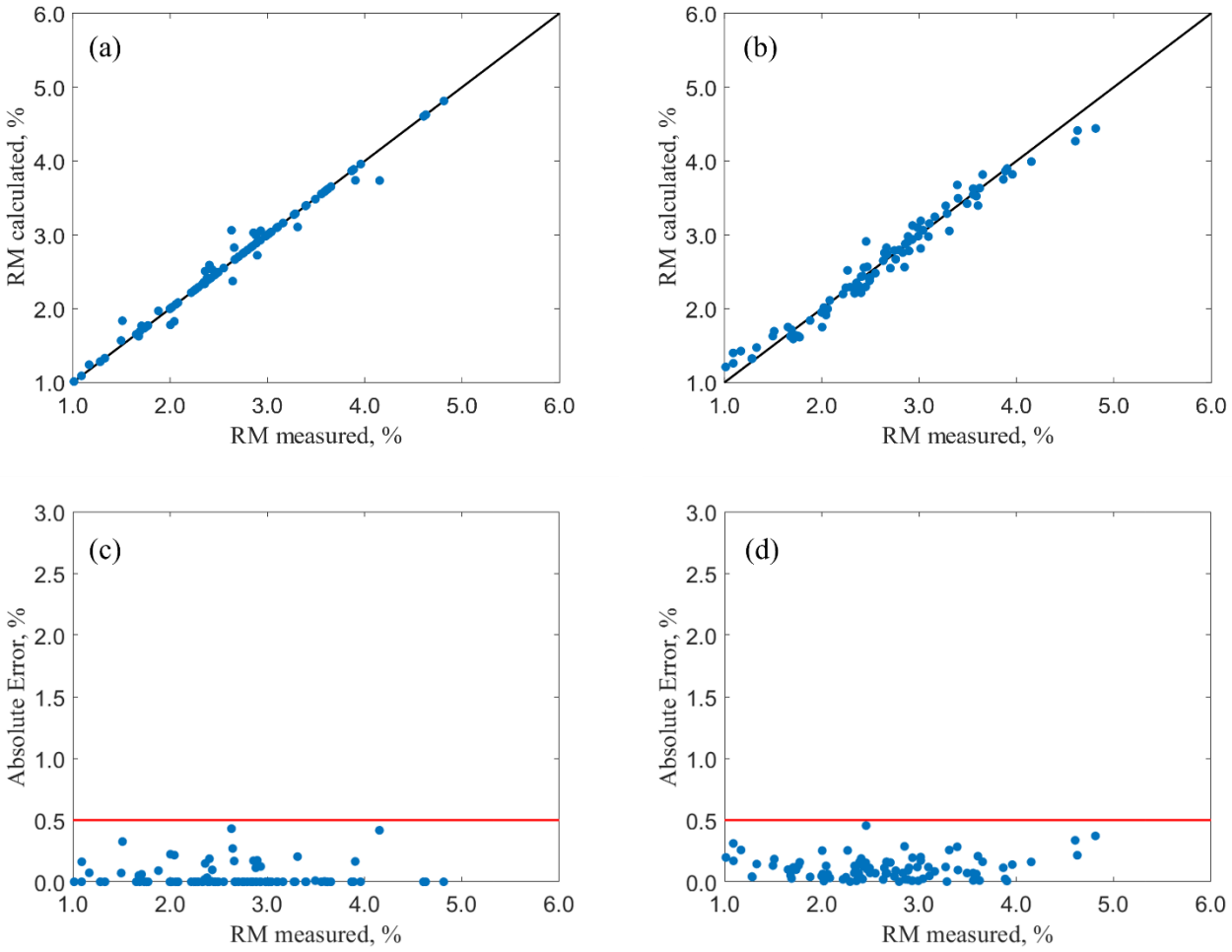

Figure 1: Parity diagrams (a, b), comparing the RM measured (%) and the RM calculated (%), and absolute error plots (c, d), obtained using the linear regression (a, c) and the neural network (b, d) models for dataset S6 for SR.

According to the data reported in Table 2, the neural network resulted in lower RMSE values than the linear model for small wavelength ranges (SR). In particular, its better performance was confirmed by the RMSE values of the arginine-sucrose mixture: for SA05 the neural network showed an RMSE value of 0.554 against 0.873 for the linear model; for SA1 the RMSE values were respectively 0.351 and 0.707. The parity diagrams and the absolute error plots were reported in the Supplementary Information in Fig. S3 and Fig. S4. For the reasons explained in the Materials and Methods section, a RMSE value not greater than 0.5 was acceptable. So, the neural network demonstrated a very good accuracy, comparable to the intrinsic error of the analytical method. Instead, both algorithms, neural network and linear model, failed in the prediction of the SA3 dataset, characterized by the highest percentage of arginine in solution (50% of the total solid fraction), with RMSE values respectively equal to 0.572 and 0.519. The spectra of the sucrose solution, the product used in the calibration of the model, is, in fact, very different from the one of sucrose-arginine mixture at high percentage. The plots of the different spectra are reported in Fig. S2 of the Supplementary

Information. The presence of arginine led to the disappearance of the peak specific of water at 5150 cm^{-1}. Since the model was analyzed in the SR wavelength range, specific for the highest peak of water, these results were considered reasonable. The bad performances of the models were also confirmed by the absolute error plots shown in Fig. 2, graphs c and d, where almost half of the points are above the acceptance threshold (red line).

Neural network described better also the dataset characterized by a different percentage of sucrose, S3 and S9 (for SR). In the first case the RMSE value calculated was 0.446 against 0.519 of the linear model. The parity diagrams and the absolute error plots were reported in Fig. S5 of the Supplementary Information. The largest difference was found for the prediction of the dataset with the highest percentage of sucrose, S9, resulting in RMSE values respectively equal to 0.305 and 0.774. The spectra of S6, S3 and S9 were similar, as reported in Fig. S1 of the Supplementary Information. However, some small differences, due to the different percentages of sucrose in water, led the linear model to give a worse prediction than the neural network. These findings were confirmed by the absolute error plots reported in Fig. 2, graphs a and b. Here, in the case of neural network only the 9% of the samples exceeded the acceptance threshold (red line); while, in the case of linear regression model 40% of the samples were above the acceptance threshold, index of the poor predictivity of the linear model.

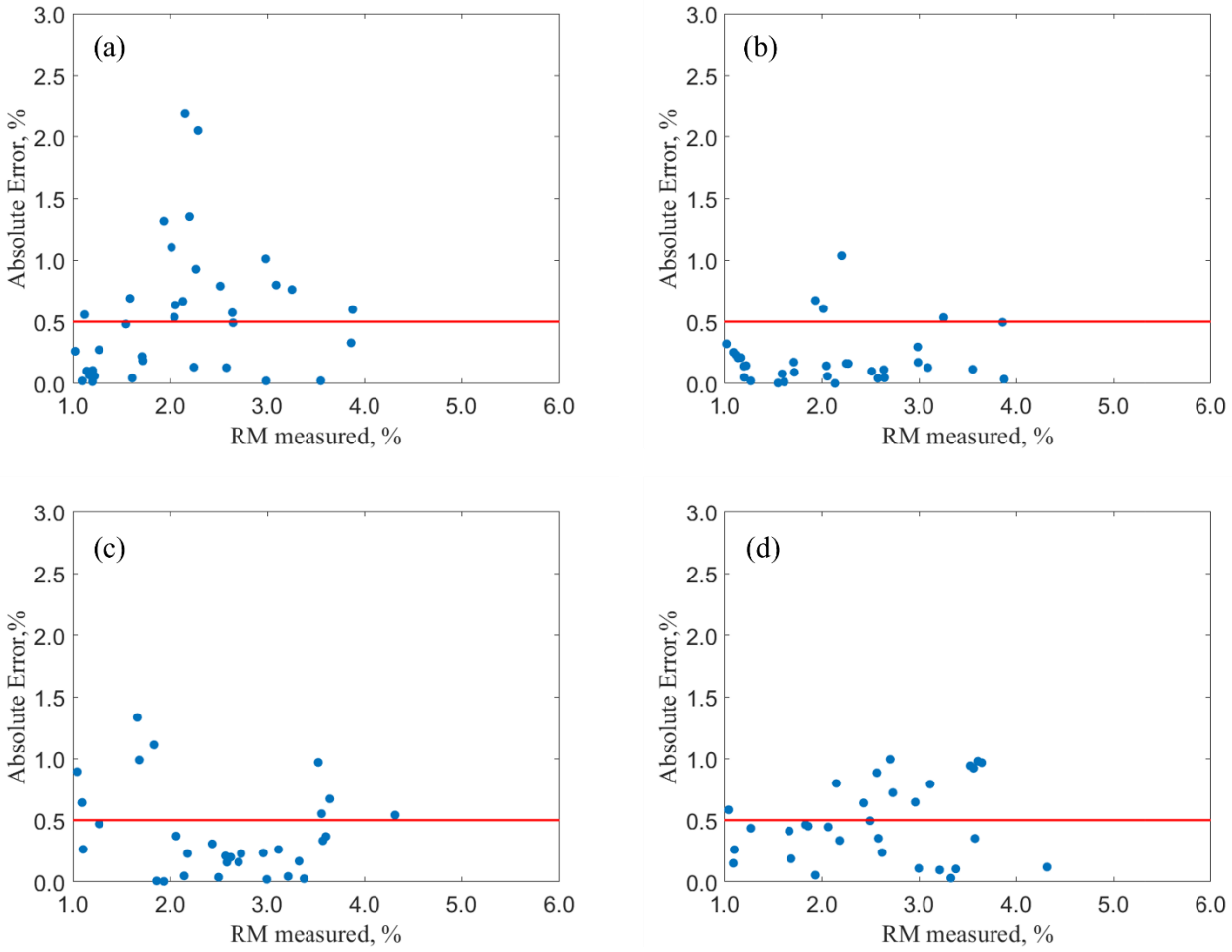

Figure 2: Comparison between the absolute error plots obtained for S9 dataset (a, b) and for SA3 dataset (c, d) in case of linear regression (a, c) and neural network (b, d) models developed with S6 database are used for SR.

The same considerations were done for the trehalose formulation, T6 (SR). The shape of the spectrum is similar to the one of S6 dataset. Therefore, the prediction turned out in accurate values of RMSE (0.199) for the neural network. A worse prediction, but still acceptable, was obtained for the linear regression, with a RMSE value equal to 0.334. The parity diagrams and the absolute error plots are reported in the Supplementary Information in Fig. S6. It has to be highlighted that the RMSE value found was slightly equal to the one obtained for S6 dataset (equal to 0.172), involved in the calibration step. This means that the model, which was calibrated only with sucrose as excipient, can perfectly predict the RM value of a formulation containing a different excipient, such as trehalose. This was made possible by the application of neural networks, which are nonlinear algorithms that can handle even more complex and challenging problems, and by focusing in the region specific of the peak of water.

The application of the model to a wider range of wavelengths yielded in worst performances for the neural network, as shown in Table 2. Obviously, the prediction of dataset S6, also included in the

calibration step, resulted in good accuracy. For all the other datasets, the RMSE values found were above the acceptable error given by the experimental method. This was considered as legitimate, since by enlarging the range of wavelengths, more peaks specific of the formulations were included.

3.2 *Ad-Hoc Models*

The poor results obtained with the Model S6 for the prediction of SA3 dataset suggested to develop an *ad hoc* model for the sucrose-arginine mixtures. The model was developed by including a certain percentage of SA05, SA1 and SA3 datasets in the calibration step, so that all the three different formulations containing the arginine were involved in the training phase of the model. Only the small range was analyzed. Also in this case, 70% of the dataset was used as training set (62 samples) and the remaining 30% as test set (27 samples). The results are shown in Table 3. For the SA3 dataset the RMSE values were 0.227 (for linear model) and 0.146 (for neural network); for SA1 they were respectively 0.336 and 0.229 and for SA05 they were equal to 0.287 and 0.233. The performances of the linear model and the neural network appeared to be comparable. This was reasonable since only few samples were involved in the calibration step and they had pretty similar spectra. The parity diagrams and the absolute error plots are reported in the Supplementary Information in Fig. S7, S8 and S9.

