Prototype-Based Label Propagation for Zero-Shot Histopathology Segmentation with Vision-Language Models

Bingan Mu[1] and Yuhao Yi[1*]

College of Computer Science, Sichuan University

Motivation: Zero-shot WSI segmentation needs slide context

A tile should not decide alone.

Frozen pathology VLMs can score tiles against text prompts, but direct tile-wise inference does not leverage the structural cues already present in unlabeled WSIs.

No fine-tuning

The encoder and prompts stay fixed at test time.

Gigapixel scale

A global tile graph is computationally impractical.

Tissue structure

Neighboring morphology should support consistent masks.

Methodology: Prototype-Based Label Propagation

1

Morphological Prototype Generation

Compress unlabeled reference tiles into **K** morphology prototypes, reducing graph size while preserving tissue patterns.

2

Global Inductive Label Propagation

Build a class-prototype-tile graph and infer soft scores for each query tile without updating the frozen VLM.

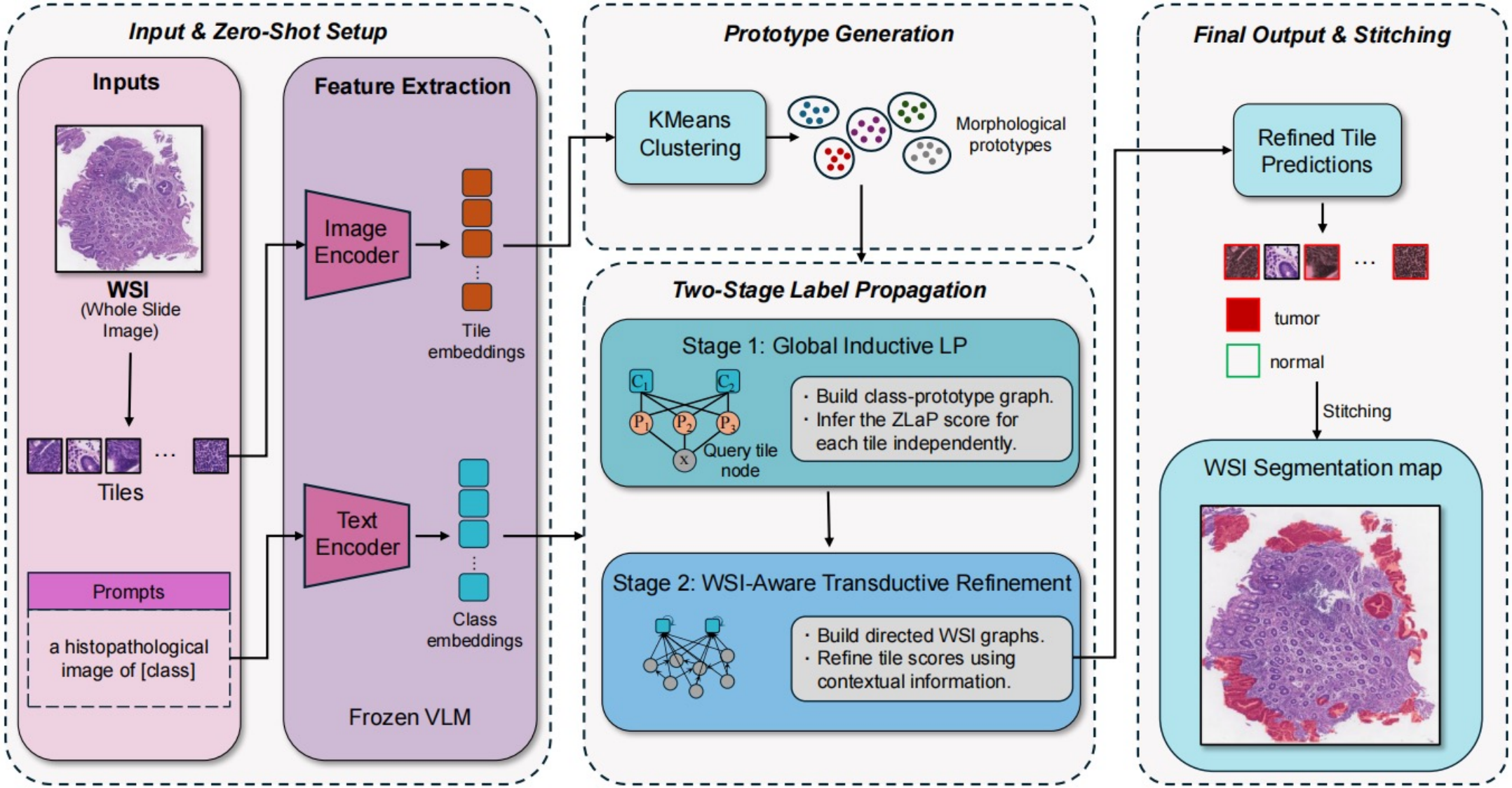
3

WSI-Aware Transductive Refinement

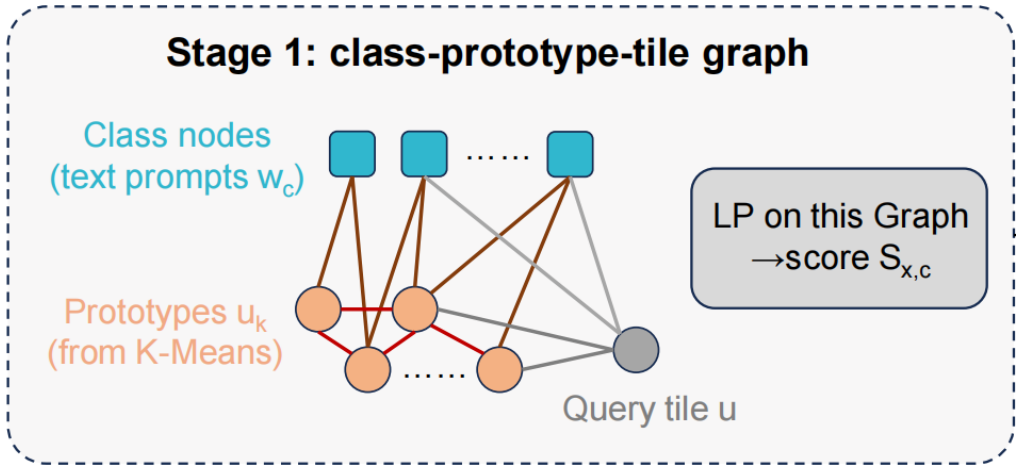
Construct a tile-to-tile graph within each WSI to propagate local evidence and improve mask consistency.

Takeaway: keep the VLM black-box, but use graph structure to make its predictions less isolated.

Overall Framework: encode, compress, propagate, stitch



Stage 1: Global Inductive Label Propagation



Prototype graph

Reference tiles are compressed into K=50 morphological prototypes.

Inductive scoring

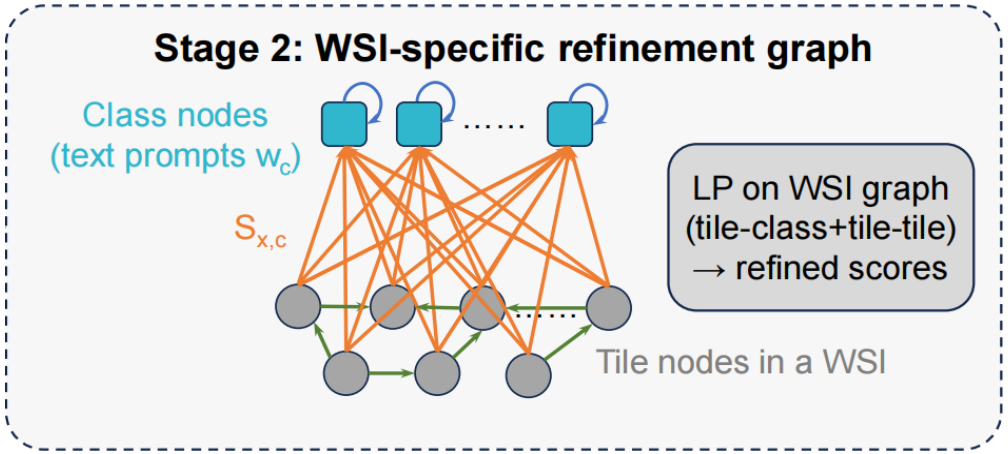
Each query tile connects to class nodes and nearest prototypes.

