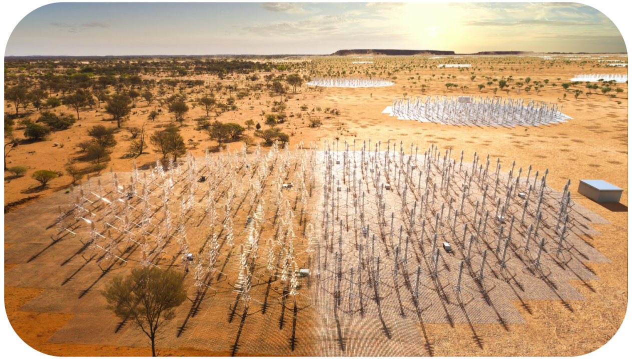
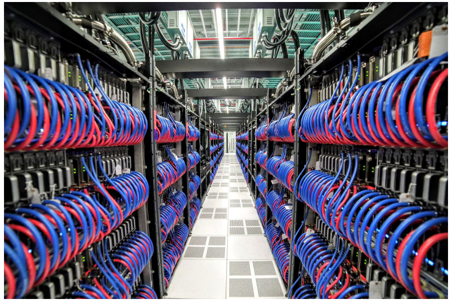
AI Based compression for radio-astronomy data

Adrien Devresse, Xiaoran Chen,
Prof. Emma Tolley, Dr Mathieu Salzmann,

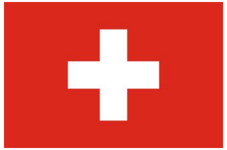
Context : The radio-Astronomy data deluge



10-100PiB



Regional sites



Archiving

ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE

Challenges

The Radio-astronomy domain:

- Very large collections of Data (10-100 PiB)
- Hard to compress: Interesting data are *IN* the noise
- Risk of scientifically relevant data losses
- Traditionally stored in legacy file format without compression (MSv2) or with basic lossless compression (Gzip)
- Random I/O pattern

ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE

Challenges (2)

- Traditional lossless algorithms (GZip, ...) are very inefficient on radio-data
- Traditional lossy image compression algorithms (JPEG, WebP) are too destructive
- Traditional A.I based image compression models are not tailored for Complex FP64 data

Project goal

- Identify compression algorithms for Radio-astronomy
  - Classical algorithms
  - A.I based algorithms
  - For Lossless and lossy
- Make it production ready & integrated with I/O toolkits
- Optimize it for HPC machines and GPUs

- Funded by

ODISSEE
horizon europe
SDSC

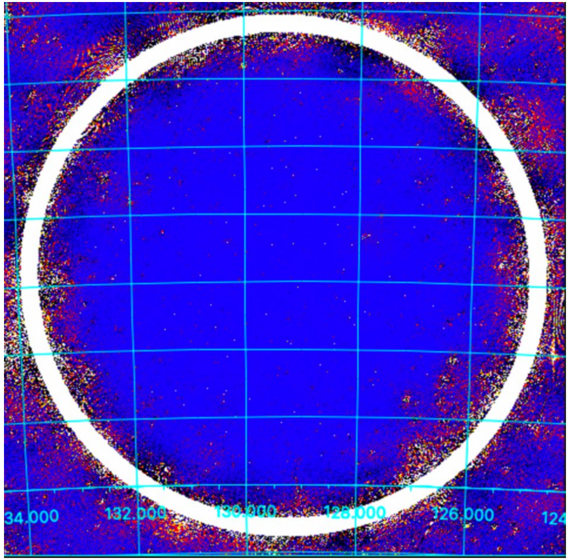
Algorithms evaluated

- **Zstandard** algorithm: classical, lossless
  - https://facebook.github.io/zstd/
- **Dysco**: classical, quantization, lossy
  - https://arxiv.org/abs/1609.02019
- **MGard**: classical, multi-grid quantization, lossy
  - https://arxiv.org/html/2401.05994v1
- **VAE, Image models**: A.I based, lossy
  - Stay tuned !
- **Bit-swap:** A.I based, HLatent Variables, lossless
  - https://arxiv.org/abs/1905.06845

