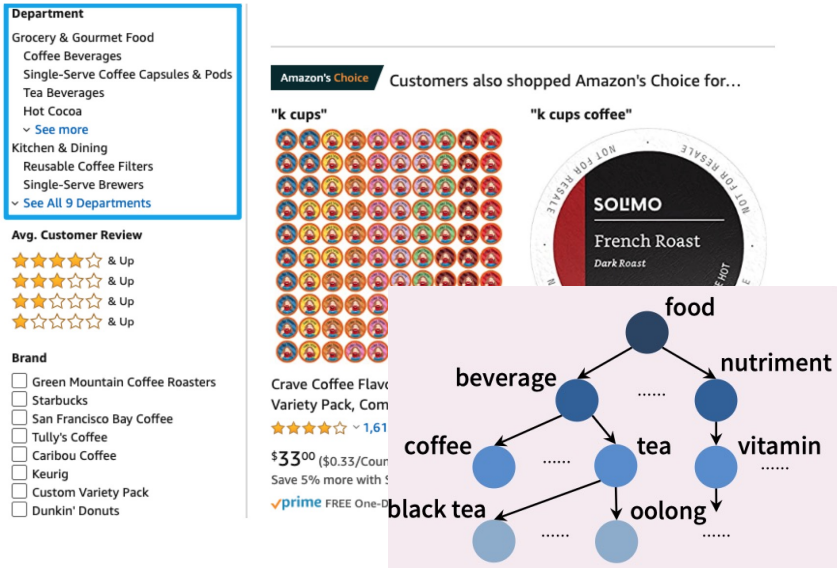
HyperExpan: Taxonomy Expansion with Hyperbolic Representation Learning

Mingyu Derek Ma, Muhao Chen*, Te-Lin Wu*, Nanyun (Violet) Peng

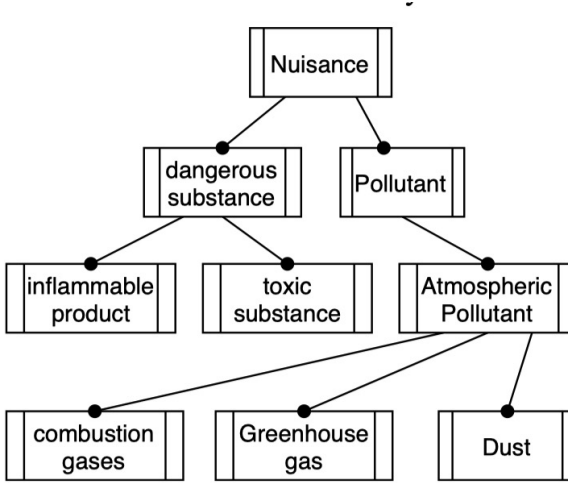



What is a taxonomy

Online catalog taxonomy



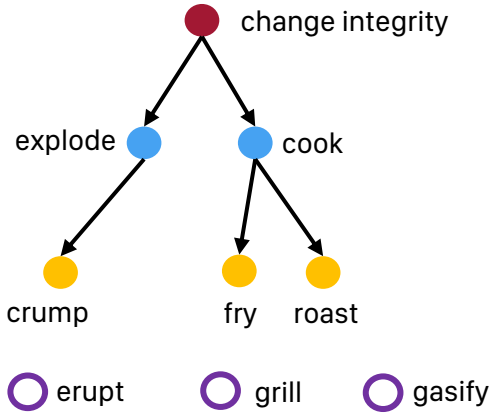
Scientific taxonomy



Wang et al., 21; Mao et al., KDD'20; Yu et al., KDD'20

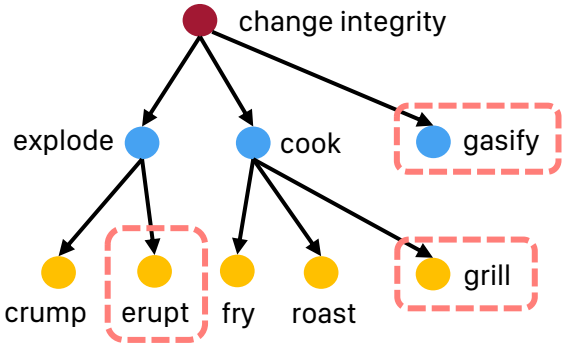
Introduction

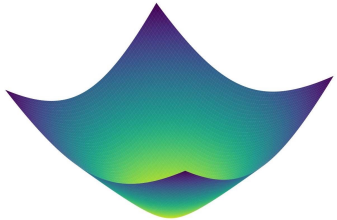
- Taxonomy curation is expensive and suffers from limited coverage
- Our task: taxonomy expansion
  - Attach new concept to an existing taxonomy

