#### PHY 835: Collider Physics Phenomenology

Machine Learning in Fundamental Physics

Gary Shiu, UW-Madison



#### Lecture 18: Topological Data Analysis

### References

- A. Cole and G. Shiu, ``Persistent Homology and Non-Gaussianity," JCAP 03, 025 (2018); https://arxiv.org/pdf/1712.08159.pdf
- A. Cole and G. Shiu, ``Topological Data Analysis for the String Landscape," JHEP 03, 054 (2019); <u>https://arxiv.org/pdf/1812.06960.pdf</u>
- A. Cole, A. Schachner and G. Shiu, ``Searching the Landscape of Flux Vacua with Genetic Algorithms," JHEP 11, 045 (2019); <u>https://arxiv.org/pdf/</u> <u>1907.10072.pdf</u>
- M. Biagetti, A. Cole and G. Shiu, ``The Persistence of Large Scale Structures I: Primordial non-Gaussianity," https://arxiv.org/pdf/ 2009.04819.pdf
- A. Cole, G.J. Loges and G. Shiu, ``Quantitative and Interpretable Order Parameters for Phase Transitions from Persistent Homology," https:// arxiv.org/pdf/2009.14231.pdf

### Data is BIG

#### **Cosmology** is marching into a big data era:

| Experimental Data | 2013                             | 2020                | 2030+                   |
|-------------------|----------------------------------|---------------------|-------------------------|
| Storage           | 1PB                              | 6PB                 | 100-1500PB              |
| Cores             | $10^{3}$                         | 70K                 | 300+K                   |
| CPU hours         | $3 \mathrm{x} 10^6 \mathrm{hrs}$ | $2 \times 10^8$ hrs | $\sim 10^9 \text{ hrs}$ |
| Simulations       | 2013                             | 2020                | 2030+                   |
| Storage           | 1-10 PB                          | 10-100PB            | > 100PB - 1EB           |
| Cores             | 0.1-1M                           | 10-100M             | > 1G                    |
| CPU hours         | 200M                             | >20G                | $> 100 {\rm G}$         |

|        | data volume | schedule     |
|--------|-------------|--------------|
| SDSS   | 40 TB       | 2000-2020    |
| DESI   | 2 PB        | 2019-2027    |
| LSST   | > 60 PB     | 2020-2030    |
| Euclid | >10 PB      | 2020-2027    |
| WFIRST | >2 PB       | 2023-2030    |
| CMB-S4 | Ø(1) (PB)   | 2020-2027(?) |
| SKA    | 4.6 EB      | 2019-2030(?) |

Table taken from 1311.2841

### Data is BIG

#### **Cosmology** is marching into a big data era:

| Experimental Data | 2013                             | 2020                | 2030 +                  |
|-------------------|----------------------------------|---------------------|-------------------------|
| Storage           | 1PB                              | 6PB                 | 100-1500PB              |
| Cores             | $10^{3}$                         | 70K                 | 300+K                   |
| CPU hours         | $3 \mathrm{x} 10^6 \mathrm{hrs}$ | $2 \times 10^8$ hrs | $\sim 10^9 \text{ hrs}$ |
| Simulations       | 2013                             | 2020                | 2030 +                  |
| Storage           | 1-10 PB                          | 10-100PB            | > 100PB - 1EE           |
| Cores             | 0.1 <b>-</b> 1M                  | 10-100M             | > 1G                    |
| CPU hours         | 200M                             | >20G                | $> 100 {\rm G}$         |

|        | data volume | schedule     |
|--------|-------------|--------------|
| SDSS   | 40 TB       | 2000-2020    |
| DESI   | 2 PB        | 2019-2027    |
| LSST   | > 60 PB     | 2020-2030    |
| Euclid | >10 PB      | 2020-2027    |
| WFIRST | >2 PB       | 2023-2030    |
| CMB-S4 | Ø(1) (PB)   | 2020-2027(?) |
| SKA    | 4.6 EB      | 2019-2030(?) |

