Carnegie Mellon University
Language
Technologies
Institute

LLab
Language Technologies Institute
Carnegie Mellon University

Semisupervised Neural Proto-Language Reconstruction

Liang Lu[1], Peirong Xie[2], David R. Mortensen[1]

[1]Carnegie Mellon University, [2]University of Southern California

lianglu@cs.cmu.edu, louisxie@usc.edu, dmortens@cs.cmu.edu

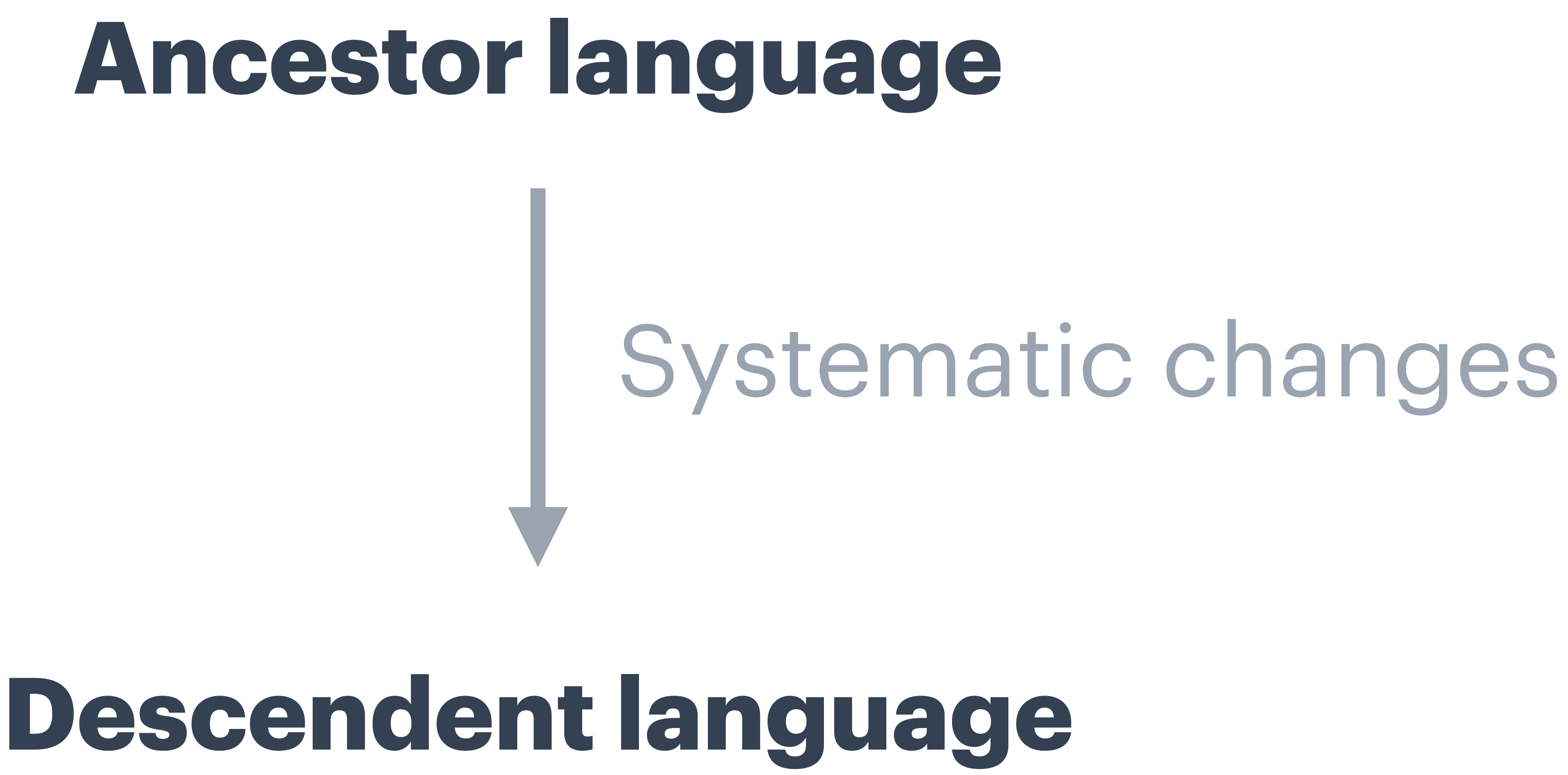
ACL 2024 Bangkok, Thailand

Background

A 19th-century Discovery

Languages change in systematic ways, and **it is possible to reproducibly reconstruct proto-languages** using these systematic patterns, even when no record of the proto-language survived.

Historical linguists use the **comparative method** to reconstruct proto-languages.

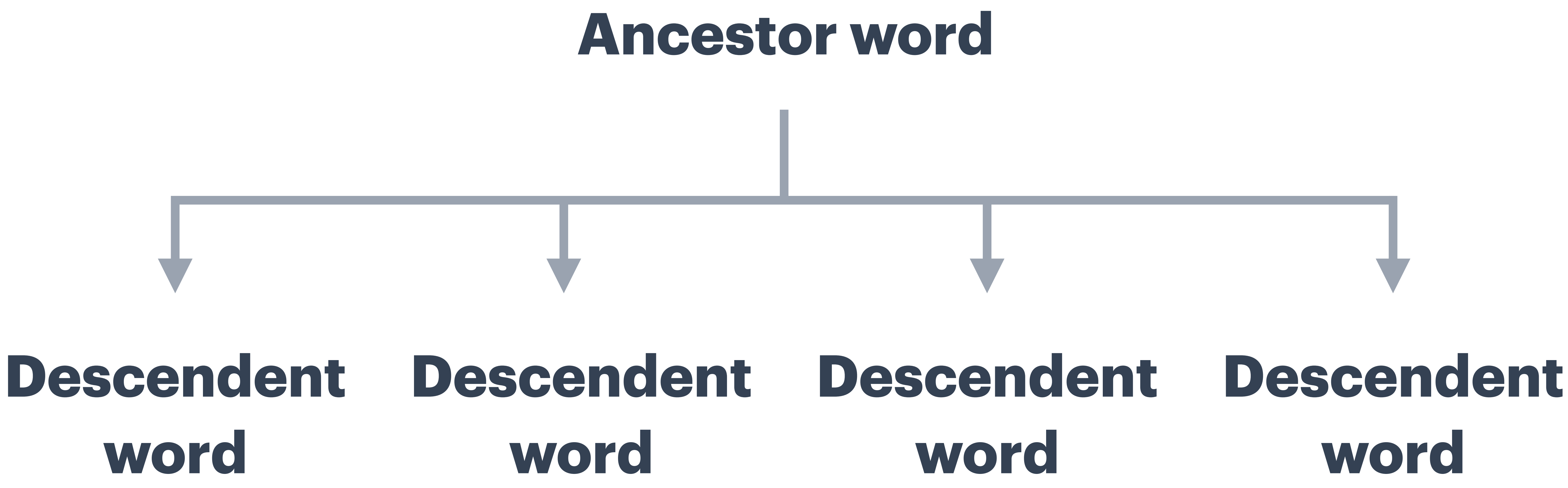


Ancestor language

Systematic changes

Descendent language

Protoform Reconstruction



Ancestor word

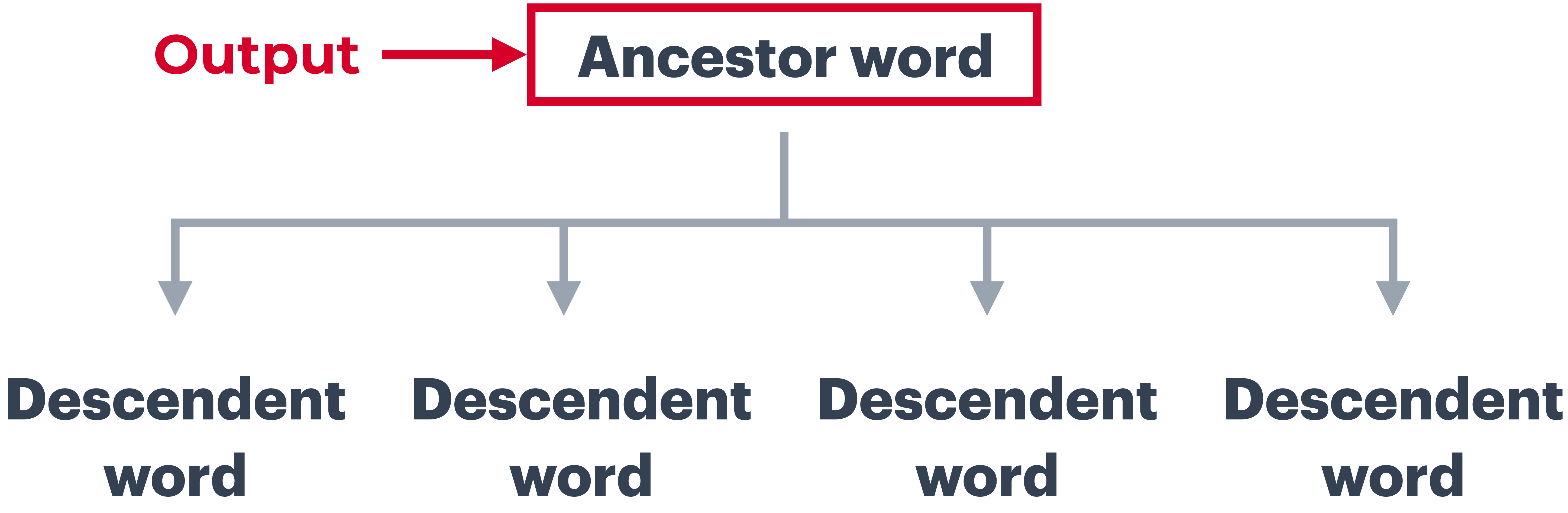
Descendent word

Descendent word

Descendent word

Descendent word

Protoform Reconstruction



Output

Ancestor word

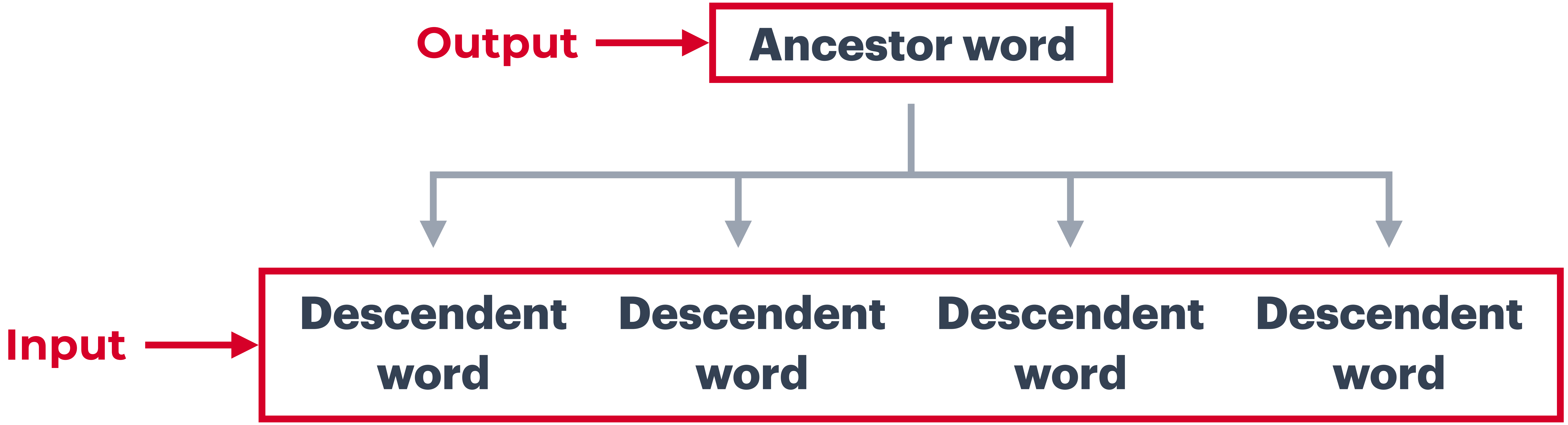
Descendent word

Descendent word

Descendent word

Descendent word

Protoform Reconstruction



Output

Ancestor word

Input

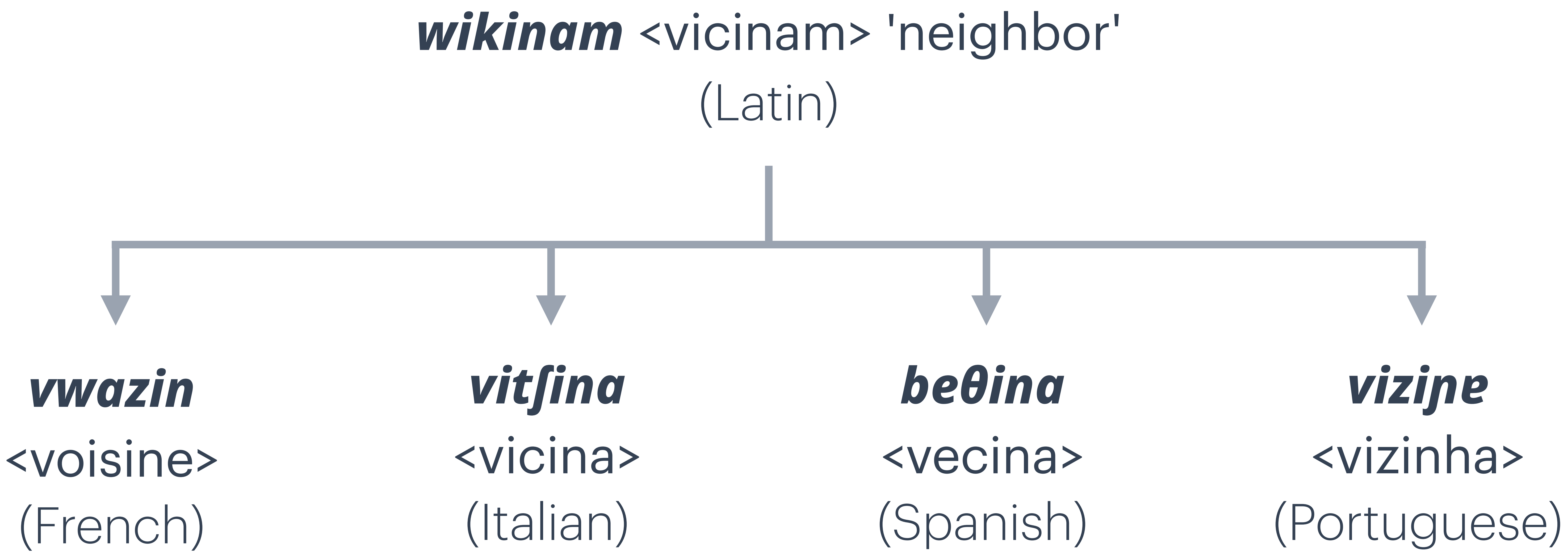
Descendent word

Descendent word

Descendent word

Descendent word

Protoform Reconstruction



wikinam <vicinam> 'neighbor'
(Latin)

vwazin
<voisine>
(French)

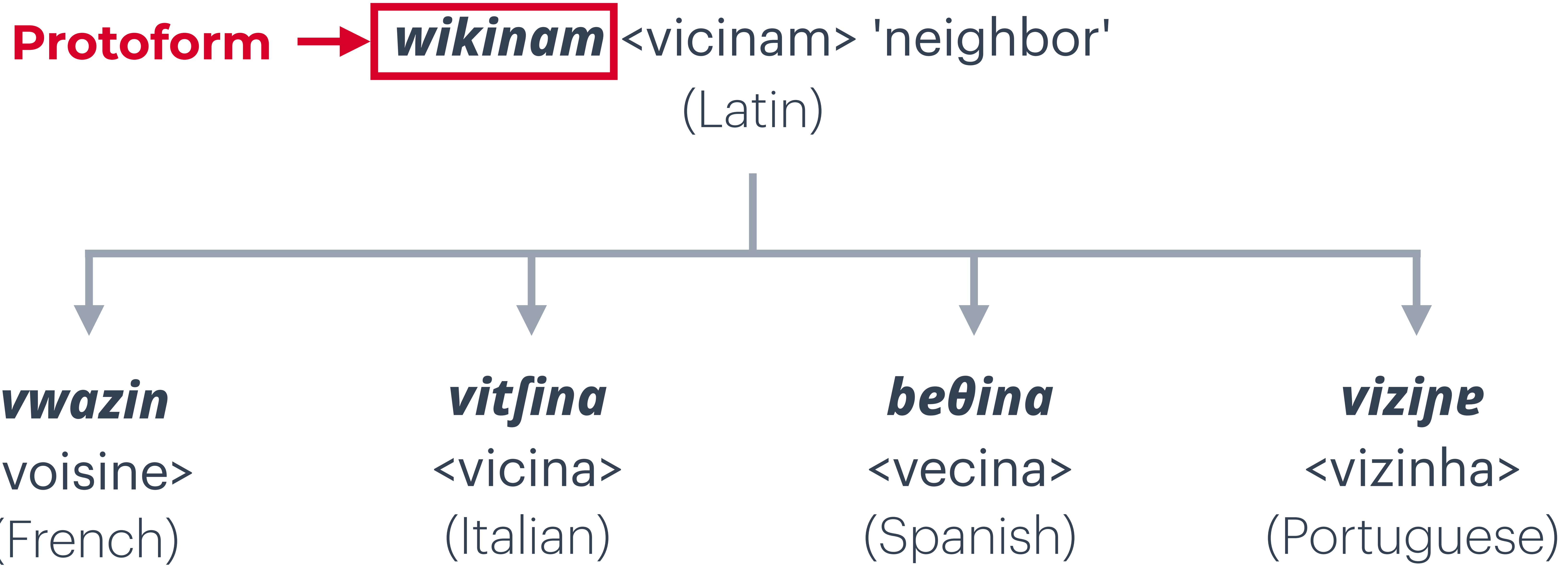
vitʃina
<vicina>
(Italian)

beθina
<vecina>
(Spanish)

viziɲɐ
<vizinha>
(Portuguese)

Example: Romance Dataset (Meloni et al., 2021; Ciobanu and Dinu, 2018)

Protoform Reconstruction



Protoform → ***wikinam*** <vicinam> 'neighbor'
(Latin)

vwazin
<voisine>
(French)

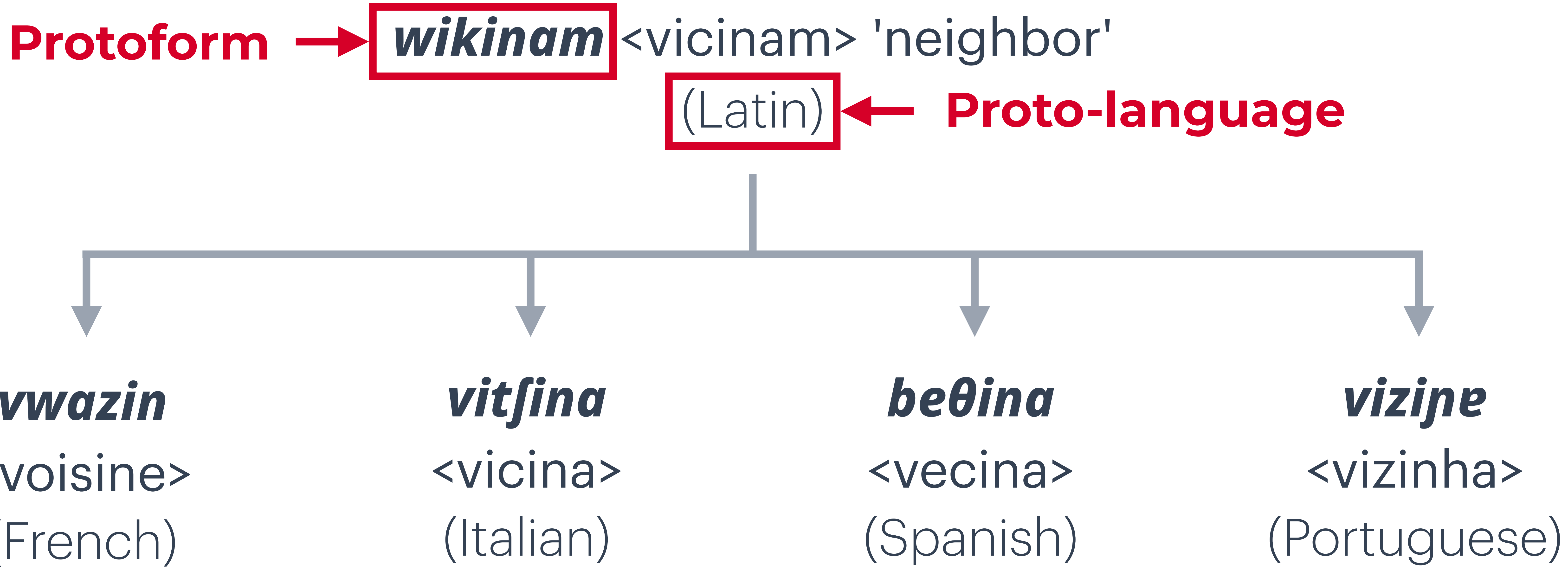
vitʃina
<vicina>
(Italian)

beθina
<vecina>
(Spanish)

viziɲe
<vizinha>
(Portuguese)

Example: Romance Dataset (Meloni et al., 2021; Ciobanu and Dinu, 2018)

Protoform Reconstruction



Protoform → ***wikinam*** <vicinam> 'neighbor'

(Latin) ← **Proto-language**

vwazin <voisine> (French)

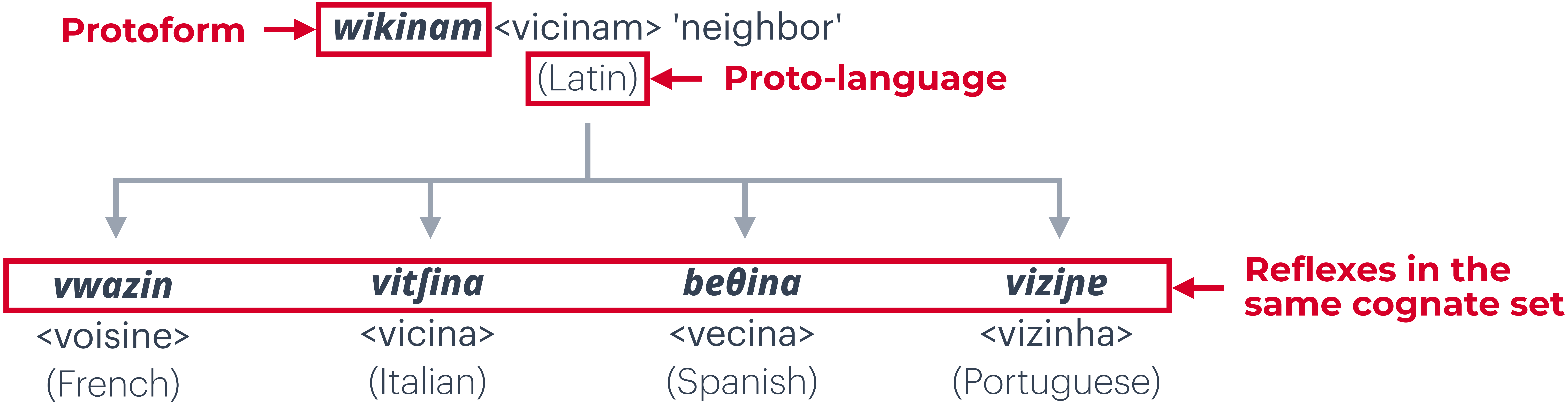
vitʃina <vicina> (Italian)

beθina <vecina> (Spanish)

viziɲe <vizinha> (Portuguese)

Example: Romance Dataset (Meloni et al., 2021; Ciobanu and Dinu, 2018)

Protoform Reconstruction



Protoform → ***wikinam*** <vicinam> 'neighbor'

