Towards Molecular-level Similarity Search based on Text Data

Presenter: Ang Yihao, Miao Yisong

NUS
National University of Singapore

Outline

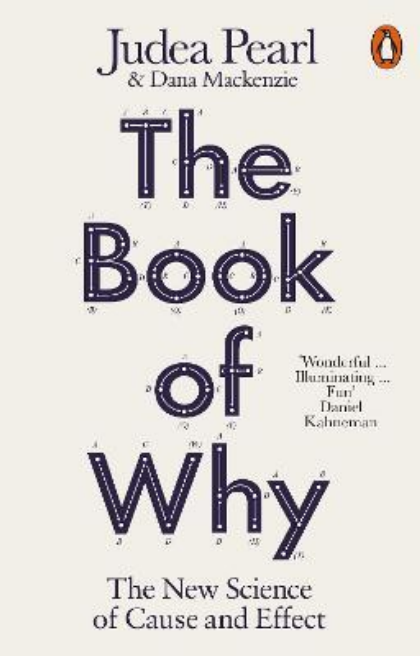
I. Motivation & Background

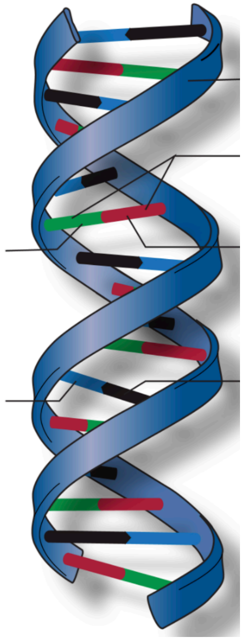
Motivation

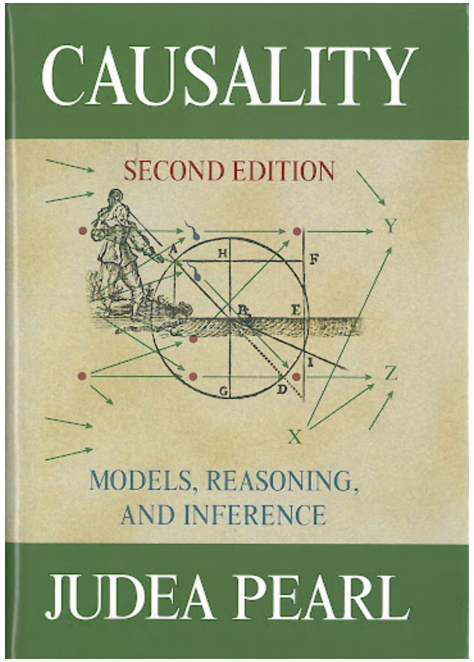
- Store text into DNA is promising.
- Information explosion.
- DNA is a promising medium for data storage.

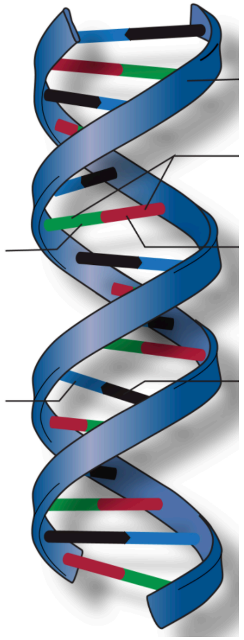
Motivation

- What if we want to retrieve two similar “books” from the “library”?