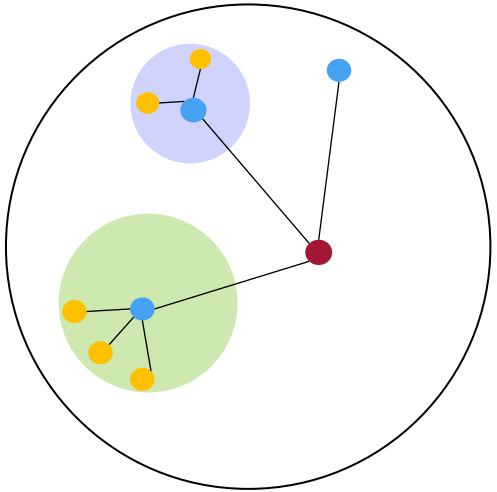


Introduction

- Taxonomy curation is expensive and suffers from limited coverage
- Our task: taxonomy expansion
  - Attach new concept to an existing taxonomy
- Taxonomy size grows exponentially
- Hyperbolic space can better capture lower-level concepts with better expressiveness





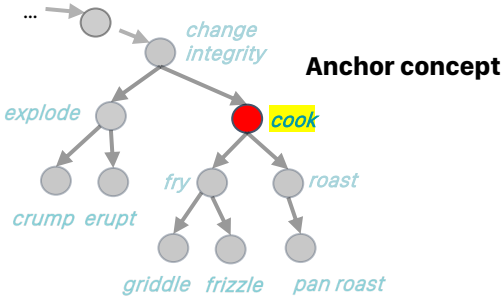


HyperExpan

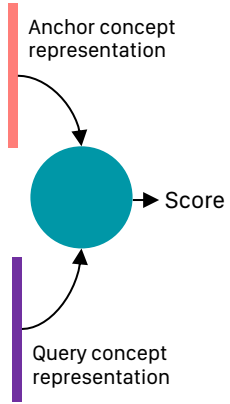
A taxonomy expansion framework based on hyperbolic representation learning

1. Better preserves the taxonomical structure in a **more expressive hyperbolic space**
2. Characterizes concepts by exploiting sparse **neighborhood information**
3. Improves inference precision and generalizability by leveraging **pretrained distributional features**

