SASBDB
Small Angle Scattering Biological Data Bank

Depositing small-angle scattering data and models to the Small-Angle Scattering Biological Data Bank (SASBDB).

Introduction.

The following guide provides a basic outline of the minimum requirements necessary to deposit small-angle scattering data to SASBDB as well as any applicable probable-real-space scattering-pair distance distributions (or $p(r)$ vs r profiles), associated 3D spatial models (e.g., dummy atom, bead, or atomistic models) and corresponding data-model fits.

This guide illustrates the deposition of small-angle X-ray scattering (SAXS) data and models obtained from a protein in solution (calmodulin). Although the data processing and analysis for this example has been performed using the ATSAS package, additional file formats from other data processing suites are welcomed (for example, GIFT, FoxS, etc). If you have developed your own data analysis and modelling tools (that are not included in the file format options already available in SASBDB), feel free to contact info@sasbdb.org and we can work with you to upload your SAS project to the database.

What you will need before your deposition.

1) The SAS data that has been reduced and background (buffer) subtracted (i.e., a 1D subtracted scattering profile). By definition SASBDB uses the convention that the scattering intensities, I, are plotted *vs* the momentum transfer, s i.e.,:

   $$I(s) \text{ vs } s,$$

   where $s = 4\pi\sin\theta/\lambda$, 2θ is the scattering angle and λ the incident radiation wavelength. **Note!** the default units of s in SASBDB are nm^{-1}. Therefore, all structural parameters (radius of gyration, R_g, maximum particle dimension, D_{max}, volume(s), etc) should be quoted in units of nm (or nm^3 in the case of volume).
2) The $p(r)$ vs r file (if applicable; in some cases uploading a $p(r)$ vs r profile for a mixture may not make sense).
3) Models (in .pdb format) and the data-model fit files. **Note!** For *ab initio* models calculated using DAMMIN, DAMMIF or GASBOR, an appropriate .fir file should be uploaded; for all other models upload the .fit file to the database. For ensemble optimisation method (EOM) it is also advised to upload the R_g distribution file. **Note!** DAMAVER.pdb models are not accepted; DAMFILT.pdb models are accepted, but must be uploaded in combination with the best-fit individual DAM model, or refined DAM model, with the corresponding individual DAM model .fir file.
4) Information about your sample.
   a. Copy and paste the ***exact*** one-letter amino-acid (or DNA/RNA) sequence of the macromolecule(s) that *produced the experimental scattering profile*. **Note!** For proteins this means including any additional N- or C-terminal amino acid extensions to the main sequence, e.g., C- or N-terminal affinity tags or additional amino acids derived from the multi-cloning site of a plasmid.
   b. The stoichiometry of your macromolecule(s).
   c. For proteins, the UNIPROT identification number of the protein, organism name, and corresponding amino-acid range relative to the wild-type protein if a truncation mutant is investigated (http://www.uniprot.org/).

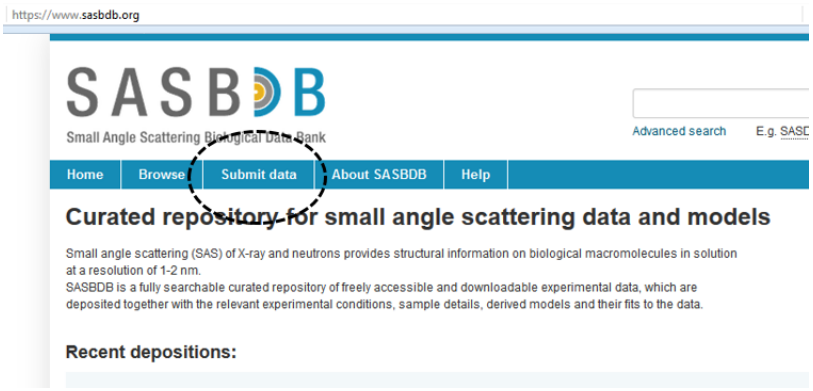
d. The exact solvent (i.e., buffer) composition including the pH.
e. Information about your SAS experiment: radiation wavelength (in nm), temperature of the sample during data collection (°C), how long the sample was exposed for (s); the number of individual exposures and, most importantly, the sample concentration in mg/ml (or the concentration range if you elect to deposit a merged or extrapolated SAS profile).

Starting a deposition.

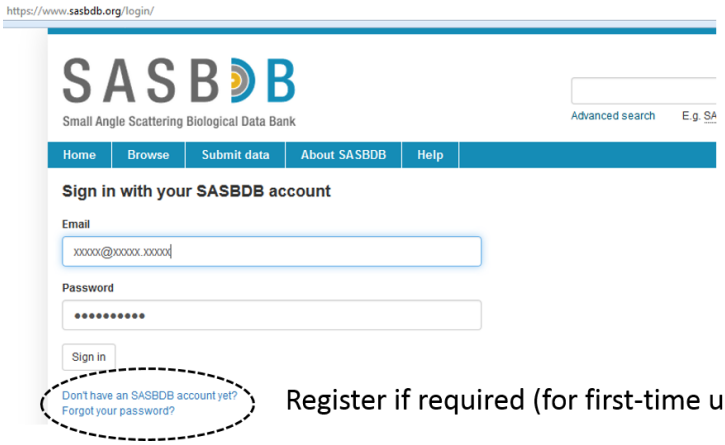
1) Go to the following:

www.sasbdb.org

and click on 'Submit data'.



2) Sign in using your user-name and password (this is different to ATSAS-online). Register for an account if you are a first time user, or if you have forgotten your password.



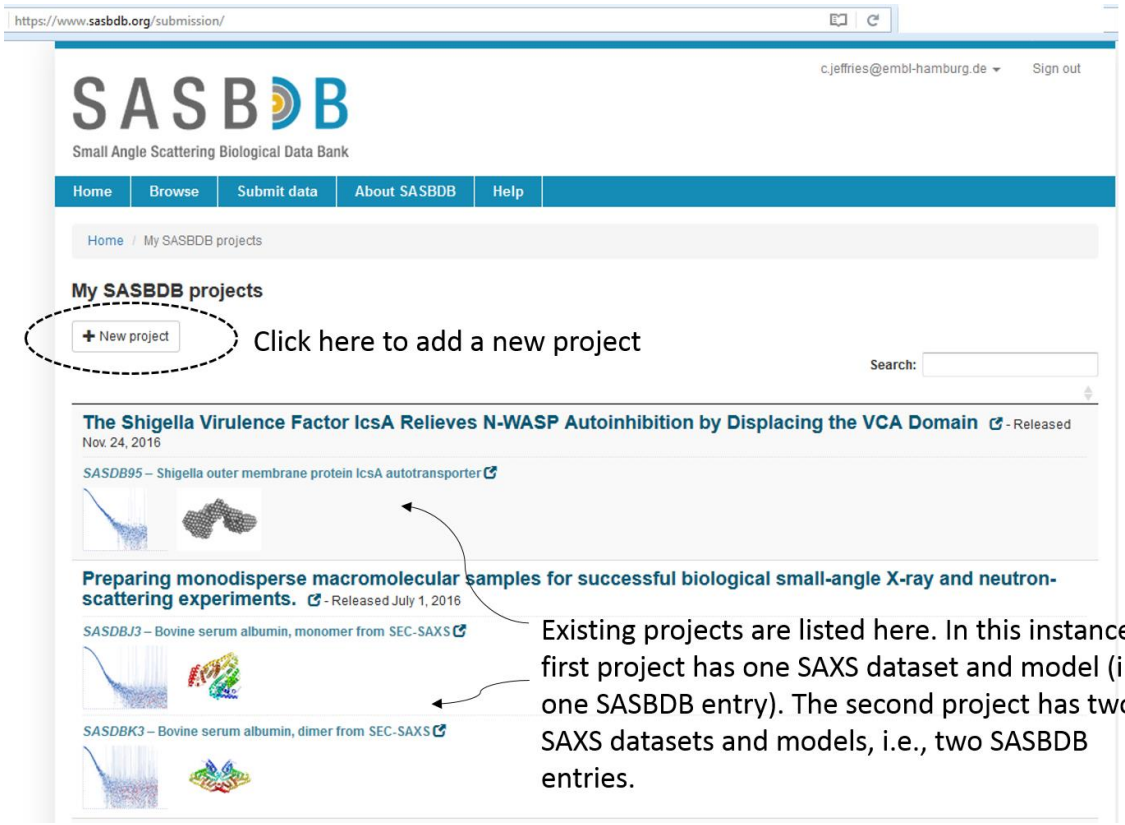
Register if required (for first-time users)

3) Once logged into SASBDB, you will be directed to your own exclusive 'Projects Page' and here you can add a new project.

