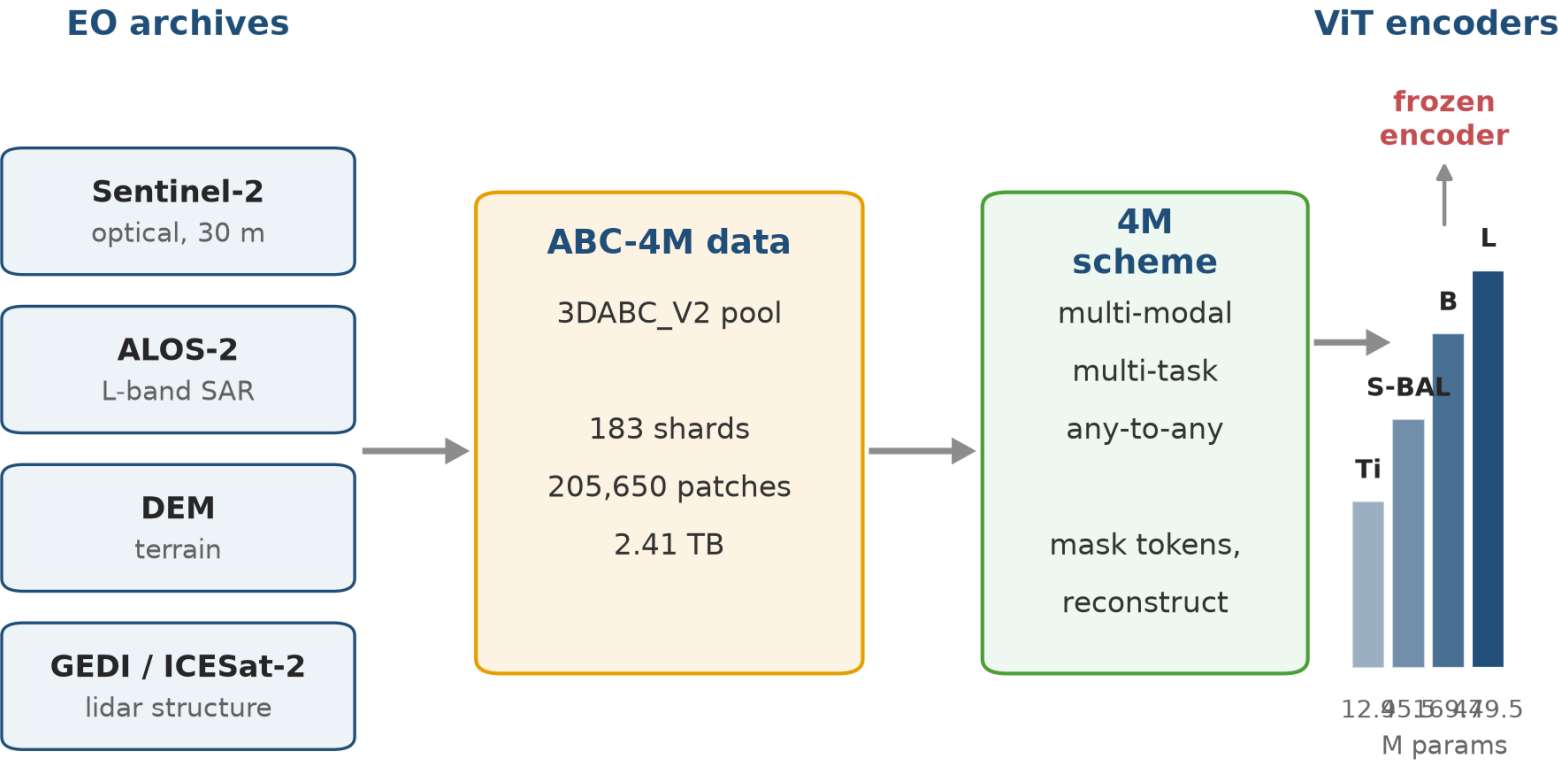


Amortized self-supervised pretraining for compute-efficient environmental monitoring: How well do EO foundation models transfer?

3D-ABC | ABC-4M pretraining on JUPITER Booster

Multi-modal EO Dataset in the Amazon, a 4M-trained ViT product



pretrained on JUPITER Booster (GH200)

how to read

left to right: the archives we pull from, the data they build, the training scheme, the encoders out

4M is the scheme - multi-modal and multi-task, any-to-any - it is not a count of four million anything

ABC-4M data - the parent corpus; the pool we actually trained on is the 3DABC_V2 cut of it

205,650 patches is the pool we ran, not the corpus - ABC-4M is roughly an order of magnitude larger

the objective is masked token reconstruction, so no labels enter pretraining anywhere

q240 / q480 = the GEDI keying the canopy target is gridded at, 240 m or 480 m

bar height is log in parameters: Ti to L is 35x for one shared recipe

the encoders are then frozen and handed to the downstream benchmarks on the next slides

Current Status of the 3DABC + Coreset Experiments

3 +

downstream benchmarks across various
domains judge the FM

3rd

of 28 models

frozen encoder on Biomasters, one
harness, no per-model tuning

4.4x

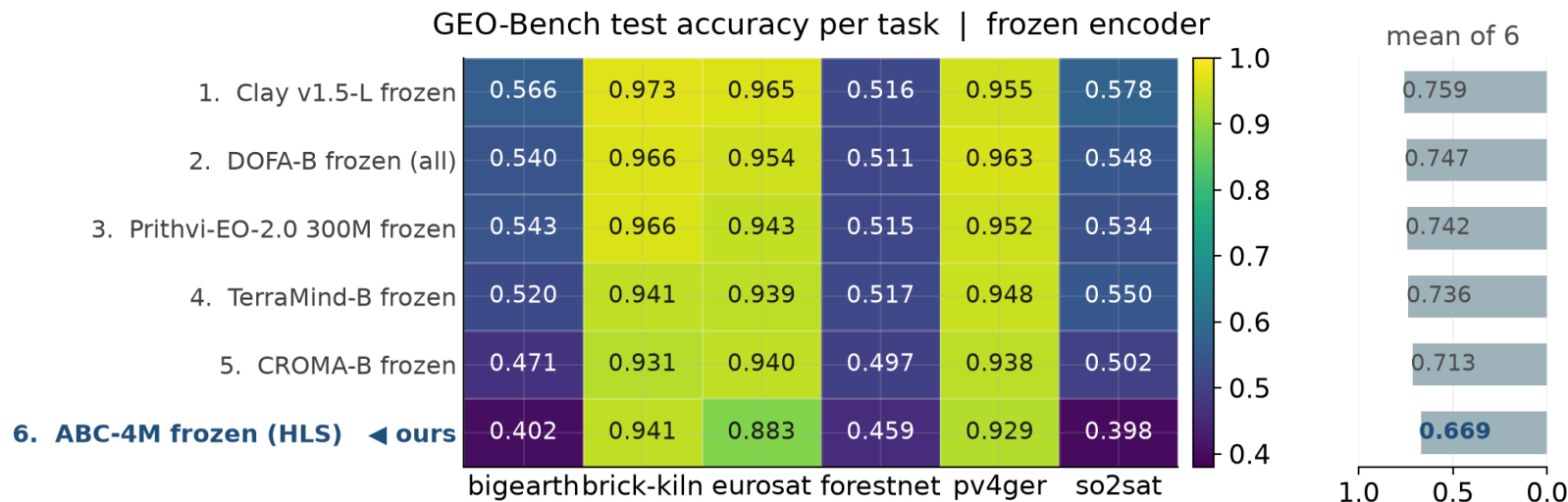
more energy for the same 107,900 steps at
8 nodes than at 1