Table taken from 1311.2841

#### ~ 200PB of *archived data* in the first 7 years of the LHC.

## Data is BIG

#### **Cosmology** is marching into a big data era:

| Experimental Data | 2013                             | 2020                | 2030 +                  |
|-------------------|----------------------------------|---------------------|-------------------------|
| Storage           | 1PB                              | 6PB                 | 100-1500PB              |
| Cores             | $10^{3}$                         | 70K                 | 300+K                   |
| CPU hours         | $3 \mathrm{x} 10^6 \mathrm{hrs}$ | $2 \times 10^8$ hrs | $\sim 10^9 \text{ hrs}$ |
| Simulations       | 2013                             | 2020                | 2030 +                  |
| Storage           | 1-10 PB                          | 10-100PB            | > 100PB - 1EE           |
| Cores             | 0.1-1M                           | 10-100M             | > 1G                    |
| CPU hours         | 200M                             | >20G                | $> 100 { m G}$          |

|        | data volume | schedule     |
|--------|-------------|--------------|
| SDSS   | 40 TB       | 2000-2020    |
| DESI   | 2 PB        | 2019-2027    |
| LSST   | > 60 PB     | 2020-2030    |
| Euclid | >10 PB      | 2020-2027    |
| WFIRST | >2 PB       | 2023-2030    |
| CMB-S4 | Ø(1) (PB)   | 2020-2027(?) |
| SKA    | 4.6 EB      | 2019-2030(?) |

Table taken from 1311.2841

#### ~ 200PB of *archived data* in the first 7 years of the LHC.

In terms of sheer volume, **not**hing **trumps** the volume of **string theoretical data**. It has been estimated that the number of vacua:

 $10^{500}$  (Type IIB flux vacua)  $10^{272,000}$  (F theory flux vacua)

### **Dimensionality Reduction**

- This data lives in high-dimensional phase space
  - N particles: dim  $\sim N_{\text{particles}}$
  - Function on sphere: dim  $\sim N_{\rm pix} \sim \ell_{\rm max}^2$
- How do we compress the data into their most relevant (and interpretable) features?





### The String Landscape



# The String Landscape





## The String Landscape





### **Distribution of Large Scale Structure**

Similar **clustering** and **void** features also appear in LSS:



#### Data is Subtle

#### N-body Simulations of the Large Scale Structure of the Universe:



$$f_{NL}^{\text{local}} = -250 \qquad \qquad f_{NL}^{\text{local}} = 0 \qquad \qquad f_{NL}^{\text{local}} = 250$$

 $f_{NL}^{local} = -0.9 \pm 5.1 \text{ (Planck)}$ 

https://mbiagetti.gitlab.io/cosmos/nbody/eos/

#### Data is Subtle

- Given a statistical system, the basic questions are:
  - How many phases are there? (unsupervised learning)
  - How are different phases distinguished? (supervised learning)
- Phase detection and classification can be subtle, e.g., the XY model:

$$H_{\rm XY} = -\sum_{\langle i,j \rangle} \cos\left(\theta_i - \theta_j\right)$$

has an **infinite order (Kosterlitz-Thouless)** phase transition with vortexantivortex pairs at low temperatures.

- First attempts to study the XY model and its KT transition using neutral networks and PCA failed in identifying vortices at low temperatures.
- Moreover, ML methods often lack the desired level of interpretability. Order parameters? Critical exponents?



Computer answer: "100 points with the following coordinates..." .



Computer answer: "100 points with the following coordinates..." Human answer: a noisy circle/ points sampled from an annulus



## **Topological Data Analysis**

- When the space of data is huge, we cannot simply "visualize" the structure of data. We need a systematic diagnostic tool.
- Topological data analysis (TDA) is a systematic tool in applied topology to diagnose the "shape" of data.
- To turn a discrete set of data points (point cloud) into a topological space, we need a notion of *persistence*.



