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 scattering (SAS) experiments are widely used for the characterization of biological macromolecules in solution. SAS patterns contain information on the size and shape of dissolved particles in nanometer resolution. Here we propose a novel method for primary SAXS data analysis based on the application of perceptrons - interconnected artificial neural networks. Trained on synthetic SAS data, the neural networks are able to predict molecular weight and maximum intraparticle distance (*Dmax*) of previously unseen experimental data. The application area of the method is of a wide range and includes SAS data from nucleic acids (DNA/RNA), as well as from compact, partially unfolded, and extended proteins, including intrinsically disordered proteins (IDPs). Potentially, this area is not limited by the described objects and can be easily extended towards e.g. data from heterocomplexes or inorganic nanoparticles by further extension of the training set. The method was rigorously tested against synthetic SAXS data and demonstrated higher accuracy and better robustness against simulated experimental noise compared to all available conventional methods. The programs for estimation of MW and *Dmax* are included in the ATSAS package and will be publicly available in the next release, while the python code for generation of NN models, for its application on SAS data, as well as many other useful utilities, are freely available at the git-hub for academic use.

# Introduction

Small-angle scattering (SAS) of X-rays (SAXS) and neutrons (SANS) 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 to the non-specialists in the area.

The SAXS scattering profile represents the scattering intensity I(s) as a function of the scattering vector s = 4πsinθ/λ, where θ is the half of the scattering angle between incoming and diffracted beams, and λ corresponds to the wavelength. The molecular weight (MW) and maximum intraparticle distance (*Dmax*) are the crucial parameters derived from SAS experiments that are typically used for further analysis, e.g. for *ab initio* model reconstruction. However, their accurate estimation remains a non-trivial task with higher potential uncertainty linked to the lower quality of experimental data.

Recently, the application of neural networks (NN) has experienced a sudden leap in almost all areas of everyday 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 by AlphaFold (Senior et al.), and even in the area of SAXS (He et al., 2020). Theoretically, the application of deep NNs becomes clearly beneficial when the amount of available training data set reaches a certain threshold. Conventional methods (including shallow NNs) are not able to comprehend the common patterns in such big data sets, and their learning curves reach saturation soon after. On contrary, deep networks have a larger capacity and are able to recognize more profound and hidden from the human eye patterns in the data. Inspired by the recent enormous progress in the field of artificial intelligence (AI), we tackled the state-of-the-art machine learning technologies for SAS data analysis.

Another major advantage of using the supervised machine learning approach for data analysis over classical mathematical or physical models is the possibility to augment the training set data: this way, one can easily adjust the area of applicability of e.g. a given NN model and tailor it for the specific objects, instrumentation features or experimental setup. One example from the SAS area is the robustness of predictions against experimental noise. The experimental noise is inevitably present in any SAS data, it reduces the information content thus increasing the ambiguity of data interpretation. In simplistic terms, the addition of experimental noise transforms data at higher angles to the white noise, thus effectively shortening the available experimental s-range from the side of the higher angles and reducing the available information on the sample.

To make a NN more robust against any specific source of distortions, e.g. gaussian or systematic noise from a beamline, one can merely generate a realistic noise and augment the training set with that noise. A NN with enough capacity will learn how to filter this noise internally and make fine predictions despite truncated data. As a byproduct, NNs allow one to practically investigate the information content that can be retrieved from SAXS data of the different angular ranges for the determination of desired parameters independently from the Shannon theorem (crystallography and 1980). Similarly, one can overcome other common shortages of SAXS data collection, such as buffer sub/over subtraction, presence of a parasitic scattering, e.g. the “flares” from beam defining slits, buffer mismatch, etc. Along with that, the machine learning-driven approaches lack the limitations of a chosen approximation (e.g. the scattering homogeneity of a model) and may bring more accurate predictions employing previously unrecognized patterns and connections between SAXS data and the derived parameters.

We employed a stack of interconnected artificial NNs trained on synthetic SAXS data generated from thousands of models from various databases including the PDB databank (Berman et al., 2000) to perform principal data analysis. Here we demonstrate, that our method has much higher accuracy and is less demanding in terms of data quality compared to all other available methods. Moreover, our method appears to be unique in the SAS field for concentration-independent estimation of MW for intrinsically disordered proteins (IDP), and the second available method for MW estimation of nucleic acids (NA). Additionally, the application of NNs gear the only human independent way of *Dmax* estimation, based solely on the one-to-one correspondence (SAXS profile – *Dmax*), without the introduction of unnecessary regularization parameters. We would like to stress, that this work primarily aims to demonstrate the proof-of-the-principle capabilities of NNs in the field of SAS, whereas further improvement in accuracy and stability is possible and limited only by the choice of the training set.

# On limitations of conventional methods.

The MW is a crucial parameter in SAXS that can be exploited e.g. for the determination of predominant oligomeric state in solution. However, the currently available methods for MW estimation are not very accurate with as a usual rule of thumb an error of 10% or more. To date, there are several well-established techniques for MW estimation, that nominally could be subdivided into two major categories: concentration-dependent and concentration-independent methods. Former ones 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. In circumstances where the protein concentration cannot be accurately determined (e.g. for in-line SEC-SAXS experiments), these methods are necessary to obtain a reliable MW assessment. In the scope of this work, we are interested in concentration-independent methods.