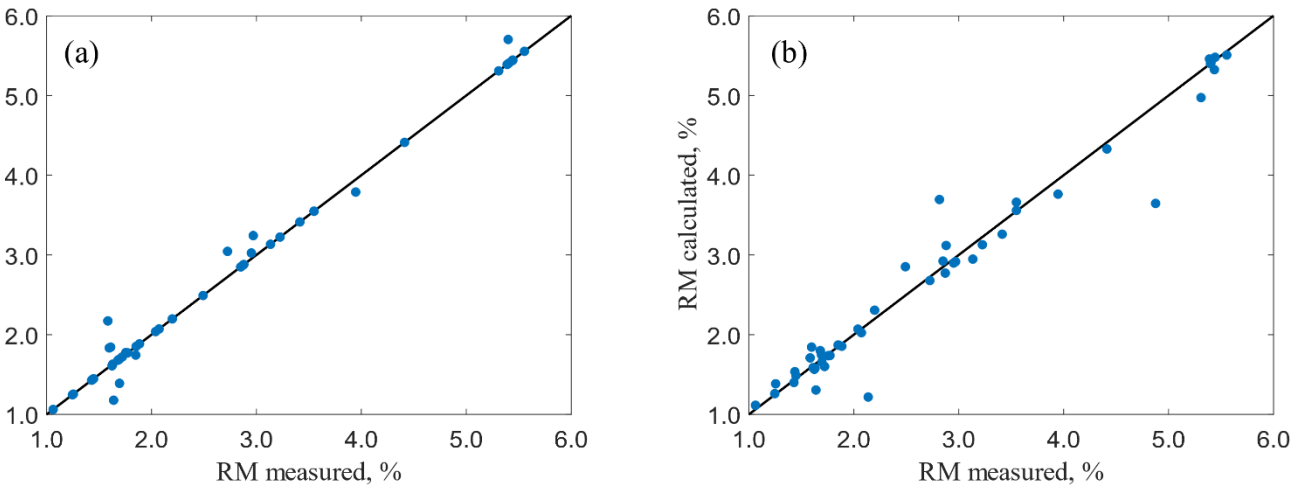

Figure 3: Parity diagrams comparing the RM measured (%) vs the RM calculated (%) using the linear regression (a) and the neural network (b) models developed for dataset T6 for SR.

Another *ad hoc* model was developed, for the trehalose containing mixtures. The training set was made of 32 samples (70%) and the test set of 13 samples (30%). The comparison between the two models is given in Table 3, with a RMSE value equal to 0.155 for the linear model and 0.128 for the neural network. As in the previous case, the performance of the two models was comparable for the

reasons discussed above. These findings were confirmed by the parity diagrams shown in Fig. 3a and 3b, indicating no or little difference between the two plots. The absolute error plots were reported in Fig. S10 of the Supplementary Information.

3.3 *Global Model*

For the global model evaluation, the whole dataset made up of S6, S3, S9, SA05, SA1, SA3 and T6, was split into two independent calibration and validation sets. The calibration set was composed of 70% of the dataset (228 samples) and the validation set of the remaining 30% (97 samples).

Firstly, the focus was on the application of the global model to the small range wavelengths. Obviously, the performance of the Global Model S6 dataset was worse than the first model, Model S6, since a lower percentage of that dataset was included into the calibration set to reach a homogeneity between all the formulations. The term homogeneity refers to having the same percentage of each formulation in the training set. In fact, the RMSE value in this case was 0.213 with the neural network and 0.358 with the linear model, higher than the previous model (respectively 0.172 and 0.122). The performance of dataset S3 globally improved with lower values of RMSE in both cases, respectively equal to 0.295 (for neural network) and 0.493 (for linear model). For S9 dataset, the situation remains the same for the neural network, while an improvement was observed for the linear model with a RMSE value of 0.444. These findings were expected, since both datasets were involved in the training set to build the model. However, the differences in the performance of the two models were not so marked. Therefore, it can be concluded that it is not necessary to build a global model to predict the RM value of S3 and S9, but the model S6 is sufficient to obtain a very good accuracy (comparable with the one obtained by the global model). Another evidence of the good performance of the model was displayed in the parity diagrams shown in Fig. 4 and Fig. 5. In Fig. 4, graphs a and b, the observations were more spread out than in the previous case, but they had similar distributions among the two models, an index of similar performances. The same observation can be made for the T6 dataset, with RMSE values equal to 0.181 (neural network) and 0.325 (linear model), values comparable with the ones obtained with the Model S6. The plots are reported in the Supplementary Information in Fig. S11, S12 and S13.

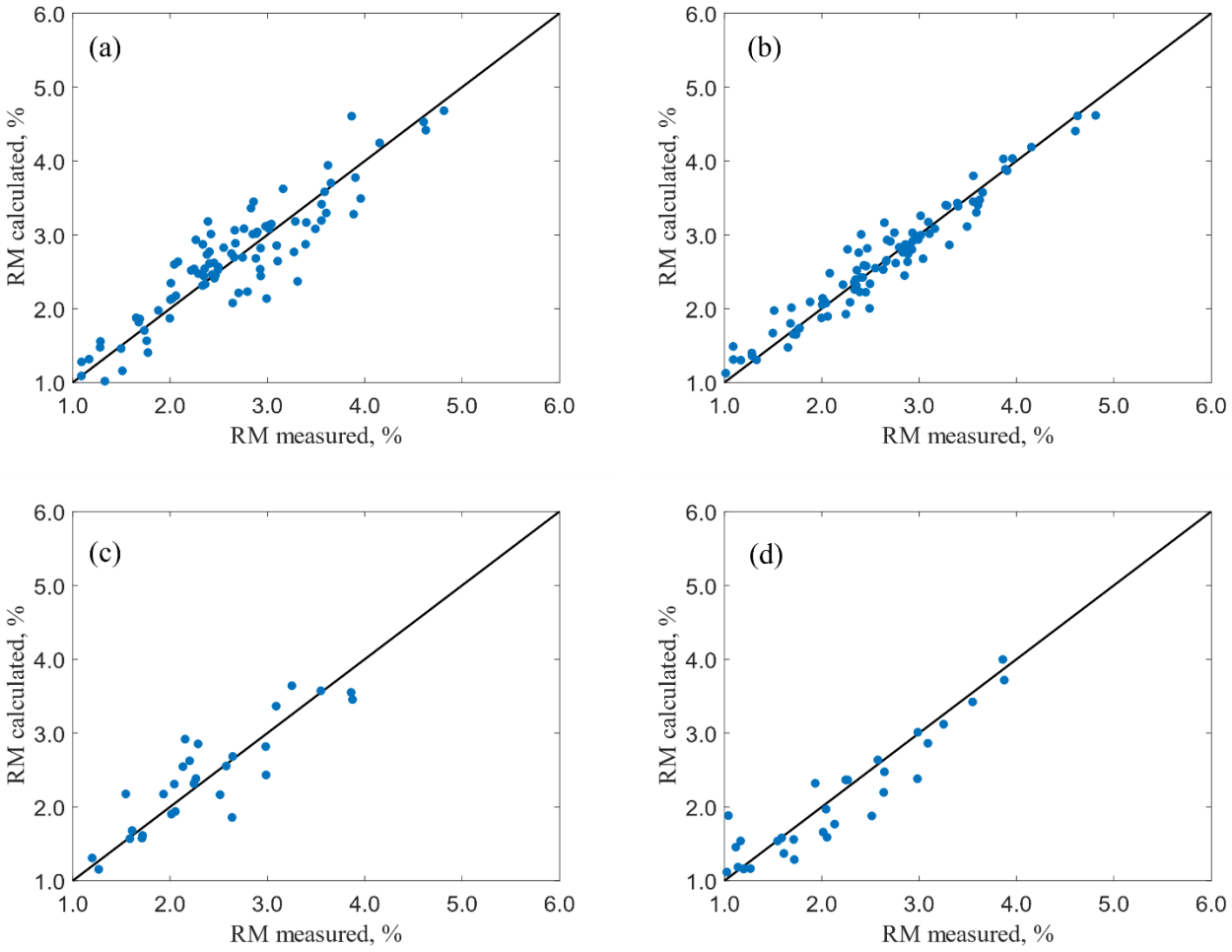

Figure 4: Parity diagrams comparing the RM measured (%) and the RM calculated (%) for dataset S6 (a, b) and for dataset S9 (c, d), obtained using the linear regression (a, c) and the neural network (b, d) global models for SR.

The situation appeared to be different for the sucrose-arginine mixtures. In this case, the performance was improved with respect to Model S6. In fact, arginine-based formulations were included in the training set, leading to an improved prediction for the three datasets (SA05, SA1 and SA3). In this way the specific peaks characteristic of arginine, which were different from the ones containing in the sucrose formulations were considered. The better performance was confirmed by the RMSE values reported in Table 2. It is evident that neural network always turned out in lower values of RMSE than the first model. For example, for SA3 the RMSE value is 0.173 against 0.572 for the previous model. So, it could be necessary to develop a model including samples of the new formulation (containing arginine in this case) in the training phase to take into account the very large variability in the spectra of the different formulations. Moreover, neural network was more suitable to deal with a huge dataset in input and to take into account their variability for the prediction of RM values of freeze-dried products. In fact, all the RMSE values were lower than the ones calculated by the linear regression model. The key reason is the non-linearity of the neural network. It is apparent

from Fig. 5, graphs a and b, where a more spread distribution is observed for the linear model with respect to the neural network. Also, the absolute error plots, in Fig. 5, graphs c and d, showed the best performance of the neural network, with no observation above the acceptance threshold.

