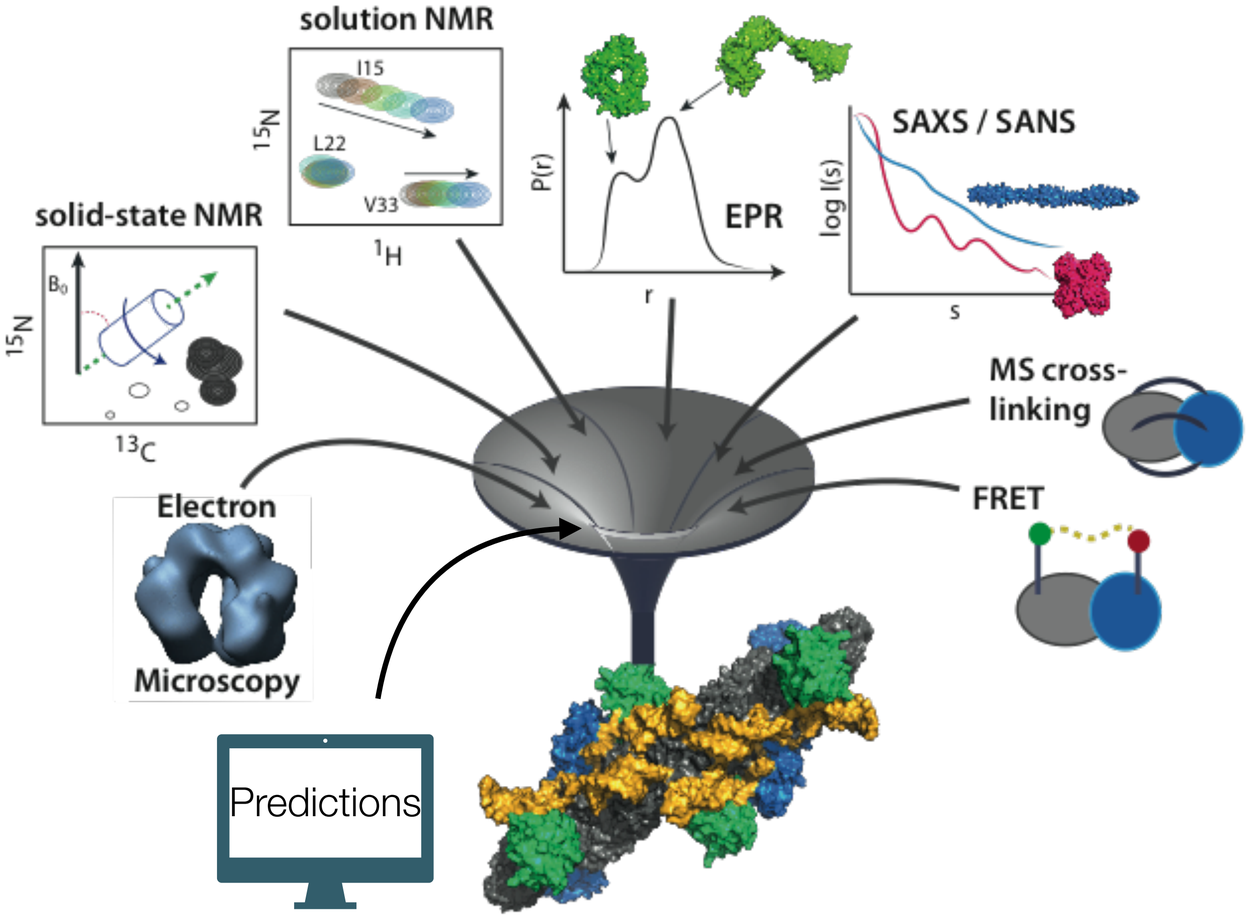
Integrative Modelling & Protein Design

Structural Bioinformatics

Integrative Structural Biology: how to determine structures using sparse and noisy data

There are cases of systems whose structure cannot be characterised using a single structural biology technique because they are either to difficult to be kept stable for long enough or because they populate very heterogeneous conformations. Yet it may be possible to accumulate enough information that if integrated may provide a structure or an ensemble of structures.

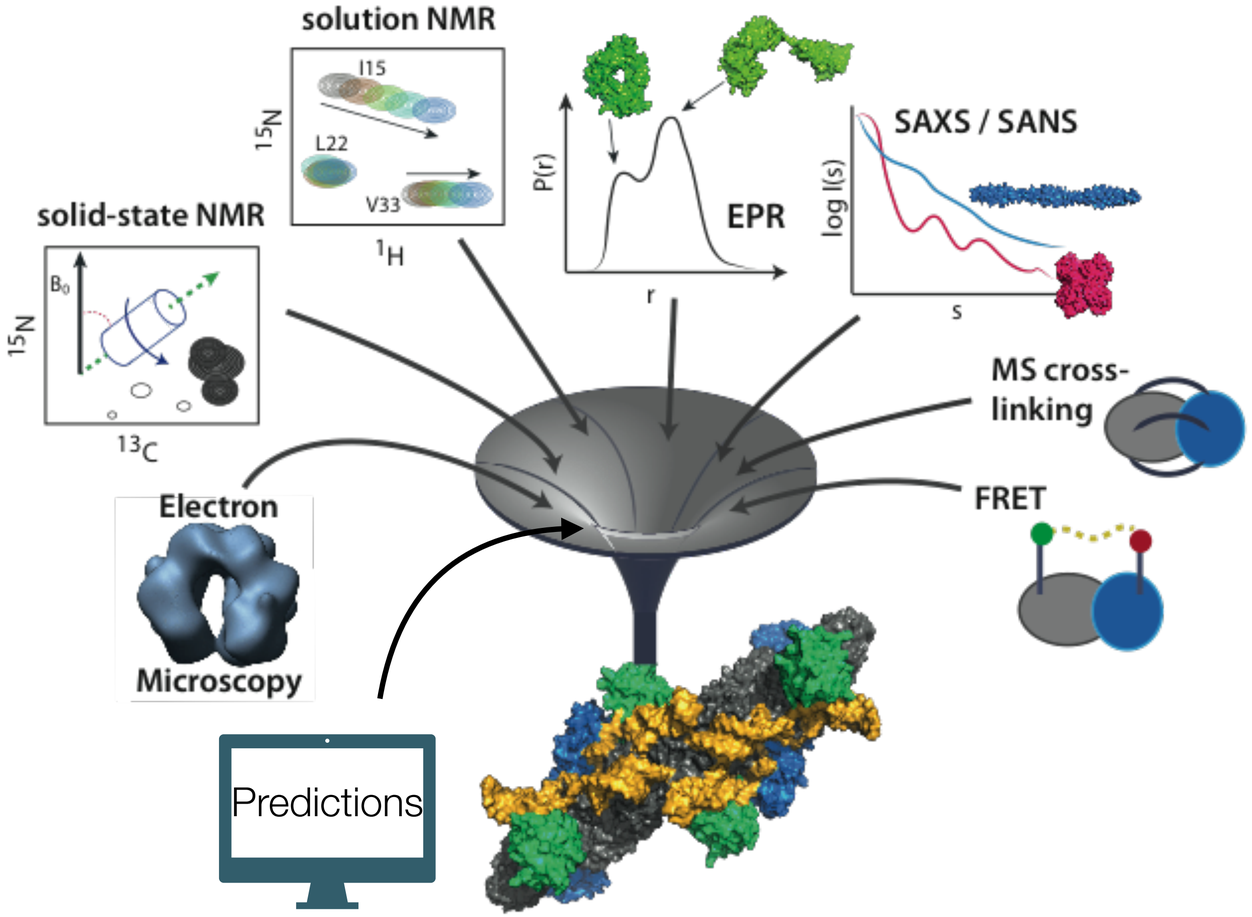


Integrative Structural Biology: weighing and optimising

The problems to be faced in integrative structural biology are essentially two and they are intertwined:

- How should we balance the information derived from different techniques?
- How do we define a score to be optimised to generate a final model?

In practice we would like to keep into account the uncertainty of the different experimental and theoretical techniques and use them to update some “prior” knowledge about our system.



Bayesian Inference: from the prior information

Let's say that we know the structures of the single proteins making a complex and that we also know some physico-chemical properties of how they may interact (for example using a force-field). We can call this our **prior information *I*** that is our starting point to build a **model *M*** for our complex (the model ***M*** will be the configuration of the complex $\boldsymbol{X_M}$ but could also include other parameters). In principle we can assemble the single structures, for which for example we do not know the stoichiometry, in many ways and we would like to score the different results in terms of their probability ***p(M|I)*. Now you know that we can write the probability of the conformation X_M of the model *M* as:**

$$p(X_M|I) \propto \exp\left(-\frac{E(X_M)}{k_B T}\right)$$

This is the probability of that configuration given the prior information, which in this case is its energy as approximated by a force field. You can imagine that **there are many many possible configurations** with low energy **and we lack a lot of information to decide which may be the correct one**.

Bayesian Inference: we update with new knowledge

$$p(X_M|I) \propto \exp\left(-\frac{E(X_M)}{k_B T}\right)$$ Prior

Now let's say that we then perform experiments to learn for example about the stoichiometry of the different components; some distance between proteins to know which ones should be close to which ones; and informations about the overall shape of the complex. **We want to supplement these data *D* to our prior information *I***. What we should do is to be able, given a configuration X_M, to calculate the **signal we would observe from that configurations $f_i(X_M)$, compare it with the experimental data point d_i, and get some probability given the accuracy of the experiment** σ_i. We can often model this probability as a Gaussian:

$$p(D|M,I) = \prod N(d_i|f_i(X_M),\sigma_i) \approx \prod \exp(-(d_i - f_i(X_M))^2/(2\sigma_i^2))$$

Considering all the experimental data point this is the **p**robability of observing the experimental **D**ata **|** given the **M**odel and the prior **I**nformation.

Bayesian Inference: to obtain a Posterior probability

$$p(X_M|I) \propto \exp\left(-\frac{E(X_M)}{k_B T}\right)$$ Prior

$$p(D|M,I) = \prod N(d_i|f_i(X_M),\sigma_i) \approx \prod \exp(-(d_i - f_i(X_M))^2/(2\sigma_i^2))$$ Likelihood

Now, **to combine the two probabilities we can multiply them, this will give the joint probability of observing a Model given the Data Likelihood and the Prior Information**

$$p(M|D,I) \propto p(D|M,I)p(M|I)$$ Posterior Probability: Bayes' Theorem

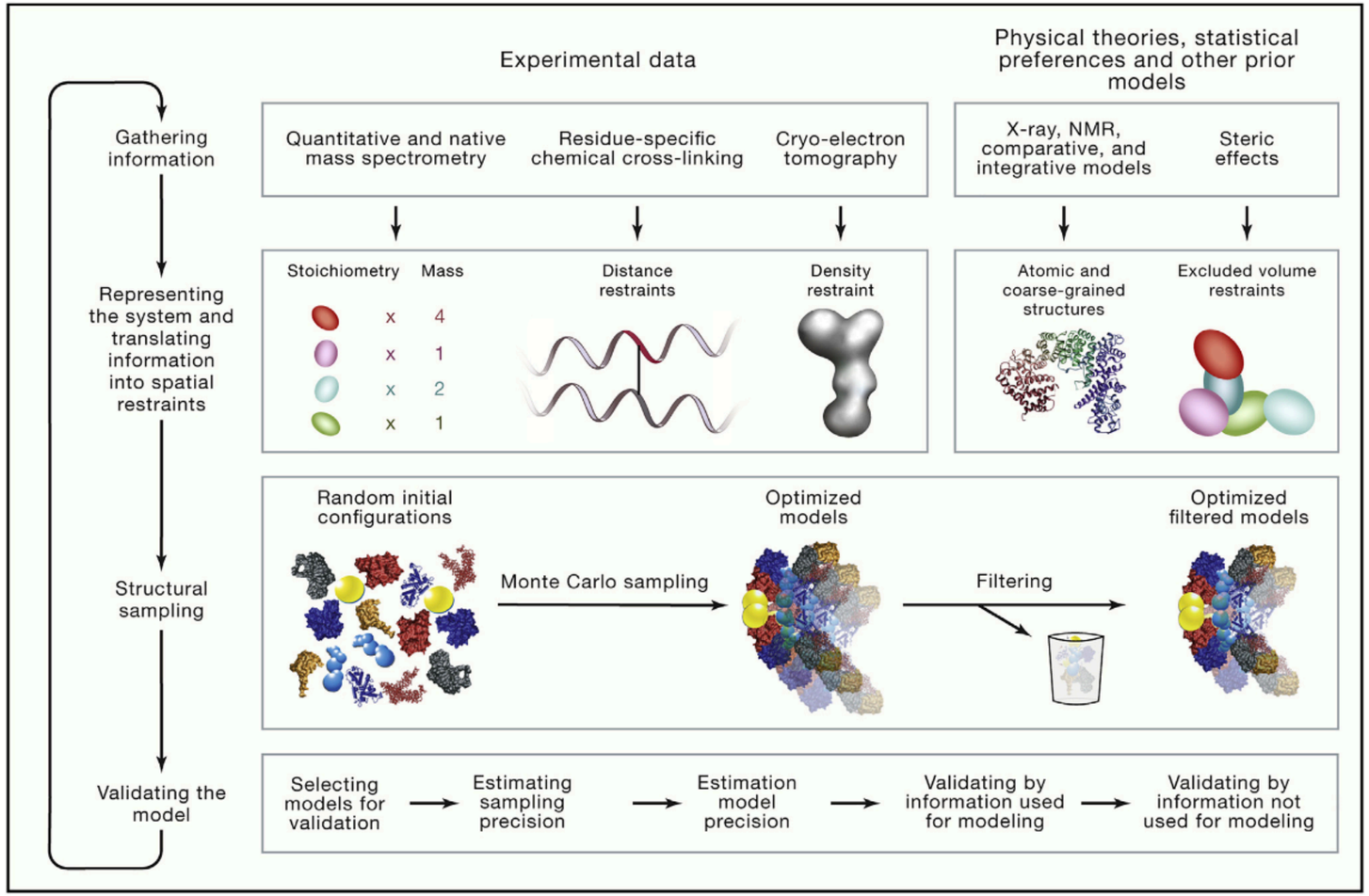
The last step is to get a scoring function, an energy, that will be the result of the integration of all our knowledge:

$$E_{int}(M) = -k_B T \log(p(D|M,I)) = -k_B T[\log(p(D|M,I)) + \log(p(M|I)) + C] =$$

$$= k_B T \sum \frac{(d_i - f_i(X_M))^2}{2\sigma_i^2} + E(X_M) + c$$

The new energy is the energy we used for prior plus a sum of quadratic restraints that try to keep our model in agreement with the experimental data accounting for their accuracy (including our ability to back calculate them).

Example of a workflow

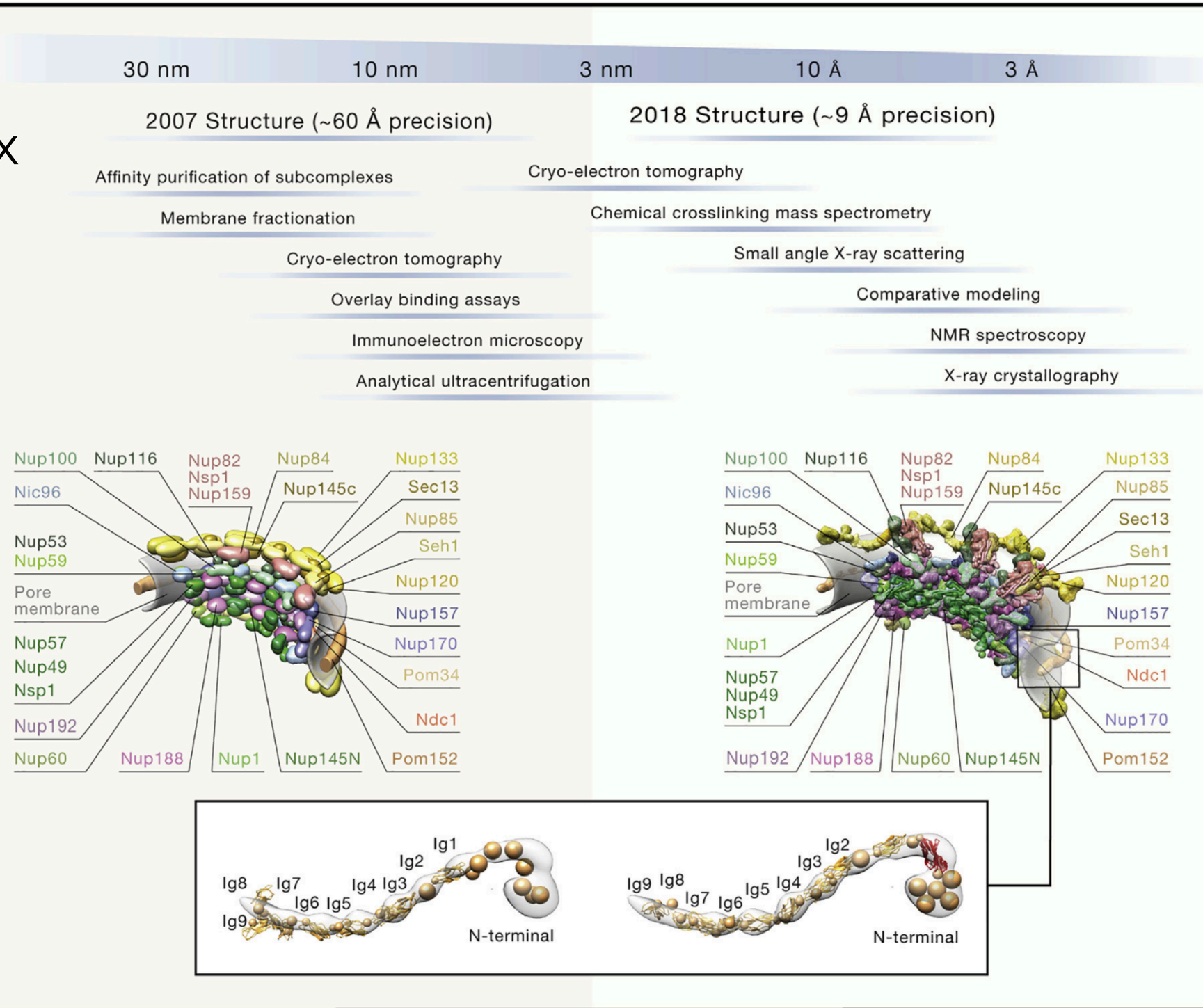


Rout, M. P. & Sali, A. Principles for Integrative Structural Biology Studies. *Cell* **177**, 1384–1403 (2019).

The structure of the Nuclear Pore Complex

Rout, M. P. & Sali, A. Principles for Integrative Structural Biology Studies. *Cell* **177**, 1384–1403 (2019).

A comparison of the integrative NPC structures determined in 2007 (Alber et al., 2007b) and 2018 (Kim et al., 2018) illustrates how the integration of a larger amount of more precise data led in turn to a structure with a higher precision. Shown in the inset is a comparison of two representative Pom152 models, without and with an atomic model of the first Ig domain (Hao et al., 2018; Upla et al., 2017), showing how incorporation of additional information (i.e., knowledge of an atomic structure of the first Ig domain [Ig0]) into the representation of a protein improves its model.



UNIVERSITÀ DEGLI STUDI DI MILANO

DBS

Rout, M. P. & Sali, A. Principles for Integrative Structural Biology Studies. *Cell* **177**, 1384–1403 (2019).

Table 3. Examples of Integrative Structures, Shown in Figure 1

System name	Input data	Accession	Citation
Polycomb repressive complex 2 (PRC2)	21-Å-resolution negative-stain EM map and ~60 intra-protein and inter-protein cross-links	N/A	Ciferri et al., 2012[1]
RNA polymerase II transcription pre-initiation complex	16-Å-resolution cryo-EM map plus 157 intra-protein and 109 inter-protein cross-links	N/A	Murakami et al., 2013[2]
$[\Psi CD]_2$	Averaged cryo-electron tomography map, NMR	PDB: 2L1F	Miyazaki et al., 2010
Actin together with the cardiac myosin binding protein C	Crystallographic and NMR structures of subunits and domains, with positions and orientations optimized against SAXS and small-angle neutron scattering data to reveal information about the quaternary interactions	N/A	Whitten et al., 2008[3]
ESCRT-I complex	SAXS, double electron-electron transfer, and FRET	N/A	Boura et al., 2011
Human and yeast TFIIH	XL-MS data, biochemical analyses, and previously published electron microscopy maps	N/A	Luo et al., 2015
HIV-1 capsid protein	Residual dipolar couplings and small-angle X-ray scattering (SAXS) data	PDB: 2M8L PDB: 2M8N PDB: 2M8P	Deshmukh et al., 2013[4]
Proteosomal lid	Native mass spectrometry and 28 cross-links	N/A	Politis et al., 2014[5]
RNA ribosome-binding element from the turnip crinkle virus genome	NMR, SAXS, EM	https://doi.org/10.6084/m9.figshare.1295199	Gong et al., 2015[6]
40S-eIF1-eIF3 translation initiation complex	X-ray crystallography, EM, and XL-MS	N/A	Erzberger et al., 2014
Cyanobacterial circadian timing KaiB-KaiC complex	Hydrogen/deuterium exchange and collision cross-section data from mass spectrometry	N/A	Snijder et al., 2014
Core of the yeast spindle pole body (SPB)	*in vivo* FRET, SAXS, X-ray crystallography, EM, anda two-hybrid analysis	N/A	Viswanath et al., 2017a[7]
Pre-pore and pore conformations of the pore-forming toxin aerolysin	Cryo-EM data and molecular dynamics simulations	N/A	Degiacomi et al., 2013[8]
Nucleosome remodeler ISWI	XL-MS, SAXS, and protein-protein docking	N/A	Harrer et al., 2018
Urease activation complex	Mobility mass spectrometry data	N/A	Eschweiler et al., 2018
ATP synthase membrane motor	cryo-EM (~7.8 Å resolution), XL-MS, and evolutionary couplings	N/A	Leone and Faraldo-Gómez, 2016[9]

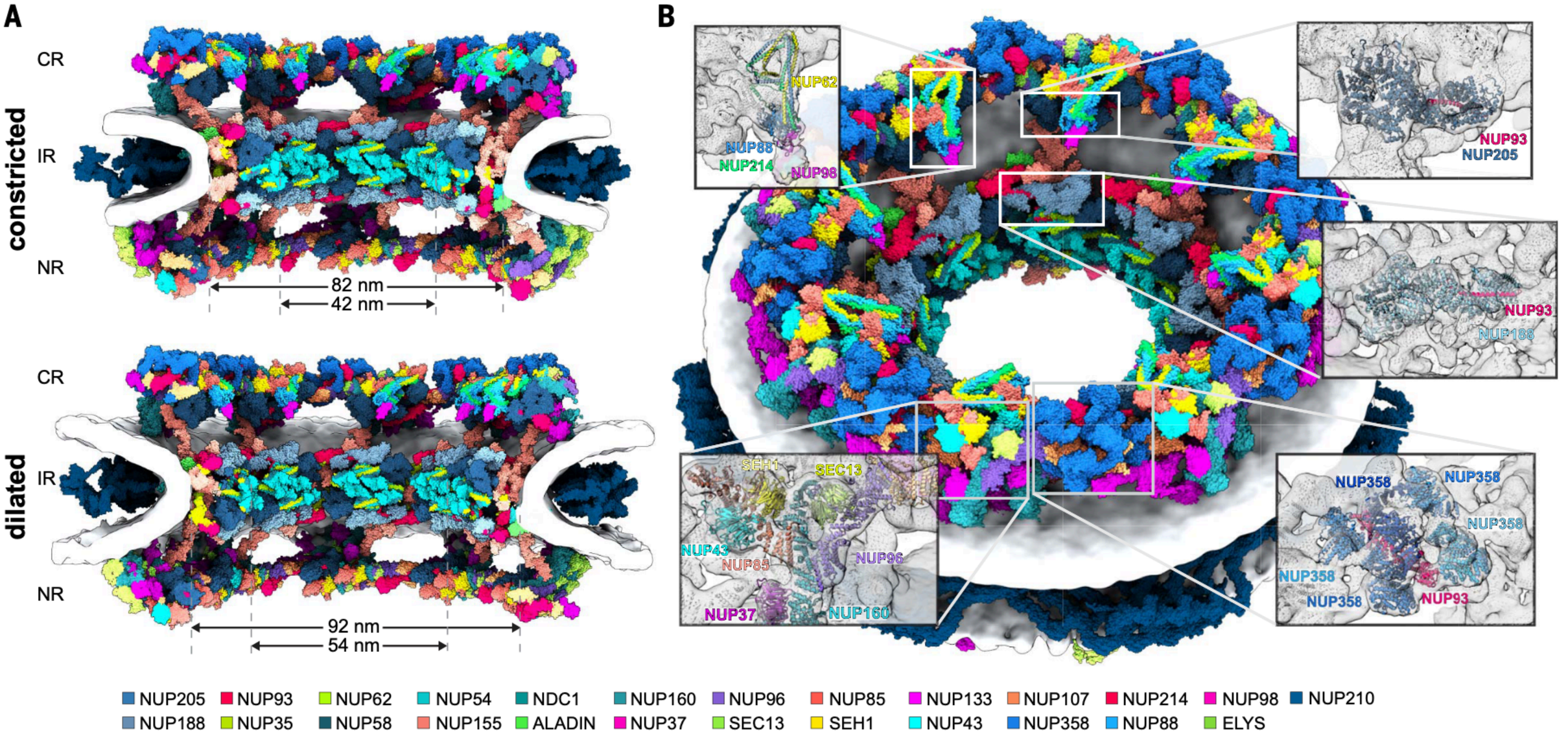


Fig. 1. Scaffold architecture of the human NPC. (**A**) The near-complete model of the human NPC scaffold is shown for the constricted and dilated states as cut-away views. High-resolution models are color coded as indicated in the color bar. The nuclear envelope is shown as a gray isosurface. (**B**) Same as (A), but shown from the cytoplasmic side for the constricted NPC. The insets show individual features of the CR and IR enlarged with secondary structures displayed as cartoons and superimposed with the isosurface-rendered cryo-ET map of the human NPC (gray).