(Latin) ← Proto-language

vwazin <voisine> (French)

vitʃina <vicina> (Italian)

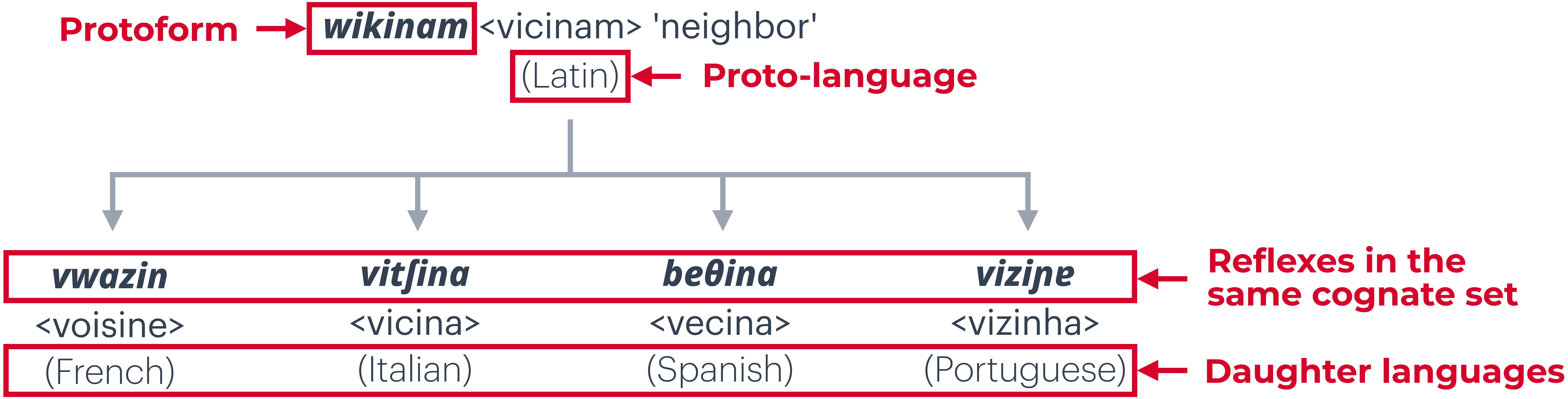
beθina <vecina> (Spanish)

viziɲe <vizinha> (Portuguese)

← Reflexes in the same cognate set

Example: Romance Dataset (Meloni et al., 2021; Ciobanu and Dinu, 2018)

Protoform Reconstruction



Protoform → ***wikinam*** <vicinam> 'neighbor'

(Latin) ← Proto-language

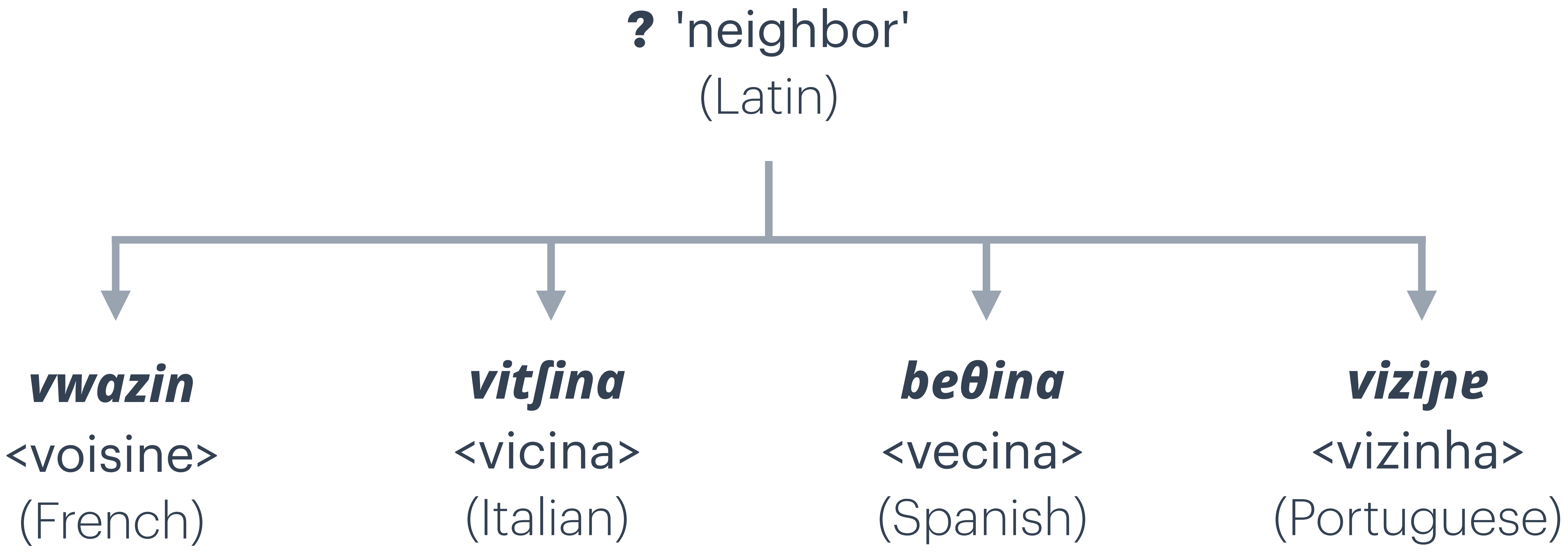
vwazin	***vitʃina***	***beθina***	***viziɲe***
<voisine>	<vicina>	<vecina>	<vizinha>
(French)	(Italian)	(Spanish)	(Portuguese)

← Reflexes in the same cognate set

← Daughter languages

Example: Romance Dataset (Meloni et al., 2021; Ciobanu and Dinu, 2018)

Protoform Reconstruction



? 'neighbor'
(Latin)

vwazin
<voisine>
(French)

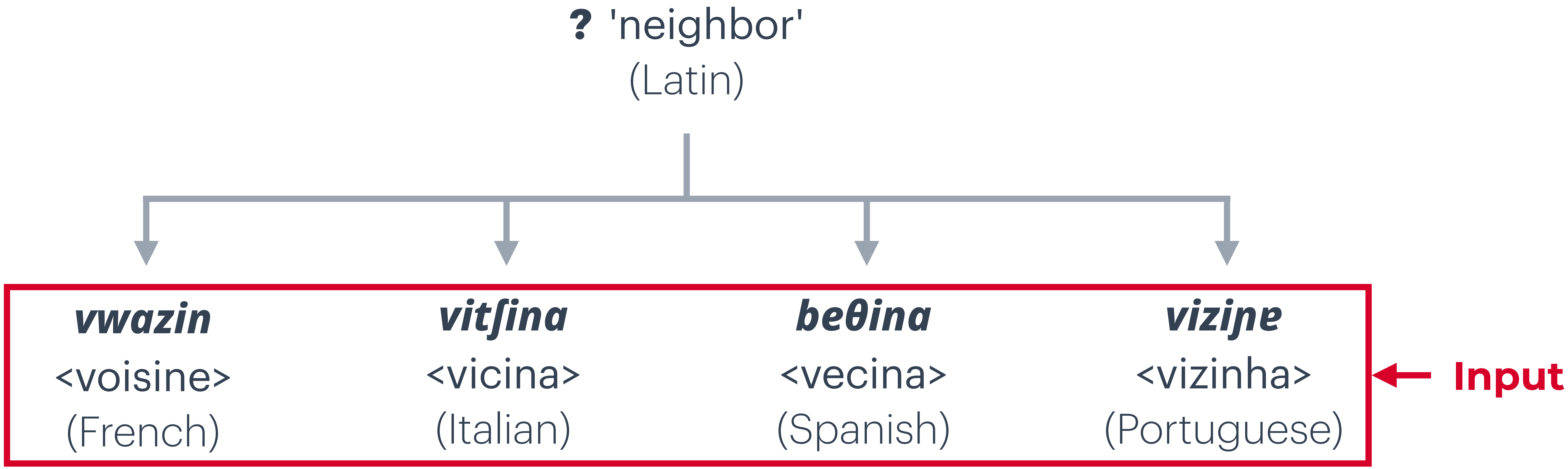
vitʃina
<vicina>
(Italian)

beθina
<vecina>
(Spanish)

viziɲɐ
<vizinha>
(Portuguese)

Example: Romance Dataset (Meloni et al., 2021; Ciobanu and Dinu, 2018)

Protoform Reconstruction



? 'neighbor'
(Latin)

vwazin
<voisine>
(French)

vitʃina
<vicina>
(Italian)

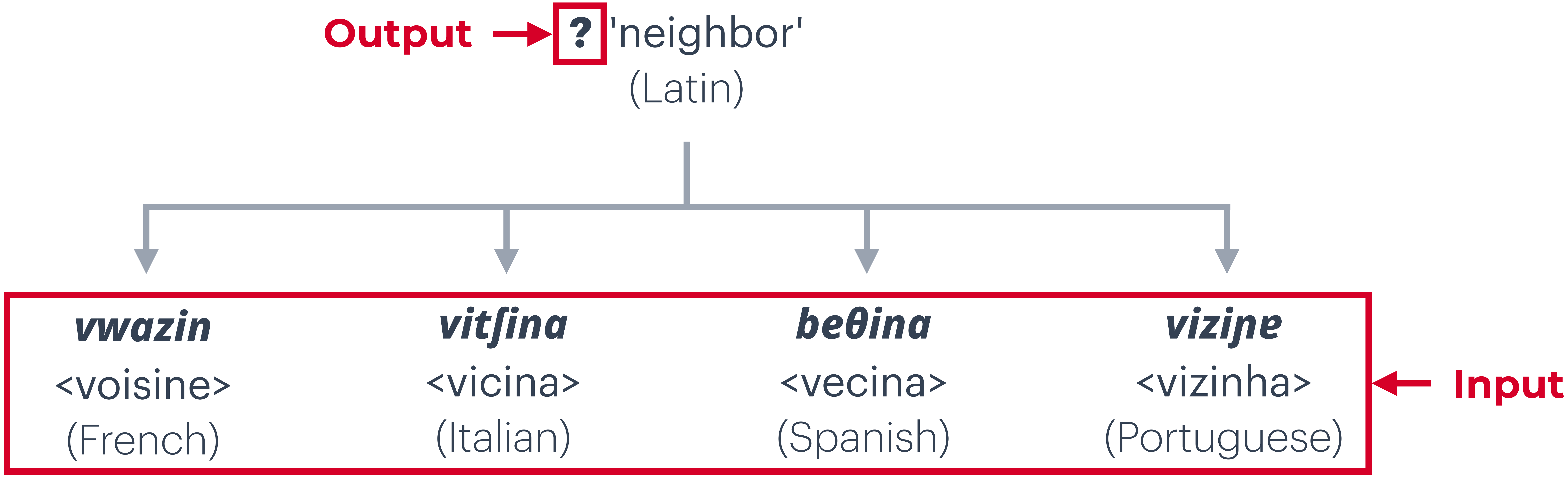
beθina
<vecina>
(Spanish)

viziɲe
<vizinha>
(Portuguese)

Input

Example: Romance Dataset (Meloni et al., 2021; Ciobanu and Dinu, 2018)

Protoform Reconstruction



Output → **?** 'neighbor'
(Latin)

vwazin
<voisine>
(French)

vitʃina
<vicina>
(Italian)

beθina
<vecina>
(Spanish)

viziɲe
<vizinha>
(Portuguese)

← Input

Example: Romance Dataset (Meloni et al., 2021; Ciobanu and Dinu, 2018)

The Comparative Method

The regularity principle:

- Sound changes are regular
- Reflexes should be derivable deterministically from reconstructions using a single set of sound change rules

> "Every sound change, in so far as it proceeds mechanically, is completed in accordance with laws admitting of no exceptions; i.e. the direction in which the change takes place is always the same for all members of a language community, apart from the case of dialect division, and all words in which the sound subject to change occurs in the same conditions are affected by the change without exception."
>
> —H. Osthoff and K. Brugmann, *Morphologische Untersuchungen auf dem Gebiete der indogermanischen Sprachen* i, Leipzig, 1878, p. xiii (quoted in Szemerényi (1996))

The Comparative Method

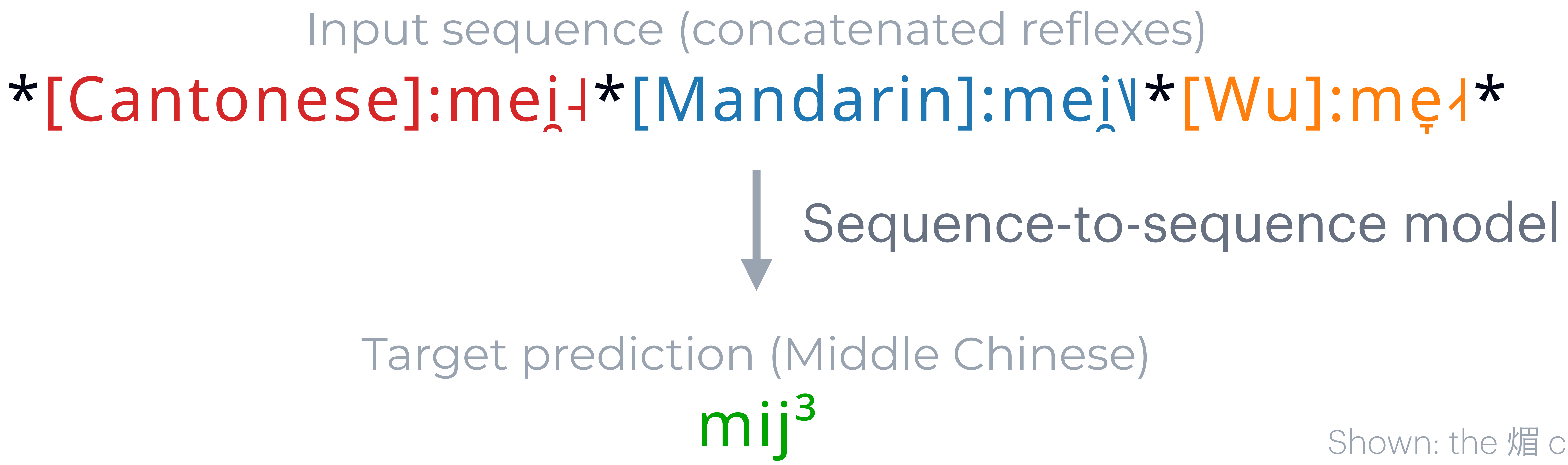
The regularity principle:

- Sound changes are regular
- Reflexes should be derivable deterministically from reconstructions using a single set of sound change rules

The comparative method is **challenging to apply in practice**, because examining a large number of cognate sets and complex combinations of sound changes can impose heavy cognitive load.

Supervised Neural Reconstruction

- **RNN with language embedding** (Meloni et al., 2021)
- **Transformer** (Kim et al., 2023)
- **VAE** (Cui et al., 2022)



Input sequence (concatenated reflexes)

[Cantonese]:mei̯˨[Mandarin]:mei̯˥˩*[Wu]:me̝˨˧*

Sequence-to-sequence model

Target prediction (Middle Chinese)

mij³

Shown: the 媚 cognate set from WikiHan

Note: Other input representations exist, such a stacked representation used by Cognate Transformer (Akavarapu and Bhattacharya, 2023)

Supervised Training

Cantonese	Gan	Hakka	Jin	Mandarin	Hokkien	Wu	Xiang	Label (Gold Protoform)
pʰʊŋ˨˩	pʰuŋ˨˦	pʰuŋ˩	pʰxə̃ŋ˩˩	pʰɤŋ˧˥	pʰaŋ˨˦	b̥ʊŋ˨˧	pʊŋ˩˧	$buŋ^{w1}$
mɔː˧˥	mo˥˧	mi̯a˨˦	-	mu̯ɔ˥	bɔŋ˥	mʊʔ˩˨	-	mak^4
saːn˧	-	-	-	ʂan˥˩	sũ̯ã˧˩	-	-	$ʂɛn^2$
kʰøy̯˥	-	-	-	t͡ɕʰy˥	kʰi̯ɤʔ˧	-	-	$kʰi^1$
siːn˧˥	-	ɕi̯en˧˩	-	ɕyan˨˩	ɕi̯ɛn˥˩	-	-	$sjen^2$
lɐu̯˩˧	-	li̯u˨˦	li̯əu̯˥˧	li̯oʊ̯˨˩	li̯u˥˩	li̯ɜ˨˧	-	$ljuw^2$
mɐn˧	-	-	-	u̯ən˥˩	bun˧	-	-	$mjun^3$
tʊŋ˧˥	tuŋ˨˩˦	tuŋ˧˩	tũŋ˥˧	tʊŋ˨˩	tɔŋ˥˩	tʊŋ˧˥	tʊŋ˥˩	$tuŋ^{w2}$
t͡sʰɐu̯˨˩	-	-	-	t͡ʂʰoʊ̯˧˥	ɕi̯u˨˦	-	-	$d͡ʑuw^1$
jœːŋ˥	-	-	-	i̯ɑŋ˥	i̯ɔŋ˥	-	-	$ʔjaŋ^1$

Examples are from WikiHan, '-' indicates missing reflex in the dataset

A More Realistic Scenario: Semisupervised Reconstruction

Cantonese	Gan	Hakka	Jin	Mandarin	Hokkien	Wu	Xiang	Label (Gold Protoform)
pʰʊŋ˨˩	pʰuŋ˨˦	pʰuŋ˩	pʰxə̃ŋ˩˩	pʰɤŋ˧˥	pʰaŋ˨˦	b̥ʊŋ˨˧	pʊŋ˩˧	(unavailable)
mɔː˧˥	mo˥˧	mi̯a˨˦	-	mu̯ɔ˥	bɔŋ˥	mʊʔ˩˨	-	(unavailable)
saːn˧	-	-	-	ʂan˥˩	sũ̯ã˧˩	-	-	(unavailable)
kʰɵy̯˥	-	-	-	t͡ɕʰy˥	kʰi̯ɤʔ˧	-	-	kʰi^{1}
siːn˧˥	-	ɕi̯en˧˩	-	ɕyan˨˩	ɕi̯ɛn˥˩	-	-	(unavailable)
lɐu̯˩˧	-	li̯u˨˦	li̯əu̯˥˧	li̯oʊ̯˨˩	li̯u˥˩	li̯ɜ˨˧	-	ljuw^{2}
mɐn˧	-	-	-	u̯ən˥˩	bun˧	-	-	(unavailable)
tʊŋ˧˥	tuŋ˩˧	tuŋ˧˩	tũŋ˥˧	tʊŋ˨˩	tɔŋ˥˩	tʊŋ˧˥	tʊŋ˥˩	tuŋʷ^{2}
t͡sʰɐu̯˨˩	-	-	-	t͡ʂʰoʊ̯˧˥	ɕi̯u˨˦	-	-	(unavailable)
jœːŋ˥	-	-	-	i̯ɑŋ˥	i̯ɔŋ˥	-	-	(unavailable)

Examples are from WikiHan, '-' indicates missing reflex in the dataset

A More Realistic Scenario: Semisupervised Reconstruction

Cantonese	Gan	Hakka	Jin	Mandarin	Hokkien	Wu	Xiang	Label (Gold Protoform)
pʰʊŋ˨˩	pʰuŋ˨˦	pʰuŋ˩	pʰxə̃ŋ˩˩	pʰɤŋ˧˥	pʰaŋ˨˦	b̥ʊŋ˨˧	pʊŋ˩˧	(unavailable)
mɔː˥˥	mo˥˧	mi̯a˨˦	-	mu̯ɔ˥	bɔŋ˥	mʊʔ˩˨	-	(unavailable)
saːn˧	-	-	-	ʂan˥˩	sũ̯ã˥˧	-	-	(unavailable)
kʰɵy̯˥	-	-	-	t͡ɕʰy˥	kʰi̯ɤʔ˧	-	-	kʰi^1
siːn˥˥	-	ɕi̯en˥˧	-	ɕyan˨˩	ɕi̯ɛn˥˩	-	-	(unavailable)
lɐu̯˩˧	-	li̯u˨˦	li̯əu̯˥˧	li̯oʊ̯˨˩	li̯u˥˩	li̯ɜ˨˧	-	ljuw^2
mɐn˧	-	-	-	u̯ən˥˩	bun˧	-	-	(unavailable)
tʊŋ˥˥	tuŋ˩˧	tuŋ˥˧	tũŋ˥˧	tʊŋ˨˩	tɔŋ˥˩	tʊŋ˥˥	tʊŋ˥˩	tuŋ$^{\text{w}2}$
t͡sʰɐu̯˨˩	-	-	-	t͡ʂʰoʊ̯˥˥	ɕi̯u˨˦	-	-	(unavailable)
jœːŋ˥	-	-	-	i̯ɑŋ˥	i̯ɔŋ˥	-	-	(unavailable)

Examples are from WikiHan, '-' indicates missing reflex in the dataset

A Hypothetical Example

Gloss	'grandchild'	'bone'	'breast'	'laugh'
Labeled?	Yes	Yes	No	No
Kachai	ð ɐ	r ɐ	n ɐ	n i
Huishu	r u k	r u k	n u k	n u k
Ukhrul	r u	r u	n u	n u
Protoform Label	d u	r u	n u	n ɨ

A Hypothetical Example

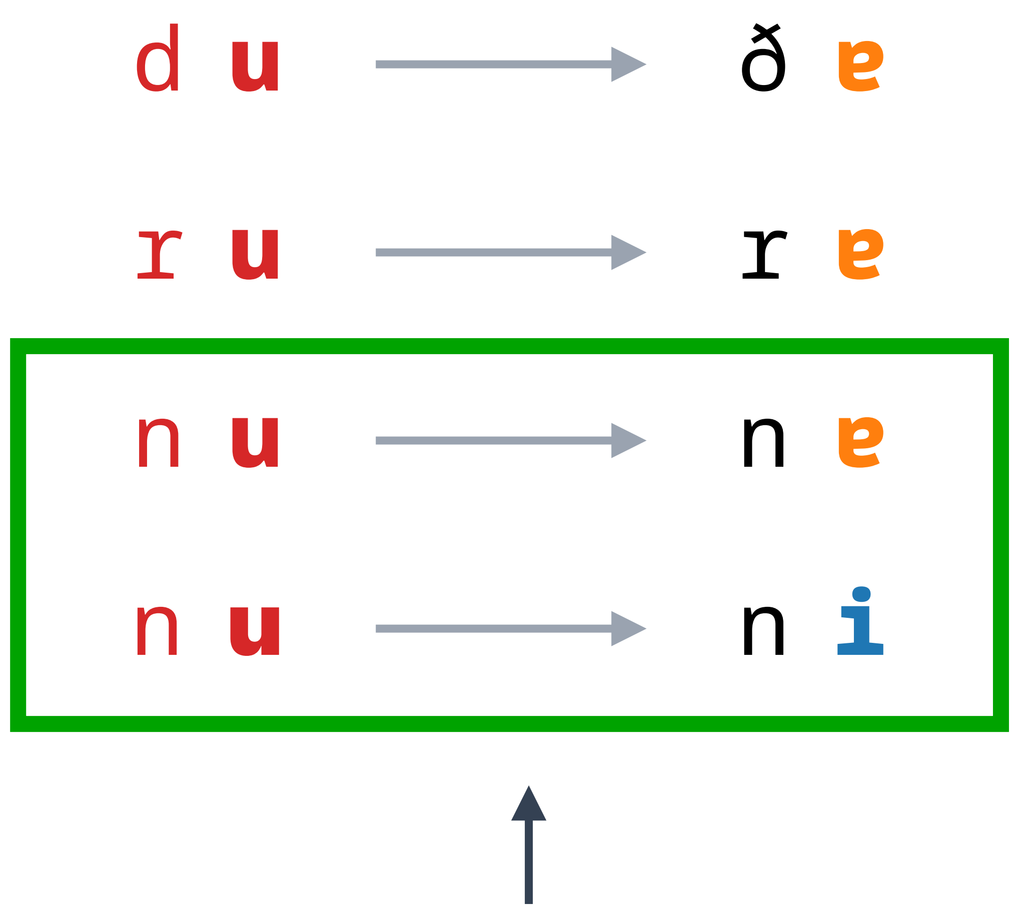
Gloss	'grandchild'	'bone'	'breast'	'laugh'
Labeled?	Yes	Yes	No	No
Kachai	ð ɐ	ɾ ɐ	n ɐ	n i
Huishu	ɾ u k	ɾ u k	n u k	n u k
Ukhrul	ɾ u	ɾ u	n u	n u
Protoform Label	d u	ɾ u	(hidden)	(hidden)

A Hypothetical Example

Gloss	'grandchild'	'bone'	'breast'	'laugh'
Labeled?	Yes	Yes	No	No
Kachai	ð ɐ	r ɐ	n ɐ	n i
Huishu	r **u** k	r **u** k	n **u** k	n **u** k
Ukhrul	r **u**	r **u**	n **u**	n **u**
Protoform Label	d **u**	r **u**	(hidden)	(hidden)
Supervised Model	d **u**	r **u**	n **u**	n **u**

A Hypothetical Example

Gloss	'grandchild'	'bone'	'breast'	'laugh'
Labeled?	Yes	Yes	No	No
Kachai	ð ɐ	ɾ ɐ	n ɐ	n i
Huishu	ɾ u k	ɾ u k	n u k	n u k
Ukhrul	ɾ u	ɾ u	n u	n u
Protoform Label	d u	ɾ u	(hidden)	(hidden)
Supervised Model	d u	ɾ u	n u	n u


d u → ð ɐ

ɾ u → ɾ ɐ

n u → n ɐ

n u → n i

Trouble: Cannot deterministically derive the reflexes!

