NEURAL INFORMATION PROCESSING SYSTEMS

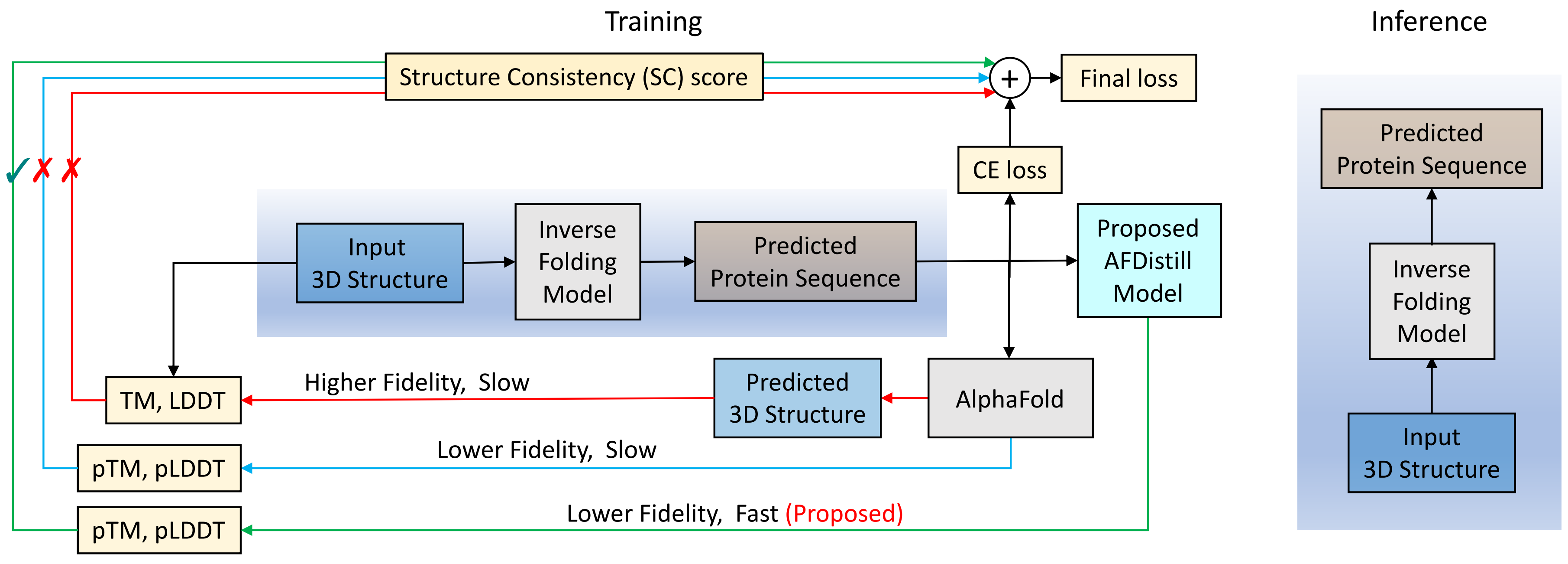
AlphaFold Distillation for Protein Design