Integrative Structural Biology: learning dynamics

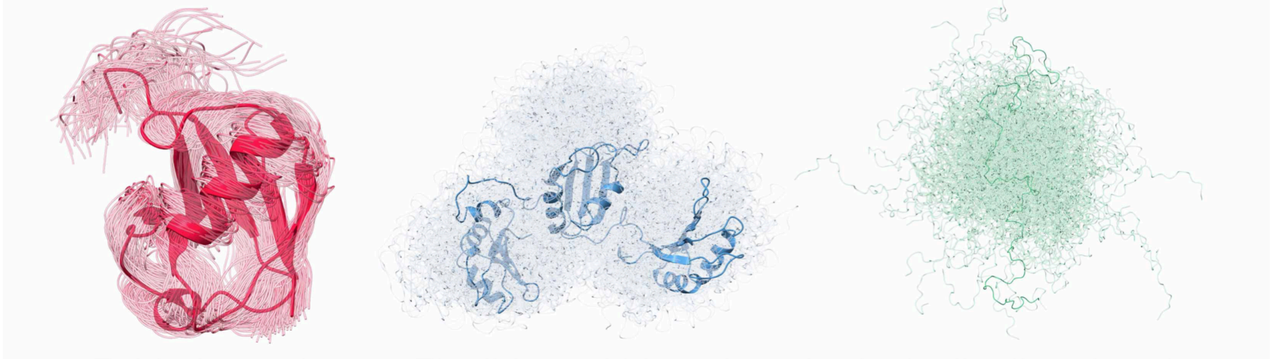


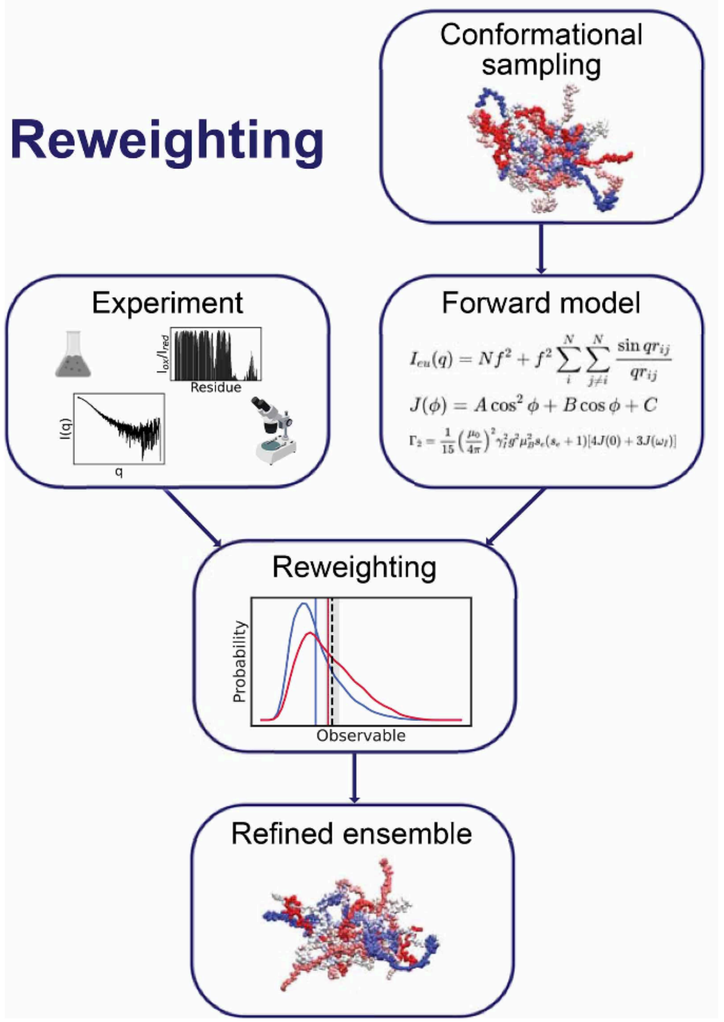
Figure 1. Conformational heterogeneity in proteins. Proteins do not exist as rigid structures in solution, but rather sample an ensemble of structures. Different proteins have variable levels of conformational heterogeneity. Stably folded proteins undergo relatively small conformational fluctuations around an average structure (left). Multidomain proteins consisting of folded domains connected by flexible linkers can display a higher level of conformational heterogeneity, as the folded domains can rearrange with respect to each other (middle). Intrinsically disordered proteins are characterized by a high level of conformational heterogeneity, as they do not fold into a well-defined structure, but rather interconvert between a range of conformations (right). The examples shown here are ubiquitin (folded protein) [22], TIA-1 (flexible multidomain protein) [23,24], and α-synuclein (intrinsically disordered protein) [24].

Bulk experiments performed at equilibrium as most NMR, SAXS, and Fluorescence experiments will report on the average behaviour over the fluctuations. In principle the correct way to interpret the data is to generate distributions, or ensembles, of conformations.

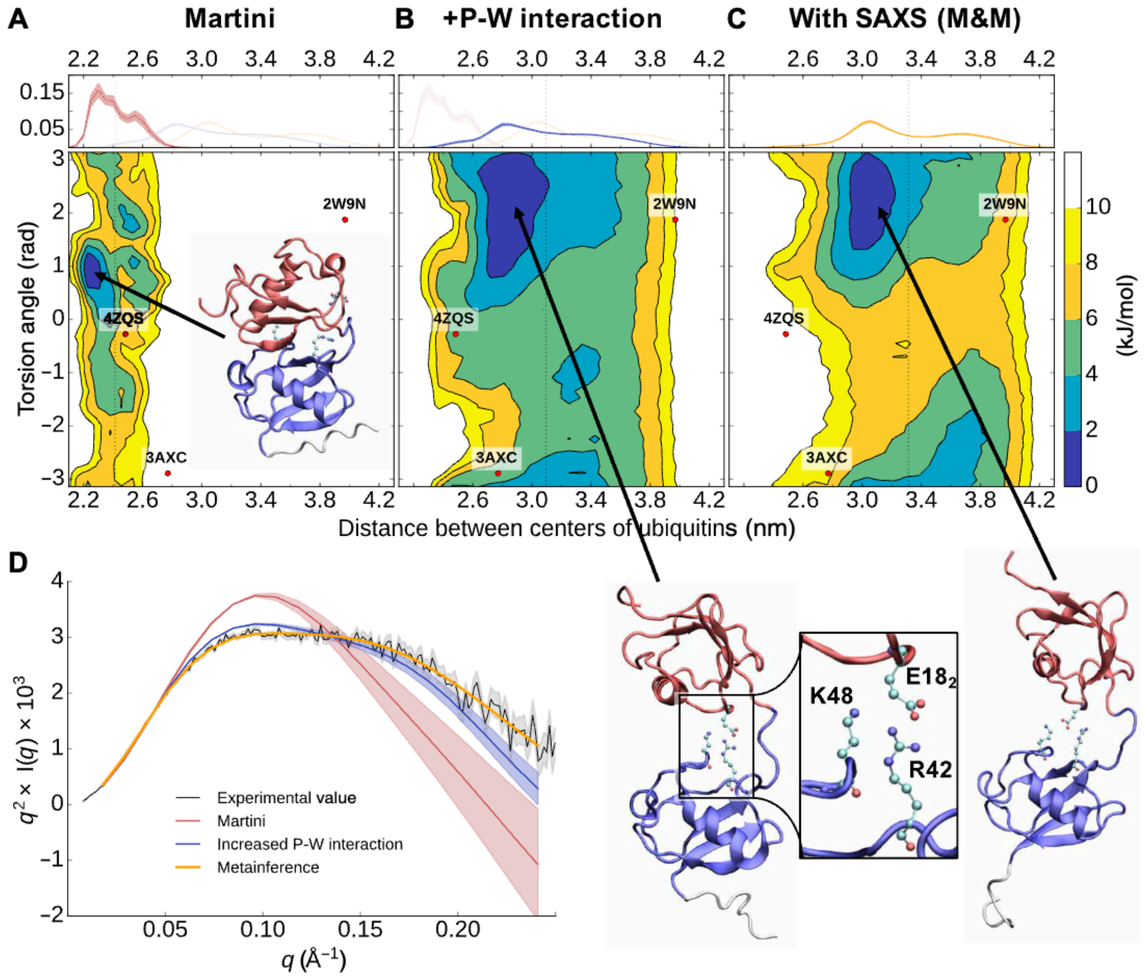
Integrative Structural Biology: learning dynamics

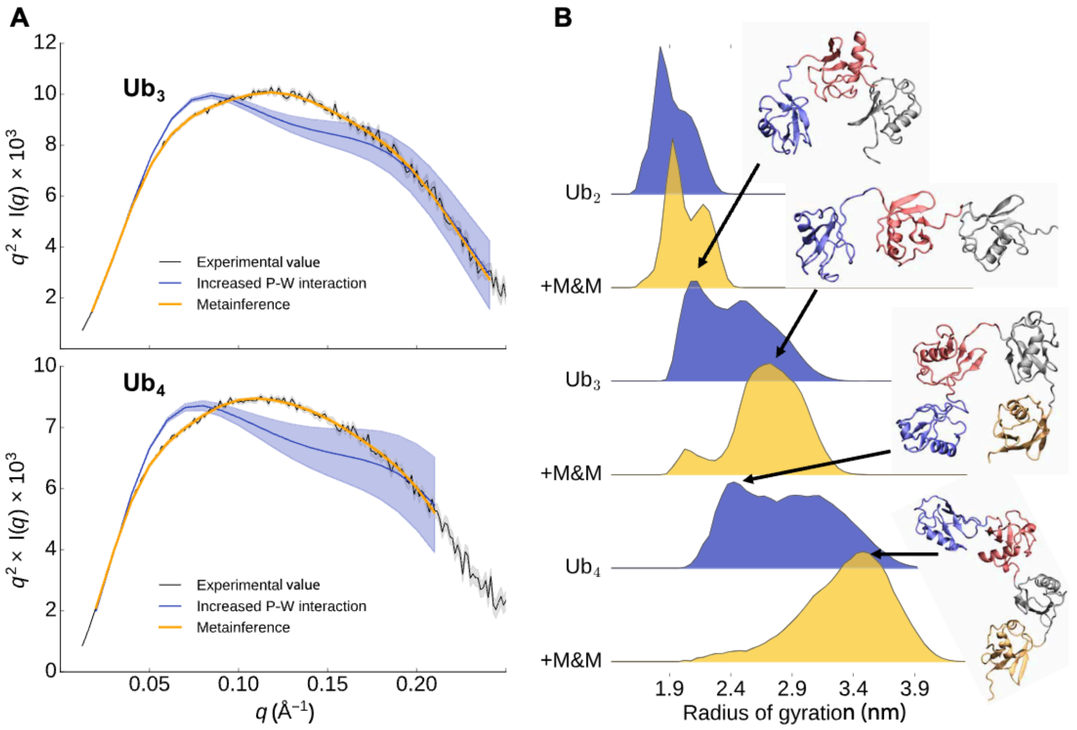
The approach is the same introduced before where the probability p(x) is not anymore the probability of a conformation but the probability distribution of all possible conformations.

Then MD simulations can be used to generate prior conformational ensembles whose probabilities can be modified to better match some available experimental information.



