The Waisman Laboratory for Brain Imaging and Behavior

University of Wisconsin
SCHOOL OF MEDICINE AND PUBLIC HEALTH

Lattice Paths for Persistent Diagrams

Applied to COVID-19 Virus Spike Protein Structures

Moo K. Chung
University of Wisconsin-Madison
www.stat.wisc.edu/~mchung

POSTECH MIND workshop, July 8 2021

Abstract

Topological data analysis, including persistent homology, has undergone significant development in recent years. However, one outstanding challenge is to build a coherent statistical inference procedure on persistent diagrams. The paired dependent data structure as birth and death in persistent diagrams adds additional complexity to developing a coherent statistical inference procedure. In this paper, we present a novel data representation that transforms persistent diagrams as lattice paths. A new exact statistical inference procedure is developed over the collection of lattice paths via combinatorial enumerations. The lattice path method is applied to the topological features of the protein structures of corona viruses. The proposed method demonstrates that there are topological changes during the conformation change of spike proteins that are need to infect host cells. The talk is based on arXiv:2105.00351.

Acknowledgement

Tananun Songdechakraiwut, Zhan Luo, Ian Carroll,
Gregory Kirk, Andrew Alexander, Hill Goldsmith, Richard Davidson
Univ. of Wisconsin-Madison, US

Yuan Wang University of South Carolina, US
Li Shen Univ. of Pennsylvania, US
Hernando Ombao KAUST, Saudi Arabia
Hyekyung Lee Seoul National University, Korea
Victor Solo University of New South Wales, Australia

Alexandra Wallas University of Washington, US
Tzu-Jing yang Academia Sinica, Taiwan
Alissa Eckert, Dan Higgins Disease Control and Prevention, US

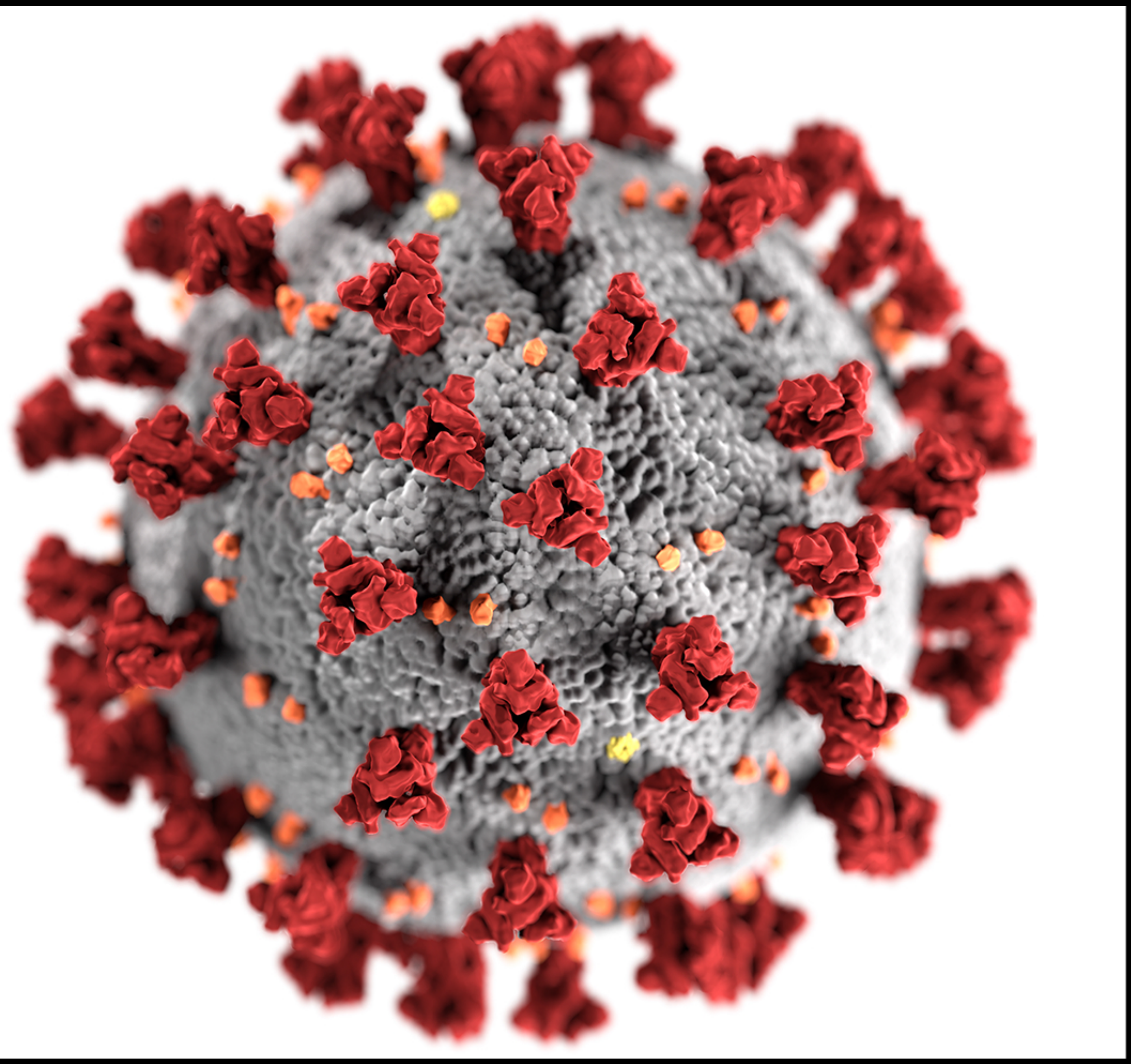
Grants:

NIH R01 Brain Initiative EB022856, R01 EB028753, NSF DMS-2010778

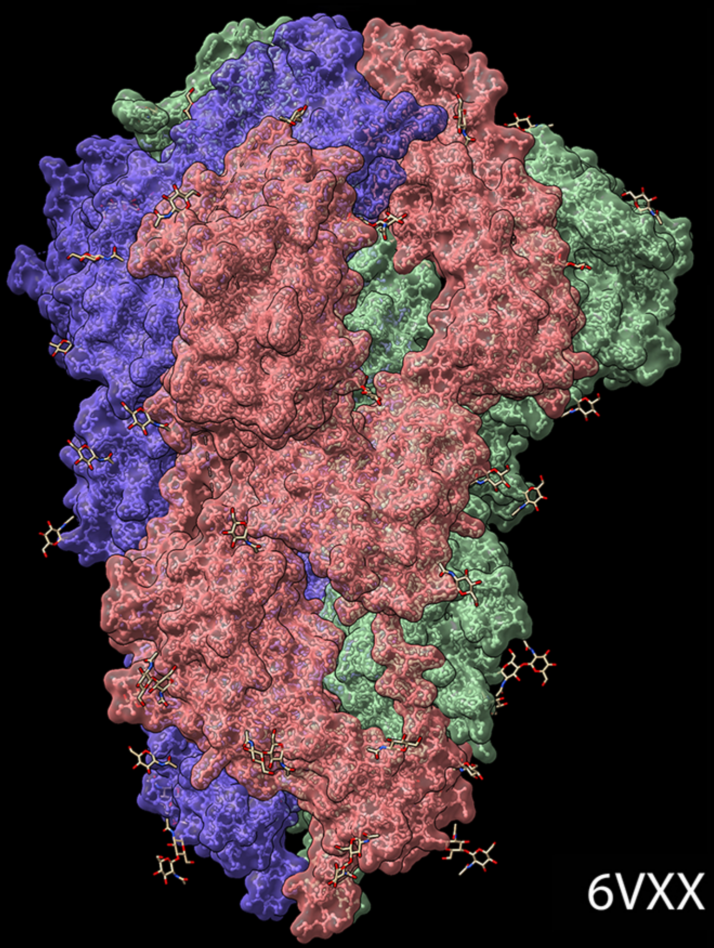
How COVID-19 virus attach to host cell?

A division of the
American Chemical Society

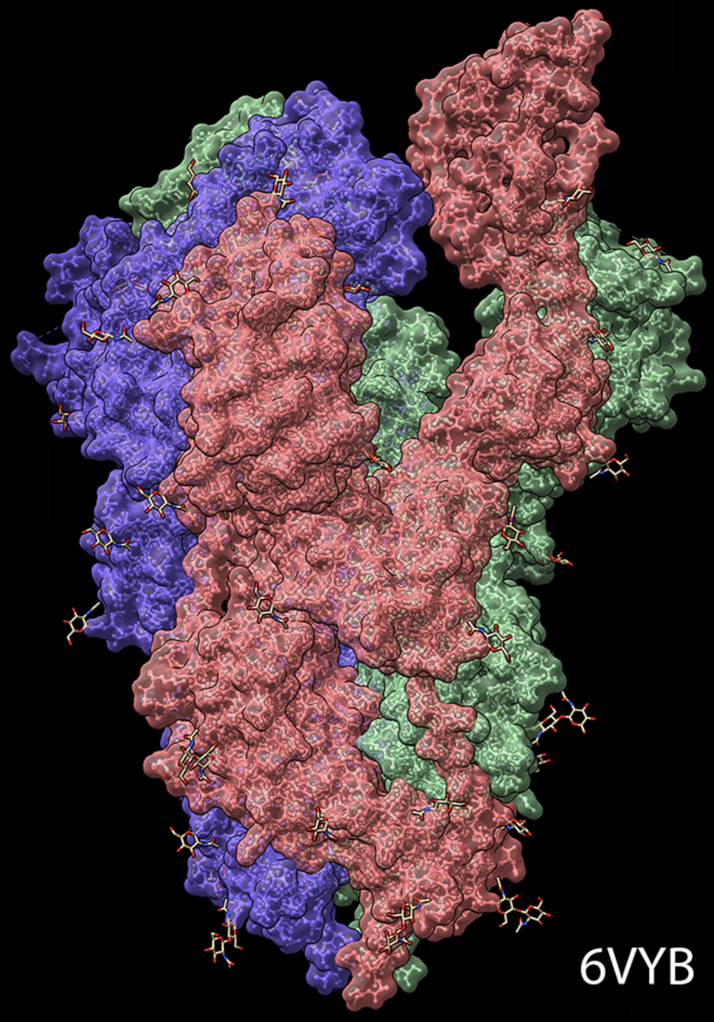
Conformational changes of spike protein



Covid-19 virus
Closed state



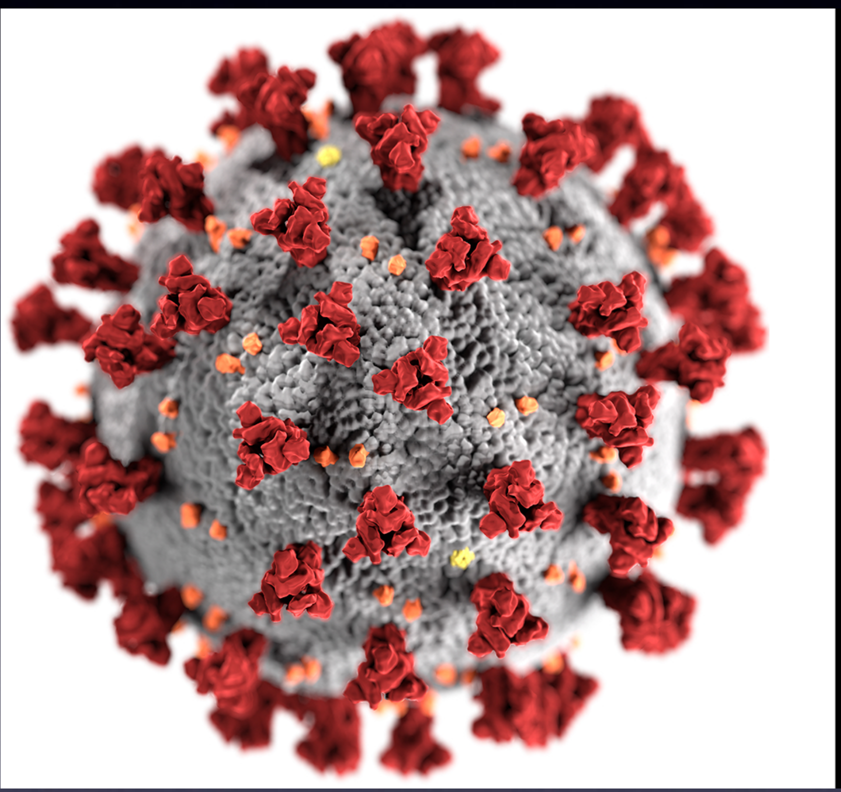
COVID-19 virus
Open state

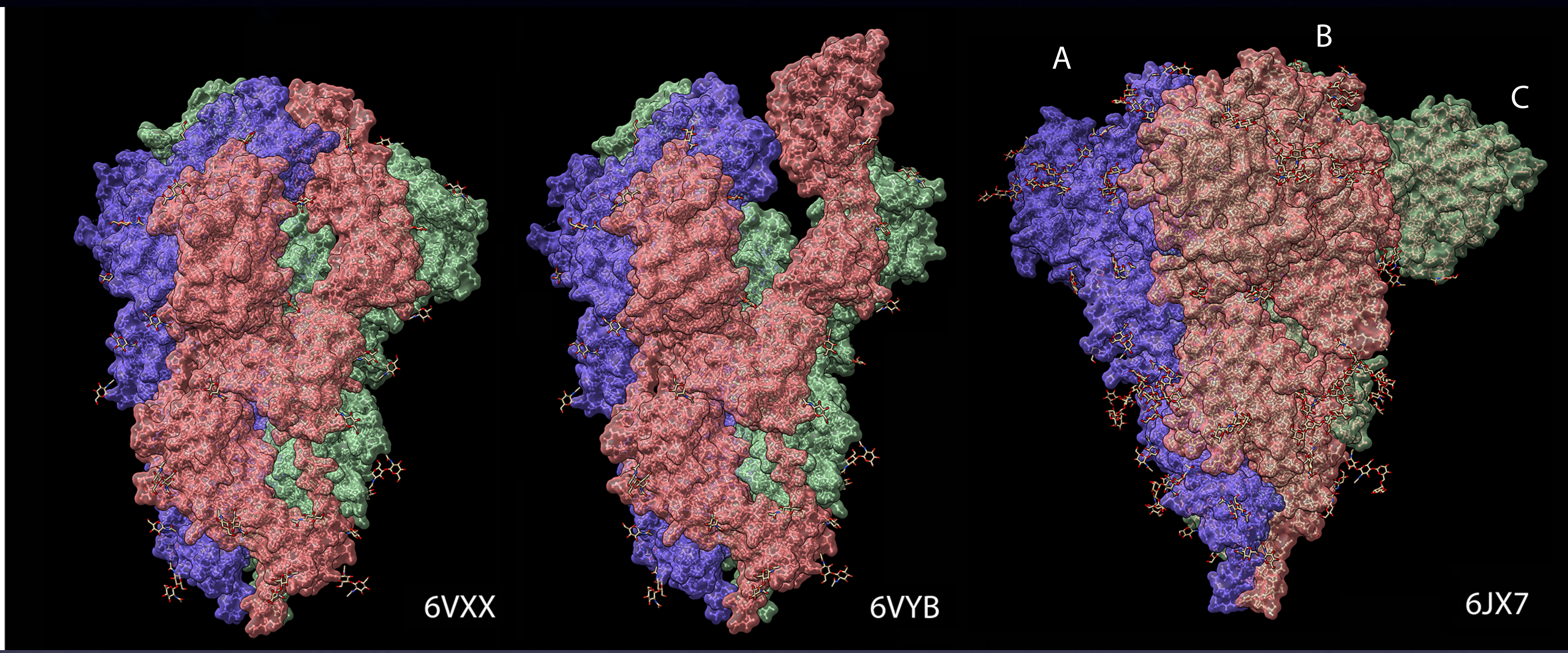


Is there topological changes?

Shape change of spike protein of COVID-19 virus

Topological distance is not enough.
We need the probability of how close they are



