Artificial neural networks for solution scattering data analysis

D. Molodenskiy1, D. Svergun1\* and A. Kikhney1\*

1 European Molecular Biology Laboratory, Hamburg Outstation, EMBL c/o DESY, Notkestrasse 85, D-22607 Hamburg, Germany

Corresponding authors: [svergun@embl-hamburg.de](mailto:svergun@embl-hamburg.de)

[a.kikhney@embl-hamburg.de](mailto:a.kikhney@embl-hamburg.de)

# Abstract

Small angle X-ray scattering (SAXS) experiments are widely used for the characterization of biological macromolecules in solution. We propose a novel method for primary SAXS data analysis based on the application of perceptrons - interconnected artificial neural networks. Trained on synthetic SAXS data, the neural networks are able to predict molecular weight and maximum intraparticle distance (Dmax) of previously unseen experimental data. The application area is of a wide range and includes data from compact proteins of various sizes and shapes, as well as nucleic acids (DNA/RNA) and intrinsically disordered proteins. The area of application of the neural network models is not limited by the described objects and can be easily enhanced by further extension or augmentation of the training set. The method was rigorously tested against synthetic SAXS data and demonstrated higher accuracy and robustness against experimental noise compared to all available conventional methods. The method was implemented as a publicly available web service with a graphical interface (<https://dara.embl-hamburg.de/gnnom.php>).

# Introduction

Small-angle scattering (SAS) of X-rays and neutrons from biological macromolecules in solution is a powerful tool, providing information on supermolecular structures and dynamics under a wide range of conditions (Feigin et al., 1987; Gräwert and Svergun, 2020; Guinier and Fournet, 1955). Due to relatively soft requirements to sample preparation for SAS experiments and huge progress in the development of data analysis software (Hopkins et al., 2017; Liu et al., 2012; Manalastas-Cantos et al., 2021), the technique became high throughput and available even to the non-specialists in the area. The molecular weight (MW) and maximum intraparticle distance (Dmax) are the crucial basic parameters derived from SAS experiment. However, its precise estimation still remains a non-trivial task with higher potential uncertainty related to the lower quality of experimental data.

Recently, the application of neural networks (NN) has experienced a sudden leap virtually in all areas of life, due in no small part to the development of deep learning technologies (Schmidhuber, 2015). Huge progress has happened in many biological applications as well, including bioinformatics (Armenteros et al., 2019), a recent breakthrough in *in silico* protein folding (Senior et al.), and even in the SAXS area (He et al., 2020). The application of deep NNs becomes clearly beneficial when the amount of data used for training is so big, that the networks may be able to recognize the hidden from the human eye patterns in the data. The major advantage of using supervised machine learning over the conventional methods is the possibility to augment the training set data, thus easily adjusting the area of applicability of a given model.

From the SAXS perspective, making a robust against experimental noise NN model is as simple as generating a realistic experimental noise and augmenting the training set accordingly. Similarly one can overcome other common shortages of SAXS data collection, such as buffer sub/over subtraction, presence of a systematic beamline noise, buffer mismatch, etc. Along with that, the machine learning-driven approaches lack the limitations of a chosen approximation (e.g. the homogeneity of a model) and may find previously unrecognized patterns and connections between SAXS data and the derived parameters.

Inspired by the huge progress in the field, we developed a novel method of SAXS data analysis based on state-of-the-art machine learning principles. We employed a stack of interconnected artificial NNs trained on synthetic SAXS data generated from thousands of models from PDB databank (Berman et al., 2000) to perform principal data analysis. The training set was chosen employing the supercube latin sampling method (McKay et al., 1979) to be as complete as possible and to make the trained networks generally applicable.

The presented method appears to be the only one for concentration-independent estimation of MW for intrinsically disordered proteins (IDP), as well as the second one for the nucleic acids (NA). Moreover, the application of NNs gear the only human independent way of Dmax calculation, based solely on the one-to-one correspondence (SAXS profile – Dmax), without the introduction of unnecessary regularization parameters.