Progress of Downstream Benchmarking and Comparison



Two downstream benchmarks, ability to compare against published encoders

Frozen, we are last of the six frozen encoders



how to read

frozen encoders only - no fine-tuned model appears in this table, so the ranking is protocol-matched

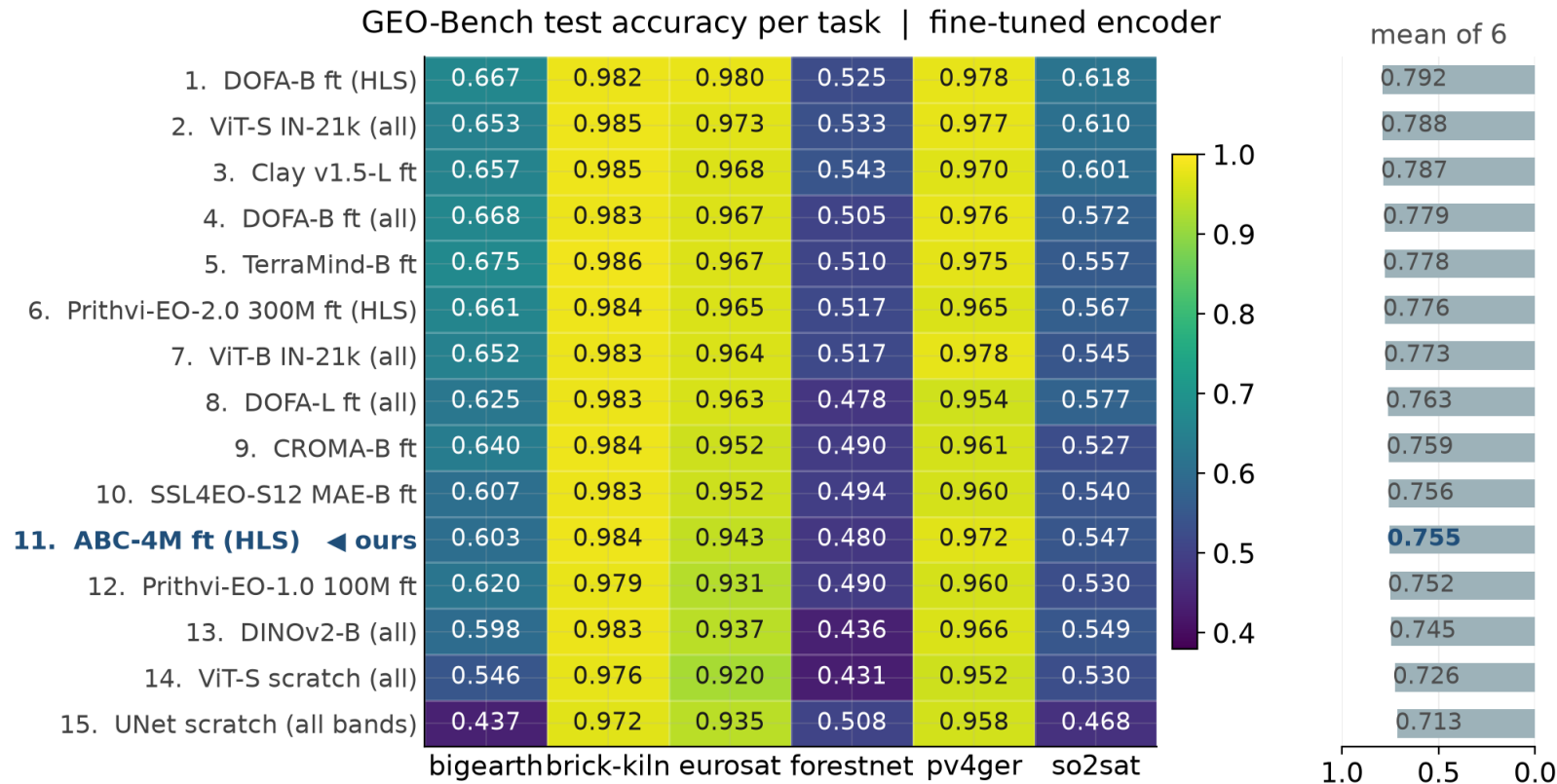
rows ranked by mean of the 6 tasks, best at the top; n/a = that model was not run on that task

ours - ABC-4M frozen, the only blue row

ours is last of the 6 frozen encoders, and last on every task here except brick-kiln

6 of GEO-Bench's 12 tasks, single seed per cell

Fine-tuned we are mid-field: 11th of 15



how to read

fine-tuned and trained-from-scratch only - no frozen encoder appears here; this is the other half of the split on the previous slide

every row ran all 6 tasks, so the mean bar and the rank are comparable top to bottom

