ProteinDJ: a high-performance and modular protein design pipeline

Dylan Silke[1,2], Julie Iskander[1,2], Junqi Pan[1,2], Andrew P. Thompson[1,2,3], Anthony T Papenfuss[1,2], Isabelle S. Lucet[1,2,3], Joshua M. Hardy[1,2,3]

[1]The Walter and Eliza Hall Institute of Medical Research, Parkville, VIC, Australia.
[2]Department of Medical Biology, University of Melbourne, Parkville, VIC, Australia.
[3]ARC Centre for Cryo-electron Microscopy of Membrane Proteins, Walter and Eliza Hall Institute of Medical Research, 1G Royal Parade, Parkville, 3052, Victoria, Australia

Leveraging artificial intelligence and deep learning to generate proteins *de novo* has unlocked new frontiers of protein design. By training deep learning models on protein sequences and experimental structures, we can sample new structural landscapes unexplored by evolution. This approach can be used to design bespoke binders that target specific proteins and domains. However, generating successful binders has a low in silico success rate, often requiring thousands of designs and hundreds of GPU hours to obtain enough hits for experimental testing. There is a lack of efficient open-source pipelines designed for high-performance computing (HPC) systems that can maximise hardware resources and parallelise the workflow efficiently.

Here, we present 'ProteinDJ'[1] - an implementation of a synthetic protein binder design workflow that is deployable on HPC systems using the Nextflow pipeline language and Singularity containerisation. It automatically batches and parallelises the workload across both GPUs and CPUs. Importantly, it allows central deployment of the workflow on shared HPC resources, rather than workstations that are restricted to individual research groups.

ProteinDJ is modular by design (see Figure 1) and currently includes RFdiffusion for fold generation, ProteinMPNN or Full-Atom MPNN for sequence design, and AlphaFold2 or Boltz-2 for prediction and validation of binder-target interfaces, with supporting packages for structural evaluation of designs. We have included structure-based filters to reject problematic designs mid-stream to minimise manual inspection of designs by the user. Our approach democratises protein binder design in an easy-to-use and robust implementation freely available on GitHub (https://github.com/PapenfussLab/proteindj).

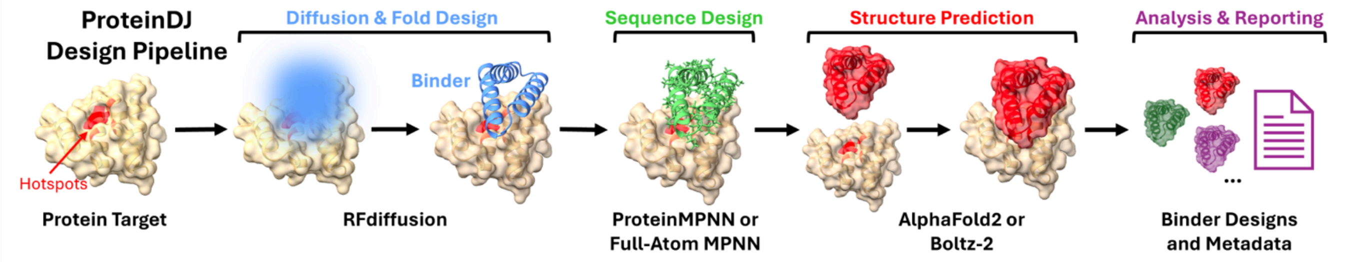

Figure 1. Overview of the ProteinDJ pipeline for de novo binder design. RFdiffusion is used to diffuse a binder fold near hotspots on a target. The user can choose from ProteinMPNN or Full-Atom MPNN for sequence design, and AlphaFold2 Initial Guess or Boltz-2 for structure prediction and validation.

1. Silke, D., et al. ProteinDJ: a high-performance and modular protein design pipeline. *bioRxiv* (2025) https://doi.org/10.1101/2025.09.24.678028