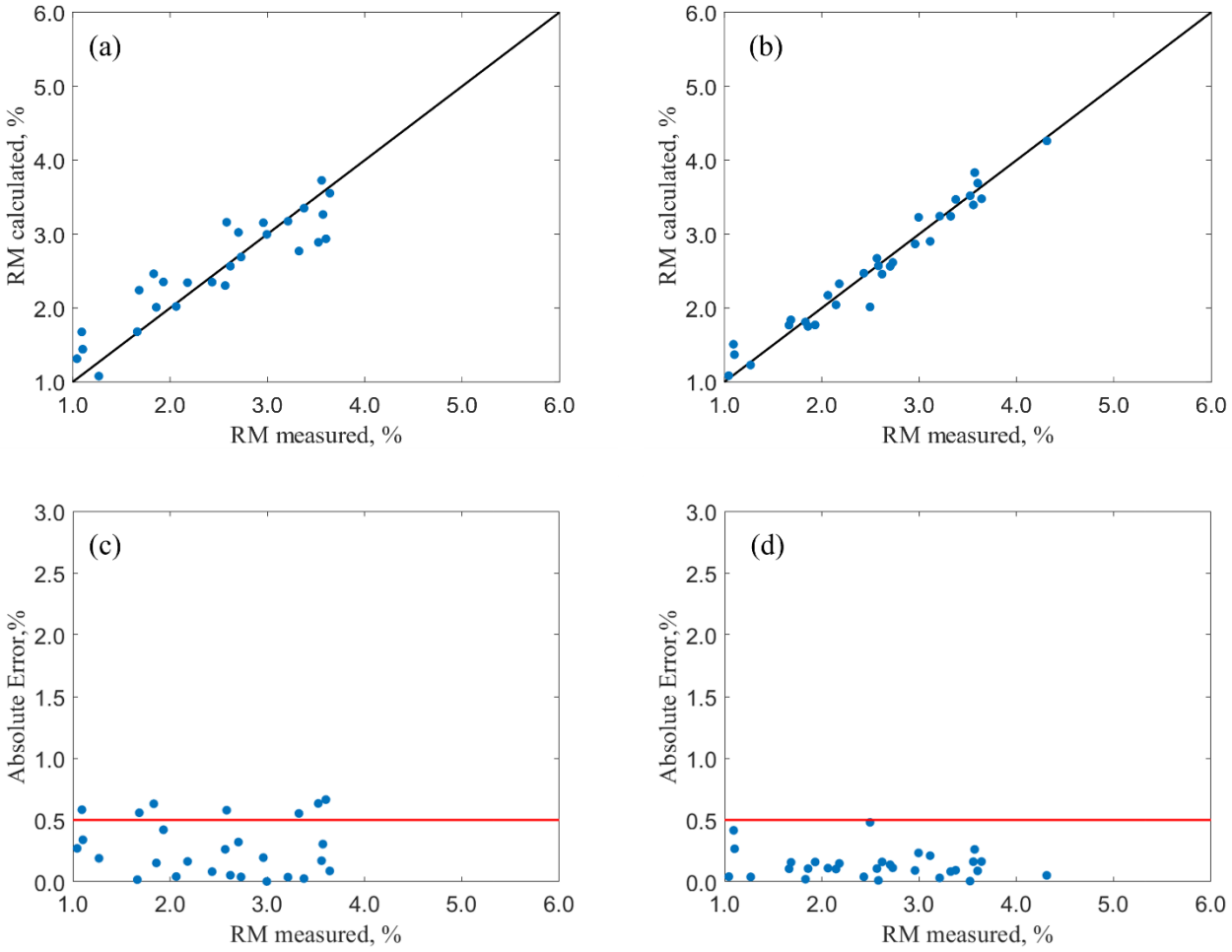

Figure 5: Parity diagrams (a, b), comparing the RM measured (%) and the RM calculated (%), and absolute error plots (c, d) obtained using the linear regression (a, c) and the neural network (b, d) global models for SR.

The global model was also applied to a wider range of wavelengths (WR) and the results obtained are reported in Table 2. In this case, the performance of the two models seemed to be quite comparable, with similar RMSE values between each other. All the values were on the order of the intrinsic error of the analytical method.

3.4 *Comparison of performances between the Neural Network and the PLS model*

The better performance of the presented models with respect to the PLS model (developed by Bobba *et al.* [30]) was observed both for the product-specific and the global model.

A comparison was made between the neural network and the PLS model for the Model S6. The higher difference was found for the dataset including the arginine product. In fact, for the SA05 dataset the

neural network gave an RMSE value of 0.554 against 0.65 for the PLS model; for SA1 the values were respectively 0.351 and 0.974 and for SA3 they were equal to 0.572 (NN) and 2.412 (PLS). These results allowed to emphasize again the improved ability of the neural network to deal with nonlinear problems, such as the prediction of a product not included in the calibration set. In this case, the tested product was arginine (an amino acid), which was characterized by specific features very different from those of sucrose. Also, for the dataset with a different percentage of sucrose, an improved prediction was reached. In fact, for S3 dataset the RMSE value calculated by the neural network was 0.446 against 0.562 calculated by the PLS model. For S9 dataset the RMSE values were respectively 0.305 (NN) and 0.741 (PLS). The same consideration could be done for the T6 dataset (0. 199 for NN and 0.578 for PLS), thus pointing out the higher accuracy of the neural network method based.
A comparison for the global model was done. Also, in this case, the results demonstrated the better accuracy of the neural network with respect to the PLS model. In fact, some examples are here reported: for dataset S3 the values were 0.295 (NN) and 0.578 (PLS); for S9 they were 0.303 (NN) and 0.555 (PLS) and for T6 they were respectively 0.181 (NN) and 0.703 (PLS).

3.5 *Effect of training set size on the performances*

The effect of the training set size on model performance was tested by calculating the RMSE values in each case. The results are summarized in Table 4 for both models used.
The RMSE values were calculated for both models by varying the size of the training set, expressed as percentage of the available data, from 70% to 40%, allowing also to point out problems of overfitting or underfitting. The results obtained by processing data with Model S6 are reported in Table 4. The highest values of RMSE were obtained for the sucrose-arginine mixture. In particular, the higher was the percentage of arginine in solution and the higher was the value of RMSE. Globally, the RMSE values obtained with the linear model were higher than the ones obtained with the neural network for all the percentages used. A close inspection of Table 4 indicates that the RMSE values remained slightly constant at the different sizes of training set analyzed for the datasets S3, S9 and T6. This is a really good result, since it allows a reduction in the experimental effort for developing the model. A moderate increase has been observed for the dataset SA05, while a huge increase is observed for the dataset with higher percentage of arginine, so SA1 and SA3. In these cases, in fact, the RMSE values ranges from 0.6 up to 2 for very low percentages of training set. The same situation can be observed for the neural network, but all the curves are shifted down, index of its better performance.
A different situation was observed for the global model in Table 4. For the linear regression model, the performances were bad at 40% of training set with very high values of RMSE ranging from 0.7

up to 1.4. The remaining trend was approximately constant, with minimal changes for the 60% of training set. For the neural network, it was clearly visible that the performances were better with very low values of RMSE whatever the percent of training set. This Table could be helpful in reaching a compromise between the accuracy of the model and the experimental effort done to develop it. Also, a confirmation of the results could be obtained by considering other statistical parameters, such as the Root Mean Square Error of Calibration (RMSEC) and the one of Cross-Validation (RMSECV). In fact, as an example, by focusing on the neural network-based model developed using the 70% of dataset as calibration set, in the case of the product-specific model for S6, their values were respectively equal to 0.131 and 0.094; while in the case of the global model, values equal to 0.105 and 0.177 have been obtained.

Table 4: Effect of training set size on model performance. In the first row the type of model is specified. In the second row the percentage of available dataset (from 40% up to 60%) used for training purposes is reported. All the values in the Table are the RMSE values corresponding at each dataset for each value of the size of the training set. The results are reported for SR.

	Model S6 LR			**Model S6 NN**			**Global Model LR**			**Global Model NN**		
Dataset	*40%*	*50%*	*60%*	*40%*	*50%*	*60%*	*40%*	*50%*	*60%*	*40%*	*50%*	*60%*
S3	0.427	0.375	0.392	0.408	0.352	0.283	0.757	0.598	0.681	0.422	0.294	0.334
S9	0.574	0.558	0.412	0.252	0.333	0.189	0.691	0.689	0.468	0.286	0.212	0.304
SA05	0.877	0.415	0.909	0.593	0.419	0.644	0.969	0.559	0.669	0.426	0.323	0.330
SA1	1.998	0.544	0.993	0.506	0.349	1.061	1.352	0.495	0.631	0.384	0.354	0.401
SA3	0.599	1.880	1.169	0.952	0.868	1.219	1.229	0.723	0.430	0.632	0.397	0.400
T6	0.494	0.479	0.479	0.491	0.291	0.358	1.363	0.465	0.473	0.660	0.487	0.578

4 Conclusions

In summary, NIR spectroscopy was coupled with machine learning techniques to quantify the residual moisture content in freeze-dried products. The first goal of the present work was the development of a model able to estimate the residual moisture in a certain reference product, the S6 dataset. Then, the robustness of this model was tested using the other different products as external validation dataset. Two different models were developed: a linear regression model and a neural network. This study clearly demonstrates that the coupling of NIR spectroscopy with chemometric techniques is a powerful tool for the quantitative prediction of RM values as an alternative to KF titration in the context of process development.

By comparing the developed models, the neural network turned out in more accurate and reliable

performance than the conventional linear models. Its better performance was assessed both for the prediction of products non-involved in the calibration step and for dealing with large dataset, as in the case of global model. So, the robustness of the neural network was demonstrated with RMSE values lower than the intrinsic error of the analytical method (KF). On the contrary, the performances of linear model and neural network were comparable for the two *ad hoc* models. Also, the results obtained with both models were compared with the PLS model developed by Bobba *et al.* [30] As highlighted from the RMSE values, the performance of the neural network was remarkably better. It was assessed that the introduction of a new component, like arginine, that gives a different contribution in the analyzed spectral region, required for the development of a global model, while the product-specific model for S6 revealed accurate in the prediction of dataset containing sucrose at different percentage and trehalose (having a similar spectra).

Obviously, machine learning tools require a higher computational cost than linear models. Hence, based on the needs, a compromise between computational cost and accuracy of the method is needed. However, in pharmaceutical processes, high accuracy is mandatory for quality control of the products. Therefore, a suitable machine learning algorithm could be more robust and powerful for this purpose.

Acknowledgment

The authors thankfully acknowledge Merck Serono S.p.A for the contribution and financial support. The previous work of Nunzio Zinfollino and Serena Bobba is fully acknowledged. The support in the experimental activities of Adamo Sulpizi (Researcher), Caterina Sapienza (Associate Researcher), Daniele Mari (Senior Laboratory Technician), and Michele Dimattia (Junior Researcher), Global Pharmaceutical Development Department, Merck Serono S.p.A, Guidonia Montecelio, Roma, is also gratefully appreciated.

References

[1] Fissore D. *Freeze Drying of Pharmaceuticals*. Encyclopedia of Pharmaceutical Science and Technology. Taylor and Francis: New York, 2013; 1723-1737. http://dx.doi.org/10.1081/E-EPT4-120050278

[2] Oddone I, Pisano R, Bullich R, & Stewart P. *Vacuum-Induced Nucleation as a Method for Freeze-Drying Cycle Optimization*. Ind. Eng. Chem. 2014; 53:18236-18244. https://doi.org/10.1021/ie502420f.

[3] De Beer T, Burggraeve A, Fonteyne M, Saerens S, Remon JP, & Vervaet C. *Near Infrared and Raman Spectroscopy for the In-Process Monitoring of Pharmaceutical Production Processes.* Int. J. Pharm. 2011a; 417: 32–47. https://doi.org/10.1016/j.ijpharm.2010.12.012.

[4] Connors KA. *The Karl Fischer Titration of Water*. Drug Dev. Ind. Pharm. 1988; 14: 1891-1903. https://doi.org/10.3109/03639048809151996.

[5] *Water: Semi-Micro Determination.* European Pharmacopoeia 8th edition Volume 1, chapter 2.5.12, 2013.

[6] *Pharmaceutical cGMPs for the 21st Century: A Risk-Based Approach.* [Online]. Available: https://www.fda.gov/media/77391. (last access: July 2022).

[7] *PAT – A Framework for Innovative Pharmaceutical Development, Manufacturing, and Quality Assurance.* [Online]. Available: https://www.fda.gov/regulatory-information/search-fda-guidance-documents/pat-framework-innovative-pharmaceutical-development-manufacturing-and-quality-assurance.

[8] Luypaert J, Massart D. L., & Vander Heyden Y. *Near-Infrared Spectroscopy Applications in Pharmaceutical Analysis.* Talanta 2007; 72: 865-883. http://dx.doi.org/10.1016/j.talanta.2006.12.023.

[9] Lakeh M. A., Karimvand S. K., Khoshayand M. R., & Abdollahi H. *Analysis of Residual Moisture in a Freeze-Dried Sample Drug Using a Multivariate Fitting Regression Model*. Microchem. J. 2020; 107: 2411-2502. http://dx.doi.org/10.1016/j.microc.2019.104516.

[10] Clua-Palau G, Jo E, Nikolic S, Coello J, & Maspoch S. *Finding a Reliable Limit of Detection in the NIR Determination of Residual Moisture in a Freeze-Dried Drug Product.* J. Pharm. Biomed. Anal. 2020; 183: 113-163. http://dx.doi.org/10.1016/j.jpba.2020.113163.

[11] Derksen M. W. J., van de Oetelaar P. J. M., & Maris F. A. *The Use of Near-Infrared Spectroscopy in the Efficient Prediction of a Specification for the Residual Moisture Content of a Freeze-Dried Product.* J. Pharm. Biomed. Anal. 1998; 17: 473-480. http://dx.doi.org/10.1016/s0731-7085(97)00216-1.