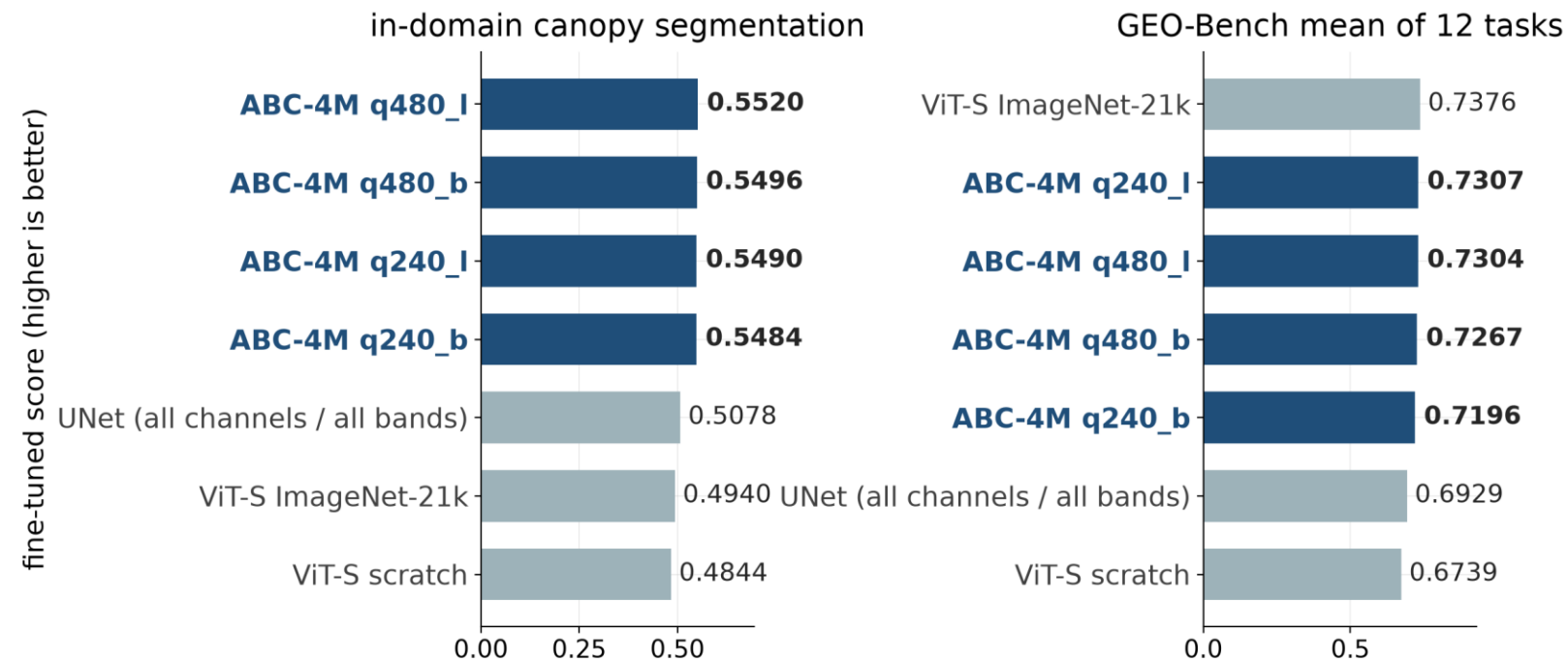
rows ranked by mean of the 6 tasks, best at the top

ours - ABC-4M fine-tuned on the 6-band HLS input recipe

mid-field, not last - the gap to the top is small next to the spread in the frozen table

excluded: HLS+extra and the RGB / HLS-6 variants, which ran only 4 or 5 of the 6 tasks and so cannot be ranked here

We lead in-domain; an ImageNet ViT edges us on GEO-Bench



■ **ours - ABC-4M q240/q480, L and B size**

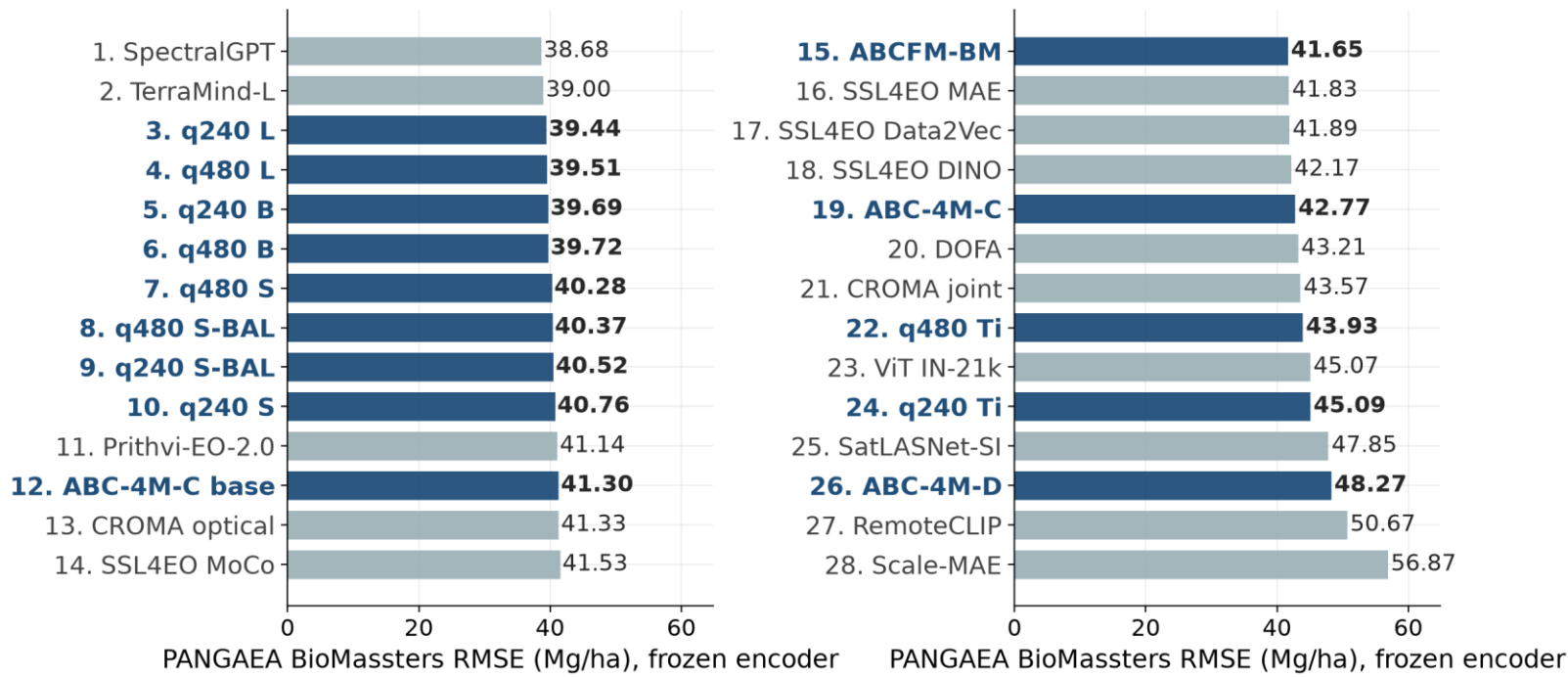
■ **external baselines in the same harness**

both panels are fine-tuned scores, higher is better, each with its own axis

U-Net here is trained from scratch with labels on the target task; ViT-S is pretrained on ImageNet-21k, not on Earth observation

we lead in-domain; ImageNet ViT-S edges us on GEO-Bench. PANGAEA is on the two ranking slides

