Artificial neural networks for solution scattering data analysis

D. Molodenskiy\*, D. Svergun\* and A. Kikhney\*

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

# 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 are of a wide range and include 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 model is not limited by the described objects and can be easily enhanced by extension or further 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 molecular 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 basic parameters derived from SAXS experiment. However its precise estimation still remains a non-trivial task with potential uncertainty closely related to the 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). A big progress has happened in many biological applications as well, including bioinformatics (Armenteros et al., 2019), recent break through in *in silico* protein folding (Senior et al.), and even in the SAXS area (He et al., 2020). The application of deep NNs becomes clerly beneficial when the amount of data used for training is so big, that the networks may be able to recognise 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 easily adjust the area of applicability of a model by augmentation of the training set data.

From the SAXS point of view, making a robust NN model against experimental noise is as simple as generating a realistic experimental noise and augment the training set accordingly. In the similar way 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 homogeneiuty of a model) and may find previously unrecognised 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 ANN trained on synthetic SAXS data 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 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 introduction of unnecessary regularization parameters. We further investigate the information content of the different angular ranges by training/applying the NNs on truncated data set. The aim of this work was to demonstrate the capability of NNs in the field of SAXS, whereas further extension of their applicability is possible and limited only by the chose of the training set.Finally, we demonstrate the higher accuracy and better robustness of our method against simulated experimental noise over the conventional methods.

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 to 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 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) – intensity of the scattering beam, Δρ – excessive electron density and Q is the Porod invariant. In case of homogeneous electron density approximation, the left part 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 particle and its mass, that equals 1.6 e.g. in 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 some power law is usually used to extrapolate the intensities on higher angles; (ii) integration is affected be 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 coeffecients A and B determined empirically for different smax values. The drawbacks for this method are essentially the same as for the 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 support vector machine approach. 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 resulting three numbers can be treated as coordinates in some 3D reference frame resulting to thousands of points (mapped models). The experimental data can be mapped on the same space and since each model in this space has known MW and Dmax, these values can be easily estimated as the weighted average between the neareast k-neighbours. This approache has an advantage of taking into account the shape of particles while computing MW and Dmax, alongside with 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 a 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 assess its credibility interval. The disadvantage is similar to the Shape&Size method: it works only for compact proteins.

**Maximum intraparticle distance Dmax.** The common and as a matter of fact the only available method to determine Dmax from SAS scattering profile is via inspection of corresponding p(r) function. Upon certain restrictions, such as non-negativity and zeroing p(0) = p(r ≥ Dmax) = 0, it is possible to find the Dmax as a first negative point of p(r) distribution. Importantly, small deviations of Dmax values typically do not significantly affect the fit to the data (inverse Fourier transform of p(r)), thus precise estimation of Dmax is left to the discretion of a human.

To the date, there are only two available methods for estimate the Dmax: Shape&Size and indirect Fourier transform. The principles of the former were described earlier, whereas the latter requires 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 [Glatter & Kratky, 1982]. Mathematically, the p(r) function is closely related to the scattering intensity versus the momentum transfer I(s) via well-known transformation [Debye, 1915]:

(1)

(2)

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. As estimation of Dmax, as well as reconstruction of a macromolecular model, is typically based on the p(r) function, therefore it is crucial to reliably and non-ambiguously compute p(r) from available experimental data. The method of solving this problem by so-called indirect Fourier transformation (IFT) has been originally proposed by Glatter [Glatter, 1977], and further enhanced by Svergun [Svergun, 1992] and Hansen [Vestergaard & Hansen, 2012]. In 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, 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 idially 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.

# Methods

Neural network architectures**.** We exploited interconnected neural networks (perceptrons) with one hidden layer (fig.2) for predicting the above-mentioned SAS parameters. Since the expected output of the NNs is a number and not a discrete value, we encounter a classical regression task for machine learning.



Fig.2 On the architectures of neural networks used for primary SAXS data analysis.

According to the Universal approximation theorem (see e.g. [G.Cybenko, 1989]), a feed-forward artificial network with a single hidden layer containing a finite number of neurons can approximate any continuous function, under mild assumption on the activation function. In particular, it was recently shown [Lu, Zhou, et al. "The expressive power of neural networks: A view from the width." *Advances in neural information processing systems*. 2017.] 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, 1980]. Therefore, in the context of SAS with an angular range smax < 5 nm-1 it is sufficient to use a NN with only one hidden layer and the width of ~40 neurons. The attempts to introduce more layers and neurons led us to increased learning time and higher instability without any gain in the NNs productivity.

Training set. In order to train the neural networks on the most realistic data set, we used the real models of deposited proteins and nucleotides from the worldwide protein databank PDB ([www.rcsb.org](http://www.rcsb.org)) for simulation. Generally, for the better performance of NNs it is recommended to prepare unbiased training set, ideally evenly distributed over the parameters for prediction. As is clearly seen from the histogram in fig.1, the deposited PDB molecules are skewed towards small and globular proteins.



(a)

(b)

*numbers*

*numbers*

Fig.1. Distribution of (a) all deposited in PDB proteins and (b) sampled proteins for the training set by their radii of gyration and molecular weights.

Intuitively, it is clear that the vast majority of the structures were solved by crystallographic methods, thus the overall distribution tends to be biassed towards small and globular proteins simply due to their ability to form a crystal. Preliminary computer experiments confirmed, that NNs trained on the whole amount of PDB data works well for the small and globular proteins and underperform for bigger and elongated models. 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 even intrinsically disordered proteins. Thus, we decided to expand 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, Beckman, et al, 1979] – a statistical method for generating a near-random sample of parameter values from a multidimensional space. In our case, we employed it to pick several thousands of models in two-dimensional space (MW, Rg) out of more than 150000 available to date in PDB. As is demonstrated in Fig.1, the chosen dataset of 1015 models has almost a step-like distribution across (MW,Rg) indicating drastically more diverse by size and shape models.

For each model, a pair of (simulated curve – predicted parameter) was prepared using CRYSOL [crysol paper] in order to apply classical supervised training of NNs. For each of the required parameters (MW, Dmax, p(r)) the dedicated NN was developed and trained. The calculation of p(r) from experimental data is typically fairly ambiguous operation due to the limited angular range of SAS data and the presence of experimental noise. The big advantage of using synthetic data for training NN is the possibility to simulate a noiseless SAS curve in an unrealistically wide range of angles. For such “ideal” data it is unnecessary to apply IFT as eq.2 readily gives a smooth and importantly unique solution. As a double-check, the Rg, Dmax, and smooth SAXS curve are computed on the fly from the predicted p(r) function and compared with the parameters estimated by more classical methods.

Validation, test set. Data augmentation.

# Results and discussion

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

Fig.3 Comparison of performances of different methods.

Fig.4 Web interface.

# Conclusion

A novel completely 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.