The Damietta Protein Design Toolkit

User Manual
v1.0.2

2024

Table of contents

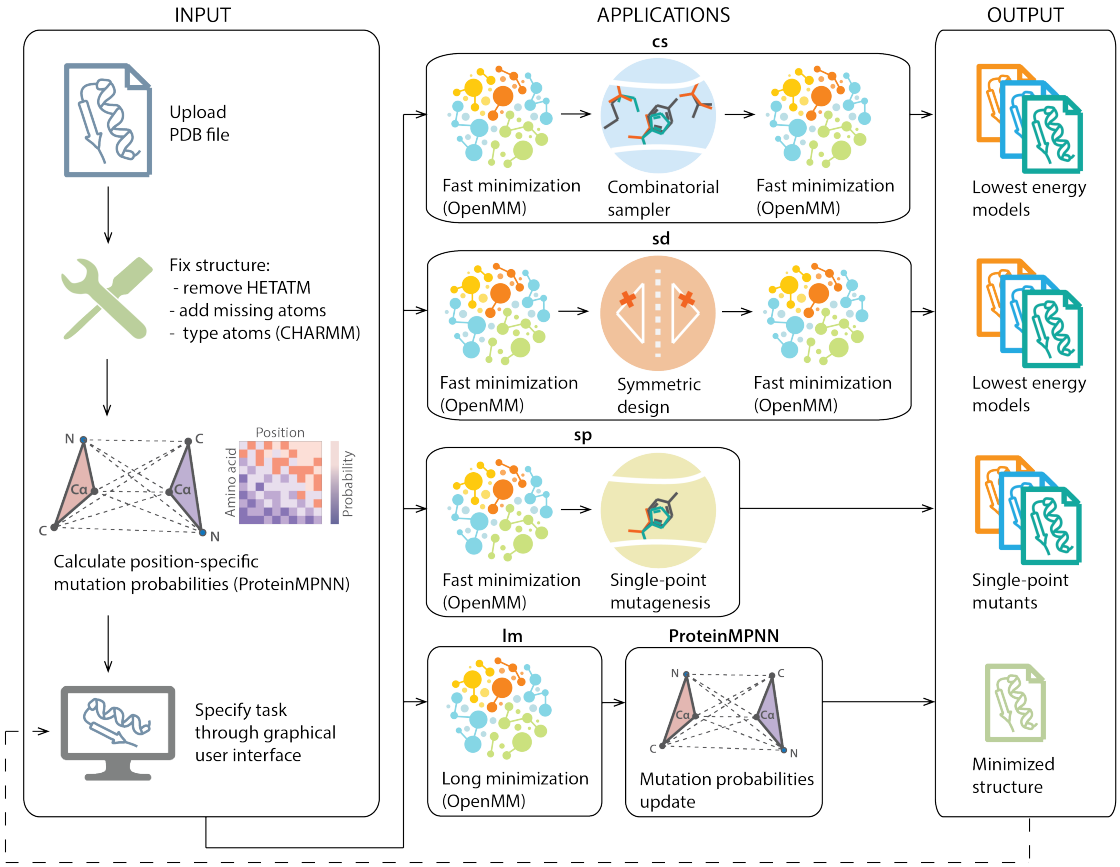
1 Availability

The Damietta toolkit is available free-of-charge for all users at the following URL: *https://damietta.de*

2 Overview

The Damietta toolkit provides a multi-tool framework, allowing the user to conduct several design and modeling processes on their protein structure. The entire framework of the toolkit is organized such that a protein structure is the center of each operation, and it is the main data flowing between different tools. This allows users to forward their structural models across different applications seamlessly.

In the current version, the toolkit provides four distinct applications, three of which are native Damietta tools. These are the *combinatorial sampler* **`(cs)`**, *symmetric design* **`(sd)`**, and *single-point mutagenesis* **`(sp)`** applications. In addition, the toolkit offers a molecular dynamics-based minimization tool **`(lm)`**, that is based on the OpenMM library.

