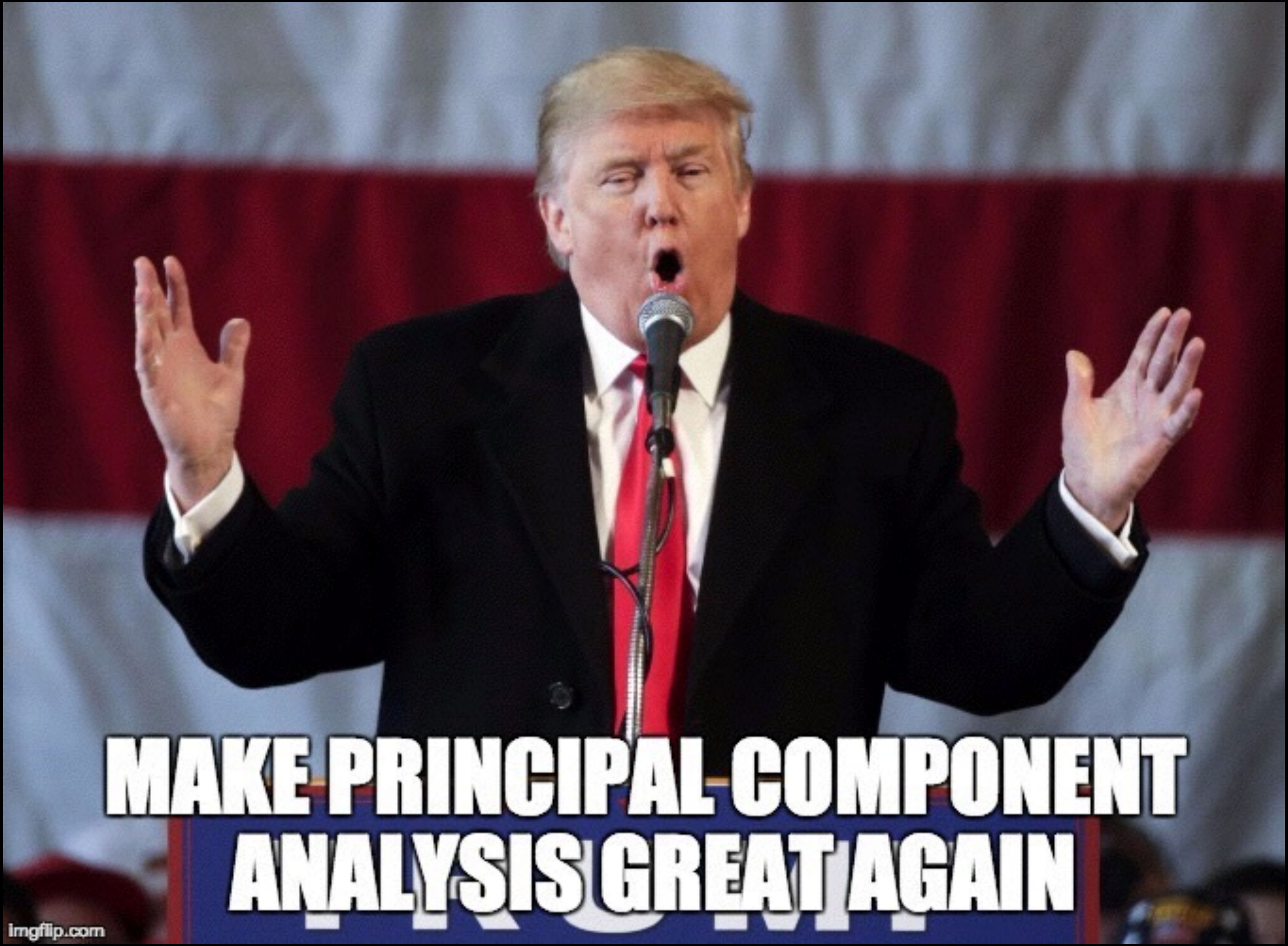
Exploring chemical space using random matrix theory

Alpha Lee

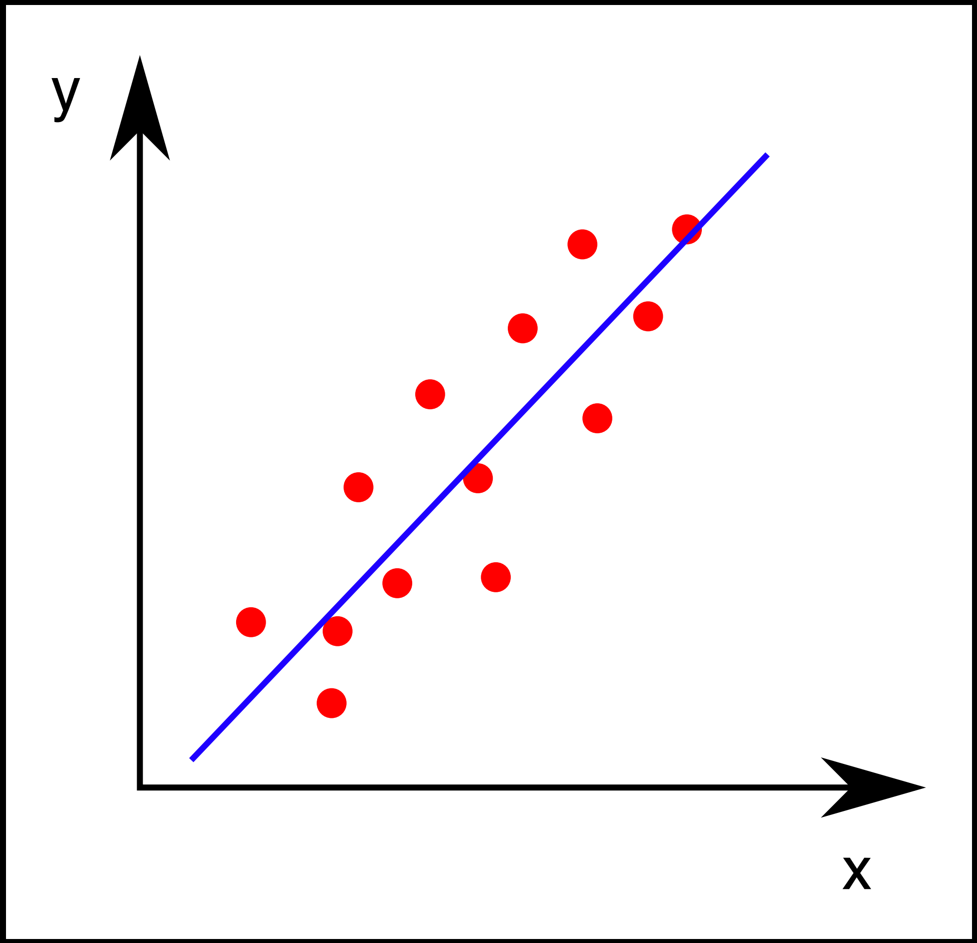
Department of Physics, University of Cambridge
aal44@cam.ac.uk
www.alpha-lee.com