Vary simplicial complexes formed by the point cloud with continuous parameters (filtration parameters)

## **Topological Data Analysis**

- TDA is widely used in other fields, e.g., imaging, neuroscience, drug design; little (if any) has been explored in physics.
- TDA can be used in conjunction with ML: presisten homology teaches us the grammar of data; natural to define a log of loss function.
- We developed TDA for a variety of physics contexts:





String Landscape [Cole, GS, '18]; [Cole, Schachner, GS, '19]

CMB [Cole, GS, '17] Large Scale Structures [Biagetti, Cole, GS, '20]

Phases of Matter [Cole, Loges, GS, '20]

# **Topological Data Analysis**

## Topology

The "shape" properties of an object that are preserved under **continuous deformations** (e.g., stretching, twisting, crumpling, bending) but **not tearing or gluing**.





## Topology

The "shape" properties of an object that are preserved under **continuous deformations** (e.g., stretching, twisting, crumpling, bending) but **not tearing or gluing**.





## Topology

The "shape" properties of an object that are preserved under **continuous deformations** (e.g., stretching, twisting, crumpling, bending) but **not tearing or gluing**.





## **Topology in Physics**



#### **String Theory**

Particle spectrum & couplings, cosmological models, ...

#### **Condensed Matter Physics**

Quantum Hall effect, topological materials, ...

## **Simplicial Complexes**

- In  $\mathbb{R}^3$ , simplices are vertices, edges, triangles, and tetrahedra
- Simplicial complexes are collections of simplices that are:
  - Closed under intersection of simplices
  - Closed under taking faces of simplices
- Combinatorial representations easy calculations for computers



Source: Wikipedia, "Simplicial Complex"

## Simplicial Homology

 Given a simplicial complex, We want to count independent p-cycles (i.e. p-loops) that are not boundaries of higher-dimensional objects

#### • Betti numbers:

- 0-th Betti number is number of connected components
- p-th Betti number is number of independent p-loops
- In practice, homology calculation is a matrix reduction.



#### Persistence

- How to choose simplicial representation of our data?
- Persistent homology: vary simplicial representation  $\Sigma_{\nu}$  of data with some filtration parameter  $\,\nu\,$  such that

$$\nu_1 \leq \nu_2 \implies \Sigma_{\nu_1} \subseteq \Sigma_{\nu_2}$$

- Track each distinct feature's lifetime (birth and death)
- Intuition: "real" topological features persist, short-lived features are noise
- Procedure is stable against perturbations to data [Cohen-Steiner 2005]









 $\nu = 5$ 



## **Delaunay Triangulation**

- The Vietoris-Rips construction has a *clique problem.*
- We can more efficiently consider subcomplexes of the Delaunay triangulation.
  - We can use an  $\alpha$ -filtration corresponding to these subcomplexes
    - Also variations using DTM function to account for outliers
- The size of the Delaunay complex grows only as N<sup>[D/2]</sup> for D≥3, in contrast to VR complex which grows exponentially with N.



VR: more simplices than necessary



Delaunay triangulation

## Visualizing Persistent Homology

 Persistence diagrams are scatter plots of birth and death times for *individual* homology generators:



- Persistence image: vector space representation useful for deriving statistics
  - Construct density function by smoothing each cycle in diagram via kernel with persistence-dependent weight, e.g.  $log(1 + \nu_{persist})$ .
  - Downsample/average over grid for lower dimension statistics.



### **Topological Curves**

- Topological curves: count cycles in particular regions of diagram.
  - $D_p$ : deaths
  - $P_p$ : persistences
  - $B_p$ : births

•  $b_p = -B_p + D_p$ : Betti numbers

 $D_p, P_p, B_p$  have interpretation of empirical distribution functions for deaths, etc. of cycles



## Applying TDA to Cosmology

![](_page_36_Picture_1.jpeg)

Matteo Biagetti

![](_page_36_Picture_3.jpeg)

Alex Cole

- "Persistent Homology and Non-Gaussianity", A. Cole and GS, JCAP 1803, 025 (2018) [arXiv:1712.08159 [astro-ph.CO]].
- "The Persistence of Large Scale Structures I: Primordial non-Gaussianity", M. Biagetti, A. Cole and GS, [arXiv:2009.04819 [astro-ph.CO]].

