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. Trained on synthetic SAXS data, the feedforward 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 SAXS 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 SAXS data, as well as many other useful utilities, are freely available at the git-hub for academic use.

# Introduction

Small angle X-ray scattering (SAXS) 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 SAXS 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 widely available.

SAXS data are obtained by illuminating a dilute, typically monodisperse solution of macromolecules with a monochromatic X-ray beam which results in an isotropic 2D scattering pattern. The 2D pattern can be azimuthally integrated into 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 scattering 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 the solution scattering. From the background-subtracted scattering profile, one can estimate structural characteristics of the scattering particle: radius of gyration (Rg), maximum intraparticle distance (Dmax), pair-distance distribution function (p(r)), molecular weight (MW). Given these parameters, it is possible to reconstruct the overall shape *ab initio* or obtain a hybrid model employing structural information from the high-resolution methods.

…

…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.

Well-established methods for the estimation of MW have been primarily developed for globular proteins and their applicability to SAXS data from disordered proteins and nucleic acids is questionable. The optimistic estimate of MW precision is 10% [Nelly, 2018]. Here, we explore the applicability of artificial neural networks (NN) to primary SAXS data analysis: estimation of MW and Dmax for data from folded, intrinsically disordered proteins and nucleic acids.

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).

Neural networks are excellent tools for supervised learning; the taskof learning a function that maps an input to the desired output based on example data sets. In the case of SAXS, the input is a vector of experimental intensities and the output is MW or Dmax value. Instead of experimental data, one could compute the scattering from known proteins and nucleic acids models for training assuming that the learned 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 any SAXS data, reduces the information content in experimental SAXS data thus increasing the ambiguity of data interpretation.

To make a NN more robust against any specific source of distortions, e.g. gaussian or systematic noise from a beamline, one can generate a realistic noise and augment the training set with that noise. 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. 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 number of feedforward artificial NNs trained on synthetic SAXS data generated from thousands of experimentally determined models to estimate MW and Dmax from folded, unfolded proteins; nucleic acids. Here we demonstrate, that our method has much higher accuracy and is less demanding in terms of data quality compared to the well-established methods evaluated in this study. Moreover, our method to the best of our knowledge is unique in the SAXS 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-principle capabilities of NNs in the field of SAXS, whereas further improvement in accuracy and stability is possible and limited only by the choice of the training set.

# Methods for MW determination.

To date, there are several well-established techniques for MW estimation, that could be subdivided into two major categories: concentration-dependent and concentration-independent methods. The former account for the dependence of forward 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) and precise measurement of the sample concentration. The latter methods utilize a single background-subtracted profile and require no additional *a priori* information on sample concentration. In circumstances where the protein concentration cannot be accurately measured (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 focus on 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 (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 equals MW/V = 0.625 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, assuming globular proteins 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. This method can be very effective when applied to globular proteins. However, it is very sensitive to the accuracy of background subtraction.

**SAXSMoW method.** The accuracy of the Porod’s method was improved by (Fischer et al., 2010). In this approach, the authors integrate the Porod invariant in (1) not up to infinity, but up to smax:

. (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 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 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 for proteins, but for RNA data. 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 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 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.

# Methods for maximum intraparticle distance Dmax determination.

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

Training/validation/test sets

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

### Folded proteins.

A total of 135238 atomic coordinate files describing protein structures from protein-only biological assemblies were obtained from the protein databank (PDB) (Berman et al.). 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 radii of gyration of those models are evenly distributed within each bin. Therefore, the selected pool of 6855 models contained both compact and extended proteins of MW in the range 4–410 kDa, Rg in the range 1–14.6 nm, 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.). A snapshot of the database was made that included 172 depositions, 269 ensembles, each ensemble contained 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 of MW in the range 0.6–92.6 kDa, Rg in the range 0.5–13.5 nm, Dmax in the range 1.2–41.3 nm.

Nucleic acids.

