PSI-KT: Predictive, Scalable and Interpretable Knowledge Tracing on Structured Domains

Hanqi Zhou[1,2,3,4], Robert Bamler[1,2,3], Charley M. Wu[1,2,3], Álvaro Tejero-Cantero[1,2]

[1] University of Tübingen [2] Cluster of Excellence Machine Learning [3] Tübingen AI Center [4] IMPRS-IS

TL;DR

We propose a **high-performant**, **scalable** model to track learners' mastery levels with the **interpretability** required for personalized education, essential for next-generation intelligent tutoring systems.

Why do we care?

What should we learn, and when to practice?

To create personalized and effective curricula, we need:

1. *Knowledge Tracing*: What do learners already know?
2. *Knowledge Mapping*: What are suitable contexts to present content, i.e. what are prerequisites?

Existing models of human learning just can't keep up!

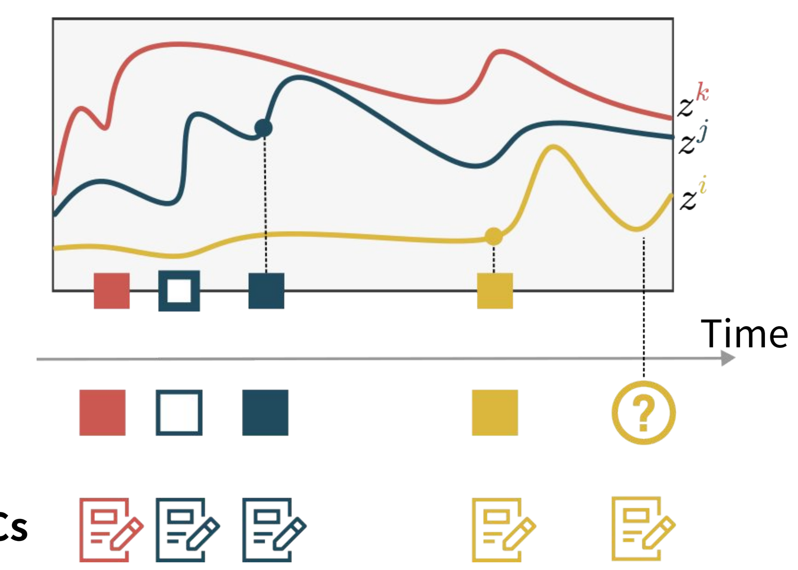
Cognitive models — Explicit forgetting & learning; Flexibility

scalability

Neural networks — Model capacity; Interpretability

Tracing knowledge over time

Knowledge Tracing (KT)[1] aims to estimate a learner's knowledge states and predict future performance given the learning interaction history.

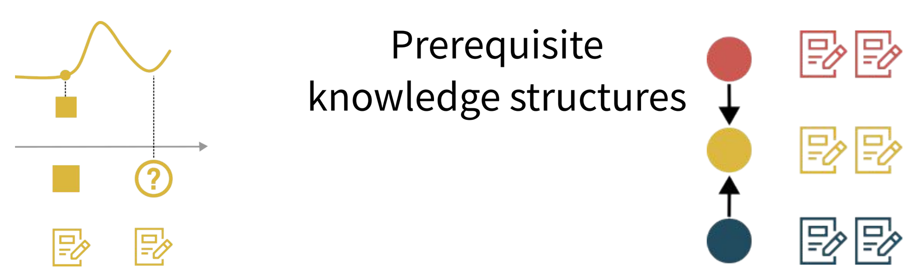


Input: Learning history $\mathcal{H}^{\ell}_{1:N} := (x_n, t_n, y_n)^{\ell}_{1:N}$