Datasets

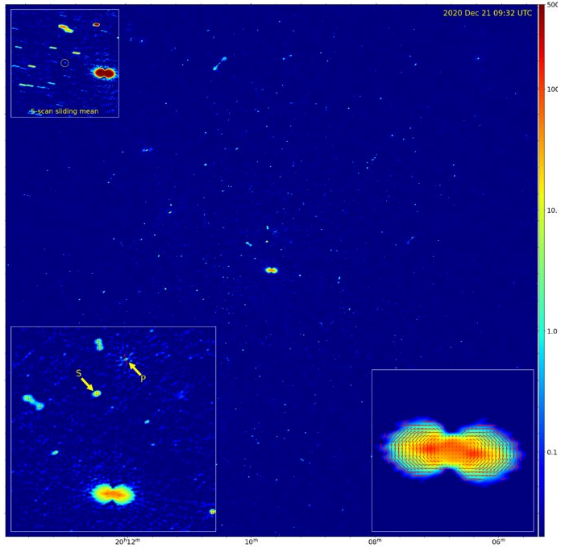
SKA Low Validation dataset

- Size: 160GiB
- SKA Low, Simulated
- Open Test dataset



RATT PARROT dataset

- Size: 411GiB
- MeerKat, Measured
- https://arxiv.org/abs/2312.12165



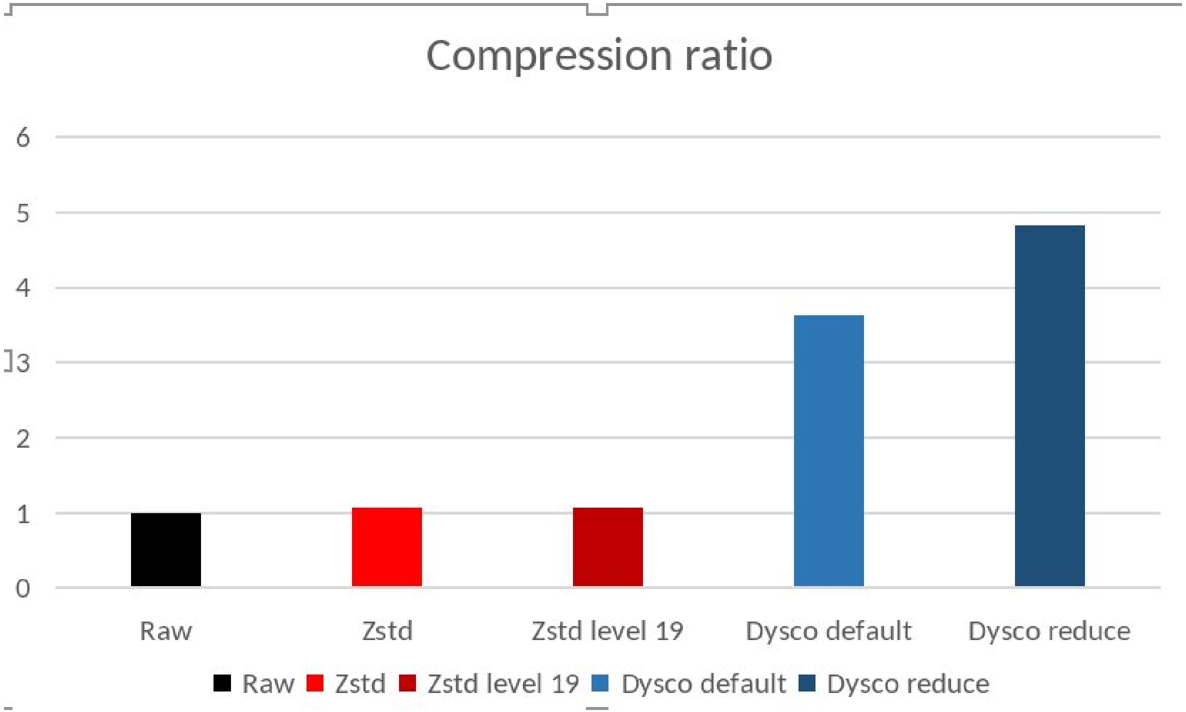
ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE

Test environment

- **CSCS Alps**
  - Grace Hopper
  - Aarch64
  - https://www.cscs.ch/computers/alps

- **Software deployed with Astro Nix**
  - Radio-astronomy software collection
  - Based on Nix
  - https://github.com/adevress/astro-nix

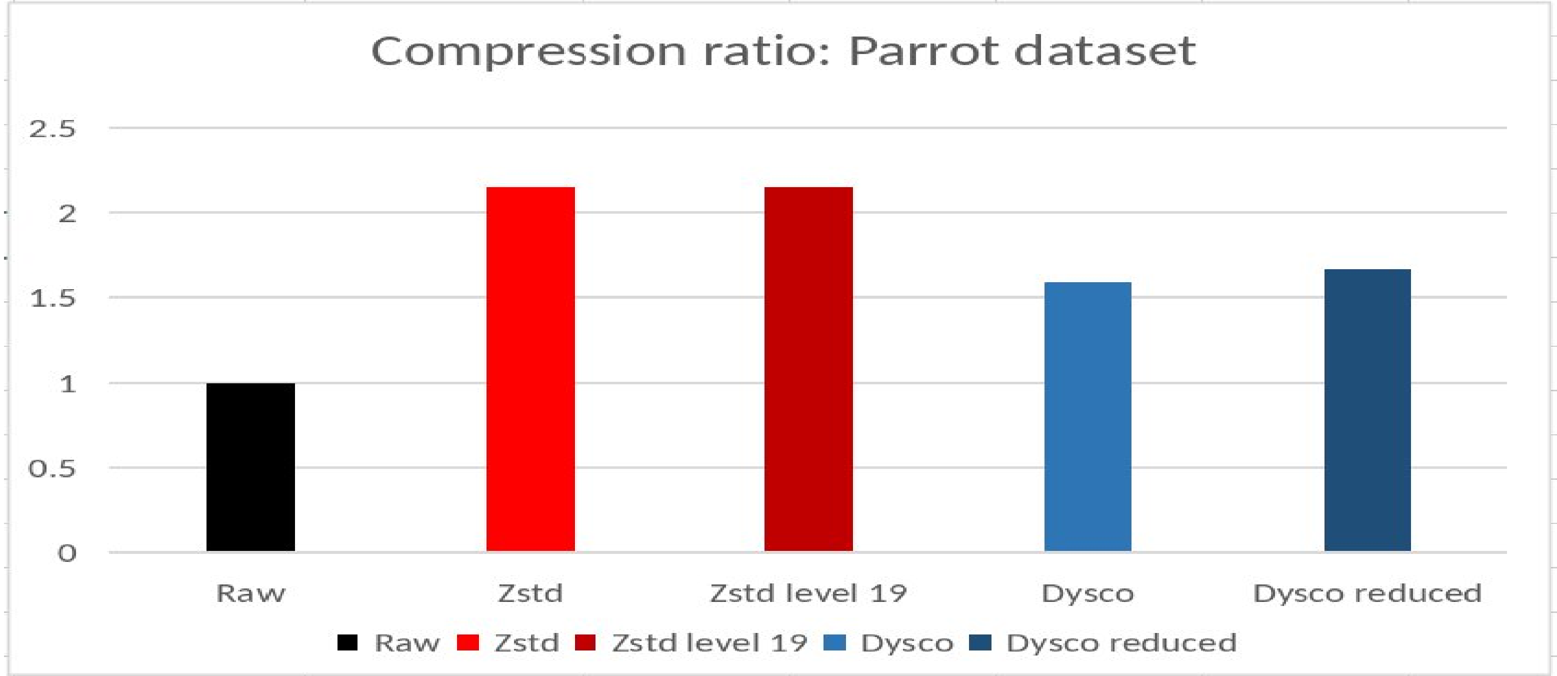
Classical Algorithms: Compression ratio