What is a project?

A project is a single-set, or combined group, of scattering data and models that are somehow related under a common/single title, e.g., a specific publication, a publication for submission, a manuscript in preparation, a report, etc. For example, for manuscript-X you may have 5 different protein SAXS data sets. Therefore, create a new project specifically for manuscript-X and then deposit all five scattering curves/models under the common 'Manuscript X' project title. You will generate five SASBDB entries for the one project (*one* scattering profile per entry). However, for a second publication, publication Y, you may have only collected one SAXS data set. Therefore, add another new project called 'Publication Y' and then deposit the SAXS data into this second unique project.

Note! It is ***not*** recommended to group SAS data and models using a common macromolecule name. For example, you have collected 25 calmodulin SAS datasets over the past ten years and have published seven calmodulin publications. It would be tempting to put all of these data into a project simply called 'Calmodulin', however this will not work. The 25 datasets actually belong to 7 different projects (one new project per publication) and need to be allocated accordingly.



4) Click on **+New Project**. You will have the choice of selecting whether the data are already-published or not published.

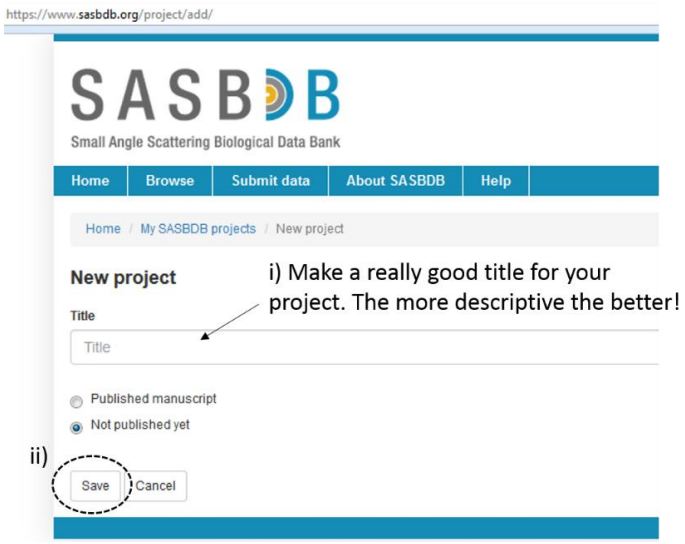
i) Published data: **Note.** Data associated with an published article will be released almost immediately after quality assurance checks and the assignment of a SASBDB accession code. The entry will be made publicly available on the SASBDB website.

Select 'Published manuscript'

The title of the publication becomes the title of the Project. You have options to delete the publication if you make a mistake.

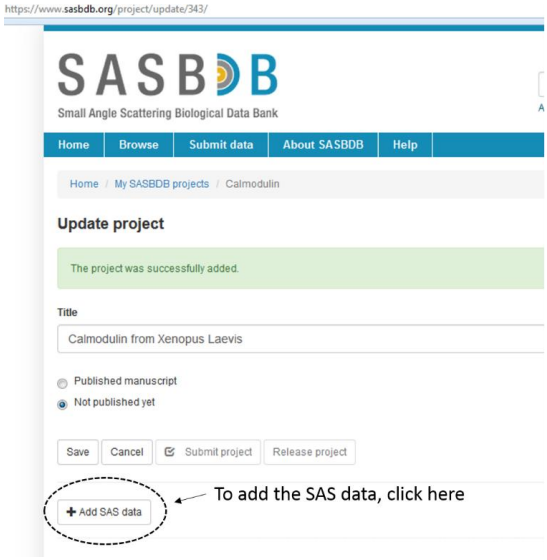
ii) Unpublished data: **Note.** Data associated with work that has not been published will *not* be released until the release is approved by you (from your SASBDB account) or after six months on hold. After quality assurance checks, the SASBDB accession code will be linked to an additional URL key-string that only you (or whoever you give the URL to) can access.

Select 'Not published yet'



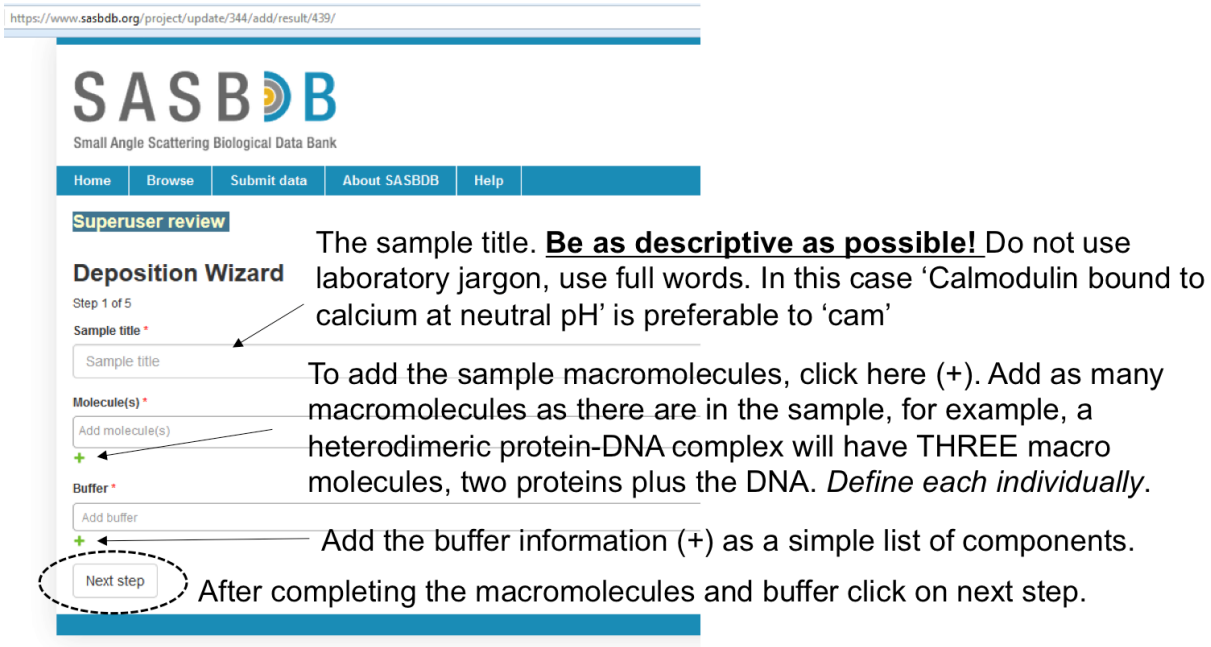
Note. The title of the unpublished project will change to the title of your manuscript if your data becomes published. However, you may release unpublished data to the public at any time; you do not need a publication to release an entry.

Whether you elect to deposit published or unpublished data, the system will direct you to the next stage: adding the SAS data and sample information. **Note.** Once a project has been written, you can return to the project at any time to draft your depositions. The deposition information for each project entry is saved on-the-fly as you progress from one step to another. There is no need to complete an entry all-in-one go. You can also delete individual entries in a project *prior to submission*. However, as soon as you have submitted your entries to SASBDB and the data and models have been accepted and assigned a SASBDB accession code, the submissions *cannot* be deleted through your SASBDB account.



The deposition process consists of five steps. When asked, ***fill in as much information as possible into the boxes of the submission forms***, especially for Steps 1, 2 and 3.

Step 1. Sample title, sample macromolecules and buffer (solvent) information.



The **'Add molecule' pop-up window** will appear when you click on + under the molecule(s). Have your sequence information at hand!

*Select the type of macromolecule in the sample (protein, RNA, etc).