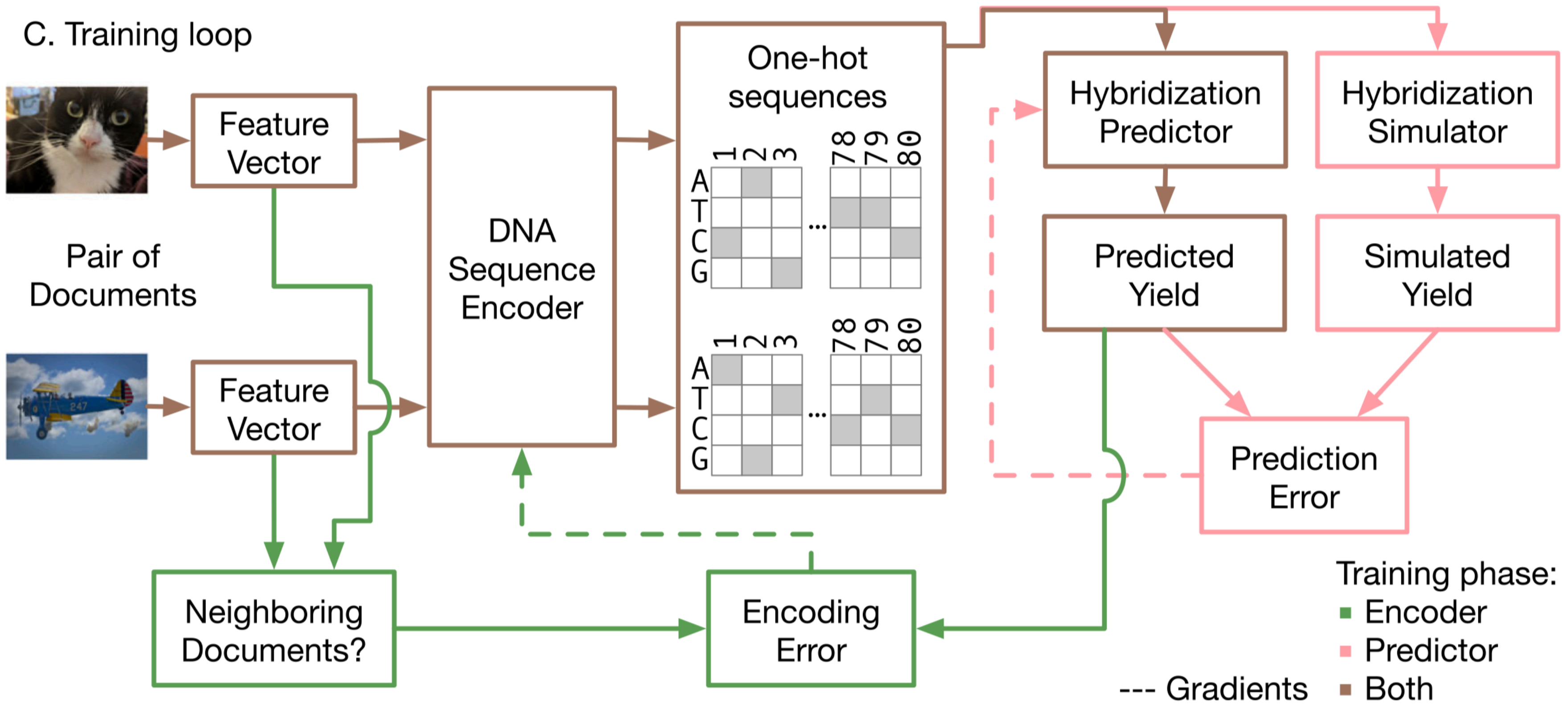








Background – Success in Image Similarity Search



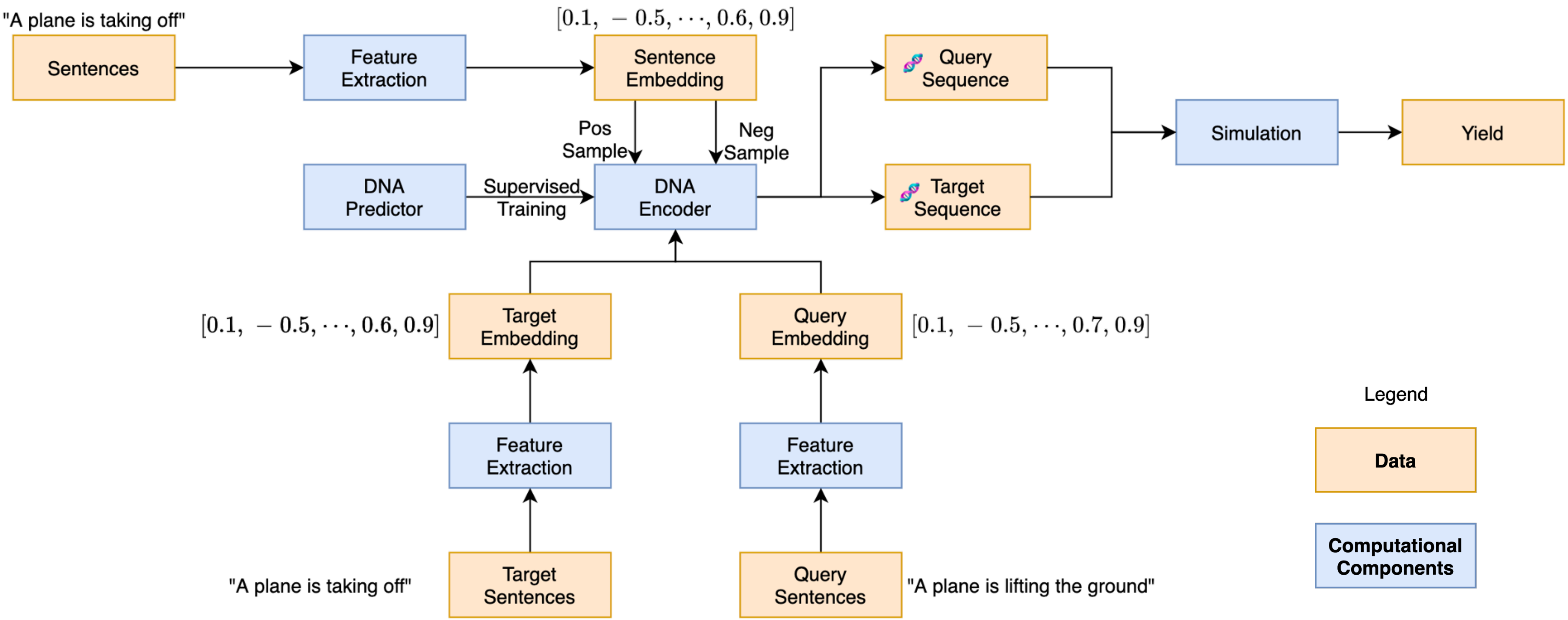
- The workflow of comparing the images of *cat* and *plane*.
- Encoder: image embedding → DNA sequence.
- Predictor: Imitate the behavior of NUPACK on DNA sequence hybridization.

II. Workflow

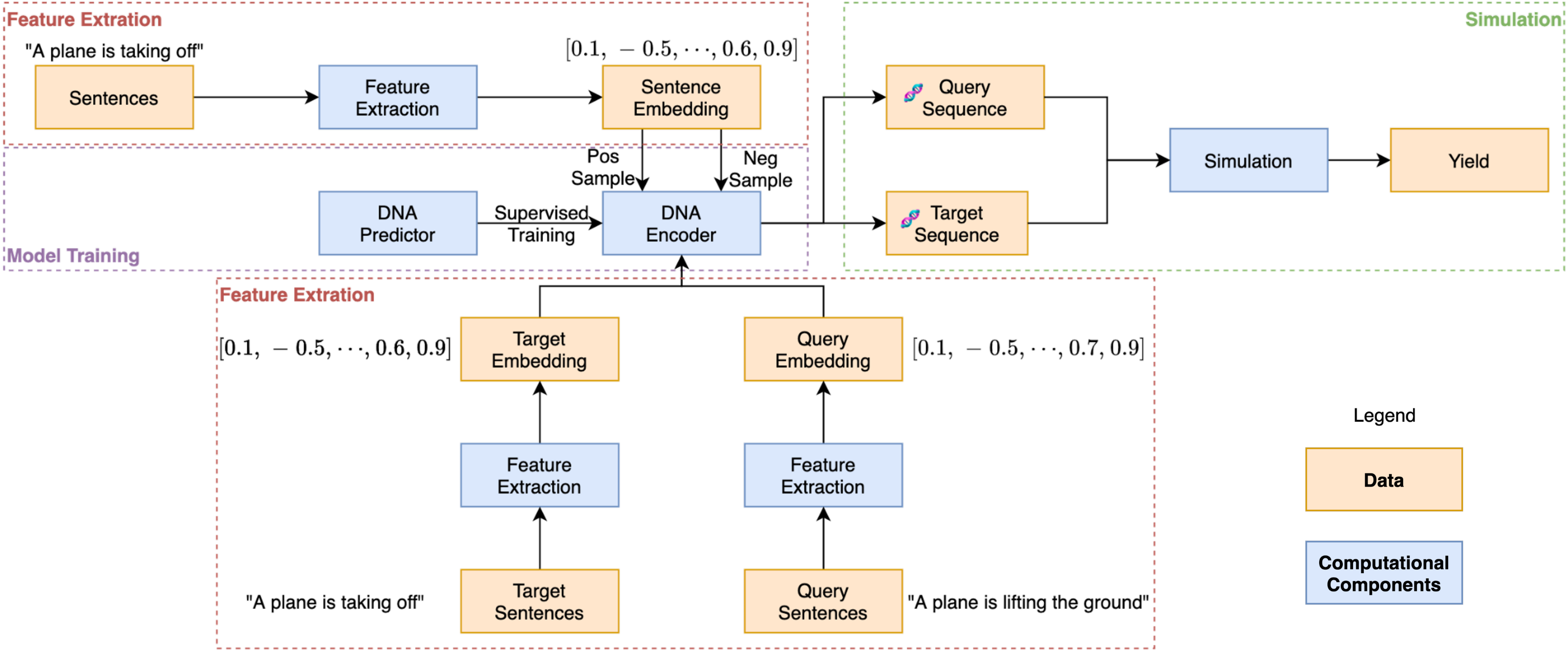
Workflow – Overview

- The workflow of embedding text into DNA data storage system and support similarity search.
  - Feature extraction
  - Training of encoder and predictor
  - Feature extraction for query and target
  - Simulation