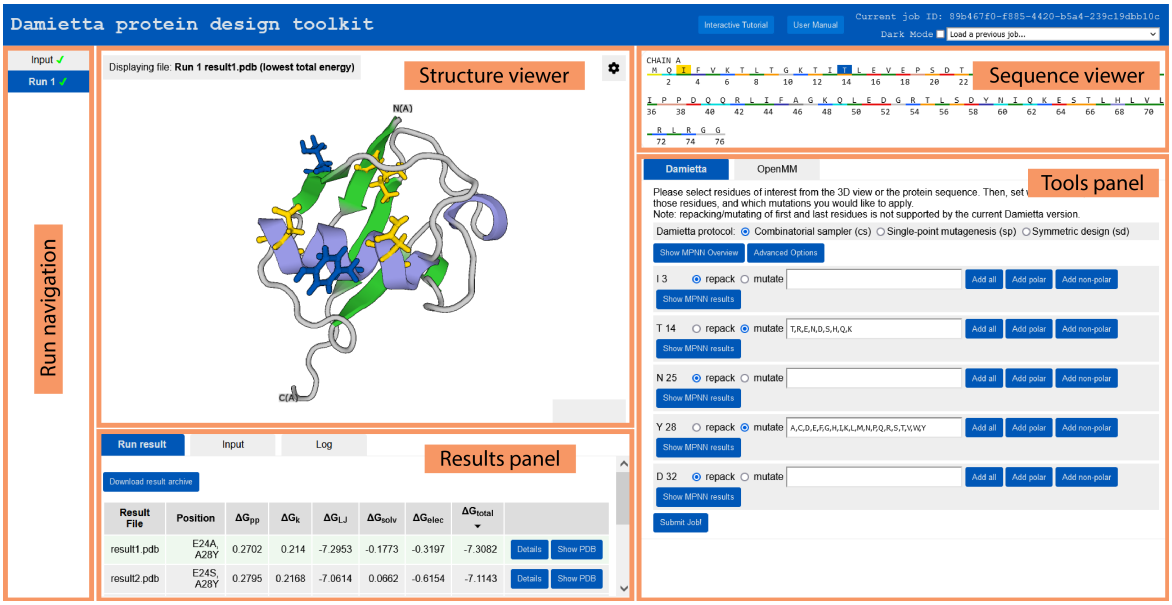


The user first uploads a PDB file, which is pre-processed and the mutation probabilities at each position are calculated using ProteinMPNN. The protein structure and its

sequence are made visually accessible through a molecular viewer. The user at this stage can choose the tool to run, and specify the run parameters. The output is one or more 3D models, with their associated energy values (i.e. average energy per residue; in kcal/mol). The selected output model can be forwarded to any other tool for further operations.

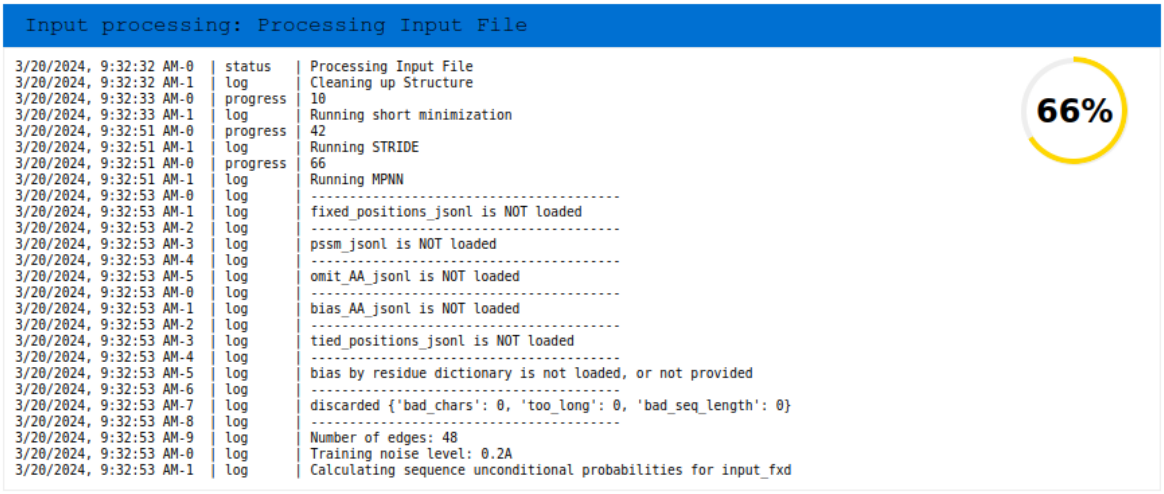
3 Quickstart guide

The graphical user interface of the toolkit includes 5 panels. These are run navigation, structure viewer, sequence viewer, tools panel, and results panel. Each of the panel is accessible depending on the current step.

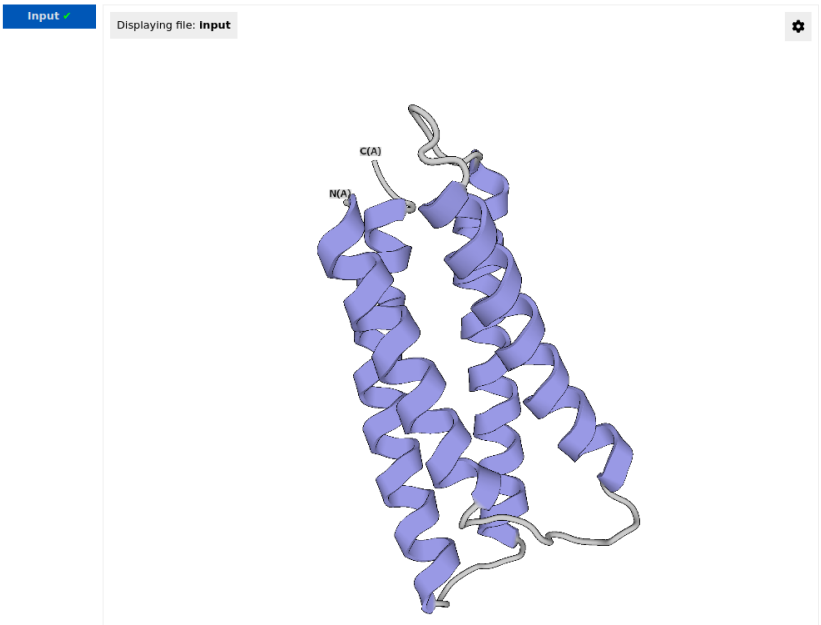


Structure pre-processing. Any workflow starts by uploading a protein structure file in PDB or mmCIF format (*https://www.wwpdb.org/documentation/file-format*). The upload form is available at the start page:

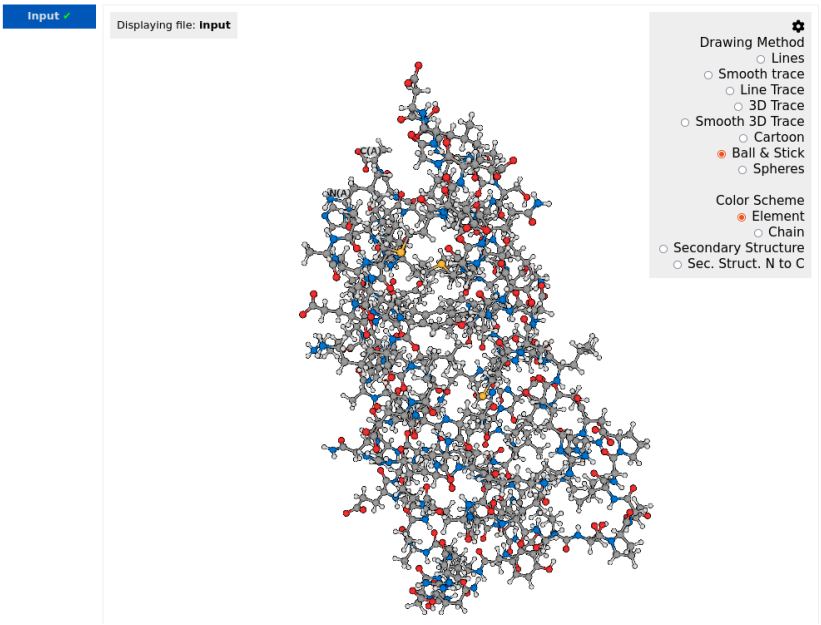
The server then takes a few seconds to process the structure file by first removing all heteroatoms, such as solvent molecules, ions, and non-proteinogenic ligands or non-standard amino acids. It then adds all hydrogens and missing atoms in standard amino acid residues. Afterwards, the atom types of all atoms in the structure are standardized according to the CHARMM36 force field. Finally, all residues are renumbered sequentially, regardless of the input numbering. In case of multi-model PDB, only the first model is processed.



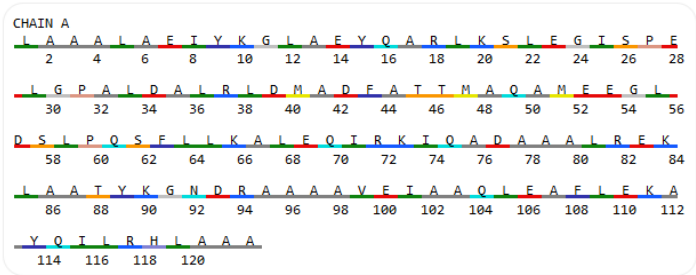
Upon successful pre-processing, the “Input” job appears in the run navigation panel, and the resulting structure is visualized in the structure viewer:



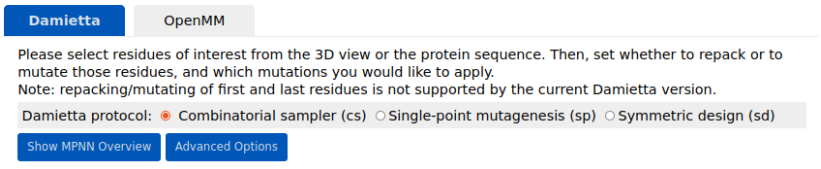
By clicking the settings button , user can select the drawing method and the color scheme for structure representation. In the example below, the drawing method is set to "Ball & Stick" and the color scheme is set to "Element":



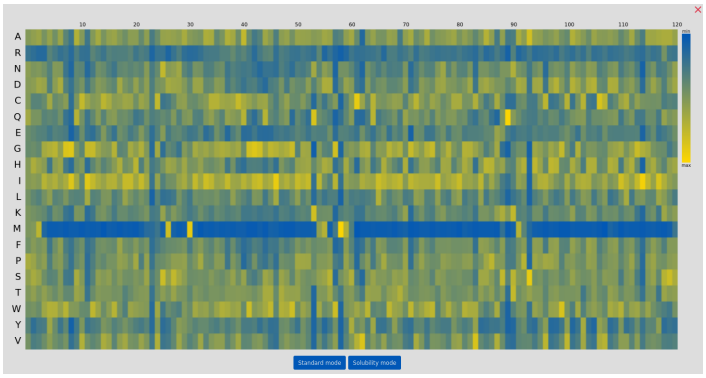
The sequence of the uploaded protein is shown in the sequence viewer:



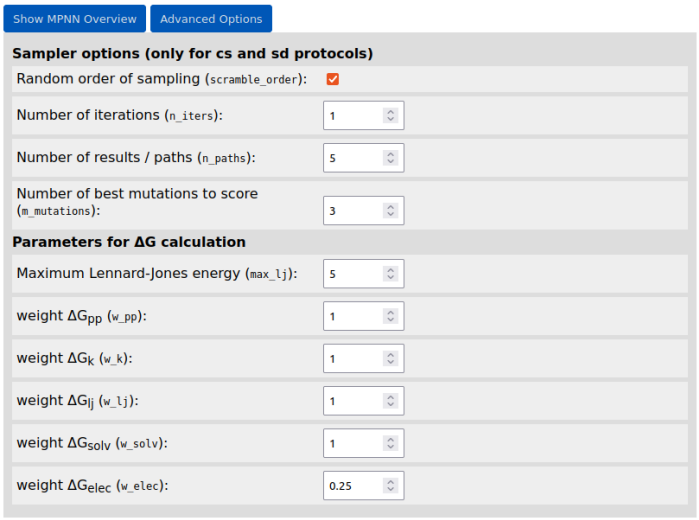
Using the tools panel, the user can choose the tool to run and specify the run parameters. The design and molecular dynamics routines are available under the “Damietta” or “OpenMM” tab, respectively:



As a part of an input file processing, the mutation probabilities at each position are calculated using ProteinMPNN. User can visualize the predicted log probabilities by clicking the "Show MPNN Overview" button:



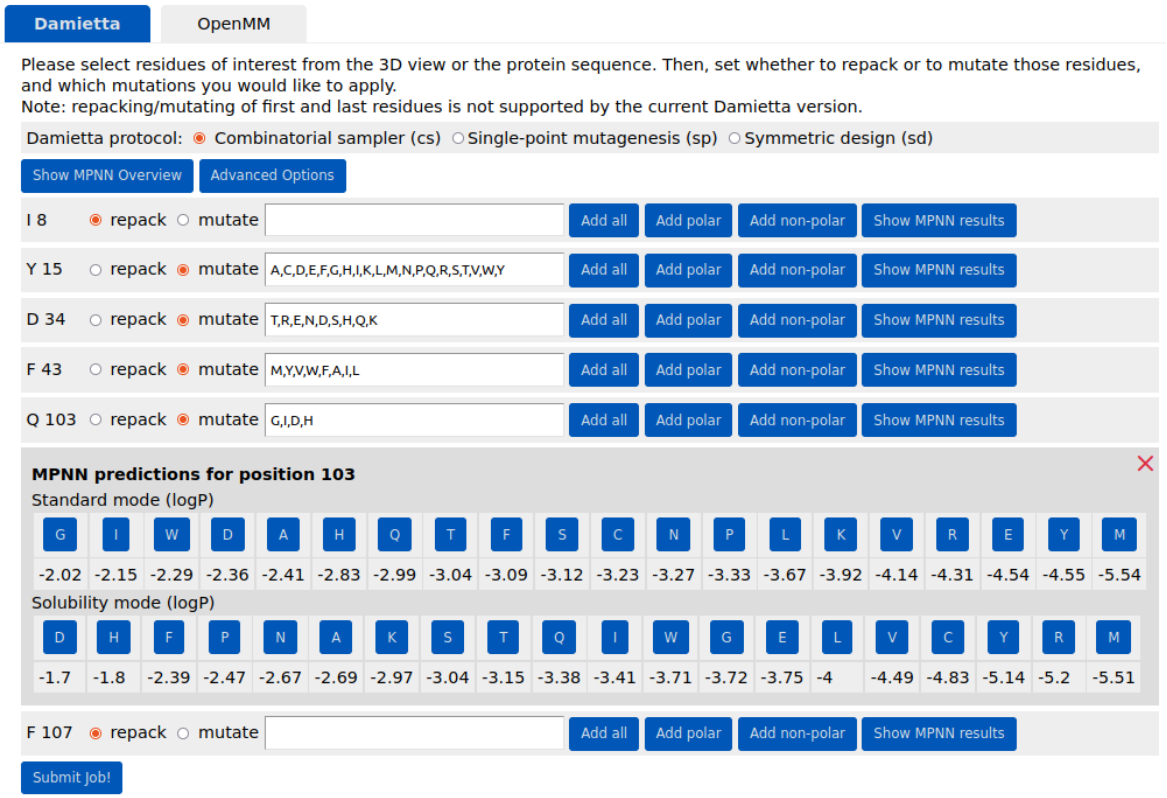
In case of using the Damietta design routines, user can change default sampler options and parameters for energy calculations by clicking the "Advanced Options" button:



Combinatorial sampler (`cs`). The combinatorial sampler is a powerful tool to introduce a large number of mutually-compatible mutations or side chain conformations simultaneously. The actual sampling is performed within a fixed backbone context, but is preceded and followed by a short spans of structure minimization to relax the backbone and side chain atoms.

By setting the radio button to **`cs`** tool and choosing the amino acid positions of interest, a new list of selected positions appears at the bottom right. The amino acid positions can be selected by clicking on the desired positions in either the structure or the sequence viewer. The user can then define the target positions for repacking or mutagenesis. Repacking a residue changes its conformation to minimize its energy, typically evaluating 100 unique conformers per a given backbone conformation. Conversely, mutating a residue evaluates all 100 conformations of the starting residue as well as of all the specified mutants, and identifies the lowest-energy mutants. The entire routine combinatorially evaluates and minimizes the energy across all of the specified residues (via a tree-swarm algorithm), and reports finally up to 5 unique-sequence designs. Reforwarding the output of the `cs` tool to itself for further design iterations is

generally recommended until the average energy per residue does not improve markedly. In the example below, two residues are repacked (I8 and F107), and four positions are mutated (15, 34, 43, and 103). The choice of mutations to sample from is critical, and fewer choices would greatly shorten the calculations. Using the associated buttons, the user can add all (e.g. position 15), polar (e.g. position 34), or non-polar (e.g. position 43) amino acids. Alternatively, the user could make guided guesses using the suggestions of ProteinMPNN, by showing the predicted log probabilities and clicking on the highest likely mutations. In the example below, the two most likely mutations in the standard mode, and the two most likely mutations in the solubility-enhancing predictions mode are selected at position 103.