![](_page_36_Picture_7.jpeg)

## Inflation

#### [Starobinsky];[Guth];[Linde];[Albrecht, Steinhardt];...

- Period of accelerated expansion in early universe
  - Solves flatness, horizon, and monopole problems
  - Predicts nearly scale-invariant, Gaussian curvature fluctuations
    - Source anisotropies in CMB, inhomogeneities in LSS
- A myriad of models. Taxonomy done mostly through their observables (n<sub>s</sub>, r)

![](_page_37_Picture_7.jpeg)

![](_page_37_Picture_8.jpeg)

### Anisotropies

• The lowest order correlation we can extract is the **power spectrum**:

$$\left\langle 0 \left| \hat{\mathcal{R}}_{\mathbf{k_1}} \hat{\mathcal{R}}_{\mathbf{k_2}} \right| 0 \right\rangle = (2\pi)^3 P_{\mathcal{R}}(k_1) \delta(\mathbf{k_1} + \mathbf{k_2}) \qquad \Delta_{\mathcal{R}}^2 = \left( \frac{k^3}{2\pi^2} \right) P_{\mathcal{R}}^2 \propto k^{n_s - 1}$$

- For a Gaussian theory, the power spectrum dictates all higher-pt correlations. But inflationary fluctuations are not perfectly Gaussian.
- The leading **non-Gaussianity** is the **bispectrum**:

$$\langle 0 | \hat{\mathcal{R}}_{\mathbf{k_1}} \hat{\mathcal{R}}_{\mathbf{k_2}} \hat{\mathcal{R}}_{\mathbf{k_3}} | 0 \rangle = (2\pi)^3 \, \delta^3 (\mathbf{k_1} + \mathbf{k_2} + \mathbf{k_3}) F(\mathbf{k_1}, \mathbf{k_2}, \mathbf{k_3})$$

- Scaling and symmetries imply that F(k<sub>1</sub>, k<sub>2</sub>, k<sub>3</sub>) is fixed by an overall size ~ f<sub>NL</sub> and its ''shape'' F(1, k<sub>2</sub>/k<sub>1</sub>, k<sub>3</sub>/k<sub>1</sub>).
- More **powerful discriminator** of inflationary models.

### Non-Gaussianities

- The bispectrum for single field slow-roll inflation was computed in [Maldacena, '02];[Acquaviva et al, '02]; its size is f<sub>NL</sub> ~ O(ε,η):
- The bispectrum for **general single field inflation** was found to be parametrized by 5 parameters [Chen, Huang, Kachru, GS, '06]:

![](_page_39_Figure_3.jpeg)

 There is also an "orthogonal shape" but it "looks" qualitatively like the equilateral shape (*challenging task for machine learning*).

### Non-Gaussianities

• More complicated models can give rise to more shapes:

![](_page_40_Figure_2.jpeg)

- Like scattering amplitudes in particle physics, non-Gaussianties can reveal interactions governing inflation: *cosmological collider*.
- In collider physics: use *different strategies* for different particles.

## Measuring Non-Gaussianity

Harmonic space: fits with <u>templates</u> of bispectrum, trispectrum, etc.
 One can define a "cosine" between distributions:

$$\cos(F_1, F_2) = \frac{F_1 \cdot F_2}{(F_1 \cdot F_1)^{1/2} (F_2 \cdot F_2)^{1/2}}$$

• Some shapes are harder to find, e.g.,

![](_page_41_Figure_4.jpeg)

Resonant shape (axion monodromy)

• Local NG: scale-dependent bias [Dalai et al]; [Matarrese, Verde]; [Slosar et al]

$$P_{hh}(k) = \left(b_g + \frac{12}{5} \frac{f_{\rm NL}^{\rm loc}}{\mathcal{M}(k)} b_{\zeta}\right)^2 P_{mm}(k) \qquad \mathcal{M}(k) \sim k^2 \qquad \text{at small k}$$
(large scales)