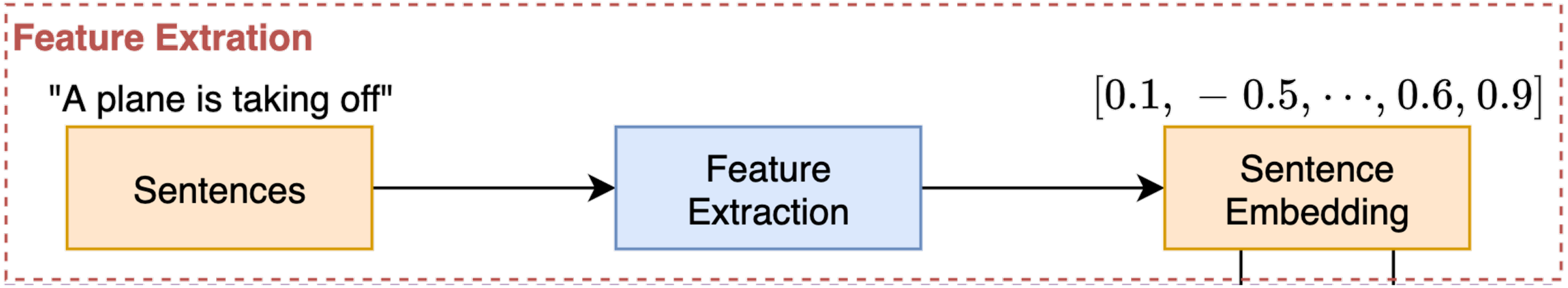
Workflow – Overview



Workflow – Overview

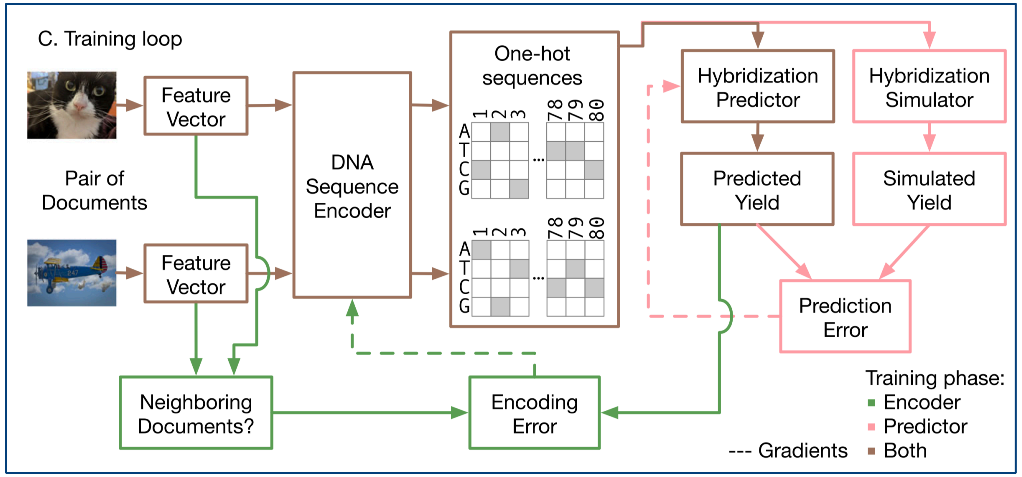


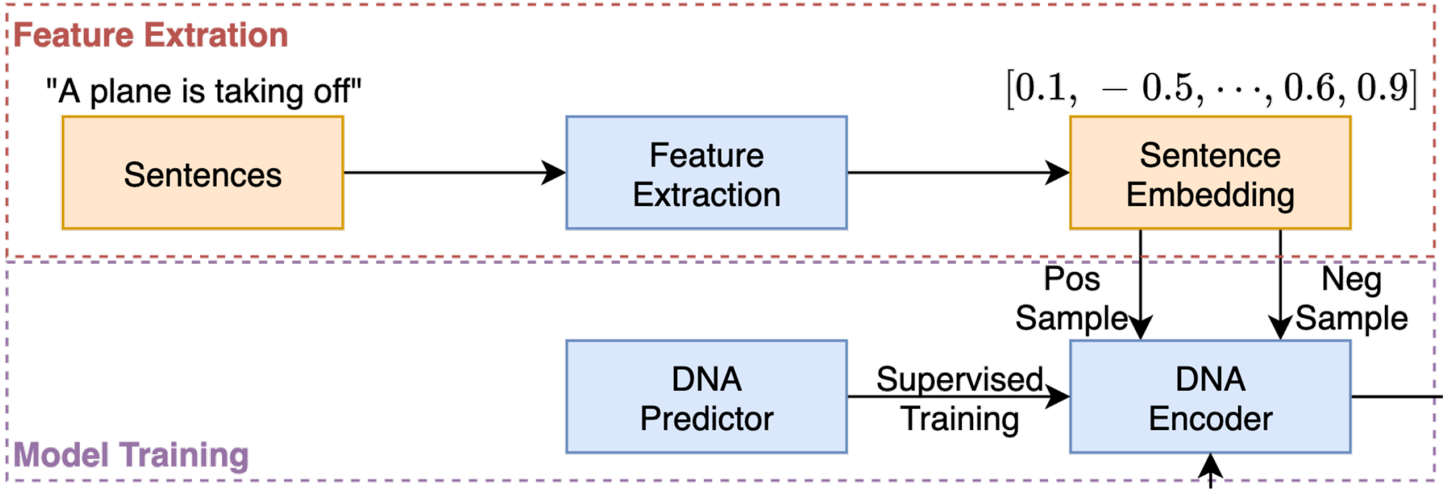
Workflow – Feature Extraction



- We choose to use Sentence-BERT model.
- This is because Sentence-BERT has been reported good performance on text similarity matching tasks.
  - Input: Sentences.
  - Output: Sentence embedding. (i.e., feature vectors)

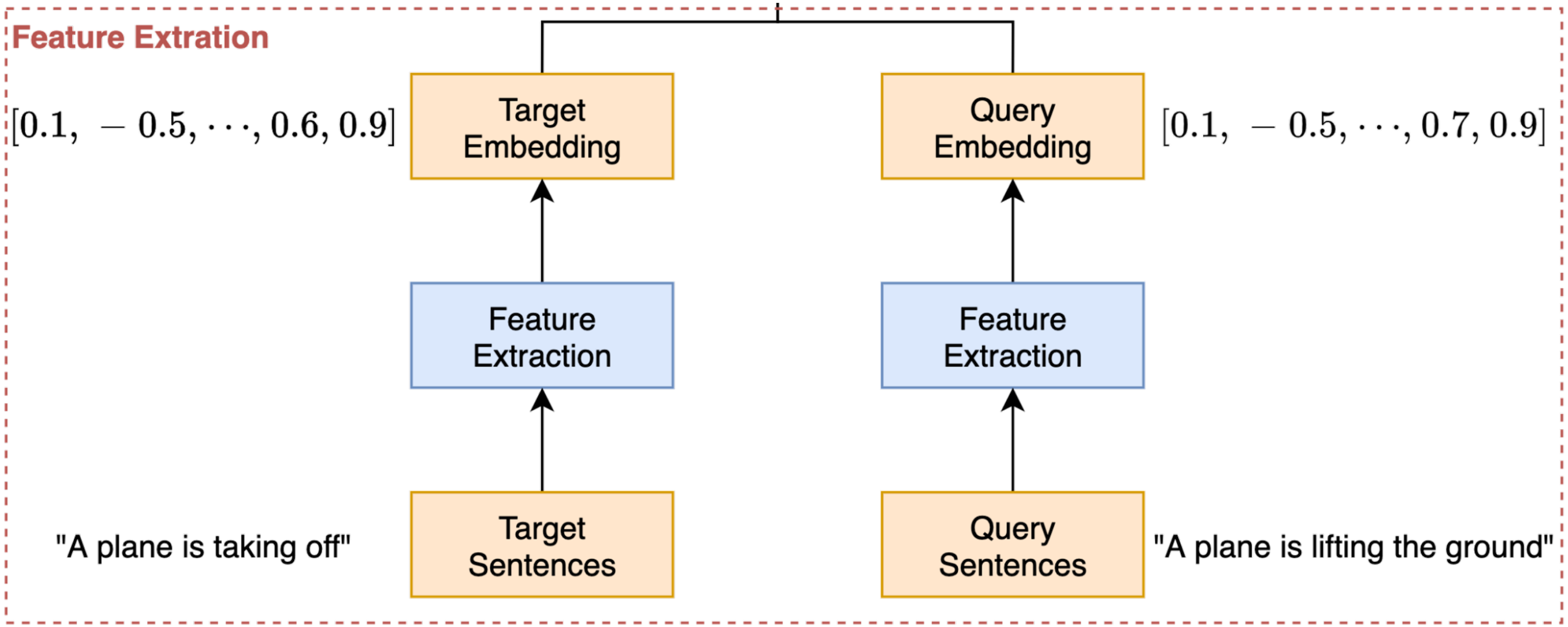
Workflow – Training of Encoder and Predictor