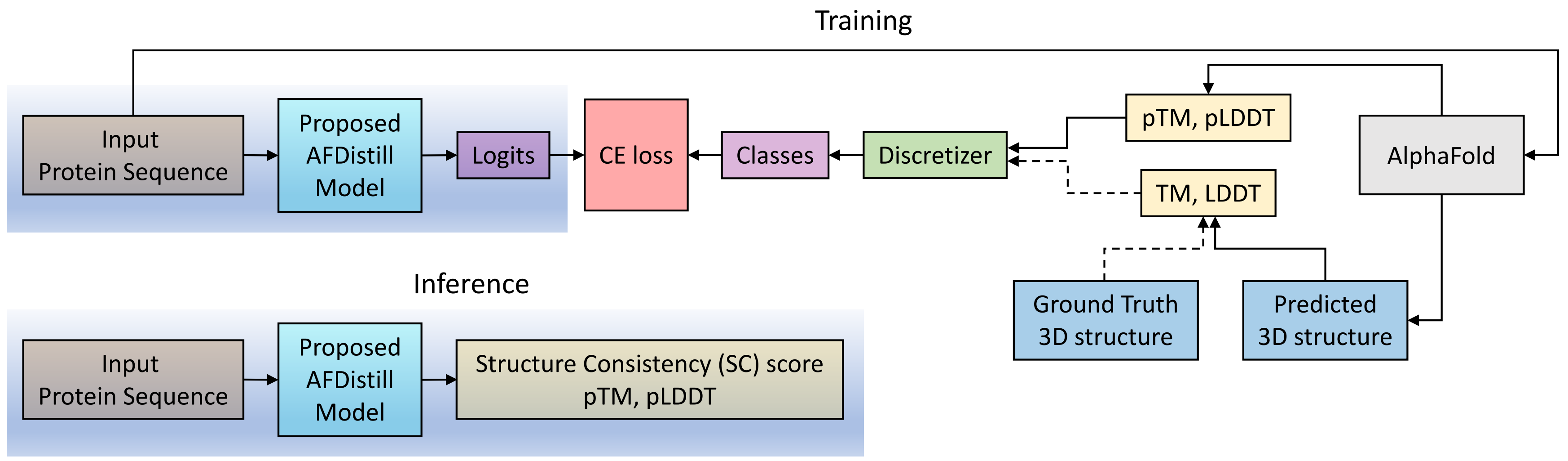
IBM Research

Igor Melnyk, Aurelie Lozano, Payel Das, Vijil Chenthamarakshan
IBM Research, Yorktown Heights, NY

Introduction

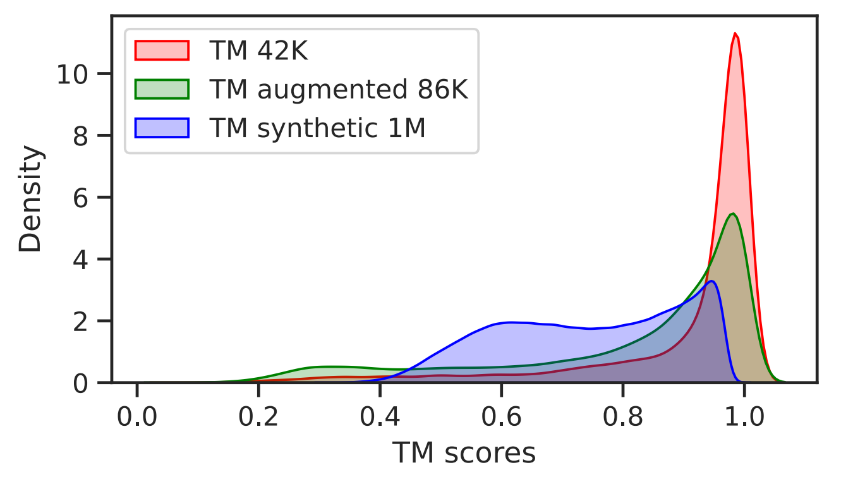
- Inverse Protein Folding
  - Design protein sequence that folds into a given 3D structure
  - Fundamental challenge in bioengineering and drug discovery
  - 8 of the top 10 best-selling drugs are engineered proteins
- Current Approaches
  - Traditionally, optimize sequences to achieve specific structures and functions
  - Recent deep generative models learn to translate structure into sequence
  - However, often lack in producing diverse, functional sequences
- AlphaFold
  - Forward folding model
  - Accurately estimates structure from sequence, provides confidence metrics (pLDDT, pTM)
  - However, very slow
- Our Work
  - Merge inverse with forward folding to provide feedback on generated sequence
  - Proposed method: **AlphaFold Distll (AFDistill)**
    - Fast, end-to-end differentiable
    - Trained on AlphaFold-generated data [sequence --> TM/LDDT score]
    - Use as part of optimization loop in the Inverse Folding Design
    - More generally, can be used in any protein optimization algorithm