The job log, input and output could be viewed in the respective tabs. The resulting design models are sorted by the average energy per residue (ΔG_{total}). The broken-down energy terms are also shown. These are the backbone conformation (ΔG_{pp}), side chain conformation (ΔG_{k}), Lennard-Jones interactions (ΔG_{LJ}), solvation (ΔG_{solv}), and electrostatic interactions (ΔG_{elec}) energies. Expanding the details button shows the per-residue energy values. The user can also download the results archive by clicking the respective button. The archive contains a job input in JSON format, a job log, the resulting design models, their FASTA sequences, and a summary table in CSV format. The FASTA file includes 1) the initial input PDB sequence as "wild type" sequence, 2) the last run's "input" sequence, and 3) the sequences of the resulting designs from the last run.

Run result | Input | Log

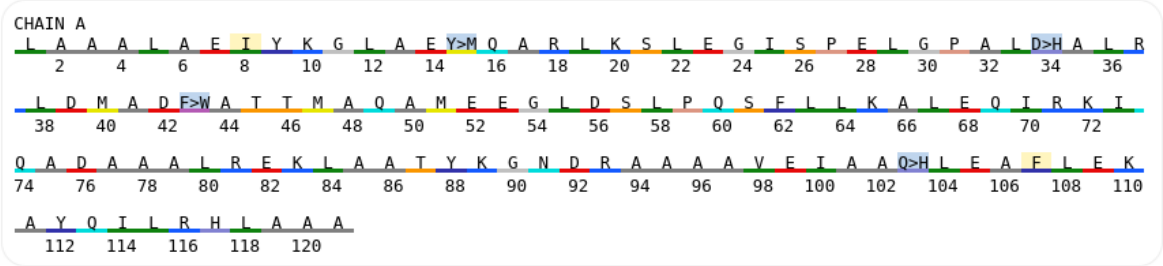
Download result archive

Result File	Position	ΔG_{pp}	ΔG_{k}	ΔG_{LJ}	ΔG_{solv}	ΔG_{elec}	ΔG_{total}	
result0.pdb	Y15M, D34H, F43W, Q103H	0.226	0.54	-13.9016	-0.2322	-0.5909	-13.9587	Details Show PDB
result2.pdb	Y15M, D34H, F43W, Q103I	0.1799	0.5846	-13.5893	-0.4564	-0.4911	-13.7723	Details Show PDB
result3.pdb	Y15M, D34H, Q103H	0.2386	0.5596	-13.5833	-0.3727	-0.544	-13.7017	Details Show PDB
result1.pdb	Y15M, D34H, F43Y, Q103H	0.2457	0.531	-13.657	0.0115	-0.7576	-13.6264	Details Show PDB
result4.pdb	Y15M, D34H, F43W	0.1794	0.5707	-13.4415	-0.2636	-0.6452	-13.6003	Details Show PDB

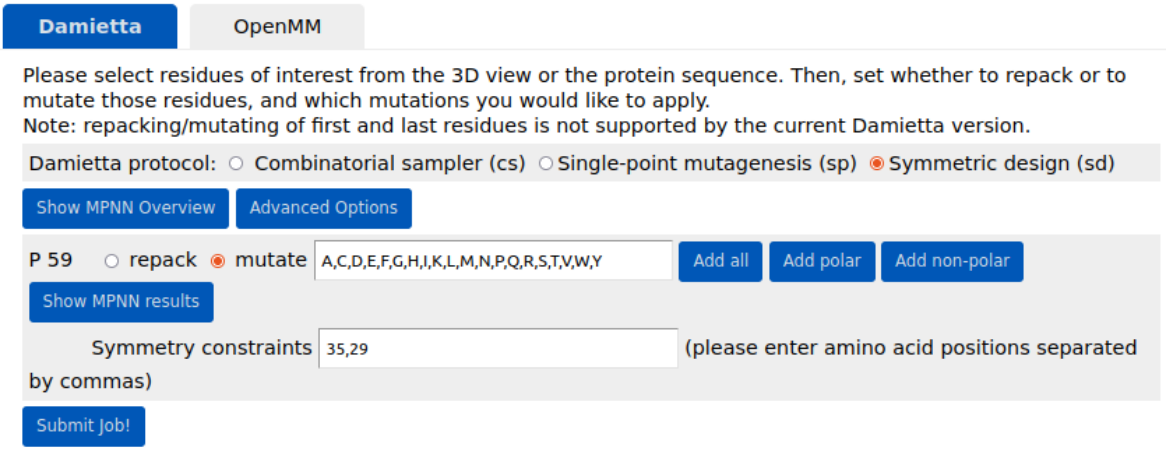
all energies in (kcal/mol)

Result File	Position	ΔG_{pp}	ΔG_{k}	ΔG_{LJ}	ΔG_{solv}	ΔG_{elec}	ΔG_{total}	
result0.pdb	Y15M, D34H, F43W, Q103H	0.226	0.54	-13.9016	-0.2322	-0.5909	-13.9587	Details Show PDB
	I8	0.1496	0.7216	-11.1797	-2.0115	0.2427	-12.0772	
	F107	0.1471	0.4159	-12.562	-0.0305	-0.56	-12.5896	
	Y15M	0.1265	0.498	-19.5899	0.8358	-0.1738	-18.3035	
	D34H	0.4332	0.8168	-11.7998	-0.3238	-1.1726	-12.0462	
	F43W	0.1083	0.3335	-19.1664	-0.2045	-0.8702	-19.7993	
	Q103H	0.3914	0.4546	-9.1117	0.3411	-1.0115	-8.9361	

The "Show PDB" button allows the user to visualize and evaluate the introduced mutations in the sequence viewer. The user can further use the selected model for another job (as specified in the tools panel).

