Exploring chemical (and model) space with random matrices and spin glasses

Alpha Lee

Department of Physics, University of Cambridge
aal44@cam.ac.uk

Two fairy tales on entropy, random matrices and spin-glasses

- **Predicting the biological activity of small molecules**
- Understanding why deep neural networks “works”

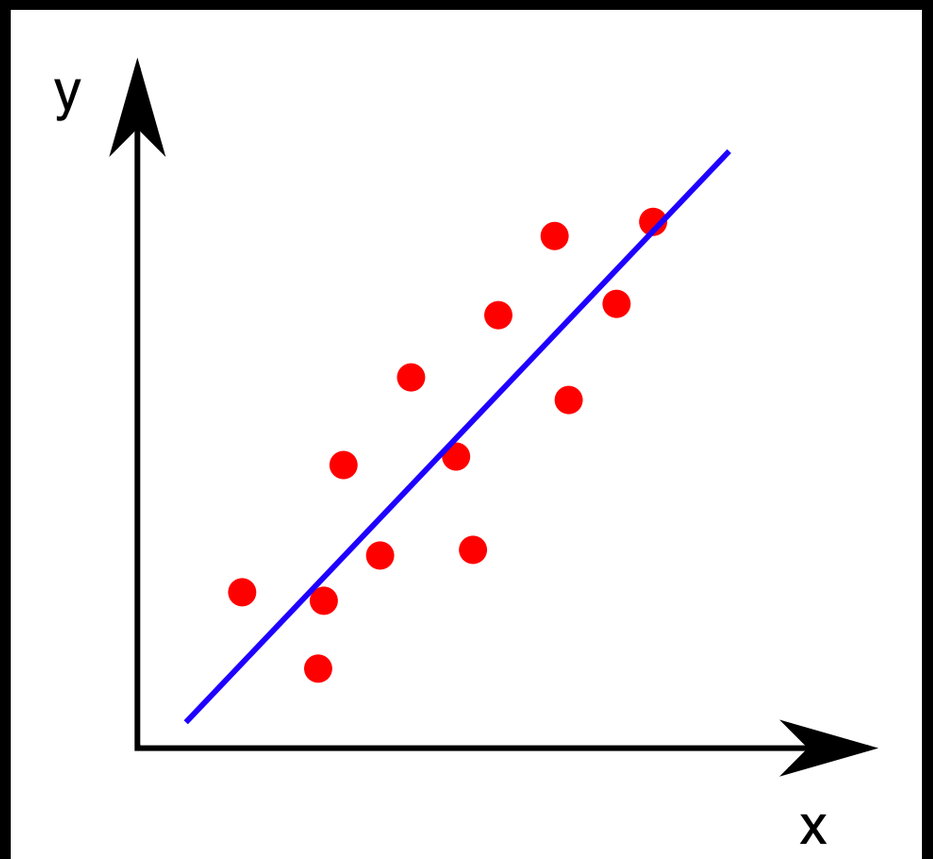
Direct interactions and indirect correlation

https://allabouthealthychoices.wordpress.com/2016/03/17/associationcorrelation-vs-causation/

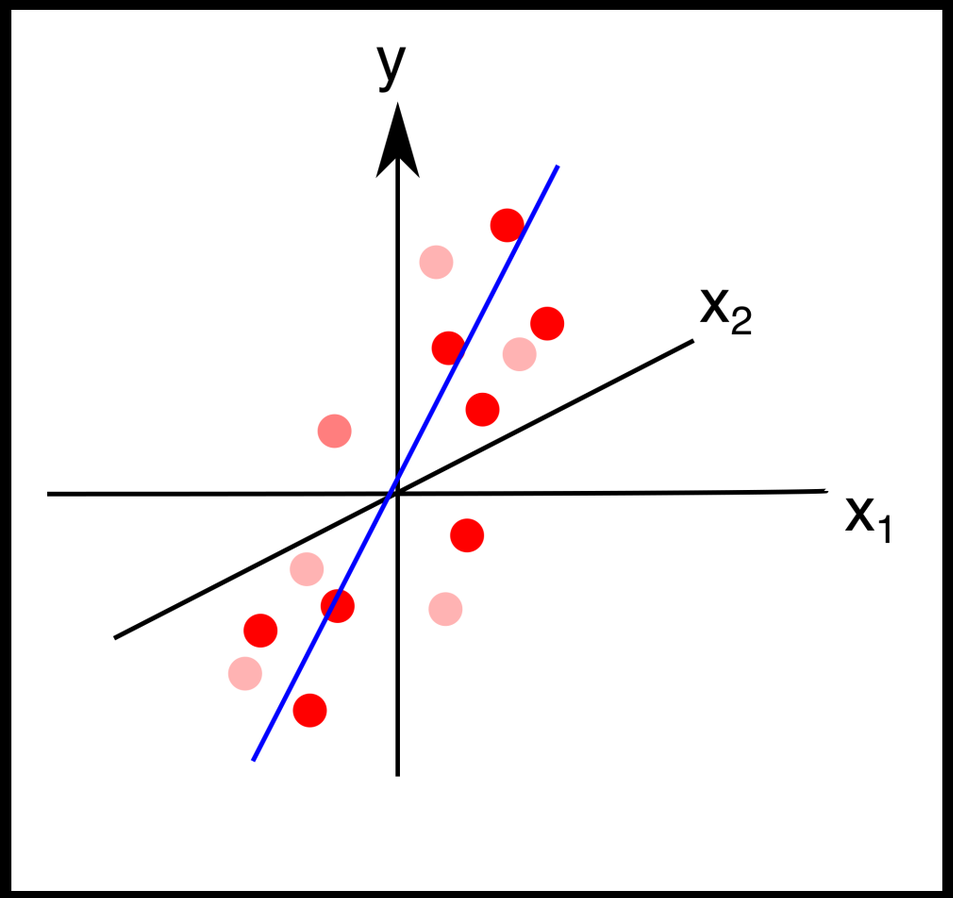
Chemical space is high dimensional

$p =$ number of variables

$n =$ number of samples

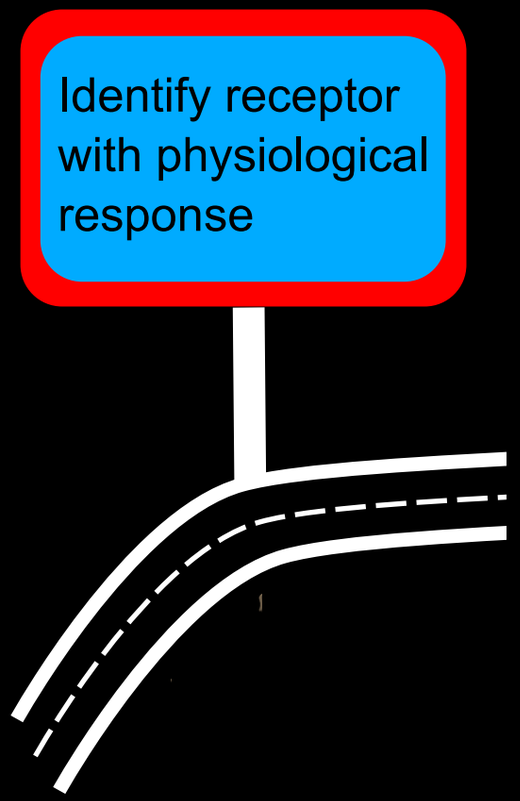


$p = O(1)$
$p/n \to 0$

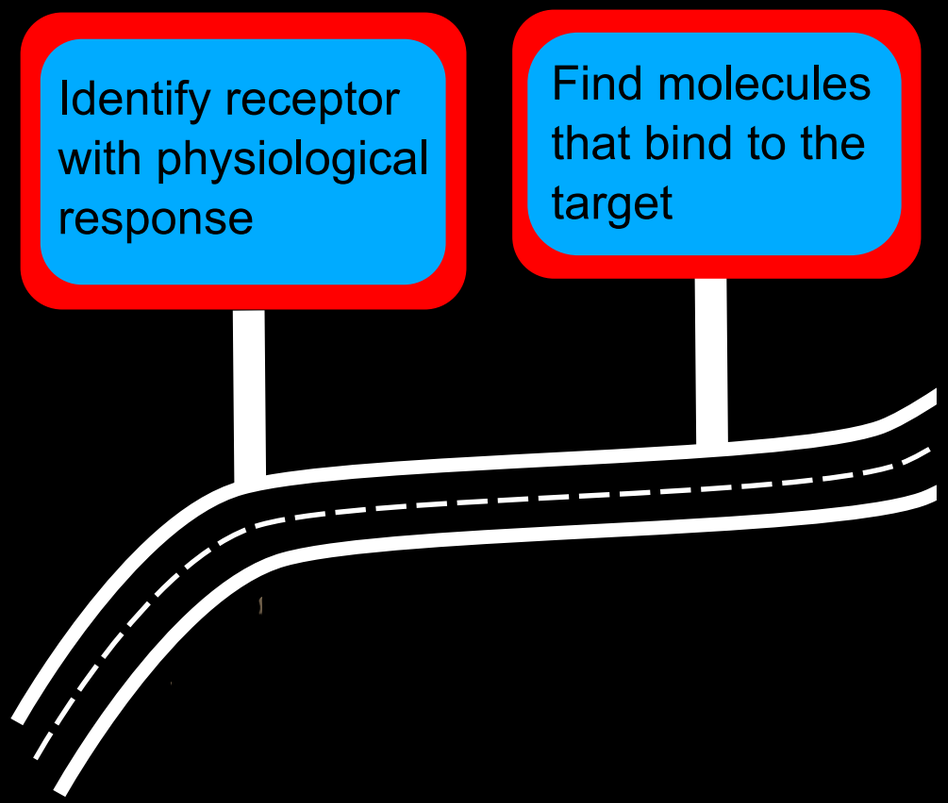


$n \to \infty$
$p/n = O(1)$

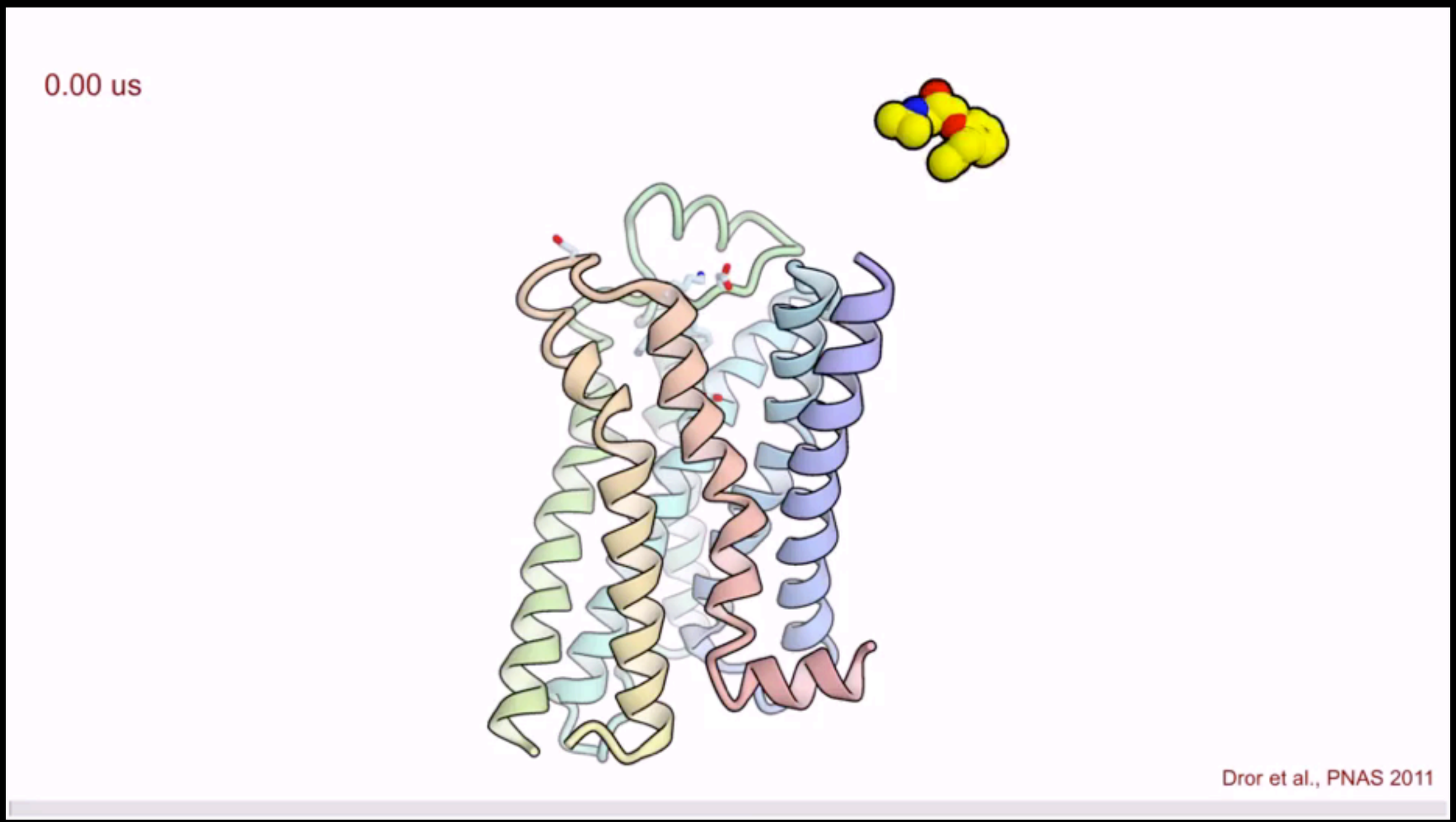
Stages in drug discovery



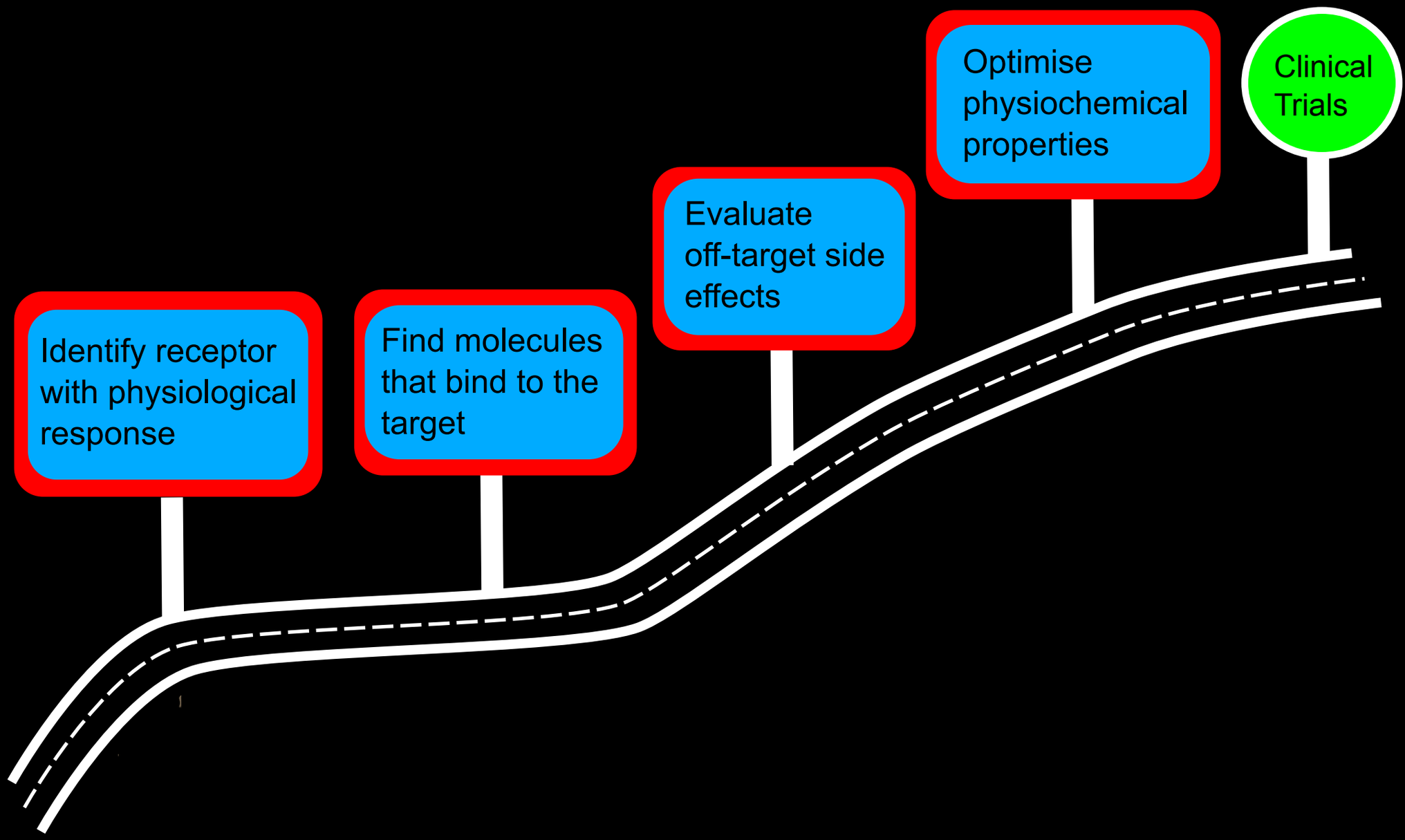
Stages in drug discovery



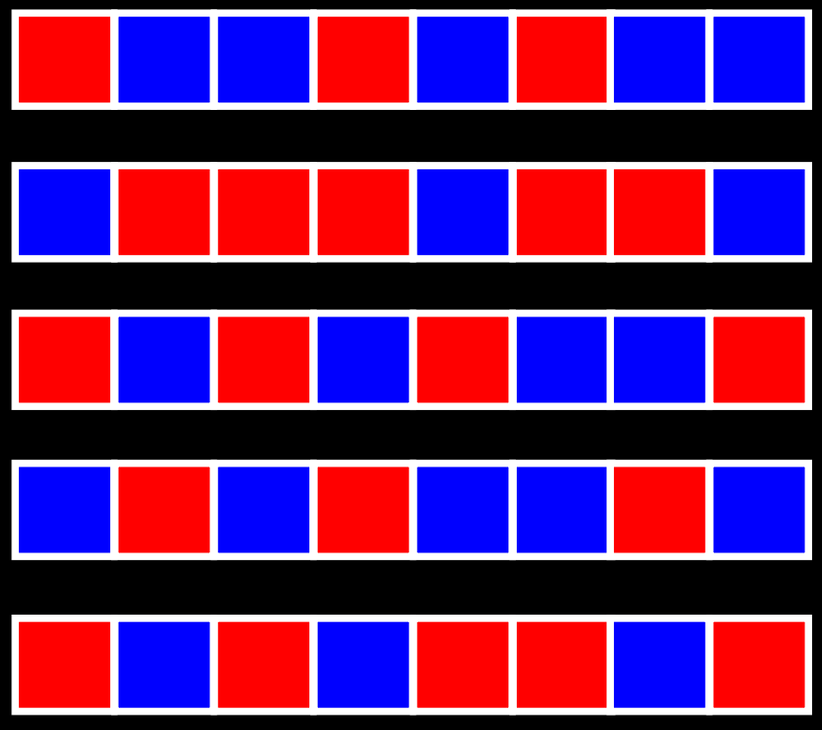
Physics-based methods are computationally expensive



Stages in drug discovery



Random matrix theory and Ising model

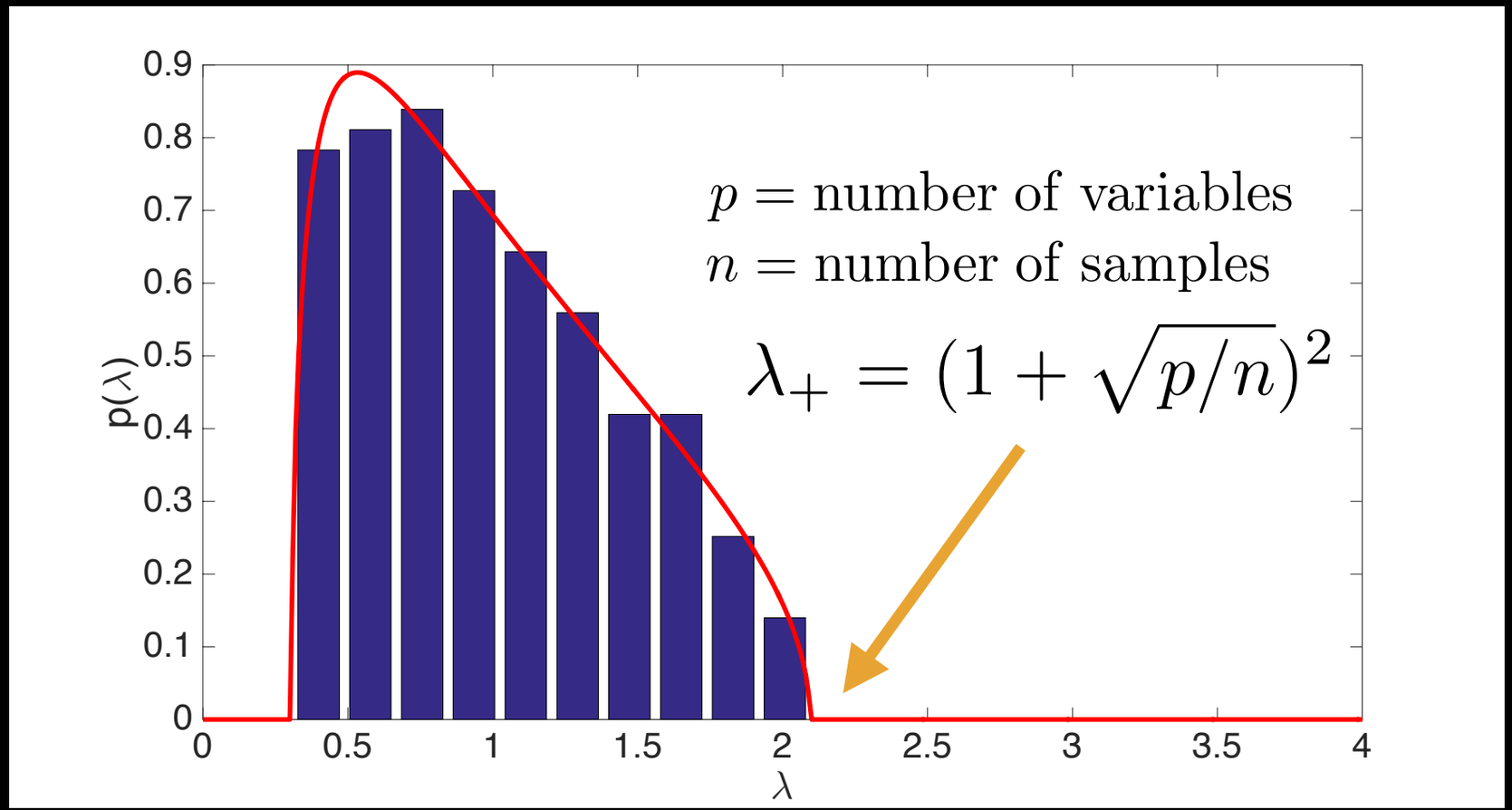


$$C_{ij} = \frac{\langle f_i f_j \rangle - \langle f_i \rangle \langle f_j \rangle}{\sigma_i \sigma_j}$$