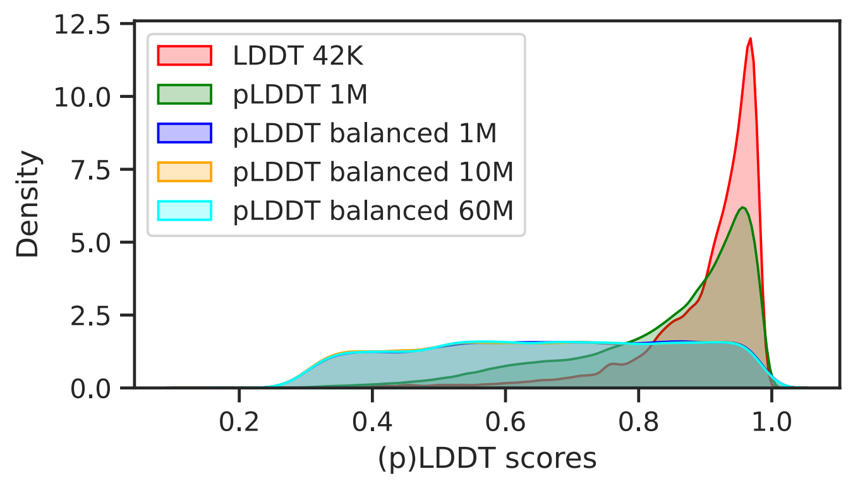


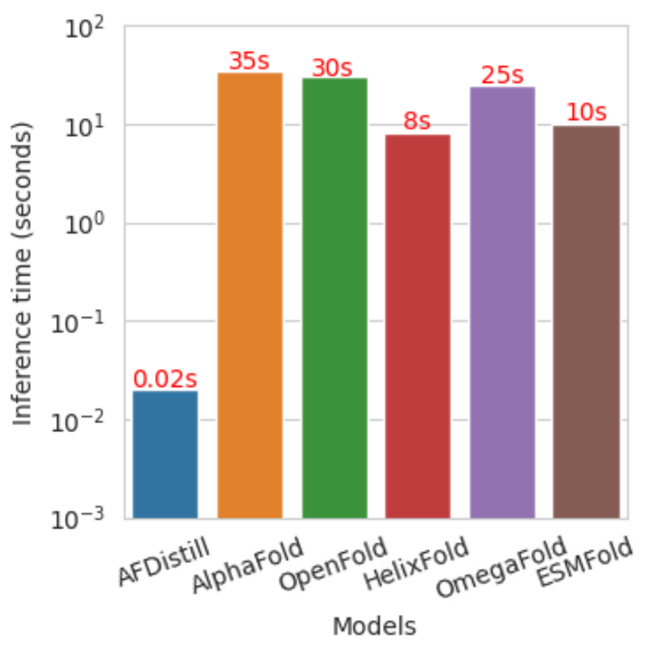


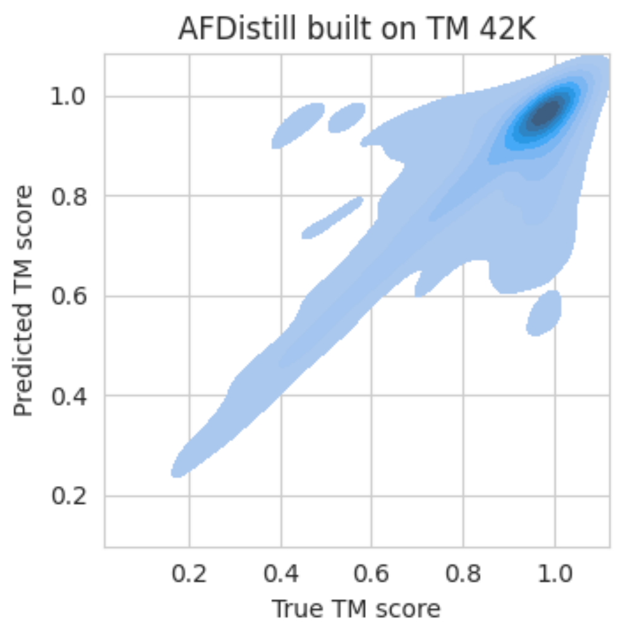
AlphaFold Distilllation

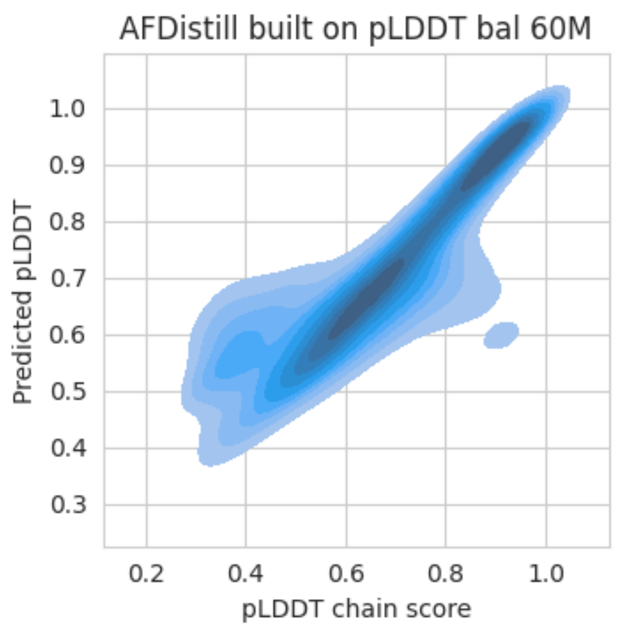
- Data
  - Sourced from AlphaFold Database Release 3 (900K+) and 4 (214M+)
  - Created multiple balanced datasets for more representative training
- Model
  - Adapted ProtBert, a BERT-based Transformer with 420M parameters
  - Adjusted ProtBert head to classify protein residue states in 50 discrete bins
  - Estimates pTM/pLDDT scores per protein sequence
- Results
  - Eval shows high accuracy with true vs. predicted scores clustering on the diagonal
  - Kernel density plots demonstrate model reliability in predicting protein structures
  - Orders of magnitude faster than existing methods