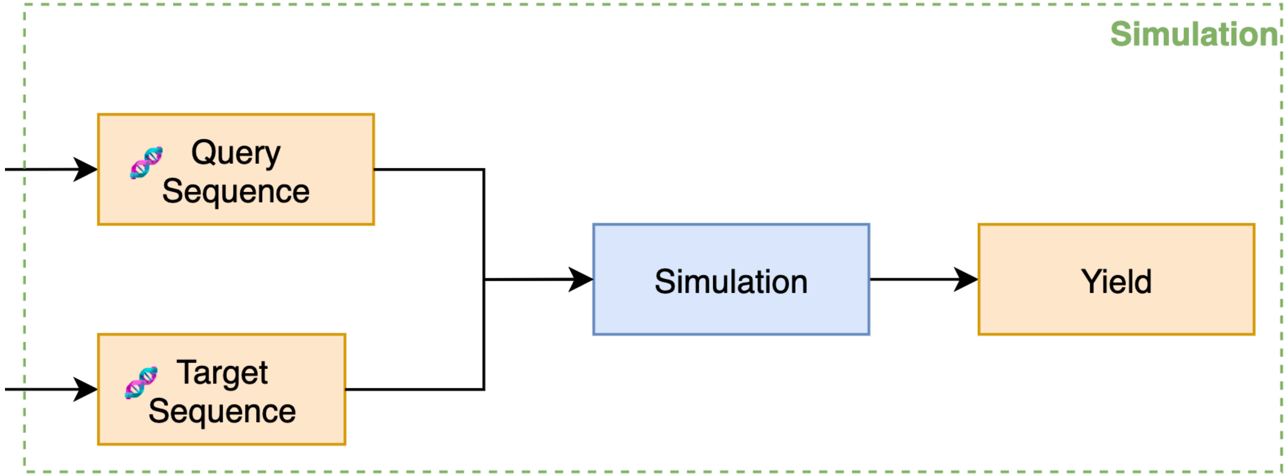
- Training Objective: Encode similar sentences into similar DNA sequences so that they have a high hybridization yield.
- Encoder:
  - **Input:** Positive pairs of sentences embeddings and Negative pairs. **Output:** DNA sequences.
  - **Supervision:** Predictor's judge on whether the two DNA sequences are going to hybridize.
- Predictor:
  - **Input:** Pairs of DNA sequences. **Output:** Hybridization yield.
  - **Supervision:** NUPACK's simulation as ground truth.

Workflow – Feature Extraction for Query and Target



- The same as processing the training data.
- We use the same feature extraction SBERT.
  - Input: Target/Query Sentence.
  - Output: Sentence embedding. (i.e., feature vectors)

Workflow – Simulation



	target_feat...	query_fea...
sentence1-0	ACGTAAACAC...	ATGGCTAAAC...
sentence1-1	AGGGACACAC...	ATGGCTAAAC...
sentence1-2	TGGCGAGCAC...	ATGGCTAAAC...
sentence1-3	TTGGGCAAAC...	ATGGCTAAAC...
sentence1-4	TATGGCTCCC...	ATGGCTAAAC...
sentence1-5	TAGTGAAAAC...	ATGGCTAAAC...
sentence1-6	TTCTGGAAAC...	ATGGCTAAAC...
sentence1-7	TGGATTACAC...	ATGGCTAAAC...
sentence1-8	TTAGTCACACT...	ATGGCTAAAC...
sentence1-9	TTAGTCACACT...	ATGGCTAAAC...

- Simulation will be using NUPACK (CUPACK interface).

III. Results Analysis

Results Analysis

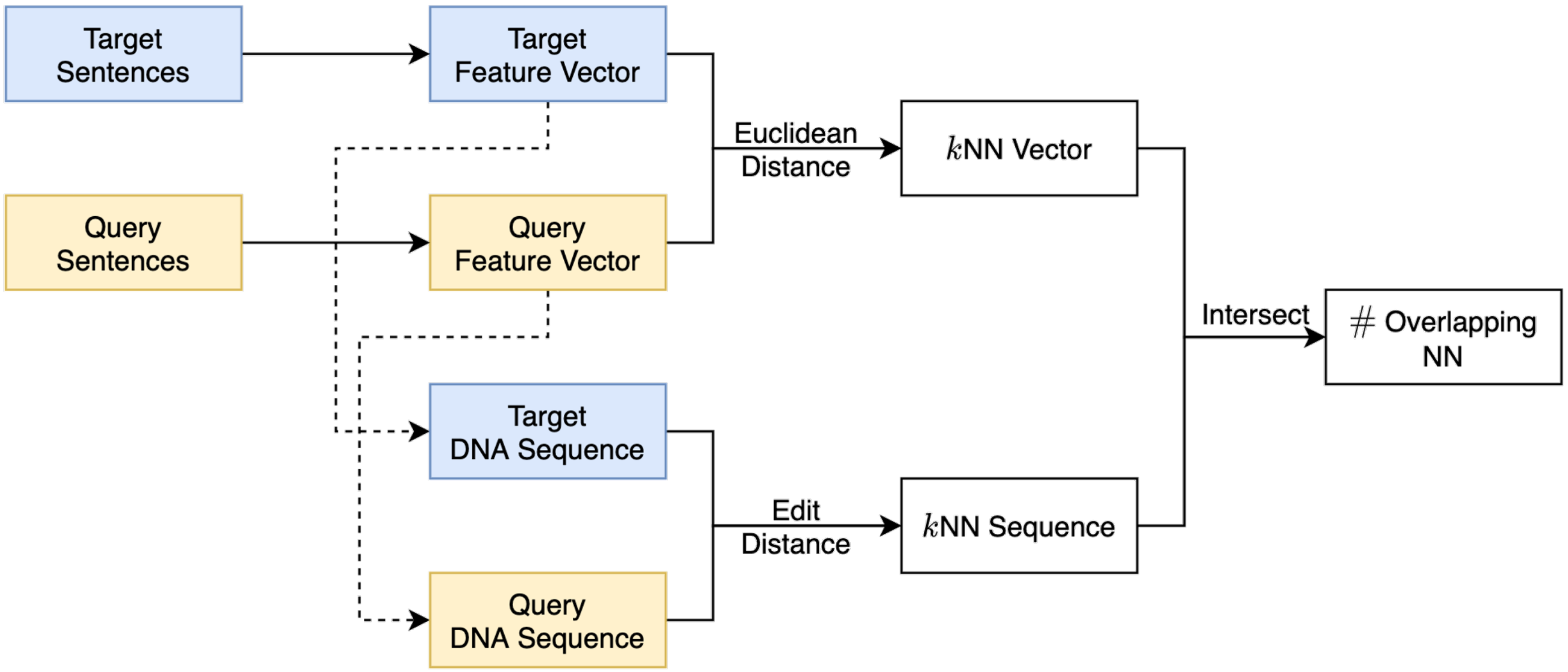
- Evaluation Process
- Experiment Setup
- Evaluation Results
- Case Study
    - Distance Analysis
    - Semantic Similarity
    - Retrieval Quality

