Artificial neural networks for solution scattering data analysis

D.S. Molodenskiy1, D.I. Svergun1\* and A.G. 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. SAXS 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 artificial neural networks (NN). Trained on synthetic SAXS data, the feedforward neural networks are able to reliably predict molecular weight and maximum intraparticle distance (Dmax) directly from the experimental data. The method is applicable to data from monodisperse solutions of folded proteins, intrinsically disordered proteins, and nucleic acids. Extensive tests on synthetic SAXS data generated in various angular ranges with varying levels of noise demonstrated a higher accuracy and better robustness of the NN approach compared to the existing methods.

# Introduction

Small angle X-ray scattering (SAXS) from biological macromolecules in solution is a powerful technique providing information on supermolecular structures and dynamics under a wide range of conditions (Gräwert and Svergun, 2020; Guinier and Fournet, 1955; Svergun et al., 2013). Due to non-restrictive requirements to sample preparation for SAXS experiments and recent progress in the instrumentation and of data analysis methods (Hopkins et al., 2017; Liu et al., 2012; Manalastas-Cantos et al., 2021), the method is widely utilized also in high throughput studies.

The SAXS data are obtained by illuminating a dilute, typically monodisperse solution of macromolecules with a monochromatic X-ray beam which results in an isotropic two-dimensional (2D) scattering pattern. The latter is azimuthally integrated into a 1D scattering profile which represents the scattering intensity I(s) as a function of the scattering vector s = 4πsinθ/λ, where θ is the half of the angle between incoming and diffracted beams, and λ is the X-ray wavelength. The “background” scattering from the pure solvent is independently measured and subtracted from that of the solution. From the background-subtracted scattering profile, one can directly evaluate structural characteristics of the scattering particle: radius of gyration (Rg), maximum intraparticle distance (Dmax), pair-distance distribution function (p(r)), molecular weight (MW). It is further possible to reconstruct the overall shape *ab initio* or obtain a hybrid model employing structural information from the high-resolution methods.

There is a number of established methods for the estimation of MW from SAXS data on a relative scale (i.e. without relying on scattering from calibrants). The accuracy of these estimates is limited (in the range of about 10% (Hajizadeh et al., 2018). These methods have been developed primarily for globular proteins and their applicability to SAXS data from disordered proteins and nucleic acids is not straightforward. Here, we explore the use of artificial neural networks (NN) for the primary SAXS data analysis to assess the MW and Dmax directly from the scattering data from folded proteins, intrinsically disordered proteins (IDP), and nucleic acids.

Recently, the application of NNs has experienced a sudden leap in almost all areas of everyday life, also thanks to the development of deep learning technologies (Schmidhuber, 2015). Massive 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 (Jumper et al., 2021), and even in the area of SAXS shape reconstruction (He et al., 2020).

NNs are excellent tools for supervised learning; the taskof learning is to find a function that maps an input to the desired output based on a training data set. In our case, the input is a vector of experimental intensities I(s) on a relative scale and the output could be a value representing an overall geometrical parameter, e.g. the MW or Dmax. Since obtaining reliable experimental SAXS data in sufficient quantities is challenging, one could compute the scattering from known protein and/or nucleic acid models for training assuming that the mapping function would be applicable to experimental data.

The simulated training set can be augmented: 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 SAXS area is the robustness of predictions against experimental noise, which is inevitably present in the experimental SAXS data and reduces its information content thus increasing the ambiguity of data interpretation.

We employed several feedforward artificial NNs trained on noise-augmented synthetic SAXS data generated from thousands of experimentally determined models to estimate MW and Dmax from folded proteins, unfolded proteins, and nucleic acids. Here we demonstrate that our method has higher accuracy and is less demanding in terms of data quality compared to the well-established methods to assess MW for folded proteins and nucleic acids. To the best of our knowledge, our method is unique for the MW estimation from SAXS data of intrinsically disordered proteins. Our method can also reliably estimate the maximum intraparticle distance Dmax directly from the SAXS profile for the above-mentioned macromolecule types.

# Estimation of the MW from SAXS data.

The approaches for MW estimation can be divided into two major categories: concentration-dependent and concentration-independent methods. The first category exploits the dependence of the forward scattering I(0) on the total number of electrons in the irradiated molecule (and, thus, on MW) and relies on the scattering from calibrants, e.g. from water or a protein with known MW (Mylonas and Svergun, 2007). These methods require the knowledge of sample concentration, partial specific volume, and scattering contrast of the solute. The second category utilizes a single background-subtracted profile on a relative scale and requires no additional information. Concentration-independent methods are more convenient to use and moreover, in some cases, the solute concentration cannot be accurately measured (e.g. for in-line size-exclusion chromatography SAXS experiments). Below we shall focus on concentration-independent methods.

**Porod’s method.** The historically first concentration-independent method is the so-called Porod’s method (Porod, 1951). 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 radiation, Δρ – excess 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 (1) 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 in the case of folded proteins is about MW/V ≈ 0.625 (Petoukhov et al., 2012). This calculation is limited by the three factors: (i) integration in (1) cannot be performed due to limitations in real experimental s-range, assuming globular proteins I ~ s-4 power law is usually applied to extrapolate the intensities on higher angles; (ii) the integration is affected by the experimental noise and the accuracy of background subtraction; and (iii) the equation (2) implies homogeneity of the scattering particle.

