

SynthStats demonstrations

SynthStats team

What a demonstration is

Task and observations

A task in English and 80 observations from simulated BoxingGym data.

Dugong length from age, and infected count from time.

Checks and predictions

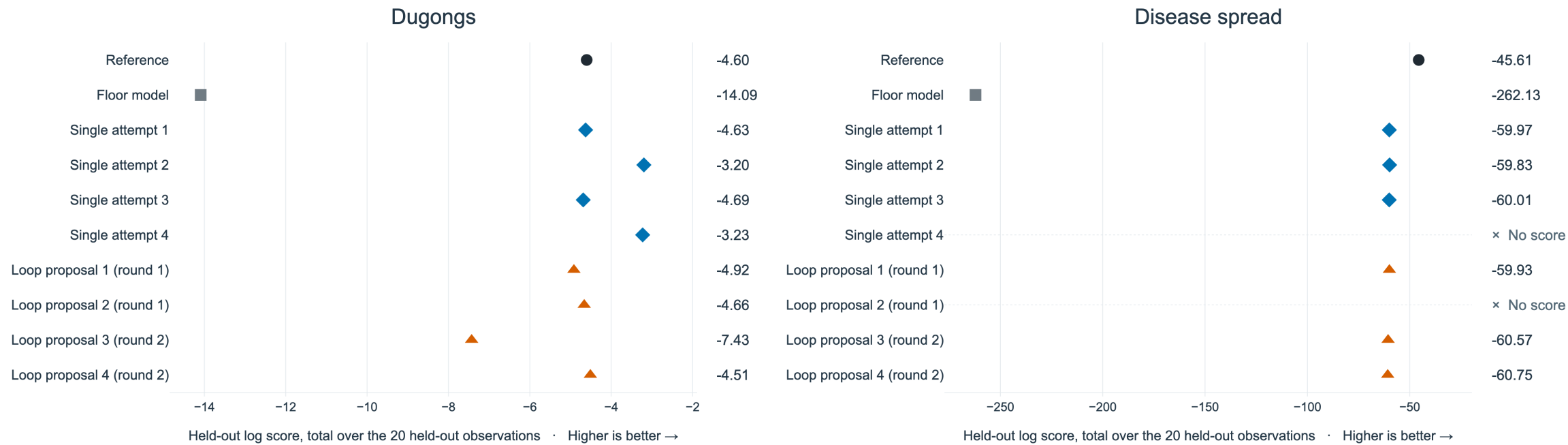
Each program is checked before fitting; those that clear the checks are fitted and scored on 20 held-out rows.

Ten candidate PyMC programs

- The reference
- The floor model, which ignores the input
- Four single attempts by Qwen3.6-35B-A3B
- Four loop proposals

The loop writes two programs, receives a critic's comments, then writes two more.

