Universität Stuttgart
Institut für Parallele und Verteilte Systeme (IPVS)

Towards Bayesian Symbolic Computational Graph Completion

A Roadmap for Scientific Knowledge Infusion into Symbolic Machine Learning

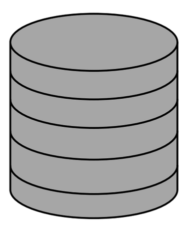
Tim Schneider M.Sc. Prof. Dr. Wolfgang Nowak Prof. Dr. Steffen Staab

Artificial Intelligence Software Academy (AISA) Kickoff 29.04.2022

Recent progress in Machine Learning is great, but . . .

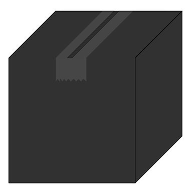
Current Machine Learning (ML) is often driven by large amounts of low quality internet data, yielding large, computationally expensive, *black box models* that provide little insight into the problem. When directly applied to *Scientific Domains*, such methods . . .

Big Data



. . . require a large amount of data to converge to a good solution. In many scientific domains only little data is available yielding to overfitting and unsatisfactory results.

Black Box Solutions



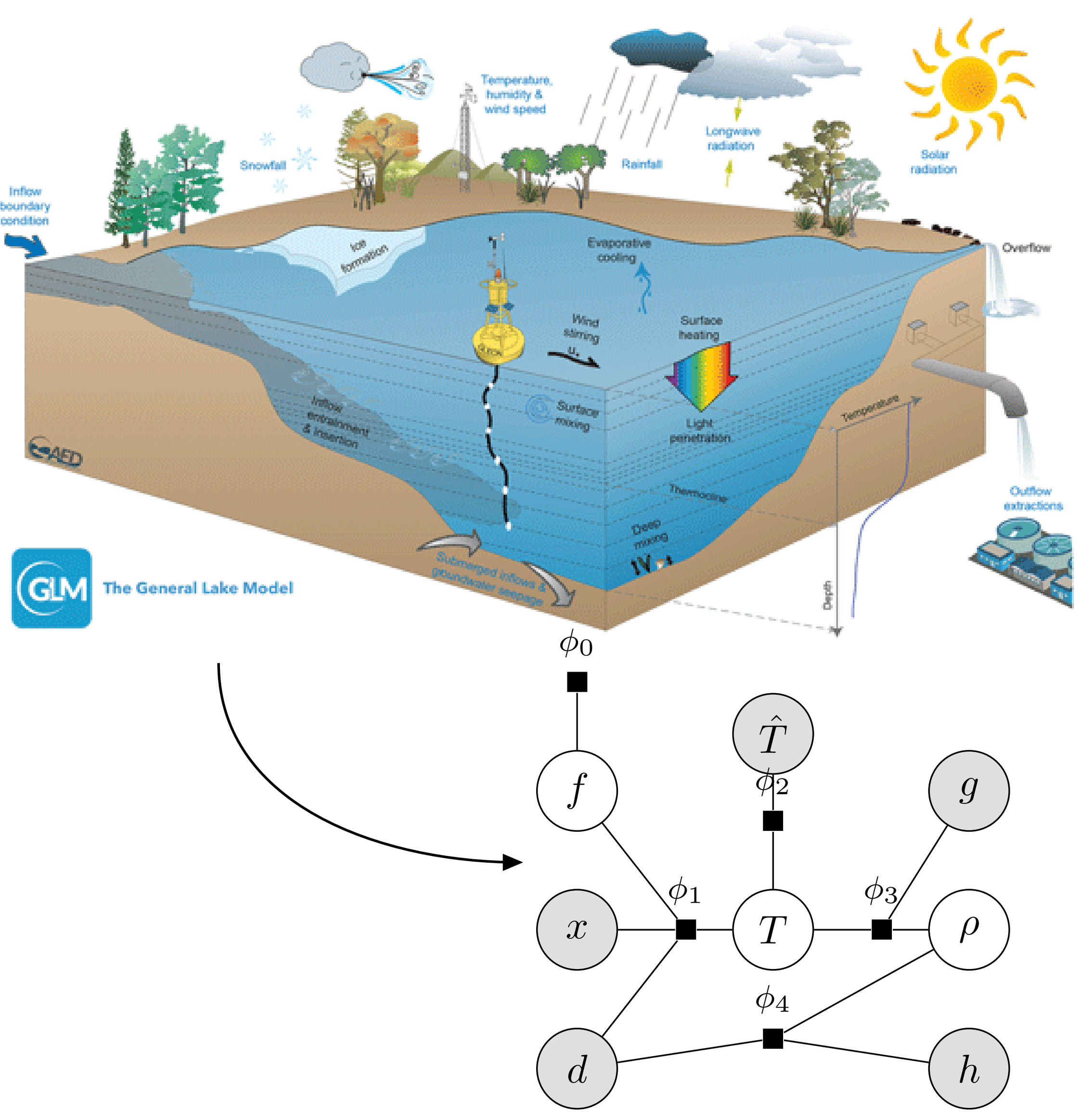
. . . provide black box solutions and therefore do not create new scientific hypothesis. But without the latter, ML only adds little value to the scientific domain research.