[1]: SKA PI26 SKA Low simulated dataset for SDP pipeline validation, ~160GiB

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

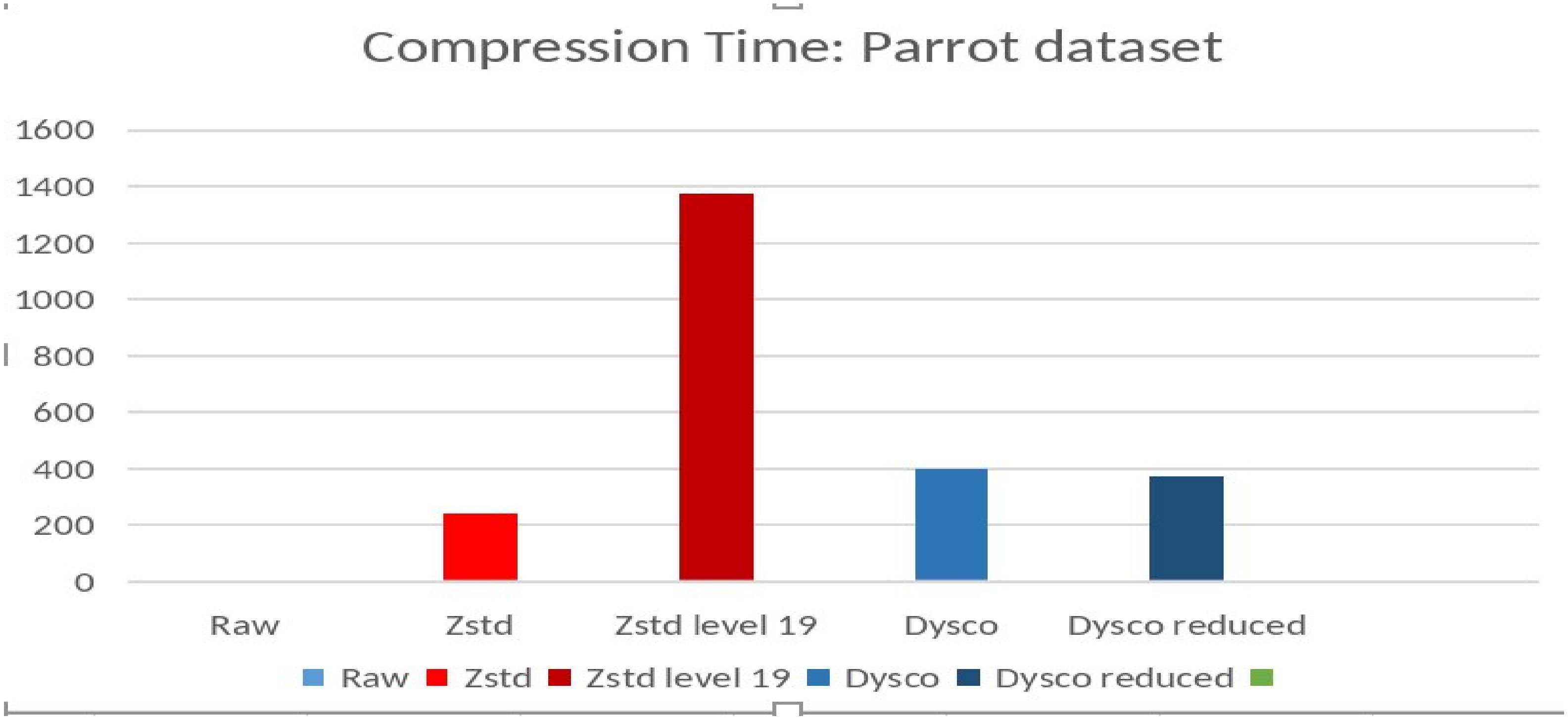
Classical Algorithms: Compression ratio



[1]: Parrot Dataset: Meerkat data for the PARROT publication https://arxiv.org/abs/2312.12165
411 GiB

ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE

Classical Algorithms: Compression time



[1]: Parrot Dataset: Meerkat data for the PARROT publication https://arxiv.org/abs/2312.12165
GiB

ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE

In summary

- Counter-intuitively, **Zstd** is worth to apply on **radio astronomy real world datasets** when losses can not be accepted
  - Should be integrated with our I/O toolkit (casacore)
- Dysco is a **very interesting solution** based on Quantization when losses are acceptable
- Dysco implementation can be optimised (GPU, threading, vectorization)
- Compression might speed up I/O compared to uncompressed read access

Learning-Based Compression Methods

- **Why Explore Learning-Based Compression Methods**
  - Can further reduce data size while preserving key scientific signals.
  - Enables flexible bit rates optimized with task-specific loss functions.
  - Adapts to complex data distributions and improves with real telescope data.

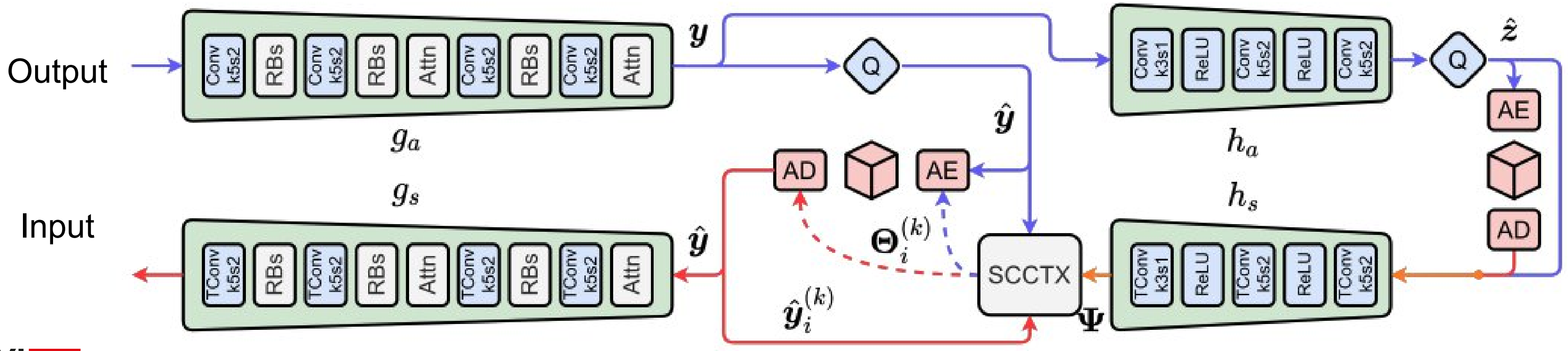
Learning-Based Compression Methods

- Visibility data sampled at each frequency can be treated as a multi-channel “image,” with channels representing polarizations.
- This allows the use of advanced **image compression methods**, leveraging correlations across channels.
- Examples include **ELIC** (autoencoder-based)[1] and **COIN** (implicit neural representations)[2].
- Methods can be optimized for specific datasets while preserving scientific fidelity.

[1] He, Dailan, et al. "Elic: Efficient learned image compression with unevenly grouped space-channel contextual adaptive coding." *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2022.
[2] Dupont, Emilien, et al. "Coin: Compression with implicit neural representations." *arXiv preprint arXiv:2103.03123* (2021).

Learning-Based Compression Methods

Efficient Learned Image Compression (ELIC) (He et al., 2022)
- Training objective: $L = R(\hat{y}) + \lambda D(x, g_s(\hat{y}))$, where $R(\hat{y}) = -E[\log p_{\hat{u}}(\hat{y})]$;
- Modelling latent codes: $p_{y|z}(y|z) = [N(\mu, \sigma^2) * U(-0.5, 0.5)]$;