We further investigate the information content of the different angular ranges for the determination of MW and Dmax by training/testing the NNs on truncated over s data sets. Finally, we demonstrate the higher accuracy and better robustness of our method against simulated experimental noise over the conventional methods. We would like to stress, that this work aims to demonstrate the capabilities of NNs in the field of SAXS, whereas further improvement in accuracy, as well as applicability extension, is possible and limited only by the choice of the training set.

The method was implemented as a publicly available web service with a graphical interface (<https://dara.embl-hamburg.de/gnnom.php>).

# On limitations of conventional methods.

The MW is a crucial parameter in SAXS that can be exploited for e.g. determination of oligomeric state. However, the currently available methods are not very accurate, a usual rule of thumb is 10% or more. To date, there are several well-established techniques for MW estimation, that nominally could be subdivided into concentration-dependent and concentration-independent methods. Former methods account for the dependence of forwarding scattering I(0) on the total number of electrons in the irradiated molecule (and, thus, on MW) and rely on the scattering from calibrants, e.g. water or a protein with known MW (Mylonas and Svergun, 2007). The latter methods utilize a single background-subtracted curve and require no additional *a priori* information on sample concentration.

**Porod’s method.** The historically first concentration-independent method is called the Porod method. It is based on the fundamental properties of the Fourier transform known as the Parseval theorem:

(1)

, where s is the scattering vector, I(s) – the intensity of the scattering beam, Δρ – excessive electron density, and Q is the Porod invariant. In the case of homogeneous electron density approximation, the right part of the equation simplifies to:

(2)

and given that intensity in the origin I(0) = (Δρ)2V2:

(3)

The MW is estimated as an empirical relation between the volume of the particle and its mass, which equals 1.6 e.g. in the case of proteins (Petoukhov et al., 2012). Therefore the precise calculation is limited by the three governing factors: (i) integration in (1) can not be performed due to limitations in experimental s-range, so I ~ s-4 power law is usually used to extrapolate the intensities on higher angles; (ii) integration is affected by the experimental noise; and (iii) the equation (2) implies homogeneity of the scattering particle. This method requires accurate knowledge of electron density, works only on proteins, and very sensitive to subtraction errors.

**Fisher’s method.** The further improvement is possible assuming the Guinier approximation at low s < smin and calculating the truncated integral up to smax (Fischer et al., 2010):

(4)

The authors introduce the apparent volume as , similarly with (3), and establish a linear connection between V and V’:

V = A + BV’ (5)

With linear and angular coefficients A and B are determined empirically for different smax values. The drawbacks of this method are essentially the same as for Porod’s method.

**Volume of correlation.** Another approach was developed by Rambo&Tainer (Rambo and Tainer, 2013) and introduces the so-called volume of correlation:

(6)

The authors found an empirical dependence between Vcand the molecular weight:

(7)

Where c and k are empirically determined constants via fitting results from theoretical scattering profiles. This is the only available method for estimating the MW of RNA, with differently determined constants c and k. However, this method is reported to be less accurate compared to the others for high signal-to-noise data, as well as for extended and small (<20 kDa) particles.

**Shape&Size.** The first machine learning-based method for MW determination was developed by (Franke et al., 2018) and represents a classical k-Nearest-Neighbor classifier. In this method, the Fisher’s truncated integral Q’ (eq.4) is calculated up to sRg = 3,4,5 for a huge set of simulated SAXS patterns from geometrical bodies, as well as from the PDB models. The obtained three numbers can be treated as coordinates in some 3D space resulting in thousands of points, where each point corresponds to a model. The experimental data can also be mapped on the same space and since each model in this space has known MW and Dmax, these values can be estimated as the weighted average between the nearest k-neighbors. This approach has the advantage of taking into account the shape of particles while computing MW and Dmax, alongside its size. The major drawback of the method is that it does not work for nucleic acids and flexible proteins.