E. T. Jaynes, *Physical Review*, 106, 620 (1957)
V. A. Machenko, L. A. Pastur, *Math. USSR Sb.*,1, 457 (1967)
S. Cocco, R. Monasson, V. Sessak,, *Phys. Rev. E*, 83, 051123 (

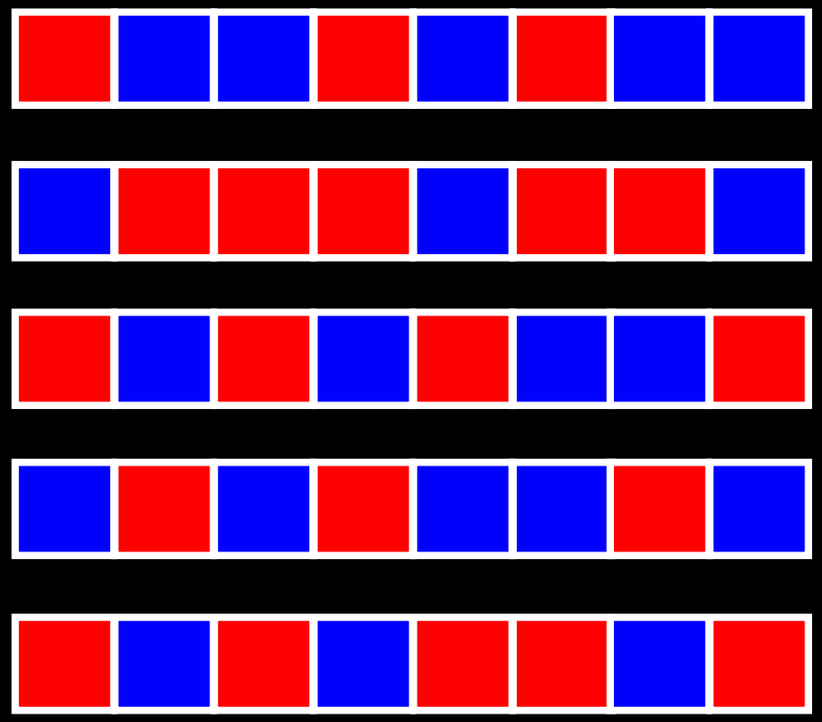
The null model

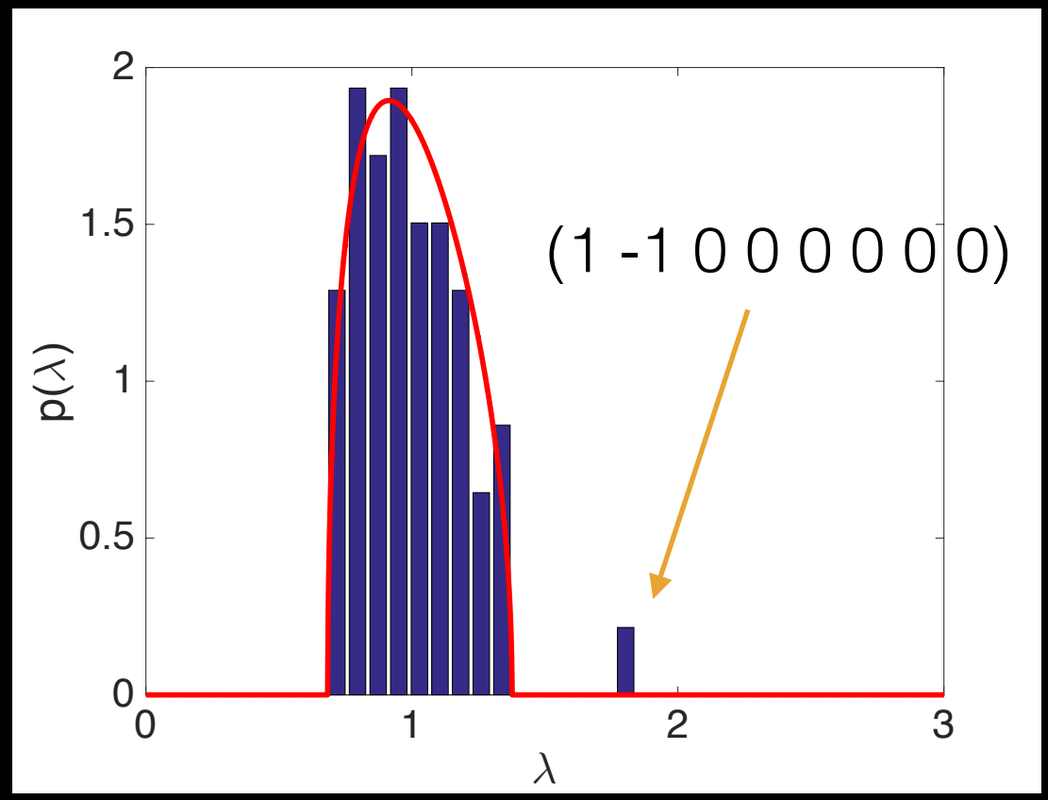
- Null model: the lattice sites are randomly coloured
- The eigenvalue distribution of the null model can be computed analytically



V. A. Machenko, L. A. Pastur, *Math. USSR Sb.*,1, 457 (1967)

Random matrix theory and Ising model

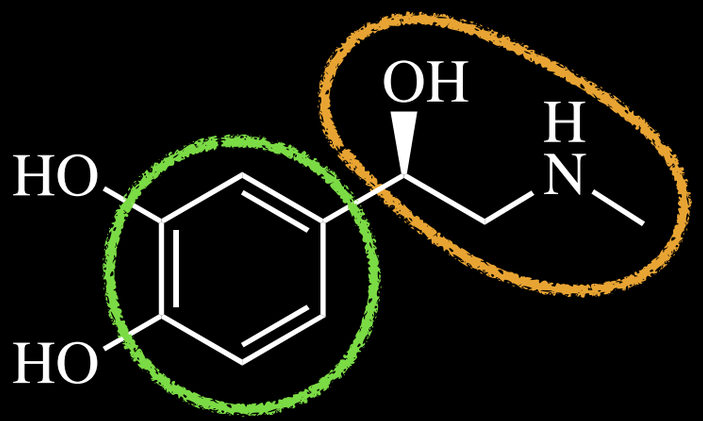




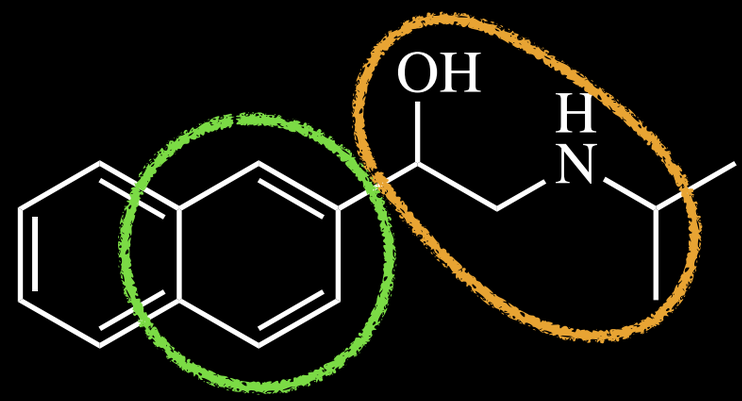
V. A. Machenko, L. A. Pastur, *Math. USSR Sb.*,1, 457 (1967)

Chemical similarity

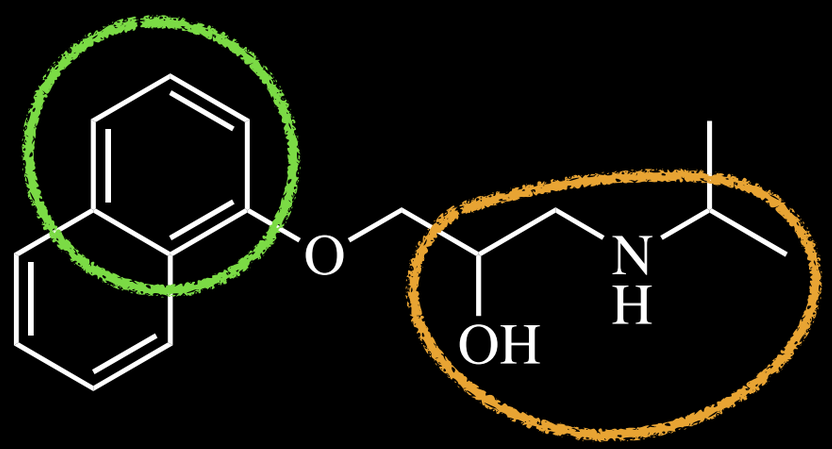
Suppose we want to design an ADRB1 antagonist



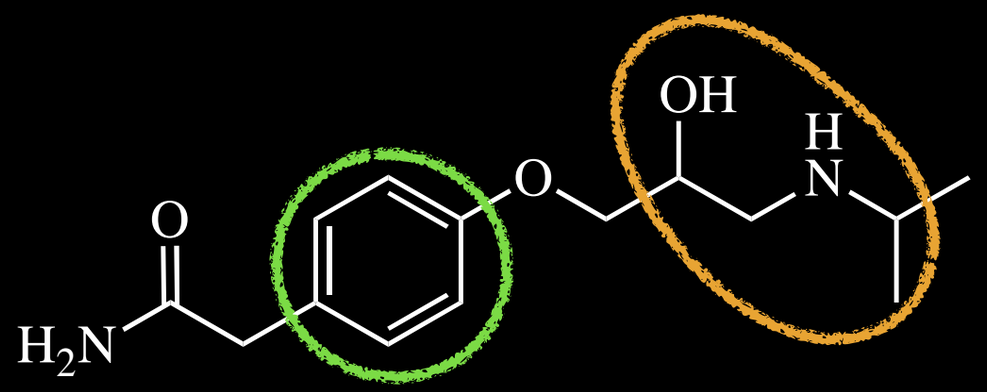
Adrenaline



Pronethalol



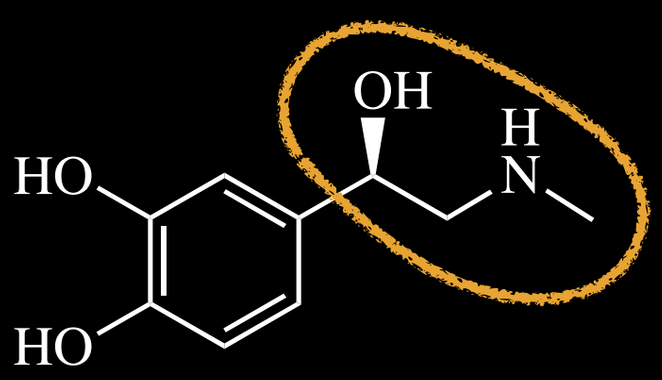
Propranolol

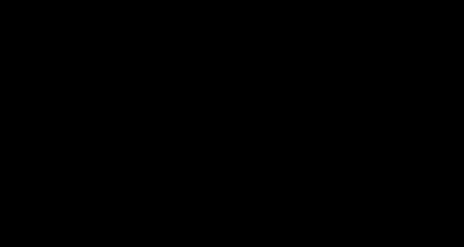


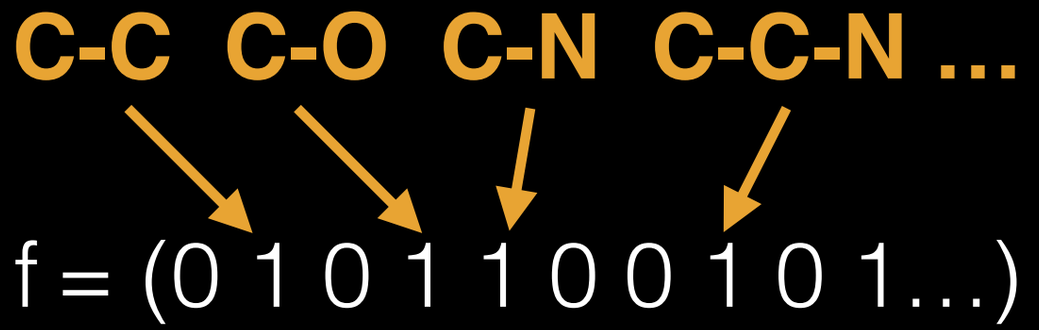
Atenolol

A. M. Johnson and G. M. Maggiora (Eds.), Concepts and Applications of Molecular Similarity, Wiley: New York, 1990.

How to extracting relevant chemical features?

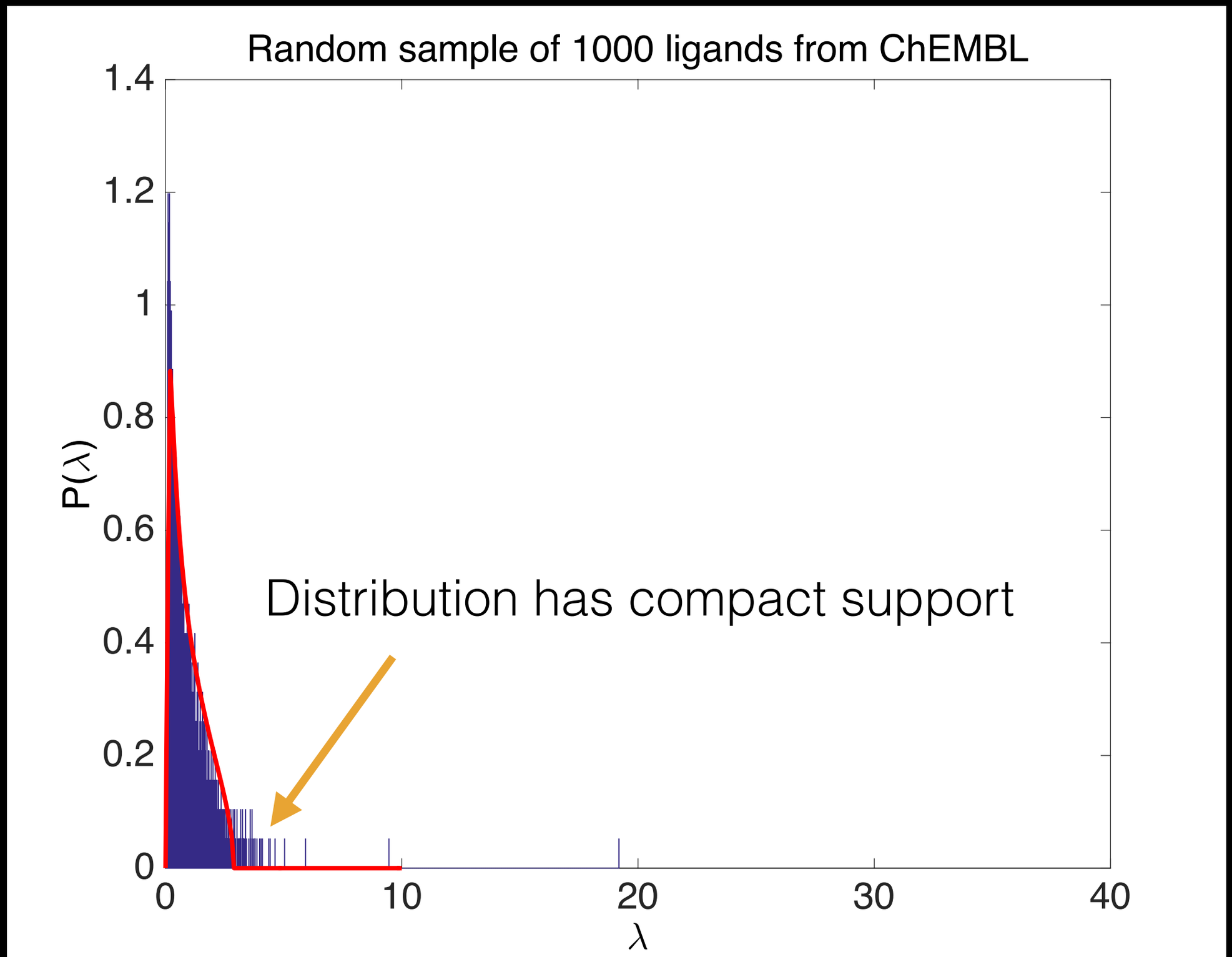




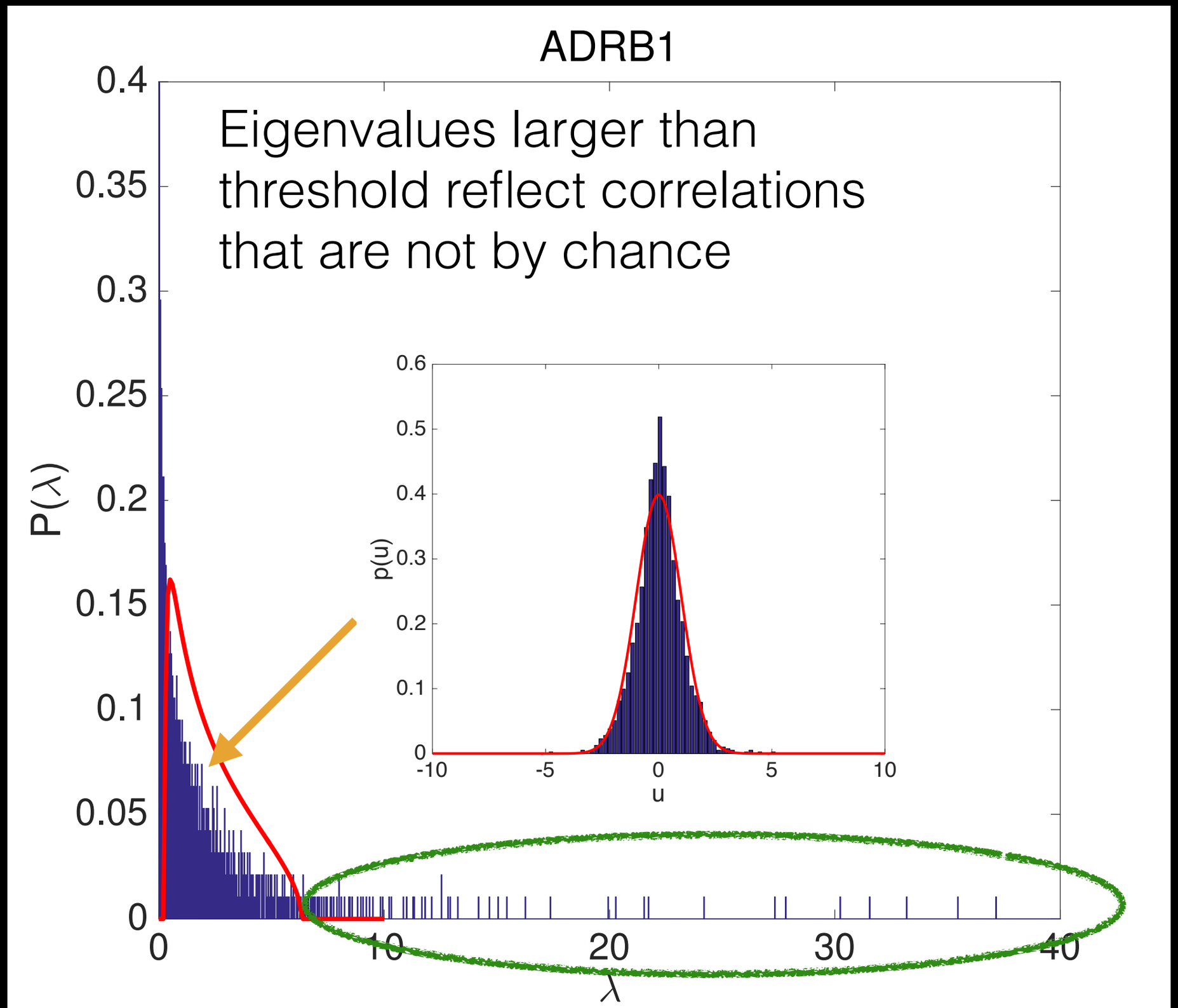


- Intuitively, there are only a few combinations chemical bonds paths (variables in the vector) that are important
- Many variables but often not many samples - data corrupted by finite sampling noise
- How do we get rid of the noise?

