Hierarchical Deep State-Space Model for Enhanced Knowledge Tracing

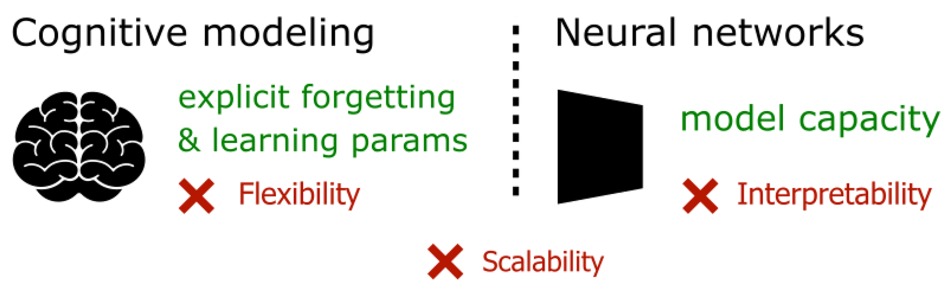
Hanqi Zhou, Robert Bamler, Charley M. Wu, Álvaro Tejero-Cantero
University of Tubingen & IMPRS-IS

imprs-is LEAD Graduate School EBERHARD KARLS UNIVERSITÄT TÜBINGEN

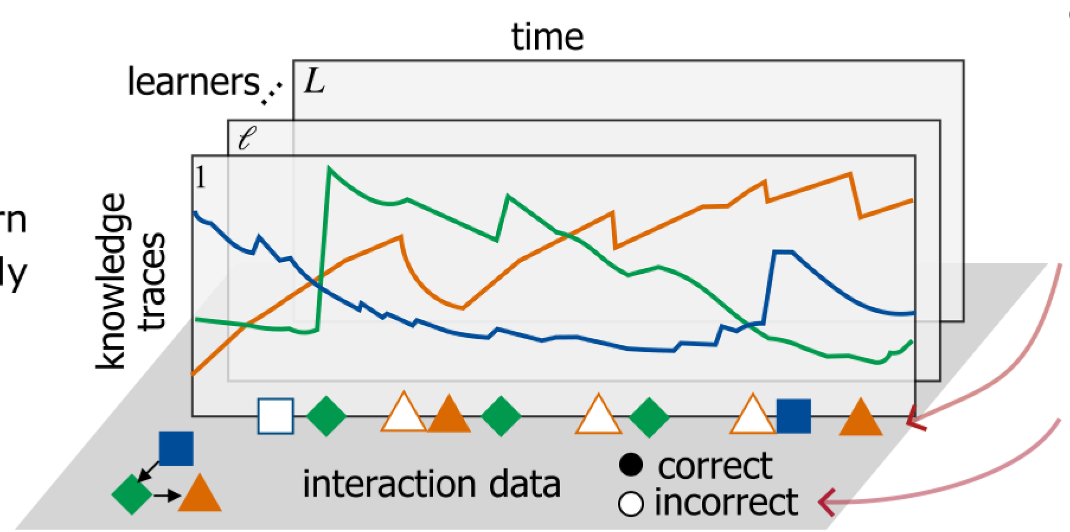
Why do we care?

What should we learn, and when to practice?
To create personalized, effective curricula, we need to find out:
1. What do learners already know, and how fast do learners forget?
2. What are suitable contexts to present content, i.e. what are prerequisites?

Existing models of human learning just can't keep up!
They fail to either grasp the structured nature of knowledge or learn interpretable variables, and meet the scalability demands of continuously updating models with new learners' data.



What is Knowledge Tracing (KT)?



Knolwedge Tracing (KT) [1] aims to estimate a learner's knowledge states given the learning interaction history.

