Application of interpretable artificial neural networks to early monoclonal antibodies development

Lorenzo Gentiluomo, Dierk Roessner, Dillen Augustijn, Hristo Svilenov, Alina Kulakova, Sujata Mahapatra, Gerhard Winter, Werner Streicher, Åsmund Rinnan, Günther H.J. Peters, Pernille Harris, Wolfgang Frieß

PII: S0939-6411(19)30318-2
DOI: https://doi.org/10.1016/j.ejpb.2019.05.017
Reference: EJPB 13058

To appear in: *European Journal of Pharmaceutics and Biopharmaceutics*

Received Date: 18 March 2019
Revised Date: 17 May 2019
Accepted Date: 17 May 2019

Please cite this article as: L. Gentiluomo, D. Roessner, D. Augustijn, H. Svilenov, A. Kulakova, S. Mahapatra, G. Winter, W. Streicher, a. Rinnan, G.H.J. Peters, P. Harris, W. Frieß, Application of interpretable artificial neural networks to early monoclonal antibodies development, *European Journal of Pharmaceutics and Biopharmaceutics* (2019), doi: https://doi.org/10.1016/j.ejpb.2019.05.017

APPLICATION OF INTERPRETABLE ARTIFICIAL NEURAL NETWORKS TO EARLY MONOCLONAL ANTIBODIES DEVELOPMENT

Lorenzo Gentiluomo[1,2], Dierk Roessner[2], Dillen Augustijn[3], Hristo Svilenov[1], Alina Kulakova[4], Sujata Mahapatra[5], Gerhard Winter[1], Werner Streicher[5], Åsmund Rinnan[3], Günther H.J. Peters[4], Pernille Harris[4], Wolfgang Frieß[1]

[1]Ludwig Maximilians-Universität München, Department of Pharmacy, Pharmaceutical Technology and Biopharmaceutics, Butenandtstrasse 5, 81377 Munich, Germany

[2]Wyatt Technology Europe GmbH, Hochstrasse 12a, 56307 Dernbach, Germany

[3]Copenhagen University, Department of Food Science, Rolighedsvej 26, 1958 Frederiksberg, Denmark

[4] Technical University of Denmark, Department of Chemistry, Kemitorvet 207, 2800 Kongens Lyngby, Denmark

[5]Novozymes A/S, Krogshoejvej 36, Bagsvaerd, Denmark

*Corresponding author: Lorenzo Gentiluomo, lorenzo.gentiluomo@wyatt.eu

Keywords

neural network(s), machine learning, protein aggregation, protein formulation, monoclonal antibody, stability.

1 ABSTRACT

The development of a new protein drug typically starts with the design, expression and biophysical characterization of many different protein constructs. The initially high number of constructs is radically reduced to a few candidates that exhibit the desired biological and physicochemical properties. This process of protein expression and characterization to find the most promising molecules is both expensive and time-consuming. Consequently, many companies adopt and implement philosophies, e.g. platforms for protein expression and formulation, computational approaches, machine learning, to save resources and facilitate protein drug development. Inspired by this, we propose the use of interpretable artificial neuronal networks (ANNs) to predict biophysical properties of therapeutic monoclonal antibodies i.e. melting temperature T_m, aggregation onset temperature T_{agg}, interaction parameter k_D as a function of pH and salt concentration from the amino acid composition. Our ANNs were trained with typical early-stage screening datasets achieving high prediction accuracy. By only using the amino acid composition, we could keep the ANNs simple which allows for high general applicability, robustness and interpretability. Finally, we propose a novel “knowledge transfer” approach, which can be readily applied due to the simple algorithm design, to understand how our ANNs come to their conclusions.

2 INTRODUCTION

Therapeutic proteins play a crucial role in the treatment for various diseases.[1-3] Currently, there are over 660 biologics with market approval worldwide. Due to the recent advances in protein engineering, it is nowadays possible to fine-tune desirable protein characteristics to find the optimal balance between efficacy, safety, stability and manufacturability. The development of a protein drug is an extremely complex process involving around 5000 critical steps[4]. During the whole development process, the stability of a protein drug is a major concern. The choice of the formulation can drastically affect the conformational, the colloidal and the chemical stability and all three have to be controlled in the final product. The high number of formulation parameters and conditions to be screened requires a significant investment of resources and time. In addition, it has been shown that only 8% of the initially investigated new drug candidates reach license application.[5] It is therefore of significant importance to efficiently use the limited resources and finally to improve the drug-candidate success rate. Nowadays, high-throughput methods are commonly used during the early stage of protein development to select promising candidates and their formulations that will be put forward to undergo forced degradation studies and real-time stability tests.[6-11] In this work, we applied Artificial Neural Networks (ANNs) to the most successful class of therapeutic proteins, the monoclonal antibodies (mAbs). ANNs are biologically inspired computer programs designed to simulate how an animal brain processes information, gathering knowledge by detecting the patterns and relationships through a trial and error procedure. There has been an increasing interest in ANNs lately since computers can now process complex shallow ANNs in minutes. The speed at which ANNs can be computed and the fact that big databases are readily available makes this approach very attractive. In recent years, this method has been applied in the pharmaceutical research area for different purposes.[12-19] Supervised ANNs were used as an alternative to response surface methodology[20] while unsupervised networks are an alternative to principal component analysis. Analysis of design of experiments is also possible by ANNs.[21] The great advantage of ANNs over classical statistical modeling is that the former can solve highly non-linear problems often encountered in pharmaceutical processes. However, when the complexity of the ANNs is increased, results from ANNs become increasingly difficult to interpret. A further drawback of ANNs is that a sufficiently big data set is usually required for the learning process.

Combined, our ANNs models provide a tool that is capable of predicting important biophysical properties commonly measured in studying protein physical stability in high throughput, namely the (melting) temperature of unfolding, T_m, the diffusion interaction parameter, k_D, and the onset temperature of aggregation, T_{agg}. These biophysical properties capture different characteristics which, taken together, define significant attributes that can be used to eliminate, or continue with, the development of a candidate. T_m values frequently correlate with the aggregation rate in accelerated stability studies.[22-24] k_D is used to characterize nonspecific protein-protein interactions in diluted solutions and is a good indicator for the solution viscosity at high protein concentrations.[25] Furthermore, the rate of aggregation upon

heating a protein solution is highly correlated to k_D.[26-27] Since the aggregation needs to be kept to a minimum level, T_{agg} is an important biophysical property. The majority of marketed antibodies have T_{agg} greater than 55°C.[28] Even though the aforementioned properties alone will not always correlate with long term stability studies, their knowledge as a function of basic formulation conditions (i.e. pH and ionic strength) allows in a high-throughput way to assess the developability for protein drug candidates in high-throughput and with minimal material consumption. Still, this approach is very labor and time intensive. Therefore *in-silico* approaches are of high interest, one of them being the use of ANNs. More importantly, our trained models are based on amino acid composition only. This would allow selecting among thousands of mAbs sequences with good predicted physical stability. The selected protein could then be expressed and purified for going into the next step of the developability assessment.

As pointed out by Ali Rahimi, a researcher in artificial intelligence at Google, machine learning has become a form of alchemy.[29] Therefore our aim was to avoid black-box algorithms. We designed networks that are manageable, and give the user an understanding of their decision-making process. The number and complexity of inputs was reduced by the use of the amino acid composition only. This simple input layer allowed a simple network design which is, compared to complex networks, more general and robust, less prone to overfitting and easier to interpret. As in most cases, we achieved accurate predictions, we confirmed that this design was suitable for our purpose. To interpret our models we design a novel “knowledge transfer” process which leads to interpretable ANNs. Additionally, Partial Least Squares Regression (PLS) was performed, and the results were compared with ANNs showing that only ANNs achieve accurate predictions.

3 MATERIAL AND METHODS

3.1 Protein and Sample preparation

Five IgG1, namely PPI-1, PPI-2, PPI-3, PPI-10, PPI-13 and one IgG2 named PPI-17, were selected based on the availability of the primary sequence, were provided by the PIPPI consortium (http://www.pippi.kemi.dtu.dk). The mAbs were dialyzed overnight using 10 kDa Slide-A-Lyzer™ cassettes (Thermo Fisher Scientific, Waltham, USA) against an excess of buffer containing 10 mM Histidine at pH 5.0, 5.5, 6.0, 6.5, 7.0, 7.5. Similarly, a buffer containing 10 mM tris(hydroxymethyl)aminomethane (Tris) was used at pH 8.0 and 9.0. Sodium chloride stock solutions were prepared in the respective buffers and diluted to a final concentration of 0, 70 and 140 mM. Protein concentration was measured on a Nanodrop 2000 (Thermo Fisher Scientific, Waltham, USA) using the respective extinction coefficients calculated from the primary sequence. Reagent chemicals were of analytical grade and were purchased from Sigma Aldrich (Steinheim, Germany) or VWR International (Darmstadt, Germany). Highly purified water (HPW, Purelab Plus, USF Elga, Germany) was used for the preparation of all buffers. Formulations including sodium chloride were prepared by mixing mAbs stock solution in the respective buffer with a stock solution of sodium chloride dissolved in the same buffer.

Finally, the formulations were sterile filtered with 0.22 µm cellulose acetate filters from VWR International (Darmstadt, Germany). The mAbs' difference in primary structures was investigated using identity and similarity scores as shown in Table S1.