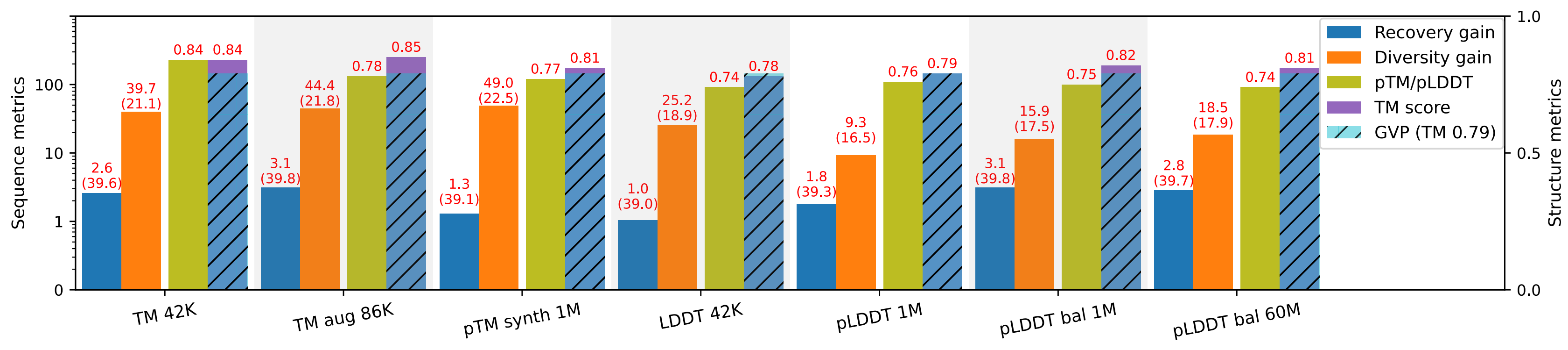






Inverse Protein Folding

- Overview
  - Use AFDistill as a Structure Consistency (SC) score in inverse protein folding
  - Evaluate protein sequence recovery, diversity, perplexity, and TM-score
  - CATH 4.2 dataset
- GVP
  - GVP+SC improves diversity without compromising TM scores
  - SC regularization induces diversity by allowing multiple high-score sequence candidates
  - Candidate protein sequences with high pTM/pLDDT drive both recovery and diversity
- ProteinMPNN
  - ProteinMPNN benefits from SC regularization, sustaining high recovery rates.
  - SC regularization enhances sequence diversity better than backbone noise alone
- PiFold
  - PiFold's performance improved by SC regularization without sacrificing recovery rates
  - SC regularization introduces significant diversity in generated protein sequences



	Recovery		Diversity		Perplexity	
	ProteinMPNN	ProteinMPNN +SC	ProteinMPNN	ProteinMPNN +SC	ProteinMPNN	ProteinMPNN +SC
Backbone Noise 0.02	47.7	47.5 (-0.4%)	22.5	24.3 (+8.0%)	5.1	5.1 (+0.0%)
Backbone Noise 0.1	43.8	44.0 (+0.5%)	28.1	30.4 (+8.2%)	5.3	5.4 (+1.9%)
Backbone Noise 0.2	39.5	39.9 (+1.0%)	31.3	34.4 (+9.9%)	5.8	5.8 (+0.0%)
Backbone Noise 0.3	36.3	36.4 (+0.0%)	33.0	37.8 (+14.6%)	6.2	6.3 (+1.6%)

	Original		TM 42K		TM aug 86K		TM synth 1M		LDDT 42K		pLDDT 1M		pLDDT bal 60M	
	Rec	Perp	Rec	Perp	Rec	Perp	Rec	Perp	Rec	Perp	Rec	Perp	Rec	Perp
Greedy	51.1	4.8	50.9 (-0.4%)	5.0 (+4.0%)	51.0 (-0.2%)	4.8 (+0.0%)	50.5 (-1.2%)	5.2 (+8.3%)	50.8 (-0.6%)	4.9 (+2.1%)	50.9 (-0.4%)	4.8 (+0.0%)	51.1 (+0.0%)	4.7 (-2.1%)
	Rec	Div	Rec	Div	Rec	Div	Rec	Div	Rec	Div	Rec	Div	Rec	Div
Sampled	42.6	52.4	42.5 (-0.2%)	60.7 (+15.8%)	42.8 (+0.5%)	60.2 (+14.9%)	42.4 (-0.5%)	61.1 (+16.6%)	42.3 (-0.7%)	60.9 (+16.2%)	42.5 (-0.2%)	60.5 (+15.5%)	42.9 (+0.7%)	60.0 (+14.5%)

Code: github.com/IBM/AFDistill

Paper: arxiv.org/abs/2210.03488