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

## Preparing the simulated SAXS data

Theoretical scattering curves were computed on the absolute with CRYSOL 2.8 (Barberato et al., 1995) from ATSAS 3.0.3 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.) that corresponds to the data acquired with the sample-to-detector distance of 3 meters, 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. The ground truth values of MW and Dmax were calculated from the models by CRYSOL. We routinely used GNU parallel [O. Tange (2018): GNU Parallel 2018, March 2018, <https://doi.org/10.5281/zenodo.1146014>] to speed up the calculations.

[Figure with simulated data from two very different models with different noise]

Neural networks architecture. A feedforward neural network consists of “dense” layers of interconnected units, each unit of each layer is connected to all units of the next layer (fig.1). A unit essentially performs a multiple linear regression operation, then 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:

, (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”).

NN contains an input layer, an output layer and one or more hidden layers (fig.1). The number of the units of the input layer corresponds to the number of I(s) points in the training set data (in this study the experimental uncertainties were not used for training). Since we expect the NN models to predict either MW or *Dmax*, the output layer contains 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” value associated with the input data. This discrepancy is measured by a loss function; in this study 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 is 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 MW prediction was just one hidden layer with five units whereas for Dmax prediction three layers with up to 80 units were necessary.

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

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 [https://www.tensorflow.org/about/bib] in Python. For benchmarking NNs against other methods we used the DATMW from ATSAS 3.0.3 (Manalastas-Cantos et al., 2021).



80 units

1 unit

256 units

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

## Application to 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.

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).

# Results and discussion

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.

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

Benchmark.

To evaluate the performance on the simulated test set data and experimental data from SASBDB [Ref.] we used the average relative error as a metric of the prediction accuracy:

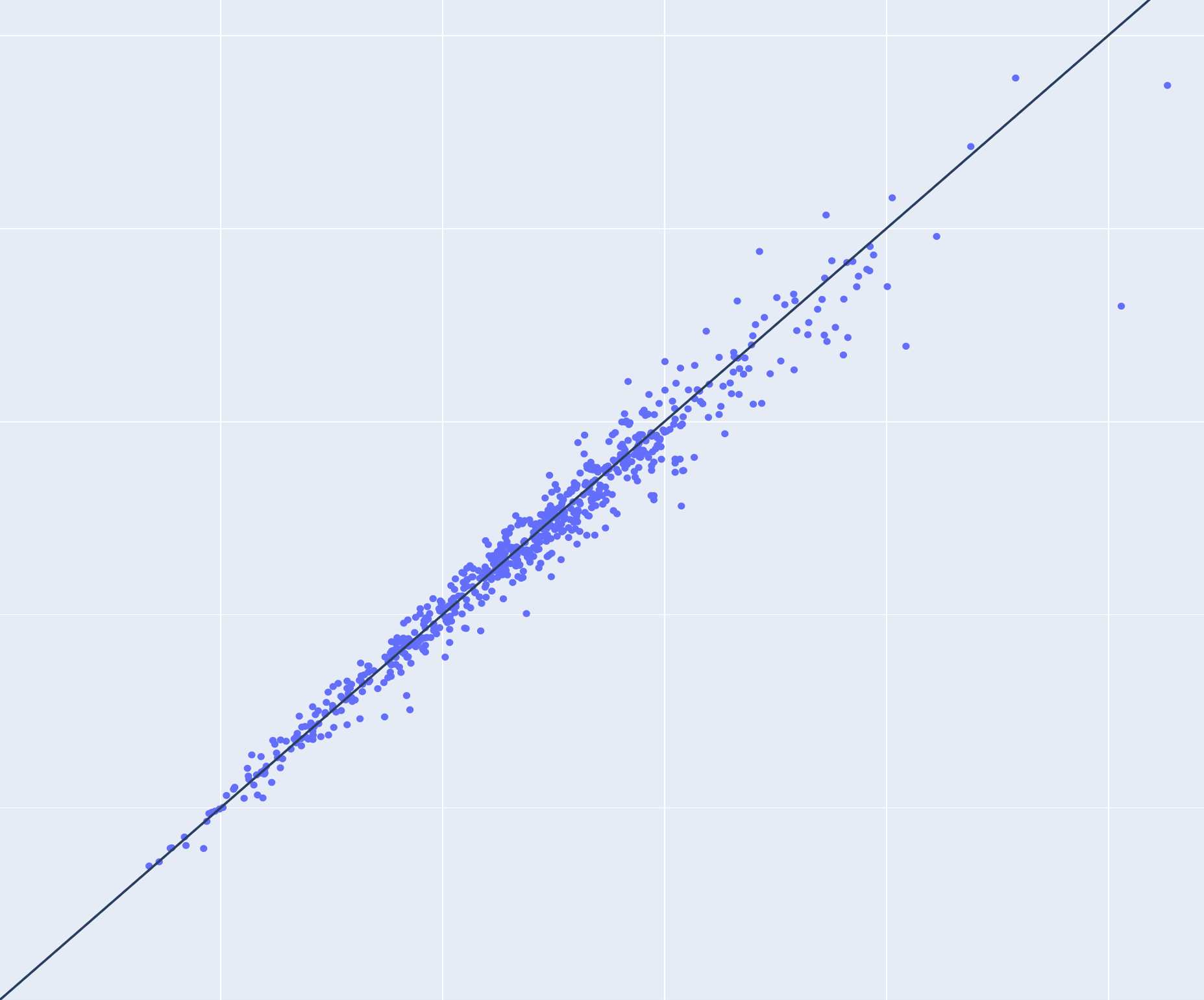
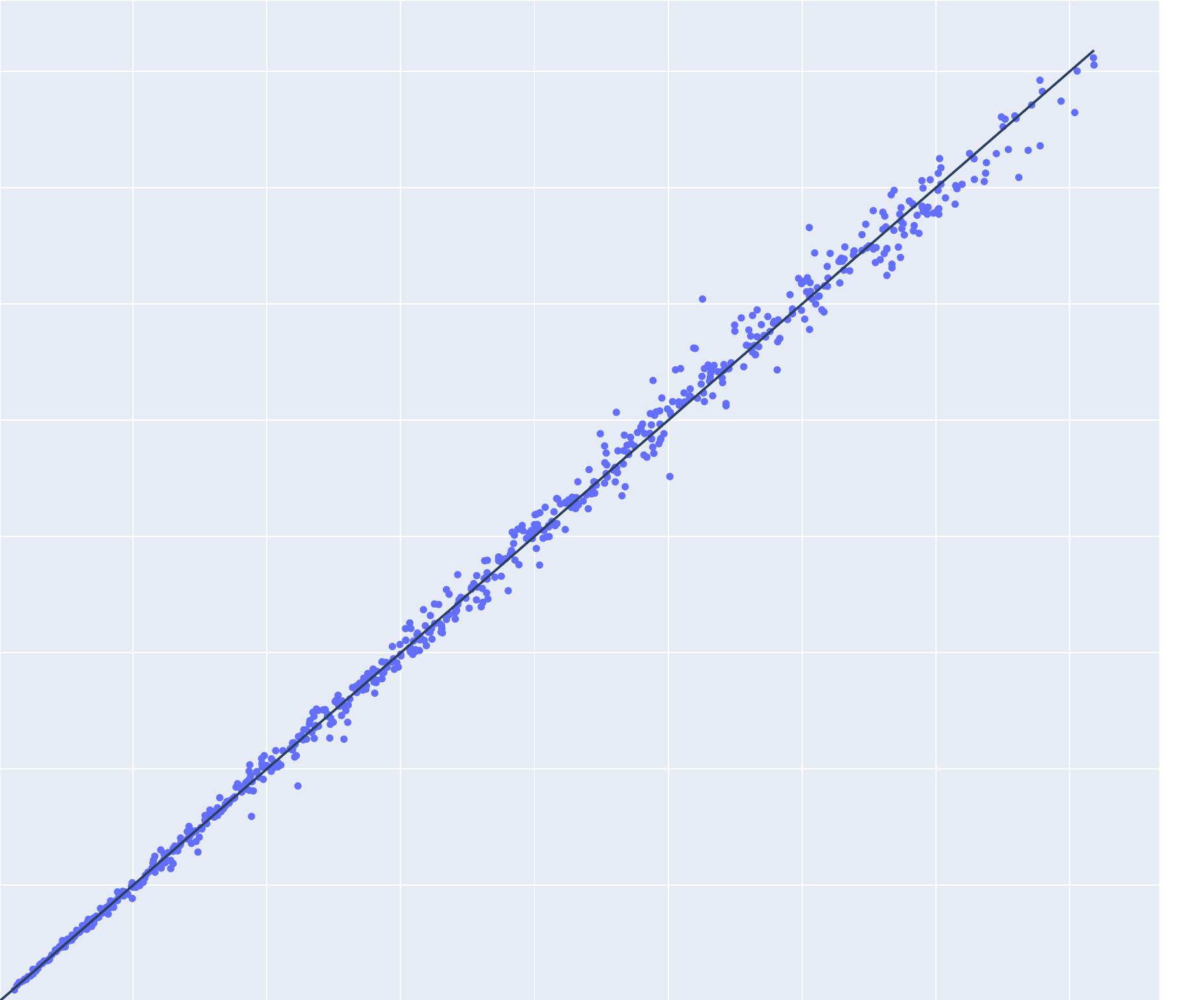
, (11)

