Steinbuch Centre for Computing (SCC)
Data Analytics, Access and Applications (D3A)

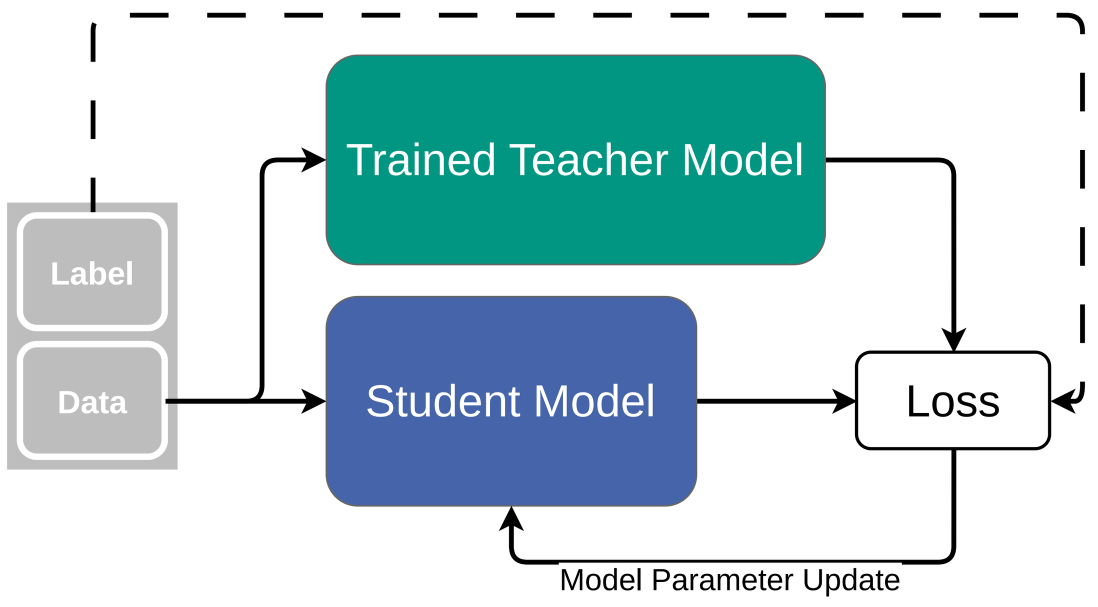
Utilizing Reversibility of Quantum Gates for Quantum Knowledge Distillation

Christof Wendenius, Eileen Kuehn, Mona Heinkelein, Achim Streit

What is Knowledge Distillation?

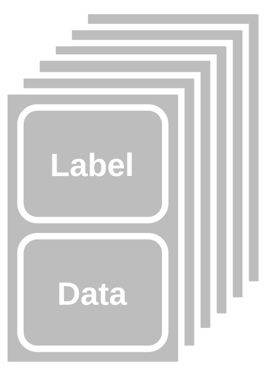
- Used to improve the runtime, size or structure of an existing model, e.g. by reducing the number of trainable parameters
- Requirement is a trained **Teacher Model**
- The parameters of the Teacher Model are frozen and used to train a smaller model
- The **Student Model** is trained to mimic the output of the Teacher Model
- We can use Knowledge Distillation to reduce the depth of Parametrized Quantum Circuits (PQCs)

Training Setup

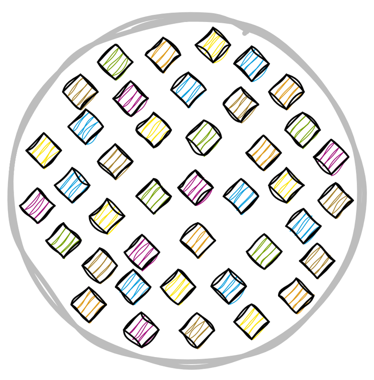


- The teacher's output indirectly contains more information than a simple label, e.g.:
  - The relationship of different classes
  - The presence and absence of certain features
  - The teacher's confidence in its output
- This *dark knowledge* is in the unprocessed output of every well-trained model
- More information per step allows to reach high(er) accuracies than training the same model size commonly from scratch

