Simplify to understand: extracting insight from high-dimensional simulation datasets of biological systems via information-theoretic approaches

Roberto Menichetti

Statistical & Biological Physics

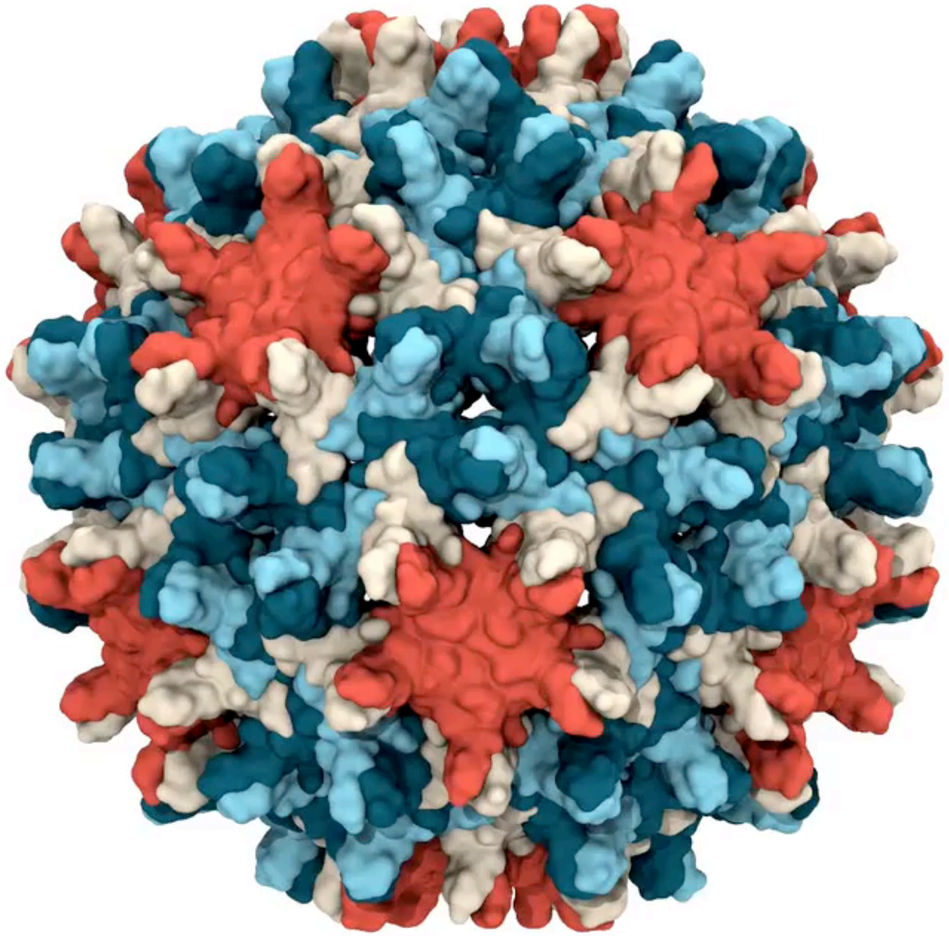
Physics Department, University of Trento, Italy
Trento Institute for Fundamental Physics and Applications (INFN-TIFPA), Italy

ALPACA Workshop
ECT*, Trento, September 2023

Data are easier and easier to get

Data are easier and easier to get

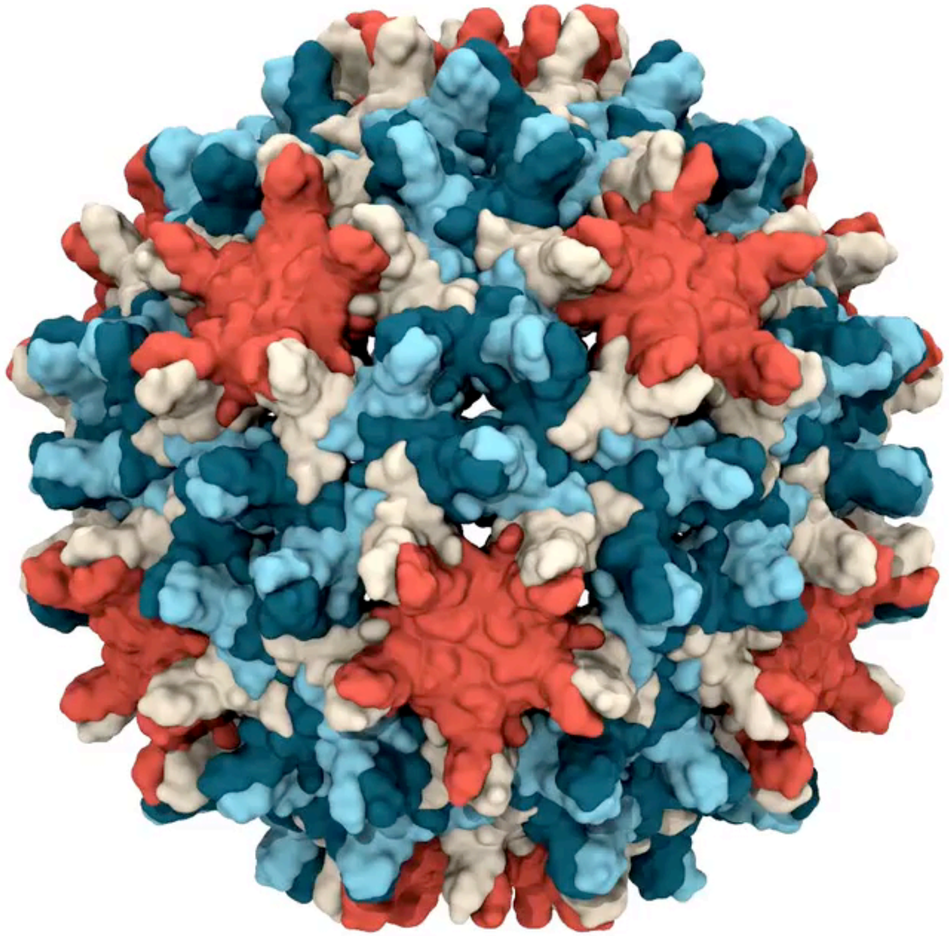
All-atom MD simulations



Hadden *et al.*, eLife 2018

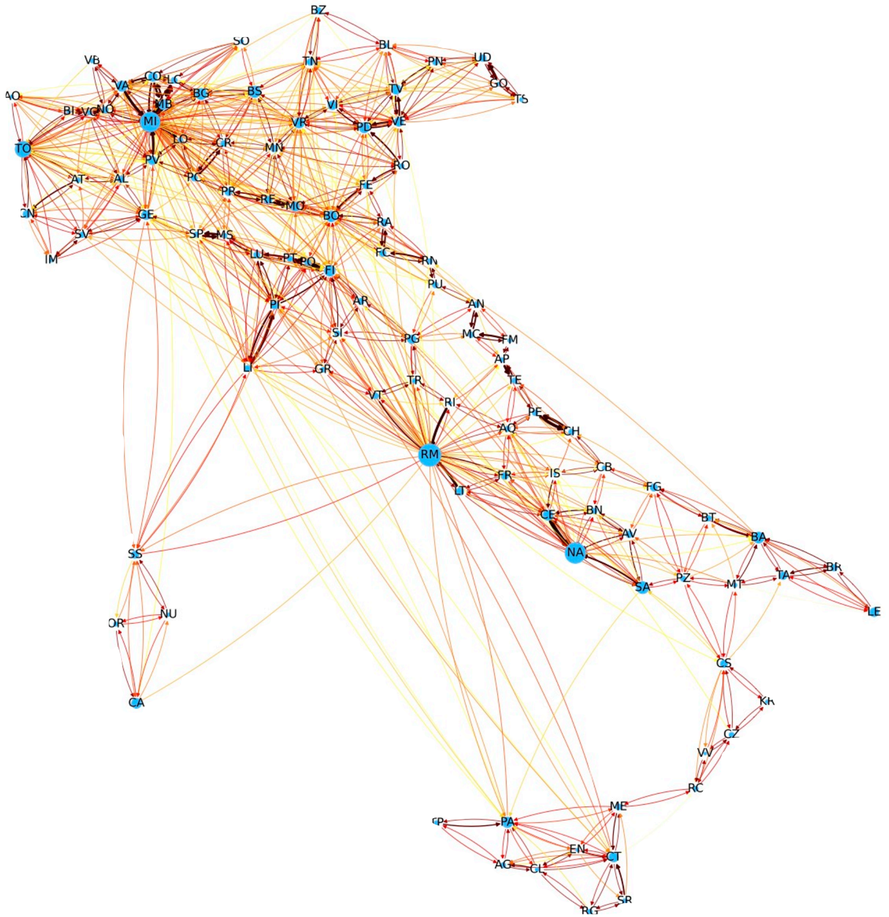
Data are easier and easier to get

All-atom MD simulations



Hadden *et al.*, eLife 2018

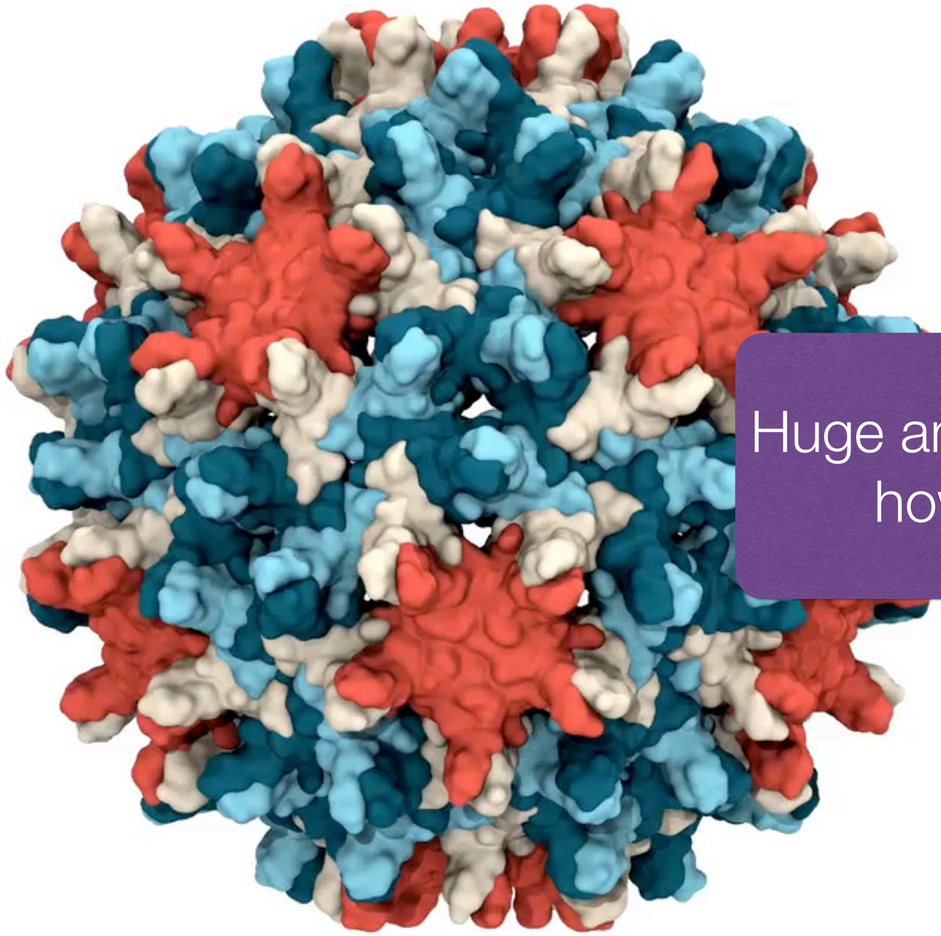
Aggregate digital databases



Morand, Yip, Velegrakis, Lattanzi, Potestio, Tubiana, *arXiv* 2023

Data are easier and easier to get

All-atom MD simulations



Hadden *et al.*, eLife 2018

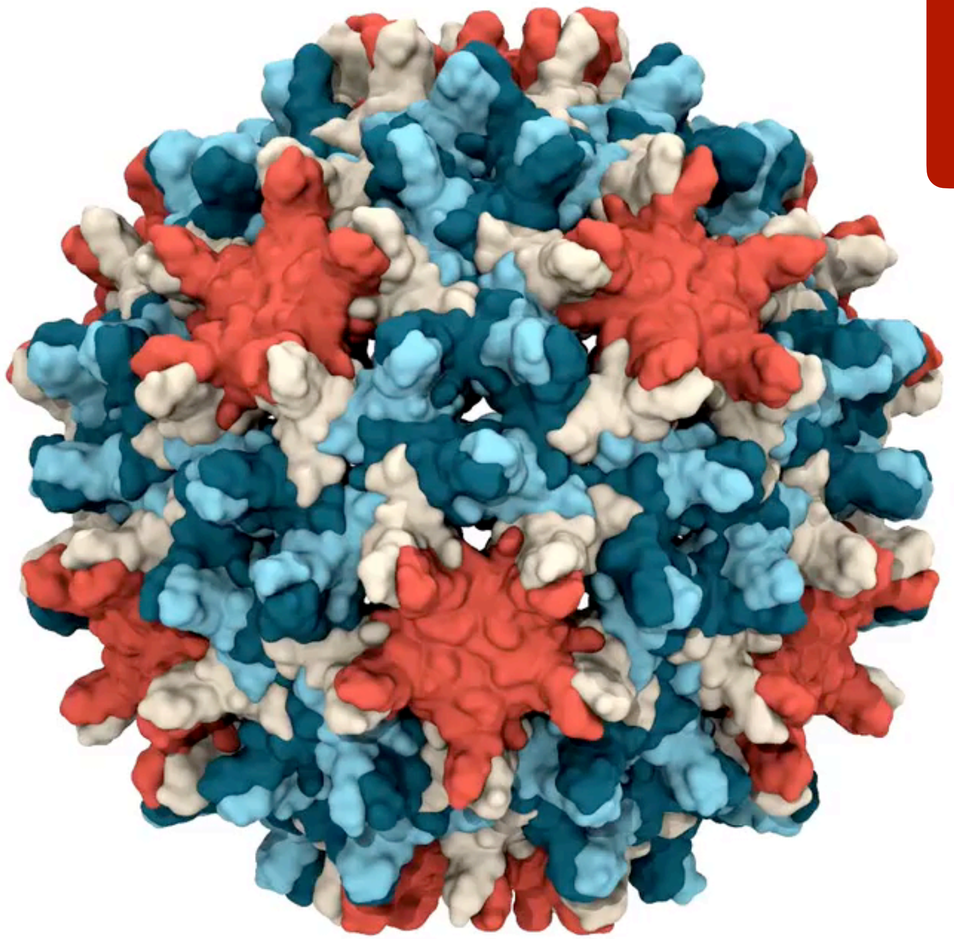
Aggregate digital databases

Morand, Yip, Velegrakis, Lattanzi, Potestio, Tubiana, *arXiv* 2023

Huge amount of generated data:
how to extract insight?

Data are easier and easier to get

All-atom MD simulations



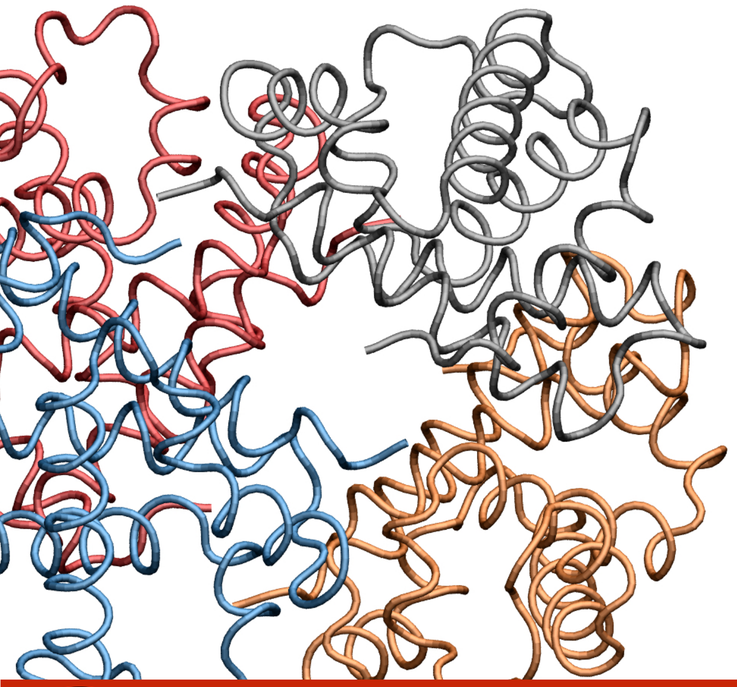
Hadden *et al.*, eLife 2018

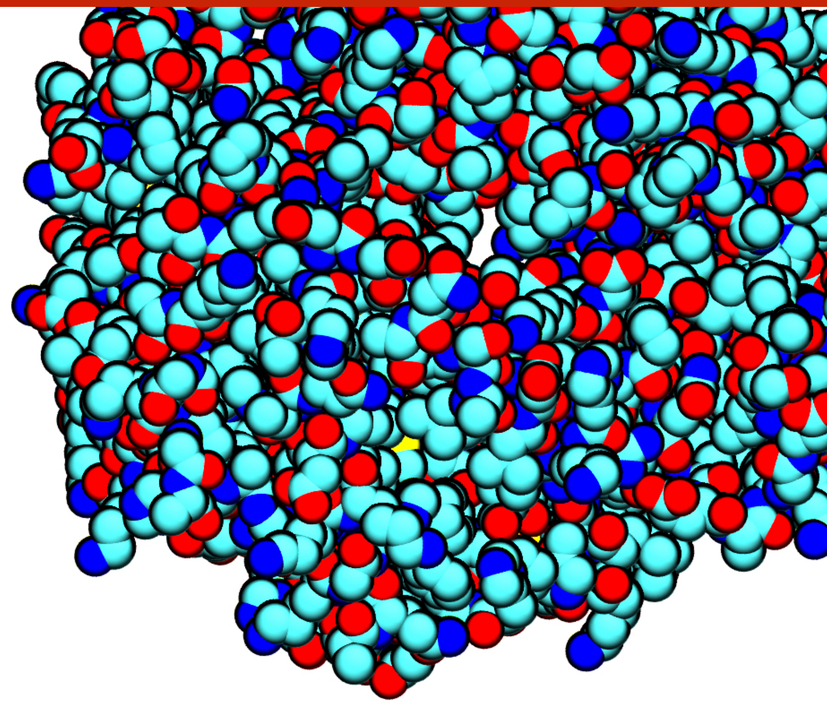
$$\vec{F} = m\vec{a}$$

10^6 atoms for 10^9 configurations of the system

```
ATOM   1153  CD   GLN X 408     110.020  92.423  42.297  0.00  0.00
ATOM   1154  OE1  GLN X 408     110.283  93.074  43.331  0.00  0.00
ATOM   1155  NE2  GLN X 408     108.983  91.618  42.373  0.00  0.00
ATOM   1156 HE21  GLN X 408     108.656  91.204  41.511  0.00  0.00
ATOM   1157 HE22  GLN X 408     108.471  91.768  43.225  0.00  0.00
ATOM   1158  C    GLN X 408     113.167  95.332  40.798  0.00  0.00
ATOM   1159  O    GLN X 408     113.764  95.563  39.715  0.00  0.00
ATOM   1160  N    GLN X 409     113.578  95.654  42.042  0.00  0.00
ATOM   1161  H    GLN X 409     113.018  95.252  42.780  0.00  0.00
ATOM   1162  CA   GLN X 409     114.881  96.418  42.260  0.00  0.00
ATOM   1163  HA   GLN X 409     115.146  96.965  41.356  0.00  0.00
ATOM   1164  CB   GLN X 409     114.444  97.423  43.339  0.00  0.00
ATOM   1165  HB1  GLN X 409     115.314  97.840  43.844  0.00  0.00
ATOM   1166  HB2  GLN X 409     113.900  96.872  44.111  0.00  0.00
ATOM   1167  CG   GLN X 409     113.759  98.671  42.779  0.00  0.00
ATOM   1168  HG1  GLN X 409     112.955  98.388  42.101  0.00  0.00
ATOM   1169  HG2  GLN X 409     114.393  99.253  42.117  0.00  0.00
ATOM   1170  CD   GLN X 409     113.157  99.520  43.821  0.00  0.00
ATOM   1171  OE1  GLN X 409     113.911 100.138  44.580  0.00  0.00
ATOM   1172  NE2  GLN X 409     111.814  99.753  43.795  0.00  0.00
ATOM   1173 HE21  GLN X 409     111.336  99.317  43.023  0.00  0.00
ATOM   1174 HE22  GLN X 409     111.343 100.484  44.321  0.00  0.00
ATOM   1175  C    GLN X 409     116.132  95.610  42.722  0.00  0.00
ATOM   1176  O    GLN X 409     117.199  96.174  42.994  0.00  0.00
ATOM   1177  N    PHE X 410     115.988  94.277  42.829  0.00  0.00
ATOM   1178  H    PHE X 410     115.077  93.891  42.653  0.00  0.00
ATOM   1179  CA   PHE X 410     116.976  93.435  43.590  0.00  0.00
ATOM   1180  HA   PHE X 410     117.124  93.938  44.557  0.00  0.00
ATOM   1181  CB   PHE X 410     116.299  92.042  43.824  0.00  0.00
ATOM   1182  HB1  PHE X 410     116.161  91.560  42.866  0.00  0.00
ATOM   1183  HB2  PHE X 410     115.259  92.135  44.154  0.00  0.00
ATOM   1184  CG   PHE X 410     117.011  91.074  44.747  0.00  0.00
ATOM   1185  CD1  PHE X 410     117.099  91.369  46.110  0.00  0.00
ATOM   1186  HD1  PHE X 410     116.520  92.142  46.584  0.00  0.00
ATOM   1187  CE1  PHE X 410     117.746  90.542  46.998  0.00  0.00
ATOM   1188  HE1  PHE X 410     117.720  90.932  48.007  0.00  0.00
ATOM   1189  CZ   PHE X 410     118.393  89.398  46.545  0.00  0.00
ATOM   1190  HZ   PHE X 410     118.896  88.709  47.208  0.00  0.00
ATOM   1191  CE2  PHE X 410     118.374  89.090  45.183  0.00  0.00
ATOM   1192  HE2  PHE X 410     118.911  88.260  44.740  0.00  0.00
ATOM   1193  CD2  PHE X 410     117.779  89.966  44.293  0.00  0.00
ATOM   1194  HD2  PHE X 410     117.894  89.832  43.227  0.00  0.00
ATOM   1195  C    PHE X 410     118.379  93.352  42.972  0.00  0.00
```

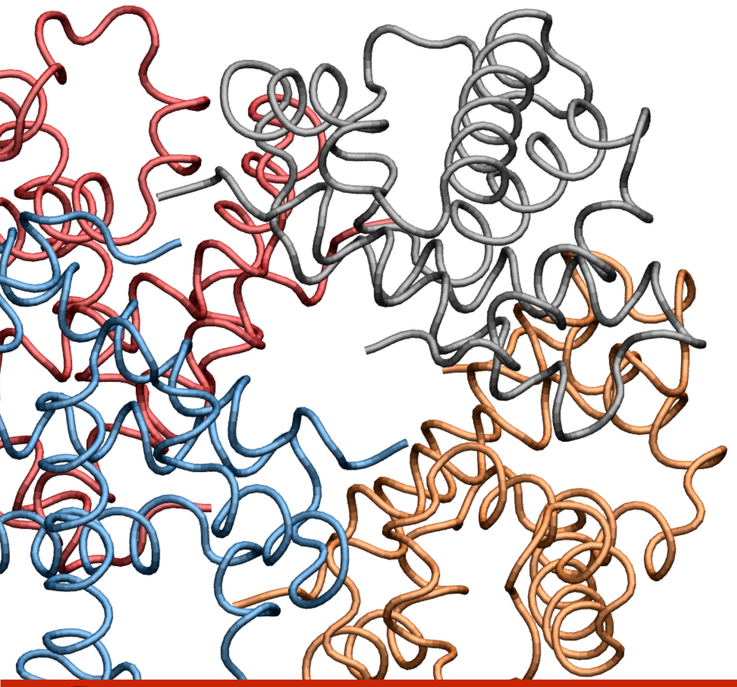
Simplify to understand

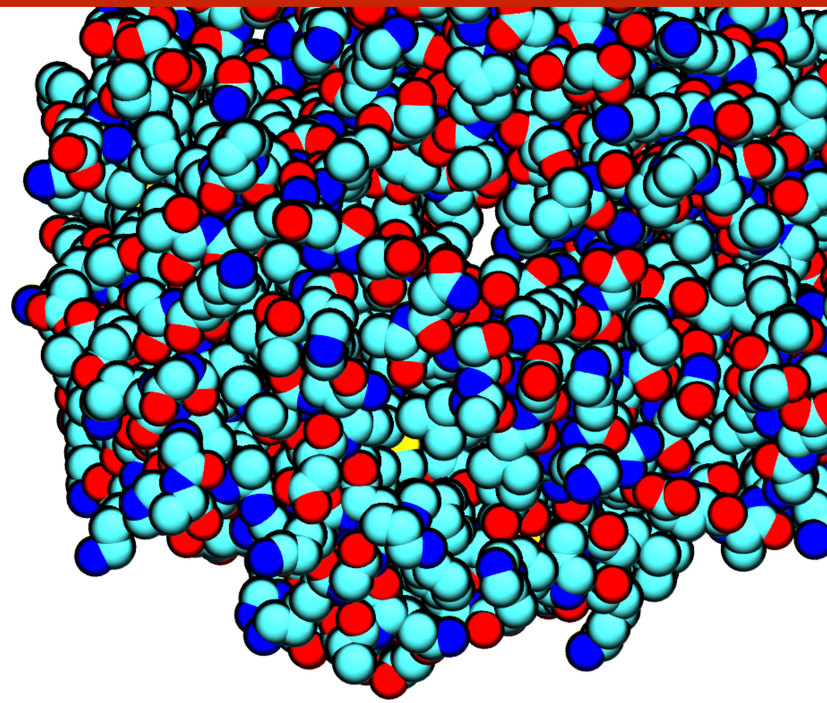




Simplify to understand

DATA != INFORMATION

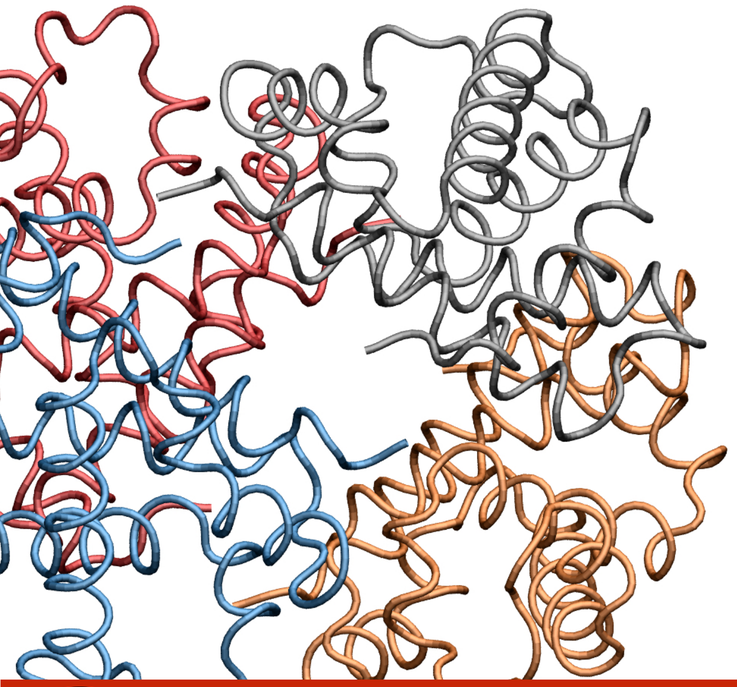


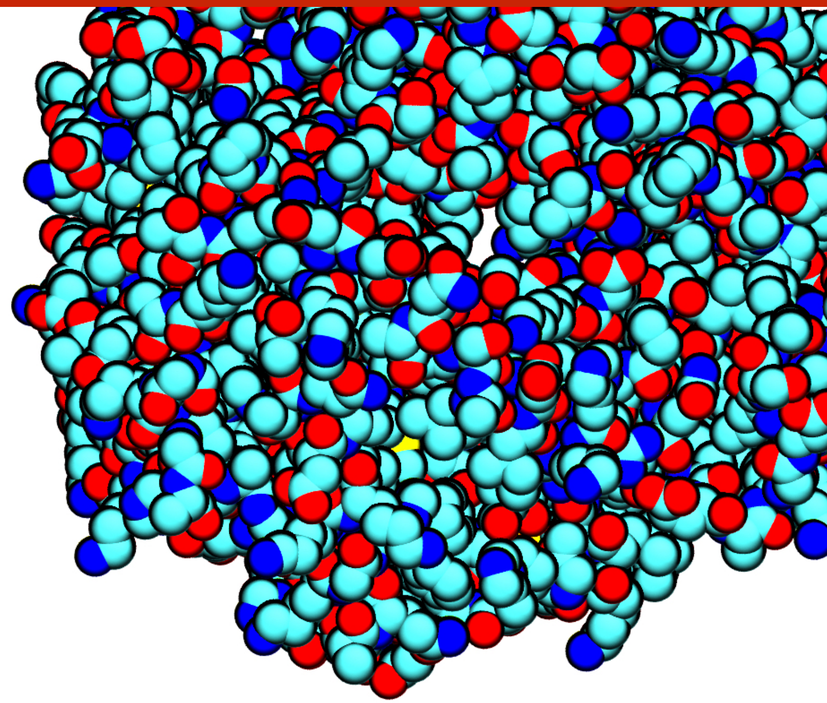


Simplify to understand

DATA != INFORMATION

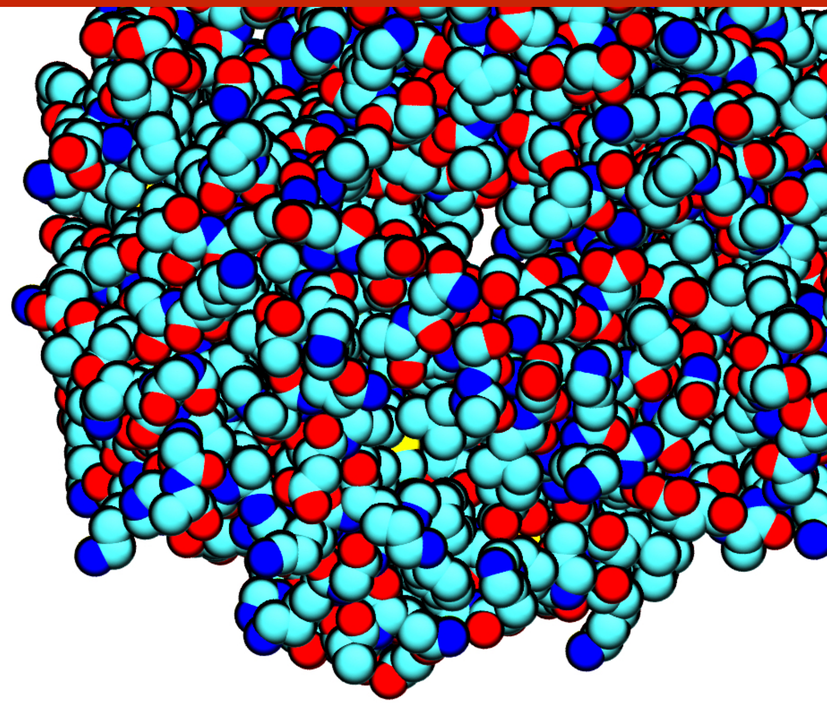
Simplify "just right"
to understand
without losing information



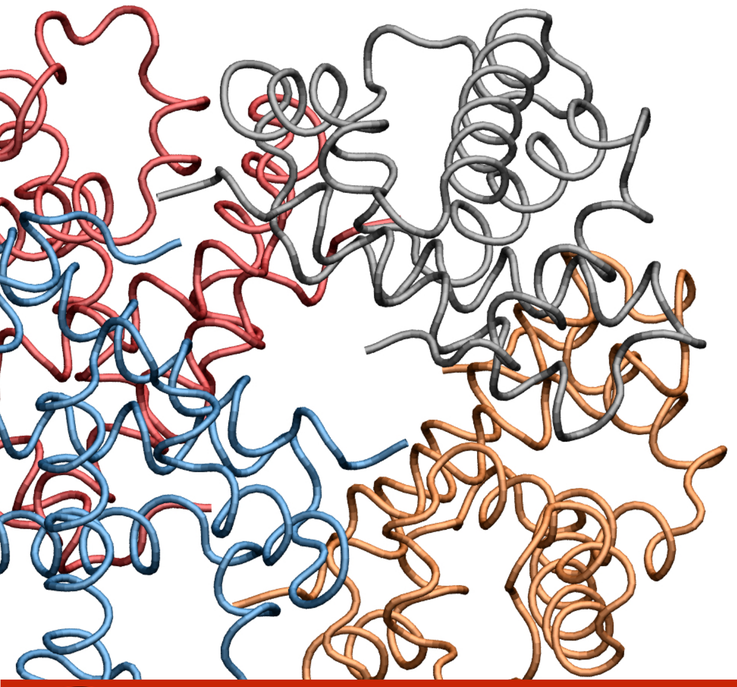


Simplify to understand

DATA != INFORMATION



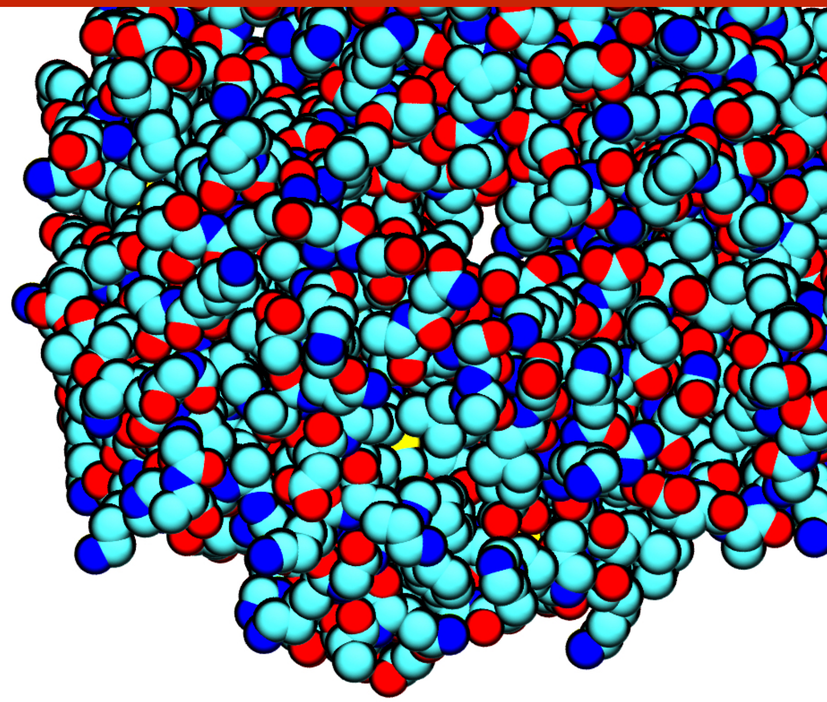
Simplify "just right"
to understand
without losing information

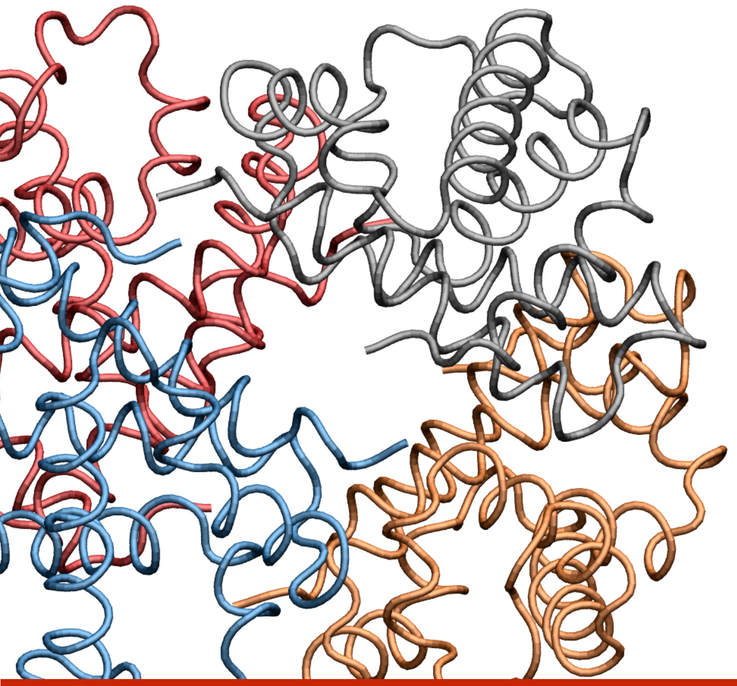


Dimensionality reduction:
in terms of what variables?

Simplify to understand

```
ATOM   1153  CD   GLN X 408     110.020  92.423  42.297  0.00  0.00
ATOM   1154  OE1  GLN X 408     110.283  93.074  43.331  0.00  0.00
ATOM   1155  NE2  GLN X 408     108.983  91.618  42.373  0.00  0.00
ATOM   1156 HE21  GLN X 408     108.656  91.204  41.511  0.00  0.00
ATOM   1157 HE22  GLN X 408     108.471  91.768  43.225  0.00  0.00
ATOM   1158  C    GLN X 408     113.167  95.332  40.798  0.00  0.00
ATOM   1159  O    GLN X 408     113.764  95.563  39.715  0.00  0.00
ATOM   1160  N    GLN X 409     113.578  95.654  42.042  0.00  0.00
ATOM   1161  H    GLN X 409     113.018  95.252  42.780  0.00  0.00
ATOM   1162  CA   GLN X 409     114.881  96.418  42.260  0.00  0.00
ATOM   1163  HA   GLN X 409     115.146  96.965  41.356  0.00  0.00
ATOM   1164  CB   GLN X 409     114.444  97.423  43.339  0.00  0.00
ATOM   1165  HB1  GLN X 409     115.314  97.840  43.844  0.00  0.00
ATOM   1166  HB2  GLN X 409     113.900  96.872  44.111  0.00  0.00
```

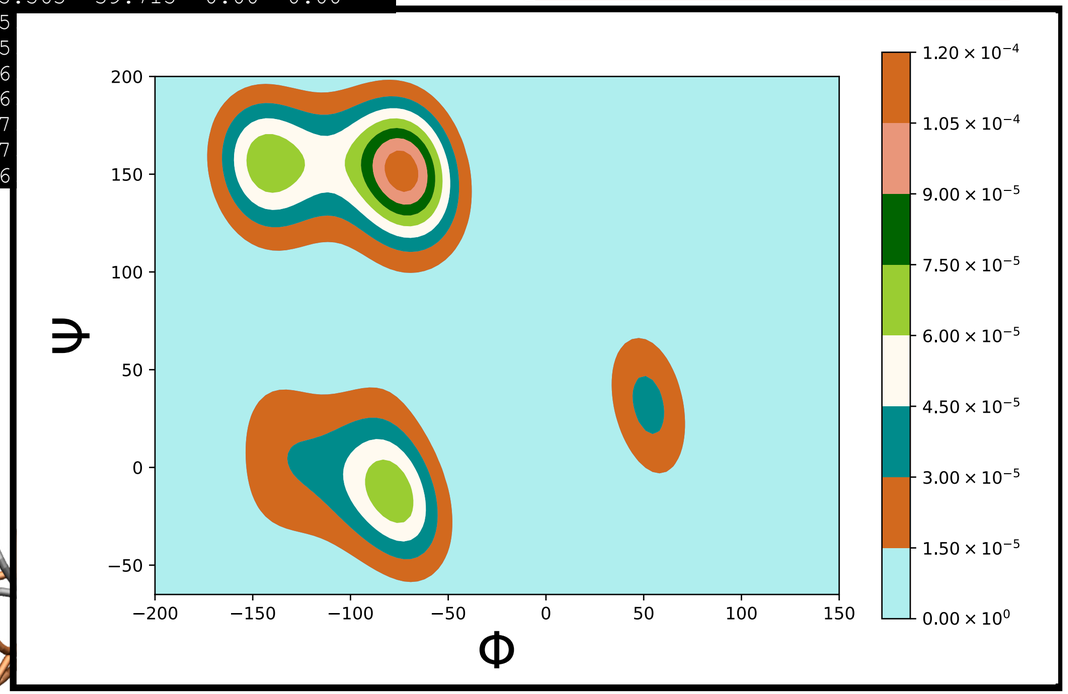


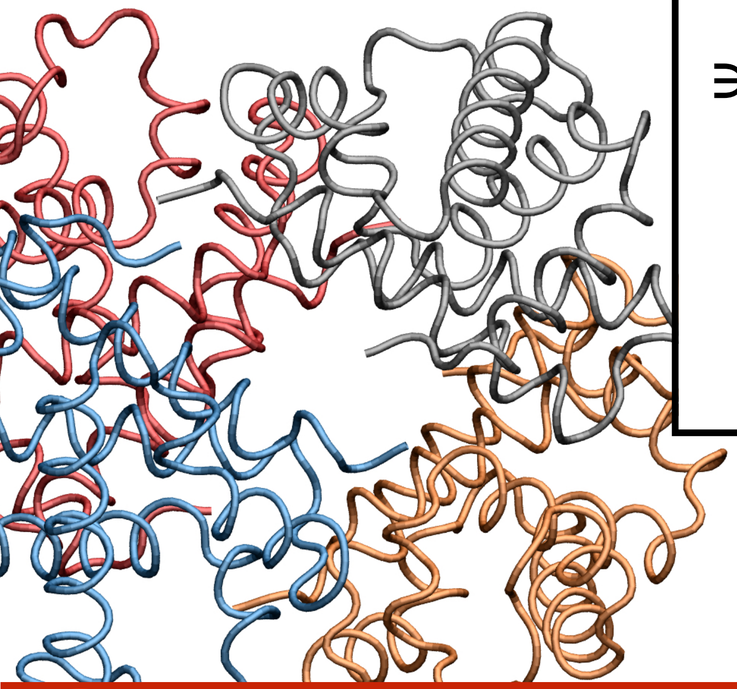


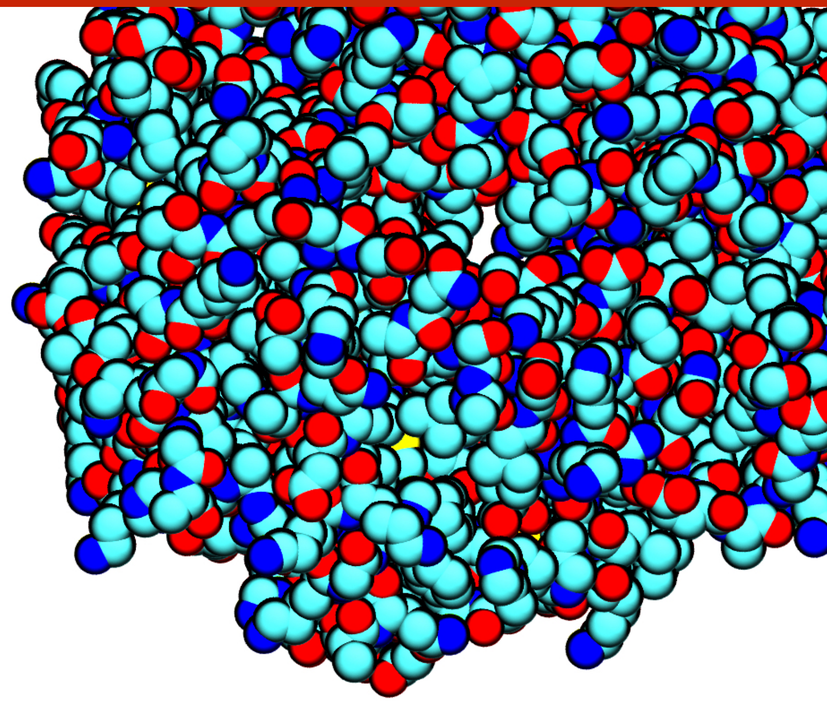
Simplify to understand

```
ATOM   1153  CD   GLN X 408     110.020  92.423  42.297  0.00  0.00
ATOM   1154  OE1  GLN X 408     110.283  93.074  43.331  0.00  0.00
ATOM   1155  NE2  GLN X 408     108.983  91.618  42.373  0.00  0.00
ATOM   1156 HE21  GLN X 408     108.656  91.204  41.511  0.00  0.00
ATOM   1157 HE22  GLN X 408     108.471  91.768  43.225  0.00  0.00
ATOM   1158  C    GLN X 408     113.167  95.332  40.798  0.00  0.00
ATOM   1159  O    GLN X 408     113.764  95.563  39.715  0.00  0.00
ATOM   1160  N    GLN X 409     113.578  95
ATOM   1161  H    GLN X 409     113.018  95
ATOM   1162  CA   GLN X 409     114.881  96
ATOM   1163  HA   GLN X 409     115.146  96
ATOM   1164  CB   GLN X 409     114.444  97
ATOM   1165  HB1  GLN X 409     115.314  97
ATOM   1166  HB2  GLN X 409     113.900  96
```

Collective variables



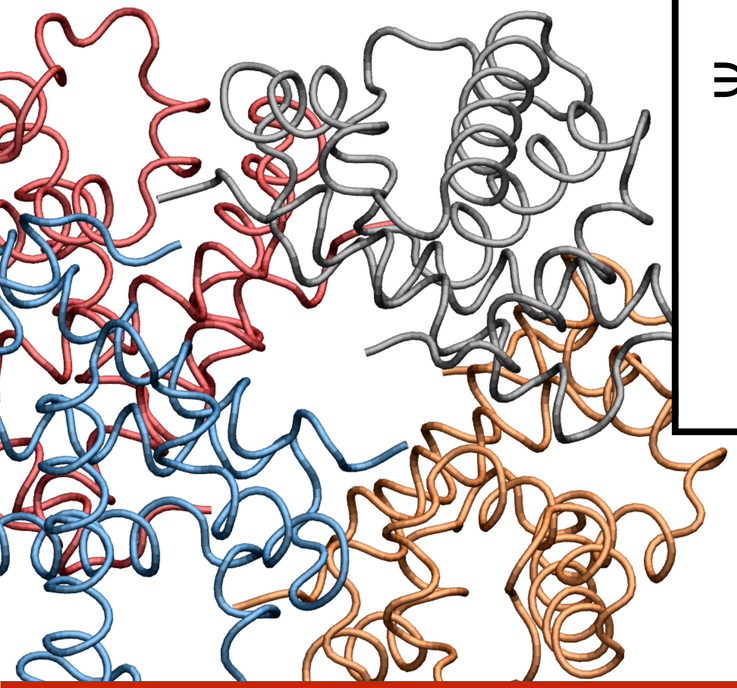


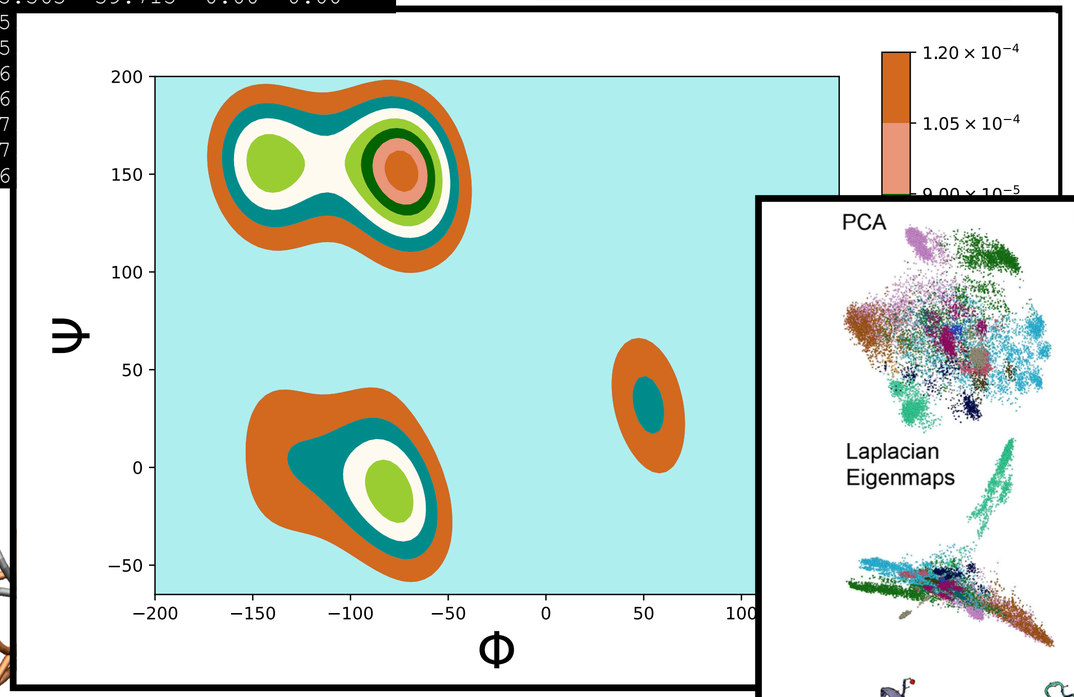


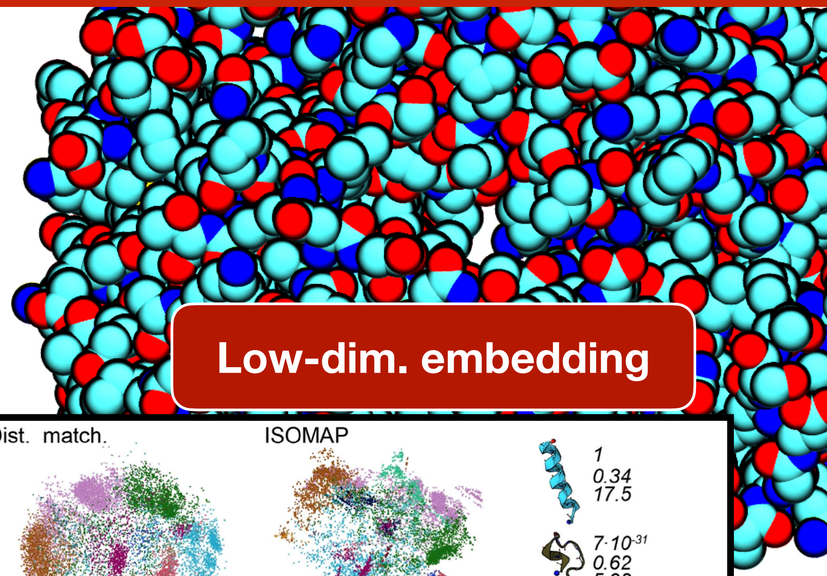
Simplify to understand

```
ATOM   1153  CD   GLN X 408     110.020  92.423  42.297  0.00  0.00
ATOM   1154  OE1  GLN X 408     110.283  93.074  43.331  0.00  0.00
ATOM   1155  NE2  GLN X 408     108.983  91.618  42.373  0.00  0.00
ATOM   1156 HE21  GLN X 408     108.656  91.204  41.511  0.00  0.00
ATOM   1157 HE22  GLN X 408     108.471  91.768  43.225  0.00  0.00
ATOM   1158  C    GLN X 408     113.167  95.332  40.798  0.00  0.00
ATOM   1159  O    GLN X 408     113.764  95.563  39.715  0.00  0.00
ATOM   1160  N    GLN X 409     113.578  95
ATOM   1161  H    GLN X 409     113.018  95
ATOM   1162  CA   GLN X 409     114.881  96
ATOM   1163  HA   GLN X 409     115.146  96
ATOM   1164  CB   GLN X 409     114.444  97
ATOM   1165  HB1  GLN X 409     115.314  97
ATOM   1166  HB2  GLN X 409     113.900  96
```

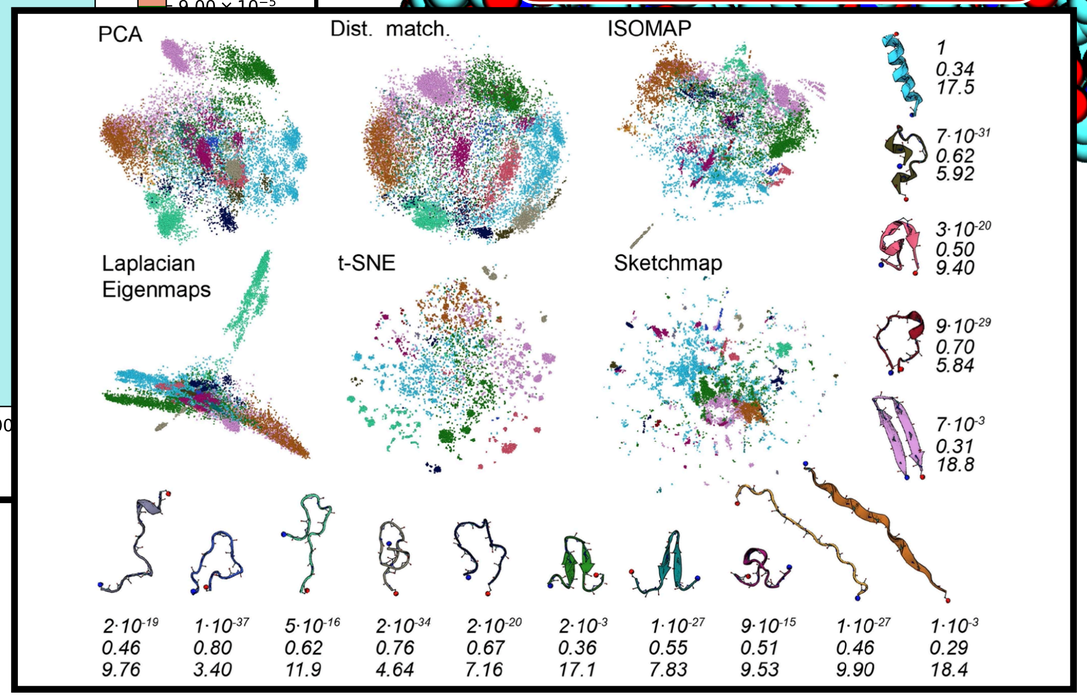
Collective variables







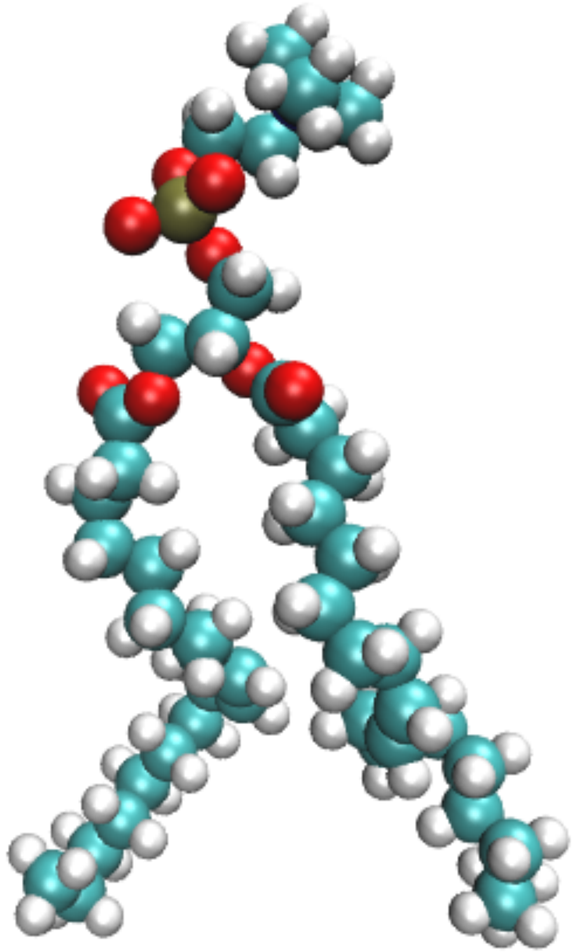
Low-dim. embedding



Tribello, Gasparotto, Front. Mol. Biosci. 2019

Coarse-graining (CG)

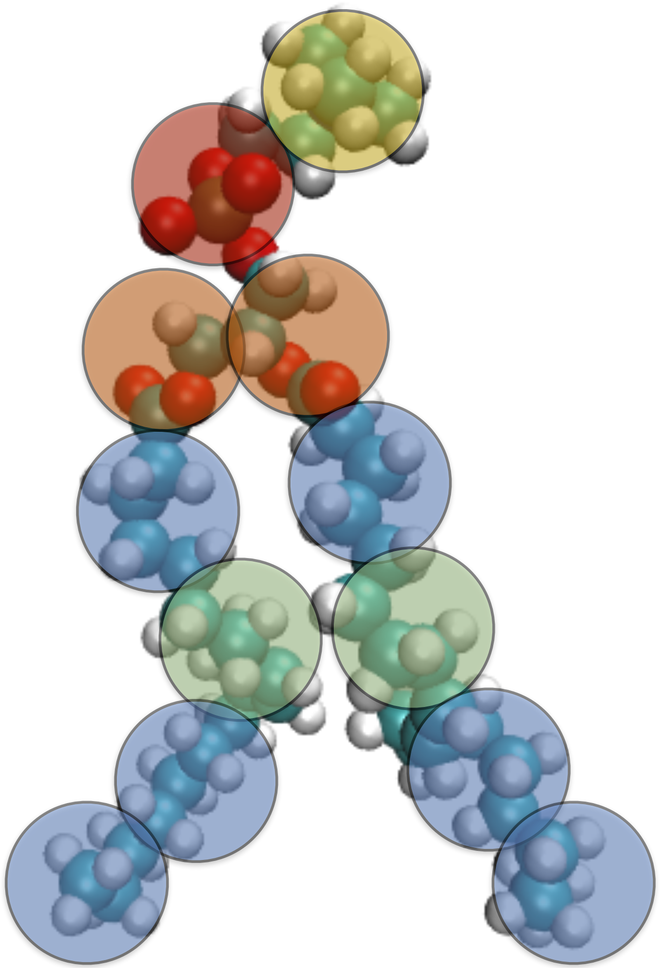
AA representation



Rudzinski, Noid, J. Chem. Phys. 2011

Coarse-graining (CG)

CG representation

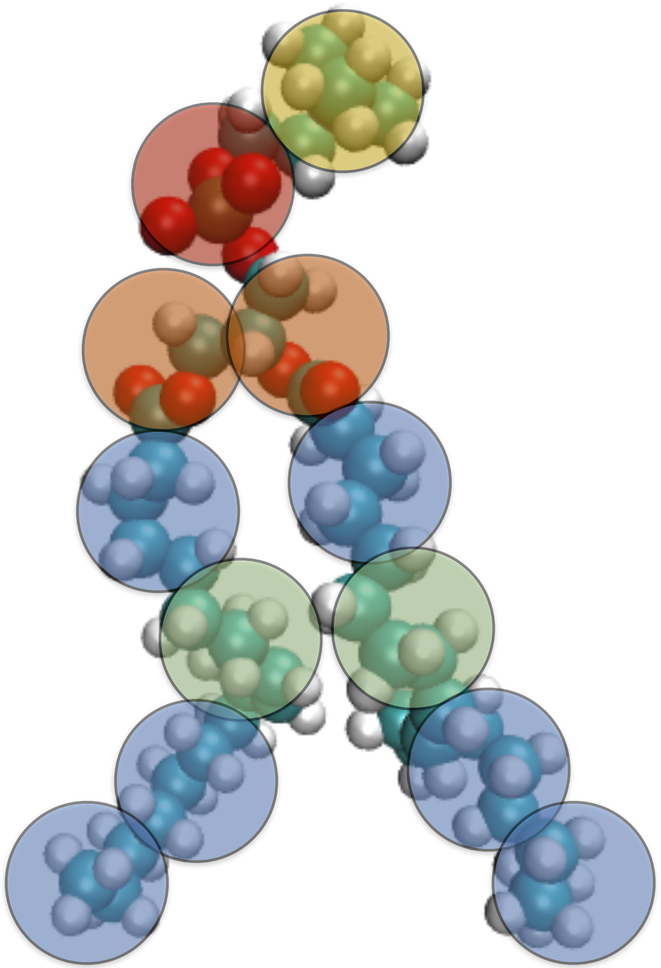


Describe the system in terms of a **reduced** number of particles: **CG filter**

Rudzinski, Noid, J. Chem. Phys. 2011

Coarse-graining (CG)

CG representation



Describe the system in terms of a **reduced** number of particles: **CG filter**

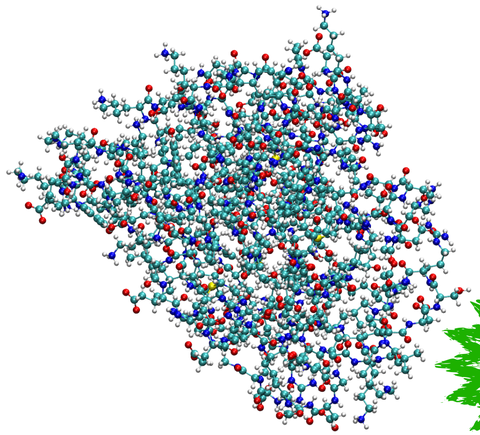
Degree of detail employed can be **modulated**

Intuitive interpretation of the outcome

Rudzinski, Noid, J. Chem. Phys. 2011

CG filters

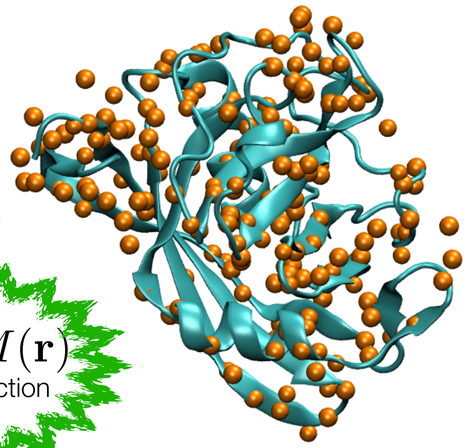
High-resolution model



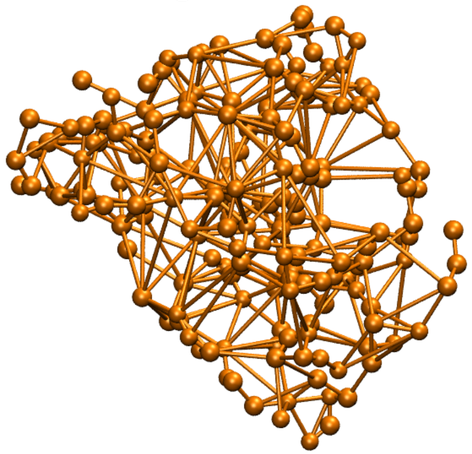
$$\mathbf{R}_I = M(\mathbf{r})$$

Mapping function

Choice of the sites

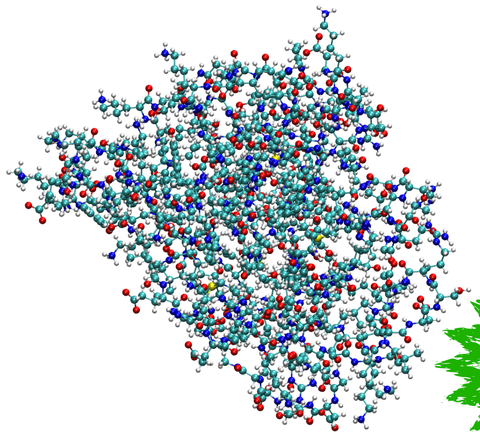


Coarse-grained model

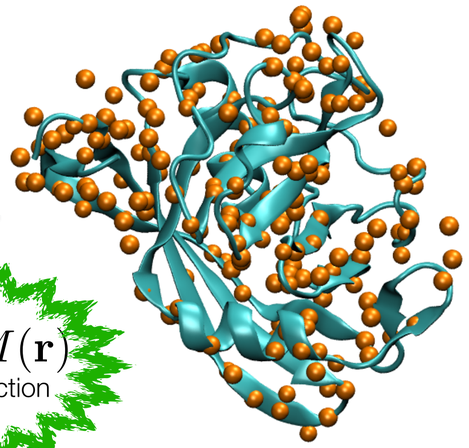


CG filters

High-resolution model



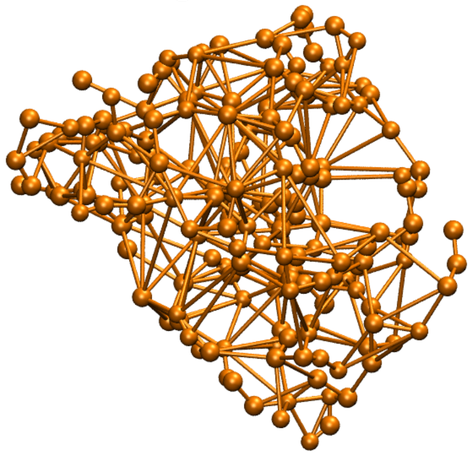
Choice of the sites

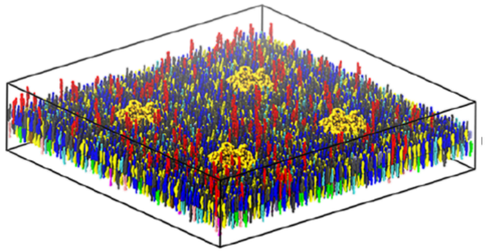


$$\mathbf{R}_I = M(\mathbf{r})$$

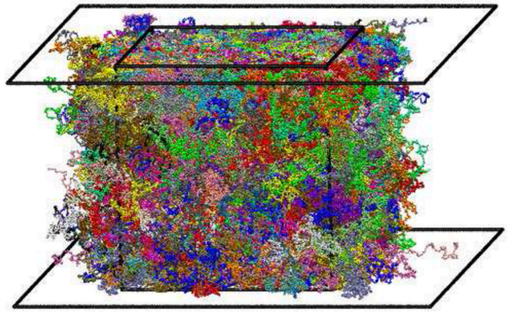
Mapping function

Coarse-grained model

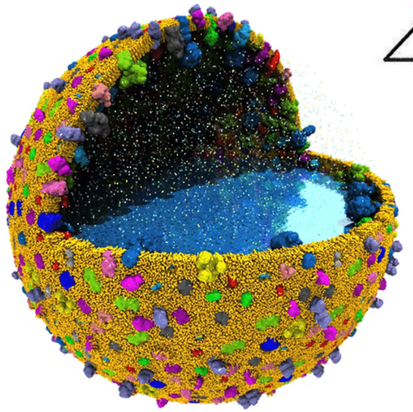




Corradi *et al*., ACS Cent. Sci. 2018

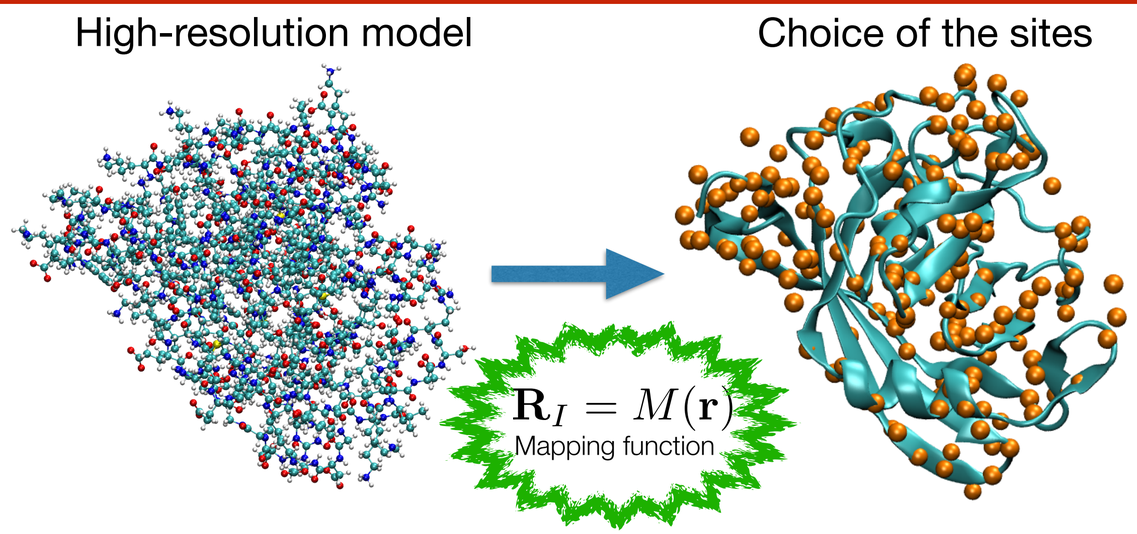


Hsu et *al.,* J. Chem. Phys. *2020*



Vermaas, *et al*., J. Chem. Inf. Model. 2022

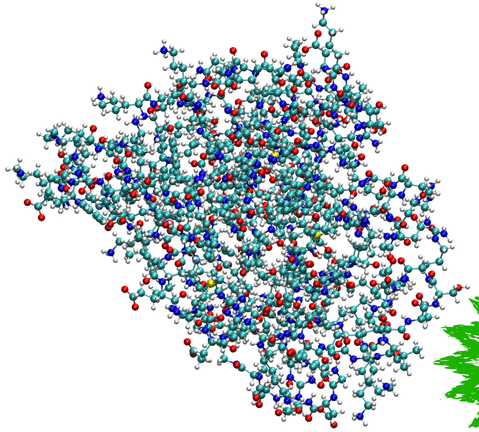
CG filters



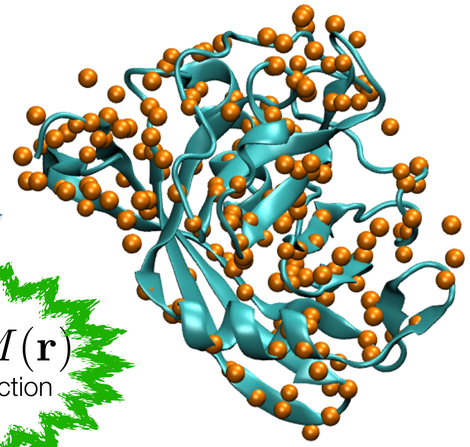
Giulini, Rigoli, RM *et al.*, Front. Mol. Biosci. 2021

CG filters

High-resolution model

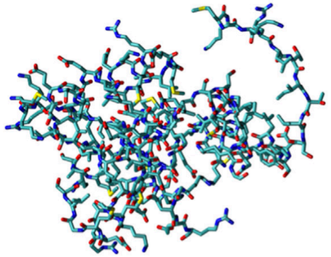


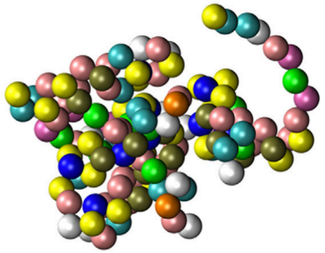
Choice of the sites



$$\mathbf{R}_I = M(\mathbf{r})$$

Mapping function



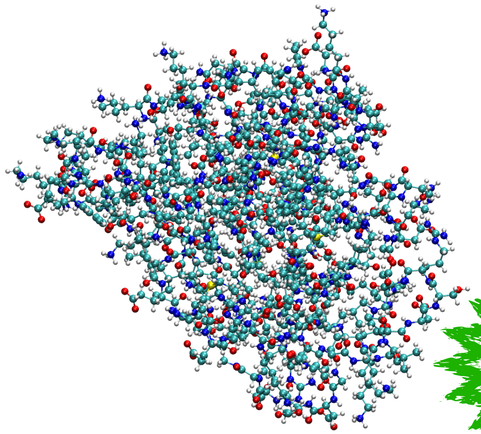


Observe the system with a pair of CG glasses: **information loss**

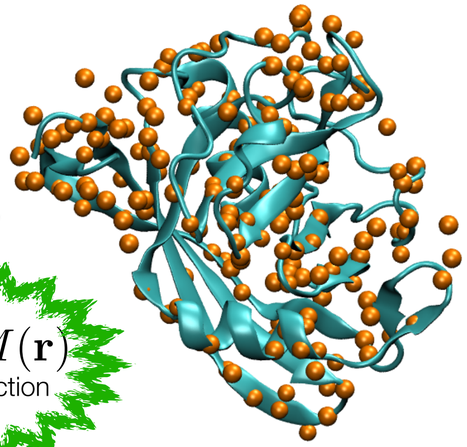
Giulini, Rigoli, RM *et al.*, Front. Mol. Biosci. 2021

CG filters

High-resolution model

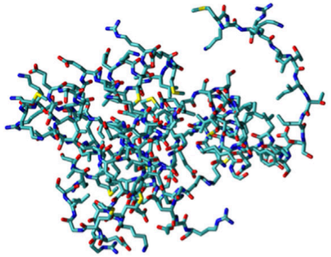


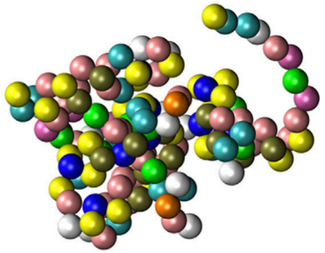
Choice of the sites



$$\mathbf{R}_I = M(\mathbf{r})$$

Mapping function





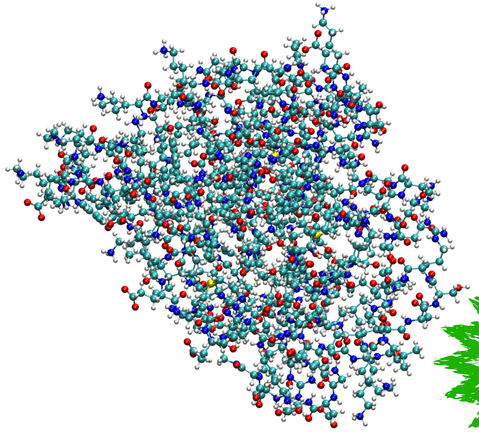
Observe the system with a pair of CG glasses: **information loss**

Optimize the glasses to extract relevant information from the data

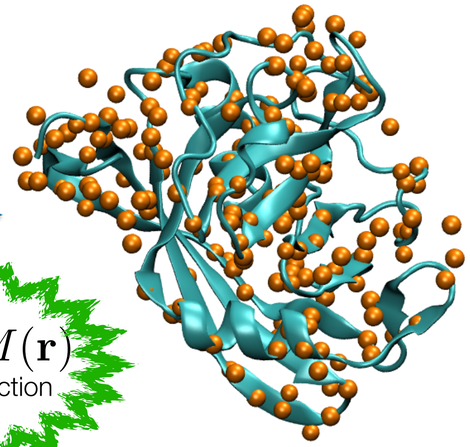
Giulini, Rigoli, RM *et al.*, Front. Mol. Biosci. 2021

CG filters

High-resolution model

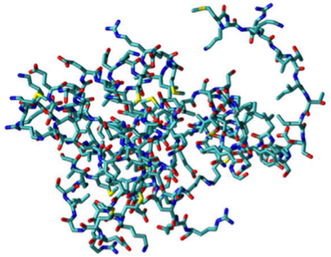


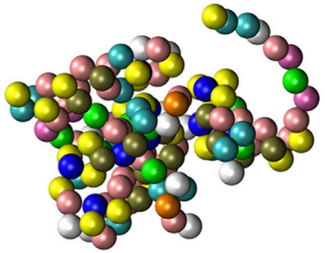
Choice of the sites



$$\mathbf{R}_I = M(\mathbf{r})$$

Mapping function

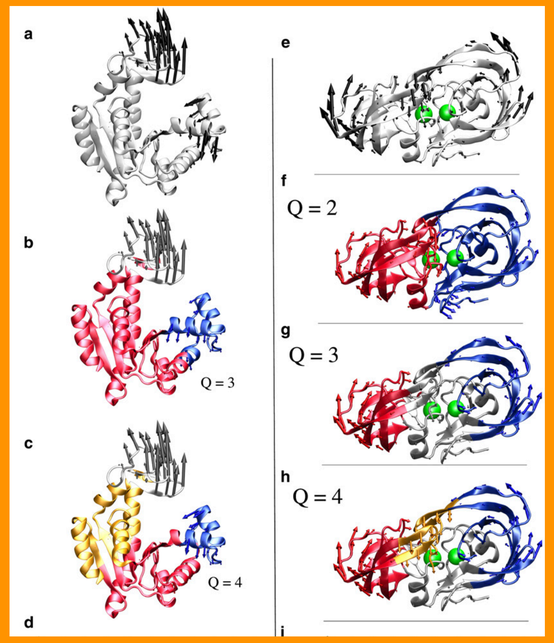




Observe the system with a pair of CG glasses: **information loss**

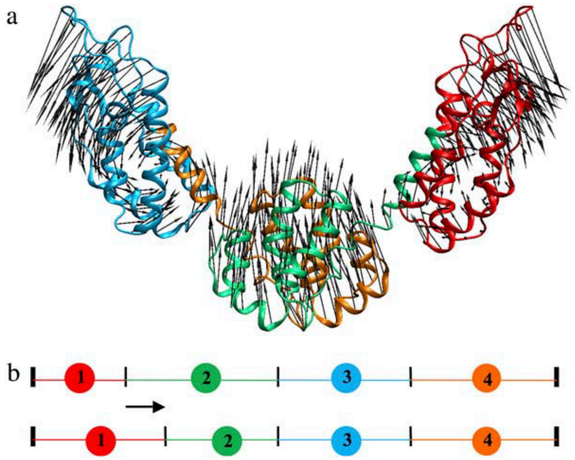
Optimize the glasses to extract relevant information from the data

Coarse-grained filters to **understand** a property or behaviour



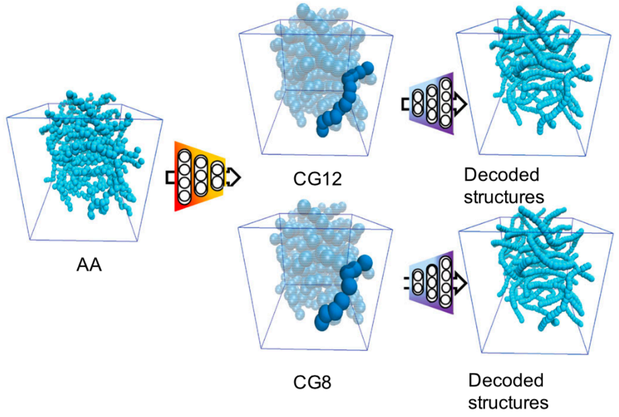
Giulini, Rigoli, RM *et al.*, Front. Mol. Biosci. 2021

CG filters

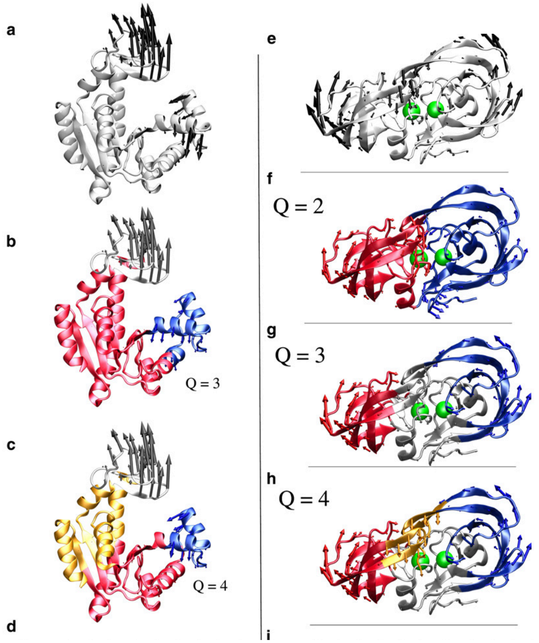


$$\chi^2 = \frac{1}{3N}\sum_{I=1}^{N}\frac{1}{n_t}\sum_{t=1}^{n_t}\left(\sum_{i\in I}\sum_{j\geq i\in I}|\Delta\mathbf{r}_i^{ED}(t)-\Delta\mathbf{r}_j^{ED}(t)|^2\right)$$

Zhang, Lu, Noid, Krishna, Pfaendtner, Voth
BJ 2008

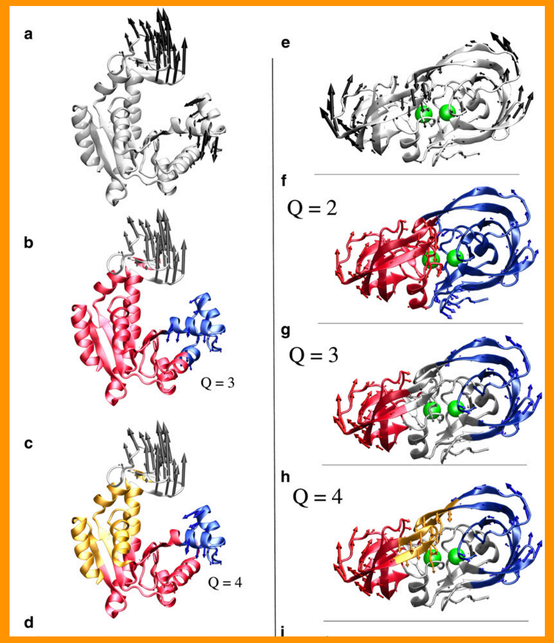


Wang, Gomez-Bombarelli
NPJ Comput. Mater. 2019



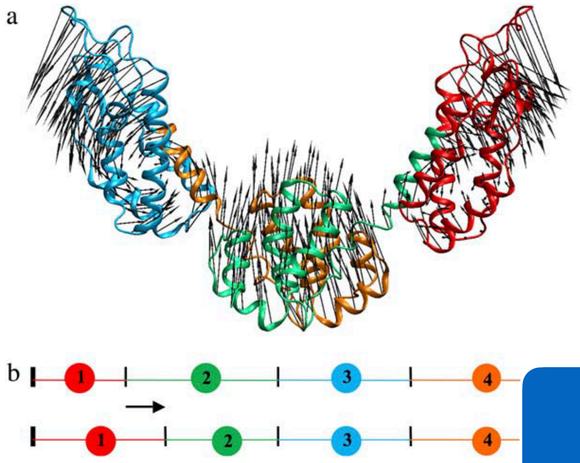
Potestio, Pontiggia, Micheletti
BJ 2009

Coarse-grained filters to
understand
a property or behaviour



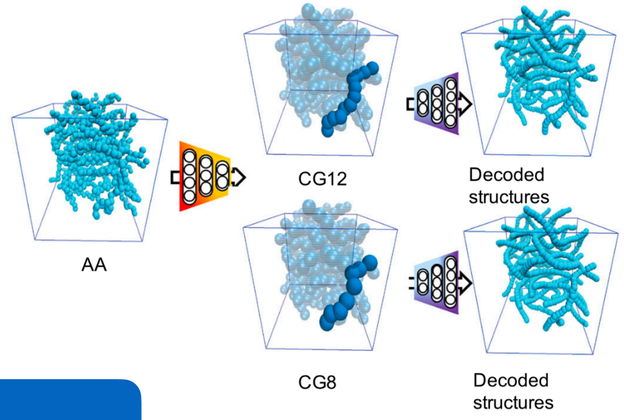
Giulini, Rigoli, RM *et al.*, Front. Mol. Biosci. 2021

CG filters



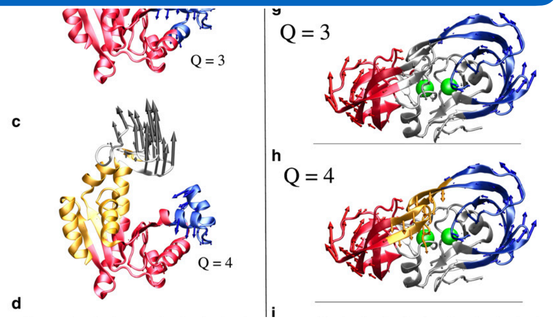
$$\chi^2 = \frac{1}{3N}\sum_{I=1}^{N}\frac{1}{n_t}\sum_{t=1}^{n_t}\left(\sum_{i\in I}\sum_{j\geq i\in I}|\Delta\mathbf{r}_i^{ED}(t) - \Delta\mathbf{r}_j^{ED}(t$$

Zhang, Lu, Noid, Krishna, Pfaendtner, Vot
BJ 2008



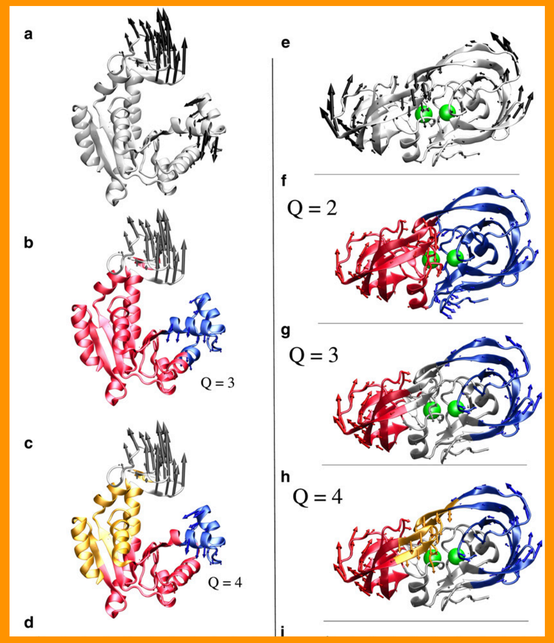
Wang, Gomez-Bombarelli
NPJ Comput. Mater. 2019

Optimize the CG filter based on quantitative criteria: In our case, **minimize the information loss arising from coarse-graining**



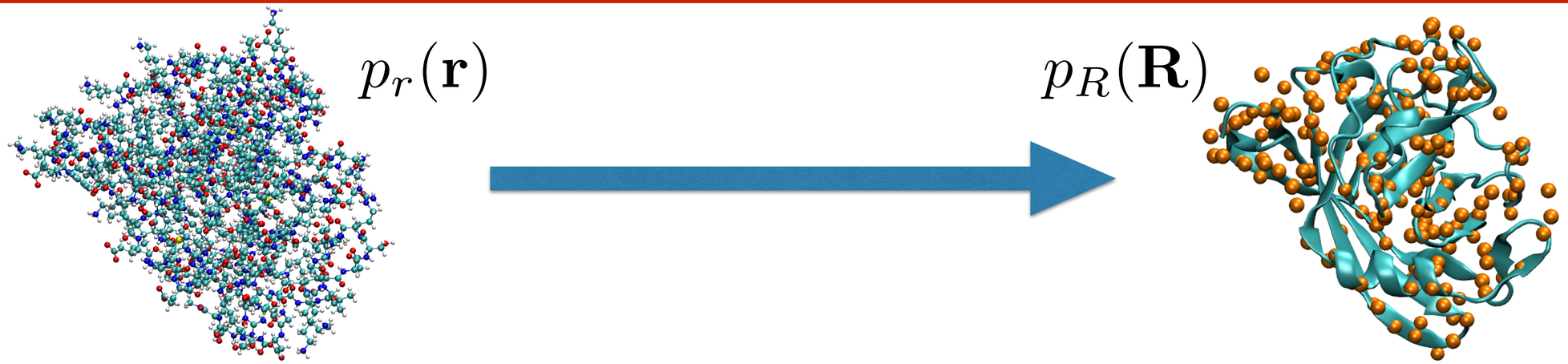
Potestio, Pontiggia, Micheletti
BJ 2009

Coarse-grained filters to **understand** a property or behaviour



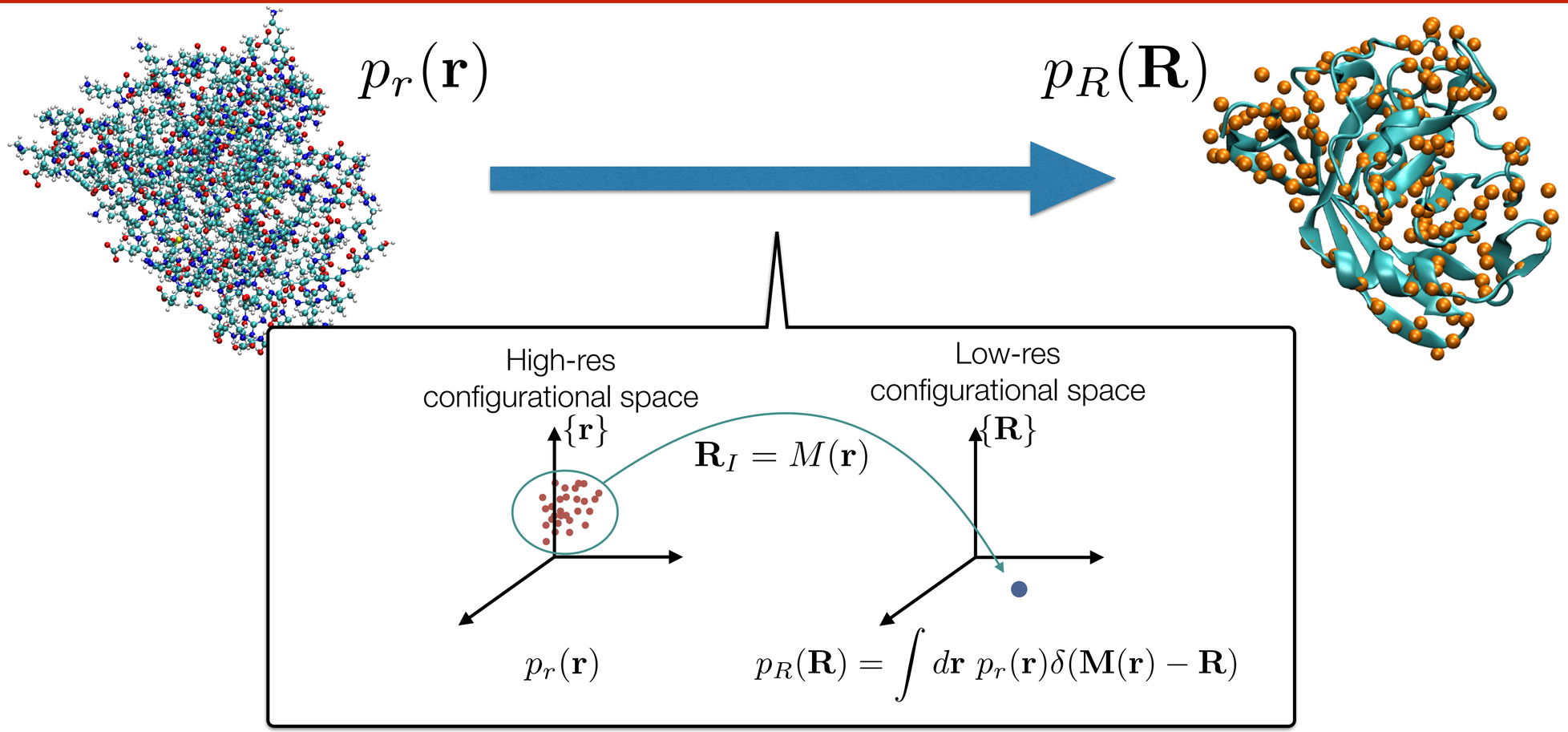
Giulini, Rigoli, RM *et al.*, Front. Mol. Biosci. 2021

Information loss in coarse-graining



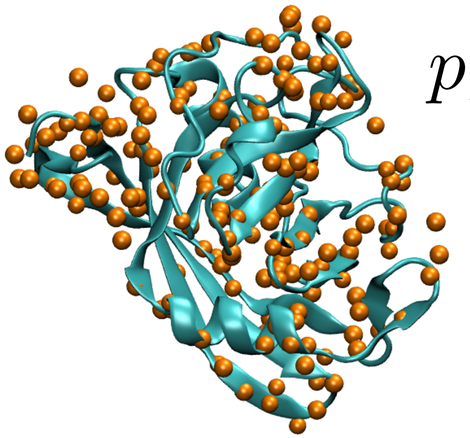
Rudzinski, Noid, J. Chem. Phys. 2011
Giulini, RM, Shell, Potestio, JCTC 2020

Information loss in coarse-graining



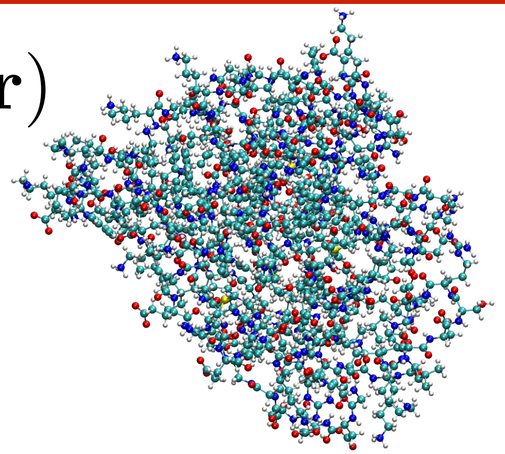
Rudzinski, Noid, J. Chem. Phys. 2011
Giulini, RM, Shell, Potestio, JCTC 2020

Information loss in coarse-graining



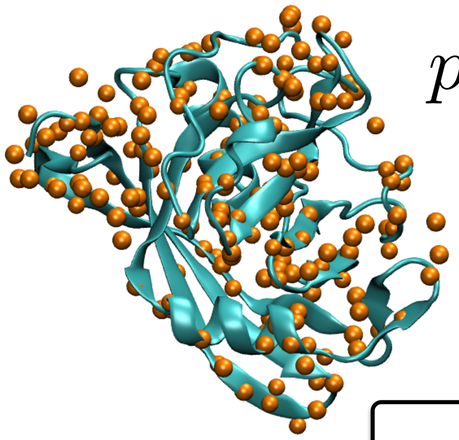
$$p_R(\mathbf{R})$$

$$\bar{p}_r(\mathbf{r}) \neq p_r(\mathbf{r})$$



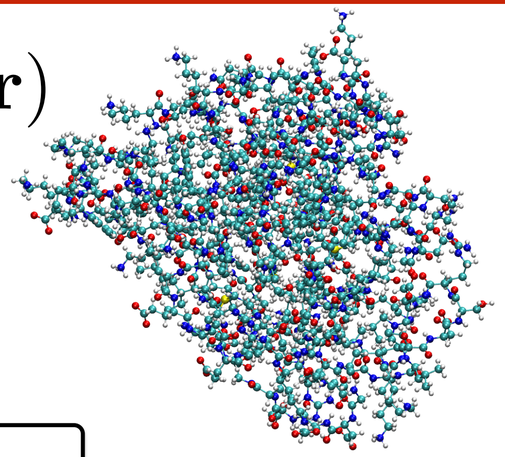
Rudzinski, Noid, J. Chem. Phys. 2011
Giulini, RM, Shell, Potestio, JCTC 2020

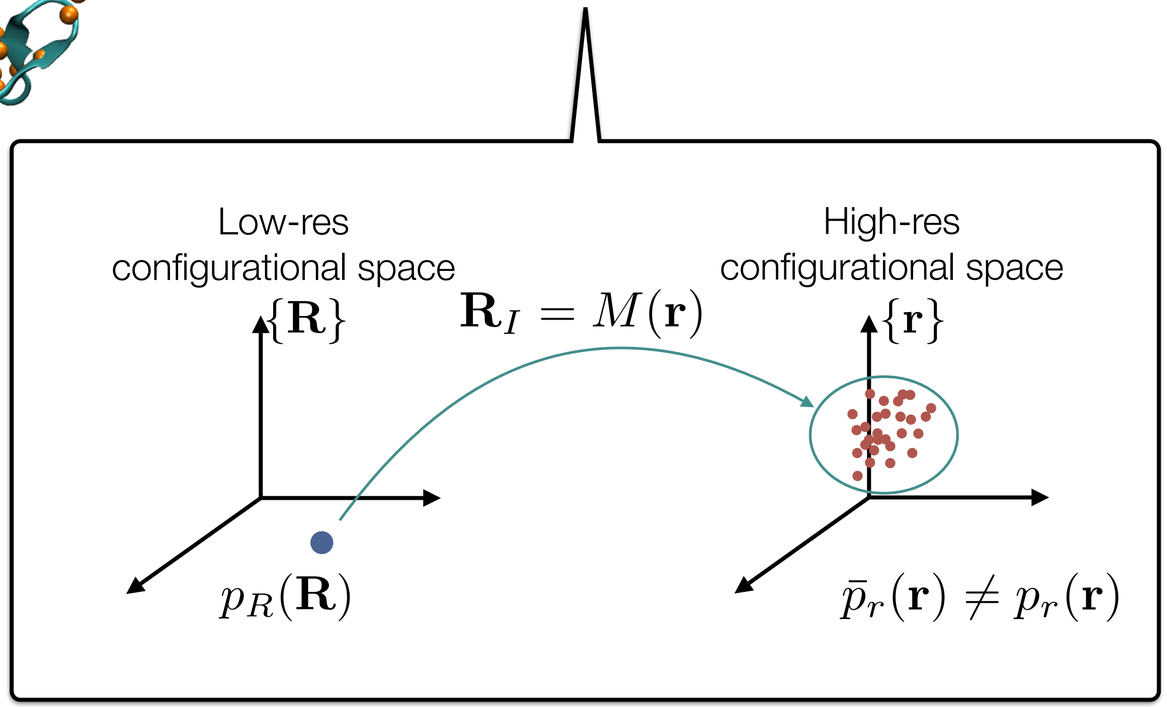
Information loss in coarse-graining



$p_R(\mathbf{R})$

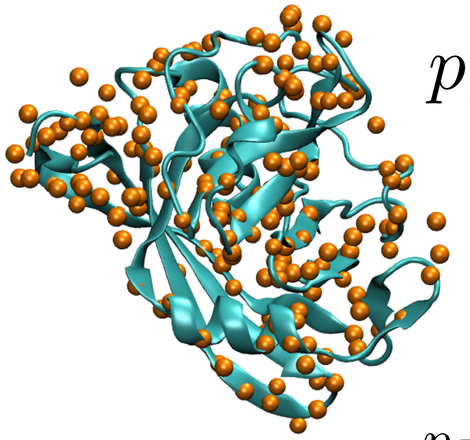
$\bar{p}_r(\mathbf{r}) \neq p_r(\mathbf{r})$





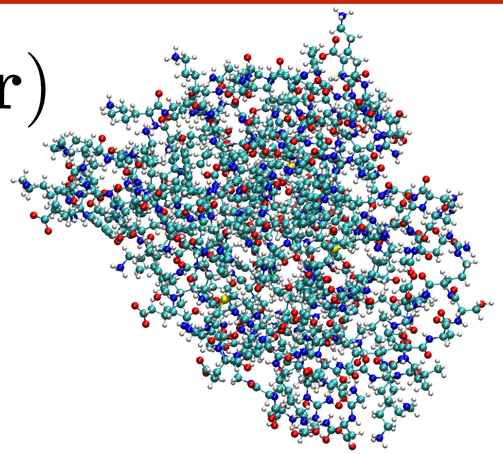
Rudzinski, Noid, J. Chem. Phys. 2011
Giulini, RM, Shell, Potestio, JCTC 2020

Information loss in coarse-graining



$$p_R(\mathbf{R})$$

$$\bar{p}_r(\mathbf{r}) \neq p_r(\mathbf{r})$$

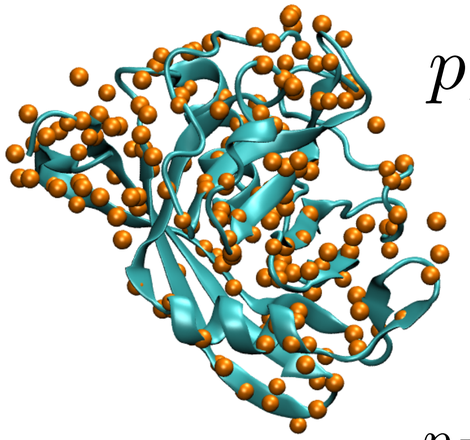


$$\bar{p}_r(\mathbf{r}) = \frac{p_R(\mathbf{M}(\mathbf{r}))}{\Sigma(\mathbf{M}(\mathbf{r}))}$$

Each microstate **r** receives the average probability of the macrostate **R** it maps onto

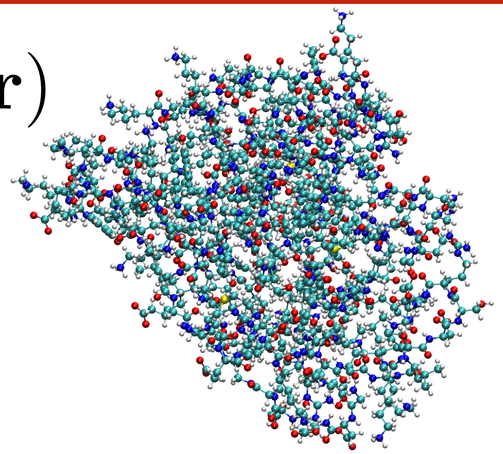
Rudzinski, Noid, J. Chem. Phys. 2011
Giulini, RM, Shell, Potestio, JCTC 2020

Information loss in coarse-graining



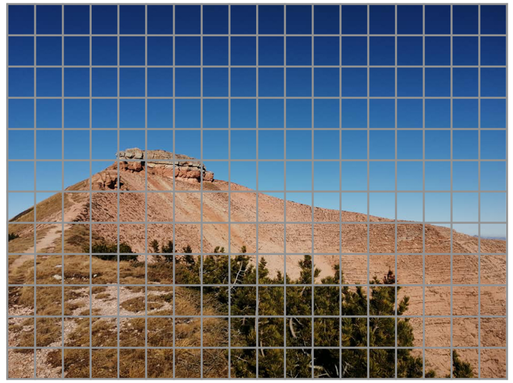
$$p_R(\mathbf{R})$$

$$\bar{p}_r(\mathbf{r}) \neq p_r(\mathbf{r})$$

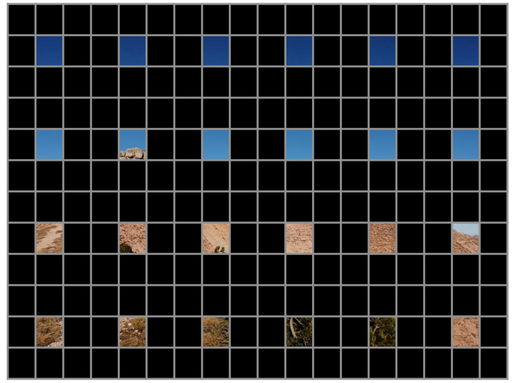


$$\bar{p}_r(\mathbf{r}) = \frac{p_R(\mathbf{M}(\mathbf{r}))}{\Sigma(\mathbf{M}(\mathbf{r}))}$$

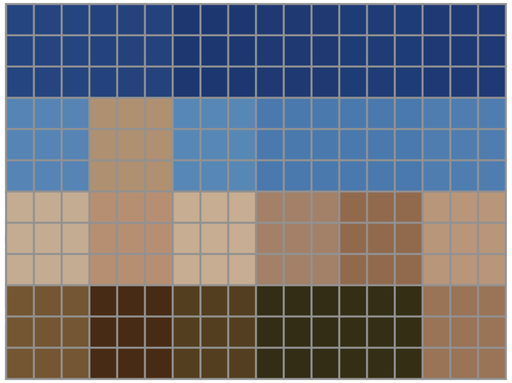
Each microstate **r** receives the average probability of the macrostate **R** it maps onto



Original high resolution representation



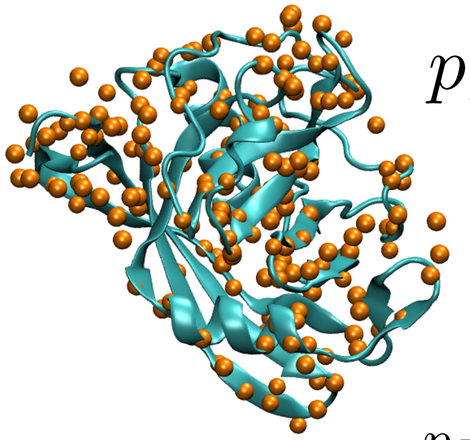
Filtering



Reconstructed high resolution representation

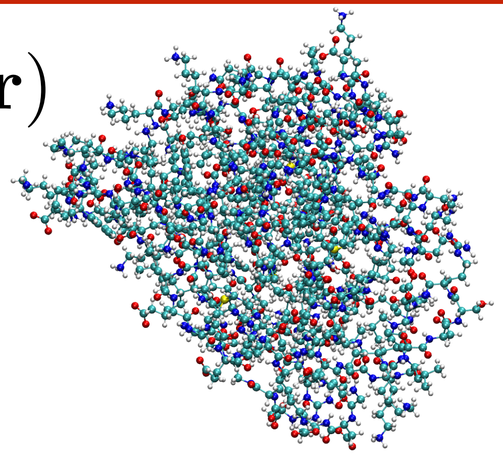
Rudzinski, Noid, J. Chem. Phys. 2011
Giulini, RM, Shell, Potestio, JCTC 2020

Information loss in coarse-graining



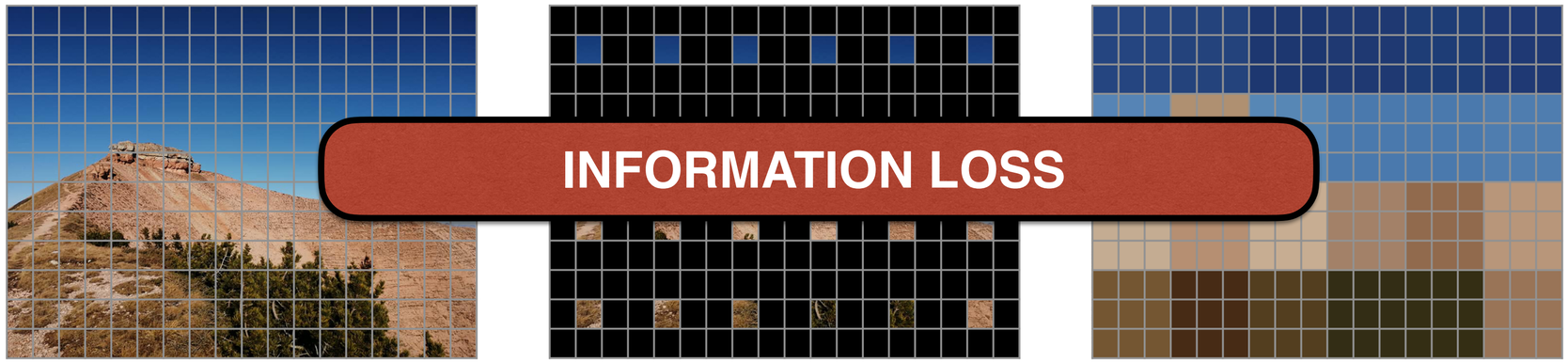
$$p_R(\mathbf{R})$$

$$\bar{p}_r(\mathbf{r}) \neq p_r(\mathbf{r})$$



$$\bar{p}_r(\mathbf{r}) = \frac{p_R(\mathbf{M}(\mathbf{r}))}{\Sigma(\mathbf{M}(\mathbf{r}))}$$

Each microstate **r** receives the average probability of the macrostate **R** it maps onto



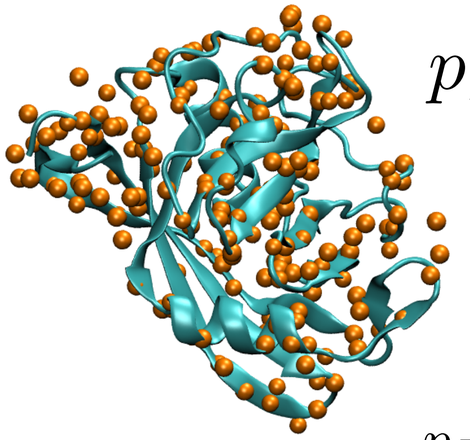
Original high resolution representation

Filtering

Reconstructed high resolution representation

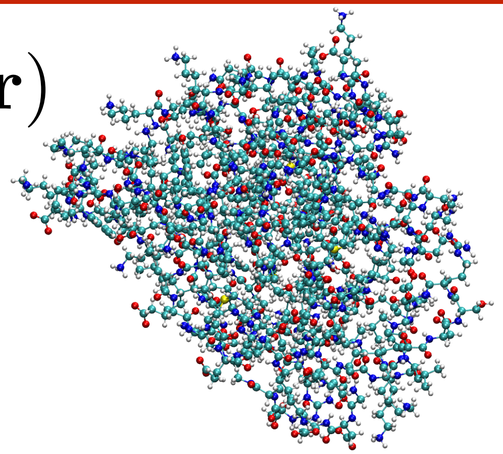
Rudzinski, Noid, J. Chem. Phys. 2011
Giulini, RM, Shell, Potestio, JCTC 2020

Information loss in coarse-graining



$$p_R(\mathbf{R})$$

$$\bar{p}_r(\mathbf{r}) \neq p_r(\mathbf{r})$$



$$\bar{p}_r(\mathbf{r}) = \frac{p_R(\mathbf{M}(\mathbf{r}))}{\Sigma(\mathbf{M}(\mathbf{r}))}$$

Each microstate **r** receives the average probability of the macrostate **R** it maps onto

Mapping entropy

$$S_{map} = k_B \int d\mathbf{r}\; p_r(\mathbf{r}) \ln \left[\frac{p_r(\mathbf{r})}{\bar{p}_r(\mathbf{r})} \right]$$

Quantifies the information loss arising from coarse-graining

DEPENDS ONLY ON THE CHOICE OF THE CG SITES

high resolution representation **Filtering** **high resolution representation**

Rudzinski, Noid, J. Chem. Phys. 2011
Giulini, RM, Shell, Potestio, JCTC 2020

Estimating the information loss

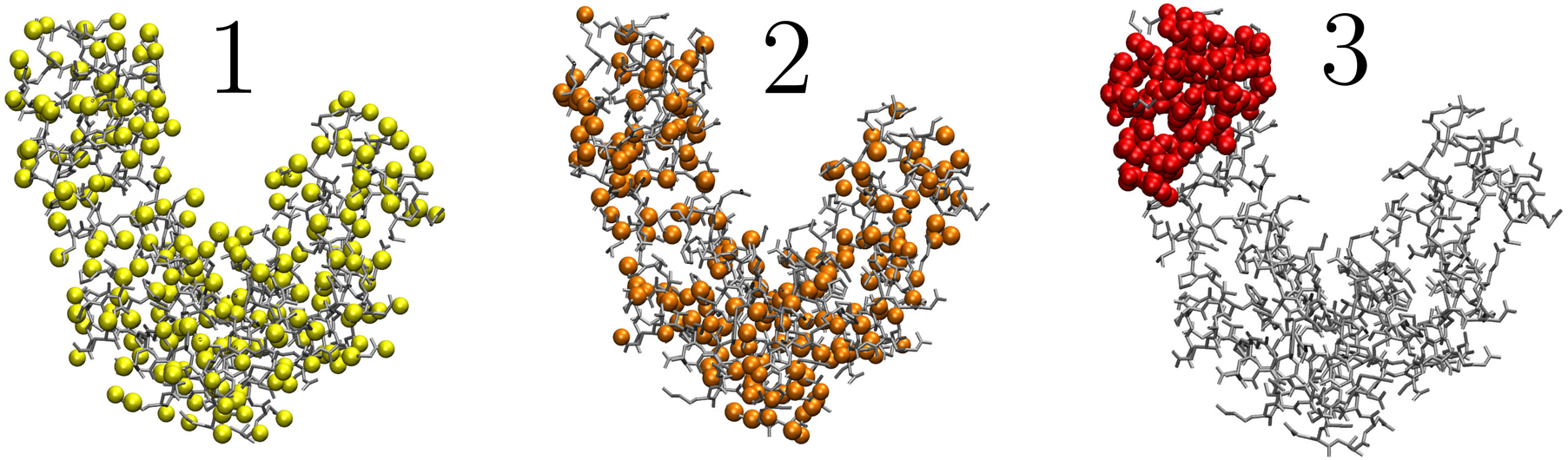
$$S_{map} \approx k_B \frac{\beta^2}{2} \int d\mathbf{R} p_R(\mathbf{R}) \langle (u - \langle u \rangle_{\beta|\mathbf{R}})^2 \rangle_{\beta|\mathbf{R}}$$

Giulini, RM, Shell, Potestio, JCTC 2020

Estimating the information loss

$$S_{map} \approx k_B \frac{\beta^2}{2} \int d\mathbf{R} p_R(\mathbf{R}) \langle (u - \langle u \rangle_{\beta|\mathbf{R}})^2 \rangle_{\beta|\mathbf{R}}$$

Calculating the mapping entropy of a CG representation

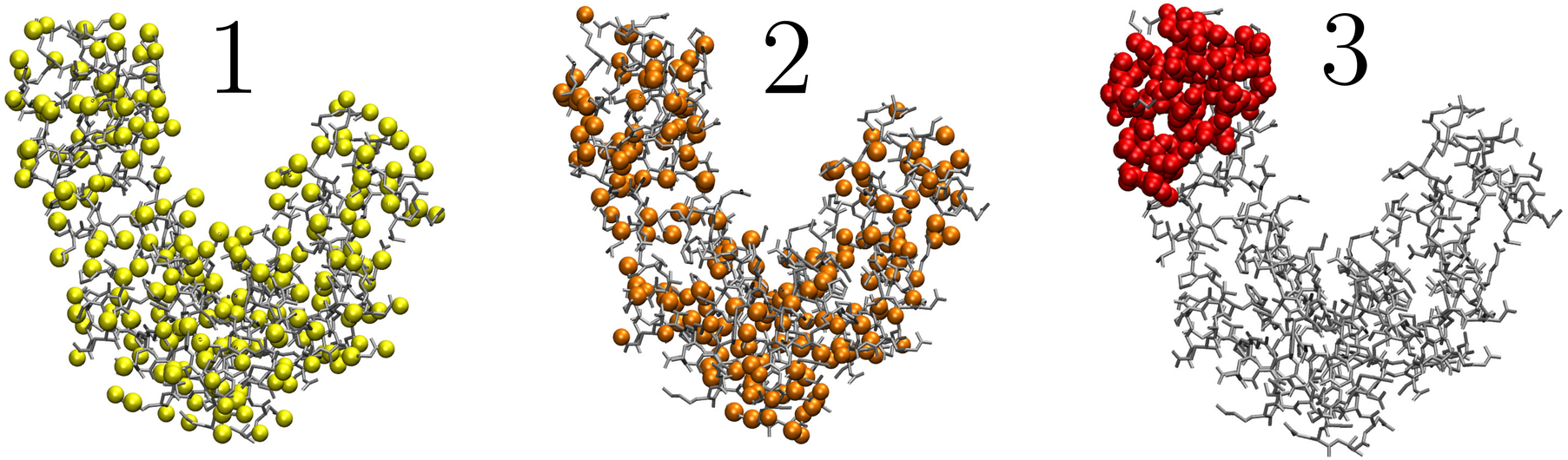


Giulini, RM, Shell, Potestio, JCTC 2020

Estimating the information loss

$$S_{map} \approx k_B \frac{\beta^2}{2} \int d\mathbf{R} p_R(\mathbf{R}) \langle (u - \langle u \rangle_{\beta|\mathbf{R}})^2 \rangle_{\beta|\mathbf{R}}$$

Calculating the mapping entropy of a CG representation



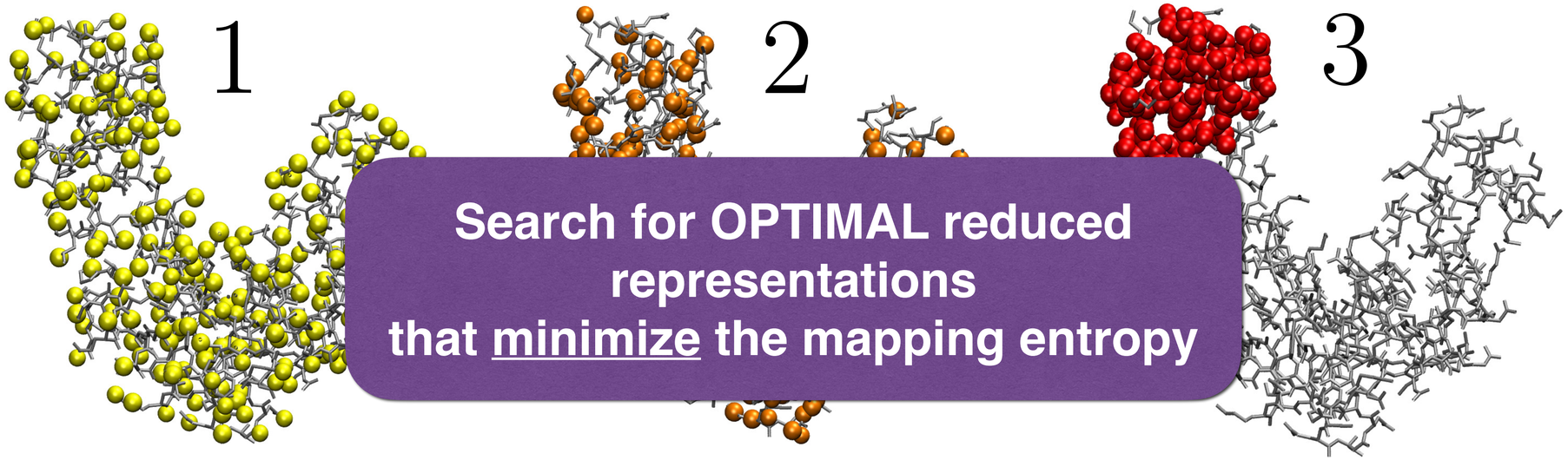
Ranking mappings according to the information they **retain: smaller S_{map}**

Giulini, RM, Shell, Potestio, JCTC 2020

Estimating the information loss

$$S_{map} \approx k_B \frac{\beta^2}{2} \int d\mathbf{R} p_R(\mathbf{R}) \langle (u - \langle u \rangle_{\beta|\mathbf{R}})^2 \rangle_{\beta|\mathbf{R}}$$

Calculating the mapping entropy of a CG representation



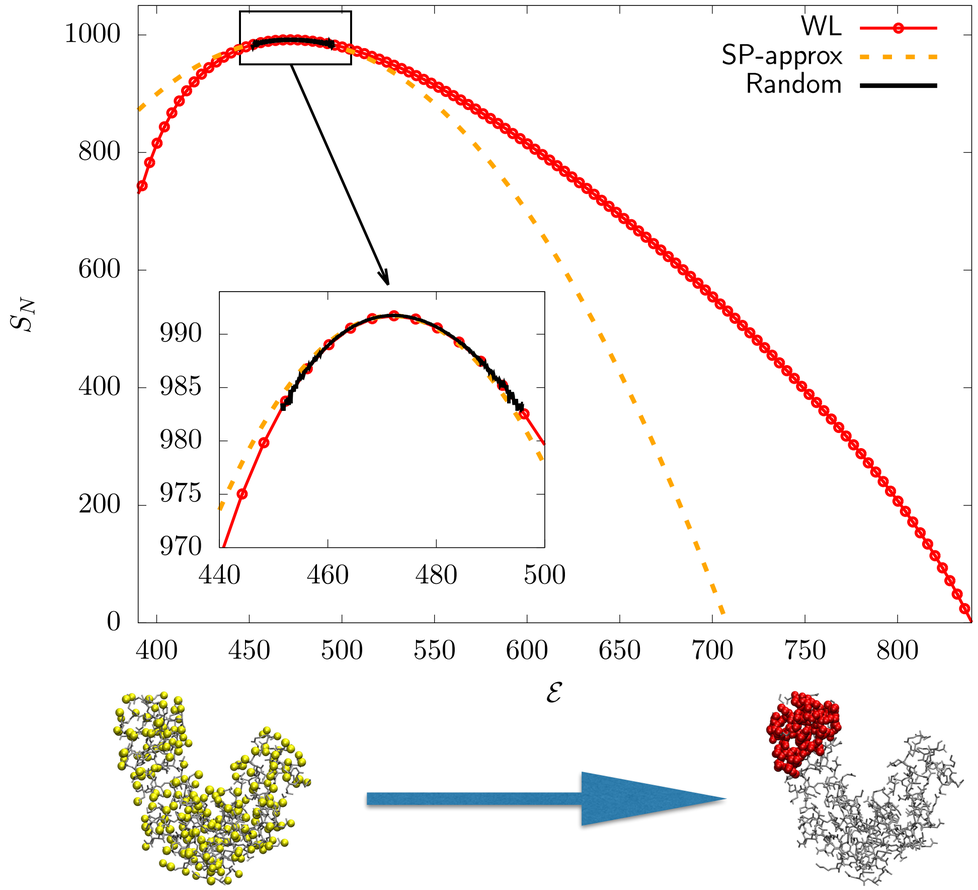
Search for OPTIMAL reduced representations that <u>minimize</u> the mapping entropy

Ranking mappings according to the information they **retain: smaller S_{map}**

Giulini, RM, Shell, Potestio, JCTC 2020

Estimating the information loss

$$\#\text{selections}(\mathcal{E}) \propto e^{S_N(\mathcal{E})}$$



RM, Giulini, Potestio, EPJB 2021

Ca

1

Ranking

ntation

3

ain: smaller S_{map}

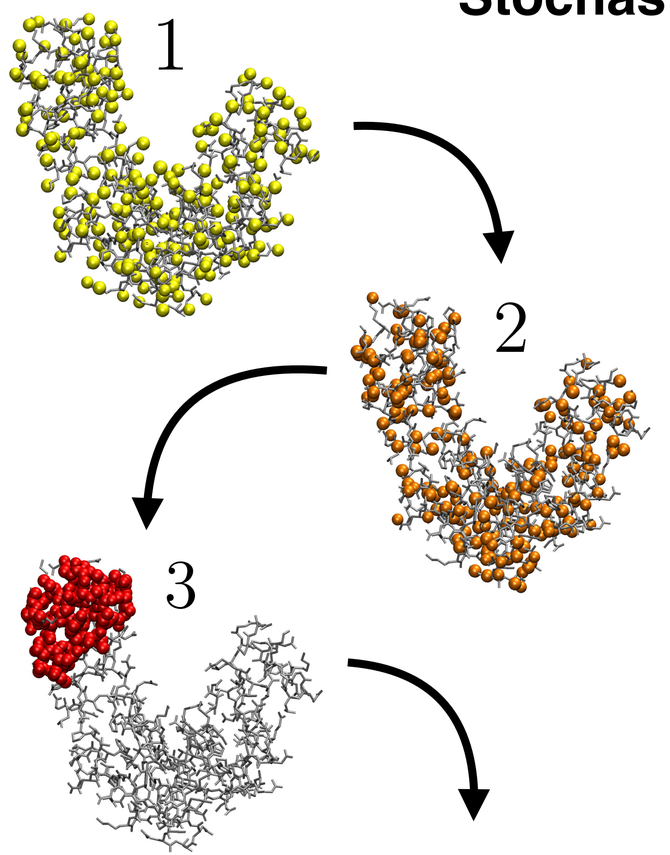
S_{map}

Giulini, RM, Shell, Potestio, JCTC 2020

Minimizing the information loss

Looking for maximally informative selections of CG sites:

Stochastic minimization of S_{map} over the mapping space

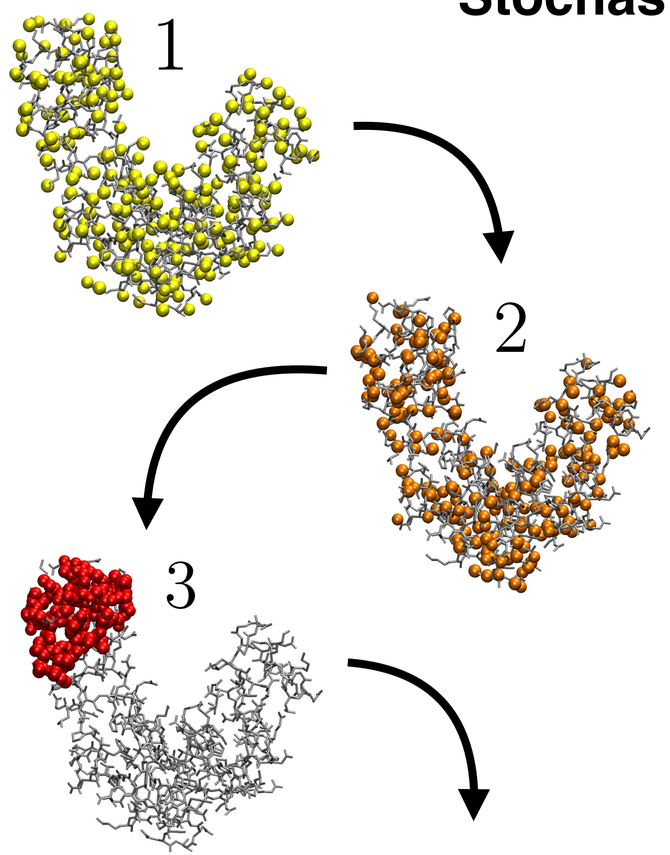


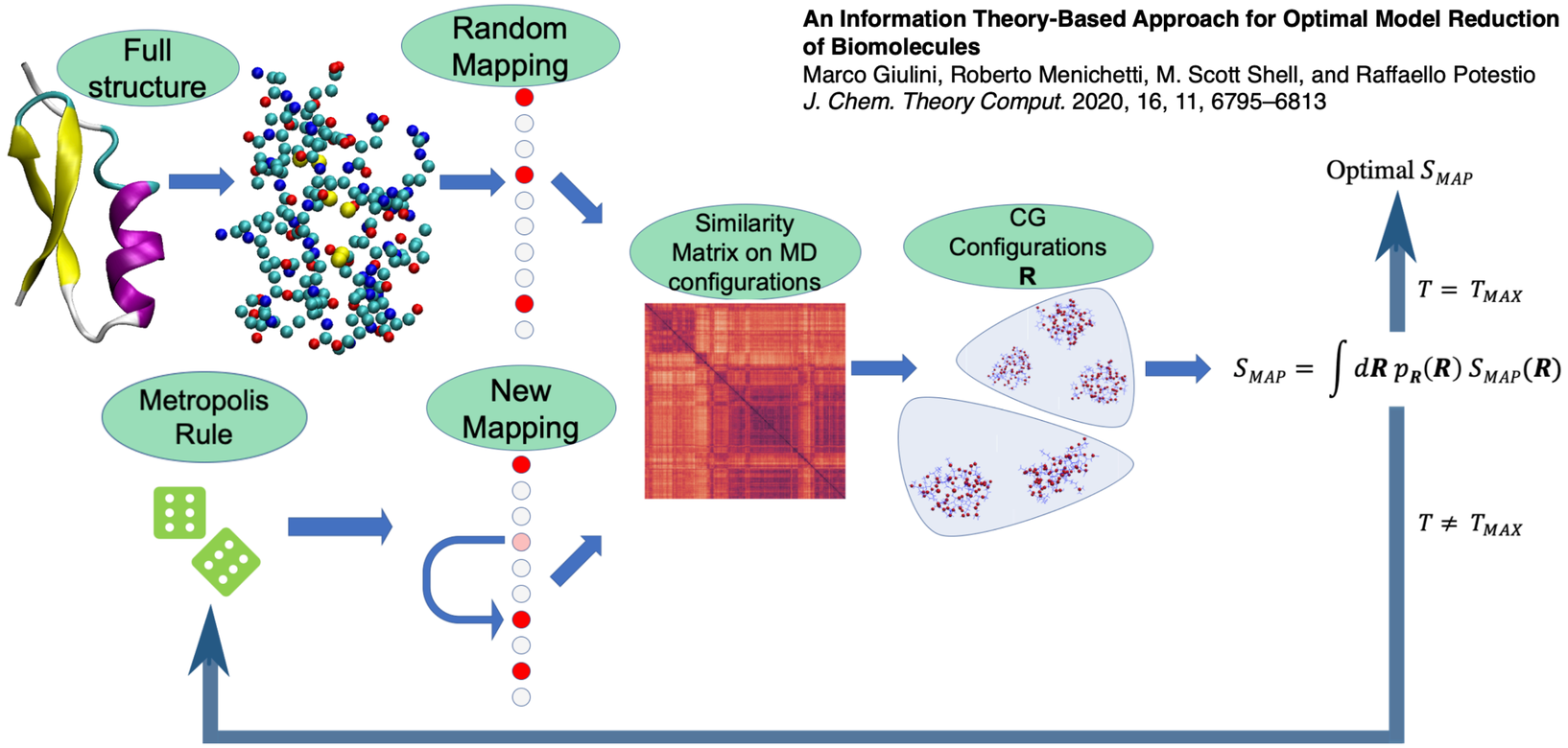
Giulini, Rigoli, RM, *et al.*, JCTC 2020
Errica, Giulini, RM *et al.*, Front. Mol. Biosci. 2021
RM, Giulini, Potestio, EPJB 2021

Minimizing the information loss

Looking for maximally informative selections of CG sites:

Stochastic minimization of S_{map} over the mapping space





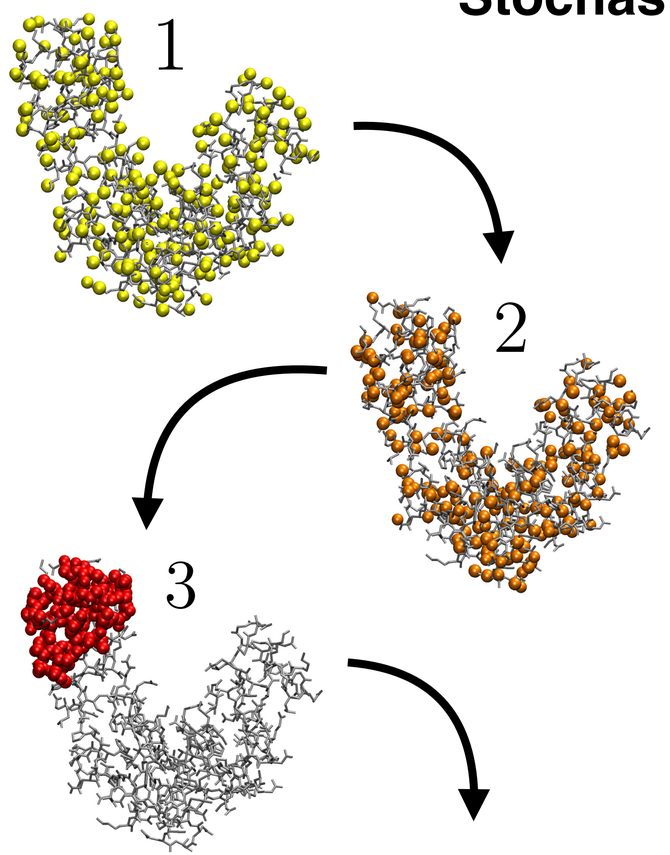
An Information Theory-Based Approach for Optimal Model Reduction of Biomolecules
Marco Giulini, Roberto Menichetti, M. Scott Shell, and Raffaello Potestio
J. Chem. Theory Comput. 2020, 16, 11, 6795–6813

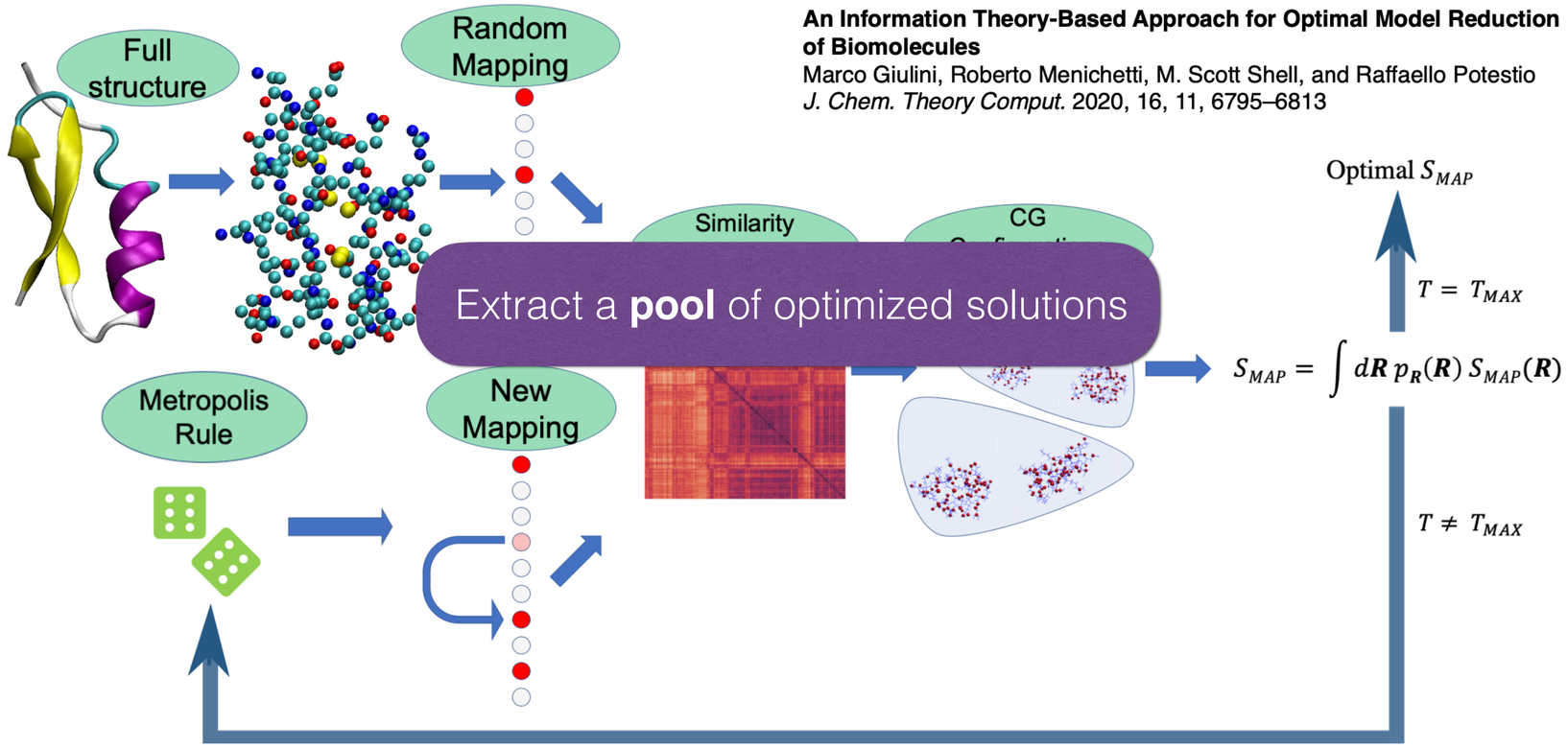
Giulini, Rigoli, RM, *et al.*, JCTC 2020
Errica, Giulini, RM *et al.*, Front. Mol. Biosci. 2021
RM, Giulini, Potestio, EPJB 2021

Minimizing the information loss

Looking for maximally informative selections of CG sites:

Stochastic minimization of S_{map} over the mapping space





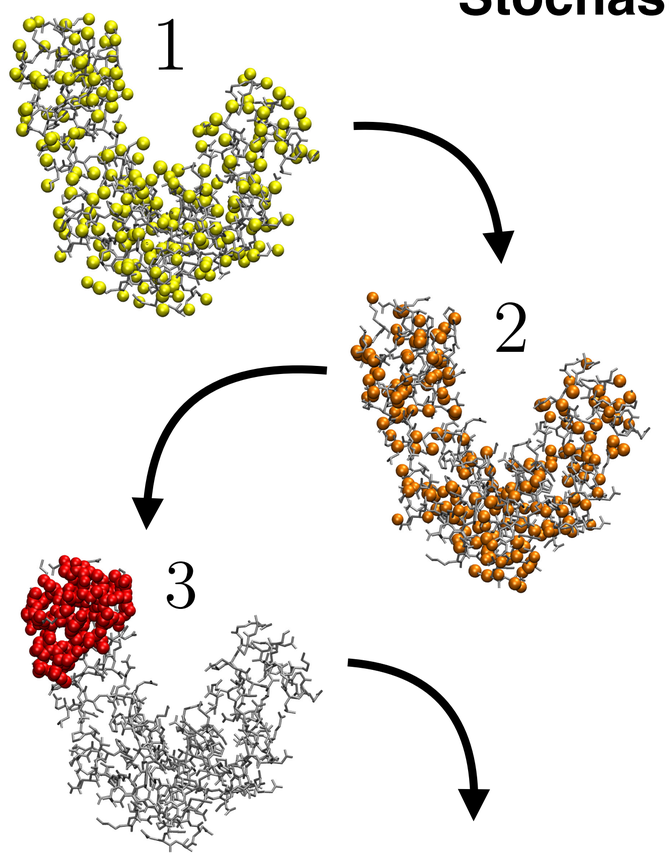
An Information Theory-Based Approach for Optimal Model Reduction of Biomolecules
Marco Giulini, Roberto Menichetti, M. Scott Shell, and Raffaello Potestio
J. Chem. Theory Comput. 2020, 16, 11, 6795–6813

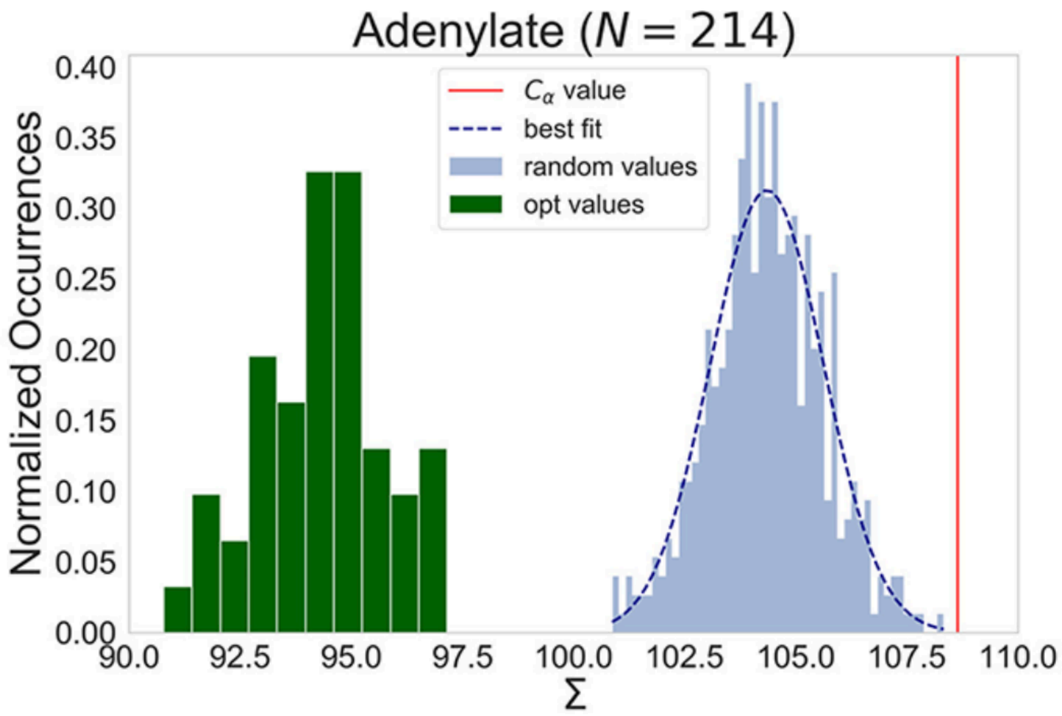
Giulini, Rigoli, RM, *et al.*, JCTC 2020
Errica, Giulini, RM *et al.*, Front. Mol. Biosci. 2021
RM, Giulini, Potestio, EPJB 2021

Minimizing the information loss

Looking for maximally informative selections of CG sites:

Stochastic minimization of S_{map} over the mapping space

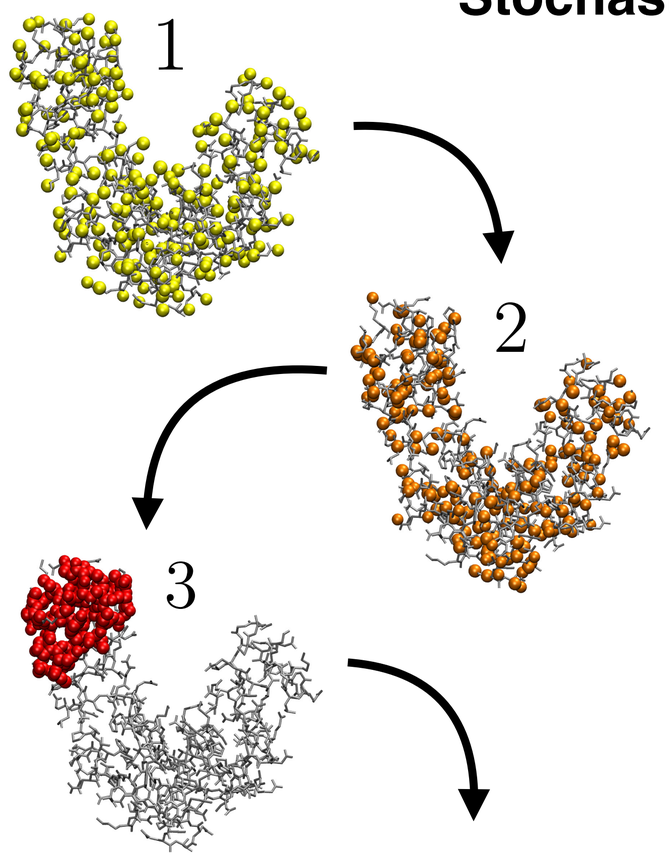


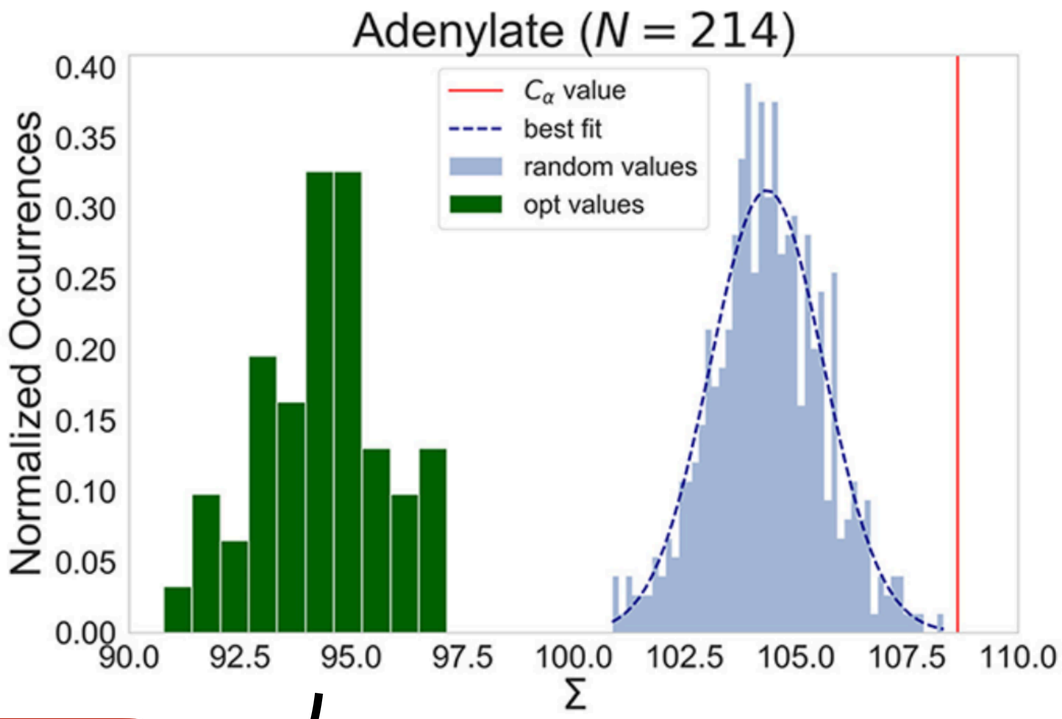


Giulini, Rigoli, RM, *et al.*, JCTC 2020
Errica, Giulini, RM *et al.*, Front. Mol. Biosci. 2021
RM, Giulini, Potestio, EPJB 2021

Minimizing the information loss

Looking for maximally informative selections of CG sites:
Stochastic minimization of S_{map} over the mapping space



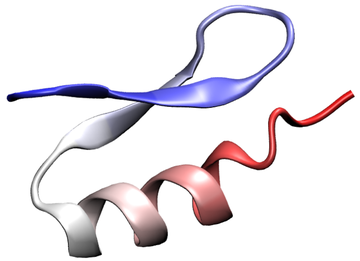


Maximally-informative CG mappings consist of a list of atoms: **statistical analysis of the pool of optimized solutions**

Giulini, Rigoli, RM, *et al.*, JCTC 2020
Errica, Giulini, RM *et al.*, Front. Mol. Biosci. 2021
RM, Giulini, Potestio, EPJB 2021

Investigating the nature of optimal mappings

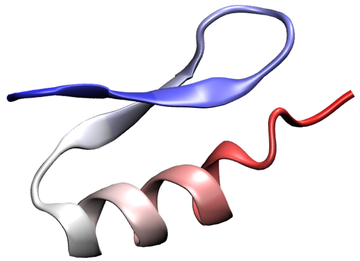
Application to Tamapin

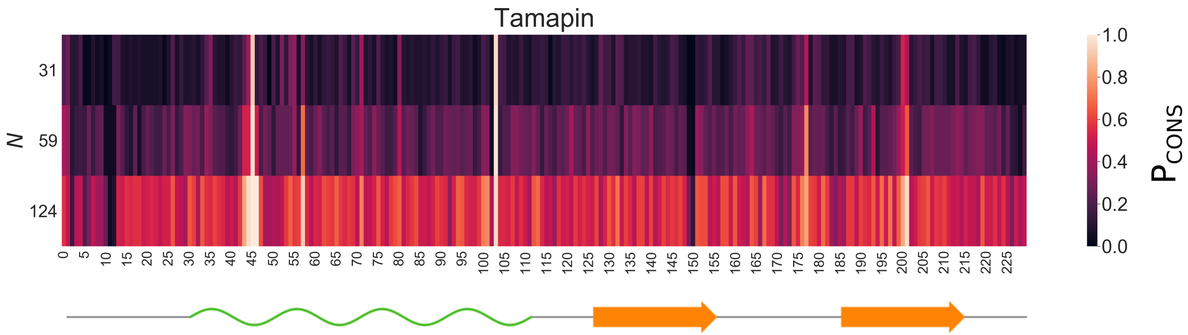


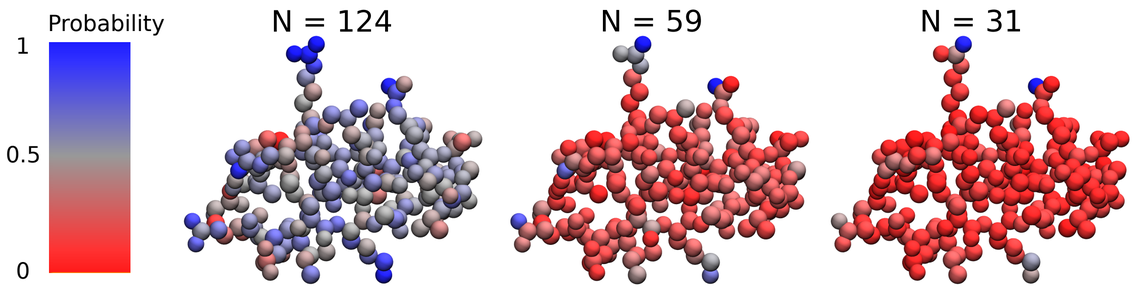
Giulini, RM, Shell, Potestio, JCTC 2020

Investigating the nature of optimal mappings

Application to Tamapin





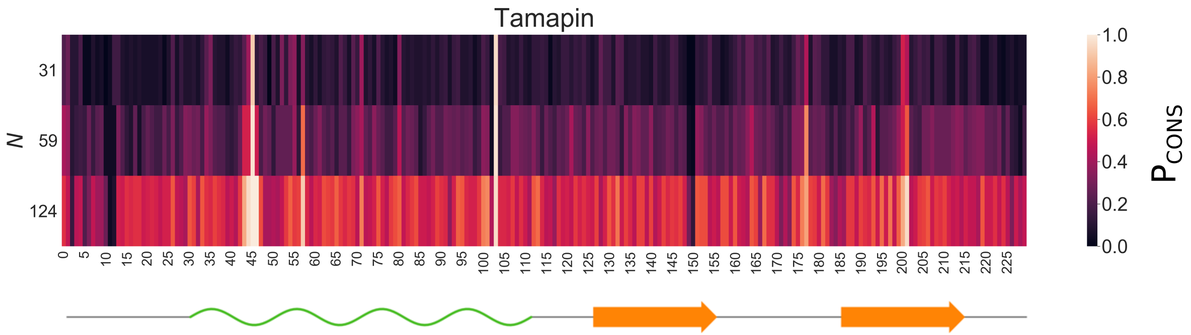


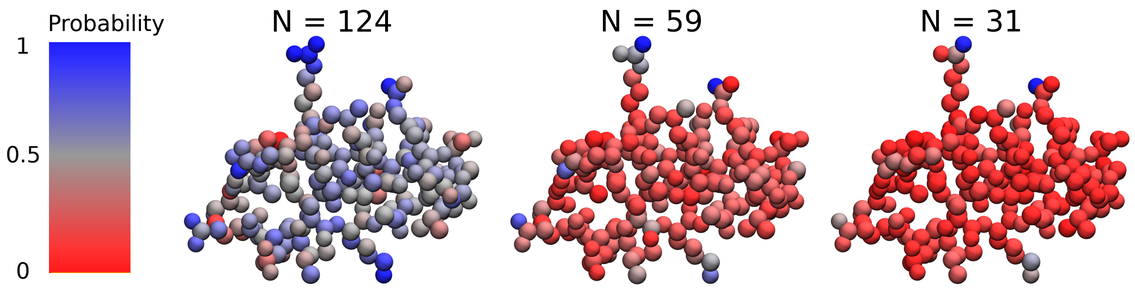
Giulini, RM, Shell, Potestio, JCTC 2020

Investigating the nature of optimal mappings

Application to Tamapin

Biologically relevant residues are singled out in an unsupervised manner





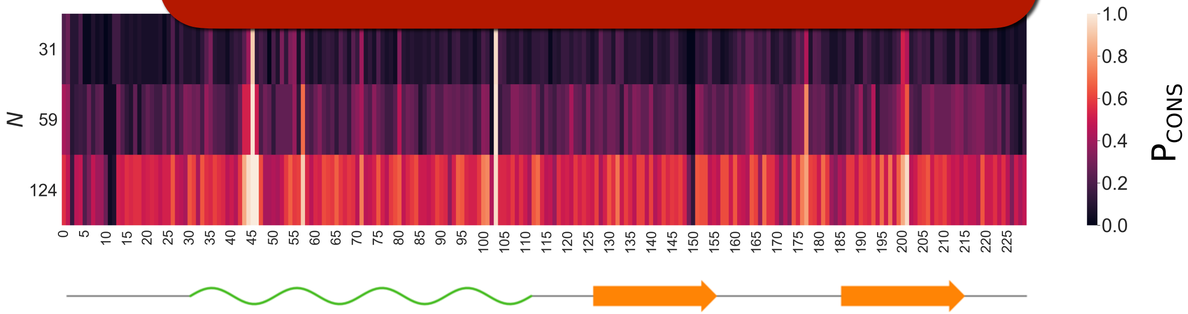
Giulini, RM, Shell, Potestio, JCTC 2020

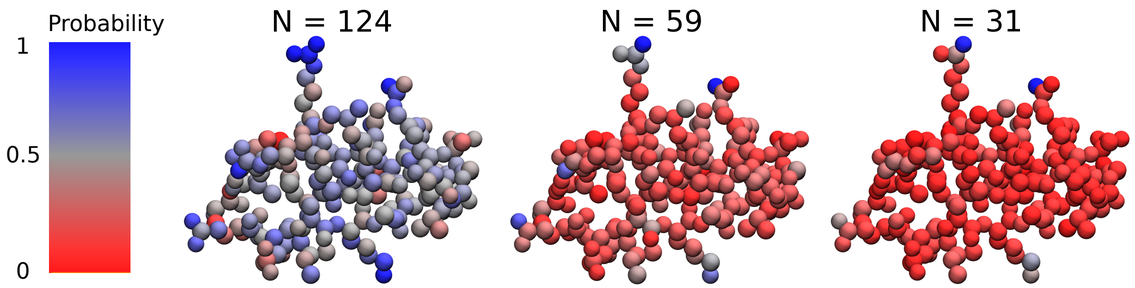
Investigating the nature of optimal mappings

Application to Tamapin

Biologically relevant residues are singled out in an unsupervised manner

Simulations were performed in absence of the substrate





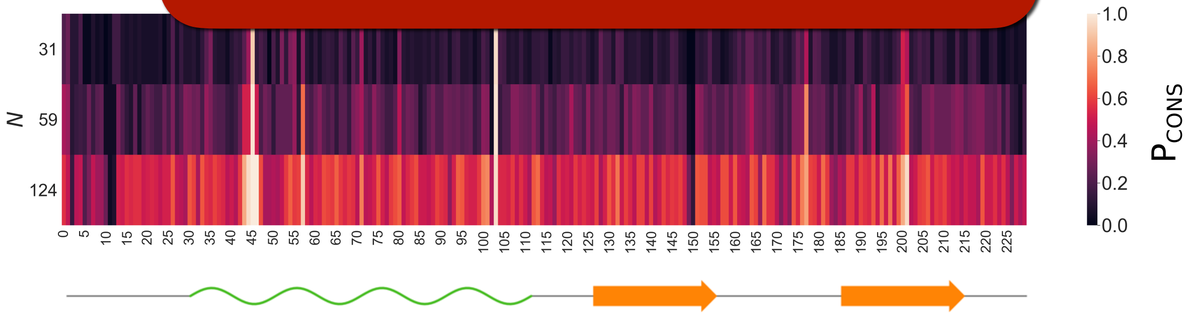
Giulini, RM, Shell, Potestio, JCTC 2020

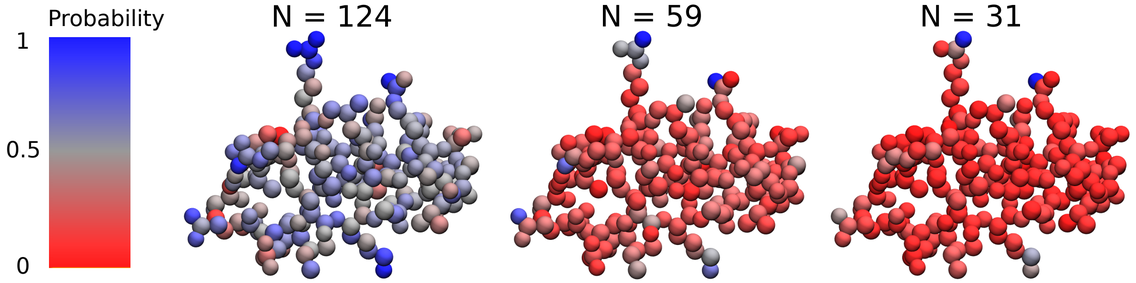
Investigating the nature of optimal mappings

Application to Tamapin

Biologically relevant residues are singled out in an unsupervised manner

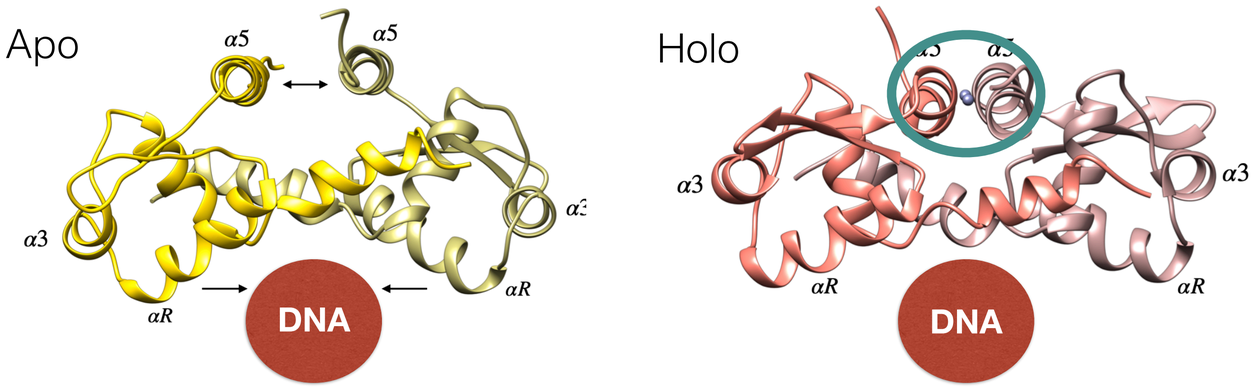
Simulations were performed in absence of the substrate





Giulini, RM, Shell, Potestio, JCTC 2020

Application to CzrA transcription repressor



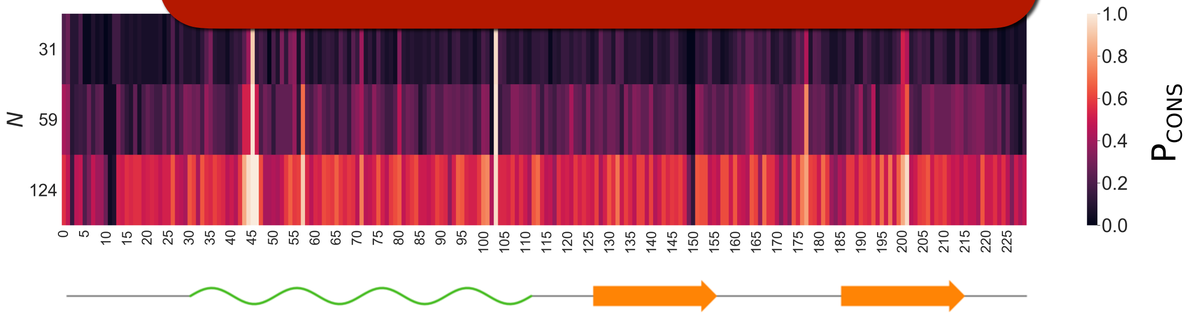
Rigoli, Giulini, RM, Potestio, in preparation

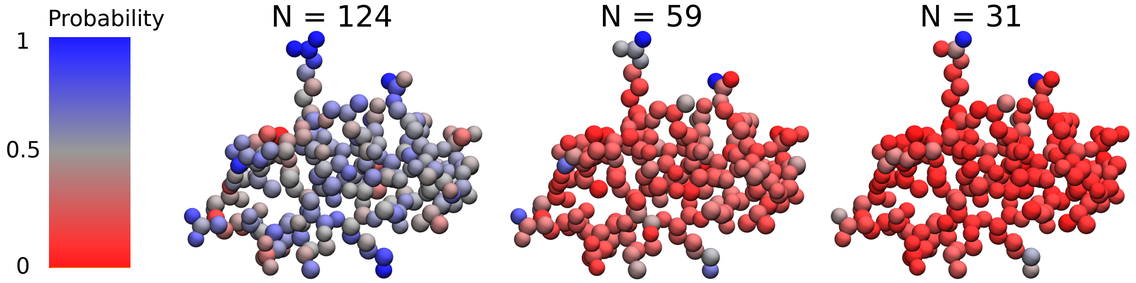
Investigating the nature of optimal mappings

Application to Tamapin

Biologically relevant residues are singled out in an unsupervised manner

Simulations were performed in absence of the substrate

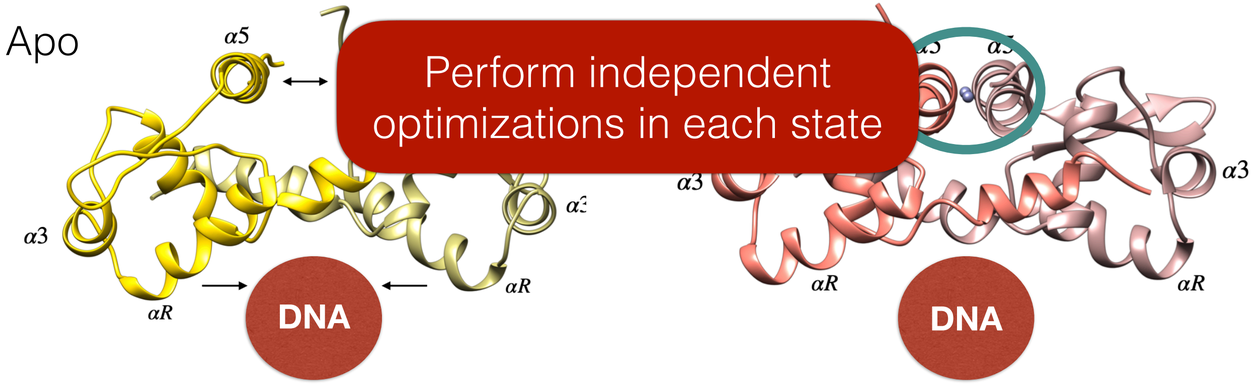




Giulini, RM, Shell, Potestio, JCTC 2020

Application to CzrA transcription repressor

Perform independent optimizations in each state



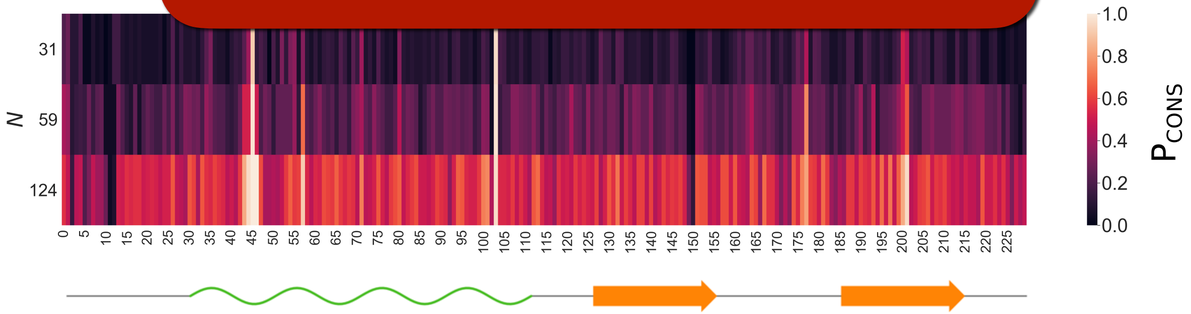
Rigoli, Giulini, RM, Potestio, in preparation

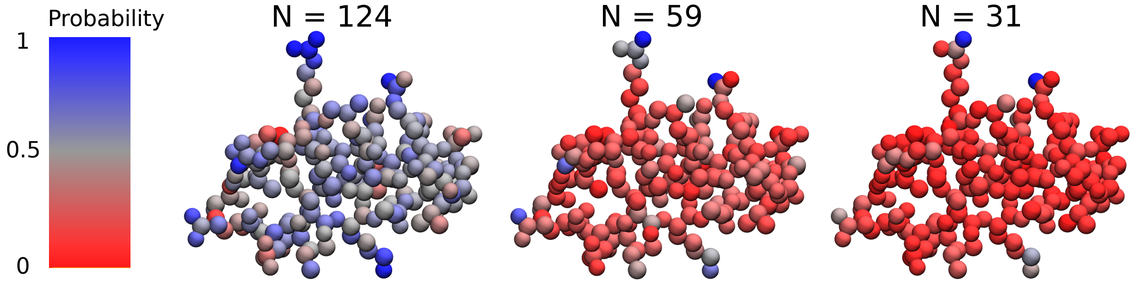
Investigating the nature of optimal mappings

Application to Tamapin

Biologically relevant residues are singled out in an unsupervised manner

Simulations were performed in absence of the substrate

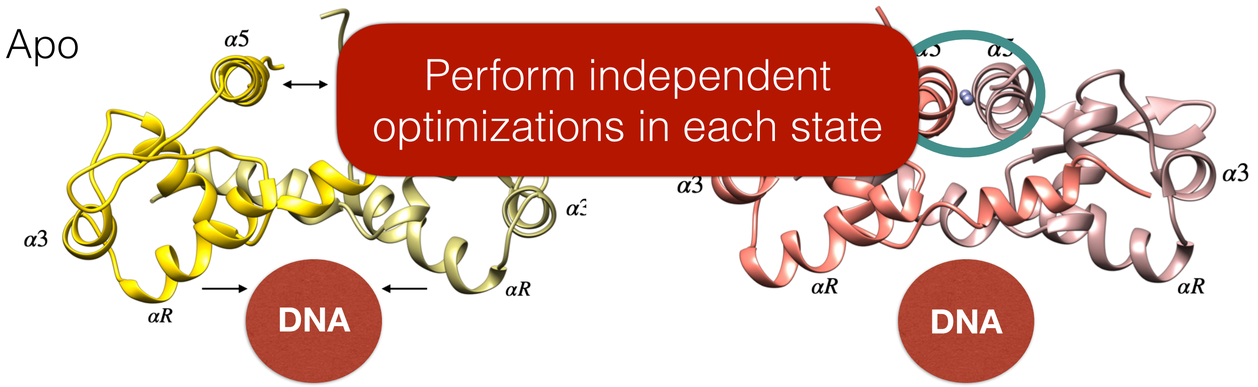


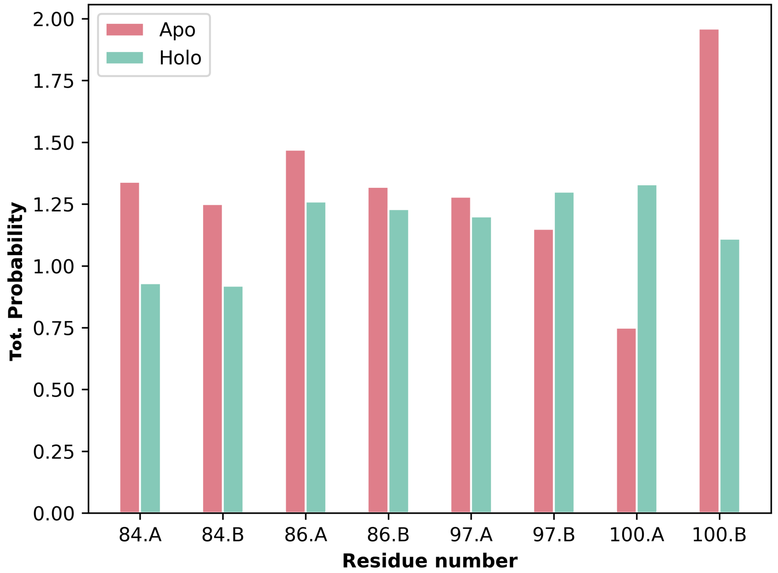


Giulini, RM, Shell, Potestio, JCTC 2020

Application to CzrA transcription repressor

Perform independent optimizations in each state





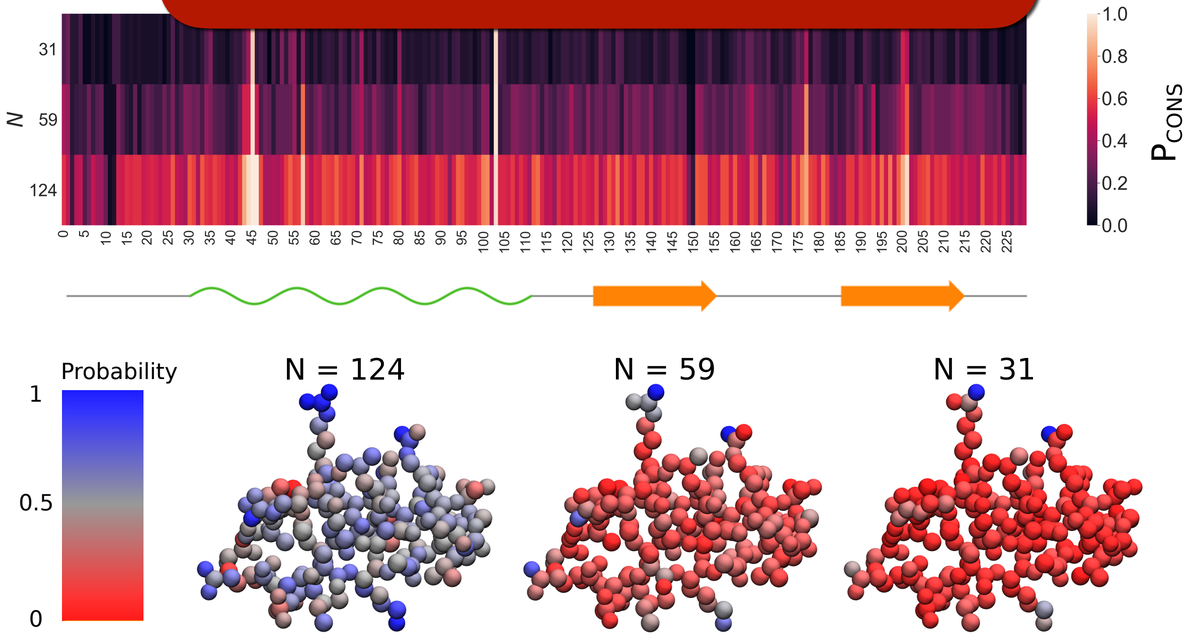
Rigoli, Giulini, RM, Potestio, in preparation

Investigating the nature of optimal mappings

Application to Tamapin

Biologically relevant residues are singled out in an unsupervised manner

Simulations were performed in absence of the substrate

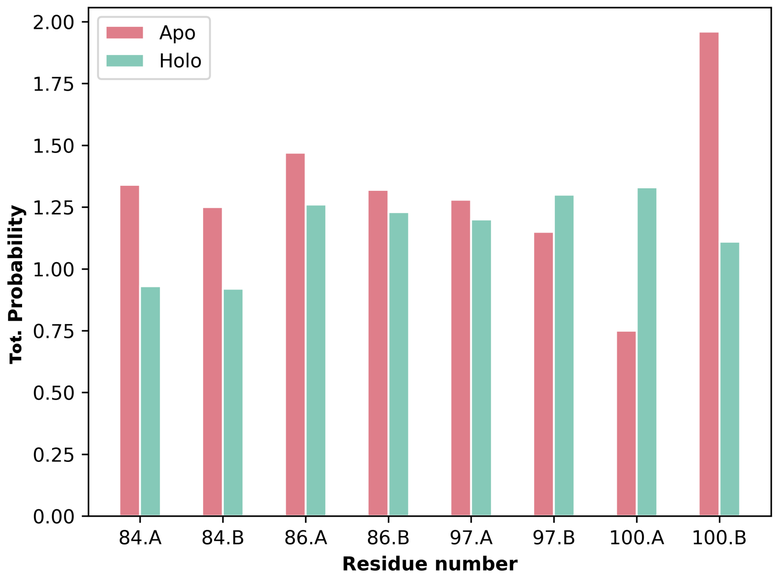


Giulini, RM, Shell, Potestio, JCTC 2020

Application to CzrA transcription repressor

Perform independent optimizations in each state

Biologically relevant residues "deactivate" in presence of the partner

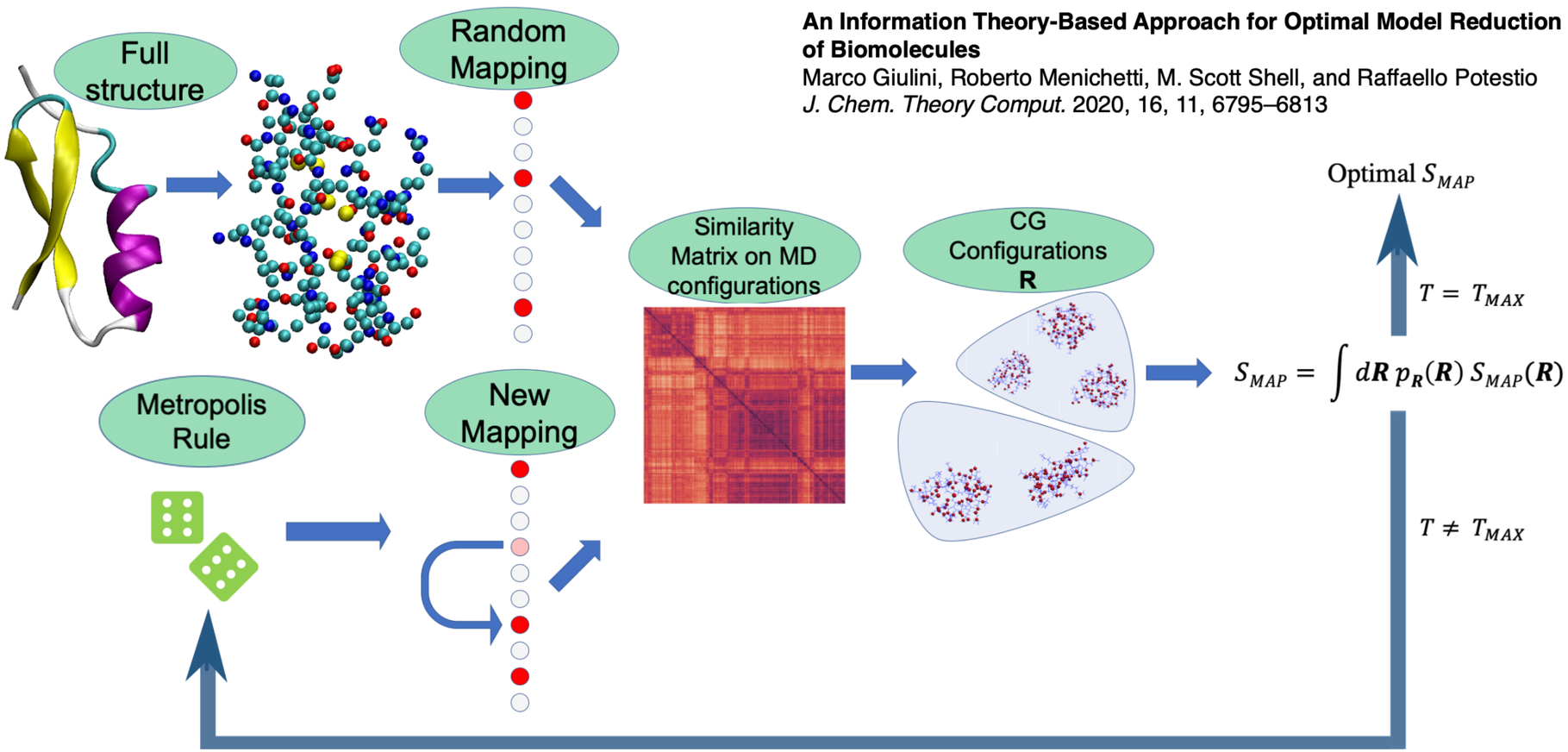


Rigoli, Giulini, RM, Potestio, in preparation

S_{map} via Deep Graph Networks

Calculating the mapping entropy is a computationally expensive task

An Information Theory-Based Approach for Optimal Model Reduction of Biomolecules
Marco Giulini, Roberto Menichetti, M. Scott Shell, and Raffaello Potestio
J. Chem. Theory Comput. 2020, 16, 11, 6795–6813

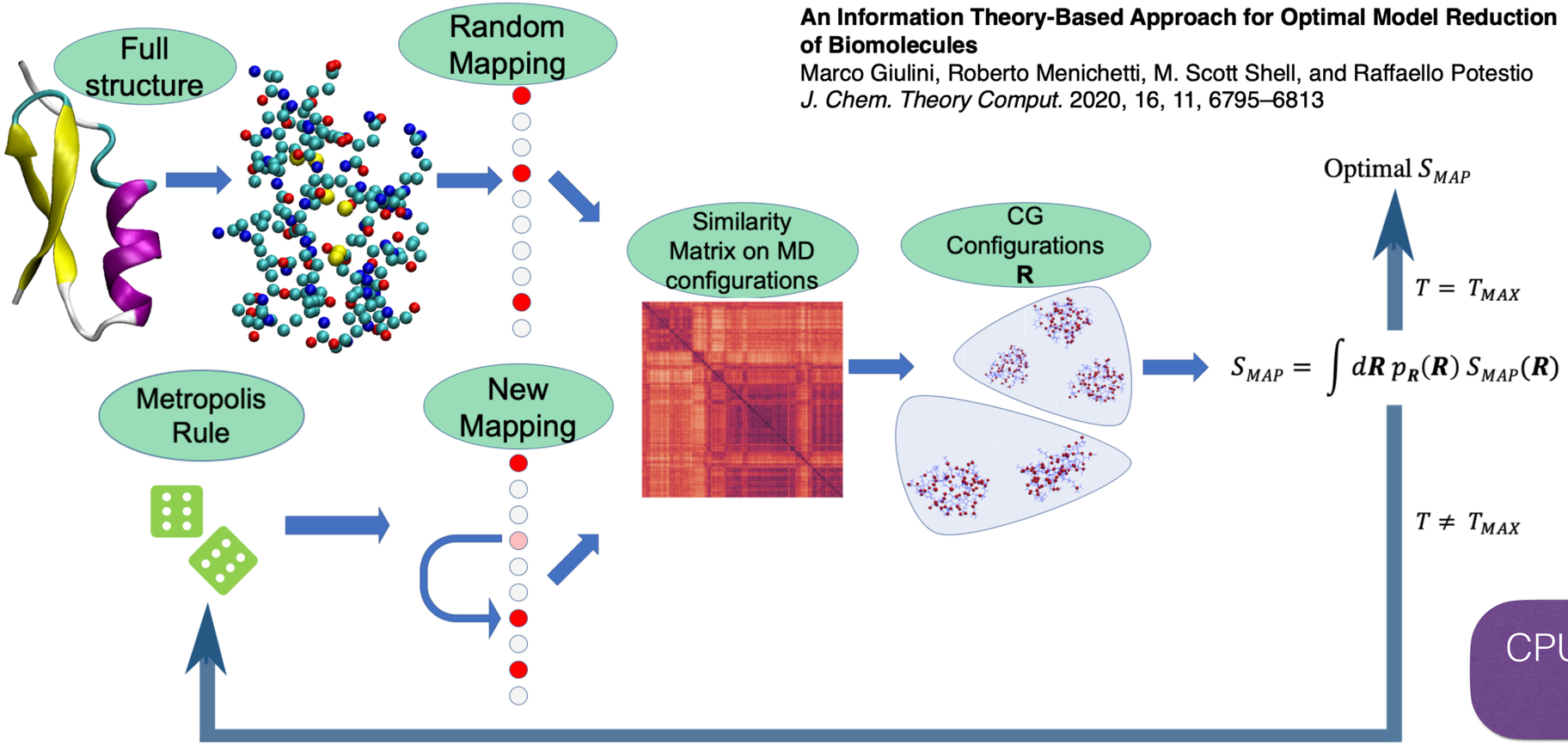


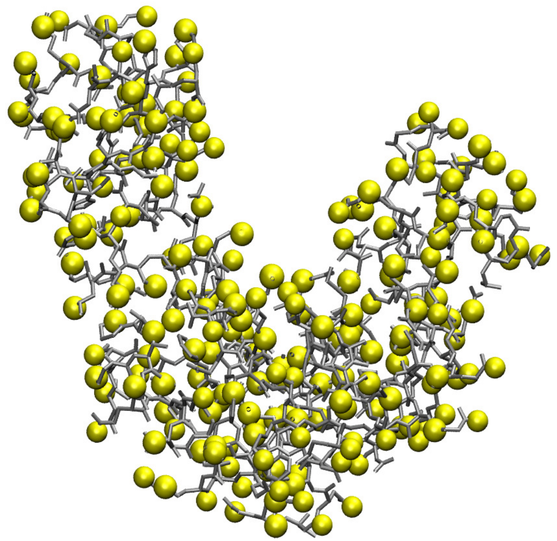
Errica, Giulini, RM *et al.*, Front. Mol. Biosci. 2021

S_{map} via Deep Graph Networks

Calculating the mapping entropy is a computationally expensive task

An Information Theory-Based Approach for Optimal Model Reduction of Biomolecules
Marco Giulini, Roberto Menichetti, M. Scott Shell, and Raffaello Potestio
J. Chem. Theory Comput. 2020, 16, 11, 6795–6813





CPU time required for a **single** S_{map} calculation: **~10 minutes**

Errica, Giulini, RM *et al.*, Front. Mol. Biosci. 2021

S_{map} via Deep Graph Networks

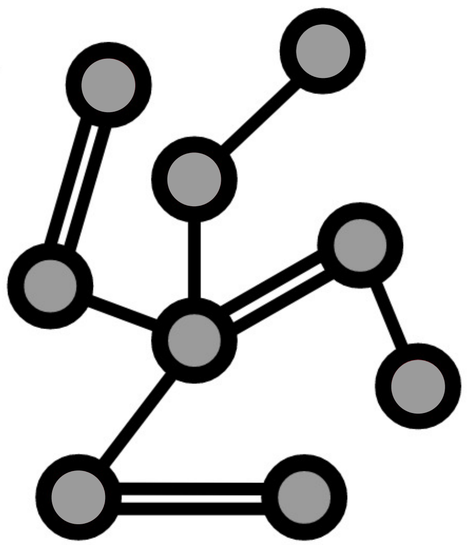
Artificial Intelligence to speed up calculations: Deep Graph Networks

Errica, Giulini, RM *et al*., Front. Mol. Biosci. 2021

S_{map} via Deep Graph Networks

Artificial Intelligence to speed up calculations: Deep Graph Networks

Encode the (static) structure of the molecule in a graph



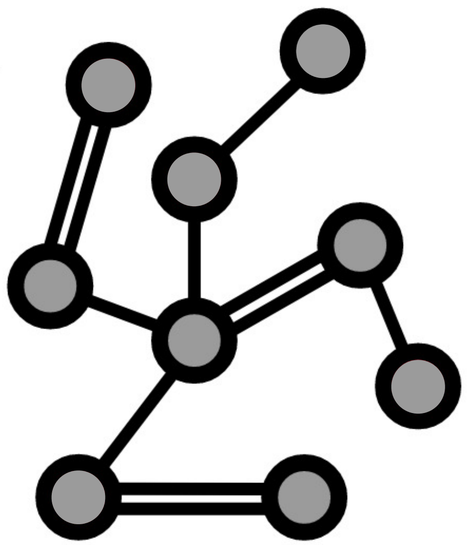
Errica, Giulini, RM *et al*., Front. Mol. Biosci. 2021

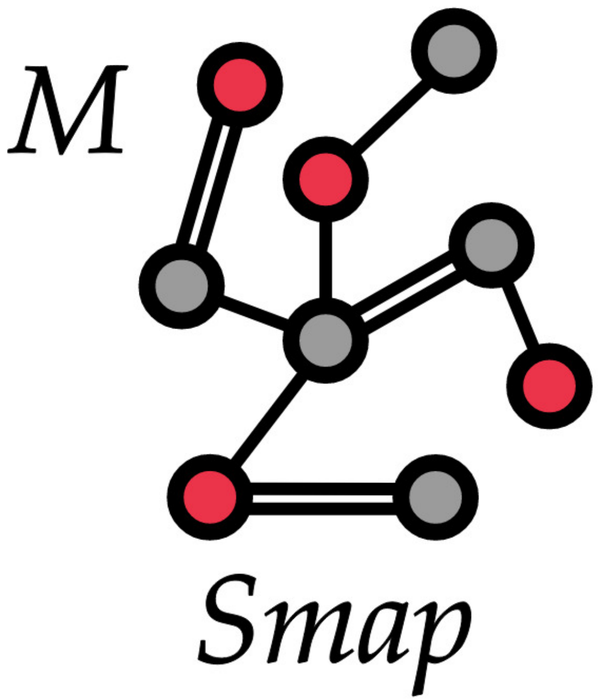
S_{map} via Deep Graph Networks

Artificial Intelligence to speed up calculations: Deep Graph Networks

Encode the (static) structure of the molecule in a graph

Employ DGNs to predict the mapping entropy





Errica, Giulini, RM *et al.*, Front. Mol. Biosci. 2021

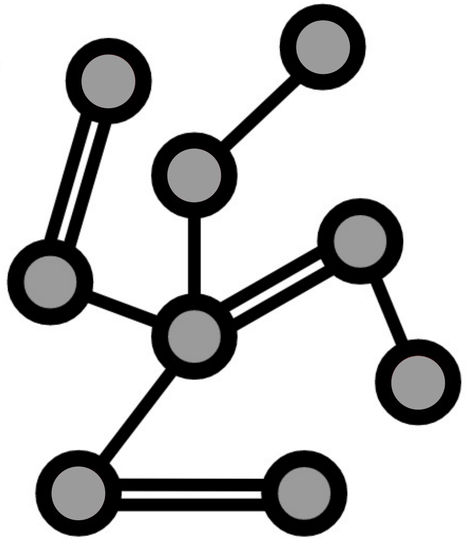
S_{map} via Deep Graph Networks

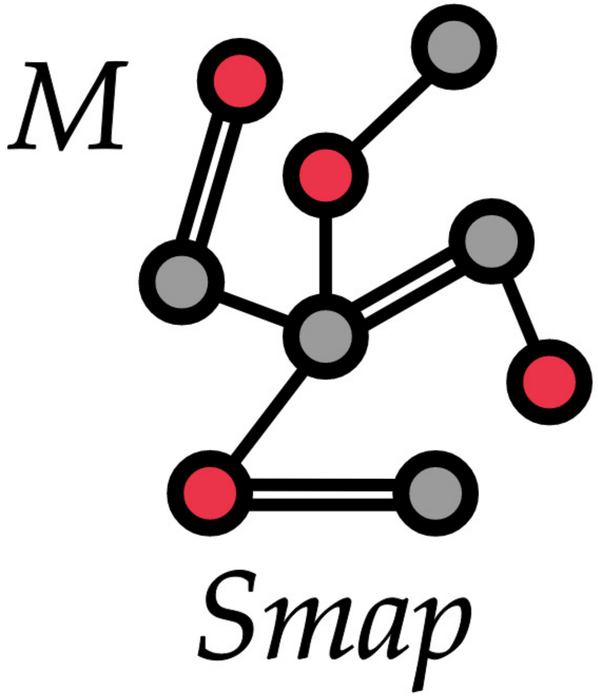
Artificial Intelligence to speed up calculations: Deep Graph Networks

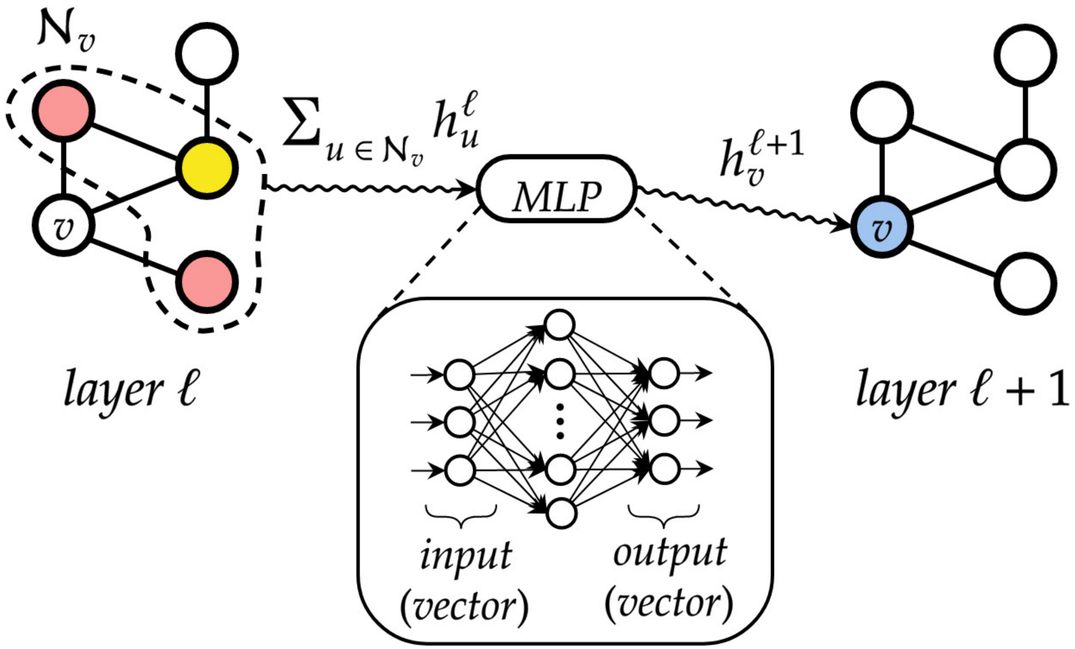
Encode the (static) structure of the molecule in a graph

Employ DGNs to predict the mapping entropy

$$\mathbf{h}_v^{\ell+1} = MLP^{\ell}\left(\left(1+\epsilon^{\ell}\right) * \mathbf{h}_v^{\ell} + \sum_{u \in \mathcal{N}_v} \mathbf{h}_u^{\ell} * e_{uv},\right)$$



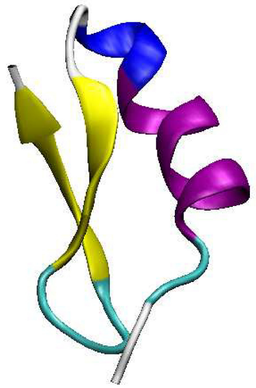




Errica, Giulini, RM *et al.*, Front. Mol. Biosci. 2021

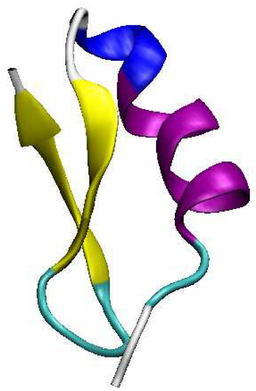
S_{map} via Deep Graph Networks

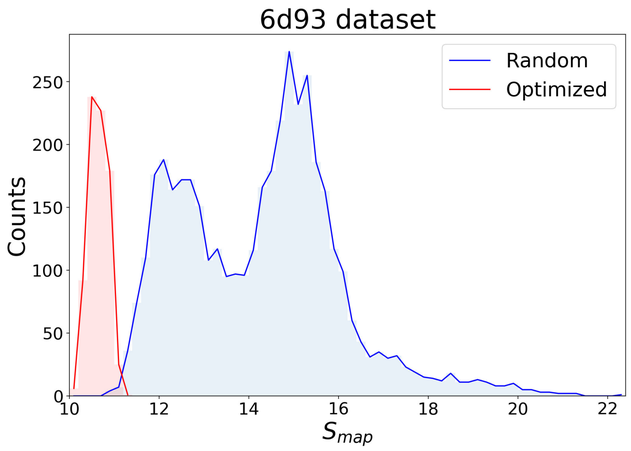
Tamapin



S_{map} via Deep Graph Networks

Tamapin

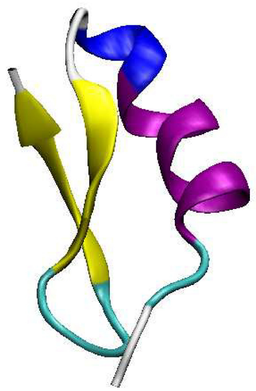


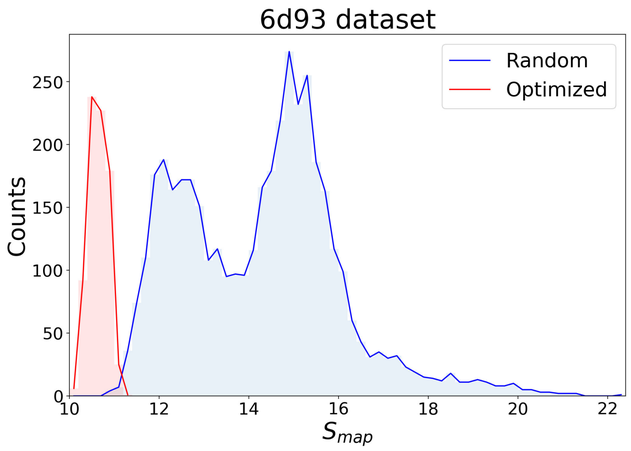


Training the DGN on the S_{map} of
- Random CG representations
- Maximally informative CG representations:
vanishing statistical weight

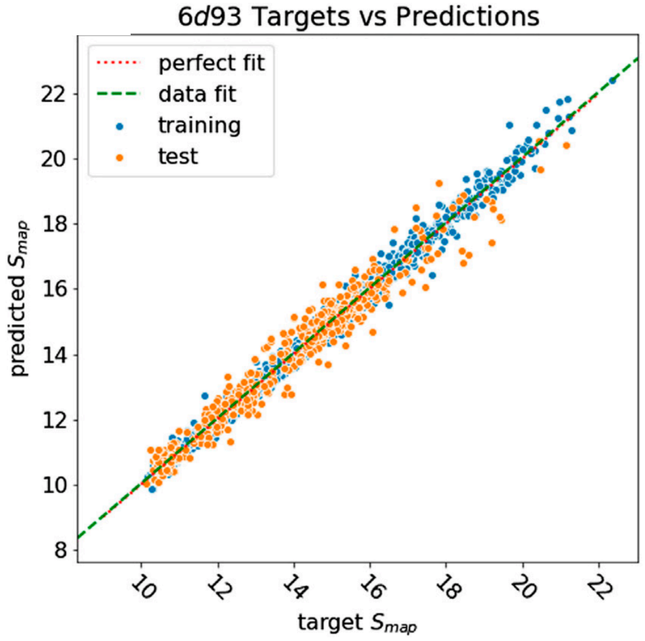
S_{map} via Deep Graph Networks

Tamapin





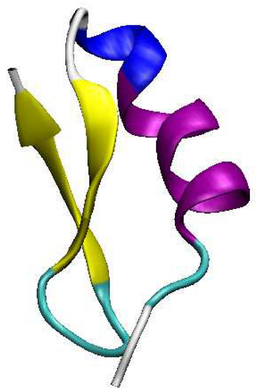
Training the DGN on the S_{map} of
- Random CG representations
- Maximally informative CG representations:
vanishing statistical weight

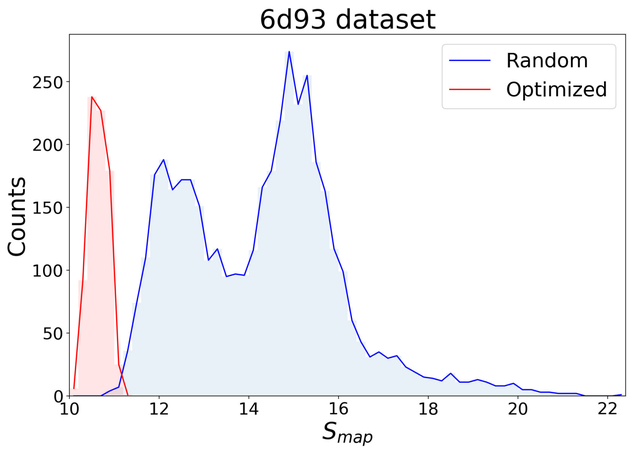


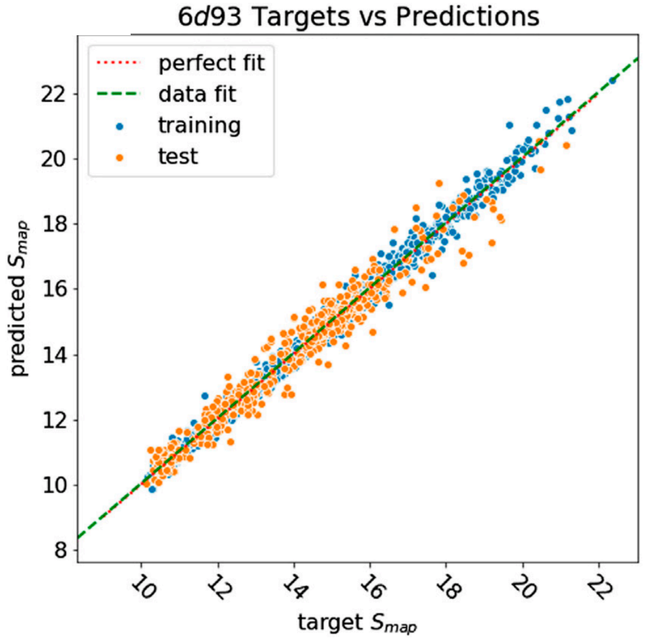
The DGN succeds in predicting
the S_{map} of both classes.
Speedup wrt standard algorithm:
10^3 (CPU) - 10^5 (GPU)

S_{map} via Deep Graph Networks

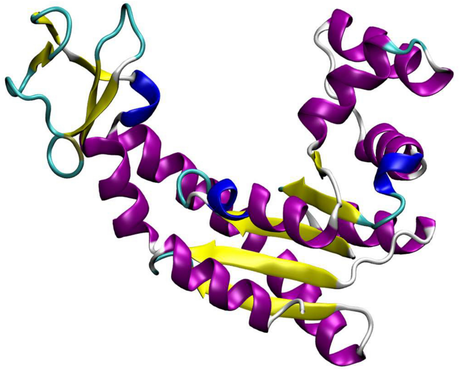
Tamapin

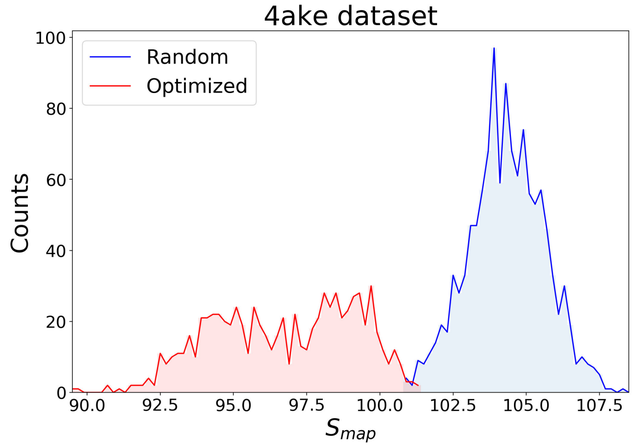


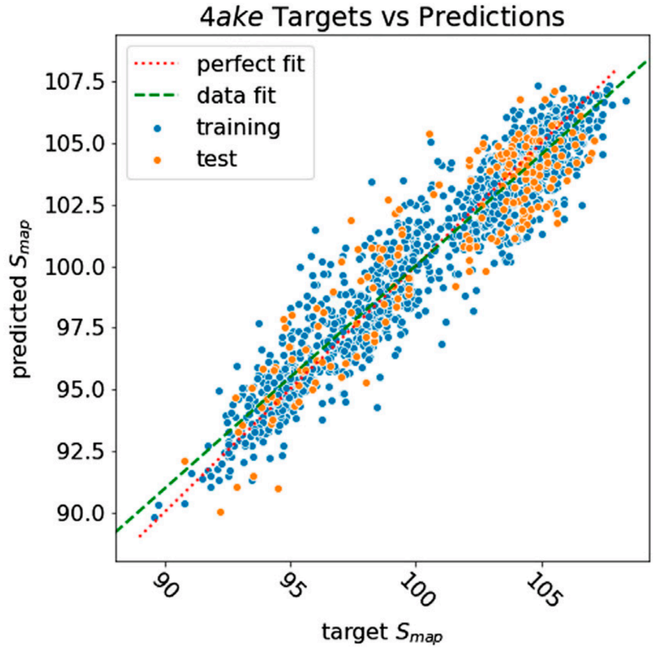




Adenylate Kinase



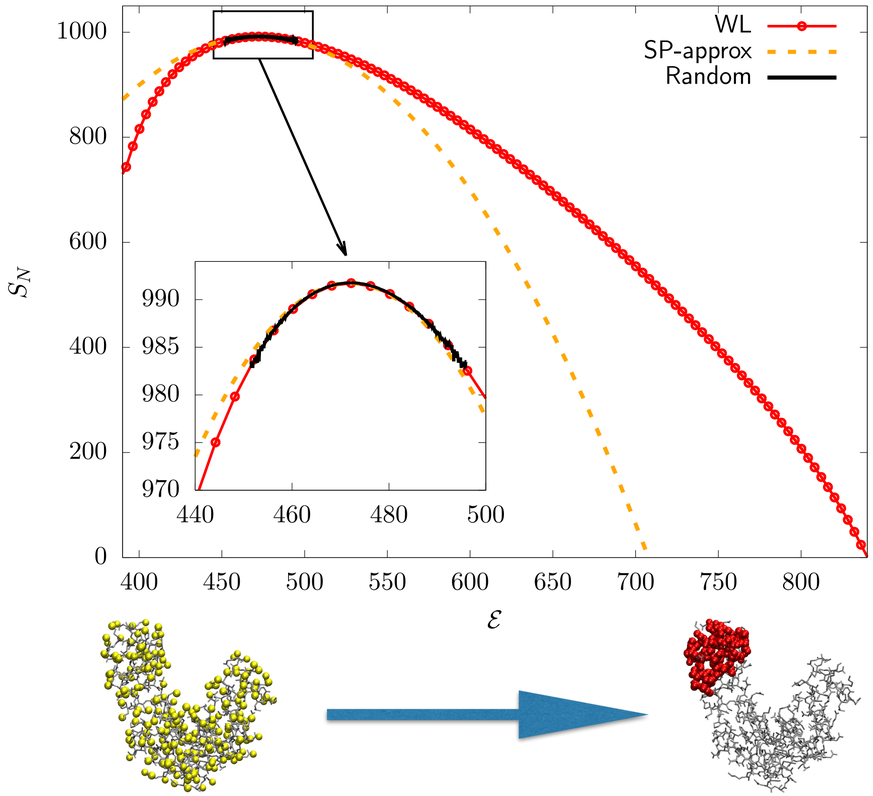




S_{map} via Deep Graph Networks

Speedup generated by DGNs: quasi-exhaustive exploration of the S_{map} landscape

$$\#\text{selections}(\mathcal{E}) \propto e^{S_N(\mathcal{E})}$$

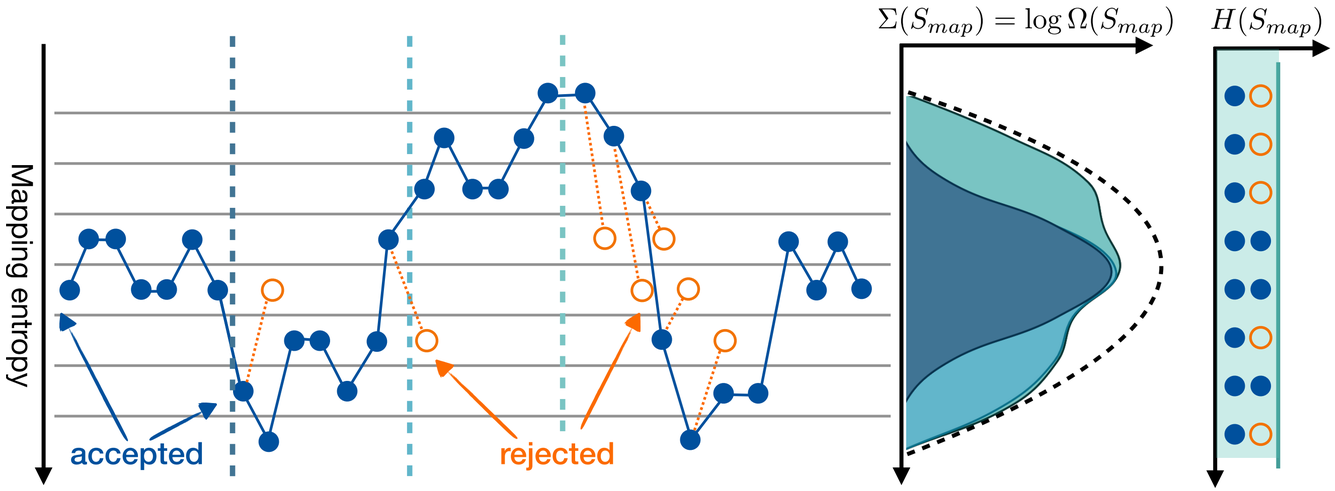


RM, Giulini, Potestio, EPJB 2021
Errica, Giulini, RM *et al.*, Front. Mol. Biosci. 2021

S_{map} via Deep Graph Networks

Speedup generated by DGNs: quasi-exhaustive exploration of the S_{map} landscape

Combine the trained DGNs with the Wang-Landau sampling algorithm **to reconstruct the density of states $\Omega(S_{map})$**

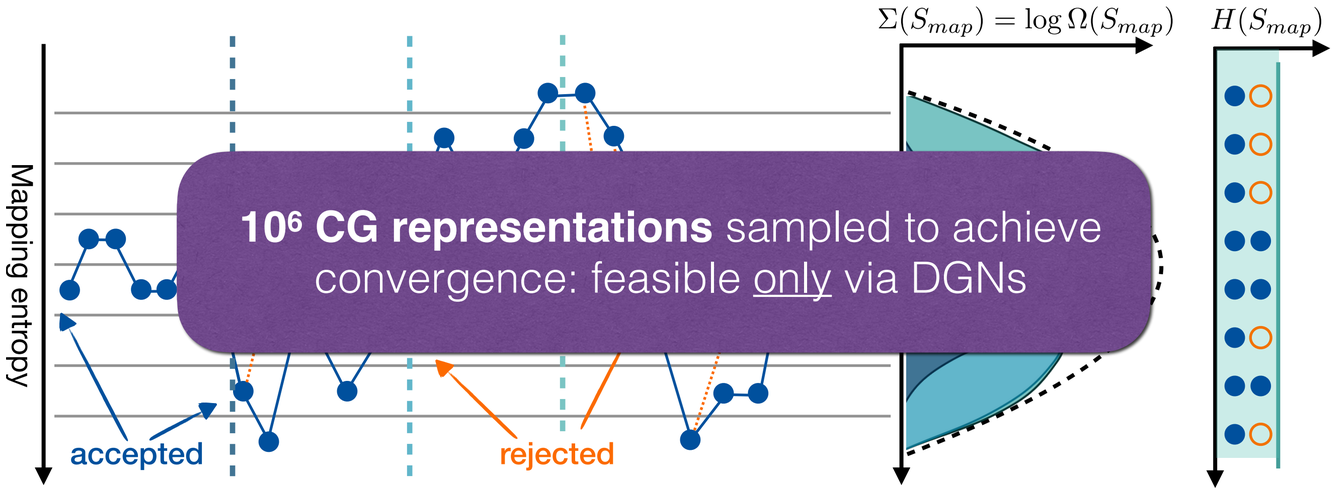


RM, Giulini, Potestio, EPJB 2021
Errica, Giulini, RM *et al.*, Front. Mol. Biosci. 2021

S_{map} via Deep Graph Networks

Speedup generated by DGNs: quasi-exhaustive exploration of the S_{map} landscape

Combine the trained DGNs with the Wang-Landau sampling algorithm **to reconstruct the density of states $\Omega(S_{map})$**



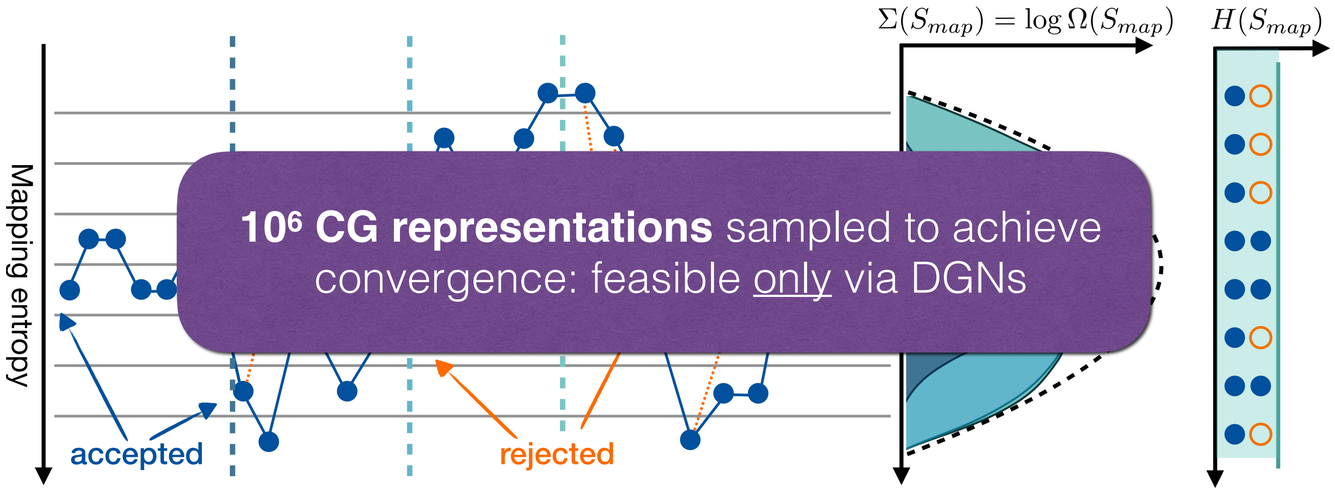
10^6 CG representations sampled to achieve convergence: feasible only via DGNs

RM, Giulini, Potestio, EPJB 2021
Errica, Giulini, RM *et al.*, Front. Mol. Biosci. 2021

S_{map} via Deep Graph Networks

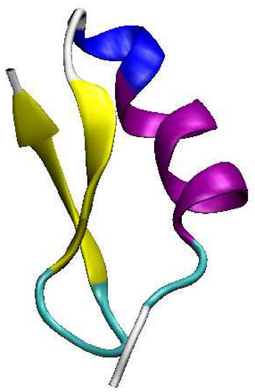
Speedup generated by DGNs: quasi-exhaustive exploration of the S_{map} landscape

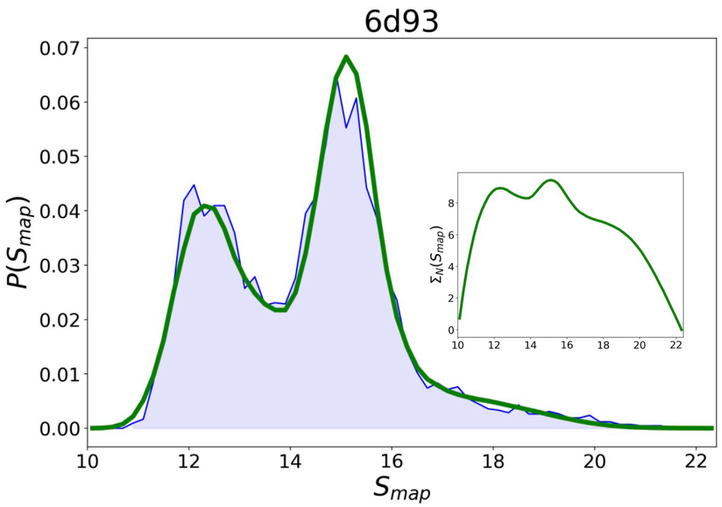
Combine the trained DGNs with the Wang-Landau sampling algorithm **to reconstruct the density of states $\Omega(S_{map})$**