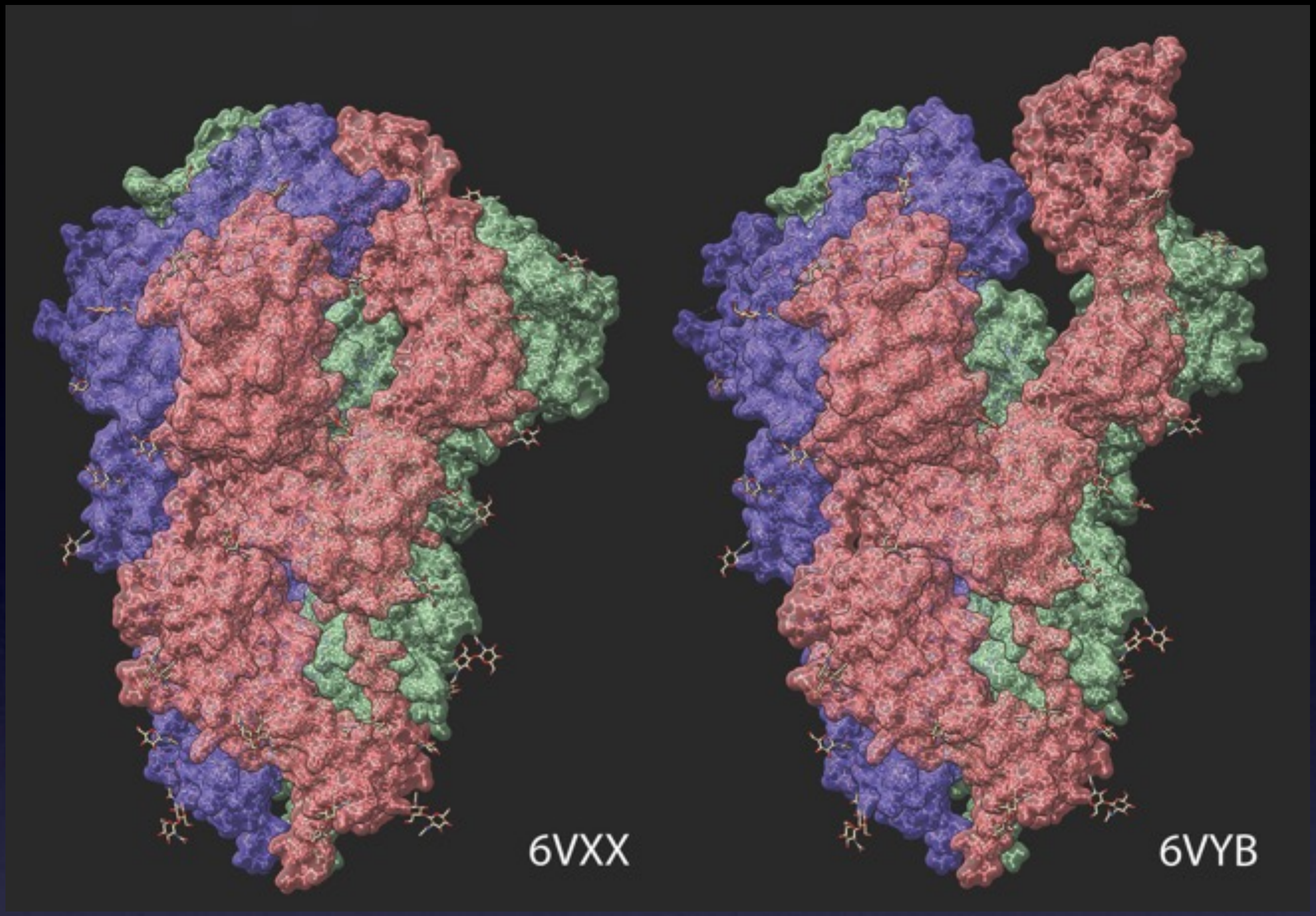


Topologically different:
p-value (probability of closeness) = 8×10^{-38}

Matlab demo

Chung and Ombao, 2021 arXiv:2105:00351

Fallacy of comparing averages



Measurements: 2 3 4 | 3 4 5

Means: 3 | 4

Incorrect conclusion: 6VYB is larger

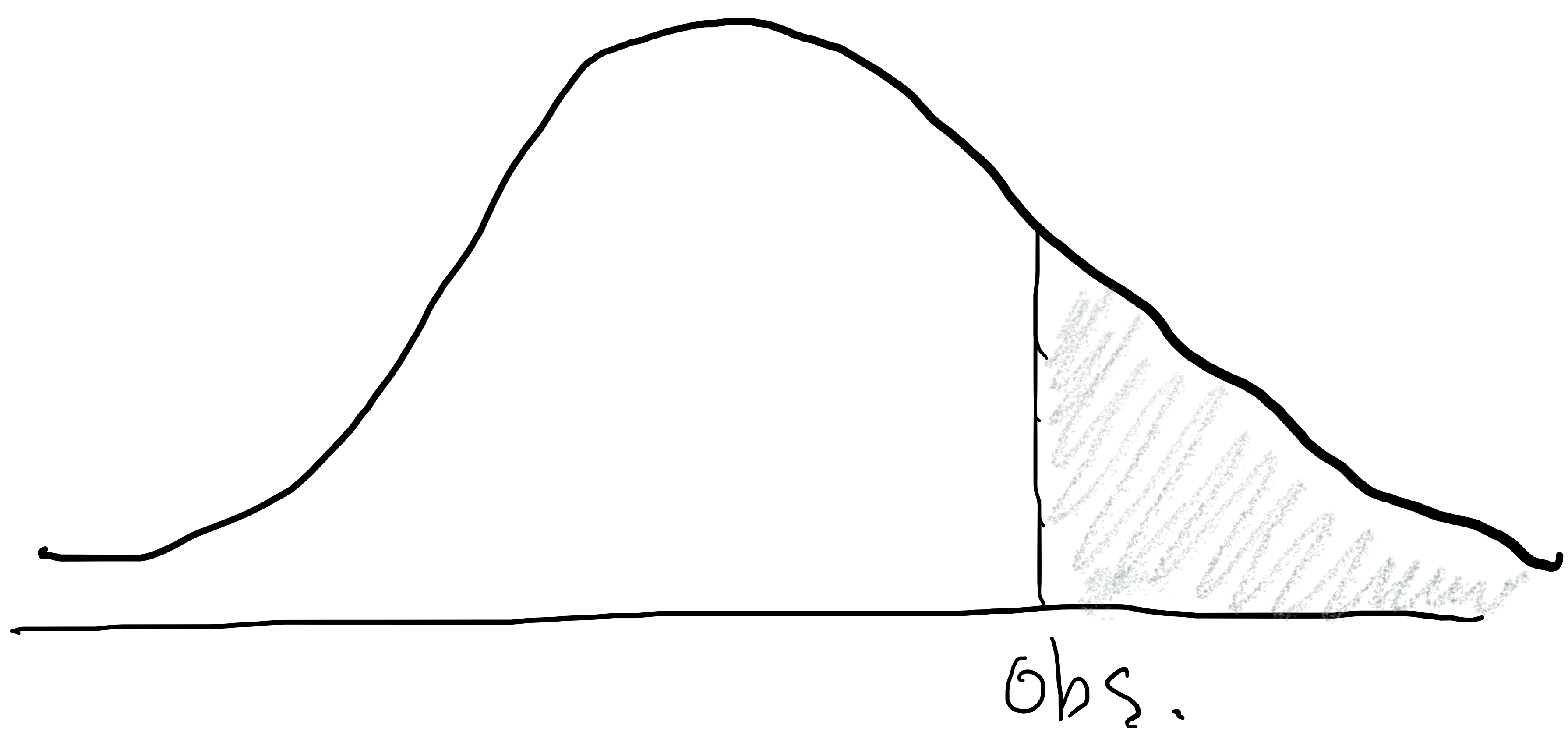
50% of time, your conclusion will be wrong!

Comparing averages is not good enough

```
a=rand(100,1)
b=rand(100,1)

count=0;
for i=1:10000
    sa = randsample(a,5);
    sb = randsample(b,5);
    count = count + (mean(sa)<mean(sb));
end

count/10000

ans =
   0.6116, 0.5340, 0.7103, 0.4858, 0.4261, 0.4295 ...
```

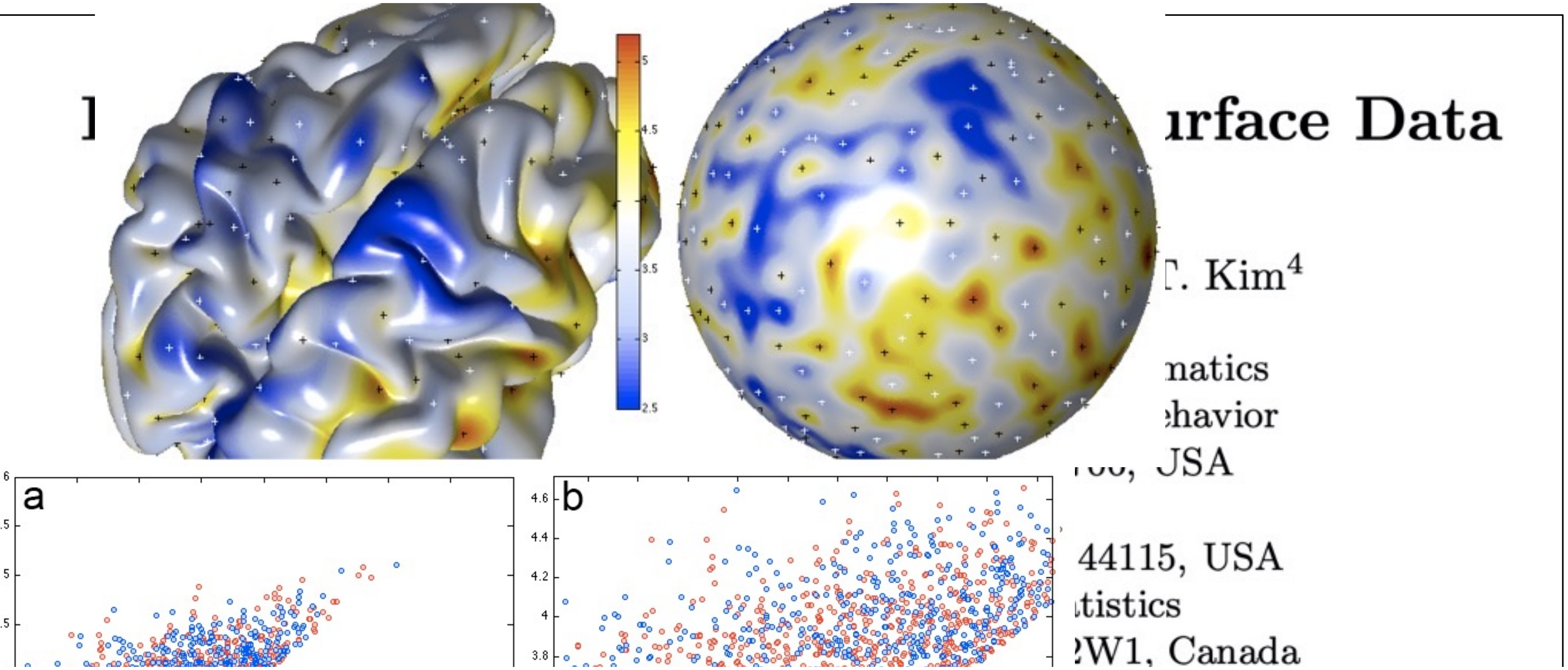
Inference on average difference

Determine the probability distribution of the average difference

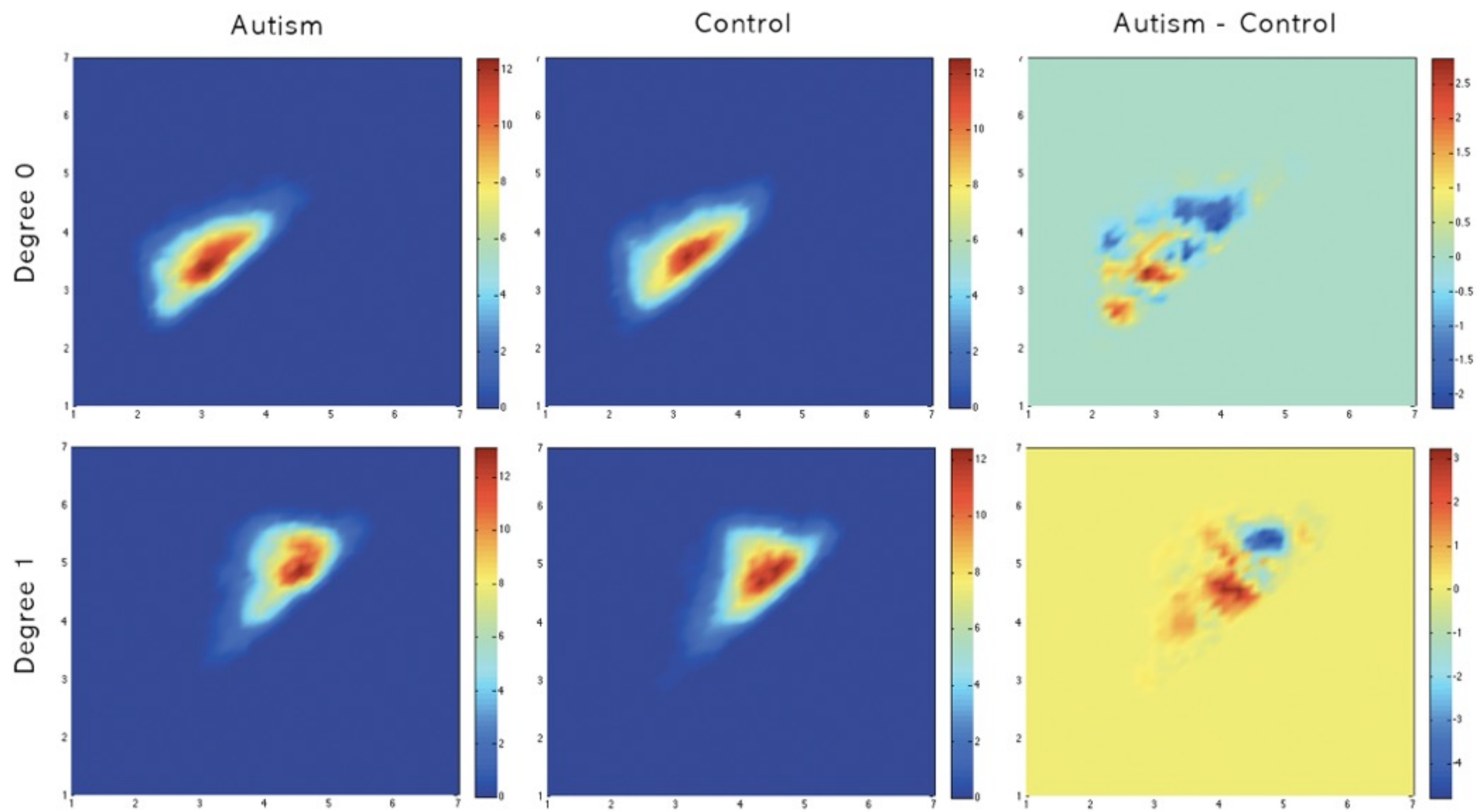


Topological inference

Topological inference is the process of using topological data analysis to infer properties of an underlying probability distribution of topological objects or features.



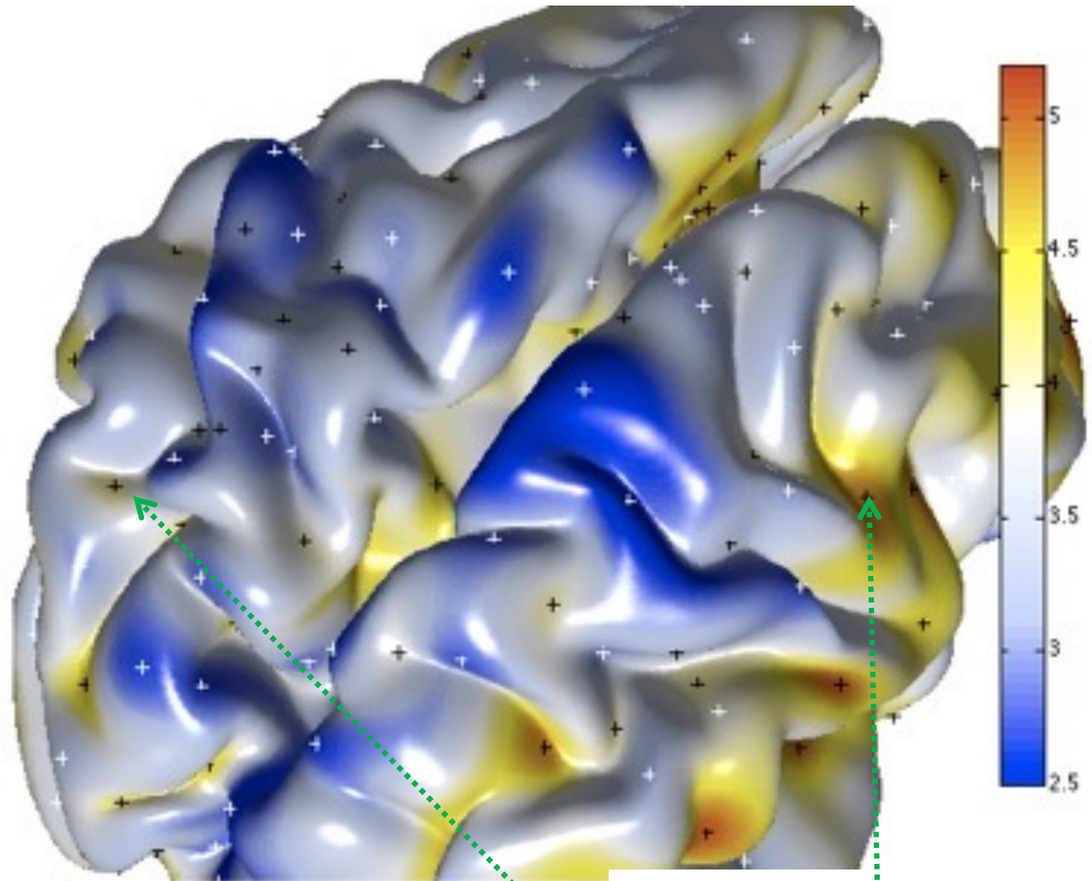
Chung et al., 2009
Information Processing in Medical Imaging (IPMI) 5636:386-397.