**Porod’s method.** The historically first concentration-independent method is known as Porod’s 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 scattered beam, Δρ – excessive electron density, and Q is the Porod invariant. If we consider the scattering particle to be of homogeneous electron density, the right part of the equation simplifies to:

, (2)

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

. (3)

The MW is typically estimated as an empirical relation between the volume of the particle and its mass, which equals V/MW = 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 real experimental s-range (so I ~ s-4 power law is usually applied 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, which could be a quite crude approximation e.g. in case of protein-RNA heterocomplexes. This method can be very effective, however, it requires accurate knowledge of electron density and works only on proteins. Additionally, it is very sensitive to subtraction errors.

**Fisher’s method.** The further performance 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 so-called apparent volume as (similarly to (3)), and establish a linear dependence between V and V’:

V = A + BV’ , (5)

where the linear and angular coefficients A and B were determined empirically for different smax values. Given the look-up table with A and B values, one can find these coefficients corresponding to the experimental smax and obtain a more accurate prediction for the MW. The drawbacks of this method are essentially the same as those of 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. The authors found different couples of c and k values for proteins and RNA, therefore, it is the only available to date method for estimating the MW of RNA. However, this approach is reported to be less accurate than 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 essentially represents a classical k-Nearest-Neighbor classifier. In this method, the Fisher’s truncated integral Q’ (eq.4) is calculated up to three values of sRg = 3,4,5 for a huge set of simulated SAXS patterns from geometrical bodies and proteins from PDB. 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 this 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 alongside their size and can be used as a classifier for the molecule type (compact, extended, flat). 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 simulated a large test dataset of SAXS profiles, then calculated 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 the most probable MW alongside 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 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 transformation (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, then the p(r) function is expressed as a sum of some analytical functions (e.g. cubic splines). Finally, the classical regularization procedure (Bertero et al., 1980) 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 somewhat arbitrary and is usually left to the discretion of a user.

# Methods

Neural networks architecture. A NN consists of layers of interconnected units (fig.1). Each unit essentially performs a multiple linear regression operation, then applies some activation function, and passes the result further to the next layer. Mathematically speaking, if we have an input vector , the unit does a dot multiplication of that vector with an internally stored vector of “weights” of the same dimensionality and adds a constant (a “bias”):

, (10)

where w is the array of weights associated with the unit, b is a scalar called “bias”, and f is some analytical activation function. For example, the very simple but widely used activation function is called “rectified Linear Unit” (ReLU) and it simply passes on the positive numbers and nullifies the negative ones. If the units of the i-th layer are interconnected with all units of the (i+1)-th layer, such architecture is called interconnected feedforward neural network, or a perceptron. In layman’s terms, one can consider a perceptron as an algorithm that can approximate any function and map its input vector to the desired output value. During the training, perceptrons usually learn how to predict output value on the basis of a big enough “training set” - couples of (input data – expected output value). The embedded minimization algorithm optimizes the weights and biases of all units in such a way, that the output value becomes as close to the known ground truth value as possible. Once trained, the NN can be used for predicting the desired parameters from previously unseen input data.

To avoid overfitting – when a NN learns the training set too well and tries to fit specific, non-general features of the training set, the separate data called “validation set” is prepared. In this way, the NN learns on the training set but assesses the loss (discrepancy between prediction and ground truth) based on validation data. Finally, to benchmark the results against other methods, the third independent “test set” is usually used. The common distribution used in machine learning is 80% training set, 10% validation set, and 10% test set.

In this work, we have exploited Keras and Tensorflow modules in the python framework to construct perceptrons capable of predicting the MW and *Dmax* The adequate choice of the NN architecture, together with the optimization of its hyperparameters, is the most crucial and typically the most time-consuming step while designing a NN for a particular purpose. Since we expect the NN models to predict either MW or *Dmax*, it seems natural to make the output layer containing a single unit with its value representing the value of the desired parameter.

As the expected output of the NNs is a numeric score and not a probability, we encounter a classical regression task of NN supervised learning. The SAXS data used for training first undergo a normalization to I(0) = 1, then the data is rebinned to a particular common angular grid (0 ≤ s ≤ 10 nm-1, 256 points) and used as an input for the first layer of the NN. Fig. 1 demonstrates an overlay of the training procedure for the determination of MW.



80 units

1 unit

256 units

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

It is a good practice to begin a NN design with an estimation of the complexity of your problem to utilize an appropriate, not overloaded in terms of units and hidden layers, architecture. 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 units 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 ReLU activation function with a width (where the width is the number of units 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 SAXS with the typical angular range up to 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 proved to work quite well on our data set, we have found that using a somewhat excessive architecture with 3 hidden layers and 80 units (33560 weights, 241 biases) leads to faster convergence and more stable solutions, as well as to the slightly better prediction results. However, the further increase of the architecture complexity only worsens the predicted results, apparently because the backpropagation algorithm struggles to find an absolute minimum of the loss function among too many parameters to optimize.

During extensive testing, we have also found, that using hyperbolic tangent as an activation function instead of ReLU marginally improves the results. The possible explanation for that is connected to the output range of tanh values (-1, +1), which allows passing negative signals in between the layers and thus pushing more units 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 proteins deposited in the worldwide protein databank PDB ([www.rcsb.org](http://www.rcsb.org)) (Berman et al.) for simulation. Theoretically, for better performance of NNs it is recommended to prepare a complete and unbiased training set, ideally evenly distributed over the parameters for prediction. In other words, the “ideal” training set shall contain an equal number of proteins of various sizes (even distribution of an MW histogram) and shapes (even distribution of a radius of gyration Rg histogram ). However, the diversity of models is limited by the database and as is seen from the fig.2(a), these distributions of deposited in PDB models are severely skewed towards small and globular proteins.