Poly-ubiquitin dynamics



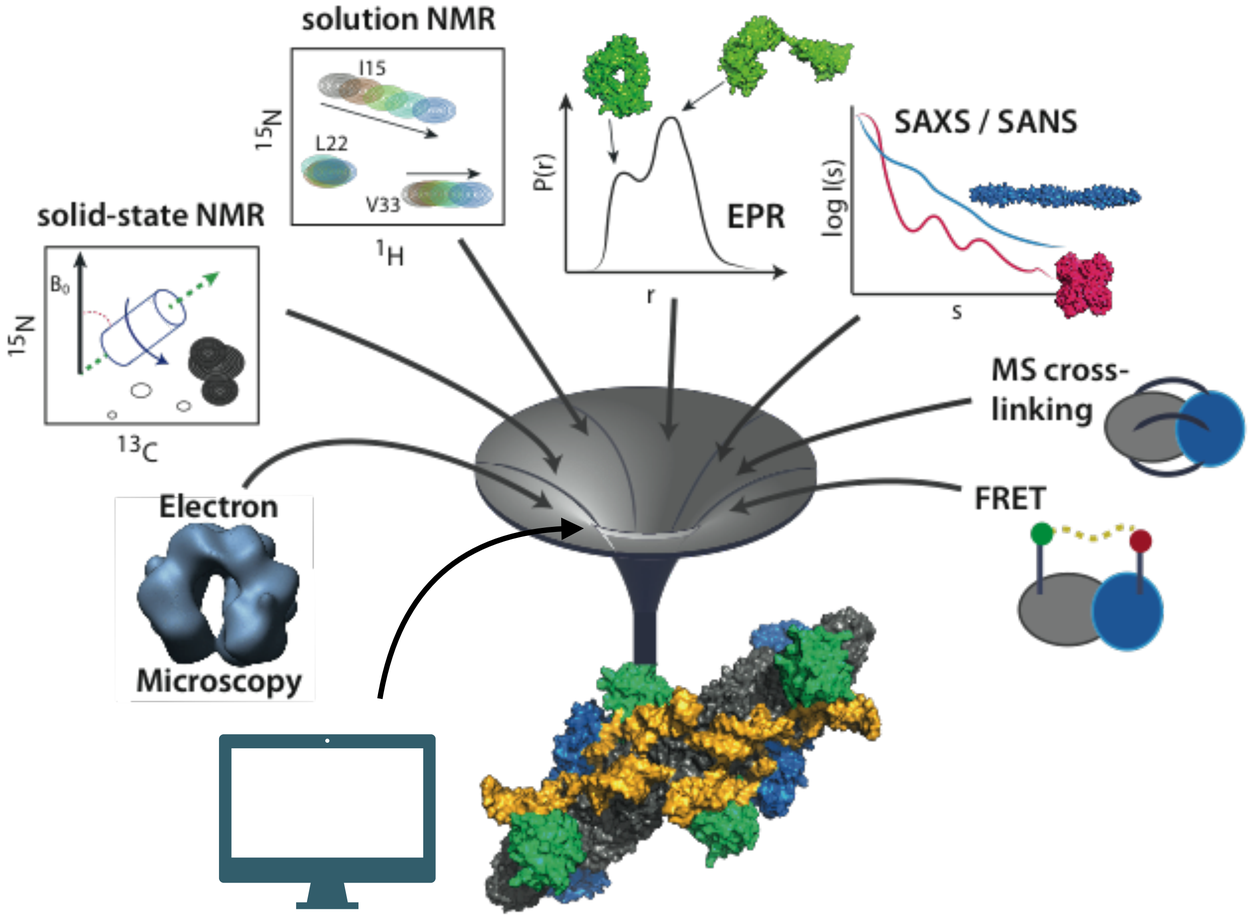


Combining the results of simulations and experiments allows to increase the resolution and the accuracy of both.

Summary

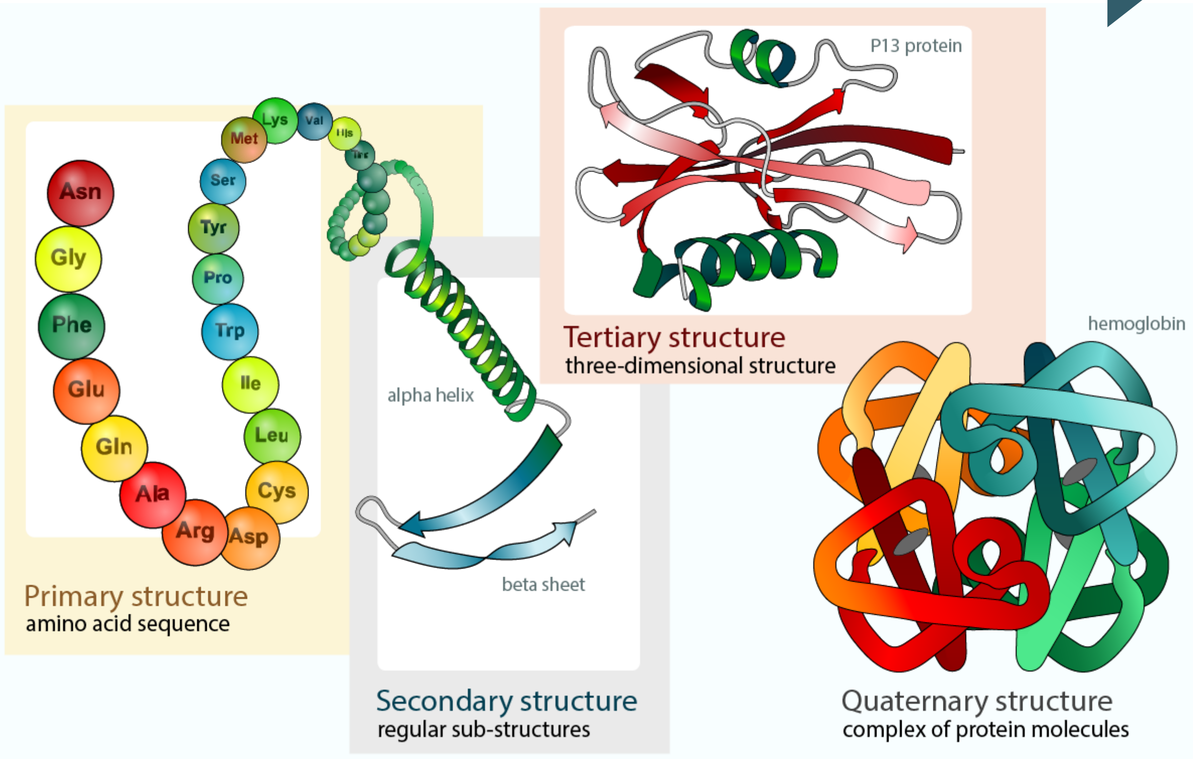
Integrative structural biology approaches allow to optimally use all the available information about a system to generate more accurate structures or simulations

They provide better lenses for our “microscopes”



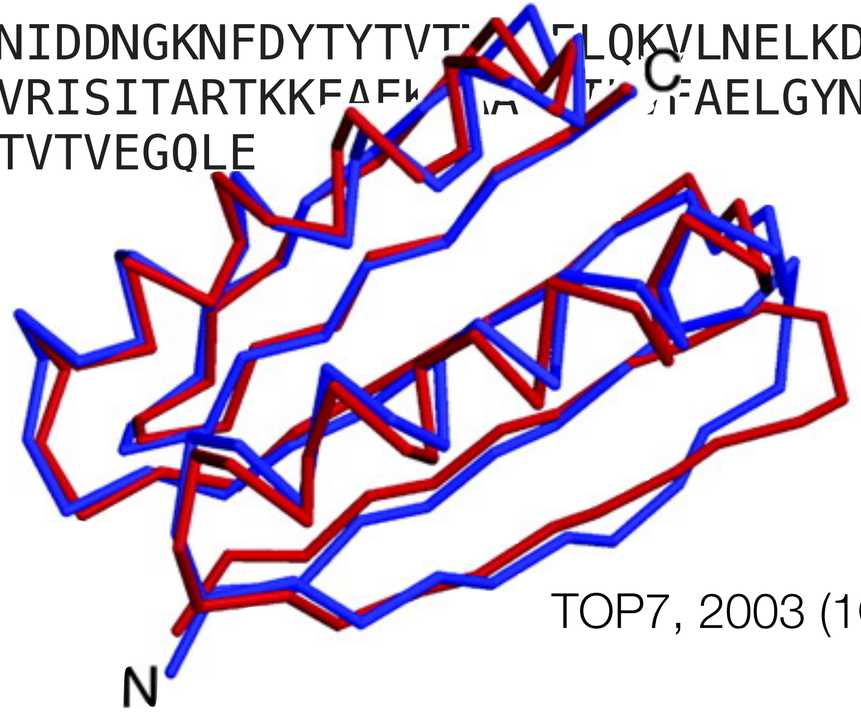
From Protein Structure Prediction to Protein Design

Structure Prediction



Protein Design

DIQVQVNIDDNGKNFDYTYTVT[illegible]LQKVLNELKDYIK
KQGAKRVRISITARTKKEAE[illegible]TFAELGYNDIN
VTFDGDTVTVEGQLE



TOP7, 2003 (1QYS)

Before AI protein design was mostly phys/chem based with a very low success rate (< 1%)

TOP7: the first example

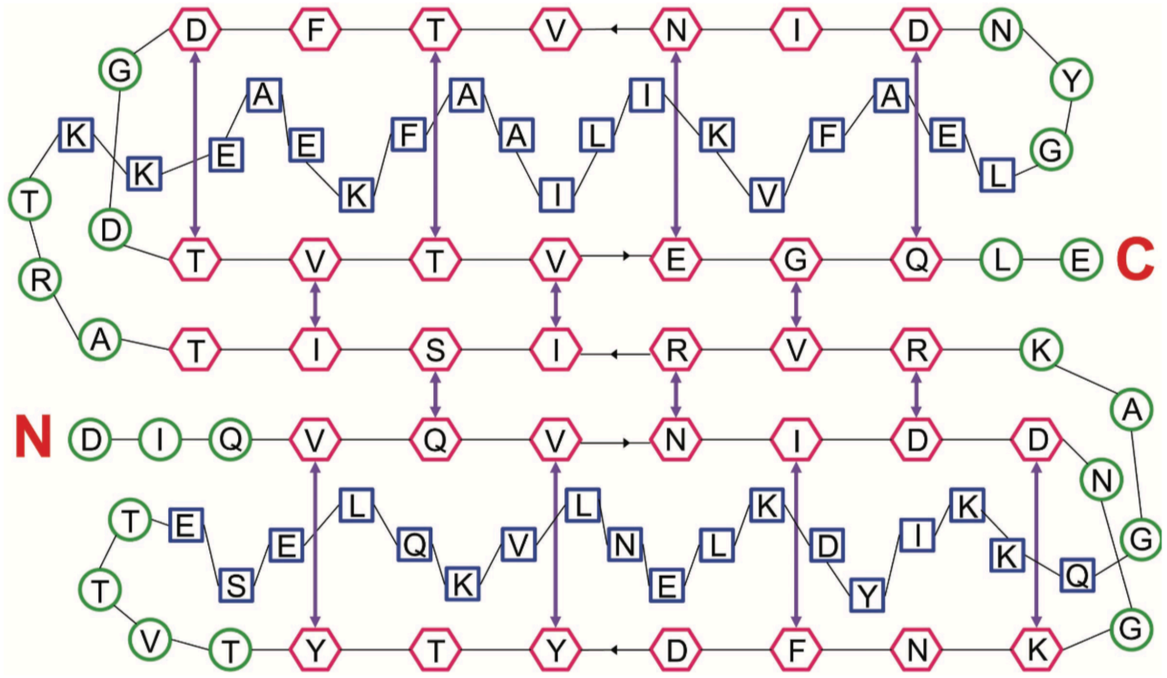


Fig. 1. A two-dimensional schematic of the target fold (hexagon, strand; square, helix; circle, other). Hydrogen bond partners are shown as purple arrows. The amino acids shown are those in the final designed (Top7) sequence.

Segments with the wanted secondary structure are selected from the PDB and assembled with rosetta to generate putative proteins → Side-chains optimisation

None of the resulting structure/sequence has a score comparable to that of natural proteins

1.Kuhlman, B. *et al*. Design of a Novel Globular Protein Fold with Atomic-Level Accuracy. *Science* **302**, 1364–1368 (2003).