**Bayesian statistics.** In the recent method (Hajizadeh et al., 2018) the authors calculate an MW using Bayesian inference with the MW calculations from all the above-mentioned methods as the evidence. The authors simulate a large test dataset of SAXS profiles, then calculate the MW for each curve using each method to build a probability distribution, that describes the original probability of obtaining a particular calculated MW given the true molecular weight. These probabilities are combined across all the methods, and the most likely molecular weight is thus estimated. The advantage of the method is that it employes all the other methods and provides not only the most probable MW, but also assesses its credibility interval. The disadvantage is similar to the Shape&Size method: it works only for compact proteins.

**Maximum intraparticle distance Dmax.** To date, there are only two available methods for an estimate the Dmax: Shape&Size and through the indirect Fourier transform (IFT). The principles of the former were described earlier, whereas the latter requires the introduction of the pair distance distribution function p(r). The p(r) function represents a histogram of distances between pairs of points in the particle, weighted by the product of their scattering contrasts (Guinier and Fournet, 1955). Mathematically, the p(r) function is closely related to the scattering intensity versus the momentum transfer I(s) via the spherically averaged Fourier transform (Debye, 1915):

(8)

(9)

The limited angular range of discretely recorded experimental data, as well as the presence of experimental noise, makes the evaluation of p(r) an ill-posed problem. The method of solving this problem by the IFT has been originally proposed by Glatter (Glatter, 1977), and further enhanced by Svergun (Svergun, 1992) and Hansen (Vestergaard and Hansen, 2006). In the IFT approach a guess on the Dmax must be given, the p(r) function is expressed as a sum of some analytical functions (e.g. cubic splines), and a classical regularization procedure (TIKHONOV and N, 1943) is applied such that p(r) i) agrees to experimental data and ii) ensures satisfaction to the imposed constraints. Most commonly the constraint is the smoothness of p(r), so that termination effects are reduced or ideally completely removed. However, in all of these approaches, the choice of the final solution remains a subjective criterion left to the discretion of the user. Therefore, small deviations of Dmax are acceptable and do not change the final solution, thus precise estimation of Dmax is left to the discretion of a user.

# Methods

Neural networks architecture. We exploited Keras and Tensorflow modules in the python framework to construct interconnected neural networks (perceptrons) with three hidden layers for predicting MW and Dmax (fig.1 demonstrates training procedure for the determination of MW). Since the expected output of the NNs is a numeric score and not a probability, we encounter a classical regression task of supervised learning.



80 units

1 unit

256 units

Fig.1 On the architecture of neural networks used for primary SAXS data analysis.

According to the universal approximation theorem (see e.g. (Cybenko, 1989)), a feed-forward artificial NN with a single hidden layer containing a finite number of neurons can approximate any continuous function, under mild assumption on the non-linearity of the activation function. In particular, it was recently shown (Lu et al.) that NNs utilizing the rectified Linear Unit (ReLU) activation function with a width (number of neurons in one layer) of n+1 is capable to approximate any continuous convex function of n-dimensional input variables to any desired degree of precision (Hanin, 2019). In SAXS one typically analyzes a hugely oversampled curve, that in fact contains only up to 15-35 Shannon channels (Moore(crystallography and 1980). Therefore, in the context of SAS with an angular range smax < 5 nm-1 it should be sufficient to use a NN with only one hidden layer and the width of ~ 35 neurons.

In practice, even though this model was able to work quite well on our data set, we have found that using a somewhat excessive architecture with 3 hidden layers and 80 units leads to faster convergence and more stable solutions, as well as to the slightly better prediction results. During extensive testing, we have also found, that using hyperbolic tangent as an activation function instead of ReLU also marginally improves the results. The possible explanation for that is probably connected to the range of tanh (-1, +1), thus allowing to pass negative signals in between the layers and pushing more neurons to be activated. The hidden layers were initialized by the ‘he-uniform’ random function with zero biases, whereas the bias for the output layer was initialized as an averaged predicted parameter (MW or Dmax) over the whole training set.