(a)

(b)

(a)

(b)

*numbers*

*numbers*



Rg, Ȧ

MW, kDa

(c)



MW, kDa

(d)

Rg, Ȧ

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 snapshot of PDB containing 135 238 proteins (d) MW vs Rg for the chosen training set.

It is intuitively clear, that smaller proteins are easier to solve and thus there are more of them in PDB, especially given the historical perspective of available methods in 1970-2000 years. Moreover, the vast majority of the structures in PDB were solved by crystallographic methods, which implies that the proteins must be able to form crystals. As is well known from crystallography, the smaller and globular proteins are more prone to be crystallized compared to large and flexible ones.

In contrast, SAXS is a low-resolution technique and is typically used for the determination of supramolecular structures. Thus, it mostly deals with relatively big proteins, homo- or heterocomplexes of 60 kDa and higher. The performance of NNs strongly depends on the chosen training set and can be significantly advanced if the diverse efficiently represented proteins are chosen, so that the broader range in (MW,*Rg*) space is covered. Preliminary experiments confirmed, that NNs trained on the whole PDB data (fig.2 (c)) works well enough for the smaller compact proteins, but underperforms for bigger and elongated models (not shown). The NNs trained on all PDB models, however of some practical value, would fall short of the SAXS community aspirations since one of the strongest sides of the method is its ability to analyze rather big, unfolded, or even intrinsically disordered proteins.

With that in mind, we decided to amend the training set to make it more diverse, complete, and evenly represented over the sizes of MW~0-400 kDa and shapes of Rg ~0-60 Ȧ. Among other techniques we tried, 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 this way and after filtering outliers we selected ~7000 models in two-dimensional space (MW, Rg) out of more than 135000 proteins available to date in PDB (fig.2(d)). As is demonstrated in fig.2(b), the chosen dataset has almost a step-like distributions across both parameters (MW, Rg) indicating a drastically better subset of models for training a NN. As a preprocessing step, the heteroatoms were deleted from all pdb models to remove binding ligands and other non-organic 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 to the I(0) = 1. It is worth noting, that additional normalization of the data such as using various SAXS plot representations (e.g. Kratky plot or log I vs s), as well as more classical normalizations such as subtracting the average curve, and dividing by standard deviation did not bring significant improvements in the NNs performance. These normalizations can marginally improve the predictions on low simulated concentrations while worsening the predictions on smoother curves.

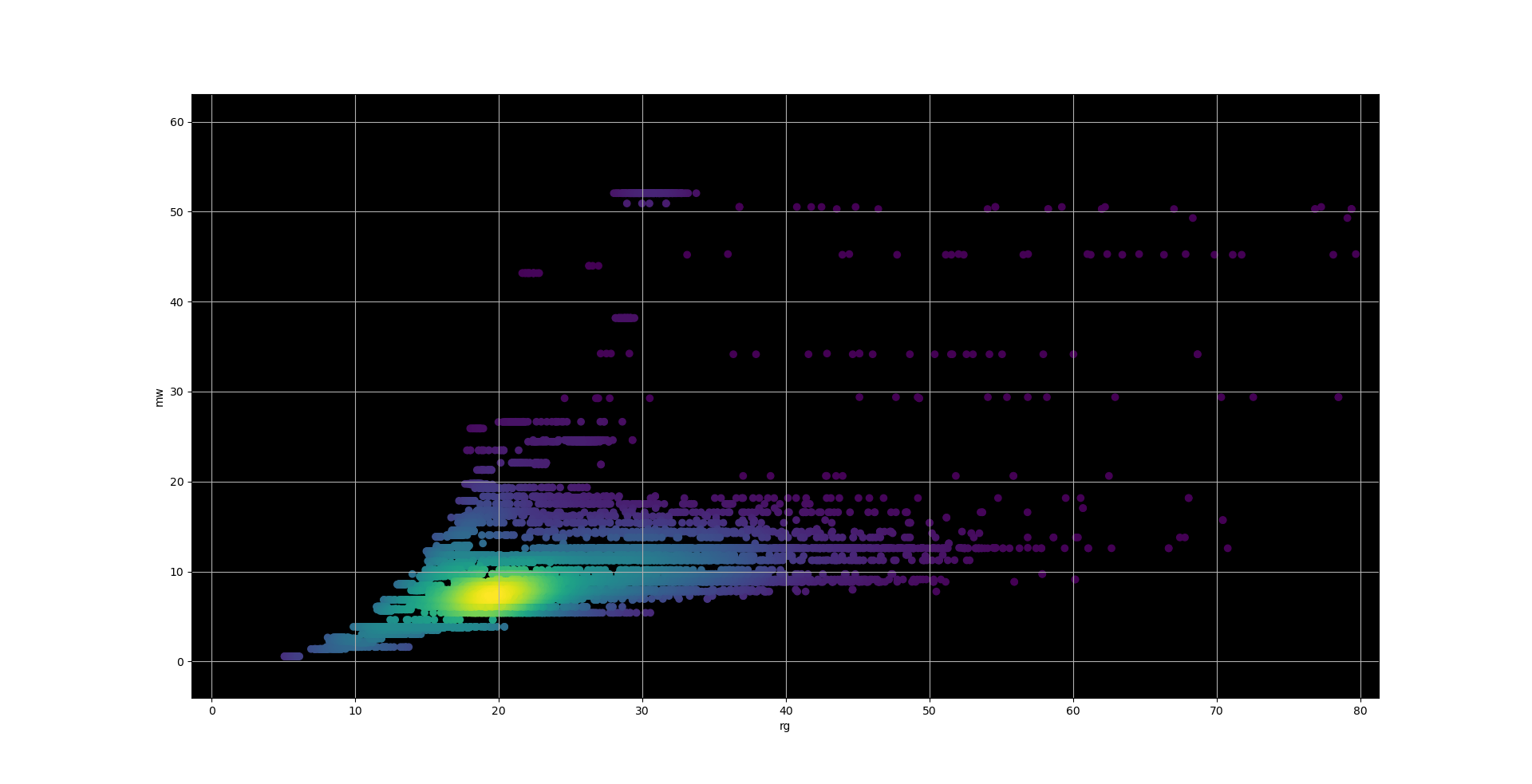
The experimental noise, simulated at 7 different protein concentrations c = 0.25, 05, 1, 2, 4, 8, 16 mg/ml, was generated based on experimental data from the p12 beamline (Blanchet et al.) that corresponds to the data acquired with the sample-to-detector distance of 1 meter, exposure time of 1 second, and X-ray energy of E = 10 keV. For MW and *Dmax* two separate NNs with similar architectures were trained.