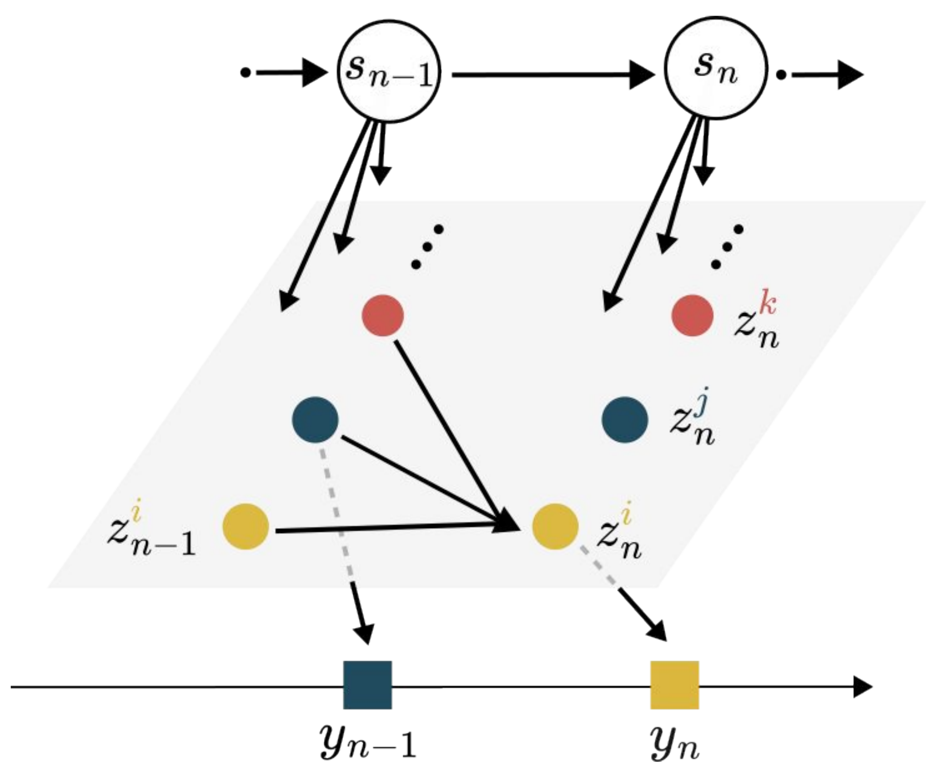
- Knowledge component (KC) x_n, e.g. pythagorean theorem
- The timestamp of the interaction t_n
- An evaluation of the learner's performance y_n

Output: Prediction of the learner's performance and, optionally, the structure of KCs.

Two key components we cared in modeling human learning in structured domains:



PSI-KT: a hierarchical model from the ground up



Per learner $s^{\ell}_n := (\alpha^{\ell}_{t_n}, \mu^{\ell}_{t_n}, \gamma^{\ell}_{t_n})$ for personalization

α^{ℓ}_n forgetting rate; μ^{ℓ}_n consolidated memory; γ^{ℓ}_n transfer ability. The Markovian evolutio is: a Kalman filtering prior.

Per learner & KC $z^{\ell,k}_n$ for memory dynamics

(1) Global structure $\mathcal{A}$ for knowledge prerequisites

The consolidated memory is shifted by the inferred structure:

$$\tilde{\mu}^{\ell,k}_n := \mu^{\ell}_n + \gamma^{\ell}_n \sum_{i\neq k} a_{ik} z^{\ell,i}_n$$

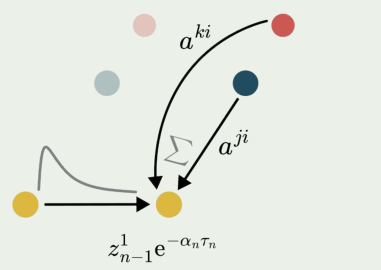
(2) Ornstein-Uhlenbeck process for temporal evolution

$$\mathrm{d}z^{\ell,k}/\mathrm{d}t = \alpha^{\ell}\left(\mu^{\ell} - z^{\ell,k}\right) + \sigma^{\ell}\eta(t)$$

Overall: Structure-aware Ornstein-Uhlenbeck process

The transition mean[2]:

$$m^{\ell,k}_n = \underbrace{\tilde{\mu}^{\ell,k}_n\left(1 - \exp\left(-\alpha^{\ell}_n \tau^{\ell}_n\right)\right)}_{\text{long-term dynamics}} + \underbrace{m^{\ell,k}_{n-1}\exp\left(-\alpha^{\ell}_n \tau^{\ell}_n\right)}_{\text{transient dynamics}}$$



Variational inference: We use amortized variational inference, selecting a distribution family q_ϕ with free parameters ϕ to approximate the posterior p_θ by minimizing their Kullback-Leibler divergence.

$$\mathrm{ELBO}(\theta,\phi) = \mathbb{H}\left(q_\phi\left(z_{1:n}, s_{1:n} \mid y_{1:n}\right)\right) + \mathbb{E}_{q_\phi(z_{1:n}, s_{1:n} \mid y_{1:n})} \log p_\theta\left(y_{1:n}, z_{1:n}, s_{1:n}\right)$$

Interpretability of traits and graphs

Inferred cognitive traits are *specific* to each learner, *consistent* across data splits, *disentangled* (i.e., component-wise meaningful) and *operationally interpretable.*

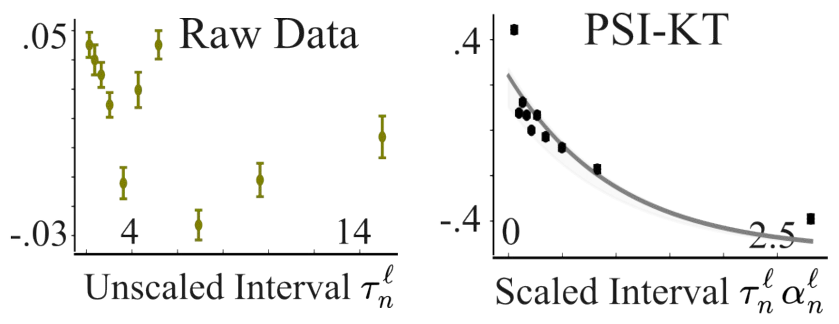
Representation capacity:

Metric	Dataset	Baseline	PSI-KT
Specificity $\mathrm{MI}(s;\ell)\uparrow$	Assist12	**8.8**	8.4
	Assist17	**10.1**	10.0
	Junyi15	13.5	**14.4**
Consistency^{-1} $\mathbb{E}_{\ell_{\text{sub}}}\mathrm{MI}(s^{\ell};\ell_{\text{sub}})\downarrow$	Assist12	12.2	**7.4**
	Assist17	**6.4**	**6.4**
	Junyi15	7.7	**5.0**
Disentanglement $D_{\mathrm{KL}}(s\|\ell)\uparrow$	Assist12	2.3	**7.4**
	Assist17	0.6	**8.4**
	Junyi15	5.0	**11.5**

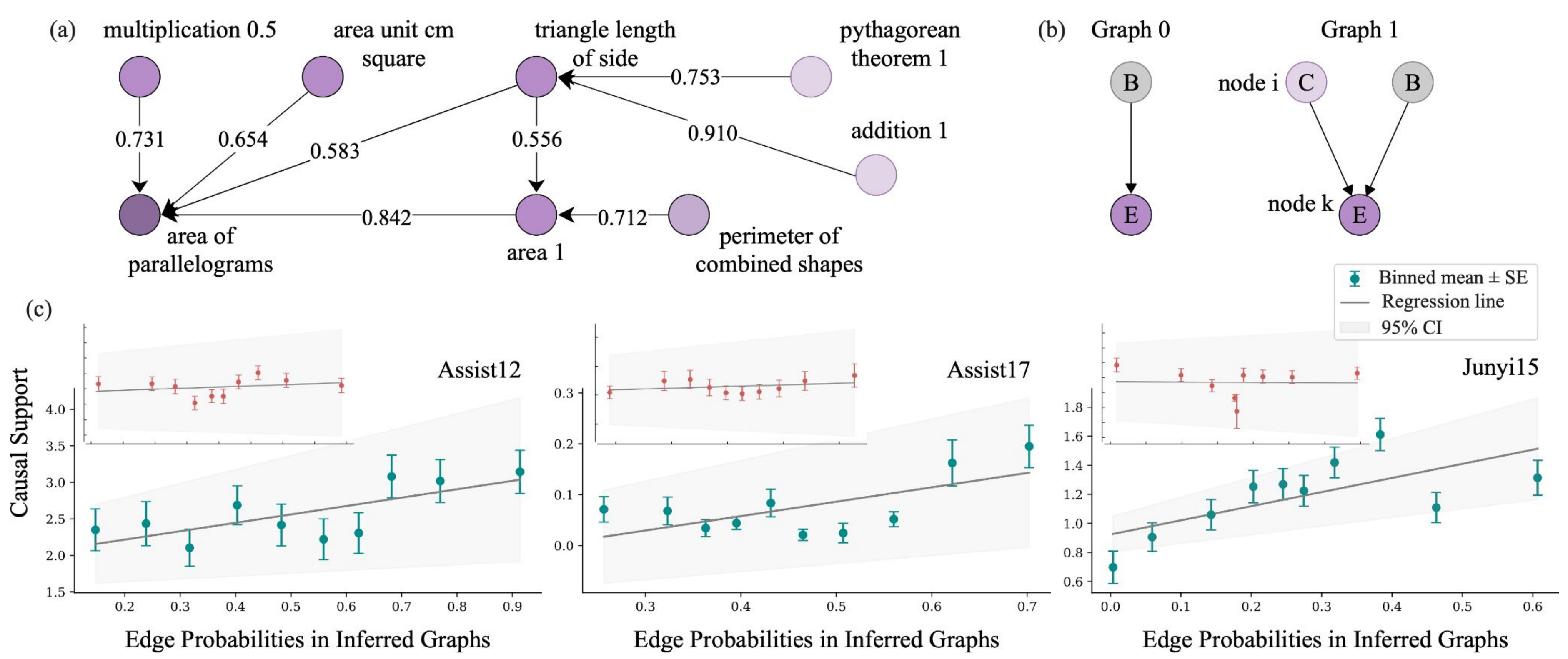
Operational interpretability:

Individual traits are able to predict meaningful behavior data in the future.

