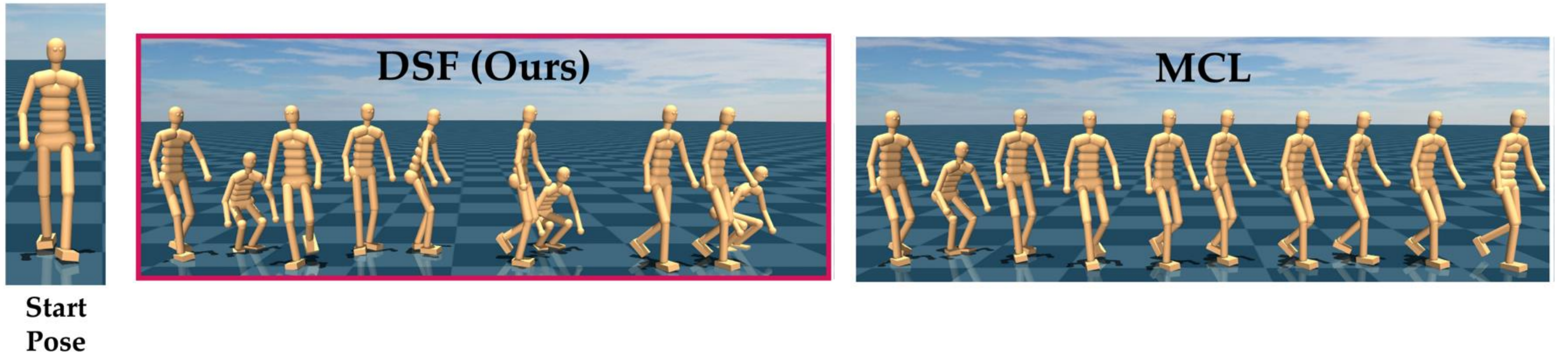
k-DPPs: Fixed Size DPPs for Diversity-Based Subsampling

Joanna Zou

18.338 Final Project, Fall 2024

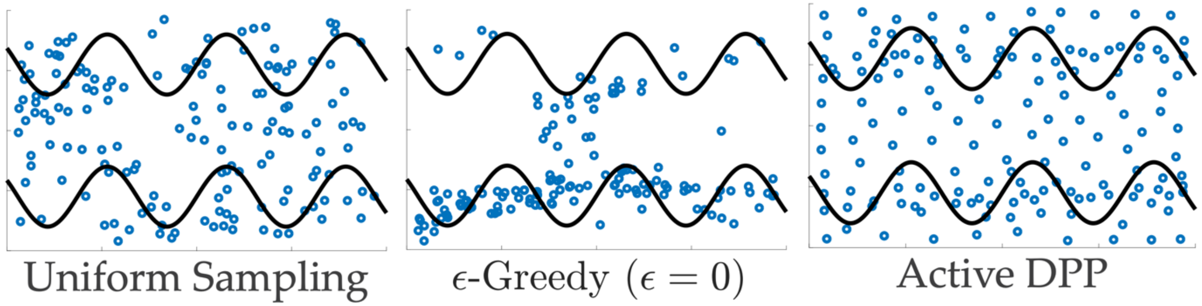
Diversity-Based Subsampling

Reduce redundancy in image search engine
(Kulesza & Taskar 2011)