Intrinsically disordered proteins. From the thermodynamical point of view, the IDPs are characterized by a low content of hydrophobic amino acids while having a high number of polar and charged amino acids. Thus IDPs cannot sufficiently bury a hydrophobic core to fold into stable globular proteins and therefore lack a stable tertiary structure in solution. Furthermore, high net charges promote disorder due to electrostatic repulsion in between charged residues (Oldfield and Dunker, 2014). Interestingly, many disordered proteins reveal regions without any regular secondary structure at all. These regions do not contain only one set of Ramachandran angles, and therefore are more flexible than the structured loops of globular proteins.

It has been predicted that more than 35% of human proteins have significant regions of disorder (Fukuchi et al., 2011) and about 25% are likely to be completely disordered (Uversky and Dunker, 2010). These proteins are functionally important for many cellular regulatory processes, and may also be involved in pathological processes associated with protein misfolding or aggregation (Chiti and Dobson, 2006), (Uversky et al., 2008). Under physiological conditions these proteins constantly fluctuate between different structural states, resulting in a dynamic mixture of conformations in a polydisperse solution. Quantitative characterization of such heterogeneous systems is a difficult task, and SAXS is among the few methods capable of providing unique information on the structural properties of the flexible macromolecules (Kikhney and Svergun, 2015).

To prepare a training data set of the IDP models, we used the only currently available open database comprising complete IDP conformer pools: the PED database (Lazar et al.). A snapshot of the database was made that included 172 ensembles. To enhance the training set we have taken the first 50 conformers from each ensemble resulting in a total number of 10 000 models (fig.3a).

(a)



(a)

MW, kDa

Rg, Ȧ



1H1K 🡪

(605 kDa)

3REC

(0.6 kDa)

4KYY

(11 kDa)

2JYH

(28 kDa)

6UES

(39 kDa)

(b)



(b)

MW, kDa

Rg, Ȧ

Fig.3. Training set chosen for IDP and nucleic acid models. The most representative DNA/RNA models are presented in (b).

The noticeable horizontal smearing effect is the result of using different conformations of the same proteins with various shapes (*Rg* values*)*, but identical MW. The parametric “hot spot” is located at circa 20 kDa and *Rg* of 7-8 Ȧ indicated the most characteristic sizes and shapes of IDPs. Finally, we have simulated the SAXS profiles and identified ground truth MW and *Dmax* similarly to how it was described previously.

For the validation set, we randomly took 10% of the training set. In order to produce a realistic pseudo-experimental test set, we averaged the simulated SAXS profiles over the ensembles (prior to adding simulated noise) resulting in 172 smeared SAXS patterns. Since the conformers within each ensemble are chemically identical molecules, the MW of each such “averaged model” was taken as the MW of the random conformer, and the *Dmax* was chosen as the *Dmax* of the most extended conformer from each ensemble. Two separate NNs were trained for the determination of MW and *Dmax* values of IDPs.

Nucleic acids. The DNA and RNA are suitable objects to be studied by SAXS as they have relatively high electron contrast and aptly scatter X-rays while being typically less prone to radiation damage than proteins. The nucleic acid models (the “pure” nucleotides, not their heterocomplexes with proteins) are not as massively populated in PDB as protein models and comprise only less than 2% of all entries.

To collect non-redundant models, we used the NDB server (<http://ndbserver.rutgers.edu/>) (Coimbatore Narayanan et al., 2014). After preliminary filtering, we fetched around 3000 models from PDB in total. As is seen in fig.3, the models are also mostly small and compact with the vast majority of them populated in between MW of 10 and 20 kDa and *Rg* of 10 and 25 Ȧ. Given the limited number of models and the fact, that according to the SASBDB database (Kikhney et al., 2019) the majority of DNA/RNA models used in SAXS are also located within the same (MW-*Rg)* interval, we decided to use those models as is without further shrinking. The potential extension of the training set is possible using one of the available software for NA secondary (e.g. Mfold (research and 2003)) and tertiary structure (e.g. OligoAnalyzer (Owczarzy et al.)) prediction. However, the 3D structure prediction from the NA sequence seems currently an active area of research and not as fully developed as protein prediction.

Therefore we decided to use experimentally determined models only. After the standard distribution of these 3000 models to 80%/10%/10% for training/validation/test sets, we trained another two NNs to predict MW and *Dmax* for NAs in solution.

# Results and discussion

Benchmark. As was discussed in detail in the previous section, all available conventional methods have different ads and procs and perform differently upon specific shapes of models and various levels of noise. In order to produce some generic metric of the prediction accuracy, we applied these methods (together with the developed NNs) to the test sets, and averaged the relative prediction error over all models:

, (11)

where *N* is the total number of models in the test set, *P* is the predicted value (either MW or *Dmax*) and *GT* is the ground truth value. Since the test set was generated to comprise models of different sizes and shapes in mind, this value represents not only the accuracy of the given method but also its ability to work on models of various MW and degrees of compactness.