3.3 Dynamic light scattering

Dynamic light scattering was conducted on a DynaPro Plate Reader II (Wyatt Technology, Santa Barbara, USA) to obtain the interaction diffusion parameter, k_D, the onset temperature of aggregation, T_{agg}, and the apparent hydrodynamic radius, R_h. 4 µL of each sample per well were pipetted in triplicates into Aurora 1536 Lobase Assay Plates (Aurora Microplates, Whitefish, USA). The samples were overlayed with Silicone oil and centrifuged at 2000 rpm for 1 minute. Data were processed by the DYNAMICS software V7.7 (Wyatt Technology, Santa Barbara, USA). From the relative autocorrelation function, the coefficient of self-diffusion, D, and the polydispersity index (PDI) were calculated. R_h, was calculated by means of the Stokes-Einstein equation.

k_D was determined using at least six different concentrations (from 1 to 10 mg/mL) in triplicates for each formulation. The samples were filtered using a Millex® 0.22 µm filter from Merk Millipore (Burlington, USA) and equilibrated at 25 °C for 10 minutes in the Plate reader. Each measurement included 20 acquisitions, each for a duration of 5 s. k_D was determined according to:

$$D = D_0(1 + k_D \cdot c)$$

where D_0 denotes the diffusion coefficient of an isolated scattering solute molecule in the solvent and c is the protein concentration.

For the determination of T_{agg}, the filtered samples at 1 mg/mL were analyzed in duplicates. To achieve high throughput while keeping a suitable point density, 48 wells were filled, and a temperature ramp rate of 0.1°C/min from 25°C to 80°C was applied. One measurement included 3 acquisitions each with a duration of 3 s. T_{agg} was calculated by the DYNAMICS software V7.7 onset algorithm from the increase in R_h.

3.5 Differential Scanning Fluorimetry with Intrinsic Protein Fluorescence Detection (nanoDSF)

Samples containing 1 mg/mL protein in the respective formulations were filled in standard nanoDSF capillaries (NanoTemper Technologies, Munich, Germany). Measurements were performed using the Prometheus NT.48 (NanoTemper Technologies, Munich, Germany) system that measures the intrinsic protein fluorescence intensity change at 330 and 350 nm (after excitation at 280 nm) as a function of temperature. A temperature ramp of 1°C/min was used from 20 to 95°C. The fluorescence intensity ratio (F350/F330) was plotted against the temperature, and the first apparent melting temperature (T_m) was

derived from the maximum of the first derivative of each measurement using the PR Control software V1.12 (NanoTemper Technologies, Munich, Germany).

3.6 Artificial Neural Networks

Artificial Neural Networks have been extensively reviewed in the literature, and they have been used successfully in the pharmaceutical industry.[12-21, 30-36] The various applications of ANNs relevant to the pharmaceutical field are classification or pattern recognition, prediction and modeling. Theoretical details can be found elsewhere.[37]

The network's fundamental parts are the neurons, also called nodes, and their connections. The diagram in **Fig. 1** shows the model of a neuron. The neuron is an information-processing unit, which is constituted of a set of connection links characterized by their weight, w_{kn}, a linear combiner, Σ, and an activation function, ψ. An externally applied bias, $b_{k,}$ is used to modify the net input received for each neuron in the network. An often used simplified description of the network is the architectural graph, depicted in **Fig. 2**.

ANNs solve problems by training, a trial and error process for optimizing the synaptic weight values. During the training, the squared error between the estimated and the experimental values is minimized by reinforcing the synaptic weights, w_{kn}. ANNs have robust performance in dealing with noisy or incomplete data sets, the ability to generalize from input data and a high fault tolerance.[38]

ANNs have a series of known limitations, namely overfitting, chance effects, overtraining, and difficult interpretability.[39-41] The first three limitations were extensively reviewed in the literature and can be prevented using various methodologies. The interpretation of ANNs is not straightforward, and it is still an open field of research. Our primary goal was therefore to build an algorithm where it was possible to follow how the networks have come to a particular conclusion. To achieve this, we used the simplest input related to the mAbs giving an accurate prediction, namely the amino acid composition. In order to comprehend the artificial decision-making procedure a novel "knowledge transfer" process was designed, which is described in section 3.7.

Our multilayer feed-forward back-propagation networks present one hidden layer, which is usually sufficient to provide adequate predictions even when continuous variables are adopted as units in the output layer.[43-45] Equation 1 (described by Carpenter[44]) was used to estimate the optimal number of neurons in the hidden layer:

Eqn. 01

$$N_{hidden} = \left(\frac{N_{sample}}{\beta} - N_{output}\right) \Big/ (N_{input} + N_{output} + 1)$$

where β, N_{hidden}, N_{output} and N_{sample} are the determination parameter, the number of hidden units, the number of output units and the number of training data pairs, respectively. Overdetermined,

underdetermined and determined parameters will be reflected by $\beta>1$, $\beta<1$ and $\beta=1$, respectively. The β value to adopt depends on the degree of quality of the data set in terms of the degree of independency among other factors. Our dataset consisted of 144 instances (24 conditions per protein) for each biophysical parameter and seven neurons were estimated to provide a β of 1. In general terms, simpler models are more general and easier to interpret. Since our aim was to have the most general and easier to interpret model possible, we selected the minimum number of neurons, 5, which provided the same result as 7 neurons. In **Table S2** the list of input parameters relative to each model is shown, while in **Fig. S3** an exemplary scheme of the model's architecture is presented. All the input parameters were normalized before the training phase by subtracting the mean and then dividing by the standard deviation. The learning rate was selected on a trial and error basis in such a way so as to keep the minimum distance between the actual and predicted value. The validation method is described in section 4.1. JMPpro® (SAS Institute Inc., Cary, USA), MATLAB® (MathWorks, Natick, USA) and Weka (Waikato University; New Zealand) were used to generate ANNs. These networks yielded highly similar results and JMPpro® v.13 was selected for its user-friendly interface and subsequently potentially easier implementation in a drug development department.[47]

3.7 Knowledge transfer to explain ANN network results

In order to understand the decision-making process of our ANN models, a novel knowledge transfer process, implying response surface methodology (RSM), was applied by evaluating the weights of the trained network to transfer the acquired knowledge of ANNs to linear models. Parameters deemed important by the networks were selected, and the interpretation of ANNs was then assessed by RSM of the linear least square regression of these "leading parameters". The scheme of this process, named "knowledge transfer", is depicted in **Fig. 3**.

None of the hidden nodes in the ANNs' prediction formulas has a weight close to zero, which means that all nodes contribute to the final output. However, around 5% of the weights of the output layer presented values which were at least twice the average mean of all the network weights. From these 5%, we selected the input parameters from the activation functions whose coefficients were at least twice the average values.

We assessed the full model using all the selected "leading parameters" from the networks, and then reduced the model to only the terms that were deemed statistically relevant. A curved response was allowed by assessing the quadratic term considering also two-way interactions. The reduced model was obtained using a backward stepwise regression. The F-statistic approach was used to perform the effect test considering a value of 0.05 or less as statistically significant. All the results were calculated using the statistical software JMP® v 13.0 (SAS Institute Inc., Cary, USA)[47], and all the analysis details can be found in the software manual

4 RESULTS AND DISCUSSION

A general flow diagram of our approach is shown in **Fig. 4**. At first, the power of our ANNs for prediction of the biophysical parameters T_m, T_{agg} and k_D at different pH as well as salt concentration was evaluated. Only the number of each amino acid species of the proteins was used as protein-related input parameters. The primary sequence was not used as an input parameter, neither were other typical molecular descriptors included e.g. charge distribution, dipole moments or solvent exposure. However, we are currently working together with other members of the PIPPI consortium (http://www.pippi.kemi.dtu.dk) to create a publicly available protein formulation database. Such a database may be used in the future to build on our findings and to generate more sophisticated deep learning models based on the amino acid sequence. We avoided the use of formulation dependent molecular descriptors (e.g. net charge) to reduce redundancy, as the formulation is always included as input. Moreover, it has been proven that even net charge cannot be accurately calculated.[48] Further, we investigated a series of molecular indices which are only protein dependent, calculated by ProtDCal,[49] listed in **Table** SI 6. However, we could not find a subset of these indices that would yield an accuracy similar to the number of amino acids. As machine learning models describe correlation and not causation - highlighted by George E. P. Box: *"Essentially, all models are wrong, but some are useful"*[50] - we selected the minimum number of input parameters to achieve high accuracy and interpretability. The number of amino acids can easily be described by only 20 input values, whereas thousands of inputs are necessary to describe the primary sequence (depending on the size of the molecule). This would drastically increase the complexity of the algorithms requiring a deep neural network with thousands to millions of data points, which are nowadays not publicly available. Such a complex approach makes the algorithm difficult to interpret and interpretability was one of our goals. As we managed to reach accurate predictions we found our model useful for its purpose: an *in-silico* tool for the selection of mAbs with predicted high physical stability from a vast number of possible candidates, which is interpretable, which is independent from other calculations (e.g. solvent exposure), and which can output experimentally accessible biophysical properties in early stage (i.e. low volume, high throughput). An additional advantage of a simple design is that such models are usually more general and robust.