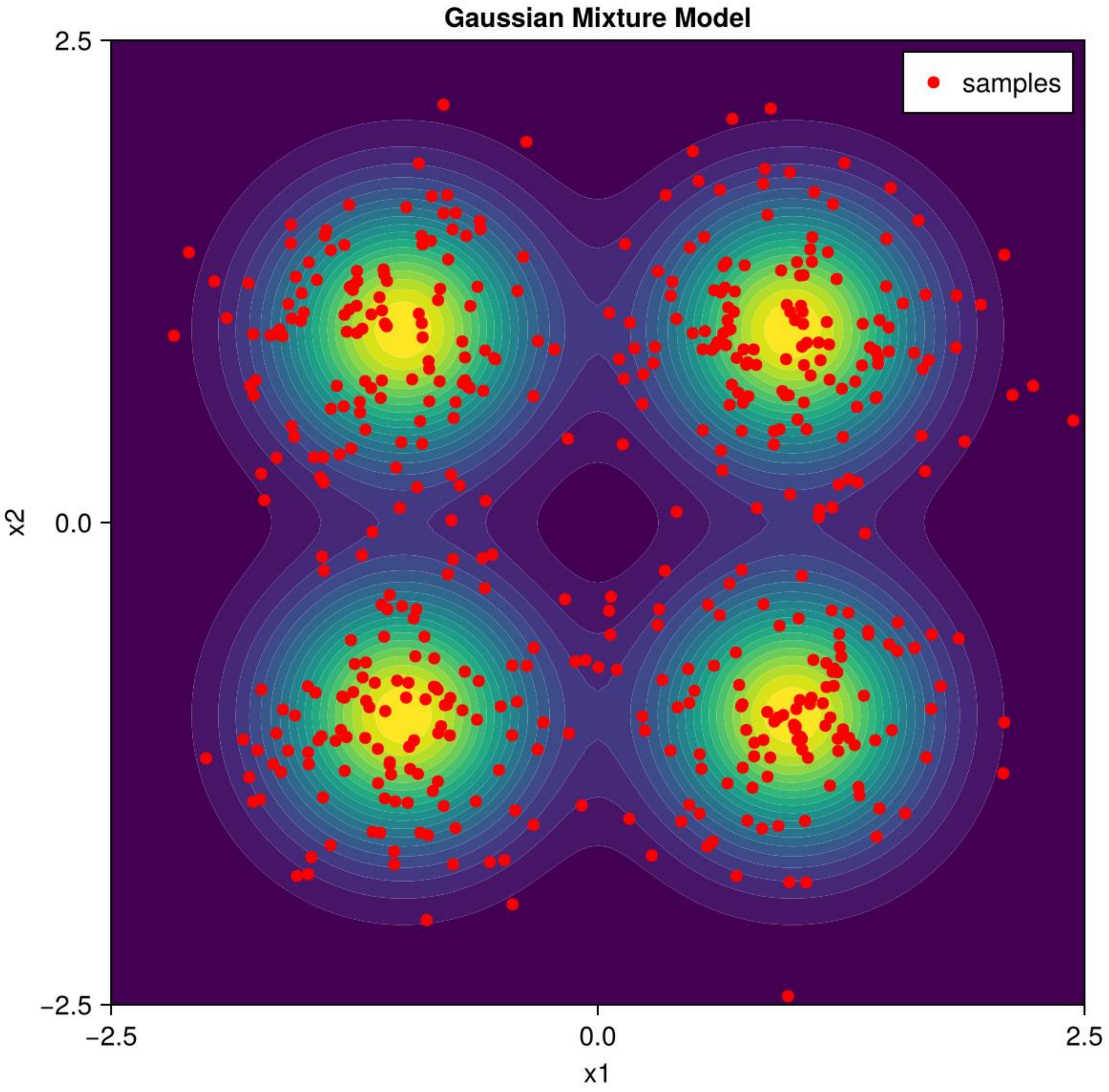
Forecast diverse trajectories
(Yuan & Kitani 2019)

Curate informative datasets for model training
(Biyik et al. 2019)



Toy Example

Given a clustered dataset, how to sample a diverse subset of a fixed size?



	DPPs	k-DPPs
Probability mass function:	$\mathcal{P}(Y = A) = \frac{\det(L_A)}{\sum_{A' \subseteq \mathcal{Y}} \det(L_{A'})}$ $= \frac{1}{\det(I+L)} \sum_{J \subseteq 1:N} \mathcal{P}^{V_J} \prod_{i \in J} \lambda_i$	$\mathcal{P}(Y = A \mid \lvert Y \rvert = k) = \frac{\det(L_A)}{\sum_{\lvert A' \rvert = k} \det(L_{A'})}$ $= \frac{1}{e_k^N} \sum_{\lvert J \rvert = k} \mathcal{P}^{V_J} \prod_{i \in J} \lambda_i$
Normalizing constant:	$\sum_{A' \subseteq \mathcal{Y}} \det(L_{A'}) = \det(I+L)$ $= \prod_{i=1}^{N} (\lambda_i + 1)$	$\sum_{\lvert A' \rvert = k} \det(L_{A'}) = \det(I+L) \sum_{\lvert A' \rvert = k} \mathcal{P}(Y = A')$ $= \sum_{\lvert J \rvert = k} \prod_{i \in J} \lambda_i = e_k^N$
Marginal probability:	$\mathcal{P}(i \in Y) = K_{ii}$	$\mathcal{P}(i \in Y \mid \lvert Y \rvert = k) = \lambda_N \frac{e_{k-1}^{N-1}}{e_k^N}$