The Dataset



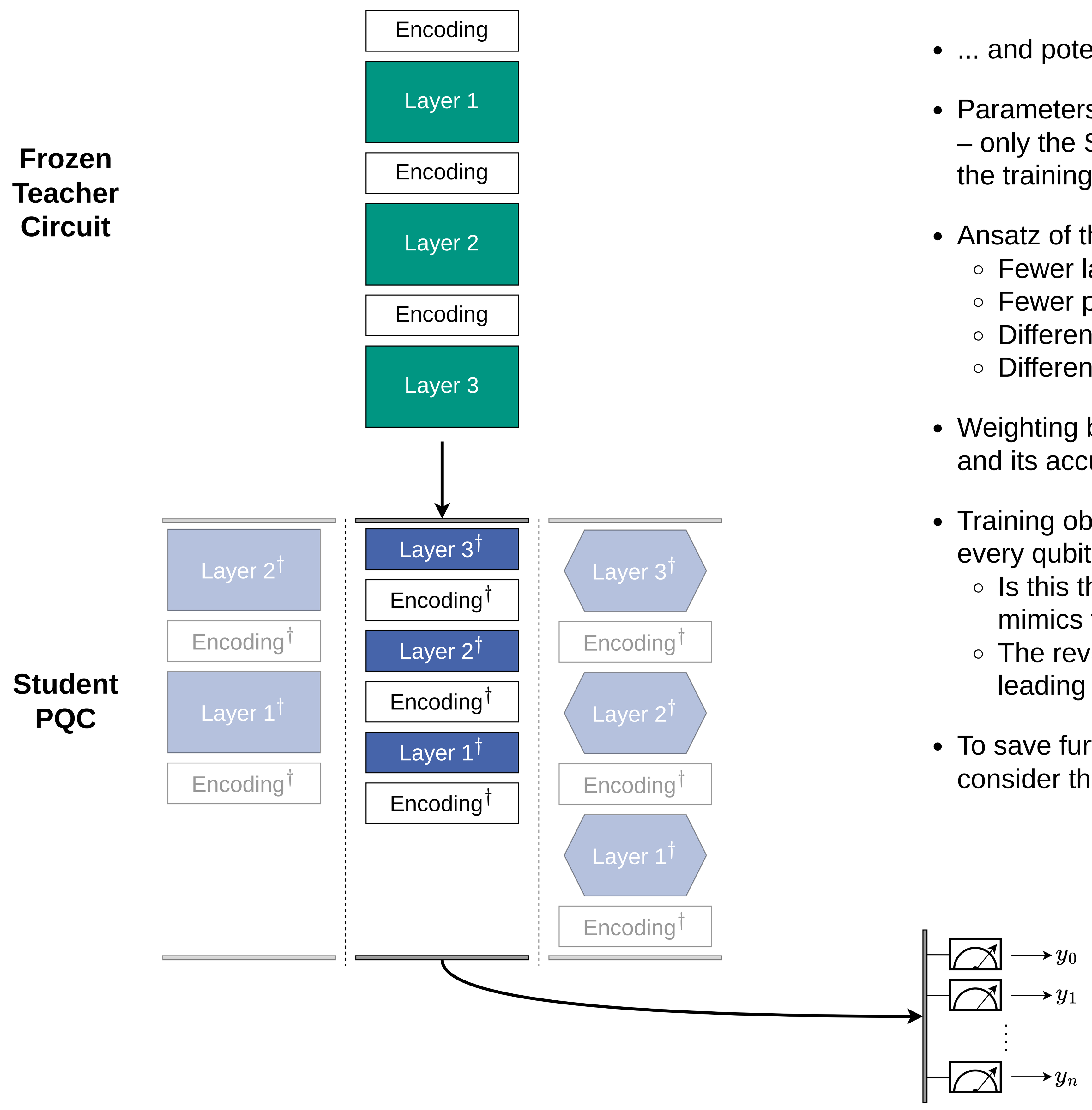
Original Dataset



Random Inputs

- The Student Model can be trained on the original dataset the Teacher Model was trained on
- The Student Model can also be trained on random inputs
  - Mimics the teacher for every input
  - More information and more training input in case of a small dataset