A Hypothetical Example

Gloss	'grandchild'	'bone'	'breast'	'laugh'
Labeled?	Yes	Yes	No	No
Kachai	ð **ɐ**	r **ɐ**	n **ɐ**	n **i**
Huishu	r **u** k	r **u** k	n **u** k	n **u** k
Ukhrul	r **u**	r **u**	n **u**	n **u**
Protoform Label	d **u**	r **u**	(hidden)	(hidden)
Supervised Model	d **u**	r **u**	n **u**	n **u**
Semisupervised Model	d **u**	r **u**	n **u**	

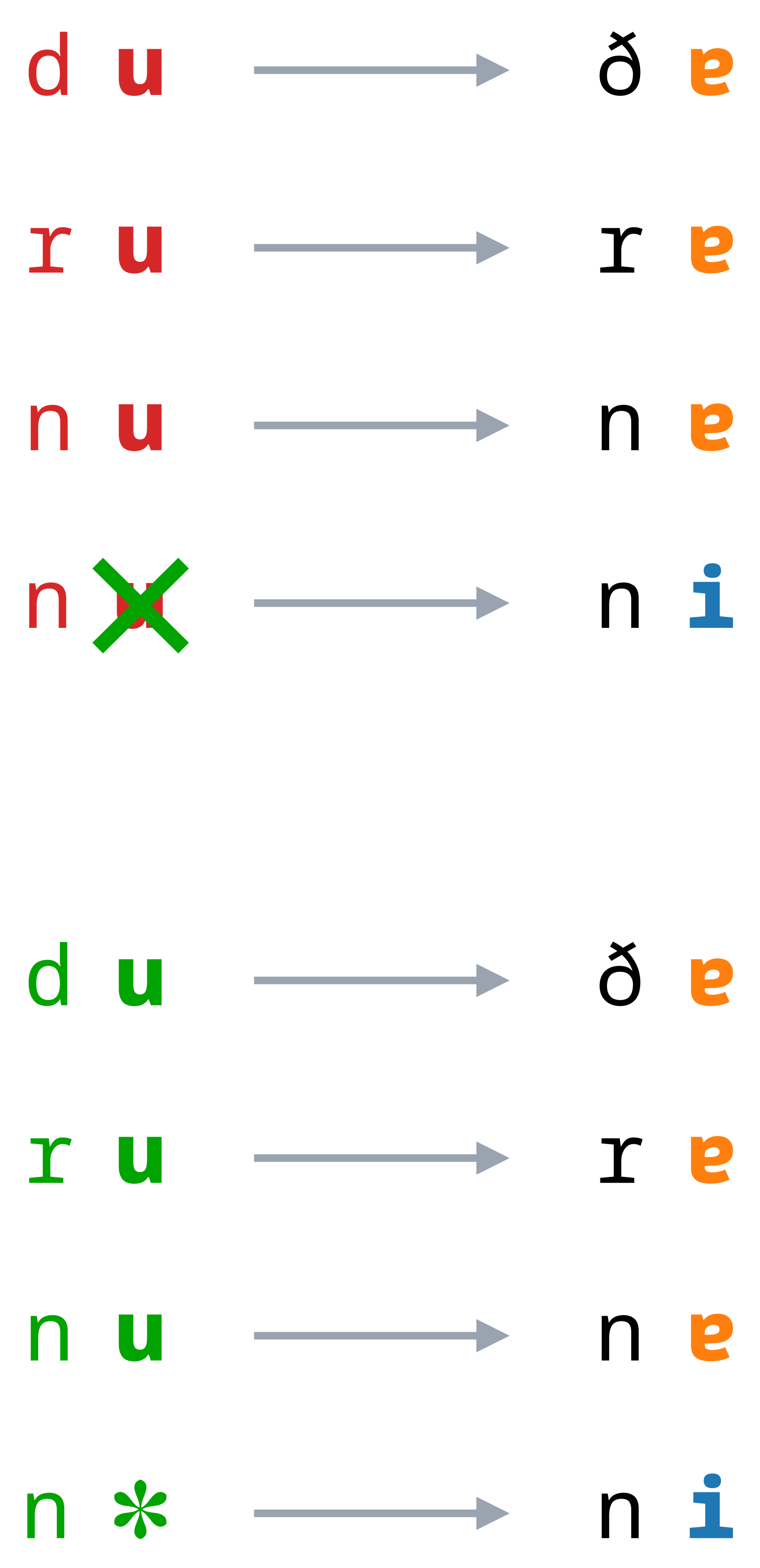
d **u** → ð **ɐ**
r **u** → r **ɐ**
n **u** → n **ɐ**
n **u** → n **i**

d **u** → ð **ɐ**
r **u** → r **ɐ**
n **u** → n **ɐ**

A Hypothetical Example

Gloss	'grandchild'	'bone'	'breast'	'laugh'
Labeled?	Yes	Yes	No	No
Kachai	ð **ɐ**	ɾ **ɐ**	n **ɐ**	n **i**
Huishu	ɾ **u** k	ɾ **u** k	n **u** k	n **u** k
Ukhrul	ɾ **u**	ɾ **u**	n **u**	n **u**
Protoform Label	d **u**	ɾ **u**	(hidden)	(hidden)
Supervised Model	d **u**	ɾ **u**	n **u**	n ~~**u**~~
Semisupervised Model	d **u**	ɾ **u**	n **u**	n *****

Something other than **u**



d **u** → ð **ɐ**

ɾ **u** → ɾ **ɐ**

n **u** → n **ɐ**

n ~~**u**~~ → n **i**

d **u** → ð **ɐ**

ɾ **u** → ɾ **ɐ**

n **u** → n **ɐ**

n ***** → n **i**

A Hypothetical Example

Gloss	'grandchild'	'bone'	'breast'	'laugh'
Labeled?	Yes	Yes	No	No
Kachai	ð ɐ	r ɐ	n ɐ	n i
Huishu	r u k	r u k	n u k	n u k
Ukhrul	r u	r u	n u	n u
Protoform Label	d u	r u	n u	n ɨ
Supervised Model	d u	r u	n u	n ~~u~~
Semisupervised Model	d u	r u	n u	n *

Indeed not **u**

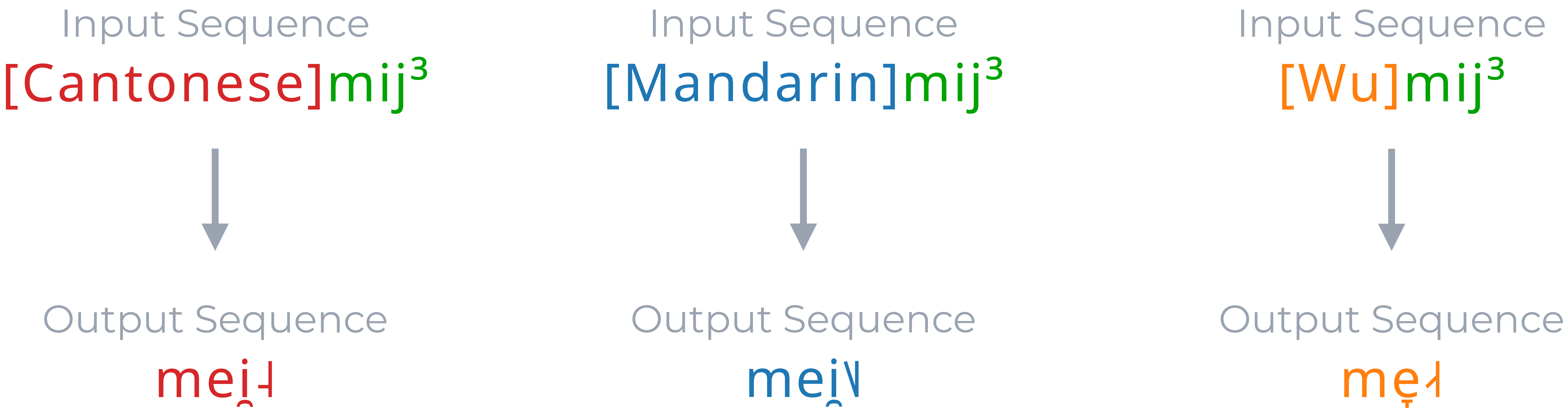
A Hypothetical Example

Gloss	'grandchild'	'bone'	'breast'	'laugh'
Labeled?	Yes	Yes	No	No
Kachai	ð **ɐ**	r **ɐ**	n **ɐ**	n **i**
Huishu	r **u** k	r **u** k	n **u** k	n **u** k
Ukhrul	r **u**	r **u**	n **u**	n **u**
Protoform Label	d **u**	r **u**	n **u**	n **ɨ**
DPD	d **u**	r **u**	n **u**	n **✻**

Reflexes $\xrightarrow{\textbf{Reconstruction}}$ Protoform? $\xrightarrow{\textbf{Reflex Prediction}}$ Reflexes?

Daughter-to-Proto-to-Daughter (DPD)

Reflex Prediction



Input Sequence

[Cantonese]mij^3

Output Sequence

mei̯˧

Input Sequence

[Mandarin]mij^3

Output Sequence

mei̯˨˩˦

Input Sequence

[Wu]mij^3

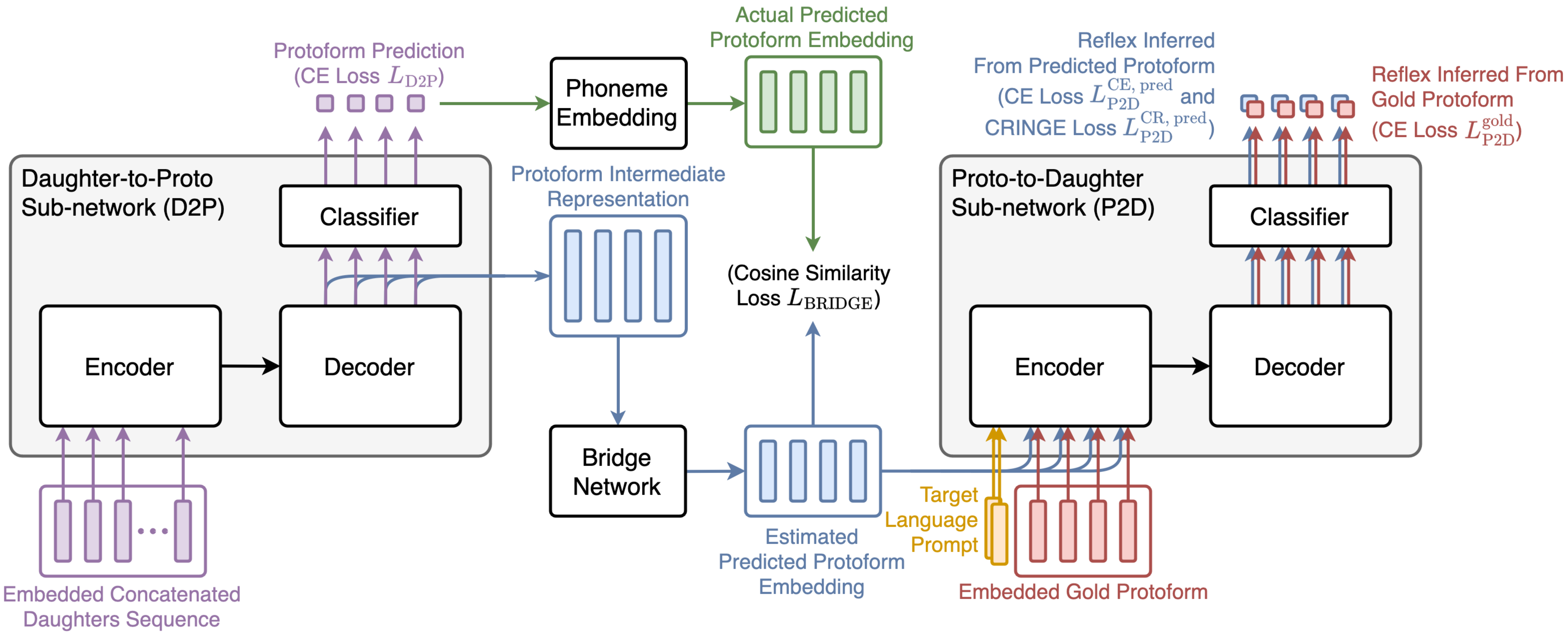
Output Sequence

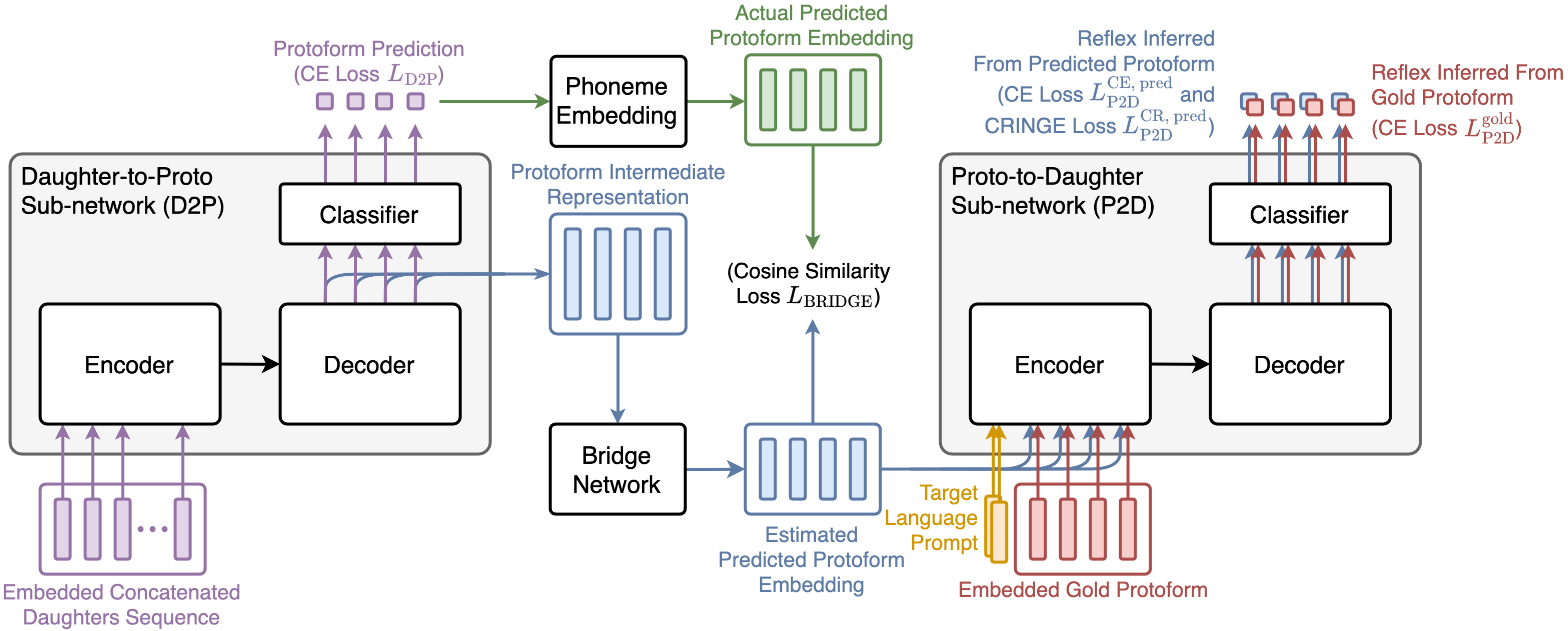
me̞˧˥

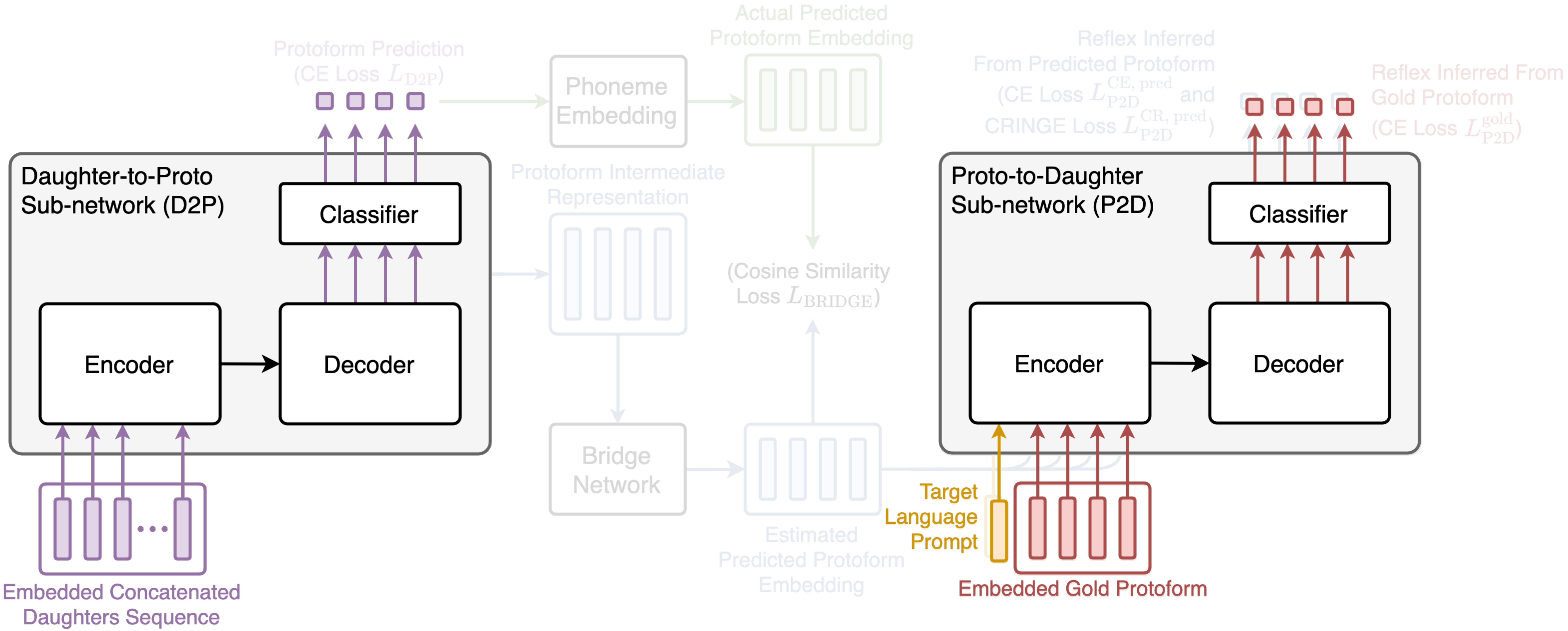
Methods

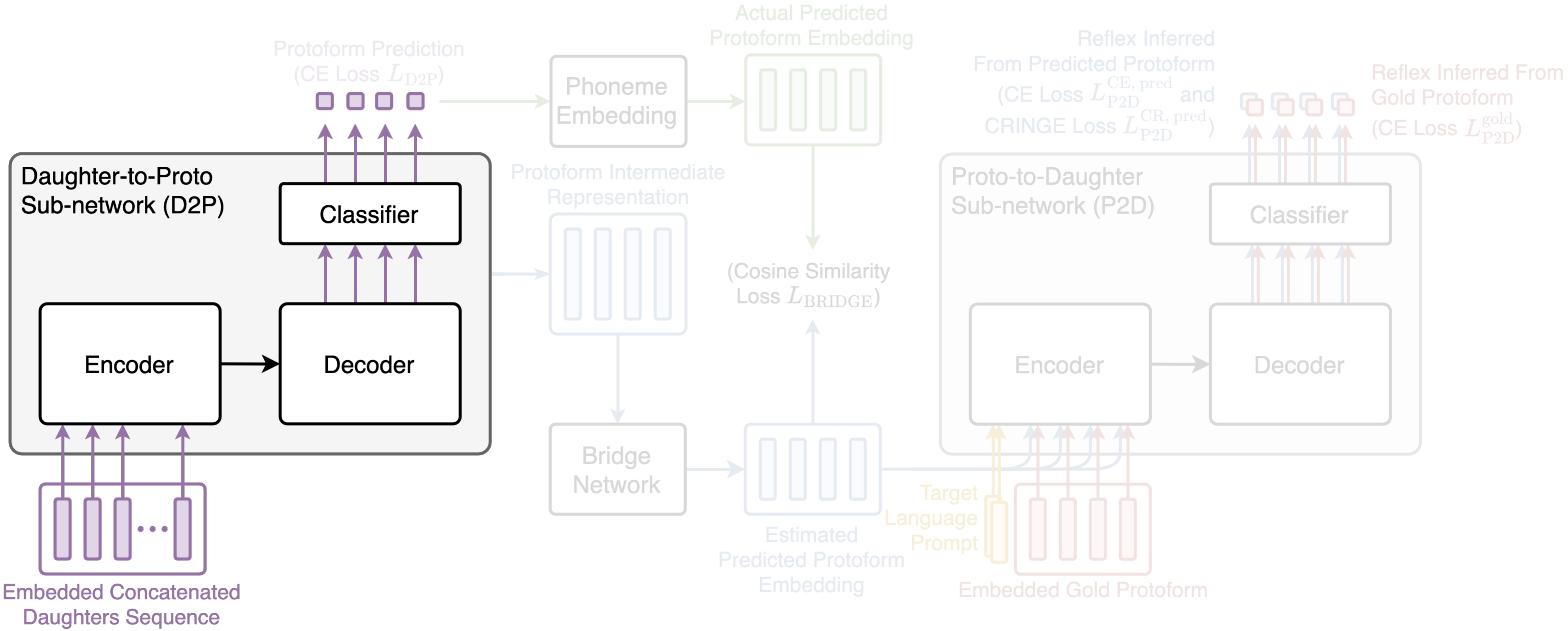
The DPD Architecture and the Experiments