10^6 CG representations sampled to achieve convergence: feasible only via DGNs

Tamapin



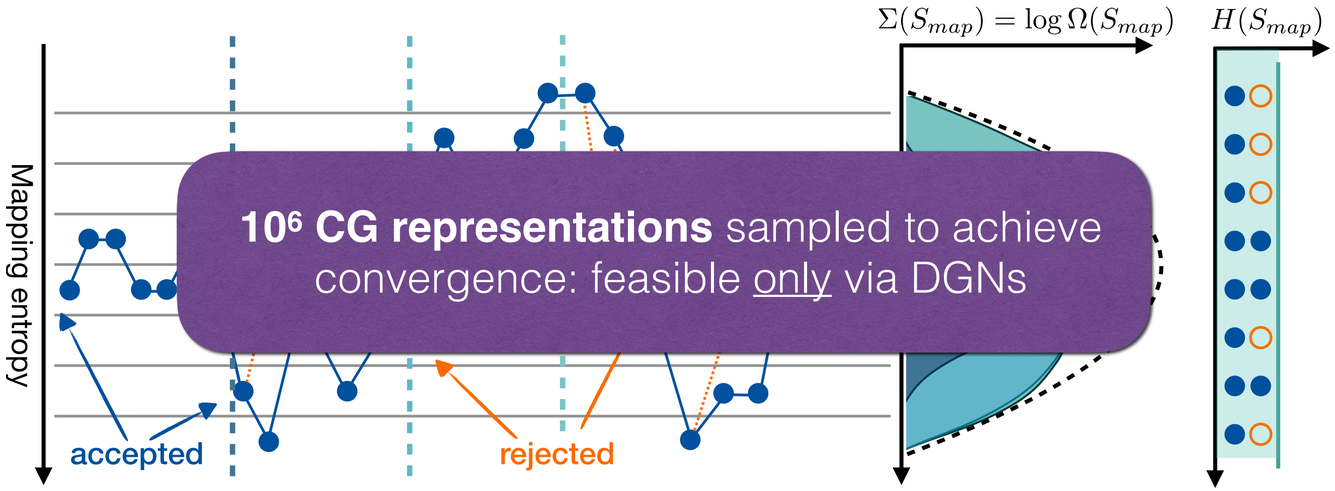


RM, Giulini, Potestio, EPJB 2021
Errica, Giulini, RM *et al.*, Front. Mol. Biosci. 2021

S_{map} via Deep Graph Networks

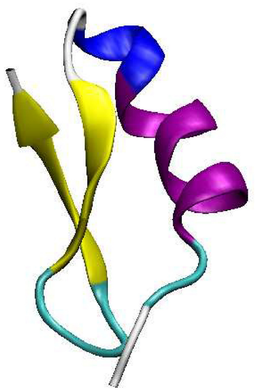
Speedup generated by DGNs: quasi-exhaustive exploration of the S_{map} landscape

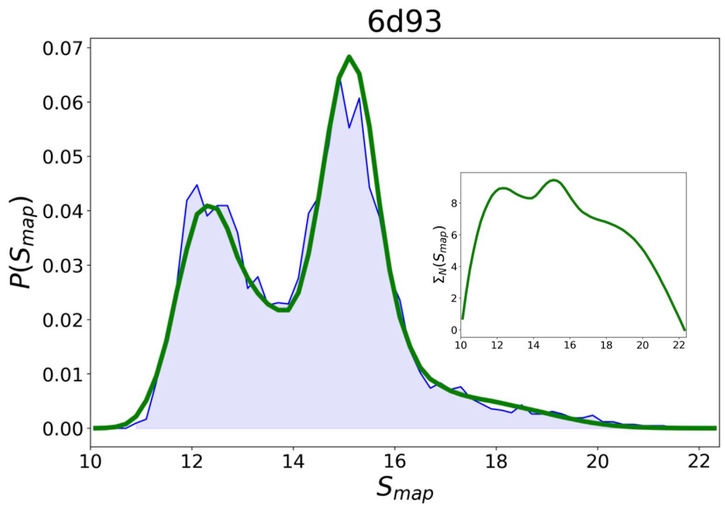
Combine the trained DGNs with the Wang-Landau sampling algorithm **to reconstruct the density of states $\Omega(S_{map})$**

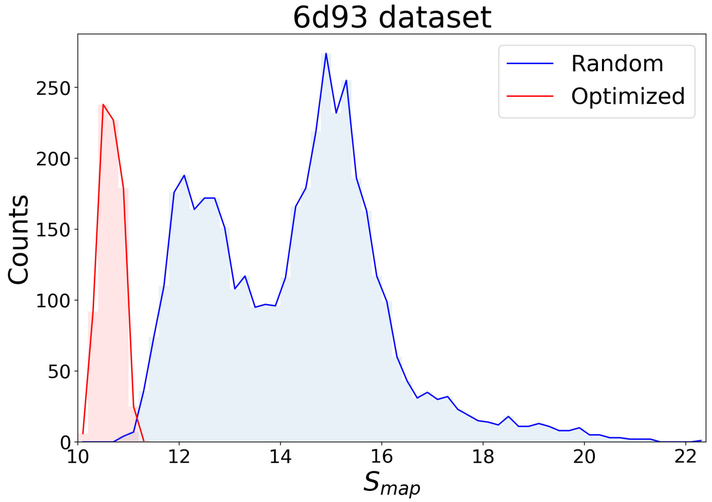


10^6 CG representations sampled to achieve convergence: feasible only via DGNs

Tamapin





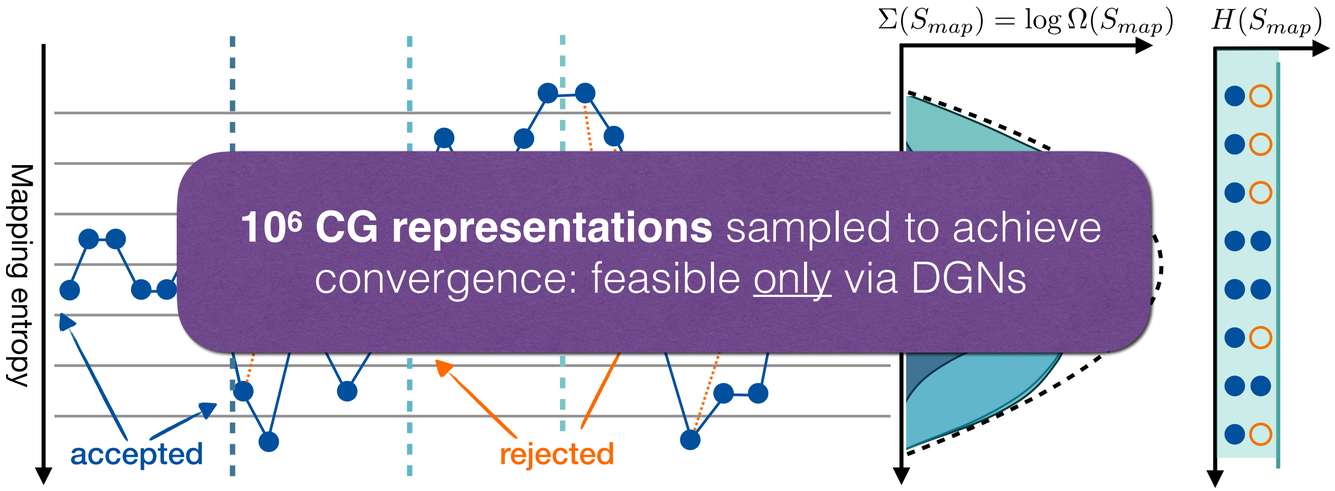


RM, Giulini, Potestio, EPJB 2021
Errica, Giulini, RM *et al.*, Front. Mol. Biosci. 2021

S_{map} via Deep Graph Networks

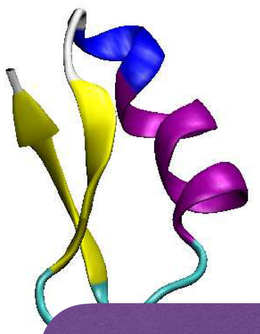
Speedup generated by DGNs: quasi-exhaustive exploration of the S_{map} landscape

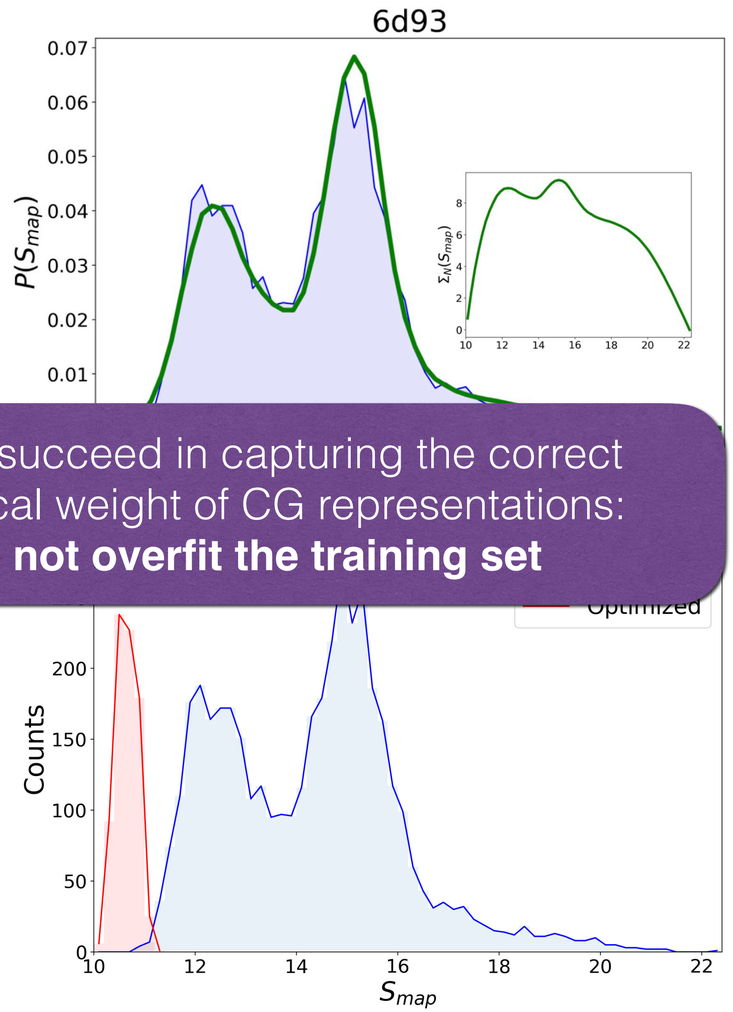
Combine the trained DGNs with the Wang-Landau sampling algorithm **to reconstruct the density of states $\Omega(S_{map})$**



10^6 CG representations sampled to achieve convergence: feasible only via DGNs

Tamapin



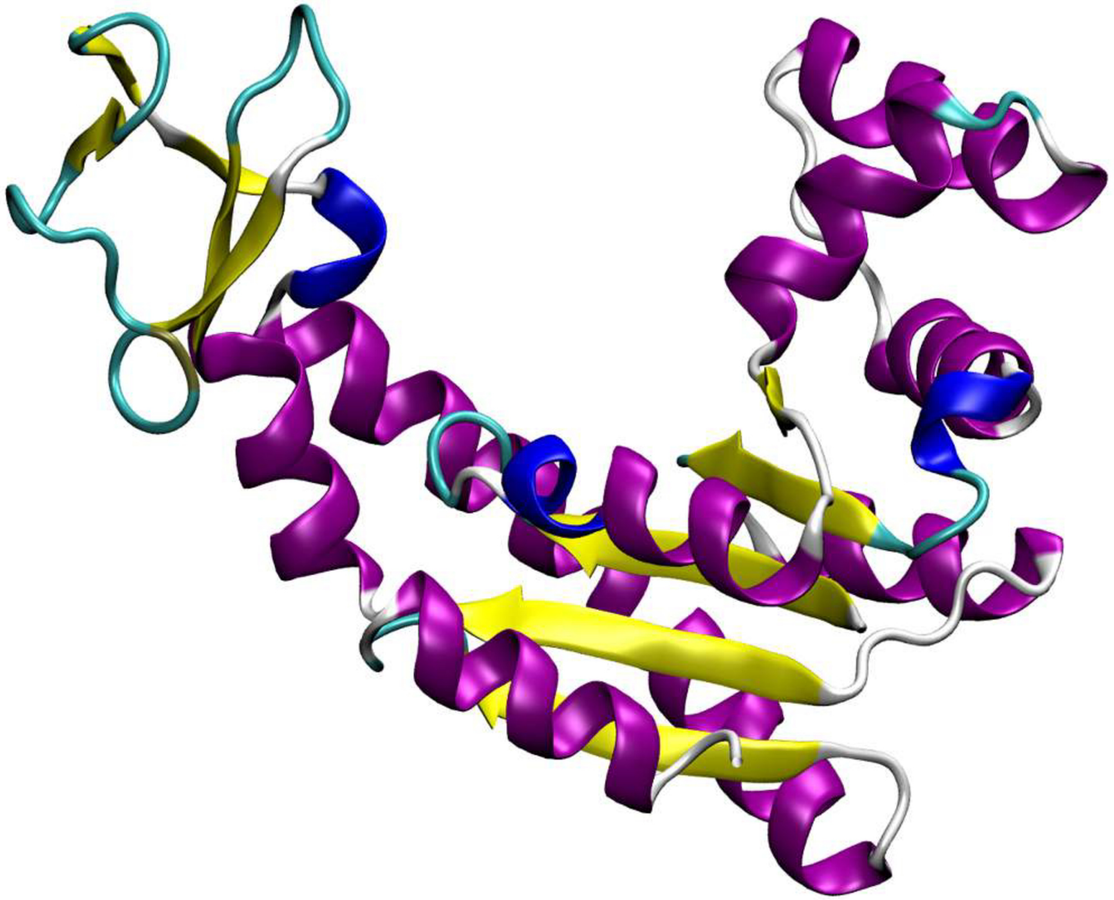


DGNs succeed in capturing the correct statistical weight of CG representations: **do not overfit the training set**

RM, Giulini, Potestio, EPJB 2021
Errica, Giulini, RM *et al.*, Front. Mol. Biosci. 2021

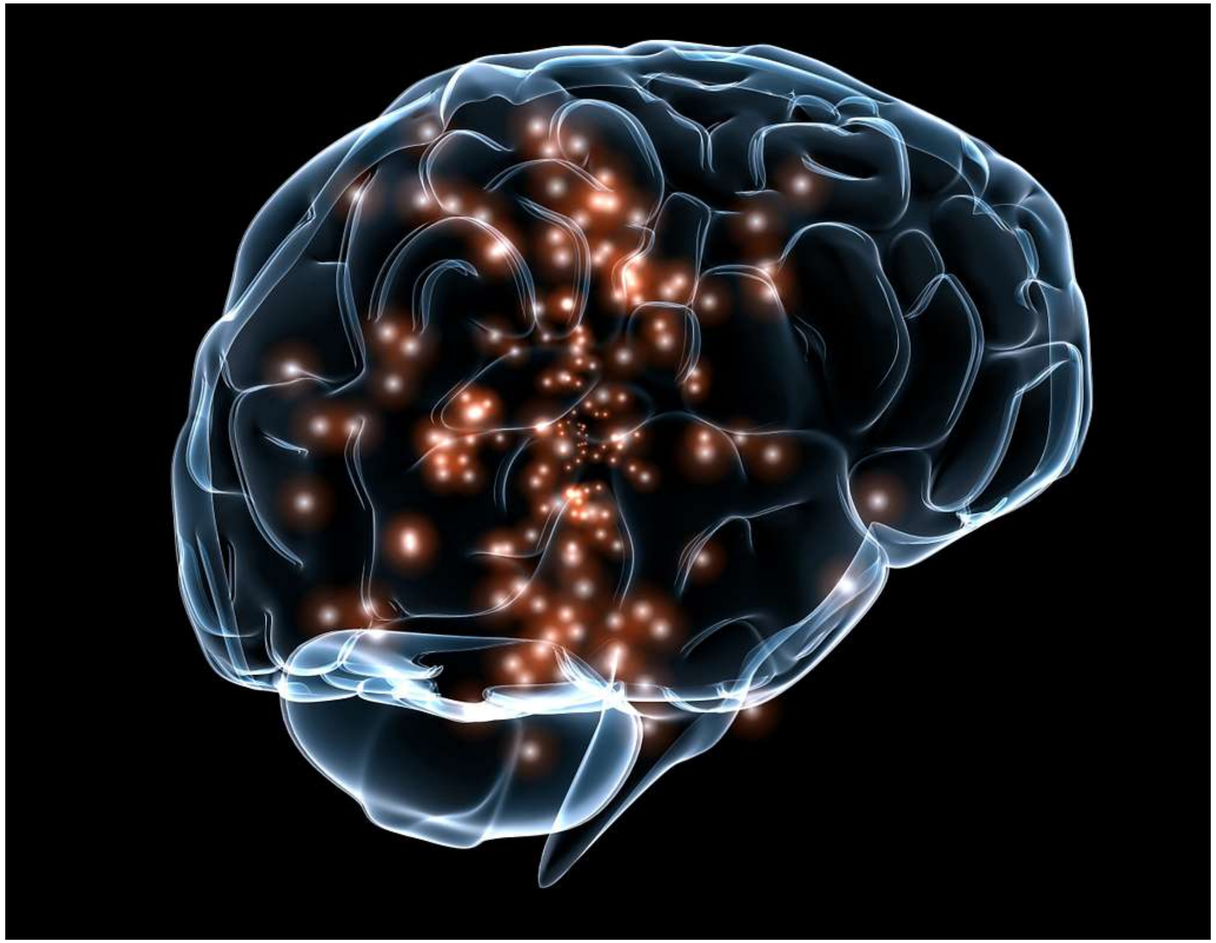
Application to neural systems

Being based on information-theoretic quantities, the mapping entropy protocol is extremely general

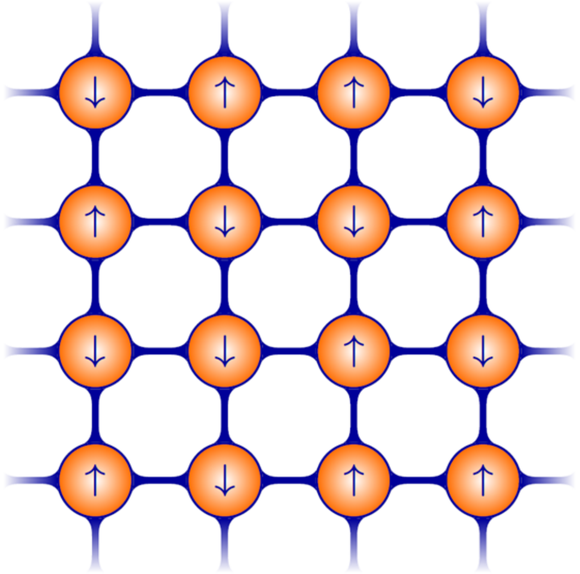


Application to neural systems

Being based on information-theoretic quantities, the mapping entropy protocol is extremely general



Neural system: Hopfield model



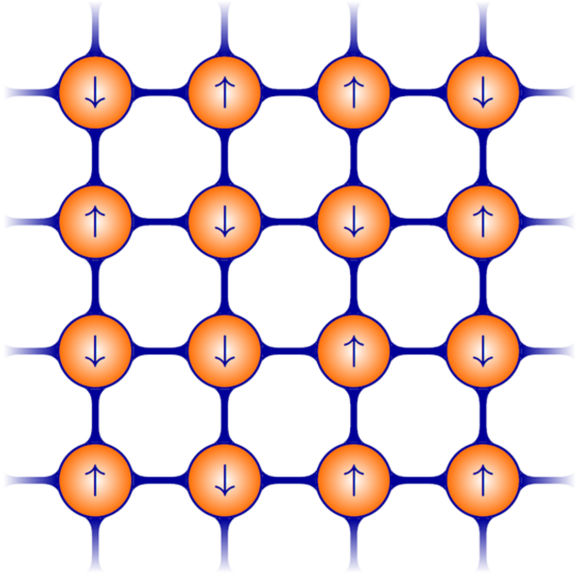
$$H(\sigma_1, ..., \sigma_N) = -\frac{1}{2}\sum_{i \neq j} J_{ij}\sigma_i\sigma_j$$

Couplings $J_{ij} = \sum_{\mu=1}^{p} \xi_i^\mu \xi_j^\mu$

Memory patterns $\xi_i^\mu, \quad \mu = 1, ..., p$

Hopfield, PNAS (1982)
Amit, *Modeling Brain Function: The World of Attractor Neural Networks* (1989)
Aldrigo, RM, Potestio (in preparation)

Neural system: Hopfield model



$$H(\sigma_1, ..., \sigma_N) = -\frac{1}{2}\sum_{i \neq j} J_{ij}\sigma_i\sigma_j$$

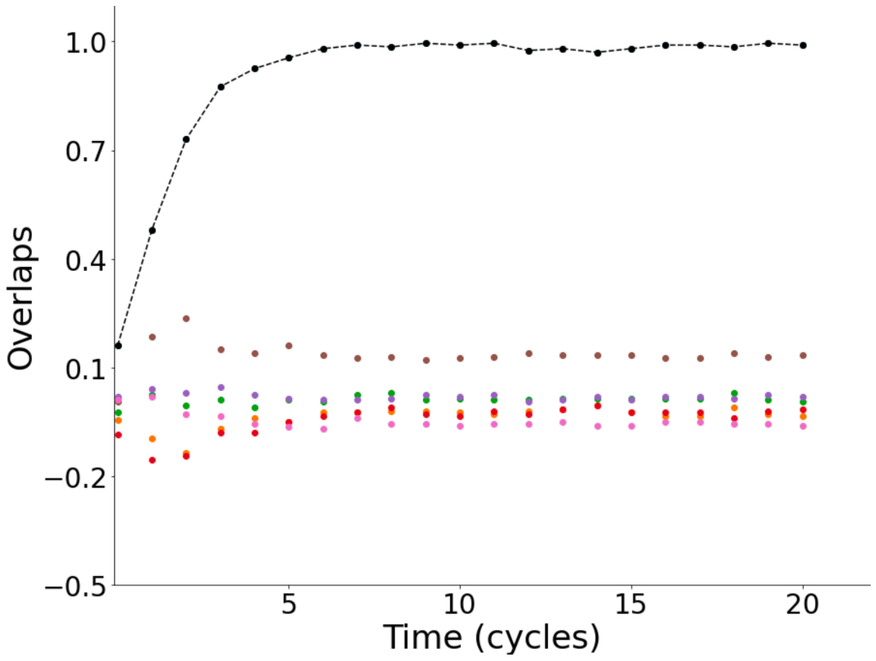
Couplings $J_{ij} = \sum_{\mu=1}^{p} \xi_i^\mu \xi_j^\mu$

Memory patterns $\xi_i^\mu, \quad \mu = 1, ..., p$

Nonlinear evolution of the configurations

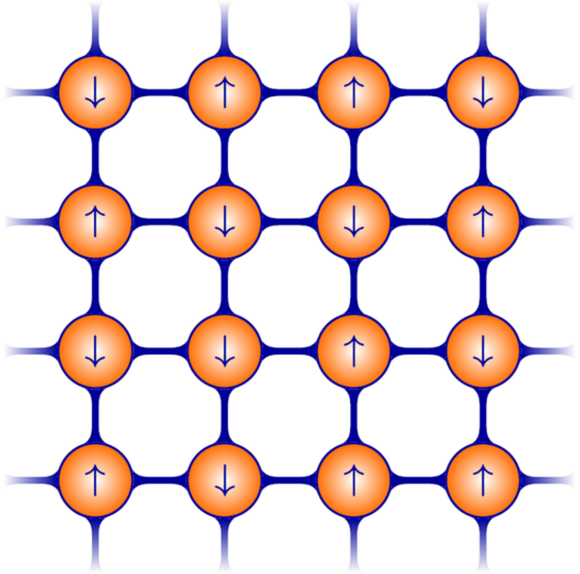
Model's dynamics retrieves memory patterns

Overlaps $m^\mu = \frac{1}{N}\sum_i \xi_i^\mu \sigma_i$



Hopfield, PNAS (1982)
Amit, *Modeling Brain Function: The World of Attractor Neural Networks* (1989)
Aldrigo, RM, Potestio (in preparation)

Neural system: Hopfield model



$$H(\sigma_1, ..., \sigma_N) = -\frac{1}{2}\sum_{i \neq j} J_{ij}\sigma_i\sigma_j$$

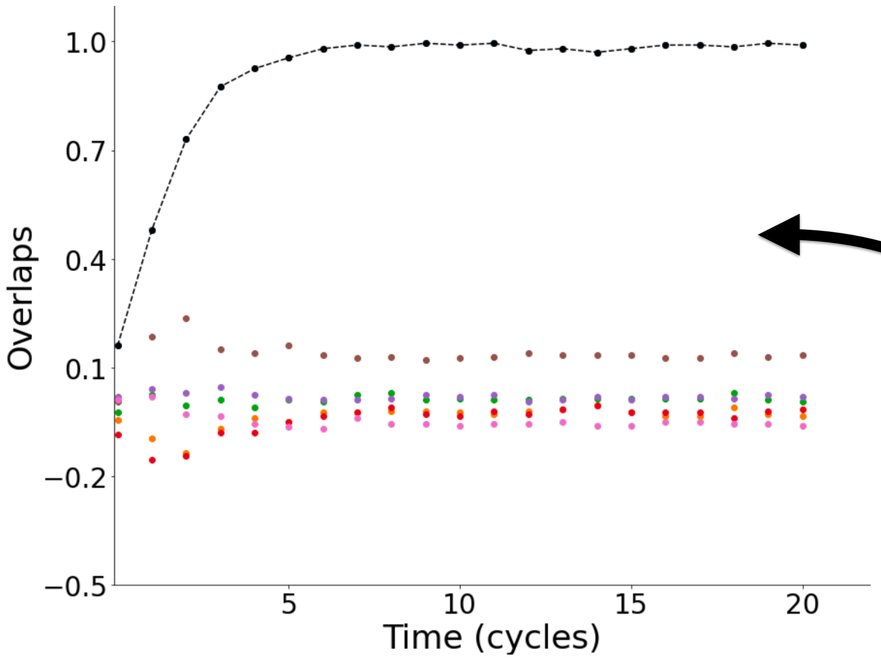
Couplings $J_{ij} = \sum_{\mu=1}^{p} \xi_i^\mu \xi_j^\mu$

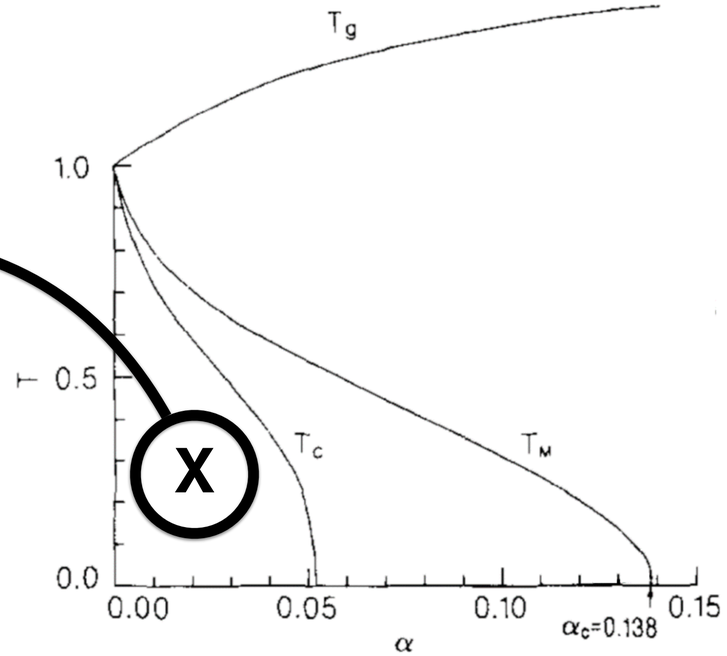
Memory patterns $\xi_i^\mu, \quad \mu = 1, ..., p$

Nonlinear evolution of the configurations

Model's dynamics retrieves memory patterns

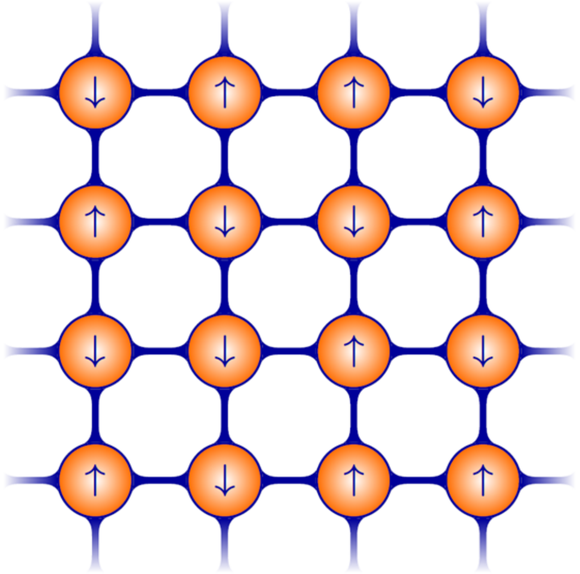
Overlaps $m^\mu = \frac{1}{N}\sum_i \xi_i^\mu \sigma_i$





Hopfield, PNAS (1982)
Amit, *Modeling Brain Function: The World of Attractor Neural Networks* (1989)
Aldrigo, RM, Potestio (in preparation)

Neural system: Hopfield model



$$H(\sigma_1, ..., \sigma_N) = -\frac{1}{2}\sum_{i \neq j} J_{ij}\sigma_i\sigma_j$$

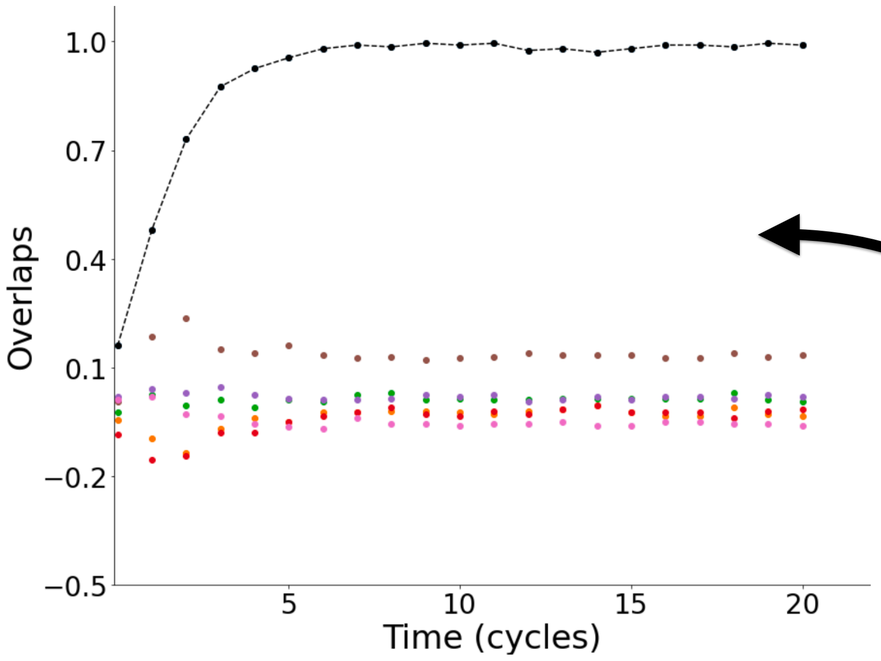
Couplings $J_{ij} = \sum_{\mu=1}^{p} \xi_i^\mu \xi_j^\mu$

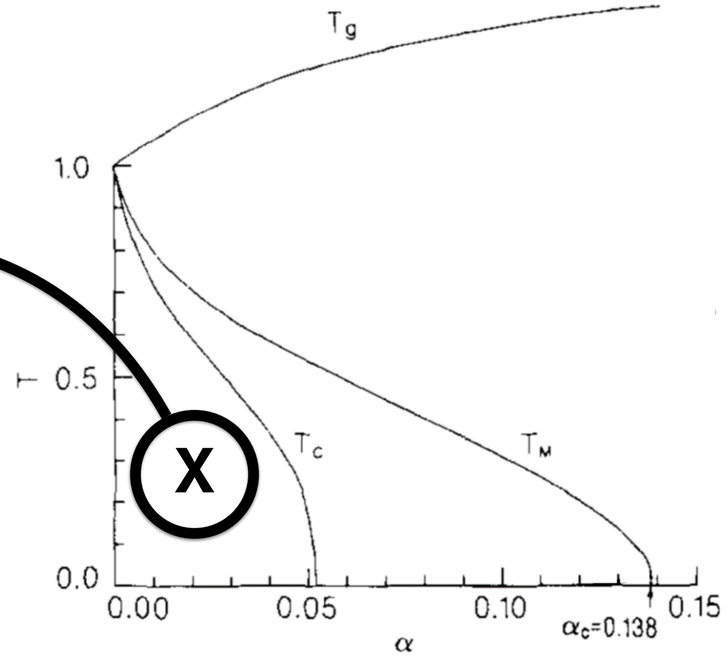
Memory patterns $\xi_i^\mu, \quad \mu = 1, ..., p$

Nonlinear evolution of the configurations

Model's dynamics retrieves memory patterns

Overlaps $m^\mu = \frac{1}{N}\sum_i \xi_i^\mu \sigma_i$





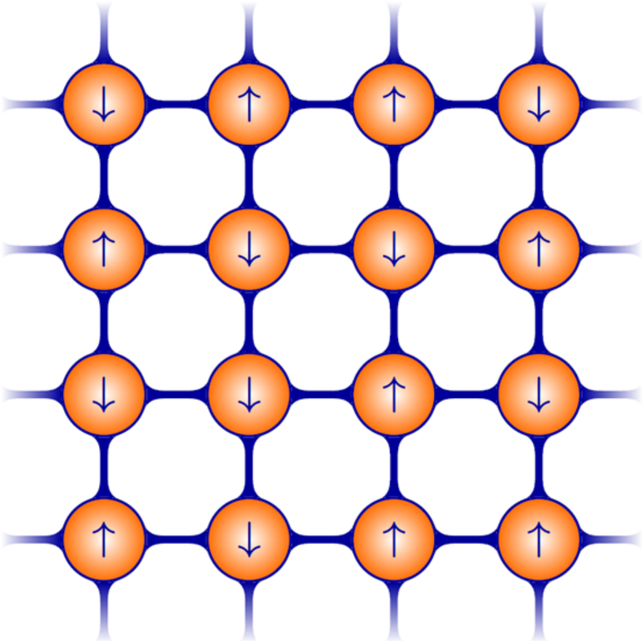
Looking for the **most informative** neurons!

Hopfield, PNAS (1982)
Amit, *Modeling Brain Function: The World of Attractor Neural Networks* (1989)
Aldrigo, RM, Potestio (in preparation)

Detection of maximally-informative neurons

CG'ing the Hopfield model

1) Simulate the Hopfield model



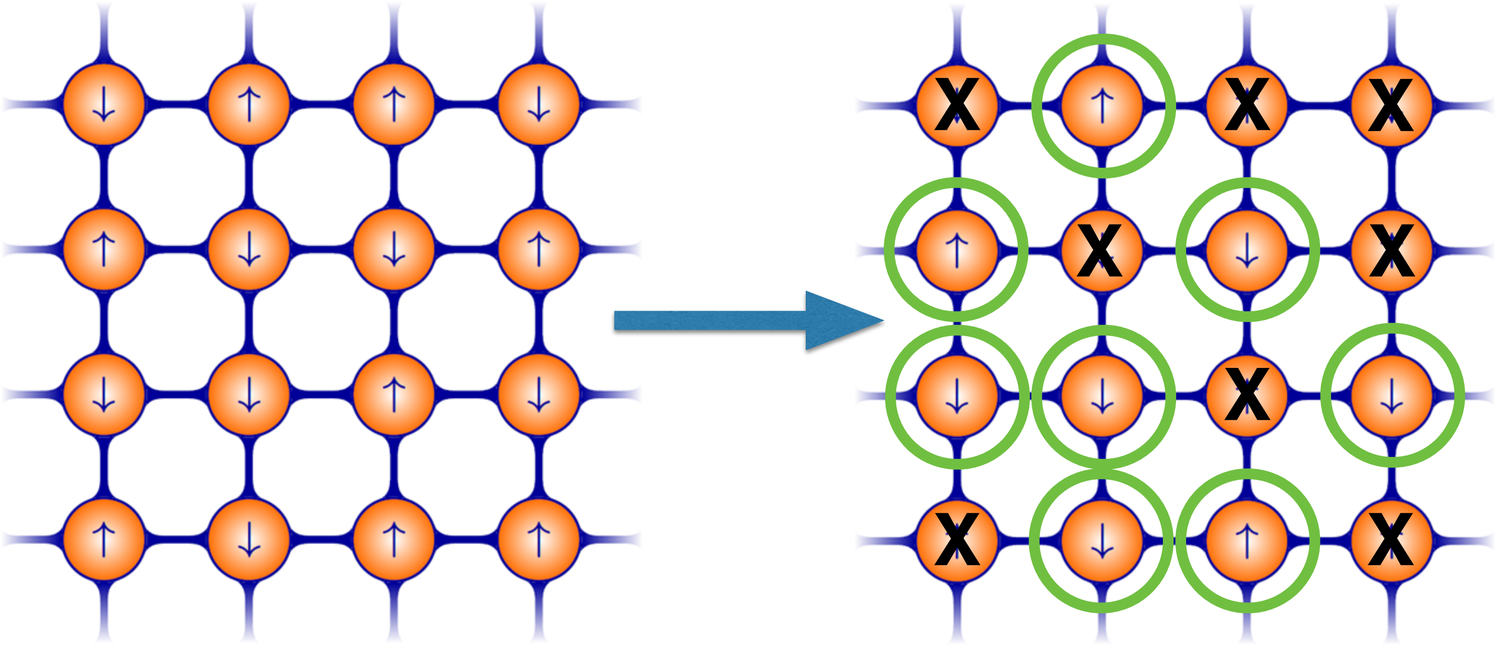
Aldrigo, RM, Potestio (in preparation)

Detection of maximally-informative neurons

CG'ing the Hopfield model

1) Simulate the Hopfield model

2) Select a subset of retained neurons



Aldrigo, RM, Potestio (in preparation)

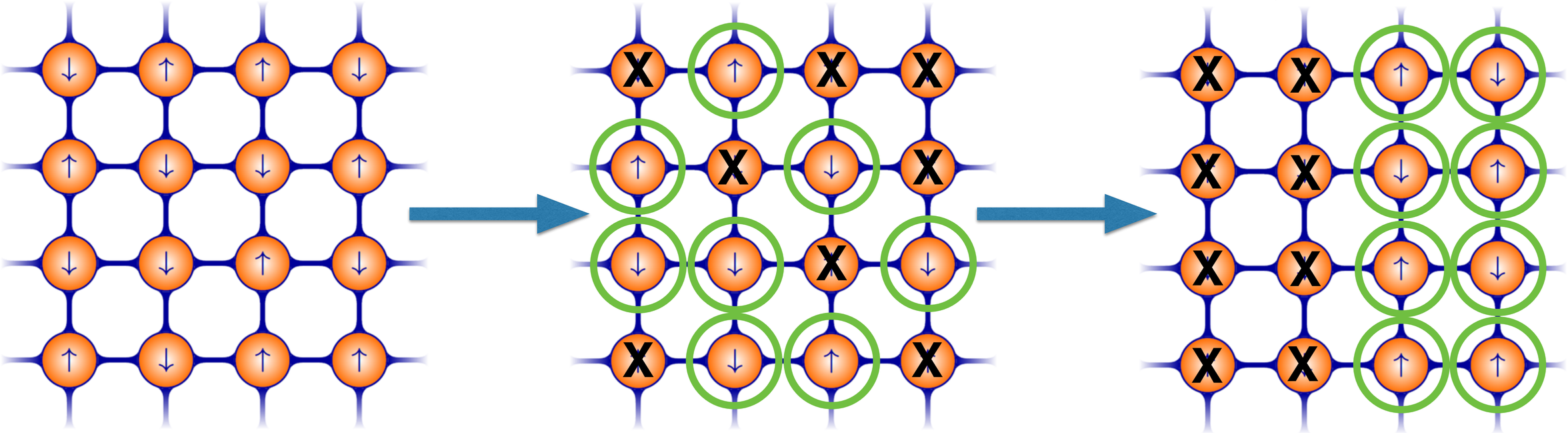
Detection of maximally-informative neurons

CG'ing the Hopfield model

1) Simulate the Hopfield model

2) Select a subset of retained neurons

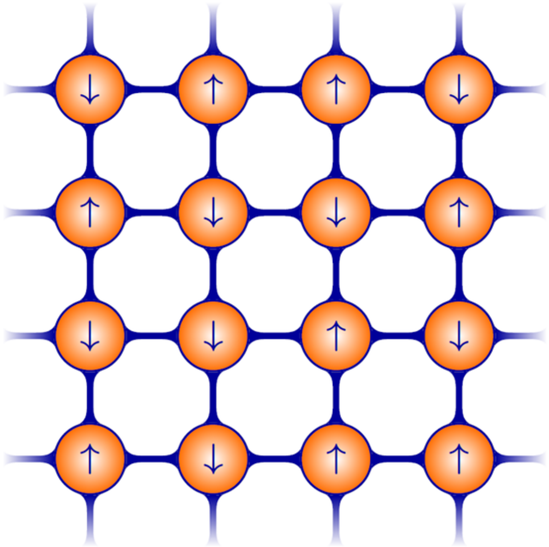
3) Optimize the selection to detect **maximally-informative** neurons



Aldrigo, RM, Potestio (in preparation)

Resolution/Mapping-Entropy analysis

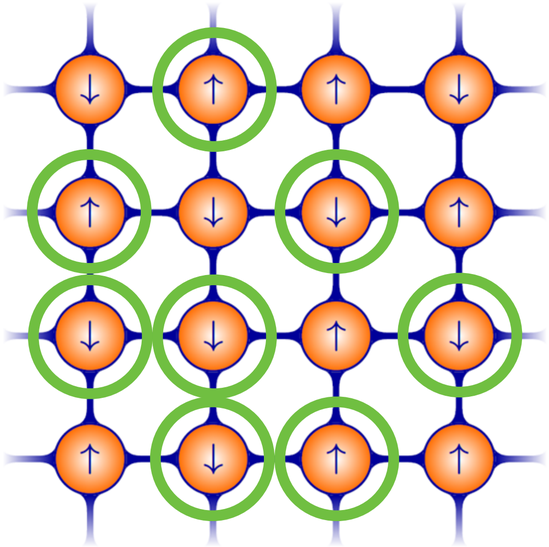
- Simulate the high-resolution Hopfield model
- **Empirical** reference probability $p(\sigma_1, ..., \sigma_N)$



Aldrigo, RM, Potestio (in preparation)

Resolution/Mapping-Entropy analysis

- Simulate the high-resolution Hopfield model
- **Empirical** reference probability $p(\sigma_1, ..., \sigma_N)$

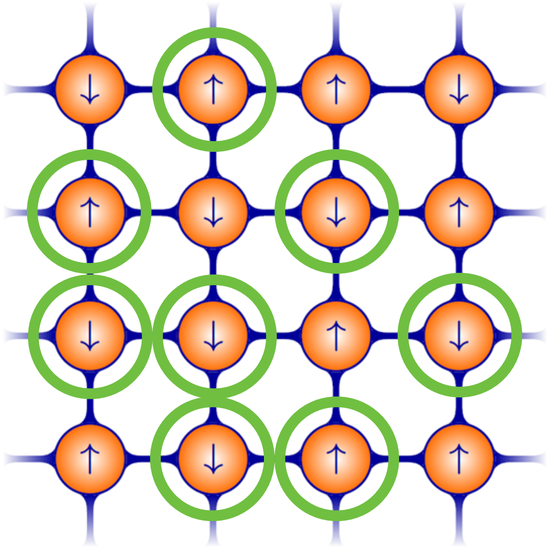


- Select n<N retained neurons S_i
- **Empirical** CG probability $P_M(S_1, ..., S_n)$
- **Empirical** backmapped probability $\bar{p}_M(\sigma_1, ..., \sigma_N)$

Aldrigo, RM, Potestio (in preparation)

Resolution/Mapping-Entropy analysis

- Simulate the high-resolution Hopfield model
- **Empirical** reference probability $p(\sigma_1, ..., \sigma_N)$



- Select n<N retained neurons S_i
- **Empirical** CG probability $P_M(S_1, ..., S_n)$
- **Empirical** backmapped probability $\bar{p}_M(\sigma_1, ..., \sigma_N)$

Resolution of the neuron selection

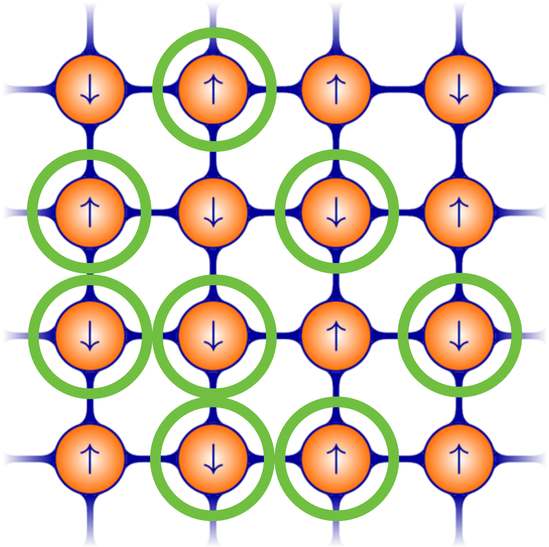
$$\mathcal{H}_M = -\sum_{\{S_i\}} P(S_1, ..., S_n) \ln P(S_1, ..., S_n)$$

- **Depends** on the specific selection
- **Decreases** by decreasing the number of retained neurons

Aldrigo, RM, Potestio (in preparation)

Resolution/Mapping-Entropy analysis

- Simulate the high-resolution Hopfield model
- **Empirical** reference probability $p(\sigma_1, ..., \sigma_N)$



- Select n<N retained neurons S_i
- **Empirical** CG probability $P_M(S_1, ..., S_n)$
- **Empirical** backmapped probability $\bar{p}_M(\sigma_1, ..., \sigma_N)$

Resolution of the neuron selection

$$\mathcal{H}_M = -\sum_{\{S_i\}} P(S_1, ..., S_n) \ln P(S_1, ..., S_n)$$

- **Depends** on the specific selection
- **Decreases** by decreasing the number of retained neurons

Information loss generated by the selection: mapping entropy

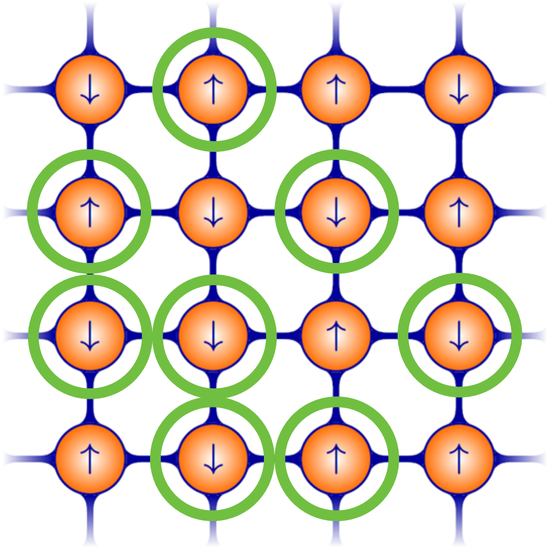
$$S_M^{map} = \sum_{\{\sigma_i\}} p(\{\sigma_i\}) \ln \left(\frac{p(\{\sigma_i\})}{\bar{p}_M(\{\sigma_i\})} \right)$$

- **Depends** on the specific selection
- **Increases** by decreasing the number of retained neurons

Aldrigo, RM, Potestio (in preparation)

Resolution/Mapping-Entropy analysis

- Simulate the high-resolution Hopfield model
- **Empirical** reference probability $p(\sigma_1, ..., \sigma_N)$



- Select n<N retained neurons S_i
- **Empirical** CG probability $P_M(S_1, ..., S_n)$
- **Empirical** backmapped probability $\bar{p}_M(\sigma_1, ..., \sigma_N)$

Resolution of the neuron selection

$$\mathcal{H}_M = -\sum_{\{S_i\}} P(S_1, ..., S_n) \ln P(S_1, ..., S_n)$$

- **Depends** on the specific selection
- **Decreases** by decreasing the number of retained neurons

Maximally informative neurons:
minimize the mapping entropy
in the space of possible selections!