[12] Clavaud M, Lema-Martinez C, Roggo Y, Bigalke M, Guillemain A, Hubert P. Ziemons E., &

Allmendinger A. *Near-Infrared Spectroscopy to Determine Residual Moisture in Freeze-Dried Products: Model Generation by Statistical Design of Experiments*. J. Pharm. Sci. 2020, 109: 719-729. http://dx.doi.org/10.1016/j.xphs.2019.08.028.

[13] Mainali D, Li J, Yehl P, & Chetwyn N. *Development of a Comprehensive Near Infrared Spectroscopy Calibration Model for Rapid Measurements of Moisture Content in Multiple Pharmaceutical Products.* J. Pharm. Biomed. Anal. 2014; 95: 169-175. http://dx.doi.org/10.1016/j.jpba.2014.03.001.

[14] De Beer T, Hansen L, Vander Heyden Y, Pieters S, Varvaet C, Remon J. P., Montenez J. P., & Daoussi R. *Near-Infrared Spectroscopy Evaluation of Lyophilized Viral Vaccine Formulations.* Biotechnol. Prog.2013; 29: 1573-1586. https://lib.ugent.be/catalog/pug01:4270700.

[15] Grohganz H, Gildemyn D, Skibsted E, Flink J.M., & Rantanen J. *Towards a Robust Water Content Determination of Freeze-Dried Samples by Near-Infrared Spectroscopy.* Anal. Chim. Acta 2010; 676: 34-40. http://dx.doi.org/10.1016/j.aca.2010.07.035.

[16] Clavaud M, Roggo Y, Degardin K, Sacrè P.Y., Hubert P, & Ziemons E. *Global Regression Model for Moisture Content Determination Using Near-Infrared Spectroscopy.* Eur. J. Pharm. Biopharm. 2017; 119: 343-352. http://dx.doi.org/10.1016/j.ejpb.2017.07.007.

[17] Goodfellow I, Bengio Y, & Courville A. *Deep learning.* MIT Press 2016.

[18] https://it.mathworks.com/content/dam/mathworks/ebook/gated/machine-learning-ebook-all-chapters.pdf (last access: July 2022).

[19] Parastar H, van Kollenburg G, Weesepoel Y, van den Doel A, Buydens L, & Jansen J. *Integration of Handheld NIR and Machine Learning to "Measure & Monitor" Chicken Meat Authenticity.* Food Control 2020; 112: 107-149. https://doi.org/10.1016/j.foodcont.2020.107149.

[20] Coronel-Reyes J, Ramirez-Morales I, Fernandez-Blanco E, Rivero D, & Pazos A. *Determination of Egg Storage Time at Room Temperature Using a Low-Cost NIR Spectrometer and Machine Learning Techniques.* Comput. Electron. Agric. 2018; 145: 1-10. http://dx.doi.org/10.1016/j.compag.2017.12.030.

[21] Richter B, Rurik M, Gurk S, Kohlbacher O, & Fischer M. *Food Monitoring: Screening of the Geographical Origin of White Asparagus Using FT-NIR and Machine Learning.* Food Control 2019; 104: 318-325. https://doi.org/10.1016/j.foodcont.2019.04.032.

[22] Martins J.A., Guerra R, Pires R, Antunes M. D., Panagopoulos T, Brazio A, Afonso A. M., Silva L, Lucas M. R., & Cavaco A. M. *Spectranet-53: A Deep Residual Learning Architecture For Predicting Soluble Solids Content with VIS-NIR Spectroscopy*. Comput. Electron. Agric. 2022. https://doi.org/10.1016/j.compag.2022.106945.

[23] Zhao J, Tian G, Qiu Y, & Qu H. *Rapid Quantification of Active Pharmaceutical Ingredient*

for Sugar-Free Yangwei Granules in Commercial Production Using FT-NIR Spectroscopy Based on Machine Learning Techniques. Spectrochim. Acta A Mol. Biomol. Spectrosc. 2021; 245: 118-878. https://doi.org/10.1016/j.saa.2020.118878.

[24] Akbar R, Bashour H, Rawat P, Robert PA, Smorodina E, et al. *Progress and Challenges for the Machine Learning-Based Design of Fit-For Purpose Monoclonal Antibodies.* MAbs 2022; 14: 2008-2790. https://doi.org/10.1080/19420862.2021.2008790.

[25] Muller A.C & Guido S. *Introduction to Machine Learning with Python – A guide for data scientists.* O'Reilly, 2016.

[26] Venkatasubramanian V. *The Promise of Artificial Intelligence in Chemical Engineering: Is it here, finally?* AIChe J. 2018; 65: 466-478. https://doi.org/10.1002/aic.16489.

[27] Geron A. *Hands-on Machine Learning with Scikit-Learn, Keras & TensorFlow – Concepts, Tools and Techniques to Build Intelligent Systems.* O'Reilly, 2019.

[28] Albon C. *Machine Learning with Python Cookbook – Practical Solutions from Preprocessing to Deep learning.* O'Reilly, 2018.

[29] Marcato A, Boccardo G, & Marchisio D. *A computational workflow to study particle transport and filtration in porous media: Coupling CFD and deep learning*. Chem. Eng. J., 2021; 417: 128-936. https://doi.org/10.1016/j.cej.2021.128936.

[30] Bobba S, Zinfollino N, & Fissore D. *Application of Near-Infrared Spectroscopy to Statistical Control in Freeze-Drying Processes.* . Eur. J. Pharm. Biopharm. 2021; 168: 26-37. https://doi.org/10.1016/j.ejpb.2021.08.009.

[31] https://towardsdatascience.com/linear-regression-using-python-b136c91bf0a2 (last access: July 2022).

[32] https://www.ibm.com/cloud/learn/neural-networks (last access: July 2022).

[33] Mittal S, Pathak S, Yadav HD, & Upadhyayula S. *A Machine Learning Approach to Improve Ignition Properties of High-Ash Indian Coals by Solvent Extraction and Coal Blending.* Chem. Eng. J. 2020; 413: 27-385. http://dx.doi.org/10.1016/j.cej.2020.127385.

[34] Gavin H.P. *The Levenberg-Marquardt Algorithm for Nonlinear Least Squares Curve-Fitting Problems*. Department of Civil and Environmental Engineering Duke University, 2020.

[35] http://matlab.izmiran.ru/help/toolbox/nnet/trainlm.html (last access: July 2022).

[36] Transtrum M. K., Machta B. B., Brown K. S. et al. *Perspective: Sloppiness and Emergent Theories in Physics, Biology, and Beyond.* J. Chem. Phys. 2015; 143: 1-10. https://doi.org/10.1063/1.4923066

[37] Ward HW & Sistare FE. *On-line determination and control of the water content in a continuous conversion reactor using NIR spectroscopy.* Anal. Chim. Acta 2007; 595: 319-322.

https://doi.org/10.1016/j.aca.2007.03.020.

[38] Clua-Palau G, Jo E, Nikolic S, Coello J, & Maspoch S. *Finding a Reliable Limit of Detection in the NIR Determination of Residual Moisture in a Freeze-Dried Drug Product.* J. Pharm. Biomed. Anal. 2020; 183: 13-163. https://doi.org/10.1016/j.jpba.2020.113163.

[39] Grohganz H, Fonteyne M, Skibsted E, Falck T, Palmqvist B, & Rantanen J. *Role of Excipients in the Quantification of Water in Lyophilized Mixtures Using NIR Spectroscopy.* J. Pharm. Biomed. Anal. 2009; 49: 901–907. https://doi.org/10.1016/j.jpba.2009.01.021.

[40] Gerzon G, Sheng Y, & Kirkitadze M. *Process Analytical Technologies – Advances in Bioprocess Integration and Future Perspectives.* J. Pharm. Biomed. Anal. 2022; 207: 114-379. https://doi.org/10.1016/j.jpba.2021.114379.

[41] Rinnan A, Norgaard L, van den Berg F, Thygesen J, Bro R, & Engelsen S. B. *Data Pre-processing.* Infrared Spectroscopy for Food Quality Analysis and Control 2009; 29-50. https://doi.org/10.1016/B978-0-12-374136-3.00002-X.

[42] Amigo J.M. *Data Mining, Machine Learning, Deep Learning, Chemometrics.* Brazilian J. Anal. Chem. 2021; 8: 45-61. http://dx.doi.org/10.30744/brjac.2179-3425.AR-38-2021.

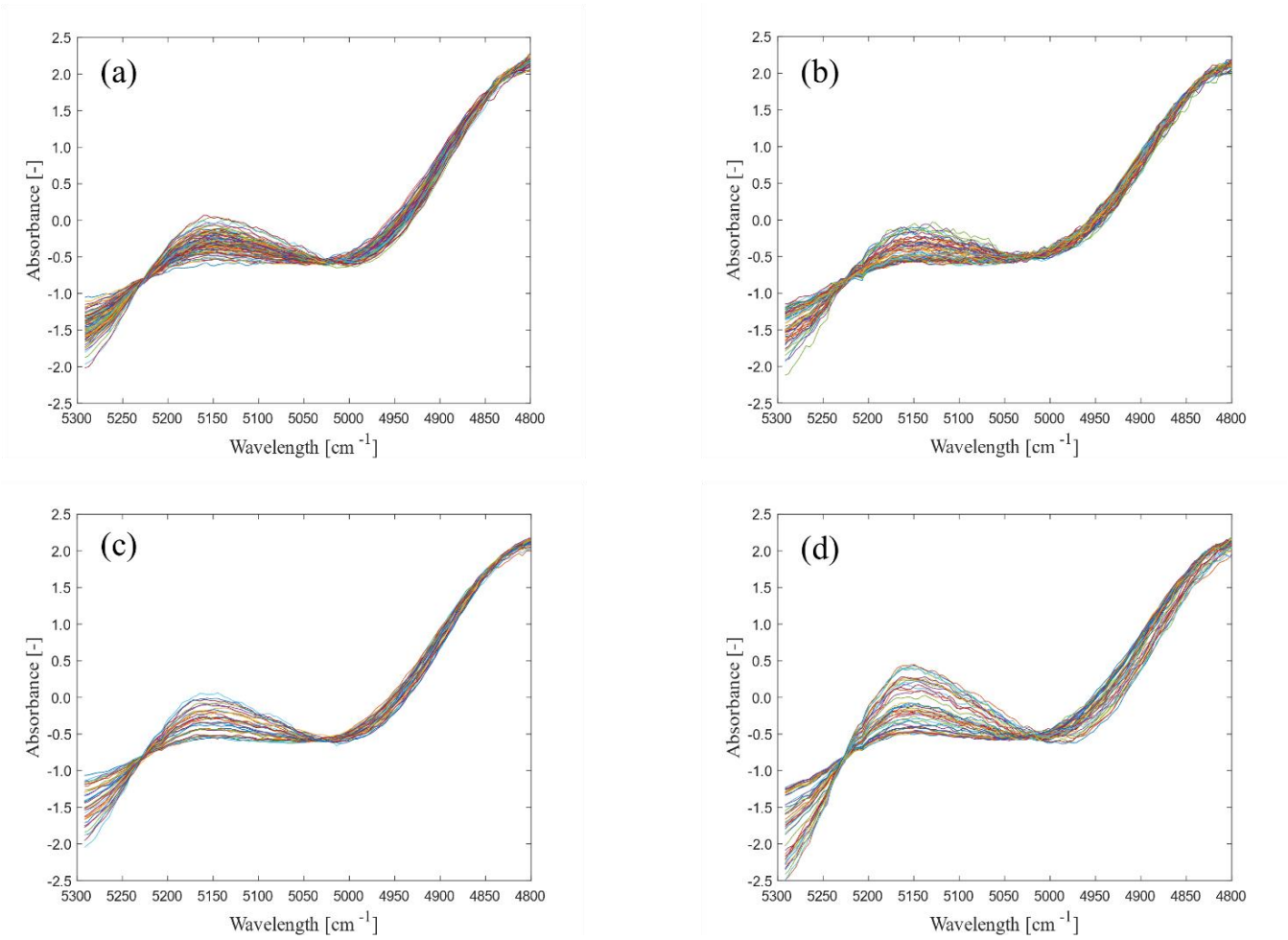

Figure S2: Spectra of different formulations after SNV pretreatment: (a) S6 dataset, (b) S3 dataset, (c) S9 dataset and (d) T6 dataset.