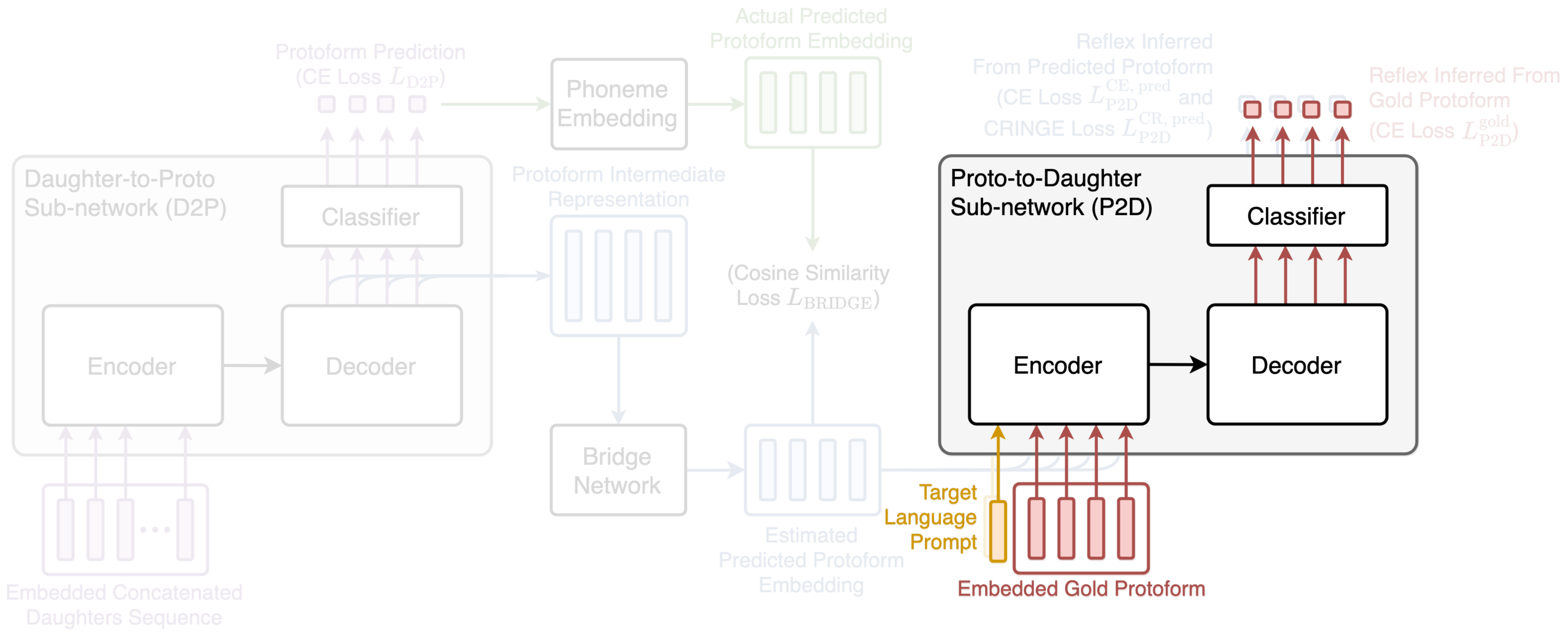
The DPD (Dughter-to-Proto-to-Daughter) Architecture


Protoform Prediction
(CE Loss L_{D2P})
Phoneme Embedding
Actual Predicted Protoform Embedding
Reflex Inferred From Predicted Protoform (CE Loss $L_{\mathrm{P2D}}^{\mathrm{CE,\ pred}}$ and CRINGE Loss $L_{\mathrm{P2D}}^{\mathrm{CR,\ pred}}$)
Reflex Inferred From Gold Protoform (CE Loss $L_{\mathrm{P2D}}^{\mathrm{gold}}$)
Daughter-to-Proto Sub-network (D2P)
Classifier
Protoform Intermediate Representation
Proto-to-Daughter Sub-network (P2D)
Classifier
(Cosine Similarity Loss L_{BRIDGE})
Encoder
Decoder
Encoder
Decoder
Bridge Network
...
Target Language Prompt
Embedded Concatenated Daughters Sequence
Estimated Predicted Protoform Embedding
Embedded Gold Protoform

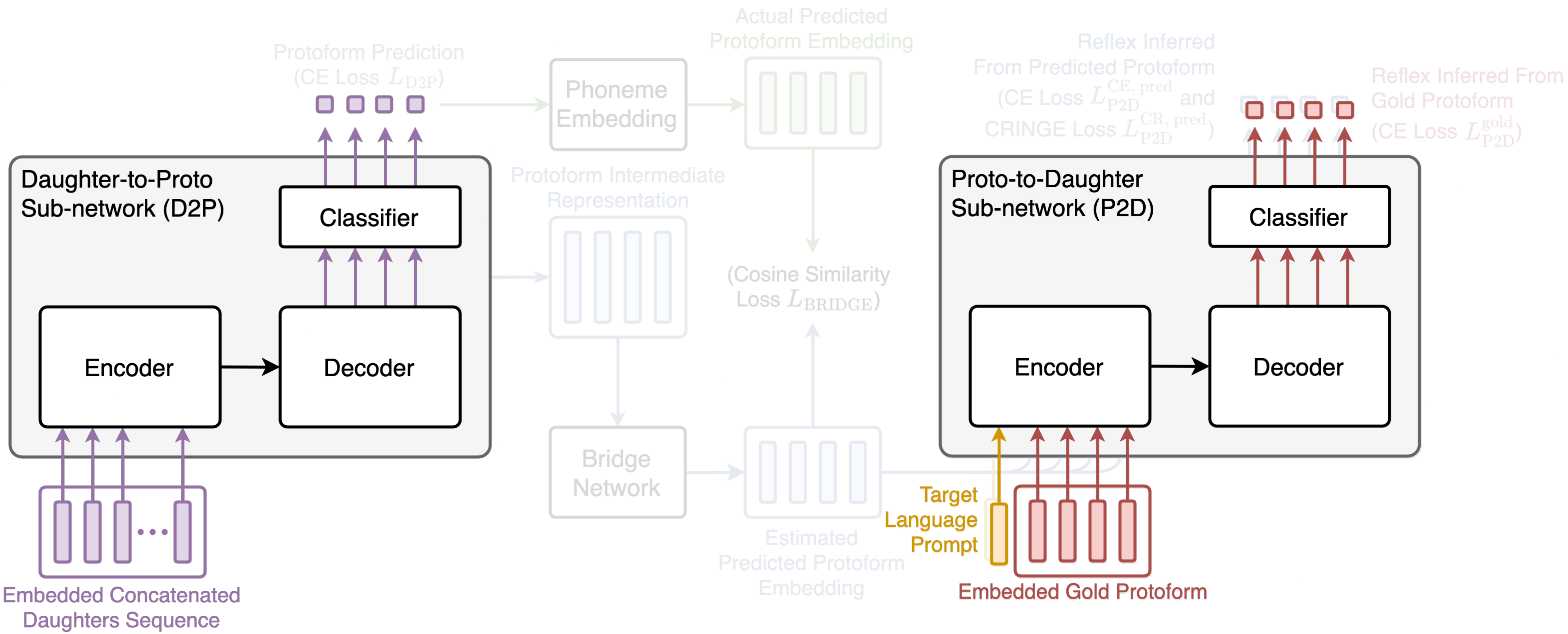

Protoform Prediction
(CE Loss L_{D2P})
Phoneme Embedding
Actual Predicted Protoform Embedding
Reflex Inferred From Predicted Protoform
(CE Loss $L_{\text{P2D}}^{\text{CE, pred}}$ and CRINGE Loss $L_{\text{P2D}}^{\text{CR, pred}}$)
Reflex Inferred From Gold Protoform
(CE Loss $L_{\text{P2D}}^{\text{gold}}$)
Daughter-to-Proto Sub-network (D2P)
Classifier
Encoder
Decoder
Protoform Intermediate Representation
(Cosine Similarity Loss L_{BRIDGE})
Proto-to-Daughter Sub-network (P2D)
Classifier
Encoder
Decoder
Bridge Network
Estimated Predicted Protoform Embedding
Target Language Prompt
Embedded Gold Protoform
Embedded Concatenated Daughters Sequence

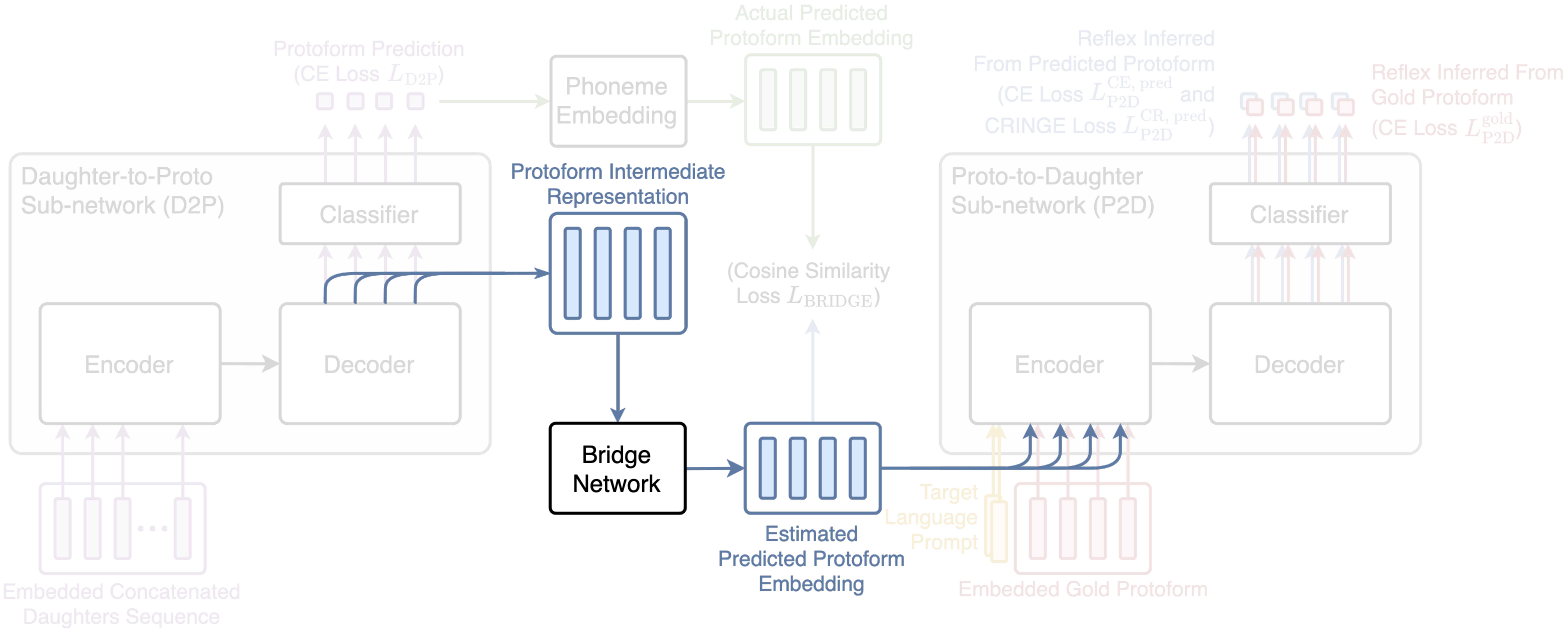

Protoform Prediction
(CE Loss L_{D2P})
Phoneme Embedding
Actual Predicted
Protoform Embedding
Reflex Inferred
From Predicted Protoform
(CE Loss $L_{\mathrm{P2D}}^{\mathrm{CE,\ pred}}$ and
CRINGE Loss $L_{\mathrm{P2D}}^{\mathrm{CR,\ pred}}$)
Reflex Inferred From
Gold Protoform
(CE Loss $L_{\mathrm{P2D}}^{\mathrm{gold}}$)
Daughter-to-Proto
Sub-network (D2P)
Classifier
Encoder
Decoder
Protoform Intermediate
Representation
(Cosine Similarity
Loss L_{BRIDGE})
Proto-to-Daughter
Sub-network (P2D)
Classifier
Encoder
Decoder
Bridge
Network
Estimated
Predicted Protoform
Embedding
Target
Language
Prompt
Embedded Gold Protoform
Embedded Concatenated
Daughters Sequence

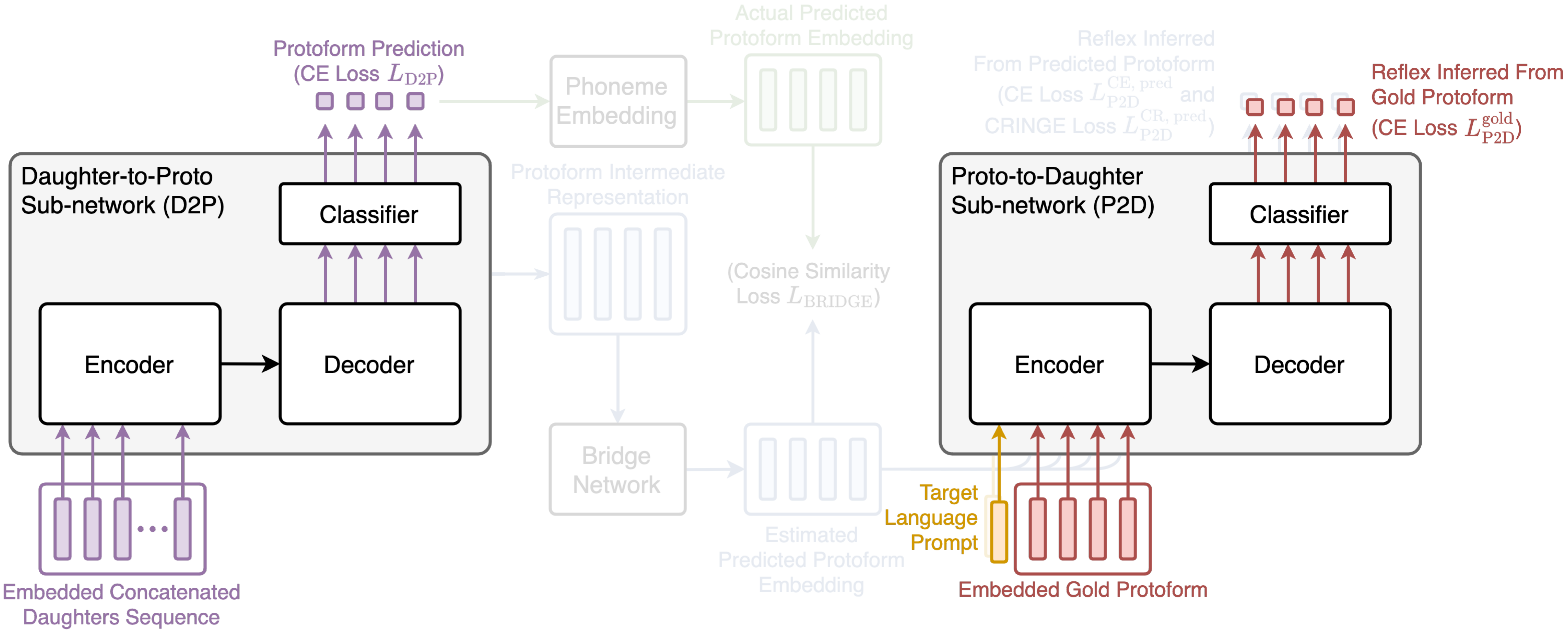

Protoform Prediction
(CE Loss L_{D2P})
Phoneme Embedding
Actual Predicted Protoform Embedding
Reflex Inferred From Predicted Protoform (CE Loss $L_{\text{P2D}}^{\text{CE, pred}}$ and CRINGE Loss $L_{\text{P2D}}^{\text{CR, pred}}$)
Reflex Inferred From Gold Protoform (CE Loss $L_{\text{P2D}}^{\text{gold}}$)
Daughter-to-Proto Sub-network (D2P)
Classifier
Encoder
Decoder
Protoform Intermediate Representation
(Cosine Similarity Loss L_{BRIDGE})
Proto-to-Daughter Sub-network (P2D)
Classifier
Encoder
Decoder
Bridge Network
Target Language Prompt
Estimated Predicted Protoform Embedding
Embedded Gold Protoform
Embedded Concatenated Daughters Sequence

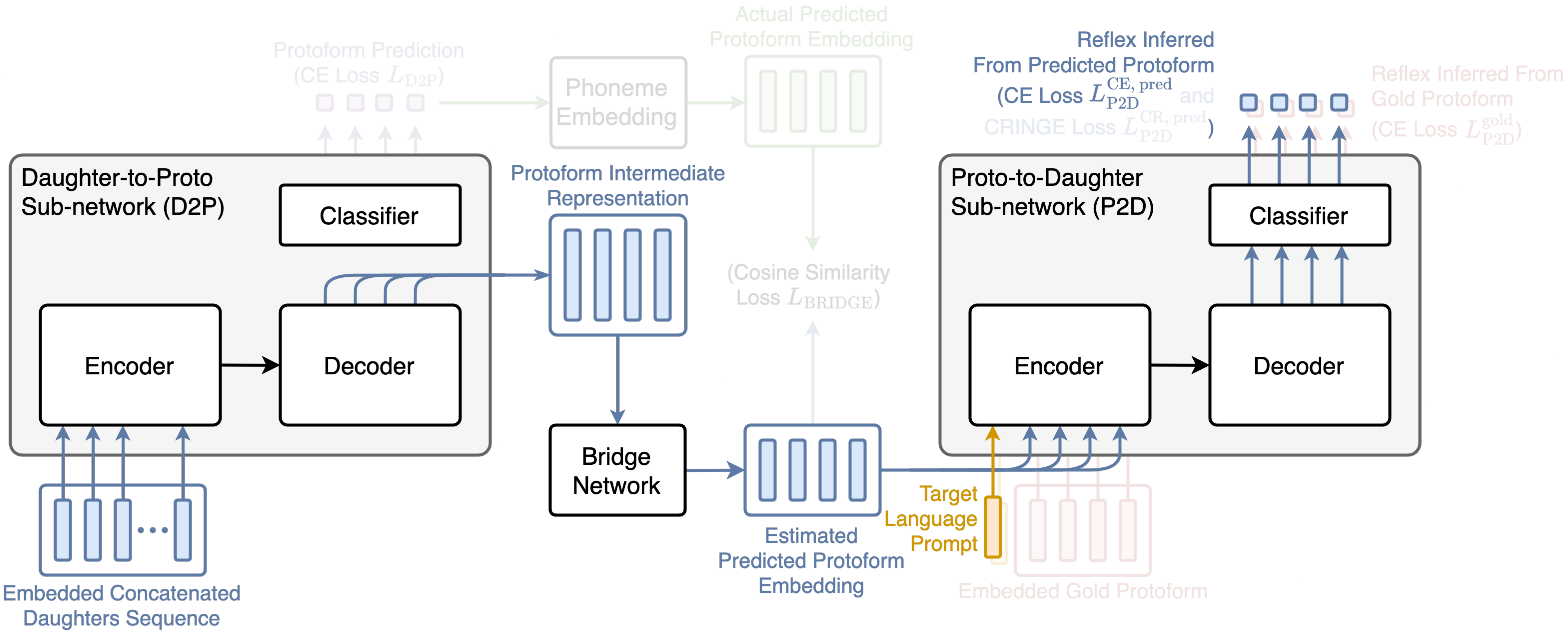


Protoform Prediction
(CE Loss L_{D2P})
Phoneme Embedding
Actual Predicted Protoform Embedding
Reflex Inferred From Predicted Protoform (CE Loss $L^{\text{CE, pred}}_{\text{P2D}}$ and CRINGE Loss $L^{\text{CR, pred}}_{\text{P2D}}$)
Reflex Inferred From Gold Protoform (CE Loss $L^{\text{gold}}_{\text{P2D}}$)
Daughter-to-Proto Sub-network (D2P)
Classifier
Protoform Intermediate Representation
Proto-to-Daughter Sub-network (P2D)
Classifier
(Cosine Similarity Loss L_{BRIDGE})
Encoder
Decoder
Encoder
Decoder
Bridge Network
Target Language Prompt
Embedded Concatenated Daughters Sequence
Estimated Predicted Protoform Embedding
Embedded Gold Protoform


Protoform Prediction
(CE Loss L_{D2P})
Phoneme Embedding
Actual Predicted
Protoform Embedding
Reflex Inferred
From Predicted Protoform
(CE Loss $L_{\text{P2D}}^{\text{CE, pred}}$ and
CRINGE Loss $L_{\text{P2D}}^{\text{CR, pred}}$)
Reflex Inferred From
Gold Protoform
(CE Loss $L_{\text{P2D}}^{\text{gold}}$)
Daughter-to-Proto
Sub-network (D2P)
Classifier
Encoder
Decoder
Protoform Intermediate
Representation
(Cosine Similarity
Loss L_{BRIDGE})
Proto-to-Daughter
Sub-network (P2D)
Classifier
Encoder
Decoder
Bridge
Network
Estimated
Predicted Protoform
Embedding
Target
Language
Prompt
Embedded Gold Protoform
Embedded Concatenated
Daughters Sequence

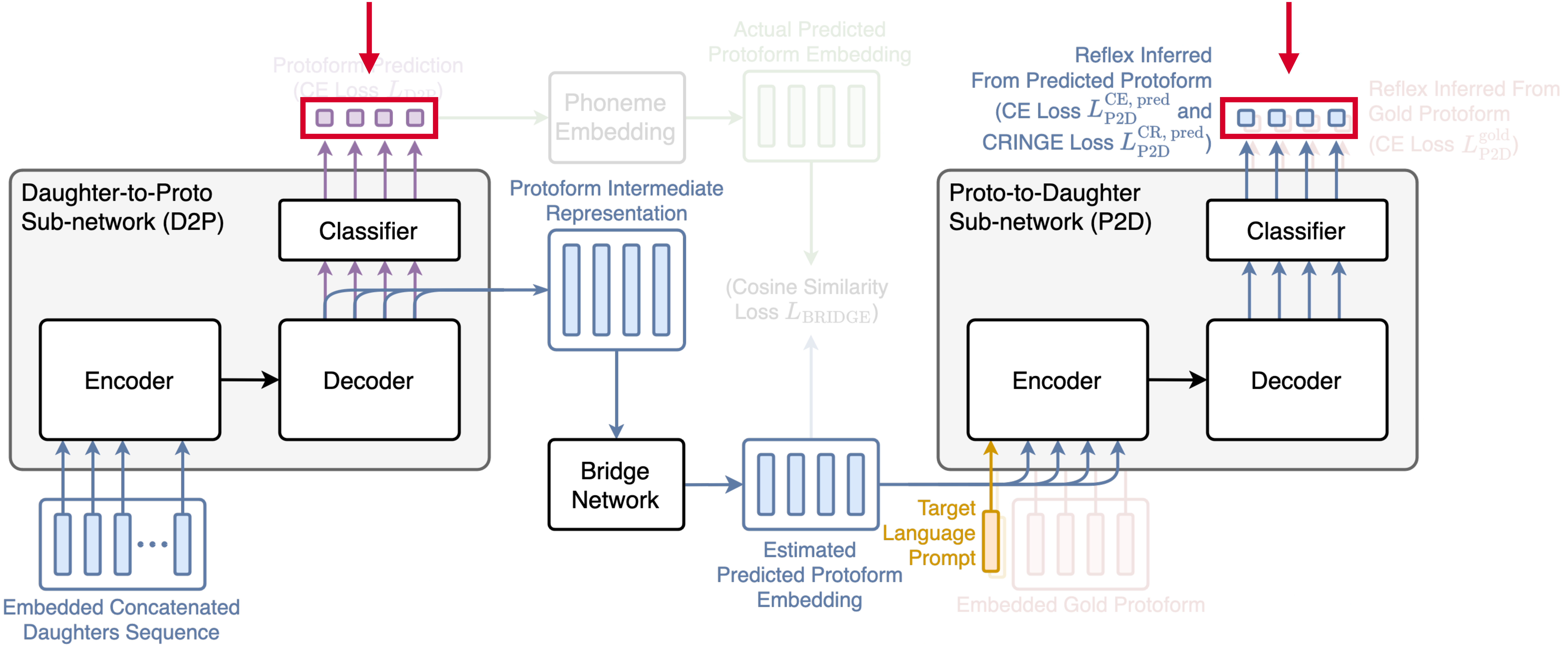

Actual Predicted
Protoform Embedding
Protoform Prediction
(CE Loss L_{D2P})
Phoneme Embedding
Reflex Inferred
From Predicted Protoform
(CE Loss $L_{\text{P2D}}^{\text{CE, pred}}$ and
CRINGE Loss $L_{\text{P2D}}^{\text{CR, pred}}$)
Reflex Inferred From
Gold Protoform
(CE Loss $L_{\text{P2D}}^{\text{gold}}$)
Daughter-to-Proto
Sub-network (D2P)
Classifier
Protoform Intermediate
Representation
Proto-to-Daughter
Sub-network (P2D)
Classifier
(Cosine Similarity
Loss L_{BRIDGE})
Encoder
Decoder
Encoder
Decoder
Bridge
Network
Target
Language
Prompt
Estimated
Predicted Protoform
Embedding
Embedded Concatenated
Daughters Sequence
Embedded Gold Protoform

