Generative Models of protein sequences

Marion Chauveau 2nd year Phd student

Supervisors: Ivan Junier & Olivier Rivoire

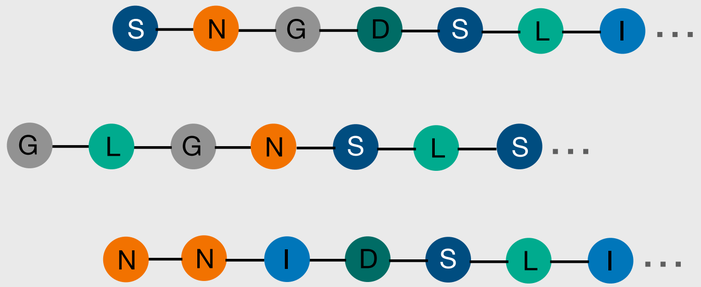
Collaborator: Yaakov Kleeorin

11/2023

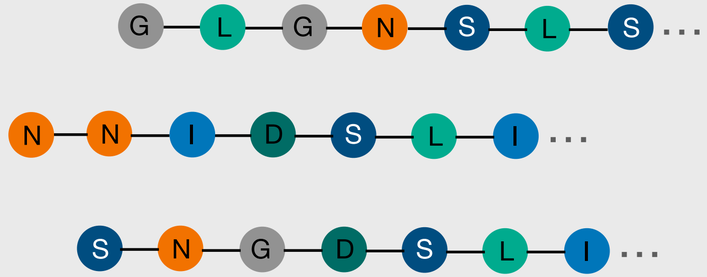
5th course on Multi-scale Integration in Biological Systems

Generative Models of Protein sequences

Natural protein sequences



Training data $\sim P_{data}$

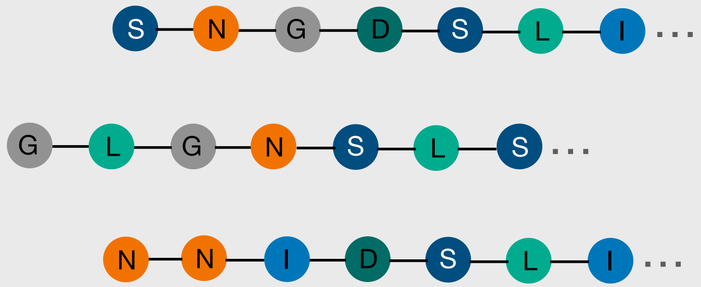


Artificial data $\sim P_{model}$

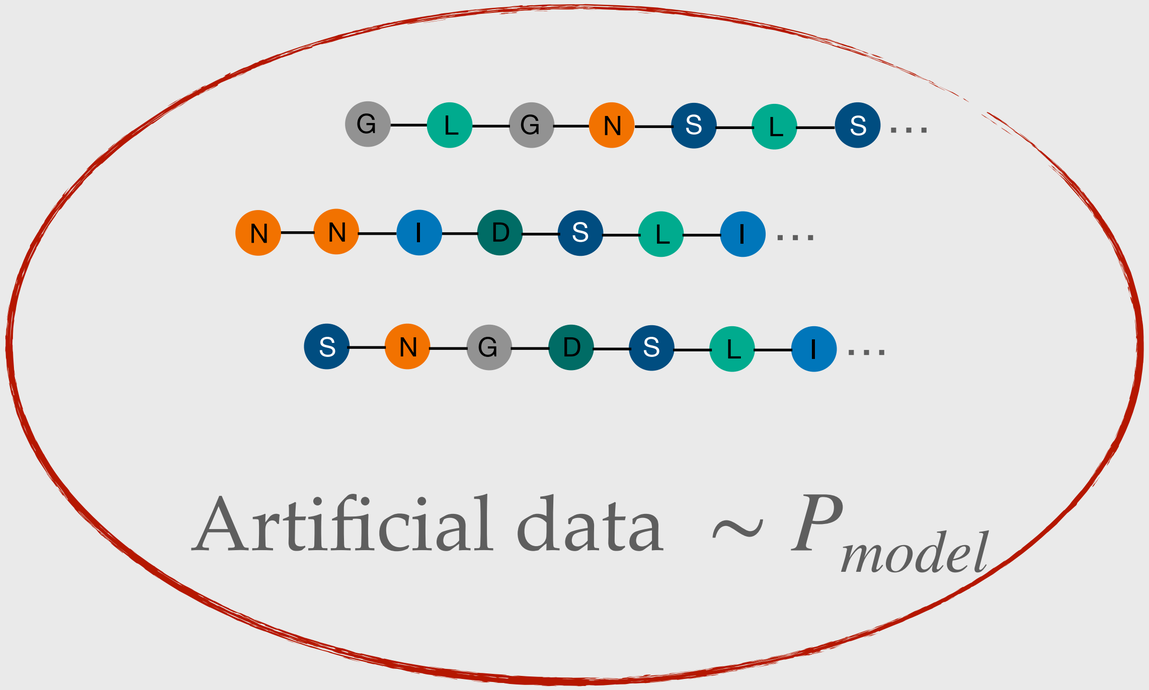
1. Learn P_{model} similar to P_{data}
2. Generate samples from P_{model}

Generative Models of Protein sequences

Natural protein sequences



Training data $\sim P_{data}$



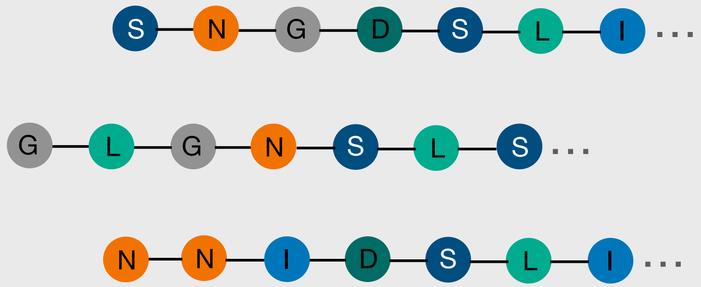
Artificial data $\sim P_{model}$

Good samples ?

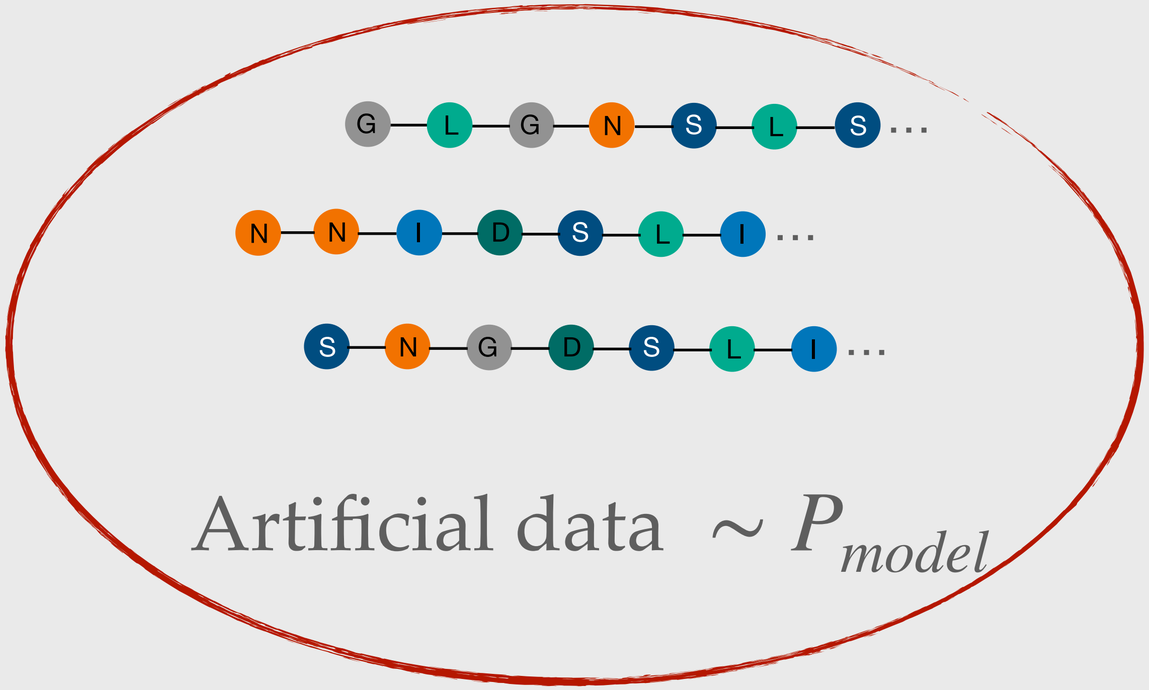
1. Learn P_{model} similar to P_{data}
2. Generate samples from P_{model}

Generative Models of Protein sequences

Natural protein sequences



Training data $\sim P_{data}$



Artificial data $\sim P_{model}$

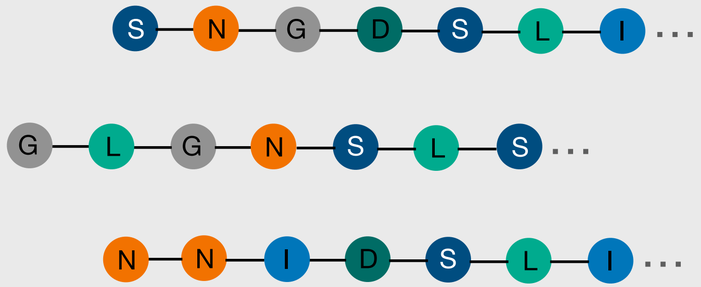
Good samples ?

- experiments

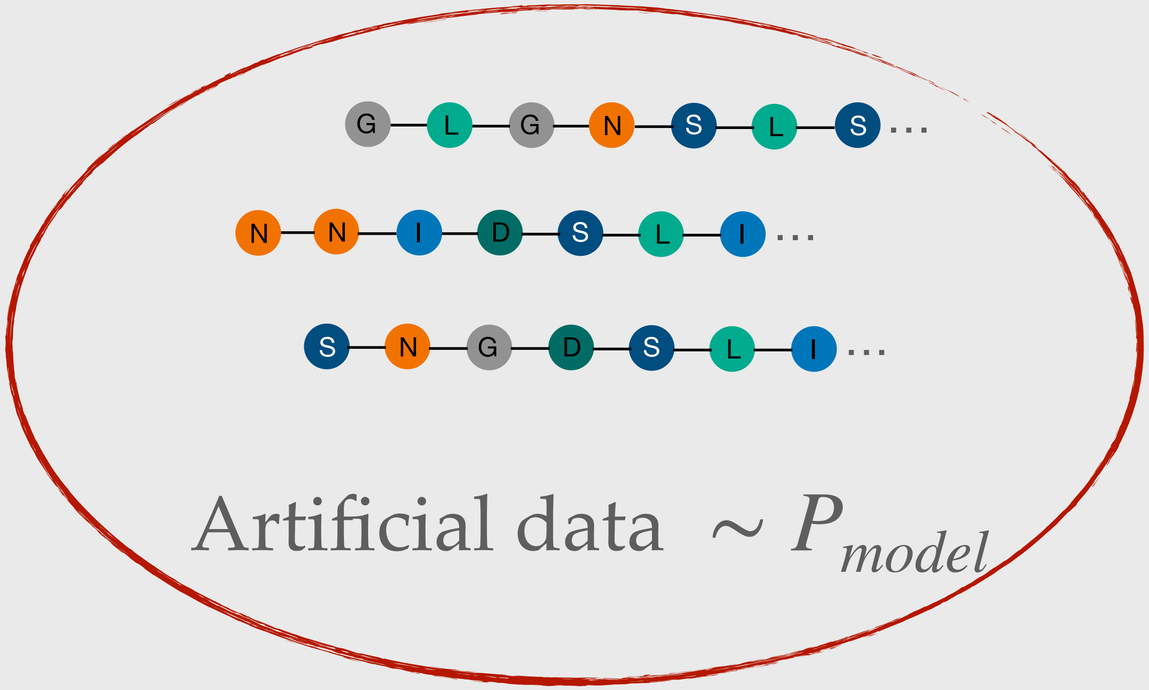
1. Learn P_{model} similar to P_{data}
2. Generate samples from P_{model}

Generative Models of Protein sequences

Natural protein sequences



Training data $\sim P_{data}$



Artificial data $\sim P_{model}$

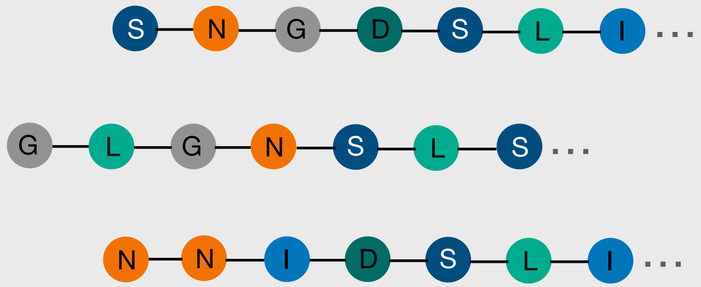
Good samples ?