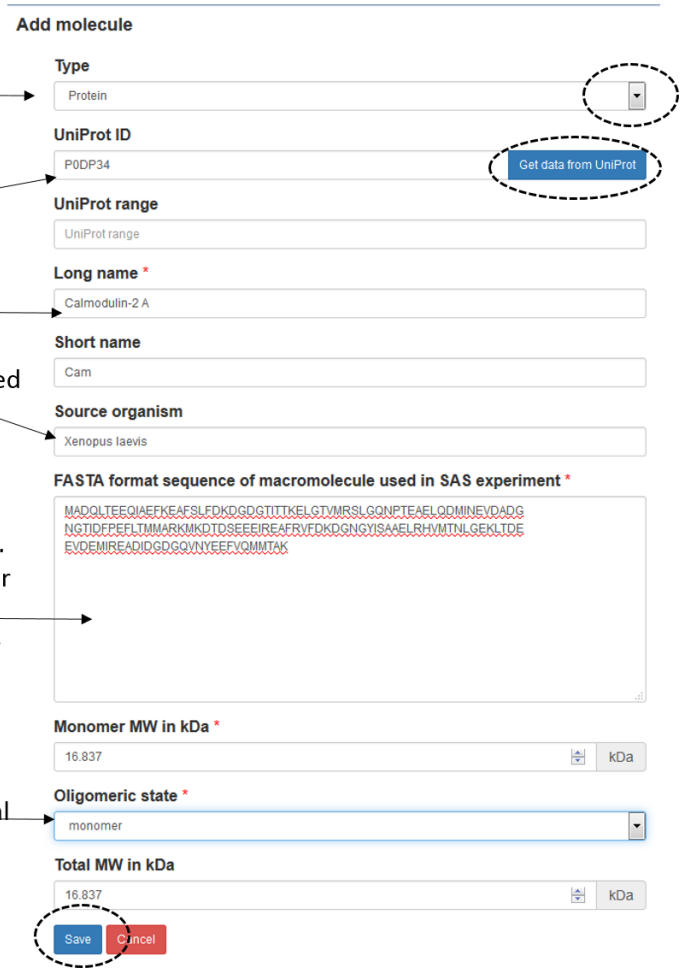
For proteins fill in the Uniprot information, including the amino acid range.

*The long name of the macromolecule is compulsory (auto filled if you use the 'Get from UniProt' option).

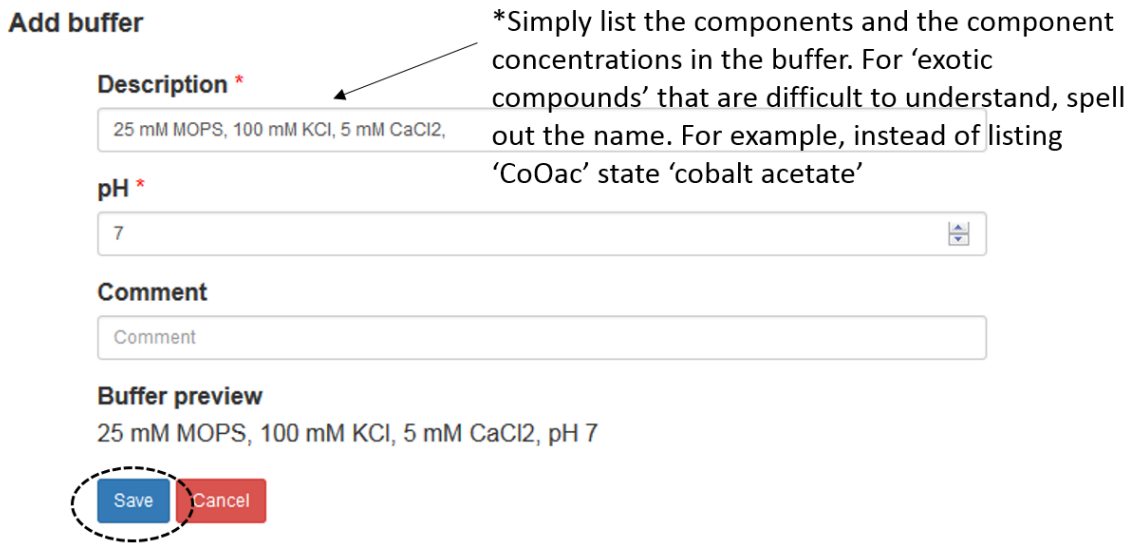
The source organism can be typed or selected from the suggestion menu. If you have a purely synthetic construct, type: synthetic.

*Copy and paste the 1-letter codes of the macromolecule ***used in the SAS experiment***. Typically type in the monomer sequence. For dsDNA remember to include the reverse complement strand. For proteins remember to include any additional amino acid sequences (affinity tags)

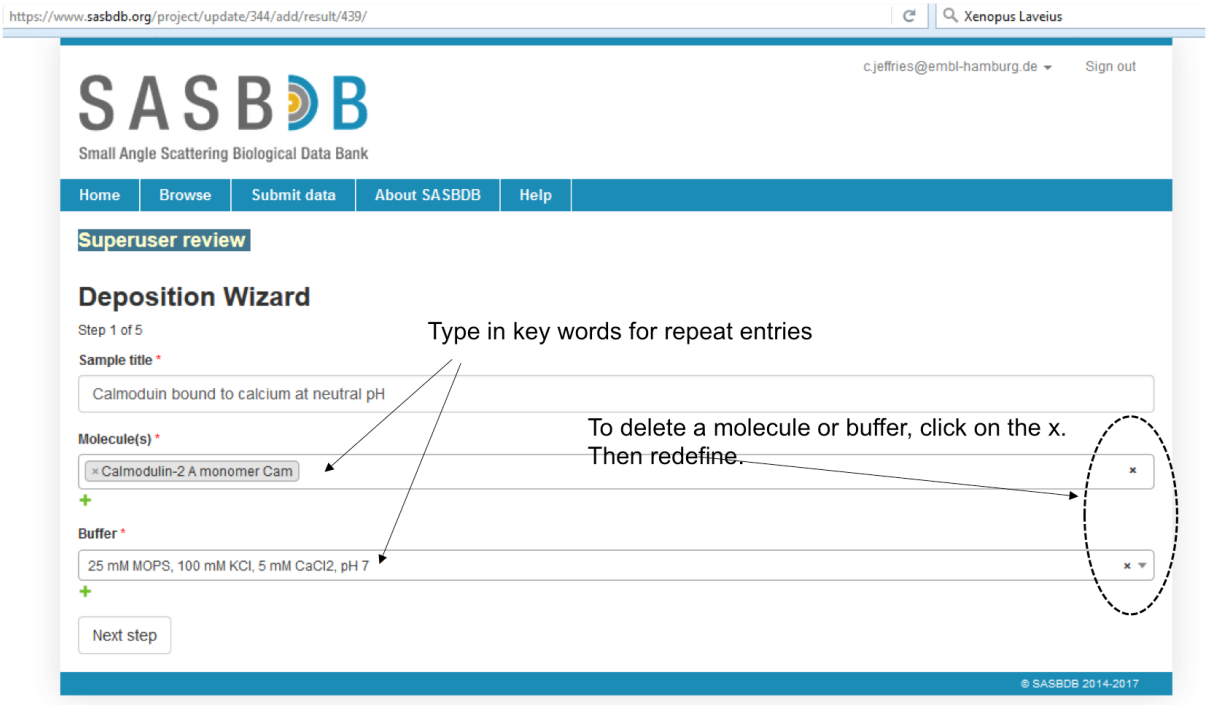
*The monomer MW will be calculated for you. Select the oligomeric state and the total expected MW from the sequence you provided will be calculated.



The **'Add buffer' pop-up window** will appear when you click on + under the buffer.



If a mistake is made…



Tip for multiple SAS entries that have the same macromolecule and buffer. It can be tedious to define the same macromolecule and buffer each-and-every time for repeat entries. Therefore, define the macromolecule and buffer once for the first deposition then for subsequent entries, type in a keyword (or series of keywords) or use the drop-down menu to locate the same macromolecule or buffer.

Step 2. Upload the SAS data and structural parameter information.