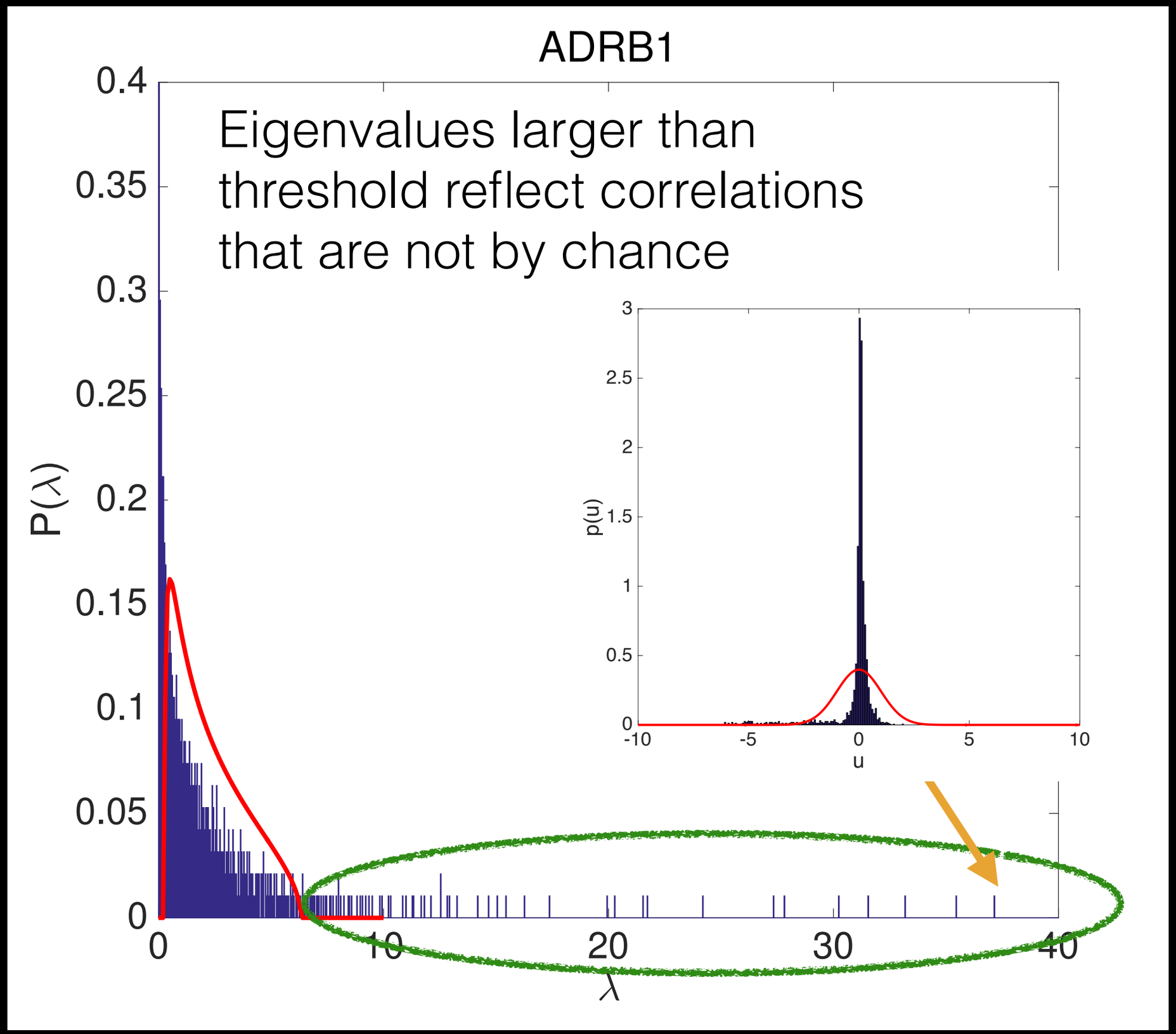
H. L. Morgan, *J. Chem. Doc.*, 5, 107 (1965)
Daylight Chemical Information Systems, Inc (since 1987)
D. Rogers and M. Hahn, *J. Chem. Inf. Model.*, 50, 742 (2010)



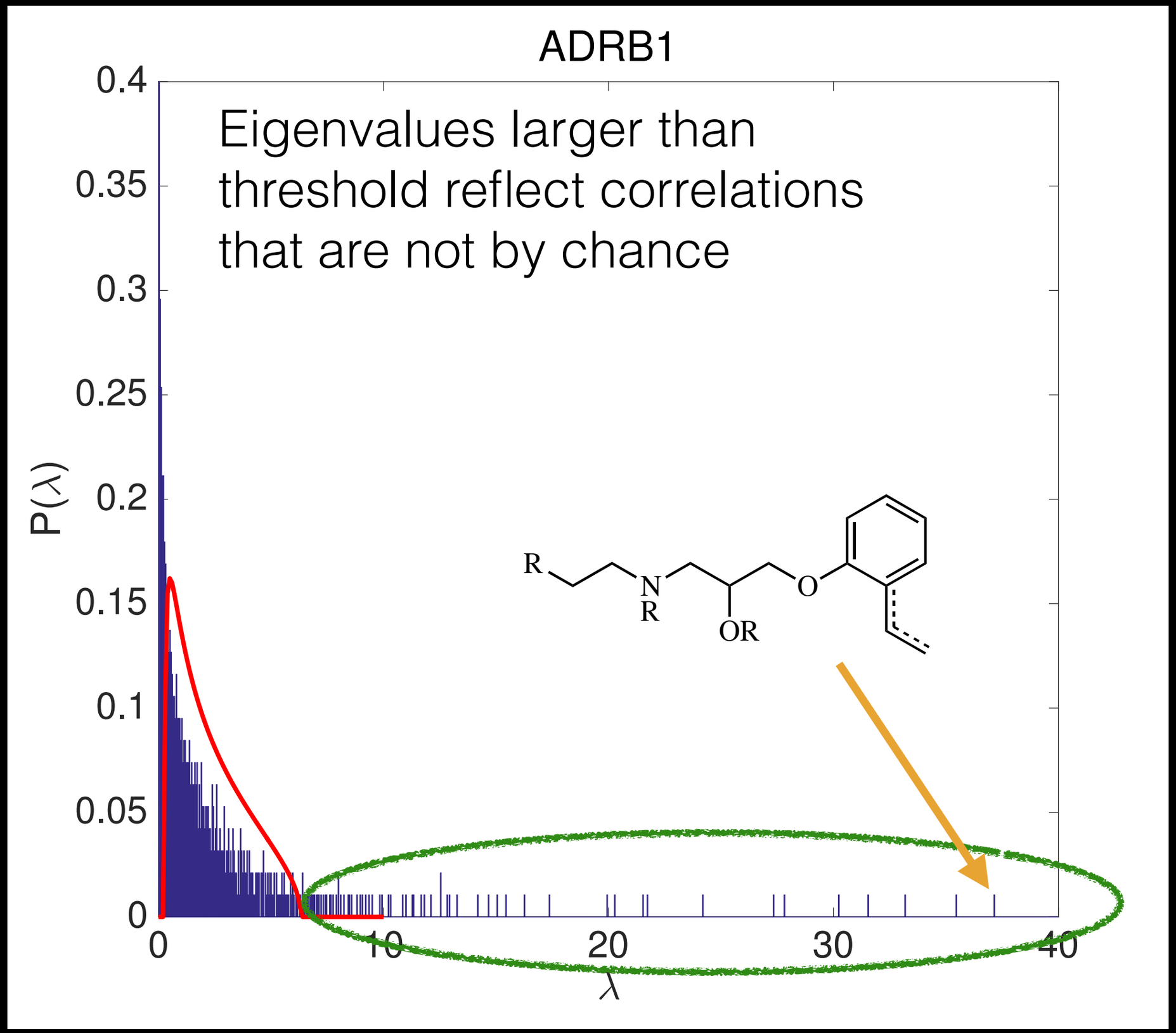
V. A. Machenko, L. A. Pastur, *Math. USSR Sb.*,1, 457 (1967)



c.f. M. Turk, A. Pentland, *J. Cognitive Neurosci.*, 3, 71 (1991)
L. Laloux et al., *Phys. Rev. Lett.*, 83, 1467 (1999)

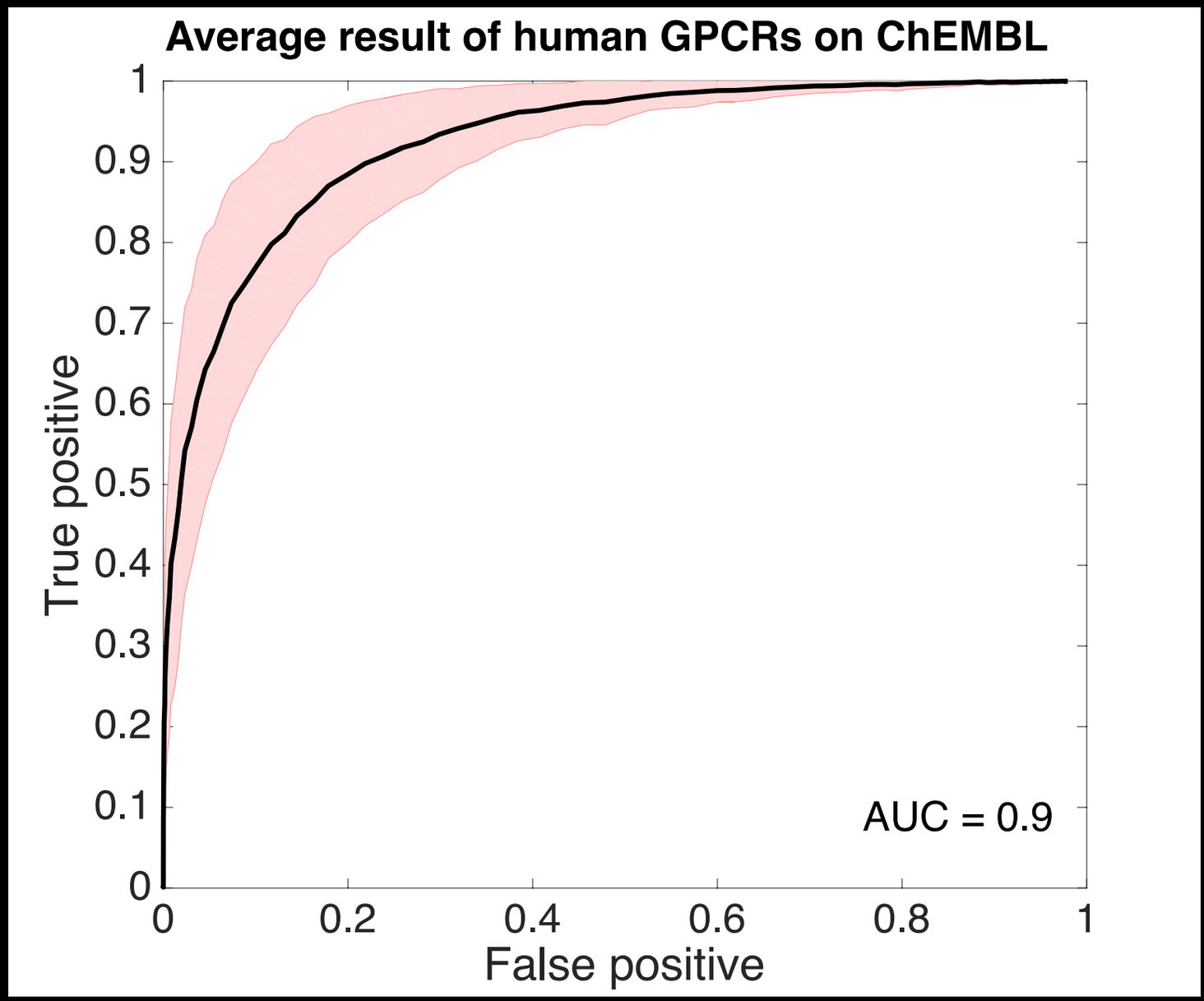


c.f. L. Laloux et al., *Phys. Rev. Lett.*, 83, 1467 (1999)



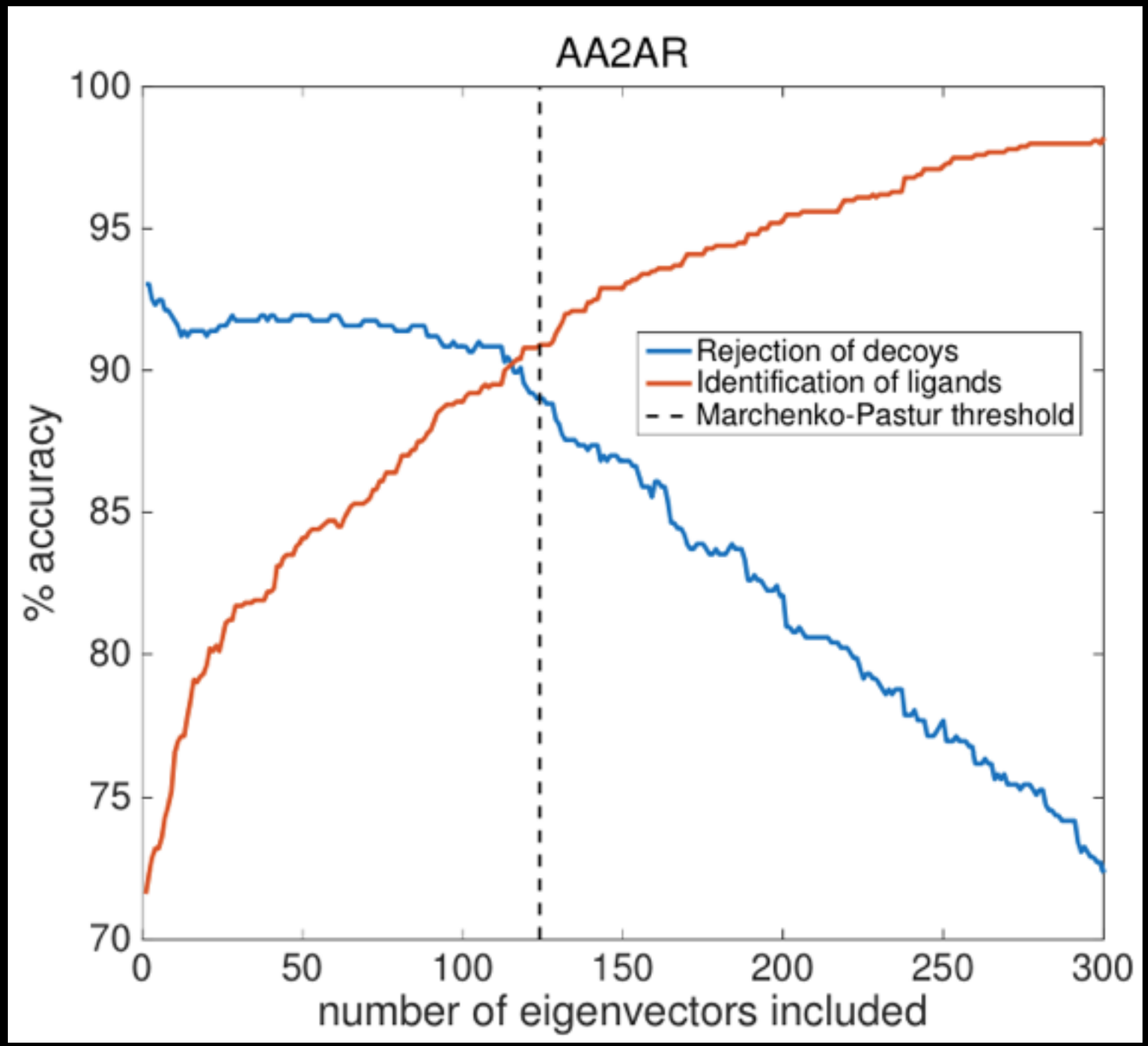
c.f. L. Laloux et al., *Phys. Rev. Lett.*, 83, 1467 (1999)

Performance of classification algorithm: GPCRs



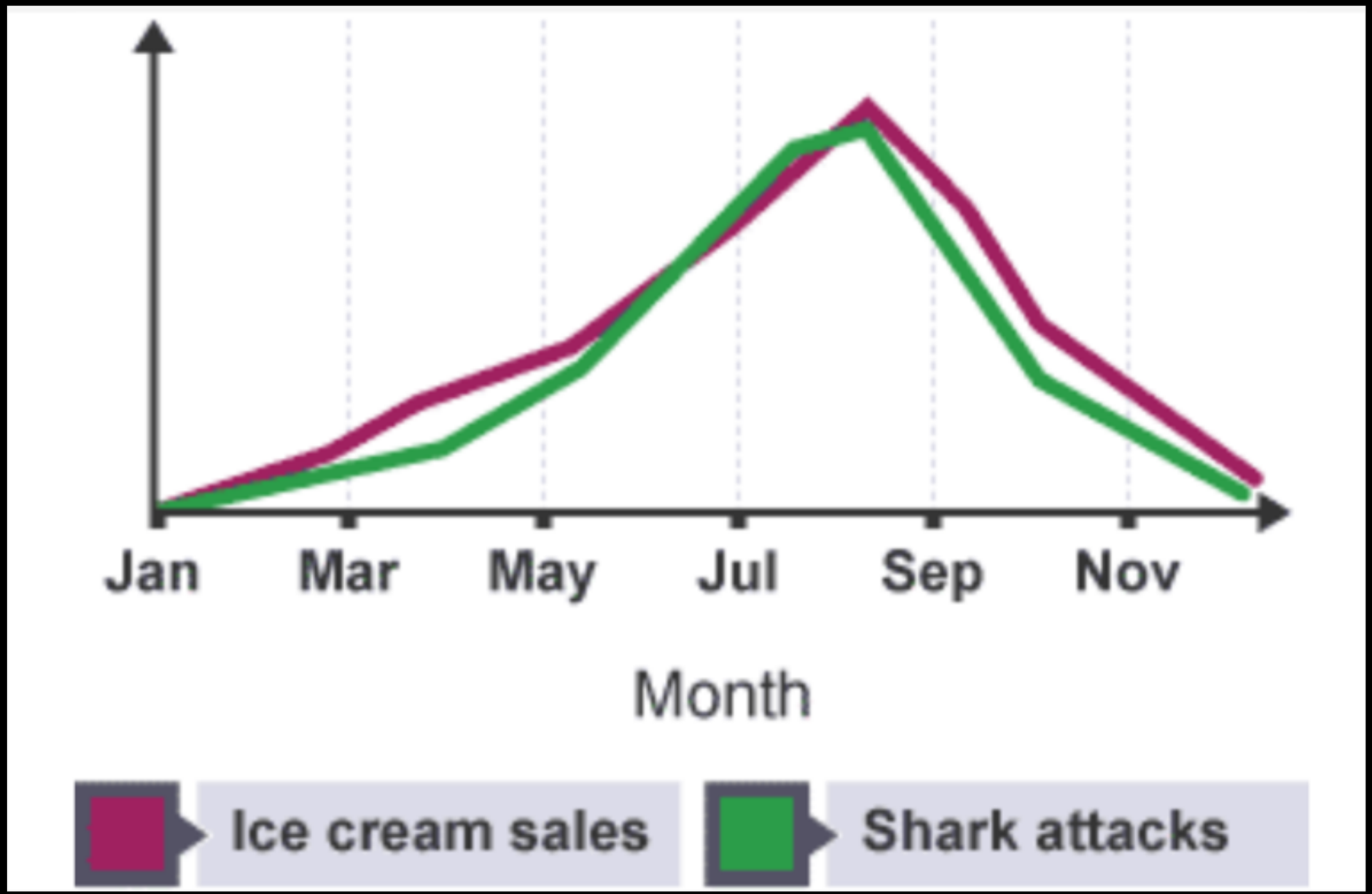
AAL, M. P. Brenner, L. J. Colwell, *PNAS*, 111, 13564 (2016)

Near optimality of random matrix bound

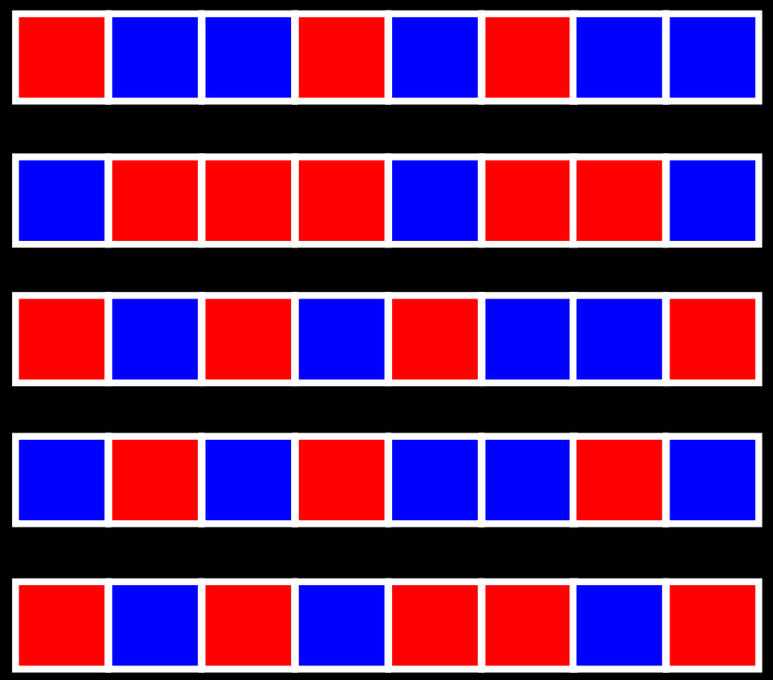


AAL, M. P. Brenner, L. J. Colwell, *PNAS*, 111, 13564 (2016)

Correlation vs causation



Posing a model for how data is generated



Q: What is the probability distribution that generates this dataset?

Maximum entropy formalism

We want the least structured model

$$S[p(\boldsymbol{\sigma})] = -\sum_{\{\boldsymbol{\sigma}\}} p(\boldsymbol{\sigma}) \log p(\boldsymbol{\sigma})$$

Whilst constraining the one and two point correlations to fit the data

$$\langle \sigma_i \rangle_{\text{data}} = \sum_{\{\boldsymbol{\sigma}\}} \sigma_i p(\boldsymbol{\sigma}) \qquad \langle \sigma_i \sigma_j \rangle_{\text{data}} = \sum_{\{\boldsymbol{\sigma}\}} \sigma_i \sigma_j p(\boldsymbol{\sigma})$$

$$p(\boldsymbol{\sigma}) = \frac{1}{Z} \exp\left(\sum_{i<j} J_{ij} \sigma_i \sigma_j + \sum_i h_i \sigma_i \right)$$

Exact inference is intractable

Maximum likelihood inference (Boltzmann Learning)
c.f. iterative Boltzmann inversion

$$h_i^{n+1} = h_i^n + \eta \left(\langle \sigma_i \rangle_D - \langle \sigma_i \rangle \right)$$

$$J_{ij}^{n+1} = J_{ij}^n + \eta \left(\langle \sigma_i \sigma_j \rangle_D - \langle \sigma_i \sigma_j \rangle \right)$$

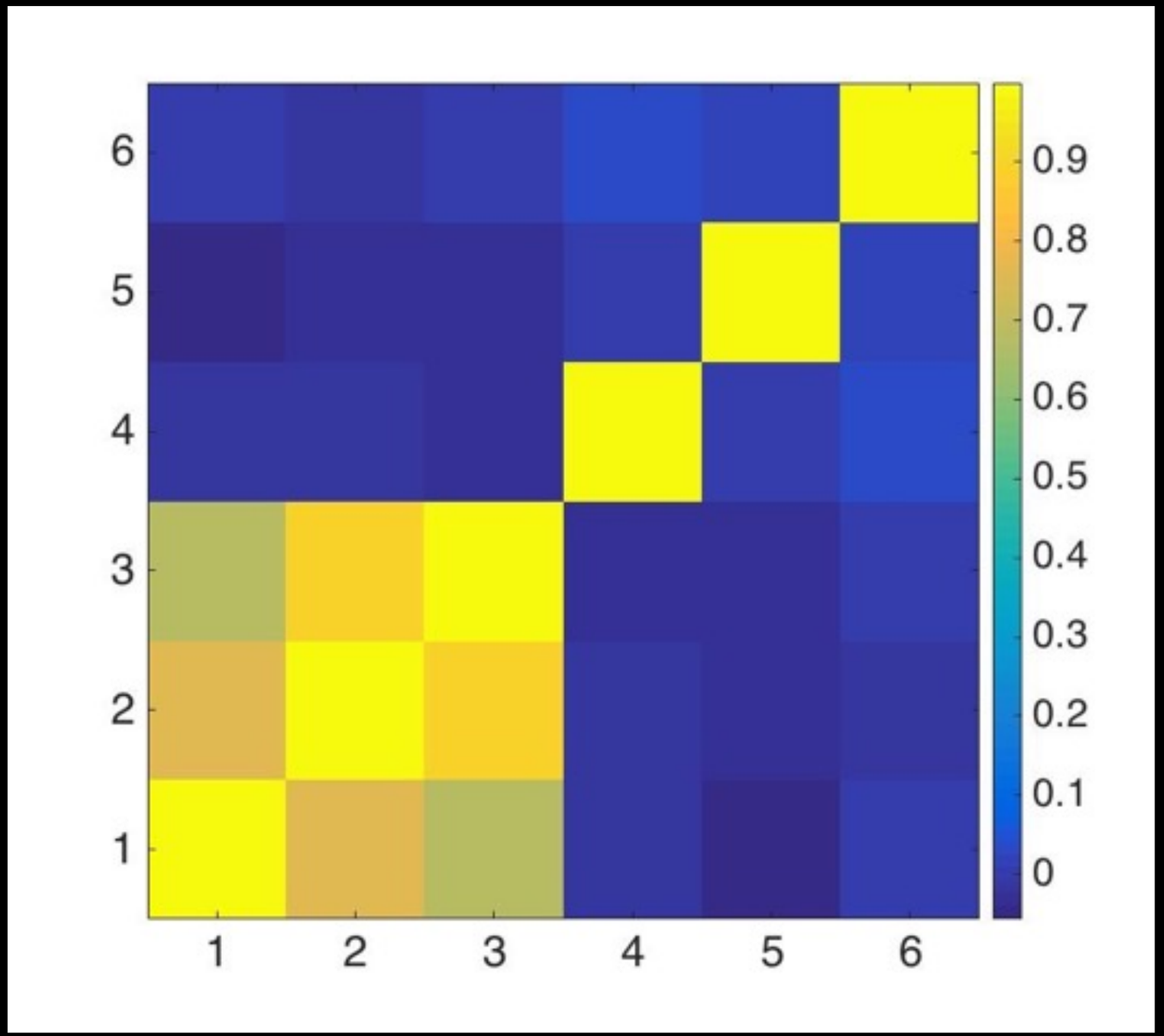
Requires sampling the probability distribution for each step!