- experiments
- Ability to capture fundamental properties of data

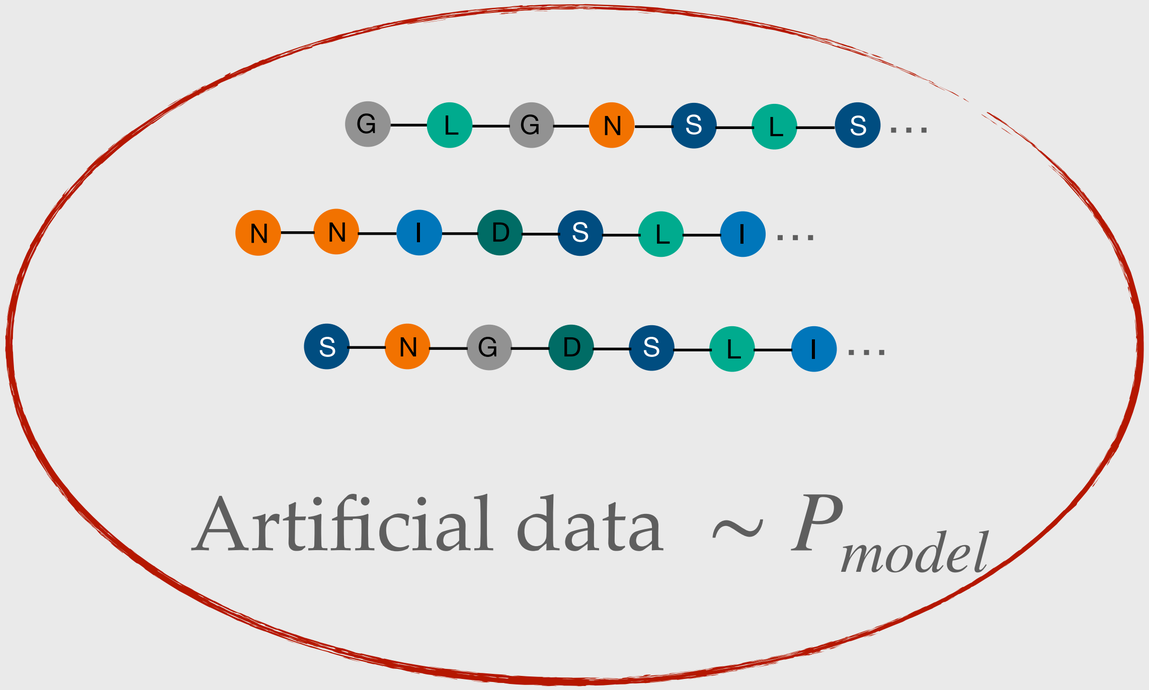
1. Learn P_{model} similar to P_{data}
2. Generate samples from P_{model}

Generative Models of Protein sequences

Natural protein sequences



Training data $\sim P_{data}$



Artificial data $\sim P_{model}$

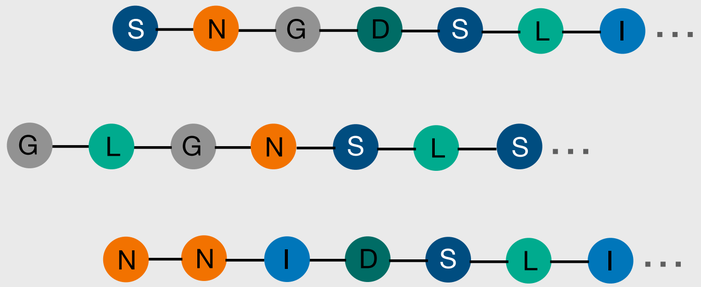
Good samples ?

- experiments
- Ability to capture fundamental properties of data
- Toy Models

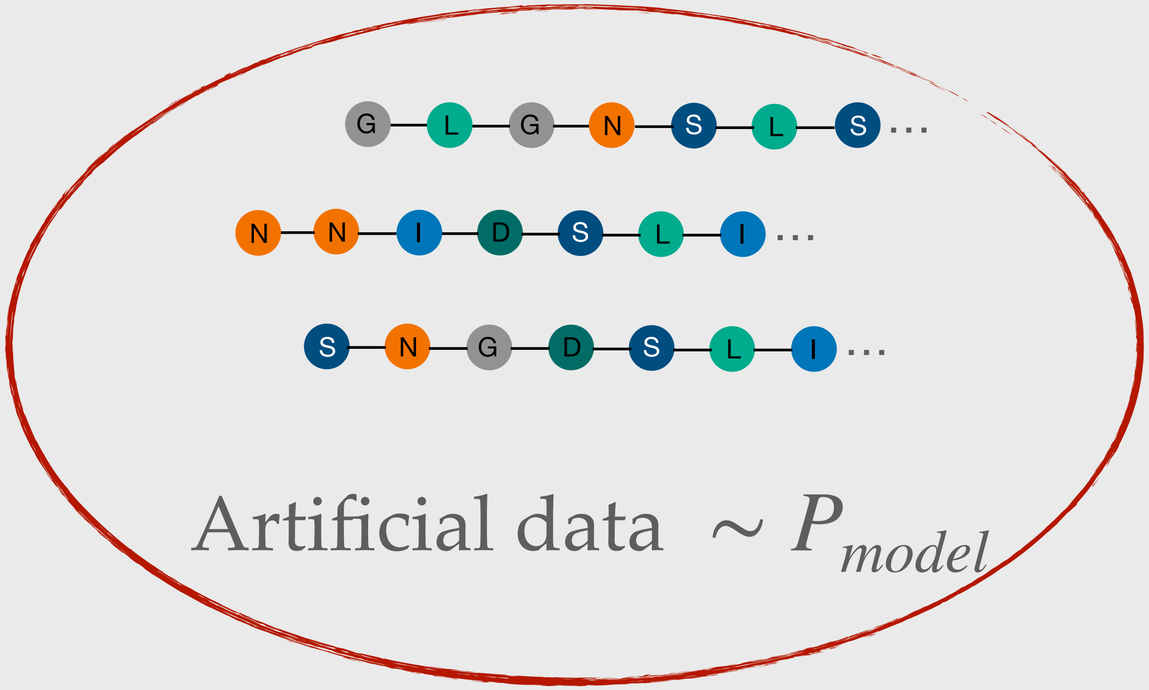
1. Learn P_{model} similar to P_{data}
2. Generate samples from P_{model}

Generative Models of Protein sequences

Natural protein sequences



Training data $\sim P_{data}$



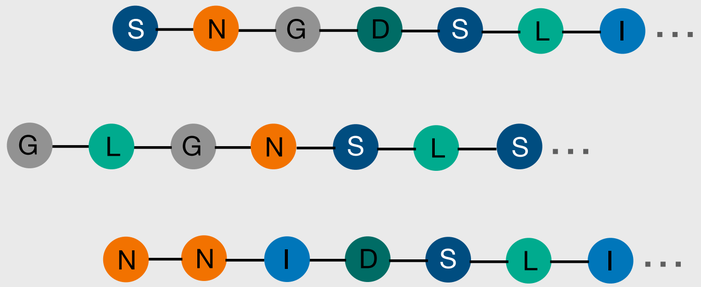
Artificial data $\sim P_{model}$

Different from training data ?

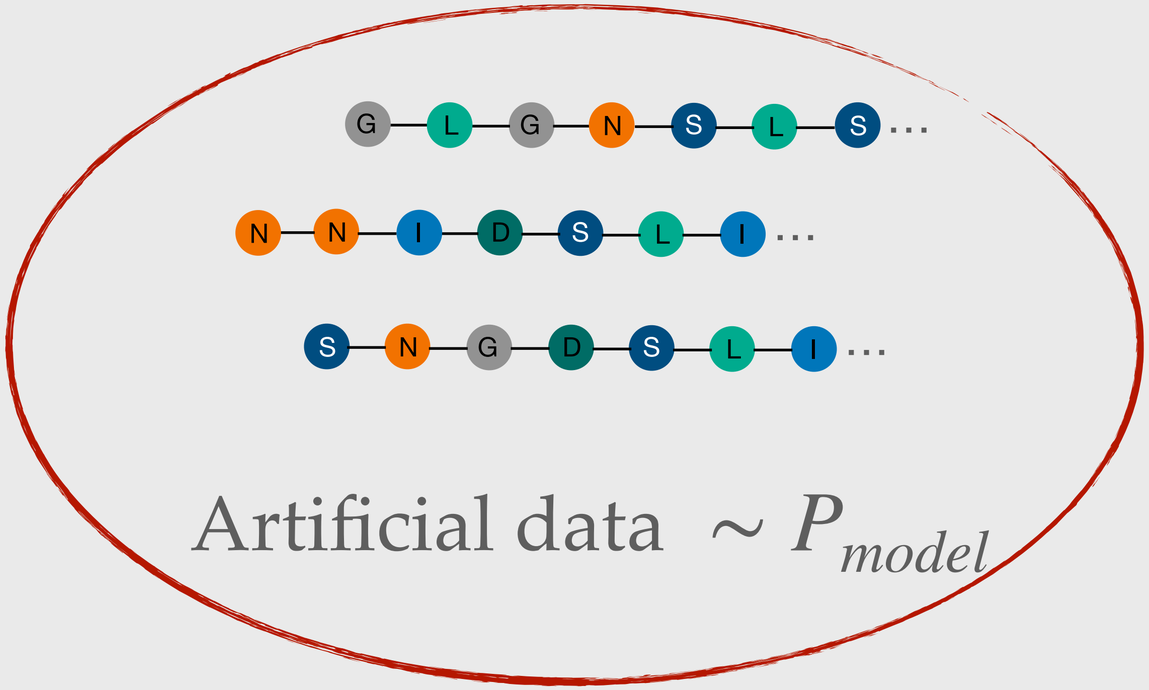
1. Learn P_{model} similar to P_{data}
2. Generate samples from P_{model}

Generative Models of Protein sequences

Natural protein sequences



Training data $\sim P_{data}$



Artificial data $\sim P_{model}$

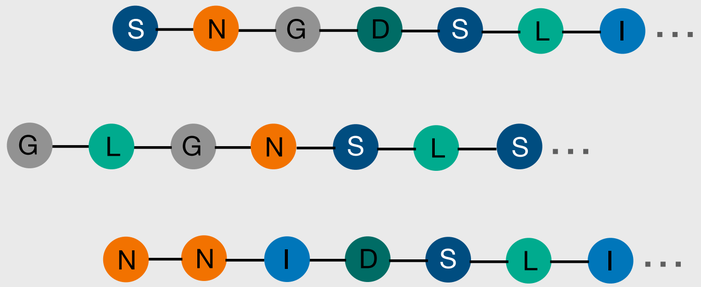
Different from training data ?