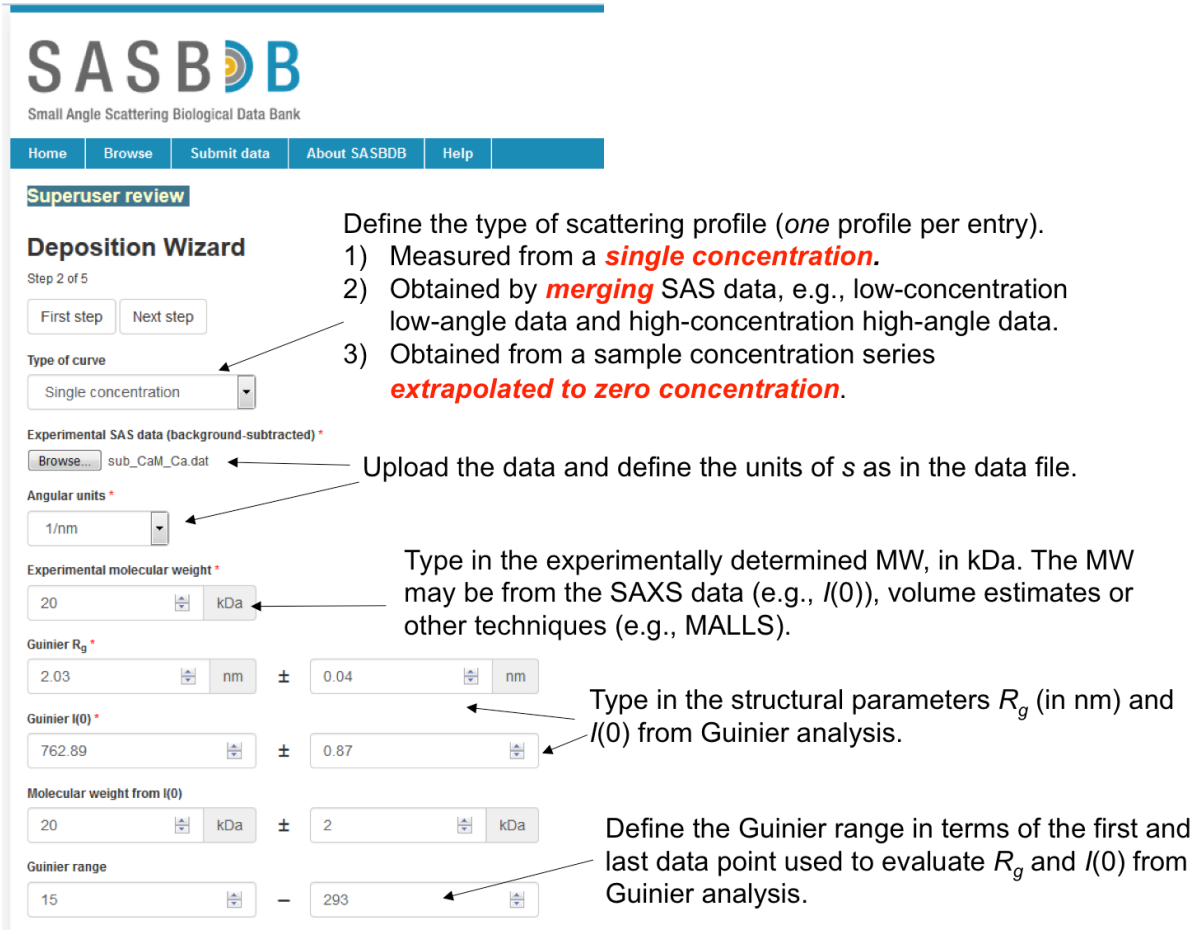
The SAS data are typically recorded in .dat format. In addition, .txt or other column formatted data files can be accepted. However, the columns must be listed the order: momentum transfer (*s*), the intensities (*I*(*s*)) and errors on the intensities, i.e.,:

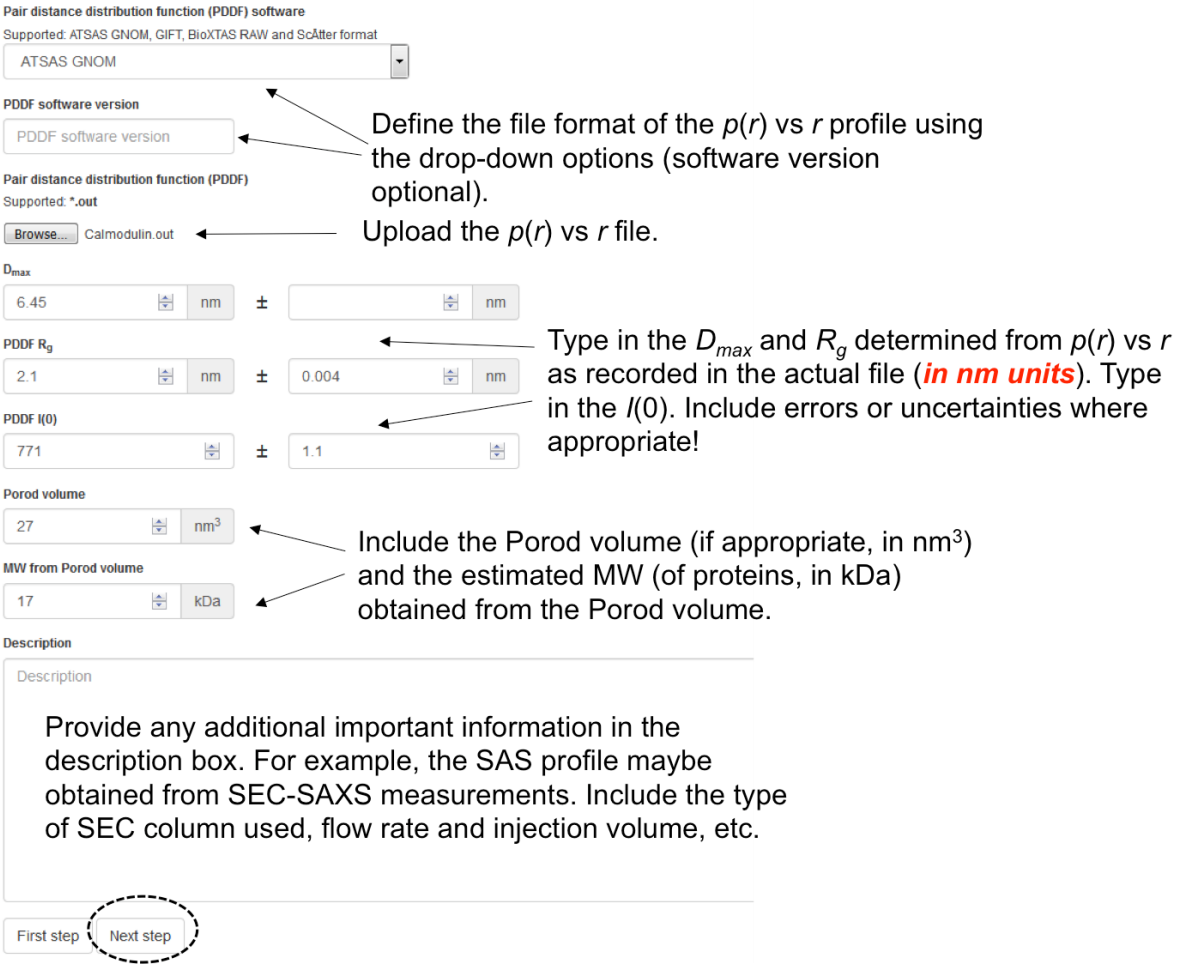
COLUMN 1	COLUMN2	COLUMN3
s	I(s)	σ(I(s))

Data maybe uploaded where *s* is in inverse nm or inverse Å. Choose the correct unit of *s* when uploading your data file. However, *and importantly*, the structural parameters typed into the boxes in Step 2 ***must*** be in nm units and the molecular weight estimate in kDa. The Porod volume is recorded in nm^3.

The *p*(*r*) vs *r* profile may also uploaded in this step. Make sure to select the file format of *p*(*r*) vs *r*, for example GNOM, GIFT, etc. Very importantly, the R_g, *I*(0) and D_{max} from *p*(*r*) vs *r* that are typed into the boxes of the submission form must be identical to the structural parameters actually listed in the *p*(*r*) vs *r* file, *but in nm*. For GNOM.out files upload the *entire* .out file and not just the *p*(*r*) vs *r* distribution at the end of the file.

To ensure speedy processing of your deposition fill in as many boxes as possible in Step 2!





