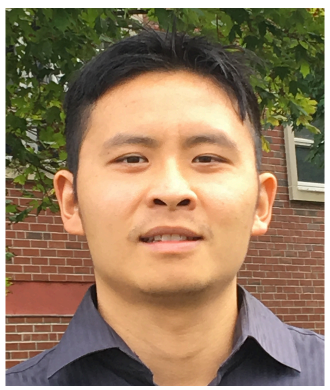
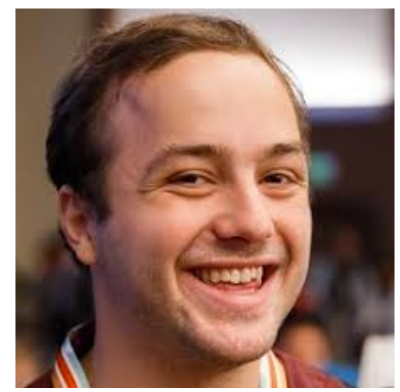
# Optimal Transport-based Alignment of Learned Character Representations for String Similarity



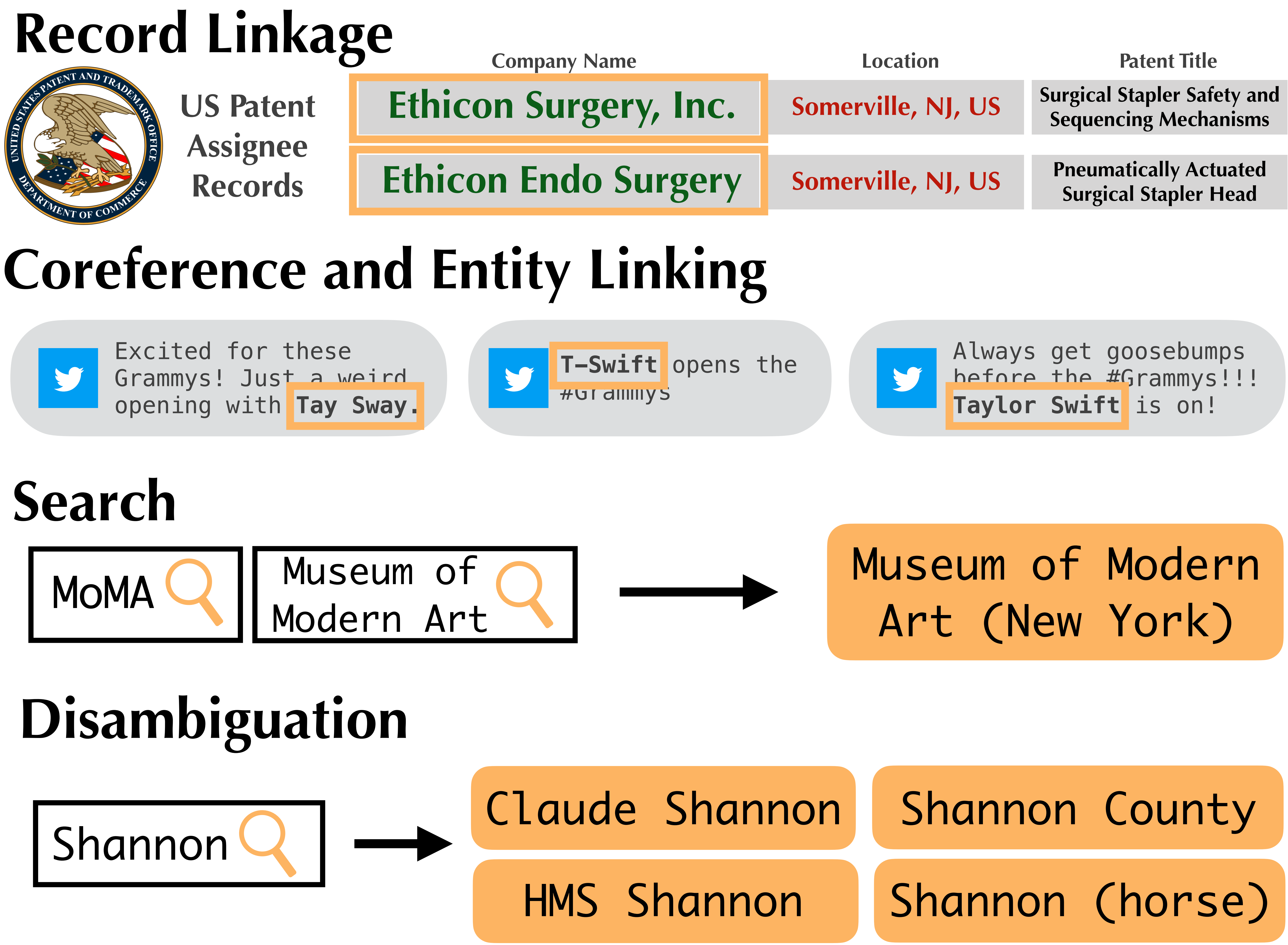



Derek Tam, Nicholas Monath, Ari Kobren, Aaron Traylor, Rajarshi Das, Andrew McCallum

IESL

UMass Amherst
College of Information and Computer Sciences

BROWN
Computer Science


Record Linkage
US Patent Assignee Records
Company Name
Location
Patent Title
Ethicon Surgery, Inc.
Somerville, NJ, US
Surgical Stapler Safety and Sequencing Mechanisms
Ethicon Endo Surgery
Somerville, NJ, US
Pneumatically Actuated Surgical Stapler Head
Coreference and Entity Linking
Excited for these Grammys! Just a weird opening with Tay Sway.
T-Swift opens the #Grammys
Always get goosebumps before the #Grammys!!! Taylor Swift is on!
Search
MoMA
Museum of Modern Art
Museum of Modern Art (New York)
Disambiguation
Shannon
Claude Shannon
Shannon County
HMS Shannon
Shannon (horse)

# Record Linkage

US Patent Assignee Records

| Company Name | Location | Patent Title |
|---|---|---|
| Ethicon Surgery, Inc. | Somerville, NJ, US | Surgical Stapler Safety and Sequencing Mechanisms |
| Ethicon Endo Surgery | Somerville, NJ, US | Pneumatically Actuated Surgical Stapler Head |

# Coreference and Entity Linking

Excited for these

T-Swift opens the

Always get goosebumps

**Similarity of mention strings informs whether or not they refer to the same entity.**

MoMA

Modern Art

Art (New York)

# Disambiguation

Shannon

Claude Shannon

Shannon County

HMS Shannon

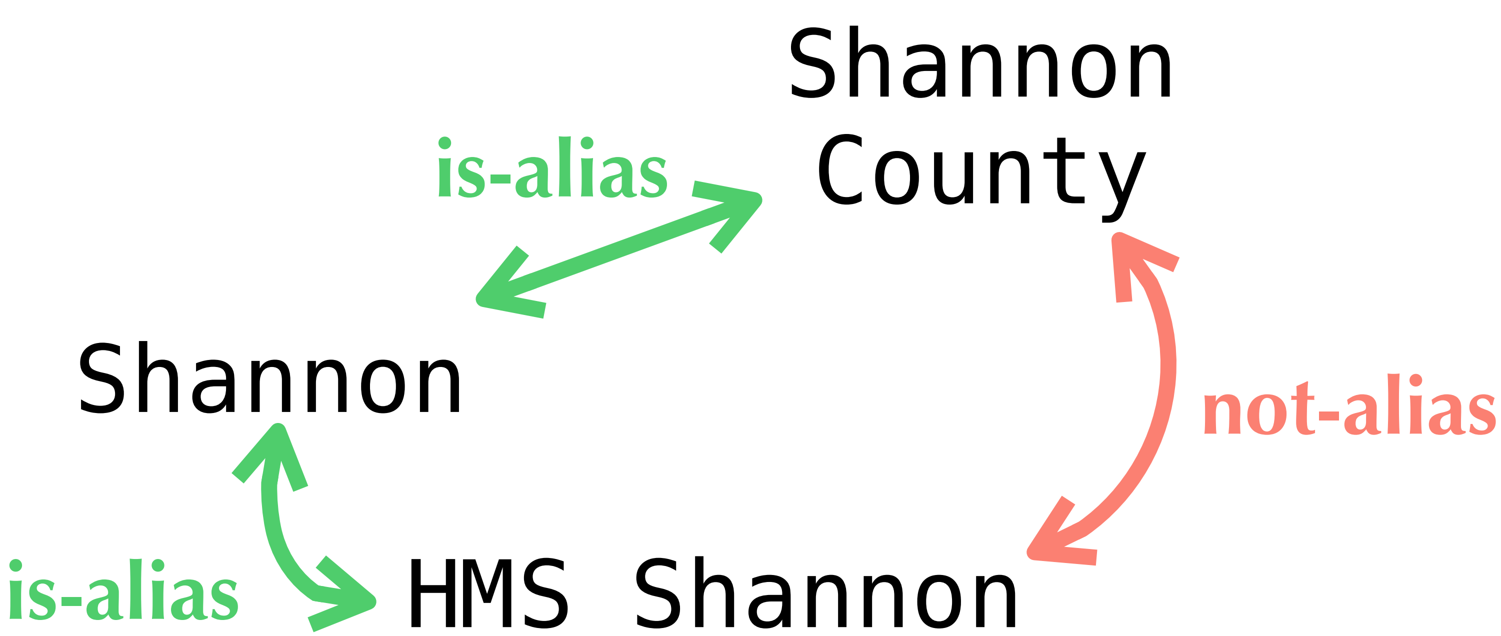
Shannon (horse)

# String Similarity for Entity Aliases

**Which strings can refer to the same entity?**

**Design similarity function *f***

| | | |
|---|---|---|
| ***f*(`string1`, `string2`)** | high similarity | if ***can refer*** to the same entity |
| ***f*(`string1`, `string2`)** | low similarity | if ***cannot refer*** to the same entity |