Frozen Encoder Results



ours - ABC-4M variants

other foundation models

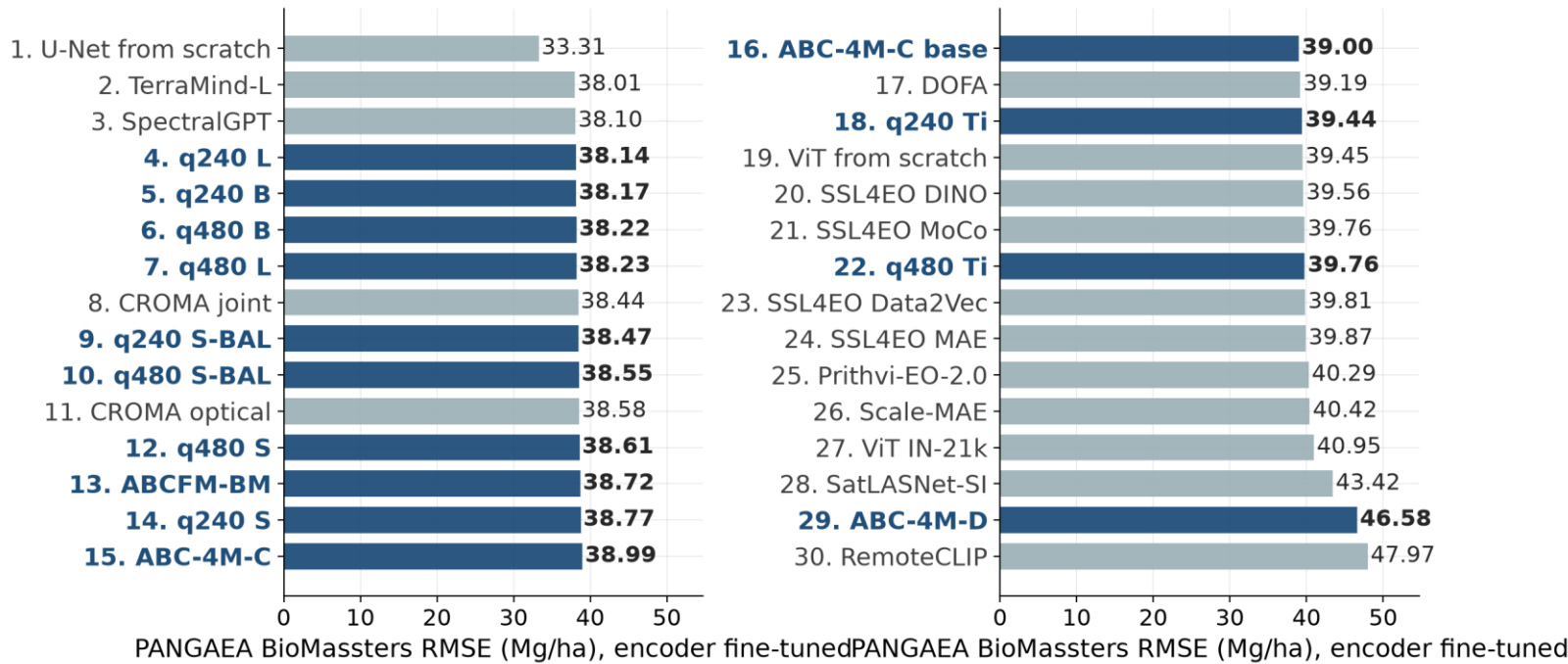
rank 1 is top-left, read down the left column then the right column

frozen encoder + one identical UPerNet head. No encoder was tuned, so this measures the representation itself

lower bar is better. One seed per cell, seed 234

ABC encoders run at 384 px, the external zoo at 256 px, token-matched - not pixel-matched

Fine-tuned Encoder Results



■ ours - ABC-4M variants
■ other FMs and from-scratch nets

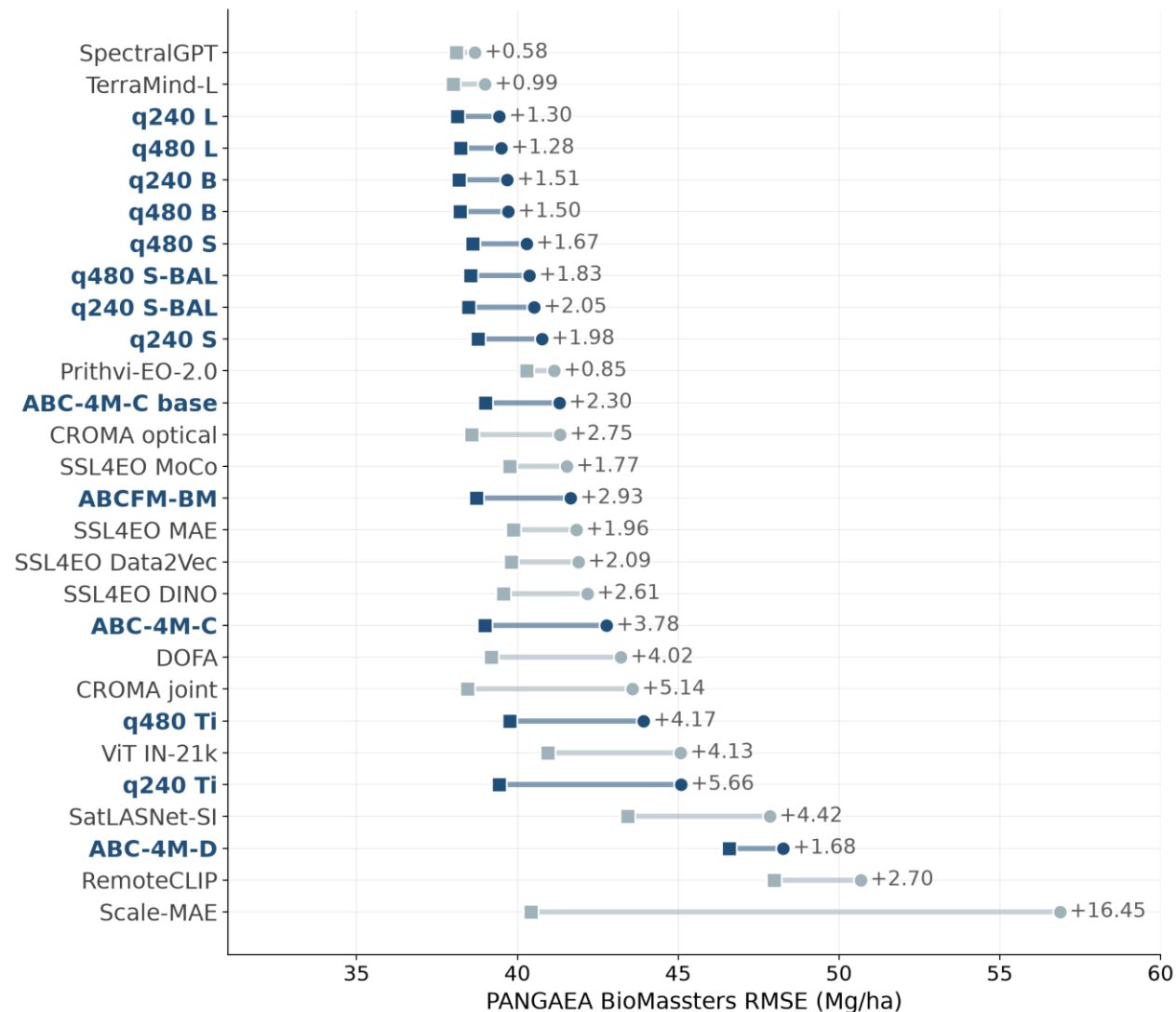
rank 1 is top-left, read down the left column then the right column