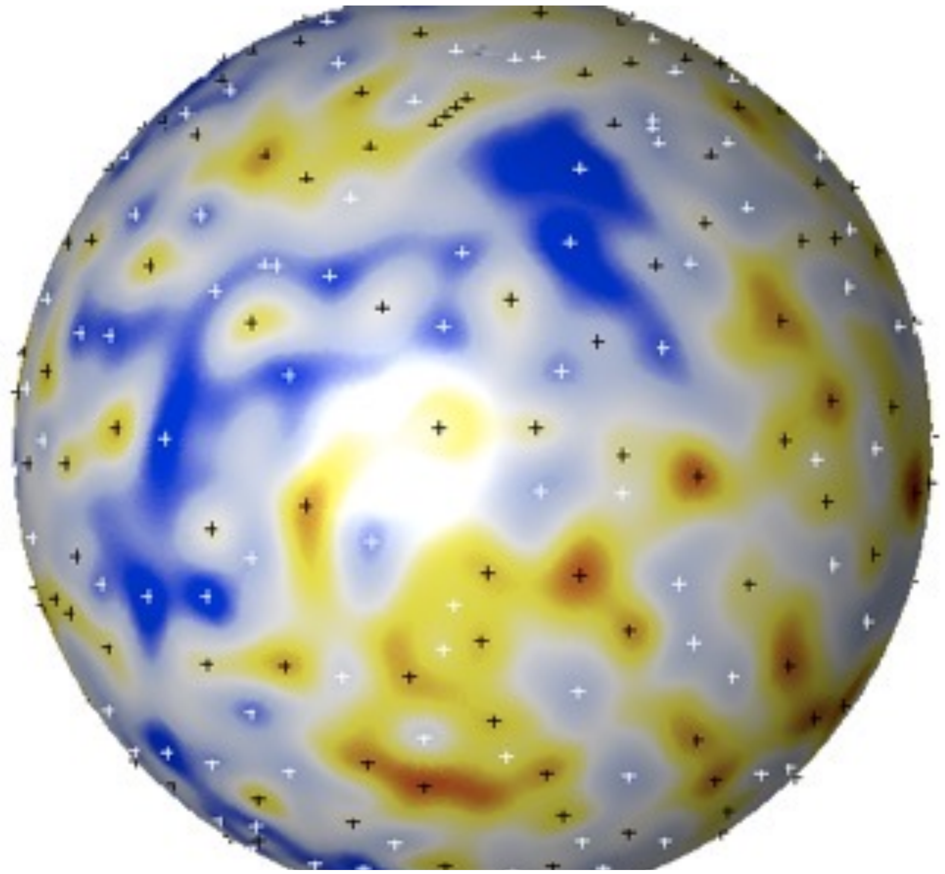


1) First persistent homology paper in medical imaging

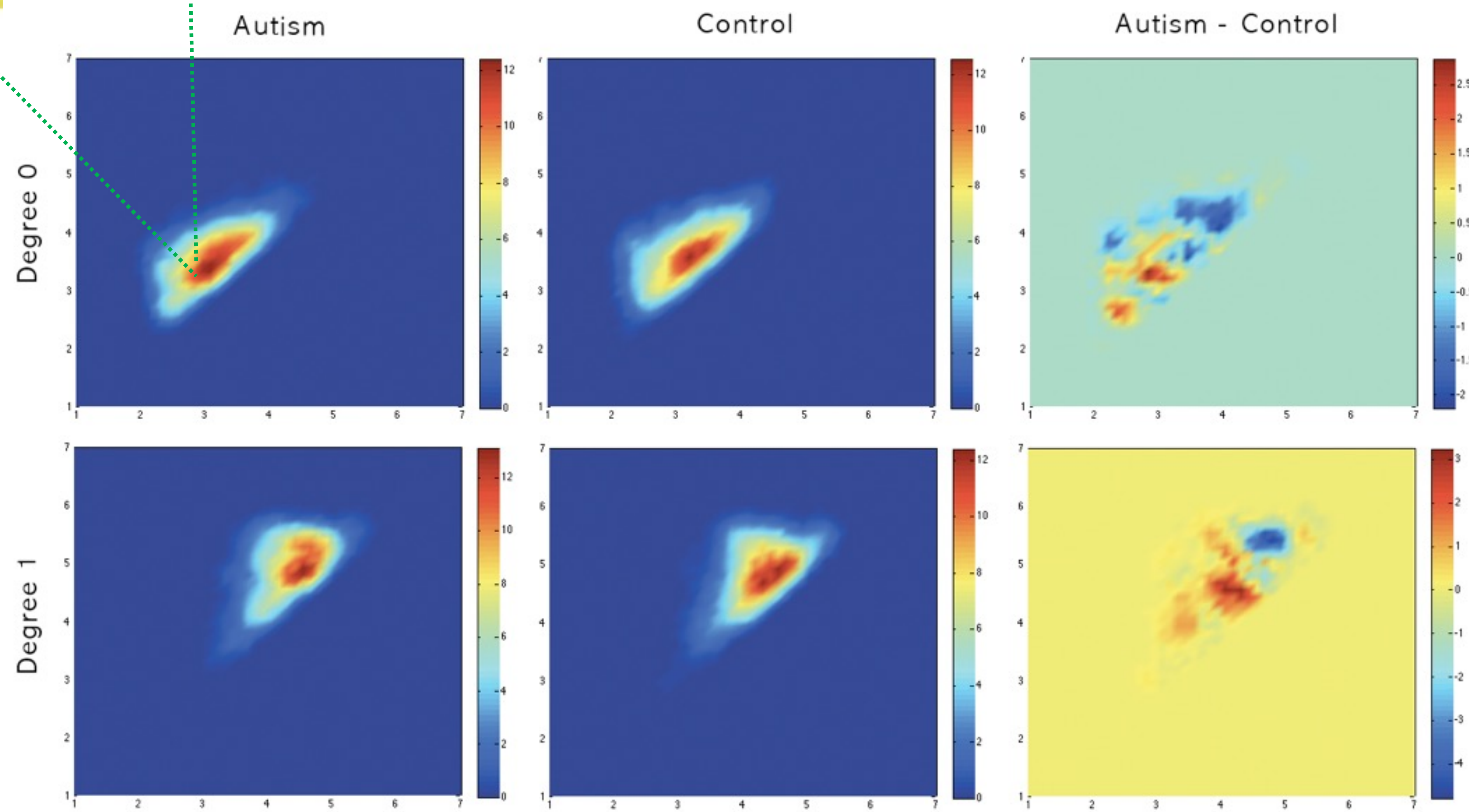
2) Smoothed PD using the uniform kernel (counting measure)

Smoothing PD is biological nonsense!

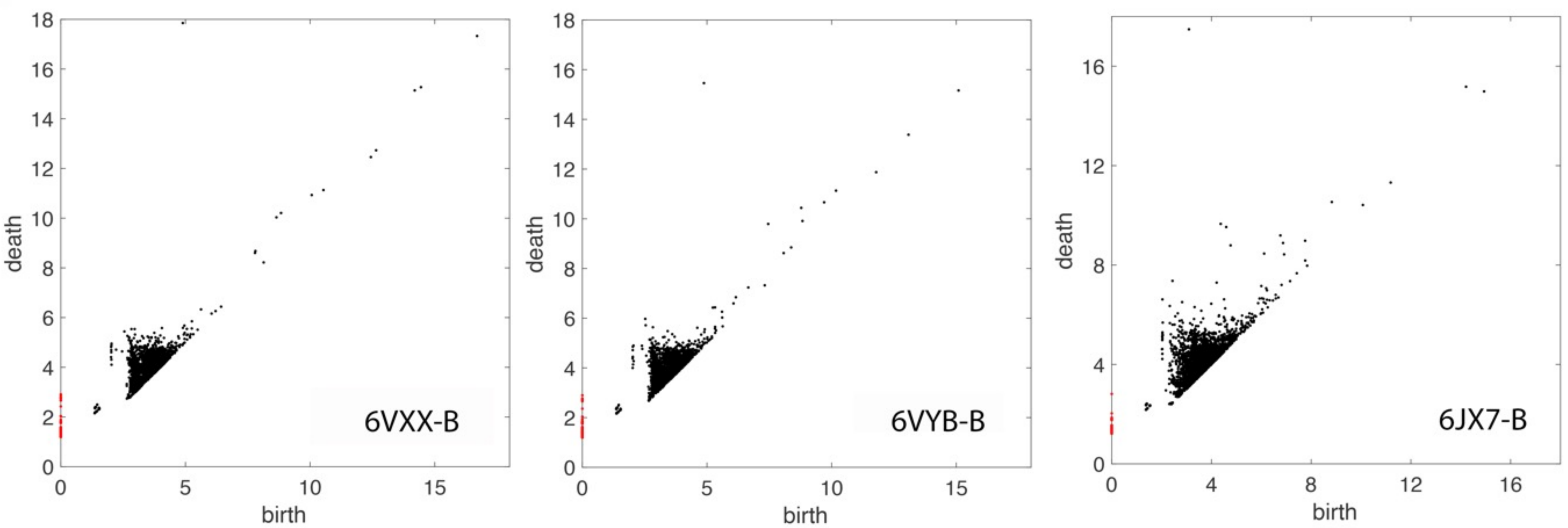




Fusing signal in different brain regions!



How we perform statistical analysis *without* smoothing persistent diagrams?



PD on Rips filtrations

Persistent diagram $\{(b_1, d_1), \cdots, (b_q, d_q)\}$

Birth-death process

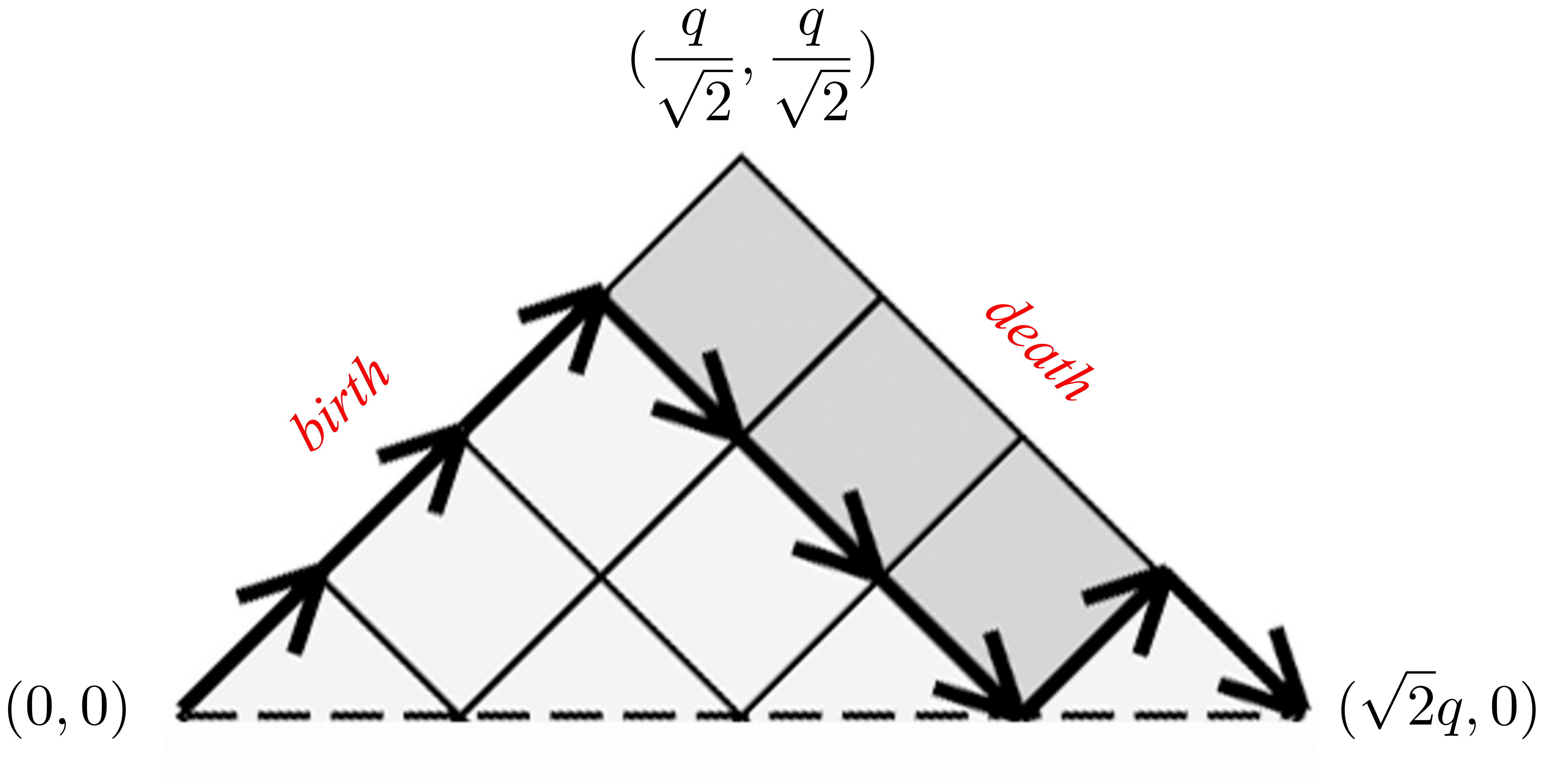
$$c_{(1)} < c_{(2)} < \cdots < c_{(2q)}$$

$c_{(i)}$: one of birth or death value

$\nearrow$ births

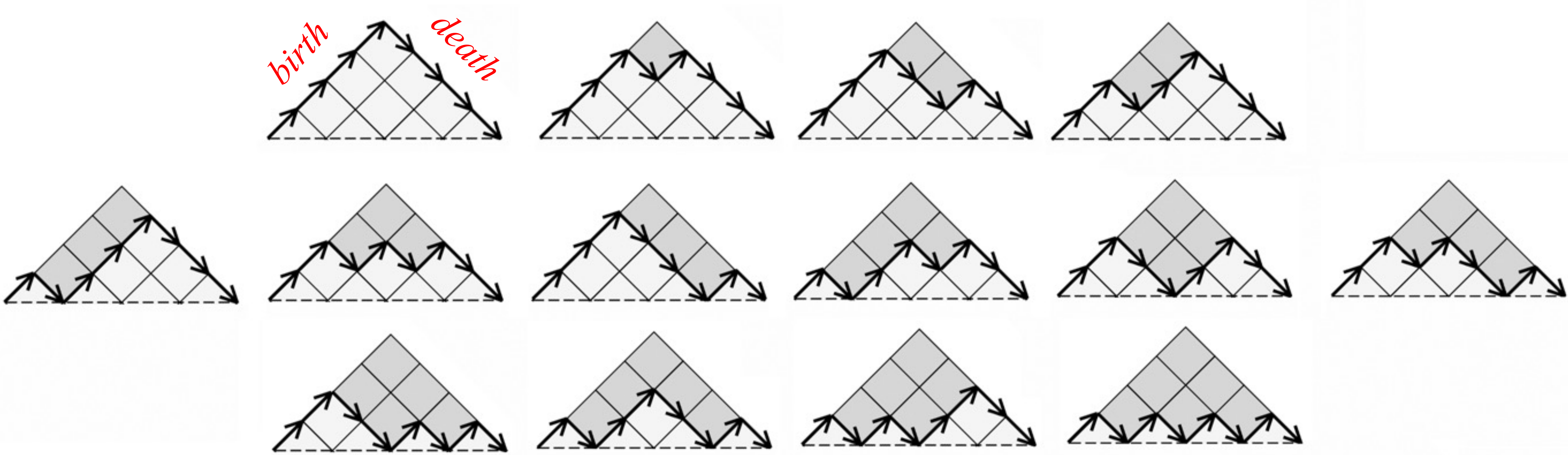
$\searrow$ deaths

Dyck paths



The path starts at $(0,0)$ and ends at $(\sqrt{2}q, 0)$.
The path stays above the horizontal line.

14 possible Dyck paths for *q*=4

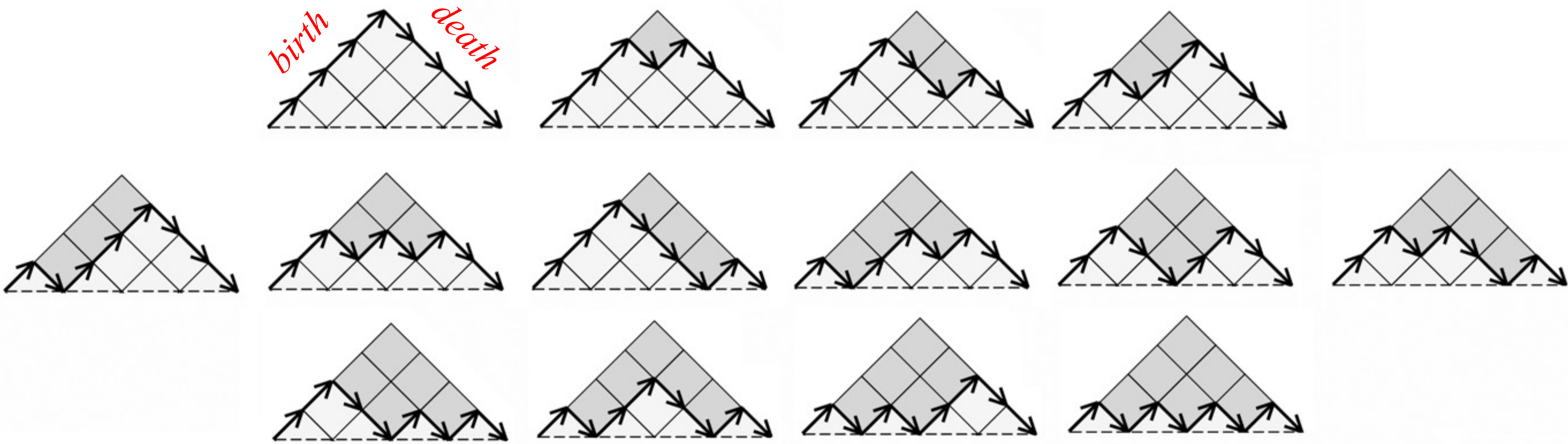


Total number of Dyck paths?

Catalan number

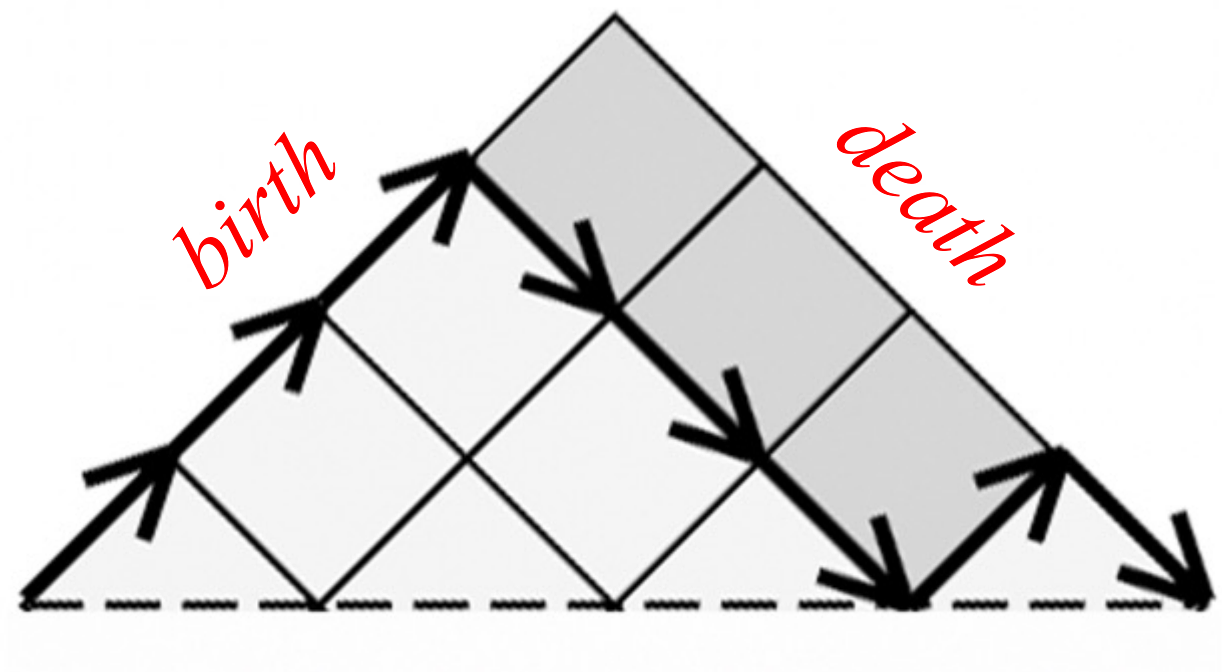
$$\kappa_p = \frac{1}{q+1}\binom{2q}{q}$$