Held-out log score · all ten programs per environment



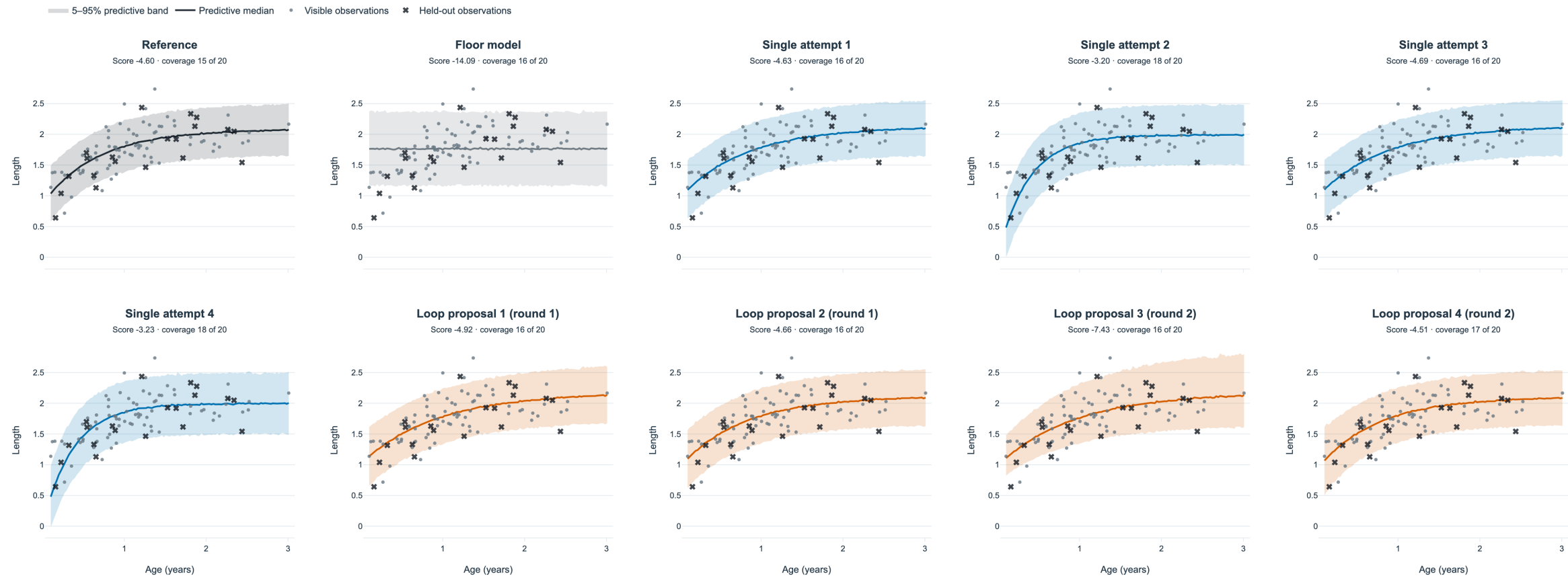
Each environment has its own score axis; compare programs within one environment. A dotted line marks a program without a score.
Data are simulated.

Dugongs: the reference program

```
import numpy as np
import pymc as pm

def model(data):
    """Dugong length against age: the growth curve and priors of BoxingGym's
    dugongs environment, with a Normal likelihood and a fixed noise scale.
    """
    age = np.asarray(data["age"], dtype=float)
    y = np.asarray(data["length"], dtype=float)
    with pm.Model() as dugongs_model:
        age_data = pm.Data("age_data", age)
        alpha = pm.Normal("alpha", mu=2.0, sigma=0.2)
        beta = pm.Normal("beta", mu=1.5, sigma=0.5)
        lam = pm.Normal("lam", mu=0.4, sigma=0.5)
        mu = alpha - beta * pm.math.abs(lam) ** age_data
        pm.Normal("length", mu=mu, sigma=0.25, observed=y)
    return dugongs_model
```