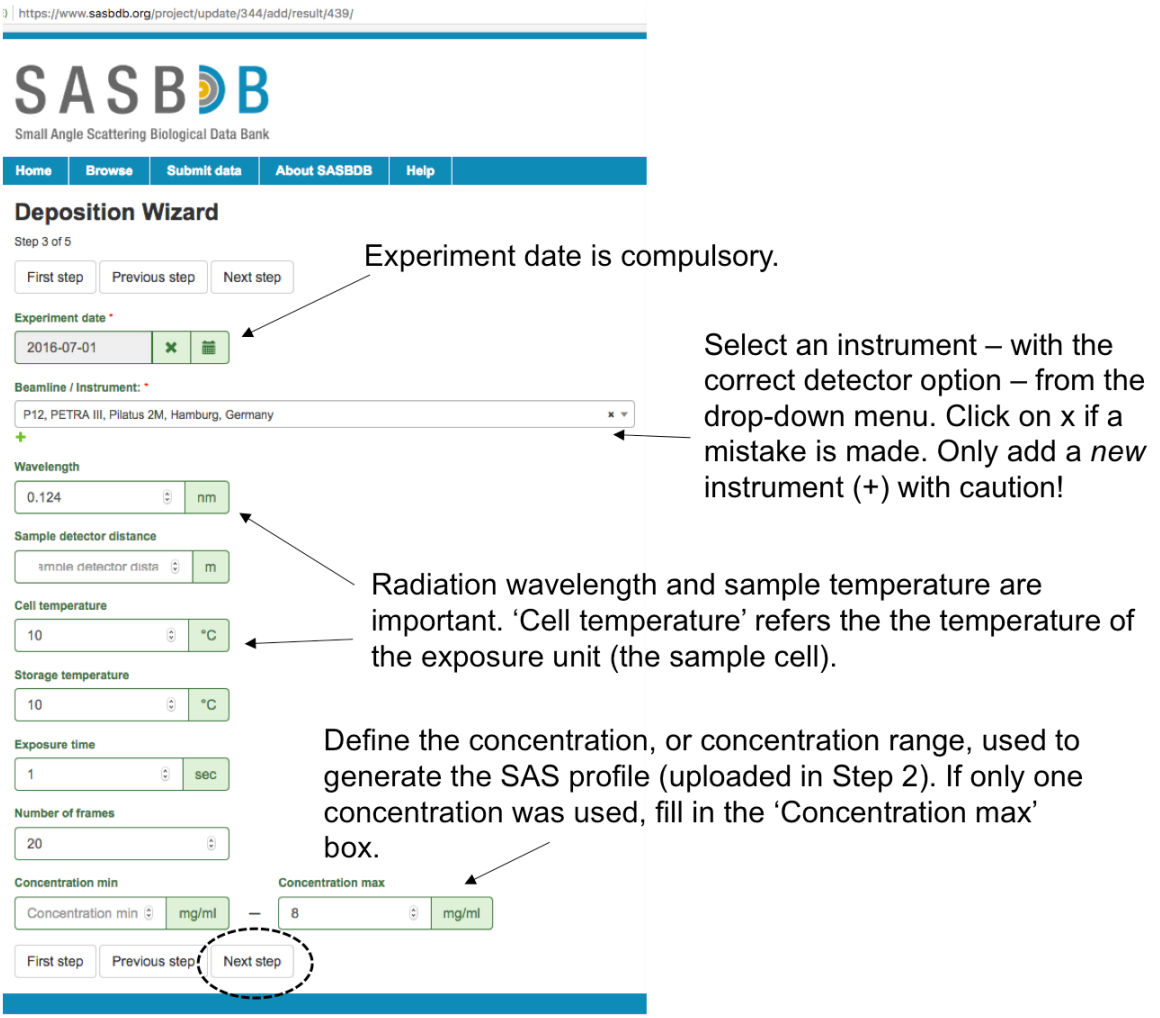
Step 3. Define the SAS experiment – wavelength, exposure time, temperature, sample concentration.

Step 3 is relatively straight-forward, however finding the correct information can be frustrating (some beam lines record the information required for Step 3 in the footer of their .dat files, so check there). It is mandatory to provide the date of the SAS experiment. The beam line or laboratory instrument used to perform the SAS measurements is also mandatory and can be selected from an extensive drop down menu. If you cannot find the beam line or specific instrument, it is possible to add a new instrument clicking on the + button and filling in *every box* in the associated pop-up window. Instrument information is assessed by SASBDB staff before being accepted, so always double check and then re-check the drop-down menu instrument options before considering adding a new instrument. Redundant or duplicated instruments will be deleted and your submission will be delayed. The radiation wavelength (in nm) is also important to include (as is routinely done in crystallography), as is the temperature of the measurement. Of most importance is the sample concentration (in mg/ml):

i) For a single concentration SAS profile, only fill in the 'concentration max' box.

ii) For a merged, or extrapolated to zero concentration, data, type the 'concentration min' and 'concentration max', i.e., the sample concentration range used to generate, or calculate, the final SAS profile (uploaded in Step 2).

iii) For SEC-SAXS record, at minimum, the initial load concentration in the 'concentration max' box. If additional biophysical characterization or experiments were performed and you know the SEC-concentration range used to generate the final SAS profile (e.g., from RI measurements), type the minimum and maximum concentrations into the relevant boxes of the submission form.



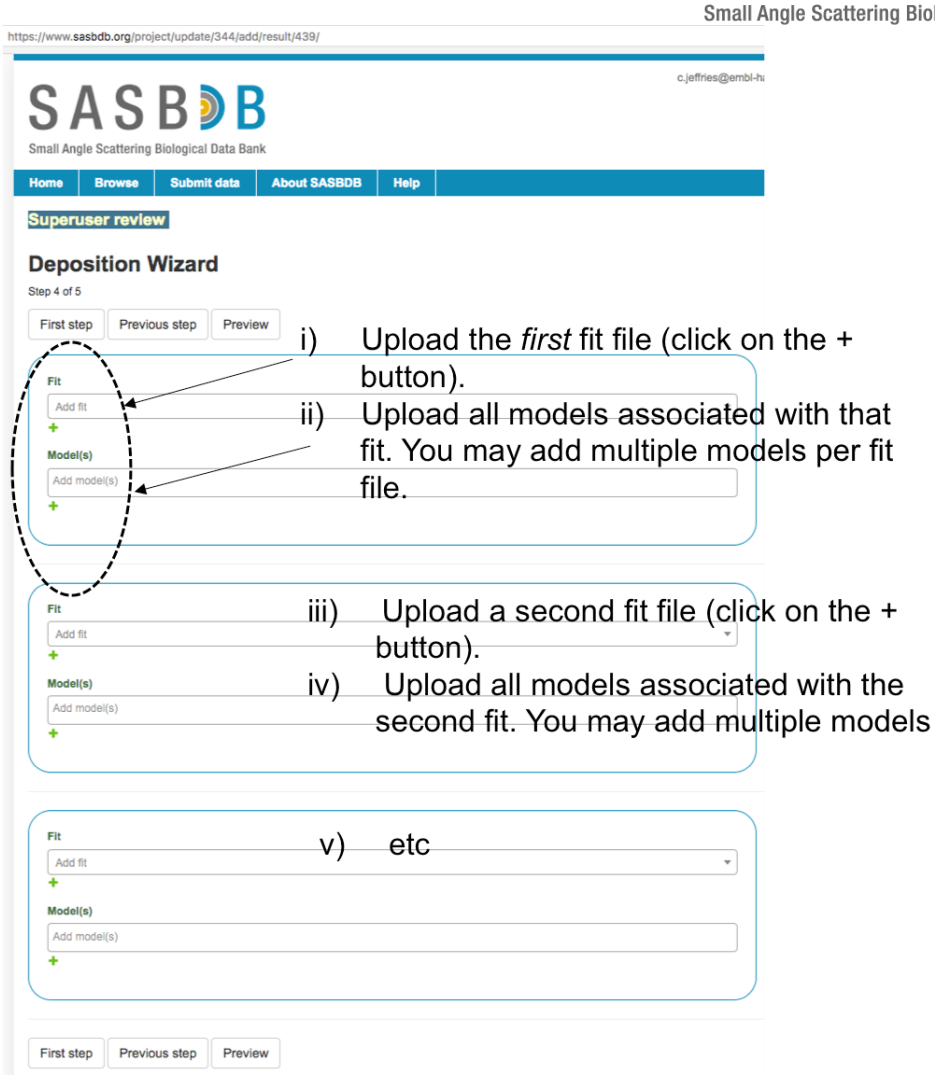
Step 4. Upload associated fit files and models for the SAS experiment.

There are many programs used to model SAS data. SASBDB attempts to accept fit files and models generated using different software packages that can be selected during the upload process. However, and in general, the files recording the model-fit to the SAS data are in simple column format (e.g., .fir, .fit, .dat) while the spatial coordinate files are in .pdb format. We do not accept DAMAVER.pdb models. Described here is the upload and display of three types of model-fits:

1) An ab initio *DAM model generated using the program DAMMIF. Files required = .fir; .pdb*

2) A set of spatial representatives from a structural ensemble calculated using EOM. Files required = .fit; .pdb; R_g distribution (.dat or .txt).