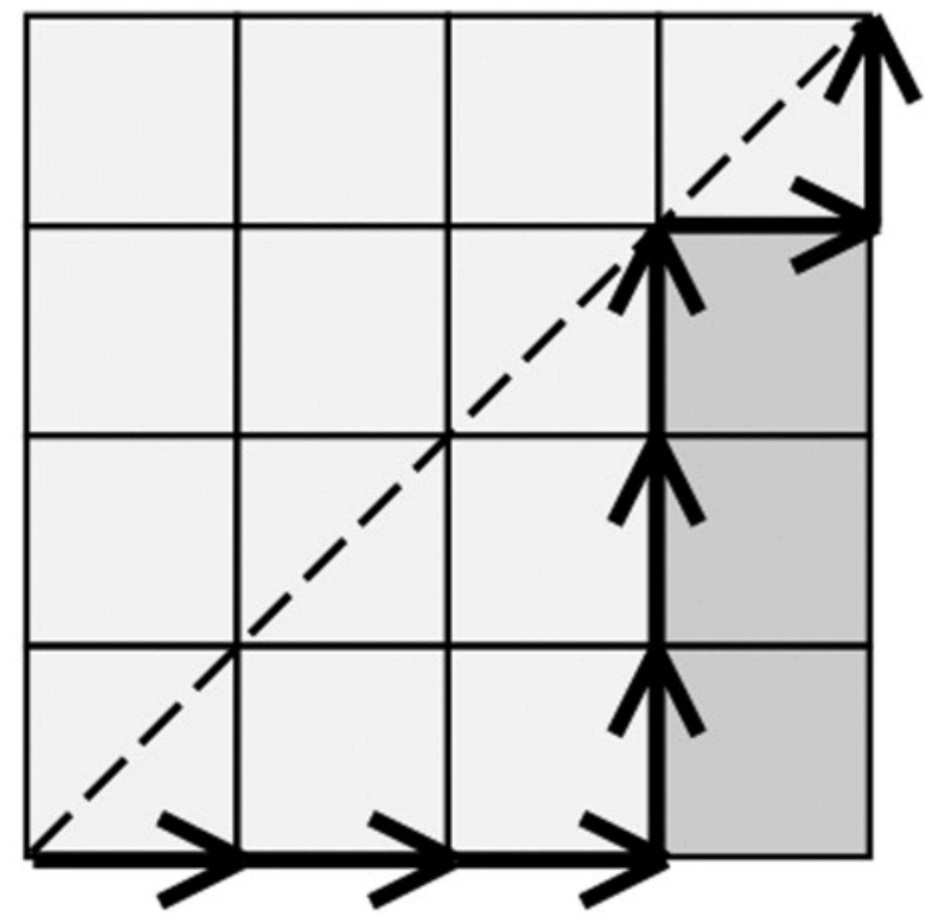
Area under Dyck path



larger area = longer persistence
smaller area = shorter persistence

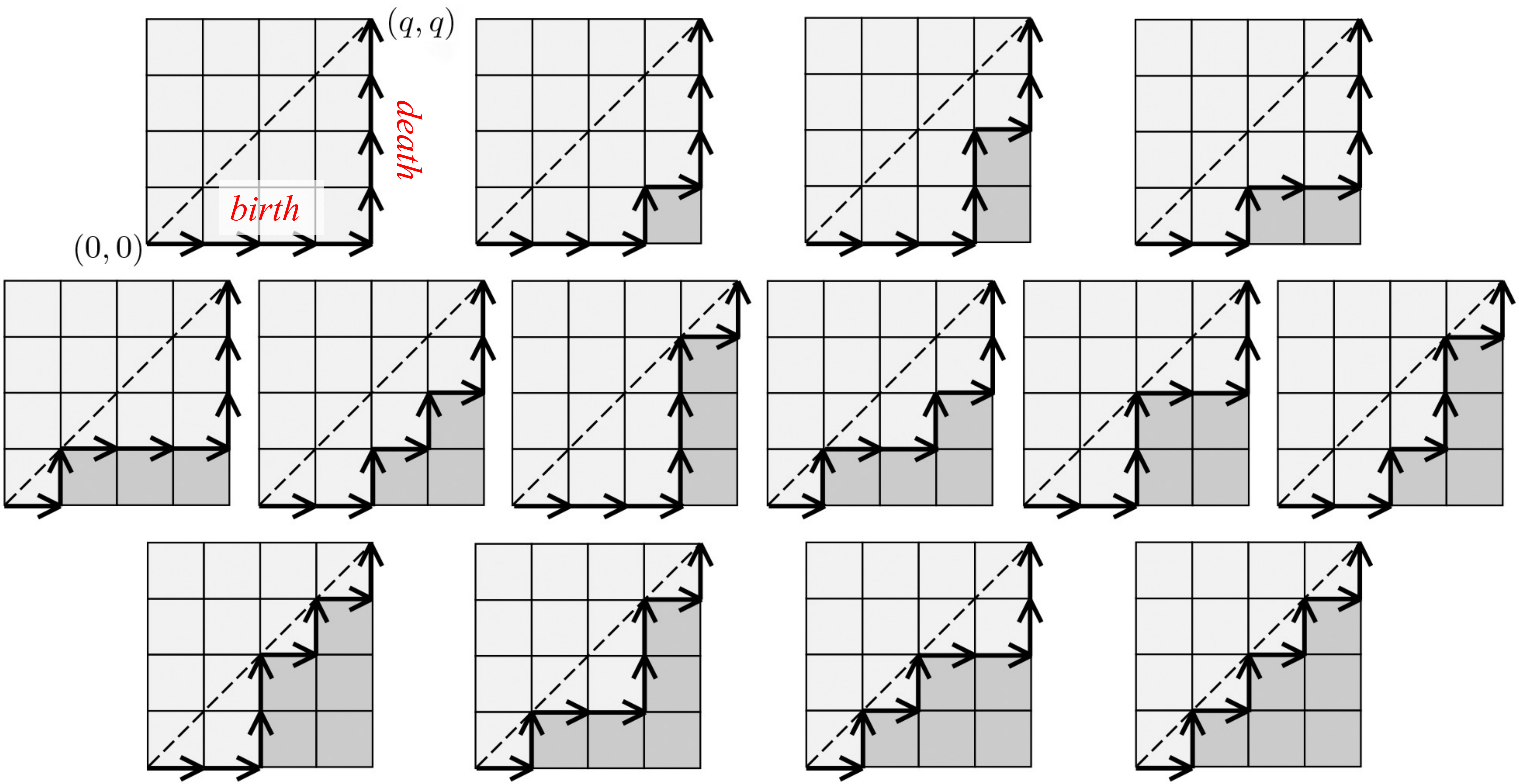
Area under Dyck path via box counting under lattice path





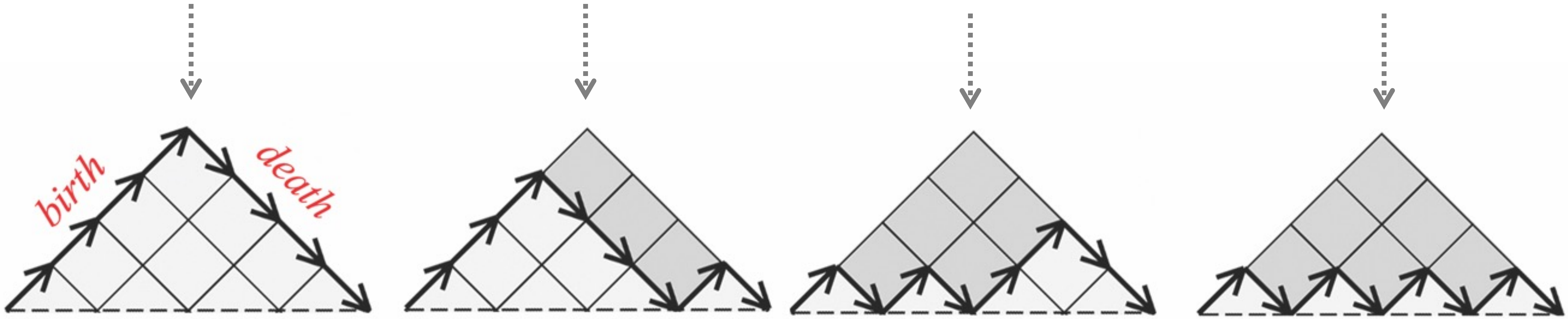
$$\text{Area below Dyck path} = \frac{q^2}{2} - \text{total area of boxes below lattice path}$$

Every possible monotone lattice path for q=4

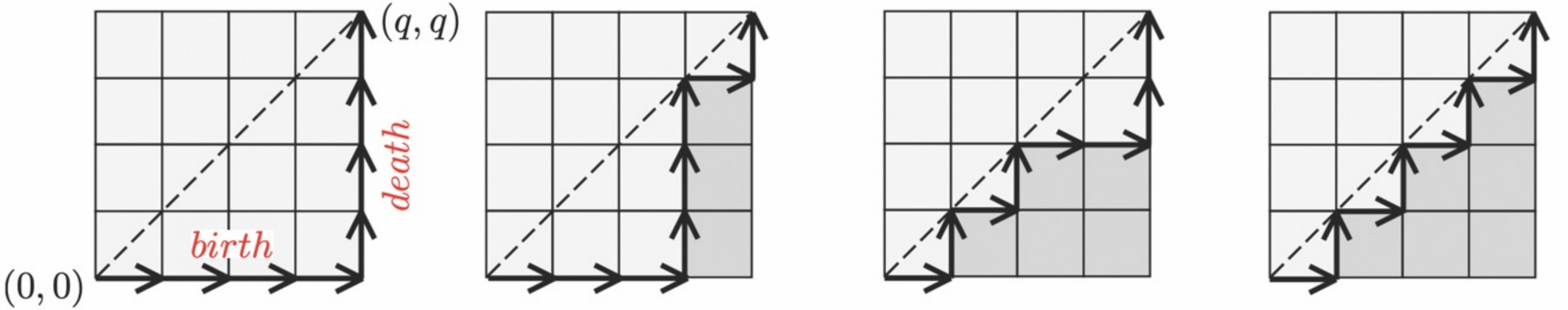


Number of boxes form a monotone sequence

Persistent diagrams



The height of boxes form a monotone sequence



0 0 0 0   0 0 0 3   0 1 2 2   0 1 2 3

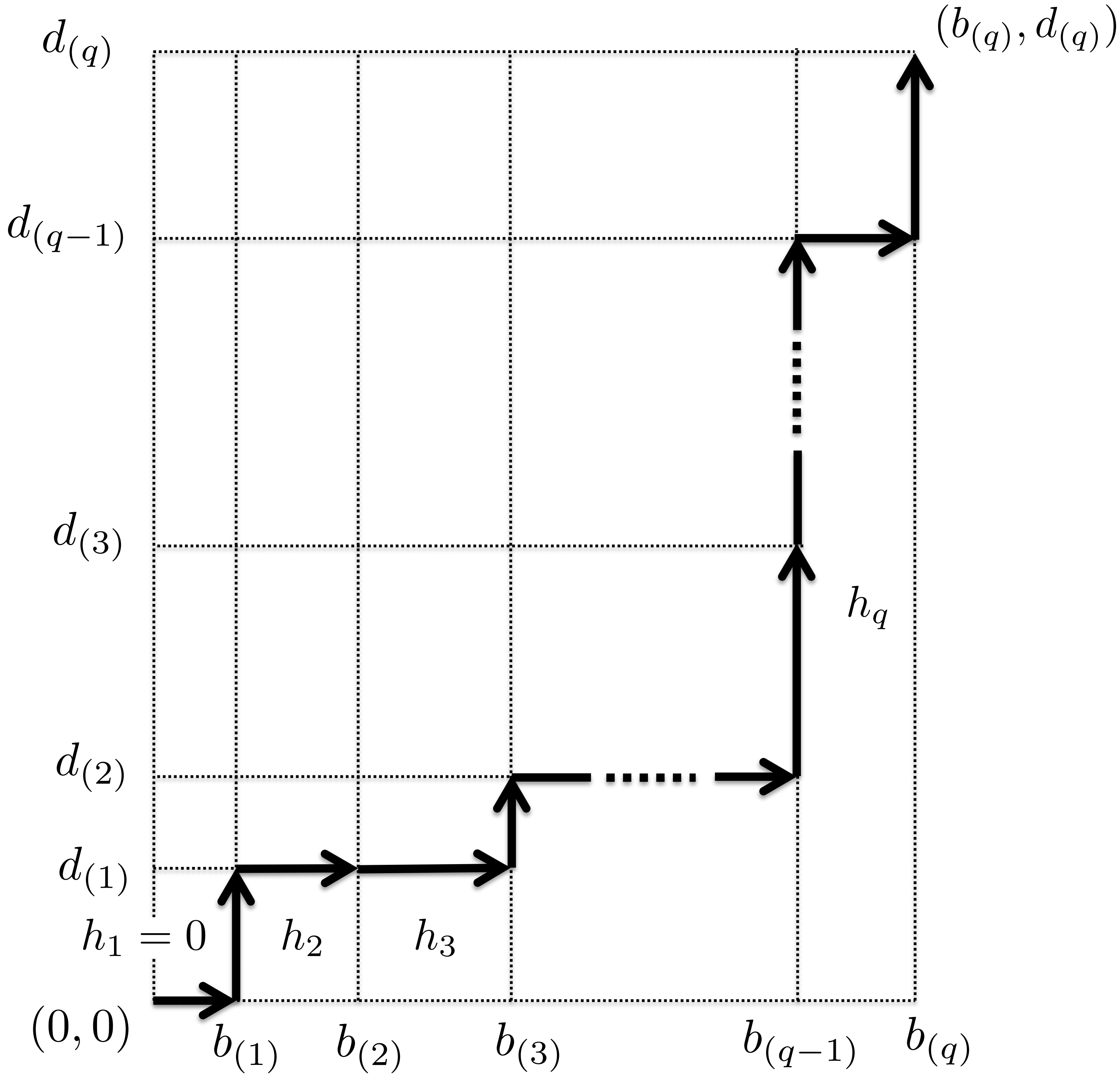
Topological learning, Topological inference

Limitation of Dyck and lattice paths

Encode the <u>order</u> of how births and deaths are paired.

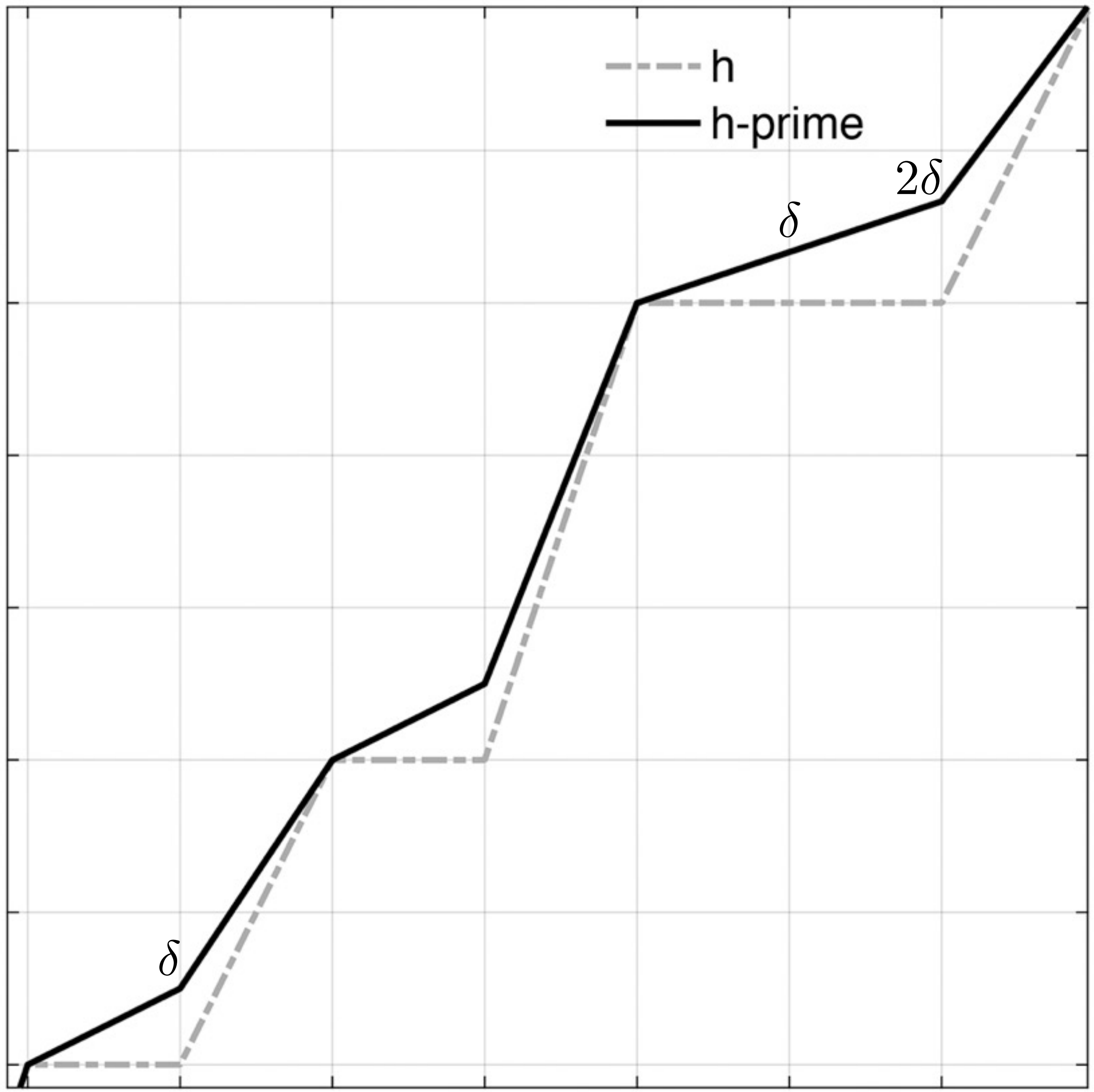
Do <u>not</u> encode the actual filtration values.