Reusable graph inference

The class-prototype graph is solved once and reused for query tiles.

Output: tile-level soft class scores Sx,c

Stage 2 restores within-slide consistency



How it works

1. Stage 1 soft scores initialize the WSI graph.
2. Each tile links to its top-m most similar tiles in the same WSI.
3. A second LP step smooths ambiguous regions without changing the VLM.

Refinement is transductive within each slide; it does not leak labels or tune parameters.

Datasets & Experimental Setup

Dataset	Task	Tiles	Reconstruction
CAMELYON16	Binary tumor	512 x 512 at 40x	Non-overlap
DigestPath	Binary colorectal	224 x 224 at 20x	75% overlap
SICAPv2	Binary + multiclass	224 x 224 at 10x	75% overlap

Frozen model

CONCH image/text encoders; 512-dimensional embeddings.

Fixed prompts

Template: a histopathological image of {class}.

Metrics

Micro-averaged Dice, precision, recall with WSI-level 5-fold splits.

Dice Results

Dice score; higher is better.

Method	DigestPath	C16	SICAPv2 Bin	SICAPv2 Multi
Zero-shot tiles	0.566	0.759	0.669	0.330
Prototype center	0.568	**0.898**	**0.713**	0.293
GMM	0.555	0.888	0.695	0.332
Ours Stage 1	0.595	0.890	0.707	**0.340**
Ours Stage 2	**0.597**	**0.895**	**0.708**	**0.338**