In order to gain insight from the ANNs decision making procedure we introduce a novel knowledge transfer process (depicted in red in **Fig. 4**). As the outputs (e.g. T_m) of our models are easily accessible in early stage, once the selected candidates are expressed and purified, it is possible to continuously re-train the network and to double check its validity. One disadvantage of such approach is that is suitable only to predict closely related protein structures to the one used for the training phase, e.g. IgG1 and IgG2.

4.1 Prediction of T_m, T_{agg} and the sign of k_D

The ability of the model to predict T_m, T_{agg} and k_D from the numbers of each different amino acid in each mAb and the formulation conditions (i.e. pH and salt concentration) was cross-validated. Data from two mAbs were selected and held back in a validation set during the training phase. Applying the model to the validation data allows an unbiased comparison between the predicted and measured values. Thus, the estimation of the prediction error for potential new mAb samples is based on the results of the validation set. This validation method was deemed superior to the random data splitting. The latter yielded better fitting and prediction. However, the model would have experienced all the molecules during the training phase. Therefore, we discarded the random data splitting as our aim was to validate a model capable of predict biophysical parameters of unknown mAbs. Using this cross-validation strategy, a total of fifteen models were built, each of them based on a different training and validation set, for each studied biophysical property. As the investigated mAbs presented different stability (i.e. different biophysical properties values) the point distribution varies depending on the validation mAbs. The models were characterised by the name of the withheld proteins (e.g., the model called PPI-1&2 is based on the validation data set of PPI-1 and PPI-2, and trained on the PPI-3, PPI-10 , PPI-13 and PPI-17 data).
In **Fig. 5**, the predicted T_m, T_{agg} and the sign of k_D of the PPI-3&13 models are shown. T_m and the sign of k_D were fitted to a very high degree of accuracy. The T_m model presented an R^2 of 0.98 and a root mean squared error (RMSE) of around 0.8°C from the reference T_m while the sign of the k_D model was classified with no false negative or false positives. The T_{agg} model presented an R^2 of 0.94 but with a higher RMSE value of around 2°C. The higher error is probably due to the high throughput fashion of the screening, which stretched the limit of necessary high data density for the determination of the onset. In other words, the input data has higher uncertainty that is reflected in the prediction error. In **Figs. S4-S5**, the predicted data point from the T_m and T_{agg} models are presented.

The robustness of the ANNs regressions was evaluated based on R^2, shown in **Fig. 6 (A)**, and RMSE values of the training and validation set. The latter was in the range of ca. 1 to 3 °C from the reference T_{agg} or T_m, with no particular trend or direction with respect to the measured values. The robustness of the classification problem, the sign of k_D, was evaluated on the misclassification rate, shown in **Fig. 6 (B)**.

Regarding the T_m models, we observe broad robustness without significant influence of the different training sets. The colloidal stability parameters, T_{agg} and sign of k_D, appear to be more sensitive to the selected training sets. Two T_{agg} models show serious deviation in prediction both involving PPI-17 and/or PPI-10. These two proteins showed extreme aggregation during temperature ramps, compared to the other mAbs. Consequently, the ANNs can easily fit PPI-17 and PPI-10 data, but in order to predict their aggregation propensity, the network would require more data representative of this kind of aggregation behavior.
The k_D data consists for ca. 70% of negative values. This unbalanced data set is caused by the charge screening effect of the added salt that occurs in two-thirds of the formulations and therefore the number of positive values is not enough to solve an ANN regression problem. One such occurrence is shown in **Fig.**

7 for the PPI-13&3 model, where all the negative values are fit well, while the positive values are not well calculated and broadly distributed. Despite this, the sign of k_D was always predicted to a high degree of accuracy as shown in **Fig. 6(B)**.

The studies on the robustness allowed us to conclude that well defined and simpler properties, such as the temperature of unfolding, are not greatly influenced by the training set. In contrast, the colloidal properties need more attention in the selection of the training set.

4.2 ANN Knowledge Transfer

The scientific community has been investigating the problem of explaining machine learning decision models and a comprehensive survey of methods for explaining black box models has been redacted.[51] In order to understand the thought process of our ANNs, a novel knowledge transfer process, depicted in **Fig**. 3, was applied. **Fig. 8** shows the results from the RSM relative to T_m, T_{agg}, k_D, while **Table 1** summarizes the effective test statistics which can be used as an indication of the relative impact of the parameters. Quadratic terms (e.g. Cys·Cys) were assessed to model potential curvature in the response. These linear models allow to understand the logic of the relative ANNs model and to follow the reasoning of the outcomes, i.e. each leading amino acid has a specific role in the physical process related to the output parameters.

The T_m linear model is primarily affected by pH, salt concentration, and the number of tryptophan, cysteine and tyrosine residues. Therefore, the main protein related contributors to the unfolding process are two hydrophobic amino acids residues and cysteine. It is known that the unfolding process is mainly guided by hydrophobic interactions,[52] while cysteine is involved in disulfide bonds, stabilizing the protein structure. Interestingly charged residues are of minor importance.

The T_{agg} linear model is mainly affected by pH, salt concentration, and the number of aspartic acid, glutamic acid and methionine residues. Therefore, the main protein related contributors to the aggregation process were charged amino acid residues and methionine. It is known that the oxidation of methionine is a critical pathway of aggregation under accelerated thermal stability stress[53]. Moreover, methionine oxidation is pactically pH independent[54], which could partially explain the minor impact of pH on the models. However, during a temperature ramp, the time of stress is relatively short and hence, the oxidation of methionine should have a minor impact. Consequently, during a temperature ramp, charged amino acids have a higher impact on the linear model.

The k_D linear model is affected by pH, salt concentration, and the number of glutamic acid, histidine, and tryptophan residues. Thus, both charged and hydrophobic amino acids are important. k_D is used to evaluate pairwise protein-protein nonspecific interactions, which can be rationalized by means of the DLVO[55,56] or proximity energy theory[57]. Both theories highlight the fact that protein-protein interactions depend heavily on hydrophobic and charged patches on the protein surface. Moreover, histidine plays a particular role in protein-protein interactions. This amino acid has a pK_a of 6.0 i.e. histidine changes

charge state under relevant formulation pH conditions. Therefore histidine doping is a common method in engineering stable proteins[58-62] and the presence of histidine residues can mediate structural transitions in binding or folding of the interacting proteins.[63-65]

Taken together, our ANN knowledge transfer process allows us to interpret the factors behind the decision-making process of the ANN to when predicting T_m, T_{agg} the sign of k_D. This process provided a global explanation of the black box through an interpretable and transparent model. By this, we build trust into our approach and are not left with a black box. As an agnostic process can explain unrelated algorithm only indifferently, our approach is not to be considered agnostic as it is tied to simple ANNs.

4.3 Prediction comparison with partial least square models

The main reason to apply ANNs comes from their prediction power using data sets with highly non-linear relationships. To demonstrate the necessity for a non-linear model, a linear regression analysis using the partial least square regression (PLS) method, was performed. PLS is probably the strongest competitor of ANNs in terms of robustness and predictive power and can be extremely powerful in fitting data and for this reason, it was compared to ANN. In fact, PLS was the only model we tested capable of fitting the dataset. As we aimed to develop an interpretable model, we tested also models usually considered readily interpretable (e.g. decision tree) without success. A detailed discussion about modeling alternatives can be found in an article by Frank and Friedmann.[66] The optimal number of latent variables was selected based on the minimum of the RMSE of the cross-validation. The same cross-validation method was applied as in the ANNs in order to make the models comparable. In **Fig. 9**, the prediction for all the proteins is shown. The results demonstrate that PLS cannot be used for our dataset and we can conclude that ANN is a far better methodology than PLS to construct models that predict the formulation behavior of unknown proteins under the conditions that we have used.

5 CONCLUSIONS

ANNs represent an interesting alternative to the classical statistical methodologies when applied to highly non-linear data sets that are frequently encountered in the pharmaceutical industry. We successfully developed interpretable models for a set of mAbs to predict important biophysical properties as a function of pH and salt concentration. In the field of mAbs development, ANNs could be a highly valuable tool to predict important biophysical properties and to support development risk assessment. This approach would allow the selection of mAbs with good physicochemical properties already before expression in cells. The only information required for our approach is the amino acid composition of each mAb. Due to the accuracy of the predictions, there was no reason to increase the complexity of the model since it would hamper the interpretability and robustness. Thanks to our design a novel knowledge transfer process allows to understand the decision-making process of our algorithm. In contrast, PLS models did not work demonstrating that a non-linear algorithm is required to analyze a data set like the one used in our study. The knowledge gathered with simpler ANNs can be used to build even more impressive

systems in the future, to confirm the reliability of ANNs and finally to highlight which factors may impact protein stability most.

6 Acknowledgements

This study was funded by a project part of the EU Horizon 2020 Research and Innovation program under the Marie Skłodowska-Curie grant agreement No 675074. The first author would like to thank Nanotemper Technologies GmbH for kindly proving support for the NanoDSF data, SAS Institute for proving JMPpro® V 13.0, and the whole PIPPI consortium (http://www.pippi.kemi.dtu.dk) for the continuous support offered and for reviewing the manuscript.