Weighted lattice path



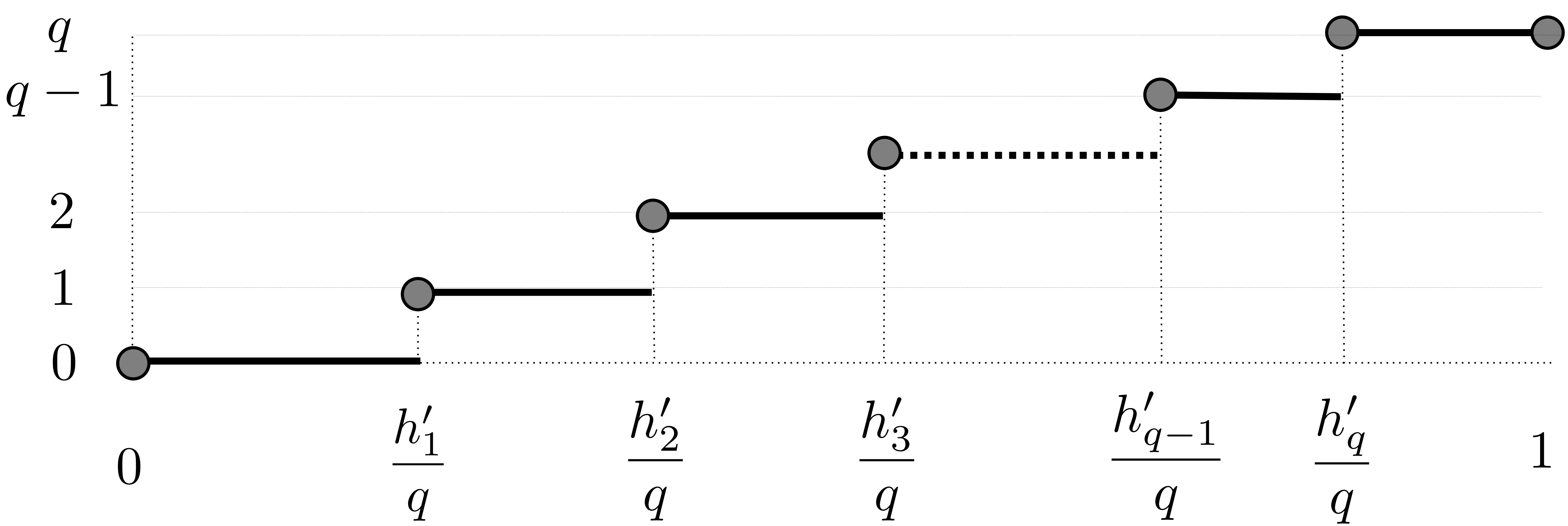
Areas below weighted lattice path

$$h' : h'_1 < h'_2 < \cdots < h'_q$$

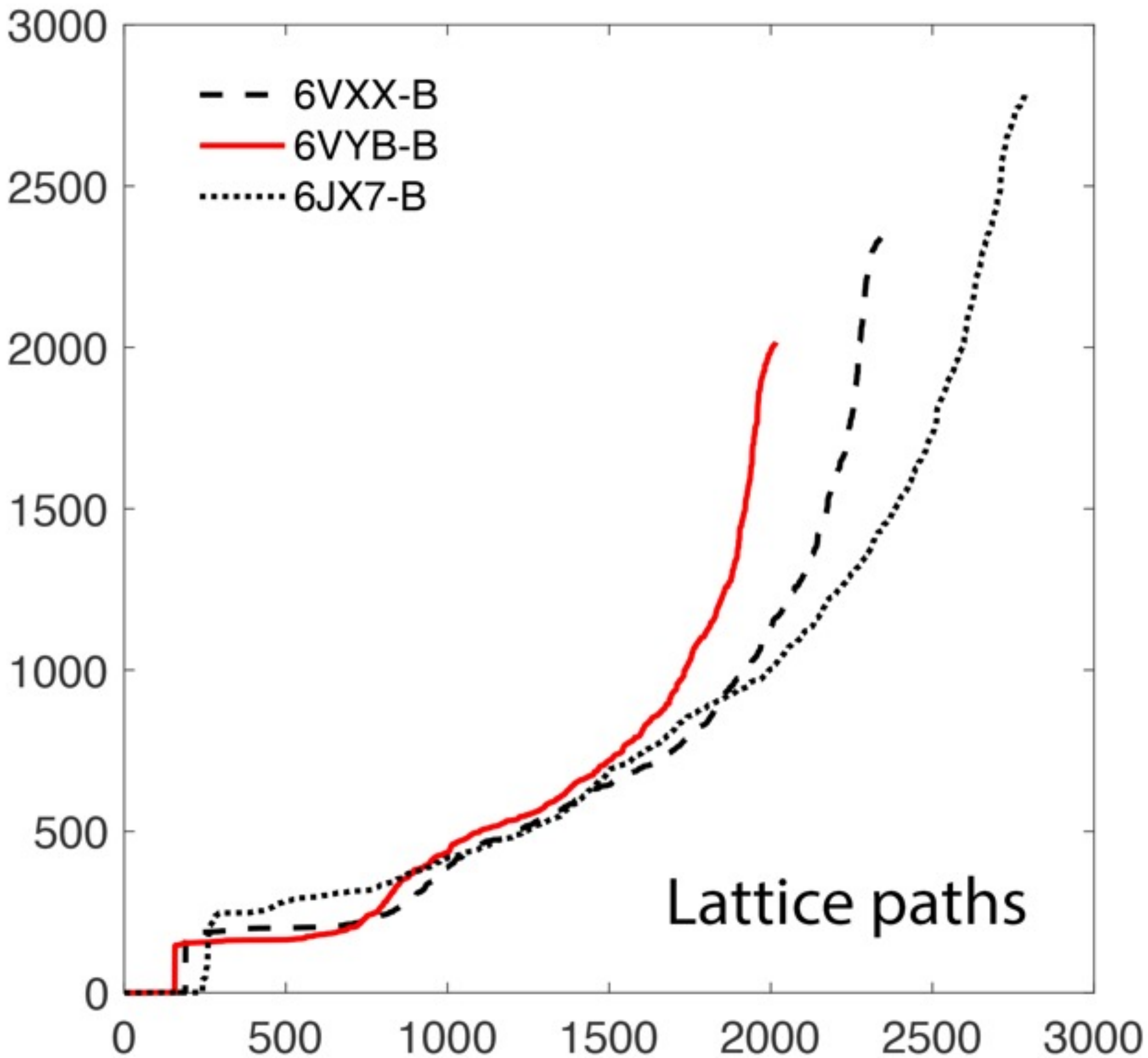


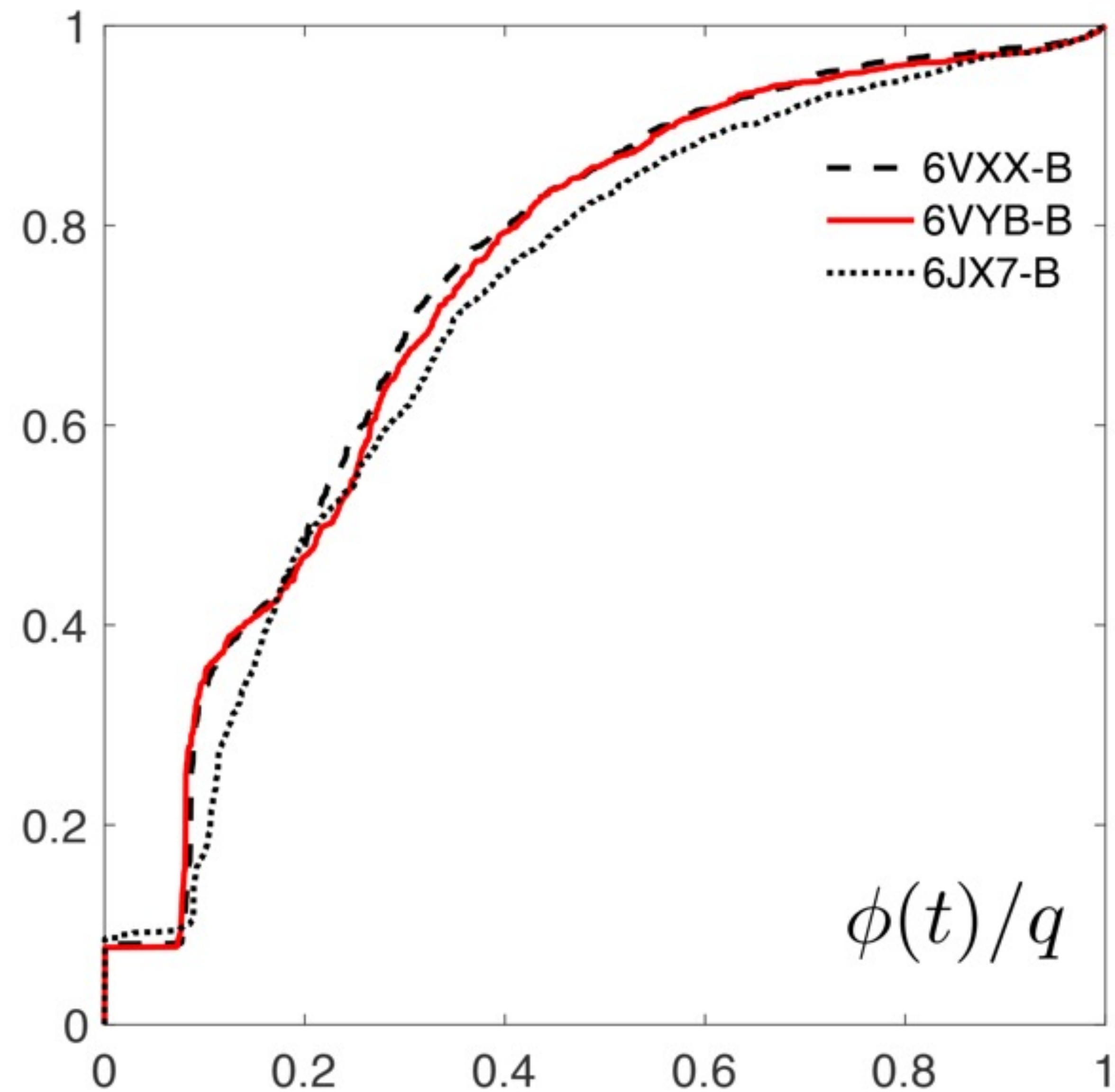
Empirical distribution like step function

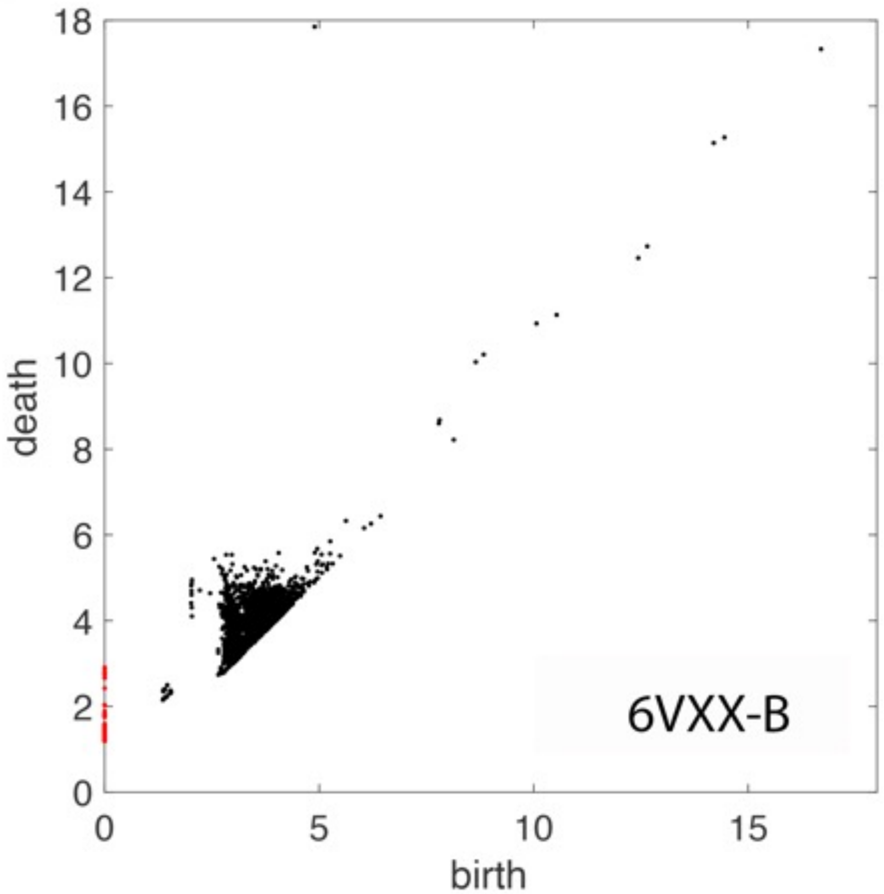
$$\phi(t) = \begin{cases} 0 & \text{if } t \in [0, \frac{h'_1}{q}) \\ j & \text{if } t \in [\frac{h'_j}{q}, \frac{h'_{j+1}}{q}) \ \text{ for } j = 1, \cdots, q-1 \\ q & \text{if } t \in [\frac{h'_q}{q}, 1] \end{cases}$$

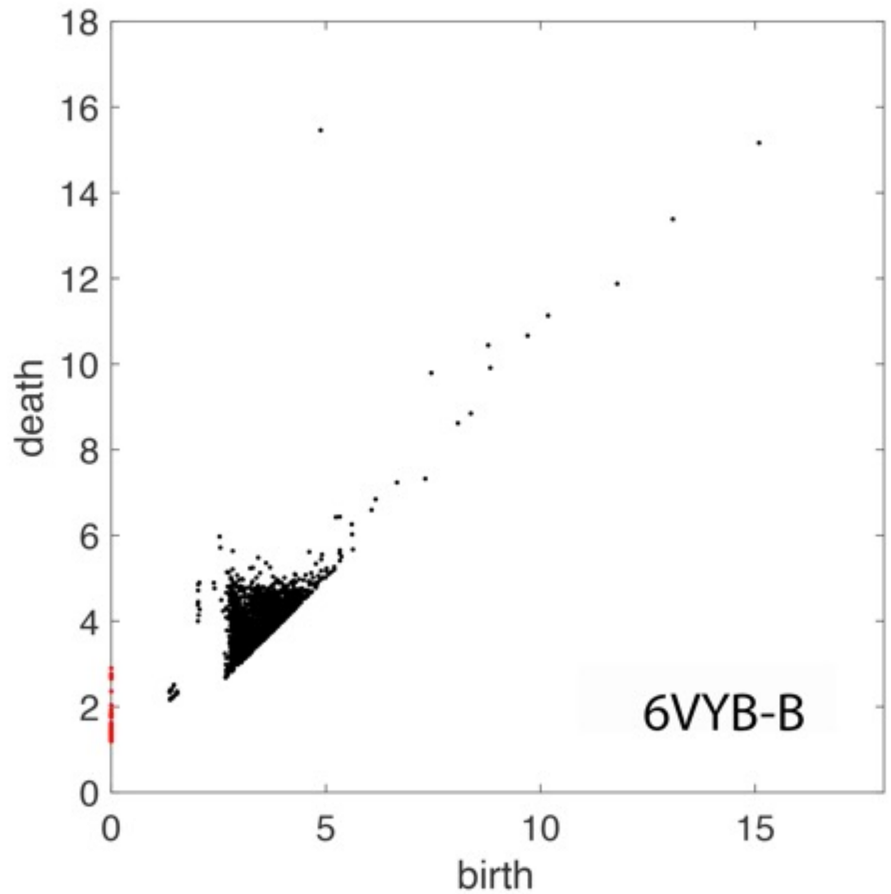


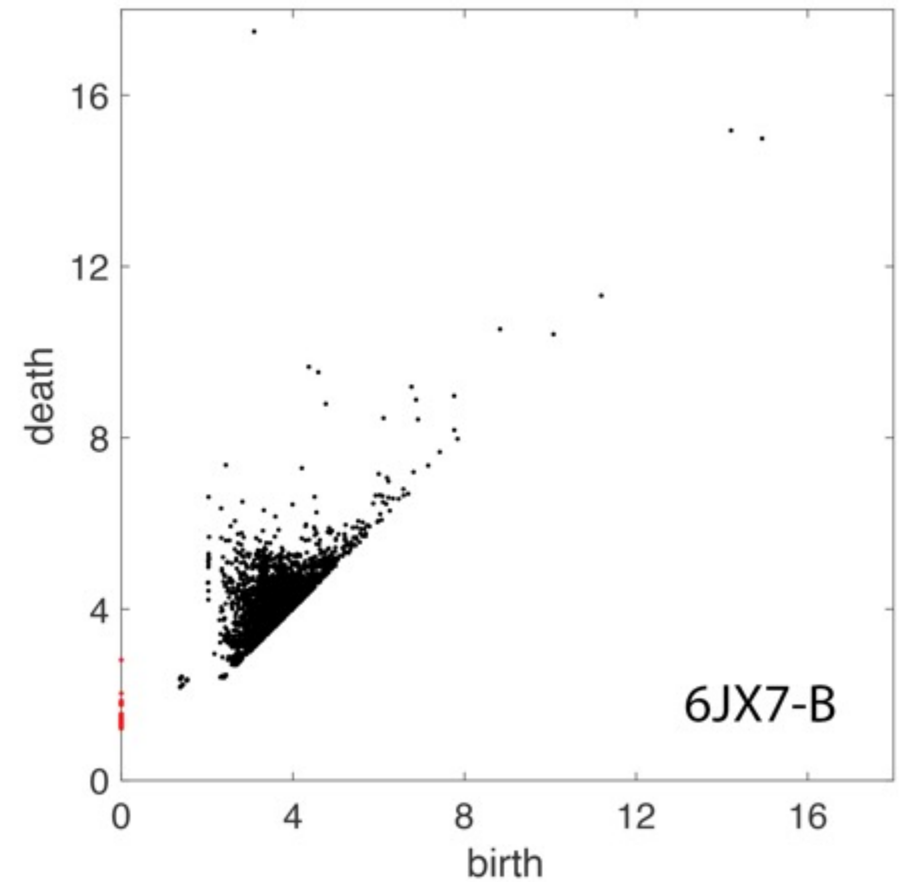
Lattice path and normalized step function

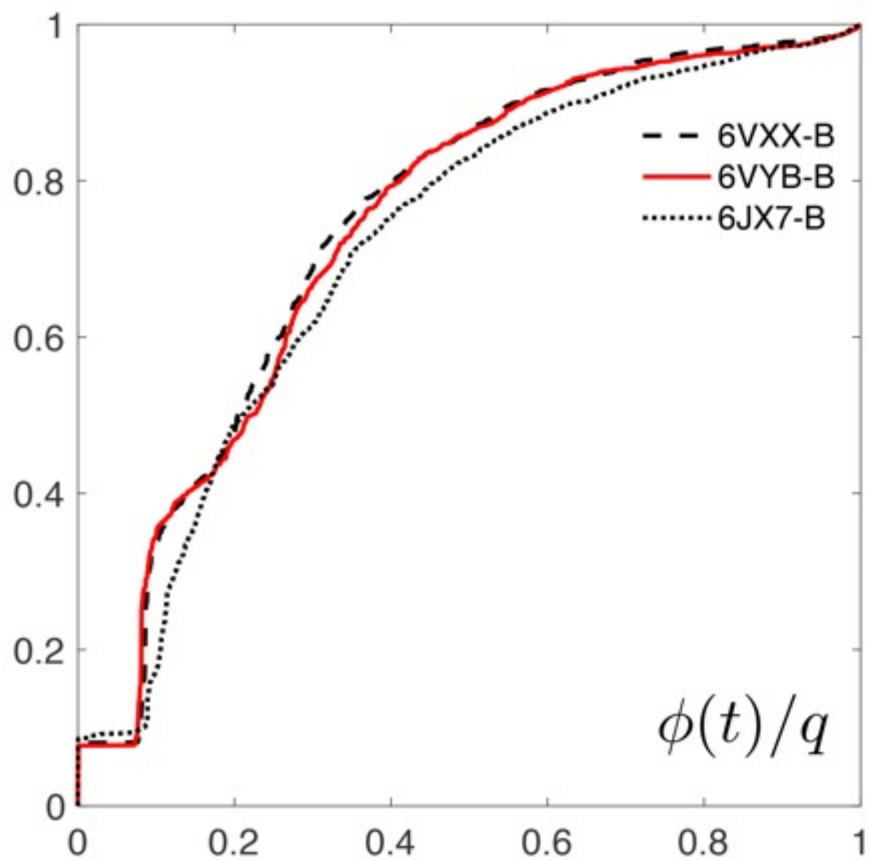












Test statistic

$$D(\phi_1, \phi_2) = \sup_{t \in [0,1]} \left| \frac{\phi_1(t)}{q_1} - \frac{\phi_2(t)}{q_2} \right|$$

Test statistic

$$D(\phi_1, \phi_2) = \sup_{t \in [0,1]} \left| \frac{\phi_1(t)}{q_1} - \frac{\phi_2(t)}{q_2} \right|$$

Upper bound of area difference

$$\int_0^1 \left| \frac{\phi_1(t)}{q_1} - \frac{\phi_2(t)}{q_2} \right| dt \leq D(\phi_1, \phi_2)$$

Birth-death processes

$$C^1 : c_1^1 < c_2^1 < \cdots < c_{q_1}^1, \quad C^2 : c_1^2 < c_2^2 < \cdots < c_{q_2}^2$$

Null hypothesis: $H_0 : C^1 = C^2$

Under null, we can interchange C^1 and C^2.