DigestPath Dice

0.568 → 0.597

prototype center to Stage 2

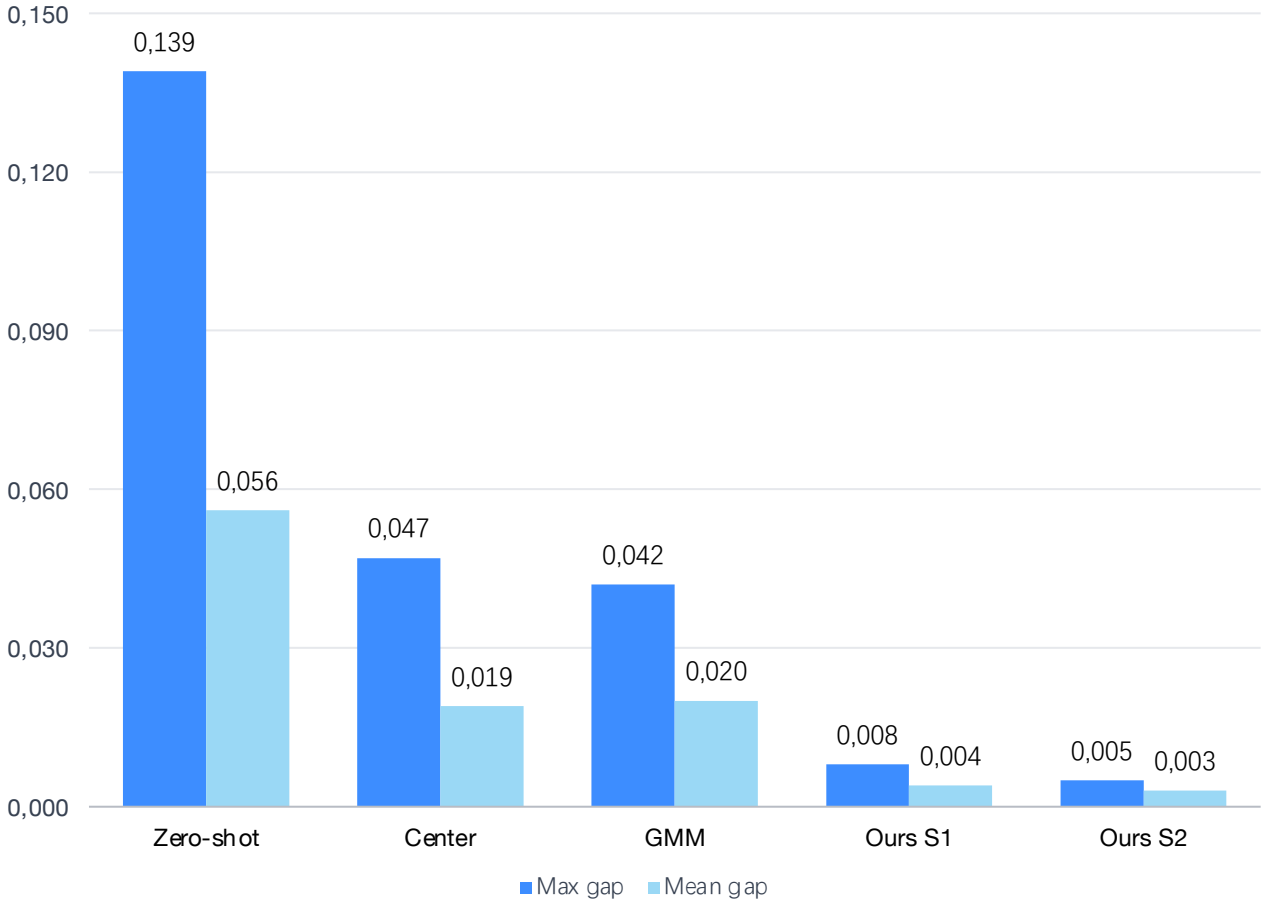
Interpretation

Propagation helps most when single-prototype labels are too brittle; when clusters are already separable, it remains close to the best baseline.

Stage 2 is best or near-best across all zero-shot tasks

Two-Stage LP Has the Smallest Worst-Case Gap

Gap-to-best Delta Dice across four zero-shot tasks; lower is better.



Compared with direct zero-shot tile prediction, Stage 2 reduces the worst-case gap from 0.139 to 0.005 and the mean gap from 0.056 to 0.003.