Normalization of k-DPPs

Computing elementary symmetric polynomial is a combinatorial problem:

$$e_k^N = e_k(\lambda_1, \ldots, \lambda_N) = \sum_{\substack{J \subseteq 1:N \\ |J|=k}} \prod_{i \in J} \lambda_i \qquad \mathcal{O}\left(k\binom{N}{k}\right)$$

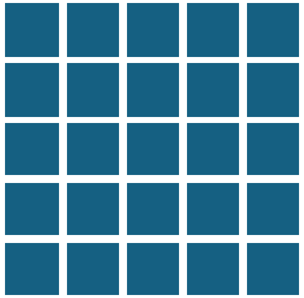
Summation algorithm using recurrence relation:

$$e_k^N = e_k^{N-1} + \lambda_N e_{k-1}^{N-1} \qquad \mathcal{O}(Nk)$$

Sampling algorithm

1 Compute kernel matrix

$$K_{ij} = \kappa(x_i, x_j)$$

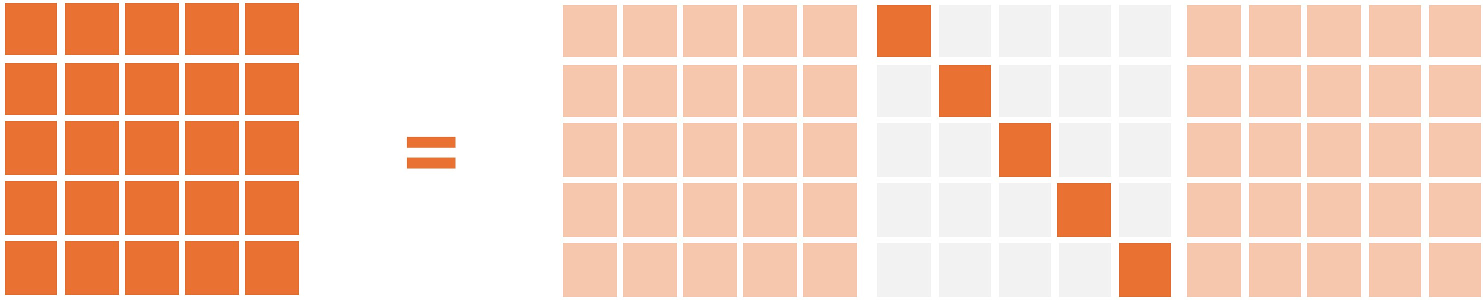


2 Compute L-ensemble

$$L = K(I - K)^{-1}$$

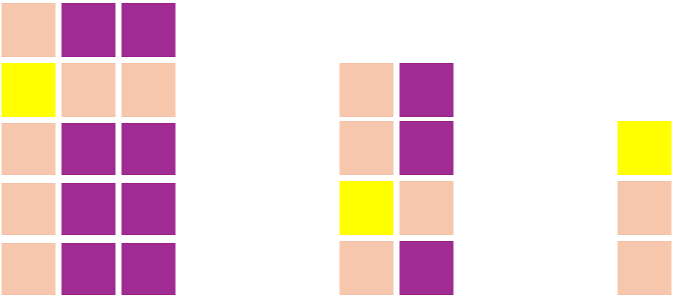
3 Compute eigendecomposition of L

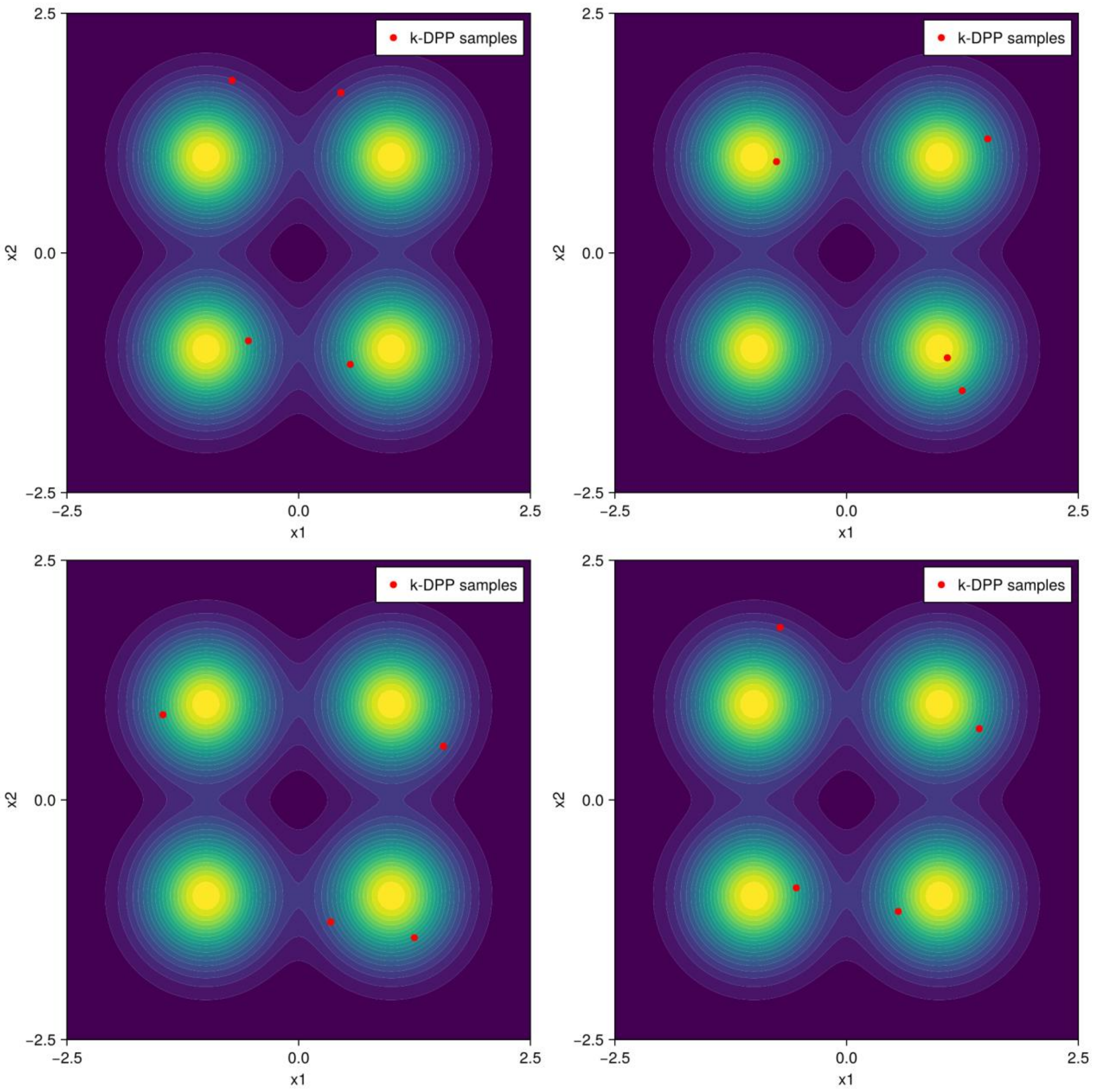