**SAXSMoW method.** The Porod’s invariant approach was extended by (Fischer et al., 2010)(Piiadov et al., 2019) Here, the authors integrate the Porod invariant in (1) not up to infinity, but up to a fixed smax value:

. (4)

The authors introduce the so-called apparent volume as (similar to (3)), and establish a linear dependence between V and V’:

V = A + BV’ , (5)

where the coefficients A and B are determined empirically for different smax values from simulated protein SAXS data. 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.

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

. (6)

The authors found an empirical dependence between Vcand the MW:

, (7)

where c and k are empirically determined constants via fitting results from theoretical scattering profiles. The authors mentioned ec = 0.1231 and 1/k = 1 for proteins and ec = 0.00934 and 1/k = 0.808 for RNA. Thus, this approach is applicable not only to SAXS data from proteins but to RNA data as well.

**Machine learning methods.** The web server for rapid search of structural neighbours DARA (Kikhney et al., 2016) accepts SAXS data from proteins, nucleic acids or their complexes, finds the closest SAXS profiles precomputed from PDB (Berman et al., 2000) models and reports the MW and Dmax of these models. If there is a structural neighbour that fits well the experimental data, then these values can be used as the estimates of the overall structural parameters.

The size&shape method (Franke et al., 2018) allows for a fast and selective lookup of structural neighbours in a database of SAXS patterns pre-computed from geometrical bodies and protein models from the PDB. This approach enables rapid multiclass shape classification (compact, extended, random-chain etc.) and estimation of Dmax and MW directly from the experimental SAXS data from proteins.

**Bayesian assessment of protein MW.** In the recent method (Hajizadeh et al., 2018) the MW is estimated using Bayesian inference with the MW calculations from the above-mentioned methods. The authors simulated a large test dataset of SAXS profiles, then calculated the MW for each profile 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 employs all the other methods and provides the most probable MW alongside its credibility interval. The disadvantage is similar to the Shape&Size method; the assessment works only for compact proteins.

# Estimation of the maximum intraparticle distance Dmax from SAXS data.

The assessment of the maximum size Dmax utilizes a pair distance distribution function p(r), which is 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 I(s) via the spherically averaged Fourier transformation (Debye, 1915):

, (8)

. (9)

(here, p(r) = 0 for r > Dmax). The limited angular range of the experimental data, as well as the presence of experimental noise, make the evaluation of p(r) an ill-posed problem. The method of solving this problem by an indirect Fourier transformation (IFT) has been originally proposed by (Glatter, 1977), and further developed by (Svergun, 1992) and (Vestergaard and Hansen, 2006). Here Dmax must be provided as an input parameter and the p(r) function is expressed as a sum of analytical functions (e.g. cubic splines). Finally, a regularization procedure (Tikhonov et al., 1977) is applied to calculate the p(r) agreeing with experimental data satisfying additional constraints. Most common constraint is the smoothness of the p(r), so that termination effects are reduced as much as possible. However, in these approaches, the choice of the final solution remains a subjective criterion left to the discretion of the user.

In the program AUTOGNOM (Petoukhov et al., 2007) (presently DATGNOM), multiple runs of GNOM (Svergun, 1992) are performed with Dmax values ranging from 2Rg to 4Rg to find the optimum Dmax and provide the p(r) function. Here Rg is the radius of gyration from the Guinier approximation.

# Methods

Training/validation/test sets

In this study, we considered three types of biological macromolecules: folded proteins, IDPs, and nucleic acids. To construct a training set we have utilized experimentally determined atomic models of the macromolecules available from respective databases. Each model was examined for connectivity and models with domains separated by more than 7Å were excluded. Heteroatoms were removed from all models.

## Folded proteins.

A total of 135 238 atomic coordinate files describing protein structures from protein-only biological assemblies were obtained from the protein databank (PDB) (Berman et al., 2005). 99% of these models have MW below 450 kDa; 80% of the models are in the range 10–86 kDa. To avoid bias towards smaller proteins, we have constructed a histogram of MW distribution for the pool of models. For each bin of this histogram, we have selected an equal number of models such that Rg values were evenly distributed within each bin. The selected 6855 models contained both compact and extended proteins with MW in the range 4–410 kDa, Rg in the range 1–14.6 nm, and Dmax in the range 3–51 nm.

## Intrinsically disordered proteins.

To prepare a set of IDP models, we used the Protein Ensemble Database for intrinsically disordered proteins (PED)(Lazar et al., 2021). A snapshot of the database was made that included 172 depositions, 269 ensembles, each ensemble containing between 3 and 29598 models. We have used up to 50 conformers from each ensemble resulting in a total number of 10 089 models. The selected pool of models contained IDPs with MW in the range 0.6–92.6 kDa, Rg in the range 0.5–13.5 nm, and Dmax in the range 1.2–41.3 nm.

Nucleic acids.

For the DNA and RNA models, we used the NDB server (Coimbatore Narayanan et al., 2014). A total of 2864 DNA-only and RNA-only models were obtained with MW in the range 0.5–314 kDa, Rg in the range 0.7–6.8 nm, and Dmax in the range 1.9–21.5 nm.

## Preparing the simulated SAXS data

Theoretical scattering curves were computed on the absolute scale with CRYSOL 2.8 (Barberato et al., 1995) from ATSAS 3.0.3 in the range of momentum transfer from s=0 to s=1.0Å-1 on a grid of 256 points using 99 spherical harmonics. The experimental noise at 7 different protein concentrations c = 0.25, 0.5, 1, 2, 4, 8, and 16 mg/ml, was simulated based on experimental data from the EMBL’s P12 beamline (Blanchet et al., 2015). This corresponded to the data acquired with the sample-to-detector distance of 3 meters, total exposure time of 1 second, and X-ray energy of 10 keV. No structure factor or polydispersity was simulated. The augmented SAXS profiles were normalized to I(0) = 1, and the examples of the simulated data are shown in Figure 1. The ground truth values of MW and Dmax were calculated from the models by CRYSOL. We routinely used GNU parallel (Tange, 2018) to speed up the calculations.