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. Since the test set was generated to comprise models of different sizes and shapes, this value represents not only the accuracy of the given method but also its ability to work on data models of various shapes and sizes.

For NNs trained on smooth (i.e. without added noise) data (0 < s ≤ 1.0 Å-1) and applied to the smooth test sets we obtained the results presented in Table 1. For folded protein data the plots of the predicted values vs. ground truth values are shown in Fig. 6.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | MW average | MW median | Dmax average | Dmax median |
| Folded proteins | 2.50% | 1.87% | 3.45% | 2.55% |
| 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: average and median relative errors applied on smooth test set



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 1. Predictions from 686 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 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 SAXS profiles and to empirically determine the dependence of the accuracy of the MW and *Dmax* estimates against the *smax*. 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: 0.8, 0.6, 0.4, 0.3, 0.2, 0.1, 0.05, 0.025 Å.

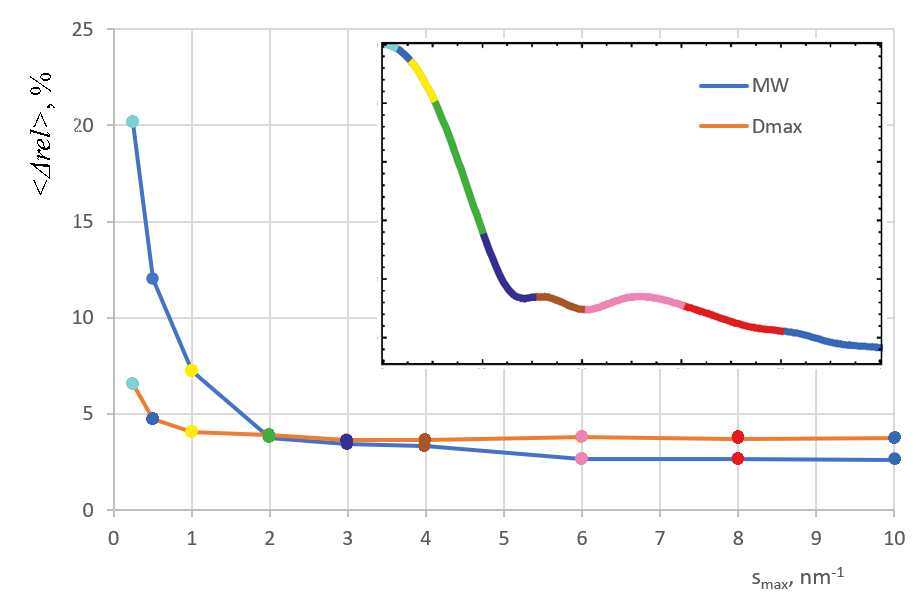


Fig.7 Errors versus maximal angular range smax for noiseless synthetic SAXS profiles

If the data is up to smax=0.1 Å-1, the accuracy of Dmax predictions on simulated data is 4.1%. It improves up to 3.7% if the angular range is increased up to smax=0.3 Å-1. The presence of higher angles does not affect the accuracy of Dmax prediction. That illustrates the fact, that lower angles in reciprocal space contain information on the larger distances in real space.