7 References

1. Gong R, Chen W, Dimitrov DS. Expression, purification, and characterization of engineered antibody CH2 and VH domains. *Methods Mol Biol* 2012;899:85-102.

2. Dimitrov DS. Therapeutic antibodies, vaccines and antibodyomes. *MAbs* 2010; 2(3):347-56.

3. Elvin JG, Couston RG, van der Walle CF. Therapeutic antibodies: market considerations, disease targets and bioprocessing. *Int J Pharm.* 2013;440(1):83-98.

4. Lagassé HA, Alexaki A, Simhadri VL, Katagiri NH, Jankowski W, Sauna ZE, Kimchi-Sarfaty C. Recent advances in (therapeutic protein) drug development. *F1000Research* 2017; 6:113.

5. US Department of Health and Human Services. Innovation or stagnation: Challenge and opportunity on the critical path to new medical products. Available at: http://wayback.archive-it.org/7993/20180125032208/https://www.fda.gov/ScienceResearch/SpecialTopics/CriticalPathInitiative/CriticalPathOpportunitiesReports/ucm077262.htm. Accessed July 3, 2017.

6. Capelle MA, Gurny R, Arvinte T. High throughput screening of protein formulation stability: practical considerations. *J Pharm Biopharm*. 2007;65(2):131-48.

7. He F, Hogan S, Latypov RF, Narhi LO, Razinkov VI. High throughput thermostability screening of monoclonal antibody formulations. *J Pharm Sci* 2010;99(4):1707-20.

8. Goldberg DS, Bishop SM, Shah AU, Sathish HA. Formulation development of therapeutic monoclonal antibodies using high-throughput fluorescence and static light scattering techniques: Role of conformational and colloidal stability. *J Pharm Sci* 2011;100(4):1306-15.

9. Goldberg DS, Lewus RA, Esfandiary R, Farkas DC, Mody N, Day KJ, Mallik P, Tracka MB, Sealey SK, Samra HS. Utility of High Throughput Screening Techniques to Predict Stability of Monoclonal Antibody Formulations During Early Stage Development. *J Pharm Sci* 2017;106(8):1971-1977

10. Chaudhuri R, Cheng Y, Middaugh CR, Volkin DB. High-throughput biophysical analysis of protein therapeutics to examine interrelationships between aggregate formation and conformational stability. *AAPS J* 2014;16(1):48-64.

11. Maddux NR, Iyer V, Cheng W, Youssef AM, Joshi SB, Volkin DB, Ralston JP, Winter G, Middaugh CR. High throughput prediction of the long-term stability of pharmaceutical

macromolecules from short-term multi-instrument spectroscopic data. *J Pharm Sci* 2014;103(3):828-39.

12. Hussain AS, Yu XQ, Johnson RD. Application of neural computing in pharmaceutical product development. *Pharm Res* 1991;8(10):1248-52

13. Murtoniemi E, Merkku P, Kinnunen P, Leiviskae K, Yliruusi J. Effect of neural network topology and training end point in modelling the fluidized bed granulation process. *Int J Pharm* 1994;110(2): 101-108.

14. Gasperlin M, Tusar L, Tusar M, Smid-Korbar J, Zupan J, Kristl J. Lipophilic semisolid emulsion systems: viscoelastic behaviour and prediction of physical stability by neural network modelling. *Int J Pharm* 2000;196(1):37-50.

15. Takayama K, Fujikawa M, Nagai T. Artificial neural network as a novel method to optimize pharmaceutical formulations. *Pharm Res* 1999;16(1):1-6.

16. Achanta AS, Kowalski JG, Rhodes CT. Artificial neural networks: implications for pharmaceutical sciences. *Drug Dev Ind Pharm* 2008;21(1): 119-155.

17. King AC, Woods M, Liu W, Lu Z, Gill D, Krebs MR. High-throughput measurement, correlation analysis, and machine-learning predictions for pH and thermal stabilities of Pfizer-generated antibodies. *Protein Sci* 2011;20(9):1546-57.

18. Yang Y, Ye Z, Su Y, Zhao Q, Li X, Ouyang D. Deep learning for in vitro prediction of pharmaceutical formulations. *Acta Pharm Sin B.* 2019 Jan;9(1):177-185.

19. Ye Z, Yang Y, Li X, Cao D, Ouyang D. An Integrated Transfer Learning and Multitask Learning Approach for Pharmacokinetic Parameter Prediction. *Mol Pharm.* 2019 Feb 4;16(2):533-541.

20. Bourquin J, Schmidli H, van Hoogevest P, Leuenberger H. Application of artificial neural networks (ANN) in the development of solid dosage forms. Pharm Dev Technol 1997;2(2):111-21.

21. Plumb AP, Rowe RC, York P, Doherty C. The effect of experimental design on the modeling of a tablet coating formulation using artificial neural networks. *Eur J Pharm Sci* 2002;16(4-5):281-8.

22. Burton L, Gandhi R, Duke G, Paborji M. Use of microcalorimetry and its correlation with size exclusion chromatography for rapid screening of the physical stability of large pharmaceutical proteins in solution. *Pharm Dev Technol* 2007;12(3):265-73.

23. Brader ML, Estey T, Bai S, Alston RW, Lucas KK, Lantz S, Landsman P, Maloney KM. Examination of thermal unfolding and aggregation profiles of a series of developable therapeutic monoclonal antibodies. *Mol Pharm* 2015;12(4):1005-17.

24. Kumar V, Dixit N, Zhou LL, Fraunhofer W. Impact of short range hydrophobic interactions and long range electrostatic forces on the aggregation kinetics of a monoclonal antibody and a dual-variable domain immunoglobulin at low and high concentrations. *Int J Pharm* 2011;421(1):82-93.

25. Neergaard MS, Kalonia DS, Parshad H, Nielsen AD, Møller EH, van de Weert M. Viscosity of high concentration protein formulations of monoclonal antibodies of the IgG1 and IgG4 subclass–Prediction of viscosity through protein–protein interaction measurements. *Eur J Pharm Sci* 2013;49(3):400-10.

26. Rubin J, Linden L, Coco WM, Bommarius AS, Behrens SH. Salt-induced aggregation of a monoclonal human immunoglobulin G1. *J Pharm Sci* 2013;102(2):377-86.

27. Rubin J, Sharma A, Linden L, Bommarius AS, Behrens SH. Gauging colloidal and thermal stability in human IgG1–sugar solutions through diffusivity measurements. *J Phys Chem B* 2014;118(11):2803-9.

28. Jarasch A, Koll H, Regula JT, Bader M, Papadimitriou A, Kettenberger H. Developability assessment during the selection of novel therapeutic antibodies. J Pharm Sci. 2015;104(6):1885-1898.

29. Hutson M. Has artificial intelligence become alchemy? *Science* 2018; 4;360(6388):478

30. Hussain AS, Yu XQ, Johnson RD. Application of neural computing in pharmaceutical product development. *Pharm Res* 1991;8(10):1248-52.

31. Ghaffari A, Abdollahi H, Khoshayand MR, Bozchalooi IS, Dadgar A, Rafiee-Tehrani M. Performance comparison of neural network training algorithms in modeling of bimodal drug delivery. *Int J Pharm* 2006;327(1-2):126-38.

32. Hussain A, Shivanand P, Johnson RD. Application of neural computing in pharmaceutical product development: computer aided formulation design. *Drug Dev Ind Pharm* 2008;20(10):1739-1752.

33. Murtoniemi E, Yliruusi J, Kinnunen P, Merkku P, Leiviskae K. The advantages by the use of neural networks in modelling the fluidized bed granulation process. Int J Pharm 1994;108(2):155-164.

34. Agatonovic-Kustrin S, Beresford R. Basic concepts of artificial neural network (ANN) modeling and its application in pharmaceutical research. *J Pharm Biomed Anal* 2000;22(5):717-27

35. Bourquin J, Schmidli H, van Hoogevest P, Leuenberger H. Advantages of Artificial Neural Networks (ANNs) as alternative modelling technique for data sets showing non-linear relationships using data from a galenical study on a solid dosage form. *Eur J Pharm Sci* 1998;7(1):5-16.

36. Chen Y, Thosar SS, Forbess RA, Kemper MS, Rubinovitz RL, Shukla AJ. Prediction of drug content and hardness of intact tablets using artificial neural network and near-infrared spectroscopy. *Drug Dev Ind Pharm* 2001;27(7):623-31.

37. Haykin SS. *Neural networks: a comprehensive foundation, 2nd ed.,* Prentice Hall PTR ; 1998.

38. Patterson DW. *Artificial neural networks: theory and applications*. Prentice Hall Asia, 1998.

39. Livingstone DJ, Manallack DT, Tetko IV. Data modelling with neural networks: advantages and limitations. *J Comput Aided Mol Des* 1997;11(2):135-42.

40. Manallack DT, Livingstone DJ. Artificial neural networks: application and chance effects for QSAR data analysis. *Med Chem Res* 1992;2:181-190.

41. Livingstone DJ, Manallack DT. Statistics using neural networks: chance effects. *J Med Chem* 1993;36(9):1295–1297.

42. Manallack DT, Ellis DD, Livingstone DJ. Analysis of linear and nonlinear QSAR data using neural networks. *J Med Chem* 1994;37(22):3758-3767.