$$L = V \Lambda V^{\mathrm{T}}$$



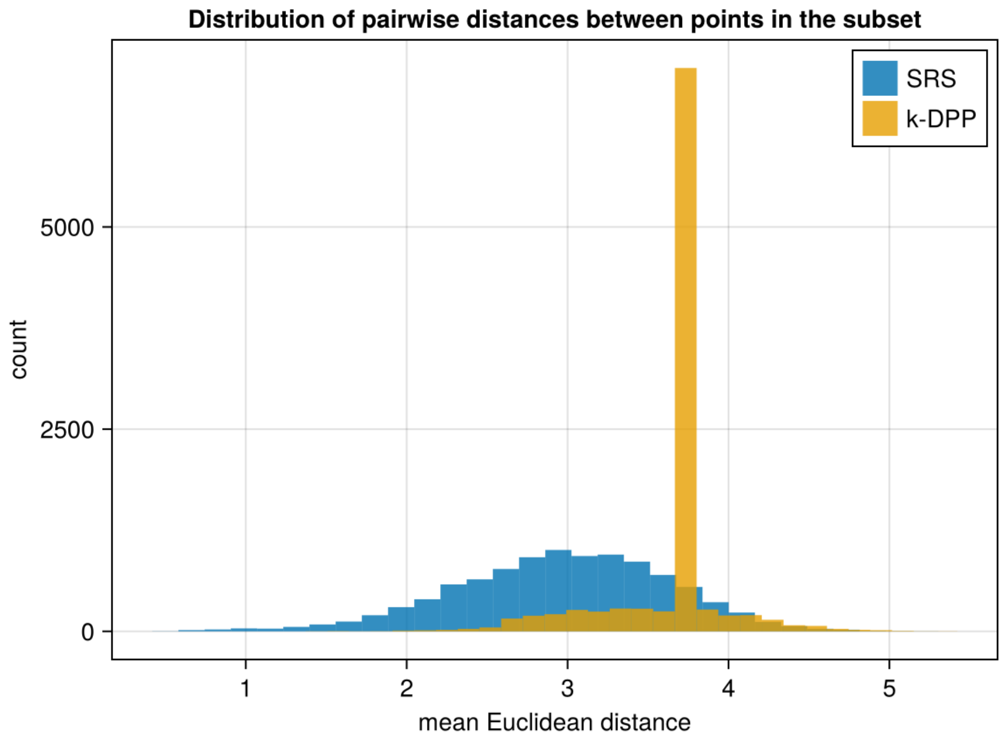
4 Subselect eigenvectors with prob. $\lambda_N \frac{e_{k-1}^{N-1}}{e_k^N}$

5 Sample indices from orthonormalized subspace of V.





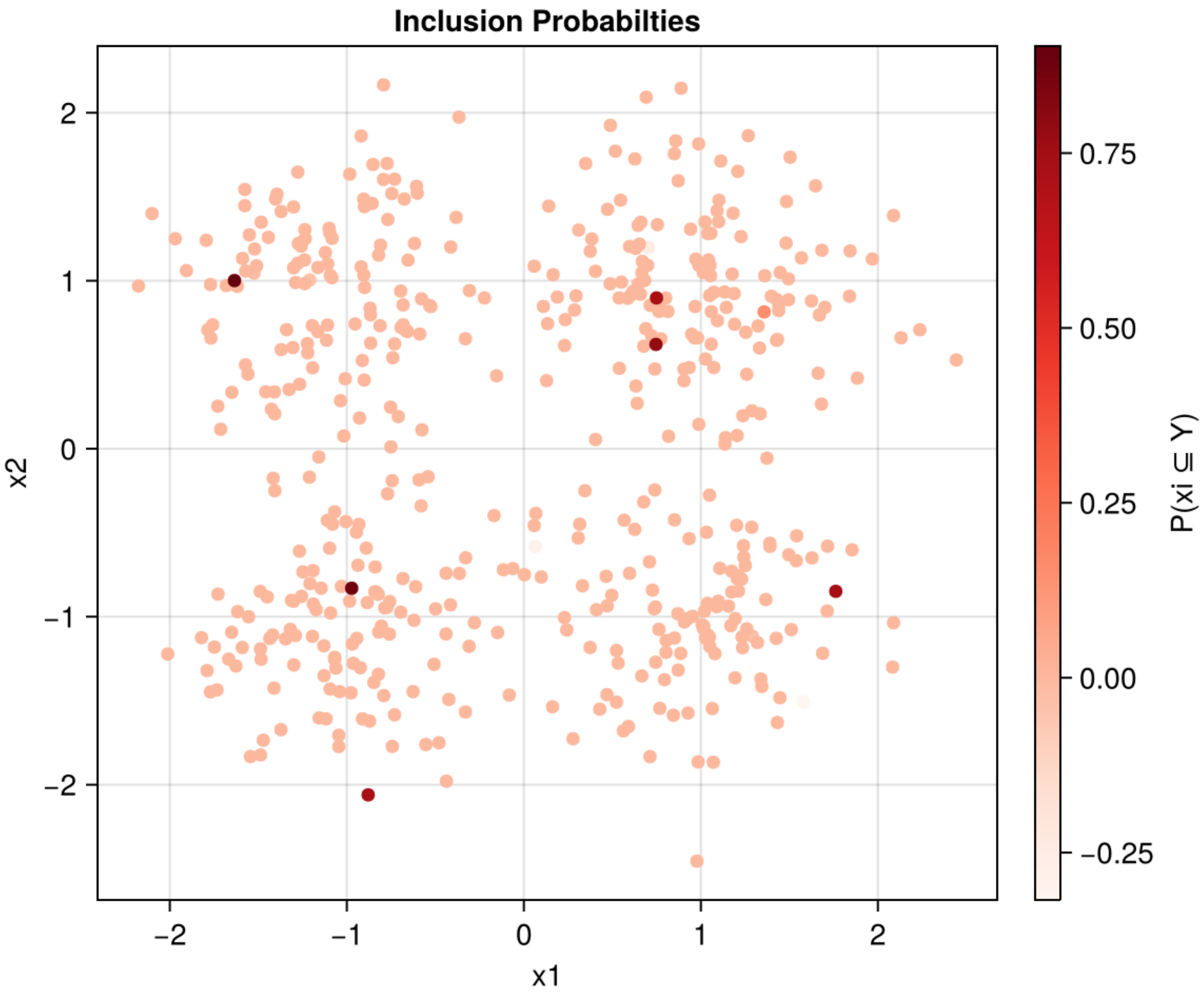
k-DPP samples more diverse subsets compared to simple random sampling.