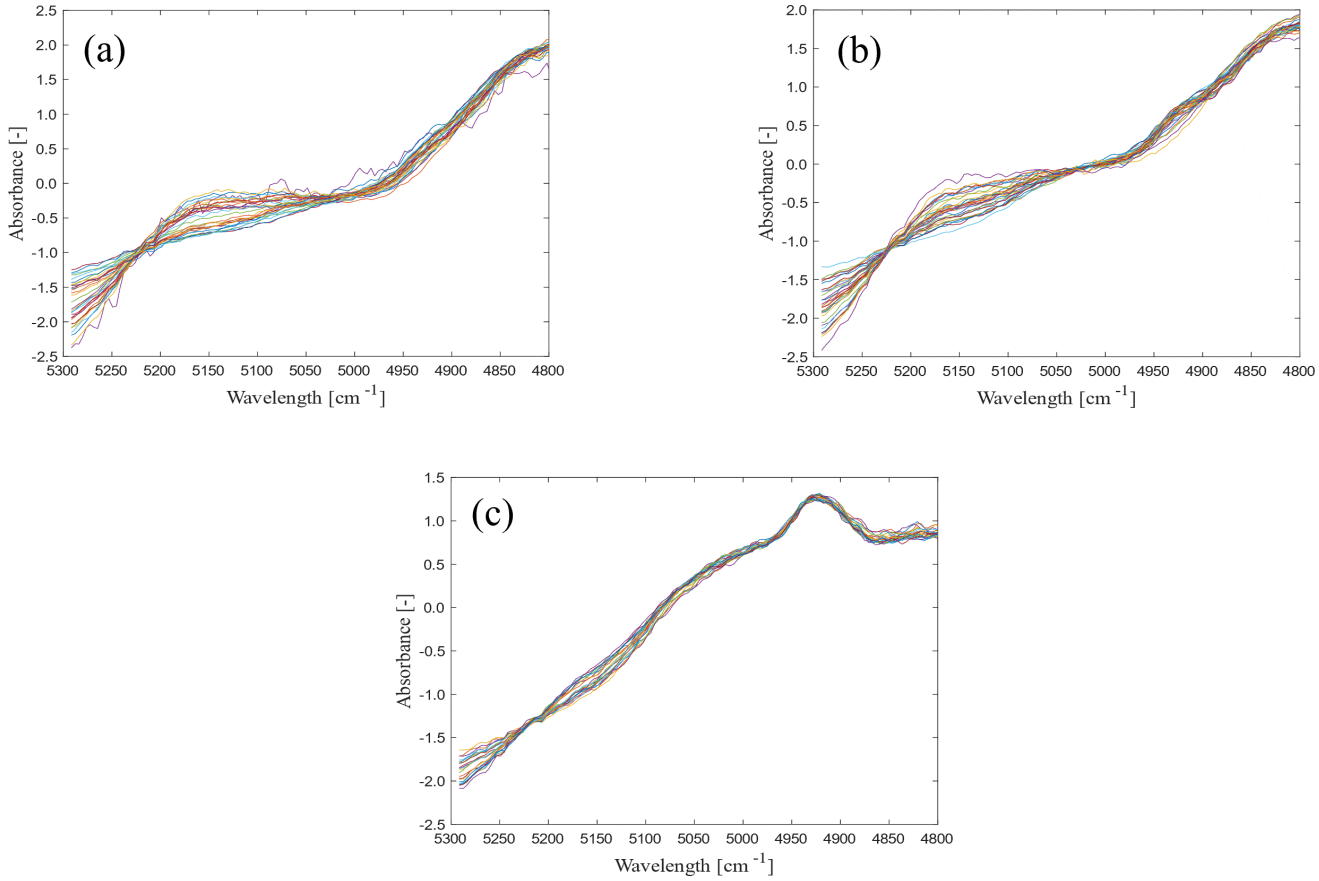

Figure S3: Spectra of the different formulations containing arginine after SNV pretreatment: (a) SA05 dataset, (b) SA1 dataset and (c) SA3 dataset.

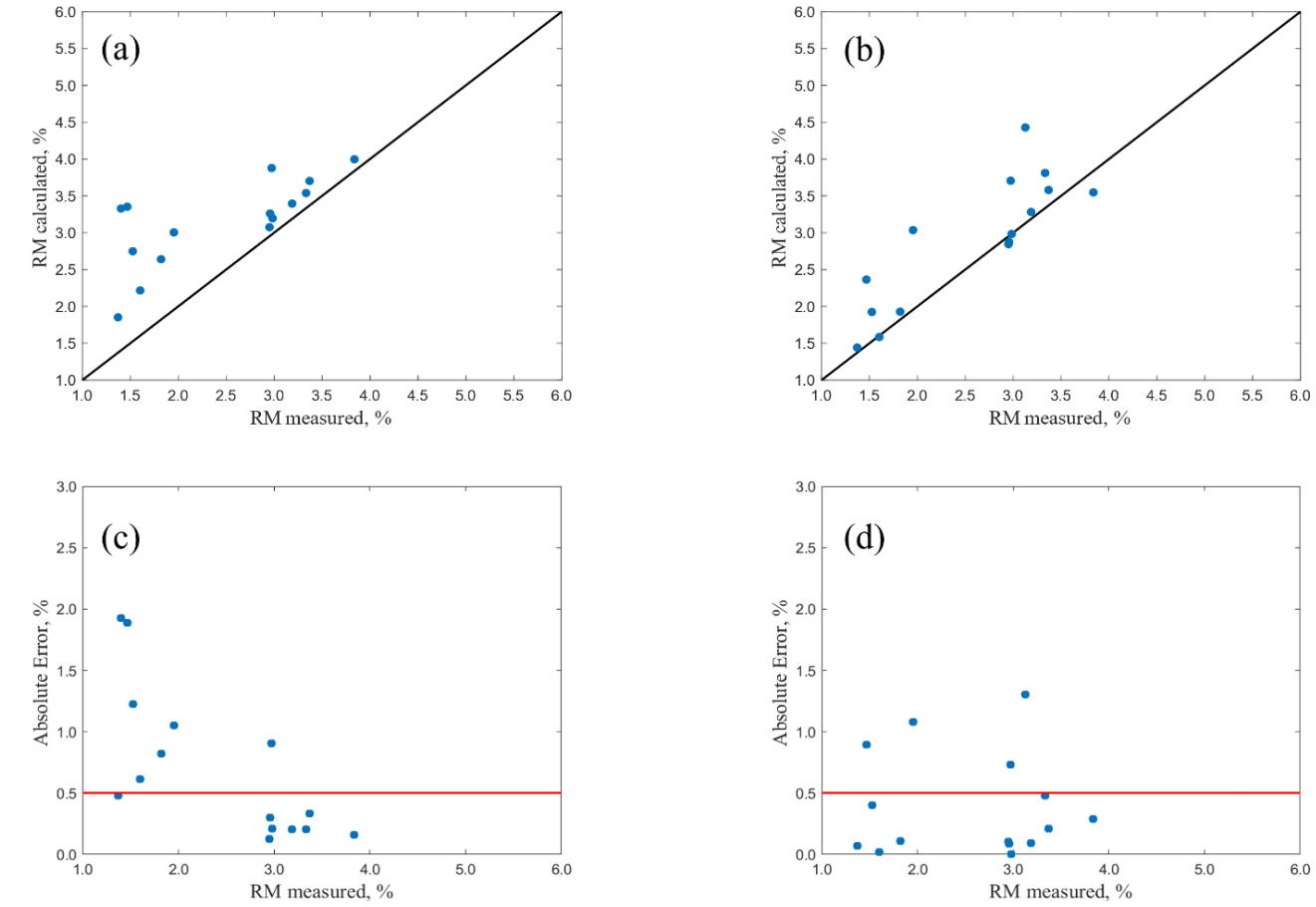

Figure S4: Parity diagrams (a, b) of dataset SA05 comparing the RM measured (%) and the RM calculated (%), and absolute error plots (c, d), obtained using the linear regression (a, c) and the neural network (b, d) models developed using dataset S6 for SR.

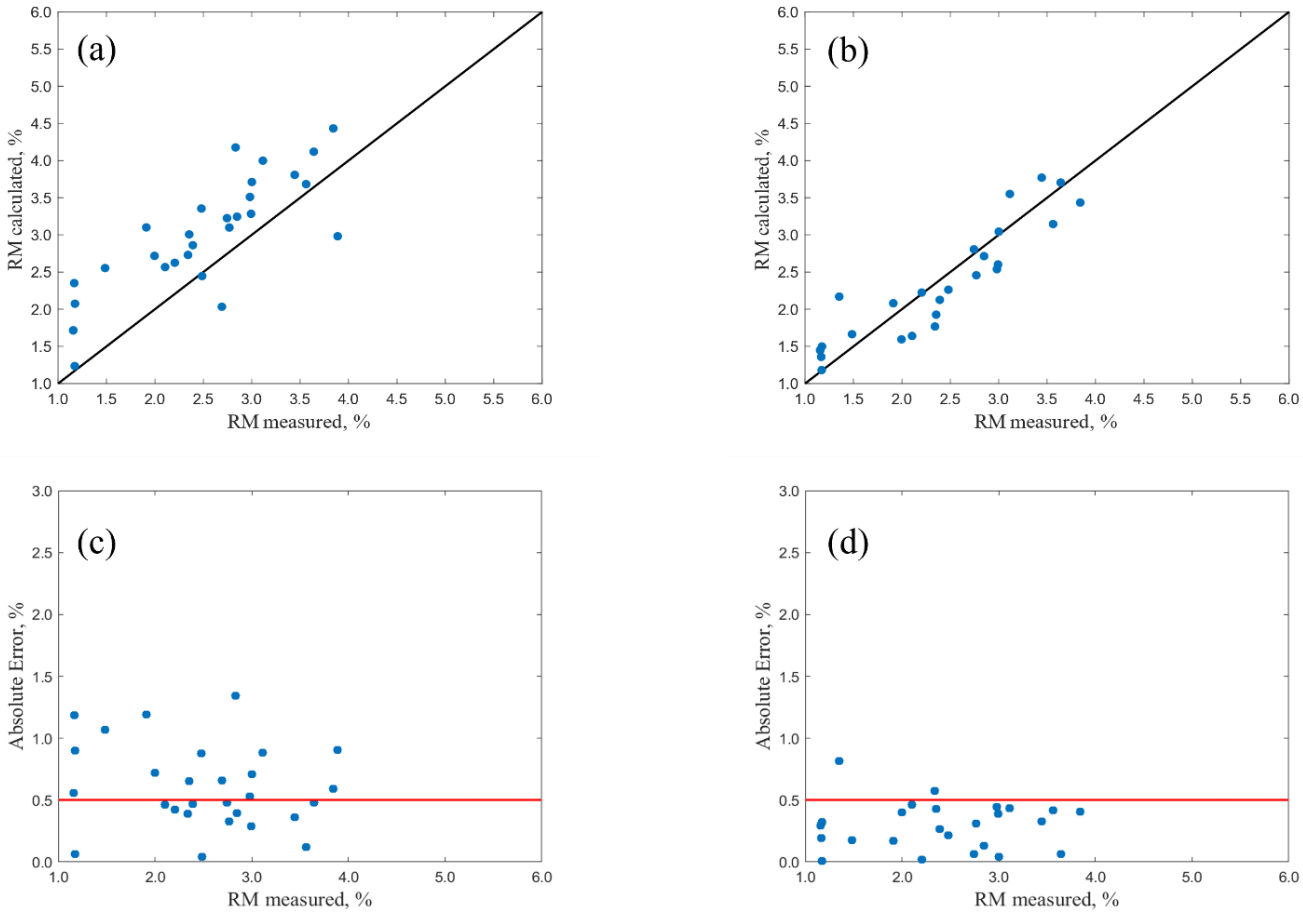

Figure S5: Parity diagrams (a, b) of dataset SA1 comparing the RM measured (%) and the RM calculated (%), and absolute error plots (c, d), obtained using the linear regression (a, c) and the neural network (b, d) models developed using dataset S6 for SR.