43. Lippman RP. An introduction to computing with neural nets. *IEEE Assp Mag* 1987;4(2):4-22.

44. Bunds DG, Lloyd PJ. A multilayer perceptron netwotk for the diagnosis of low back pain. *IEEE Int Conf Neur Net* ,1988;2:481-489.

45. Cybenko G. Approximation by superpositions of a sigmoidal function. *Math Control Signals Syst* 1989;2(4):303-314.

46. Carpenter WC. Understanding Neural network approximations and polynomial approximations helps neural network performance. *AI Expert March* 1995;31-33.

47. Lehman A. *JMP for basic univariate and multivariate statistics: a step-by-step guide*, SAS Institute; 2005.

48. Filoti DI, Shire SJ, Yadav S, Laue TM. Comparative study of analytical techniques for determining protein charge. J Pharm Sci. 2015 Jul;104(7):2123-31.

49. Ruiz-Blanco YB, Paz W, Green J, Marrero-Ponce Y. ProtDCal : A program to compute general-purpose - numerical descriptors for sequences and 3D-structures of proteins. BMC Bioinformatics. 2015 May 16;16:162.

50. Box G. Science and statistic. *J Am Stat Assoc* 1976 Apr 05;791:799

51. Guidotti R, Monreale A, Ruggieri S, Turini F, Giannotti F, Pedreschi Dino. A survey of methods for explaining black box models. ACM computing surveys (CSUR), 2018, 51(5): 93.

52. Pratt C, Cornely K. *Essential Biochemistry, 3rd ed*, Wiley; 2004.

53. Vogt W. Oxidation of methionyl residues in proteins: tools, targets, and reversal. *Free Radic Biol Med* 1995;18(1):93-105.

54. Devanaboyina SC, Lynch SM, Ober RJ, Ram S, Kim D, Puig-Canto A, Breen S, Kasturirangan S, Fowler S, Peng L, Zhong H, Jermutus L, Wu H, Webster C, Ward ES, Gao C. The effect of pH dependence of antibody-antigen interactions on subcellular trafficking dynamics. *MAbs* 2013;5(6):851-9.

55. Israelachvili JN. *Intermolecular and surface forces, 3rd ed,* Elsevier; 2011.

56. Nicoud L, Owczarz M, Arosio P, Morbidelli M. A multiscale view of therapeutic protein aggregation: A colloid science perspective. Biotechnol J 2015;10(3)10:367-78.

57. Laue T. Proximity energies: a framework for understanding concentrated solutions. *J Mol Recognit* 2012;25(3):165-73

58. Schroeter C, Guenther R, Rhiel L, Becker S, Toleikis L, Doerner A, Becker J, Schoenemann A, Nasu D, Neuteboom B, Kolmar H, Hock B. A generic approach to engineer antibody pH-switches using combinatorial histidine scanning libraries and yeast display. *MAbs* 2015;7(1):138-51.

59. Chaparro-Riggers J, Liang H, DeVay RM, Bai L, Sutton JE, Chen W, Geng T, Lindquist K, Casas MG, Boustany LM, Brown CL, Chabot J, Gomes B, Garzone P, Rossi A, Strop P, Shelton D, Pons J, Rajpal A. Increasing serum half-life and extending cholesterol lowering in vivo by engineering antibody with pH-sensitive binding to PCSK9. *J Biol Chem* 2012;287(14):11090-7.

60. Gera N, Hill AB, White DP, Carbonell RG, Rao BM. Design of pH sensitive binding proteins from the hyperthermophilic Sso7d scaffold. *PLoS One* 2012;7(11):e48928.

61. Igawa T, Ishii S, Tachibana T, Maeda A, Higuchi Y, Shimaoka S, Moriyama C, Watanabe T, Takubo R, Doi Y, Wakabayashi T, Hayasaka A, Kadono S, Miyazaki T, Haraya K, Sekimori Y, Kojima T, Nabuchi Y, Aso Y, Kawabe Y, Hattori K. Antibody recycling by engineered pH-dependent antigen binding improves the duration of antigen neutralization. *Nat Biotechnol* 2010;28(11):1203-7.

62. Kulkarni MV, Tettamanzi MC, Murphy JW, Keeler C, Myszka DG, Chayen NE, Lolis EJ, Hodsdon ME. Two independent histidines, one in human prolactin and one in its receptor, are critical for pH-dependent receptor recognition and activation. *J Biol Chem* 2010;285(49):38524-33.

63. Maeda K1 Kato Y, Sugiyama Y. pH-dependent receptor/ligand dissociation as a determining factor for intracellular sorting of ligands for epidermal growth factor receptors in rat hepatocytes. *J Control Release* 2002 Jul 18;82(1):71-82.

64. Roopenian DC, Akilesh S. FcRn: the neonatal Fc receptor comes of age. *Nat Rev Immunol* 2007;7(9):715-25.

65. Tesar DB, Bjoerkman PJ. An intracellular traffic jam: Fc receptor-mediated transport of immunoglobulin G. *Curr Opin Struct Biol* 2010;20(2):226-33.

66. Ildiko FE, Friedman JH. A statistical view of some chemometrics regression tools. *Technometrics* 1993;35(2):109-135.

8 FIGURES

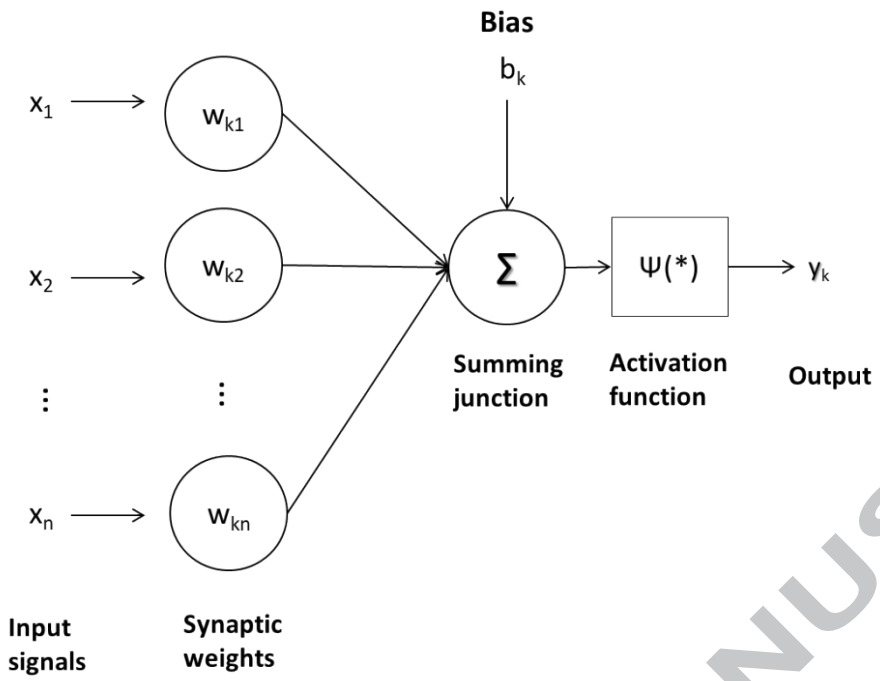

Figure 1. Model of a neuron. x_n represent the inputs connected to the neuron, k, by the weights, w_{kn}, which multiply the corresponding input signal. All the weighted signals are summed by a summing junction Σ. An external bias b_k can be applied to Σ, to increase or lower the output signal. Finally, Σ is connected to an activation function, ψ(*), which limits the amplitude of a signal to the output, y_k. Picture modified from: Neural networks: a comprehensive foundation, S. Haykin.[45]

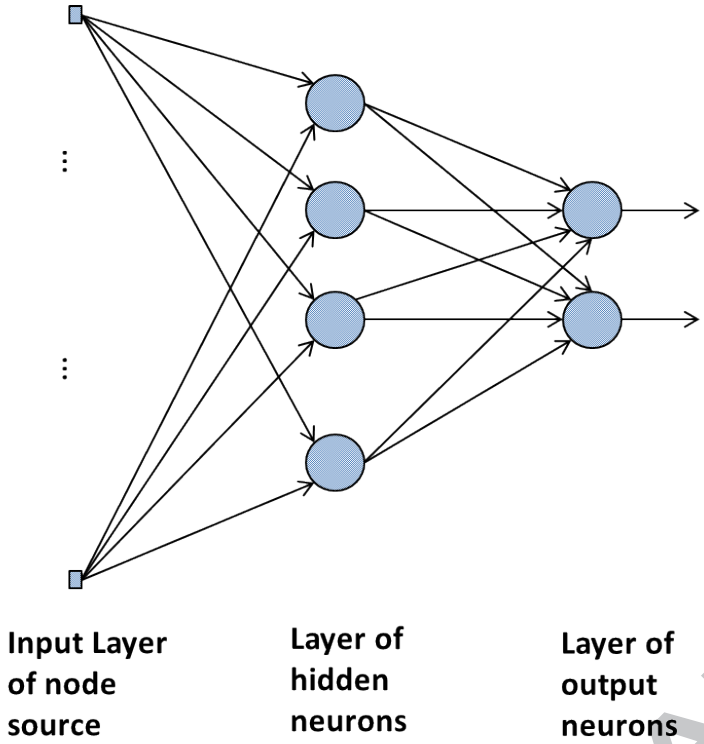

Figure 2. Signal-flow graph of a fully connected feedforward network with one hidden layer and one output layer. The signal-flow graph provides a neat description of the neural networks describing the links between the various nodes of the model. Picture adapted from: Neural networks: a comprehensive foundation, S. Haykin.[45]