**Designed to inform coreference decisions**

# Classic Approaches

Similarity determined by number and type of edits

```
Music in Chile
```

**# Edits = 2**

```
Music in China
```

**# Edits = 12**

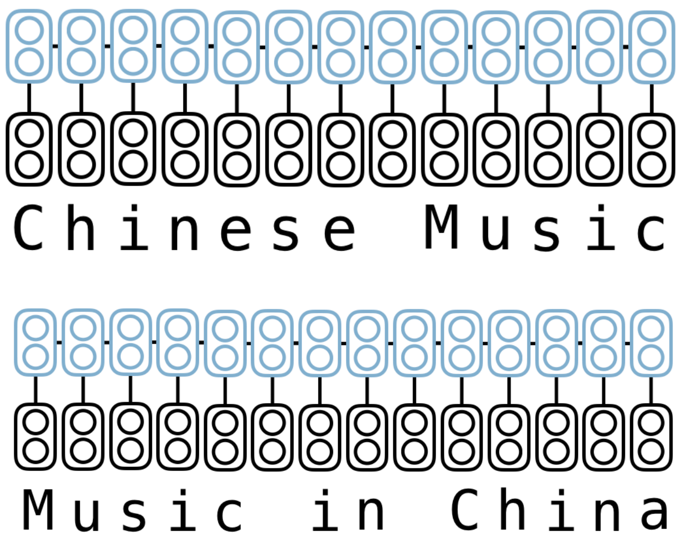
```
Chinese Music
```

***Character edits alone insufficient!***

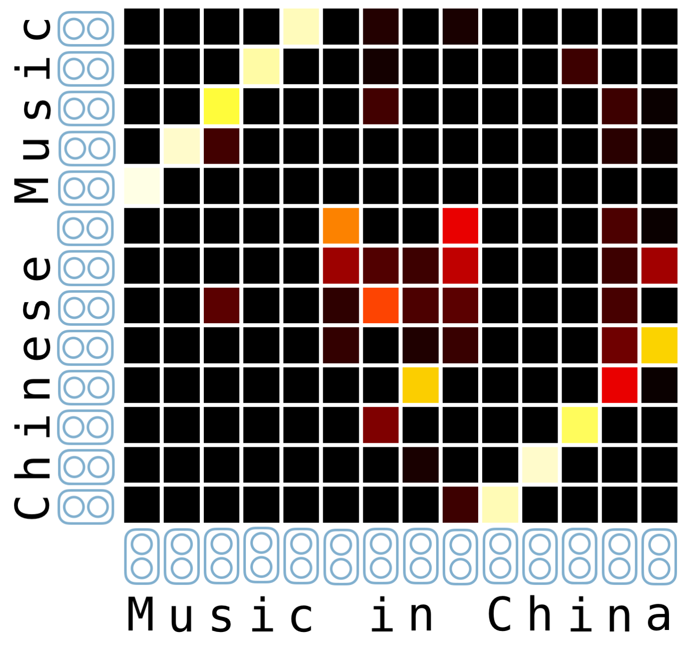
# STANCE

## Similarity of Transport Aligned Neural Character Encodings

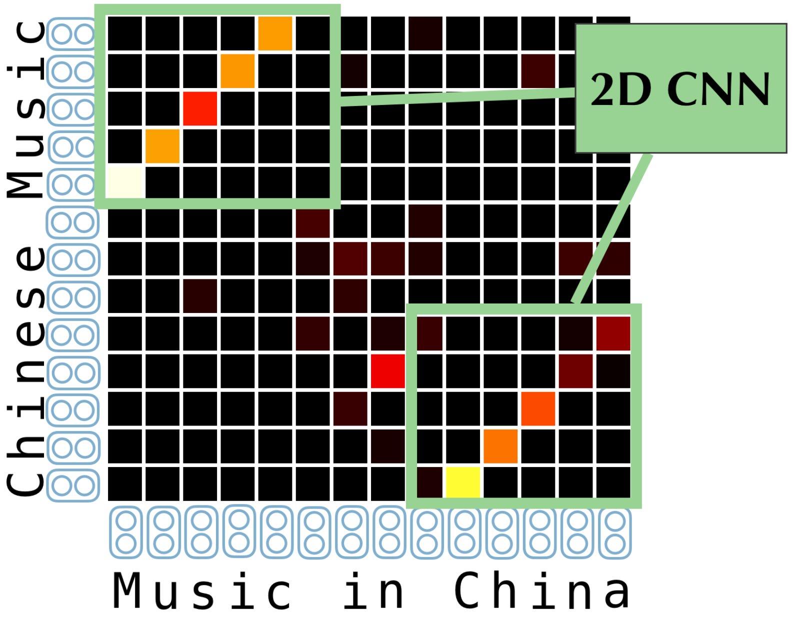
### Character Embeddings




### Optimal Transport based Alignment



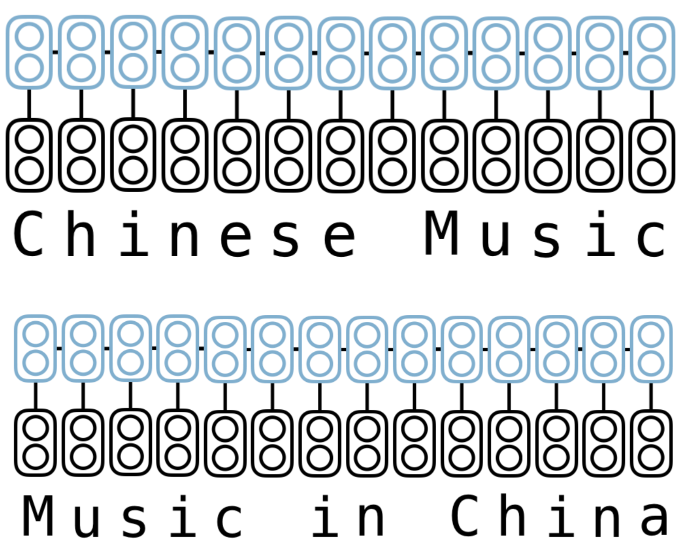

### CNN Scoring Function



# STANCE

## Similarity of Transport Aligned Neural Character Encodings

### Character Embeddings



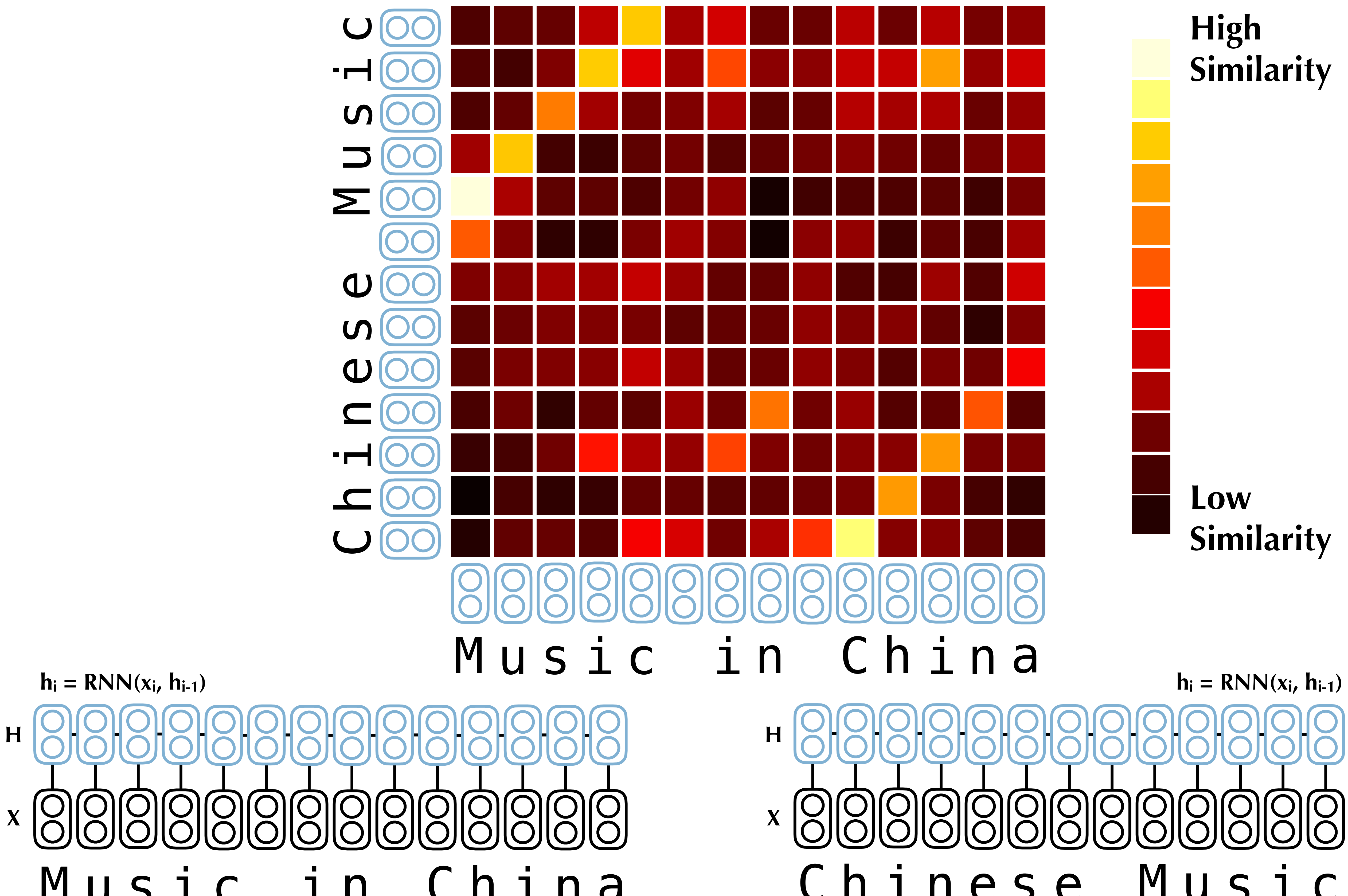

### Optimal Transport based Alignment




### CNN Scoring Function

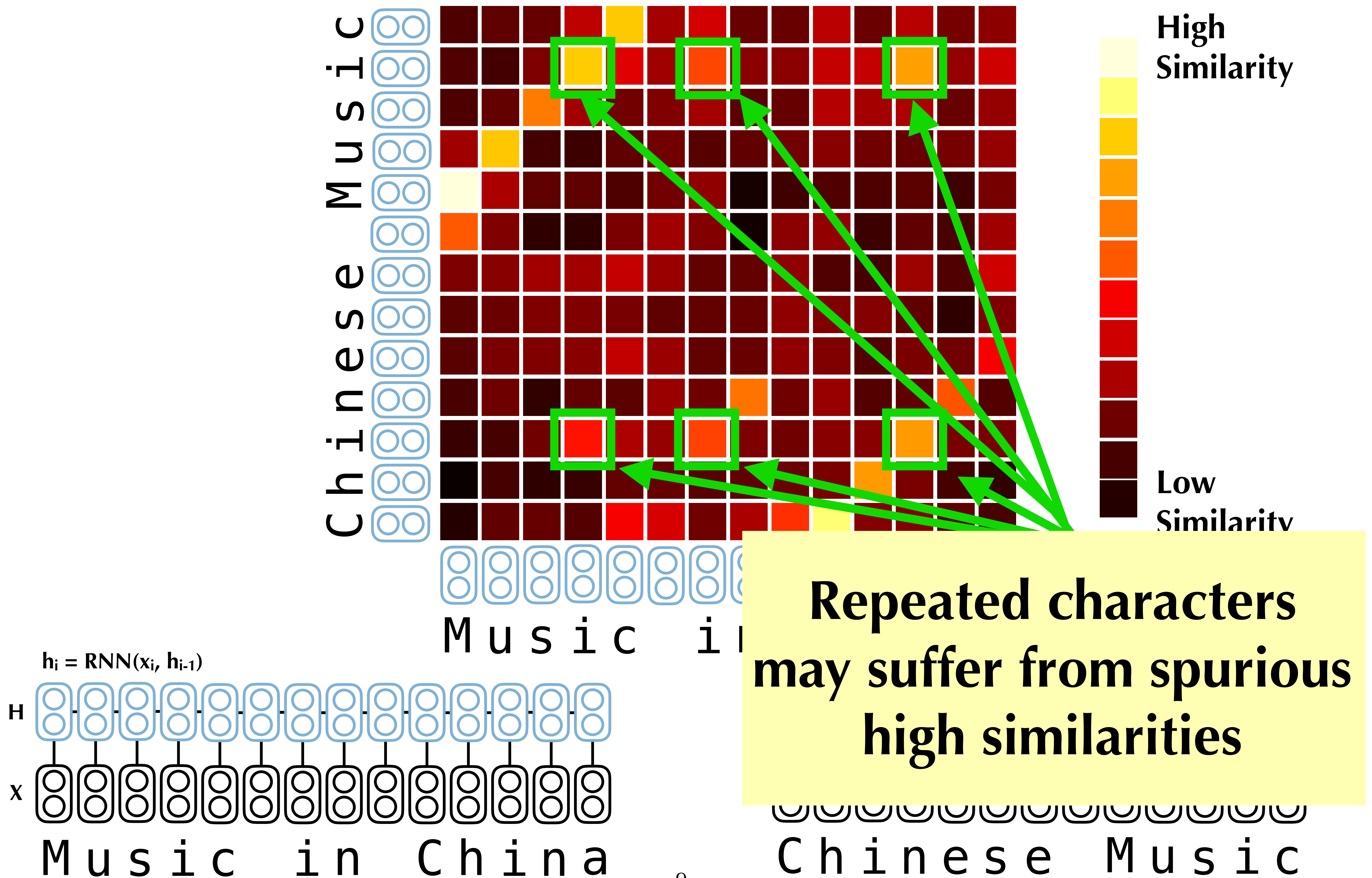