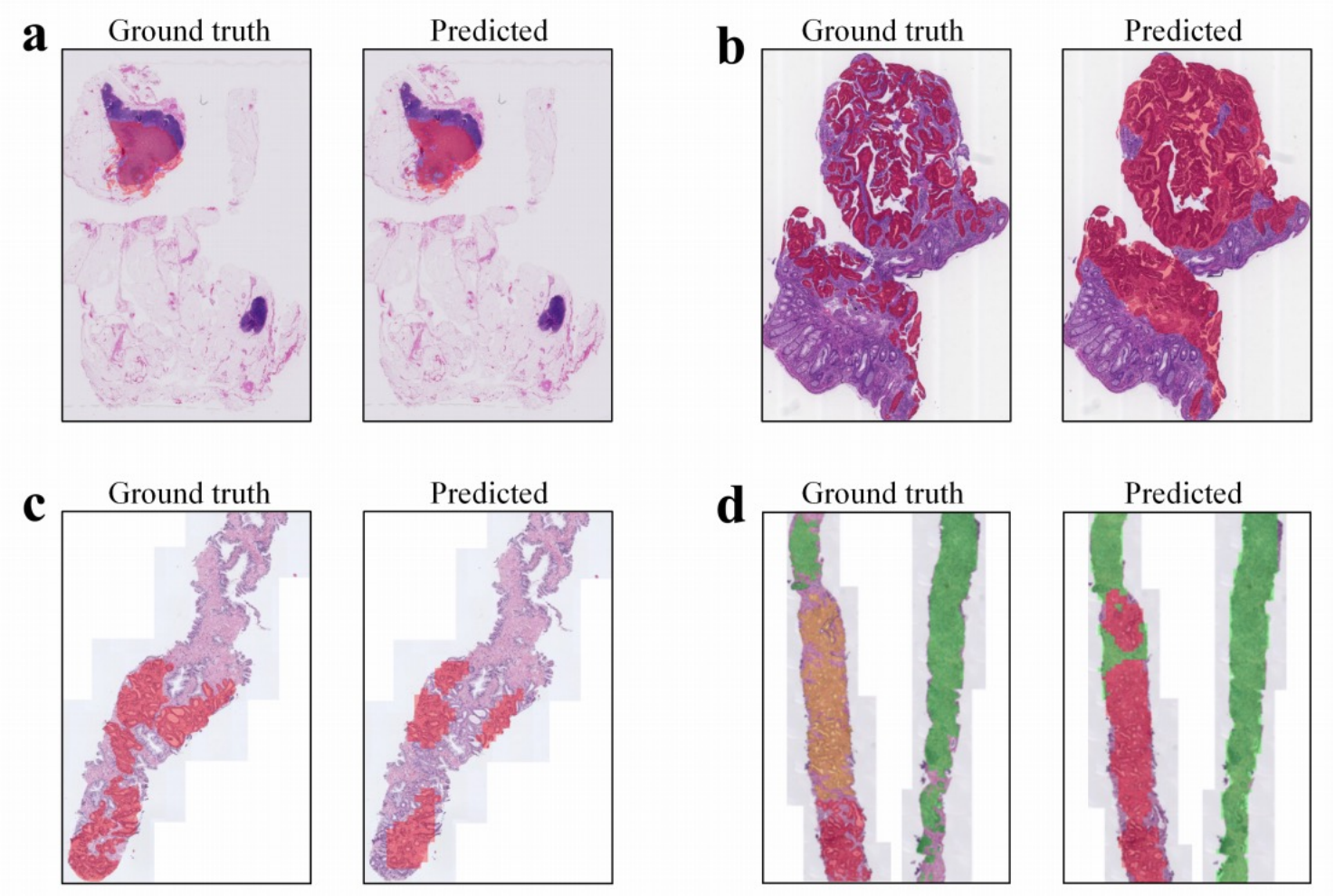
Worst-case ΔDice

0.139 → 0.005

Mean ΔDice

0.056 → 0.003

Qualitative WSI Segmentation Maps



Binary Tasks (a-c)

Red: positive class region in ground truth and prediction.
Uncolored: benign tissue and background.
Predictions largely follow the ground-truth tissue contours.

Multiclass SICAPv2 (d)

Red: well differentiated adenocarcinoma.

Orange: moderately differentiated adenocarcinoma.

Green: poorly differentiated adenocarcinoma.

Key Takeaways

Thanks!

2024223045015@stu.scu.edu.cn

Key Takeaways：

- Frozen VLM, no task-specific finetuning.
- Prototype compression enables scalable label propagation.
- Intra-WSI refinement produces coherent and robust segmentation masks.

Backup: full precision and recall metrics

Table 1. Zero-shot WSI level segmentation results on DigestPath and C16 (binary). Best in bold, second best underlined. Dice is reported as *score* (Δ), where Δ is the gap to the best Dice on that dataset (lower is better).

	DigestPath			C16		
Method	Dice (Δ)	Precision	Recall	Dice (Δ)	Precision	Recall
Zero-shot tiles baseline	0.566 (0.031)	0.605	0.680	0.759 (0.139)	0.696	0.908
Prototype nearest center	0.568 (0.029)	0.596	0.663	**0.898 (-)**	**0.878**	0.921
GMM baseline	0.555 (0.042)	0.603	0.676	0.888 (0.010)	0.850	0.935
Ours (Stage 1)	0.595 (0.002)	0.631	0.723	0.890 (0.008)	0.858	0.930
Ours (Stage 2)	**0.597 (-)**	**0.635**	**0.731**	0.895 (0.003)	0.863	**0.936**

Table 2. Zero-shot WSI level segmentation results on SICAPv2 (binary and multi-class). Best in bold, second best underlined. Dice is reported as *score* (Δ), where Δ is the gap to the best Dice within each task (lower is better).

	SICAPv2 (Binary)			SICAPv2 (Multi-class)		
Method	Dice (Δ)	Precision	Recall	Dice (Δ)	Precision	Recall
Zero-shot tiles baseline	0.669 (0.044)	0.665	0.795	0.330 (0.010)	0.397	0.560
Prototype nearest center	**0.713 (-)**	**0.696**	0.833	0.293 (0.047)	0.341	0.439
GMM baseline	0.695 (0.018)	0.685	0.826	0.332 (0.008)	**0.464**	0.565
Ours (Stage 1)	0.707 (0.006)	0.693	0.834	**0.340 (-)**	0.408	0.604
Ours (Stage 2)	0.708 (0.005)	0.694	**0.835**	0.338 (0.002)	0.411	**0.606**

Backup evidence from Tables 1-2