Probability of element being included in the k-DPP set:

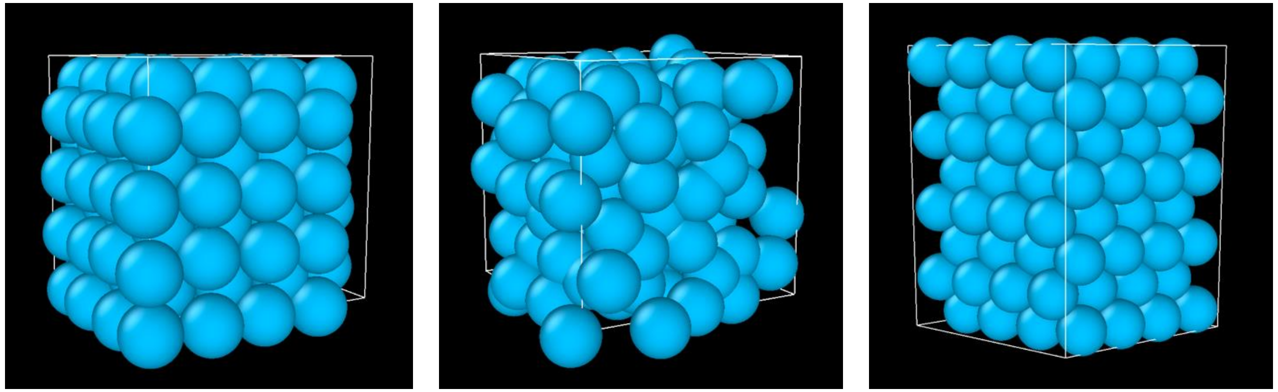
“DPP mode”

$$\mathcal{P}(i \in \color{blue}{Y} \mid |\color{blue}{Y}| = k) = \lambda_N \frac{e_{k-1}^{N-1}}{e_k^N}$$