Also assumes accurate covariance estimation from finite number of datapoints (more on that later)

D. H. Ackley, G. E. Hinton, and T. J. Sejnowski, Cognitive Science, 9, 147 (1985)

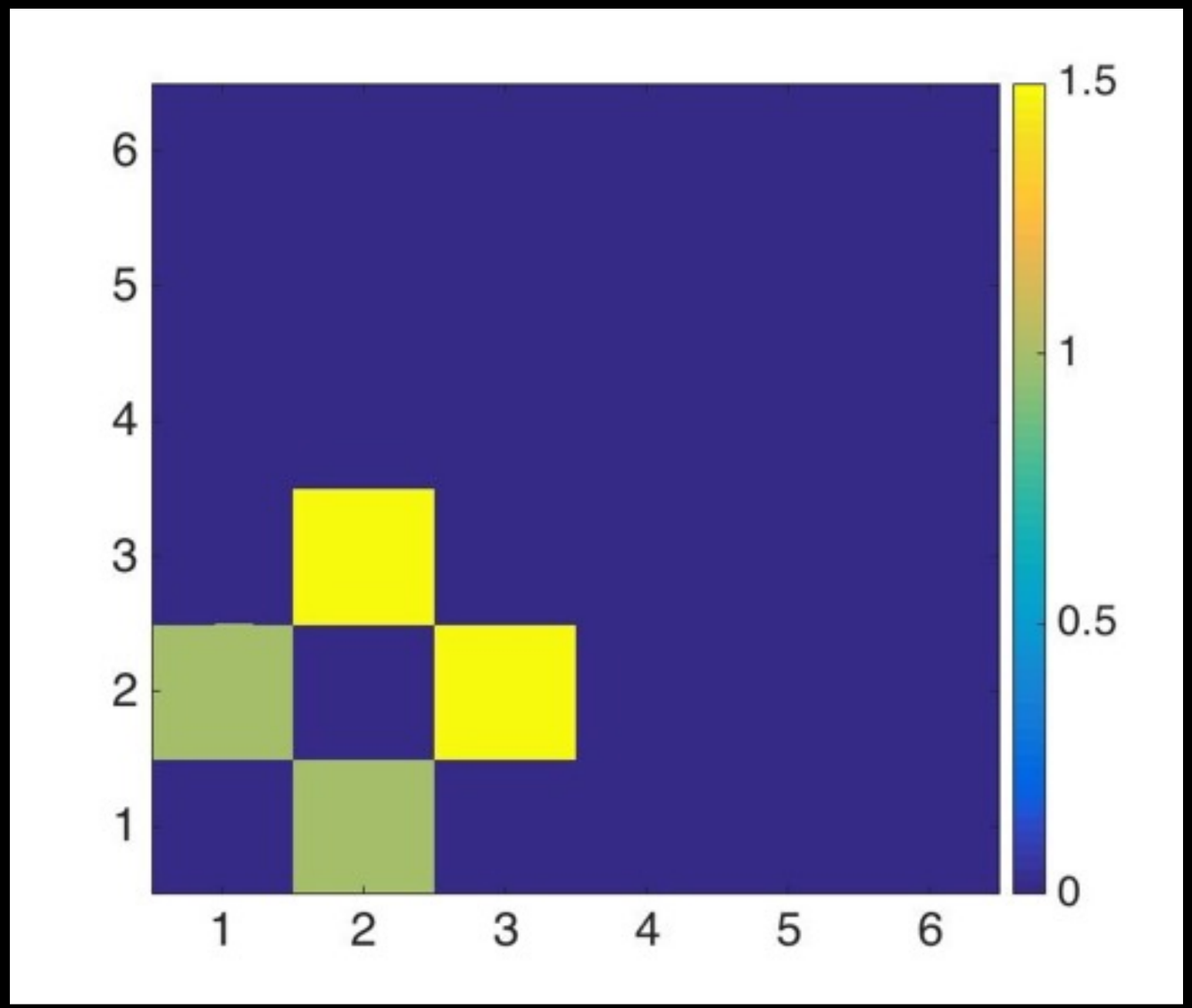
Inverse Ising model and indirect correlations

Observed correlation matrix

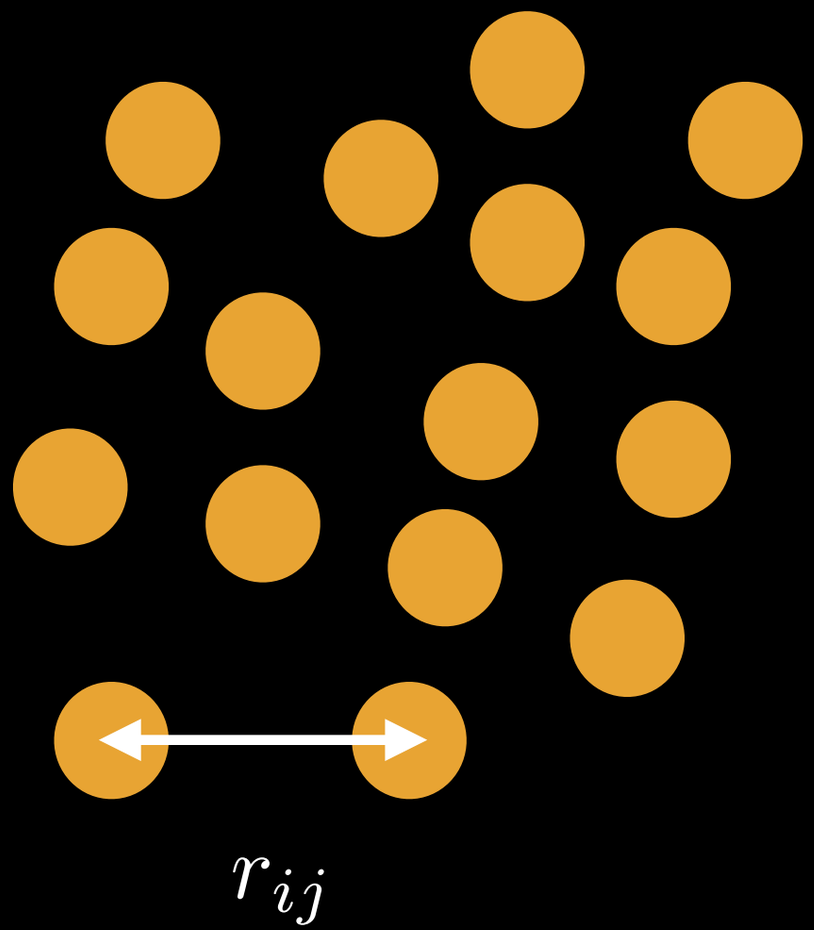


Inverse Ising model and indirect correlations

Actual coupling matrix



A detour in liquid state physics



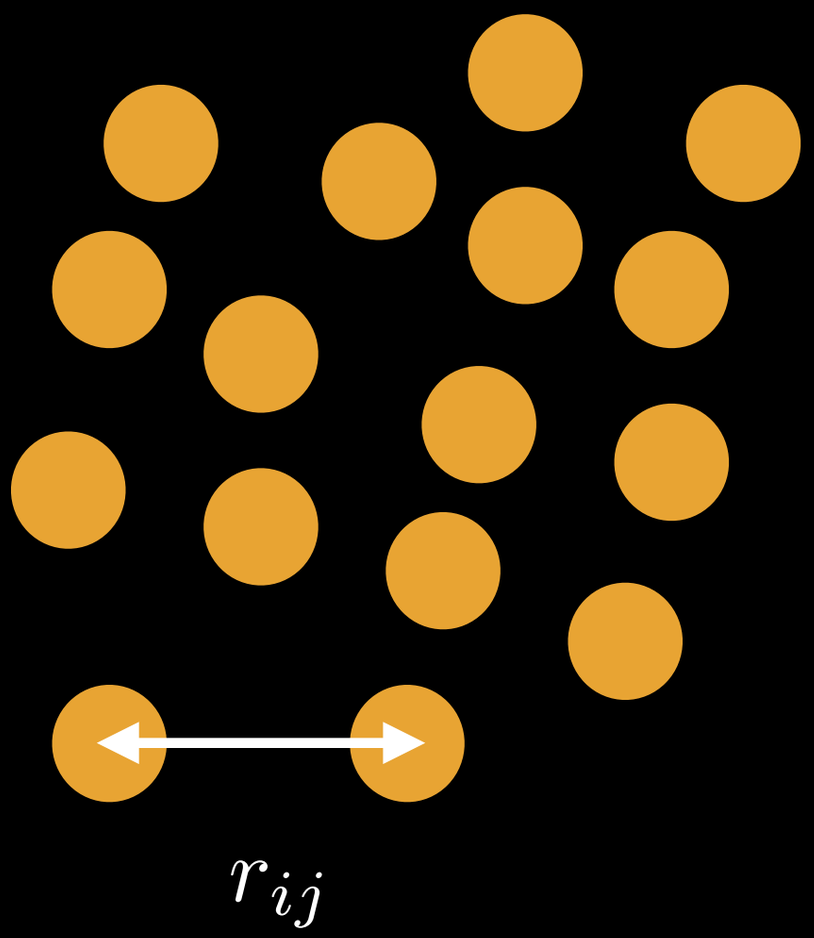
Intermolecular potential: $v(r_{ij})$

Radial distribution function: $g(r_{ij})$

Puzzle: $g(r_{ij})$ Long range

$v(r_{ij})$ Short range

Separating direct and indirect correlations



Total correlation: $h(r_{ij}) \equiv g(r_{ij}) - 1$

$$h(r_{ij}) = c(r_{ij}) + \int \mathrm{d}\mathbf{r}_k \; c(r_{ik})c(r_{kj}) + \int \mathrm{d}\mathbf{r}_k \int \mathrm{d}\mathbf{r}_l \; c(r_{ik})c(r_{kl})c(r_{lj}) + \cdots$$

$$\boxed{h(r_{ij}) = c(r_{ij}) + \int \mathrm{d}\mathbf{r}_k \; c(r_{ik})h(r_{kj})}$$

Ornstein and Zernike, 1914

Closing the OZ equation

Closure relations are approximately *local*

$$f(h(r_{ij}), c(r_{ij}), v(r_{ij}); \rho) = 0$$

e.g. the Hypernetted-chain approximation

$$c(r)/\rho = h(r)/\rho - \log(h(r)/\rho + 1) - v(r)$$

Ising meets Ornstein-Zernike

Covariance matrix C_{ij} $\longleftrightarrow$ $h(r_{ij})$ Total correlation function

"Direct coupling" matrix D_{ij} $\longleftrightarrow$ $c(r_{ij})$ Direct correlation function

$$C_{ij} = \delta_{ij} + D_{ij} + \sum_k D_{ik} D_{kj} + \sum_{k,l} D_{ik} D_{kl} D_{lj} + \cdots$$

$$D = \mathbb{I} - C^{-1}$$

Ornstein-Zernike meets Deep Learning

"OZ-like" closure for the inverse Ising problem

$$J_{ij} \approx F\left(C_{ij}, [C^{-1}]_{ij}, \langle s_i \rangle, \langle s_j \rangle\right)$$

$$h_i \approx G\left(\tanh^{-1}\langle s_i \rangle, [C^{-1}]_{ii}, \sum_{j\neq i} J_{ij}\langle s_j \rangle, \sum_{j\neq i} C_{ij}\langle s_j \rangle\right)$$

The OZ problem can be interpreted as non-linear regression

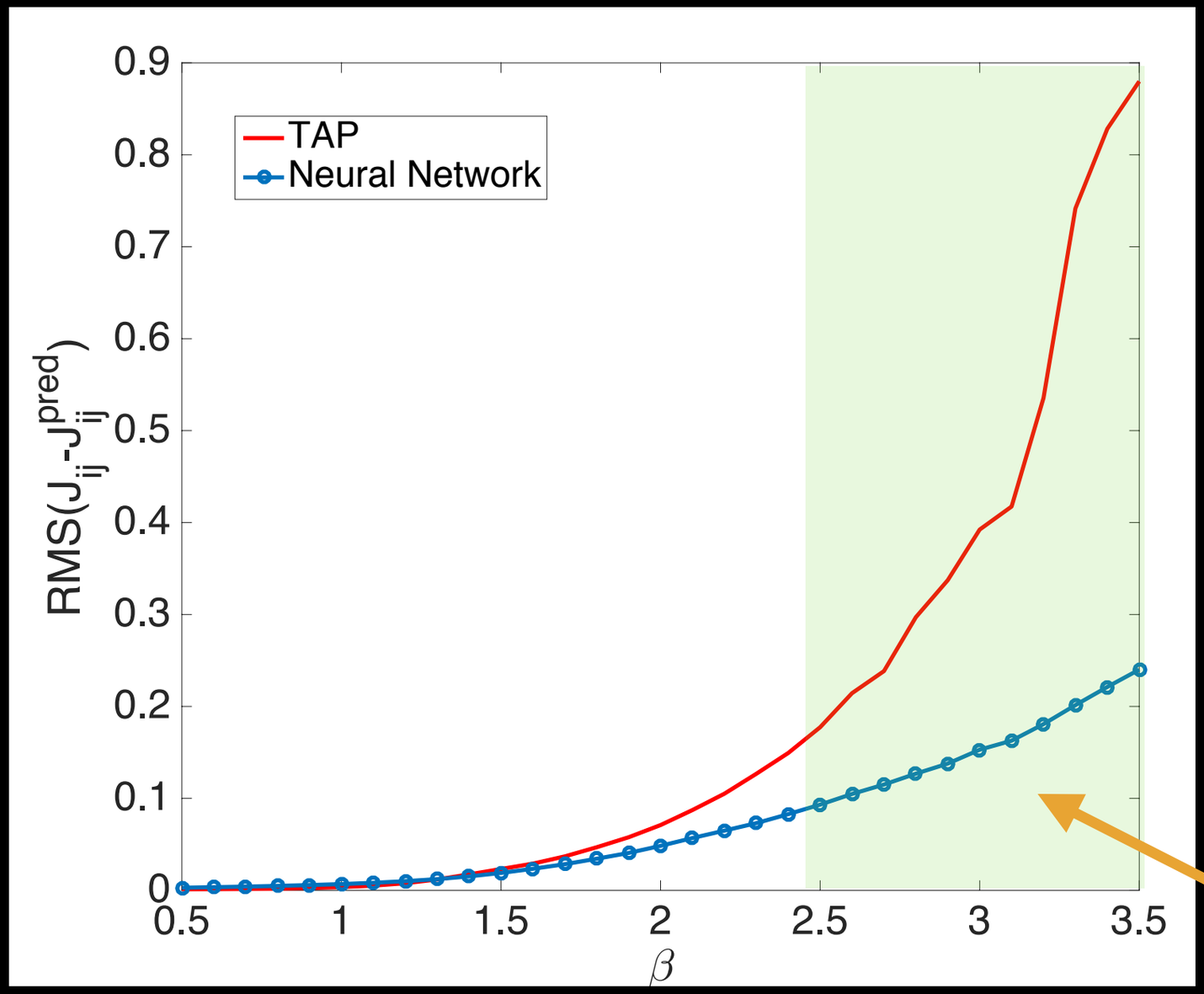
MCMC sampling till convergence

$$\{J_{ij}, h_i\} \rightleftarrows \{C_{ij}, \langle s_i \rangle\}$$

Non-linear regression

S. Turi, **AAL**, arXiv:1706.08466

The deep learning approach is generalisable



Validation data:

- Gaussian distributed J_{ij} and h_i with zero mean and unit variance, p = 70 sites
- "Inverse temperature" between 0.5-3.5

Extrapolation outside training set

S. Turi, **AAL**, arXiv:1706.08466

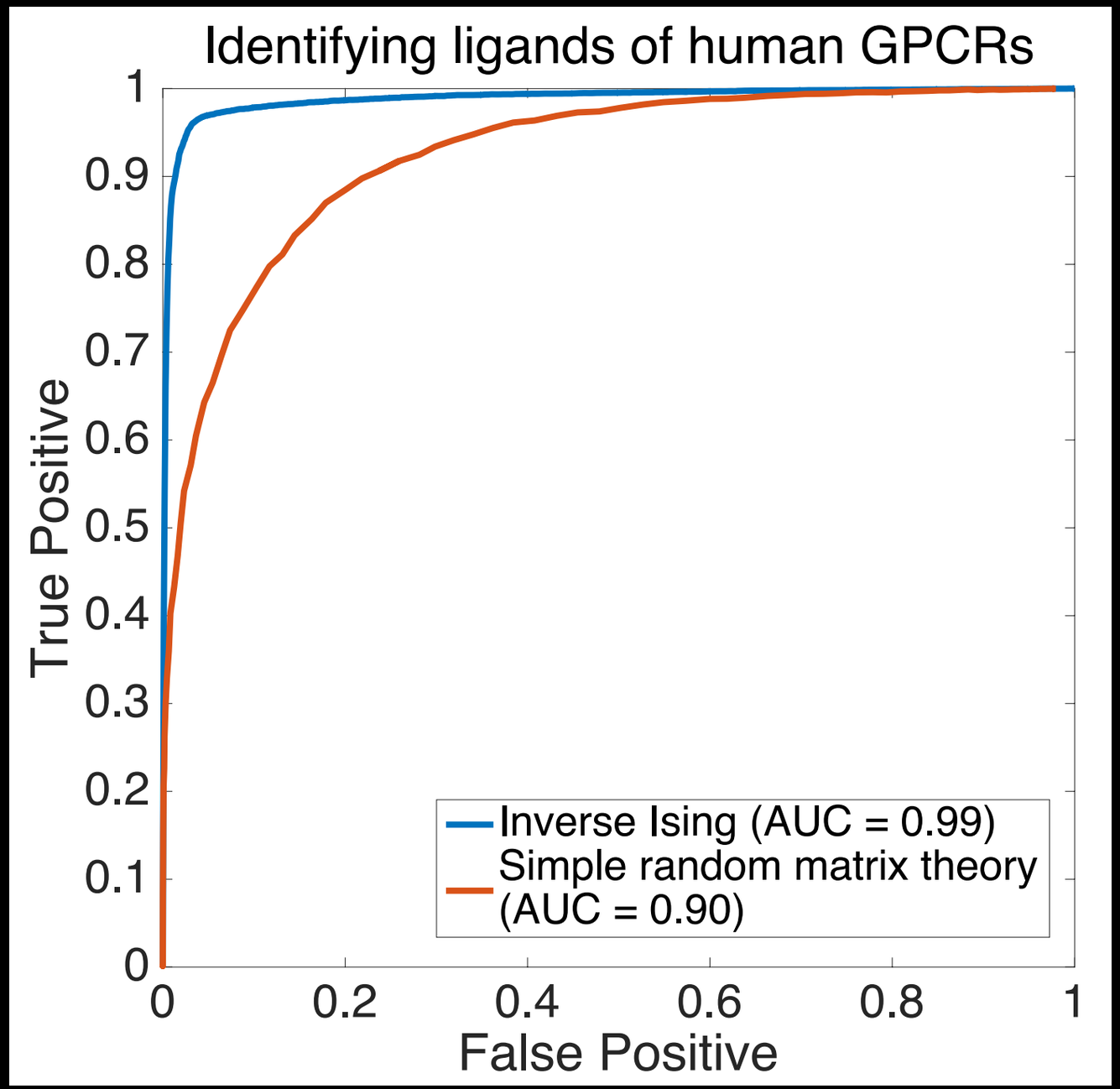
Using random matrices again to remove finite sampling noise

$$J_{ij} \approx F\left(C_{ij}, [C^{-1}]_{ij}, \langle s_i \rangle, \langle s_j \rangle\right)$$

$$h_i \approx G\left(\tanh^{-1}\langle s_i \rangle, [C^{-1}]_{ii}, \sum_{j \neq i} J_{ij}\langle s_j \rangle, \sum_{j \neq i} C_{ij}\langle s_j \rangle\right)$$

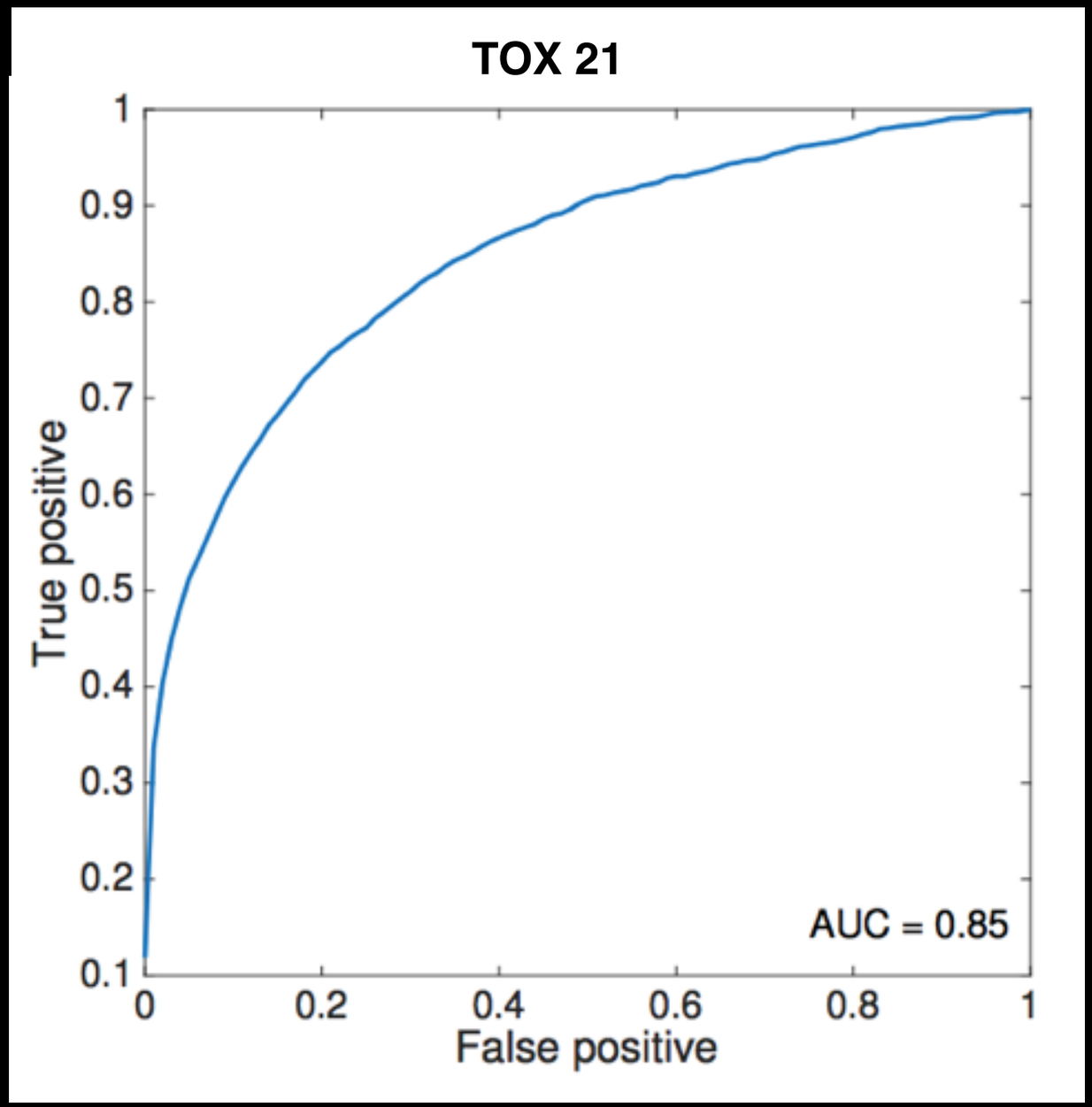
We need to remove the eigenvectors below the random matrix bound, else the noise will dominate C^{-1}