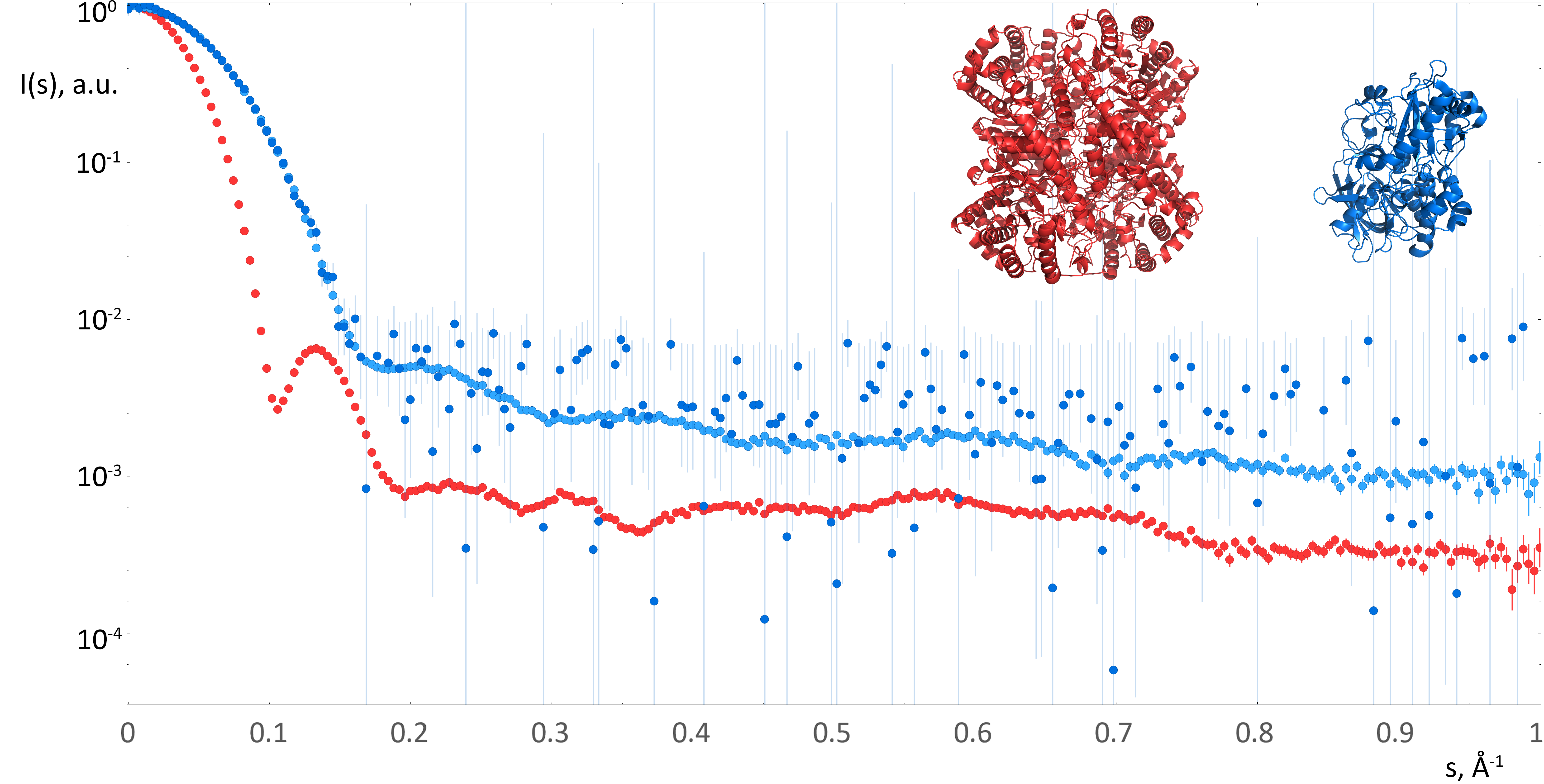


Figure 1. Examples of noise-augmented simulated data from the training set. Red dots: SAXS data computed from xylose isomerase (red model, pdb:1a0d, MW=198 kDa, Dmax=101.5 Å), concentration 16 mg/ml. Blue dots: data computed from oxidoreductase (blue model, pdb: 3b3r, MW=55.7 kDa, Dmax=79.1 Å), light blue dots correspond to the concentration 16 mg/ml, dark blue dots correspond to 0.5 mg/ml.

Neural networks architecture. A feedforward neural network consists of “dense” layers of interconnected units, and each unit of each layer is connected to all units of the next layer (Figure 2). A unit essentially performs a multiple linear regression operation, applies some activation function, and passes the result further to the next layer. Given an input vector , the unit does a dot multiplication of that vector with an internally stored vector of “weights” of the same dimensionality and (optionally) adds a scalar value. The output of the operation reads as:

, (10)

where w is the array of weights associated with the unit, b is the scalar (“bias”), and f is an analytical activation function. In this study, we considered two activation functions: “rectified linear unit” (ReLU) and hyperbolic tangent (tanh).