# Character Representations

Encode with RNN, Measure Pairwise Similarities



# Character Representations

Encode with RNN, Measure Pairwise Similarities




**Repeated characters may suffer from spurious high similarities**

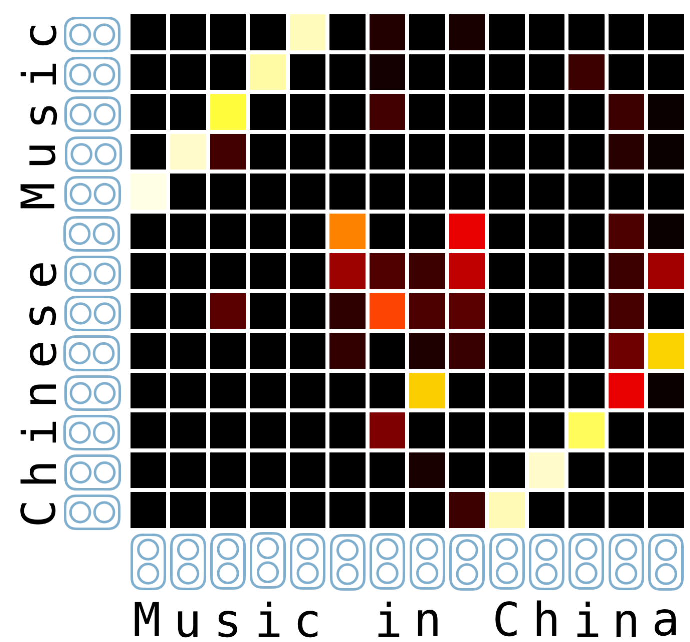
# STANCE

## Similarity of Transport Aligned Neural Character Encodings

### Character Embeddings



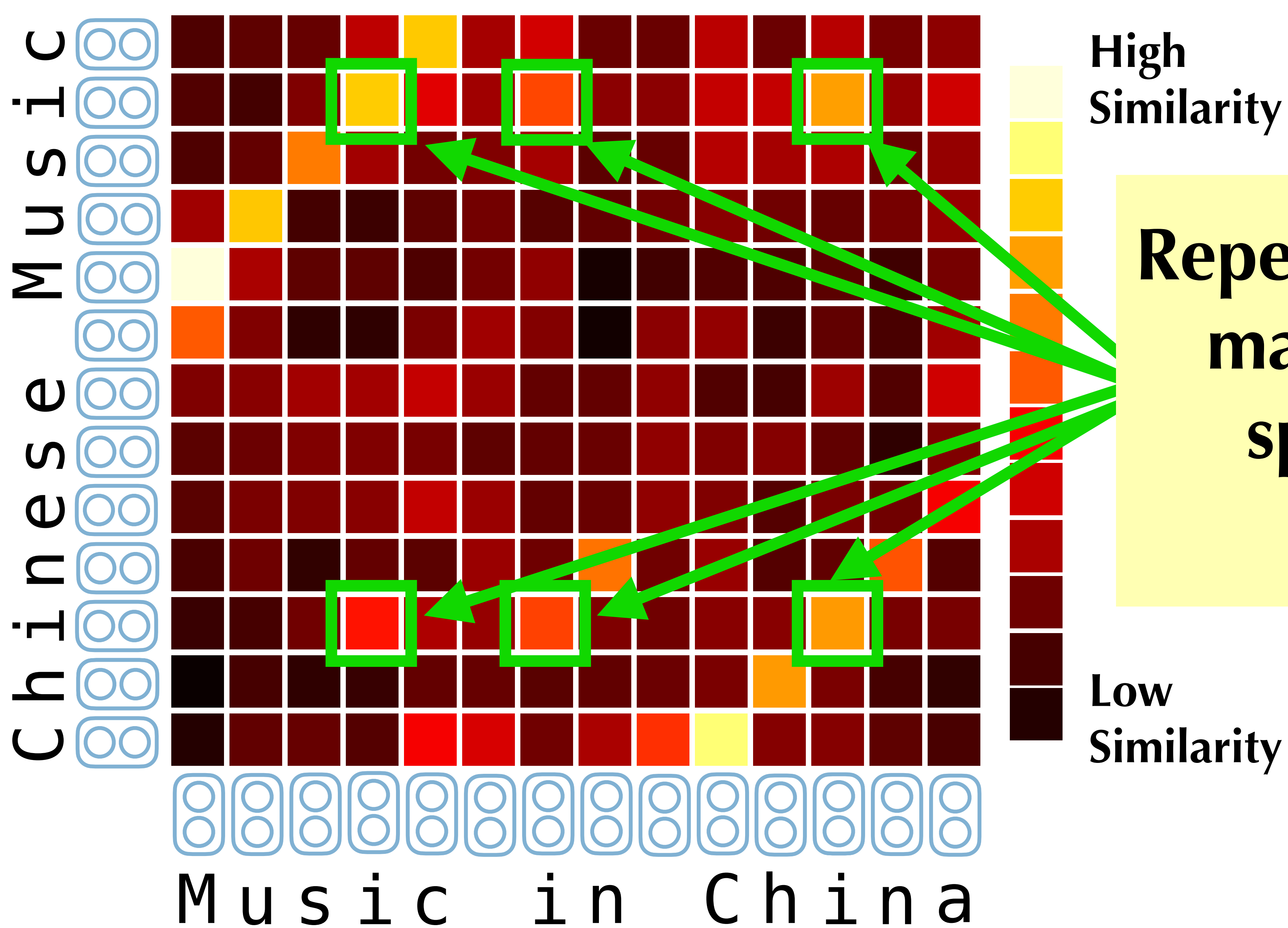

### Optimal Transport based Alignment




### CNN Scoring Function

# Optimal Transport-based Alignment

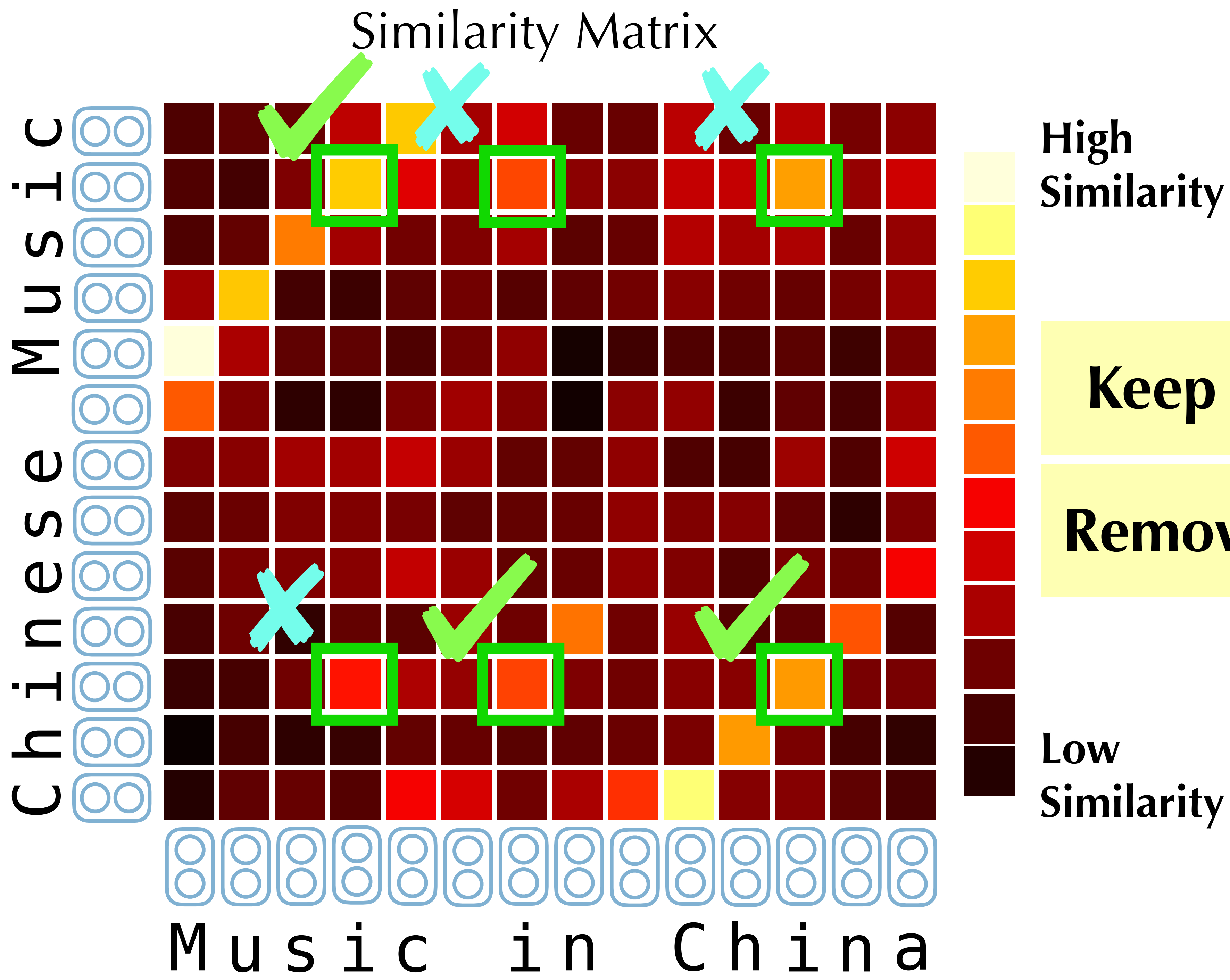
Similarity Matrix




**Repeated characters may suffer from spurious high similarities**

# Optimal Transport-based Alignment

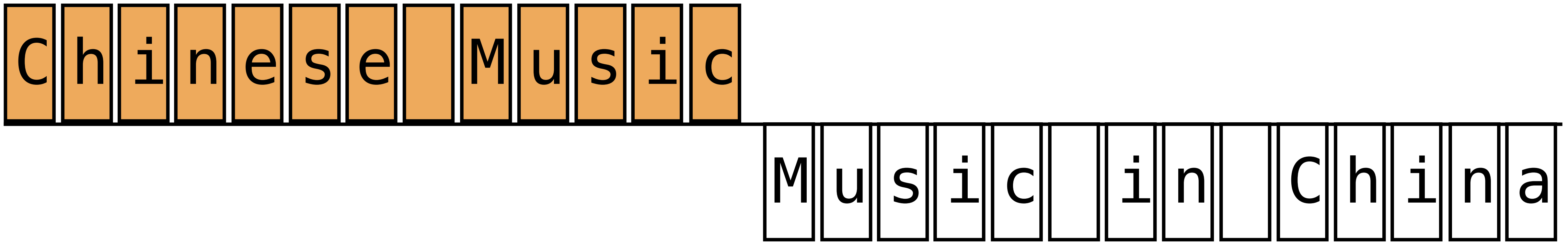
**Each character aligned to closest character(s) in other string**