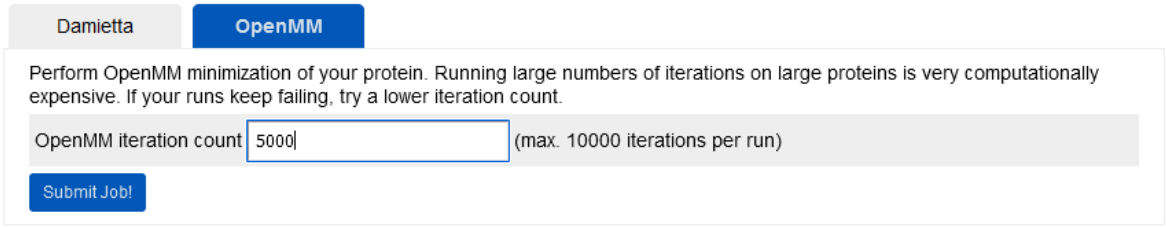


Symmetric design (`sd`). The **`sd`** protocol performs the combinatorial sampling operations as **`cs`**, while allowing for introducing sequence symmetry constraints. A mutable position's mutation is synchronized with one or more specified positions, which can be provided through an additional field, as numbers separated by commas. In the example below, position 59 is mutated to all amino acids, whereby the mutation is also evaluated at positions 35 and 29. In this situation the symmetrically-linked mutations are accepted only if the average energy across all mutable and repackable positions achieved a lower energy as a result. It worth nothing that the term symmetry here only refers to the imposed sequence symmetry, whereas the sampler will minimize the conformations in an asymmetric manner.



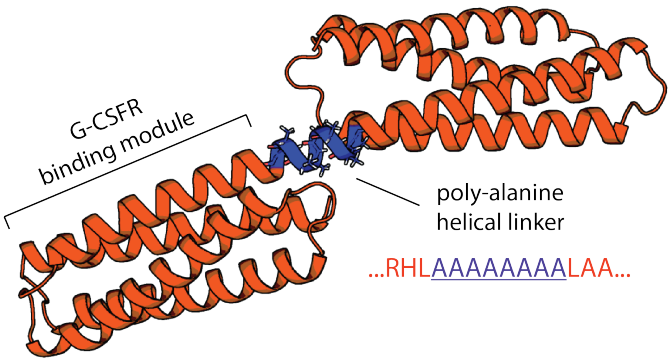
Single-point mutagenesis (`sp`). The **`sp`** protocol attempts to enforce all the listed mutations for all the positions individually, and generates the single-point mutants even if they possess higher energies than the starting model. However, if the introduced mutation exhibits substantial steric incompatibility, it will be omitted.

Long minimization (`lm`). Currently conjugate gradient minimization is made available under the OpenMM routines, which is especially useful for relaxing the designed models after a large number of mutations. The minimized structures can also be forwarded for further design as needed, and the ProteinMPNN-derived probabilities will be automatically updated after this long minimization protocol. The example below specifies 5000 minimization steps for the selected structure:



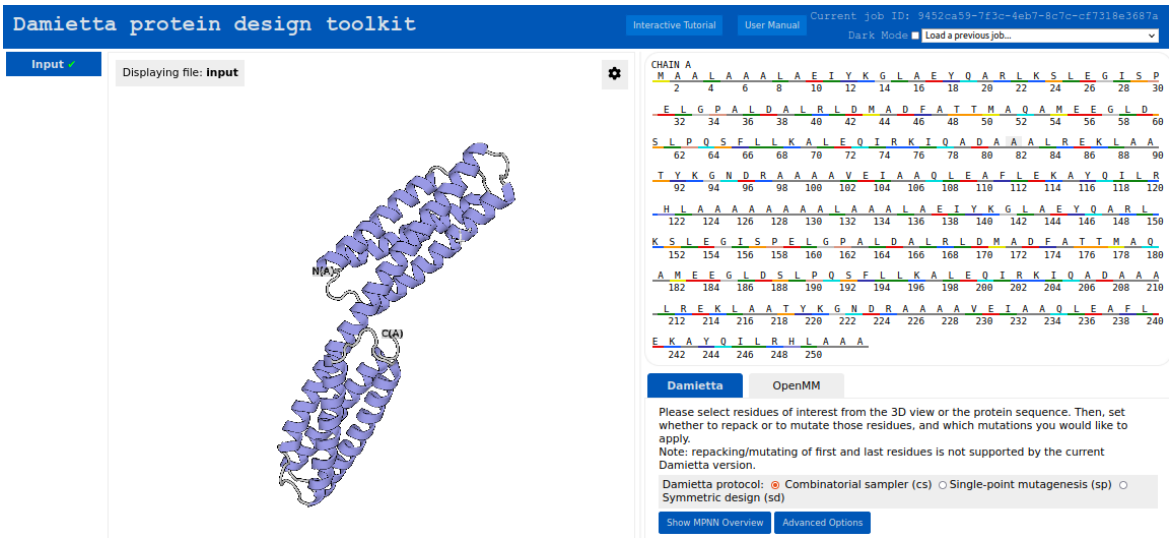
4 Case-study: Design of a rigid helical linker

This section presents an example of using the Damietta toolkit to design a rigid helical linker between two monomeric proteins in order to create a bivalent agonist able to dimerize the granulocyte colony-stimulating factor receptor (G-CSFR) at a bespoke angle. The example is taken from the study by Ullrich et. al (URL: *http://dx.doi.org/10.1101/2023.11.25.568662*). Two previously designed G-CSFR binding modules were connected N- to C-terminally with a poly-alanine helical segment (AlphaFold2 model is shown below). The task is to design the sequence of the helical linker to improve the conformational stability of the molecule.