top bar is U-Net trained from scratch ON THIS TASK with labels - it has no frozen row because it has no pretrained encoder to freeze

so the frozen-vs-fine-tuned gap is not a like-for-like race: U-Net spends supervised labels the FMs never saw

best pretrained encoder here is TerraMind-L, then SpectralGPT - both within 0.13 of our q240-L

A short frozen-to-tuned gap means the features were already useful



● circle = frozen RMSE
 ■ square = fine-tuned RMSE
 — the gap fine-tuning closes

number at right = RMSE gained by fine-tuning. A SHORT bar means the pretrained features were already useful frozen

sorted by frozen RMSE, best at top

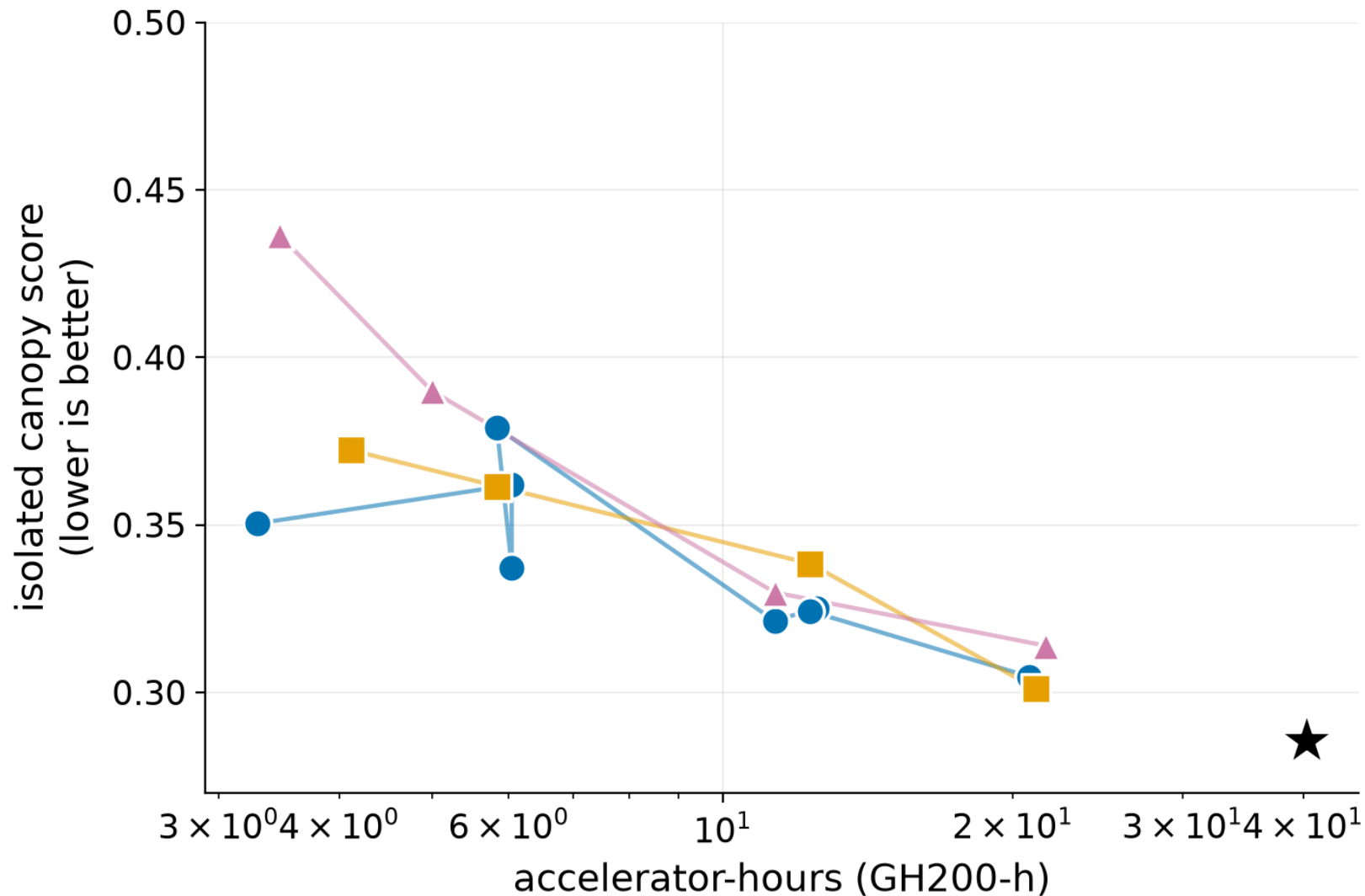
U-Net appears only fine-tuned - it has no pretrained encoder, so it cannot be frozen

Coresets



Can a fraction of ABC-4M train a foundation model as well as all of it?