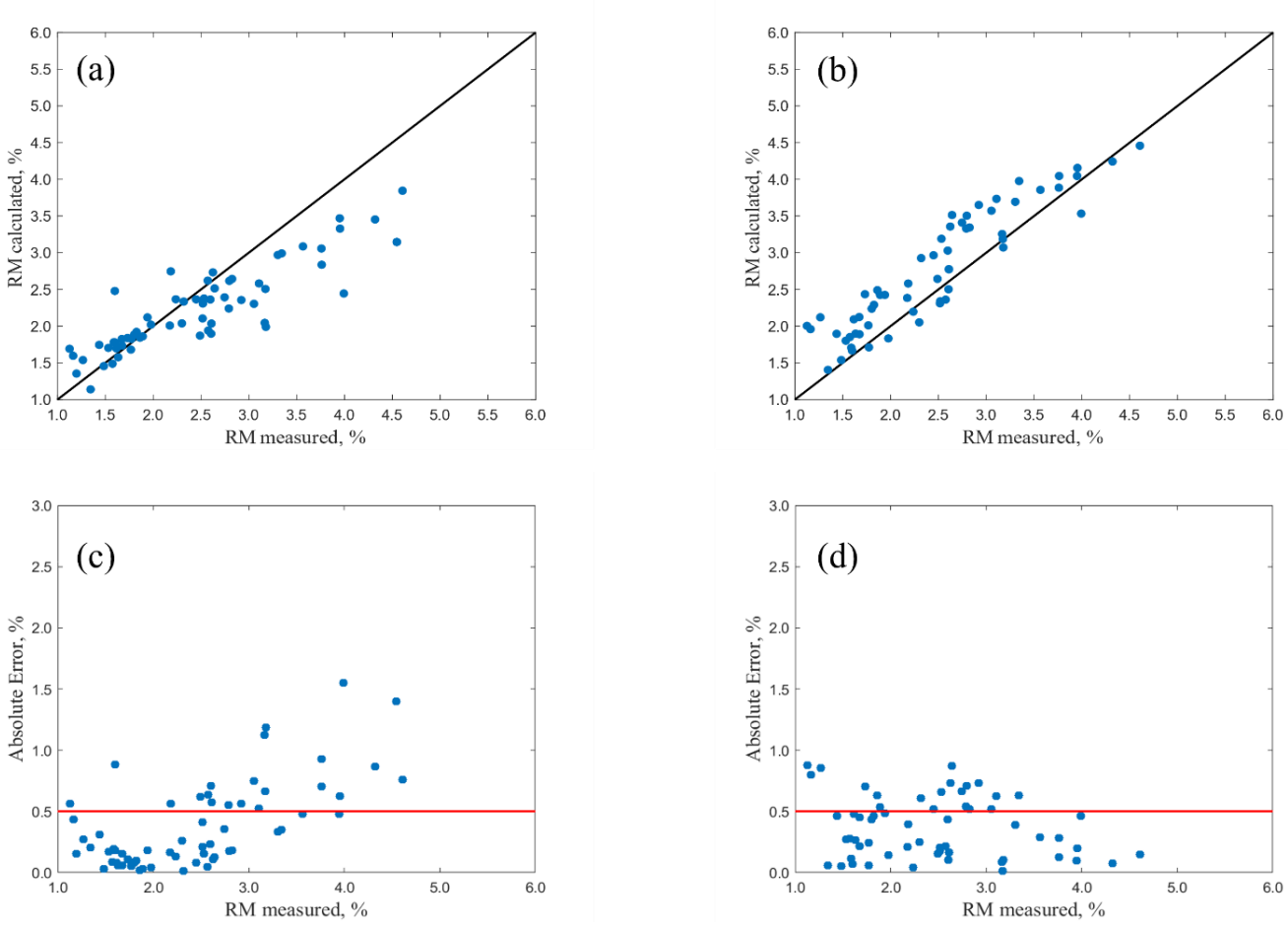

Figure S6: Parity diagrams (a, b) of dataset S3 comparing the RM measured (%) and the RM calculated (%), and absolute error plots (c, d), obtained using the linear regression (a, c) and the neural network (b, d) models developed using dataset S6 for SR.

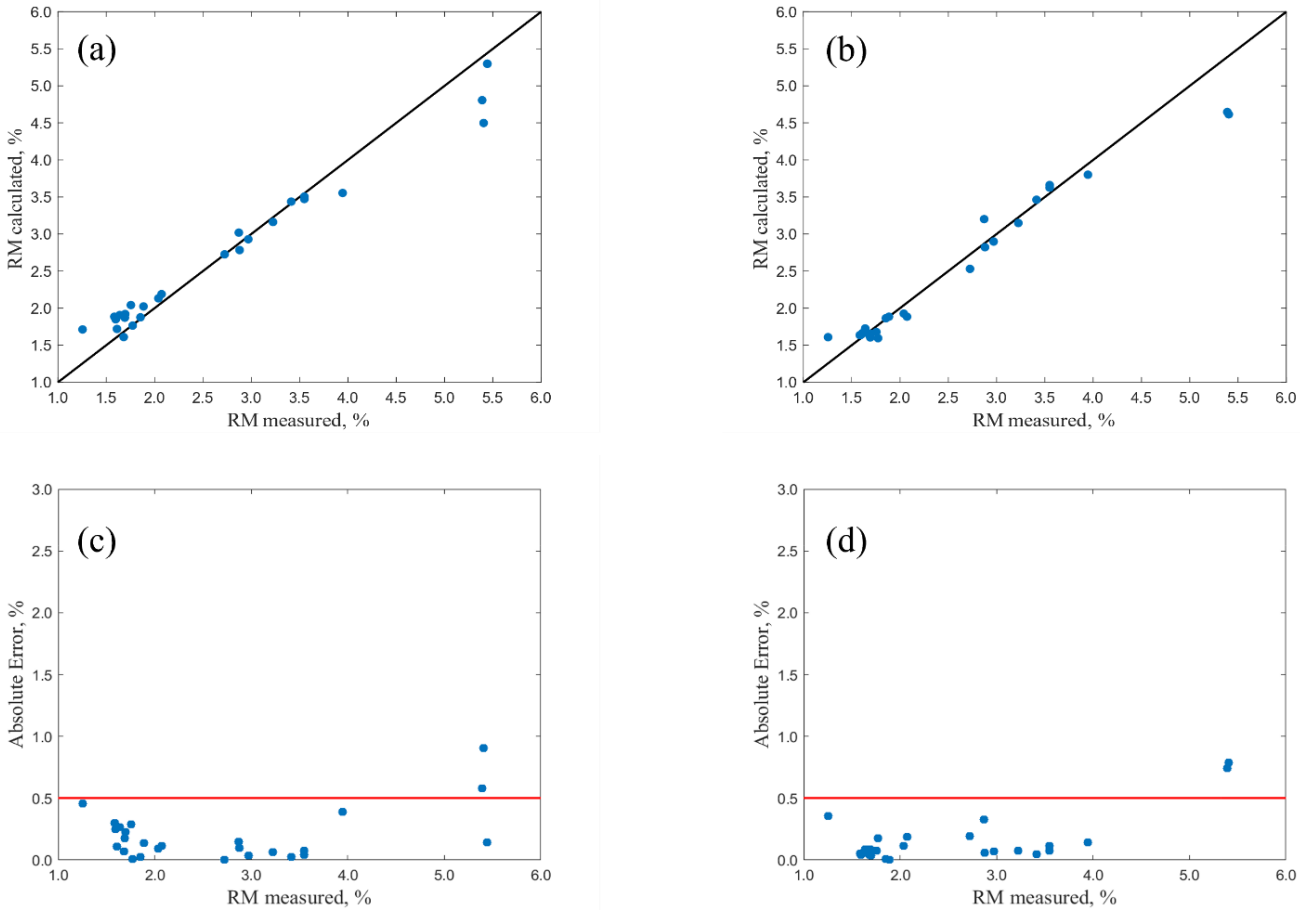

Figure S7: Parity diagrams (a, b) of dataset T6 comparing the RM measured (%) and the RM calculated (%), and absolute error plots (c, d), obtained using the linear regression (a, c) and the neural network (b, d) models developed using dataset S6 for SR.

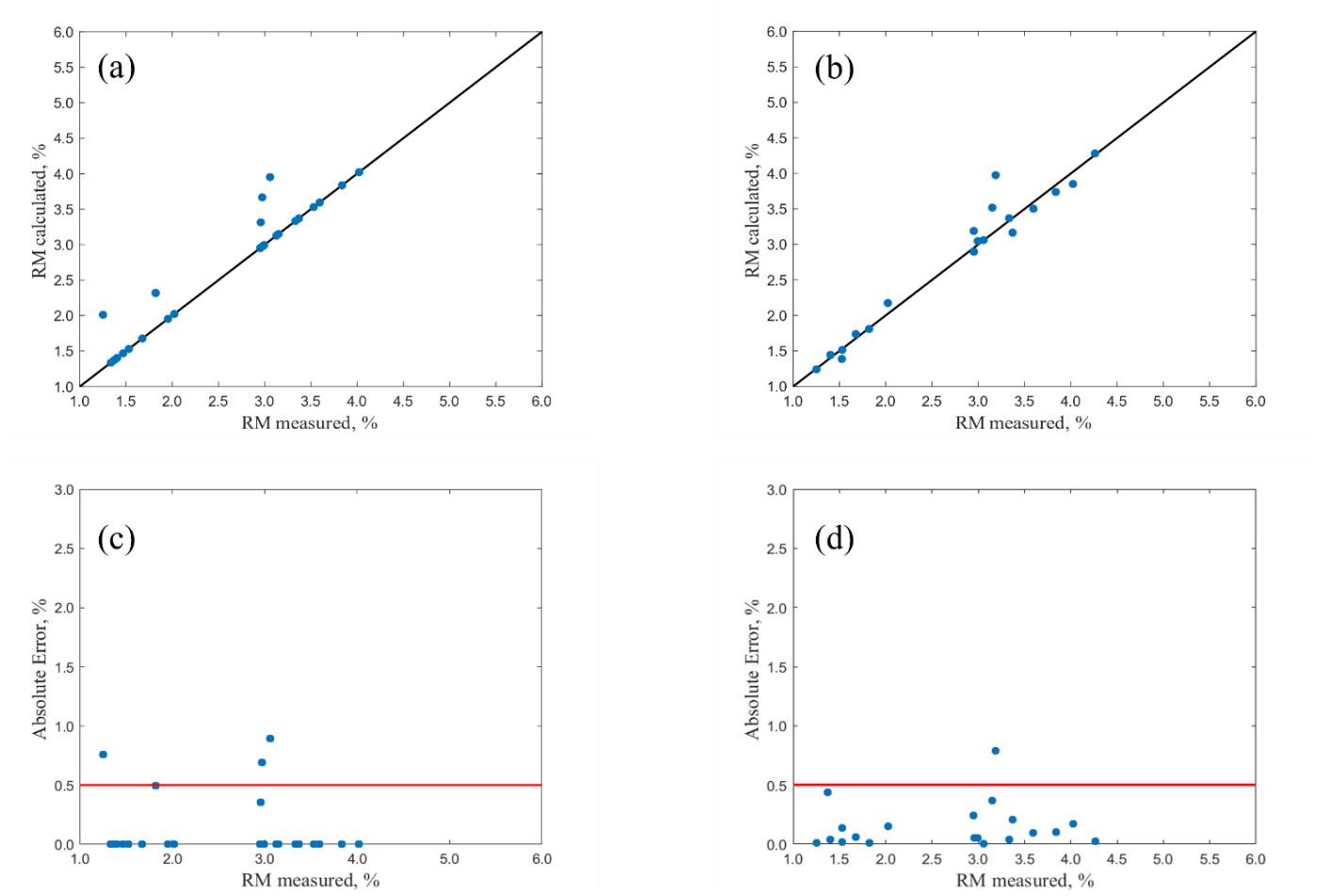

Figure S8: Parity diagrams (a, b) of dataset SA05 comparing the RM measured (%) and the RM calculated (%), and absolute error plots (c, d), obtained using the linear regression (a, c) and the neural network (b, d) models developed using datasets of arginine-sucrose mixtures for SR.