The Training Setup and how to choose your Student

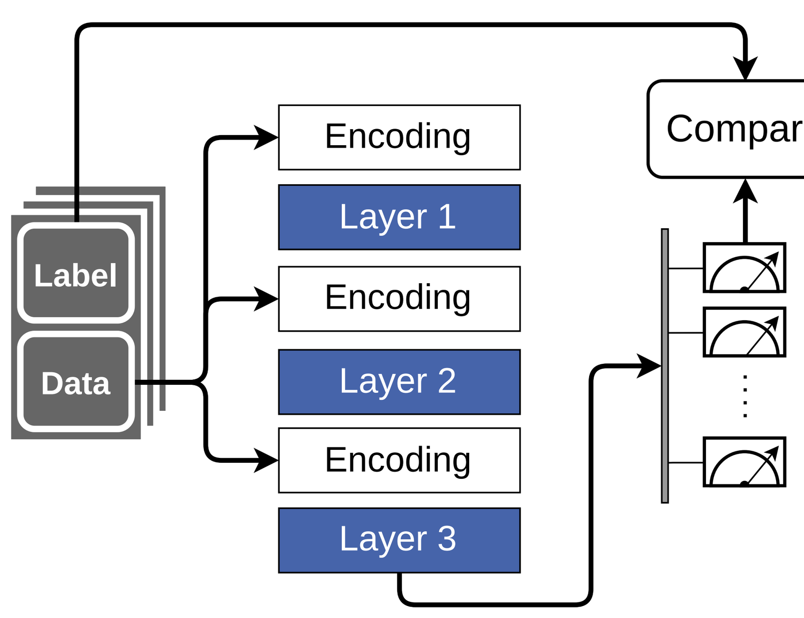


- Setup allows to compare the two outputs without measuring both outputs ...
- ... and potentially lose information
- Parameters of the Teacher Model are frozen – only the Student Model is updated during the training
- Ansatz of the student can be chosen freely
  - Fewer layers
  - Fewer parameters in the layers
  - Different gate set
  - Different ansatz
- Weighting between the desired model/ansatz and its accuracy is possible
- Training objective is to measure zero on every qubit
  - Is this the case the student completely mimics the teacher
  - The reversed student undoes the teacher leading to the initialization state
- To save further circuit evaluations, we do not consider the label in the loss function

Loss Calculation

$$\max(\langle 0^{\otimes n}|UMU^{\dagger}|0^{\otimes n}\rangle)$$

Evaluation



- Evaluation is done on the main dataset
- The Student Model is no longer reversed
- In some (Quantum) Knowledge Distillation setups, the Student Model can even surpass the Teacher Model

More on Knowledge Distillation:
[1] Gou, Jianping, et al. "Knowledge distillation: A survey." *International Journal of Computer Vision* 129 (2021): 1789-1819.
[2] Hinton, Geoffrey, Oriol Vinyals, and Jeff Dean. "Distilling the knowledge in a neural network." *arXiv preprint arXiv:1503.02531* (2015).
[3] Romero, Adriana, et al. "Fitnets: Hints for thin deep nets." *arXiv preprint arXiv:1412.6550* (2014).

More on Quantum Knowledge Distillation:
[4] Alam, Mahabubul, Satwik Kundu, and Swaroop Ghosh. "Knowledge Distillation in Quantum Neural Network using Approximate Synthesis." *Proceedings of the 28th Asia and South Pacific Design Automation Conference*. 2023.

wendenius@kit.edu