It is worth noting, that even though some of the methods are not directly applicable to IDPs and NAs, we decided nevertheless to demonstrate their performance just for the completeness of the picture. In essence, there is only the volume of correlation (Vc) method that can estimate MW of RNA (not DNA!) from SAXS and DATGNOM’s IFT method that works equally well for the particles of different chemical nature. The all-to-all comparison is presented in fig.5, where it is seen that the NNs not only outperform the conventional methods for all types of particles but are also much more robust against simulated noise.

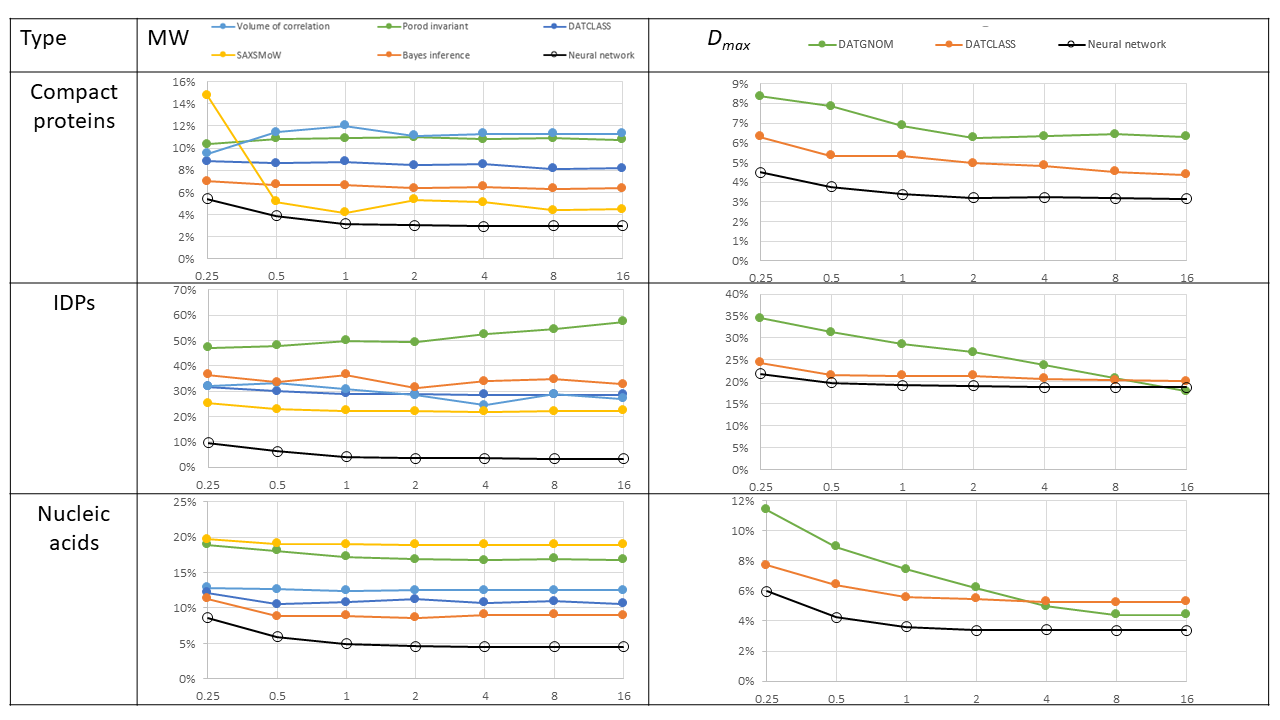


Fig.5 Comparison of performances of different methods (proteins, NAs, IDPs).

Interestingly, the accuracy of predictions by NNs for both MW and *Dmax* improves gradually with the simulated concentration and mostly reaches a plateau at concentrations as low as 1-2 mg/ml. The conventional methods work somewhat successfully on globular proteins and worse on proteins with more exotic shapes, with the best average result provided by the Fisher’s (MoW) method. It also appeared to be the only method where its behavior is strongly dependent on the amount of simulated noise, indicating the usage of higher angles for predictions. Otherwise, the errors are mostly within the claimed by authors 10%. Surprisingly, the NN almost two times outperformed all these methods with the average error just beyond 3% for the higher simulated concentrations. The *Dmax* estimate by the NN has also improved the previous best result of the Size&Shape method (DATCLASS) almost two-fold with the lowest <Δrel> touching 3%.

Understandably, as these methods are not specifically tailored for the nucleic acids, the relative performance of NNs is even better in the case of DNA/RNA models. MoW can not compete with NN in this case and demonstrates the worst results of <Δrel> about 20%. The Bayesian inference gives the best predictions in MW, however, it comes short of the classical IFT method for *Dmax* prediction on the highest simulated concentrations.

IDPs happen to be the most challenging objects for predictions. The conventional methods failed to reproduce reasonable MW estimates with the <Δrel> as high as 50%, while the NN showed much better results of 3-10% enabling for the first time to reliably estimate the MW of IDPs from SAS data. Unfortunately, all methods could not make precise predictions for the *Dmax* estimate, with NN result slightly better than DATGNOM and DATCLASS, but with <Δrel> still staying beyond 20%. This bottleneck may be connected with the preparation of the “averaged-over-ensemble model”: if we take into consideration the SAS curve averaged over all conformers and estimate the *Dmax* utilizing any available methods, the predicted value will be drastically underestimated and closer rather to averaged over ensemble *Dmax* than to the longest conformer present in the ensemble. However, given that *Dmax* for IDPs is a rather arbitrary number and does not play a significant role in further SAS analysis, we decided not to overfocus on this issue and leave it beyond the scope of this article.