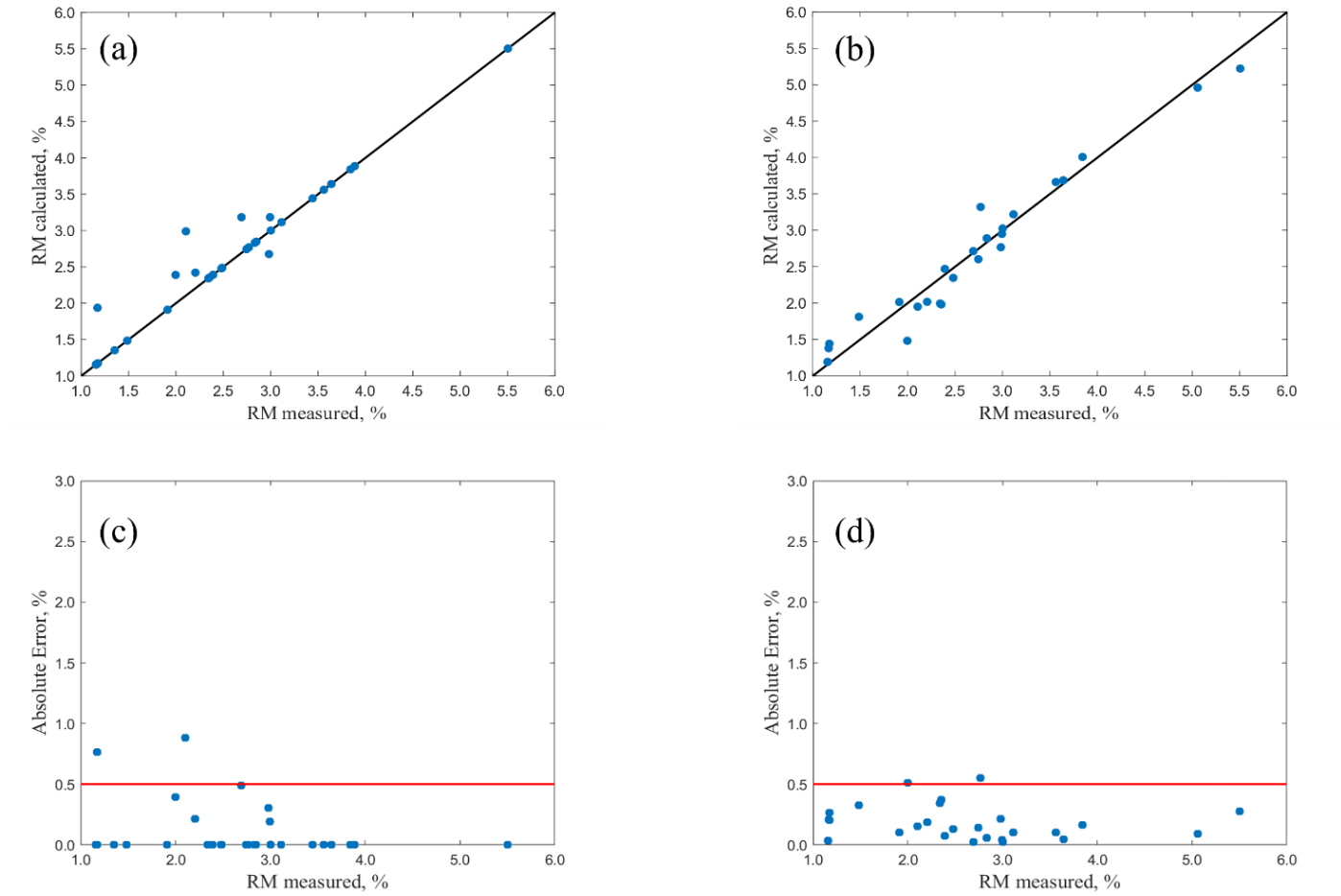

Figure S9: Parity diagrams (a, b) of dataset SA1 comparing the RM measured (%) and the RM calculated (%), and absolute error plots (c, d), obtained using the linear regression (a, c) and the neural network (b, d) models developed using datasets of arginine-sucrose mixtures for SR.

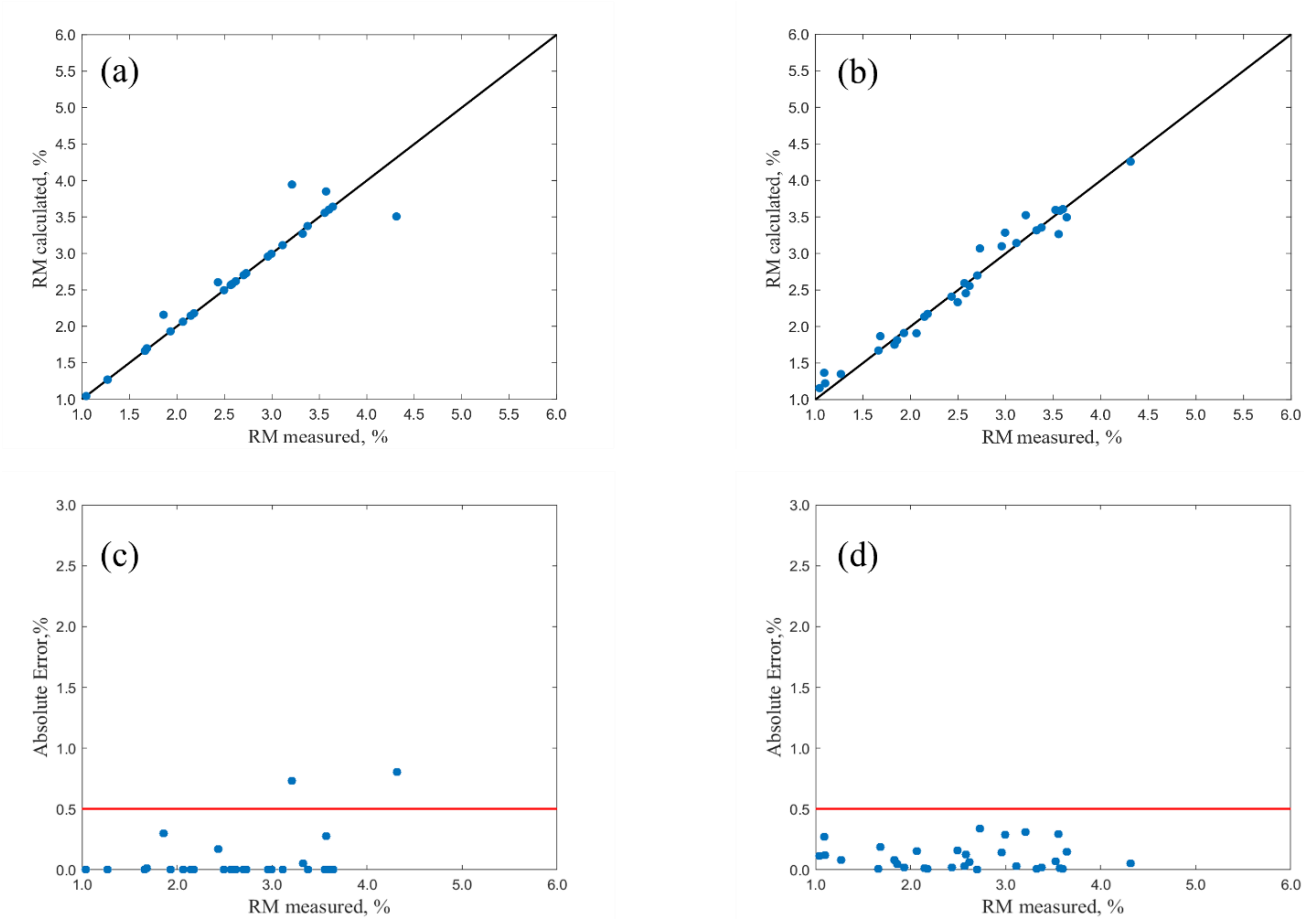

Figure S10: Parity diagrams (a, b) of dataset SA3 comparing the RM measured (%) and the RM calculated (%), and absolute error plots (c, d), obtained using the linear regression (a, c) and the neural network (b, d) models developed using datasets of arginine-sucrose mixtures for SR.

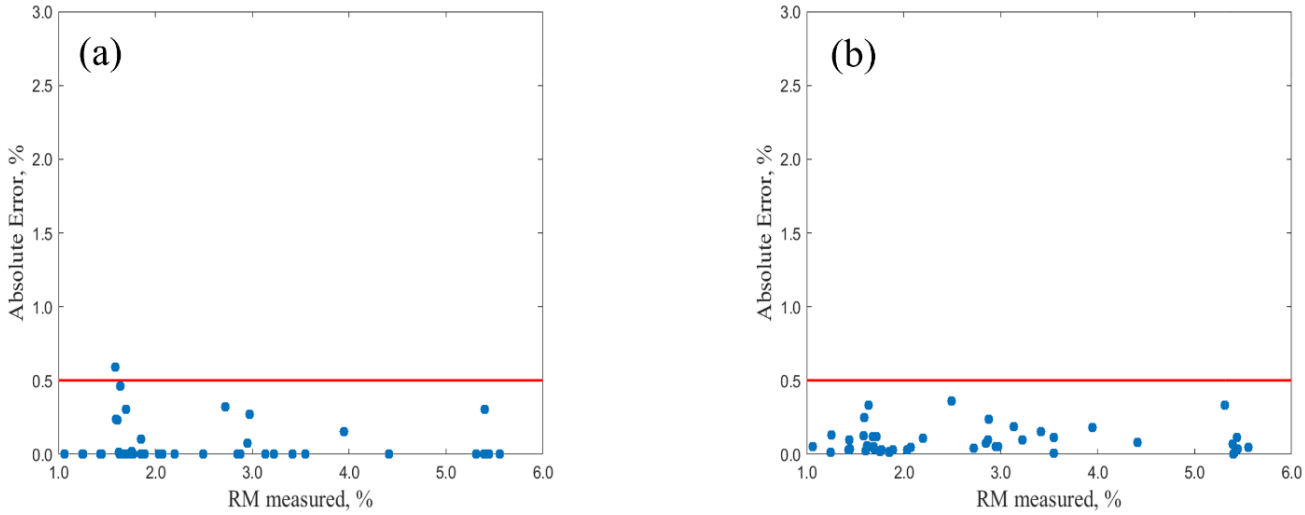

Figure S11: Absolute error plots obtained suing (a) linear regression and (b) neural network models for T6 dataset. The data were processed by the ad-hoc model for trehalose solutions for SR.