**Keep "good" alignments**

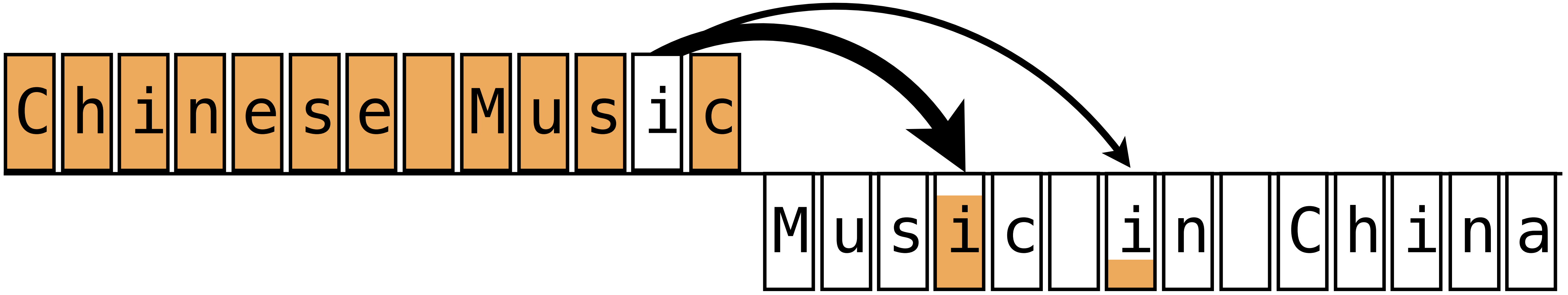
**Remove "bad" alignments**

# Alignment as Optimal Transport



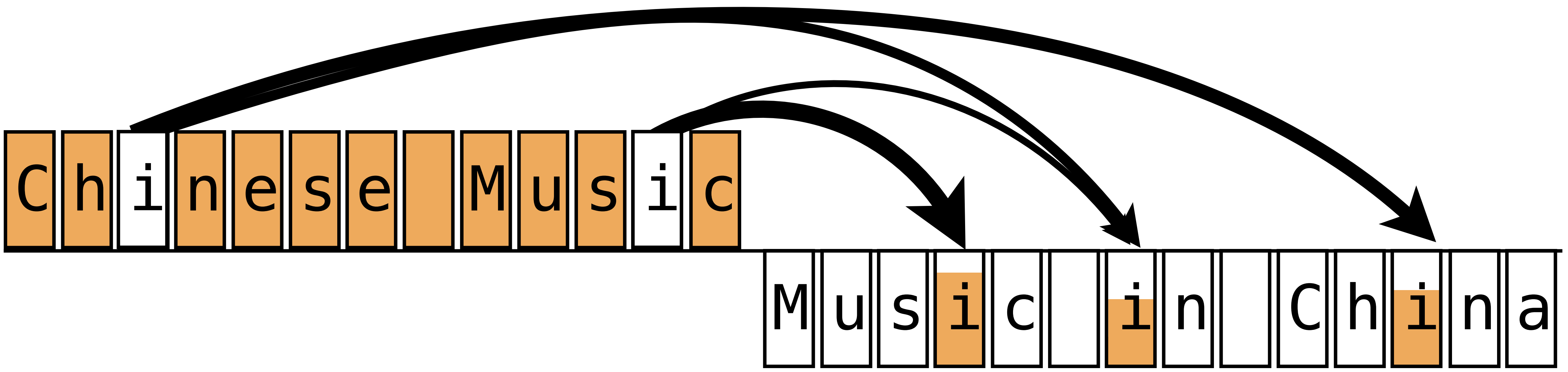

**The amount of transported mass indicates degree of alignment.**

# Alignment as Optimal Transport




**The amount of transported mass indicates degree of alignment.**

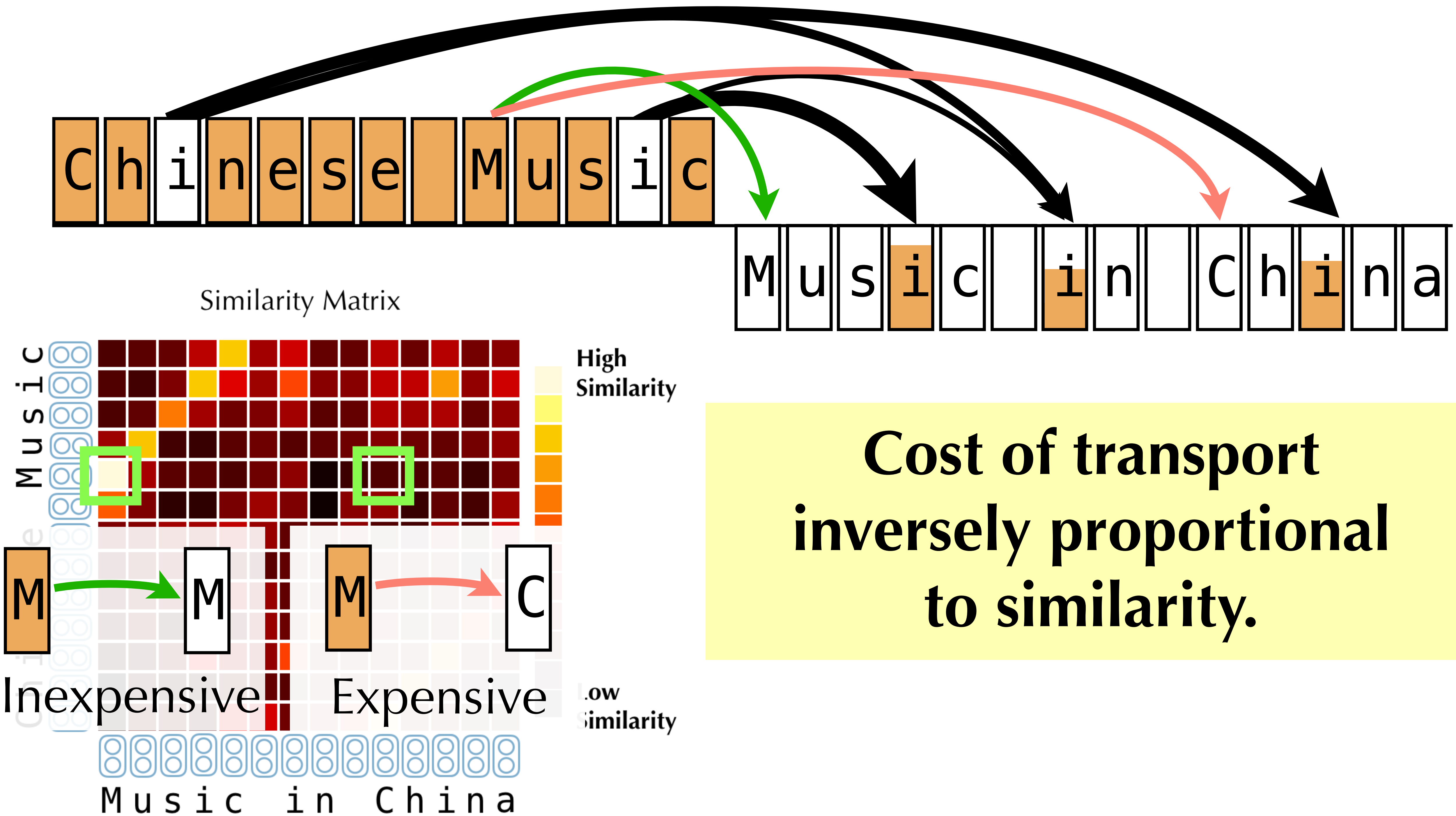
# Alignment as Optimal Transport




**The amount of transported mass indicates degree of alignment.**

# Alignment as Optimal Transport

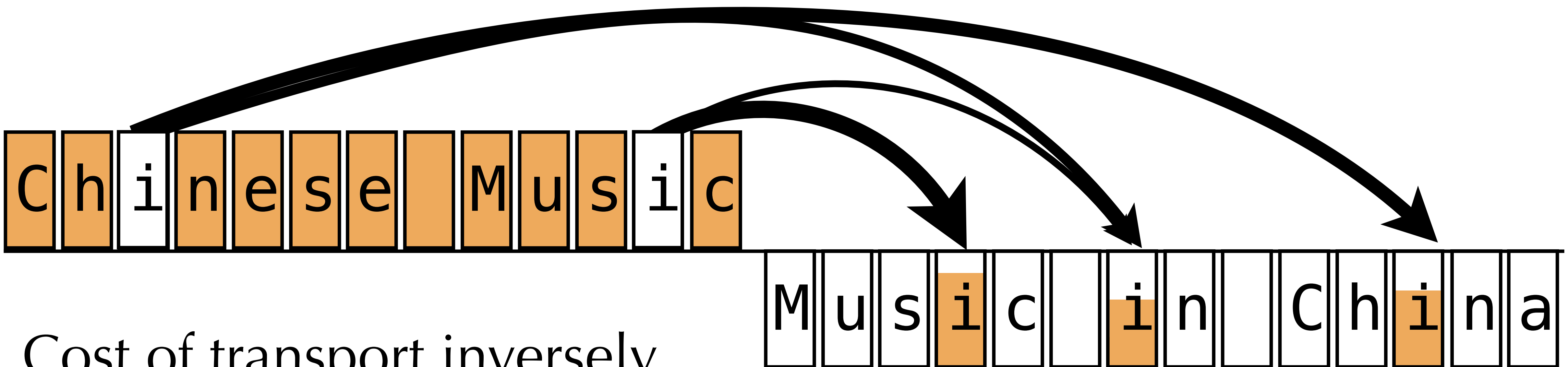
The amount of transported mass indicates degree of alignment.




**Cost of transport inversely proportional to similarity.**

# Alignment as Optimal Transport

The amount of transported mass indicates degree of alignment.




Cost of transport inversely proportional to similarity.

**Characters have fixed amount of mass to transport (or receive).**

To transport:

mass( M ) = 1/ StringLength

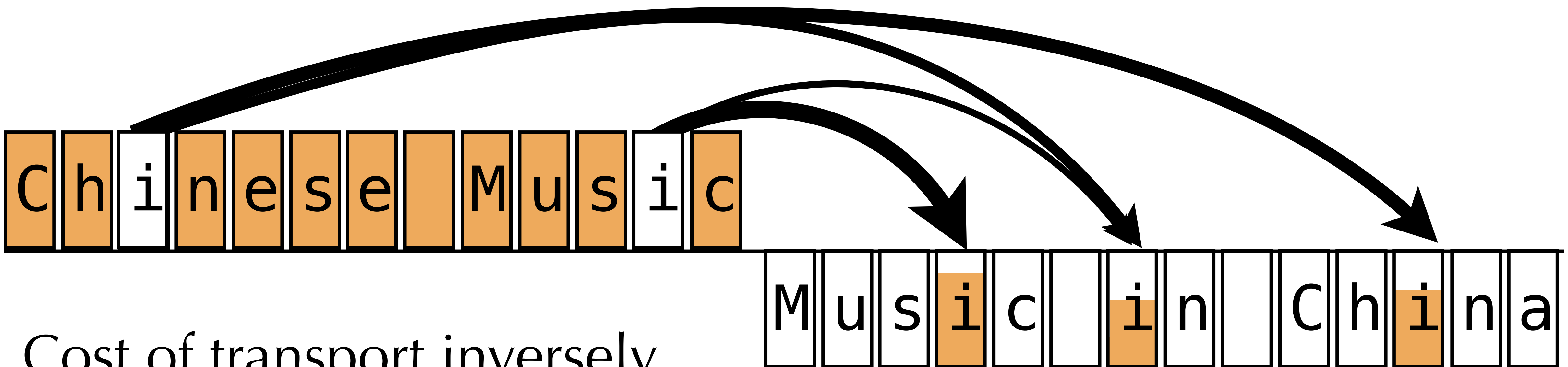
To receive:

mass( M ) = 1/ StringLength

**All characters must transport (or receive) entire mass.**

# Alignment as Optimal Transport

The amount of transported mass indicates degree of alignment.




Cost of transport inversely proportional to similarity.

$$\text{AlignmentCost}(S1,S2) = \sum_{i\in S1}\sum_{j\in S2} T_{i\to j}\ \text{Cost}(i, j)$$

Characters in S1 (e.g., "Chinese Music")

Characters in S2 (e.g., "Music in China")

How much of i is transported to j

Inversely proportional to similarity of i & j

# Alignment as Optimal Transport

The amount of transported mass indicates degree of alignment.




**Find minimum cost alignment between characters of the two strings**

$$\sum_{i \in S1} \sum_{j \in S2} T_{i \to j} \, \mathrm{Cost}(i, j)$$

Characters in S1 (e.g., "Chinese Music")

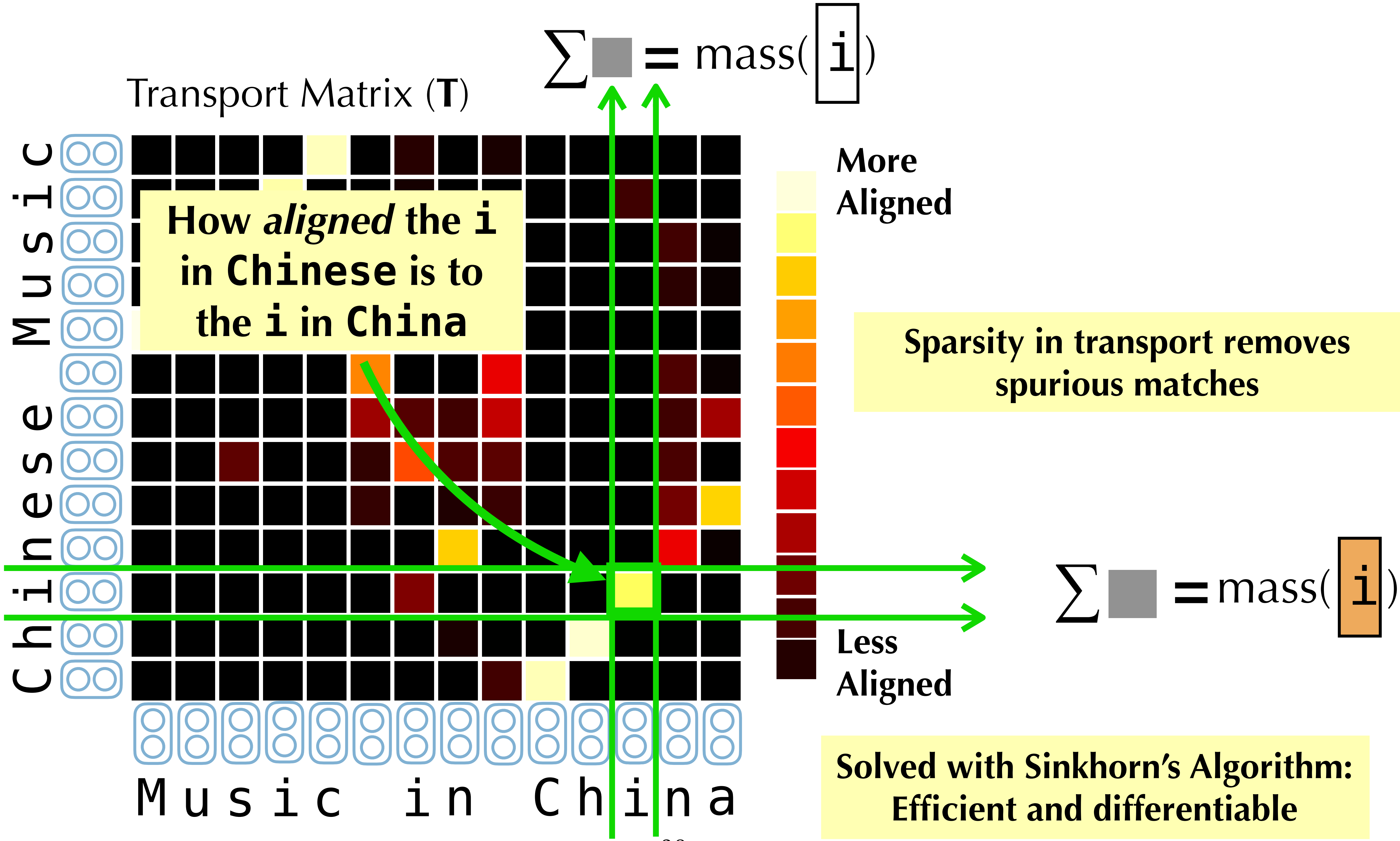
Characters in S2 (e.g., "Music in China")

How much of i is transported to j

Inversely proportional to similarity of i & j

# Optimal Transport-based Alignment

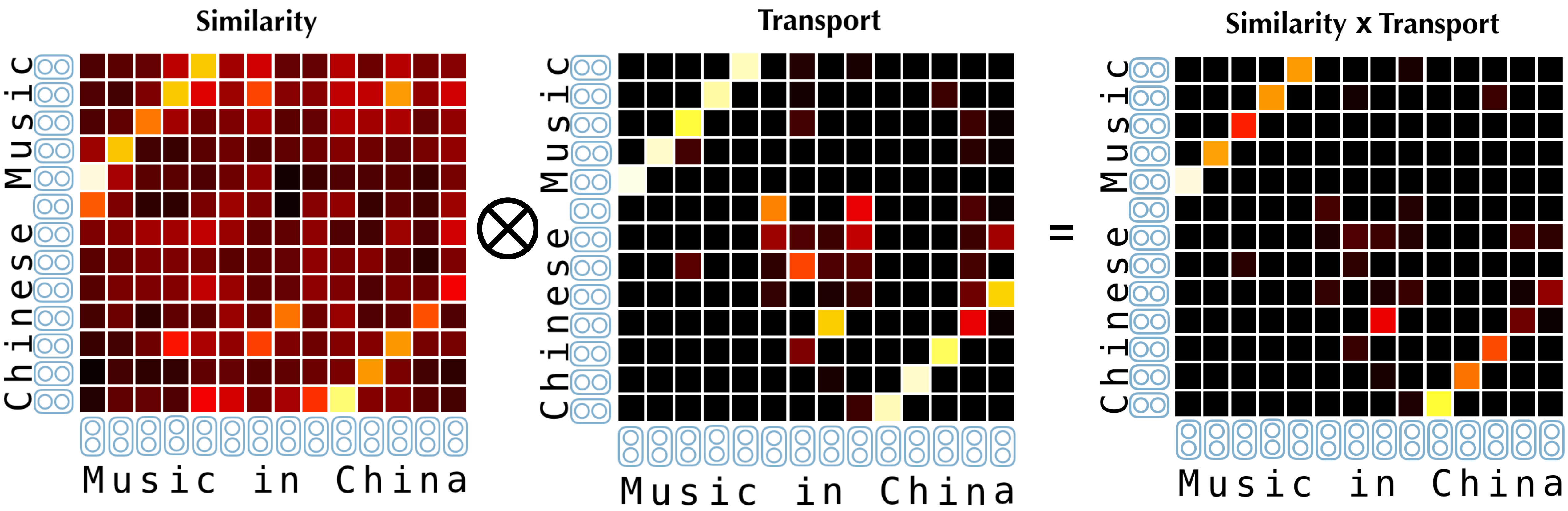
Minimum cost soft alignment btw characters of the two strings