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

x_n Knolwdege component (KC), e.g. an exercise on pythagorean theorem

t_n The timestamp of the interaction

y_n An evaluation of the learner's performance

Output: Prediction of the probability of learner's performance $p\left(y_{t_{n+1}} \mid x_{t_{n+1}}, \mathcal{H}_{t'<t_{n+1}}\right)$

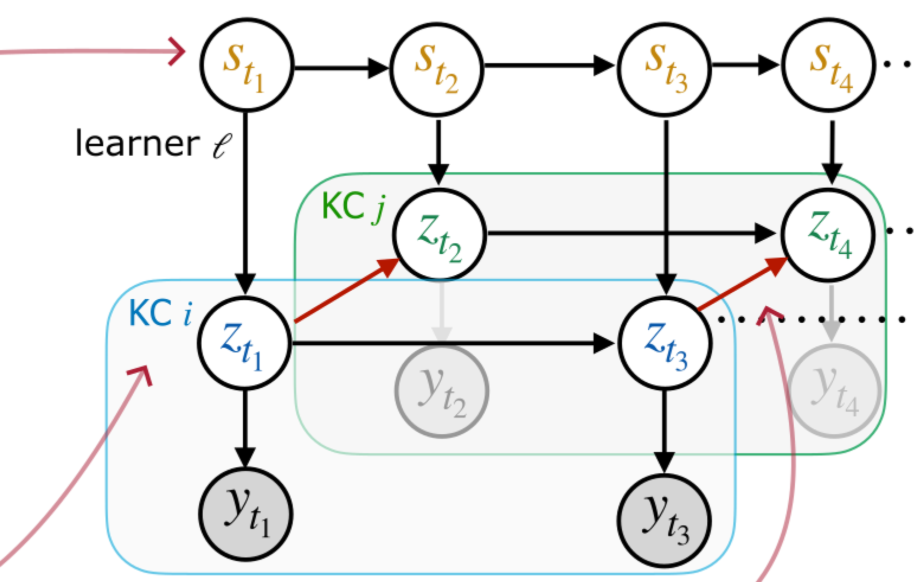
We propose GroupKT - a generative KT model.

Generative model

Cognitive traits - per learner $s^{\ell}_{t_n} := (\alpha^{\ell}_{t_n}, \mu^{\ell}_{t_n}, \gamma^{\ell}_{t_n})$
for personalization

$\alpha^{\ell}_{t_n}$ forgetting rate
$\mu^{\ell}_{t_n}$ long-term convergence level
$\gamma^{\ell}_{t_n}$ transfer ability

The Markovian evolution is modeled via a Kalman filtering prior.



Knowledge states - per learner and per KC $z^{k,\ell}_{t_n}$
for memory dynamics

the evolution is modeled via an Ornstein-Uhlenbeck process [4,5]

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

the transition distribution $p(z^{k,\ell}_{t_n} \mid s^{\ell}_{t_n}, z^{k,\ell}_{t_{n-1}}) = \mathcal{N}(z^{k,\ell}_{t_n} \mid m^{k,\ell}_{t_n}, w^{k,\ell}_{t_n})$ has mean

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

where $\tau^{\ell}_n = t^{\ell}_n - t^{\ell}_{n-1}$ is the time lag of two consecutive interactions, $\tilde{\mu}^{k,\ell}_{t_n} = \mu^{\ell}_{t_n}$ for single KC

Global Prerequisite graph A
for knowledge structure

for connected KCs, we shift the long-term convergence level in Eq.(1) by using the inferred structure

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

the existence and direction of edges are parameterized by KC embeddings U and transformation matrix M [2]

$$a_{ik} := \underbrace{\sigma((u^i)^{\intercal}u^k)}_{\substack{\text{the probability that} \\ \text{an edge exists at all}}}\ \underbrace{\sigma((u^i)^{\intercal}(M - M^{\intercal})u^k)}_{\substack{\text{the probability that the edge goes} \\ \text{from i to k given that it exists}}}$$

Inference of latent variables

In variational inference, we approximate an intractable posterior distribution $p_\theta(z \mid y) = p_\theta(y, z)/p_\theta(y)$ with $q_\phi(z \mid y)$ from a tractable distribution class.

