Guiding diffusion models for antibody sequence and structure co-design with developability properties

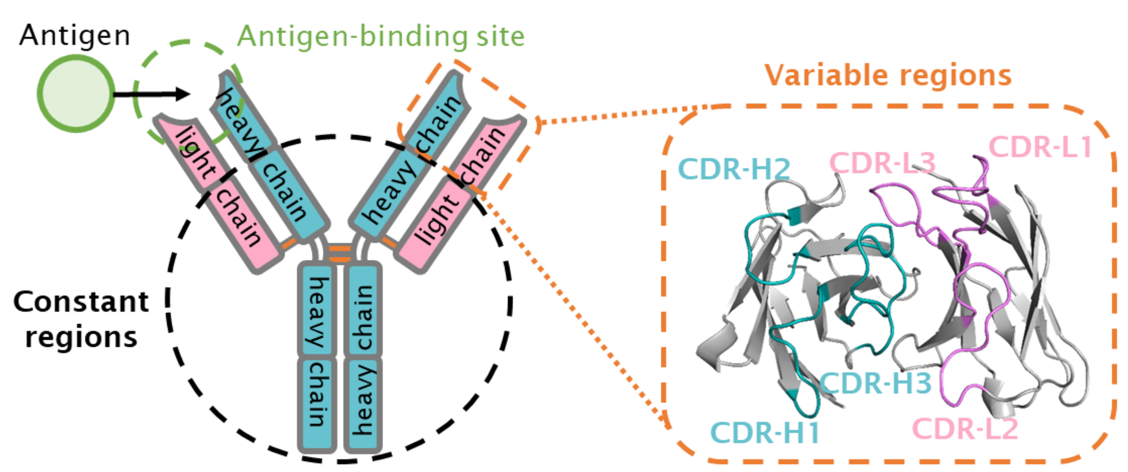
Amelia Villegas-Morcillo, Jana M. Weber, and Marcel J.T. Reinders

Pattern Recognition & Bioinformatics, Dept. of Intelligent Systems, EEMCS, TU Delft, Netherlands

Contact: a.o.villegasmorcillo@tudelft.nl

Antibody structure and composition

Antibodies are Y-shaped proteins produced by the immune system in response to pathogens called **antigens**, composed of two **heavy** and two **light chains**, with a **constant** and **variable region**. The antigen-binding site includes six **complementarity-determining regions (CDRs)** denoted as {H1, H2, H3, L1, L2, L3}. CDRs are **highly variable** domains (especially **CDR-H3**) and determine the **specificity** of an antibody for a particular antigen.



Developing therapeutic antibodies

Antibody engineering

- Antigen-targeting properties
  - Binding affinity
  - Epitope analysis
  - Target specificity
- *Developability* properties
  - Solubility (aggregation)
  - Thermal stability
  - Immunogenicity

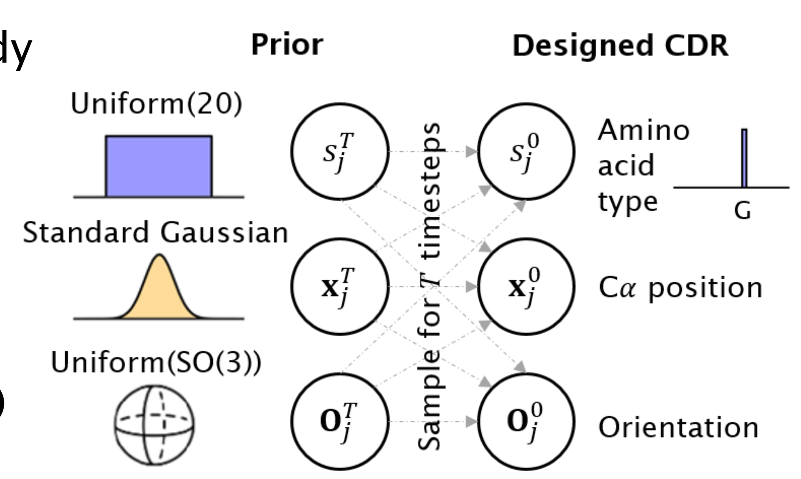
Developability properties are needed to **ensure manufacturability and clinical use. Poor developability profiles** may prevent an antibody from becoming a therapeutic.

Diffusion model for antibody design

Prior **DiffAb** model for antibody sequence-structure co-design

Inputs for amino acid j:

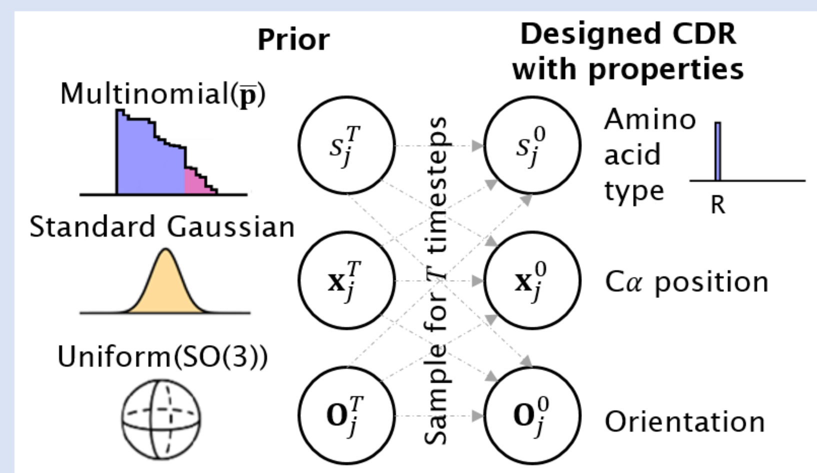
- Amino acid types (s_j)
- C_α atom positions ($\mathbf{x}_j$)
- Amino acid orientations ($\mathbf{O}_j$)

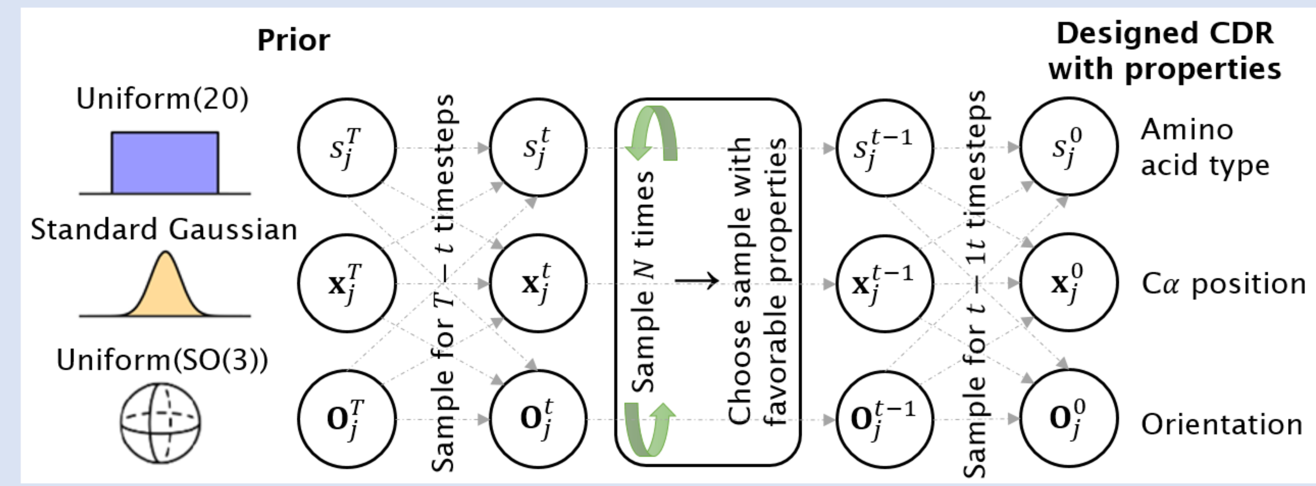


Generate **one CDR loop**, $R = \{(s_j, \mathbf{x}_j, \mathbf{O}_j) \mid j = l+1, \dots, l+m\}$, given the **antibody-antigen complex**, $C = \{(s_i, \mathbf{x}_i, \mathbf{O}_i) \mid i \neq j\}$.

$T = 100$ timesteps of generation

Antibody design guided on properties



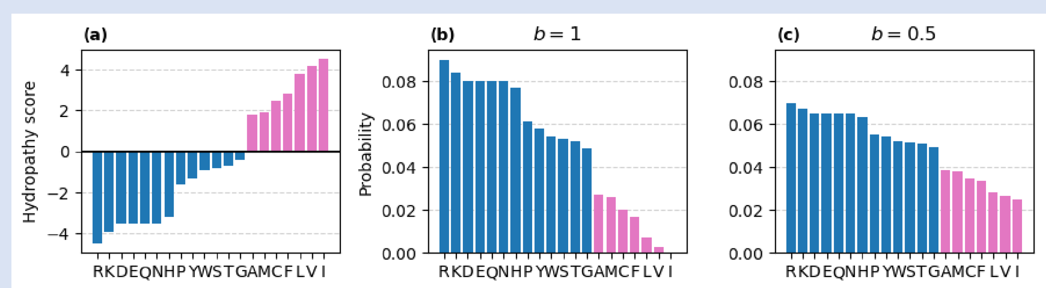


Property-aware prior

Sample from prior: $s_j^T \sim \text{Multinomial}(\overline{\mathbf{p}})$

$= (1-b) \cdot \text{Uniform}(20) + b \cdot \text{Multinomial}(\mathbf{p})$

Property: **Hydropathy** (proxy for solubility)



Sampling by property

At each generation timestep:

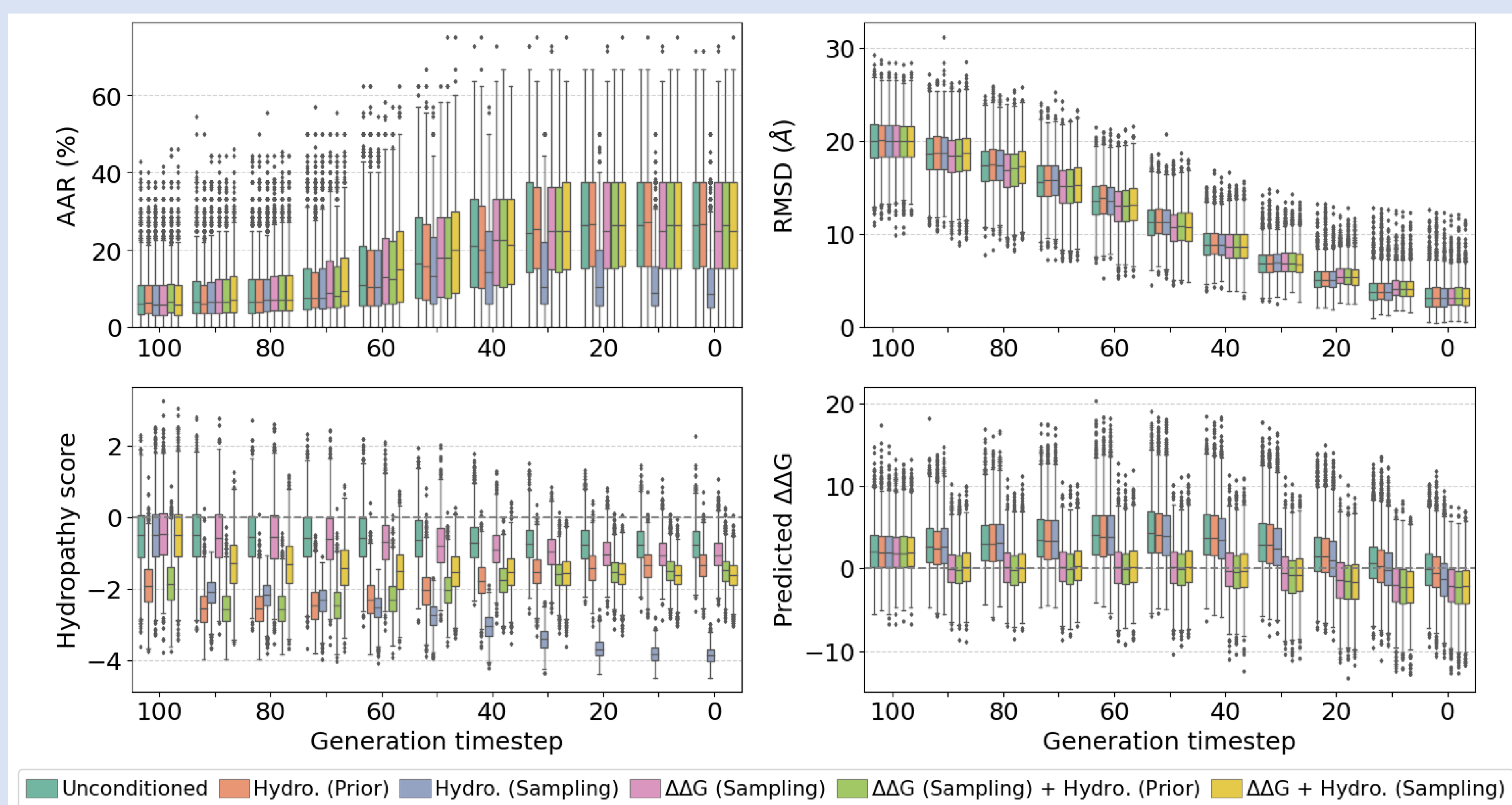
1. Sample N times
2. Select the sample with min. property value

With multiple properties, select Pareto optimal

Properties: **Hydropathy** and **folding energy** (predicted ΔΔG)

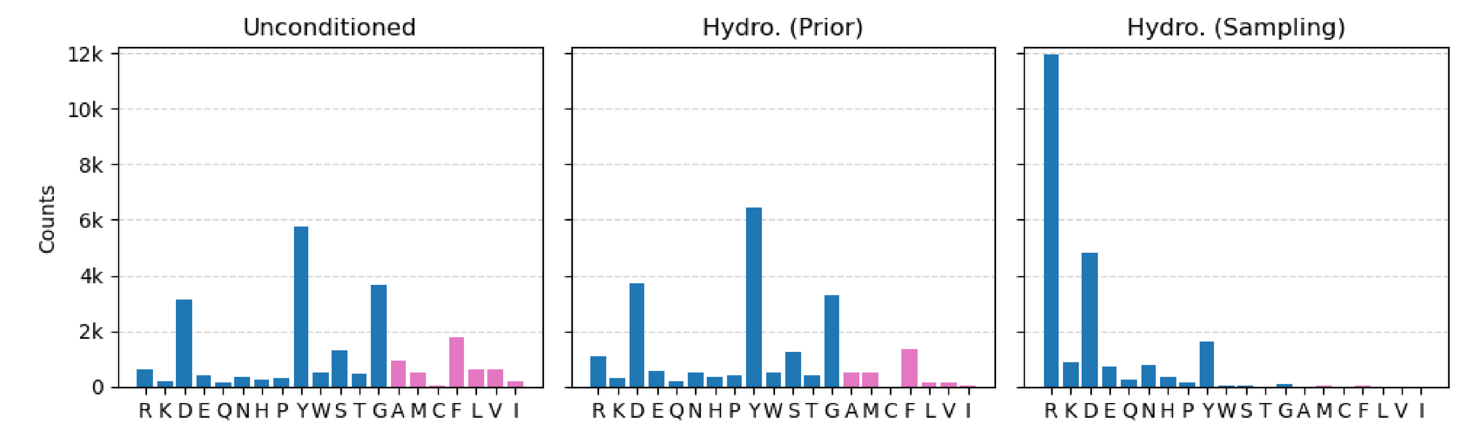
No retraining required!

Guidance on properties is effective



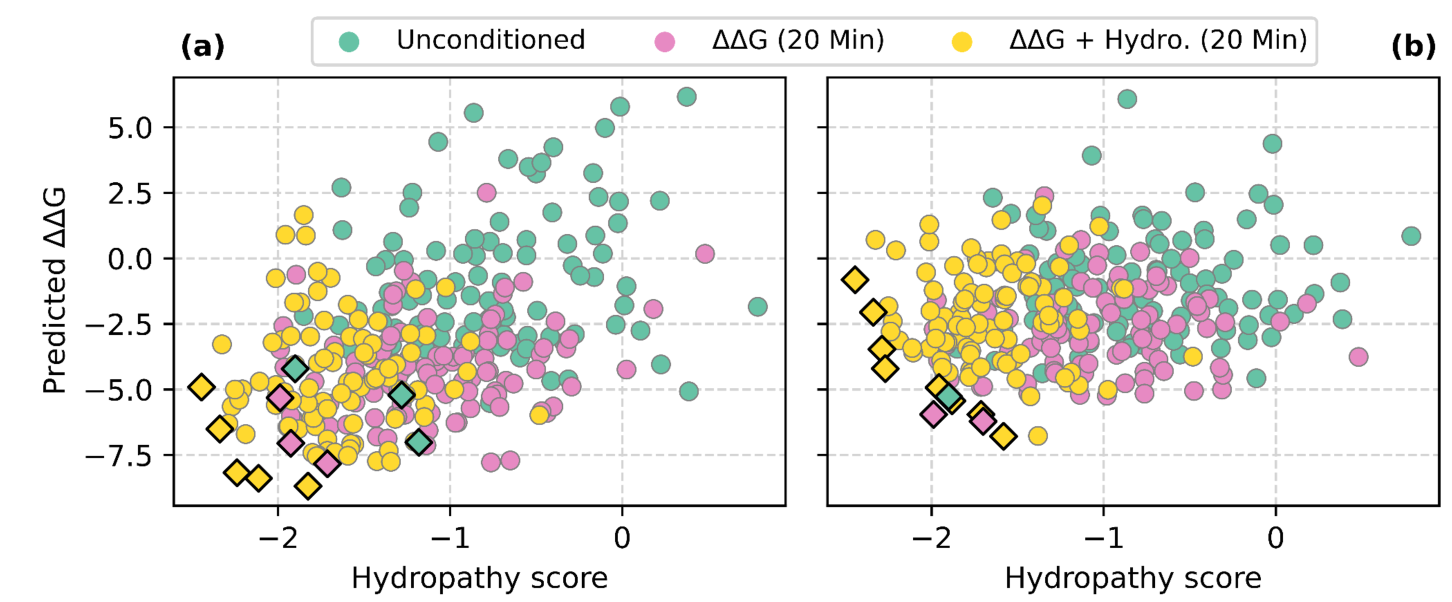
- Sampling by ΔΔG improves the hydropathy score, sampling by hydropathy improves the pred. ΔΔG.
- Most favorable outcomes are achieved when **both properties are combined**.
- **AAR** (amino acid recovery) **and RMSD** (root mean square deviation) **consistent** with unconditioned.

