Generative Models of protein sequences

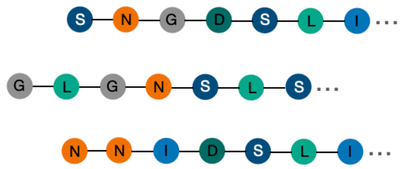
Marion Chauveau, 2nd year Phd student

Supervisors: Ivan Junier & Olivier Rivoire / Collaborator: Yaakov Kleeorin

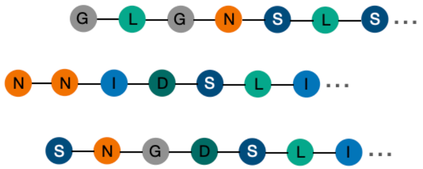
01/2024
Journées de la Physique Statistique

UMR 7083 Gulliver ESPCI PARIS TIMC UGA Université Grenoble Alpes

Generative Models



Training data $\sim P_{data}$

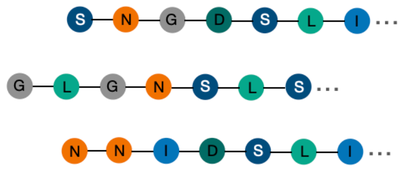


Artificial data $\sim P_{model}$

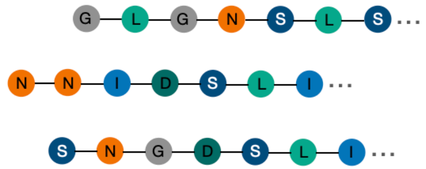
1. Learn P_{model} similar to P_{data}
2. Generate samples from P_{model}

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

Generative Models



Training data $\sim P_{data}$

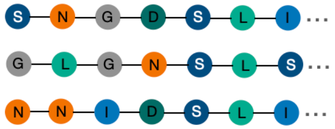


Artificial data $\sim P_{model}$

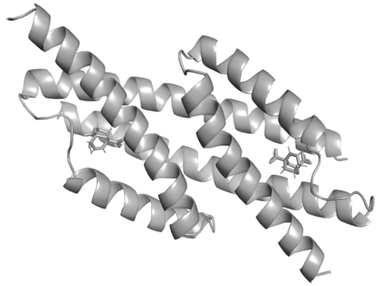
1. Learn P_{model} similar to P_{data}
2. Generate samples from P_{model}

Homologous sequences

Amino-acid chains



3d structure

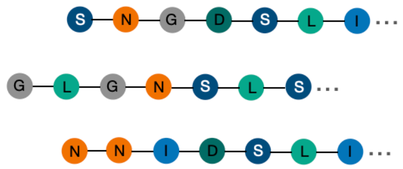


Function

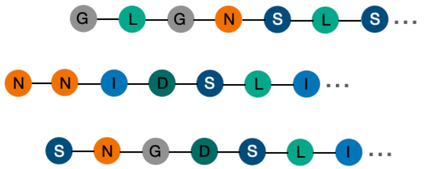
Catalyse a specific reaction

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

Generative Models



Training data $\sim P_{data}$

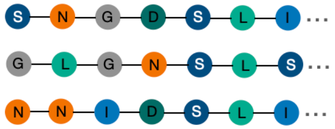


Artificial data $\sim P_{model}$

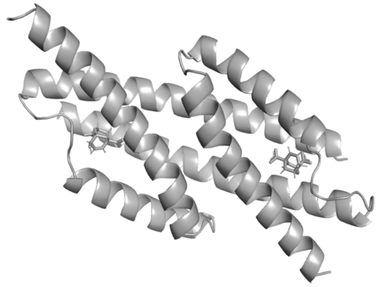
1. Learn P_{model} similar to P_{data}
2. Generate samples from P_{model}

Homologous sequences

Amino-acid chains



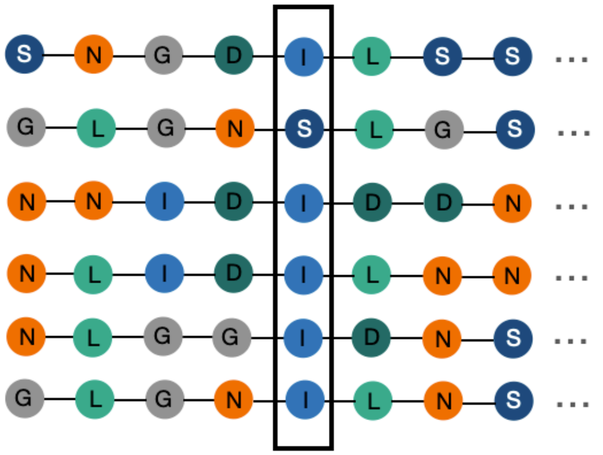
3d structure



Function

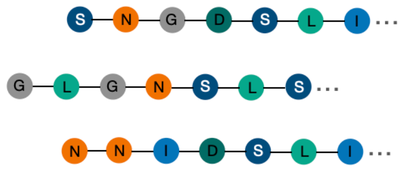
Catalyse a specific reaction

Conservation

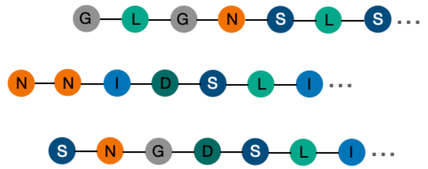


[1] Russ WP et al., 2020 [2] Morcos F et al., 2011 [3] Marks DS et al., 2012 [4] Bitbol AF et al., 2016 [5] Hopf TA et al., 2017

Generative Models



Training data $\sim P_{data}$

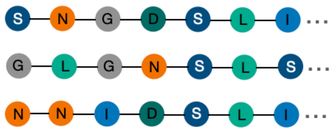


Artificial data $\sim P_{model}$

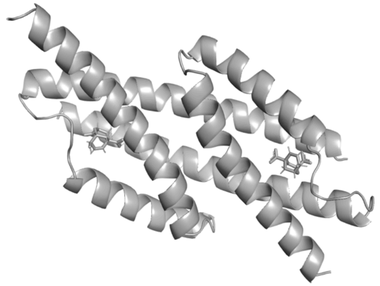
1. Learn P_{model} similar to P_{data}
2. Generate samples from P_{model}

Homologous sequences

Amino-acid chains



3d structure

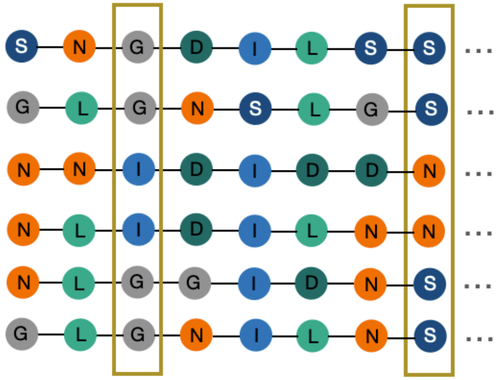


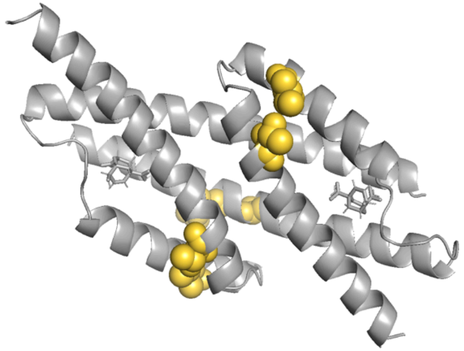
Function

Catalyse a specific reaction

Correlations

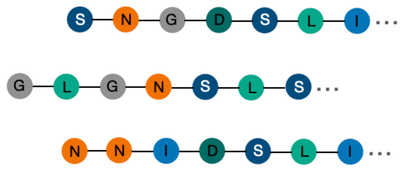
Contacts



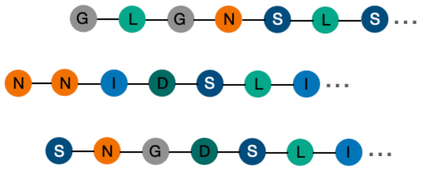


[1] Russ WP et al., 2020 [2] Morcos F et al., 2011 [3] Marks DS et al., 2012 [4] Bitbol AF et al., 2016 [5] Hopf TA et al., 2017