MAKE PRINCIPAL COMPONENT
ANALYSIS GREAT AGAIN
imgflip.com

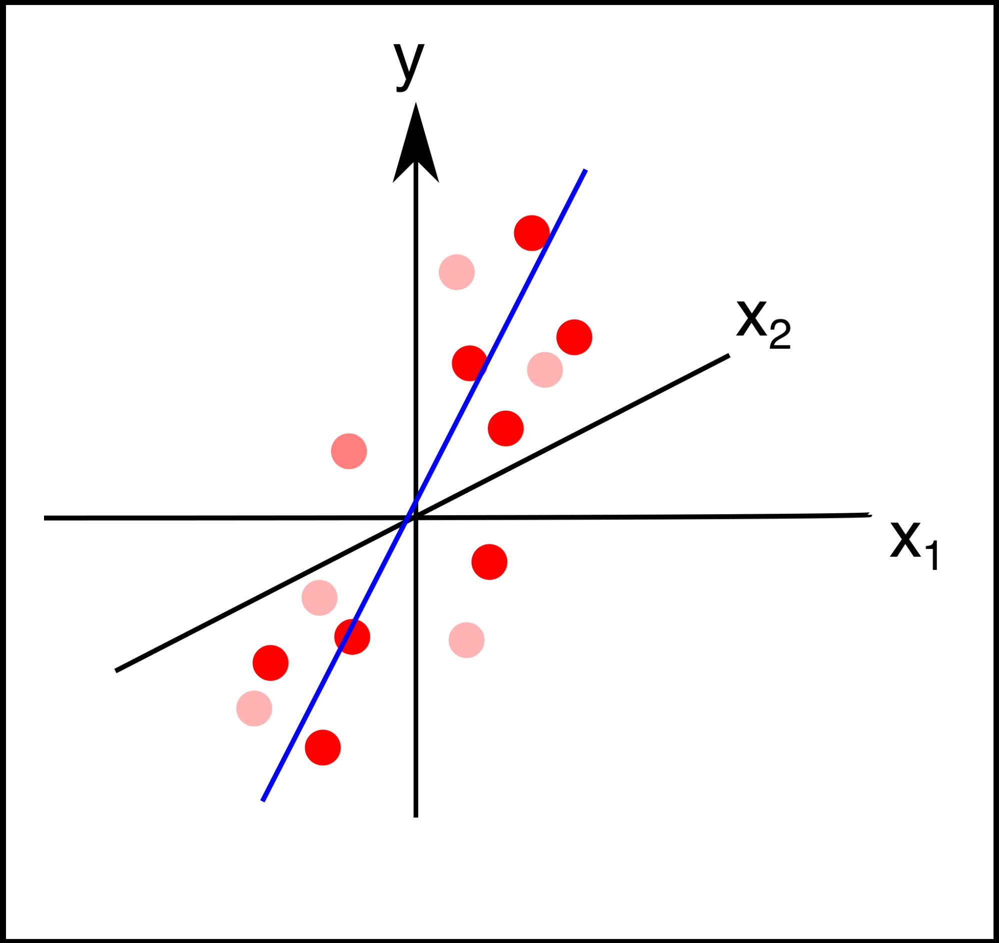
Chemical space is high dimensional

$p = \text{number of variables}$

$n = \text{number of samples}$

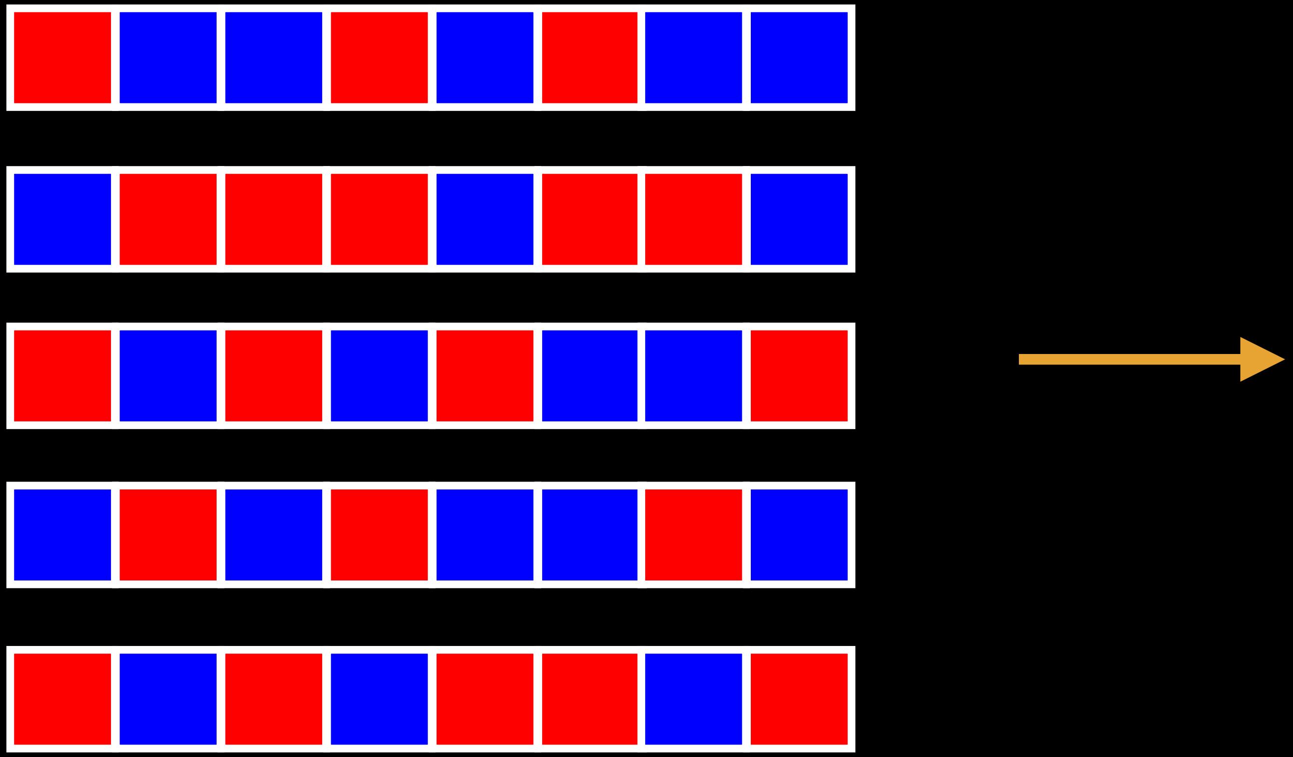


$$p = O(1)$$
$$p/n \to 0$$



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

Random matrix theory

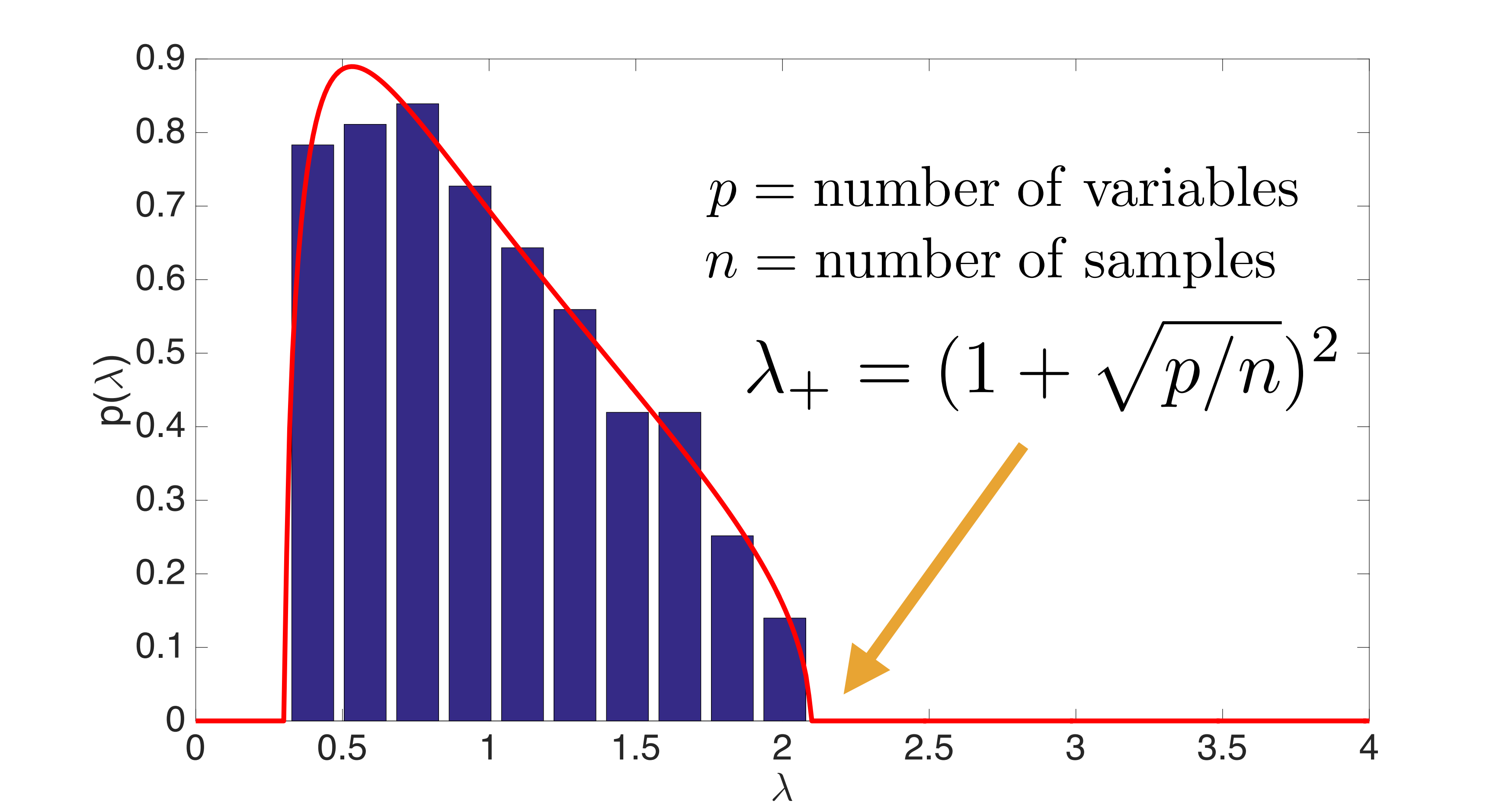


$$C_{ij} = \frac{\langle f_i f_j \rangle - \langle f_i \rangle \langle f_j \rangle}{\sigma_i \sigma_j}$$