- Consolidated memory vs. initial performance
- **Forgetting rate vs. performance difference**

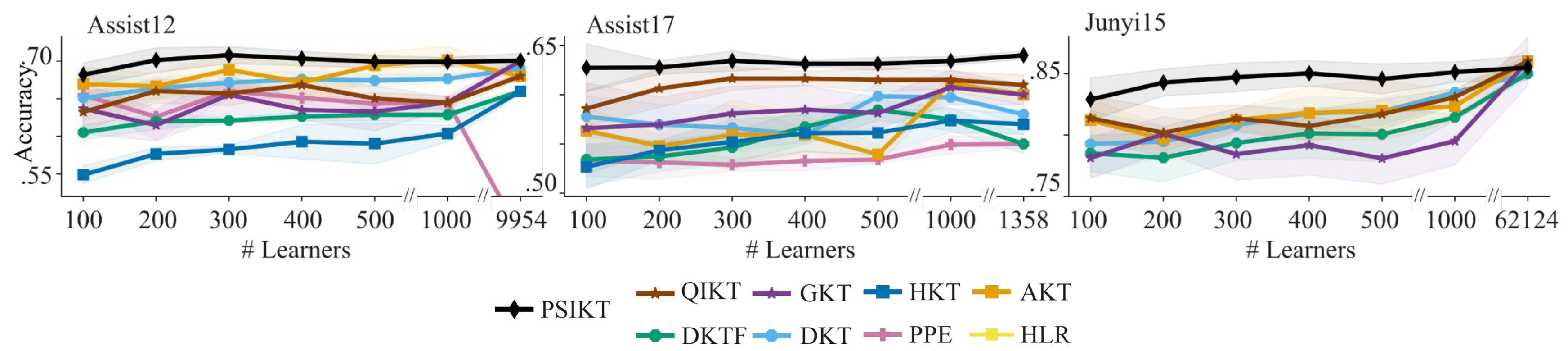


The prerequisite structures with *inferred edge probabilities* are shown to *causally support* learning[3].



Predictive accuracy on limited data

Within-learner prediction performance as a function of cohort sizes:

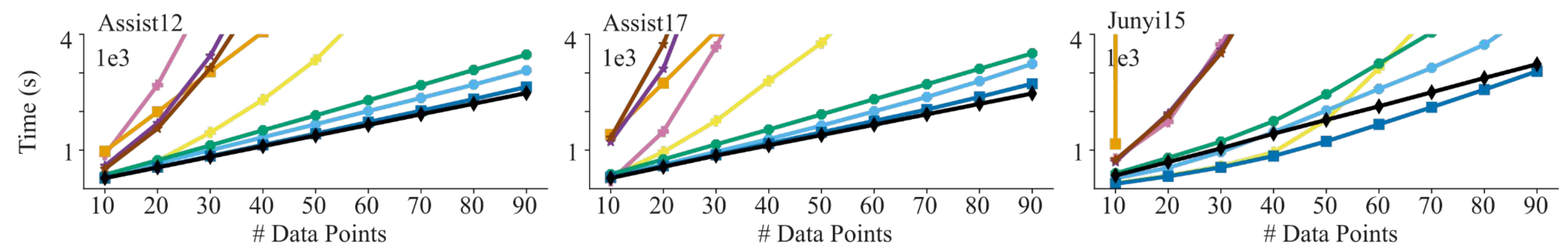


Scalability with more interaction

The Bayesian nature of PSI-KT allows it to be updated with new interaction without retraining[4].

$$\mathrm{ELBO}^{\mathrm{VCL}}(\theta,\phi_n) = \underbrace{\mathbb{E}_{q_\phi(z_{1:n},s_{1:n}\mid y_{1:n})}\left[\log p_\theta\left(y_n \mid z_n, s_n\right)\right]}_{\text{log-likelihood } p(D_t\mid\omega_t)} \underbrace{- \mathbb{E}_{q_\phi(z_{1:n},s_{1:n}\mid y_{1:n})}\left[\underbrace{\log q_{\phi_n}\left(z_n, s_n \mid y_{1:n}\right)}_{\text{posterior at time } t_n} - \underbrace{q_{\phi_{n-1},\theta}(z_n,s_n\mid y_{1:n-1})}_{\text{prior from time } t_{n-1}}\right]}_{\mathrm{KL}\left(q_{\phi_n}(\omega)\|q_{\phi_{n-1}}(\omega)\right)}$$

Cumulative training time of continual learning:



References:

[1] Abdelrahman, G., Wang, Q., & Nunes, B. (2023). Knowledge tracing: A survey. ACM Computing Surveys, 55(11), 1-37.
[2] Särkkä, S., & Solin, A. (2019). Applied stochastic differential equations (Vol. 10). Cambridge University Press.
[3] Griffiths, T. L., & Tenenbaum, J. B. (2009). Theory-based causal induction. Psychological review, 116(4), 661.
[4] Nguyen, C. V., Li, Y., Bui, T. D., & Turner, R. E. (2018, February). Variational Continual Learning. In International Conference on Learning Representations.

Contact: hanqi.zhou@uni-tuebingen.de