TOP7: the first example

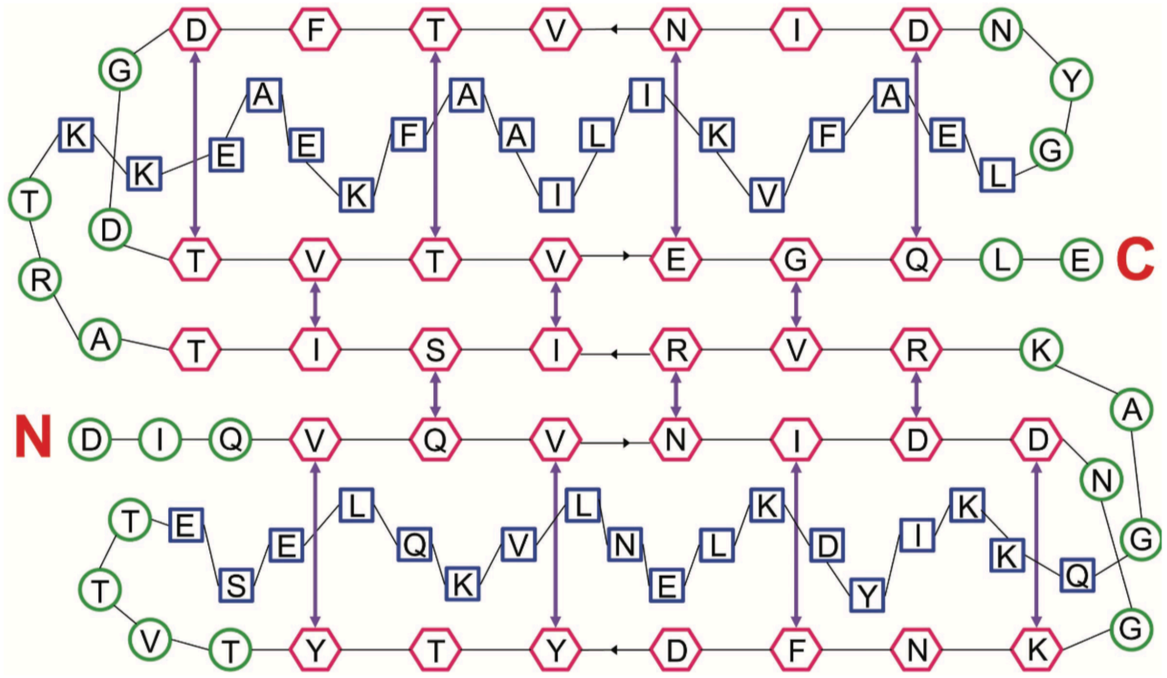
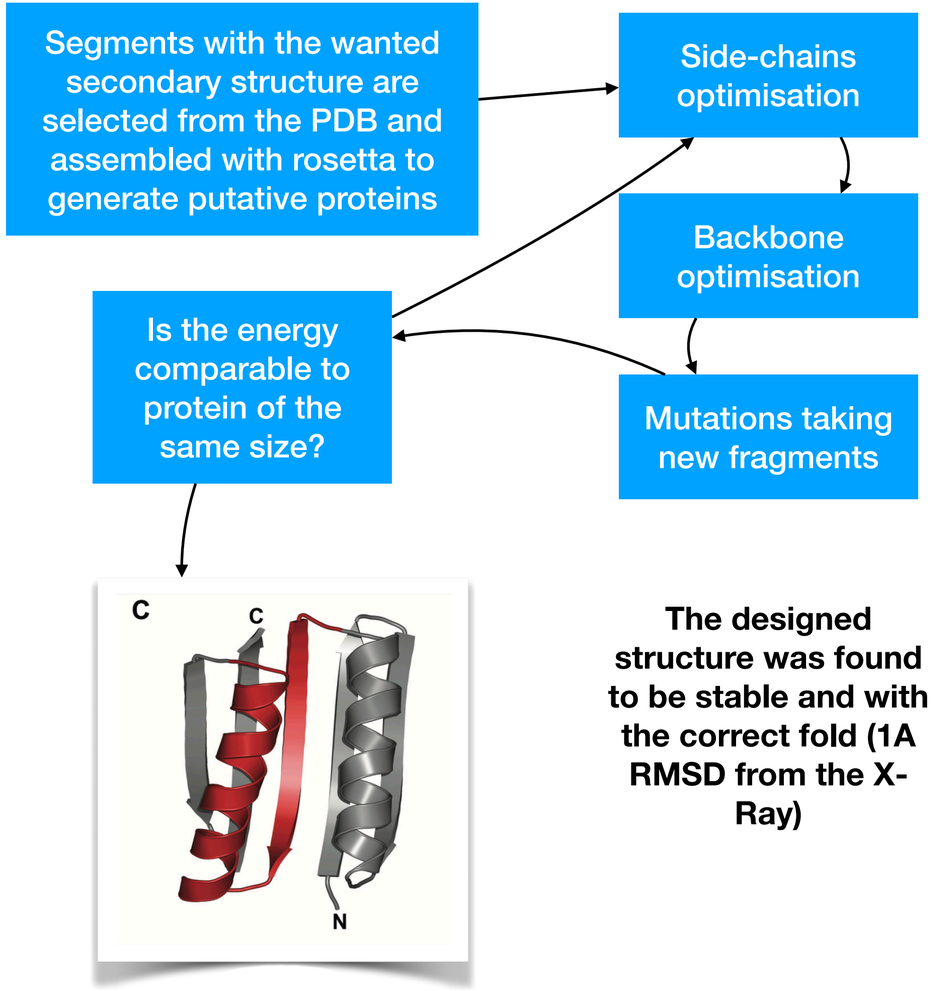


Fig. 1. A two-dimensional schematic of the target fold (hexagon, strand; square, helix; circle, other). Hydrogen bond partners are shown as purple arrows. The amino acids shown are those in the final designed (Top7) sequence.

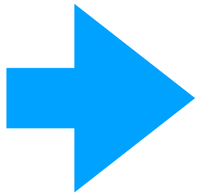


The designed structure was found to be stable and with the correct fold (1A RMSD from the X-Ray)

1.Kuhlman, B. *et al*. Design of a Novel Globular Protein Fold with Atomic-Level Accuracy. *Science* **302**, 1364–1368 (2003).

Protein Design before the advent of AI methods

From a designed structure



To a Protein Sequence

1. **Define your goal (what do you want to design)**
2. **Find suitable restraints (an active site geometry, a binding surface shape, …)**
3. **Search for scaffolds or design new ones that can accomodate your restraints**
4. **Optimise protein sequences/structures**

If you find solution that:

- **Respects the restraints**
- **Is not far from the scaffold (<2-3A)**
- **Has an energy comparable to natural proteins of the same size**

Produce your protein! If it works then improve it by directed evolution.

The success rate of this approach was low <1% of the sequences could be actually successfully expressed.

1.Chevalier, A. *et al*. Massively parallel de novo protein design for targeted therapeutics. *Nature* **550**, 74–79 (2017).

The first designed enzyme

If it is possible to design a protein with a given fold maybe it is also possible to design an enzyme that performs a chemical reaction that is not catalysed by any known protein?

Given the chemical reaction, one could think of a possible active site organisation (in the transition state)

↓

DFT calculations to optimise the geometry

↓

Search the PDB to find scaffold that could accomodate in their active site the same geometry

↓

Overall scaffold optimisation

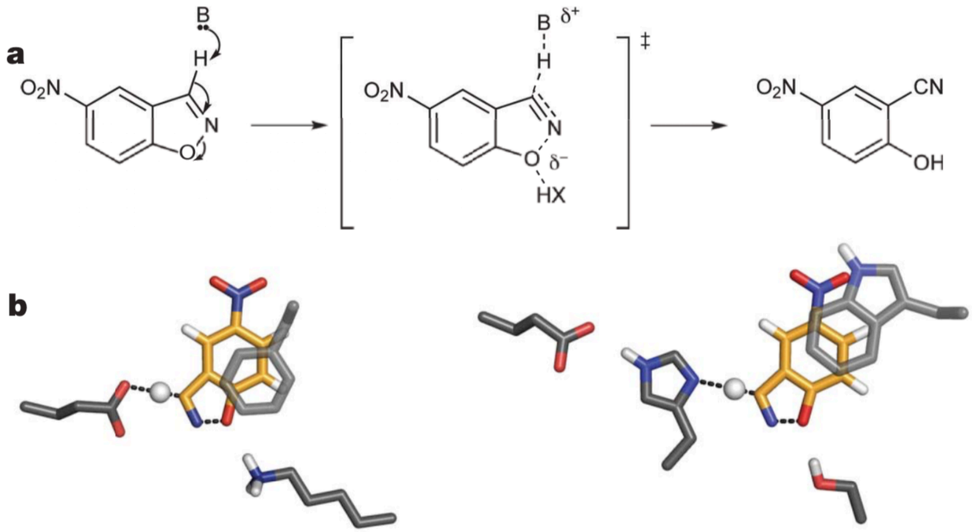


Figure 1 | Reaction scheme and catalytic motifs used in design. **a**, The Kemp elimination proceeds by means of a single transition state, which can be stabilized by a base deprotonating the carbon and the dispersion of the resulting negative charge; a hydrogen bond donor can also be used to stabilize the partial negative charge on the phenolic oxygen. **b**, Examples of active site motifs highlighting the two choices for the catalytic base (a carboxylate (left) or a His–Asp dyad (right)) used for deprotonation, and a π-stacking aromatic residue for transition state stabilization. For each catalytic base, all combinations of hydrogen bond donor groups (Lys, Arg, Ser, Tyr, His, water or none) and π-stacking interactions (Phe, Tyr, Trp) were input as active site motifs into RosettaMatch.

1.Röthlisberger, D. *et al*. Kemp elimination catalysts by computational enzyme design. *Nature* **453**, 190–195 (2008).

The first designed enzyme

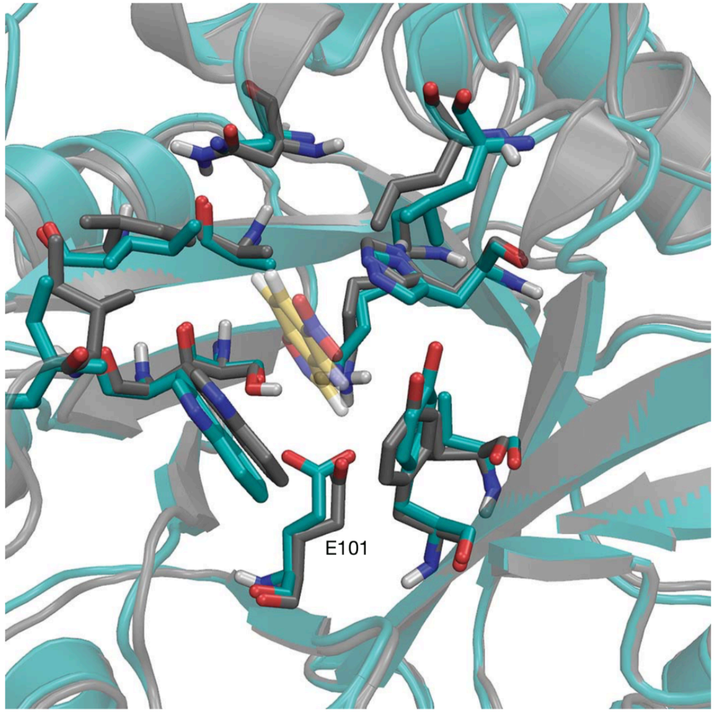


Figure 4 | Comparison of the designed model of KE07 and the crystal structure. The crystal structure (cyan) was solved in the unbound state and shows only modest rearrangement of active site side chains compared to the designed structure (grey) modelled in the presence of the transition state (yellow, transparent). (Backbone r.m.s.d. for the active site is 0.32 Å versus 0.95 Å for the active site including the side chains.) The observed electron density around relevant amino acids in the active site is shown in Supplementary Fig. 6. KE07 contains 13 mutations compared to the starting template scaffold (PDB code 1thf).

Table 2 | Kinetic parameters of KE07 variants

Variant	Mutations	k_{cat} (s^{-1})	K_m, (mM)	k_{cat}/K_m ($M^{-1} s^{-1}$)
KE07 WT	-	0.018 ± 0.001	1.4 ± 0.1	12.2 ± 0.1
R2 11/10D†	K19E Q123R K146T G202R N224D	0.021 ± 0.001	0.31 ± 0.02	66 ± 2
R3 I3/10A	I7Q F86L K146T G202R N224D F229S	0.206 ± 0.003	0.48 ± 0.03	425 ± 16
R3 I3/10A E101A				≤3.9
R4 1E/11H	I7D K146E G202R N224D	0.699 ± 0.001	2.40 ± 0.07	291 ± 9
R4 1E/11H E101A				≤2.4
R5 10/3B	I7D V12M G202R N224D	0.49 ± 0.01	0.59 ± 0.03	836 ± 18
R6 3/7F	I7D K19E K146T G202R N224D	0.60 ± 0.07	0.69 ± 0.09	872 ± 25
R7 2/5B	I7D F77I G202R N224D	1.20 ± 0.08	0.86 ± 0.08	1,388 ± 44
R7 10/11G	I7D V12M F77I I102F K146T G202R N224D F229S	1.37 ± 0.14	0.54 ± 0.12	2,590 ± 302

The enzyme worked as designed, improvements in its efficiency was then obtained using directed-evolution experiments.