Results Analysis

- Observations from NLP
  - Sentences that have close Euclidean distances of feature vectors are semantically similar.
  - Examples
    - The cat is licking a bottle.
    - A cat is licking itself.
    - A cat plays with a small bottle.
- Adopt Euclidean distances of feature vectors as ground truth
- Aim to find pairs of sentences with similar feature vectors and similar DNA sequence

Results Analysis

- Evaluation Process
  - Given k, Overlapping Ratio $= \frac{\#\ k\text{NN Sequence}}{\#\ k\text{NN Vector}}$
  - Higher the overlapping ratio, the more precise the model retrieves

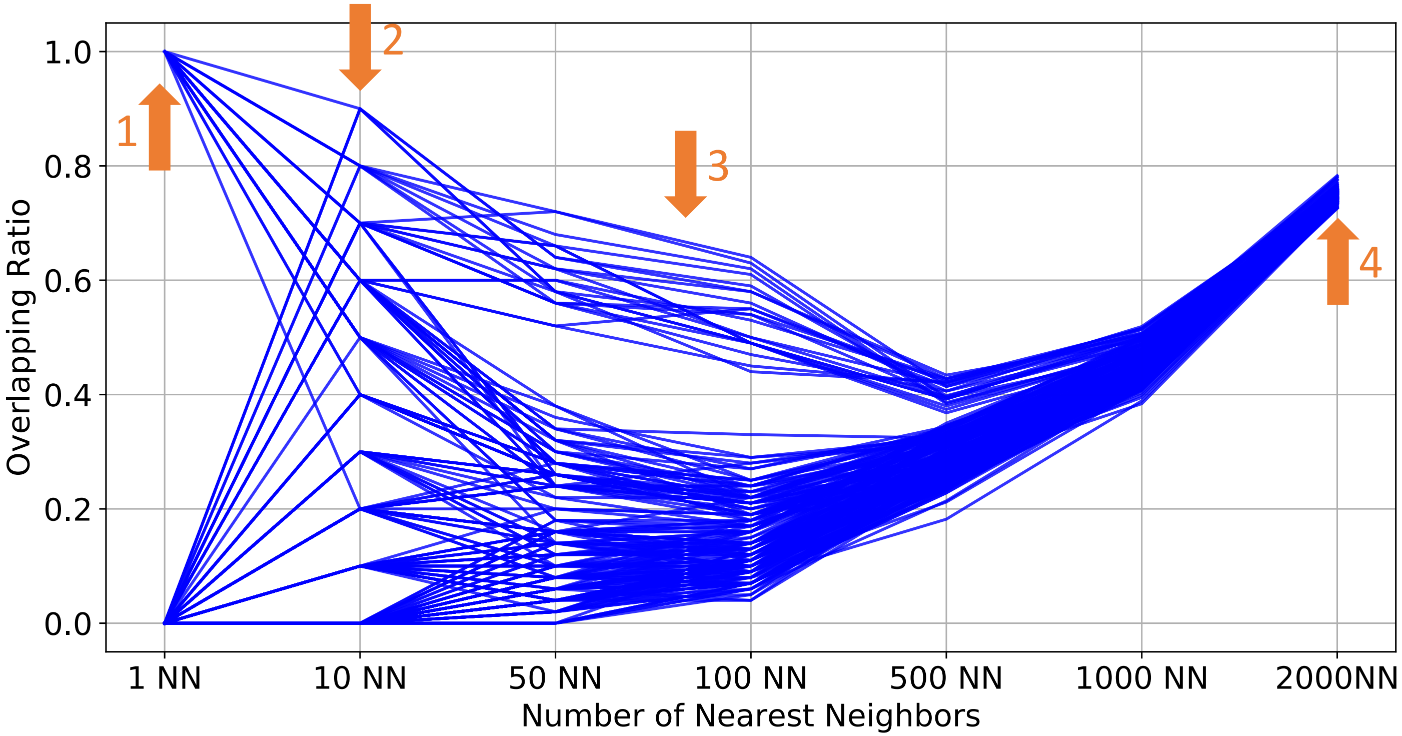


Results Analysis

- Experiment Setup
  - Query: 9 sentences
    - 'A plane is taking off.',
    - 'A woman is peeling a potato.',
    - 'The cat is licking a bottle.',
    - 'Steve Jobs is the CEO of Apple Inc. She hold many dollars of money.',
    - 'Computer science is one of the most revolutionary fields in scientific research.',
    - 'The all-* models where trained on all available training data (more than 1 billion training pairs) and are designed as general purpose models.',
    - 'The church has cracks in the top.',
    - 'The statue is offensive and people are mad that it is on display.',
    - 'A group of people are playing in a symphony.’
  - Target Dataset: STSB (2758 sentences), SNLI (20000 sentences)
  - Feature Extractor: MPNet, MiniLM
  - Encoder: early stop at different checkpoints
  - $k = 1, 10, 50, 100, 500, 1000, 2000$

Results Analysis

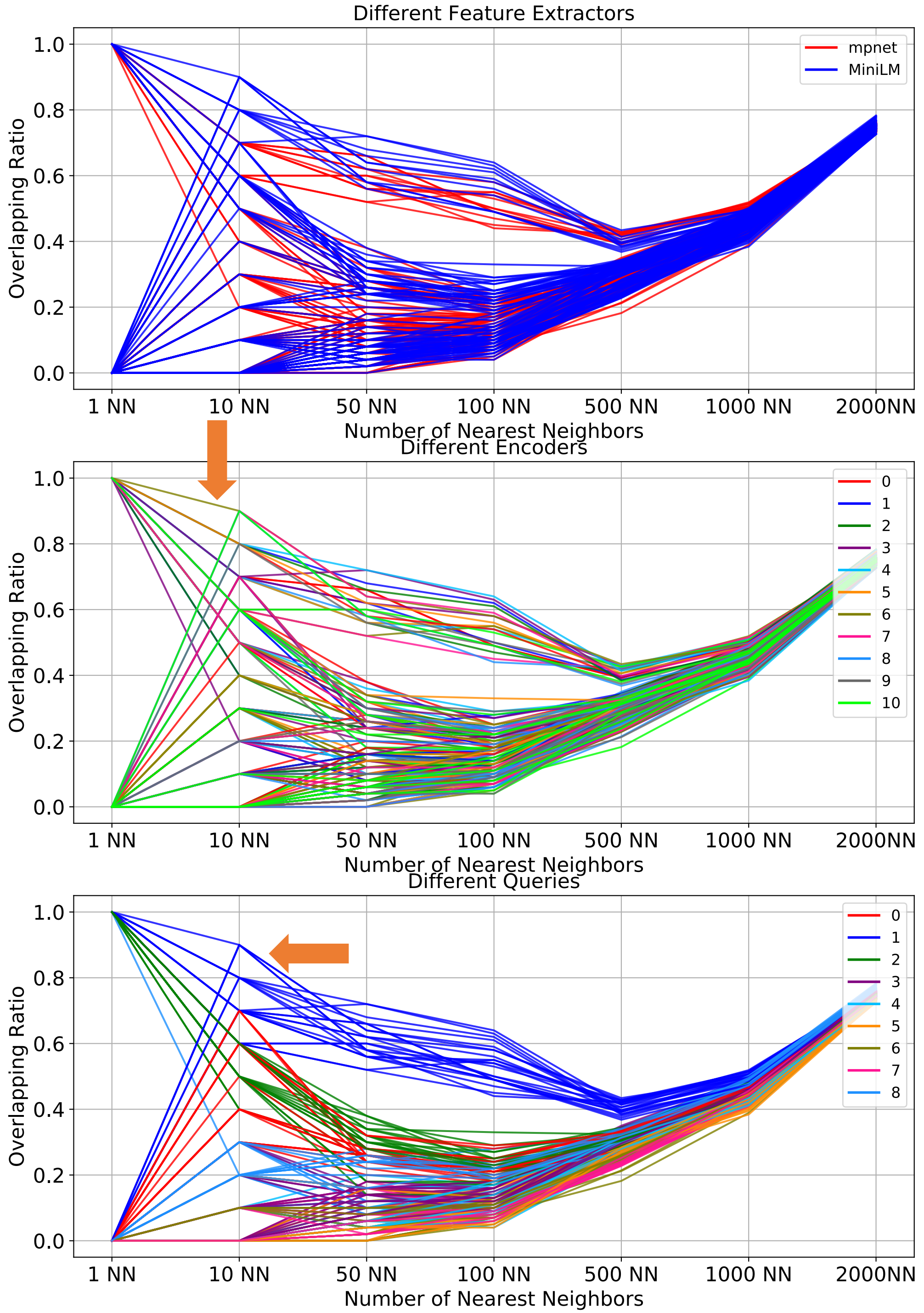
- Evaluation Results
  1. Some parameter combinations can get accurate 1 NN
  2. Some parameter combinations can get over 80% accurate 10 NN
  3. Although the initial NNs are accurate, the ratio drops as k increases
  4. As k continues to increase, the ratio will converge at around 75%