A neural network contains an input layer, an output layer, and one or more hidden layers (Figure 2). Here, the number of the units of the input layer corresponds to the number of angular intensity points I(s) in the training set data (the experimental uncertainties were not used for training). Since we expect the NN models to predict either MW or Dmax, the output layer consists of a single unit. The minimization algorithm optimizes the weights and biases of all units such that the output layer value becomes as close as possible to the “ground truth” values associated with the input data. This discrepancy is measured by a loss function; here, we used the mean absolute percentage error. Once trained, the NN can be used for predicting the desired parameters from previously unseen input data.

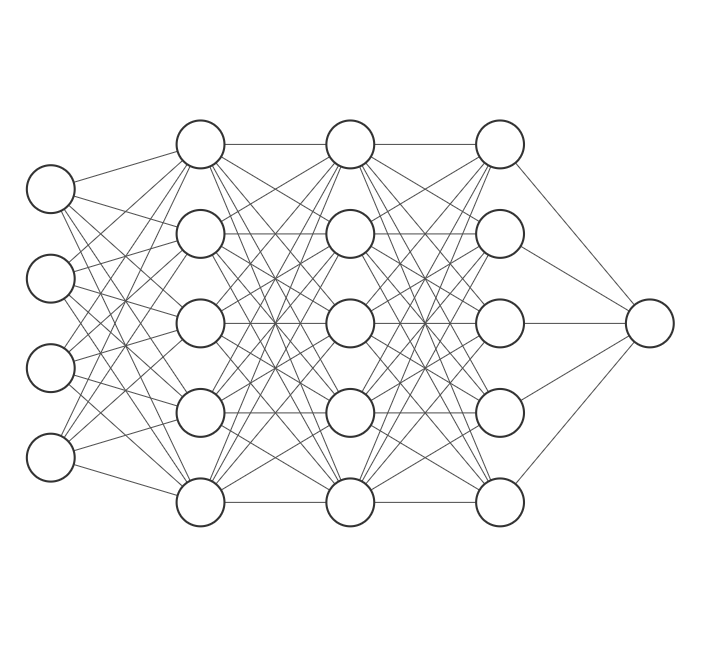
To avoid overfitting – when a NN learns the training set too well and tries to fit specific, non-general features of the training set – a separate validation data set was prepared. During training, the performance of NN is evaluated by applying the loss function to the validation set. Each simulated data set was randomly split into 80% training set and 10 % validation set. The remaining 10% (test set) were used to benchmark the results against other methods.

To find the optimal architecture, we tried different numbers of units and hidden layers to accurately predict MW and Dmax*.* The minimal architecture for the MW prediction was just one hidden layer with five units whereas for Dmax prediction three layers with up to 80 units each were necessary.

Various preprocessing normalizations were tested for input I(s) data and output MW or Dmax values. A simple condition I(0)=1 was found optimal for MW determination; additional subtraction of the mean training set SAXS profile improved the results for Dmax. In both cases, the output values were normalized by the maximum value in the set.

Initially, six NNs were trained (three types of biological macromolecules, for MW and Dmax) on the angular range up to 1.0 Å-1 using smooth data. The accuracy and robustness of MW/Dmax predictions were investigated by re-training the NNs using noisy data and different angular ranges.

In this work, we used the TensorFlow software library with Keras interface (Abadi et al., 2016)in Python. For benchmarking the NNs against other methods we used the DATMW from ATSAS 3.0.3 (Hopkins et al., 2017; Liu et al., 2012; Manalastas-Cantos et al., 2021).



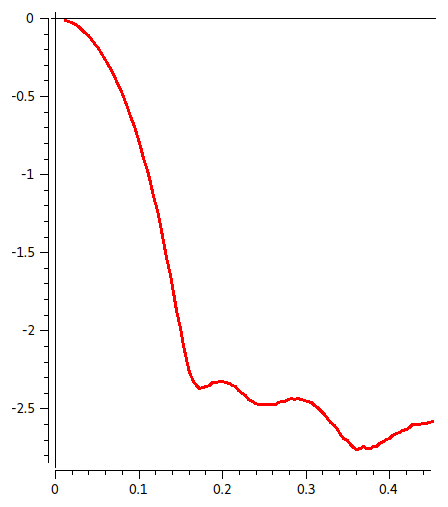
I(s1)

I(s2)

I(s3)

I(s4)

Log I



15.2 nm

Input

layer

Dmax

Output

layer

Hidden layers

256 units

80 units

1 unit

s

Figure 2. The architecture of the neural networks trained for MW/Dmax estimation (an example for Dmax is shown)

## Application to the experimental data

To prepare the input data for the format of the NNs, the further steps are required:

1. estimate I(0) from the Guinier approximation using AUTORG (Petoukhov, 2007);
2. normalize the data to I(0) = 1;
3. convert to Å-1 if necessary; rebin to the grid of the training set.

The angular range of the input SAXS data must match the range used for NN training. The sample type (folded protein/IDP/nucleic acid) must match the applied NN type.