Every coreset arm sits below and left of full data

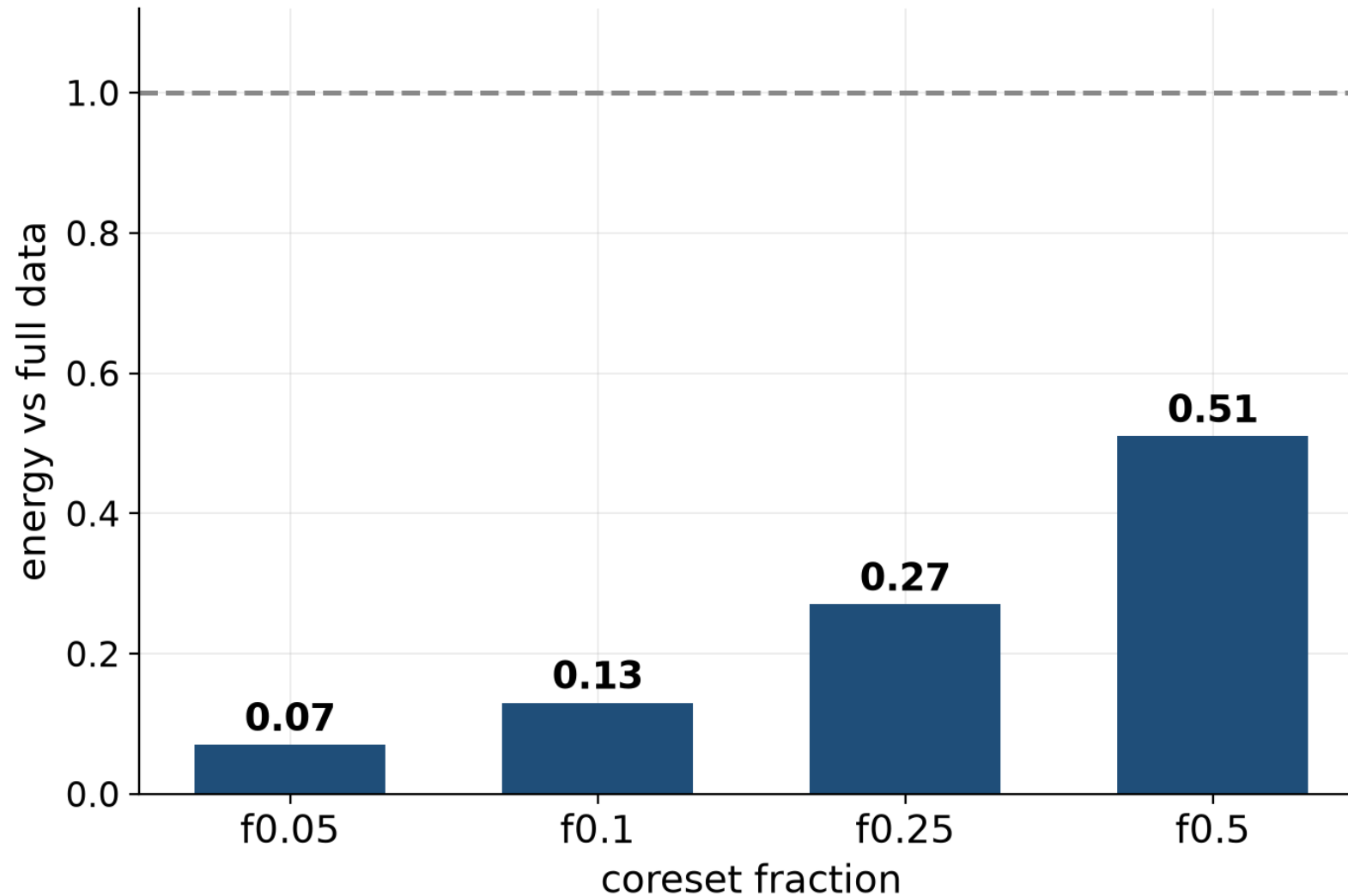


how to read

- random - one point per coreset fraction, joined by fraction
- stratified zone - same, sampled across climate zones
- ▲ first - same, ordered truncation
- ★ full data - the 100 % reference arm - up and left of it is the goal

x axis is log-scaled - equal spacing is a x10 step in energy

Energy falls almost linearly with fraction



how to read

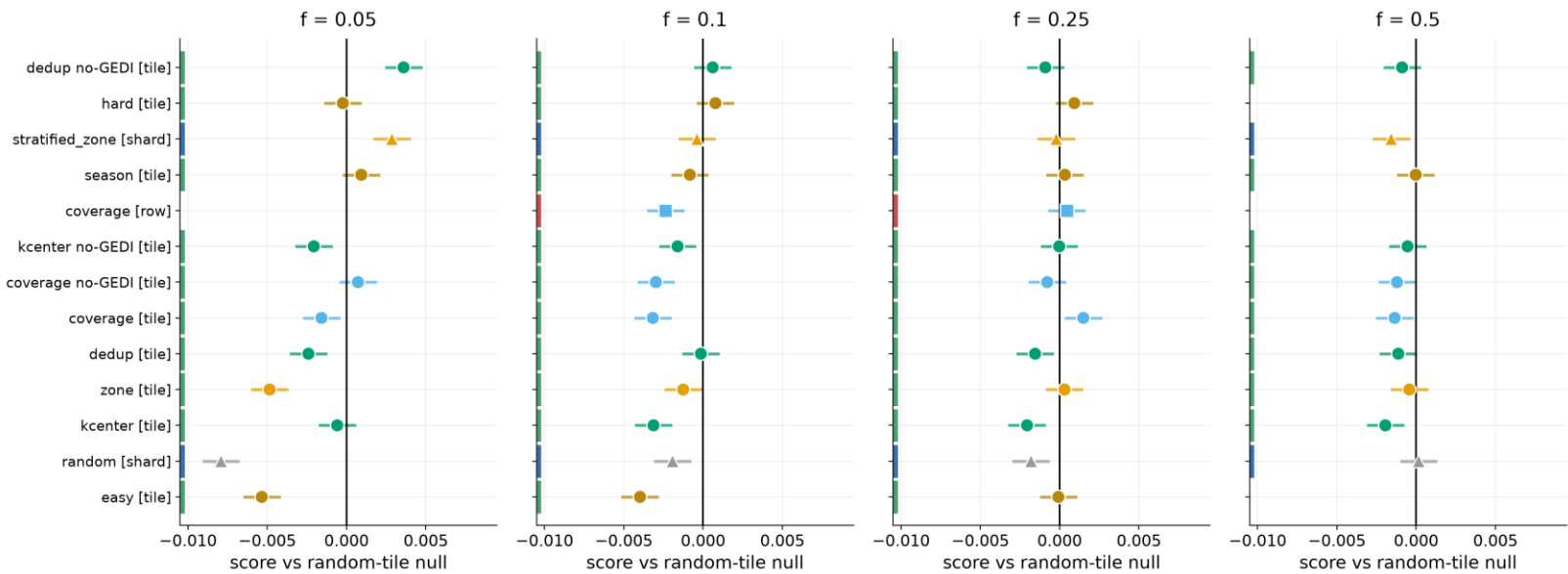
bar = energy used - measured kWh for that fraction, as a multiple of the full-data run

dashed = full data (1.00x) - the f1.0 reference arm, 20.8 kWh

numbers on the bars are measured ratios, not modelled ones

equal-epoch rule: each arm sees its own data once, so this is the rule that actually saves energy