One line:
one set of parameter combination

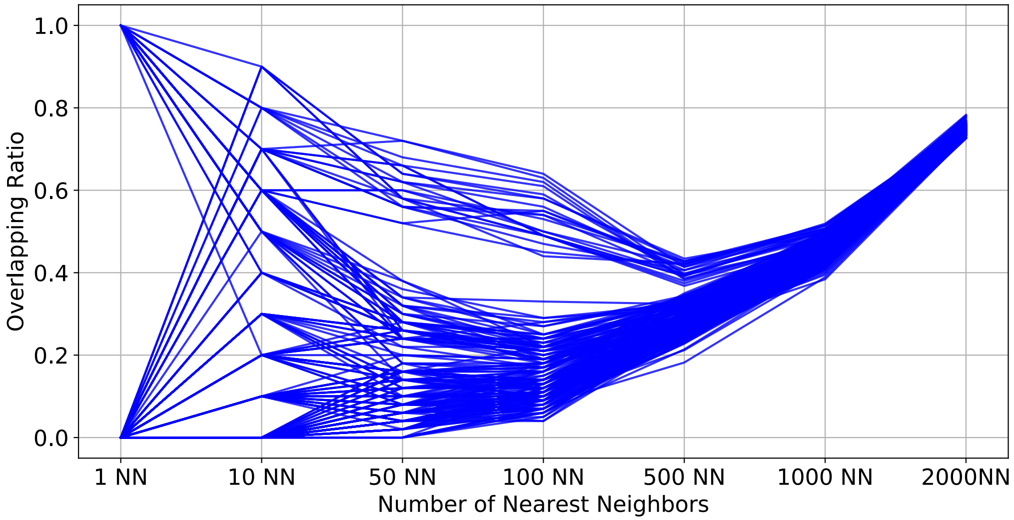
Results Analysis

- Evaluation Results
  - Effect of each kind of parameters (colors)
    - Feature Extractor: MPNet, MiniLM
    - Encoder: early stop at different checkpoints
      - Can select some models with better performance
    - Query ID
      - Queries that are related to this dataset perform better



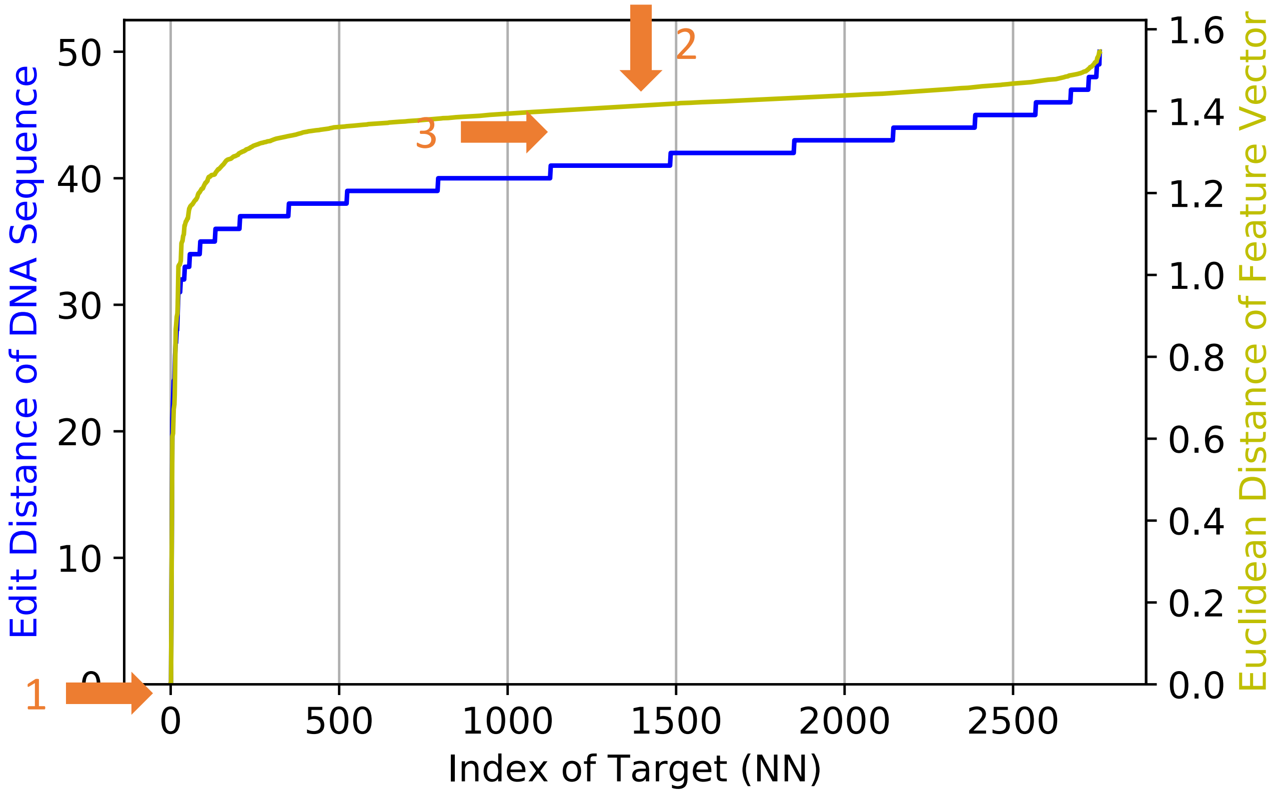
Results Analysis

- Evaluation Results
  - How to select a better set of parameters?
  - For all parameter combinations, given each k value, get the mean and median of the ratio, apply the following rules:
    1. 1 NN should be accurate
    2. 10 NN should be high
    3. 50—100 NN should be relatively high