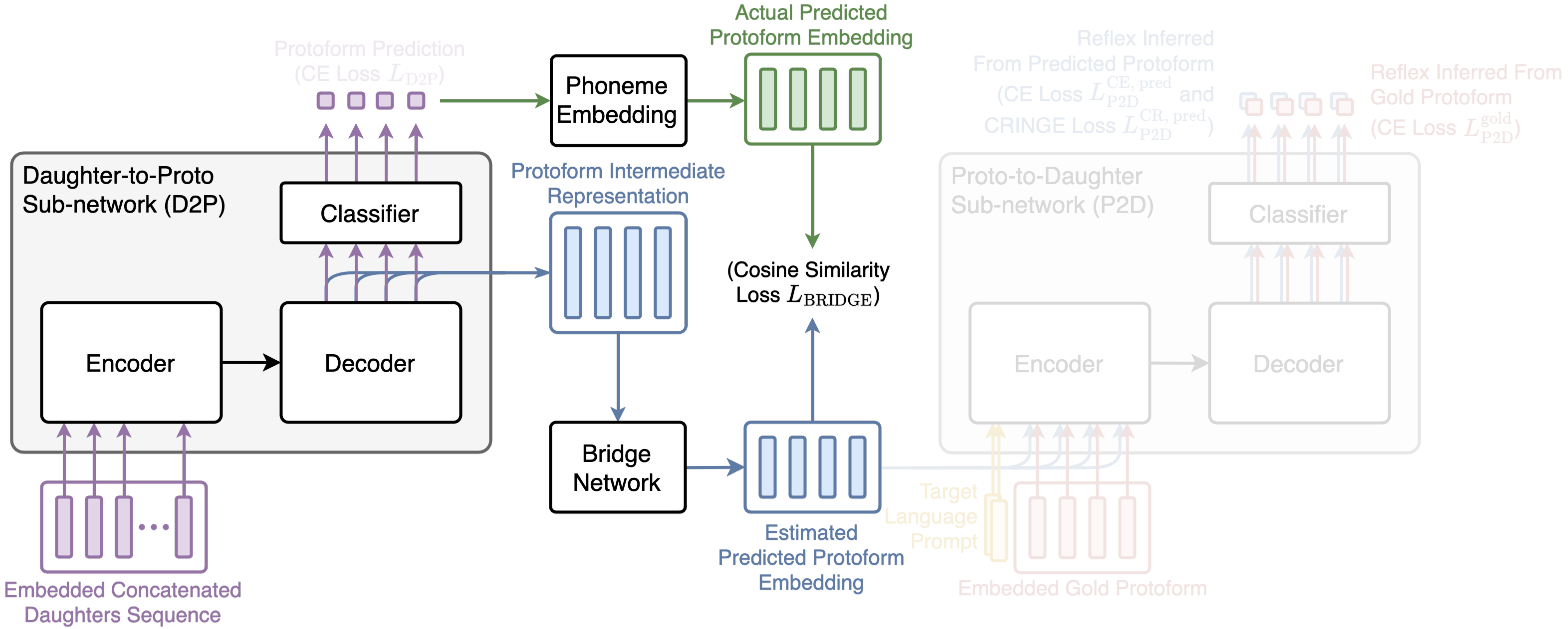

Protoform Prediction
(CE Loss L_{D2P})
Phoneme Embedding
Actual Predicted Protoform Embedding
Reflex Inferred From Predicted Protoform (CE Loss $L_{\text{P2D}}^{\text{CE, pred}}$ and CRINGE Loss $L_{\text{P2D}}^{\text{CR, pred}}$)
Reflex Inferred From Gold Protoform
(CE Loss $L_{\text{P2D}}^{\text{gold}}$)
Daughter-to-Proto Sub-network (D2P)
Classifier
Encoder
Decoder
Protoform Intermediate Representation
(Cosine Similarity Loss L_{BRIDGE})
Proto-to-Daughter Sub-network (P2D)
Classifier
Encoder
Decoder
Bridge Network
Estimated Predicted Protoform Embedding
Target Language Prompt
Embedded Gold Protoform
Embedded Concatenated Daughters Sequence

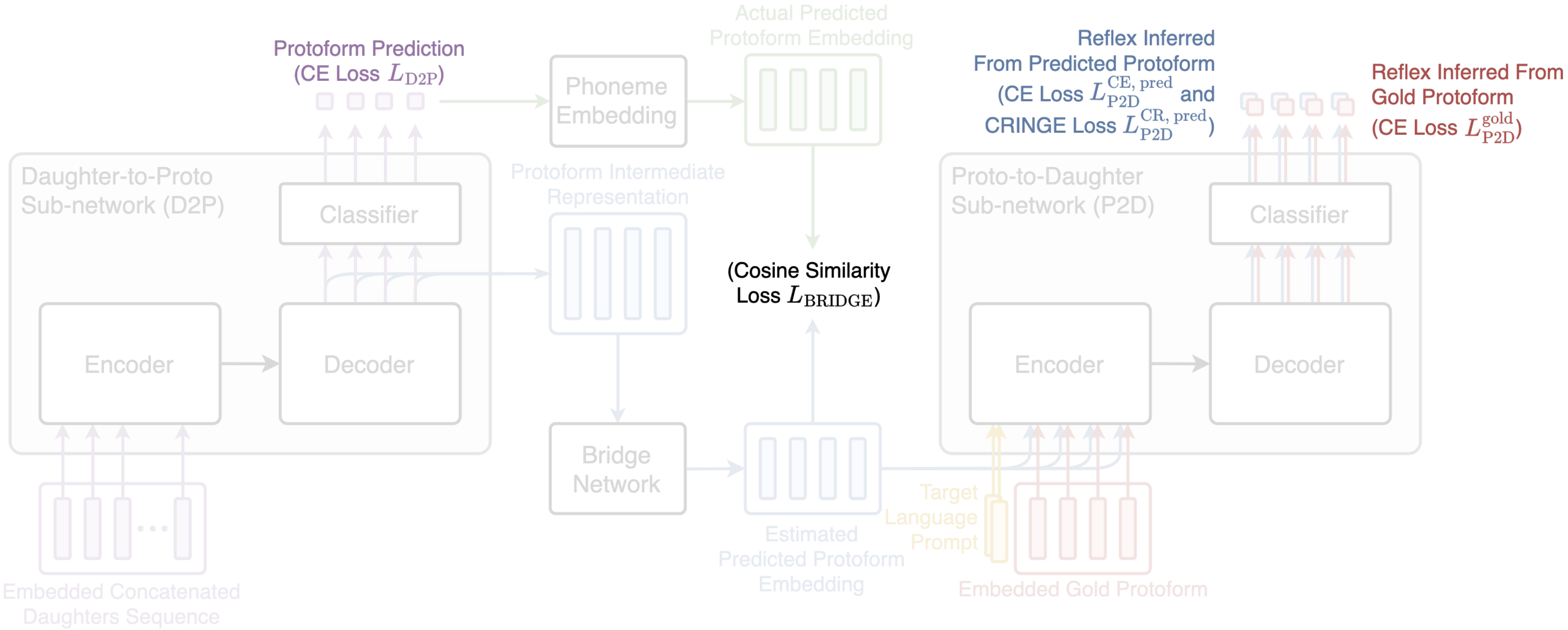

Protoform Prediction
(CE Loss L_{D2P})
Phoneme Embedding
Actual Predicted Protoform Embedding
Reflex Inferred From Predicted Protoform
(CE Loss $L_{\text{P2D}}^{\text{CE, pred}}$ and
CRINGE Loss $L_{\text{P2D}}^{\text{CR, pred}}$)
Reflex Inferred From Gold Protoform
(CE Loss $L_{\text{P2D}}^{\text{gold}}$)
Daughter-to-Proto Sub-network (D2P)
Classifier
Encoder
Decoder
Protoform Intermediate Representation
(Cosine Similarity Loss L_{BRIDGE})
Proto-to-Daughter Sub-network (P2D)
Classifier
Encoder
Decoder
Bridge Network
Embedded Concatenated Daughters Sequence
Estimated Predicted Protoform Embedding
Target Language Prompt
Embedded Gold Protoform

Incorrect protoform

Correct reflexes derived from incorrect protoform


Protoform Prediction
(CE Loss L_{D2P})
Phoneme Embedding
Actual Predicted Protoform Embedding
Reflex Inferred From Predicted Protoform (CE Loss $L_{\text{P2D}}^{\text{CE, pred}}$ and CRINGE Loss $L_{\text{P2D}}^{\text{CR, pred}}$)
Reflex Inferred From Gold Protoform (CE Loss $L_{\text{P2D}}^{\text{gold}}$)
Daughter-to-Proto Sub-network (D2P)
Classifier
Encoder
Decoder
Protoform Intermediate Representation
(Cosine Similarity Loss L_{BRIDGE})
Proto-to-Daughter Sub-network (P2D)
Classifier
Encoder
Decoder
Bridge Network
Target Language Prompt
Embedded Concatenated Daughters Sequence
Estimated Predicted Protoform Embedding
Embedded Gold Protoform


Protoform Prediction
(CE Loss L_{D2P})
Phoneme Embedding
Actual Predicted Protoform Embedding
Reflex Inferred From Predicted Protoform (CE Loss $L_{\text{P2D}}^{\text{CE, pred}}$ and CRINGE Loss $L_{\text{P2D}}^{\text{CR, pred}}$)
Reflex Inferred From Gold Protoform (CE Loss $L_{\text{P2D}}^{\text{gold}}$)
Daughter-to-Proto Sub-network (D2P)
Classifier
Encoder
Decoder
Protoform Intermediate Representation
(Cosine Similarity Loss L_{BRIDGE})
Proto-to-Daughter Sub-network (P2D)
Classifier
Encoder
Decoder
Bridge Network
Estimated Predicted Protoform Embedding
Target Language Prompt
Embedded Gold Protoform
Embedded Concatenated Daughters Sequence



Protoform Prediction (CE Loss L_{D2P})

Phoneme Embedding

Actual Predicted Protoform Embedding

Reflex Inferred From Predicted Protoform (CE Loss $L_{\text{P2D}}^{\text{CE, pred}}$ and CRINGE Loss $L_{\text{P2D}}^{\text{CR, pred}}$)

Reflex Inferred From Gold Protoform (CE Loss $L_{\text{P2D}}^{\text{gold}}$)

Daughter-to-Proto Sub-network (D2P)

Classifier

Encoder

Decoder

Protoform Intermediate Representation

(Cosine Similarity Loss L_{BRIDGE})

Proto-to-Daughter Sub-network (P2D)

Classifier

Encoder

Decoder

Bridge Network

Embedded Concatenated Daughters Sequence

Estimated Predicted Protoform Embedding

Target Language Prompt

Embedded Gold Protoform

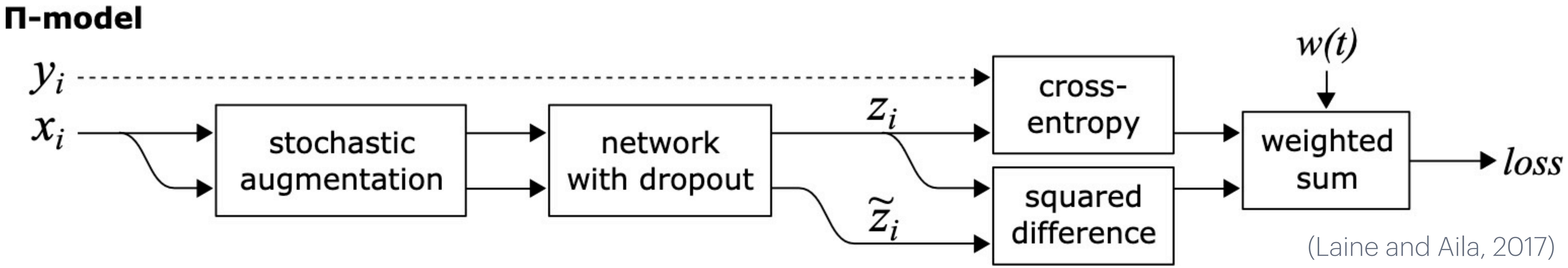
$$L_{\text{overall}} = \alpha_1 L_{\text{D2P}} + \alpha_2 L_{\text{P2D}}^{\text{CE, pred}} + \alpha_3 L_{\text{P2D}}^{\text{CR, pred}} + \alpha_4 L_{\text{P2D}}^{\text{gold}} + \alpha_5 L_{\text{BRIDGE}} \quad \text{where} \quad \alpha_{\{1\ldots5\}} \text{ are constants}$$

Weak Baseline Strategies

Supervised only (SUPV): only train the model on the labeled training examples

Bootstrapping (BST): A form of **proxy-labelling** in which the model's **most confident predictions are added as pseudo-labels** to the train set (Lee, 2013)

Π-Model (ΠM): An implementation of **consistency regularization** by training the model to **produce similar outputs on stochastically augmented inputs** (Laine and Aila, 2017)



(Laine and Aila, 2017)

Implementing Stochastic Augmentation for Π-Model

Original input sequence

[Cantonese]:mei̯˧[Mandarin]:mei̯˥˩*[Wu]:me̞˨˧*

↓

Randomly reorder the reflexes

[Wu]:me̞˨˧[Mandarin]:mei̯˥˩*[Cantonese]:mei̯˧*

↓

Drop a daughter language with a 50% probability (unless there is only one)

[Mandarin]:mei̯˥˩[Cantonese]:mei̯˧*

Architectures and Training Strategies

1. Supervised only (**SUPV**)
2. Bootstrapping (**BST**)
3. Π-model (**ΠM**)

Weak baselines

4. Π-model with Bootstrapping (**ΠM-BST**)

Strong baseline

5. **DPD**
6. DPD with Bootstrapping (**DPD-BST**)
7. DPD merged with Π-model (**DPD-ΠM**)
8. DPD-ΠM with Bootstrapping (**DPD-ΠM-BST**)

DPD-based strategies

Architectures and Training Strategies

1. Supervised only (**SUPV**)
2. Bootstrapping (**BST**)
3. Π-model (**ΠM**)
4. Π-model with Bootstrapping (**ΠM-BST**)
5. **DPD**
6. DPD with Bootstrapping (**DPD-BST**)
7. DPD merged with Π-model (**DPD-ΠM**)
8. DPD-ΠM with Bootstrapping (**DPD-ΠM-BST**)

×

cartesian product

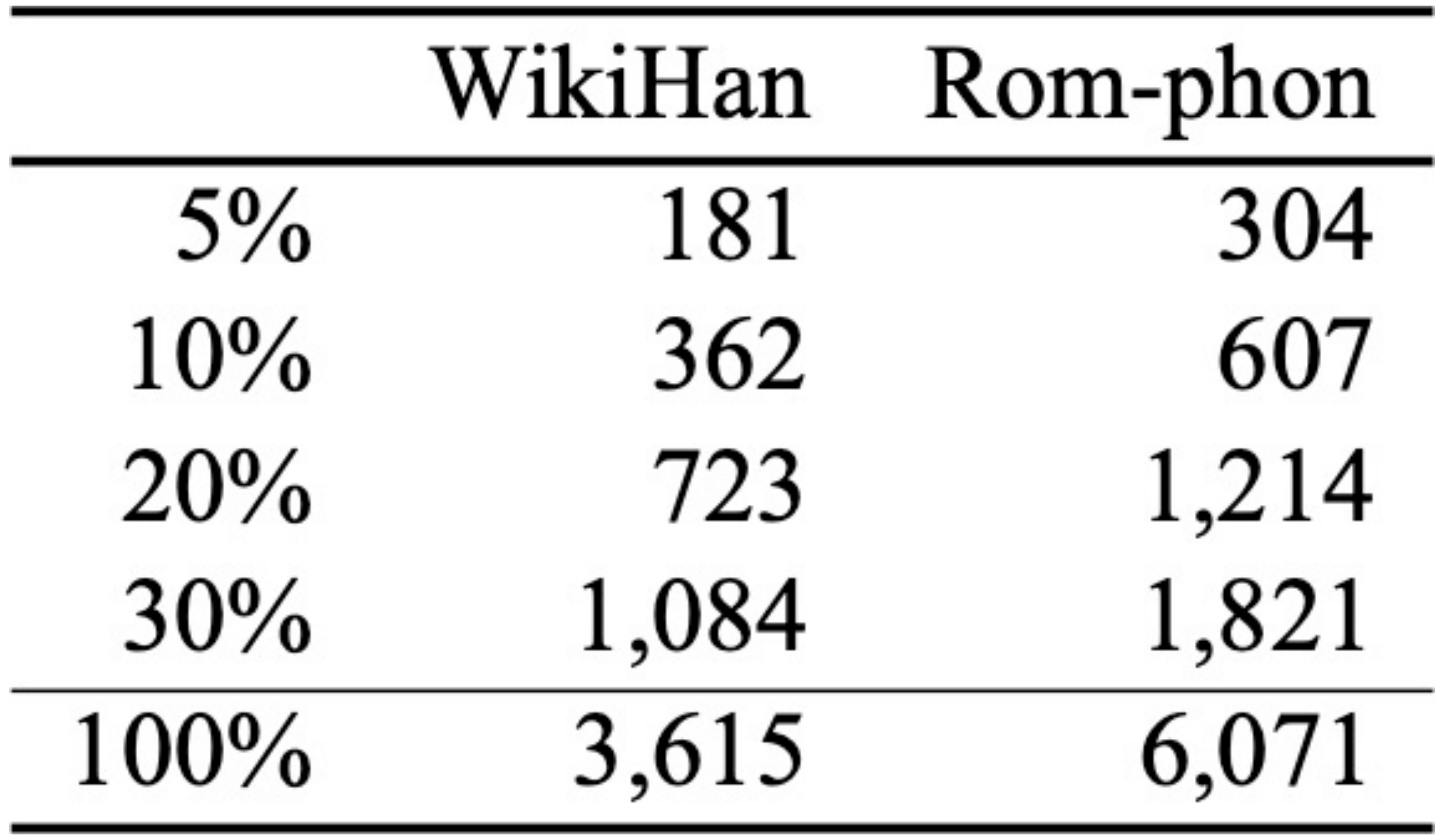
1. GRU (**GRU**)
2. Transformer (**Trans**)

Datasets

Dataset	Language Family	Ancestor Language	Number of Cognate Sets
WikiHan (phonetic) (Chang et al., 2022)	Sinitic	Middle Chinese	8,703
Rom-phon (Romance, phonetic version) (Meloni et al., 2021; Ciobanu and Dinu, 2018)	Romance	Latin	5,165

Semisupervised Datasets

We **take away labels** to simulate a semisupervised situation.

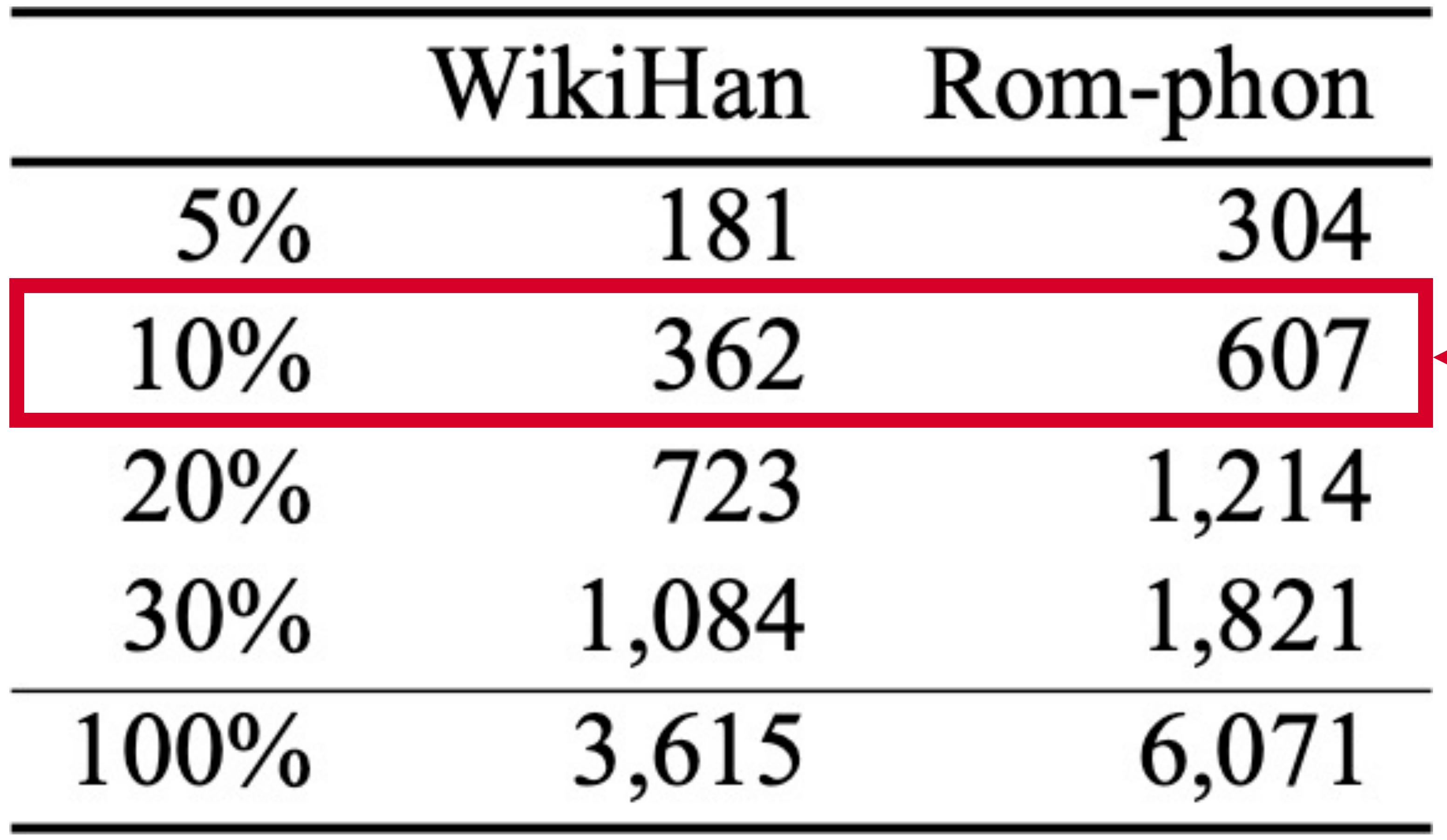


	WikiHan	Rom-phon
5%	181	304
10%	362	607
20%	723	1,214
30%	1,084	1,821
100%	3,615	6,071

Number of labeled training examples (i.e. cognate sets with an associated gold protoform) in the train set for each labeling setting and dataset, as well as the total number of cognate sets for reference (100%).

Semisupervised Datasets

We **take away labels** to simulate a semisupervised situation.



	WikiHan	Rom-phon
5%	181	304
10%	362	607
20%	723	1,214
30%	1,084	1,821
100%	3,615	6,071

Our focus

Number of labeled training examples (i.e. cognate sets with an associated gold protoform) in the train set for each labeling setting and dataset, as well as the total number of cognate sets for reference (100%).