ELBO of hierachical state-space model

The ELBO of our two-layer state space model is given

$$\begin{aligned}\mathrm{ELBO}(\theta, \phi) &= \mathbb{H}(q_\phi(z_{t_1:t_n}, s_{t_1:t_n} \mid y_{t_1:t_n})) \\ &\quad + \mathbb{E}_{q_\phi(z_{t_1:t_n}, s_{t_1:t_n} \mid y_{t_1:t_n})} \log p_\theta(y_{t_1:t_n}, z_{t_1:t_n}, s_{t_1:t_n})\end{aligned}$$

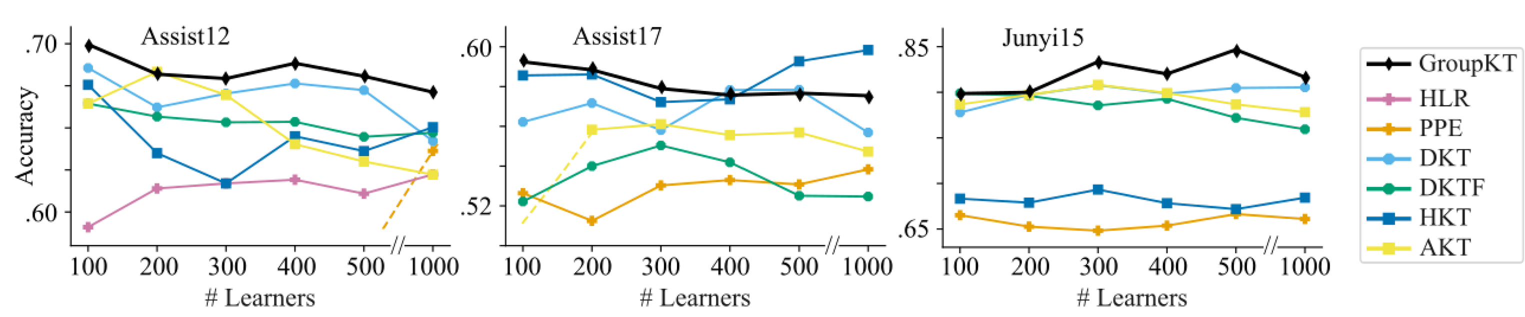
with joint log-likelihood of observations and latent variables

$$\begin{aligned}\log p_\theta\left(y_{t_1:t_n}, z_{t_1:t_n}, s_{t_1:t_n}\right) &= \log p_\theta(s_{t_1}) + \log p_\theta(z_{t_1}) \\ &\quad + \sum_{t_2}^{t_n}\left[\log p_\theta(s_{t_n} \mid s_{t_{n-1}}) + \log p_\theta(z_{t_n} \mid z_{t_{n-1}}, s_{t_n})\right] \\ &\quad + \sum_{t_1}^{t_n} p_\theta(y_{t_n} \mid z_{t_n})\end{aligned}$$

GroupKT enhances prediction, interpretability, and scalability

Prediction performance

Within-learner



Cross-learner

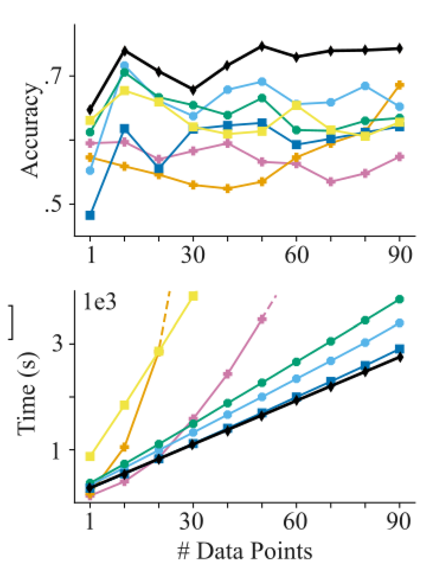
Dataset	Experiment	HLR	PPE	DKT	DKTF	HKT	AKT	Ours
Assist12	Within ↑	.591	.501	<u>.686</u>	.664	.676	.664	**.700**
	Between ↑	.503	.500	.552	.513	.552	<u>.588</u>	**.609**
	Between w/ FT ↑	.520	.500	.583	.549	.569	<u>.612</u>	**.620**
Assist17	Within	.471	.526	.562	.522	<u>.586</u>	.498	**.592**
	Between	.331	.512	.514	.482	<u>.519</u>	.472	**.525**
	Between w/ FT	.406	.513	.511	.534	<u>.551</u>	.507	**.563**
Junyi15	Within	.551	.665	.778	**.799**	.683	.787	**.799**
	Between	.481	.559	.760	<u>.762</u>	.619	.734	**.791**
	Between w/ FT	.522	.649	.817	**.843**	.646	.841	<u>.841</u>