Information loss generated by the selection: mapping entropy

$$S_M^{map} = \sum_{\{\sigma_i\}} p(\{\sigma_i\}) \ln \left(\frac{p(\{\sigma_i\})}{\bar{p}_M(\{\sigma_i\})} \right)$$

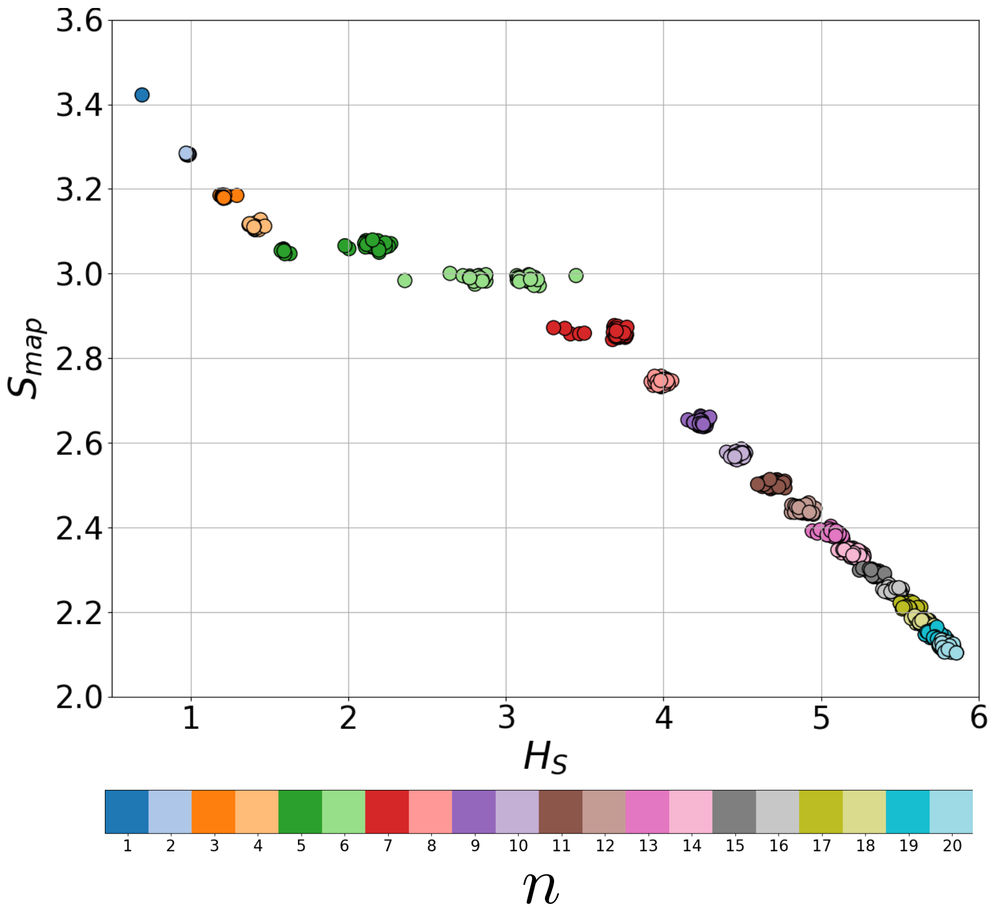
- **Depends** on the specific selection
- **Increases** by decreasing the number of retained neurons

Aldrigo, RM, Potestio (in preparation)

Resolution/Mapping-Entropy analysis

Maximally informative selection of neurons that **minimize the mapping entropy**

Hopfield model with N=100 neurons and 5 memory patterns

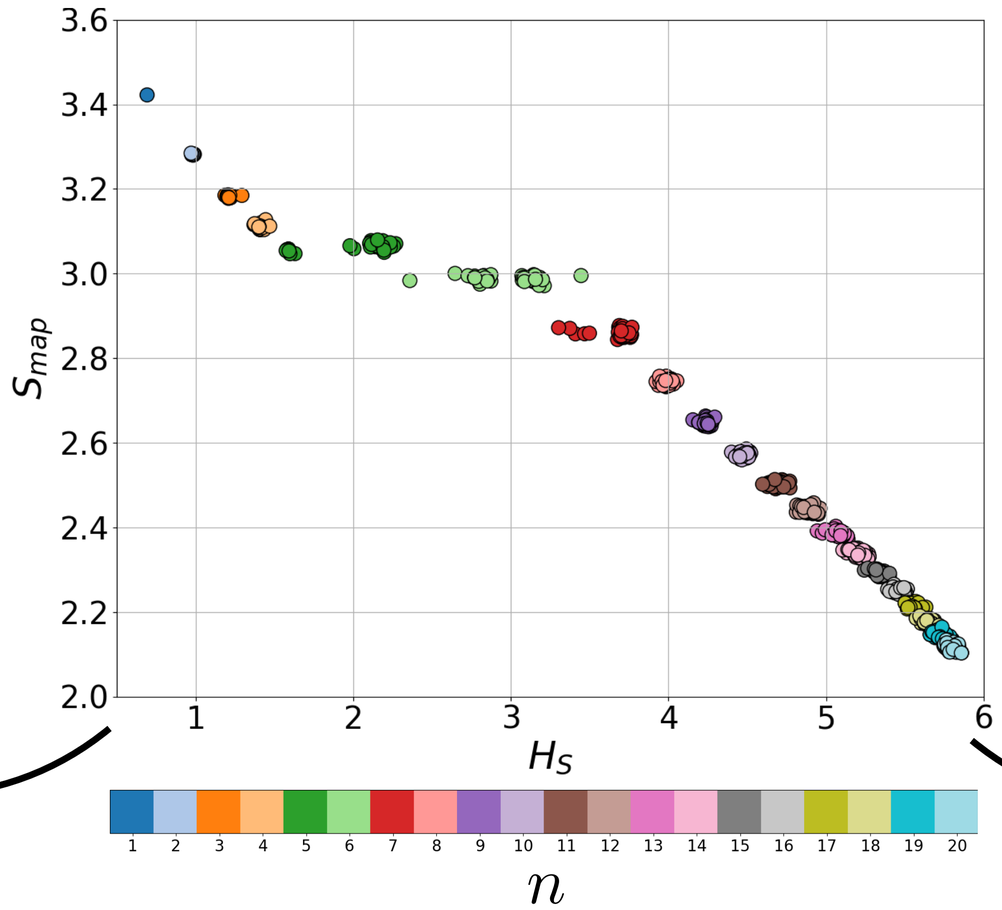


Aldrigo, RM, Potestio (in preparation)

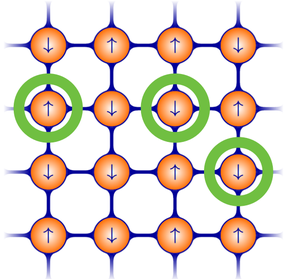
Resolution/Mapping-Entropy analysis

Maximally informative selection of neurons that **minimize the mapping entropy**

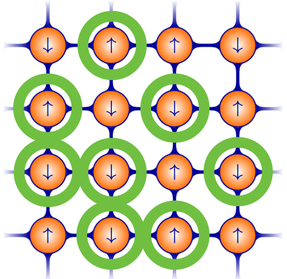
Hopfield model with N=100 neurons and 5 memory patterns



Low resolution



High resolution

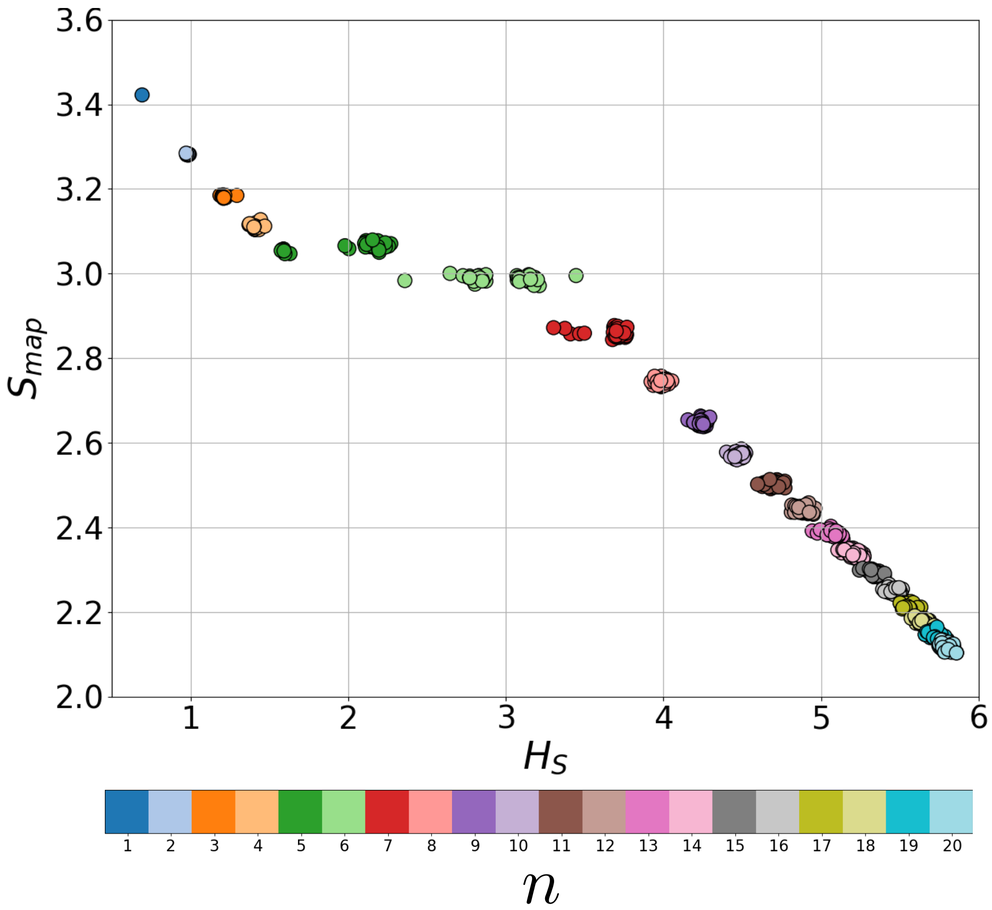


Aldrigo, RM, Potestio (in preparation)

Resolution/Mapping-Entropy analysis

Maximally informative selection of neurons that **minimize the mapping entropy**

Hopfield model with N=100 neurons and 5 memory patterns



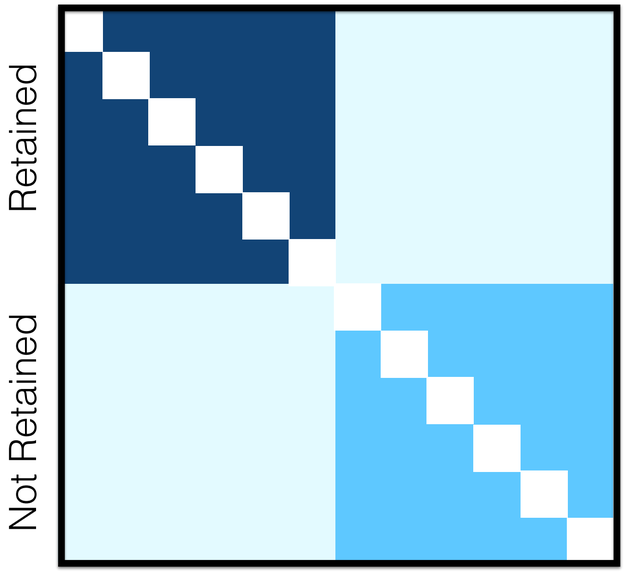
Aldrigo, RM, Potestio (in preparation)

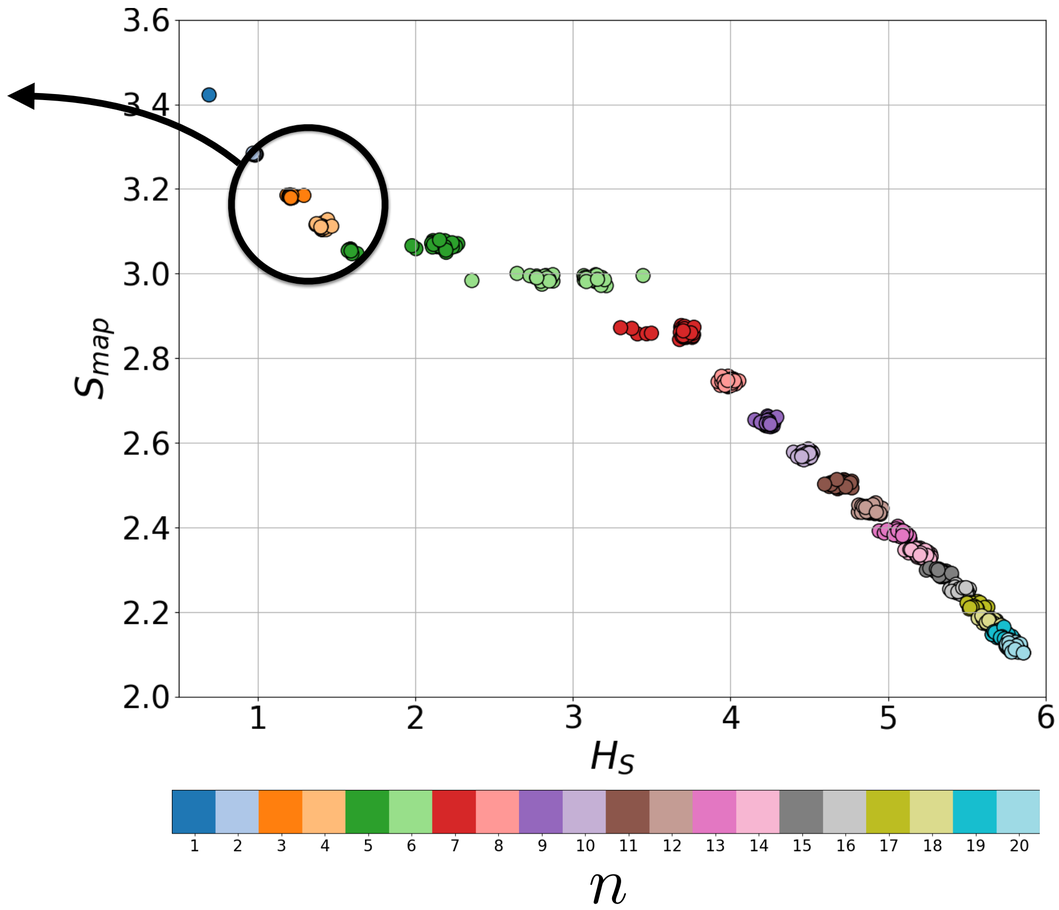
Resolution/Mapping-Entropy analysis

Maximally informative selection of neurons that **minimize the mapping entropy**

Hopfield model with N=100 neurons and 5 memory patterns

Ordered coupling matrix $|J_{ij}|$





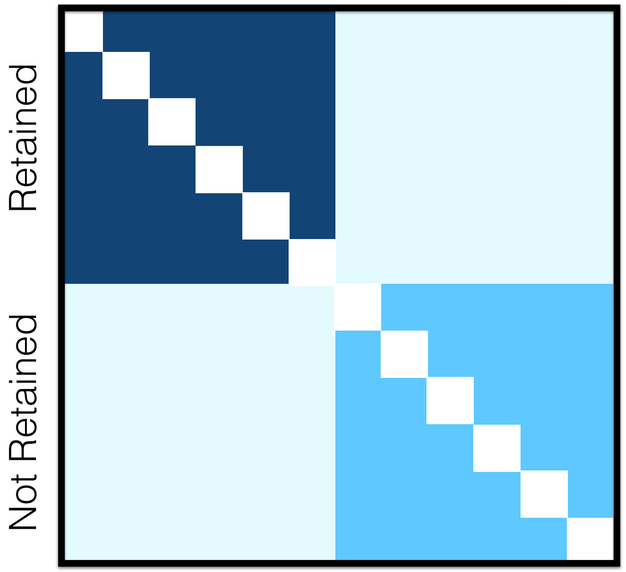
Aldrigo, RM, Potestio (in preparation)

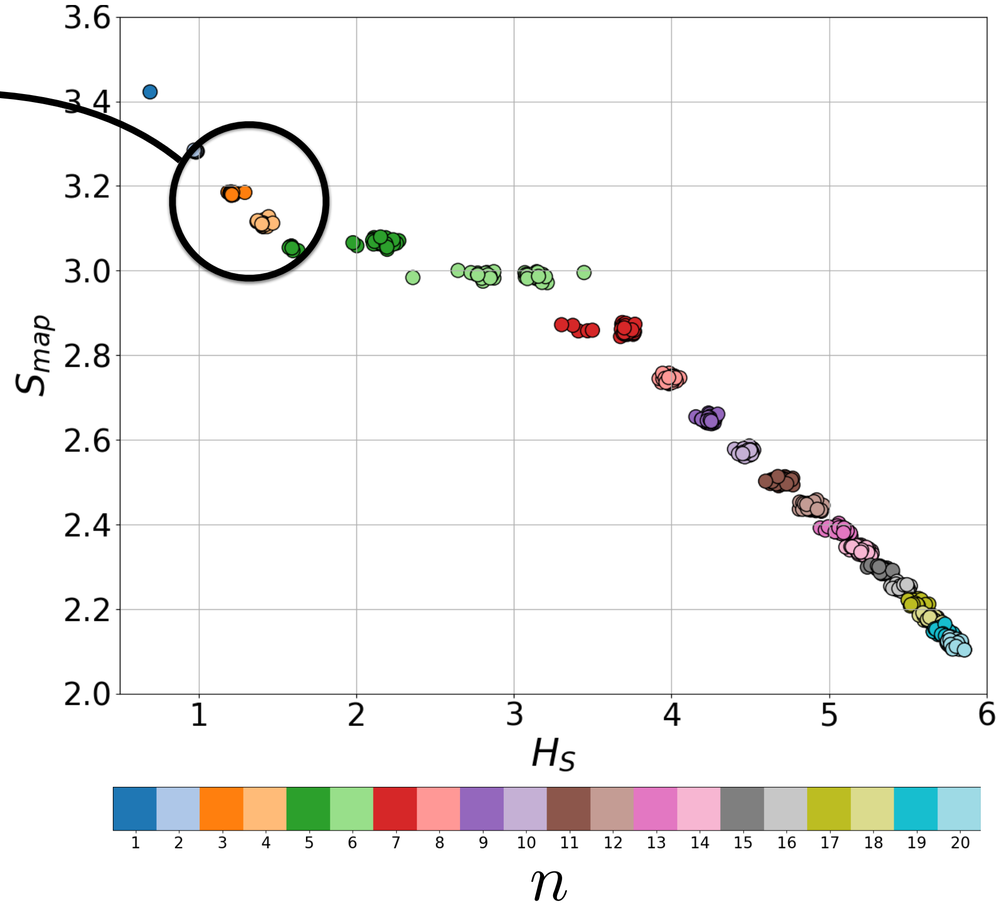
Resolution/Mapping-Entropy analysis

Maximally informative selection of neurons that **minimize the mapping entropy**

Hopfield model with N=100 neurons and 5 memory patterns

Ordered coupling matrix $|J_{ij}|$





Strongly interacting retained neurons, weakly coupled with the integrated ones

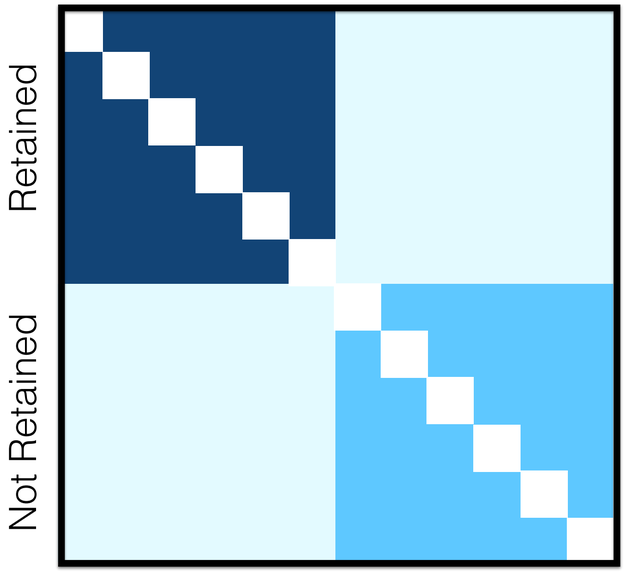
Aldrigo, RM, Potestio (in preparation)

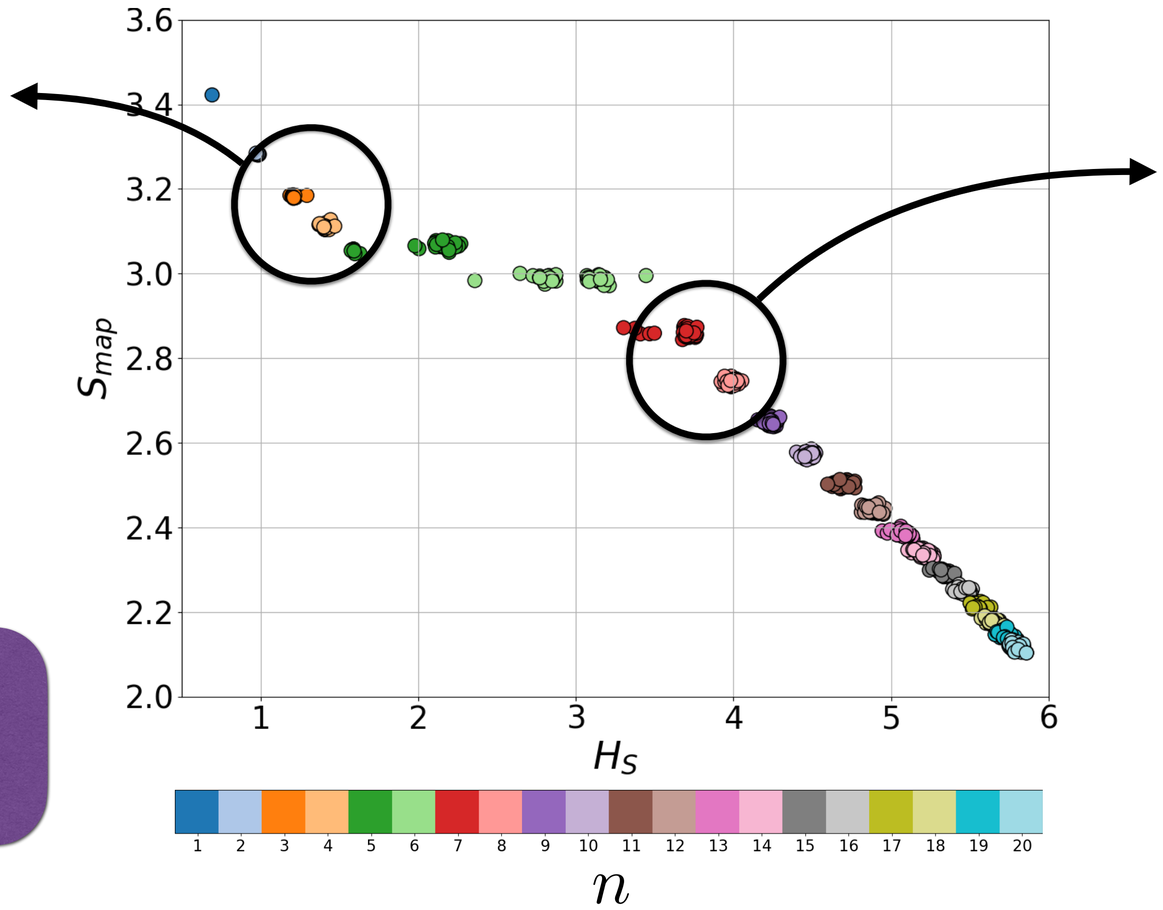
Resolution/Mapping-Entropy analysis

Maximally informative selection of neurons that **minimize the mapping entropy**

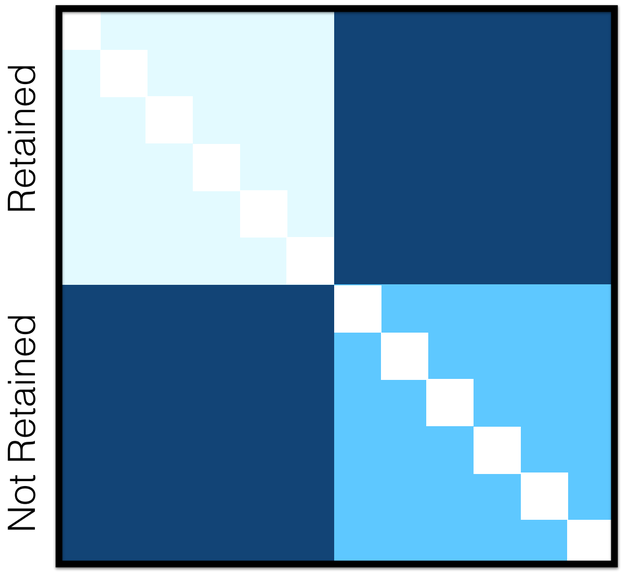
Hopfield model with N=100 neurons and 5 memory patterns

Ordered coupling matrix $|J_{ij}|$





Ordered coupling matrix $|J_{ij}|$



Strongly interacting retained neurons, weakly coupled with the integrated ones

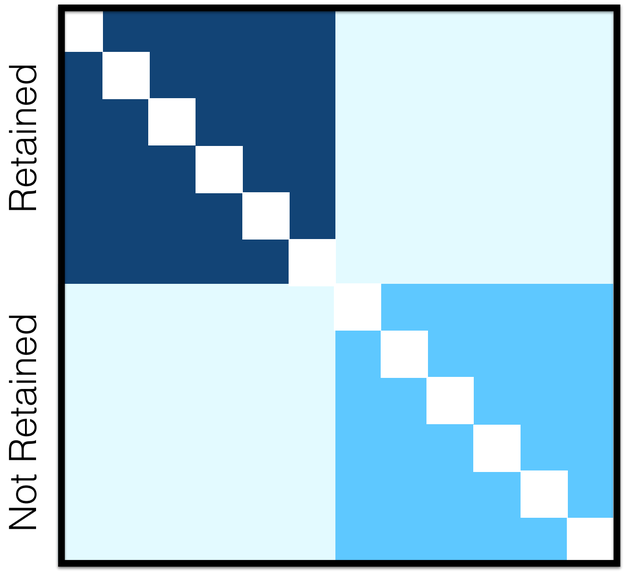
Aldrigo, RM, Potestio (in preparation)

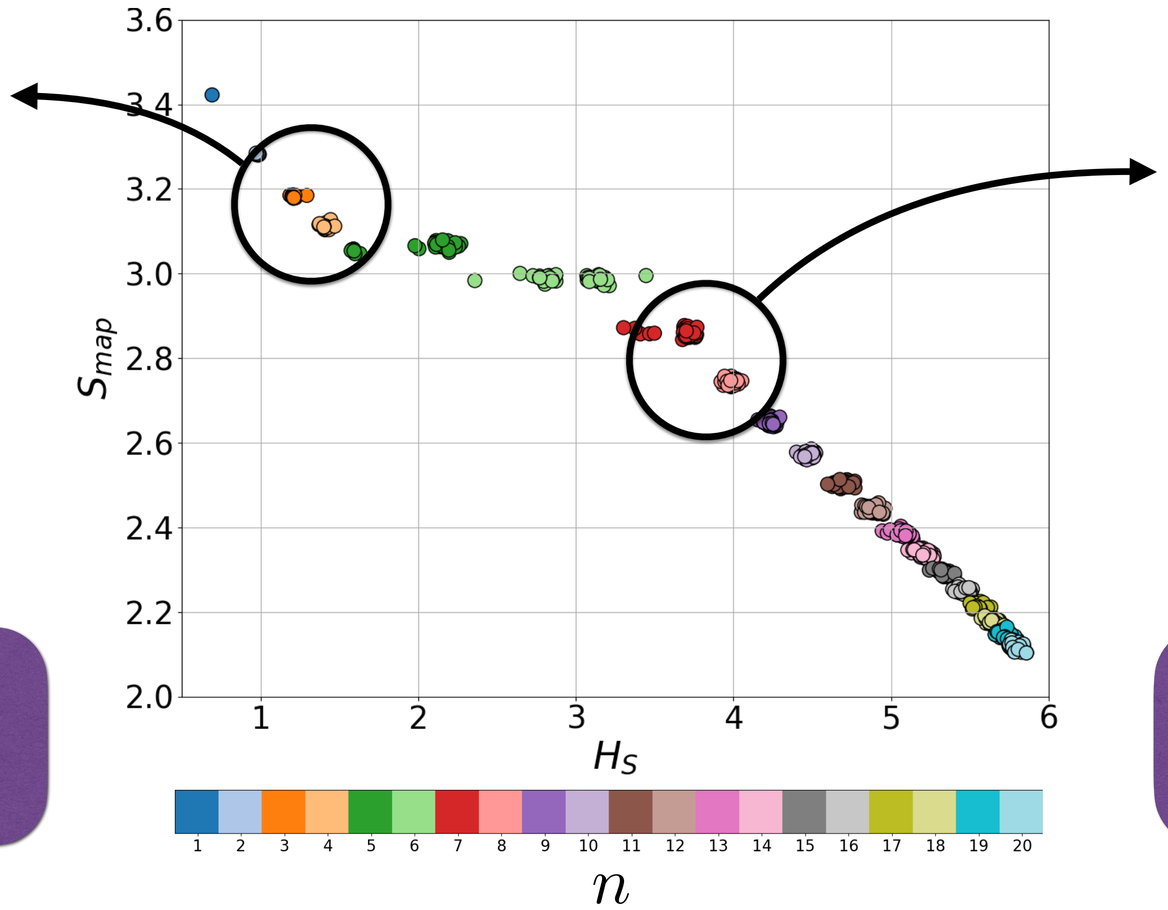
Resolution/Mapping-Entropy analysis

Maximally informative selection of neurons that **minimize the mapping entropy**

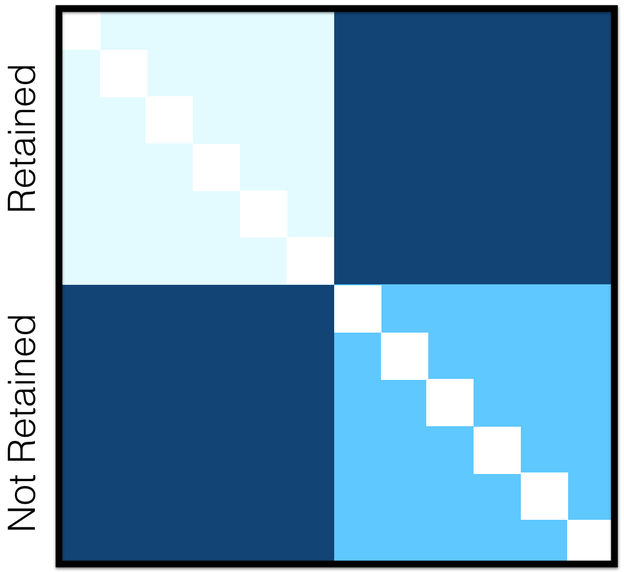
Hopfield model with N=100 neurons and 5 memory patterns

Ordered coupling matrix $|J_{ij}|$





Ordered coupling matrix $|J_{ij}|$



Strongly interacting retained neurons, weakly coupled with the integrated ones

Weakly interacting retained neurons, strongly coupled with the integrated ones

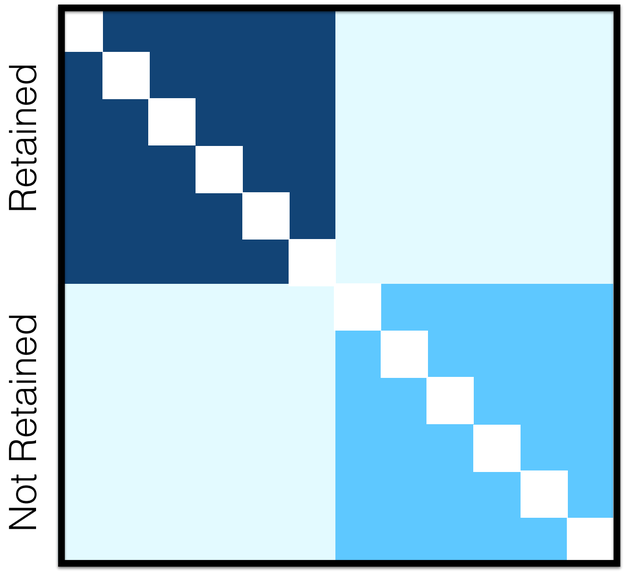
Aldrigo, RM, Potestio (in preparation)

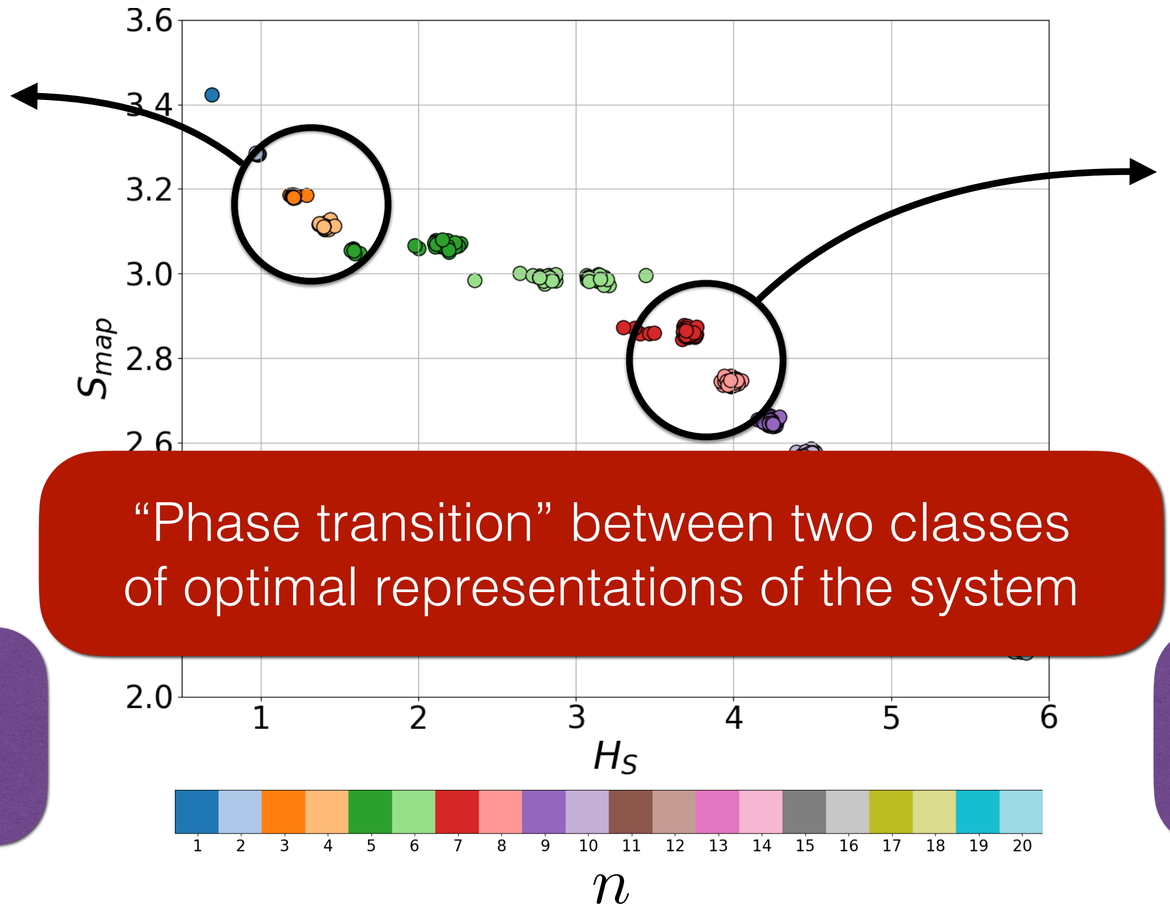
Resolution/Mapping-Entropy analysis

Maximally informative selection of neurons that **minimize the mapping entropy**

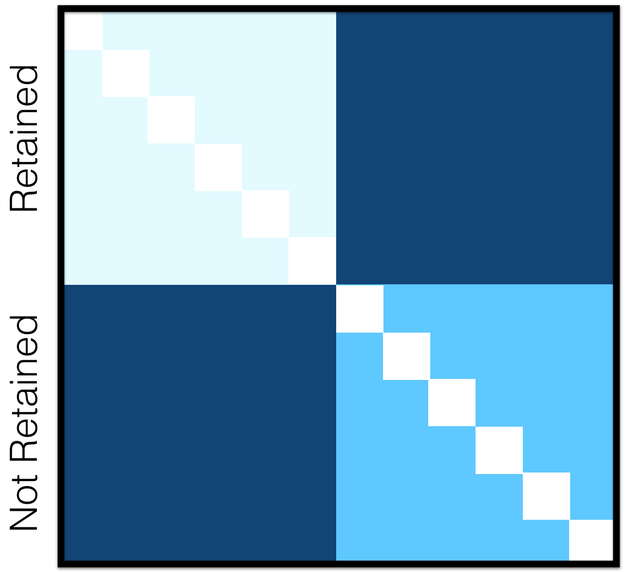
Hopfield model with N=100 neurons and 5 memory patterns

Ordered coupling matrix $|J_{ij}|$





Ordered coupling matrix $|J_{ij}|$



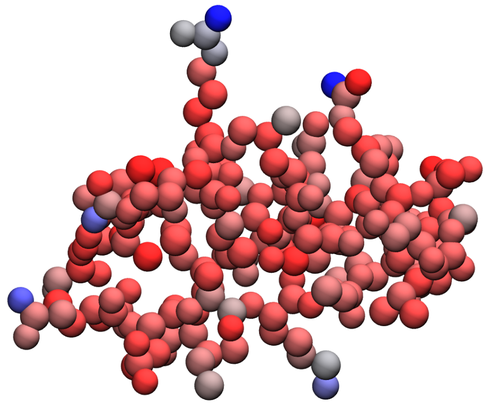
"Phase transition" between two classes of optimal representations of the system

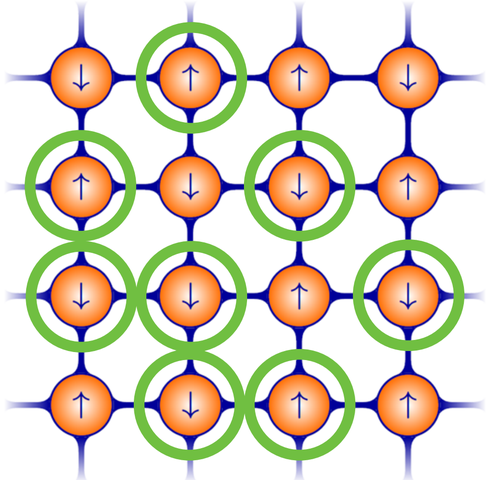
Strongly interacting retained neurons, weakly coupled with the integrated ones

Weakly interacting retained neurons, strongly coupled with the integrated ones

Aldrigo, RM, Potestio (in preparation)

Conclusions





- CG filters as a mean to gain insight on the behavior of complex systems
- CG'ing procedures result in a loss of statistical information
- Maximally-informative CG representations single out biologically relevant residues of a macromolecular system in an unsupervised manner
- Deep Graph networks can be employed to speed up information loss calculations and to achieve a quasi-exhaustive exploration of the mapping space
- Application of the mapping entropy protocol to a neural system: maximally informative neurons in a Hopfield model
- Phase separation of the maximally informative representations of the system depending on the observational level of detail

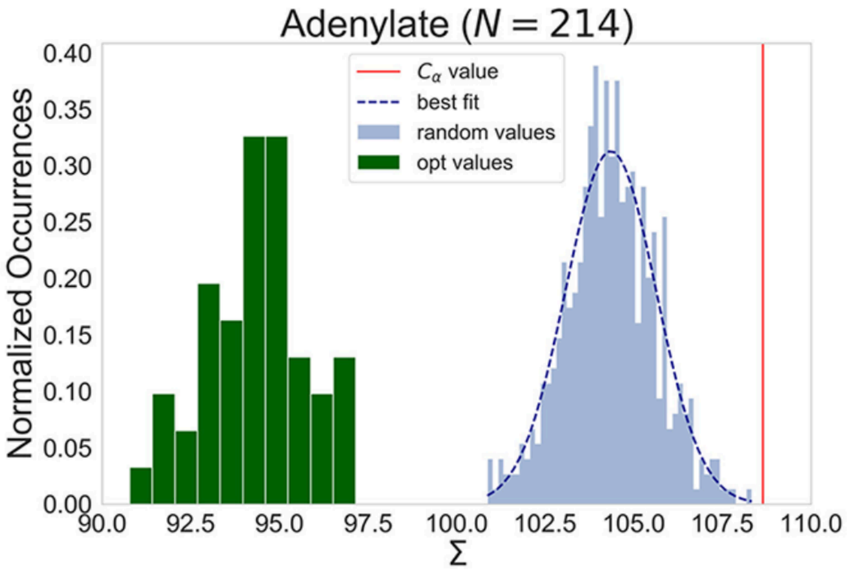
Thank you for the attention!

Acknowledgments

Potestio Lab:
(& former members)

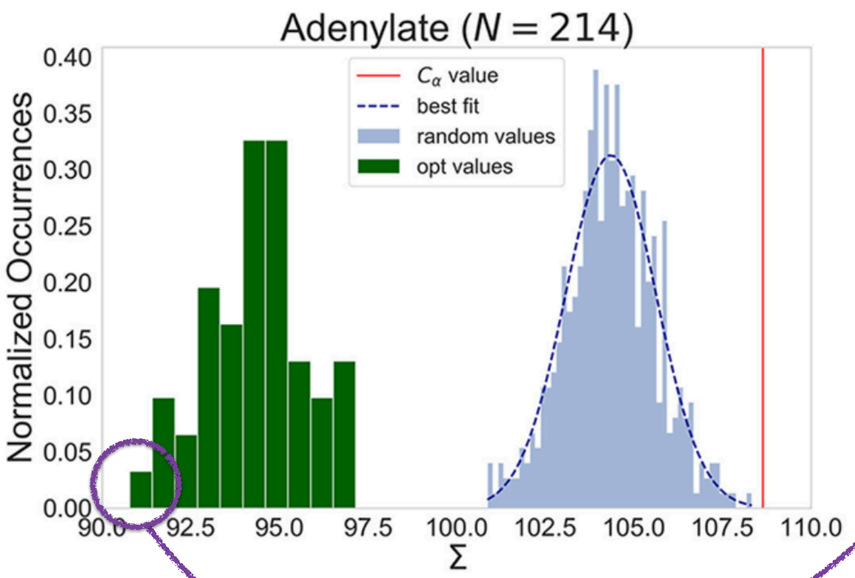
Raffaello Potestio
Thomas Tarenzi, now at UoB (UK)
Giovanni Mattiotti
Marco Giulini, now at UU (NL)
Roberto Menichetti
Elio Fiorentini, now at AM (IT)
Marta Rigoli, now at CIBIO (IT)
Margherita Mele
Lorenzo Petrolli
Manuel Micheloni
Camilla Spreti

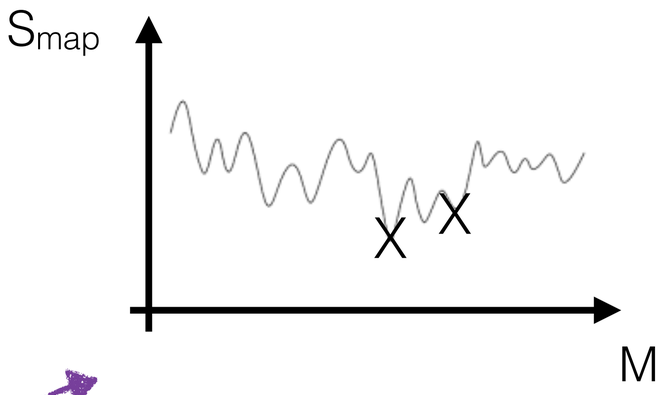
Characterizing the space of solutions



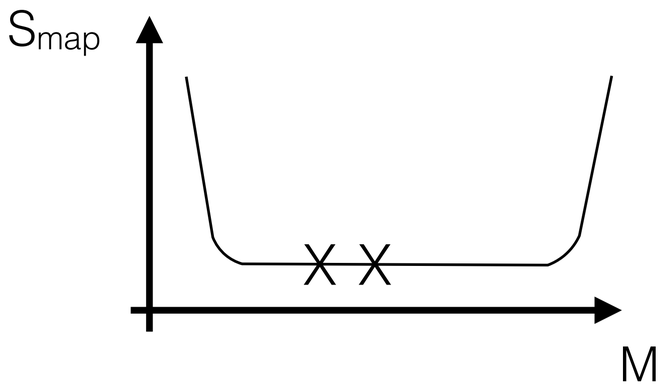
Giulini, RM, Shell, Potestio, JCTC 2020

Characterizing the space of solutions



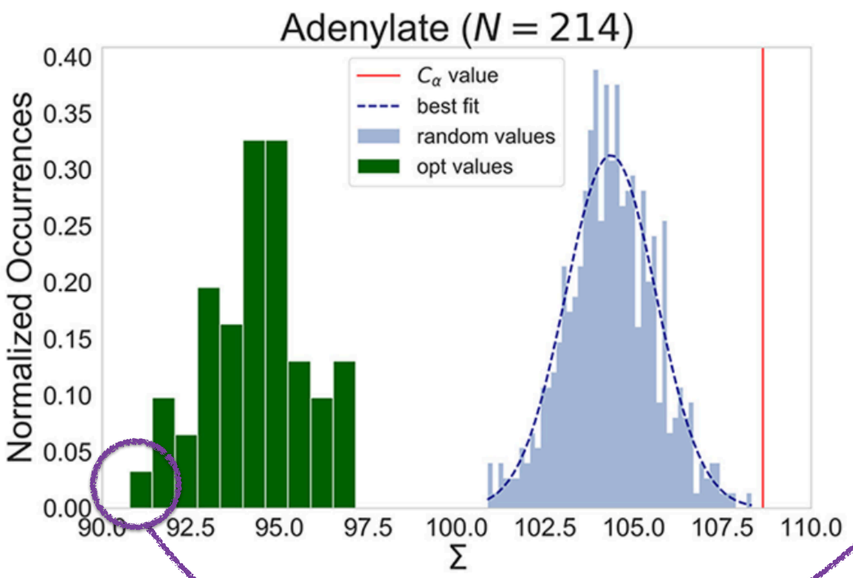


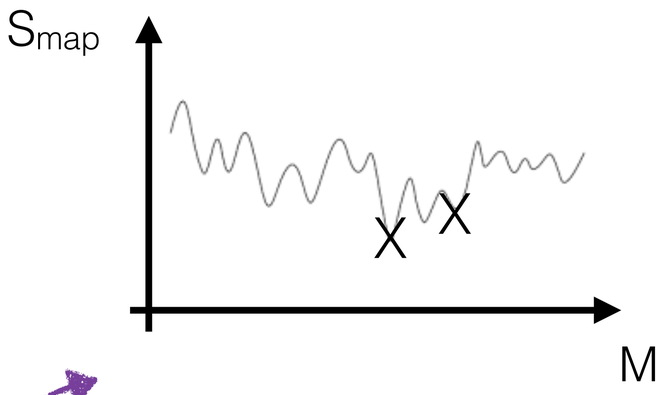
VS



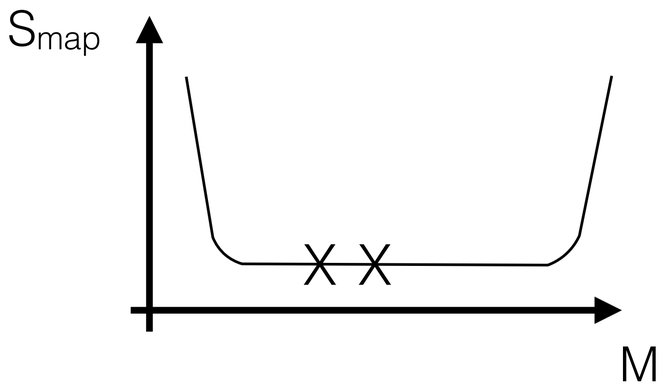
Giulini, RM, Shell, Potestio, JCTC 2020

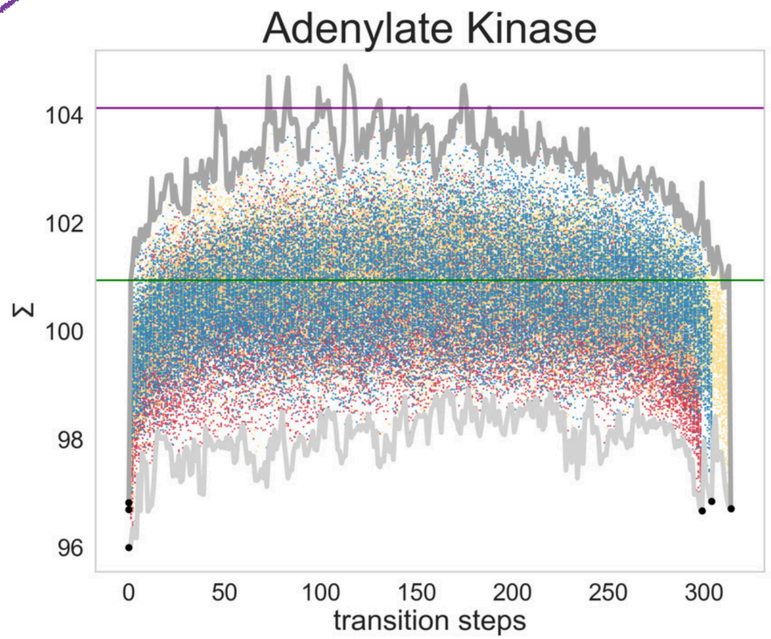
Characterizing the space of solutions

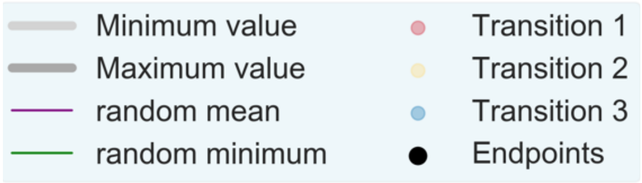




VS

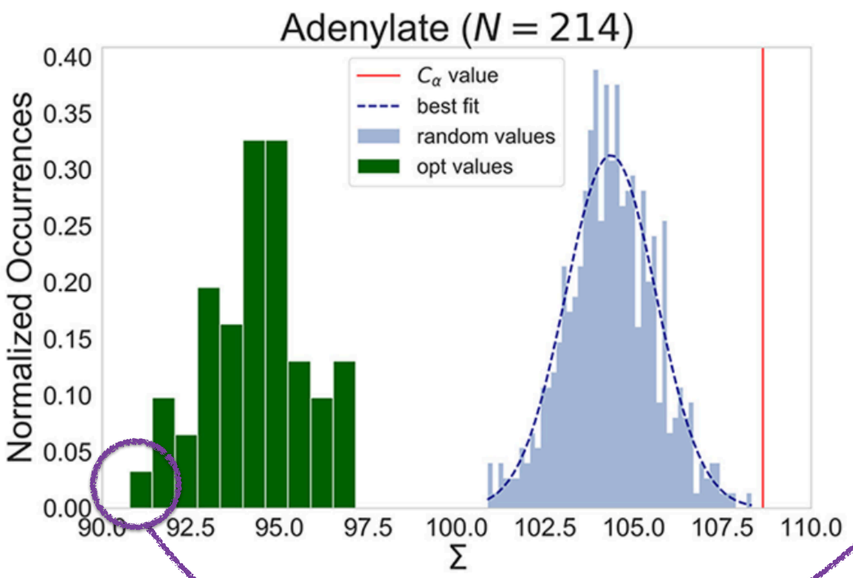


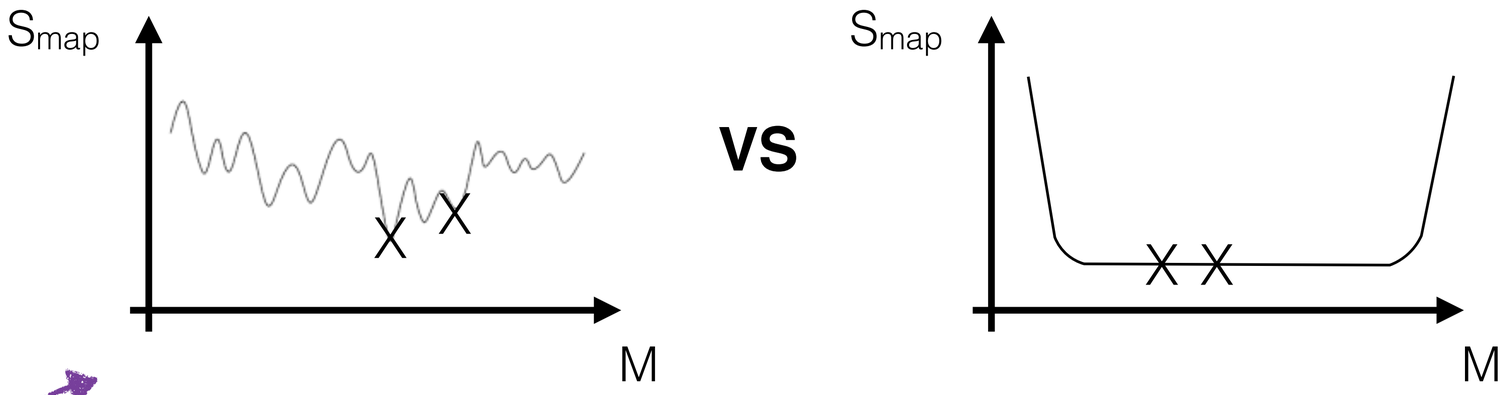


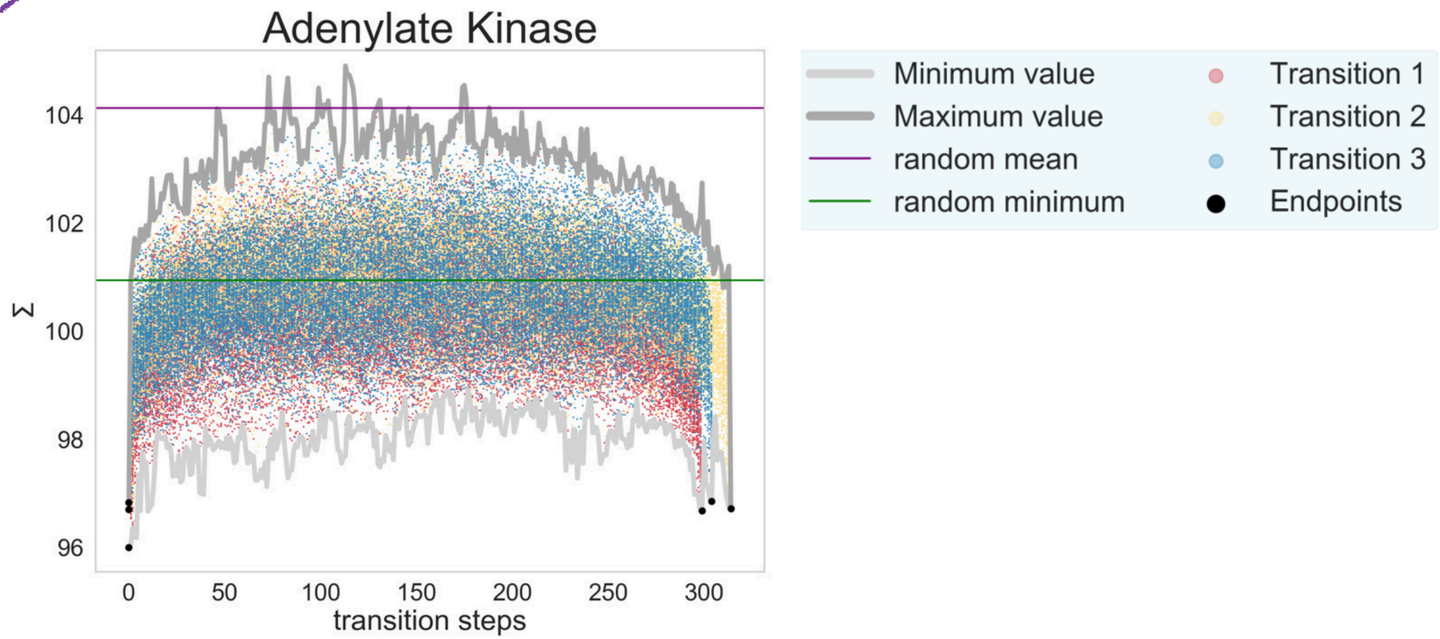


Giulini, RM, Shell, Potestio, JCTC 2020

Characterizing the space of solutions





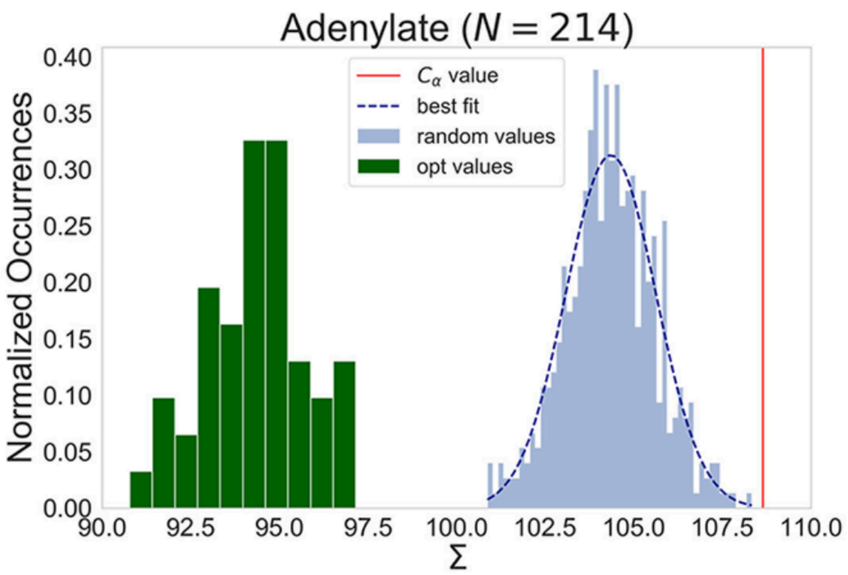


Optimized solutions are isolated points in the mapping entropy landscape

Atoms conserved with high probability are likely present in many uncorrelated solutions