Generative Models



Training data $\sim P_{data}$

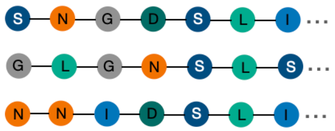


Artificial data $\sim P_{model}$

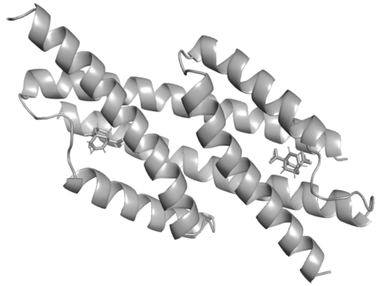
1. Learn P_{model} similar to P_{data}
2. Generate samples from P_{model}

Homologous sequences

Amino-acid chains



3d structure

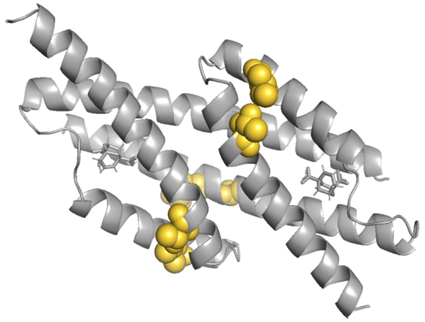


Function

Catalyse a specific reaction

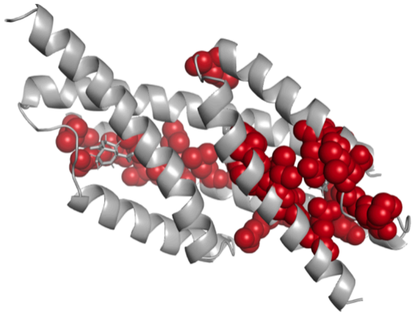
Correlations

Contacts



Local

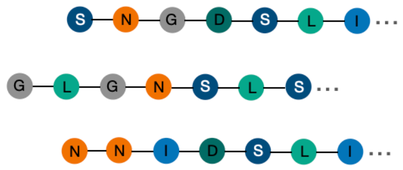
Functional Positions



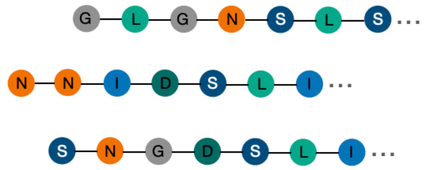
Collective

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

Generative Models



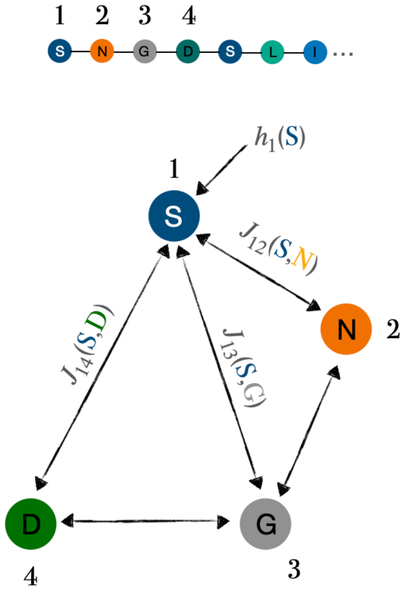
Training data $\sim P_{data}$



Artificial data $\sim P_{model}$

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

Potts Model

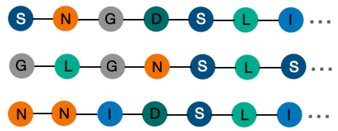


$$p(\{\sigma_i\}_{i=1,\dots,L}) = \frac{1}{Z(\boldsymbol{h},\boldsymbol{J})} \prod_{i=1}^{L} e^{h_i(\sigma_i)} \prod_{i<j} e^{J_{ij}(\sigma_i,\sigma_j)}$$