Ignores Scientific Knowledge

. . . ignore existing prior knowledge in the domain and learn everything from scratch often resulting in solutions that violate common knowledge.

To address these problems . . .

① Encode Prior Knowledge in a (Bayesian) Graph



- *computational graphs* [6] represent prior knowledge and relations between variables in computational tasks
- a *factor graph* [4] notation allows probabilities (=factors ϕ_i) for variables as well as knowledge like
  - equations and auxiliary variables [1]
  - differential equations [3]
  - structural priors (with grammars) [9]

- *Example:* 1*D* Lake Temperature Model

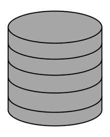
  Task: find the unknown function $f(x, d) = T$

  AND

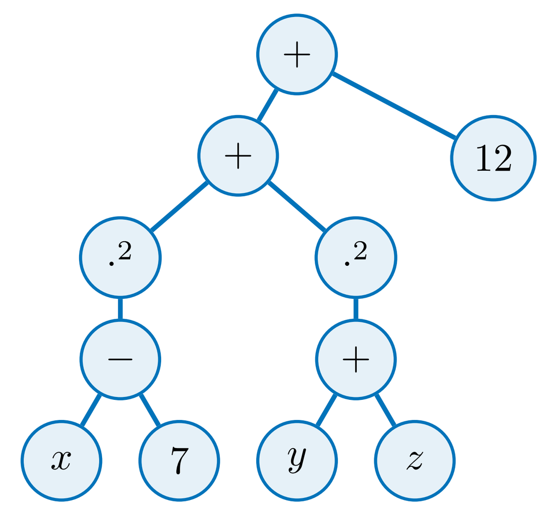
  "I know that temperature T is related to density ρ via g and the density monotony h holds with increasing depth d." (compare [1])

② Input Additional Data Evidence

- measurements or observations for certain variables in the graph



③ Output Symbolic Solution Candidates



- for the unknown functions in the graph
- composes complex expressions out of a *library* of known operations
- yields a *scientific hypothesis* that can be interpreted and debated
- can be found with Bayesian Symbolic Regression (BSR) [2] by performing MCMC steps in the space of symbolic representations.