Information content. Upper bound for predictions.

Aside from practical value, the usage of NNs offers a new challenge to practically estimate the fundamental limit of the accuracy of MW and *Dmax* predictions from SAS data. When employing NNs, one does not rely on mathematical models and may hope to push the precision of predictions further compared with the conventional methods. If we take as an axiom, that NN can always find the absolute minima of the loss function, the accuracy of predictions is limited only by the information content of a given SAS curve.

To study this question, we simulated the “ideal” training set of noiseless SAXS profiles, determined on a wide s-range up to smax=10 nm-1,and normalized it on I(0) = 1. Further, we trained the NNs on these smooth data and applied them to the similar “ideal” smooth simulated test curves to predict MW and *Dmax*.

Surprisingly, as is seen from fig. 6, the obtained result is not drastically better than the ones we obtained for the noisy data and equal to 2.7% for MW and 3% for *Dmax*. This simple experiment demonstrates two important facts: (1) the described in fig.5 results almost reached the theoretical limit of prediction accuracy using NNs (and maybe the fundamental precision limit of SAS as a method); (2) the augmentation of the training set with experimental noise, as is seen from fig.5, indeed helps to deal with noisy data and only marginally reduces the overall accuracy of predictions.



(b)

(a)

Fig.6. Predictions for “ideal” smooth datasets versus ground truth for all models from the test set: (a) MW, (b) *Dmax*

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

Angular range. In theory, the information contained in a scattering curve depends only on two factors: maximum experimental angle *smax* and maximal intraparticle distance of the scattering molecule *Dmax*. All features and local behavior of a SAS curve may be completely described by the N equidistant Shannon channels with N = smax·*Dmax*/ π and the distance between channels equal to π/*Dmax* (crystallography and 1980). However, since the intensity of SAS curve rapidly decreases with angle, the useful signal at higher angles may be rather weak compared with the background and the effective *smax* be smaller than the nominally recorded one. On the other hand, this noise can be partially compensated for by oversampling, due to a much smaller angular increment in the SAXS profile compared with the π/*Dmax*. Therefore the estimation of effective *smax* depends on many factors and is far from being trivial.

One possible way to determine the real number *smax* was suggested by (Konarev and Svergun, 2015). It relies on the fact, that the usage of the unfeasibly high number of terms in the Shannon interpolation formula does not improve the fit to the SAXS data, but leads to high-frequency oscillations in p(r). The authors introduced a new criterion:

*f(N) = χ2(N) + αΩ(r)*. (11)

This criterion accounts for *χ2(N)* - fit to the data by the Shannon interpolation formula in reciprocal space and *Ω(r)* - the smoothness of p(r) function in direct space, with α being a scaling coefficient. This way, it is possible to estimate the effective *smax* of experimental SAS data and use only meaningful s-range.

An important question arises: given the effective *smax*, what is the maximum precision of MW and *Dmax* predictions that one can expect? The usage of NNs enables a convenient opportunity to get a deeper insight into the information content of different angular ranges of SAS curves and to empirically determine the dependence of the accuracy of the MW and *Dmax* estimates against the *smax*.

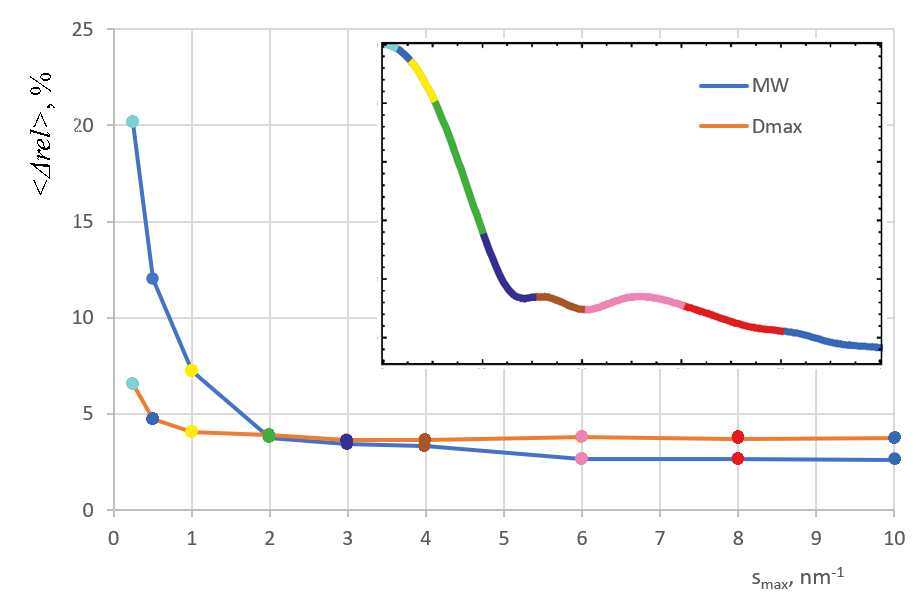


Fig.7 Errors versus maximal angular range smax for noiseless synthetic SAXS curves

To do so, we repeated the previously described experiment on smooth curves for various *smax* values. As is shown in fig.7, the presence of higher angles does not drastically improve both estimates, while the plot for MW demonstrates more pronounced degradation upon the lower smax values. That illustrates the well-known fact, that lower angles in reciprocal space contain information on the larger distances in real space, whereas the estimation of maximal intraparticle distance essentially requires only the Guinier region. Therefore, the *Dmax* prediction curve at fig.7 rapidly reaches a plateau of ~4% soon after smax goes beyond the Guinier region. For MW prediction the impact of higher angles is more pronounced, and the angles up to smax = 6 nm-1 are used to slightly improve the prediction. The angles on this scale mostly contain information on the molecule interface and molecule’s inner structure, which indeed may add a somewhat small correction to the overall estimation of the molecule mass.