Evaluation Metrics

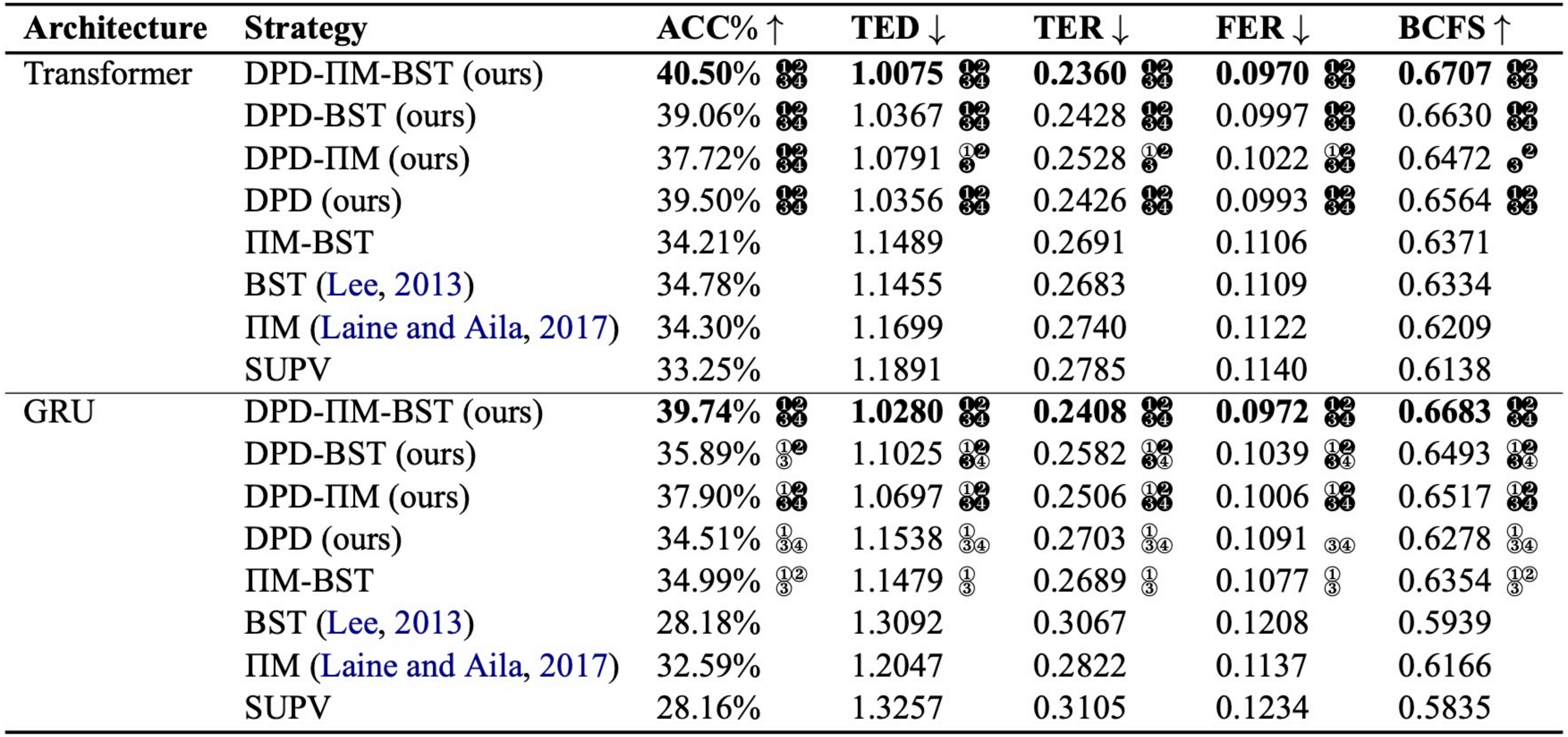
- **Accuracy (ACC):** The percentage of exactly correct predictions
- **Token edit distance (TED):** The number of token insertions, deletions, or substitutions between predictions and targets (Levenshtein et al., 1966)
- **Token error rate (TER):** Length-normalized edit distance (Cui et al., 2022)
- **Feature error rate (FER):** Length-normalized phonological edit distance measured by PanPhon (Mortensen et al., 2016)
- **B-Cubed F Score (BCFS):** A measure of the structural similarity between predictions and targets (Amigó et al., 2009; List, 2019)

Results

DPD Performs Well

Results:
10% Labeled WikiHan

Bold: the best-performing model for each metric
①**:** significantly better than all weak baselines (SUPV, BST, and ΠM) on dataset seed 1 with $p < 0.01$
❶**:** significantly better than the ΠM-BST strong baseline and all weak baselines on dataset seed 1 with $p < 0.01$
②**,** ③**,** ④**,** ❷**,** ❸**,** ❹**:** likewise for dataset seeds 2–4.

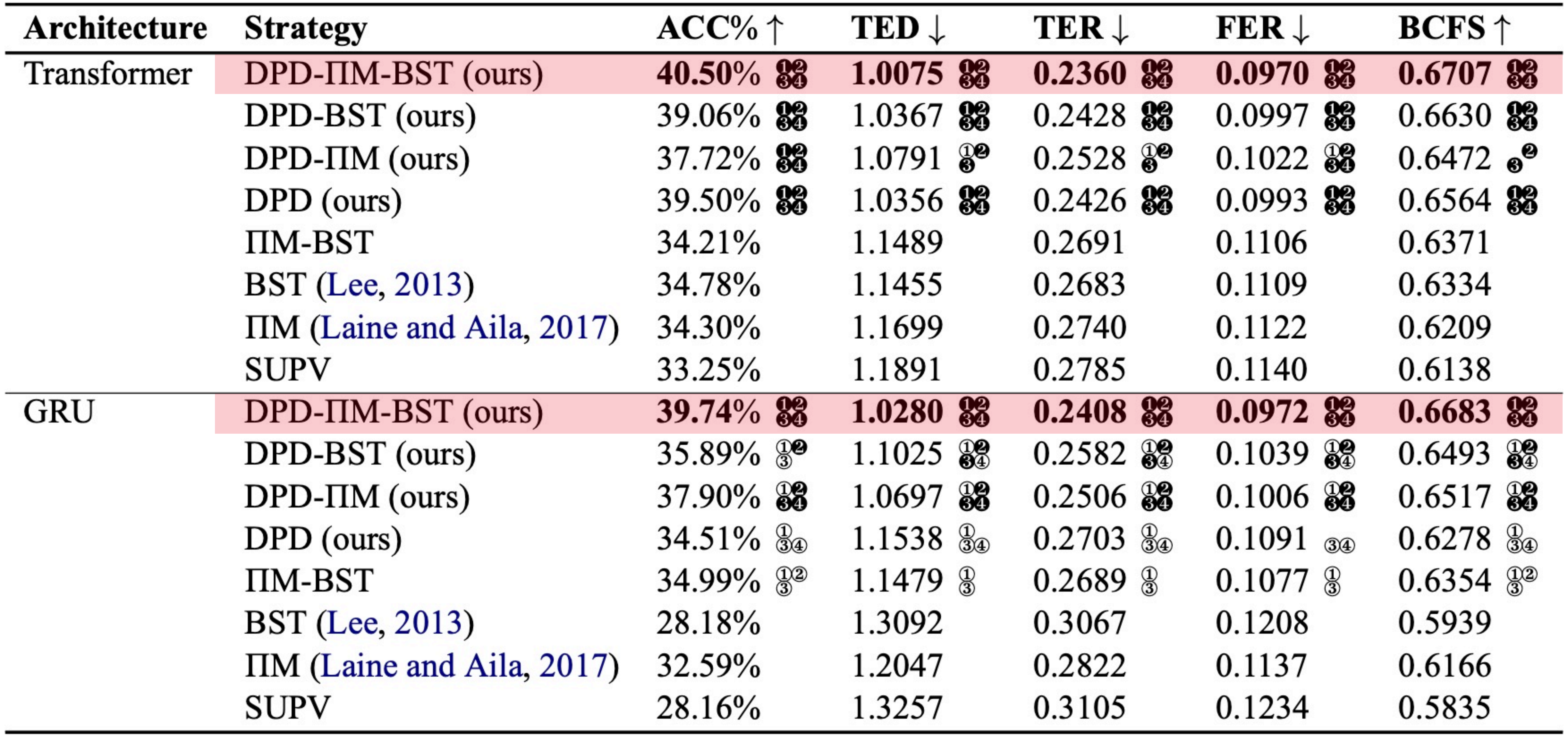


Architecture	Strategy	ACC% ↑	TED ↓	TER ↓	FER ↓	BCFS ↑
Transformer	DPD-ΠM-BST (ours)	**40.50%** ❶❷❸❹	**1.0075** ❶❷❸❹	**0.2360** ❶❷❸❹	**0.0970** ❶❷❸❹	**0.6707** ❶❷❸❹
	DPD-BST (ours)	39.06% ❶❷❸❹	1.0367 ❶❷❸❹	0.2428 ❶❷❸❹	0.0997 ❶❷❸❹	0.6630 ❶❷❸❹
	DPD-ΠM (ours)	37.72% ❶❷❸❹	1.0791 ①❷❸	0.2528 ①❷❸	0.1022 ①❷❸❹	0.6472 ❷❸
	DPD (ours)	39.50% ❶❷❸❹	1.0356 ❶❷❸❹	0.2426 ❶❷❸❹	0.0993 ❶❷❸❹	0.6564 ❶❷❸❹
	ΠM-BST	34.21%	1.1489	0.2691	0.1106	0.6371
	BST (Lee, 2013)	34.78%	1.1455	0.2683	0.1109	0.6334
	ΠM (Laine and Aila, 2017)	34.30%	1.1699	0.2740	0.1122	0.6209
	SUPV	33.25%	1.1891	0.2785	0.1140	0.6138
GRU	DPD-ΠM-BST (ours)	**39.74%** ❶❷❸❹	**1.0280** ❶❷❸❹	**0.2408** ❶❷❸❹	**0.0972** ❶❷❸❹	**0.6683** ❶❷❸❹
	DPD-BST (ours)	35.89% ①❷③	1.1025 ①❷❸④	0.2582 ①❷❸④	0.1039 ①❷❸④	0.6493 ①❷❸④
	DPD-ΠM (ours)	37.90% ①❷❸❹	1.0697 ①❷❸❹	0.2506 ①❷❸❹	0.1006 ①❷❸❹	0.6517 ①❷❸❹
	DPD (ours)	34.51% ①③④	1.1538 ①③④	0.2703 ①③④	0.1091 ③④	0.6278 ①③④
	ΠM-BST	34.99% ①②③	1.1479 ①③	0.2689 ①③	0.1077 ①③	0.6354 ①②③
	BST (Lee, 2013)	28.18%	1.3092	0.3067	0.1208	0.5939
	ΠM (Laine and Aila, 2017)	32.59%	1.2047	0.2822	0.1137	0.6166
	SUPV	28.16%	1.3257	0.3105	0.1234	0.5835

Results:
10% Labeled WikiHan

Bold: the best-performing model for each metric
①: significantly better than all weak baselines (SUPV, BST, and ΠM) on dataset seed 1 with $p < 0.01$
❶: significantly better than the ΠM-BST strong baseline and all weak baselines on dataset seed 1 with $p < 0.01$
②, ③, ④, ❷, ❸, ❹: likewise for dataset seeds 2–4.

DPD-ΠM-BST performs the best and significantly better than all baselines on all metrics.



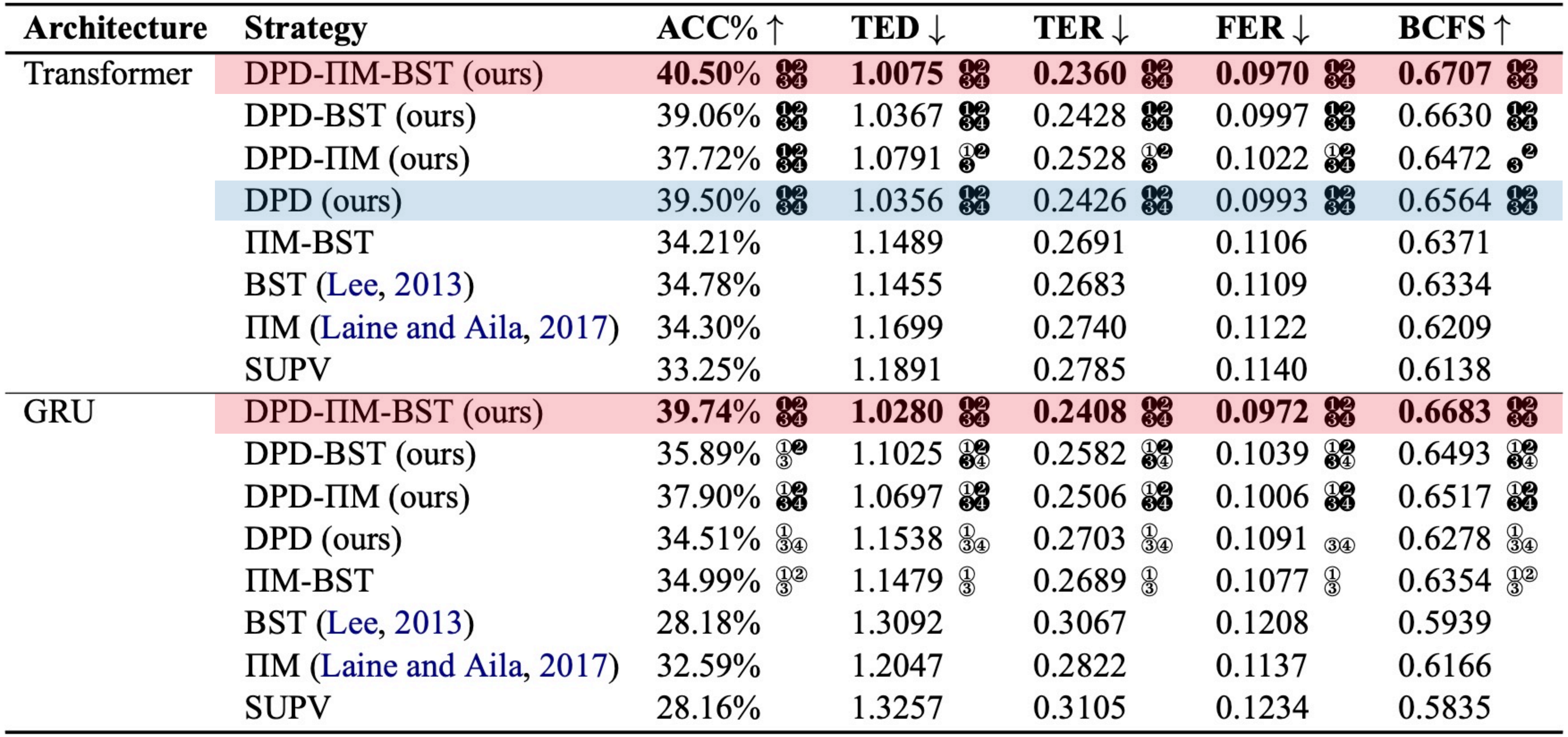
Architecture	Strategy	ACC% ↑	TED ↓	TER ↓	FER ↓	BCFS ↑
Transformer	DPD-ΠM-BST (ours)	**40.50%** ❶❷❸❹	**1.0075** ❶❷❸❹	**0.2360** ❶❷❸❹	**0.0970** ❶❷❸❹	**0.6707** ❶❷❸❹
	DPD-BST (ours)	39.06% ❶❷❸❹	1.0367 ❶❷❸❹	0.2428 ❶❷❸❹	0.0997 ❶❷❸❹	0.6630 ❶❷❸❹
	DPD-ΠM (ours)	37.72% ❶❷❸❹	1.0791 ①❷❸	0.2528 ①❷❸	0.1022 ①❷❸❹	0.6472 ❷❸
	DPD (ours)	39.50% ❶❷❸❹	1.0356 ❶❷❸❹	0.2426 ❶❷❸❹	0.0993 ❶❷❸❹	0.6564 ❶❷❸❹
	ΠM-BST	34.21%	1.1489	0.2691	0.1106	0.6371
	BST (Lee, 2013)	34.78%	1.1455	0.2683	0.1109	0.6334
	ΠM (Laine and Aila, 2017)	34.30%	1.1699	0.2740	0.1122	0.6209
	SUPV	33.25%	1.1891	0.2785	0.1140	0.6138
GRU	DPD-ΠM-BST (ours)	**39.74%** ❶❷❸❹	**1.0280** ❶❷❸❹	**0.2408** ❶❷❸❹	**0.0972** ❶❷❸❹	**0.6683** ❶❷❸❹
	DPD-BST (ours)	35.89% ①❷③	1.1025 ①❷❸④	0.2582 ①❷❸④	0.1039 ①❷❸④	0.6493 ①❷❸④
	DPD-ΠM (ours)	37.90% ①❷❸❹	1.0697 ①❷❸❹	0.2506 ①❷❸❹	0.1006 ①❷❸❹	0.6517 ①❷❸❹
	DPD (ours)	34.51% ①③④	1.1538 ①③④	0.2703 ①③④	0.1091 ③④	0.6278 ①③④
	ΠM-BST	34.99% ①②③	1.1479 ①③	0.2689 ①③	0.1077 ①③	0.6354 ①②③
	BST (Lee, 2013)	28.18%	1.3092	0.3067	0.1208	0.5939
	ΠM (Laine and Aila, 2017)	32.59%	1.2047	0.2822	0.1137	0.6166
	SUPV	28.16%	1.3257	0.3105	0.1234	0.5835

Results:
10% Labeled WikiHan

Bold: the best-performing model for each metric
①: significantly better than all weak baselines (SUPV, BST, and ΠM) on dataset seed 1 with $p < 0.01$
❶: significantly better than the ΠM-BST strong baseline and all weak baselines on dataset seed 1 with $p < 0.01$
②, ③, ④, ❷, ❸, ❹: likewise for dataset seeds 2–4.

DPD-ΠM-BST performs the best and significantly better than all baselines on all metrics.

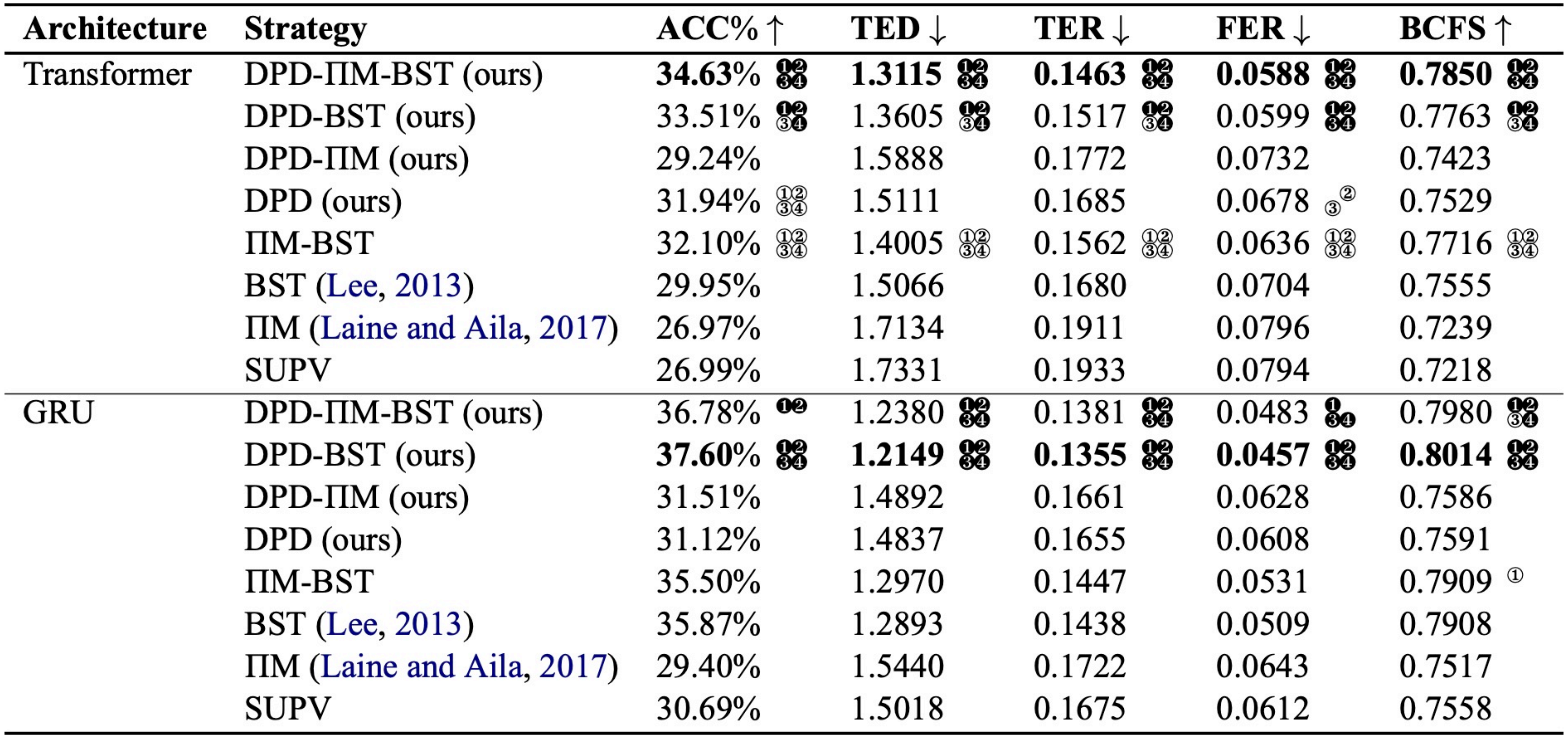
Transformer trained with DPD performs similarly well.



Architecture	Strategy	ACC% ↑	TED ↓	TER ↓	FER ↓	BCFS ↑
Transformer	DPD-ΠM-BST (ours)	**40.50%** ❶❷❸❹	**1.0075** ❶❷❸❹	**0.2360** ❶❷❸❹	**0.0970** ❶❷❸❹	**0.6707** ❶❷❸❹
	DPD-BST (ours)	39.06% ❶❷❸❹	1.0367 ❶❷❸❹	0.2428 ❶❷❸❹	0.0997 ❶❷❸❹	0.6630 ❶❷❸❹
	DPD-ΠM (ours)	37.72% ❶❷❸❹	1.0791 ①❷❸	0.2528 ①❷❸	0.1022 ①❷❸❹	0.6472 ❷❸
	DPD (ours)	39.50% ❶❷❸❹	1.0356 ❶❷❸❹	0.2426 ❶❷❸❹	0.0993 ❶❷❸❹	0.6564 ❶❷❸❹
	ΠM-BST	34.21%	1.1489	0.2691	0.1106	0.6371
	BST (Lee, 2013)	34.78%	1.1455	0.2683	0.1109	0.6334
	ΠM (Laine and Aila, 2017)	34.30%	1.1699	0.2740	0.1122	0.6209
	SUPV	33.25%	1.1891	0.2785	0.1140	0.6138
GRU	DPD-ΠM-BST (ours)	**39.74%** ❶❷❸❹	**1.0280** ❶❷❸❹	**0.2408** ❶❷❸❹	**0.0972** ❶❷❸❹	**0.6683** ❶❷❸❹
	DPD-BST (ours)	35.89% ①❷③	1.1025 ①❷❸④	0.2582 ①❷❸④	0.1039 ①❷❸④	0.6493 ①❷❸④
	DPD-ΠM (ours)	37.90% ①❷❸❹	1.0697 ①❷❸❹	0.2506 ①❷❸❹	0.1006 ①❷❸❹	0.6517 ①❷❸❹
	DPD (ours)	34.51% ①③④	1.1538 ①③④	0.2703 ①③④	0.1091 ③④	0.6278 ①③④
	ΠM-BST	34.99% ①②③	1.1479 ①③	0.2689 ①③	0.1077 ①③	0.6354 ①②③
	BST (Lee, 2013)	28.18%	1.3092	0.3067	0.1208	0.5939
	ΠM (Laine and Aila, 2017)	32.59%	1.2047	0.2822	0.1137	0.6166
	SUPV	28.16%	1.3257	0.3105	0.1234	0.5835

Results:
10% Labeled Rom-phon

Bold: the best-performing model for each metric
①**:** significantly better than all weak baselines (SUPV, BST, and ΠM) on dataset seed 1 with p < 0.01
❶**:** significantly better than the ΠM-BST strong baseline and all weak baselines on dataset seed 1 with p < 0.01
②, ③, ④, ❷, ❸, ❹**:** likewise for dataset seeds 2–4.