 Persistent homology probes multi-scale topology of the LSS data; turns out to be most sensitive to smaller scales [Biagetti, Cole, GS]

### **TDA on Non-Gaussianity**

![](_page_42_Picture_1.jpeg)

[Cole, GS, '17]

- We can neglect non-Gaussanity from non-linearities of gravity.
- Current bound from Planck:

$$f_{NL}^{local} = -0.9 \pm 5.1$$
  $f_{NL}^{equil} = -26 \pm 47$ 

![](_page_42_Picture_6.jpeg)

[Biagetti, Cole, GS, '20]

- Larger number of modes
- Higher dimensional topologies: clusters, filaments, and voids are naturally phrased in topology.
- Topology probes tail of distribution
   →enhanced in late time observables.

## TDA for CMB

#### [Cole, GS, '17]

![](_page_43_Figure_2.jpeg)

- Sublevel topology changes when threshold passes critical points (cf. Morse theory).
- Topological simplification: approximate function by its network of critical points.

(Hotter points are deeper red)

![](_page_44_Picture_2.jpeg)

![](_page_45_Figure_1.jpeg)

Many distinct components, no loops

![](_page_45_Figure_3.jpeg)

![](_page_45_Figure_4.jpeg)

 $\nu = 0$ 

Many loops, fewer distinct components

(Sublevel set in black)

![](_page_46_Picture_4.jpeg)

![](_page_47_Figure_1.jpeg)

One connected component, many loops have been filled in

(Sublevel set in black)

![](_page_47_Picture_4.jpeg)

## TDA for CMB

#### [Cole, GS, '17]

- As warmup, we carried out TDA for local NG and with low-resolution CMB simulations [Elsner, Wandelt] ( $\ell_{max} \sim 1024$ ) with varying  $f_{NL}^{\rm local}$ .
- We binned the persistence diagrams & performed likelihood analysis:

![](_page_48_Figure_4.jpeg)

 NB: WMAP-resolution simulations. More important is the improvement of the sensitivity of topological statistics by factor of ~2 (compare to Betti numbers).

### TDA for LSS

#### [Biagetti, Cole, GS, '20]

![](_page_49_Picture_2.jpeg)

Cosmology surveys provide access to ~10 billion galaxies

Data's topology is more efficiently computed with  $\alpha$ -filtration.

![](_page_49_Picture_5.jpeg)

## Scales of Topology

![](_page_50_Figure_1.jpeg)

- Features at  $\mathcal{O}(10) \ll 1000$  Mpc/h
- Contrast with scale-dependent bias in halo power spectrum which is sensitive to largest scales.
- We can compute persistence in sub-boxes of full simulations and subsample to uniform halo number.
- Besides voids, filament loops provide a new, competitive observable.

## **TDA for LSS: Pipeline**

#### [Biagetti, Cole, GS]

- N-body EoS dataset: <u>https://mbiagetti.gitlab.io/cosmos/</u> <u>nbody/eos/</u>
- Compute the α-filtration (and its variants), persistence diagrams for subsampled N-body simulations.
   Process these into topological curves & persistence images.
- Subsampling accounts for observational unknowns and allows for use of templates with large  $f_{NL}$ .

![](_page_51_Figure_5.jpeg)

![](_page_51_Figure_6.jpeg)

#### TDA for LSS: Anomaly Detection with Templates

 Template method: compute templates corresponding to deviations from the fiducial cosmology.

$$\overrightarrow{T^{X}} \equiv \frac{1}{N_{r}} \sum_{i=1}^{N_{r}} \overrightarrow{S}_{NG_{i}}^{X} - \overrightarrow{S}_{G_{i}}^{X}$$
$$D_{\text{template}} = \frac{\left(\overrightarrow{S}_{\text{survey}} \cdot \overrightarrow{T} - \overrightarrow{S}_{\text{mock,avg}} \cdot \overrightarrow{T}\right)}{\sigma}$$

- Compare D<sub>template</sub> to results from fiducial cosmology to account for cosmic variance. Set threshold for detection at e.g. 97.5%.
- Detection rate of 85% for  $f_{NL} = 10$ . Competitive w.r.t. detection rate of method using halo bias (72.7%)

