

A number alone says little. **85 kWh/m² per year** or **92% accuracy** only means something next to a reference: similar buildings, or other methods on the same test. **Benchmarking is comparing against a fixed reference, the same way for everyone.** In the built environment the word has two meanings.

MEANING A

Building performance benchmarking

"How does my building compare with buildings like it?"

- **Measure:** energy use intensity, $EUI = \text{annual energy} \div \text{floor area (kWh/m}^2\text{-yr)}$
- **Normalise:** for weather, use, size and hours, so the comparison is fair
- **Compare:** to a peer group of similar buildings, as a percentile or score (e.g. ENERGY STAR 1-100 in the US)
- **Used for:** disclosure rules, spotting the worst performers, choosing what to retrofit first

MEANING B

Model and method benchmarking

"Is my method really better, compared fairly?"

- **Same task** and **same data splits** for every method
- **Same metric** and **same rules** (what data and compute are allowed)
- **Baselines** to beat, from trivial to state of the art
- **Used for:** papers, model selection, tracking progress in a field

Anatomy of a method benchmark

Part	What it fixes	Building example
Task	The question to answer	Forecast next-day hourly electricity
Dataset + splits	Train, validation and a hidden test set	Train on 2016, test on 2017; test buildings unseen
Metric	How answers are scored	CV(RMSE) and NMBE
Protocol	The rules of the game	No weather data from the future
Baselines	What you must beat	"Tomorrow = same hour last week"
Leaderboard	Public, comparable results	A competition ranking

Common metrics

Task	Metrics
Forecasting	MAE, RMSE; CV(RMSE) and NMBE in building energy
Classification	accuracy, F1 score
Similarity search	precision@k: share of the top k that are truly alike
Control	energy cost and comfort violations together

REMEMBER

ASHRAE Guideline 14 calibration targets: hourly $CV(RMSE) \leq 30\%$ and NMBE within $\pm 10\%$; monthly $\leq 15\%$ and $\pm 5\%$.

Good practice

- **Beat a simple baseline first.** If "same as last week" wins, the model adds nothing.
- **Split by time and by place.** Neighbouring buildings leak information into the test set.
- **Run several seeds** and report the spread, not one lucky run.
- **Same budget** of data and compute for every method.
- **Share code and splits** so others can repeat the result.

Well-known benchmarks

Name	Tests
ImageNet	image recognition (2009)
GLUE, MMLU	language understanding and knowledge
GEPIII	ASHRAE energy prediction competition on Kaggle (2019)
BuildingsBench	energy forecasting; pretraining on simulated buildings
BOPTTEST	building control in simulation
CityLearn	multi-building demand response with RL

Pitfalls

- **Goodhart's law:** when a measure becomes a target, it stops being a good measure.
- **Overfitting the test set** by trying many versions against it.
- **Leakage:** future or test information sneaks into training.
- **Saturation:** everyone scores $\sim 100\%$, so the benchmark no longer separates methods.
- **Benchmark $\neq$ reality:** a leaderboard win may not survive deployment.

FOR OUR PROJECT · A MINI BENCHMARK

Task: "find buildings like this one". Labels: the V-World use code BPRP_SE. Metric: precision@5. Baselines: random pick, then size-only. Split by district so neighbours cannot leak.

SAY IT LIKE THIS

"A benchmark is a fixed exam: the same questions and the same marking scheme for every student."

IN ONE LINE

Benchmark = task + data + metric + rules + baselines.