**Artificial neural networks for solution small angle scattering data analysis**

Abstract

We propose a novel method of SAXS data analysis based on the application of interconnected neural networks (perceptrons). For a given experimental data from proteins, RNA, or DNA our stack of networks evaluates model-free parameters: molecular weight, maximum intraparticle distance, a radius of gyration, pair-distance distribution function p(r), and a noise-free scattering curve. This completely automatic approach has proved to be robust against experimental errors, applicable to data from particles of various nature, size, and shape. The method was implemented as a publicly available web service with a graphical interface, providing the possibility to inspect and download the results (<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 [Grawert & Svergun, 2020; Koch, 2003; Glatter & Kratky, 1982]. Due to relatively soft requirements to sample preparation for SAS experiment and to the intense development of synchrotron radiation sources, the technique became high throughput and publicly available even to the non-specialists in the area. Despite the huge success in the development of data analysis software [ATSAS, SASTBX (J. Appl. Cryst. (2012). 45, 587-593), BioXTAS], calculation of crucial primary SAXS-derived parameters such as radius of gyration Rg, maximum intraparticle distance Dmax, molecular weight MW, and pair-distance distribution function p(r) still strongly depends on the quality of the experimental data and sometimes stays non-trivial task. 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. Attempts to solve this problem by so-called indirect Fourier transformation (IFT) have been originally proposed by Glatter [Glatter, 1977], and further developed by Svergun [Svergun, 1992] and Hansen [Vestergaard & Hansen, 2012]. In IFT approach a guess on the Dmax must be given, a p(r) function is expressed as a sum of some functions (e.g. cubic splines), and a classical regularization procedure [Tikhonov, 1943] is applied such that i) p(r) agrees to experimental data and ii) ensuring satisfaction to the imposed constraints. Most commonly it is a constraint on the smoothness of p(r), allowing to reduce/remove termination effect. However, in all of these approaches, the choice of the final solution remains a subjective criterion left to the discretion of the user.

Recently, artificial intelligence (AI) technologies applied to structural biology have experienced a sudden leap with regard not only to the business and industrial applications but also to academic research [Senior et al, 2019], [Armenteros et al, 2019], [Liu et al, 2020]. We propose a novel method of SAXS data analysis based on state-of-the-art machine learning principles. We employed a stack of interconnected neural networks trained on synthetic SAXS data from thousands of models from PDB databank [Berman et al, 2000] to perform principal data analysis. For a given experimental data from proteins, RNA, or DNA our stack of networks independently evaluates overall parameters: molecular weight (MW), Dmax, Rg; as well as p(r) function and a noise-free scattering curve. Here we demonstrate the robustness of the novel method and test predicted parameters against those estimated by well-established independent methods. This completely automatic approach has proved to be robust against experimental errors and applicable to data from particles of various chemical nature, size, and shape. The method was implemented as a publicly available web service with a graphical interface, providing the possibility to inspect and download the results (<https://dara.embl-hamburg.de/gnnom.php>).

Methods:

**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 the better performance it is generally recommended to prepare unbiased training set, ideally evenly distributed over the parameter for prediction. As is seen from the histogram in Fig.1, deposited to PDB molecules are clearly shifted towards small and globular proteins.



*numbers*

*numbers*

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

Intuitively, it is clear that the vast majority of the structures were solved by crystallographic methods, thus the overall distribution tends to be biased towards small and globular proteins simply due to their ability to form a crystal. Computer experiments confirmed, that neural networks trained on the whole amount of data works well for the small and globular proteins and underperform for bigger and elongated models. To make the training set as diverse and complete as possible, we applied 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 out of more than 150000 in two-dimensional space (MW, Rg). As is demonstrated in Fig.1, the chosen dataset of 1015 models have a step-like distribution of drastically more diverse by size and shape models.

**Neural networks architectures.** We exploit supervised learning interconnected neural networks (perceptrons) with one hidden layer (Fig.1) for predicting MW, Dmax and p(r) function. Since expected output is a number and not a discrete value, we encounter a classical regression task.