#### [Biagetti, Cole, GS]

![](_page_52_Figure_6.jpeg)

## **TDA for LSS: Degeneracies**

#### [Biagetti, Cole, GS]

- Degeneracy test: to avoid false positives, compute template optimality via "cosine"
- Included in EoS: variation of  $\sigma_{8}$ . Cutoff for template optimality removes false detections of  $f_{\rm NL}^{\rm loc} \neq 0$ .
- Of the 3003 draws from NG10L, 2548 lie beyond the 97.5% confidence level. Of the 2548, all of them are assigned to correct template, none assigned to  $\sigma_8$  template.
- Ongoing: compute degeneracies for wider range of cosmological parameters.

![](_page_53_Figure_6.jpeg)

![](_page_53_Figure_7.jpeg)

## NB: results degrade significantly without subsampling

#### Phase Detection and Classification

![](_page_54_Picture_1.jpeg)

**Alex Cole** 

![](_page_54_Picture_3.jpeg)

**Gregory Loges** 

![](_page_54_Figure_5.jpeg)

 "Quantitative and Interpretable Order Parameters for Phase Transitions from Persistent Homology," A. Cole, G. Loges and GS, [arXiv:2009.14231 [cond-mat.stat-mech]].

### **Phase Transitions**

- Unsupervised (supervised) ML techniques have been used to detect (classify) phases of matter.
- Clustering algorithms, support vector machines and (deep) neural networks have been used to study such critical phenomena.
- **Drawbacks:** lack interpretability and often face difficulties in identifying order parameters.
- [Cole, Loges, GS, 20]: TDA can efficiently distinguish phases and provides interpretable order parameters for phase transitions.
- Encodes multiscale info: captures a system's approach towards scale invariance, enabling quantitative study of critical exponents.

## **Curse of Dimensionality**

 For reasonably large systems, can we simply enumerate all states and go through the exercise of classifying them?

![](_page_56_Figure_2.jpeg)

- Already for 1  $\downarrow$  spins on a 50  $\times$  50 lattice in 2D, there are  $2^{50} \times 2^{50} \sim 10^{752}$  states.
- We want to identify useful patterns by sampling a relatively small fraction of the configuration space.

## Spin Systems and ML

- Success of ML (esp. for images) suggests naive dimensionality doesn't always spell doom.
- Statistical physics presents clean lab for studying ML techniques: we know the Hamiltonian, RG, ...
- Exchange between statistical physics and ML is rich (restricted Boltzmann machine, softmax, ...).
- Phase classification with ML [Carrasquilla, Melko;...].
- For illustrations, we applied TDA to 4 spin systems [Cole, Loges, GS].

![](_page_57_Picture_6.jpeg)

|            | Unfrustrated | Frustrated          |
|------------|--------------|---------------------|
| Discrete   | Ising        | Square-ice          |
| Continuous | XY           | Fully-frustrated XY |

#### Detecting and Characterizing Phases [Cole, Loges, GS, '20]

• From the persistence images of spin systems, we can identify such varied phenomena as magnetization, frustration, (anti)vortices:

![](_page_58_Figure_2.jpeg)

- Simplicity of ML architecture: phase classification and extraction of order parameters achieved via a simple logistic regression.
- Persistent homology compresses the data sets into their most relevant (and interpretable) features, i.e., the grammar of data.

# Ising Model

• A system of **discrete spins** with Hamiltonian:

$$H_{Is} = -\sum_{\langle i,j \rangle} s_i s_j , \quad s_i \in \{-1,1\}$$

- Spontaneous magnetization below  $T_c\approx 2.27,$  breaking  $\mathbb{Z}_2$  symmetry.
- Well-studied (including critical exponents) largely in part to Onsager's exact solution, a good warmup.
- $\alpha$ -filtration on locations of spins align with majority

![](_page_59_Figure_6.jpeg)

![](_page_59_Picture_7.jpeg)

![](_page_59_Picture_8.jpeg)

## Ising Model: Topology

![](_page_60_Figure_1.jpeg)

- For low-temp phase: 1-cycles are very small, either lattice spacingsized or barely bigger (isolated minority spin)
  - For high-temp phase: there is a distribution of births/deaths corresponding to randomly oriented spins.