# Results and discussion

To evaluate the performance on the simulated test set data and experimental data from SASBDB (Kikhney et al., 2019) an average relative error was used as a metric of the prediction accuracy

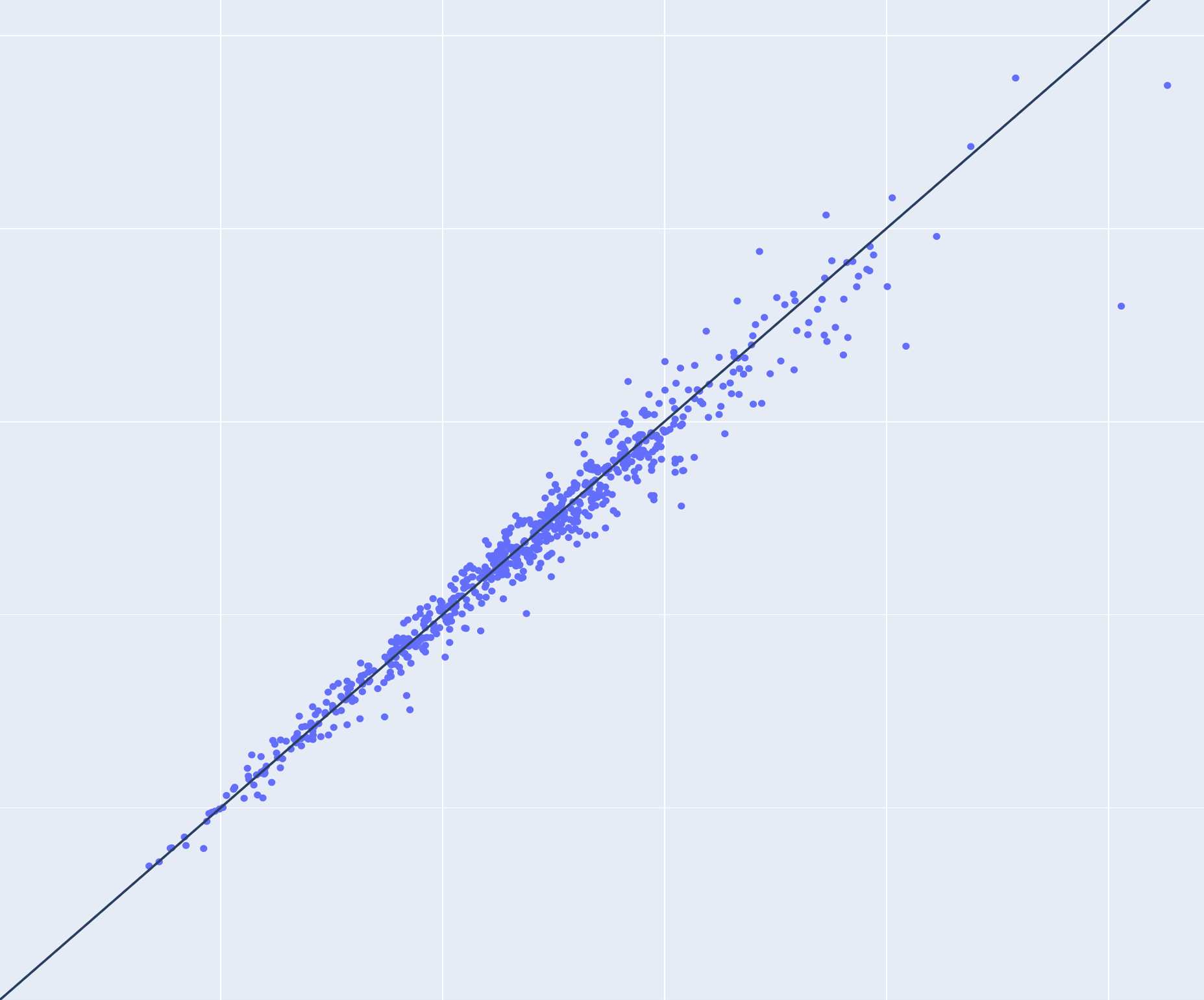
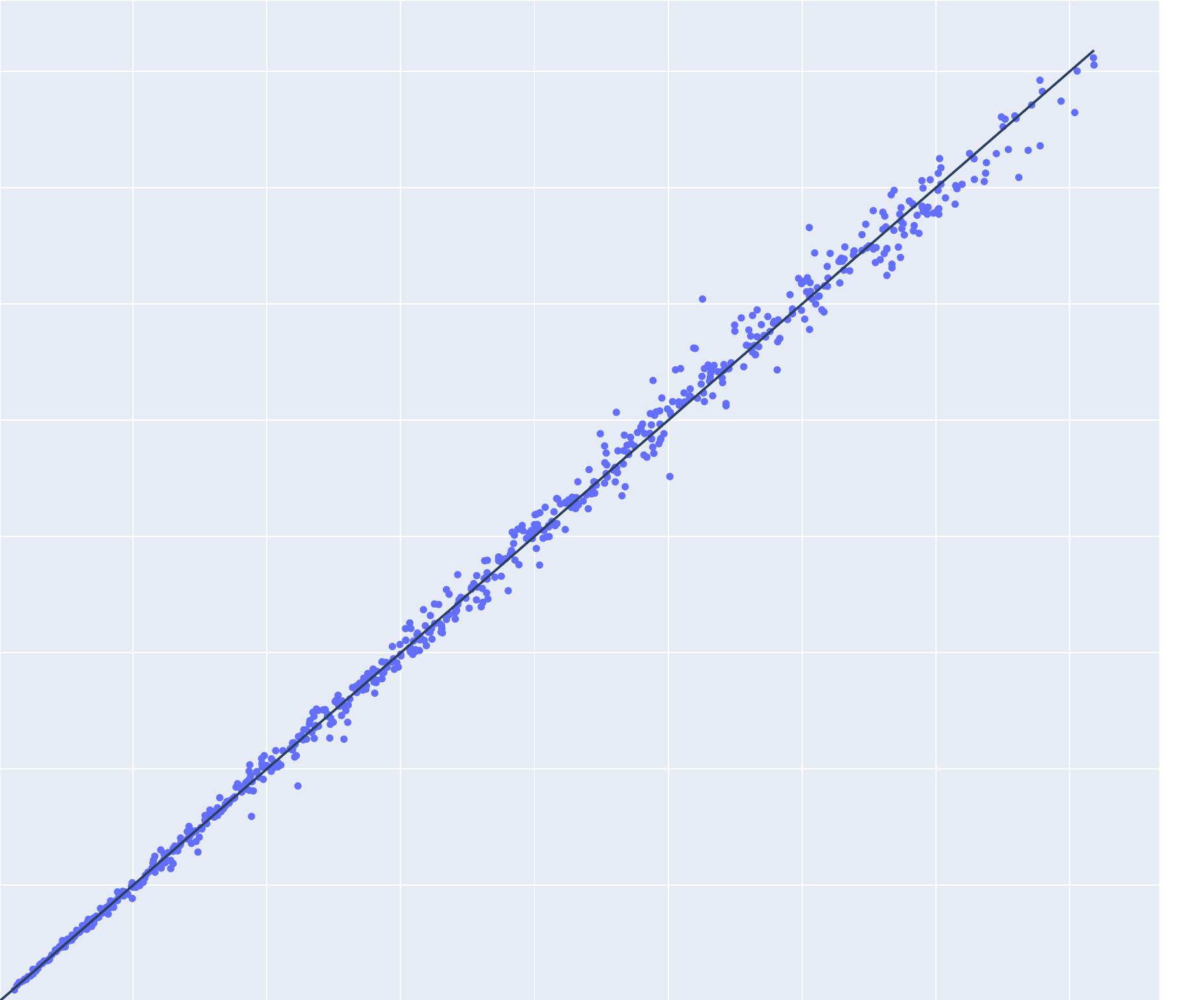
, (11)