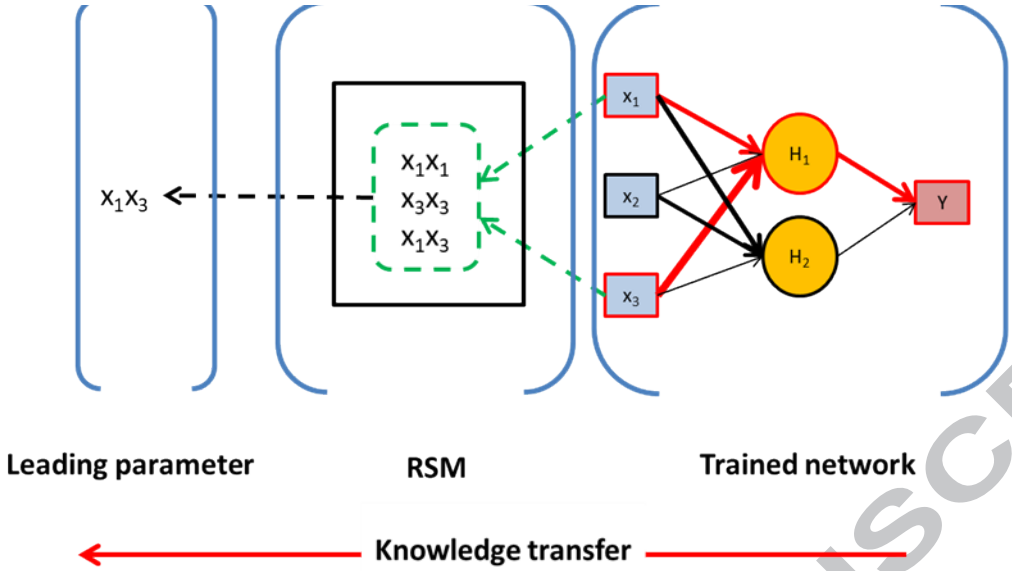

Figure 3. Scheme of the knowledge transfer procedure. On a trained network, where the arrow thickness represents the weight value (i.e. smaller arrow present lower weights), the input parameters with the higher impact, in red, are selected. These inputs are used for a least square linear regression where the RSM is applied considering only two-way interactions. From the analysis, leading parameters are selected and discussed to interpret the network decision-making process.

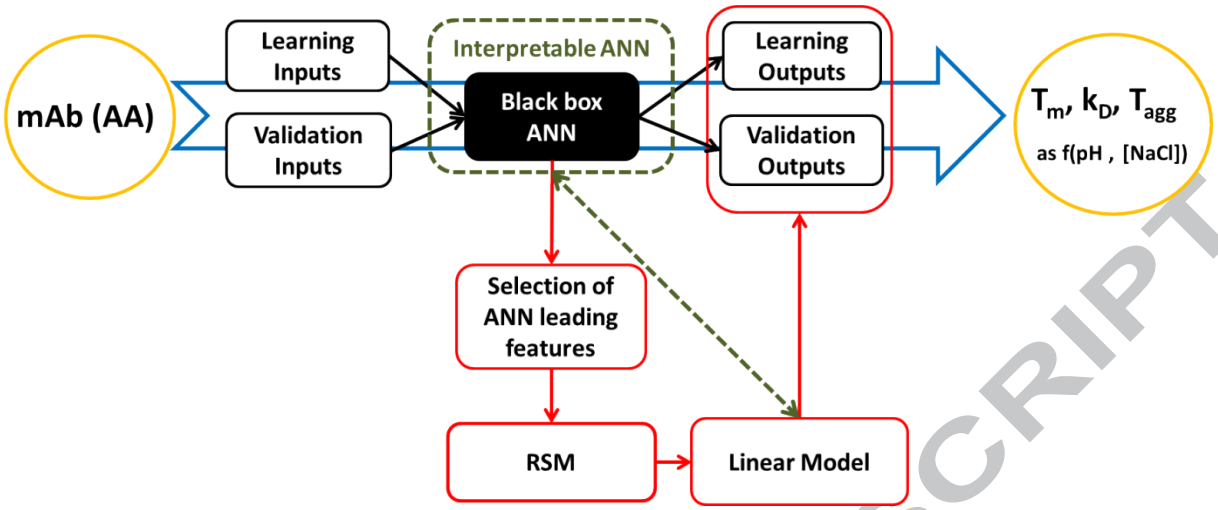

Figure 4. Diagram describing the process applied to achieve an interpretable prediction by ANNs. The knowledge transfer process is highlighted in red. The model explanation (dashed green lines) is aimed at understanding the overall logic behind the black box. Once trained and validated the interpretable ANN can be applied to new mAb candidates, even before cell expression. This allows to predict important biophysical parameters (i.e. T_m, k_D and T_{agg}) as a function of pH and salt concentration.

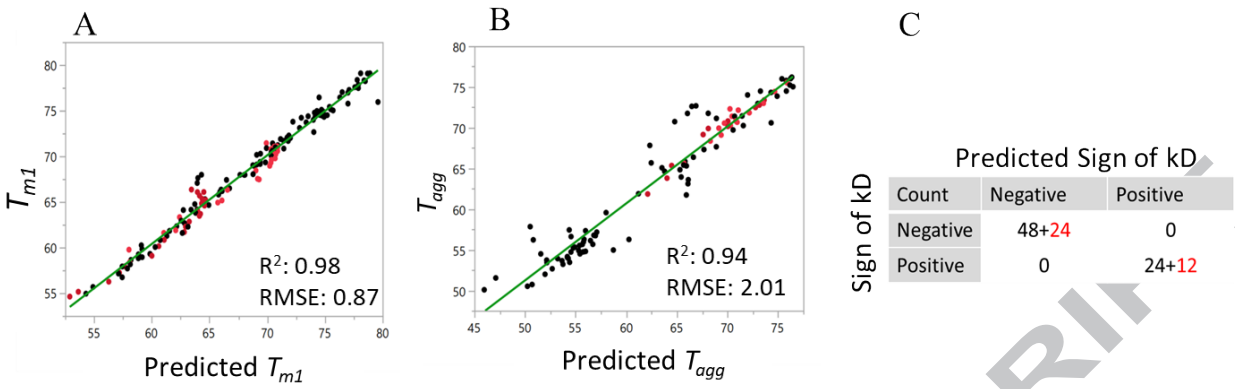

Figure 5. Results from PPI-13&3 models for the prediction of T_m, T_{agg} and the sign of k_D are shown in graphs A, B and C respectively. Black dots and numbers represent the training set, while red dots and numbers represent the validation set.

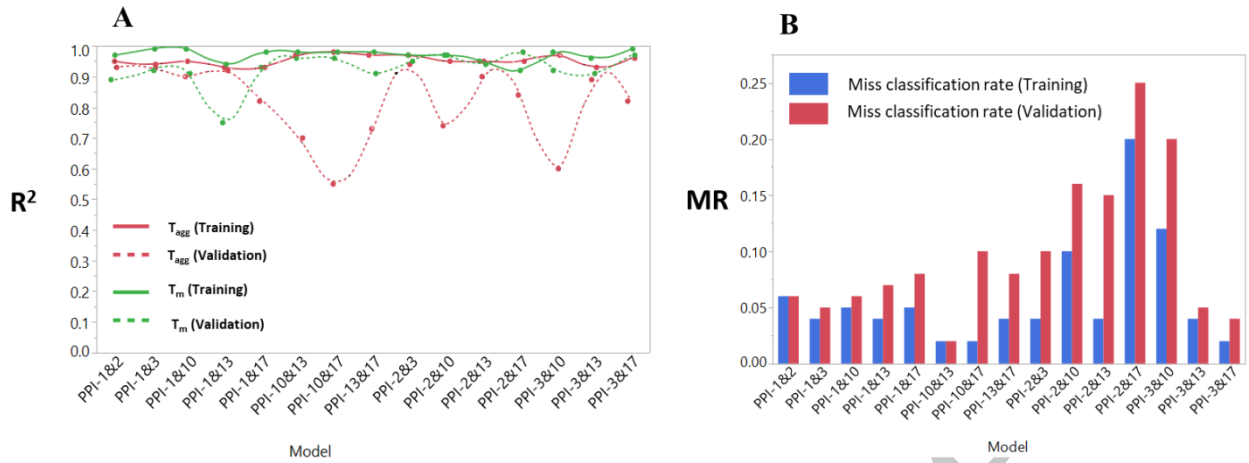

Figure 6. ANN robustness study of ANNs. In graph A, the R^2 values for the T_m and T_{agg} models are shown. In graph B, the misclassification rate (MR) of the sign of k_D models are shown. Blue bars represent the validation set while red bars represent the validation set. The models were classified by the name of the proteins used for the validation.