1. Learn P_{model} similar to P_{data}
2. Generate samples from P_{model}

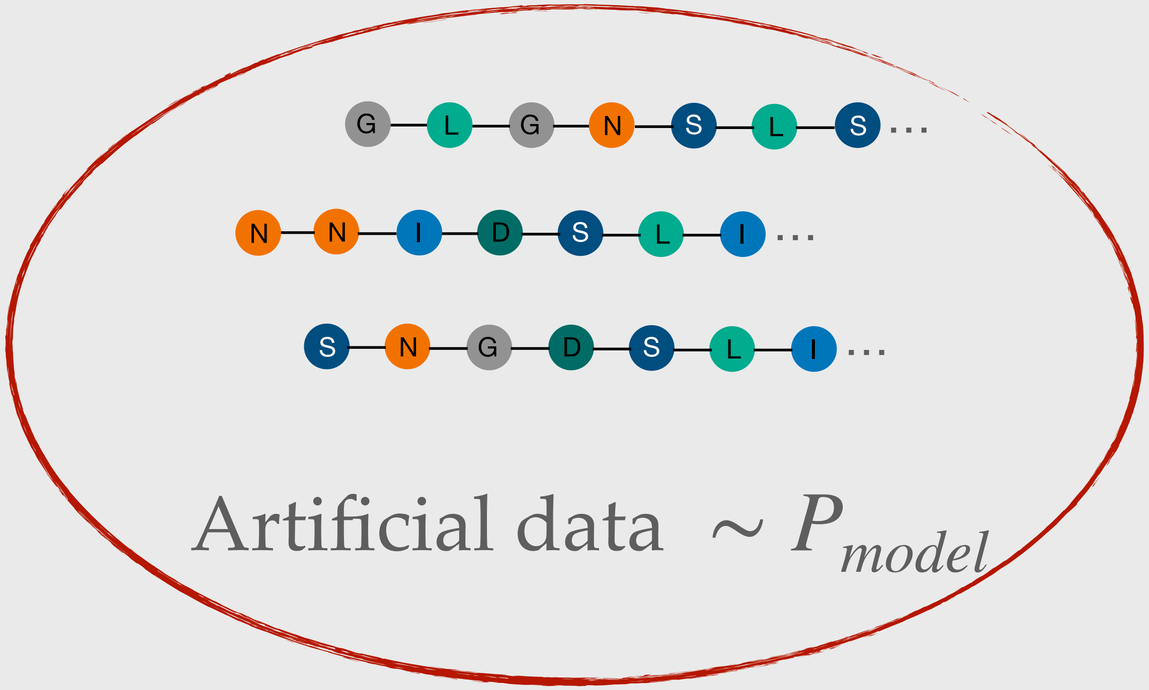
Overfitting

Generative Models of Protein sequences

Natural protein sequences



Training data $\sim P_{data}$



Artificial data $\sim P_{model}$

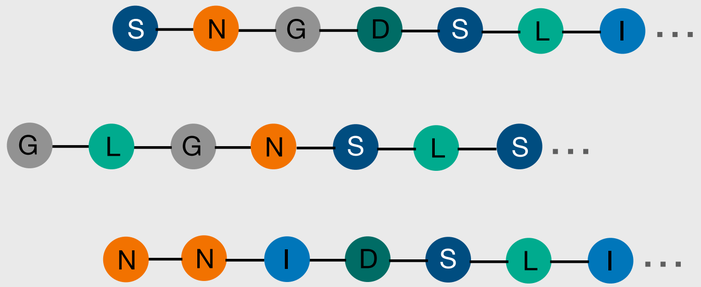
Different from training data ?

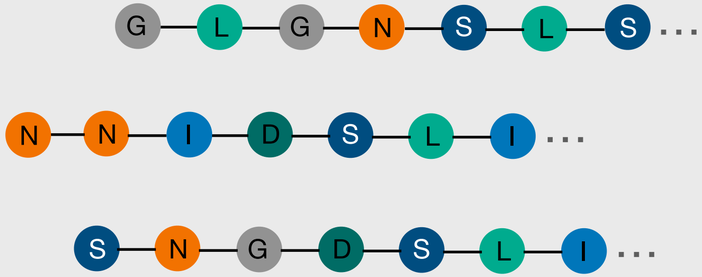
- Regularization

1. Learn P_{model} similar to P_{data}
2. Generate samples from P_{model}

Generative Models of Protein sequences

Natural protein sequences





Training data $\sim P_{data}$

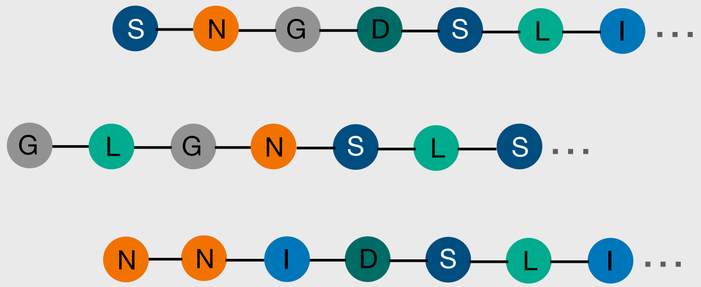
Artificial data $\sim P_{model}$

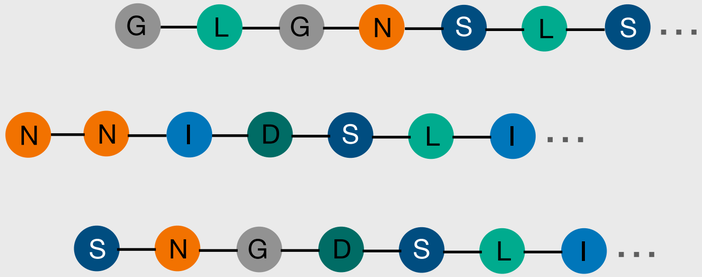
Can we learn something ?

1. Learn P_{model} similar to P_{data}
2. Generate samples from P_{model}

Generative Models of Protein sequences

Natural protein sequences





Training data $\sim P_{data}$

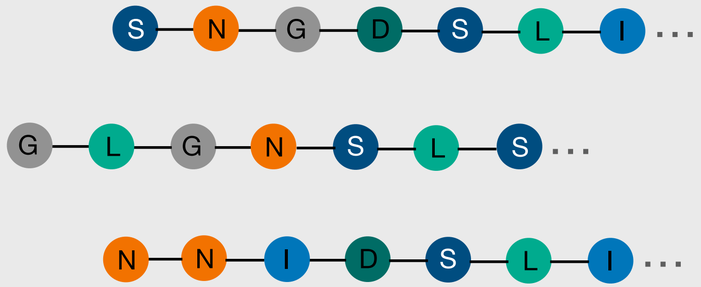
Artificial data $\sim P_{model}$

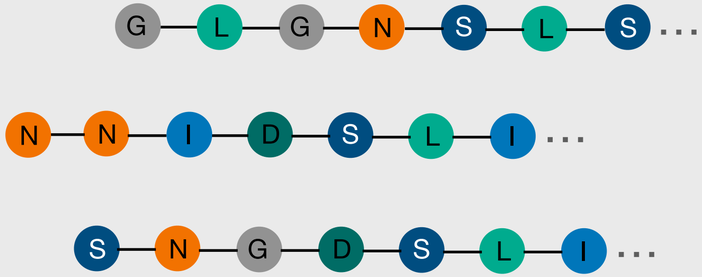
Can we learn something ?

1. Learn P_{model} similar to P_{data}
2. Generate samples from P_{model}

Generative Models of Protein sequences

Natural protein sequences





Training data $\sim P_{data}$

Artificial data $\sim P_{model}$

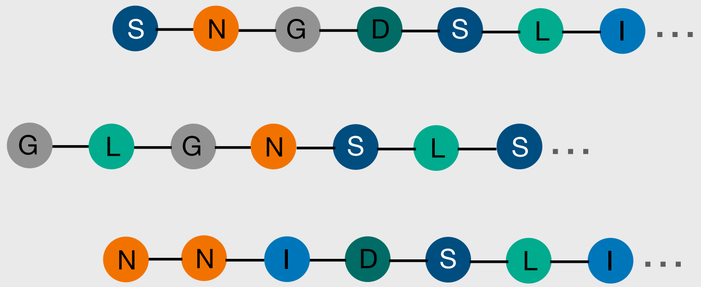
Can we learn something ?

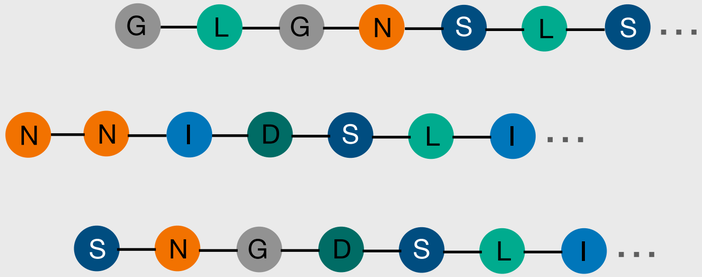
- Work on interpretability

1. Learn P_{model} similar to P_{data}
2. Generate samples from P_{model}

Generative Models of Protein sequences

Natural protein sequences





Training data $\sim P_{data}$

Artificial data $\sim P_{model}$

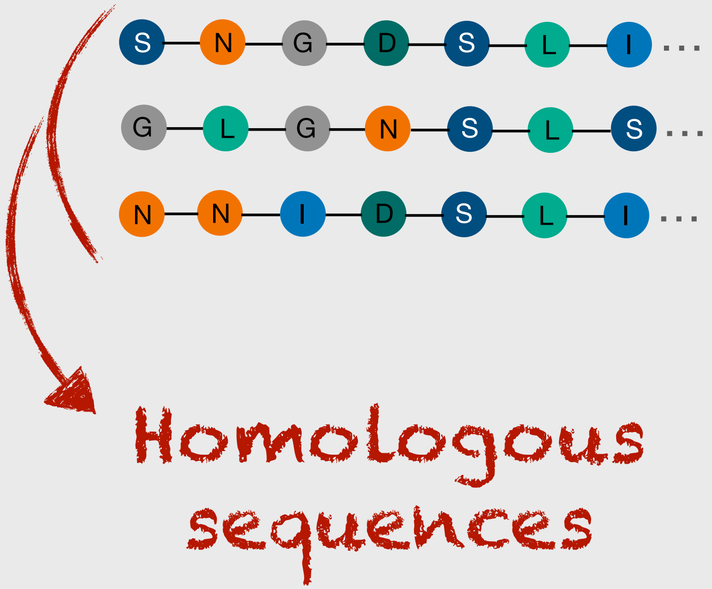
Can we learn something ?