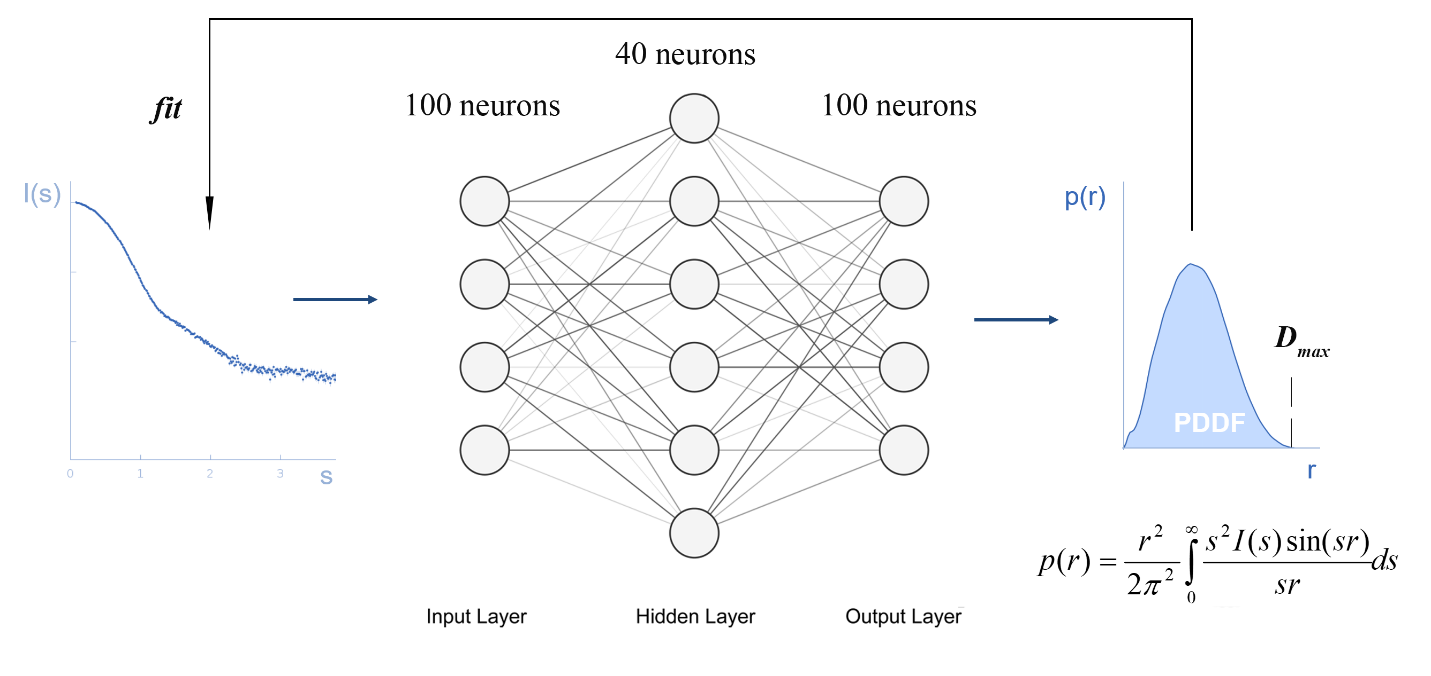
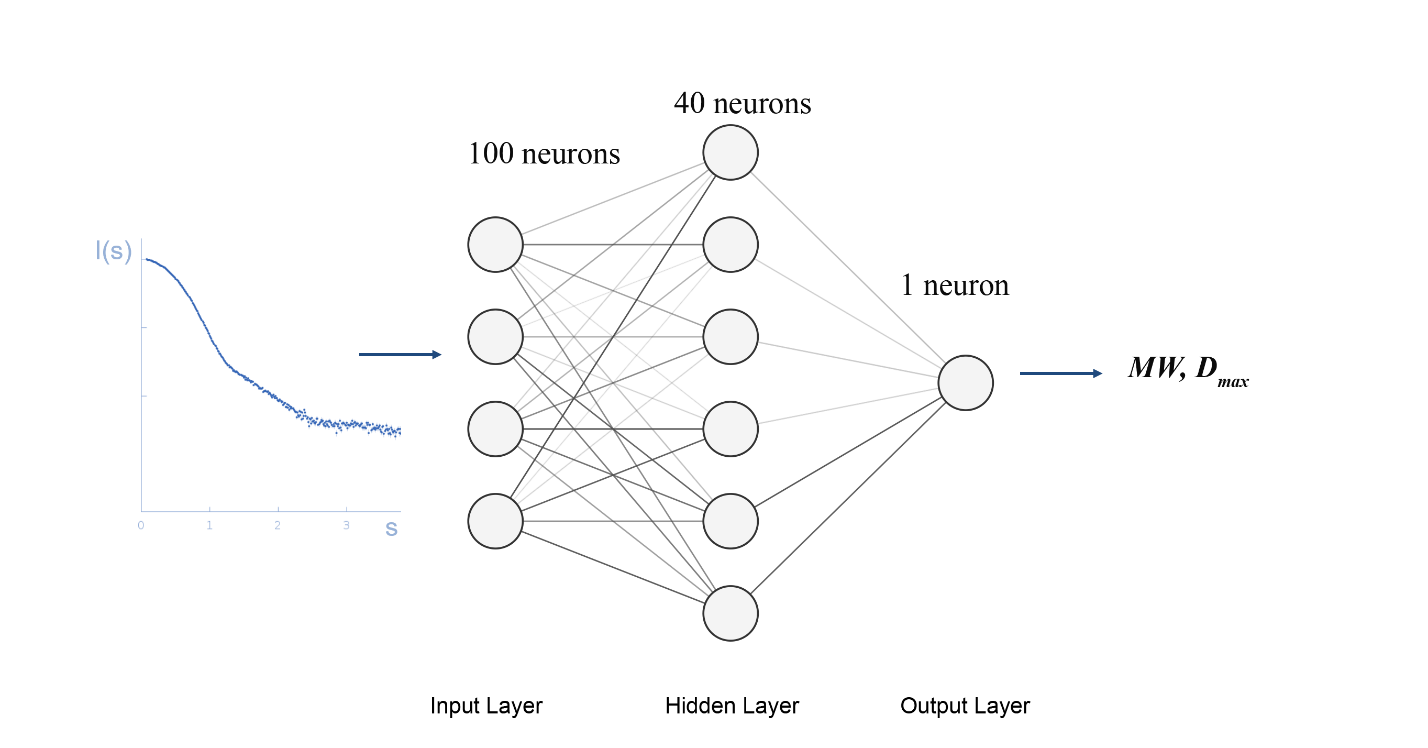