E. T. Jaynes, *Physical Review*, 106, 620 (1957)
V. A. Machenko, L. A. Pastur, *Math. USSR Sb.*,1, 457 (1967)

The null model

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



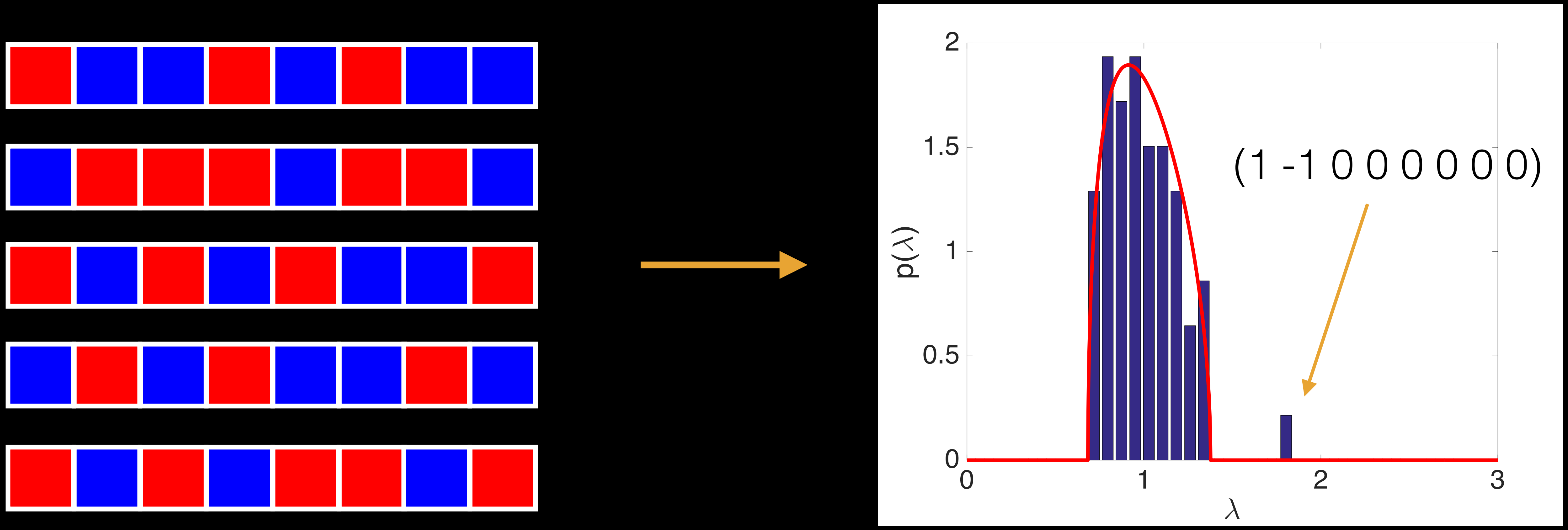
$p =$ number of variables

$n =$ number of samples

$$\lambda_+ = (1 + \sqrt{p/n})^2$$

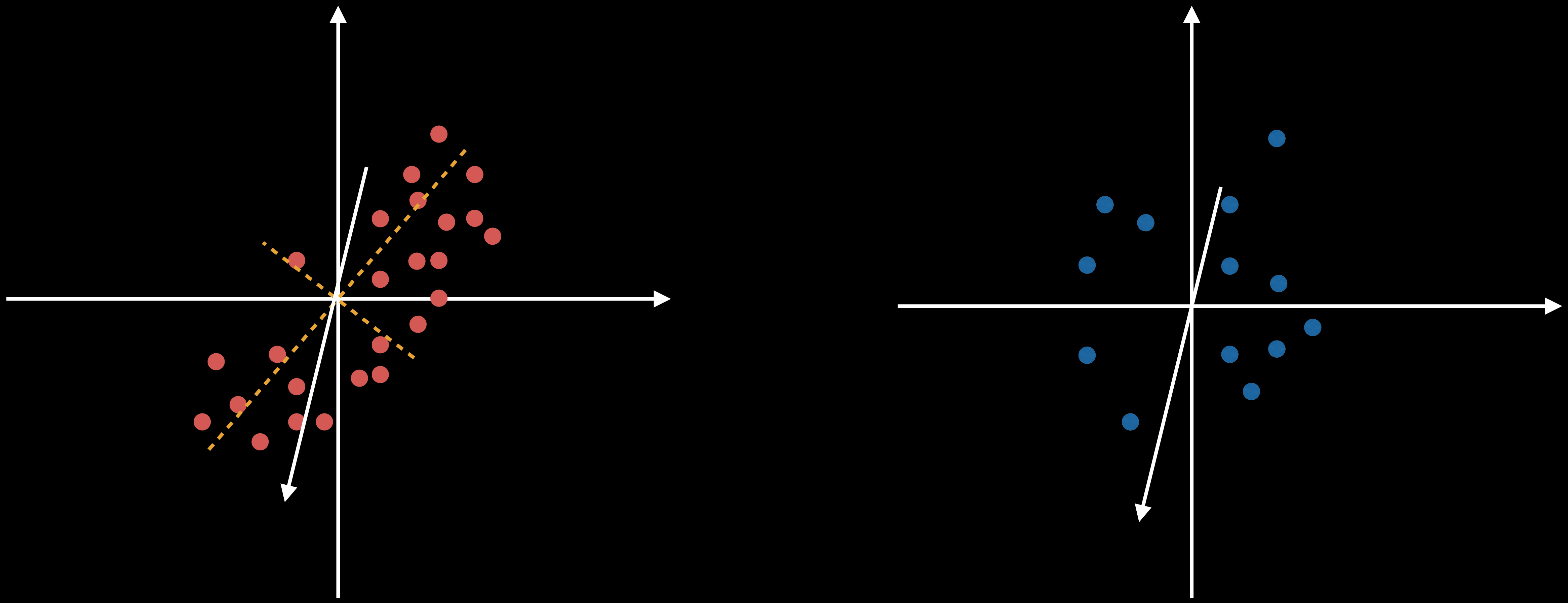
E. T. Jaynes, *Physical Review*, 106, 620 (1957)
V. A. Machenko, L. A. Pastur, *Math. USSR Sb.*,1, 457 (1967)

Random matrix theory



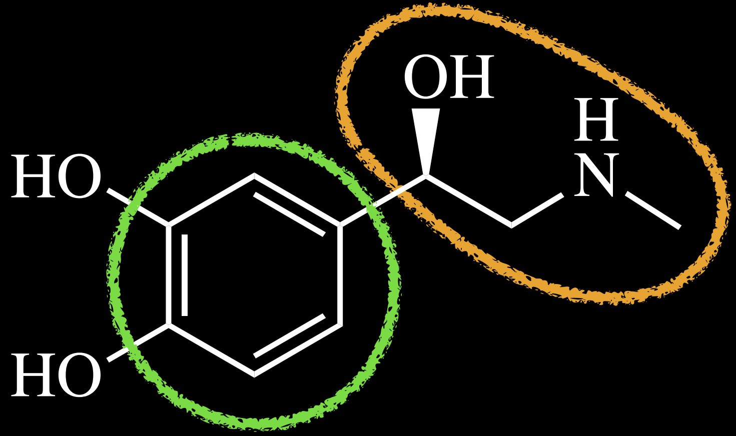
E. T. Jaynes, *Physical Review*, 106, 620 (1957)
V. A. Machenko, L. A. Pastur, *Math. USSR Sb.*,1, 457 (1967)

Connection to PCA

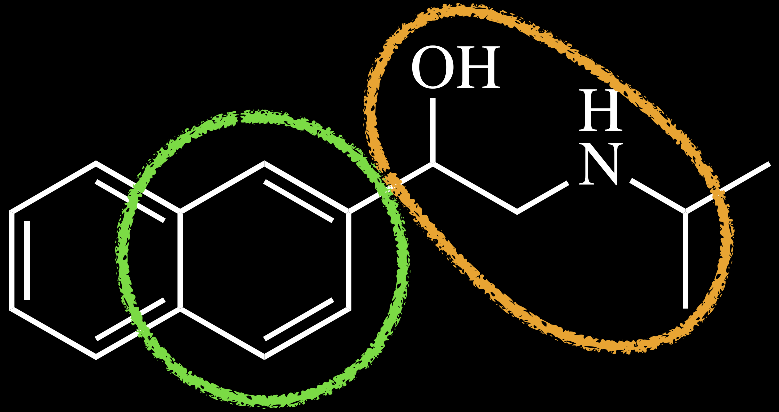


Chemical similarity

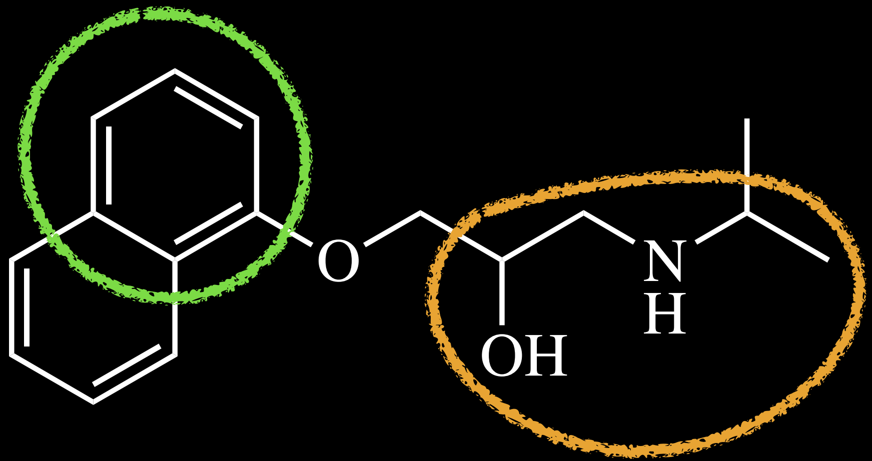
Suppose we want to design an ADRB1 antagonist