Selection signal matters at extreme compression - and nowhere else



rows = same arms in all 4 panels,
best average at top; y has no
units

a row missing a marker in a panel
= that arm was not run at that
fraction

x = 0 = random-tile null - right
of it beats random

bar = 95 % CI - 7 seeds/arm,
df=318

colour = selection signal

- coverage
- redundancy
- geography
- season
- none (random)
- difficulty

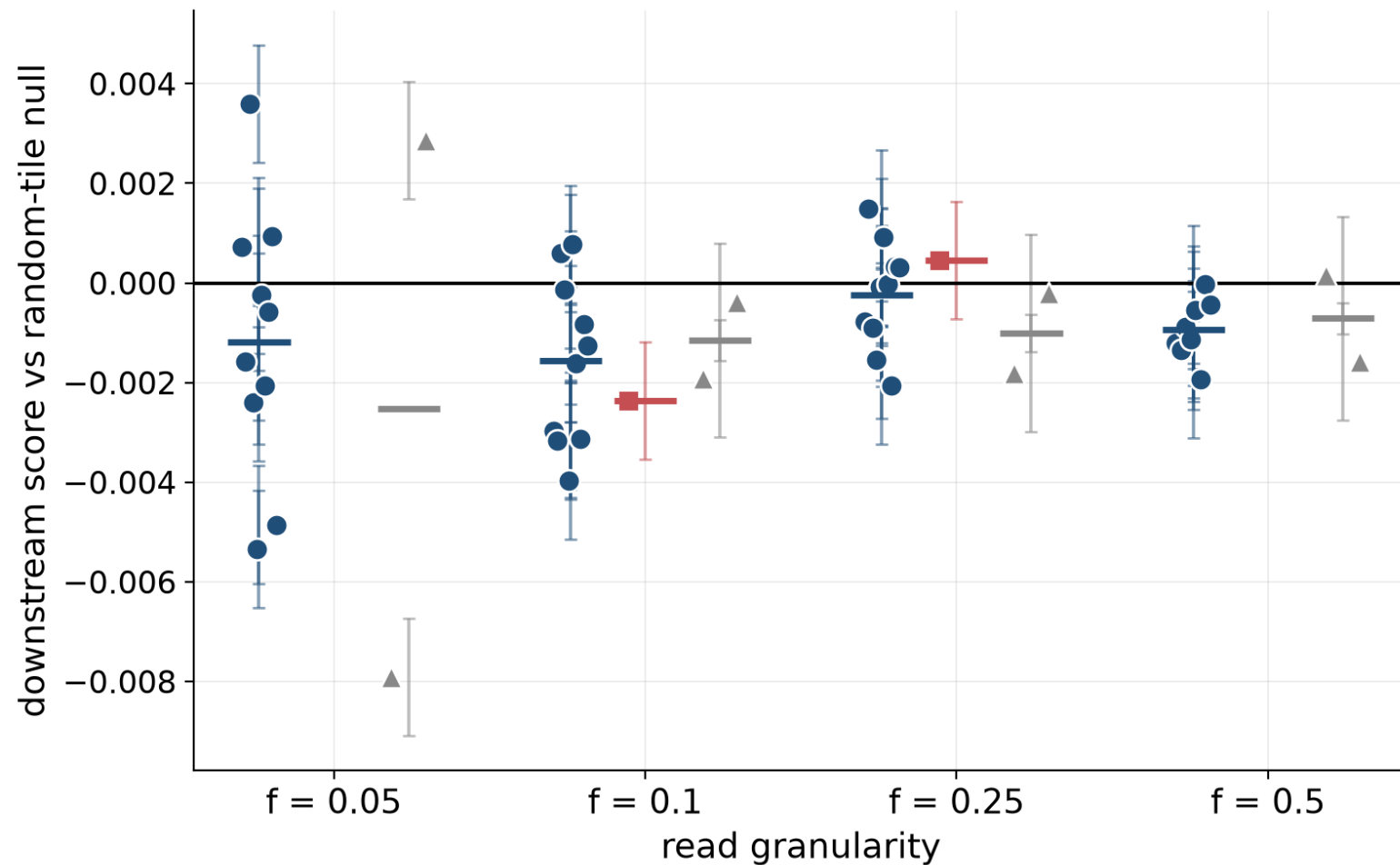
marker + left band = granularity

- tile [t] - cheap
- row [r] - ~20x reads
- shard [s] - uneven

'no-GEDI' = leak-FREE twin (GEDI
descriptors removed)

the canopy target is itself GEDI-
derived, so a GEDI rule can win on
exposure - read each rule with its
twin

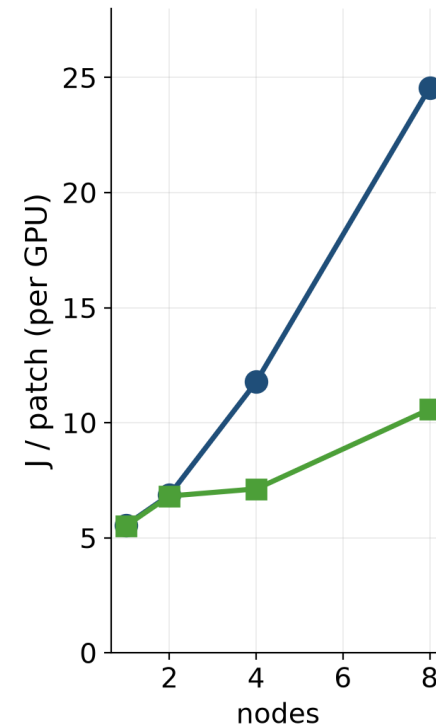
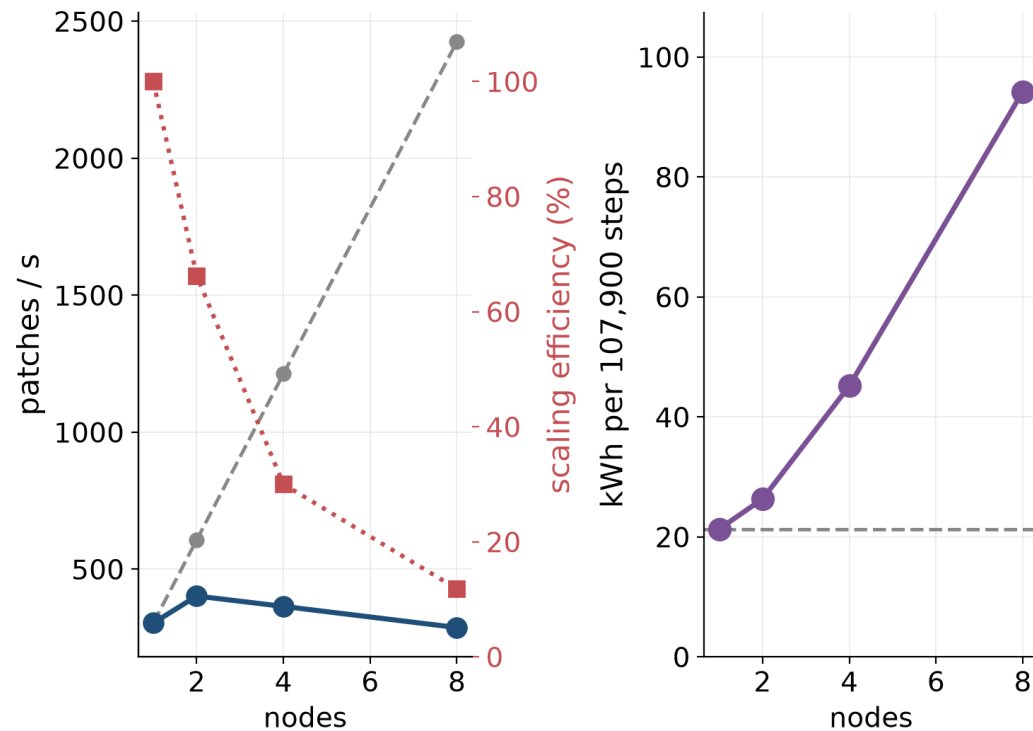
Granularity moves cost far more than selection method