Changes in amino acid composition

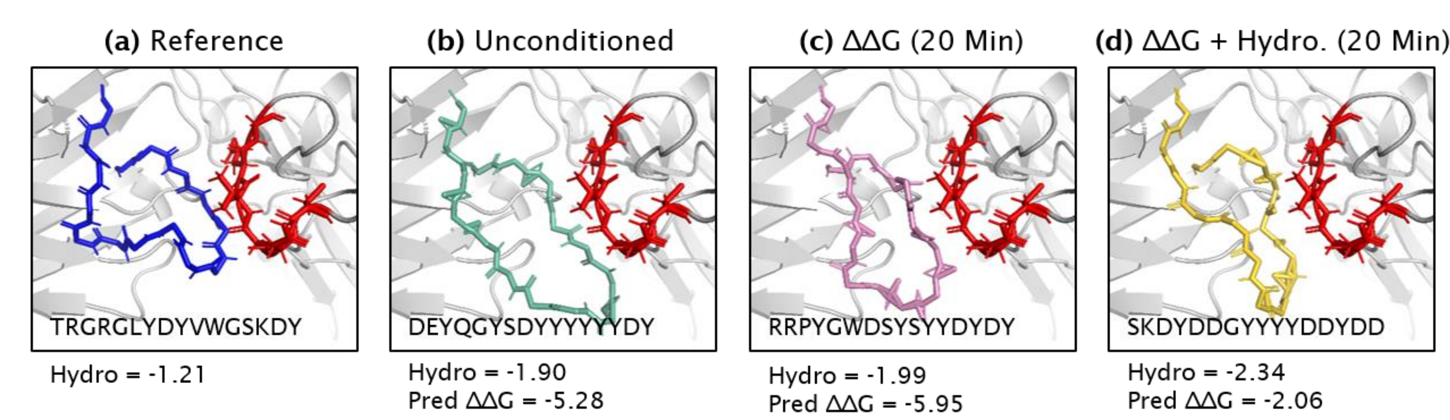


Pareto optimal solutions

Guided approaches exhibit a trend towards the **lowest values of hydropathy and predicted ΔΔG**, (a) before and (b) after Rosetta relaxation.



For designs in the Pareto frontier, different CDR sequences lead to **similar structures** compared to the reference, but with **improved hydropathy and predicted ΔΔG.**



References

- J. Kim, *et al.* "Computational and artificial intelligence-based methods for antibody development." *Trends in Pharmacological Sciences*, 2023.
- S. Luo, *et al.* "Antigen-specific antibody design and optimization with diffusion-based generative models." In *NeurIPS* 2022.
- J. Kyte and R.F. Doolittle. "A simple method for displaying the hydropathic character of a protein." *Journal of Molecular Biology*, 1982.
- S. Shan, *et al.* "Deep learning guided optimization of human antibody against SARS-CoV-2 variants with broad neutralization." *Proceedings of the National Academy of Sciences*, 2022.
- R.F. Alford, *et al.* "The rosetta all-atom energy function for macromolecular modeling and design." *Journal of chemical theory and computation*, 2017.