# A platform for Protein Design

**Nicolás López Carranza[1] • Marcin J. Skwark[1] • Thomas Pierrot[1] • Hippolyte Jacomet[1] • Ashraf Guitouni[1] • David Sehnal[1,2,3] • Boulbeba Mallouli[1]**
**Aliou Kayantao[1] • Aurelien Delfosse[1] • Nacef Labidi[1] • Beyrem Makhlouf[1] • Roser Gonzalez Franco[1] • Alexander Laterre[1] • Zohra Slim[1] • Karim Beguir[1]**
(1) InstaDeep™ - (2) CEITEC - (3) Protein Data Bank in Europe (PDBe)

DeepChain™ by InstaDeep™

## Introduction

Here we present a novel, accessible and user-friendly platform for protein design. The user of the platform inputs a PDB file describing the 3D structure of the protein complex and through a simple, visual user interface selects the positions at the interface to apply AI directed mutagenesis to. DeepChain™ will find a sequence of mutations that improve the binding energy between both parts, or any metric designed by the user in a form of a scorer.

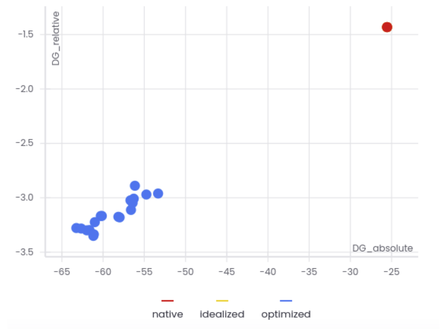
This scoring metric, together with state-of-the-art multiple objective optimisation algorithms and pre-trained transformer models allow DeepChain™ to mutate protein sequences from a PDB file and propose highly plausible designs within a few hundred thousand steps.

## Protein Design using statistical biophysics and macromolecular scoring functions

Multiple objective optimisation algorithms, state-of-the-art bioinformatics and pre-trained transformer models allow DeepChain™ to mutate protein sequences from a PDB file and propose highly plausible designs within a few hundred thousand steps. Key use cases include:

• Analysing mutations of interest, to discover how these affect the binding and stability of a protein complex, to study evolution and/or disease associated protein variants.

• Uncovering mutations that increase or decrease binding between proteins of interest to optimize target specificity, improve enzyme activity, and vaccine immunogenicity, as well as build sensors or facilitate research of big protein complexes.



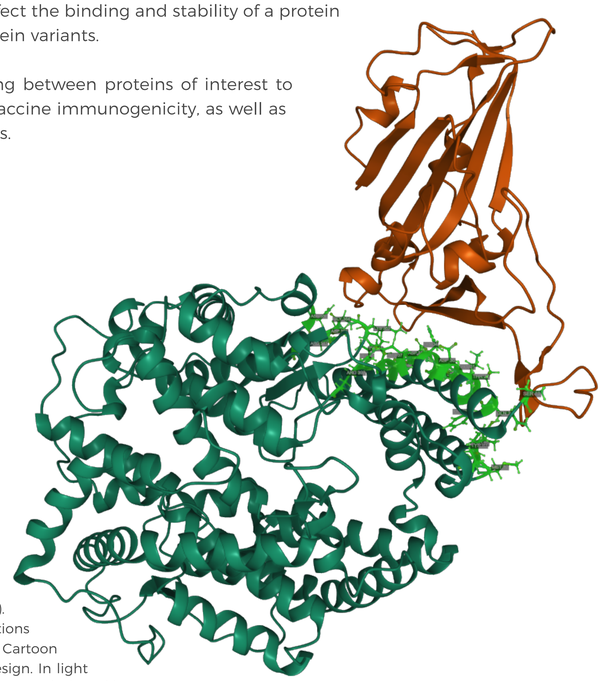



Figure: Left : DeepChain™ results of an AI-directed mutagenesis run on the complex 6lzg (Spike glycoprotein from SARS-CoV2 in contact with the human ACE2 receptor) in terms of the interface free energy (δG). The red dot is the input native structure and the blue dots are mutations proposed by DeepChain™ in terms of the interface free energy. Right: Cartoon representation of the complex pose for energetically most favorable design. In light green we see the positions highlighted by the user, depicting a mutated ACE2 designed by the platform. DeepChain uses Molstar [2] to visualize 3D structures.