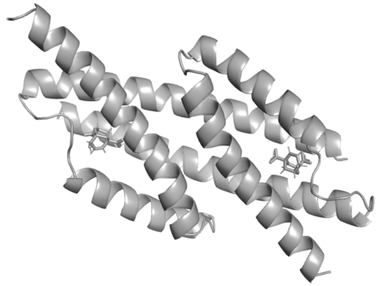
- Parameters inferred with Gradient descent algorithm
- Boltzmann Machine algorithm (BM)

Homologous sequences

Amino-acid chains



3d structure

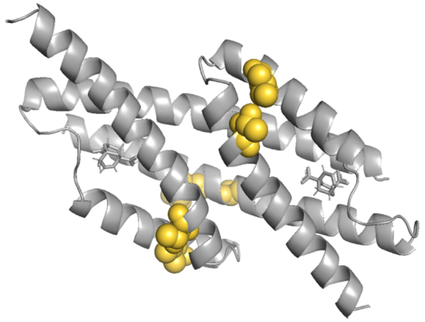


Function

Catalyse a specific reaction

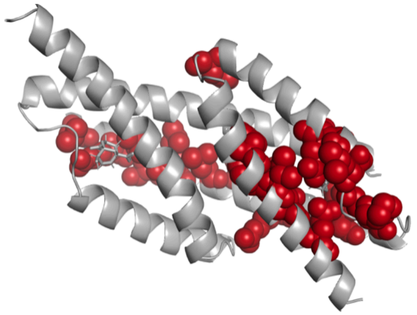
Correlations

Contacts



Local

Functional Positions

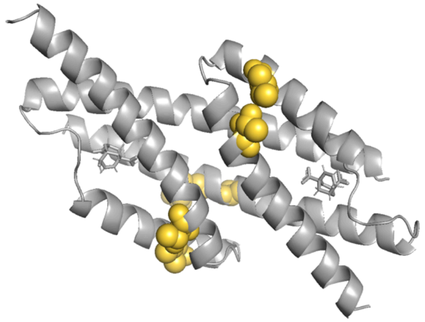


Collective

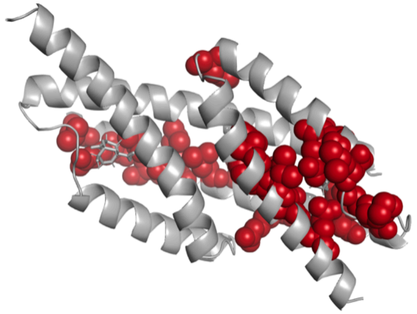
[1] Russ WP et al., 2020 [2] Morcos F et al., 2011 [3] Marks DS et al., 2012 [4] Bitbol AF et al., 2016 [5] Hopf TA et al., 2017

Different features

Contacts



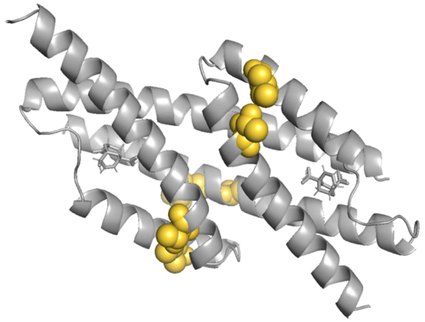
Functional Positions



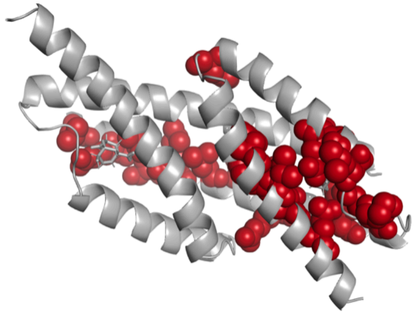
Y. Kleeorin, W. P. Russ, O. Rivoire, and R. Ranganathan, 2021

Different features

Contacts



Functional Positions



Undersampling

$$\text{Number of natural sequences} << \begin{cases} \text{Number of possible sequences} \\ \text{Number of parameters} \end{cases}$$