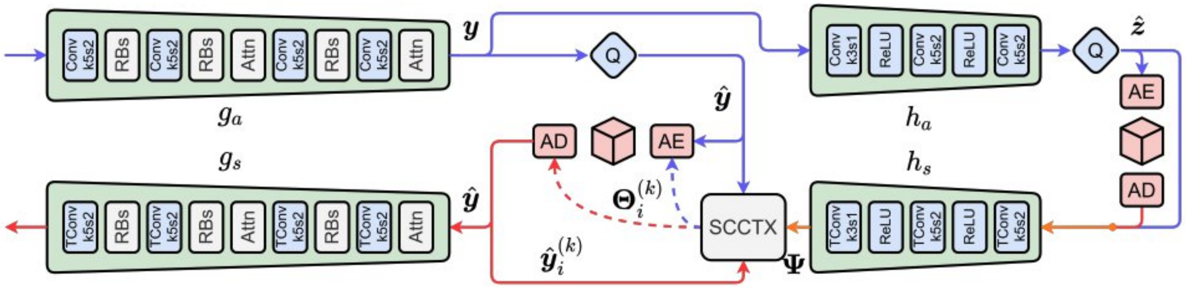


Learning-Based Compression Methods

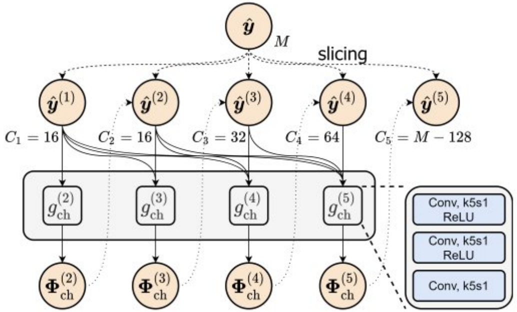
Efficient Learned Image Compression (ELIC) (He et al., 2022)

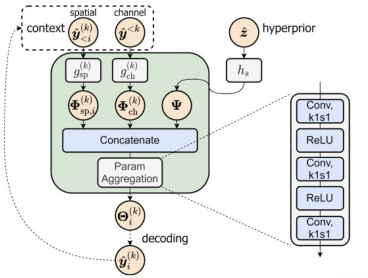
Input

Output



Uneven grouped context to speed up encoding

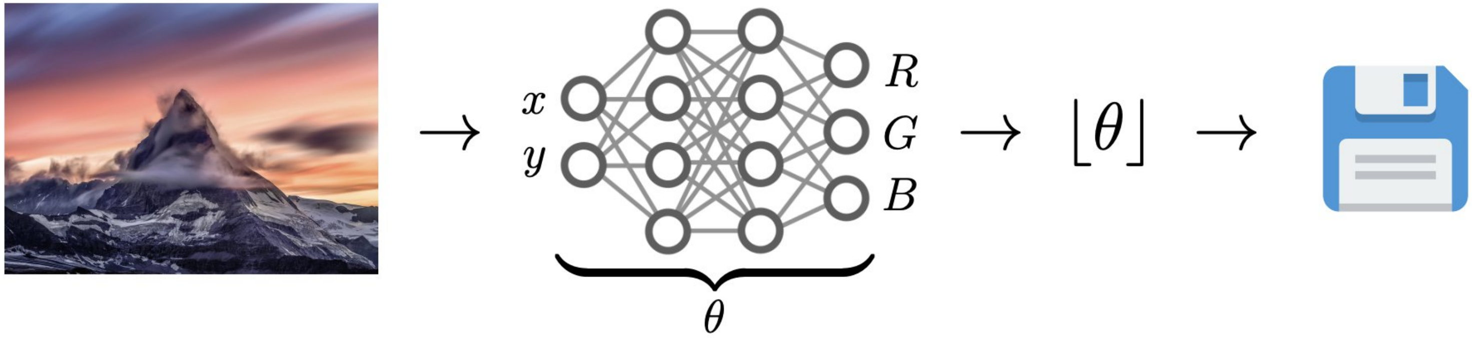




Parallel backward adaption coding in the spatial and channel axes

Learning-Based Compression Methods

- COIN (Dupont et al., 2021), Implicit neural representation model for data compression without entropy coding.
- Represents the signal as a neural network rather than storing explicit values.
- Data values are reconstructed by querying coordinates through **MLP** and **SIREN layers**.