- Work on interpretability
- Handcrafted methods

1. Learn P_{model} similar to P_{data}
2. Generate samples from P_{model}

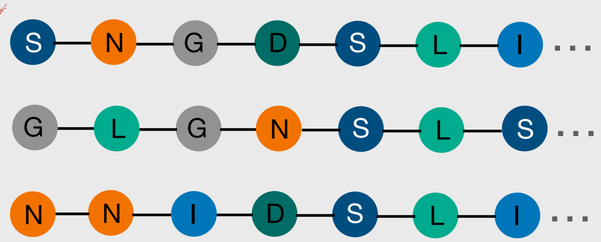
Generative Models of Protein sequences

Amino-acid chains



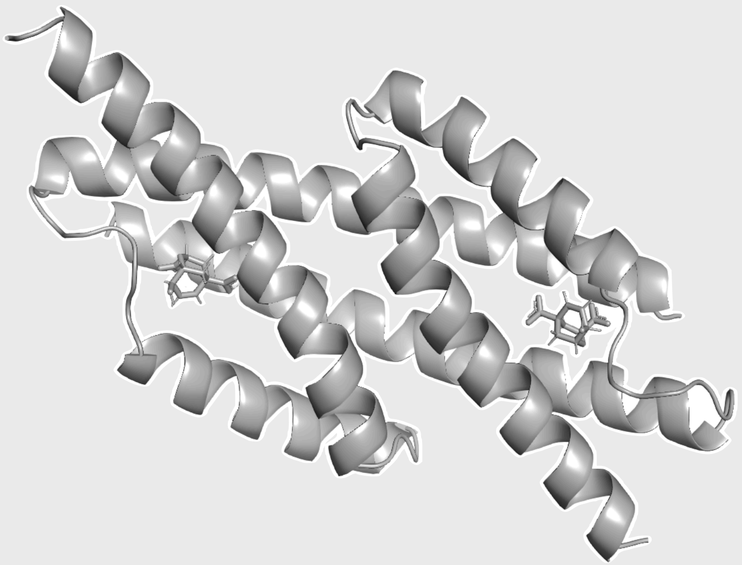
Generative Models of Protein sequences

Amino-acid chains



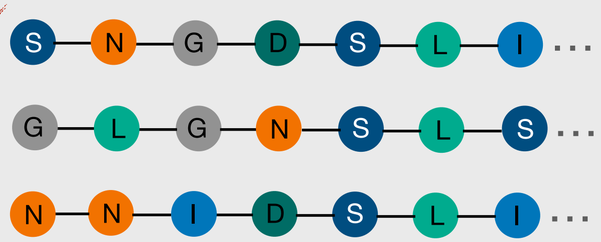
Homologous sequences

3d structure



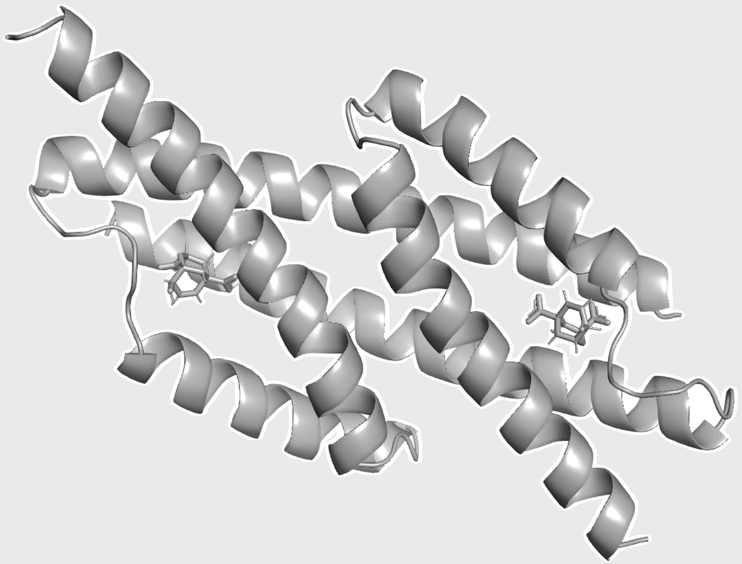
Generative Models of Protein sequences

Amino-acid chains



Homologous sequences

3d structure

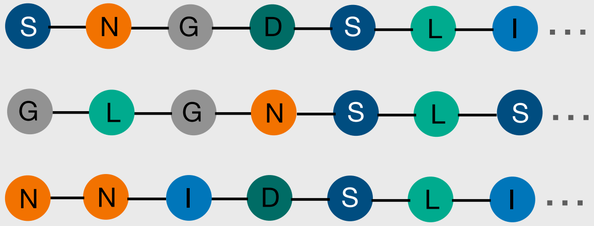


Function

Catalyse a specific reaction

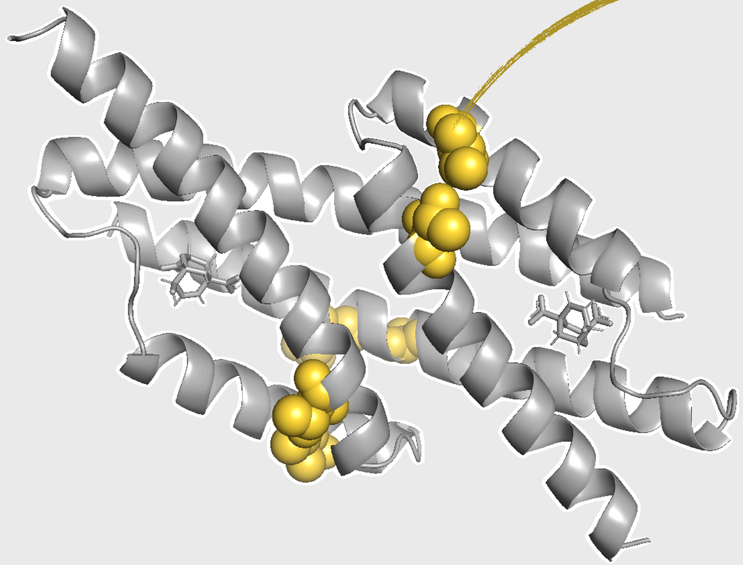
Generative Models of Protein sequences

Amino-acid chains



Homologous sequences

3d structure



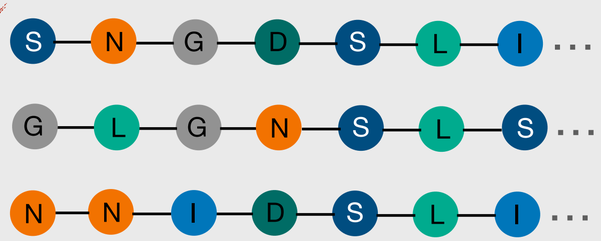
Contacts

Function

Catalyse a specific reaction

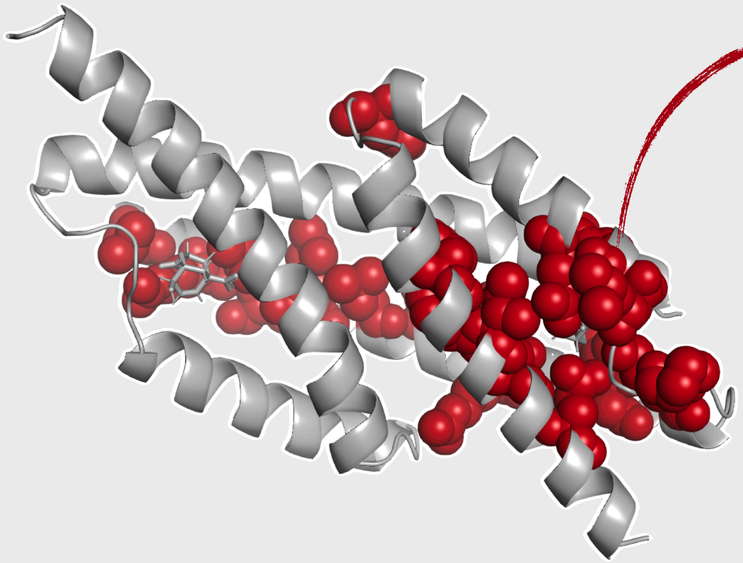
Generative Models of Protein sequences

Amino-acid chains



Homologous sequences

3d structure

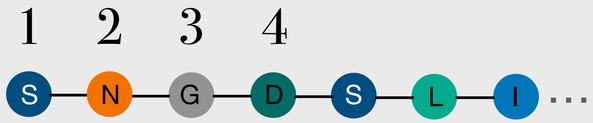


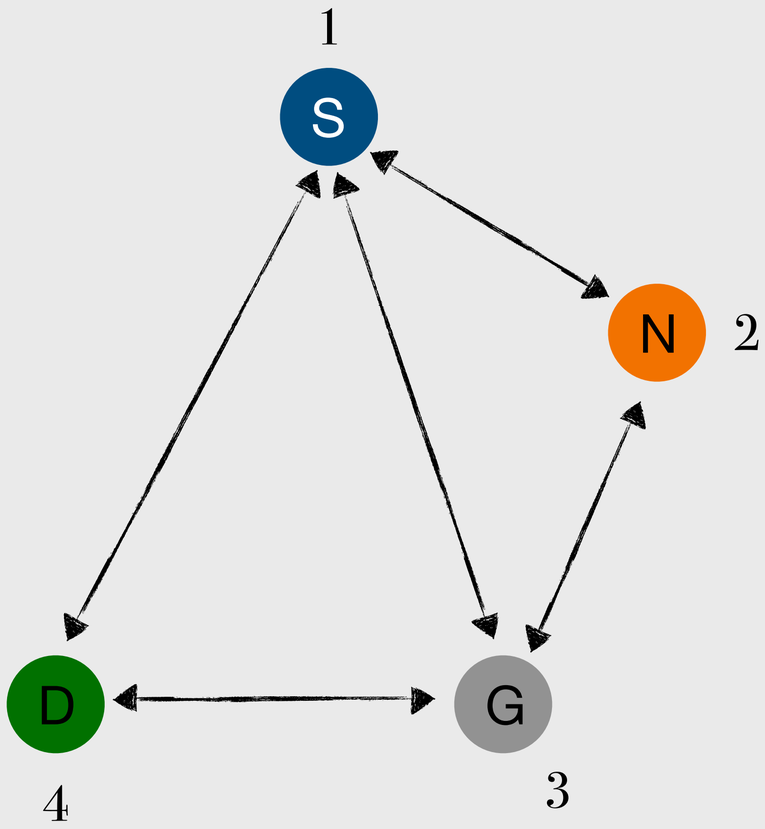
Functional Positions

Function

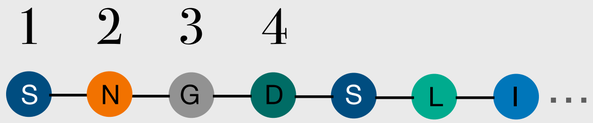
Catalyse a specific reaction

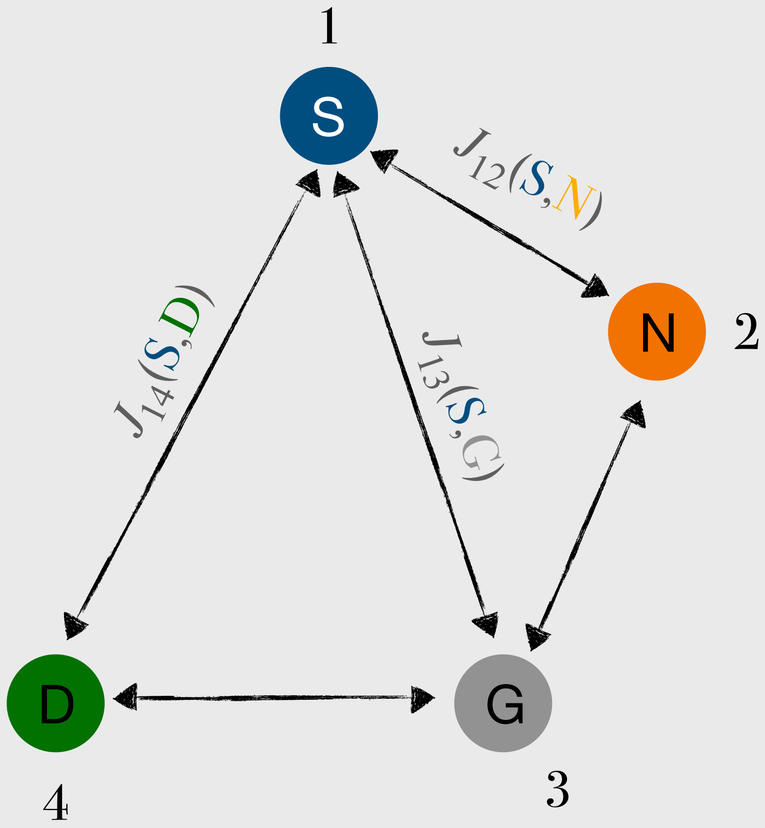
Generative Models of Protein sequences



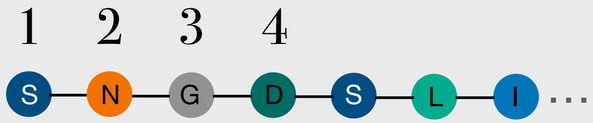


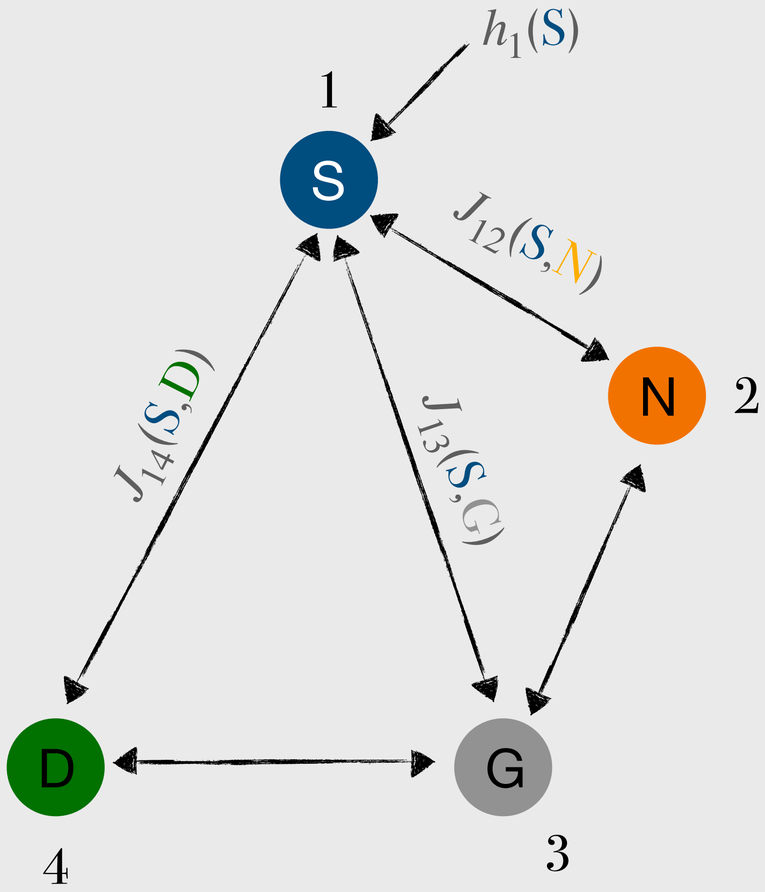
Generative Models of Protein sequences



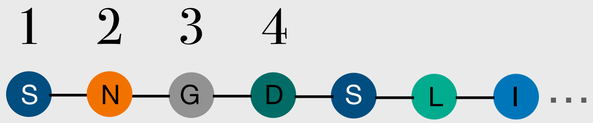


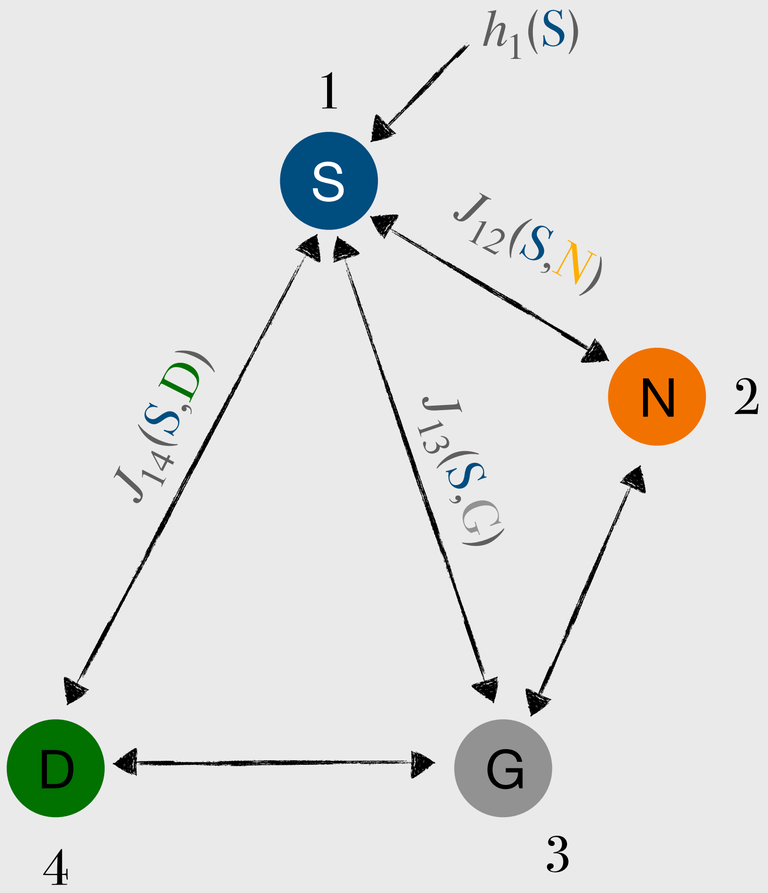