Performance of classification algorithm: Human GPCRs



AAL, arXiv:1706.08466

Performance of classification algorithm: Tox 21



MoleculeNet benchmark
AUC = 0.83 with deep neural networks (multitasking model)

S. Turi, **AAL**, arXiv:1706.08466

Prospective discovery of novel CHAM1 agonists

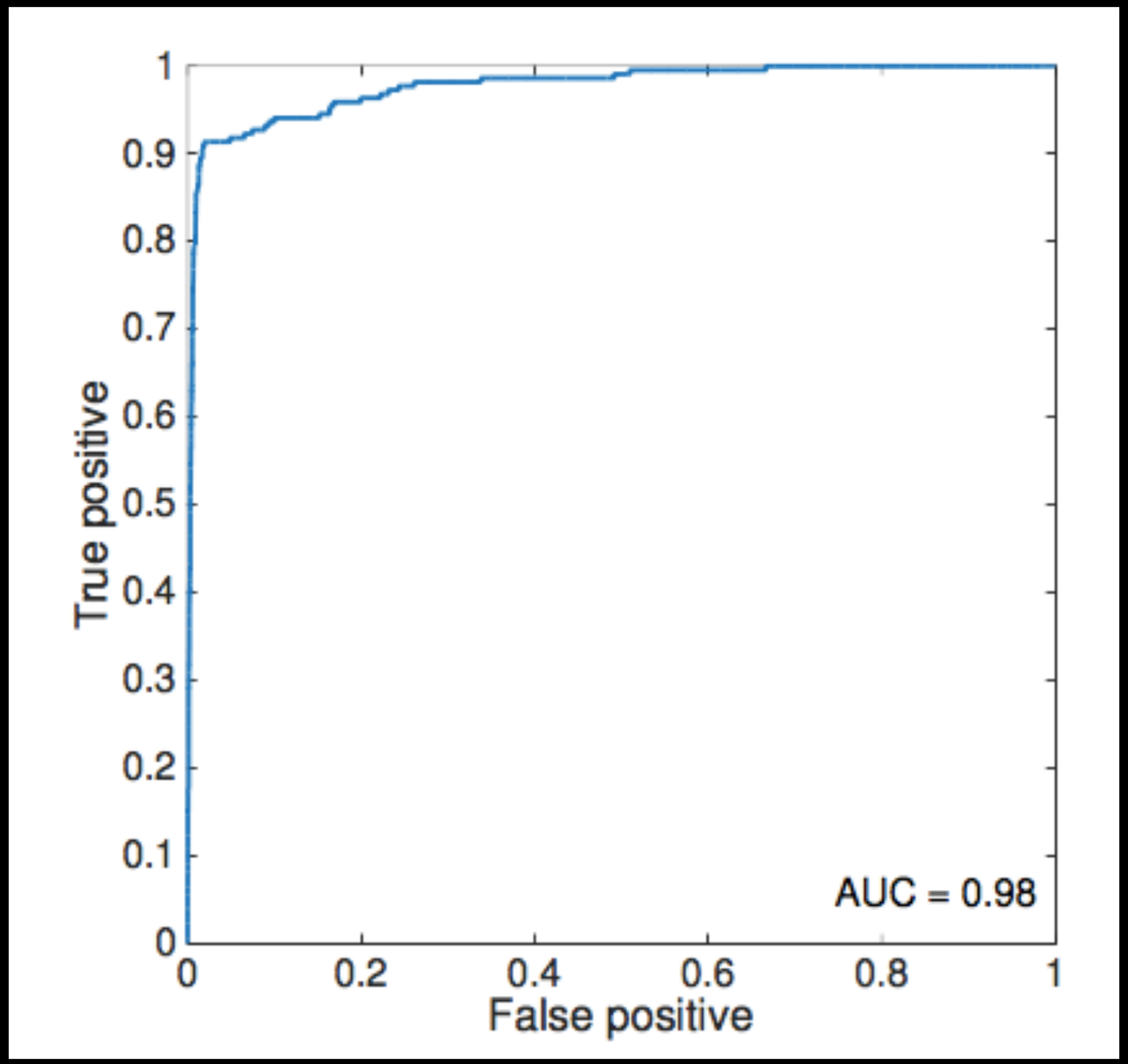
- Human CHRM1 is a receptor that is responsible for central nervous system functions

- M1 agonists are hypothesized to ameliorate the symptoms of Alzheimers disease and Schizophrenia

- Our random matrix algorithm is trained on internal Pfizer data: 222 active, 5223 inactive

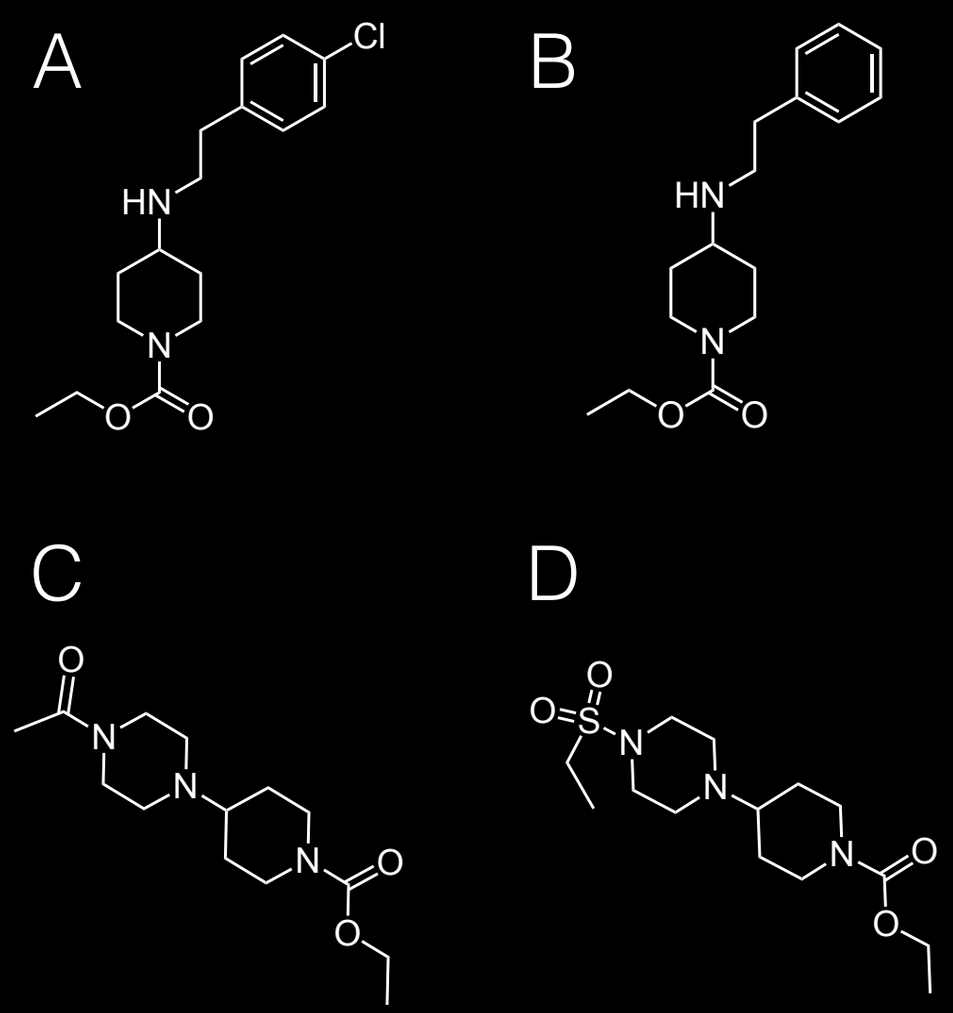
AAL and Pfizer collaboration, unpublished (2018)

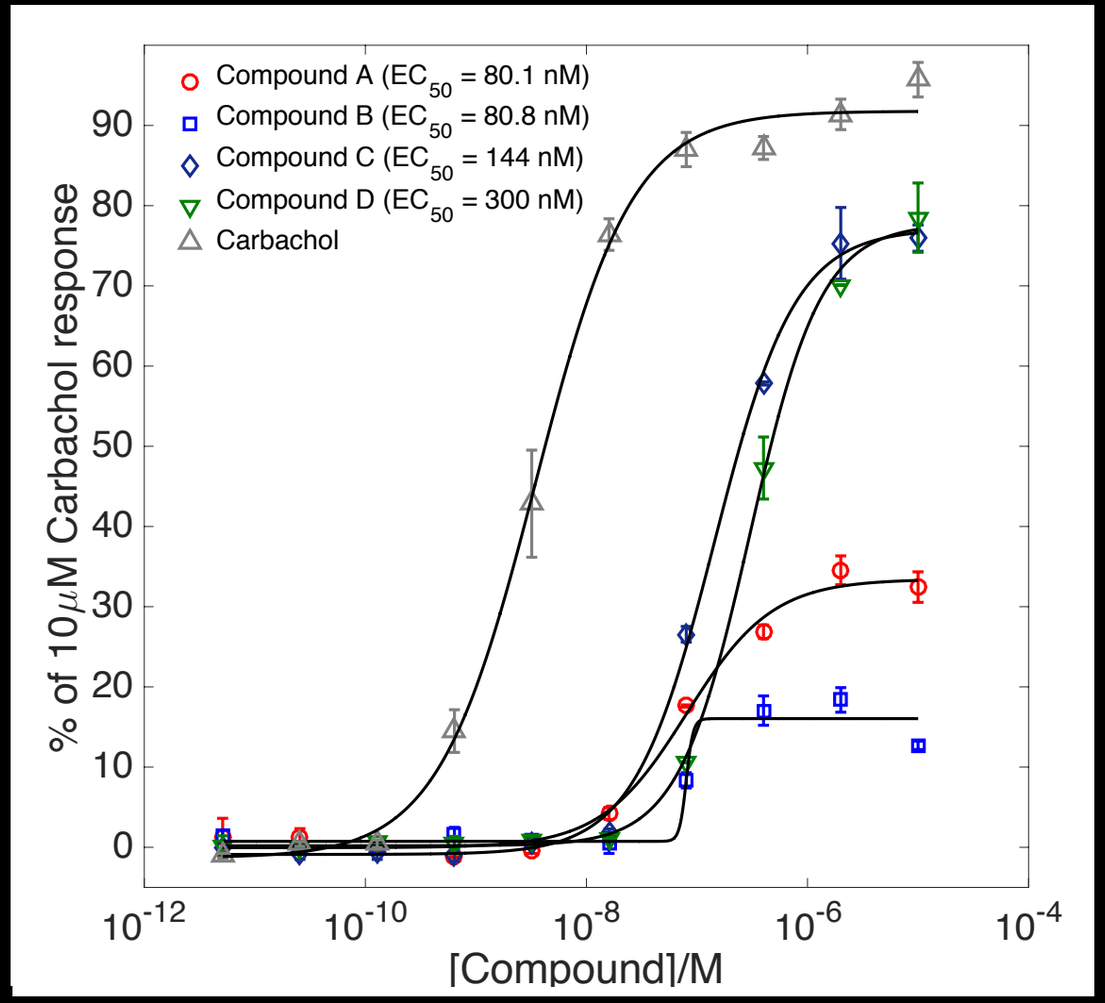
Prospective discovery of novel CHAM1 agonists



Prospective discovery of novel CHAM1 agonists

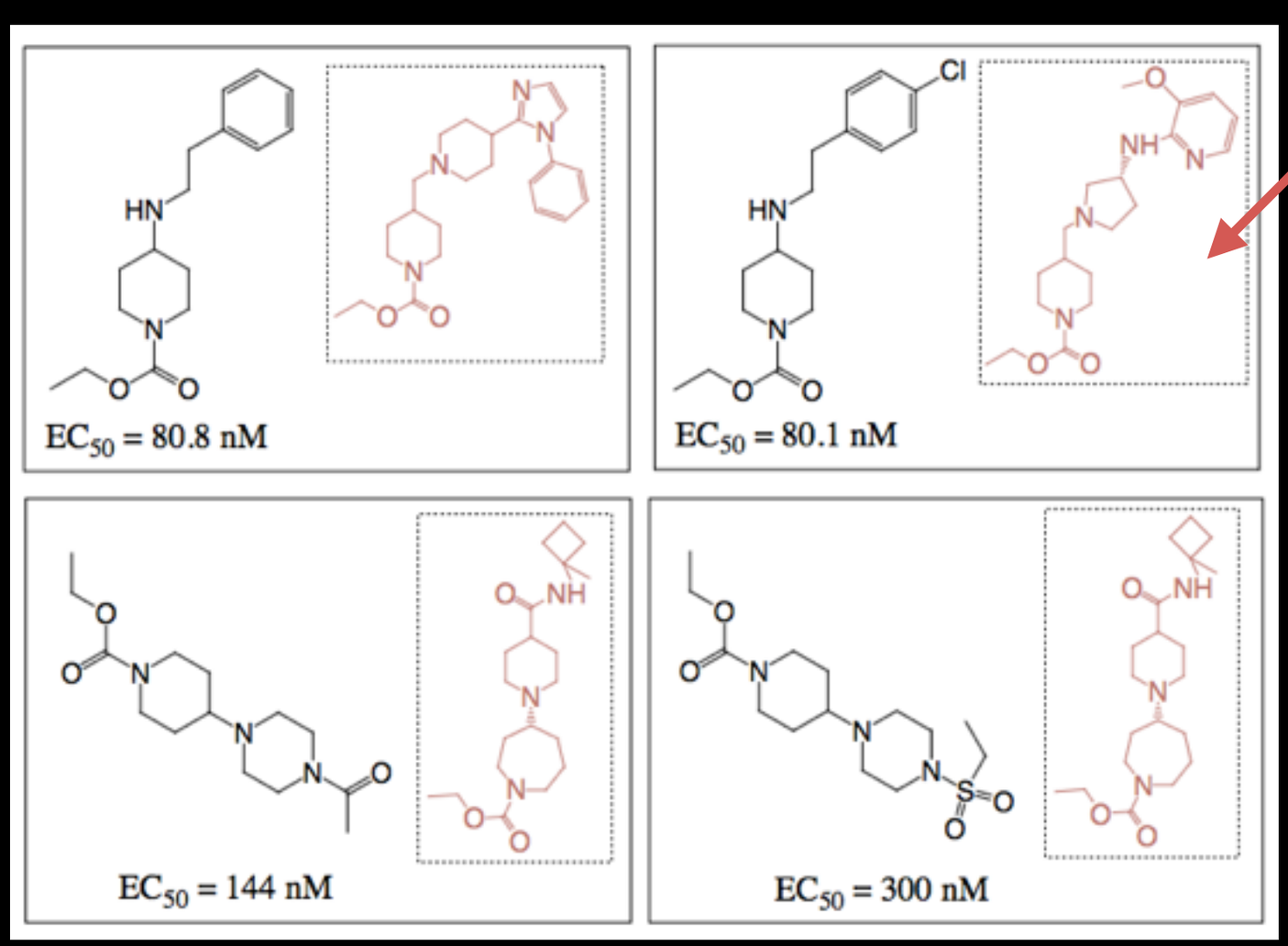
- The algorithm is used to screen the e-Molecules database of ~5.9 million purchasable molecules. The top 118 were prospectively experimentally tested for activity





Prospective discovery of novel CHAM1 agonists

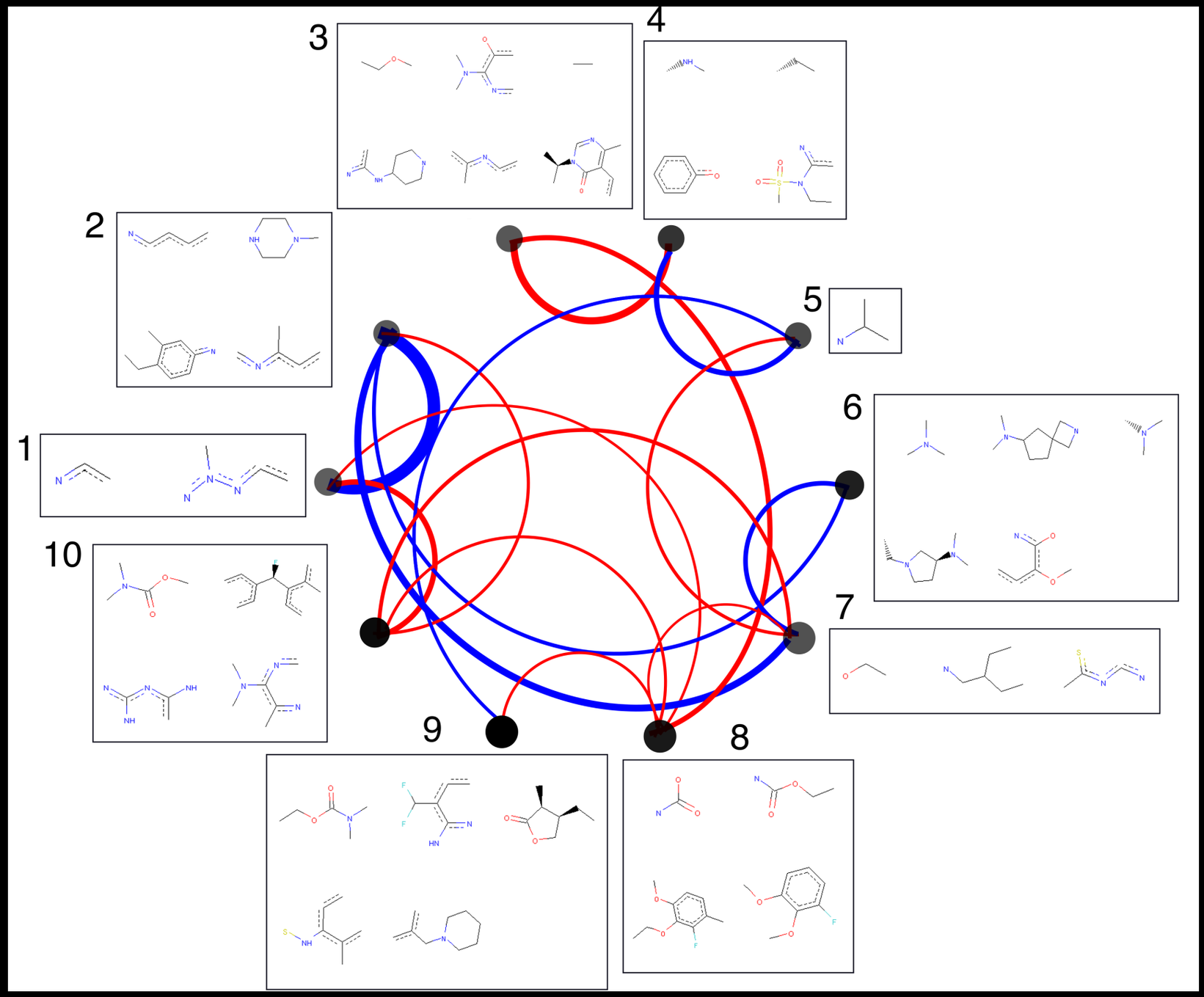
- The algorithm is used to screen the e-Molecules database of ~5.9 million purchasable molecules. The top 118 were prospectively experimentally tested for activity



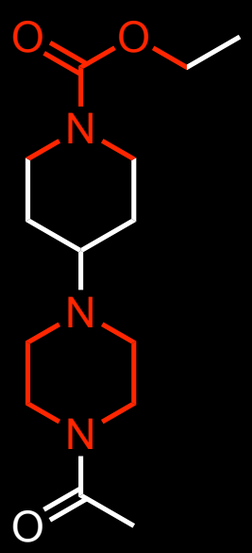
How good is the model?

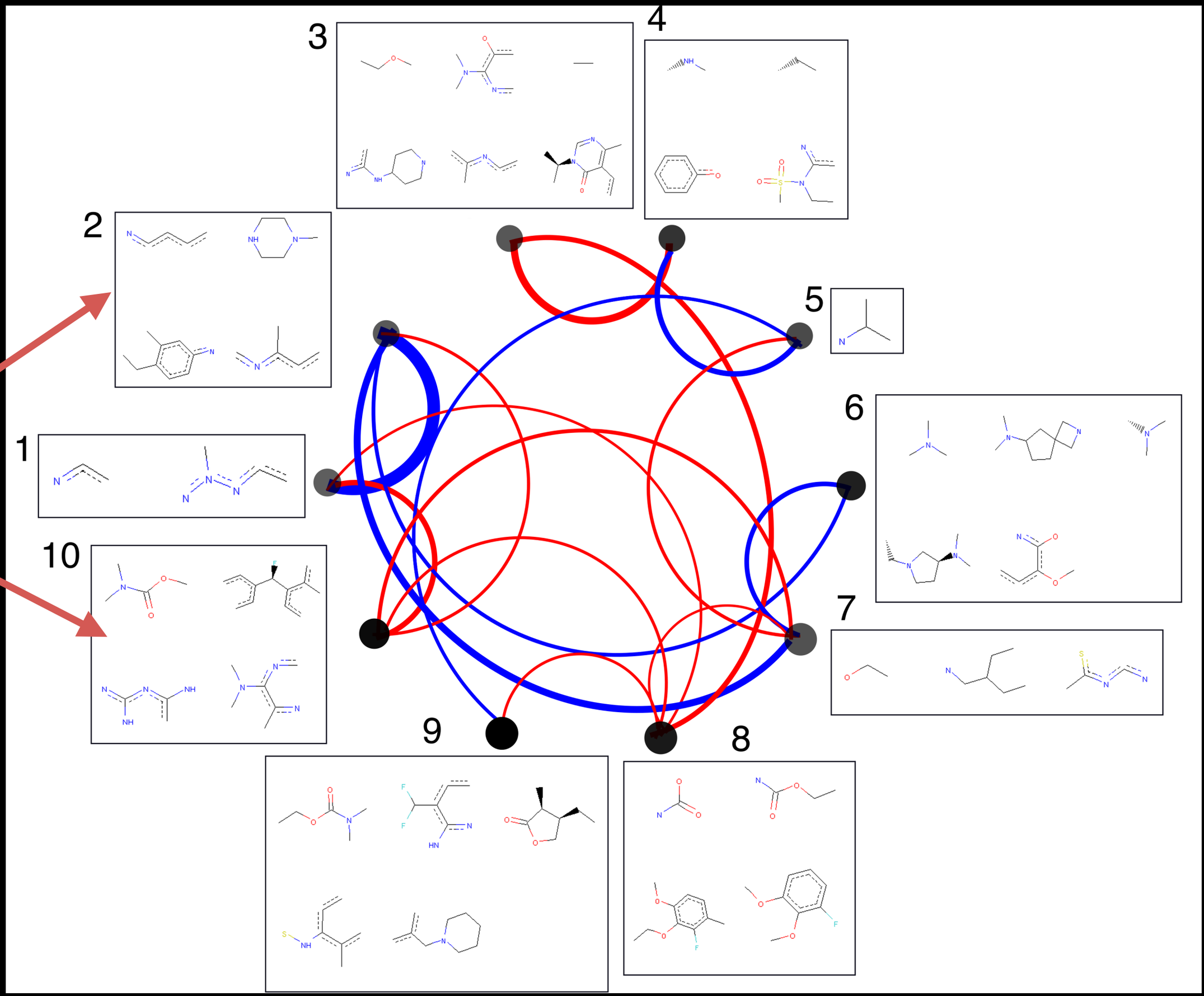
- Hit rate of the random matrix algorithm: ~4%
- Hit rate at random: <0.4%
- Hit rate of algorithms used on a day-to-day basis at Pfizer: 0% (Tanimoto similarity), 2% (naive Bayes)

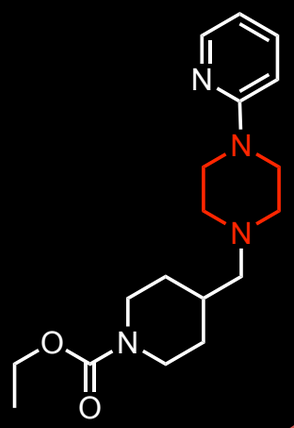
How does the algorithm make predictions?