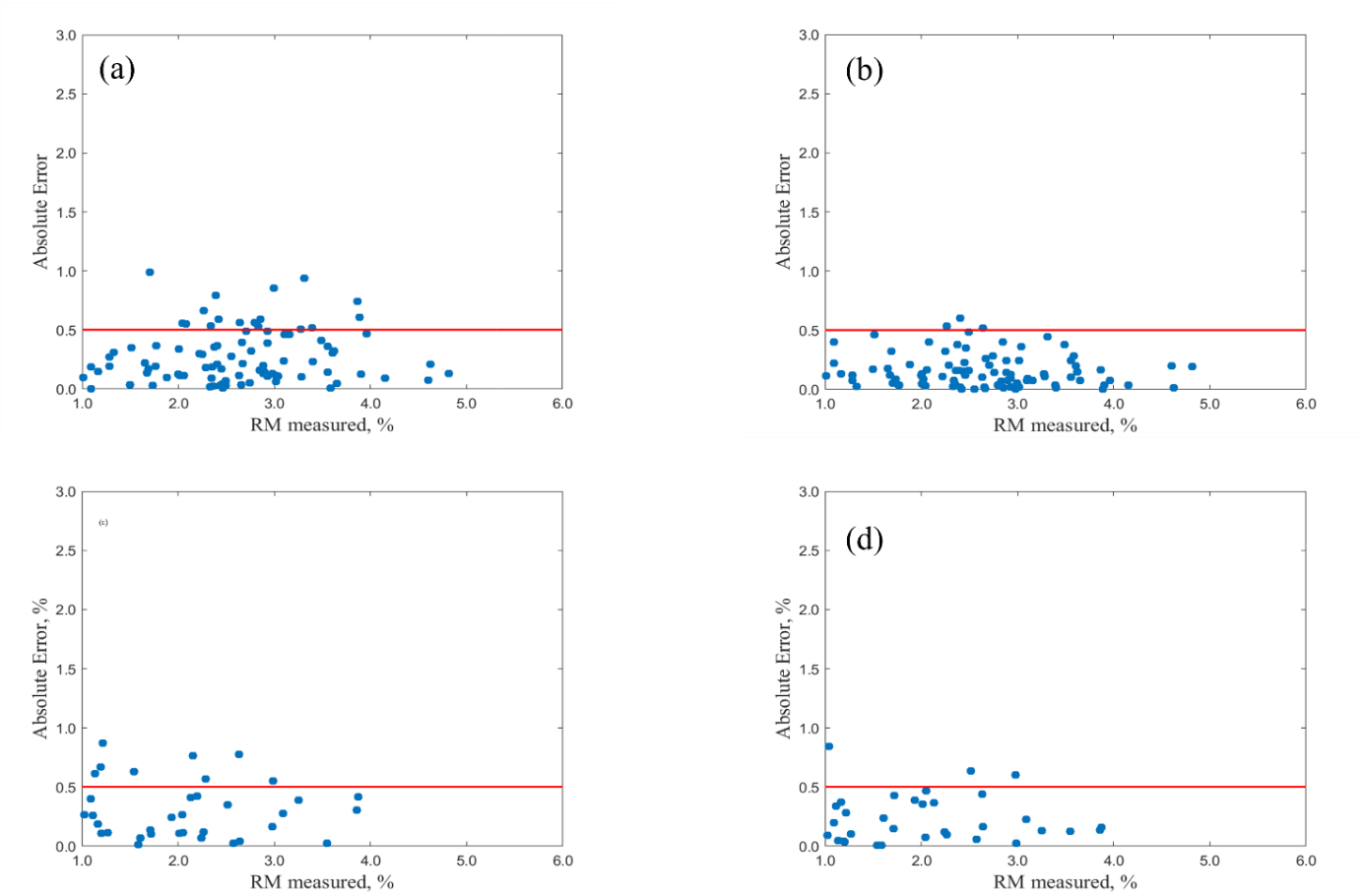

Figure S11: Absolute error plots for S6 dataset (a, b) and for S9 dataset (c, d), obtained using (a) linear regression and (b) neural network global models for SR.

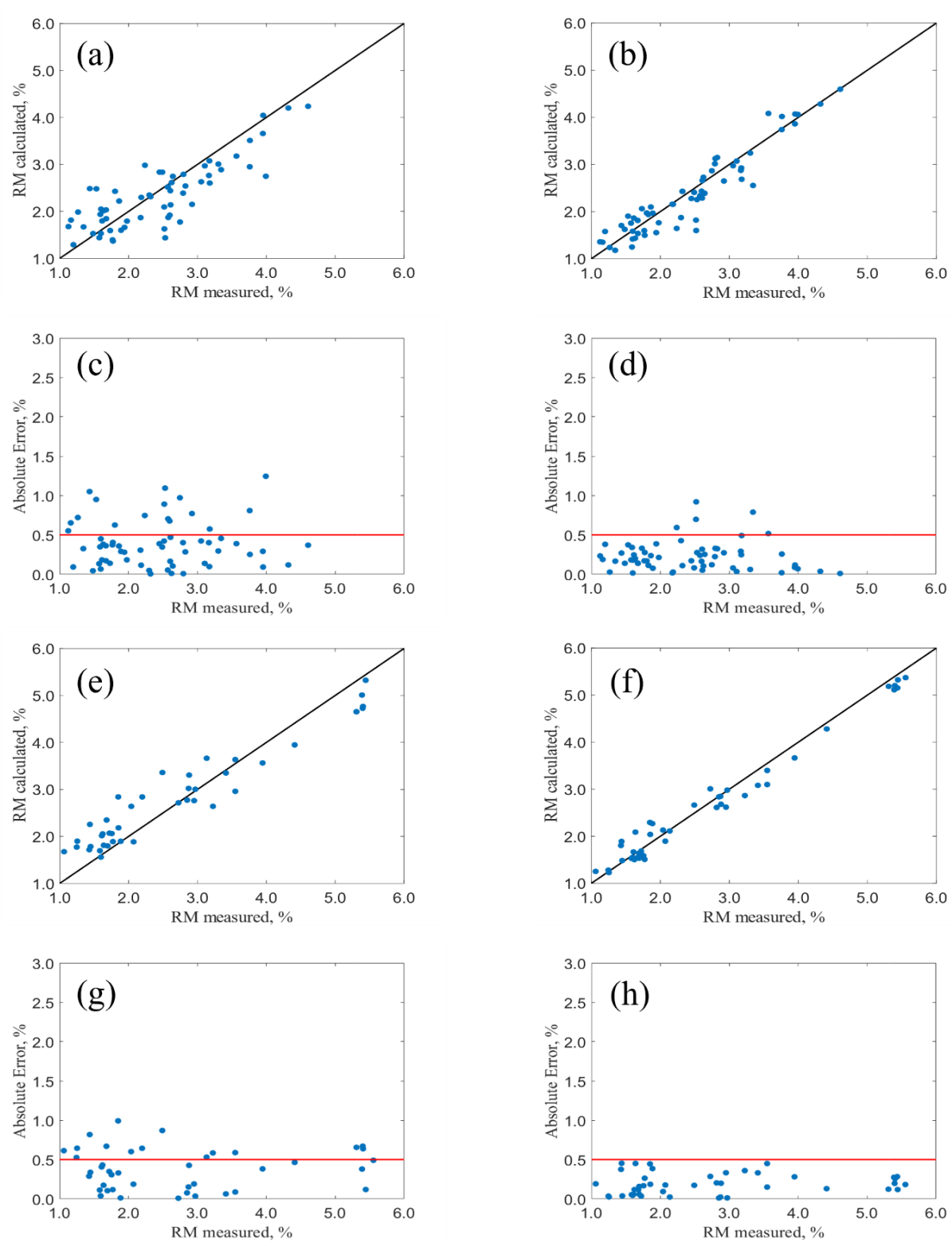

Figure S12: Parity diagrams (a, b, e, f) comparing the RM measured (%) and the RM calculated (%) and absolute error plots (c, d, g, h) obtained using the linear regression (a, c, e, g) and the neural network (b, d, f, h) global models for dataset S3 (a, b, c, d) and for dataset T6 (e, f, g, h) for SR.

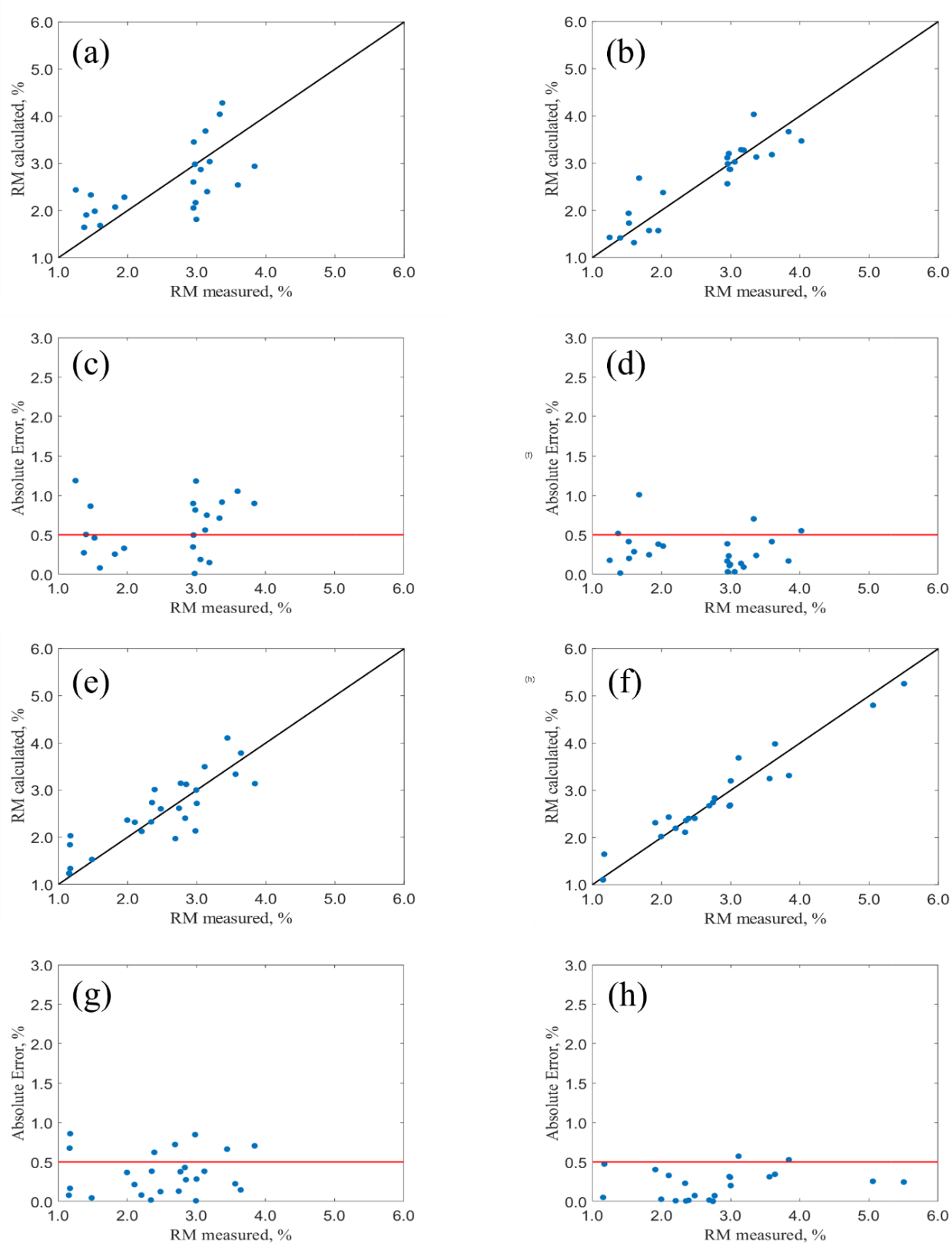

Figure S13: Parity diagrams (a, b, e, f) comparing the RM measured (%) and the RM calculated (%) and absolute error plots (c, d, g, h) obtained using the linear regression (a, c, e, g) and the neural network (b, d, f, h) global models for dataset SA05 (a, b, c, d) and for dataset SA1 (e, f, g, h) for SR.