Generative Models of Protein sequences





Generative Models of Protein sequences

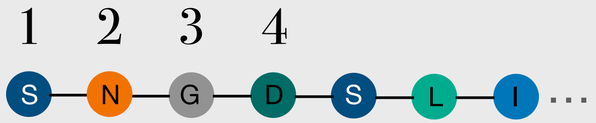


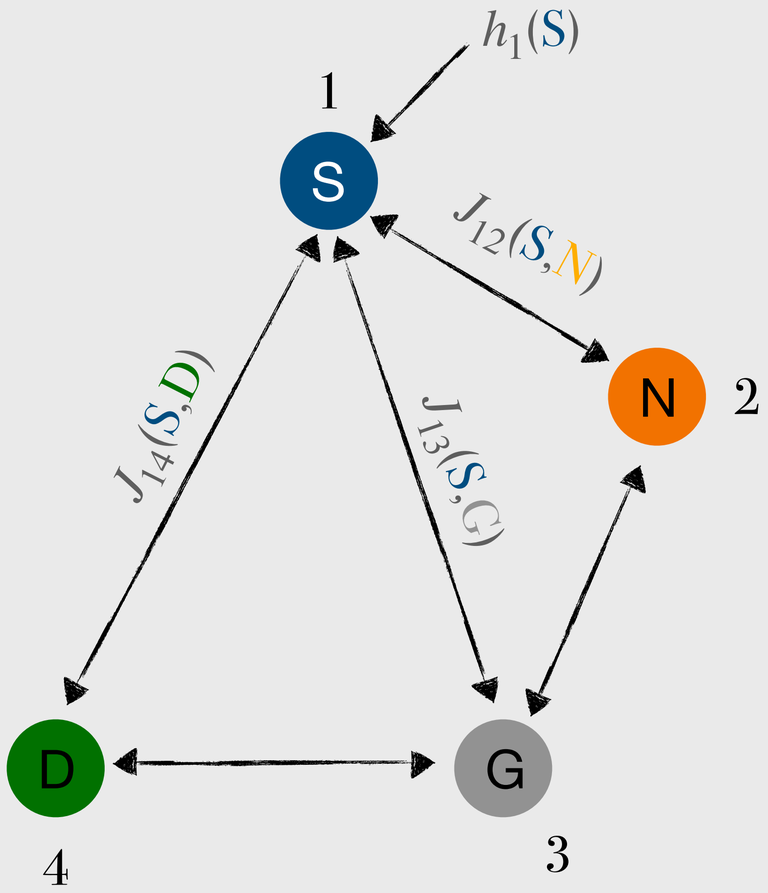


Potts Model :

$$p(\{\sigma_i\}_{i=1,\dots,L}) = \frac{1}{Z(\boldsymbol{h},\boldsymbol{J})} \prod_{i=1}^{L} e^{h_i(\sigma_i)} \prod_{i<j} e^{J_{ij}(\sigma_i,\sigma_j)}$$

Generative Models of Protein sequences





Potts Model :

$$p(\{\sigma_i\}_{i=1,\dots,L}) = \frac{1}{Z(\boldsymbol{h},\boldsymbol{J})} \prod_{i=1}^{L} e^{h_i(\sigma_i)} \prod_{i<j} e^{J_{ij}(\sigma_i,\sigma_j)}$$

- Maximum entropy model trained to match the empirical **one and two-body frequencies** of amino acids
- Parameters inferred with Gradient descent algorithm
- Boltzmann Machine algorithm (BM)

[1] Russ WP et al., 2020 [2] Morcos F et al., 2011 [4] Bitbol AF et al., 2016
[3] Marks DS et al., 2012 [5] Hopf TA et al., 2017

Undersampling induced-biases

Number of natural sequences $<<$ Number of possible sequences

$\sim 10^3 - 10^5$ $\sim 10^{66} - 10^{650}$

Undersampling induced-biases

Number of natural sequences $<<$ Number of possible sequences

Overfitting

→ Regularization

Undersampling induced-biases

Number of natural sequences $<<$ Number of possible sequences

Overfitting

→ Regularization

L2-Regularization :

Undersampling induced-biases

Number of natural sequences $<<$ Number of possible sequences

Overfitting

Regularization

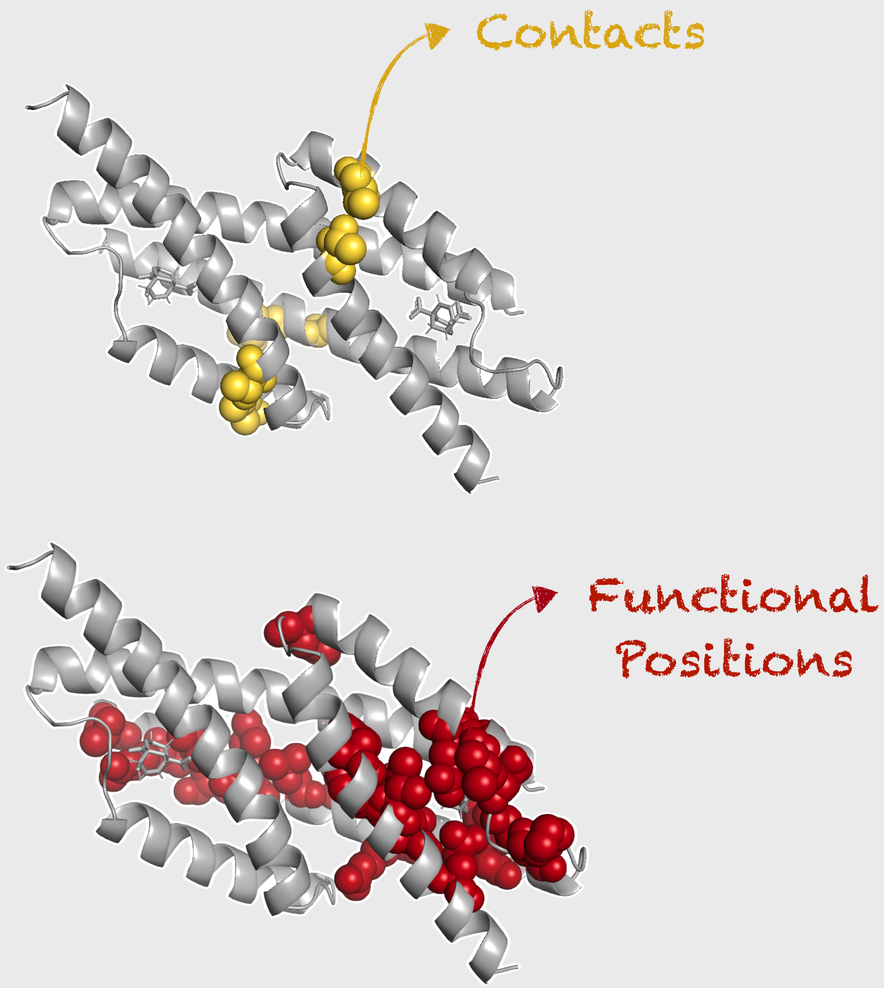
L2-Regularization : Loss $+ \lambda \times$ Penalty