Training set. Proteins.To train the neural networks, we used the models of solved deposited proteins and nucleotides from the worldwide protein databank PDB for simulation ([www.rcsb.org](http://www.rcsb.org)) (Berman et al.). Generally, for the better performance of NNs it is recommended to prepare a maximally unbiased training set, ideally evenly distributed over the parameters for prediction. However, as is seen from the histogram in fig.2, the deposited PDB molecules are severely skewed towards small and globular proteins.



(a)

(b)

(a)

(b)

*numbers*

*numbers*



(c)



(d)

Fig.2. Distribution over the radius of gyration and MW of (a) all deposited in PDB proteins; (b) sampled for the training set by the Latin supercube sampling (c) MW vs Rg heat map for the 135 238 proteins (d) for the chosen training set.

It is intuitively clear, that the vast majority of the structures in PDB were solved by crystallographic methods, thus the overall distribution tends to be biassed towards small and globular proteins due to their ability to form crystals. On the contrarary, in SAXS one usually deals with relatively big proteins or their complexes, since SAXS is a low-resolution technique and is typically used for the determination of supramolecular structures. Preliminary experiments confirmed, that NNs trained on the whole PDB data works well for the smaller compact proteins, but underperforms for bigger and elongated models (the right-hand part of fig.2 (c)). These results, however commendable, would fall short of the SAXS community aspirations, as one of the strongest sides of the method is its ability to analyze unfolded and intrinsically disordered proteins.

Thus, we decided to amend the training set to make it as diverse and complete as possible. Among other techniques, the most successful one turned out to be a Latin hypercube sampling (McKay et al., 1979) – a statistical method for generating a most representative near-random sample of parameter values from a multidimensional space. In our case, the method was simplified to a classical 2D Latin square over two parameters – MW and radius of gyration (Rg) as a degree of protein compactness. In this way, after filtering outliers we picked ~7000 of models in two-dimensional space (MW, Rg) out of 135238 proteins available to date in PDB. As is demonstrated in fig.2, the chosen dataset has almost a step-like distribution across (MW, Rg) indicating drastically more diverse by size and shape models. The heteroatoms were removed from all models to remove binding ligands and other molecules. Then the data were randomly distributed as to 80% training, 10% validation, and 10% test sets.

For each model, a pair of (simulated SAXS curve – predicted parameter) was prepared using CRYSOL (Barberato et al., 1995), then the SAXS curves were augmented with the experimental noise and normalized on the I(0) = 1 and. It is worth noting, that additional normalization of the data such as using the other representations (e.g. Kratky plot or log I vs s), subtracting the average curve, and dividing by standard deviation does not bring significant improvements in the NNs performance. The experimental noise, simulated at 7 different protein concentrations c = 0.25, 05, 1, 2, 4, 8, 16 mg/ml, was generated based on experimental data from the p12 beamline (Blanchet et al.) and corresponds to the data acquired with the sample-to-detector distance of 1 meter, exposure time of 1 second, and X-ray energy of E = 10 keV. For MW and Dmax two separate NNs with similar architectures were trained.

Nucleic acids. The nucleic acid models (pure DNA/RNA, not heterocomplexes) are not that massively populated in PDB as protein models and comprise only less than 2% of all entries. To collect only non-redundant models, we used the NDB server (<http://ndbserver.rutgers.edu/>) (Coimbatore Narayanan et al., 2014). After preliminary filtering, we fetched ~3000 models and distributed them as 80%/10%/10% for training/validation/test sets. As is seen in fig.3, the models are also mostly small and compact with the majority populated in between MW of 10 and 20 kDa and Rg of 10 and 25 Ȧ. However, given the limited number of models and the fact, that according to the biggest SAXS database SASBDB (Kikhney et al., 2019) the majority of DNA/RNA models used in SAXS are within this interval, we decided to use those models as is without further shrinking.