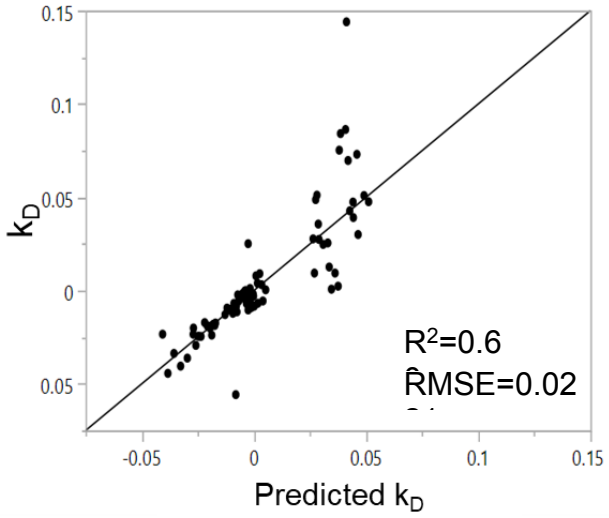

Figure 7. Correlation between experimentally determined and predicted k_D values for the PPI-13&3 model.

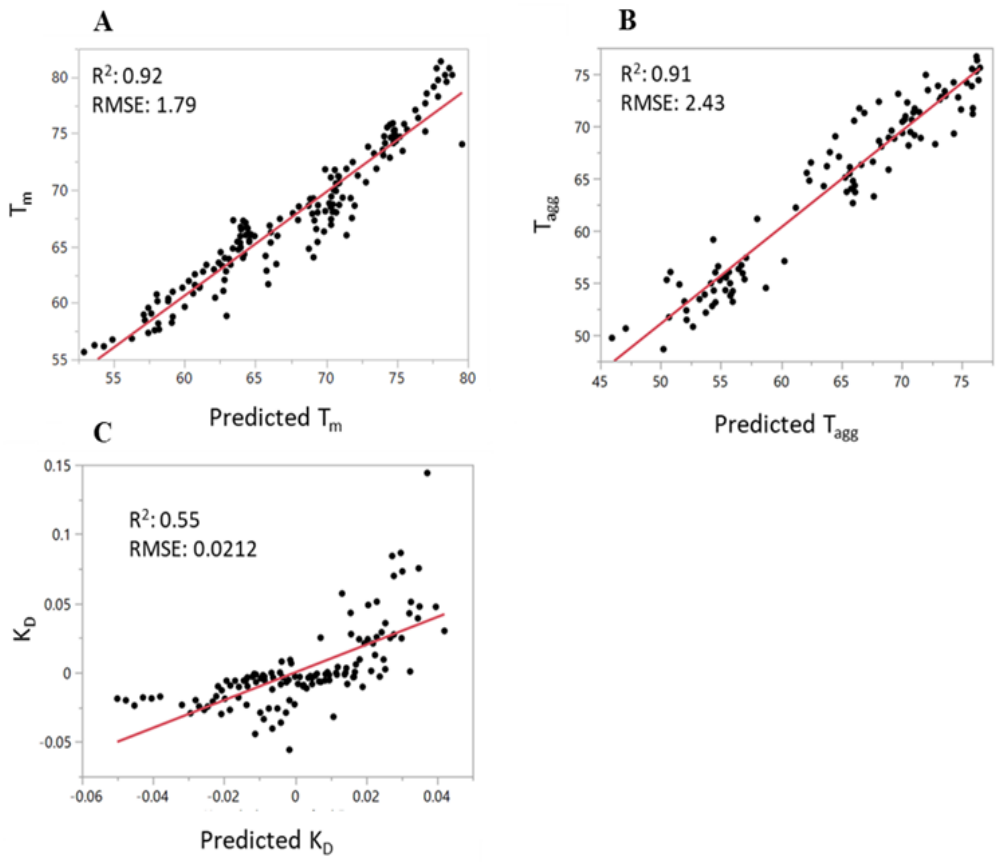

Figure 8. Results of T_m, T_{agg}, k_D linear models from the network knowledge transfer are shown respectively in graph A, B and C. The 3 graphs are generated by RSM using the selected leading parameter. The relative effect test is presented in Table 1.

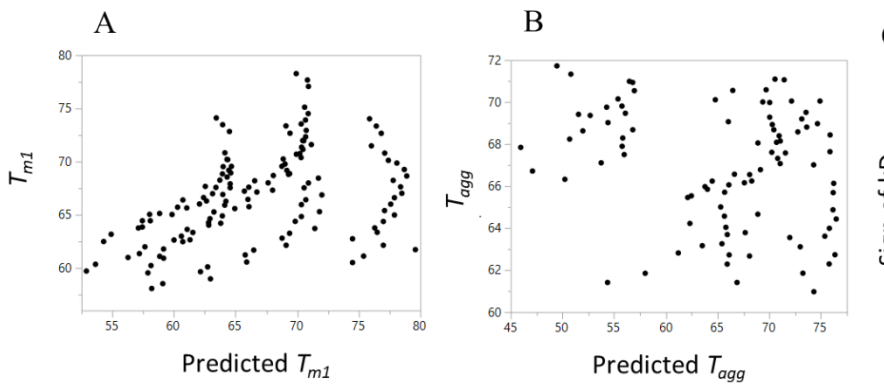

Figure 9. Results of the validation sets from the PLS model of T_m, T_{agg} and the sign of k_D are shown respectively in graphs A, B and C. The graphs show that the models cannot accurately predict protein properties that were not involved in the training set.

9 TABLES

Table 1. Effect tests results of the RSM applied to the linear least square regression from the knowledge transfer of ANNs' models. In Fig. 6 the relative graphs are shown. Information on the inputs can be found in Table S6. The quadratic terms (e.g. Cys·Cys) and the cross terms (e.g. pH·Cys) from the RSM were selected by reducing the full model using a backward stepwise regression where a value of p<0.05 is deemed statistically significant. LogWorth is defined as -log10(p-value).

T_m		T_{agg}		k_D	
Input	LogWorth	Input	LogWorth	Input	LogWorth
Trp	27.942	Glu	36.173	[NaCl]	11.608
pH	25.425	Met·Met	26.675	Glu	9.529
pH·Cys	13.701	Met	19.023	Trp	9.151
pH·pH	13.256	Asp	6.996	His	8.828
Cys·Cys	8.528	pH	6.084	pH	2.490
Cys	4.024	pH·pH	4.881		
Tyr·Tyr	3.813	Asp·Asp	4.199		
Tyr	3.284	[NaCl]	2.474		
[NaCl]	2.753				

10 SUPPORTIVE INFORMATION

Tables S1. Identity and similarity scores, respectively in red and yellow cells, from the primary sequences of the heavy chains, light chains, and the complete mAb with the relative statistics. The similarity is considered as: GAVLI, FYW, CM, ST, KRH, DENQ, P, where the single letter represents the standard single letter amino acid code. The identity scores were calculate by the Sequence Manipulation Suite (Stothard P (2000) The Sequence Manipulation Suite: JavaScript programs for analyzing and formatting protein and DNA sequences. Biotechniques 28:1102-1104).

Score Legend: Similarity identity

Heavy Chain (HC)

PPI-1	100%	15.36%	88.69%	29.94%	17.29%	23.09%
PPI-2	8.46%	100%	16.70%	14.69%	25.16%	42.15%
PPI-3	86.25%	10.24%	100%	32.73%	17.73%	29.14%
PPI-10	18.04%	9.27%	27.39%	100%	16.03%	29.14%
PPI-13	7.98%	18.48%	8.86%	9.35%	100%	16.14%
PPI-17	15.47%	37.21%	23.76%	23.31%	11.21%	100%
	PPI-1	PPI-2	PPI-3	PPI-10	PPI-13	PPI-17

Light chain (LC)

PPI-1	100%	11.73%	13.08%	13.55%	12.61%	11.62%
PPI-2	8.45%	100%	23.94%	23.00%	23.94%	24.88%
PPI-3	7.94%	18.30%	100%	94.39%	95.79%	48.59%
PPI-10	8.41%	18.43%	91.58%	100%	94.85%	44.85%
PPI-13	7.94%	18.77%	92.05%	93.92%	100%	45.79%
PPI-17	7.90%	16.43%	44.39%	38.78%	38.78%	100%
	PPI-1	PPI-2	PPI-3	PPI-10	PPI-13	PPI-17

mAb

PPI-1	100%	14%	51%	22%	15%	17%
PPI-2	8.46%	100%	20%	19%	25%	34%
PPI-3	47.10%	14.27%	100%	64%	57%	39%
PPI-10	13.23%	13.85%	59.49%	100%	55%	37%
PPI-13	7.96%	18.63%	50.46%	51.64%	100%	31%
PPI-17	11.69%	26.82%	34.08%	31.05%	25.00%	100%
	PPI-1	PPI-2	PPI-3	PPI-10	PPI-13	PPI-17

Statistic	HC	LC	mAb	HC	LC	mAb
Minimum	7.98%	7.90%	7.96%	15%	12%	14%
Maximum	86.25%	93.92%	59.49%	89%	96%	64%

Mean	21.02%	34.14%	27.58%	28%	39%	28%
Std deviation	19%	31%	17%	18%	31%	16%
Variance	4%	10%	3%	4%	10%	3%

Table S2. List of the input parameters with corresponding statistics. Input considered as discrete are only listed and no statistics is applied. To the right it is highlighted if the input is implemented to predict the corresponding protein stability indicator.