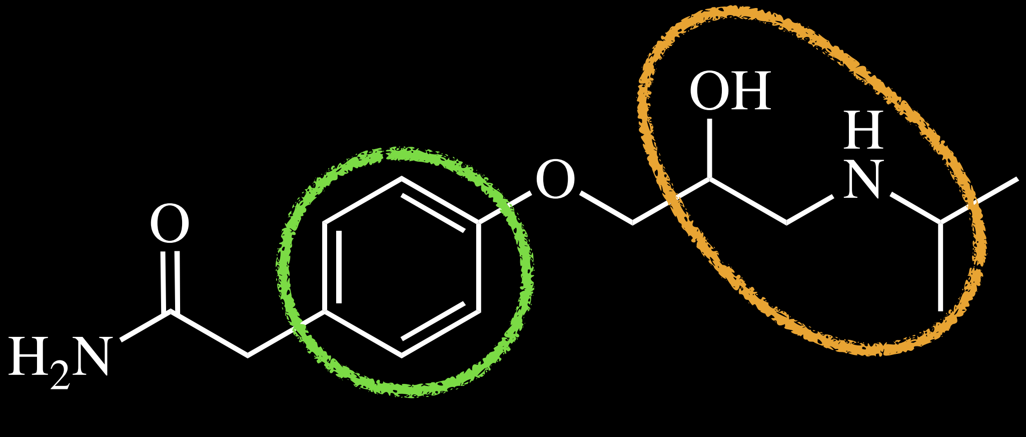
Adrenaline



Pronethalol



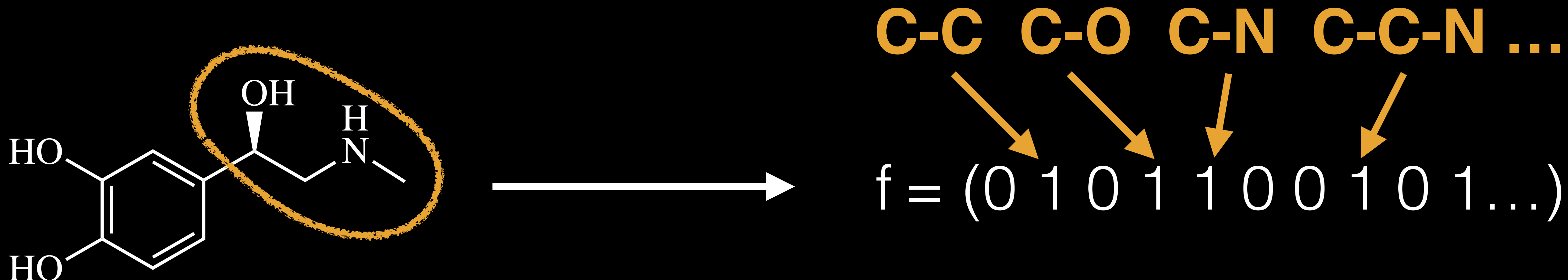
Propranolol



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?

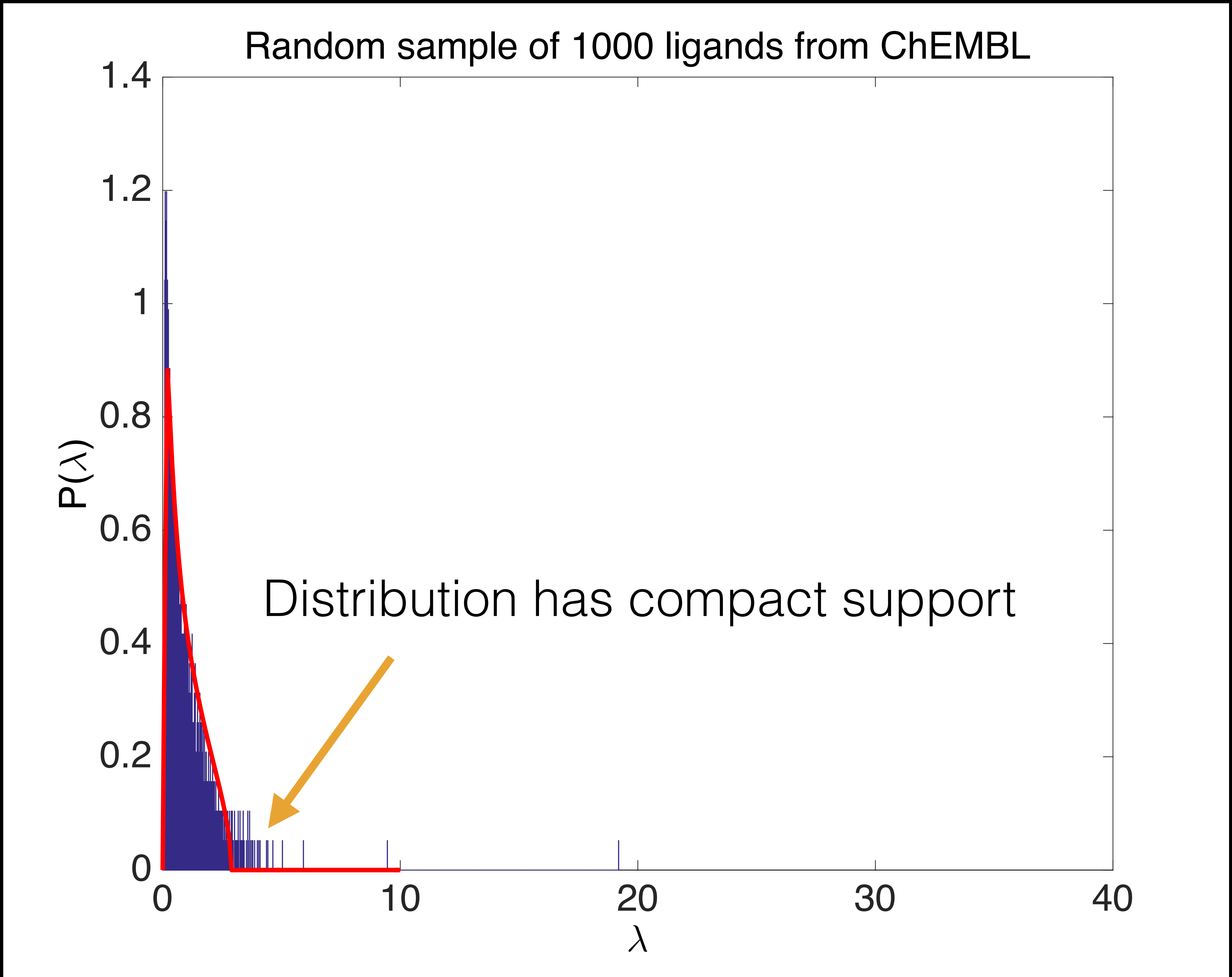


C-C C-O C-N C-C-N …

f = (0 1 0 1 1 0 0 1 0 1…)

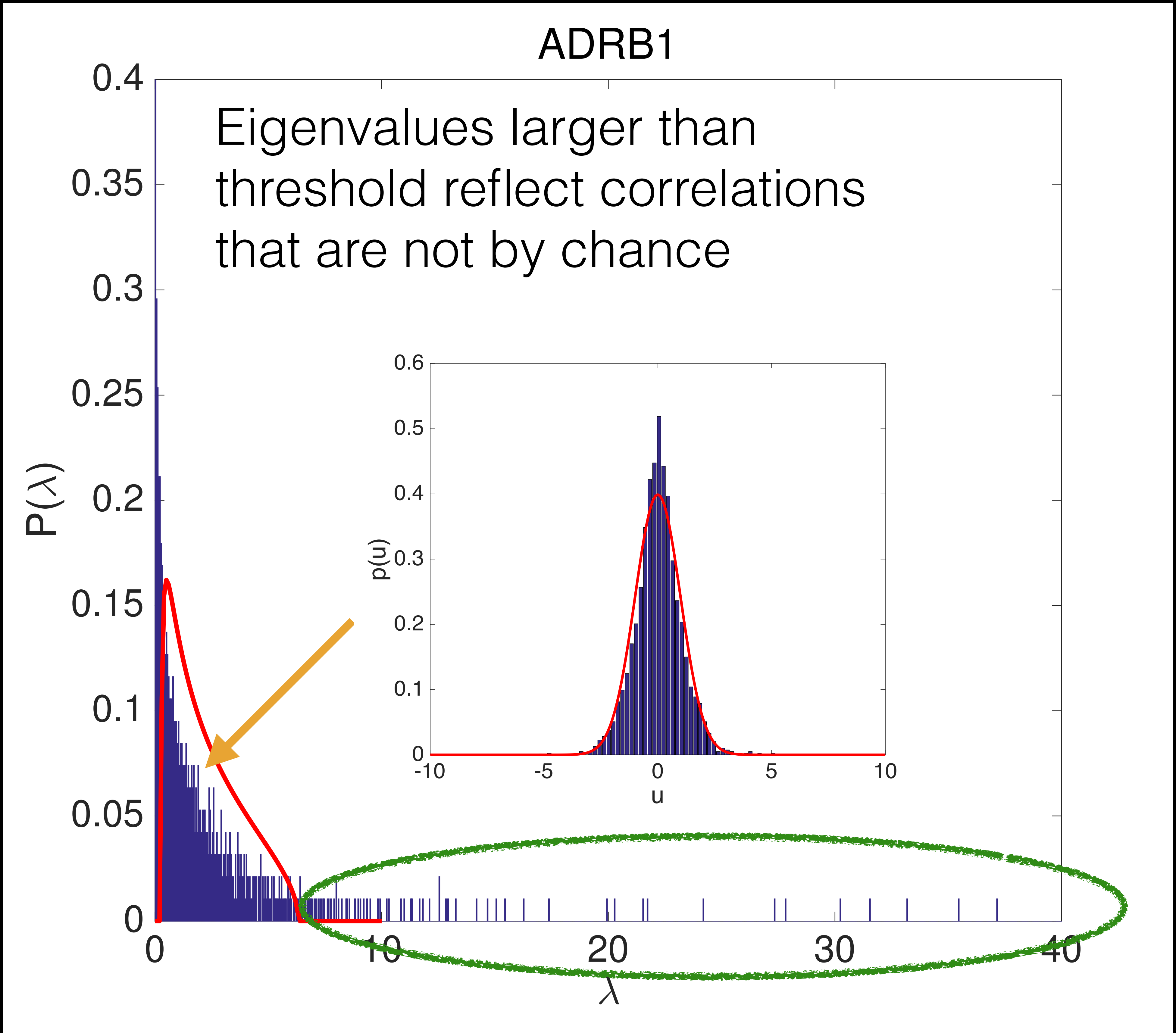
- Intuitively, there are only a few combinations chemical bonds (variables in the vector) that are important
- Many variables but often not many samples - data corrupted by finite sampling noise
- How do we get rid of the noise?