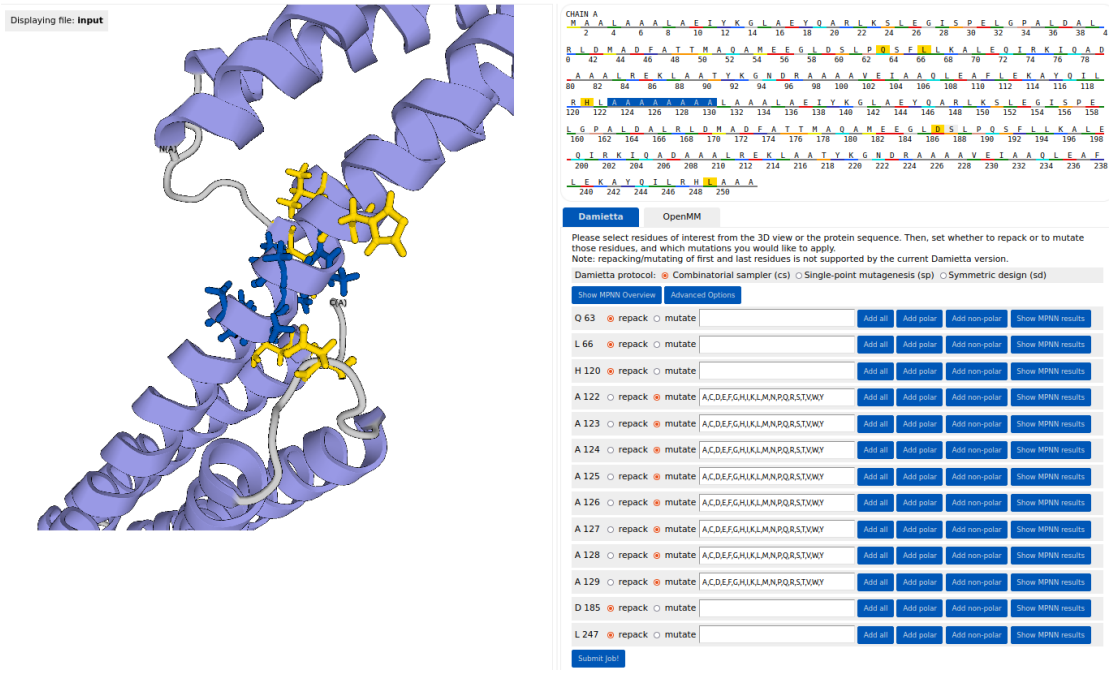


First, an AlphaFold2 model for the following sequence was uploaded using the upload form at the start page:

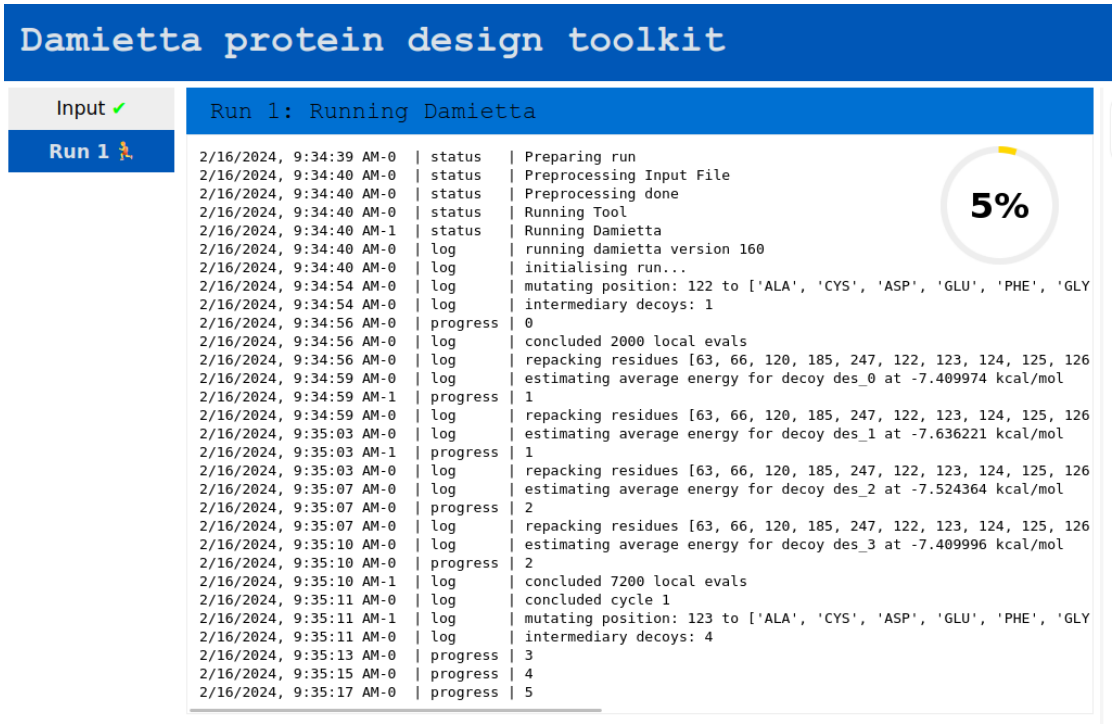
MAALAAALAEIYKGLAEYQARLKSLEGISPELGPALDALRYDMADFAILMAQAM
EEGLDSLPQSFLRKALEMIRKIQADAAALREKLAATYKGNDRAAAAVEIAAQLE
AFLEKAYQILRHLAAAAAAAALAAALAEIYKGLAEYQARLKSLEGISPELGPALD
ALRYDMADFAILMAQAMEEGLDSLPQSFLRKALEMIRKIQADAAALREKLAATY
KGNDRAAAAVEIAAQLEAFLEKAYQILRHLAAA