L2-norm of the parameters

The larger λ is, the stronger the regularization

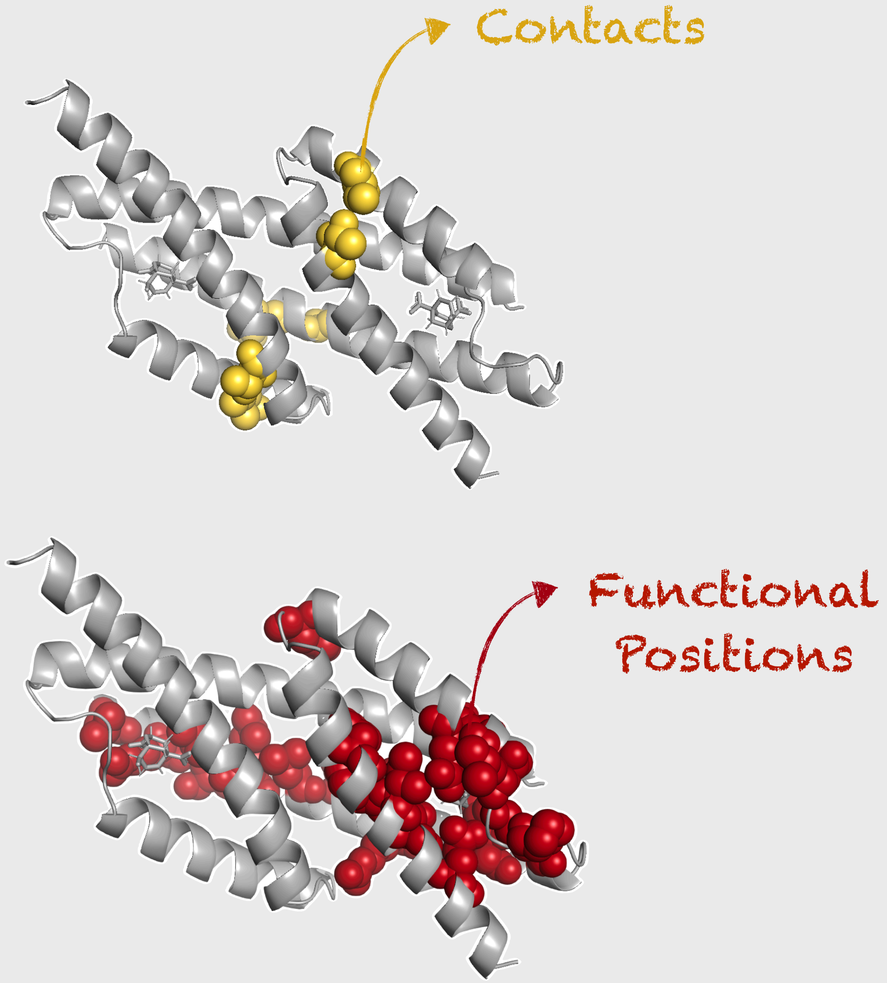
Undersampling induced-biases

Different kind of features

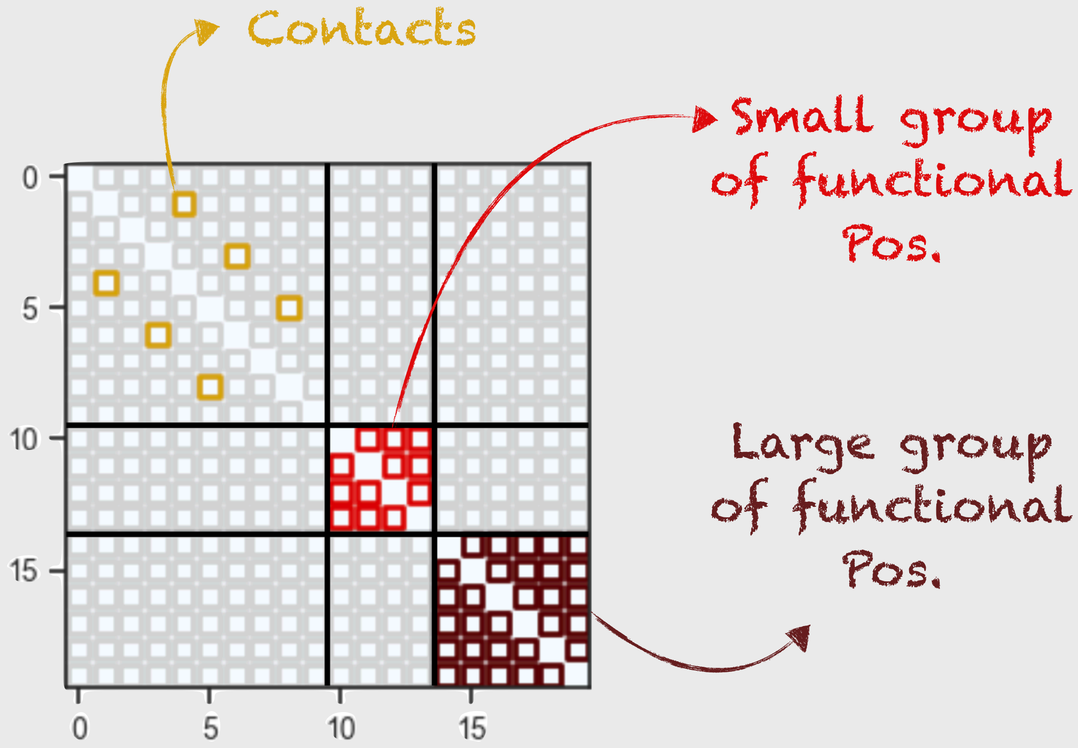


Undersampling induced-biases

Different kind of features

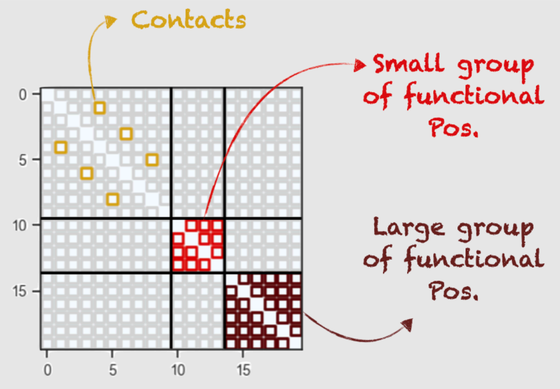


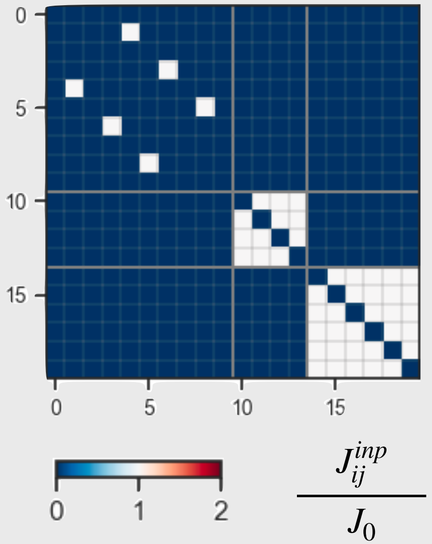
Let's play with a Toy Model…



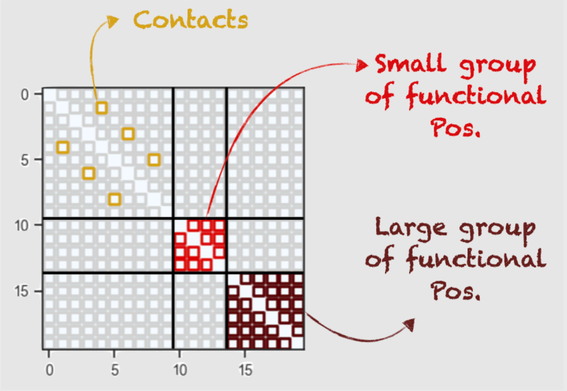
Y. Kleeorin, W. P. Russ, O. Rivoire, and R. Ranganathan, 2021.