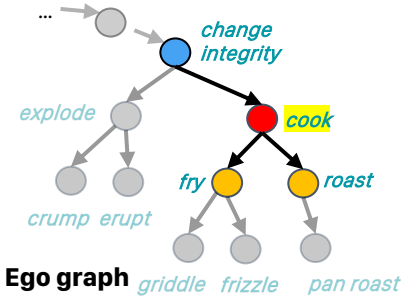
Model design



Query concept *grill*

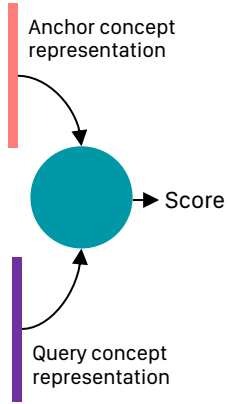


Model design

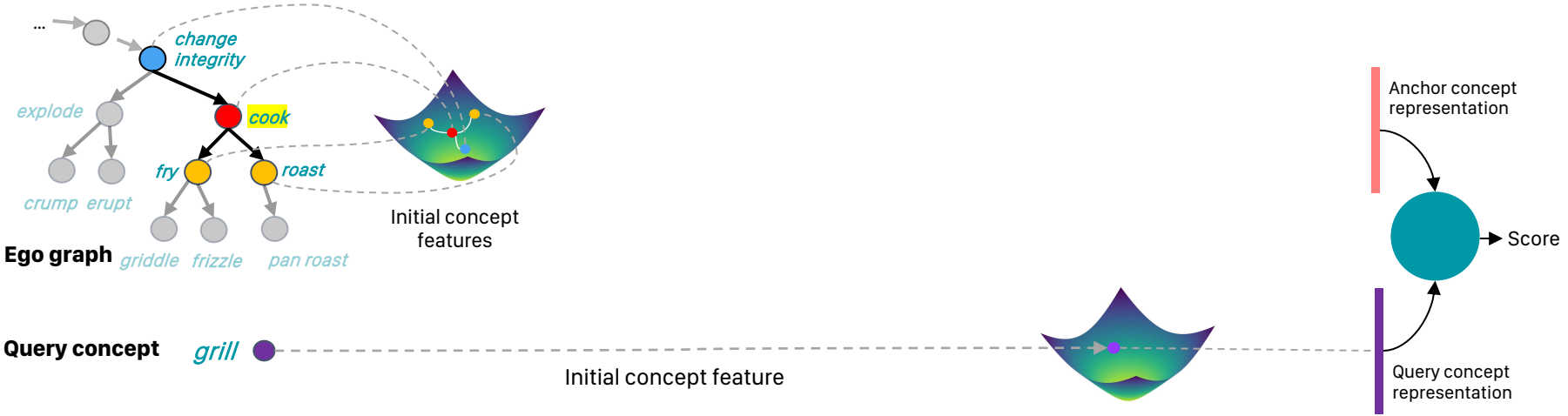


Ego graph

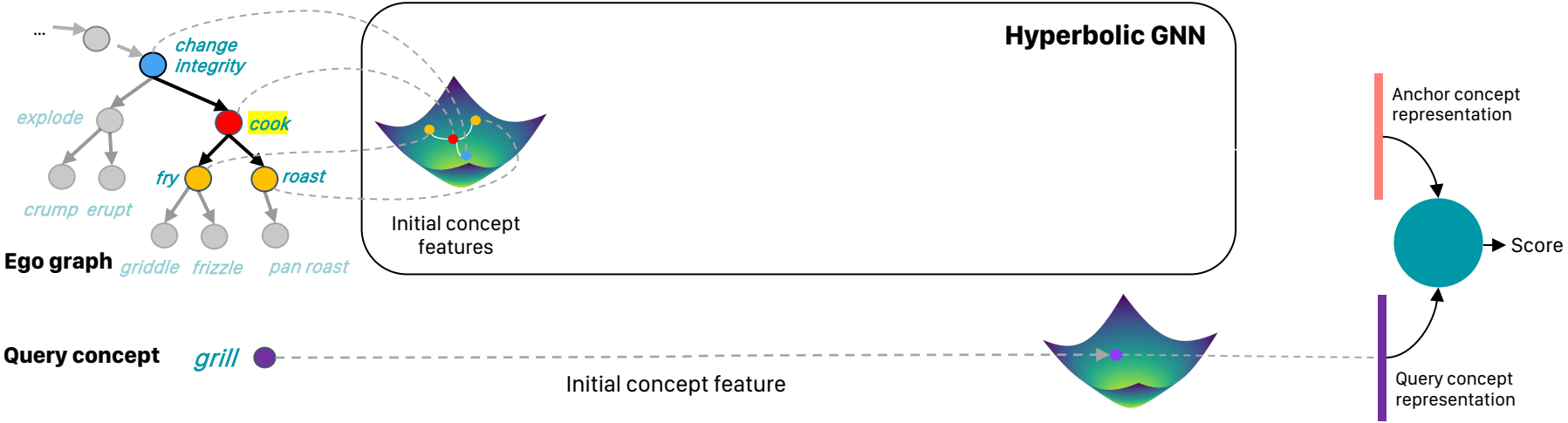
Query concept *grill*



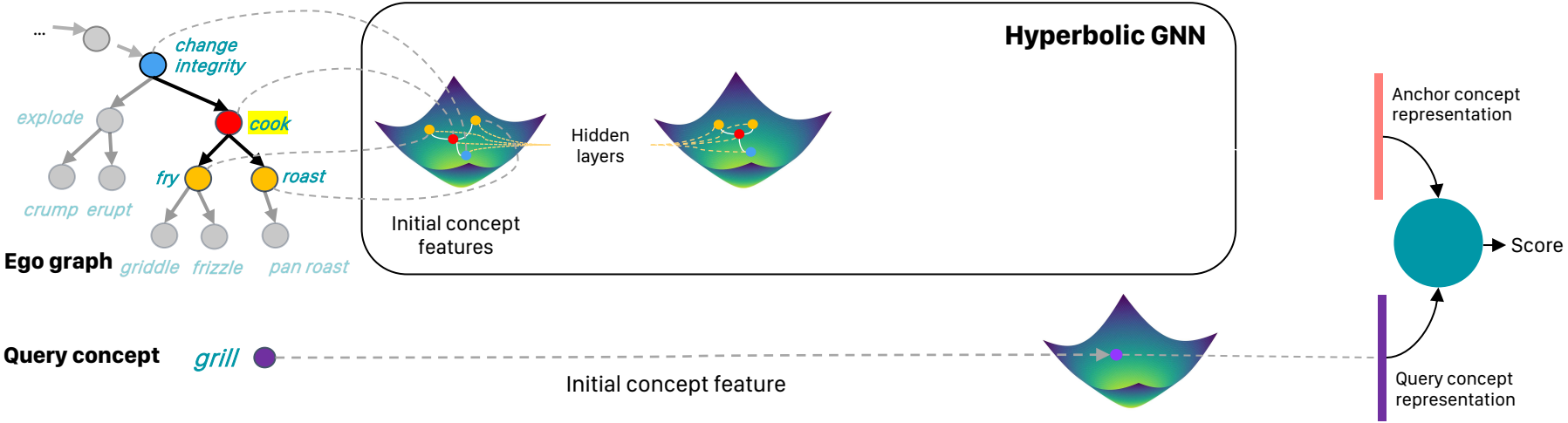