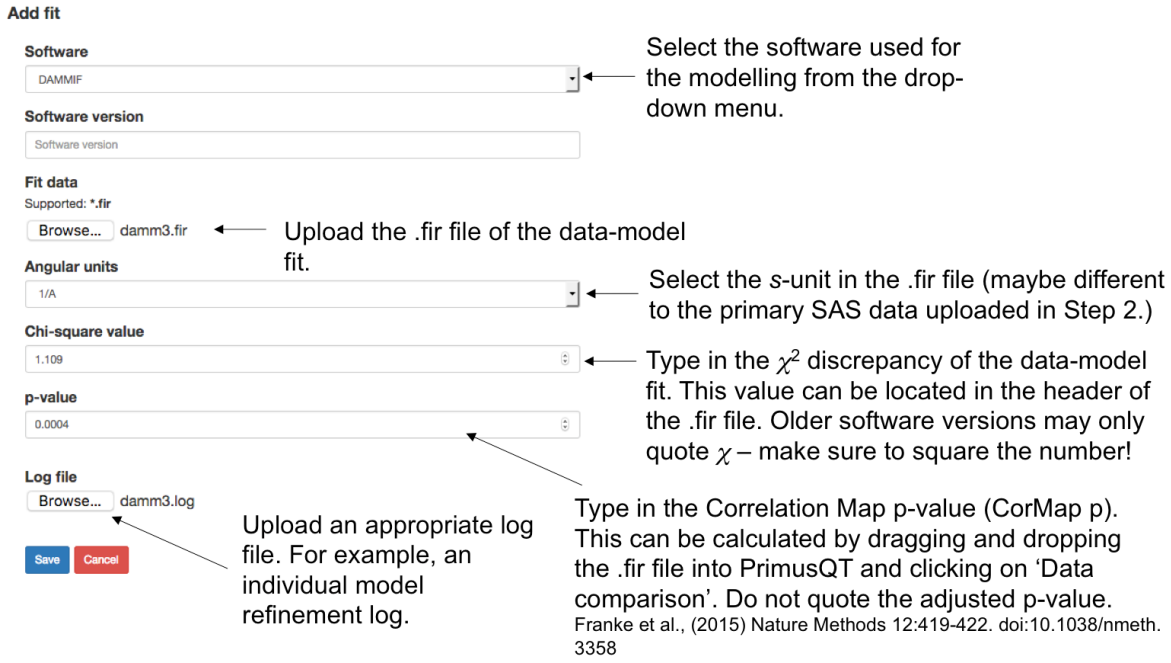
3) A X-ray crystal structure fit to the SAS data. Files required = .fit; .pdb; PDB accession code.



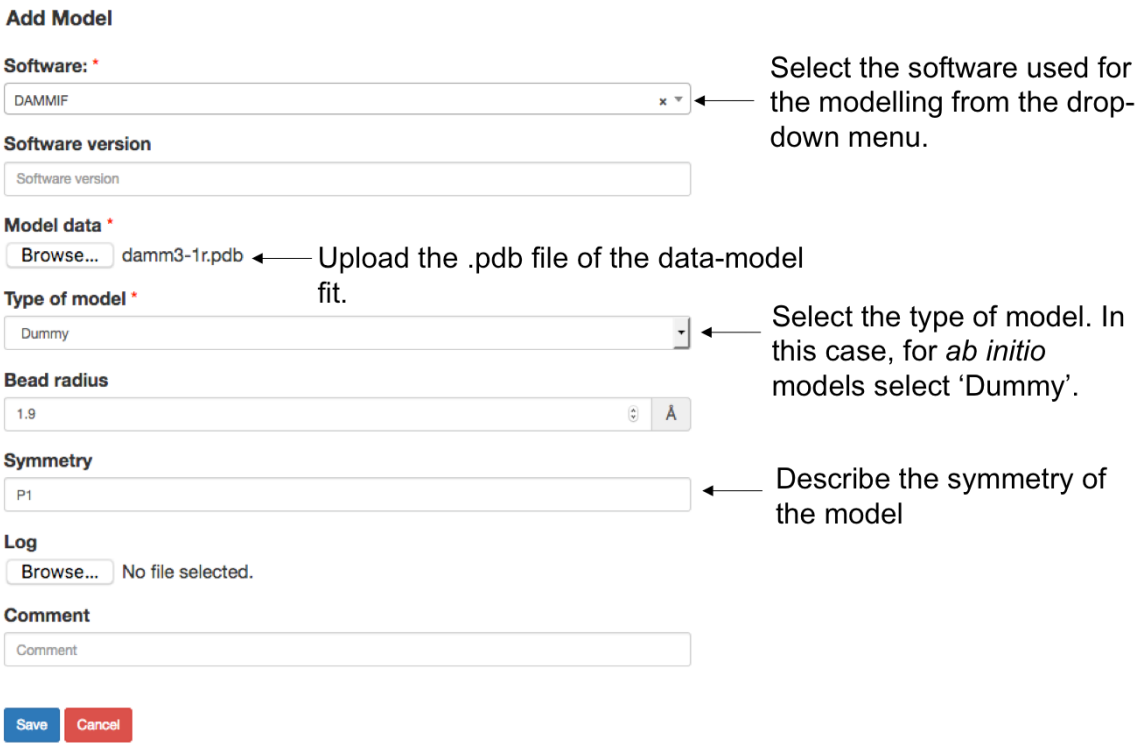
On clicking 'Add fit' a popup menu will appear. Once again it is advised to fill in as much information into the boxes as possible.

Ab initio fits and models

*Pop-up **fit** window for* ab initio *models (dummy atoms/beads; DAMMIN, DAMMIF, GASBOR).*



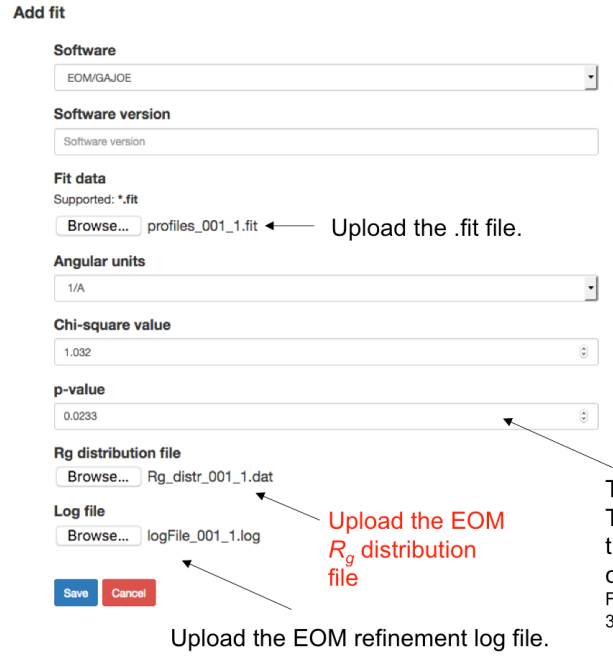
*Pop-up **model** window for* ab initio *models (dummy atoms or beads).*



Note. For DAMFILT.pdb models upload the damsel.log file.

EOM fits and models

*Pop-up **fit** window for EOM models (mix of dummy atoms and atomic high-resolution rigid-bodies).*



Select the software used for the modelling from the drop-down menu.

Upload the .fit file.

Select the *s*-unit in the .fit file (maybe different to the primary SAS data uploaded in Step 2.)

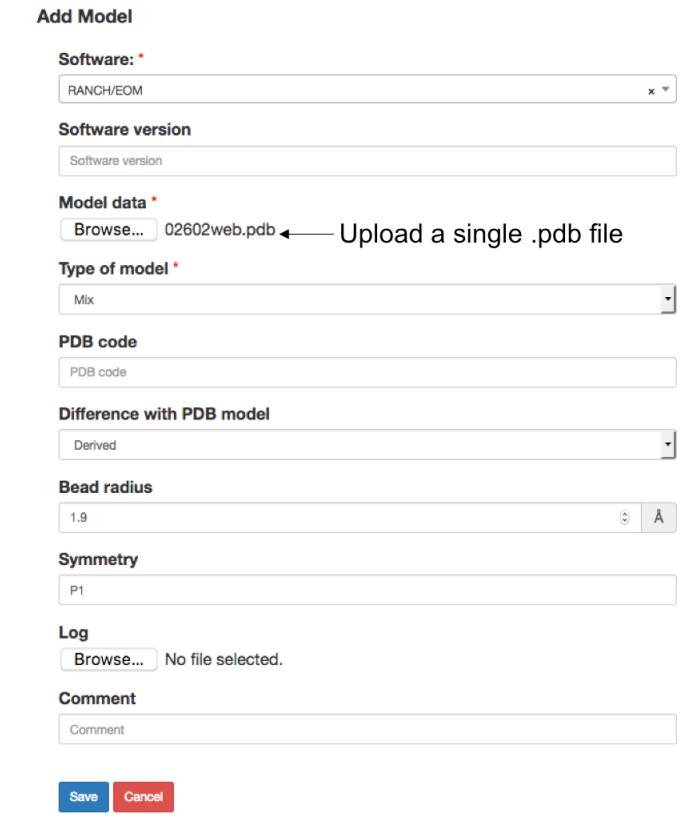
Type in the χ^2 discrepancy of the data-model fit. This value can be located in the header of the .fit file. Older software versions may only quote χ – make sure to square the number!