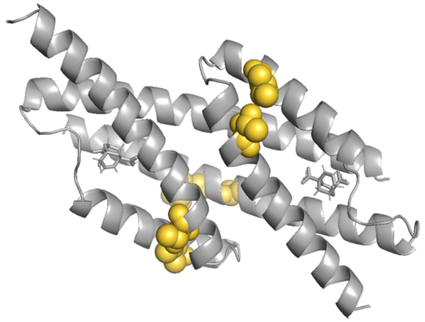
$$\text{L2-Regularization}: \quad \text{Loss} + \lambda \times \text{Penalty}$$

L2-norm of the parameters

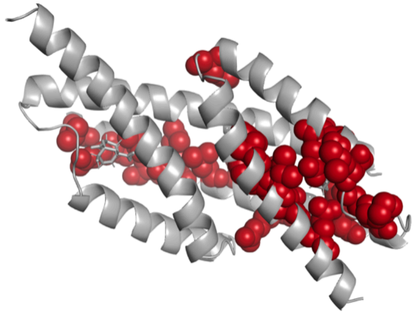
Y. Kleeorin, W. P. Russ, O. Rivoire, and R. Ranganathan, 2021

Different features

Contacts



Functional Positions



Undersampling

$$\text{Number of natural sequences} << \begin{cases} \text{Number of possible sequences} \\ \text{Number of parameters} \end{cases}$$

$$\text{L2-Regularization}: \quad \text{Loss} + \lambda \times \text{Penalty}$$

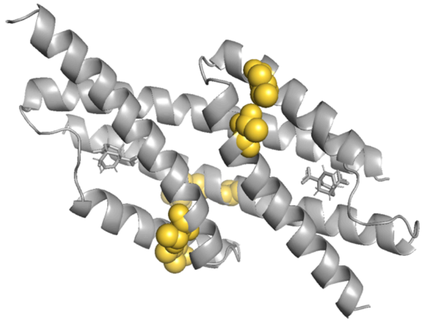
L2-norm of the parameters

→ Introduce a bias

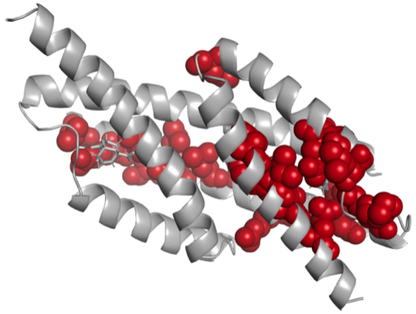
Y. Kleeorin, W. P. Russ, O. Rivoire, and R. Ranganathan, 2021

Different features

Contacts

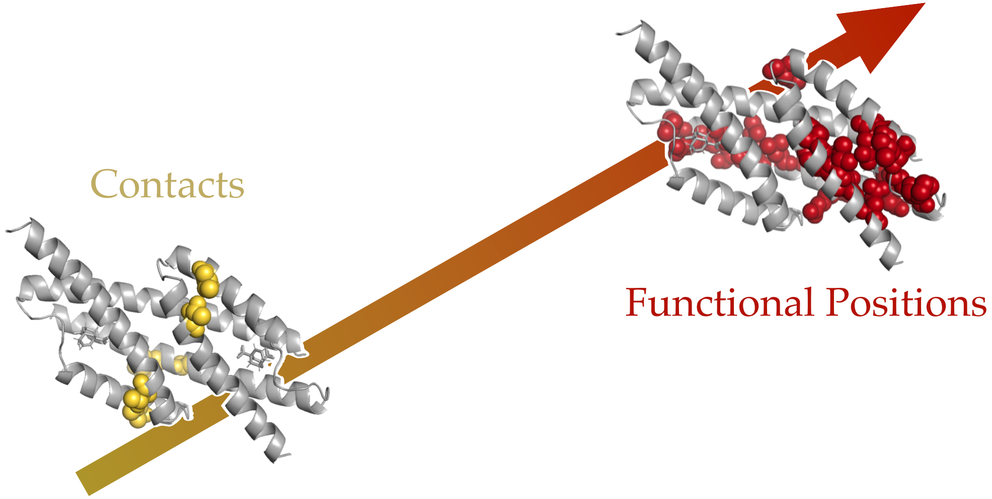


Functional Positions



High regularization

Contacts



Functional Positions

Low regularization

Undersampling

$$\text{Number of natural sequences} << \begin{cases} \text{Number of possible sequences} \\ \text{Number of parameters} \end{cases}$$

