Typical properties of efficient representations of high dimensional data

Matteo Marsili
The Abdus Salam ICTP, Trieste

+ A. Haimovici UBA/NETADIS
+ S. Grigolon, S. Franz LPTMS/NETADIS
+ Y. Roudi, N. Bulso, C. Battistin NTNU/NETADIS
+ I. Mastromatteo, E. Politech. Paris
+ Vincenzo Carnevale, Daniele Granata (Temple U)
+ J A Ryan Cubero, Mojtaba Khodadadi (ICTP)
+ Nicola Prezza, Alberto Policriti (Univ. Udine)
+ Clelia De Mulatier (ICTP)
+ Juyong Song, Junghyo Jo APCTP, Pohang

Outline

- Motivation and questions:
  - science in the Big Data era. What are the relevant variables?
  - why are broad abundance distributions so ubiquitous?

- Main result: Relevance = entropy of the frequency distribution
  - minimally sufficient representations
  - maximally informative samples have broad distributions

- Where is this realised and how can it be used?
  - Minimum Description Length (optimal coding)
  - deep learning
  - relevant neurons in rat's navigation
  - relevant residues in protein sequences

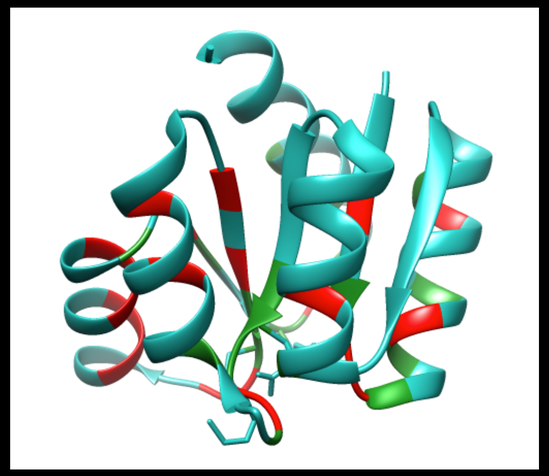
- Conclusions

Science in the Big Data era

e.g.

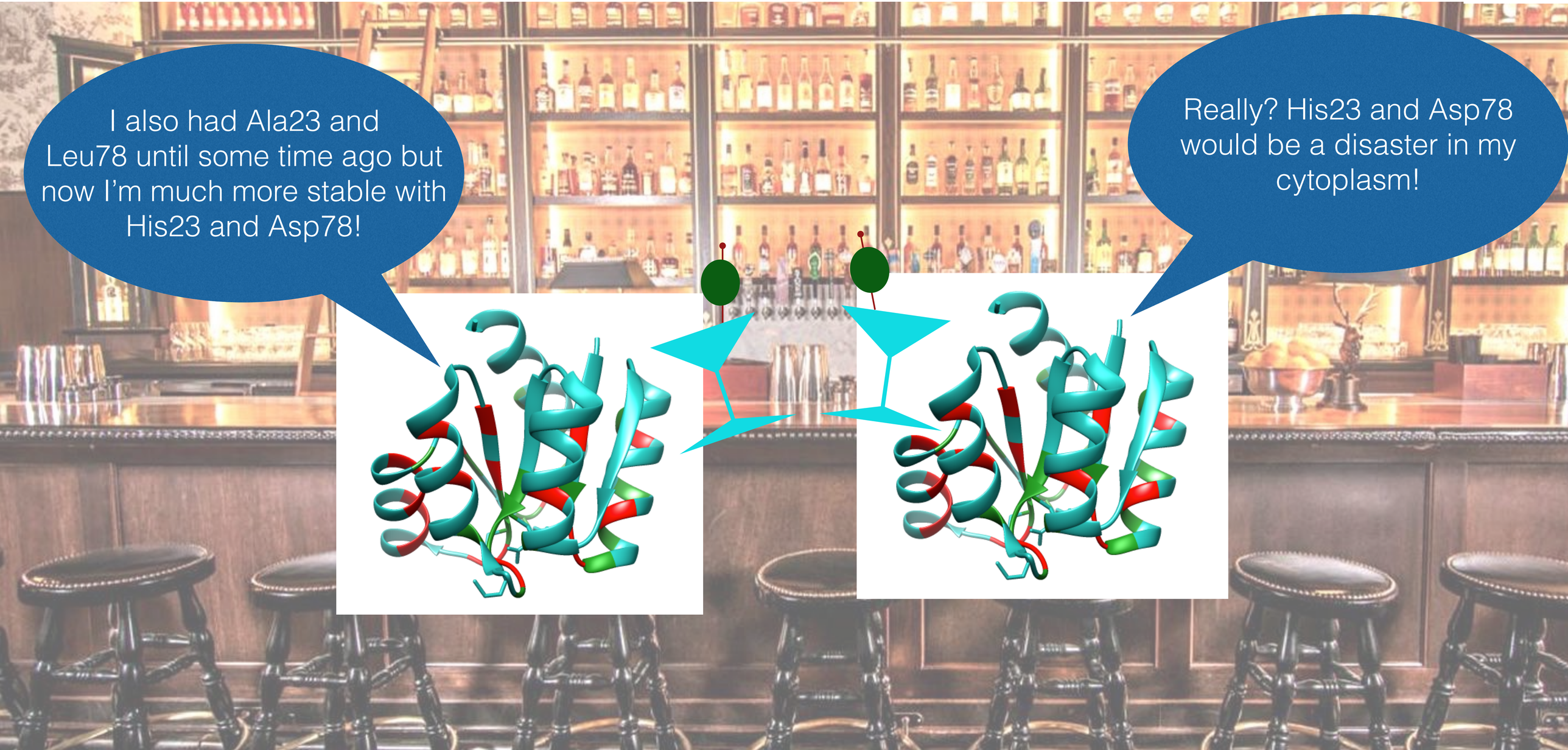
```
-CTMPMEAGSCDGKLARWHFARDDNKCMPFYYTGCGGNHNQFISLDQCEEQC-
-CTNTLAQGEGPLSVTRYYFNAQSRTCDEFMFRGLKGNSNNFNSLAECEKAC-
-CTQPKDSGVCSGSQRSFYFDTRMKVCQPFLYSGCGGNENRFSTSKECRDACQ
-CTQPRDNGSCSENGRAFYFDTRTKVCQPFLYSGCGGNDNRFATSKECRDSCQ
-CTQTLAQGEGPLSVARFYFNAQSRTCDEFMFRGLKGNSNNFKSQEDCEKAC-
-CTQTLAQGEGPLSVARFYFNAQSRTCDEFMFRGLKGNSNNFKSQEDCEKAC-
-CTQTLAQGEGPLSVARFYFNAQSRTCDEFMFRGLKGNSNNFKSQEDCEKAC-
-CTSPPVTGPCRAGFKRYNYNTRTKQCEPFKYGGCKGNGNRYKSEQDCLDAC-
-CTVLPSEGYCKKRYFRFLYDSNTKTCQLFWYRGCGGTENNFPTYYSCLDRC-
-CTVLPSEGYCRKKYFRFLYDSNTKTCQLFWYRGCGGTENNFPTYYACLDRC-
-CTVQPTNGLCVPSTLGIYFDVETQHCR---FLGC-GNKRLFASLEDCEKIC-
-CTVQPTNGLCVPSTLGIYFDVETQHCR---FLGC-GNKRLFASLEDCEKIC-
-CVAKPDAGPCRAAFPAFFYDPDTNSCQPFIYGGCRGNGNRYNSREECLSRC-
-CVAPLDKCP--GNVIIYYYN-RTSGCQQMHRGNCSDN-GNYPTLQECQEYCL
-CVDLPDTGLCKESIPRWYYNPFSEHCARFTYGGCYGNKNNFEEEQQCLESCR
-CVDLPDTGLCKESIPRWYYNPFSEHCARFTYGGCYGNKNNFEEEQQCLESCR
…
```

?

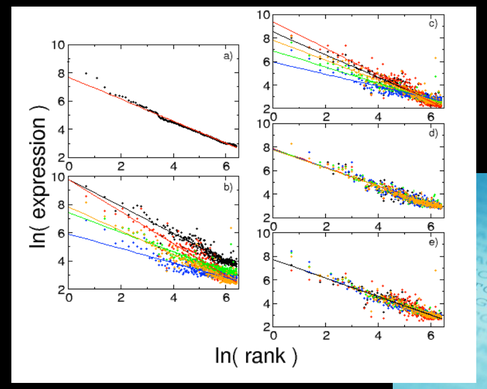


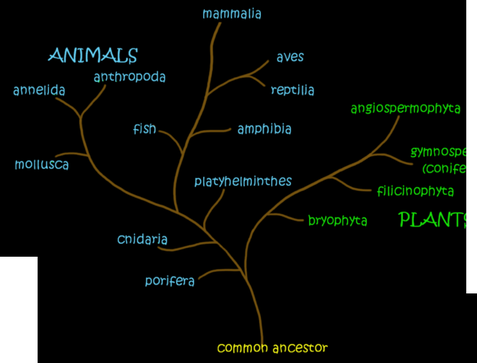
- Big Data is not that big (e.g.$s \in \{A, B, \ldots, W\}^D$, D~10^2, N~10^4 << 20^D)
- Model is unknown
- What are the relevant variables (e.g. residues along the sequence)?
- How much information does the data contains about the model?
- How to quantify relevant information?

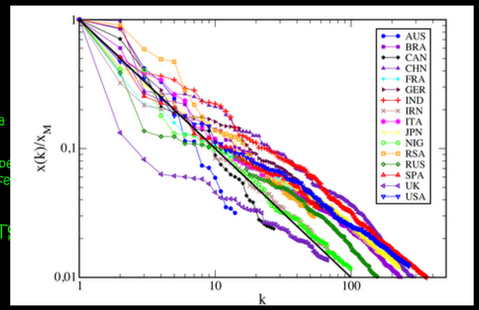
Intrinsic relevance: e.g. relevant positions for proteins are those they talk about when they meet

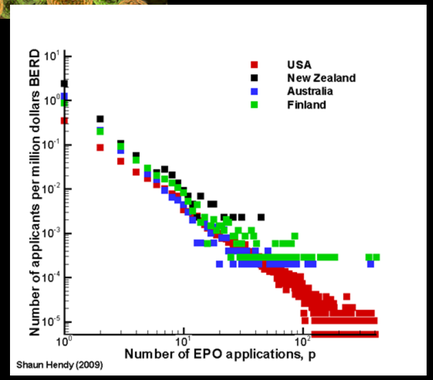


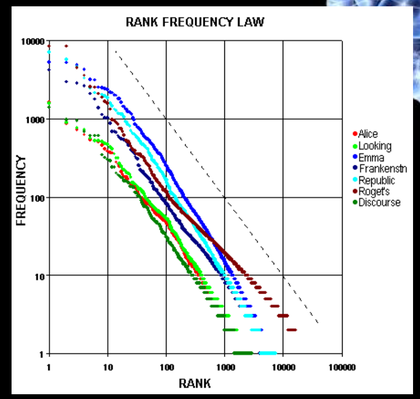
Complexity ~ broad frequency distributions: why?

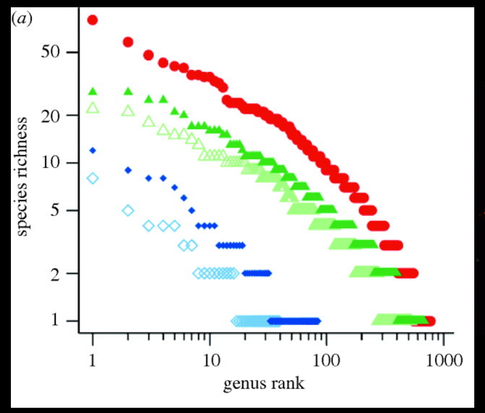


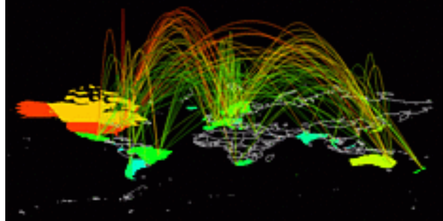


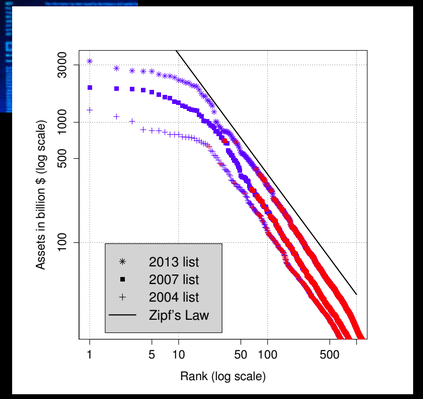










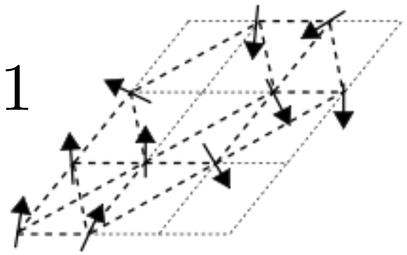


Broad distributions are the exception in physics

- Statistical mechanics: order and disorder

$$\underline{s} = (s_1, \ldots, s_N), \qquad s_i = \pm 1$$

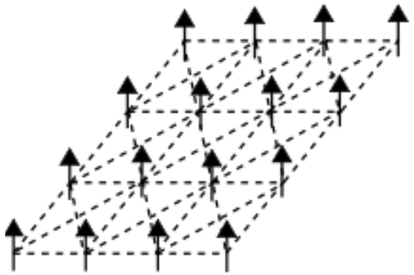
$$p\{\underline{s}|\hat{g}\} = \frac{1}{Z} e^{-E_{\hat{g}}[\underline{s}]/T}$$



$T \gg T_c$

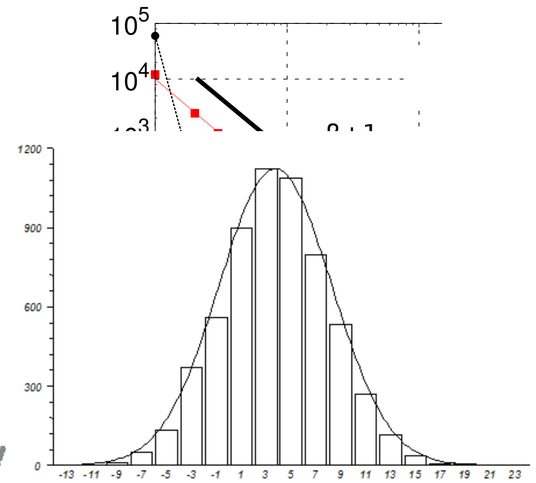
Weak interaction
Large entropy

critical point T_c



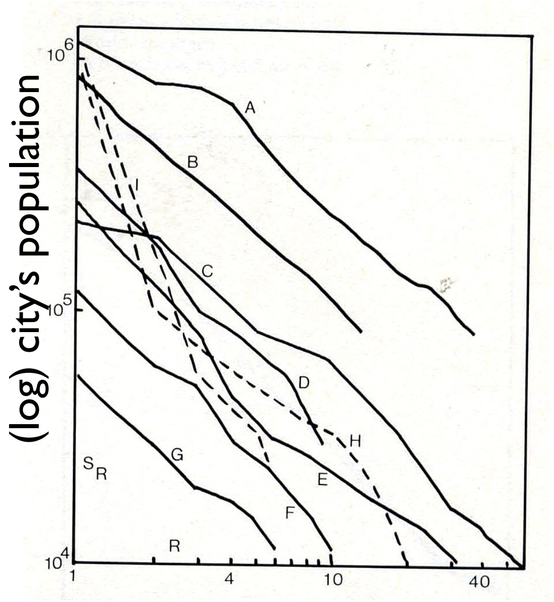
$T \ll T_c$

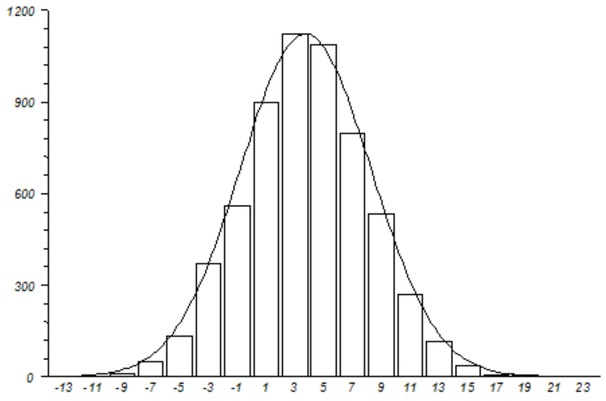
Strong interaction
Small entropy



- Self-Organized Criticality (P. Bak, How Nature Works, 1996)
- Statistical Criticality (?)

The distribution depends on which variables we measure
e.g. where do you live?





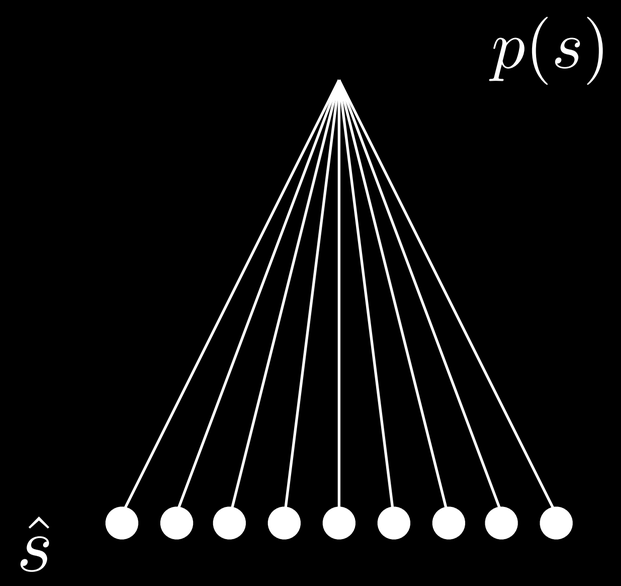
population by zip code

The relevant variables for a complex systems
- may not be those an engineer would choose
- appear with a non-trivial frequency distribution (**statistical citicality**)

btw: where do I live? At my zip-code 34151=13 x 37 x 71 which nicely decomposes in primes

Minimally sufficient representations

Data: $\hat{s} = (s^{(1)}, \ldots, s^{(N)})$ → Model: $P(\hat{s}) = \prod_{i=1}^{N} p(s^{(i)})$



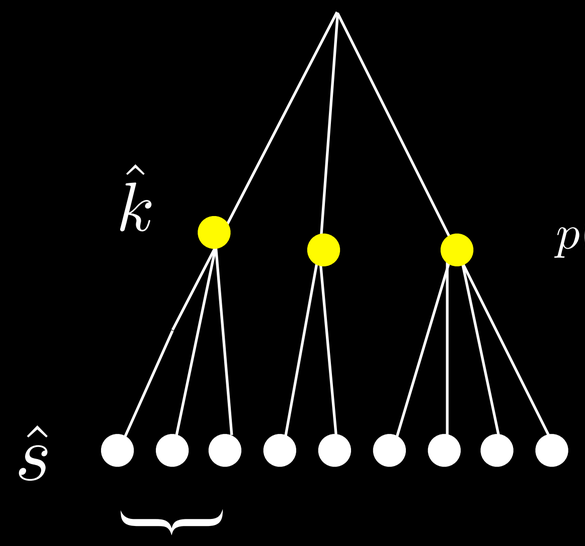
Information content (coding cost/data point):

$$\hat{H}[s] = -\sum_{s} \frac{k_s}{N} \log \frac{k_s}{N} \qquad k_s = |\{i : s_i = s\}|$$

How many of these bits are just noise and how many of them can provide useful information on p(s)?

Minimally sufficient representations

Data: $\hat{s} = (s^{(1)}, \ldots, s^{(N)})$ → Model: $P(\hat{s}) = \prod_{i=1}^{N} p(s^{(i)})$



$$p(k) = \frac{k m_k}{N}$$

$$p(s|k) = \frac{1}{m_k} \delta_{k_s,k}$$

$$m_k = |\{s : \ k_s = k\}|$$

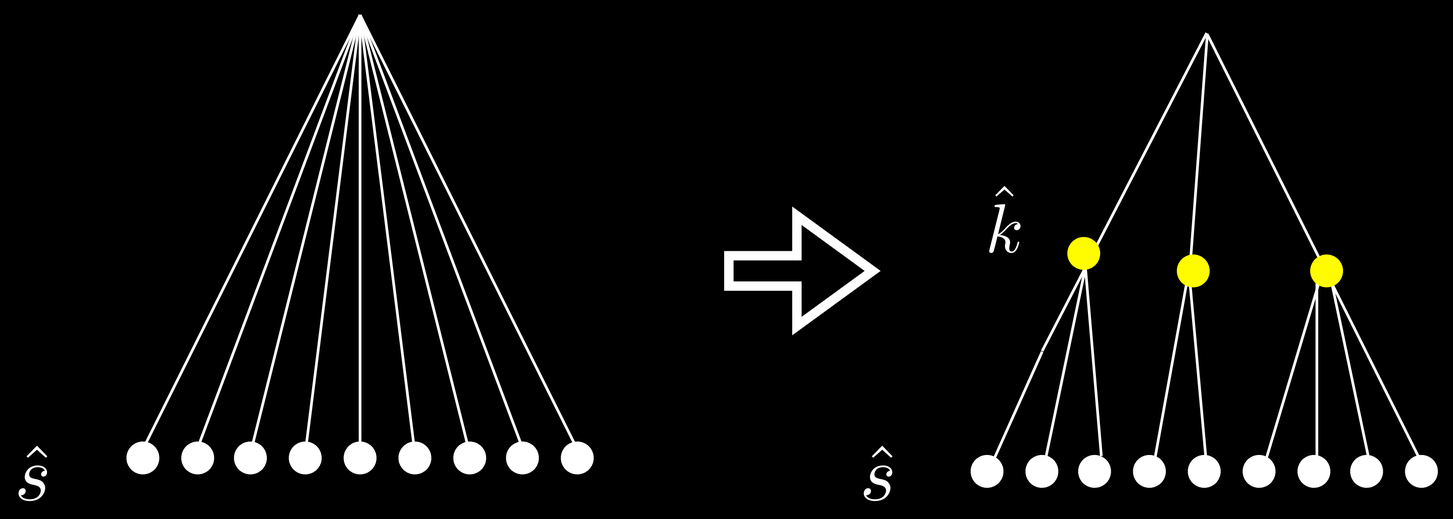
Information content (coding cost/data point):

$$\hat{H}[s] = \hat{H}[k] + \hat{H}[s|k]$$

Useful information

Noise

H[s]=Resolution, H[k]= Relevance



$\rightarrow H[k]$ intrinsic resolution = relevance

$\rightarrow H[s]$ user defined resolution
(e.g. classification of products, firms, species, molecules, …)

Bayesian point of view:

(+A. Haimovici UBA/NETADIS)

$H[k] \sim$ number of parameters that can be estimated with $\hat{s}$

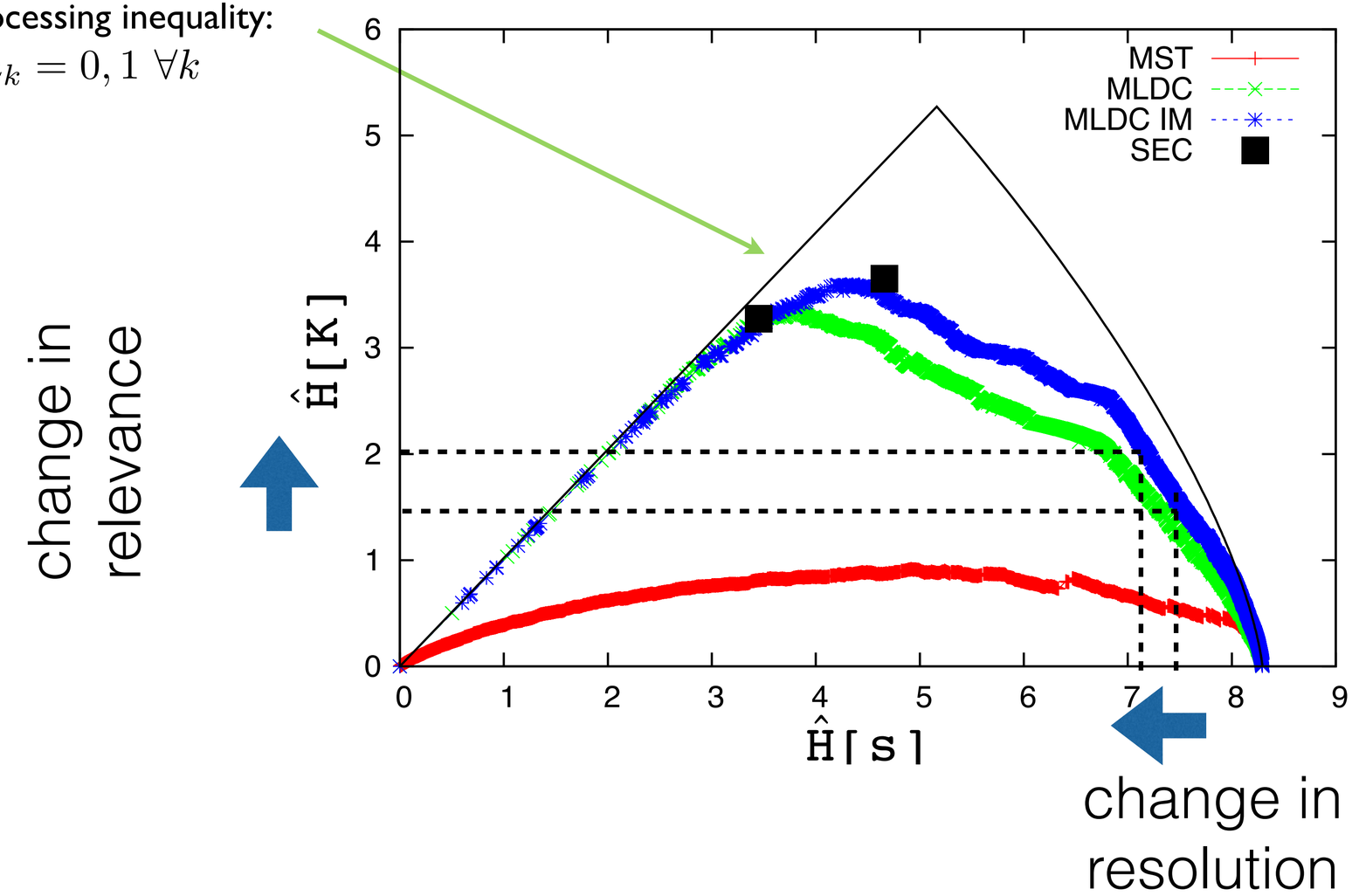
Idea: if states s and s' have $k_s = k_{s'}$ then they have the same probability

$\Rightarrow$ s and s' can be distinguished only if $k_s \neq k_{s'}$

Resolution - Relevance trade-off

Data processing inequality:

$m_k = 0, 1 \; \forall k$



e.g. clustering of 4000 stocks in NYSE

Maximally informative samples look critical

$$m_k^* = \arg\max_{m_k} H[k]$$

$$\text{s.t.} \quad H[s] = H_0, \quad \sum_k k m_k = N$$

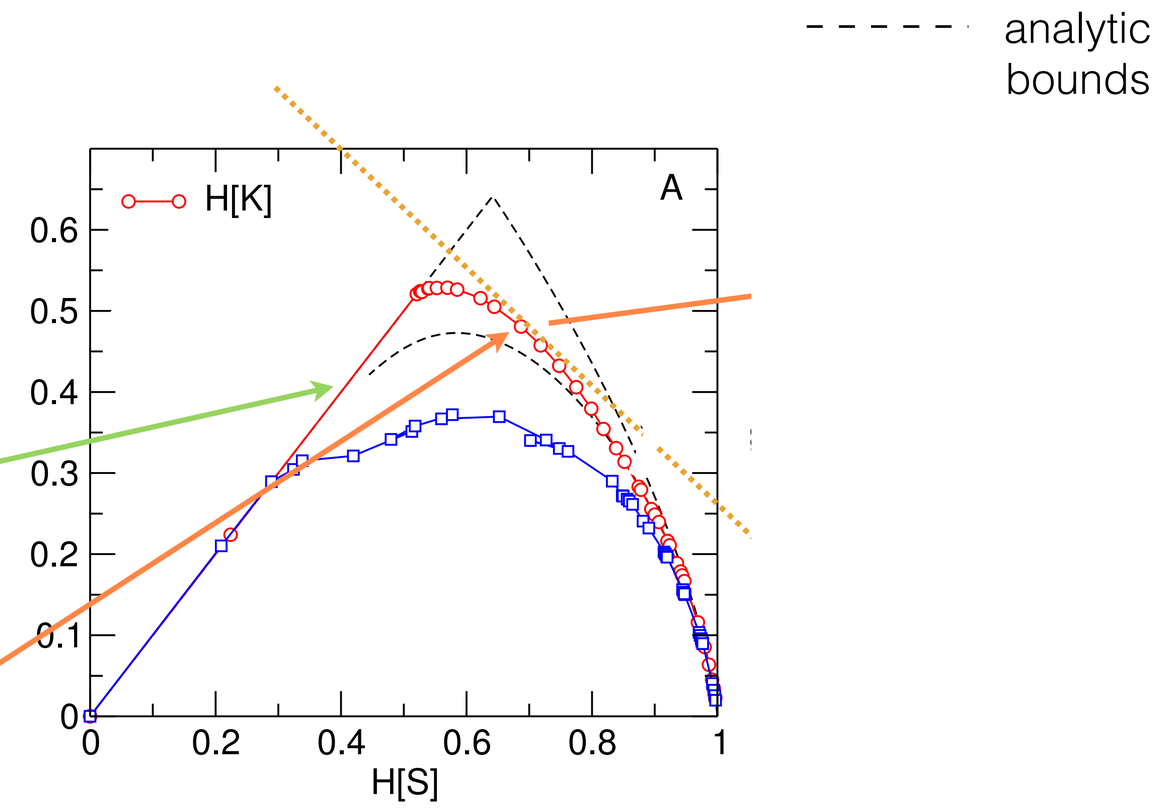
Data processing inequality:

$$m_k = 0, 1 \ \forall k$$

Resolution-relevance tradeoff:

$$\max_{m_k} \{H[k] + \mu H[s]\}$$

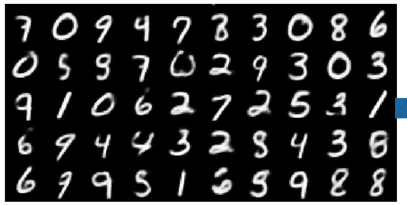
$\Rightarrow$ Zipf's law (μ=1) $\Leftrightarrow$ optimal compression



Where do we expect to find maximally informative samples

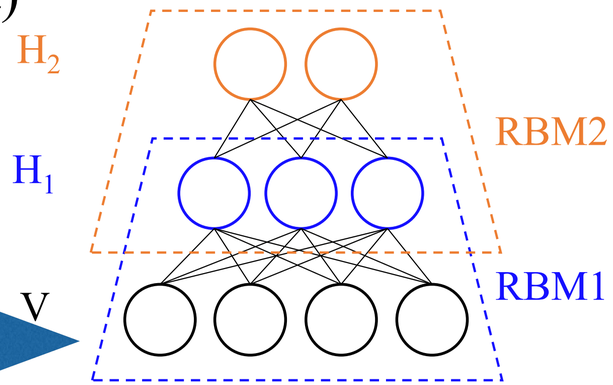
and how can we use it to find relevant variables?

Why does deep learning works so well?



data (e.g. MNIST)

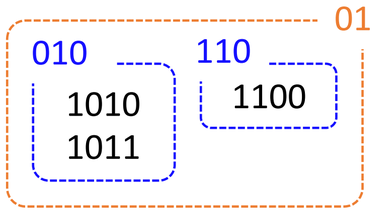
(a)



(b)

V	H_1	H_2
1010	010	01
1011	010	01
1100	110	01

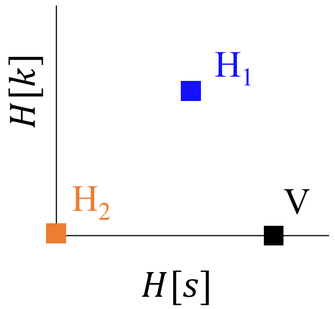
(c)



(d)

	s	k
V	1010	1
	1011	1
	1100	1
H_1	010	2
	110	1
H_2	01	3

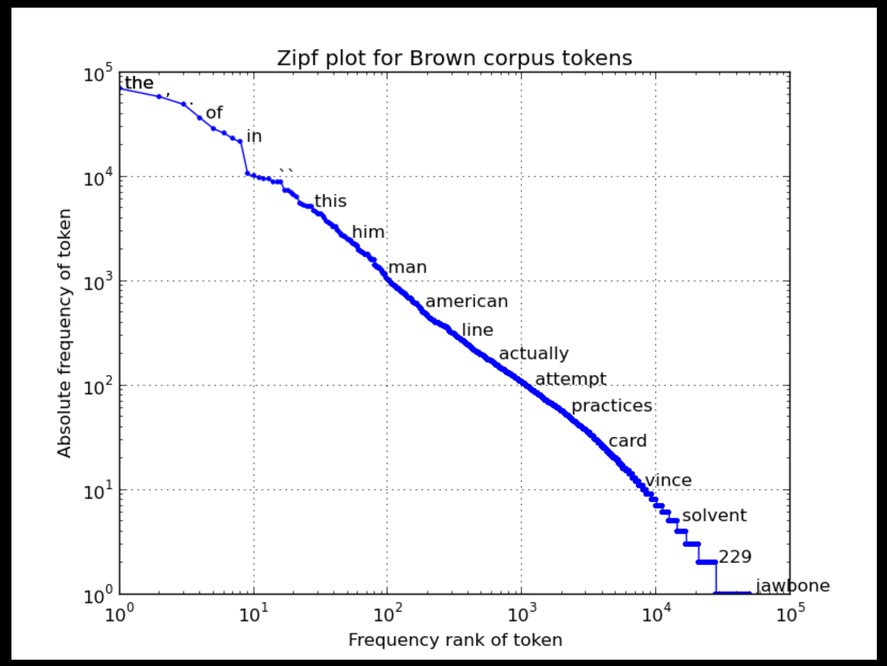
	k	m_k	km_k
V	1	3	3
H_1	1	1	1
	2	1	2
H_2	3	1	3



(Juyong Song, Junghyo Jo APCTP, Pohang)

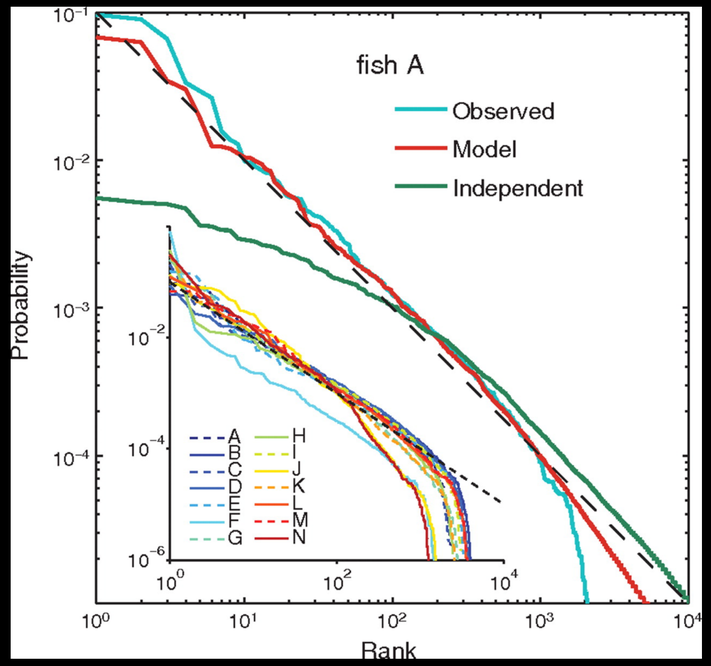
Zipf's law in efficient representations

Language
Frequency of words



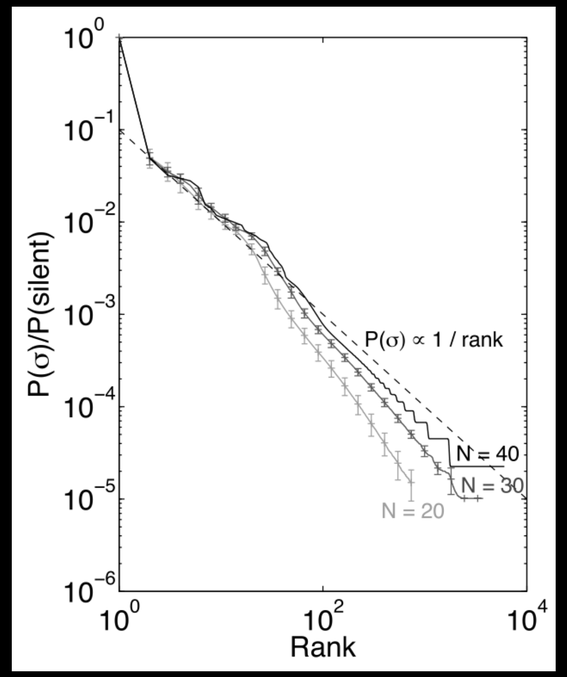
Zipf 1949
(from Finn Årup Nielsen's blog)

Immune system
Antibody binding seqs



Mora, et al. PNAS (2010)

Neurons (retina)
Firing patterns



Tkacik et al., 2007
Mora and Bialek, 2011

… and in Google matrix

Ermann, Frahm, and Shepelyansky: Google matrix analysis of directed networks

Rev. Mod. Phys., Vol. 87, No. 4, October–December 2015

L.Ermann, D.L.Shepelyansky: Google matrix of the world trade network EPJB 2015

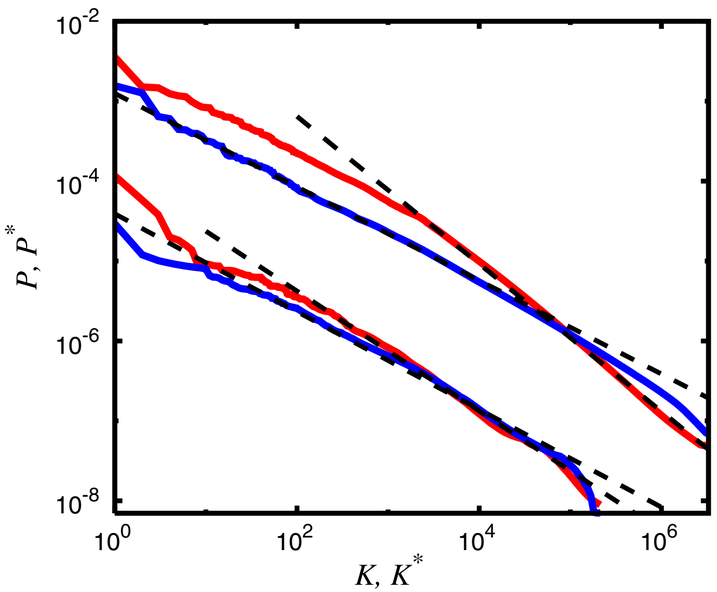


FIG. 5 (color online). Dependence of probabilities of PageRank P [gray (red) curves] and CheiRank P^* [black (blue) curves] vectors on the corresponding rank indices K and K^* for networks of Wikipedia August 2009 (top curves) and University of Cambridge (bottom curves, moved down by a factor of 100). The straight dashed lines show the power law fits for PageRank and CheiRank with the slopes $\beta = 0.92$ and 0.58, respectively, corresponding to $\beta = 1/(\mu_{\text{in,out}} - 1)$ for Wikipedia (see Fig. 2), and $\beta = 0.75$ and 0.61 for Cambridge. From Zhirov, Zhirov, and Shepelyansky, 2010 and Frahm, Georgeot, and Shepelyansky, 2011.

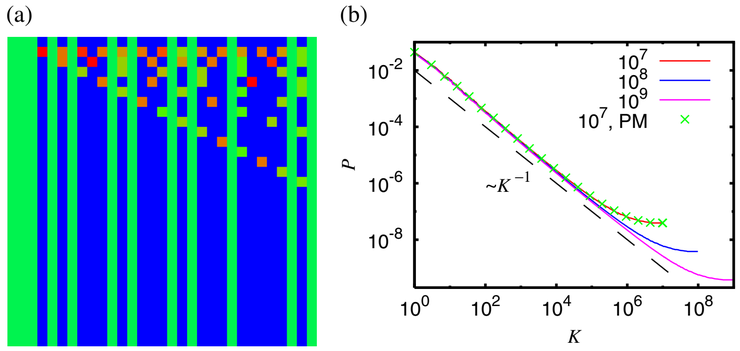


FIG. 42 (color online). (a) The Google matrix of integers. The amplitudes of matrix elements S_{mn} are shown by color black (blue) for minimal zero elements and gray (red) maximal unity elements, with $1 \le n \le 31$ corresponding to the x axis (with $n = 1$ corresponding to the left column) and $1 \le m \le 31$ for the y axis (with $m = 1$ corresponding to the upper row). (b) The full lines correspond to the dependence of the PageRank probability $P(K)$ on index K for the matrix sizes $N = 10^7$, 10^8, and 10^9 with the PageRank evaluated by the exact expression $P \propto \sum_{j=0}^{l-1} v^{(j)}$. The gray (green) crosses correspond to the PageRank obtained by the power method for $N = 10^7$; the dashed straight line shows the Zipf law dependence $P \sim 1/K$. From Frahm, Chepelianskii, and Shepelyansky, 2012.

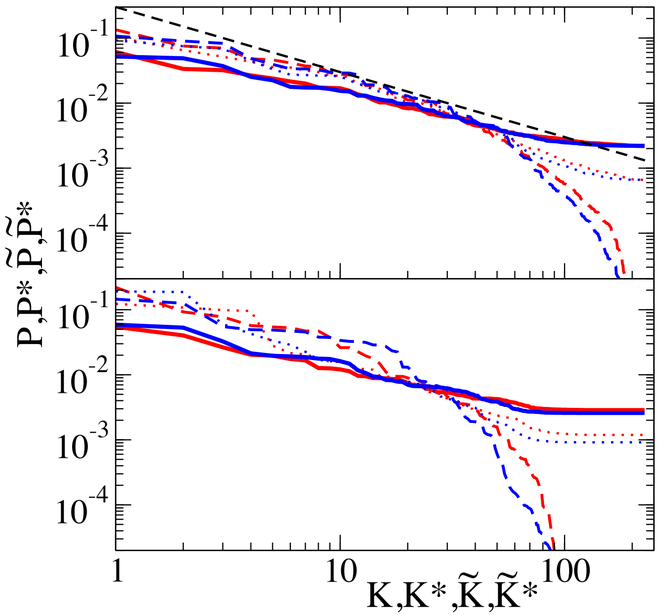


Fig. 2. (Color online) Probability distributions of PageRank $P(K)$, CheiRank $P^*(K^*)$, ImportRank $\tilde{P}(\tilde{K})$, and ExportRank $\tilde{P}^*(\tilde{K}^*)$ are shown as function of their indexes in logarithmic scale for *all commodities* (top panel) and *crude petroleum* (bottom panel) for WTN in 2008 with $N = 227$. Here $P(K)$ and $P^*(K^*)$ are shown by red and blue curves respectively, for $\alpha = 0.5$ (solid curves) and $\alpha = 0.85$ (dotted curves); $\tilde{P}(\tilde{K})$ and $\tilde{P}^*(\tilde{K}^*)$ are displayed by dashed red and blue curves respectively. For both commodities the distributions $P(K)$ and $P^*(K^*)$ follow a power law dependence like $P \propto 1/K^\beta$ (see text), the Zipf law is shown by the straight dashed line with $\beta = 1$ in top panel.

A minimal description of MDL

Alice

$$f(s|\theta) \to \hat{s} = (s_1, \ldots, s_N)$$

$\Rightarrow$

Bob does not know θ

$$\hat{s} \quad \to -\log_2 P(\hat{s}) \quad \text{bits}$$

Regret: $\mathcal{R}(\hat{s}, P) = -\log_2 P(\hat{s}) - \left[-\log_2 f(\hat{s}|\hat{\theta})\right]$

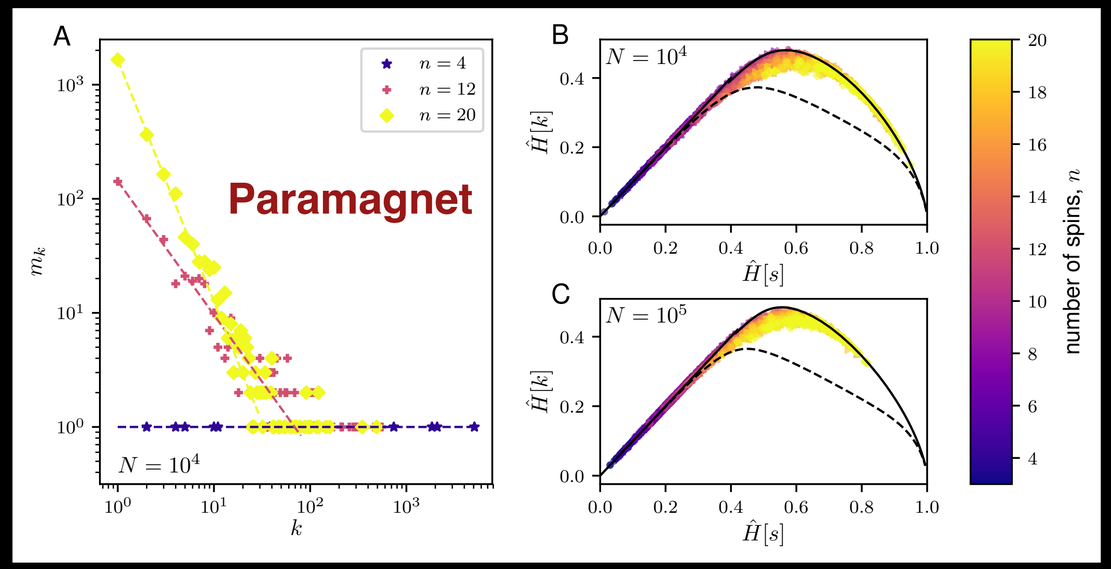
Normalised Maximum Likelihood (NML)

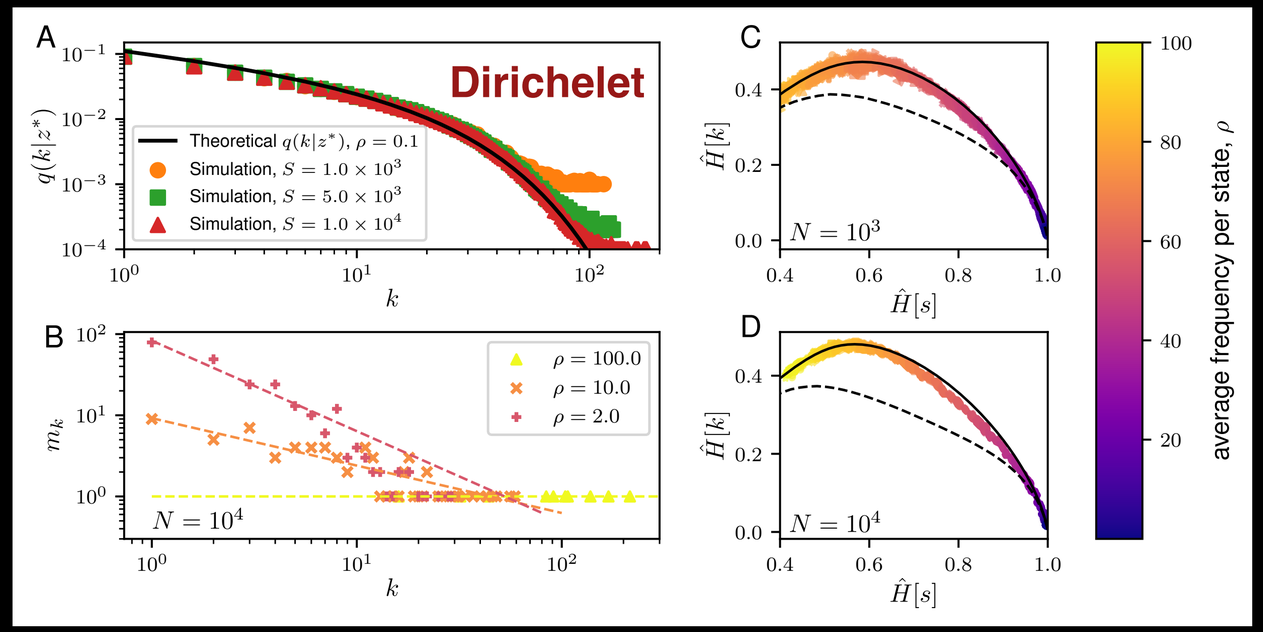
Optimal coding: $\bar{P} = \arg\min_P \max_{\hat{s}} \mathcal{R}(\hat{s}, P) \quad \Rightarrow \quad \bar{P}(\hat{s}) = \dfrac{f(\hat{s}|\hat{\theta})}{\sum_{\hat{s}'} f(\hat{s}'|\hat{\theta}')}$

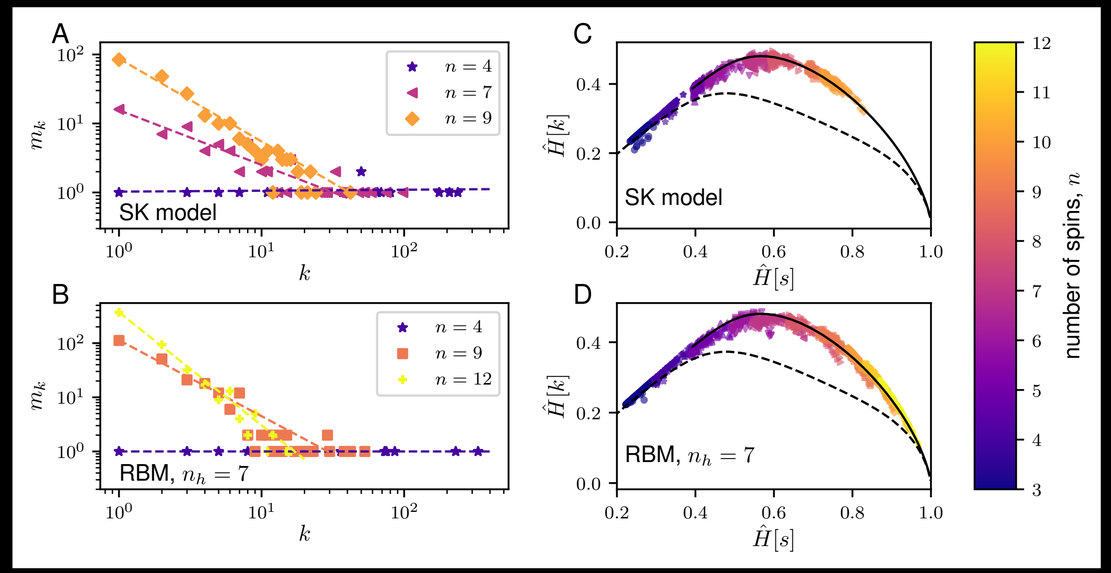
Complexity: $\mathcal{R}(\hat{s}, \bar{P}) = -\log_2 \sum_{\hat{s}'} f(\hat{s}'|\hat{\theta}') \simeq \dfrac{K}{2} \log_2 N + \log_2 \int d\theta \sqrt{\det J(\theta)}$

MDL samples are maximally informative

In a large variety of models, samples generated from Normalised Maximum Likelihood exhibit broad distributions and a high value of H[k]







MDL samples are “poised at criticality”

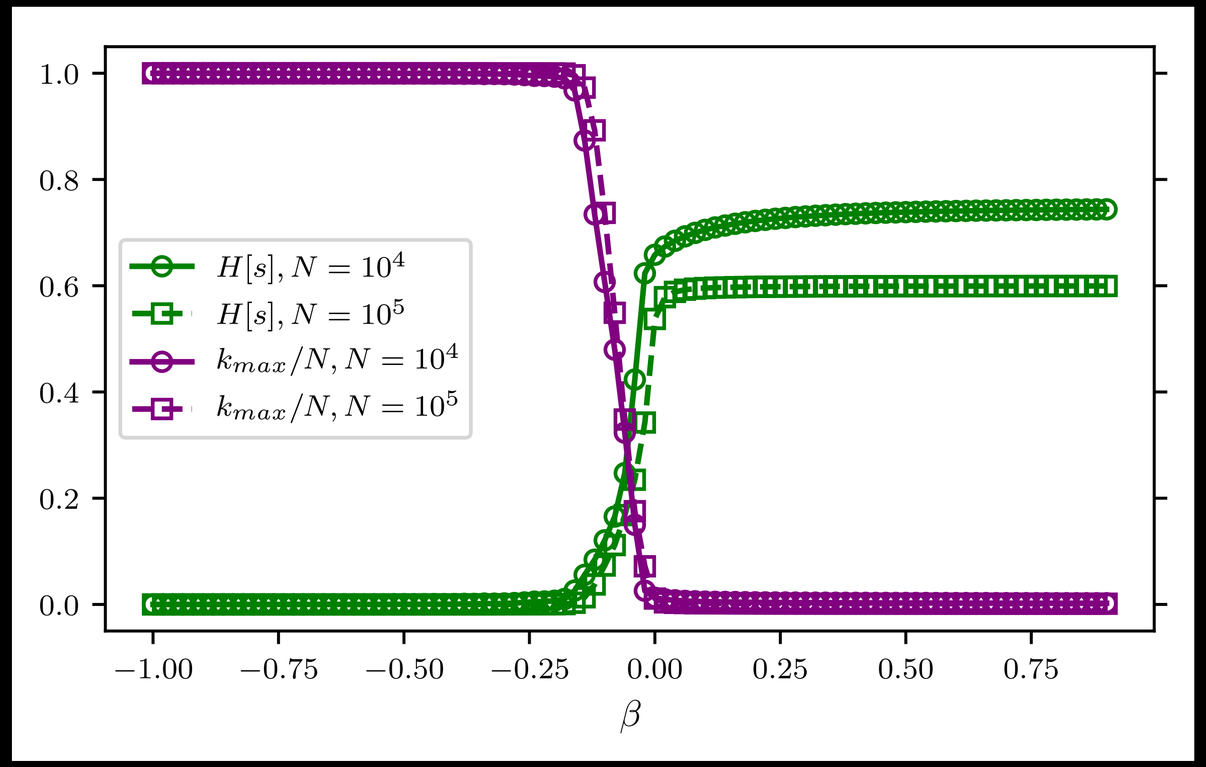
Atypical samples with anomalously low coding cost do not exist.

Proof: Large Deviation Theory

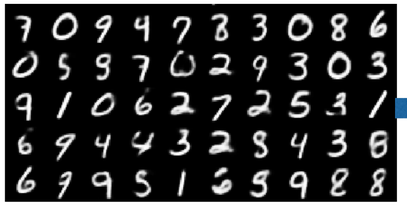
$$P\{\hat{H}[s] = E\} \sim e^{-N \max_\beta [\beta E - \phi(\beta)]}$$

$$\phi(\beta) = \frac{1}{N} \sum_{\hat{s}} \bar{P}(\hat{s}) e^{\beta \hat{H}[s]}$$

Samples with $\hat{H}[s]$ less than the typical value condensate on a single outcome.

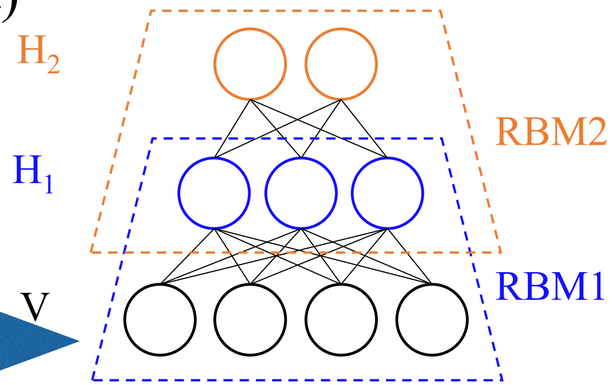


How informative are the representations that deep neural networks extract?



data (e.g. MNIST)

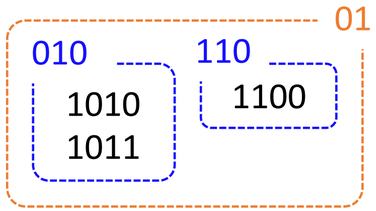
(a)



(b)

V	H_1	H_2
1010	010	01
1011	010	01
1100	110	01

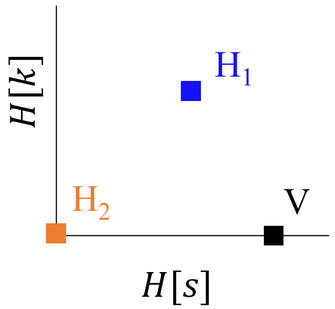
(c)



(d)

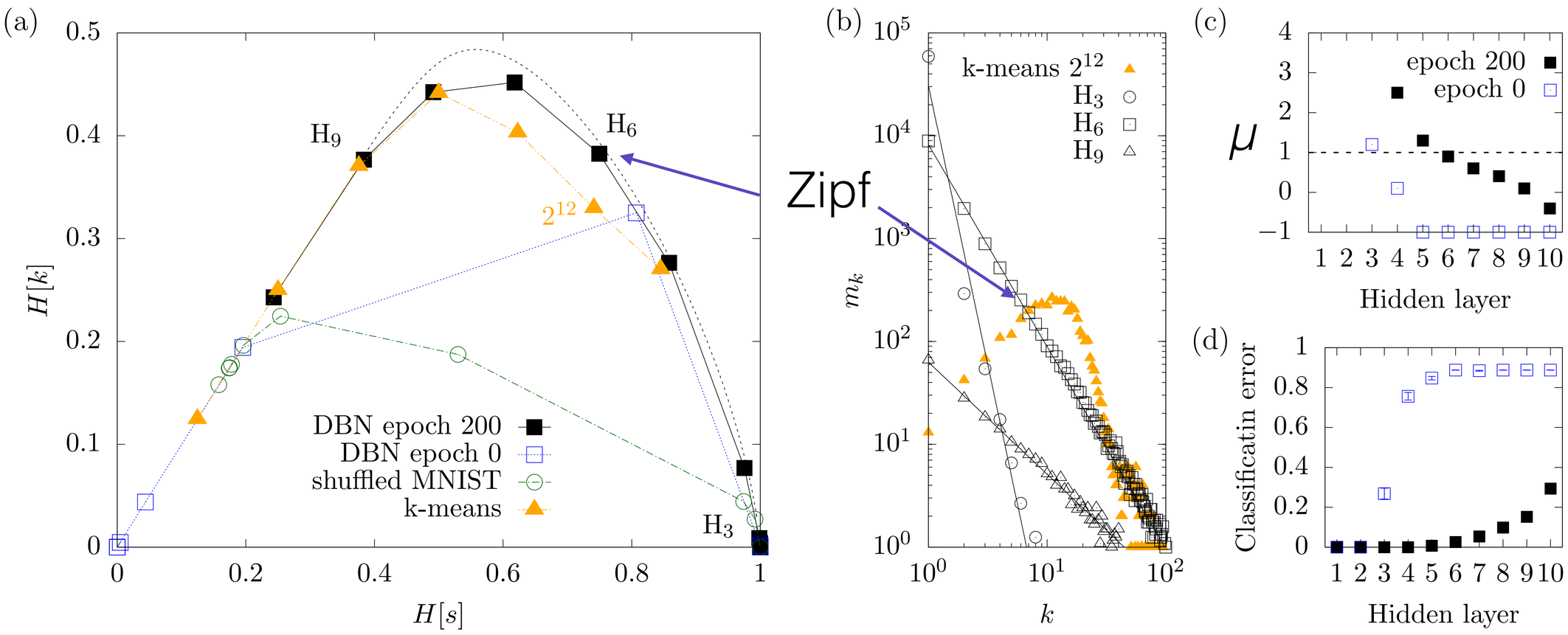
	s	k
V	1010	1
	1011	1
	1100	1
H_1	010	2
	110	1
H_2	01	3

	k	m_k	km_k
V	1	3	3
H_1	1	1	1
	2	1	2
H_2	3	1	3



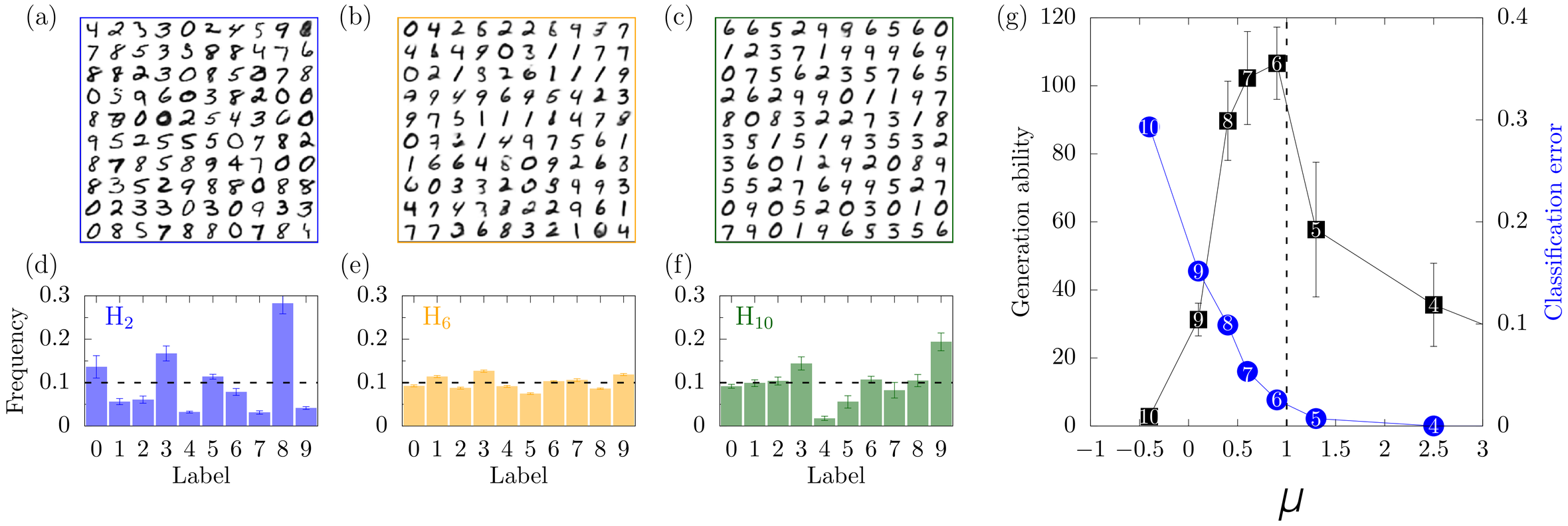
(Juyong Song, Junghyo Jo APCTP, Pohang)

Maximally informative representations in deep layers (MNIST)



(Juyong Song, Junghyo Jo APCTP, Pohang)

Zipf = optimal generalisation



(Juyong Song, Junghyo Jo APCTP, Pohang)

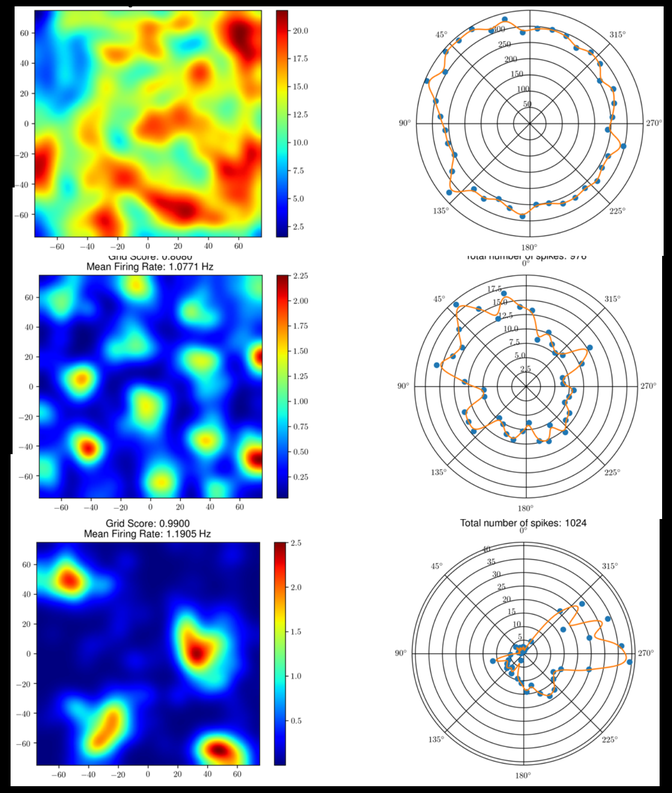
Searching for relevant neurons in the brain

e.g. recording neurons responsible for spatial navigation

neurons

```
0001000011110000100001000101010101000001
1111110100111110111010101111010101111111
1010001000101010010100100100000100001001
1001000100001000000001000010000000010000
1111010111111111000111111110111111110000
0000000111110000000000011111111000010001 1
1010001000101010110100100100000100001000
1010111100110010010111111000001001111001
```

time

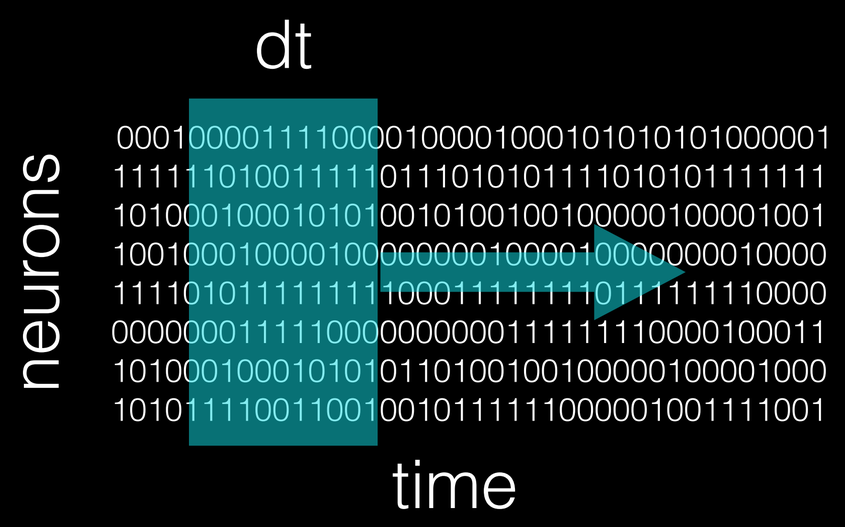


How to find relevant neurons when the co-variate is unknown?