Type in the Correlation Map p-value (CorMap p). This can be calculated by dragging and dropping the .fit file into PrimusQT and clicking on 'Data comparison'. Do not quote the adjusted p-value.
Franke et al., (2015) Nature Methods 12:419-422. doi:10.1038/nmeth.3358

Upload the EOM R_g distribution file

Upload the EOM refinement log file.

*Pop-up **model** window for EOM models (mix of dummy atoms and atomic high-resolution rigid-bodies). For EOM upload one model at a time!*



Select the software used for the modelling from the drop-down menu.

Upload a single .pdb file

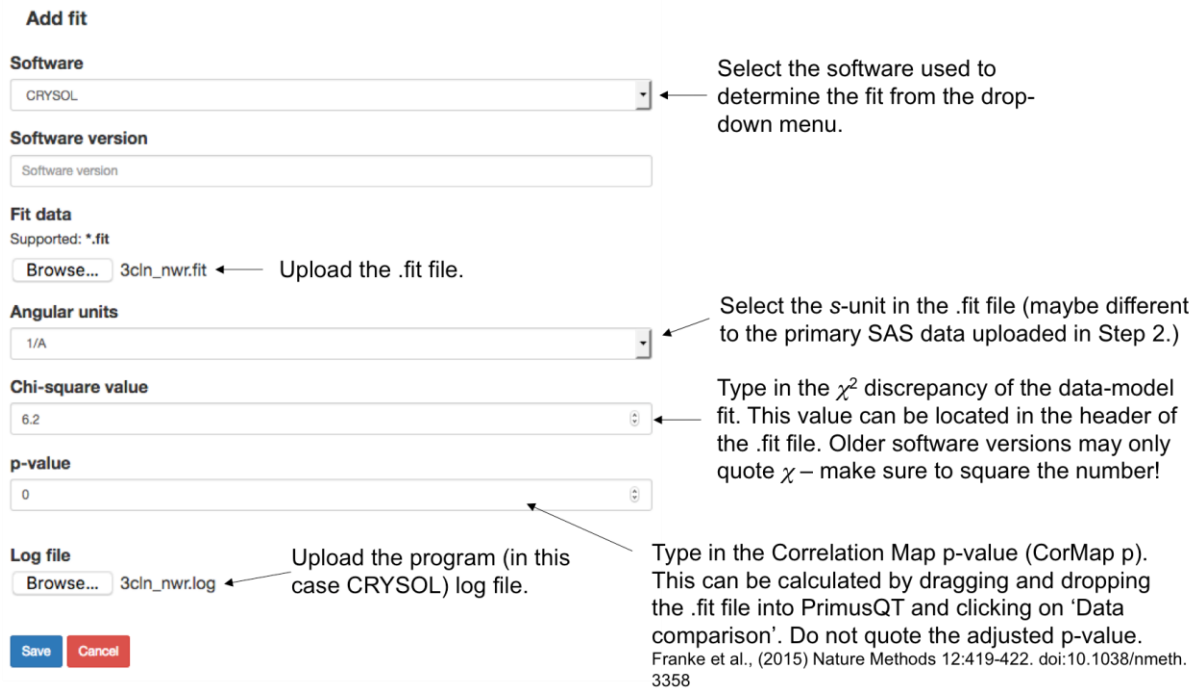
Select the type of model. In this case, the EOM modelling used a *mixture* of beads and rigid-bodies, therefore 'Mix' is selected. Other ATSAS modelling programs that produce 'Mix' models are CORAL and BUNCH.

Describe the symmetry of the model

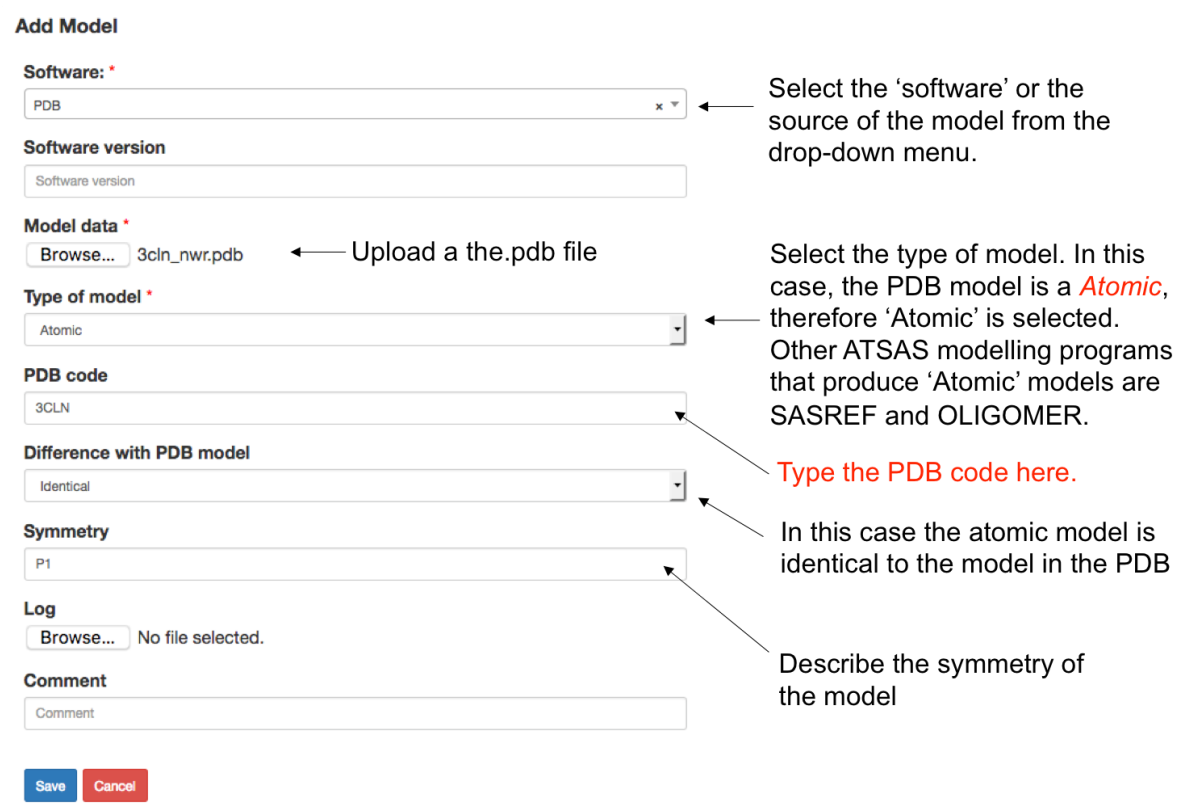
SASBDB
Small Angle Scattering Biological Data Bank

X-ray crystal, NMR, EM homology-model structure fits and models

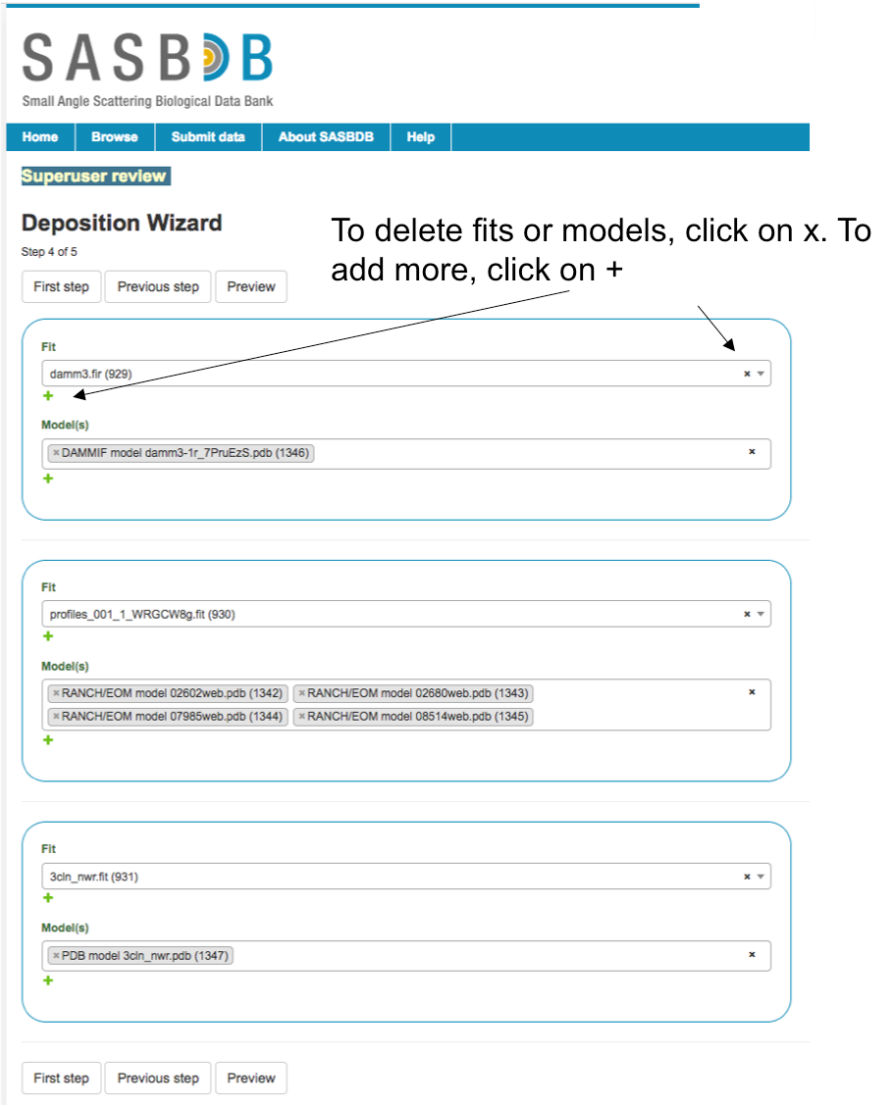
*Pop-up **fit** window for 'PDB models'.*



*Pop-up **model** window for 'PDB models' (from the Protein Data Bank).*



After uploading the fits and models you can remove or add additional models and fits.



Tip for depositing MONSA models. MONSA also generates *ab initio* models, however in this instance it is necessary to deposit the associated .fit files (not .fir) as well as the refined DAM-phase models MONSA produces. MONSA is often used to model multi-component systems with different phases against different SAS profiles in parallel. To deposit a MONSA model of a complex, that consists of three phases (solvent and two individual components), it is necessary to set up *multiple* fits associated to *one* model.

i) Check that the primary SAS dataset for the entry (uploaded in Step2) is the scattering from the entire complex.

ii) Upload *multiple* .fit files produced by MONSA, one phase per 'fit box' in STEP 4.

iii) Upload a single *combined* MONSA phase model (MONSA.pdb) linked to the fit to the complex data.

MONSA outputs a number of .pdb files depending on the number of phases used during refinement. For example: MONSA-0.pdb (solvent); MONSA-1.pdb (phase 1); MONSA-2.pdb (phase 2); MONSA.pdb (phase 1, 2, and 3). Upload the complete MONSA.pdb file (containing all three phases).

iv) Generate additional SASBDB entries for the SAS data obtained from the isolated components used in the MONSA modelling. For example, a protein:DNA complex will have three SASBDB entries. The complex (with the complete MONSA.pdb model and fits); The protein component; The DNA component.

Step 5. Preview your entry before submitting!

After the models and fits have been uploaded to SASBDB, the next step is to preview the entry. The preview page is what the entry will essentially look like on the web page! Therefore check the following:

1) The data and Guinier plot are displayed correctly – check the s-units, etc. If the Guinier plot is not displayed correctly either the R_g, $I(0)$ or s-units typed into STEP2 are incorrect or the SAS data are in a non-standard format. Go back to STEP2 and revise.

2) The dimensionless Kratky plot is generated automatically. If the Kratky plot is not displayed correctly either the R_g, $I(0)$ or s-units typed into STEP2 are incorrect or the SAS data are in a non-standard format. Go back to STEP2 and revise.

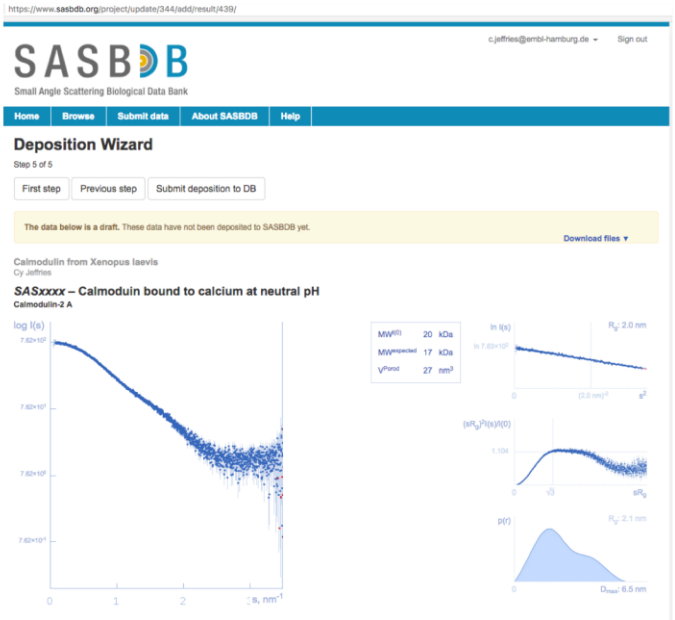
3) The $p(r)$ vs r profile is displayed correctly and the D_{max} and R_g are in the correct units (and that the R_g is similar to that reported in the Guinier plot).

4) Make sure that the MW from $I(0)$, MW expected (from sequence) and Porod volume make sense.

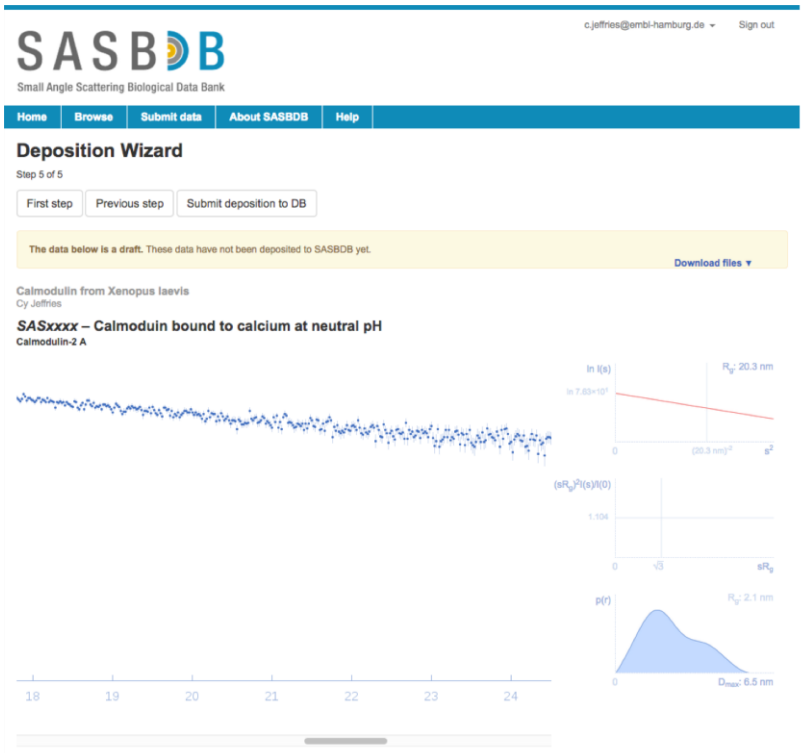
5) Check that the models are associated with the correct fit. Note! For rigid-body or atomistic model pictures. If your model does not display correctly in ribbon format using PyMOL, your model will not display correctly in the picture in SASBDB. Go back and change the chain identifiers in the PDB file to correct this issue and re-upload the model in STEP4.

6) At the bottom of the preview page, review the molecule name, UniProt ID/link, FASTA sequence and links to the PDB.

Check that the data displays correctly and the *s*-units are are correct.



Check that the Guinier, Kratky and $p(r)$ vs r plots plot are displayed correctly and the structural parameters are correct.



The data, Guinier and Kratky plots do not display correctly

✖

Errors have occurred in STEP2 of the submission form. The source data and structural parameters are in nm, however, the R_g was typed in as Å, the *s*-unit of the data was selected as $Å^{-1}$ and the $I(0)$ was incorrectly typed into the box (out by a factor of 10).

When you are happy with your entry…

Click Submit deposition to DB and wait for a member of the SASBDB team to contact you. This may take 24-48 hours (for correct entries) or up to one week for those depositions that require significant revision.