$$\text{L2-Regularization}: \quad \text{Loss} + \lambda \times \text{Penalty}$$

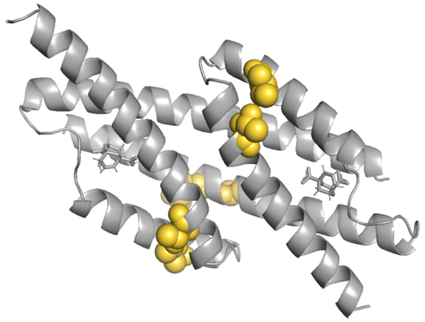
L2-norm of the parameters

→ Introduce a bias

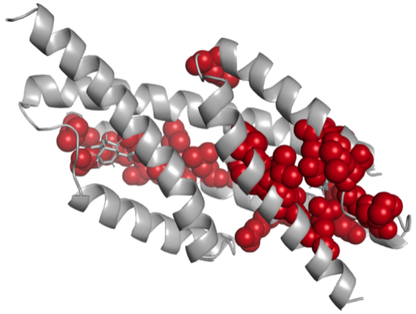
Y. Kleeorin, W. P. Russ, O. Rivoire, and R. Ranganathan, 2021

Different features

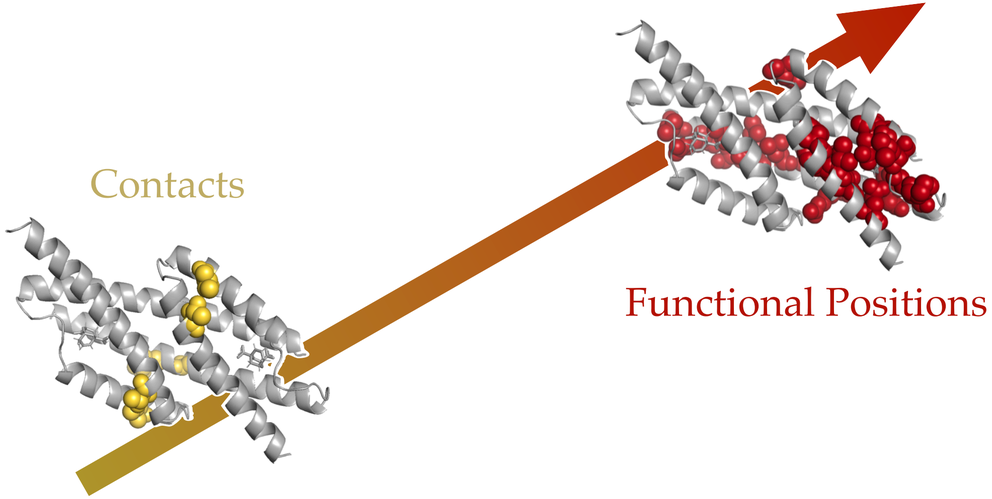
Contacts



Functional Positions



High regularization



Contacts

Functional Positions

Low regularization

Undersampling

$$\text{Number of natural sequences} << \begin{cases} \text{Number of possible sequences} \\ \text{Number of parameters} \end{cases}$$

L2-Regularization : $\text{Loss} + \lambda \times \text{Penalty}$

L2-norm of the parameters

→ Introduce a bias

Infer Generative Models that:

- Combine the inference of both local & Collective features
- Reproduce the diversity of natural protein families
- Capture other statistics

  (1^{st}, 2^{nd}, 3^{rd} order statistics, PCA, energy distributions…)

→ **SBM (stochastic Boltzmann Machine) ***

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

*Yaakov Kleeorin