Scalability regarding new interactions

Extend ELBO to online setup with new interactions [3]

$$\mathrm{ELBO}^{\mathrm{VCL}}(\theta, \phi_{t_n}) = \underbrace{\mathbb{E}_{q_\phi(z_{t_1:t_n}, s_{t_1:t_n} \mid y_{t_1:t_n})}\left[\log p_\theta(y_{t_n} \mid z_{t_n}, s_{t_n})\right]}_{\text{log-likelihood } p(D_t \mid \omega_t)}$$

$$- \underbrace{\mathbb{E}_{q_\phi(z_{t_1:t_n}, s_{t_1:t_n} \mid y_{t_1:t_n})}\left[\log(\underbrace{q_{\phi_{t_n}}(z_{t_n}, s_{t_n} \mid y_{t_1:t_n})}_{\text{posterior at time } t_n} - \underbrace{q_{\phi_{t_{n-1}}, \theta}(z_{t_n}, s_{t_n} \mid y_{t_1:t_{n-1}})}_{\text{prior from time } t_{n-1}}\right]}_{\mathrm{KL}(q_{\phi_t}(\omega) \| q_{\phi_{t-1}}(\omega))}$$



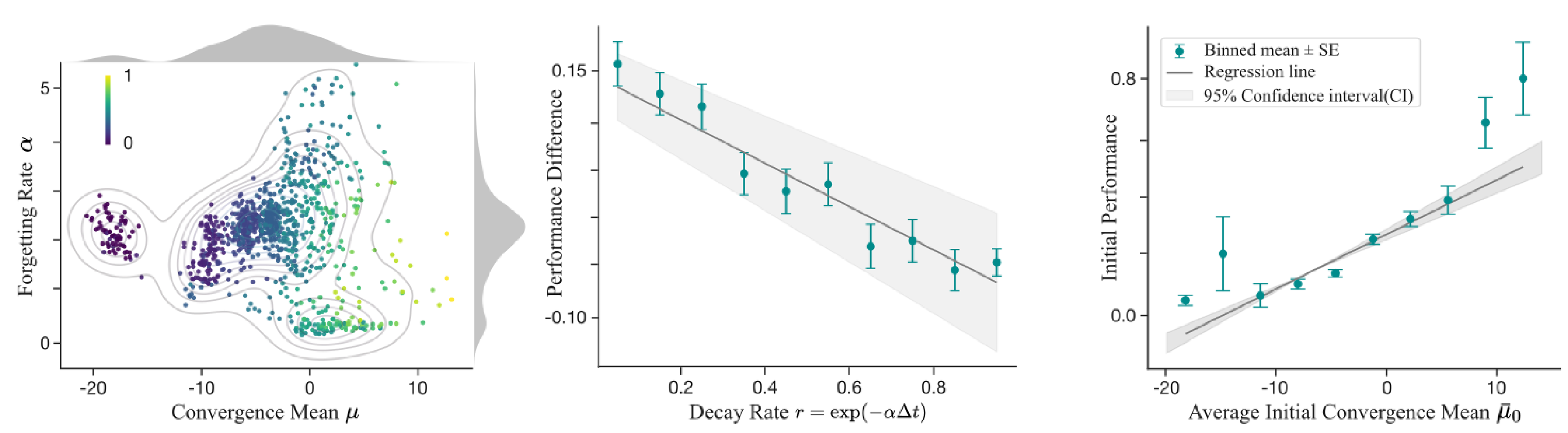
Interpretability of inferred variables and structure

Cognitive traits

Inferred cognitive traits identify different **clusters** of leaners.