H. L. Morgan, *J. Chem. Doc.*, 5, 107 (1965)
Daylight Chemical Information Systems, Inc (since 1987)
A. Bender et al, *J. Chem. Inf. Comput. Sci.*, 44, 170 (2004)
D. Rogers and M. Hahn, *J. Chem. Inf. Model.*, 50, 742 (2010)



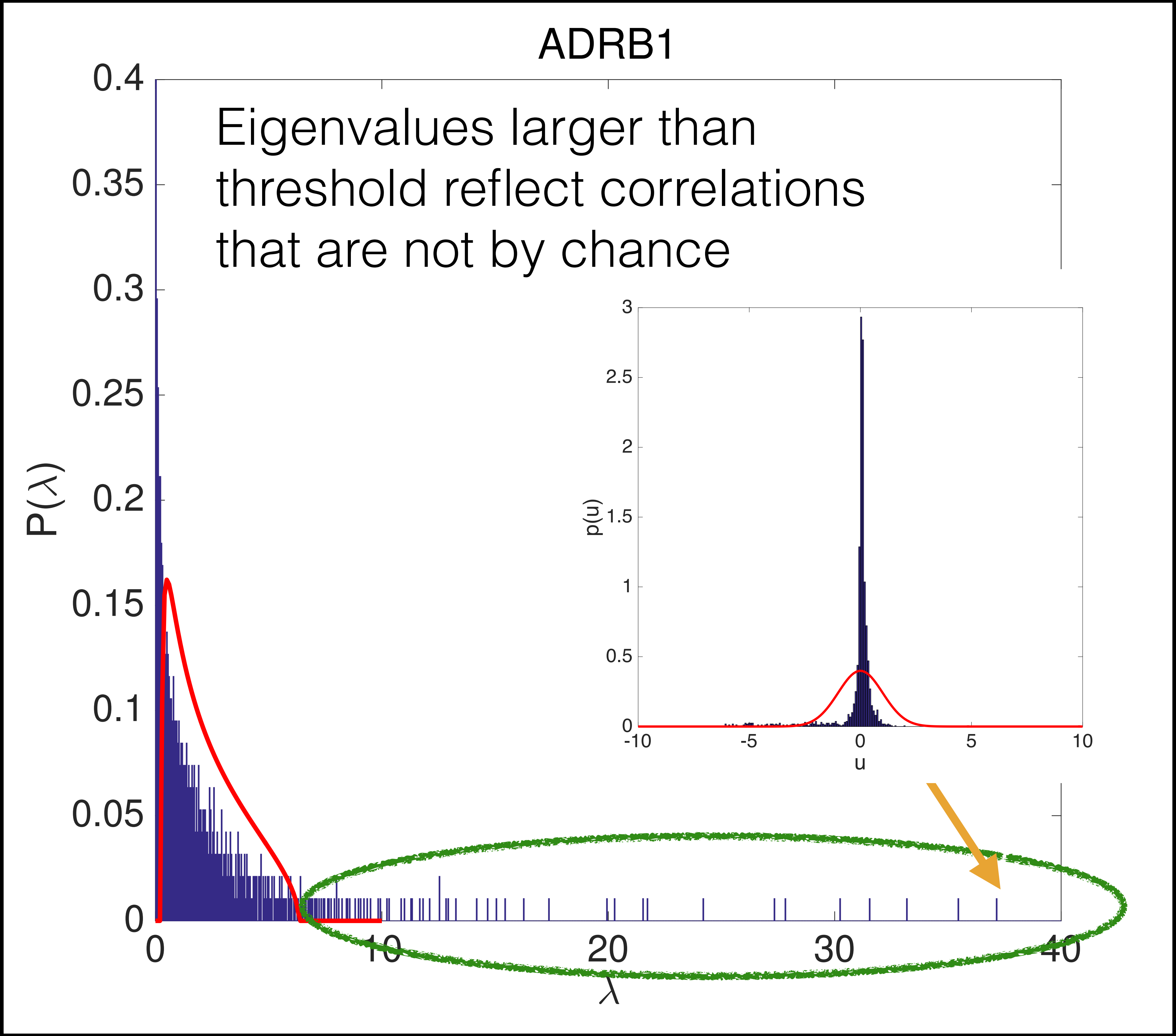
Distribution has compact support

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



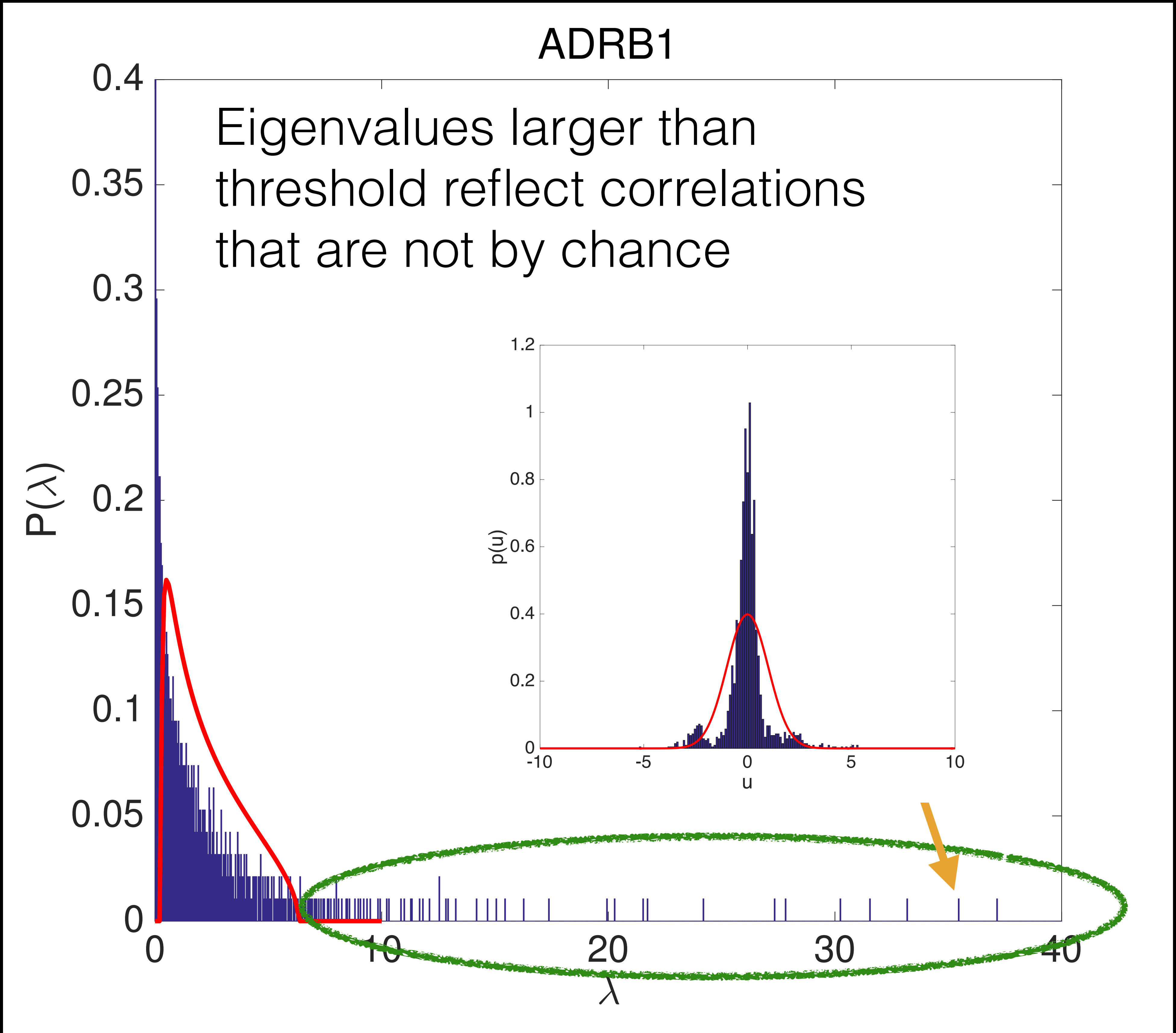
Eigenvalues larger than threshold reflect correlations that are not by chance