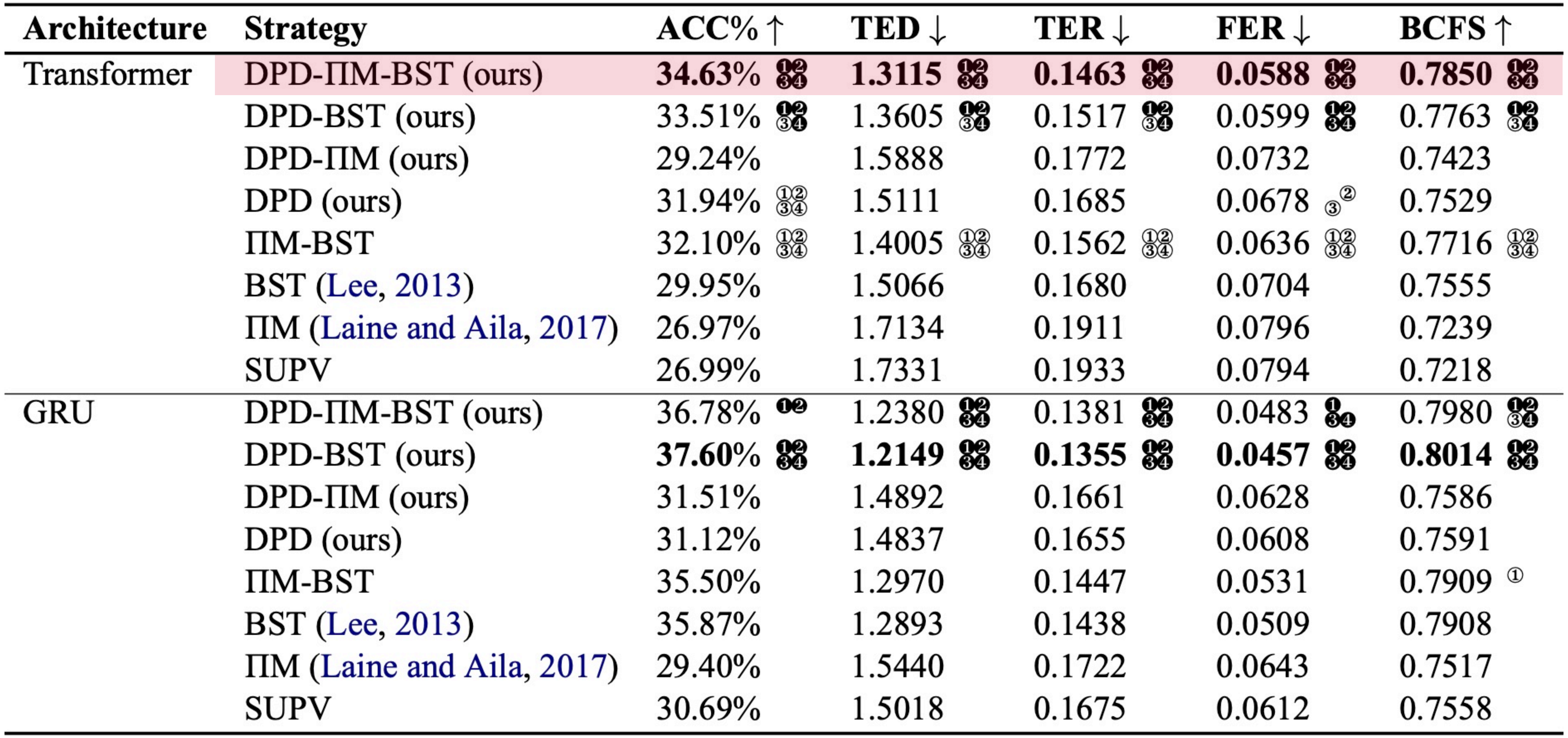


Architecture	Strategy	ACC% ↑	TED ↓	TER ↓	FER ↓	BCFS ↑
Transformer	DPD-ΠM-BST (ours)	**34.63%** ❶❷❸❹	**1.3115** ❶❷❸❹	**0.1463** ❶❷❸❹	**0.0588** ❶❷❸❹	**0.7850** ❶❷❸❹
	DPD-BST (ours)	33.51% ❶❷③❹	1.3605 ❶❷③❹	0.1517 ❶❷③❹	0.0599 ❶❷❸❹	0.7763 ❶❷③❹
	DPD-ΠM (ours)	29.24%	1.5888	0.1772	0.0732	0.7423
	DPD (ours)	31.94% ①②③④	1.5111	0.1685	0.0678 ②③	0.7529
	ΠM-BST	32.10% ①②③④	1.4005 ①②③④	0.1562 ①②③④	0.0636 ①②③④	0.7716 ①②③④
	BST (Lee, 2013)	29.95%	1.5066	0.1680	0.0704	0.7555
	ΠM (Laine and Aila, 2017)	26.97%	1.7134	0.1911	0.0796	0.7239
	SUPV	26.99%	1.7331	0.1933	0.0794	0.7218
GRU	DPD-ΠM-BST (ours)	36.78% ❶❷	1.2380 ❶❷❸❹	0.1381 ❶❷❸❹	0.0483 ❶❸❹	0.7980 ❶❷③❹
	DPD-BST (ours)	**37.60%** ❶❷❸❹	**1.2149** ❶❷❸❹	**0.1355** ❶❷❸❹	**0.0457** ❶❷❸❹	**0.8014** ❶❷❸❹
	DPD-ΠM (ours)	31.51%	1.4892	0.1661	0.0628	0.7586
	DPD (ours)	31.12%	1.4837	0.1655	0.0608	0.7591
	ΠM-BST	35.50%	1.2970	0.1447	0.0531	0.7909 ①
	BST (Lee, 2013)	35.87%	1.2893	0.1438	0.0509	0.7908
	ΠM (Laine and Aila, 2017)	29.40%	1.5440	0.1722	0.0643	0.7517
	SUPV	30.69%	1.5018	0.1675	0.0612	0.7558

Results:
10% Labeled Rom-phon

Bold: the best-performing model for each metric
①**:** significantly better than all weak baselines (SUPV, BST, and ΠM) on dataset seed 1 with p < 0.01
❶**:** significantly better than the ΠM-BST strong baseline and all weak baselines on dataset seed 1 with p < 0.01
②**,** ③**,** ④**,** ❷**,** ❸**,** ❹**:** likewise for dataset seeds 2–4.

Transformer performed the best when trained with DPD-ΠM-BST



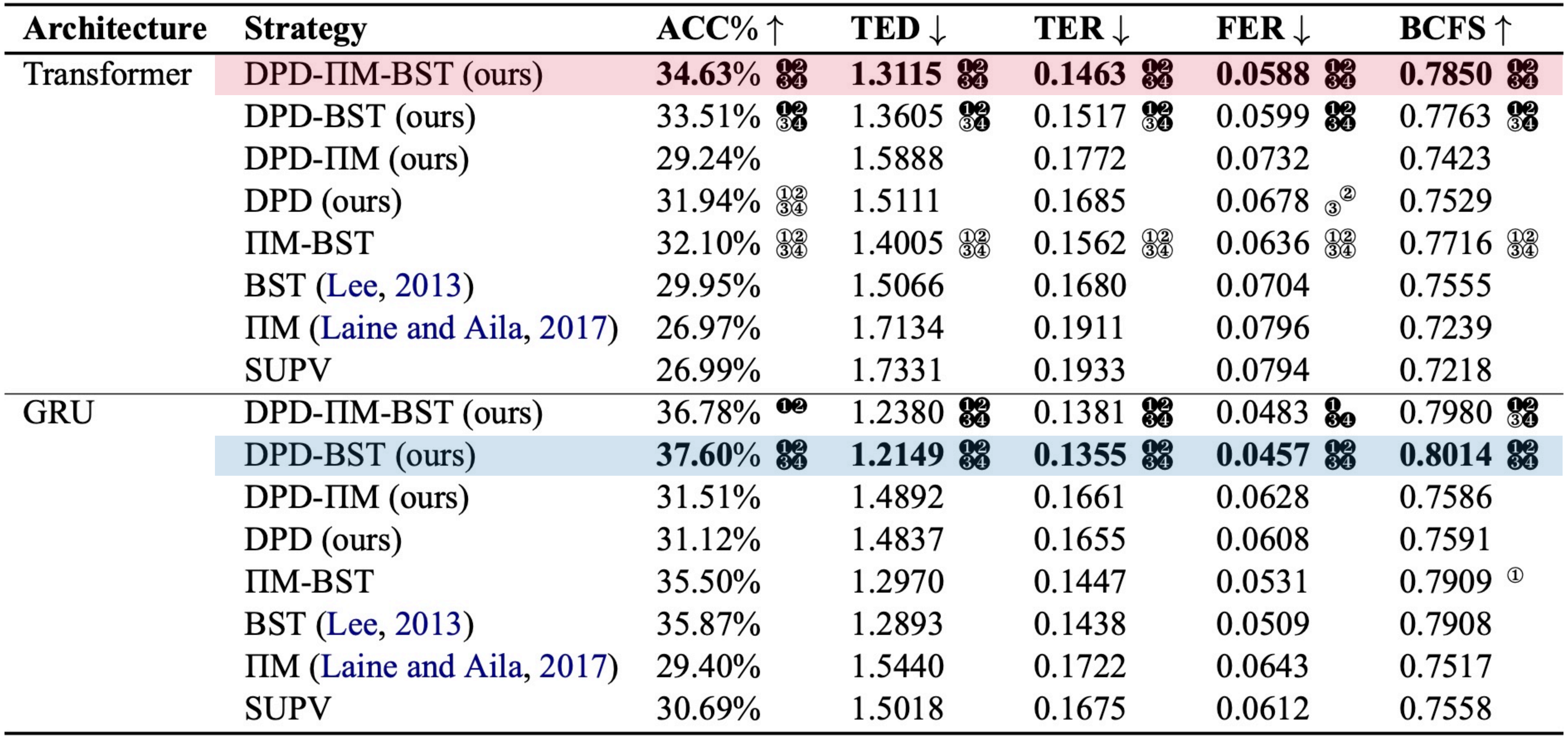
Architecture	Strategy	ACC% ↑	TED ↓	TER ↓	FER ↓	BCFS ↑
Transformer	DPD-ΠM-BST (ours)	**34.63%** ❶❷❸❹	**1.3115** ❶❷❸❹	**0.1463** ❶❷❸❹	**0.0588** ❶❷❸❹	**0.7850** ❶❷❸❹
	DPD-BST (ours)	33.51% ❶❷③❹	1.3605 ❶❷③❹	0.1517 ❶❷③❹	0.0599 ❶❷❸❹	0.7763 ❶❷③❹
	DPD-ΠM (ours)	29.24%	1.5888	0.1772	0.0732	0.7423
	DPD (ours)	31.94% ①②③④	1.5111	0.1685	0.0678 ②③	0.7529
	ΠM-BST	32.10% ①②③④	1.4005 ①②③④	0.1562 ①②③④	0.0636 ①②③④	0.7716 ①②③④
	BST (Lee, 2013)	29.95%	1.5066	0.1680	0.0704	0.7555
	ΠM (Laine and Aila, 2017)	26.97%	1.7134	0.1911	0.0796	0.7239
	SUPV	26.99%	1.7331	0.1933	0.0794	0.7218
GRU	DPD-ΠM-BST (ours)	36.78% ❶❷	1.2380 ❶❷❸❹	0.1381 ❶❷❸❹	0.0483 ❶❸❹	0.7980 ❶❷③❹
	DPD-BST (ours)	**37.60%** ❶❷❸❹	**1.2149** ❶❷❸❹	**0.1355** ❶❷❸❹	**0.0457** ❶❷❸❹	**0.8014** ❶❷❸❹
	DPD-ΠM (ours)	31.51%	1.4892	0.1661	0.0628	0.7586
	DPD (ours)	31.12%	1.4837	0.1655	0.0608	0.7591
	ΠM-BST	35.50%	1.2970	0.1447	0.0531	0.7909 ①
	BST (Lee, 2013)	35.87%	1.2893	0.1438	0.0509	0.7908
	ΠM (Laine and Aila, 2017)	29.40%	1.5440	0.1722	0.0643	0.7517
	SUPV	30.69%	1.5018	0.1675	0.0612	0.7558

Results:
10% Labeled Rom-phon

Bold: the best-performing model for each metric
①**:** significantly better than all weak baselines (SUPV, BST, and ΠM) on dataset seed 1 with p < 0.01
❶**:** significantly better than the ΠM-BST strong baseline and all weak baselines on dataset seed 1 with p < 0.01
②, ③, ④, ❷, ❸, ❹**:** likewise for dataset seeds 2–4.

Transformer performed the best when trained with DPD-ΠM-BST

GRU performed the best when trained with DPD-BST

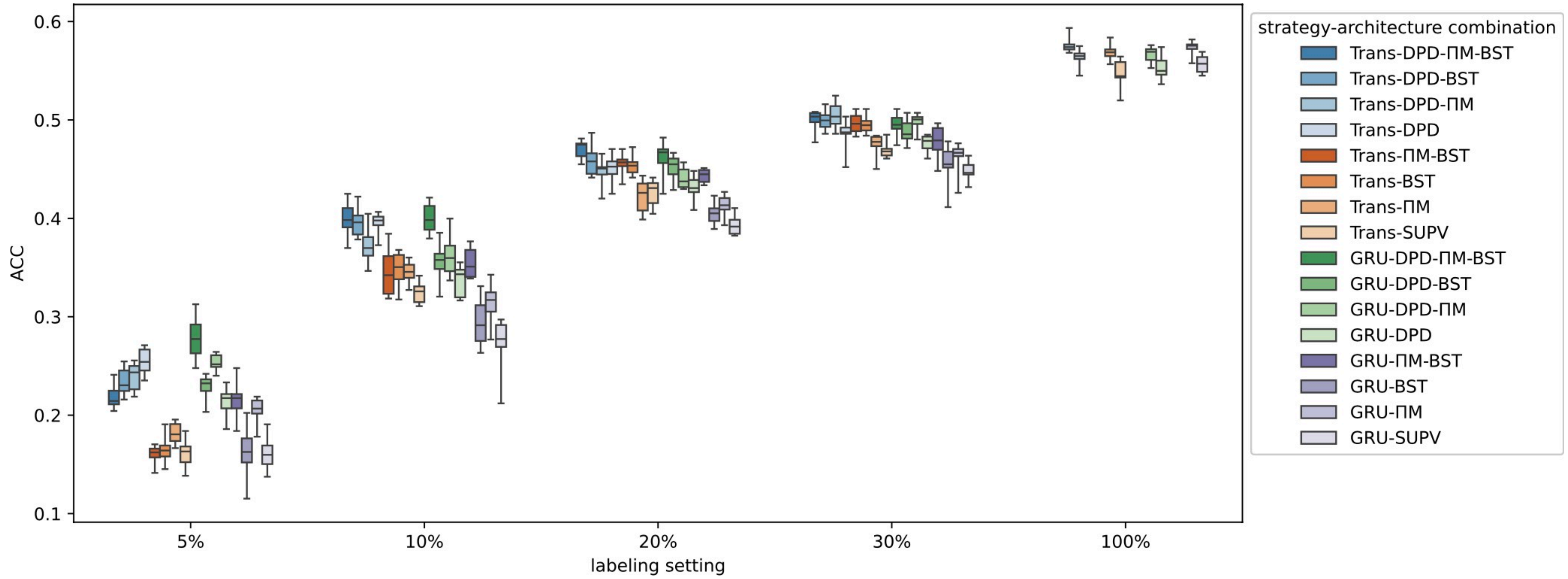


Architecture	Strategy	ACC% ↑	TED ↓	TER ↓	FER ↓	BCFS ↑
Transformer	DPD-ΠM-BST (ours)	**34.63%** ❶❷❸❹	**1.3115** ❶❷❸❹	**0.1463** ❶❷❸❹	**0.0588** ❶❷❸❹	**0.7850** ❶❷❸❹
	DPD-BST (ours)	33.51% ❶❷③❹	1.3605 ❶❷③❹	0.1517 ❶❷③❹	0.0599 ❶❷❸❹	0.7763 ❶❷③❹
	DPD-ΠM (ours)	29.24%	1.5888	0.1772	0.0732	0.7423
	DPD (ours)	31.94% ①②③④	1.5111	0.1685	0.0678 ②③	0.7529
	ΠM-BST	32.10% ①②③④	1.4005 ①②③④	0.1562 ①②③④	0.0636 ①②③④	0.7716 ①②③④
	BST (Lee, 2013)	29.95%	1.5066	0.1680	0.0704	0.7555
	ΠM (Laine and Aila, 2017)	26.97%	1.7134	0.1911	0.0796	0.7239
	SUPV	26.99%	1.7331	0.1933	0.0794	0.7218
GRU	DPD-ΠM-BST (ours)	36.78% ❶❷	1.2380 ❶❷❸❹	0.1381 ❶❷❸❹	0.0483 ❶❸❹	0.7980 ❶❷③❹
	DPD-BST (ours)	**37.60%** ❶❷❸❹	**1.2149** ❶❷❸❹	**0.1355** ❶❷❸❹	**0.0457** ❶❷❸❹	**0.8014** ❶❷❸❹
	DPD-ΠM (ours)	31.51%	1.4892	0.1661	0.0628	0.7586
	DPD (ours)	31.12%	1.4837	0.1655	0.0608	0.7591
	ΠM-BST	35.50%	1.2970	0.1447	0.0531	0.7909 ①
	BST (Lee, 2013)	35.87%	1.2893	0.1438	0.0509	0.7908
	ΠM (Laine and Aila, 2017)	29.40%	1.5440	0.1722	0.0643	0.7517
	SUPV	30.69%	1.5018	0.1675	0.0612	0.7558

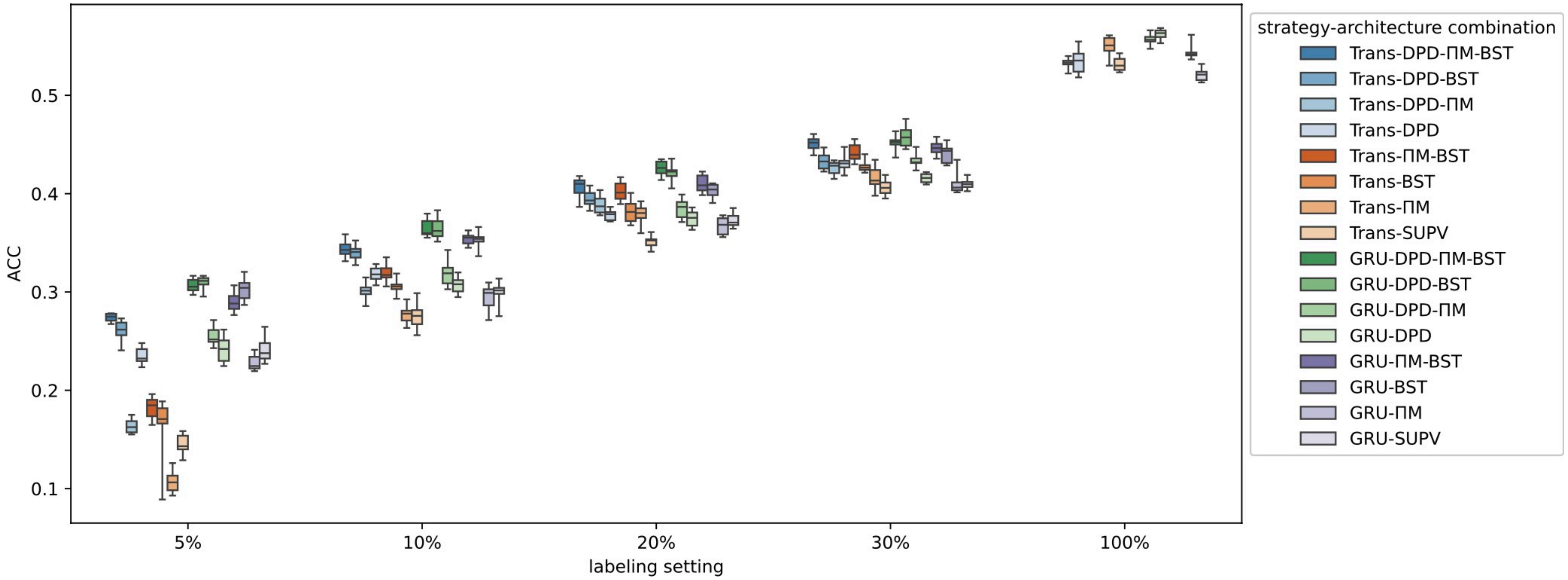
Performance on Different Labeling Settings (See Paper)

Performance distribution for varied labeling settings (dataset seed 1).

x-axis: various labeling settings including semisupervised situations at 5%, 10%, 20%, and 30% and fully supervised reference at 100% (not drawn to scale).



WikiHan

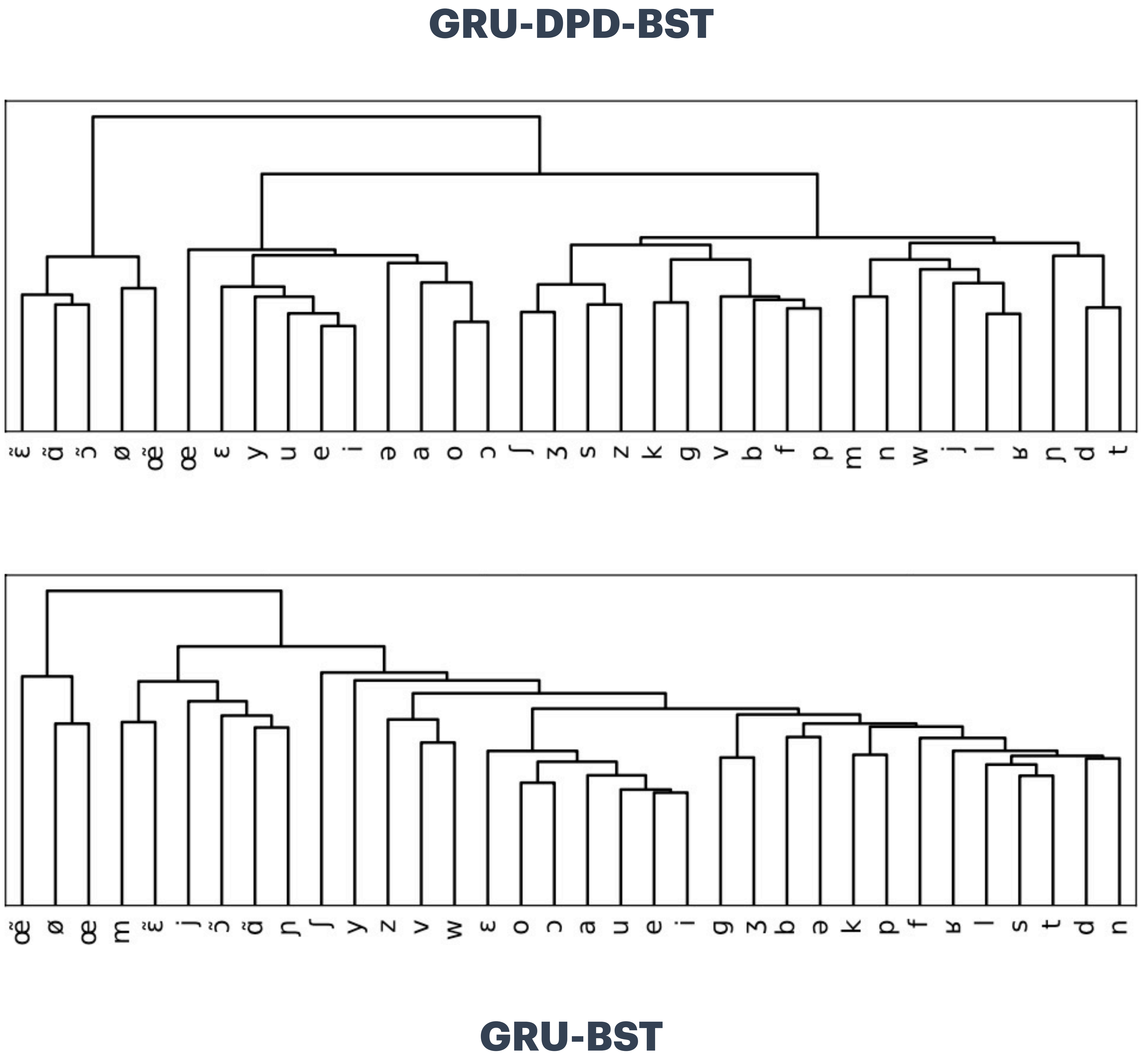


Rom-phon

Analysis

Hierarchical Clustering of Phoneme Embeddings: French

Hierarchical clustering of French phoneme embeddings obtained from the best run (within dataset seed 1 of 10% labeling setting) in the best DPD-based strategy-architecture combination (top) and the best run from their non-DPD counterpart (bottom).