Dugongs · predictions from all ten programs

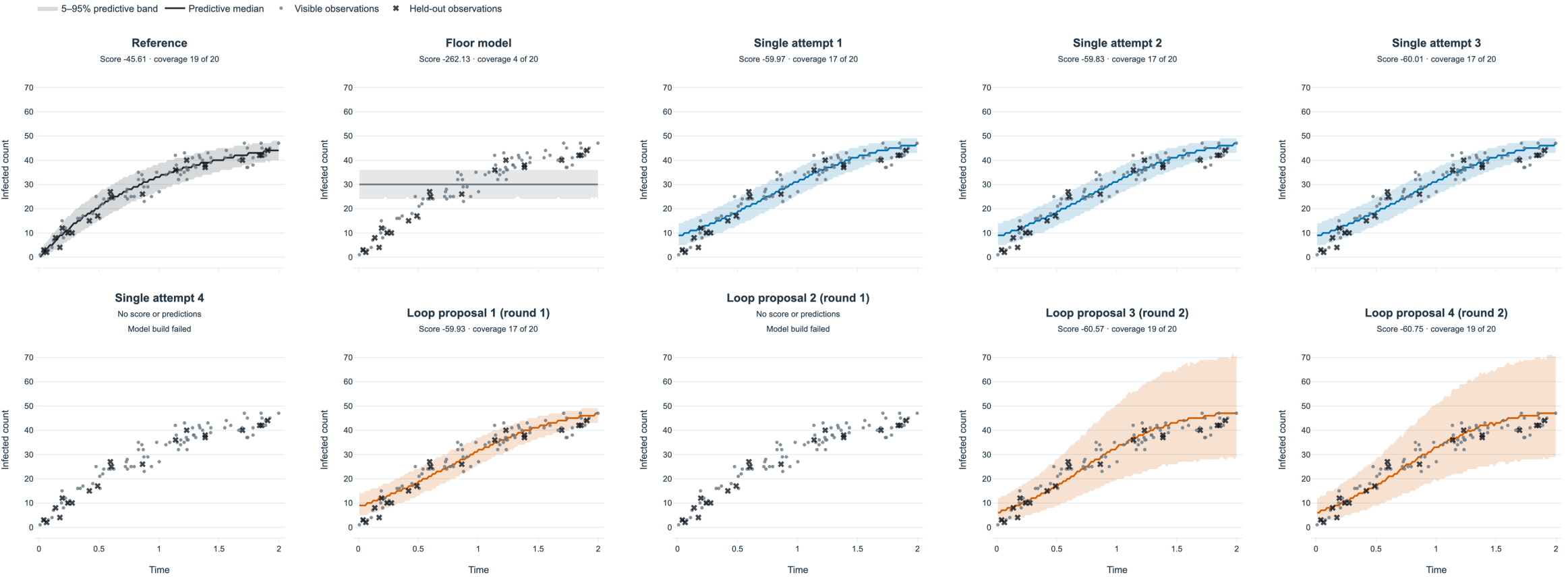


All panels share the same x and y ranges. Shading shows the 5th–95th percentiles of predictive draws.
Score: total held-out log score over the 20 held-out observations (higher is better). Coverage: held-out points inside the 5–95% band, bounds inclusive.
Data are simulated.

Dugongs: score and coverage

Program	Held-out log score	Coverage
Reference	-4.6028	15 of 20
Floor model	-14.0898	16 of 20
Single attempt 1	-4.6292	16 of 20
Single attempt 2	-3.1969	18 of 20
Single attempt 3	-4.6876	16 of 20
Single attempt 4	-3.2267	18 of 20
Loop proposal 1 (round 1)	-4.9166	16 of 20
Loop proposal 2 (round 1)	-4.6634	16 of 20
Loop proposal 3 (round 2)	-7.4306	16 of 20
Loop proposal 4 (round 2)	-4.5121	17 of 20

Disease spread · all ten programs



All panels share the same x and y ranges. Shading shows the 5th–95th percentiles of predictive draws.
Score: total held-out log score over the 20 held-out observations (higher is better). Coverage: held-out points inside the 5–95% band, bounds inclusive.
Data are simulated.

Disease spread: score and coverage

Program	Held-out log score	Coverage
Reference	-45.6085	19 of 20
Floor model	-262.1291	4 of 20
Single attempt 1	-59.9733	17 of 20
Single attempt 2	-59.8279	17 of 20
Single attempt 3	-60.0128	17 of 20
Single attempt 4	failed at model build	
Loop proposal 1 (round 1)	-59.9324	17 of 20
Loop proposal 2 (round 1)	failed at model build	
Loop proposal 3 (round 2)	-60.5719	19 of 20
Loop proposal 4 (round 2)	-60.7518	19 of 20

What the numbers mean

The score is the total log predictive density over 20 held-out rows. Higher is better within one task.

Coverage is the number of the 20 held-out observations inside the 90% predictive interval (5th to 95th percentile, bounds inclusive). The nominal coverage is 18 of 20.

Disease spread

Program	Held-out log score	Coverage
Single attempt 2	-59.8279	17 of 20
Loop proposal 3 (round 2)	-60.5719	19 of 20

Single attempt 2 has the higher score. Loop proposal 3 covers more held-out observations.