6UES

(39 kDa)

2JYH

(28 kDa)

4KYY

(11 kDa)

3REC

(0.6 kDa)

1H1K 🡪

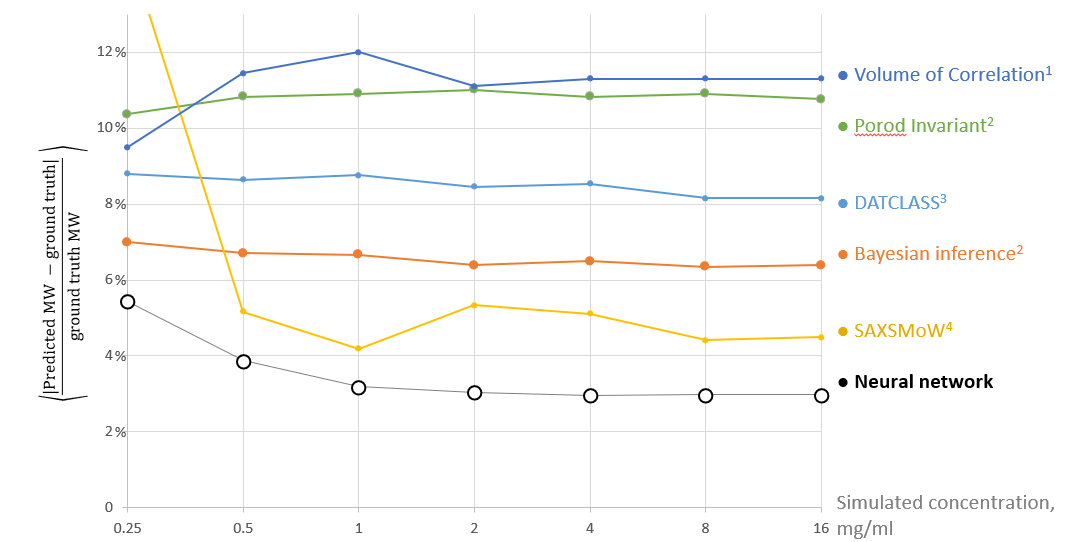
(605 kDa)

Fig.3. Training set from PDB for the nucleic acids

Intrinsically disordered proteins. For IDPs, we used the program RANCH and validated the results against models from the PED database… (Fig.4)

# Results and discussion

Benchmark. Comparison of accuracy of predicted SAXS parameters with the other available methods. Fig.5