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



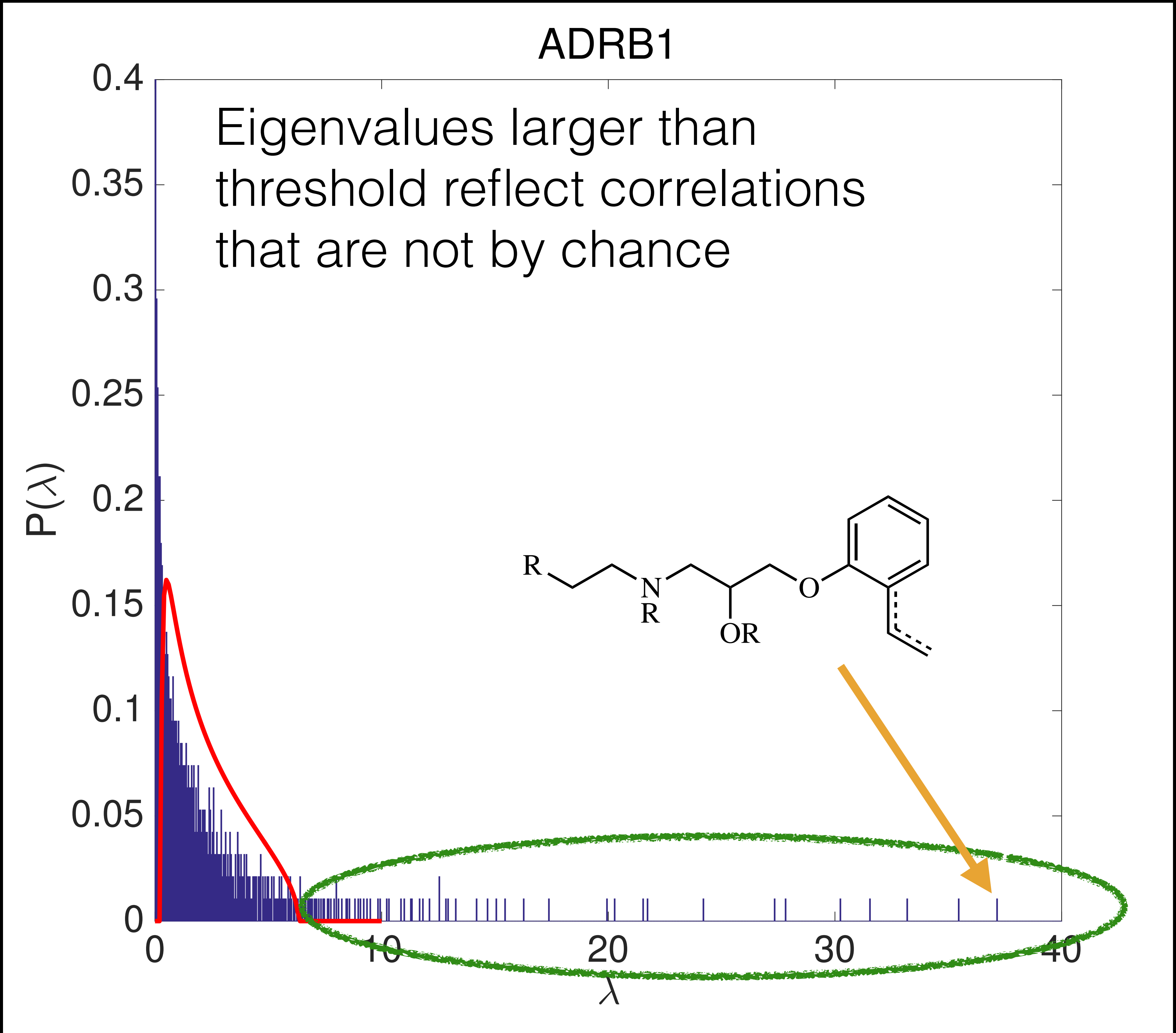
Eigenvalues larger than threshold reflect correlations that are not by chance

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



Eigenvalues larger than threshold reflect correlations that are not by chance

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



Eigenvalues larger than threshold reflect correlations that are not by chance

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

Predicting protein-ligand affinity

1. Let $\{v_i\}_{i=1}^q$ be eigenvectors with eigenvalues above threshold

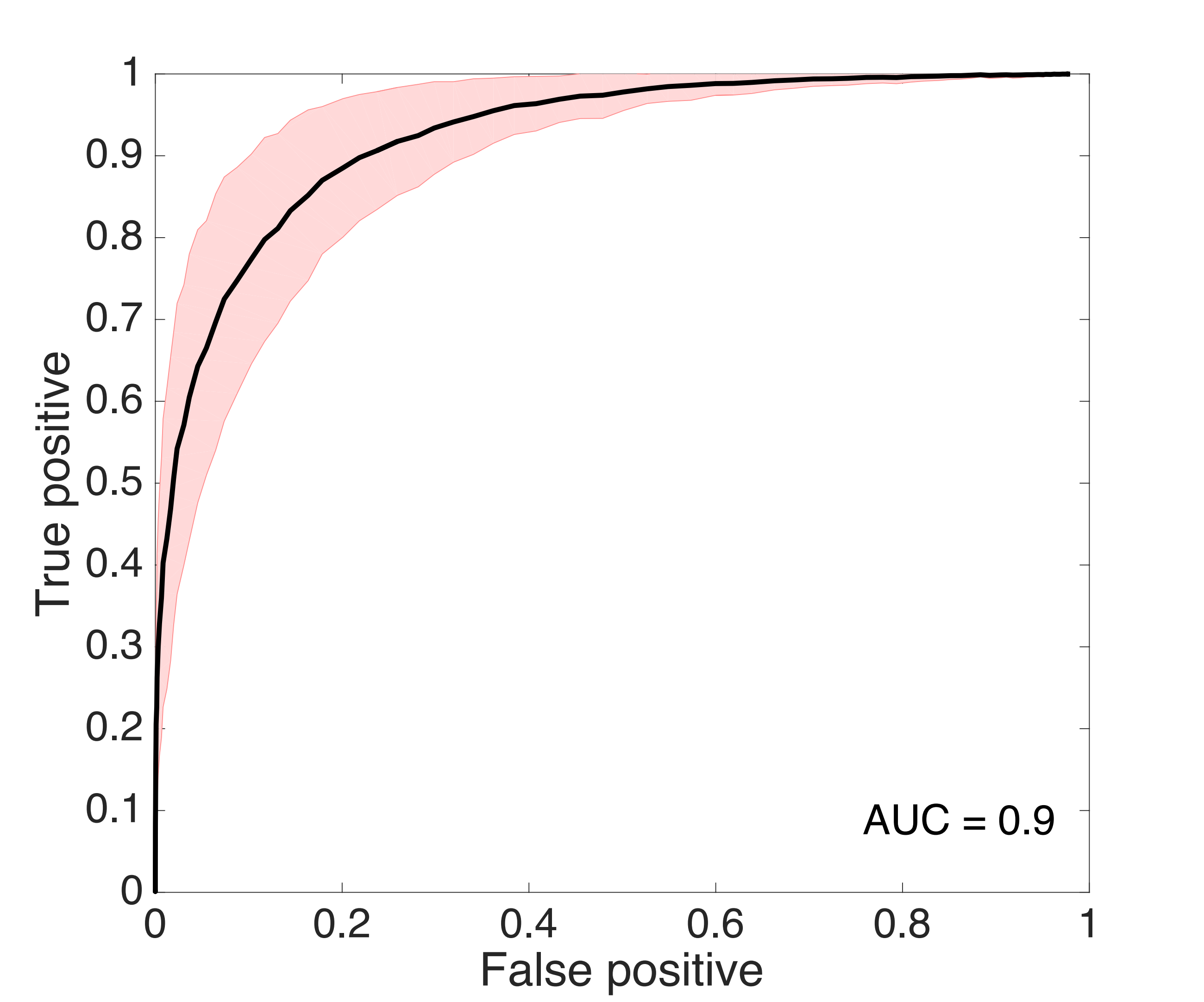
2. Convert unknown molecules into vector u

3. Compute how “close” is u to $\mathrm{span}\{v_i\}_{i=1}^q$

Criterion for binding $$\left\| u - \sum_{i=1}^{q} (u \cdot v_i) v_i \right\|_2 < \epsilon$$

c.f. M. Turk, A. Pentland, *J. Cognitive Neurosci.*, 3, 71 (1991)