1.Röthlisberger, D. *et al*. Kemp elimination catalysts by computational enzyme design. *Nature* **453**, 190–195 (2008).

Protein-Protein interactions

Is it possible to design protein-protein interactions? So design proteins that bind to specific regions of other proteins?

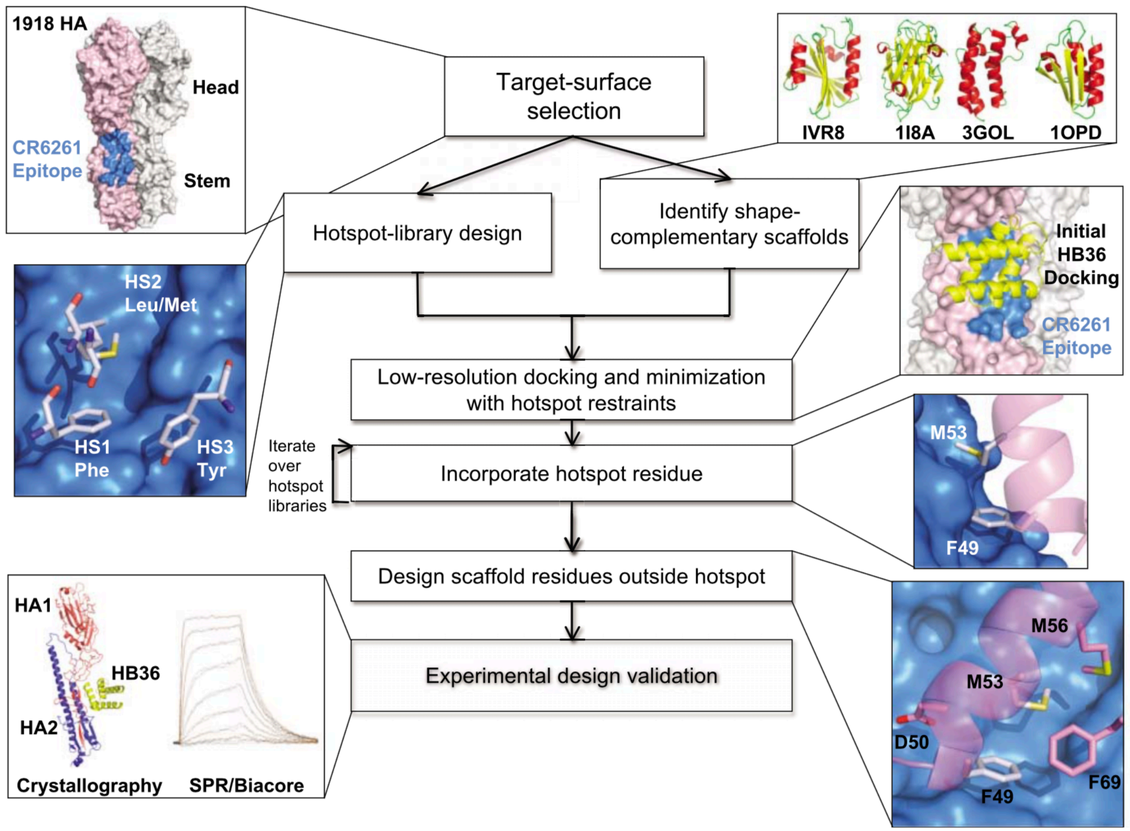


Fig. 1. Flow chart illustrating the key steps in the design of novel binding proteins. The thumbnails illustrate each step in the creation of binders that target the stem of the 1918 HA. Abbreviations (*29*).

Given the target surface, one select specific interactions to provide most of the energy (hotspot, that are salt-bridges, h-bonds, hydrophobic contacts between specific couples of aminoacids). These restraints are then inserted into complementary surfaces found from existing scaffolds. So that the final results is shape complementarity + specific interactions.

UNIVERSITÀ DEGLI STUDI DI MILANO | DBS

1.Fleishman, S. J. *et al*. Computational Design of Proteins Targeting the Conserved Stem Region of Influenza Hemagglutinin. *Science* **332**, 816–821 (2011).

Protein-Protein interactions

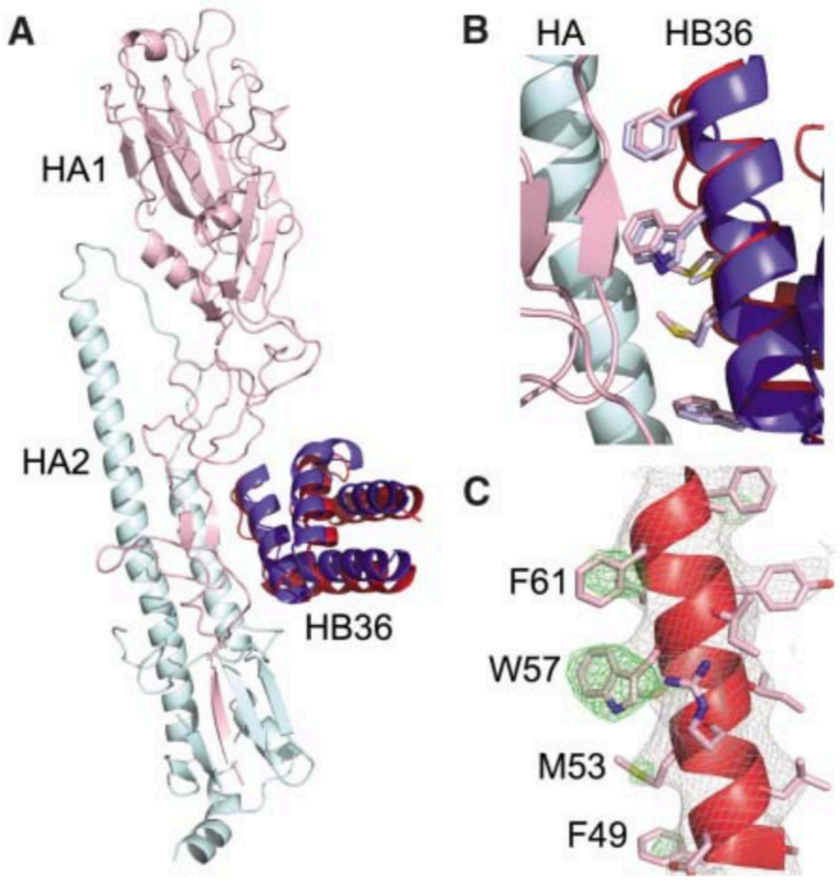


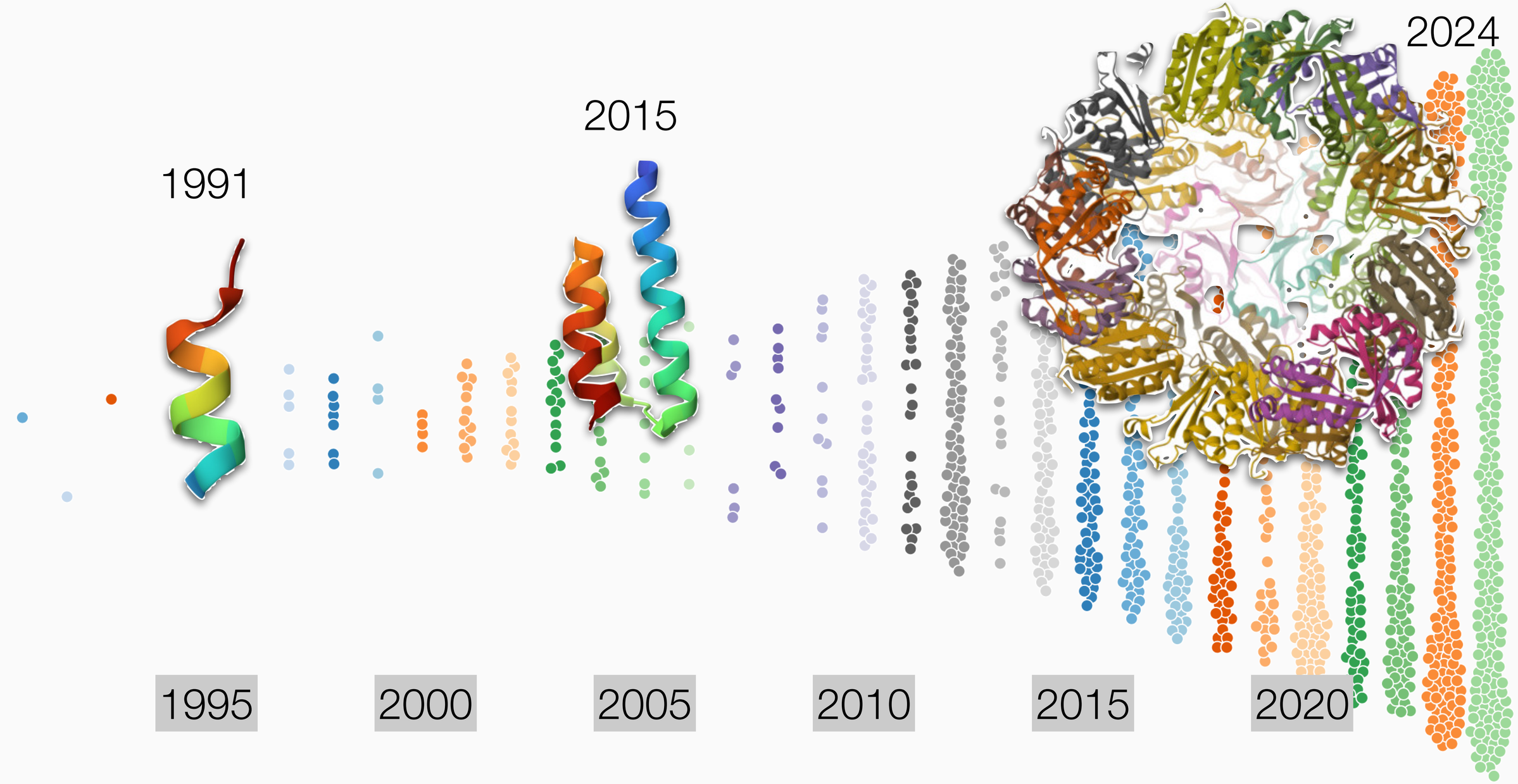
Table 1. Summary of dissociation constants between SC1918/H1 HA and selected design variants. Apparent K_d was determined using yeast surface display titrations. Numbers in parentheses indicate K_d determined by SPR. NB, no binding.

Design	K_d (nM)
1U84 (HB36 scaffold)	NB (NB)
HB36	200 (>2000)
HB36 D47S	5
HA36 A60V	8
HB36.3 (HB36 D47S, A60V)	4 (29)
HB36.4 (HB36 D47S, A60V, N64K)	4 (22)
2CJJ (HB80 scaffold)	NB
HB80	>5000
HB80 M26T	100
HB80 N36K	300
HB80 M26T N36K	7.5
HB80 Δ54-95, M26T, N36K	5
HB80.3 (HB80 Δ54-95, D12Gly, A24S, M26T, N36K)	3 (38)

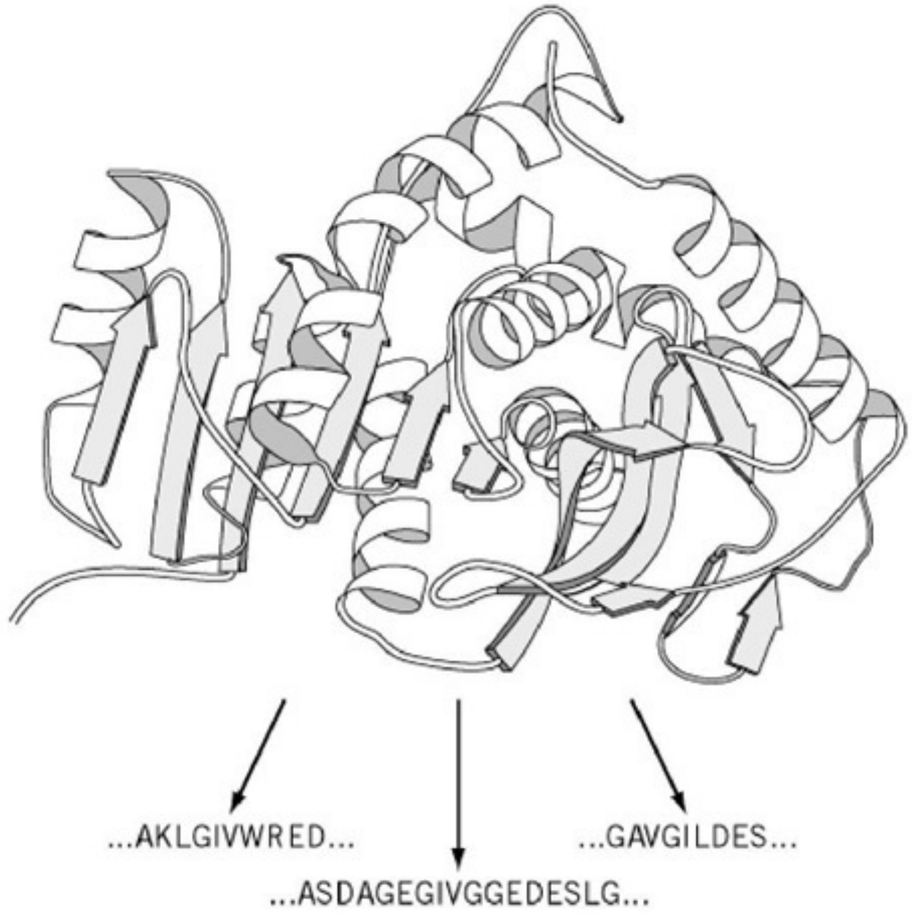
The scaffold chosen doesn't originally bind, after the design there is a good binding that is then optimised by affinity maturation (directed evolution)

1.Fleishman, S. J. *et al*. Computational Design of Proteins Targeting the Conserved Stem Region of Influenza Hemagglutinin. *Science* **332**, 816–821 (2011).