Combine C^1 and C^2:

$$c_1^1 < c_1^2 < c_2^2 < c_2^1 < \cdots < c_{q_1}^1 < c_{q_2}^2$$

$$\rightarrow \quad \uparrow \quad \uparrow \quad \rightarrow \quad \quad \rightarrow \quad \uparrow$$

Birth-death processes

$$C^1 : c_1^1 < c_2^1 < \cdots < c_{q_1}^1, \quad C^2 : c_1^2 < c_2^2 < \cdots < c_{q_2}^2$$

Null hypothesis: $H_0 : C^1 = C^2$

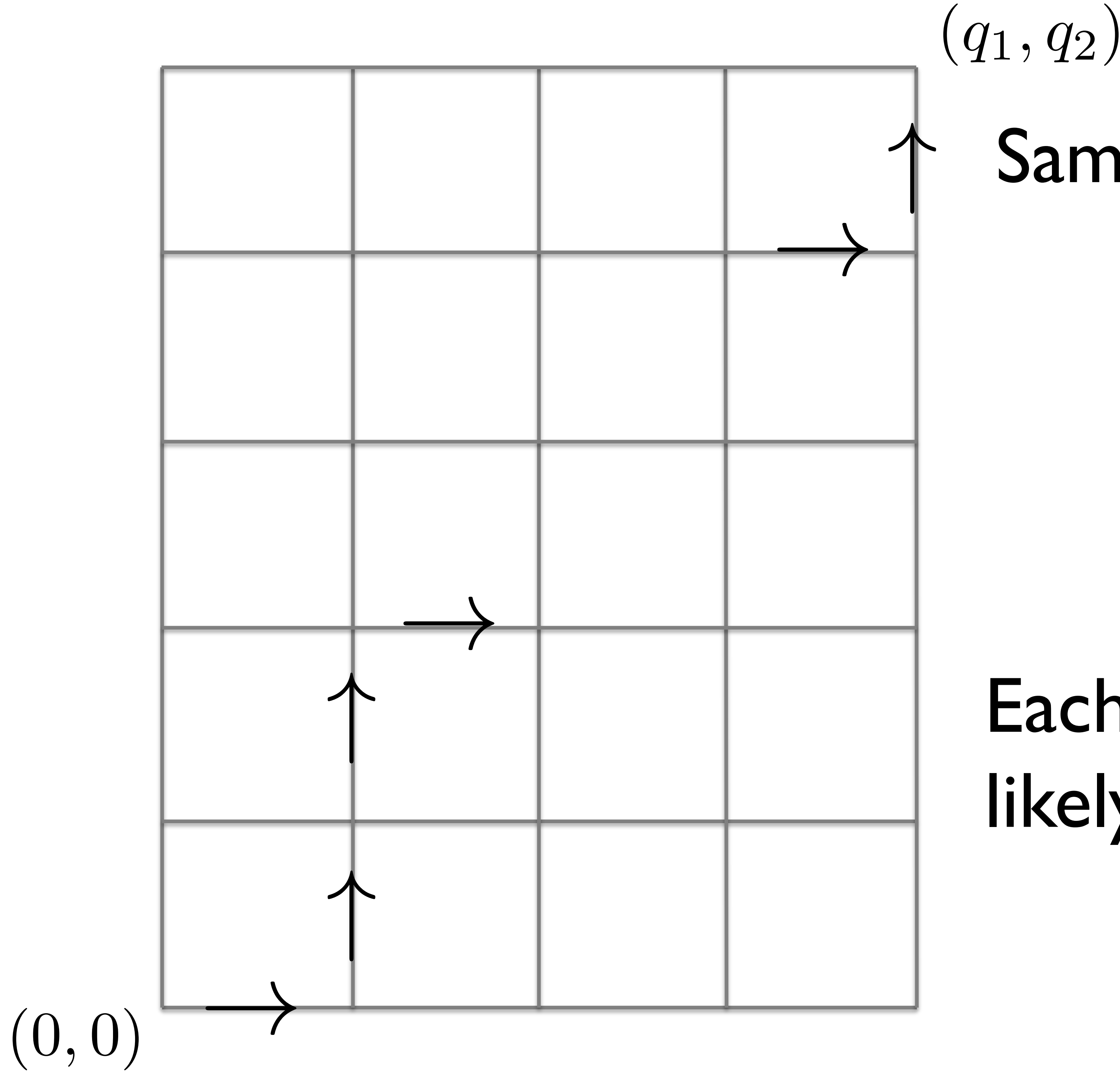
Under null, we can interchange C^1 and C^2.

Combine C^1 and C^2:

$$c_1^1 < c_1^2 < c_2^2 < c_2^1 < \cdots < c_{q_1}^1 < c_{q_2}^2$$

$$\rightarrow \quad \uparrow \quad \uparrow \quad \rightarrow \quad \cdots \quad \rightarrow \quad \uparrow$$

$\rightarrow \quad \uparrow \quad \uparrow \quad \rightarrow \quad \rightarrow \quad \uparrow$



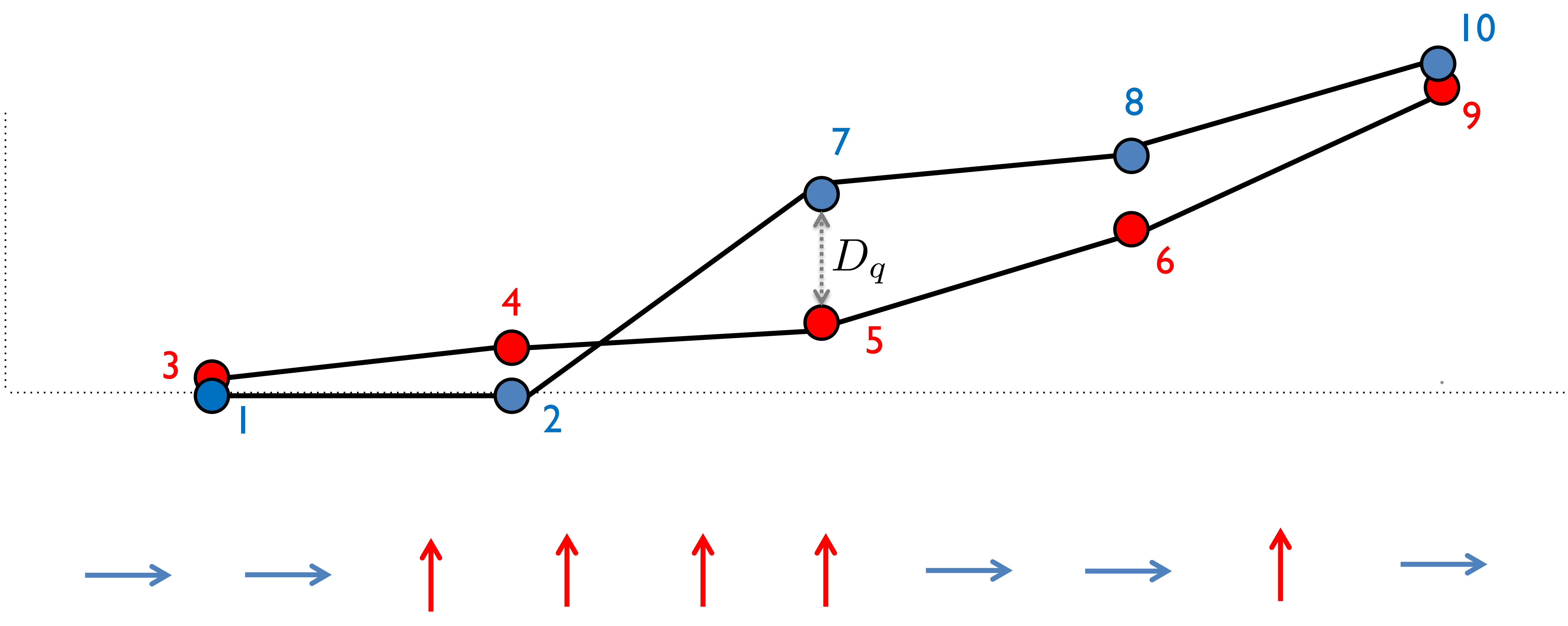
Sample space:

$\binom{q_1 + q_2}{q_1}$ number of lattice paths

Each path is equally likely with probability

$$\frac{1}{\binom{q_1+q_2}{q_1}}$$

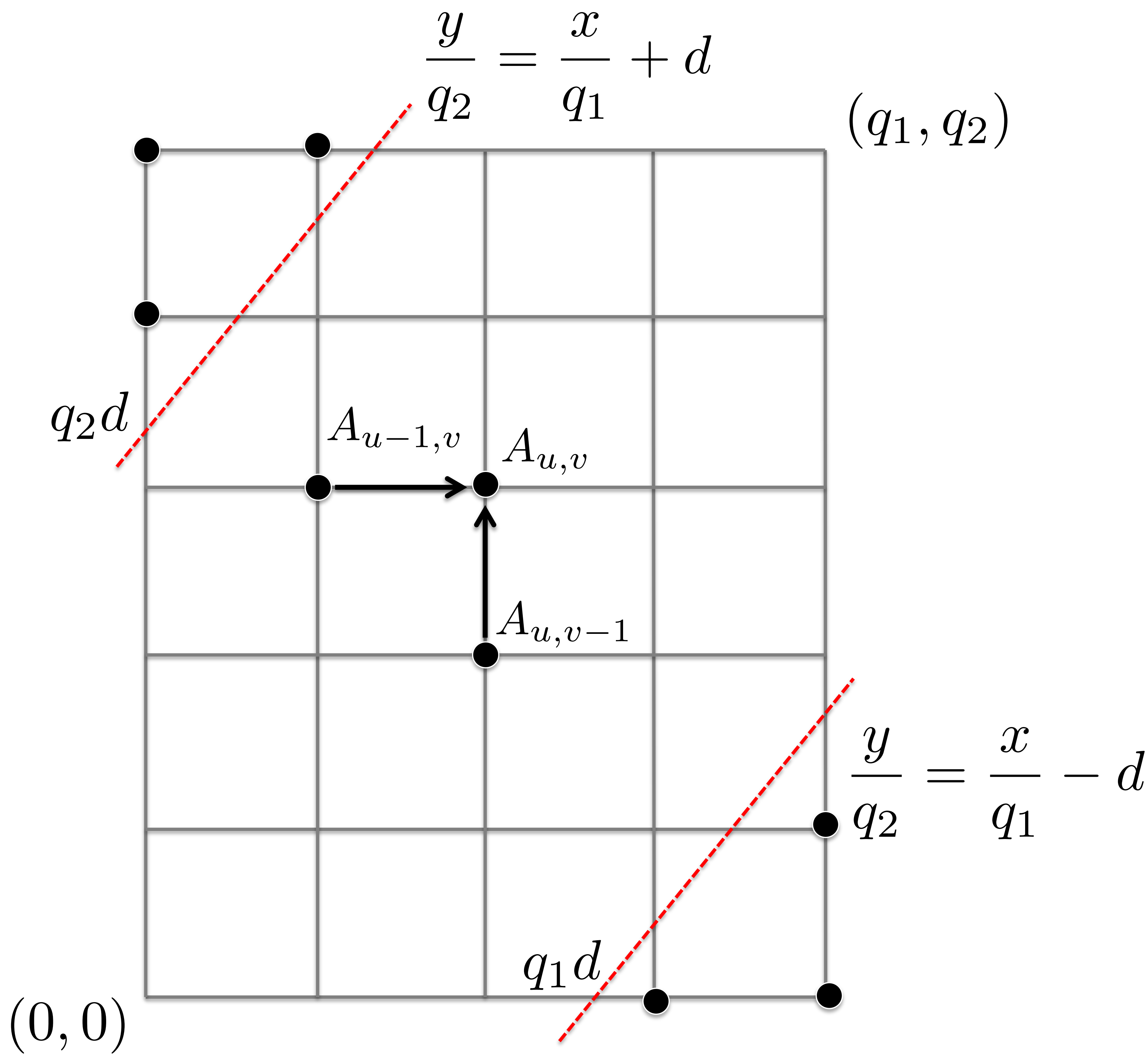
$$D(\phi_1, \phi_2) = \sup_{t \in [0,1]} \left| \frac{\phi_1(t)}{q_1} - \frac{\phi_2(t)}{q_2} \right|$$

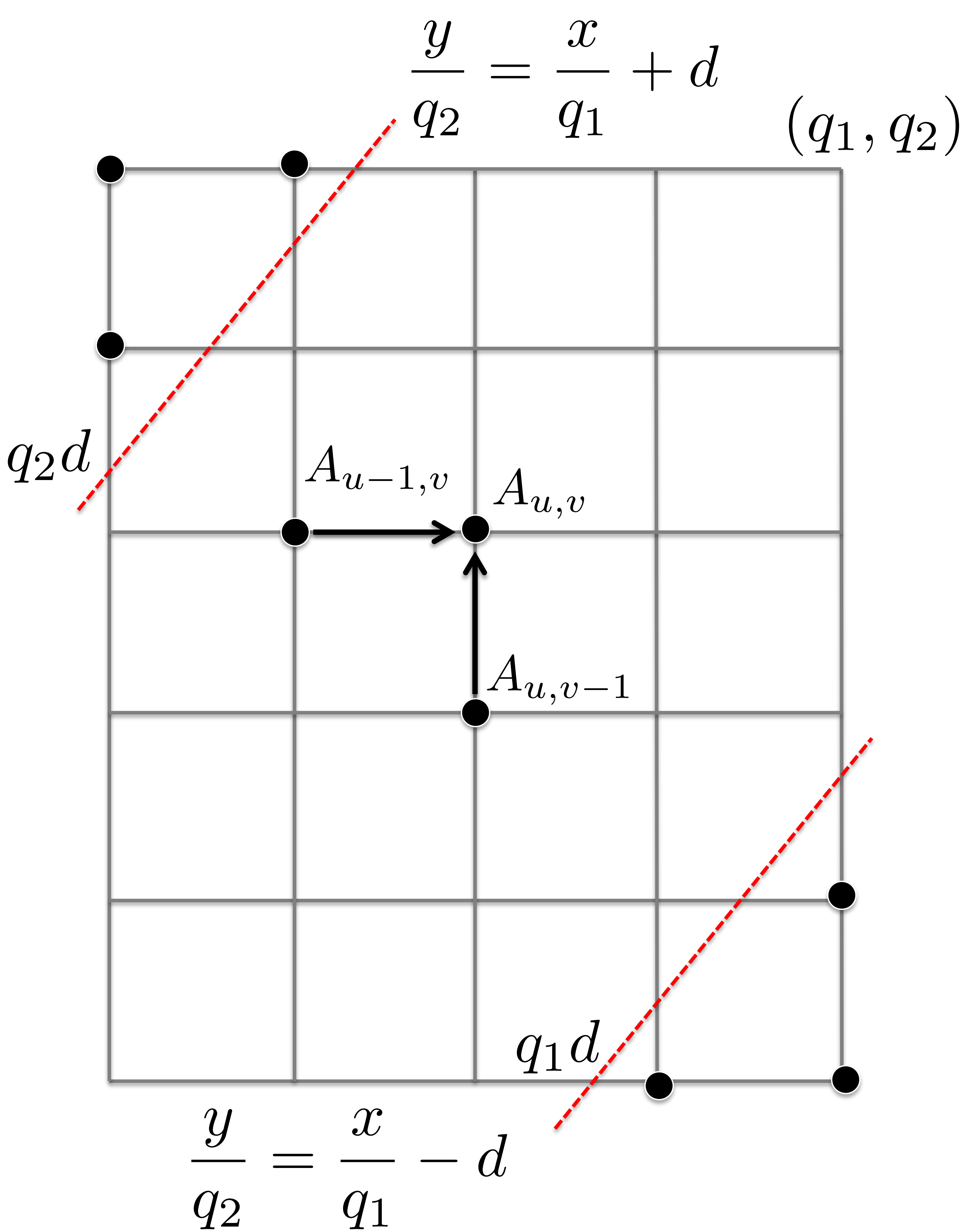


$$D(\phi_1, \phi_2) = \sup_{t \in [0,1]} \left| \frac{\phi_1(t)}{q_1} - \frac{\phi_2(t)}{q_2} \right|$$