Giulini, RM, Shell, Potestio, JCTC 2020

Characterizing the space of solutions



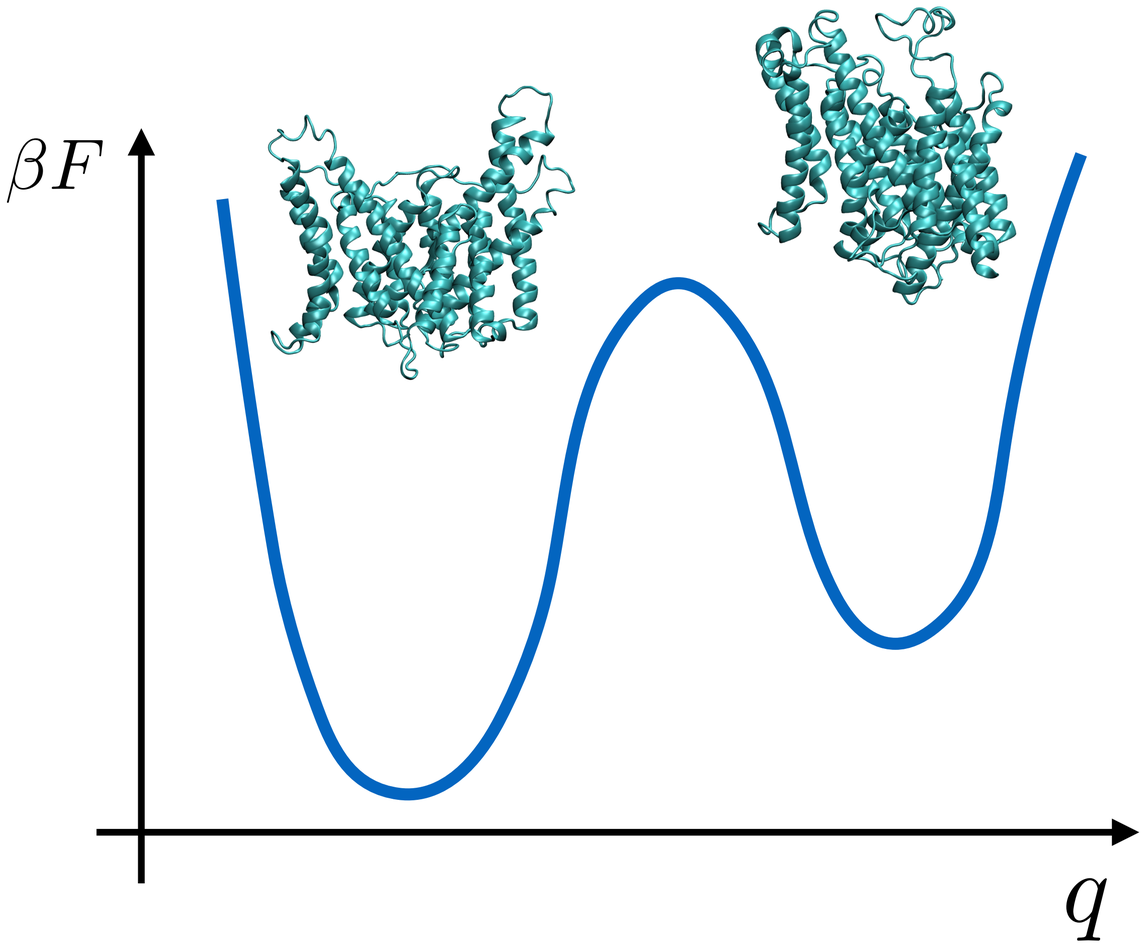
Optimized solutions are isolated points in the mapping entropy landscape

Atoms conserved with high probability are likely present in many uncorrelated solutions

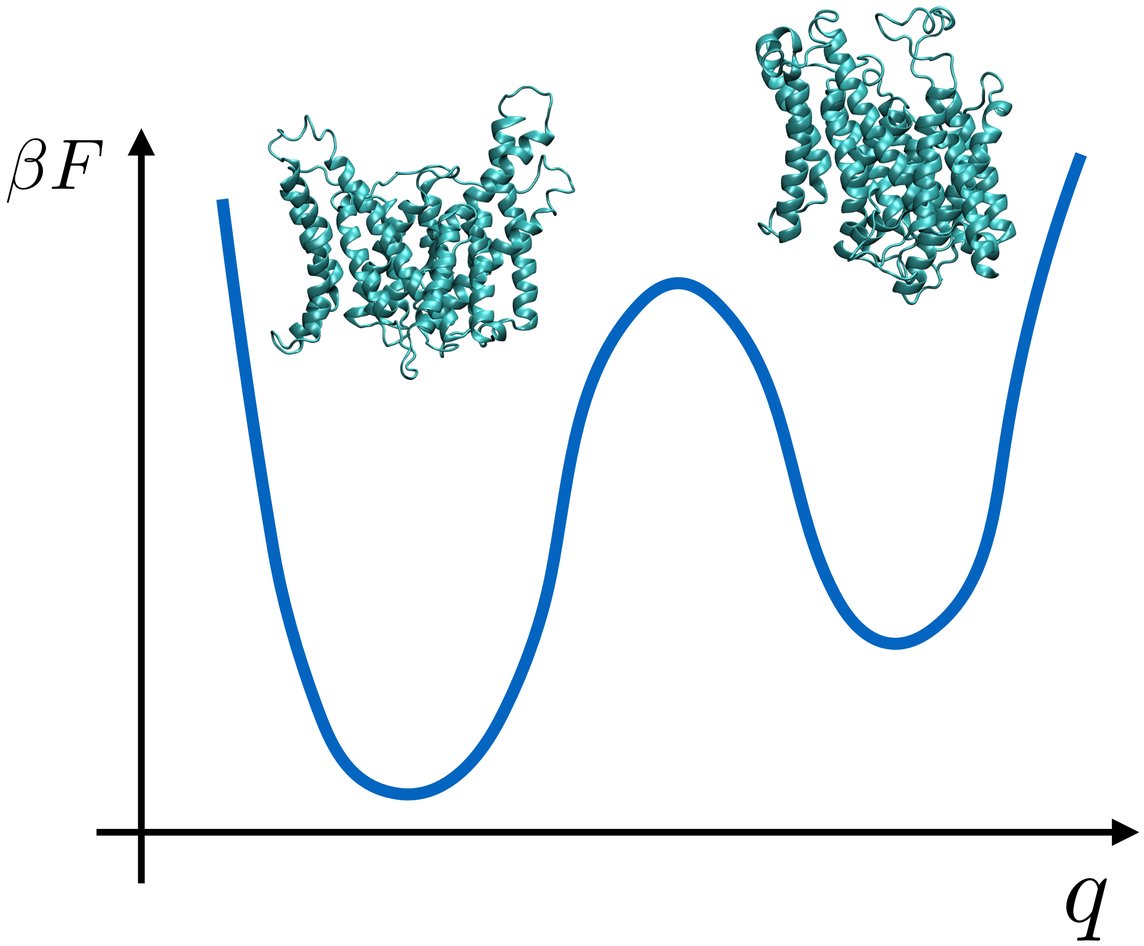
Giulini, RM, Shell, Potestio, JCTC 2020

State-dependence of optimal mappings

State-dependence of optimal mappings

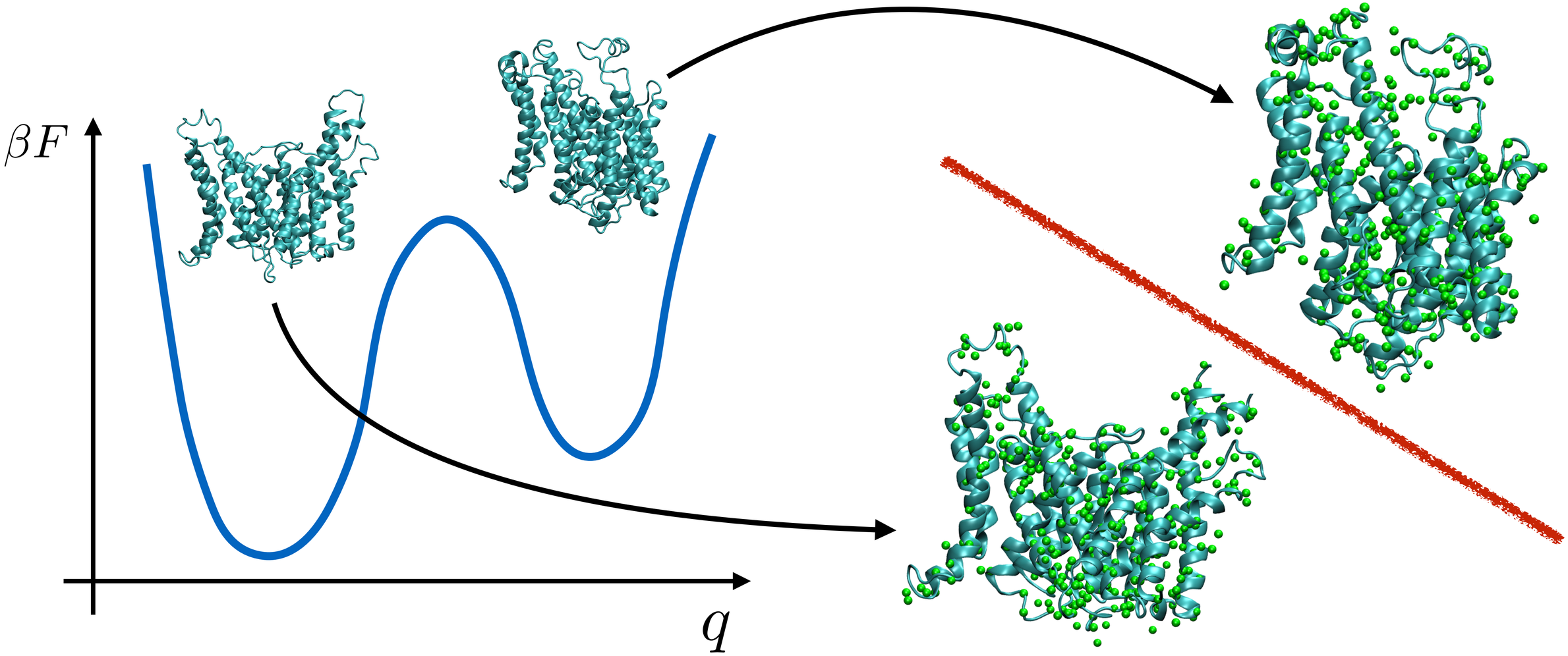


State-dependence of optimal mappings

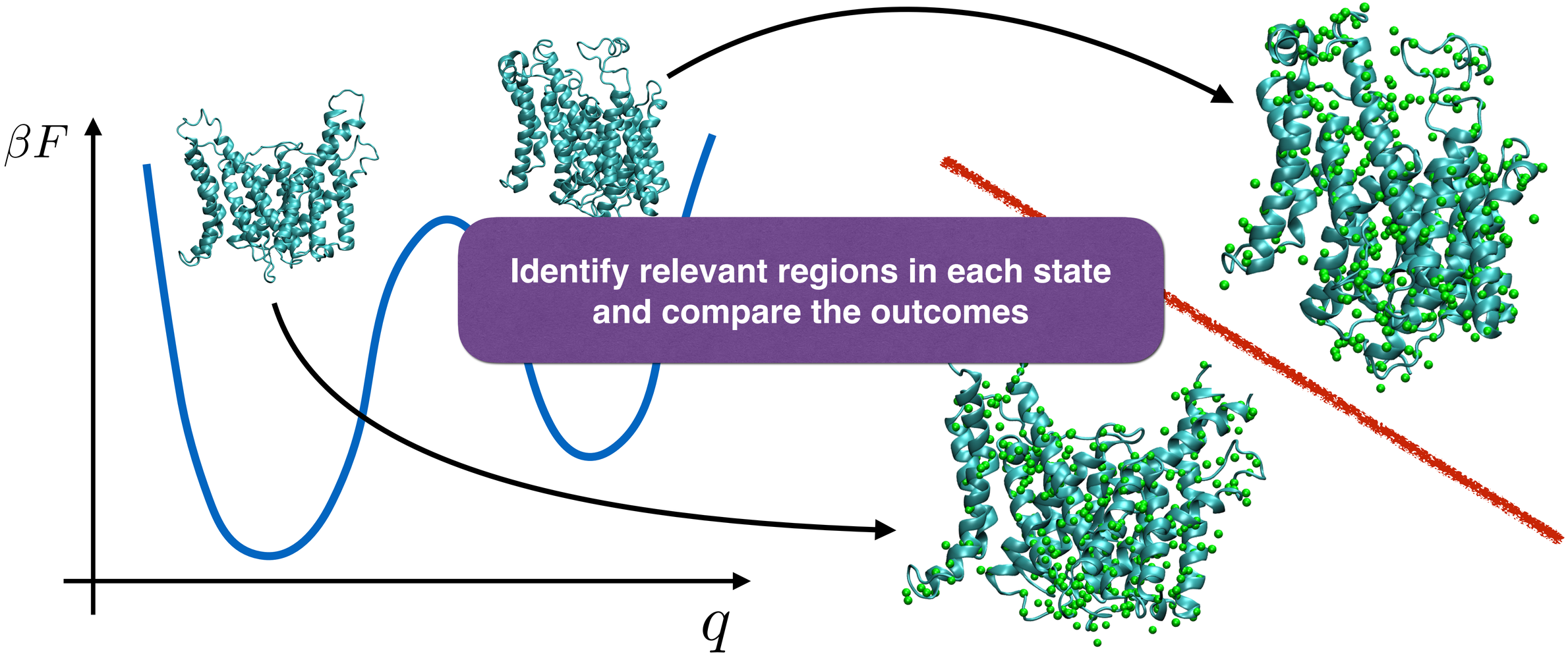


Search for mappings that minimize the information loss within each basin

State-dependence of optimal mappings



State-dependence of optimal mappings



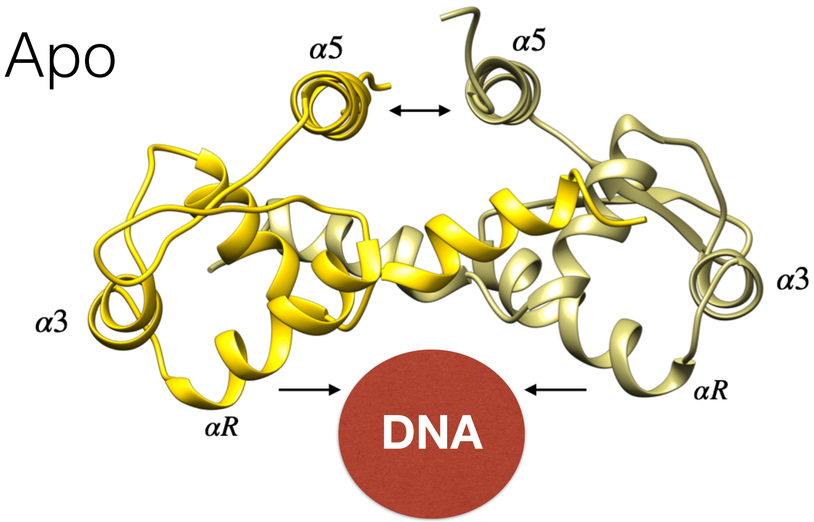
Identify relevant regions in each state and compare the outcomes

Application: CzrA transcription repressor

Rigoli, Giulini, RM, Potestio, in preparation

Application: CzrA transcription repressor

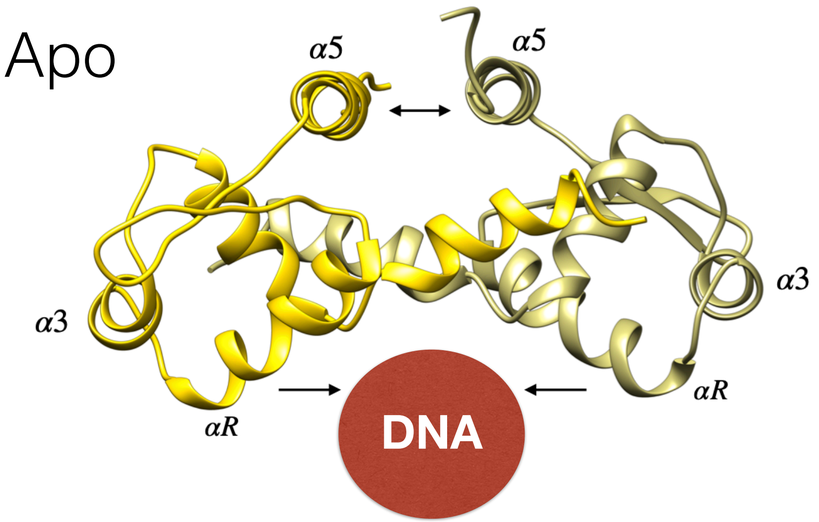
Apo



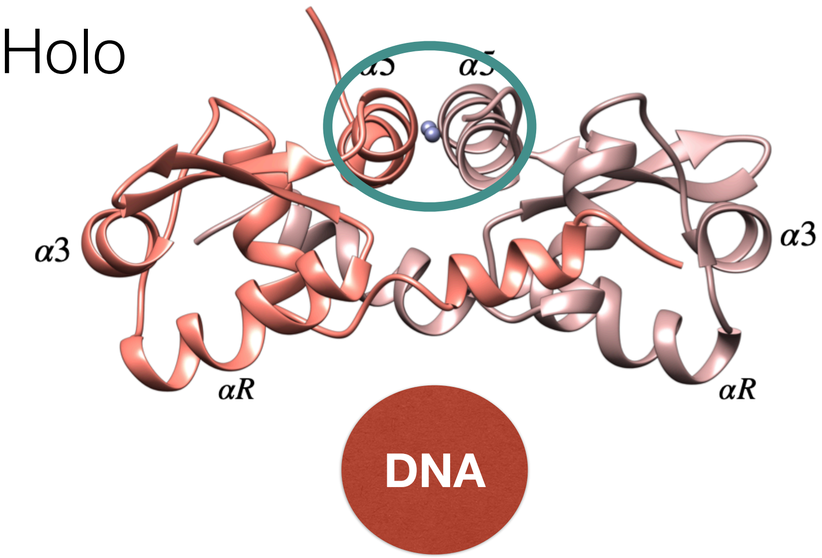
Rigoli, Giulini, RM, Potestio, in preparation

Application: CzrA transcription repressor

Apo



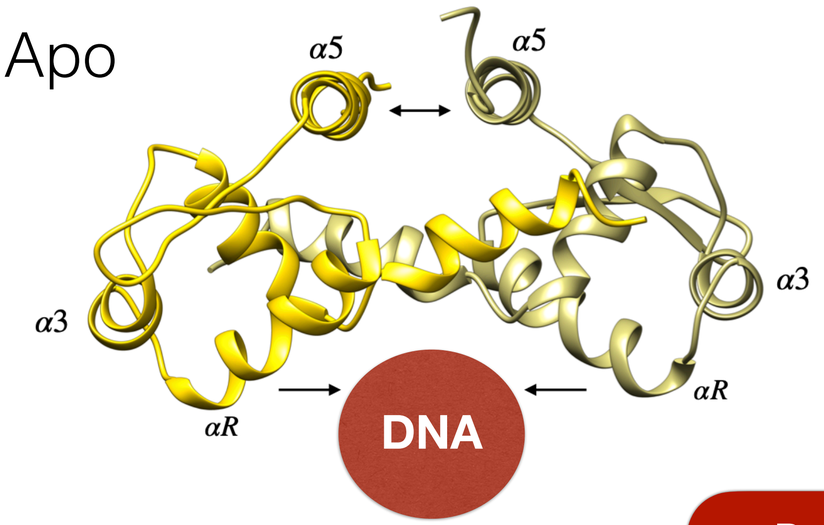
Holo



Rigoli, Giulini, RM, Potestio, in preparation

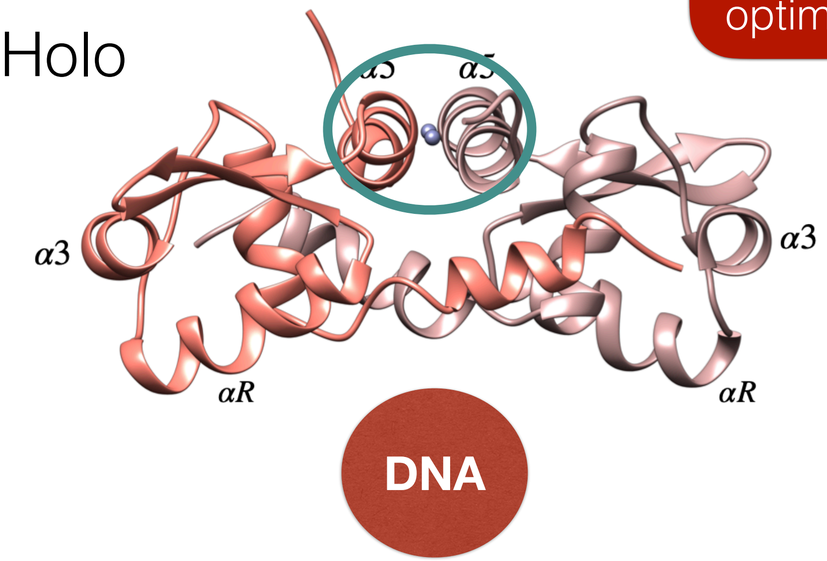
Application: CzrA transcription repressor

Apo



Perform independent optimizations in each state

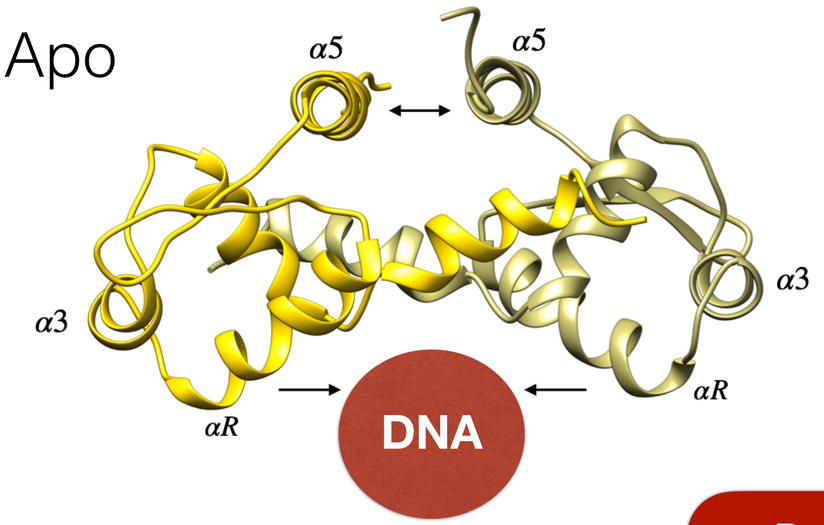
Holo



Rigoli, Giulini, RM, Potestio, in preparation

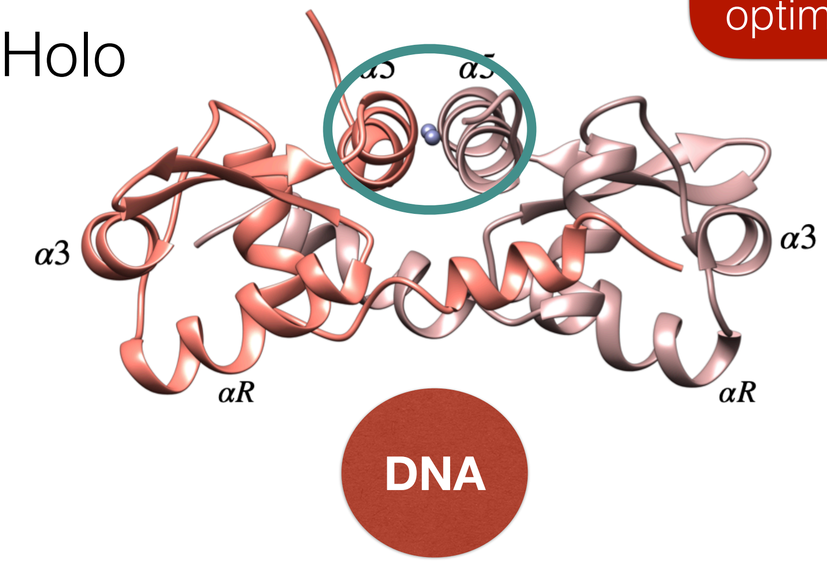
Application: CzrA transcription repressor

Apo



Perform independent optimizations in each state

Holo

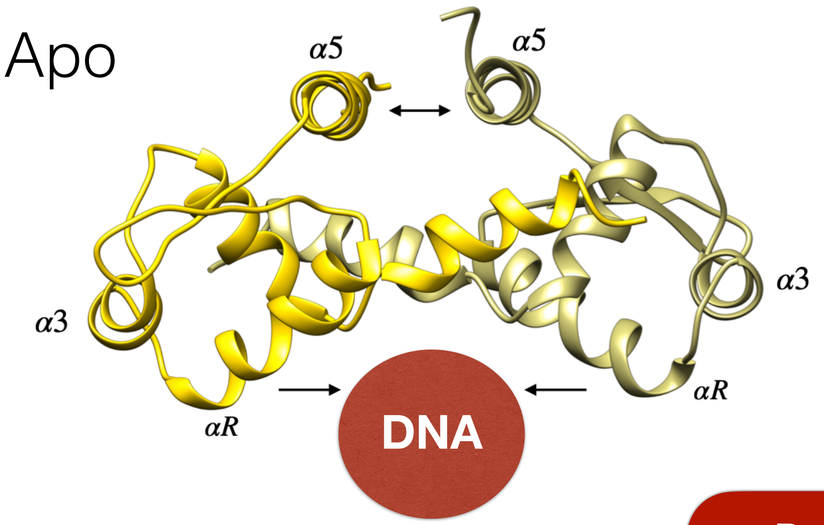


Biologically relevant residues slightly "deactivate" between Apo and Holo

Rigoli, Giulini, RM, Potestio, in preparation

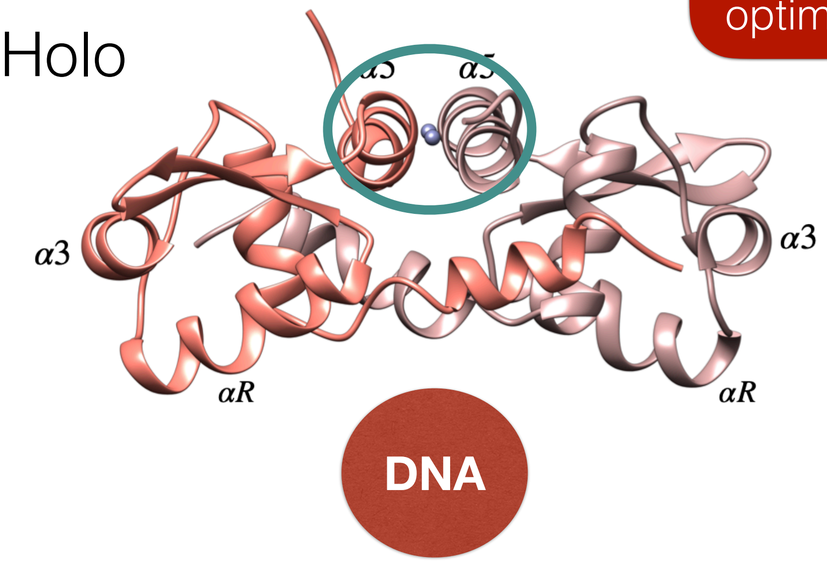
Application: CzrA transcription repressor

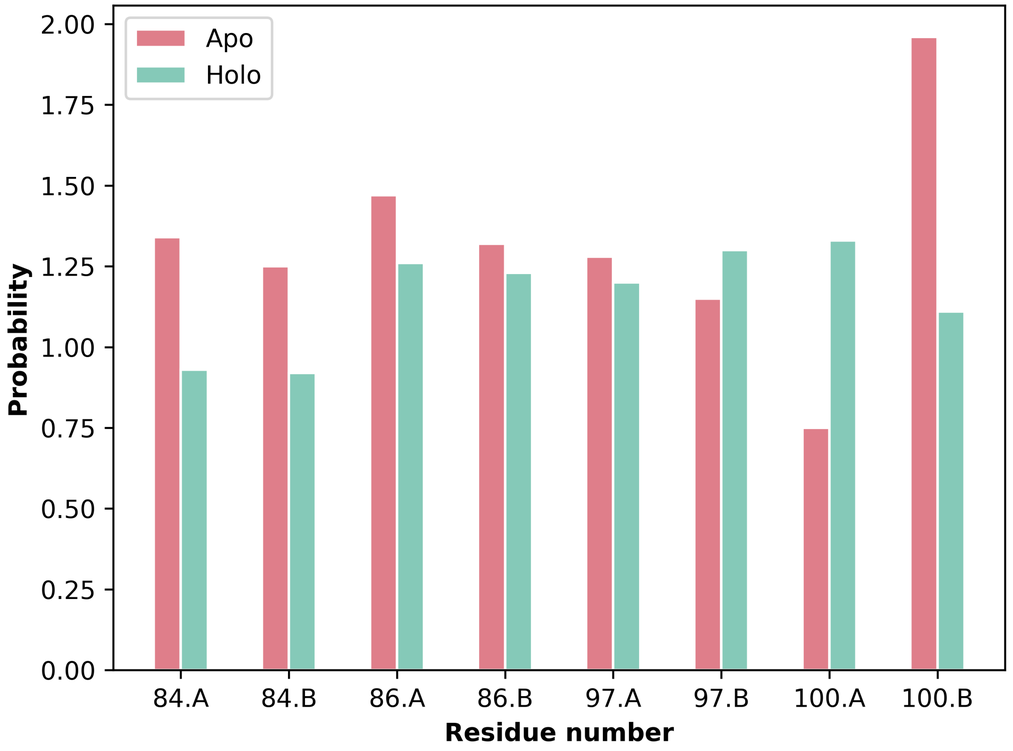
Apo



Perform independent optimizations in each state

Holo

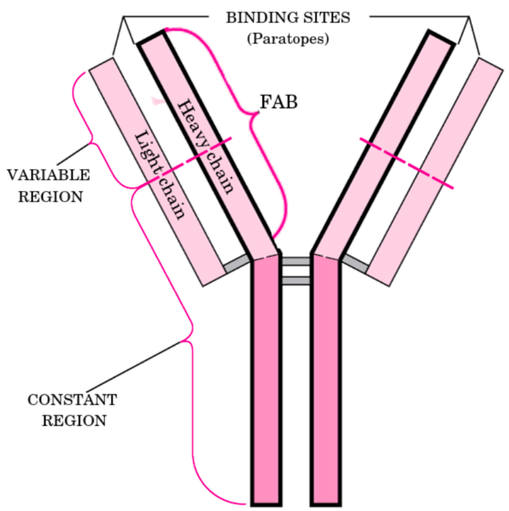




Biologically relevant residues slightly "deactivate" between Apo and Holo

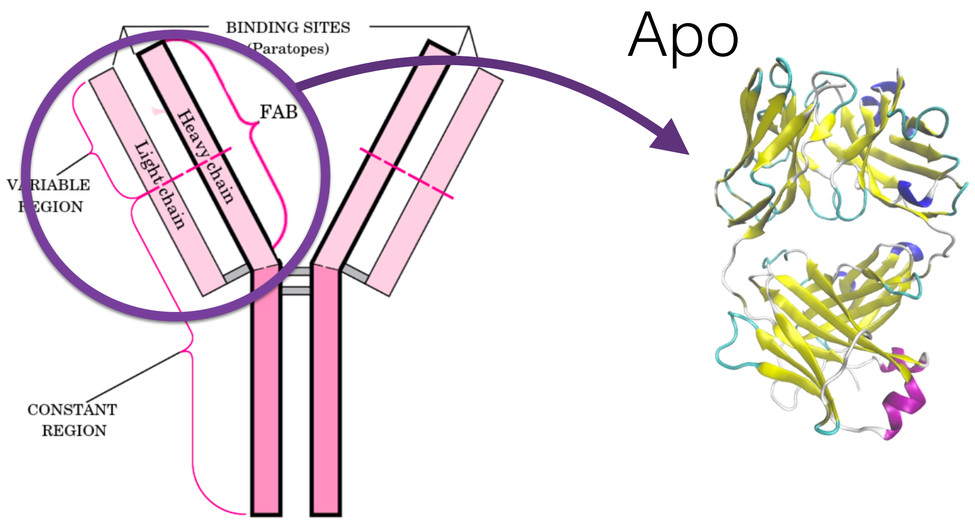
Rigoli, Giulini, RM, Potestio, in preparation

Application: G6 Antibody and VEGF-A antigen



Galante, Tarenzi, Potestio, preliminary results

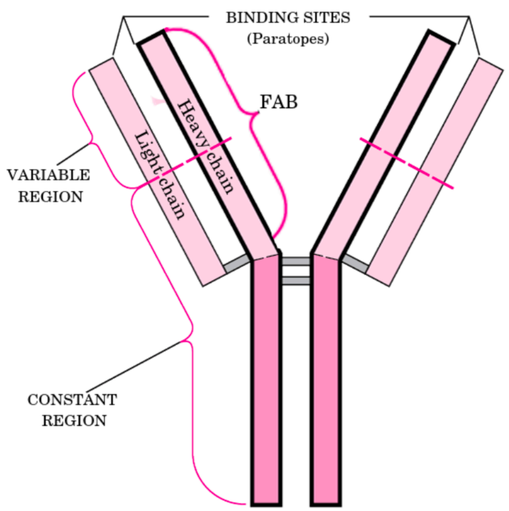
Application: G6 Antibody and VEGF-A antigen



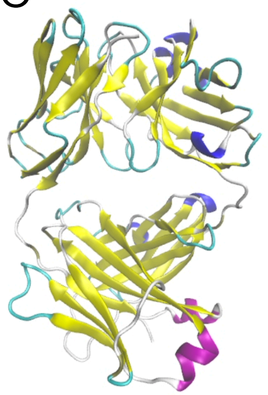
Apo

Galante, Tarenzi, Potestio, preliminary results

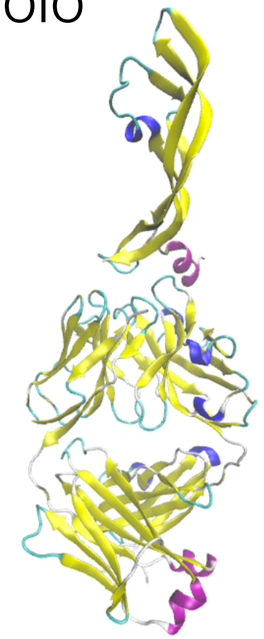
Application: G6 Antibody and VEGF-A antigen



Apo

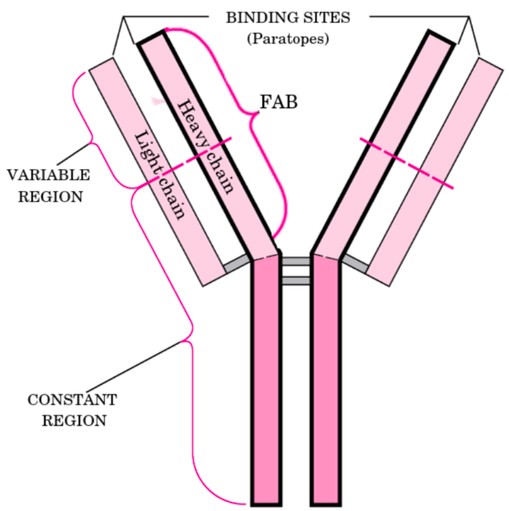


Holo

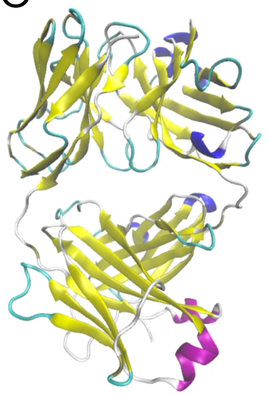


Galante, Tarenzi, Potestio, preliminary results

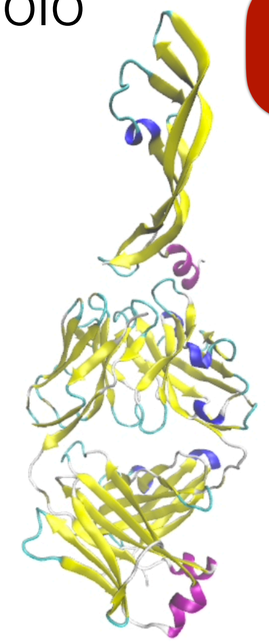
Application: G6 Antibody and VEGF-A antigen



Apo



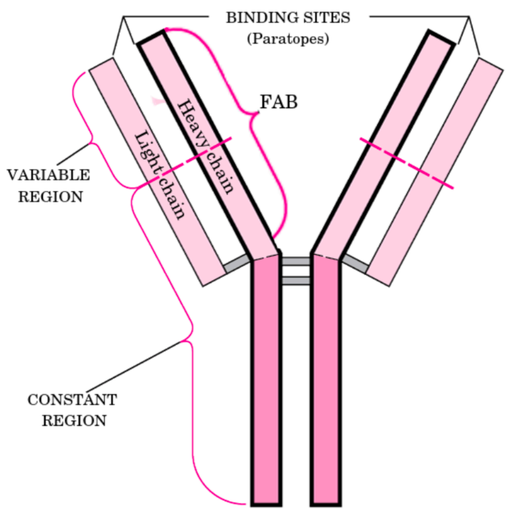
Holo



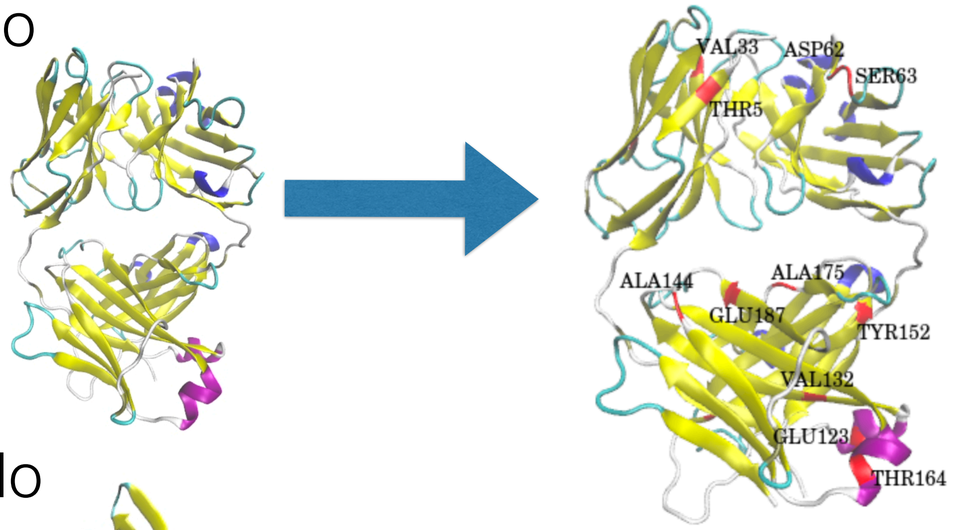
Perform independent optimizations in each state

Galante, Tarenzi, Potestio, preliminary results

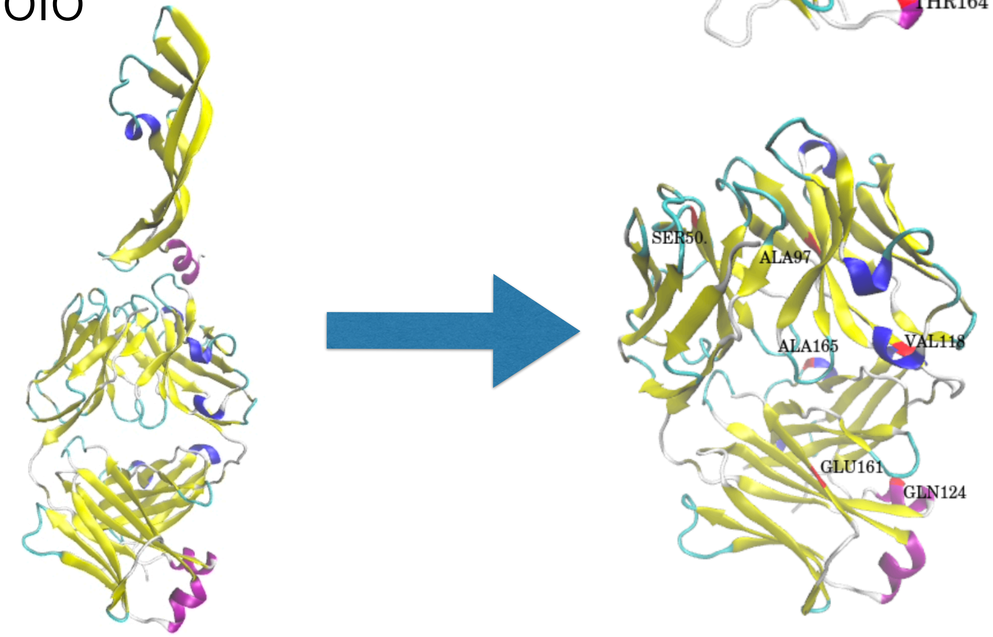
Application: G6 Antibody and VEGF-A antigen



Apo

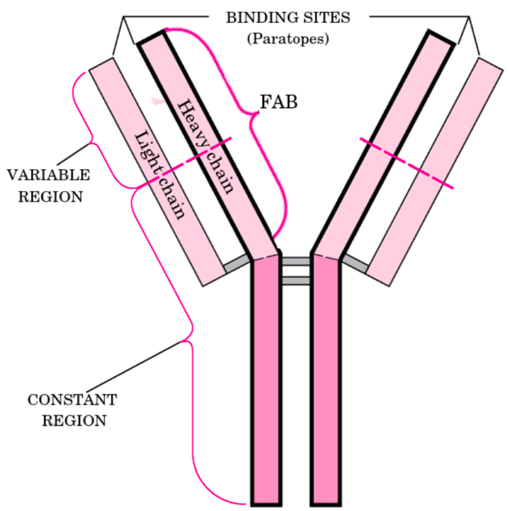


Holo

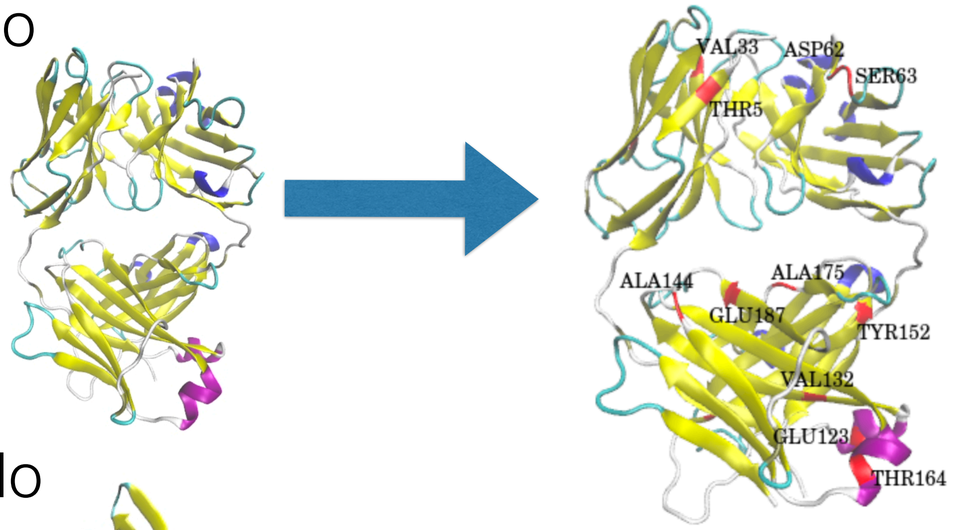


Galante, Tarenzi, Potestio, preliminary results

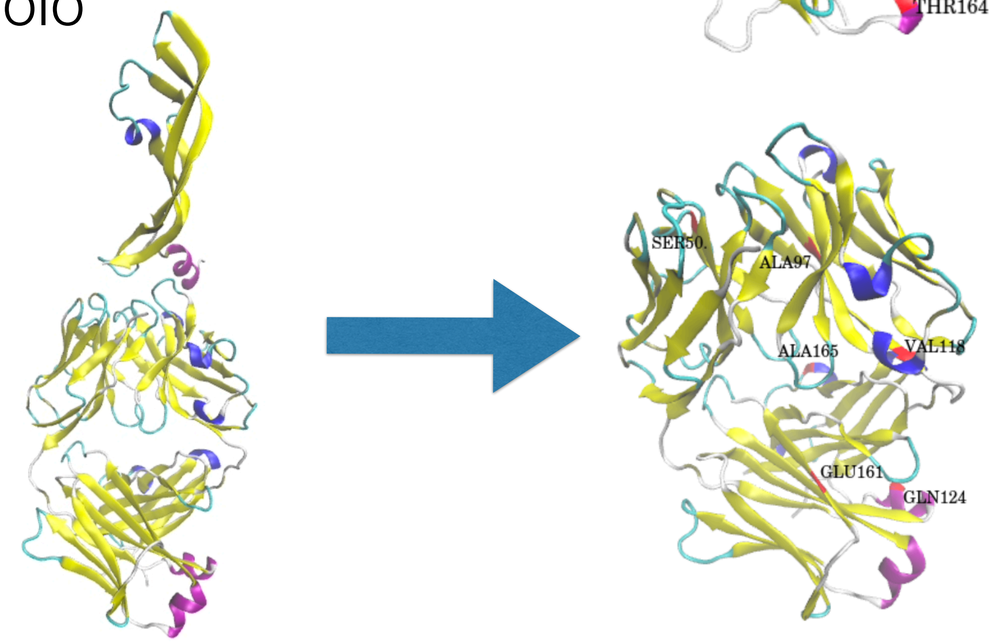
Application: G6 Antibody and VEGF-A antigen



Apo

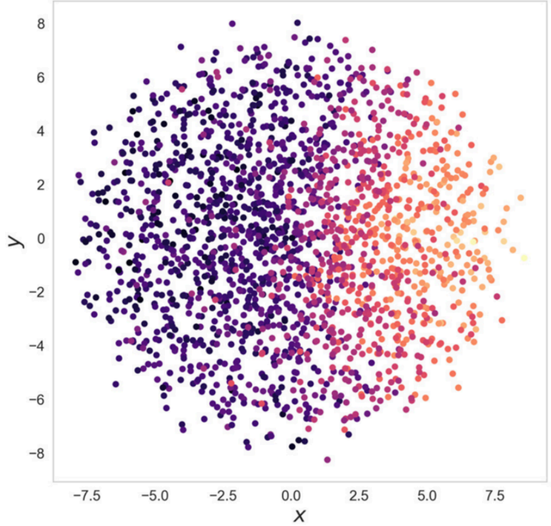


Holo



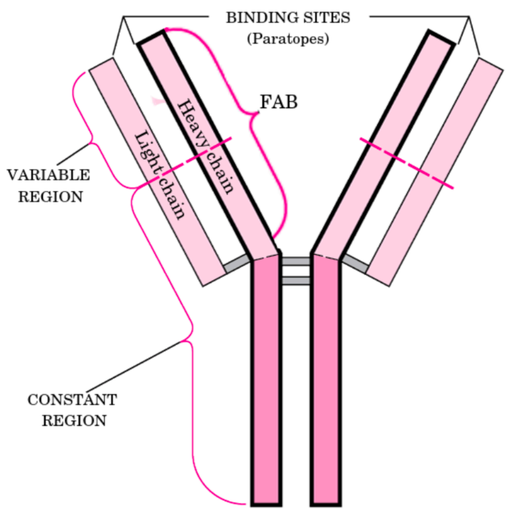
The curse of the space dimensionality

An exhaustive sampling of all mappings is unfeasible

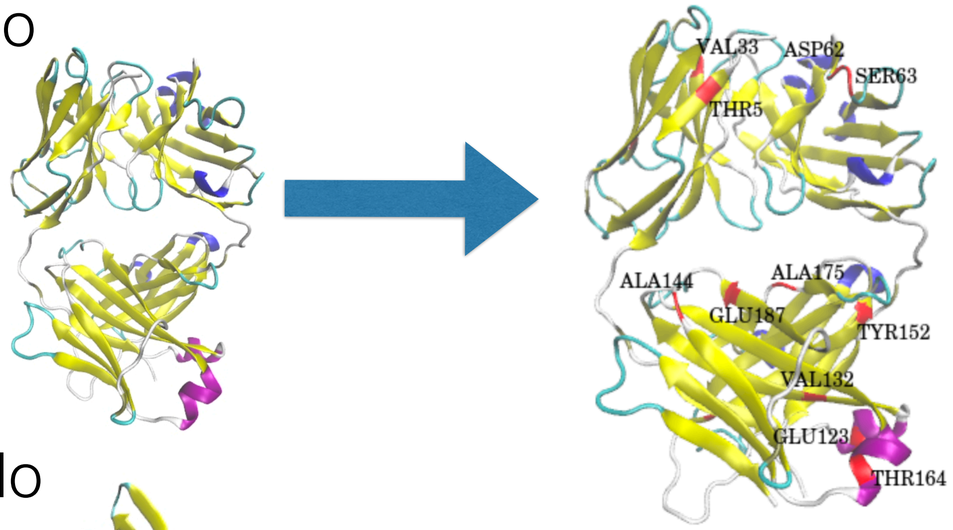


Galante, Tarenzi, Potestio, preliminary results

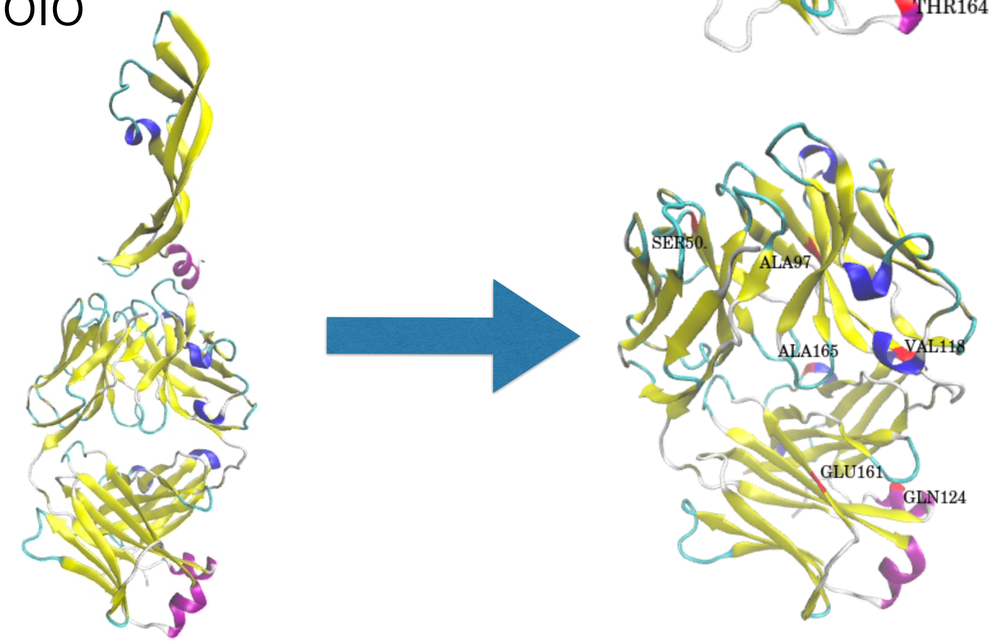
Application: G6 Antibody and VEGF-A antigen