**Solved with Sinkhorn's Algorithm: Efficient and differentiable**

# Optimal Transport-based Alignment

**Re-weight similarity by transport matrix**



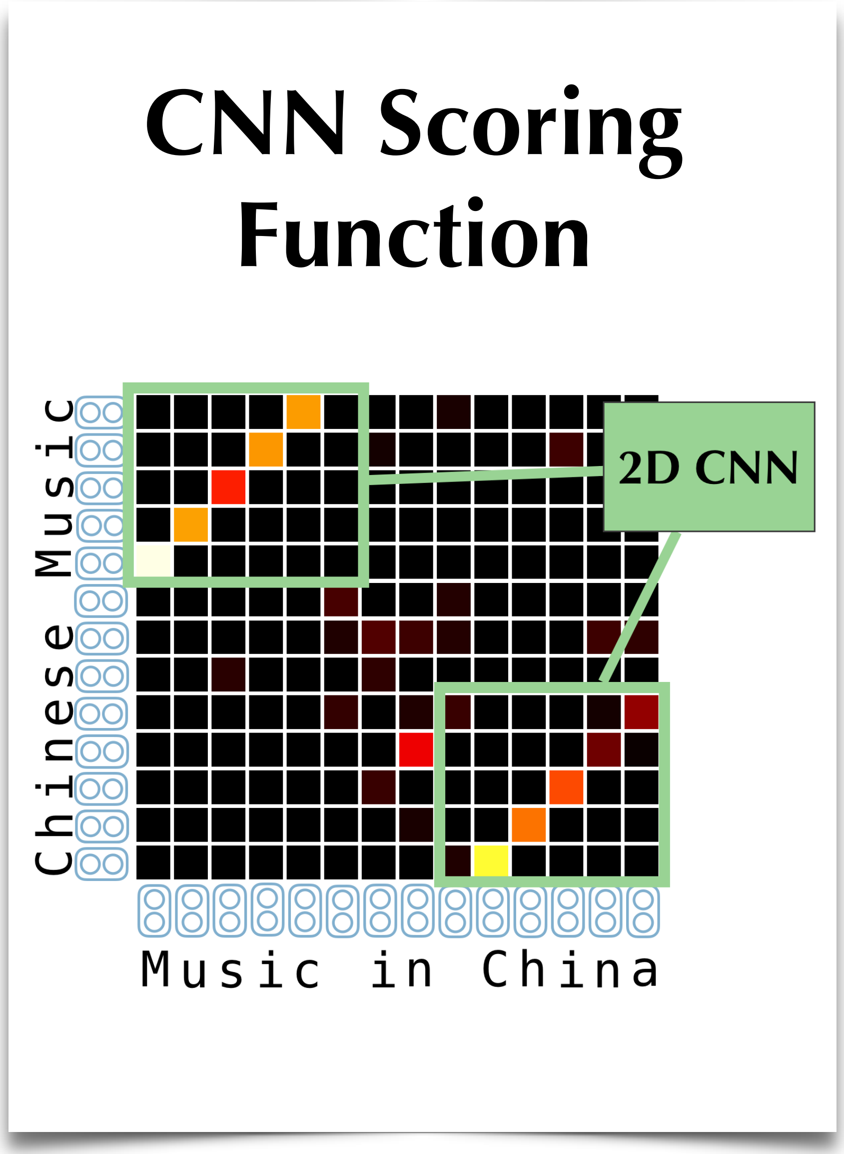
# STANCE

## Similarity of Transport Aligned Neural Character Encodings



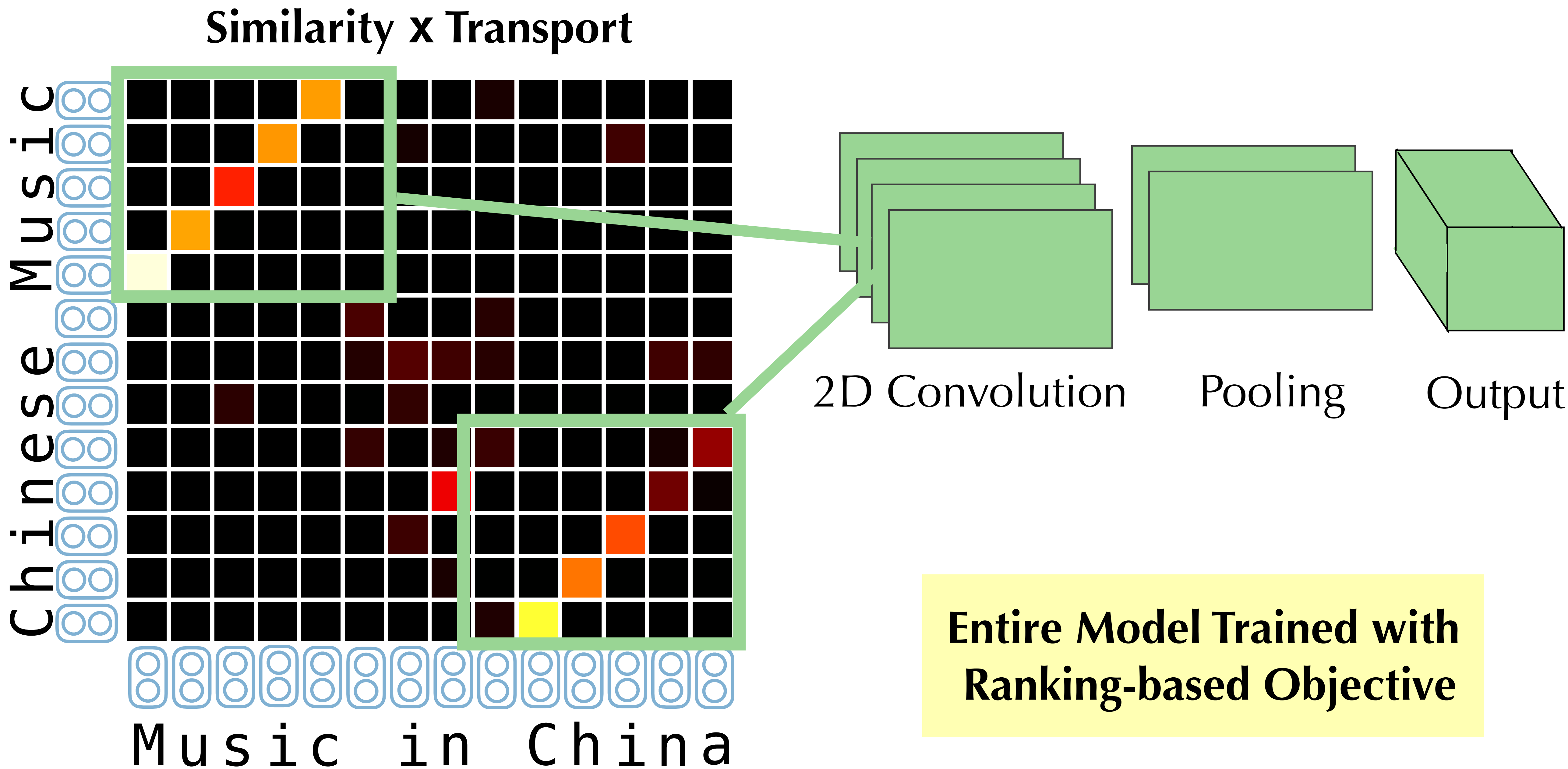






# CNN Scoring Function

Capture patterns of sequential alignment between characters.




**Entire Model Trained with Ranking-based Objective**

# Experimental Results

## Task 1: Alias Detection

## Task 2: Cross Document Coreference

## Qualitative Analysis & Ablation Study

# Alias Detection

**Aliases - Two strings that *can* refer to the same entity**

**Given a *query* string, *rank* candidate aliases.**

***Query***

Peace Agreement

***Candidates***

Peace Treaty

Peace Pact

Lease Agreement

Peacekeeping Troops

***Ranking***

Peace Treaty

Peace Pact

Peacekeeping Troops

Lease Agreement

# Datasets

**Built 5 datasets for alias detection from open KBs**

**Wikipedia**

Irish music **is-alias** Irish Folk

# Datasets

**Built 5 datasets for alias detection from open KBs**

**Wikipedia** **Wikipedia-People**

Queen Elizabeth II **is-alias**
Queen Elizabeth the Second

# Datasets

**Built 5 datasets for alias detection from open KBs**

**Wikipedia** | **Wikipedia-People** | **Music Artist**

Red Hot Chili Peppers **is-alias** RHCP

# Datasets

**Built 5 datasets for alias detection from open KBs**

**Wikipedia** | **Wikipedia-People** | **Music Artist**

**Patent Assignee**

The Proctor & Gamble Company **is-alias** Proctor and Gamble

# Datasets

**Built 5 datasets for alias detection from open KBs**

**Wikipedia** **Wikipedia-People** **Music Artist**

**Patent Assignee** **Disease**

black water fever **is-alias** hemolytic malaria

# Alias Detection Experiments

Compare STANCE to 8 baseline methods including:

## Alignment Methods

- Levenshtein Similarity
- Learned Dynamic Time Warping - LDTW (Cuturi et al. 2017)

## Neural Methods

- Deep Conflation Model - DCM (Gan et al. 2017)

# Alias Detection Experiments

Compare STANCE to 8 baseline methods including:

**Alignment Methods**

- Levenshtein Similarity
- Learned Dynamic Time Warping - LDTW (Cuturi et al. 2017)

**Neural Methods**

- Deep Conflation Model - DCM (Gan et al. 2017)

# Alias Detection Experiments

Compare STANCE to 8 baseline methods including:

## Alignment Methods

- Levenshtein Similarity
- Learned Dynamic Time Warping - LDTW (Cuturi et al. 2017)

## Neural Methods

- Deep Conflation Model - DCM (Gan et al. 2017)

# Alias Detection Experiments

Compare STANCE to 8 baseline methods including:

## Alignment Methods

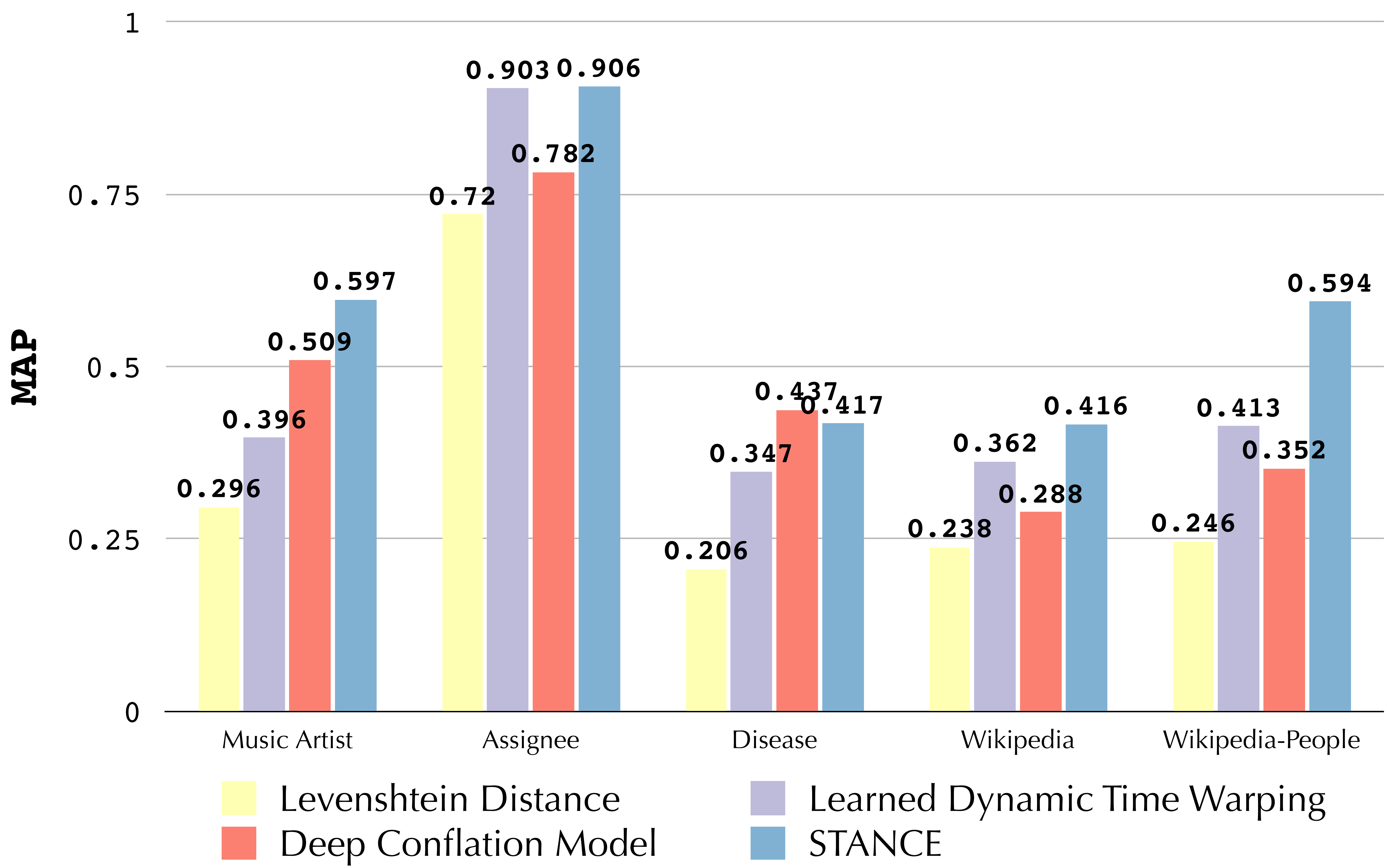
- Levenshtein Similarity

- Learned Dynamic Time Warping - LDTW (Cuturi et al. 2017)