$$P(D \geq d)$$

$$= 1 - \frac{A_{q_1,q_2}}{\binom{q_1+q_2}{q_1}}$$





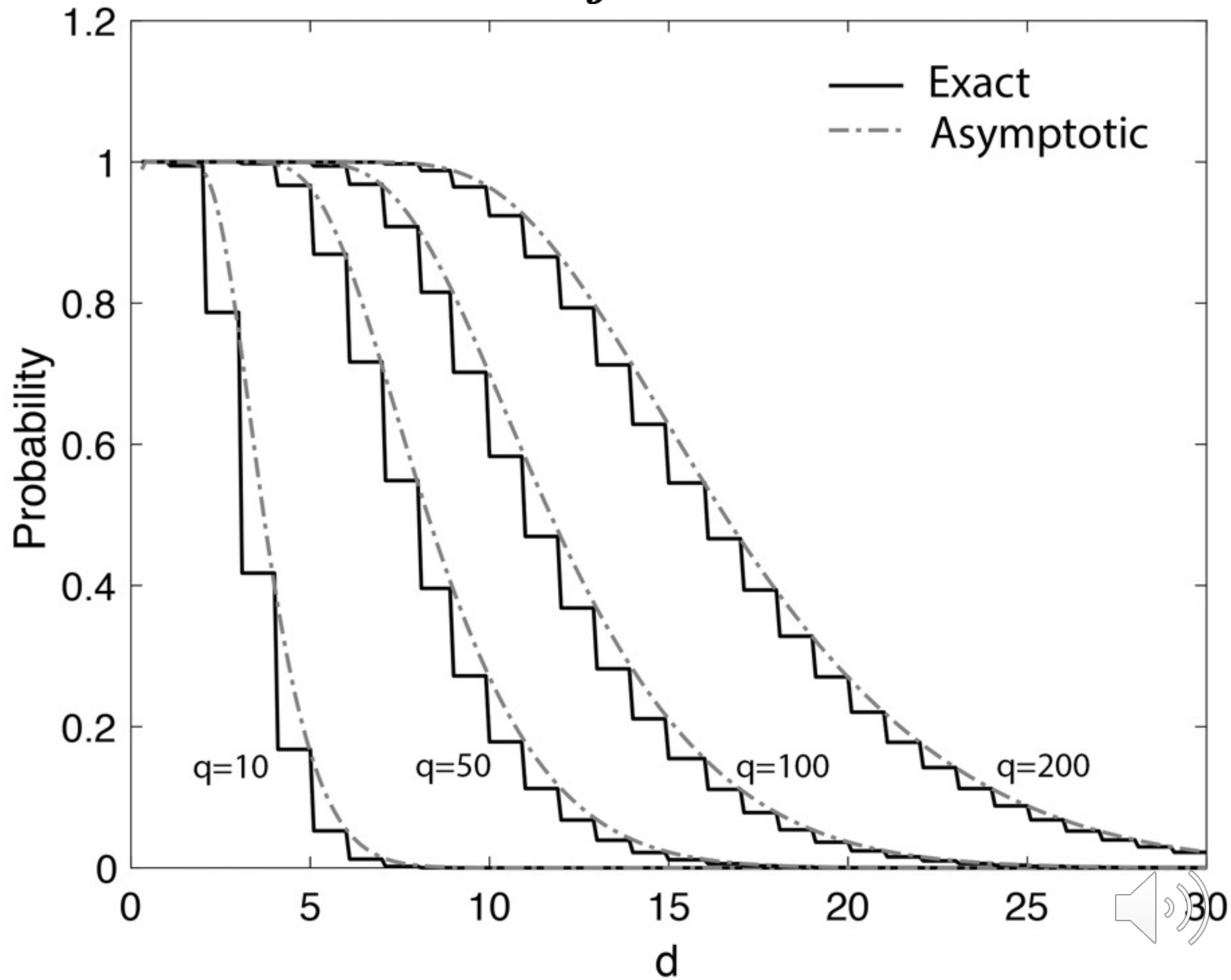
$$A_{u,v} = A_{u-1,v} + A_{u,v-1}$$

$$A_{q_1,0} = A_{0,q_2} = 1$$

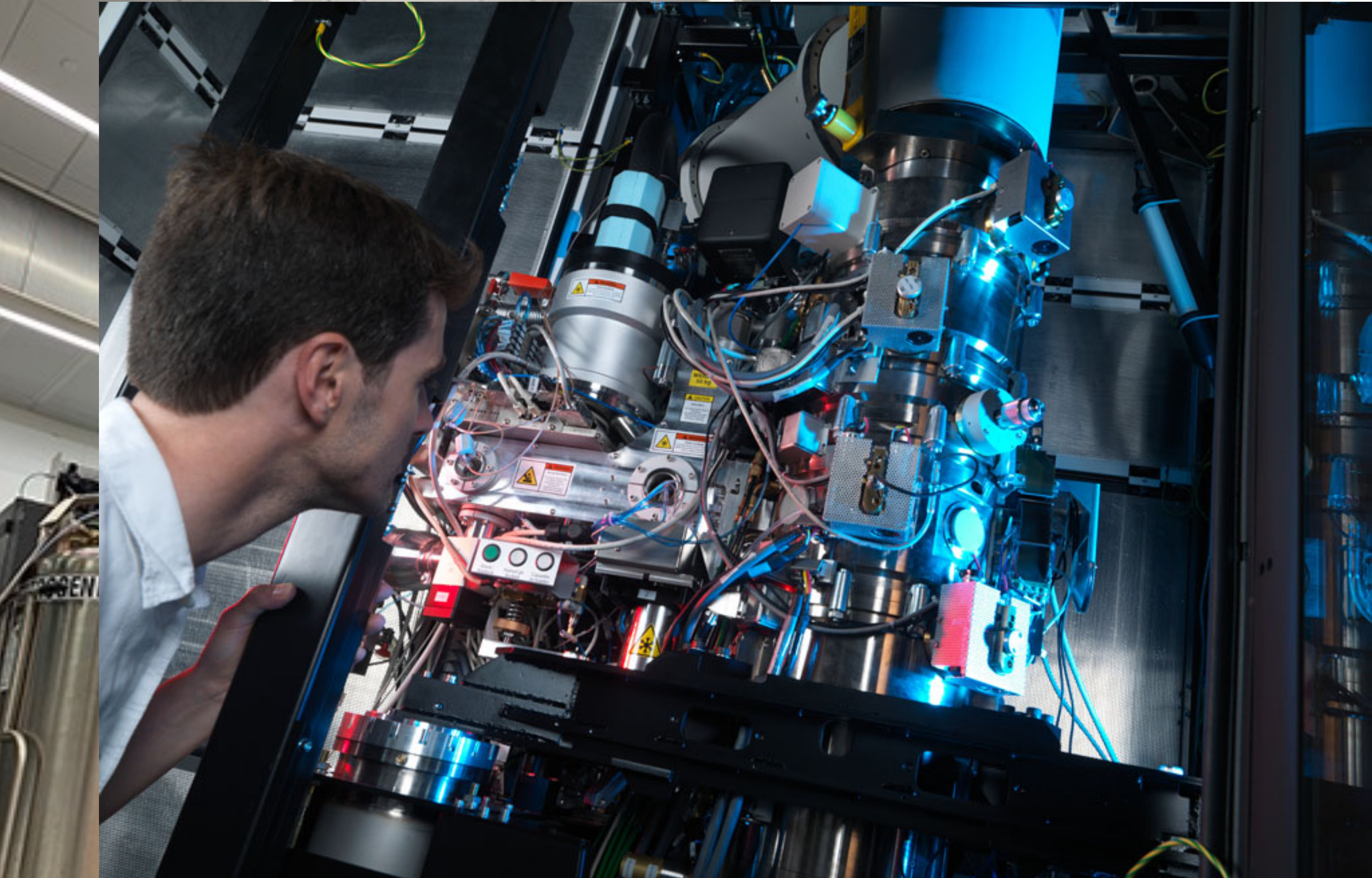
$$A_{0,0} = 0 \to A_{1,0} = A_{0,1} = 1$$
$$A_{1,1} = 2 \to \cdots \to A_{q_1,q_2}$$

Asymptotic for large-scale data

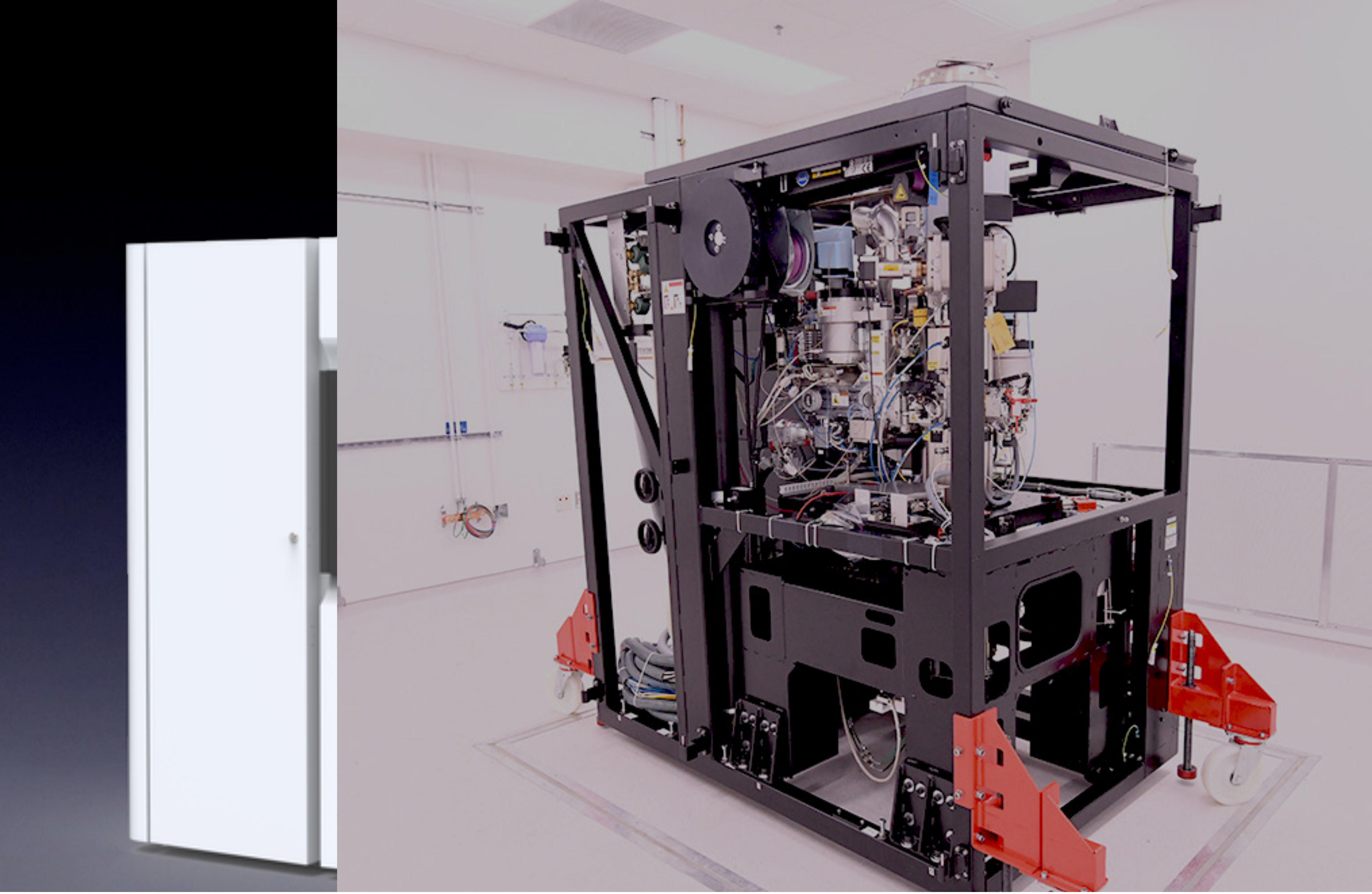
$$\lim_{q_1, q_2 \to \infty} P\left(\sqrt{\frac{q_1 q_2}{q_1 + q_2}} D \geq d\right) = 2 \sum_{j=1}^{\infty} (-1)^{j-1} e^{-2j^2 d^2}$$



$$q = q_1 = q_2$$



Thermo Fisher Scientific 300kV Titan Krios
Univ. of Wisocnsin-Madison Cyro-EM Research Center

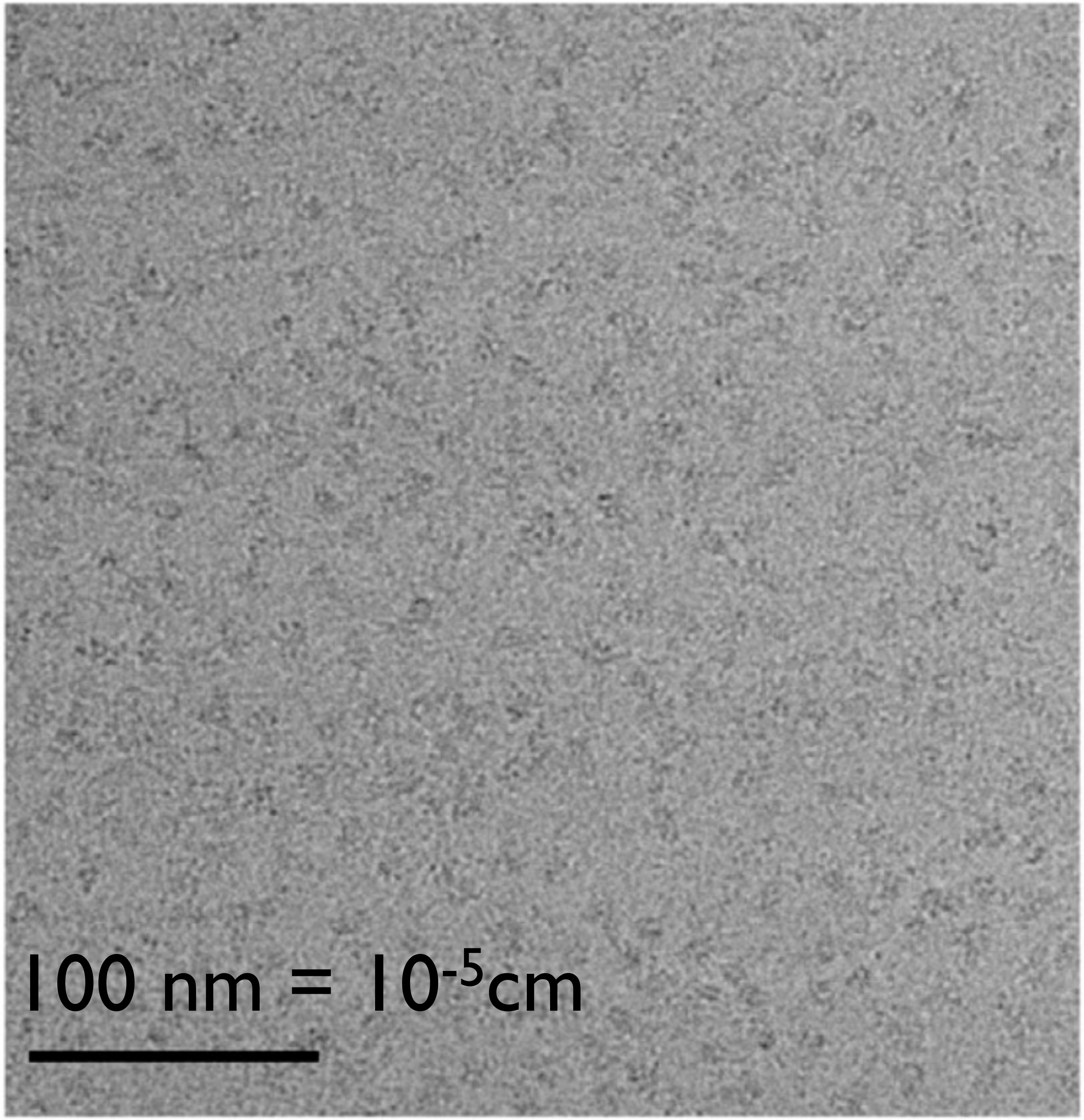


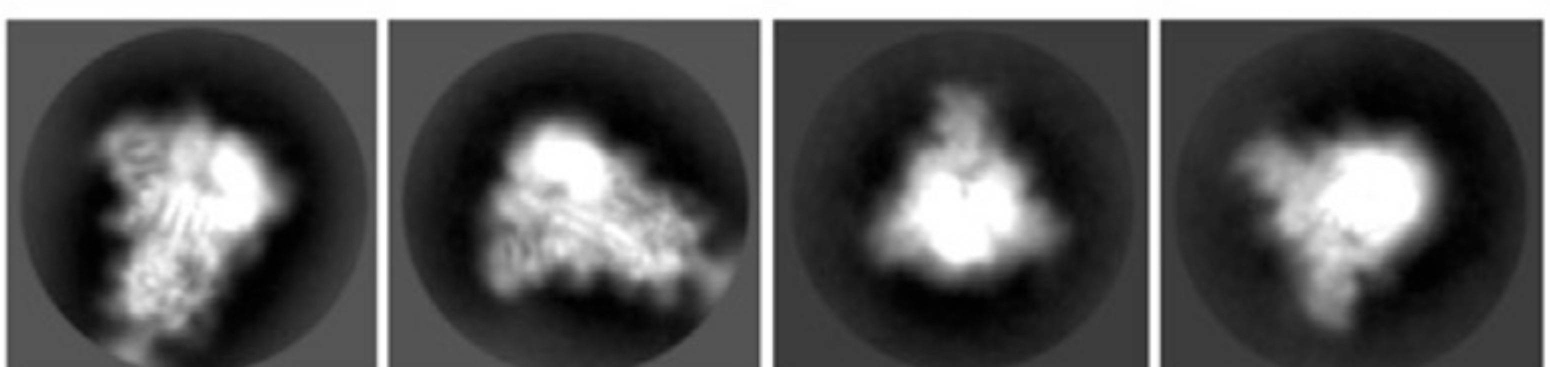
Thermo Fisher Scientific 200kV Talos Arctica Cryo-TEM
Univreristy of Wisconsin-Madison Cyro-EM Research Center

Cryogenic electron microscopy (Cryo-EM)

Cryogenic electron microscopy (cryo-EM) is an electron microscopy (EM) technique on samples at cryogenic temperatures and embedded in frozen amorphous waster with liquid ethane.