(+ Ryan Cubero, Yasser Roudi, NTNU/SISSA)

Multi-scale Relevance

e.g. recording neurons responsible for spatial navigation



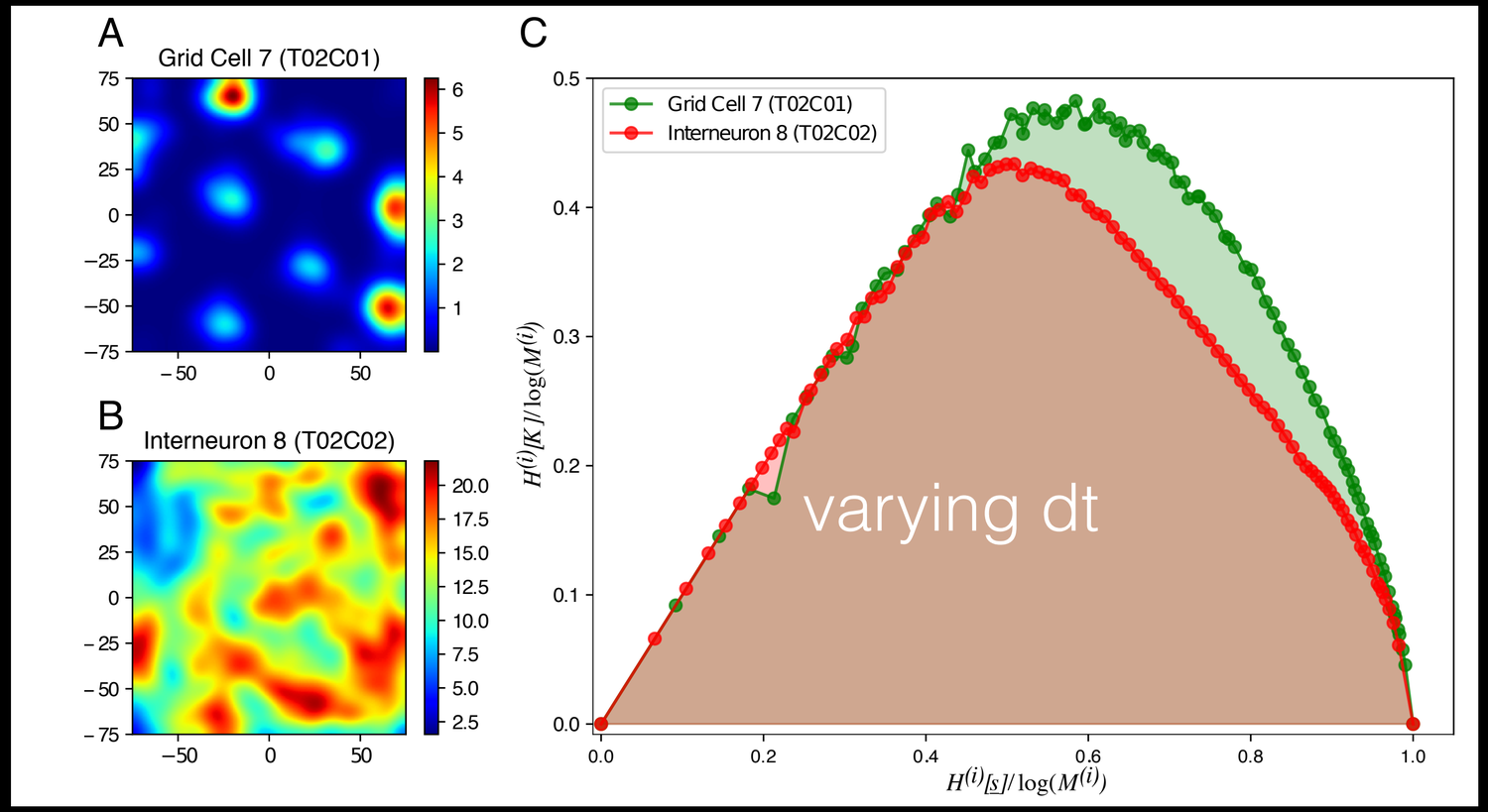
(H[s], H[k])

dt sets resolution H[s]
H[k] ~ variation in
dynamical states

(+ Ryan Cubero, Yasser Roudi, NTNU/SISSA)

Multi-scale Relevance

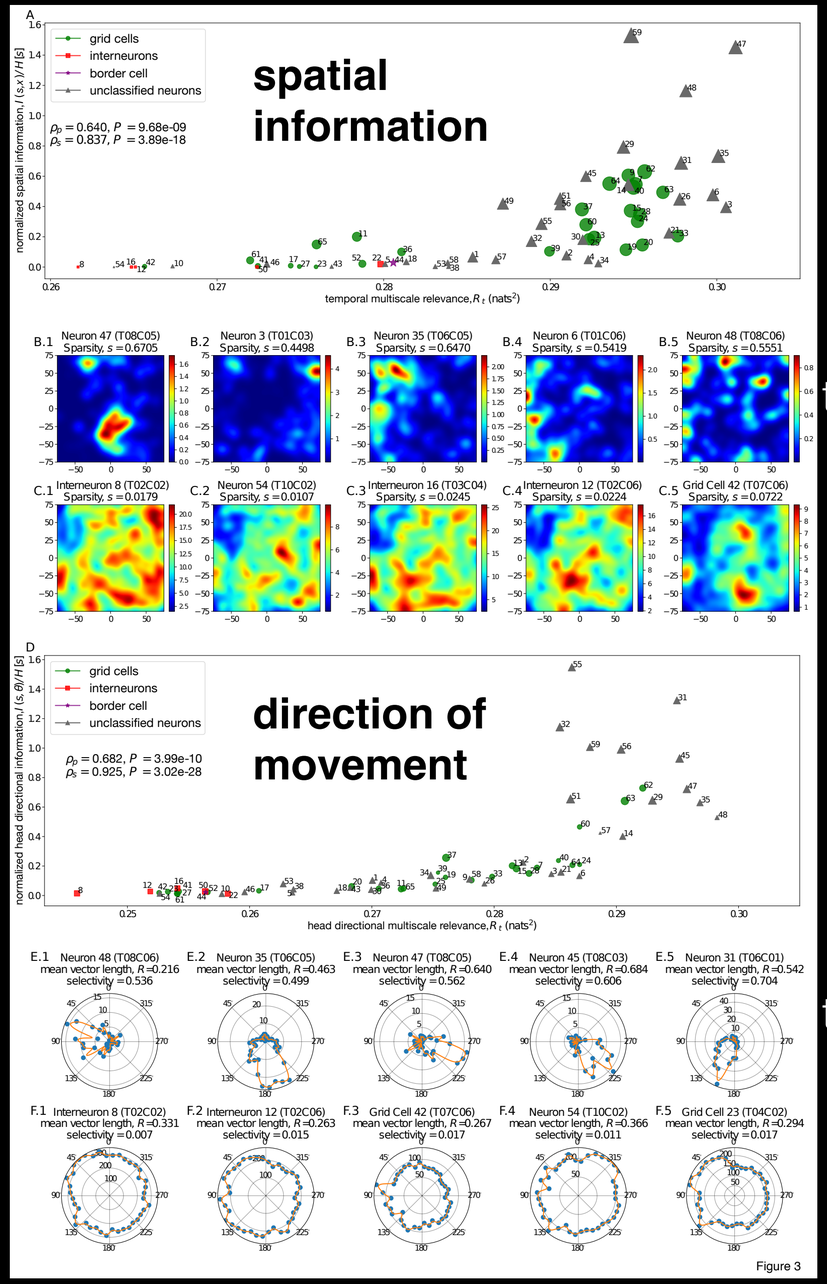
e.g. recording neurons responsible for spatial navigation



Multi-scale Relevance = Area under the curve

(+ Ryan Cubero, Yasser Roudi, NTNU/SISSA)

65 cells in medial Enthorinal Cortex



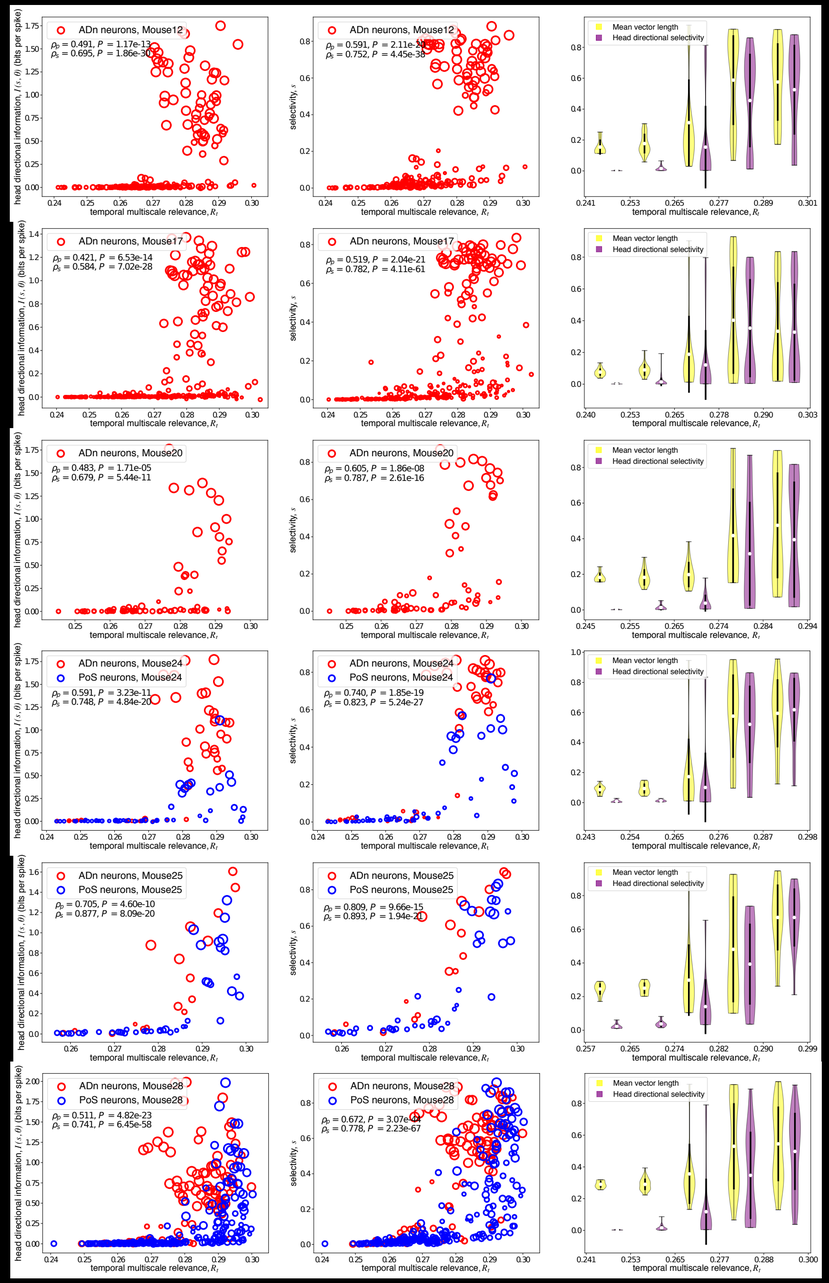
top 5

bottom 5

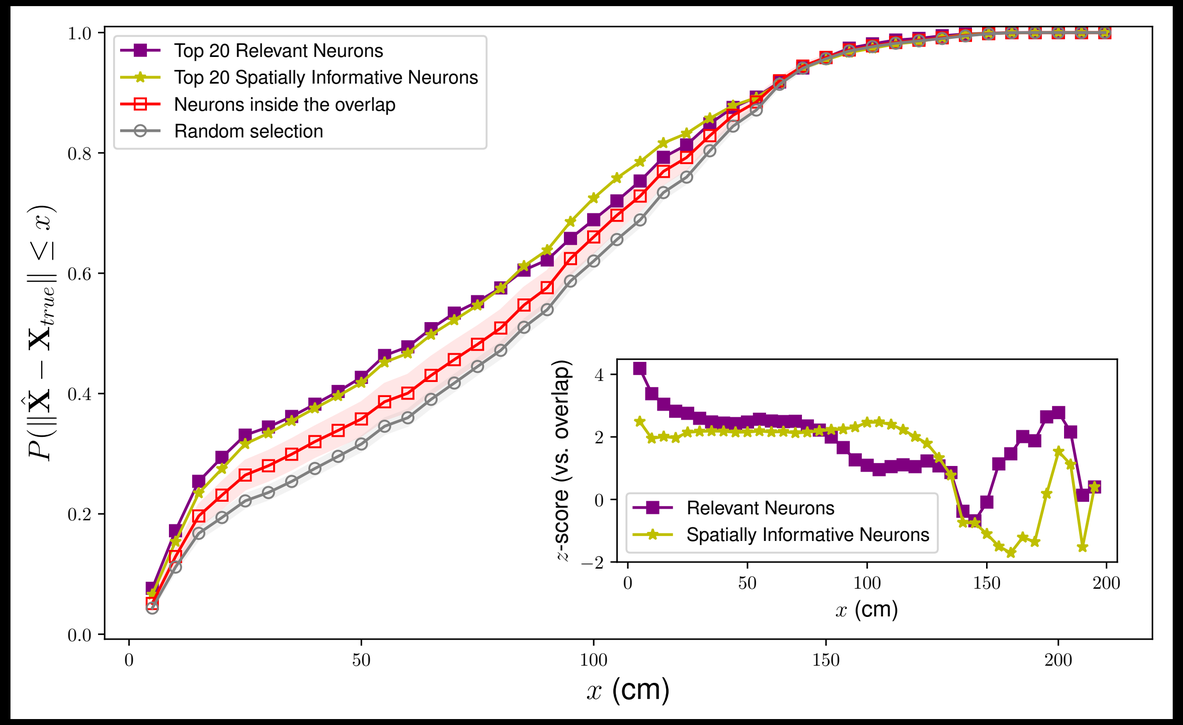
top 5

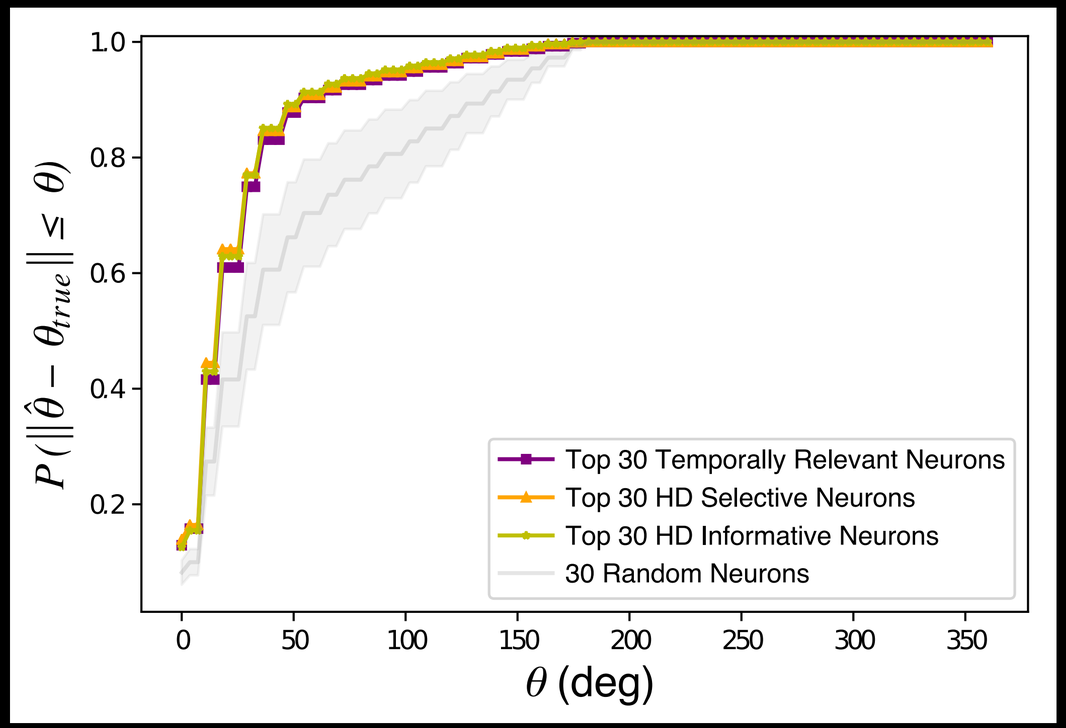
bottom 5

746 cells in AD nucleus and PoS



Neurons with high Multi-scale Relevance contain as much information on position/direction as those whose neural activity has the highest mutual information with position/direction