Undersampling induced-biases

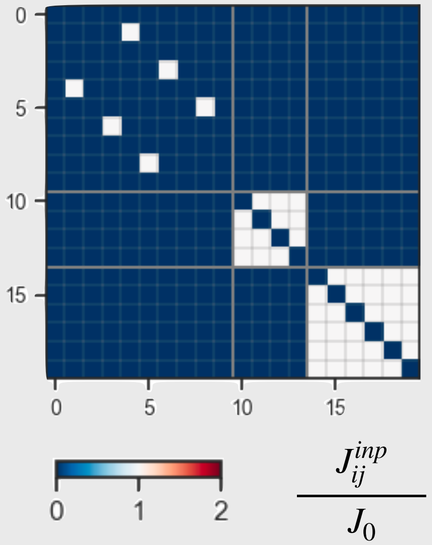


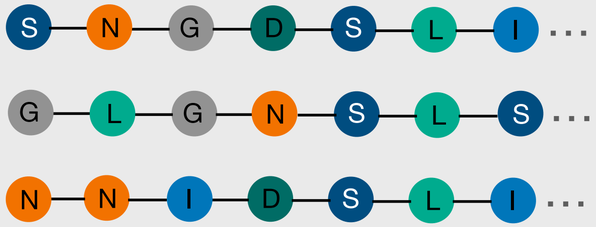


Undersampling induced-biases

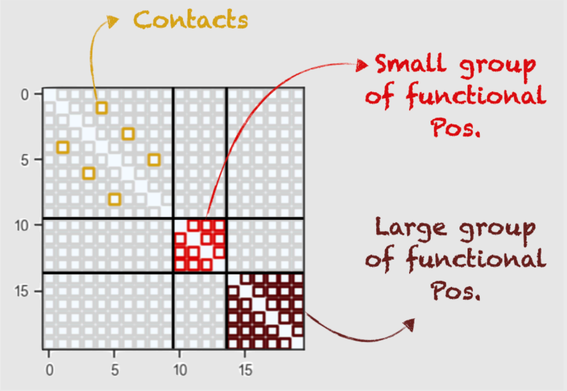


Sample sequences



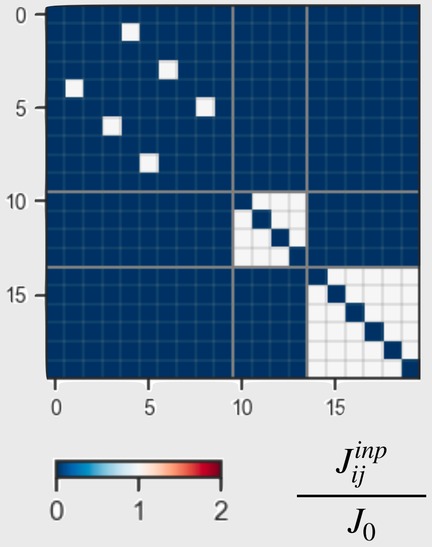


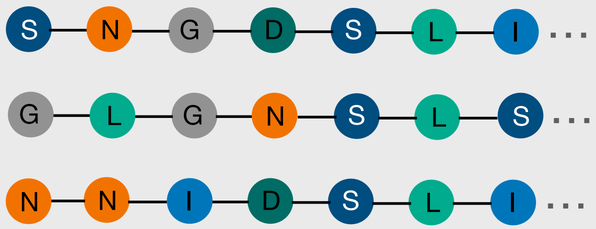
Undersampling induced-biases



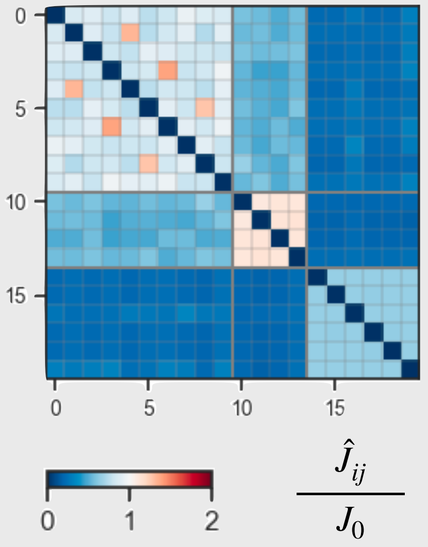
Sample sequences

Inference with a regularized BM

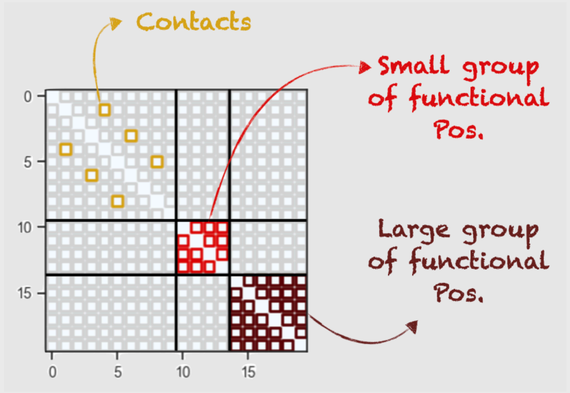




BM

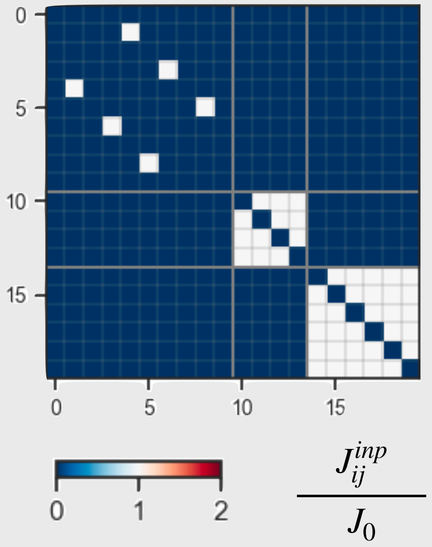


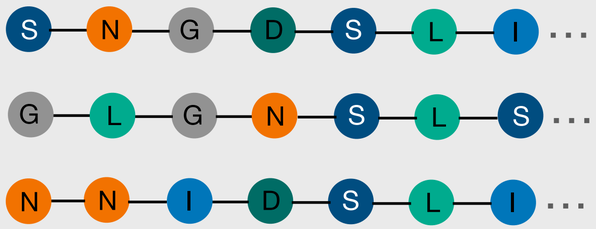
Undersampling induced-biases



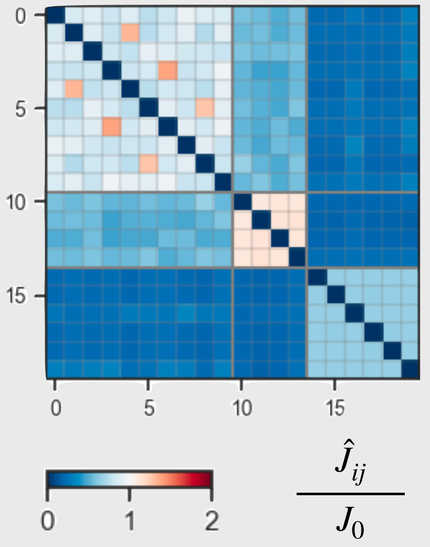
Sample sequences

Inference with a regularized BM



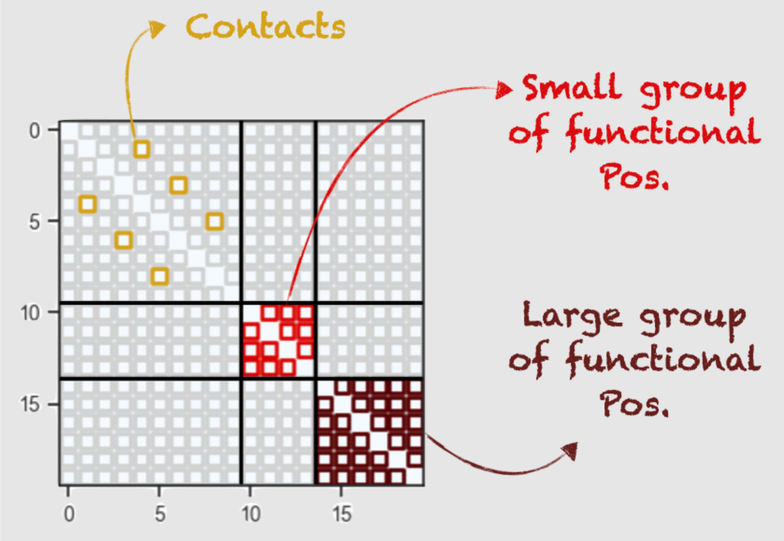


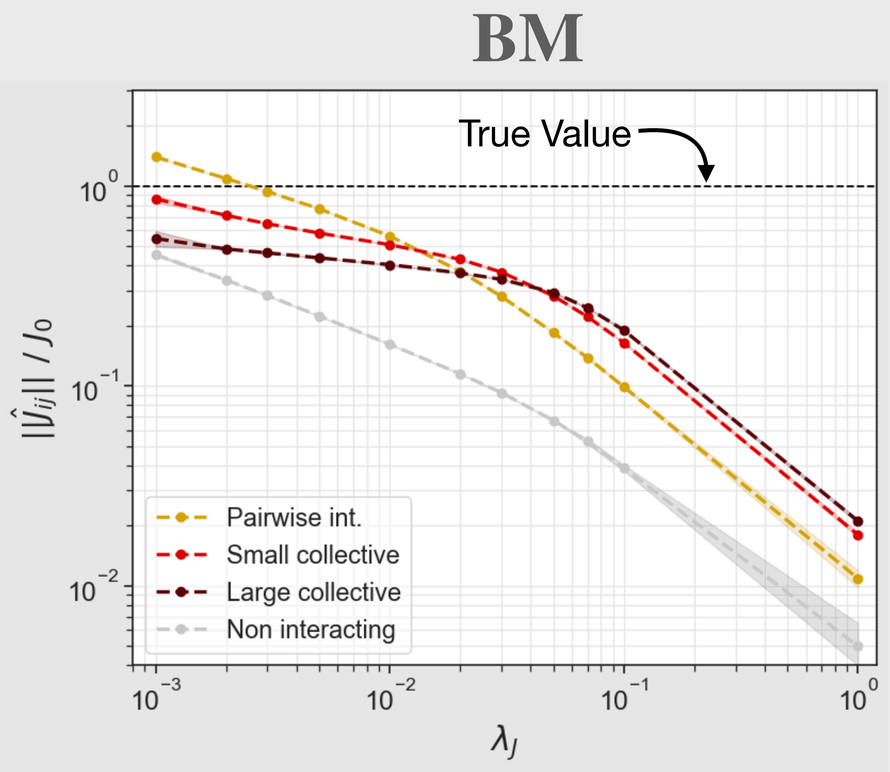
BM



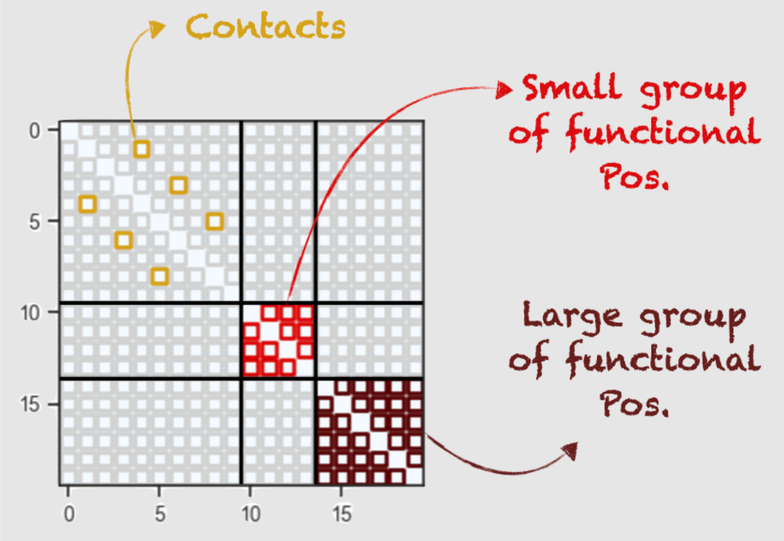
Bias between features of different scales !

Undersampling induced-biases as function of regularization

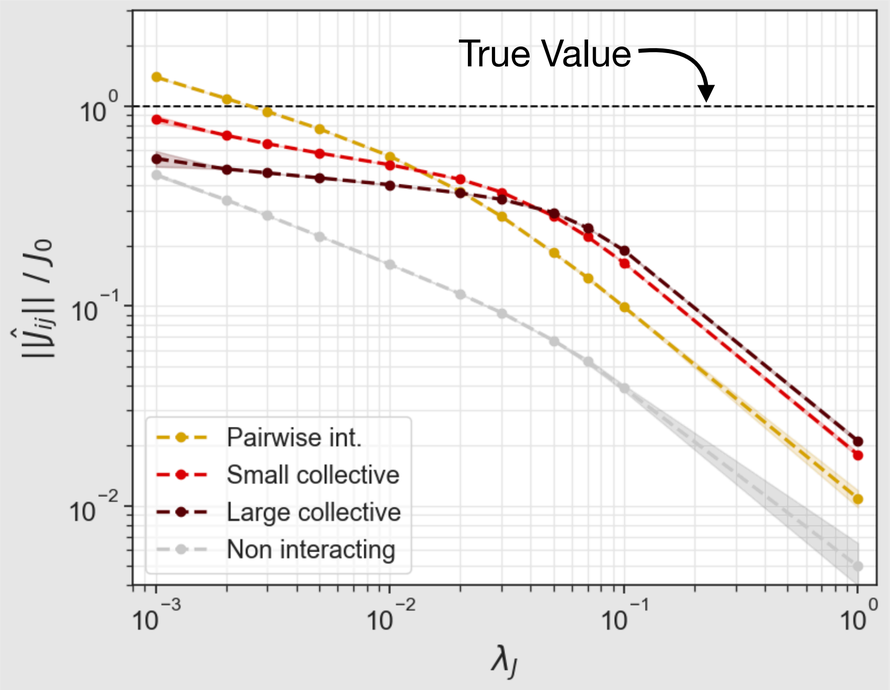




Undersampling induced-biases as function of regularization



BM



New Method

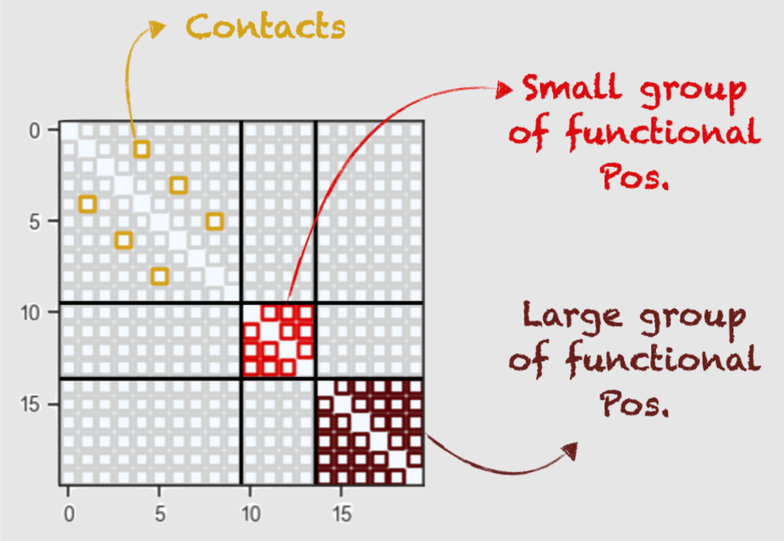
SBM

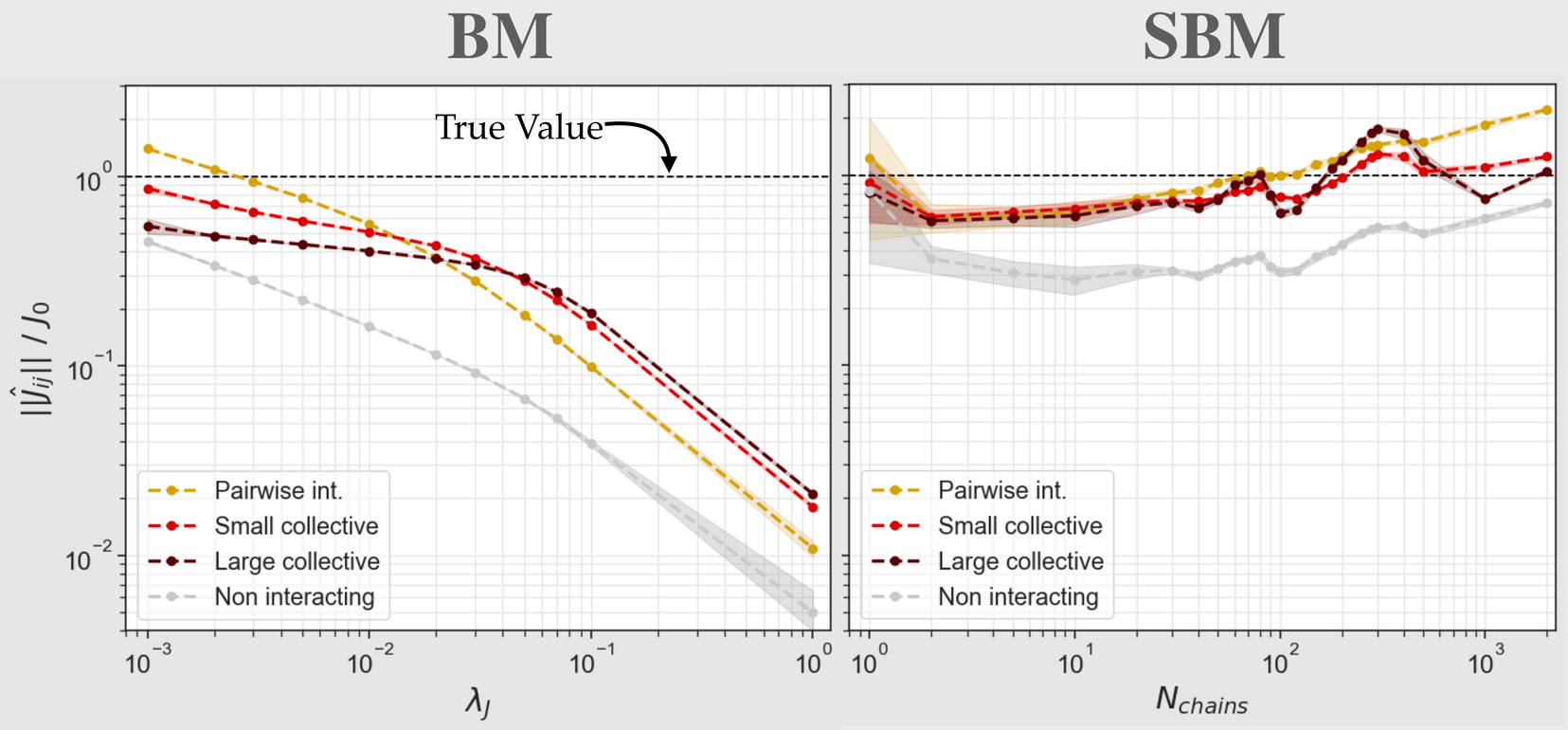
Stochastic Boltzmann Machine

- Quasi-Newton Gradient Descent *
- Implicit regularization
- Undersample to compute model statistics (N_{chains})