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. In addition to the average, we computed the median relative error to control for the skewness of the error distribution.

For NNs trained on smooth (i.e. without added noise) data up to smax = 1.0 Å-1 and applied to the smooth test sets, we obtained the results presented in Table 1. For folded proteins, the plots of the predicted values vs. ground truth values are shown in Figure 3.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | MW average | MW median | Dmax average | Dmax median |
| Folded proteins | 2.50% | 1.87% | 2.78% | 2.13% |
| IDPs | 3.94% | 2.37% | 8.52% | 4.15% |
| Nucleic acids | 2.86% | 2.00% | 2.82% | 1.89% |

Table 1. Performance of the neural networks trained on smooth data and applied on smooth test sets: average and median relative errors.



Predicted MW, kDa

Ground truth MW, kDa

Ground truth Dmax, Å

Predicted Dmax, Å

100

200

300

400

100

200

300

400

100

200

200

100

Figure 3. Predictions from 684 test data sets simulated from folded protein models (without added noise) versus ground truth. Left: molecular weight (MW), right: maximum intra-particle distance (Dmax). Lines of equality are in black.

Angular range. An important question arises: given the maximum angle *smax*, what is the maximum precision of MW and Dmax predictions that one can expect? The use of NNs enables a convenient opportunity to get a deeper insight into the information content of different angular ranges of the SAXS profiles. To evaluate the impact of the angular range on the accuracy of the MW and Dmax predictions, we re-trained the same NNs on smooth data computed from the folded proteins up to various smax values, namely 0.8, 0.6, 0.4, 0.3, 0.2, 0.1, 0.05 and 0.025 Å-1.

For the data cropped at smax=0.1 Å-1, the accuracy of Dmax predictions was 3.3%; it improves up to 2.8% with the angular range increased up to smax=0.4 Å-1 (see Figure 4, purple circles), but the further increase of the angular range did not affect the accuracy of the Dmax predictions. That illustrates the fact, that lower angles in reciprocal space contain information on the larger distances in real space.

­­­­­­­

Figure 4. Relative average MW error (green) and Dmax error (purple) estimated from smooth data decreases with increasing angular range. The light blue and red SAXS profiles (computed from the same models as in fig.1) are shown for demonstrative purposes only.

For MW prediction the impact of higher angles was more pronounced: the accuracy improves from 8% to 2.8% with smax increasing from 0.1 Å-1 to 0.6 Å-1 (see Figure 4, green circles). This is an interesting observation reflecting the fact that the curves normalized to I(0)=1 were utilized in the training such that the direct information about the MW was effectively lost. The network was trained “indirectly” through the geometry of the curve, e.g. the Parseval theorem relations in Equations (1-3). As seen from Figure 4, the intensities at higher angles do provide important information contributing to the overall estimation of the MW.

Effects of the experimental noise. Depending on the sample concentration, contrast, molecule volume, intensity of the X-ray beam, the noise in I(s) may vary drastically. To evaluate how the amount of noise impacts the prediction accuracy, we have added simulated noise to the folded proteins test data set (with known ground truth MW and Dmax) and first applied the above-mentioned NNs trained on the smooth data up to smax = 1.0 Å-1.

For simulated concentrations 4, 8, and 16 mg/ml the average relative MW error was below 3% i.e. comparable to the MW accuracy of the smooth data set, but for lower concentrations, the accuracy decreased significantly, see Figure 5 (blue circles). For the lowest concentrations (0.5 and 0.25 mg/ml) about 2% of the predictions were negative or very close to zero, i.e. the NN failed to produce an MW estimate; without these outliers, the average relative errors were 9.5% (0.5 mg/ml) and 18% (0.25 mg/ml).