For MW prediction the impact of higher angles is more pronounced: the accuracy increases from 7.2% to 2.6% with smax increasing from 0.1 Å-1 to 0.6 Å-1. The intensities in this angular range mostly contain information on the molecule surface and inner structure, which indeed may contribute to the overall estimation of the MW.

Experimental noise. Depending on the sample concentration, contrast, molecule volume, intensity of the X-ray beam, the amount of noise may vary drastically. To evaluate how the amount of noise impacts the accuracy of prediction, we have added simulated noise to the smooth data with known ground truth MW and Dmax and applied NNs trained on the smooth data. For simulated concentrations above 2 mg/ml the median relative MW error was 2%, but for lowest concentrations (c < 1 mg/ml) median values were 7-12% and about 2% of the predictions were negative or very close to zero. For Dmax the predictions were uncorrelated with the ground truth values. We have re-trained the NNs using the noise/augmented training set which led to a significant improvement of the MW predictions on the lower concentrations c < 4 mg/ml (fig. MM) and made the Dmax predictions work. For concentrations above c=2mg/ml the median relative error was 2%.

Figure 2. Trained on smooth data vs. trained on noise-augmented data. median relative error

Similarly, we trained the NNs on noise-augmented data simulated from IDPs and nucleic acids. To benchmark our results, we applied the NNs and conventional methods implemented in ATSAS 3.0 [Ref] to noise-augmented test sets. 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 more robust against simulated noise.