The Protein Design Archive



The inverse folding problem (ProteinMPNN)



Given a structure, how to find a sequence that will fold it. ProteinMPNN is a succesfull attempt, it rebuilds a protein sequence ~50% of correctness.

Improving Protein Expression, Stability, and Function with ProteinMPNN

Kiera H. Sumida, Reyes Núñez-Franco, Indrek Kalvet, Samuel J. Pellock, Basile I. M. Wicky, Lukas F. Milles, Justas Dauparas, Jue Wang, Yakov Kipnis, Noel Jameson, Alex Kang, Joshmyn De La Cruz, Banumathi Sankaran, Asim K. Bera, Gonzalo Jiménez-Osés, and David Baker*

Cite This: *J. Am. Chem. Soc.* 2024, 146, 2054–2061 Read Online

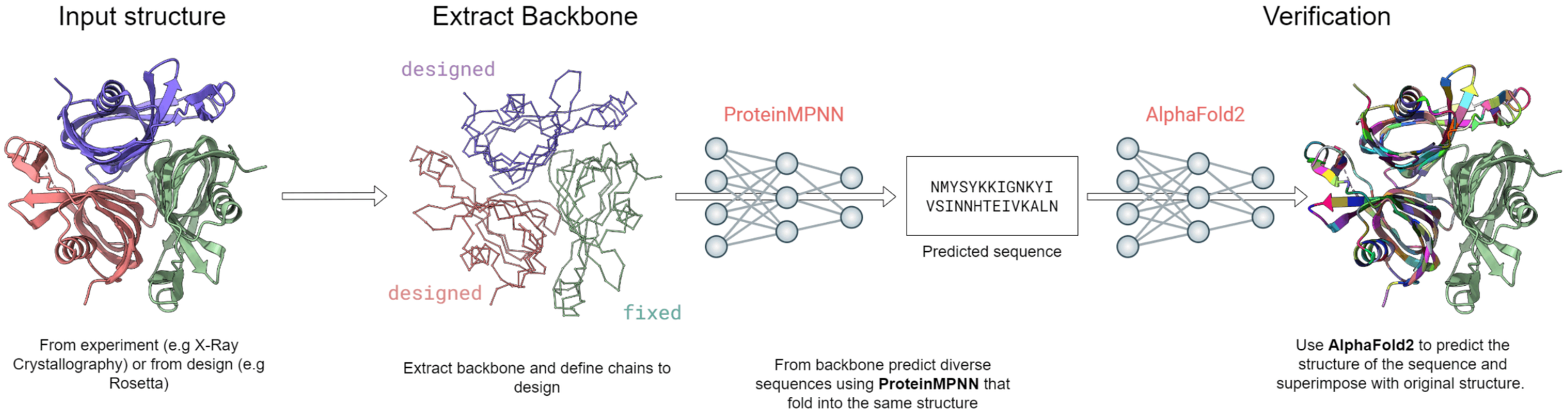
"Your protein, but better!"

This is a key step, shift focus on design folds without sequence.

Predicted sequences scored by AF2

Robust deep learning-based protein sequence design using ProteinMPNN

This model takes as input a protein structure and based on its backbone predicts new sequences that will fold into that backbone. Optionally, we can run AlphaFold2 on the predicted sequence to check whether the predicted sequences adopt the same backbone (WIP).

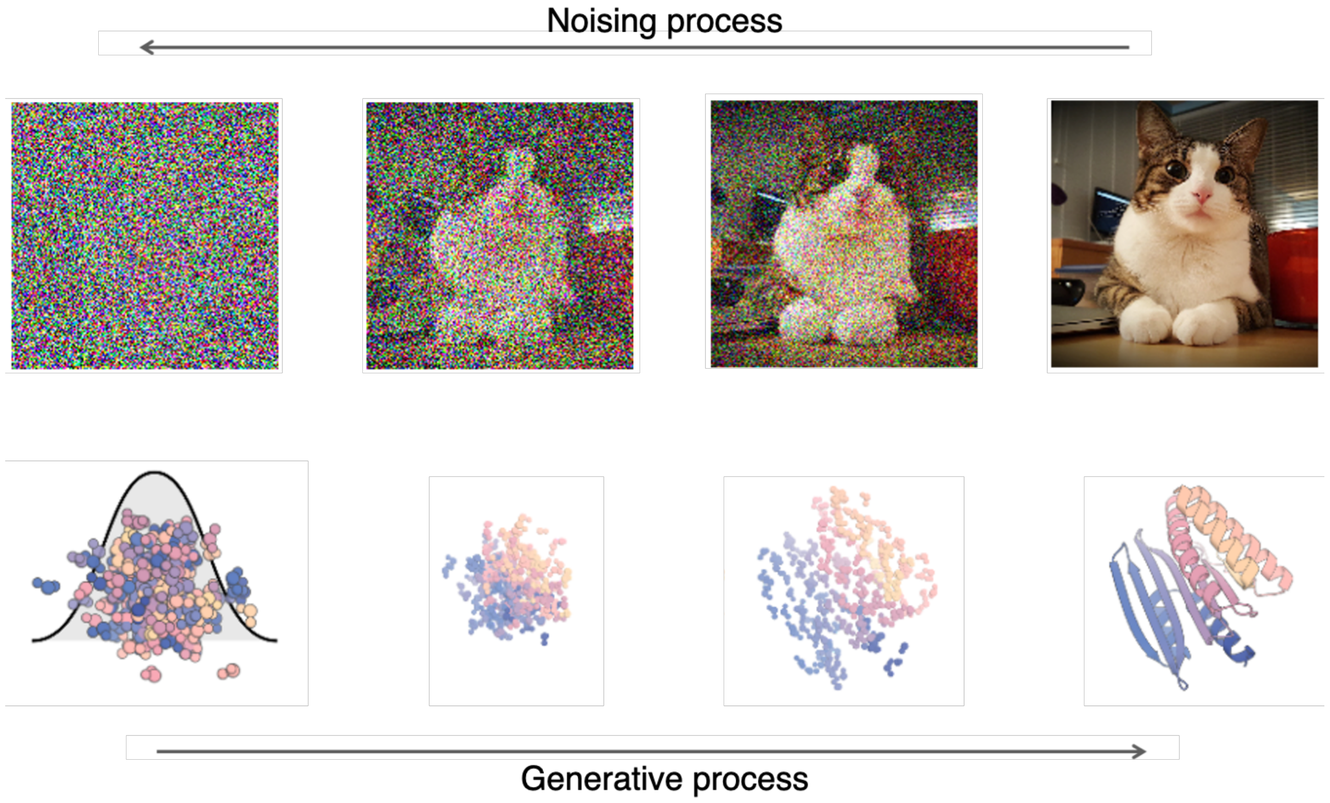


This is a network that has learned how to find an optimal sequence given a protein backbone configuration. On native protein backbones, ProteinMPNN has a sequence recovery of 52.4% compared with 32.9% for Rosetta and it is orders of magnitude faster. Expression success rate > 50%.

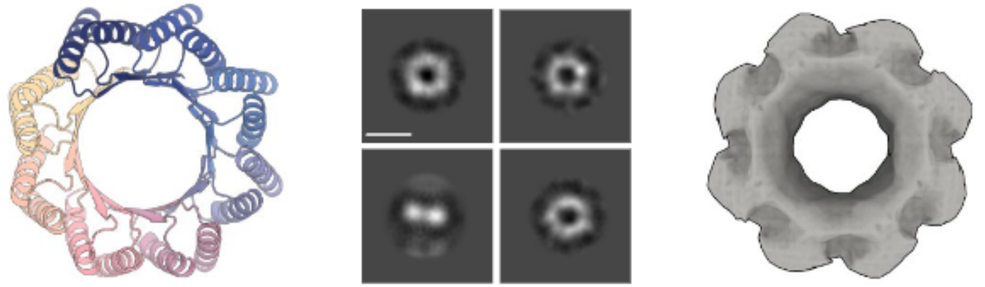
UNIVERSITÀ DEGLI STUDI DI MILANO | DBS

1.Wang, J. *et al*. Scaffolding protein functional sites using deep learning. *Science* **377**, 387–394 (2022).

But how to design ad hoc folds? 1. Diffusion models (RFdiffusion)

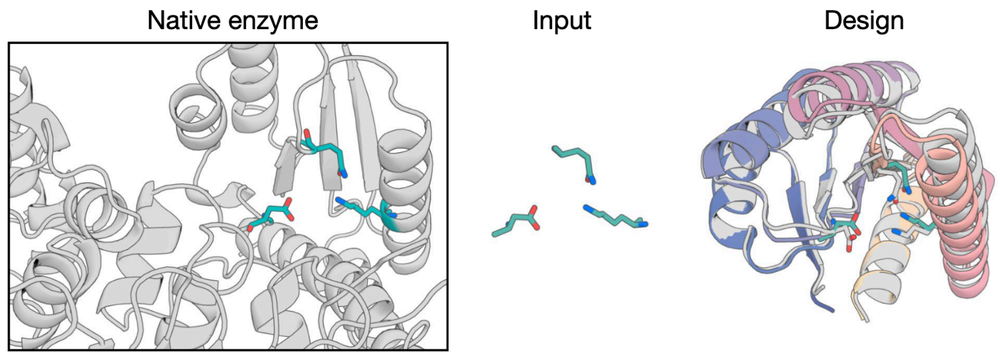


Nanocages (inputs are sequence length and olig symmetry)



Enzymes (inputs are sequence length and active site residues)

Oxidoreductase (EC1)



success rate 5–30%

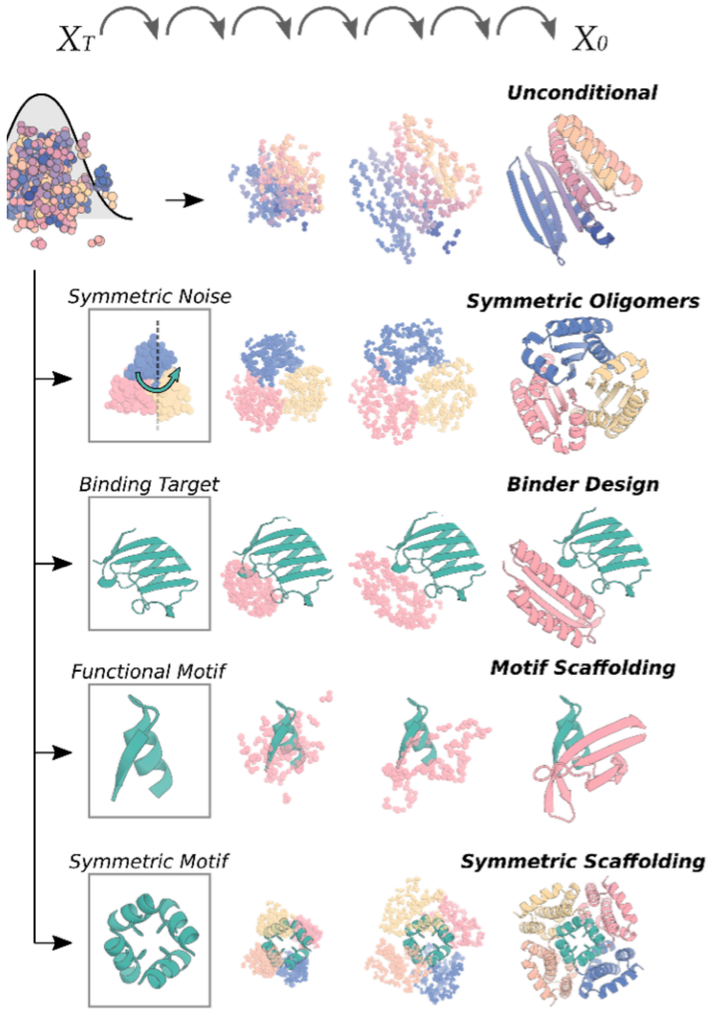
RFdiffusion: generation of protein structures

The idea of the method is to train to RosettaFold to fix errors in protein structure, that is a protein structure is made noisy by adding random errors to the positions of the atoms and the network is trained to recover the correct structure.