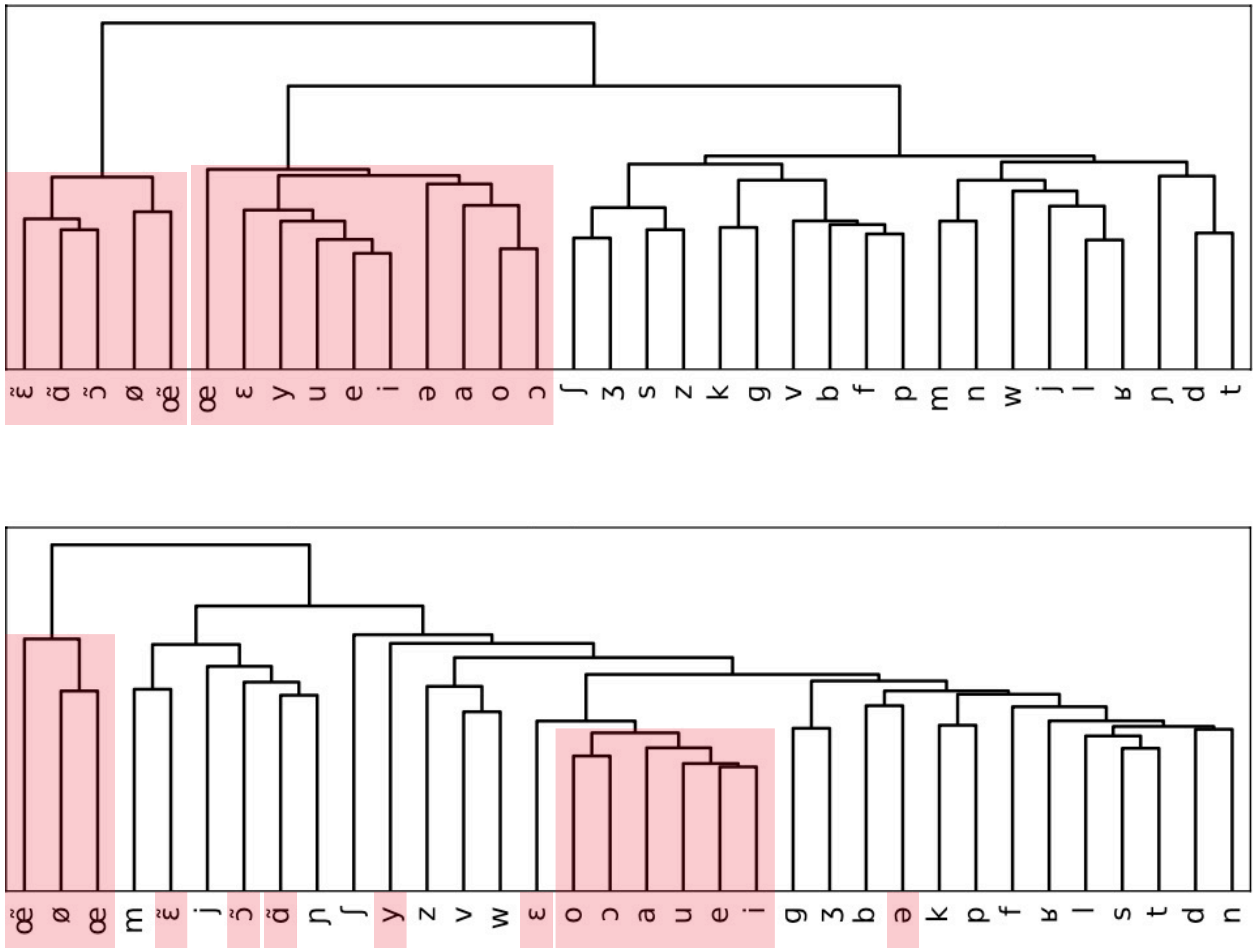
GRU-DPD-BST

GRU-BST

Hierarchical Clustering of Phoneme Embeddings: French

Hierarchical clustering of French phoneme embeddings obtained from the best run (within dataset seed 1 of 10% labeling setting) in the best DPD-based strategy-architecture combination (top) and the best run from their non-DPD counterpart (bottom).

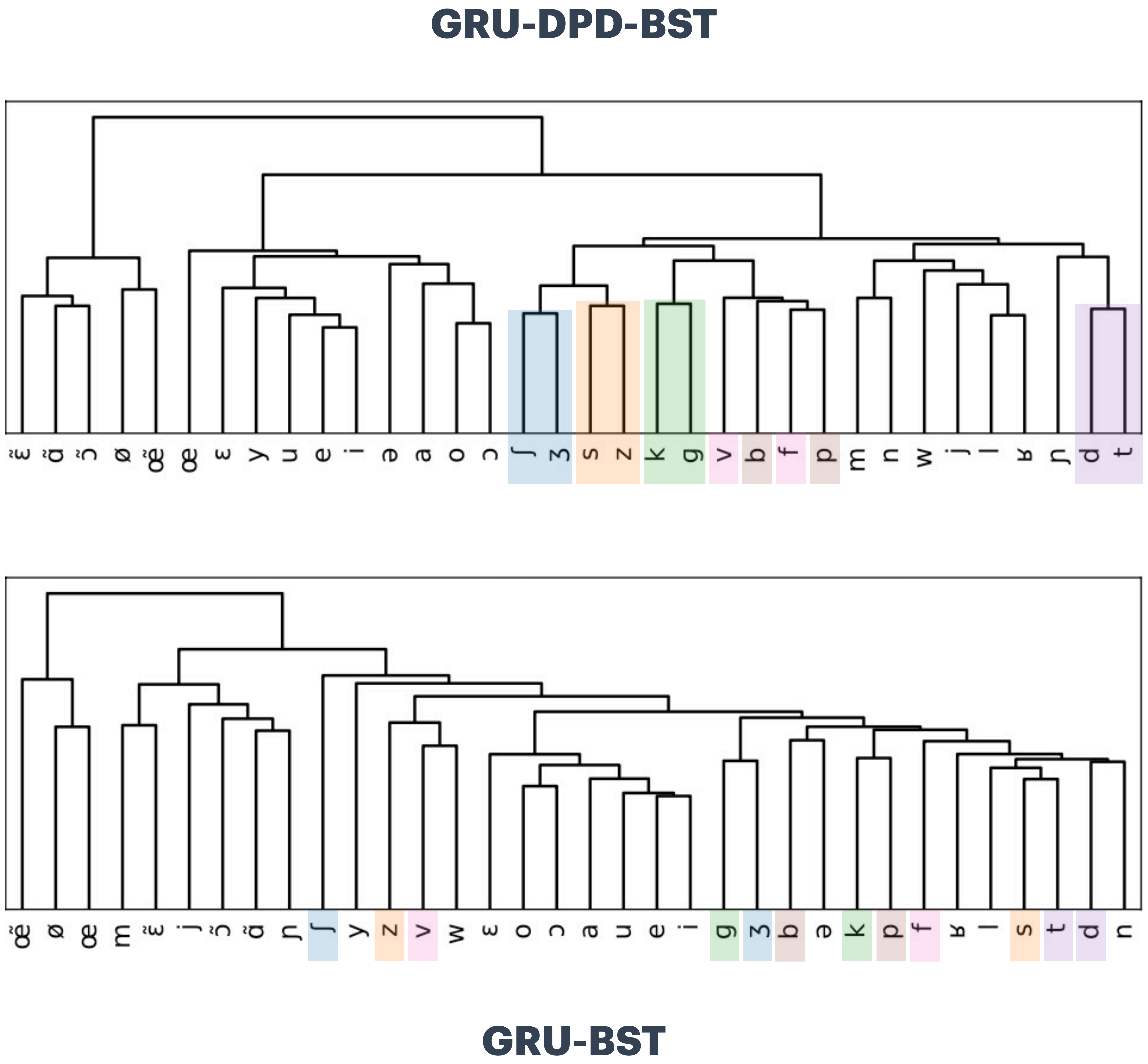
GRU-DPD-BST



GRU-BST

Hierarchical Clustering of Phoneme Embeddings: French

Hierarchical clustering of French phoneme embeddings obtained from the best run (within dataset seed 1 of 10% labeling setting) in the best DPD-based strategy-architecture combination (top) and the best run from their non-DPD counterpart (bottom).

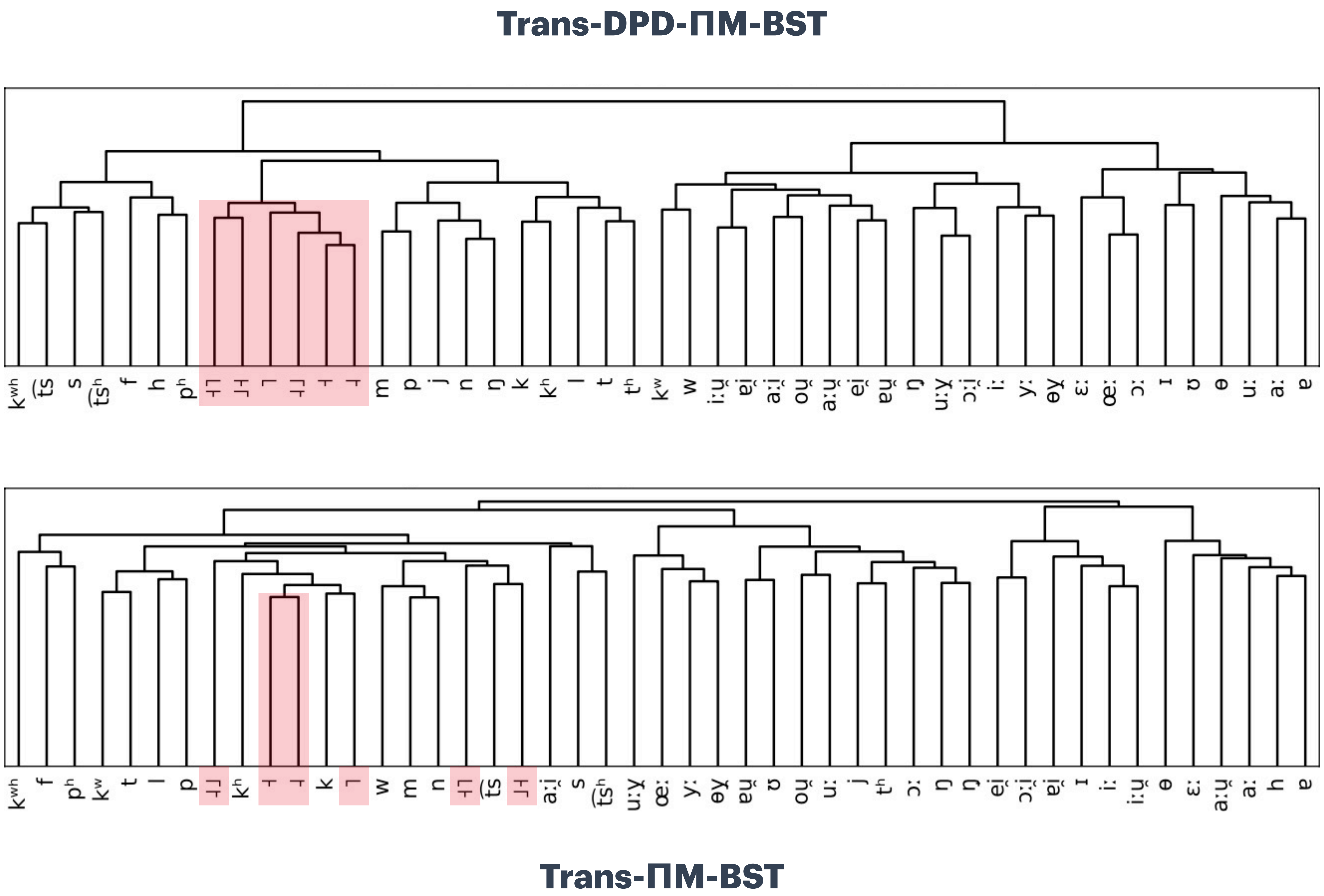


GRU-DPD-BST

GRU-BST

Hierarchical Clustering of Phoneme Embeddings: Cantonese

Hierarchical clustering of Cantonese phoneme embeddings obtained from the best run (within dataset seed 1 of 10% labeling setting) in the best DPD-based strategy-architecture combination (top) and the best run from their non-DPD counterpart (bottom).



Additional Analyses (See Paper)

- The interaction between D2P and P2D during training
- The error patterns of DPD-based vs. non-DPD-based models
- Transductive evaluation of reconstruction performance
- Ablation studies removing the unlabeled data
- Generalizing DPD to supervised reconstruction

Conclusion

Conclusion

We **introduce the new task of semisupervised reconstruction**, marking a step forward toward **building practical computational reconstruction systems** that can assist early-stage proto-language reconstruction projects.

We **design the DPD architecture** to **implement historical linguists' comparative method** and **learn effectively from unlabeled cognate sets**, yielding performance that surpasses existing sequence-to-sequence reconstruction models and established semisupervised learning techniques, especially when protoform labels are scarce.

Links

Links

Paper:	https://arxiv.org/abs/2406.05930 (or conference site)
Code:	https://github.com/cmu-llab/dpd
Checkpoints:	https://huggingface.co/chaosarium/dpd

References

References

Leonard Adolphs, Tianyu Gao, Jing Xu, Kurt Shuster, Sainbayar Sukhbaatar, and Jason Weston. 2022. The CRINGE Loss: Learning what language not to model.

V. S. D. S. Mahesh Akavarapu and Arnab Bhattacharya. 2023. Cognate Transformer for Automated Phonological Reconstruction and Cognate Reflex Prediction. In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, pages 6852–6862.

Enrique Amigó, Julio Gonzalo, Javier Artiles, and Felisa Verdejo. 2009. A comparison of extrinsic clustering evaluation metrics based on formal constraints. Information retrieval, 12:461–486.

Aryaman Arora, Adam Farris, Samopriya Basu, and Suresh Kolichala. 2023. Jambu: A historical linguistic database for South Asian languages. In Proceedings of the 20th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology, pages 68–77, Toronto, Canada. Association for Computational Linguistics.

Siddhant Arora, Siddharth Dalmia, Brian Yan, Florian Metze, Alan W. Black, and Shinji Watanabe. 2022. Token-level Sequence Labeling for Spoken Language Understanding using Compositional Endto-End Models.

Lukas Biewald. 2020. Experiment tracking with weights and biases.

Timotheus A. Bodt and Johann-Mattis List. 2022. Reflex prediction: A case study of Western Kho-Bwa. Diachronica, 39(1):1–38.

Alexandre Bouchard-Côté, Thomas L. Griffiths, and Dan Klein. 2009. Improved Reconstruction of Protolanguage Word Forms. In Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, pages 65–73, Boulder, Colorado. Association for Computational Linguistics.

Alexandre Bouchard-Côté, David Hall, Thomas L. Griffiths, and Dan Klein. 2013. Automated reconstruction of ancient languages using probabilistic models of sound change. Proceedings of the National Academy of Sciences, 110(11):4224–4229.

Alexandre Bouchard-Côté, Percy Liang, Thomas Griffiths, and Dan Klein. 2007a. A Probabilistic Approach to Diachronic Phonology. In Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLPCoNLL), pages 887–896, Prague, Czech Republic. Association for Computational Linguistics.

Alexandre Bouchard-Côté, Percy S Liang, Dan Klein, and Thomas Griffiths. 2007b. A Probabilistic Approach to Language Change. In Advances in Neural Information Processing Systems, volume 20. Curran Associates, Inc.

L. Campbell. 2021. Historical Linguistics: An Introduction. Edinburgh University Press.

Chundra Cathcart and Taraka Rama. 2020. Disentangling dialects: A neural approach to Indo-Aryan historical phonology and subgrouping. In Proceedings of the 24th Conference on Computational Natural Language Learning, pages 620–630, Online. Association for Computational Linguistics.

Kalvin Chang, Chenxuan Cui, Youngmin Kim, and David R. Mortensen. 2022. WikiHan: A New Comparative Dataset for Chinese Languages. In Proceedings of the 29th International Conference on Computational Linguistics, pages 3563–3569, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Yong Cheng, Wei Xu, Zhongjun He, Wei He, Hua Wu, Maosong Sun, and Yang Liu. 2016. SemiSupervised Learning for Neural Machine Translation. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 1965–1974, Berlin, Germany. Association for Computational Linguistics.

Alina Maria Ciobanu and Liviu P. Dinu. 2018. Ab Initio: Automatic Latin Proto-word Reconstruction. In Proceedings of the 27th International Conference on Computational Linguistics, pages 1604–1614, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Alina Maria Ciobanu, Liviu P. Dinu, and Laurentiu Zoicas. 2020. Automatic Reconstruction of Missing Romanian Cognates and Unattested Latin Words. In Proceedings of the Twelfth Language Resources and Evaluation Conference, pages 3226–3231, Marseille, France. European Language Resources Association.

Chenxuan Cui, Ying Chen, Qinxin Wang, and David R Mortensen. Neural Proto-Language Reconstruction.

Sergey Edunov, Myle Ott, Michael Auli, and David Grangier. 2018. Understanding Back-Translation at Scale. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pages 489–500, Brussels, Belgium. Association for Computational Linguistics.

Clémentine Fourrier. 2022. Neural Approaches to Historical Word Reconstruction. Ph.D. thesis, Université PSL (Paris Sciences & Lettres).

Clémentine Fourrier and Benoît Sagot. 2022. Probing multilingual cognate prediction models. In Findings of the Association for Computational Linguistics: ACL 2022, pages 3786–3801.

Çaglar Gülçehre, Orhan Firat, Kelvin Xu, Kyunghyun Cho, Loïc Barrault, Huei-Chi Lin, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2015. On Using Monolingual Corpora in Neural Machine Translation. ArXiv.

Andre He, Nicholas Tomlin, and Dan Klein. 2023. Neural Unsupervised Reconstruction of Protolanguage Word Forms. In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 1636– 1649, Toronto, Canada. Association for Computational Linguistics.

Wilbert Heeringa and Brian Joseph. 2007. The Relative Divergence of Dutch Dialect Pronunciations from their Common Source: An Exploratory Study. In Proceedings of Ninth Meeting of the ACL Special Interest Group in Computational Morphology and Phonology, pages 31–39, Prague, Czech Republic. Association for Computational Linguistics.

Young Min Kim, Kalvin Chang, Chenxuan Cui, and David R. Mortensen. 2023. Transformed Protoform Reconstruction. In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), pages 24– 38, Toronto, Canada. Association for Computational Linguistics.

Samuli Laine and Timo Aila. 2017. Temporal Ensembling for Semi-Supervised Learning.

Dong-Hyun Lee. 2013. Pseudo-Label : The Simple and Efficient Semi-Supervised Learning Method for Deep Neural Networks.

Vladimir I. Levenshtein. 1966. Binary codes capable of correcting deletions, insertions, and reversals. In Soviet Physics Doklady, volume 10, pages 707–710. Soviet Union.

Johann-Mattis List. 2019. Beyond edit distances: Comparing linguistic reconstruction systems. Theoretical Linguistics, 45(3-4):247–258.

Johann-Mattis List and Robert Forkel. 2021. LingPy. A Python Library for Historical Linguistics. Zenodo.

Johann-Mattis List, Robert Forkel, and Nathan Hill. 2022. A New Framework for Fast Automated Phonological Reconstruction Using Trimmed Alignments and Sound Correspondence Patterns. In Proceedings of the 3rd Workshop on Computational Approaches to Historical Language Change, pages 89– 96, Dublin, Ireland. Association for Computational Linguistics.

Yucen Luo, Jun Zhu, Mengxi Li, Yong Ren, and Bo Zhang. 2018. Smooth Neighbors on Teacher Graphs for Semi-Supervised Learning. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 8896–8905.

Clayton Marr and David Mortensen. 2023. Largescale computerized forward reconstruction yields new perspectives in French diachronic phonology. Diachronica, 40(2):238–285.

Clayton Marr and David R. Mortensen. 2020. Computerized Forward Reconstruction for Analysis in Diachronic Phonology, and Latin to French Reflex Prediction. In Proceedings of LT4HALA 2020 1st Workshop on Language Technologies for Historical and Ancient Languages, pages 28–36, Marseille, France. European Language Resources Association (ELRA).

Carlo Meloni, Shauli Ravfogel, and Yoav Goldberg. 2021. Ab Antiquo: Neural Proto-language Reconstruction. In Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 4460–4473, Online. Association for Computational Linguistics.

David R. Mortensen, Patrick Littell, Akash Bharadwaj, Kartik Goyal, Chris Dyer, and Lori Levin. 2016. PanPhon: A Resource for Mapping IPA Segments to Articulatory Feature Vectors. In Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers, pages 3475–3484, Osaka, Japan. The COLING 2016 Organizing Committee.

Yassine Ouali, Céline Hudelot, and Myriam Tami. 2020. An Overview of Deep Semi-Supervised Learning.

Michael Saxon, Samridhi Choudhary, Joseph P. McKenna, and Athanasios Mouchtaris. 2021. Endto-End Spoken Language Understanding for Generalized Voice Assistants. In Interspeech 2021, pages 4738–4742.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Improving Neural Machine Translation Models with Monolingual Data.

Siva Sivaganesan. 1994. An Introduction to the Bootstrap (Bradley Efron and Robert J. Tibshirani). SIAM Review, 36(4):677–678.

Ivan Skorokhodov, Anton Rykachevskiy, Dmitry Emelyanenko, Sergey Slotin, and Anton Ponkratov. 2018. Semi-Supervised Neural Machine Translation with Language Models. In Proceedings of the AMTA 2018 Workshop on Technologies for MT of Low Resource Languages (LoResMT 2018), pages 37–44, Boston, MA. Association for Machine Translation in the Americas.

Szemerényi, O. J. L. (1996). Introduction to Indo-European Linguistics. Oxford University Press UK.

Kuen-Han Tsai and Hsuan-Tien Lin. 2019. Learning from Label Proportions with Consistency Regularization. ArXiv.

Joe H. Ward, Jr. 1963. Hierarchical Grouping to Optimize an Objective Function. Journal of the American Statistical Association, 58(301):236–244.

Frank Wilcoxon. 1992. Ranking Methods. pages 196–202.