Step 1: initial concept features



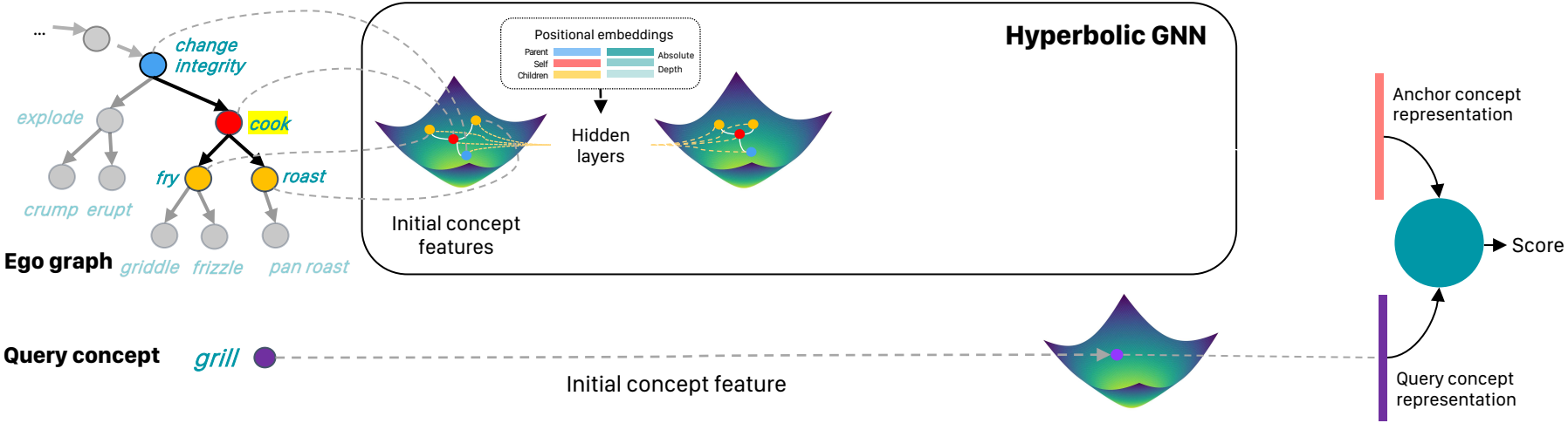
Step 2: anchor concept representation



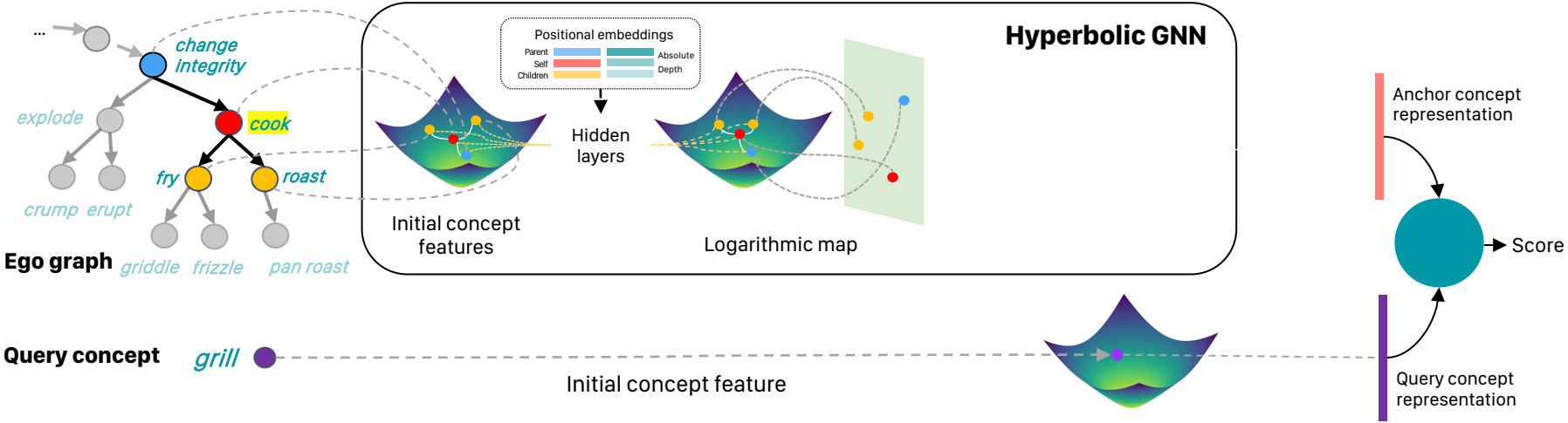
Step 2: anchor concept representation



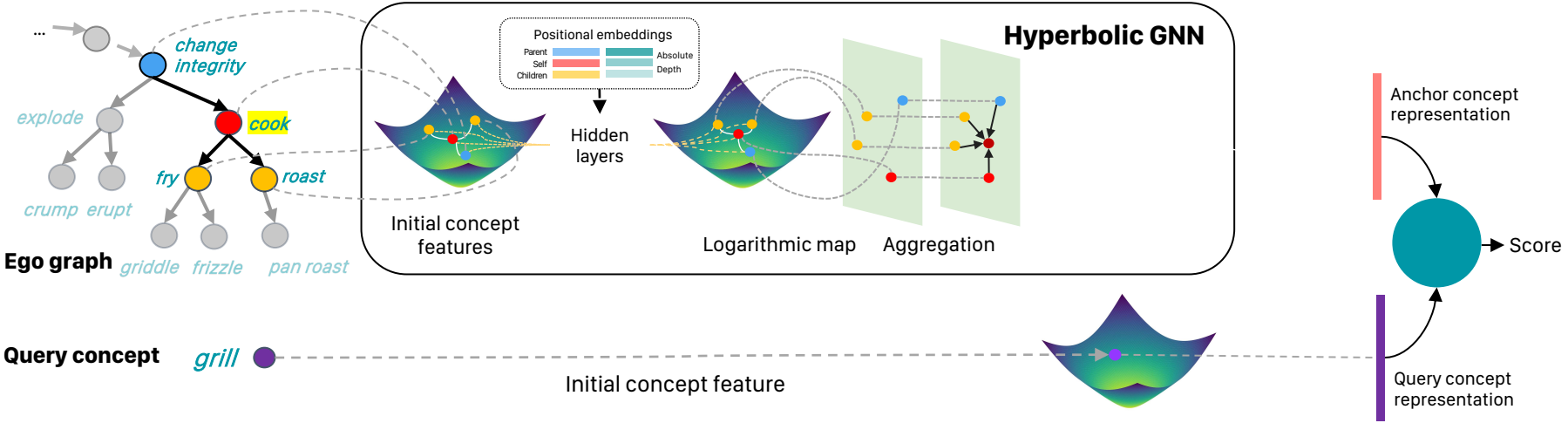