Apo

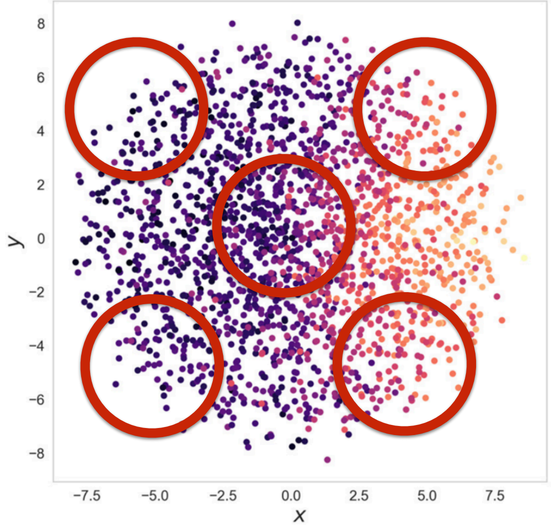


Holo



The curse of the space dimensionality

An exhaustive sampling of all mappings is unfeasible



Explore according to the distance

$$\mathcal{D}(M, M')$$

Galante, Tarenzi, Potestio, preliminary results

S_{map} via deep graph networks

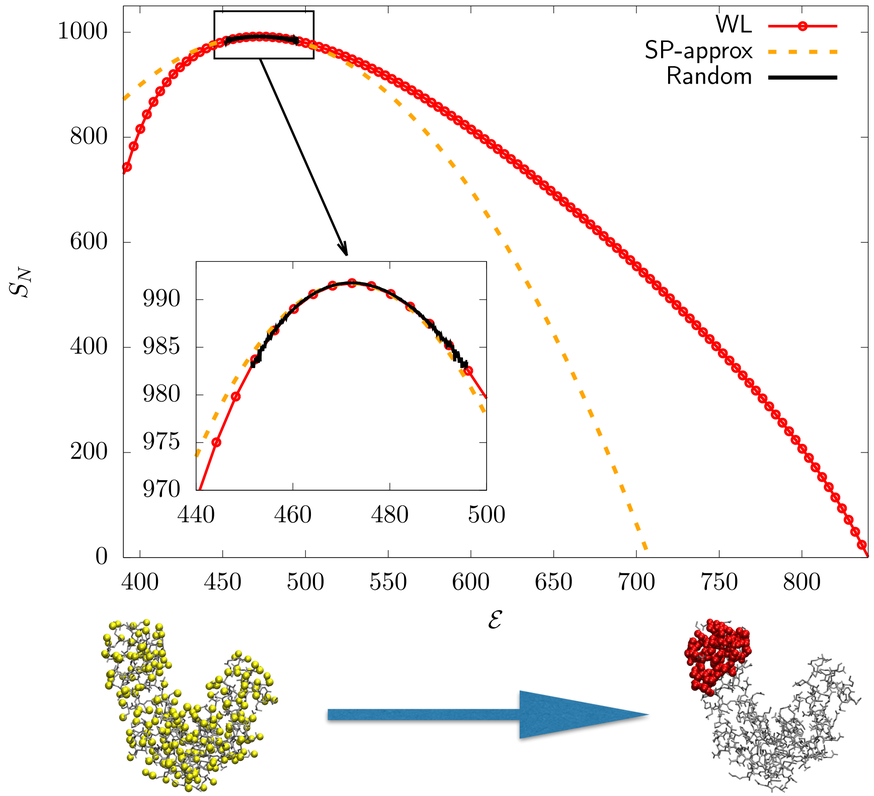
Speedup generated by DGNs: quasi-exhaustive exploration of the S_{map} landscape

RM, Giulini, Potestio, EPJB 2021
Errica, Giulini, RM *et al*., Front. Mol. Biosci. 2021

S_{map} via deep graph networks

Speedup generated by DGNs: quasi-exhaustive exploration of the S_{map} landscape

$$\#\text{selections}(\mathcal{E}) \propto e^{S_N(\mathcal{E})}$$

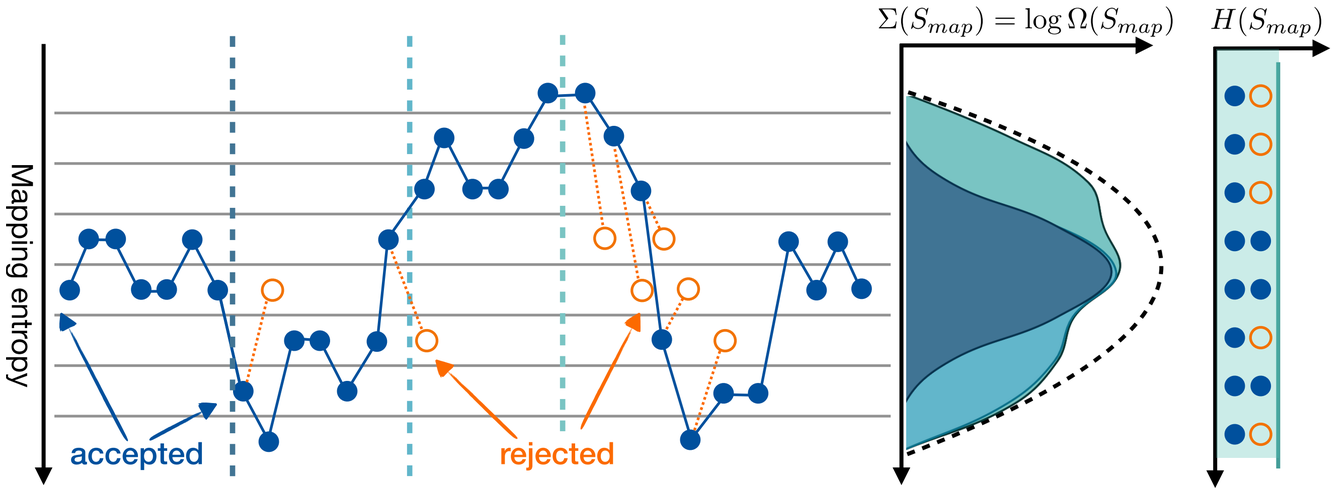


RM, Giulini, Potestio, EPJB 2021
Errica, Giulini, RM *et al.*, Front. Mol. Biosci. 2021

S_{map} via deep graph networks

Speedup generated by DGNs: quasi-exhaustive exploration of the S_{map} landscape

Combine the trained DGN with the Wang-Landau sampling algorithm **to reconstruct the density of states of S_{map}**

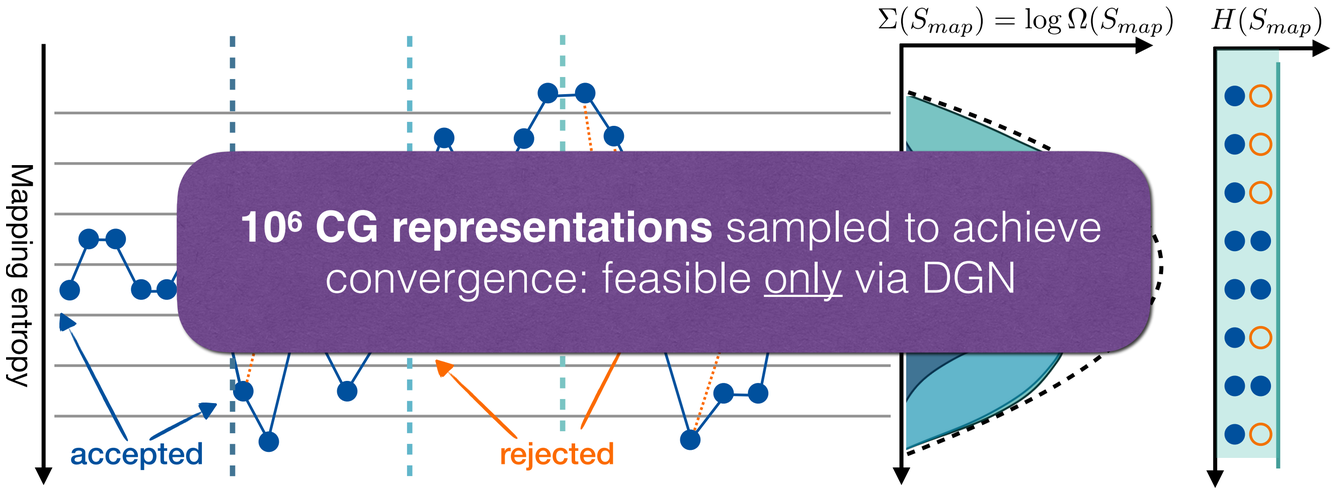


RM, Giulini, Potestio, EPJB 2021
Errica, Giulini, RM *et al.*, Front. Mol. Biosci. 2021

S_{map} via deep graph networks

Speedup generated by DGNs: quasi-exhaustive exploration of the S_{map} landscape

Combine the trained DGN with the Wang-Landau sampling algorithm **to reconstruct the density of states of S_{map}**



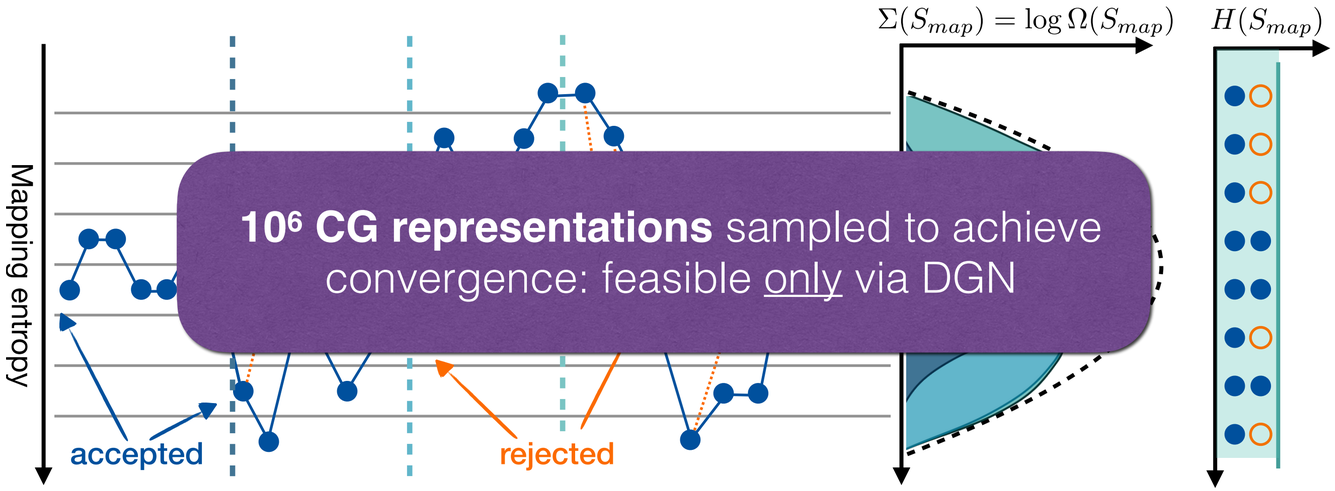
10^6 CG representations sampled to achieve convergence: feasible only via DGN

RM, Giulini, Potestio, EPJB 2021
Errica, Giulini, RM *et al.*, Front. Mol. Biosci. 2021

S_{map} via deep graph networks

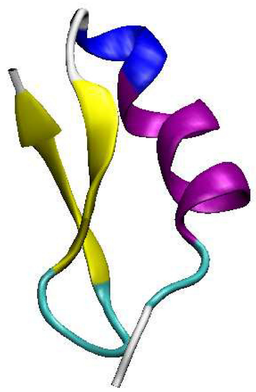
Speedup generated by DGNs: quasi-exhaustive exploration of the S_{map} landscape

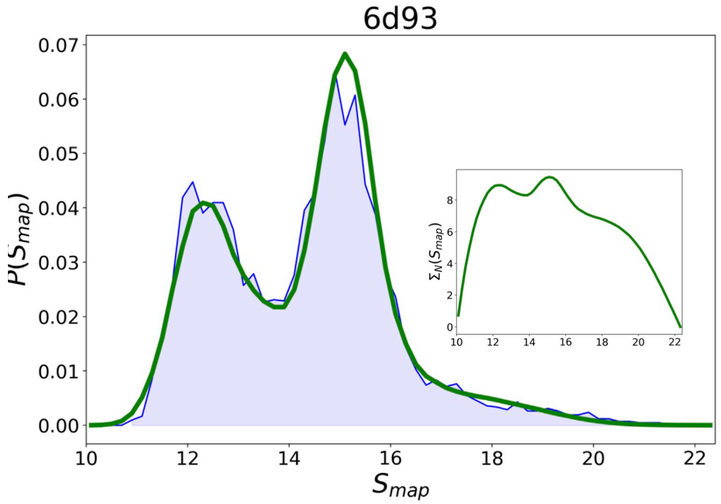
Combine the trained DGN with the Wang-Landau sampling algorithm **to reconstruct the density of states of S_{map}**



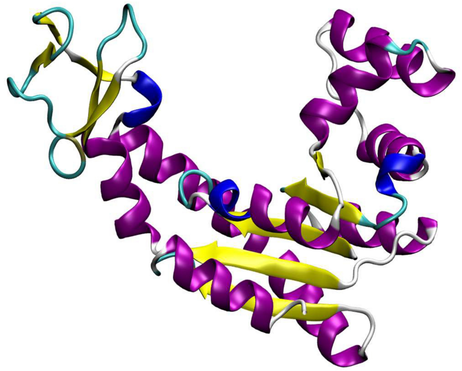
10^6 CG representations sampled to achieve convergence: feasible only via DGN

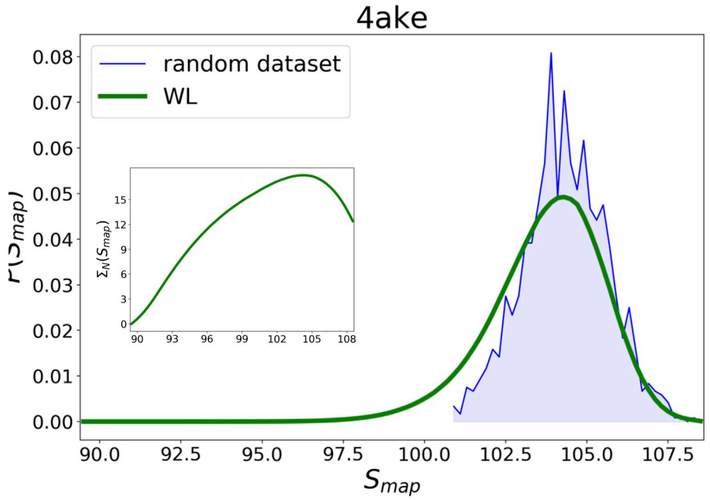
Tamapin





Adenylate Kinase



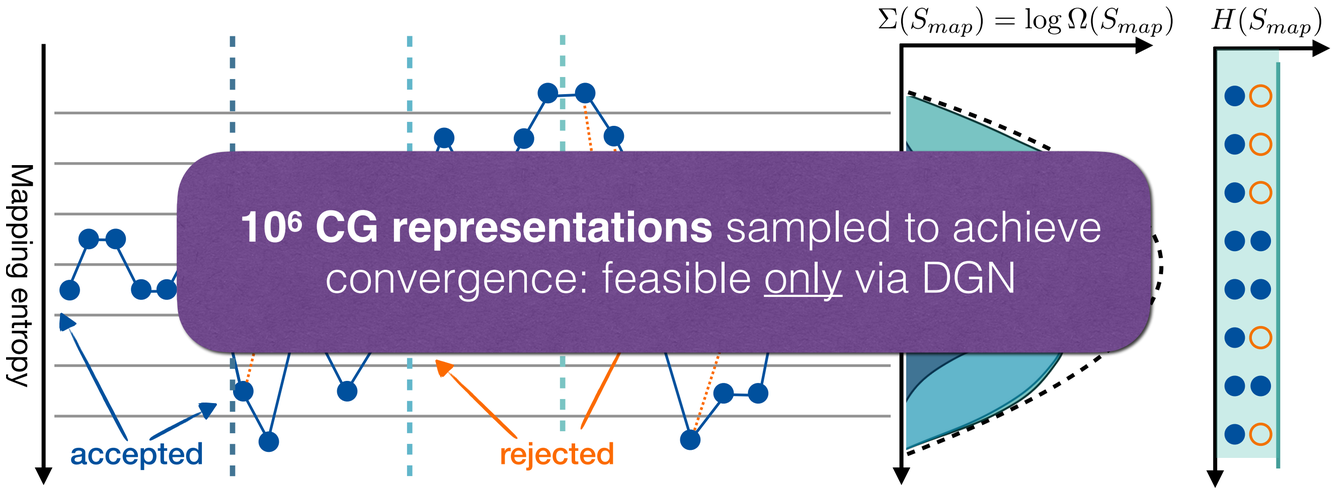


RM, Giulini, Potestio, EPJB 2021
Errica, Giulini, RM *et al.*, Front. Mol. Biosci. 2021

S_{map} via deep graph networks

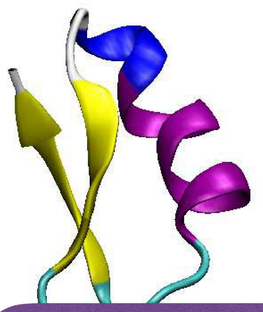
Speedup generated by DGNs: quasi-exhaustive exploration of the S_{map} landscape

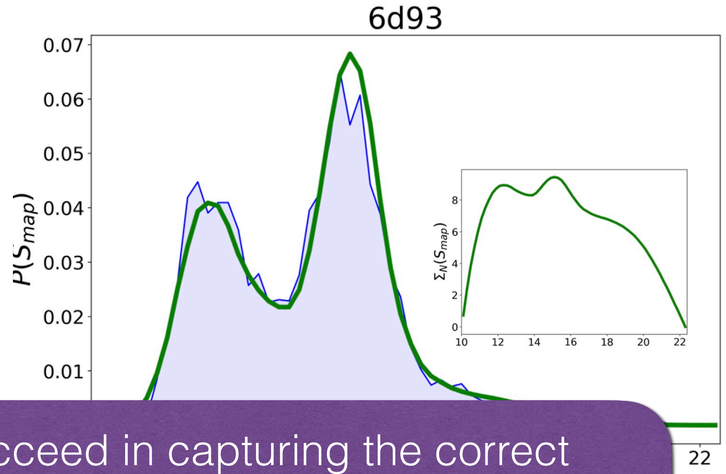
Combine the trained DGN with the Wang-Landau sampling algorithm **to reconstruct the density of states of S_{map}**



10^6 CG representations sampled to achieve convergence: feasible only via DGN

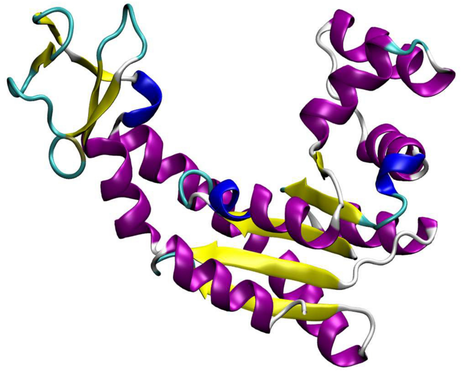
Tamapin

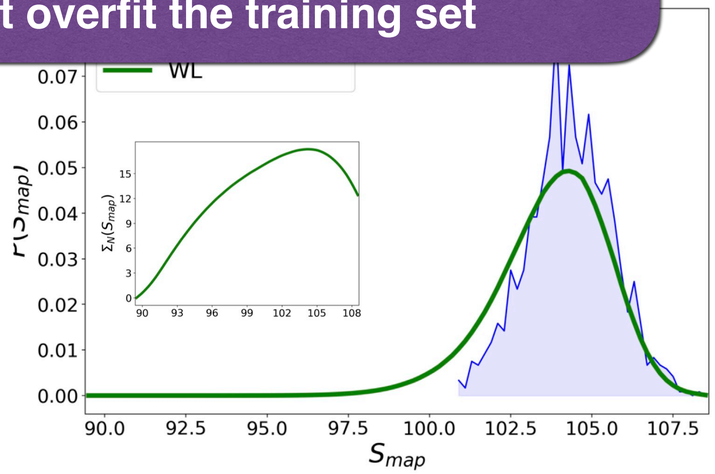




DGNs succeed in capturing the correct statistical weight of CG representations: **do not overfit the training set**

Ader



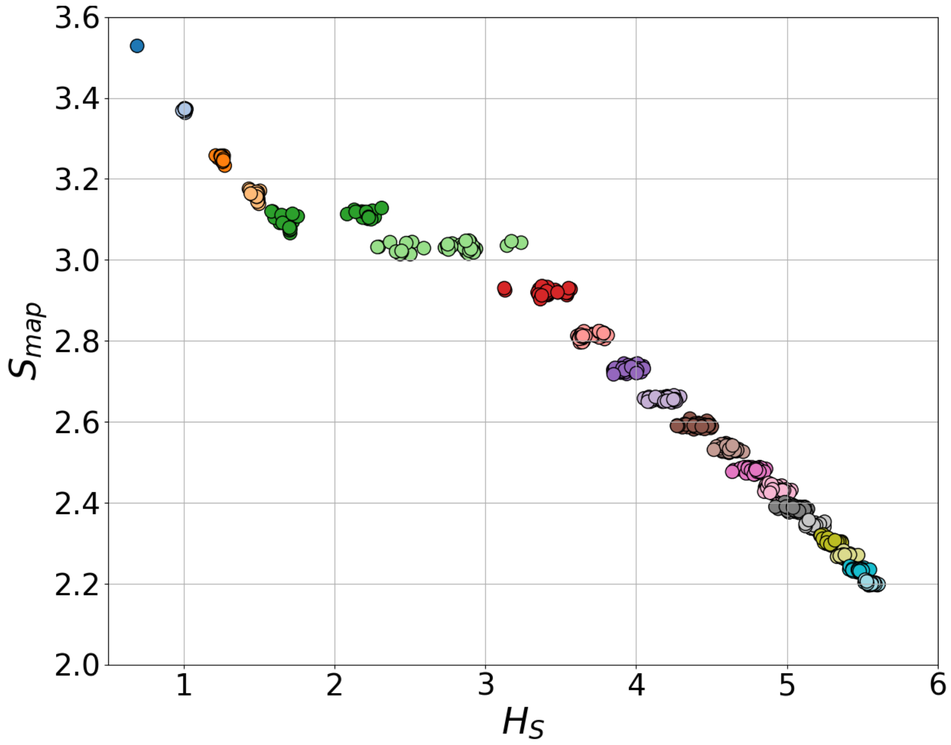


RM, Giulini, Potestio, EPJB 2021

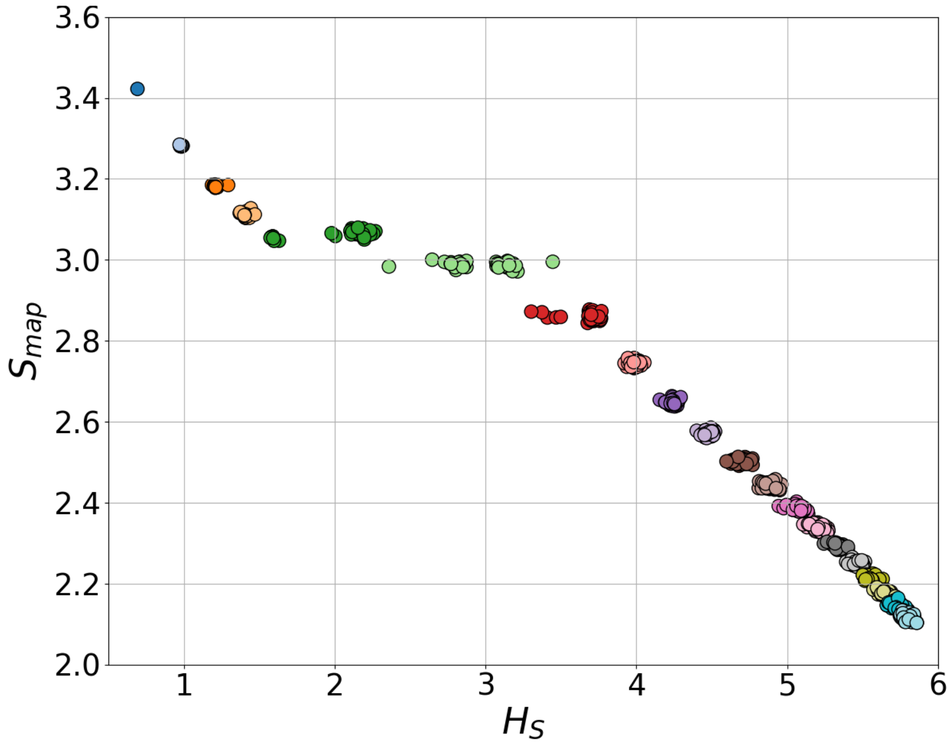
Errica, Giulini, RM *et al.*, Front. Mol. Biosci. 2021

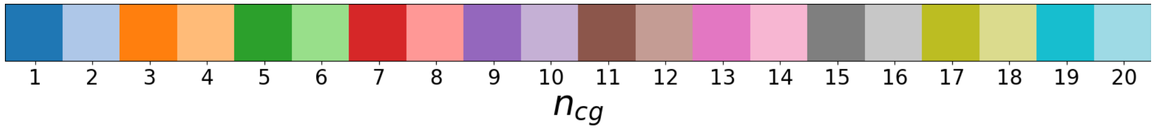
Moving towards the SG phase

N=100 neurons, p= 4 patterns

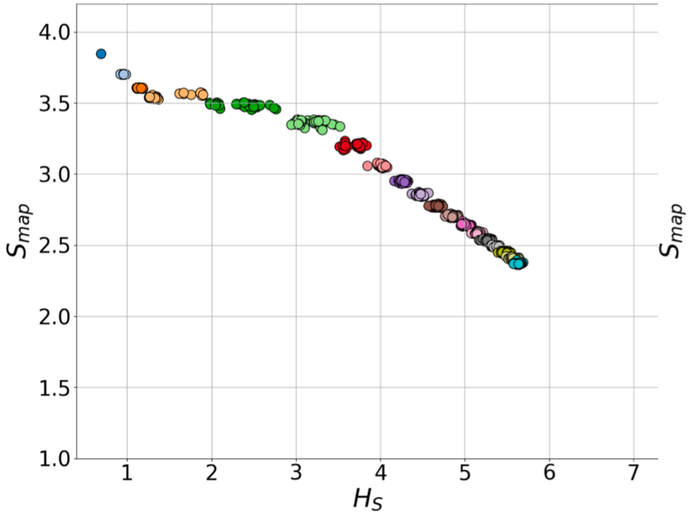


N=100 neurons, p= 5 patterns

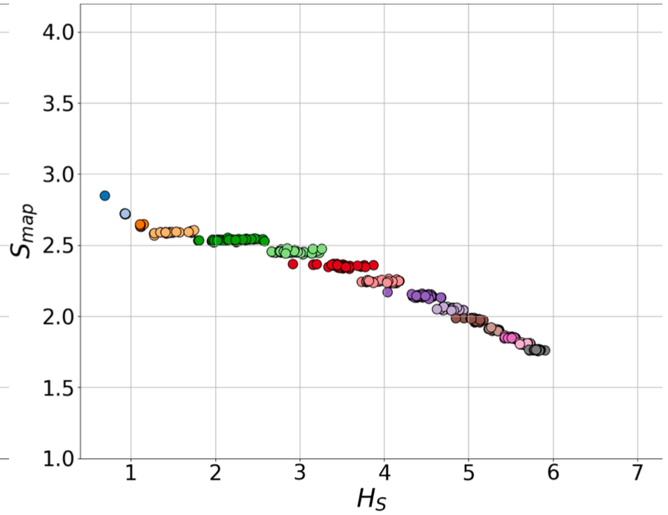




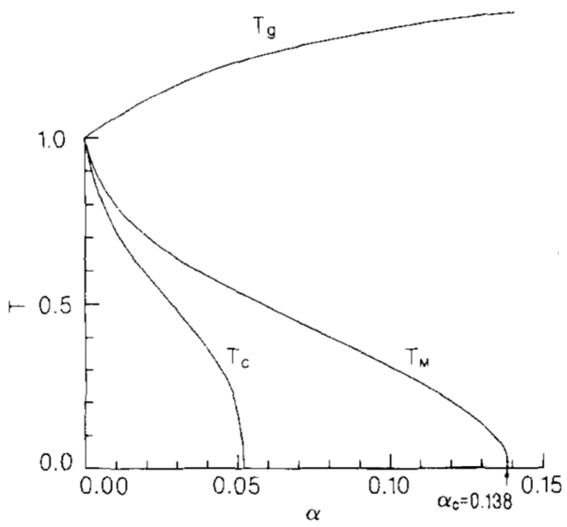
Moving towards the SG phase

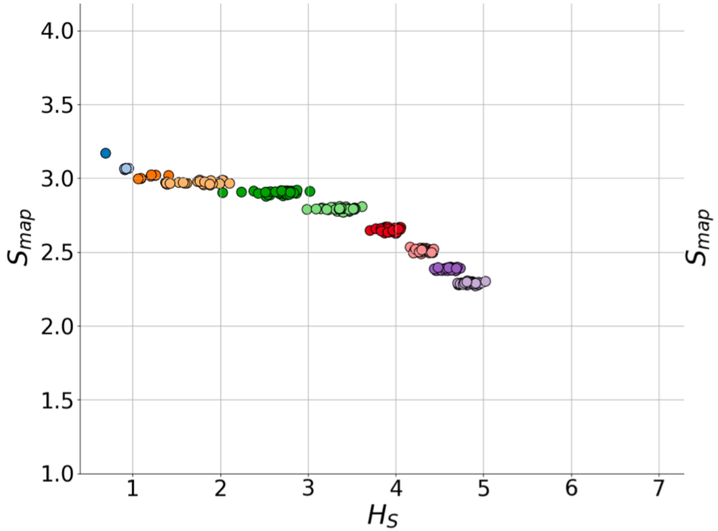


(e) $p = 6$

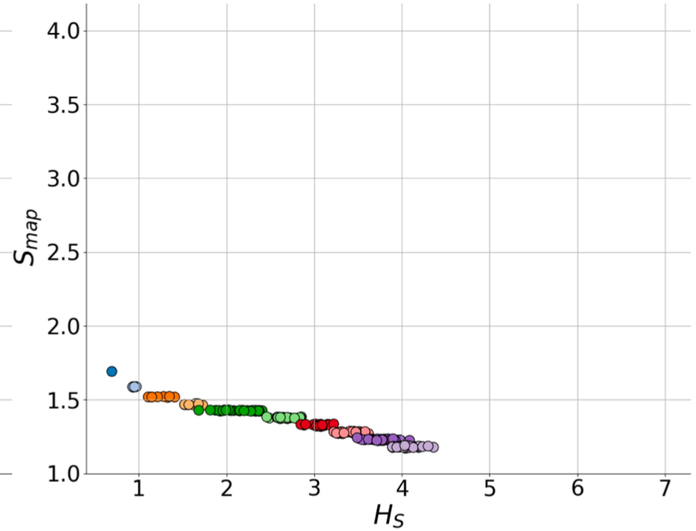


(f) $p = 7$

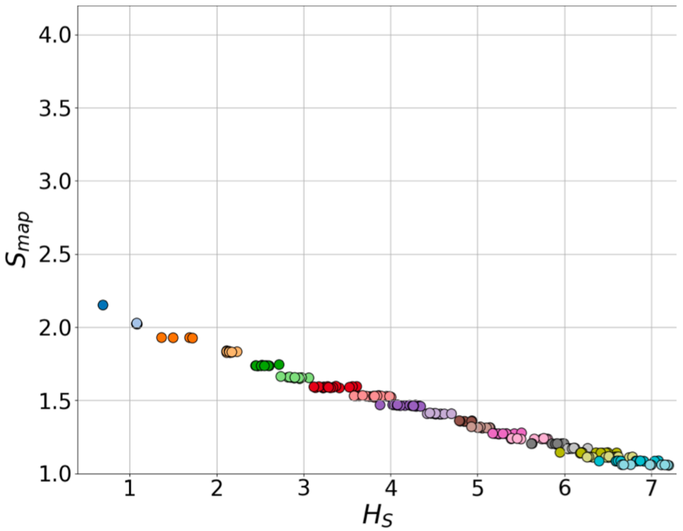




(a) $p = 8$



(b) $p = 9$

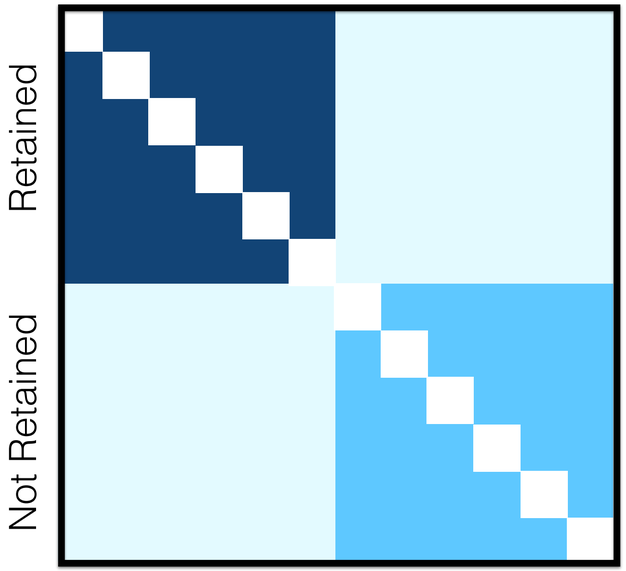


(c) $p = 10$

Semidispersion spectrum

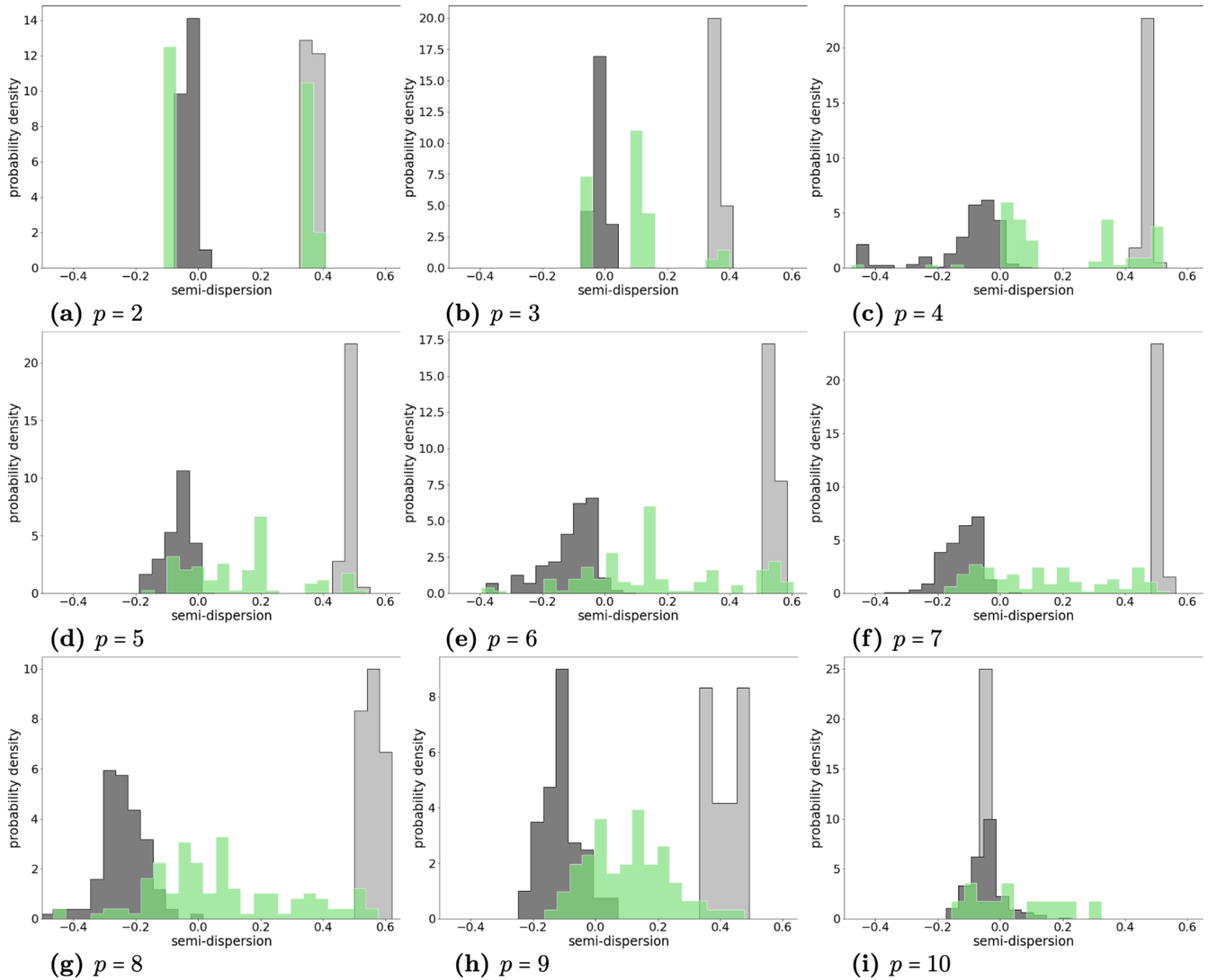
$$S = \left(\begin{array}{c|c} s^* s^* & s^* s \\ \hline s\ s^* & s\ s \end{array}\right)$$

Ordered coupling matrix $|J_{ij}|$



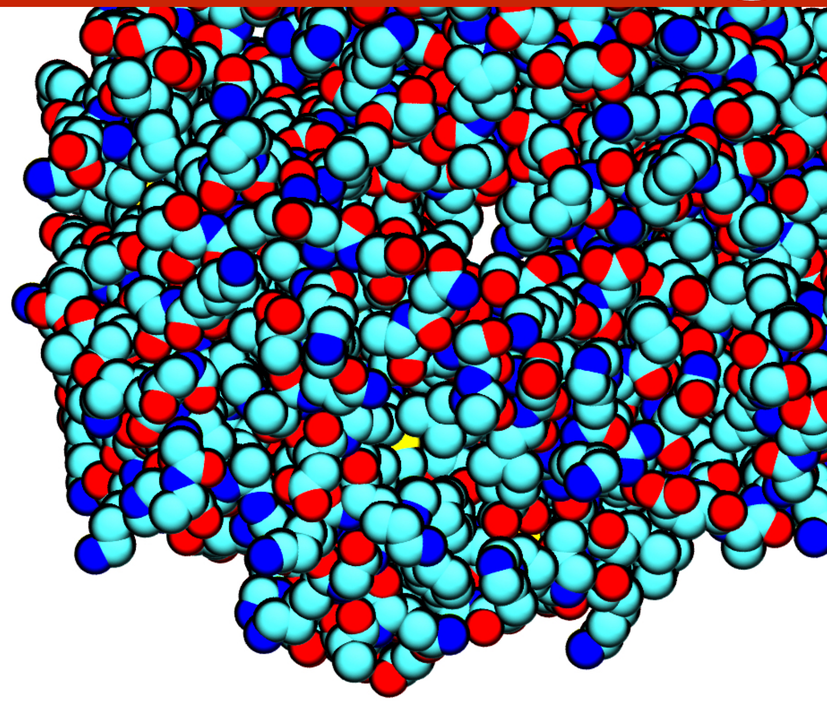
Semidispersion

$$\rho = \frac{\langle |J| \rangle_{s^* s^*} - \langle |J| \rangle_{s s^*}}{\langle |J| \rangle_{s^* s^*} + \langle |J| \rangle_{s s^*}}$$

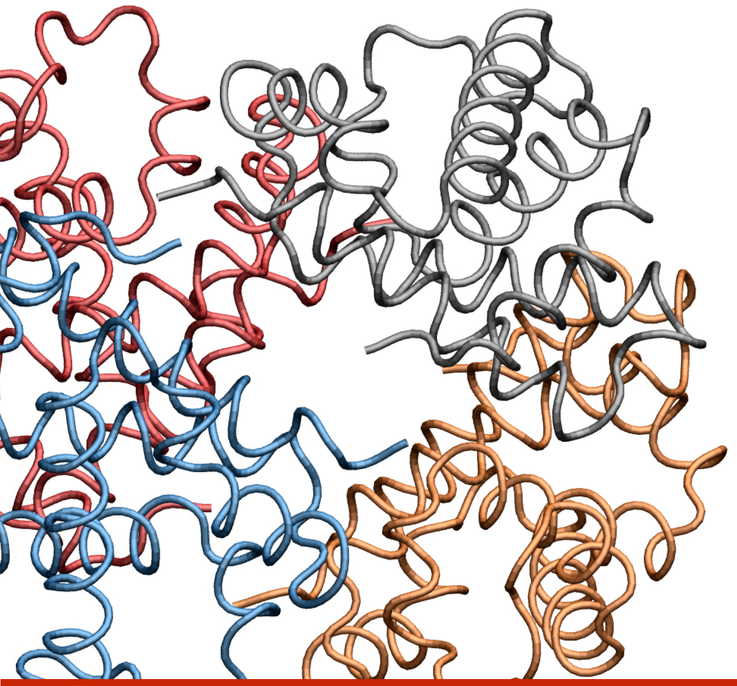


Simplify to understand

DATA != INFORMATION



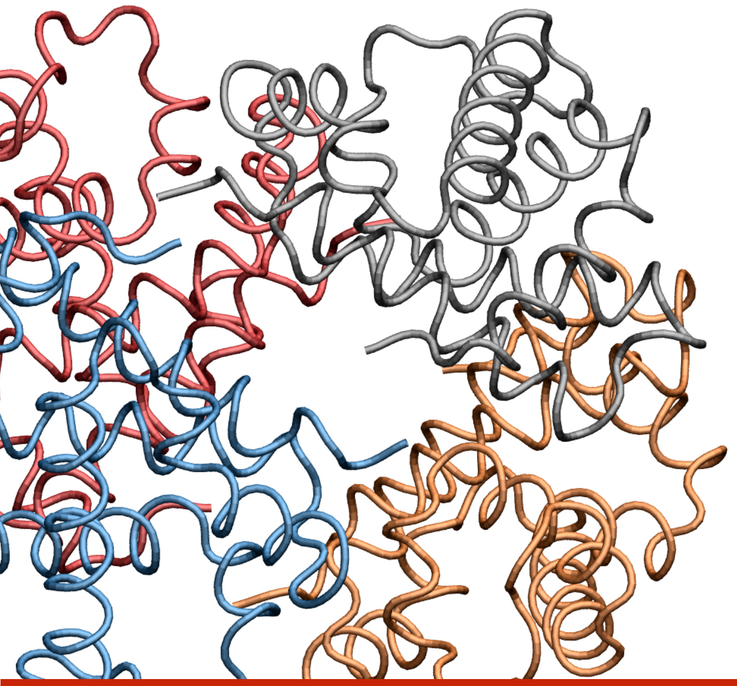
Simplify "just right"
to understand
without losing information



Dimensionality reduction:
in terms of what variables?

Simplify to understand

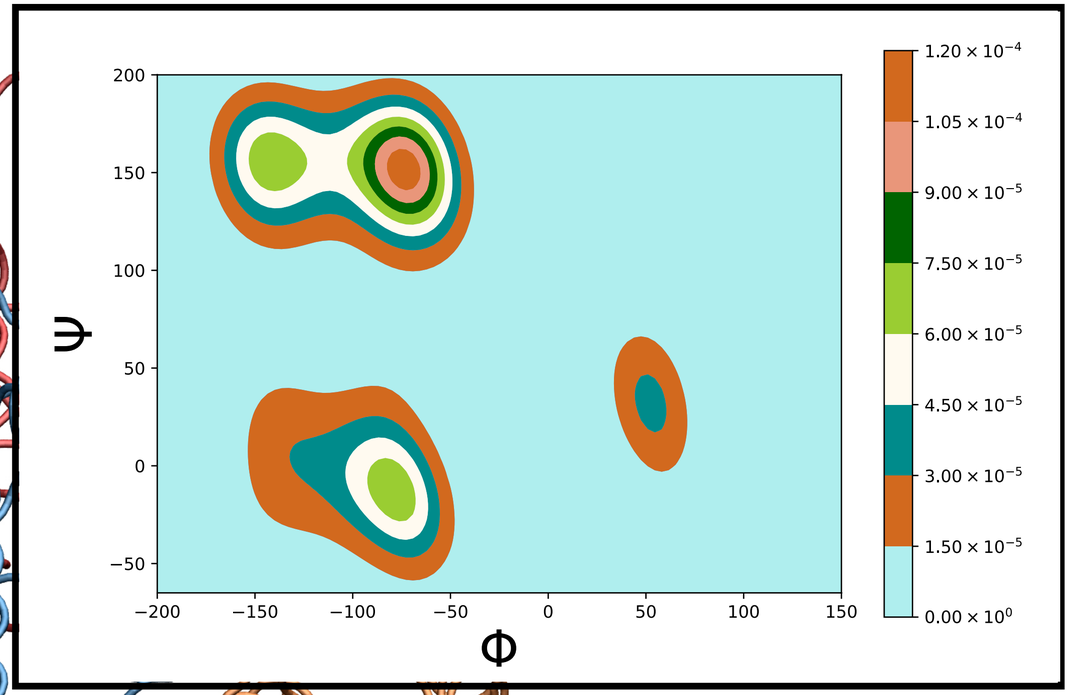
```
ATOM   1153  CD   GLN X 408     110.020  92.423  42.297  0.00  0.00
ATOM   1154  OE1  GLN X 408     110.283  93.074  43.331  0.00  0.00
ATOM   1155  NE2  GLN X 408     108.983  91.618  42.373  0.00  0.00
ATOM   1156 HE21  GLN X 408     108.656  91.204  41.511  0.00  0.00
ATOM   1157 HE22  GLN X 408     108.471  91.768  43.225  0.00  0.00
ATOM   1158  C    GLN X 408     113.167  95.332  40.798  0.00  0.00
ATOM   1159  O    GLN X 408     113.764  95.563  39.715  0.00  0.00
ATOM   1160  N    GLN X 409     113.578  95.654  42.042  0.00  0.00
ATOM   1161  H    GLN X 409     113.018  95.252  42.780  0.00  0.00
ATOM   1162  CA   GLN X 409     114.881  96.418  42.260  0.00  0.00
ATOM   1163  HA   GLN X 409     115.146  96.965  41.356  0.00  0.00
ATOM   1164  CB   GLN X 409     114.444  97.423  43.339  0.00  0.00
ATOM   1165  HB1  GLN X 409     115.314  97.840  43.844  0.00  0.00
ATOM   1166  HB2  GLN X 409     113.900  96.872  44.111  0.00  0.00
```



Simplify to understand

Collective variables

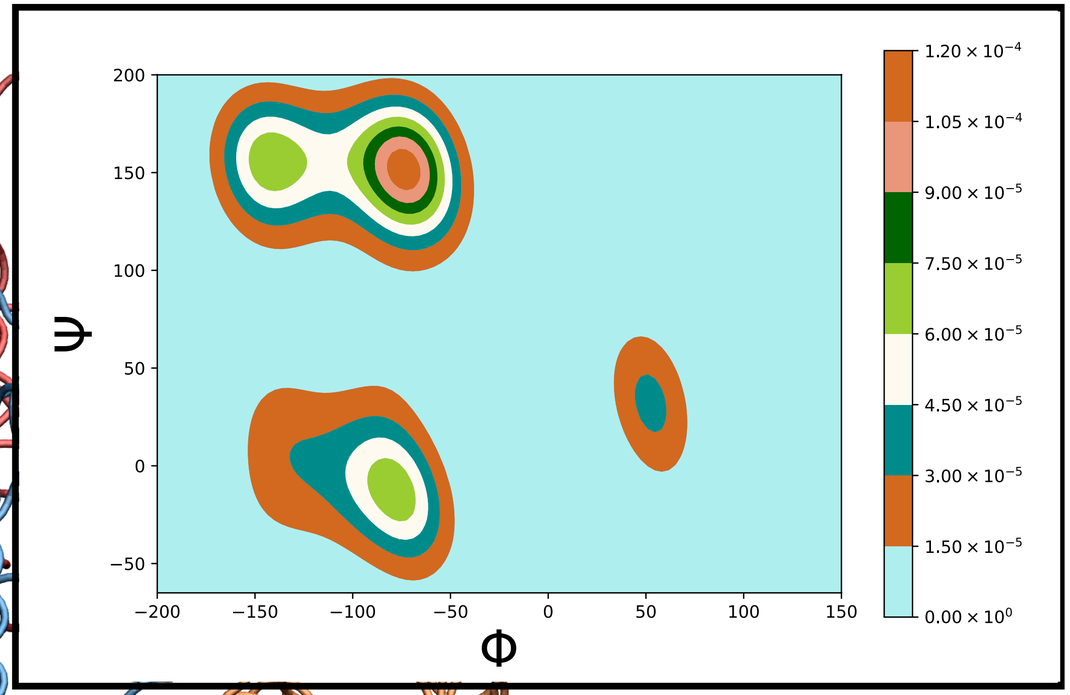
```
ATOM   1153  CD   GLN X 408     110.020  92.423  42.297  0.00  0.00
ATOM   1154  OE1  GLN X 408     110.283  93.074  43.331  0.00  0.00
ATOM   1155  NE2  GLN X 408     108.983  91.618  42.373  0.00  0.00
ATOM   1156 HE21  GLN X 408     108.656  91.204  41.511  0.00  0.00
ATOM   1157 HE22  GLN X 408     108.471  91.768  43.225  0.00  0.00
ATOM   1158  C    GLN X 408     113.167  95.332  40.798  0.00  0.00
ATOM   1159  O    GLN X 408     113.764  95.563  39.715  0.00  0.00
ATOM   1160  N    GLN X 409     113.578  95.654  42.042  0.00  0.00
ATOM   1161  H    GLN X 409     113.018  95.252  42.780  0.00  0.00
ATOM   1162  CA   GLN X 409     114.881  96.418  42.260  0.00  0.00
ATOM   1163  HA   GLN X 409     115.146  96.965  41.356  0.00  0.00
ATOM   1164  CB   GLN X 409     114.444  97.423  43.339  0.00  0.00
ATOM   1165  HB1  GLN X 409     115.314  97.840  43.844  0.00  0.00
ATOM   1166  HB2  GLN X 409     113.900  96.872  44.111  0.00  0.00
```



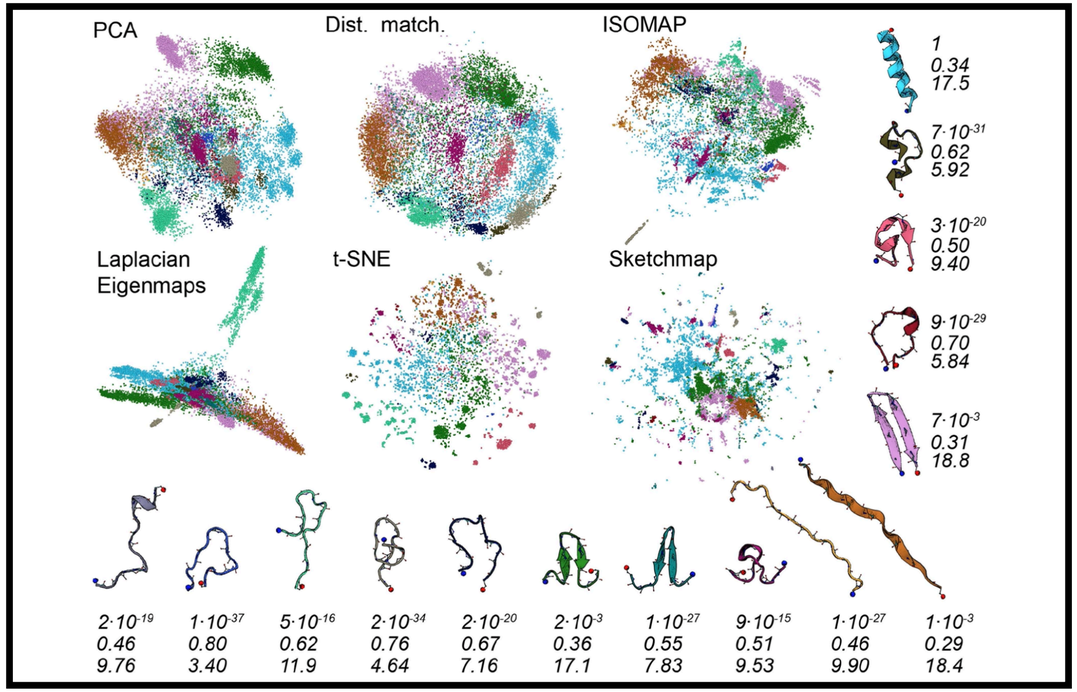
Simplify to understand

```
ATOM   1153  CD   GLN X 408     110.020  92.423  42.297  0.00  0.00
ATOM   1154  OE1  GLN X 408     110.283  93.074  43.331  0.00  0.00
ATOM   1155  NE2  GLN X 408     108.983  91.618  42.373  0.00  0.00
ATOM   1156 HE21  GLN X 408     108.656  91.204  41.511  0.00  0.00
ATOM   1157 HE22  GLN X 408     108.471  91.768  43.225  0.00  0.00
ATOM   1158  C    GLN X 408     113.167  95.332  40.798  0.00  0.00
ATOM   1159  O    GLN X 408     113.764  95.563  39.715  0.00  0.00
ATOM   1160  N    GLN X 409     113.578  95.654  42.042  0.00  0.00
ATOM   1161  H    GLN X 409     113.018  95.252  42.780  0.00  0.00
ATOM   1162  CA   GLN X 409     114.881  96.418  42.260  0.00  0.00
ATOM   1163  HA   GLN X 409     115.146  96.965  41.356  0.00  0.00
ATOM   1164  CB   GLN X 409     114.444  97.423  43.339  0.00  0.00
ATOM   1165  HB1  GLN X 409     115.314  97.840  43.844  0.00  0.00
ATOM   1166  HB2  GLN X 409     113.900  96.872  44.111  0.00  0.00
```

Collective variables



Low-D embedding



Tribello, Gasparotto, Front. Mol. Biosci. 2019