* L-BFGS (D. C. Liu and J. Nocedal, 1989)

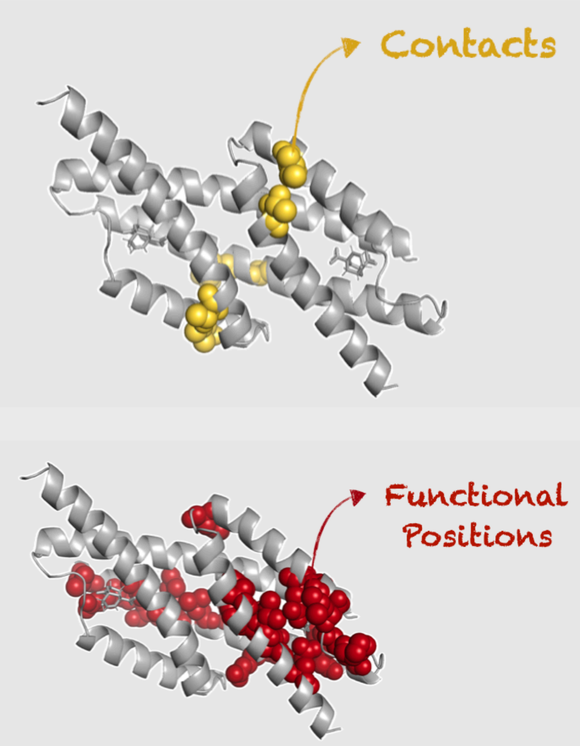
Undersampling induced-biases as function of regularization

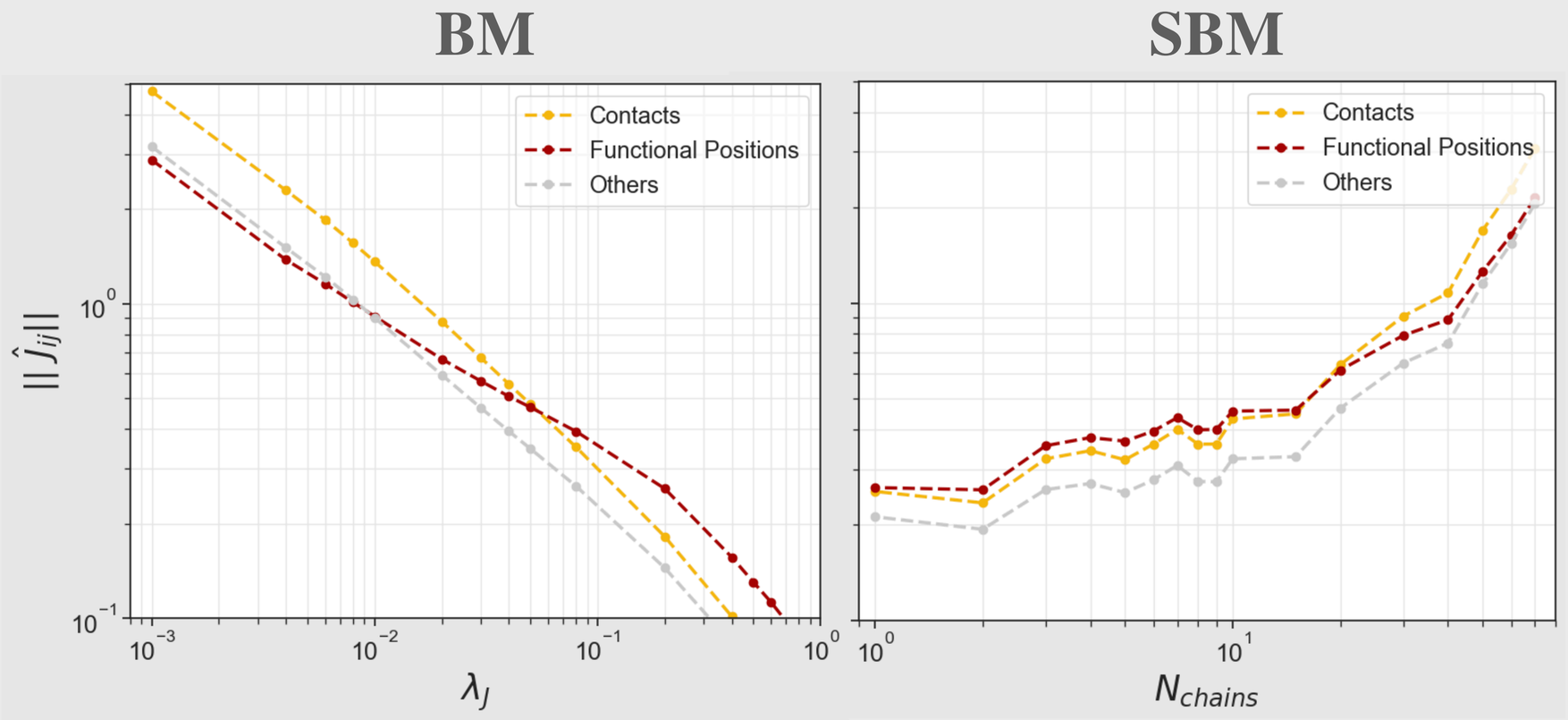


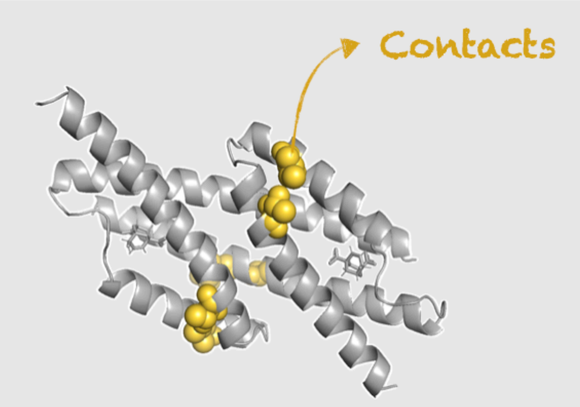


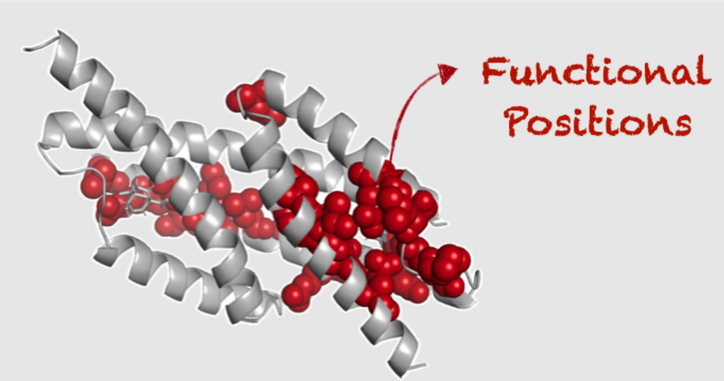
Undersampling induced-biases

Chorismate Mutase family

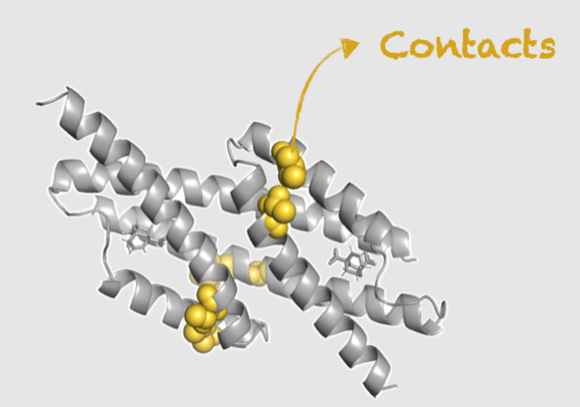


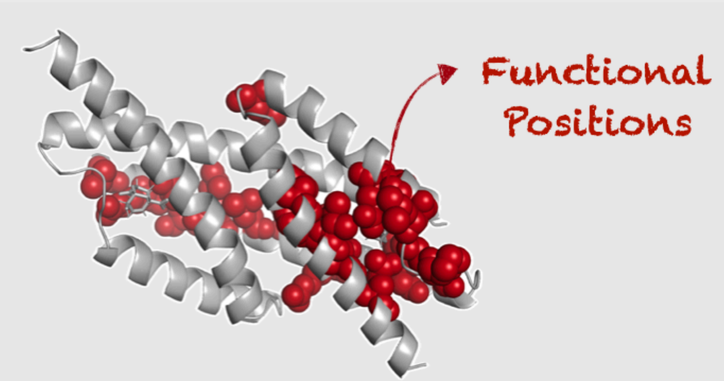






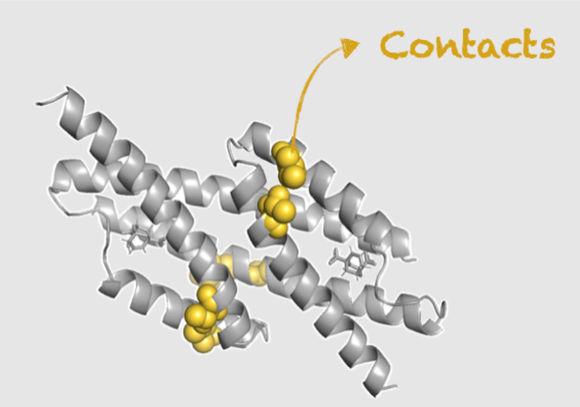
→ SBM allows to combine the inference of both local & Collective features

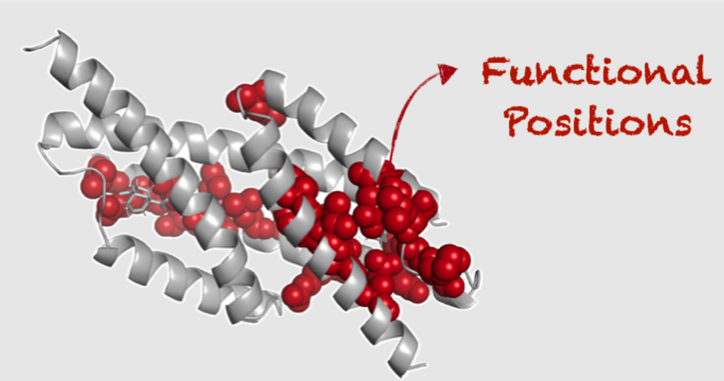




→ SBM allows to combine the inference of both local & Collective features

→ SBM allows reproduce the diversity of natural protein families





→ SBM allows to combine the inference of both local & Collective features

→ SBM allows reproduce the diversity of natural protein families

→ Other statistics are just as well reproduced with SBM as with BM

(1^{st}, 2^{nd}, 3^{rd} order statistics, and PCA)

→ SBM converges very fast

Thanks for your attention

CompBio@TrEE

UMR 7083 Gulliver

ESPCI PARIS

Supervisors: Olivier Rivoire & Ivan Junier

Collaborator: Yaakov Kleeorin

TIMC

UGA Université Grenoble Alpes