- **tile - whole tiles, cheap reads**
- **row - ~20x read amplification**
- ▲ **shard - not strictly equal-budget**
- **thick bar = group mean**
- **whisker = 95 % CI per arm**
- | **zero = random-tile null**

only ONE row strategy exists (coverage_row), so the row column is that arm, not a granularity average

Appendix A: Energy Scaling



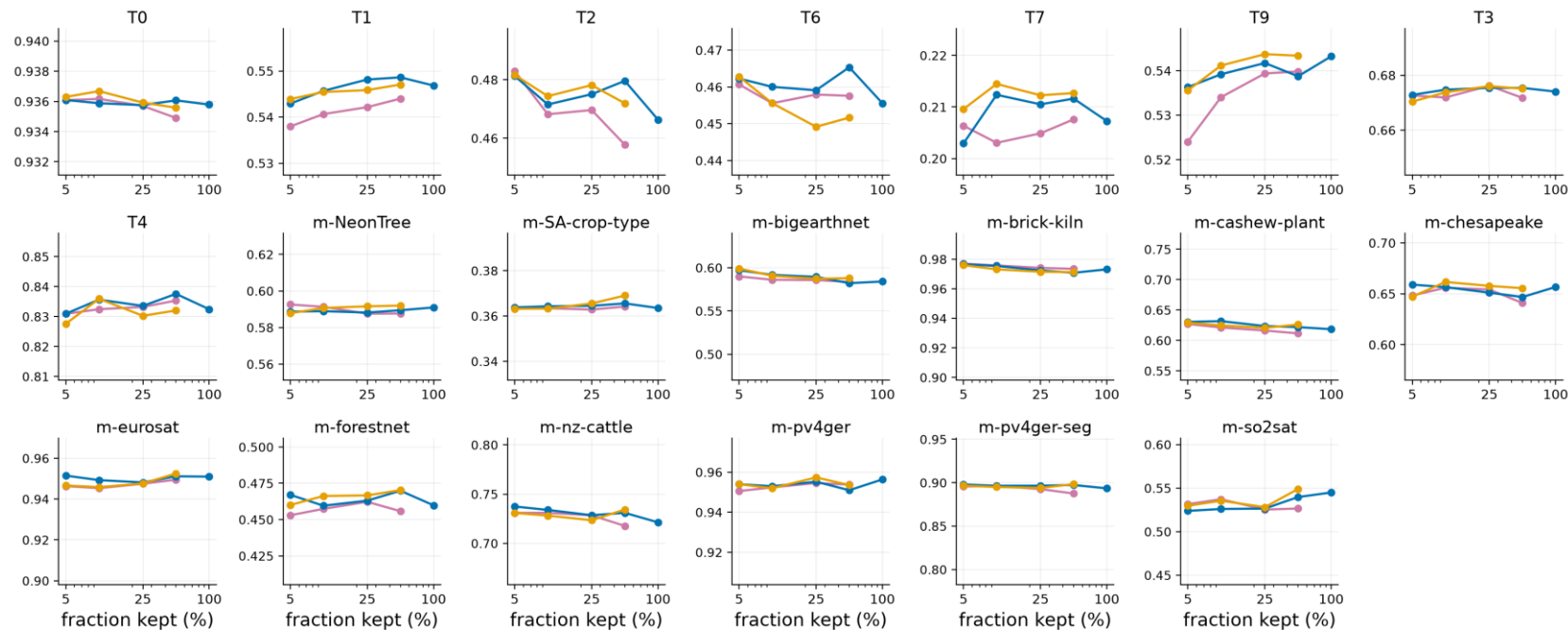
- total throughput / fixed global
- ideal throughput
- scaling efficiency, right axis
- metered kWh, same workload
- J/patch at fixed per-GPU batch

all three panels share the x axis:
node count, at a fixed global
batch unless stated

left: eight nodes deliver LESS
total throughput than one. middle:
the same steps cost 4.4x the
energy

right: holding per-GPU batch
instead makes the cost look flat -
but the global batch grew 8x, so
it is a different experiment

Appendix B: Iso-compute



ISO-COMPUTE (fixed 53,950 updates at every fraction), not equal-epoch: a 5% arm sees ~20 epochs, a 100% arm ~1. Fraction is confounded with epochs - read shapes only, not coreset-vs-full-data.

colour = source domain

in-domain - our own bench, 4-class land-cover segmentation (6 tasks)

out-of-domain - bioclimatic regression, shifted drivers (2 tasks)

GEO-Bench - external public benchmark, 12 classification tasks

line = selection rule

random - randomly kept shards

stratified zone - proportional across climate zones

first - ordered truncation of the corpus

each panel is one task on its own y axis - compare shapes within a panel, never heights across panels

ISO-COMPUTE (fixed 53,950 updates at every fraction), NOT equal-epoch: a 5% arm sees ~20 epochs, a 100% arm ~1. Fraction is confounded with epochs, so a coreset 'beating' full data here is not a selection result