## Neural Methods

- Deep Conflation Model - DCM (Gan et al. 2017)

# Alias Detection - Mean Average Precision (MAP)



# Experimental Results

Task 1: Alias Detection

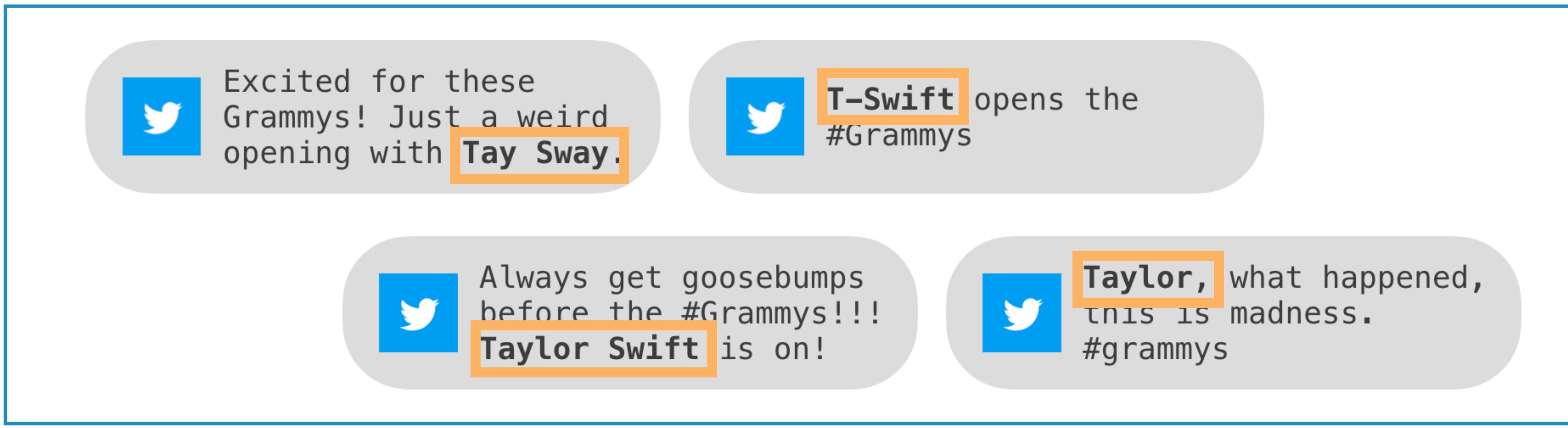
Task 2: Cross Document Coreference

Qualitative Analysis & Ablation Study

# Cross-Document Coreference

**Twitter at the Grammy's Dataset (Dredze et al, 2016)**

**4577 Mentions, 273 Entities**



# Cross-Document Coreference

## Twitter at the Grammy's Dataset (Dredze et al, 2016)

## 4577 Mentions, 273 Entities

Excited for these Grammys! Just a weird opening with **Tay Sway**.

**T-Swift** opens the #Grammys

Always get goosebumps before the #Grammys!!! **Taylor Swift** is on!

**Taylor**, what happened, this is madness. #grammys

---

**LL Cool J** has swag for days. No better person to host the #Grammys!

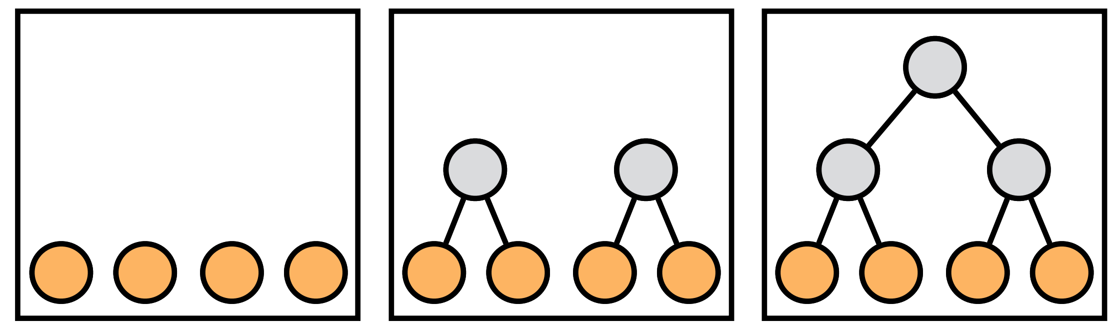
**El-El Cool John**. #Grammy

**LL Cool James** just mispronounced @edsheeran's name AGAIN at the #Grammys!

# Cross-Document Coreference

**Our approach**

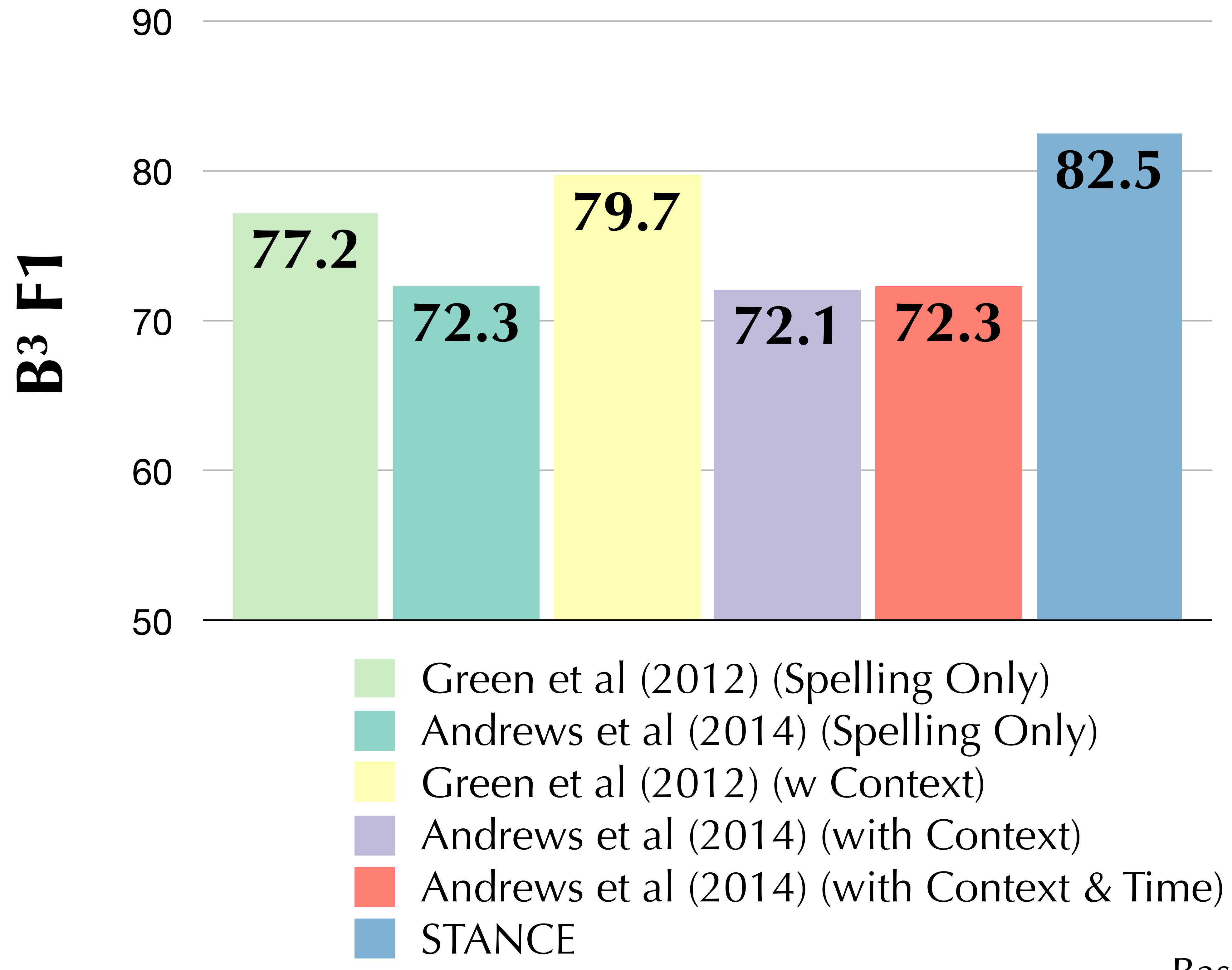
Average-Linkage
**Hierarchical Agglomerative Clustering**



Use **pre-trained STANCE** model on Wikipedia-People as **pairwise similarity function**.

Tune **threshold** to cut tree for **predicting entities** on dev set

# Cross-Document Coreference Performance




Baseline Results from Dredze et al (2016)

# Cross-Document Coreference

## Twitter at the Grammy's Dataset (Dredze et al, 2016)

Excited for these Grammys! Just a weird opening with **Tay Sway.**

**T-Swift** opens the #Grammys

Always get goosebumps before the #Grammys!!! **Taylor Swift** is on!

**Taylor,** what happened, this is madness. #grammys

**LL Cool J** has swag for days. No better person to host the #Grammys!

**El-El Cool John.** #Grammy

**LL Cool James** just mispronounced @edsheeran's name AGAIN at the #Grammys!

**Name variation more informative than context**

# Experimental Results

## Task 1: Alias Detection

## Task 2: Cross Document Coreference

## Qualitative Analysis & Ablation Study

# Qualitative Analysis

**Query:** Boom Microphones

**Nearest Neighbors:**

| | | | | |
|---|---|---|---|---|
| **STANCE** | Boom mike |  | Boom mics |  |
| **LDTW** | Open Microphone |  | Shotgun Microphone |  |
| **DCM** | Open Microphone |  | Condensor Microphone |  |