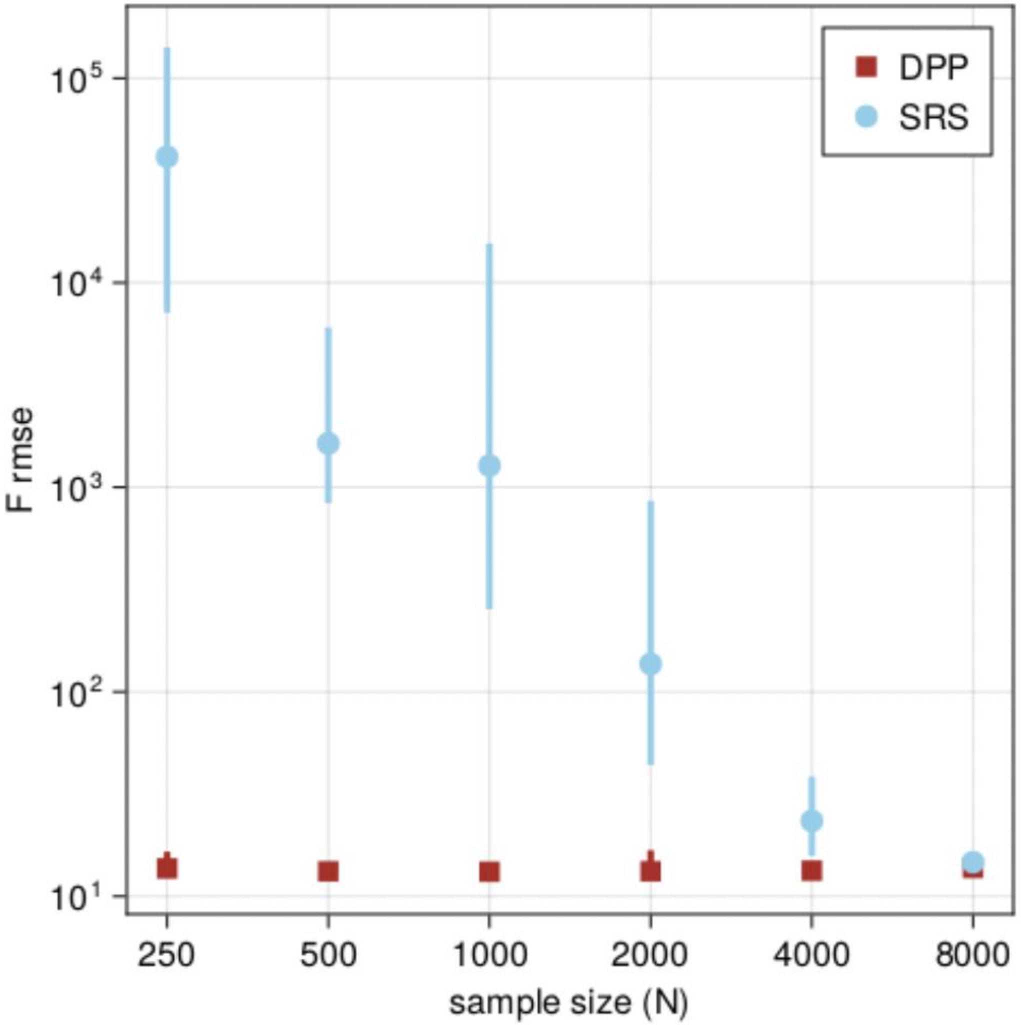
Application:

Curating training dataset for machine learning force field for Hafnium



Starting configurations of Hf atoms composing the training set

Test error for varying size training sets, chosen by SRS and k-DPP



References

[1] A. Kulesza, B. Taskar (2011). "k-DPPs: Fixed-Sized Determinantal Point Processes". *ICML 2011*.

[2] A. Kulesza, B. Taskar (2012). "Determinantal point processes for machine learning." *Foundations and Trends in Machine Learning* 5 (2-3), pp. 123-286.

[3] S. Barthelme, P. Amblard, N. Tremblay (2018). "Asymptotic equivalence of fixed-size and varying-size determinantal point processes." *Bernoulli* 25 (4B).

[4] A. Edelman. "Random Matrix Theory". *Work-in-progress*.

[5] Y. Yuan, K. Kitani (2019). "Diverse trajectory forecasting with determinantal point processes." *ICLR 2020*.

[6] E. Biyik, K. Wang, N. Anari, D. Sadigh (2019). "Batch active learning using determinantal point processes." *Preprint*.