Protein Design with Hybrid AI

Design with Intuition[1] and Logic[2]

An Hybrid AI protein design architecture

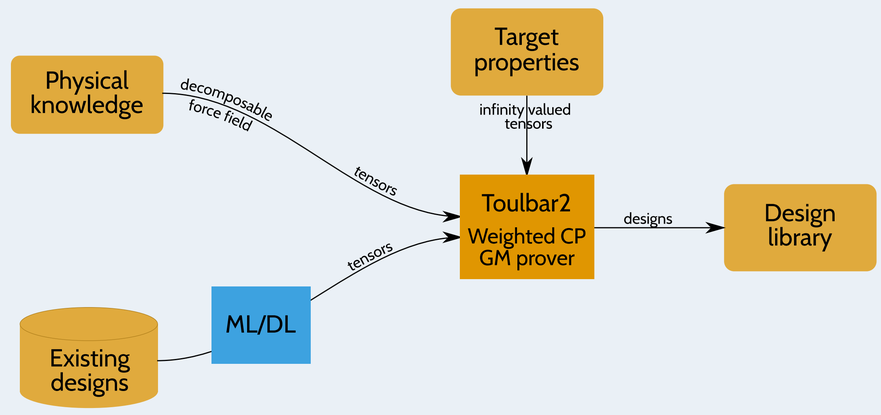
T. Schiex, S. Barbe, S. de Givry, G. Katsirelos, D. Simoncini

Designing physical objects typically starts from:

- a "design target", a set of original design properties that the design must satisfy, expressed as (possibly non-pariwise) constraints.
- to make the final design feasible, these properties must be combined with the laws of physics.
- working or successful past designs from which features characterizing working or successful designs can be extracted.

Our protein design architecture uses our award-winning guaranteed graphical model (GM) solver, TOULBAR2 to integrate physical pairwise decomposable force fields (laws of physics), target design properties (expressed as constraints) and machine/deep-learned information extracted from experimental data on existing protein with sequence and/or their 3D structure and/or their fitness for the design target.

The ML/DL technology we leverage (or create) must produce features in a mathematical format that suits GMs, where functions are typically described as a sum of (possibly infinite-valued) tensors. Energy-based model losses, combined with discrete, convex or stochastic gradient optimization are our favorite tools here.



1 - Deep Learning a coarse-grain protein energy function[1]

Physical knowledge about proteins is summarized into approximate empirical energy functions. With the large amount of available protein data, Deep Learning offers an appealing approach to learn a decomposable energy function directly from data. However, the specificity of protein structures must be accounted for to design an efficient neural architecture:

- the architecture should be insensitive to global rotations and translations (SE(3)-equivariant).
- it should take into account local geometrical arrangements
- and be able to capture interactions between the basic elements of the protein (amino acids)

So far, no consensus has emerged on the most adequate representation of protein structures: hand-crafted features, voxels, distance maps, graphs and point clouds have all been used[1].

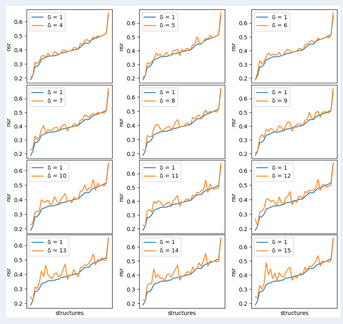
We chose an approach based on features representing local environments, processed using attention mechanisms to extract the important interactions between amino acids and their neighbours. A second neural net processes pairs of amino acids to produce a full energy matrix, optimized to maximize amino acid recovery.

After training on an independent dataset of 18,000 proteins, we obtained an energy function which, on the task of separating natural proteins from artificial ones, using the authors' benchmark set, is able to outperform KORP, a state-of-the-art statistical potential (Bioinformatics, 2019).

2 - Generating libraries of diverse proteins[2]

Computational protein design fundamentally relies on several approximations, possibly combined with learned information extracted from data, including sampling noise. The design process remains therefore partly unreliable, requiring the generation of libraries of sequences for experimental characterization. In this work, relying on our recent results on guaranteed diverse optimal solutions (see block 4, 5 in associated poster), we observe that the generation of increasingly diverse minimum energy sequences allows to improve the correct reconstruction of natural protein sequences.

From a collection of 15 natural proteins structures extracted from the Protein Data Bank, we designed a library of 10 minimum energy sequences separated from each other by a Hamming distance of $\delta \in \{4 \ldots, 15\}$. For each δ, the best obtained sequence recovery is compared to $\delta = 1$. The X axis represents the 15 proteins, ordered by increasing $\delta = 1$ best sequence recovery. The Y axis is the sequence recovery. One can clearly observe that increasing diversity increases sequence recovery, especially for proteins with poor $\delta = 1$ recoveries.



3 - Positive Multi-State Design[5]

Proteins are flexible objects and this flexibility may be crucial for their function. In this work, we show that multi-state formulations of Protein Design that account for flexibility define a challenging Σ_p^2 problem. We identify an NP-complete class and exploit the capability of Toulbar2 to enforce constraints to encode the problem in a Cost Function Network that TOULBAR2 can solve. On various benchmark problems, Positive Multi-State Design significantly improves the accuracy of reconstruction of natural proteins and also shows improved efficiency (given the increased search space). It also outperforms a recent guaranteed multi-state design tool (Karimi et al. Bioinformatics, 2018) by several orders of magnitude.