Step 2: anchor concept representation



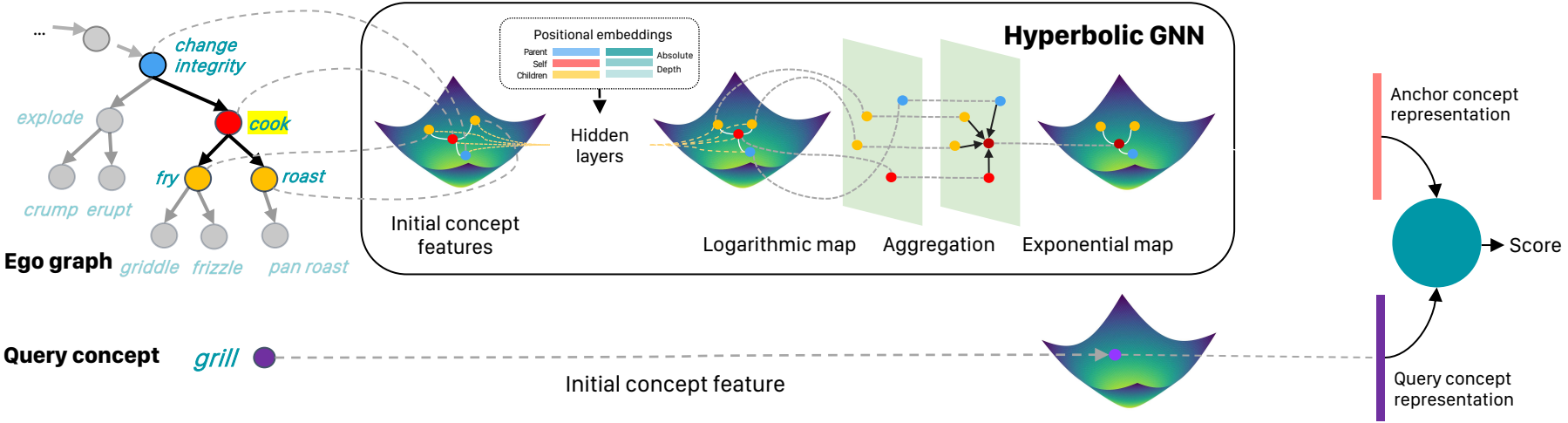
Step 2: anchor concept representation



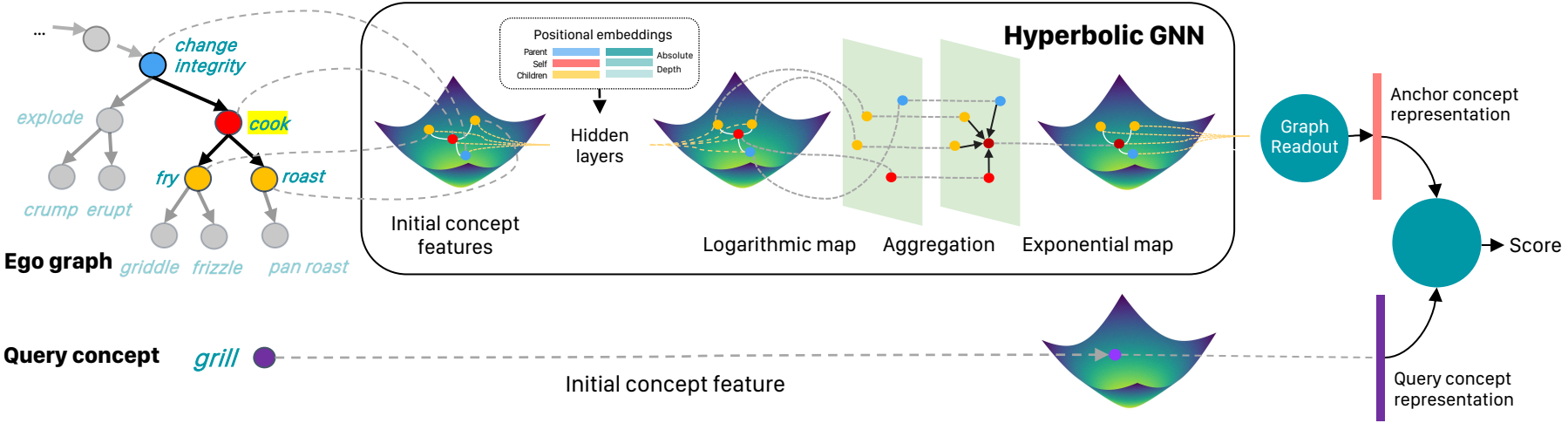
Step 2: anchor concept representation



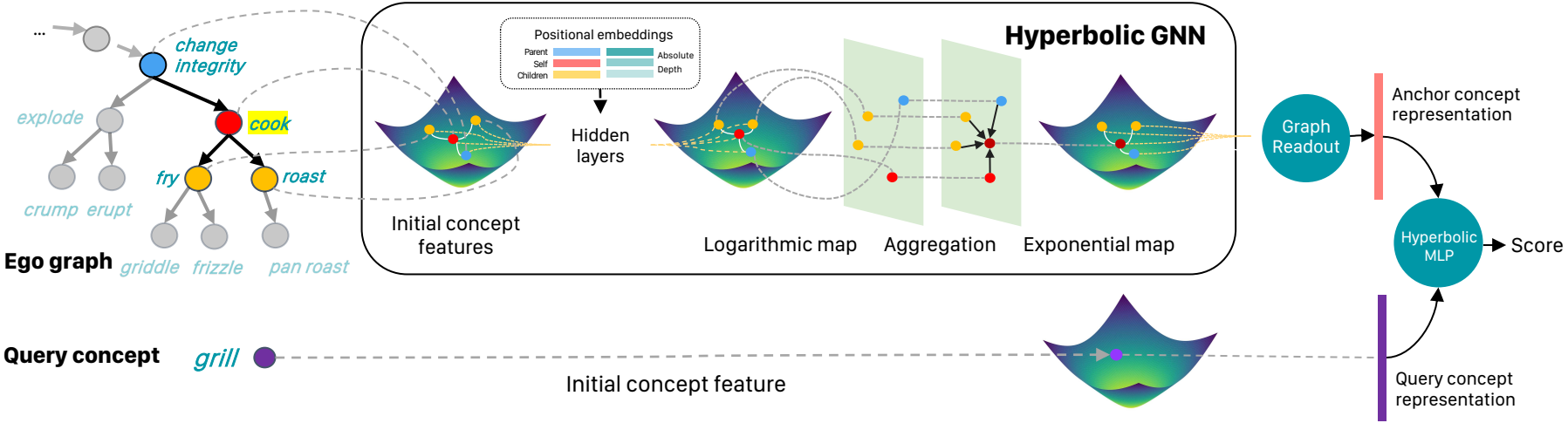
Step 2: anchor concept representation