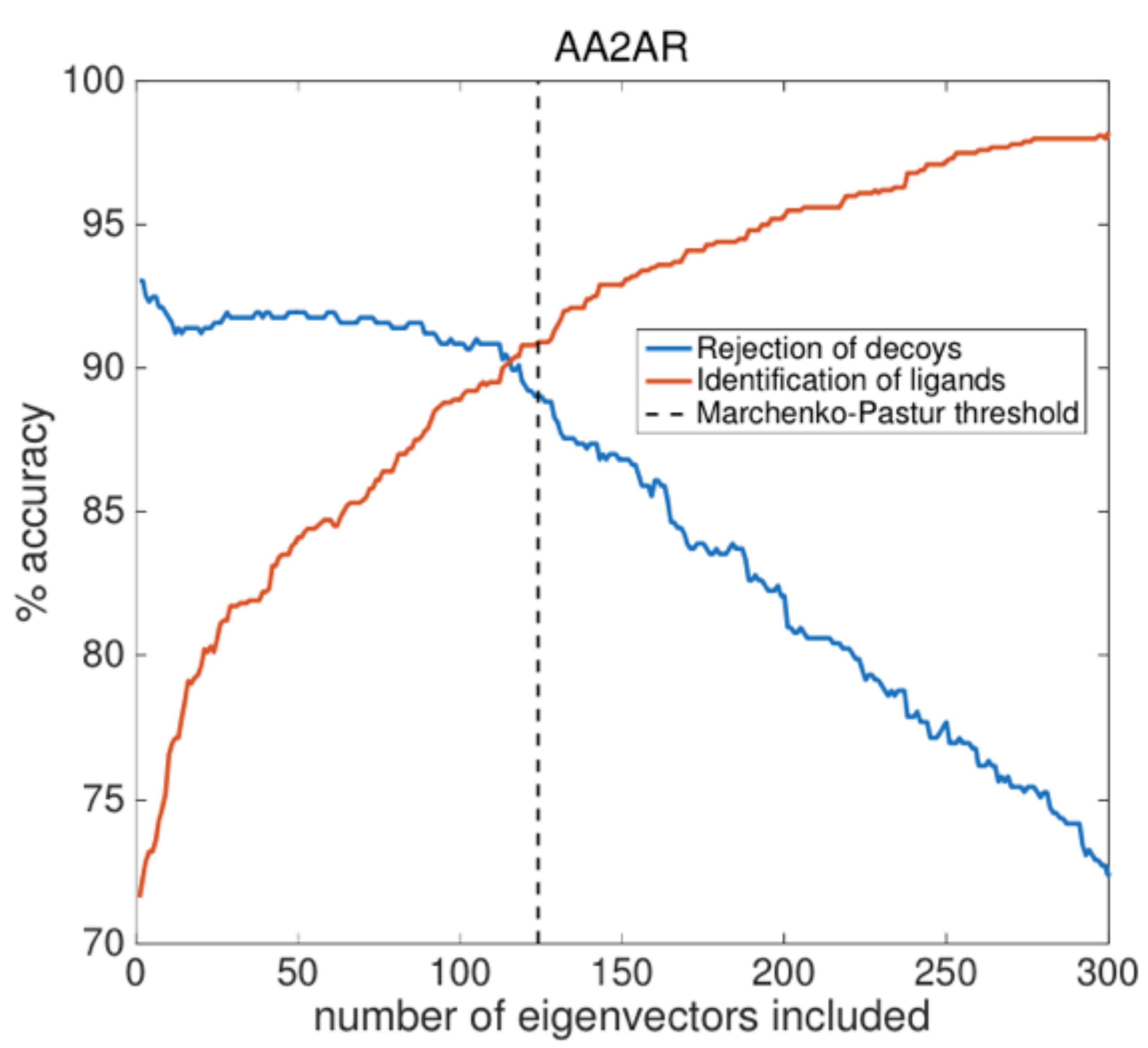
Performance of classification algorithm



The algorithm outperforms all algorithms that we are aware of

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)

Many shades of grey: Turning classification into regression

- It is costly to do measurements precisely!
- It is often much easier to measure whether the property of a compound is above/below a threshold

molecule$_1$	soluble
molecule$_2$	soluble
molecule$_3$	insoluble
molecule$_4$	insoluble
…	

+

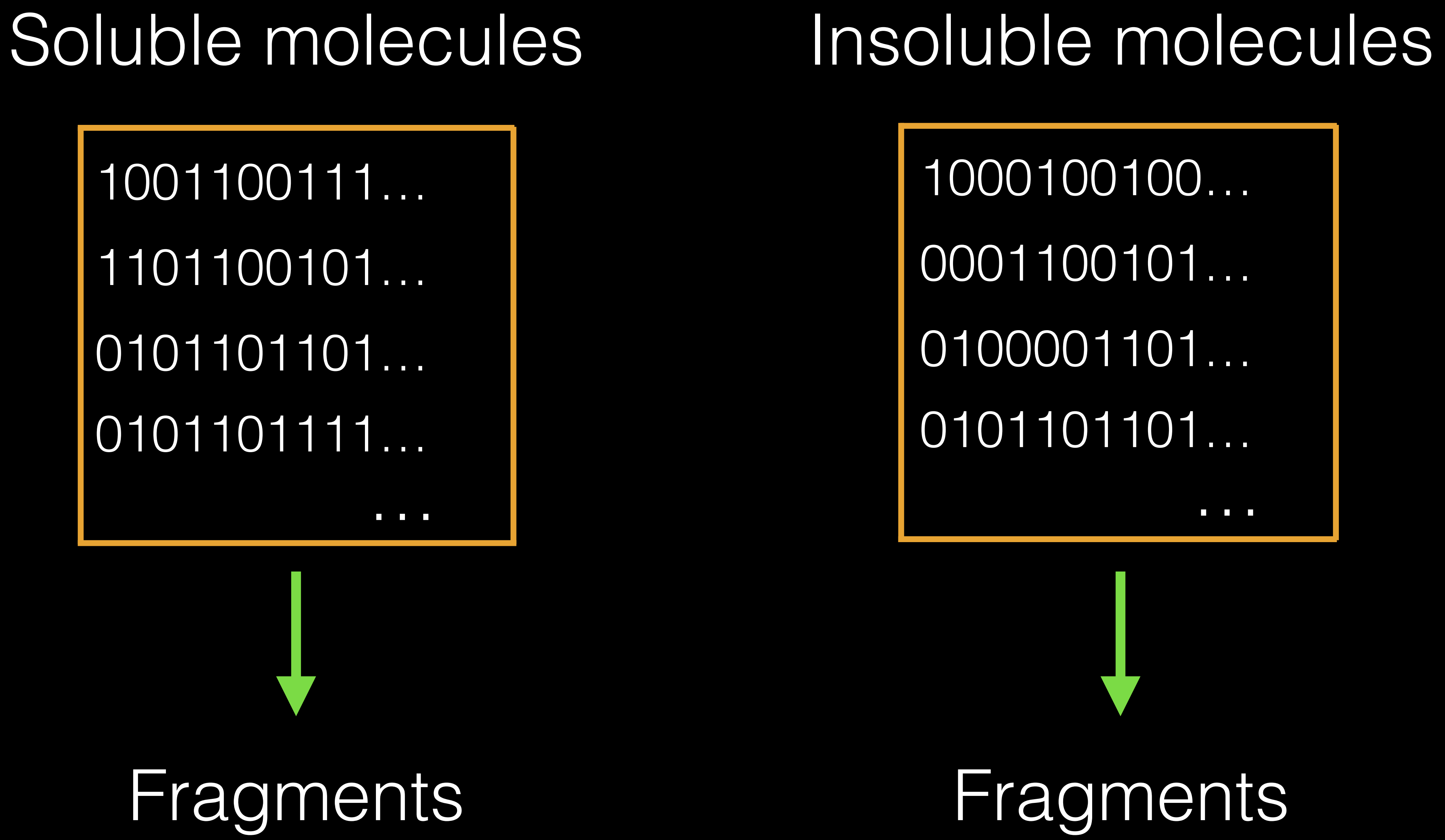
molecule$_j$	s = 1 mol/L
molecule$_{j+1}$	s = 0.1 mol/L
molecule$_{j+2}$	s = 0.5 mol/L
molecule$_{j+3}$	s = 0.01 mol/L
…	

= ?

c.f. A. Llinàs, R. C. Glen and J.M. Goodman, *J. Chem. Inf. Model.*, 48, 1289 (2008)

Back to correlation analysis

We know that there are chemical functional groups contributing to a molecule being soluble/insoluble



Soluble molecules

1001100111…

1101100101…

0101101101…

0101101111…

…

Fragments

Insoluble molecules

1000100100…

0001100101…

0100001101…

0101101101…

…

Fragments

Combining imprecise and precise measurements

Posit a quadratic model:

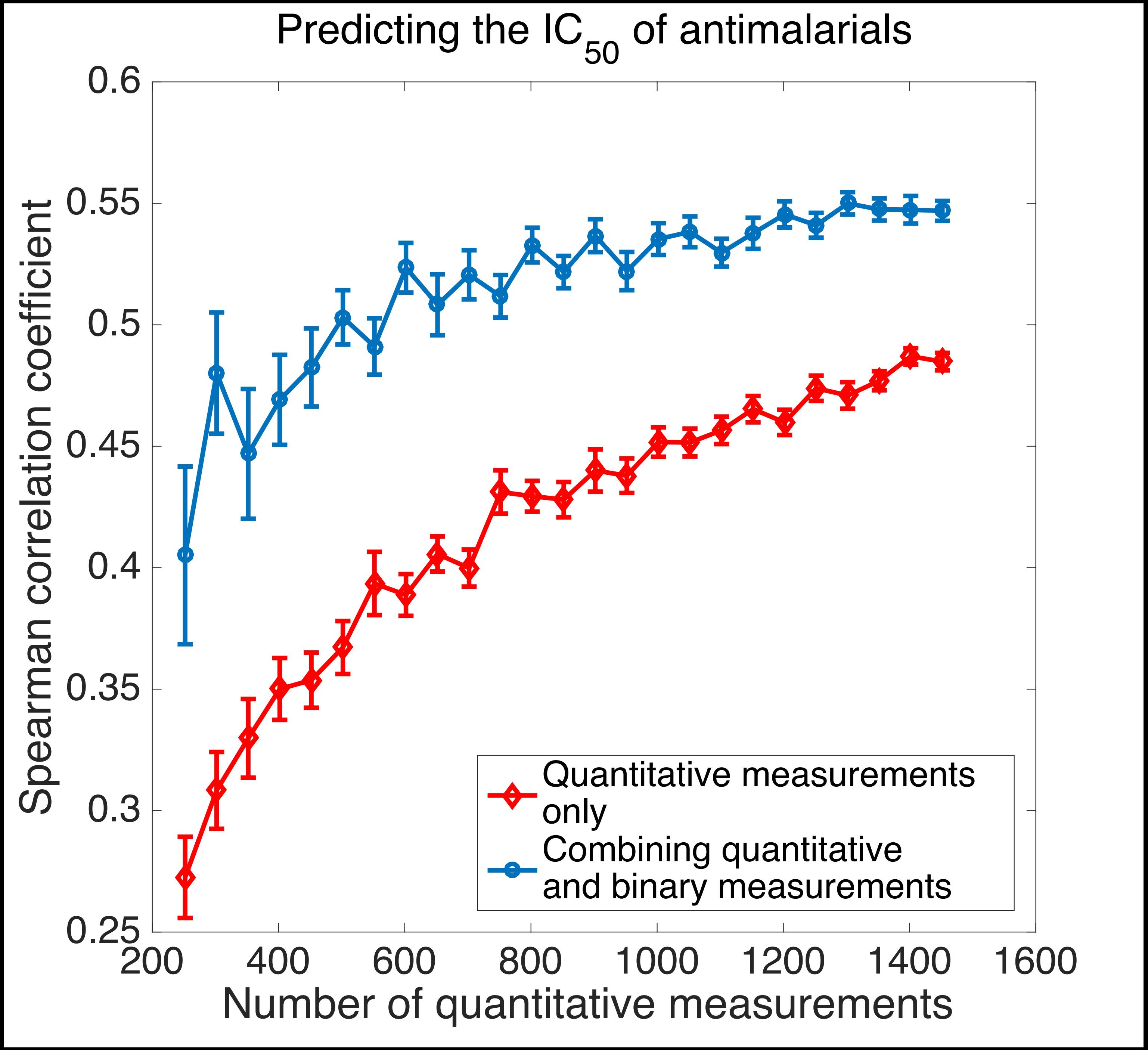
$$y_i = \mathbf{h}^T \mathbf{f}_i + \mathbf{f}_i^T J \mathbf{f}_i + \epsilon_i$$