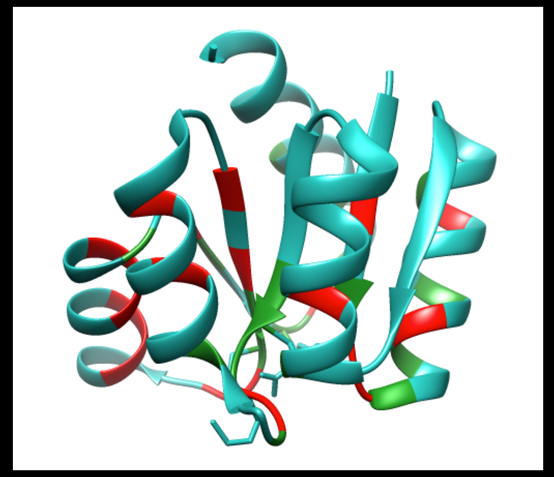


Identifying relevant positions in proteins
Critical Variable Selection

(S. Grigolon *et al*, Mol. Biosys, 2016)

```
-CTMPMEAGSCDGKLARWHFARDDNKCMPFYYTGCGGNHNQFISLDQCEEQC-
-CTNTLAQGEGPLSVTRYYFNAQSRTCDEFMFRGLKGNSNNFNSLAECEKAC-
-CTQPKDSGVCSGSQRSFYFDTRMKVCQPFLYSGCGGNENRFSTSKECRDACQ
-CTQPRDNGSCSENGRAFYFDTRTKVCQPFLYSGCGGNDNRFATSKECRDSCQ
-CTQTLAQGEGPLSVARFYFNAQSRTCDEFMFRGLKGNSNNFKSQEDCEKAC-
-CTQTLAQGEGPLSVARFYFNAQSRTCDEFMFRGLKGNSNNFKSQEDCEKAC-
-CTQTLAQGEGPLSVARFYFNAQSRTCDEFMFRGLKGNSNNFKSQEDCEKAC-
-CTSPPVTGPCRAGFKRYNYNTRTKQCEPFKYGGCKGNGNRYKSEQDCLDAC-
-CTVLPSEGYCKKRYFRFLYDSNTKTCQLFWYRGCGGTENNFPTYYSCLDRC-
-CTVLPSEGYCRKKYFRFLYDSNTKTCQLFWYRGCGGTENNFPTYYACLDRC-
-CTVQPTNGLCVPSTLGIYFDVETQHCR---FLGC-GNKRLFASLEDCEKIC-
-CTVQPTNGLCVPSTLGIYFDVETQHCR---FLGC-GNKRLFASLEDCEKIC-
-CVAKPDAGPCRAAFPAFFYDPDTNSCQPFIYGGCRGNGNRYNSREECLSRC-
-CVAPLDKCP--GNVIIYYYN-RTSGCQQMHRGNCSDN-GNYPTLQECQEYCL
-CVDLPDTGLCKESIPRWYYNPFSEHCARFTYGGCYGNKNNFEEEQQCLESCR
-CVDLPDTGLCKESIPRWYYNPFSEHCARFTYGGCYGNKNNFEEEQQCLESCR
…
```

?

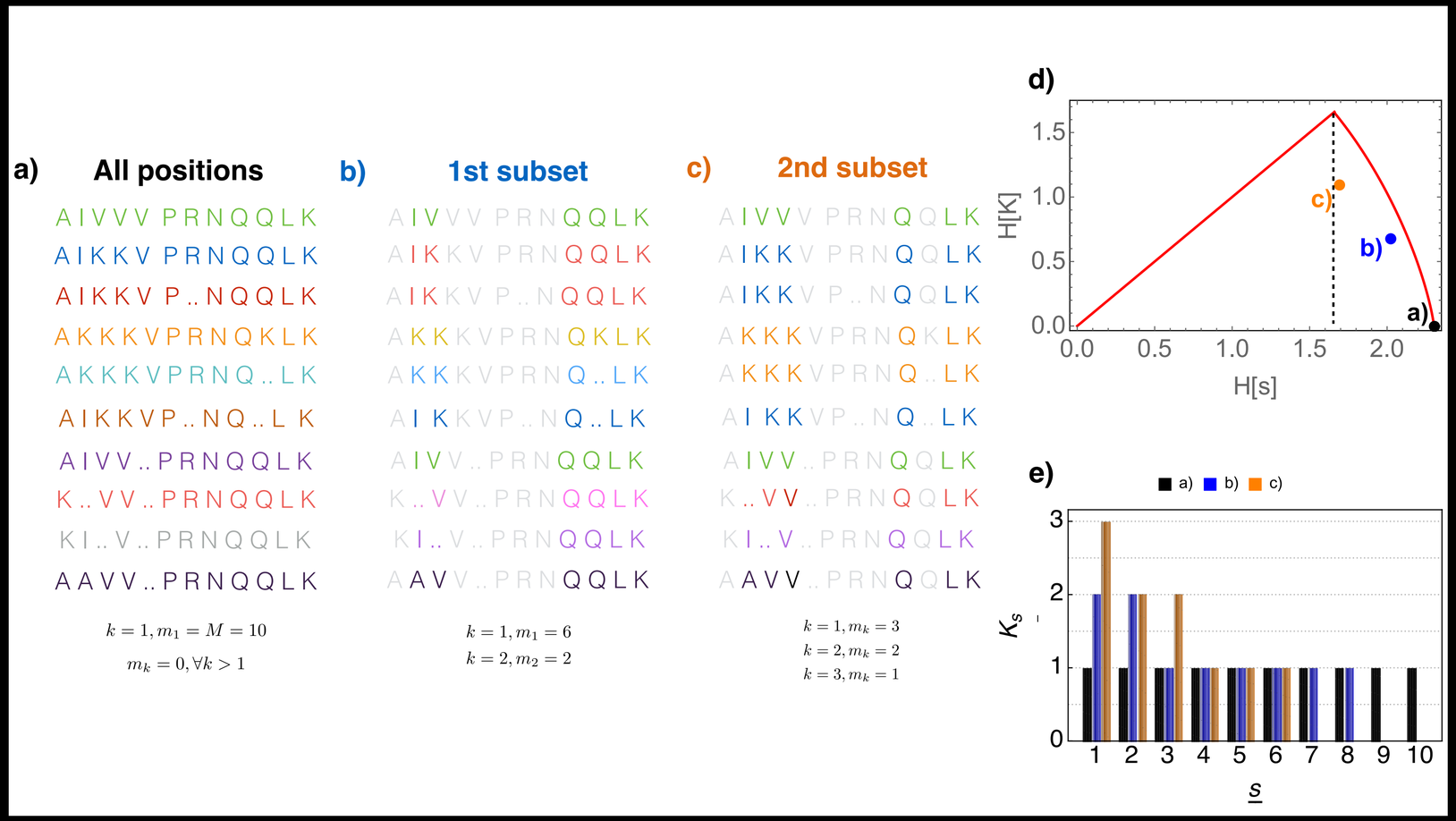


$$s \in \{A, B, \ldots, W\}^L$$

$L \sim 10^2$, $M \sim 10^4 << 20^L$

Identifying relevant positions in proteins
Critical Variable Selection

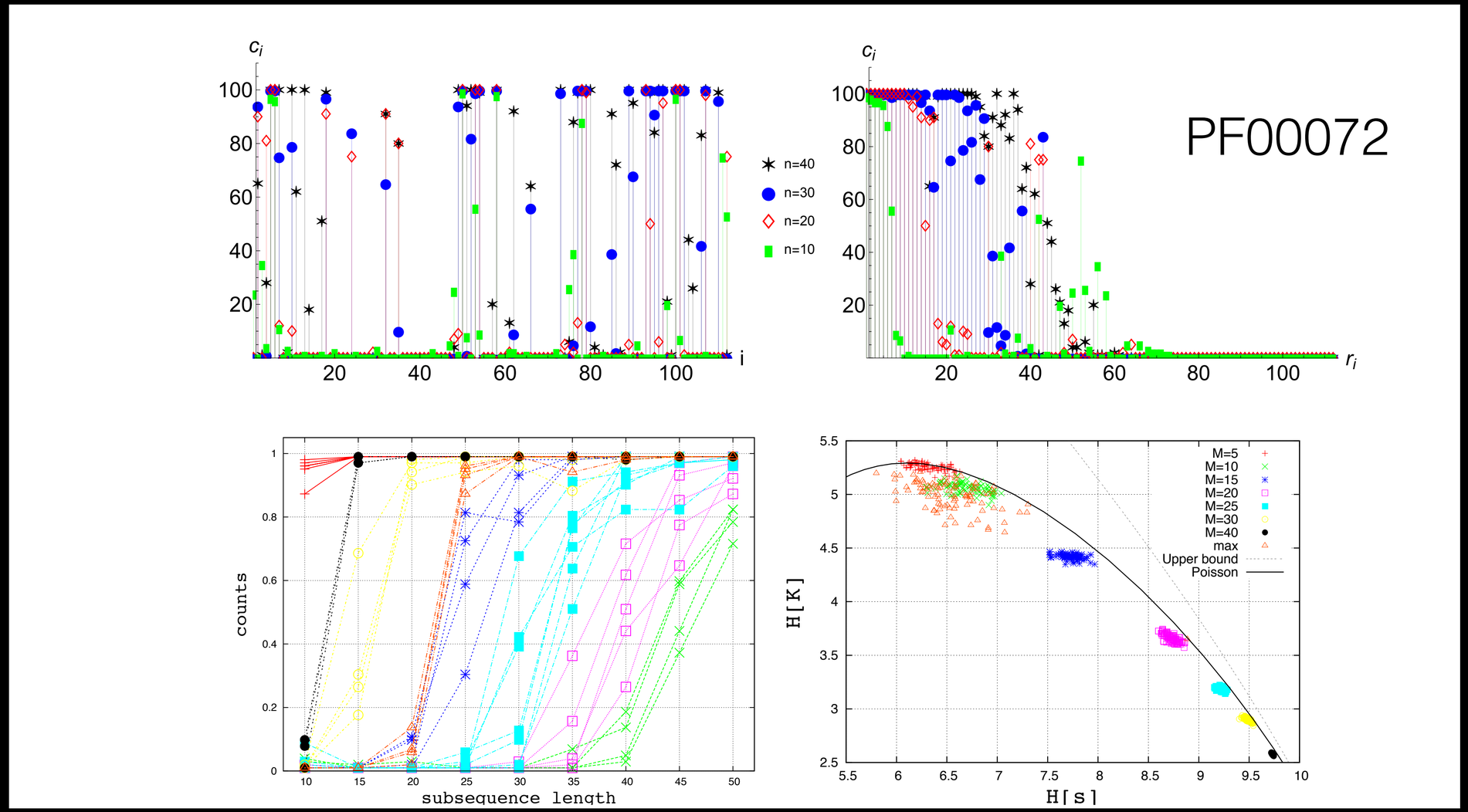
(S. Grigolon *et al*, Mol. Biosys, 2016)



repeat many times and count how many times position x is selected in max H[k]