# Qualitative Analysis

**Query:** RPM

**Nearest Neighbors:**

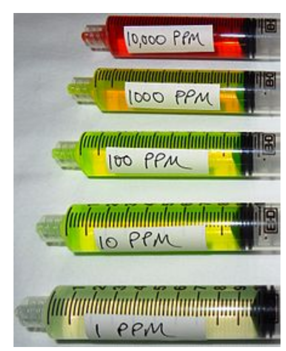
**STANCE**

RPM Weekly

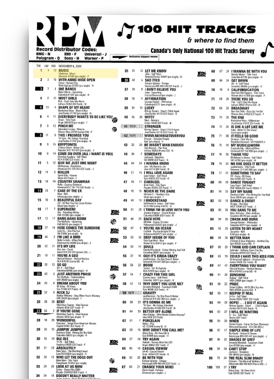

Randle Patrick McMurphy

**LDTW**

PPM




RPM Alternative 30



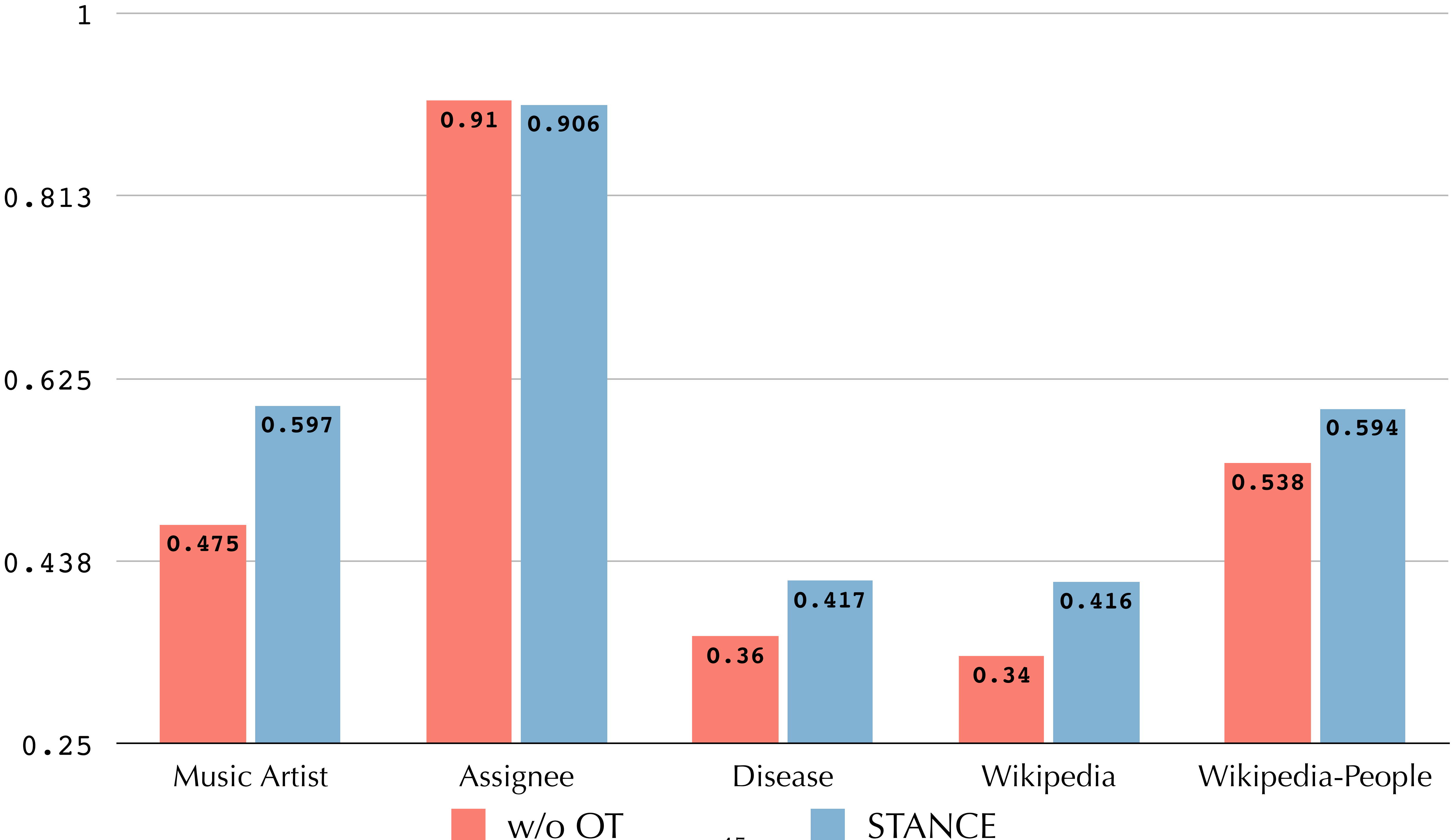

**DCM**

RP1

PRM

# Impact of Optimal Transport in STANCE

## OT component improves results on 4 of 5 datasets.



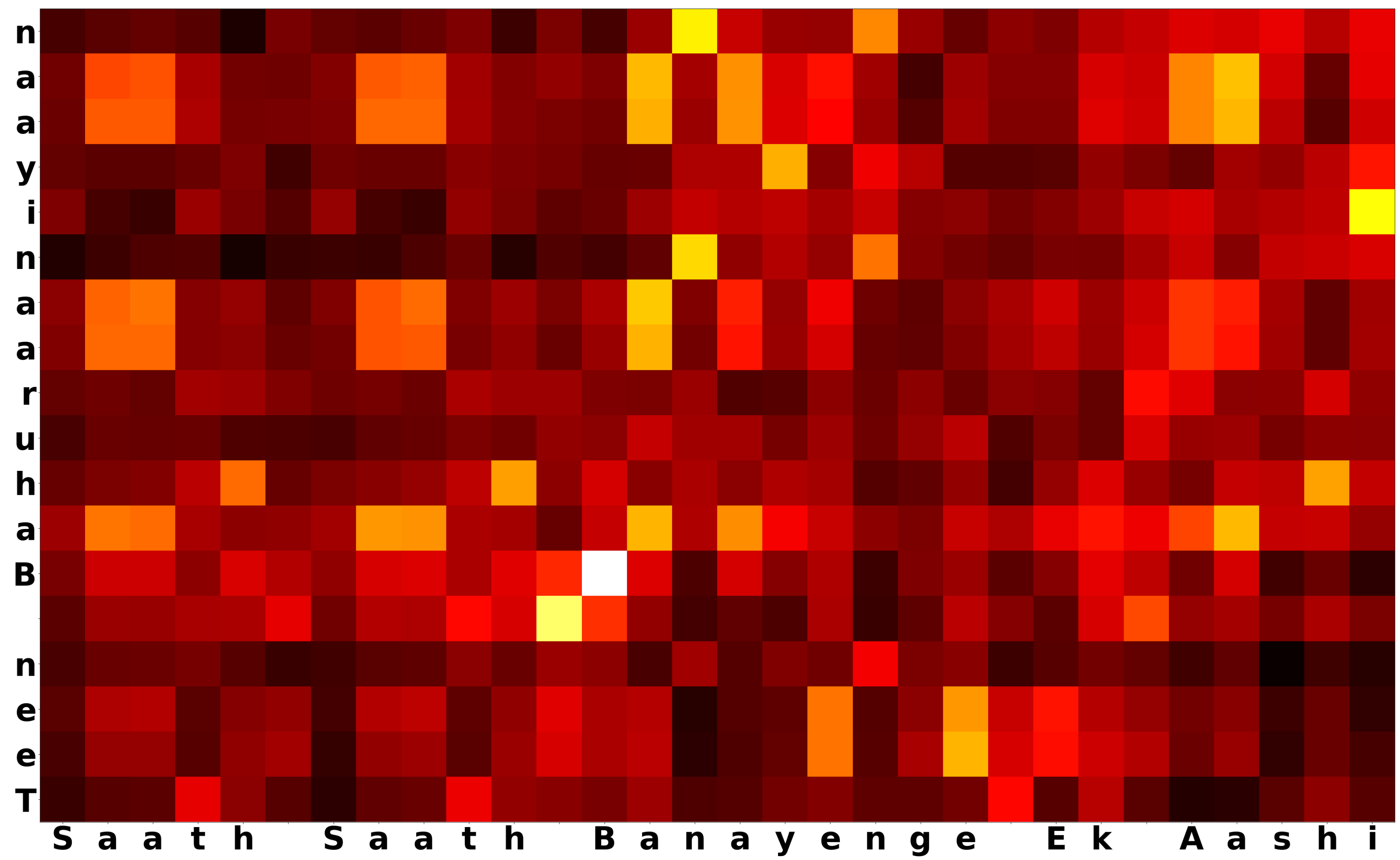
# Benefit of OT - Noise Reduction

**Query:** `Saath Saath Banayenger Ek Aashi`

**Non-Alias Candidate**: `Teen Bahuraaniyaan`

**Significant number of repeated characters and character bigrams**

**Similarity Matrix - w/o OT**



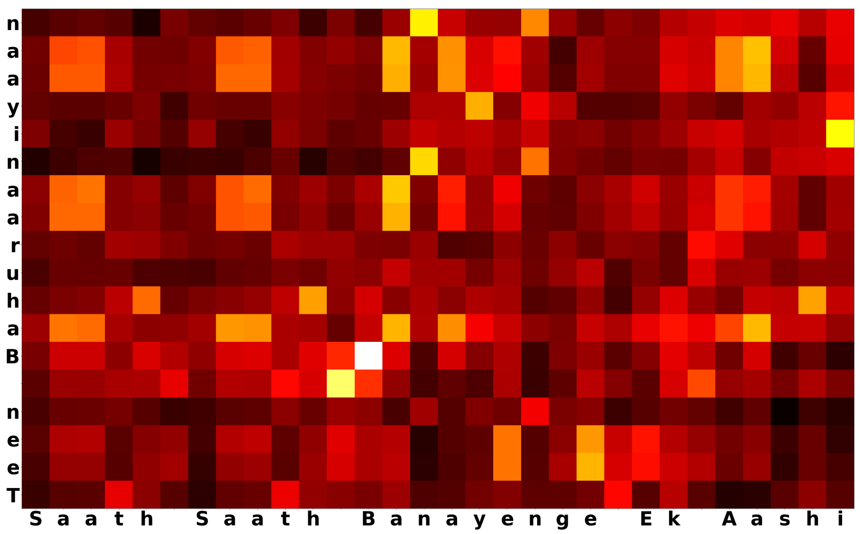
# Benefit of OT - Noise Reduction

**Query:** `Saath Saath Banayenger Ek Aashi`

**Non-Alias Candidate**: `Teen Bahuraaniyaan`

**Significant number of repeated characters and character bigrams**

**Similarity Matrix - w/o OT**



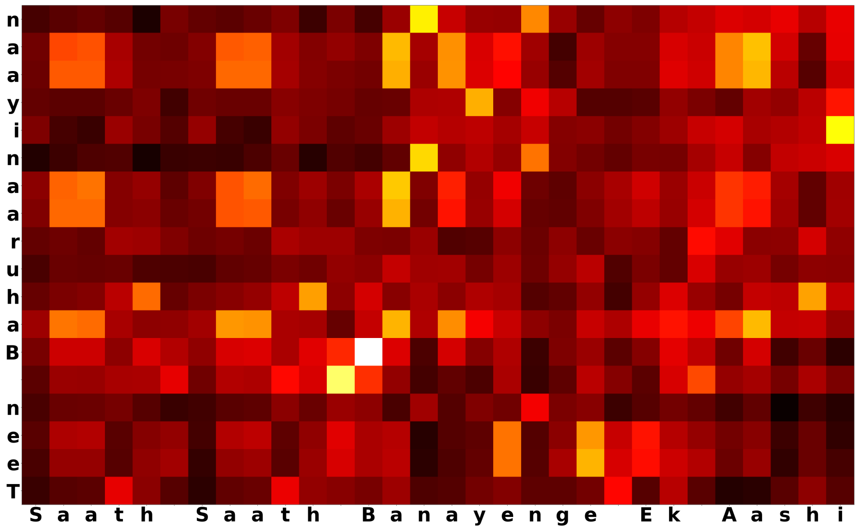
# Benefit of OT - Noise Reduction

**Query:** `Saath Saath Banayenger Ek Aashi`

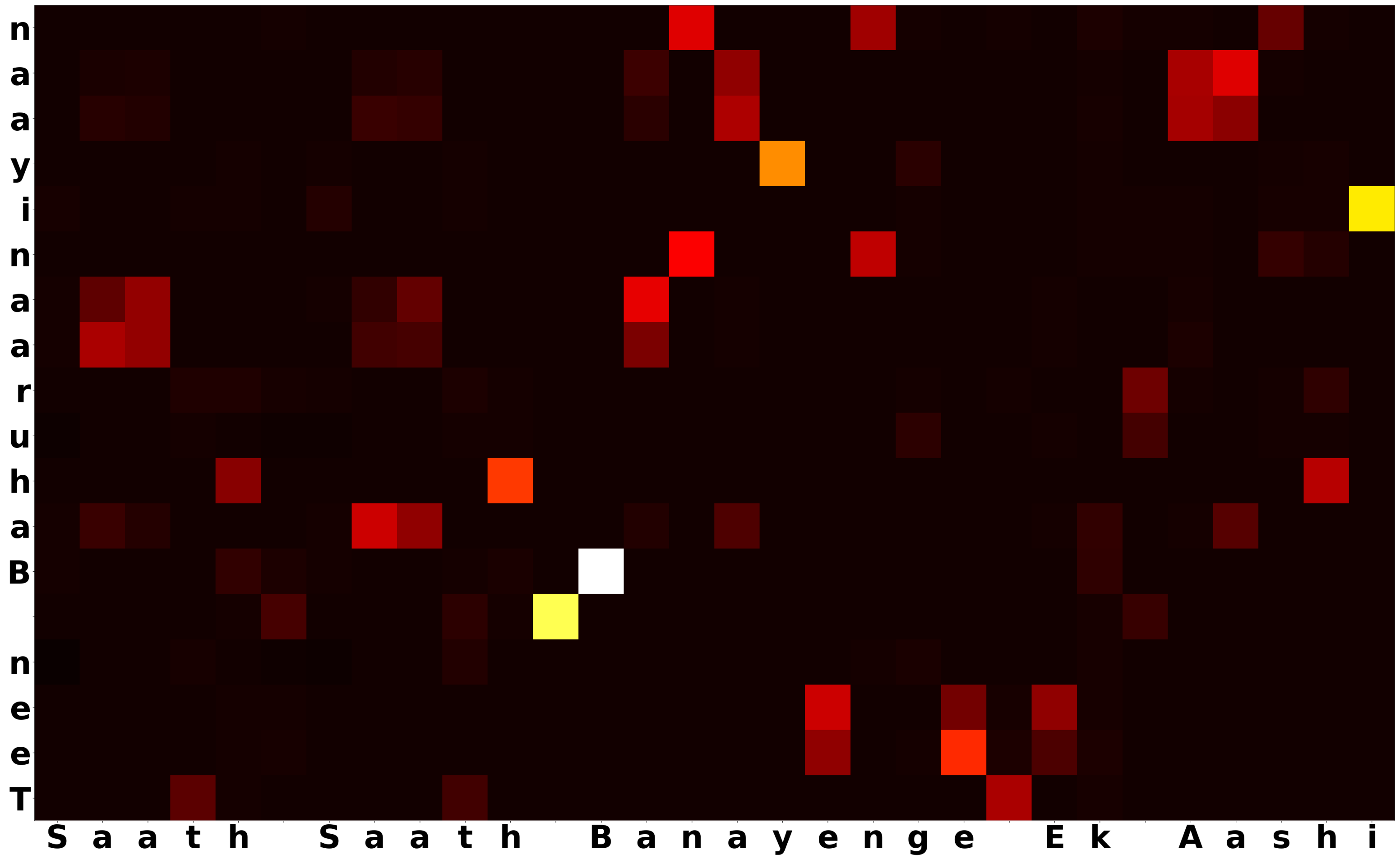
**Non-Alias Candidate:** `Teen Bahuraaniyaan`