Next, under the "Damietta" tab the **`cs`** tool was chosen to design the sequence of a rigid helical linker. The linker residues from A122 to A129 were selected as mutable (highlighted in blue). Using the "Add all" button, all amino acids were specified as possible options at every mutable position. Residues Q63, L66, H120, D185, and L247 were selected as repackable (highlighted in yellow).



After submitting the job, the job log is shown, which allows to track the progress of the run:

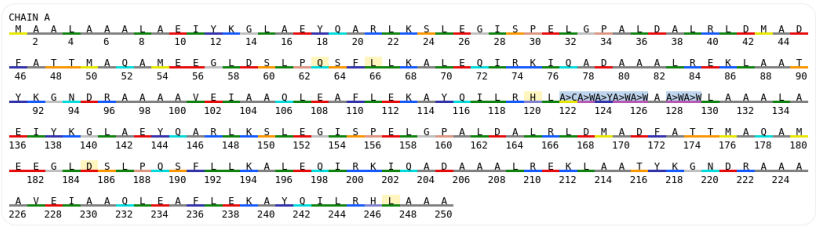


Five design candidates were reported in the result table. The lowest-energy candidate (i.e. result2.pdb) had the following mutations: A122C, A123W, A124Y, A125W, A126W, A128W, and A129W.

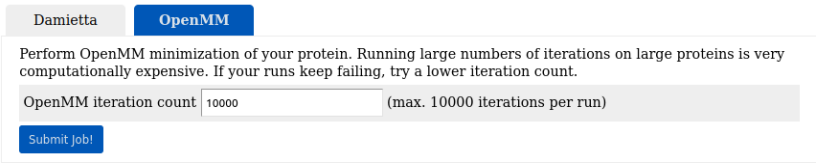
Result File	Position	ΔG_{pp}	ΔG_{k}	ΔG_{LJ}	ΔG_{solv}	ΔG_{elec}	ΔG_{total} ▾	
result2.pdb	A122C, A123W, A124Y, A125W, A126W, A128W, A129W	0.1955	0.3284	-12.033	1.1089	-0.4819	-10.8821	Details Show PDB
result1.pdb	A122C, A123W, A124Y, A125W, A126W, A128W, A129R	0.1954	0.3607	-12.1776	1.1917	-0.4164	-10.8462	Details Show PDB
result0.pdb	A122C, A123W, A124Y, A125W, A126W, A128W	0.1947	0.3248	-11.8344	0.8284	-0.3501	-10.8366	Details Show PDB
result3.pdb	A122C, A123W, A124W, A125W, A126W, A128W, A129W	0.1863	0.3076	-11.9781	1.1595	-0.4998	-10.8245	Details Show PDB
result4.pdb	A122R, A123M, A124Y, A125W, A126W, A128W, A129W	0.1937	0.3992	-12.3118	1.4761	-0.5476	-10.7904	Details Show PDB

all energies in (kcal/mol)

The introduced mutations could be also checked in the sequence viewer:



To relax the designed model it was forwarded to the long minimization **`(lm)`** tool with 10000 minimization steps:



By clicking "Download result archive" button, the output containing the PDB file for the minimized structure together with the job log was downloaded.

5 Limitations of the Damietta toolkit

The current version of the Damietta toolkit can handle protein structures with maximum 1000 residues. For processing bigger proteins, please contact us or download the full version of the Damietta software: *https://bio.mpg.de/damietta*

The first (N-terminal) and the last (C-terminal) residues of the protein can not be mutated, since either ϕ or ψ dihedral angle is not defined for them.

The current version of the Damietta toolkit does not account for any interactions with heteroatoms (e.g. ligands, cofactors, ions, solvent molecules).

6 References

The reference for the Damietta toolkit will be provided soon.

If you used one of the Damietta tools, please cite:

Maksymenko et al., The design of functional proteins using tensorized energy calculations, 2023, Cell Reports Methods (doi:10.1016/j.crmeth.2023.100560).

If you used OpenMM, please cite:

Eastman et al., OpenMM 7: Rapid development of high performance algorithms for molecular dynamics, 2017, PloS Computational Biology (doi:10.1371/journal.pcbi.1005659).

If you relied on the ProteinMPNN-provided suggestions, please cite:

Dauparas et al., Robust deep learning–based protein sequence design using ProteinMPNN, 2022, Science (doi: 10.1126/science.add2187).