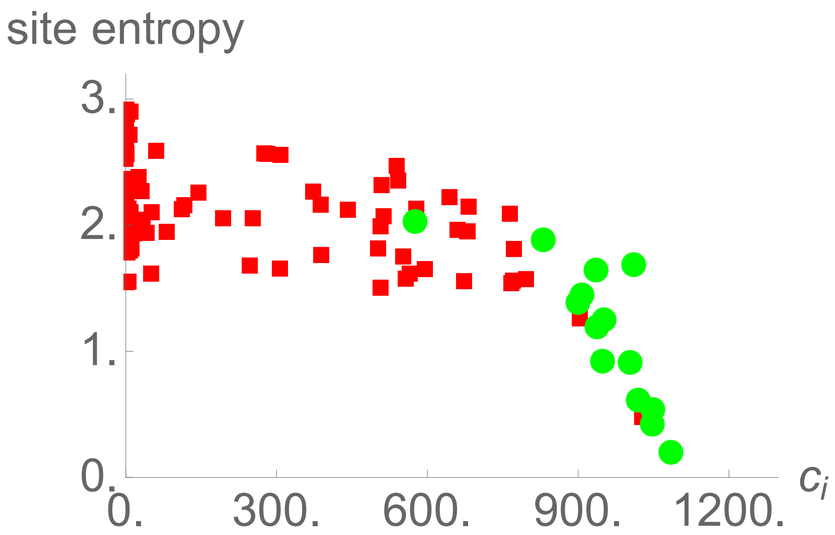
Sharp separation between relevant and irrelevant sites

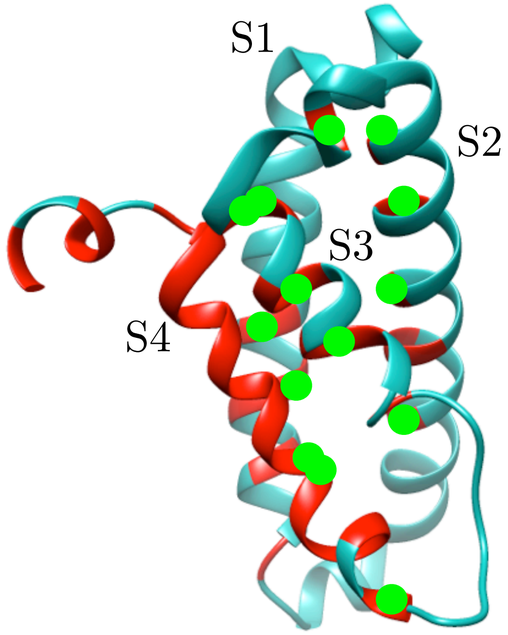
(+S. Grigolon LPTMS/NETADIS)



Conserved and biologically relevant sites

(+S. Grigolon LPTMS/NETADIS)



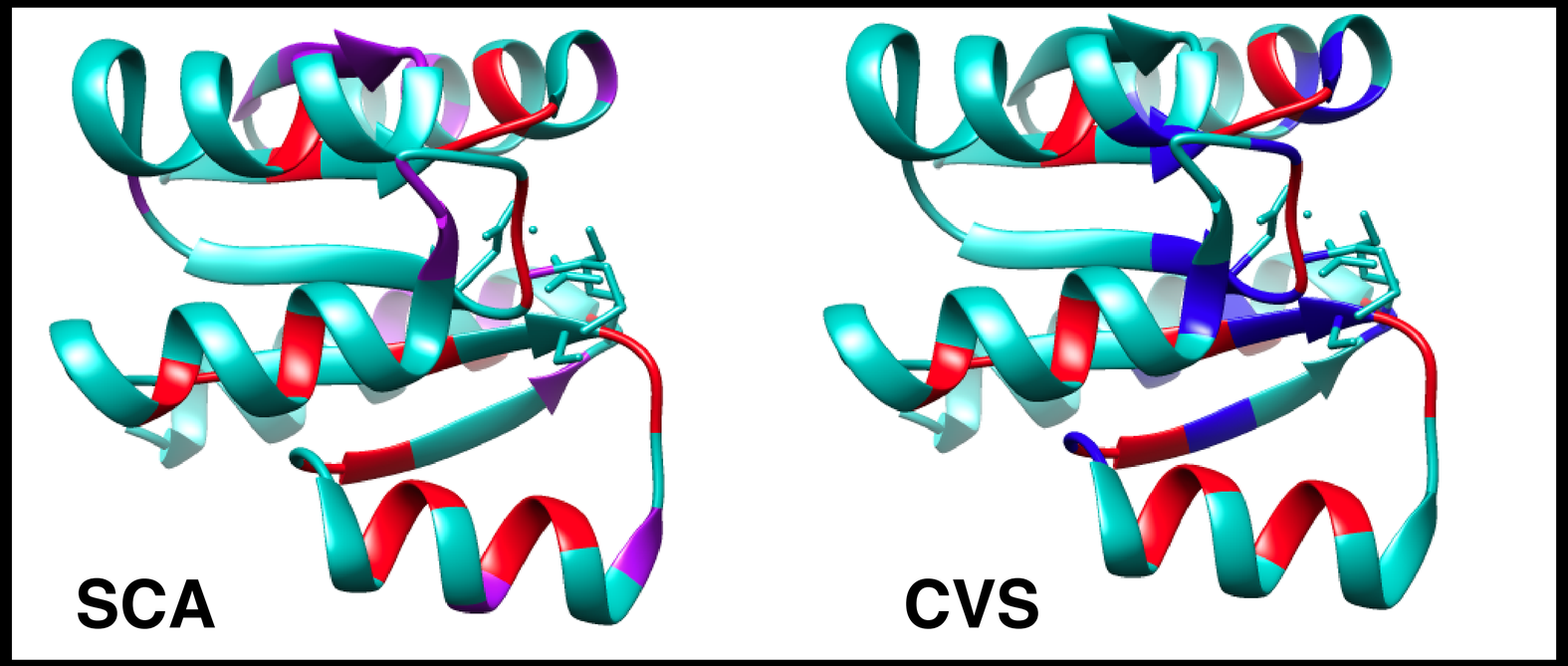


Voltage Sensor Domain PF000520 M=6651 (M.L. Klein et al. 2014)

Orthogonal to correlation based methods (SCA)

(+S. Grigolon LPTMS/NETADIS)

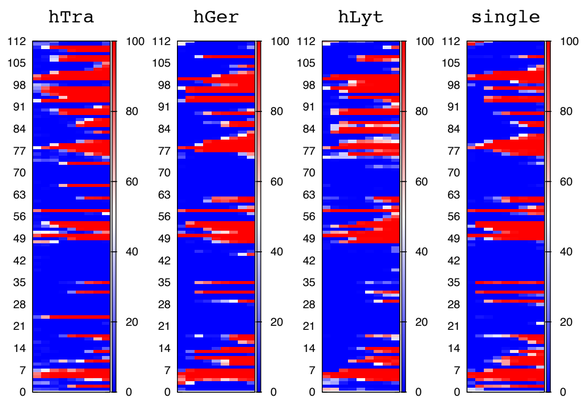
- No overlap for the 18 most relevant sites
- Maximal overlap (51%) for 41 sites (random 33%)

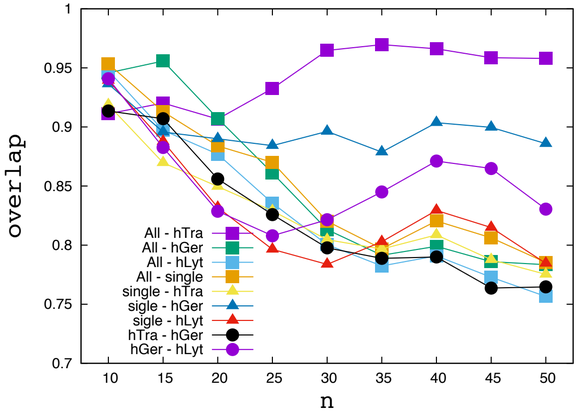


SCA=Statistical Coupling Analysis, Lockless & Ranganathan, Science (1999)

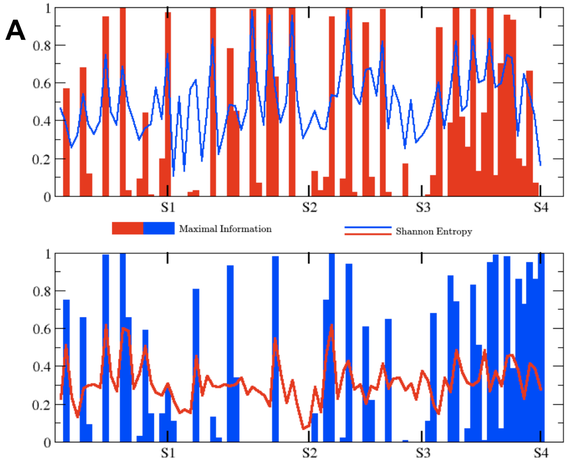
Subfamilies and families with similar structures

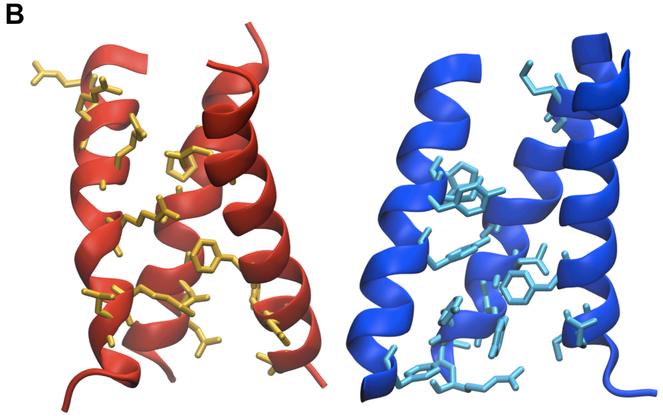
(+S. Grigolon LPTMS/NETADIS)





Response regulator domains



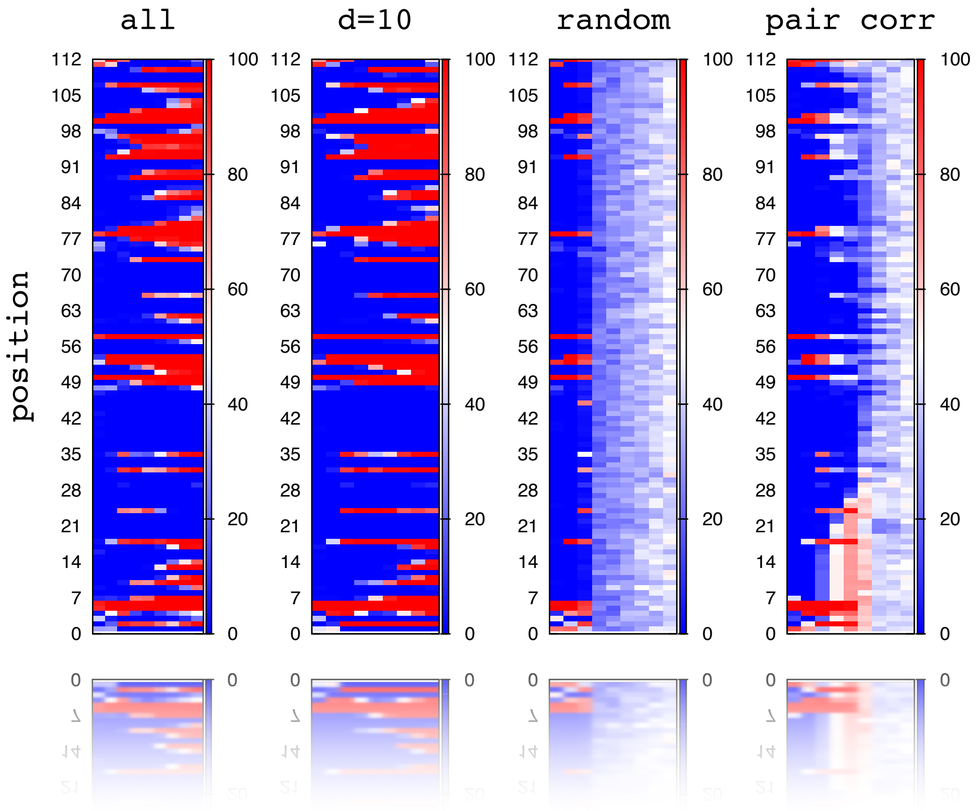


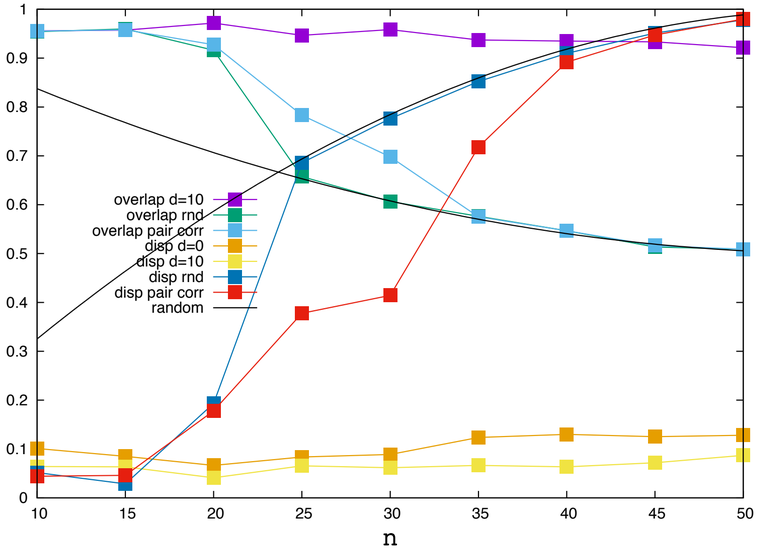
Voltage gated cation channels and TRP channels

(M.L. Klein et al. 2014)

Beyond pairwise correlation

(+S. Grigolon LPTMS/NETADIS)





Conclusions

- H[k] as a model free measure of intrinsic relevance, very easy to compute

- Maximally informative samples (maximal relevance at fixed resolution) typically exhibit power laws frequency distributions

- Applications:
  - Maximally informative samples and criticality in Minimum Description Length
  - Understanding deep learning
  - Featureless selection of relevant neurons in spatial navigation
  - Prediction of relevant positions in proteins

- Extensions:
  - Inference beyond correlations (higher order interactions)
  - Heuristics for Bayesian model selection
  - Relevance for dynamical data
  - Continuous variables