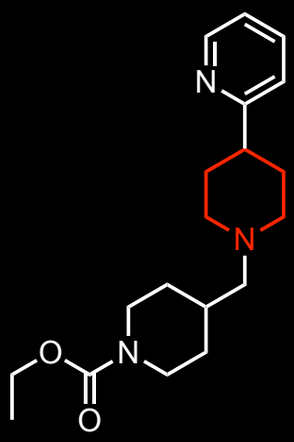
- The algorithm infers correlations between presence/ absence of chemical groups
- Red denotes positive correlations, blue denotes negative correlations



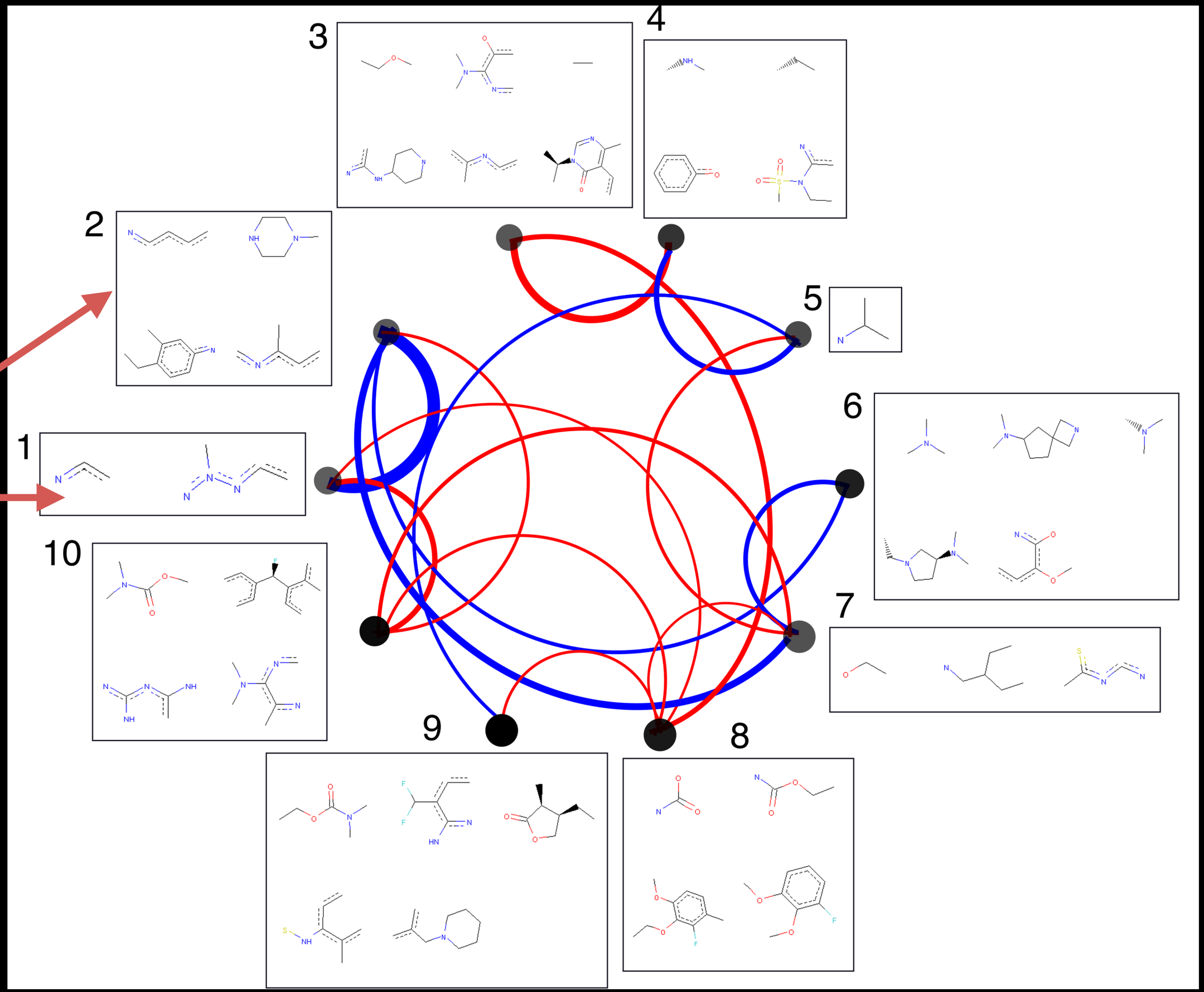

1
2
3
4
5
6
7
8
9
10



$EC_{50} > 10\ \mu M$



$EC_{50} = 0.150\ \mu M$

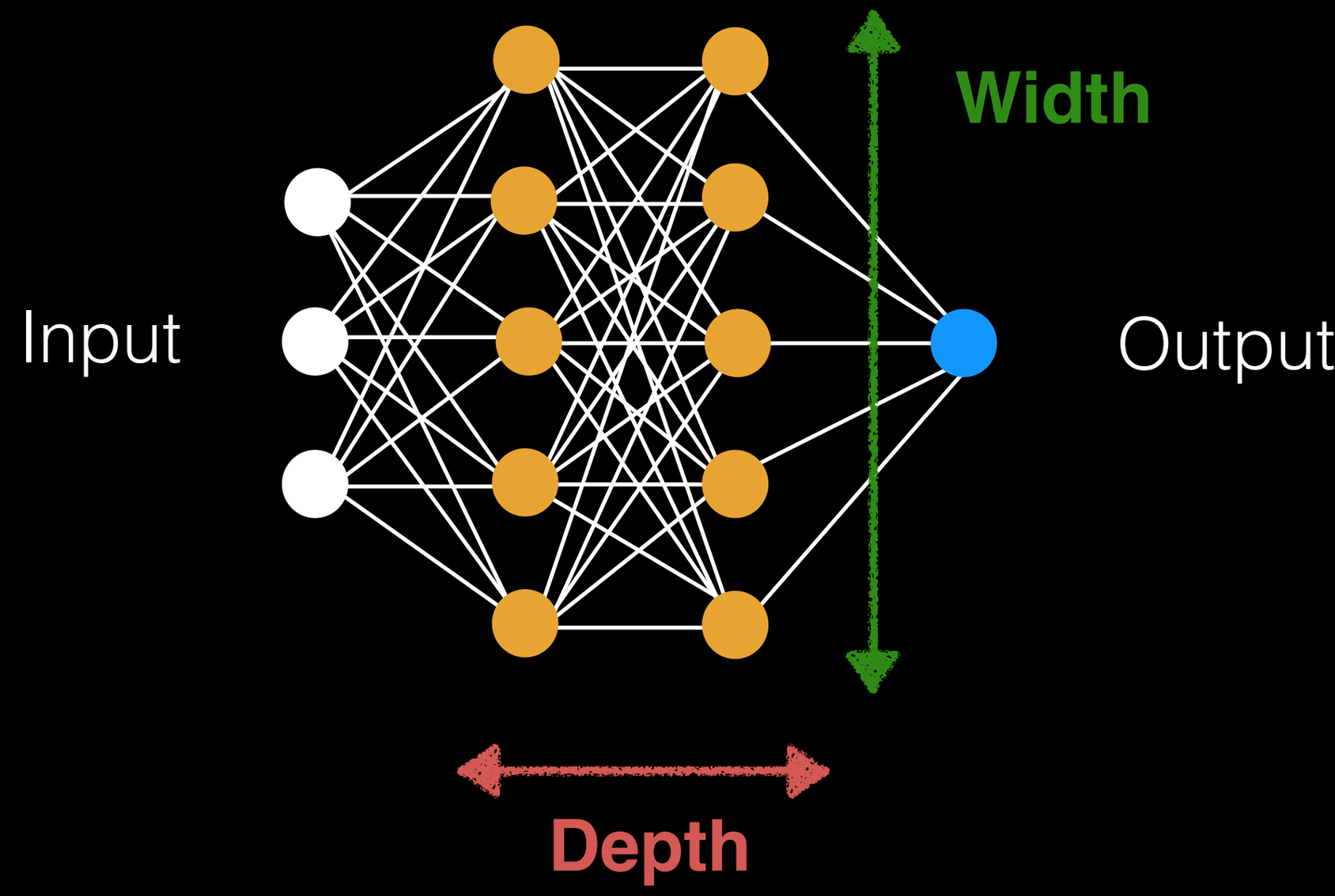


Two fairy tales on entropy, random matrices and spin-glasses

- Predicting the biological activity of small molecules
- **Understanding why deep neural networks “works”**

What is under the hood?

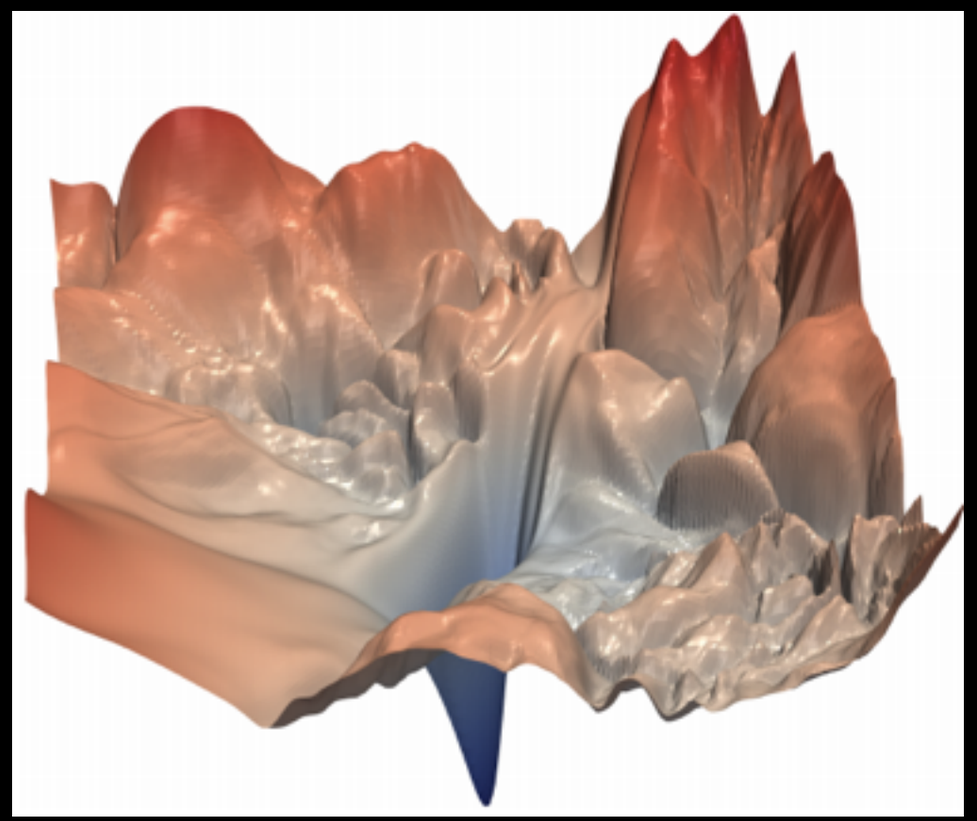
$$y = \sigma(\mathbf{W}_H \sigma(\mathbf{W}_{H-1} \cdots \sigma(\mathbf{W}_2 \sigma(\mathbf{W}_1 \mathbf{x}))))$$



The mechanics of deep learning

- We want to find parameters that minimise the loss function

$$L(\mathbf{\Theta}) = \frac{1}{P}\sum_{i=1}^{P} l(y_i, f(\mathbf{x}_i, \mathbf{\Theta}))$$



Li et al., arXiv:1712.09913 (2018)

The mechanics of deep learning

- We want to find parameters that minimise the loss function

$$y_{\text{pred}} = f(\mathbf{x}_i, \boldsymbol{\Theta}) \qquad L(\boldsymbol{\Theta}) = \frac{1}{P}\sum_{i=1}^{P} l(y_i, f(\mathbf{x}_i, \boldsymbol{\Theta}))$$

- Stochastic gradient descent

$$\boldsymbol{\Theta}^{t+1} = \boldsymbol{\Theta}^{t} - \eta \frac{1}{|\mathcal{B}_t|} \sum_{i \in \mathcal{B}_t} \nabla_{\boldsymbol{\Theta}} l(y_i, f(\mathbf{x}_i, \boldsymbol{\Theta}^t))$$

Two mysteries of deep learning

- The loss function is non-convex, yet gradient descent can find “sufficiently good” parameters
- An infinitely wide 1 layer neural network can approximate any smooth function
    - BUT it is known that a deep networks attain lower error than wide networks, fixing the number of parameters

Questions about DNNs

- **Why stochastic gradient descent?**
- Why deep networks?

The continuous time limit of SGD

$$\boldsymbol{\Theta}^{t+1} = \boldsymbol{\Theta}^{t} - \eta \frac{1}{|\mathcal{B}_t|} \sum_{i \in \mathcal{B}_t} \nabla_{\boldsymbol{\Theta}} l(y_i, f(\mathbf{x}_i, \boldsymbol{\Theta}^t))$$

$\eta \to 0$

$$\frac{\mathrm{d}\boldsymbol{\Theta}}{\mathrm{d}t} = -\nabla_{\boldsymbol{\Theta}} L(\boldsymbol{\Theta}) + \boldsymbol{\eta}(t)$$

$$\boldsymbol{\eta}(t) = \nabla_{\boldsymbol{\theta}} \left[\frac{1}{P} \sum_i l_i(\mathbf{x}_i, \boldsymbol{\theta}) - \frac{1}{|B_t|} \sum_{i \in B_t} l_i(\mathbf{x}_i, \boldsymbol{\theta}) \right]$$

The noise is the signal

The noise is mean zero

$\langle \boldsymbol{\eta}(t) \rangle = \mathbf{0}$

BUT the variance is non-trivial

$$\langle \eta_\mu(t)\eta_\nu(t') \rangle = \frac{1}{b}\left(1 - \frac{b}{P}\right)\left[\left\langle \frac{\partial l}{\partial \theta_\mu}\frac{\partial l}{\partial \theta_\nu}\right\rangle - \left\langle \frac{\partial l}{\partial \theta_\mu}\right\rangle\left\langle \frac{\partial l}{\partial \theta_\nu}\right\rangle\right]\delta(t-t')$$

Effective temperature

Near a critical point $\langle \partial l/\partial \theta_\nu \rangle \approx 0$

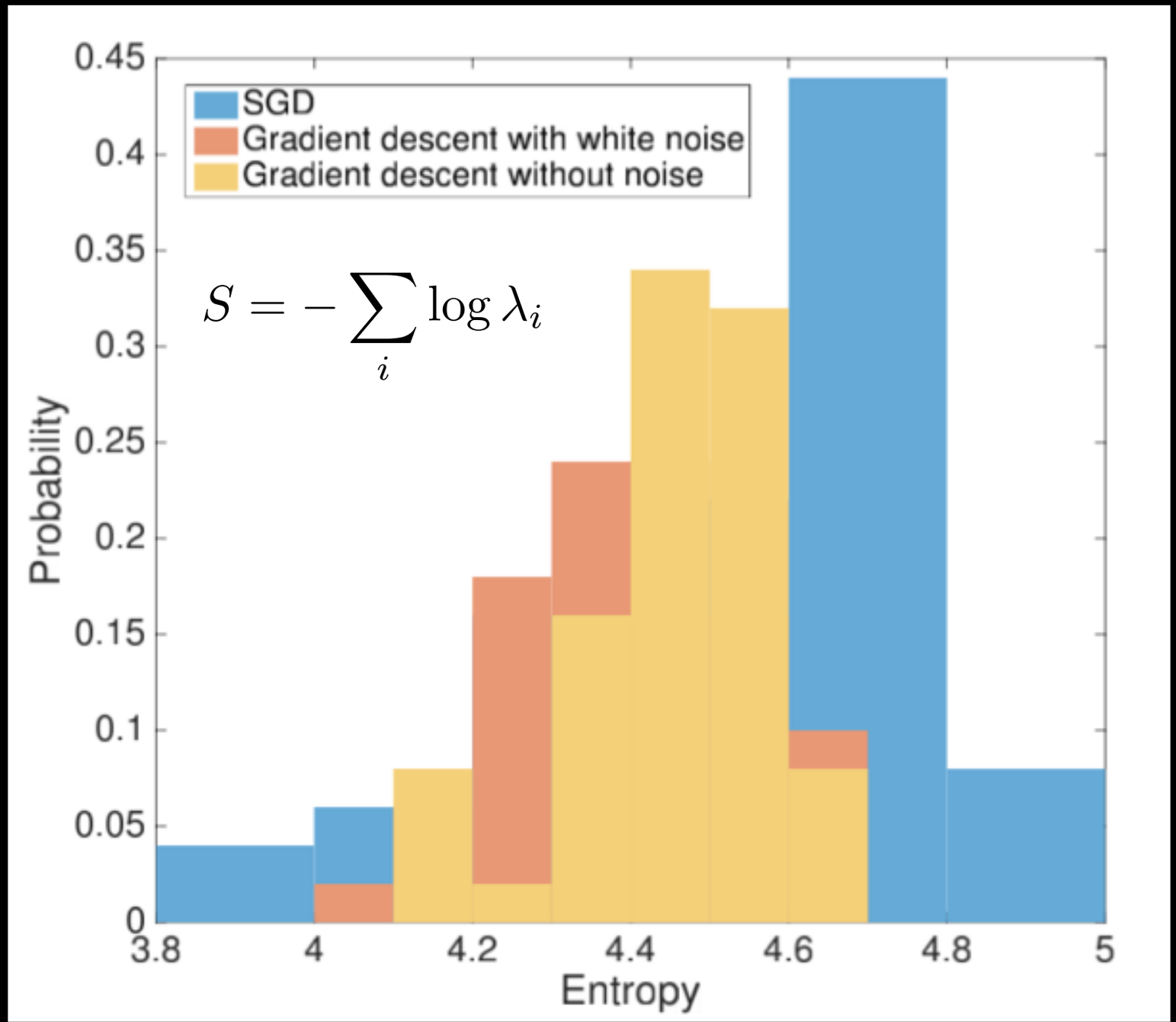
If l is a log likelihood function

$$\left\langle \frac{\partial l}{\partial \theta_\mu}\frac{\partial l}{\partial \theta_\nu}\right\rangle = \left\langle \frac{\partial^2 l}{\partial \theta_\mu \partial \theta_\nu}\right\rangle$$

Noise pointing at directions with large eigenvalues, i.e. stiff directions

Thus SGD preferentially selects **wide** minima

SGD preferentially selects wide minima



Y. Zhang, A. M. Saxe, M. Advani, **AAL**, Molecular Physics (2018)

Why are wide minima good?