Fig.1 Architectures of neural networks and their relations in the overall workflow.

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 or 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 ANNs utilizing ReLU activation function with a width (or a number of neurons) 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 context of SAXS it is sufficient to use a perceptron with only one hidden layer and the width of 40 neurons.

**Molecular weight**. Estimation of molecule weight of suspended macromolecules in solution is one of the most important overall parameters derived from SAXS experiment. To date, there is a number of well-established techniques, that could be nominally subdivided into concentration dependent and concentration independent methods. Former account for dependence of forward scattering I(0) on the MW and utilize the scattering from calibrants (e.g. known protein or water) [Mylonas, Svergun, 2007]. The latter methods allow to determine MW of a molecule from a single background subtracted curve and require no *a priori* information on sample concentration. The historically first method calculates hydrated volume (also known as the Porod volume) from ratio of I(0) to Porod invariant Qp, and estimates MW as an empirical relation between protein volume and mass [Petoukhov et al, 2012]. Further there were attempts to improve the accuracy of MW determination by varying integration range over *s* [Fisher, 2010], modification of Porod invariant and introduction of a more complicated empirical dependences between dimensionless volume of correlation and MW [Rambo&Tainer, 2013]. The recent developments account not only for molecular size, but also for its shape using the classical support vector machine approach [Franke et al, 2018]. At the same time there was an attempt to combine all described methods together employing Bayesian statistics, enabling one to calculate the most probable molecular weight and assess its credibility interval [Hajizadeh et al, 2018].

Application of neural networks is a novel very promising method, that takes into account not only molecular size, but its shape and conformational state (folded/unfolded), as it uses the whole range SAXS curve for calculation. Moreover, it appears to be the first method for concentration independent estimation of MW from SAXS data applicable to DNA/RNA samples, and not only to proteins.

**Maximum intraparticle distance Dmax.** The common and as a matter of fact the only available method to determine Dmax from SAXS scattering profile is via inspection of corresponding p(r) function. Upon certain restrictions, such as non-negativity and 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.

Application of ANNs gear human independent way of Dmax calculation, based solely on the one-to-one SAXS profile – Dmax correspondence. As a matter of fact, in experimental SAXS data there is no a single Dmax value, as the scattering is always averaged over the huge ensemble of slightly different (even in case of monodisperse solution) particles. Additionally, hydration layers may vary among different particles and introduce uncertainties to Dmax to some extent. Nevertheless, as we demonstrate below, our method seems to be extremely precise and robust against experimental errors.

**Pair-distance distribution function: Dmax, Rg and noiseless curve**. Pair distance distribution function represents the histogram of distances between pairs of points within the particle. , weigh… it is possible to validate .a as a byproduct we get another way of calculation of Dmax and Rg. The noiseless curve can be readily obtained by inverse Fourier transform of predicted p(r) function (eq.1).

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

Fig.2

*Conclusion.*

A novel completely independent technique for estimation of primary SAXS parameters was developed. The comparison with well-established methods showed high consistency and robustness against the experimental noise.