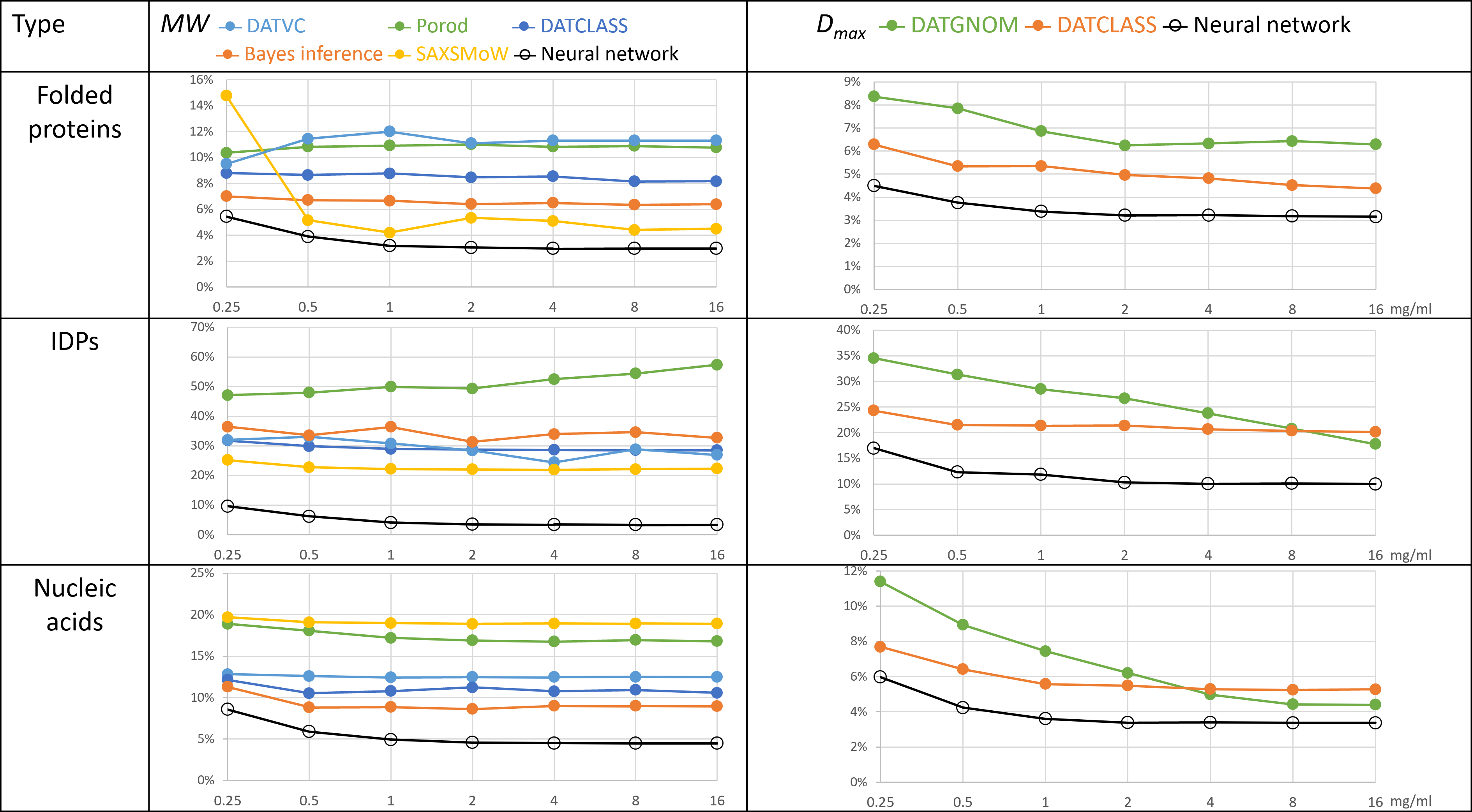


Figure 5. Average relative errors of MW and Dmax predictions – comparison of conventional methods (colored circles/lines) with the NN predictions (black circles/lines). six neural network models trained on noise-augmented data Comparison of performances of different methods (proteins, NAs, IDPs).

­­­­­­

The accuracy of predictions by NNs for both MW and *Dmax* improves gradually with the simulated concentration and reaches a plateau at concentrations above 1-2 mg/ml.

For folded proteins the accuracy of prediction for higher simulated concentrations is just

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 SAXSMoW 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 SAXS 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 SAXS profile 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 SAXS analysis, we decided not to overfocus on this issue and leave it beyond the scope of this article.

Experimental data.

To evaluate the performance on real experimental data, we needed data collected from well-characterized monodisperse solutions with reliably determined MW and Dmax “ground truth” values. For folded proteins we used data from 29 SASBDB [ref] entries that were tagged “Benchmark” and, with some 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 on the least common angular range 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 most inaccurate. In 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 case of ribonuclease (16.5 kDa) the MW was underestimated by 30% and Dmax was overestimated by 11% – possibly because 17% of the protein is flexible and not present in the model (PDB: 3MZQ). The detailed results are summarized in supplementary Table s1.

To study the effects of experimental noise on MW and Dmax predictions, we used 100 background subtracted experimental data sets from SASDDN3 [ref machine learning]. The data were collected at the EMBL P12 beam line [ref] from 2.25 mg/ml solution of bovine serum albumin, 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.

### Current limitations and perspectives

The presented approach works only for macromolecules within the MW/Dmax ranges that were used for training. One could expand the applicability of the trained NNs by scaling the input data angular range and adjust the predicted parameters accordingly.

To expand the applicability, one could expand the training set. In this study 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 [ref.] or other structure prediction approaches; the IDPs training set is amendable by RANCH [ref. EOM 2.0]. The extension of the nucleic acids training set is possible by using software for secondary (e.g. Mfold (research and 2003)) and tertiary structure (e.g. OligoAnalyzer (Owczarzy et al.)) prediction.

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 [arXiv:1704.00109]. Alternatively, one could determine the variability of the predicted values by resampling of the input data (i.e. adding pseudo experimental noise) using DATRESAMPLE [ref to ATSAS 2020].

Train on data with SF and/or systematic noise. Polydisperse data.

Estimate experimental data quality: noise level, resolution(?!), problems with sample/instrument.

# Conclusions

We presented a novel 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).

# Acknowledgments

The work was supported by the BMBF grant 16QK10A (SAXS-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](https://www.cell.com/structure/article-types" \t "_blank). 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" \t "_blank). If you have any questions or concerns, please let me know and I’d be happy to assist you.