# Conclusions

Here we presented a novel independent method for the estimation of primary SAXS parameters using modern NN technologies. Through a systematic analysis, we found that well-established methods for MW and *Dmax* evaluation demonstrate variable performance depending on the size, shape, chemical nature, and amount of simulated noise. The comparison of our method with existing methods demonstrated much higher accuracy and robustness of our method against simulated noise. Nevertheless, we believe that this work is merely a demonstration of the real capacities of NNs applications to the SAXS data analysis, and the potential of the method can be easily extended and improved under specific needs.

Each conventional method for MW determination utilizes its own physical and mathematical assumptions thus harbours its own advantages and limitations. Due to the fact, that our method is not confined within the frames of any approximations (e.g. the homogeneity of electron density), it allows us to anew assess the real capacities of SAXS data in terms of information content and to push further the accuracy of SAXS primary data analysis beyond the commonly accepted uncertainty of 10%.

It should be noted that so far MW of RNA was only available through the Vc method, but with the new approach, the MW of DNA/RNA and IDPs are also accessible for predictions. The *Dmax* estimations by our method do not require IFT and can be done directly from experimental data. So far it was only possible utilizing the Shape&Size method of (Franke et al., 2018). The impact of angular range on the prediction accuracy was also discussed and the empirical dependence was established.

The developed methods for MW and *Dmax* estimation will be included in the next ATSAS release, free for academic use (https://www.embl-hamburg.de/biosaxs/download.html). The python code for preparing test sets, analyzing the parametric space, generating and augmenting the data with pseudo-realistic experimental noise, training and applying NNs, comparing the results using multiple metrics and representations are openly available in the git-hub (<https://git.embl.de/grp-svergun/gnnom>).

# Acknowledgments

The work was supported by the BMBF grant 16QK10A (SAS-BSOFT).

# References

Armenteros, J.J.A., Salvatore, M., Emanuelsson, O., Winther, O., Von Heijne, G., Elofsson, A., and Nielsen, H. (2019). Detecting sequence signals in targeting peptides using deep learning. Life Sci. Alliance *2*.

Barberato, C., Henri, M., Koch, J., Svergun, D., Barberato, C., and Koch, M.H.J. (1995). CRYSOL-a Program to Evaluate X-ray Solution Scattering of Biological Macromolecules from Atomic Coordinates Projet View project Projet4 View project CRYSOL-a Program to Evaluate X-ray Solution Scattering of Biological Macromolecules from Atomic Coordinates. Artic. J. Appl. Crystallogr. *28*, 768–773.

Berman, H., Westbrook, J., … Z.F.-N. acids, and 2000, undefined The protein data bank. Academic.Oup.Com.

Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N., and Bourne, P.E. (2000). The Protein Data Bank. Nucleic Acids Res. *28*, 235–242.

Bertero, M., De Mol, C., and Viano, G.A. (1980). The Stability of Inverse Problems. In Dokl. Akad. Nauk SSSR, pp. 161–214.

Blanchet, C., Spilotros, A., … F.S.-J. of applied, and 2015, undefined Versatile sample environments and automation for biological solution X-ray scattering experiments at the P12 beamline (PETRA III, DESY). Scripts.Iucr.Org.

Chiti, F., and Dobson, C.M. (2006). Protein misfolding, functional amyloid, and human disease. Annu. Rev. Biochem. *75*, 333–366.

Coimbatore Narayanan, B., Westbrook, J., Ghosh, S., Petrov, A.I., Sweeney, B., Zirbel, C.L., Leontis, N.B., and Berman, H.M. (2014). The Nucleic Acid Database: new features and capabilities. Nucleic Acids Res. *42*, D114–D122.

crystallography, P.M.-J. of applied, and 1980, undefined Small-angle scattering. Information content and error analysis. Scripts.Iucr.Org.

Cybenko, G. (1989). Approximation by superpositions of a sigmoidal function. Math. Control. Signals, Syst. *2*, 303–314.

Debye, P. (1915). Zerstreuung von Röntgenstrahlen. Ann. Phys. *351*, 809–823.

Feigin, L.A., Svergun, D.I., and Taylor, G.W. (1987). Principles of the Theory of X-Ray and Neutron Scattering. In Structure Analysis by Small-Angle X-Ray and Neutron Scattering, (Springer US), pp. 3–24.

Fischer, H., De Oliveira Neto, M., Napolitano, H.B., Polikarpov, I., and Craievich, A.F. (2010). Determination of the molecular weight of proteins in solution from a single small-angle X-ray scattering measurement on a relative scale. J. Appl. Crystallogr. *43*, 101–109.

Franke, D., Jeffries, C.M., and Svergun, D.I. (2018). Machine Learning Methods for X-Ray Scattering Data Analysis from Biomacromolecular Solutions. Biophys. J. *114*, 2485–2492.

Fukuchi, S., Hosoda, K., Homma, K., Gojobori, T., and Nishikawa, K. (2011). Binary classification of protein molecules into intrinsically disordered and ordered segments. BMC Struct. Biol. *11*.