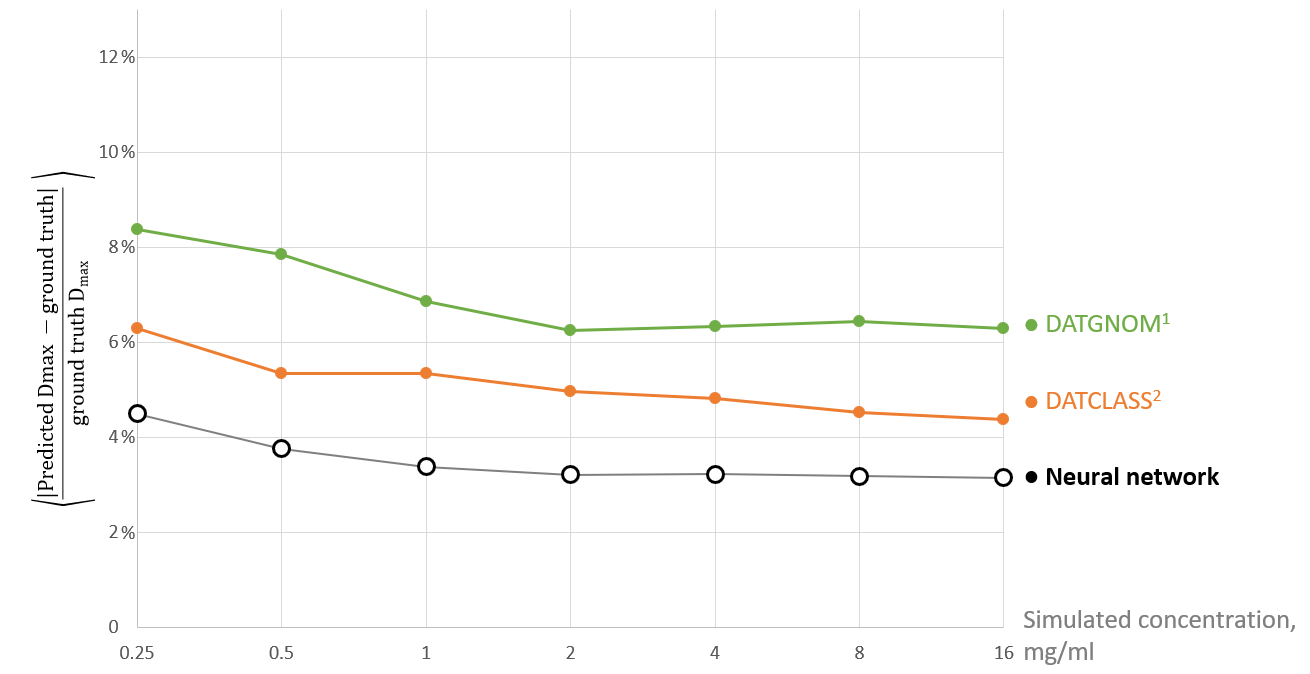


Fig.5 Comparison of performances of different methods (proteins, NAs, IDPs).

Information content. Lower bound. The usage of NNs offers the challenge to fundamentally estimate the accuracy of MW and Dmax estimations for ideal cases of noiseless SAXS profile, determined on a very wide s-range. To do so, we trained the NNs on smooth curves up to smax=10 nm-1 and applied them also on the smooth curves from the test set. The obtained result (Fig6) is not drastically different from what we obtained for the noisy data and equals 2.7% for MW and 3% for Dmax.



(b)

(a)

Fig.6. Predictions for “ideal” smooth datasets versus ground truth ((a) – MW, (b) – Dmax)

Interestingly, the highest deviations in MW were observed for big (> 360 kDa) proteins, potentially indicating the lack of such proteins in PDB, and consequently, in our training set.

Angular range. Another convenient opportunity that opens the usage of NNs is to get a deeper insight into the information content of the different angular ranges of SAXS curves and their impact in terms of MW and Dmax estimates. (Fig.7)

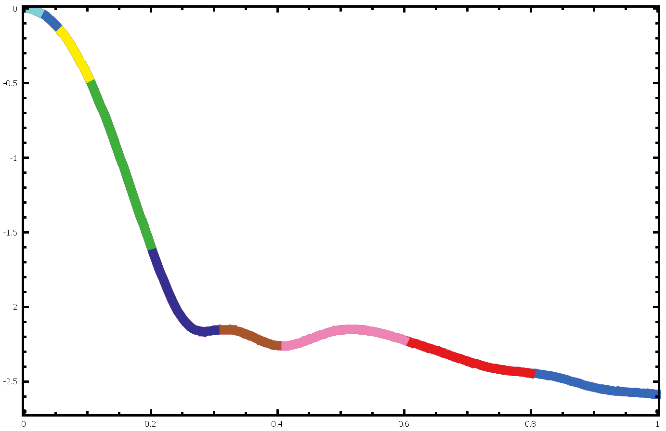


Fig.7 Error versus angular range for abs

It is seen from fig.4, that Dmax estimation requires predominantly low angles, and an increase of angular range after 6 nm-1 does not improve the predictions at all, whereas for MW the situation is different

**Web interface** (or we just say “to be included in the new release of ATSAS”??).

# Conclusion

A novel independent method for the estimation of primary SAXS parameters was developed. The comparison with well-established methods demonstrated higher accuracy and robustness of the method against the experimental noise.

# References

11

Cell / structure:

A basic guideline in terms of the length for our Resource and Review articles is 8000-8500 words (including references and figure legends), and up to 7 display items (figures + tables). Further details can be found at: <https://www.cell.com/structure/article-types>. If needed, we are happy to provide feedback to our authors on outlines and drafts.

When you are ready, you may submit your manuscript to our [online submission system](https://www.editorialmanager.com/structure/default.aspx). If you have any questions or concerns, please let me know and I’d be happy to assist you.