Surprisingly, the NN trained to predict Dmax on noise-free data produced almost random outputs when applied to data with noise. Even for the 16 mg/ml test data, the number of negative predictions was 17% and the rest had an average relative Dmax error of 15%. For the lower concentrations, the predictions were practically uncorrelated with the ground truth values.

We have re-trained both NNs using the noise-augmented training set. This led to a significant improvement of the MW predictions for lower concentrations c < 4 mg/ml (see Figure 5, orange circles) and there were no negative output values (i.e. failures). For simulated concentrations ≥ 1 mg/ml, the accuracy of prediction was below 3%. The Dmax predictions became reliable as well with less than 1% failures and average errors below 3.3% for the concentrations higher than 1 mg/ml; at 0.25 mg/ml the average error was 5.8% (which was comparable to the performance of the MW NN) and 2% failures.

Figure 5. Performance of neural networks trained to predict molecular weight on smooth data (blue circles) and trained on noise-augmented data (orange circles) applied to the noise-augmented test set.

Similarly, we trained the NNs on noise-augmented data simulated from IDPs and nucleic acids. To benchmark our results, we applied the NNs and the conventional methods implemented in ATSAS 3.0 (Manalastas-Cantos et al., 2021) to the noise-augmented test sets. The all-to-all comparison is presented in Figure 6, where it is seen that the NNs not only outperform the conventional methods for all types of particles but are also more robust against simulated noise. Indeed, the prediction accuracy by NNs for both MW and Dmax improves gradually with the simulated concentration and reaches a plateau at concentrations above 1 mg/ml.



Figure 6. Average relative errors of the molecular weight (MW, left) and maximum intra-particle distance (Dmax, right) predictions for folded proteins (top), intrinsically disordered proteins (IDPs, middle), and nucleic acids (RNA and DNA, bottom) vs. simulated concentration. Comparison of conventional methods (colored circles/lines) with the NNs predictions (black circles/lines). Dashed lines represent methods not directly applicable for estimating MW from nucleic acids data.

­­­­­­

Overall, IDPs happened to be the most challenging objects for predictions. The conventional methods failed to produce reasonable MW estimates with the <Δrel> in the range of 20%–50%, while the NN showed much better results of 3-10% enabling to reliably estimate the MW of IDPs from SAXS data.

As a note, conventional methods for MW estimation were developed for proteins and are thus not directly applicable to data from nucleic acids. In the case of Vc method, we have used the empirically determined coefficients (eq.7) reported by (Rambo and Tainer, 2013). Based on our training set, we have also established empirical correction factors for MW estimation for nucleic acids for Porod’s method and SAXSMoW.

Experimental data.

To evaluate the performance of our approach for experimental data, one requires the SAXS data collected from well-characterized monodisperse solutions with reliably determined MW and Dmax as “ground truth” values. For folded proteins, we used data from 29 SASBDB (Kikhney et al., 2019) entries that were tagged “Benchmark” and, with a few exceptions, fitted by atomic models. The “ground truth” MW values were calculated from the protein sequence, the “ground truth” Dmax values were obtained from the models. The neural networks were retrained using the same training set but on the shortest common experimental data angular range of the SASBDB-deposited data, namely 0.02 < s < 0.3 Å-1.

The average relative MW and Dmax errors were 10% and 7%. We have inspected the cases where the predictions were the least accurate. In the case of apoferritin, the MW was underestimated by 22% which was expected because the MW of apoferritin (479 kDa) is beyond the range of the training set (up to 410 kDa). In the case of ribonuclease (16.5 kDa) the MW was underestimated by 30% and Dmax was overestimated by 11% because a large part of the protein (17% in sequence) is flexible and not present in the model (PDB: 3MZQ). The detailed results are summarized in Supplementary Table 1.

To study the reproducibility of MW and Dmax predictions from experimental data, we used 100 background-subtracted data sets from bovine serum albumin (BSA), entry SASDDN3 (Franke et al., 2018). The data were collected at the EMBL P12 beam line (Blanchet et al., 2015) from 2.25 mg/ml solution of BSA, exposure time 50 ms. For MW the obtained average prediction was 73.8 kDa, standard deviation 2.3 kDa, for Dmax the average was 108 Å and the standard deviation 4 Å. The determined values somewhat exceed those expected for a monomeric protein in agreement with the fact that the BSA sample reveals a partial dimerization in solution, as indicated by (Franke et al., 2018).

### Current limitations and perspectives

The present approach works only for macromolecules within the MW and Dmax ranges covered by the training sets. The predicted values might become meaningless if the NN fails to make a reasonable prediction e.g. if the input data are too different from the training set. One could further expand the applicability of the trained NNs by scaling the input data angular range and adjust the predicted parameters accordingly.

Several approaches can be used to expand the applicability of the approach. First, to generalize the applications to experimental data the NN can be pre-trained using several typical angular ranges such that the most suitable network can be automatically employed for the given experimental data set. Moreover, the basis for “ground truth” can be further extended. Here, we have used only experimentally determined models of proteins and nucleic acids. It is possible to further enhance the folded proteins training set by using models computed by Alpha-Fold (Jumper et al.) or other structure prediction approaches; the IDPs training set is amendable e.g. by RANCH (Tria et al., 2015). Similarly, one may generate training sets for the models of folded proteins containing significant proportions of flexible chains (see the above case of ribonuclease). The extension of the nucleic acids training set is possible by using software for secondary (e.g. Mfold (Zuker, 2003)) and tertiary (e.g. OligoAnalyzer (Owczarzy et al., 2008)) structure predictions.