**Significant number of repeated characters and character bigrams**

**Similarity Matrix - w/o OT**

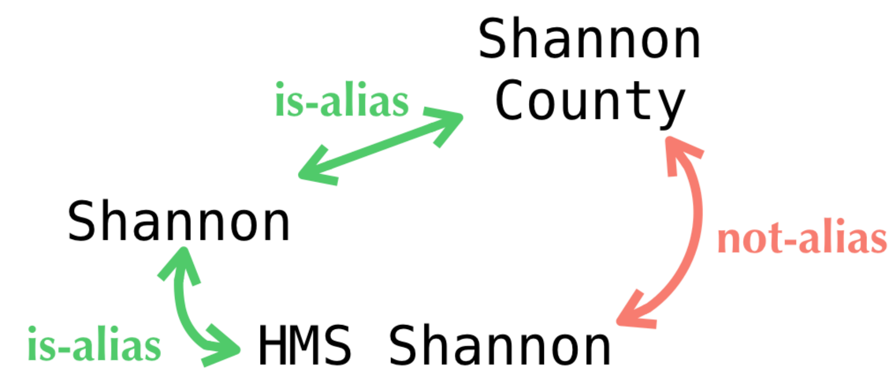



**Similarity Matrix - STANCE**



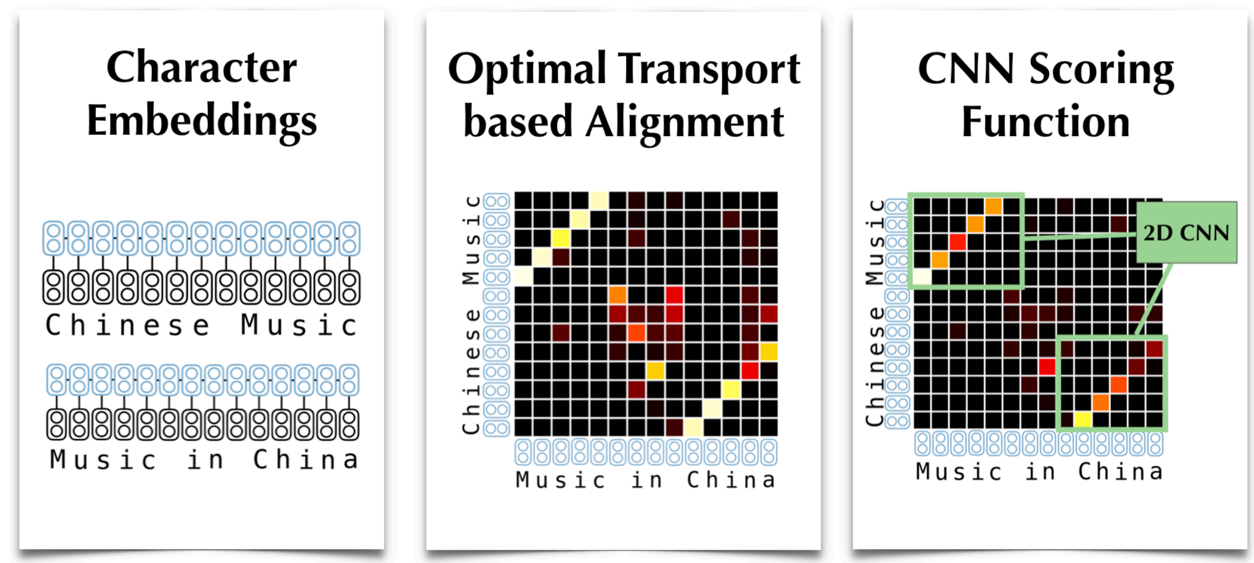
# Summary

## Learned String Similarity

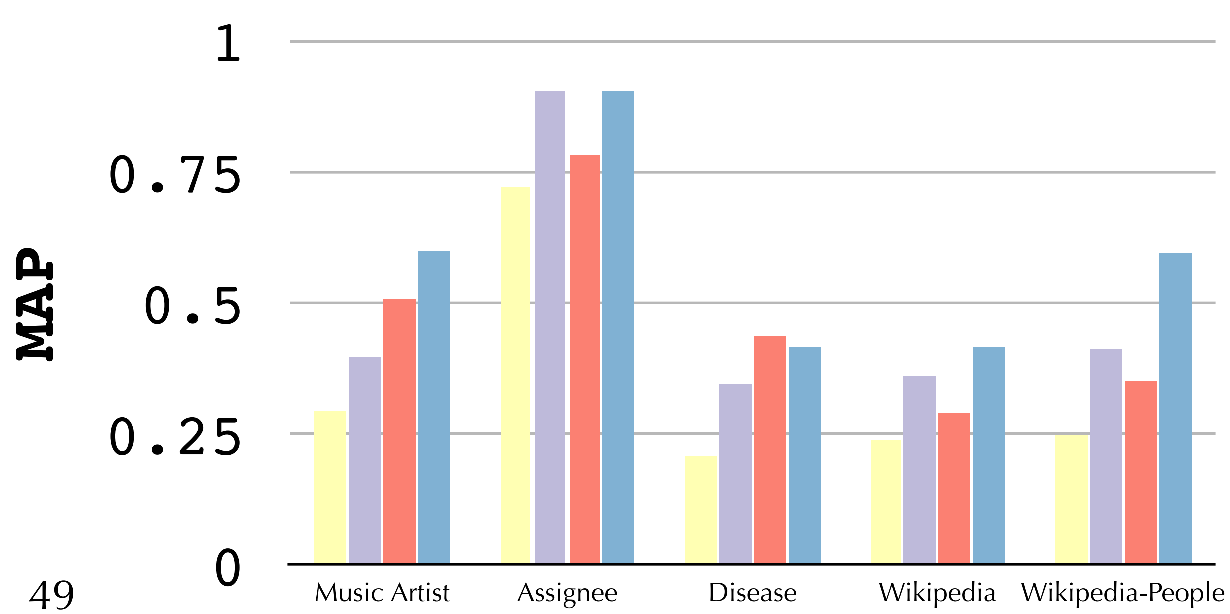


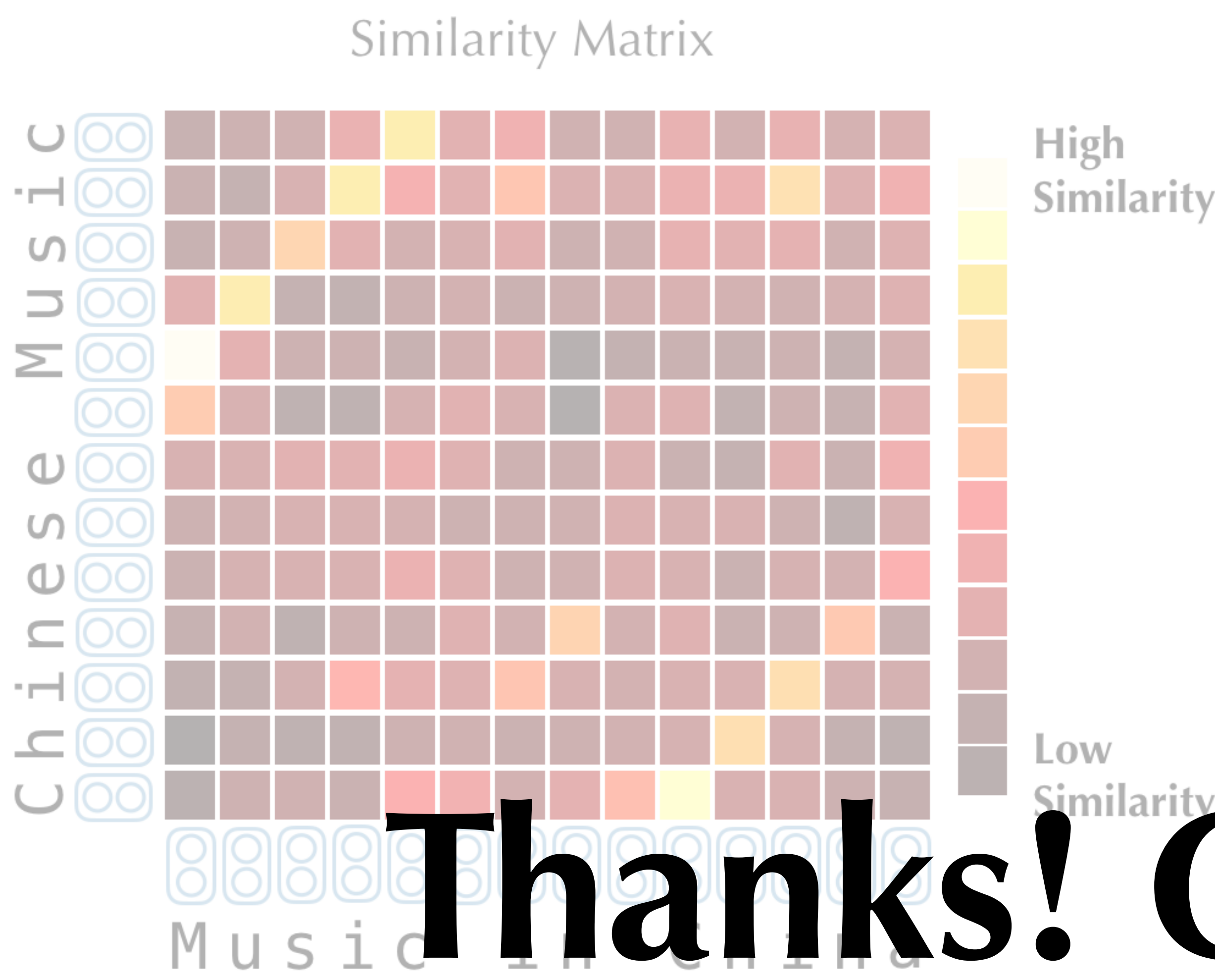

## STANCE

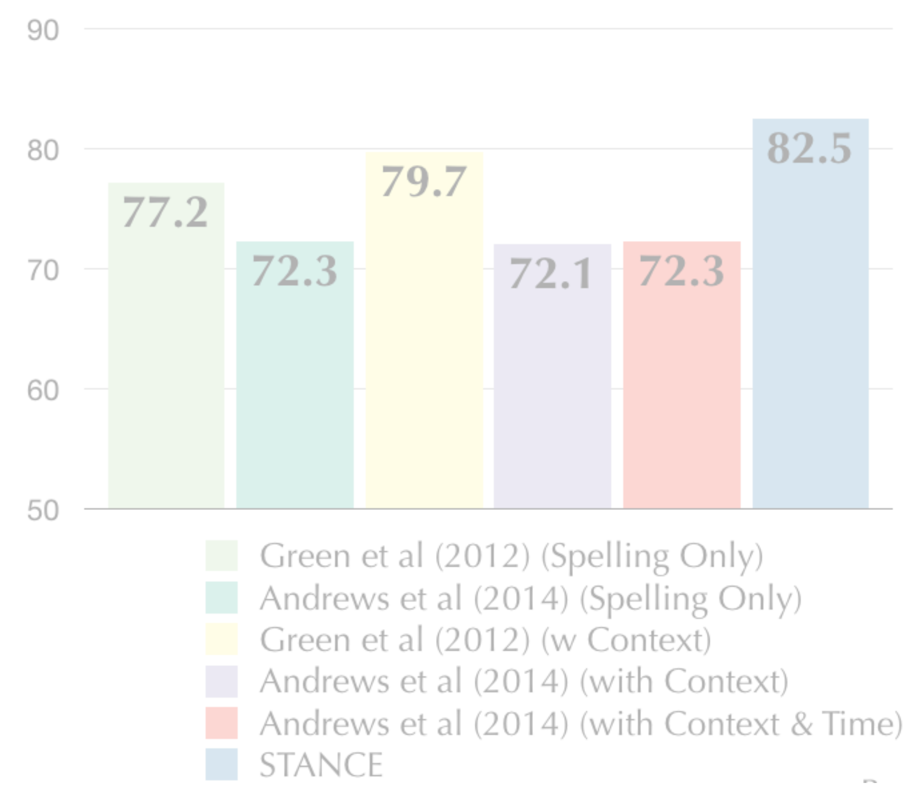
**S**imilarity of **T**ransport **A**ligned **N**eural **C**haracter **E**ncodings




## New Datasets and Results






# Thanks! Questions?



## Code: https://github.com/iesl/stance