Let the ground truth be

$$y_i = f(\mathbf{x}_i, \boldsymbol{\Theta}) + \boldsymbol{\epsilon}_i \qquad \boldsymbol{\epsilon}_i \sim \mathcal{N}(0, \sigma^2)$$

The posterior (assuming a uniform prior)

$$p(\boldsymbol{\Theta}|\{\mathbf{x}_i, y_i\}_{i=1}^P) = \frac{1}{Z} \exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^{P} (f(\mathbf{x}_i, \boldsymbol{\Theta}) - y_i)^2\right)$$

Given a new data point

$$\langle \hat{y} \rangle = \int p(\boldsymbol{\Theta}|\{\mathbf{x}_i, y_i\}_{i=1}^P) f(\hat{\mathbf{x}}, \boldsymbol{\Theta}) \, \mathrm{d}\boldsymbol{\Theta}$$

Laplace approximation

Define the (intensive) quantity

$$u(\mathbf{\Theta}) = \frac{1}{P}\sum_{i=1}^{P}\frac{(f(\mathbf{x}_i,\mathbf{\Theta})-y_i)^2}{2\sigma^2}$$

Harmonic expansion around each minima

$$\langle\hat{y}\rangle \approx \frac{1}{Z}\sum_q \frac{\exp(-Pu(\mathbf{\Theta}_q))}{\sqrt{\det H(\mathbf{\Theta}_q)}} f(\hat{\mathbf{x}},\mathbf{\Theta}_q)$$

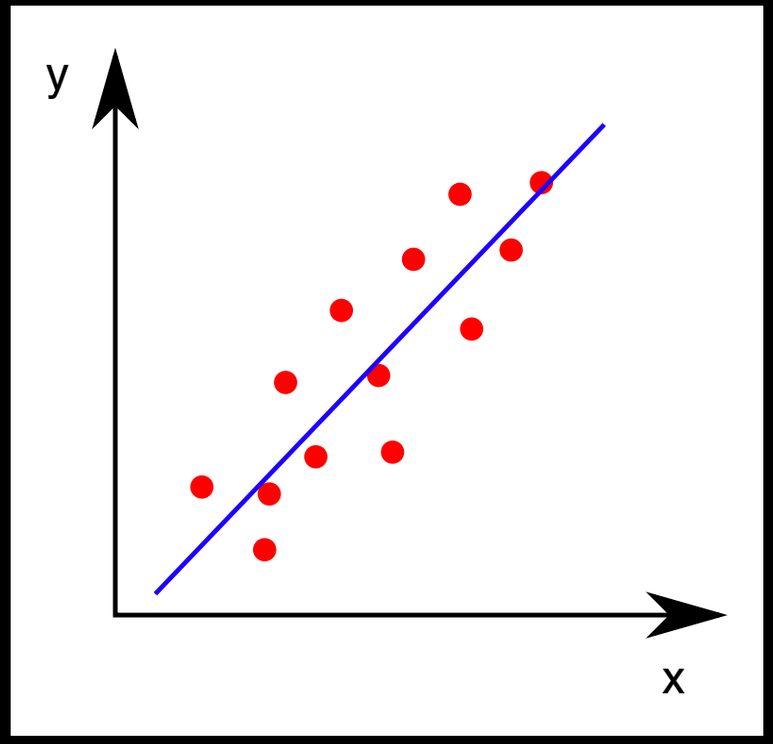
$$= \frac{1}{Z}\sum_q \exp\left(-Pu(\mathbf{\Theta}_q) - N\underbrace{\frac{1}{N}\sum_i \log\lambda_i(\mathbf{\Theta}_q)}_{-s(\mathbf{\Theta}_q)}\right) f(\hat{\mathbf{x}},\mathbf{\Theta}_q)$$

c.f. Sloppiness: Mechta et al., Science, 342, 604 (2013)
Heat capacity: R. Das and D. J. Wales, Physical Review E 93, 063310 (2016)

Entropy is the key

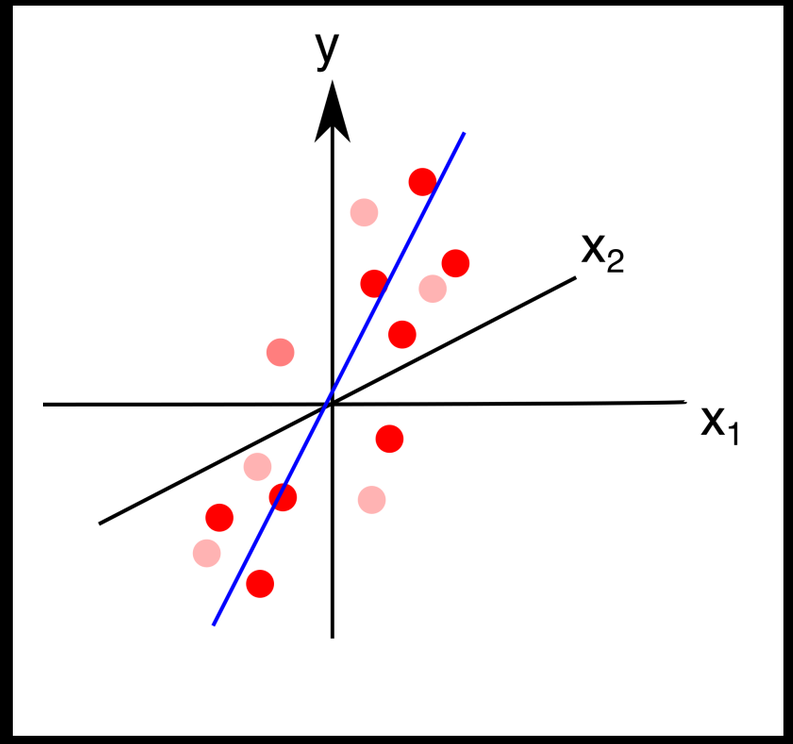
$$\langle y \rangle = \frac{1}{Z} \sum_q \exp(-PF(\mathbf{\Theta}_q)) f(\hat{\mathbf{x}}, \mathbf{\Theta}_q)$$

$$F(\mathbf{\Theta}_q) = u(\mathbf{\Theta}_q) - \frac{N}{P} s(\mathbf{\Theta}_q)$$



Classical statistics

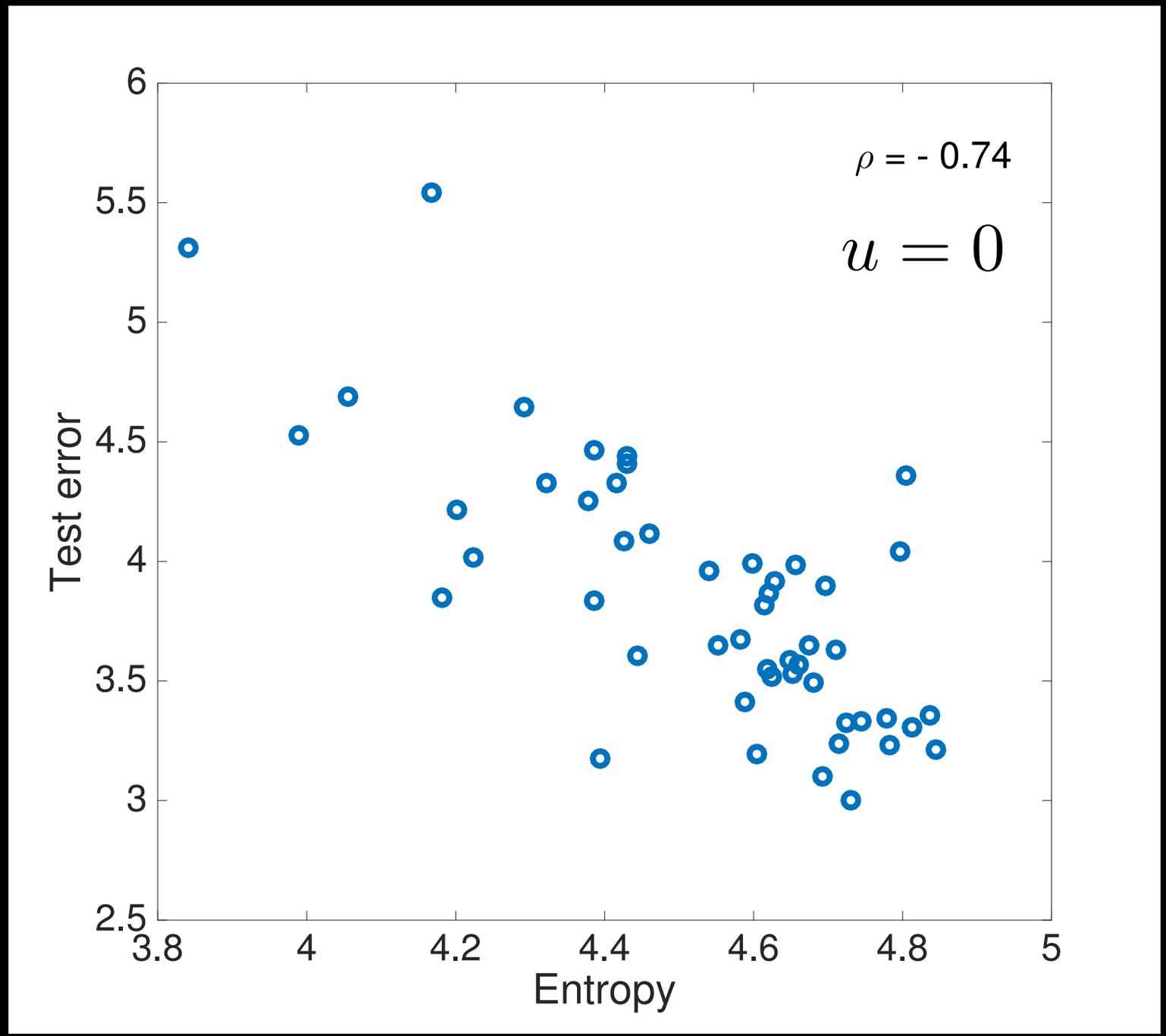
$P \to \infty \quad N/P \to 0$



Deep learning

$P \to \infty \quad N/P = O(1)$

Numerical experiment with overparameterised neural networks



Roadmap

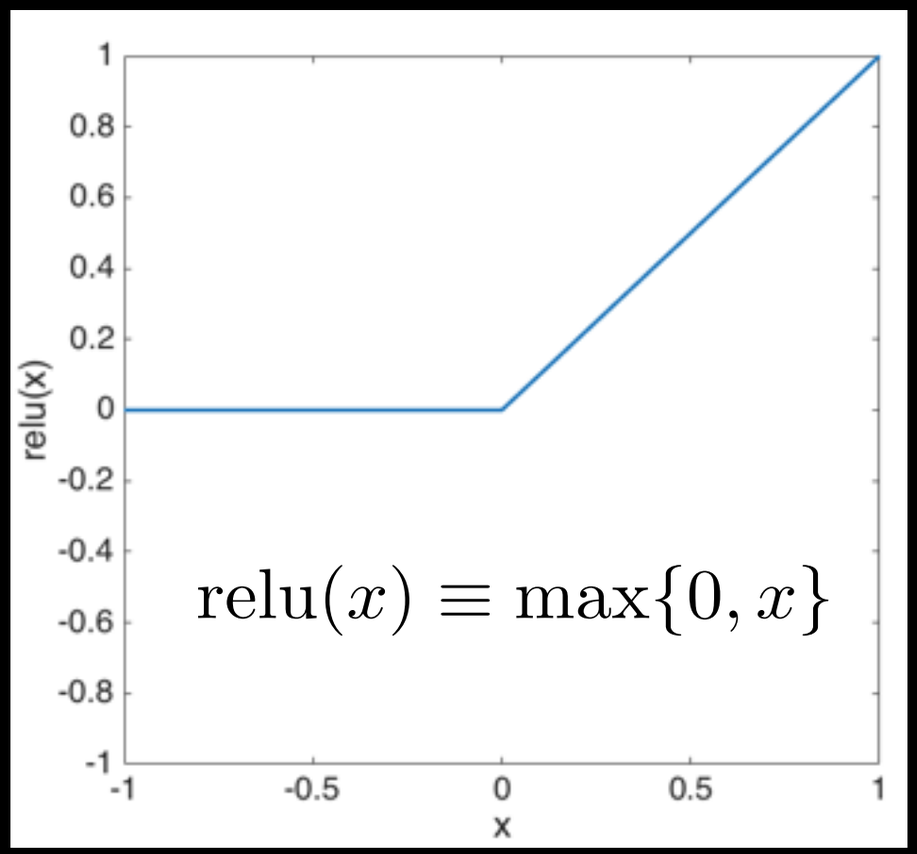
- Why stochastic gradient descent?
- **Why deep networks?**

Three salient assumptions

- A model with less critical points is easier to optimise than a model with more critical points
- A model with minima closer together in parameter space is easier to optimise than a model with minima far apart
- A model where the entropy of low energy minima is similar to high energy minima is easier to optimise than a model where low energy minima have much lower entropy

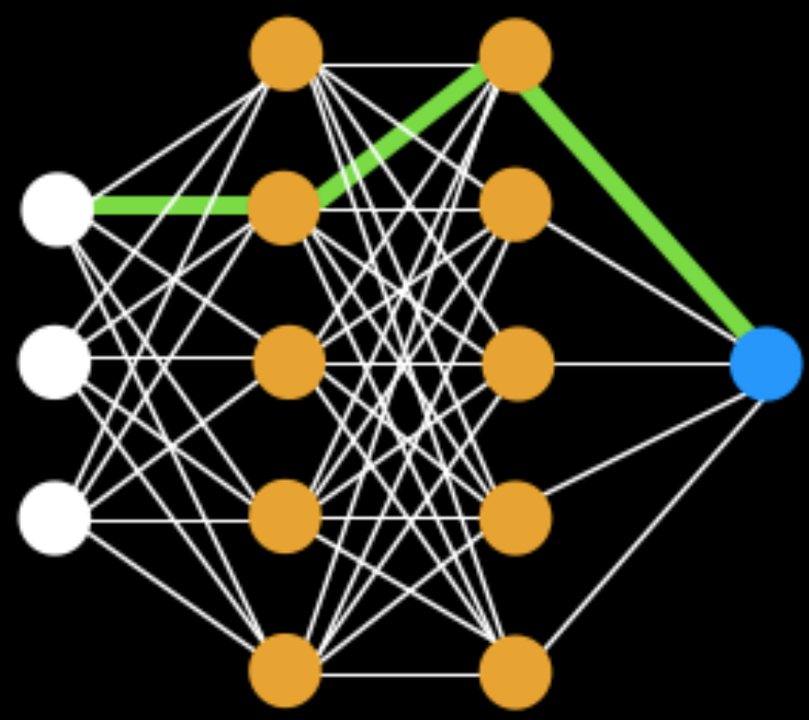
An analytical model of neural networks

$$y = \sigma(\mathbf{W}_H \sigma(\mathbf{W}_{H-1} \cdots \sigma(\mathbf{W}_2 \sigma(\mathbf{W}_1 \mathbf{x}))))$$



$$\text{relu}(x) \equiv \max\{0, x\}$$

Sum over p paths



A. Choromanska et al, AISTATS 2015

From neural network to a spin glass

$$L(\mathbf{w}) = \frac{1}{N}\sum_{i=1}^{N} |Y(\mathbf{x}_i, \mathbf{w}) - y_i|^2$$

$\mathbf{x}, y$ uncorrelated

…and other simplifying assumptions

$$\mathcal{H}_\Lambda(\mathbf{w}) = \frac{1}{\Lambda^{(H-1)/2}} \sum_{i_1,..,i_H=1}^{\Lambda} Z_{i_1,\ldots,i_H} \prod_{k=1}^{H} w_{i_k}$$

$$Z_{i_1,\ldots,i_H} \sim \mathcal{N}(0,1)$$

A. Choromanska et al, AISTATS 2015

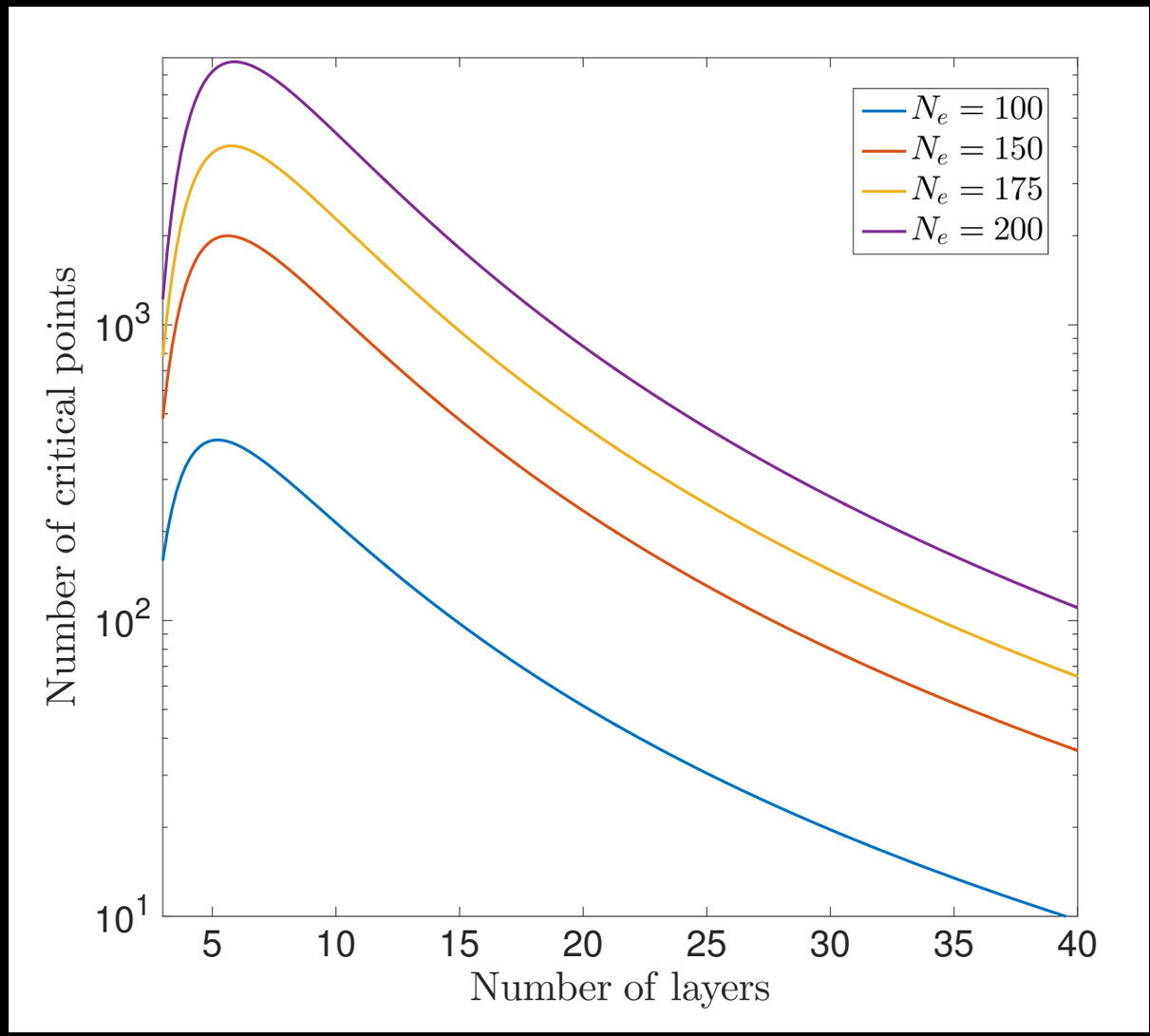
How many critical points are there?

$$\mathcal{N} = \frac{(H-1)^{\Lambda} - 1}{H-2}$$

The relevant limit is fixing the number of parameters whilst increasing network depth, c.f.

$$\Lambda = \frac{\sqrt{4N_e(H-1)+1}+1}{2(H-1)}$$

$$\sim \sqrt{\frac{N_e}{H}}$$ (number of nodes per layer)



S. Becker, Y. Zhang, **AAL**, arXiv:1808.00408 (2018)

How far are the critical points from each other?

- $\mathrm{Crt}(-\infty, \mathcal{E})$: the set of critical points for which the loss function takes values in $(-\infty, \Lambda\mathcal{E})$

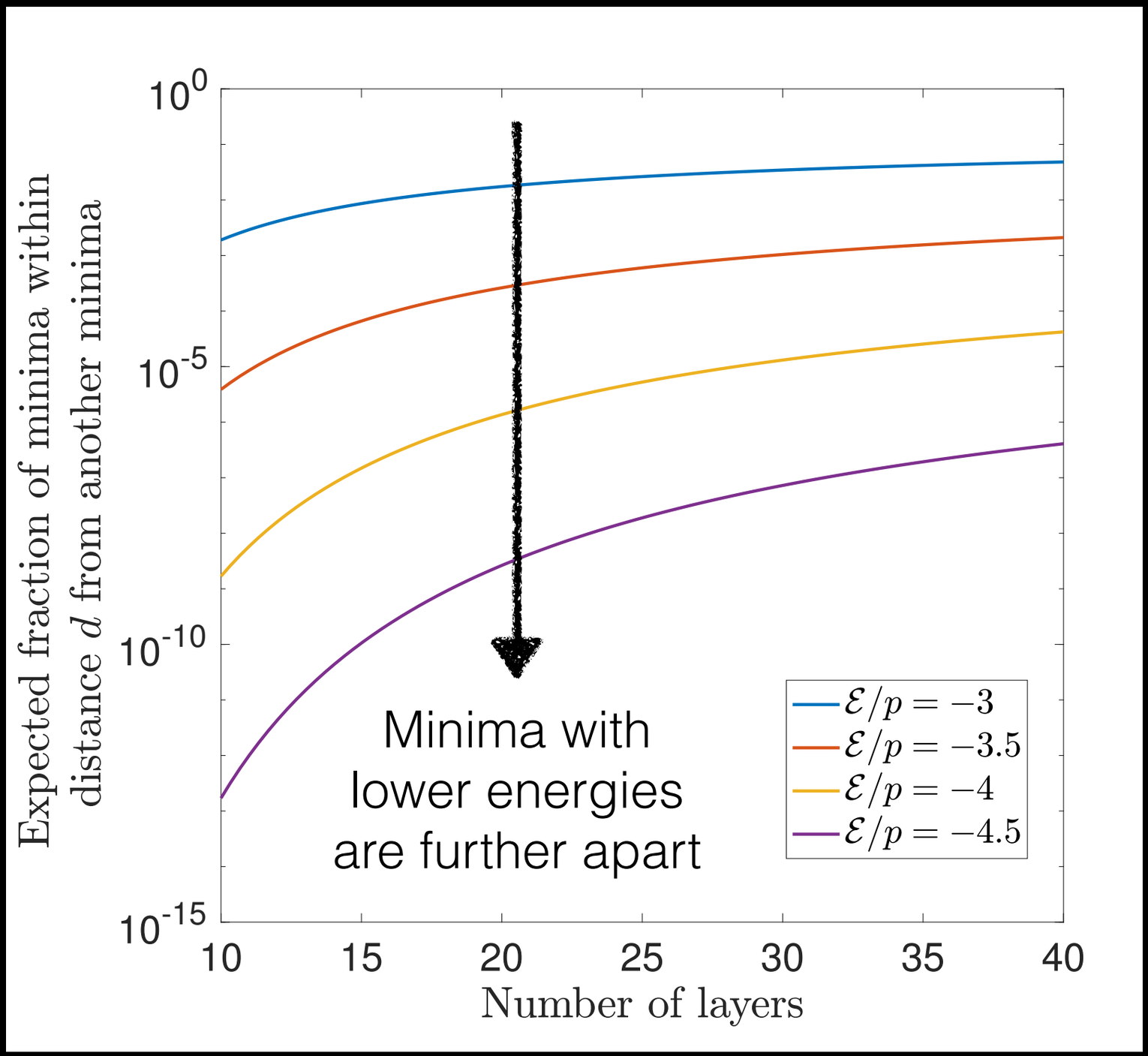
- $[\mathrm{Crt}((\infty, \mathcal{E}), d)]_2$: the set of pairs of points in $\mathrm{Crt}(-\infty, \mathcal{E})$ for which the distance between them is less than d

- We can prove a bound of the form

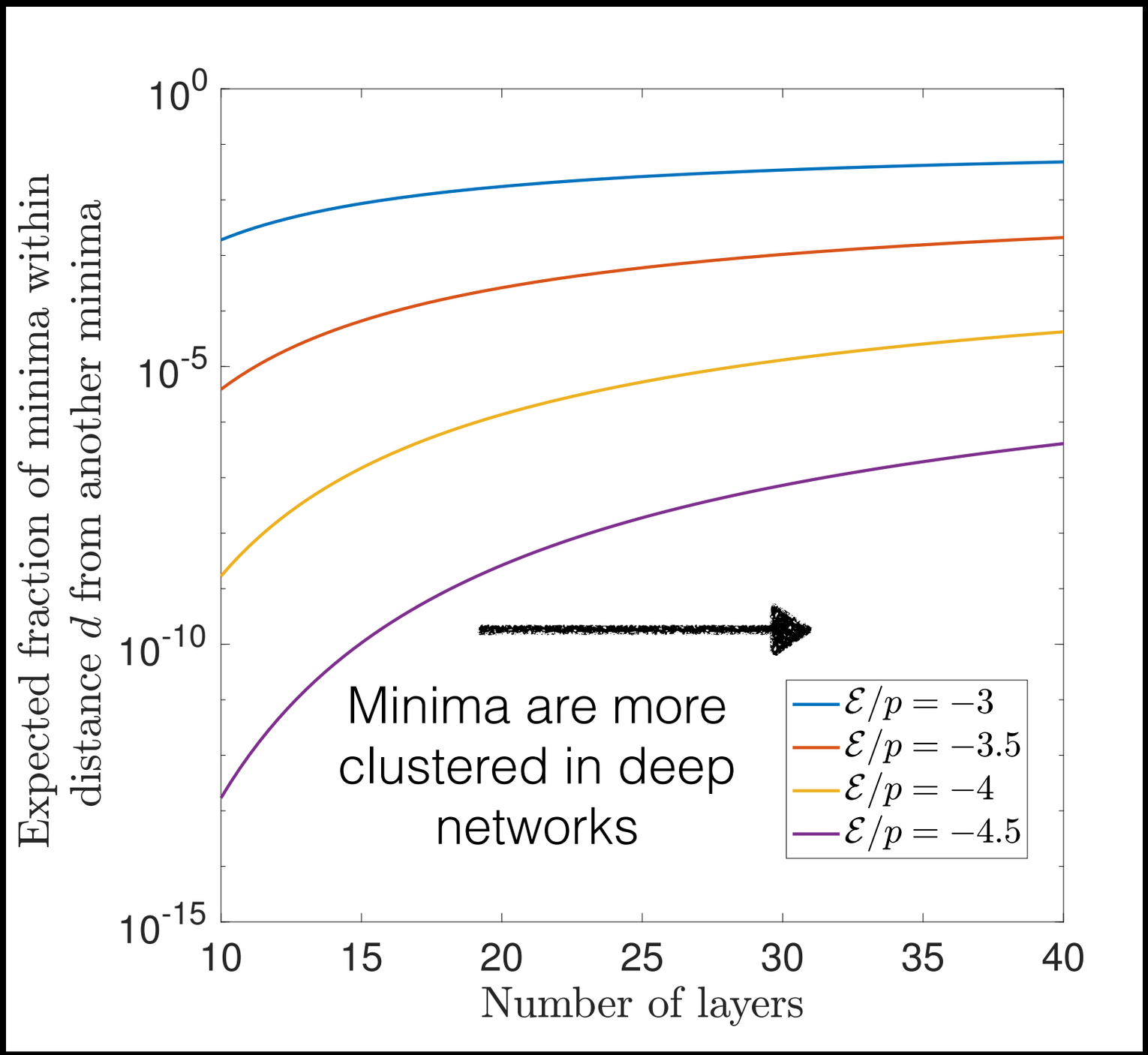
$$\limsup_{\Lambda\to\infty} \frac{1}{\Lambda} \log\left(\frac{\mathbb{E}\,|[\mathrm{Crt}((-\infty, \mathcal{E}), d)]_2|}{\mathbb{E}\,|\mathrm{Crt}(-\infty, \mathcal{E})|}\right) \leq \sup_{r\in I} \sup_{v\in(-\infty, \mathcal{E}/p)} \Psi_H(r, v, \mathcal{E})$$

S. Becker, Y. Zhang, **AAL**, arXiv:1808.00408 (2018)

How far are the critical points from each other?