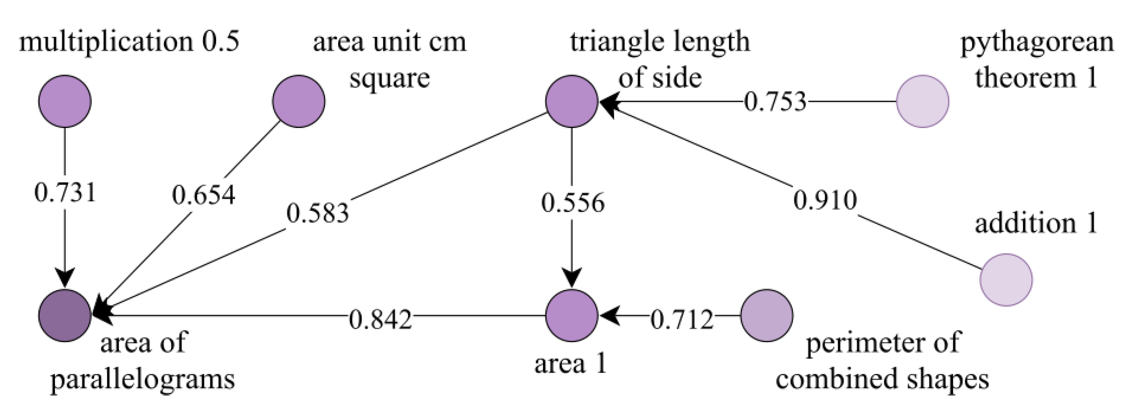
Individual **decay** rates indicate temporal performance difference.

Convergence variables indicate learners' overall familarity of the learning domain.

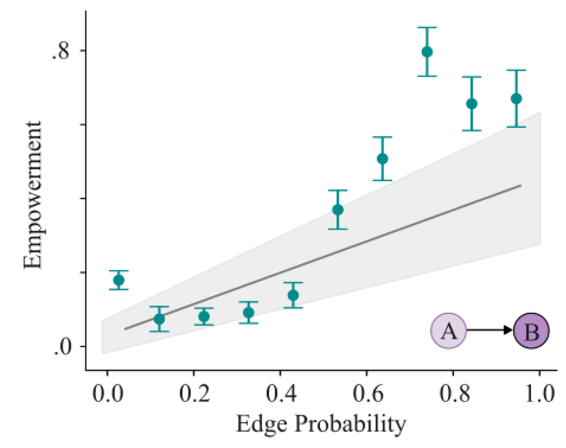


Inferred prerequisite graph

The prerequisite graph structure can be extracted with computed **edge probabilities**.



Edge probabilities indicate **empowerment** of learners' performance on target KC from its prerequisite.

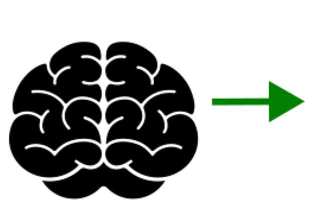


	Jaccard ↑	Pearson's r ↑, p-value ↓	nLL ↓	MRR ↑
HKT	.0034	<u>.0034</u>, .0056	<u>3.5389</u>	<u>.0087</u>
AKT	<u>.0027</u>	-.0114, .4213	5.3354	.0079
Ours	**.0079**	**.0890**, 3e-10	**2.2439**	**.0091**

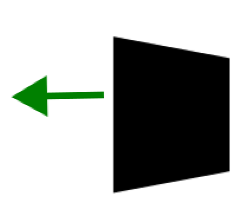
Why choose GroupKT?

Cognitive modeling

Neural networks



Flexibility in capturing human learning processes
Interpretability in connection with cognitive traits
Scalability and efficiency in dealing with new interaction data



References

[1] Abdelrahman, G., Wang, Q., & Nunes, B. (2023). Knowledge tracing: A survey. ACM Computing Surveys, 55(11), 1-37.
[2] Lippe, P., Cohen, T., & Gavves, E. (2021). Efficient neural causal discovery without acyclicity constraints. arXiv preprint arXiv:2107.10483.
[3] Nguyen, C. V., Li, Y., Bui, T. D., & Turner, R. E. (2017). Variational continual learning. arXiv preprint arXiv:1710.10628.
[4] Särkkä, S., & Solin, A. (2019). Applied stochastic differential equations (Vol. 10). Cambridge University Press.
[5] Zhou, H., Tejero-Cantero, Á., & Wu, C. M. (in press). The Dynamic and Structured Nature of Learning and Memory. In L. Hunt, C. Summerfield, T. Konkle, E. Fedorenko, & T. Naselaris (Eds.), Proceedings of the 2023 Conference on Cognitive Computational Neuroscience. Oxford, UK.