Step 2: anchor concept representation



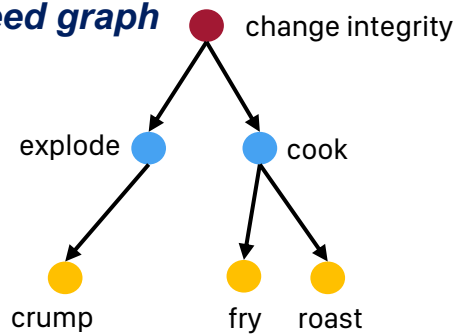
Step 3: matching module



Learning and inference

- Training
  - Self-supervision: positive + negative pairs

Training/seed graph



Self-supervised data points

🙂 roast, cook

🙃 change integrity, cook
🙃 explode, cook
🙃 crump, cook

…

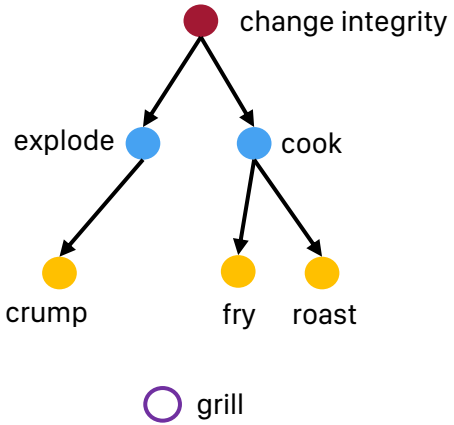
  - Loss function

$$\mathcal{L}(\Theta) = -\frac{1}{|\mathbb{X}|} \sum_{\mathbf{X}_i \in \mathbb{X}} \left[\log \frac{f(n_p, n_c)}{\sum_{\langle n_j, n_c \rangle \in \mathbf{X}_i} f(n_j, n_c)} \right]$$