RFdiffusion start from random position of the backbone and iteratively assemble a backbone structure. This can also happen under some constraint like the presence of specific motif taken from another PDB structure or the interaction with the surface of another protein or other input feature.

Once you have the backbone proteinMPNN is used to recover the sequence and AF2 to determine the full structure.

The process is stochastic so if you repeat the process you can always get new combinations of sequences and structures.



2. Montecarlo sampling with some reasonable score

De novo protein design by deep network hallucination

https://doi.org/10.1038/s41586-021-04184-w

Received: 18 September 2020

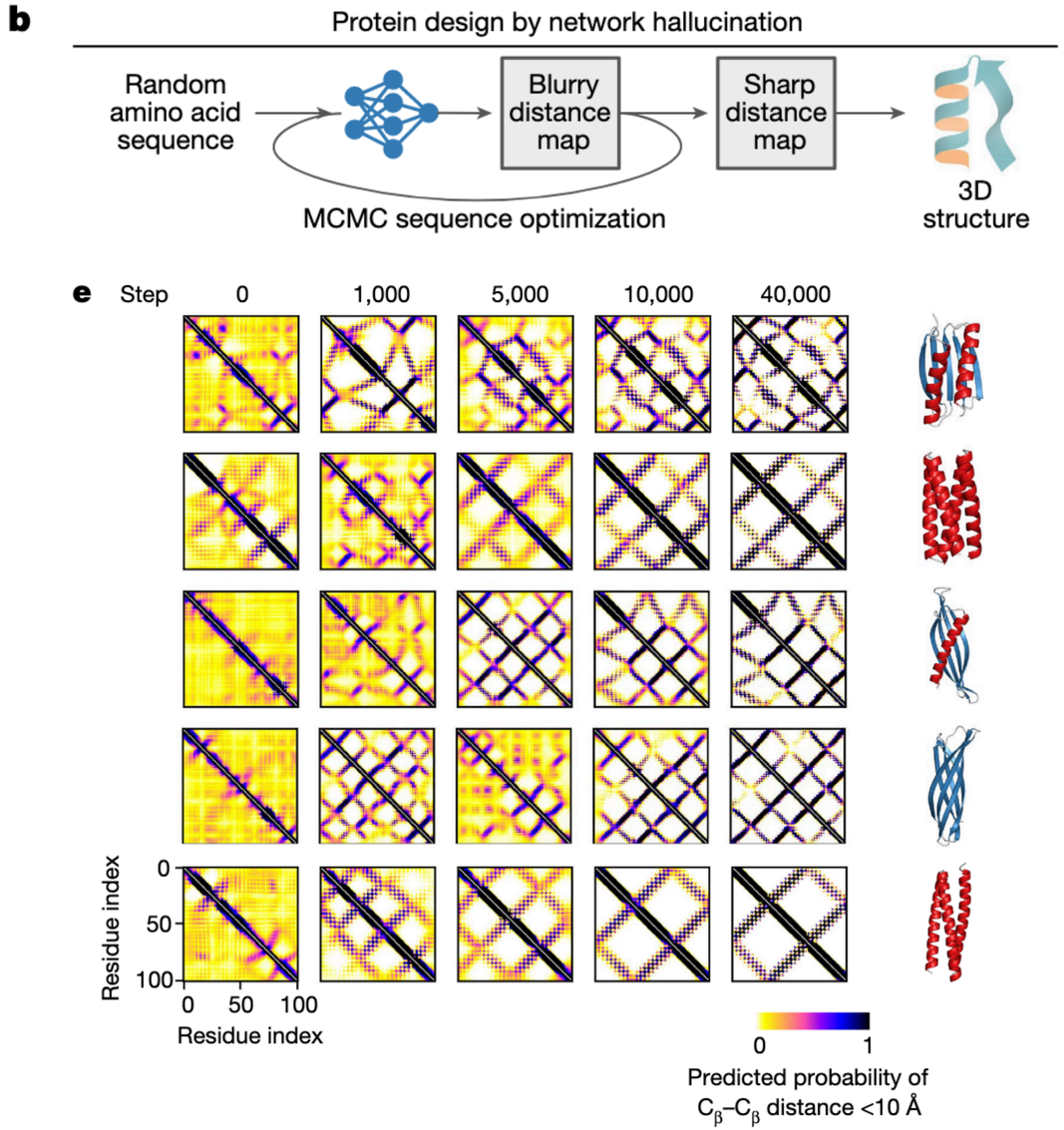
Accepted: 21 October 2021

Published online: 01 December 2021

Check for updates

Ivan Anishchenko[1,2,7], Samuel J. Pellock[1,2,7], Tamuka M. Chidyausiku[1,2], Theresa A. Ramelot[3,4], Sergey Ovchinnikov[5], Jingzhou Hao[3,4], Khushboo Bafna[3,4], Christoffer Norn[1,2], Alex Kang[1,2], Asim K. Bera[1,2], Frank DiMaio[1,2], Lauren Carter[1,2], Cameron M. Chow[1,2], Gaetano T. Montelione[3,4] & David Baker[1,2,6]✉

There has been considerable recent progress in protein structure prediction using deep neural networks to predict inter-residue distances from amino acid sequences[1–3]. Here we investigate whether the information captured by such networks is sufficiently rich to generate new folded proteins with sequences unrelated to those of the naturally occurring proteins used in training the models. We generate random amino acid sequences, and input them into the trRosetta structure prediction network to predict starting residue–residue distance maps, which, as expected, are quite featureless. We then carry out Monte Carlo sampling in amino acid sequence space, optimizing the contrast (Kullback–Leibler divergence) between the inter-residue distance distributions predicted by the network and background distributions averaged over all proteins. Optimization from different random starting points resulted in novel proteins spanning a wide range of sequences and predicted structures. We obtained synthetic genes encoding 129 of the network-'hallucinated' sequences, and expressed and purified the proteins in *Escherichia coli*; 27 of the proteins yielded monodisperse species with circular dichroism spectra consistent with the hallucinated structures. We determined the three-dimensional structures of three of the hallucinated proteins, two by X-ray crystallography and one by NMR, and these closely matched the hallucinated models. Thus, deep networks trained to predict native protein structures from their sequences can be inverted to design new proteins, and such networks and methods should contribute alongside traditional physics-based models to the de novo design of proteins with new functions.

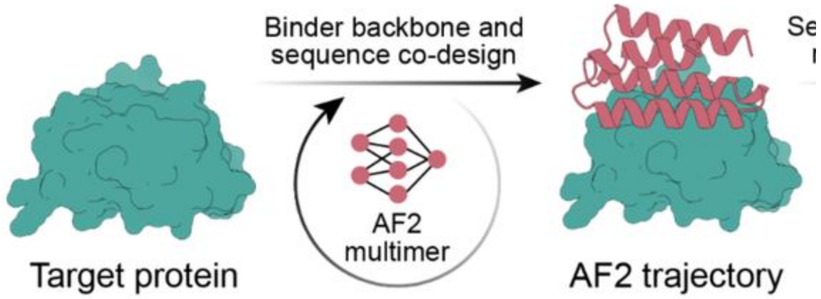


UNIVERSITÀ DEGLI STUDI DI MILANO

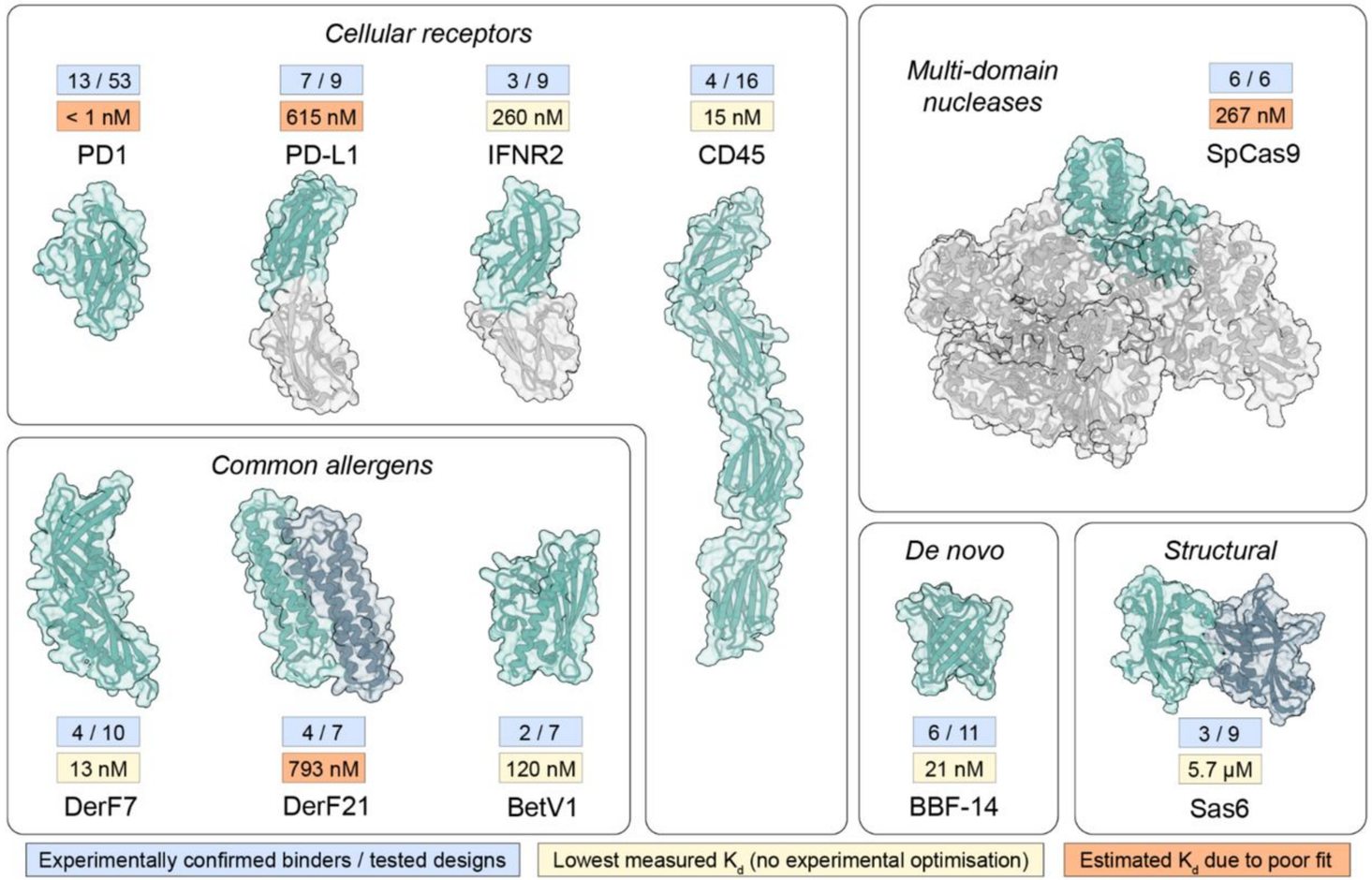
DBS

Success rate ~20%

3. Backpropagate a structure prediction network (BindCraft)



A random sequence and the target sequence are predicted together and then the sequence to be designed is modified following the gradient of the per residue AF2 confidence.



success rate 10–100%

Protein design competitions:

Adaptyv Bio

Design binders to the EGFR extracellular receptor.

First round:

- **700 sequence submited**
- 202 tested experimentally
- 147 sucessfully produced
- **11 succesfully binding < 100 µM**
- 5 with K_D bewetween 0.1-100 nM

Second round:

- **>1000 sequence submited**
- 202 tested experimentally
-

Timeline

Submission deadline: November 4, 2024 at 11:59 pm PT.

Results announcement: December 4, 2024

More challanges ongoing: The BioML Challenge 2024, Rosetta Winter challange, …

UNIVERSITÀ DEGLI STUDI DI MILANO

DBS

Questions:

- How does Bayesian statistics allow to integrate experiments and computational models?
- How can you improve the results of a MD simulations?
- What are the different aspects of protein design?
- What is the inverse folding problem?
- What input should you provide to make an enzyme?
- What input should you provide to design a protein binder?
- What are the current approches to protein design?
- …