•

## Ising Model: Phase Classification

- For  $T \in \{1.00, 1.05, \dots 3.50\}$ , generate 1000 configurations.
- Train logistic regression using 25% of simulations at extreme temperatures:

![](_page_61_Figure_3.jpeg)

- The trained logistic regression estimates  $T_c \approx 2.37$  (lattice effects).
- Logistic regression coefficients identify the magnetization (features at lattice scale) as order parameter.

## Ising Model: Critical Exponents

- Persistent homology captures multi-scale behavior near criticality.
- At criticality, the 1-cycle death probability density  $D_T(d)$ :

$$D_T(d) = Ad^{-\mu}e^{-d/a_{death}}, \quad a_{death} \sim |T - T_c|^{-\nu_{death}}, \quad \mu \approx 2, \, \nu_{death} \approx 1$$

![](_page_62_Figure_4.jpeg)

• Using scaling arguments, proportion of clusters of k aligned spins:

 $P(\text{clusters of size } k) \sim k^{-\tau}$ ,  $\tau \approx 2.032$  and  $\xi \sim |T - T_c|^{-1}$ 

## XY Model

• A system of **continuous spins** with Hamiltonian:

$$H_{XY} = -\sum_{\langle i,j \rangle} \cos(\theta_i - \theta_j)$$

- Kosterlitz-Thouless transition at  $T_{XY} = 0.892$
- At low temp, there are bound **vortex-antivortex** pairs, while at high temp, free  $y_{\text{ortices}}$  proliferate and spins farter and pmly oriented.  $\langle f \rangle = 0$
- Sublevel filtration: given Morse function  $f: \Lambda \to S^1$ , consider the sublevel sets  $f^{-1}[-\pi, \nu]$ :

![](_page_63_Figure_6.jpeg)

![](_page_63_Picture_7.jpeg)

![](_page_63_Figure_8.jpeg)

T=1.50

## XY Model: Topology

![](_page_64_Figure_1.jpeg)

- In low-temp phase, vortex-antivortex pairs manifest as early-born 0cycles and late-born 1-cycles.
- Lattice periodicity: two immortal 1-cycles and one immortal 0-cycle.
- High-temp phase has more critical points, worth keeping in mind when comparing normalized persistence images.

#### **XY Model: Phase Classification**

![](_page_65_Figure_1.jpeg)

- Critical temperature estimated as  $T_{XY} \sim 0.9$
- Logistic regression coefficients: in low-temperature phase, lots of probability density assigned to vortex features, which fall in specific regions of persistence images.

![](_page_66_Figure_0.jpeg)

![](_page_66_Picture_1.jpeg)

**Andreas Schachner** 

- "Topological Data Analysis for the String Landscape," A. Cole and GS, JHEP 03 (2019), 054 [arXiv:1812.06960 [hep-th]].
- "Searching the Landscape of Flux Vacua with Genetic Algorithms," A. Cole, A. Schachner and GS, JHEP 11 (2019), 045 [arXiv:1907.10072 [hep-th]].

## Summary

- Persistent homology compresses data in an info rich way: interpretable observables amendable to statistical analysis.
- TDA for LSS [Biagetti, Cole, GS]:
  - Competitive constraints on local NG at smaller scales; can break degeneracies with other cosmological parameters.
  - Reproduce previous results on void size function and suggests new observables: filament loops formed by dark matter halos.
  - Future: other NG shapes and other cosmological parameters.
- TDA for Phase Transitions [Cole, Loges, GS]:
  - Quantitative & interpretable order parameters and critical exponents obtained by a simple logistic regression. Method can be improved further with advanced ML architecture.
  - Future: explore in light of persistent homology how ML pipelines utilize multiscale info to form internal representations of datasets.