Let $\{\mathbf{u}^{\pm}\}$ be the set of eigenvectors of the correlation matrix of soluable/insoluable molecules

$$J = \sum_{i=1}^{\hat{p}_+} c_i^+ \mathbf{u}_i^+ \otimes \mathbf{u}_i^+ + \sum_{i=1}^{\hat{p}_-} c_i^- \mathbf{u}_i^- \otimes \mathbf{u}_i^-$$

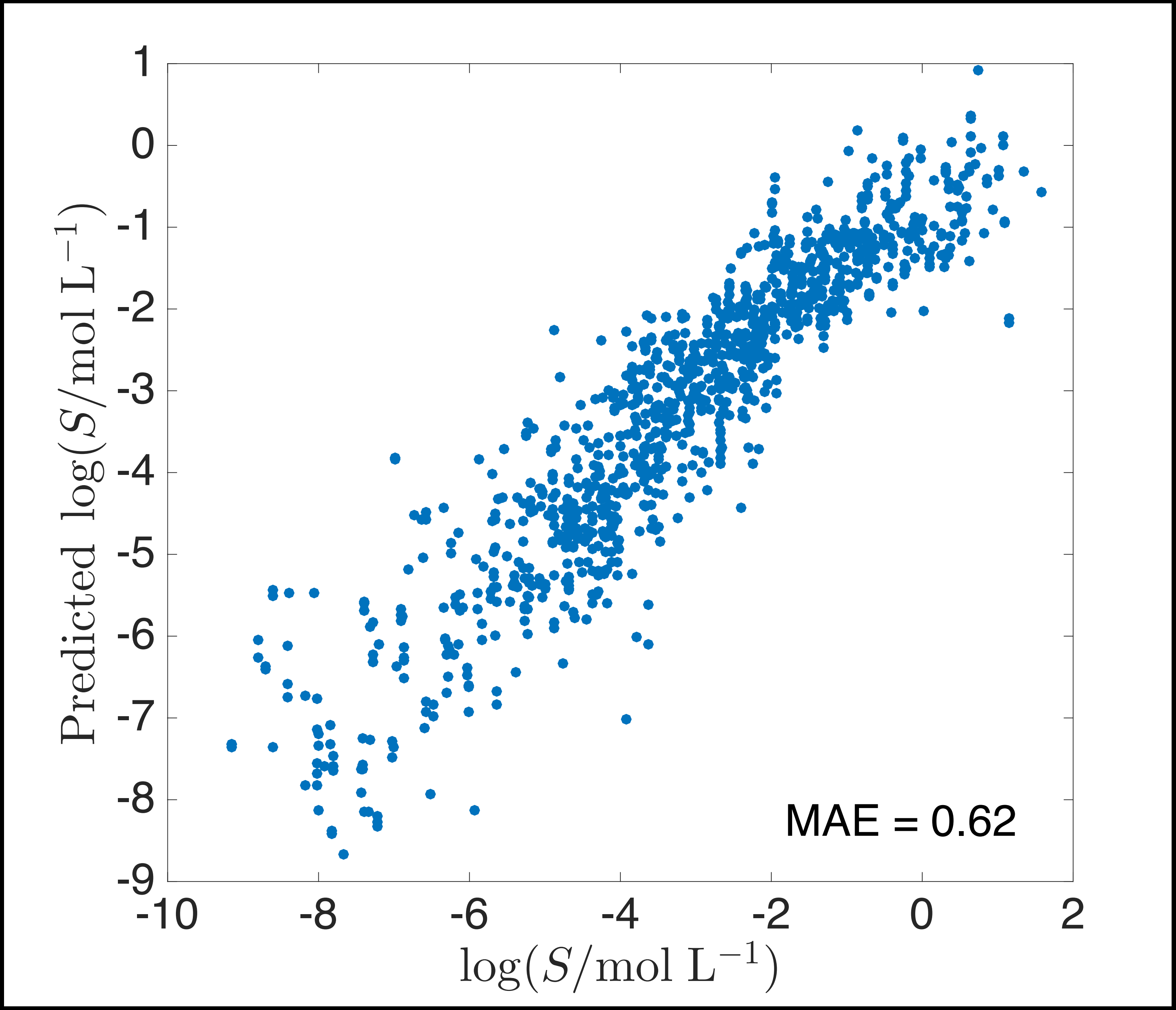
Use regression to find $\left\{\mathbf{h}, \mathbf{c}^+, \mathbf{c}^-\right\}$

The malaria TDT challenge



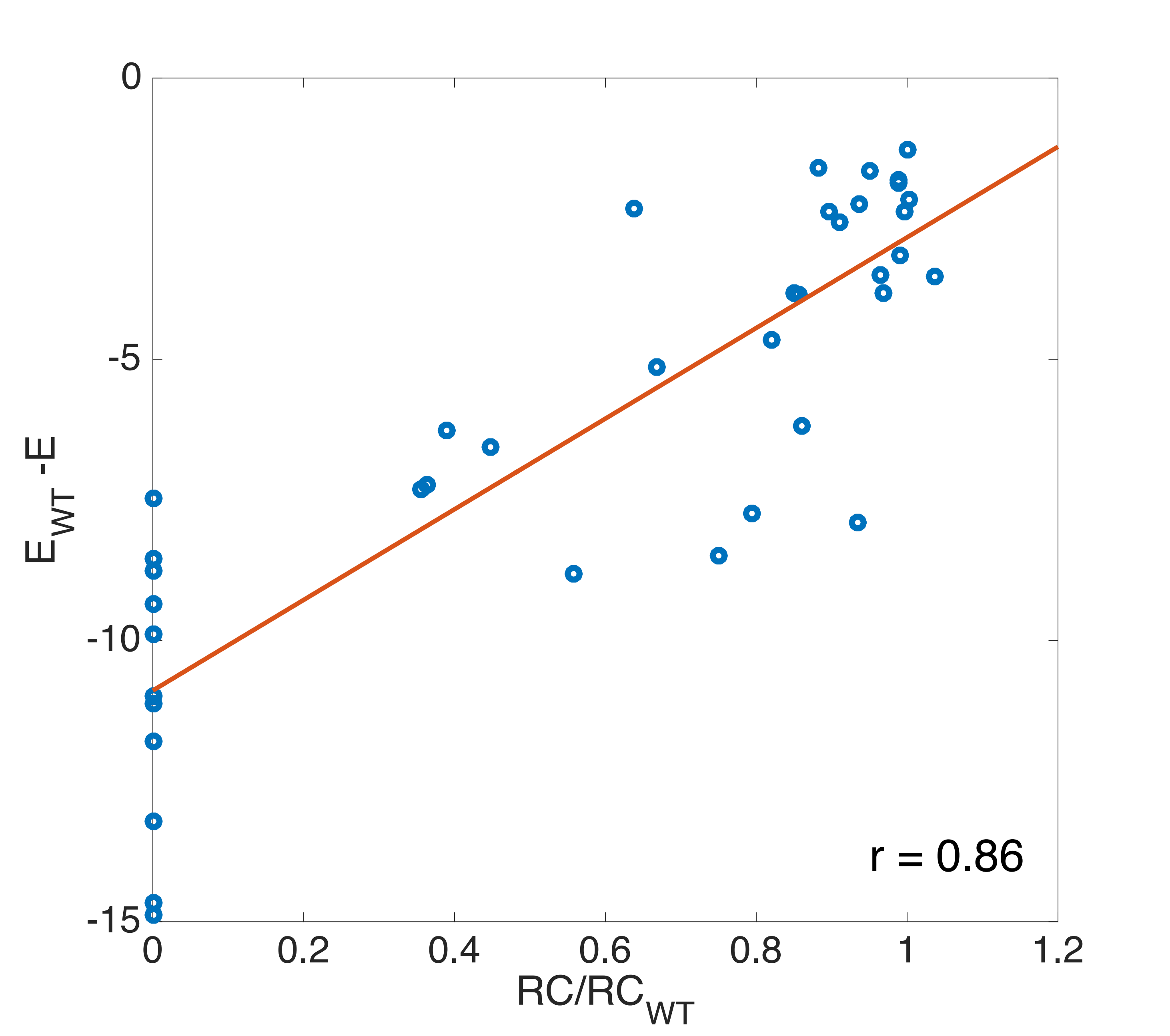
AAL, M. P. Brenner, L. J. Colwell, arXiv:1702.06001

Solubility prediction



AAL, M. P. Brenner, L. J. Colwell, arXiv:1702.06001

Fitness landscape of HIV-1 Gag



AAL, arXiv:1706.08466 (2017)

Conclusion

- Finite data effects are prevalent in chemical space exploration
- Random matrix theory provides a useful null model to undress sampling noise
- Precise and imprecise measurements can be combined to yield a predictive model

We are hiring!

contact me: aal44@cam.ac.uk