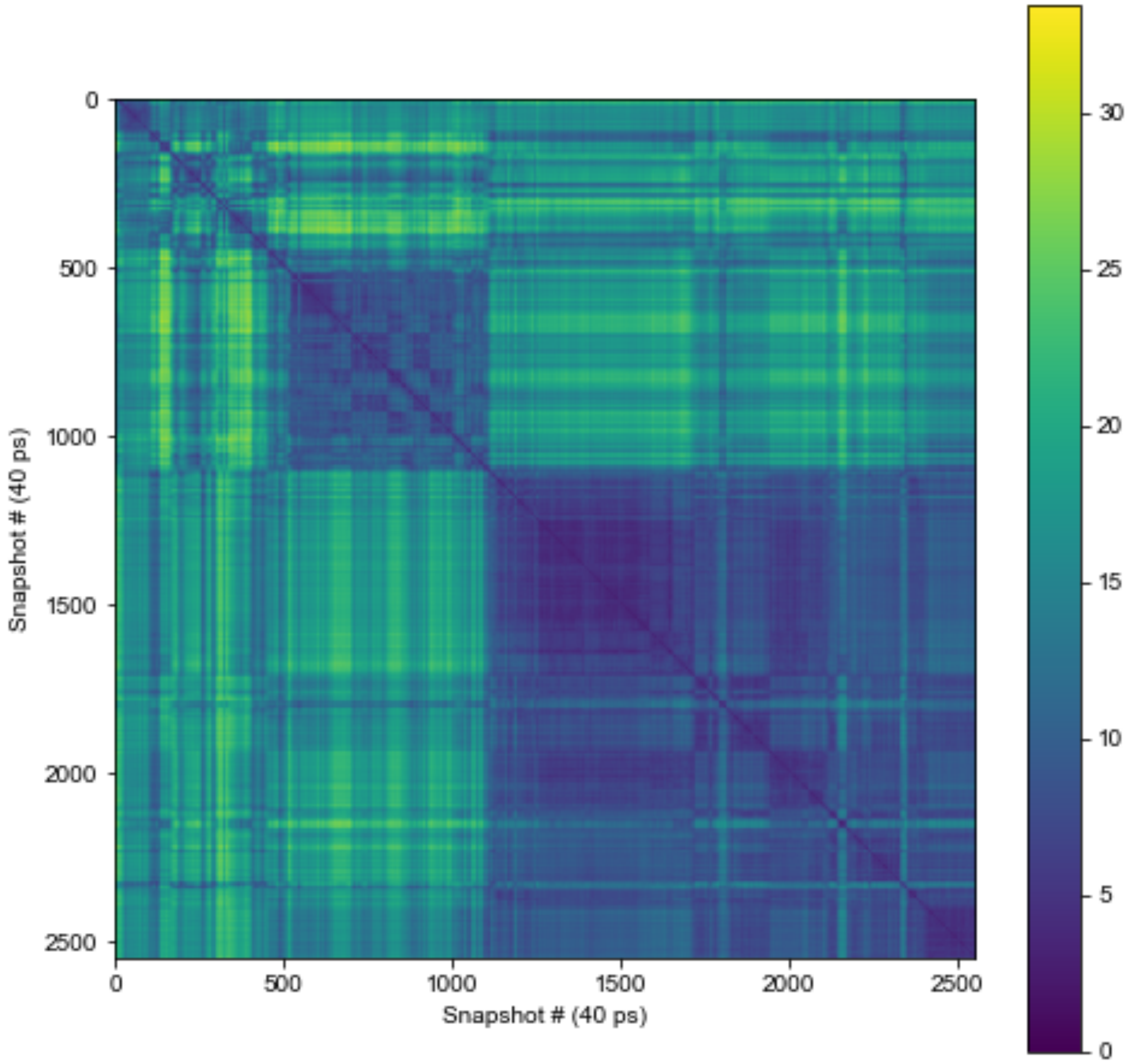
Clustering

- MD simulations yield configurations in continuous space
- Clustering methods group together similar configurations (or, in a more general data science context, observations)
- Clustering is useful
  - interpreting simulation results
  - calculating thermodynamic and kinetic properties
    - predicted populations of conformations
    - predicted rates of transitions (e.g. Markov state models [1, 2])
  - selecting representative configurations for molecular docking [3]

Distance matrices in clustering

- Almost all clustering algorithms employ a distance matrix
- In a matrix **D**, D_{kl} denotes the distance between observation k and l
- Distance matrices include [3]
  - the RMSD
    - between alpha carbons/all heavy atoms
    - in a entire protein/in a region of the protein
  - based on occupancy fingerprints
    - a 3D grid with zero or one depending whether a point is close to an atom
    - If M_{ab} is the number of points where one grid has *a* and the second *b*,
      - the overlap is $M_{10} + M_{01}$
      - the Tanimoto similarity is $-\log_2[M_{11}/(M_{11} + M_{10} + M_{01})]$
      - the Jaccard distance is $[(M_{11} + M_{01})/(M_{11} + M_{10} + M_{01})]$
  - Euclidean distance between principal components (like the RMSD, PCA can be based on different subsets of coordinates)



Heat map of Euclidean distances between top 20 principal components in a simulation of ubiquitin