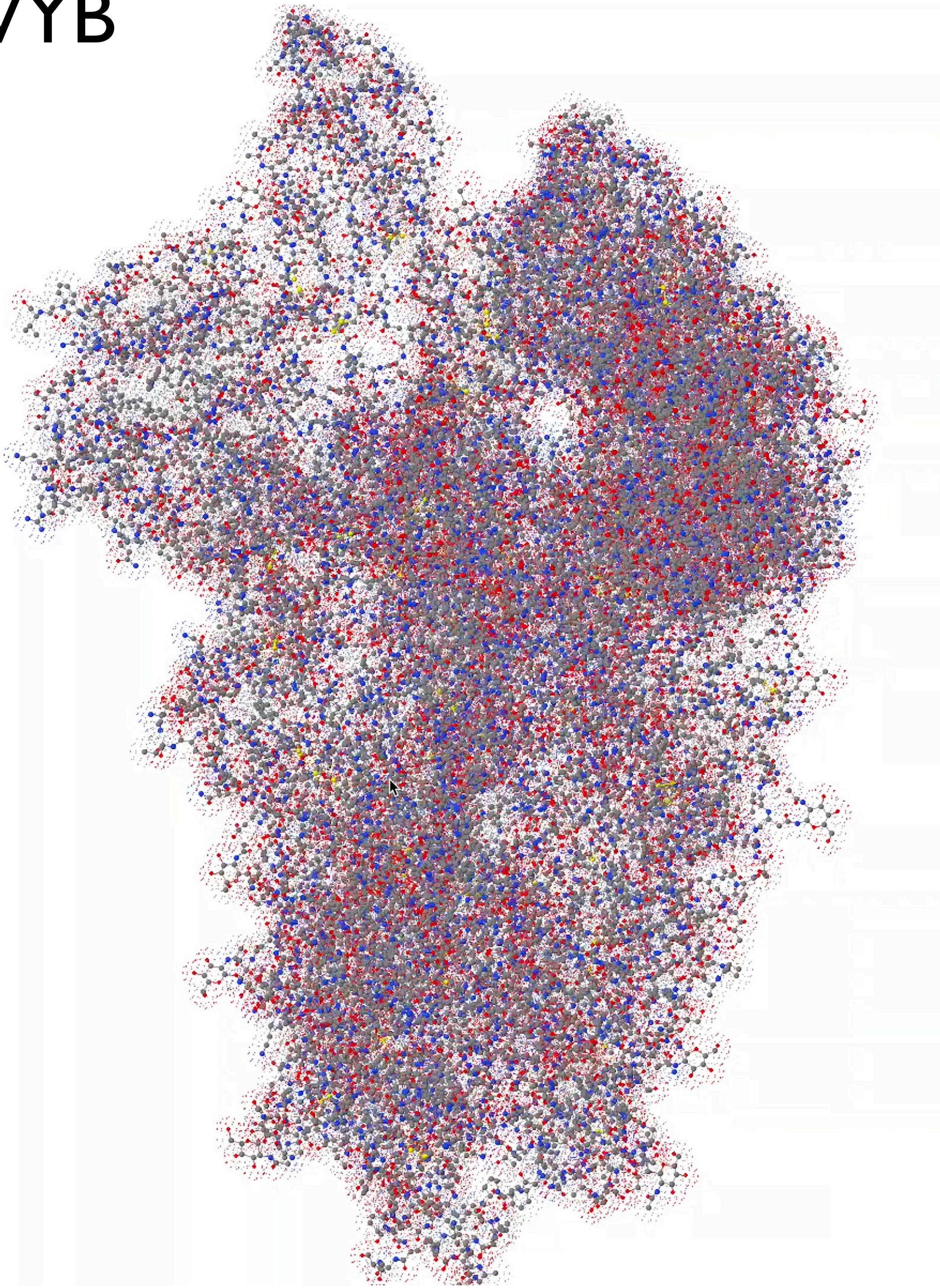
Novel prize in chemistry in 2017 to Jacques Dubochet (Univ. of Lausanne), Joachim Frank (Columbia Univ.), and Richard Henderson (Cambridge Univ.)

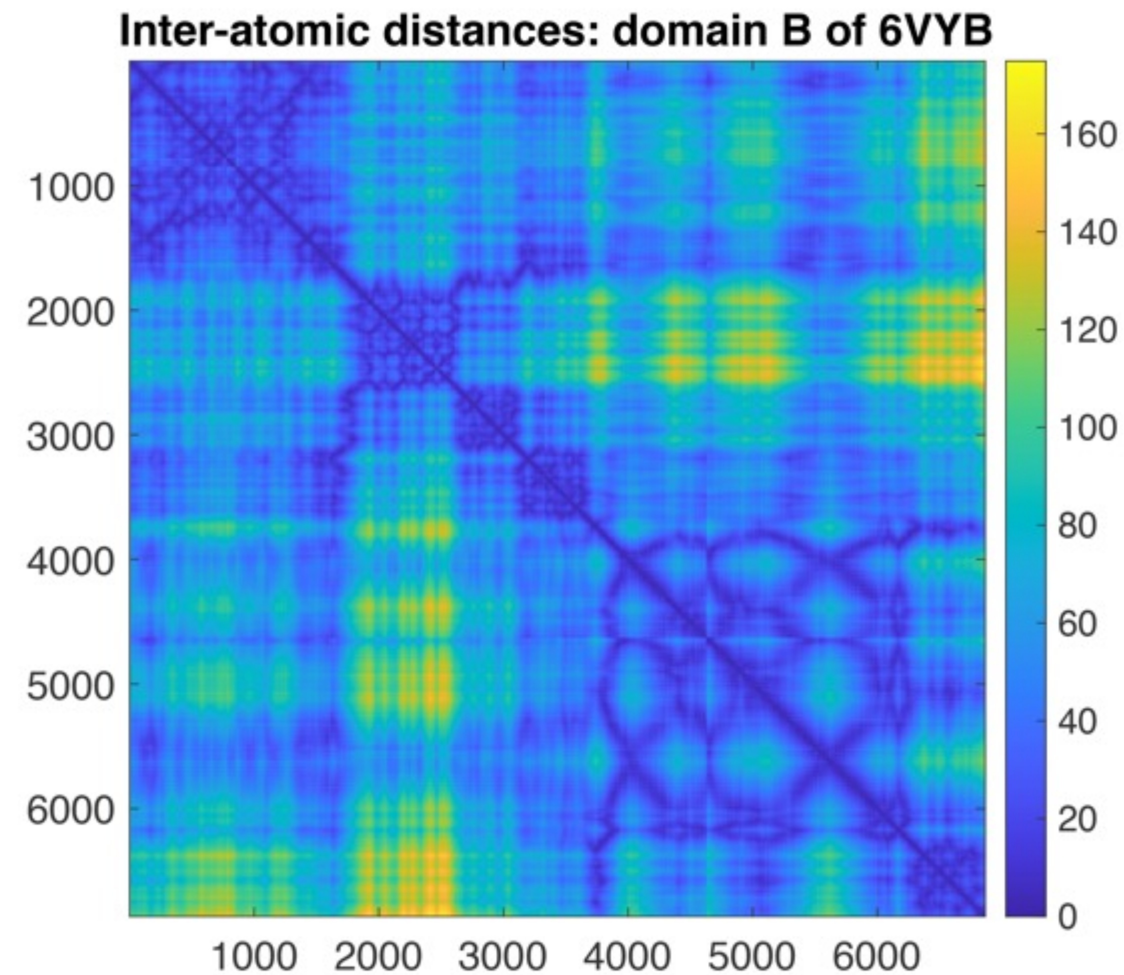




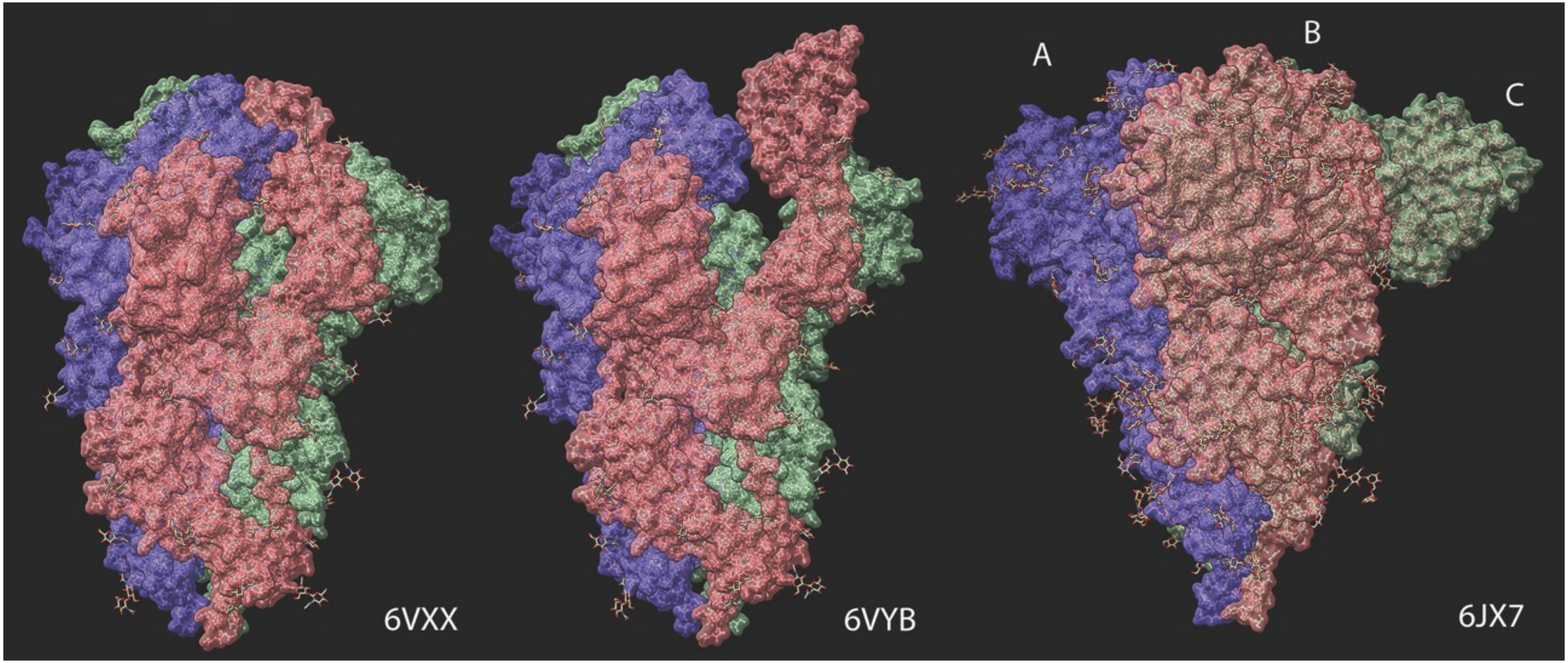
Spike protein image
From 300kV Titan Krios
Walls et al. 2020, Cell

6VYB

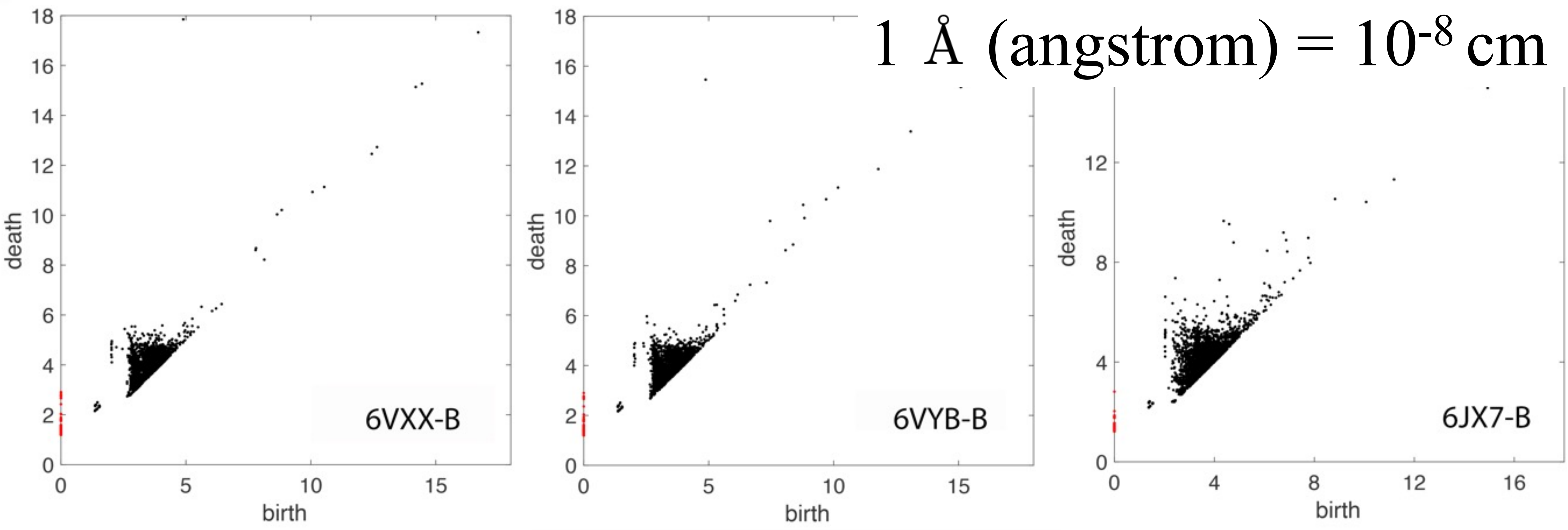




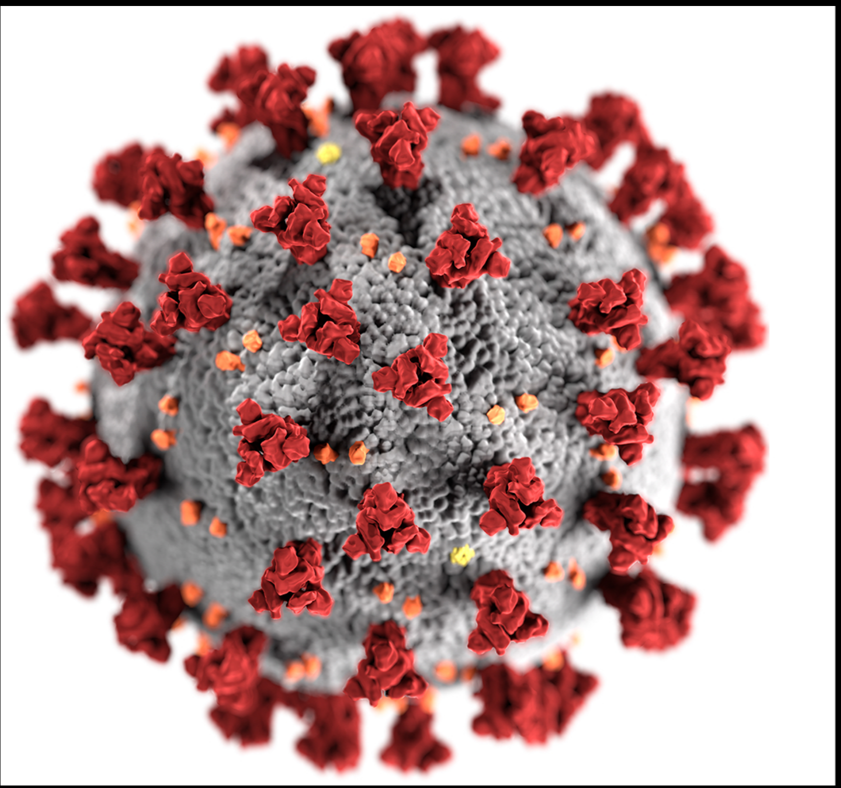
Inter-atomic distances: domain B of 6VYB
1000
2000
3000
4000
5000
6000
1000
2000
3000
4000
5000
6000
160
140
120
100
80
60
40
20
0



1 Å (angstrom) = 10^{-8} cm



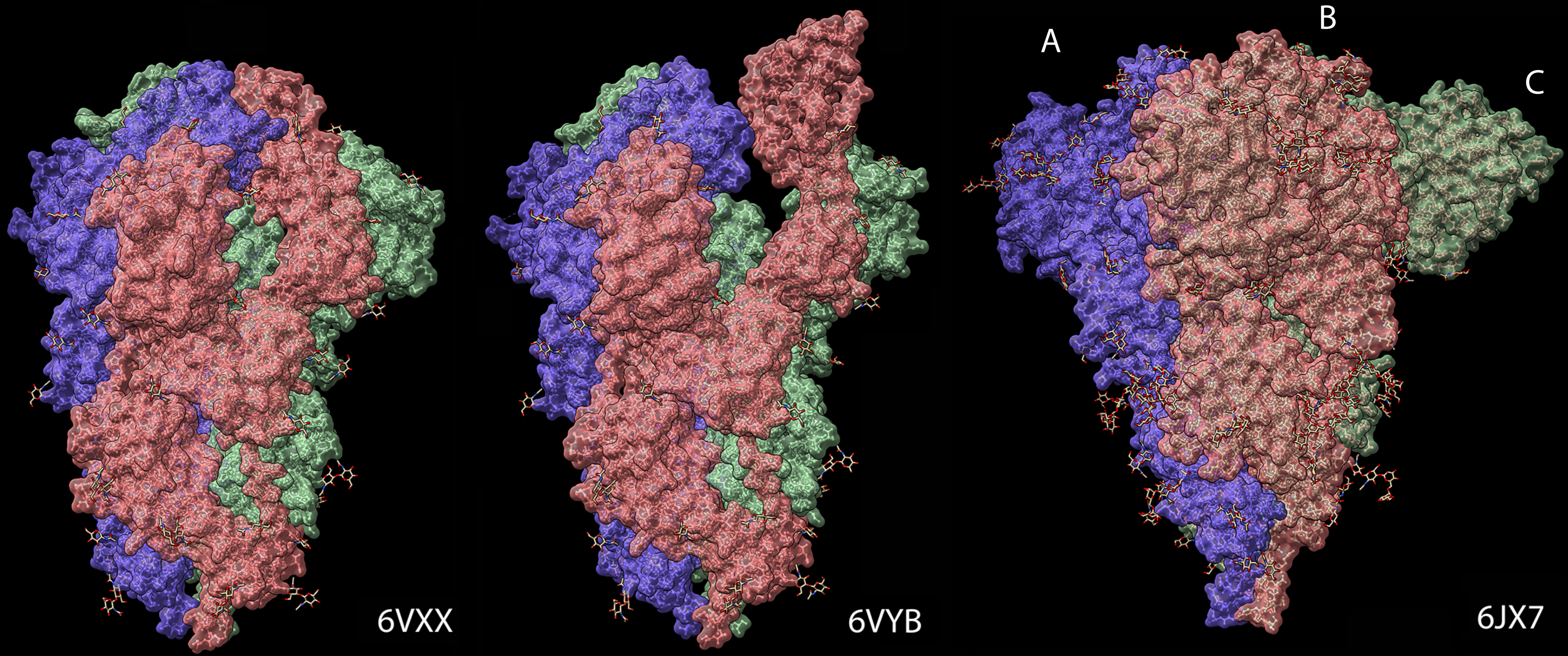
We conclude



Covid-19 virus
Closed state

COVID-19 virus
Open state

Feline corona virus



Topologically different:
p-value = 8×10^{-38}

Interpretation:

The probability of this event occurring by the random chance alone is extremely small.

Chung and Ombao, 2021 arXiv:2105:00351

Thank you! Ready for more TDA?

Sept. 27

First MICCAI workshop

http://sites.google.com/view/tda-for-medical-data

<u>Postdoc position:</u>

Combinatorial enumeration, Boltzmann machine, Bayesian learning, Ising model, interacting particles, spectral geometry, dynamical systems