To estimate the confidence intervals of the predicted values, one could apply an ensemble of independently trained NNs or snapshots of a single neural network, converging to several local minima along its optimization path (Huang et al., 2017). Alternatively, one may determine the variability of the predicted values by resampling the input data (i.e. adding pseudo experimental noise) using DATRESAMPLE (Manalastas-Cantos et al., 2021).

To further expand the applicability of NNs to experimental data on systems with significant interaction effects, one could augment the training set by simulating the structure factor, adding systematic noise, simulating polydispersity.

# Conclusions

We presented a novel method for the estimation of primary SAXS parameters using neural networks. A comparison with existing methods applicable to folded proteins demonstrated that the NN approach provides higher accuracy and is robust against noise. The NN method is not restricted by assumptions (e.g. homogeneity of the electron density), and it does therefore allow one to further reassess the real capacities of SAXS data in terms of information content and to improve the accuracy of primary SAXS data analysis beyond the commonly accepted uncertainty (e.g. about 10% for MW).

To the best of our knowledge, our method is a conceptually new approach to reliably estimate the molecular weight from the SAXS data by intrinsically disordered proteins and nucleic acids. The Dmax estimations by our method do not require the analysis by indirect Fourier transformation and can therefore be conducted directly from experimental data. The developed programs for MW and Dmax estimation are being included in the next ATSAS package release, free for academic use.

# Acknowledgments

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

# References

Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., Devin, M., Ghemawat, S., Irving, G., Isard, M., et al. (2016). TensorFlow: A System for Large-Scale Machine Learning.

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, U. (2005). The Protein Data Bank. In Structural Bioinformatics, pp. 181–198.

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.

Blanchet, C.E., Spilotros, A., Schwemmer, F., Graewert, M.A., Kikhney, A., Jeffries, C.M., Franke, D., Mark, D., Zengerle, R., Cipriani, F., et al. (2015). Versatile sample environments and automation for biological solution X-ray scattering experiments at the P12 beamline (PETRA III, DESY). J. Appl. Crystallogr. *48*, 431–443.

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.

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

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.

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.

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.

Huang, G., Li, Y., Pleiss, G., Liu, Z., Hopcroft, J.E., and Weinberger, K.Q. (2017). Snapshot Ensembles: Train 1, get M for free. 5th Int. Conf. Learn. Represent. ICLR 2017 - Conf. Track Proc.

Jumper, J., Evans, R., Pritzel, A., Green, T. (2021). Highly accurate protein structure prediction with AlphaFold. Nature, 1-11.

Kikhney, A.G., Panjkovich, A., Sokolova, A. V., and Svergun, D.I. (2016). DARA: a web server for rapid search of structural neighbours using solution small angle X-ray scattering data. Bioinformatics *32*, 616–618.

Kikhney, A.G., Borges, C.R., D.S., 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.

Lazar, T., Martínez-Pérez, E., Quaglia, F., Hatos, A., Chemes, L.B., Iserte, J.A., Méndez, N.A., Garrone, N.A., Saldaño, T.E., Marchetti, J., et al. (2021). PED in 2021: a major update of the protein ensemble database for intrinsically disordered proteins. Nucleic Acids Res. *49*, D404–D411.

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.

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.

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.

Owczarzy, R., Tataurov, A. V., Wu, Y., Manthey, J.A., McQuisten, K.A., Almabrazi, H.G., Pedersen, K.F., Lin, Y., Garretson, J., McEntaggart, N.O., et al. (2008). IDT SciTools: a suite for analysis and design of nucleic acid oligomers. Nucleic Acids Res. *36*.

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.

Petoukhov, M.V., Konarev, P.V., Kikhney, A.G., and Svergun, D.I. (2007). ATSAS 2.1 – towards automated and web-supported small-angle scattering data analysis. Urn:Issn:0021-8898 *40*, s223–s228.

Piiadov, V., Ares de Araújo, E., Oliveira Neto, M., Craievich, A.F., and Polikarpov, I. (2019). SAXSMoW 2.0: Online calculator of the molecular weight of proteins in dilute solution from experimental SAXS data measured on a relative scale. Protein Sci. *28*, 454–463.

Porod, G. (1951). Die Röntgenkleinwinkelstreuung von dichtgepackten kolloiden Systemen. Kolloid-Zeitschrift 1951 1242 *124*, 83–114.

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

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

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

Svergun, D., Koch, M., Timmins, P., and May, R. (2013). Small angle X-ray and neutron scattering from solutions of biological macromolecules.

Tange, O. (2018). GNU Parallel 2018, March 2018.

Tikhonov, A. N., & Arsenin, V. Y. (1977). Solutions of ill-posed problems. *New York*, *1*(30), 48Tikhonov, A. N., & Arsenin, V. Y. (1977). Solutions of ill-posed problems. New York, 1(30), 487.7.

Tria, G., Mertens, H.D.T., Kachala, M., and Svergun, D.I. (2015). Advanced ensemble modelling of flexible macromolecules using X-ray solution scattering. IUCrJ *2*, 207–217.

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

Zuker, M. (2003). Mfold web server for nucleic acid folding and hybridization prediction. Nucleic Acids Res. *31*, 3406–3415.