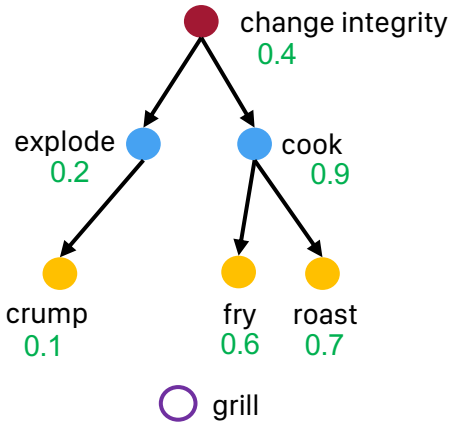
Learning and inference

- Inference

Query node pending to attach



Calculate matching scores with candidates



Ranking list

cook
roast
fry
change integrity
…

Experiments

Model	MR ↓	MRR ↑	Recall % ↑ @1	Recall % ↑ @5	Recall % ↑ @10	Precision % ↑ @1	Precision % ↑ @5	Precision % ↑ @10	MR ↓	MRR ↑	Recall % ↑ @1	Recall % ↑ @5	Recall % ↑ @10	Precision % ↑ @1	Precision % ↑ @5	Precision % ↑ @10
	WordNet-Verb (Candidates #: 11,936)								WordNet-Noun (Candidates #: 81,073)							
ARBORIST	608.7	0.280	10.8	24.0	27.7	6.7	4.8	3.2	1095.1	0.435	16.5	28.4	34.1	16.8	5.8	3.5
TaxoExpan	502.8	0.439	12.4	28.2	35.2	12.4	5.6	3.5	649.6	0.562	19.7	38.2	47.4	20.1	7.8	4.8
TMN	465.0	0.479	14.9	31.6	37.9	13.2	6.4	4.0	**501.0**	0.595	20.7	40.5	50.1	21.1	8.3	5.1
GCN	456.9	0.445	10.9	27.2	34.5	10.9	5.4	3.5	684.1	0.563	20.9	39.8	47.3	21.3	8.1	4.8
GAT	471.7	0.449	11.6	28.7	35.6	11.6	5.7	3.6	640.7	0.585	22.3	40.9	49.7	22.7	8.3	5.1
HYPEREXPAN	**400.8**	**0.517**	**15.0**	**32.8**	**42.7**	**15.0**	**6.6**	**4.3**	573.6	**0.607**	**23.9**	**42.1**	**52.5**	**24.4**	**8.6**	**5.4**
	MAG-PSY (Candidates #: 21,187)								MAG-CS (Candidates #: 22,754)							
ARBORIST	119.9	0.722	21.0	48.4	62.9	25.8	12.5	7.7	284.7	0.602	15.1	38.9	49.4	24.6	12.6	8.0
TaxoExpan	68.5	0.775	26.1	56.9	69.5	33.8	14.7	9.0	189.8	0.661	15.9	42.9	55.4	25.8	13.9	9.0
TMN	73.0	0.781	25.8	58.7	70.5	33.4	15.2	9.1	160.5	0.667	16.0	43.1	56.3	26.0	14.0	9.1
GCN	51.4	0.742	23.8	52.5	64.3	30.8	13.6	7.4	90.3	0.653	14.5	39.6	53.3	23.6	12.9	8.7
GAT	48.6	0.751	23.6	52.4	65.8	30.5	13.5	8.5	92.2	0.676	15.9	41.9	56.0	25.9	13.6	9.1
HYPEREXPAN	**38.4**	**0.827**	**28.8**	**63.0**	**75.3**	**37.2**	**16.3**	**9.7**	**74.4**	**0.689**	**16.1**	**44.6**	**58.0**	**26.1**	**14.5**	**9.4**

- HyperExpan get large performance increase compared with GCN and GAT due to expressiveness of the hyperbolic space
- HyperExpan outperforms previous SOTA TMN

Ablation study

Model	MRR ↑	Rec ↑ @10	Prec ↑ @1
w/o trainable curvature	0.490	40.8	14.4
anchor + parent + children	0.506	42.2	15.0
#4 + anchor's ancestors	0.505	42.5	15.5
#5 + anchor's descendants	0.517	42.7	15.0
#6 + anchor's siblings	0.502	41.7	14.5
w/o Relative Pos Emb	0.497	40.8	13.0
w/o Absolute Pos Emb	0.503	41.2	14.3
w/o both Positional Emb	0.482	38.8	12.5
HYPEREXPAN	0.517	42.7	15.0

MRR is scaled by 10,
i/o means instead of, w/o means without

- Trainable curvature leads fine-grained manifold setting
- Adding descendant or ancestors of the anchor node is helpful, anchor's sibling nodes are not
- Positional embeddings are helpful

Conclusion

- HyperExpan: a taxonomy expansion model which better preserves the taxonomical structure in an expressive hyperbolic space
- Use HGNN to incorporate neighborhood information and positional features of concepts
- Experimental results show that HyperExpan performs better than its Euclidean counterparts and achieves the state-of-the-art

Thank You

HyperExpan: Taxonomy Expansion with Hyperbolic Representation Learning

Mingyu Derek Ma, Muhao Chen*, Te-Lin Wu*, Nanyun (Violet) Peng

Code available at:
github.com/PlusLabNLP/HyperExpan