## Protein exploration leveraging transformer models trained on millions of protein sequences

Transformer models trained on millions of protein sequences [3,4] contain powerful insights on protein grammar and semantics, where good grammar indicates des feasible biophysical characteristics, and semantics refers to the functional and structural attributes. Transformers have a clear use case on predicting mutation tolerance and studying evolutionary preserved variants.

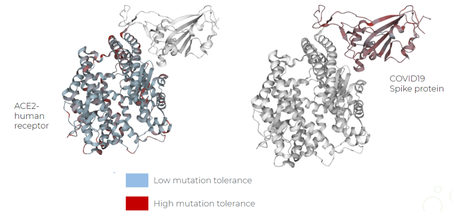



Figure: The transformer model displays protein variability, which is greater in viral proteins vs human ones

## Build machine-learning scorers to personalise protein design and analysis

DeepChain™ Apps is a collaborative framework that allows the user to create personalised ML-enabled scorers to evaluate protein sequences. These scorers can be either Classifiers or Predictors trained on protein embeddings generated by powerful transformer models trained on billion of amino acid sequences.

• Classifiers identify and classify a particular characteristic to explore proteins and datasets: Location, pathogen vs human sequence, protein family.

• Predictors score for a particular characteristic and serve to design new proteins or analyse known variants: Toxicity, Immunogenicity.

We have open-sourced 3 repositories to share biological datasets, trained apps and easily work with bio-transformers.

## Validate with Molecular Dynamics Simulations

DeepChain™ allows the user to run and visualise MD simulations powered by GROMACS[6] MD simulator.

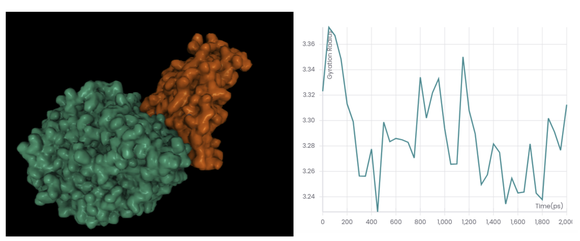



Figure: Left: 3D MD simulation trajectory viewer
Right: radius of gyration as a function of time as displayed in DeepChain™

## Conclusion

We introduce a novel platform for protein design, by combining state-of-the-art optimisation techniques, natural language processing applied to amino acid sequences, and computational structural biology. This approach yields highly prospective binders which result in stable, affine complexes. The platform also allows the optimisation of custom scores generated by the user, based on Neural Networks leveraging the predictive power of protein embeddings generated by Transformer Models trained on billions of amino acid sequences. DeepChain™ allows cross validation of designs through Gromacs MD simulations. DeepChain™ has been proved successful in many protein design problems with an entirely hands-off approach.

**References:**
(1) Designing a Prospective COVID-19 Therapeutic with Reinforcement Learning arXiv:2012.01736
(2) Mol* : A modern web-based open-source toolkit for visualisation and analysis of large-scale molecular data - https://molstar.org/
(3) https://github.com/DeepChainBio/ **bio-datasets:** Free collection of Bio datasets and embeddings, bio-transformers: A place where transformers meet biology. **deep-chain-apps:** Build powerful predictors and classifiers, in minutes.
(4) ProtTrans: Towards Cracking the Language of Life's Code Through Self-Supervised Deep Learning and High Performance Computing - bioRxiv 2020.07.12.199554; doi: https://doi.org/10.1101/2020.07.12.199554
(5) Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences - PNAS April 13, 2021 118 (15) e2016239118;
(6) GROMACS 2021.1 (Version 2021.1) http://doi.org/10.5281/zenodo.4561625