Sequence recoveries					
NMR structures			X-ray structures		
SSD	min-MSD	Σ-MSD	SSD	min-MSD	Σ-MSD
50.8%	50.3%	66.4%	56.7%	58.0%	64.7%

A comparison of the proposed guaranteed multi-state design approach (Σ-MSD) on two data-sets of natural proteins with either an ensemble of NMR-resolved structures or a backrub-generated ensemble generated from a single X-ray structure. In both cases, Σ-MSD offers significantly improved reconstruction accuracy.

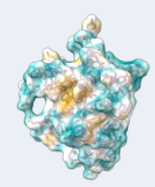
4 - Designing proteins for real

To get feedback on our hybrid AI design engine and to develop visible and impactful solutions for biotech, we push our design technology to the real world by collaborating with biochemists, favoring applications on potential answers to the challenging questions defined by climate change and pandemics.

Coll. A. Olichon, CRCT/INSERM: nanobodies are small antibody-like proteins. Our Hybrid AI design capacities were crucial for this redesign of nanobody scaffolds.

- Constraints: chemical composition, patentable, hydrophilic
- Data: existing nanobody sequences (llamas, sharks, humanized)
- Physics: Rosetta beta_nov16 score function
- Multi-state: to account for multiple CDR loop sets.

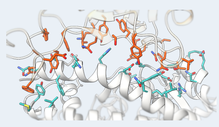
Several functional (able to carry CDR loops with experimentally checked affinity) nanobody scaffolds were produced, some having very high yields.



Coll. C. Bahl, IPI, Boston: we explored the evolutionnary landscape of CoViD-SARS-2 (with early polyclonal vaccine design as a long term target). The design relied purely on the automated reasoning capacities of Toulbar2.

- Constraints: Affinity for human receptor, stability
- Data: None beyond the X-ray structure (RBD bound to ACE2)
- Physics: Rosetta genpot_full score function, glycosylation
- Query: exhaustive sequence enumeration (91 million produced)

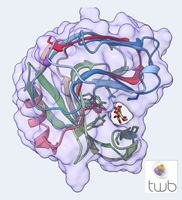
Among 59 predicted variants, 7 infectious antibodies-resistant pseudo-viruses have been experimentally characterized, some very different from known variants.



Toulouse White Biotech support: we redesigned an engineered biomass degrading enzyme (GH11 xylanase) with many applications in bio-energies and green chemistry.

- Targets: thermo-stability, thermo-resistance, catalytic activity, hydrophilic
- Data: MD simulations[3] (no ML)
- Physics: Rosetta beta_nov16 score function

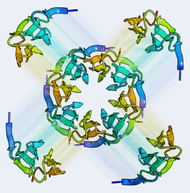
We proposed several mutants with improved thermostability, thermo-resistance and catalytic activity of an already engineered and very active enzyme mutant.



Coll. A. Voet (KU Leuven, Belgium): full design of a self-assembling symmetric beta-propeller[6] (evidence for gene duplication in evolution).

- Constraints: conserved residues in existing designs, symmetry (Rosetta)
- Data: existing designs (no ML, predated its availability)
- Physics: Rosetta talaris_14 score function

X-ray structure determined. With 2 different design engines (Rosetta + evolution vs. Toulbar2), the Toulbar2 design was the only one able to self assemble, also showing extreme stability (both thermal and chemical).



Coll. S Tagami (Riken, Japan): full design of a (self-assembling) symmetric double-psi beta-barrels[4] (evidence for early apparition of DPBBs in evolution).

- Constraints: limited chemical variety, symmetry (Rosetta)
- Data: existing DPBB structures (no ML)
- Physics: Rosetta beta_nov16 score function

Several protein sequences that could fold into DPBBs structure, some using only 7 different amino acid types, were characterized (by X-ray diffraction), reinforcing the evidence for the possible early apparition, during evolution, of DPBB-structures, an important RNA polymerase domain.

Several other proteins have been or are in the process of being co-developed with other research institutes or for biotech companies, targeting functions such as plastic depolymerisation, CO2 to O2 conversion, antibiotic resistance or CoViD-blocking,...

1. Defresne, M. *et al.* Protein Design with Deep Learning. *International Journal of Molecular Sciences* **22,** 11741 (2021).
2. Ruffini, M. *et al.* Guaranteed diversity and optimality in cost function network based computational protein design methods. *Algorithms* **14,** 168 (2021).
3. Vucinic, J. *et al.* A Comparative Study to Decipher the Structural and Dynamics Determinants Underlying the Activity and Thermal Stability of GH-11 Xylanases. *Int. J. Molecular Sciences* **22,** 5961 (2021).
4. Yagi, S. *et al.* Seven Amino Acid Types Suffice to Create the Core Fold of RNA Polymerase. *Journal of the American Chemical Society* **143,** 15998–16006 (2021).
5. Vucinic, J. *et al.* Positive multistate protein design. *Bioinformatics* **36,** 122–130 (2020).
6. Noguchi, H. *et al.* Computational design of symmetrical eight-bladed β-propeller proteins. *IUCrJ* **6,** 46–55 (2019).