How far are the critical points from each other?



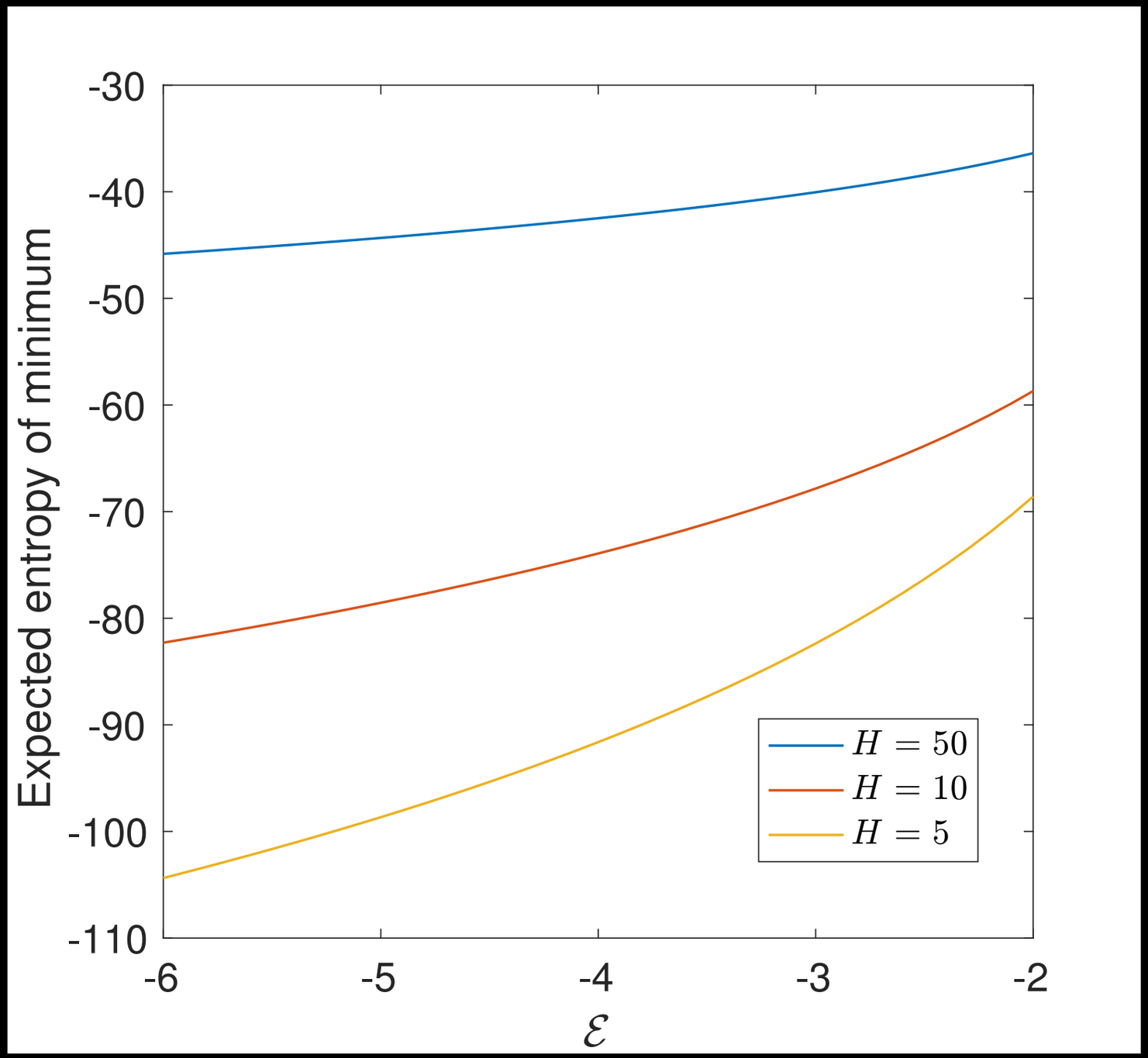
How does the energy of a minimum scale with its basin volume?

- We operationalise basin volume by the log determinant of the Hessian (harmonic approximation)
- We can prove

$$\mathbb{E}\left(S(\mathrm{Hess}\,\mathcal{L})|\Lambda\mathcal{E}\right) = -(\Lambda-1)\log(p) + \tfrac{\Lambda-1}{2}\log\left(\tfrac{\Lambda}{2(\Lambda-1)H(H-1)}\right)$$
$$-\tfrac{\Lambda-1}{\pi}\int_{-\sqrt{2}}^{\sqrt{2}}\log\left|\sigma\sqrt{\tfrac{\Lambda}{\Lambda-1}}\tfrac{\mathcal{E}}{p}-t\right|\sqrt{2-t^2}\;dt$$

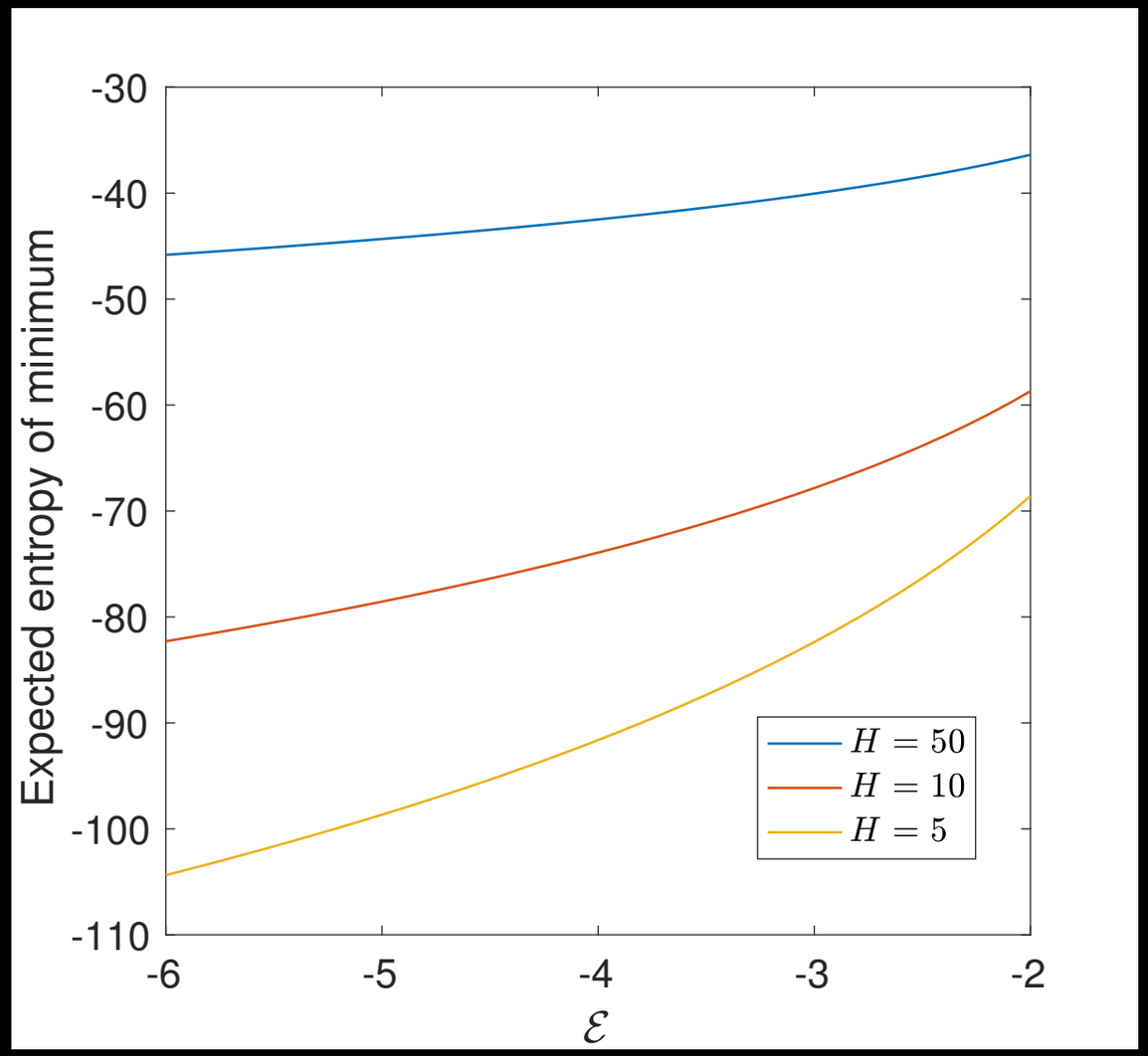
S. Becker, Y. Zhang, **AAL**, arXiv:1808.00408 (2018)

How does the energy of a minimum scale with its basin volume?



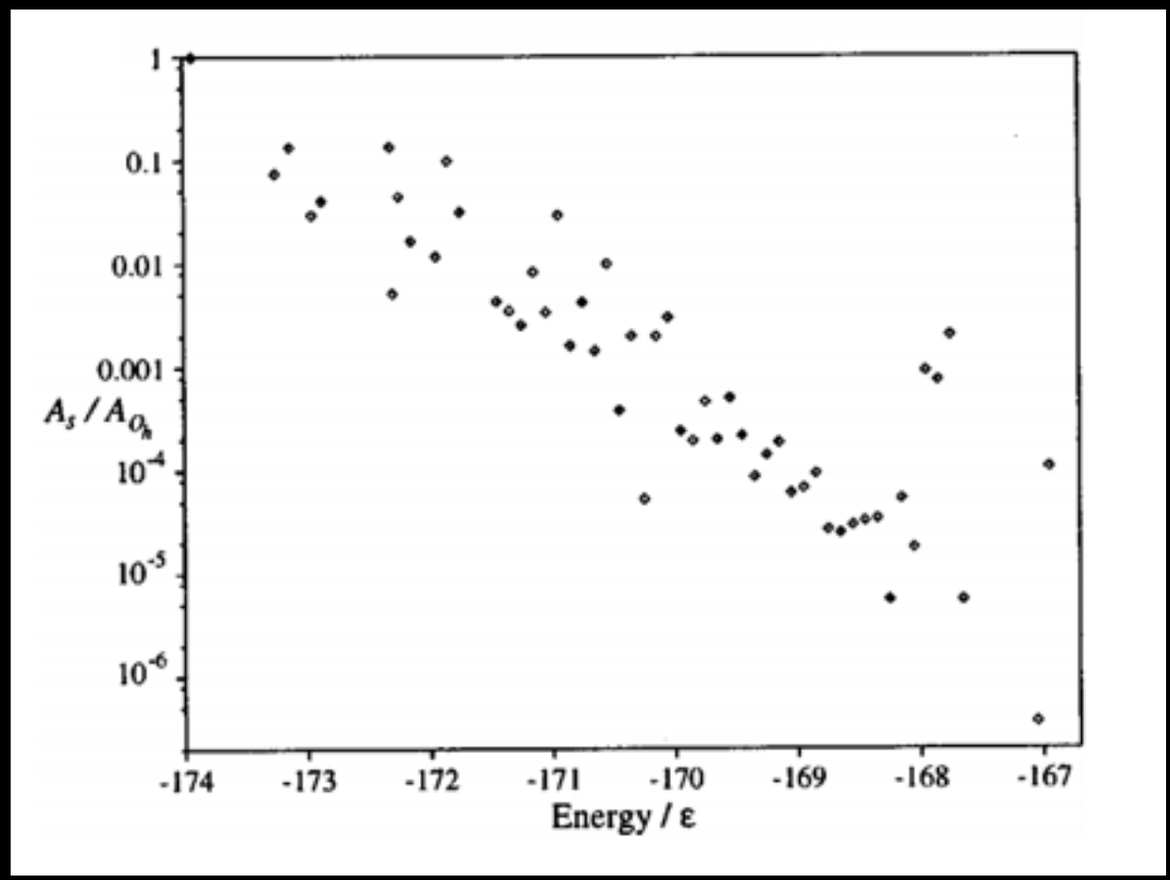
- Energy-entropy competition!
- This competition is eased for deeper networks

How does the energy of a minimum scale with its basin volume?



VS

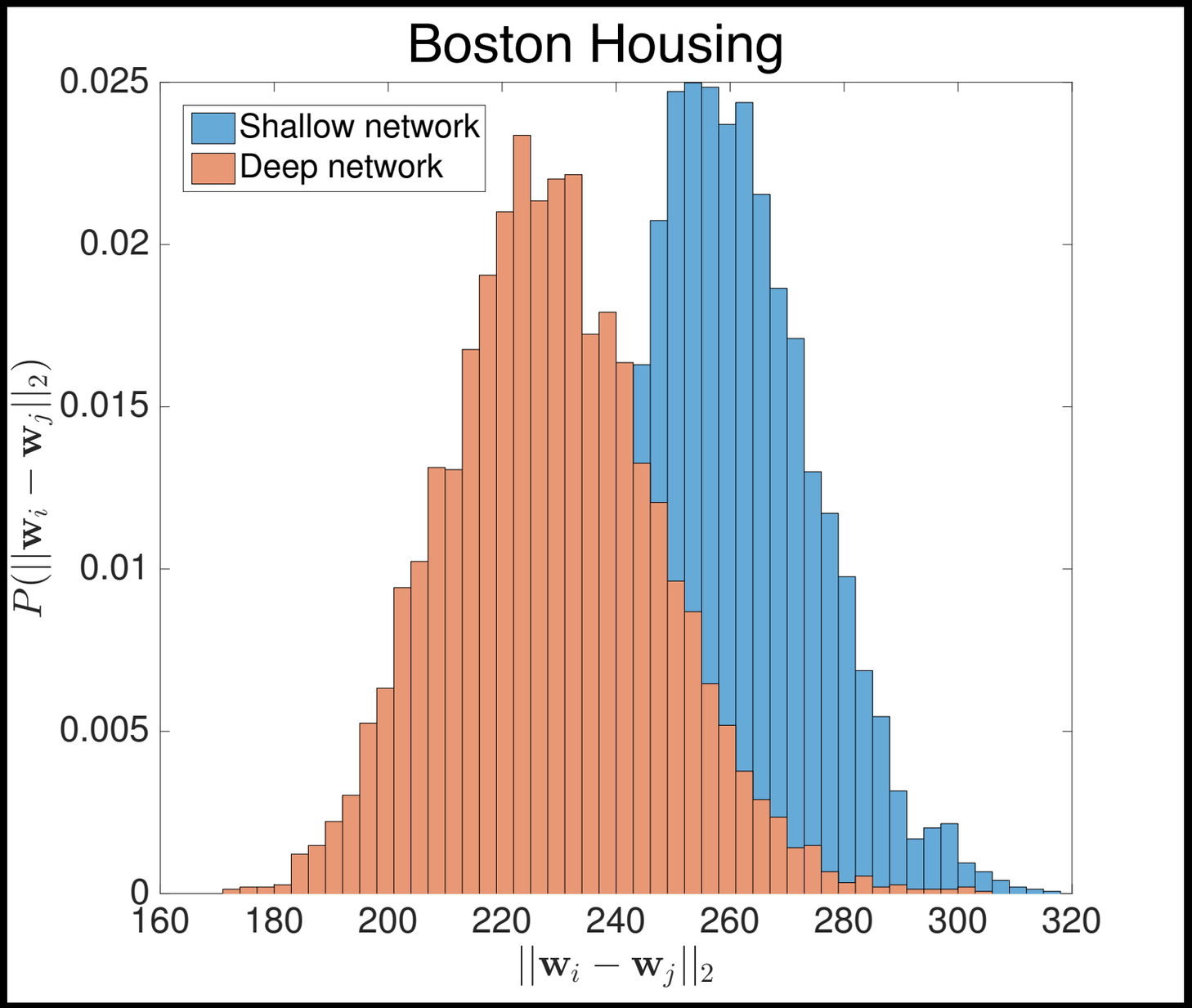
LJ_{55}



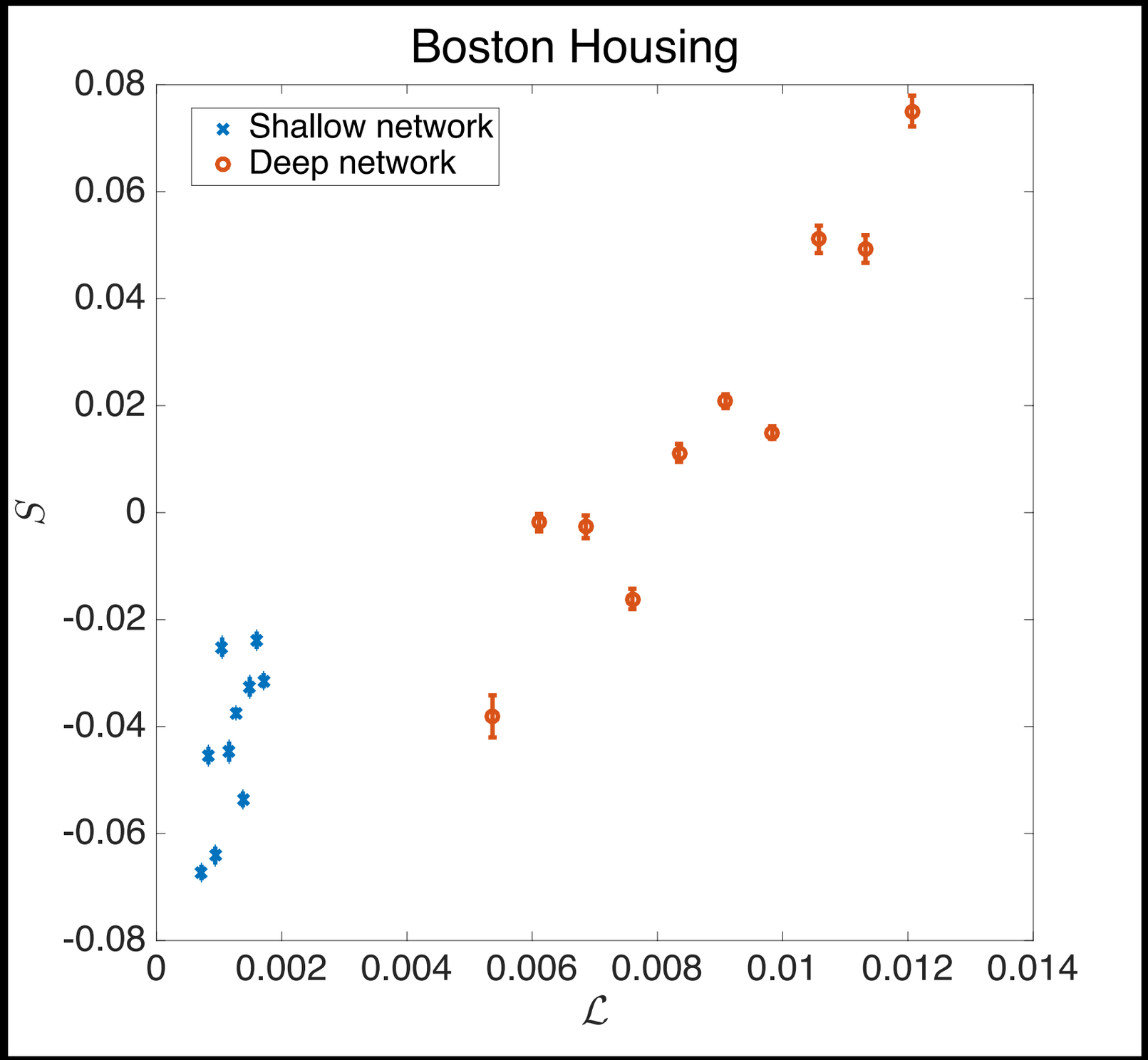
J. P. K. Doye, D. J. Wales and M. Miller, JCP, 109, 8143 (1998)

Numerical experiments

We ran SGD followed by full gradient descent starting from random initial conditions for a range of standard datasets



Numerical experiments



Conclusion

- Modelling undersampling noise and disentangling correlations vs causations allow us to build accurate models for drug discovery

- We can understand deep neural networks and stochastic gradient descent using thermodynamics

- Curiously, both questions are related to spin glasses, random matrices, and entropy!

Acknowledgements

- Bioactivity prediction
  - Soma Turi (Cambridge)
- DNN and spin glass
  - Yao Zhang, Simon Becker (Cambridge)
- CHRM1 agonist discovery
  - David Price, Xinjun Hou, Chris Butler and Joy Yang (Pfizer)
- Random matrix theory
  - Michael Brenner (Harvard)
  - Lucy Cowell (Cambridge)

THE WINTON PROGRAMME FOR THE
Physics of Sustainability

We are hiring!

contact me: aal44@cam.ac.uk