Input parameters relative to the mAbs						
Amino acid	**Code**	**Minimum**	**Maximum**	**Standard deviation**	**Variance**	**Mean**
Alanine	Ala	64	80	5.62	31.56	69.33
Cysteine	Cys	30	38	2.75	7.56	32.67
Aspartic acid	Asp	52	62	3.54	12.56	54.33
Glutamic Acid	Glu	58	68	3.77	14.22	62.67
Phenylalanine	Phe	38	54	5.22	27.22	45.67
Glycine	Gly	82	98	5.63	31.67	91.00
Histidine	His	18	26	2.75	7.56	23.33
Isoleucine	Ile	28	36	2.52	6.33	31.00
Lysine	Lys	76	96	6.30	39.67	89.00
Glutamine	Glu	88	108	6.26	39.22	97.67
Methionine	Met	8	16	3.06	9.33	12.00
Asparagine	Asn	44	52	2.69	7.22	48.33
Proline	Pro	88	106	5.85	34.22	94.67
Glutammine	Gln	54	66	4.23	17.89	59.67
Arginine	Arg	30	50	6.43	41.33	38.00
Serine	Ser	158	188	10.13	102.67	172.00
Threonine	Thr	98	120	7.61	57.89	109.67
Valine	Val	110	120	3.14	9.89	115.67
Tryptophan	Trp	20	26	2.24	5.00	23.00
Tyrosine	Tyr	52	64	4.27	18.22	58.67
Input parameters relative to the formulation	**List**					
pH	5, 5.5, 6, 6.5, 7, 7.5, 8, 9	-	-	-	-	-
[NaCl] (mM)	0, 70 , 140	-	-	-	-	-

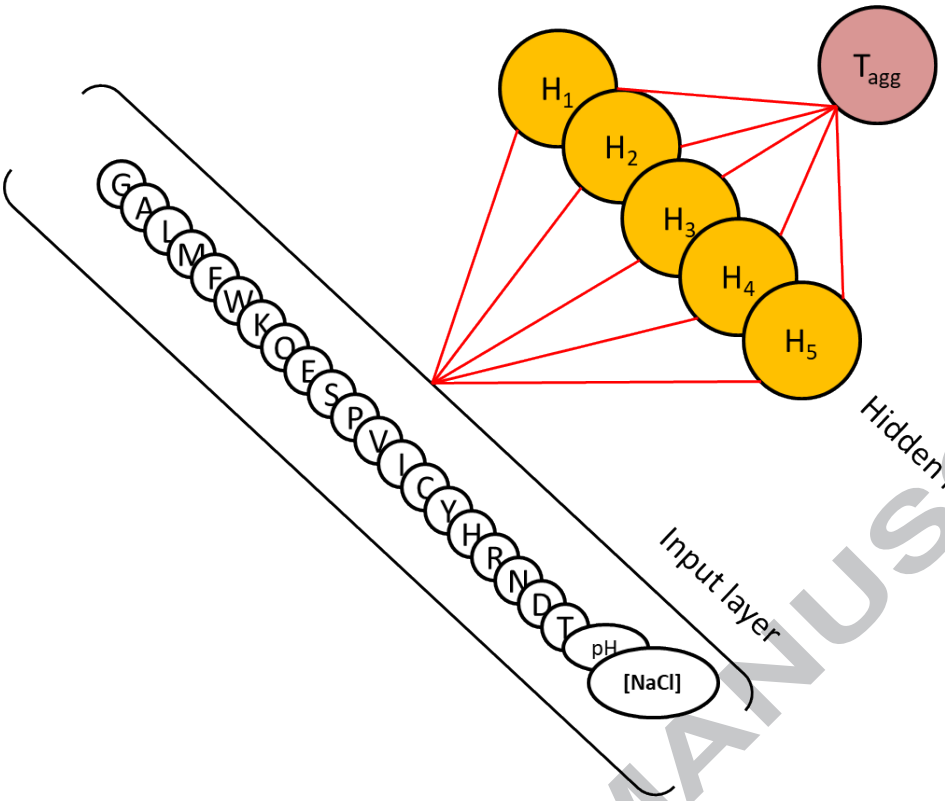

Figure S3. Exemplary picture of applied network architecture. The brackets containing the input layer represent a complete connection of the input layer with the hidden one (i.e. each input is connected with all the neurons of the hidden layer).

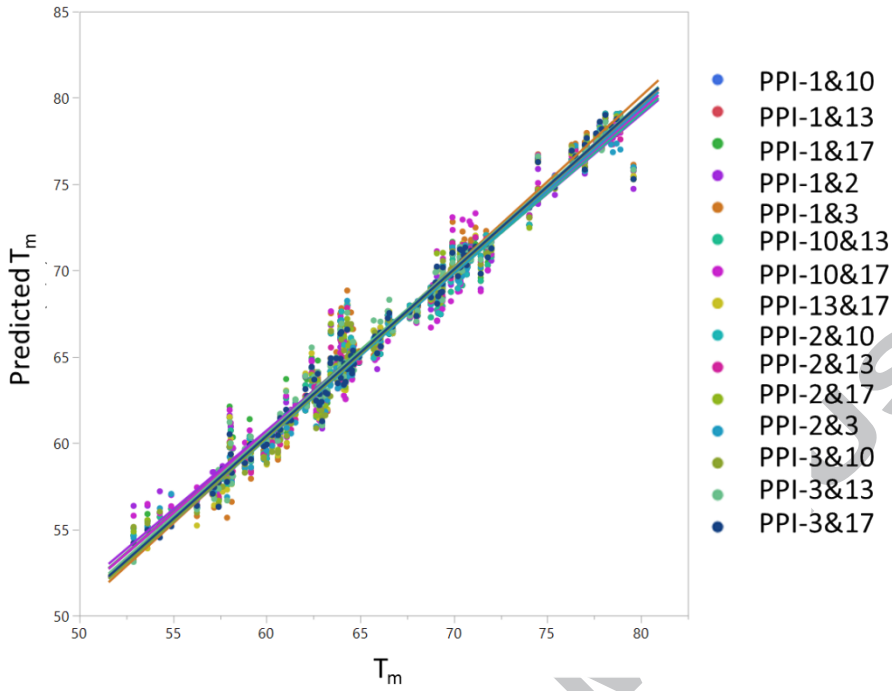

Figure S4. ANNs' T_m models results of the 15 different training sets.

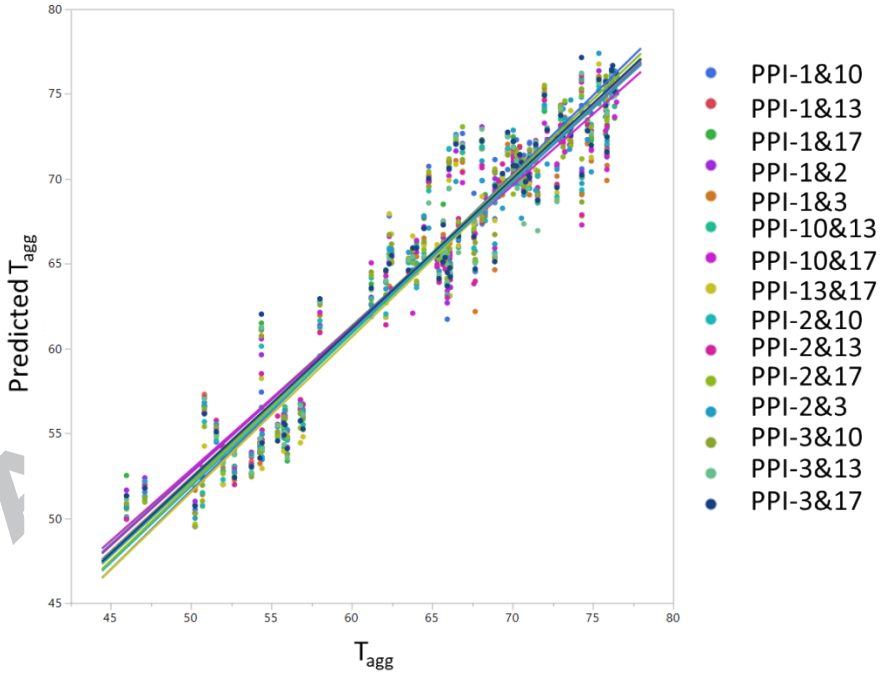

Figure S5. ANNs' T_{agg} models results of the 15 different training sets.

Table SI 6. List of the molecular descriptor calculated by ProDCal. The description of the molecular indices can be found in the relative software manual.

ProtDCal	ProtDCal
dGc(F)	wRWCO
dGw(F)	wdHBd
Gs(F)	wLCO
W(F)	wCo
HBd	wFLC
dGs	wPsiH
dGw	wPsiS
dGel	wPSil
dGLJ	Psi
dGtor	wR2
Gs(U)	wPjiH
Gw(U)	wPhiS
W(U)	wPhil
Mw	Phi
Ap	LnFD
Ecl	wCLQ
HP	wCTP
IP	wSP
ISA	WNc
Pa	Ap
Pb	dA
Pa	dAnp
Pt	WNLC
z1	wFLC
z2	wR2
z3	lnFD
dHf	
Xi	
L1-9	

GRAPHICAL ABSTRACT