Results Analysis

- Case Study 1
  - Query: A woman is peeling a potato.
  - Dataset: STSB
  - Feature Extractor: MPNet
  - Two Distance @ kNN
    1. Have accurate 1 NN
    2. A plateau in the middle
    3. Close match of two curves



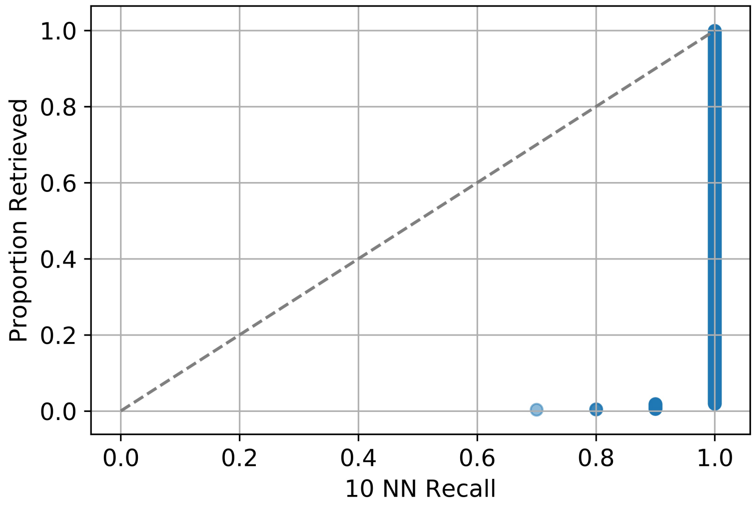
Results Analysis

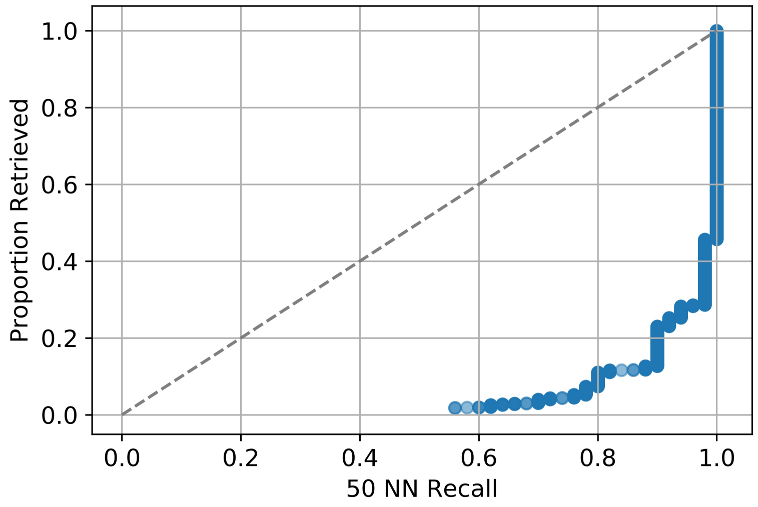
- Case Study 1
  - Query: A woman is peeling a potato.
  - Dataset: STSB
  - Feature Extractor: MPNet
  - Semantic Similarity

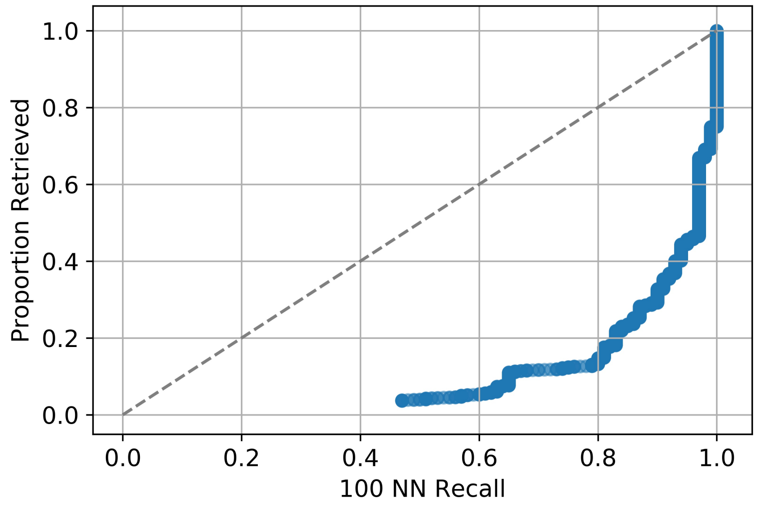
Feature Vector NN	DNA Sequence NN
A woman is peeling potato.	A woman is peeling potato.
A person is peeling a potato.	A person is peeling a potato.
A man is peeling a potato.	A woman is cutting potatoes.
The lady peeled the potato.	A woman is chopping a peeled potato into slices.
…	…
A person is peeling a potato with a potato peeler.	A man is peeling a potato.

Results Analysis

- Case Study 1
  - Query: A woman is peeling a potato.
  - Dataset: STSB
  - Feature Extractor: MPNet
  - Retrieval quality @ 10NN, 50NN, 100NN