Proposed Framework: Symbolic Computational Graph Completion

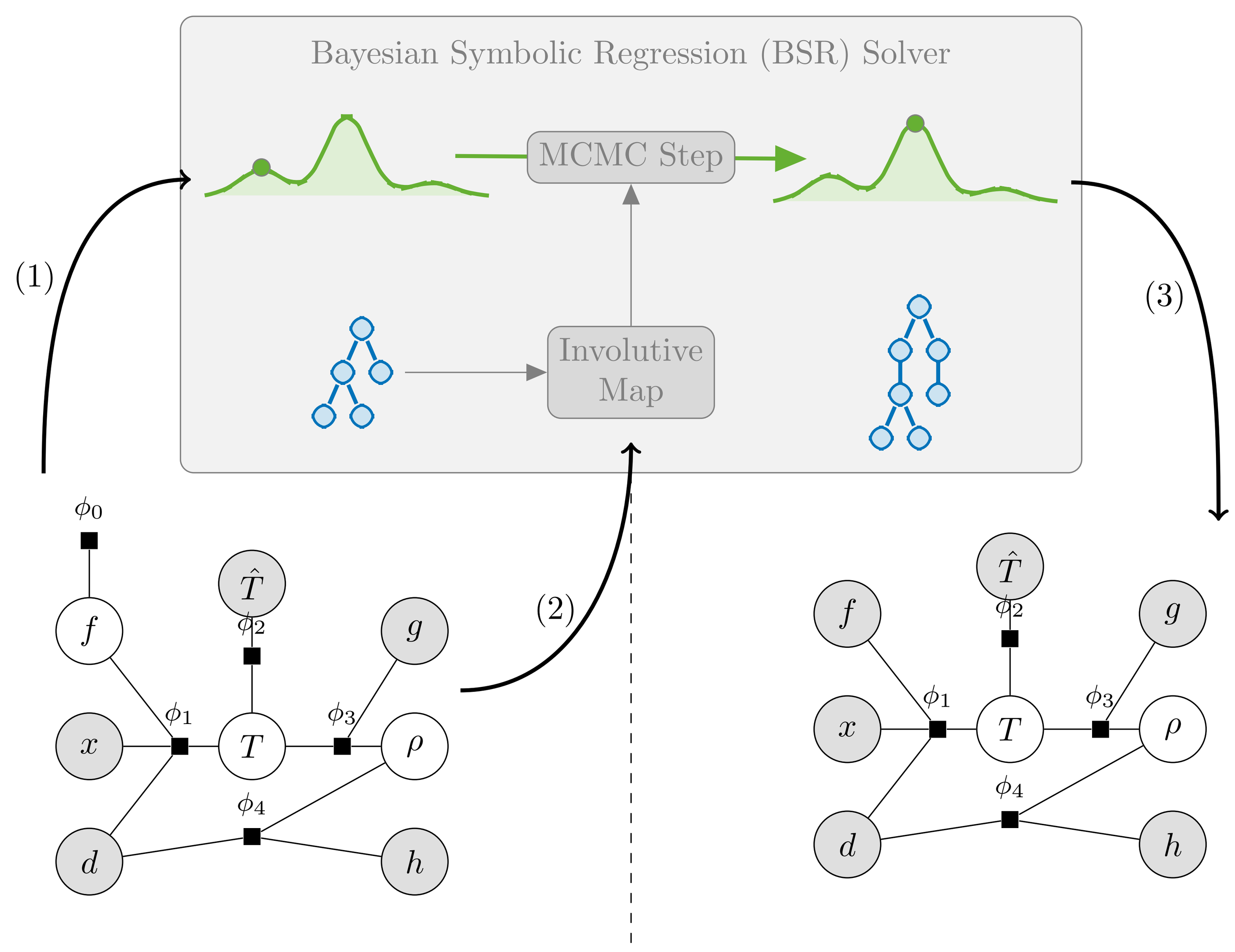


Figure: Overview of the proposed framework for Symbolic Computational Graph Completions

Our Approach

- Encode the problem as well as all prior knowledge in a *computational (factor) graph* representation.
- Fill in symbolic solutions with data and *Bayesian Reasoning*.

Scientific Knowledge Exploitation

Hereby, the framework considers *three steps* that exploit the prior knowledge in the graph:

(1) *Unnormalized Posterior:*
use the graph to evaluate the unnormalized posterior distribution of potential symbolic expressions in order to sample from it using MCMC (see [2])

(2) *Informed Sampler:*
involutive maps [5] encode additional information about the problem to design informed samplers that produce samples of high likelihood. Especially learned maps [8] allow a connection to recent advances in deep learning (e.g. [7]).

(3) *Graph Update:*
The Bayesian solver yields symbolic hypothesis that enables a *human researcher* to
- accept one of the hypothesis / candidates or
- come up with new prior knowledge in the graph

References

[1] A. Daw, A. Karpatne, W. Watkins, J. Read, and V. Kumar. Physics-guided neural networks (pgnn): An application in lake temperature modeling. *arXiv preprint arXiv:1710.11431*, 2017.
[2] Y. Jin, W. Fu, J. Kang, J. Guo, and J. Guo. Bayesian symbolic regression. *arXiv preprint arXiv:1910.08892*, 2019.
[3] S. Karimpouli and P. Tahmasebi. Physics informed machine learning: Seismic wave equation. *Geoscience Frontiers*, 11(6):1993–2001, 2020.
[4] D. Koller and N. Friedman. *Probabilistic graphical models: principles and techniques.* MIT press, 2009.
[5] K. Neklyudov, M. Welling, E. Egorov, and D. Vetrov. Involutive mcmc: a unifying framework. In *International Conference on Machine Learning*, pages 7273–7282. PMLR, 2020.
[6] H. Owhadi. Computational graph completion. *arXiv preprint arXiv:2110.10323*, 2021.
[7] B. K. Petersen, M. L. Larma, T. N. Mundhenk, C. P. Santiago, S. K. Kim, and J. T. Kim. Deep symbolic regression: Recovering mathematical expressions from data via risk-seeking policy gradients. *arXiv preprint arXiv:1912.04871*, 2019.
[8] J. Song, S. Zhao, and S. Ermon. A-nice-mc: Adversarial training for mcmc. *Advances in Neural Information Processing Systems*, 30, 2017.
[9] K. Xu, A. Srivastava, D. Gutfreund, F. Sosa, T. Ullman, J. Tenenbaum, and C. Sutton. A bayesian-symbolic approach to reasoning and learning in intuitive physics. *Advances in Neural Information Processing Systems*, 34, 2021.

AISA
Artificial Intelligence Software Academy