Glatter, O. (1977). Data evaluation in small angle scattering: calculation of the radial electron density distribution by means of indirect Fourier transformation. Acta Phys. Austriaca *47*, 83–102.

Gräwert, T.W., and Svergun, D.I. (2020). Structural Modeling Using Solution Small-Angle X-ray Scattering (SAXS). J. Mol. Biol. *432*, 3078–3092.

Guinier, A., and Fournet, G. (1955). Small-angle scattering of X-rays (Translation by C. B. Walker).

Hajizadeh, N.R., Franke, D., Jeffries, C.M., and Svergun, D.I. (2018). Consensus Bayesian assessment of protein molecular mass from solution X-ray scattering data. Sci. Rep. *8*, 1–13.

Hanin, B. (2019). Universal function approximation by deep neural nets with bounded width and ReLU activations. Mathematics *7*, 1–9.

He, H., Liu, C., and Liu, H. (2020). Model Reconstruction from Small-Angle X-Ray Scattering Data Using Deep Learning Methods. IScience *23*, 100906.

Hopkins, J.B., Gillilan, R.E., and Skou, S. (2017). BioXTAS RAW: Improvements to a free open-source program for small-angle X-ray scattering data reduction and analysis. J. Appl. Crystallogr. *50*, 1545–1553.

Kikhney, A.G., and Svergun, D.I. (2015). A practical guide to small angle X-ray scattering (SAXS) of flexible and intrinsically disordered proteins. FEBS Lett. *589*, 2570–2577.

Kikhney, A.G., Borges, C.R., Dmitry, |, Molodenskiy, S., Jeffries, C.M., and Svergun, D.I. (2019). SASBDB: Towards an automatically curated and validated repository for biological scattering data. Wiley Online Libr. *29*, 66–75.

Konarev, P. V, and Svergun, D.I. (2015). A posteriori determination of the useful data range for small-angle scattering experiments on dilute monodisperse systems. *2*, 352–360.

Lazar, T., Martínez-Pérez, E., … F.Q.-N. acids, and 2021, undefined PED in 2021: a major update of the protein ensemble database for intrinsically disordered proteins. Academic.Oup.Com.

Liu, H., Hexemer, A., and Zwart, P.H. (2012). The Small Angle Scattering ToolBox (SASTBX): An open-source software for biomolecular small-angle scattering. J. Appl. Crystallogr. *45*, 587–593.

Lu, Z., Pu, H., Wang, F., Hu, Z., and Wang, L. The Expressive Power of Neural Networks: A View from the Width.

Manalastas-Cantos, K., Konarev, P. V., Hajizadeh, N.R., Kikhney, A.G., Petoukhov, M. V., Molodenskiy, D.S., Panjkovich, A., Mertens, H.D.T., Gruzinov, A., Borges, C., et al. (2021). ATSAS 3.0 : expanded functionality and new tools for small-angle scattering data analysis . J. Appl. Crystallogr. *54*, 343–355.

McKay, M.D., Beckman, R.J., and Conover, W.J. (1979). A Comparison of Three Methods for Selecting Values of Input Variables in the Analysis of Output from a Computer Code. Technometrics *21*, 239.

Mylonas, E., and Svergun, D.I. (2007). Accuracy of molecular mass determination of proteins in solution by small-angle X-ray scattering. In Journal of Applied Crystallography, (International Union of Crystallography), pp. s245–s249.

Oldfield, C.J., and Dunker, A.K. (2014). Intrinsically Disordered Proteins and Intrinsically Disordered Protein Regions. Http://Dx.Doi.Org/10.1146/Annurev-Biochem-072711-164947 *83*, 553–584.

Owczarzy, R., Tataurov, A., … Y.W.-N. acids, and 2008, undefined IDT SciTools: a suite for analysis and design of nucleic acid oligomers. Academic.Oup.Com.

Petoukhov, M. V., Franke, D., Shkumatov, A. V., Tria, G., Kikhney, A.G., Gajda, M., Gorba, C., Mertens, H.D.T., Konarev, P. V., and Svergun, D.I. (2012). New developments in the ATSAS program package for small-angle scattering data analysis. J. Appl. Crystallogr. *45*, 342–350.

Rambo, R.P., and Tainer, J.A. (2013). Accurate assessment of mass, models and resolution by small-angle scattering. Nature *496*, 477–481.

research, M.Z.-N. acids, and 2003, undefined Mfold web server for nucleic acid folding and hybridization prediction. Academic.Oup.Com.

Schmidhuber, J. (2015). Deep Learning in neural networks: An overview. Neural Networks *61*, 85–117.

Senior, A.W., Evans, R., Jumper, J., Kirkpatrick, J., Sifre, L., Green, T., Qin, C., Zídek, A., Nelson, A.W.R., Bridgland, A., et al. AlphaFold: Improved protein structure prediction using 1 potentials from deep learning 2.

Svergun, D.I. (1992). Determination of the regularization parameter in indirect-transform methods using perceptual criteria. J. Appl. Cryst *25*, 495–503.

Uversky, V.N., and Dunker, A.K. (2010). Understanding protein non-folding. Biochim. Biophys. Acta - Proteins Proteomics *1804*, 1231–1264.

Uversky, V.N., Oldfield, C.J., and Dunker, A.K. (2008). Intrinsically disordered proteins in human diseases: Introducing the D 2 concept. Annu. Rev. Biophys. *37*, 215–246.

Vestergaard, B., and Hansen, S. (2006). Application of Bayesian analysis to indirect Fourier transformation in small-angle scattering. J. Appl. Crystallogr. *39*, 797–804.

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.