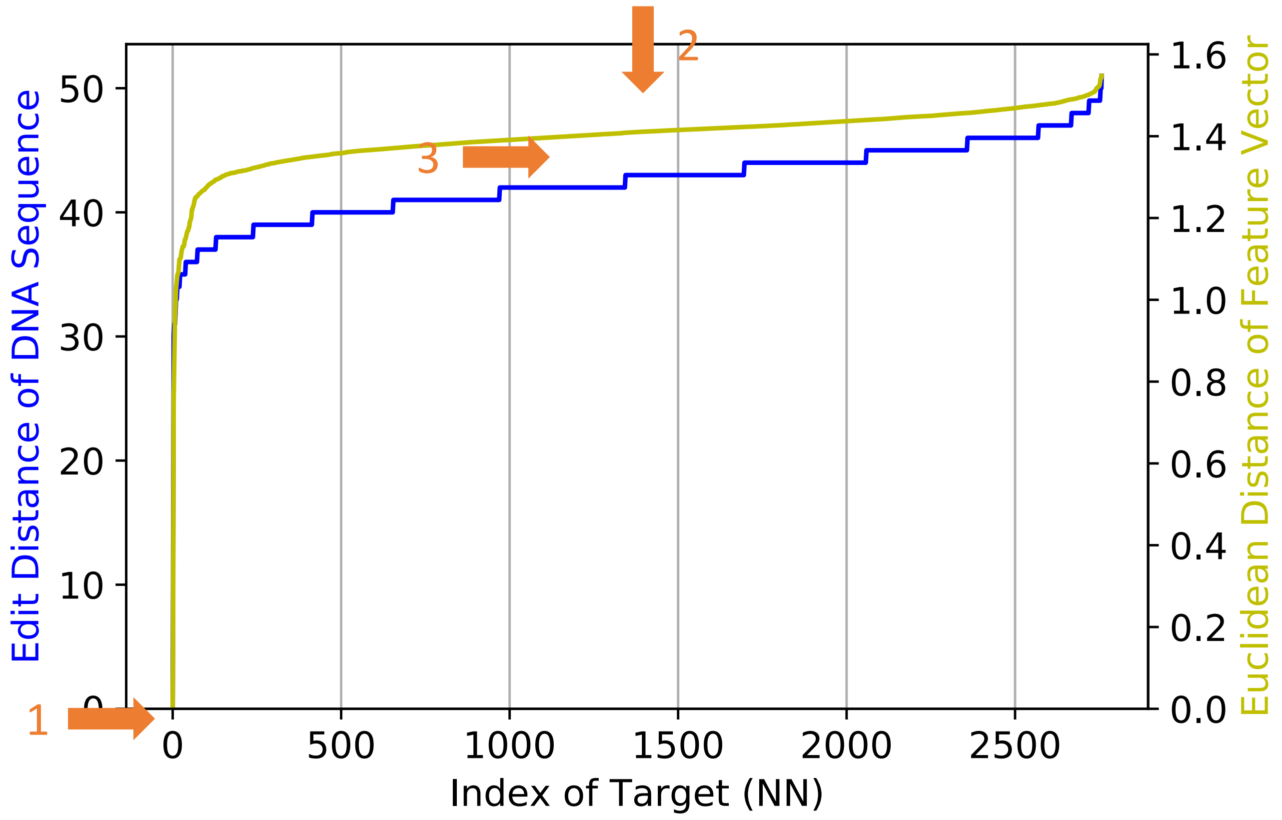






Results Analysis

- Case Study 2
  - Query: The cat is licking a bottle.
  - Dataset: STSB
  - Feature Extractor: MiniLM
  - Two Distance @ kNN
    1. Have accurate 1 NN
    2. A plateau in the middle
    3. Close match of two curves



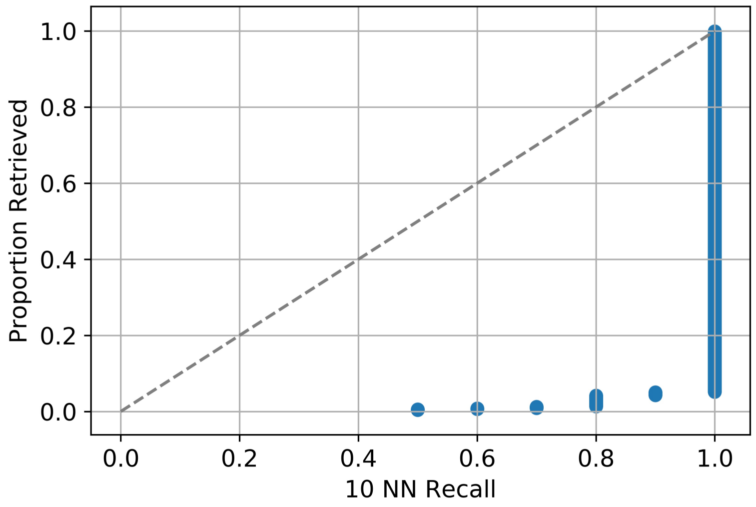
Results Analysis

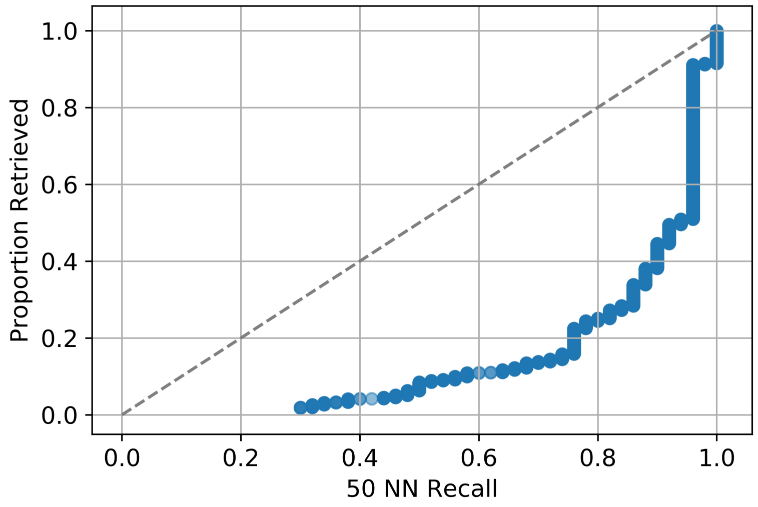
- Case Study 2
  - Query: The cat is licking a bottle.
  - Dataset: STSB
  - Feature Extractor: MiniLM
  - Semantic Similarity

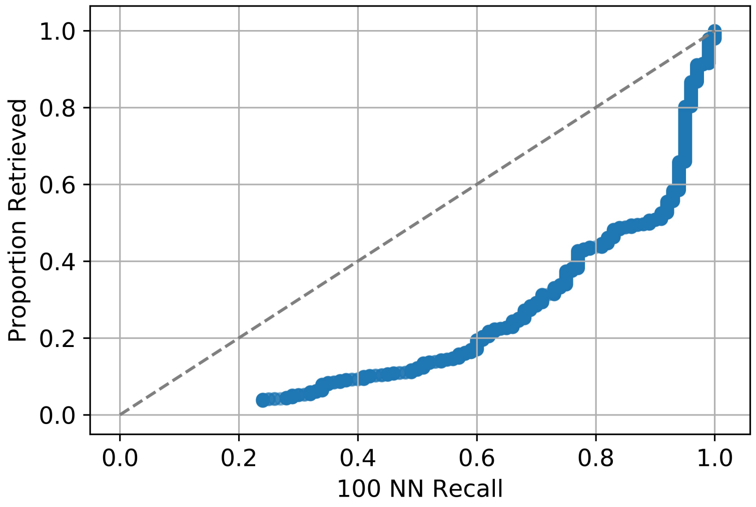
Feature Vector NN	DNA Sequence NN
A cat is licking a bottle.	A cat is licking a bottle.
A cat is licking itself.	A cat plays with a small bottle.
A cat plays with a small bottle.	A cat is licking itself.
A white cat is licking and drinking milk kept on a plate.	A white cat is licking and drinking milk kept on a plate.
…	…
A kitten is drinking milk from a bowl.	A cat is eating some corn.

Results Analysis

- Case Study 2
  - Query: The cat is licking a bottle.
  - Dataset: STSB
  - Feature Extractor: MiniLM
  - Retrieval quality @ 10NN, 50NN, 100NN







IV. Discussions

Discussions

- Limitations
  - Scalability
  - Efficiency
  - Hybridizations

Conclusion

- Contributions
  - Replicate the workflow on image data
  - Extend the framework to text data
  - Evaluate empirically the results of similarity search from text data

- The workflow introduced in the original paper can be extended to other modality.
- In particular, text data can be input in this framework and training process needs to be modified.
- The results show the encoded DNA sequence of sentences preserve semantic similarity

Reference

1. Bee, C., Chen, Y. J., Queen, M., Ward, D., Liu, X., Organick, L., ... & Ceze, L. (2021). Molecular-level similarity search brings computing to DNA data storage. *Nature communications*, *12*(1), 1-9.
2. Reimers, N., & Gurevych, I. (2019). Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.
3. Song, K., Tan, X., Qin, T., Lu, J., & Liu, T. Y. (2020). Mpnet: Masked and permuted pre-training for language understanding. *arXiv preprint arXiv:2004.09297*.
4. Wang, W., Wei, F., Dong, L., Bao, H., Yang, N., & Zhou, M. (2020). Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. *arXiv preprint arXiv:2002.10957*.

THANK YOU

National University of Singapore