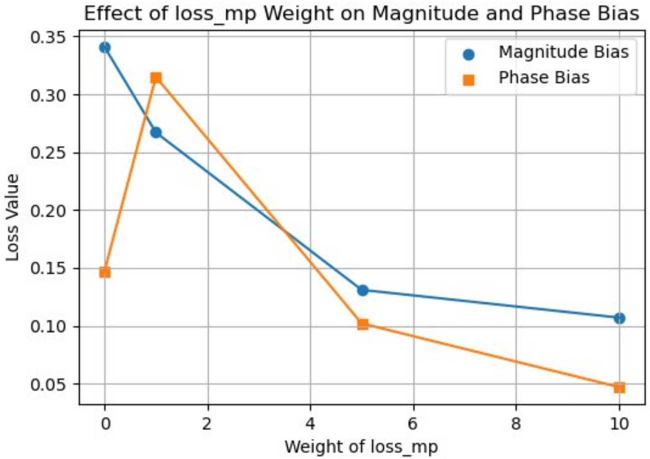


ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE

Learning-Based Compression Methods

- Compressing **synthetic data** per-frequency visibility data with ELIC
- Loss = λ_1*recons_loss + bpp_loss + λ_2* $loss_{mp}$

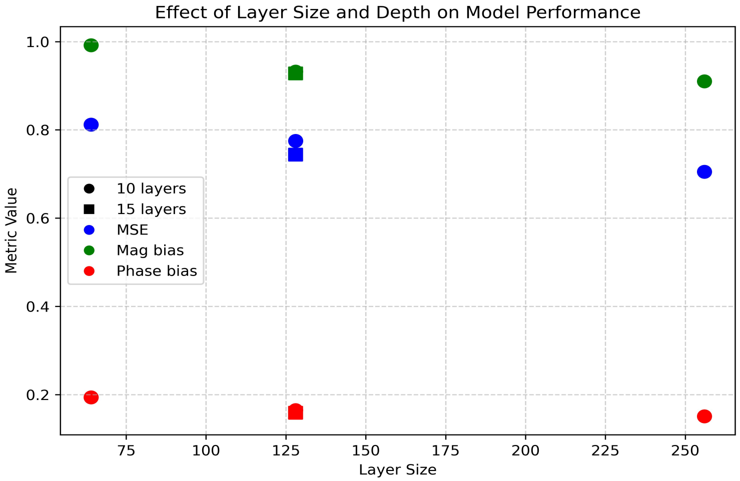
	MSE	BPP	Mag. bias	Phase bias
λ_1=1, λ_2=0	0.274	**0.121**	0.341	0.147
λ_1=1, λ_2=1	0.201	0.139	0.267	0.315
λ_1=1, λ_2=5	0.231	0.196	0.131	0.102
λ_1=1, λ_2=10	**0.108**	0.271	**0.107**	**0.047**



Learning-Based Compression Methods

- Compressing **synthetic data** per-frequency visibility data with COIN

	Laye r size	#layer	MSE	Mag bias	Phase bias (rad)
λ_1=1, λ_2=1	64	10	0.812	0.992	0.194
λ_1=1, λ_2=1	128	10	0.775	0.932	0.165
λ_1=1, λ_2=1	256	10	**0.705**	**0.910**	**0.151**
λ_1=1, λ_2=1	128	15	0.744	0.928	0.159



ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

Learning-Based Compression Methods

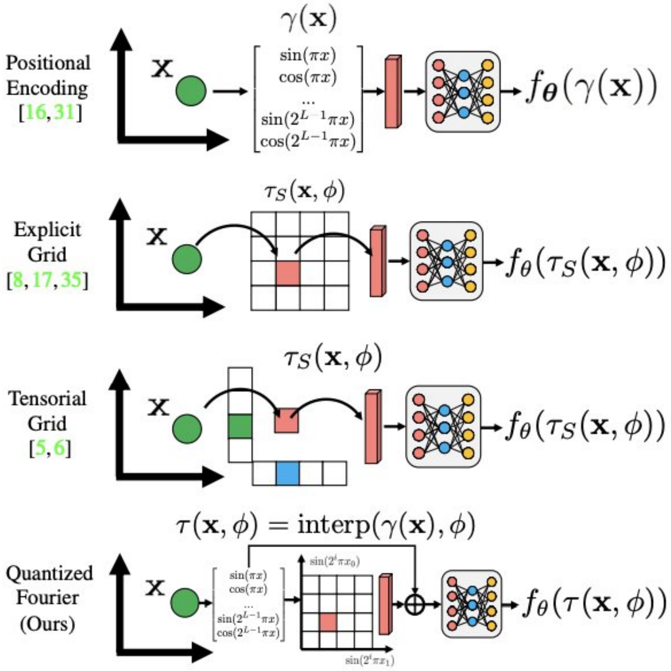
- Compressing **real data** with ELiC and COIN

Methods	**MSE**	**Mag bias**	**Phase bias (rad)**	**Compression rate**
DYSCO (Offringa)	**0.001**	**0.026**	**0.160**	34.31%
ELIC (He et al.) – per frequency	0.039	0.084	0.561	**0.11%**
ELIC (He et al.) – per frequency + per antenna	0.011	0.031	0.396	3.00%
COIN (Dupont et al.) – per frequency	0.009	0.031	0.779	0.16%
COIN (Dupont et al.) – per frequency + per antenna	0.005	0.064	0.575	8.76%

Learning-Based Compression Methods

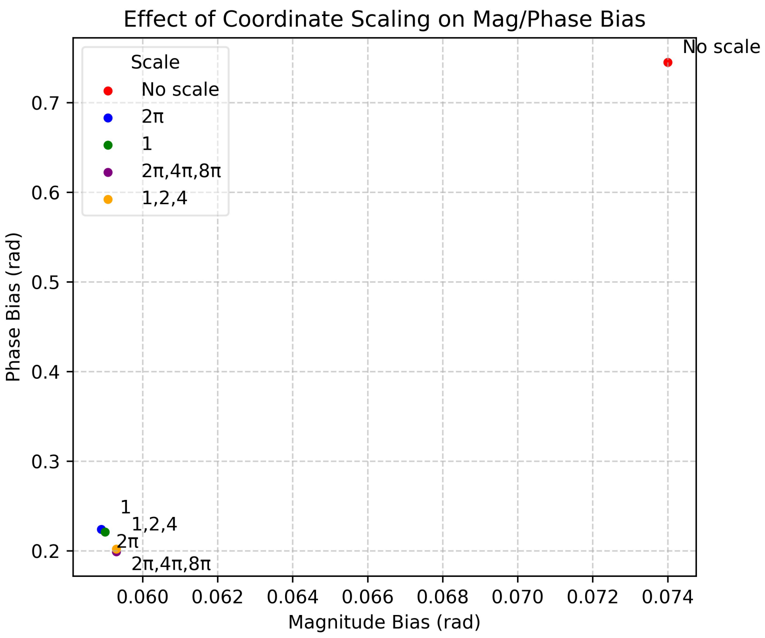
Positional encoding for coordinates

- Raw coordinates has difficulty handling oscillatory data;
- Encoding coordinate space for the data;
    - Less straight-forward for ELIC, but existing related work from neural field represenation [1] for INR



Learning-Based Compression Methods

Coordinates	Scale	Layer size	#layers	Mag/phase bias (rad)
Raw	/	128	10	0.0740 / 0.745
Encoded	2π	128	10	**0.0589** / 0.224
Encoded	1	128	10	0.0590 / 0.221
Encoded	$[2\pi, 4\pi, 8\pi]$	128	10	0.0593 / **0.199**
Encoded	[1, 2, 4]	128	10	0.0593 / 0.202



Adrien Devresse, Xiaoran Chen

Thank you!

ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE