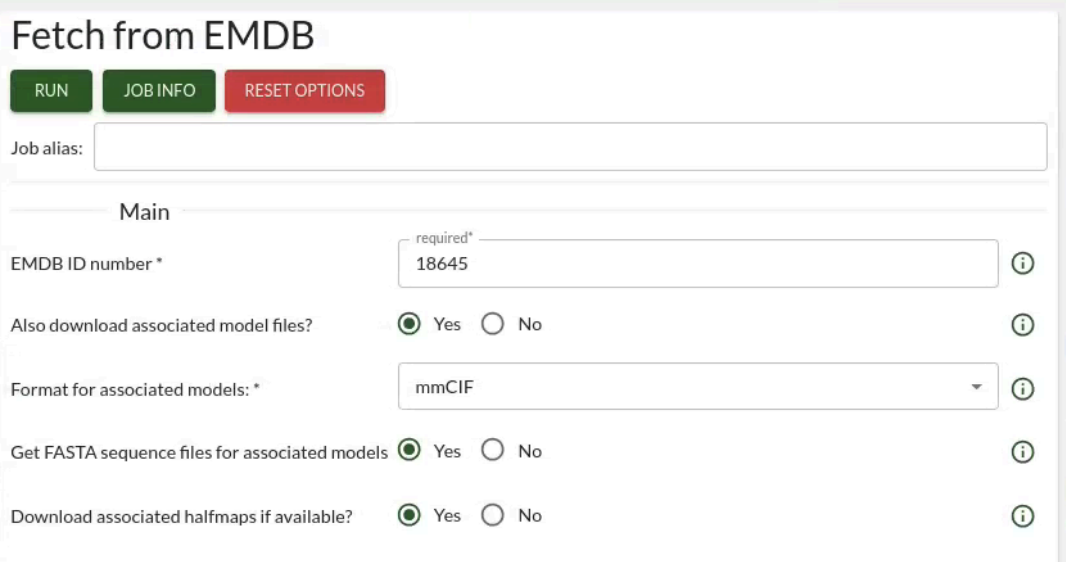
ModelAngelo tutorial

Building a map with FASTA sequence

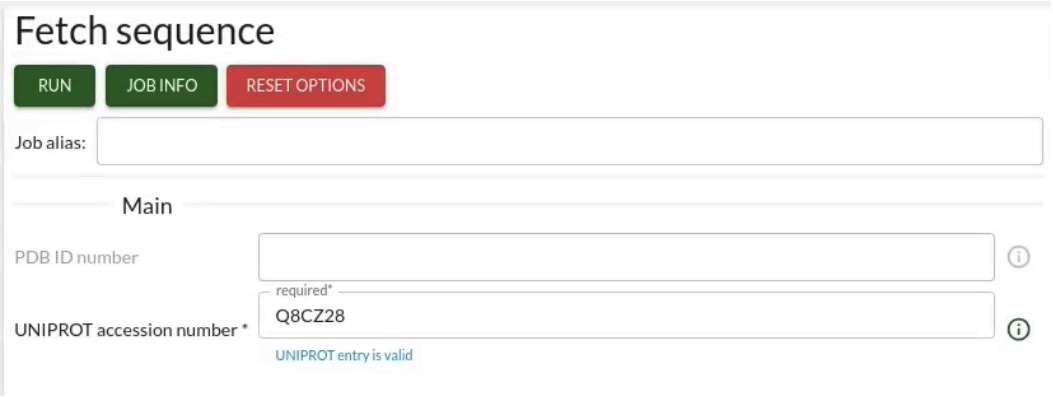
This is the recommended use case - when you have access to a medium-high resolution cryo-EM map (resolutions exceeding 4 Å) as well as FASTA files with all of your protein, RNA, and DNA sequences.

If you would like to build nucleotides as well, you need to provide the RNA and DNA portions of your sequences in different files

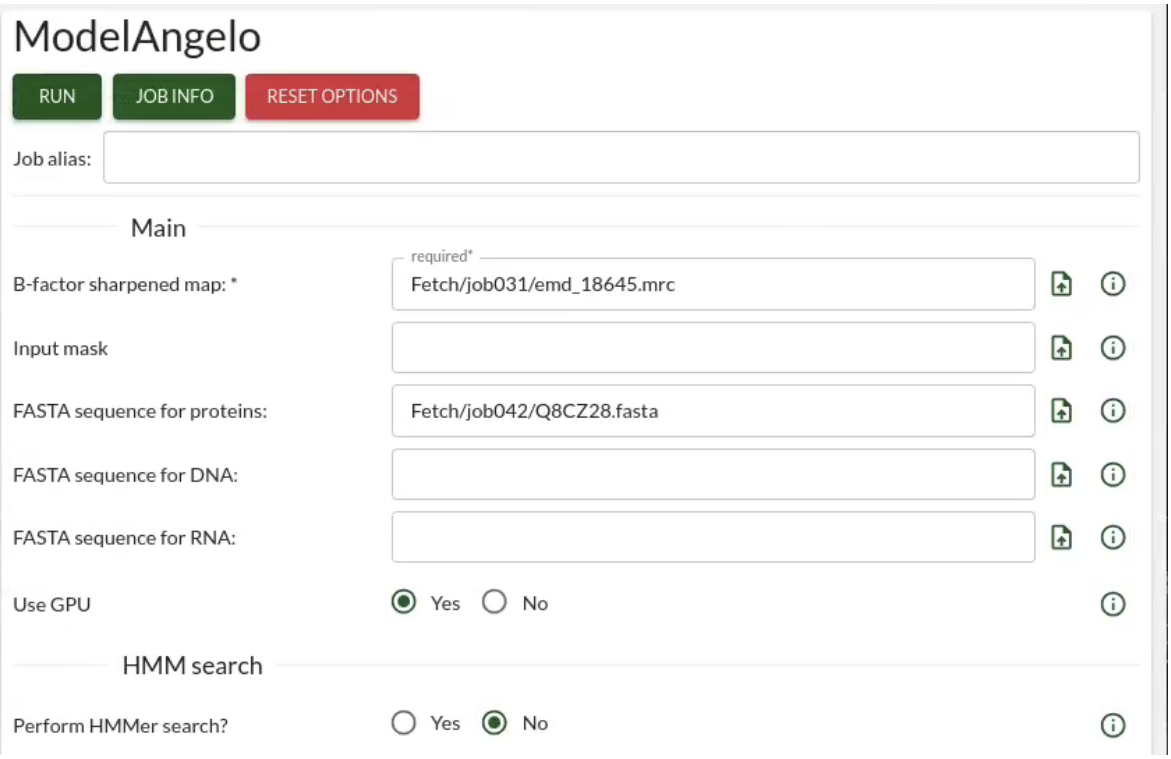
1) Fetch the 2.2Å map of Streptococcus pneumoniae NADPH oxidase from EMDB (ID number 18645). Download the associated (deposited) model and sequence to compare against the ModelAngelo results.



2) Fetch the sequence of the protein from the UNIPROT database (Q8CZ28 https://www.uniprot.org/uniprotkb/Q8CZ28/entry#sequences)

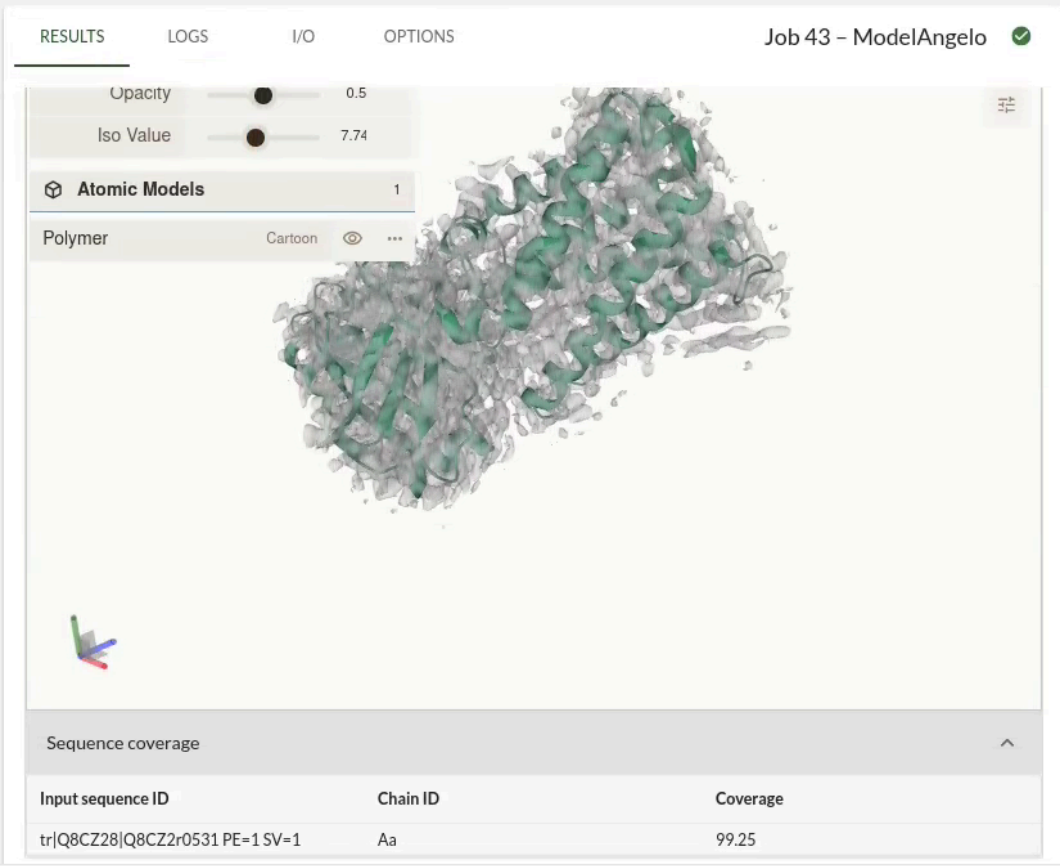


3) Run ModelAngelo with the downloaded map and protein sequence as inputs



Use GPU and the Icknield_gpu submission script in the Queue submission options

The Results shows the model built in the map and the sequence coverage with respect to the input protein sequence (99.25%)



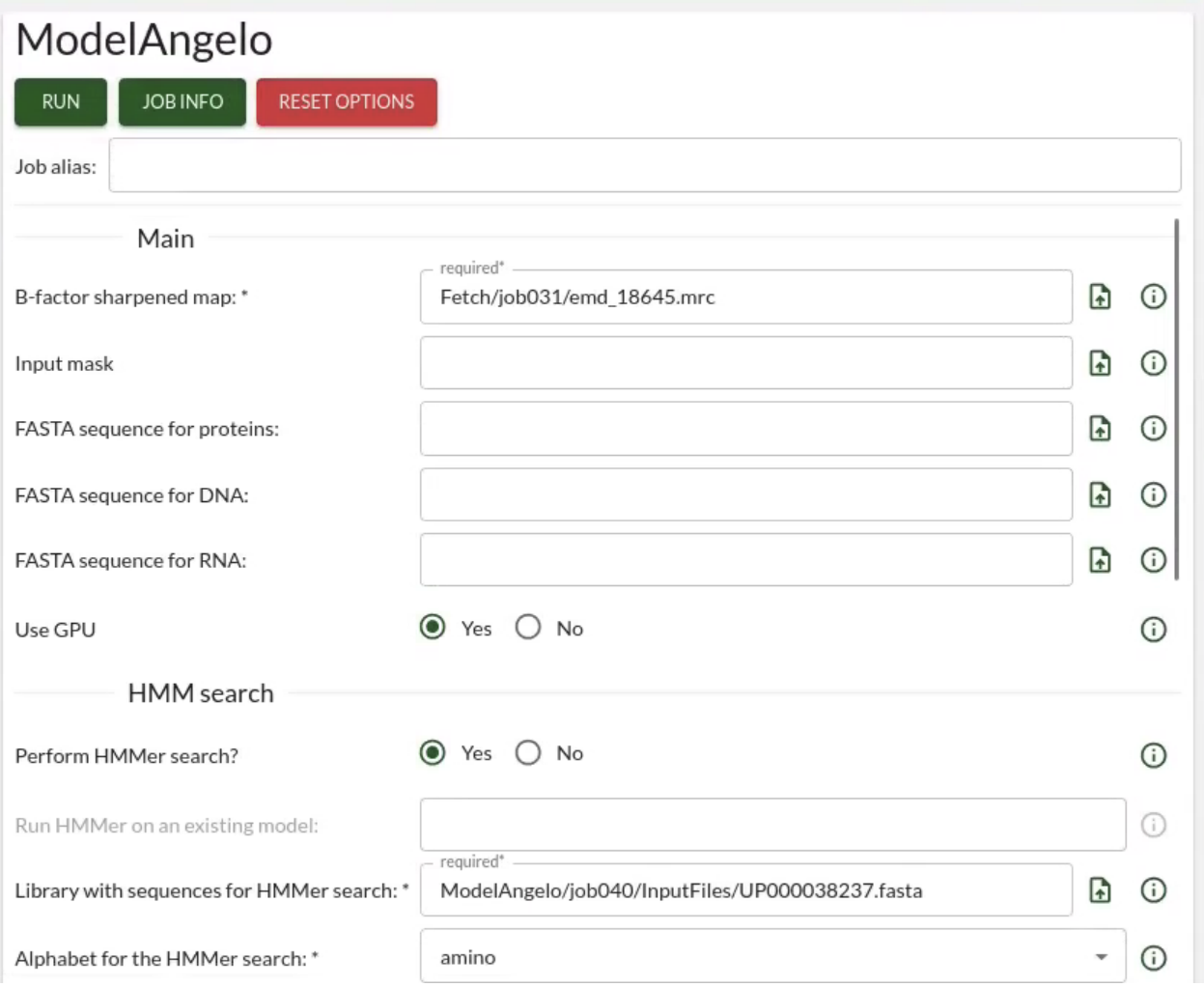
To view the confidence scores associated with the model, set an appropriate *Iso value* for the map and select *Atom property; Uncertainty/Disorder* as the Color Theme. (Click on the *Actions* button to the right of the *Polymer* text in the ***Atomic Models*** panel, then click on the *Color theme* selection (which is set to *Chain Id* by default) and then expand the *Atom property* drop-down list and select *Uncertainty/Disorder*).

Building a map with no FASTA sequence

If the sequence is not known, ModelAngelo can build a model based on the best residue predictions at each C-alpha position identified. In addition to a built model, it provides you with HMM (hidden Markov model) profile files that you can use to search a database.

We can use the same example above to build a model without using any sequence information as input. We can search for known sequences from a database that fits the sequence profile identified by ModelAngelo. Here we search against Streptococcus pneumoniae proteome (https://www.uniprot.org/proteomes/UP000038237) . The FASTA file is already downloaded in the tutorial data folder.

Run the job and compare the results against the previous run with sequence input.



If you would like to compare the predicted atomic models from each run of ModelAngelo you can:

- Go to one of your finished ModelAngelo jobs. Go to the ***I/O*** tab and select the *Fetch/jobXXX/emd_18645.mrc* and the *ModelAngelo/jobXXX/jobXXX.cif* files and open them with the ***Moorhen (beta)*** button.
- Load the other predicted atomic model (from the other ModelAngelo job you ran). You can do this by opening the Moorhen menu (top left), clicking on the ***File*** button and then click on the ***Browse*** button underneath *Coordinates* (at the top). Locate the other *ModelAngelo/jobXXX/jobXXX.cif* file and open it.
- You can open the ***Maps*** window from the Moorhen menu and use the “Eye” button to toggle visualisation of the density map on/off. You can also adjust the contour level of the map and the radius shown around the point you are centered on

- You can open the ***Models*** window from the Moorhen menu and similarly use the "Eye" button to turn on/off the visualisations of each atomic model.
    - You might find that without an exact sequence that ModelAngelo has been able to successfully build some parts of the model but not others. There may be some disconnected regions of the density map that ModelAngelo has built a model into. Each disconnected region built into will be labelled as a different chain.
    - If this is the case, have a look at some of the regions where ModelAngelo struggles. You can most easily find these by going to the ***Models*** window and clicking on the *Sequences* tab to expand it. If you have multiple chains, look at the ends of them and compare the atomic models. These are regions where ModelAngelo may have struggled without the provided FASTA sequence.

Advanced options for HMM search:

Multiple filters are used to control the quality of alignment between the residue profile and the database sequence.

The F1, F2, F3 and E-value control the statistical significance for alignment between the residue profile and any sequence from the database. Increasing these will find more sequence hits but also bring in false positives. F1, F2 and F3 are increasingly stringent (lower P-value thresholds).